Systems, methods, and apparatuses relating to in-network storage for a configurable spatial accelerator are described. In one embodiment, a configurable spatial accelerator includes a plurality of processing elements; a circuit switched interconnect network between the plurality of processing elements to receive an input of a dataflow graph comprising a plurality of nodes, wherein the dataflow graph is to be overlaid into the circuit switched interconnect network and the plurality of processing elements with each node represented as a dataflow operator in the plurality of processing elements, and the plurality of processing elements are to perform an operation by a respective, incoming operand set arriving at each of the dataflow operators of the plurality of processing elements; and an in-network storage element of the circuit switched interconnect network comprising a queue coupled to an output queue of a first processing element, and a controller that switches the in-network storage element into a first mode that provides a value stored in the queue of the in-network storage element by the output queue of the first processing element to an input queue of a second processing element when a configuration value is a first value, and into a second mode that bypasses the queue of the in-network storage element and provides a value from the output queue of the first processing element to the input queue of the second processing element when the configuration value is a second value.
FIG. 1

MEMORY INTERFACE 102

INTERCONNECT NETWORK 104

IN-FABRIC STORAGE

INTEGER PROCESSING ELEMENT (P.E.)

COMMUNICATION CIRCUITRY (E.G., NETWORK DATAFLOW ENDPOINT CIRCUIT)
PROCESSOR 200

CORE 0  ...  CORE N

ACCELERATOR 0  ...  ACCELERATOR M

MEMORY 202

FIG. 2
```c
void func ( int x , y ) {
    x = x * y;
    return x;
}
```
void memcpy(void *A, void *B, int N) {
    for(int index = 0; index < N; index++) {
        a[index] = b[index]
    }
}
MEMORY/CACHE HIERARCHY INTERFACE

Fig. 5
if (CondDeq && NotEmpty) {
    if (Tail + 1 > capacity) {
        Tail <= 0;
    } else {
        Tail <= Tail + 1
    }
}

if (Ready && Valid) {
    if (Head + 1 > capacity) {
        Head <= 0;
    } else {
        Head <= Head + 1
    }
}
Count Determiner:
\[ \text{Count} \leq \text{Count} + (\text{Ready} \&\& \text{Valid})^* - (\text{CondDeq} \&\& \text{NotEmpty})^* \]
*Assumes conversion of boolean value true to integer 1 and false to integer 0

FIG. 16

Enqueue = Valid \&\& Ready

FIG. 17

Not Full (Ready) = \text{Count} < \text{CAPACITY}

FIG. 18

Not Empty = \text{Count} > 0

FIG. 19

Valid = \text{Count} > 0

FIG. 20
if (Ready && Valid) {
    if (NotFull && CondEng) {
        if (Tail + 1 > capacity) {
            Tail <= 0;
        } else {
            Tail <= Tail + 1
        }
    } else {
        Head <= Head + 1
    }
} else {
    Head <= 0;
}
Count Determiner:
\[ \text{Count} \leq \text{Count} + (\text{NotFull} \land \text{CondEnq})^* - (\text{Ready} \land \text{Valid})^* \]
*Assumes conversion of boolean value true to integer 1 and false to integer 0*

**FIG. 26**

\[ \text{Enqueue} = \text{Valid} \land \text{Ready} \]

**FIG. 27**

\[ \text{NotFull (Ready)} = \text{Count} < \text{CAPACITY} \]

**FIG. 28**

\[ \text{NotEmpty} = \text{Count} > 0 \]

**FIG. 29**

\[ \text{Valid} = \text{Count} > 0 \]

**FIG. 30**
if ( Ready && NotEmpty ) {
  if ( Queue Ready && Valid ) {
    if ( Head + 1 > capacity ) {
      Head < = 0;
    } else {
      Head < = Head + 1;
    }
  } else {
    Tail < = Tail + 1;
  }
}

if ( Tail + 1 > capacity ) {
  Tail < = 0;
} else {
  Tail < = Tail + 1;
}
Count Determiner:
Count <= Count + (QueueReady && Valid) - (Ready && NotEmpty)*
*Assumes conversion of boolean value true to integer 1 and false to integer 0

FIG. 38

Enqueue = Valid && QueueReady

FIG. 39

QueueReady (Not Full) = Count < CAPACITY

FIG. 40

QueueValid (Not Empty) = Count > 0

FIG. 41
STORING A FIRST CONFIGURATION VALUE IN A CONFIGURATION REGISTER WITHIN A FIRST PROCESSING ELEMENT THAT CAUSES THE FIRST PROCESSING ELEMENT TO PERFORM AN OPERATION ACCORDING TO THE FIRST CONFIGURATION VALUE, AND A SECOND CONFIGURATION VALUE IN A CONFIGURATION REGISTER WITHIN A SECOND PROCESSING ELEMENT THAT CAUSES THE SECOND PROCESSING ELEMENT TO PERFORM AN OPERATION ACCORDING TO THE SECOND CONFIGURATION VALUE

COUPLING A QUEUE OF AN IN-NETWORK STORAGE ELEMENT, OF A CIRCUIT SWITCHED INTERCONNECT NETWORK BETWEEN THE FIRST PROCESSING ELEMENT AND THE SECOND PROCESSING ELEMENT, TO AN OUTPUT QUEUE OF A PLURality OF OUTPUT QUEUES OF THE FIRST PROCESSING ELEMENT

CONTROLLING ENQUEUE AND DEQUEUE OF VALUES INTO THE PLURality OF OUTPUT QUEUES OF THE FIRST PROCESSING ELEMENT ACCORDING TO THE FIRST CONFIGURATION VALUE WITH AN OUTPUT CONTROLLER IN THE FIRST PROCESSING ELEMENT

CONTROLLING ENQUEUE AND DEQUEUE OF VALUES INTO A PLURality OF INPUT QUEUES OF THE SECOND PROCESSING ELEMENT ACCORDING TO THE SECOND CONFIGURATION VALUE WITH AN INPUT CONTROLLER IN THE SECOND PROCESSING ELEMENT

STORING A THIRD CONFIGURATION VALUE IN A CONFIGURATION REGISTER WITHIN THE IN-NETWORK STORAGE ELEMENT


FIG. 45
int foo(int *a, int *b);
return (*a) * (*b);
**BASIC NETWORK CONFIGURATION**

SEND:  
- DEST: 5102A
- CHANNEL: 5102B
- INPUT: 5102C

RECV:  
- OUTPUTS: 5104

**FIG. 51**

**NETWORK PACKET FORMAT**

SEND:  
- TYPE: 5202A
- DESTINATION: 5202B
- CHANNEL: 5202C
- INPUT (E.G., PAYLOAD): 5202D

**FIG. 52**

**HIGH-RADIX CONFIGURATION**

SEND (E.G., SWITCH):  
- DEST: 5302A
- CHANNEL: 5302B
- INPUT(S): 5302C
- OP: 5302D

RECV (E.G., PICK):  
- OUTPUT(S): 5304A
- INPUT(S): 5304B
- OP: 5304C

**FIG. 53**

SEND CONTROL:
- DEST: 5402A
- CHANNEL: 5402B
- INPUT: 5402C

**FIG. 54**
SWITCH ANY CONTROL:

INPUT(S) (E.G., VALID BITS) 5702A → 5702B

SELECTION OP

INPUT STATUS

CREDIT STATUS

PERFORMABLE

SELECTION DECODER 5706

MUX SELECT BITS

FIG. 57

PICK CONTROL:

OUTPUT(S) 5802A → INPUT(S) 5802B → 5802C

OP 5802

INPUT (E.G., NETWORK INGRESS BUFFER) STATUS

OUTPUT STATUS

SELECTION OP

SELECTION OP STATUS

PERFORMABLE

SELECTION DECODER 5806

MUX SELECT BITS

FIG. 58
PICK ANY CONTROL:

OUTPUT(S) 5902A  INPUT(S) 5902B  OP 5902C

INPUT (E.G., NETWORK INGRESS BUFFER)  STATUS

OUTPUT STATUS

PERFORMABLE

FIG. 59

DECIDING AN EXECUTION:

PENDING OPERATIONS 6001

6002  6004  6006

MUX 6008  MUX 6010

CONTROL LINES  DEQUEUE CONTROL LINES

PRIORITY ENCODER 6012

PE STATUS

FIG. 60
Egress Queue Management:

Example: Pick (Input 0)

Network Ingress Buffer 6224
Ingress Control

Egress Configuration

Network Egress Buffer 6222
Send 0,0

Scheduler 6228A
Data

Scheduler 6228B
Index

Output Buffer(s) (E.g., Spatial Array Egress Buffers) 6226B

Index

(Input) Queue(s) (E.g., Spatial Array Egress Buffers) 6202

Op Argument 6202A
Send 0,0

Egress Credit Counters

Networks

FIG. 62
EGRESS QUEUE MANAGEMENT:

EXAMPLE: PICK (INPUT 1)

NETWORK INGRESS BUFFER 6324
EGRESS CONFIGURATION 6332
INGRESS CONTROL

SEND 0,1 6326A

INDEX

DATA

FROM MEM

SEND 0,1

DATA INDEX INDEX DATA

-6330 EGRESS CREDIT COUNTERS

(INPUT) QUEUE (S) (E.G., SPATIAL ARRAY EGRESS BUFFERS) 6302

OUTPUT BUFFER(S) (E.G., SPATIAL ARRAY EGRESS BUFFERS) 6308

FIG. 63A
Routing, with a packet switched communications network, data within a spatial array of processing elements between the processing elements according to a dataflow graph.

Performing a first dataflow operation of the dataflow graph with the processing elements.

Performing a second dataflow operation of the dataflow graph with a plurality of network dataflow endpoint circuits of the packet switched communications network.

FIG. 64
6800

C & C++ FORTRAN OTHER

COMPILE TIME

COMPILE (LLVM), INCLUDING OUTLINING

RUN TIME OR LATER

FAT BINARY: LLVM IR

JIT

LLVM => CSA CODE (KEY LOOPS)

BUFFERING INSERTION

PLACE AND ROUTE GRAPHS

GENERATE CONFIGURATION CODE

RUN CODE

FEEDBACK

FIG. 68
6900

COMPILER FRONT END

SEQ INS GEN
CONTROL FLOW (CF) TO DF
DF OPT

CSA BACKEND

SEQUENTIAL OPERATION SEMANTICS
CONVERSION TO DATAFLOW (DF)

DF ASSEMBLY CODE

BUFFERING INSERTION

UPDATED ASSEMBLY

PLACE & ROUTE

UPDATED ASSEMBLY

"SYMBOLIC" CSA ASM + LINKER
FUNCTIONAL SIMULATOR
TIMING SIMULATOR

FIG. 69
ld32 Rdata, Raddr
ld32 Rdata2, Raddr2
mul32 Rv0, Rdata, 17
mul32 Rv1, Rdata2, Rdata2
add32 Rres, Rv0, Rv1
st32 Raddr, Rres
ld32 Rdata3, Raddr3

SEQUENTIAL ASSEMBLY 7002

FIG. 70A

.lic .i32 data; .lic .i64 addr;
.lic .i32 data2; .lic .i64 addr2;
.lic .i32 data3; .lic .i64 addr3;
.lic .i32 v0; .lic .i32 v1; .lic .i32 res
ld32 data, addr
ld32 data2, addr2
mul32 v0, data, 17
mul32 v1, data2, data2
add32 res, v0, v1
st32 addr, res, done, %ign
ld32 data3, addr3, %ign, done

DATAFLOW ASSEMBLY 7004

FIG. 70B
DATAFLOW GRAPH 7006

FIG. 70C
if (i < n)
    y = x + a;
else
    y = i + x;

C SOURCE CODE 7102

FIG. 71A

.lic .il test
cmplts32 test, i, n
switch32 %ign, aT, test, a
switch32 iF, %ign, test, i
switch32 xF, xT, test, x
add32 yT, xT, aT  # True path
add32 yF, iF, xF   # False path
pick32 y, test, yF, yT
add32 z, y, 1

DATAFLOW ASSEMBLY 7104

FIG. 71B
```c
int i = 0;
int sum = 0;
do {
    sum = sum + i;
i = i + 1;
} while (i < n);
return sum;
```

```
# Loop control channels.
.lic .il picker
.lic .il switcher

# Offset values in picker with an initial 0.
.curr picker; .value 0; .avail 0

# Generate value of i for each loop iteration
pick32 top_i, picker, init_i, loopback_i
add32 bottom_i, top_i, 1
switch32 %ign, loopback_i, switcher, bottom_i

# Repeat value of n for each execution of the loop.
pick32 loop_n, picker, init_n, loopback_n
switch32 %ign, loopback_n, switcher, loop_n

# Comparison at the bottom of the loop.
cmplts32 switcher, bottom_i, loop_n
movl picker, switcher

# Add up the sum around the loop iteration.
pick32 top_sum, picker, init_sum, loopback_sum
add32 bottom_sum, top_sum, top_i
switch32 out_sum, loopback_sum, switcher, bottom_sum
```

```
C SOURCE CODE 7202

FIG. 72A
```

```
DATAFLOW ASSEMBLY 7204

FIG. 72B
```
DECODING AN INSTRUCTION WITH A DECODER OF A CORE OF A PROCESSOR INTO A DECODED INSTRUCTION 7302

EXECUTING THE DECODED INSTRUCTION WITH AN EXECUTION UNIT OF THE CORE OF THE PROCESSOR TO PERFORM A FIRST OPERATION 7304

RECEIVING AN INPUT OF A DATAFLOW GRAPH COMPRISING A PLURALITY OF NODES 7306


FIG. 73A
RECEIVING AN INPUT OF A DATAFLOW GRAPH COMPRISING A PLURALITY OF NODES 7303


FIG. 73B
FIG. 80

RECONFIGURATION CIRCUIT
8018

CONFIGURATION STATE
8020
DECODING AN INSTRUCTION WITH A DECODER OF A CORE OF A PROCESSOR INTO A DECODED INSTRUCTION 8702

EXECUTING THE DECODED INSTRUCTION WITH AN EXECUTION UNIT OF THE CORE OF THE PROCESSOR TO PERFORM A FIRST OPERATION 8704

RECEIVING AN INPUT OF A DATAFLOW GRAPH COMPRISING A PLURALITY OF NODES 8706

OVERLAYING THE DATAFLOW GRAPH INTO AN ARRAY OF PROCESSING ELEMENTS OF THE PROCESSOR WITH EACH NODE REPRESENTED AS A DATAFLOW OPERATOR IN THE ARRAY OF PROCESSING ELEMENTS 8708

PERFORMING A SECOND OPERATION OF THE DATAFLOW GRAPH WITH THE ARRAY OF PROCESSING ELEMENTS WHEN AN INCOMING OPERAND SET ARRIVES AT THE ARRAY OF PROCESSING ELEMENTS 8710

FIG. 87
DECODING AN INSTRUCTION WITH A DECODER OF A
CORE OF A PROCESSOR INTO A DECODED
INSTRUCTION 8802

EXECUTING THE DECODED INSTRUCTION WITH AN
EXECUTION UNIT OF THE CORE OF THE PROCESSOR
TO PERFORM A FIRST OPERATION 8804

RECEIVING AN INPUT OF A DATAFLOW GRAPH
COMPRISING A PLURALITY OF NODES 8806

OVERLAYING THE DATAFLOW GRAPH INTO A PLURALITY
OF PROCESSING ELEMENTS OF THE PROCESSOR AND
AN INTERCONNECT NETWORK BETWEEN THE
PLURALITY OF PROCESSING ELEMENTS OF THE
PROCESSOR WITH EACH NODE REPRESENTED AS A
DATAFLOW OPERATOR IN THE PLURALITY OF
PROCESSING ELEMENTS 8808

PERFORMING A SECOND OPERATION OF THE
DATAFLOW GRAPH WITH THE INTERCONNECT
NETWORK AND THE PLURALITY OF PROCESSING
ELEMENTS WHEN AN INCOMING OPERAND SET
ARRIVES AT THE PLURALITY OF PROCESSING
ELEMENTS 8810

FIG. 88
FIG. 89A

MEMORY ORDERING CIRCUIT 8905

ACCELERATION HARDWARE 8902

MEMORY SUBSYSTEM 8910

FIG. 89B

ACCELERATION 8902

12A 12B 12C 12D
for(i) {
    temp = p[i];
    P[i+2] = temp;
}

FIG. 97A

for(i) {
    temp = p[i];
    P[i+2] = temp;
}

FIG. 97B
<table>
<thead>
<tr>
<th>COUNTER_0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>COUNTER_1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>INCOMING VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OUTGOING VALUE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LOAD ADDR.</th>
<th>STORE ADDR.</th>
<th>STORE DATA</th>
<th>COMPLETION</th>
</tr>
</thead>
<tbody>
<tr>
<td>p[0]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>p[2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FIG. 99A**

<table>
<thead>
<tr>
<th>COUNTER_0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>COUNTER_1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>INCOMING VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OUTGOING VALUE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LOAD ADDR.</th>
<th>STORE ADDR.</th>
<th>STORE DATA</th>
<th>COMPLETION</th>
</tr>
</thead>
<tbody>
<tr>
<td>p[0]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FIG. 99B**

<table>
<thead>
<tr>
<th>COUNTER_0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>COUNTER_1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>INCOMING VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OUTGOING VALUE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LOAD ADDR.</th>
<th>STORE ADDR.</th>
<th>STORE DATA</th>
<th>COMPLETION</th>
</tr>
</thead>
<tbody>
<tr>
<td>p[1]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[3]</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>p[4]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FIG. 99C**
<table>
<thead>
<tr>
<th>COUNTER_0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>COUNTER_1</td>
<td>1</td>
</tr>
</tbody>
</table>

**FIG. 99D**

<table>
<thead>
<tr>
<th>LOAD ADDR</th>
<th>STORE ADDR</th>
<th>STORE DATA</th>
<th>COMPLETION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>p[6]</td>
<td></td>
</tr>
</tbody>
</table>

**FIG. 99E**

<table>
<thead>
<tr>
<th>LOAD ADDR</th>
<th>STORE ADDR</th>
<th>STORE DATA</th>
<th>COMPLETION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>p[6]</td>
<td></td>
</tr>
</tbody>
</table>

**FIG. 99F**

<table>
<thead>
<tr>
<th>LOAD ADDR</th>
<th>STORE ADDR</th>
<th>STORE DATA</th>
<th>COMPLETION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>p[6]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>p[7]</td>
<td></td>
</tr>
</tbody>
</table>
### FIG. 99G

<table>
<thead>
<tr>
<th>COUNTER_0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>COUNTER_1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LOAD ADDR.</th>
<th>STORE ADDR.</th>
<th>STORE DATA</th>
<th>COMPLETION</th>
</tr>
</thead>
<tbody>
<tr>
<td>p[1]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[2]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[3]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[4]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[5]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[6]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[7]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[8]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### FIG. 99H

<table>
<thead>
<tr>
<th>COUNTER_0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>COUNTER_1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LOAD ADDR.</th>
<th>STORE ADDR.</th>
<th>STORE DATA</th>
<th>COMPLETION</th>
</tr>
</thead>
<tbody>
<tr>
<td>p[3]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[3]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[4]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[4]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[5]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[6]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[7]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[8]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p[9]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

"INCOMING VALUE" OUTGOING VALUE
START

QUEUE MEMORY OPERATIONS

RECEIVE, IN ADDRESS QUEUE, AN ADDRESS OF MEMORY ASSOCIATED WITH A SECOND MEMORY OPERATION

RECEIVE, FROM ACCELERATION HARDWARE, A DEPENDENCY TOKEN ASSOCIATED WITH THE ADDRESS, THE DEPENDENCY TOKEN INDICATING A DEPENDENCY ON DATA GENERATED BY A PRECEDING FIRST MEMORY OPERATION

SCHEDULE ISSUANCE OF SECOND MEMORY OPERATION, TO THE MEMORY IN RESPONSE TO RECEIVING DEPENDENCY TOKEN AND THE ADDRESS FOR THE SECOND MEMORY OPERATION

ISSUE THE SECOND MEMORY OPERATION IN RESPONSE TO COMPLETION OF THE FIRST MEMORY OPERATION

START

FIG. 100
**FIG. 102D**

<table>
<thead>
<tr>
<th>AUGMENTATION OPERATION FIELD 10150</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U=0</strong></td>
</tr>
<tr>
<td>RS FIELD 10152A</td>
</tr>
<tr>
<td>ROUND 10152A.1</td>
</tr>
<tr>
<td>SAE FIELD 10156</td>
</tr>
<tr>
<td>ROUND OPERATION FIELD 10158</td>
</tr>
<tr>
<td>ROUND CONTROL FIELD 10154A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MOD FIELD 10242</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U=1</strong></td>
</tr>
<tr>
<td>WRITE MASK CONTROL FIELD 10152C</td>
</tr>
<tr>
<td>MERGING</td>
</tr>
<tr>
<td>ZEROING</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>CLASS FIELD</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ALPHA FIELD</strong></td>
</tr>
<tr>
<td><strong>BETA FIELD</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>MOD FIELD</strong> 10242</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DATA TRANSFORM FIELD</strong></td>
</tr>
<tr>
<td><strong>VECTOR LENGTH FIELD</strong></td>
</tr>
<tr>
<td><strong>BROADCAST FIELD</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>VECTOR LENGTH FIELD</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BROADCAST FIELD</strong></td>
</tr>
</tbody>
</table>
FIG. 106

PROCESSOR 10600

CORE 10602N

SPECIAL PURPOSE LOGIC 10608

BUS CONTROLLER INTEGRATED UNIT(S) 10616

CORE 10602A CACHE UNIT(S) 10604A CACHE UNIT(S), 10604N

SYSTEM AGENT UNIT 10610

MEMORY CONTROLLER UNIT(S) 10614

SHARED CACHE UNIT(S) 10606 RING 10612
FIG. 107
APPARATUSES, METHODS, AND SYSTEMS
FOR IN-NETWORK STORAGE IN A
CONFIGURABLE SPATIAL ACCELERATOR
STATEMENT REGARDING FEDERALLY
SPONSORED RESEARCH AND
DEVELOPMENT

[0001] This invention was made with Government support
under contract number H98230-13-D-0124 awarded by the
Department of Defense. The Government has certain rights
in this invention.

TECHNICAL FIELD

[0002] The disclosure relates generally to electronics, and,
more specifically, an embodiment of the disclosure relates to
a configurable spatial accelerator with in-network storage
circuitry.

BACKGROUND

[0003] A processor, or set of processors, executes instructions
from an instruction set, e.g., the instruction set architecture (ISA).
The instruction set is the part of the computer architecture related to
programming, and generally includes the native data types, instructions,
register architecture, addressing modes, memory architecture, interrupt
and exception handling, and external input and output (I/O). It
should be noted that the term instruction herein may refer to
a macro-instruction, e.g., an instruction that is provided to
the processor for execution, or to a micro-instruction, e.g.,
an instruction that results from a processor’s decoder decoding
macros-instructions.

BRIEF DESCRIPTION OF THE DRAWINGS

[0004] The present disclosure is illustrated by way of
example and not limitation in the figures of the accompa-
nying drawings, in which like references indicate similar
elements and in which:

[0005] FIG. 1 illustrates an accelerator tile according to
embodiments of the disclosure.

[0006] FIG. 2 illustrates a hardware processor coupled to
a memory according to embodiments of the disclosure.

[0007] FIG. 3A illustrates a program source according to
embodiments of the disclosure.

[0008] FIG. 3B illustrates a dataflow graph for the pro-
gram source of FIG. 3A according to embodiments of the
disclosure.

[0009] FIG. 3C illustrates an accelerator with a plurality of
processing elements configured to execute the dataflow
graph of FIG. 3B according to embodiments of the disclo-
sure.

[0010] FIG. 3D illustrates an example execution of a
dataflow graph according to embodiments of the disclosure.

[0011] FIG. 4 illustrates a program source according to
embodiments of the disclosure.

[0012] FIG. 5 illustrates an accelerator tile comprising an
array of processing elements according to embodiments of
the disclosure.

[0013] FIG. 6A illustrates a configurable data path network
according to embodiments of the disclosure.

[0014] FIG. 6B illustrates a configurable flow control path
network according to embodiments of the disclosure.

[0015] FIG. 7 illustrates a circuit switched network according to embodiments of the disclosure.

[0016] FIG. 8 illustrates a hardware processor tile comprising
an accelerator according to embodiments of the disclosure.

[0017] FIG. 9 illustrates a processing element according to
embodiments of the disclosure.

[0018] FIG. 10 illustrates data paths and control paths of
a processing element according to embodiments of the disclo-
sure.

[0019] FIG. 11 illustrates input controller circuitry of input
controller and/or input controller of processing element in
FIG. 10 according to embodiments of the disclosure.

[0020] FIG. 12 illustrates enqeneue circuitry of input con-
troller and/or input controller in FIG. 11 according to
embodiments of the disclosure.

[0021] FIG. 13 illustrates a status determiner of input
controller and/or input controller in FIG. 10 according to
embodiments of the disclosure.

[0022] FIG. 14 illustrates a head determiner state machine
according to embodiments of the disclosure.

[0023] FIG. 15 illustrates a tail determiner state machine
according to embodiments of the disclosure.

[0024] FIG. 16 illustrates a count determiner state
machine according to embodiments of the disclosure.

[0025] FIG. 17 illustrates an enqeneue determiner state
machine according to embodiments of the disclosure.

[0026] FIG. 18 illustrates a Not Full determiner state
machine according to embodiments of the disclosure.

[0027] FIG. 19 illustrates a Not Empty determiner state
machine according to embodiments of the disclosure.

[0028] FIG. 20 illustrates a valid determiner state machine
according to embodiments of the disclosure.

[0029] FIG. 21 illustrates output controller circuitry of
output controller and/or output controller of processing

[0030] FIG. 22 illustrates enqeneue circuitry of output
controller and/or output controller in FIG. 11 according to
embodiments of the disclosure.

[0031] FIG. 23 illustrates a status determiner of output
controller and/or output controller in FIG. 10 according to
embodiments of the disclosure.

[0032] FIG. 24 illustrates a head determiner state machine
according to embodiments of the disclosure.

[0033] FIG. 25 illustrates a tail determiner state machine
according to embodiments of the disclosure.

[0034] FIG. 26 illustrates a count determiner state
machine according to embodiments of the disclosure.

[0035] FIG. 27 illustrates an enqeneue determiner state
machine according to embodiments of the disclosure.

[0036] FIG. 28 illustrates a Not Full determiner state
machine according to embodiments of the disclosure.

[0037] FIG. 29 illustrates a Not Empty determiner state
machine according to embodiments of the disclosure.

[0038] FIG. 30 illustrates a valid determiner state machine
according to embodiments of the disclosure.

[0039] FIG. 31 illustrates an accelerator tile comprising an
array of processing elements and in-network storage ele-
ments according to embodiments of the disclosure.

[0040] FIG. 32A illustrates a first processing element, a
second processing element, a third processing element, a
first in-network storage element, and a second in-network
storage element coupled together by a network according to
embodiments of the disclosure.
FIG. 32B illustrates an in-network storage element of a circuit switched network of FIG. 32A configured to provide in-network storage according to embodiments of the disclosure.

FIG. 33 illustrates controller circuitry of a controller of an in-network storage element according to embodiments of the disclosure.

FIG. 34 illustrates enqueue and dequeue circuitry for use with the controller in FIG. 33 according to embodiments of the disclosure.

FIG. 35 illustrates a status determiner of the controller in FIG. 33 according to embodiments of the disclosure.

FIG. 36 illustrates a head determiner state machine according to embodiments of the disclosure.

FIG. 37 illustrates a tail determiner state machine according to embodiments of the disclosure.

FIG. 38 illustrates a count determiner state machine according to embodiments of the disclosure.

FIG. 39 illustrates an enqueue determiner state machine according to embodiments of the disclosure.

FIG. 40 illustrates a Queue Ready (Not Full) determiner state machine according to embodiments of the disclosure.

FIG. 41 illustrates a Queue Valid (Not Empty) determiner state machine according to embodiments of the disclosure.

FIG. 42 illustrates controller circuitry of a controller of an in-network storage element for multiple networks according to embodiments of the disclosure.

FIG. 43 illustrates enqueue and dequeue circuitry for use with the controller in FIG. 42 according to embodiments of the disclosure.

FIG. 44A illustrates a configurable data path network for use with an in-network storage element in a buffer mode according to embodiments of the disclosure.

FIG. 44B illustrates a configurable data path network for use with an in-network storage element in a bypass mode according to embodiments of the disclosure.

FIG. 44C illustrates a configurable flow control path network for use with an in-network storage element in a buffer mode according to embodiments of the disclosure.

FIG. 44D illustrates a configurable flow control path network for use with an in-network storage element in a bypass mode according to embodiments of the disclosure.

FIG. 45 illustrates a flow diagram according to embodiments of the disclosure.

FIG. 46 illustrates a request address file (RAF) circuit according to embodiments of the disclosure.

FIG. 47 illustrates a plurality of request address file (RAF) circuits coupled between a plurality of accelerator tiles and a plurality of cache banks according to embodiments of the disclosure.

FIG. 48 illustrates a data flow graph of a pseudo-code function call according to embodiments of the disclosure.

FIG. 49 illustrates a spatial array of processing elements with a plurality of network dataflow endpoint circuits according to embodiments of the disclosure.

FIG. 50 illustrates a network dataflow endpoint circuit according to embodiments of the disclosure.

FIG. 51 illustrates data formats for a send operation and a receive operation according to embodiments of the disclosure.

FIG. 52 illustrates another data format for a send operation according to embodiments of the disclosure.

FIG. 53 illustrates to configure a circuit element (e.g., network dataflow endpoint circuit) data formats to configure a circuit element (e.g., network dataflow endpoint circuit) for a send (e.g., switch) operation and a receive (e.g., pick) operation according to embodiments of the disclosure.

FIG. 54 illustrates a configuration data format to configure a circuit element (e.g., network dataflow endpoint circuit) for a send operation with its input, output, and control data annotated on a circuit according to embodiments of the disclosure.

FIG. 55 illustrates a configuration data format to configure a circuit element (e.g., network dataflow endpoint circuit) for a selected operation with its input, output, and control data annotated on a circuit according to embodiments of the disclosure.

FIG. 56 illustrates a configuration data format to configure a circuit element (e.g., network dataflow endpoint circuit) for a Switch operation with its input, output, and control data annotated on a circuit according to embodiments of the disclosure.

FIG. 57 illustrates a configuration data format to configure a circuit element (e.g., network dataflow endpoint circuit) for a Pick operation with its input, output, and control data annotated on a circuit according to embodiments of the disclosure.

FIG. 58 illustrates a configuration data format to configure a circuit element (e.g., network dataflow endpoint circuit) for a PickAny operation with its input, output, and control data annotated on a circuit according to embodiments of the disclosure.

FIG. 59 illustrates a configuration data format to configure a circuit element (e.g., network dataflow endpoint circuit) for a PickAny operation with its input, output, and control data annotated on a circuit according to embodiments of the disclosure.

FIG. 60 illustrates selection of an operation by a network dataflow endpoint circuit for performance according to embodiments of the disclosure.

FIG. 61 illustrates a network dataflow endpoint circuit according to embodiments of the disclosure.

FIG. 62 illustrates a network dataflow endpoint circuit receiving input zero (0) while performing a pick operation according to embodiments of the disclosure.

FIG. 63A illustrates a network dataflow endpoint circuit receiving input one (1) while performing a pick operation according to embodiments of the disclosure.

FIG. 63B illustrates a network dataflow endpoint circuit outputting the selected input while performing a pick operation according to embodiments of the disclosure.

FIG. 64 illustrates a flow diagram according to embodiments of the disclosure.

FIG. 65 illustrates a floating point multiplier partitioned into three regions (the result region, three potential carry regions, and the gated region) according to embodiments of the disclosure.

FIG. 66 illustrates an in-flight configuration of an accelerator with a plurality of processing elements according to embodiments of the disclosure.

FIG. 67 illustrates a snapshot of an in-flight, pipelined extraction according to embodiments of the disclosure.

FIG. 68 illustrates a compilation toehl chain for an accelerator according to embodiments of the disclosure.
[0082] FIG. 69 illustrates a compiler for an accelerator according to embodiments of the disclosure.

[0083] FIG. 70A illustrates sequential assembly code according to embodiments of the disclosure.

[0084] FIG. 70B illustrates dataflow assembly code for the sequential assembly code of FIG. 70A according to embodiments of the disclosure.

[0085] FIG. 70C illustrates a dataflow graph for the dataflow assembly code of FIG. 70B for an accelerator according to embodiments of the disclosure.

[0086] FIG. 71A illustrates C source code according to embodiments of the disclosure.

[0087] FIG. 71B illustrates dataflow assembly code for the C source code of FIG. 71A according to embodiments of the disclosure.

[0088] FIG. 71C illustrates a dataflow graph for the dataflow assembly code of FIG. 71B for an accelerator according to embodiments of the disclosure.

[0089] FIG. 72A illustrates C source code according to embodiments of the disclosure.

[0090] FIG. 72B illustrates dataflow assembly code for the C source code of FIG. 72A according to embodiments of the disclosure.

[0091] FIG. 72C illustrates a dataflow graph for the dataflow assembly code of FIG. 72B for an accelerator according to embodiments of the disclosure.

[0092] FIG. 73A illustrates a flow diagram according to embodiments of the disclosure.

[0093] FIG. 73B illustrates a flow diagram according to embodiments of the disclosure.

[0094] FIG. 74 illustrates a throughput versus energy per operation graph according to embodiments of the disclosure.

[0095] FIG. 75 illustrates an accelerator tile comprising an array of processing elements and a local configuration controller according to embodiments of the disclosure.

[0096] FIGS. 76A-76C illustrate a local configuration controller configuring a data path network according to embodiments of the disclosure.

[0097] FIG. 77 illustrates a configuration controller according to embodiments of the disclosure.

[0098] FIG. 78 illustrates an accelerator tile comprising an array of processing elements, a configuration cache, and a local configuration controller according to embodiments of the disclosure.

[0099] FIG. 79 illustrates an accelerator tile comprising an array of processing elements and a configuration and exception handling controller with a reconfiguration circuit according to embodiments of the disclosure.

[0100] FIG. 80 illustrates a reconfiguration circuit according to embodiments of the disclosure.

[0101] FIG. 81 illustrates an accelerator tile comprising an array of processing elements and a configuration and exception handling controller with a reconfiguration circuit according to embodiments of the disclosure.

[0102] FIG. 82 illustrates an accelerator tile comprising an array of processing elements and a mezzanine exception aggregator coupled to a tile-level exception aggregator according to embodiments of the disclosure.

[0103] FIG. 83 illustrates a processing element with an exception generator according to embodiments of the disclosure.

[0104] FIG. 84 illustrates an accelerator tile comprising an array of processing elements and a local extraction controller according to embodiments of the disclosure.

[0105] FIGS. 85A-85C illustrate a local extraction controller configuring a data path network according to embodiments of the disclosure.

[0106] FIG. 86 illustrates an extraction controller according to embodiments of the disclosure.

[0107] FIG. 87 illustrates a flow diagram according to embodiments of the disclosure.

[0108] FIG. 88 illustrates a flow diagram according to embodiments of the disclosure.

[0109] FIG. 89A is a block diagram of a system that employs a memory ordering circuit interposed between a memory subsystem and acceleration hardware according to embodiments of the disclosure.

[0110] FIG. 89B is a block diagram of the system of FIG. 89A, but which employs multiple memory ordering circuits according to embodiments of the disclosure.

[0111] FIG. 90 is a block diagram illustrating general functioning of memory operations into and out of acceleration hardware according to embodiments of the disclosure.

[0112] FIG. 91 is a block diagram illustrating a spatial dependency flow for a store operation according to embodiments of the disclosure.

[0113] FIG. 92 is a detailed block diagram of the memory ordering circuit of FIG. 89 according to embodiments of the disclosure.

[0114] FIG. 93 is a flow diagram of a microarchitecture of the memory ordering circuit of FIG. 89 according to embodiments of the disclosure.

[0115] FIG. 94 is a block diagram of an executable determiner circuit according to embodiments of the disclosure.

[0116] FIG. 95 is a block diagram of a priority encoder according to embodiments of the disclosure.

[0117] FIG. 96 is a block diagram of an exemplary load operation, both logical and in binary according to embodiments of the disclosure.

[0118] FIG. 97A is a flow diagram illustrating logical execution of an example code according to embodiments of the disclosure.

[0119] FIG. 97B is the flow diagram of FIG. 97A, illustrating memory-level parallelism in an unfolded version of the example code according to embodiments of the disclosure.

[0120] FIG. 98A is a block diagram of exemplary memory arguments for a load operation, and for a store operation according to embodiments of the disclosure.

[0121] FIG. 98B is a block diagram illustrating flow of load operations and the store operations, such as those of FIG. 98A, through the microarchitecture of the memory ordering circuit of FIG. 93 according to embodiments of the disclosure.


[0123] FIG. 100 is a flow chart of a method for ordering memory operations between a acceleration hardware and an out-of-order memory subsystem according to embodiments of the disclosure.

[0124] FIG. 101A is a block diagram illustrating a generic vector friendly instruction format and class A instruction templates thereof according to embodiments of the disclosure.
FIG. 101B is a block diagram illustrating the generic vector friendly instruction format and class B instruction templates thereof according to embodiments of the disclosure.

FIG. 102A is a block diagram illustrating fields for the generic vector friendly instruction formats in FIGS. 101A and 101B according to embodiments of the disclosure.

FIG. 102B is a block diagram illustrating the fields of the specific vector friendly instruction format in FIG. 102A that make up a full opcode field according to one embodiment of the disclosure.

FIG. 102C is a block diagram illustrating the fields of the specific vector friendly instruction format in FIG. 102A that make up a register index field according to one embodiment of the disclosure.

FIG. 102D is a block diagram illustrating the fields of the specific vector friendly instruction format in FIG. 102A that make up the augmentation operation field 10150 according to one embodiment of the disclosure.

FIG. 103 is a block diagram of a register architecture according to one embodiment of the disclosure.

FIG. 104A is a block diagram illustrating both an exemplary in-order pipeline and an exemplary register renaming, out-of-order issue/execution pipeline according to embodiments of the disclosure.

FIG. 104B is a block diagram illustrating both an exemplary embodiment of an in-order architecture core and an exemplary register renaming, out-of-order issue/execution architecture core to be included in a processor according to embodiments of the disclosure.

FIG. 105A is a block diagram of a single processor core, along with its connection to the on-die interconnect network and with its local subset of the Level 2 (L2) cache, according to embodiments of the disclosure.

FIG. 105B is an expanded view of part of the processor core in FIG. 105A according to embodiments of the disclosure.

FIG. 106 is a block diagram of a processor that may have more than one core, may have an integrated memory controller, and may have integrated graphics according to embodiments of the disclosure.

FIG. 107 is a block diagram of a system in accordance with one embodiment of the present disclosure.

FIG. 108 is a block diagram of a more specific exemplary system in accordance with an embodiment of the present disclosure.

FIG. 109 shows a block diagram of a second more specific exemplary system in accordance with an embodiment of the present disclosure.

FIG. 110, shown is a block diagram of a system on a chip (SoC) in accordance with an embodiment of the present disclosure.

FIG. 111 is a block diagram contrasting the use of a software instruction converter to convert binary instructions in a source instruction set to binary instructions in a target instruction set according to embodiments of the disclosure.

DETAILED DESCRIPTION

In the following description, numerous specific details are set forth. However, it is understood that embodiments of the disclosure may be practiced without these specific details. In other instances, well-known circuits, structures and techniques have not been shown in detail in order not to obscure the understanding of this description.

References in the specification to “one embodiment,” “an embodiment,” “an example embodiment,” etc., indicate that the embodiment described may include a particular feature, structure, or characteristic, but every embodiment may not necessarily include the particular feature, structure, or characteristic. Moreover, such phrases are not necessarily referring to the same embodiment. Further, when a particular feature, structure, or characteristic is described in connection with an embodiment, it is submitted that it is within the knowledge of one skilled in the art to affect such feature, structure, or characteristic in connection with other embodiments whether or not explicitly described.

A processor (e.g., having one or more cores) may execute instructions (e.g., a thread of instructions) to operate on data, for example, to perform arithmetic, logic, or other functions. For example, software may request an operation and a hardware processor (e.g., a core or cores thereof) may perform the operation in response to the request. One non-limiting example of an operation is a blend operation to input a plurality of vectors elements and output a vector with a blended plurality of elements. In certain embodiments, multiple operations are accomplished with the execution of a single instruction.

Exascale performance, e.g., as defined by the Department of Energy, may require system-level floating point performance to exceed 10^18 floating point operations per second (exaFLOPs) or more within a given (e.g., 20 MW) power budget. Certain embodiments herein are directed to a spatial array of processing elements (e.g., a configurable spatial accelerator (CSA)) that targets high performance computing (HPC), for example, of a processor. Certain embodiments herein of a spatial array of processing elements (e.g., a CSA) target the direct execution of a dataflow graph to yield a computationally dense yet energy-efficient spatial microarchitecture which far exceeds conventional roadmap architectures. Certain embodiments herein are directed to a spatial array of processing elements, memory, etc. and/or the communications network performing other communications (e.g., not data processing) operations. Certain embodiments herein are directed to a communications network (e.g., a packet switched network) of a (e.g., coupled to) spatial array of processing elements (e.g., a CSA) to perform certain dataflow operations, e.g., in addition to the communications network data between the processing elements, memory, etc. or the communications network performing other communications operations. Certain embodiments herein are directed to network dataflow endpoints that (e.g., each) perform (e.g., a portion or all) a dataflow operation or operations, for example, a pick or switch dataflow operation, e.g., of a dataflow graph. Certain embodiments herein include augmented network endpoints, e.g., network dataflow endpoints to support the control for (e.g., a plurality of or a subset of) dataflow operation(s), e.g., utilizing the network endpoints to perform a (e.g., dataflow) operation instead of a processing element (e.g., core) or arithmetic/logic unit (e.g. to perform arithmetic and logic operations) performing that (e.g., dataflow) operation. In one embodiment, a net-
work dataflow endpoint circuit is separate from a spatial array (e.g., an interconnect or fabric thereof) and/or processing elements.

[0145] Below also includes a description of the architectural philosophy of embodiments of a spatial array of processing elements (e.g., a CSA) and certain features thereof. As with any revolutionary architecture, programmability may be a risk. To mitigate this issue, embodiments of the CSA architecture have been co-designed with a compilation tool chain, which is also discussed below.

INTRODUCTION

[0146] Exascale computing goals may require enormous system-level floating point performance (e.g., 1 ExaFLOPs) within an aggressive power budget (e.g., 20 MW). However, simultaneously improving the performance and energy efficiency of program execution with classical von Neumann architectures has become difficult: out-of-order scheduling, simultaneous multi-threading, complex register files, and other structures provide performance, but at high energy cost. Certain embodiments herein achieve performance and energy requirements simultaneously. Exascale computing power-performance targets may demand both high throughput and low energy consumption per operation. Certain embodiments herein provide this by providing for large numbers of low-complexity, energy-efficient processing (e.g., computational) elements which largely eliminate the control overheads of previous processor designs. Guided by this observation, certain embodiments herein include a spatial array of processing elements, for example, a configurable spatial accelerator (CSA), e.g., comprising an array of processing elements (PEs) connected by a set of lightweight, high-speed interconnects (e.g., communication) networks. One example of a CSA tile is depicted in FIG. 1. Certain embodiments of processing (e.g., compute) elements are dataflow operators, e.g., multiple of a dataflow accelerator that only processes input data when both (i) the input data has arrived at the dataflow operation and (ii) there is space available for storing the output data, e.g., otherwise no processing is occurring. Certain embodiments (e.g., of an accelerator or CSA) do not utilize a triggered instruction.

[0147] FIG. 1 illustrates an accelerator tile 100 embodiment of a spatial array of processing elements according to embodiments of the disclosure. Accelerator tile 100 may be a portion of a larger tile. Accelerator tile 100 executes a dataflow graph or graphs. A dataflow graph may generally refer to an explicitly parallel program description which arises in the compilation of sequential codes. Certain embodiments herein (e.g., CSAs) allow dataflow graphs to be directly configured onto the CSA array, for example, rather than being transformed into sequential instruction streams. Certain embodiments herein allow a first (e.g., type of) dataflow operation to be performed by one or more processing elements (PEs) of the spatial array and, additionally or alternatively, a second (e.g., different, type of) dataflow operation to be performed by one or more of the network communication circuits (e.g., endpoints) of the spatial array.

[0148] The derivation of a dataflow graph from a sequential compilation flow allows embodiments of a CSA to support familiar programming models and to directly (e.g., without using a table of work) execute existing high performance computing (HPC) code. CSA processing elements (PEs) may be energy efficient. In FIG. 1, memory interface 102 may couple to a memory (e.g., memory 202 in FIG. 2) to allow accelerator tile 100 to access (e.g., load and store) data to the (e.g., off die) memory. Depicted accelerator tile 100 is a heterogeneous array comprised of several kinds of PEs coupled together via an interconnect network 104. Accelerator tile 100 may include one or more of integer arithmetic PEs, floating point arithmetic PEs, communication circuitry (e.g., network dataflow endpoint circuits), and in-fabric storage, e.g., as part of spatial array of processing elements 101. Dataflow graphs (e.g., compiled dataflow graphs) may be overlaid on the accelerator tile 100 for execution. In one embodiment, for a particular dataflow graph, each PE handles only one or two (e.g., dataflow) operations of the graph. The array of PEs may be heterogeneous, e.g., such that no PE supports the full CSA dataflow architecture and/or one or more PEs are programmed (e.g., customized) to perform only a few, but highly efficient operations. Certain embodiments herein thus yield a processor or accelerator having an array of processing elements that is computationally dense compared to roadmap architectures and yet achieves approximately an order-of-magnitude gain in energy efficiency and performance relative to existing HPC offerings.

[0149] Certain embodiments herein provide for performance increases from parallel execution within a (e.g., dense) spatial array of processing elements (e.g., CSA) where each PE and/or network dataflow endpoint circuit utilized may perform its operations simultaneously, e.g., if input data is available. Efficiency increases may result from the efficiency of each PE and/or network dataflow endpoint circuit, e.g., where each PE’s operation (e.g., behavior) is fixed once per configuration (e.g., mapping) step and execution occurs on local data available at the PE, e.g., without considering other fabric activity, and/or where each network dataflow endpoint circuit’s operation (e.g., behavior) is variable (e.g., not fixed) when configured (e.g., mapped). In certain embodiments, a PE and/or network dataflow endpoint circuit is (e.g., each a single) dataflow operator, for example, a dataflow operator that only operates on input data when both (i) the input data has arrived at the dataflow operator and (ii) there is space available for storing the output data, e.g., otherwise no operation is occurring.

[0150] Certain embodiments herein include a spatial array of processing elements as an energy-efficient and high-performance way of accelerating user applications. In one embodiment, applications are mapped in an extremely parallel manner. For example, inner loops may be unrolled multiple times to improve parallelism. This approach may provide high performance, e.g., when the occupancy (e.g., use) of the unrolled code is high. However, if there are less used code paths in the loop body unrolled (for example, an exceptional code path like floating point de-normalized mode) then (e.g., fabric area of) the spatial array of processing elements may be wasted and throughput consequently lost.

[0151] One embodiment herein to reduce pressure on (e.g., fabric area of) the spatial array of processing elements (e.g., in the case of unoptimized code segments) is time multiplexing. In this mode, a single instance of the less used (e.g., colder) code may be shared among several loop bodies, for example, analogous to a function call in a shared library. In one embodiment, spatial arrays (e.g., of processing elements) support the direct implementation of multiplexed codes. However, e.g., when multiplexing or demul-
multiplexing in a spatial array involves choosing among many and distant targets (e.g., sharers), a direct implementation using dataflow operators (e.g., using the processing elements) may be inefficient in terms of latency, throughput, implementation area, and/or energy. Certain embodiments herein describe hardware mechanisms (e.g., network circuitry) supporting (e.g., high-radix) multiplexing or demultiplexing. Certain embodiments herein (e.g., of network dataflow endpoint circuits) permit the aggregation of many targets (e.g., sharers) with little hardware overhead or performance impact. Certain embodiments herein allow for compiling of (e.g., legacy) sequential codes to parallel architectures in a spatial array.

[0152] In one embodiment, a plurality of network dataflow endpoint circuits combine as a single dataflow operator, for example, as discussed in reference to FIG. 49 below. As non-limiting examples, certain (for example, high (e.g., 4-6) radix) dataflow operators are listed below.

[0153] An embodiment of a “Pick” dataflow operator is to select data (e.g., a token) from a plurality of input channels and provide that data as its (e.g., single) output according to control data. Control data for a Pick may include an input selector value. In one embodiment, the selected input channel is to have its data (e.g., token) removed (e.g., discarded), for example, to complete the performance of that dataflow operation (or its portion of a dataflow operation). In one embodiment, additionally, those non-selected input channels are also to have their data (e.g., token) removed (e.g., discarded), for example, to complete the performance of that dataflow operation (or its portion of a dataflow operation).

[0154] An embodiment of a “PickSingLeEg” dataflow operator is to select data (e.g., a token) from a plurality of input channels and provide that data as its (e.g., single) output according to control data, but in certain embodiments, the non-selected input channels are ignored, e.g., those non-selected input channels are not to have their data (e.g., token) removed (e.g., discarded), for example, to complete the performance of that dataflow operation (or its portion of a dataflow operation). Control data for a PickSingLeEg may include an input selector value. In one embodiment, the selected input channel is also to have its data (e.g., token) removed (e.g., discarded), for example, to complete the performance of that dataflow operation (or its portion of a dataflow operation).

[0155] An embodiment of a “PickAny” dataflow operator is to select the first available (e.g., to the circuit performing the operation) data (e.g., a token) from a plurality of input channels and provide that data as its (e.g., single) output. In one embodiment, PickSingLeEg is also to output the index (e.g., indicating which of the plurality of input channels) had its data selected. In one embodiment, the selected input channel is to have its data (e.g., token) removed (e.g., discarded), for example, to complete the performance of that dataflow operation (or its portion of a dataflow operation). In certain embodiments, the non-selected input channels (e.g., with or without input data) are ignored, e.g., those non-selected input channels are not to have their data (e.g., token) removed (e.g., discarded), for example, to complete the performance of that dataflow operation (or its portion of a dataflow operation). Control data for a PickAny may include a value corresponding to the PickAny, e.g., without an input selector value.

[0156] An embodiment of a “Switch” dataflow operator is to steer (e.g., single) input data (e.g., a token) so as to provide that input data to one or a plurality of (e.g., less than all) outputs according to control data. Control data for a Switch may include an output(s) selector value or values. In one embodiment, the input data (e.g., from an input channel) is to have its data (e.g., token) removed (e.g., discarded), for example, to complete the performance of that dataflow operation (or its portion of a dataflow operation).

[0157] An embodiment of a “SwitchAny” dataflow operator is to steer (e.g., single) input data (e.g., a token) so as to provide that input data to one or a plurality of (e.g., less than all) outputs that may receive that data, e.g., according to control data. In one embodiment, SwitchAny may provide the input data to any coupled output channel that has availability (e.g., available storage space) in its ingress buffer, e.g., network ingress buffer in FIG. 50. Control data for a SwitchAny may include a value corresponding to the SwitchAny, e.g., without an output(s) selector value or values. In one embodiment, the input data (e.g., from an input channel) is to have its data (e.g., token) removed (e.g., discarded), for example, to complete the performance of that dataflow operation (or its portion of a dataflow operation). In one embodiment, SwitchAny also is to output the index (e.g., indicating which of the plurality of output channels) that it provided (e.g., sent) the input data to. SwitchAny may be utilized to manage replicated sub-graphs in a spatial array, for example, an unrolled loop.

[0158] Certain embodiments herein thus provide paradigm-shifting levels of performance and tremendous improvements in energy efficiency across a broad class of existing single-stream and parallel programs, e.g., while preserving familiar HPC programming models. Certain embodiments herein may target HPC such that floating point energy efficiency is extremely important. Certain embodiments herein not only deliver compelling improvements in performance and reductions in energy, they also deliver these gains to existing HPC programs written in mainstream HPC languages and for mainstream HPC frameworks. Certain embodiments of the architecture herein (e.g., with compilation in mind) provide several extensions in direct support of the control-dataflow internal representations generated by modern compilers. Certain embodiments herein are direct to a CSA dataflow compiler, e.g., which can accept C, C++, and Fortran programming languages, to target a CSA architecture.

[0159] FIG. 2 illustrates a hardware processor 200 coupled to (e.g., connected to) a memory 202 according to embodiments of the disclosure. In one embodiment, hardware processor 200 and memory 202 are a computing system 201. In certain embodiments, one or more of accelerators is a CSA according to this disclosure. In certain embodiments, one or more of the cores in a processor are those cores disclosed herein. Hardware processor 200 (e.g., each core thereof) may include a hardware decoder (e.g., decode unit) and a hardware execution unit. Hardware processor 200 may include registers. Note that the figures herein may not depict all data communication couplings (e.g., connections). One of ordinary skill in the art will appreciate that this is to not obscure certain details in the figures. Note that a double headed arrow in the figures may not require two-way communication, for example, it may indicate one-way communication (e.g., to or from that component or device). Any or all combinations of communications paths may be utilized in certain embodiments herein. Depicted hardware processor 200 includes a plurality of cores (0 to N, where N
may be 1 or more) and hardware accelerators (0 to M, where M may be 1 or more) according to embodiments of the disclosure. Hardware processor 200 (e.g., accelerator(s) and/or core(s) thereof) may be coupled to memory 202 (e.g., data storage device). Hardware decoder (e.g., of core) may receive an (e.g., single) instruction (e.g., macro-instruction) and decode the instruction, e.g., into micro-instructions and/or micro-operations. Hardware execution unit (e.g., of core) may execute the decoded instruction (e.g., macro-instruction) to perform an operation or operations.

Section 1 below discloses embodiments of CSA architecture. In particular, novel embodiments of integrating memory within the dataflow execution model are disclosed. Section 2 delves into the microarchitectural details of embodiments of a CSA. In one embodiment, the main goal of a CSA is to support compiler produced programs. Section 3 below examines embodiments of a CSA compilation tool chain. The advantages of embodiments of a CSA are compared to other architectures in the execution of compiled code in Section 4. Finally, the performance of embodiments of a CSA microarchitecture is discussed in Section 5. Further CSA details are discussed in Section 6, and a summary is provided in Section 7.

1. CSA Architecture

The goal of certain embodiments of a CSA is to rapidly and efficiently execute programs, e.g., programs produced by compilers. Certain embodiments of the CSA architecture provide programming abstractions that support the needs of compiler technologies and programming paradigms. Embodiments of the CSA execute dataflow graphs, e.g., a program manifestation that closely resembles the compiler’s own internal representation (IR) of compiled programs. In this model, a program is represented as a dataflow graph comprised of nodes (e.g., vertices) drawn from a set of architecturally-defined dataflow operators (e.g., that encompass both computation and control operations) and edges which represent the transfer of data between dataflow operators. Execution may proceed by injecting dataflow tokens (e.g., those that are or represent data values) into the dataflow graph. Tokens may flow between and be transformed at each node (e.g., vertex), for example, forming a complete computation. A sample dataflow graph and its derivation from high-level source code is shown in FIGS. 3A-3C, and FIG. 4 shows an example of the execution of a dataflow graph.

Embodiments of the CSA are configured for dataflow graph execution by providing exactly those dataflow-graph-execution supports required by compilers. In one embodiment, the CSA is an accelerator (e.g., an accelerator in FIG. 2) and it does not seek to provide some of the necessary but infrequently used mechanisms available on general purpose processing cores (e.g., a core in FIG. 2), such as system calls. Therefore, in this embodiment, the CSA can execute many codes, but not all codes. In exchange, the CSA gains significant performance and energy advantages. To enable the acceleration of code written in commonly used sequential languages, embodiments herein also introduce several novel architectural features to assist the compiler. One particular novelty is CSA’s treatment of memory, a subject which has been ignored or poorly addressed previously. Embodiments of the CSA are also unique in the use of dataflow operators, e.g., as opposed to lookup tables (LUTs), as their fundamental architectural interface.

Turning to embodiments of the CSA, dataflow operators are discussed next.

1.1 Dataflow Operators

The key architectural interface of embodiments of the accelerator (e.g., CSA) is the dataflow operator, e.g., as a direct representation of a node in a dataflow graph. From an operational perspective, dataflow operators behave in a streaming or data-driven fashion. Dataflow operators may execute as soon as their incoming operands become available. CSA dataflow execution may depend (e.g., only) on highly localized status, for example, resulting in a highly scalable architecture with a distributed, asynchronous execution model. Dataflow operators may include arithmetic dataflow operators, for example, one or more of floating point addition and multiplication, integer addition, subtraction, and multiplication, various forms of comparison, logical operators, and shift. However, embodiments of the CSA may also include a rich set of control operators which assist in the management of dataflow tokens in the program graph. Examples of these include a “pick” operator, e.g., which multiplexes two or more logical input channels into a single output channel, and a “switch” operator, e.g., which operates as a channel demultiplexor (e.g., outputting a single channel from two or more logical input channels). These operators may enable a compiler to implement control paradigms such as conditional expressions. Certain embodiments of a CSA may include a limited dataflow operator set (e.g., to relatively small number of operations) to yield dense and energy efficient PE microarchitectures. Certain embodiments may include dataflow operators for complex operations that are common in HPC code. The CSA dataflow operator architecture is highly amenable to deployment-specific extensions. For example, more complex mathematical dataflow operators, e.g., trigonometry functions, may be included in certain embodiments to accelerate certain mathematics-intensive HPC workloads. Similarly, a neural-network tuned extension may include dataflow operators for vectorized, low precision arithmetic.

FIG. 3A illustrates a program source according to embodiments of the disclosure. Program source code includes a multiplication function (func). FIG. 3B illustrates a dataflow graph 300 for the program source of FIG. 3A according to embodiments of the disclosure. Dataflow graph 300 includes a pick node 304, switch node 306, and multiplication node 308. A buffer may optionally be included along one or more of the communication paths. Depicted dataflow graph 300 may perform an operation of selecting input X with pick node 304, multiplying X by Y (e.g., multiplication node 308), and then outputting the result from the left output of the switch node 306. FIG. 3C illustrates an accelerator (e.g., CSA) with a plurality of processing elements 301 configured to execute the dataflow graph of FIG. 3B according to embodiments of the disclosure. More particularly, the dataflow graph 300 is overlaid into the array of processing elements 301 (e.g., and the (e.g., interconnect) network(s) therebetween), for example, such that each node of the dataflow graph 300 is represented as a dataflow operator in the array of processing elements 301. For example, certain dataflow operations may be achieved with a processing element and/or certain dataflow operations may
be achieved with a communications network (e.g., a network dataflow endpoint circuit thereof). For example, a Pick, PickSingleEl, PickAny, Switch, and/or SwitchAny operation may be achieved with one or more components of a communications network (e.g., a network dataflow endpoint circuit thereof), e.g., in contrast to a processing element.

[0166] In one embodiment, one or more of the processing elements in the array of processing elements 301 is to access memory through memory interface 302. In one embodiment, pick node 304 of dataflow graph 300 thus corresponds (e.g., is represented by) to pick operator 304A, switch node 306 of dataflow graph 300 thus corresponds (e.g., is represented by) to switch operator 306A, and multiplier node 308 of dataflow graph 300 thus corresponds (e.g., is represented by) to multiplier operator 308A. Another processing element and/or a flow control path network may provide the control signals (e.g., control tokens) to the pick operator 304A and switch operator 306A to perform the operation in FIG. 3A. In one embodiment, array of processing elements 301 is configured to execute the dataflow graph 300 of FIG. 3B before execution begins. In one embodiment, compiler performs the conversion from FIG. 3A-3B. In one embodiment, the input of the dataflow graph nodes into the array of processing elements logically embeds the dataflow graph into the array of processing elements, e.g., as discussed further below, such that the input/output paths are configured to produce the desired result.

1.2 Latency Insensitive Channels

[0167] Communications arcs are the second major component of the dataflow graph. Certain embodiments of a CSA describe these arcs as latency insensitive channels, for example, in-order, back-pressured (e.g., not producing or sending output until there is a place to store the output), point-to-point communications channels. As with dataflow operators, latency insensitive channels are fundamentally asynchronous, giving the freedom to compose many types of networks to implement the channels of a particular graph. Latency insensitive channels may have arbitrarily long latencies and still faithfully implement the CSA architecture. However, in certain embodiments there is strong incentive in terms of performance and energy to make latencies as small as possible. Section 2.2 herein discloses a network microarchitecture in which dataflow graph channels are implemented in a pipelined fashion with no more than one cycle of latency. Embodiments of latency-insensitive channels provide a critical abstraction layer which may be leveraged with the CSA architecture to provide a number of runtime services to the applications programmer. For example, a CSA may leverage latency-insensitive channels in the implementation of the CSA configuration (the loading of a program onto the CSA array).

[0168] FIG. 3D illustrates an example execution of a dataflow graph 300 according to embodiments of the disclosure. At step 1, input values (e.g., 1 for X in FIG. 3B and 2 for Y in FIG. 3B) may be loaded in dataflow graph 300 to perform a 1*2 multiplication operation. One or more of the data input values may be static (e.g., constant) in the operation (e.g., 1 for X and 2 for Y in reference to FIG. 3B) or updated during the operation. At step 2, a processing element (e.g., on a flow control path network) or circuit outputs a zero to control input (e.g., multiplexer control signal) of pick node 304 (e.g., to source a one from port "0" to its output) and outputs a zero to control input (e.g., multiplexer control signal) of switch node 306 (e.g., to provide its output of port "0" to a destination (e.g., a downstream processing element). At step 3, the data value of 1 is output from pick node 304 (e.g., and consumes its control signal "0" at the pick node 304) to multiplier node 308 to be multiplied with the data value of 2 at step 4. At step 4, the output of multiplier node 308 arrives at switch node 306, e.g., which causes switch node 306 to consume a control signal "0" to output the value of 2 from port "0" of switch node 306 at step 5. The operation is then complete. A CSA may thus be programmed accordingly such that a corresponding dataflow operator for each node performs the operations in FIG. 3. Although execution is serialized in this example, in principle all dataflow operations may execute in parallel. Steps are used in FIG. 3 to differentiate dataflow execution from any physical microarchitectural manifestation. In one embodiment a downstream processing element is to send a signal (or not send a ready signal) (for example, on a flow control path network) to the switch 306 to stall the output from the switch 306, e.g., until the downstream processing element is ready (e.g., has storage room) for the output.

1.3 Memory

[0169] Dataflow architectures generally focus on communication and data manipulation with less attention paid to state. However, enabling real software, especially programs written in legacy sequential languages, requires significant attention to interfacing with memory. Certain embodiments of a CSA use architectural memory operations as their primary interface to (e.g., large) stateful storage. From the perspective of the dataflow graph, memory operations are similar to other dataflow operations, except that they have the side effect of updating a shared store. In particular, memory operations of certain embodiments herein have the same semantics as every other dataflow operator, for example, they “execute” when their operands, e.g., an address are available and, after some latency, a response is produced. Certain embodiments herein explicitly decouple the operand input and result output such that memory operations are naturally pipelined and have the potential to produce many simultaneous outstanding requests, e.g., making them exceptionally well suited to the latency and bandwidth characteristics of a memory subsystem. Embodiments of a CSA provide basic memory operations such as load, which takes an address channel and populates a response channel with the values corresponding to the addresses, and a store. Embodiments of a CSA may also provide more advanced operations such as in-memory atomicity and consistency operators. These operations may have similar semantics to their von Neumann counterparts. Embodiments of a CSA may accelerate existing programs described using sequential languages such as C and Fortran. A consequence of supporting these language models is addressing program memory order, e.g., the serial ordering of memory operations typically prescribed by these languages.

[0170] FIG. 4 illustrates a program source (e.g., C code) 400 according to embodiments of the disclosure. According to the memory semantics of the C programming language, memory copy (memcpy) should be serialized. However, memcpy may be parallelized with an embodiment of the CSA if arrays A and B are known to be disjoint. FIG. 4 further illustrates the problem of program order. In general, compilers cannot prove that array A is different from array...
B, e.g., either for the same value of index or different values of index across loop bodies. This is known as pointer or memory aliasing. Since compilers are to generate statically correct code, they are usually forced to serialize memory accesses. Typically, compilers targeting sequential von Neumann architectures use instruction ordering as a natural means of enforcing program order. However, embodiments of the CSA have no notion of instruction or instruction-based program ordering as defined by a program counter. In certain embodiments, incoming dependency tokens, e.g., which contain no architecturally visible information, are like all other dataflow tokens and memory operations may not execute until they have received a dependency token. In certain embodiments, memory operations produce an outgoing dependency token once their operation is visible to all logically subsequent, dependent memory operations. In certain embodiments, dependency tokens are similar to other dataflow tokens in a dataflow graph. For example, since memory operations occur in conditional contexts, dependency tokens may also be manipulated using control operators described in Section 1.1, e.g., like any other tokens. Dependency tokens may have the effect of serializing memory accesses, e.g., providing the compiler a means of architecturally defining the order of memory accesses.

1.4 Runtime Services

[0171] A primary architectural considerations of embodiments of the CSA involve the actual execution of user-level programs, but it may also be desirable to provide several support mechanisms which underpin this execution. Chief among these are configuration (in which a dataflow graph is loaded into the CSA), extraction (in which the state of an executing graph is moved to memory), and exceptions (in which mathematical, soft, and other types of errors in the fabric are detected and handled, possibly by an external entity). Section 2.8 below discusses the properties of a latency-insensitive dataflow architecture of an embodiment of a CSA to yield efficient, largely pipelined implementations of these functions. Conceptually, configuration may load the state of a dataflow graph into the interconnect (and/or communications network (e.g., a network dataflow endpoint circuit thereof)) and processing elements (e.g., fabric), e.g., generally from memory. During this step, all structures in the CSA may be loaded with a new dataflow graph and any dataflow tokens live in that graph, for example, as a consequence of a context switch. The latency-insensitive semantics of a CSA may permit a distributed, asynchronous initialization of the fabric, e.g., as soon as PEs are configured, they may begin execution immediately. Unconfigured PEs may backpressure their channels until they are configured, e.g., preventing communications between configured and unconfigured elements. The CSA configuration may be partitioned into privileged and user-level state. Such a two-level partitioning may enable primary configuration of the fabric to occur without invoking the operating system. During an embodiment of extraction, a logical view of the dataflow graph is captured and committed into memory, e.g., including all live control and dataflow tokens and state in the graph.

[0172] Extraction may also play a role in providing reliability guarantees through the creation of fabric checkpoints. Exceptions in a CSA may generally be caused by the same events that cause exceptions in processors, such as illegal operator arguments or reliability, availability, and serviceability (RAS) events. In certain embodiments, exceptions are detected at the level of dataflow operators, for example, checking argument values or through modular arithmetic schemes. Upon detecting an exception, a dataflow operator (e.g., circuit) may halt and emit an exception message, e.g., which contains both an operation identifier and some details of the nature of the problem that has occurred. In one embodiment, the dataflow operator will remain halted until it has been reconfigured. The exception message may then be communicated to an associated processor (e.g., core) for service, e.g., which may include extracting the graph for software analysis.

1.5 Tile-Level Architecture

[0173] Embodiments of the CSA computer architectures (e.g., targeting HPC and datacenter uses) are tiled. FIGS. 5 and 8 show tile-level deployments of a CSA. FIG. 8 shows a full-tile implementation of a CSA, e.g., which may be an accelerator of a processor with a core. A main advantage of this architecture is that the potential demand from the core may be reduced design risk, e.g., such that the CSA and core are completely decoupled in manufacturing. In addition to allowing better component reuse, this may allow the design of components like the CSA cache to consider only the CSA, e.g., rather than needing to incorporate the stricter latency requirements of the core. Finally, separate tiles may allow for the integration of CSA with small or large cores. One embodiment of the CSA captures most vector-parallel workloads such that most vector-style workloads run directly on the CSA, but in certain embodiments vector-style instructions in the core may be included, e.g., to support legacy binaries.

2. Microarchitecture

[0174] In one embodiment, the goal of the CSA microarchitecture is to provide a high quality implementation of each dataflow operator specified by the CSA architecture. Embodiments of the CSA microarchitecture provide that each processing element (and/or communications network (e.g., a network dataflow endpoint circuit thereof)) of the microarchitecture corresponds to approximately one node (e.g., entity) in the architectural dataflow graph. In one embodiment, a node in the dataflow graph is distributed in multiple network dataflow endpoints. In certain embodiments, this results in microarchitectural elements that are not only compact, resulting in a dense computation array, but also energy efficient, for example, where processing elements (PEs) are both simple and largely unmultiplied, e.g., executing a single dataflow operator for a configuration (e.g., programming) of the CSA. To further reduce energy and implementation area, a CSA may include a configurable, heterogeneous fabric style in which each PE thereof implements only a subset of dataflow operators (e.g., with a separate subset of dataflow operators implemented with network dataflow endpoint circuit(s)). Peripheral and support subsystems, such as the CSA cache, may be provisioned to support the distributed parallelism incumbent in the main CSA processing fabric itself. Implementation of CSA microarchitectures may utilize dataflow and latency-insensitive communications abstractions present in the architecture. In certain embodiments, there is (e.g., substantially) a one-to-one correspondence between nodes in the compiler generated graph and the dataflow operators (e.g., dataflow operator compute elements) in a CSA.
Below is a discussion of an example CSA, followed by a more detailed discussion of the microarchitecture. Certain embodiments herein provide a CSA that allows for easy compilation, e.g., in contrast to an existing FPGA compilers that handle a small subset of a programming language (e.g., C or C++) and require many hours to compile even small programs.

Certain embodiments of a CSA architecture admits of heterogeneous coarse-grained operations, like double precision floating point. Programs may be expressed in fewer coarse grained operations, e.g., such that the disclosed compiler runs faster than traditional spatial compilers. Certain embodiments include a fabric with new processing elements to support sequential concepts like program ordered memory accesses. Certain embodiments implement hardware to support coarse-grained dataflow-style communication channels. This communication model is abstract, and very close to the control-dataflow representation used by the compiler. Certain embodiments herein include a network implementation that supports single-cycle latency communications, e.g., utilizing (e.g., small) PEs which support single control-dataflow operations. In certain embodiments, not only does this improve energy efficiency and performance, it simplifies compilation because the compiler makes a one-to-one mapping between high-level dataflow constructs and the fabric. Certain embodiments herein thus simplify the task of compiling existing (e.g., C, C++, or Fortran) programs to a CSA (e.g., fabric).

Energy efficiency may be a first order concern in modern computer systems. Certain embodiments herein provide a new schema of energy-efficient spatial architectures. In certain embodiments, these architectures form a fabric with a unique composition of a heterogeneous mix of small, energy-efficient, data-flow oriented processing elements (PEs) (and/or a packet switched communications network (e.g., a network dataflow endpoint circuit thereof)) with a lightweight circuit switched communications network (e.g., interconnect), e.g., with hardened support for low control. Due to the energy advantages of each, the combination of these components may form a spatial accelerator (e.g., as part of a computer) suitable for executing compiler-generated parallel programs in an extremely energy efficient manner. Since this fabric is heterogeneous, certain embodiments may be customized for different application domains by introducing new domain-specific PEs. For example, a fabric for high-performance computing might include some customization for double-precision, fused multiply-add, while a fabric targeting deep neural networks might include low-precision floating point operations.

An embodiment of a spatial architecture schema, e.g., as exemplified in FIG. 5, is the composition of lightweight processing elements (PE) connected by an inter-PE network. Generally, PEs may comprise dataflow operators, e.g., where once (e.g., all) input operands arrive at the dataflow operator, some operation (e.g., micro-instruction or set of micro-instructions) is executed, and the results are forwarded to downstream operators. Control, scheduling, and data storage may therefore be distributed amongst the PEs, e.g., removing the overhead of the centralized structures that dominate classical processors.

Programs may be converted to dataflow graphs that are mapped onto the architecture by configuring PEs and the network to express the control-dataflow graph of the program. Communication channels may be flow-controlled and fully back-pressured, e.g., such that PEs will stall if either source communication channels have no data or destination communication channels are full. In one embodiment, at runtime, data flow through the PEs and channels that have been configured to implement the operation (e.g., an accelerated algorithm). For example, data may be streamed in from memory, through the fabric, and then back out to memory.

Embodiments of such an architecture may achieve remarkable performance efficiency relative to traditional multicore processors: compute (e.g., in the form of PEs) may be simpler, more energy efficient, and more plentiful than in larger cores, and communications may be direct and mostly short-haul, e.g., as opposed to occurring over a wide, full-chip network as in typical multicore processors. Moreover, because embodiments of the architecture are extremely parallel, a number of powerful circuit and device level optimizations are possible without seriously impacting throughput, e.g., low leakage devices and low operating voltage. These lower-level optimizations may enable even greater performance advantages relative to traditional cores. The combination of efficiency at the architectural, circuit, and device levels yields of these embodiments are compelling. Embodiments of this architecture may enable larger active areas as transistor density continues to increase.

Embodiments herein offer a unique combination of dataflow support and circuit switching to enable the fabric to be smaller, more energy-efficient, and provide higher aggregate performance as compared to previous architectures. FPGAs are generally tuned towards fine-grained bit manipulation, whereas embodiments herein are tuned toward the double-precision floating point operations found in HPC applications. Certain embodiments herein may include a FPGA in addition to a CSA according to this disclosure.

Certain embodiments herein combine a lightweight network with energy efficient dataflow processing elements (and/or communications network (e.g., a network dataflow endpoint circuit thereof)) to form a high-throughput, low-latency, energy-efficient HPC fabric. This low-latency network may enable the building of processing elements (and/or communications network (e.g., a network dataflow endpoint circuit thereof)) with fewer functionalities, for example, only one or two instructions and perhaps one architecturally visible register, since it is efficient to gang multiple PEs together to form a complete program.

Relative to a processor core, CSA embodiments herein may provide for more computational density and energy efficiency. For example, when PEs are very small (e.g., compared to a core), the CSA may perform many more operations and have much more computational parallelism than a core, e.g., perhaps as many as 16 times the number of FMs as a vector processing unit (VPV). To utilize all of these computational elements, the energy per operation is very low in certain embodiments.

The energy advantages of our embodiments of this dataflow architecture are many. Parallelism is explicit in dataflow graphs and embodiments of the CSA architecture spend no or minimal energy to extract it, e.g., unlike out-of-order processors which must re-discover parallelism each time an instruction is executed. Since each PE is responsible for a single operation in one embodiment, the register files and ports counts may be small, e.g., often only one, and therefore use less energy than their counterparts in core. Certain CSAs include many PEs, each of which holds
Live program values, giving the aggregate effect of a huge register file in a traditional architecture, which dramatically reduces memory accesses. In embodiments where the memory is multi-ported and distributed, a CSA may sustain many more outstanding memory requests and utilize more bandwidth than a core. These advantages may combine to yield an energy level per watt that is only a small percentage over the cost of the bare arithmetic circuitry. For example, in the case of an integer multiply, a CSA may consume no more than 25% more energy than the underlying multiplication circuit. Relative to one embodiment of a core, an integer operation in that CSA fabric consumes less than 1/8th of the energy per integer operation.

From a programming perspective, the application-specific malleability of embodiments of the CSA architecture yields significant advantages over a vector processing unit (VPU). In traditional, inflexible architectures, the number of functional units, like floating divide or the various transcendental mathematical functions, must be chosen at design time based on some expected use case. In embodiments of the CSA architecture, such functions may be configured (e.g., by a user and not a manufacturer) into the fabric based on the requirement of each application. Application throughput may thereby be further increased. Simultaneously, the compute density of embodiments of the CSA improves by avoiding hardening such functions, and instead provision more instances of primitive functions like floating multiplication. These advantages may be significant in HPC workloads, some of which spend 75% of floating execution time in transcendental functions.

Certain embodiments of the CSA represents a significant advance as a dataflow-oriented spatial architectures, e.g., the PEs of this disclosure may be smaller, but also more energy-efficient. These improvements may directly result from the combination of dataflow-oriented PEs with a lightweight, circuit switched interconnect, for example, which has single-cycle latency, e.g., in contrast to a packet switched network (e.g., with, at a minimum, a 300% higher latency). Certain embodiments of PEs support 32-bit or 64-bit operation. Certain embodiments herein permit the introduction of new application-specific PEs, for example, for machine learning or security, and not merely a homogeneous combination. Certain embodiments herein combine lightweight dataflow-oriented processing elements with a lightweight, low-latency network to form an energy efficient computational fabric.

In order for certain spatial architectures to be successful, programmers are to configure them with relatively little effort, e.g., while obtaining significant power and performance superiority over sequential cores. Certain embodiments herein provide for a CSA (e.g., spatial fabric) that is easily programmed (e.g., by a compiler), power efficient, and highly parallel. Certain embodiments herein provide for a (e.g., interconnect) network that achieves these three goals. From a programmability perspective, certain embodiments of the network provide flow controlled channels, e.g., which correspond to the control-dataflow graph (CDFG) model of execution used in compilers. Certain network embodiments utilize dedicated, circuit switched links, such that program performance is easier to reason about, both by a human and a compiler, because performance is predictable. Certain network embodiments offer both high bandwidth and low latency. Certain network embodiments (e.g., static, circuit switching) provides a latency of 0 to 1 cycle (e.g., depending on the transmission distance.) Certain network embodiments provide for a high bandwidth by laying out several networks in parallel, e.g., and in low-level metals. Certain network embodiments communicate in low-level metals and over short distances, and thus are very power efficient.

Certain embodiments of networks include architectural support for flow control. For example, in spatial accelerators composed of small processing elements (PEs), communications latency and bandwidth may be critical to overall program performance. Certain embodiments herein provide for a lightweight, circuit switched network which facilitates communication between PEs in spatial processing arrays, such as the spatial array shown in FIG. 5, and the micro-architectural control features necessary to support this network. Certain embodiments of a network enable the construction of point-to-point, flow controlled communications channels which support the communications of the dataflow oriented processing elements (PEs). In addition to point-to-point communications, certain networks herein also support multiast communications. Communications channels may be formed by statically configuring the network to from virtual circuits between PEs. Circuit switching techniques herein may decrease communications latency and commensurately minimize network buffering, e.g., resulting in both high performance and high energy efficiency. In certain embodiments of a network, inter-PE latency may be as low as a zero cycles, meaning that the downstream PE may operate on data in the cycle after it is produced. To obtain even higher bandwidth, and to admit more programs, multiple networks may be laid out in parallel, e.g., as shown in FIG. 5.

Spatial architectures, such as the one shown in FIG. 5, may be the composition of lightweight processing elements connected by an inter-PE network (and/or communications network (e.g., a network dataflow endpoint circuit thereof)). Programs, viewed as dataflow graphs, may be mapped onto the architecture by configuring PEs and the network. Generally, PEs may be configured as dataflow operators, and once (e.g., all) input operands arrive at the PE, some operation may then occur, and the result are forwarded to the desired downstream PEs. PEs may communicate over dedicated virtual circuits which are formed by statically configuring a circuit switched communications network. These virtual circuits may be flow controlled and fully back-pressured, e.g., such that PEs will stall if either the source has no data or the destination is full. At runtime, data may flow through the PEs implementing the mapped algorithm. For example, data may be streamed in from memory, through the fabric, and then back out to memory. Embodiments of this architecture may achieve remarkable performance efficiency relative to traditional multicore processors: for example, where compute, in the form of PEs, is simpler and more numerous than larger cores and communication are direct, e.g., as opposed to an extension of the memory system.

FIG. 5 illustrates an accelerator tile comprising an array of processing elements (PES) according to embodiments of the disclosure. The interconnect network is depicted as circuit switched, statically configured communications channels. For example, a set of channels coupled together by a switch (e.g., switch 510 in a first network and switch 511 in a second network). The first network and second network may be separate or coupled together. For
example, switch 510 may couple one or more of the four data paths (512, 514, 516, 518) together, e.g., as configured to perform an operation according to a dataflow graph. In one embodiment, the number of data paths is any plurality. Processing element (e.g., processing element 504) may be as disclosed herein, for example, as in FIG. 9. Accelerator tile 500 includes a memory/cache hierarchy interface 502, e.g., to interface the accelerator tile 500 with a memory and/or cache. A data path (e.g., 518) may extend to another tile or terminate, e.g., at the edge of a tile. A processing element may include an input buffer (e.g., buffer 506) and an output buffer (e.g., buffer 508).

[0191] Operations may be executed based on the availability of their inputs and the status of the PE. A PE may obtain operands from input channels and write results to output channels, although internal register state may also be used. Certain embodiments herein include a configurable dataflow-friendly PE. FIG. 9 shows a detailed block diagram of one such PE: the integer PE. This PE consists of several I/O buffers, an ALU, a storage register, some instruction registers, and a scheduler. Each cycle, the scheduler may select an instruction for execution based on the availability of the input and output buffers and the status of the PE. The result of the operation may then be written to either an output buffer or to a (e.g., local to the PE) register. Data written to an output buffer may be transported to a downstream PE for further processing. This style of PE may be extremely energy efficient, for example, rather than reading data from a complex, multi-ported register file, a PE reads the data from a register. Similarly, instructions may be stored directly in a register, rather than in a virtualized instruction cache.

[0192] Instruction registers may be set during a special configuration step. During this step, auxiliary control wires and state, in addition to the inter-PE network, may be used to stream in configuration across the several PEs comprising the fabric. As result of parallelism, certain embodiments of such a network may provide for rapid reconﬁguration, e.g., a tiled sized fabric may be conﬁgured in less than about 10 microseconds.

[0193] FIG. 9 represents one example conﬁguration of a processing element, e.g., in which all architectural elements are minimally sized. In other embodiments, each of the components of a processing element is independently scaled to produce new PEs. For example, to handle more complicated programs, a larger number of instructions that are executable by a PE may be introduced. A second dimension of conﬁgurability is in the function of the PE arithmetic logic unit (ALU). In FIG. 9, an integer PE is depicted which may support addition, subtraction, and various logic operations. Other kinds of PEs may be created by substituting different kinds of functional units into the PE. An integer multiplication PE, for example, might have no registers, a single instruction, and a single output buffer. Certain embodiments of a PE decompose a fused multiply add (FMA) into separate, but tightly coupled ﬂoating multiply and ﬂoating add units to improve support for multiply-accumulate workloads. PEs are discussed further below.

[0194] FIG. 6A illustrates a conﬁgurable data path network 600 (e.g., of network one or network two discussed in reference to FIG. 5) according to embodiments of the disclosure. Network 600 includes a plurality of multiplexers (e.g., multiplexers 602, 604, 606) that may be conﬁgured (e.g., via their respective control signals) to connect one or more data paths (e.g., from PEs) together. FIG. 6B illustrates a conﬁgurable ﬂow control path network 601 (e.g., network one or network two discussed in reference to FIG. 5) according to embodiments of the disclosure. A network may be a lightweight PE-to-PE network. Certain embodiments of a network may be thought of as a set of composable primitives for the construction of distributed, point-to-point data channels. FIG. 6A shows a network that has two channels enabled, the bold black line and the dotted black line. The bold black line channel is multica...
in a single cycle. In one embodiment, to improve routing bandwidth, several networks may be laid out in parallel between rows of PEs. [0198] Relative to FPGAs, certain embodiments of networks herein have three advantages: area, frequency, and programmation. Certain embodiments of networks herein operate at a coarse grain, e.g., which reduces the number configuration bits, and thereby the area of the network. Certain embodiments of networks also obtain area reduction by implementing flow control logic directly in circuitry (e.g., silicon). Certain embodiments of hardened network implementations also enjoy a frequency advantage over FPGA. Because of an area and frequency advantage, a power advantage may exist where a lower voltage is used at throughputs parity. Finally, certain embodiments of networks provide better high-level semantics than FPGA wires, especially with respect to variable timing, and thus those certain embodiments are more easily targeted by compilers. Certain embodiments of networks herein may be thought of as a set of composables primitives for the construction of distributed, point-to-point data channels. [0199] In certain embodiments, a multistage source may not assert its data valid unless it receives a ready signal from each sink. Therefore, an extra conjunction and control bit may be used in the multistage case. [0200] Like certain PEs, the network may be statically configured. During this step, configuration bits are set at each network component. These bits control, for example, the multiplexer selection and flow control function. The forward path of our network requires some bits to swing their muxes. In the example shown in FIG. 6A, four bits per hop are required: the east and west muxes utilize one bit each, while the southbound multiplexer utilizes two bits. In this embodiment, four bits may be utilized for the data path, but 7 bits may be utilized for the flow control function (e.g., in the flow control path network). Other embodiments may utilize more bits, for example, if a CSA further utilizes a north-south direction. The flow control function may utilize a control bit for each direction from which flow control can come. This may enable the setting of the sensitivity of the flow control function statically. The table 1 below summarizes the Boolean algebraic implementation of the flow control function for the network in FIG. 6B, with configuration bits capitalized. In this example, seven bits are utilized.

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>Flow Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>readyToEast</td>
<td>(EAST_WEST_SENSITIVE + readyFromWest) *</td>
</tr>
<tr>
<td>readyToWest</td>
<td>(WEST_EAST_SENSITIVE + readyFromSouth) *</td>
</tr>
<tr>
<td>readyToNorth</td>
<td>(NORTH_SOUTH_SENSITIVE + readyFromWest) *</td>
</tr>
<tr>
<td></td>
<td>(NORTH_EAST_SENSITIVE + readyFromEast) *</td>
</tr>
<tr>
<td></td>
<td>(NORTH_SOUTHSENSITIVE + readyFromSouth)</td>
</tr>
</tbody>
</table>

For the third flow control box from the left in FIG. 6B, EAST_WEST_SENSITIVE and NORTH_SOUTH_SENSITIVE are depicted as set to implement the flow control for the bold line and dotted line channels, respectively. [0201] FIG. 7 illustrates a circuit switched network 700 according to the embodiment of the disclosure. Circuit switched network 700 is coupled to a CSA component (e.g., a processing element (PE)) 702, and may likewise couple to other CSA component(s) (e.g., PEs), for example, over one or more channels that are created from switches (e.g., multiplexers) 704-726. This may include horizontal (H) switches and/or vertical (V) switches. Depicted switches may be switches in FIG. 6. Switches may include one or more registers 704A-726A to store the control values (e.g., configuration bits) to control the selection of input(s) and/or output(s) of the switch to allow values to pass from an input(s) to an output(s). In one embodiment, the switches are selectively coupled to one or more of networks 730 (e.g., sending data to the right (east (E)), 732 (e.g., sending data downward (south (S))), 734 (e.g., sending data to the left (west (W))), and/or 736 (e.g., sending data upward (north (N))). Networks 730, 732, 734, and/or 736 may be coupled to another instance of the components (or a subset of the components) in FIG. 7, for example, to create flow controlled communications channels (e.g., paths) which support communications between components (e.g., PEs) of a configurable spatial accelerator (e.g., a CSA as discussed herein). In one embodiment, a network (e.g., networks 730, 732, 734, and/or 736 or a separate network) receive a control value (e.g., configuration bits) from a source (e.g., a core) and cause that control value (e.g., configuration bits) to be stored in registers 704A-728A to cause the corresponding switches 704-728 to form the desired channels (e.g., according to a dataflow graph). Processing element 702 may also include control register(s) 702A, for example, as operation configuration register 709 in FIG. 9. Switches and other components may thus be set in certain embodiments to create data path or data paths between processing elements and/or backpressure paths for those data paths, e.g., as discussed herein. In one embodiment, the values (e.g., configuration bits) in these (core) registers 704A-728A are depicted with variables names that refer to the mux selection for the inputs, for example, with the values having a number which refers to the port number, and a letter which refers to the direction or PE output the data is coming from, e.g., where E1 in 706A refers to port number 1 coming from the east side of the network. [0202] The network(s) may be statically configured, e.g., in addition to PEs being statically configured during configuration for a dataflow graph. During the configuration step, configuration bits may be set at each network component. These bits may control, for example, the multiplexer selections to control the flow of a dataflow token (e.g., on a data path network) and its corresponding backpressure token (e.g., on a flow control path network). A network may comprise a plurality of networks, e.g., a data path network and a flow control path network. A network or plurality of networks may utilize paths of different widths (e.g., a first width, and a narrower or wider second width). In one embodiment, a data path network has a wider (e.g., bit transport) width than the width of a flow control path network. In one embodiment, each of a first network and a second network includes its own data paths and flow control paths, e.g., data path A and flow control path A and wider data path B and flow control path B. For example, a data path and flow control path for a single output buffer of a producer PE that couples to a plurality of input buffers of consumer PEs. In one embodiment, to improve routing bandwidth, several networks are laid out in parallel between rows of PEs. Like certain PEs, the network may be statically configured. During this step, configuration bits may be set at each network component. These bits control, for example,
the data path (e.g., multiplexer created data path) and/or flow control path (e.g., multiplexer created flow control path). The forward (e.g., data) path may utilize control bits to swing its switches and/or logic gates.

[0203] FIG. 8 illustrates a hardware processor tile 800 comprising an accelerator 802 according to embodiments of the disclosure. Accelerator 802 may be a CSA according to this disclosure. Tile 800 includes a plurality of cache banks (e.g., cache bank 808). Request address file (RAF) circuits 810 may be included, e.g., as discussed below in Section 2.2. ODIF may refer to an On-Die Interconnect, e.g., an interconnect stretching across an entire die connecting up all the tiles. OTI may refer to an On-Tile Interconnect, for example, stretching across a tile, e.g., connecting cache banks on the tile together.

2.1 Processing Elements

[0204] In certain embodiments, a CSA includes an array of heterogeneous PEs, in which the fabric is composed of several types of PEs each of which implement only a subset of the dataflow operators. By way of example, FIG. 9 shows a provisional implementation of a PE capable of implementing a broad set of the integer and control operations. Other PEs, including those supporting floating point addition, floating point multiplication, buffering, and certain control operations may have a similar implementation style, e.g., with the appropriate (dataflow operator) circuitry substituted for the ALU. PEs (e.g., dataflow operators) of a CSA may be configured (e.g., programmed) before the beginning of execution to implement a particular dataflow operation from among the set that the PE supports. A configuration may include one or two control words which specify an opcode controlling the ALU, steer the various multiplexers within the PE, and activate dataflow into and out of the PE channels. Dataflow operators may be implemented by microwiring these configurations bits. The depicted integer PE 900 in FIG. 9 is organized as a single-stage logical pipeline flowing from top to bottom. Data enters PE 900 from one of the local networks, where it is registered in an input buffer for subsequent operation. Each PE may support a number of wide, data-oriented and narrow, control-oriented channels. The number of provisioned channels may vary based on PE functionality, but one embodiment of an integer-oriented PE has 2 wide and 1-2 narrow input and output channels. Although the integer PE is implemented as a single-cycle pipeline, other pipelining choices may be utilized. For example, multiplication PEs may have multiple pipeline stages.

[0205] PE execution may proceed in a dataflow style. Based on the configuration microcode, the scheduler may examine the status of the PE ingress and egress buffers, and, when all the inputs for the configured operation have arrived and the egress buffer of the operation is available, orchestrates the actual execution of the operation by a dataflow operator (e.g., on the ALU). The resulting value may be placed in the configured egress buffer. Transfers between the egress buffer of one PE and the ingress buffer of another PE may occur asynchronously as buffering becomes available. In certain embodiments, PEs are provisioned such that at least one dataflow operation completes per cycle. Section 2 discussed dataflow operator encompassing primitive operations, such as add, xor, or pick. Certain embodiments may provide advantages in energy, area, performance, and latency. In one embodiment, with an extension to a PE control path, more fused combinations may be enabled. In one embodiment, the width of the processing elements is 64 bits, e.g., for the heavy utilization of double-precision floating point computation in HPC and to support 64-bit memory addressing.

2.2 Communications Networks

[0206] Embodiments of the CSA microarchitecture provide a hierarchy of networks which together provide an implementation of the architectural abstraction of latency-insensitive channels across multiple communications scales. The lowest level of CSA communications hierarchy may be the local network. The local network may be statically switched, e.g., using configuration registers to swing multiplexer(s) in the local network data-path to form fixed electrical paths between communicating PEs. In one embodiment, the configuration of the local network is set once per dataflow graph, e.g., at the same time as the PE configuration. In one embodiment, static, circuit switching optimizes for energy, e.g., when a large majority (perhaps greater than 95%) of CSA communications traffic will cross the local network. A program may include terms which are used in multiple expressions. To optimize for this case, embodiments herein provide for hardware support for multicast within the local network. Several local networks may be ganged together to form routing channels, e.g., which are interspersed (as a grid) between rows and columns of PEs. As an optimization, several local networks may be included to carry control tokens. In comparison to a FPGA interconnect, a CSA local network may be routed at the granularity of the data-path, and another difference may be a CSA’s treatment of control. One embodiment of a CSA local network is explicitly flow controlled (e.g., back-pressure). For example, for each forward data-path and multiplexer set, a CSA is to provide a backward-flowing flow control path that is physically paired with the forward data-path. The combination of the two microarchitectural paths may provide a low-latency, low-energy, low-area, point-to-point implementation of the latency-insensitive channel abstraction. In one embodiment, a CSA’s flow control lines are not visible to the user program, but they may be manipulated by the architecture in service of the user program. For example, the exception handling mechanisms described in Section 1.2 may be achieved by pulling flow control lines to a “not present” state upon the detection of an exceptional condition. This action may not only gracefully stalls those parts of the pipeline which are involved in the offending computation, but also trigger the exception for diagnostic analysis. A second network layer, e.g., the mezzanine network, may be a shared, packet switched network. Mezzanine network may include a plurality of distributed network controllers, network dataflow endpoint circuits. The mezzanine network (e.g., the network schematically indicated by the dotted box in FIG. 77) may provide more general, long range communications, e.g., at the cost of latency, bandwidth, and energy. In some programs, most communications may occur on the local network, and thus mezzanine network provisioning will be considerably reduced in comparison, for example, each PE may connect to multiple local networks, but the CSA will provision only one mezzanine endpoint per logical neighborhood of PEs. Since the mezzanine is effectively a shared network, each mezzanine network may carry multiple logically independent channels, e.g., and be provisioned with
multiple virtual channels. In one embodiment, the main function of the mezzanine network is to provide wide-range communications in-between PE(s) and between PE(s) and memory. In addition to this capability, the mezzanine may also include network dataflow endpoint circuit(s), for example, to perform certain dataflow operations. In addition to this capability, the mezzanine may also operate as a runtime support network, e.g., by which various services may access the complete fabric in a user-program-transparent manner. In this capacity, the mezzanine endpoint may function as a controller for its local neighborhood, for example, during CSA configuration. To form channels spanning a CSA tile, three subchannels and two local network channels (which carry traffic to and from a single channel in the mezzanine network) may be utilized. In one embodiment, one mezzanine channel is utilized, e.g., one mezzanine and two local-3 total network hops.

[0207] The composability of channels across network layers may be extended to higher level network layers at the inter-tile, inter-die, and fabric granularities.

[0208] FIG. 9 illustrates a processing element 900 according to embodiments of the disclosure. In one embodiment, operation configuration register 919 is loaded during configuration (e.g., mapping) and specifies the particular operation (or operations) this processing (e.g., compute) element is to perform. Register 920 activity may be controlled by that operation an output of multiplexer 916, e.g., controlled by the scheduler 914. Scheduler 914 may schedule an operation or operations of processing element 900, for example, when input data and control input arrives. Control input buffer 922 is connected to local network 902 (e.g., and local network 902 may include a data path network as in FIG. 6A and a flow control path network as in FIG. 6B) and is loaded with a value when it arrives (e.g., the network has a data bit(s) and valid bit(s)). Control output buffer 932, data output buffer 934, and/or data output buffer 936 may receive an output of processing element 900, e.g., as controlled by the operation (an output of multiplexer 916). Status register 938 may be loaded whenever the ALU 918 executes (also controlled by output of multiplexer 916). Data in control input buffer 922 and control output buffer 932 may be a single bit of multiplexer 921 (e.g., operand A) and multiplexer 923 (e.g., operand B) may source inputs.

[0209] For example, suppose the operation of this processing (e.g., compute) element is (or includes) what is called a call a pick in FIG. 3B. The processing element 900 then is to select data from either data input buffer 924 or data input buffer 926, e.g., to go to data output buffer 934 (e.g., default) or data output buffer 936. The control bit in 922 may thus indicate a 0 if selecting from data input buffer 924 or a 1 if selecting from data input buffer 926.

[0210] For example, suppose the operation of this processing (e.g., compute) element is (or includes) what is called call a switch in FIG. 3B. The processing element 900 is to output data to data output buffer 934 or data output buffer 936, e.g., from data input buffer 924 (e.g., default) or data input buffer 926. The control bit in 922 may thus indicate a 0 if outputting to data output buffer 934 or a 1 if outputting to data output buffer 936.

[0211] Multiple networks (e.g., interconnects) may be connected to a processing element, e.g., (input) networks 902, 904, 906 and (output) networks 908, 910, 912. The connections may be switches, e.g., as discussed in reference to FIGS. 6A and 6B. In one embodiment, each network includes two sub-networks (or two channels on the network), e.g., one for the data path network in FIG. 6A and one for the flow control (e.g., backpressure) path network in FIG. 6B. As one example, local network 902 (e.g., set up as a control interconnect) is depicted as being switched (e.g., connected) to control input buffer 922. In this embodiment, a data path (e.g., network as in FIG. 6A) may carry the control input value (e.g., bit or bits) (e.g., a control token) and the flow control path (e.g., network) may carry the backpressure signal (e.g., backpressure or no-backpressure token) from control input buffer 922, e.g., to indicate to the upstream producer (e.g., PE) that a new control input value is not to be loaded into (e.g., sent to) control input buffer 922 until the backpressure signal indicates there is room in the control input buffer 922 for the new control input value (e.g., from a control output buffer of the upstream producer). In one embodiment, the new control input value may not enter control input buffer 922 until both (i) the upstream producer receives the “space available” backpressure signal from “control input” buffer 922 and (ii) the new control input value is sent from the upstream producer, e.g., and this may stall the processing element 900 until that happens (and space in the target, output buffer(s) is available).

[0212] Data input buffer 924 and data input buffer 926 may perform similarly, e.g., local network 904 (e.g., set up as a data (as opposed to control) interconnect) is depicted as being switched (e.g., connected) to data input buffer 924. In this embodiment, a data path (e.g., network as in FIG. 6A) may carry the data input value (e.g., bit or bits) (e.g., a dataflow token) and the flow control path (e.g., network) may carry the backpressure signal (e.g., backpressure or no-backpressure token) from data input buffer 924, e.g., to indicate to the upstream producer (e.g., PE) that a new data input value is not to be loaded into (e.g., sent to) data input buffer 924 until the backpressure signal indicates there is room in the data input buffer 924 for the new data input value (e.g., from a data output buffer of the upstream producer). In one embodiment, the new data input value may not enter data input buffer 924 until both (i) the upstream producer receives the “space available” backpressure signal from “data input” buffer 924 and (ii) the new data input value is sent from the upstream producer, e.g., and this may stall the processing element 900 until that happens (and space in the target, output buffer(s) is available). A control output value and/or data output value may be stalled in their respective output buffers (e.g., 932, 934, 936) until a backpressure signal indicates there is available space in the output buffer for the downstream processing element(s).

[0213] A processing element 900 may be stalled from execution until its operands (e.g., a control input value and its corresponding data input value or values) are received and/or until there is room in the output buffer(s) of the processing element 900 for the data that is to be produced by the execution of the operation on those operands.

[0214] In certain embodiments, a significant source of area and energy reduction is the customization of the dataflow operations supported by each type of processing element. In one embodiment, a proper subset (e.g., most) processing elements support only a few operations (e.g., one, two, three, or four operation types), for example, an implementation choice where a floating point PE only supports one of floating point multiply or floating point add, but not both.

[0215] FIG. 10 depicts a processing element (PE) 1000 that supports (e.g., only) two operations, although the below
discussion is equally applicable for a PE that supports a single operation or more than two operations. In one embodiment, processing element 1000 supports two operations, and the configuration value being set selects a single operation for performance, e.g., to perform one or multiple instances of a single operation type for that configuration.

[0216] FIG. 10 illustrates data paths and control paths of a processing element 1000 according to embodiments of the disclosure. A processing element may include one or more of the components discussed herein, e.g., as discussed in reference to FIG. 9. Processing element 1000 includes operation configuration storage 1019 (e.g., register) to store an operation configuration value that causes the PE to perform the selected operation when its requirements are met, e.g., when the incoming operands become available (e.g., from input storage 1024 and/or input storage 1026) and when there is space available to store the output (resultant) operand or operands (e.g., in output storage 1034 and/or output storage 1036). In certain embodiments, operation configuration value (e.g., corresponding to the mapping of a dataflow graph to that PE(s)) is loaded (e.g., stored) in operation configuration storage 1019 as described herein, e.g., in section 3 below.

[0217] Operation configuration value may be a (e.g., unique) value, for example, according to the format discussed in section 3.5 below, e.g., for the operations discussed in section 3.6 below. In certain embodiments, operation configuration value includes a plurality of bits that cause processing element 1000 to perform a desired (e.g., preselected) operation, for example, performing the desired (e.g., preselected) operation when the incoming operands become available (e.g., in input storage 1024 and/or input storage 1026) and when there is space available to store the output (resultant) operand or operands (e.g., in output storage 1034 and/or output storage 1036). The depicted processing element 1000 includes two sets of operation circuitry 1025 and 1027, for example, to each perform a different operation. In certain embodiments, a PE includes status (e.g., state) storage, for example, within operation circuitry or a status register. Status storage may be modified during the operation in the course of execution. Status storage may be shared among several operations. See, for example, the status register 938 in FIG. 9, the state stored in schedulers in FIGS. 32A-32B, or the state stored in the scheduler in FIGS. 61-63B.

[0218] Depicted processing element 1000 includes an operation configuration storage 1019 (e.g., register(s)) to store an operation configuration value. In one embodiment, all of or a proper subset of a (e.g., single) operation configuration value is sent from the operation configuration storage 1019 (e.g., register(s)) to the multiplexers (e.g., multiplexer 1021 and multiplexer 1023) and/or demultiplexers (e.g., demultiplexer 1041 and demultiplexer 1043) of the processing element 1000 to steer the data according to the configuration.

[0219] Processing element 1000 includes a first input storage 1024 (e.g., input queue or buffer) coupled to (e.g., circuit switched) network 1002 and a second input storage 1026 (e.g., input queue or buffer) coupled to (e.g., circuit switched) network 1004. Network 1002 and network 1004 can be the same network (e.g., different circuit switched paths of the same network). Although two input storages are depicted, a single input storage or more than two input storages (e.g., any integer or proper subset of integers) may be utilized (e.g., with their own respective input controllers). Operation configuration value may be sent via the same network that the input storage 1024 and/or input storage 1026 are coupled to.

[0220] Depicted processing element 1000 includes input controller 1001, input controller 1003, output controller 1005, and output controller 1007 (e.g., together forming a scheduler for processing element 1000). Embodiments of input controllers are discussed in reference to FIGS. 11-20. Embodiments of output controllers are discussed in reference to FIGS. 11-20. In certain embodiments, operation circuitry (e.g., operation circuitry 1025 or operation circuitry 1027 in FIG. 10) includes a coupling to a scheduler to perform certain actions, e.g., to activate certain logic circuitry in the operations circuitry based on control provided from the scheduler.

[0221] In FIG. 10, the operation configuration value (e.g., set according to the operation that is to be performed) or a subset of less than all of the operation configuration value causes the processing element 1000 to perform the programmed operation, for example, when the incoming operands become available (e.g., from input storage 1024 and/or input storage 1026) and when there is space available to store the output (resultant) operand or operands (e.g., in output storage 1034 and/or output storage 1036). In the depicted embodiment, the input controller 1001 and/or input controller 1003 are to cause a supplying of the input operand(s) and the output controller 1005 and/or output controller 1007 are to cause a storing of the resultant of the operation on the input operand(s). In one embodiment, a plurality of input controllers are combined into a single input controller. In one embodiment, a plurality of output controllers are combined into a single output controller.

[0222] In certain embodiments, the input data (e.g., dataflow token or tokens) is sent to input storage 1024 and/or input storage 1026 by networks 1002 or networks 1004. In one embodiment, input data is stalled until there is available storage (e.g., in the targeted storage input storage 1024 or input storage 1026) in the storage that is to be utilized for that input data. In the depicted embodiment, operation configuration value (or a portion thereof) is sent to the multiplexers (e.g., multiplexer 1021 and multiplexer 1023) and/or demultiplexers (e.g., demultiplexer 1041 and demultiplexer 1043) of the processing element 1000 as control values(s) to steer the data according to the configuration. In certain embodiments, input operand selection switches 1021 and 1023 (e.g., multiplexers) allow data (e.g., dataflow tokens) from input storage 1024 and input storage 1026 as inputs to either of operation circuitry 1025 or operation circuitry 1027. In certain embodiments, result (e.g., output operand) selection switches 1037 and 1039 (e.g., multiplexers) allow data from either of operation circuitry 1025 or operation circuitry 1027 into output storage 1034 and/or output storage 1036. Storage may be a queue (e.g., first-in-first-out (FIFO) queue). In certain embodiments, an operation takes one input operand (e.g., from either of input storage 1024 and input storage 1026) and produce two resultants (e.g., stored in output storage 1034 and output storage 1036). In certain embodiments, an operation takes two or more input operands (for example, one from each input storage queue, e.g., one from each of input storage 1024 and input storage 1026) and produces a single (or plurality of) resultants (for example, stored in output storage, e.g., output storage 1034 and/or output storage 1036).
In certain embodiments, processing element 1000 is stalled from execution until there is input data (e.g., dataflow token or tokens) in input storage and there is storage space for the resultant data available in the output storage (e.g., as indicated by a backpressure value sent that indicates the output storage is not full). In the depicted embodiment, the input storage (queue) status value from path 1009 indicates (e.g., by asserting a “not empty” indication value or an “empty” indication value) when input storage 1024 contains (e.g., new) input data (e.g., dataflow token or tokens) and the input storage (queue) status value from path 1011 indicates (e.g., by asserting a “not empty” indication value or an “empty” indication value) when input storage 1026 contains (e.g., new) input data (e.g., dataflow token or tokens). In one embodiment, the input storage (queue) status value from path 1009 for input storage 1024 and the input storage (queue) status value from path 1011 for input storage 1026 is steered to the operation circuitry 1025 and/or operation circuitry 1027 (e.g., along with the input data from the input storage that is to be operated on) by multiplexer 1021 and multiplexer 1023.

In the depicted embodiment, the output storage (queue) status value from path 1013 indicates (e.g., by asserting a “not full” indication value or a “full” indication value) when output storage 1034 has available storage for (e.g., new) output data (e.g., as indicated by a backpressure token or tokens) and the output storage (queue) status value from path 1015 indicates (e.g., by asserting a “not full” indication value or a “full” indication value) when output storage 1036 has available storage for (e.g., new) output data (e.g., as indicated by a backpressure token or tokens). In the depicted embodiment, operation configuration value (or a portion thereof) is sent to both multiplexer 1041 and multiplexer 1043 to source the output storage (queue) status value(s) from the output controllers 1005 and/or 1007. In certain embodiments, operation configuration value includes a bit or bits to cause a first output storage status value to be asserted, where the first output storage status value indicates the output storage (queue) is not full or a second, different output storage status value to be asserted, where the second output storage status value indicates the output storage (queue) is full. The first output storage status value (e.g., “not full”) or second output storage status value (e.g., “full”) may be output from output controller 1005 and/or output controller 1007, e.g., as discussed below. In one embodiment, a first output storage status value (e.g., “not full”) is sent to the operation circuitry 1025 and/or operation circuitry 1027 to cause the operation circuitry 1025 and/or operation circuitry 1027, respectively, to perform the programmed operation when an input value is available in input storage (queue) and a second output storage status value (e.g., “full”) is sent to the operation circuitry 1025 and/or operation circuitry 1027 to cause the operation circuitry 1025 and/or operation circuitry 1027, respectively, to not perform the programmed operation even when an input value is available in input storage (queue).

In the depicted embodiment, dequeue (e.g., conditional dequeue) multiplexers 1029 and 1031 are included to cause a dequeue (e.g., removal) of a value (e.g., token) from a respective input storage (queue), e.g., based on operation completion by operation circuitry 1025 and/or operation circuitry 1027. The operation configuration value includes a bit or bits to cause the dequeue (e.g., conditional dequeue) multiplexers 1029 and 1031 to dequeue (e.g., remove) a value (e.g., token) from a respective input storage (queue).

In the depicted embodiment, enqueue (e.g., conditional enqueue) multiplexers 1033 and 1035 are included to cause an enqueue (e.g., insertion) of a value (e.g., token) into a respective output storage (queue), e.g., based on operation completion by operation circuitry 1025 and/or operation circuitry 1027. The operation configuration value includes a bit or bits to cause the enqueue (e.g., conditional enqueue) multiplexers 1033 and 1035 to enqueue (e.g., insert) a value (e.g., token) into a respective output storage (queue).

Certain operations hereinafter allow the manipulation of the control values sent to these queues, e.g., based on local values computed and/or stored in the PE.

In one embodiment, the dequeue multiplexers 1029 and 1031 are conditional dequeue multiplexers 1029 and 1031 that, when a programmed operation is performed, the consumption (e.g., dequeuing) of the input value from the input storage (queue) is conditionally performed. In one embodiment, the enqueue multiplexers 1033 and 1035 are conditional enqueue multiplexers 1033 and 1035 that, when a programmed operation is performed, the storing (e.g., enqueuing) of the output value for the programmed operation into the output storage (queue) is conditionally performed.

For example, as discussed herein, certain operations may make dequeuing (e.g., consumption) decisions for an input storage (queue) conditionally (e.g., based on token values) and/or enqueuing (e.g., output) decisions for an output storage (queue) conditionally (e.g., based on token values). An example of a conditional enqueue operation is a PredMerge operation that conditionally writes its outputs, so conditional enqueue multiplexer(s) will be swung, e.g., to store or not store the predmerge result into the appropriate output queue. An example of a conditional dequeue operation is a PredProp operation that conditionally reads its inputs, so conditional dequeue multiplexer(s) will be swung, e.g., to store or not store the predprop result into the appropriate input queue.

In certain embodiments, control input value (e.g., bit or bits) (e.g., a control token) is input into a respective input storage (queue), e.g., into a control input buffer as discussed herein (e.g., control input buffer 922 in FIG. 9). In one embodiment, control input value is used to make dequeuing (e.g., consumption) decisions for an input storage (queue) conditionally based on the control input value and/or enqueuing (e.g., output) decisions for an output storage (queue) conditionally based on the control input value. In certain embodiments, control output value (e.g., bit or bits) (e.g., a control token) is output into a respective output storage (queue), e.g., for example, into a control output buffer as discussed herein (e.g., control output buffer 932 in FIG. 9).

Input Controllers

FIG. 11 illustrates input controller circuitry 1100 of input controller 1001 and/or input controller 1003 of processing element 1000 in FIG. 10 according to embodiments of the disclosure. In one embodiment, each input queue (e.g., buffer) includes its own instance of input controller circuitry 1100, for example, 2, 3, 4, 5, 6, 7, 8, or more (e.g., any integer) of instances of input controller circuitry 1100. Depicted input controller circuitry 1100 includes a queue status register 1102 to store a value representing the current status of that queue (e.g., the queue status register 1102...
storing any combination of a head value (e.g., pointer) that represents the head (beginning) of the data stored in the queue, a tail value (e.g., pointer) that represents the tail (ending) of the data stored in the queue, and a count value that represents the number of (e.g., valid) values stored in the queue. For example, a count value may be an integer (e.g., two) where the queue is storing the number of values indicated by the integer (e.g., storing two values in the queue). The capacity of data (e.g., storage slots for data, e.g., for data elements) in a queue may be preselected (e.g., during programming), for example, depending on the total bit capacity of the queue and the number of bits in each element. Queue status register 1102 may be updated with the initial values, e.g., during configuration time.

[0231] Depicted input controller circuitry 1100 includes a Status determiner 1104, a Not Full determiner 1106, and a Not Empty determiner 1108. A determiner may be implemented in software or hardware. A hardware determiner may be a circuit implementation, for example, a logic circuit programmed to produce an output based on the inputs into the state machine(s) discussed below. Depicted (e.g., new) Status determiner 1104 includes a port coupled to queue status register 1102 to read and/or write to input queue status register 1102.

[0232] Depicted Status determiner 1104 includes a first input to receive a Valid value (e.g., a value indicating valid) from a transmitting component (e.g., an upstream PE) that indicates if (e.g., when) there is data (valid data) to be sent to the PE that includes input controller circuitry 1100. The Valid value may be referred to as a dataflow token. Depicted Status determiner 1104 includes a second input to receive a value or values from queue status register 1102 that represents that current status of the input queue that input controller circuitry 1100 is controlling. Optionally, Status determiner 1104 includes a third input to receive a value (from within the PE that includes input controller circuitry 1100) that indicates if (when) there is a conditional dequeue, e.g., from operation circuitry 1025 and/or operation circuitry 1027 in FIG. 10.

[0233] As discussed further below, the depicted Status determiner 1104 includes a first output to send a value on path 1110 that will cause input data (transmitted to the input queue that input controller circuitry 1100 is controlling) to be enqueued into the input queue or not enqueued into the input queue. Depicted Status determiner 1104 includes a second output to send an updated value to be stored in queue status register 1102, e.g., where the updated value represents the updated status (e.g., head value, tail value, count value, or any combination thereof) of the input queue that input controller circuitry 1100 is controlling.

[0234] Input controller circuitry 1100 includes a Not Full determiner 1106 that determines a Not Full (e.g., Ready) value and outputs the Not Full value to a transmitting component (e.g., an upstream PE) to indicate if (e.g., when) there is storage space available for input data in the input queue being controlled by input controller circuitry 1100. The Not Full (e.g., Ready) value may be referred to as a backpressure token, e.g., a backpressure token from a receiving PE sent to a transmitting PE.

[0235] Input controller circuitry 1100 includes a Not Empty determiner 1108 that determines an input storage (queue) status value and outputs (e.g., on path 1099 or path 1041 in FIG. 10) the input storage (queue) status value that indicates (e.g., by asserting a “not empty” indication value or an “empty” indication value) when the input queue being controlled contains (e.g., new) input data (e.g., dataflow token or tokens). In certain embodiments, the input storage (queue) status value (e.g., being a value that indicates the input queue is not empty) is one of the two control values (with the other being that storage for the resultants is not full) that is to stall a PE (e.g., operation circuitry 1025 and/or operation circuitry 1027 in FIG. 10) until both of the control values indicate the PE may proceed to perform its programmed operation (e.g., with a Not Empty value for the input queue(s) that provide the inputs to the PE and a Not Full value for the output queue(s) that are to store the resultant(s) for the PE operation). An example of determining the Not Full value for an output queue is discussed below in reference to FIG. 21. In certain embodiments, input controller circuitry includes any one or more of the inputs and any one or more of the outputs discussed herein.

[0236] For example, assume that the operation that is to be performed is to source data from both input storage 1024 and input storage 1026 in FIG. 10. Two instances of input controller circuitry 1100 may be included to cause a respective input value to be enqueued into input storage 1024 and input storage 1026 in FIG. 10. In this example, each input controller circuitry instance may send a Not Empty value within the PE containing input storage 1024 and input storage 1026 (e.g., to operation circuitry) to cause the PE to operate on the input values (e.g., when the storage for the resultants is also not full).

[0237] FIG. 12 illustrates enqueue circuitry 1200 of input controller 1001 and/or input controller 1003 in FIG. 11 according to embodiments of the disclosure. Depicted enqueue circuitry 1200 includes a queue status register 1202 to store a value representing the current status of the input queue 1204. Input queue 1204 may be any input queue, e.g., input storage 1024 or input storage 1026 in FIG. 10. Enqueue circuitry 1200 includes a multiplexer 1206 coupled to queue register enable ports 1208. Enqueue input 1210 is to receive a value indicating to enqueue (e.g., store) an input value into input queue 1204 or not. In one embodiment, enqueue input 1210 is coupled to path 1110 of an input controller that causes input data (e.g., transmitted to the input queue 1204 that input controller circuitry 1100 is controlling) to be enqueued into. In the depicted embodiment, the tail value from queue status register 1202 is used as the control value to control whether the input data is stored in the first slot 1204A or the second slot 1204B of input queue 1204. In one embodiment, input queue 1204 includes three or more slots, e.g., with that same number of queue register enable ports as the number of slots. Enqueue circuitry 1200 includes a multiplexer 1212 coupled to input queue 1204 that causes data from a particular location (e.g., slot) of the input queue 1204 to be output into a processing element. In the depicted embodiment, the head value from queue status register 1202 is used as the control value to control whether the output data is sourced from the first slot 1204A or the second slot 1204B of input queue 1204. In one embodiment, input queue 1204 includes three or more slots, e.g., with that same number of input ports of multiplexer 1212 as the number of slots. A Data In value may be the input data (e.g., payload) for an input storage, for example, in contrast to a Valid value which may (e.g., only) indicate (e.g., by a single bit) that input data is being sent or ready.
to be sent but does not include the input data itself. Data Out value may be sent to multiplexer 1021 and/or multiplexer 1023 in FIG. 10.

[0238] Queue status register 1202 may store any combination of a head value (e.g., pointer) that represents the head (beginning) of the data stored in the queue, a tail value (e.g., pointer) that represents the tail (ending) of the data stored in the queue, and a count value that represents the number of (e.g., valid) values stored in the queue). For example, a count value may be an integer (e.g., two) where the queue is storing the number of values indicated by the integer (e.g., storing two values in the queue). The capacity of data (e.g., storage slots for data, e.g., for data elements) in a queue may be preselected (e.g., during programming), for example, depending on the total bit capacity of the queue and the number of bits in each element. Queue status register 1202 may be updated with the initial values, e.g., during configuration time. Queue status register 1202 may be updated as discussed in reference to FIG. 11.

[0239] FIG. 13 illustrates a status determiner 1300 of input controller 1001 and/or input controller 1003 in FIG. 10 according to embodiments of the disclosure. Status determiner 1300 may be used as status determiner 1104 in FIG. 11. Depicted status determiner 1300 includes a head determiner 1302, a tail determiner 1304, a count determiner 1306, and an enqueue determiner 1308. A status determiner may include one or more (e.g., any combination) of a head determiner 1302, a tail determiner 1304, a count determiner 1306, or an enqueue determiner 1308. In certain embodiments, head determiner 1302 provides a head value that that represents the current head (e.g., starting) position of input data stored in an input queue, tail determiner 1304 provides a tail value (e.g., pointer) that represents the current tail (e.g., ending) position of the input data stored in that input queue, count determiner 1306 provides a count value that represents the number of (e.g., valid) values stored in the input queue, and enqueue determiner 1308 provides an enqueue value that indicates whether to enqueue (e.g., store) input data (e.g., an input value) into the input queue or not.

[0240] FIG. 14 illustrates a head determiner state machine 1400 according to embodiments of the disclosure. In certain embodiments, head determiner 1302 in FIG. 13 operates according to state machine 1400. In one embodiment, head determiner 1302 in FIG. 13 includes logic circuitry that is programmed to perform according to state machine 1500. State machine 1500 includes inputs for an input queue of the input queue’s: current head value (e.g., from queue status register 1102 in FIG. 11 or queue status register 1202 in FIG. 12), capacity (e.g., a fixed number), conditional dequeue value (e.g., output from conditional dequeue multiplexers 1029 and 1031 in FIG. 10), and not empty value (e.g., from Not Empty determiner 1108 in FIG. 11), state machine 1400 outputs an updated head value based on those inputs. The & symbol indicates a logical AND operation. The <= symbol indicates assignment of a new value, e.g., head <= 0 assigns the value of zero as the updated head value. In FIG. 12, an (e.g., updated) head value is used as a control input to multiplexer 1212 to select a head value from the input queue 1204.

[0241] FIG. 15 illustrates a tail determiner state machine 1500 according to embodiments of the disclosure. In certain embodiments, tail determiner 1304 in FIG. 13 operates according to state machine 1500. In one embodiment, tail determiner 1304 in FIG. 13 includes logic circuitry that is programmed to perform according to state machine 1500. State machine 1500 includes inputs for an input queue of the input queue’s: current tail value (e.g., from queue status register 1102 in FIG. 11 or queue status register 1202 in FIG. 12), capacity (e.g., a fixed number), ready value (e.g., output from Not Full determiner 1106 in FIG. 11), and valid value (for example, from a transmitting component (e.g., an upstream PE) as discussed in reference to FIG. 11 or FIG. 20). State machine 1500 outputs an updated tail value based on those inputs. The & symbol indicates a logical AND operation. The <= symbol indicates assignment of a new value, e.g., tail <= tail + 1 assigns the value of the previous tail value plus one as the updated tail value. In FIG. 12, an (e.g., updated) tail value is used as a control input to multiplexer 1206 to help select a tail slot of the input queue 1204 to store new input data into.

[0242] FIG. 16 illustrates a count determiner state machine 1600 according to embodiments of the disclosure. In certain embodiments, count determiner 1306 in FIG. 13 operates according to state machine 1600. In one embodiment, count determiner 1306 in FIG. 13 includes logic circuitry that is programmed to perform according to state machine 1600. State machine 1600 includes inputs for an input queue of the input queue’s: current count value (e.g., from queue status register 1102 in FIG. 11 or queue status register 1202 in FIG. 12), ready value (e.g., output from Not Full determiner 1106 in FIG. 11), valid value (for example, from a transmitting component (e.g., an upstream PE) as discussed in reference to FIG. 11 or FIG. 20), conditional dequeue value (e.g., output from conditional dequeue multiplexers 1029 and 1031 in FIG. 10), and not empty value (e.g., from Not Empty determiner 1108 in FIG. 11). State machine 1600 outputs an updated count value based on those inputs. The & symbol indicates a logical AND operation. The + symbol indicates an addition operation. The - symbol indicates a subtraction operation. The <= symbol indicates assignment of a new value, e.g., to the count field of queue status register 1102 in FIG. 11 or queue status register 1202 in FIG. 12. Note that the asterisk symbol indicates the conversion of a Boolean value of true to an integer 1 and a Boolean value of false to an integer 0.

[0243] FIG. 17 illustrates an enqueue determiner state machine 1700 according to embodiments of the disclosure. In certain embodiments, enqueue determiner 1308 in FIG. 13 operates according to state machine 1700. In one embodiment, enqueue determiner 1308 in FIG. 13 includes logic circuitry that is programmed to perform according to state machine 1700. State machine 1700 includes inputs for an input queue of the input queue’s: ready value (e.g., output from Not Full determiner 1106 in FIG. 11), and valid value (for example, from a transmitting component (e.g., an upstream PE) as discussed in reference to FIG. 11 or FIG. 20). State machine 1700 outputs an updated enqueue value based on those inputs. The & symbol indicates a logical AND operation. The - symbol indicates assignment of a new value. In FIG. 12, an (e.g., updated) enqueue value is used as an input on path 1210 to multiplexer 1206 to cause the tail slot of the input queue 1204 to store new input data therein.

[0244] FIG. 18 illustrates a Not Full determiner state machine 1800 according to embodiments of the disclosure. In certain embodiments, Not Full determiner 1106 in FIG. 11 operates according to state machine 1800. In one embodiment, Not Full determiner 1106 in FIG. 11 includes logic
circuitry that is programmed to perform according to state machine 1800. State machine 1800 includes inputs for an input queue of the input queue's count value (e.g., from queue status register 1102 in FIG. 11 or queue status register 1202 in FIG. 12) and capacity (e.g., a fixed number indicating the total capacity of the input queue). The < symbol indicates a less than operation, such that a ready value (e.g., a Boolean one) indicating the input queue is not full is asserted as long as the current count of the input queue is less than the input queue's capacity. In FIG. 11, an (e.g., updated) Ready (e.g., Not Full) value is sent to a transmitting component (e.g., an upstream PE) to indicate if (e.g., when) there is storage space available for additional input data in the input queue.

[0245] FIG. 19 illustrates a Not Empty determiner state machine 1900 according to embodiments of the disclosure. In certain embodiments, Not Empty determiner 1108 in FIG. 11 operates according to state machine 1900. In one embodiment, Not Empty determiner 1108 in FIG. 11 includes logic circuitry that is programmed to perform according to state machine 1900. State machine 1900 includes an input for an input queue of the input queue's count value (e.g., from queue status register 1102 in FIG. 11 or queue status register 1202 in FIG. 12). The < symbol indicates a greater than operation, such that a Not Empty value (e.g., a Boolean one) indicating the input queue is not empty is asserted as long as the current count of the input queue is greater than zero (or whatever number indicates an empty input queue). In FIG. 11, an (e.g., updated) Not Empty value is to cause the PE (e.g., the PE that includes the input queue) to operate on the input value(s), for example, when the storage for the resultant of that operation is also not full.

[0246] FIG. 20 illustrates a valid determiner state machine 2000 according to embodiments of the disclosure. In certain embodiments, Not Empty determiner 2108 in FIG. 21 operates according to state machine 2000. In one embodiment, Not Empty determiner 2108 in FIG. 21 includes logic circuitry that is programmed to perform according to state machine 2000. State machine 2000 includes an input for an output queue of the output queue’s count value (e.g., from queue status register 2102 in FIG. 21 or queue status register 2202 in FIG. 22). The > symbol indicates a greater than operation, such that a Not Empty value (e.g., a Boolean one) indicating the output queue is not empty is asserted as long as the current count of the output queue is greater than zero (or whatever number indicates an empty output queue). In FIG. 21, an (e.g., updated) valid value is sent from a transmitting (e.g., upstream PE) to the receiving PE (e.g., the receiving PE that includes the input queue being controlled by input controller 1100 in FIG. 11), e.g., and that valid value is used as the valid value in state machines 1500, 1600, and/or 1700.

Output Controllers

[0247] FIG. 21 illustrates output controller circuitry 2100 of output controller 1005 and/or output controller 1007 of processing element 1000 in FIG. 10 according to embodiments of the disclosure. In one embodiment, each output queue (e.g., buffer) includes its own instance of output controller circuitry 2100, for example, 2, 3, 4, 5, 6, 7, 8, or more (e.g., any integer) instances of output controller circuitry 2100. Depicted output controller circuitry 2100 includes a queue status register 2102 to store a value representing the current status of that queue (e.g., the queue status register 2102 storing any combination of a head value (e.g., pointer) that represents the head (beginning) of the data stored in the queue, a tail value (e.g., pointer) that represents the tail (ending) of the data stored in the queue, and a count value that represents the number of (e.g., valid) values stored in the queue). For example, a count value may be an integer (e.g., two) where the queue is storing the number of values indicated by the integer (e.g., storing two values in the queue). The capacity of data (e.g., storage slots for data, e.g., for data elements) in a queue may be prespecified (e.g., during programming), for example, depending on the total bit capacity of the queue and the number of bits in each element. Queue status register 2102 may be updated with the initial values, e.g., during configuration time. Count value may be set at zero during initialization.

[0248] Depicted output controller circuitry 2100 includes a Status determiner 2104, a Not Full determiner 2106, and a Not Empty determiner 2108. A determiner may be implemented in software or hardware. A hardware determiner may be a circuit implementation, for example, a logic circuit programmed to produce an output based on the inputs into the state machine(s) discussed below. Depicted (e.g., new) Status determiner 2104 includes a port coupled to queue status register 2102 to read and/or write to output queue status register 2102.

[0249] Depicted Status determiner 2104 includes a first input to receive a Ready value from a receiving component (e.g., a downstream PE) that indicates if (e.g., when) there is space (e.g., in an input queue thereof) for new data to be sent to the PE. In certain embodiments, the Ready value from the receiving component is sent by an input controller that includes input controller circuitry 1100 in FIG. 11. The Ready value may be referred to as a backpressure token, e.g., a backpressure token from a receiving PE sent to a transmitting PE. Depicted Status determiner 2104 includes a second input to receive a value or values from queue status register 2102 that represents that current status of the output queue that output controller circuitry 2100 is controlling. Optionally, Status determiner 2104 includes a third input to receive a value (from within the PE that includes output controller circuitry 1100) that indicates if (when) there is a conditional enqueuence, e.g., from operation circuitry 1025 and/or operation circuity 1027 in FIG. 10.

[0250] As discussed further below, the depicted Status determiner 2104 includes a first output to send a value on path 2110 that will cause output data (sent to the output queue that output controller circuitry 2100 is controlling) to be enqueued into the output queue or not enqueued into the output queue. Depicted Status determiner 2104 includes a second output to send an updated value to be stored in queue status register 2102, e.g., where the updated value represents the updated status (e.g., head value, tail value, count value, or any combination thereof) of the output queue that output controller circuitry 2100 is controlling.

[0251] Output controller circuitry 2100 includes a Not Full determiner 2106 that determines a Not Full (e.g., Ready) value and outputs the Not Full value, e.g., within the PE that includes output controller circuitry 2100, to indicate if (e.g., when) there is storage space available for output data in the output queue being controlled by output controller circuitry 2100. In one embodiment, for an output queue of a PE, a Not Full value that indicates there is no storage space available in that output queue is to cause a stall of execution of the PE (e.g., stall execution that is to cause a resultant to be stored
into the storage space) until storage space is available (e.g., and when there is available data in the input queue(s) being sourced from in that PE).

[0252] Output controller circuitry 2100 includes a Not Empty Determiner 2108 that determines an output storage (queue) status value and outputs (e.g., on path 1045 or path 1047 in FIG. 10) an output storage (queue) status value that indicates (e.g., by asserting a “not empty” indication value or an “empty” indication value) when the output queue being controlled contains (e.g., new) output data (e.g., dataflow token or tokens), for example, so that output data may be sent to the receiving PE. In certain embodiments, the output storage (queue) status value (e.g., being a value that indicates the output queue of the sending PE is not empty) is one of the two control values (with the other being that input storage of the receiving PE coupled to the output storage is not full) that is to stall transmittal of that data from the sending PE to the receiving PE until both of the control values indicate the components (e.g., PEs) may proceed to transmit that (e.g., payload) data (e.g., with a Ready value for the input queue(s) that is to receive data from the transmitting PE and a Valid value for the output queue(s) in the receiving PE that is to store the data). An example of determining the Ready value for an input queue is discussed above in reference to FIG. 11. In certain embodiments, output controller circuitry includes any one or more of the inputs and any one or more of the outputs discussed herein.

[0253] For example, assume that the operation that is to be performed is to send (e.g., sink) data into both output storage 1034 and output storage 1036 in FIG. 10. Two instances of output controller circuitry 2100 may be included to cause a respective output value(s) to be enqueued into output storage 1034 and output storage 1036 in FIG. 10. In this example, each output controller circuitry instance may send a Not Full value within the PE containing output storage 1034 and output storage 1036 (e.g., to operation circuitry) to cause the PE to operate on its input values (e.g., when the input storage to source the operation input(s) is also not empty).

[0254] FIG. 22 illustrates enqueuing circuitry 2200 of output controller 1005 and/or output controller 1007 in FIG. 10 according to embodiments of the disclosure. Depicted enqueuing circuitry 2200 includes a queue status register 2202 to store a value representing the current status of the output queue 2204. Output queue 2204 may be any output queue, e.g., output storage 1034 or output storage 1036 in FIG. 10. Enqueuing circuitry 2200 includes a multiplexer 2206 coupled to queue register enable ports 2208. Enqueuing input 2210 is to receive a value indicating to enqueue (e.g., store) an output value into output queue 2204 or not. In one embodiment, enqueuing input 2210 is coupled to path 2110 of an output controller that causes output data (e.g., transmitted to the output queue 2204 that output controller circuitry 2100 is controlling) to be enqueued into. In the depicted embodiment, the tail value from queue status register 2202 is used as the control value to control whether the output data is stored in the first slot 2204A or the second slot 2204B of output queue 2204. In one embodiment, output queue 2204 includes three or more slots, e.g., with that same number of queue register enable ports as the number of slots. Enqueuing circuitry 2200 includes a multiplexer 2212 coupled to output queue 2204 that causes data from a particular location (e.g., slot) of the output queue 2204 to be output to a network (e.g., to a downstream processing element). In the depicted embodiment, the head value from queue status register 2202 is used as the control value to control whether the output data is sourced from the first slot 2204A or the second slot 2204B of output queue 2204. In one embodiment, output queue 2204 includes three or more slots, e.g., with that same number of output ports of multiplexer 2212 as the number of slots. A data in value may be the output data (e.g., payload) for an output storage, for example, in contrast to a Valid value which may (e.g., only) indicate (e.g., by a single bit) that output data is being sent or ready to be sent but does not include the output data itself. Data out value may be sent to multiplexer 1041 and/or multiplexer 1043 in FIG. 10.

[0255] Queue status register 2202 may store any combination of a head value (e.g., pointer) that represents the head (beginning) of the data stored in the queue, a tail value (e.g., pointer) that represents the tail (ending) of the data stored in the queue, and a count value that represents the number of (e.g., valid) values stored in the queue. For example, a count value may be an integer (e.g., two) where the queue is storing the number of values indicated by the integer (e.g., storing two values in the queue). The capacity of data (e.g., storage slots for data, e.g., for data elements) in a queue may be preselected (e.g., during programming), for example, depending on the total bit capacity of the queue and the number of bits in each element. Queue status register 2202 may be updated with the initial values, e.g., during configuration time. Queue status register 2202 may be updated as discussed in reference to FIG. 21.

[0256] FIG. 23 illustrates a status determiner 2300 of output controller 1005 and/or output controller 1007 in FIG. 10 according to embodiments of the disclosure. Status determiner 2300 may be used as status determiner 2104 in FIG. 21. Depicted status determiner 2300 includes a head determiner 2302, a tail determiner 2304, a count determiner 2306, and an enqueue determiner 2308. A status determiner may include one or more (e.g., any combination) of a head determiner 2302, a tail determiner 2304, a count determiner 2306, or an enqueue determiner 2308. In certain embodiments, head determiner 2302 provides a head value that represents the current head (e.g., starting) position of output data stored in an output queue, tail determiner 2304 provides a tail value (e.g., pointer) that represents the current tail (e.g., ending) position of the output data stored in that output queue, count determiner 2306 provides a count value that represents the number of (e.g., valid) values stored in the output queue, and enqueue determiner 2308 provides an enqueue value that indicates whether to enqueue (e.g., store) output data (e.g., an output value) into the output queue or not.

[0257] FIG. 24 illustrates a head determiner state machine 2400 according to embodiments of the disclosure. In certain embodiments, head determiner 2302 in FIG. 23 operates according to state machine 2400. In one embodiment, head determiner 2302 in FIG. 23 includes logic circuitry that is programmed to perform according to state machine 2400. State machine 2400 includes inputs for an output queue of: a current head value (e.g., from queue status register 2102 in FIG. 21 or queue status register 2202 in FIG. 22), capacity (e.g., a fixed number), ready value (e.g., output from a Not Full determiner 2106 in FIG. 21 from a receiving component (e.g., a downstream PE) for its input queue), and valid value (for example, from a Not Empty determiner of the PE as discussed in reference to FIG. 21 or FIG. 29). State machine 2400 outputs an updated head value based on those inputs. The symbol indicates a logical AND operation. The <=
symbol indicates assignment of a new value, e.g., head <= 0 assigns the value of zero as the updated head value. In FIG. 22, an (e.g., updated) head value is used as a control input to multiplexer 2212 to select a head value from the output queue 2204.

[0258] FIG. 25 illustrates a tail determiner state machine 2500 according to embodiments of the disclosure. In certain embodiments, tail determiner 2304 in FIG. 23 operates according to state machine 2500. In one embodiment, tail determiner 2304 in FIG. 23 includes logic circuitry that is programmed to perform according to state machine 2500. State machine 2500 includes inputs for an output queue of: a current tail value (e.g., from queue status register 2102 in FIG. 21 or queue status register 2202 in FIG. 22), capacity (e.g., a fixed number), a Not Full value (e.g., from a Not Full determiner of the PE as discussed in reference to FIG. 21 or FIG. 28), and a Conditional Enqueue value (e.g., output from conditional enqueuing multiplexers 1033 and 1035 in FIG. 10). State machine 2500 outputs an updated tail value based on those inputs. The && symbol indicates a logical AND operation. The <= symbol indicates assignment of a new value, e.g., tail := tail+1 assigns the value of the previous tail value plus one as the updated tail value. In FIG. 22, an (e.g., updated) tail value is used as a control input to multiplexer 2212 to help select a tail slot of the output queue 2204 to store new output data into.

[0259] FIG. 26 illustrates a count determiner state machine 2600 according to embodiments of the disclosure. In certain embodiments, count determiner 2306 in FIG. 23 operates according to state machine 2600. In one embodiment, count determiner 2306 in FIG. 23 includes logic circuitry that is programmed to perform according to state machine 2600. State machine 2600 includes inputs for an output queue of: current count value (e.g., from queue status register 2102 in FIG. 21 or queue status register 2202 in FIG. 22), ready value (e.g., output from a Not Full determiner 2106 in FIG. 21 from a receiving component (e.g., a downstream PE) for its input queue), valid value (for example, from a Not Empty determiner of the PE as discussed in reference to FIG. 21 or FIG. 29), Conditional Enqueue value (e.g., output from conditional enqueuing multiplexers 1033 and 1035 in FIG. 10), and Not Full value (e.g., from a Not Full determiner of the PE as discussed in reference to FIG. 21 or FIG. 28). State machine 2600 outputs an updated count value based on those inputs. The && symbol indicates a logical AND operation. The => symbol indicates an add operation. The ~ symbol indicates a subtraction operation. The <= symbol indicates assignment of a new value, e.g., to the count field of queue status register 2102 in FIG. 21 or queue status register 2202 in FIG. 22. Note that the asterisk symbol indicates the conversion of a Boolean value of true to an integer 1 and a Boolean value of false to an integer 0.

[0260] FIG. 27 illustrates an enqueue determiner state machine 2700 according to embodiments of the disclosure. In certain embodiments, enqueue determiner 2308 in FIG. 23 operates according to state machine 2700. In one embodiment, enqueue determiner 2308 in FIG. 23 includes logic circuitry that is programmed to perform according to state machine 2700. State machine 2700 includes inputs for an output queue of: ready value (e.g., output from a Not Full determiner 1106 in FIG. 11 from a receiving component (e.g., a downstream PE) for its input queue), and valid value (for example, from a Not Empty determiner of the PE as discussed in reference to FIG. 21 or FIG. 29). State machine 2700 outputs an updated enqueue value based on those inputs. The && symbol indicates a logical AND operation. The ~ symbol indicates assignment of a new value. In FIG. 22, an (e.g., updated) enqueue value is used as an input on path 2210 to multiplexer 2206 to cause the tail slot of the output queue 2204 to store new output data therein.

[0261] FIG. 28 illustrates a Not Full determiner state machine 2800 according to embodiments of the disclosure. In certain embodiments, Not Full determiner 2106 in FIG. 21 operates according to state machine 2800. In one embodiment, Not Full determiner 2106 in FIG. 21 includes logic circuitry that is programmed to perform according to state machine 2800. State machine 2800 includes inputs for an output queue of the output queue’s count value (e.g., from queue status register 2102 in FIG. 21 or queue status register 2202 in FIG. 22) and capacity (e.g., a fixed number indicating the total capacity of the output queue). The <= symbol indicates a less than operation, such that a ready value (e.g., a Boolean one) indicating the output queue is not full is asserted as long as the current count of the output queue is less than the output queue’s capacity. In FIG. 21, an (e.g., updated) Not Full value is produced and used within the PE to indicate if (e.g., when) there is storage space available for additional output data in the output queue.

[0262] FIG. 29 illustrates a Not Empty determiner state machine 2900 according to embodiments of the disclosure. In certain embodiments, Not Empty determiner 1108 in FIG. 11 operates according to state machine 2900. In one embodiment, Not Empty determiner 1108 in FIG. 11 includes logic circuitry that is programmed to perform according to state machine 2900. State machine 2900 includes an input for an input queue of the input queue’s count value (e.g., from queue status register 1102 in FIG. 11 or queue status register 1202 in FIG. 12). The > symbol indicates a greater than operation, such that a Not Empty value (e.g., a Boolean one) indicating the input queue is not empty is asserted as long as the current count of the input queue is greater than zero (or whatever number indicates an empty input queue). In FIG. 11, an (e.g., updated) Not Empty value is to cause the PE (e.g., the PE that includes the input queue) to operate on the input value(s), for example, when the storage for the resultant of that operation is also not full.

[0263] FIG. 30 illustrates a valid determiner state machine 3000 according to embodiments of the disclosure. In certain embodiments, Not Empty determiner 2106 in FIG. 21 operates according to state machine 3000. In one embodiment, Not Empty determiner 2108 in FIG. 21 includes logic circuitry that is programmed to perform according to state machine 3000. State machine 3000 includes an input for an output queue of the output queue’s count value (e.g., from queue status register 2102 in FIG. 21 or queue status register 2202 in FIG. 22). The > symbol indicates a greater than operation, such that a Not Empty value (e.g., a Boolean one) indicating the output queue is not empty is asserted as long as the current count of the output queue is greater than zero (or whatever number indicates an empty output queue). In FIG. 21, an (e.g., updated) valid value is sent from a transmitting (e.g., upstream) PE to the receiving PE (e.g., sent by the transmitting PE that includes the output queue being controlled by output controller 2100 in FIG. 21), e.g., and that valid value is used as the valid value in state machines 2400, 2600, and/or 2700.
In certain embodiments, a state machine includes a plurality of single bit width input values (e.g., 0s or 1s), and produces a single output value that has a single bit width (e.g., a 0 or a 1).

In certain embodiments, a first LIC channel may be formed between an output of a first PE to an input of a second PE, and a second LIC channel may be formed between an output of the second PE and an input of a third PE. As an example, a ready value may be sent on a first path of a LIC channel by a receiving PE to a transmitting PE and a valid value may be sent on a second path of the LIC channel by the transmitting PE to the receiving PE. As an example, see FIGS. 11 and 21. Additionally, a LIC channel in certain embodiments may include a third path for transmittal of the (e.g., payload) data, e.g., transmitted after the ready value and valid value are asserted.

2.3 in-Network Storage Element

In certain embodiments of configurable spatial accelerators, the placement of operators within a physical array may affect the performance and/or efficiency of the accelerators. In one embodiment, if communicating operators (e.g., PEs configured as dataflow operators) are assigned to physical locations (e.g., PEs) that are too far apart, then the accelerator may operate at a greatly reduced performance level, e.g., by reducing the operational frequency of data being clocked into storage.

Certain embodiments herein provide for one or more in-network storage elements that allow the accelerator to move data in a pipelined fashion. In certain embodiments, this preserves the design performance and operating frequency where only the lengthened communication path is degraded. In certain embodiments, an in-network storage elements includes a queue to solve the timing closure problem. In one embodiment, a plurality of in-network storage elements are disposed periodically within the network (e.g., along one or more edges of a tile or accelerator). In certain embodiments, a queue within an in-network storage element is guarded by one or more switches (e.g., multiplexers) that can bypass the queue, e.g., if the queue is not needed.

Certain embodiments of in-network storage elements discussed herein improve operation timing and/or reduce tool chain complexity. Certain embodiments of in-network storage elements discussed herein simplifies the process of mapping programs to a configurable spatial accelerator and/or reduce the performance penalty of long-distance communications (e.g., communications that take more than one of a desired clock cycle to complete).

Certain embodiments of in-network storage elements discussed herein allow for easy incorporation into place-and-route tools as these in-network storage resources are exclusively reserved for the place-and-route tool. Certain embodiments of in-network storage elements discussed herein are more (e.g., power and/or circuit area space) efficient than using a processing element (PE) to merely store data (e.g., not to operate on the data), for example, when the in-network storage element is more power efficient and/or has a smaller area that it occupies. Certain embodiments of in-network storage elements utilize back pressure (e.g., ready) and/or other control values discussed herein (e.g., enqueue, valid, etc.). Certain embodiments of in-network storage elements operate in an asynchronous fashion, e.g., are not limited to operating in a synchronous fashion.

Certain embodiments of in-network storage elements discussed herein allow for the use of bypassable storage queues in a stably circuit-switched communications network within a coarse-grained array of an accelerator. In one embodiment, the in-network storage elements are fully participatory in the communications protocols of the accelerator (e.g., processing elements thereof), e.g., including back-pressure. In one embodiment, the in-network storage elements guarantee timing closure within a configurable spatial accelerator (e.g., network thereof).

Certain embodiments of in-network storage elements discussed herein reduce software complexity by simplifying any timing problems of a software place-and-route tool. For example, instead of provisioning PEs merely for storage, the PEs can be used for computations, e.g., where the place-and-route tool preserves a fixed number of the PEs (e.g., at fixed locations in the fabric) for its own usage and/or the place and route tool is to comprehend how to route through or around these resources. Certain embodiments of in-network storage elements discussed herein allow a software place-and-route tool to add storage to any path that utilizes it, for example, by simply configuring an in-network storage element accordingly, e.g., with no complexity in the tool required for this use.

Certain embodiments of in-network storage elements discussed herein reduce hardware complexity by, rather than provisioning one or more PEs for the purpose of assisting in timing closure, instead one or more in-network storage elements are provisioned (e.g., with their own controller). In one embodiment, an in-network storage elements takes up about ¼th of the area of a single processing element, e.g., and thus confers both a substantial energy and area savings above using a processing element merely for in-network storage.

Certain embodiments of in-network storage elements discussed herein allow the in-network storage elements to be bypassed, e.g., to reduce latency and energy use in the case that two communicating PEs are physically local.

Certain embodiments of in-network storage elements discussed herein allow a configurable spatial accelerator to guarantee frequency (e.g., iso-frequency) operating points. Certain embodiments of in-network storage elements discussed herein allow for an increase in the storage on dataflow arcs, e.g., to balance path latency and improve performance.

In certain embodiments, a connection between a first PE and a second PE in a (e.g., circuit switched) network includes an in-network storage element that includes a bypass feature. Certain embodiments herein include bypassable storage buffers on certain of the networks. In one embodiment, the in-network storage element or elements are added to any of the networks 730, 732, 734, or 736 discussed in reference to FIG. 7.

FIG. 31 illustrates an accelerator tile 3100 comprising an array of processing elements and in-network storage elements according to embodiments of the disclosure. The interconnect network is depicted as circuit switched, statically configured communications channels. For example, a set of channels coupled together by a switch (e.g., switch 3110 in a first network and switch 3111 in a second network). The first network and second network may be separate or coupled together. For example, switch 3110 may couple one or more of the four data paths (3112, 3114, 3116, 3118) together, e.g., as configured to perform an
operation according to a dataflow graph. In one embodiment, the number of data paths is any plurality. Processing element (e.g., processing element 3104) may be as disclosed herein, for example, as in FIG. 9. Accelerator tile 3100 includes a memory/cache hierarchy interface 3102, e.g., to interface the accelerator tile 3100 with a memory and/or cache. A data path (e.g., 3118) may extend to another tile or terminate, e.g., at the edge of a tile. A processing element may include an input buffer (e.g., buffer 3106) and an output buffer (e.g., buffer 3108). In certain embodiments, one or more in-network storage elements (3120, 3124, 3126) are included to store values therein (e.g., instead of taking up storage in the PEs). In one embodiment, an in-network storage element includes a queue (e.g., queue 3122 of in-storage network element 3120) that, in a first mode, allows for loading of data into the queue from a (e.g., transmitter) processing element and storing of data from the queue to a (e.g., receiver) processing element (e.g., or other network storage element). In certain embodiments, an in-network storage element includes a queue bypass pathway therein, for example, as enabled by a configuration value stored in the in-network storage element, for example, during configuration (e.g., before runtime).

0276] FIG. 32A illustrates a first processing element 3200A, a second processing element 3200B, a third processing element 3200C, a first in-network storage element 3201A, and a second in-network storage element 3201M coupled together by a network 3210 according to embodiments of the disclosure. In one embodiment, network 3210 is a circuit switched network, e.g., configured to send data from first PE 3200A to second PE 3200B and/or third PE 3200C.

0277] In one embodiment, a circuit switched network 3210 includes (i) a path to send data from first PE 3200A to second PE 3200B and/or third PE 3200C, e.g., for operations to be performed on that data by second PE 3200B and/or third PE 3200C, and (ii) a flow control path to send control data that controls (or is used to control) the sending of that data from first PE 3200A to second PE 3200B and/or third PE 3200C.

0278] In certain embodiments, one or more of in-network storage elements 3201A-3201M (e.g., where M is any integer) may be coupled (for example, via setting one or more switches, e.g., as discussed in reference to FIG. 7) between processing elements (e.g., in the data path and the flow control path) to allow (i) storage of data within a queue(s) of an in-network storage element and/or (ii) bypass of the queue(s) of the in-network storage element.

0279] Data path may send a data (e.g., valid) value when data is in a buffer of an in-network storage element (e.g., when data is in buffer 3242 of in-network storage element 3201A) and/or an output buffer or buffers of a PE (e.g., when data is in control output buffer 3232A, first data output buffer 3234A, or second data output buffer 3236A of first PE 3200A). In one embodiment, each output buffer includes its own data path, e.g., for its own data value from producer PE to consumer PE or producer to in-network storage element and in-network storage element to consumer PE. In certain embodiments, each data path may include an in-network storage element coupled therebetween PEs. Components in PE are examples, for example, a PE may include only a single (e.g., data) input buffer and/or a single (e.g., data) output buffer, or any plurality of buffers. Flow control path may send control data that controls (or is used to control) the sending of corresponding data from first PE 3200A (e.g., control output buffer 3232A, first data output buffer 3234A, or second data output buffer 3236A thereof) to second PE 3200B and/or third PE 3200C. Flow control path may send control data that controls (or is used to control) the sending of corresponding data from PE 3200A (e.g., control output buffer 3232A, first data output buffer 3234A, or second data output buffer 3236A thereof) to in-network storage element 3201A (e.g., buffer 3242 thereof) and/or send control data that controls (or is used to control) the sending of corresponding data from in-network storage element 3201A (e.g., buffer 3242 thereof) to the second PE 3200B (e.g., control input buffer 3222B, first data input buffer 3224B, or second data input buffer 3226B) and/or the third PE 3200C (e.g., control input buffer 3222C, first data input buffer 3224C, or second data input buffer 3226C).

0280] Flow control data may include a backpressure value from each consumer PE (or aggregated from all consumer PEs, e.g., with an AND logic gate) and/or an in-network storage element. Flow control data may include a backpressure value, for example, indicating the buffer of the second PE 3200B (e.g., control input buffer 3222B, first data input buffer 3224B, or second data input buffer 3226B) and/or the buffer of the third PE 3200C (e.g., control input buffer 3222C, first data input buffer 3224C, or second data input buffer 3226C) where the data (e.g., from control output buffer 3232A, first data output buffer 3234A, or second data output buffer 3236A of first PE 3200A) is to-be-stored is (e.g., in the current cycle) full or has an empty slot (e.g., empty in the current cycle or next cycle) (e.g., transmission attempt). Flow control data may include a backpressure (e.g., ready or not ready value) from an in-network storage element, e.g., indicating the queue thereof is full or has an empty slot.

0281] Data path and flow path may extend directly from the producer PE to the consumer PE or indirectly from the producer PE to an in-network storage element, and from the in-network storage element to the consumer PE.

0282] Flow control data may include a speculation value and/or success value. Network 3210 may include a speculation path (e.g., to transport a speculation value) and/or success path (e.g., to transport a success value). In one embodiment, a success path follows (e.g., is parallel to) the data path, e.g., is sent from the producer PE to the consumer PEs. In one embodiment, a speculation path follows (e.g., is parallel to) the backpressure path, e.g., is sent from a consumer PE to the producer PE. In one embodiment, each consumer PE has its own flow control path, e.g., in a circuit switched network 3210, to its producer PE. In one embodiment, each consumer PEs flow control path is combined into an aggregated flow control path for its producer PE.

0283] Turning to the depicted PEs, processing elements 3200A-C include operation configuration registers 3219A-C that may be loaded during configuration (e.g., mapping) and specify the particular operation or operations (for example, and indicate whether to enable multicast mode and/or the in-network operations discussed herein) that processing (e.g., compute) element and network is to perform.

0284] In-network storage element 3201A includes a configuration storage (e.g., register) 3256 to set the operations (e.g., bypass mode or buffer mode) of the in-network storage element. In certain embodiments, the value in configuration storage 3256 is stored during configuration, e.g., by a core coupled to the accelerator.
In certain embodiments, a buffer of an in-network storage element includes a plurality of slots to permit full throughput, for example, where a transmitter PE is not allowed to send data into an occupied storage slot, e.g., where the transmitter PE does not have knowledge of the pre-existing value will move. In one embodiment, the transmission latency is one cycle to transmit between each component (e.g., PE and in-network storage element) within a configurable spatial accelerator.

FIG. 32B illustrates an in-network storage element 3201A of the circuit switched network of FIG. 32A configured to provide in-network storage according to embodiments of the disclosure. Although certain paths from in-network storage element 3201A are shown as connected to two downstream PEs 3200B-3-C, in other embodiments, in-network storage element 3201A may be connected between a single upstream PE (e.g., PE 3200A) and a single downstream PE (e.g., PEs 3200B or PEs 3200C).

Depicted in-network storage element 3201A includes a buffer 3242. Buffer may have multiple slots therein (e.g., first-in-first-out buffer) or have only a single slot. In certain embodiments, configuration storage 3256 (e.g., register) has a configuration value stored therein (e.g., during configuration time), and when the configuration value is a first value, a data value from upstream path 3250 is stored into a slot of buffer 3242 (and/or sent from a slot of buffer 3242 to output path 3252, e.g., in a following cycle) and, when the configuration value is a second value, a data value from upstream path 3250 is steered downstream (e.g., with no delay greater than the delay caused by the physical path) via bypass path 3244 to downstream path 3252 (e.g., to a downstream PE (e.g., receiver PE)), for example, without being stored within (and/or modified by) in-network storage element 3201A. In certain embodiments, controller 3240 controls the selection of (i) bypass mode that utilizes bypass path 3244 or (ii) buffer mode that utilizes buffer 3242 storage based on the configuration value stored in configuration storage 3256. In certain embodiments, a data value may include the data itself (e.g., payload) along with an “valid out” value that indicates the data itself is valid for a buffer to store. In certain embodiments, a “valid out” value is sent with the flow control values, e.g., from port 3208A of PE 3200A to port 3208B of PE 3200B or port 3208C of PE 3200C. As discussed further herein (e.g., in reference to FIG. 34), ports (e.g., port 3208A of upstream (TX) PE 3200A to port 3208B of downstream (RX) PE 3200B or port 3208C of downstream (RX) PE 3200C) may be used to send ready, valid in, and/or valid out values to PEs and/or in-network storage elements (e.g., in-network element 3201A). In one embodiment, the path 3250 extending from buffer 3232A carries a “valid out” value and the data value itself (e.g., the path 3250 may include two, respective wires). In certain embodiments, port 3208A of PE 3200A, port 3208B of PE 3200B, and/or port 3208C of PE 3200C can send and receive data. For example, port 3208A of PE 3200A can (i) receive a flow control value(s) (e.g., a back-pressure value) from a downstream component (e.g., in-network storage element 3201A, PE 3200B, or PE 3200C) and/or (ii) send a “valid out” value to a downstream component.

In certain embodiments, (i) when in bypass mode, a flow control value(s) (e.g., a back-pressure value) received on downstream path 3248 is steered upstream by controller 3240 to upstream path 3246 (e.g., to an upstream PE (e.g., transmitter PE)), for example, without being modified by in-network storage element 3201A, when in bypass mode or (ii) when in buffer mode, controller 3240 is to send a flow control value to upstream path 3246 (e.g., to an upstream PE (e.g., transmitter PE)) based on the status of buffer 3242 (e.g., a “queue ready” value when storage space is available in buffer 3242).

As one example, first (e.g., as producer) PE 3200A includes a (e.g., input) port 3208A(1-6) coupled to network 3210, e.g., to receive a backpressure value from second (e.g., as consumer) PE 3200B or third (e.g., as consumer) PE 3200C. In one circuit switched configuration, (e.g., input) port 3208A(1-6) (e.g., having a plurality of parallel inputs (1), (2), (3), (4), (5), and (6)) is to receive a respective backpressure value from each one of control input buffer 3222B, first data input buffer 3224B, and second data input buffer 3226B and/or control input buffer 3222C, first data input buffer 3224C, and second data input buffer 3226C.

In one embodiment, a circuit switched backpressure path (e.g., channel) is formed by setting switches coupled to wires between an input (e.g., input 1, 2, or 3) of port 3208A and an output (e.g., output 1, 2, or 3) of port 3208B to send a backpressure token (e.g., a value indicating no storage is available in an input buffer/queue) for one of control input buffer 3222B, first data input buffer 3224B, or second data input buffer 3226B of second PE 3200B. Additionally or alternatively, a (e.g., different) circuit switched backpressure path (e.g., channel) is formed by setting switches coupled to wires between an input (e.g., a different input of input 1, 2, or 3 (or one of more than 3 inputs in another embodiment) of port 3208A and an output (e.g., output 1, 2, or 3) of port 3208C to send a backpressure token (e.g., a value indicating no storage is available in an input buffer/queue) for one of control input buffer 3222C, first data input buffer 3224C, or second data input buffer 3226C of third PE 3200C.

In certain embodiments, output buffer 3232A of PE 3200A is coupled to in-network storage element, and, (i) when the configuration value in configuration storage 3256 is a first value, a data value stored in output buffer 3232A of PE 3200A is sent through upstream path 3250 (e.g., upstream from in-network storage element 3201A) and is steered downstream (e.g., with no delay greater than the delay caused by the physical path) via bypass path 3244 to downstream path 3252 (e.g., downstream from in-network storage element 3201A) to an input buffer of a downstream PE (e.g., PE 3200B or PE 3200C), for example, without being stored within (and/or modified by) in-network storage element 3201A, and (ii) when the configuration value in configuration storage 3256 is a second value, a (e.g., different) data value stored in buffer 3232A is sent through upstream path 3250 into a slot of buffer 3242 of in-network storage element 3201A. In one embodiment, that (e.g., different) value stored in buffer 3242 is sent to downstream PE (e.g., PE 3200B or PE 3200C), e.g., when the input buffer of the downstream PE has an available storage space (e.g., slot). In certain embodiments, the controller 3240 is an instance of the controller 3300 in FIG. 33.

FIG. 33 illustrates controller circuitry 3300 of controller 3240 of in-network storage element 3201A in FIG. 32B according to embodiments of the disclosure. In one embodiment, each queue (e.g., buffer) of an in-network storage element (e.g., having multiple queues) includes its own instance of controller circuitry 3300, for example, 2, 3, 4, 5, 6, 7, 8, or more (e.g., any integer) instances of
controller circuitry 3300. Depicted controller circuitry 3300 includes a queue status register 3302 to store a value representing the current status of that queue (e.g., the queue status register 3302 storing any combination of a head value (e.g., pointer) that represents the head (beginning) of the data stored in the queue, a tail value (e.g., pointer) that represents the tail value (ending) of the data stored in the queue, and a count value that represents the number of (e.g., valid) values stored in the queue). For example, a count value may be an integer (e.g., two) where the queue is storing the number of values indicated by the integer (e.g., storing two values in the queue). The capacity of data (e.g., storage slots for data, e.g., for data elements) in a queue may be preselected (e.g., during programming), for example, depending on the total bit capacity of the queue and the number of bits in each element. Queue status register 3302 may be updated with the initial values, e.g., during configuration time.

[0293] Depicted controller circuitry 3300 includes a Status determiner 3304, a Not Full determiner 3306, and a Not Empty determiner 3308. A determiner may be implemented in software or hardware. A hardware determiner may be a circuit implementation, for example, a logic circuit programmed to produce an output based on the inputs into the state machine(s) discussed below. Depicted (e.g., new) Status determiner 3304 includes a port coupled to queue status register 3302 to read and/or write to queue status register 3302.

[0294] Depicted Status determiner 3304 includes a first input to receive a Valid value (e.g., a Not Empty value) from a transmitting component (e.g., an upstream PE) that indicates if (e.g., when) there is data (valid data) to be sent to the in-network storage element that includes controller circuitry 3300. The Valid value may be referred to as a dataflow token. Depicted Status determiner 3304 includes a second input to receive a value or values from queue status register 3302 that represents the current status of the queue that controller circuitry 3300 is controlling. Status determiner 3304 includes a third input to receive a Ready value (e.g., a Not Full value) from a receiving component (e.g., a downstream PE) that indicates if (e.g., when) there is available space in the receiving component. The Ready value may be referred to as a backpressure token.

[0295] As discussed further below, the depicted Status determiner 3304 includes a first output to send a value on path 3310 that will cause input data (transmitted to the queue that controller circuitry 3300 is controlling) to be enqueued into the queue or not enqueued into the queue. Depicted Status determiner 3304 includes a second output to send an updated value to be stored in queue status register 3302, e.g., where the updated value represents the updated status (e.g., head value, tail value, count value, or any combination thereof) of the queue that controller circuitry 3300 is controlling.

[0296] Controller circuitry 3300 includes a Not Full determiner 3306 that determines a Not Full (e.g., Queue Ready) value and outputs the Not Full (e.g., Queue Ready) value within the in-network storage element that includes controller circuitry 3300 to indicate if (e.g., when) the queue of the in-network storage element that includes controller circuitry 3300 has storage space for a new value. The Not Full (e.g., Queue Ready) value may be referred to as a backpressure token, e.g., a backpressure token for the queue within an in-network storage element for internal use by the controller of the in-network storage element.

[0297] Controller circuitry 3300 includes a Not Empty determiner 3308 that determines an input storage (queue) status value and outputs a (queue valid) status value that indicates (e.g., by asserting a “not empty” indication value or an “empty” indication value) when the queue being controlled contains (e.g., new) input data. In certain embodiments, controller circuitry includes any one or more of the inputs and any one or more of the outputs discussed herein.

[0298] FIG. 34 illustrates enqueue and dequeue circuitry for use with the controller in FIG. 33 according to embodiments of the disclosure. Depicted enqueue and dequeue circuitry 3400 includes a queue status register 3402 to store a value representing the current status of the queue 3404. Queue 3404 may be any queue, e.g., buffer 3242 in FIG. 323. Enqueue and dequeue circuitry 3400 includes a multiplexer 3406 coupled to queue register enable ports 3408. Enqueue input 3410 is to receive a value indicating to enqueue (e.g., store) an input value into queue 3404 or not. In one embodiment, enqueue input 3410 is coupled to path 3310 of a controller that causes input data (e.g., transmitted to the queue 3404 that controller circuitry 3300 is controlling) to be enqueued into. In the depicted embodiment, the tail value from queue status register 3402 is used as the control value to control whether the input data is stored in the first slot 3404A or the second slot 3404B of queue 3404. In one embodiment, queue 3404 includes three or more slots, e.g., with that same number of queue register enable ports as the number of slots. Enqueue and dequeue circuitry 3400 includes a multiplexer 3412 coupled to queue 3404 that causes data from a particular location (e.g., slot) of the queue 3404 to be output into bypass switch 3414. In the depicted embodiment, the head value from queue status register 3402 is used as the control value to control whether the output data is sourced from the first slot 3404A or the second slot 3404B of queue 3404. In one embodiment, queue 3404 includes three or more slots, e.g., with that same number of input ports of multiplexer 3412 as the number of slots. A Data In value may be the input data (e.g., payload) for an input storage, for example, in contrast to a Valid value which may (e.g., only) indicate (e.g., by a single bit) that input data is being sent or ready to be sent but does not include the input data itself. In certain embodiments, other queue implementations may be used.

[0299] Queue status register 3402 may store any combination of a head value (e.g., pointer) that represents the head (beginning) of the data stored in the queue, a tail value (e.g., pointer) that represents the tail (ending) of the data stored in the queue, and a count value that represents the number of (e.g., valid) values stored in the queue. For example, a count value may be an integer (e.g., two) where the queue is storing the number of values indicated by the integer (e.g., storing two values in the queue). The capacity of data (e.g., storage slots for data, e.g., for data elements) in a queue may be preselected (e.g., during programming), for example, depending on the total bit capacity of the queue and the number of bits in each element. Queue status register 3402 may be updated with the initial values, e.g., during configuration time. Queue status register 3402 may be updated as discussed in reference to FIG. 33.

[0300] In certain embodiments, (i) when the configuration value in configuration storage 3419 is a first (e.g., buffer mode) value, (e.g., output) bypass switch 3414 outputs a
data value from a slot of queue 3404 (e.g., a Data In value stored in queue 3404 from a transmitter (TX) PE), and (ii) when the configuration value in configuration storage 3419 is a second (e.g., bypass mode) value, bypass switch 3414 outputs a data value (e.g., Data In) from an upstream component (e.g., transmitter (TX) PE) on bypass path 3444 (e.g., and bypass switch 3414 outputs the corresponding Valid In value from the upstream component (e.g., transmitter (TX) PE)) to a downstream component (e.g., receiver (RX) PE). In one embodiment, the storage of data into queue 3404 is controlled by controller 3300 in FIG. 33, e.g., with the Data In value and Valid In value sourced from the upstream component (e.g., transmitter (TX) PE) and the Ready (Queue Ready) value sourced from controller 3300 instead of from the downstream component (e.g., receiver (RX) PE).

In certain embodiments, (i) when the configuration value in configuration storage 3419 is a first (e.g., buffer mode) value, (e.g., backpressure) bypass switch 3416 outputs (e.g., to the transmitter (TX) PE) a Ready value that indicates when the queue 3404 can accept a new value (e.g., when queue 3404 is not full), e.g., the Queue Ready value sourced from controller 3300 in FIG. 33, and (ii) when the configuration value in configuration storage 3419 is a second (e.g., bypass mode) value, (e.g., backpressure) bypass switch 3416 outputs (e.g., to the transmitter (TX) PE) a Ready value that indicates when the downstream component (e.g., receiver (RX) PE) can accept a new value (e.g., when the targeted input buffer of the RX PE is not full). Values from (e.g., TX and RX) PE input and output controllers may be generated and/or sourced according to the discussion of FIGS. 10-30.

FIG. 35 illustrates a status determiner 3500 of the controller 3300 in FIG. 33 according to embodiments of the disclosure. Depicted status determiner 3500 includes a head determiner 3502, a tail determiner 3504, a count determiner 3506, and an enqueue determiner 3508. A status determiner may include one or more (e.g., any combination) of a head determiner 3502, a tail determiner 3504, a count determiner 3506, or an enqueue determiner 3508. In certain embodiments, head determiner 3502 provides a head value that represents the current head (e.g., starting) position of input data stored in a queue, tail determiner 3504 provides a tail value (e.g., pointer) that represents the current tail (e.g., ending) position of the input data stored in that queue, count determiner 3506 provides a count value that represents the number of (e.g., valid) values stored in the queue, and enqueue determiner 3508 provides an enqueue value that indicates whether to enqueue (e.g., store) input data (e.g., an input value) into the queue or not in the next cycle (e.g., state).

FIG. 36 illustrates a head determiner state machine 3600 according to embodiments of the disclosure. In certain embodiments, head determiner 3502 in FIG. 35 operates according to state machine 3600. In one embodiment, head determiner 3502 in FIG. 35 includes logic circuitry that is programmed to perform according to state machine 3600. State machine 3600 includes inputs for a queue of the queue’s current head value (e.g., from queue status register 3302 in FIG. 33 or queue status register 3402 in FIG. 34), capacity (e.g., a fixed number), Ready value (for example, output from a receiver (RX) PE, e.g., sourced from port 32083 in FIG. 32B), and not empty value (e.g., from Not Empty determiner 3308 in FIG. 33). State machine 3600 outputs an updated head value based on those inputs. The && symbol indicates a logical AND operation. The <= symbol indicates assignment of a new value, e.g., head <= 0 assigns the value of zero as the updated head value. In FIG. 34, an (e.g., updated) head value is used as a control input to multiplexer 3412 to select a head value from the queue 3404.

FIG. 37 illustrates a tail determiner state machine 3700 according to embodiments of the disclosure. In certain embodiments, tail determiner 3504 in FIG. 35 operates according to state machine 3700. In one embodiment, tail determiner 3504 in FIG. 35 includes logic circuitry that is programmed to perform according to state machine 3700. State machine 3700 includes inputs for a queue of the queue’s current tail value (e.g., from queue status register 3302 in FIG. 33 or queue status register 3402 in FIG. 34), capacity (e.g., a fixed number or set to a value during configuration), ready value (e.g., output from Not Full determiner 3306 in FIG. 33), and valid value (for example, from a transmitting component (e.g., an upstream PE) as discussed in reference to FIG. 33 or FIG. 20). State machine 3700 outputs an updated tail value based on those inputs. The && symbol indicates a logical AND operation. The <= symbol indicates assignment of a new value, e.g., tail <= tail+1 assigns the value of the previous tail value plus one as the updated tail value. In FIG. 34, an (e.g., updated) tail value is used as a control input to multiplexer 3406 to help select a tail slot of the queue 3404 to store new input data into.

FIG. 38 illustrates a count determiner state machine 3800 according to embodiments of the disclosure. In certain embodiments, count determiner 3506 in FIG. 35 operates according to state machine 3800. In one embodiment, count determiner 3506 in FIG. 35 includes logic circuitry that is programmed to perform according to state machine 3800. State machine 3800 includes inputs for a queue of the queue’s current count value (e.g., from queue status register 3302 in FIG. 33 or queue status register 3402 in FIG. 34), ready value (e.g., output from Not Full determiner 3306 in FIG. 33), valid value (for example, from a transmitting component (e.g., an upstream PE) as discussed in reference to FIG. 33 or FIG. 20), Queue Ready value (e.g., output from Not Full determiner 3306 in FIG. 33), and not empty value (e.g., from Not Empty determiner 3308 in FIG. 33). State machine 3800 outputs an updated count value based on those inputs. The && symbol indicates a logical AND operation. The + symbol indicates an addition operation. The -= symbol indicates a subtraction operation. The <=
symbol indicates assignment of a new value, e.g., to the count field of queue status register 3302 in FIG. 33 or queue status register 3402 in FIG. 34. Note that the asterisk symbol indicates the conversion of a Boolean value of true to an integer 1 and a Boolean value of false to an integer 0.

[0306] FIG. 39 illustrates an enqueuer determiner state machine 3900 according to embodiments of the disclosure. In certain embodiments, enqueuer determiner 3508 in FIG. 35 operates according to state machine 3900. In one embodiment, enqueuer determiner 3508 in FIG. 35 includes logic circuitry that is programmed to perform according to state machine 3900. State machine 3900 includes inputs for a queue of the queue’s ready value (e.g., output from Not Full determiner 3306 in FIG. 33), and valid value (for example, from a transmitting component (e.g., an upstream PE) as discussed in reference to FIG. 33 or FIG. 20). State machine 3900 outputs an updated enqueuer value based on those inputs. The && symbol indicates a logical AND operation. The ~ symbol indicates assignment of a new value. In FIG. 34, an (e.g., updated) enqueuer value is used as an input to enqueuer input path 3410 to multiplexer 3406 to cause the tail slot of the queue 3404 to store new input data therein.

[0307] FIG. 40 illustrates a Queue Ready (Not Full) determiner state machine 4000 according to embodiments of the disclosure. In certain embodiments, Not Full determiner 3306 in FIG. 33 operates according to state machine 4000. In one embodiment, Not Full determiner 3306 in FIG. 33 includes logic circuitry that is programmed to perform according to state machine 4000. State machine 4000 includes inputs for a queue of the queue’s count value (e.g., from queue status register 3302 in FIG. 33 or queue status register 3402 in FIG. 34) and capacity (e.g., a fixed number indicating the total capacity of the queue). The < symbol indicates a less than operation, such that a ready value (e.g., a Boolean one) indicating the queue is not full is asserted as long as the current count of the queue is less than the queue’s capacity. In FIG. 33, an (e.g., updated) Ready (e.g., Not Full) value is sent to a transmitting component (e.g., an upstream PE) to indicate if (e.g., when) there is storage space available for additional input data in the queue.

[0308] FIG. 41 illustrates a Queue Valid (Not Empty) determiner state machine 4100 according to embodiments of the disclosure. In certain embodiments, Not Empty determiner 3308 in FIG. 33 operates according to state machine 4100. In one embodiment, Not Empty determiner 3308 in FIG. 33 includes logic circuitry that is programmed to perform according to state machine 4100. State machine 4100 includes an input for a queue of the queue’s count value (e.g., from queue status register 3302 in FIG. 33 or queue status register 3402 in FIG. 34). The > symbol indicates a greater than operation, such that a Not Empty value (e.g., a Boolean one) indicating the queue is not empty is asserted as long as the current count of the queue is greater than zero (or whatever number indicates an empty queue). In FIG. 33, an (e.g., updated) Not Empty value is used to cause the PE (e.g., the PE that includes the queue) to operate on the input value(s), for example, when the storage for the resultant of that operation is also not full.

[0309] In one embodiment, in-network storage elements belong uniquely to one, directional network. In another embodiment, a single, in-network storage element switches between accessing data in two networks (e.g., the first network in the foreground of FIG. 31 and the second network in the background of FIG. 31), for example, by adding a switch at the head of the storage and inbound control. In one embodiment, the switch is configured depending on which network is using the storage. Similarly, additional configuration may be provided by a switch for outbound control and data selection.

[0310] FIG. 42 illustrates controller circuitry 4200 of a controller of an in-network storage element for multiple networks according to embodiments of the disclosure. In comparison to FIG. 33, the inputs and outputs are switched between a first set of a transmitter PE and receiver PE pair (e.g., TX 0 and RX 0) and a second set of a transmitter PE and receiver PE pair (e.g., TX 1 and RX 1). In one embodiment, each queue (e.g., buffer) of an in-network storage element (e.g., having multiple queues) includes its own instance of controller circuitry 4200, for example, 2, 3, 4, 5, 6, 7, 8, or more (e.g., any integer) instances of controller circuitry 4200. Depicted controller circuitry 4200 includes a queue status register 4202 to store a value representing the current status of that queue (e.g., the queue status register 4202 storing any combination of a head value (e.g., pointer) that represents the head (beginning) of the data stored in the queue, a tail value (e.g., pointer) that represents the tail (ending) of the data stored in the queue, and a count value that represents the number of (e.g., valid) values stored in the queue). For example, a count value may be an integer (e.g., two) where the queue is storing the number of values indicated by the integer (e.g., storing two values in the queue). The capacity of data (e.g., storage slots for data, e.g., for data elements) in a queue may be preselected (e.g., during programming), for example, depending on the total bit capacity of the queue and the number of bits in each element. Queue status register 4202 may be updated with the initial values, e.g., during configuration time.

[0311] Depicted controller circuitry 4200 includes a Status determiner 4204, a Not Full determiner 4206, and a Not Empty determiner 4208. A determiner may be implemented in software or hardware. A hardware determiner may be a circuit implementation, for example, a logic circuit programmed to produce an output based on the inputs into the state machine(s) discussed below. Depicted (e.g., new) Status determiner 4204 includes a port coupled to queue status register 4202 to read and/or write to queue status register 4202.

[0312] Depicted Status determiner 4204 includes a first input to receive a Valid value (e.g., a Not Empty value) from a transmitting component (e.g., an upstream PE) that indicates if (e.g., when) there is data (valid data) to be sent to the in-network storage element that includes controller circuitry 4200. The Valid value may be referred to as a dataflow token. Depicted Status determiner 4204 includes a second input to receive a value or values from queue status register 4202 that represents the current status of the queue that controller circuitry 4200 is controlling. Status determiner 4204 includes a third input to receive a Ready value (e.g., a Not Full value) from a receiving component (e.g., a downstream PE) that indicates if (e.g., when) there is available space in the receiving component. The Ready value may be referred to as a backpressure token.

[0313] As discussed further below, the depicted Status determiner 4204 includes a first output to send a value on path 4210 that will cause input data (transmitted to the queue that controller circuitry 4200 is controlling) to be enqueued into the queue or not enqueued into the queue. Depicted Status determiner 4204 includes a second output to send an
updated value to be stored in queue status register 4202, e.g., where the updated value represents the updated status (e.g., head value, tail value, count value, or any combination thereof) of the queue that controller circuitry 4200 is controlling.

[0314] Controller circuitry 4200 includes a Not Full determiner 4206 that determines a Not Full (e.g., Queue Ready) value and outputs the Not Full (e.g., Queue Ready) value within the in-network storage element that includes controller circuitry 4200 to indicate if (e.g., when) the queue of the in-network storage element that includes controller circuitry 4200 has storage space for a new value. The Not Full (e.g., Queue Ready) value may be referred to as a backpressure token, e.g., a backpressure token for the queue within an in-network storage element for internal use by the controller of the in-network storage element.

[0315] Controller circuitry 4200 includes a Not Empty determiner 4208 that determines an input storage (queue) status value and outputs a (queue valid) status value that indicates (e.g., by asserting a “not empty” indication value or an “empty” indication value) when the queue being controlled contains (e.g., new) input data. In certain embodiments, controller circuitry includes any one or more of the inputs and any one or more of the outputs discussed herein.

[0316] FIG. 43 illustrates enqueue and dequeue circuitry 4300 for use with the controller in FIG. 42 according to embodiments of the disclosure. Depicted enqueue and dequeue circuitry 4300 includes a queue status register 4302 to store a value representing the current status of the queue 4304. In one embodiment, configuration value includes a network select field that, when storing a first value, sources data from a first network for a first set of a transmitter PE and receiver PE pair (e.g., TX 0 and RX 0) and from a second network for a second set of a transmitter PE and receiver PE pair (e.g., TX 1 and RX 1). Multiple switches (4314, 4316, 4318, 4320, 4322, 4324, and 4326) are included to switch enqueue and dequeue circuitry 4300 between the first network and the second network.

[0317] Enqueue and dequeue circuitry 4300 includes a multiplexer 4306 coupled to queue register enable ports 4308. Enqueue input 4310 is to receive a value indicating to enqueue (e.g., store) an input value into queue 4304 or not. In one embodiment, enqueue input 4310 is coupled to path 3310 of a controller that causes input data (e.g., transmitted to the queue 4304 that controller circuitry 3300 is controlling) to be enqueued into. In the depicted embodiment, the tail value from queue status register 4302 is used as the control value to control whether the input data is stored in the first slot 4304A or the second slot 4304B of queue 4304. In one embodiment, queue 4304 includes three or more slots, e.g., with that same number of queue register enable ports as the number of slots. Enqueue and dequeue circuitry 4300 includes a multiplexer 4312 coupled to queue 4304 that causes data from a particular location (e.g., slot) of the queue 4304 to be output into bypass switch 4314. In the depicted embodiment, the head value from queue status register 4302 is used as the control value to control whether the output data is sourced from the first slot 4304A or the second slot 4304B of queue 4304. In one embodiment, queue 4304 includes three or more slots, e.g., with that same number of input ports of multiplexer 4312 as the number of slots. A Data In value may be the input data (e.g., payload) for an input storage, for example, in contrast to a Valid value which may (e.g., only) indicate (e.g., by a single bit) that input data is being sent or ready to be sent but does not include the input data itself.

[0318] Queue status register 4302 may store any combination of a head value (e.g., pointer) that represents the head (beginning) of the data stored in the queue, a tail value (e.g., pointer) that represents the tail (ending) of the data stored in the queue, and a count value that represents the number of (e.g., valid) values stored in the queue. For example, a count value may be an integer (e.g., two) where the queue is storing the number of values indicated by the integer (e.g., storing two values in the queue). The capacity of data (e.g., storage slots for data, e.g., for data elements) in a queue may be preselected (e.g., during programming), for example, depending on the total bit capacity of the queue and the number of bits in each element. Queue status register 4302 may be updated with the initial values, e.g., during configuration time. Queue status register 4302 may be updated as discussed in reference to FIG. 33.

[0319] In certain embodiments, the configuration value includes a first, mode field to indicate buffer or bypass mode and a second, network field to indicate network one or network two. In certain embodiments, a first set of transmitter PE (TX 0) and receiver PE (RX 0) are coupled to a first network (e.g., network 0) in FIG. 31 (e.g., the network in the foreground in FIG. 31 that includes switch 3110) and a second set of transmitter PE (TX 1) and receiver PE (RX 1) are coupled to a second network (e.g., network 1) in FIG. 31 (e.g., the network in the background in FIG. 31 that includes switch 3111).

[0320] In certain embodiments, (i) when the configuration value in configuration storage 4319 includes a first (e.g., network 0) network selection value, the “data network select” switch 4314 outputs data into queue 4304 from network 0 (e.g., transmitter PE 0), and (ii) when the configuration value in configuration storage 4319 includes a second (e.g., network 1) network selection value, the “data network select” switch 4314 outputs data into queue 4304 from network 1 (e.g., transmitter PE 1).

[0321] In certain embodiments, (i) when the configuration value in configuration storage 4319 includes a first (e.g., buffer mode) mode value, (e.g., output) bypass switch 4316 inputs a data value from a slot of queue 4304 (e.g., a Data In value stored in queue 4304 from a transmitter (TX) PE) and bypass switch 4316 outputs that data value via network 0 to receiver PE 0 when the configuration value in configuration storage 4319 includes a first (e.g., network 0) network selection value and bypass switch 4318 inputs a data value from a slot of queue 4304 (e.g., a Data In value stored in queue 4304 from a transmitter (TX) PE) and bypass switch 4318 outputs that data value via network 1 to receiver PE 1 when the configuration value in configuration storage 4319 includes a second (e.g., network 1) network selection value, and (ii) when the configuration value in configuration storage 4319 includes a second (e.g., bypass mode) mode value, bypass switch 4316 outputs a data value (e.g., Data In) from a network 0 upstream component (e.g., transmitter (TX) PE 0) when the configuration value in configuration storage 4319 includes a first (e.g., network 0) network selection value and bypass switch 4318 outputs a data value (e.g., Data In) from a network 1 upstream component (e.g., transmitter (TX) PE 1) when the configuration value in configuration storage 4319 includes a second (e.g., network 1) network selection value.
In one embodiment, the storage of data into queue 4304 is controlled by controller 4200 in FIG. 42. e.g., with the Data In value and Valid In value sourced from the respective upstream component (e.g., transmitter (TX) PE) and the Ready (Queue Ready) value sourced from controller 3300 instead of from the respective downstream component (e.g., receiver (RX) PE).

In certain embodiments, (i) when the configuration value in configuration storage 4319 includes a first (e.g., buffer mode) mode value and a second (e.g., network 0) network selection value, (e.g., backpressure) bypass switch 4320 outputs (e.g., to the transmitter (TX) PE 0) a Ready value that indicates when the queue 4304 can accept a new value (e.g., when queue 4304 is not full), e.g., the Queue Ready value sourced from controller 4200 in FIG. 42, and (ii) when the configuration value in configuration storage 4319 includes a second (e.g., bypass mode) mode value and a first (e.g., network 0) network selection value, (e.g., backpressure) bypass switch 4320 outputs (e.g., to the transmitter (TX) PE 0) a Ready value that indicates when the downstream component (e.g., receiver (RX) PE 0) can accept a new value (e.g., when the targeted input buffer of the RX PE 0 is not full).

In certain embodiments, (i) when the configuration value in configuration storage 4319 includes a first (e.g., buffer mode) mode value and a second (e.g., network 1) network selection value, (e.g., backpressure) bypass switch 4322 outputs (e.g., to the transmitter (TX) PE 1) a Ready value that indicates when the queue 4304 can accept a new value (e.g., when queue 4304 is not full), e.g., the Queue Ready value sourced from controller 4200 in FIG. 42, and (ii) when the configuration value in configuration storage 4319 includes a second (e.g., bypass mode) mode value and a first (e.g., network 1) network selection value, (e.g., backpressure) bypass switch 4322 outputs (e.g., to the transmitter (TX) PE 1) a Ready value that indicates when the downstream component (e.g., receiver (RX) PE 1) can accept a new value (e.g., when the targeted input buffer of the RX PE 1 is not full).

In certain embodiments, (i) when the configuration value in configuration storage 4319 includes a first (e.g., network 0) network selection value, valid (e.g., valid in) select switch 4324 outputs a Valid value (Valid In) from network 0 for transmitter (TX) PE 0, and (ii) when the configuration value in configuration storage 4319 includes a second (e.g., network 1) network selection value, Valid (e.g., Valid In) select switch 4324 outputs a Valid value (Valid In) from network 1 for transmitter (TX) PE 1, e.g., switch 4324 outputs its value to controller 4200 in FIG. 42.

In certain embodiments, (i) when the configuration value in configuration storage 4319 includes a first (e.g., network 0) network selection value, Ready (e.g., Not Full) select switch 4326 outputs a Ready value (Not Full) from network 0 for transmitter (TX) PE 0, and (ii) when the configuration value in configuration storage 4319 includes a second (e.g., network 1) network selection value, Ready (e.g., Not Full) select switch 4326 outputs a Ready (e.g., Not Full) from network 1 for transmitter (TX) PE 1, e.g., switch 4326 outputs its value to controller 4200 in FIG. 42.

Values from (e.g., TX and RX) PE input and output controllers may be generated and/or sourced according to the discussion of FIGS. 10-30.

One example of the integration of bypass-able storage into a coarse grained fabric network is shown in reference to FIGS. 44A-44B. Although two separate configuration values are shown, it is possible to combine them. In one embodiment, an in-network storage element includes a queue (e.g., buffer) tied into the directional routing network, e.g., the networks shown in FIG. 7, along with some new control, e.g., as described herein. Certain embodiments include a switch (e.g., multiplexer) which selects either the original network path or the storage-augmented path. In certain embodiments, the choice of provisioning of in-network storage is a function of the timing of a particular microarchitecture. For example, networks which can transmit and receive data that traverses a distance (e.g., wire length) between a certain number of PEs (e.g., 3, 4, 5, 6, 7, 8, 9, 10, or any other number) in a single cycle may place storage at every 3rd, given number of PEs (e.g., every 3rd, 4th, 5th, etc. PE). This choice can be highly asymmetric within a single routing microarchitecture, for example, horizontal and vertical networks may have different provisioning. As noted herein, certain components of a CSA use the disclosed back-pressure and communications protocols (micro-protocols), and thus in those embodiments, an in-network storage element may be included that obeys the communications protocol (e.g., in which transmitters and receivers send signals to one another to transmit data transfers). A two and four wire protocol are discussed herein. Certain embodiments of in-network storage elements include a controller to allow bypassing of the storage component (e.g., buffer of) an in-network storage element, e.g., to select whether to use the inbound control from upstream and/or downstream PEs or to use the control of the network storage itself.

FIG. 44A illustrates a configurable data path network 4401 for use with an in-network storage element 4420 in a buffer mode according to embodiments of the disclosure.

FIG. 44B illustrates a configurable data path network 4402 for use with an in-network storage element 4420 in a bypass mode according to embodiments of the disclosure.

FIG. 44C illustrates a configurable flow control path network 4403 for use with an in-network storage element 4420 in a buffer mode according to embodiments of the disclosure.

FIG. 44D illustrates a configurable flow control path network 4404 for use with an in-network storage element 4420 in a bypass mode according to embodiments of the disclosure.

Depicted networks include a plurality of multiplexers that may be configured (e.g., via their respective control signals) to connect one or more data paths (e.g., from PEs and in-network storage elements) together. In one embodiment, in-network storage element 4420 is an instance of any of the in-network storage elements discussed herein, e.g., in-network storage element 3201A in FIG. 32B. Depicted in-network storage element 4420 includes a controller 4440 to control the selection of (i) bypass mode that utilizes bypass path 4444 or (ii) buffer mode that utilizes buffer 4442 storage based on the configuration value stored in configuration storage 4456.

In FIG. 44A, in-network storage element 4420 is in the buffer mode such that input data (e.g., payload) received by in-network storage element 4420 is stored into buffer 4442, e.g., and output from the buffer 4442 from switch 4414 as controlled by controller 4440.
In FIG. 44C, in-network storage element 4420 is in the buffer mode such that flow control data (for example, backpressure (e.g., ready) from a RX and/or valid from a TX) is sent to in-network storage element 4420, e.g., to controller 4440 to control the flow of data into and out of buffer 4442.

In FIG. 44B, in-network storage element 4420 is in the bypass mode such that input data (e.g., payload) received by in-network storage element 4420 is sent through the in-network storage element 4420 on bypass path 4444 and not stored into buffer 4442, e.g., and output from the buffer 4442 from switch 4414 as controlled by controller 4440.

In FIG. 44D, in-network storage element 4420 is not used (e.g., it is in the bypass mode) such that flow control data (for example, backpressure (e.g., ready) from a RX and/or valid from a TX) is not sent into in-network storage element 4420, e.g., is not sent to controller 4440 to control the flow of data into and out of buffer 4442.

FIG. 45 illustrates a flow diagram 4500 according to embodiments of the disclosure. Depicted flow 4500 includes storing a first configuration value in a configuration register within a first processing element that causes the first processing element to perform an operation according to the first configuration value, and a second configuration value in a configuration register within a second processing element that causes the second processing element to perform an operation according to the second configuration value 4502; coupling a queue on an in-network storage element, of a circuit switched interconnect network between the first processing element and the second processing element, to an output queue of a plurality of output queues of the first processing element 4504; controlling enqueue and dequeue of values into the plurality of output queues of the first processing element according to the first configuration value with an output controller in the first processing element 4506; controlling enqueue and dequeue of values into a plurality of input queues of the second processing element according to the second configuration value with an input controller in the second processing element 4508; storing a third configuration value in a configuration register within the in-network storage element 4510; and switching the in-network storage element between a first mode that provides a value stored in the queue of the in-network storage element by the output queue of the first processing element to an input queue of the plurality of input queues of the second processing element when the third configuration value is a first value, and into a second mode that bypasses the queue of the in-network storage element and provides a value from the output queue of the first processing element to the input queue of the second processing element when the configuration value is a second value.

In one embodiment, an apparatus (e.g., CSA) includes a plurality of processing elements that each comprise: a configuration register within a respective processing element to store a configuration value that causes the respective processing element to perform an operation according to the configuration value, a plurality of input queues, an input controller to control enqueue and dequeue of values into the plurality of input queues according to the configuration value, a plurality of output queues, and an output controller to control enqueue and dequeue of values into the plurality of output queues according to the configuration value; a circuit switched interconnect network between the plurality of processing elements to transfer values between the plurality of processing elements; and an in-network storage element of the circuit switched interconnect network comprising: a queue coupled to an output queue of a plurality of output queues of a first processing element of the plurality of processing elements, a switch, a configuration register of the in-network storage element to store a configuration value, and a controller that switches the switch into a first mode that provides a value stored in the queue of the in-network storage element by the output queue of the first processing element to an input queue of a plurality of input queues of a second processing element of the plurality of processing elements when the configuration value is a first value, and into a second mode that bypasses the queue of the in-network storage element and provides a value from the output queue of the first processing element to the input queue of the second processing element when the configuration value is a second value. The controller of the in-network storage element may send a valid out value to the second processing element: when the output queue of the first processing element stores the value and the configuration value in the configuration register of the in-network storage element is the first value; when the output queue of the first processing element stores the value and the configuration value in the configuration register of the in-network storage element is the second value. When the input queue of the second processing element stores the value from the output queue of the first processing element or the value from the queue of the in-network storage element, an input controller of the second processing element may send a not empty value to operation circuitry of the second processing element to indicate to the second processing element to begin the operation on the value stored in the input queue of the second processing element. The controller of the in-network storage element may send a ready value to the first processing element: when the input queue of the second processing element is not full and the configuration value in the configuration register of the in-network storage element is the first value; and when the queue of the in-network storage element is not full and the configuration value in the configuration register of the in-network storage element is the second value. The output controller of the first processing element may dequeue the value from the output queue of the first processing element after: the ready value is received from the in-network storage element by the first processing element; and a valid out value is asserted by the first processing element, through the in-network storage element, to the second processing element when the configuration value in the configuration register of the in-network storage element is the first value, and asserted by the first processing element to the in-network storage element when the configuration value in the configuration register of the in-network storage element is the second value. The controller of the in-network storage element may send a ready value to the first processing element: when the input queue of the second processing element is not full and the configuration value in the configuration register of the in-network storage element is the first value; and when the queue of the in-network storage element is not full and the configuration value in the configuration register of the in-network storage element is the second value. When the configuration value in the configuration register of the in-network storage element is the first value, the input controller of the second processing element may store the value within the input queue of the second processing element.
element in response to receipt of the ready value; and when
the configuration value in the configuration register of the
in-network storage element is the second value, the control-
er of the in-network storage element may store the value
within the queue of the in-network storage element in
response to receipt of the ready value. The in-network
storage element may include a second switch, and the
configuration value may include a network select field to
switch between a first network having a first set of the first
processing element and the second processing element when
the network select field is a first value, and a second network
having a second, different set of the first processing element
and the second processing element when the network select
field is a second value.

[0340] In another embodiment, a method includes storing a
first configuration value in a configuration register within
a first processing element that causes the first processing
element to perform an operation according to the first
configuration value, and a second configuration value in a
configuration register within a second processing element
that causes the second processing element to perform an
operation according to the second configuration value; cou-
ping a queue of an in-network storage element, of a circuit
switched interconnect network between the first processing
element and the second processing element, to an output
queue of a plurality of output queues of the first processing
element; controlling enqueue and dequeue of values into the
plurality of output queues of the first processing element
according to the first configuration value with an output
controller in the first processing element; controlling
enqueue and dequeue of values into a plurality of input
queues of the second processing element according to the
second configuration value with an input controller in the
second processing element; storing a third configuration
value in a configuration register within the in-network
storage element; and switching the in-network storage ele-
ment between a first mode that provides a value stored in the
queue of the in-network storage element by the output queue
of the first processing element to an input queue of the
plurality of input queues of the second processing element
when the third configuration value is a first value, and into
a second mode that bypasses the queue of the in-network
storage element and provides a value from the output queue
of the first processing element to the input queue of the
second processing element when the third configuration
value is a second value. The method may include sending a
valid out value to the second processing element from the
in-network storage element when the input queue of the
second processing element is not full and the third configu-
ration value in the configuration register of the in-network
storage element is the first value; and when the queue of the
in-network storage element is not full and the third configu-
ration value in the configuration register of the in-network
storage element is the second value. The method may
include the output controller of the first processing element
dequeuing the value from the output queue of the first
processing element after the ready value is received from
the in-network storage element by the first processing ele-
ment; and a valid out value is asserted by the first processing
element, through the in-network storage element, to the
second processing element when the third configuration
value in the configuration register of the in-network storage
element is the first value, and asserted by the first processing
element to the in-network storage element when the third
configuration value in the configuration register of the
in-network storage element is the second value. The method
may include sending a ready value to the first processing
element from the in-network storage element: when the
input queue of the second processing element is not full and
the third configuration value in the configuration register of the
in-network storage element is the first value; and when the
queue of the in-network storage element is not full and the third
configuration value in the configuration register of the in-network
storage element is the second value. The method may
include the input controller of the second processing ele-
ment storing the value within the input queue of the
second processing element in response to receipt of the
ready value; and, when the third configuration value in the
configuration register of the in-network storage element is
the second value, the in-network storage element storing the
value within the queue of the in-network storage element in
response to receipt of the ready value. The method may
include switching the in-network storage element between
a first network having a first set of the first processing
 element and the second processing element when a network select
field of the third configuration value is a first value, and a
second network having a second, different set of the first
processing element and the second processing element when
the network select field of the third configuration value is a
second value.

[0341] In yet another embodiment, a processor includes a
core with a decoder to decode an instruction into a decoded
instruction and an execution unit to execute the decoded
instruction; and a hardware accelerator coupled to the core,
the hardware accelerator comprising: a plurality of process-
ing elements that each comprise: a configuration register
within a respective processing element to store a configura-
tion value that causes the respective processing element to
perform an operation according to the configuration value, a
plurality of input queues, an input controller to control
enqueue and dequeue of values into the plurality of input
queues according to the configuration value, a plurality of
output queues, and an output controller to control enqueue
and dequeue of values into the plurality of output queues
according to the configuration value, a circuit switched
interconnect network between the plurality of processing
elements to transfer values between the plurality of process-
ing elements, and an in-network storage element of the
circuit switched interconnect network comprising: a queue
coupled to an output queue of a plurality of output queues of a first processing element of the plurality of processing elements, a switch, a configuration register of the in-network storage element to store a configuration value, and a controller that switches the switch into a first mode that provides a value stored in the queue of the in-network storage element by the output queue of the first processing element to an input queue of a plurality of input queues of a second processing element of the plurality of processing elements when the configuration value is a first value, and into a second mode that bypasses the queue of the in-network storage element and provides a value from the output queue of the first processing element to the input queue of the second processing element when the configuration value is a second value. The controller of the in-network storage element may send a valid out value to the second processing element: when the output queue of the first processing element stores the value and the configuration value in the configuration register of the in-network storage element is the first value; and when the queue of the in-network storage element stores the value and the configuration value in the configuration register of the in-network storage element is the second value. When the input queue of the second processing element stores the value from the output queue of the first processing element or the value from the queue of the in-network storage element, an input controller of the second processing element may send a not empty value to operation circuitry of the second processing element to indicate to the second processing element to begin the operation on the value stored in the input queue of the second processing element. The controller of the in-network storage element may send a ready value to the first processing element: when the input queue of the second processing element is not full and the configuration value in the configuration register of the in-network storage element is the first value; and when the queue of the in-network storage element is not full and the configuration value in the configuration register of the in-network storage element is the second value. The output controller of the first processing element may dequeue the value from the output queue of the first processing element after: the ready value is received from the in-network storage element by the first processing element; and a valid out value is asserted by the first processing element, through the in-network storage element, to the second processing element when the configuration value in the configuration register of the in-network storage element is the first value, and asserted by the first processing element to the in-network storage element when the configuration value in the configuration register of the in-network storage element is the second value. The controller of the in-network storage element may send a ready value to the first processing element: when the input queue of the first processing element is not full and the configuration value in the configuration register of the in-network storage element is the first value; and when the queue of the in-network storage element is not full and the configuration value in the configuration register of the in-network storage element is the second value. When the configuration value in the configuration register of the in-network storage element is the first value, the input controller of the second processing element may store the value within the input queue of the second processing element in response to receipt of the ready value; and when the configuration value in the configuration register of the in-network storage element is the second value, the controller of the in-network storage element may store the value within the queue of the in-network storage element in response to receipt of the ready value. The in-network storage element may include a second switch, and the configuration value may include a network select field to switch between a first network having a first set of the first processing element and the second processing element when the network select field is a first value, and a second network having a second, different set of the first processing element and the second processing element when the network select field is a second value.

[0342] In another embodiment, an apparatus (e.g., CSA) includes a plurality of processing elements that each comprise: a configuration register within a respective processing element to store a configuration value that causes the respective processing element to perform an operation according to the configuration value, a plurality of input queues, an input controller to control enqueque and dequeue of values into the plurality of input queues according to the configuration value, a plurality of output queues, and an output controller to control enqueue and dequeue of values into the plurality of output queues according to the configuration value; a circuit switched interconnect network between the plurality of processing elements to transfer values between the plurality of processing elements; and an in-network storage element of the circuit switched interconnect network comprising: a queue coupled to an output queue of a plurality of output queues of a first processing element of the plurality of processing elements, a configuration register of the in-network storage element to store a configuration value, and means that switches the in-network storage element into a first mode that provides a value stored in the queue of the in-network storage element by the output queue of the first processing element to an input queue of a plurality of input queues of a second processing element of the plurality of processing elements when the configuration value is a first value, and into a second mode that bypasses the queue of the in-network storage element and provides a value from the output queue of the first processing element to the input queue of the second processing element when the configuration value is a second value.

[0343] In another embodiment, an apparatus comprises a data storage device that stores code that when executed by a hardware processor causes the hardware processor to perform any method disclosed herein. An apparatus may be as described in the detailed description. A method may be as described in the detailed description.

[0344] In yet another embodiment, a non-transitory machine readable medium that stores code that when executed by a machine causes the machine to perform a method comprising any method disclosed herein.

Multicast Example

[0345] Referring again to FIGS. 32A and 32B, the following may occur by setting the configuration value in configuration storage 3256 to cause the bypass path 3244 to be utilized instead of buffer 3242. A processing element (e.g., or in the network itself) may include a conditional queue (e.g., having only a single slot, of having multiple slots in each conditional queue) as discussed herein. In one embodiment, a single buffer (e.g., or queue) may include its own, respective conditional queue. In the depicted embodiment, optional conditional queue 3207 is included for control input buffer 3222, optional conditional queue 3209 is included
for first data input buffer 3224B, and optional conditional queue 3211 is included for second data input buffer 3226B, optional conditional queue 3213 is included for control input buffer 3222C, optional conditional queue 3215 is included for first data input buffer 3224C, and optional conditional queue 3217 is included for second data input buffer 3226C.

[0346] Register 3220A-C activity may be controlled by that operation (an output of multiplexer 3216A-C, e.g., controlled by the scheduler 3214A-C). Scheduler 3214A-C may schedule an operation or operations of processing element 3200A-C, respectively, for example, when a dataflow token arrives (e.g., input data and/or control input). Control input buffer 3222A, first data input buffer 3224A, and second data input buffer 3226A are connected to local network 3202 for first PE 3200A. In one embodiment, control output buffer 3232A is connected to network 3210 for first PE 3200A, control input buffer 3222B is connected to local network 3210 for second PE 3200B, and control input buffer 3222C is connected to local network 3210 for third PE 3200C (e.g., and each local network may include a data path as in FIG. 7A and a flow control path as in FIG. 7B) and is loaded with a value when it arrives (e.g., the network has a data bit(s) and valid bit(s)). In one embodiment, first data output buffer 3234A is connected to network 3210 for first PE 3200A, first data input buffer 3224B is connected to local network 3210 for second PE 3200B, and first data input buffer 3224C is connected to local network 3210 for third PE 3200C (e.g., and each local network may include a data path as in FIG. 7A and a flow control path as in FIG. 7B) and is loaded with a value when it arrives (e.g., the network has a data bit(s) and valid bit(s)). In one embodiment, second data output buffer 3236A is connected to network 3210 for first PE 3200A, second data input buffer 3226B is connected to local network 3210 for second PE 3200B, and second data input buffer 3226C is connected to local network 3210 for third PE 3200C (e.g., and each local network may include a data path as in FIG. 7A and a flow control path as in FIG. 7B) and is loaded with a value when it arrives (e.g., the network has a data bit(s) and valid bit(s)).

Control output buffer 3232A-C, data output buffer 3234A-C, and/or data output buffer 3236A-C may receive an output of processing element 3200A-C (respectively), e.g., as controlled by the operation (an output of multiplexer 3216A-C). Status register 3238A-C may be loaded whenever the ALU 3218A-C executes (e.g., also controlled by output of multiplexer 3216A-C). Data in control input buffer 3222A-C and control output buffer 3232A-C may be a single bit. Multiplexer 3221A-C (e.g., operand A) and multiplexer 3223A-C (e.g., operand B) may source inputs.

[0347] For example, suppose the operation of first processing (e.g., compute) element 3200A is (or includes) what is called call a switch in FIG. 3B. The processing element 3200A may output data to data output buffer 3234A or data output buffer 3236A, e.g., from data input buffer 3224A (e.g., default) or data input buffer 3226A. The control bit in 3222A may thus indicate a 0 if outputting to data output buffer 3234A or a 1 if outputting to data output buffer 3236A. The output data may be the result of an operation by the ALU in certain embodiments. In one embodiment, a conditional value is sent by a different PE (e.g., none of PEs 3200A, 3200B, or 3200C). In one embodiment, the conditional value is sent from an output buffer of fourth PE 3200D (e.g., which may include the circuitry as in any PE discussed herein), for example, on a circuit switched path formed in a circuit switched embodiment of network 3210. Additional PEs may be coupled to network 3200.

[0348] However, in certain embodiments herein, the network 3210 and one or more of the conditional queues (conditional queue 3207, conditional queue 3209, conditional queue 3211, conditional queue 3213, conditional queue 3215, and/or conditional queue 3217) may be utilized to perform the switch operation, e.g., saving PE from being consumed merely for the switch operation. For example, in a multicast mode, a conditional value received from a PE may be used to cause the multicast (dataflow) token to be used or discarded by a consumer PE or multiple consumer PEs. Other conditional queues may be included (e.g., conditional output queues 3229, 3231, 3233).

[0349] Multiple networks (e.g., interconnects) may be connected to a processing element, e.g., networks 3202, 3204, 3206, and 3210. The connections may be switches, e.g., as discussed herein. In one embodiment, PEs and a circuit switched network 3210 are configured (e.g., control settings are selected) such that circuit switched network 3210 includes (i) a data path to send data from first PE 3200A to both second PE 3200B and third PE 3200C, e.g., for operations to be performed on that data by second PE 3200B and third PE 3200C, and (ii) a flow control path to send control data that controls (or is used to control) the sending of that data from first PE 3200A to both second PE 3200B and third PE 3200C. First PE 3200A includes a scheduler 3214A. A scheduler or other PE and/or network circuitry may include control circuitry to control a multicast operation. A scheduler or other PE and/or network circuitry may include control circuitry to control the in-network operations discussed herein. Flow control data may include a backpressure value, a speculation value, and/or a success value.

[0350] First (e.g., as producer) PE 3200A includes a (e.g., input) port 3208A(1-6) coupled to network 3210, e.g., to receive a backpressure value from second (e.g., as consumer) PE 3200B and/or third (e.g., as consumer) PE 3200C. In one circuit switched configuration, (e.g., input) port 3208A(1-6) (e.g., having a plurality of parallel inputs (1), (2), (3), (4), (5), and (6)) is to receive a respective backpressure value from each one of control input buffer 3222B, first data input buffer 3224B, and second data input buffer 3226B and/or control input buffer 3222C, first data input buffer 3224C, and second data input buffer 3226C. In one embodiment, (e.g., input) port 3208A(1-6) is to receive an aggregated (e.g., single) respective backpressure value of each of (i) a backpressure value from control input buffer 3222B logically ANDed (e.g., it returns the Boolean value true (e.g., binary high, e.g., binary 1) if both input operands are true and returns false (e.g., binary 0) otherwise) with a backpressure value from control input buffer 3222C (e.g., on input 3208A(1)), (ii) a backpressure value from first data input buffer 3224B logically ANDed with a backpressure value from first data input buffer 3224C (e.g., on input 3208A(2)), and (iii) a backpressure value from second data input buffer 3226B logically ANDed with a backpressure value from second data input buffer 3226C (e.g., on input 3208A(3)). In one embodiment, an input or output marked as a (1), (2), or (3) is its own respective wire or other coupling.

[0351] In one circuit switched configuration, (e.g., input) port 3208A(1-6) (e.g., having a plurality of parallel inputs (1), (2), (3), (4), (5), and (6)) is to receive a respective
backpressure value from any of control input buffer 3222B, first data input buffer 3224B, and second data input buffer 3226B and/or control input buffer 3222C, first data input buffer 3224C, and second data input buffer 3226C. In one embodiment, a circuit switched backpressure path (e.g., channel) is formed by setting switches coupled to wires between an input (e.g., input 1, 2, or 3) of port 3208A and an output (e.g., output 1, 2, or 3) of port 3208B to send a backpressure token (e.g., a value indicating no storage is available in an input buffer/queue) for one of control input buffer 3222B, first data input buffer 3224B, or second data input buffer 3226B of second PE 3200B. Additionally or alternatively, a (e.g., different) circuit switched backpressure path (e.g., channel) is formed by setting switches coupled to wires between an input (e.g., a different input of input 1, 2, or 3 or one of more than 3 inputs in another embodiment) of port 3208A and an output (e.g., output 1, 2, or 3) of port 3208C to send a backpressure token (e.g., a value indicating no storage is available in an input buffer/queue) for one of control input buffer 3222C, first data input buffer 3224C, or second data input buffer 3226C of third PE 3200C.

[0352] In one circuit switched configuration, a multicast data path is formed from (i) control output buffer 3232A to control input buffer 3222B and control input buffer 3222C, (ii) first data output buffer 3234A to first data input buffer 3224B and first data input buffer 3224C, (iii) second data output buffer 3236A to second data input buffer 3226B and second data input buffer 3226C, or any combination thereof. A data path may be used to send a data token from the producer PE to the consumer PEs.

[0353] In one embodiment, second PE 3200B includes any one of (e.g., any combination of) conditional queue 3207 for control input buffer 3222B, conditional queue 3209 for first data input buffer 3224B, and conditional queue 3211 for second data input buffer 3226B. In one circuit switched configuration, (e.g., output) port 3208B(1-3) is to send a respective backpressure value for each of one control input buffer 3222B (e.g., on output 3208B(1), first data input buffer 3224B (e.g., on output 3208B(2)), and second data input buffer 3226B (e.g., on output 3208B(3)), e.g., by scheduler 3214B.

[0354] In one embodiment, third PE 3200C includes any one of (e.g., any combination of) conditional queue 3213 for control input buffer 3222C, conditional queue 3215 for first data input buffer 3224C, and conditional queue 3217 for second data input buffer 3226C. In one circuit switched configuration, (e.g., output) port 3208C(1-3) is to send a respective backpressure value for each of one control input buffer 3222C (e.g., on output 3208C(1)), first data input buffer 3224C (e.g., on output 3208C(2)), and second data input buffer 3226C (e.g., on output 3208C(3)), e.g., by scheduler 3214C. A port may include a plurality of inputs and/or outputs. A processing element may include a single port into network 3210, or any plurality of ports.

[0355] First (e.g., as consumer) PE 3200A may include an (e.g., output) port 3225(1-3) coupled to network 3202, e.g., to send a backpressure value from first (e.g., as consumer) PE 3200A to an upstream (e.g., as producer) PE. In one circuit switched configuration, (e.g., output) port 3225(1-3) is to send a respective backpressure value for each of one control input buffer 3222A (e.g., on output 3225(1)), first data input buffer 3224A (e.g., on output 3225(2)), and second data input buffer 3226A (e.g., on output 3225(3)), e.g., by scheduler 3214A.

[0356] Any of input ports (e.g., input ports to conditional queue 3207, conditional queue 3209, conditional queue 3211, conditional queue 3213, conditional queue 3215, and/or conditional queue 3217) may include a backpressure path to the component (e.g., output port thereof) that is to send data to the input port.

[0357] Second (e.g., as producer) PE 3200B may include a (e.g., input) port 3235(1-6) coupled to network 3204 (e.g., which may be the same network as network 3206), e.g., to receive a backpressure value from a downstream (e.g., as consumer) PE or PEs. In one circuit switched configuration, (e.g., input) port 3235(1-6) (e.g., having a plurality of parallel inputs (1), (2), (3), (4), (5), and (6)) is to receive a respective backpressure value from each one of control input buffer, first data input buffer, and second data input buffer of a first downstream PE and/or control input buffer, first data input buffer, and second data input buffer of a second downstream PE. In one embodiment, (e.g., input) port 3235(1-6) is to receive an aggregated (e.g., single) respective backpressure value of each of (i) a backpressure value from control input buffer for first downstream PE logically ANDd (e.g., it returns the Boolean value true (e.g., binary high, e.g., binary 1) if both input operands are true and returns false (e.g., binary 0) otherwise) with a backpressure value from control input buffer for second downstream PE (e.g., on input 3235(1)), (ii) a backpressure value from first data input buffer for first downstream PE logically ANDd with a backpressure value from first data input buffer for first downstream PE (e.g., on input 3235(2)), and (iii) a backpressure value from second data input buffer for first downstream PE logically ANDd with a backpressure value from second data input buffer for first downstream PE (e.g., on input 3235(3)). In one embodiment, an input or output marked as a (1), (2), or (3) is its own respective wire or other coupling. In one embodiment, each PE includes the same circuitry and/or components. In certain embodiments, a (e.g., output) port is to receive a backpressure value that is determined (e.g., not just ANDed unconditionally) according to a configurable flow control path network (e.g., and flow control functions thereof), e.g., a backpressure value on the configurable flow control path network in FIG. 7B from an output port(s) to an input port(s)).

[0358] Third (e.g., as producer) PE 3200C may include a (e.g., input) port 3245(1-6) coupled to network 3206 (e.g., which may be the same network as network 3204), e.g., to receive a backpressure value from a downstream (e.g., as consumer) PE or PEs. In one circuit switched configuration, (e.g., input) port 3245(1-6) (e.g., having a plurality of parallel inputs (1), (2), (3), (4), (5), and (6)) is to receive a respective backpressure value from each one of control input buffer, first data input buffer, and second data input buffer of a first downstream PE and/or control input buffer, first data input buffer, and second data input buffer of a second downstream PE. In one embodiment, (e.g., input) port 3245(1-6) is to receive an aggregated (e.g., single) respective backpressure value of each of (i) a backpressure value from control input buffer for first downstream PE logically ANDd (e.g., it returns the Boolean value true (e.g., binary high, e.g., binary 1) if both input operands are true and returns false (e.g., binary 0) otherwise) with a backpressure value from control input buffer for second downstream PE (e.g., on input 3245(1)), (ii) a backpressure value from first data input buffer for first downstream PE logically ANDd with a backpressure value from first data input buffer for first downstream PE (e.g., on input 3245(2)), and (iii) a backpressure value from second data input buffer for first downstream PE logically ANDd with a backpressure value from second data input buffer for first downstream PE (e.g., on input 3245(3)).
downstream PE (e.g., on input 3245(2)), and (iii) a backpressure value from second data input buffer for first downstream PE logically AND'd with a backpressure value from second data input buffer for first downstream PE (e.g., on input 3245(3)). In one embodiment, an input or output marked as a (1), (2), or (3) is its own respective wire or other coupling. In one embodiment, each PE includes the same circuitry and/or components.

A processing element may include two sub-networks (or two channels on the network), e.g., one for a data path and one for a flow control path. A processing element (e.g., PE 3200A, PE 3200B, and PE 3200C) may function and/or include the components as in any of the disclosure herein. A processing element may be stalled from execution until its operands (e.g., in its input buffer(s)) are received and/or until there is room in the output buffer(s) of the processing element for the data that is to be produced by the execution of the operation on those operands. A conditional queue may be added to manipulate the backpressure value(s) and/or queueing/dequeueing of data (e.g., in or out of a buffer or queue).

In one embodiment, a data token is received in control output buffer 3232A which causes the multicast critical path of the first example to begin operation. In one embodiment, the data token’s reception therein causes the producer PE 3200A (e.g., transmitter) to drive its dataflow (e.g., valid) value (e.g., on the path from control output buffer 3232A to control input buffer 3222B) (e.g., through network 3210) and the path from control output buffer 3232A to control input buffer 3222C (e.g., through network 3210) to a value (e.g., binary high) to indicate that it has data to be transmitted. In one embodiment, the dataflow value (e.g., valid) is the transmittal of the dataflow token (e.g., payload data) itself. In one embodiment, a first path is included from producer PE to (e.g., each) consumer PE through network 3210 for the dataflow token and a second path is included from producer PE to (e.g., each) consumer PE through network 3210 for a dataflow value to indicate if that dataflow token (e.g., in storage coupled to the first path) is valid or invalid.

In the first transmission attempt for this dataflow token, if the backpressure value (e.g., ready value) on the path from port 3200B(1) of second PE 3200B to port 3200A(1) of first PE 3200A and the backpressure value (e.g., ready value) on the path from port 3200C(1) of third PE 3200C to port 3208A(1) of first PE 3200A both indicate (e.g., as the output from AND logic gate at 3252) there is no backpressure (e.g., there is storage available in each of control input buffer 3222B and control input buffer 3222C), then the first PE (e.g., scheduler 3214A) determines that this transmission attempt will be successful, for example, and the dataflow token is to bedequeued (e.g., in the next cycle) from the control output buffer 3232A of the first PE 3200A.

An (e.g., input) port 3208A(1, 2, or 3) may receive an aggregated (e.g., single) respective backpressure value of a backpressure value from port 3208B(1, 2, or 3) of second PE 3200B for one of the input buffers (1122B, 3224B, and/or 3226B) of second PE 3200B or a backpressure value from port 3208C(1, 2, or 3) of third PE 3200C for one of the input buffers (1122C, 3224C, and/or 3226C) of third PE 3200C by AND logic gate 3260.

In the depicted embodiment, (e.g., input) port 3208A(1, 2, or 3) may receive an aggregated (e.g., single) respective backpressure value of a backpressure value from port 3208B(3) of second PE 3200B for second data input buffer 3226B of second PE 3200B logically AND'd (e.g., it returns the Boolean value true (e.g., binary high, e.g., binary 1) if both input operands are true and returns false (e.g., binary low, e.g., binary 0) otherwise) with a backpressure value from port 3208C(2) of third PE 3200C for first data input buffer 3224C of third PE 3200C by AND logic gate 3260. Conditional queue 3211 of second data input buffer 3226B of second PE 3200B may be coupled to backpressure path and/or second data input buffer 3226B, e.g., to control the backpressure value and/or queueing/dequeueing of data (e.g., in or out of buffer 3226B). Conditional queue 3215 of first data input buffer 3224C of third PE 3200C may be coupled to backpressure path and/or first data input buffer 3224C, e.g., to control the backpressure value and/or queueing/dequeueing of data (e.g., in or out of buffer 3224C). In one embodiment, a separate conditional queue is used for each buffer. In one embodiment, a plurality of queues are shared among the ports of a PE.

In the depicted embodiment, first data output buffer 3234A of first PE 3200A is coupled via data path in network 3210 to both of the second data input buffer 3226B of second PE 3200B and first data input buffer 3224C of third PE 3200C, e.g., via multiplexer 3257 sending two simultaneous outputs for a single input. Datapath and/or backpressure path may include multiple switches (e.g., multiplexers) and/or logic gates (e.g., as discussed below) illustrated as boxes, e.g., boxes 3254, 3255, 3258. In one embodiment, an output buffer of first PE 3200A is to assert a dataflow token on its dataflow path when the output buffer (e.g., data output buffer 3234A) providing that dataflow token receives it into a slot thereof. The dataflow token may remain asserted until the receiving PE (all receiving PEs) asserts their backpressure value on their backpressure path (or paths) to indicate if storage is available in their input buffer for the dataflow token.

In certain embodiments, a dataflow token that is multicast may also be controlled by a conditional value (e.g., at each of the consumer PEs) to cause a PE that is configured to perform operations on that dataflow token, to either (i) release the dataflow token for use (e.g., as an input to processing) by a consumer PE or (ii) not accept or discard the dataflow token. In certain embodiments, each endpoint (e.g., consumer PE) is to have its own conditional value such that each endpoint decides to either (i) use or (ii) not use the dataflow token based on the conditional value (e.g., as an independent choice from the other consumer PEs that are part of the multicast operation). Certain embodiments herein insert a conditional (e.g., Boolean) queue associated with an (e.g., each) input of the consumer (e.g., receiver) PEs that is to receive a dataflow token (e.g., value), and the condition queue is to receive (e.g., store) a conditional token (e.g., value) to control the (i) use or (ii) not use of that dataflow token.

In certain embodiments of multicast operations and in-network operations using a conditional queue, a consumer (e.g., receiver) PEs may signal the transmitter (e.g., on a backpressure path) according to the multicast discussions herein.
2.4 Memory Interface

[0367] The request address file (RAF) circuit, a simplified version of which is shown in FIG. 46, may be responsible for executing memory operations and serves as an intermediary between the CSA fabric and the memory hierarchy. As such, the main microarchitectural task of the RAF may be to rationalize the cut-of-order memory subsystem with the in-order semantics of CSA fabric. In this capacity, the RAF circuit may be provisioned with completion buffers, e.g., queue-like structures that reorder memory responses and return them to the fabric in the request order. The second major functionality of the RAF circuit may be to provide support in the form of address translation and a page walker. Incoming virtual addresses may be translated to physical addresses using a channel-associative translation lookaside buffer (TLB). To provide ample memory bandwidth, each CSA tile may include multiple RAF circuits. Like the various PEs of the fabric, the RAF circuits may operate in a dataflow style by checking for the availability of input arguments and output buffering, if required, before selecting a memory operation to execute. Unlike some PEs, however, the RAF circuit is multiplexed among several colocated memory operations. A multiplexed RAF circuit may be used to minimize the area overhead of its various subcomponents, e.g., to share the Accelerator Cache Interface (ACI) network (described in more detail in Section 2.3), shared virtual memory (SVM) support hardware, mezzanine network interface, and other hardware management facilities. However, there are some program characteristics that may also motivate this choice. In one embodiment, a (e.g., valid) dataflow graph is to poll memory in a shared virtual memory system. Memory-latency-bound programs, like graph traversals, may utilize many separate memory operations to saturate memory bandwidth due to memory-dependent control flow. Although each RAF may be multiplexed, a CSA may include multiple (e.g., between 8 and 32) RAFs at a tile granularity to ensure adequate cache bandwidth. RAFs may communicate with the rest of the fabric via both the local network and the mezzanine network. Where RAFs are multiplexed, each RAF may be provisioned with several ports into the local network. These ports may serve as a minimum-latency, highly-deterministic path to memory for use by latency-sensitive or high-bandwidth memory operations. In addition, a RAF may be provisioned with a mezzanine network endpoint, e.g., which provides memory access to runtime services and distant user-level memory accessors.

[0368] FIG. 46 illustrates a request address file (RAF) circuit 4600 according to embodiments of the disclosure. In one embodiment, at configuration time, the memory load and store operations that were in a dataflow graph are specified in registers 4610. The arcs to those memory operations in the dataflow graphs may then be connected to the input queues 4622, 4624, and 4626. The arcs from those memory operations are thus to leave completion buffers 4628, 4630, or 4632. Dependency tokens (which may be single bits) arrive into queues 4618 and 4620. Dependency tokens are to leave from queue 4616. Dependency token counter 4614 may be a compact representation of a queue and track a number of dependency tokens used for any given input queue. If the dependency token counters 4614 saturate, no additional dependency tokens may be generated for new memory operations. Accordingly, a memory ordering circuit (e.g., a RAF in FIG. 47) may stall scheduling new memory operations until the dependency token counters 4614 becomes unsaturated.

[0369] As an example for a load, an address arrives into queue 4622 which the scheduler 4612 matches up with a load in 4610. A completion buffer slot for this load is assigned in the order the address arrived. Assuming this particular load in the graph has no dependencies specified, the address and completion buffer slot are sent off to the memory system by the scheduler (e.g., via memory command 4642). When the result returns to multiplexer 4640 (shown schematically), it is stored into the completion buffer slot it specifies (e.g., as it carried the target slot all along though the memory system). The completion buffer sends results back into local network (e.g., local network 4602, 4604, 4606, or 4608) in the order the addresses arrived.

[0370] Stores may be similar except both address and data have to arrive before any operation is sent off to the memory system.

2.5 Cache

[0371] Dataflow graphs may be capable of generating a profusion of (e.g., word granularity) requests in parallel. Thus, certain embodiments of the CSA provide a cache subsystem with sufficient bandwidth to service the CSA. A heavily banked cache microarchitecture, e.g., as shown in FIG. 47 may be utilized. FIG. 47 illustrates a circuit 4700 with a plurality of request address file (RAF) circuits (e.g., RAF circuit (1)) coupled between a plurality of accelerator tiles (4708, 4710, 4712, 4714) and a plurality of cache banks (e.g., cache bank 4702) according to embodiments of the disclosure. In one embodiment, the number of RAFs and cache banks may be in a ratio of either 1:1 or 1:2. Cache banks may contain full cache lines (e.g., as opposed to sharding by word), with each line having exactly one home in the cache. Cache lines may be mapped to cache banks via a pseudo-random function. The CSA may adopt the shared virtual memory (SVM) model to integrate with other tiled architectures. Certain embodiments include an Accelerator Cache Interface (ACI) network connecting the RAFs to the cache banks. This network may carry address and data between the RAFs and the cache. The topology of the ACI may be a cascaded crossbar, e.g., as a compromise between latency and implementation complexity.

2.6 Network Resources, e.g., Circuitry, to Perform (e.g., Dataflow) Operations

[0372] In certain embodiments, processing elements (PEs) communicate using dedicated virtual circuits which are formed by statically configuring a (e.g., circuit switched) communications network. These virtual circuits may be flow controlled and fully back-pressure, e.g., such that a PE will stall if either the source has no data or its destination is full. At runtime, data may flow through the PEs implementing the mapped dataflow graph (e.g., mapped algorithm). For example, data may be streamed in from memory, through the (e.g., fabric area of a) spatial array of processing elements, and then back out to memory.

[0373] Such an architecture may achieve remarkable performance efficiency relative to traditional multicore processors: compute, e.g., in the form of PEs, may be simpler and more numerous than cores and communications may be direct, e.g., as opposed to an extension of the memory
system. However, the (e.g., fabric area of) spatial array of processing elements may be tuned for the implementation of compiler-generated expression trees, which may feature little multiplexing or demultiplexing. Certain embodiments herein extend (for example, via network resources, such as, but not limited to, network dataflow endpoint circuits) the architecture to support (e.g., high-radix) multiplexing and/or demultiplexing, for example, especially in the context of function calls.

Spatial arrays, such as the spatial array of processing elements 101 in FIG. 1, may use (e.g., packet switched) networks for communications. Certain embodiments herein provide circuitry to overlay high-radix dataflow operations on these networks for communications. For example, certain embodiments herein utilize the existing network for communications (e.g., interconnect network 104 described in reference to FIG. 1) to provide data routing capabilities between processing elements and other components of the spatial array, but also augment the network (e.g., network endpoints) to support the performance and/or control of some (e.g., less than all) of dataflow operations (e.g., without utilizing the processing elements to perform those dataflow operations). In one embodiment, (e.g., high radix) dataflow operations are supported with special hardware structures (e.g. network dataflow endpoint circuits) within a spatial array, for example, without consuming processing resources or degrading performance (e.g., of the processing elements).

In one embodiment, a circuit switched network between two points (e.g., between a producer and consumer of data) includes a dedicated communication line between those two points, for example, with (e.g., physical) switches between the two points set to create a (e.g., exclusive) physical circuit between the two points. In one embodiment, a circuit switched network between two points is set up at the beginning of use of the connection between the two points and maintained throughout the use of the connection. In another embodiment, a packet switched network includes a shared communication line (e.g., channel) between two (e.g., or more) points, for example, where packets from different connections share that communication line (for example, routed according to data of each packet, e.g., in the header of a packet including a header and a payload). An example of a packet switched network is discussed below, e.g., in reference to a mezzanine network.

FIG. 48 illustrates a data flow graph 4800 of a pseudocode function call 4801 according to embodiments of the disclosure. Function call 4801 is to load two input data operands (e.g., indicated by pointers α and β, respectively), and multiply them together, and return the resultant data. This or other functions may be performed multiple times (e.g., in a dataflow graph). The dataflow graph in FIG. 48 illustrates a PacketAny dataflow operator 4802 to perform the operation of selecting a control data (e.g., an index) (for example, from call sites 4802A) and copying with copy dataflow operator 4804 that control data (e.g., index) to each of the first Pick dataflow operator 4806, second Pick dataflow operator 4806, and Switch dataflow operator 4816. In one embodiment, an index (e.g., from the PacketAny thus inputs and outputs data to the same index position, e.g., of [0, 1, ..., M], where M is an integer. First Pick dataflow operator 4806 may then pull one input data element of a plurality of input data elements 4806A according to the control data, and use the one input data element as (*a) to then load the input data value stored at *a with load dataflow operator 4810. Second Pick dataflow operator 4808 may then pull one input data element of a plurality of input data elements 4808A according to the control data, and use the one input data element as (*b) to then load the input data value stored at *b with load dataflow operator 4812. Those two input data values may then be multiplied by multiplication dataflow operator 4814 (e.g., as a part of a processing element). The resultant data of the multiplication may then be routed (e.g., to a downstream processing element or component) by Switch dataflow operator 4816, e.g., to call sites 4816A, for example, according to the control data (e.g., index) to Switch dataflow operator 4816.

FIG. 48 is an example of a function call where the number of dataflow operators used to manage the steering of data (e.g., tokens) may be significant, for example, to steer the data to and/or from call sites. In one example, one or more of PickAny dataflow operator 4802, first Pick dataflow operator 4806, second Pick dataflow operator 4808, and Switch dataflow operator 4816 may be utilized to route (e.g., steer) data, for example, when there are multiple (e.g., many) call sites. In an embodiment where a (e.g., main) goal of introducing a multiplexed and/or demultiplexed function call is to reduce the implementation area of a particular dataflow graph, certain embodiments herein (e.g., of micro-architecture) reduce the area overhead of such multiplexed and/or demultiplexed (e.g., portions of) dataflow graphs.

FIG. 49 illustrates a spatial array 4901 of processing elements (PEs) with a plurality of network dataflow endpoint circuits (4902, 4904, 4906) according to embodiments of the disclosure. Spatial array 4901 of processing elements may include a communications (e.g., interconnect) network in between components, for example, as discussed herein. In one embodiment, communications network is one or more (e.g., channels of a) packet switched communications network. In one embodiment, communications network is one or more circuit switched, statically configured communications channels. For example, a set of channels coupled together by a switch (e.g., switch 4910 in a first network and switch 4911 in a second network). The first network and second network may be separate or coupled together. For example, switch 4910 may couple one or more of a plurality (e.g., four) data paths therein together, e.g., as configured to perform an operation according to a dataflow graph. In one embodiment, the number of data paths is any plurality. Processing element (e.g., processing element 4908) may be as disclosed herein, for example, as in FIG. 9. Accelerator tile 4900 includes a memory/cache hierarchy interface 4912, e.g., to interface the accelerator tile 4900 with a memory and/or cache. A data path may extend to another tile or terminate, e.g., at the edge of a tile. A processing element may include an input buffer (e.g., buffer 4909) and an output buffer.

Operations may be executed based on the availability of their inputs and the status of the PE. A PE may obtain operands from input channels and write results to output channels, although internal register state may also be used. Certain embodiments herein include a configurable dataflow-friendly PE. FIG. 9 shows a detailed block diagram of one such PE: the integer PE. This PE consists of several I/O buffers, an ALU, a storage register, some instruction registers, and a scheduler. Each cycle, the scheduler may select an instruction for execution based on the availability of the input and output buffers and the status of the PE. The
result of the operation may then be written to either an output buffer or to a (e.g., local to the PE) register. Data written to an output buffer may be transported to a downstream PE for further processing. This style of PE may be extremely energy efficient, for example, rather than reading data from a complex, multi-ported register file, a PE reads the data from a register. Similarly, instructions may be stored directly in a register, rather than in a virtualized instruction cache.

[0380] Instruction registers may be set during a special configuration step. During this step, auxiliary control wires and state, in addition to the inter-PE network, may be used to stream in configuration across the several PEs comprising the fabric. As result of parallelism, certain embodiments of such a network may provide for rapid reconfiguration, e.g., a tile sized fabric may be configured in less than about 10 microseconds.

[0381] Further, depicted accelerator tile 4900 includes packet switched communications network 4914, for example, as part of a mezzanine network, e.g., as described below. Certain embodiments herein allow for (e.g., a distributed) dataflow operations (e.g., operations that only route data) to be performed on (e.g., within) the communications network (e.g., and not in the processing element(s)). As an example, a distributed Pick dataflow operation of a dataflow graph is depicted in FIG. 49. Particularly, distributed pick is implemented using three separate configurations on three separate networks (e.g., global) networks (e.g., network dataflow end point circuits (4902, 4904, 4906)). Dataflow operations may be distributed, e.g., with several endpoints to be configured in a coordinated manner. For example, a compilation tool may understand the need for coordination. Endpoints (e.g., network dataflow end point circuits) may be shared among several distributed operations, for example, a dataflow operation (e.g., Pick) endpoint may be collated with several sends related to the dataflow operation (e.g., Pick). A distributed dataflow operation (e.g., Pick) may generate the same result the same as a non-distributed dataflow operation (e.g., Pick). In certain embodiments, a difference between distributed and non-distributed dataflow operations is that in the distributed dataflow operations, they data (e.g., data to be routed, but which may not include control data) over a packet switched communications network, e.g., with associated flow control and distributed coordination. Although different sized processing elements (PE) are shown, in one embodiment, each processing element is of the same size (e.g., silicon area). In one embodiment, a buffer element to buffer data may also be included, e.g., separate from a processing element.

[0382] As one example, a pick dataflow operation may have a plurality of inputs and steer (e.g., route) one of them as an output, e.g., as in FIG. 48. Instead of utilizing a processing element to perform the pick dataflow operation, it may be achieved with one or more of network communication resources (e.g., network dataflow end point circuits). Additionally or alternatively, the network dataflow end point circuits may route data between processing elements, e.g., for the processing elements to perform processing operations on the data. embodiments herein may thus utilize to the communications network to perform (e.g., steering) dataflow operations. Additionally or alternatively, the network dataflow end point circuits may perform as a mezzanine network discussed below.

[0383] In the depicted embodiment, packet switched communications network 4914 may handle certain (e.g., configuration) communications, for example, to program the processing elements and/or circuit switched network (e.g., network 4913, which may include switches). In one embodiment, a circuit switched network is configured (e.g., programmed) to perform one or more operations (e.g., dataflow operations of a dataflow graph).

[0384] Packet switched communications network 4914 includes a plurality of endpoints (e.g., network dataflow end point circuits (4902, 4904, 4906)). In one embodiment, each endpoint includes an address or other indicator value to allow data to be routed to and/or from that endpoint, e.g., according to (e.g., a header of) a data packet.

[0385] Additionally or alternatively to performing one or more of the above, packet switched communications network 4914 may perform dataflow operations. Network dataflow end point circuits (4902, 4904, 4906) may be configured (e.g., programmed) to perform a (e.g., distributed pick) operation of a dataflow graph. Programming of components (e.g., a circuit) are described herein. An embodiment of configuring a network dataflow end point circuit (e.g., an operation configuration register thereof) is discussed in reference to FIG. 50.

[0386] As an example of a distributed pick dataflow operation, network dataflow end point circuits (4902, 4904, 4906) in FIG. 49 may be configured (e.g., programmed) to perform a distributed pick operation of a dataflow graph. An embodiment of configuring a network dataflow end point circuit (e.g., an operation configuration register thereof) is discussed in reference to FIG. 50. Additionally or alternatively to configuring remote endpoint circuits, local endpoint circuits may also be configured according to this disclosure.

[0387] Network dataflow end point circuit 4902 may be configured to receive input data from a plurality of sources (e.g., network dataflow end point circuit 4904 and network dataflow end point circuit 4906), and to output resultant data, e.g., as in FIG. 48, for example, according to control data. Network dataflow end point circuit 4904 may be configured to provide (e.g., send) input data to network dataflow end point circuit 4902, e.g., on receipt of the input data from processing element 4922. This may be referred to as Input 0 in FIG. 49. In one embodiment, circuit switched network is configured (e.g., programmed) to provide a dedicated communication line between processing element 4922 and network dataflow end point circuit 4904 along path 4924. Network dataflow end point circuit 4906 may be configured to provide (e.g., send) input data to network dataflow end point circuit 4902, e.g., on receipt of the input data from processing element 4920. This may be referred to as Input 1 in FIG. 49. In one embodiment, circuit switched network is configured (e.g., programmed) to provide a dedicated communication line between processing element 4920 and network dataflow end point circuit 4906 along path 4916.

[0388] When network dataflow end point circuit 4904 is to transmit input data to network dataflow end point circuit 4902 (e.g., when network dataflow end point circuit 4902 has available storage room for the data and/or network dataflow end point circuit 4904 has its input data), network dataflow end point circuit 4904 may generate a packet (e.g., including the input data and a header to steer that data to network dataflow end point circuit 4902 on the packet switched communications network 4914 (e.g., as a stop on that (e.g., ring) network 4914). This is illustrated schematically with
dashed line 4926 in FIG. 49. Although the example shown in FIG. 49 utilizes two sources (e.g., two inputs) a single or any plurality (e.g., greater than two) of sources (e.g., inputs) may be utilized.

[0389] When network dataflow endpoint circuit 4906 is to transmit input data to network dataflow endpoint circuit 4902 (e.g., when network dataflow endpoint circuit 4902 has available storage room for the data and/or network dataflow endpoint circuit 4906 has its input data), network dataflow endpoint circuit 4904 may generate a packet (e.g., including the input data and a header to steer that data to network dataflow endpoint circuit 4902 on the packet switched communications network 4914 (e.g., as a stop on that (e.g., ring) network 4914). This is illustrated schematically with dashed line 4918 in FIG. 49. Though a mesh network is shown, other network topologies may be used.

[0390] Network dataflow endpoint circuit 4902 (e.g., on receipt of the Input 0 from network dataflow endpoint circuit 4904, Input 1 from network dataflow endpoint circuit 4906, and/or control data) may then perform the programmed dataflow operation (e.g., a Pick operation in this example). The network dataflow endpoint circuit 4902 may then output the according resultant data from the operation, e.g., to processing element 4908 in FIG. 49. In one embodiment, circuit switched network is configured (e.g., programmed) to provide a dedicated communication line between processing element 4908 (e.g., a buffer thereof) and network dataflow endpoint circuit 4902 along path 4928. A further example of a distributed Pick operation is discussed below in reference to FIG. 62-633.

[0391] In one embodiment, the control data to perform an operation (e.g., pick operation) comes from other components of the spatial array, e.g., a processing element or through network. An example of this is discussed below in reference to FIG. 50. Note that Pick operation is shown schematically in endpoint 4902, and may not be a multiplexer circuit, for example, see the discussion below of network dataflow endpoint circuit 5000 in FIG. 50.

[0392] In certain embodiments, a dataflow graph may have certain operations performed by a processing element and certain operations performed by a communication network (e.g., network dataflow endpoint circuit or circuits).

[0393] FIG. 50 illustrates a network dataflow endpoint circuit 5000 according to embodiments of the disclosure. Although multiple components are illustrated in network dataflow endpoint circuit 5000, one or more instances of each component may be utilized in a single network dataflow endpoint circuit. An embodiment of a network dataflow endpoint circuit may include any (e.g., not all) of the components in FIG. 50.

[0394] FIG. 50 depicts the microarchitecture of a (e.g., mezzanine) network interface showing embodiments of main data (solid line) and control data (dotted line) paths. This microarchitecture provides a configuration storage and scheduler to enable (e.g., high-radix) dataflow operators. Certain embodiments herein include data paths to the scheduler to enable leg selection and description. FIG. 50 shows a high-level microarchitecture of a network (e.g., mezzanine) endpoint (e.g., stop), which may be a member of a ring network for context. To support (e.g., high-radix) dataflow operations, the configuration of the endpoint (e.g., operation configuration storage 5026) to include configurations that examine multiple network (e.g., virtual) channels (e.g., as opposed to single virtual channels in a baseline implementation). Certain embodiments of network dataflow endpoint circuit 5000 include data paths from ingress and to egress to control the selection of (e.g., pick and switch types of operations), and/or to describe the choice made by the scheduler in the case of PickAny dataflow operators or SwitchAny dataflow operators. Flow control and backpressure behavior may be utilized in each communication channel, e.g., in a (e.g., packet switched communications) network and (e.g., circuit switched) network (e.g., fabric of a spatial array of processing elements).

[0395] As one description of an embodiment of the microarchitecture, a pick dataflow operator may function to pick one output of resultant data from a plurality of inputs of input data, e.g., based on control data. A network dataflow endpoint circuit 5000 may be configured to consider one of the spatial array ingress buffer(s) 5002 of the circuit 5000 (e.g., data from the fabric being control data) as selecting among multiple input data elements stored in network ingress buffer(s) 5024 of the circuit 5000 to steer the resultant data to the spatial array egress buffer 5008 of the circuit 5000. Thus, the network ingress buffer(s) 5024 may be thought of as inputs to a virtual mix, the spatial array ingress buffer 5002 as the multiplexer select, and the spatial array egress buffer 5008 as the multiplexer output. In one embodiment, when a (e.g., control data) value is detected and/or arrives in the spatial array ingress buffer 5002, the scheduler 5028 (e.g., as programmed by an operation configuration in storage 5026) is sensitized to examine the corresponding network ingress channel. When data is available in that channel, it is removed from the network ingress buffer 5024 and moved to the spatial array egress buffer 5008. The control bits of both ingresses and egress may then be updated to reflect the transfer of data. This may result in control flow tokens or credits being propagated in the associated network. In certain embodiment, all inputs (e.g., control or data) may arise locally or over the network.

[0396] Initially, it may seem that the use of packet switched networks to implement the (e.g., high-radix staging) operators of multiplexed and/or demultiplexed codes harms performance. For example, in one embodiment, a packet-switched network is generally shared and the caller and callee dataflow graphs may be distant from one another. Recall, however, that in certain embodiments, the intention of supporting multiplexing and/or demultiplexing is to reduce the area consumed by infrequent code paths within a dataflow operator (e.g., by the spatial array). Thus, certain embodiments herein reduce area and avoid the consumption of more expensive fabric resources, for example, like PEs, e.g., without (substantially) affecting the area and efficiency of individual PEs to supporting those (e.g., infrequent) operations.

[0397] Turning now to further detail of FIG. 50, depicted network dataflow endpoint circuit 5000 includes a spatial array (e.g., fabric) ingress buffer 5002, for example, to input data (e.g., control data) from a (e.g., circuit switched) network. As noted above, although a single spatial array (e.g., fabric) ingress buffer 5002 is depicted, a plurality of spatial array (e.g., fabric) ingress buffers may be in a network dataflow endpoint circuit. In one embodiment, spatial array (e.g., fabric) ingress buffer 5002 is to receive data (e.g., control data) from a communications network of a spatial array (e.g., a spatial array of processing elements),
for example, from one or more of network 5004 and network 5006. In one embodiment, network 5004 is part of network 4913 in FIG. 49.

[0398] Depicted network dataflow endpoint circuit 5000 includes a spatial array (e.g., fabric) egress buffer 5008, for example, to output data (e.g., control data) to a (e.g., circuit switched) network. As noted above, although a single spatial array (e.g., fabric) egress buffer 5008 is depicted, a plurality of spatial arrays (e.g., fabric) egress buffers may be in a network dataflow endpoint circuit. In one embodiment, spatial array (e.g., fabric) egress buffer 5008 is to send (e.g., transmit) data (e.g., control data) onto a communications network of a spatial array (e.g., a spatial array of processing elements), for example, onto one or more of network 5010 and network 5012. In one embodiment, network 5010 is part of network 4913 in FIG. 49.

[0399] Additionally or alternatively, network dataflow endpoint circuit 5000 may be coupled to another network 5014, e.g., a packet switched network. Another network 5014, e.g., a packet switched network, may be used to transmit (e.g., send or receive) (e.g., input and/or resultant) data to processing elements or other components of a spatial array and/or to transmit one or more of input data or resultant data. In one embodiment, network 5014 is part of the packet switched communications network 4914 in FIG. 49, e.g., a time multiplexed network.

[0400] Network buffer 5018 (e.g., register(s)) may be a stop on (e.g., ring) network 5014, for example, to receive data from network 5014.

[0401] Depicted network dataflow endpoint circuit 5000 includes a network egress buffer 5022, for example, to output data (e.g., resultant data) to a (e.g., packet switched) network. As noted above, although a single network egress buffer 5022 is depicted, a plurality of network egress buffers may be in a network dataflow endpoint circuit. In one embodiment, network egress buffer 5022 is to send (e.g., transmit) data (e.g., resultant data) onto a communications network of a spatial array (e.g., a spatial array of processing elements), for example, onto network 5014. In one embodiment, network 5014 is part of packet switched network 4914 in FIG. 49. In certain embodiments, network egress buffer 5022 is to output data (e.g., from spatial array egress buffer 5002) to (e.g., packet switched) network 5014, for example, to be routed (e.g., steered) to other components (e.g., other network dataflow endpoint circuit(s)).

[0402] Depicted network dataflow endpoint circuit 5000 includes a network ingress buffer 5022, for example, to input data (e.g., input data) from a (e.g., packet switched) network. As noted above, although a single network ingress buffer 5024 is depicted, a plurality of network ingress buffers may be in a network dataflow endpoint circuit. In one embodiment, network ingress buffer 5024 is to receive (e.g., transmit) data (e.g., input data) from a communications network of a spatial array (e.g., a spatial array of processing elements), for example, from network 5014. In one embodiment, network 5014 is part of packet switched network 4914 in FIG. 49. In certain embodiments, network ingress buffer 5024 is to receive data (e.g., input data) from a communications network of a spatial array (e.g., a spatial array of processing elements), for example, from network 5014. In one embodiment, network 5014 is part of packet switched network 4914 in FIG. 49. In certain embodiments, network ingress buffer 5024 is to receive data (e.g., input data) from a communications network of a spatial array (e.g., a spatial array of processing elements), for example, from network 5014. In one embodiment, network 5014 is part of packet switched network 4914 in FIG. 49. In certain embodiments, network ingress buffer 5024 is to receive data (e.g., input data) from a communications network of a spatial array (e.g., a spatial array of processing elements), for example, from network 5014.

[0403] In one embodiment, the data format (e.g., of the data on network 5014) includes a packet having data and a header (e.g., with the destination of that data). In one embodiment, the data format (e.g., of the data on network 5004 and/or 5006) includes only the data (e.g., not a packet having data and a header (e.g., with the destination of that data)). Network dataflow endpoint circuit 5000 may add (e.g., data output from circuit 5000) or remove (e.g., data input into circuit 5000) a header (or other data) to or from a packet. Coupling 5020 (e.g., wire) may send data received from network 5014 (e.g., from network buffer 5018) to network ingress buffer 5002. Multiplexer 5016 may (e.g., via a control signal from the scheduler 5028) output data from network buffer 5018 or from network egress buffer 5022. In one embodiment, one or more of multiplexer 5016 or network buffer 5018 are separate components from network dataflow endpoint circuit 5000. A buffer may include a plurality of (e.g., discrete) entries, for example, a plurality of registers.

[0404] In one embodiment, operation configuration storage 5026 (e.g., register or registers) is loaded during configuration (e.g., mapping) and specifies the particular operation (or operations) this network dataflow endpoint circuit 5000 (e.g., not a processing element of a spatial array) is to perform (e.g., data steered operations in contrast to logic or arithmetic operations). Buffer(s) (e.g., 5002, 5008, 5022, and/or 5024) activity may be controlled by that operation (e.g., controlled by the scheduler 5028). Scheduler 5028 may schedule an operation or operations of network dataflow endpoint circuit 5000, for example, when (e.g., all) input (e.g., payload data and/or control data) arrives. Dotted lines and from scheduler 5028 indicate paths that may be utilized for control data, e.g., to and/or from scheduler 5028. Scheduler may also control multiplexer 5016, e.g., to steer data to and/or from network dataflow endpoint circuit 5000 and network 5014.

[0405] In reference to the distributed pick operation in FIG. 49 above, network dataflow endpoint circuit 4902 may be configured (e.g., as an operation in its operation configuration register 5026 as in FIG. 50) to receive (e.g., in two storage locations in) its network ingress buffer 5024 as in FIG. 50 input data from each of network dataflow endpoint circuit 4904 and network dataflow endpoint circuit 4906, and to output resultant data (e.g., from its spatial array egress buffer 5008 as in FIG. 50), for example, according to control data (e.g., in its spatial array ingress buffer 5002 as in FIG. 50). Network dataflow endpoint circuit 4904 may be configured (e.g., as an operation in its operation configuration register 5026 as in FIG. 50) to provide (e.g., send via circuit 4904’s network egress buffer 5022 as in FIG. 50) input data to network dataflow endpoint circuit 4902, e.g., on receipt (e.g., in circuit 4904’s spatial array ingress buffer 5002 as in FIG. 50) of the input data from processing element 4922. This may be referred to as Input 0 in FIG. 49. In one embodiment, circuit switched network is configured (e.g., programmed) to provide a dedicated communication line between processing element 4922 and network dataflow endpoint circuit 4904 along path 4924. Network dataflow endpoint circuit 4904 may include (e.g., add) a header packet to the received data (e.g., in its network egress buffer 5022 as in FIG. 50) to steer the packet (e.g., input data) to network dataflow endpoint circuit 4902. Network dataflow endpoint circuit 4906 may be configured (e.g., as an operation in its operation configuration register 5026 as in FIG. 50) to provide (e.g., send via circuit 4906’s network egress buffer 5022 as in FIG. 50) input data to network
dataflow endpoint circuit 4902, e.g., on receipt (e.g., in circuit 4906’s spatial array ingress buffer 5002 as in FIG. 50) of the input data from processing element 4920. This may be referred to as Input 1 in FIG. 49. In one embodiment, circuit switched network is configured (e.g., programmed) to provide a dedicated communication line between processing element 4920 and network dataflow endpoint circuit 4906 along path 4916. Network dataflow endpoint circuit 4906 may include (e.g., add) a header packet with the received data (e.g., in its network egress buffer 5022 as in FIG. 50) to steer the packet (e.g., input data) to network dataflow endpoint circuit 4902.

[0406] When network dataflow endpoint circuit 4904 is to transmit input data to network dataflow endpoint circuit 4902 (e.g., when network dataflow endpoint circuit 4902 has available storage room for the data and/or network dataflow endpoint circuit 4904 has its input data), network dataflow endpoint circuit 4904 may generate a packet (e.g., including the input data and a header to steer that data to network dataflow endpoint circuit 4902 on the packet switched communications network 4914 (e.g., as a stop on that (e.g., ring) network). This is illustrated schematically with dashed line 4926 in FIG. 49. Network 4914 is shown schematically with multiple dotted boxes in FIG. 49. Network 4914 may include a network controller 4914A, e.g., to manage the ingress and/or egress of data on network 4914A.

[0407] When network dataflow endpoint circuit 4906 is to transmit input data to network dataflow endpoint circuit 4902 (e.g., when network dataflow endpoint circuit 4902 has available storage room for the data and/or network dataflow endpoint circuit 4906 has its input data), network dataflow endpoint circuit 4906 may generate a packet (e.g., including the input data and a header to steer that data to network dataflow endpoint circuit 4902 on the packet switched communications network 4914 (e.g., as a stop on that (e.g., ring) network). This is illustrated schematically with dashed line 4918 in FIG. 49.

[0408] Network dataflow endpoint circuit 4902 (e.g., on receipt of the Input 0 from network dataflow endpoint circuit 4904 in circuit 4902’s network ingress buffer(s), Input 1 from network dataflow endpoint circuit 4906 in circuit 4902’s network ingress buffer(s), and/or control data from processing element 4908 in circuit 4902’s spatial array ingress buffer) may then perform the programmed dataflow operation (e.g., a Pick operation in this example). The network dataflow endpoint circuit 4902 may then output the according resultant data from the operation, e.g., to processing element 4908 in FIG. 49. In one embodiment, circuit switched network is configured (e.g., programmed) to provide a dedicated communication line between processing element 4908 (e.g., a buffer therein) and network dataflow endpoint circuit 4902 along path 4928. A further example of a distributed Pick operation is discussed below in reference to FIG. 62-63B. Buffers in FIG. 49 may be the small, unlabeled boxes in each PE.

[0409] FIGS. 51-58 below include example data formats, but other data formats may be utilized. One or more fields may be included in a data format (e.g., in a packet). Data format may be used by network dataflow endpoint circuits, e.g., to transmit (e.g., send and/or receive) data between a first component (e.g., between a first network dataflow endpoint circuit and a second network dataflow endpoint circuit, component of a spatial array, etc.).

[0410] FIG. 51 illustrates data formats for a send operation 5102 and a receive operation 5104 according to embodiments of the disclosure. In one embodiment, send operation 5102 and receive operation 5104 are data formats of data transmitted on a packet switched communication network. Depicted send operation 5102 data format includes a destination field 5102A (e.g., indicating which component in a network the data is to be sent to), a channel field 5102B (e.g., indicating which channel in the network the data is to be sent on), and an input field 5102C (e.g., the payload or input data that is to be sent). Depicted receive operation 5104 includes an output field, e.g., which may also include a destination field (not depicted). These data formats may be used (e.g., for packet(s)) to handle moving data in and out of components. These configurations may be separable and/or happen in parallel. These configurations may use separate resources. The term channel may generally refer to the communication resources (e.g., in management hardware) which is used to associate with the request. Association of configuration and queue management hardware may be explicit.

[0411] FIG. 52 illustrates another data format for a send operation 5202 according to embodiments of the disclosure. In one embodiment, send operation 5202 is a data format of data transmitted on a packet switched communication network. Depicted send operation 5202 data format includes a type field 5202A (e.g., used to annotate special control packets, such as, but not limited to, configuration, extraction, or exception packets), destination field 5202B (e.g., indicating which component in a network the data is to be sent to), a channel field 5202C (e.g., indicating which channel in the network the data is to be sent on), and an input field 5202D (e.g., the payload or input data that is to be sent).

[0412] FIG. 53 illustrates configuration data formats to configure a circuit (e.g., network dataflow endpoint circuit) for a send (e.g., switch) operation 5302 and a receive (e.g., pick) operation 5304 according to embodiments of the disclosure. In one embodiment, send operation 5302 and receive operation 5304 are configuration data formats for data to be transmitted on a packet switched communication network, for example, between network dataflow endpoint circuits. Depicted send operation configuration data format 5302 includes a destination field 5302A (e.g., indicating which component(s) in a network the (input) data is to be sent to), a channel field 5302B (e.g., indicating which channel in the network the (input) data is to be sent on), an input field 5302C (for example, an identifier of the component(s) that is to send the input data, e.g., the set of inputs in the (e.g., fabric ingress) buffer that this element is sensitive to), and an operation field 5302D (e.g., indicating which of a plurality of operations are to be performed). In one embodiment, the (e.g., outbound) operation is one of a Switch or SwitchAny dataflow operation, e.g., corresponding to a (e.g., same) dataflow operator of a dataflow graph.

[0413] Depicted receive operation configuration data format 5304 includes an output field 5304A (e.g., indicating which component(s) in a network the (resultant) data is to be sent to), an input field 5304B (e.g., an identifier of the component(s) that is to send the input data), and an operation field 5304C (e.g., indicating which of a plurality of operations are to be performed). In one embodiment, the (e.g., inbound) operation is one of a Pick, PickSingle, PickAny, or Merge dataflow operation, e.g., corresponding to a (e.g., same) dataflow operator of a dataflow graph.
one embodiment, a merge dataflow operation is a pick that requires and dequeues all operands (e.g., with the egress endpoint receiving control).

[0414] A configuration data format utilized herein may include one or more of the fields described herein, e.g., in any order.

[0415] FIG. 54 illustrates a configuration data format 5402 to configure a circuit element (e.g., network dataflow endpoint circuit) for a send operation with its input, output, and control data annotated on a circuit 5400 according to embodiments of the disclosure. Depicted is send operation configuration data format 5402 includes a destination field 5402A (e.g., indicating which component in a network the data is to be sent to), a channel field 5402B (e.g., indicating which channel on the (packet switched) network the data is to be sent on), and an input field 5402C (e.g., an identifier of the component(s) that is to send the input data). In one embodiment, circuit 5400 (e.g., network dataflow endpoint circuit) is to receive packet of data in the data format ofSend operation configuration data format 5402, for example, with the destination indicating which circuit of a plurality of circuits the resultant is to be sent to, the channel indicating which channel of the (packet switched) network the data is to be sent on, and the input identifying which circuit of a plurality of circuits the input data is to be received from. The AND gate 5404 is to allow the operation to be performed when both the input data is available and the credit status is a yes (for example, the dependency token indicates) indicating there is room for the output data to be stored, e.g., in a buffer of the destination. [0418] In one embodiment, the send operation does not utilize control beyond checking its input(s) are available for sending. This may enable switch to perform the operation without credit on all legs. In one embodiment, the Switch and/or SwitchAny operation includes a multiplexer controlled by the value stored in the operation field 5502D to select the correct queue management circuitry.

[0419] Value stored in operation field 5502D may select among control options, e.g., with different control (e.g., logic) circuitry for each operation, for example, as in FIGS. 56-59. In some embodiments, credit (e.g., credit on a network) status is another input (e.g., as depicted in FIGS. 56-57 here).

[0420] FIG. 56 illustrates a configuration data format to configure a circuit element (e.g., network dataflow endpoint circuit) for a Switch operation configuration data format 5602 with its input, output, and control data annotated on a circuit 5600 according to embodiments of the disclosure. In one embodiment, the (e.g., outbound) operation value stored in the operation field 5602D is for a Switch operation, e.g., corresponding to a Switch dataflow operation of a dataflow graph. In one embodiment, circuit 5600 (e.g., network dataflow endpoint circuit) is to receive a packet of data in the data format of Switch operation 5602, for example, with the input in input field 5602A being what component(s) are to be sent the data and the operation field 5602B indicating which operation is to be performed (e.g., shown schematically as Switch). Depicted circuit 5600 may select the operation to be executed from a plurality of available operations based on the operation field 5602B. In one embodiment, circuit 5500 is to perform that operation when both the input data (for example, according to the input status, e.g., there is room for the data in the destination(s)) is available and the credit status (e.g., selection operation (OP) status) is a yes (for example, the network credit indicates that there is availability on the network to send that data to the destination(s)). For example, multiplexers 5610, 5612, 5614 may be used with a respective input status and credit status for each input (e.g., where the output data is to be sent to in the switch operation), e.g., to prevent an input from showing as available until both the input status (e.g., room for data in the destination) and the credit status (e.g., there is room on the network to get to the destination) are true (e.g., yes). In one embodiment, input status is an indication there is or is not room for the (output) data to be stored, e.g., in a buffer of the destination. In certain embodiments, AND gate 5606 is to allow the operation to be performed when both the input data is available (e.g., as output from multiplexer 5604) and the selection operation (e.g., control data) status is a yes, for example, indicating the selection operation (e.g., which of a plurality of outputs an input is to be sent to, see, e.g., FIG. 48). In certain embodiments, the performance of the operation with the control data (e.g., selection op) is to cause input data from one of the inputs to be output on one or more (e.g., a plurality of) outputs (e.g., as indicated by the control data), e.g., according to the multiplexer selection bits from multiplexer 5608. In one embodiment, selection op chooses which leg of the switch output will be used and/or selection decoder creates multiplexer selection bits.

[0421] FIG. 57 illustrates a configuration data format to configure a circuit element (e.g., network dataflow endpoint circuit) for a SwitchAny operation configuration data format...
5702 with its input, output, and control data annotated on a circuit 5700 according to embodiments of the disclosure. In one embodiment, the (e.g., outbound) operation value stored in the operation field 5502D is for a SwitchAny operation, e.g., corresponding to a SwitchAny dataflow operator of a dataflow graph. In one embodiment, circuit 5700 (e.g., network dataflow endpoint circuit) is to receive a packet of data in the data format of SwitchAny operation configuration data format 5702, for example, with the input in input field 5702A being what component(s) are to be sent the data and the operation field 5702B indicating which operation is to be performed (e.g., shown schematically as SwitchAny) and/or the source of the control data for that operation. In one embodiment, circuit 5500 is to perform that operation when any of the input data (for example, according to the input status, e.g., there is room for the data in the destination (s)) is available and the credit status is a yes (for example, the network credit indicates that there is availability on the network to send that data to the destination(s)). For example, multiplexers 5710, 5712, 5714 may be used with a respective input status and credit status for each input (e.g., where the output data is to be sent to in the SwitchAny operation), e.g., to prevent an input from showing as available until both the input status (e.g., room for data in the destination) and the credit status (e.g., there is room on the network to get to the destination) are true (e.g., yes). In one embodiment, input status is an indication there is room or is not room for the (output) data to be stored, e.g., in a buffer of the destination. In certain embodiments, OR gate 5704 is to allow the operation to be performed when any one of the outputs are available. In certain embodiments, the performance of the operation is to be the first available input data from one of the inputs to be output on one or more (e.g., a plurality of) outputs, e.g., according to the multiplexer selection bits from multiplexer 5706. In one embodiment, SwitchAny occurs as soon as any output credit is available (e.g., as opposed to a Switch that utilizes a selection op). Multiplexer select bits may be used to steer an input to an (e.g., network) egress buffer of a network dataflow endpoint circuit.

[0423] FIG. 59 illustrates a configuration data format to configure a circuit element (e.g., network dataflow endpoint circuit) for a PickAny operation configuration data format 5802 with its input, output, and control data annotated on a circuit 5800 according to embodiments of the disclosure. In one embodiment, the (e.g., inbound) operation value stored in the operation field 5802D is for a Pick operation, e.g., corresponding to a Pick dataflow operator of a dataflow graph. In one embodiment, circuit 5800 (e.g., network dataflow endpoint circuit) is to receive a packet of data in the data format of Pick operation configuration data format 5802, for example, with the data in input field 5802A being what component(s) are to send the input data, the data in output field 5802B being what component(s) are to be sent the input data, and the operation field 5802C indicating which operation is to be performed (e.g., shown schematically as Pick) and/or the source of the control data for that operation. Depicted circuit 5800 may select the operation to be executed from a plurality of available operations based on the operation field 5802C. In one embodiment, circuit 5800 is to perform that operation when any of the input data is available (e.g., as output from multiplexer 5804) and/or selection decoder creates multiplexer selection bits.

[0424] In one embodiment, PickAny executes on the presence of any data and/or selection decoder creates multiplexer selection bits.

[0425] FIG. 60 illustrates selection of an operation (6002, 6004, 6006) by a network dataflow endpoint circuit 6000 for performance according to embodiments of the disclosure. Pending operations storage 6001 (e.g., in scheduler 5028 in FIG. 50) may store one or more dataflow operations, e.g., according to the format(s) discussed herein. Scheduler (for example, based on a fixed priority or the oldest of the operations, e.g., that have all of their operands) may schedule an operation for performance. For example, scheduler
may select operation 6002, and according to a value stored in operation field, send the corresponding control signals from multiplexer 6008 and/or multiplexer 6010. As an example, several operations may be simultaneously executable in a single network dataflow endpoint circuit. Assuming all data is there, the “performable” signal (e.g., as shown in FIGS. 54–59) may be input as a signal into multiplexer 6012. Multiplexer 6012 may send as an output control signals for a selected operation (e.g., one of operation 6002, 6004, and 6006) that cause multiplexer 6008 to configure the connections in a network dataflow endpoint circuit to perform the selected operation (e.g., to source from or send data to buffer(s)). Multiplexer 6012 may send as an output control signals for a selected operation (e.g., one of operation 6002, 6004, and 6006) that cause multiplexer 6010 to configure the connections in a network dataflow endpoint circuit to remove data from the queue(s), e.g., consumed data. As an example, see the discussion herein about having data (e.g., token) removed. The “PE status” in FIG. 60 may be the control data coming from a PE, for example, the empty indicator and full indicators of the queues (e.g., backpressure signals and/or network credit). In one embodiment, the PE status may include the empty or full bits for all the buffers and/or datapaths, e.g., in FIG. 50 herein. FIG. 60 illustrates generalized scheduling for embodiments hereinafter, e.g., with specialized scheduling for embodiments discussed in reference to FIGS. 56–59.

[0426] In one embodiment, (e.g., as with scheduling) the choice of dequence is determined by the operation and its dynamic behavior, e.g., to dequence the operation after performance. In one embodiment, a circuit is to use the operand selection bits to dequence data (e.g., input, output and/or control data).

[0427] FIG. 61 illustrates a network dataflow endpoint circuit 6100 according to embodiments of the disclosure. In comparison to FIG. 50, network dataflow endpoint circuit 6100 has split the configuration and control into two separate schedulers. In one embodiment, egress scheduler 6128A is to schedule an operation on data that is to enter (e.g., from a circuit switched communication network coupled to) the dataflow endpoint circuit 6100 (e.g., at argument queue 6102, for example, spatial array ingress buffer 5002 as in FIG. 50) and output (e.g., from a packet switched communication network coupled to) the dataflow endpoint circuit 6100 (e.g., at network egress buffer 6122, for example, network egress buffer 5022 as in FIG. 50). In one embodiment, ingress scheduler 6128B is to schedule an operation on data that is to enter (e.g., from a packet switched communication network coupled to) the dataflow endpoint circuit 6100 (e.g., at network ingress buffer 6124, for example, network ingress buffer 5024 as in FIG. 50) and output (e.g., from a circuit switched communication network coupled to) the dataflow endpoint circuit 6100 (e.g., at output buffer 6108, for example, spatial array egress buffer 5008 as in FIG. 50). Scheduler 6128A and/or scheduler 6128B may include as an input the (e.g., operating) status of circuit 6100, e.g., fullness level of inputs (e.g., buffers 6102A, 6102D), fullness level of outputs (e.g., buffers 6108), values (e.g., value in 6102A), etc. Scheduler 6128B may include a credit return circuit, for example, to denote that credit is returned to sender, e.g., after receipt in network ingress buffer 6124 of circuit 6100.

[0428] Network 6114 may be a circuit switched network, e.g., as discussed herein. Additionally or alternatively, a packet switched network (e.g., as discussed herein) may also be utilized, for example, coupled to network egress buffer 6122, network ingress buffer 6124, or other components herein. Argument queue 6102 may include a control buffer 6102A, for example, to indicate when a respective input queue (e.g., buffer) includes a (new) item of data, e.g., as a single bit. Turning now to FIGS. 62–63B, in one embodiment, these cumulatively show the configurations to create a distributed pick.

[0429] FIG. 62 illustrates a network dataflow endpoint circuit 6200 receiving input zero (0) while performing a pick operation according to embodiments of the disclosure, for example, as discussed above in reference to FIG. 49. In one embodiment, egress configuration 6226A is loaded (e.g., during a configuration step) with a portion of a pick operation that is to send data to a different network dataflow endpoint circuit (e.g., circuit 63B00 in FIG. 63B). In one embodiment, egress scheduler 6228A is to monitor the argument queue 6202 (e.g., data queue) for input data (e.g., from a processing element). According to an embodiment of the depicted data format, the “send” (e.g., a binary value therefor) indicates data is to be sent according to fields X, Y, with X being the value indicating a particular target network dataflow endpoint circuit (e.g., 0 being network dataflow endpoint circuit 63B300 in FIG. 63B) and Y being the value indicating which network ingress buffer (e.g., buffer 63B24) location the value is to be stored. In one embodiment, X is the value indicating a particular channel of a multiple channel (e.g., packet switched) network (e.g., 0 being channel 0 and/or buffer element 0 of network dataflow endpoint circuit 63B300 in FIG. 63B). When the input data arrives, it is then to be sent (e.g., from network egress buffer 6222) by network dataflow endpoint circuit 6200 to a different network dataflow endpoint circuit (e.g., network dataflow endpoint circuit 63B300 in FIG. 63B).

[0430] FIG. 63A illustrates a network dataflow endpoint circuit 63A00 receiving input one (1) while performing a pick operation according to embodiments of the disclosure, for example, as discussed above in reference to FIG. 49. In one embodiment, egress configuration 63A26A is loaded (e.g., during a configuration step) with a portion of a pick operation that is to send data to a different network dataflow endpoint circuit (e.g., circuit 63B300 in FIG. 63B). In one embodiment, egress scheduler 63A28A is to monitor the argument queue 63A20 (e.g., data queue 63A02B) for input data (e.g., from a processing element). According to an embodiment of the depicted data format, the “send” (e.g., a binary value therefor) indicates data is to be sent according to fields X, Y, with X being the value indicating a particular target network dataflow endpoint circuit (e.g., 0 being network dataflow endpoint circuit 63B300 in FIG. 63B) and Y being the value indicating which network ingress buffer (e.g., buffer 63B24) location the value is to be stored. In one embodiment, Y is the value indicating a particular channel of a multiple channel (e.g., packet switched) network (e.g., 1 being channel 1 and/or buffer element 1 of network dataflow endpoint circuit 63B300 in FIG. 63B). When the input data arrives, it is then to be sent (e.g., from network egress buffer 6222) by network dataflow endpoint circuit 63A00 to a different network dataflow endpoint circuit (e.g., network dataflow endpoint circuit 63B300 in FIG. 63B).

[0431] FIG. 63B illustrates a network dataflow endpoint circuit 63B00 outputting the selected input while performing a pick operation according to embodiments of the disclo-
sure, for example, as discussed above in reference to FIG. 49. In one embodiment, other network dataflow endpoint circuits (e.g., circuit 6200 and circuit 63A00) are to send their input data to network ingress buffer 63B24 of circuit 63B300. In one embodiment, ingress configuration 633263 is loaded (e.g., during a configuration step) with a portion of a pick operation that is to pick the data sent to network dataflow endpoint circuit 63B300, e.g., according to a control value. In one embodiment, control value is to be received in ingress control 63332 (e.g., buffer). In one embodiment, ingress scheduler 63A28A is to monitor the receipt of the control value and the input values (e.g., in network ingress buffer 63B324). For example, if the control value says pick from buffer element A (e.g., 0 or 1 in this example) (e.g., from channel A) of network ingress buffer 63B324, the value stored in that buffer element A is then output as a resultant of the operation by circuit 63B300, for example, into an output buffer 63B308, e.g., when output buffer has storage space (e.g., as indicated by a backpressure signal). In one embodiment, circuit 63B300’s output data is sent out when the egress buffer has a token (e.g., input data and control data) and the receiver asserts that it has buffer (e.g., indicating storage is available, although other assignments of resources are possible, this example is simply illustrative).

[0432] FIG. 64 illustrates a flow diagram 6400 according to embodiments of the disclosure. Depicted flow 6400 includes providing a spatial array of processing elements; routing, with a packet switched communications network, data within the spatial array between processing elements according to a dataflow graph 6402; performing a first dataflow operation of the dataflow graph with the processing elements 6404; and performing a second dataflow operation of the dataflow graph with a plurality of network dataflow endpoint circuits of the packet switched communications network 6406.

[0433] Referring again to FIG. 8, accelerator (e.g., CSA) 802 may perform (e.g., or request performance of) an access (e.g., a load and/or store) of data to one or more of plurality of cache banks (e.g., cache bank 808). A memory interface circuit (e.g., request address file (RAF) circuit(s)) may be included, e.g., as discussed herein, to provide access between memory (e.g., cache banks) and the accelerator 802. Referring again to FIG. 47, a requesting circuit (e.g., a processing element) may perform (e.g., or request performance of) an access (e.g., a load and/or store) of data to one or more of plurality of cache banks (e.g., cache bank 4702). A memory interface circuit (e.g., request address file (RAF) circuit(s)) may be included, e.g., as discussed herein, to provide access between memory (e.g., one or more banks of the cache memory) and the accelerator (e.g., one or more of accelerator tiles (4708, 4710, 4712, 4714)). Referring again to FIGS. 49 and/or 50, a requesting circuit (e.g., a processing element) may perform (e.g., or request performance of) an access (e.g., a load and/or store) of data to one or more of a plurality of cache banks. A memory interface circuit (for example, request address file (RAF) circuit(s), e.g., RAF/ interface cache 4912) may be included, e.g., as discussed herein, to provide access between memory (e.g., one or more banks of the cache memory) and the processor (e.g., one or more of the processing elements and/or network dataflow endpoint circuits (e.g., circuits 4902, 4904, 4906)).

[0434] In certain embodiments, an accelerator (e.g., a PE thereof) couples to a RAF circuit or a plurality of RAF circuits through (i) a circuit switched network (for example, as discussed herein, e.g., in reference to FIGS. 5-47) or (ii) through a packet switched network (for example, as discussed herein, e.g., in reference to FIGS. 48-64).

[0435] In certain embodiments, a circuit (e.g., a request address file (RAF) circuit) (e.g., each of a plurality of RAF circuits) includes a translation lookaside buffer (TLB) (e.g., TLB circuit). TLB may receive an input of a virtual address and output a physical address corresponding to the mapping (e.g., address mapping) of the virtual address to the physical address (e.g., different than any mapping of a dataflow graph to hardware). A virtual address may be an address as seen by a program running on a circuit (e.g., on an accelerator and/or processor). A physical address may be an (e.g., different than the virtual) address in memory hardware. A TLB may include a data structure (e.g., table structure) (e.g., recently used) virtual-to-physical memory address translations, e.g., such that the translation does not have to be performed on each virtual address present to obtain the physical memory address corresponding to that virtual address. If the virtual address entry is not in the TLB, a circuit (e.g., a TLB manager circuit) may perform a page walk to determine the virtual-to-physical memory address translation. In one embodiment, a circuit (e.g., a RAF circuit) is to receive an input of a virtual address for translation in a TLB (e.g., TLB in RAF circuit) from a requesting entity (e.g., a PE or other hardware component) via a circuit switched network, e.g., as in FIGS. 5-47. Additionally or alternatively, a circuit (e.g., a RAF circuit) may receive an input of a virtual address for translation in a TLB (e.g., TLB in RAF circuit) from a requesting entity (e.g., a PE, network dataflow endpoint circuit, or other hardware component) via a packet switched network, e.g., as in FIGS. 48-64. In certain embodiments, data received for a memory (e.g., cache) access request is a memory command. A memory command may include the virtual address to be accessed, operation to be performed (e.g., a load or a store), and/or payload data (e.g., for a store).

2.7 Floating Point Support

[0436] Certain HPC applications are characterized by their need for significant floating point bandwidth. To meet this need, embodiments of a CSA may be provisioned with multiple (e.g., between 128 and 256 each) of floating add and multiplication PEs, e.g., depending on tile configuration. A CSA may provide a few other extended precision modes, e.g., to simplify math library implementation. CSA floating point PEs may support both single and double precision, but lower precision PEs may support machine learning workloads. A CSA may provide an order of magnitude more floating point performance than a processor core. In one embodiment, in addition to increasing floating point bandwidth, in order to power all of the floating point units, the energy consumed in floating point operations is reduced. For example, to reduce energy, a CSA may selectively gate the low-order bits of the floating point multiplier array. In examining the behavior of floating point arithmetic, the low order bits of the multiplication array may often not influence the final, rounded product. FIG. 65 illustrates a floating point multiplier 6500 partitioned into three regions (the result region, three potential carry regions (6502, 6504, 6506), and the gated region) according to embodiments of the disclosure. In certain embodiments, the carry region is likely to influence the result region and the gated region is unlikely to
influence the result region. Considering a gated region of g bits, the maximum carry may be:

\[
carry_g \leq \frac{1}{2^g} \sum_{i=0}^{g-1} \frac{1}{2^i} = \sum_{i=0}^{g-1} \frac{1}{2^i} \quad |g| \geq 1
\]

Given this maximum carry, if the result of the carry region is less than \(2^r-g\), where the carry region is c bits wide, then the gated region may be ignored since it does not influence the result region. Increasing g means that it is more likely the gated region will be needed, while increasing c means that, under random assumption, the gated region will be unused and may be disabled to avoid energy consumption. In embodiments of a CSA floating multiplication PE, a two stage pipelined approach is utilized in which first the carry region is determined and then the gated region is determined if it is found to influence the result. If more information about the context of the multiplication is known, a CSA more aggressively tunes the size of the gated region. In FMA, the multiplication result may be added to an accumulator, which is often much larger than either of the multiplicands. In this case, the addend exponent may be observed in advance of multiplication and the CSDA may adjust the gated region accordingly. One embodiment of the CSA includes a scheme in which a context value, which bounds the minimum result of a computation, is provided to related multipliers, in order to select minimum energy gating configurations.

2.8 Runtime Services

In certain embodiment, a CSA includes a heterogeneous and distributed fabric, and consequently, runtime service implementations are to accommodate several kinds of PEs in a parallel and distributed fashion. Although runtime services in a CSA may be critical, they may be infrequent relative to user-level computation. Certain implementations, therefore, focus on overlaying services on hardware resources. To meet these goals, CSA runtime services may be cast as a hierarchy, e.g., with each layer corresponding to a CSA network. At the tile level, a single external-facing controller may accept or sends service commands to an associated core with the CSA tile. A tile-level controller may serve to coordinate regional controllers at the RAIs, e.g., using the ACPI network. In turn, regional controllers may coordinate local controllers at certain mezzanine network stops (e.g., network dataflow endpoint circuits). At the lowest level, service specific micro-protocols may execute over the local network, e.g., during a special mode controlled through the mezzanine controllers. The micro-protocols may permit each PE (e.g., PE class by type) to interact with the runtime service according to its own needs. Parallelism is thus implicit in this hierarchical organization, and operations at the lowest levels may occur simultaneously. This parallelism may enable the configuration of a CSA tile in between hundreds of nanoseconds to a few microseconds, e.g., depending on the configuration size and its location in the memory hierarchy. Embodiments of the CSA thus leverage properties of dataflow graphs to improve implementation of each runtime service. One key observation is that runtime services may need only to preserve a legal logical view of the dataflow graph, e.g., a state that can be produced through some ordering of dataflow operator executions. Services may generally not need to guarantee a temporal view of the dataflow graph, e.g., the state of a dataflow graph in a CSA at a specific point in time. This may permit the CSA to conduct most runtime services in a distributed, pipelined, and parallel fashion, e.g., provided that the service is orchestrated to preserve the logical view of the dataflow graph. The local configuration micro-protocol may be a packet-based protocol overlaid on the local network. Configuration targets may be organized into a configuration chain, e.g., which is fixed in the microarchitecture. Fabric (e.g., PE) targets may be configured one at a time, e.g., using a single extra register per target to achieve distributed coordination. To start configuration, a controller may drive an out-of-band signal which places all fabric targets in its neighborhood into an unconfigured, paused state and swings multiplexors in the local network to a pre-defined configuration. As the fabric (e.g., PE) targets are configured, that is they completely receive their configuration packet, they may set their configuration microprotocol registers, notifying the immediately succeeding target (e.g., PE) that it may proceed to configure using the subsequent packet. There is no limitation to the size of a configuration packet, and packets may have dynamically variable length. For example, PEs configuring constant operands may have a configuration packet that is lengthened to include the constant field (e.g., X and Y in FIGS. 3B-3C). FIG. 66 illustrates an in-flight configuration of an accelerator 6600 with a plurality of processing elements (e.g., PEs 6602, 6604, 6606, 6608) according to embodiments of the disclosure. Once configured, PEs may execute subject to dataflow constraints. However, channels involving unconfigured PEs may be disabled by the microarchitecture, e.g., preventing any undefined operations from occurring. These properties allow embodiments of a CSA to initialize and execute in a distributed fashion with no centralized control whatsoever. From an unconfigured state, configuration may occur completely in parallel, e.g., in perhaps as few as 200 nanoseconds. However, due to the distributed initialization of embodiments of a CSA, PEs may become active, for example sending requests to memory, well before the entire fabric is configured. Extraction may proceed in much the same way as configuration. The local network may be conformed to extract data from one target at a time, and state bits used to achieve distributed coordination. A CSA may orchestrate extraction to be non-destructive, that is, at the completion of extraction each extractable target has returned to its starting state. In this implementation, all state in the target may be circulated to an egress register tied to the local network in a scan-like fashion. Although in-place extraction may be achieved by introducing new paths at the register-transfer level (RTL), or using existing lines to provide the same functionalities with lower overhead. Like configuration, hierarchical extraction is achieved in parallel. FIG. 67 illustrates a snapshot 6700 of an in-flight, pipelined extraction according to embodiments of the disclosure. In some use cases of extraction, such as checkpointing, latency may not be a concern so long as fabric throughput is maintained. In these cases, extraction may be orchestrated in a pipelined fashion. This arrangement, shown in FIG. 67, permits most of the fabric to continue executing, while a narrow region is disabled for extraction. Configuration and extraction may be coordinated and composed to achieve a pipelined context switch. Exceptions may differ qualitatively from configuration and extraction in that,
rather than occurring at a specified time, they arise anywhere in the fabric at any point during runtime. Thus, in one embodiment, the exception micro-protocol may not be overlaid on the local network, which is occupied by the user program at runtime, and utilizes its own network. However, by nature, exceptions are rare and insensitive to latency and bandwidth. Thus, certain embodiments of CSA utilize a packet-switched network to carry exceptions to the local mezzanine stop, e.g., where they are forwarded up the service hierarchy (e.g., as in FIG. 82). Packets in the local exception network may be extremely small. In many cases, a PE identification (ID) of only two to eight bits suffices as a complete packet, e.g., since the CSA may create a unique exception identifier as the packet traverses the exception service hierarchy. Such a scheme may be desirable because it also reduces the area overhead of producing exceptions at each PE.

3. Compilation

[0043] The ability to compile programs written in high-level languages onto a CSA may be essential for industry adoption. This section gives a high-level overview of compilation strategies for embodiments of a CSA. First is a proposal for a CSA software framework that illustrates the desired properties of an ideal production-quality toolchain. Next, a prototype compiler framework is discussed. A “control-to-dataflow conversion” is then discussed, e.g., to convert ordinary sequential control-flow code into CSA dataflow assembly code.

3.1 Example Production Framework

[0044] FIG. 68 illustrates a compiler toolchain 6800 for an accelerator according to embodiments of the disclosure. This toolchain compiles high-level languages (such as C, C++, and Fortran) into a combination of host code (LLVM) intermediate representation (IR) for the specific regions to be accelerated. The CSA-specific portion of this compilation toolchain takes LLVM IR as its input, optimizes and compiles this IR into a CSA assembly, e.g., adding appropriate buffering on latency-insensitive channels for performance. It then places and routes the CSA assembly on the hardware fabric, and configures the PEs and network for execution. In one embodiment, the toolchain supports the CSA-specific compilation as a just-in-time (JIT), incorporating potential runtime feedback from actual executions. One of the key design characteristics of the framework is compilation of LLVM IR for the CSA, rather than using a higher-level language as input. While programs written in a high-level programming language designed specifically for the CSA might achieve maximal performance and/or energy efficiency, the adoption of new high-level languages or programming frameworks may be slow and limited in practice because of the difficulty of converting existing code bases. Using (LLVM) IR as input enables a wide range of existing programs to potentially execute on a CSA, e.g., without the need to create a new language or significantly modify the front-end of new languages that want to run on the CSA.

3.2 Prototype Compiler

[0045] FIG. 69 illustrates a compiler 6900 for an accelerator according to embodiments of the disclosure. Compiler 6900 initially focuses on ahead-of-time compilation of C and C++ through the (e.g., Clang) front-end. To compile (LLVM) IR, the compiler implements a CSA back-end target within LLVM with three main stages. First, the CSA back-end lowers LLVM IR into a target-specific machine instructions for the sequential unit, which implements most CSA operations combined with a traditional RISC-like control-flow architecture (e.g., with branches and a program counter). The sequential unit in the toolchain may serve as a useful aid for both compiler and application developers, since it enables an incremental transformation of a program from control flow (CF) to dataflow (DF), e.g., converting one section of code at a time from control-flow to dataflow and validating program correctness. The sequential unit may also provide a model for handling code that does not fit in the spatial array. Next, the compiler converts these control-flow instructions into dataflow operators (e.g., code) for the CSA. This phase is described later in Section 3.3. Then, the CSA back-end may run its own optimization passes on the dataflow instructions. Finally, the compiler may dump the instructions in a CSA assembly format. This assembly format is taken as input to late-stage tools which place and route the dataflow instructions on the actual CSA hardware.

3.3 Control to Dataflow Conversion

[0046] A key portion of the compiler may be implemented in the control-to-dataflow conversion pass, or dataflow conversion pass for short. This pass takes in a function represented in control flow form, e.g., a control-flow graph (CFG) with sequential machine instructions operating on virtual registers, and converts it into a dataflow function that is conceptually a graph of dataflow operations (instructions) connected by latency-insensitive channels (LICs). This section gives a high-level description of this pass, describing how it conceptually deals with memory operations, branches, and loops in certain embodiments.

Straight-Line Code

[0047] FIG. 70A illustrates sequential assembly code 7002 according to embodiments of the disclosure. FIG. 70A illustrates dataflow assembly code 7004 for the sequential assembly code 7002 of FIG. 70A according to embodiments of the disclosure. FIG. 70B illustrates a dataflow graph 7006 for the dataflow assembly code 7004 of FIG. 70A for an accelerator according to embodiments of the disclosure.

[0048] First, consider the simple case of converting straight-line sequential code to dataflow. The dataflow conversion pass may convert a basic block of sequential code, such as the code shown in FIG. 70A into CSA assembly code, shown in FIG. 70B. Conceptually, the CSA assembly in FIG. 70B represents the dataflow graph shown in FIG. 70C. In this example, each sequential instruction is translated into a matching CSA assembly. The .lic statements (e.g., for data) declare latency-insensitive channels which correspond to the virtual registers in the sequential code (e.g., Rdata). In practice, the input to the dataflow conversion pass may be in numerical virtual registers. For clarity, however, this section uses descriptive register names. Note that load and store operations are supported in the CSA architecture in this embodiment, allowing for more programs to run than an architecture supporting only pure dataflow. Since the sequential code input to the compiler is in SSA (single-statistic assignment) form, for a simple basic block, the control-to-dataflow pass may convert each virtual register definition into the production of a single value on a
latency-insensitive channel. The SSA form allows multiple uses of a single definition of a virtual register, such as in Rdata2). To support this model, the CSA assembly code supports multiple uses of the same LIC (e.g., data2), with the simulator implicitly creating the necessary copies of the LICs. One key difference between sequential code and dataflow code is in the treatment of memory operations. The code in FIG. 70A is conceptually serial, which means that the load32 (ld32) of addr3 should appear to happen after the load32 of addr, in case that addr and addr3 addresses overlap.

Branches

[0445] To convert programs with multiple basic blocks and conditionals to dataflow, the compiler generates special dataflow operators to replace the branches. More specifically, the compiler uses switch operators to steer outgoing data at the end of a basic block in the original CFG, and pick operators to select values from the appropriate incoming channel at the beginning of a basic block. As a concrete example, consider the code and corresponding dataflow graph in FIGS. 71A-71C, which conditionally computes a value of y based on several inputs: a, i, x, and n. After computing the branch condition test, the dataflow code uses a switch operator (e.g., see FIGS. 3B-3C) steers the value in channel x to channel xF if test is 0, or channel xT if test is 1. Similarly, a pick operator (e.g., see FIGS. 3B-3C) is used to send channel yF to y if test is 0, or send channel yT to y if test is 1. In this example, it turns out that even though the value of a is only used in the true branch of the conditional, the CSA is to include a switch operator which steers it to channel aT when test is 1, and consumes (eats) the value when test is 0. This latter case is expressed by setting the false output of the switch to % ign. It may not be correct to simply connect channel a directly to the true path, because in the cases where execution actually takes the false path, this value of “a” will be left over in the graph, leading to incorrect value of a for the next execution of the function. This example highlights the property of control equivalence, a key property in embodiments of correct dataflow conversion.

[0446] Control Equivalence: Consider a single-entry-single-exit control flow graph G with two basic blocks A and B. A and B are control-equivalent if all complete control flow paths through G visit A and B the same number of times.

[0447] LIC Replacement: In a control flow graph G, suppose an operation in basic block A defines a virtual register x, and an operation in basic block B that uses x. Then a correct control-to-dataflow transformation can replace x with a latency-insensitive channel only if A and B are control equivalent. The control-equivalence relation partitions the basic blocks of a CFG into strong control-dependence regions. FIG. 71A illustrates C source code WXX02 according to embodiments of the disclosure. FIG. 71B illustrates dataflow assembly code WXX04 for the C source code WXX02 of FIG. 71A according to embodiments of the disclosure. FIG. 71C illustrates a dataflow graph WXX06 for the dataflow assembly code WXX04 of FIG. 71B for an accelerator according to embodiments of the disclosure. In the example in FIGS. 71A-71B, the basic block before and after the conditionals are control-equivalent to each other, but the basic blocks in the true and false paths are each in their own control dependence region. One correct algorithm for converting a CFG to dataflow is to have the compiler insert (1) switches to compensate for the mismatch in execution frequency for any values that flow between basic blocks which are not control equivalent, and (2) picks at the beginning of basic blocks to choose correctly from any incoming values to a basic block. Generating the appropriate control signals for these picks and switches may be the key part of dataflow conversion.

Loops

[0448] Another important class of CGs in dataflow conversion are CFGs for single-entry-single-exit loops, a common form of loop generated in (LLVM) IR. These loops may be almost acyclic, except for a single back edge from the end of the loop back to a loop header block. The dataflow conversion pass may use some high level strategy to convert loops as for branches, e.g., it inserts switches at the end of the loop to direct values out of the loop (either out the loop exit or around the back-edge to the beginning of the loop), and inserts picks at the beginning of the loop to choose between initial values entering the loop and values coming through the back edge. FIG. 72A illustrates C source code WXX02 according to embodiments of the disclosure. FIG. 72B illustrates dataflow assembly code WXX04 for the C source code WXX02 of FIG. 72A according to embodiments of the disclosure. Figure WXXC illustrates a dataflow graph WXX06 for the dataflow assembly code WXX04 of FIG. 72B for an accelerator according to embodiments of the disclosure. FIGS. 72A-WXXC shows C and CSA assembly code for an example do-while loop that adds up values of a loop induction variable i, as well as the corresponding dataflow graph. For each variable that conceptually cycles around the loop (i and sum), this graph has a corresponding pick/pick switch pair that controls the flow of these values. Note that this example also uses a pick/pick switch pair to cycle the value of n around the loop, even though n is loop-invariant. This repetition of n enables conversion of n’s virtual register into a LIC, since it matches the execution frequencies between a conceptual definition of n outside the loop and the one or more uses of n inside the loop. In general, for a correct dataflow conversion, registers that are live-in into a loop are to be repeated once for each iteration inside the loop body when the register is converted into a LIC. Similarly, registers that are updated inside a loop are live-out from the loop are to be consumed, e.g., with a single final value sent out of the loop. Loops introduce a wrinkle into the dataflow conversion process, namely that the control for a pick at the top of the loop and the switch for the bottom of the loop are offset. For example, if the loop in FIG. 71A executes three iterations and exits, the control to picker should be 0, 1, 1, while the control to switcher should be 1, 1, 0. This control is implemented by starting the picker channel with an initial extra 0 when the function begins on cycle 0 (which is specified in the assembly by the directives .value 0 and .avail 0), and then copying the output switcher into picker. Note that the last 0 in switcher restores a final 0 into picker, ensuring that the final state of the dataflow graph matches its initial state.

[0449] FIG. 73A illustrates a flow diagram 7300 according to embodiments of the disclosure. Depicted flow 7300 includes decoding an instruction with a decoder of a core of a processor into a decoded instruction 7302, executing the decoded instruction with an execution unit of the core of the processor to perform a first operation 7304, receiving an input of a dataflow graph comprising a plurality of nodes
7306: overlaying the dataflow graph into a plurality of processing elements of the processor and an interconnect network between the plurality of processing elements of the processor with each node represented as a dataflow operator in the plurality of processing elements 7308; and performing a second operation of the dataflow graph with the interconnect network and the plurality of processing elements by a respective, incoming operand set arriving at each of the dataflow operators of the plurality of processing elements 7310.

[0450] FIG. 73B illustrates a flow diagram 7301 according to embodiments of the disclosure. Depicted flow 7301 includes receiving an input of a dataflow graph comprising a plurality of nodes 7303; and overlaying the dataflow graph into a plurality of processing elements of a processor, a data path network between the plurality of processing elements, and a flow control path network between the plurality of processing elements with each node represented as a dataflow operator in the plurality of processing elements 7305.

[0451] In one embodiment, the core writes a command into a memory queue and a CSA (e.g., the plurality of processing elements) monitors the memory queue and begins executing when the command is read. In one embodiment, the core executes a first part of a program and a CSA (e.g., the plurality of processing elements) executes a second part of the program. In one embodiment, the core does other work while the CSA is executing its operations.

4. CSA Advantages

[0452] In certain embodiments, the CSA architecture and microarchitecture provides profound energy, performance, and usability advantages over roadmap processor architectures and FPGAs. In this section, these architectures are compared to embodiments of the CSA and highlights the superiority of CSA in accelerating parallel dataflow graphs relative to each.

4.1 Processors

[0453] FIG. 74 illustrates a throughput versus energy per operation graph 7400 according to embodiments of the disclosure. As shown in FIG. 74, small cores are generally more energy efficient than large cores, and, in some workloads, this advantage may be translated to absolute performance through higher core counts. The CSA microarchitecture follows these observations to their conclusion and removes (e.g., most) energy-hungry control structures associated with von Neumann architectures, including most of the instruction-side microarchitecture. By removing these overheads and implementing simple, single operation PEs, embodiments of a CSA obtains a dense, efficient spatial array. Unlike small cores, which are usually quite serial, a CSA may gang its PEs together, e.g., via the circuit switched local network, to form explicitly parallel aggregate dataflow graphs. The result is performance in not only parallel applications, but also serial applications as well. Unlike cores, which may pay dearly for performance in terms area and energy, a CSA is already parallel in its native execution model. In certain embodiments, a CSA neither requires speculation to increase performance nor does it need to repeatedly re-extract parallelism from a sequential program representation, thereby avoiding two of the main energy taxes in von Neumann architectures. Most structures in embodiments of a CSA are distributed, small, and energy efficient, as opposed to the centralized, bulky, energy hungry structures found in cores. Consider the case of registers in the CSA: each PE may have a few (e.g., 10 or less) storage registers. Taken individually, these registers may be more efficient that traditional register files. In aggregate, these registers may provide the effect of a large, in-fabric register file. As a result, embodiments of a CSA avoids most of stack spills and fills incurred by classical architectures, while using much less energy per state access. Of course, applications may still access memory. In embodiments of a CSA, memory access request and response are architecturally decoupled, enabling workloads to sustain many more outstanding memory accesses per unit of area and energy. This property yields substantially higher performance for cache-bound workloads and reduces the area and energy needed to saturate main memory in memory-bound workloads. Embodiments of a CSA expose new forms of energy efficiency which are unique to non-von Neumann architectures. One consequence of executing a single operation (e.g., instruction) at a (e.g., most) PEs is reduced operand entropy. In the case of an increment operation, each execution may result in a handful of circuit-level toggles and little energy consumption, a case examined in detail in Section 5.2. In contrast, von Neumann architectures are multiplexed, resulting in large numbers of bit transitions. The asynchronous style of embodiments of a CSA also enables microarchitectural optimizations, such as the floating point optimizations described in Section 2.7 that are difficult to realize in tightly scheduled core pipelines. Because PEs may be relatively simple and their behavior in a particular dataflow graph be stochastically known, clock gating and power gating techniques may be applied more effectively than in coarser architectures. The graph-execution style, small size, and malleability of embodiments of CSA PEs and the network together enable the expression many kinds of parallelism: instruction, data, pipeline, vector, memory, thread, and task parallelism may all be implemented. For example, in embodiments of a CSA, one application may use arithmetic units to provide a high degree of address bandwidth, while another application may use those same units for computation. In many cases, multiple kinds of parallelism may be combined to achieve even more performance. Many key HPC operations may be both replicated and pipelined, resulting in orders-of-magnitude performance gains. In contrast, von Neumann-style cores typically optimize for one style of parallelism, carefully chosen by the architects, resulting in a failure to capture all important application kernels. Just as embodiments of a CSA expose and facilitates many forms of parallelism, it does not mandate a particular form of parallelism, or, worse, a particular subroutine be present in an application in order to benefit from the CSA. Many applications, including single-stream applications, may obtain both performance and energy benefits from embodiments of a CSA, e.g., even when compiled without modification. This reverses the long trend of requiring significant programmer effort to obtain a substantial performance gain in single-stream applications. Indeed, in some applications, embodiments of a CSA obtain more performance from functionally equivalent, but less “modern” codes than from their convoluted, contemporary cousins which have been tortured to target vector instructions.

4.2 Comparison of CSA Embodiments and FPGAs

[0454] The choice of dataflow operators as the fundamental architecture of embodiments of a CSA differentiates
those CSAs from a FPGA, and particularly the CSA is as superior accelerator for HPC dataflow graphs arising from traditional programming languages. Dataflow operators are fundamentally asynchronous. This enables embodiments of a CSA not only to have great freedom of implementation in the microarchitecture, but it also enables them to simply and succinctly accommodate abstract architectural concepts. For example, embodiments of a CSA naturally accommodate many memory microarchitectures, which are essentially asynchronous, with a simple load-store interface. One need only examine an FPGA DRAM controller to appreciate the difference in complexity. Embodiments of a CSA also leverage asynchrony to provide faster and more fully-featured runtime services like configuration and extraction, which are believed to be four to six orders of magnitude faster than an FPGA. By narrowing the architectural interface, embodiments of a CSA provide control over most timing paths at the microarchitectural level. This allows embodiments of a CSA to operate at a much higher frequency than the more general control mechanism offered in a FPGA. Similarly, clock and reset, which may be architecturally fundamental to FPGAs, are microarchitectural in the CSA, e.g., obviating the need to support them as programmable entities. Dataflow operators may be, for the most part, coarse-grained. By only dealing in coarse operators, embodiments of a CSA improve both the density of the fabric and its energy consumption: CSA executes operations directly rather than emulating them with look-up tables. A second consequence of coarseness is a simplification of the place and route problem. CSA dataflow graphs are many orders of magnitude smaller than FPGA net-lists and place and route time are commensurately reduced in embodiments of a CSA. The significant differences between embodiments of a CSA and a FPGA make the CSA superior as an accelerator, e.g., for dataflow graphs arising from traditional programming languages.

5. Evaluation

The CSA is a novel computer architecture with the potential to provide enormous performance and energy advantages relative to roadmap processors. Consider the case of computing a single stranded address for walking across an array. This case may be important in HPC applications, e.g., which spend significant integer effort in computing address offsets. In address computation, and especially stranded address computation, one argument is constant and the other varies only slightly per computation. Thus, only a handful of bits per cycle toggle in the majority of cases. Indeed, it may be shown, using a derivation similar to the bound on floating point carry bits described in Section 2.7, that less than two bits of input toggle per computation in average for a stride calculation, reducing energy by 50% over a random toggle distribution. Were a time-multiplexed approach used, much of this energy savings may be lost. In one embodiment, the CSA achieves approximately 3x energy efficiency over a core while delivering an 8x performance gain. The parallelism gains achieved by embodiments of a CSA may result in reduced program run times, yielding a proportionate, substantial reduction in leakage energy. At the PE level, embodiments of a CSA are extremely energy efficient. A second important question for the CSA is whether the CSA consumes a reasonable amount of energy at the tile level. Since embodiments of a CSA are capable of exercising every floating point PE in the fabric at every cycle, it serves as a reasonable upper bound for energy and power consumption, e.g., such that most of the energy goes into floating point multiply and add.

6. Further CSA Details

6.1 Microarchitecture for Configuring a CSA

This section discusses further details for configuration and exception handling. Configuring a CSA (e.g., fabric), how to achieve this configuration quickly, and how to minimize the resource overhead of configuration. Configuring the fabric quickly may be of preeminent importance in accelerating small portions of a larger algorithm, and consequently in broadening the applicability of a CSA. The section further discloses features that allow embodiments of a CSA to be programmed with configurations of different length.

6.2 Embodiments of a CSA that may differ from traditional cores in that they make use of a configuration step in which (e.g., large) parts of the fabric are loaded with program configuration in advance of program execution. An advantage of static configuration may be that very little energy is spent at runtime on the configuration, e.g., as opposed to sequential cores which spend energy fetching configuration information (an instruction) nearly every cycle. The previous disadvantage of configuration is that it was a coarse-grained step with a potentially large latency, which places an under-bound on the size of program that can be accelerated in the fabric due to the cost of context switching. This disclosure describes a scalable microarchitecture for rapidly configuring a spatial array in a distributed fashion, e.g., that avoids the previous disadvantages.

As discussed above, a CSA may include lightweight processing elements connected by an inter-PE network. Programs, viewed as control-dataflow graphs, are then mapped onto the architecture by configuring the configurable fabric elements (CFEs), for example PEs and the interconnect (fabric) networks. Generally, PEs may be configured as dataflow operators and once all input operands arrive at the PE, some operation occurs, and the results are forwarded to another PE or PEs for consumption or output. PEs may communicate over dedicated virtual circuits which are formed by statically configuring the circuit switched communications network. These virtual circuits may be flow controlled and fully back-pressure, e.g., such that PEs will stall if either the source has no data or destination is full. At runtime, data may flow through the PEs implementing the mapped algorithm. For example, data may be streamed in from memory, through the fabric, and then back out to memory. Such a spatial architecture may achieve remarkable performance efficiency relative to traditional multicore processors: compute, in the form of PEs, may be simpler and more numerous than larger cores and communications may be direct, as opposed to an extension of the memory system.

Embodiments of a CSA may not utilize (e.g., software controlled) packet switching, e.g., packet switching that requires significant software assistance to realize, which slows configuration. Embodiments of a CSA include out-of-band signaling in the network (e.g., of only 2-3 bits, depending on the feature set supported) and a fixed configuration topology to avoid the need for significant software support.

One key difference between embodiments of a CSA and the approach used in FPGAs is that a CSA approach may use a wide data word, is distributed, and
includes mechanisms to fetch program data directly from memory. Embodiments of a CSA may not utilize JTAG-style single bit communications in the interest of area efficiency, e.g., as that may require milliseconds to completely configure a large FPGA fabric.

[0462] Embodiments of a CSA include a distributed configuration protocol and microarchitecture to support this protocol. Initially, configuration state may reside in memory. Multiple (e.g., distributed) local configuration controllers (LCCs) may stream portions of the overall program into their local region of the spatial fabric, e.g., using a combination of a small set of control signals and the fabric-provided network. State elements may be used at each CFE to form configuration chains, e.g., allowing individual CFEs to self-program without global addressing.

[0463] Embodiments of a CSA include specific hardware support for the formation of configuration chains, e.g., not software establishing these chains dynamically at the cost of increasing configuration time. Embodiments of a CSA are not purely packet switched and do include extra out-of-band control wires (e.g., control is not sent through the data path requiring extra cycles to strobe this information and resequence this information). Embodiments of a CSA decreases configuration latency by fixing the configuration ordering and by providing explicit out-of-band control (e.g., by at least a factor of two), while not significantly increasing network complexity.

[0464] Embodiments of a CSA do not use a serial mechanism for configuration in which data is streamed bit by bit into the fabric using a JTAG-like protocol. Embodiments of a CSA utilize a coarse-grained fabric approach. In certain embodiments, adding a few control wires or state elements to a 64 or 32-bit-oriented CSA fabric has a lower cost relative to adding those same control mechanisms to a 4 or 6 bit fabric.

[0465] FIG. 77 illustrates an accelerator tile 7700 comprising an array of processing elements (PE) and a local configuration controller (7702, 7706) according to embodiments of the disclosure. Each PE, each network controller (e.g., network dataflow endpoint circuit), and each switch may be a configurable fabric elements (CFEs), e.g., which are configured (e.g., programmed) by embodiments of the CSA architecture.

[0466] Embodiments of a CSA include hardware that provides for efficient, distributed, low-latency configuration of a heterogeneous spatial fabric. This may be achieved according to four techniques. First, a hardware entity, the local configuration controller (LCC) is utilized, for example, as in FIGS. 77-79. An LCC may fetch a stream of configuration information from (e.g., virtual) memory. Second, a configuration data path may be included, e.g., that is as wide as the native width of the PE fabric and which may be overlaid on top of the PE fabric. Third, new control signals may be received into the PE fabric which orchestrate the configuration process. Fourth, state elements may be located (e.g., in a register) at each configurable endpoint which track the status of adjacent CFEs, allowing each CFE to unambiguously self-configure without extra control signals. These four microarchitectural features may allow a CSA to configure chains of its CFEs. To obtain low configuration latency, the configuration may be partitioned by building many LCCs and CFE chains. At configuration time, these may operate independently to load the fabric in parallel, e.g., dramatically reducing latency. As a result of these combinations, fabrics configured using embodiments of a CSA architecture, may be completely configured (e.g., in hundreds of nanoseconds). In the following, the detailed the operation of the various components of embodiments of a CSA configuration network are disclosed.

[0467] FIGS. 76A-76C illustrate a local configuration controller WX002 configuring a data path network according to embodiments of the disclosure. Depicted network includes a plurality of multiplexers (e.g., multiplexers WX006, WX008, WX010) that may be configured (e.g., via their respective control signals) to connect one or more data paths (e.g., from PEs) together. FIG. 76A illustrates the network WX000 (e.g., fabric) configured (e.g., set) for some previous operation or program. FIG. 76B illustrates the local configuration controller WX002 (e.g., including a network interface circuit WX004 to send and/or receive signals) strobing a configuration signal and the local network is set to a default configuration (e.g., as depicted) that allows the LCC to send configuration data to all configurable fabric elements (CFEs), e.g., muxes. FIG. 76C illustrates the LCC strobing configuration information across the network, configuring CFEs in a predetermined (e.g., silicon-defined) sequence. In one embodiment, when CFEs are configured they may begin operation immediately. In another embodiments, the CFEs wait to begin operation until the fabric has been completely configured (e.g., as signaled by configuration terminator (e.g., configuration terminator 7804 and configuration terminator 7808 in FIG. 78) for each local configuration controller). In one embodiment, the LCC obtains control over the network fabric by sending a special message, or driving a signal. It then strobes configuration data (e.g., over a period of many cycles) to the CFEs in the fabric. In these figures, the multiplexer networks are analogues of the “Switch” shown in certain Figures (e.g., FIG. 5).

Local Configuration Controller

[0468] FIG. 77 illustrates a (e.g., local) configuration controller 7702 according to embodiments of the disclosure. A local configuration controller (LCC) may be the hardware entity which is responsible for loading the local portions (e.g., in a subset of a tile or otherwise) of the fabric program, interpreting these program portions, and then loading these program portions into the fabric by driving the appropriate protocol on the various configuration wires. In this capacity, the LCC may be a special-purpose, sequential microcontroller.

[0469] LCC operation may begin when it receives a pointer to a code segment. Depending on the LCB microarchitecture, this pointer (e.g., stored in pointer register 7706) may come either over a network (e.g., from within the CSA (fabric itself)) or through a memory system access to the LCC. When it receives such a pointer, the LCC optionally drains relevant state from its portion of the fabric for context storage, and then proceeds to immediately reconfigure the portion of the fabric for which it is responsible. The program loaded by the LCC may be a combination of configuration data for the fabric and control commands for the LCC, e.g., which are lightly encoded. As the LCC streams in the program portion, it may interprets the program as a command stream and perform the appropriate encoded action to configure (e.g., load) the fabric.

[0470] Two different microarchitectures for the LCC are shown in FIG. 77, e.g., with one or both being utilized in a
CSA. The first places the LCC 7702 at the memory interface. In this case, the LCC may make direct requests to the memory system to load data. In the second case the LCC 7706 is placed on a memory network; in which it may make requests to the memory only indirectly. In both cases, the logical operation of the LCB is unchanged. In one embodiment, an LCCs is informed of the program to load, for example, by a set of (e.g., OS-visible) control-status-registers which will be used to inform individual LCCs of new program pointers, etc.

Extra Out-of-Band Control Channels (e.g., Wires)

[0471] In certain embodiments, configuration relies on 2-8 extra, out-of-band control channels to improve configuration speed, as defined below. For example, configuration controller 7702 may include the following control channels, e.g., CFG_START control channel 7708, CFG_VALID control channel 7710, and CFG_DONE control channel 7712, with examples of each discussed in Table 2 below.

<table>
<thead>
<tr>
<th>Control Channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFG_START</td>
</tr>
<tr>
<td>CFG_VALID</td>
</tr>
<tr>
<td>CFG_DONE</td>
</tr>
</tbody>
</table>

[0472] Generally, the handling of configuration information may be left to the implementer of a particular CFE. For example, a selectable function CFE may have a provision for setting registers using an existing data path, while a fixed function CFE might simply set a configuration register.

[0473] Due to long wire delays when programming a large set of CFEs, the CFG_VALID signal may be treated as a clock/latch enable for CFE components. Since this signal is used as a clock, in one embodiment the duty cycle of the line is at most 50%. As a result, configuration throughput is approximately halved. Optionally, a second CFG_VALID signal may be added to enable continuous programming.

[0474] In one embodiment, only CFG_START is strictly communicated on an independent coupling (e.g., wire), for example, CFG_VALID and CFG_DONE may be overlaid on top of other network couplings.

Reuse of Network Resources

[0475] To reduce the overhead of configuration, certain embodiments of a CSA make use of existing network infrastructure to communicate configuration data. A LCC may make use of both a chip-level memory hierarchy and a fabric-level communications networks to move data from storage into the fabric. As a result, in certain embodiments of a CSA, the configuration infrastructure adds no more than 2% to the overall fabric area and power.

[0476] Reuse of network resources in certain embodiments of a CSA may cause a network to have some hardware support for a configuration mechanism. Circuit switched networks of embodiments of a CSA cause an LCC to set their multiplexors in a specific way for configuration when the ‘CFG_START’ signal is asserted. Packet switched networks do not require extension, although LCC endpoints (e.g., configuration terminators) use a specific address in the packet switched network. Network reuse is optional, and some embodiments may find dedicated configuration buses to be more convenient.

Per CFE State

[0477] Each CFE may maintain a bit denoting whether or not it has been configured (see, e.g., FIG. 66). This bit may be de-asserted when the configuration start signal is driven, and then asserted once the particular CFE has been configured. In one configuration protocol, CFEs are arranged to form chains with the CFE configuration state bit determining the topology of the chain. A CFE may read the configuration state bit of the immediately adjacent CFE. If this adjacent CFE is configured and the current CFE is not configured, the CFE may determine that any current configuration data is targeted at the current CFE. When the ‘CFG_DONE’ signal is asserted, the CFE may set its configuration bit, e.g., enabling upstream CFEs to configure. As a base case to the configuration process, a configuration terminator (e.g., configuration terminator 7704 for LCC 7702 or configuration terminator 7708 for LCC 7706 in FIG. 77) which asserts that it is configured may be included at the end of a chain.

[0478] Internal to the CFE, this bit may be used to drive flow control ready signals. For example, when the configuration bit is de-asserted, network control signals may automatically be clamped to a values that prevent data from flowing, while, within PEs, no operations or other actions will be scheduled.

Dealing with High-Delay Configuration Paths

[0479] One embodiment of an LCC may drive a signal over a long distance, e.g., through many multiplexors and with many loads. Thus, it may be difficult for a signal to arrive at a distant CFE within a short clock cycle. In certain embodiments, configuration signals are at some division (e.g., fraction of) of the main (e.g., CSA) clock frequency to ensure digital timing discipline at configuration. Clock division may be utilized in an out-of-band signaling protocol, and does not require any modification of the main clock tree.

Ensuring Consistent Fabric Behavior During Configuration

[0480] Since certain configuration schemes are distributed and have non-deterministic timing due to program and memory effects, different portions of the fabric may be configured at different times. As a result, certain embodiments of a CSA provide mechanisms to prevent inconsistent operation among configured and unconfigured CFEs. Generally, consistency is viewed as a property required of and maintained by CFEs themselves, e.g., using the internal CFE state. For example, when a CFE is in an unconfigured state, it may claim that its input buffers are full, and that its output is invalid. When configured, these values will be set to the true state of the buffers. As enough of the fabric comes out of configuration, these techniques may permit it to begin operation. This has the effect of further reducing context switching latency, e.g., if long-latency memory requests are issued early.

Variable-Width Configuration

[0481] Different CFEs may have different configuration word widths. For smaller CFE configuration words, implementers may balance delay by equitably assigning CFE
configuration loads across the network wires. To balance loading on network wires, one option is to assign configuration bits to different portions of network wires to limit the net delay on any one wire. Wide data words may be handled by using serialization/deserialization techniques. These decisions may be taken on a per-fabric basis to optimize the behavior of a specific CSA (e.g., fabric). Network controller (e.g., one or more of network controller 7710 and network controller 7712 may communicate with each domain (e.g., subset) of the CSA (e.g., fabric), for example, to send configuration information to one or more LCCs. Network controller may be part of a communications network (e.g., separate from circuit switched network). Network controller may include a network dataflow endpoint circuit.

6.2 Microarchitecture for Low Latency Configuration of a CSA and for Timely Fetching of Configuration Data for a CSA

[0482] Embodiments of a CSA may be an energy-efficient and high-performance means of accelerating user applications. When considering whether a program (e.g., a dataflow graph thereof) may be successfully accelerated by an accelerator, both the time to configure the accelerator and the time to run the program may be considered. If the run time is short, then the configuration time may play a large role in determining successful acceleration. Therefore, to maximize the domain of accelerated programs, in some embodiments the configuration time is made as short as possible. One or more configuration caches may be included in a CSA, e.g., such that the high bandwidth, low-latency store enables rapid reconfiguration. Next is a description of several embodiments of a configuration cache.

[0483] In one embodiment, during configuration, the configuration hardware (e.g., LCC) optionally accesses the configuration cache to obtain new configuration information. The configuration cache may operate either as a traditional address based cache, or in an OS managed mode, in which configurations are stored in the local address space and addressed by reference to that address space. If configuration state is located in the cache, then no requests to the backing store are to be made in certain embodiments. In certain embodiments, this configuration cache is separate from any (e.g., lower level) shared cache in the memory hierarchy.

[0484] FIG. 78 illustrates an accelerator tile 7800 comprising an array of processing elements, a configuration cache (e.g., 7818 or 7820), and a local configuration controller (e.g., 7802 or 7806) according to embodiments of the disclosure. In one embodiment, configuration cache 7814 is co-located with local configuration controller 7802. In one embodiment, configuration cache 7818 is located in the configuration domain of local configuration controller 7806, e.g., with a first domain ending at configuration terminator 7804 and a second domain ending at configuration terminator 7808. A configuration cache may allow a local configuration controller to refer to the configuration cache during configuration, e.g., in the hope of obtaining configuration state with lower latency than a reference to memory. A configuration cache (storage) may either be dedicated or may be accessed as a configuration mode of an in-fabric storage element, e.g., local cache 7816.

Caching Modes

[0485] 1. Demand Caching—In this mode, the configuration cache operates as a true cache. The configuration controller issues address-based requests, which are checked against tags in the cache. Misses are loaded into the cache and then may be re-referenced during subsequent reprogramming.

[0486] 2. In-Fabric Storage (Scratchpad) Caching—In this mode the configuration cache receives a reference to a configuration sequence in its own, small address space, rather than the larger address space of the host. This may improve memory density since the portion of cache used to store tags may instead be used to store configuration.

[0487] In certain embodiments, a configuration cache may have the configuration data pre-loaded into it, e.g., either by external direction or internal direction. This may allow reduction in the latency to load programs. Certain embodiments herein provide for an interface to a configuration cache which permits the loading of new configuration state into the cache, e.g., even if a configuration is running in the fabric already. The initiation of this load may occur from either an internal or external source. Embodiments of a pre-loading mechanism further reduce latency by removing the latency of cache loading from the configuration path.

Pre Fetching Modes

[0488] 1. Explicit Prefetching—A configuration path is augmented with a new command, Configuration-CachePrefetch. Instead of programming the fabric, this command simply cause a load of the relevant program configuration into a configuration cache, without programming the fabric. Since this mechanism piggybacks on the existing configuration infrastructure, it is exposed both within the fabric and externally, e.g., to cores and other entities accessing the memory space.

[0489] 2. Implicit prefetching—A global configuration controller may maintain a prefetch predictor, and use this to initiate the explicit prefetching to a configuration cache, e.g., in an automated fashion.

6.3 Hardware for Rapid Reconfiguration of a CSA in Response to an Exception

[0490] Certain embodiments of a CSA (e.g., a spatial fabric) include large amounts of instruction and configuration state, e.g., which is largely static during the operation of the CSA. Thus, the configuration state may be vulnerable to soft errors. Rapid and error-free recovery of these soft errors may be critical to the long-term reliability and performance of spatial systems.

[0491] Certain embodiments herein provide for a rapid configuration recovery loop, e.g., in which configuration errors are detected and portions of the fabric immediately reconfigured. Certain embodiments herein include a configuration controller, e.g., with reliability, availability, and serviceability (RAS) reprogramming features. Certain embodiments of CSA include circuitry for high-speed configuration, error reporting, and parity checking within the spatial fabric. Using a combination of these three features, and optionally, a configuration cache, a configuration/exception handling circuit may recover from soft errors in configuration. When detected, soft errors may be conveyed to a configuration cache which initiates an immediate reconfiguration of (e.g., that portion of) the fabric. Certain embodiments provide for a dedicated reconfiguration circuit, e.g., which is faster than any solution that would be indi-
rectly implemented in the fabric. In certain embodiments, co-located exception and configuration circuit cooperates to reload the fabric on configuration error detection.

FIG. 79 illustrates an accelerator tile 7900 comprising an array of processing elements and a configuration and exception handling controller (7902, 7906) with a reconfiguration circuit (7918, 7922) according to embodiments of the disclosure. In one embodiment, when a PE detects a configuration error through its local RAS features, it sends a (e.g., configuration error or reconfiguration error) message by its exception generator to the configuration and exception handling controller (e.g., 7902 or 7906). On receipt of this message, the configuration and exception handling controller (e.g., 7902 or 7906) initiates the co-located reconfiguration circuit (e.g., 7918 or 7922, respectively) to reload configuration state. The configuration microarchitecture proceeds and reloads (e.g., only) configuration state, and in certain embodiments, only the configuration state for the PE reporting the RAS error. Upon completion of reconfiguration, the fabric may resume normal operation. To decrease latency, the configuration state used by the configuration and exception handling controller (e.g., 7902 or 7906) may be sourced from a configuration cache. As a base case to the configuration or reconfiguration process, a configuration terminator (e.g., configuration terminator 7904 for configuration and exception handling controller 7902 or configuration terminator 7908 for configuration and exception handling controller 7906) in FIG. 79 which asserts that it is configured (or reconfigures) may be included at the end of a chain.

FIG. 80 illustrates a reconfiguration circuit 8018 according to embodiments of the disclosure. Reconfiguration circuit 8018 includes a configuration state register 8020 to store the configuration state (or a pointer thereto).

7.4 Hardware for Fabric-Initiated Reconfiguration of a CSA

Some portions of an application targeting a CSA (e.g., spatial array) may be run infrequently or may be mutually exclusive with other parts of the program. To save area, to improve performance, and/or reduce power, it may be useful to time multiplex portions of the spatial fabric among several different parts of the program dataflow graph. Certain embodiments herein include an interface by which a CSA (e.g., via the spatial program) may request that part of the fabric be reprogrammed. This may enable the CSA to dynamically change itself according to dynamic control flow. Certain embodiments herein allow for fabric initiated reconfiguration (e.g., reprogramming). Certain embodiments herein provide for a set of interfaces for triggering configuration from within the fabric. In some embodiments, a PE issues a reconfiguration request based on some decision in the program dataflow graph. This request may travel a network to our new configuration interface, where it triggers reconfiguration. Once reconfiguration is completed, a message may optionally be returned notifying of the completion. Certain embodiments of a CSA thus provide for a program (e.g., dataflow graph) directed reconfiguration capability.

FIG. 81 illustrates an accelerator tile 8100 comprising an array of processing elements and a configuration and exception handling controller 8106 with a reconfiguration circuit 8118 according to embodiments of the disclosure. Here, a portion of the fabric issues a request for (re)configuration to a configuration domain, e.g., of configuration and exception handling controller 8106 and/or reconfiguration circuit 8118. The domain (re)configures itself, and when the request has been satisfied, the configuration and exception handling controller 8106 and/or reconfiguration circuit 8118 issues a response to the fabric, to notify the fabric that (re)configuration is complete. In one embodiment, configuration and exception handling controller 8106 and/or reconfiguration circuit 8118 disables communication during the time that (re)configuration is ongoing, so the program has no consistency issues during operation.

Configuration Modes

Configure-by-address—In this mode, the fabric makes a direct request to load configuration data from a particular address.

Configure-by-reference—In this mode the fabric makes a request to load a new configuration, e.g., by a pre-determined reference ID. This may simplify the determination of the code to load, since the location of the code has been abstracted.

Configuring Multiple Domains

A CSA may include a higher level configuration controller to support a multicast mechanism to cast (e.g., via network indicated by the dotted box) configuration requests to multiple (e.g., distributed or local) configuration controllers. This may enable a single configuration request to be replicated across larger portions of the fabric, e.g., triggering a broad reconfiguration.

6.5 Exception Aggregators

Certain embodiments of a CSA may also experience an exception (e.g., exceptional condition), for example, floating point underflow. When these conditions occur, a special handlers may be invoked to either correct the program or to terminate it. Certain embodiments herein provide for a system-level architecture for handling exceptions in spatial fabrics. Since certain spatial fabrics emphasize area efficiency, embodiments herein minimize total area while providing a general exception mechanism. Certain embodiments herein provide a low area means of signaling exceptional conditions occurring in within a CSA (e.g., a spatial array). Certain embodiments herein provide an interface and signaling protocol for conveying such exceptions, as well as a PE-level exception semantics. Certain embodiments herein are dedicated exception handling capabilities, e.g., and do not require explicit handling by the programmer.

One embodiments of a CSA exception architecture consists of four portions, e.g., shown in FIGS. 82-83. These portions may be arranged in a hierarchy, in which exceptions flow from the producer, and eventually up to the tile-level exception aggregator (e.g., handler), which may rendezvous with an exception servicing, e.g., of a core. The four portions may be:

1. PE Exception Generator
2. Local Exception Network
3. Mezzanine Exception Aggregator
4. Tile-Level Exception Aggregator

FIG. 82 illustrates an accelerator tile 8200 comprising an array of processing elements and a mezzanine exception aggregator 8204 coupled to a tile-level exception aggregator 8202 according to embodiments of the disclo-
sure. FIG. 83 illustrates a processing element 8300 with an exception generator 8344 according to embodiments of the disclosure.

PE Exception Generator

[0506] Processing element 8300 may include processing element 900 from FIG. 9, for example, with similar numbers being similar components, e.g., local network 902 and local network 8302. Additional network 8313 (e.g., channel) may be an exception network. A PE may implement an interface to an exception network (e.g., exception network 8313 (e.g., channel) on FIG. 83). For example, FIG. 83 shows the microarchitecture of such an interface, wherein the PE has an exception generator 8344 (e.g., initiate an exception finite state machine (FSM) 8340 to strobe an exception packet (e.g., BOXID 8342) out on to the exception network. BOXID 8342 may be a unique identifier for an exception producing entity (e.g., a PE or box) within a local exception network. When an exception is detected, exception generator 8344 senses the exception network and strobes out the BOXID when the network is found to be free. Exceptions may be caused by many conditions, for example, but not limited to, arithmetic error, failed ECC check on state, etc. however, it may also be that an exception data flow operation is introduced, with the idea of support constructs like breakpoints.

[0507] The initiation of the exception may either occur explicitly, by the execution of a programmer supplied instruction, or implicitly when a hardware error condition (e.g., a floating point underflow) is detected. Upon an exception, the PE 8300 may enter a waiting state, in which it waits to be serviced by the event exception handler, e.g., external to the PE 8300. The contents of the exception packet depend on the implementation of the particular PE, as described below.

Local Exception Network

[0508] A (e.g., local) exception network steers exception packets from PE 8300 to the mezzanine exception network. Exception network (e.g., 8313) may be a serial, packet switched network consisting of (a e.g., single) control wire and one or more data wires, e.g., organized in a ring or tree topology, e.g., for a subset of PEs. Each PE may have a (e.g., ring) stop in the (e.g., local) exception network, e.g., where it can arbitrate to inject messages into the exception network.

[0509] PE endpoints needing to inject an exception packet may observe their local exception network egress point. If the control signal indicates busy, the PE is to wait to commence inject its packet. If the network is not busy, that is, the downstream stop has no packet to forward, then the PE will proceed commence injection.

[0510] Network packets may be of variable or fixed length. Each packet may begin with a fixed length header field identifying the source PE of the packet. This may be followed by a variable number of PE-specific field containing information, for example, including error codes, data values, or other useful status information.

Mezzanine Exception Aggregator

[0511] The mezzanine exception aggregator 8204 is responsible for assembling local exception network into larger packets and sending them to the tile-level exception aggregator 8202. The mezzanine exception aggregator 8204 may pre-pend the local exception packet with its own unique ID, e.g., ensuring that exception messages are unambiguous. The mezzanine exception aggregator 8204 may interface to a special exception-only virtual channel in the mezzanine network, e.g., ensuring the deadlock-free operation of exceptions.

[0512] The mezzanine exception aggregator 8204 may also be able to directly service certain classes of exception. For example, a configuration request from the fabric may be served out of the mezzanine network using caches local to the mezzanine network stop.

Tile-Level Exception Aggregator

[0513] The final stage of the exception system is the tile-level exception aggregator 8202. The tile-level exception aggregator 8202 is responsible for collecting exceptions from the various mezzanine-level exception aggregators (e.g., 8204) and forwarding them to the appropriate servicing hardware (e.g., core). As such, the tile-level exception aggregator 8202 may include some internal tables and controller to associate particular messages with handler routines. These tables may be indexed either directly or with a small state machine in order to steer particular exceptions.

[0514] Like the mezzanine exception aggregator, the tile-level exception aggregator may service some exception requests. For example, it may initiate the reprogramming of a large portion of the PE fabric in response to a specific exception.

6.6 Extraction Controllers

[0515] Certain embodiments of a CSA include an extraction controller(s) to extract data from the fabric. The below discusses embodiments of how to achieve this extraction quickly and how to minimize the resource overhead of data extraction. Data extraction may be utilized for such critical tasks as exception handling and context switching. Certain embodiments herein extract data from a heterogeneous spatial fabric by introducing features that allow extractable fabric elements (EFEs) (for example, PEs, network controllers, and/or switches) with variable and dynamically variable amounts of state to be extracted.

[0516] Embodiments of a CSA include a distributed data extraction protocol and microarchitecture to support this protocol. Certain embodiments of a CSA include multiple local extraction controllers (LECs) which stream program data out of their local region of the spatial fabric using a combination of (a e.g., small) set of control signals and the fabric-provided network. State elements may be used at each extractable fabric element (EFE) to form extraction chains, e.g., allowing individual EFEs to self-extract without global addressing.

[0517] Embodiments of a CSA do not use a local network to extract program data. Embodiments of a CSA include specific hardware support (e.g., an extraction controller) for the formation of extraction chains, for example, and do not rely on software to establish these chains dynamically, e.g., at the cost of increasing extraction time. Embodiments of a CSA are not purely packet switched and do include extra out-of-band control wires (e.g., control is not sent through the data path requiring extra cycles to strobe and re-serialize this information). Embodiments of a CSA decrease extraction latency by fixing the extraction ordering and by pro-
viding explicit out-of-band control (e.g., by at least a factor of two), while not significantly increasing network complexity.

[0518] Embodiments of a CSA do not use a serial mechanism for data extraction, in which data is streamed bit by bit from the fabric using a JTAG-like protocol. Embodiments of a CSA utilize a coarse-grained fabric approach. In certain embodiments, adding a few control wires or state elements to a 64 or 32-bit-oriented CSA fabric has a lower cost relative to adding those same control mechanisms to a 4 or 6 bit fabric.

[0519] FIG. 84 illustrates an accelerator tile 8400 comprising an array of processing elements and a local extraction controller 8402, 8406 according to embodiments of the disclosure. Each PE, each network controller, and each switch may be an extractable fabric elements (EFFs), e.g., which are configured (e.g., programmed) by embodiments of the CSA architecture.

[0520] Embodiments of a CSA include hardware that provides for efficient, distributed, low-latency extraction from a heterogeneous spatial fabric. This may be achieved according to four techniques. First, a hardware entity, the local extraction controller (LEC) is utilized, for example, as in FIGS. 84-86. A LEC may accept commands from a host (for example, a processor core), e.g., extracting a stream of data from the spatial array, and writing this data back to virtual memory for inspection by the host. Second, a extraction data path may be included, e.g., that is as wide as the native width of the PE fabric and which may be overlaid on top of the PE fabric. Third, new control signals may be received into the PE fabric which orchestrate the extraction process. Fourth, state elements may be located (e.g., in a register) at each configurable endpoint which track the status of adjacent EFFs, allowing each EFE to unambiguously export its state without extra control signals. These four microarchitectural features may allow a CSA to extract data from chains of EFFs. To obtain low data extraction latency, certain embodiments may partition the extraction problem by including multiple (e.g., many) LECs and EFE chains in the fabric. At extraction time, these chains may operate independently to extract data from the fabric in parallel, e.g., dramatically reducing latency. As a result of these combinations, a CSA may perform a complete state dump (e.g., in hundreds of nanoseconds).

[0521] FIGS. 85A-85C illustrate a local extraction controller 8502 configuring a data path network according to embodiments of the disclosure. Depicted network includes a plurality of multiplexers (e.g., multiplexers 8506, 8508, 8510) that may be configured (e.g., via their respective control signals) to connect one or more data paths (e.g., from PEs) together. FIG. 85A illustrates the network 8500 (e.g., fabric) configured (e.g., set) for some previous operation. FIG. 85B illustrates the local extraction controller 8502 (e.g., including a network interface circuit 8504 to send and/or receive signals) strobing an extraction signal and all PEs controlled by the LEC enter into extraction mode. The last PE in the extraction chain (or an extraction terminator) may master the extraction channels (e.g., bus) and being sending data according to either (1) signals from the LEC or (2) internally produced signals (e.g., from a PE). Once completed, a PE may set its completion flag, e.g., enabling the next PE to extract its data. FIG. 85C illustrates the most distant PE has completed the extraction process and as a result it has set its extraction state bit or bits, e.g., which swinging the muxes into the adjacent network to enable the next PE to begin the extraction process. The extracted PE may resume normal operation. In some embodiments, the PE may remain disabled until other action is taken. In these figures, the multiplexer networks are analogues of the “Switch” shown in certain Figures (e.g., FIG. 5).

[0522] The following sections describe the operation of the various components of embodiments of an extraction network.

Local Extraction Controller

[0523] FIG. 86 illustrates an extraction controller 8602 according to embodiments of the disclosure. A local extraction controller (LEC) may be the hardware entity which is responsible for accepting extraction commands, coordinating the extraction process with the EFFs, and/or storing extracted data, e.g., to virtual memory. In this capacity, the LEC may be a special-purpose, sequential microcontroller.

[0524] LEC operation may begin when it receives a pointer to a buffer (e.g., in virtual memory) where fabric state will be written, and, optionally, a command controlling how much of the fabric will be extracted. Depending on the LEC microarchitecture, this pointer (e.g., stored in pointer register 8604) may come either over a network or through a memory system access to the LEC. When it receives such a pointer (e.g., command), the LEC proceeds to extract state from the portion of the fabric for which it is responsible. The LEC may stream this extracted data out of the fabric into the buffer provided by the external caller.

[0525] Two different microarchitectures for the LEC are shown in FIG. 84. The first places the LEC 8402 at the memory interface. In this case, the LEC may make direct requests to the memory system to write extracted data. In the second case the LEC 8406 is placed on a memory network, in which it may make requests to the memory only indirectly. In both cases, the logical operation of the LEC may be unchanged. In one embodiment, LECs are informed of the desire to extract data from the fabric, for example, by a set of (e.g., OS-visible) control-status registers which will be used to inform individual LECs of new commands.

Extra Out-of-Band Control Channels (e.g., Wires)

[0526] In certain embodiments, extraction relies on 2-8 extra, out-of-band signals to improve configuration speed, as defined below. Signals driven by the LEC may be labelled LEC. Signals driven by the EFE (e.g., PE) may be labelled EFE. Configuration controller 8602 may include the following control channels, e.g., LEC_EXTRACT control channel 8606, LEC_START control channel 8608, LEC_STROBE control channel 8610, and EFE_COMPLETE control channel 8612, with examples of each discussed in Table 3 below.

<table>
<thead>
<tr>
<th>Extraction Channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEC_EXTRACT</td>
</tr>
<tr>
<td>LEC_START</td>
</tr>
<tr>
<td>LEC_STROBE</td>
</tr>
</tbody>
</table>
TABLE 3-continued

<table>
<thead>
<tr>
<th>Extraction Channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFE_COMPLETE</td>
</tr>
<tr>
<td>Optional signal strobed when EFE has completed dumping state. This helps LEC identify the completion of individual EFE dumps.</td>
</tr>
</tbody>
</table>

[0527] Generally, the handling of extraction may be left to the implementer of a particular EFE. For example, selectable function EFE may have a provision for dumping registers using an existing data path, while a fixed function EFE might simply have a multiplexer.

[0528] Due to long wire delays when programming a large set of EFFs, the LEC_STROBE signal may be treated as a clock/latch enable for EFE components. Since this signal is used as a clock, in one embodiment the duty cycle of the line is at most 50%. As a result, extraction throughput is approximately halved. Optionally, a second LEC_STROBE signal may be added to enable continuous extraction.

[0529] In one embodiment, only LEC_START is strictly communicated on an independent coupling (e.g., wire), for example, other control channels may be overlayed on existing network (e.g., wires).

Reusing Network Resources

[0530] To reduce the overhead of data extraction, certain embodiments of a LEC may use the existing network infrastructure to communicate extraction data. A LEC may make use of both a chip-level memory hierarchy and a fabric-level communications networks to move data from the fabric into storage. As a result, in certain embodiments of a CSA, the extraction infrastructure adds no more than 2% to the overall fabric area and power.

[0531] Reuse of network resources in certain embodiments of a CSA may cause a network to have some hardware support for an extraction protocol. Circuit switched networks require of certain embodiments of a CSA cause a LEC to set their multiplexors in a specific way for configuration when the ‘LEC_START’ signal is asserted. Packet switched networks do not require configuration, although LEC endpoints (e.g., extraction terminators) use a specific address in the packet switched network. Network reuse is optional, and some embodiments may find dedicated configuration buses to be more convenient.

Per EFE State

[0532] Each EFE may maintain a bit denoting whether or not it has exported its state. This bit may be asserted when the extraction start signal is driven, and then asserted once the particular EFE finished extraction. In one extraction protocol, EFFs are arranged to form chains with the EFE extraction state bit determining the topology of the chain. A LEC may read the extraction state bit of the immediately adjacent EFE. If this adjacent EFE has its extraction bit set and the current EFE does not, the EFE may determine that it owns the extraction bus. When an EFE dumps its last data value, it may drive the ‘FFE DONE’ signal and sets its extraction bit, e.g., enabling upstream EFFs to configure for extraction. The network adjacent to the EFE may observe this signal and also adjust its state to handle the transition. As a base case to the extraction process, an extraction terminator (e.g., extraction terminator 8404 for LEC 8406 or extraction terminator 8408 for LEC 8406 in FIG. 84) which asserts that extraction is complete may be included at the end of a chain.

[0533] Internal to the EFE, this bit may be used to drive flow control ready signals. For example, when the extraction bit is de-asserted, network control signals may automatically be clamped to a values that prevent data from flowing, while, within PEs, no operations or actions will be scheduled.

Dealing with High-Delay Paths

[0534] One embodiment of a LEC may drive a signal over a long distance, e.g., through many multiplexors and with many loads. Thus, it may be difficult for a signal to arrive at a distant EFE within a short clock cycle. In certain embodiments, extraction signals are at some division (e.g., fraction of) of the main (e.g., CSA) clock frequency to ensure digital timing discipline at extraction. Clock division may be utilized in an out-of-band signaling protocol, and does not require any modification of the main clock tree.

Ensuring Consistent Fabric Behavior During Extraction

[0535] Since certain extraction scheme are distributed and have non-deterministic timing due to program and memory effects, different members of the fabric may be under extraction at different times. While LEC_EXTRACT is driven, all network flow control signals may be driven logically low, e.g., thus freezing the operation of a particular segment of the fabric.

[0536] An extraction process may be non-destructive. Therefore a set of PEs may be considered operational once extraction has completed. An extension to an extraction protocol may allow PEs to optionally be disabled post extraction. Alternatively, beginning configuration during the extraction process will have similar effect in embodiments.

Single PE Extraction

[0537] In some cases, it may be expedient to extract a single PE. In this case, an optional address signal may be driven as part of the commencement of the extraction process. This may enable the PE targeted for extraction to be directly enabled. Once the PE has been extracted, the extraction process may cease with the lowering of the LEC_EXTRACT signal. In this way, a single PE may be selectively extracted, e.g., by the local extraction controller.

Handling Extraction Backpressure

[0538] In an embodiment where the LEC writes extracted data to memory (for example, for post-processing, e.g., in software), it may be subject to limited memory bandwidth. In the case that the LEC exhausts its buffering capacity, or expects that it will exhaust its buffering capacity, it may stop strobing the LEC_STROBE signal until the buffering issue has resolved.

[0539] Note that in certain figures (e.g., FIGS. 77, 78, 79, 81, 82, and 84) communications are shown schematically. In certain embodiments, those communications may occur over the (e.g., interconnect) network.

6.7 Flow Diagrams

[0540] FIG. 87 illustrates a flow diagram 8700 according to embodiments of the disclosure. Depicted flow 8700 includes decoding an instruction with a decoder of a core of a processor into a decoded instruction 8702, executing the decoded instruction with an execution unit of the core of the
processor to perform a first operation \(8704\); receiving an input of a dataflow graph comprising a plurality of nodes \(8706\); overlaying the dataflow graph into an array of processing elements of the processor with each node represented as a dataflow operator in the array of processing elements \(8708\); and performing a second operation of the dataflow graph with the array of processing elements when an incoming operand set arrives at the array of processing elements \(8710\).

[0541] FIG. 88 illustrates a flow diagram \(8800\) according to embodiments of the disclosure. Depicted flow \(8800\) includes decoding an instruction with a decoder of a core of a processor into a decoded instruction \(8802\); executing the decoded instruction with an execution unit of the core of the processor to perform a first operation \(8804\); receiving an input of a dataflow graph comprising a plurality of nodes \(8806\); overlaying the dataflow graph into a plurality of processing elements of the processor and an interconnect network between the plurality of processing elements of the processor with each node represented as a dataflow operator in the plurality of processing elements \(8808\); and performing a second operation of the dataflow graph with the interconnect network and the plurality of processing elements when an incoming operand set arrives at the plurality of processing elements \(8810\).

6.8 Memory

[0542] FIG. 89A is a block diagram of a system \(8900\) that employs a memory ordering circuit \(8905\) interposed between a memory subsystem \(8910\) and acceleration hardware \(8902\), according to an embodiment of the present disclosure. The memory subsystem \(8910\) may include known memory components, including cache, memory, and one or more memory controller(s) associated with a processor-based architecture. The acceleration hardware \(8902\) may be coarse-grained spatial architecture made up of lightweight processing elements (or other types of processing components) connected by an inter-processing element (PE) network or another type of inter-component network.

[0543] In one embodiment, programs, viewed as control data flow graphs, are mapped onto the spatial architecture by configuring PEs and a communications network. Generally, PEs are configured as dataflow operators, similar to functional units in a processor: once the input operands arrive at the PE, some operation occurs, and results are forwarded to downstream PEs in a pipelined fashion. Dataflow operators (or other types of operators) may choose to consume incoming data on a per-operator basis. Simple operators, like those handling the unconditional evaluation of arithmetic expressions often consume all incoming data. It is sometimes useful, however, for operators to maintain state, for example, in accumulation.

[0544] The PEs communicate using dedicated virtual circuits, which are formed by statically configuring a circuit-switched communications network. These virtual circuits are flow controlled and fully back pressured, such that PEs will stall if either the source has no data or the destination is full. At runtime, data flows through the PEs implementing a mapped algorithm according to a dataflow graph, also referred to as a subprogram herein. For example, data may be streamed in from memory, through the acceleration hardware \(8902\), and then back out to memory. Such an architecture can achieve remarkable performance efficiency relative to traditional multicore processors: compute, in the form of PEs, is simpler and more numerous than larger cores and communication is direct, as opposed to an extension of the memory subsystem \(8910\). Memory system parallelism, however, helps to support parallel PE computation. If memory accesses are serialized, high parallelism is likely unachievable. To facilitate parallelism of memory accesses, the disclosed memory ordering circuit \(8905\) includes memory ordering architecture and microarchitecture, as will be explained in detail. In one embodiment, the memory ordering circuit \(8905\) is a request address file circuit (or “RAF”) or other memory request circuitry.

[0545] FIG. 89B is a block diagram of the system \(8900\) of FIG. 89A but which employs multiple memory ordering circuits \(8905\), according to an embodiment of the present disclosure. Each memory ordering circuit \(8905\) may function as an interface between the memory subsystem \(8910\) and a portion of the acceleration hardware \(8902\) (e.g., spatial array of processing elements or tile). The memory subsystem \(8910\) may include a plurality of cache slices \(12\) (e.g., cache slices \(12A, 12B, 12C\), and \(12D\) in the embodiment of FIG. 89B), and a certain number of memory ordering circuits \(8905\) (four in this embodiment) may be used for each cache slice \(12\). A crossbar \(8904\) (e.g., RAF circuit) may connect the memory ordering circuits \(8905\) to banks of cache that make up each cache slice \(12A, 12B, 12C\), and \(12D\). For example, there may be eight banks of memory in each cache slice in one embodiment. The system \(8900\) may be instantiated on a single die, for example, as a system on a chip (SoC). In one embodiment, the SoC includes the acceleration hardware \(8902\). In an alternative embodiment, the acceleration hardware \(8902\) is an external programmable chip such as an FPGA or GPGPU, and the memory ordering circuits \(8905\) interface with the acceleration hardware \(8902\) through an input/output hub or the like.

[0546] Each memory ordering circuit \(8905\) may accept read and write requests to the memory subsystem \(8910\). The requests from the acceleration hardware \(8902\) arrive at the memory ordering circuit \(8905\) in a separate channel for each node of the dataflow graph that initiates read or write accesses, also referred to as load or store accesses herein. Buffering is provided so that the processing of loads will return the requested data to the acceleration hardware \(8902\) in the order it was requested. In other words, iteration six data is returned before iteration seven data, and so forth. Furthermore, note that the request channel from a memory ordering circuit \(8905\) to a particular cache bank may be implemented as an ordered channel and any first request that leaves before a second request will arrive at the cache bank before the second request.

[0547] FIG. 90 is a block diagram \(9000\) illustrating general functioning of memory operations into and out of the acceleration hardware \(8902\), according to an embodiment of the present disclosure. The operations occurring out the top of the acceleration hardware \(8902\) are understood to be made to and from a memory of the memory subsystem \(8910\). Note that two load requests are made, followed by corresponding load responses. While the acceleration hardware \(8902\) performs processing on data from the load responses, a third load request and response occur, which trigger additional acceleration hardware processing. The results of the acceleration hardware processing for these three load operations are then passed into a store operation, and thus a final result is stored back to memory.
By considering this sequence of operations, it may be evident that spatial arrays more naturally map to channels. Furthermore, the acceleration hardware 8902 is latency-insensitive in terms of the request and response channels, and inherent parallel processing that may occur. The acceleration hardware may also decouple execution of a program from implementation of the memory subsystem 8910 (FIG. 89A), as interfacing with the memory occurs at discrete moments separate from multiple processing steps taken by the acceleration hardware 8902. For example, the load request to and a load response from memory are separate actions, and may be scheduled differently in different circumstances depending on dependency flow of memory operations. The use of spatial fabric, for example, for processing instructions facilitates spatial separation and distribution of such a load request and a load response.

FIG. 91 is a block diagram 9100 illustrating a spatial dependency flow for a store operation 9101, according to an embodiment of the present disclosure. Reference to a store operation is exemplary, as the same flow may apply to a load operation (but without incoming data), or to other operators such as a fence. A fence is an ordering operation for memory subsystems that ensures that all prior memory operations of a type (such as all stores or all loads) have completed. The store operation 9101 may receive an address 9102 (of memory) and data 9104 received from the acceleration hardware 8902. The store operation 9101 may also receive an incoming dependency token 9108, and in response to the availability of these three items, the store operation 9101 may generate an outgoing dependency token 9112. The incoming dependency token, which may, for example, be an initial dependency token of a program, may be provided in a compiler-supplied configuration for the program, or may be provided by execution of memory-mapped input/output (I/O). Alternatively, if the program has already been running, the incoming dependency token 9108 may be received from the acceleration hardware 8902, e.g., in association with a preceding memory operation from which the store operation 9101 depends. The outgoing dependency token 9112 may be generated based on the address 9102 and data 9104 being required by a program-subsequent memory operation.

FIG. 92 is a detailed block diagram of the memory ordering circuit 8905 of FIG. 89A, according to an embodiment of the present disclosure. The memory ordering circuit 8905 may be coupled to an out-of-order memory subsystem 8910, which as discussed, may include cache 12 and memory 18, and associated out-of-order memory controller (s). The memory ordering circuit 8905 may include, or be coupled to, a communications network interface 20 that may be either an inter-tile or an intra-tile network interface, and may be a circuit switched network interface (as illustrated), and thus include circuit-switched interconnects. Alternatively, or additionally, the communications network interface 20 may include packet-switched interconnects.

The memory ordering circuit 8905 may further include, but not be limited to, a memory interface 9210, an operations queue 9212, input queue(s) 9216, a completion queue 9220, an operation configuration data structure 9224, and an operations manager circuit 9230 that may further include a scheduler circuit 9232 and an execution circuit 9234. In one embodiment, the memory interface 9210 may be circuit-switched, and in another embodiment, the memory interface 9210 may be packet-switched, or both may exist simultaneously. The operations queue 9212 may buffer memory operations (with corresponding arguments) that are being processed for request, and may, therefore, correspond to addresses and data coming into the input queues 9216.

More specifically, the input queues 9216 may be an aggregation of at least the following: a load address queue, a store address queue, a store data queue, and a dependency queue. When implementing the input queue 9216 as aggregated, the memory ordering circuit 8905 may provide for sharing of logical queues, with additional control logic to logically separate the queues, which are individual channels with the memory ordering circuit. This may maximize input queue usage, but may also require additional complexity and space for the logic circuitry to manage the logical separation of the aggregated queue. Alternatively, as will be discussed with reference to FIG. 93, the input queues 9216 may be implemented in a segregated fashion, with a separate hardware queue for each. Whether aggregated (FIG. 92) or disaggregated (FIG. 93), implementation for purposes of this disclosure is substantially the same, with the former using additional logic to logically separate the queues within a single, shared hardware queue.

When shared, the input queues 9216 and the completion queue 9220 may be implemented as ring buffers of a fixed size. A ring buffer is an efficient implementation of a circular queue that has a first-in-first-out (FIFO) data characteristic. These queues may, therefore, enforce a semantical order of a program for which the memory operations are being requested. In one embodiment, a ring buffer (such as for the store address queue) may have entries corresponding to entries flowing through an associated queue (such as the store data queue or the dependency queue) at the same rate. In this way, a store address may remain associated with corresponding store data.

More specifically, the load address queue may buffer an incoming address of the memory 18 from which to retrieve data. The store address queue may buffer an incoming address of the memory 18 to which to write data, which is buffered in the store data queue. The dependency queue may buffer dependency tokens in association with the addresses of the load address queue and the store address queue. Each queue, representing a separate channel, may be implemented with a fixed or dynamic number of entries. When fixed, the more entries that are available, the more efficient complicated loop processing may be made. But, having too many entries costs more area and energy to implement. In some cases, e.g., with the aggregated architecture, the disclosed input queue 9216 may share queue slots. Use of the slots in a queue may be statically allocated.

The completion queue 9220 may be a separate set of queues to buffer data received from memory in response to memory commands issued by load operations. The completion queue 9220 may be used to hold a load operation that has been scheduled but for which data has not yet been received (and thus has not yet completed). The completion queue 9220, may therefore, be used to reorder data and operation flow.

The operations manager circuit 9230, which will be explained in more detail with reference to FIGS. 93 through 98, may provide logic for scheduling and executing queued memory operations when taking into account dependency tokens used to provide correct ordering of the memory operations. The operation manager circuit 9230 may access
the operation configuration data structure 9224 to determine which queues are grouped together to form a given memory operation. For example, the operation configuration data structure 9224 may include that a specific dependency counter (or queue), input queue, output queue, and completion queue are all grouped together for a particular memory operation. As each successive memory operation may be assigned a different group of queues, access to varying queues may be interleaved across a sub-program of memory operations. Knowing all of these queues, the operations manager circuit 9230 may interface with the operations queue 9212, the input queue(s) 9216, the completion queue(s) 9220, and the memory subsystem 8910 to initially issue memory operations to the memory subsystem 8910 when successive memory operations become “executable,” and to next complete the memory operation with some acknowledgement from the memory subsystem. This acknowledgement may be, for example, data in response to a load operation command or an acknowledgement of data being stored in the memory in response to a store operation command.

[0557] FIG. 93 is a flow diagram of a microarchitecture 9300 of the memory ordering circuit 8905 of FIG. 89A, according to an embodiment of the present disclosure. The memory subsystem 8910 may allow illegal execution of a program in which ordering of memory operations is wrong, due to the semantics of C language (and other object-oriented program languages). The microarchitecture 9300 may enforce the ordering of the memory operations (sequences of loads from and stores to memory) so that results of instructions that the accelerator hardware 8902 executes are properly ordered. A number of local networks 50 are illustrated to represent a portion of the accelerator hardware 8902 coupled to the microarchitecture 9300.

[0558] From an architectural perspective, there are at least two goals: first, to run general sequential codes correctly, and second, to obtain high performance in the memory operations performed by the microarchitecture 9300. To ensure program correctness, the compiler expresses the dependency between the store operation and the load operation to an array, p, in some fashion, which are expressed via dependency tokens as will be explained. To improve performance, the microarchitecture 9300 finds issues as many load commands of an array in parallel as is legal with respect to program order.

[0559] In one embodiment, the microarchitecture 9300 may include the operations queue 9212, the input queues 9216, the completion queues 9220, and the operations manager circuit 9230 discussed with reference to FIG. 92, above, where individual queues may be referred to as channels. The microarchitecture 9300 may further include a plurality of dependency token counters 9314 (e.g., one per input queue), a set of dependency queues 9318 (e.g., one each per input queue), an address multiplexer 9332, a store data multiplexer 9334, a completion queue index multiplexer 9336, and a load data multiplexer 9338. The operations manager circuit 9230, in one embodiment, may direct these various multiplexers in generating a memory command 9350 (to be sent to the memory subsystem 8910) and in receipt of responses of load commands back from the memory subsystem 8910, as will be explained.

[0560] The input queues 9216, as mentioned, may include a load address queue 9322, a store address queue 9324, and a store data queue 9326. (The small numbers 0, 1, 2 are channel labels and will be referred to later in FIG. 96 and FIG. 99A.) In various embodiments, these input queues may be multiplied to contain additional channels, to handle additional parallelization of memory operation processing. Each dependency queue 9318 may be associated with one of the input queues 9216. More specifically, the dependency queue 9318 labeled B0 may be associated with the load address queue 9322 and the dependency queue labeled B1 may be associated with the store address queue 9324. If additional channels of the input queues 9216 are provided, the dependency queues 9318 may include additional, corresponding channels.

[0561] In one embodiment, the completion queues 9220 may include a set of output buffers 9344 and 9346 for receipt of load data from the memory subsystem 8910 and a completion queue 9342 to buffer addresses and data for load operations according to an index maintained by the operations manager circuit 9230. The operations manager circuit 9230 can manage the index to ensure in-order execution of the load operations, and to identify data received into the output buffers 9344 and 9346 that may be moved to scheduled load operations in the completion queue 9342.

[0562] More specifically, because the memory subsystem 8910 is out of order, but the accelerator hardware 8902 completes operations in order, the microarchitecture 9300 may re-order memory operations with use of the completion queue 9342. Three different sub-operations may be performed in relation to the completion queue 9342, namely to allocate, enqueue, and dequeue. For allocation, the operations manager circuit 9230 may allocate an index into the completion queue 9342 in an in-order next slot of the completion queue. The operations manager circuit may provide this index to the memory subsystem 8910, which may then know the slot to which to write data for a load operation. To enqueue, the memory subsystem 8910 may write data as an entry to the indexed, in-order next slot in the completion queue 9342 like random access memory (RAM), setting a status bit of the entry to valid. To dequeue, the operations manager circuit 9230 may present the data stored in this in-order next slot to complete the load operation, setting the status bit of the entry to invalid. Invalid entries may then be available for a new allocation.

[0563] In one embodiment, the status signals 9248 may refer to statuses of the input queues 9216, the completion queues 9220, the dependency queues 9318, and the dependency token counters 9314. These statuses, for example, may include an input status, an output status, and a control status, which may refer to the presence or absence of a dependency token in association with an input or an output. The input status may include the presence or absence of addresses and the output status may include the presence or absence of store values and available completion buffer slots. The dependency token counters 9314 may be a compact representation of a queue and track a number of dependency tokens used for any given input queue. If the dependency token counters 9314 saturate, no additional dependency tokens may be generated for new memory operations. Accordingly, the memory ordering circuit 8905 may stall scheduling new memory operations until the dependency token counters 9314 becomes unsaturated.

[0564] With additional reference to FIG. 94, FIG. 94 is a block diagram of an executable determiner circuit 9400, according to an embodiment of the present disclosure. The
memory ordering circuit 8905 may be set up with several different kinds of memory operations, for example a load and a store:

[0565]  ldNo[x,d] result.outN, addr.in64, order.in0, order.out0
[0566]  stNo[x,d] addr.in64, data.inN, order.in0, order.out0
[0567]  The executable determinant circuit 9400 may be integrated as a part of the scheduler circuit 9232 and which may perform a logical operation to determine whether a given memory operation is executable, and thus ready to be issued to memory. A memory operation may be executed when the queues corresponding to its memory arguments have data and an associated dependency token is present. These memory arguments may include, for example, an input queue identifier 9410 (indicative of a channel of the input queue 9216), an output queue identifier 9420 (indicative of a channel of the completion queues 9220), a dependency queue identifier 9430 (e.g., what dependency queue or counter should be referenced), and an operation type indicator 9440 (e.g., load operation or store operation). A field (e.g., of a memory request) may be included, e.g., in the above format, that stores a bit or bits to indicate to use the hazard checking hardware.

[0568]  These memory arguments may be queued within the operations queue 9212, and used to schedule issuance of memory operations in association with incoming addresses and data from memory and the acceleration hardware 8902. (See FIG. 95.) Incoming status signals 9248 may be logically combined with these identifiers and then the results may be added (e.g., through an AND gate 9450) to output an executable signal, e.g., which is asserted when the memory operation is executable. The incoming status signals 9248 may include an input status 9412 for the input queue identifier 9410, an output status 9422 for the output queue identifier 9420, and a control status 9432 (related to dependency tokens) for the dependency queue identifier 9430.

[0569]  For a load operation, and by way of example, the memory ordering circuit 8905 may issue a load command when the load operation has an address (input status) and room to buffer the load result in the completion queue 9342 (output status). Similarly, the memory ordering circuit 8905 may issue a store command for a store operation when the store operation has both an address and data value (input status). Accordingly, the status signals 9248 may communicate a level of emptiness (or fullness) of the queues to which the status signals pertain. The operation type may then dictate whether the logic results in an executable signal depending on what address and data should be available.

[0570]  To implement dependency ordering, the scheduler circuit 9232 may extend memory operations to include dependency tokens as underlined above in the example load and store operations. The control status 9432 may indicate whether a dependency token is available within the dependency queue identified by the dependency queue identifier 9430, which could be one of the dependency queues 9318 (for an incoming memory operation) or a dependency token counter 9314 (for a completed memory operation). Under this formulation, a dependent memory operation requires an additional ordering token to execute and generates an additional ordering token upon completion of the memory operation, where completion means that data from the result of the memory operation has become available to program-subsequent memory operations.

[0571]  In one embodiment, with further reference to FIG. 93, the operations manager circuit 9230 may direct the address multiplexer 9332 to select an address argument that is buffered within either the load address queue 9322 or the store address queue 9324, depending on whether a load operation or a store operation is currently being scheduled for execution. If it is a store operation, the operations manager circuit 9230 may also direct the store data multiplexer 9334 to select corresponding data from the store data queue 9326. The operations manager circuit 9230 may also direct the completion queue index multiplexer 9336 to retrieve a load operation entry, indexed according to queue status and/or program order, within the completion queues 9220, to complete a load operation. The operations manager circuit 9230 may also direct the load data multiplexer 9336 to select data received from the memory subsystem 8910 into the completion queues 9220 for a load operation that is awaiting completion. In this way, the operations manager circuit 9230 may direct selection of inputs that go into forming the memory command 9350, e.g., a load command or a store command, or that the execution circuit 9234 is waiting for to complete a memory operation.

[0572]  FIG. 95 is a block diagram the execution circuit 9234 that may include a priority encoder 9506 and selection circuitry 9508 and which generates output control line(s) 9510, according to one embodiment of the present disclosure. In one embodiment, the execution circuit 9234 may access queued memory operations (in the operations queue 9212) that have been determined to be executable (FIG. 94). The execution circuit 9234 may also receive the schedules 9504A, 9504B, 9504C for multiple of the queued memory operations that have been queued and also indicated as ready to issue to memory. The priority encoder 9506 may thus receive an identity of the executable memory operations that have been scheduled and execute certain rules (or follow particular logic) to select the memory operation from those coming in that has priority to be executed first. The priority encoder 9506 may output a selector signal 9507 that identifies the scheduled memory operation that has a highest priority, and has thus been selected.

[0573]  The priority encoder 9506, for example, may be a circuit (such as a state machine or a simpler converter) that compresses multiple binary inputs into a smaller number of outputs, including possibly just one output. The output of a priority encoder is the binary representation of the original number starting from zero of the most significant input bit. So, in one example, when memory operation 0 ("zero"), memory operation one ("1"), and memory operation two ("2") are executable and scheduled, corresponding to 9504A, 9504B, and 9504C, respectively. The priority encoder 9506 may be configured to output the selector signal 9507 to the selection circuitry 9508 indicating the memory operation zero as the memory operation that has highest priority. The selection circuitry 9508 may be a multiplexer in one embodiment, and be configured to output its selection (e.g., of memory operation zero) onto the control lines 9510, as a control signal, in response to the selector signal from the priority encoder 9506 (and indicative of selection of memory operation of highest priority). This control signal may go to the multiplexers 9332, 9334, 9336, and/or 9338, as discussed with reference to FIG. 93, to populate the memory command 9350 that is next to issue (be sent) to the memory subsystem 8910. The transmission of the memory
command may be understood to be issuance of a memory operation to the memory subsystem 8910.

[0574] FIG. 96 is a block diagram of an exemplary load operation 9600, both logical and in binary form, according to an embodiment of the present disclosure. Referring back to FIG. 94, the logical representation of the load operation 9600 may include channel zero (“0”) (corresponding to the load address queue 9322) as the input queue identifier 9410 and completion channel one (“1”) (corresponding to the output buffer 9344) as the output queue identifier 9420. The dependency queue identifier 9430 may include two identifiers, channel 0O (corresponding to the first of the dependency queues 9318) for incoming dependency tokens and counter CO for outgoing dependency tokens. The operation type 9440 has an indication of “Load,” which could be a numerical indicator as well, to indicate the memory operation is a load operation. Below the logical representation of the logical memory operation is a binary representation for exemplary purposes, e.g., where a load is indicated by “00.” The load operation of FIG. 96 may be extended to include other configurations such as a store operation (FIG. 98A) or other type of memory operations, such as a fence.

[0575] An example of memory ordering by the memory ordering circuit 8905 will be illustrated with a simplified example for purposes of explanation with relation to FIGS. 97A-97B, 98A-98B, and 99A-99G. For this example, the following code includes an array, p, which is accessed by indices i and i+2:

```c
for(i) {
    temp = p[i];
    p[i+2] = temp;
}
```

[0576] Assume, for this example, that array p contains 0, 1, 2, 3, 4, 5, 6, and at the end of loop execution, array p will contain 0, 1, 0, 1, 0, 1, 0, 0. This code may be transformed by unrolling the loop, as illustrated in FIGS. 97A and 97B. True address dependencies are annotated by arrows in FIG. 97A, which in each case, a load operation is dependent on a store operation to the same address. For example, for the first of such dependencies, a store (e.g., a write) to p[2] needs to occur before a load (e.g., a read) from p[2], and second of such dependencies, a store to p[3] needs to occur before a load from p[3], and so forth. As a compiler is to be pessimistic, the compiler annotates dependencies between two memory operations, load p[i] and store p[i+2]. Note that only sometimes do reads and writes conflict. The microarchitecture 9300 is designed to extract memory-level parallelism where memory operations may move forward at the same time when there are no conflicts to the same address. This is especially the case for load operations, which expose latency in code execution due to waiting for preceding dependent store operations to complete. In the example code in FIG. 97B, safe reorderings are noted by the arrows on the left of the unfolded code.

[0577] The way the microarchitecture may perform this reordering is discussed with reference to FIGS. 98A-98B and 99A-99G. Note that this approach is not as optimal as possible because the microarchitecture 9300 may not send a memory command to memory every cycle. However, with minimal hardware, the microarchitecture supports dependency flows by executing memory operations when operands (e.g., address and data, for a store, or address for a load) and dependency tokens are available.

[0578] FIG. 98A is a block diagram of exemplar memory arguments for a load operation 9802 and for a store operation 9804, according to an embodiment of the present disclosure. These, or similar, memory arguments were discussed in relation to FIG. 96 and will not be repeated here. Note, however, that the store operation 9804 has no indicator for the output queue identifier because no data is being output to the acceleration hardware 8902. Instead, the store address in channel 1 and the data in channel 2 of the input queues 9216, as identified in the input queue identifier memory argument, are to be scheduled for transmission to the memory subsystem 8910 in a memory command to complete the store operation 9804. Furthermore, the input channels and output channels of the dependency queues are both implemented with counters. Because the load operations and the store operations as displayed in FIGS. 97A and 97B are interdependent, the counters may be cycled between the load operations and the store operations within the flow of the code.

[0579] FIG. 98B is a block diagram illustrating flow of the load operations and store operations, such as the load operation 9802 and the store operation 8904 of FIG. 98A, through the microarchitecture 9300 of the memory ordering circuit of FIG. 93, according to an embodiment of the present disclosure. For simplicity of explanation, not all of the components are displayed, but reference may be made back to the additional components displayed in FIG. 93. Various ovals indicating “Load” for the load operation 9802 and “Store” for the store operation 9804 are overlaid on some of the components of the microarchitecture 9300 as indication of how various channels of the queues are being used as the memory operations are queued and ordered through the microarchitecture 9300.

[0580] FIGS. 99A, 99B, 99C, 99D, 99E, 99F, 99G, and 99H are block diagrams illustrating functional flow of load operations and store operations for the exemplary program of FIGS. 97A and 97B through queues of the microarchitecture of FIG. 98B, according to an embodiment of the present disclosure. Each figure may correspond to a next cycle of processing by the microarchitecture 9300. Values that are italicized are incoming values (into the queues) and values that are bolded are outgoing values (out of the queues). All other values with normal fonts are retained values already existing in the queues.

[0581] In FIG. 99A, the address p[0] is incoming into the load address queue 9322, and the address p[2] is incoming into the store address queue 9324, starting the control flow process. Note that counter CO, for dependency input for the load address queue, is “1” and counter CI, for dependency output, is zero. In contrast, the “1” of CI indicates a dependency output value for the store operation. This indicates an incoming dependency for the load operation of p[0] and an outgoing dependency for the store operation of p[2]. These values, however, are not yet active, but will become active, in this way, in FIG. 99B.

[0582] In FIG. 99B, address p[0] is bolded to indicate it is outgoing in this cycle. A new address p[1] is incoming into the load address queue and a new address p[3] is incoming into the store address queue. A zero (“0”)-valued bit in the completion queue 9342 is also incoming, which indicates any data present for that indexed entry is invalid. As
mentioned, the values for the counters C0 and C1 are now indicated as incoming, and are thus now active this cycle.

[0583] In FIG. 99C, the outgoing address p[0] has now left the load address queue and a new address p[2] is incoming into the load address queue. And, the data ("0") is incoming into the completion queue for address p[0]. The validity bit is set to "1" to indicate that the data in the completion queue is valid. Furthermore, a new address p[4] is incoming into the store address queue. The value for counter C0 is indicated as outgoing and the value for counter C1 is indicated as incoming. The value of "1" for C1 indicates an incoming dependency for store operation to address p[4].

[0584] Note that the address p[2] for the newest load operation is dependent on the value that first needs to be stored by the store operation for address p[2], which is at the top of the store address queue. Later, the indexed entry in the completion queue for the load operation from address p[2] may remain buffered until the data from the store operation to the address p[2] is completed (see FIGS. 99F-99I).

[0585] In FIG. 99D, the data ("0") is outgoing from the completion queue for address p[0], which is therefore being sent out to the acceleration hardware 8902. Furthermore, a new address p[3] is incoming into the load address queue and a new address p[5] is incoming into the store address queue. The values for the counters C0 and C1 remain unchanged.

[0586] In FIG. 99E, the value ("0") for the address p[2] is incoming into the store data queue, while a new address p[4] comes into the load address queue and a new address p[6] comes into the store address queue. The counter values for C0 and C1 remain unchanged.

[0587] In FIG. 99F, the value ("0") for the address p[2] in the store data queue, and the address p[2] in the store address queue are both outgoing values. Likewise, the value for the counter C1 is indicated as outgoing, while the value ("0") for counter C0 remain unchanged. Furthermore, a new address p[5] is incoming into the load address queue and a new address p[7] is incoming into the store address queue.

[0588] In FIG. 99G, the value ("0") is incoming to indicate the indexed value within the completion queue 9342 is invalid. The address p[1] is bolded to indicate that it is outgoing from the load address queue while a new address p[6] is incoming into the load address queue. A new address p[8] is also incoming into the store address queue. The value of counter C0 is incoming as a "1," corresponding to an incoming dependency for the load operation of address p[6] and an outgoing dependency for the store operation of address p[8]. The value of counter C1 is now "0," and is indicated as outgoing.

[0589] In FIG. 99H, a data value of "1" is incoming into the completion queue 9342 while the validity bit is also incoming as a "1," meaning that the buffered data is valid. This is the data needed to complete the load operation for p[2]. Recall that this data had to first be stored to address p[2], which happened in FIG. 99F. The value of "0" for counter C0 is outgoing, and a value of "1," for counter C1 is incoming. Furthermore, a new address p[7] is incoming into the load address queue and a new address p[9] is incoming into the store address queue.

[0590] In the present embodiment, the process of executing the code of FIGS. 97A and 97B may continue on with bouncing dependency tokens between "0" and "1" for the load operations and the store operations. This is due to the tight dependencies between p[1] and p[1+2]. Other code with less frequent dependencies may generate dependency tokens at a slower rate, and thus reset the counters C0 and C1 at a slower rate, causing the generation of tokens of higher values (corresponding to further semantically-separated memory operations).

[0591] FIG. 100 is a flow chart of a method 10000 for ordering memory operations between acceleration hardware and an out-of-order memory subsystem, according to an embodiment of the present disclosure. The method 10000 may be performed by a system that may include hardware (e.g., circuitry, dedicated logic, and/or programmable logic), software (e.g., instructions executable on a computer system to perform hardware simulation), or a combination thereof. In an illustrative example, the method 10000 may be performed by the memory ordering circuit 8905 and various subcomponents of the memory ordering circuit 8905.

[0592] More specifically, referring to FIG. 100, the method 10000 may start with the memory ordering circuit queuing memory operations in an operations queue of the memory ordering circuit (10010). Memory operation and control arguments may make up the memory operations, as queued, where the memory operation and control arguments are mapped to certain queues within the memory ordering circuit as discussed previously. The memory ordering circuit may work to issue the memory operations to a memory in association with acceleration hardware, to ensure the memory operations complete in program order. The method 10000 may continue with the memory ordering circuit receiving, in set of input queues, from the acceleration hardware an address of the memory associated with a second memory operation of the memory operations (10020). In one embodiment, a load address queue of the set of input queues is the channel to receive the address. In another embodiment, a store address queue of the set of input queues is the channel to receive the address. The method 10000 may continue with the memory ordering circuit receiving, from the acceleration hardware, a dependency token associated with the address, wherein the dependency token indicates a dependency on data generated by a first memory operation, of the memory operations, which precedes the second memory operation (10030). In one embodiment, a channel of a dependency queue is a receiving the dependency token. The first memory operation may be either a load operation or a store operation.

[0593] The method 10000 may continue with the memory ordering circuit scheduling issuance of the second memory operation to the memory in response to receiving the dependency token and the address associated with the dependency token (10040). For example, when the load address queue receives the address for an address argument of a load operation and the dependency queue receives the dependency token for a control argument of the load operation, the memory ordering circuit may schedule issuance of the second memory operation as a load operation. The method 10000 may continue with the memory ordering circuit issuing the second memory operation (e.g., in a command) to the memory in response to completion of the first memory operation (10050). For example, if the first memory operation is a store, completion may be verified by acknowledgement that the data in a store data queue of the set of input queues has been written to the address in the memory. Similarly, if the first memory operation is a load operation, completion may be verified by receipt of data from the memory for the load operation.
7. Summary

Supercomputing at the ExaFLOP scale may be a challenge in high-performance computing, a challenge which is not likely to be met by conventional von Neumann architectures. To achieve ExaFLOPs, embodiments of a CSA provide a heterogeneous spatial array that targets direct execution of (e.g., compiler-produced) dataflow graphs. In addition to laying out the architectural principles of embodiments of a CSA, the above also describes and evaluates embodiments of a CSA which showed performance and energy of larger than 10x over existing products. Compiler-generated code may have significant performance and energy gains over roadmap architectures. As a heterogeneous, parametric architecture, embodiments of a CSA may be readily adapted to all computing uses. For example, a mobile version of CSA might be tuned to 32-bits, while a machine-learning focused array might feature significant numbers of vectorized 8-bit multiplication units. The main advantages of embodiments of a CSA are high performance and extreme energy efficiency, characteristics relevant to all forms of computing ranging from supercomputing and datacenter to the internet-of-things.

In one embodiment, a processor includes a spatial array of processing elements; and a packet switched communications network to route data within the spatial array between processing elements according to a dataflow graph to perform a first dataflow operation of the dataflow graph, wherein the packet switched communications network further comprises a plurality of network dataflow endpoint circuits to perform a second dataflow operation of the dataflow graph. A network dataflow endpoint circuit of the plurality of network dataflow endpoint circuits may include a network ingress buffer to receive input data from the packet switched communications network, and a spatial array egress buffer to output resultant data to the spatial array of processing elements according to the second dataflow operation on the input data. The spatial array egress buffer may output the resultant data based on a scheduler within the network dataflow endpoint circuit monitoring the packet switched communications network. The spatial array egress buffer may output the resultant data based on the scheduler within the network dataflow endpoint circuit monitoring a selected channel of multiple network virtual channels of the packet switched communications network. A network dataflow endpoint circuit of the plurality of network dataflow endpoint circuits may include a spatial array ingress buffer to receive control data from the spatial array that causes a network ingress buffer of the network dataflow endpoint circuit that received input data from the packet switched communications network to output resultant data to the spatial array of processing elements according to the second dataflow operation on the input data and the control data. A network dataflow endpoint circuit of the plurality of network dataflow endpoint circuits may stall an output of resultant data of the second dataflow operation from a spatial array egress buffer of the network dataflow endpoint circuit when a backpressure signal from a downstream processing element of the spatial array of processing elements indicates that storage in the downstream processing element is not available for the output of the network dataflow endpoint circuit. A network dataflow endpoint circuit of the plurality of network dataflow endpoint circuits may send a backpressure signal to stall a source from sending input data on the packet switched communications network into a network ingress buffer of the network dataflow endpoint circuit when the network ingress buffer is not available. The spatial array of processing elements may include a plurality of processing elements; and an interconnect network between the plurality of processing elements to receive an input of the dataflow graph comprising a plurality of nodes, wherein the dataflow graph is to be overlaid into the interconnect network, the plurality of processing elements, and the plurality of network dataflow endpoint circuits with each node represented as a dataflow operator in either of the plurality of processing elements and the plurality of network dataflow endpoint circuits, and the plurality of processing elements and the plurality of network dataflow endpoint circuits are to perform an operation by an incoming operand set arriving at each of the dataflow operators of the plurality of processing elements and the plurality of network dataflow endpoint circuits. The spatial array of processing elements may include a circuit switched network to transport the data within the spatial array between processing elements according to the dataflow graph.

In another embodiment, a method includes providing a spatial array of processing elements; routing, with a packet switched communications network, data within the spatial array between processing elements according to a dataflow graph; performing a first dataflow operation of the dataflow graph with the processing elements; and performing a second dataflow operation of the dataflow graph with a plurality of network dataflow endpoint circuits of the packet switched communications network. The performing the second dataflow operation may include receiving input data from the packet switched communications network with a network ingress buffer of a network dataflow endpoint circuit of the plurality of network dataflow endpoint circuits; and outputting resultant data from a spatial array egress buffer of the network dataflow endpoint circuit to the spatial array of processing elements according to the second dataflow operation on the input data. The outputting may include outputting the resultant data based on a scheduler within the network dataflow endpoint circuit monitoring the packet switched communications network. The outputting may include outputting the resultant data based on the scheduler within the network dataflow endpoint circuit monitoring a selected channel of multiple network virtual channels of the packet switched communications network. The performing the second dataflow operation may include receiving control data, with a spatial array ingress buffer of a network dataflow endpoint circuit of the plurality of network dataflow endpoint circuits, from the spatial array, and configuring the network dataflow endpoint circuit to cause a network ingress buffer of the network dataflow endpoint circuit that received input data from the packet switched communications network to output resultant data to spatial array of processing elements according to the second dataflow operation on the input data and the control data. The performance of the second dataflow operation may include stalling an output of the second dataflow operation from a spatial array egress buffer of a network dataflow endpoint circuit of the plurality of network dataflow endpoint circuits when a backpressure signal from a downstream processing element of the spatial array of processing elements indicates that storage in the downstream processing element is not available for the output of the network dataflow endpoint circuit.
circuit of the plurality of network dataflow endpoint circuits to stall a source from sending input data on the packet switched communications network into a network ingress buffer of the network dataflow endpoint circuit when the network ingress buffer is not available. The routing, performing the first dataflow operation, and performing the second dataflow operation may include receiving an input of a dataflow graph comprising a plurality of nodes; overlaying the dataflow graph into the spatial array of processing elements and the plurality of network dataflow endpoint circuits with each node represented as a dataflow operator in either of the processing elements and the plurality of network dataflow endpoint circuits; and performing the first dataflow operation with the processing elements and performing the second dataflow operation with the plurality of network dataflow endpoint circuits when an incoming operand set arrives at each of the dataflow operators of the processing elements and the plurality of network dataflow endpoint circuits. The method may include transporting the data within the spatial array between processing elements according to the dataflow graph with a circuit switched network of the spatial array.

[0597] In yet another embodiment, a non-transitory machine readable medium that stores code that when executed by a machine causes the machine to perform a method including providing a spatial array of processing elements; routing, with a packet switched communications network, data within the spatial array between processing elements according to a dataflow graph; performing a first dataflow operation of the dataflow graph with the processing elements; and performing a second dataflow operation of the dataflow graph with a plurality of network dataflow endpoint circuits of the packet switched communications network. The performing the second dataflow operation may include receiving input data from the packet switched communications network with a network ingress buffer of a network dataflow endpoint circuit of the plurality of network dataflow endpoint circuits; and outputting resultant data from a spatial array egress buffer of the network dataflow endpoint circuit to the spatial array of processing elements according to the second dataflow operation on the input data. The outputting may include outputting the resultant data based on a scheduler within the network dataflow endpoint circuit monitoring the packet switched communications network. The outputting may include outputting the resultant data based on the scheduler within the network dataflow endpoint circuit monitoring a selected channel of multiple network virtual channels of the packet switched communications network. The performing the second dataflow operation may include receiving control data, with a spatial array ingress buffer of a network dataflow endpoint circuit of the plurality of network dataflow endpoint circuits, from the spatial array; and configuring the network dataflow endpoint circuit to cause a network ingress buffer of the network dataflow endpoint circuit that received input data from the packet switched communications network to output resultant data to the spatial array of processing elements according to the second dataflow operation on the input data and the control data. The performing the second dataflow operation may include stalling an output of the second dataflow operation from a spatial array egress buffer of a network dataflow endpoint circuit of the plurality of network dataflow endpoint circuits when a backpressure signal from a downstream processing element of the spatial array of processing elements indicates that storage in the downstream processing element is not available for the output of the network dataflow endpoint circuit. The performing the second dataflow operation may include sending a backpressure signal from a network dataflow endpoint circuit of the plurality of network dataflow endpoint circuits to stall a source from sending input data on the packet switched communications network into a network ingress buffer of the network dataflow endpoint circuit when the network ingress buffer is not available. The routing, performing the first dataflow operation, and performing the second dataflow operation may include receiving an input of a dataflow graph comprising a plurality of nodes; overlaying the dataflow graph into the spatial array of processing elements and the plurality of network dataflow endpoint circuits with each node represented as a dataflow operator in either of the processing elements and the plurality of network dataflow endpoint circuits when an incoming operand set arrives at each of the dataflow operators of the processing elements and the plurality of network dataflow endpoint circuits. The method may include transporting the data within the spatial array between processing elements according to the dataflow graph with a circuit switched network of the spatial array.

[0598] In another embodiment, a processor includes a spatial array of processing elements; and a packet switched communications network to route data within the spatial array between processing elements according to a dataflow graph to perform a first dataflow operation of the dataflow graph, wherein the packet switched communications network further comprises means to perform a second dataflow operation of the dataflow graph.

[0599] In one embodiment, a processor includes a core with a decoder to decode an instruction into a decoded instruction and an execution unit to execute the decoded instruction to perform a first operation; a plurality of processing elements; and an interconnect network between the plurality of processing elements to receive an input of a dataflow graph comprising a plurality of nodes, wherein the dataflow graph is to be overlaid into the interconnect network and the plurality of processing elements with each node represented as a dataflow operator in the plurality of processing elements, and the plurality of processing elements are to perform a second operation by a respective, incoming operand set arriving at each of the dataflow operators of the plurality of processing elements. A processing element of the plurality of processing elements may stall execution when a backpressure signal from a downstream processing element indicates that storage in the downstream processing element is not available for an output of the processing element. The processor may include a flow control path network to carry the backpressure signal according to the dataflow graph. A dataflow token may cause an output from a dataflow operator receiving the dataflow token to be sent to an input buffer of a particular processing element of the plurality of processing elements. The second operation may include a memory access and the plurality of processing elements comprises a memory-accessing dataflow operator that is not to perform the memory access until receiving a memory dependency token from a logically previous dataflow operator. The plurality of processing
elements may include a first type of processing element and a second, different type of processing element.

[0600] In another embodiment, a method includes decoding an instruction with a decoder of a core of a processor into a decoded instruction; executing the decoded instruction with an execution unit of the core of the processor to perform a first operation; receiving an input of a dataflow graph comprising a plurality of nodes; overlaying the dataflow graph into a plurality of processing elements of the processor and an interconnect network between the plurality of processing elements of the processor with each node represented as a dataflow operator in the plurality of processing elements; and performing a second operation of the dataflow graph with the interconnect network and the plurality of processing elements by a respective, incoming operand set arriving at each of the dataflow operators of the plurality of processing elements. The method may include stalling execution by a processing element of the plurality of processing elements when a backpressure signal from a downstream processing element indicates that storage in the downstream processing element is not available for an output of the processing element. The method may include sending the backpressure signal on a flow control path network according to the dataflow graph. A dataflow token may cause an output from a dataflow operator receiving the dataflow token to be sent to an input buffer of a particular processing element of the plurality of processing elements. The method may include not performing a memory access until receiving a memory dependency token from a logically previous dataflow operator, wherein the second operation comprises the memory access and the plurality of processing elements comprises a memory-accessing dataflow operator. The method may include providing a first type of processing element and a second, different type of processing element of the plurality of processing elements.

[0601] In yet another embodiment, an apparatus includes a data path network between a plurality of processing elements; and a flow control path network between the plurality of processing elements, wherein the data path network and the flow control path network are to receive an input of a dataflow graph comprising a plurality of nodes, the dataflow graph is to be overlaid into the data path network, the flow control path network, and the plurality of processing elements with each node represented as a dataflow operator in the plurality of processing elements, and the plurality of processing elements are to perform a second operation by a respective, incoming operand set arriving at each of the dataflow operators of the plurality of processing elements. The flow control path network may carry backpressure signals to a plurality of dataflow operators according to the dataflow graph. A dataflow token sent on the data path network to a dataflow operator may cause an output from the dataflow operator to be sent to an input buffer of a particular processing element of the plurality of processing elements on the data path network. The data path network may include a static, circuit switched network to carry the respective, incoming operands to each of the dataflow operators according to the dataflow graph. The flow control path network may transmit a backpressure signal according to the dataflow graph from a downstream processing element to indicate that storage in the downstream processing element is not available for an output of the processing element. At least one data path of the data path network and at least one flow control path of the flow control path network may form a channelized circuit with backpressure control. The flow control path network may pipeline at least two of the plurality of processing elements in series.

[0602] In another embodiment, a method includes receiving an input of a dataflow graph comprising a plurality of nodes; and overlaying the dataflow graph into a plurality of processing elements of a processor, a data path network between the plurality of processing elements, and a flow control path network between the plurality of processing elements with each node represented as a dataflow operator in the plurality of processing elements. The method may include carrying backpressure signals with the flow control path network to a plurality of dataflow operators according to the dataflow graph. The method may include sending a dataflow token on the data path network to a dataflow operator to cause an output from the dataflow operator to be sent to an input buffer of a particular processing element of the plurality of processing elements on the data path network. The method may include setting a plurality of switches of the data path network and/or a plurality of switches of the flow control path network to carry the respective, incoming operand set to each of the dataflow operators according to the dataflow graph, wherein the data path network is a static, circuit switched network. The method may include transmitting a backpressure signal with the flow control path network according to the dataflow graph from a downstream processing element to indicate that storage in the downstream processing element is not available for an output of the processing element. The method may include forming a channelized circuit with backpressure control with at least one data path of the data path network and at least one flow control path of the flow control path network.

[0603] In yet another embodiment, a processor includes a core with a decoder to decode an instruction into a decoded instruction and an execution unit to execute the decoded instruction to perform a first operation; a plurality of processing elements; and a network means between the plurality of processing elements to receive an input of a dataflow graph comprising a plurality of nodes, wherein the dataflow graph is to be overlaid into the network means and the plurality of processing elements with each node represented as a dataflow operator in the plurality of processing elements, and the plurality of processing elements are to perform a second operation by a respective, incoming operand set arriving at each of the dataflow operators of the plurality of processing elements.

[0604] In another embodiment, an apparatus includes a data path means between a plurality of processing elements; and a flow control path means between the plurality of processing elements, wherein the data path means and the flow control path means are to receive an input of a dataflow graph comprising a plurality of nodes, the dataflow graph is to be overlaid into the data path means, the flow control path means, and the plurality of processing elements with each node represented as a dataflow operator in the plurality of processing elements, and the plurality of processing elements are to perform a second operation by a respective, incoming operand set arriving at each of the dataflow operators of the plurality of processing elements.

[0605] In one embodiment, a processor includes a core with a decoder to decode an instruction into a decoded instruction and an execution unit to execute the decoded instruction to perform a first operation; and an array of processing elements to receive an input of a dataflow graph.
comprising a plurality of nodes, wherein the dataflow graph is to be overlaid into the array of processing elements with each node represented as a dataflow operator in the array of processing elements, and the array of processing elements is to perform a second operation when an incoming operand set arrives at the array of processing elements. The array of processing elements may not perform the second operation until the incoming operand set arrives at the array of processing elements and storage in the array of processing elements is available for output of the second operation. The array of processing elements may include a network (or channel(s)) to carry dataflow tokens and control tokens to a plurality of dataflow operators. The second operation may include a memory access and the array of processing elements may include a memory-accessing dataflow operator that is not to perform the memory access until receiving a memory dependency token from a logically previous dataflow operator. Each processing element may perform only one or two operations of the dataflow graph.

[0606] In another embodiment, a method includes decoding an instruction with a decoder of a core of a processor into a decoded instruction; executing the decoded instruction with an execution unit of the core of the processor to perform a first operation; receiving an input of a dataflow graph comprising a plurality of nodes; overlaying the dataflow graph into an array of processing elements of the processor with each node represented as a dataflow operator in the array of processing elements; and performing a second operation of the dataflow graph with the array of processing elements when an incoming operand set arrives at the array of processing elements. The array of processing elements may not perform the second operation until the incoming operand set arrives at the array of processing elements and storage in the array of processing elements is available for output of the second operation. The array of processing elements may include a network carrying dataflow tokens and control tokens to a plurality of dataflow operators. The second operation may include a memory access and the array of processing elements comprises a memory-accessing dataflow operator that is not to perform the memory access until receiving a memory dependency token from a logically previous dataflow operator. Each processing element may perform only one or two operations of the dataflow graph.

[0608] In another embodiment, a processor includes a core with a decoder to decode an instruction into a decoded instruction and an execution unit to execute the decoded instruction to perform a first operation; and means to receive an input of a dataflow graph comprising a plurality of nodes, wherein the dataflow graph is to be overlaid into the means with each node represented as a dataflow operator in the means, and the means is to perform a second operation when an incoming operand set arrives at the means.

[0609] In one embodiment, a processor includes a core with a decoder to decode an instruction into a decoded instruction and an execution unit to execute the decoded instruction to perform a first operation; a plurality of processing elements; an interconnect network between the plurality of processing elements to receive an input of a dataflow graph comprising a plurality of nodes; and each configuration controller is to load configuration information from storage and cause coupling of the respective subset of the plurality of processing elements according to the configuration information. The processor may include a plurality of configuration caches, and each configuration controller is coupled to a respective configuration cache to fetch the configuration information for the respective subset of the plurality of processing elements. The first operation performed by the execution unit may prefetch configuration information into each of the plurality of configuration caches. Each of the plurality of configuration controllers may include a reconfiguration circuit to cause a reconfiguration for at least one processing element of the respective subset of the plurality of processing elements on receipt of a configuration error message from the at least one processing element. Each of the plurality of configuration controllers may reconfiguration circuit to cause a reconfiguration for the respective subset of the plurality of processing elements on receipt of a reconfiguration request message, and disable communication with the respective subset of the plurality of processing elements until the reconfiguration is complete. The processor may include a plurality of exception aggregators, and each exception aggregator is coupled to a respective subset of the plurality of processing elements to collect exceptions from the respective subset of the plurality of processing elements and forward the exceptions to the core for servicing. The processor may include a plurality of extraction controllers, each extraction controller is coupled to a respective subset of the plurality of processing elements, and each extraction controller is to cause state data from the respective subset of the plurality of processing elements to be saved to memory.
[0610] In another embodiment, a method includes decoding an instruction with a decoder of a core of a processor into a decoded instruction; executing the decoded instruction with an execution unit of the core of the processor to perform a first operation; receiving an input of a dataflow graph comprising a plurality of nodes; overlaying the dataflow graph into a plurality of processing elements of the processor and an interconnect network between the plurality of processing elements of the processor with each node represented as a dataflow operator in the plurality of processing elements; and performing a second operation of the dataflow graph with the interconnect network and the plurality of processing elements when an incoming operand set arrives at the plurality of processing elements. The method may include loading configuration information from storage for respective subsets of the plurality of processing elements and causing coupling for each respective subset of the plurality of processing elements according to the configuration information. The method may include fetching the configuration information for the respective subset of the plurality of processing elements from a respective configuration cache of a plurality of configuration caches. The first operation performed by the execution unit may be prefetching configuration information into each of the plurality of configuration caches. The method may include causing a reconfiguration for at least one processing element of the respective subset of the plurality of processing elements on receipt of a configuration error message from the at least one processing element. The method may include causing a reconfiguration for the respective subset of the plurality of processing elements on receipt of a reconfiguration request message; and disabling communication with the respective subset of the plurality of processing elements until the reconfiguration is complete. The method may include collecting exceptions from a respective subset of the plurality of processing elements; and forwarding the exceptions to the core for servicing. The method may include causing state data from a respective subset of the plurality of processing elements to be saved to memory.

[0612] In another embodiment, a processor includes a core with a decoder to decode an instruction into a decoded instruction and an execution unit to execute the decoded instruction to perform a first operation; a plurality of processing elements; and means between the plurality of processing elements to receive an input of a dataflow graph comprising a plurality of nodes, wherein the dataflow graph is to be overlaid into the m and the plurality of processing elements with each node represented as a dataflow operator in the plurality of processing elements, and the plurality of processing elements to perform a second operation when an incoming operand set arrives at the plurality of processing elements.

[0613] In one embodiment, an apparatus (e.g., a processor) includes: a spatial array of processing elements comprising a communications network to receive an input of a dataflow graph comprising a plurality of nodes, wherein the dataflow graph is to be overlaid into the spatial array of processing elements with each node represented as a dataflow operator in the spatial array of processing elements, and the spatial array of processing elements to perform an operation by a respective, incoming operand set arriving at each of the dataflow operators; a plurality of request address file circuits coupled to the spatial array of processing elements and a cache memory, each request address file circuit of the plurality of request address file circuits to access data in the cache memory in response to a request for data access from the spatial array of processing elements; a plurality of translation lookaside buffers comprising a translation lookaside buffer in each of the plurality of request address file circuits to provide an output of a physical address for an input of a virtual address; and a translation lookaside buffer manager circuit comprising a higher level translation lookaside buffer than the plurality of translation lookaside buffers, the translation lookaside buffer manager circuit to perform a first page walk in the cache memory for a miss of an input of a virtual address into a first translation lookaside buffer and into the higher level translation lookaside buffer to determine a physical address mapped to the virtual address, store a mapping of the virtual address to the physical address from the first page walk in the higher level translation lookaside buffer to cause the higher level translation lookaside buffer to send the physical address to the first translation lookaside buffer in a first request address file circuit. The translation lookaside buffer manager circuit may simultaneously, with the first page walk, perform a second page walk in the cache memory, wherein the second page walk is for
a miss of an input of a virtual address into a second translation lookaside buffer and into the higher level translation lookaside buffer to determine a physical address mapped to the virtual address, store a mapping of the virtual address to the physical address from the second page walk in the higher level translation lookaside buffer to cause the higher level translation lookaside buffer to send the physical address to the second translation lookaside buffer in a second request address file circuit. The receipt of the physical address in the first translation lookaside buffer may cause the first request address file circuit to perform a data access for the request for data access from the spatial array of processing elements on the physical address in the cache memory. The translation lookaside buffer manager circuit may insert an indicator in the higher level translation lookaside buffer for the miss of the input of the virtual address in the first translation lookaside buffer and the higher level translation lookaside buffer to prevent an additional page walk for the input of the virtual address during the first page walk. The translation lookaside buffer manager circuit may receive a shutdown message from a requesting entity for a mapping of a physical address to a virtual address, invalidate the mapping in the higher level translation lookaside buffer, and send shutdown messages to only those of the plurality of request address file circuits that include a copy of the mapping in a respective translation lookaside buffer, wherein each of those of the plurality of request address file circuits are to send an acknowledgement message to the translation lookaside buffer manager circuit, and the translation lookaside buffer manager circuit is to send a shutdown completion acknowledgement message to the requesting entity when all acknowledgement messages are received. The translation lookaside buffer manager circuit may receive a shutdown message from a requesting entity for a mapping of a physical address to a virtual address, invalidate the mapping in the higher level translation lookaside buffer, and send shutdown messages to all of the plurality of request address file circuits, wherein each of the plurality of request address file circuits are to send an acknowledgement message to the translation lookaside buffer manager circuit, and the translation lookaside buffer manager circuit is to send a shutdown completion acknowledgement message to the requesting entity when all acknowledgement messages are received.

In another embodiment, a method includes overlaying an input of a dataflow graph comprising a plurality of nodes into a spatial array of processing elements comprising a communications network with each node represented as a dataflow operator in the spatial array of processing elements; coupling a plurality of request address file circuits to the spatial array of processing elements and a cache memory with each request address file circuit of the plurality of request address file circuits accessing data in the cache memory in response to a request for data access from the spatial array of processing elements; providing an output of a physical address for an input of a virtual address into a translation lookaside buffer of a plurality of translation lookaside buffers comprising a translation lookaside buffer in each of the plurality of request address file circuits; coupling a translation lookaside buffer manager circuit comprising a higher level translation lookaside buffer than the plurality of translation lookaside buffers to the plurality of request address file circuits and the cache memory; and performing a first page walk in the cache memory for a miss of an input of a virtual address into a first translation lookaside buffer and into the higher level translation lookaside buffer with the translation lookaside buffer manager circuit to determine a physical address mapped to the virtual address, store a mapping of the virtual address to the physical address from the first page walk in the higher level translation lookaside buffer to cause the higher level translation lookaside buffer to send the physical address to the first translation lookaside buffer in a first request address file circuit. The method may include simultaneously, with the first page walk, performing a second page walk in the cache memory with the translation lookaside buffer manager circuit, wherein the second page walk is for a miss of an input of a virtual address into a second translation lookaside buffer and into the higher level translation lookaside buffer to determine a physical address mapped to the virtual address, and storing a mapping of the virtual address to the physical address from the second page walk in the higher level translation lookaside buffer to cause the higher level translation lookaside buffer to send the physical address to the second translation lookaside buffer in a second request address file circuit. The method may include causing the first request address file circuit to perform a data access for the request for data access from the spatial array of processing elements on the physical address in the cache memory in response to receipt of the physical address in the first translation lookaside buffer. The method may include inserting, with the translation lookaside buffer manager circuit, an indicator in the higher level translation lookaside buffer for the miss of the input of the virtual address in the first translation lookaside buffer and the higher level translation lookaside buffer to prevent an additional page walk for the input of the virtual address during the first page walk. The method may include receiving, with the translation lookaside buffer manager circuit, a shutdown message from a requesting entity for a mapping of a physical address to a virtual address, invalidating the mapping in the higher level translation lookaside buffer, and sending shutdown messages to only those of the plurality of request address file circuits that include a copy of the mapping in a respective translation lookaside buffer, wherein each of those of the plurality of request address file circuits are to send an acknowledgement message to the translation lookaside buffer manager circuit, and the translation lookaside buffer manager circuit is to send a shutdown completion acknowledgement message to the requesting entity when all acknowledgement messages are received. The method may include receiving, with the translation lookaside buffer manager circuit, a shutdown message from a requesting entity for a mapping of a physical address to a virtual address, invalidating the mapping in the higher level translation lookaside buffer, and sending shutdown messages to all of the plurality of request address file circuits, wherein each of the plurality of request address file circuits are to send an acknowledgement message to the translation lookaside buffer manager circuit, and the translation lookaside buffer manager circuit is to send a shutdown completion acknowledgement message to the requesting entity when all acknowledgement messages are received.

In another embodiment, an apparatus includes a spatial array of processing elements comprising a communications network to receive an input of a dataflow graph comprising a plurality of nodes, wherein the dataflow graph is to be overlaid into the spatial array of processing elements
with each node represented as a dataflow operator in the spatial array of processing elements, and the spatial array of processing elements is to perform an operation by a respective, incoming operand set arriving at each of the dataflow operators; a plurality of request address file circuits coupled to the spatial array of processing elements and a plurality of cache memory banks, each request address file circuit of the plurality of request address file circuits to access data in (e.g., each of) the plurality of cache memory banks in response to a request for data access from the spatial array of processing elements; a plurality of translation lookaside buffers comprising a translation lookaside buffer in each of the plurality of request address file circuits to provide an output of a physical address for an input of a virtual address; a plurality of higher level, than the plurality of translation lookaside buffers, translation lookaside buffers comprising a higher level translation lookaside buffer in each of the plurality of cache memory banks to provide an output of a physical address for an input of a virtual address; and a translation lookaside buffer manager circuit to perform a first page walk in the plurality of cache memory banks for a miss of an input of a virtual address into a first translation lookaside buffer and into a first higher level translation lookaside buffer to determine a physical address mapped to the virtual address, store a mapping of the virtual address to the physical address from the first page walk in the first higher level translation lookaside buffer to cause the first higher level translation lookaside buffer to send the physical address to the first translation lookaside buffer in a first request address file circuit. The translation lookaside buffer manager circuit may simultaneously, with the first page walk, perform a second page walk in the plurality of cache memory banks, wherein the second page walk is for a miss of an input of a virtual address into a second translation lookaside buffer and into a second higher level translation lookaside buffer to determine a physical address mapped to the virtual address, store a mapping of the virtual address to the physical address from the second page walk in the second higher level translation lookaside buffer to cause the second higher level translation lookaside buffer to send the physical address to the second translation lookaside buffer in a second request address file circuit. The receipt of the physical address in the first translation lookaside buffer may cause the first request address file circuit to perform a data access for the request for data access from the spatial array of processing elements on the physical address in the plurality of cache memory banks. The translation lookaside buffer manager circuit may insert an indicator in the first higher level translation lookaside buffer for the miss of the input of the virtual address in the first translation lookaside buffer and the first higher level translation lookaside buffer to prevent an additional page walk for the input of the virtual address during the first page walk. The translation lookaside buffer manager circuit may receive a shutdown message from a requesting entity for a mapping of a physical address to a virtual address, invalidate the mapping in a higher level translation lookaside buffer storing the mapping, and send shutdown messages to only those of the plurality of request address file circuits that include a copy of the mapping in a respective translation lookaside buffer, wherein each of those of the plurality of request address file circuits are to send an acknowledgement message to the translation lookaside buffer manager circuit, and the translation lookaside buffer manager circuit is to send a shutdown completion acknowledgement message to the requesting entity when all acknowledgement messages are received. The translation lookaside buffer manager circuit may receive a shutdown message from a requesting entity for a mapping of a physical address to a virtual address, invalidate the mapping in a higher level translation lookaside buffer storing the mapping, and send shutdown messages to all of the plurality of request address file circuits, wherein each of the plurality of request address file circuits are to send an acknowledgement message to the translation lookaside buffer manager circuit, and the translation lookaside buffer manager circuit is to send a shutdown completion acknowledgement message to the requesting entity when all acknowledgement messages are received.

[0616] In yet another embodiment, a method includes: overlaying an input of a dataflow graph comprising a plurality of nodes into a spatial array of processing elements comprising a communications network with each node represented as a dataflow operator in the spatial array of processing elements; coupling a plurality of request address file circuits to the spatial array of processing elements and a plurality of cache memory banks with each request address file circuit of the plurality of request address file circuits accessing data in the plurality of cache memory banks in response to a request for data access from the spatial array of processing elements; providing an output of a physical address for an input of a virtual address into a translation lookaside buffer of a plurality of translation lookaside buffers comprising a translation lookaside buffer in each of the plurality of request address file circuits; providing an output of a physical address for an input of a virtual address into a higher level, than the plurality of translation lookaside buffers, translation lookaside buffer of a plurality of higher level translation lookaside buffers comprising a higher level translation lookaside buffer in each of the plurality of cache memory banks; coupling a translation lookaside buffer manager circuit to the plurality of request address file circuits and the plurality of cache memory banks; and performing a first page walk in the plurality of cache memory banks for a miss of an input of a virtual address into a first translation lookaside buffer and into a first higher level translation lookaside buffer with the translation lookaside buffer manager circuit to determine a physical address mapped to the virtual address, store a mapping of the virtual address to the physical address from the first page walk in the first higher level translation lookaside buffer to cause the first higher level translation lookaside buffer to send the physical address to the first translation lookaside buffer in a first request address file circuit. The method may include simultaneously, with the first page walk, performing a second page walk in the plurality of cache memory banks, wherein the second page walk is for a miss of an input of a virtual address into a second translation lookaside buffer and into a second higher level translation lookaside buffer with the translation lookaside buffer manager circuit to determine a physical address mapped to the virtual address, store a mapping of the virtual address to the physical address from the second page walk in the second higher level translation lookaside buffer to cause the second higher level translation lookaside buffer to send the physical address to the second translation lookaside buffer in a second request address file circuit. The method may include causing the first request address file circuit to perform a data access for the request for data access from the spatial array of processing elements...
on the physical address in the plurality of cache memory banks in response to receipt of the physical address in the first translation lookaside buffer. The method may include inserting, with the translation lookaside buffer manager circuit, an indicator in the first higher level translation lookaside buffer for the miss of the input of the virtual address in the first translation lookaside buffer and the first higher level translation lookaside buffer to prevent an additional page walk for the input of the virtual address during the first page walk. The method may include receiving, with the translation lookaside buffer manager circuit, a shootdown message from a requesting entity for a mapping of a physical address to a virtual address, invalidating the mapping in a higher level translation lookaside buffer storing the mapping, and sending shootdown messages to only those of the plurality of request address file circuits that include a copy of the mapping in a respective translation lookaside buffer, wherein each of those of the plurality of request address file circuits are to send an acknowledgement message to the translation lookaside buffer manager circuit, and the translation lookaside buffer manager circuit is to send a shootdown completion acknowledgement message to the requesting entity when all acknowledgement messages are received. The method may include receiving, with the translation lookaside buffer manager circuit, a shootdown message from a requesting entity for a mapping of a physical address to a virtual address, invalidating the mapping in a higher level translation lookaside buffer storing the mapping, and sending shootdown messages to all of the plurality of request address file circuits, wherein each of the plurality of request address file circuits are to send an acknowledgement message to the translation lookaside buffer manager circuit, and the translation lookaside buffer manager circuit is to send a shootdown completion acknowledgement message to the requesting entity when all acknowledgement messages are received.

[0617] In another embodiment, a system includes a core with a decoder to decode an instruction into a decoded instruction and an execution unit to execute the decoded instruction to perform a first operation; a spatial array of processing elements comprising a communications network to receive an input of a dataflow graph comprising a plurality of nodes, wherein the dataflow graph is to be overlaid into the spatial array of processing elements with each node represented as a dataflow operator in the spatial array of processing elements, and the spatial array of processing elements is to perform a second operation by a respective incoming operand set arriving at each of the dataflow operators; a plurality of request address file circuits coupled to the spatial array of processing elements and a cache memory, each request address file circuit of the plurality of request address file circuits to access data in the cache memory in response to a request for data access from the spatial array of processing elements; a plurality of translation lookaside buffers comprising a translation lookaside buffer in each of the plurality of request address file circuits to provide an output of a physical address for an input of a virtual address; and a translation lookaside buffer manager circuit comprising a higher level translation lookaside buffer than the plurality of translation lookaside buffers, the translation lookaside buffer manager circuit to perform a first page walk in the cache memory for a miss of an input of a virtual address into a first translation lookaside buffer and into the higher level translation lookaside buffer to determine a physical address mapped to the virtual address, store a mapping of the virtual address to the physical address from the first page walk in the higher level translation lookaside buffer to cause the higher level translation lookaside buffer to send the physical address to the first translation lookaside buffer in a first request address file circuit. The translation lookaside buffer manager circuit may simultaneously, with the first page walk, perform a second page walk in the cache memory, wherein the second page walk is for a miss of an input of a virtual address into a second translation lookaside buffer and into the higher level translation lookside buffer to determine a physical address mapped to the virtual address, store a mapping of the virtual address to the physical address from the second page walk in the higher level translation lookside buffer to cause the higher level translation lookside buffer to send the physical address to the second translation lookside buffer in a second request address file circuit. The receipt of the physical address in the first translation lookside buffer may cause the first request address file circuit to perform a data access for the request for data access from the spatial array of processing elements on the physical address in the cache memory. The translation lookside buffer manager circuit may insert an indicator in the higher level translation lookside buffer for the miss of the input of the virtual address in the first translation lookside buffer and the higher level translation lookside buffer to prevent an additional page walk for the input of the virtual address during the first page walk. The translation lookside buffer manager circuit may receive a shootdown message from a requesting entity for a mapping of a physical address to a virtual address, invalidate the mapping in the higher level translation lookside buffer, and send shootdown messages to only those of the plurality of request address file circuits that include a copy of the mapping in a respective translation lookside buffer, wherein each of those of the plurality of request address file circuits are to send an acknowledgement message to the translation lookside buffer manager circuit, and the translation lookside buffer manager circuit is to send a shootdown completion acknowledgement message to the requesting entity when all acknowledgement messages are received.

[0618] In yet another embodiment, a system includes a core with a decoder to decode an instruction into a decoded instruction and an execution unit to execute the decoded instruction to perform a first operation; a spatial array of processing elements comprising a communications network to receive an input of a dataflow graph comprising a plurality of nodes, wherein the dataflow graph is to be overlaid into the spatial array of processing elements with each node represented as a dataflow operator in the spatial array of processing elements, and the spatial array of processing elements is to perform a second operation by a respective,
incoming operand set arriving at each of the dataflow operators; a plurality of request address file circuits coupled to the spatial array of processing elements and a plurality of cache memory banks, each request address file circuit of the plurality of request address file circuits to access data in (e.g., each of) the plurality of cache memory banks in response to a request for data access from the spatial array of processing elements; a plurality of translation looksaside buffers comprising a translation looksaside buffer in each of the plurality of request address file circuits to provide an output of a physical address for an input of a virtual address; a plurality of higher level, than the plurality of translation looksaside buffers, translation looksaside buffers comprising a higher level translation looksaside buffer in each of the plurality of cache memory banks to provide an output of a physical address for an input of a virtual address; and a translation looksaside buffer manager circuit to perform a first page walk in the plurality of cache memory banks for a miss of an input of a virtual address into a first translation looksaside buffer and into a first higher level translation looksaside buffer to determine a physical address mapped to the virtual address, store a mapping of the virtual address to the physical address from the first page walk in the first higher level translation looksaside buffer to send the physical address to the first translation looksaside buffer in a first request address file circuit. The translation looksaside buffer manager circuit may simultaneously, with the first page walk, perform a second page walk in the plurality of cache memory banks, wherein the second page walk is for a miss of an input of a virtual address into a second translation looksaside buffer and into a second higher level translation looksaside buffer to determine a physical address mapped to the virtual address, store a mapping of the virtual address to the physical address from the second page walk in the second higher level translation looksaside buffer to cause the second higher level translation looksaside buffer to send the physical address to the second translation looksaside buffer in a second request address file circuit. The receipt of the physical address in the first translation looksaside buffer may cause the first request address file circuit to perform a data access for the request for data access from the spatial array of processing elements on the physical address in the plurality of cache memory banks. The translation looksaside buffer manager circuit may insert an indicator in the first higher level translation looksaside buffer for the miss of the input of the virtual address in the first translation looksaside buffer and the first higher level translation looksaside buffer to prevent an additional page walk for the input of the virtual address during the first page walk. The translation looksaside buffer manager circuit may receive a shutdown message from a requesting entity for a mapping of a physical address to a virtual address, invalidate the mapping in a higher level translation looksaside buffer storing the mapping, and send shutdown messages to all of the plurality of request address file circuits, wherein each of the plurality of request address file circuits are to send an acknowledgement message to the translation looksaside buffer manager circuit, and the translation looksaside buffer manager circuit is to send a shutdown completion acknowledgment message to the requesting entity when all acknowledgement messages are received. The translation looksaside buffer manager circuit may receive a shutdown message from a requesting entity for a mapping of a physical address to a virtual address, invalidate the mapping in a higher level translation looksaside buffer storing the mapping, and send shutdown messages to all of the plurality of request address file circuits, wherein each of the plurality of request address file circuits are to send an acknowledgement message to the translation looksaside buffer manager circuit, and the translation looksaside buffer manager circuit is to send a shutdown completion acknowledgment message to the requesting entity when all acknowledgement messages are received.

In another embodiment, an apparatus (e.g., a processor) includes: a spatial array of processing elements comprising a communications network to receive an input of a dataflow graph comprising a plurality of nodes, wherein the dataflow graph is to be overlaid into the spatial array of processing elements with each node represented as a dataflow operator in the spatial array of processing elements, and the spatial array of processing elements is to perform an operation by a respective, incoming operand set arriving at each of the dataflow operators; a plurality of request address file circuits coupled to the spatial array of processing elements and a cache memory, each request address file circuit of the plurality of request address file circuits to access data in the cache memory in response to a request for data access from the spatial array of processing elements; a plurality of translation looksaside buffers comprising a translation looksaside buffer in each of the plurality of request address file circuits to provide an input of a physical address for an input of a virtual address; and a means comprising a higher level translation looksaside buffer than the plurality of translation looksaside buffers, the means to perform a first page walk in the cache memory for a miss of an input of a virtual address into a first translation looksaside buffer and into the higher level translation looksaside buffer to determine a physical address mapped to the virtual address, store a mapping of the virtual address to the physical address from the first page walk in the higher level translation looksaside buffer to send the physical address to a translation looksaside buffer in a first request address file circuit. In yet another embodiment, an apparatus includes a spatial array of processing elements comprising a communications network to receive an input of a dataflow graph comprising a plurality of nodes, wherein the dataflow graph is to be overlaid into the spatial array of processing elements with each node represented as a dataflow operator in the spatial array of processing elements, and the spatial array of processing elements is to perform an operation by a respective, incoming operand set arriving at each of the dataflow operators; a plurality of request address file circuits coupled to the spatial array of processing elements and a plurality of cache memory banks, each request address file circuit of the plurality of request address file circuits to access data in (e.g., each of) the plurality of cache memory banks in response to a request for data access from the spatial array of processing elements; a plurality of translation looksaside buffers comprising a translation looksaside buffer in each of the plurality of request address file circuits to provide an output of a physical address for an input of a virtual address; a plurality of higher level, than the plurality of translation looksaside buffers, translation looksaside buffers comprising a higher level translation looksaside buffer in each of the plurality of cache memory banks to provide an output of a physical address for an input of a virtual address.
physical address for an input of a virtual address; and a
means to perform a first page walk in the plurality of cache
memory banks for a miss of an input of a virtual address into
a first translation lookaside buffer and into a first higher level
translation lookaside buffer to determine a physical address
mapped to the virtual address, store a mapping of the virtual
address to the physical address from the first page walk in
the first higher level translation lookaside buffer to cause the
first higher level translation lookaside buffer to send the
physical address to the first translation lookaside buffer in a
first request address file circuit.

[0621] In another embodiment, an apparatus comprises a
data storage device that stores code that when executed by
a hardware processor causes the hardware processor to
perform any method disclosed herein. An apparatus may be
described as detailed description. A method may be as
described in the detailed description.

[0622] In yet another embodiment, a non-transitory
machine readable medium that stores code that when
executed by a machine causes the machine to perform a
method comprising any method disclosed herein.

[0623] An instruction set (e.g., for execution by a core)
may include one or more instruction formats. A given
instruction format may define various fields (e.g., number of
bits, location of bits) to specify, among other things, the
operation to be performed (e.g., opcode, the operand(s)
on which that operation is to be performed and/or other data
field(s) (e.g., mask). Some instruction formats are further
defined as described in the detailed description. For example, the instruction templates of a
given instruction format may be defined to have different
subsets of the instruction format’s fields (the included fields
are typically in the same order, but at least some have
different bit positions because there are less fields included)
and/or defined to have a given field interpreted differently.
Thus, each instruction of an ISA is expressed using a given
instruction format (and, if defined, in a given one of the
instruction templates of that instruction format) and includes
fields for specifying the operation and the operands. For
example, an exemplary ADD instruction has a specific
opcode and an instruction format that includes an opcode
field to specify that opcode and operand fields to select
operands (source1/destination and source2); and an occurrence
of this ADD instruction in an instruction stream will have specific contents in the operand fields that select
specific operands. A set of SIMD extensions referred to as
the Advanced Vector Extensions (AVX) (AVX1 and AVX2)
and using the Vector Extensions (VEX) coding scheme has
been released and/or published (e.g., see Intel® 64 and
IA-32 Architecture Software Developer’s Manual, June
2016; and see Intel® Architecture Instruction Set Extensions
Programming Reference, February 2016).

Exemplary Instruction Formats

[0624] Embodiments of the instruction(s) described herein
may be embodied in different formats. Additionally, examplary
systems, architectures, and pipelines are detailed
below. Embodiments of the instruction(s) may be executed
on such systems, architectures, and pipelines, but are not
limited to those detailed.

Generic Vector Friendly Instruction Format

[0625] A vector friendly instruction format is an instruc
tion format that is suited for vector instructions (e.g., there
are certain fields specific to vector operations). While
embodiments are described in which both vector and scalar
operations are supported through the vector friendly instruc
tion format, alternative embodiments use only vector opera
tions the vector friendly instruction format.

[0626] FIGS. 101A-101B are block diagrams illustrating a
generic vector friendly instruction format and instruction
templates thereof according to embodiments of the disclo
sure. FIG. 101A is a block diagram illustrating a generic
vector friendly instruction format and class A instruction
templates thereof according to embodiments of the disclo
sure; while FIG. 101B is a block diagram illustrating the
generic vector friendly instruction format and class B
instruction templates thereof according to embodiments of the
disclosure. Specifically, a generic vector friendly instruc
tion format 10100 for which are defined class A and class B
instruction templates, both of which include no memory
access 10105 instruction templates and memory access
10120 instruction templates. The term generic in the context
of the vector friendly instruction format refers to the instruc
tion format not being tied to any specific instruction set.

[0627] While embodiments of the disclosure will be
described in which the vector friendly instruction format
supports the following: a 64 byte vector operand length (or
size) with 32 bit (4 byte) or 64 bit (8 byte) data element
widths (or sizes) (and thus, a 64 byte vector consists of either
16 doubleword-size elements or alternatively, 8 quadword-
size elements); a 64 byte vector operand length (or size) with
16 bit (2 byte) or 8 bit (1 byte) data element widths (or
sizes); a 32 byte vector operand length (or size) with 32 bit
(4 byte), 64 bit (8 byte), 16 bit (2 byte), or 8 bit (1 byte) data
element widths (or sizes); and a 16 byte vector operand
length (or size) with 32 bit (4 byte), 64 bit (8 byte), 16 bit
(2 byte), or 8 bit (1 byte) data element widths (or sizes); and
alternative embodiments may support more, less and/or
different vector operand sizes (e.g., 256 byte vector oper-
ands) with more, less, or different data element widths (e.g.,
128 bit (16 byte) data element widths).

[0628] The class A instruction templates in FIG. 101A
include: 1) within the no memory access 10105 instruction
templates there is shown a no memory access, full round
control type operation 10110 instruction template and a
no memory access, data transform type operation 10115
instruction template; and 2) within the memory access
10120 instruction templates there is shown a memory
access, temporal 10125 instruction template and a memory
access, non-temporal 10130 instruction template. The class
B instruction templates in FIG. 101B include: 1) within the
no memory access 10105 instruction templates there is
shown a no memory access, write mask control, partial
round control type operation 10112 instruction template and
a no memory access, write mask control, vsize type operation
10117 instruction template; and 2) within the memory
access 10120 instruction templates there is shown a memory
access, write mask control 10127 instruction template.

[0629] The generic vector friendly instruction format
10100 includes the following fields listed below in the order
illustrated in FIGS. 101A-101B.

[0630] Format field 10140—a specific value (an instruc
tion format identifier value) in this field uniquely identifies
the vector friendly instruction format, and thus occurrences
of instructions in the vector friendly instruction format in
instruction streams. As such, this field is optional in the
sense that it is not needed for an instruction set that has only the generic vector friendly instruction format.

[0631] Base operation field 10142—its content distinguishes different base operations.

[0632] Register index field 10144—its content, directly or through address generation, specifies the locations of the source and destination operands, be they in registers or in memory. These include a sufficient number of bits to select N registers from a P×Q (e.g., 32x512, 16x128, 32x1024, 64x1024) register file. While in one embodiment N may be up to three sources and one destination register, alternative embodiments may support more or less sources and destination registers (e.g., may support up to two sources where one of these sources also acts as the destination, may support up to three sources where one of these sources also acts as the destination, may support up to two sources and one destination).

[0633] Modifier field 10146—its content distinguishes occurrences of instructions in the generic vector instruction format that specify memory access from those that do not; that is, between no memory access 10105 instruction templates and memory access 10120 instruction templates. Memory access operations read and/or write to the memory hierarchy (in some cases specifying the source and/or destination addresses using values in registers), while non-memory access operations do not (e.g., the source and destinations are registers). While in one embodiment this field also selects between three different ways to perform memory address calculations, alternative embodiments may support more, less, or different ways to perform memory address calculations.

[0634] Augmentation operation field 10150—its content distinguishes which one of a variety of different operations to be performed in addition to the base operation. This field is context specific. In one embodiment of the disclosure, this field is divided into a class field 10168, an alpha field 10152, and a beta field 10154. The augmentation operation field 10150 allows common groups of operations to be performed in a single instruction rather than 2, 3, or 4 instructions.

[0635] Scale field 10160—its content allows for the scaling of the index field’s content for memory address generation (e.g., for address generation that uses $2^{scale} + index + base$).

[0636] Displacement Field 10162A—its content is used as part of memory address generation (e.g., for address generation that uses $2^{scale} + index + base + displacement$).

[0637] Displacement Factor Field 10162B (note that the juxtaposition of displacement field 10162A directly over displacement factor field 10162B indicates one or the other is used)—its content is used as part of address generation; it specifies a displacement factor that is to be scaled by the size of a memory access (N)—where N is the number of bytes in the memory access (e.g., for address generation that uses $2^{scale} + index + base + scaled displacement$). Redundant low-order bits are ignored and hence, the displacement factor field’s content is multiplied by the memory operands total size (N) in order to generate the final displacement to be used in calculating an effective address. The value of N is determined by the processor hardware at runtime based on the full opcode field 10174 (described later herein) and the data manipulation field 10154C. The displacement field 10162A and the displacement factor field 10162B are optional in the sense that they are not used for the no memory access 10105 instruction templates and/or different embodiments may implement only one or none of the two.

[0638] Data element width field 10164—its content distinguishes which one of a number of data element widths is to be used (in some embodiments for all instructions; in other embodiments for only some of the instructions). This field is optional in the sense that it is not needed if only one data element width is supported and/or data element widths are supported using some aspect of the opcodes.

[0639] Write mask field 10170—its content controls, on a per data element position basis, whether that data element position in the destination operand reflects the result of the base operation and augmentation operation. Class A instruction templates support merging-waremasking, while class B instruction templates support both merging- and zeroing-waremasking. When merging, vector masks allow any set of elements in the destination to be protected from updates during the execution of any operation (specified by the base operation and the augmentation operation); in other one embodiment, preserving the old value of each element of the destination where the corresponding mask bit has a 0. In contrast, when zeroing vector masks allow any set of elements in the destination to be zeroed during the execution of any operation (specified by the base operation and the augmentation operation); in one embodiment, an element of the destination is set to 0 when the corresponding mask bit has a 0 value. A subset of this functionality is the ability to control the length of the operation being performed (that is, the span of elements being modified, from the first to the last one); however, it is not necessary that the elements that are modified be consecutive. Thus, the write mask field 10170 allows for partial vector operations, including loads, stores, arithmetic, logical, etc. While embodiments of the disclosure are described in which the write mask field’s 10170 content selects one of a number of write mask registers that contain the write mask to be used (and thus the write mask field’s 10170 content indirectly identifies that masking is to be performed), alternative embodiments instead or additionally allow the mask write field’s 10170 content to directly specify the masking to be performed.

[0640] Immediate field 10172—its content allows for the specification of an immediate. This field is optional in the sense that is it not present in an implementation of the generic vector friendly format that does not support immediate and it is not present in instructions that do not use an immediate.

[0641] Class field 10168—its content distinguishes between different classes of instructions. With reference to FIGS. 101-A-B, the contents of this field select between class A and class B instructions. In FIGS. 101-A-B, rounded corner squares are used to indicate a specific value is present in a field (e.g., class A 10168A and class B 10168B for the class field 10168 respectively in FIGS. 101-A-B).

Instruction Templates of Class A

[0642] In the case of the non-memory access 10105 instruction templates of class A, the alpha field 10152 is interpreted as an RS field 10152A, whose content distinguishes which one of the different augmentation operation types are to be performed (e.g., round 10152A.1 and data transform 10152A.2 are respectively specified for the no memory access, round type operation 10110 and the no memory access, data transform type operation 10115 instruction templates), while the beta field 10154 distin-
guishes which of the operations of the specified type is to be performed. In the no memory access 10105 instruction templates, the scale field 10160, the displacement field 10162A, and the displacement scale filed 10162B are not present.

No-Memory Access Instruction Templates—Full Round Control Type Operation

[0643] In the no memory access full round control type operation 10110 instruction template, the beta field 10154 is interpreted as a round control field 10154A, whose content (s) provide static rounding. While the described embodiments of the disclosure the round control field 10154A includes a suppress all floating point exceptions (SAE) field 10156 and a round operation control field 10158; alternative embodiments may support may encode both these concepts into the same field or only have one or the other of these concepts/fields (e.g., may have only the round operation control field 10158).

[0644] SAE field 10156—its content distinguishes whether or not to disable the exception event reporting; when the SAE field's 10156 content indicates suppression is enabled, a given instruction does not report any kind of floating-point exception flag and does not raise any floating point exception handler.

[0645] Round operation control field 10158—its content distinguishes which one of a group of rounding operations to perform (e.g., Round-up, Round-down, Round-towards-zero and Round-to-nearest). Thus, the round operation control field 10158 allows for the changing of the rounding mode on a per instruction basis. In one embodiment of the disclosure where a processor includes a control register for specifying rounding modes, the round operation control field's 10150 content overrides that register value.

No Memory Access Instruction Templates—Data Transform Type Operation

[0646] In the no memory access data transform type operation 10115 instruction template, the beta field 10154 is interpreted as a data transform field 10154B, whose content distinguishes which one of a number of data transforms is to be performed (e.g., no data transform, swizzle, broadcast).

[0647] In the case of a memory access 10120 instruction template of class A, the alpha field 10152 is interpreted as an eviction hint field 10152B, whose content distinguishes which one of the eviction hints is to be used (in FIG. 101A, temporal 10152B.1 and non-temporal 10152B.2 are respectively specified for the memory access, temporal 10125 instruction template and the memory access, non-temporal 10130 instruction template), while the beta field 10154 is interpreted as a data manipulation field 10154C, whose content distinguishes which one of a number of data manipulation operations (also known as primitives) is to be performed (e.g., no manipulation; broadcast; up conversion of a source; and down conversion of a destination). The memory access 10120 instruction templates include the scale field 10160, and optionally the displacement field 10162A or the displacement scale field 10162B.

[0648] Vector memory instructions perform vector loads from and vector stores to memory, with conversion support. As with regular vector instructions, vector memory instructions transfer data from/to memory in a data element-wise fashion, with the elements that are actually transferred is dictated by the contents of the vector mask that is selected as the write mask.

Memory Access Instruction Templates—Temporal

[0649] Temporal data is data likely to be reused soon enough to benefit from caching. This is, however, a hint, and different processors may implement it in different ways, including ignoring the hint entirely.

Memory Access Instruction Templates—Non-Temporal

[0650] Non-temporal data is data unlikely to be reused soon enough to benefit from caching in the 1st-level cache and should be given priority for eviction. This is, however, a hint, and different processors may implement it in different ways, including ignoring the hint entirely.

Instruction Templates of Class B

[0651] In the case of the instruction templates of class B, the alpha field 10152 is interpreted as a write mask control (Z) field 10152C, whose content distinguishes whether the write masking controlled by the write mask field 10170 should be a merging or a zeroing.

[0652] In the case of the non-memory access 10105 instruction templates of class B, part of the beta field 10154 is interpreted as an RL field 10157A, whose content distinguishes which one of the different augmentation operation types are to be performed (e.g., round 10157A.1 and vector length (VSIZE) 10157A.2 are respectively specified for the no-memory access, write mask control, partial round control type operation 10112 instruction template and the no-memory access, write mask control, VSIZE type operation 10117 instruction template), while the rest of the beta field 10154 distinguishes which of the operations of the specified type is to be performed. In the no memory access 10105 instruction templates, the scale field 10160, the displacement field 10162A, and the displacement scale filed 10162B are not present.

[0653] In the no memory access, write mask control, partial round control type operation 10110 instruction template, the rest of the beta field 10154 is interpreted as a round operation field 10159A and exception event reporting is disabled (a given instruction does not report any kind of floating-point exception flag and does not raise any floating point exception handler).

[0654] Round operation control field 10159A—just as round operation control field 10158, its content distinguishes which one of a group of rounding operations to perform (e.g., Round-up, Round-down, Round-towards-zero and Round-to-nearest). Thus, the round operation control field 10159A allows for the changing of the rounding mode on a per instruction basis. In one embodiment of the disclosure where a processor includes a control register for specifying rounding modes, the round operation control field's 10150 content overrides that register value.

[0655] In the no memory access, write mask control, VSIZE type operation 10117 instruction template, the rest of the beta field 10154 is interpreted as a vector length field 10159B, whose content distinguishes which one of a number of data vector lengths is to be performed on (e.g., 128, 256, or 512 byte).

[0656] In the case of a memory access 10120 instruction template of class B, part of the beta field 10154 is interpreted
as a broadcast field 10157B, whose content distinguishes whether or not the broadcast type data manipulation operation is to be performed, while the rest of the beta field 10154 is interpreted the vector length field 10159B. The memory access 10120 instruction templates include the scale field 10160, and optionally the displacement field 10162A or the displacement scale field 10162B.

[0657] With regard to the generic vector friendly instruction format 10100, a full opcode field 10174 is shown including the format field 10140, the base operation field 10142, and the data element width field 10164. While one embodiment is shown where the full opcode field 10174 includes all of these fields, the full opcode field 10174 includes less than all of these fields in embodiments that do not support all of them. The full opcode field 10174 provides the operation code (opcode).

[0658] The augmentation operation field 10150, the data element width field 10164, and the write mask field 10170 allow these features to be specified on a per instruction basis in the generic vector friendly instruction format.

[0659] The combination of write mask field and data element width field create typed instructions in that they allow the mask to be applied based on different data element widths.

[0660] The various instruction templates found within class A and class B are beneficial in different situations. In some embodiments of the disclosure, different processors or different cores within a processor may support only class A, only class B, or both classes. For instance, a high performance general purpose out-of-order core intended for general-purpose computing may support only class B, a core intended primarily for graphics and/or scientific (throughput) computing may support only class A, and a core intended for both may support both (of course, a core that has some mix of templates and instructions from both classes but not all templates and instructions from both classes is within the purview of the disclosure). Also, a single processor may include multiple cores, all of which support the same class or in which different cores support different class. For instance, in a processor with separate graphics and general purpose cores, one of the graphics cores intended primarily for graphics and/or scientific computing may support only class A, while one or more of the general purpose cores may be high performance general purpose cores with out of order execution and register renaming intended for general-purpose computing that support only class B. Another processor that does not have a separate graphics core, may include one more general purpose in-order or out-of-order cores that support both class A and class B. Of course, features from one class may also be implement in the other class in different embodiments of the disclosure. Programs written in a high level language would be put (e.g., just in time compiled or statically compiled) into an array of different executable forms, including: 1) a form having only instructions of the class(es) supported by the target processor for execution; or 2) a form having alternative routines written using different combinations of the instructions of all classes and having control flow code that selects the routines to execute based on the instructions supported by the processor which is currently executing the code.

Exemplary Specific Vector Friendly Instruction Format

[0661] FIG. 102A is a block diagram illustrating an exemplary specific vector friendly instruction format according to embodiments of the disclosure. FIG. 102A shows a specific vector friendly instruction format 10200 that is specific in the sense that it specifies the location, size, interpretation, and order of the fields, as well as values for some of those fields. The specific vector friendly instruction format 10200 may be used to extend the x86 instruction set, and thus some of the fields are similar or the same as those used in the existing x86 instruction set and extension thereof (e.g., AVX). This format remains consistent with the prefix encoding field, real opcode byte field, MOD R/M field, SIB field, displacement field, and immediate fields of the existing x86 instruction set with extensions. The fields from FIG. 101 into which the fields from FIG. 102A map are illustrated.

[0662] It should be understood that, although embodiments of the disclosure are described with reference to the specific vector friendly instruction format 10200 in the context of the generic vector friendly instruction format 10100 for illustrative purposes, the disclosure is not limited to the specific vector friendly instruction format 10200 except where claimed. For example, the generic vector friendly instruction format 10100 contemplates a variety of possible sizes for the various fields, while the specific vector friendly instruction format 10200 is shown as having fields of specific sizes. By way of specific example, while the data element width field 10164 is illustrated as a one bit field in the specific vector friendly instruction format 10200, the disclosure is not so limited (that is, the generic vector friendly instruction format 10100 contemplates other sizes of the data element width field 10164).

[0663] The generic vector friendly instruction format 10100 includes the following fields listed below in the order illustrated in FIG. 102A.

[0664] EVEX Prefix (Bytes 0-3) 10202—is encoded in a four-byte form.

[0665] Format Field 10140 (EVEX Byte 0, bits [7:0])—the first byte (EVEX Byte 0) is the format field 10140 and it contains 0x62 (the unique value used for distinguishing the vector friendly instruction format in one embodiment of the disclosure).

[0666] The second-fourth bytes (EVEX Bytes 1-3) include a number of bit fields providing specific capability.

[0667] REX field 10205 (EVEX Byte 1, bits [7-5])—consists of a EVEX.R bit field (EVEX Byte 1, bit [7]—R), EVEX.X bit field (EVEX byte 1, bit [6]—X), and 10157BEX byte 1, bit [5]—B. The EVEX.R, EVEX.X, and EVEX.B bit fields provide the same functionality as the corresponding VEX bit fields, and are encoded using is complement form, i.e. ZMM0 is encoded as 4711B, ZMM15 is encoded as 0000B. Other fields of the instructions encode the lower three bits of the register indexes as is known in the art (rrr, xxx, and bbb), so that Rrrr, Xxxx, and Bbbb may be formed by adding EVEX.R, EVEX.X, and EVEX.B.

[0668] REX’ field 10110—is this the first part of the REX’ field 10110 and is the EVEX.R’ bit field (EVEX Byte 1, bit [4]—R’) that is used to encode either the upper 16 or lower 16 of the extended 32 register set. In one embodiment of the disclosure, this bit, along with others as indicated below, is stored in bit inverted format to distinguish (in the well-known x86 32-bit mode) from the BOUND instruction, whose real opcode byte is 62, but does not accept in the MOD R/M field (described below) the value of 11 in the
MOD field; alternative embodiments of the disclosure do not store this and the other indicated bits below in the inverted format. A value of 1 is used to encode the lower 16 registers. In other words, R'Rrrr is formed by combining EVEX.R', EVEX.R, and the other RRR from other fields.

[0669] Opcode map field 10215 (EVEX byte 1, bits [3:0]—mmmm)—its content encodes an implied leading opcode byte (0F, 0F' 38, or 0F' 3).

[0670] Data element width field 10164 (EVEX byte 2, bit [7]—W)—is represented by the notation EVEX.W. EVEX.W is used to define the granularity (size) of the datatype (either 32-bit data elements or 64-bit data elements).

[0671] EVEX.vvvv 10220 (EVEX Byte 2, bits [6:3]—vvvv)—the role of EVEX.vvvv may include the following: 1) EVEX.vvvv encodes the first source register operand, specified in inverted (1s complement) form and is valid for instructions with 2 or more source operands; 2) EVEX.vvvv encodes the destination register operand, specified in 1s complement form for certain vector shifts; or 3) EVEX.vvvv does not encode any operand, the field is reserved and should contain 4111b. Thus, EVEX.vvvv field 10220 encodes the 4 low-order bits of the first source register specifier stored in inverted (1s complement) form. Depending on the instruction, an extra different EVEX bit field is used to extend the specifier size to 32 registers.

[0672] EVEX.0 10168 Class field (EVEX byte 2, bit [2]—U)—if EVEX.0 = 0, it indicates class A or EVEX.10; if EVEX.0 = 1, it indicates class B or EVEX.11.

[0673] Prefix encoding field 10225 (EVEX byte 2, bits [1:0]—pp)—provides additional bits for the base operation field. In addition to providing support for the legacy SSE instructions in the EVEX prefix format, this also has the benefit of compacting the SIMD prefix (rather than requiring a byte to express the SIMD prefix, the EVEX prefix requires only 2 bits). In one embodiment, to support legacy SSE instructions that use a SIMD prefix (MMX, SSE, SSE2, SSE3) both the legacy format and in the EVEX prefix format, these legacy SIMD prefixes are encoded into the SIMD prefix encoding field; and at runtime are expanded into the legacy SIMD prefix prior to being provided to the decoder’s PLA (so the PLA can execute both the legacy and EVEX format of these legacy instructions without modification). Although newer instructions could use the EVEX prefix encoding field’s content directly as an opcode extension, certain embodiments expand in a similar fashion for consistency but allow for different meanings to be specified by these legacy SIMD prefixes. An alternative embodiment may redesign the PLA to support the 2 bit SIMD prefix encodings, and thus not require the expansion.

[0674] Alpha field 10152 (EVEX byte 3, bit [7]—EH; also known as EVEX.EH, EVEX.rs, EVEX.rl, EVEX.mask control, and EVEX.N; also illustrated with α)—as previously described, this field is context specific.

[0675] Beta field 10154 (EVEX byte 3, bits [6:4]—S, also known as EVEX.z, EVEX.z2, EVEX.z2o, EVEX.o, EVEX.l, EVEX.LL, EVEX.LLB; also illustrated with PP)[3]—as previously described, this field is context specific.

[0676] REX’ field 10110—is this the remainder of the REX’ field and is the EVEX.vV’ bit field (EVEX Byte 3, bit [3]—V’V) that may be used to encode either the upper 16 or lower 16 of the extended 32 register set. This bit is stored in bit inverted format. A value of 1 is used to encode the lower 16 registers. In other words, VVVVV is formed by combining EVEX.V’, EVEX.vvvv.

[0677] Write mask field 10170 (EVEX byte 3, bits [2:0]—kkk)—its content specifies the index of a register in the write mask registers as previously described. In one embodiment of the disclosure, the specific value EVEX.kkk=000 has a special behavior implying no write mask is used for the particular instruction (this may be implemented in a variety of ways including the use of a write mask hardwired to all ones or hardware that bypasses the masking hardware).

[0678] Real Opcode Field 10230 (Byte 4) is also known as the opcode byte. Part of the opcode is specified in this field.

[0679] MOD R/M Field 10240 (Byte 5) includes MOD field 10242, Reg field 10244, and R/M field 10246. As previously described, the MOD field’s 10242 content distinguishes between memory access and non-memory access operations. The role of Reg field 10244 can be summarized to two situations: encoding either the destination register operand or a source register operand, or be treated as an opcode extension and not used to encode any instruction operand. The role of R/M field 10246 may include the following: encoding the instruction operand that references a memory address, or encoding either the destination register operand or a source register operand.

[0680] Scale, Index, Base (SIB) Byte (6)—As previously described, the scale field’s 5450 content is used for memory address generation. SIB.xxx 10254 and SIB.xxx 10256—the contents of these fields have been previously referred to with regard to the register indexes XXX and BBB.

[0681] Displacement field 10162A (Bytes 7-10)—When MOD field 10242 contains 10, bytes 7-10 are the displacement field 10162A, and it works the same as the legacy 32-bit displacement (disp32) and works at byte granularity.

[0682] Displacement factor field 10162B (Byte 7)—When MOD field 10242 contains 01, byte 7 is the displacement factor field 10162B. The location of this field is that same as that of the legacy xs6 instruction set 8-bit displacement (disp8), which works at byte granularity. Since disp8 is sign extended, it can only address between –128 and 127 bytes offsets; in terms of 64 byte cache lines, disp8 uses 8 bits that can be set to only four really useful values –128, –64, 0, and 64; since a greater range is often needed, disp32 is used; however, disp32 requires 4 bytes. In contrast to disp8 and disp32, the displacement factor field 10162B is a reinterpretation of disp8; when using displacement factor field 10162B, the actual displacement is determined by the content of the displacement factor field multiplied by the size of the memory operand access (N). This type of displacement is referred to as disp8*N. This reduces the average instruction length (a single byte of used for the displacement but with a much greater range). Such compressed displacement is based on the assumption that the effective displacement is multiple of the granularity of the memory access, and hence, the redundant low-order bits of the address offset do not need to be encoded. In other words, the displacement factor field 10162B substitutes the legacy xs6 instruction set 8-bit displacement. Thus, the displacement factor field 10162B is encoded the same way as an xs6 instruction set 8-bit displacement (so no changes in the ModRM/SIB encoding rules) with the only exception that disp8 is overloaded to disp8*N. In other words, there are no changes in the encoding rules or encoding lengths but only in the interpretation of the displacement value by hardware (which needs
to scale the displacement by the size of the memory operand to obtain a byte-wise address offset. Immediate field 10172 operates as previously described.

Full Opcode Field

[0683] FIG. 102B is a block diagram illustrating the fields of the specific vector friendly instruction format 10200 that make up the full opcode field 10174 according to one embodiment of the disclosure. Specifically, the full opcode field 10174 includes the format field 10140, the base operation field 10142, and the data element width (W) field 10164. The base operation field 10142 includes the prefix encoding field 10225, the opcode map field 10215, and the real opcode field 10230.

Register Index Field

[0684] FIG. 102C is a block diagram illustrating the fields of the specific vector friendly instruction format 10200 that make up the register index field 10144 according to one embodiment of the disclosure. Specifically, the register index field 10144 includes the REX field 10205, the REX' field 10210, the MODR/M.reg field 10244, the MODR/M.r/m field 1026, the VVVV field 10220, xxx field 10254, and the bbb field 10256.

Augmentation Operation Field

[0685] FIG. 102D is a block diagram illustrating the fields of the specific vector friendly instruction format 10200 that make up the augmentation operation field 10150 according to one embodiment of the disclosure. When the class (U) field 10168 contains 0, it signifies EVEX.U1 (class B 10168B); when it contains 1, it signifies EVEX.U1 (class B 10168B). When U=0 and the MOD field 10242 contains 11 (signifying a no memory access operation), the alpha field 10152 (EVEX byte 3, bit [7]—EH) is interpreted as the rs field 10152A. When the rs field 10152A contains 0, it signifies EVEX.U1 (class A 10168A); when it contains 1, it signifies EVEX.U1 (class B 10168B). When U=0 and the MOD field 10242 contains 11 (signifying a no memory access operation), the alpha field 10152 (EVEX byte 3, bit [7]—EH) is interpreted as the rs field 10152A. When the rs field 10152A contains 0, it signifies EVEX.U1 (class A 10168A); when it contains 1, it signifies EVEX.U1 (class B 10168B). When U=0 and the MOD field 10242 contains 11 (signifying a no memory access operation), the alpha field 10152 (EVEX byte 3, bit [7]—EH) is interpreted as the rs field 10152A. When the rs field 10152A contains 0, it signifies EVEX.U1 (class A 10168A); when it contains 1, it signifies EVEX.U1 (class B 10168B).

Exemplary Register Architecture

[0687] FIG. 103 is a block diagram of a register architecture 10300 according to one embodiment of the disclosure. In the embodiment illustrated, there are 32 vector registers 10310 that are 512 bits wide; these registers are referenced as zmn0 through zmn31. The lower order 256 bits of the lower 16 zmn registers are overlaid on registers zmn0-16. The lower order 128 bits of the lower 16 zmn registers (the lower order 128 bits of the zmn registers) are overlaid on registers zmn0-15. The specific vector friendly instruction format 10200 operates on these overlaid register file as illustrated in the below tables.

<table>
<thead>
<tr>
<th>Adjustable</th>
<th>Class</th>
<th>Operations</th>
<th>Registers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instruction Templates</td>
<td>A (FIG. 5410, 10115)</td>
<td>101A, 10125, 10130</td>
<td>vector length is 64 byte</td>
</tr>
<tr>
<td>that do not include the vector length field</td>
<td>B (FIG. 101B)</td>
<td>5412</td>
<td>zmn registers (the vector length is 64 byte)</td>
</tr>
<tr>
<td>10159B</td>
<td>U = 1)</td>
<td>B (FIG. 5417, 10127)</td>
<td>zmn, ymm, or xmm registers (the vector length is 64 byte, 32 byte, or 16 byte)</td>
</tr>
<tr>
<td>Instruction templates that do include the vector length field 10159B</td>
<td>U = 1)</td>
<td>101B</td>
<td>depending on the vector length field 10159B</td>
</tr>
</tbody>
</table>

[0688] In other words, the vector length field 10159B selects between a maximum length and one or more other shorter lengths, where each such shorter length is half the length of the preceding length; and instructions templates without the vector length field 10159B operate on the maximum vector length. Further, in one embodiment, the class B instruction templates of the specific vector friendly instruction format 10200 operate on packed or scalar single/double-precision floating point data and packed or scalar integer data. Scalar operations are operations performed on the lowest order data element position in an zmn/ymm/xmm register; the higher order data element positions are either left the same as they were prior to the instruction or zeroed depending on the encoding.

[0689] Write mask registers 10315—in the embodiment illustrated, there are 8 write mask registers (k0 through k7), each 64 bits in size. In an alternate embodiment, the write mask registers 10315 are 16 bits in size. As previously described, in one embodiment of the disclosure, the vector mask register k0 cannot be used as a write mask; when the encoding that would normally indicate k0 is used for a write mask, it selects a hardware write mask of 0xFFFF, effectively disabling write masking for that instruction.

[0690] General-purpose registers 10325—in the embodiment illustrated, there are sixteen 64-bit general-purpose registers that are used along with the existing x86 addressing modes to address memory operands. These registers are referenced by the names RAX, RBX, RCX, RDX, RBP, RSI, RDI, RSP, and R8 through R15.

[0691] Scalar floating point stack register file (x87 stack) 10345, on which is aliased the MMX packed integer flat register file 10350—in the embodiment illustrated, the x87
stack is an eight-element stack used to perform scalar floating-point operations on 32/64/80-bit floating point data using the x87 instruction set extension; while the MMX registers are used to perform operations on 64-bit packed integer data, as well as to hold operands for some operations performed between the MMX and XMM registers.

Alternative embodiments of the disclosure may use wider or narrower registers. Additionally, alternative embodiments of the disclosure may use more, less, or different register files and registers.

Exemplary Core Architectures, Processors, and Computer Architectures

Processor cores may be implemented in different ways, for different purposes, and in different processors. For instance, implementations of such cores may include: 1) a general purpose in-order core intended for general-purpose computing; 2) a high performance general purpose out-of-order core intended for general-purpose computing; 3) a special purpose core intended primarily for graphics and/or scientific (throughput) computing. Implementations of different processors may include: 1) a CPU including one or more general purpose in-order cores intended for general-purpose computing and/or one or more general purpose out-of-order cores intended for general-purpose computing; and 2) a coprocessor including one or more special purpose cores intended primarily for graphics and/or scientific (throughput). Such different processors lead to different computer system architectures, which may include: 1) the coprocessor on a separate chip from the CPU; 2) the coprocessor on a separate die in the same package as a CPU; 3) the coprocessor on the same die as a CPU (in which case, such a coprocessor is sometimes referred to as a special purpose logic, such as integrated graphics and/or scientific (throughput) logic, or as special purpose cores); and 4) a system on a chip that may include on the same die the described CPU (sometimes referred to as the application core(s) or application processor(s)), the above described coprocessor, and additional functionality. Exemplary core architectures are described next, followed by descriptions of exemplary processors and computer architectures.

Exemplary Core Architectures

In-Order and Out-of-Order Core Block Diagram

FIG. 104A is a block diagram illustrating both an exemplary in-order pipeline and an exemplary register renaming, out-of-order issue/execution pipeline according to embodiments of the disclosure. FIG. 104B is a block diagram illustrating both an exemplary embodiment of an in-order architecture core and an exemplary register renaming, out-of-order issue/execution architecture core to be included in a processor according to embodiments of the disclosure. The solid lined boxes in FIGS. 104A-B illustrate the in-order pipeline and in-order core, while the optional addition of the dashed lined boxes illustrates the register renaming, out-of-order issue/execution pipeline and core. Given that the in-order aspect is a subset of the out-of-order aspect, the out-of-order aspect will be described.

In FIG. 104A, a processor pipeline stage 10400 includes a fetch stage 10402, a length decode stage 10404, a decode stage 10406, an allocation stage 10408, a renaming stage 10410, a scheduling (also known as a dispatch or issue) stage 10412, a register read/memory read stage 10414, an execute stage 10416, a write back/memory write stage 10418, an exception handling stage 10422, and a commit stage 10424.

FIG. 104B shows processor core 10490 including a front end unit 10430 coupled to an execution engine unit 10450, and both are coupled to a memory unit 10470. The core 10490 may be a reduced instruction set computing (RISC) core, a complex instruction set computing (CISC) core, a very long instruction word (VLIW) core, or a hybrid or alternative core type. As yet another option, the core 10490 may be a special-purpose core, such as, for example, a network or communication core, compression engine, coprocessor core, general purpose computing graphics processing unit (GPGPU) core, graphics core, or the like.

The front end unit 10430 includes a branch prediction unit 10432 coupled to an instruction cache unit 10434, which is coupled to an instruction translation lookaside buffer (TLB) 10436, which is coupled to an instruction fetch unit 10438, which is coupled to a decode unit 10440. The decode unit 10440 (or decoder or decoder unit) may decode instructions (e.g., macro-instructions), and generate as an output one or more micro-operations, micro-code entry points, micro-instructions, other instructions, or other control signals, which are decoded from, or which otherwise reflect, or are derived from, the original instructions. The decode unit 10440 may be implemented using various different mechanisms. Examples of suitable mechanisms include, but are not limited to, look-up tables, hardware implementations, programmable logic arrays (PLAs), microcode read only memories (ROMs), etc. In one embodiment, the core 10490 includes a microcode ROM or other medium that stores microcode for certain macro-instructions (e.g., in decode unit 10440 or otherwise within the front end unit 10430). The decode unit 10440 is coupled to a rename/allocator unit 10452 in the execution engine unit 10450.

The execution engine unit 10450 includes the rename/allocator unit 10452 coupled to a retirement unit 10454 and a set of one or more scheduler unit(s) 10456. The scheduler unit(s) 10456 represents any number of different schedulers, including reservations stations, central instruction window, etc. The scheduler unit(s) 10456 is coupled to the physical register file(s) unit(s) 10458. Each of the physical register file(s) units 10458 represents one or more physical register files, different ones of which store one or more different data types, such as scalar integer, scalar floating point, packed integer, packed floating point, vector integer, vector floating point, etc. (e.g., an instruction pointer that is the address of the next instruction to be executed), etc. In one embodiment, the physical register file(s) unit(s) 10458 comprises a vector registers unit, a write mask registers unit, and a scalar registers unit. These register units may provide architectural vector registers, vector mask registers, and general purpose registers. The physical register file(s) unit(s) 10458 is overlapped by the retirement unit 10454 to illustrate various ways in which register renaming and out-of-order execution may be implemented (e.g., using a reorder buffer(s) and a retirement register file(s); using a future file(s), a history buffer(s), and a retirement register file(s); using a register maps and a pool of registers, etc.). The retirement unit 10454 and the physical register file(s) unit(s) 10458 are coupled to the execution cluster(s) 10460. The execution cluster(s) 10460 includes a set of one or more execution units 10462 and a set of one or more memory
access units 10464. The execution units 10462 may perform various operations (e.g., shifts, addition, subtraction, multiplication) and on various types of data (e.g., scalar floating point, packed integer, integer vector, floating point). While some embodiments may include a number of execution units dedicated to specific functions or sets of functions, other embodiments may include only one execution unit or multiple execution units that all perform all functions. The scheduler unit(s) 10456, physical register file(s) unit(s) 10458, and execution cluster(s) 10460 are shown as being possibly plural because certain embodiments create separate pipelines for certain types of operations (e.g., a scalar integer pipeline, a scalar floating point/packed integer/packed floating point pipeline, and/or a memory access pipeline that are their own scheduler unit, physical register file(s) unit, and/or execution cluster—and in the case of a separate memory access pipeline, certain embodiments are implemented in which only the execution cluster of this pipeline has the memory access unit(s) 10464). It should also be understood that where separate pipelines are used, one or more of these pipelines may be out-of-order execution and the rest in-order.

The set of memory access units 10464 is coupled to the memory unit 10470, which includes a data TLB unit 10472 coupled to a data cache unit 10474 coupled to a level 2 (L2) cache unit 10476. In one exemplary embodiment, the memory access units 10464 may include a load unit, a store address unit, and a store data unit, each of which is coupled to the data TLB unit 10472 in the memory unit 10470. The instruction cache unit 10434 is further coupled to a level 2 (L2) cache unit 10476 in the memory unit 10470. The L2 cache unit 10476 is coupled to one or more other levels of cache and eventually to a main memory.

By way of example, the exemplary register renaming, out-of-order issue/exec core architecture may implement the pipeline 10400 as follows: 1) the instruction fetch 10438 performs the fetch and length decoding stages 10402 and 10404; 2) the decode unit 10404 performs the decode stage 10406; 3) the rename/alloc unit 10452 performs the allocation stage 10408 and renaming stage 10410; 4) the scheduler unit(s) 10456 performs the schedule stage 10412; 5) the physical register file(s) unit(s) 10458 and the memory unit 10470 perform the register read/memory read stage 10414; the execution cluster 10460 performs the execute stage 10416; 6) the memory unit 10470 and the physical register file(s) unit(s) 10458 perform the write back/memory write stage 10418; 7) various units may be involved in the exception handling stage 10422; and 8) the retirement unit 10454 and the physical register file(s) unit(s) 10458 perform the commit stage 10424.

The core 10490 may support one or more instructions (e.g., the x86 instruction set with some extensions that have been added with newer versions); the MIPS instruction set of MIPS Technologies of Sunnyvale, Calif.; the ARM instruction set (with optional additional extensions such as NEON) of ARM Holdings of Sunnyvale, Calif.), including the instruction(s) described herein. In one embodiment, the core 10490 includes logic to support a packed data instruction set extension (e.g., AVX1, AVX2), thereby allowing the operations used by many multimedia applications to be performed using packed data.

It should be understood that the core may support multithreading (executing two or more parallel sets of operations or threads), and may do so in a variety of ways including time sliced multithreading, simultaneous multithreading (where a single physical core provides a logical core for each of the threads that physical core is simultaneously multithreading), or a combination thereof (e.g., time sliced fetching and decoding and simultaneous multithreading thereafter such as in the Intel® Hyperthreading technology).

While register renaming is described in the context of out-of-order execution, it should be understood that register renaming may be used in an in-order architecture. While the illustrated embodiment of the processor also includes separate instruction and data cache units 10434/10474 and a shared L2 cache unit 10476, alternative embodiments may have a single internal cache for both instructions and data, such as, for example, a Level 1 (L1) internal cache, or multiple levels of internal cache. In some embodiments, the system may include a combination of an internal cache and an external cache that is external to the core and/or the processor. Alternatively, all of the cache may be external to the core and/or the processor.

Specific Exemplary In-Order Core Architecture

Figs. 105A-B illustrate a block diagram of a more specific exemplary in-order core architecture, which core would be one of several logic blocks (including other cores of the same type and/or different types) in a chip. The logic blocks communicate through a high-bandwidth interconnect network (e.g., a ring network) with some fixed function logic, memory I/O interfaces, and other necessary I/O logic, depending on the application.

Fig. 105A is a block diagram of a single processor core, along with its connection to the on-die interconnect network 10502 and with its local subset of the Level 2 (L2) cache 10504, according to embodiments of the disclosure. In one embodiment, an instruction decode unit 10500 supports the x86 instruction set with a packed data instruction set extension. An L1 cache 10506 allows low-latency accesses to cache memory into the scalar and vector units. While in one embodiment (to simplify the design), a scalar unit 10508 and a vector unit 10510 use separate register sets (respectively, scalar registers 10512 and vector registers 10514) and data transferred between them is written to memory and then read back in from a level 1 (L1) cache 10506, alternative embodiments of the disclosure may use a different approach (e.g., use a single register set or include a communication path that allow data to be transferred between the two register files without being written and read back).

The local subset of the L2 cache 10504 is part of a global L2 cache that is divided into separate local subsets, one per processor core. Each processor core has a direct access path to its own local subset of the L2 cache 10504. Data read by a processor core is stored in its L2 cache subset 10504 and can be accessed quickly, in parallel with other processor cores accessing their own local L2 cache subsets. Data written by a processor core is stored in its own L2 cache subset 10504 and is flushed from other subsets, if necessary. The ring network ensures coherency for shared data. The ring network is bi-directional to allow agents such as processor cores, L1 caches and other logic blocks to communicate with each other within the chip. Each ring data-path is 1012-bits wide per direction.

Fig. 105B is an expanded view of part of the processor core in Fig. 105A, according to embodiments of the disclosure. Fig. 105B includes an L1 data cache 10506A.
part of the L1 cache 10506, as well as more detail regarding the vector unit 10510 and the vector registers 10514. Specifically, the vector unit 10510 is a 16-wide vector processing unit (VPU) (see the 16-wide ALU 10528), which executes one or more of integer, single-precision float, and double-precision float instructions. The VPU supports swizzling the register inputs with swizzle unit 10520, numeric conversion with numeric convert units 10522-A, and replication with replication unit 10524 on the memory input. Write mask registers 10526 allow predicated resulting vector writes.

[0708] FIG. 106 is a block diagram of a processor 10600 that may have more than one core, may have an integrated memory controller, and may have integrated graphics according to embodiments of the disclosure. The solid lined boxes in FIG. 106 illustrate a processor 10600 with a single core 10602-A, a system agent 10610, a set of one or more bus controller units 10616, while the optional addition of the dashed lined boxes illustrates an alternative processor 10600 with multiple cores 10602-A-N, a set of one or more integrated memory controller unit(s) 10614 in the system agent unit 10610, and special purpose logic 10608.

[0709] Thus, different implementations of the processor 10600 may include: 1) a CPU with the special purpose logic 10608 being integrated graphics and/or scientific (throughput) logic (which may include one or more cores), and the cores 10602-A-N being one or more general purpose cores (e.g., general purpose in-order cores, general purpose out-of-order cores, a combination of the two); 2) a coprocessor with the cores 10602-A-N being a large number of special purpose cores intended primarily for graphics and/or scientific (throughput); and 3) a coprocessor with the cores 10602-A-N being a large number of general purpose in-order cores. Thus, the processor 10600 may be a general-purpose processor, coprocessor or special-purpose processor, such as, for example, a network or communication processor, compression engine, graphics processor, GPGPU (general purpose graphics processing unit), a high-throughput many integrated core (MIC) coprocessor (including 30 or more cores), embedded processor, or the like. The processor may be implemented on one or more chips. The processor 10600 may be a part of and/or may be implemented on one or more substrates using any of a number of process technologies, such as, for example, BICMOS, CMOS, or NMOS.

[0710] The memory hierarchy includes one or more levels of cache within the cores, a set or one or more shared cache units 10606, and external memory (not shown) coupled to the set of integrated memory controller units 10614. The set of shared cache units 10606 may include one or more mid-level caches, such as level 2 (L2), level 3 (L3), level 4 (L4), or other levels of cache, a last level cache (LLC), and/or combinations thereof. While in one embodiment a ring based interconnect unit 10612 interconnects the integrated graphics logic 10608, the set of shared cache units 10606, and the system agent unit 10610/integrated memory controller unit(s) 10614, alternative embodiments may use any number of well-known techniques for interconnecting such units. In one embodiment, coherency is maintained between one or more cache units 10606 and cores 10602-A-N.

[0711] In some embodiments, one or more of the cores 10602-A-N are capable of multi-threading. The system agent 10610 includes those components coordinating and operating cores 10602-A-N. The system agent unit 10610 may include for example a power control unit (PCU) and a display unit. The PCU may be or include logic and components needed for regulating the power state of the cores 10602-A-N and the integrated graphics logic 10608. The display unit is for driving one or more externally connected displays.

[0712] The cores 10602-A-N may be homogenous or heterogeneous in terms of architecture instruction set; that is, two or more of the cores 10602-A-N may be capable of executing the same instruction set, while others may be capable of executing only a subset of that instruction set or a different instruction set.

Exemplary Computer Architectures

[0713] FIGS. 107-110 are block diagrams of exemplary computer architectures. Other system designs and configurations known in the arts for laptops, desktops, handheld PCs, personal digital assistants, engineering workstations, servers, network devices, network hubs, switches, embedded processors, digital signal processors (DSPs), graphics devices, video game devices, set-top boxes, micro controllers, cell phones, portable media players, hand held devices, and various other electronic devices, are also suitable. In general, a huge variety of systems or electronic devices capable of incorporating a processor and/or other execution logic as disclosed herein are generally suitable.

[0714] Referring now to FIG. 107, shown is a block diagram of a system 10700 in accordance with one embodiment of the present disclosure. The system 10700 may include one or more processors 10710, 10715, which are coupled to a controller hub 10720. In one embodiment the controller hub 10720 includes a graphics memory controller hub (GMCH) 10790 and an Input/Output Hub (IOH) 10750 (which may be on separate chips); the GMCH 10790 includes memory and graphics controllers to which are coupled memory 10740 and a coprocessor 10745; the IOH 10750 is couples input/output (I/O) devices 10760 to the GMCH 10790. Alternatively, one or both of the memory and graphics controllers are integrated within the processor (as described herein), the memory 10740 and the coprocessor 10745 are coupled directly to the processor 10710, and the controller hub 10720 in a single chip with the IOH 10750. Memory 10740 may include a compiler module 10740A, for example, to store code that when executed causes a processor to perform any method of this disclosure.

[0715] The optional nature of additional processors 10715 is denoted in FIG. 107 with broken lines. Each processor 10710, 10715 may include one or more of the processing cores described herein and may be version of the processor 10600.

[0716] The memory 10740 may be, for example, dynamic random access memory (DRAM), phase change memory (PCM), or a combination of the two. For at least one embodiment, the controller hub 10720 communicates with the processor(s) 10710, 10715 via a multi-drop bus, such as a frontside bus (FSB), point-to-point interface such as QuickPath Interconnect (QPI), or similar connection 10795.

[0717] In one embodiment, the coprocessor 10745 is a special-purpose processor, such as, for example, a high-throughput MIC processor, a network or communication processor, compression engine, graphics processor, GPGPU, embedded processor, or the like. In one embodiment, controller hub 10720 may include an integrated graphics accelerator.
[0718] There can be a variety of differences between the physical resources 10710, 10715 in terms of a spectrum of metrics of merit including architectural, microarchitectural, thermal, power consumption characteristics, and the like.

[0719] In one embodiment, the processor 10710 executes instructions that control data processing operations of a general type. Embedded within the instructions may be coprocessor instructions. The processor 10710 recognizes these coprocessor instructions as being of a type that should be executed by the attached coprocessor 10745. Accordingly, the processor 10710 issues these coprocessor instructions (or control signals representing coprocessor instructions) on a coprocessor bus or other interconnect, to coprocessor 10745. Coprocessor(s) 10745 accept and execute the received coprocessor instructions.

[0720] Referring now to FIG. 108, shown is a block diagram of a first more specific exemplary system 10800 in accordance with an embodiment of the present disclosure. As shown in FIG. 108, multiprocessor system 10800 is a point-to-point interconnect system, and includes a first processor 10870 and a second processor 10880 coupled via a point-to-point interconnect 10850. Each of processors 10870 and 10880 may be some version of the processor 10600. In one embodiment of the disclosure, processors 10870 and 10880 are respectively processors 10710 and 10715, while coprocessor 10830 is coprocessor 10745. In another embodiment, processors 10870 and 10880 are respectively processor 10710 coprocessor 10745.

[0721] Processors 10870 and 10880 are shown including integrated memory controller (IMC) units 10872 and 10882, respectively. Processor 10870 also includes as part of its bus controller units point-to-point (P-P) interfaces 10876 and 10878; similarly, second processor 10880 includes P-P interfaces 10886 and 10888. Processors 10870, 10880 may exchange information via a point-to-point (P-P) interface 10850 using P-P interface circuits 10878, 10888. As shown in FIG. 108, IMCs 10872 and 10882 couple the processors to respective memories, namely a memory 10832 and a memory 10834, which may be portions of main memory locally attached to the respective processors.

[0722] Processors 10870, 10880 may each exchange information with a chipset 10890 via individual P-P interfaces 10852, 10854 using point to point interface circuits 10876, 10894, 10886, 10898. Chipset 10890 may optionally exchange information with the coprocessor 10838 via a high-performance interface 10839. In one embodiment, the coprocessor 10838 is a special-purpose processor, such as, for example, a high-throughput MIC processor, a network or communication processor, compression engine, graphics processor, GPGPU, embedded processor, or the like.

[0723] A shared cache (not shown) may be included in either processor or outside of both processors, yet connected with the processors via P-P interconnect, such that either or both processors’ local cache information may be stored in the shared cache if a processor is placed into a low power mode.

[0724] Chipset 10890 may be coupled to a first bus 10816 via an interface 10896. In one embodiment, first bus 10816 may be a Peripheral Component Interconnect (PCI) bus, or a bus such as a PCI Express bus or another third generation I/O interconnect bus, although the scope of the present disclosure is not so limited.

[0725] As shown in FIG. 108, various I/O devices 10814 may be coupled to first bus 10816, along with a bus bridge 10818 which couples first bus 10816 to a second bus 10820. In one embodiment, one or more additional processor(s) 10815, such as coprocessors, high-throughput MIC processors, GPGPU’s, accelerators such as, e.g., graphics accelerators or digital signal processing (DSP) units, field programmable gate arrays, or any other processor, are coupled to first bus 10816. In one embodiment, second bus 10820 may be a low pin count (LPC) bus. Various devices may be coupled to a second bus 10820 including, for example, a keyboard and/or mouse 10822, communication devices 10827 and a storage unit 10828 such as a disk drive or other mass storage device which may include instructions/code and data 10830, in one embodiment. Further, an audio I/O 10824 may be coupled to the second bus 10820. Note that other architectures are possible. For example, instead of the point-to-point architecture of FIG. 108, a system may implement a multi-drop bus or other such architecture.

[0726] Referring now to FIG. 109, shown is a block diagram of a second more specific exemplary system 10900 in accordance with an embodiment of the present disclosure. Like elements in FIGS. 108 and 109 bear like reference numerals, and certain aspects of FIG. 108 have been omitted from FIG. 109 in order to avoid obscuring other aspects of FIG. 109.

[0727] FIG. 109 illustrates that the processors 10870, 10880 may include integrated memory and I/O control logic (“CCL”) 10872 and 10882, respectively. Thus, the CL 10872, 10882 include integrated memory controller units and include I/O control logic. FIG. 109 illustrates that not only are the memories 10832, 10834 coupled to the CL 10872, 10882, but also that I/O devices 10914 are also coupled to the control logic 10872, 10882. Legacy I/O devices 10915 are coupled to the chipset 10890.

[0728] Referring now to FIG. 110, shown is a block diagram of a SoC 11000 in accordance with an embodiment of the present disclosure. Similar elements in FIG. 106 bear like reference numerals. Also, dashed lined boxes are optional features on more advanced SoCs. In FIG. 110, an interconnect unit(s) 11002 is coupled to: an application processor 11010 which includes a set of one or more cores 11060A-N and shared cache unit(s) 11060; a system agent unit 110610; a bus controller unit(s) 110610; an integrated memory controller unit(s) 110614; a set or one or more coprocessors 11020 which may include integrated graphics logic, an image processor, an audio processor, and a video processor; an static random access memory (SRAM) unit 11030; a direct memory access (DMA) unit 11032; and a display unit 11040 for coupling to one or more external displays. In one embodiment, the coprocessor(s) 11020 include a special-purpose processor, such as, for example, a network or communication processor, compression engine, GPGPU, a high-throughput MIC processor, embedded processor, or the like.

[0729] Embodiments (e.g., of the mechanisms) disclosed herein may be implemented in hardware, software, firmware, or a combination of such implementation approaches. Embodiments of the disclosure may be implemented as computer programs or program code executing on programmable systems comprising at least one processor, a storage system (including volatile and non-volatile memory and/or storage elements), at least one input device, and at least one output device.

[0730] Program code, such as code 10830 illustrated in FIG. 108, may be applied to input instructions to perform the
functions described herein and generate output information. The output information may be applied to one or more output devices, in known fashion. For purposes of this application, a processing system includes any system that has a processor, such as, for example; a digital signal processor (DSP), a microcontroller, an application specific integrated circuit (ASIC), or a microprocessor.

[0731] The program code may be implemented in a high level procedural or object oriented programming language to communicate with a processing system. The program code may also be implemented in assembly or machine language, if desired. In fact, the mechanisms described herein are not limited in scope to any particular programming language. In any case, the language may be a compiled or interpreted language.

[0732] One or more aspects of at least one embodiment may be implemented by representative instructions stored on a machine-readable medium which represents various logic within the processor, which when read by a machine causes the machine to fabricate logic to perform the techniques described herein. Such representations, known as “IP cores” may be stored on a tangible, machine readable medium and supplied to various customers or manufacturing facilities to load into the fabrication machines that actually make the logic or processor.

[0733] Such machine-readable storage media may include, without limitation, non-transitory, tangible arrangements of articles manufactured or formed by a machine or device, including storage media such as hard disks, any other type of disk including floppy disks, optical disks, compact disk read-only memories (CD-ROMs), compact disk rewritable’s (CD-RWs), and magneto-optical disks, semiconductor devices such as read-only memories (ROMs), random access memories (RAMs) such as dynamic random access memories (DRAMs), static random access memories (SRAMs), erasable programmable read-only memories (EPROMs), flash memories, electrically eraseable programmable read-only memories (EEPROMs), phase change memory (PCM), magnetic or optical cards, or any other type of media suitable for storing electronic instructions.

[0734] Accordingly, embodiments of the disclosure also include non-transitory, tangible machine-readable media containing instructions or containing design data, such as Hardware Description Language (HDL), which defines structures, circuits, apparatuses, processors and/or system features described herein. Such embodiments may also be referred to as program products.

Emulation (Including Binary Translation, Code Morphing, Etc.)

[0735] In some cases, an instruction converter may be used to convert an instruction from a source instruction set to a target instruction set. For example, the instruction converter may translate (e.g., using static binary translation, dynamic binary translation including dynamic compilation), morph, emulate, or otherwise convert an instruction to one or more other instructions to be processed by the core. The instruction converter may be implemented in software, hardware, firmware, or a combination thereof. The instruction converter may be on processor, off processor, or part on and part off processor.

[0736] FIG. 111 is a block diagram contrasting the use of a software instruction converter to convert binary instructions in a source instruction set to binary instructions in a target instruction set according to embodiments of the disclosure. In the illustrated embodiment, the instruction converter is a software instruction converter, although alternatively the instruction converter may be implemented in software, firmware, hardware, or various combinations thereof. FIG. 111 shows a program in a high level language 11102 may be compiled using an x86 compiler 11104 to generate x86 binary code 11106 that may be natively executed by a processor with at least one x86 instruction set core 11116. The processor with at least one x86 instruction set core 11116 represents any processor that can perform substantially the same functions as an Intel® processor with at least one x86 instruction set core by compatibly executing or otherwise processing (1) a substantial portion of the instruction set of the Intel® x86 instruction set core or (2) object code versions of applications or other software targeted to run on an Intel® processor with at least one x86 instruction set core, in order to achieve substantially the same result as an Intel® processor with at least one x86 instruction set core. The x86 compiler 11104 represents a compiler that is operable to generate x86 binary code 11106 (e.g., object code) that can, with or without additional linkage processing, be executed on the processor with at least one x86 instruction set core 11116. Similarly, FIG. 111 shows the program in the high level language 11102 may be compiled using an alternative instruction set compiler 11108 to generate alternative instruction set binary code 11110 that may be natively executed by a processor without at least one x86 instruction set core 11114 (e.g., a processor with cores that execute the MIPS instruction set of MIPS Technologies of Sunnyvale, Calif. and/or that execute the ARM instruction set of ARM Holdings of Sunnyvale, Calif.). The instruction converter 11112 is used to convert the x86 binary code 11106 into code that may be natively executed by the processor without an x86 instruction set core 11114. This converted code is not likely to be the same as the alternative instruction set binary code 11110 because an instruction converter capable of this is difficult to make; however, the converted code will accomplish the general operation and be made up of instructions from the alternative instruction set. Thus, the instruction converter 11112 represents software, firmware, hardware, or a combination thereof that, through emulation, simulation or any other process, allows a processor or other electronic device that does not have an x86 instruction set processor or core to execute the x86 binary code 11106.

What is claimed is:

1. An apparatus comprising:
   a plurality of processing elements that each comprise:
   a configuration register within a respective processing element to store a configuration value that causes the respective processing element to perform an operation according to the configuration value,
   a plurality of input queues,
   an input controller to control enqueue and dequeue of values into the plurality of input queues according to the configuration value,
   a plurality of output queues, and
   an output controller to control enqueue and dequeue of values into the plurality of output queues according to the configuration value;
   a circuit switched interconnect network between the plurality of processing elements to transfer values between the plurality of processing elements; and
an in-network storage element of the circuit switched interconnect network comprising:
a queue coupled to an output queue of a plurality of
output queues of a first processing element of the
plurality of processing elements,
a switch,
a configuration register of the in-network storage
element to store a configuration value, and
a controller that switches the switch into a first mode
that provides a value stored in the queue of the
in-network storage element by the output queue of
the first processing element to an input queue of a
plurality of input queues of a second processing
element of the plurality of processing elements when
the configuration value is a first value, and into a
second mode that bypasses the queue of the in-
network storage element and provides a value from
the output queue of the first processing element to
the input queue of the second processing element
when the configuration value is a second value.

2. The apparatus of claim 1, wherein the controller of the
in-network storage element is to send a valid out value to the
second processing element:
when the output queue of the first processing element
stores the value and the configuration value in the
configuration register of the in-network storage ele-
ment is the first value; and
when the queue of the in-network storage element stores
the value and the configuration value in the configura-
tion register of the in-network storage element is the
second value.

3. The apparatus of claim 2, wherein, when the input
queue of the second processing element stores the value
from the output queue of the first processing element or the
value from the queue of the in-network storage element, an
input controller of the second processing element is to send
a not empty value to operation circuitry of the second
processing element to indicate to the second processing
element to begin the operation on the value stored in the
input queue of the second processing element.

4. The apparatus of claim 2, wherein the controller of the
in-network storage element is to send a ready value to the
first processing element:
when the input queue of the second processing element is
not full and the configuration value in the configuration
register of the in-network storage element is the first
value; and
when the queue of the in-network storage element is not
full and the configuration value in the configuration
register of the in-network storage element is the second
value.

5. The apparatus of claim 4, wherein the output controller
of the first processing element dequeues the value from the
output queue of the first processing element after:
the ready value is received from the in-network storage
element by the first processing element; and
a valid out value is asserted by the first processing
element, through the in-network storage element, to the
second processing element when the configuration
value in the configuration register of the in-network
storage element is the first value, and asserted by the
first processing element to the in-network storage ele-
ment when the configuration value in the configuration
register of the in-network storage element is the second
value.

6. The apparatus of claim 1, wherein the controller of the
in-network storage element is to send a ready value to the
first processing element:
when the input queue of the second processing element is
not full and the configuration value in the configuration
register of the in-network storage element is the first
value; and
when the queue of the in-network storage element is not
full and the configuration value in the configuration
register of the in-network storage element is the second
value.

7. The apparatus of claim 6, wherein:
when the configuration value in the configuration register
of the in-network storage element is the first value, the
input controller of the second processing element stores
the value within the input queue of the second process-
ing element in response to receipt of the ready value;
and
when the configuration value in the configuration register
of the in-network storage element is the second value,
the controller of the in-network storage element stores
the value within the queue of the in-network storage
element in response to receipt of the ready value.

8. The apparatus of claim 1, wherein the in-network
storage element comprises a second switch, and the con-
figuration value in the configuration register of the in-
network storage element includes a network select field
to switch between a first network having a first set of the first
processing element and the second processing element when
the network select field is a first value, and a second network
having a second, different set of the first processing element
and the second processing element when the network select
field is a second value.

9. A method comprising:
storing a first configuration value in a configuration
register within a first processing element that causes the
first processing element to perform an operation
according to the first configuration value, and a second
configuration value in a configuration register within a
second processing element that causes the second pro-
cessing element to perform an operation according to
the second configuration value;
coupling a queue of an in-network storage element, of a
circuit switched interconnect network between the first
processing element and the second processing element,
to an output queue of a plurality of output queues of the
first processing element;
controlling enqueue and dequeue of values into the plu-
rality of output queues of the first processing element;
controlling enqueue and dequeue of values into a plurality
of input queues of the second processing element
according to the first configuration value with an output
controller in the first processing element;
controlling enqueue and dequeue of values into a plurality
of input queues of the second processing element
according to the second configuration value with an input
controller in the second processing element;
storing a third configuration value in a configuration
register within the in-network storage element; and
switching the in-network storage element between a first
mode that provides a value stored in the queue of the
in-network storage element by the output queue of the
first processing element to an input queue of the
plurality of input queues of the second processing element when the third configuration value is a first value, and into a second mode that bypasses the queue of the in-network storage element and provides a value from the output queue of the first processing element to the input queue of the second processing element when the third configuration value is a second value.

10. The method of claim 9, further comprising sending a valid out value to the second processing element from the in-network storage element:
   when the output queue of the first processing element stores the value and the third configuration value in the configuration register of the in-network storage element is the first value; and
   when the queue of the in-network storage element stores the value and the third configuration value in the configuration register of the in-network storage element is the second value.

11. The method of claim 10, further comprising sending a not empty value to operation circuitry of the second processing element from the input controller of the second processing element to indicate to the second processing element to begin the operation on the value stored in the input queue of the second processing element when the input queue of the second processing element stores the value from the output queue of the first processing element or the value from the queue of the in-network storage element.

12. The method of claim 10, further comprising sending a ready value to the first processing element from the in-network storage element:
   when the input queue of the second processing element is not full and the third configuration value in the configuration register of the in-network storage element is the first value; and
   when the queue of the in-network storage element is not full and the third configuration value in the configuration register of the in-network storage element is the second value.

13. The method of claim 12, further comprising the output controller of the first processing element dequeuing the value from the output queue of the first processing element after:
   the ready value is received from the in-network storage element by the first processing element; and
   a valid out value is asserted by the first processing element, through the in-network storage element, to the second processing element when the third configuration value in the configuration register of the in-network storage element is the first value, and asserted by the first processing element to the in-network storage element when the third configuration value in the configuration register of the in-network storage element is the second value.

14. The method of claim 9, further comprising sending a ready value to the first processing element from the in-network storage element:
   when the input queue of the second processing element is not full and the third configuration value in the configuration register of the in-network storage element is the first value; and
   when the queue of the in-network storage element is not full and the third configuration value in the configuration register of the in-network storage element is the second value.

15. The method of claim 14, further comprising:
   when the third configuration value in the configuration register of the in-network storage element is the first value, the input controller of the second processing element storing the value within the input queue of the second processing element in response to receipt of the ready value; and
   when the third configuration value in the configuration register of the in-network storage element is the second value, the in-network storage element storing the value within the queue of the in-network storage element in response to receipt of the ready value.

16. The method of claim 9, further comprising switching the in-network storage element between a first network having a first set of the first processing element and the second processing element when a network select field of the third configuration value is a first value, and a second network having a second, different set of the first processing element and the second processing element when the network select field of the third configuration value is a second value.

17. A processor comprising:
   a core with a decoder to decode an instruction into a decoded instruction and an execution unit to execute the decoded instruction; and
   a hardware accelerator coupled to the core, the hardware accelerator comprising:
   a plurality of processing elements that each comprise:
   a configuration register within a respective processing element to store a configuration value that causes the respective processing element to perform an operation according to the configuration value,
   a plurality of input queues,
   an input controller to control enqueue and dequeue of values into the plurality of input queues according to the configuration value,
   a plurality of output queues, and
   an output controller to control enqueue and dequeue of values into the plurality of output queues according to the configuration value,
   a circuit switched interconnect network between the plurality of processing elements to transfer values between the plurality of processing elements, and
   an in-network storage element of the circuit switched interconnect network comprising:
   a queue coupled to an output queue of a plurality of output queues of a first processing element of the plurality of processing elements,
   a switch,
   a configuration register of the in-network storage element to store a configuration value, and
   a controller that switches the switch into a first mode that provides a value stored in the queue of the in-network storage element to the output queue of the first processing element to an input queue of a plurality of input queues of a second processing element of the plurality of processing elements when the configuration value is a first value, and into a second mode that bypasses the queue of the in-network storage element and provides a value from the output queue of the first processing
element to the input queue of the second processing element when the configuration value is a second value.

18. The processor of claim 17, wherein the controller of the in-network storage element is to send a valid out value to the second processing element:

- when the output queue of the first processing element stores the value and the configuration value in the configuration register of the in-network storage element is the first value; and
- when the queue of the in-network storage element stores the value and the configuration value in the configuration register of the in-network storage element is the second value.

19. The processor of claim 18, wherein, when the input queue of the second processing element stores the value from the output queue of the first processing element or the value from the queue of the in-network storage element, an input controller of the second processing element is to send a not empty value to operation circuitry of the second processing element to indicate to the second processing element to begin the operation on the value stored in the input queue of the second processing element.

20. The processor of claim 18, wherein the controller of the in-network storage element is to send a ready value to the first processing element:

- when the input queue of the second processing element is not full and the configuration value in the configuration register of the in-network storage element is the first value; and
- when the queue of the in-network storage element is not full and the configuration value in the configuration register of the in-network storage element is the second value.

21. The processor of claim 20, wherein the output controller of the first processing element dequeues the value from the output queue of the first processing element after: the ready value is received from the in-network storage element by the first processing element; and
- a valid out value is asserted by the first processing element, through the in-network storage element, to the second processing element when the configuration value in the configuration register of the in-network storage element is the first value, and asserted by the first processing element to the in-network storage element when the configuration value in the configuration register of the in-network storage element is the second value.

22. The processor of claim 17, wherein the controller of the in-network storage element is to send a ready value to the first processing element:

- when the input queue of the second processing element is not full and the configuration value in the configuration register of the in-network storage element is the first value; and
- when the queue of the in-network storage element is not full and the configuration value in the configuration register of the in-network storage element is the second value.

23. The processor of claim 22, wherein:

- when the configuration value in the configuration register of the in-network storage element is the first value, the input controller of the second processing element stores the value within the input queue of the second processing element in response to receipt of the ready value; and
- when the configuration value in the configuration register of the in-network storage element is the second value, the controller of the in-network storage element stores the value within the queue of the in-network storage element in response to receipt of the ready value.

24. The processor of claim 17, wherein the in-network storage element comprises a second switch, and the configuration value in the configuration register of the in-network storage element includes a network select field to switch between a first network having a first set of the first processing element and the second processing element when the network select field is a first value, and a second network having a second, different set of the first processing element and the second processing element when the network select field is a second value.

* * * * *