(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 April 2024 (18.04.2024)

WIPO | PCT

(10) International Publication Number
WO 2024/079687 A1

(51) International Patent Classification:
G06F 8/41 (2018.01)

(21) International Application Number:
PCT/IB2023/060303

(22) International Filing Date:
12 October 2023 (12.10.2023)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/415,305 12 October 2022 (12.10.2022) US

(71) Applicant: **MOBILEYE VISION TECHNOLOGIES LTD.** [IL/IL]; 13 Hartom St., 9777513 Jerusalem (IL).

(72) Inventors: **MARJIEH, Michael**; 22 Safuriya, 1650622 Nazareth (IL). **KOM, Alon**; 8 Zohar Street, 3540234 Haifa (IL). **BENITA BEN-SIMHON, Oren**; 38 Abba Hushi Street, 3478117 Haifa (IL).

(74) Agent: **SHICHRUR, Naim Avraham**; SHICHRUR & CO., Horowitz Tower, Suite 521, 2 Moti Kind Street, 7638523 Rehovot (IL).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH,

(54) Title: APPARATUS, SYSTEM, AND METHOD OF COMPILING CODE FOR A PROCESSOR

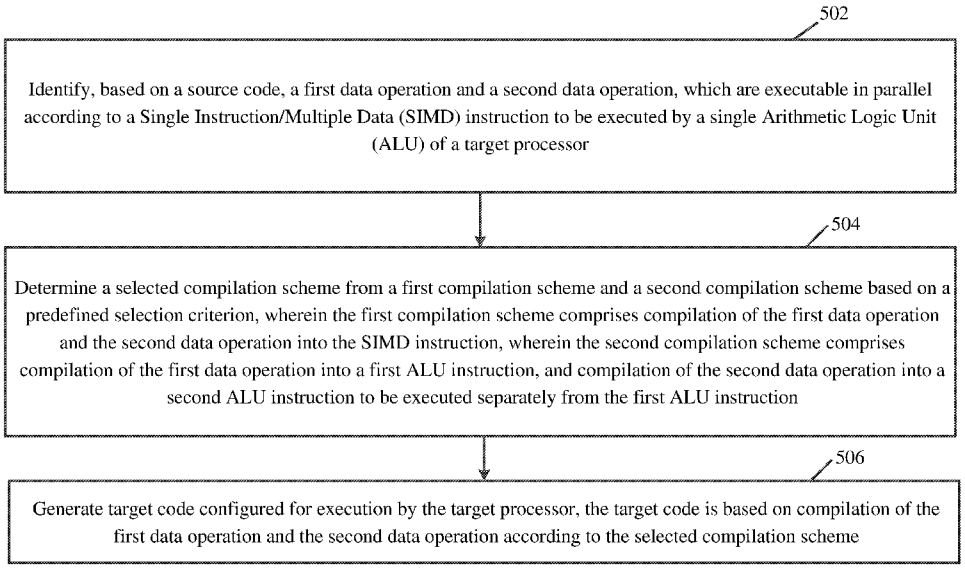

Fig. 5

(57) **Abstract:** For example, a compiler may be configured to identify a first data operation and a second data operation, which are executable in parallel according to a SIMD instruction to be executed by a single ALU of a target processor; to determine a selected compilation scheme from a first compilation scheme and a second compilation scheme based on a predefined selection criterion, wherein the first compilation scheme includes compilation of the first data operation and the second data operation into the SIMD instruction, wherein the second compilation scheme includes compilation of the first data operation into a first ALU instruction, and compilation of the second data operation into a second ALU instruction to be executed separately from the first ALU instruction; and to generate target code based on compilation of the first data operation and the second data operation according to the selected compilation scheme.

WO 2024/079687 A1

[Continued on next page]

TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
— *of inventorship (Rule 4.17(iv))*

Published:

— *with international search report (Art. 21(3))*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*
— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

APPARATUS, SYSTEM, AND METHOD OF COMPILING CODE FOR A PROCESSOR

CROSS REFERENCE

[0001] This Application claims the benefit of and priority from US Provisional Patent Application No. 63/415,305 entitled "APPARATUS, SYSTEM, AND METHOD OF COMPILING CODE FOR A PROCESSOR", filed October 12, 2022, the entire disclosure of which is incorporated herein by reference.

BACKGROUND

[0002] A compiler may be configured to compile source code into target code configured for execution by a processor.

[0003] There is a need to provide a technical solution to support efficient processing functionalities.

BRIEF DESCRIPTION OF THE DRAWINGS

[0004] For simplicity and clarity of illustration, elements shown in the figures have not necessarily been drawn to scale. For example, the dimensions of some of the elements may be exaggerated relative to other elements for clarity of presentation. Furthermore, reference numerals may be repeated among the figures to indicate corresponding or analogous elements. The figures are listed below.

[0005] Fig. 1 is a schematic block diagram illustration of a system, in accordance with some demonstrative aspects.

[0006] Fig. 2 is a schematic illustration of a compiler, in accordance with some demonstrative aspects.

[0007] Fig, 3 is a schematic illustration of a vector processor, in accordance with some demonstrative aspects.

[0008] Fig. 4 is a schematic flow-chart illustration of a method of compiling code for a processor, in accordance with some demonstrative aspects.

[0009] Fig. 5 is a schematic flow-chart illustration of a method of compiling code for a processor, in accordance with some demonstrative aspects.

[00010] Fig. 6 is a schematic illustration of a product, in accordance with some demonstrative aspects.

DETAILED DESCRIPTION

[00011] In the following detailed description, numerous specific details are set forth in order to provide a thorough understanding of some aspects. However, it will be understood by persons of ordinary skill in the art that some aspects may be practiced without these specific details. In other instances, well-known methods, procedures, components, units and/or circuits have not been described in detail so as not to obscure the discussion.

[00012] Some portions of the following detailed description are presented in terms of algorithms and symbolic representations of operations on data bits or binary digital signals within a computer memory. These algorithmic descriptions and representations may be the techniques used by those skilled in the data processing arts to convey the substance of their work to others skilled in the art.

[00013] An algorithm is here, and generally, considered to be a self-consistent sequence of acts or operations leading to a desired result. These include physical manipulations of physical quantities. Usually, though not necessarily, these quantities capture the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated. It has proven convenient at times, principally for reasons of common usage, to refer to these signals as bits, values, elements, symbols, characters, terms, numbers or the like. It should be understood, however, that all of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities.

[00014] Discussions herein utilizing terms such as, for example, “processing”, “computing”, “calculating”, “determining”, “establishing”, “analyzing”, “checking”, or the like, may refer to operation(s) and/or process(es) of a computer, a computing platform, a computing system, or other electronic computing device, that manipulate and/or transform data represented as physical (e.g., electronic) quantities within the computer’s registers and/or memories into other data similarly represented as physical quantities within the computer’s registers and/or memories or other information storage medium that may store instructions to perform operations and/or processes.

[00015] The terms “plurality” and “a plurality”, as used herein, include, for example, “multiple” or “two or more”. For example, “a plurality of items” includes two or more

items.

[00016] References to “one aspect”, “an aspect”, “demonstrative aspect”, “various aspects” etc., indicate that the aspect(s) so described may include a particular feature, structure, or characteristic, but not every aspect necessarily includes the particular feature, structure, or characteristic. Further, repeated use of the phrase “in one aspect” does not necessarily refer to the same aspect, although it may.

[00017] As used herein, unless otherwise specified the use of the ordinal adjectives “first”, “second”, “third” etc., to describe a common object, merely indicate that different instances of like objects are being referred to, and are not intended to imply that the objects so described must be in a given sequence, either temporally, spatially, in ranking, or in any other manner.

[00018] Some aspects, for example, may capture the form of an entirely hardware aspect, an entirely software aspect, or an aspect including both hardware and software elements. Some aspects may be implemented in software, which includes but is not limited to firmware, resident software, microcode, or the like.

[00019] Furthermore, some aspects may capture the form of a computer program product accessible from a computer-usable or computer-readable medium providing program code for use by or in connection with a computer or any instruction execution system. For example, a computer-usable or computer-readable medium may be or may include any apparatus that can contain, store, communicate, propagate, or transport the program for use by or in connection with the instruction execution system, apparatus, or device.

[00020] In some demonstrative aspects, the medium may be an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system (or apparatus or device) or a propagation medium.

[00021] In some demonstrative aspects, a data processing system suitable for storing and/or executing program code may include at least one processor coupled directly or indirectly to memory elements, for example, through a system bus. The memory elements may include, for example, local memory employed during actual execution of the program code, bulk storage, and cache memories which may provide temporary storage of at least some program code in order to reduce the number of times code must

be retrieved from bulk storage during execution.

[00022] In some demonstrative aspects, input/output or I/O devices (including but not limited to keyboards, displays, pointing devices, etc.) may be coupled to the system either directly or through intervening I/O controllers. In some demonstrative aspects, network adapters may be coupled to the system to enable the data processing system to become coupled to other data processing systems or remote printers or storage devices, for example, through intervening private or public networks. In some demonstrative aspects, modems, cable modems and Ethernet cards are demonstrative examples of types of network adapters. Other suitable components may be used.

[00023] Some aspects may be used in conjunction with various devices and systems, for example, a computing device, a computer, a mobile computer, a non-mobile computer, a server computer, or the like.

[00024] As used herein, the term "circuitry" may refer to, be part of, or include, an Application Specific Integrated Circuit (ASIC), an integrated circuit, an electronic circuit, a processor (shared, dedicated or group), and/or memory (shared. Dedicated, or group), that execute one or more software or firmware programs, a combinational logic circuit, and/or other suitable hardware components that provide the described functionality. In some aspects, some functions associated with the circuitry may be implemented by, one or more software or firmware modules. In some aspects, circuitry may include logic, at least partially operable in hardware.

[00025] The term “logic” may refer, for example, to computing logic embedded in circuitry of a computing apparatus and/or computing logic stored in a memory of a computing apparatus. For example, the logic may be accessible by a processor of the computing apparatus to execute the computing logic to perform computing functions and/or operations. In one example, logic may be embedded in various types of memory and/or firmware, e.g., silicon blocks of various chips and/or processors. Logic may be included in, and/or implemented as part of, various circuitry, e.g., processor circuitry, control circuitry, and/or the like. In one example, logic may be embedded in volatile memory and/or non-volatile memory, including random access memory, read only memory, programmable memory, magnetic memory, flash memory, persistent memory, and the like. Logic may be executed by one or more processors using memory, e.g., registers, stuck, buffers, and/or the like, coupled to the one or more processors,

e.g., as necessary to execute the logic.

[00026] Reference is now made to Fig. 1, which schematically illustrates a block diagram of a system 100, in accordance with some demonstrative aspects.

[00027] As shown in Fig. 1, in some demonstrative aspects system 100 may include a computing device 102.

[00028] In some demonstrative aspects, device 102 may be implemented using suitable hardware components and/or software components, for example, processors, controllers, memory units, storage units, input units, output units, communication units, operating systems, applications, or the like.

[00029] In some demonstrative aspects, device 102 may include, for example, a computer, a mobile computing device, a non-mobile computing device, a laptop computer, a notebook computer, a tablet computer, a handheld computer, a Personal Computer (PC), or the like.

[00030] In some demonstrative aspects, device 102 may include, for example, one or more of a processor 191, an input unit 192, an output unit 193, a memory unit 194, and/or a storage unit 195. Device 102 may optionally include other suitable hardware components and/or software components. In some demonstrative aspects, some or all of the components of one or more of device 102 may be enclosed in a common housing or packaging, and may be interconnected or operably associated using one or more wired or wireless links. In other aspects, components of one or more of device 102 may be distributed among multiple or separate devices.

[00031] In some demonstrative aspects, processor 191 may include, for example, a Central Processing Unit (CPU), a Digital Signal Processor (DSP), one or more processor cores, a single-core processor, a dual-core processor, a multiple-core processor, a microprocessor, a host processor, a controller, a plurality of processors or controllers, a chip, a microchip, one or more circuits, circuitry, a logic unit, an Integrated Circuit (IC), an Application-Specific IC (ASIC), or any other suitable multi-purpose or specific processor or controller. Processor 191 may execute instructions, for example, of an Operating System (OS) of device 102 and/or of one or more suitable applications.

[00032] In some demonstrative aspects, input unit 192 may include, for example, a

keyboard, a keypad, a mouse, a touch-screen, a touch-pad, a track-ball, a stylus, a microphone, or other suitable pointing device or input device. Output unit 193 may include, for example, a monitor, a screen, a touch-screen, a flat panel display, a Light Emitting Diode (LED) display unit, a Liquid Crystal Display (LCD) display unit, a plasma display unit, one or more audio speakers or earphones, or other suitable output devices.

[00033] In some demonstrative aspects, memory unit 194 includes, for example, a Random Access Memory (RAM), a Read Only Memory (ROM), a Dynamic RAM (DRAM), a Synchronous DRAM (SD-RAM), a flash memory, a volatile memory, a non-volatile memory, a cache memory, a buffer, a short term memory unit, a long term memory unit, or other suitable memory units. Storage unit 195 may include, for example, a hard disk drive, a Solid State Drive (SSD), or other suitable removable or non-removable storage units. Memory unit 194 and/or storage unit 195, for example, may store data processed by device 102.

[00034] In some demonstrative aspects, device 102 may be configured to communicate with one or more other devices via at least one network 103, e.g., a wireless and/or wired network.

[00035] In some demonstrative aspects, network 103 may include a wired network, a local area network (LAN), a wireless network, a wireless LAN (WLAN) network, a radio network, a cellular network, a WiFi network, an IR network, a Bluetooth (BT) network, and the like.

[00036] In some demonstrative aspects, device 102 may be configured to perform and/or to execute one or more operations, modules, processes, procedures and/or the like, e.g., as described herein.

[00037] In some demonstrative aspects, device 102 may include a compiler 160, which may be configured to generate a target code 115, for example, based on a source code 112, e.g., as described below.

[00038] In some demonstrative aspects, compiler 160 may be configured to translate the source code 112 into the target code 115, e.g., as described below.

[00039] In some demonstrative aspects, compiler 160 may include, or may be implemented as, software, a software module, an application, a program, a subroutine,

instructions, an instruction set, computing code, words, values, symbols, and/or the like.

[00040] In some demonstrative aspects, the source code 112 may include computer code written in a source language.

[00041] In some demonstrative aspects, the source language may include a programing language. For example, the source language may include a high-level programming language, for example, such as, C language, C++ language, and/or the like.

[00042] In some demonstrative aspects, the target code 115 may include computer code written in a target language.

[00043] In some demonstrative aspects, the target language may include a low-level language, for example, such as, assembly language, object code, machine code, or the like.

[00044] In some demonstrative aspects, the target code 115 may include one or more object files, e.g., which may create and/or form an executable program.

[00045] In some demonstrative aspects, the executable program may be configured to be executed on a target computer. For example, the target computer may include a specific computer hardware, a specific machine, and/or a specific operating system.

[00046] In some demonstrative aspects, the executable program may be configured to be executed on a processor 180, e.g., as described below.

[00047] In some demonstrative aspects, processor 180 may include a vector processor 180, e.g., as described below. In other aspects, processor 180 may include any other type of processor.

[00048] Some demonstrative aspects are described herein with respect to a compiler, e.g., compiler 160, configured to compile source code 112 into target code 115 configured to be executed by a vector processor 180, e.g., as described below. In other aspects, a compiler, e.g., compiler 160, configured to compile source code 112 into target code 115 configured to be executed by any other type of processor 180.

[00049] In some demonstrative aspects, processor 180 may be implemented as part of device 102.

[00050] In other aspects, processor 180 may be implemented as part of any other

device, e.g., separate from device 102.

[00051] In some demonstrative aspects, vector processor 180 (also referred to as an "array processor") may include a processor, which may be configured to process an entire vector in one instruction, e.g., as described below.

[00052] In other aspects, the executable program may be configured to be executed on any other additional or alternative type of processor.

[00053] In some demonstrative aspects, the vector processor 180 may be designed to support high-performance image and/or vector processing. For example, the vector processor 180 may be configured to processes 1/2/3/4D arrays of fixed point data and/or floating point arrays, e.g., very quickly and/or efficiently.

[00054] In some demonstrative aspects, the vector processor 180 may be configured to process arbitrary data, e.g., structures with pointers to structures. For example, the vector processor 180 may include a scalar processor to compute the non-vector data, for example, assuming the non-vector data is minimal.

[00055] In some demonstrative aspects, compiler 160 may be implemented as a local application to be executed by device 102. For example, memory unit 194 and/or storage unit 195 may store instructions resulting in compiler 160, and/or processor 191 may be configured to execute the instructions resulting in compiler 160 and/or to perform one or more calculations and/or processes of compiler 160, e.g., as described below.

[00056] In other aspects, compiler 160 may include a remote application to be executed by any suitable computing system, e.g., a server 170.

[00057] In some demonstrative aspects, server 170 may include at least a remote server, a web-based server, a cloud server, and/or any other server.

[00058] In some demonstrative aspects, the server 170 may include a suitable memory and/or storage unit 174 having stored thereon instructions resulting in compiler 160, and a suitable processor 171 to execute the instructions, e.g., as descried below.

[00059] In some demonstrative aspects, compiler 160 may include a combination of a remote application and a local application.

[00060] In one example, compiler 160 may be downloaded and/or received by the user of device 102 from another computing system, e.g., server 170, such that compiler

160 may be executed locally by users of device 102. For example, the instructions may be received and stored, e.g., temporarily, in a memory or any suitable short-term memory or buffer of device 102, e.g., prior to being executed by processor 191 of device 102.

[00061] In another example, compiler 160 may include a client-module to be executed locally by device 102, and a server module to be executed by server 170. For example, the client-module may include and/or may be implemented as a local application, a web application, a web site, a web client, e.g., a Hypertext Markup Language (HTML) web application, or the like.

[00062] For example, one or more first operations of compiler 160 may be performed locally, for example, by device 102, and/or one or more second operations of compiler 160 may be performed remotely, for example, by server 170.

[00063] In other aspects, compiler 160 may include, or may be implemented by, any other suitable computing arrangement and/or scheme.

[00064] In some demonstrative aspects, system 100 may include an interface 110, e.g., a user interface, to interface between a user of device 102 and one or more elements of system 100, e.g., compiler 160.

[00065] In some demonstrative aspects, interface 110 may be implemented using any suitable hardware components and/or software components, for example, processors, controllers, memory units, storage units, input units, output units, communication units, operating systems, and/or applications.

[00066] In some aspects, interface 110 may be implemented as part of any suitable module, system, device, or component of system 100.

[00067] In other aspects, interface 110 may be implemented as a separate element of system 100.

[00068] In some demonstrative aspects, interface 110 may be implemented as part of device 102. For example, interface 110 may be associated with and/or included as part of device 102.

[00069] In one example, interface 110 may be implemented, for example, as middleware, and/or as part of any suitable application of device 102. For example,

interface 110 may be implemented as part of compiler 160 and/or as part of an OS of device 102.

[00070] In some demonstrative aspects, interface 110 may be implemented as part of server 170. For example, interface 110 may be associated with and/or included as part of server 170.

[00071] In one example, interface 110 may include, or may be part of a Web-based application, a web-site, a web-page, a plug-in, an ActiveX control, a rich content component, e.g., a Flash or Shockwave component, or the like.

[00072] In some demonstrative aspects, interface 110 may be associated with and/or may include, for example, a gateway (GW) 113 and/or an Application Programming Interface (API) 114, for example, to communicate information and/or communications between elements of system 100 and/or to one or more other, e.g., internal or external, parties, users, applications and/or systems.

[00073] In some aspects, interface 110 may include any suitable Graphic-User-Interface (GUI) 116 and/or any other suitable interface.

[00074] In some demonstrative aspects, interface 110 may be configured to receive the source code 112, for example, from a user of device 102, e.g., via GUI 116, and/or API 114.

[00075] In some demonstrative aspects, interface 110 may be configured to transfer the source code 112, for example, to compiler 160, for example, to generate the target code 115, e.g., as described below.

[00076] Reference is made to Fig. 2, which schematically illustrates a compiler 200, in accordance with some demonstrative aspects. For example, compiler 160 (Fig. 1) may be implement one or more elements of compiler 200, and/or may perform one or more operations and/or functionalities of compiler 200.

[00077] In some demonstrative aspects, as shown in Fig. 2, compiler 200 may be configured to generate a target code 233, for example, by compiling a source code 212 in a source language.

[00078] In some demonstrative aspects, as shown in Fig. 2, compiler 200 may include a front-end 210 configured to receive and analyze the source code 212 in the

source language.

[00079] In some demonstrative aspects, front-end 210 may be configured to generate an intermediate code 213, for example, based on the source code 212.

[00080] In some demonstrative aspects, intermediate code 213 may include a lower level representation of the source code 212.

[00081] In some demonstrative aspects, front-end 210 may be configured to perform, for example, lexical analysis, syntax analysis, semantic analysis, and/or any other additional or alternative type of analysis, of the source code 212.

[00082] In some demonstrative aspects, front-end 210 may be configured to identify errors and/or problems with an outcome of the analysis of the source code 212. For example, front-end 210 may be configured to generate error information, e.g., including error and/or warning messages, for example, which may identify a location in the source code 212, for example, where an error or a problem is detected.

[00083] In some demonstrative aspects, as shown in Fig. 2, compiler 200 may include a middle-end 220 configured to receive and process the intermediate code 213, and to generate an adjusted, e.g., optimized, intermediate code 223.

[00084] In some demonstrative aspects, middle-end 220 may be configured to perform one or more adjustment, e.g., optimizations, to the intermediate code 213, for example, to generate the adjusted intermediate code 223.

[00085] In some demonstrative aspects, middle-end 220 may be configured to perform the one or more optimizations on the intermediate code 213, for example, independent of a type of the target computer to execute the target code 233.

[00086] In some demonstrative aspects, middle-end 220 may be implemented to support use of the optimized intermediate code 223, for example, for different machine types.

[00087] In some demonstrative aspects, middle-end 220 may be configured to optimize the intermediate representation of the intermediate code 223, for example, to improve performance and/or quality of the produced target code 233.

[00088] In some demonstrative aspects, the one or more optimizations of the intermediate code 213, may include, for example, inline expansion, dead-code

elimination, constant propagation, loop transformation, parallelization, and/or the like.

[00089] In some demonstrative aspects, as shown in Fig. 2, compiler 200 may include a back-end 230 configured to receive and process the adjusted intermediate code 213, and to generate the target code 233 based on the adjusted intermediate code 213.

[00090] In some demonstrative aspects, back-end 230 may be configured to perform one or more operations and/or processes, which may be specific for the target computer to execute the target code 233. For example, back-end 230 may be configured to process the optimized intermediate code 213 by applying to the adjusted intermediate code 213 analysis, transformation, and/or optimization operations, which may be configured, for example, based on the target computer to execute the target code 233.

[00091] In some demonstrative aspects, the one or more analysis, transformation, and/or optimization operations applied to the adjusted intermediate code 213 may include, for example, resource and storage decisions, e.g., register allocation, instruction scheduling, and/or the like.

[00092] In some demonstrative aspects, the target code 233 may include target-dependent assembly code, which may be specific to the target computer and/or a target operating system of the target computer, which is to execute the target code 233.

[00093] In some demonstrative aspects, the target code 233 may include target-dependent assembly code for a processor, e.g., vector processor 180 (Fig. 1).

[00094] In some demonstrative aspects, compiler 200 may include a Vector Micro-Code Processor (VMP) Open Computing Language (OpenCL) compiler, e.g., as described below. In other aspects, compiler 200 may include, or may be implemented as part of, any other type of vector processor compiler.

[00095] In some demonstrative aspects, the VMP OpenCL compiler may include a Low Level Virtual Machine (LLVM) based (LLVM-based) compiler, which may be configured according to an LLVM-based compilation scheme, for example, to lower OpenCL C-code to VMP accelerator assembly code, e.g., suitable for execution by vector processor 180 (Fig. 1).

[00096] In some demonstrative aspects, compiler 200 may include one or more technologies, which may be required to compile code to a format suitable for a VMP

architecture, e.g., in addition to open-sourced LLVM compiler passes.

[00097] In some demonstrative aspects, FE 210 may be configured to parse the OpenCL C-code and to translate it, e.g., through an Abstract Syntax Tree (AST), for example, into an LLVM Intermediate Representation (IR).

[00098] In some demonstrative aspects, compiler 200 may include a dedicated API, for example, to detect a correct pattern for compiler pattern matching, for example, suitable for the VMP. For example, the VMP may be configured as a Complex Instruction Set Computer (CISC) machine implementing a very complex Instruction Set Architecture (ISA), which may be hard to target from standard C code. Accordingly, compiler pattern matching may not be able to easily detect the correct pattern, and for this case the compiler may require a dedicated API.

[00099] In some demonstrative aspects, FE 210 may implement one or more vendor extension built-ins, which may target VMP-specific ISA, for example, in addition to standard OpenCL built-ins, which may be optimized to a VMP machine.

[000100] In some demonstrative aspects, FE 210 may be configured to implement OpenCL structures and/or work item functions.

[000101] In some demonstrative aspects, ME 220 may be configured to process LLVM IR code, which may be general and target-independent, for example, although it may include one or more hooks for specific target architectures.

[000102] In some demonstrative aspects, ME 220 may perform one or more custom passes, for example, to support the VMP architecture, e.g., as described below.

[000103] In some demonstrative aspects, ME 220 may be configured to perform one or more operations of a Control Flow Graph (CFG) Linearization analysis, e.g., as described below.

[000104] In some demonstrative aspects, the CFG Linearization analysis may be configured to linearize the code, for example, by converting if-statements to select patterns, for example, in case VMP vector code does not support standard control flow.

[000105] In one example, ME 220 may receive a given code, e.g., as follows:

```
If (x > 0) {
        A = A + 5;
```

```
} else    {
            B = B * 2;
}
```

According to this example, ME 220 may be configured to apply the CFG Linearization analysis to the given code, e.g., as follows:

```
tmpA = A + 5;
tmpB = B * 2;
mask = x > 0;
A = Select mask, tmpA, A
B = Select not mask, tmpB, B
```

Example (1)

[000106] In some demonstrative aspects, ME 220 may be configured to perform one or more operations of an auto-vectorization analysis, e.g., as described below.

[000107] In some demonstrative aspects, the auto-vectorization analysis may be configured to vectorize, e.g., auto-vectorize, a given code, e.g., to utilize vector capabilities of the VMP.

[000108] In some demonstrative aspects, ME 220 may be configured to perform the auto-vectorization analysis, for example, to vectorize code in a scalar form. For example, some or all operations of the auto-vectorization analysis may not be performed, for example, in case the code is already provided in a vectorized form.

[000109] In some demonstrative aspects, for example, in some use cases and/or scenarios, a compiler may not always be able to auto-vectorize a code, for example, due to data dependencies between loop iterations.

[000110] In one example, ME 220 may receive a given code, e.g., as follows:

```
char* a,b,c;
for (int i=0; i < 2048; i++) {
    a[i]=b[i]+c[i];
```

```
}
```

According to this example, ME 220 may be configured to perform the CFG auto-vectorization analysis by applying a first conversion, e.g., as follows:

```
char* a,b,c;
for (int i=0; i < 2048; i+=32) {
    a[i.i+31]=b[i…i+31]+c[i…i+31];
}
```

Example (2a)

For example, ME 220 may be configured to perform the CFG auto-vectorization analysis by applying a second conversion, for example, following the first conversion, e.g., as follows:

```
char32* a,b,c;
for (int i=0; i < 64; i++) {
    a[i]=b[i]+c[i];
}
```

Example (2b)

[000111] In some demonstrative aspects, ME 220 may be configured to perform one or more operations of a Scratch Pad Memory Loop Access Analysis (SPMLAA), e.g., as described below.

[000112] In some demonstrative aspects, the SPMLAA may define Processing Blocks (PB), e.g., that should be outlined and compiled for VMP later.

[000113] In some demonstrative aspects, the processing blocks may include accelerated loops, which may be executed by the vector unit of the VMP.

[000114] In some demonstrative aspects, a PB, e.g., each PB, may include memory references. For example, some or all memory accesses may refer to local memory banks.

[000115] In some demonstrative aspects, the VMP may enable access to memory banks through AGUs, e.g., AGUs 320 as described below with reference to Fig. 3, and Scatter Gather units (SG).

[000116] In some demonstrative aspects, the AGUs may be pre-configured, e.g., before loop execution. For example, a loop trip count may be calculated, e.g., ahead of running a processing block.

[000117] In some demonstrative aspects, image references, e.g., some or all image references, may be created at this stage, and may be followed by calculation of strides and offsets, e.g., per dimension for each reference.

[000118] In some demonstrative aspects, ME 220 may be configured to perform one or more operations of an AGU planner analysis, e.g., as described below.

[000119] In some demonstrative aspects, the AGU Planner analysis may include iterator assignment, which may cover image references, e.g., all image references, from the entire Processing Block.

[000120] In some demonstrative aspects, an iterator may cover a single reference or a group of references.

[000121] In some demonstrative aspects, one or more memory references may be coalesced and/or reuse a same access through shuffle instructions, and/or saving values read from previous iterations.

[000122] In some demonstrative aspects, other memory references, e.g., which have no linear access pattern, may be handled using a Scatter-Gather (SG) unit, which may have a performance penalty, e.g., as it may require maintaining indices and/or masks.

[000123] In some demonstrative aspects, a plan may be configured as an arrangement of iterators in a processing block. For example, a processing block may have multiple plans, e.g., theoretically.

[000124] In some demonstrative aspects, the AGU Planner analysis may be configured to build all possible plans for all PBs, and to select a combination, e.g., a best combination, e.g., from all valid combinations.

[000125] In some demonstrative aspects, a total number of iterators in a valid combination may be limited, e.g., not to exceed a number of available AGUs on a VMP.

[000126] In some demonstrative aspects, one or more parameters, e.g., including stride, width and/or base, may be defined for an iterator, e.g., for each iterator for example, as part of the AGU Planner analysis. For example, min-max ranges for the iterators may be defined in a dimension, e.g., in each dimension, for example, as part of the AGU Planner analysis.

[000127] In some demonstrative aspects, the AGU Planner analysis may be configured to track and evaluate a memory reference, e.g., each memory reference, to an image, e.g., to understand its access pattern.

[000128] In one example, according to Examples 2a/2b, the image 'a' which is the base address, may be accessed with steps of 32 bytes for 64 iterations.

[000129] In some demonstrative aspects, the LLVM may include a scalar evaluation analysis (SCEV), which may compute an access pattern, e.g., to understand every image reference.

[000130] In some demonstrative aspects, ME 220 may utilize masking capabilities of the AGUs, for example, to avoid maintaining an induction variable, which may have a performance penalty.

[000131] In some demonstrative aspects, ME 220 may be configured to perform one or more operations of a rewrite analysis, e.g., as described below.

[000132] In some demonstrative aspects, the rewrite analysis may be configured to transform the code of a processing block, for example, while setting iterators and/or modifying memory access instructions.

[000133] In some demonstrative aspects, setting of the iterators, e.g., of all iterators, may be implemented in IR in target-specific intrinsics. For example, the setting of the iterators may reside in a pre-header of an outermost loop.

[000134] In some demonstrative aspects, the rewrite analysis may include loop-perfectization analysis, e.g., as described below.

[000135] In some demonstrative aspects, the code may be compiled with a target that substantially all calculations should be executed inside the innermost loop.

[000136] For example, the loop-perfectization analysis may hoist instructions, e.g., to move into a loop an operation performed after a last iteration of the loop.

[000137] For example, the loop-perfectization analysis may sink instructions, e.g., to move into a loop an operation performed before a first iteration of the loop.

[000138] For example, the loop-perfectization analysis may hoist instructions and/or sink instructions, for example, such that substantially all instructions are moved from outer loops to the innermost loops.

[000139] For example, the loop-perfectization analysis may be configured to provide a technical solution to support VMP iterators, e.g., to work on perfectly nested loops only.

[000140] For example, the loop-perfectization analysis may result in a situation where there are no instructions between the "for" statements that compose the loop, e.g., to support VMP iterators, which cannot emulate such cases.

[000141] In some demonstrative aspects, the loop-perfectization analysis may be configured to collapse a nested loop into a single collapsed loop.

[000142] In one example, ME 220 may receive a given code, e.g., as follows:

```
for (int i = 0; i < N; i++) {
    int sum = 0;
    for (int j = 0; j < M; j++)
    {
        sum += a[j + stride * i];
    }
        res[i] = sum;
}
```

According to this example, ME 220 may be configured to perform the loop-perfectization analysis to collapse the nested loop in the code to a single collapsed loop, e.g., as follows:

```
for (int k = 0; k < N * M; k++) {
    sum = (k % M == 0 ? 0 : sum);
    sum += a[k % M + stride * ( k / M )];
```

```
        res[k/M] = sum;
    }
```

Example (3)

[000143] In some demonstrative aspects, ME 220 may be configured to perform one or more operations of a Vector Loop Outlining analysis, e.g., as described below.

[000144] In some demonstrative aspects, the Vector Loop Outlining analysis may be configured to divide a code between a scalar subsystem and a vector subsystem, e.g., vector processing block 310 (Fig. 3) and scalar processor 330 (Fig. 3) as described below with reference to Fig. 3.

[000145] In some demonstrative aspects, the VMP accelerator may include the scalar and/or vector subsystems, e.g., as described below. For example, each of the subsystems may have different compute units/processors. Accordingly, a scalar code may be compiled on a scalar compiler, e.g., an SSC compiler, and/or an accelerated vector code may run on the VMP vector processor.

[000146] In some demonstrative aspects, the Vector Loop Outlining analysis may be configured to create a separate function for a loop body of the accelerated vector code. For example, these functions may be marked for the VMP and/or may continue to the VMP backend, for example, while the rest of the code may be compiled by the SSC compiler.

[000147] In some demonstrative aspects, one or more parts of a vector loop, e.g., configuration of the vector unit and/or initialization of vector registers, may be performed by a scalar unit. However, these parts may be performed in a later stage, for example, by performing backpatching into the scalar code, e.g., as the scalar code may still be in LLVM IR before processing by the SSC compiler.

[000148] In some demonstrative aspects, BE 230 may be configured to translate the LLVM IR into machine instructions. For example, the BE 230 may not be target agnostic and may be familiar with target-specific architecture and optimizations, e.g., compared to ME 220, which may be agnostic to a target-specific architecture.

[000149] In some demonstrative aspects, BE 230 may be configured to perform one or more analyses, which may be specific to a target machine, e.g., a VMP machine, to which the code is lowered, e.g., although BE 230 may use common LLVM.

[000150] In some demonstrative aspects, BE 230 may be configured to perform one or more operations of an instruction lowering analysis, e.g., as described below.

[000151] In some demonstrative aspects, the instruction lowering analysis may be configured to translate LLVM IR into target-specific instructions Machine IR (MIR), for example, by translating the LLVM IR into a Directed Acyclic Graph (DAG).

[000152] In some demonstrative aspects, the DAG may go through a legalization process of instructions, for example, based on the data types and/or VMP instructions, which may be supported by a VMP HW.

[000153] In some demonstrative aspects, the instruction lowering analysis may be configured to perform a process of pattern-matching, e.g., after the legalization process of instructions, for example, to lower a node, e.g., each node, in the DAG, for example, into a VMP-specific machine instruction.

[000154] In some demonstrative aspects, the instruction lowering analysis may be configured to generate the MIR, for example, after the process of pattern-matching.

[000155] In some demonstrative aspects, the instruction lowering analysis may be configured to lower the instruction according to machine Application Binary Interface (ABI) and/or calling conventions.

[000156] In some demonstrative aspects, BE 230 may be configured to perform one or more operations of a unit balancing analysis, e.g., as described below.

[000157] In some demonstrative aspects, the unit balancing analysis may be configured to balance instructions between VMP compute units, e.g., data processing units 316 (Fig. 3) as described below with reference to Fig. 3.

[000158] In some demonstrative aspects, the unit balancing analysis may be familiar with some or all available arithmetic transformations, and/or may perform transformations according to an optimal algorithm.

[000159] In some demonstrative aspects, BE 230 may be configured to perform one or more operations of a modulo scheduler (pipeliner) analysis, e.g., as described below.

[000160] In some demonstrative aspects, the pipeliner may be configured to schedule the instructions according to one or more constraints, e.g., data dependency, resource bottlenecks and/or any other constrains, for example, using Swing Modulo Scheduling (SMS) heuristics and/or any other additional and/or alternative heuristic.

[000161] In some demonstrative aspects, the pipeliner may be configured to schedule a set, e.g., an Initiation Interval (II), of Very Long Instruction Word (VLIW) instructions that the program will iterate on, e.g., during a steady state.

[000162] In some demonstrative aspects, a performance metric, which may be based on a number of cycles a typical loop may execute, may be measured, e.g., as follows:

*(Size of Input data in bytes) * II / (Bytes consumed/produced every iteration)*

[000163] In some demonstrative aspects, the pipeliner may try to minimize the II, e.g., as much as possible, for example, to improve performance.

[000164] In some demonstrative aspects, the pipeliner may be configured to calculate a minimum II, and to schedule accordingly. For example, if the pipeliner fails the scheduling, the pipeliner may try to increase the II and retry scheduling, e.g., until a predefined II threshold is violated.

[000165] In some demonstrative aspects, BE 230 may be configured to perform one or more operations of a register allocation analysis, e.g., as described below.

[000166] In some demonstrative aspects, the register allocation analysis may be configured to attempt to assign a register in an efficient, e.g., optimal, way.

[000167] In some demonstrative aspects, the register allocation analysis may assign values to bypass vector registers, general purpose vector registers, and/or scalar registers.

[000168] In some demonstrative aspects, the values may include private variables, constants, and/or values that are rotated across iterations.

[000169] In some demonstrative aspects, the register allocation analysis may implement an optimal heuristic that suites one or more VMP register file (regfile) constraints. For example, in some use cases, the register allocation analysis may not use a standard LLVM register allocation.

[000170] In some demonstrative aspects, in some cases, the register allocation analysis may fail, which may mean that the loop cannot be compiled. Accordingly, the register allocation analysis may implement a retry mechanism, which may go back to the modulo scheduler and may attempt to reschedule the loop, e.g., with an increased initiation interval. For example, increasing the initiation interval may reduce register pressure, and/or may support compilation of the vector loop, e.g., in many cases.

[000171] In some demonstrative aspects, BE 230 may be configured to perform one or more operations of an SSC configuration analysis, e.g., as described below.

[000172] In some demonstrative aspects, the SSC configuration analysis may be configured to set a configuration to execute the kernel, e.g., the AGU configuration.

[000173] In some demonstrative aspects, the SSC configuration analysis may be performed at a late stage, for example, due to configurations calculated after legalization, the register allocation analysis, and/or the modulo scheduling analysis.

[000174] In some demonstrative aspects, the SSC configuration analysis may include a Zero Overhead Loop (ZOL) mechanism in the vector loop. For example, the ZOL mechanism may configure a loop trip count based on an access pattern of the memory references in the loop, for example, to avoid running instructions that check the loop exit condition every iteration.

[000175] In some demonstrative aspects, a VMP Compilation Flow may include one or more, e.g., a few, steps, which may be invoked during the compilation flow in a test library (testlib), e.g., a wrapper script for compilation, execution, and/or program testing. For example, these steps may be performed outside of the LLVM Compiler.

[000176] In some demonstrative aspects, a PCB Hardware Description Language (PHDL) simulator may be implemented to perform one or more roles of an assembler, encoder, and/or linker.

[000177] In some demonstrative aspects, compiler 200 may be configured to provide a technical solution to support robustness, which may enable compilation of a vast selection of loops, with HW limitations. For example, compiler 200 may be configured to support a technical solution, which may not create verification errors.

[000178] In some demonstrative aspects, compiler 200 may be configured to provide a technical solution to support programmability, which may provide a user an ability to express code in multiple ways, which may compile correctly to the VMP architecture.

[000179] In some demonstrative aspects, compiler 200 may be configured to provide a technical solution to support an improved user-experience, which may allow the user capability to debug and/or profile code. For example, the improved user-experience may provide informative error messages, report tools, and/or a profiler.

[000180] In some demonstrative aspects, compiler 200 may be configured to provide a technical solution to support improved performance, for example, to optimize a VMP assembly code and/or iterator accesses, which may lead to a faster execution. For example, improved performance may be achieved through high utilization of the compute units and usage of its complex CISC.

[000181] Reference is made to Fig. 3, which schematically illustrates a vector processor 300, in accordance with some demonstrative aspects. For example, vector processor 180 (Fig. 1) may be implement one or more elements of vector processor 300, and/or may perform one or more operations and/or functionalities of vector processor 300.

[000182] In some demonstrative aspects, vector processor 300 may include a Vector Microcode Processor (VMP).

[000183] In some demonstrative aspects, vector processor 300 may include a Wide Vector machine, for example, supporting Very Long Instruction Word (VLIW) architectures, and/or Single Instruction/Multiple Data (SIMD) architectures.

[000184] In some demonstrative aspects, vector processor 300 may be configured to provide a technical solution to support high performance for short integral types, which may be common, for example, in computer-vision and/or deep-learning algorithms.

[000185] In other aspects, vector processor 300 may include any other type of vector processor, and/or may be configured to support any other additional or alternative functionalities.

[000186] In some demonstrative aspects, as shown in Fig. 3, vector processor 300 may include a vector processing block (vector processor) 310, a scalar processor 330, and a Direct Memory Access (DMA) 340, e.g., as described below.

[000187] In some demonstrative aspects, as shown in Fig. 3, vector processing block 310 may be configured to process, e.g., efficiently process, image data and/or vector data. For example, the vector processing block 310 may be configured to use vector computation units, for example, to speed up computations.

[000188] In some demonstrative aspects, scalar processor 330 may be configured to perform scalar computations. For example, the scalar processor 330 may be used as a "glue logic" for programs including vector computations. For example, some, e.g., even most, of the computation of the programs may be performed by the vector processing block 310. However, several tasks, for example, some essential tasks, e.g., scalar computations, may be performed by the scalar processor 330.

[000189] In some demonstrative aspects, the DMA 340 may be configured to interface with one or more memory elements in a chip including vector processor 300.

[000190] In some demonstrative aspects, the DMA 340 may be configured to read inputs from a main memory, and/or write outputs to the main memory.

[000191] In some demonstrative aspects, the scalar processor 330 and the vector processing block 310 may use respective local memories to process data.

[000192] In some demonstrative aspects, as shown in Fig. 3, vector processor 300 may include a fetcher and decoder 350, which may be configured to control the scalar processor 330 and/or the vector processing block 310.

[000193] In some demonstrative aspects, operations of the scalar processor 330 and/or the vector processing block 310 may be triggered by instructions stored in a program memory 352.

[000194] In some demonstrative aspects, the DMA 340 may be configured to transfer data, for example, in parallel with the execution of the program instructions in memory 352.

[000195] In some demonstrative aspects, DMA 340 may be controlled by software, e.g., via configuration registers, for example, rather than instructions, and, accordingly, may be considered as a second "thread" of execution in vector processor 300.

[000196] In some demonstrative aspects, the scalar processor 330, the vector processing block 310, and/or the DMA 340 may include one or more data processing units, for example, a set of data processing units, e.g., as described below.

[000197] In some demonstrative aspects, the data processing units may include hardware configured to preform computations, e.g., an Arithmetic Logic Unit (ALU).

[000198] In one example, a data processing unit may be configured to add numbers, and/or to store the numbers in a memory.

[000199] In some demonstrative aspects, the data processing units may be controlled by commands, e.g., encoded in the program memory 352 and/or in configuration registers. For example, the configuration registers may be memory mapped, and may be written by the memory store commands of the scalar processor 330.

[000200] In some demonstrative aspects, the scalar processor 330, the vector processing block 310, and/or the DMA 340 may include a state configuration including a set of registers and memories, e.g., as described below.

[000201] In some demonstrative aspects, as shown in Fig. 3, vector processor block 310 may include a set of vector memories 312, which may be configured, for example, to store data to be processed by vector processor block 310.

[000202] In some demonstrative aspects, as shown in Fig. 3, vector processor block 310 may include a set of vector registers 314, which may be configured, for example, to be used in data processing by vector processor block 310.

[000203] In some demonstrative aspects, the scalar processor 330, the vector processing block 310, and/or the DMA 340 may be associated with a set of memory maps.

[000204] In some demonstrative aspects, a memory map may include a set of addresses accessible by a data processing unit, which may load and/or store data from/to registers and memories.

[000205] In some demonstrative aspects, as shown in Fig. 3, the vector processing block 310 may include a plurality of Address Generation Units (AGUs) 320, which may include addresses accessible to them, e.g., in one or more of memories 312.

[000206] In some demonstrative aspects, as shown in Fig. 3, vector processor block 310 may include a plurality of data processing units 316, e.g., as described below.

[000207] In some demonstrative aspects, data processing units 316 may be configured to process commands, e.g., including several numbers at a time. In one example, a command may include 8 numbers. In another example, a command may include 4 numbers, 16 numbers, or any other count of numbers.

[000208] In some demonstrative aspects, two or more data processing units 316 may be used simultaneously. In one example, data processing units 316 may process and execute a plurality of different command, e.g., 3 different commands, for example, including 8 numbers, at a throughout of a single cycle.

[000209] In some demonstrative aspects, data processing units 316 may be asymmetrical. For example, first and second data processing units 316 may support different commands. For example, addition may be performed by a first data processing unit 316, and/or multiplication may be performed by a second data processing unit 316. For example, both operations may be performed by one or more additional other data processing units 316.

[000210] In some demonstrative aspects, data processing units 316 may be configured to support arithmetic operations for many combinations of input & output data types.

[000211] In some demonstrative aspects, data processing units 316 may be configured to support one or more operations, which may be less common. For example, processing units 316 may support operations working with a Look Up Table (LUT) of vector processor 300, and/or any other operations.

[000212] In some demonstrative aspects, data processing units 316 may be configured to support efficient computation of non-linear functions, histograms, and/or random data access, e.g., which may be useful to implement algorithms like image scaling, Hough transforms, and/or any other algorithms.

[000213] In some demonstrative aspects, vector memories 312 may include, for example, memory banks having a size of 16K or any other size, which may be accessed at a same cycle.

[000214] In one example, a maximal memory access size may be 64 bits. According to this example, a peak throughput may be 256 bits, e.g., 64x4 = 256. For example, high

memory bandwidth may be implemented to utilize computation capabilities of the data processing units 316.

[000215] In one example, two data processing units 316 may support 16 8-bit multiply & accumulate operations (MACs) per cycle. According to this example, the two data processing units 316 may not be useful, for example, in case the input numbers are not fetched at this speed, and/or there isn't exactly 256 bits of input, e.g., *16x8x2 = 256.*

[000216] In some demonstrative aspects, AGUs 320 may be configured to perform memory access operations, e.g., loading and storing data from/to vector memories 314.

[000217] In some demonstrative aspects, AGUs 320 may be configured to compute addresses of input and output data items, for example, to handle I/O to utilize the data processing units 316, e.g., in case sheer bandwidth is not enough.

[000218] In some demonstrative aspects, AGUs 320 may be configured to compute the addresses of the input and/or output data items, for example, based on configuration registers written by the scalar processor 330, for example, before a block of vector commands, e.g., a loop, is entered.

[000219] For example, AGUs 320 may be configured to write an image base pointer, a width, a height and/or a stride to the configuration registers, for example, in order to iterate over an image.

[000220] In some demonstrative aspects, AGUs 320 may be configured to handle addressing, e.g., all addressing, for example, to provide a technical solution in which data processing units 316 may not have the burden of incrementing pointers or counters in a loop, and/or the burden to check for end-of-row conditions, e.g., to zero a counter in the loop.

[000221] In some demonstrative aspects, as shown in Fig. 3, AGUs 320 may include 4 AGUs, and, accordingly, four memories 312 may be accessed at a same cycle. In other aspects, any other count of AGUs 32 may be implemented.

[000222] In some demonstrative aspects, AGUs 320 may not be "tied" to memory banks 312. For example, an AGU 320, e.g., each AGU 320, may access a memory bank 312, e.g., every memory bank 312, for example, as long as two or more AGUs 320 do not try to access the same memory bank 312 at the same cycle.

[000223] In some demonstrative aspects, vector registers 314 may be configured to support communication between the data processing units 316 and AGUs 320.

[000224] In one example, a total number of vector registers 314 may be 28, which may be divided into several subsets, e.g., based on their function. For example, a first subset of vector registers 314 may be used for inputs/outputs, e.g., of all data processing units 316 and/or AGUs 320; and/or a second subset of vector registers 314 may not be used for outputs of some operations, e.g., most operations, and may be used for one or more other operations, e.g., to store loop-invariant inputs.

[000225] In some demonstrative aspects, a data processing unit 316, e.g., each data processing unit 316, may have one or more registers to host an output of a last executed operation, e.g., which may be fed as inputs to other data processing units 316. For example, these registers may "bypass" the vector registers 314, and may work faster than writing these outputs to first set of vector registers 314.

[000226] In some demonstrative aspects, fetcher and decoder 350 may be configured to support low-overhead vector loops, e.g., very low overhead vector loops (also referred to as “zero-overhead vector loops”), for example, where there may be no need to check a termination (exit) condition of a vector loop during an execution of the vector loop.

[000227] For example, a termination (exit) condition may be signaled by an AGU 320, for example, when the AGU 320 finishes iterating over a configured memory region.

[000228] For example, fetcher and decoder 350 may quit the loop, for example, when the AGU 320 signals the termination condition.

[000229] For example, the scalar processor 330 may be utilized to configure the loop parameters, e.g., first & last instructions and/or the exit condition.

[000230] In one example, vector loops may be utilized, for example, together with high memory bandwidth and/or cheap addressing, for example, to solve a control and data flow problem, for example, to provide a technical solution to allow the data processing units 316 to process data, e.g., without substantially additional overhead.

[000231] In some demonstrative aspects, scalar processor 330 may be configured to provide one or more functionalities, which may be complementary to those of the vector

processing block 310. For example, a large portion, e.g., most, of the work in a vector program may be performed by the data processing units 316. For example, the scalar processor 330 may be utilized, for example, for "gluing" together the various blocks of vector code of the vector program.

[000232] In some demonstrative aspects, scalar processor 330 may be implemented separately from vector processing block 310. In other aspects, scalar processor 330 may be configured to share one or more components and/or functionalities with vector processing block 310.

[000233] In some demonstrative aspects, scalar processor 330 may be configured to perform operations, which may not be suitable for execution on vector processing block 310.

[000234] For example, scalar processor 330 may be utilized to execute 32 bit C programs. For example, scalar processor 330 may be configured to support 1, 2, and/or 4 byte data types of C code, and/or some or all arithmetic operators of C code.

[000235] For example, scalar processor 330 may be configured to provide a technical solution to perform operations that cannot be executed on vector processing block 310, for example, without using a full-blown CPU.

[000236] In some demonstrative aspects, scalar processor 330 may include a scalar data memory 332, e.g., having a size of 16K or any other size, which may be configured to store data, e.g., variables used by the scalar parts of a program.

[000237] For example, scalar processor 330 may store local and/or global variables declared by portable C code, which may be allocated to scalar data memory by a compiler, e.g., compiler 200 (Fig. 2).

[000238] In some demonstrative aspects, as shown in Fig. 3, scalar processor 330 may include, or may be associated with, a set of vector registers 334, which may be used in data processing performed by the scalar processor 330.

[000239] In some demonstrative aspects, scalar processor 330 may be associated with a scalar memory map, which may support scalar processor 330 in accessing substantially all states of vector processor 300. For example, the scalar processor 330 may configure the vector units and/or the DMA channels via the scalar memory map.

[000240] In some demonstrative aspects, scalar processor 330 may not be allowed to access one or more block control registers, which may be used by external processors to run and debug vector programs.

[000241] In some demonstrative aspects, DMA 340 may be configured to communicate with one or more other components of a chip implementing the vector processor 300, for example, via main memory. For example, DMA 340 may be configured to transfer blocks of data, e.g., large, contiguous, blocks of data, for example, to support the scalar processor 330 and/or the vector processing block, which may manipulate data stored in the local memories. For example, a vector program may be able to read data from the main chip memory using DMA 340.

[000242] In some demonstrative aspects, DMA 340 may be configured to communicate with other elements of the chip, for example, via a plurality of DMA channels, e.g., 8 DMA channels or any other count of DMA channels. For example, a DMA channel, e.g., each DMA channel, may be capable of transferring a rectangular patch from the local memories to the main chip memory, or vice versa. In other aspects, the DMA channel may transfer any other type of data block between the local memories and the main chip memory.

[000243] In some demonstrative aspects, a rectangular patch may be defined by a base pointer, a width, a height, and astride.

[000244] For example, at peak throughput, 8 bytes per cycle may be transferred, however, there may be overheads for each patch and/or for each row in a patch.

[000245] In some demonstrative aspects, DMA 340 may be configured to transfer data, for example, in parallel with computations, e.g., via the plurality of DMA channels, for example, as long as executed commands do not access a local memory involved in the transfer.

[000246] In one example, as all channels may access the same memory bus, using several channels to implement a transfer may not save I/O cycles, e.g., compared to the case when a single channel is used. However, the plurality of DMA channels may be utilized to schedule several transfers and execute them in parallel with computations. This may be advantageous, for example, compared to a single channel, which may not allow scheduling a second transfer before completion of the first transfer.

[000247] In some demonstrative aspects, DMA 340 may be associated with a memory map, which may support the DMA channels in accessing vector memories and/or the scalar data. For example, access to the vector memories may be performed in parallel with computations. For example, access to the scalar data may usually not be allowed in parallel, e.g., as the scalar processor 330 may be involved in almost any sensible program, and may likely access it's local variables while the transfer is performed, which may lead to a memory contention with the active DMA channel.

[000248] In some demonstrative aspects, DMA 340 may be configured to provide a technical solution to support parallelization of I/O and computations. For example, a program performing computations may not have to wait for I/O, for example, in case these computations may run fast by vector processing block 310.

[000249] In some demonstrative aspects, an external processor, e.g., a CPU, may be configured to initiate execution of a program on vector processor 300. For example, vector processor 300 may remain idle, e.g., as long as program execution is not initiated.

[000250] In some demonstrative aspects, the external processor may be configured to debug the program, e.g., execute a single step at a time, halt when the program reaches breakpoints, and/or inspect contents of registers and memories storing the program variables.

[000251] In some demonstrative aspects, an external memory map may be implemented to support the external processor in controlling the vector processor 300 and/or debugging the program, for example, by writing to control registers of the vector processor 300.

[000252] In some demonstrative aspects, the external memory map may be implemented by a superset of the scalar memory map. For example, this implementation may make all registers and memories defined by the architecture of the vector processor 300 accessible to a debugger back-end running on the external processor.

[000253] In some demonstrative aspects, the vector processor 300 may raise an interrupt signal, for example, when the vector processor 300 terminates a program.

[000254] In some demonstrative aspects, the interrupt signal may be used, for example to implement a driver to maintain a queue of programs scheduled for execution

by the vector processor 300, and/or to launch a new program, e.g., by the external processor, for example, upon the completion of a previously executed program.

[000255] Referring back to Fig. 1, in some demonstrative aspects, compiler 160 may be configured to generate the target code 115 configured, for example, to utilize registers of a processor, for example, a vector processor, e.g., vector processor 180, according to a register allocation scheme, e.g., as described below.

[000256] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to reduce a usage of allocated registers for executing a program by a processor, for example, a vector processor, e.g., as described below.

[000257] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to utilize a reduced, e.g., optimized and/or minimal, number of allocated registers for executing a program by a processor, for example, a vector processor, e.g., as described below.

[000258] In one example, the register allocation scheme may be configured to provide a technical solution to utilize a reduced, e.g., optimized and/or minimal, number of allocated registers, which may be allocated from the plurality of vector registers 314 (Fig. 3), for execution of a program by vector processor 300 (Fig. 3), e.g., as described below.

[000259] In some demonstrative aspects, a compiler, e.g., compiler 160, may be configured to generate target code, e.g., the target code 115, which may be configured to utilize registers of a vector processor, e.g., vector processor 180, according to a register allocation scheme, e.g., as described below.

[000260] In other aspects, a compiler, e.g., compiler 160, may be configured to generate target code, e.g., the target code 115, which may be configured to utilize registers of any other suitable type of processor according to the register allocation scheme, e.g., as described below.

[000261] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution improve, e.g., to optimize, an allocation of registers to execute the executable program, e.g., as described below.

[000262] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to support an improved allocation, for example, an efficient allocation, e.g., an optimized allocation, of registers to execute the executable program, e.g., as described below.

[000263] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to utilize a reduced number, for example, an optimized number, e.g., a minimal number, of allocated registers to execute the executable program, e.g., as described below.

[000264] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to improve performance of the executable program, for example, by reducing the number of allocated registers to execute the executable program, e.g., as described below.

[000265] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to reduce the usage of the allocated registers, for example, by reducing ranges of live intervals of one or more variables, e.g., as described below.

[000266] In some demonstrative aspects, a live range of a variable may include a range of cycles of the executable program, for example, between a first cycle, e.g., which includes a first use and/or a production of the variable, and a second cycle, e.g., which includes a second use, e.g., subsequent to the first use, of the variable, e.g., as described below.

[000267] In some demonstrative aspects, there may be a need to provide a technical solution to efficiently allocate registers of a processor, e.g., a vector processor, for execution of a program, for example, in order to reduce a usage of the allocated registers, for example, allocated vector registers, e.g., as described below.

[000268] For example, a number of physical registers implemented by a chip including a processor, e.g., a vector processor or any other processor, may be limited, for example, according to a design and/or layout of the chip. For example, a number of vector registers implemented by the vector processor may be limited by the number of physical registers on the chip implementing the vector processor.

[000269] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to reduce, e.g., optimize and/or minimize, the number of the allocated registers, for example, for CPUs, e.g., vector processors, having limited storage capabilities, e.g., a limited register pool.

[000270] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to reduce, e.g., optimize and/or minimize, the number of the allocated registers, for example, for CPUs, e.g., vector processors, having limited support of, or not having support of, memory spill/fill operations, e.g., to store live values.

[000271] For example, processors having no fill/spill capabilities and/or limited storage capabilities may be forced to use compute resources instead.

[000272] In one example, a processor may identify an unsuccessful register allocation according to an instruction scheduling, e.g., due to a limited number of registers, which may not be able to support the register allocation according to the instruction scheduling. One option to address this situation may be to relax the instruction scheduling, e.g., in attempt to reduce the number of live variables sharing the same execution cycles. However, this option may result in degraded performance.

[000273] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to reduce the number of the allocated registers, for example, to avoid, or even eliminate, the use of these extra compute resources.

[000274] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to reduce a register pressure for execution of a program by a processor, for example, a vector processor, and/or any other processor.

[000275] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to reduce the register pressure, for example, by reducing the number of allocated registers for execution of the program, e.g., as described below.

[000276] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to support efficient execution of programs, e.g., complex programs, which may be sensitive to register pressure. For example, for

some programs, a register pressure issue may be a bottleneck, and, accordingly, the register pressure may affect performance.

[000277] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to reduce, e.g., optimize and/or minimize, the number of allocated registers for execution of a program, for example, while providing a suitable allocation of registers, e.g., vector registers, for execution of the program, e.g., as described below.

[000278] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to reduce, e.g., minimize and/or optimize, ranges of live intervals for one or more variables, e.g., as described below.

[000279] In some demonstrative aspects, the reduction of the ranges of live intervals may provide a technical solution to support reduced register usage and, accordingly, to support reduction, e.g., minimization and/or optimization, of register pressure.

[000280] For example, the register allocation scheme may be implemented to provide a technical solution to support efficient execution of programs, which may be subject to a register pressure problem.

[000281] In some demonstrative aspects, the register allocation scheme may be configured to provide a technical solution to support improved performance of programs, which may be executed, for example, by exposed pipeline processors, and/or any other additional or alternative type of processors.

[000282] In some demonstrative aspects, complier 160 may be configured to process a given instruction scheduling, e.g., based on the source code 112, and to generate target code 115, which may be configured to utilize a reduced, e.g., optimized and/or minimized, number of allocated registers, for example, for a successful register allocation, e.g., as described below.

[000283] In some demonstrative aspects, complier 160 may be configured to identify in the given instruction scheduling a pair of instructions, e.g., including a first instruction and a second instruction, which may be independent and/or similar.

[000284] In some demonstrative aspects, compiler 160 may be configured to determine whether to perform the first instruction and the second instruction in parallel,

e.g., using a single SIMD instruction, or to perform the first instruction and the second instruction using two separate instructions, e.g., as descried below.

[000285] In one example, in some use cases, implementations and/or scenarios, outputs and/or inputs of an SIMD instruction may have asymmetric uses. For example, execution of such a SIMD instruction may raise one or more constraints for an instruction scheduler. For example, these constraints may be reflected in a resulting register pressure, e.g., as described below.

[000286] In some demonstrative aspects, complier 160 may be configured selectively allocate a pair of instructions to be executed as a SIMD instruction, for example, based on a criterion, which may be configured to reduce, e.g., minimize, live intervals corresponding to the pair of instructions, e.g., as described below.

[000287] In some demonstrative aspects, the criterion may be configured to provide a technical solution to reduce, e.g., minimize and/or optimize, register usage, e.g., as described below.

[000288] In some demonstrative aspects, complier 160 may be configured to provide a technical solution to support reduced, e.g., minimal and/or optimal, usage of registers allocated for execution of a program, for example, for a given instruction scheduling, e.g., based on the source code 112, e.g., as described below.

[000289] In some demonstrative aspects, compiler 160 may be configured to combine a pair of instructions, e.g., a pair of similar yet independent instructions, into an SIMD instruction.

[000290] In some demonstrative aspects, compiler 160 may be configured to generate a valid instruction scheduling based on the SIMD instruction, e.g., as described below.

[000291] In some demonstrative aspects, compiler 160 may be configured to determine whether or not the valid instruction scheduling may result in an invalid register allocation, for example, due to register pressure and/or any other reason, e.g., as described below.

[000292] In some demonstrative aspects, compiler 160 may be configured to identify a valid instruction scheduling, which may result in an invalid register allocation, e.g., due to register pressure and/or any other reason.

[000293] In some demonstrative aspects, compiler 160 may be configured to attempt finding in the valid instruction scheduling one or more identified SIMD instructions, e.g., SIMD output instructions, having asymmetric uses, for example, based on a determination that the valid instruction scheduling may result in the invalid register allocation, e.g., as described below.

[000294] In some demonstrative aspects, compiler 160 may be configured to split an identified SIMD instruction, e.g., each identified SIMD instruction or only some identified SIMD instructions, into two or more instructions, for example, if possible, e.g., as described below.

[000295] In some demonstrative aspects, compiler 160 may be configured to generate a new instruction set including the two or more instructions, and to perform instruction scheduling and register allocation on the new instruction set, e.g., as described below.

[000296] In some demonstrative aspects, the selective splitting of the identified SIMD instructions into multiple instructions may provide a technical solution to avoid and/or remove asymmetry from the identified SIMD instructions, e.g., as described below.

[000297] For example, the removal of asymmetry from the identified SIMD instructions may provide a technical solution to release one or more constraints in the instruction scheduling, e.g., as described below.

[000298] In some demonstrative aspects, compiler 160 may be configured to identify, for example, based on source code 112, a first data operation and a second data operation, which may be executable in parallel, e.g., as described below.

[000299] In some demonstrative aspects, compiler 160 may be configured to identify the first data operation and the second data operation, which may be executable in parallel, for example, according to a SIMD instruction, e.g., as described below.

[000300] In some demonstrative aspects, the SIMD instruction may be configured to be executed, for example, by a single data processing unit (ALU) of a target processor, for example, processor 180, e.g., as described below.

[000301] In one example, compiler 200 (Fig. 2) may be configured to identify a first data operation and a second data operation, which may be executed in parallel, for example, according to a SIMD instruction to be executed by a single data processing unit 316 (Fig. 3) of vector processor 300 (Fig. 3).

[000302] In some demonstrative aspects, the first data operation and the second data operation may include data operations of a same instruction, e.g., as described below.

[000303] In some demonstrative aspects, the first data operation may include a data operation of a first instruction, e.g., as described below.

[000304] In some demonstrative aspects, the second data operation may include a data operation of a second instruction, for example, separate from, and/or independent of, the first instruction, e.g., as described below.

[000305] In some demonstrative aspects, compiler 160 may be configured to determine a selected compilation scheme to be applied for compiling the first data operation and the second data operation, e.g., as described below.

[000306] In some demonstrative aspects, compiler 160 may be configured to determine the selected compilation scheme, for example, from a first compilation scheme and a second compilation scheme, e.g., as described below.

[000307] In some demonstrative aspects, compiler 160 may be configured to determine the selected compilation scheme, for example, by selecting the selected compilation scheme from the first compilation scheme and the second compilation scheme, for example, based on a predefined selection criterion, e.g., as described below.

[000308] In some demonstrative aspects, the first compilation scheme may include compilation of the first data operation and the second data operation into a SIMD instruction, e.g., as described below.

[000309] In some demonstrative aspects, the second compilation scheme may include compilation of the first data operation into a first ALU instruction, and compilation of the second data operation into a second ALU instruction, e.g., as described below.

[000310] In some demonstrative aspects, the second ALU instruction may be configured to be executed separately from the first ALU instruction, e.g., as described below.

[000311] In some demonstrative aspects, compiler 160 may be configured to generate the target code 115, for example, based on compilation of the first data operation and the second data operation according to the selected compilation scheme, e.g., as described below.

[000312] In some demonstrative aspects, compiler 160 may be configured to generate the target code 115 configured, for example, for execution by a Very Long Instruction Word (VLIW) Single Instruction/Multiple Data (SIMD) target processor, e.g., processor 180.

[000313] In other aspects, compiler 160 may be configured to generate the target code 115 configured, for example, for execution by any other suitable type of processor.

[000314] In some demonstrative aspects, compiler 160 may be configured to generate the target code 115, for example, based on the source code 112 including Open Computing Language (OpenCL) code.

[000315] In other aspects, compiler 160 may be configured to generate the target code 115, for example, based on the source code 112 including any other suitable type of code.

[000316] In some demonstrative aspects, compiler 160 may be configured to compile the source code 112 into the target code 115, for example, according to a Low Level Virtual Machine (LLVM) based (LLVM-based) compilation scheme.

[000317] In other aspects, compiler 160 may be configured to compile the source code 112 into the target code 115 according to any other suitable compilation scheme.

[000318] In some demonstrative aspects, the first ALU instruction may be configured to be executed in a first execution cycle, e.g., as described below.

[000319] In some demonstrative aspects, the second ALU instruction may be configured to be executed in a second execution cycle, for example, different from the first execution cycle, e.g., as described below.

[000320] In some demonstrative aspects, the second compilation scheme may include compilation of the first data operation into the first ALU instruction to be executed, for example, during a first cycle, and compilation of the second data operation into the second ALU instruction to be executed, for example, during a second cycle, which may be different from the first cycle, e.g., as described below.

[000321] In some demonstrative aspects, compiler 160 may be configured to identify an ALU to be allocated to execute the second ALU instruction, for example, based on

a determination that the ALU is available during the second cycle, e.g., as described below.

[000322] In some demonstrative aspects, the first ALU instruction may be configured to be executed by a first ALU of the target processor 180, e.g., as described below.

[000323] In some demonstrative aspects, the second ALU instruction may be configured to be executed by a second ALU of the target processor 180, e.g., as described below.

[000324] In some demonstrative aspects, the first ALU instruction and the second ALU instruction may be configured to be executed by a same ALU of the target processor 180, e.g., as described below.

[000325] In some demonstrative aspects, compiler 160 may be configured to generate target code, e.g., target code 115, which may be configured for execution by the target processor, e.g., processor 180, e.g., as described below.

[000326] In some demonstrative aspects, compiler 160 may be configured to determine the selected compilation scheme for compilation of the first and second data operations, for example, according to a predefined selection criterion, which may include, for example, a register-utilization criterion corresponding, for example, to a register utilization of one or more registers of the target processor 180, e.g., as described below.

[000327] In some demonstrative aspects, the register-utilization criterion may be based, for example, on the register utilization of the one or more registers of the target processor 180 according to the first compilation scheme, for example, based on utilizing the SIMD instruction to execute the first and second data operations, e.g., as described below.

[000328] In some demonstrative aspects, the register-utilization criterion may be configured, for example, such that the selected compilation scheme may include the second compilation scheme, for example, when a first register utilization of the one or more registers of the target processor 180 according to the first compilation scheme is greater than a second register utilization of the one or more registers of the target processor 180 according to the second compilation scheme, e.g., as described below.

[000329] In some demonstrative aspects, compiler 160 may be configured to determine that the selected compilation scheme is to include the second compilation scheme, for example, based on a determination that the first register utilization of the one or more registers of the target processor 180 according to the first compilation scheme is greater than the second register utilization of the one or more registers of the target processor 180 according to the second compilation scheme, e.g., as described below.

[000330] In some demonstrative aspects, compiler 160 may be configured to determine that the selected compilation scheme is to include the second compilation scheme, for example, based on a determination that the first register utilization of the one or more registers of the target processor 180 according to the first compilation scheme is greater than a predefined register utilization threshold, e.g., as described below.

[000331] In other aspects, the register-utilization criterion may include any other additional or alternative criterion relating to the register utilization of the target processor 180.

[000332] In some demonstrative aspects, compiler 160 may be configured to determine the selected compilation scheme, for example, according to a predefined selection criterion, which may be based, for example, on a live range of at least one variable corresponding to at least one of the first data operation and/or the second data operation, e.g., as described below.

[000333] In some demonstrative aspects, the at least one variable corresponding to the first data operation and/or the second data operation may include, for example, at least one input variable of the first data operation and/or the second data operation, e.g., as described below.

[000334] In some demonstrative aspects, the at least one variable corresponding to the first data operation and/or the second data operation may include, for example, at least one output variable resulting from the first data operation and/or the second data operation, e.g., as described below.

[000335] In some demonstrative aspects, the predefined selection criterion may be based, for example, on a live range of an input variable of the SIMD instruction, for example, according to the first compilation scheme, e.g., as described below.

[000336] In some demonstrative aspects, the predefined selection criterion may be based, for example, on a live range of an output variable resulting from the SIMD instruction, for example, according to the first compilation scheme, e.g., as described below.

[000337] In some demonstrative aspects, the predefined selection criterion may be configured, for example, such that the selected compilation scheme is to include the second compilation scheme, for example, when a first live range of the variable according to the first compilation scheme is greater than a second live range of the variable according to the second compilation scheme, e.g., as described below.

[000338] In some demonstrative aspects, compiler 160 may be configured to determine that the selected compilation scheme is to include the second compilation scheme, for example, based on a determination that the first live range of the variable according to the first compilation scheme is greater than the second live range of the variable according to the second compilation scheme, e.g., as described below.

[000339] In some demonstrative aspects, compiler 160 may be configured to determine that the selected compilation scheme is to include the second compilation scheme, for example, based on a determination that the live range of the variable according to the first compilation scheme is greater than a predefined live range threshold, e.g., as described below.

[000340] In some demonstrative aspects, compiler 160 may be configured to determine the selected compilation scheme, for example, according to a predefined selection criterion, which may be based, for example, on a difference between a live range of a first variable and a live range of a second variable, e.g., as described below.

[000341] In some demonstrative aspects, the first variable may result from execution of the first data operation by the SIMD instruction, e.g., according to the first compilation scheme, e.g., as described below.

[000342] In some demonstrative aspects, the second variable may result from execution of the second data operation by the SIMD instruction, e.g., according to the first compilation scheme, e.g., as described below.

[000343] In some demonstrative aspects, the first variable may include an input for execution of the first data operation by the SIMD instruction, e.g., according to the first compilation scheme, e.g., as described below.

[000344] In some demonstrative aspects, the second variable may include an input for execution of the second data operation by the SIMD instruction, e.g., according to the first compilation scheme, e.g., as described below.

[000345] In some demonstrative aspects, the predefined selection criterion may be based on a register-utilization corresponding to the first compilation scheme, e.g., as described below.

[000346] In some demonstrative aspects, the register-utilization corresponding to the first compilation scheme may be based, for example, on a count of cycles that a register is to be occupied by a result of the SIMD instruction, e.g., as described below.

[000347] In some demonstrative aspects, the register-utilization corresponding to the first compilation scheme may be based, for example, on a count of cycles between a first cycle, e.g., an initial cycle, in which the result of the SIMD instruction is to be available in the register, and a second cycle, e.g., subsequent to the first cycle, in which the result of the SIMD instruction is to be retrieved from the register, e.g., as described below.

[000348] In some demonstrative aspects, the register-utilization corresponding to the first compilation scheme may be based, for example, on a count of cycles that a register is to be occupied by an input variable of the SIMD instruction, e.g., as described below.

[000349] In some demonstrative aspects, the register-utilization corresponding to the first compilation scheme may be based, for example, on a count of cycles between a first cycle, e.g., an initial cycle, in which the input variable of the SIMD instruction is to be available in the register, and a second cycle, e.g., subsequent to the first cycle, in which the input variable of the SIMD instruction is to be retrieved from the register, for example for execution of the SIMD instruction, e.g., as described below.

[000350] In other aspects, the predefined selection criterion may include any other additional or alternative criterion relating to the live range of the variable resulting from the first data operation and/or the second data operation.

[000351] In some demonstrative aspects, the predefined selection criterion may be based, for example, on a latency of the ALU the target processor to execute the SIMD instruction, e.g., as described below.

[000352] In other aspects, the predefined selection criterion may include any other additional or alternative criteria, e.g., relating to any other additional or alternative parameter and/or attribute.

[000353] In some demonstrative aspects, compiler 160 may be configured to provide a compiled code, e.g., target code 115, for example, based on compilation of the first and second data operations according to the selected compilation scheme, e.g., as described below.

[000354] In some demonstrative aspects, compiler 160 may be configured to determine the selected compilation scheme, for example, according to a predefined selection criterion, which may be based, for example, on a live range of at least one output variable to result from execution of the SIMD instruction, for example, according to the first compilation scheme, e.g., as described below.

[000355] In some demonstrative aspects, compiler 160 may be configured to process a source code 112 of an executed program.

[000356] In one example, source code 112 may include an instruction scheduling, which may be configured, for example, to calculate an expression based on a plurality of variables, for example, including four variables, denoted $a,\ b,\ c,\ d$, e.g., as follows:

$$(a \cdot b \cdot 3) + c \cdot d \tag{1}$$

[000357] For example, the variables $a,\ b,\ c,\ d$, may be stored in four registers, denoted, *R0, R1, R2* and *R3*,respectively.

[000358] For example, compiler 160 may be configured to compile the source code 112 for a processor (SIMD processor) supporting SIMD instructions, e.g., processor 180. For example, the processor may include a first ALU and a second ALU, which

may be configured to perform add and multiply instructions, for example, with a latency of 2 cycles.

[000359] For example, the SIMD processor may have a memory unit with a latency of 1 cycle for a memory access operation, e.g., a store operation to store a value to the memory unit, or load operation to load a value from the memory unit.

[000360] For example, execution of the Expression 1 may be implemented according to the following instruction scheduling, e.g., utilizing a SIMD instruction:

Cycle	**ALU1**	**ALU2**	**Comments**
0	%0, %1 = mul(a, b, c, d)		$\%0 = a \cdot b, \%1 = c \cdot d$ At First, we assume that it is beneficial to use the SIMD mul instruction.
1			
2	%2 = mul(%0, 3)		
3			
4	%3 = add(%2, %1)		

TABLE (1)

[000361] As shown in Table 1, the instructions of the executed program may be performed using only the first ALU (ALU1).

[000362] As shown in Table 1, the second ALU (ALU2) may not be required to perform any operations of the executed program.

[000363] As shown in Table 1, the execution of the Expression 1 may be performed during 5 cycles.

[000364] As Shown in Table 1, the instruction scheduling may include a SIMD instruction to be executed in cycle 0. The SIMD instruction may include a combination of a first data operation, e.g., *%0=a*b*, and a second data operation, e.g., *%1=c*d*, which may be executed in parallel by the ALU1.

[000365] In some demonstrative aspects, compiler 160 may be configured to identify live ranges of variables in the instruction scheduling according to Table 1, e.g., as follows:

Value	**Live Range in Cycles [Start, End] (inclusive)**
%0	[2, 2]
%1	[2, 4]
%2	[4, 4]
%3	[6, 6]

TABLE (2)

[000366] As shown in Table 2, the output variable *%1*, which may result from the execution of the second data operation in the SIMD instruction, may be live between the cycle 2 and the cycle 4. This live range of the variable *%1* may require occupying a register to store the value of the variable *%1*, for example, between the cycle 2 and the cycle 4.

[000367] In one example, the requirement to reserve the register for storing the variable *%1* between the cycle 2 and the cycle 4 may result in register pressure.

[000368] In some demonstrative aspects, compiler 160 may identify that using the SIMD instruction may result in register pressure.

[000369] In some demonstrative aspects, compiler 160 may identify that the SIMD instruction may result in asymmetric use of registers. For example, compiler 160 may identify that the SIMD instruction may result in the output variable *%0*, which may be live only in cycle 2, while the output variable *%1* may be required to be maintained live between the cycles 2-4.

[000370] In some demonstrative aspects, compiler 160 may be configured to determine that it may be preferable to compile the first data operation, e.g., %0=*a*b*, into a first instruction (first ALU instruction), which may be executed, for example, by the first ALU (ALU1), and to compile the second data operation, e.g., *%1=c*d*, into a

second instruction (second ALU instruction), which may be executed, for example, by the second ALU (ALU2).

[000371] In some demonstrative aspects, compiler 160 may be configured to determine that it may be preferable to split the SIMD instruction into the first ALU instruction to be executed by the first ALU, and a second ALU instruction to be executed by the second ALU, e.g., as described below.

[000372] In some demonstrative aspects, compiler 160 may be configured generate another instruction scheduling including the first ALU instruction to be executed by the first ALU (ALU1), and the second ALU instruction to be executed by the second ALU (ALU2), e.g., as follows:

Cycle	**ALU1**	**ALU2**	**Comments**
0	%0, _= mul(a, b, _, _)		_ is the annotation of unused
1			
2	%2 = mul(%0, 3)	%1 = mul(c, d,_, _)	_ is the annotation of unused
3			
4	%3 = add(%2, %1)		

TABLE (3)

[000373] As shown in Table 3, the first ALU instruction, e.g., *%0=a*b*, may be executed, e.g., by the first ALU, during the cycle 0.

[000374] As shown in Table 3, the second ALU instruction, e.g., *%1=c*d*, may be executed, e.g., by the second ALU, during the cycle 2.

[000375] In some demonstrative aspects, one or more of the variables according to the instruction scheduling of Table 3 may have live ranges, which may be different, for example, from the live ranges of Table 2, e.g., as follows.

Value	Live Range in Cycles [Start, End] (inclusive)
%0	[2, 2]
%1	[4, 4]
%2	[4, 4]
%3	[6, 6]

TABLE (4)

[000376] As shown in Table 4, the variable *%1* may be live only during the cycle 4, and/or may be free during other cycles.

[000377] For example, the variable *%1* may occupy a register only during the cycle 4.

[000378] Accordingly, the instruction scheduling of Table 3 may provide a technical solution to reduce utilization of registers, reduce register pressure, and/or reduce a number of used registers.

[000379] In some demonstrative aspects, a number of cycles of the executed program according to the instruction set of Table 1 may be equal to a number of cycles of the executed program according to the instruction set of Table 3, e.g., 5 cycles.

[000380] In some demonstrative aspects, the executed program according to the instruction set of Table 3 may utilize a reduced number of registers, e.g., compared to the executed program according to the instruction set of Table 1.

[000381] Accordingly, the instruction set of Table 3 may be implemented to provide a technical solution to improve performance of the executed program.

[000382] In some demonstrative aspects, compiler 160 may be configured to determine the selected compilation scheme, for example, according to a predefined selection criterion, which may be based, for example, on a live range of at least one input variable to be input for execution of the SIMD instruction, for example, according to the first compilation scheme, e.g., as described below.

[000383] In one example, source code 112 may include an instruction scheduling, which may be configured to calculate a first expression (data operation) and a second expression (data operation) based on a plurality of variables, for example, including two variables, denoted $a,\ b$, e.g., as follows:

$$(3 - a) \cdot b \quad (2)$$

$$4 \cdot a \quad (3)$$

[000384] In some demonstrative aspects, compiler 160 may be configured to load the variable *a,* for example, from a first memory location *(a_add),* and/or to load the variable *b,* for example, from a second memory location (*b_add).*

[000385] In some demonstrative aspects, compiler 160 may be configured to store the variables $a,\ b,$ in two registers, denoted, *R0,* and *R1,* respectively.

[000386] For example, compiler 160 may be configured to compile the source code 112 for a processor, e.g., a SIMD processor, which may include a single ALU, e.g., ALU1.

[000387] For example, the ALU of the processor may be configured to perform SIMD instructions, e.g., add, multiply and negative instructions, for example, with a latency of 2 cycles.

[000388] For example, the SIMD processor may have a memory unit with a latency of 1 cycle for a memory access operation, e.g., a store operation to store a value to the memory unit, or a load operation to load a value from the memory unit.

[000389] For example, execution of the Expressions 2 and 3 may be implemented according to the following instruction scheduling, e.g., using a SIMD instruction:

Cycle	**ALU1**	**Load/Store Unit**	**Comments**
0		a= load(a_add)	
1	%0 = neg(a)		
2			

3	%1 = add(%0, 3)		
4		b = load(b_add)	
5	%2, %3 = mul(%1, b, 4, a)		$\%2 = (3 - a) \cdot b,$ $\%3 = 4 \cdot a$ At First, we assume that it is beneficial to use the SIMD mul instruction.
6			
7		store(%2, $res1\_add$)	
8		store(%3, $res2\_add$)	

TABLE (5)

[000390] As shown in Table 5, the instructions of the executed program may be performed using the single ALU1.

[000391] As shown in Table 5, the execution of the Expression 2 and the Expression 3 may be performed during 9 cycles.

[000392] As Shown in Table 5, the instruction scheduling may include a SIMD instruction to be executed in cycle 5.

[000393] For example, as shown in Table 5, the SIMD instruction may include a combination of a first instruction, e.g., *%2=%1*b=(3-a)*b*, and a second instruction, e.g., *%3=4*a*, which may be executed in parallel by the ALU1.

[000394] In some demonstrative aspects, compiler 160 may be configured to identify live ranges of variables in the instruction scheduling, and/or to assign registers to the variables, for example, according to Table 5, e.g., as follows:

Value	**Live Range in Cycles [Start, End] (Inclusive)**	**Assigned Register**
a	[1, 5]	R0
b	[5, 5]	R1
%0	[3, 3]	R1
%1	[5, 5]	R2
%2	[7, 7]	R0
%3	[7, 8]	R1

TABLE (6)

[000395] As shown in Table 6, the register R0 may be occupied between the cycle 1 and the cycle 5, for example, to store the variable *a*, which may be used as an input for the second instruction in the SIMD instruction.

[000396] As shown in Table 6, the register R1 may be occupied during the cycle 5, for example, to store the variable *b*.

[000397] As shown in Table 6, the variable *%1,* which may result from the data operation *%1 = add(%0, 3)*, which may be used as an input for the first instruction in the SIMD instruction, may be live during the cycle 5. For example, this may require occupying a register, e.g. the register R2, to store the value of the variable *%1*, for example, during the cycle 5. For example, the variable *%1* may occupy the register R2, for example, as the register R0 and the register R1 may be occupied during the cycle 5.

[000398] In one example, the requirement to reserve the register for storing the variable *%1* during the cycle 5 may result in register pressure.

[000399] In some demonstrative aspects, compiler 160 may identify that using the SIMD instruction may result in register pressure.

[000400] In some demonstrative aspects, compiler 160 may identify that the SIMD instruction results in asymmetric use of registers, e.g., the input variable *b* is live only in cycle 5, while the input variable *a* is live between the cycles 1-5.

[000401] In some demonstrative aspects, compiler 160 may be configured to determine that it may be preferable to compile the second data operation, e.g., *%3=4*a*, into a first multiply instruction to be executed by ALU1, and to compile the first data operation, e.g., *%2=(3-a)*b*, into a second multiply instruction to be executed by ALU1, e.g., after the first multiply instruction.

[000402] In some demonstrative aspects, compiler 160 may be configured to determine that it may be preferable to split the SIMD instruction into the first multiply instruction and the second multiply instruction.

[000403] In some demonstrative aspects, compiler 160 may be configured generate another instruction scheduling including the first multiply instruction to be executed by ALU1, and the second multiply instruction to be executed by ALU1, for example, after the first multiply instruction, e.g., as follows:

0		a = load(a_add)	
1	%0 = neg(a)		
2	%3, _ = mul(4, a, _ , _)		_ is the annotation of unused
3	%1 = add(%0, 3)	b = load(b_add)	
4		store(%3, $res2\_add$)	
5	%2, _ = mul(%1, b, _ , _)		$\%2 = (3 - a) \cdot b$
6			
7		store(%2, $res1\_add$)	

TABLE (7)

[000404] As shown in Table 7, the first multiply instruction, e.g., *%3=4*a*, may be executed by ALU1 during the cycle 2.

[000405] As shown in Table 7, the second multiply instruction, e.g., *%2=%1*b=(3-a)*b*, may be executed by ALU1 during the cycle 5.

[000406] In some demonstrative aspects, the variables according to the instruction scheduling of Table 7 may have live ranges, which may be different, for example, from the live ranges of Table 6, and/or an assignment of registers to the variables according to the instruction scheduling of Table 7 may be different, for example, from the assignment of registers according to the Table 6, e.g., as follows.

Value	Live Range in Cycles [Start, End] (Including)	Assigned Register
a	[1, 2]	R0
b	[4, 5]	R0
%0	[3, 3]	R0
%1	[5, 5]	R1
%2	[7, 7]	R0
%3	[4, 4]	R1

TABLE (8)

[000407] As shown in Table 8, the variable *a* may be live between the cycle 1 and the cycle 2, and the variable *b* may be live between the cycle 4 and the cycle 5. For example, the separate, non-overlapping, live ranges of the variables *a* and *b* may be utilized, for example, to assign the variable *a* and the variable *b* to occupy a same register, e.g., *R0*, between the cycle 1 and the cycle 5, for example, instead of occupying two registers, e.g., the registers *R0* and *R1*. Accordingly, the instruction scheduling of Table 7 may

be implemented to provide a technical solution to utilize the register R1, e.g., instead of the register R2, for example, to store the variable *%1* during the cycle 5.

[000408] Accordingly, the instruction scheduling of Table 7 may provide a technical solution to reduce utilization of registers, reduce register pressure, and/or reduce a number of used registers.

[000409] In some demonstrative aspects, a number of cycles of the executed program according to the instruction set of Table 7, e.g., 8 cycles, may be less than a number of cycles of the executed program according to the instruction set of Table 5, e.g., 9 cycles.

[000410] In some demonstrative aspects, the executed program according to the instruction set of Table 7 may utilize a reduced number of registers, e.g., 2 registers, for example, compared to the executed program according to the instruction set of Table 5, e.g., 3 registers.

[000411] Accordingly, the instruction set of Table 7 may be implemented to provide a technical solution to improve both performance of the executed program as well as register pressure of the executed program.

[000412] In one example, an attempt to schedule execution of the Expressions 2 and 3 according to the instruction scheduling of Table 5 may result in an unsuccessful register allocation, e.g., if only two registers are available. One option to address this situation may be to relax the instruction scheduling, e.g., in attempt to reduce the number of live variables sharing the same execution cycles. However, this option may result in degraded performance.

[000413] In some demonstrative aspects, execution of the Expressions 2 and 3 according to the register allocation scheme described above, e.g., using the instruction scheduling of Table 7, may provide a technical solution to support successful register allocation, e.g., even if only two registers are available, for example, while avoiding the performance degradation resulting from the instruction scheduling of Table 4.

[000414] Reference is made to Fig. 4, which schematically illustrates a method of compiling code for a processor. For example, one or more operations of the method of Fig. 4 may be performed by a system, e.g., system 100 (Fig. 1); a device, e.g., device 102 (Fig. 1); a server, e.g., server 170 (Fig. 1); and/or a compiler, e.g., compiler 160 (Fig. 1), and/or compiler 200 (Fig. 2).

[000415] In some demonstrative aspects, as indicated at block 402, the method may include processing code to identify one or more pairs of data operations and/or instructions, e.g., similar and independent data operations and/or instructions, which may be potentially combined into an SIMD instruction. For example, compiler 160 (Fig. 1) may identify a pair of similar independent instructions, which may potentially be combined into a single SIMD, e.g., as described above.

[000416] In some demonstrative aspects, as indicated at block 404, the method may include performing instruction scheduling and register allocation, for example, using the SIMD instruction. For example, compiler 160 (Fig. 1) may perform instruction scheduling and register allocation using the SIMD instruction, e.g., as described above.

[000417] In some demonstrative aspects, as indicated at block 406, the method may include determining whether the SIMD instruction may, e.g., should, be replaced by two separate instructions, e.g., as described below.

[000418] In some demonstrative aspects, as indicated at block 406, the method may include observing the instructions, e.g., each of the instructions, which are combined into the SIMD instruction, for example, to determine if the single SIMD has an asymmetric use-definition chain, for example, such that the parallel execution of the pair of instructions adds constraints to the scheduling, which may lead to large live ranges. For example. compiler 160 (Fig. 1) may determine if an instruction in the SIMD instruction leads to a large live range, e.g., as described above.

[000419] In some demonstrative aspects, as indicated at block 406, the method may include splitting the SIMD instruction into separate instructions, for example, based on a determination that there are available resources, e.g., ALU resources, to perform the separate instructions. For example, compiler 160 (Fig. 1) may split the SIMD instruction into separate instructions, for example, if it is determined that ALU resources are available to perform the separate instructions, e.g., as described above.

[000420] In some demonstrative aspects, as indicated at block 408, the method may include generating an instruction scheduling and register allocation based on the separate instructions, e.g., after splitting the SIMD instruction, for example, to reduce ranges of the live intervals. For example, compiler 160 (Fig. 1) may generate target code 115 (Fig. 1) based on an instruction scheduling and the register allocation, which

may be configured to reduce the size of the live range of the variable *%1*, e.g., as described above.

[000421] Reference is made to Fig. 5, which schematically illustrates a method of compiling code for a processor. For example, one or more operations of the method of Fig. 5 may be performed by a system, e.g., system 100 (Fig. 1); a device, e.g., device 102 (Fig. 1); a server, e.g., server 170 (Fig. 1); and/or a compiler, e.g., compiler 160 (Fig. 1), and/or compiler 200 (Fig. 2).

[000422] In some demonstrative aspects, as indicated at block 502, the method may include identify, e.g., based on a source code, a first data operation and a second data operation, which may be executable in parallel according to a SIMD instruction, e.g., to be executed by a single ALU of a target processor. For example, compiler 160 (Fig. 1) may be configured identify the first and second data operations, for example, based on the source code 112 (Fig. 1), e.g., as descried above.

[000423] In some demonstrative aspects, as indicated at block 504, the method may include determining a selected compilation scheme from a first compilation scheme and a second compilation scheme based on a predefined selection criterion. For example, the first compilation scheme may include compilation of the first data operation and the second data operation into the SIMD instruction, and the second compilation scheme may include compilation of the first data operation into a first ALU instruction, and compilation of the second data operation into a second ALU instruction to be executed separately from the first ALU instruction. For example, compiler 160 (Fig. 1) may be configured to identify the selected compilation scheme to compile the first and second data operations, e.g., as described above.

[000424] In some demonstrative aspects, as indicated at block 506, the method may include generating target code, which may be configured for execution by the target processor. For example, the target code may be based on compilation of the first data operation and the second data operation according to the selected compilation scheme. For example, compiler 160 (Fig. 1) may be configured to generate the target code 115 (Fig. 1), for example, by compilation of the first and second data operations according to the selected compilation scheme, e.g., as described above.

[000425] Reference is made to Fig. 6, which schematically illustrates a product of manufacture 600, in accordance with some demonstrative aspects. Product 600 may include one or more tangible computer-readable ("machine-readable") non-transitory storage media 602, which may include computer-executable instructions, e.g., implemented by logic 604, operable to, when executed by at least one computer processor, enable the at least one computer processor to implement one or more operations at device 102 (Fig. 1), server 170 (Fig. 1), and/or compiler 160 (Fig. 1), to cause device 102 (Fig. 1), server 170 (Fig. 1), and/or compiler 160 (Fig. 1) to perform, trigger and/or implement one or more operations and/or functionalities, and/or to perform, trigger and/or implement one or more operations and/or functionalities described with reference to the Figs. 1-5, and/or one or more operations described herein. The phrases "non-transitory machine-readable medium" and "computer-readable non-transitory storage media" may be directed to include all computer-readable media, with the sole exception being a transitory propagating signal.

[000426] In some demonstrative aspects, product 600 and/or machine-readable storage media 602 may include one or more types of computer-readable storage media capable of storing data, including volatile memory, non-volatile memory, removable or non-removable memory, erasable or non-erasable memory, writeable or re-writeable memory, and the like. For example, machine-readable storage media 602 may include, RAM, DRAM, Double-Data-Rate DRAM (DDR-DRAM), SDRAM, static RAM (SRAM), ROM, programmable ROM (PROM), erasable programmable ROM (EPROM), electrically erasable programmable ROM (EEPROM), flash memory (e.g., NOR or NAND flash memory), content addressable memory (CAM), polymer memory, phase-change memory, ferroelectric memory, silicon-oxide-nitride-oxide-silicon (SONOS) memory, a disk, a hard drive, and the like. The computer-readable storage media may include any suitable media involved with downloading or transferring a computer program from a remote computer to a requesting computer carried by data signals embodied in a carrier wave or other propagation medium through a communication link, e.g., a modem, radio or network connection.

[000427] In some demonstrative aspects, logic 604 may include instructions, data, and/or code, which, if executed by a machine, may cause the machine to perform a method, process and/or operations as described herein. The machine may include, for

example, any suitable processing platform, computing platform, computing device, processing device, computing system, processing system, computer, processor, or the like, and may be implemented using any suitable combination of hardware, software, firmware, and the like.

[000428] In some demonstrative aspects, logic 604 may include, or may be implemented as, software, a software module, an application, a program, a subroutine, instructions, an instruction set, computing code, words, values, symbols, and the like. The instructions may include any suitable type of code, such as source code, compiled code, interpreted code, executable code, static code, dynamic code, and the like. The instructions may be implemented according to a predefined computer language, manner or syntax, for instructing a processor to perform a certain function. The instructions may be implemented using any suitable high-level, low-level, object-oriented, visual, compiled and/or interpreted programming language, machine code, and the like.

EXAMPLES

[000429] The following examples pertain to further aspects.

[000430] Example 1 includes a product comprising one or more tangible computer-readable non-transitory storage media comprising computer-executable instructions operable to, when executed by at least one processor, enable the at least one processor to cause a computing device to identify, based on a source code, a first data operation and a second data operation, which are executable in parallel according to a Single Instruction/Multiple Data (SIMD) instruction to be executed by a single Arithmetic Logic Unit (ALU) of a target processor; determine a selected compilation scheme from a first compilation scheme and a second compilation scheme based on a predefined selection criterion, wherein the first compilation scheme comprises compilation of the first data operation and the second data operation into the SIMD instruction, wherein the second compilation scheme comprises compilation of the first data operation into a first ALU instruction, and compilation of the second data operation into a second ALU instruction to be executed separately from the first ALU instruction; and generate target code configured for execution by the target processor, the target code is based on

compilation of the first data operation and the second data operation according to the selected compilation scheme.

[000431] Example 2 includes the subject matter of Example 1, and optionally, wherein the predefined selection criterion comprises a register-utilization criterion corresponding to a register utilization of one or more registers of the target processor.

[000432] Example 3 includes the subject matter of Example 2, and optionally, wherein the register-utilization criterion is based on the register utilization of the one or more registers of the target processor according to the first compilation scheme.

[000433] Example 4 includes the subject matter of Example 2 or 3, and optionally, wherein the register-utilization criterion is configured such that the selected compilation scheme is to include the second compilation scheme when a first register utilization of the one or more registers of the target processor according to the first compilation scheme is greater than a second register utilization of the one or more registers of the target processor according to the second compilation scheme.

[000434] Example 5 includes the subject matter of any one of Examples 1-4, and optionally, wherein the predefined selection criterion is based on a live range of at least one variable corresponding to at least one of the first data operation or the second data operation.

[000435] Example 6 includes the subject matter of Example 5, and optionally, wherein the at least one variable comprises at least one of an input variable of the first data operation or the second data operation, or an output variable resulting from the first data operation or the second data operation.

[000436] Example 7 includes the subject matter of Example 5 or 6, and optionally, wherein the predefined selection criterion is based on a live range of an input variable of the SIMD instruction, or an output variable resulting from the SIMD instruction.

[000437] Example 8 includes the subject matter of any one of Examples 5-7, and optionally, wherein the predefined selection criterion is configured such that the selected compilation scheme is to include the second compilation scheme when a first live range of the variable according to the first compilation scheme is greater than a second live range of the variable according to the second compilation scheme.

[000438] Example 9 includes the subject matter of any one of Examples 1-8, and optionally, wherein the predefined selection criterion is based on a count of cycles between a first cycle in which a result of the SIMD instruction is to be available in a register of the target processor, and a second cycle subsequent to the first cycle, in which the result of the SIMD instruction is to be retrieved from the register of the target processor.

[000439] Example 10 includes the subject matter of any one of Examples 1-9, and optionally, wherein the predefined selection criterion is based on a count of cycles between a first cycle in which an input variable of the SIMD instruction is to be available in a register of the target processor, and a second cycle subsequent to the first cycle, in which the input variable of the SIMD instruction is to be retrieved from the register of the target processor.

[000440] Example 11 includes the subject matter of any one of Examples 1-10, and optionally, wherein the predefined selection criterion is based on a difference between a live range of a first variable and a live range of a second variable, the first variable resulting from execution of the first data operation by the SIMD instruction, the second variable resulting from execution of the second data operation by the SIMD instruction.

[000441] Example 12 includes the subject matter of any one of Examples 1-11, and optionally, wherein the predefined selection criterion is based on a difference between a live range of a first variable and a live range of a second variable, the first variable comprising an input for execution of the first data operation by the SIMD instruction, the second variable comprising an input for execution of the second data operation by the SIMD instruction.

[000442] Example 13 includes the subject matter of any one of Examples 1-12, and optionally, wherein the predefined selection criterion is based on a latency of the ALU the target processor to execute the SIMD instruction.

[000443] Example 14 includes the subject matter of any one of Examples 1-13, and optionally, wherein the first ALU instruction is configured to be executed in a first execution cycle, and the second ALU instruction is configured to be executed in a second execution cycle different from the first execution cycle.

[000444] Example 15 includes the subject matter of any one of Examples 1-14, and optionally, wherein the second compilation scheme comprises compilation of the first data operation into the first ALU instruction to be executed during a first execution cycle, and compilation of the second data operation into the second ALU instruction to be executed during a second cycle, which is different from the first cycle.

[000445] Example 16 includes the subject matter of any one of Examples 1-15, and optionally, wherein the first ALU instruction is to be executed by a first ALU, and the second ALU instruction is to be executed by a second ALU.

[000446] Example 17 includes the subject matter of any one of Examples 1-15, and optionally, wherein the first ALU instruction and the second ALU instruction are to be executed by a same ALU.

[000447] Example 18 includes the subject matter of any one of Examples 1-17, and optionally, wherein the first data operation and the second data operation comprise data operations of a same instruction.

[000448] Example 19 includes the subject matter of any one of Examples 1-17, and optionally, wherein the first data operation comprises a data operation of a first instruction, and the second data operation comprises a data operation of a second instruction separate from the first instruction.

[000449] Example 20 includes the subject matter of any one of Examples 1-19, and optionally, wherein the source code comprises Open Computing Language (OpenCL) code.

[000450] Example 21 includes the subject matter of any one of Examples 1-20, and optionally, wherein the computer-executable instructions, when executed, cause the computing device to compile the source code into the target code according to a Low Level Virtual Machine (LLVM) based (LLVM-based) compilation scheme.

[000451] Example 22 includes the subject matter of any one of Examples 1-21, and optionally, wherein the target code is configured for execution by a Very Long Instruction Word (VLIW) Single Instruction/Multiple Data (SIMD) target processor.

[000452] Example 23 includes the subject matter of any one of Examples 1-22, and optionally, wherein the target code is configured for execution by a target vector processor.

[000453] Example 24 includes a compiler configured to perform any of the described operations of any of Examples 1-23.

[000454] Example 25 includes a computing device configured to perform any of the described operations of any of Examples 1-23.

[000455] Example 26 includes a computing system comprising at least one memory to store instructions; and at least one processor to retrieve instructions from the memory and execute the instructions to cause the computing system to perform any of the described operations of any of Examples 1-23.

[000456] Example 27 includes a computing system comprising a compiler to generate target code according to any of the described operations of any of Examples 1-23, and a processor to execute the target code.

[000457] Example 28 comprises an apparatus comprising means for executing any of the described operations of any of Examples 1-23.

[000458] Example 29 comprises an apparatus comprising: a memory interface; and processing circuitry configured to: perform any of the described operations of any of Examples 1-23.

[000459] Example 30 comprises a method comprising any of the described operations of any of Examples 1-23.

[000460] Functions, operations, components and/or features described herein with reference to one or more aspects, may be combined with, or may be utilized in combination with, one or more other functions, operations, components and/or features described herein with reference to one or more other aspects, or vice versa.

[000461] While certain features have been illustrated and described herein, many modifications, substitutions, changes, and equivalents may occur to those skilled in the art. It is, therefore, to be understood that the appended claims are intended to cover all such modifications and changes as fall within the true spirit of the disclosure.

CLAIMS

What is claimed is:

1. A product comprising one or more tangible computer-readable non-transitory storage media comprising computer-executable instructions operable to, when executed by at least one processor, enable the at least one processor to cause a computing device to:

identify, based on a source code, a first data operation and a second data operation, which are executable in parallel according to a Single Instruction/Multiple Data (SIMD) instruction to be executed by a single Arithmetic Logic Unit (ALU) of a target processor;

determine a selected compilation scheme from a first compilation scheme and a second compilation scheme based on a predefined selection criterion, wherein the first compilation scheme comprises compilation of the first data operation and the second data operation into the SIMD instruction, wherein the second compilation scheme comprises compilation of the first data operation into a first ALU instruction, and compilation of the second data operation into a second ALU instruction to be executed separately from the first ALU instruction; and

generate target code configured for execution by the target processor, the target code is based on compilation of the first data operation and the second data operation according to the selected compilation scheme.

2. The product of claim 1, wherein the predefined selection criterion comprises a register-utilization criterion corresponding to a register utilization of one or more registers of the target processor.

3. The product of claim 2, wherein the register-utilization criterion is based on the register utilization of the one or more registers of the target processor according to the first compilation scheme.

4. The product of claim 2, wherein the register-utilization criterion is configured such that the selected compilation scheme is to include the second compilation scheme when a first register utilization of the one or more registers of the target processor according to the first compilation scheme is greater than a second register utilization of

the one or more registers of the target processor according to the second compilation scheme.

5. The product of claim 1, wherein the predefined selection criterion is based on a live range of at least one variable corresponding to at least one of the first data operation or the second data operation.

6. The product of claim 5, wherein the at least one variable comprises at least one of an input variable of the first data operation or the second data operation, or an output variable resulting from the first data operation or the second data operation.

7. The product of claim 5, wherein the predefined selection criterion is based on a live range of an input variable of the SIMD instruction, or an output variable resulting from the SIMD instruction.

8. The product of claim 5, wherein the predefined selection criterion is configured such that the selected compilation scheme is to include the second compilation scheme when a first live range of the variable according to the first compilation scheme is greater than a second live range of the variable according to the second compilation scheme.

9. The product of any one of claims 1-8, wherein the predefined selection criterion is based on a count of cycles between a first cycle in which a result of the SIMD instruction is to be available in a register of the target processor, and a second cycle subsequent to the first cycle, in which the result of the SIMD instruction is to be retrieved from the register of the target processor.

10. The product of any one of claims 1-8, wherein the predefined selection criterion is based on a count of cycles between a first cycle in which an input variable of the SIMD instruction is to be available in a register of the target processor, and a second cycle subsequent to the first cycle, in which the input variable of the SIMD instruction is to be retrieved from the register of the target processor.

11. The product of any one of claims 1-8, wherein the predefined selection criterion is based on a difference between a live range of a first variable and a live range of a second variable, the first variable resulting from execution of the first data

operation by the SIMD instruction, the second variable resulting from execution of the second data operation by the SIMD instruction.

12. The product of any one of claims 1-8, wherein the predefined selection criterion is based on a difference between a live range of a first variable and a live range of a second variable, the first variable comprising an input for execution of the first data operation by the SIMD instruction, the second variable comprising an input for execution of the second data operation by the SIMD instruction.

13. The product of any one of claims 1-8, wherein the predefined selection criterion is based on a latency of the ALU the target processor to execute the SIMD instruction.

14. The product of any one of claims 1-8, wherein the first ALU instruction is configured to be executed in a first execution cycle, and the second ALU instruction is configured to be executed in a second execution cycle different from the first execution cycle.

15. The product of any one of claims 1-8, wherein the second compilation scheme comprises compilation of the first data operation into the first ALU instruction to be executed during a first execution cycle, and compilation of the second data operation into the second ALU instruction to be executed during a second cycle, which is different from the first cycle.

16. The product of any one of claims 1-8, wherein the first ALU instruction is to be executed by a first ALU, and the second ALU instruction is to be executed by a second ALU.

17. The product of any one of claims 1-8, wherein the first ALU instruction and the second ALU instruction are to be executed by a same ALU.

18. The product of any one of claims 1-8, wherein the first data operation and the second data operation comprise data operations of a same instruction.

19. The product of any one of claims 1-8, wherein the first data operation comprises a data operation of a first instruction, and the second data operation comprises a data operation of a second instruction separate from the first instruction.

20. The product of any one of claims 1-8, wherein the source code comprises Open Computing Language (OpenCL) code.

21. The product of any one of claims 1-8, wherein the computer-executable instructions, when executed, cause the computing device to compile the source code into the target code according to a Low Level Virtual Machine (LLVM) based (LLVM-based) compilation scheme.

22. The product of any one of claims 1-8, wherein the target code is configured for execution by a Very Long Instruction Word (VLIW) Single Instruction/Multiple Data (SIMD) target processor.

23. The product of any one of claims 1-8, wherein the target code is configured for execution by a target vector processor.

24. A computing system comprising:

at least one memory to store instructions; and

at least one processor to retrieve the instructions from the memory and to execute the instructions to cause the computing system to:

identify, based on a source code, a first data operation and a second data operation, which are executable in parallel according to a Single Instruction/Multiple Data (SIMD) instruction to be executed by a single Arithmetic Logic Unit (ALU) of a target processor;

determine a selected compilation scheme from a first compilation scheme and a second compilation scheme based on a predefined selection criterion, wherein the first compilation scheme comprises compilation of the first data operation and the second data operation into the SIMD instruction, wherein the second compilation scheme comprises compilation of the first data operation into a first ALU instruction, and compilation of the second data operation into a second ALU instruction to be executed separately from the first ALU instruction; and

generate target code configured for execution by the target processor, the target code is based on compilation of the first data operation and the second data operation according to the selected compilation scheme.

25. The computing system of claim 24, wherein the predefined selection criterion comprises a register-utilization criterion corresponding to a register utilization of one or more registers of the target processor.

26. The computing system of claim 24 or 25 comprising the target processor.

27. A method comprising:

identifying, based on a source code, a first data operation and a second data operation, which are executable in parallel according to a Single Instruction/Multiple Data (SIMD) instruction to be executed by a single Arithmetic Logic Unit (ALU) of a target processor;

determining a selected compilation scheme from a first compilation scheme and a second compilation scheme based on a predefined selection criterion, wherein the first compilation scheme comprises compilation of the first data operation and the second data operation into the SIMD instruction, wherein the second compilation scheme comprises compilation of the first data operation into a first ALU instruction, and compilation of the second data operation into a second ALU instruction to be executed separately from the first ALU instruction; and

generating target code configured for execution by the target processor, the target code is based on compilation of the first data operation and the second data operation according to the selected compilation scheme.

28. The method of claim 27, wherein the predefined selection criterion is based on a live range of at least one variable corresponding to at least one of the first data operation or the second data operation.

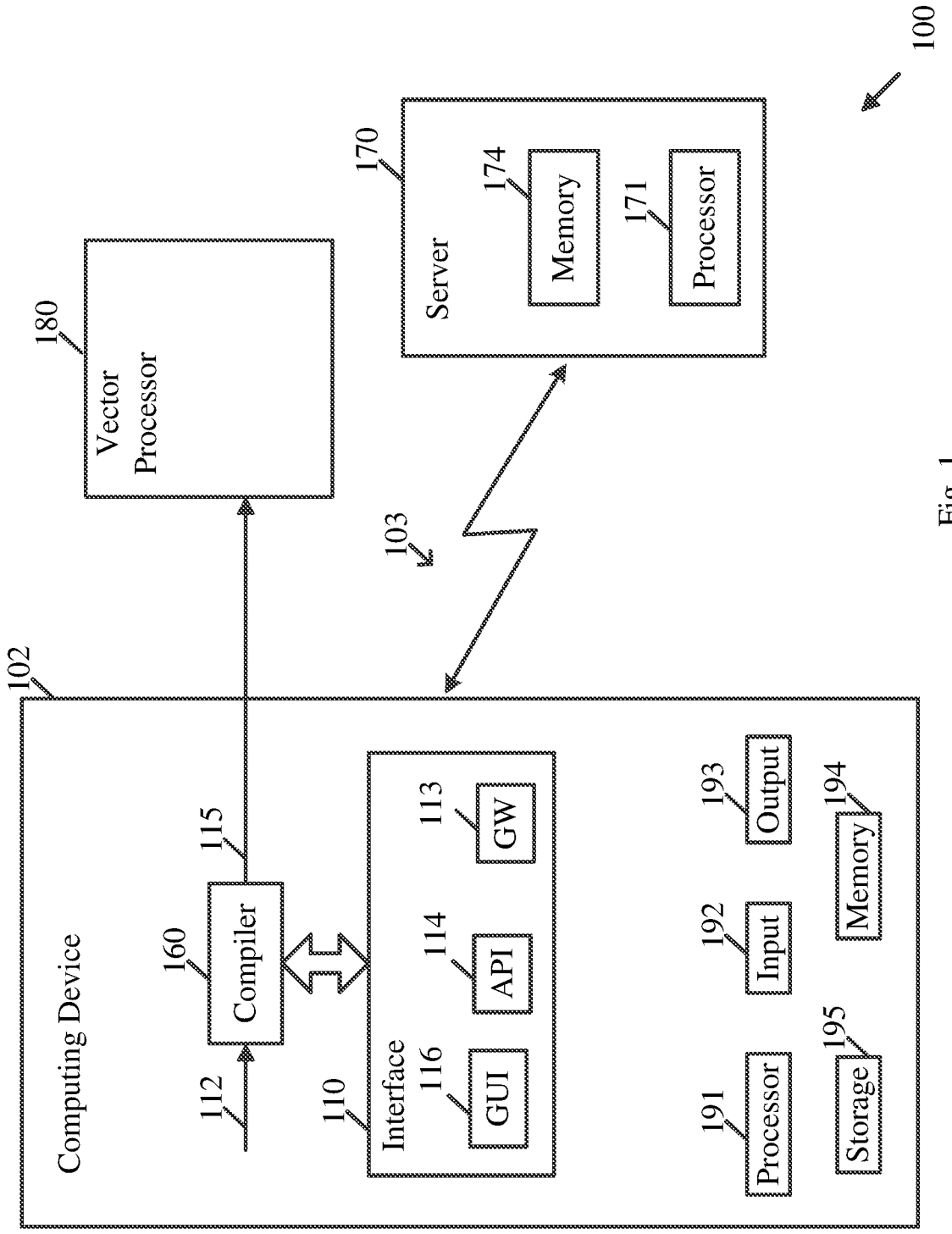
100
102
Computing Device
103
180
Vector Processor
170
Server
174
Memory
171
Processor
160
Compiler
115
112
110
Interface
116
GUI
114
API
113
GW
191
Processor
192
Input
193
Output
194
Memory
195
Storage

Fig. 1

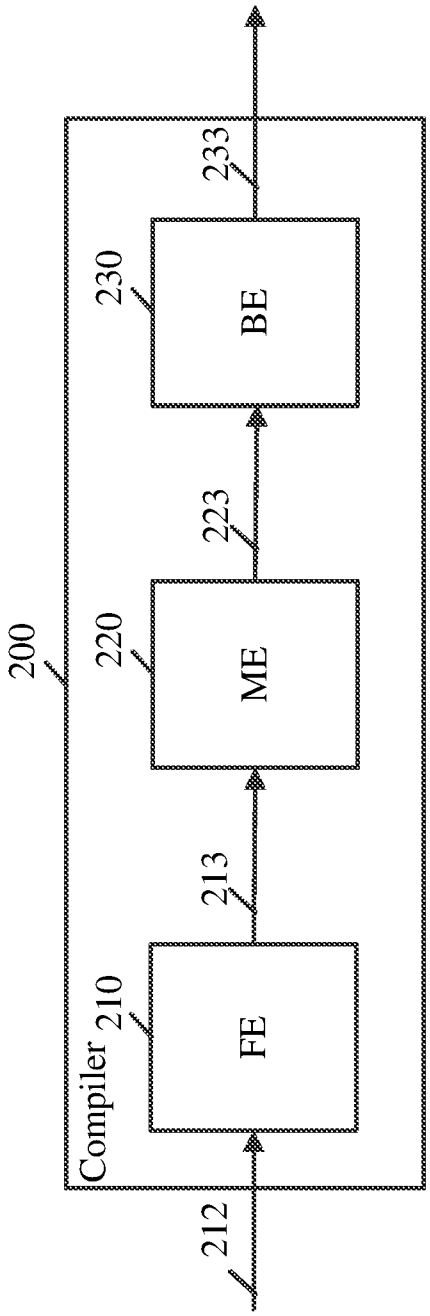

Fig. 2

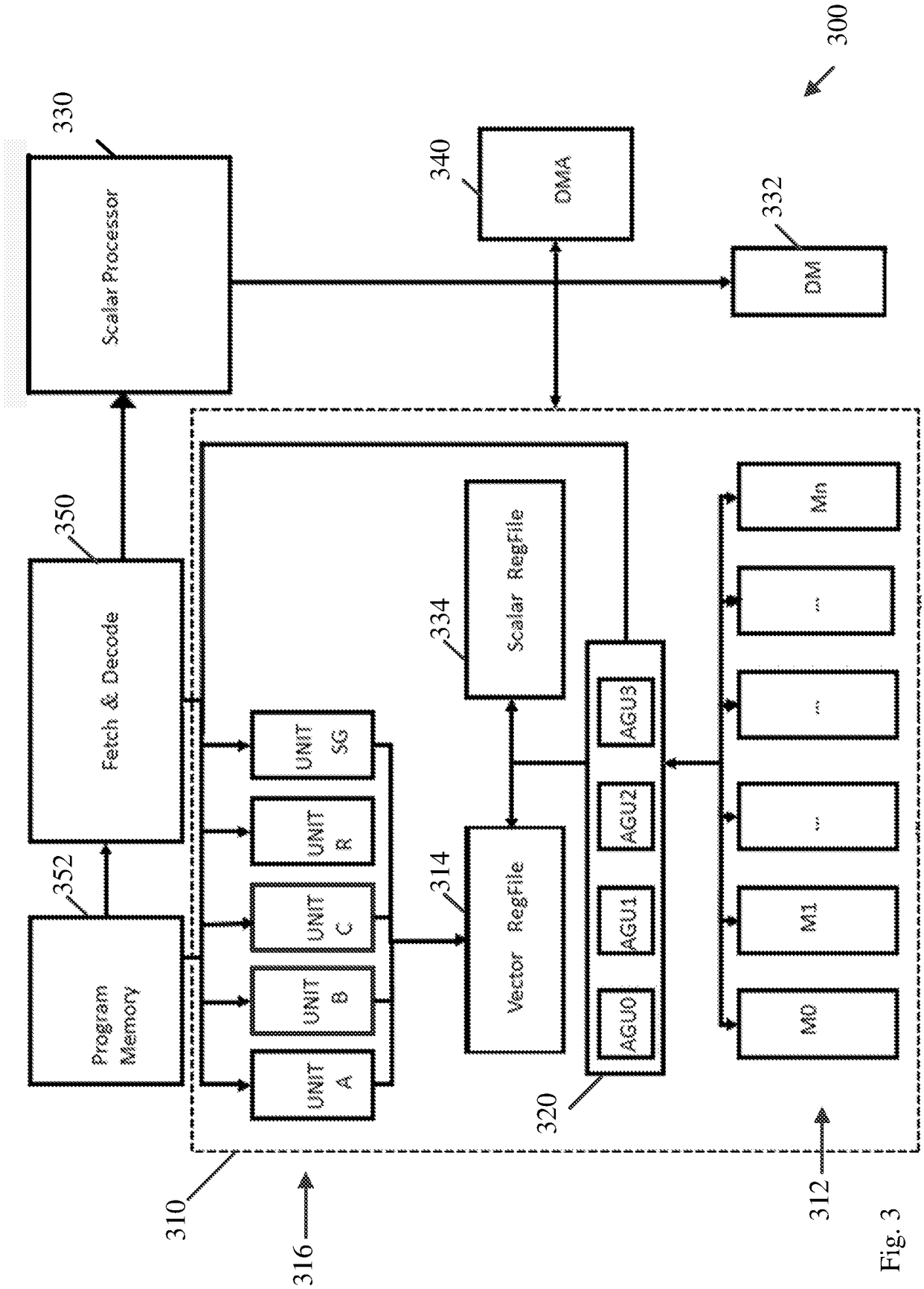
300
330
Scalar Processor
340
DMA
332
DM
350
Fetch & Decode
352
Program Memory
UNIT A
UNIT B
UNIT C
UNIT R
UNIT SG
314
Vector RegFile
334
Scalar RegFile
320
AGU0
AGU1
AGU2
AGU3
M0
M1
...
...
...
Mn
310
316
312

Fig. 3

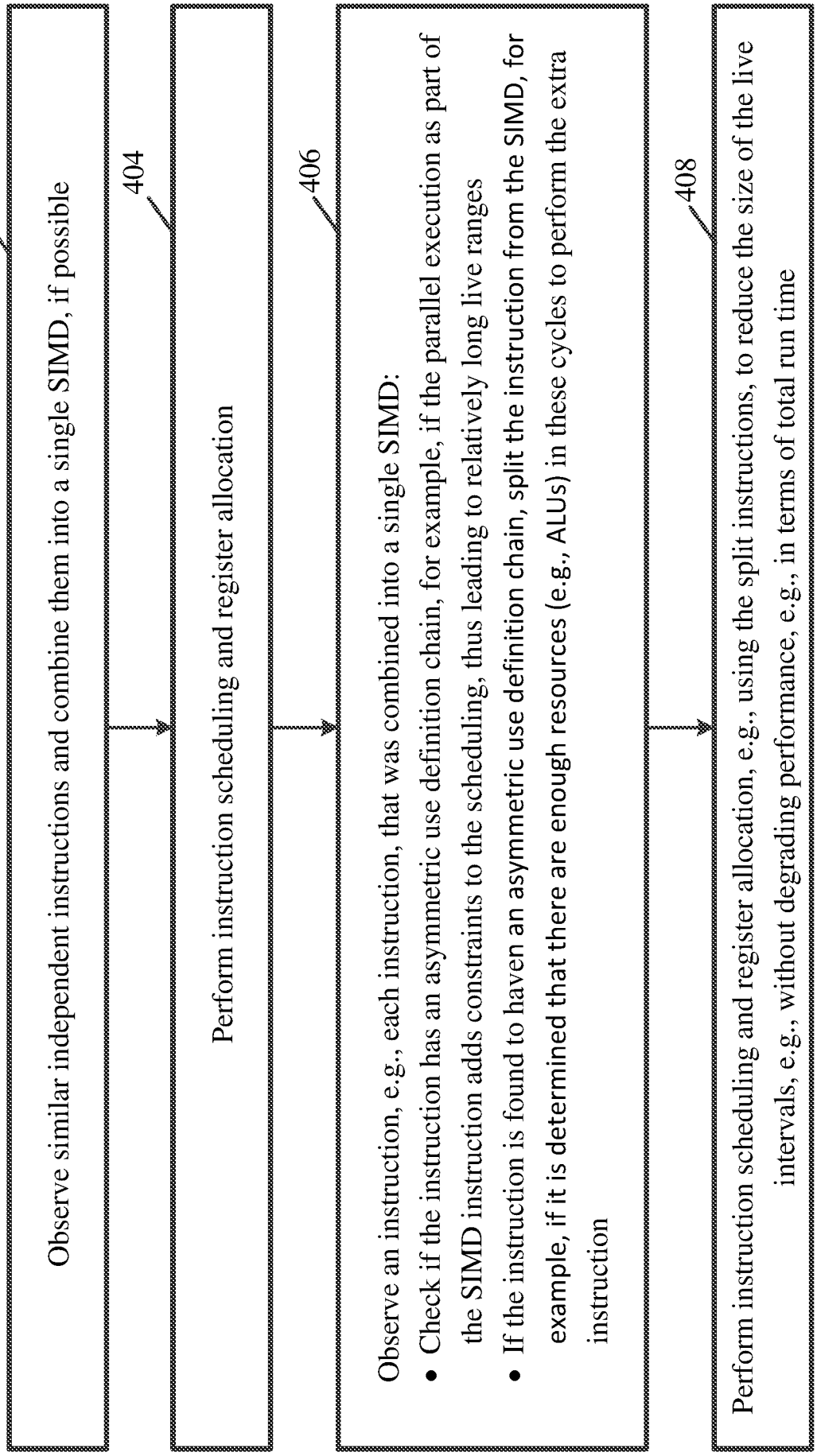
402
Observe similar independent instructions and combine them into a single SIMD, if possible
404
Perform instruction scheduling and register allocation
406
Observe an instruction, e.g., each instruction, that was combined into a single SIMD:
• Check if the instruction has an asymmetric use definition chain, for example, if the parallel execution as part of the SIMD instruction adds constraints to the scheduling, thus leading to relatively long live ranges
• If the instruction is found to haven an asymmetric use definition chain, split the instruction from the SIMD, for example, if it is determined that there are enough resources (e.g., ALUs) in these cycles to perform the extra instruction
408
Perform instruction scheduling and register allocation, e.g., using the split instructions, to reduce the size of the live intervals, e.g., without degrading performance, e.g., in terms of total run time

Fig. 4

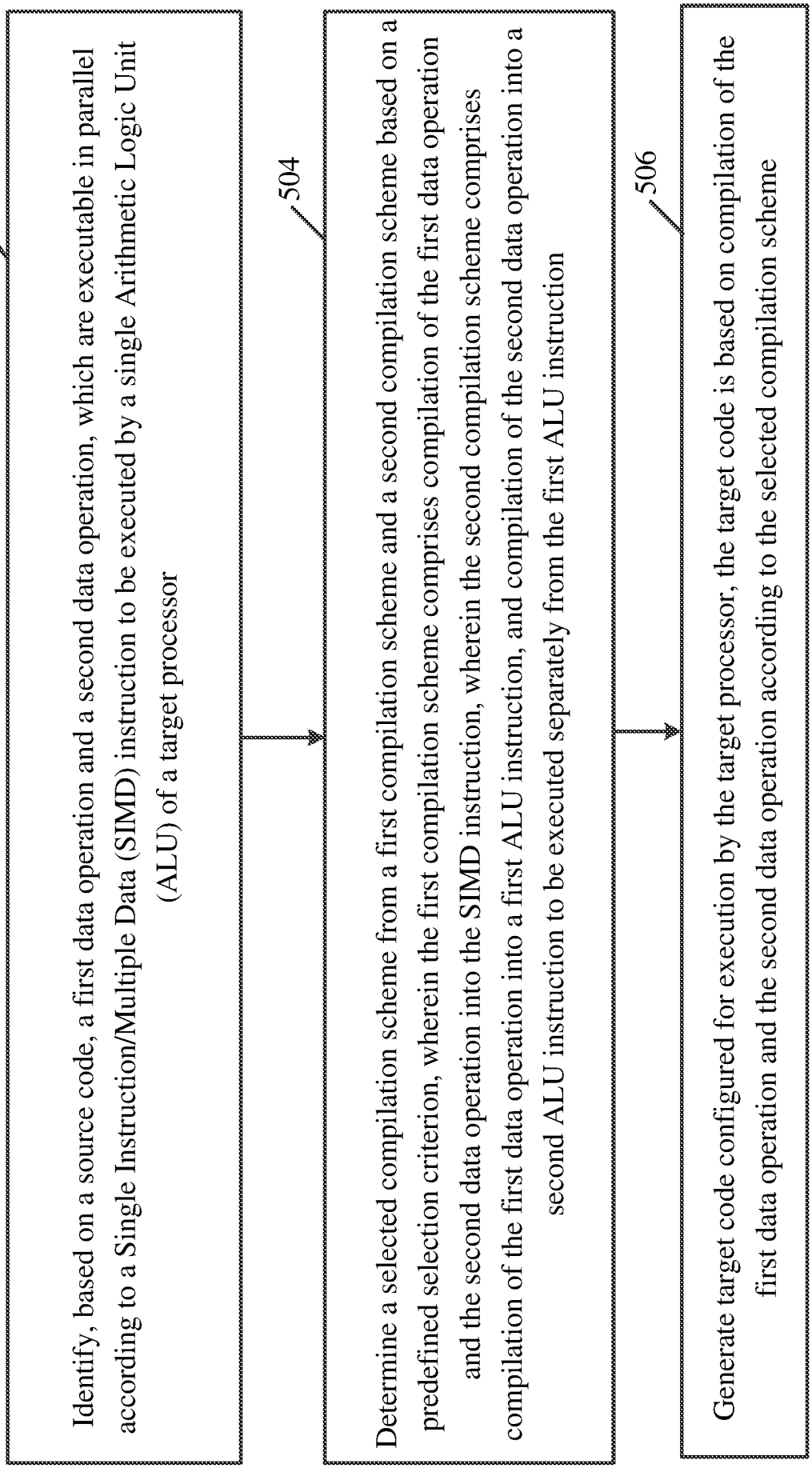
502
Identify, based on a source code, a first data operation and a second data operation, which are executable in parallel according to a Single Instruction/Multiple Data (SIMD) instruction to be executed by a single Arithmetic Logic Unit (ALU) of a target processor
504
Determine a selected compilation scheme from a first compilation scheme and a second compilation scheme based on a predefined selection criterion, wherein the first compilation scheme comprises compilation of the first data operation and the second data operation into the SIMD instruction, wherein the second compilation scheme comprises compilation of the first data operation into a first ALU instruction, and compilation of the second data operation into a second ALU instruction to be executed separately from the first ALU instruction
506
Generate target code configured for execution by the target processor, the target code is based on compilation of the first data operation and the second data operation according to the selected compilation scheme

Fig. 5

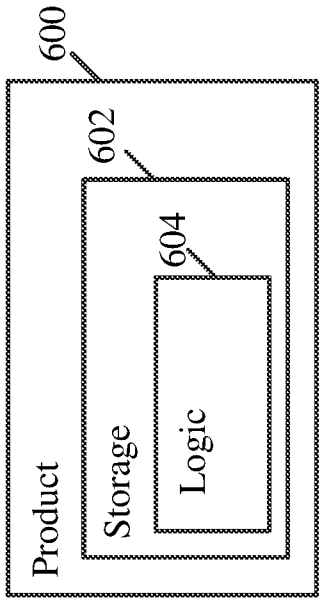

Fig. 6

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2023/060303

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F8/41
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, INSPEC, IBM-TDB, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	PORPODAS VASILEIOS VASILEIOS PORPODAS@INTEL COM ET AL: "VW-SLP auto-vectorization with adaptive vector width", 2005 43RD ACM/IEEE DESIGN AUTOMATION CONFERENCE, IEEE, PISCATAWAY, NJ, USA, 1 November 2018 (2018-11-01), pages 1-15, XP058498312, DOI: 10.1145/3243176.3243189 ISBN: 978-1-59593-381-2	1,5-7, 9-24, 26-28
Y	abstract page 1, right-hand column, paragraph 1 - page 3, right-hand column, paragraph 3 page 4, left-hand column, paragraph 2 page 9, right-hand column, paragraph 3 - page 9, right-hand column, paragraph 4	2-4,8,25

-/--

[X] Further documents are listed in the continuation of Box C.

[ ] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
9 February 2024	19/02/2024

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Lelait, Sylvain

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2023/060303

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WILLIAM JALBY ET AL: "An efficient memory operations optimization technique for vector loops on Itanium 2 processors", CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE, WILEY, LONDON, GB, vol. 18, no. 11, 17 January 2006 (2006-01-17), pages 1485-1508, XP072307510, ISSN: 1532-0626, DOI: 10.1002/CPE.1017 abstract page 1487, paragraph 1 - page 1488, paragraph 3 page 1492, paragraph 3 - page 1494, paragraph 2 page 1496, paragraph 3 - page 1499, paragraph 2	2-4,8,25

1