

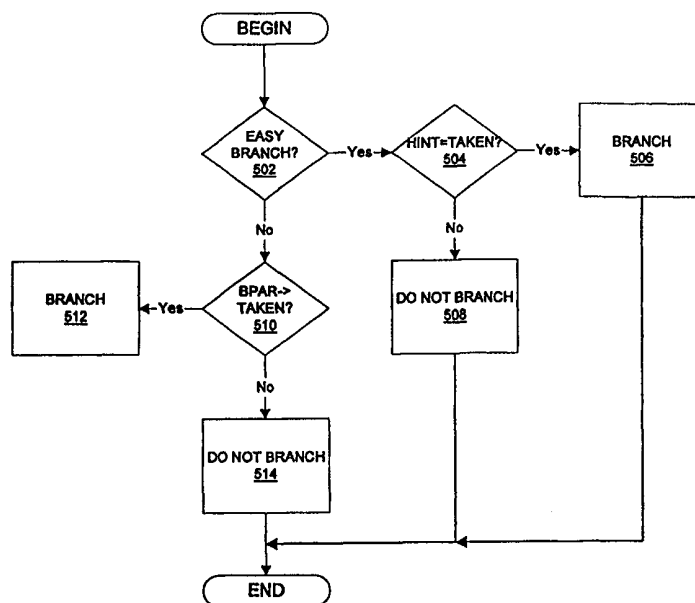


INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<p>(51) International Patent Classification ⁷ : G06F 9/32</p>	<p>A2</p>	<p>(11) International Publication Number: WO 00/33182 (43) International Publication Date: 8 June 2000 (08.06.00)</p>
---	------------------	--

<p>(21) International Application Number: PCT/US99/28876 (22) International Filing Date: 3 December 1999 (03.12.99) (30) Priority Data: 09/204,792 3 December 1998 (03.12.98) US (71) Applicant: SUN MICROSYSTEMS, INC. [US/US]; 901 San Antonio Road, M/S PAL01-521, Palo Alto, CA 94303 (US). (72) Inventor: TREMBLAY, Marc; 140 Hanna Way, Menlo Park, CA 94025 (US). (74) Agents: TERRILE, Stephen, A. et al.; Skjerven, Morrill, MacPherson, Franklin & Friel LLP, Suite 700, 25 Metro Drive, San Jose, CA 95110 (US).</p>	<p>(81) Designated States: JP, KR, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). Published <i>Without international search report and to be republished upon receipt of that report.</i></p>
---	--

(54) Title: SOFTWARE BRANCH PREDICTION FILTERING FOR A MICROPROCESSOR



(57) Abstract

The present invention provides software branch prediction filtering for a microprocessor. In one embodiment, a method for a software branch prediction filtering for a microprocessor includes determining whether a branch is "easy" to predict, and predicting the branch using software branch prediction if the branch is easy to predict. Otherwise (i.e., the branch is "hard" to predict), the branch is predicted using hardware branch prediction. Accordingly, more accurate but space-limited hardware branch prediction resources are conserved for hard-to-predict branches.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

SOFTWARE BRANCH PREDICTION FILTERING FOR A MICROPROCESSOR

TECHNICAL FIELD

The present invention relates generally to microprocessors, and more particularly, to branch prediction for a microprocessor.

5 BACKGROUND ART

Reduced Instruction Set Computing (RISC) microprocessors are well known. RISC microprocessors are characterized by a smaller number of instructions, which are relatively simple to decode, and by having all arithmetic/logic operations be performed register-to-register. RISC instructions are generally of only one length (e.g., 32-bit instructions). RISC instruction execution is of the direct hardwired type, as opposed to microcoding.

10 There is a fixed instruction cycle time, and the instructions are defined to be relatively simple so that each instruction generally executes in one relatively short cycle.

A RISC microprocessor typically includes an instruction for a conditional branch operation. I.e., if a certain condition is present, then branch to a given location. It is known that a relatively small number of branch operations cause most of the branch mispredictions. For example, it has been suggested that 80 percent of the

15 branch mispredictions result from 20 percent of the branch instructions for a given processor. Other branch operations are relatively easy to predict. For example, if an array access is preceded by a check for a valid array access, the check for a valid array access is accomplished in a typical RISC microprocessor by executing multiple conditional branches. These branches are generally easy to predict.

DISCLOSURE OF INVENTION

20 Speed of execution is highly dependent on the sequentiality of the instruction stream executed by the microprocessor. Branches in the instruction stream disrupt the sequentiality of the instruction stream executed by the microprocessor and generate stalls while the prefetched instruction stream is flushed and a new instruction stream begun.

Accordingly, the present invention provides software branch prediction filtering for a microprocessor.

25 For example, the present invention provides a cost-effective and high performance implementation of software branch prediction filtering executed on a microprocessor that performs branch operations. By providing the software branch prediction filtering, many easy-to-predict branches can be eliminated from a hardware-implemented branch prediction table thereby freeing up space in the branch prediction table that would otherwise be occupied by the easy-to-predict branches. In other words, easy-to-predict branches waste entries in a limited-

- 2 -

size branch prediction table and, thus, are eliminated from the branch prediction table. This robust approach to software branch prediction filtering provides for improved branch prediction, which is desired in various environments, such as a Java™ computing environment. For example, this method can be used for various instruction sets such as Sun Microsystems, Inc.'s UltraJava™ instruction set.

5 In one embodiment, a method for software branch prediction filtering for a microprocessor includes determining whether a conditional branch operation is “easy”-to-predict and predicting whether to execute the branch operation based on software branch prediction. However, “hard”-to-predict branches are predicted using a hardware branch prediction (e.g., a limited size hardware branch prediction table).

10 Other aspects and advantages of the present invention will become apparent from the following detailed description and accompanying drawings.

BRIEF DESCRIPTION OF DRAWINGS

Figure 1 is a schematic block diagram illustrating a single integrated circuit chip implementation of a processor in accordance with an embodiment of the present invention.

Figure 2 is a schematic block diagram showing the core of the processor.

15 Figure 3 shows a format of a branch instruction in accordance with one embodiment of the present invention.

Figure 4 is a block diagram of an implementation of the branch instruction of Figure 3 in accordance with one embodiment of the present invention.

20 Figure 5 is a flow diagram of the operation of the branch instruction of Figure 3 in accordance with one embodiment of the present invention.

MODES FOR CARRYING OUT THE INVENTION

The present invention provides software branch prediction filtering for branch operations for a microprocessor. In one embodiment, software branch prediction filtering uses hardware branch prediction only for “hard”-to-predict branches (a branch in which historical operation of the branch taken is important in
25 determining whether the branch will be taken this time, e.g., an if..then statement) and uses software branch prediction for “easy” to prediction branches (a branch in which the history is not important in determining whether the branch will be taken for this particular branch, e.g., a loop). For example, the branch instruction can be used in a computing environment in which compiled programs include a significant number of branch

operations, such as in a Java™ computing environment or in a computing environment that is executing compiled “C” programs.

For example, branch mispredictions generally slow down Java™ code executing on a typical microprocessor, which is due to the time wasted fetching the branched to instruction(s). Even with advanced
5 compiler optimizations, it is difficult to eliminate all such branch mispredictions. Well-known Just-In-Time (JIT) Java™ compilers that generate software branch predictions for a typical Reduced Instruction Set Computing (RISC) microprocessor are currently about 75% accurate. Current hardware branch prediction is more accurate at about 85-93% accurate. Hardware branch prediction is typically implemented using a hardware branch
10 prediction table. Because the hardware branch prediction table is limited in size (e.g., 512 entries), this approach is not desirable if there are a significant number of branches (e.g., more than 1000 branches) that can lead to aliasing effects (e.g., two different branches sharing the same entries will corrupt each others prediction state).

The present invention solves this problem by providing a branch instruction that includes a bit for indicating whether the branch is easy to predict or hard to predict in accordance with one embodiment. If the branch is hard to predict, then hardware branch prediction is used. Otherwise, software branch prediction is used.
15 Thus, the more accurate hardware branch prediction is efficiently reserved for hard-to-predict branches. For example, a compiler can determine whether a branch is labeled as hard to predict or easy to predict (e.g., about 80% of the branches can be labeled easy to predict, and mechanisms may be added to update or modify these predictions based on mispredictions, as further discussed below).

Referring to Figure 1, a schematic block diagram illustrates a single integrated circuit chip
20 implementation of a processor 100 that includes a memory interface 102, a geometry decompressor 104, two media processing units 110 and 112, a shared data cache 106, and several interface controllers. The interface controllers support an interactive graphics environment with real-time constraints by integrating fundamental components of memory, graphics, and input/output bridge functionality on a single die. The components are mutually linked and closely linked to the processor core with high bandwidth, low-latency communication
25 channels to manage multiple high-bandwidth data streams efficiently and with a low response time. The interface controllers include a an UltraPort Architecture Interconnect (UPA) controller 116 and a peripheral component interconnect (PCI) controller 120. The illustrative memory interface 102 is a direct Rambus dynamic RAM (DRDRAM) controller. The shared data cache 106 is a dual-ported storage that is shared among the media processing units 110 and 112 with one port allocated to each media processing unit. The data cache 106 is four-
30 way set associative, follows a write-back protocol, and supports hits in the fill buffer (not shown). The data cache 106 allows fast data sharing and eliminates the need for a complex, error-prone cache coherency protocol between the media processing units 110 and 112.

The UPA controller 116 is a custom interface that attains a suitable balance between high-performance computational and graphic subsystems. The UPA is a cache-coherent, processor-memory interconnect. The

- 4 -

UPA attains several advantageous characteristics including a scaleable bandwidth through support of multiple bus-based interconnects for data and addresses, packets that are switched for improved bus utilization, higher bandwidth, and precise interrupt processing. The UPA performs low latency memory accesses with high throughput paths to memory. The UPA includes a buffered cross-bar memory interface for increased bandwidth and improved scaleability. The UPA supports high-performance graphics with two-cycle single-word writes on the 64-bit UPA interconnect. The UPA interconnect architecture utilizes point-to-point packet switched messages from a centralized system controller to maintain cache coherence. Packet switching improves bus bandwidth utilization by removing the latencies commonly associated with transaction-based designs.

The PCI controller 120 is used as the primary system I/O interface for connecting standard, high-volume, low-cost peripheral devices, although other standard interfaces may also be used. The PCI bus effectively transfers data among high bandwidth peripherals and low bandwidth peripherals, such as CD-ROM players, DVD players, and digital cameras.

Two media processing units 110 and 112 are included in a single integrated circuit chip to support an execution environment exploiting thread level parallelism in which two independent threads can execute simultaneously. The threads may arise from any sources such as the same application, different applications, the operating system, or the runtime environment. Parallelism is exploited at the thread level since parallelism is rare beyond four, or even two, instructions per cycle in general purpose code. For example, the illustrative processor 100 is an eight-wide machine with eight execution units for executing instructions. A typical "general-purpose" processing code has an instruction level parallelism of about two so that, on average, most (about six) of the eight execution units would be idle at any time. The illustrative processor 100 employs thread level parallelism and operates on two independent threads, possibly attaining twice the performance of a processor having the same resources and clock rate but utilizing traditional non-thread parallelism.

Thread level parallelism is particularly useful for Java™ applications which are bound to have multiple threads of execution. Java™ methods including "suspend", "resume", "sleep", and the like include effective support for threaded program code. In addition, Java™ class libraries are thread-safe to promote parallelism. Furthermore, the thread model of the processor 100 supports a dynamic compiler which runs as a separate thread using one media processing unit 110 while the second media processing unit 112 is used by the current application. In the illustrative system, the compiler applies optimizations based on "on-the-fly" profile feedback information while dynamically modifying the executing code to improve execution on each subsequent run. For example, a "garbage collector" may be executed on a first media processing unit 110, copying objects or gathering pointer information, while the application is executing on the other media processing unit 112.

Although the processor 100 shown in Figure 1 includes two processing units on an integrated circuit chip, the architecture is highly scaleable so that one to several closely-coupled processors may be formed in a message-based coherent architecture and resident on the same die to process multiple threads of execution. Thus,

in the processor 100, a limitation on the number of processors formed on a single die thus arises from capacity constraints of integrated circuit technology rather than from architectural constraints relating to the interactions and interconnections between processors.

Referring to Figure 2, a schematic block diagram shows the core of the processor 100. The media processing units 110 and 112 each include an instruction cache 210, an instruction aligner 212, an instruction buffer 214, a pipeline control unit 226, a split register file 216, a plurality of execution units, and a load/store unit 218. In the illustrative processor 100, the media processing units 110 and 112 use a plurality of execution units for executing instructions. The execution units for a media processing unit 110 include three media functional units (MFU) 222 and one general functional unit (GFU) 220. The media functional units 222 are multiple single-instruction-multiple-datapath (MSIMD) media functional units. Each of the media functional units 222 is capable of processing parallel 16-bit components. Various parallel 16-bit operations supply the single-instruction-multiple-datapath capability for the processor 100 including add, multiply-add, shift, compare, and the like. The media functional units 222 operate in combination as tightly-coupled digital signal processors (DSPs). Each media functional unit 222 has an separate and individual sub-instruction stream, but all three media functional units 222 execute synchronously so that the subinstructions progress lock-step through pipeline stages.

The general functional unit 220 is a RISC processor capable of executing arithmetic logic unit (ALU) operations, loads and stores, branches, and various specialized and esoteric functions such as parallel power operations, reciprocal squareroot operations, and many others. The general functional unit 220 supports less common parallel operations such as the parallel reciprocal square root instruction.

The illustrative instruction cache 210 has a 16 Kbyte capacity and includes hardware support to maintain coherence, allowing dynamic optimizations through self-modifying code. Software is used to indicate that the instruction storage is being modified when modifications occur. The 16K capacity is suitable for performing graphic loops, other multimedia tasks or processes, and general-purpose Java™ code. Coherency is maintained by hardware that supports write-through, non-allocating caching. Self-modifying code is supported through explicit use of "store-to-instruction-space" instructions *store2i*. Software uses the *store2i* instruction to maintain coherency with the instruction cache 210 so that the instruction caches 210 do not have to be snooped on every single store operation issued by the media processing unit 110.

The pipeline control unit 226 is connected between the instruction buffer 214 and the functional units and schedules the transfer of instructions to the functional units. The pipeline control unit 226 also receives status signals from the functional units and the load/store unit 218 and uses the status signals to perform several control functions. The pipeline control unit 226 maintains a scoreboard, generates stalls and bypass controls. The pipeline control unit 226 also generates traps and maintains special registers.

Each media processing unit 110 and 112 includes a split register file 216, a single logical register file including 128 thirty-two bit registers. The split register file 216 is split into a plurality of register file segments 224 to form a multi-ported structure that is replicated to reduce the integrated circuit die area and to reduce access time. A separate register file segment 224 is allocated to each of the media functional units 222 and the general functional unit 220. In the illustrative embodiment, each register file segment 224 has 128 32-bit registers. The first 96 registers (0-95) in the register file segment 224 are global registers. All functional units can write to the 96 global registers. The global registers are coherent across all functional units (MFU and GFU) so that any write operation to a global register by any functional unit is broadcast to all register file segments 224. Registers 96-127 in the register file segments 224 are local registers. Local registers allocated to a functional unit are not accessible or "visible" to other functional units.

The media processing units 110 and 112 are highly structured computation blocks that execute software-scheduled data computation operations with fixed, deterministic and relatively short instruction latencies, operational characteristics yielding simplification in both function and cycle time. The operational characteristics support multiple instruction issue through a pragmatic very large instruction word (VLIW) approach that avoids hardware interlocks to account for software that does not schedule operations properly. Such hardware interlocks are typically complex, error-prone, and create multiple critical paths. A VLIW instruction word always includes one instruction that executes in the general functional unit (GFU) 220 and from zero to three instructions that execute in the media functional units (MFU) 222. A MFU instruction field within the VLIW instruction word includes an operation code (opcode) field, three source register (or immediate) fields, and one destination register field.

Instructions are executed in-order in the processor 100 but loads can finish out-of-order with respect to other instructions and with respect to other loads, allowing loads to be moved up in the instruction stream so that data can be streamed from main memory. The execution model eliminates the usage and overhead resources of an instruction window, reservation stations, a re-order buffer, or other blocks for handling instruction ordering. Elimination of the instruction ordering structures and overhead resources is highly advantageous since the eliminated blocks typically consume a large portion of an integrated circuit die. For example, the eliminated blocks consume about 30% of the die area of a Pentium II processor.

To avoid software scheduling errors, the media processing units 110 and 112 are high-performance but simplified with respect to both compilation and execution. The media processing units 110 and 112 are most generally classified as a simple 2-scalar execution engine with full bypassing and hardware interlocks on load operations. The instructions include loads, stores, arithmetic and logic (ALU) instructions, and branch instructions so that scheduling for the processor 100 is essentially equivalent to scheduling for a simple 2-scalar execution engine for each of the two media processing units 110 and 112.

The processor 100 supports full bypasses between the first two execution units within the media processing unit 110 and 112 and has a scoreboard in the general functional unit 220 for load operations so that the compiler does not need to handle nondeterministic latencies due to cache misses. The processor 100 scoreboards long latency operations that are executed in the general functional unit 220, for example a reciprocal square-root operation, to simplify scheduling across execution units. The scoreboard (not shown) operates by tracking a record of an instruction packet or group from the time the instruction enters a functional unit until the instruction is finished and the result becomes available. A VLIW instruction packet contains one GFU instruction and from zero to three MFU instructions. The source and destination registers of all instructions in an incoming VLIW instruction packet are checked against the scoreboard. Any true dependencies or output dependencies stall the entire packet until the result is ready. Use of a scoreboarded result as an operand causes instruction issue to stall for a sufficient number of cycles to allow the result to become available. If the referencing instruction that provokes the stall executes on the general functional unit 220 or the first media functional unit 222, then the stall only endures until the result is available for intra-unit bypass. For the case of a *load* instruction that hits in the data cache 106, the stall may last only one cycle. If the referencing instruction is on the second or third media functional units 222, then the stall endures until the result reaches the writeback stage in the pipeline where the result is bypassed in transmission to the split register file 216.

The scoreboard automatically manages load delays that occur during a load hit. In an illustrative embodiment, all loads enter the scoreboard to simplify software scheduling and eliminate NOPs in the instruction stream.

The scoreboard is used to manage most interlocks between the general functional unit 220 and the media functional units 222. All loads and non-pipelined long-latency operations of the general functional unit 220 are scoreboarded. The long-latency operations include division *idiv*, *fdiv* instructions, reciprocal squareroot *frecsqrt*, *precsqrt* instructions, and power *ppower* instructions. None of the results of the media functional units 222 is scoreboarded. Non-scoreboarded results are available to subsequent operations on the functional unit that produces the results following the latency of the instruction.

The illustrative processor 100 has a rendering rate of over fifty million triangles per second without accounting for operating system overhead. Therefore, data feeding specifications of the processor 100 are far beyond the capabilities of cost-effective memory systems. Sufficient data bandwidth is achieved by rendering of compressed geometry using the geometry decompressor 104, an on-chip real-time geometry decompression engine. Data geometry is stored in main memory in a compressed format. At render time, the data geometry is fetched and decompressed in real-time on the integrated circuit of the processor 100. The geometry decompressor 104 advantageously saves memory space and memory transfer bandwidth. The compressed geometry uses an optimized generalized mesh structure that explicitly calls out most shared vertices between triangles, allowing the processor 100 to transform and light most vertices only once. In a typical compressed

- 8 -

mesh, the triangle throughput of the transform-and-light stage is increased by a factor of four or more over the throughput for isolated triangles. For example, during processing of triangles, multiple vertices are operated upon in parallel so that the utilization rate of resources is high, achieving effective spatial software pipelining. Thus operations are overlapped in time by operating on several vertices simultaneously, rather than overlapping several loop iterations in time. For other types of applications with high instruction level parallelism, high trip count loops are software-pipelined so that most media functional units 222 are fully utilized.

Figure 3 shows a format of a branch instruction in accordance with one embodiment of the present invention. The branch instruction 300 includes a bit 302 for indicating that the branch is easy to predict (e.g., 0) or hard to predict (e.g., 1). The branch instruction includes a bit 304 for indicating a software branch prediction that the branch is taken (e.g., 0) or not taken (e.g., 1). The software branch prediction loaded in bit 304 is used to predict the outcome of the branch if the branch is easy to predict (e.g., bit 302 is set to 0). Branch instruction 300 also includes opcode 306, which corresponds to the opcode for a branch instruction, destination portion 308, which sets forth the destination register (e.g., where the condition resides), and relative offset portion 310, which sets forth the relative offset of the branch target when the branch is taken.

Accordingly, software branch prediction filtering migrates some of the complexity associated with conditional branches to the compiler. It is observed that, for example: graphics code has few branches, or very predictable branches; JAVA applications have more unconditional branches than typical C or Fortran applications (mainly due to the extensive usage of jumps or calls); a dynamic compiler has better observability and has the capability to update software-controlled prediction bits; software branch prediction with simple heuristics can predict branches successfully > 75% of the time, or possibly even > 83% for brute force heuristics. See, e.g., Thomas Ball, James Larus, *Branch Prediction for Free, Programming Languages Design & Implementation*, 1993, New Mexico, pp300 - 312.

Based on these observations, branch instructions have 2 bits that the compiler can set to let the processor know (a) if the branch is easy or hard to predict, and (b) the branch is predicted taken, which is a software branch prediction (e.g., determined by the compiler at compile time). In this way, when the microprocessor encounters an easy-to-predict branch, it simply uses the software branch prediction provided by the other bit. On the other hand, when the microprocessor encounters a hard-to-predict branch, it can use a simple hardware-based branch prediction or a more robust hardware-based branch prediction. In this way it is possible to dedicate a hardware-based branch prediction mechanism only to those branches that the software cannot predict very well. Measurements show that a reduction of the number of mispredictions between 20 - 40 percent is achievable. Alternately, the prediction efficiency can be kept at the same level, while the size of the branch prediction table can be reduced.

Figure 4 is a block diagram of an implementation of the branch instruction of Figure 4 in accordance with one embodiment of the present invention. MPU 400 includes an instruction fetch unit 402, which fetches

- 9 -

instruction data from an instruction cache unit (see Figure 1). Instruction fetch unit 402 is coupled to a branch prediction circuit 404. Branch prediction circuit 404 includes a branch prediction table 406, such as a conventional 512-entry branch prediction table. Instruction fetch unit 402 is also coupled to a decoder 408, which decodes an instruction for execution by execution unit 410. One of ordinary skill in the art will recognize that there are various way to implement the circuitry and logic for performing the branch prediction operation in a microprocessor, such as a pipelined microprocessor.

Figure 5 is a flow diagram of the operation of the branch instruction of Figure 5 in accordance with one embodiment of the present invention. The operation of the branch instruction begins at stage of operation 502. At stage 502, whether the branch is easy to predict is determined. If so, then software branch prediction is used to predict whether the branch is taken. At stage 504, whether the software branch prediction predicts that the branch is taken is determined. If so, then the branch is taken at stage 506. Otherwise, the branch is not taken at stage 508.

Otherwise (i.e., the branch is hard to predict), a hardware branch prediction mechanism (e.g., the branch prediction circuit of Figure 5) is used to determine if the branch is predicted to be taken. If the branch is predicted take by the hardware branch prediction circuit (e.g., branch prediction array (bpar)), then the branch is taken at stage 512 (e.g., the offset is added to the current program counter to provide a new address sequence to be fetched). Otherwise, the branch is not taken at stage 514 (e.g., the present instruction stream is continued in sequence).

In one embodiment, a branch misprediction by the software branch prediction causes a modification of the software branch prediction bit (e.g., toggles bit 504 of Figure 5 using self-modifying code). A hardware branch misprediction causes a modification in the hardware branch prediction table (e.g., an entry in branch prediction table 506 of Figure 5 is modified).

In one embodiment, the software branch prediction utilizes heuristics involving code analysis such as that set forth in Ball et al.

In one embodiment, the hardware branch prediction utilizes the following branch prediction scheme:

if offset < 0 (backward branch) then predict taken

else (i.e., offset > 0) (forward branch) predict not taken.

The displacement of an unconditional branch is treated as an offset and added to the program counter (not shown) to form the target address of the next instruction if the branch is taken. Alternatively, a more robust hardware branch prediction approach utilizes a branch prediction table (e.g., 512-entry branch prediction table) and associates a state machine to each branch. For example, a 2-bit counter is used to describe four states:

- 10 -

strongly taken, likely taken, likely not taken, and strongly not taken. The branch prediction table is implemented as a branch prediction array.

5 In one embodiment, a JIT compiler for JAVA™ source code provides software branch prediction (e.g., sets bit 504) and indicates whether a compiled branch is easy to predict or hard to predict (e.g., sets bit 502). The software branch prediction filtering can reduce misprediction rates by about 25% and considering that about 20% of compiled JAVA™ code can be branches, this embodiment provides a significant improvement. The present invention can also be applied to statically compiled C code or to static compilation of other computer programming languages. Also, this approach reduces the risk of polluting the hardware branch prediction table by conserving the hardware branch prediction table for hard-to-predict branches.

10 Although particular embodiments of the present invention have been shown and described, it will be obvious to those skilled in the art that changes and modifications can be made without departing from the present invention in its broader aspects. For example, different approaches to software branch prediction and to hardware branch prediction can be used. Also, dynamic software branch prediction or dynamic hardware branch prediction (or both) can be utilized in accordance with one embodiment of the present invention. The present invention is not limited by any particular processor architecture, the presence or structure of caches or memory, 15 or the number of bits in any register or memory location. Therefore, the appended claims are to encompass within their scope all such changes and modifications that fall within the true scope of the present invention.

WE CLAIM

- 1 1. A process for software branch prediction filtering for a microprocessor, comprising:
2 determining whether the branch is easy to predict; and
3 predicting the branch using software branch prediction if the branch is easy to predict.

- 1 2. The process of Claim 1 further comprising:
2 predicting the branch using hardware branch prediction if the branch is hard to predict.

- 1 3. The process of Claim 2 further comprising:
2 checking a first bit of an instruction that indicates whether the branch is easy to predict or hard to
3 predict.

- 1 4. The process of Claim 3 further comprising:
2 checking a second bit of the instruction that indicates whether the branch is predicted taken or not taken
3 by the software branch prediction.

- 1 5. The process of Claim 4 further comprising:
2 modifying the second bit if the software branch prediction mispredicts the branch.

- 1 6. The process of Claim 2 further comprising:
2 modifying a branch prediction table if the hardware branch prediction mispredicts the branch.

- 1 7. The process of Claim 6 wherein the hardware branch prediction comprises incrementing and
2 decrementing a counter based on a state machine.

- 1 8. The process of Claim 6 wherein the software branch prediction comprises utilizing heuristics.

- 1 9. The process of Claim 3 wherein the first bit is set by a compiler that compiled the instruction.

- 1 10. The process of Claim 9 wherein the compiler is a Java™ Just-In-Time compiler.

- 1 11. An apparatus for software branch prediction filtering for a microprocessor, comprising:
2 branch prediction circuitry, the branch prediction circuitry comprising a branch prediction table; and

3 software branch prediction filtering logic coupled to the branch prediction circuitry, the software branch
4 prediction filtering logic executing a branch instruction and determining whether the branch is
5 easy to predict, and the software branch prediction filtering logic predicting the branch using
6 the software branch prediction if the branch is easy to predict.

1 12. The apparatus of Claim 11, wherein the software branch prediction filtering logic further
2 comprises predicting the branch using the hardware branch prediction circuitry if the branch is hard to predict.

1 13. The apparatus of Claim 12 wherein a first bit of the branch instruction provides an indication of
2 whether the branch is easy to predict, and a second bit provides an indication of the software branch prediction.

1 14. The apparatus of Claim 13 wherein the software branch prediction filtering logic further
2 comprises modifying the second bit if the software branch prediction mispredicts the branch.

1 15. The apparatus of Claim 14 wherein the hardware branch prediction circuitry comprises a 512-
2 entry branch prediction table.

1 16. The apparatus of Claim 15 wherein the branch instruction comprises a compiled Java™
2 instruction.

FIG. 1

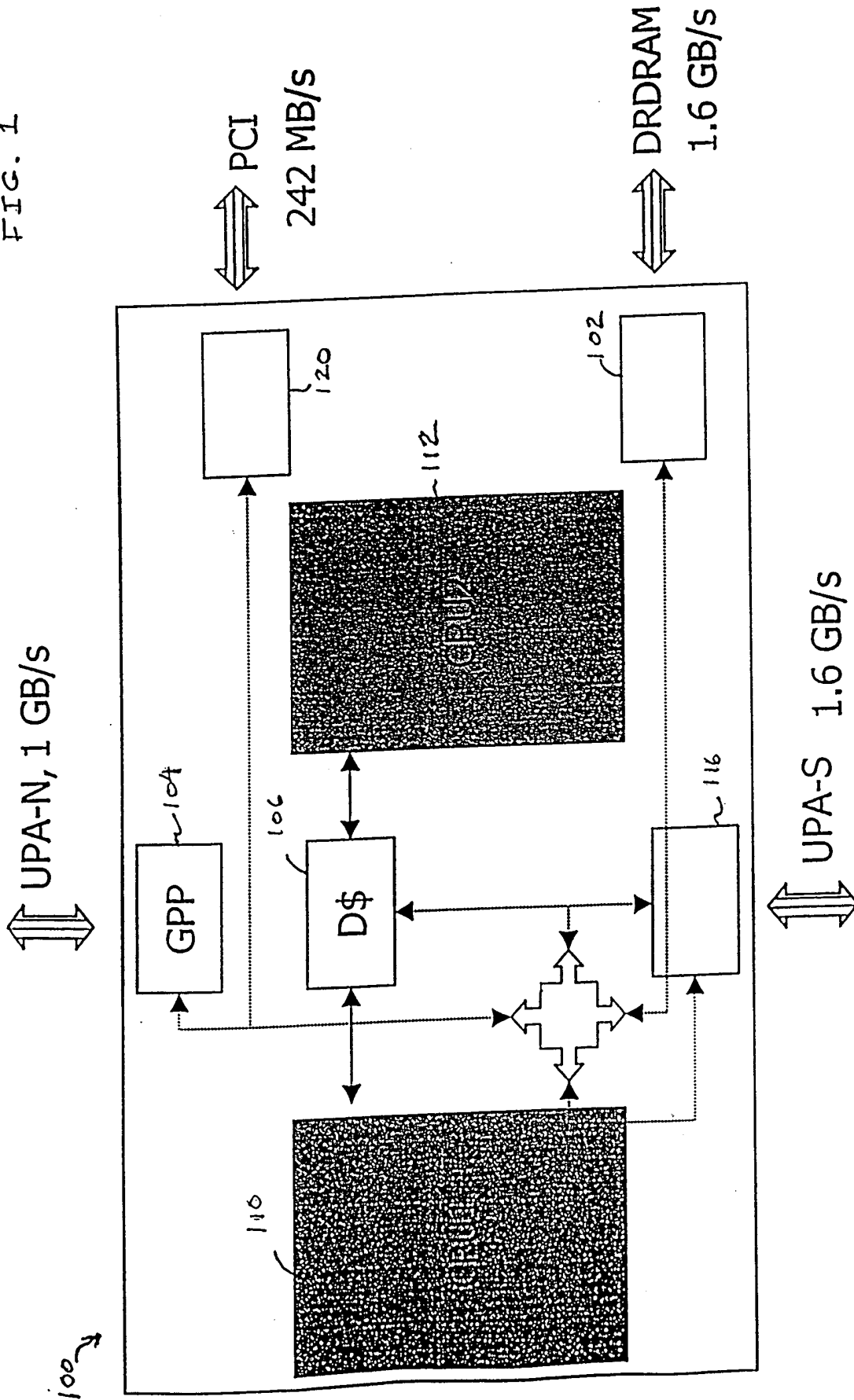
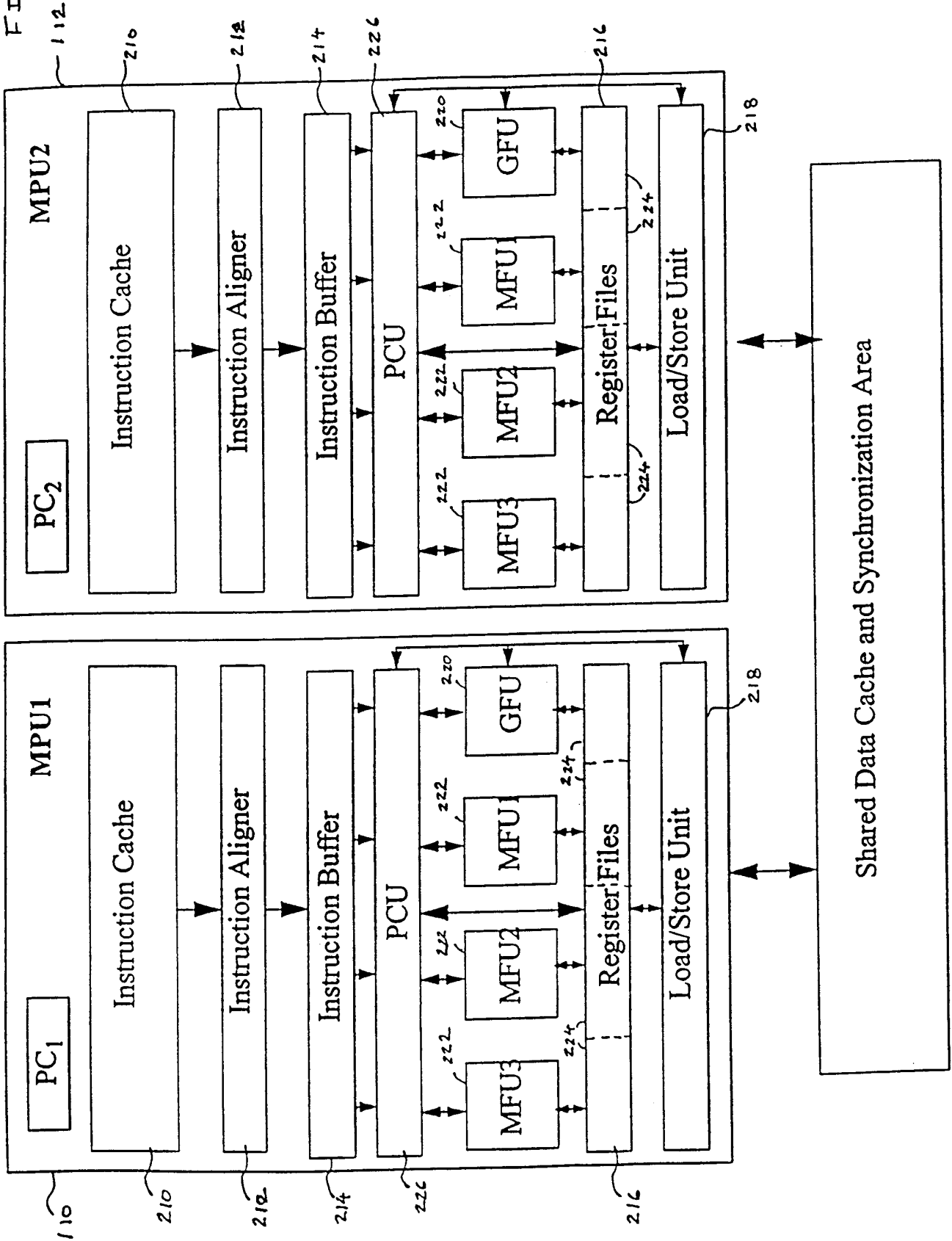


FIG. 2



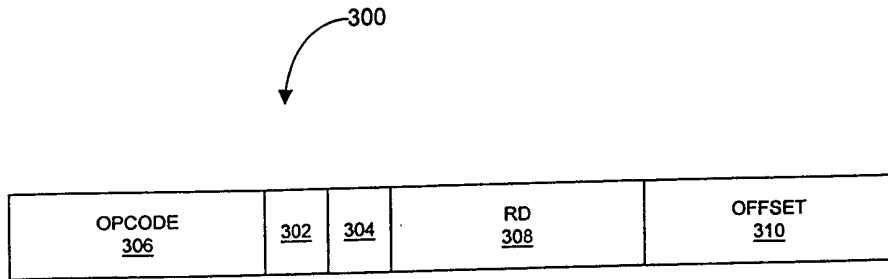


Figure 3

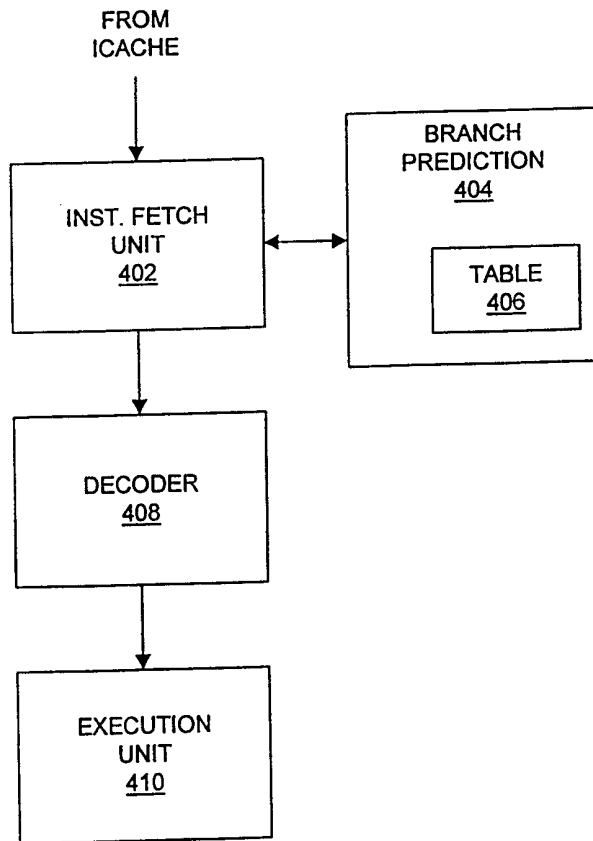


Figure 4

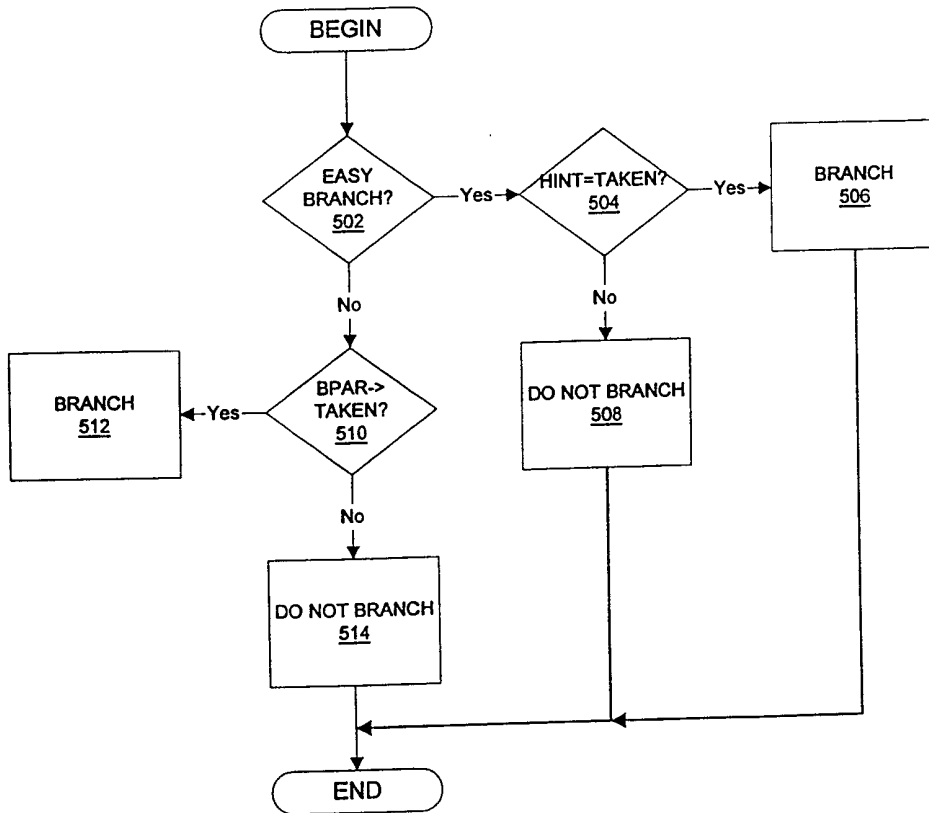


Figure 5