

# Glossary

- absolute address** A variable's or routine's actual address in memory.
- abstraction** A model that renders lower-level details of computer systems temporarily invisible in order to facilitate design of sophisticated systems.
- activation record** *See* procedure frame.
- address** A value used to delineate the location of a specific data element within a memory array.
- address mapping** *See* address translation.
- address translation** Also called address mapping. The process by which a virtual address is mapped to an address used to access memory.
- addressing mode** One of several addressing regimes delimited by their varied use of operands and/or addresses.
- aliasing** A situation in which the same object is accessed by two addresses; can occur in virtual memory when there are two virtual addresses for the same physical page.
- ALU** *See* arithmetic logic unit (ALU).
- Amdahl's law** A rule stating that the performance enhancement possible with a given improvement is limited by the amount that the improved feature is used.
- AND gate** Hardware that performs the AND operation on input signals yielding a single signal result.
- AND operation** An operation that leaves a 1 in the result only if both bits of the operands are 1.
- architecture** *See* instruction set architecture.
- arithmetic logic unit (ALU)** Hardware that performs arithmetic and logical operations.
- arithmetic mean** The average of the execution times that is directly proportional to total execution time.
- assembler** A program that translates a symbolic version of an instruction into the binary version.
- assembler directive** An operation that tells the assembler how to translate a program but does not produce machine instructions; always begins with a period.
- assembly language** A symbolic language that can be translated into binary.
- asserted signal** A signal that is (logically) true, or 1.

**asynchronous bus** A bus that uses a handshaking protocol for coordinating usage rather than a clock; can accommodate a wide variety of devices of differing speeds.

**atomic operation** An operation in which the processor can both read a location and write it in the same bus operation, preventing any other processor or I/O device from reading or writing memory until it completes.

**backpatching** A method for translating from assembly language to machine instructions in which the assembler builds a (possibly incomplete) binary representation of every instruction in one pass over a program and then returns to fill in previously undefined labels.

**backplane bus** A bus that is designed to allow processors, memory, and I/O devices to coexist on a single bus.

**barrier synchronization** A synchronization scheme in which processors wait at the barrier and do not proceed until every processor has reached it.

**base addressing** Also called displacement addressing. An addressing regime in which the operand is at the memory location whose address is the sum of a register and an address in the instruction.

**basic block** A sequence of instructions without branches (except possibly at the end) and without branch targets or branch labels (except possibly at the beginning).

**biased notation** A notation that represents the most negative value by  $00 \dots 000_{\text{two}}$  and the most positive value by  $11 \dots 11_{\text{two}}$ , with 0 typically having the value  $10 \dots 00_{\text{two}}$ , thereby biasing the number such that the number plus the bias has a nonnegative representation.

**binary bit** *See* binary digit.

**binary digit** Also called binary bit. One of the two numbers in base 2, 0 or 1, that are the components of information.

**block** The minimum unit of information that can be either present or not present in the two-level hierarchy.

- Booth's algorithm** An algorithm based on the observation that the ability to both add and subtract allows for multiple ways to compute a product, so that by looking at multiple bits we potentially save arithmetic operations.
- branch delay slot** The slot directly after a delayed branch instruction, which in the MIPS architecture is filled by an instruction that does not affect the branch.
- branch hazard** Also called control hazard. An occurrence in which the proper instruction cannot execute in the proper clock cycle because the instruction that was fetched is not the one that is needed; that is, the flow of instruction addresses is not what the pipeline expected.
- branch history table** *See* branch prediction buffer.
- branch not taken** A branch where the branch condition is false and the program counter (PC) becomes the address of the instruction that sequentially follows the branch.
- branch prediction** A method of resolving a branch hazard that assumes a given outcome for the branch and proceeds from that assumption rather than waiting to ascertain the actual outcome.
- branch prediction buffer** Also called branch history table. A small memory that is indexed by the lower portion of the address of the branch instruction and that contains one or more bits indicating whether the branch was recently taken or not.
- branch taken** A branch where the branch condition is satisfied and the program counter (PC) becomes the branch target. All unconditional branches are taken branches.
- branch target address** The address specified in a branch, which becomes the new program counter (PC) if the branch is taken. In the MIPS architecture the branch target is given by the sum of the offset field of the instruction and the address of the instruction following the branch.
- bubble** *See* pipeline stall.
- bus** In logic design, a collection of data lines that is treated together as a single logical signal; also, a shared collection of lines with multiple sources and uses.
- bus arbitration** The process of deciding which bus master gets to use a bus next.
- bus master** A unit on the bus that can initiate bus requests.
- bus request** A signal on the bus requesting access to a bus.

**bus transaction** A sequence of bus operations that includes a request and may include a response, either of which may carry data. A transaction is initiated by a single request and may take many individual bus operations.

**bypassing** *See* forwarding.

**cache coherency** Consistency in the value of data between the versions in the caches of several processors.

**cache memory** A small, fast memory that acts as a buffer for a slower, larger memory.

**cache miss** A request for data from the cache that cannot be filled because the data is not present in the cache.

**callee** A procedure that executes a series of stored instructions based on parameters provided by the caller and then returns control to the caller.

**callee-saved register** A register saved by the routine making a procedure call.

**caller** The program that instigates a procedure and provides the necessary parameter values.

**caller-saved register** A register saved by the routine being called.

**capacity miss** A cache miss that occurs because the cache, even with full associativity, cannot contain all the block needed to satisfy the request.

**central processor unit (CPU)** Also called processor. The active part of the computer, which contains the datapath and control and which adds numbers, tests numbers, signals I/O devices to activate, and so on.

**centralized, parallel arbitration** A bus arbitration scheme that employs multiple request lines by which the devices independently request the bus and that uses a centralized arbiter to choose from the devices requesting bus access and to notify the selected device that it is now bus master.

**chip** *See* integrated circuit.

**clock** *See* clock cycle.

**clock cycle** Also called tick, clock tick, clock period, clock, cycle. The time for one clock period, usually of the processor clock, which runs at a constant rate. The clock cycle is often used to measure the speed at which hardware can perform basic functions.

**clock cycles per instruction (CPI)** Average number of clock cycles per instruction for a program or program fragment.

**clock period** *See* clock cycle.

**clock rate** The speed of the processor or system clock measured as the number of clock cycles per second and usually stated in megahertz or millions of clock cycles per second. The clock rate is the inverse of the clock period. Designers refer to the clock cycle time both as the duration of one clock period, measured as seconds per clock cycle (e.g., 2 ns) and as the clock rate, measured as clock cycles per second (e.g., 500 MHz).

**clock skew** The difference in absolute time between the times when two state elements see a clock edge.

**clock tick** *See* clock cycle.

**clocking methodology** The approach used to determine when data is valid and stable relative to the clock.

**cluster** A set of computers connected over a local area network (LAN) that function as a single large multiprocessor.

**cold start miss** *See* compulsory miss.

**collision miss** *See* conflict miss.

**combinational logic** A logic system whose blocks do not contain memory and hence compute the same output given the same input.

**commit unit** The unit in a dynamic or out-of-order execution pipeline that decides when it is safe to release the result of an operation to programmer-visible registers and memory.

**compiler** A program that translates high-level language statements into assembly language statements.

**compulsory miss** Also called cold start miss. A cache miss caused by the first access to a block that has never been in the cache.

**computer generation** A classification of computers often based on the implementation technology used in each generation, originally lasting eight to ten years.

**conditional branch** An instruction that requires the comparison of two values and that allows for a subsequent transfer of control to a new address in the program based on the outcome of the comparison.

**conflict miss** Also called collision miss. A cache miss that occurs in a set-associative or direct-mapped cache when multiple blocks compete for the same set and that are eliminated in a fully associative cache of the same size.

**context switch** A changing of the internal state of the processor to allow a different process to use the processor that includes saving the state needed to return to the currently executing process.

**control** The component of the processor that commands the datapath, memory, and I/O devices according to the instructions of the program.

**control hazard** *See* branch hazard.

**control signal** A signal used for multiplexor selection or for directing the operation of a functional unit; contrasts with a **data signal**, which contains information that is operated on by a functional unit.

**control value** *See* selector value.

**CPI** *See* clock cycles per instruction (CPI).

**CPU** *See* central processor unit (CPU).

**CPU execution time** Also called CPU time. The actual time the CPU spends computing for a specific task.

**CPU time** *See* CPU execution time.

**crossbar network** A network that allows any node to communicate with any other node in one pass through the network.

**cycle** *See* clock cycle.

**D flip-flop** A flip-flop with one data input that stores the value of that input signal in the internal memory when the clock edge occurs.

**daisy chain arbitration** A bus arbitration scheme in which the bus grant line is run through the devices from highest priority to lowest (the priorities are determined by the position on the bus) so that when the bus is requested the highest priority device sees the bus grant signal first.

**data dependencies** The need for specific data at a given point in a pipeline.

**data hazard** Also called pipeline data hazard. An occurrence in which a planned instruction cannot execute in the proper clock cycle because data that is needed to execute the instruction is not yet available.

**data parallelism** Parallelism achieved by having massive data.

**data segment** The segment of a Unix object or executable file that contains a binary representation of the initialized data used by the program.

**data transfer instruction** A command that moves data between memory and registers.

**datapath** The component of the processor that performs arithmetic operations.

**datapath element** A functional unit used to operate on or hold data within a processor. In the MIPS implementation the datapath elements include the instruction and data **memories**, the register file, the arithmetic logic unit (ALU), and adders.

**deasserted signal** A signal that is (logically) false, or 0.

**decoder** A logic block that has an  $n$ -bit input and  $2^n$  outputs where only one output is asserted for each input combination.

**defect** A microscopic flaw in a wafer or in patterning steps that can result in the failure of the die containing that defect.

**delay** *See* rotation latency.

- delayed branch** A type of branch where the instruction immediately following the branch is always executed, independent of whether the branch condition is true or false.
- delayed load** A software format that requires load instructions to be followed by an instruction independent of the load.
- die** The individual rectangular sections that are cut from a *wafer*, more informally known as chips.
- die area** The size of a die.
- direct-mapped cache** A cache structure in which each memory location is mapped to exactly one location in the cache.
- direct memory access (DMA)** A mechanism that provides a device controller the ability to transfer data directly to or from the memory without involving the processor.
- directory** A repository for information on the state of every block in main memory, including which caches have copies of the block, whether it is dirty, and so on.
- dispatch** An operation in a microprogrammed control unit in which the next microinstruction is selected on the basis of one or more fields of a macroinstruction, usually by creating a table containing the addresses of the target microinstructions and indexing the table using a field of the macroinstruction. The dispatch tables are typically implemented in ROM or programmable logic array (PLA). The term *dispatch* is also used in dynamically scheduled processors to refer to the process of sending an instruction to a queue.
- displacement addressing** *See* base addressing.
- distributed arbitration by collision detection** A bus arbitration scheme that allows each device to independently request the bus and that uses a scheme for retrying the arbitration when multiple simultaneous requests occur.
- distributed arbitration by self-selection** A bus arbitration scheme that gives the devices requesting the bus the ability to determine which device gets the bus by having each requester detect whether it should receive the bus allocation.
- distributed memory** Physical memory that is divided into modules, with some placed near each processor in a multiprocessor.
- distributed shared memory (DSM)** A memory scheme that uses addresses to access remote data when demanded rather than retrieving the data in case it might be used.
- divisor** A number that the dividend is divided by; produces the dividend when multiplied by the quotient and added to the remainder.

**DMA** See direct memory access (DMA).

**don't-care term** An element of a logical function in which the output does not depend on the values of all the inputs. Don't-care terms may be specified in different ways.

**double precision** A floating-point value represented in two 32-bit words.

**DRAM** See dynamic random access memory (DRAM).

**DSM** See distributed shared memory (DSM).

**dynamic pipeline scheduling** A form of scheduling that goes past stalls in order to find later instructions to execute while waiting for the stalls to be resolved.

**dynamic random access memory (DRAM)** Memory that contains the instructions and data of a program while it is running, which allows faster access than accessing a magnetic disk.

**edge-triggered clocking** A clocking scheme in which all state changes occur on a clock edge.

**Ethernet** A computer network whose length is limited to about a kilometer. Originally capable of transferring up to 10 million bits per second, newer versions can run up to 100 million bits per second and even 1000 million bits per second. It treats the wire like a bus with multiple masters and uses collision detection and a back-off scheme for handling simultaneous accesses.

**exception** Also called interrupt. An unscheduled event that disrupts program execution; used to detect overflow.

**exception enable** Also called interrupt enable. A signal or action that controls whether the process responds to an exception or not; necessary for preventing the occurrence of exceptions during intervals before the processor has safely saved the state needed to restart.

**exclusive OR gate** Hardware that performs the exclusive OR operation on input signals yielding a single signal result exclusive OR operation; also, an operation that leaves a 1 in the result only if two bits of the operands are unequal.

**executable file** A functional program in the format of an object file that contains no unresolved references, relocation information, symbol table, or debugging information.

**execution time** See response time.

**exponent** In the numerical representation system of floating-point arithmetic, the value that is placed in the exponent field.

**external label** Also called global label. A label referring to an object that can be referenced from files other than the one in which it is defined.



**fairness** A property of an allocation scheme, such as a bus arbitration protocol, that ensures that no device, even one with low priority, ever be completely locked out from the bus.

**false sharing** A sharing situation in which two unrelated shared variables are located in the same cache block and the full block is exchanged between processors even though the processors are accessing different variables.

**finite state machine** A sequential logic function consisting of a set of inputs and outputs, a next-state function that maps the current state and the inputs to a new state, and an output function that maps the current state and possibly the inputs to a set of asserted outputs.

**firmware** Microcode implemented in a memory structure, typically ROM or RAM.

**flip-flop** A memory element for which the output is equal to the value of the stored state inside the element and for which the internal state is changed only on a clock edge.

**floating point** Computer arithmetic that represents numbers in which the binary point is not fixed.

**floppy disk** A portable form of secondary memory composed of a rotating mylar platter coated with a magnetic recording material.

**flush (instructions)** To discard instructions in a pipeline, usually due to an unexpected event.

**formal parameter** A variable that is the argument to a procedure or macro; replaced by that argument once the macro is expanded.

**forward reference** A label that is used before it is defined.

**forwarding** Also called bypassing. A method of resolving a data hazard by retrieving the missing data element from internal buffers rather than waiting for it to arrive from programmer-visible registers or memory.

**frame pointer** A value denoting the location of the saved registers and local variables for a given procedure.

**fully associative cache** A cache structure in which a block can be placed in any location in the cache.

**fully connected network** A network that connects processor-memory nodes by supplying a dedicated communication link between every node.

**gate** A device that implements basic logic functions, such as AND or OR.

**general-purpose electronic computer** A computer that has not been constructed for one specific function.

**general-purpose register (GPR)** A register that can be used for addresses or for data with virtually any instruction.

**geometric mean**  $\sqrt[n]{\prod_{i=1}^n \text{Execution time ratio}_i}$  A formula useful for summarizing execution times that have been normalized.

**gigabyte** Traditionally 1,073,741,824 ( $2^{30}$ ) bytes, although some communications and secondary storage systems have redefined it to mean 1,000,000,000 ( $10^9$ ) bytes.

**global label** See external label.

**global miss rate** The fraction of references that miss in all levels of a multilevel cache.

**global pointer** The register that is reserved for static data.

**GPR** See general-purpose register (GPR).

**guard** The first of two extra bits kept on the right during intermediate calculations of floating-point numbers; used to improve rounding accuracy.

**handshaking protocol** A series of steps used to coordinate asynchronous bus transfers in which the sender and receiver proceed to the next step only when both parties agree that the current step has been completed.

**hard disk** A form of secondary memory composed of rotating metal platters coated with a magnetic recording material.

**hardwired control** An implementation of finite state machine control typically using programmable logic arrays (PLAs) or collections of PLAs and random logic.

$$\text{HM} = \frac{n}{\sum_{i=1}^n \frac{1}{\text{Rate}_i}}$$

**harmonic mean of rates** A summary that tracks execution time when the data is given as rates rather than as a times.

**hexadecimal** Numbers in base 16.

**high-level programming language** A portable language such as C, Fortran, or Java composed of English words and algebraic notation that can be translated by a compiler into assembly language.

**hit rate** The fraction of memory accesses found in a cache.

**hit time** The time required to access a level of the memory hierarchy, including the time needed to determine whether the access is a hit or a miss.

- hold time** The minimum time during which the input must be valid after the clock edge.
- horizontal microcode** Use of microinstructions containing many fields that can control the datapath units in parallel and require little additional decoding. The use of many fields makes the microinstructions wider or more horizontal.
- immediate addressing** An addressing regime in which the operand is a constant within the instruction itself.
- implementation** Hardware that obeys the architecture abstraction.
- imprecise exception** See imprecise interrupt.
- imprecise interrupt** Also called imprecise exception. Interrupts or exceptions in pipelined computers that are not associated with the exact instruction that was the cause of the interrupt or exception.
- in-order commit** A commit in which the results of pipelined execution are written to the programmer-visible state in the same order that instructions are fetched.
- in-order execution** A conventional pipelined execution, in which all following instructions must wait when an instruction is blocked from executing.
- input device** A mechanism through which the computer is fed information, such as the keyboard or mouse.
- instruction format** A form of representation of an instruction composed of fields of binary numbers.
- instruction latency** The inherent execution time for an instruction.
- instruction mix** A measure of the dynamic frequency of instructions across one or many programs.
- instruction set** The vocabulary of commands understood by a given architecture.
- instruction set architecture** Also called architecture. An abstract interface between the hardware and the lowest level software of a machine that encompasses all the information necessary to write a machine language program that will run correctly, including instructions, registers, memory size, and so on.
- integrated circuit** Also called chip. A device combining dozens to millions of transistors.
- interrupt** An exception that comes from outside of the processor. (Some architectures use the term *interrupt* for all exceptions.)
- interrupt-driven I/O** An I/O scheme that employs interrupts to indicate to the processor that an I/O device needs attention.
- interrupt enable** See exception enable.
- interrupt handler** A piece of code that is run as a result of an exception or an interrupt.
- I/O instruction** A dedicated instruction that is used to give a command to an I/O device and that specifies both the device number and the command word (or the location of the command word in memory).

**jump address table** Also called jump table. A table of addresses of alternative instruction sequences.

**jump-and-link instruction** An instruction that jumps to an address and simultaneously saves the address of the following instruction in a register (\$ra in MIPS).

**jump table** See jump address table.

**kernel benchmark** A small, time-intensive code fragment from a real program that is used for performance evaluation.

**kernel mode** Also called supervisor mode. A mode indicating that a running process is an operating system process.

**kilobyte**  $1024 (2^{10})$  bytes.

**LAN** See local area network (LAN).

**latch** A memory element in which the output is equal to the value of the stored state inside the element and the state is changed whenever the appropriate inputs change and the clock is asserted.

**latency (pipeline)** The number of stages in a pipeline or the number of stages between two instructions during execution.

**least recently used (LRU)** A replacement scheme in which the block replaced is the one that has been unused for the longest time.

**least significant bit** The rightmost bit in a MIPS word.

**level-sensitive clocking** A timing methodology in which state changes occur at either high or low clock levels but are not instantaneous, as such changes are in edge-triggered designs.

**link editor** See linker.

**linker** Also called link editor. A systems program that combines independently assembled machine language programs and resolves all undefined labels into an executable file.

**load-store machine** Also called register-register machine. An instruction set architecture in which all operations are between registers and data memory may only be accessed via loads or stores.

**load-use data hazard** A specific form of data hazard in which the data requested by a load instruction has not yet become available when it is requested.

**loader** A systems program that places an object program in main memory so that it is ready to execute.

**local area network (LAN)** A network designed to carry data within a geographically confined area, typically within a single building.

**local label** A label referring to an object that can be used only within the file in which it is defined.

**local miss rate** The fraction of references to one level of a cache that miss; used in multilevel hierarchies.

**lock** A synchronization device that allows access to data to only one processor at a time.

- logic minimization** A technique for reducing the number of gates needed to implement a set of logic functions.
- loop unrolling** A technique to get more performance from loops that access arrays, in which multiple copies of the loop body are made and instructions from different iterations are scheduled together.
- LRU** *See* least recently used (LRU).
- machine language** Binary representation used for communication within a computer system.
- macro** A pattern-matching and replacement facility that provides a simple mechanism to name a frequently used sequence of instructions.
- macroinstruction** An instruction in the instruction set architecture being implemented, used to distinguish the instructions visible to the programmer from the microinstructions of a microprogrammed control unit.
- magnetic disk** A form of nonvolatile secondary memory composed of rotating platters coated with a magnetic recording material.
- main memory** *See* primary memory.
- main-memory coherence** Consistency in the value of data in memory in a network-connected multiprocessor.
- massively parallel** A computer with at least 100 processors.
- maximally encoded** Use of encoded forms of control that require multiple levels of decode; vertical microcode is maximally encoded.
- megabyte** Traditionally 1,048,576 ( $2^{20}$ ) bytes, although some communications and secondary storage systems have redefined it to mean 1,000,000 ( $10^6$ ) bytes.
- megaFLOPS** *See* million floating-point operations per second (MFLOPS).
- memory** The storage area in which programs are kept when they are running and that contains the data needed by the running programs.
- memory hierarchy** A structure that uses multiple levels of memories; as the distance from the CPU increases, the size of the memories and the access time both increase.
- memory-mapped I/O** An I/O scheme in which portions of address space are assigned to I/O devices and reads and writes to those addresses are interpreted as commands to the I/O device.
- MESI cache coherency protocol** A write-invalidate protocol whose name is an acronym for the four states of the protocol: Modified, Exclusive, Shared, Invalid.
- message passing** Communicating between multiple processors by explicitly sending and receiving information.

**metastability** A situation that occurs if a signal is sampled when it is not stable for the required set-up and hold times, possibly causing the sampled value to fall in the indeterminate region between a high and low value.

**MFLOPS** *See* million floating-point operations per second (MFLOPS).

**microcode** The set of microinstructions that control a processor.

**microcode assembler** A program that translates microprograms into microinstructions that can be implemented in a ROM or PLA.

**microinstruction** A representation of control using low-level instructions, each of which asserts a set of control signals that are active on a given clock cycle as well as specifies what microinstruction to execute next.

**microprogram** A symbolic representation of control in the form of instructions, called microinstructions, that are executed on a simple micromachine.

**microprogrammed control** A method of specifying control that uses microcode rather than a finite state representation.

**million floating-point operations per second (MFLOPS)** Also called megaFLOPS. A measurement of program execution speed based on the number of millions of floating-point operations executed per second. MFLOPS is computed as the number of floating-point operations in a program divided by the product of the execution time and  $10^6$ .

**million instructions per second (MIPS)** A measurement of program execution speed based on the number of millions of instructions. MIPS is computed as the instruction count divided by the product of the execution time and  $10^6$ .

**MIMD** *See* multiple instruction streams, multiple data streams (MIMD).

**minimally encoded** Use of an unencoded control format that can directly control a datapath; horizontal microcode is minimally encoded.

**minterms** Also called product terms. A set of logic inputs joined by conjunction (AND operations); the product terms form the first logic stage of the programmable logic array (PLA).

**MIPS** *See* million instructions per second (MIPS).

**miss penalty** The time required to fetch a block into a level of the memory hierarchy from the lower level, including the time to access the block, transmit it from one level to the other, and insert it in the level that experienced the miss.

**miss rate** The fraction of memory accesses not found in a level of the memory hierarchy.

- most significant bit** The leftmost bit in a MIPS word.
- motherboard** A plastic board containing packages of integrated circuits or chips, including processor, cache, memory, and connectors for I/O devices such as networks and disks.
- multicomputer** Parallel processors with multiple private addresses.
- multicycle implementation** Also called multiple clock cycle implementation. An implementation in which an instruction is executed in multiple clock cycles.
- multilevel cache** A memory hierarchy with multiple levels of caches, rather than just a cache and main memory.
- multiple clock cycle implementation** *See* multicycle implementation.
- multiple-instruction issue** A procedure in which the instruction fetch unit can send multiple instructions to the next pipeline stage in a single clock cycle.
- multiple instruction streams, multiple data streams (MIMD)** A computer classification in Flynn's taxonomy referring to computers that use multiple instruction streams and multiple data streams.
- multiprocessor** Parallel processors with a single shared address.
- multistage network** A network that supplies a small switch at each node.
- NAND gate** An inverted AND gate.
- network bandwidth** Informally, the peak transfer rate of a network; can refer to the speed of a single link or the collective transfer rate of all links in the network.
- next-state counter** A counter that supplies the sequential next state.
- next-state function** A combinational function that, given the inputs and the current state, determines the next state of a finite state machine.
- next-state output** An output of the combinational logic that specifies the next-state number.
- nonblocking cache** A cache that allows the processor to make references to the cache while the cache is handling an earlier miss.
- nonuniform memory access (NUMA)** A type of single-address space multiprocessor in which some memory accesses are faster than others depending which processor asks for which word.
- nonvolatile memory** A form of memory that retains data even in the absence of a power source and that is used to store programs between runs. Magnetic disk is nonvolatile and DRAM is not.
- nop** An instruction that does no operation to change state.

**NOR gate** An inverted OR gate.

**normalized** A number in floating-point notation that has no leading 0s.

**NUMA** *See* nonuniform memory access (NUMA).

**object program** A combination of machine language instructions, data, and information needed to place them properly in memory.

**opcode** The field that denotes the operation and format of an instruction.

**operating system** Supervising program that manages the resources of a computer for the benefit of the programs that run on that machine.

**out-of-order commit** A commit in which the results of pipelined execution need not be written to the programmer visible state in the same order that instructions are fetched.

**out-of-order execution** A situation in pipelined execution when an instruction blocked from executing does not cause the following instructions to wait.

**output device** A mechanism that conveys the result of a computation to the user.

**overflow (floating-point)** A situation in which a positive exponent becomes too large to fit in the exponent field.

**page fault** An event that occurs when an accessed page is not present in main memory.

**page mode** A mechanism in DRAM that provides the ability to access multiple bits of a row by changing the column address only and, hence, is faster than a normal access cycle that changes row and column addresses.

**page table** The table containing the virtual to physical address translations in a virtual memory system. The table, which is stored in memory, is typically indexed by the virtual page number; each entry in the table contains the physical page number for that virtual page if the page is currently in memory.

**parallel processing program** A single program that runs on multiple processors simultaneously.

**PC** *See* program counter (PC).

**PC-relative addressing** An addressing regime in which the address is the sum of the program counter (PC) and a constant in the instruction.

**personal computer** A general-purpose computer designed to be manufactured in high volume and at a cost affordable enough to allow for use in the home.

**physical address** An address in main memory.

**physically addressed cache** A cache that is addressed by a physical address.

**pipeline data hazard** *See* data hazard.



**pipeline stall** Also called bubble. A stall initiated in order to resolve a hazard.

**pipelining** An implementation technique in which multiple instructions are overlapped in execution, much like to an assembly line.

**pipelining stage** A step in executing an instruction that occurs simultaneously with other steps in other instructions and typically lasts one clock cycle.

**pixel** The smallest individual picture element. Screen are composed of hundreds of thousands to millions of pixels, organized in a matrix.

**PLA** *See* programmable logic array (PLA).

**polling** The process of periodically checking the status of an I/O device to determine the need to service the device.

**precise exception** *See* precise interrupt.

**precise interrupt** Also called precise exception. An interrupt or exception that is always associated with the correct instruction in pipelined computers.

**prefetching** A technique in which data blocks needed in the future are brought into the cache early by the use of special instructions that specify the address of the block.

**primary memory** Also called main memory. Volatile memory used to hold programs while they are running; typically consists of DRAM in today's computers.

**procedure** A stored subroutine that performs a specific task based on the parameters with which it is provided.

**procedure call convention** *See* register-use convention.

**procedure call frame** A block of memory that is used to hold values passed to a procedure as arguments, to save registers that a procedure may modify but that the procedure's caller does not want changed, and to provide space for variables local to a procedure.

**procedure frame** Also called activation record. The segment of the stack containing a procedure's saved registers and local variables.

**processor-memory bus** A bus that connects processor and memory and that is short, generally high speed, and matched to the memory system so as to maximize memory-processor bandwidth.

**product terms** *See* minterms.

**program counter (PC)** The register containing the address of the instruction in the program being executed

**programmable logic array (PLA)** A structured-logic element composed of a set of inputs and corresponding input complements and two stages of logic: the first generating product terms of the inputs and input complements and the second generating sum terms of the product terms. Hence, PLAs implement logic functions as a sum of products.

**programmable ROM (PROM)** A form of read-only memory that can be programmed when a designer knows its contents.

**PROM** *See* programmable ROM (PROM).

**propagation time** The time required for an input to a flip-flop to propagate to the outputs of the flip-flop.

**protection** A set of mechanisms for ensuring that multiple processes sharing the processor, **memory**, or I/O devices cannot **interfere**, intentionally or unintentionally, with one another by reading or writing each other's data. These mechanisms also isolate the operating system from a user process.

**pseudoinstruction** A common variation of assembly language instructions often treated as if it were an instruction in its own right.

**quotient** The primary result of a division; a number that when multiplied by the divisor and added to the remainder produces the dividend.

**RAID** *See* redundant arrays of inexpensive disks (RAID).

**raster cathode ray tube (CRT) display** A display, such as a television set, that scans an image one line at a time, 30 to 75 times per second.

**read-only memory (ROM)** A memory whose contents are designated at creation time, after which the contents can only be read. ROM is used as structured logic to implement a set of logic functions by using the terms in the logic functions as address inputs and the outputs as bits in each word of the memory.

**receive message routine** A routine used by a processor in machines with private memories to accept a message from another processor.

**recursive procedures** Procedures that call themselves either directly or indirectly through a chain of calls.

**redundant arrays of inexpensive disks (RAID)** An organization of disks that uses an array of small and inexpensive disks so as to increase both performance and reliability.

**reference bit** Also called use bit. A field that is set whenever a page is accessed and that is used to implement LRU or other replacement schemes.

**register addressing** A mode of addressing in which the operand is a register.

**register file** A state element that consists of a set of registers that can be read and written by supplying a register number to be accessed.

**register-register machine** *See* load-store machine.

**register use** *See* register-use convention.

**register-use convention** Also called procedure call convention. A software protocol governing the use of registers by procedures.

**relocation information** The segment of a Unix object file that identifies instructions and data words that depend on absolute addresses.

**remainder** The secondary result of a division; a number that when added to the product of the quotient and the divisor produces the dividend.

**rename buffer** Also called rename register. An extra internal register within processors that is used to hold results while waiting for the commit unit to commit the result to one of the real registers.

**rename register** *See* rename buffer.

**reorder buffer** A register that holds instructions in a dynamic pipelined machine whose results have not yet been committed to programmer-visible registers or memory; machines with out-of-order execution and in-order commit will retire an instruction from the reorder buffer only when the instruction has finished execution and all instructions ahead of it have been completed.

**reservation station** A buffer within a functional unit that holds the operands and the operation.

**response time** Also called execution time. The total time required for the computer to complete a task, including disk accesses, memory accesses, I/O activities, operating system overhead, CPU execution time, and so on.

**restartable instruction** An instruction that can resume execution after an exception is resolved without the exception's affecting the result of the instruction.

**return address** A link to the calling site that allows a procedure to return to the proper address; in MIPS it is stored in register `$ra`.

**ROM** *See* read-only memory (ROM).

**rotation latency** Also called delay. The time required for the desired sector of a disk to rotate under the read/write head; usually assumed to be half the rotation time.

**round** Method to make the intermediate floating-point result fit the floating-point format; the goal is typically to find the nearest number that can be represented in the format.

**scientific notation** A notation that renders numbers with a single digit to the left of the decimal point.

**SCSI** *See* small computer systems interface (SCSI).

**secondary memory** Nonvolatile memory used to store programs and data between runs; typically consists of magnetic disks in today's computers.

**sector** One of the segments that make up a track on a magnetic disk; a sector is the smallest amount of information that is read or written on a disk.

**seek** The process of positioning a read/write head over the proper track on a disk.

**segmentation** A variable-size address mapping scheme in which an address consists of two parts: a segment number, which is mapped to a physical address, and a segment offset.

**selector value** Also called control value. The control signal that is used to select one of the input values of a multiplexor as the output of the multiplexor.

**semiconductor** A substance that does not conduct electricity well.

**send message routine** A routine used by a processor in machines with private memories to pass to another processor.

**separate compilation** Splitting a program across many files, each of which can be compiled without knowledge of what is in the other files.

**sequential access memory** Memory whose access time differs depending on the location of the data being retrieved because data is stored sequentially so that all data must be passed over to access the final bit of information; contrasts with random access memory, in which any bit may be accessed in the same time.

**sequential logic** A group of logic elements that contain memory and hence whose value depends on the inputs as well as the current contents of the memory.

**set-associative cache** A cache that has a fixed number of locations (at least two) where each block can be placed.

**set-up time** The minimum time that the input to a memory device must be valid before the clock edge.

**shared memory** A memory for a parallel processor with a single address space, implying implicit communication with loads and stores.

**sign-extend** To increase the size of a data item by replicating the high-order sign bit of the original data item in the high-order bits of the larger, destination data item.

**significand** In the numerical representation system of floating-point arithmetic, the value in that is placed in the significand field.

**silicon** A substance found in sand that does not conduct electricity well.

**silicon crystal ingot** A rod composed of silicon crystal that is between 6 and 12 inches in diameter and about 12 to 24 inches long.

**SIMD** *See* single instruction stream, multiple data streams (SIMD).

**SIMM** *See* single in-line memory module (SIMM).

**single clock cycle implementation** *See* single-cycle implementation.

**single-cycle implementation** Also called single clock cycle implementation. An implementation in which an instruction is executed in one clock cycle.

**single in-line memory module (SIMM)** A small printed circuit board containing 4 to 24 DRAM integrated circuits. Today's computers use SIMMs to allow main memory to be upgraded and expanded over time by the customer.

single instruction stream, multiple data streams

**(SIMD)** A computer classification in Flynn's taxonomy that refers to computers with single instruction streams but multiple data streams and in which a single instruction operates on many data elements at the same time.

single instruction stream, single data stream **(SISD)**

A computer classification in Flynn's taxonomy that refers to computers with single instruction streams and single data streams. (SISD is the conventional processor covered in the first eight chapters.)

single precision A floating-point value represented in a single 32-bit word.

**SISD** *See* single instruction stream, single data stream (SISD).

slave A device that responds to read and write requests but does not generate them and hence cannot be a bus master.

small computer systems interface **(SCSI)** A bus used as a standard for I/O devices.

**SMP** *See* symmetric multiprocessor (SMP).

snooping cache coherency A method for maintaining cache coherency in which all cache controllers monitor or snoop on the bus to determine whether or not they have a copy of the desired block.

source language The high-level language in which a program is originally written.

spatial locality The locality principle stating that if a data location is referenced, data locations with nearby addresses will tend to be referenced soon.

**SPEC benchmark** *See* system performance evaluation cooperative (SPEC) benchmark.

speculative execution A pipelining technique that combines dynamic scheduling with branch prediction.

speedup The measure of how a machine performs relative to how it previously performed before an enhancement was implemented. Speedup is equal to the ratio of execution time before the enhancement to execution time after the enhancement.

split cache A scheme in which a level of the memory hierarchy is composed of two independent caches that operate in parallel with each other with one handling instructions and one handling data.

split transaction protocol A protocol in which the bus is released during a bus transaction while the requester is waiting for the data to be transmitted, which frees the bus for access by another requester.

**SRAM** *See* static random access memory (SRAM).

stack A data structure for spilling registers organized as a last-in-first-out queue.

**stack frame** See procedure call frame.

**stack pointer** A value denoting the most recently allocated address in a stack that shows where registers should be spilled or where old register values can be found.

**stack segment** The portion of memory used by a program to hold procedure call frames.

**state assignment** A control optimization that works by attempting to choose the state numbers such that the resulting logic equations contain more redundancy and can thus be simplified.

**state element** A memory element.

**state input** An input to the combinational logic that specifies the current state.

**static data** The portion of memory that contains data whose size is known to the compiler and whose lifetime is the program's entire execution.

**static random access memory (SRAM)** A memory where data is stored statically (as in flip-flops) rather than dynamically (as in DRAM). SRAMs are faster than DRAMs, but less dense and more expensive per bit.

**sticky bit** A bit used in rounding in addition to guard and round that is set whenever there are nonzero bits to the right of the round bit.

**stored-program computer** A computer whose instructions are represented as numbers, allowing the same memory to contain instructions and data and thus allowing programs to produce programs.

**stored-program concept** The idea that instructions and data of many types can be stored in memory as numbers, leading to the stored program computer.

**structural hazard** An occurrence in which a planned instruction cannot execute in the proper clock cycle because the hardware cannot support the combination of instructions that are set to execute in the given clock cycle.

**subroutine library** A collection of commonly used programs.

**sum of products** A form of logical representation that employs a logical sum (OR) of products (terms joined using the AND operator).

**supercomputer** The fastest and most expensive computer, typically used for scientific computation. Supercomputers generally cost between \$1 and \$30 million.

**superpipelining** A technique that increases processor speed by lengthening pipelines.

**superscalar** An advanced pipelining technique that enables the processor to execute more than one instruction per clock cycle.

**superscalar pipelining** A technique that replicates internal components of the computer in order to launch and execute multiple instructions in every pipeline stage.

**supervisor mode** *See* kernel mode.

**symbol table** A table that matches names of labels to the addresses of the memory words that instructions occupy.

**symmetric multiprocessor (SMP)** Also called **UMA** machine. A multiprocessor in which accesses to main memory take the same amount of time no matter which processor requests the access and no matter which word is asked.

**synchronization** The process of coordinating the behavior of two or more processes, which may be running on different processors.

**synchronizer failure** A situation in which a flip-flop enters a metastable state and where some logic blocks reading the output of the flip-flop see a 0 while others see a 1.

**synchronous bus** A bus that includes a clock in the control lines and a fixed protocol for communicating that is relative to the clock.

**synchronous system** A memory system that employs clocks and where data signals are read only when the clock indicates that the signal values are stable.

**system call** A special instruction that transfers control from user mode to a dedicated location in supervisor code space, invoking the exception mechanism in the process.

**system CPU time** The CPU time spent in the operating system performing tasks on behalf of the program.

**system performance evaluation cooperative (SPEC) benchmark** A set of standard **CPU-intensive**, integer and floating point benchmarks based on real programs.

**systems software** Software that provides services that are commonly useful, including operating systems, compilers, and assemblers.

**tag** A field in a table used for a memory hierarchy that contains the address information required to identify whether the associated block in the hierarchy corresponds to a requested word.

**temporal locality** The principle stating that if a data location is referenced then it will tend to be referenced again soon.

**terabyte** Originally 1,099,511,627,776 ( $2^{40}$ ) bytes, although some communications and secondary storage systems have redefined it to mean 1,000,000,000,000 ( $10^{12}$ ) bytes.

**text segment** The segment of a Unix object file that contains the machine language code for routines in the source file.

**three Cs model** A cache model in which all cache misses are classified into one of three categories: compulsory misses, capacity misses, and conflict misses.

**tick** See clock cycle.

**TLB** See translation-lookaside buffer (TLB).

**track** One of 1000 to 5000 concentric circles that makes up the surface of a magnetic disk.

**transaction processing** A type of application that involves handling small short operations (called transactions) that typically require both I/O and computation. Transaction processing applications typically have both response time requirements and a performance measurement based on the throughput of transactions.

**transfer time** The time required to transfer a block of bits, typically a sector, during disk access.

**transistor** An on/off switch controlled by electricity.

**translation-lookaside buffer (TLB)** A cache that keeps track of recently used address mappings to avoid an access to the page table.

**ulp** See units in the last place (ulp).

**UMA** See uniform memory access (UMA).

**UMA machine** See symmetric multiprocessor (SMP).

**underflow (floating-point)** A situation in which a negative exponent becomes too large to fit in the exponent field.

**uniform memory access (UMA)** Memory access that takes the same amount of time no matter which processor requests the access and no matter which word is asked for.

**units in the last place (ulp)** The number of bits in error in the least significant bits of the significand between the actual number and the number that can be represented.

**unresolved reference** A reference that requires more information from an outside source in order to be complete.

**use bit** See reference bit.

**user CPU time** The CPU time spent in a program itself.

**vacuum tube** An electronic component, predecessor of the transistor, that consists of a hollow glass tube about 5 to 10 cm long from which as much air has been removed as possible.

**valid bit** A field in the tables of a memory hierarchy that indicates that the associated block in the hierarchy contains valid data.



**vector processor** An architecture and compiler model that was popularized by supercomputers in which high-level operations work on linear arrays of numbers.

**vector supercomputer** A supercomputer whose instructions operate on vectors of numbers, typically 64 floating-point numbers at a time.

**vectored interrupt** An interrupt for which the address to which control is transferred is determined by the cause of the exception.

**vertical microcode** Use of microinstructions containing many fewer fields that require additional decoding before being used to control the datapath units. The use of fewer fields makes the microinstructions narrower or more vertical.

**very large scale integrated (VLSI) circuit** A device containing tens of thousands to millions of transistors.

**virtual address** An address that corresponds to a location in virtual space and is translated by address mapping to a physical address when memory is accessed.

**virtual machine** A virtual computer that appears to have nondelayed branches and loads and a richer instruction set than the actual hardware.

**virtual memory** A technique that uses main memory as a "cache" for secondary storage.

**virtually addressed cache** A cache that is accessed with a virtual address rather than a physical address.

**VLSI circuit** *See* very large scale integrated (VLSI) circuit.

**volatile memory** Storage, such as DRAM, that only retains data if it is receiving power.

**wafer** A slice from a silicon ingot no more than 0.1 inch thick, used to create chips.

**weighted arithmetic mean** A summary that tracks the execution time of a workload with weighting factors designed to reflect the presence of the programs in a workload; computed as the sum of the products of weighting factors and execution times.

**wide area network** A network extended over hundreds of kilometers which can span a continent.

**word** The natural unit of access in a computer, usually a group of 32 bits; corresponds to the size of a register in the MIPS architecture.

**workload** A set of programs run on a computer that is either the actual collection of applications run by a user or is constructed from real programs to approximate such a mix. A typical workload specifies both the programs as well as the relative frequencies.

**write-back** A scheme that handles writes by updating values only to the block in the cache, then writing the modified block to the lower level of the hierarchy when the block is replaced.

**write-broadcast** A snooping protocol scheme in which the writing processor disseminates the new data over the bus, allowing all copies to be updated with the new value.

**write buffer** A queue that holds data while the data are waiting to be written to memory.

**write-invalidate** A type of snooping protocol in which the writing processor causes all copies in other caches to be invalidated before changing its local copy, which allows it to update the local data until another processor asks for it.

**write-through** A scheme in which writes always update both the cache and the memory, ensuring that data is always consistent between the two.

**yield** The percentage of good dies from the total number of dies on the wafer.