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 to facilitate design of sophisticated systems.

acronym A word constructed by taking the initial letters of a string of words. For example: RAM is an acronym for Random Access Memory, and CPU is an acronym for Central Processing Unit.

active matrix display A liquid crystal display using a transistor to control the transmission of light at each individual pixel.

address A value used to delineate the location of a specific data element within a memory array.

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.

advanced load In IA-64, a speculative load instruction with support to check for aliases that could invalidate the load.

AGP An extended version of the original PCI I/O bus, which provided up to eight times the bandwidth of the original PCI bus to a single card slot. Its primary purpose was to connect graphics subsystems into PC systems.

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.

alignment restriction A requirement that data be aligned in memory on natural boundaries.

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.

antidependence Also called name dependence. An ordering forced by the reuse of a name, typically a register, rather than by a true dependence that carries a value between two instructions.

antifuse A structure in an integrated circuit that when programmed makes a permanent connection between two wires.

application binary interface (ABI) The user portion of the instruction set plus the operating system interfaces used by application programmers. Defines a standard for binary portability across computers.

application programming interface (API) A set of function and data structure definitions providing an interface to a library of functions.

architectural registers The instruction set visible registers of a processor; for example, in MIPS, these are the 32 integer and 16 floating-point registers.

arithmetic intensity The ratio of floating-point operations in a program to the number of data bytes accessed by a program from main memory.

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 instructions 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 memory operation A memory read-modify-write operation sequence that completes without any intervening access.

atomic swap 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.

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).

behavioral specification Describes how a digital system operates functionally.

biased notation A notation that represents the most negative value by $00\ldots000_{two}$ and the most positive value by $11\ldots11_{two}$, with 0 typically having the value $10\ldots00_{two}$, thereby biasing the number such that the number plus the bias has a nonnegative representation.

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

bisection bandwidth The bandwidth between two equal parts of a multiprocessor. This measure is for a worst-case split of the multiprocessor.

bit error rate The fraction in bits of a message or collection of messages that is incorrect.

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

blocking assignment In Verilog, an assignment that completes before the execution of the next statement.

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 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.

branch target buffer A structure that caches the destination PC or destination instruction for a branch. It is usually organized as a cache with tags, making it more costly than a simple prediction buffer.

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 master A unit on the bus that can initiate bus requests.

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.

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

cache-coherent NUMA (CC-NUMA) A nonuniform memory access multiprocessor that maintains coherence for all caches.

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 blocks needed to satisfy the request.

carrier signal A continuous signal of a single frequency capable of being modulated by a second data-carrying signal.

cathode ray tube (CRT) display A display, such as a television set, that displays an image using an electron beam scanned across a screen.

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.

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.

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

clock period The length of each clock cycle.

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

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

clusters Collections of computers connected via I/O over standard network switches to form a message-passing multiprocessor.

coarse-grained multithreading A version of hardware multithreading that suggests switching between threads only after significant events, such as a cache 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.
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 is eliminated in a fully associative cache of the same size.
constellation A cluster that uses an SMP as the building block.
context switch A changing of the internal state of the processor to allow a different process to use the processor; 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 Also called branch 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.
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.
cooperative thread array (CTA) A set of concurrent threads in one multiprocessor that execute the same thread program and may cooperate to compute a result. A GPU CTA implements a CUDA thread block.
correlating predictor A branch predictor that combines local behavior of a particular branch and global information about the behavior of some recent number of executed branches.
CPU execution time Also called CPU time. The actual time the CPU spends computing for a specific task.
crossbar network A network that allows any node to communicate with any other node in one pass through the network.
CUDA A scalable parallel programming model and language based on C/C++. It is a parallel programming platform for GPUs and multicore CPUs.
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.
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-level parallelism Parallelism achieved by operating on independent data.
data parallelism Parallelism achieved by having massive data.
data rate Performance measure of bytes per unit time, such as GB/second.
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 $2n$ 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.

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.

desktop computer A computer designed for use by an individual, usually incorporating a graphics display, keyboard, and mouse.

die The individual rectangular sections that are cut from a wafer, more informally known as chips.

DIMM (dual inline memory module) A small board that contains DRAM chips on both sides. SIMMs have DRAMs on only one side. Both DIMMs and SIMMs are meant to be plugged into memory slots, usually on a motherboard.

Direct3D A graphics API defined by Microsoft and partners.

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 with 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. Used for cache coherence.

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.

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.

dividend A number being divided.

divisor A number that the dividend is divided by.

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.

dynamic branch prediction Prediction of branches at runtime using runtime information.

dynamic multiple issue An approach to implementing a multiple-issue processor in which many decisions are made during execution by the processor.

dynamic pipeline scheduling Hardware support for reordering the order of instruction execution so as to avoid stalls.

dynamic random access memory (DRAM) Memory built as an integrated circuit; it provides random access to any location.

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

embedded computer A computer inside another device used for running one predetermined application or collection of software.

error detection code A code that enables the detection of an error in data, but not the precise location and, hence, correction of the error.

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.

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

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.

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.

field programmable device (FPD) An integrated circuit containing combinational logic, and possibly memory devices, that are configurable by the end user.

field programmable gate array (FPGA) A configurable integrated circuit containing both combinational logic blocks and flip-flops.

fine-grained multithreading A version of hardware multithreading that suggests switching between threads after every instruction.

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.

flat panel display, liquid crystal display A display technology using a thin layer of liquid polymers that can be used to transmit or block light according to whether a charge is applied.

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.

fraction The value, generally between 0 and 1, placed in the fraction field.

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

global memory Per-application memory shared by all threads.

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

global pointer The register that is reserved to point to static data.

GPGPU Using a GPU for general-purpose computation via a traditional graphics API and graphics pipeline.

GPU computing Using a GPU for computing via a parallel programming language and API.

graphics processing unit (GPU) A processor optimized for 2D and 3D graphics, video, visual computing, and display.

grid A set of thread blocks that execute the same kernel program.

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

handler Name of a software routine invoked to "handle" an exception or interrupt.

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.

hardware description language A programming language for describing hardware, used for generating simulations of a hardware design and also as input to synthesis tools that can generate actual hardware.

hardware multithreading Increasing utilization of a processor by switching to another thread when one thread is stalled.

hardware synthesis tools Computer-aided design software that can generate a gate-level design based on behavioral descriptions of a digital system.

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

heterogeneous system A system combining different processor types. A PC is a heterogeneous CPU–GPU system.

hexadecimal Numbers in base 16.

high-level programming language A portable language such as C, Fortran, or Java composed of 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.
hot-swapping Replacing a hardware component while the system is running.
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).
I/O rate Performance measure of I/Os per unit time, such as reads per second.
I/O requests Reads or writes to I/O devices.
implementation Hardware that obeys the architecture abstraction.
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.
input device A mechanism, such as the keyboard or mouse, through which the computer is fed information.
instruction format A form of representation of an instruction composed of fields of binary numbers.
instruction group In IA-64, a sequence of consecutive instructions with no register data dependences among them.
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 access, I/O, and so on.
instruction-level parallelism The parallelism among instructions.
integrated circuit Also called chip. A device combining dozens to millions of transistors.
interrupt An exception that comes from outside 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 handler A piece of code that is run as a result of an exception or an interrupt.
issue packet The set of instructions that issues together in one clock cycle; the packet may be determined statically by the compiler or dynamically by the processor.
issue slots The positions from which instructions could issue in a given clock cycle; by analogy, these correspond to positions at the starting blocks for a sprint.
Java bytecode Instruction from an instruction set designed to interpret Java programs.
Java Virtual Machine (JVM) The program that interprets Java bytecodes.
job-level parallelism or **process-level parallelism** Utilizing multiple processors by running independent programs simultaneously.
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).

Just In Time compiler (JIT) The name commonly given to a compiler that operates at runtime, translating the interpreted code segments into the native code of the computer.

kernel A program or function for one thread, designed to be executed by many threads.

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

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.

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 memory Per-thread local memory private to the thread.

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 only one processor at a time to access data.

lookup tables (LUTs) In a field programmable device, the name given to the cells, because they consist of a small amount of logic and RAM.

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.

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.

magnetic disk Also called hard disk. A form of nonvolatile secondary memory composed of rotating platters coated with a magnetic recording material.

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.

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 setup and hold times, possibly causing the sampled value to fall in the indeterminate region between a high and low value.

microarchitecture The organization of the processor, including the major functional units, their interconnection, and control.

microcode The set of microinstructions that control a processor.

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.

micro-operations The RISC-like instructions directly executed by the hardware in recent Pentium implementations.

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 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 or Multiple Instruction streams, Multiple Data streams. A multiprocessor.

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).

MIP-map A latin phrase *multum in parvo*, or much in a small space. A MIP-map contains precalculated images of different sizes, used to increase rendering speed and reduce artifacts.

mirroring Writing identical data to multiple disks to increase data availability.

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.

multicore microprocessor A microprocessor containing multiple processors (**cores**) in a single integrated circuit.

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 issue A scheme whereby multiple instructions are launched in one clock cycle.
multiply add (MAD) A single floating-point instruction that performs a compound operation: multiplication followed by addition.
multiprocessor A computer system with at least two processors. This is in contrast to a **uniprocessor**, which has one.
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 function A combinational function that, given the inputs and the current state, determines the next state of a finite-state machine.
nonblocking assignment An assignment that continues after evaluating the right-hand side, assigning the left-hand side the value only after all right-hand sides are evaluated.
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 on which processor asks for which word.
nonvolatile Storage device where data retains its value even when power is removed.
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 A logical bit-by-bit operation with two operands that calculates the NOT of the OR of the two operands.
NOR gate An inverted OR gate.
normalized A number in floating-point notation that has no leading 0s.
NOT A logical bit-by-bit operation with one operand that inverts the bits; that is, it replaces every 1 with a 0, and every 0 with a 1.
object-oriented language A programming language that is oriented around objects rather than actions or data versus logic.
opcode The field that denotes the operation and format of an instruction.
OpenGL An open standard graphics API.
OpenMP An API for shared memory multiprocessing in C, C++, or Fortran that runs on UNIX and Microsoft platforms. It includes compiler directives, a library, and runtime directives.
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 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 a user or another computer.
overflow (floating-point) A situation in which a positive exponent becomes too large to fit in the exponent field.
package Basically a directory that contains a group of related classes.
page fault An event that occurs when an accessed page is not present in main memory.
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.
PCI-Express A standard system I/O bus that uses point-to-point interconnect.
PC-relative addressing An addressing regime in which the address is the sum of the program counter (PC) and a constant in the instruction.
physical address An address in main memory.
physically addressed cache A cache that is addressed by a physical address.
pipeline stall Also called bubble. A stall initiated to resolve a hazard.
pipelining An implementation technique in which multiple instructions are overlapped in execution, much like an assembly line.
pixel The smallest individual picture element. Screens are composed of hundreds of thousands to millions of pixels, organized in a matrix.
poison A result generated when a speculative load yields an exception, or an instruction uses a poisoned operand.
polling The process of periodically checking the status of an I/O device to determine the need to service the device.
precise interrupt Also called precise exception. An interrupt or exception that is always associated with the correct instruction in pipelined computers.
predication A technique to make instructions dependent on predicates rather than on branches.
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 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.
process-level parallelism or **job-level parallelism** Utilizing multiple processors by running independent programs simultaneously.
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.
program counter (PC) The register containing the address of the instruction in the program being executed.
programmable array logic (PAL) Contains a programmable and-plane followed by a fixed or-plane.
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 logic device (PLD) An integrated circuit containing combinational logic whose function is configured by the end user.
programmable ROM (PROM) A form of read-only memory that can be programmed when a designer knows its contents.

propagation time The time required for an input to a flip-flop to propagate to the outputs of the flip-flop.
protected A Java keyword that restricts invocation of a method to other methods in that package.
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.
protection group The group of data disks or blocks that share a common check disk or block.
pseudoinstruction A common variation of assembly language instructions, often treated as if it were an instruction in its own right.
Pthreads A UNIX API for creating and manipulating threads. It comes with a library.
public A Java keyword that allows a method to be invoked by any other method.
quotient The primary result of a division; a number that when multiplied by the divisor and added to the remainder produces the dividend.
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.
reduction A function that processes a data structure and returns a single value.
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.
reg In Verilog, 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 renaming The renaming of registers, by the compiler or hardware, to remove antidependences.
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.
reorder buffer The buffer that holds results in a dynamically scheduled processor until it is safe to store the results to memory or a register.
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`.

rotational 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.

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.

sensitivity list The list of signals that specifies when an `always` block should be re-evaluated.

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

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.

server A computer used for running larger programs for multiple users, often simultaneously, and typically accessed only via a network.

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

setup time The minimum time that the input to a memory device must be valid before the clock edge.

shader A program that operates on graphics data, such as a vertex or a pixel fragment.

shading language A graphics rendering language, usually having a dataflow or streaming programming model.

shared memory Per-block memory shared by all threads of the block.

shared memory multiprocessor (SMP) 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.

silicon A natural element that is a semiconductor.

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

SIMD or Single Instruction stream, Multiple Data streams. A multiprocessor. The same instruction is applied to many data streams, as in a vector processor or array processor.

simple programmable logic device (SPLD) Programmable logic device usually containing either a single PAL or PLA.
simultaneous multithreading (SMT) A version of multithreading that lowers the cost of multithreading by utilizing the resources needed for multiple issue, dynamically schedule microarchitecture.
single precision A floating-point value represented in a single 32-bit word.
single-program multiple data (SPMD) A parallel programming model in which all threads execute the same program. SPMD threads typically coordinate with barrier synchronization.
single-cycle implementation Also called single clock cycle implementation. An implementation in which an instruction is executed in one clock cycle.
single instruction multiple thread (SIMT) A processor architecture that applies one instruction to multiple independent threads in parallel.
SISD or Single Instruction stream, Single Data stream. A uniprocessor.
small computer systems interface (SCSI) A bus used as a standard for I/O devices.
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.
software as a service Rather than selling software that is installed and run on customers own computers, the software is run at a remote site and made available over the Internet typically via a Web interface to customers. Customers are charged based on use rather than buying the program for local unlimited use.
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.
special function unit (SFU) A hardware unit that supports the computation of both transcendental functions and planar attribute interpolation.
speculation An approach whereby the compiler or processor guesses the outcome of an instruction to remove it as a dependence in executing other instructions.
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 requestor is waiting for the data to be transmitted, which frees the bus for access by another requestor.
SPMD Single Program, Multiple Data streams. The conventional MIMD programming model, where a single program runs across all processors.
stack A data structure for spilling registers organized as a last-in-first-out queue.
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.
standby spares Reserve hardware resources that can immediately take the place of a failed component.
state element A memory element.
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 method A method that applies to the whole class rather than to an individual object. It is unrelated to static in C.

static multiple issue An approach to implementing a multiple-issue processor where many decisions are made by the compiler before 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.
stop In IA-64, an explicit indicator of a break between independent and dependent instructions.
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.
striping Allocation of logically sequential blocks to separate disks to allow higher performance than a single disk can deliver.
strong scaling Speed-up achieved on a multiprocessor without increasing the size of the problem.
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.
structural specification Describes how a digital system is organized in terms of a hierarchical connection of elements.
sum of products A form of logical representation that employs a logical sum (OR) of products (terms joined using the AND operator).
supercomputer A class of computers with the highest performance and cost; they are configured as servers and typically cost millions of dollars.
superscalar An advanced pipelining technique that enables the processor to execute more than one instruction per clock cycle.
swap space The space on the disk reserved for the full virtual memory space of a process.
switched network A network of dedicated point-to-point links that are connected to each other with a switch.
symbol table A table that matches names of labels to the addresses of the memory words that instructions occupy.
symmetric multiprocessor (SMP) or uniform memory access (UMA). 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.
synchronization barrier Threads wait at a synchronization barrier until all participating threads arrive at the barrier.
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 in which 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, 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.

texture A 1D, 2D, or 3D array that supports sampled and filtered lookups with interpolated coordinates.

thread block A set of concurrent threads that execute the same thread program and may cooperate to compute a result.

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.

tournament branch predictor A branch predictor with multiple predictions for each branch and a selection mechanism that chooses which predictor to enable for a given branch.

trace cache An instruction cache that holds a sequence of instructions with a given starting address; in recent Pentium implementations, the trace cache holds micro-operations rather than IA-32 instructions.

track One of thousands of concentric circles that make 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.

transistor An on/off switch controlled by an electric signal.

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

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

uniform memory access (UMA) See SMP.

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.

unmapped A portion of the address space that cannot have page faults.

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

untaken branch One that falls through to the successive instruction. A taken branch is one that causes transfer to the branch target.

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 and that uses an electron beam to transfer data.

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.

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

Verilog One of the two most common hardware description languages.

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

VHDL One of the two most common hardware description languages.

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.

visual computing A mix of graphics processing and computing that lets you visually interact with computed objects via graphics, images, and video.

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

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

warp The set of parallel threads that execute the same instruction together in a SIMT architecture.

weak scaling Speed-up achieved on a multiprocessor while increasing the size of the problem proportionally to the increase in the number of processors.

weighted arithmetic mean An average of 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 (WAN) A network extended over hundreds of kilometers that can span a continent.

wire In Verilog, specifies a combinational signal.

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 constructed from real programs to approximate such a mix. A typical workload specifies both the programs and 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 buffer A queue that holds data while the data is 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.