CPUs & GPUs: The Processors Behind Computing and AI

Figures current as of October 2026

How processors work, who makes them, and why the chips behind artificial intelligence became the most valuable products in technology. The course covers what a processor does and how one core is made fast; the memory every chip waits on; the instruction sets x86, Arm, and RISC-V; the graphics processor and how it became a parallel computer; the chips in phones, personal computers (PCs), consoles, and cars; the processors that run data centers; AI chips from Nvidia, Advanced Micro Devices (AMD), and their challengers, and the cloud companies' own; the networks, power, and cooling that join chips into clusters; the software that programs them; how computing is measured and priced; and the business of selling it. A dated watch list closes the course.

MODULE 1 — What a Processor Does

Goal: Understand what a processor is made of and how it runs a program.

Transistors as switches, logic gates, and binary arithmetic; the stored-program computer and the cycle of fetching, decoding, and executing instructions; registers, the clock, and the instruction set, the processor's contract with software; from Intel's 4004 of 1971 to the chips of 2026; Moore's law and the end of free speed-ups about 2005; and how chips are made, in brief.

MODULE 2 — Making One Core Fast

Goal: See the tricks that make a single core fast, and why they hit a wall.

Pipelines, superscalar and out-of-order execution, and branch prediction; the speculation flaws found in 2018; vector units and multithreading; the performance equation; the power wall that stalled clock speeds near 5-6 gigahertz; and the multicore chips, mixed big and small cores, and Amdahl's law that followed.

MODULE 3 — Memory: The Wall Every Processor Hits

Goal: Learn why memory, not arithmetic, now limits most processors.

The ladder of memory from registers to disk; caches and locality; why memory's delay barely improved while its bandwidth multiplied; DDR5, LPDDR, GDDR7, and the stacked high-bandwidth memory beside AI chips; capacity against bandwidth; the roofline model; and the memory shortage of 2026.

MODULE 4 — Instruction Sets: x86, Arm, and RISC-V

Goal: Compare the three instruction sets and the business models behind them.

Complex and reduced instruction sets, and how micro-operations ended that argument; the Intel-AMD duopoly in x86; Arm's licenses and royalties, its first chip of its own, and its court fight with Qualcomm; RISC-V's open standard and China's bet on it; Apple's switch of 2020; and whether the instruction set still matters for efficiency.

MODULE 5 — The GPU: From Pixels to Parallel Computing

Goal: Understand how a graphics processor works and how it became the engine of AI.

The graphics problem and the rendering pipeline; ray tracing, upscaling, and generated frames; the history from the GeForce 256 of 1999 through programmable shaders, CUDA, AlexNet, and tensor cores; how a GPU trades delay for throughput; low-precision arithmetic; and the graphics-card market of 2026.

MODULE 6 — Phones, PCs, Consoles, and Cars

Goal: See how the chips in everyday devices are built, and why they cost more in 2026.

The system on a chip; the phone chips of Apple, Qualcomm, MediaTek, Samsung, and Google; Apple's M5 and M6, Intel's Panther Lake, AMD's laptop chips, Arm's spread into PCs, and Nvidia's RTX Spark; neural processing units and "AI PCs"; the chips in consoles, cars, and robots; and how AI's demand for memory squeezed devices in 2026.

MODULE 7 — Server CPUs

Goal: Learn what data-center processors do and who is winning their market.

The jobs of a server processor, including its new one as host for AI chips; the core counts of AMD, Intel, and the clouds' own chips; AMD's rise to a third of server chips and nearly half their revenue; Arm in the cloud; China's server processors, and the CPU-only machine atop the June 2026 list of supercomputers; and software licensed by the core.

MODULE 8 — AI Accelerators: Nvidia, AMD, and the Challengers

Goal: Read an AI chip's specification sheet, and know who makes the leading parts.

Training and inference, and the two phases of answering a prompt; the specification sheet decoded; Nvidia's yearly cadence from Hopper to Rubin and beyond, and its inference rack built on Groq's technology; AMD's MI455X and Helios rack and their customers; Intel's next attempt; the start-ups; and Huawei's Ascend chips, with export controls in brief.

MODULE 9 — Custom Silicon: TPUs, Trainium, and the Clouds' Own Chips

Goal: Understand why the largest buyers design their own chips, and what that costs.

Why build a chip for one job; the systolic array; Google's TPUs from 2015 to the eighth generation, now sold to outside customers; Amazon's Trainium; Microsoft's Maia; Meta's MTIA; OpenAI's Jalapeño; Tesla's AI5; the design houses, led by Broadcom; and whether the clouds' own chips will displace Nvidia's.

MODULE 10 — From Chip to Cluster: The Networks

Goal: See how thousands of chips are joined to work as one machine.

The four ways a training job is split across chips; the links inside a rack, from NVLink to UALink; the networks between racks, InfiniBand against Ethernet, and optics moved into the package; the largest clusters of 2026; and why failures are routine at that scale.

MODULE 11 — Power and Heat

Goal: Understand why power and cooling now set the pace of computing.

Chip power from 700 watts to about 2 kilowatts; racks from about 11 kilowatts to a planned 600; air and liquid cooling, and water; the move to 800-volt direct current; power usage effectiveness; the energy in a prompt; and what a 100-megawatt hall can hold.

MODULE 12 — Software: CUDA and the Moat

Goal: Learn what CUDA is, and judge how far Nvidia's software protects its lead.

What a programming model is; CUDA's kernels, threads, blocks, and memory; the libraries and serving software around it; the alternatives from AMD, Intel, Google, OpenAI, and Apple; independent tests of how far AMD has closed the gap; and whether CUDA is a durable moat.

MODULE 13 — Measuring and Pricing Compute

Goal: Read a benchmark claim critically, and price an hour of computing.

Peak against delivered performance; the benchmarks, from SPEC and Geekbench to MLPerf and the TOP500; how to read a marketing claim; the prices of chips, cloud instances, and rented GPUs; the cost of a million tokens; and how long an AI chip lasts, on the books and in use.

MODULE 14 — The Processor Business: Companies, Markets, and Money

Goal: Map who earns what from processors, and why their margins differ.

The market by segment and who holds it; the latest quarters of Nvidia, Broadcom, Intel, AMD, Qualcomm, and Arm; why their margins differ; customer concentration and the web of investments; China's market; Intel's turnaround under partial government ownership; and whether the AI chip boom is a bubble.

MODULE 15 — Where Things Stand, and What to Watch

Goal: Close with the argument, the contradictions, and a dated watch list.

The course's argument on a page; the contradictions no single module states; what to watch from November 2026 on; and a last word.

A note on the sources: company figures come from the companies' own reports to investors and their filings with the Securities and Exchange Commission (SEC), America's market regulator, each dated to its quarter. Specifications, performance figures, and plans for future products are the makers' own unless the text names an independent test, and the course marks them as claims. Market shares come from Mercury Research for x86 processors, Jon Peddie Research for graphics, and IDC (International Data Corporation), Gartner, and Counterpoint, three research firms, for devices. Estimates of the world's computing power for AI come from Epoch AI, a research group.

Independent measurements come from MLPerf, the benchmarks run by MLCommons, an industry consortium; from the TOP500, a twice-yearly ranking of the fastest supercomputers; from the Uptime Institute's annual survey of data centers; and from SemiAnalysis, a research firm that tests chips and the software that runs on them. Prices for renting a graphics processor come from trackers that measure different things, some transacted prices and some list prices, so the course names the tracker beside each figure. Prices for chips and racks are reported estimates, since Nvidia and AMD publish none.

Fiscal years differ. Nvidia's ends in late January, so its fiscal 2027 is mostly calendar 2026; Broadcom's ends in early November, Marvell's at the start of February, Arm's on March 31, Qualcomm's in late September, and Microsoft's on June 30. Computing speed is counted in floating-point operations per second (FLOPS), arithmetic on numbers with a decimal point: a teraflop is a trillion a second, a petaflop a thousand trillion, and an exaflop a million trillion. The size of the numbers matters as much as the count, since a chip does several times as many operations on 8-bit or 4-bit numbers as on 32-bit ones (Module 5). Amounts of memory are in kilobytes (KB), megabytes (MB), gigabytes (GB), and terabytes (TB), thousands, millions,

billions, and trillions of bytes, and the speeds of memory and links in gigabytes or terabytes a second. Power is in watts (W), kilowatts (kW), and megawatts (MW), and energy in watt-hours (Wh).

The field argues over five questions: whether the instruction set still matters, whether the cloud companies' own chips will displace Nvidia's, whether Nvidia's software is a lasting moat, how long an AI chip lasts, and whether the spending on AI chips is a bubble. On each, the course makes each case as its advocates make it, gives the evidence, and then states its own reading once.

The State of the Art, as of October 2026

The Leading Processors, Who Makes Them, and When They Shipped

This table is the course's reference point: the most advanced processor in each class as of October 30, 2026, who makes it, when it shipped, and the module that explains it. A central processing unit (CPU) is a computer's general-purpose processor, and a graphics processing unit (GPU) is the parallel processor that now runs most AI work. A node is a generation of chipmaking process, named in nanometers (nm), though the names no longer measure anything (Module 1). Dates in the future are the companies' own plans.

Class	Leading product	Maker	Shipped	Module
Phone processors	A20 Pro; Snapdragon 8 Elite Gen 6; Dimensity 9600, all on TSMC's 2nm-class processes; Samsung's Exynos 2600 on its own 2nm	Apple; Qualcomm; MediaTek; Samsung	September 2026 for the first three; early 2026 for the Exynos	6
PC processors	M6, Apple's first 2nm chip; Core Ultra Series 3 (Panther Lake), the first PC chip on Intel's 18A process; RTX Spark, on Arm	Apple; Intel; Nvidia	August 2026, first in the Mac mini; January 2026; systems from autumn 2026	6
Graphics card	GeForce RTX 5090	Nvidia	January 30, 2025; no newer Nvidia gaming GPU by October 2026	5
Game consoles	PlayStation 5 Pro; Nintendo Switch 2	Sony, with AMD's chip; Nintendo, with Nvidia's	November 2024; June 2025	6
Server processors	EPYC Venice; Xeon 6+ (Clearwater Forest)	AMD; Intel	July 2026; mid-2026	7
Cloud company's own CPU	Graviton5	Amazon Web Services	June 2026	7
AI accelerator	Rubin GPU, in the Vera Rubin NVL72 rack with Vera CPUs	Nvidia	Production shipments from August 2026	8
Rival AI accelerator	Instinct MI455X, in the Helios rack	AMD	Launched in the second quarter of 2026; volume due in the second half	8
Inference rack	Groq 3 LPX	Nvidia, on technology licensed from Groq	In full production by August 2026	8
Custom AI chips	TPU 8t and 8i; Trainium3; Maia 200; MTIA 300; Jalapeño	Google; Amazon; Microsoft; Meta; OpenAI, each with partners	TPU 8 announced April 2026 for later that year; Trainium3 December 2025; Maia 200 January 2026; MTIA 300 in 2026; Jalapeño's first chips by September 2026	9
China's AI accelerator	Ascend 950 and the Atlas 950 SuperPoD	Huawei, made by SMIC	950PR in 2026; 950DT and the SuperPoD due in the fourth quarter of 2026	8
Supercomputer	LineShine, a system without accelerators	National Supercomputing Center, Shenzhen	First on the TOP500 list of June 2026	7, 13
Memory for AI chips	HBM4	SK hynix, Samsung, Micron	In production from 2026, in Rubin and the MI455X	3, 8
Link between AI chips	NVLink 6; UALink, the open alternative	Nvidia; the UALink consortium	NVLink 6 in Vera Rubin, 2026; UALink products expected in 2027	10

Class	Leading product	Maker	Shipped	Module
Leading-edge processes	TSMC's N2; Intel's 18A	TSMC; Intel	N2 in volume from late 2025; 18A in Intel's chips from 2026	1, 6, 7

Each module gives the specifications and sources. SMIC is China's leading chipmaker; HBM is the high-bandwidth memory stacked beside AI chips (Module 3).

Who's Who

The Companies, People, and Organizations This Course Names, as of October 30, 2026

Companies

Name	What it does in processors	Based in	Modules
Nvidia	Leading maker of GPUs and AI accelerators, with its own Arm CPUs, networks, and CUDA software	United States	5, 8, 10, 12, 14
Advanced Micro Devices (AMD)	x86 CPUs for PCs and servers, Instinct AI accelerators, Radeon graphics, and console chips	United States	4, 6, 7, 8, 14
Intel	x86 CPUs, AI chips, and chip factories that also make chips for others	United States	4, 6, 7, 8, 14
Arm	Designs and licenses the Arm instruction set and cores; makes its first chip, the AGI CPU	United Kingdom	4, 7, 14
Qualcomm	Snapdragon chips for phones, PCs, and cars; modems	United States	4, 6, 14
Apple	Designs the A-series and M-series chips in its own devices	United States	4, 6
MediaTek	Phone processors; a design partner for Google's TPUs	Taiwan	6, 9
Samsung	Memory, Exynos phone chips, and chip factories	South Korea	3, 6
TSMC (Taiwan Semiconductor Manufacturing Company)	Makes most of the world's leading-edge processors for other companies	Taiwan	1
SK hynix and Micron	Memory makers, with Samsung the three makers of HBM	South Korea; United States	3
Broadcom	Designs custom AI chips for other companies, and network chips	United States	9, 10, 14
Marvell	Designs custom chips and network chips for data centers	United States	9, 14
Alphabet (Google)	Designs TPUs and Axion CPUs; runs Google Cloud	United States	7, 9, 11
Amazon (Amazon Web Services, AWS)	Designs Trainium AI chips and Graviton CPUs through Annapurna Labs	United States	7, 9, 13
Microsoft	Designs Maia AI chips and Cobalt CPUs; runs the Azure cloud	United States	7, 9
Meta	Designs MTIA chips; a large buyer of Nvidia's and AMD's	United States	9, 10
OpenAI and Anthropic	AI labs; large buyers of chips, and in OpenAI's case a designer of its own	United States	8, 9, 14
xAI	AI company founded by Elon Musk; builder of the Colossus clusters	United States	10
Tesla	Designs AI chips for its cars and robots	United States	6, 9
Huawei	Designs Ascend AI chips and Kunpeng CPUs	China	7, 8
SMIC (Semiconductor Manufacturing International Corporation)	China's leading chipmaker, maker of Huawei's chips	China	6, 8
Cambricon, Hygon, Moore Threads, and Loongson	Chinese designers of AI chips, x86 CPUs, GPUs, and CPUs on their own instruction set	China	7, 8, 14
Alibaba (T-Head)	Designs XuanTie RISC-V cores and AI chips	China	4

Name	What it does in processors	Based in	Modules
Cerebras, Groq, Etched, MatX, Positron, SambaNova, Tenstorrent, d-Matrix	Start-ups building AI chips; Cerebras is listed, the rest private	United States; Tenstorrent in Canada	8
Sony, Nintendo, Valve, and Microsoft's Xbox	Console makers, all using AMD's chips except Nintendo, which uses Nvidia's	Japan; United States	6
CoreWeave, Nebius, Oracle, and Crusoe	Cloud companies renting out GPUs	United States; Nebius in the Netherlands	8, 13

People

Name	Who they are	Modules
Jensen Huang	Co-founder and chief executive of Nvidia	5, 8, 14
Colette Kress	Nvidia's chief financial officer	8, 14
Lisa Su	Chief executive of AMD since 2014	6, 8
Lip-Bu Tan	Chief executive of Intel	14
Rene Haas	Chief executive of Arm since 2022	4
Hock Tan	Chief executive of Broadcom	9
Andy Jassy	Chief executive of Amazon	9, 13
Kevin Scott	Microsoft's chief technology officer	9
Thomas Kurian	Chief executive of Google Cloud	9
Elon Musk	Chief executive of Tesla and head of xAI	6, 9, 10
Sam Altman	Chief executive of OpenAI	11
Jonathan Ross	Founder of Groq, who joined Nvidia under its licensing deal	8
Jim Keller	Chief executive of Tenstorrent; earlier a chip architect at AMD, Apple, Intel, and Tesla	4, 8
Shuntaro Furukawa	President of Nintendo	6
Maryellen Noreika	Federal judge in Delaware hearing the cases between Arm and Qualcomm	4
Michael Burry	Investor and critic of the clouds' depreciation of AI chips	13
John von Neumann	Mathematician who described the stored-program computer in 1945	1
Gordon Moore and Robert Dennard	Authors of the forecasts and rules on shrinking transistors	1
Federico Faggin, Ted Hoff, Stanley Mazor, and Masatoshi Shima	Designers of Intel's 4004, the first commercial microprocessor	1
Gene Amdahl	Computer architect whose law limits the gain from parallel processors	2
John Hennessy and David Patterson	Pioneers of reduced instruction sets; Patterson also co-founded RISC-V	2, 4
Sophie Wilson and Steve Furber	Designers of the first Arm processor	4
Krste Asanović	Leader of the Berkeley project that created RISC-V	4
Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton	Authors of AlexNet, trained on Nvidia gaming cards in 2012	5

Terms

The Vocabulary of Processors

Term	Meaning
Transistor	A tiny electronic switch, the basic part of every chip; a processor of 2026 holds billions
Core	One complete processor within a chip, able to run its own stream of instructions; chips hold from a few cores to hundreds
Clock and gigahertz (GHz)	The signal that keeps a chip's circuits in step; a gigahertz is a billion ticks a second
Instruction set architecture (ISA)	The instructions a processor understands, its contract with software; x86, Arm, and RISC-V are the main ones
Microarchitecture	How a chip carries out its instruction set inside, which changes with each generation
Cache	Small, fast memory on the chip that keeps recently used data close to the cores
DRAM	Dynamic random-access memory, a computer's main memory, dense and cheap but slow to answer
High-bandwidth memory (HBM)	DRAM dies stacked and placed beside a processor, delivering terabytes a second; HBM4 is the newest generation
Bandwidth and latency	Bandwidth is how much data moves each second; latency is the delay before the first of it arrives
FLOPS, petaflop, exaflop	Floating-point operations per second; a petaflop is a thousand trillion a second and an exaflop a million trillion
Precision (FP64, FP16, FP8, FP4)	The number of bits used for each number; fewer bits mean more operations a second but coarser numbers
Dense and sparse	Dense figures count every operation; sparse figures assume half the numbers are zero and double the headline rate
Vector unit (SIMD)	Hardware that applies one instruction to many numbers at once
Streaming multiprocessor and warp	A GPU's building block, and the bundle of threads in it that run the same instruction together
Tensor core and matrix engine	Units that multiply small matrices in one step, the source of most AI arithmetic
Systolic array	A grid of multiply-add units through which numbers pulse, the core of Google's TPUs
System on a chip (SoC)	One chip, usually a single die, carrying a device's CPU, GPU, NPU, and other units
Neural processing unit (NPU)	A block for running AI models on little power, in phones and laptops
Chiplet	One of several dies joined in a single package to make a processor
Node	A generation of chipmaking process, labeled in nanometers, such as 3nm or 2nm
Training and inference	Training builds a model by adjusting its parameters; inference runs the finished model to answer requests
Parameter and token	A parameter is one of a model's learned numbers; a token is a word or piece of a word that a model reads or writes
Prefill and decode	The two phases of answering a prompt: reading it all at once, then writing the reply one token at a time
Mixture of experts (MoE)	A model split into many sub-networks, of which only a few are used for each token
Scale-up and scale-out	Links that join chips within a rack, such as NVLink, and networks that join racks into a cluster
InfiniBand and Ethernet	The two families of network used between racks; Ethernet is the ordinary data-center standard
CUDA	Nvidia's programming model and software for its GPUs, unveiled in 2006
Kernel	In GPU programming, a function that thousands of threads run at once

Term	Meaning
Application-specific integrated circuit (ASIC)	A chip built for one job, such as a cloud company's own AI chip
Hyperscaler and neocloud	The largest cloud companies, such as Amazon, Microsoft, and Google; and newer clouds that mainly rent GPUs
Rack	A cabinet holding servers; AI racks draw from about 100 kilowatts upward
Power usage effectiveness (PUE)	A data center's total power divided by the power reaching its computers; 1.0 would be perfect
GPU-hour	One GPU used for one hour, the unit in which computing is rented
Gigawatt (GW)	A billion watts, the unit in which AI data centers and chip deals are now sized
Depreciation	The yearly charge that spreads a machine's cost over its expected life

Module 1 — What a Processor Does

Switches, Instructions, and Fifty Years of Shrinking

Recommended Reading / Reference

Book: Charles Petzold, *Code: The Hidden Language of Computer Hardware and Software* (second edition, 2022), which builds a working computer up from switches, one idea at a time

Reference: Our World in Data, “Moore's law: the number of transistors per microprocessor”, a chart of transistor counts in processors from 1970 to the present

1a. A Machine That Follows Instructions

A processor is a chip that reads instructions from memory and carries them out, one after another, billions of times a second. Each instruction is tiny: add two numbers, copy a number from memory, compare two values, jump to another instruction if the comparison came out a certain way. Everything a computer does, from showing a web page to answering a question put to a chatbot, is built from long sequences of such steps.

The central processing unit (CPU) is the general-purpose processor that runs a computer's operating system and most of its programs. It is built to finish any single task as fast as possible, whatever the task is. The graphics processing unit (GPU) was built to color the millions of dots on a screen, and it works differently: thousands of simple arithmetic units apply the same operation to a great many numbers at once (Module 5). That design turned out to suit the arithmetic of artificial intelligence (AI), and most AI computing now runs on GPUs or on chips built in their image.

Other processors fill narrower roles. A neural processing unit (NPU) in a phone or laptop runs AI models on little power (Module 6). A microcontroller, a small, cheap processor with its own memory on the chip, runs a washing machine, a car's brakes, or a computer mouse, and tens of billions are sold every year. A modern phone carries several kinds on one piece of silicon. Different as they are, all of them rest on the ideas in this module: switches, instructions, a clock, and memory.

1b. Switches, Gates, and Binary

The basic part is the transistor, a switch with no moving parts. In the kind used in processors, a voltage on one terminal, the gate, opens or closes a channel between two others, letting current flow or stopping it. A switch has two states, so a processor works in two symbols, 0 and 1, called bits. Two states are easy to tell apart even when voltages wobble, which is why computers count in binary rather than in tens.

Transistors wired together make logic gates, circuits that combine bits by fixed rules. A NOT gate turns 1 into 0 and 0 into 1. An AND gate gives 1 only if both of its inputs are 1, and an OR gate gives 1 if either is. The NAND gate, an AND followed by a NOT, needs only four transistors in the design used by nearly every chip, and any logic at all can be built from NAND gates alone.

Gates do arithmetic. Adding two single bits takes just two gates: one gives the sum digit and one gives the carry, much as 7 plus 5 gives 2 and carries 1 in ordinary arithmetic. Chain 64 slightly larger stages, each taking in the carry from the one before, and the circuit adds two 64-bit numbers, values up to about 18 billion billion. Real adders use cleverer wiring to get the carries across faster, but the principle holds, and multiplication is built from additions in the same way.

Gates also remember. Two gates whose outputs feed back into each other's inputs hold a bit for as long as power flows, and this loop, a latch, is the root of the fastest memory in a processor. So a chip needs only one kind of part, the transistor, arranged by the billions into circuits that compute and circuits that remember.

1c. The Stored-Program Computer

The design nearly every processor follows was set out in 1945. In a report on the EDVAC, an early American computer, the mathematician John von Neumann described a machine that keeps its program in the same memory as its data. A program then becomes just more numbers, which can be loaded, changed, or replaced without rewiring anything. Earlier machines were often reprogrammed by moving cables and switches by hand.

Such a processor runs a loop. It fetches the next instruction from memory, using a counter that holds that instruction's address. It decodes the instruction to work out what is wanted, executes it, and writes the result back. Then the counter moves on, unless the instruction was a jump that sent it elsewhere. This fetch-decode-execute cycle is the heartbeat of every CPU.

The work happens in a few places. Registers are a handful of storage slots inside the processor, each holding one number, where values sit while they are being worked on. The 64-bit version of x86, the instruction set of most personal computers (PCs) and servers, has 16 general-purpose registers, and Arm's 64-bit design has 31 (Module 4). The arithmetic logic unit does the sums and comparisons, and the control unit steers values between registers, the arithmetic unit, and memory.

A clock keeps everything in step: a signal that ticks at a fixed rate, with each tick letting the circuits settle on new values. Clock rates are counted in gigahertz (GHz), billions of ticks a second. At 5 GHz a tick lasts 0.2 billionths of a second, a fifth of a nanosecond, in which light covers about 6 centimeters. That limit is physical, which is one reason chips are small.

1d. The Instruction Set: The Contract With Software

An instruction set architecture (ISA), or instruction set, is the list of instructions a processor understands, the registers it offers, and the rules for reaching memory. It is a contract: any program written for an instruction set runs on any chip that honors it, whoever designed the chip and however it works inside. The inside, how many instructions it handles at once and how it guesses what comes next, is the microarchitecture, and it changes with every generation while the contract stays put.

Programmers rarely write instructions by hand. They write in languages such as C, Python, or Java, and a compiler, itself a program, translates their code into machine code for a particular instruction set. A short line that adds two values may become an instruction to load the first number into a register, a second load, an add, and a store of the result back to memory.

Because the contract outlives the hardware, software piles up behind it. Intel's 8086 began the x86 line, and programs written for it can still run on an Intel or AMD chip of 2026. That backward compatibility is x86's greatest asset and its heaviest burden, and the instruction set a company picks ties it to a body of software for decades. Module 4 compares the three instruction sets that matter today: x86, Arm, and RISC-V.

1e. From 2,300 Transistors to 336 Billion

The first commercial microprocessor, a whole CPU on one chip, was Intel's 4004, announced on November 15, 1971. It was designed for a Japanese calculator maker, Busicom, by Federico Faggin, Ted Hoff, Stanley Mazor, and Busicom's Masatoshi Shima. It worked on 4 bits at a time; the table below gives its transistor count and clock. Its features were 10 micrometers across, about a tenth of the width of a human hair.

Each later generation packed in more. Intel's 8080 of 1974 powered the first hobbyist computers, and a version of its 8086 went into the Personal Computer of IBM (International Business Machines) in 1981. The x86 line moved to 32 bits with Intel's 386 and to 64 bits with AMD's Opteron and Athlon 64 in 2003. Both companies later began putting two processors, or cores, on one chip, and core counts have climbed ever since (Module 2).

Nvidia's Rubin GPU of 2026 holds about 146 million times as many transistors as the 4004, by Nvidia's count, 55 years on: 27 doublings, one roughly every two years. Few technologies have improved at that pace for that long, and the improvement paid for almost everything else in this course.

Chip	Year	Transistors	Clock	What it ran
Intel 4004	1971	2,300	740 kHz	A calculator
Intel 8086	1978	29,000	5-10 MHz	The first x86 programs; a variant, the 8088, ran IBM's PC
Intel 386	1985	275,000	12-33 MHz	32-bit PCs
Intel Pentium 4	2000-04	42-125 million	Up to 3.8 GHz	The peak of the race for clock speed
Nvidia Rubin	2026	336 billion	Not published	AI training and inference, in racks of 72

Transistor counts and clocks are the makers' figures; the Pentium 4's range runs from its first design to its 2004 version. kHz is thousands of ticks a second, MHz millions, and GHz billions.

1f. Moore's Law and the End of Free Speed-Ups

In 1965 Gordon Moore, then at Fairchild Semiconductor and later a co-founder of Intel, noticed that the number of components that could be put on a chip at the lowest cost per component was doubling every year. In 1975 he revised the pace to every two years. The forecast became known as Moore's law, and the industry treated it as a schedule. Its pace has slowed, and each new process now costs more to develop, but the counts are still rising.

Shrinking transistors also made them faster and thriftier, which mattered as much as the count. In 1974 Robert Dennard of IBM and his colleagues showed that if every dimension of a transistor shrinks by a factor, its voltage and current shrink with it. The power drawn by each square millimeter then stays about the same while the transistors switch faster. This rule, Dennard scaling, meant that each generation ran faster on the same power, and clock speeds rose from a few megahertz to several gigahertz.

Dennard scaling broke down around 2005. Below a certain size, transistors leak current even when they are switched off, and the supply voltage could not fall much below about one volt without making the switches unreliable. Power density began to climb with each shrink, and a chip running much faster would have melted. Intel's Pentium 4 reached its top clock in 2004, and Intel cancelled its planned successor that year and turned to chips with two cores instead.

Since then, extra transistors have gone into doing more things at once rather than doing one thing faster. Their main uses have been more cores, wider arithmetic units, larger caches (Module 3), and circuits built

for one kind of work, such as graphics, video, or AI. That shift from making one stream of instructions faster to running many streams, or specializing, is the turn on which this course rests.

1g. How Chips Are Made, in Brief

A chip is built on a wafer, a polished disc of pure silicon 300 millimeters across. Machines print patterns onto it with light, a process called lithography, then deposit, etch, and dope the silicon layer by layer to make transistors. A dozen or more layers of copper wiring are laid on top to connect them. The finest patterns are printed with extreme-ultraviolet light, whose wavelength is 13.5 nanometers, from machines that only ASML of the Netherlands makes.

A leading-edge wafer passes through more than a thousand steps over about three to four months. It is then cut into dies, the individual chips, which are tested and packaged. The share that works, the yield, decides much of the cost, and a large die is likelier to contain a flaw than a small one. That is why the biggest chips are increasingly split into several smaller dies, or chiplets, joined in one package (Module 2).

Generations of manufacturing process are named in nanometers, such as 5nm, 3nm, and 2nm, but the names are labels for generations, not measurements: nothing on a 3nm chip is 3 nanometers long. The Taiwan Semiconductor Manufacturing Company (TSMC) makes most of the world's leading-edge processors, for Apple, Nvidia, AMD, Qualcomm, and many others. Intel makes most of its own and sells space in its factories to others, and Samsung makes its own chips and other companies'. This course takes the factories as given and looks at the processors they make.

1h. The Course's Thread

The course follows what came after Dennard scaling ended (1f). Speed has since come from parallel and specialized chips, and the money has followed: the GPU's leading maker, Nvidia, became the world's most valuable company (Module 14). Its largest customers now design chips of their own to replace it (Module 9).

The limits have moved too, to memory (Module 3), the networks between chips (Module 10), power and cooling (Module 11), and software (Module 12). The course asks how far specialization will go, and whether today's leaders hold moats or only leads. Modules 1 to 5 explain how processors work, Modules 6 to 9 cover the chips and their makers, and Modules 10 to 14 cover the systems, the software, and the money.

Worked example — A CPU Adding Up 1,000 Numbers

Take a list of 1,000 numbers in memory and a program that adds them up. On a simple processor the loop takes five instructions per number: load the number into a register, add it to the running total, move the address on to the next number, count down how many are left, and jump back if any remain. That is 5,000 instructions. A simple processor that finishes one instruction per tick, with a 1 GHz clock, needs about 5,000 ticks, or 5 millionths of a second (microseconds). Now suppose each load must wait for the computer's main memory, which takes about 100 nanoseconds (Module 3), or 100 ticks at 1 GHz. The 1,000 loads then cost about 100,000 ticks, and the job takes about 100 microseconds, twenty times as long, almost all of it spent waiting. Module 3 explains how caches, small fast memories near the processor, remove most of that wait. A modern core at 5 GHz goes further still. Its vector unit adds 16 numbers in one instruction (Module 2), so the loop runs about 63 times instead of 1,000, and the core completes about four instructions per tick. The whole sum then takes roughly 80 ticks, about 16 nanoseconds, some 300 times faster than the simple processor. These are illustrative figures, but the lesson is real: the clock explains only a small part of the speed-up, and most of the rest comes from parallel work and from keeping the arithmetic fed with data.

Module 1 Key Points

What a Processor Does

• A processor fetches, decodes, and executes instructions; at 5 GHz its clock ticks every fifth of a nanosecond.

• Every circuit is built from transistor switches; a NAND gate takes four, and any logic can be built from NAND gates.

• The instruction set is the contract with software: programs written for Intel's 8086 of 1978 can still run today.

• Transistor counts rose from 2,300 in Intel's 4004 (1971) to 336 billion in Nvidia's Rubin (2026), doubling about every two years.

• Dennard scaling ended about 2005: shrinks no longer held power density steady, so clocks stalled and chips turned parallel and specialized.

Module 1 Quiz

Check Your Understanding

1. What does a processor's instruction set define?

A) The chip's clock speed
B) The instructions software can use
C) Its transistor count
D) The factory process that makes the chip

2. Why did clock speeds stop rising quickly after about 2005?

A) Factories could no longer make smaller transistors at all
B) Software could not keep up
C) Power and heat rose too fast
D) Customers preferred cheaper chips with fewer features, so the makers stopped trying to raise clock speeds

3. What did Dennard scaling describe?

A) Transistor counts doubling every year, as Gordon Moore first predicted in his article of 1965.
B) New chips running all of the old software without changes, because the instruction set never changes.
C) Factories cutting the price of each wafer in half with every new generation of process.
D) Smaller transistors keeping power per area constant while switching faster.

4. In the worked example, why does the simple processor take about 100 microseconds rather than 5 when it has no cache?

A) Each of the 1,000 loads waits about 100 ticks for main memory, so the waiting dwarfs the 5,000 ticks of real work.
B) The processor must decode each instruction twice when there is no cache to hold it.
C) Without a cache the adder can handle only one bit at a time.
D) Its clock slows to a tenth of its speed to save power.

5. About how often have transistor counts per processor doubled since Intel's 4004?

A) About every six months, faster than any other technology.
B) About every ten years, as the factories slowed.
C) About every five years.
D) About every two years, close to Moore's 1975 revision.

Module 2 — Making One Core Fast

Pipelines, Prediction, and the Power Wall

Recommended Reading / Reference

Book: John L. Hennessy and David A. Patterson, *Computer Architecture: A Quantitative Approach* (sixth edition, 2017), the standard text on how processors are designed and measured, by two of the pioneers of the reduced instruction set computer (RISC)

Reference: Agner Fog, *The Microarchitecture of Intel, AMD and VIA CPUs*, an independent guide to how real cores work, updated as new chips appear

2a. The Performance Equation

How long a program takes on a processor comes down to three numbers multiplied together. The first is how many instructions the program needs. The second is how many clock ticks, or cycles, each instruction takes on average. The third is the length of each tick, the inverse of the clock rate. Engineers write it as time equals instructions times cycles per instruction, divided by the clock rate.

Each term is a lever. The instruction count falls with better compilers and with instructions that do more at once, such as the vector instructions in 2e. The cycles per instruction fall when a core finishes several instructions in each tick; its inverse, instructions per cycle (IPC), is the figure chip makers quote when they claim a new core is faster. The clock rate rises with faster transistors and shorter circuits.

A worked line shows the scale. A program of 10 billion instructions, on a core that averages four instructions per cycle at 5 GHz, takes half a second. Until Dennard scaling broke down, the clock did much of the work, rising from 4.77 megahertz in IBM's first PC to the Pentium 4's peak, about 800 times (Module 1). Since then clocks have crept up by about two and a half gigahertz, to 6.2 GHz at most (2g), and almost all the gain has come from the other two levers and from adding cores (2h).

2b. Pipelines

A pipeline works like an assembly line. Running one instruction involves several steps: fetch it, decode it, execute it, reach memory if it needs to, and write the result back. Instead of finishing one instruction before starting the next, a pipelined core starts a new instruction every cycle, so that while one is being executed the next is being decoded and the one after is being fetched. A five-stage pipeline holds five instructions at once, each at a different stage, and can finish one per cycle even though each takes five cycles from start to end.

Splitting the work into more, shorter stages lets the clock run faster, because each stage has less to do in a tick. Intel pushed this hardest: its Pentium 4 of 2000 had a pipeline of 20 stages, and the 2004 version 31. The cost of a deep pipeline appears when the flow breaks, because an instruction may need a result that an earlier one has not yet produced. The core must then stall or pass the result straight along from one stage to the next, a trick called forwarding.

2c. Superscalar and Out-of-Order Execution

A superscalar core runs several pipelines side by side and starts several instructions in each cycle. The widest cores of 2026 decode and dispatch eight or more instructions a cycle and have more than a dozen execution units: several for integer arithmetic, several for floating-point and vector math, and several for loads and stores to memory. The hard part is keeping them all busy, since programs are written as one long sequence in which many instructions depend on the ones just before them.

Out-of-order execution solves this by looking ahead. The core reads far down the instruction stream, holding the instructions in a window, and executes each one as soon as its inputs are ready, whatever its place in the program. Intel's Lion Cove core of 2024 tracks up to 576 instructions in flight and AMD's Zen 5 up to 448, by the companies' figures. When a slow load is waiting on memory, the core runs hundreds of independent instructions around it.

The program must still see results exactly as if each instruction ran in order. So the core renames registers behind the scenes, giving each value its own physical slot among hundreds, which stops unrelated instructions from tripping over the same register name. A structure called the reorder buffer then retires the instructions in program order, making their results final one by one. Out-of-order designs have been standard in fast cores since the mid-1990s, and the windows have grown with almost every generation since.

2d. Branch Prediction and Speculation

About one instruction in every five or six in ordinary code is a branch, a decision to jump elsewhere or carry on. A deep, wide core cannot wait to learn which way each branch goes, because by the time the answer is known it would have fetched dozens of instructions down one path. So it guesses. Branch predictors keep a history of how each branch behaved and of the patterns among branches, and modern ones guess right well over 95% of the time on typical programs.

The core then runs ahead on the predicted path, a technique called speculative execution. If the guess was right, the work is kept. If it was wrong, the core throws away everything after the branch and starts again on the correct path, losing roughly 15 to 20 cycles. With hundreds of instructions in flight, even a few percent of wrong guesses costs a noticeable share of performance, so designers spend a large part of each core on prediction.

Speculation turned out to have a flaw. On January 3, 2018, researchers from Google's Project Zero team and several universities disclosed two attacks, Spectre and Meltdown. Work done ahead of time, on a wrongly guessed path or, in Meltdown's case, before a check on whether the memory may be read, is discarded, but it leaves traces, such as data pulled into a cache, and a program can detect those traces by timing how quickly memory responds. That lets one program read secrets that should be off limits, such as another program's data or the operating system's.

The fixes were patches to operating systems and browsers, then changes to chip designs. They slowed some kinds of work, notably programs that call on the operating system often, and new variants of the attack have been found in most years since, in chips from every major designer. The episode showed that the tricks that make one core fast also make it hard to secure, and the trade between speed and isolation is now part of every design.

2e. Vector Units

A vector instruction applies one operation to a row of numbers at once, an approach called single instruction, multiple data (SIMD). A 512-bit vector register holds sixteen 32-bit numbers, so one vector add does sixteen additions. Most such units can also fuse a multiplication and an addition into one step, a multiply-accumulate that counts as two operations. A core with two such units does 64 floating-point operations each cycle, and at 5 GHz that is about 320 billion a second from a single core.

The x86 instruction set added vector instructions in widening steps: 128 bits with the Streaming SIMD Extensions (SSE) in 1999, 256 bits with the Advanced Vector Extensions (AVX) in 2011, and 512 bits with AVX-512 from 2016. Intel and AMD have since agreed to converge on AVX10 (Module 4). Arm's equivalents are NEON, 128 bits wide, and the Scalable Vector Extension (SVE), whose width can vary from chip to chip. Vector units power video, image editing, scientific code, and much of the AI inference that runs on CPUs.

CPUs have recently gained matrix engines too, units that multiply small blocks of numbers in one operation. Intel's Advanced Matrix Extensions (AMX) arrived in its Xeon server chips of January 2023, and Arm's Scalable Matrix Extension (SME) appears in Apple's M4 and later chips. They narrow but do not close the gap with GPUs on AI work (Module 5).

2f. Multithreading

Even a wide out-of-order core leaves units idle, for instance while it waits on memory. Simultaneous multithreading (SMT) lets one core run two streams of instructions, called threads, at once, drawing on the same execution units, so that one thread's instructions fill gaps the other leaves. Intel introduced it as Hyper-Threading in 2002. Typical gains run from about 10% to 30% more work per core, depending on the program, at a small cost in chip area.

The trade-off is contested. Two threads sharing a core share its caches and predictors, a channel for the leaks in 2d, and a thread gets less of the core than it would alone. Intel dropped Hyper-Threading from its Lunar Lake and Arrow Lake PC chips of 2024, and its server chips built of small cores never had it, while AMD keeps two threads per core in its server chips. Nvidia's Vera server processor also runs two threads on each of its cores (Module 7).

2g. The Power Wall

The power a chip burns as its transistors switch rises with three things: the electrical load being switched, the square of the voltage, and the clock frequency. Engineers write it as capacitance times voltage squared times frequency. Running faster usually needs a higher voltage too, so the cost compounds. A 20% rise in clock speed that needs 10% more voltage raises power by about 45%, because 1.1 squared times 1.2 is about 1.45.

That arithmetic is why the last few hundred megahertz are so expensive. Desktop chips of 2026 boost to between 5 and 6 GHz for short bursts, and the fastest clock on a standard chip, Intel's Core i9-14900KS of March 2024, is 6.2 GHz. Such chips can draw about 250 watts, which a large cooler must carry away from a piece of silicon the size of a fingernail. Phones get by on a few watts, so their cores reach top speed only in short bursts and lean on efficiency.

The wall explains the shape of modern chips. A design that hits its power limit gains nothing from more transistors running at full speed, so the extra transistors are spent on cores that run slower and more

efficiently, on circuits that switch off when idle, and on specialized units that do one job on far less energy than a general-purpose core. Power, not transistor count, became the binding limit for most designs once Dennard scaling ended (Module 1), and Module 11 follows it to the data center.

2h. Many Cores, Big and Small

When the clock stalled, chip makers began spending transistors on more cores. AMD and Intel both shipped two-core PC chips in 2005, and core counts have doubled every few years since. In 2026 a desktop chip has up to 24 cores, and the largest server chips have hundreds (Module 7). More cores help only if a program can be split into parts that run at the same time, and that is where the limits lie.

In 1967 Gene Amdahl, an architect of IBM's mainframes, pointed out that the part of a job that cannot be split sets a ceiling on the gain. If 5% of a program must run in sequence, then even infinitely many cores leave that 5%, so the program can never run more than 20 times faster. The rule, Amdahl's law, explains why single-core speed still matters even in a many-core chip. In 1988 John Gustafson added the counterpoint: people given more cores tend to tackle bigger problems, whose parallel part grows while the serial part stays small.

Designers have answered with mixed cores. Arm introduced its big.LITTLE scheme in 2011, pairing fast, power-hungry cores with small, efficient ones, and nearly every phone chip now uses some version of it. Apple brought it to its own chips in 2016 and Intel to its PC chips in 2021. Intel's Panther Lake of 2026, for example, combines 4 performance cores, 8 efficient cores, and 4 low-power ones (Module 6). The operating system steers demanding work to the big cores and background work to the small ones.

The cores themselves are increasingly built as chiplets. AMD's Zen 2 chips of 2019 split a processor into several small dies of cores, made on the newest process, and one larger die for the connections to memory and the rest of the system, made on an older and cheaper one. Small dies yield better (Module 1), and each part can use the process that suits it. Intel followed with chips built of tiles from 2023, and the largest AI chips are chiplet designs too (Module 8). The cost is the energy and delay of moving data between dies.

Worked example — What 8, 64, and 1,000 Cores Buy

Amdahl's law gives the speed-up as 1 divided by the sum of two parts: the share of the work that is serial, plus the parallel share divided by the number of cores. Take a program that is 95% parallel. On 8 cores the speed-up is 1 / (0.05 + 0.95/8) = 1 / 0.169, about 5.9 times. On 64 cores it is 1 / (0.05 + 0.95/64) = 1 / 0.0648, about 15.4 times. On 1,000 cores it is 1 / (0.05 + 0.00095), about 19.6 times, almost at the ceiling of 20. The last 936 cores lifted the speed-up only from 15.4 to 19.6, so most of them sit idle waiting for the serial 5%. Now take a program that is 99.9% parallel, as the arithmetic of graphics and neural networks nearly is. On 1,000 cores the speed-up is 1 / (0.001 + 0.000999), about 500 times, and on 10,000 cores about 909 times. That difference is the whole case for the GPU (Module 5): a design with thousands of simple cores pays off only for work that is almost entirely parallel, and the work that most needs computing power today is exactly that kind.

Module 2 Key Points

Making One Core Fast

• Run time is instructions times cycles per instruction over clock rate; since 2005, gains have come from more work per cycle and more cores.

• Out-of-order cores keep hundreds of instructions in flight, 576 in Intel's Lion Cove, and retire them in program order.

• Branch predictors guess right more than 95% of the time, but Spectre and Meltdown showed that speculation can leak secrets.

• Power rises with voltage squared times frequency, so the last few hundred megahertz cost the most; standard chips top out at 6.2 GHz.

• Amdahl's law: a program that is 95% parallel can never run more than 20 times faster, however many cores it gets.

Module 2 Quiz

Check Your Understanding

1. Why does a core execute instructions out of order but retire them in order?

A) To save power by switching off the units that wait.

B) Because memory can accept results only in the order they were requested.

C) So it can run whichever instruction has its inputs ready, while the program still sees the results exactly as if each had run in sequence.

D) To keep two programs on the same core from seeing each other's data.

2. What does a chip's switching power rise with?

A) Voltage squared times frequency

B) The number of registers it has, squared

C) Its transistor count divided by its clock

D) The square root of its clock speed

3. By Amdahl's law, about how much faster does a program that is 95% parallel run on 64 cores?

A) Exactly 64 times, one for each core

B) About 15 times

C) About 3 times

D) About 61 times, since only the serial 5% is lost and the other 95% scales perfectly with the number of cores

4. What made the Spectre and Meltdown attacks possible?

A) Compilers produced faulty instructions that let programs run each other's code without permission.

B) Work done on a wrongly guessed path left traces in caches that timing could reveal.

C) The chips' clocks ran faster than their designers had certified.

D) Operating systems stored passwords in registers that any program could read.

5. Why have AMD and Intel built processors from chiplets?

A) To raise clock speeds above 6 GHz

B) Because chiplets need no software

C) To use a different instruction set in each part

D) Small dies yield better, and each part can use the manufacturing process that suits it, which lowers the cost

Module 3 — Memory: The Wall Every Processor Hits

Caches, Main Memory, High-Bandwidth Memory, and the Shortage of 2026

Recommended Reading / Reference

Book: Bruce Jacob, Spencer W. Ng, and David T. Wang, *Memory Systems: Cache, DRAM, Disk* (2008), the standard reference on how computer memory is built and why it behaves as it does

Reference: TrendForce, the price and market reports on memory chips of its DRAMeXchange service, the most widely cited tracker of memory prices

3a. The Ladder of Memory

A processor can do arithmetic far faster than it can fetch the numbers to do it on. The fix is a ladder of memories, each larger, cheaper, and slower than the one above it. At the top sit the registers, a few hundred bytes reached instantly. Below them come caches built into the chip from static random-access memory (SRAM), then the main memory on separate chips of dynamic random-access memory (DRAM), and at the bottom the storage drives that keep data when the power is off.

The table gives typical sizes and delays in 2026, and the last column scales time so that one clock tick at 5 GHz lasts one second. Nothing else in a computer explains as much about its speed as that ladder.

Level	Typical size	Typical delay	If one tick were one second	Built from
Registers	Hundreds of bytes per core	Within a tick	Under a second	Latches in the core
Level 1 (L1) cache	32-64 KB per core, for data	About 1 nanosecond	About 5 seconds	SRAM in the core
Level 2 (L2) cache	1-3 MB per core	3-4 nanoseconds	About 15-20 seconds	SRAM beside the core
Level 3 (L3) cache	Tens of MB, shared; up to about 1 GB in AMD's largest server chip	10-20 nanoseconds	About 50-100 seconds	SRAM on or stacked on the chip
Main memory	Gigabytes to terabytes	About 100 nanoseconds	About 8 minutes	DRAM chips
Solid-state drive	Terabytes	50-100 microseconds	About 3-6 days	Flash memory
Hard disk	Tens of terabytes	5-10 milliseconds	About 10 months to 1.5 years	Spinning magnetic platters

Typical figures for a 2026 desktop or server core; real chips vary. A nanosecond is a billionth of a second and a microsecond a millionth; a tick is one cycle at 5 GHz (Module 1).

3b. Caches and Locality

Caches work because programs are predictable in how they use memory. A value used once is likely to be used again soon, which engineers call temporal locality, and a value's neighbors are likely to be used next, which they call spatial locality. A loop runs the same instructions again and again, and a program reading a list reads its entries in order. A cache keeps recently used data close, so most requests never reach main memory.

SRAM stores each bit in a loop of six transistors, which is fast but bulky. Caches move data in blocks called lines, usually 64 bytes, so a request for one number brings its neighbors along. A well-behaved program finds what it needs in the first-level cache well over 90% of the time. Prefetchers, circuits that spot patterns such as a program walking through a list, fetch lines before they are asked for.

Because a miss is so costly, caches take up a large share of a modern chip, often more area than the cores' arithmetic. AMD stacks an extra layer of cache on some of its chips, a design it calls 3D V-Cache; in its Ryzen 7 9800X3D of 2024 the cache sits beneath the cores, giving 96 MB of third-level cache. Games, which jump around large amounts of data, run markedly faster on such chips.

3c. Why Memory's Delay Barely Improved

DRAM stores each bit as a charge in a tiny capacitor beside a single transistor. That design is dense and cheap, so a server can hold terabytes of it. But the charge leaks away, so every bit must be read and rewritten many times a second, and reading a bit means sensing a minute charge on a long wire. The time to reach a random location, tens of nanoseconds inside the DRAM chip itself, has hardly changed since the late 1990s.

What did change is bandwidth, the amount of data delivered each second once a transfer is under way. Each generation of DRAM moves more bits per tick down more wires at once. A single channel of DDR5-6400, a common kind in 2026, transfers 6.4 billion times a second, eight bytes at a time, so it delivers 51.2 GB a second. A large server chip has 12 or 16 such channels.

The gap between processor speed and memory delay grew for decades. In 1995 the computer scientists William Wulf and Sally McKee named the problem the memory wall, warning that processors would come to spend most of their time waiting. Out-of-order execution (Module 2), caches, and prefetchers are all ways of hiding that wait. The trend toward parallel chips has made the wall higher, because a thousand cores need a thousand times the data.

3d. Kinds of DRAM

The same basic DRAM cell is packaged four ways for different jobs. DDR5, the fifth generation of double-data-rate memory, sits on removable modules in desktops and servers and balances capacity, speed, and price. LPDDR, the low-power version, is soldered beside the processor in phones and laptops; in its current generation, LPDDR5X, it is the memory of Apple's chips, which share one pool between the CPU and GPU, and the M5 of October 2025 reaches 153 GB a second. GDDR7, the graphics version, is built for speed on a wide bus, and Nvidia's RTX 5090 graphics card carries 32 GB of it at 1,792 GB a second.

High-bandwidth memory (HBM) is the fourth and newest form. It stacks up to a dozen or more DRAM dies on top of one another, connected by thousands of vertical wires through the silicon, and places each stack right beside the processor on a shared base of silicon. A very wide connection, 1,024 wires per stack in the HBM3E generation and twice that in HBM4, the generation that entered production in 2026, lets each stack deliver more than a terabyte a second. HBM first appeared in a graphics card in 2015.

The AI chips of 2026 carry from four to twelve stacks each, with totals given in Modules 8 and 9. HBM costs several times as much per gigabyte as ordinary DRAM and is hard to make, since one bad die can spoil a whole stack. Three companies make it: SK hynix of South Korea, the leader, Samsung, and Micron of the United States. The amount of HBM a maker can supply limits how many AI chips can be built.

3e. Capacity Against Bandwidth

Memory sets two separate limits. Capacity decides what fits: a model of 70 billion parameters, the learned numbers in a neural network, takes 140 GB stored at 16 bits each, 70 GB at 8 bits, and 35 GB at 4 bits. Bandwidth decides how fast what fits can be read. A design can be rich in one and poor in the other, and much of the variety among AI chips comes from where each sits on that trade-off.

HBM gives the most bandwidth but limited capacity at a high price. LPDDR gives large capacity cheaply but less bandwidth: Nvidia pairs each Vera CPU with up to 1.5 TB of it at 1.2 TB a second, by Nvidia's figures, as a store that its GPUs can reach (Module 8). Laptops and small workstations with large pools of LPDDR, such as Nvidia's RTX Spark, can hold big models that no gaming card can, but they read them more slowly (Module 6).

3f. The Roofline

Whether a job is limited by arithmetic or by memory depends on its arithmetic intensity: the number of operations it performs for each byte it reads from memory. The roofline model, published in 2009 by Samuel Williams, Andrew Waterman, and David Patterson, turns this into a simple rule. A chip's speed on a job is the lower of two ceilings: its peak arithmetic rate, or its memory bandwidth times the job's intensity.

The crossover point, where the two ceilings meet, is the peak arithmetic rate divided by the bandwidth. For Nvidia's H100, about 2 petaflops of 8-bit arithmetic over 3.35 TB a second puts the crossover near 600 operations per byte, by Nvidia's figures (Module 8). A job doing fewer operations than that per byte is limited by memory, however fast the arithmetic units are. Large matrix multiplications, which reuse each number many times, sit above the line; scanning a database, or generating text one word at a time, sits far below it.

The roofline explains why chip designers chase bandwidth as hard as arithmetic. Each new GPU's arithmetic has grown faster than its memory bandwidth, so the crossover keeps rising and more jobs fall below it. Nvidia's Rubin raised bandwidth by 2.8 times over Blackwell, by Nvidia's figures, partly to keep up. Much of the clever software of recent years, such as fusing several steps so that data is read once, exists to raise arithmetic intensity (Module 12).

3g. The Memory Shortage of 2026

In 2026 the demand for AI chips collided with the supply of memory. Each AI accelerator needs hundreds of gigabytes of HBM, and the servers around it need terabytes of ordinary DRAM. The memory makers shifted their factories toward HBM and server memory, which pay best, and every stack of HBM takes several times the wafer area of the same capacity in ordinary DRAM. Micron, one of the three makers, chose to leave consumer memory altogether: it announced in December 2025 that it would stop selling memory under its Crucial brand, and shipments ended in February 2026.

Prices for everyone else soared. Gartner, a research firm, estimated that the cost of DRAM rose about fivefold in a year to the second quarter of 2026. IDC counted memory prices up as much as 500%, with a 128 GB kit of DDR5 for a desktop costing up to $3,399, by its count of October 2026. Module 6 follows the effects on phones, PCs, graphics cards, and consoles, whose makers raised prices or shipped fewer.

How long the squeeze lasts is a matter of forecasts. Analysts at Gartner and at chip companies including Intel do not expect memory prices to ease before late 2027 or 2028, because new factories take years to build. Acer's chief executive, Jason Chen, argued in October 2026 that the shortage forecasts were

overblown and that PC prices would fall by the second half of 2027, as new capacity in China comes online. The memory makers, for their part, are earning record profits and building as fast as they can.

Worked example — Why a Chatbot's Speed Is Set by Memory

When a language model writes a reply, it produces one token, a word or piece of a word, at a time, and for each token it must read every one of its parameters from memory. For each parameter it does just one multiply and one add, two operations per parameter read, far below the roofline's crossover. So a single user's speed is capped by bandwidth divided by the bytes read per token. Take a laptop with an Apple M5, at the 153 GB a second given in 3d, running a model of 8 billion parameters stored at 4 bits, half a byte each: 4 GB per token, so at most about 38 tokens a second. A Blackwell B200 GPU, at 8 TB a second (Module 8), running a 70-billion-parameter model at 8 bits reads 70 GB per token, so at most about 114 tokens a second. A Rubin GPU, at the 22 TB a second given in Module 8, would allow about 314. Real systems reach perhaps 60-80% of these ceilings. Two tricks change the picture. A service that answers 64 users at once reads the parameters once per step for all of them, so its total output rises many times over, which is why AI services batch requests. The gain falls short of 64-fold, because each user's own record of the conversation so far, kept in memory, must still be read for every token. And a mixture-of-experts model reads only part of itself for each token: DeepSeek-V3 has 671 billion parameters, but only 37 billion are used for each token, so it reads about 37 GB per token at 8 bits.

Module 3 Key Points

Memory: The Wall Every Processor Hits

• Main memory takes about 100 nanoseconds to answer, some 500 ticks of a 5 GHz clock, so caches must catch most requests.

• Memory's delay barely improved for decades while its bandwidth multiplied; one DDR5-6400 channel delivers 51.2 GB a second.

• High-bandwidth memory stacks DRAM dies beside the processor; Rubin's eight HBM4 stacks deliver up to 22 TB a second.

• A chatbot's speed for one user is capped at memory bandwidth divided by the bytes read for each token.

• AI's demand for memory raised DRAM costs about fivefold in the year to mid-2026, by Gartner's estimate.

Module 3 Quiz

Check Your Understanding

1. Which kind of memory sits in stacks right beside AI chips?

A) DDR5 modules
B) Flash chips in solid-state drives
C) High-bandwidth memory
D) Graphics memory soldered on the board

2. In the roofline model, what limits a job with low arithmetic intensity?

A) Memory bandwidth
B) The chip's peak arithmetic rate
C) The number of instructions in its instruction set
D) The clock speed of its slowest core

3. Why is generating text one token at a time limited by memory rather than arithmetic?

A) Because models are stored on disks that spin too slowly.
B) Because each token needs more multiplications than the chip can do.
C) Because every parameter must be read from memory for each token, while the arithmetic done per byte read is tiny.
D) Because caches erase themselves after each token.

4. What does the course give as the main cause of the memory shortage of 2026?

A) A fire at the largest memory factory in South Korea halted its output for most of the year.
B) New tariffs on memory chips imported into America and Europe cut the supply sharply.
C) Phone makers hoarded a year's supply.
D) AI data centers absorbed the supply as makers shifted output to HBM.

5. In the worked example, about how many tokens a second can one user get from a 70-billion-parameter model at 8 bits on a Rubin GPU?

A) About 314, an upper bound
B) About 2
C) About 31
D) About 3,140, since Rubin's arithmetic is so fast

Module 4 — Instruction Sets: x86, Arm, and RISC-V

Contracts, Licenses, and Royalties

Recommended Reading / Reference

Book: James Ashton, *The Everything Blueprint: The Microchip Design That Changed the World* (2023), a history of Arm from a Cambridge start-up to the designs inside nearly every phone

Reference: Mercury Research, quarterly estimates of x86 processor market share, reported each quarter by the trade press

4a. Complex and Reduced Instruction Sets

Instruction sets come in two broad styles. A complex instruction set computer (CISC), such as x86, offers many instructions, some of which do several things at once, such as reading a number from memory, adding to it, and writing it back. Its instructions vary in length, from 1 to 15 bytes in x86. A reduced instruction set computer (RISC) offers simpler instructions of a fixed length, usually 4 bytes, and only two kinds of instruction touch memory: a load, which copies a number into a register, and a store, which copies one back.

RISC came out of university research in the early 1980s. John Hennessy's group at Stanford and David Patterson's at the University of California, Berkeley, argued that simple instructions could be pipelined (Module 2) and run much faster, and that compilers could do the work that complex instructions had done. Their designs led to the MIPS and SPARC chips, and the idea runs through Arm and RISC-V today. Hennessy and Patterson shared computing's highest prize, the Turing Award, for 2017.

The argument between the two styles was largely settled by a compromise. From the Pentium Pro of 1995, Intel's x86 chips have translated each complex instruction, as it arrives, into one or more simple internal steps called micro-operations, which an out-of-order engine much like a RISC core then executes. AMD's chips do the same. The front of an x86 core still pays a price in power and complexity for decoding instructions of uneven length, and the debate over how much that matters is taken up in 4f.

4b. x86: A Duopoly by License

The x86 instruction set began with Intel's 8086 (Module 1). When IBM chose an Intel chip for its PC, it insisted on a second supplier, and in 1982 Intel licensed AMD to make x86 chips. Years of lawsuits over the license followed, settled in 2009 when Intel paid AMD $1.25 billion and the two signed a new cross-license. AMD, for its part, designed the 64-bit extension that every x86 chip now uses (Module 1), and Intel adopted it in 2004.

The result is a duopoly that no newcomer can easily join. Each company needs the other's patents to build a modern x86 chip, and the licenses are not for sale. Two Chinese firms make x86-compatible chips under older arrangements: Zhaoxin, through a joint venture with Taiwan's VIA Technologies, and Hygon, through a joint venture with AMD (Module 7). Everyone else who wants a custom processor turns to Arm or RISC-V.

Rivals for decades, Intel and AMD now coordinate. In October 2024 they formed the x86 Ecosystem Advisory Group with Microsoft, Google, Meta, Dell, HP, Lenovo, and others, to keep the instruction set consistent as both companies add to it. A year on, in October 2025, the group announced agreement on four

shared features. They were a faster way of handling interrupts, called FRED (Flexible Return and Event Delivery); AVX10, a common set of vector instructions; ACE, a standard form of matrix acceleration; and ChkTag, memory tagging to catch a common class of software bugs, whose full specification was still to come.

Within the duopoly, the balance has shifted toward AMD. Mercury Research, which tracks the market, gave AMD a record share, shown in the table, of all x86 processors shipped in the second quarter of 2026, up from about 29% a year earlier, and a larger share still of the money, because AMD sells more of the costliest server chips. Arm-based processors, led by Apple's, reached a record 15.3% of PC processors in the same quarter, by Mercury's count as reported by ComputerBase, a German trade site. Modules 6 and 7 take up the PC and server markets in turn.

Segment	AMD's share of units, second quarter of 2026	Intel's share of units, second quarter of 2026	AMD's share of revenue, first quarter of 2026
All x86 processors	34.1%	65.9%	38.1%
Servers	34.5%	65.5%	46.2%
All PCs	30.3%	69.7%	31.4%
Desktops	34.9%	65.1%	37.6%
Laptops	28.9%	71.1%	28.9%

Mercury Research estimates, as reported by HotHardware (August 21, 2026), Basic Tutorials (August 24, 2026), and Tom's Hardware (May 14, 2026). The revenue split for the second quarter had not been reported when this course was written. Mercury counts x86 processors only, so Arm-based chips are left out.

4c. Arm: Licenses and Royalties

Arm began as a processor for Acorn, a British computer maker, whose first chip, the ARM1, was designed by Sophie Wilson and Steve Furber and worked first time in 1985. In 1990 Acorn, Apple, and VLSI Technology spun the design team out as Advanced RISC Machines, based near Cambridge, England. Its chips used little power, which suited mobile phones, and Arm designs came to run nearly every phone in the world. SoftBank of Japan bought Arm in 2016, Nvidia's attempt to buy it collapsed in 2022, and Arm returned to the stock market in 2023 with SoftBank keeping about 90%.

For most of its history Arm made no chips. It sells three kinds of license. An architecture license lets a company design its own cores that run Arm's instruction set, as Apple, Qualcomm, and Nvidia do. A core license sells Arm's own ready-made designs, its Cortex cores for phones and Neoverse cores for servers. A compute subsystem (CSS) is a bigger block, several cores already wired together and checked, which lets a customer build a chip faster.

Arm earns twice from each design: a fee when a company licenses it, and a royalty on every chip that ships with it, often tied to the chip's selling price. Its partners have shipped more than 350 billion Arm-based chips in all, by Arm's count in its annual report of May 2026. The company publishes no rates, but its chief executive, Rene Haas, wrote in 2025 that the typical royalty for its newer Armv9 architecture was about double that for Armv8, and that compute subsystems command more still. Raising the royalty per chip, rather than selling more chips, is the heart of Arm's strategy.

In March 2026 Arm crossed a line it had held for 35 years and launched a chip of its own, named for artificial general intelligence (AGI): the AGI CPU, a server processor for AI data centers. It carries 136 Neoverse V3 cores on two dies made on TSMC's 3nm process and draws 300 watts, by The Register's account of the launch. Meta helped design it and is its first customer, and OpenAI, Cerebras, and Cloudflare signed on as

launch partners. By July 2026 Arm said demand for it had passed $2 billion across its fiscal years 2027 and 2028.

Selling chips makes Arm a competitor of the companies that pay it royalties, and the tension is already in court. Arm sued Qualcomm in 2022, arguing that Qualcomm's purchase of Nuvia, a start-up that designed Arm cores, did not carry over Nuvia's license. A jury found in December 2024 that Qualcomm had not breached Nuvia's license, and the judge entered judgment for Qualcomm in September 2025; Arm appealed.

Qualcomm countersued, and in October 2026 a federal jury in Wilmington, Delaware, found that Arm had breached its architecture license with Qualcomm and had interfered with Qualcomm's business with Meta. Arm prevailed on four of Qualcomm's six claims, and the jury awarded damages of $2 in all. The larger questions were left to the judge, Maryellen Noreika. She must decide whether the breach lets Qualcomm stop paying royalties for five years, a sum that could run from hundreds of millions to billions of dollars, and whether Arm failed to negotiate in good faith. Both sides claimed victory, and either may appeal.

4d. RISC-V: The Open Standard

RISC-V, pronounced risk-five, began in 2010 as a research project at Berkeley, led by Krste Asanović with his students Yunsup Lee and Andrew Waterman and with David Patterson. It is an open standard: anyone may build a chip that runs it, without a license or a royalty. A non-profit body, RISC-V International, maintains it; it moved its legal home from the United States to Switzerland in 2020, partly to reassure members abroad that American export rules could not cut them off from the standard.

RISC-V is modular. A small base set of integer instructions is required, and extensions for multiplication, floating point, vectors, and more are optional. That flexibility suited microcontrollers but made it hard for software to know what a given chip could do. The fix is the profile, a required bundle of extensions. RVA23, the profile ratified in October 2024, makes vector instructions and support for virtual machines mandatory for application processors, so an operating system can target every compliant chip at once. Canonical's Ubuntu began requiring RVA23 with its October 2025 release, and the first compliant chips, such as SpacemiT's K3, reached boards in 2026.

Most RISC-V cores sit where nobody sees them, managing power, security, or storage inside larger chips. Nvidia uses RISC-V microcontrollers inside its GPUs, and Alibaba's XuanTie cores are in more than 6 billion chips shipped, by DigiTimes' report of October 20, 2026. The SHD Group, a research firm, forecast in April 2026 that RISC-V chips would reach 35.9 billion units a year by 2031. Big buyers are investing too: Meta agreed to buy Rivos, a RISC-V start-up, in 2025, and Qualcomm bought Ventana Micro Systems, another, in December 2025.

China has the strongest reason to back RISC-V. American export controls can bar Chinese firms from the most advanced Arm and x86 designs, but no one can revoke an open standard. Chinese companies are among the largest contributors to RISC-V, and Alibaba's chip unit, T-Head, now offers server-class XuanTie cores. China also has an instruction set of its own, LoongArch, built by Loongson for government and state-company computers (Module 7).

4e. Switching Instruction Sets

Changing instruction sets is hard because existing software must keep running. Apple has done it three times on the Mac, the last in June 2020, when it announced its move from Intel's x86 chips to its own Arm-based ones. Its first Arm-based chip, the M1, arrived in November 2020, and the move was complete within three years. Apple's translator, Rosetta 2, converts x86 programs into Arm instructions the first time they run, so most users noticed little.

Translation has a cost, since translated code usually runs slower than code built for the chip, and it cannot cover everything: programs that reach deep into the system, such as some anti-cheat software for games, often fail. So a switch succeeds only when the main programs are rebuilt for the new instruction set. Microsoft's own translator for Windows on Arm, Prism, improved in 2024, and Qualcomm's Snapdragon X chips gave Windows its first competitive Arm laptops that year (Module 6).

4f. Contested: Does the Instruction Set Still Matter?

Those who say it matters point to Apple. Its Arm-based Mac chips have led on performance per watt since 2020, and phones, where every milliwatt counts, run almost entirely on Arm. They argue that x86 carries decades of baggage, including instructions of uneven length that make it hard to decode many at once, so Arm cores can be built wider for the same power. The x86 makers' own moves show that the baggage is real: one shared feature of 4b, FRED, replaces an old way of handling interrupts, and Intel proposed in 2023 to drop the oldest operating modes, though it withdrew the plan in 2024.

Those who say it hardly matters point to the micro-operations of 4a: inside, every fast core is much the same machine. A careful academic study published in 2013, comparing Arm and x86 chips of similar design, found that the instruction set made little measurable difference to performance or energy. The gaps came instead from the design's goals, its process, and its software. Intel's Lunar Lake of 2024 and Panther Lake of 2026, x86 laptop chips built for efficiency, narrowed Apple's lead considerably, and Jim Keller, an architect of chips for AMD, Apple, Intel, and Tesla, has long argued that the instruction set is a small factor.

The evidence favors the second view on engineering and the first on business. A team that sets out to build an efficient x86 core can come close to an Arm one, so the instruction set alone does not explain Apple's lead. But Arm's licensing lets any company design its own chip, and x86's duopoly lets almost no one, which is why every cloud company's custom processor runs Arm (Module 7). The course's reading is that the instruction set now matters less for how fast a chip runs than for who may build one, and on what terms.

Worked example — What Arm Earns From a Chip

Arm's royalties were $715 million in its quarter to June 2026, about $2.9 billion a year at that pace. The company publishes no rates, so take illustrative ones consistent with what it has said: 1% of a chip's price for an older design, double that for Armv9, and 4% for a design built on a compute subsystem. A microcontroller that sells for $1, on an older design, then earns Arm a cent. A phone processor that sells for $150, on Armv9, earns it $3, and $6 if built from a compute subsystem. A server processor that sells for $5,000, at 2%, earns it $100. Those differences explain the strategy: Arm earns far more by moving phone and server chips to Armv9 and its subsystems than by adding microcontrollers, however many billions of them ship. They also explain the AGI CPU. When Arm sells its own chip it collects the whole price, less what TSMC charges to make it, rather than a few percent of the price. Its reported demand of more than $2 billion over two fiscal years (4c), about $1 billion a year, would equal about a third of its royalty income of about $2.9 billion a year, though a chip earns far less profit per dollar than a royalty. That is the prize, and the risk is that licensees such as Qualcomm come to treat Arm as a rival.

Module 4 Key Points

Instruction Sets: x86, Arm, and RISC-V

• Since the Pentium Pro of 1995, x86 cores split complex instructions into simple micro-operations, so fast cores look alike inside.

• Only Intel and AMD make mainstream x86 chips; AMD held a record 34.1% of x86 processors shipped in mid-2026.

• Arm earns a royalty on every chip built on its designs; its partners have shipped more than 350 billion.

• In October 2026 a jury found Arm breached Qualcomm's license, awarding $2; the judge may still suspend royalties for five years.

• RISC-V is open and royalty-free; Alibaba's XuanTie cores alone are in more than 6 billion chips.

• The instruction set matters less for a chip's speed than for who may build one, and on what terms.

Module 4 Quiz

Check Your Understanding

1. Since Intel's Pentium Pro of 1995, what do x86 cores do with their complex instructions?

A) Send them to a separate RISC chip on the motherboard
B) Split them into micro-operations
C) Run them more slowly in software emulation
D) Execute each one directly in hardware exactly as written, which is why x86 cores need no decoder at all

2. Which companies make mainstream x86 processors?

A) Intel, AMD, Arm, and Qualcomm, under a shared license
B) Intel and AMD
C) Any company, since x86 is an open standard
D) Intel alone, since AMD's license lapsed in 2009

3. What did the jury decide in Qualcomm's case against Arm in October 2026?

A) It found that Qualcomm had breached its license by using Nuvia's designs, and ordered it to pay billions.
B) It dismissed every claim, finding that Arm's license terms were fair and lawful.
C) It found Arm breached its license, but awarded $2.
D) It ordered Arm to open its designs.

4. What does the RVA23 profile, ratified in October 2024, do for RISC-V?

A) It sets a common baseline, with vector instructions and virtual-machine support required, so that software can target many chips at once.
B) It makes RISC-V royalty-bearing
C) It merges RISC-V with Arm
D) It limits RISC-V to microcontrollers in cars

5. In the worked example, what does Arm gain by selling its own AGI CPU rather than licensing designs?

A) It avoids paying TSMC to make the chip.
B) It no longer needs Meta as a customer.
C) It gains the right to stop its licensees from designing their own Arm cores.
D) It earns the chip's whole selling price, not a few percent.

Module 5 — The GPU: From Pixels to Parallel Computing

Rendering, Warps, Tensor Cores, and the Graphics Market

Recommended Reading / Reference

Book: Stephen Witt, *The Thinking Machine: Jensen Huang, Nvidia, and the World's Most Coveted Microchip* (2025), a history of Nvidia and of how its graphics chips came to run AI

Reference: Jon Peddie Research, quarterly reports on shipments of graphics cards and of all graphics processors in PCs

5a. The Graphics Problem

A screen is a grid of tiny dots, or pixels, each lit in a mix of red, green, and blue. A 4K screen has about 8.3 million of them, and a game running at 120 frames a second must work out the color of each one 120 times every second, about a billion pixels a second. Working out a pixel's color means deciding what object it shows, how that object is lit, and what lies in shadow, which can take hundreds or thousands of arithmetic operations.

The saving grace is that each pixel can be worked out on its own. The color of one has no bearing on the color of the next, so the work can be split among as many arithmetic units as a chip can hold. That is the kind of job Amdahl's law (Module 2) rewards: almost entirely parallel. A processor built for it trades the CPU's few clever cores for thousands of simple ones, and that design, the graphics processing unit, is the subject of this module.

5b. The Rendering Pipeline

A game builds its scenes out of triangles, often millions of them, each defined by three corner points called vertices. The GPU turns them into an image in stages. First it moves every vertex from the scene's coordinates to the screen's, which means multiplying each one by small grids of numbers called matrices. Then it rasterizes: for each triangle it works out which pixels the triangle covers. Then it shades each covered pixel, applying textures, the images painted onto surfaces, and computing the light that falls on it. Finally it blends the results into the frame.

The vertex and pixel stages run small programs called shaders, written by game developers and run once for every vertex or pixel, millions of times a frame. Their arithmetic is mostly multiplying and adding numbers in matrices. That detail shaped the history that followed: the operation at the heart of graphics, many multiply-adds applied in parallel to grids of numbers, is also the operation at the heart of neural networks (5f).

5c. Ray Tracing, Upscaling, and Generated Frames

Rasterizing is fast, but it fakes light, approximating shadows and reflections with tricks. Ray tracing simulates light instead: for each pixel the GPU traces a ray into the scene to find what it hits, then traces more rays toward lights and shiny surfaces. Path tracing goes further, following many randomly chosen rays as they bounce several times, the way film studios render. The costly step is finding which of millions of

triangles a ray hits, and Nvidia added hardware for that, its ray-tracing (RT) cores, with its Turing generation of 2018; AMD's chips have ray accelerators for the same job.

Even with that hardware, tracing every pixel of a 4K frame many times is too slow, so games now lean on AI. Upscaling renders a frame at a lower resolution and lets a neural network rebuild it at full size, using the motion of objects and earlier frames to fill in detail. Nvidia's version, DLSS (Deep Learning Super Sampling), arrived in 2018. AMD's FidelityFX Super Resolution (FSR) 4, its first upscaler built on machine learning, runs only on its newest Radeon cards; Intel has XeSS, and Sony has PlayStation Spectral Super Resolution (PSSR) for the PlayStation 5 Pro.

Frame generation goes further, inserting frames that a neural network predicts between the frames the game actually renders. Nvidia's DLSS 4 of January 2025 generates up to three frames for each rendered one, and DLSS 4.5, announced in January 2026, up to five, on its RTX 50 cards only. Generated frames make motion look smoother, but they do not make the game respond faster to the player, since responsiveness still follows the rendered rate.

That distinction matters when reading claims. At the launch of the RTX 50 series (5g), Nvidia's chief executive, Jensen Huang, said the RTX 5070 delivered the performance of the previous generation's $1,599 RTX 4090 for far less money. Reviewers found that held only with frame generation and aggressive upscaling switched on, not in frames actually rendered. Much of what a gamer sees in 2026 is inferred by a neural network rather than drawn, which makes graphics itself an AI workload.

5d. From Graphics Card to AI Engine

In the mid-1990s, add-in cards such as 3dfx's Voodoo took over the 3D drawing that CPUs had done in software. Nvidia, founded in 1993 by Jensen Huang, Chris Malachowsky, and Curtis Priem, launched the GeForce 256 on August 31, 1999, and marketed it as the world's first GPU, a term it popularized for a single chip that handled the geometry and lighting as well as the drawing. Its GeForce 3 of 2001 made the pipeline programmable with shaders, and researchers soon began bending those shaders to scientific arithmetic, disguising their numbers as pixels.

Nvidia made that easier in two steps. Its GeForce 8800 GTX of November 2006 replaced its separate vertex and pixel processors with unified shaders, identical processors that could run any kind of work. The same year it unveiled CUDA, a way of programming its GPUs in a version of the C language, with no pretense that the numbers were pixels (Module 12). CUDA ran on every Nvidia gaming card, so a graduate student with a few hundred dollars had a parallel computer. AMD, which bought the graphics maker ATI in 2006, offered similar tools but less consistently.

The payoff came in 2012. Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton of the University of Toronto trained a neural network, later called AlexNet, on two Nvidia GTX 580 gaming cards over about five to six days. It won that year's ImageNet contest for recognizing objects in photos with an error rate of 15.3% on its five best guesses, against 26.2% for the next best entry. The result launched the modern boom in deep learning, and the GPU became its engine.

Nvidia then built GPUs for AI on purpose. Its V100 of 2017, in the Volta generation, added tensor cores, units built only to multiply matrices (5f), and each generation since has added more of them: the A100 of 2020, the H100 of 2022, Blackwell from 2024, and Rubin in 2026 (Module 8). In the data-center chips the graphics hardware has largely been dropped, so that a chip still called a GPU is now an AI accelerator in all but name.

5e. Throughput Against Latency

A CPU core is built for latency, finishing one stream of instructions as fast as possible, with out-of-order execution, prediction, and large caches (Module 2). A GPU is built for throughput, finishing as much total work as possible, and it does so with many simple cores. Nvidia groups its cores into streaming multiprocessors (SMs), each with its own scheduler, registers, and small memories: the RTX 5090 gaming card has 170, and the Rubin data-center GPU 224, by Nvidia's figures. AMD calls the equivalent unit a compute unit.

The GPU runs threads, small copies of a program each working on its own piece of the data, in bundles. Nvidia's bundle, the warp, is 32 threads that execute the same instruction at the same moment on different data, a scheme it calls single instruction, multiple threads (SIMT). If the threads in a warp take different branches, the warp must run both paths in turn, with the threads on the other path sitting idle each time, so GPU code avoids branching on the data.

The trick that makes this work is how a GPU handles waiting. A CPU core stalls when it waits for memory unless it can find other work in its own instruction stream. Each streaming multiprocessor keeps dozens of warps resident at once, with all their registers, and when one warp waits for memory the scheduler switches to another whose data has arrived, in a single cycle. With enough threads in flight, the wait is hidden entirely, which is why a GPU needs tens of thousands of threads to run at full speed.

The cost is everything a GPU does badly. A single thread runs far slower than on a CPU, code full of unpredictable branches wastes most of the hardware, and programs must be written to expose massive parallelism. GPUs also depend on fast memory to feed so many cores, which is why they use graphics memory and HBM (Module 3). The two designs are complementary: in every AI server, CPUs run the operating system and coordinate the work while GPUs do the arithmetic (Module 7).

5f. Tensor Cores and Low Precision

A neural network is mostly matrix multiplication. Each layer multiplies a grid of incoming numbers by a grid of the network's parameters, adds the results, and passes them on. A tensor core is a unit that multiplies two small matrices and adds a third in a single operation, doing in one step what would take an ordinary core hundreds of multiply-adds. Most of a modern AI chip's arithmetic comes from these units, and its headline petaflops are tensor-core figures.

The other lever is precision, the number of bits used for each number. Science and engineering use 64-bit floating-point numbers (FP64), the computer's form of scientific notation. Neural networks turned out to tolerate far less, and AI has moved step by step from 32 bits (FP32) to 16 (FP16, and BF16, a variant devised at Google), 8 (FP8), and 4 (FP4, including NVFP4, a 4-bit format of Nvidia's own). Halving the bits roughly doubles the arithmetic a chip can do in the same area and power and halves the memory and bandwidth each number needs. The table sets out the main formats.

Format	Bits	Roughly what it holds	Main use	When AI adopted it
FP64 (double precision)	64	About 16 significant digits	Science and engineering simulation	Not used for AI
FP32 (single precision)	32	About 7 digits	Graphics; AI training until the mid-2010s	The first neural networks on GPUs

Format	Bits	Roughly what it holds	Main use	When AI adopted it
FP16 and BF16 (half precision)	16	About 3 digits for FP16; about 2 for BF16, which keeps FP32's range	AI training	FP16 from 2017, in Nvidia's Volta; BF16 from 2017 in Google's TPU v2 and from 2020 in Nvidia's A100
FP8	8	About 1 digit	Training and inference	From 2022, in Nvidia's Hopper
FP4, including Nvidia's NVFP4	4	Just 16 distinct values, with a shared scale for each small block	Inference, and increasingly training	From 2024-25, in Nvidia's Blackwell

The 4-bit formats work only because each small block of numbers shares a scale factor, which stretches the format's few values over the range the block needs. 8-bit formats also use shared scales, often one for a whole matrix.

Low precision is not free. Fewer bits mean coarser numbers, so a model squeezed to 4 bits must be trained or adjusted to cope, and results can suffer if it is done carelessly. Most scientific simulations need 64-bit accuracy, which has opened a gap between chips built for AI, whose 64-bit arithmetic has fallen behind, and chips built for supercomputers. AMD's MI430X is built for that market (Module 8).

5g. The Graphics-Card Market of 2026

Gaming is now the GPU's smaller business, but it is still large. Jon Peddie Research counted 12.5 million desktop graphics cards shipped in the second quarter of 2026, the most in any quarter since early 2022, as buyers rushed to beat further price rises. Nvidia shipped about 90% of them, AMD about 8%, and Intel about 2%, by Tom's Hardware's reading of the report. Counting the graphics built into laptop and desktop processors as well, about three-quarters of the 75.5 million PC graphics processors shipped that quarter were built into processors (Module 6) rather than separate cards, by this course's arithmetic from Jon Peddie's counts.

Nvidia's current gaming line, the RTX 50 series, launched in early 2025: the RTX 5090 at $1,999 on January 30, the 5080 at $999, the 5070 Ti at $749 in February, the 5070 at $549 in March, and the 5060 at $299 in May. The 5090 holds about 92 billion transistors and draws up to 575 watts, by Nvidia's specifications. AMD answered with the Radeon RX 9070 XT at $599 in March 2025, aimed at the middle of the market rather than the top, and Intel's Arc B580 launched at $249 in December 2024.

No new Nvidia gaming GPU had appeared in 2026 by October. The Information reported in February that Nvidia had shelved a refresh of the RTX 50 cards and would launch no new gaming GPU that year, cutting production to send memory to its AI chips. It also reported that mass production of the next generation had slipped to the end of 2027; Nvidia confirmed none of it. Board makers at the Computex trade show in June expected new cards from both Nvidia and AMD in the second half of 2027.

Meanwhile prices rose far above the launch prices, driven by the memory shortage (Module 3) and by buyers who use gaming cards for AI. On October 11, 2026, the lowest asking price for a new RTX 5090 on BuySellRam's marketplace tracker, run by a hardware reseller, was $4,921, about two and a half times its launch price. The cheapest RTX 5070 was listed at $725, about a third above its launch price. In 2026 Nvidia also stopped reporting gaming revenue as a line of its own (Module 14), a sign of how far graphics now trails AI in its business.

Worked example — One Matrix Multiplication on Three Chips

Multiply two square matrices of 8,192 rows and columns each. Every entry of the result needs 8,192 multiplications and 8,192 additions, and there are 8,192 times 8,192 entries, so the job takes about 2 times 8,192 cubed, or 1.1 trillion operations. A 16-core desktop CPU with two 512-bit vector units per core (Module 2) peaks at about 16 times 64 operations a tick at 5 GHz, roughly 5 trillion 32-bit operations a second. At a realistic 70% of peak it needs about a third of a second. An RTX 5090's tensor cores are rated by Nvidia at about 209 trillion 16-bit operations a second with the running sums kept at full precision, so at 70% they finish in about 7.5 milliseconds, some 40 times faster than the CPU, though on 16-bit rather than 32-bit numbers. A Blackwell B200 data-center GPU is rated by Nvidia at about 2.5 quadrillion 16-bit operations a second in its fastest configuration, without the sparsity explained in Module 8, so it finishes in about 0.6 milliseconds, some 500 times faster than the CPU. The arithmetic, not the clock, explains the gap: the B200's clock is slower than the CPU's, but its tensor cores do hundreds of thousands of multiply-adds in each tick, against about 500 for the CPU's 16 cores. Each matrix, about 134 MB at 16 bits, is small beside the GPU's memory, and every number in it is used 8,192 times, high on the roofline (Module 3), so for once memory is not the limit.

Module 5 Key Points

The GPU: From Pixels to Parallel Computing

• A 4K screen at 120 frames a second needs about a billion pixels colored each second, every one independently.

• GPUs hide memory delays by switching among thousands of threads, which run in warps of 32 sharing one instruction.

• AlexNet's 2012 win, trained on two Nvidia gaming cards, began the GPU's career in AI.

• Tensor cores multiply small matrices in one step; each halving of a number's bits roughly doubles throughput, down to 4 bits.

• Nvidia shipped about 90% of desktop graphics cards in mid-2026, and had launched no new gaming GPU in 2026 by October.

Module 5 Quiz

Check Your Understanding

1. How does a GPU keep its arithmetic units busy while some threads wait for memory?

A) It predicts branches far ahead
B) It raises its clock speed
C) It keeps all its data in registers
D) It switches at once to other groups of threads whose data has already arrived

2. What did Nvidia add to its Volta generation, the V100 of 2017?

A) Hardware for tracing rays of light
B) The first programmable pixel shaders, which let developers write their own lighting effects
C) Tensor cores
D) Support for the CUDA language

3. What do generated frames do in a game?

A) They raise the displayed frame rate and also cut the delay between a press of a button and its effect.
B) They smooth motion on screen but do not make the game respond faster.
C) They replace the need for any rendered frames at all.
D) They lower power use by switching off the GPU between frames.

4. Why did AlexNet matter to GPUs?

A) It showed gaming GPUs could train neural networks
B) It was the first program written in CUDA
C) It introduced the tensor core, which Nvidia then copied into its chips
D) It proved CPUs were faster than GPUs for large models, until Nvidia changed its designs

5. In the worked example, why does the B200 finish the multiplication so much faster than the desktop CPU?

A) Its clock runs a hundred times faster.
B) It skips half the arithmetic by guessing results.
C) Its tensor cores do far more multiply-adds in each tick, on 16-bit numbers.
D) It reads both matrices from disk instead of from memory, which is much faster for large files.

Module 6 — Phones, PCs, Consoles, and Cars

Systems on a Chip, and the Squeeze of 2026

Recommended Reading / Reference

Book: Chris Miller, *Chip War: The Fight for the World's Most Critical Technology* (2022), on how the chips in phones and PCs came to be designed in America and made in Asia

Reference: IDC and Gartner, quarterly estimates of PC shipments, and Counterpoint Research, quarterly reports on smartphone chips

6a. The System on a Chip

The processor in a phone is not just a CPU. It is a system on a chip (SoC): one die that carries the CPU cores, a GPU, a neural processing unit, a signal processor for the camera, engines that compress and play video, a secure area for keys and passwords, and the controllers for memory and displays, often with the radio modem as well. A laptop chip from Apple, Intel, AMD, or Qualcomm is built the same way, on a larger scale, though Intel's and some of AMD's join several dies in one package (Module 2).

The main reason is energy. Moving a bit across a chip costs a small fraction of the energy of moving it between chips, so putting everything on one die saves power, space, and cost. Apple goes a step further with unified memory: the CPU, GPU, and neural engine share one pool of memory in the same package, so data need not be copied between them (Module 3). The cost is flexibility, since nothing can be upgraded later, and size, since the die grows with every unit added.

6b. Phone Chips

Four companies design most of the world's leading phone processors, and all four use Arm's instruction set (Module 4). Apple designs its A-series chips for the iPhone alone. Qualcomm sells its Snapdragon chips to most makers of premium Android phones. MediaTek of Taiwan, which ships the most phone processors of anyone, sells across every price range, and Samsung uses its own Exynos chips in some of its Galaxy phones and Qualcomm's in others. Google designs the Tensor chips in its Pixel phones, and from the Tensor G5 of August 2025 TSMC rather than Samsung makes them.

The first phone chips made on 2nm-class processes arrived in 2026. Samsung's Exynos 2600, built on Samsung's own 2nm process, reached some Galaxy S26 phones early in the year. Apple's A20 Pro, announced on September 9, 2026 for the iPhone 18 Pro, has six CPU cores, two of them new "super cores", and a neural engine that handles 8-bit floating-point numbers. Apple claims its fast cores are up to 20% quicker and its GPU up to 40% quicker than the A19 Pro's.

Qualcomm's Snapdragon 8 Elite Gen 6 and MediaTek's Dimensity 9600, both announced in September 2026 on TSMC's 2nm process, followed in Chinese phones from Xiaomi and Vivo. Qualcomm's has two cores that reach 5.0 GHz, faster than most laptop chips. Huawei, cut off from TSMC by American export rules, makes its Kirin chips at SMIC, the leading Chinese chipmaker, on an older process.

Apple is also replacing other people's chips inside its phones. Its first cellular modem of its own, the C1, appeared in the iPhone 16e in February 2025, and its own wireless chip, the N1, in the iPhone 17 family that September, each displacing a supplier: Qualcomm for modems, Broadcom for wireless. Qualcomm's

revenue from phone chips fell 20% from a year earlier, to $5.09 billion, in its quarter to June 2026, while its chips for cars and other devices grew.

6c. PC Chips

Apple's Mac chips grow out of its phone chips. The M5 of October 2025 has a 10-core CPU and a 10-core GPU with a neural accelerator in each GPU core. The M5 Pro and M5 Max followed in March 2026 and the M5 Ultra in August, alongside the M6, Apple's first chip made on a 2nm process. The M6 has 12 CPU cores, two of them super cores like the A20 Pro's (6b), and 12 GPU cores, and Apple claims up to 1.2 times the M5's speed when all its cores work at once. It came first to the Mac mini, then to the 14-inch MacBook Pro and the iMac, announced on October 27, 2026. Apple's Macs held 10.5% of PC shipments in the third quarter of 2026, by Gartner's count, and its chips have set the benchmark for performance per watt (Module 4).

Intel's answer is Panther Lake, sold as Core Ultra Series 3, launched in January 2026 as its first PC chip made on its own 18A process. The top model has 16 cores, boosts to 5.1 GHz, and carries Intel's Arc B390 integrated GPU, which Notebookcheck measured as roughly matching a 60-watt RTX 4050 laptop graphics card. More than 200 laptop designs use it. Intel's next family, Nova Lake, was meant for late 2026, and leaked plans put its first desktop chips in early 2027.

AMD's laptop chips at the start of 2026 were refreshes. Its Ryzen AI 400 series, shown in January, raised clocks and the neural engine's speed over the Ryzen AI 300 chips without changing their design. Its Ryzen AI Max chips, known as Strix Halo, pair 16 cores with a large integrated GPU and up to 128 GB of shared memory, popular for running AI models locally. Chips on AMD's new Zen 6 design had reached servers (Module 7) but not PCs by October 2026.

Arm's push into Windows PCs gained strength in 2026. Qualcomm's Snapdragon X2 Elite, announced in September 2025 with up to 18 cores, was due in laptops in the first half of 2026. Then Nvidia entered with RTX Spark, unveiled at the Computex trade show on June 1, 2026: up to 20 Arm cores and a Blackwell GPU with 6,144 cores, on one package with up to 128 GB of memory, built on Arm's compute subsystems. Nvidia expected more than 30 laptops and about 10 desktops with it from the autumn, from Dell, HP, Lenovo, Asus, MSI, and Microsoft.

Nvidia is also working with Intel. In September 2025 it agreed to invest $5 billion in Intel and to develop x86 PC chips with Nvidia graphics built in, a purchase completed in December 2025. No such chip had been announced by October 2026, and leaked plans point to early 2028.

6d. Neural Processing Units and the AI PC

A neural processing unit is a block of matrix engines, like a small cousin of a GPU's tensor cores (Module 5), built to run AI models on as little power as possible. Its speed is quoted in trillions of operations a second (TOPS), usually on 8-bit integers. Apple has put one, its Neural Engine, in every iPhone since 2017, where it handles face recognition, photo processing, and dictation without draining the battery.

In May 2024 Microsoft set a bar for Windows PCs: to carry its Copilot+ label, a PC needs an NPU of at least 40 TOPS, along with at least 16 GB of memory. Every major PC chip now clears it, and the chip makers compete on bigger totals. Qualcomm's X2 Elite Extreme has an 80-TOPS NPU, and Intel claims 180 TOPS for Panther Lake by adding its NPU, GPU, and CPU together, a sum that no single program uses at once.

What buyers do with the NPU is a fair question. The common uses are modest: blurring a video call's background, live captions, searching photos by what they show, and running small language models. The largest models still run in the cloud, because a laptop's memory and bandwidth cap the model it can hold and how fast it runs (Module 3). The AI PC has so far mattered more to marketing than to purchases, which in 2026 turned on price (6g).

6e. Consoles

Game consoles run on semi-custom chips, designed by a chip company to a console maker's order. AMD makes them for Sony's PlayStation 5 and Microsoft's Xbox Series machines, both launched in November 2020, each pairing eight Zen 2 CPU cores with a Radeon GPU; for the PlayStation 5 Pro of November 2024; and for Valve's Steam Deck handheld and Steam Machine. Nintendo's Switch 2 of June 2025 uses a custom Nvidia chip with an Arm CPU and an Nvidia GPU that supports DLSS (Module 5).

Sony had shipped 95.3 million PlayStation 5s by June 30, 2026, and Nintendo 23.68 million Switch 2s in their first 13 months, its fastest-selling console. Xbox hardware is shrinking: its revenue fell 29% in Microsoft's fiscal year to June 2026. The next consoles are set: AMD's chief executive, Lisa Su, said in February 2026 that AMD's chip for the next Xbox was ready for a launch in 2027. Sony had not announced a PlayStation 6 by October 2026.

Consoles usually get cheaper as they age. This generation has gone the other way, first because of tariffs in 2025 and then because of memory in 2026. The disc PlayStation 5 rose to $649.99 in the United States on April 2, 2026, and the Xbox Series X to $799.99 and Series S to $499.99 on August 1. The Switch 2 rose from $449.99 to $499.99 on September 1. Nintendo's president, Shuntaro Furukawa, named the AI-driven memory shortage as the biggest cause and said the increase still would not cover about ¥100 billion of added costs.

Valve felt the same pressure. It announced the Steam Machine, a small PC for the living room, in November 2025, then delayed it in February 2026 to rethink its price because of memory costs. It shipped on June 29, 2026, at $1,049 and $1,349, prices Valve said were far above its plans. In July 2026 Americans spent 29% less on console hardware than a year earlier, by Circana's count.

6f. Cars, Robots, and Small Chips

A car has long carried dozens of microcontrollers, each running one function such as the brakes, the windows, or the airbags. Newer cars consolidate them into a few central computers, with powerful processors that run driver assistance, the dashboard, and the infotainment screen together. Nvidia sells Drive Thor, a single car computer built for this, and Qualcomm sells its Snapdragon car chips; Qualcomm's automotive revenue grew 61% from a year earlier, to $1.59 billion, in its quarter to June 2026.

Tesla designs its own. Its AI5 chip, made by TSMC and Samsung, was finished in design in April 2026, nearly two years behind its original schedule by press reports. On Tesla's earnings call of October 21, 2026, its chief executive, Elon Musk, said AI5 would go into Optimus, Tesla's humanoid robot, around April 2027 (Module 9 follows Tesla's other chips). Robots and drones from other makers mostly use Nvidia's Jetson modules, whose Jetson AGX Thor developer kit sells for $3,499.

At the small end, microcontrollers matter for their sheer number rather than their speed (Module 1). Many now run Arm's Cortex-M cores or RISC-V cores, and they set the pace for whole industries: the chip

shortage of 2021, which idled car factories around the world, was mostly a shortage of cheap microcontrollers and power chips, not of advanced processors.

6g. The Squeeze of 2026

The memory shortage (Module 3) hit every kind of device in 2026. PC shipments fell 20.1% from a year earlier in the third quarter, 15.8 million fewer units, by IDC's count; Gartner put the fall at 22.4%, to 55.4 million, back to the level of early 2023. Gartner's analyst Namrata Banerjee blamed higher prices, scarce parts, and a shift to costlier configurations rather than unsold stock. Apple's Macs fell least, by 7.1% on Gartner's count, while HP's shipments fell by almost a third on IDC's. Apple raised prices across its Macs and iPads on June 25, 2026, blaming the shortage of memory chips: its cheapest MacBook Pro went from $1,699 to $1,999, the price at which the M6 model arrived in October.

Phones and their chips fell too. Shipments of smartphone processors dropped 15% from a year earlier in the first half of 2026 as memory costs rose, by Counterpoint's count. Qualcomm told investors in July that it was raising prices to offset a difficult memory and supply market, and the iPhone 18 Pro launched at a higher price than its predecessor. Desktop processor shipments fell more than 20% in the second quarter, by Mercury Research's estimate.

The squeeze shows how AI reaches into ordinary devices: until new memory factories open (Module 3), the price of a personal computer depends in part on the appetite of the largest AI companies.

Worked example — A Laptop's Energy Budget

Take a thin laptop with a 70 watt-hour battery, a common size. Its power draw depends on the job, and these figures are illustrative. Playing a video, with the screen on and the video engine doing the work, it draws about 5 watts, so the battery lasts about 70 divided by 5, or 14 hours. Browsing the web at about 7 watts gives about 10 hours. A game that pushes the integrated GPU to 40 watts drains it in under 2 hours. Now run a small language model locally, writing a reply of 500 tokens. On the NPU, at 5 watts and 20 tokens a second, the reply takes 25 seconds and 125 joules, about 0.035 watt-hours, so the battery could produce about 2,000 such replies counting the chip alone. On the GPU, at 30 watts and 40 tokens a second, the reply takes 12.5 seconds but 375 joules, about 0.1 watt-hours, so the battery would give about 670 replies. The GPU answers twice as fast and spends three times the energy. That trade explains why phones and laptops carry NPUs at all: on a battery, energy per task matters more than speed.

Module 6 Key Points

Phones, PCs, Consoles, and Cars

• A system on a chip puts the CPU, GPU, NPU, and more on one die, because moving data on the chip costs far less energy.

• The first 2nm phone chips arrived in 2026: Samsung's Exynos 2600, then Apple's A20 Pro and chips from Qualcomm and MediaTek.

• Microsoft's Copilot+ label requires an NPU of at least 40 trillion operations a second.

• Consoles got dearer in 2026: the Switch 2 rose to $499.99 and the Xbox Series X to $799.99.

• PC shipments fell about a fifth in the third quarter of 2026 as memory costs passed into prices.

Module 6 Quiz

Check Your Understanding

1. Why put a CPU, GPU, and NPU on one system on a chip?

A) Because a single die can run only one instruction set at a time, which forces every part to share it
B) Because separate chips cannot share a battery
C) To let each part use a different clock speed
D) Moving data on-chip uses less energy

2. What does Microsoft's Copilot+ PC label require?

A) A 2nm processor
B) An NPU of at least 40 trillion operations a second
C) An Arm processor
D) At least 32 GB of memory

3. Which company makes the chip in the Nintendo Switch 2?

A) Nvidia
B) AMD
C) Qualcomm
D) Nintendo itself, with a chip designed in Kyoto and made by TSMC

4. By how much did Gartner say PC shipments fell in the third quarter of 2026?

A) About 2%
B) About 45%, the worst on record
C) About 22%
D) Hardly at all, since prices rose instead

5. Why did Valve delay its Steam Machine in February 2026?

A) Its chip failed testing at AMD.
B) Microsoft blocked its operating system.
C) It waited for Nvidia's new graphics cards.
D) It needed to rethink the price after memory costs rose sharply.

Module 7 — Server CPUs

Cores, Clouds, and the Host for AI

Recommended Reading / Reference

Book: Luiz André Barroso, Urs Hölzle, and Parthasarathy Ranganathan, *The Datacenter as a Computer: Designing Warehouse-Scale Machines* (third edition, 2018), Google engineers' account of how a data center is built as one computer

Reference: The TOP500 list of supercomputers, at top500.org, published every June and November with the processors, speed, and power of each system

7a. What a Server Processor Does

A server processor runs the internet's back rooms: websites and the programs behind apps, databases, email, storage, and the virtual machines that cloud companies rent out by the hour. It differs from a PC chip in what it is built to sustain. It has many more cores, many more memory channels to feed them, dozens of lanes of the PCI Express links that connect drives, network cards, and accelerators, and memory that corrects its own errors. It runs flat out, all day, for years.

Most servers hold two processors, or sockets, sharing the work. Cloud companies divide each server into virtual machines and sell them by the virtual CPU, usually one thread of a core (Module 2). The market for server processors grew nearly 20% from a year earlier in the second quarter of 2026, by Mercury Research's estimate, even as PC processors shrank, and both Intel and AMD said demand ran ahead of supply. In March 2026 the two warned Chinese customers of longer waits for server chips. In October Intel forecast that shipments of server processors across the industry would grow more than 30% in both 2026 and 2027, with supply short into 2028.

AI has given server processors a new job: managing the accelerators. Every AI server needs CPUs to load data, run the operating system and the scheduling software, move results across the network, and handle everything the GPUs cannot. Nvidia's rack of 72 GPUs carries 36 of its own Arm-based CPUs, one for every two GPUs, and an ordinary server with eight GPUs carries two x86 CPUs. AI agents, programs that use tools such as code and web searches to complete tasks, add more CPU work still, since the tools run on CPUs.

7b. The Race for Cores

Server chips have the most cores of any CPUs, and the count has climbed fast; the table below gives the leading chips. AMD's EPYC Venice, the first high-performance server processor on TSMC's 2nm process, has a third more cores than its predecessor, Turin, and its memory channels deliver about 1.6 TB a second per socket. AMD claims more than 70% better performance and efficiency than Turin.

Intel's Xeon 6+, known as Clearwater Forest, its first server chip made on its own 18A process, uses small, efficient cores without multithreading, and a two-socket server holds 576 of them. Intel claims about 30% more performance per thread than AMD's densest Turin chip. Its larger cores come in Granite Rapids, with up to 128 cores, and Diamond Rapids, on 18A-P, with more still.

Chip	Maker	Cores	Process	Status, October 2026
EPYC Venice (6th generation)	AMD	Up to 256 (512 threads)	TSMC 2nm	Debuted July 22, 2026
Xeon 6+ Clearwater Forest	Intel	Up to 288 small cores	Intel 18A	Shipping since mid-2026
Xeon 7 Diamond Rapids	Intel	About 192 large cores	Intel 18A-P	Planned for 2027
Graviton5	Amazon Web Services	192	TSMC 3nm	Generally available June 10, 2026
Cobalt 200	Microsoft	132 (Arm Neoverse V3)	TSMC 3nm	In deployment for 2026
Axion	Google	Arm Neoverse V2	Not disclosed	In service since 2024; hosts Google's newest TPUs (Module 9)
Vera	Nvidia	88 custom Arm cores (176 threads)	Not disclosed	Shipping with the Vera Rubin rack (Module 8); also sold alone
AmpereOne	Ampere, owned by SoftBank	Up to 192	TSMC 5nm	On sale; SoftBank's $6.5 billion purchase closed November 25, 2025

Core counts are the makers' figures. A thread is a stream of instructions; with multithreading, one core runs two (Module 2). Neoverse V2 and V3 are Arm's own server core designs.

The cloud companies' own chips match the merchant parts on cores. Amazon Web Services (AWS), Amazon's cloud business, claims up to 25% better performance for Graviton5 than for Graviton4. Microsoft's Cobalt 200 is built from Arm's ready-made Neoverse designs, while Nvidia's Vera, the CPU of its Rubin generation, uses cores of Nvidia's own design and can be bought on its own, in racks of 256 CPUs.

More cores need more memory bandwidth, or they starve (Module 3). That is why Venice and Diamond Rapids move to 16 memory channels and why server chips carry the largest caches of any processors (Module 3). It is also why the newest server chips use PCI Express 6.0, the latest version of the links to accelerators and network cards, which doubles the speed of its predecessor.

7c. AMD's Climb

In 2017 AMD held about 1% of the market for server processors. Its comeback began that year with the first EPYC chips, built on its new Zen design. The second generation, Rome, was the chiplet design of Module 2, made by TSMC on a 7nm process, and offered 64 cores when Intel's mainstream server chips topped out at 28. Intel had stumbled: its 10nm process ran years late, and its server chips lagged in cores and efficiency for half a decade.

By 2026 AMD's share of server processor revenue ran well ahead of its share of units, by Mercury Research's estimates (Module 4). The gap between the two shares says that AMD sells the larger, costlier chips. Intel's comeback rests on its 18A process, Clearwater Forest, and Diamond Rapids. Its Data Center and AI unit grew fastest of its businesses in the third quarter of 2026 (Module 14), though Intel said the output of its Intel 3 process and a shortage of the substrates on which chips are packaged kept it from making enough Xeons.

7d. Arm in the Cloud

The largest buyers of servers have moved to Arm. AWS says Graviton chips made up more than half of all the new CPU capacity it added over the past three years, and that its custom-chip business, Graviton and its AI chips together, passed an annual run rate of $25 billion by July 2026; both are AWS's figures. Meta committed in 2026 to tens of millions of Graviton5 cores. Google uses its Axion chips for general work and, from its eighth-generation TPUs, as their host processors, replacing x86 there for the first time (Module 9).

The reasons echo Module 4. An Arm license lets a cloud company design exactly the chip it needs, without paying Intel's or AMD's margin, and Arm's efficient cores lower the power bill. Arm reported that its royalties from data-center chips more than doubled from a year earlier in its quarter to June 2026. It also cited IDC's finding that spending on servers built around accelerators with Arm host processors had nearly doubled over two quarters and passed spending on those with x86 hosts.

x86 remains strong where software must run unchanged. Companies running their own data centers, with decades of programs and licenses written for x86, mostly stay with Intel and AMD, and so do many of the cloud's rented machines, since customers ask for them. The split is roughly this: Arm is winning the cloud companies' own workloads and the hosts for AI chips, while x86 holds the enterprise.

7e. China's Server Processors

China's government wants its data centers on domestic processors, and four families compete. Hygon makes x86 chips derived from AMD's first Zen design, under a 2016 joint venture, and its revenue rose 66.5% from a year earlier, to 9.1 billion yuan, in the first half of 2026, by Caixin's report. Zhaoxin makes x86 chips through a joint venture between VIA Technologies and the Shanghai government. Huawei's Kunpeng chips use Arm's instruction set, and its roadmap promises a Kunpeng 960 with up to 256 cores in 2028. Loongson's chips run LoongArch, an instruction set of its own (Module 4).

The clearest sign of China's progress came in June 2026, when a new Chinese system, LineShine, at the National Supercomputing Center in Shenzhen, took first place on the TOP500 list of supercomputers, the first Chinese system to lead it since 2017. It runs on LingKun LX2 processors, each with 304 cores at 1.55 GHz, and lists no accelerators. Its speed and power, and what the list measures, are in Module 13. A machine of CPUs alone beat GPU-based rivals on the list's benchmark, though with more power.

7f. Software Licensed by the Core

Much business software is priced by the core. Microsoft sells its SQL Server database in licenses of two cores each, with a minimum of four cores per processor. Oracle counts processors with a table of factors per core, and VMware, owned by Broadcom since 2023, moved in 2024 to subscriptions charged per core with a minimum of 16 cores per processor. For such software the license bill can far exceed the cost of the hardware and its electricity.

Per-core pricing shapes which chips sell. A buyer running licensed databases wants fewer, faster cores, so both companies sell chips with modest core counts and high clocks, such as AMD's 16-core EPYC 9175F, which boosts to 5 GHz. Clearwater Forest and the densest Venice chips suit the opposite case, software built to spread across many small cores and licensed by the server or not at all, which describes much of what the cloud companies run for themselves.

Worked example — Replacing 100 Five-Year-Old Servers

Take 100 two-socket servers bought in 2021, each with two 32-core processors: 6,400 cores, each server drawing about 700 watts at full load. Suppose, as an illustrative assumption, that a 2026 core does 1.6 times the work of a 2021 core; the makers claim more for their best chips. The same work then needs 4,000 new cores, which fit in 8 two-socket servers with two 256-core Venice chips each (7b), 4,096 cores in all. If each new server draws about 1.6 kilowatts at full load, the 8 draw about 13 kilowatts against the old fleet's 70, a saving of about 57 kilowatts. Over a year of 8,760 hours that is about 500 megawatt-hours, about $50,000 at 10 cents a kilowatt-hour, and closer to $75,000 once the building's cooling and power losses are counted (Module 11). The space falls from about five racks to under half of one. The largest saving may be licenses: software charged per core would need 2,304 fewer core licenses. At an illustrative $7,500 per core, that is about $17 million, far more than the electricity. These figures are assumptions for the example, not any vendor's quote, but the pattern is real, and it is why companies replace servers long before they wear out.

Module 7 Key Points

Server CPUs

• AMD's EPYC Venice reached 256 cores on a 2nm process in July 2026; Intel's Clearwater Forest packs 288 small cores.

• AMD rose from about 1% of server processors in 2017 to about a third of units and nearly half of revenue in 2026.

• AWS says its Graviton chips made up more than half of the new CPU capacity it added over three years.

• In Nvidia's rack of 72 GPUs, 36 Arm-based CPUs manage the work, one for every two GPUs.

• LineShine, a CPU-only Chinese system, topped the June 2026 TOP500, the first Chinese leader since 2017.

• Software licensed per core rewards fewer, faster cores; its bill can far exceed the hardware's cost.

Module 7 Quiz

Check Your Understanding

1. How many cores does AMD's EPYC Venice offer at most?

A) 128 cores, as before
B) 256 cores
C) 288 cores, like Intel's
D) 512 cores in one socket

2. What new job do CPUs have in AI servers?

A) Hosting the GPUs: feeding them data and managing the work
B) Doing the matrix math
C) Cooling the rack
D) Replacing the GPUs entirely for training, since CPUs now carry tensor cores of their own

3. What share of the new CPU capacity AWS added over three years does it say Graviton made up?

A) About 5%
B) More than half
C) All of it, since AWS no longer buys x86 chips
D) About a tenth, mostly for testing

4. Why does software licensed per core push some buyers toward fewer, faster cores?

A) Because faster cores need no licenses at all.
B) Because per-core software cannot run on chips with many cores.
C) Because cloud providers ban per-core software.
D) Because the license bill grows with the number of cores, so doing the same work on fewer cores cuts it.

5. What was unusual about LineShine, the top system on the June 2026 TOP500 list?

A) It ran on CPUs alone, with no accelerators listed
B) It was built in America
C) It used Nvidia's Rubin GPUs
D) It was the first machine ever to pass an exaflop, years before any other system on the list

Module 8 — AI Accelerators: Nvidia, AMD, and the Challengers

Spec Sheets, Roadmaps, and Rivals

Recommended Reading / Reference

Book: Tae Kim, *The Nvidia Way: Jensen Huang and the Making of a Tech Giant* (2024), on how Nvidia's culture and bets prepared it for the AI boom

Reference: Epoch AI, data on machine-learning hardware and on the production of AI chips, at epoch.ai

8a. Training and Inference

AI chips do two jobs. Training builds a model: it runs trillions of words, images, or other examples through a neural network and nudges its parameters after each batch to reduce its errors. Training a frontier model takes months on tens of thousands of chips working as one (Module 10). It needs fast arithmetic, a great deal of memory, for the parameters and the bookkeeping that adjusts them, and fast links between chips.

Inference runs the finished model to answer requests, and it has two phases. Prefill reads the whole prompt at once, processing every token in parallel; it reuses each parameter across many tokens, so it is limited by arithmetic. Decode then writes the reply one token at a time, reading every parameter for each one, so it is limited by memory bandwidth (Module 3). Reasoning models, which write long chains of working before answering, spend most of their time in decode.

Because the two phases need different things, the newest systems split them, running prefill on chips rich in arithmetic and decode on chips rich in memory bandwidth, an approach called disaggregated serving. Inference has grown to rival training as a use of AI chips, since every user of a chatbot or an agent generates it, and the chip makers now design for both. The rest of this module looks at the chips, starting with how to read their specifications.

8b. Reading a Specification Sheet

An AI chip's headline number is its peak arithmetic rate, in petaflops, and it hides two choices. The first is precision. The rate rises as the number of bits falls (Module 5), so a figure means little until its precision is stated. The second is sparsity. Nvidia's chips can skip multiplications when two of every four parameters in a block are zero, a pattern called structured sparsity, and the company often quotes rates that assume it, which doubles the figure. Few models in use are built that way, so dense figures, without sparsity, are the fair comparison.

The other lines matter as much. Memory capacity sets the size of model a chip can hold and memory bandwidth the speed of decode (Module 3). The bandwidth of the chip's links to other chips sets how well it works in a group (Module 10). Power sets what the data center must supply and cool (Module 11). Names count differently too: at its conference in March 2025 Nvidia called its next rack the Vera Rubin NVL144, counting each of 144 GPU dies, and from January 2026 it called the same rack the NVL72, counting the 72 packages that hold them.

Chip	Memory	Memory bandwidth	Dense 8-bit (petaflops)	Dense 4-bit (petaflops)	Power	Available
Nvidia H100	80 GB HBM3	3.35 TB/s	2	None	Up to 700 W	2022-23

Chip	Memory	Memory bandwidth	Dense 8-bit (petaflops)	Dense 4-bit (petaflops)	Power	Available
Nvidia H200	141 GB HBM3E	4.8 TB/s	2	None	Up to 700 W	2024
Nvidia B200 (Blackwell)	192 GB HBM3E	8 TB/s	5	10	Up to 1,200 W	2025
Nvidia B300 (Blackwell Ultra)	288 GB HBM3E	8 TB/s	5	15	Up to 1,400 W	From late 2025
Nvidia Rubin	288 GB HBM4	Up to 22 TB/s	Not published	50 for inference, 35 for training	Not published; 1.8-2.3 kW reported	Shipping (8c)
AMD MI355X	288 GB HBM3E	8 TB/s	About 5	About 10	Up to 1,400 W	June 2025
AMD MI455X	432 GB HBM4	19.6 TB/s	About 20	About 40	Not disclosed	Due in volume (8d)

Makers' figures for each chip package, without sparsity; the Blackwell figures are for its fastest configuration. Rubin's 4-bit figures use Nvidia's own format, NVFP4. AMD's page for its Helios rack also gives up to 23.3 TB/s for the MI455X's memory. The reported Rubin power figures are press reports of two settings, not Nvidia's.

8c. Nvidia's Yearly Cadence

Nvidia now releases a new data-center generation every year, alternating new designs with upgraded versions. Hopper, in the H100 of 2022 and the H200 of 2024, powered the first wave of large language models. Blackwell, announced in March 2024, joined two dies in one package and was followed by Blackwell Ultra; the table gives when each reached customers. Most were sold not as single chips but in racks: the GB200 and GB300 NVL72, each joining 72 GPUs and Nvidia's Grace CPUs into one machine (Modules 7 and 10).

Rubin is the 2026 generation. Each Rubin GPU joins two compute dies and eight stacks of HBM4, and in the Vera Rubin NVL72 rack it is paired with Nvidia's Vera CPU (Module 7). Nvidia declared the platform in full production at the CES trade show in January 2026, but that was its own term: production shipments began in August, its chief financial officer, Colette Kress, said on August 26. By then racks were running at CoreWeave, Google Cloud, Microsoft Azure, Oracle, and Nebius.

Kress expected Vera Rubin to bring in about 20% of Nvidia's data-center revenue in the quarter to late October and called it the fastest product ramp in the company's history. Nvidia claims the rack cuts the cost of each token to a tenth of the GB200 rack's. A trade analysis of Nvidia's figures found that the claim rested on a single model of the mixture-of-experts kind, which suits the new rack's memory and links unusually well, so the gain on other models is likely smaller.

The roadmap runs on. Rubin Ultra, planned for the second half of 2027, is meant to come in a new rack called Kyber, drawing 600 kilowatts, with 144 GPU packages. SemiAnalysis reported in July 2026, as relayed by the German site Hardwareluxx, that Nvidia had dropped the planned four-die Rubin Ultra package because it warped, that Kyber faced delays, and that the largest version might not arrive before 2028. Nvidia said its roadmap was intact. After Rubin comes Feynman, planned for 2028 on TSMC's A16 process, with chips stacked in three dimensions.

Nvidia has also bought its way into new kinds of inference. On December 24, 2025 it licensed the technology of Groq, a start-up whose chips keep a model's parameters in fast on-chip SRAM rather than in HBM, which makes decode very quick at the cost of capacity. The deal was not exclusive, but Groq's

founder, Jonathan Ross, and its president, Sunny Madra, joined Nvidia. The price was not disclosed and was widely reported at about $20 billion.

By August 2026 Nvidia's Groq 3 LPX rack, with 256 of the chips and 128 GB of on-chip memory per rack, was in full production. Nvidia claims it raises inference output per megawatt by up to 35 times, a figure no one has checked independently. A chip announced in September 2025 for the prefill phase, Rubin CPX, was missing from Nvidia's 2026 product lists, and its status was unclear in October 2026.

8d. AMD's Challenge

AMD has competed on memory. Its MI300X of December 2023 offered 192 GB, far more than the H100 (8b), and its successors kept the lead in capacity. Its weakness was scale: until 2026 its GPUs could be joined in groups of only eight, against Nvidia's 72. Its answer is Helios, its first rack-scale system, which joins 72 MI455X GPUs with its Venice CPUs (Module 7) and its own network cards. It uses the UALink protocol over Ethernet to link the GPUs (Module 10). AMD rates Helios at 2.9 exaflops of 4-bit arithmetic and 31 TB of HBM4.

The MI400 series launched in the second quarter of 2026, and AMD expects Helios to reach volume deployments in the second half. The family includes the MI430X, which keeps full 64-bit arithmetic for supercomputers (Module 5), and the MI500 series is planned for 2027. AMD calls Helios a reference design that partners build, and it does not report its GPU revenue separately from its server CPUs (Module 14).

AMD's customers are the largest buyers in AI, and it has paid to win some of them with its own shares. In October 2025 OpenAI agreed to deploy 6 gigawatts of AMD GPUs over several generations, starting with 1 gigawatt of the MI450 series in the second half of 2026, and received warrants for up to 160 million AMD shares, about a tenth of the company, vesting as the deployment proceeds and AMD's share price rises. Meta signed a similar deal for up to 6 gigawatts in February 2026, with a warrant for another 160 million shares.

Anthropic agreed in July 2026 to take up to 2 gigawatts of MI450-series GPUs, the first gigawatt from the first half of 2027, while AMD agreed to invest up to $5 billion in Anthropic. Oracle was due to start deploying 50,000 MI450s in the third quarter of 2026, and Microsoft committed in July to deploy Helios racks on its Azure cloud. A gigawatt here means the power the chips and their data centers draw, a measure the industry now uses to size deals (Module 11).

8e. Intel and the Start-ups

Intel has struggled in AI chips. Its Gaudi 3 accelerator, with 128 GB of an older kind of HBM, is still sold but won little business, and it kept its next design, Falcon Shores, for internal development. Its new plan rests on GPUs it can program like its graphics chips. The first, Crescent Island, is a chip for inference with 160 GB of LPDDR5X rather than HBM, which Intel began sampling to customers in October 2026 as the GPU D-series; a larger design, Jaguar Shores, follows later.

Start-ups attack where GPUs are weakest, most often decode. Cerebras builds a single chip from a whole wafer, with 4 trillion transistors and 44 GB of SRAM on its third-generation design, by its figures. It listed on Nasdaq on May 14, 2026, raising about $6.4 billion including shares sold to its underwriters, and its revenue rose 74% from a year earlier, to $180.1 million, in its first quarter as a public company. It says more than 600 megawatts of its capacity is running or under contract, for delivery by the end of 2027.

Others are private and raising large sums, by press and investor reports. Etched builds chips only for the transformer, the design behind today's language models. MatX designs chips for large language models, and Positron builds inference systems. Tenstorrent, led by Jim Keller, builds AI chips around RISC-V cores and was valued at $6.2 billion in October 2026, while SambaNova and d-Matrix also stayed independent. Groq itself continues under a new chief executive, selling its chips through its own cloud.

8f. Huawei and China's Accelerators

American export controls since October 2022 have barred sales of the most advanced AI chips to China, and the rules have swung since. In April 2025 Washington required licenses for the H20, a chip Nvidia had cut down for China, costing Nvidia a $4.5 billion charge. In December 2025 it agreed to allow sales of the more powerful H200 to approved buyers, with a quarter of the revenue going to the American government. On May 14, 2026 the Commerce Department licensed about ten Chinese firms, including Alibaba, Tencent, ByteDance, and JD.com, to buy up to 75,000 each.

Beijing pulled the other way, discouraging purchases of Nvidia's chips and steering buyers to domestic ones. In July 2026 it was reported to be approving fewer than 200,000 H200s in all, for training only. Nvidia's forecast for the quarter to late October assumed no revenue at all from data-center computing chips in China.

Huawei leads China's alternatives with its Ascend chips, made by SMIC. Its 2026 parts are the Ascend 950PR, for prefill, with 128 GB of Huawei's own HBM, and the 950DT, due in the fourth quarter, with 1 petaflop of 8-bit arithmetic, about half an H100. Huawei makes up for weaker chips with scale: its CloudMatrix 384 links 384 chips across 16 racks by optical cables, and its Atlas 950 SuperPoD, due in the fourth quarter of 2026, joins 8,192. In September 2026 Huawei pulled its roadmap forward: the Ascend 960DT is due in the first quarter of 2027, three quarters earlier than planned, the 960PR in the third quarter, and the 970 in 2028.

The gap remains large. Epoch AI estimated in September 2026 that Huawei would make about 1.5 million Ascend chips in 2026, up from about 800,000 in 2025, against about 5.9 million Nvidia accelerators, and that Huawei's chips would supply under 4% of Nvidia's computing power. SMIC's best process packs transistors about half as densely as TSMC's 3nm, and China's stockpile of foreign-made HBM was expected to run out during 2026. Cambricon, another Chinese designer, more than doubled its revenue to 6.0 billion yuan in the first half of 2026, by Caixin's report.

Worked example — Dollars per Petaflop, and What "Sparse" Does

Nvidia and AMD publish no prices, so use reported ones, with the specifications in the table in 8b. An H100 sold for about $25,000 to $31,000 in 2026; take $28,000. For its 2 dense 8-bit petaflops that is $14,000 a petaflop, and for its 3.35 TB/s of bandwidth about $8,400 per TB/s. A B200 sold for about $30,000 to $50,000; take $40,000. For its 5 dense 8-bit petaflops that is $8,000 a petaflop, and for 8 TB/s, $5,000 per TB/s. So one generation cut the price of arithmetic by about 43% and of bandwidth by about 40%. A whole GB300 NVL72 rack sold for just under $5.0 million in purchase orders reported by Wing Venture Capital in August 2026, about $69,000 per GPU once the CPUs, switches, and cooling are counted. On its 15 dense 4-bit petaflops per GPU that is about $4,600 a petaflop; at 8 bits, where it does no more than the B200's 5 petaflops, it is about $13,800. No reliable prices for Rubin or the MI455X were public by October 2026. Now the trap. Nvidia also rates the B200 at 20 petaflops at 4 bits with sparsity. Against the H100's 2 dense 8-bit petaflops that looks like a tenfold gain, but halving the precision doubles the figure and sparsity doubles it again (Module 5 and 8b). Compared fairly, at 8 bits and dense, the gain is 2.5 times, 5 against 2. A buyer comparing chips should line them up at the same precision, without sparsity, and per dollar and per watt, not per chip.

Module 8 Key Points

AI Accelerators: Nvidia, AMD, and the Challengers

• Prefill reads a whole prompt at once and is limited by arithmetic; decode writes one token at a time and is limited by memory.

• Compare dense figures at one precision: a 4-bit figure with sparsity is four times the same chip's dense 8-bit figure.

• Nvidia began shipping Vera Rubin in August 2026 and expected it to bring about 20% of that quarter's data-center revenue.

• AMD's MI455X carries 432 GB of HBM4, half again as much as Rubin's 288 GB.

• Epoch AI estimates Huawei will make about 1.5 million Ascend chips in 2026, under 4% of Nvidia's computing power.

Module 8 Quiz

Check Your Understanding

1. Which phase of answering a prompt is limited mainly by memory bandwidth?

A) Prefill, reading the prompt
B) Training the model
C) Decode
D) Tokenizing the text

2. What does "with sparsity" mean in a chip's headline figure?

A) It counts idle time
B) It is measured on 64-bit numbers for science
C) It is the speed reached when only a few cores are switched on, to save power
D) It assumes half the weights are zero

3. How much HBM4 does AMD's MI455X carry, by AMD's figures?

A) 192 GB
B) 288 GB, like Rubin
C) 432 GB, half again Rubin's
D) 1 TB, enough to hold any model in use today

4. What did Nvidia gain from its December 2025 deal with Groq?

A) A license to Groq's inference technology, plus its founder.
B) Full ownership of Groq, which became a division of Nvidia under its old brand.
C) The right to make Groq's chips in its own factories in Taiwan.
D) A seat on Groq's board in exchange for supplying it with Blackwell GPUs.

5. Why can a headline of 20 petaflops at 4 bits with sparsity mislead?

A) Because 4-bit numbers are too coarse for any model.
B) Because sparsity doubles the figure only when half the weights are zero, and a 4-bit operation does less work than an 8-bit one.
C) Because petaflops are measured on CPUs, not GPUs.
D) Because Nvidia counts each chip twice in its racks.

Module 9 — Custom Silicon: TPUs, Trainium, and the Clouds' Own Chips

Why the Largest Buyers Design Their Own

Recommended Reading / Reference

Paper: Norman P. Jouppi and others, “In-Datacenter Performance Analysis of a Tensor Processing Unit” (2017), Google's own account of its first TPU and why it built one

Reference: Google Cloud's documentation of its TPU system architecture, with the specifications of each TPU generation

9a. Why Build a Chip for One Job

A GPU is built to sell to everyone, so it carries features that any one buyer may not need: 64-bit arithmetic for scientists, remnants of graphics hardware, and support for every kind of model. A custom chip, called an application-specific integrated circuit (ASIC), is built for one company's work; in the AI trade it is often called an XPU, a catch-all name for custom accelerators. It keeps the matrix engines, memory, and links that the company's own models use and drops the rest, so it can do the same work on less silicon and less power.

The bigger saving is the margin. Nvidia's gross margin (Module 14) means that a buyer of its data-center chips pays roughly four times what they cost to make. A company that designs its own pays its design partner and TSMC instead, and it can tune the chip to its own software, networks, and data centers, as Google has done for a decade. It also gains a second source of supply when GPUs are scarce.

The costs are steep. Reuters reported in February 2025, citing industry sources, that one version of a leading-edge AI chip could cost about $500 million to design, and that the software and supporting hardware around it could double that. A design takes two to three years, may fail or arrive late, and needs software of its own to run (Module 12). So only companies that buy chips by the hundreds of thousands build their own, and nearly all of them work with a design partner, such as Broadcom or Marvell, which handles the physical layout, the packaging with HBM, and the dealings with TSMC (9f).

9b. The Systolic Array

Google's TPU, or tensor processing unit, is built around a systolic array, an idea published by the computer scientists H. T. Kung and Charles Leiserson in 1978. It is a grid of multiply-accumulate units wired to their neighbors. Numbers flow in from two edges and pulse through the grid step by step, the way blood pulses through the body, which gives the design its name. Each unit multiplies the two numbers passing through it, adds the product to a running sum, and passes the numbers on.

The design's strength is reuse. Each number is read from memory once and then used by every unit along its row or column, which gives the array a very high arithmetic intensity (Module 3) and needs almost no control logic. Google's matrix units were grids of 128 by 128 in most generations, 16,384 multiply-accumulates in each tick, and 256 by 256 in Trillium and Ironwood, 65,536. The array does one thing, multiply matrices, extremely efficiently; anything else goes to the TPU's vector and scalar units.

Google's first TPU used a 256-by-256 grid of 8-bit units. Google reported in 2017 that on its own inference work it ran 15 to 30 times faster than the CPUs and GPUs of the day, with 30 to 80 times the operations per

watt. Those are Google's figures for its own workloads, but they showed what a chip built for one job could do.

9c. Google's TPUs

Google has built a TPU generation every year or two since, adding training, liquid cooling, and ever larger pods, groups of chips linked into one machine. Its pods use optical circuit switches, which reroute beams of light between blocks of 64 chips, so a pod can be reshaped or a failed block bypassed without rewiring (Module 10). The table traces the line, by Google's figures.

Generation	Introduced	Peak per chip	Memory per chip	Largest pod
TPU v1	In service 2015; disclosed 2016	92 trillion 8-bit integer operations a second (inference only)	8 GiB DDR3	A single card
TPU v2	2017; the first for training	45 teraflops (BF16)	16 GiB HBM	256 chips
TPU v3	2018; liquid-cooled	123 teraflops (BF16)	32 GiB HBM	1,024 chips
TPU v4	In service 2020; announced 2021	275 teraflops (BF16)	32 GiB HBM	4,096 chips
TPU v6e, Trillium	Announced May 2024	918 teraflops (BF16)	32 GiB HBM	256 chips
TPU7x, Ironwood	Announced April 2025; launched November 2025	4,614 teraflops (FP8)	192 GiB HBM at 7.37 TB/s	9,216 chips
TPU 8t (training)	Announced April 22, 2026; for customers later in 2026	12.6 petaflops (FP4)	216 GB HBM3E at 6.5 TB/s	9,600-chip superpods
TPU 8i (inference)	Announced April 22, 2026; for customers later in 2026	10.1 petaflops (FP4)	288 GB HBM3E at 8.6 TB/s, plus 384 MB of SRAM	1,024 chips

Google's figures. A gibibyte (GiB) is 2 to the 30th bytes, about 7% more than a gigabyte. The v5 generation of 2023 is left out for brevity. Precisions differ by row (Module 5), so the figures are not directly comparable.

Ironwood showed the strategy. Each chip is modest beside Nvidia's, but a full pod joins about 42.5 exaflops of 8-bit arithmetic, by this course's arithmetic, and nearly 2 petabytes of HBM. With its eighth generation Google split the line in two, the 8t for training and the 8i for inference. Both are built on TSMC's 3nm family and hosted by Google's own Axion processors (Module 7). Google kept to HBM3E rather than the newer HBM4, by trade reports for reasons of cost and supply, and bets on scale rather than speed per chip.

Broadcom has designed TPUs with Google since 2015 and, by a filing of April 2026, is to supply them through at least 2031. MediaTek joined as a second design partner in December 2025 and is reported to design the 8i, though Broadcom's own remarks in September 2026 spoke of shipping that chip; the division of work was not public.

TPUs now reach other companies. In October 2025 Anthropic agreed to use up to a million of them, with well over a gigawatt of capacity coming online in 2026, a deal worth tens of billions of dollars. In April 2026 it agreed with Google and Broadcom on multiple gigawatts more from 2027, and Broadcom told investors in September that Anthropic would add 5 gigawatts of eighth-generation TPUs in 2027 and 10 more in 2028. Meta agreed to rent TPUs through Google's cloud, The Information reported in February 2026.

By October 2026 Google was also placing TPU systems in customers' own data centers and in other companies' facilities, with Oracle and Crux AI as new partners. On Alphabet's results call of October 28, the chief executive of Google Cloud, Thomas Kurian, said nearly 100 customers used TPUs and named Apple

among those training their models on them. Alphabet's finance chief expected sales of TPU systems to bring a meaningful rise in revenue in 2027.

9d. Amazon's Trainium

Amazon designs its chips through Annapurna Labs, an Israeli start-up it bought in 2015, which also makes its Graviton CPUs (Module 7). Its AI chips began with Inferentia, for inference, announced in 2018, and Trainium, for training, announced in 2020. Trainium2 became generally available in December 2024, and Amazon built its largest cluster from it: Project Rainier in Indiana, about 500,000 Trainium2 chips for Anthropic, switched on in October 2025 (Module 10). Anthropic calls Amazon its primary cloud and training partner.

Trainium3, Amazon's first 3nm AI chip, became generally available on December 2, 2025. Each chip delivers 2.52 petaflops at 8 bits with 144 GB of HBM3E at 4.9 TB/s, and an UltraServer joins 144 of them. AWS claims 4.4 times the performance of the Trainium2 servers and 4 times the performance per watt, about half of the gain coming from putting more chips in each server, as The Register noted. Trainium4, shown in December 2025 with no date, will connect through Nvidia's NVLink Fusion, the version of Nvidia's chip link that it licenses to other companies' chips (Module 10).

Amazon, too, plans to sell its chips to others. On its results call of October 29, 2026, its chief executive, Andy Jassy, said revenue commitments for Trainium and for Graviton had each grown more than twentyfold in a year, and that Amazon intended to offer Trainium outside AWS from late 2027, a business he expected to become very significant.

9e. Microsoft, Meta, OpenAI, and Others

Microsoft came late. Its first AI chip, Maia 100, announced in November 2023, ran only internal work at limited scale. Its second, Maia 200, unveiled on January 26, 2026 for inference, is made on TSMC's 3nm process and delivers more than 10 petaflops at 4 bits and 5 at 8 bits. It carries 216 GB of HBM3E at 7 TB a second and 272 MB of SRAM, and draws 750 watts, by Microsoft's figures.

Microsoft claims Maia 200 has three times the 4-bit speed of Trainium3 and more 8-bit speed than Ironwood, and 30% better performance per dollar than the newest hardware in its own fleet. It runs in Microsoft's data centers near Des Moines, Iowa, with Phoenix next. Microsoft's chief technology officer, Kevin Scott, said in October 2025 that the company aims in the long run to use mainly its own chips.

Meta's chips, the Meta Training and Inference Accelerator (MTIA) line, began in 2023 with chips that rank posts and advertisements in its apps. In March 2026 it set out four generations in two years. The MTIA 300 was in mass production for ranking and recommendation, and the MTIA 400, for generative AI as well, was being readied for racks of 72 chips. The MTIA 450 was due in early 2027 and the MTIA 500 later that year. Broadcom co-designs them, and in April 2026 Meta and Broadcom extended their partnership to 2029. Meta still buys Nvidia's and AMD's chips in huge numbers (Module 8 and 9g).

OpenAI agreed with Broadcom in October 2025 to deploy 10 gigawatts of accelerators of its own design by the end of 2029. On June 24, 2026 it unveiled the first, Jalapeño, a chip for inference that it says took nine months from design to tape-out, the moment a finished design goes to the factory. Broadcom said in September that it had shipped the first Jalapeño chips and expected OpenAI to deploy about 1.3 gigawatts of them in 2027 and more than 5 in 2028. The first deployment is to come with Microsoft.

Tesla disbanded the team behind Dojo, its training supercomputer, in August 2025 and now builds its own inference chips for cars and robots, the AI5 (Module 6) and the AI6, which Samsung is to make under a $16.5 billion contract signed in July 2025. Its chief executive, Elon Musk, says Tesla will also build a chip factory of its own, Terafab, in Austin, with test chips hoped for in mid-2027 and volume about 2028, by his own timetable. Apple runs its cloud AI on servers built from its own Mac chips and was reported in September 2026 to be considering larger servers that would link to Nvidia's racks.

9f. The Design Houses

Behind most custom chips stands Broadcom, the largest designer of custom chips for other companies. Its revenue from AI chips rose 221% from a year earlier, to $16.7 billion, in its quarter to August 2, 2026, and custom accelerators made up about 73% of it, by its own account; the rest was mostly networking chips (Module 10). It guided to about $21.7 billion of AI revenue the next quarter. Its chief executive, Hock Tan, set out a goal of about $115 billion of AI revenue in fiscal 2027, up from about $58 billion in fiscal 2026, by summaries of the call. Counterpoint Research expects Broadcom to hold about 60% of the market for custom AI chips through 2027.

Broadcom's custom customers include Google, Meta, Anthropic, which buys racks of TPUs through it, and OpenAI. Marvell is second: its data-center revenue was $2.17 billion in its quarter to August 1, 2026, and it expects its custom business to more than double in its fiscal 2028; AWS signed a five-year agreement with it in December 2024. MediaTek, Google's second TPU partner, expected about $1 billion of such sales in 2026. Alchip and GUC, both of Taiwan, take customers' designs through TSMC's factories and packaging.

9g. Contested: Will the Clouds' Own Chips Displace Nvidia's?

Those who say yes point to the money. A buyer paying Nvidia's margin on a hundred thousand chips a year can fund a design team many times over, and the builders claim the savings are real: Microsoft's figure for Maia 200 (9e), and AWS's claim in 2024 that Trainium2 gave 30-40% better price-performance than its GPU servers. They point to Anthropic's commitments to TPUs and Trainium, to Broadcom's forecast for its AI revenue (9f), to Google's and Amazon's moves to sell their chips to other companies (9c and 9d), and to Goldman Sachs's forecast that demand for custom chips will match demand for GPUs by 2027.

Those who say no point to Nvidia's pace and breadth. It ships a new generation every year, sells whole racks with their networks and software (Modules 10 and 12), and runs every model a customer might want, while a custom chip runs its designer's own. Custom programs have slipped often: Maia's second generation came years after its first, Meta's earlier training chip ran into trouble, and Apple's server chip was reported delayed. Nvidia says it is a generation ahead, and every builder of custom chips still buys Nvidia's: Meta agreed in February 2026 to buy millions of Nvidia GPUs, and Google rents Vera Rubin racks in its own cloud.

The evidence so far supports both, in different places. Custom chips are winning the inference of their owners' own largest models, where the work is known in advance and cost per token decides, and Anthropic shows that a lab can run frontier work on them. Nvidia's share of AI chip revenue is slipping only slowly (Module 14), because it keeps the frontier training runs and almost every buyer outside the largest few. The course's reading is that the clouds' own chips will take a growing share of the largest companies' work and press on Nvidia's prices and margins, but will not displace it in this decade.

Worked example — Build or Buy

Suppose a cloud company can buy a GPU for $40,000, roughly a B200's reported price (Module 8). Alternatively it can design its own chip for $1 billion, the Reuters estimate of $500 million doubled for software, and then pay $12,000 for each chip, covering TSMC, the HBM, packaging, and its design partner's margin. These are illustrative figures. If its chip does 70% of the GPU's useful work, which allows for less mature software, each GPU's worth of work costs it $12,000 divided by 0.7, about $17,100, a saving of about $22,900. The $1 billion is then repaid after about 44,000 GPUs' worth of work, roughly 62,000 of its own chips. A gigawatt of AI data centers holds several hundred thousand accelerators, and the largest buyers are building several gigawatts each, so for them the design cost is small. Change one assumption, though, and the picture shifts. If the chip does only 40% of a GPU's work, the saving falls to $10,000 and the break-even rises to 100,000 GPUs' worth. If it arrives a year late, against a GPU that has doubled in speed meanwhile, it may never pay back. So the risk in a custom chip lies less in its design cost than in its speed and its software.

Module 9 Key Points

Custom Silicon: TPUs, Trainium, and the Clouds' Own Chips

• A custom chip strips out what its owner does not need; one design can cost about $500 million, Reuters reported.

• Google's TPUs use systolic arrays that reuse each number many times; an Ironwood pod joins 9,216 chips.

• In October 2025 Anthropic agreed to use up to a million Google TPUs; Amazon built it a cluster of about 500,000 Trainium2s.

• Broadcom's AI chip revenue rose 221%, to $16.7 billion, in its quarter to August 2026.

• The clouds' chips will take a growing share of their owners' work, pressing on Nvidia's margins more than displacing it.

Module 9 Quiz

Check Your Understanding

1. What is a systolic array?

A) A grid of multiply-adders
B) A cache
C) A cooling system for TPU pods
D) A network linking chips in a pod

2. Which company has designed Google's TPUs with it since 2015?

A) MediaTek, from the start
B) Broadcom
C) Marvell Technology
D) Nvidia, under license

3. What did Anthropic agree with Google in October 2025?

A) To buy Google's chip unit.
B) To rent 10,000 TPUs for a year.
C) To use up to a million TPUs, with well over a gigawatt online in 2026.
D) To build its own TPUs in Google's factories, sharing the designs with Broadcom and MediaTek.

4. In the build-or-buy example, what decides whether a custom chip pays off?

A) The design cost alone
B) Whether TSMC can make it
C) Its clock speed
D) Volume, and how much of a GPU's useful work it does per dollar

5. How did Broadcom's AI chip revenue change in its quarter to August 2026?

A) It rose 221% from a year earlier, to $16.7 billion, mostly from custom accelerators.
B) It fell, as Google moved its TPUs to MediaTek.
C) It doubled, to about $5 billion.
D) It was flat, since customers waited for HBM4.

Module 10 — From Chip to Cluster: The Networks

Scale-Up, Scale-Out, and the Largest Clusters

Recommended Reading / Reference

Paper: Meta's Llama team, “The Llama 3 Herd of Models” (2024), whose sections on infrastructure describe training on 16,000 GPUs, its networks, and its failures

Reference: Epoch AI, its database of frontier AI data centers, with estimates of their chips and power from satellite images and filings

10a. Why One Chip Is Not Enough

Training a frontier model takes more arithmetic than any chip could do in a lifetime. Meta's Llama 3.1, with 405 billion parameters, took about 3.8 times 10^{25} operations to train, by Meta's account, and the frontier models of 2026 take more. A single GPU sustaining a petaflop would need more than a thousand years for that. The model would not fit on one chip either: its 405 billion parameters at 16 bits fill 810 GB, and training needs several times that again for the running statistics that adjust each parameter.

So the work is split across tens of thousands of chips, and those chips must constantly exchange results. Every step of training ends with the chips combining what each has learned, and the slowest link sets the pace for all of them. A cluster's network therefore matters as much as its chips. A chip that waits on the network does no useful work, and the share of time chips spend waiting is one of the main reasons real systems deliver well under their peak (Module 13).

10b. Four Ways to Split the Work

Engineers combine four ways of dividing a model across chips. In data parallelism, each group of chips holds a full copy of the model and trains on a different slice of the data; after each step the groups average their adjustments, an exchange called an all-reduce. In tensor parallelism, each layer's matrices are split across several chips, which must swap partial results many times within every step, so it needs the fastest links and usually stays inside one rack.

In pipeline parallelism, different layers sit on different chips, and batches of data flow through them like an assembly line, so each chip talks mainly to its neighbors. In expert parallelism, used for mixture-of-experts models (Module 3), different experts sit on different chips and each token is sent to the ones it needs, a pattern called all-to-all that loads every link at once. A large training run uses all four at once, mapped onto the network so that the heaviest traffic travels over the fastest links.

The same splits apply to inference. A model too big for one chip's memory is divided across a rack, and the chips must exchange results for every token they generate, which is why the size of the fast domain, the number of chips joined by the quickest links, has become a selling point. Nvidia's rack of 72 GPUs, Google's pods, and AMD's Helios are each attempts to make that domain larger.

10c. Scale-Up: Inside the Rack

Links within a rack, called scale-up links, join chips closely enough that they can share each other's memory. Nvidia's NVLink has roughly doubled with each generation: 600 GB a second per GPU in the A100 of 2020, 900 in the H100, 1.8 TB a second in Blackwell, and 3.6 TB a second in Rubin, counting both directions. For comparison, the standard PCI Express link that connects a card to a computer carries about 128 GB a second in its fifth version.

Nvidia's NVL72 racks join all 72 GPUs to each other through NVLink switch chips. In the GB200 rack, 18 switch chips in nine trays connect the GPUs over a spine of copper cables in the back of the rack, giving about 130 TB a second in all; the Vera Rubin rack reaches about 260 TB a second, by Nvidia's figures. Copper is used because it needs no power-hungry lasers over such short distances, and that choice is one reason AI racks are packed so densely (Module 11).

NVLink belonged to Nvidia alone until May 2025, when it announced NVLink Fusion, a version other companies can build into their chips. Amazon's Trainium4 (Module 9) and the custom x86 CPUs that Intel is to build for Nvidia will use it. The industry's open alternative is UALink, short for Ultra Accelerator Link, backed by a consortium founded in May 2024 by AMD, Broadcom, Cisco, Google, HPE, Intel, Meta, and Microsoft, with Alibaba, Apple, and AWS joining its board in 2025.

UALink's first specification, in April 2025, allows up to 1,024 accelerators in one domain, and its second, on April 7, 2026, added the ability to compute inside the switches. The consortium's chair, Kurtis Bowman, said the first UALink chips would reach laboratories in the second half of 2026 and products later in 2027, and conceded that the first versions would not fully match NVLink. AMD's Helios meanwhile carries the UALink protocol over Ethernet hardware (Module 8). Google takes its own route, linking TPUs in three-dimensional grids reshaped by optical switches (Module 9).

10d. Scale-Out: Between Racks

Links between racks, called scale-out networks, join thousands of racks into one cluster. Each GPU gets a network card of its own, 800 gigabits a second in the Blackwell generation and 1.6 terabits in Rubin's, and the cards feed layers of switches. The traffic uses remote direct memory access (RDMA), which lets one chip read another's memory across the network without either computer's CPU taking part, so data moves with little delay.

Two technologies compete. InfiniBand, built for supercomputers, was long the choice for AI clusters, and Nvidia owns its leading maker, Mellanox, which it bought for $6.9 billion in 2020. Ethernet, the standard of ordinary data centers, has caught up through RDMA over Converged Ethernet (RoCE) and through switch chips from Broadcom, whose Tomahawk 6, shipping since June 2025, carries 102.4 terabits a second. The Ultra Ethernet Consortium published a specification in June 2025 to tune Ethernet for AI and supercomputing.

The evidence points to Ethernet at the largest scales. Meta's Prometheus cluster joins its buildings over Ethernet, the largest system ever entered in the MLPerf inference benchmark ran on RoCE (Module 13), and Nvidia's own Ethernet line, Spectrum-X, grew 2.6 times in a year, its finance chief said in August 2026. Nvidia sells both, and its networking revenue, counting NVLink, reached $14.8 billion in its quarter to April 2026, up 199% from a year earlier, by its own report.

Every link between racks needs a laser and a detector at each end, in plug-in optical modules that consume a large share of a network's power. The fix now arriving is co-packaged optics, which moves the optics into

the same package as the switch chip. Nvidia claims its co-packaged switches use 3.5 times less power than plug-in optics and said in May 2026 that its Ethernet version was in production; Broadcom sells a co-packaged Tomahawk 6. Putting optics inside NVLink itself was reported delayed in July 2026.

10e. The Largest Clusters of 2026

The largest AI clusters are now measured in gigawatts of power and hundreds of thousands of chips. The table gives the leading ones as of October 2026. Most of the figures come from the owners or from estimates, because the companies publish little and enter none of these machines in public rankings (Module 13).

Cluster	Owner and place	Chips	Power	Source and status
Colossus 2	xAI, Memphis, Tennessee	About 550,000 Nvidia GB200 and GB300 GPUs	About 1 gigawatt, in a site planned for about 2	Elon Musk's statements to January 2026
Stargate, Abilene	OpenAI, with Oracle and the developer Crusoe, Texas	About 400,000-450,000 GB200s, by press reports	1.2 gigawatts across eight buildings when complete	First buildings live from October 2025
Prometheus	Meta, New Albany, Ohio	"Hundreds of thousands" of GPUs, growing toward about 1 million	Aiming at 1 gigawatt	First phase running in October 2026, by Meta
Project Rainier	Amazon, for Anthropic, New Carlisle, Indiana	Trainium2 chips (Module 9)	About 1 gigawatt of computing load by October 2026	Epoch AI's estimate from satellite images
Fairwater	Microsoft, Mount Pleasant, Wisconsin, linked to a second site near Atlanta	"Hundreds of thousands" of GB200s, by Microsoft	Not disclosed	Announced September 2025; online from early 2026

Counts and power figures are the companies' statements or outside estimates, rounded; they mix chips installed with chips planned. Epoch AI also estimates that Rainier's chips, at their October 2026 size, equal about 818,000 H100s in computing power.

10f. Failure as Routine

With so many parts, something is always breaking. Meta's report on training Llama 3 counted 466 interruptions to its training job in a 54-day period on up to 16,000 H100 GPUs: 47 planned and 419 unexpected. About 78% of the unexpected ones were confirmed or suspected hardware faults. Faulty GPUs caused 148 of the 419 and their HBM memory 72, and software bugs, network switches and cables, and host machines caused most of the rest. Six were silent data corruption, wrong answers with no error signal at all.

That is one unexpected interruption about every three hours, or roughly one for every 50,000 hours of GPU time, by this course's arithmetic. Meta reported a similar figure, about 47,000 GPU-hours between interruptions, for its Prometheus cluster in October 2026. At that rate a job on 100,000 GPUs would be interrupted about every half hour. Even so, automated recovery kept Llama 3's effective training time above 90%, with only three interruptions needing significant manual work.

The defenses are now standard. Training jobs save checkpoints, snapshots of the model's state, every few minutes so a failure loses little work. Spare servers stand ready to replace failed ones, and software spots slow or faulty chips before they spoil a run. Meta reported that its Prometheus jobs spent more than 85% of their time doing useful work in pre-training. The scale brings odder problems too: Llama 3's throughput

varied by 1-2% between day and night with the temperature, and thousands of GPUs pausing together could swing the data center's power draw (Module 11).

Worked example — Sizing a Frontier Training Run

Suppose a lab wants to train a model with 10^{27} operations, about 26 times the compute of Llama 3.1 405B, on Blackwell Ultra GPUs. Each is rated at 5 dense 8-bit petaflops (Module 8). Real runs sustain perhaps 40% of peak once the time spent waiting on the network, on memory, and on recoveries is counted, so each GPU delivers about 2 petaflops of useful work. A run of 100 days is 8.64 million seconds, so the lab needs 10^{27} divided by (2 times 10^{15} times 8.64 million), about 58,000 GPUs. At about 2.5 kilowatts per GPU, counting its share of the CPUs, network, and cooling (Module 11), the cluster draws about 145 megawatts. Over 100 days that is about 350 gigawatt-hours, roughly $28 million of electricity at 8 cents a kilowatt-hour. The hardware is the larger bill: at about $69,000 per GPU in a GB300 rack (Module 8), 58,000 GPUs cost about $4 billion. Two levers stand out. If better networking and software raised the sustained share from 40% to 50%, the same run would need a fifth fewer GPUs, about 11,600, saving about $800 million of hardware. And at Prometheus's rate of failure (10f), 58,000 GPUs would suffer an interruption about every 50 minutes, so fast, automatic recovery is part of the design, not an afterthought.

Module 10 Key Points

From Chip to Cluster: The Networks

• A frontier training run splits the work four ways: by data, within layers, across layers, and among experts.

• NVLink roughly doubled each generation, to 3.6 TB a second per GPU in Rubin, joining 72 GPUs as one machine.

• Nvidia's networking revenue, counting NVLink, reached $14.8 billion in its quarter to April 2026, up 199%.

• Ethernet is overtaking InfiniBand at the largest scales; Meta's Prometheus joins its buildings over Ethernet.

• Llama 3's training saw 419 unexpected interruptions in 54 days, mostly hardware faults, about one every three hours.

Module 10 Quiz

Check Your Understanding

1. Why is tensor parallelism usually kept inside one rack?

A) Because racks cannot hold more than eight GPUs.
B) Because splitting each layer across chips needs the fastest links, which reach only within a rack.
C) Because it needs no communication between chips at all.
D) Because power rules forbid linking racks, and Ethernet cannot carry the data needed by any model larger than a few billion parameters.

2. What is UALink?

A) Nvidia's new name for NVLink.
B) A form of InfiniBand built by Mellanox for supercomputers in the 2000s.
C) A satellite link that connects data centers across oceans, built by a group of cloud companies.
D) An open standard for linking AI chips in a rack.

3. In Meta's Llama 3 report, what caused most of the unexpected interruptions?

A) Software bugs in the training code
B) Power cuts at the data center, from storms and grid trouble across the summer
C) Hardware
D) Network switches and cables

4. About how much was Nvidia's networking revenue in its quarter to April 2026?

A) About $1.5 billion
B) $14.8 billion
C) About $48 billion
D) It was not reported, since Nvidia counts all networking inside its gaming and edge products

5. Why do large training runs save checkpoints often?

A) To free memory
B) To satisfy auditors
C) So failures cost minutes, not days
D) To let each GPU run a different model, so that the results can be compared at the end

Module 11 — Power and Heat

Watts, Water, and the 100-Megawatt Hall

Recommended Reading / Reference

Report: Lawrence Berkeley National Laboratory, *2024 United States Data Center Energy Usage Report* (December 2024), the standard estimate of how much electricity American data centers use and why

Reference: Uptime Institute, *Global Data Center Survey 2026* (July 2026), the annual survey of data-center operators on power, density, cooling, and efficiency

11a. Where the Watts Go

Every operation a chip performs costs a little energy, and moving data costs far more than computing with it. Mark Horowitz of Stanford University estimated in 2014, for chips of that time, that multiplying two 32-bit numbers took under 4 picojoules, trillionths of a joule, while fetching one 32-bit number from DRAM took about 640. The ratio has shifted with newer processes, but its lesson holds: most of a modern chip's power goes into moving bits, between memory and the arithmetic units and between chips, not into the arithmetic itself.

That is why the techniques of earlier modules double as energy savings. Caches keep data close (Module 3), systems on a chip avoid trips between chips (Module 6), systolic arrays reuse each number many times (Module 9), and lower precision moves fewer bits for each number (Module 5). The industry's figure of merit has shifted from operations per second to operations per joule, or for AI tokens per joule, because power is now the limit that binds.

The chips themselves draw more each generation. Nvidia's data-center chips have roughly tripled in power in four years, from the H100 of 2022 to Rubin (Module 8). A desktop gaming card draws a few hundred watts at most (Module 5), a laptop processor tens, and a phone's a few. Chips draw more because a data center with limited space can earn more from each rack by packing in more computing, and because the power wall (Module 2) is now met by spending power rather than saving it.

11b. Racks: From 11 Kilowatts to 600

The rack, a cabinet about 2 meters tall holding servers, is the unit a data center plans around. The Uptime Institute's 2026 survey found that the typical rack drew a little over 11 kilowatts on average, up from 9 in 2025, and 7.8 kilowatts leaving out the most densely packed sites. About a quarter of operators reported their densest racks at 30 kilowatts or more, and about one in seven at 50 or more.

AI racks are another category. Industry estimates put Nvidia's GB200 NVL72 at about 120 to 132 kilowatts and the GB300 version at about 135 to 140. The Vera Rubin rack is estimated at about 190 to 230 kilowatts; Nvidia's own guide lists four power shelves of 110 kilowatts each, which is the supply's capacity rather than the rack's draw. The Kyber rack planned for Rubin Ultra would draw about three times as much as a Vera Rubin rack (Module 8). One AI rack of 2026 can therefore draw as much as 10 to 20 ordinary ones in the same floor space, and a Kyber rack about 50.

The density is deliberate. Nvidia's copper NVLink spine reaches only about a meter or two (Module 10), so the 72 GPUs that must work as one machine have to sit in one rack. Packing them that tightly saves the cost

and power of optical links, and it moves the problem to power delivery and cooling, which most existing data centers were not built for. Many operators have had to rebuild halls or build new ones to take AI racks.

11c. Air, Liquid, and Water

Fans and cold air can cool a rack of a few tens of kilowatts. Beyond that, air cannot carry heat away fast enough, and the industry has moved to liquid. In direct-to-chip cooling, metal cold plates sit on each processor and coolant flows through them, carrying the heat to a coolant distribution unit and then out of the building. Nvidia's GB300 rack is about 90% liquid-cooled, and the Vera Rubin rack is cooled by liquid entirely, with no option for air. Immersion, sinking whole servers in a non-conducting fluid, remains a niche.

The Vera Rubin rack accepts inlet water as warm as 45 degrees Celsius, about 113 degrees Fahrenheit. Water that warm can be cooled by outdoor air in most climates without energy-hungry chillers, and it leaves warm enough to heat buildings: Europe's JUPITER supercomputer, cooled by warm water, feeds its waste heat to its campus's heating network.

Water is a concern because many data centers cool themselves by evaporating it. Google's data centers consumed 10.9 billion gallons in 2025, 34% more than a year earlier, while its electricity use rose 37%, its largest annual rise, by its 2026 environmental report. Closed-loop liquid systems, such as those in Meta's Prometheus cluster, circulate the same coolant and use little water on site. The trade-off is that dry cooling uses more electricity on hot days, so siting a data center means choosing between water and power.

11d. Power Delivery and 800 Volts

Getting power into a 200-kilowatt rack is itself hard. Today's racks take alternating current from the building and convert it to direct current at about 50 volts, and at that voltage 200 kilowatts means about 4,000 amperes, which needs thick copper bars and wastes power as heat. Nvidia proposed in May 2025 to distribute direct current at 800 volts instead, cutting the current for the same power by a factor of about 16.

By August 2026 Nvidia said it had developed the design with Google and Microsoft through the Open Compute Project, an industry group, and that more than 80 companies were building equipment to the specification. It expected an 800-volt power rack in the second half of 2026 and a power center for a whole row of racks, up to 2 megawatts, in 2027. The first Vera Rubin racks ship with conventional power shelves, with 800 volts an option for operators preparing for Rubin Ultra.

Power also swings. When thousands of GPUs finish a training step together and wait for the network, the draw drops, then jumps when the next step starts, by tens of megawatts in a large cluster (Module 10), which can disturb the grid. Batteries and smarter power supplies smooth the swings: Nvidia says Rubin's power system cuts average power by about 10% and brief 50-millisecond peaks by about 20%. Grid operators have begun writing rules for such loads.

11e. Power Usage Effectiveness

Power usage effectiveness (PUE) is a data center's total power divided by the power that reaches its computers. A PUE of 1.0 would mean no power lost to cooling, conversion, or lighting. The average across the Uptime Institute's 2026 survey was 1.52, and 1.45 for respondents' largest sites, with almost no improvement in seven years, as older buildings dominate the average. Google reported a fleet-wide PUE of 1.09 for 2025, and new liquid-cooled AI halls aim for similar figures.

PUE has limits as a measure. It says nothing about how much useful work the computers do with the power they get, and a hall of old, inefficient servers can score well. For AI, the measure that matters is the energy per token or per training run, which depends far more on the chips, the precision, and the software than on the building. A new GPU generation that doubles tokens per watt does more for efficiency than any cut in PUE now possible.

11f. The Energy in a Prompt

Estimates of the energy in a single AI query have converged on fractions of a watt-hour for ordinary text. Google reported in August 2025 that a median text prompt in its Gemini apps used 0.24 watt-hours of energy, 0.03 grams of carbon-dioxide emissions, and 0.26 milliliters of water. It counted idle machines, host processors, and cooling, and said the energy per prompt had fallen 33-fold in a year. These are Google's own measurements, not independently checked.

OpenAI's chief executive, Sam Altman, put an average ChatGPT query at about 0.34 watt-hours in June 2025, and Epoch AI estimated about 0.3 for a typical query to one of OpenAI's models that February. A third of a watt-hour runs a 10-watt light-emitting diode (LED) bulb for about two minutes. Long reasoning answers, images, and above all video take far more, so the energy per task varies by orders of magnitude with what is asked.

The totals are what strain the grid. Data centers of every kind used about 485 terawatt-hours of electricity in 2025, about 1.5% of the world's electricity, and the International Energy Agency (IEA) projected in April 2026 that they would use about 950 terawatt-hours by 2030, with AI the largest driver. Falling energy per query and rising queries pull against each other, and so far the queries are winning.

Worked example — What a 100-Megawatt Hall Holds

Take a new data hall with a grid connection of 100 megawatts and a PUE of 1.3, typical for a new liquid-cooled build. Its computers get 100 divided by 1.3, about 77 megawatts. Vera Rubin racks drawing about 200 kilowatts each, an estimate (11b), mean about 385 racks, holding about 27,700 GPUs. The hall uses 100 megawatts for 8,760 hours a year, 876 gigawatt-hours, which at 8 cents a kilowatt-hour costs about $70 million a year. Divided among its GPUs, the facility draws about 3.6 kilowatts per GPU, so each GPU-hour uses about 3.6 kilowatt-hours, about 29 cents of electricity. That is a small share of what such a GPU rents for (Module 13), most of which pays for the hardware itself. Two lessons follow. For an AI cloud, the scarce thing is not cheap electricity but power that can be connected at all, since a hall's grid connection caps how many GPUs it can run. And as the chips grow more efficient, a fixed connection holds more computing, which is why operators upgrade chips in old halls rather than wait for new grid connections that can take years.

Module 11 Key Points

Power and Heat

• Fetching a number from DRAM took over a hundred times the energy of multiplying it, so moving data dominates a chip's power.

• A typical data-center rack drew about 11 kilowatts in 2026; Nvidia's AI racks draw over 100, with 600 planned for 2027.

• Vera Rubin racks are cooled entirely by liquid, taking inlet water as warm as 45 degrees Celsius.

• The industry's average power usage effectiveness was 1.52 in 2026, against Google's fleet-wide 1.09.

• Google measured a median Gemini text prompt at 0.24 watt-hours in August 2025.

Module 11 Quiz

Check Your Understanding

1. Why are AI racks moving to 800-volt direct current?

A) Because GPUs will not run at any voltage below 800.

B) Because utilities charge less for power delivered at higher voltages, whatever the rack's size or the distance.

C) Because 800-volt power needs no cooling, which lets data centers remove their water systems entirely.

D) To cut the current, and so the copper, needed for such power.

2. What does power usage effectiveness measure?

A) A data center's total power divided by its computers' power

B) Chips' speed per watt

C) Water used for each prompt

D) The price of each kilowatt-hour

3. What did Google report for a median Gemini text prompt in August 2025?

A) About 3 watt-hours, ten times a web search, and half a liter of water.

B) About 0.24 watt-hours, but rising 33 times over the previous year as models grew larger.

C) About 0.24 watt-hours of energy.

D) About 24 watt-hours, enough to charge a phone several times.

4. What did the Uptime Institute's 2026 survey find about typical racks?

A) About 600 kW

B) About 130 kW, like Nvidia's

C) Under 2 kW

D) A little over 11 kW on average, up from 9 kW in 2025

5. In the worked example, about what does electricity add to each GPU-hour in a 100-megawatt hall?

A) About 29 cents

B) About $3

C) About $1.50, half the rent

D) About $29, more than the GPU's rent, since power is the largest cost of running AI

Module 12 — Software: CUDA and the Moat

Programming Models, Libraries, and the Fight to Run Anywhere

Recommended Reading / Reference

Book: Wen-mei W. Hwu, David B. Kirk, and Izzat El Hajj, *Programming Massively Parallel Processors: A Hands-on Approach* (fourth edition, 2022), the standard introduction to programming GPUs with CUDA

Reference: SemiAnalysis, InferenceX, an open benchmark run every night that compares AI chips serving models on each vendor's software

12a. What a Programming Model Is

A CPU's instruction set is a contract that lets old programs run on new chips (Module 1). GPUs needed a contract of their own, and a harder one, because their instruction sets change with almost every generation. A programming model supplies it: a stable way of describing parallel work that the vendor's software translates for whatever chip is present. Nvidia's CUDA (Module 5) compiles programs to an intermediate form, which the driver turns into instructions for the particular GPU when the program runs.

That design is why CUDA spread. A program written for a GeForce card in 2012 runs, recompiled, on a Rubin GPU in 2026 and gets faster without being rewritten. Nvidia also kept CUDA on every product it sold, from cheap gaming cards to supercomputers, so students, researchers, and companies all learned the same tools. Twenty years of that consistency built up a body of code, skills, and habits that is the core of what investors call Nvidia's moat.

12b. Kernels, Threads, Blocks, and Memory

In CUDA the programmer writes a kernel, a function that thousands of threads run at once, each on its own piece of the data. The threads are launched in a grid of blocks. A block holds up to 1,024 threads that run on the same streaming multiprocessor (Module 5), can share a fast on-chip memory, and can wait for one another. Different blocks run independently, in any order, so the same program spreads across a small GPU or a large one without change, simply running more blocks at once on the larger chip.

The memory follows the same layers. Each thread has its own registers. Each block shares a scratchpad on its streaming multiprocessor, up to 228 KB in the H100, and all the threads share a large second-level cache, 50 MB in the H100, and then the HBM. Most of the skill in writing a fast kernel lies in moving data through those layers so that the arithmetic units never wait. A matrix multiplication, for instance, loads a tile of each matrix into the shared scratchpad and reuses it many times, raising its arithmetic intensity (Module 3).

The biggest gains often come from fusing steps. Running a model's operations one at a time sends every intermediate result out to HBM and back. A fused kernel keeps it on the chip. FlashAttention, published by Tri Dao and colleagues in 2022, did this for attention, the step in a transformer where each token weighs every other, and it ran that step two to four times faster. Such kernels are why the same chip can run a model much faster a year after it ships.

12c. The Libraries and the Serving Stack

Few AI developers write CUDA themselves. They use libraries Nvidia built on it: cuBLAS for linear algebra, cuDNN, from 2014, for the building blocks of neural networks, and NCCL, the Nvidia Collective Communications Library, from about 2016, for the exchanges among GPUs in a cluster (Module 10). For inference, TensorRT-LLM, from 2023, tunes models for Nvidia's chips, and Dynamo, released as open source in March 2025, splits prefill and decode across GPUs (Module 8).

Above the libraries sit the frameworks in which models are written. PyTorch, begun at Meta and run since 2022 by a foundation under the non-profit Linux Foundation, is the most widely used, and JAX, from Google, is the main route to TPUs. Models are then served by inference engines, of which two open-source ones dominate: vLLM, which grew out of research at Berkeley in 2023, and SGLang, from 2024. Nvidia says more than 7.5 million developers use CUDA and its other tools, by its annual report for the year to January 2026.

12d. The Alternatives

Every rival has built its own stack. AMD's ROCm is open-source software for its GPUs, with a language called HIP (Heterogeneous-computing Interface for Portability) that closely mirrors CUDA and tools that translate CUDA code. Its seventh version brought large gains: in the MLPerf inference results of September 2026, software updates alone raised the MI355X's throughput on one large model by 28% to 38%, by AMD's own submissions. Intel's oneAPI is built on SYCL, an open standard, and Intel gave it to the UXL Foundation, set up under the Linux Foundation in 2023 to offer a vendor-neutral alternative to CUDA.

Google's compiler, XLA (Accelerated Linear Algebra), turns JAX and PyTorch programs into code for TPUs as well as GPUs. OpenAI's Triton, released in 2021, is a language for writing GPU kernels in a style close to Python, which compiles to both Nvidia's and AMD's GPUs. Since PyTorch's second version in 2023, PyTorch has generated Triton kernels automatically, so many programs run on AMD hardware without hand-written CUDA. Amazon's Neuron software serves Trainium, Huawei's CANN serves Ascend, and Apple's Metal and its MLX framework, from 2023, serve its own chips. Qualcomm bought Modular, a start-up building a language and engine meant to run models on any chip, in 2026.

Portability is no longer theory. In the MLPerf inference round of September 2026, Cisco entered the first system mixing chips from two vendors, eight Nvidia H200s and eight AMD MI350Xs, and another entry split one cluster across continents, with AMD GPUs in South Korea and the United States. The open inference engines run on Nvidia, AMD, and other hardware, so a model served through them can, in principle, move between chips.

12e. What Independent Tests Find

The best public evidence comes from SemiAnalysis. In December 2024 it found AMD's software for training on its MI300X far behind Nvidia's out of the box. Its InferenceX tests of February 16, 2026, on serving models, found a narrower but real gap. AMD's MI355X was competitive on a single server and with basic disaggregated serving. Using 8-bit numbers with disaggregated prefill and models spread widely across GPUs by expert, it matched Nvidia's B200 on SGLang, and on one server it beat Nvidia on 8-bit performance per dollar of total cost of ownership, the lifetime cost of buying and running the hardware.

The gap opened when the techniques were combined. With 4-bit numbers, disaggregated serving, and wide expert parallelism together, which is how frontier labs actually serve their models, AMD fell well behind, because its optimizations lost their edge when stacked. SemiAnalysis also found that AMD's own inference

engine lacked key features and had no production customers, that the vLLM project ran no tests on the MI355X in its automated testing, and that Nvidia led on energy per token. The pattern recurs: AMD catches up on each technique, and Nvidia stays ahead on their combination.

12f. Contested: Is CUDA a Durable Moat?

Those who say it is point to depth and speed. CUDA's long record of libraries, tools, and documentation (12a), and the millions of developers trained on them, mean that a new technique appears on Nvidia's chips first, because that is where researchers work. The edge that matters most, combining many optimizations at the frontier, is exactly where independent tests find rivals furthest behind (12e). Nvidia's software now spans networking and whole racks (Module 10), so a rival must match an entire system, not a chip, and Nvidia keeps raising the performance of chips already sold through software updates.

Those who say it is not point to the layers above CUDA. Most developers write in PyTorch, JAX, or Triton, which already target several kinds of chip, and the compilers are getting better at it. The largest buyers have the engineers to port their own software: Google runs on TPUs, Anthropic says it serves its models on Google's TPUs, Amazon's Trainium, and Nvidia's GPUs, and Meta, which created PyTorch, remains its largest contributor. AMD's software gains of 2025-26 were large, and inference, the growing part of the market, runs a few kinds of models at enormous scale, which makes porting each one worthwhile.

The evidence supports a narrower claim than either. CUDA protects Nvidia firmly among the many buyers without large engineering teams and at the leading edge of new techniques. It protects Nvidia only weakly among the handful of companies that buy most AI chips, who have shown that they can move. The course's reading is that CUDA is less a moat around Nvidia's chips than a head start renewed every year, and that it lasts only as long as Nvidia's hardware stays a generation ahead.

Worked example — Splitting a Million-Element Job

Take the simplest parallel job: add two lists of 1,048,576 numbers, element by element, in 32-bit floating point. In CUDA the programmer writes a kernel that adds one pair, and launches one thread per pair. With 256 threads in each block, that is 4,096 blocks. An RTX 5090 has 170 streaming multiprocessors (Module 5), each able to hold about 1,500 threads at once, so roughly 255,000 threads run together and the 4,096 blocks pass through in about four waves. Now the timing. The job reads two lists of 4 MB each and writes one, 12 MB in all, which at the card's 1,792 GB a second (Module 3) takes about 7 microseconds. The card's rating of about 105 trillion 32-bit operations a second counts each multiply-add as two, so it does about 52 trillion plain additions a second. The million additions take about two hundredths of a microsecond, some 350 times less. Merely launching the kernel costs a few microseconds more. So this job is limited by memory and by overhead, never by arithmetic, and a GPU running it is mostly idle. Real programs therefore fuse such steps into larger kernels, adding the lists while the numbers are already on the chip for some other purpose. That is the daily work of the libraries in 12c, and the reason the software matters as much as the silicon.

Module 12 Key Points

Software: CUDA and the Moat

- CUDA lets programmers write kernels that thousands of threads run at once, in blocks of up to 1,024.
- Fast GPU code is mostly about moving data: tiling and fused kernels keep numbers on the chip.
- Nvidia says more than 7.5 million developers use CUDA and its other tools.
- AMD's ROCm 7 raised the MI355X's throughput on one large model by 28-38% in MLPerf, on the same hardware, by AMD's own submissions.
- SemiAnalysis found AMD competitive on single servers but behind when frontier serving techniques are combined.
- CUDA is a head start renewed each year: firm for most buyers, crossable by the largest.

Module 12 Quiz

Check Your Understanding

1. What is a kernel in CUDA?

A) A chip's operating system.
B) A function that thousands of GPU threads run at once.
C) The cooling plate on a GPU.
D) The part of a CPU that schedules tasks for the GPU, running inside the operating system's core.

2. How many developers does Nvidia say use CUDA and its other tools?

A) About 100,000
B) About 75 million, one in every hundred people online
C) About 750,000, mostly at universities
D) More than 7.5 million

3. What did SemiAnalysis find in February 2026 about AMD's MI355X?

A) It beat Nvidia on every test.
B) It could not run SGLang.
C) It was competitive on single servers but fell behind when frontier serving techniques were combined.
D) It used less energy per token than Nvidia's B200.

4. What is Triton?

A) AMD's GPU driver
B) OpenAI's language for writing GPU kernels
C) A Google TPU
D) A benchmark that ranks supercomputers by how fast they solve dense systems of linear equations

5. In the worked example, what limits the speed of adding two lists of a million numbers?

A) Memory traffic and launch overhead.
B) The number of additions the GPU can do, since each thread adds slowly.
C) The clock speed of the CPU that launches the kernel.
D) The tensor cores, which cannot add numbers without multiplying them first.

Module 13 — Measuring and Pricing Compute

Benchmarks, Claims, Rents, and the Life of a Chip

Recommended Reading / Reference

Reference: MLCommons, the results of each MLPerf Training and MLPerf Inference round, at mlcommons.org, with every submission's system and software

Reference: The GPU rental price indexes of Silicon Data and Ornn, which track what it costs to rent each kind of GPU by the hour

13a. Peak and Delivered

A chip's rated petaflops are a ceiling, reached only if every arithmetic unit is busy every tick. Real work falls well short. The share of peak that a training run actually turns into useful work, called model FLOPs utilization, typically runs from about 30% to 50% on large systems. Meta reported 38% to 43% for training Llama 3.1 on the cluster described in Module 10. Generating text one token at a time uses a far smaller share of the arithmetic, because it waits on memory (Module 3).

The gap has many causes, all covered earlier: waiting for data from memory, waiting on the network (Module 10), recovering from failures, and software that leaves the units idle (Module 12). It means that the same chip can be worth very different amounts to different buyers. A team that raises its utilization from 35% to 50% gets the effect of 43% more chips at no cost, which is one reason the largest labs employ so many engineers on software and systems.

13b. The Benchmarks

For CPUs, the standard tests come from the Standard Performance Evaluation Corporation (SPEC), whose suite of real programs, SPEC CPU 2017, is still its current version, and from Geekbench 6, a free test of phones and PCs from Primate Labs, released in 2023. For AI, the reference is MLPerf, run by MLCommons. Companies submit results on set tasks under set rules, and every result is checked by the other submitters before publication, so the figures are measured, not claimed, though each company chooses which tests to enter.

The training round published on June 16, 2026 drew 24 organizations and 95 systems, and added two tests: DeepSeek-V3, the mixture-of-experts model of Module 3, and GPT-OSS 20B. Nvidia's systems trained Llama 3.1 405B in 7.07 minutes on 8,192 GB200 GPUs and DeepSeek-V3 in 2.02 minutes on 8,192 GB300s, and Nvidia was the only company to enter every test. The test runs a fixed slice of training, not a whole model from scratch, so the minutes measure a system's speed rather than the cost of building a model.

The inference round of September 16, 2026 drew a record 30 organizations and 486 results. The best result per accelerator on the DeepSeek-R1 reasoning model was 5.7 times better than a year earlier, gains that came from both hardware and software. It included the first peer-reviewed results for Nvidia's Vera Rubin rack, in a preview category, where Nvidia claimed up to 2.5 times the GB300 rack's throughput on DeepSeek-R1. Crusoe, a cloud company, entered the largest inference system yet tested, 512 AMD MI355X GPUs across 64 servers linked by Ethernet.

Supercomputers have their own ranking. The TOP500 measures each system on HPL, a benchmark that solves a huge system of linear equations in 64-bit arithmetic, the work of physics simulations. The June 2026 list, the 67th, counted five systems above an exaflop. A companion list, the Green500, ranks systems by operations per watt; it was led by KAIROS at the University of Toulouse, with 73.3 billion 64-bit operations a second per watt. The next lists come out at the SC26 conference (Module 15).

Rank	System	Site	Processors	Speed on HPL	Power
1	LineShine	National Supercomputing Center, Shenzhen, China	LingKun LX2 CPUs only (Module 7)	2.198 exaflops	42.2 MW
2	El Capitan	Lawrence Livermore National Laboratory, United States	AMD EPYC and Instinct MI300A	1.809 exaflops	29.7 MW
3	Frontier	Oak Ridge National Laboratory, United States	AMD EPYC and Instinct MI250X	1.353 exaflops	24.6 MW
4	Aurora	Argonne National Laboratory, United States	Intel Xeon CPU Max and Data Center GPU Max	1.012 exaflops	38.7 MW
5	JUPITER Booster	EuroHPC and Jülich, Germany	Nvidia GH200 Grace Hopper	1.000 exaflops	15.8 MW

TOP500, June 2026.

The list shows what it measures and what it misses. JUPITER, Europe's first exascale system, delivers its exaflop on less than half of Aurora's power. The giant AI clusters of Module 10 enter neither MLPerf's largest tests nor the TOP500, because their owners have no reason to spend days running a benchmark, and their AI work uses 4-bit and 8-bit arithmetic that HPL does not count. So the public rankings now describe science's machines well and industry's largest machines not at all.

13c. Reading a Claim

Chip makers' claims are true on their own terms and often misleading on any other. Five questions expose most of them. At what precision, and with sparsity or without (Module 8)? Per chip, per package, or per rack? Measured, or peak? Compared with what, in what configuration? On which workload, one model or many? A sixth question asks who measured it. Each of the course's examples fails at least one: the RTX 5070 that matched an RTX 4090 only with generated frames (Module 5), and Vera Rubin's claimed cut in token cost, which rests on one model (Module 8).

Google claimed that an Ironwood pod has more than 118 times the 8-bit exaflops of its nearest competitor, a comparison that, on the most natural reading, sets a whole pod (Module 9) against a rival's largest single system of linked chips. Microsoft's claim against Trainium3 compares one chip with one chip (Module 9), a fairer basis. Apple's claims compare each chip with its own predecessor. The safest numbers are independent ones, from MLPerf, the TOP500, SemiAnalysis, or published reviews, and the safest comparison is work done per dollar and per watt.

13d. What Chips and Racks Cost

Nvidia and AMD publish no prices, and buyers sign agreements that keep theirs secret, so prices come from resellers, analysts, and leaked purchase orders. Module 8 gives reported prices for the H100, the B200, and the GB300 rack. Others fill in the picture. A server board with eight H100s cost about $250,000 to $320,000 in 2026, and used H100s sold for about $15,000 to $28,000. Estimates for a GB200 NVL72 rack ran from about $2.8 million to $3.4 million. AMD's MI300X was estimated at about $10,000 to $15,000 in early 2026 by Silicon Analysts, an aggregator of estimates.

The pattern behind the prices is that buyers pay for systems, not chips. A GPU's share of a rack's cost is perhaps two-thirds, with the rest going to CPUs, memory, network cards, switches, cables, cooling, and assembly. The largest buyers negotiate discounts and pay partly in long commitments, and some, like AMD's customers, receive shares in the seller (Module 8). Published prices are therefore a rough guide to what anyone actually paid.

13e. Renting by the Hour

Most buyers rent. The large clouds publish list prices for GPU servers, and a market of smaller clouds, often called neoclouds, rents GPUs at prices tracked by several indexes. In June 2025 AWS cut its on-demand prices for H100 servers by 44%, H200 servers by 25%, and A100 servers by 33%. In January 2026 it went the other way, raising its prices for reserved blocks of GPU time by about 15%, so that a block of eight H200s rose from $34.61 to $39.80 an hour.

Market rents rose through 2026. Silicon Data's index for an H100 rose 44%, from $2.00 an hour on December 9, 2025, to the reading in the table, which gives the latest figures from three trackers for each GPU. Because the trackers measure different things, each range is wide.

GPU	Silicon Data, October 24, 2026	Ornn, October 25, 2026	Median on-demand listing, October 25, 2026
A100 (2020)	Not given	$1.09	Not given
H100	$2.88	$3.03	$3.99
H200	$3.48	$6.30	$5.95
B200	$6.45	$9.70	$7.50
B300	$7.53	Not given	$8.02

Dollars per GPU-hour. Silicon Data indexes smaller clouds' prices, Ornn averages transacted prices, and the medians of listings come from priceofcompute.com and GetDeploying's survey of 88 providers, as compiled by Singularity Research.

The reason for the rise is scarcity. Amazon's chief executive, Andy Jassy, said that AWS's capacity for 2027 was largely sold out, and SemiAnalysis reported that the newest GPUs were sold out or booked six months ahead, with one-year contracts for H100s back near their levels of early 2023, about $3 an hour. A chip that the industry treats as two generations old was earning more in October 2026 than at the end of 2025.

13f. The Price of a Million Tokens

What users of AI pay for is tokens, and the cost of producing them has fallen steeply. SemiAnalysis measured in February 2026 that serving DeepSeek-R1 on Nvidia's B200, with 4-bit numbers and Nvidia's software, cost about $0.56 per million tokens written, at 50 tokens a second for each user. A cloud company, Crusoe, charged $1.35 per million tokens read and $5.40 per million written for the same model. SemiAnalysis put Crusoe's own costs at no more than about $0.23 per million tokens read and $2.96 per million written, far above the B200 benchmark for writing, which gives gross margins of up to about 83% on input and 45% on output.

The cost per token depends on the chip, the precision, the software, and the batch size as much as on the model. Better software on the same hardware cut the cost several times over in 2025 and 2026 (Module 12), and Nvidia promises a further steep cut with Rubin on some models (Module 8). Epoch AI found in 2025 that the price of reaching a given level of capability had been falling between 9 and 900 times a year, depending on the task. Falling prices have so far raised total demand rather than cut total spending.

13g. Contested: How Long Does an AI Chip Last?

The question matters because it sets the depreciation, the annual charge that spreads a machine's cost over its useful life. A longer life means a smaller charge and higher reported profit. The table gives the lives the largest buyers use for their servers. In the critics' view, those lives are too long for chips that Nvidia replaces every year.

Company	Useful life for servers	Change and effect
Amazon	5 years for a subset of servers and network gear, from 6	From January 2025; cut its 2025 net income by about $1.0 billion
Meta	5.5 years for most servers and network assets	From January 2025; raised its 2025 net income by about $2.59 billion
Microsoft	6 years	Extended from 4 in its fiscal 2023
Alphabet	6 years	Extended from 4 in 2023
CoreWeave	6 years	Extended from 5 in 2023
Nebius	4 years	—

Companies' filings, as compiled by Altss (October 18, 2026) and beancount (July 10, 2026). None of the largest buyers changed its lives in 2026.

The critics' case was put most sharply by the investor Michael Burry, in posts in November 2025. He argued that the clouds inflate their earnings by stretching useful lives while Nvidia's product cycle runs two to three years, and estimated about $176 billion of understated depreciation across 2026-2028. He projected that 2028 earnings would be overstated by about 27% at Oracle and 21% at Meta, and disclosed bets against Nvidia's shares; he alleged nothing illegal. Goldman Sachs's research arm noted that books can show orderly depreciation while hardware becomes obsolete much faster, and the International Monetary Fund (IMF) warned of the same risk in April 2026.

The defenders' case rests on use. A chip trains frontier models for a year or two, then moves to inference and lighter work for years more, the value cascade that worked for earlier servers. The market supports them for now: A100s from 2020 still earned rent in October 2026, and H100 rents rose through 2026 (13e). As long as an old chip earns more than its running costs, it stays in service.

The evidence of 2026 favors the defenders, in a market where demand exceeds supply. But a chip's economic life depends less on its age than on the power it uses. A data center with a fixed grid connection (Module 11) earns more by replacing an old chip with one that produces several times as many tokens from the same watts, so chips may be retired well before they wear out once supply catches up. The course's reading is that lives of five to six years are defensible while chips are scarce, as they were in 2026, and too long once supply catches up.

Worked example — The Cost of a GPU-Hour

Build up the cost of running one GPU in a GB300 rack. Its share of the rack costs about $69,000 (Module 8). Spread over five years, 43,800 hours, that is about $1.58 an hour. Financing at 8% a year on half the purchase price, the average balance outstanding, adds about $0.32. Electricity and cooling at the Rubin rate worked out in Module 11, a little high for a GB300 GPU, add about $0.29, and space, staff, and maintenance, here an illustrative $0.50, bring the total to about $2.68 an hour. But no GPU is rented every hour; at 70% of hours sold, each sold hour must cover $2.68 divided by 0.7, about $3.83. Against the B300's market rents of about $7.50 to $8 in October 2026 (13e), that leaves a healthy margin. Now shorten the life to three years, as the critics urge: the hardware cost rises to about $2.63 an hour, the total to about $3.73, and the cost of each sold hour to about $5.33, still under the market rate, but with far less room. And if rents fell by about 30%, enough to undo the 44% rise in H100 rents since December 2025, the three-year case would barely break even. That sensitivity, to the chip's life and to the rent, is what the argument in 13g is about.

Module 13 Key Points

Measuring and Pricing Compute

• Large training runs sustain only about 30-50% of their chips' peak; Meta's Llama 3.1 run reached 38-43%.

• In MLPerf's June 2026 round, Nvidia trained Llama 3.1 405B in 7.07 minutes on 8,192 GPUs.

• Before believing a claim, check its precision, sparsity, unit, baseline, workload, and who measured it.

• Renting an H100 cost about $2.88-3.03 an hour in late October 2026, up about 44% since December 2025.

• The clouds book servers over 5-6 years; critics say AI chips last 2-3; A100s from 2020 still rent for $1.09.

Module 13 Quiz

Check Your Understanding

1. What share of their chips' peak do large training runs typically sustain?

A) About 5%
B) About 40%
C) About 95%, near the peak
D) More than 100%, with boost clocks

2. What does the HPL benchmark behind the TOP500 measure?

A) Solving dense linear equations
B) Tokens generated per second on a set of language models
C) How many images a system can classify in an hour
D) Electricity used per year, since the list ranks supercomputers by how little power they draw for their size

3. Why did H100 rental prices rise in 2026?

A) Nvidia stopped making GPUs.
B) Cloud companies cut capacity.
C) New chips ran old models worse.
D) Demand ran ahead of supply: AWS said its 2027 capacity was largely sold out, and the newest GPUs were booked months ahead.

4. Over how many years does Meta depreciate most of its servers and network gear?

A) Two
B) Ten years
C) Five and a half years
D) Three years, matching Nvidia's product cycle, as critics urge

5. In the worked example, what is the cost of each sold GPU-hour at 70% utilization over five years?

A) About $3.83
B) About $0.29, the electricity alone
C) About $1.58, the hardware alone
D) About $7.53, the market rate for renting a B300

Module 14 — The Processor Business: Companies, Markets, and Money

Who Earns What, and Whether It Lasts

Recommended Reading / Reference

Reference: Each company's quarterly results release and financial commentary, filed with the SEC on Form 8-K and posted on its investor-relations site

Reference: Epoch AI, its estimates of AI chip production, sales, and computing power by company, at epoch.ai

14a. The Market by Segment

Processors are several markets, each with its own leader, and the table sets them side by side with the module that gives the detail. One market now dwarfs the others. Nvidia's data-center revenue in its quarter to July 2026 (14b) was more than Intel or AMD takes in during a whole year at its current pace, and its growth has reordered the industry around AI.

Market	Leader	Its position	Main challengers	Detail
AI accelerators	Nvidia	Most of the revenue, by the estimate below	Custom chips (Google, Amazon, Microsoft, Meta, OpenAI, with Broadcom and Marvell); AMD; Huawei in China	Modules 8 and 9
Custom AI chip design	Broadcom	The largest share, by Counterpoint's estimate	Marvell, MediaTek, Alchip, GUC	Module 9
Server CPUs	Intel by units; AMD close on revenue	AMD gaining in units and revenue	Arm-based chips from the clouds and Nvidia	Modules 4 and 7
PC CPUs	Intel	Most x86 PC processors	AMD; Apple and Qualcomm on Arm	Modules 4 and 6
Phone processors	MediaTek by units; Apple and Qualcomm at the top end	MediaTek ships the most	Samsung, Google, Huawei	Module 6
Graphics cards	Nvidia	The great majority of desktop cards	AMD, Intel	Module 5

The modules named give the shares, which are estimates from different firms and periods.

The estimate of Nvidia's share comes from Silicon Analysts, an aggregator that blends estimates from Bloomberg Intelligence, IDC, and its own work. It put Nvidia at about 87% of AI accelerator revenue in 2024, 80-81% in 2025, and about 75% in 2026, with AMD at about 7-10% and custom chips taking most of the rest. Epoch AI counts units instead (Module 8). Neither figure is precise, but both show Nvidia's lead shrinking slowly from a great height.

14b. The Latest Quarters

The table gives each company's most recent results as of October 30, 2026, and its guidance for the next quarter; Module 15 gives the dates of the next reports. Fiscal quarters differ, as the note on the sources at the front explains.

Company	Latest quarter	Revenue	Change on a year earlier	Notable lines	Gross margin	Guidance for next quarter
Nvidia	Second quarter of fiscal 2027, to July 26, 2026	$96.2 billion	+106%	Data Center $89.0 billion (+117%); Edge Computing $7.2 billion (+27%)	75.0%	$108 billion
Broadcom	Third quarter of fiscal 2026, to August 2, 2026	$29.6 billion	+86%	Semiconductors $20.8 billion; software $8.8 billion; AI chips in Module 9	Not used here	About $34.8 billion
Intel	Third quarter of 2026, to September 26, 2026	$16.5 billion	+21%	Client $8.1 billion (-5%); Data Center and AI $7.3 billion (+78%); Foundry $6.7 billion (+58%)	43.0% (44.1% adjusted)	$15.0-16.0 billion
AMD	Second quarter of 2026, to June 27, 2026	$11.5 billion	+50%	Data Center $6.7 billion (+107%); Client $3.1 billion (+23%); Gaming $0.8 billion (-31%)	54% (56% adjusted)	About $13.0 billion
Qualcomm	Third quarter of fiscal 2026, to late June	$9.95 billion	-4%	Chips $8.5 billion; licensing $1.28 billion; phones and cars in Module 6	Not used here	$9.7-10.5 billion
Arm	First quarter of fiscal 2027, to June 30, 2026	$1.29 billion	+22%	Licensing $574 million; royalties in Module 4; adjusted operating margin 41.2%	Not used here	Not used here
Marvell	Second quarter of fiscal 2027, to August 1, 2026	$2.74 billion	+37%	Data center in Module 9	Not used here	About $3.15 billion

Companies' releases. Adjusted margins leave out items such as stock pay and acquisition costs. AMD's Data Center line includes both EPYC CPUs and Instinct GPUs. Nvidia's Data Center and Edge Computing are market platforms it began reporting in 2026, which sum to its revenue; its formal accounting segments remain Compute and Networking and Graphics.

Nvidia first became the world's most valuable listed company in June 2024. At the close of October 23, 2026 it was worth about $5.09 trillion, by StockAnalysis's figures, after a record close of $238.90 a share on October 5 and a fall of about 12% since. In September it added $150 billion to its authority to buy back its own shares.

14c. Why Margins Differ

Gross margin, the share of revenue left after the cost of making what was sold, separates these companies sharply. Nvidia's margin is extraordinary for a hardware maker. It reflects scarcity, since buyers have queued for its chips since 2023, along with the software that makes them hard to replace (Module 12) and the whole systems it sells, racks and networks included (Module 10). Arm's licensing has almost no cost of goods at all, which is why it reports its profitability as an operating margin (14b).

AMD's margin reflects a challenger's prices and a mix weighted toward PCs and consoles. Intel's, the lowest in the table, reflects its factories. Unlike the others, which design chips and pay TSMC to make them, Intel owns its plants, so it carries their enormous fixed costs whether or not they are full, and its foundry business sold only $293 million to outside customers in the second quarter. A factory owner's margin swings with how full its factories are: Intel's rose about 12 points in the year to June as demand recovered, and it expected to lose about 6 points in the fourth quarter as it cut its output of PC chips (14f).

14d. Customers and the Web of Investments

A few buyers account for most of the money. Amazon, Microsoft, Alphabet, and Meta guided in July 2026 to about $720-745 billion of capital spending for the year, and about $800 billion with Oracle; much of it goes on AI chips and the data centers to house them. Nvidia's finance chief, Colette Kress, said in August that those five would invest $1.3 trillion in 2027, and that the cloud industry's backlog of contracted work exceeded $2 trillion.

The four companies' results of late October left their plans for 2026 little changed. Amazon kept about $220 billion and Alphabet $195-205 billion, Meta narrowed its range to $135-145 billion, and Microsoft expected about $170 billion for the calendar year, a little less than before, because capacity it had contracted from other providers arrived late. By this course's sum the four will spend about $720-740 billion in 2026. None gave a figure for 2027, but Amazon expected to spend meaningfully more and Alphabet significantly more, and Microsoft expected its spending to speed up in the first half of 2027; Module 15 lists their next reports.

The suppliers increasingly finance their customers. Nvidia invested about $30 billion in OpenAI's funding round of early 2026, in place of a plan for up to $100 billion announced in 2025 that was never completed. In November 2025 it agreed to invest up to $10 billion in Anthropic. It owns about 4% of Intel (Module 6). In the quarter to July it signed memoranda with six large investment firms, among them BlackRock and KKR, to raise more than $500 billion of outside capital for AI infrastructure.

AMD has paid its largest customers in its own shares, through warrants to OpenAI and Meta, and agreed to invest in Anthropic (Module 8). Critics call the pattern circular: money flows from chip makers to AI companies and back as chip purchases, inflating the revenue of both. Defenders reply that the investments are small beside the purchases and that customers such as Microsoft, Alphabet, and Meta pay mainly from their own profits. The pattern is a reason to weigh each supplier's revenue by who is paying with what.

14e. China's Market

China was once a large market for Nvidia. American export controls and Beijing's push for domestic chips (Module 8) have made it a small and uncertain one, and Nvidia's guidance now assumes none of its data-center computing sales there. The market itself is growing fast: IDC expects China's spending on accelerated servers to rise from $37.7 billion in 2025 to $150.1 billion in 2030, as reported by Caixin, and domestic suppliers are taking it.

The winners are the Chinese designers that investors have rushed to fund. Cambricon's and Hygon's revenues surged in the first half of 2026 (Modules 7 and 8). Moore Threads, a maker of GPUs that listed on Shanghai's STAR Market for technology companies in December 2025, more than doubled its revenue to 1.7 billion yuan in the first half and plans to list in Hong Kong as well. Their limits are the factories and memory available to them, since their output depends on what SMIC can make (Module 8).

14f. Intel's Turnaround

Intel spent 2024 in crisis: losing share in servers and PCs, years late on its processes, and short of cash for its factories. Lip-Bu Tan became chief executive in March 2025. That August the American government took about 9.9% of the company, 433.3 million new shares at $20.47, about $8.9 billion, paid for with grants already promised under the CHIPS Act, the 2022 law that subsidizes chipmaking in America. The stake is passive, with no seat on the board. SoftBank invested $2 billion, and Nvidia bought shares too (Module 6).

The second quarter of 2026 showed both sides. Revenue grew 25%, its fastest pace in years, and its 18A process reached volume in Panther Lake and Clearwater Forest, with management saying output ran about 25% above target. Yet it reported a net loss of $11.0 billion, because of a $12.5 billion accounting charge on government shares held in escrow, which are released as Intel meets the terms of a defense program, as the value of those shares rose with Intel's stock. That charge reflected Intel's recovery rather than its business.

The third quarter, reported on October 29, showed how far the recovery still had to go. Revenue grew again (14b), and chips made on 18A rose from under 5% of Intel's PC processor revenue in the first quarter to nearly a third. But Intel expected lower revenue in the fourth quarter, for two reasons. PC makers were cutting their stocks of its chips, in a slowdown it expected to last at least through the first half of 2027. And an explosion that killed two workers at its factory in Kiryat Gat, Israel, on October 14 shut the plant for about two weeks, which Intel expected to cost about $300 million of revenue in the fourth quarter and probably as much in the first. Its shares fell by more than a fifth after the report.

The test ahead is the factories. Intel's next process, 14A, is due for risk production in the second half of 2027, and Tan expects outside customers to decide whether to commit to it between late 2026 and mid-2027. In October Intel gave selected customers version 0.9 of the design kit for 14A, the rules and models they need to lay out a chip on it, but it named no customers. Intel has said it could slow or cancel 14A without enough outside demand. A foundry with only its own chips to make cannot match TSMC's scale, so the decisions of companies such as Apple, Nvidia, and Microsoft will decide whether Intel remains a chipmaker or becomes mainly a designer.

14g. Contested: Is the AI Chip Boom a Bubble?

Those who see a bubble point to the scale of the spending (14d), against revenue from AI services that remains a fraction of it. They point to the circular financing of 14d, to the depreciation question (Module 13), and to concentration: a handful of buyers whose plans can change at once. The fall in Nvidia's shares since their recent record (14b) is a reminder of how much optimism their price contains.

Those who see a sound market point to the physical evidence. Every chip that can be made is sold, rents for old GPUs rose through 2026, and Amazon says its capacity for 2027 is largely sold out (Module 13). Nvidia's chief executive, Jensen Huang, said in March 2026 that Nvidia was targeting $1 trillion of orders for Blackwell and Rubin through 2027, double the figure of a year earlier. The buyers are among the most profitable companies in history, and the cost of AI per token keeps falling while use rises faster (Module 13).

The evidence answers two different questions differently. The market for chips in 2026 shows shortage, not glut: no sign of idle GPUs, falling rents, or cancelled orders, the signals that marked the fiber glut that followed the boom of 2000. Whether the spending earns a return is a separate question about the future, and it will be answered by the revenue of AI services, not by chip sales. The course's reading is that the chips are not a bubble in 2026, but the expectations built into the prices of the companies may be, and Module 15 names the signals that would tell.

Worked example — Where a Gigawatt's Dollars Go

Nvidia now sizes its business in gigawatts of data-center capacity. On its call of August 26, 2026, Kress and Huang put Nvidia's revenue for each gigawatt at about $18 billion for Hopper, $25 billion for Grace Blackwell, and $40 billion for Vera Rubin, and Huang separately called a gigawatt a "$60 billion investment." If both figures describe the same thing, Nvidia's share of a gigawatt rises to about two-thirds with Vera Rubin, leaving about $20 billion for the land, buildings, power equipment, cooling, and everything else; Nvidia did not reconcile them. The more revealing figure is the rise: from $18 billion to $40 billion is a gain of about 120% in Nvidia's revenue for each watt it supplies. Because buyers are limited by power more than by money (Module 11), a supplier's revenue grows with what it can sell per watt, and Nvidia's strategy is to raise that number each year. The same logic explains why buyers accept it. A Vera Rubin gigawatt costs more than a Blackwell gigawatt, but if it produces several times the tokens, as Nvidia claims (Module 8), each token costs less. The price of AI per watt is rising for buyers even as its price per token falls, and the gap between the two is where Nvidia's margin lives.

Module 14 Key Points

The Processor Business: Companies, Markets, and Money

• Nvidia's revenue was $96.2 billion in its quarter to July 2026, 92.5% of it from data centers, at a 75% gross margin.

• Nvidia held an estimated three-quarters of AI accelerator revenue in 2026, down from about 87% in 2024.

• Intel lost $11.0 billion in its second quarter of 2026, mostly from a $12.5 billion charge on government shares.

• Nvidia's revenue per gigawatt rose from about $18 billion for Hopper to about $40 billion for Vera Rubin.

• The chip market of 2026 shows shortage, not glut; the bubble question is about expectations, not chips.

Module 14 Quiz

Check Your Understanding

1. What share of Nvidia's revenue came from data centers in its quarter to July 2026?

A) About half
B) About 92.5%, $89.0 billion of $96.2 billion
C) About 60%
D) About a third, mostly from gaming

2. Why did Intel report a net loss in its second quarter of 2026 despite strong sales?

A) It sold its factories.
B) It paid a fine to the European Union.
C) It wrote off its 18A process as a failure, closing its newest factory in Arizona and cancelling Panther Lake.
D) A $12.5 billion charge on the government's shares held in escrow.

3. About what share of AI accelerator revenue did Nvidia hold in 2026, on one estimate?

A) About a quarter
B) About half, as custom chips caught up
C) About three-quarters
D) All of it except Google's TPUs, which it does not count as accelerators

4. What did Nvidia say it earns for each gigawatt of Vera Rubin?

A) About $40 billion
B) About $4 billion
C) About $18 billion, as for Hopper
D) About $400 billion, the whole capital budget of a cloud company

5. What is the course's reading on whether the AI chip boom is a bubble?

A) It is a bubble about to burst.
B) Chips are in glut.
C) It cannot be a bubble, because every chip is sold and every rent is rising, so the prices of AI companies must be justified.
D) The market shows shortage; the risk lies in expectations.

Module 15 — Where Things Stand, and What to Watch

The Argument, the Contradictions, and the Calendar

Recommended Reading / Reference

Reference: The investor-relations pages of Nvidia, AMD, Intel, Broadcom, Arm, and Qualcomm, where each posts its dates, results, and the transcripts of its calls

Reference: The TOP500 and MLPerf results pages, top500.org and mlcommons.org, where the next lists and rounds appear

15a. The Argument on One Page

For decades processors got faster almost for free, because each shrink of the transistor made chips denser, quicker, and thriftier at once (Module 1). That ended when power density stopped falling, and since then speed has come from doing many things at once and from chips built for one kind of work (Module 2). The general-purpose CPU did not disappear: it still runs every computer and now manages the accelerators beside it (Module 7). But the growth, and the money, moved to parallel and specialized processors.

The GPU won the first round of AI because it arrived ready. Its design, thousands of simple cores hiding memory delays behind threads, fit the arithmetic of neural networks, and CUDA had put a programming model on every Nvidia card years before anyone needed it for AI (Modules 5 and 12). Nvidia then turned a chip into a system, selling racks, networks, and software together (Modules 8 and 10), and became one of the most profitable companies in history (Module 14).

The limits have moved. Arithmetic is cheap; memory is slow and, in 2026, scarce (Module 3). The largest systems are bound by their networks, their power, and their cooling (Modules 10 and 11), and by the software that keeps the arithmetic busy (Module 12). The instruction set matters less for speed than for who may build a chip, which is why the clouds build on Arm and China bets on RISC-V (Module 4). And the largest buyers now design chips of their own, pressing on Nvidia's prices (Module 9).

The course asked whether today's leaders hold moats or only leads. Its answer is a lead, renewed every year, that holds only as long as Nvidia's hardware stays a generation ahead (Modules 9 and 12).

15b. The Contradictions

AI made the consumer's computer dearer. The chips that run AI live in data centers, yet their appetite for memory raised the price of phones, PCs, consoles, and graphics cards in 2026, and cut how many were sold (Modules 3, 5, and 6). Consumers paid for an AI boom whose largest models they mostly do not run on their own devices, and whose most capable machines they will never see.

The most valuable chip company sells to the customers building its replacements. Google, Amazon, Microsoft, Meta, and OpenAI all buy Nvidia's chips by the hundreds of thousands while designing rivals to them, and Nvidia licenses its own chip link to one of those rivals, Amazon's Trainium4 (Modules 8, 9, and 10). Each side needs the other for now: the buyers cannot get enough chips, and Nvidia cannot find better customers.

Old chips gained value as new ones made them obsolete on paper. Nvidia claims Vera Rubin cuts the cost of a token steeply, yet rents for the H100, two generations older, rose through 2026, and A100s from 2020 still earn money (Modules 8 and 13). In a shortage the oldest working chip is still worth renting, and its value depends on the market for computing, not on the calendar.

The fastest public supercomputer runs on CPUs alone, while the largest AI machines appear on no list. LineShine topped the TOP500 with no accelerators listed, while clusters many times its size in AI computing never run the benchmark (Modules 7, 10, and 13).

Every generation uses less energy per token and more power per rack. Each new GPU produces more tokens per joule, yet racks have climbed from tens of kilowatts toward hundreds, and data centers' total electricity keeps rising (Modules 11 and 13). Efficiency has lowered the price of each answer and so raised the number of answers asked for, and the savings on each token have never shown up as savings on the grid.

15c. What to Watch

When	What	Why it matters
November 3	AMD's third-quarter results	The first revenue from the MI450 series and Helios, and whether AMD's GPU business is becoming a second source at scale or stays a niche
November 4	Arm's and Qualcomm's results	Royalty rates per chip and orders for Arm's own chip; Qualcomm's moves into data centers. Judge Noreika's rulings on royalties in the two companies' case were pending, with no date set by October 30
November 15-20	The SC26 supercomputing conference in Chicago, with new TOP500 and Green500 lists	Whether LineShine keeps first place, and how many new systems use AI chips for science
November 17	Nvidia's results for the quarter to October 25	Vera Rubin's share of revenue, gross margin under the new ramp, and any sales in China: the clearest single reading of AI demand
November 30-December 4	AWS re:Invent, Las Vegas	Amazon usually unveils its next Trainium and Graviton chips there
December	Broadcom's fourth quarter and Marvell's third	Whether custom-chip revenue keeps doubling, and Broadcom's outlook for fiscal 2027
January 6-9, 2027	CES, Las Vegas	PC chips for 2027, including Intel's Nova Lake and AMD's first Zen 6 laptops, and whether memory prices ease
Late January to early February 2027	Microsoft's, Alphabet's, Amazon's, and Meta's fourth-quarter results; dates not yet announced	Their first figures for 2027 capital spending (Module 14), which will set the size of that year's market for AI chips
March 2027	Nvidia's GTC conference, San Jose	Whether Rubin Ultra and the Kyber rack are on time, and details of Feynman
First half of 2027	Anthropic's first gigawatt of AMD GPUs; Intel's 14A customers decide; Huawei's Ascend 960DT, due in the first quarter	A frontier lab at scale on AMD, the future of Intel's factories, and how fast China's own chips advance
2027	Rubin Ultra; AMD's MI500; the next Xbox; Intel's Diamond Rapids; first UALink products; OpenAI's Jalapeño at gigawatt scale	The next round of the race for racks, and the first open rival to NVLink

When	What	Why it matters
Late 2027 to 2028	Memory prices expected to ease; Amazon's first sales of Trainium outside AWS; Nvidia's RTX 60 and AMD's RDNA 5 gaming GPUs, as reported; Nvidia's Feynman in 2028	Relief for consumers, if new memory factories arrive; whether the clouds' chips become products sold to others; and the next generation after Rubin

Dates as announced or reported by October 30, 2026; products and dates beyond 2026 are the companies' plans or press reports.

15d. The Last Word

The 4004 of 1971 added numbers for a calculator. Its descendants of 2026 are measured in gigawatts and priced like power stations, and the question asked of them has changed from how fast a chip runs to how much useful work it does for each watt and each dollar. The answer now depends as much on memory, links, cooling, and software as on transistors. The companies that understood that first, Nvidia above all, have been rewarded on a scale the industry had never seen. What happens next turns on whether the largest buyers can do the same for themselves, and on whether the uses of all that computing grow as fast as the machines built to supply it.

Module 15 Key Points

Where Things Stand, and What to Watch

• Since 2005, speed has come from parallel and specialized chips; memory, power, and software now set the limits.

• AI's demand for memory made phones, PCs, consoles, and graphics cards dearer in 2026.

• Nvidia's lead is a head start renewed yearly; its largest customers buy from it while designing replacements.

• Old GPUs gained value: H100 rents rose about 44% from December 2025 while newer chips promised far cheaper tokens.

• Watch AMD on November 3, Nvidia on November 17, and the cloud companies' first figures for 2027 spending early in 2027.

Module 15 Quiz

Check Your Understanding

1. Which contradiction does the course draw from the memory shortage?

A) Memory became cheaper as AI grew.

B) AI chips stopped using memory.

C) The boom in AI chips made phones, PCs, consoles, and graphics cards dearer for consumers who mostly do not run large models locally.

D) Consoles became cheaper while PCs rose.

2. Which system topped the June 2026 TOP500 list?

A) El Capitan, on AMD GPUs

B) LineShine

C) Colossus 2

D) Meta's Prometheus, the first AI cluster ever entered, which ran the benchmark on the first phase of its GPUs in October

3. When does AMD report its third-quarter 2026 results?

A) October 29, with Intel

B) Mid-to-late November

C) December, after Broadcom

D) November 3

4. What does the course name as the binding limits on computing in 2026?

A) Transistors

B) Clock speeds and cores

C) Memory, power, and software

D) The number of engineers who can design chips, since every company now builds its own

5. How does the course describe Nvidia's position?

A) A permanent monopoly

B) A lead renewed each year, secure only while its hardware stays a generation ahead

C) A company already overtaken by custom chips

D) A maker of gaming cards that dabbles in AI

Answer Sheet

Answers to every module's Check Your Understanding quiz.

Module	Answers
Module 1 Quiz	1. B 2. C 3. D 4. A 5. D
Module 2 Quiz	1. C 2. A 3. B 4. B 5. D
Module 3 Quiz	1. C 2. A 3. C 4. D 5. A
Module 4 Quiz	1. B 2. B 3. C 4. A 5. D
Module 5 Quiz	1. D 2. C 3. B 4. A 5. C
Module 6 Quiz	1. D 2. B 3. A 4. C 5. D
Module 7 Quiz	1. B 2. A 3. B 4. D 5. A
Module 8 Quiz	1. C 2. D 3. C 4. A 5. B
Module 9 Quiz	1. A 2. B 3. C 4. D 5. A
Module 10 Quiz	1. B 2. D 3. C 4. B 5. C
Module 11 Quiz	1. D 2. A 3. C 4. D 5. A
Module 12 Quiz	1. B 2. D 3. C 4. B 5. A
Module 13 Quiz	1. B 2. A 3. D 4. C 5. A
Module 14 Quiz	1. B 2. D 3. C 4. A 5. D
Module 15 Quiz	1. C 2. B 3. D 4. C 5. B