

GPUs, TPUs, and Why AI Needs Special Chips

FREE EDUCATIONAL BRIEF · LEARN AI. UNCOVER THE DATA. PROTECT THE TRUTH.

Modern AI runs on a simple mathematical diet: enormous grids of numbers multiplied together, billions of times per second. Ordinary CPUs execute a few complex tasks quickly; AI needs millions of simple multiplications simultaneously. Graphics processors — built to color millions of pixels in parallel — happened to be exactly that machine, and the discovery reshaped an industry.

THE CAST OF PROCESSORS

» CPU

The generalist: superb at sequential logic, the conductor of every system — but a bottleneck for bulk math.

» GPU

Thousands of small cores working in parallel; the workhorse of both AI training and inference.

» TPU and custom accelerators

Chips designed specifically for neural-network math, trading flexibility for efficiency at scale.

» NPU

Small accelerators inside phones and laptops that run compact models locally — powering on-device AI.

TRAINING VERSUS INFERENCE

Training a frontier model is a months-long computation across thousands of interconnected accelerators — the exclusive province of well-funded labs. Inference — actually using a trained model — is far lighter, which is why capable AI runs from ordinary cloud services and increasingly on the device in your pocket. Businesses buy inference; only labs buy training runs.

WHY CHIPS BECAME GEOPOLITICS

Cutting-edge accelerators depend on a handful of fabrication facilities and one dominant lithography supplier. Export controls on advanced chips are now instruments of national strategy, and “compute” is discussed alongside oil and grain. When you read about chip restrictions, you are reading about who gets to train the next generation of AI.

INVESTIGATOR'S TAKEAWAYS

- AI's core workload is massive parallel multiplication — GPUs fit it natively.
- Training is rare and colossal; inference is cheap and everywhere.
- On-device NPUs are bringing private, offline AI to consumer hardware.
- The accelerator supply chain concentrates in very few hands.
- Compute access is now a strategic resource — expect policy to follow it.