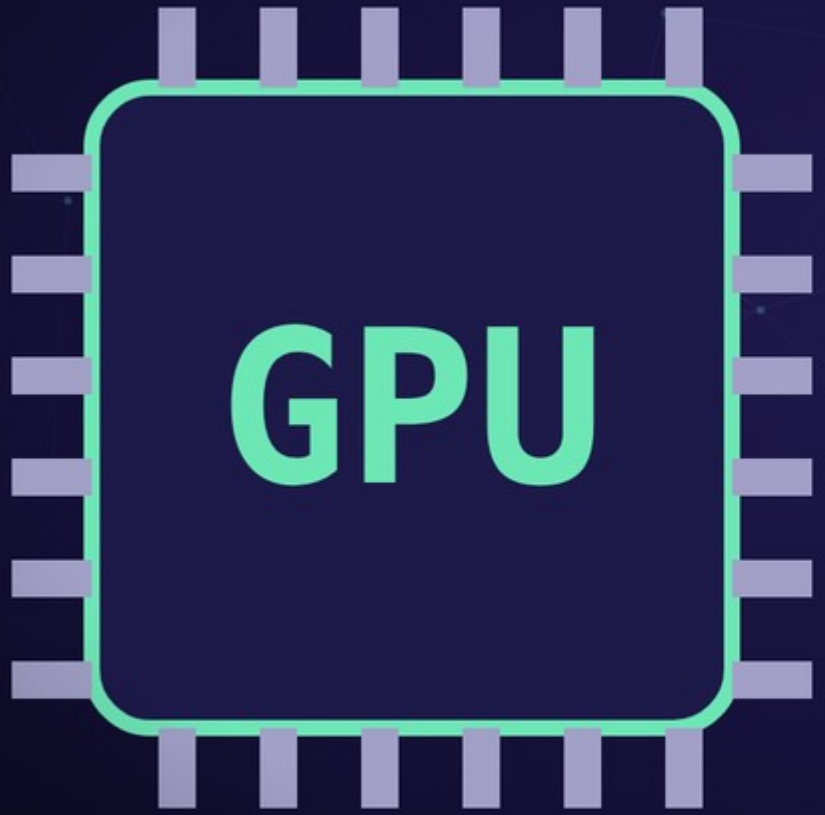
@MyWebUniversity

#4

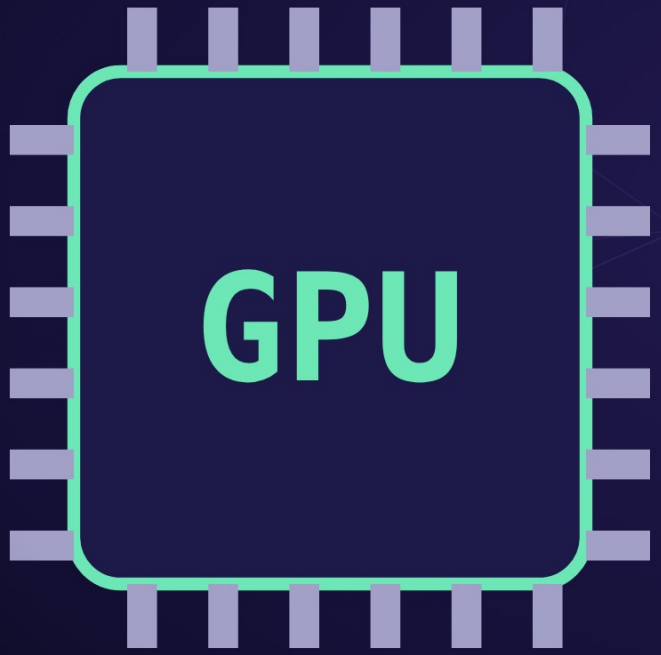


GPU

vs CPU

for AI

AI EXPLAINED · A FREE SERIES FOR STUDENTS AND DEVELOPERS



A MYWEBUNIVERSITY SERIES

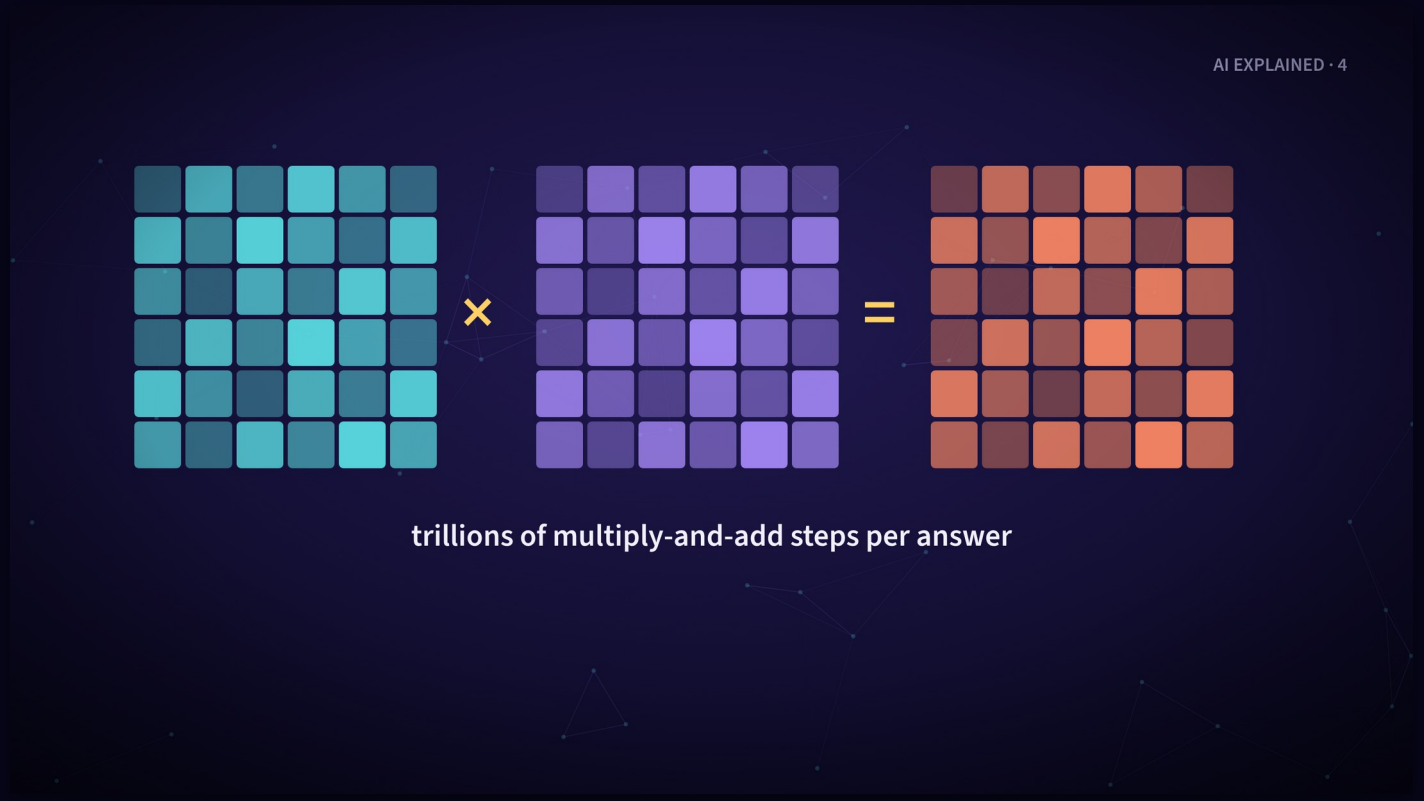
AI EXPLAINED

EPISODE 4

Hardware for AI

CPU · GPU · TPU · VRAM

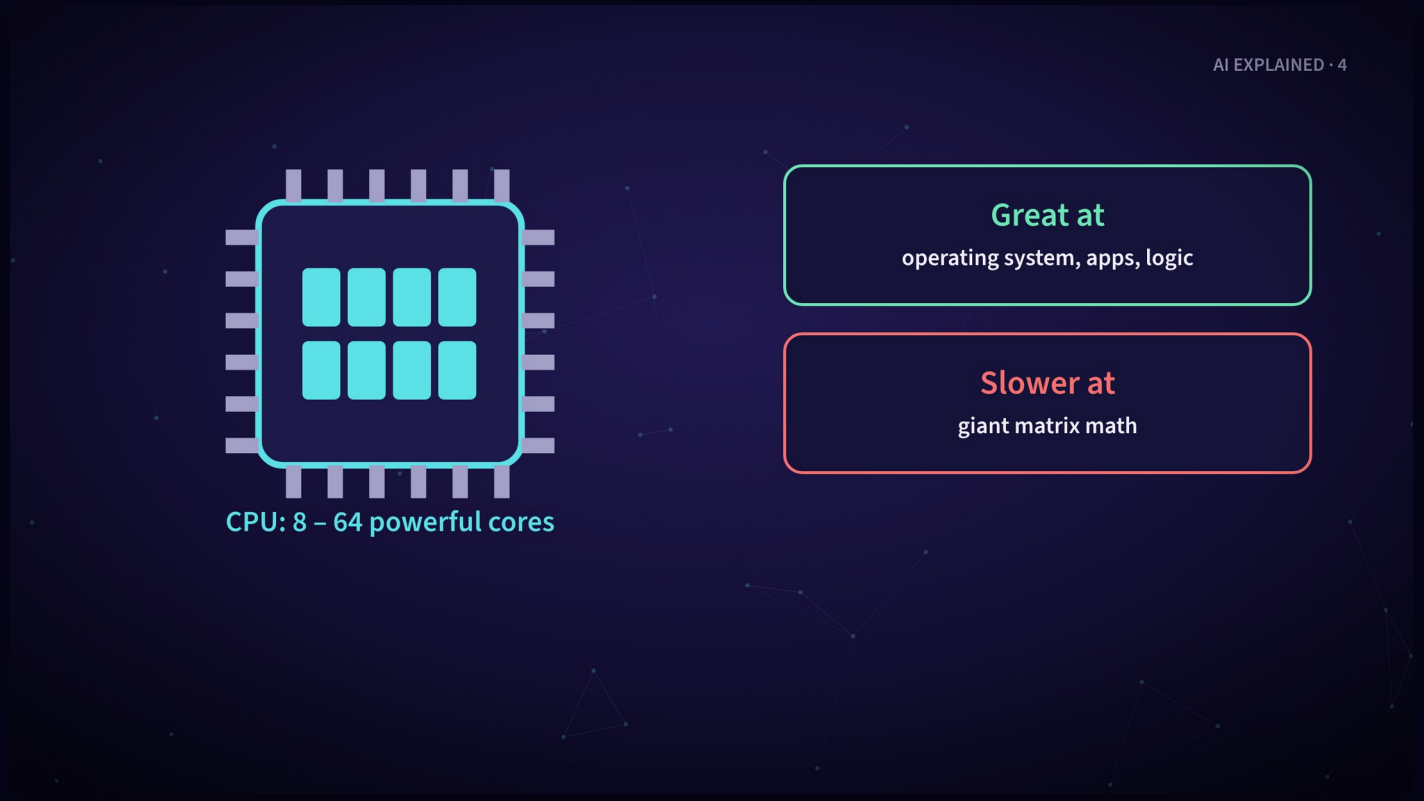
AI Is Mostly Multiplication



- Neural networks = matrix multiplication
- Trillions of operations per answer
- Parallel hardware wins

01

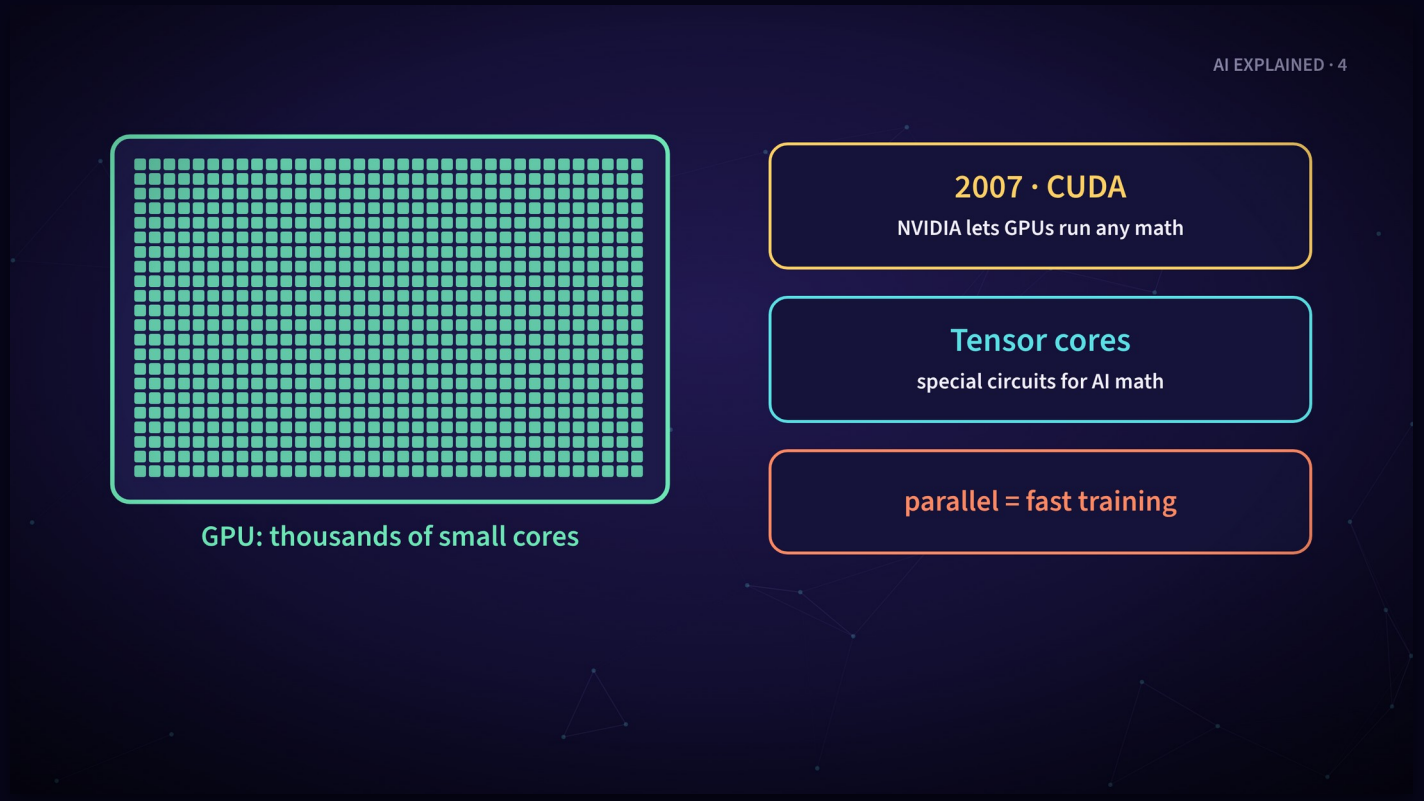
The CPU: A Few Brilliant Workers



- CPU: 8–64 powerful cores
- Great for general, sequential work
- Slow for giant matrix math

02

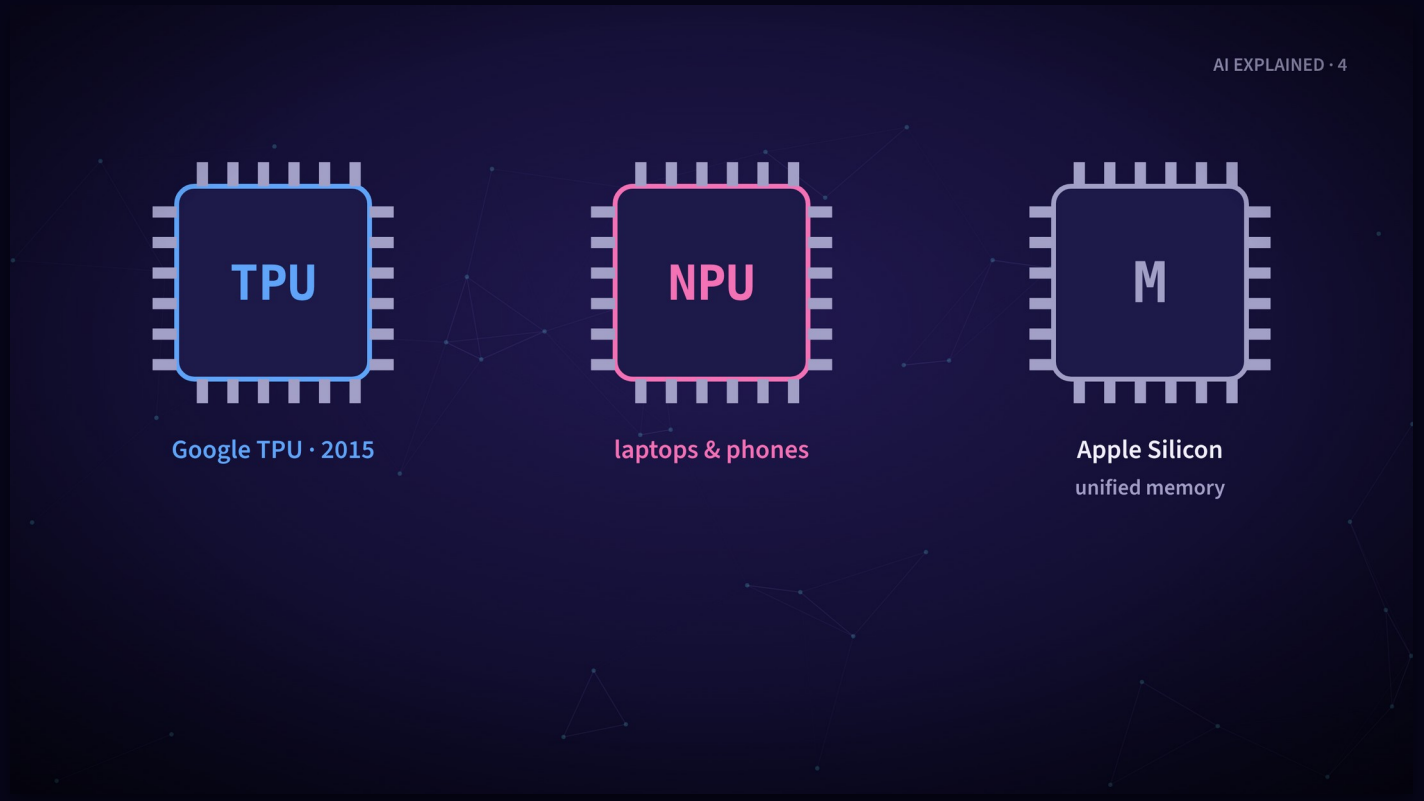
The GPU: Thousands of Workers



- GPU: thousands of parallel cores
- 2007: NVIDIA CUDA → general GPU computing
- Tensor cores for AI math

03

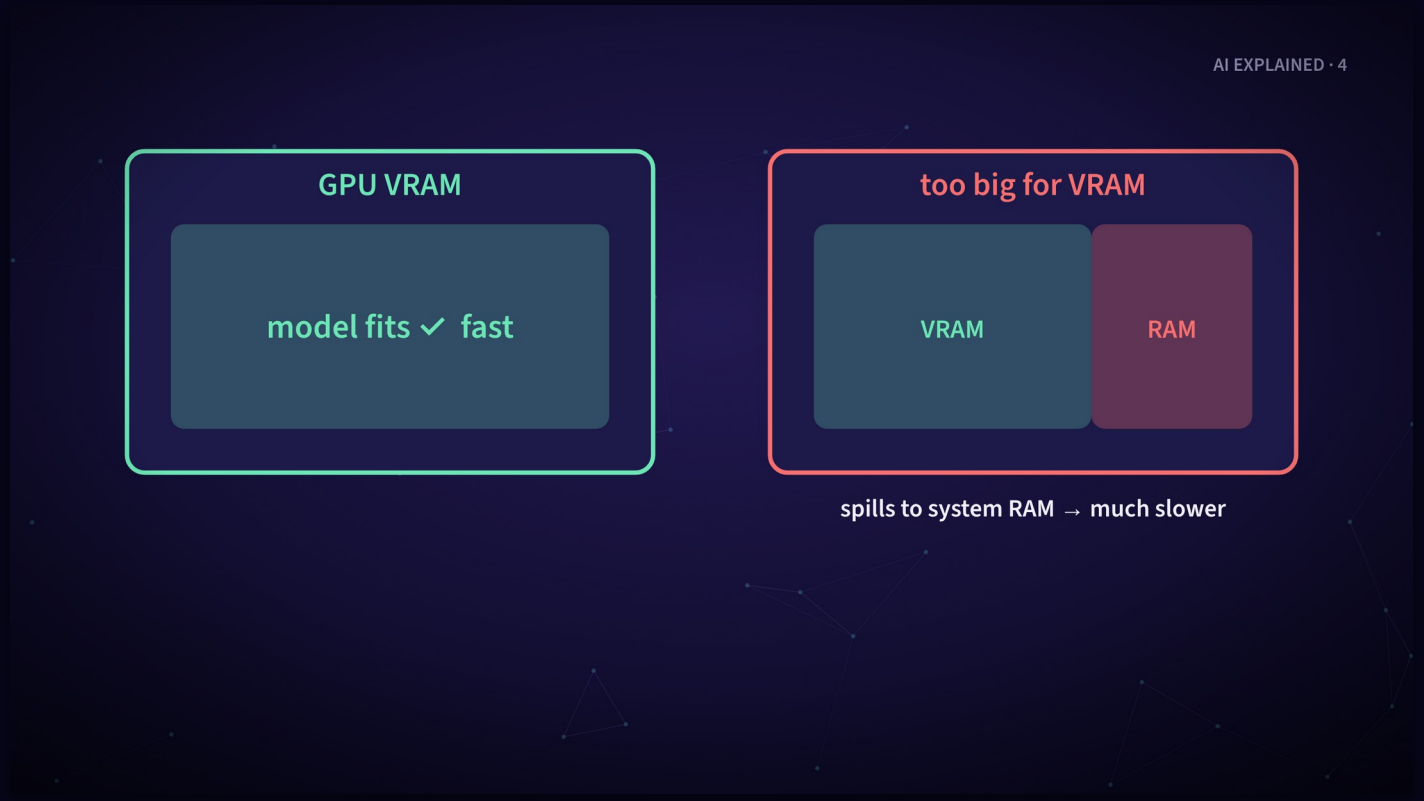
TPUs, NPUs and Apple Silicon



- Google TPU (in use since 2015)
- NPUs in laptops and phones
- Apple Silicon: unified memory

04

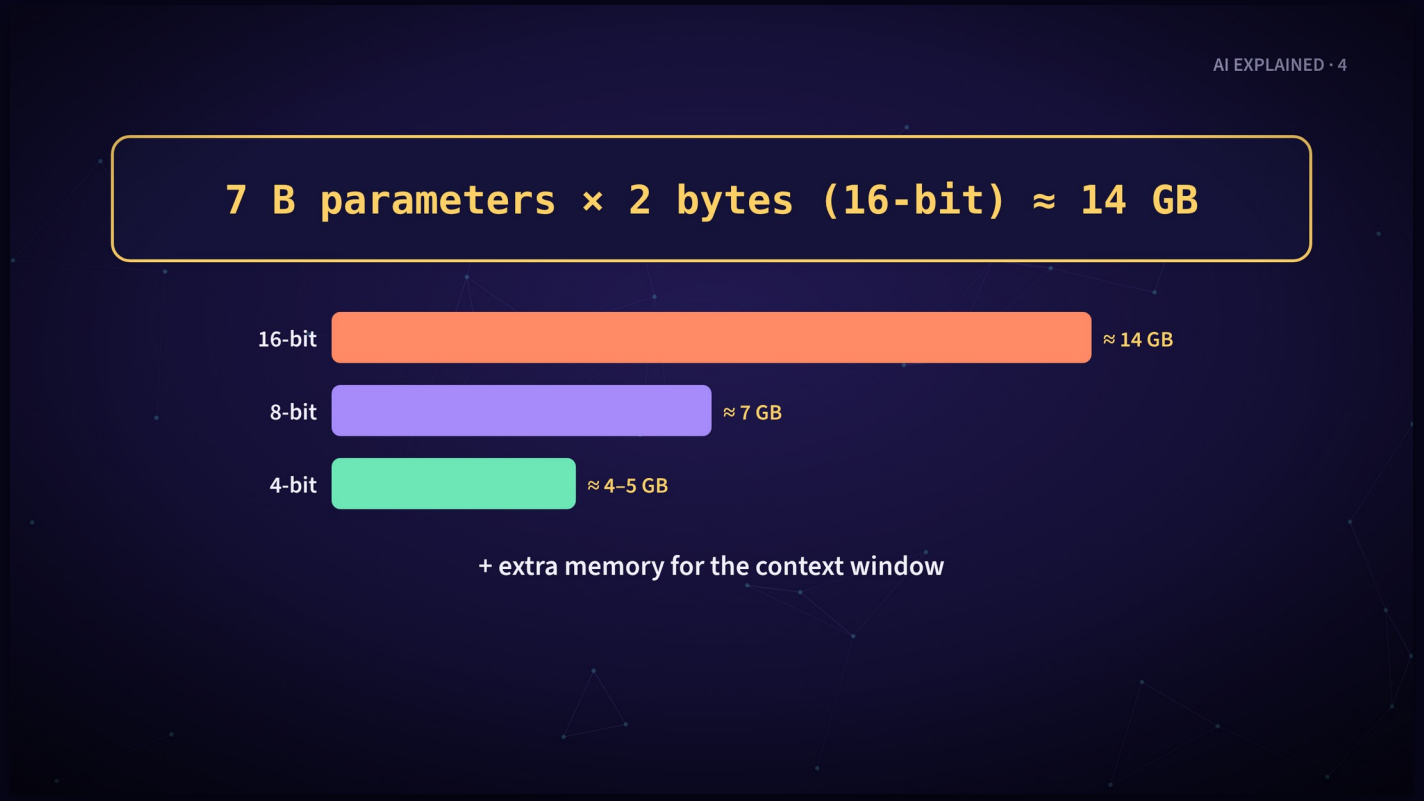
VRAM: The Model Must Fit



- The model must fit in VRAM to be fast
- Spill to system RAM → slow

05

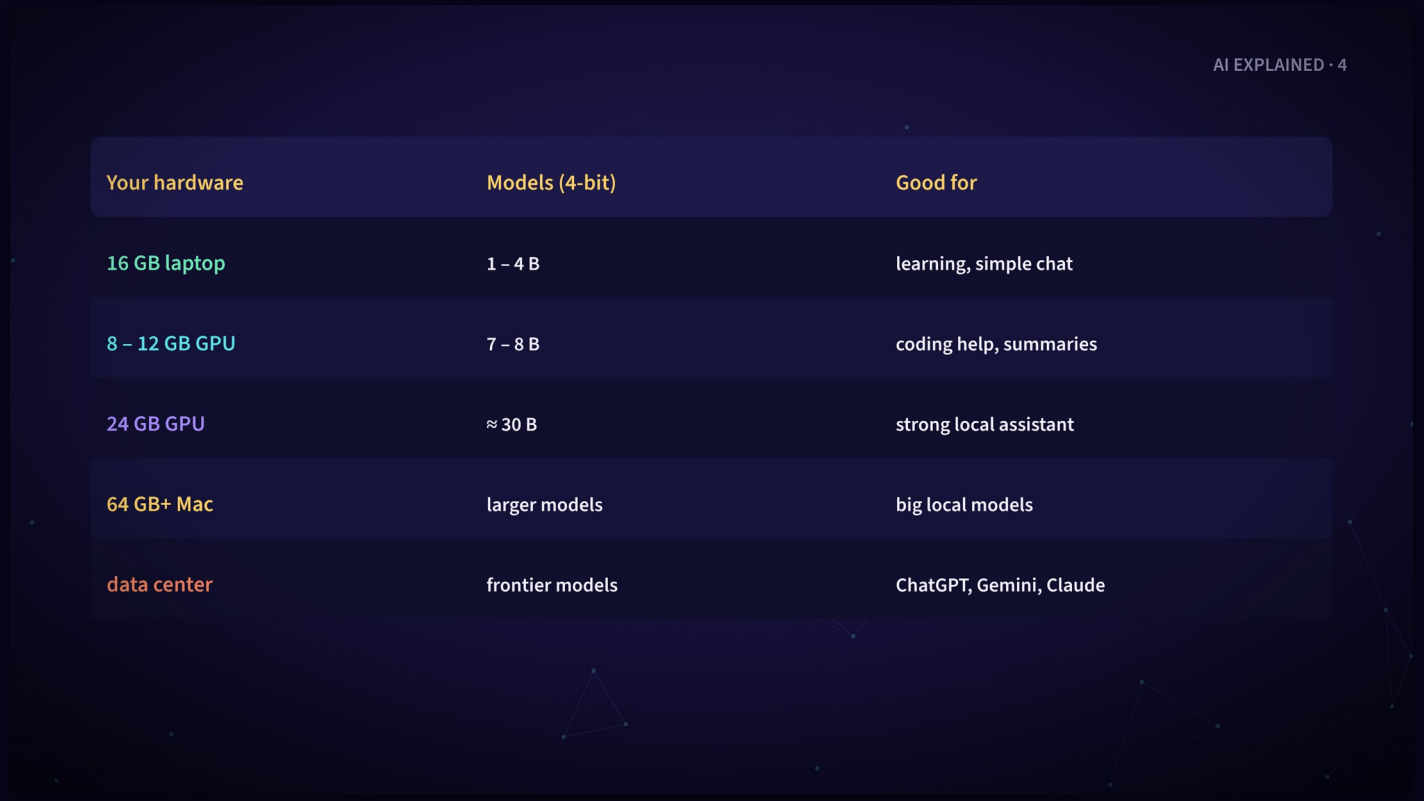
A Simple Memory Rule



- 16-bit: 2 bytes/parameter → 7B ≈ 14 GB
- 4-bit quantization → 7B ≈ 4–5 GB
- Add memory for context

06

What Can Your Computer Run?



- 16 GB laptop: 1–4B models
- 8–12 GB GPU: 7–8B · 24 GB GPU: ~30B (4-bit)
- 64 GB+ Mac: larger models

07

Inside an AI Data Center



- Clusters of 80 GB+ GPUs
- High-speed networking
- Electricity and cooling limits

08

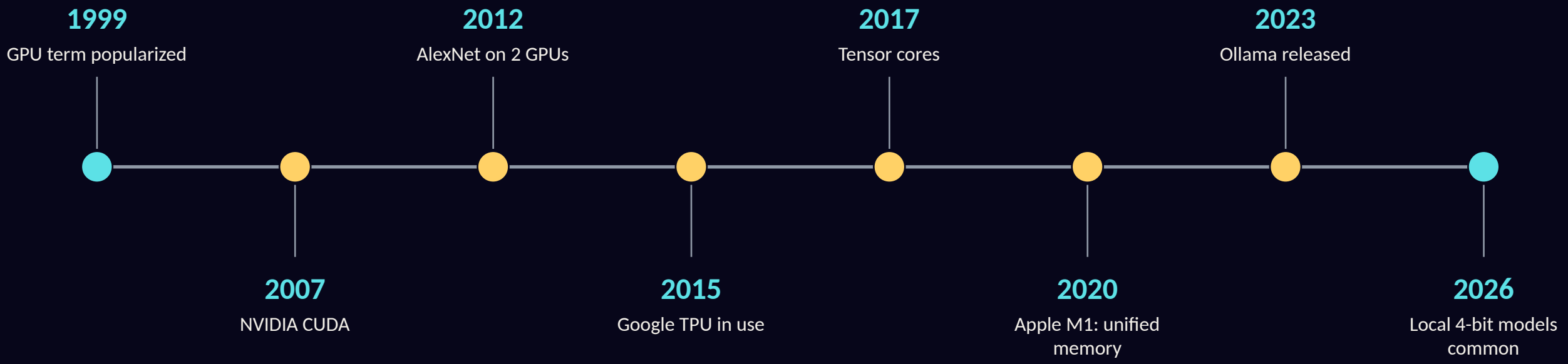
Good News: You Don't Need a Supercomputer



- Cloud chat: any device works
- Local: Ollama + open models
- Measure tokens per second

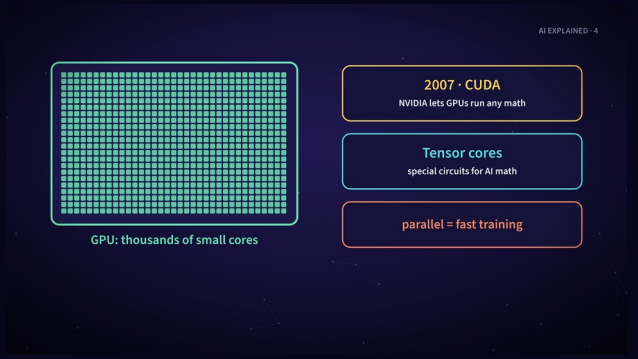
09

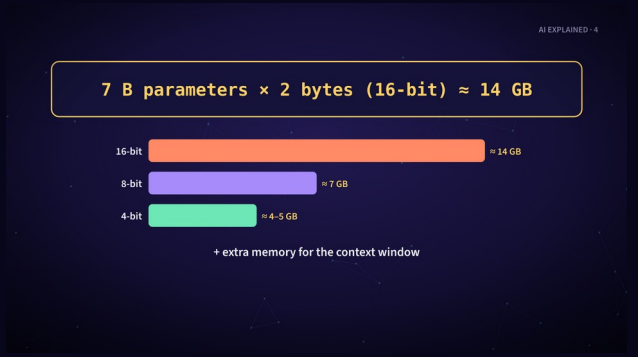
Timeline: 1999 to 2026



Sources and Further Reading

1. NVIDIA CUDA documentation — developer.nvidia.com/cuda-zone
2. Jouppi et al., “In-Datacenter Performance Analysis of a Tensor Processing Unit,” 2017
3. Ollama — ollama.com (model library and system requirements)
4. Hugging Face, quantization guide — huggingface.co/docs
5. Apple, Apple silicon and unified memory — apple.com





Thank you for watching

MyWebUniversity.com

Free technical education since 2000 — English and Dari

@MyWebUniversity · @MyWebUniversity-Dari