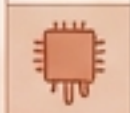


From Math to Memory

LEGACY COMPUTE

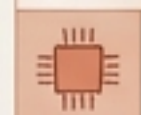


HIGH TOP

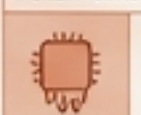


$$\frac{\partial \mathcal{L}}{\partial t} = \frac{1}{2} (z - h_0) y \approx 0 \quad z = \frac{d\alpha}{dt^2} - \frac{3}{2\pi} v_{lv}(t)$$
$$\Delta P_L = \ell_{\text{can}}(x_1) - \frac{\partial \ell}{\partial x_1} f_0(x) \approx \left(+ \frac{\partial m}{\partial m_0} x f_0(x^2) + p^{f_{lv}} t^2 \right) \cdot e^{-b_2 M_L}$$
$$-\frac{\partial \ell}{\partial t} \left(\ell_1^2 + \frac{1}{m_{10}} \ell_u(x) \right) = \begin{bmatrix} 0 & -1 & 0 & +0 \\ 0 & 1 & 1 & -0 \\ 0 & -1 & 0 & 0 \end{bmatrix}$$
$$\beta \pi r_i = \pi \alpha_i^2$$
$$\epsilon_0 = [dt_{w_0} + 1F_{001} \frac{\partial \ell}{\partial t}]$$
$$-\frac{dt}{dt} \left(\frac{1}{\delta} + \frac{1}{\delta} \frac{\partial \ell}{\partial t} \right)$$
$$\frac{d\mathcal{L}^{(0)}}{\partial t}(x) = \frac{\partial \ell}{\partial t} \left(\frac{1}{\delta} + \frac{1}{\delta} \frac{\partial \ell}{\partial t} \right)$$
$$\tau_0 = (H_0 + a_{01}/\partial \ell) \ell_0$$
$$X = \begin{bmatrix} x_{01} & x_{11} & x_{21} & x_{31} \\ x_{02} & x_{12} & x_{22} & x_{32} \\ x_{03} & x_{13} & x_{23} & x_{33} \end{bmatrix} = 0$$
$$P_c = \frac{m \cdot \partial_{\text{ck}}}{2} t_0$$
$$P_c = \frac{m \cdot \partial_{\text{ck}}}{2} t_0 \left(1 + \sum_{i=0}^m x_i^0 \right)$$
$$h_{01}(\alpha) \approx q^1 g x_1^{(2)}$$

HEAT BOTTLENECK



POWER CONSTRAINT



READY-CENTRIC LOGIC



DISTRIBUTED AI

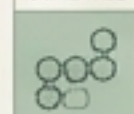


EFFICIENT DATA FLOW

LOW LATENCY

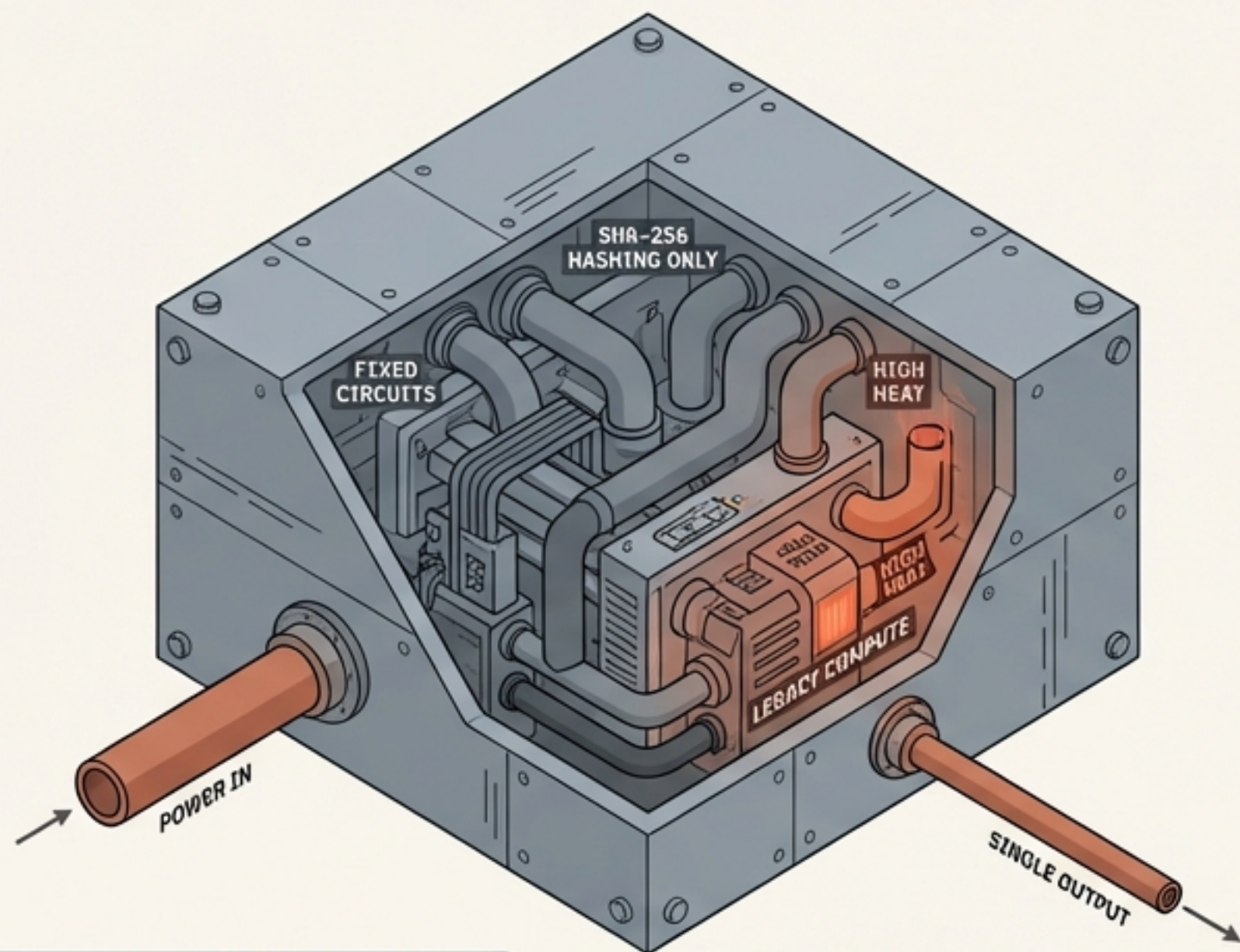


SCALABLE GRID



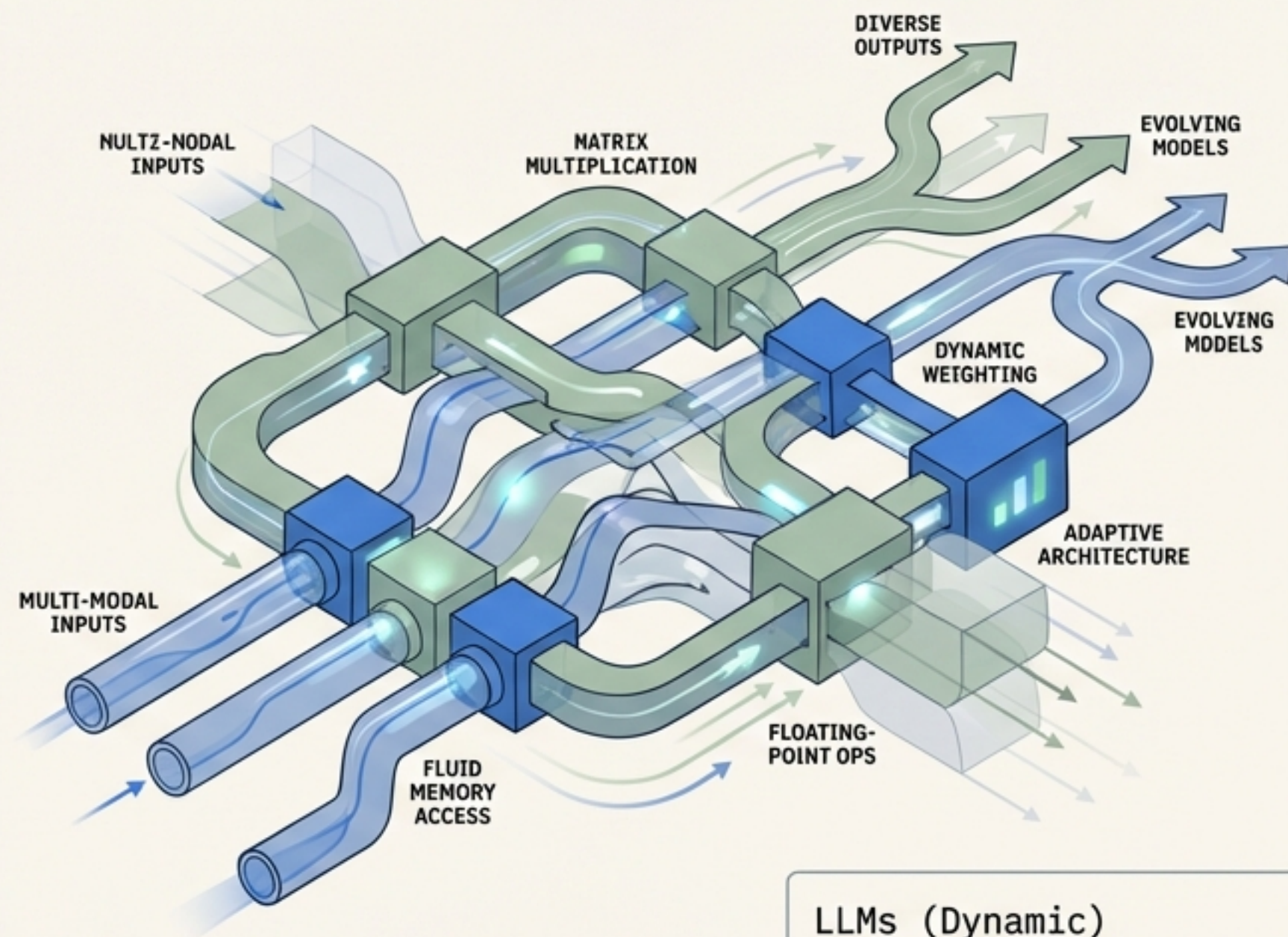
A visual document exploring the distributed AI revolution and the end of the compute-bound era.

The Fixed Logic Illusion



ASIC Miners (Fixed)

Etched silicon. Built for single hashing algorithms. Cannot process matrix multiplication.

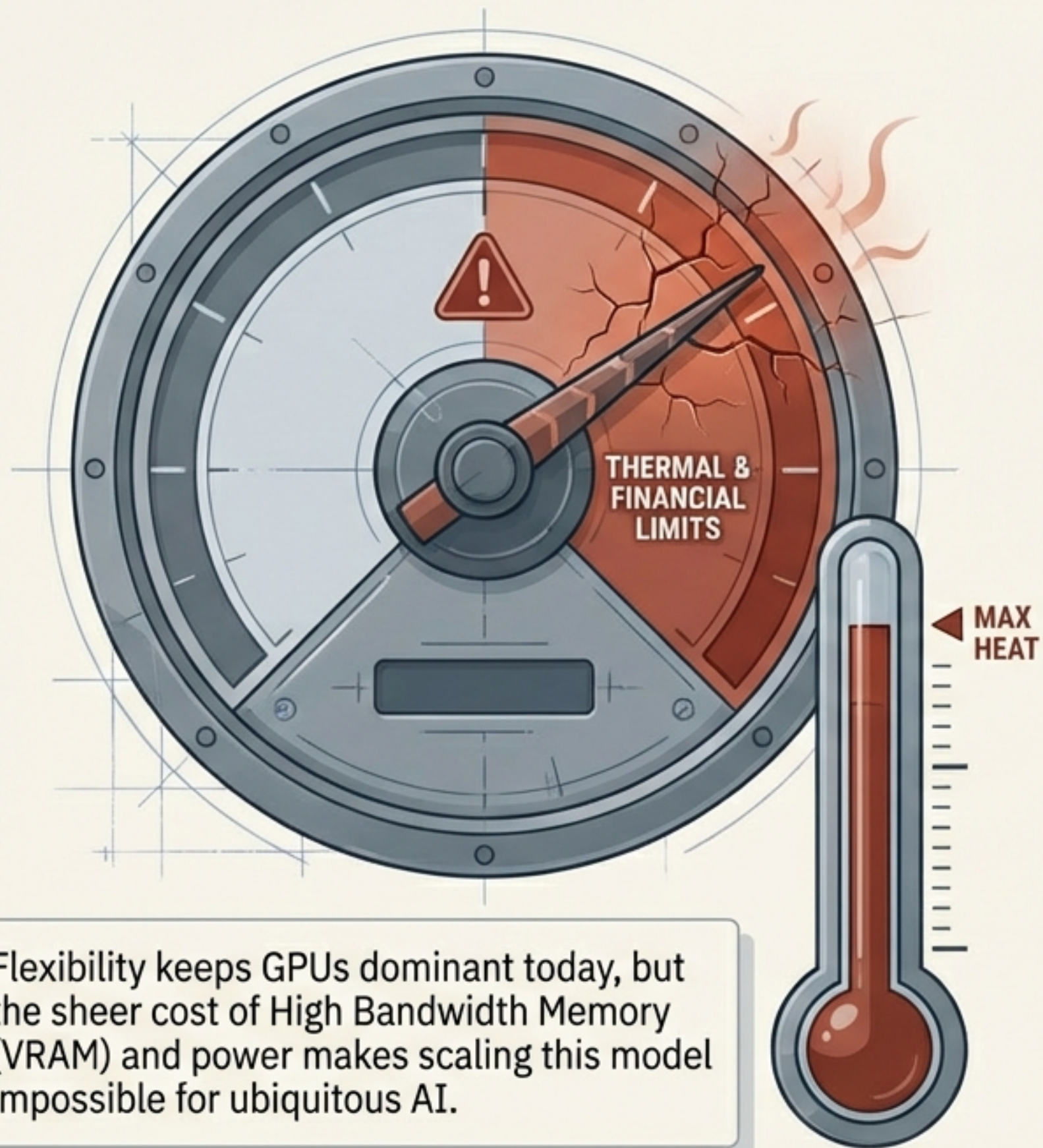


LLMs (Dynamic)

Require fluid weighting, billions of floating-point operations, and constantly evolving architectures.

You cannot update a Bitcoin miner to run an LLM any more than you can update a toaster to be a refrigerator.

The GPU Wall is Nearing



Flexibility keeps GPUs dominant today, but the sheer cost of High Bandwidth Memory (VRAM) and power makes scaling this model impossible for ubiquitous AI.

Power Draw

700W+ per card

(Requires massive AC cooling infrastructure)

Cost Barrier

\$40,000+

per high-end Nvidia H100

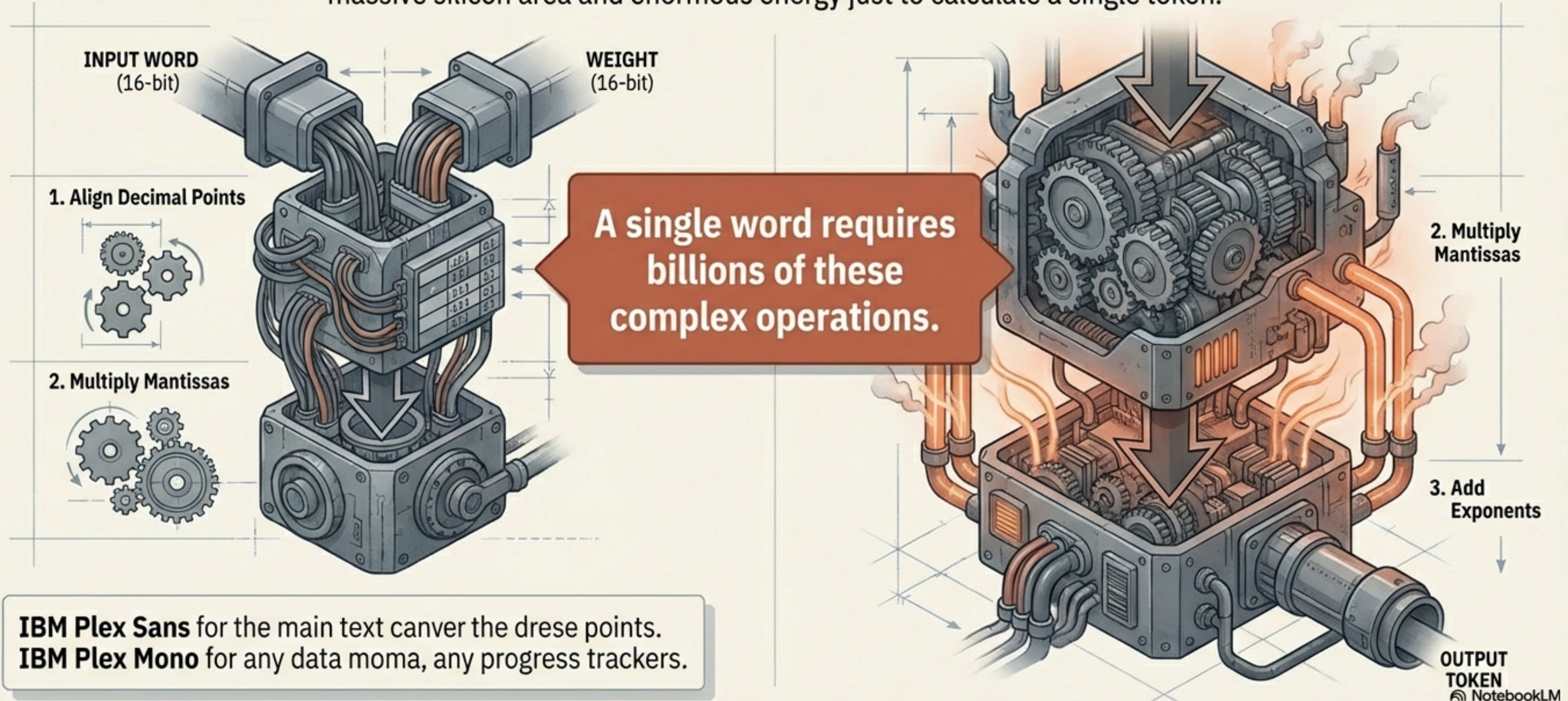
The Bottleneck

Compute-Bound Architecture

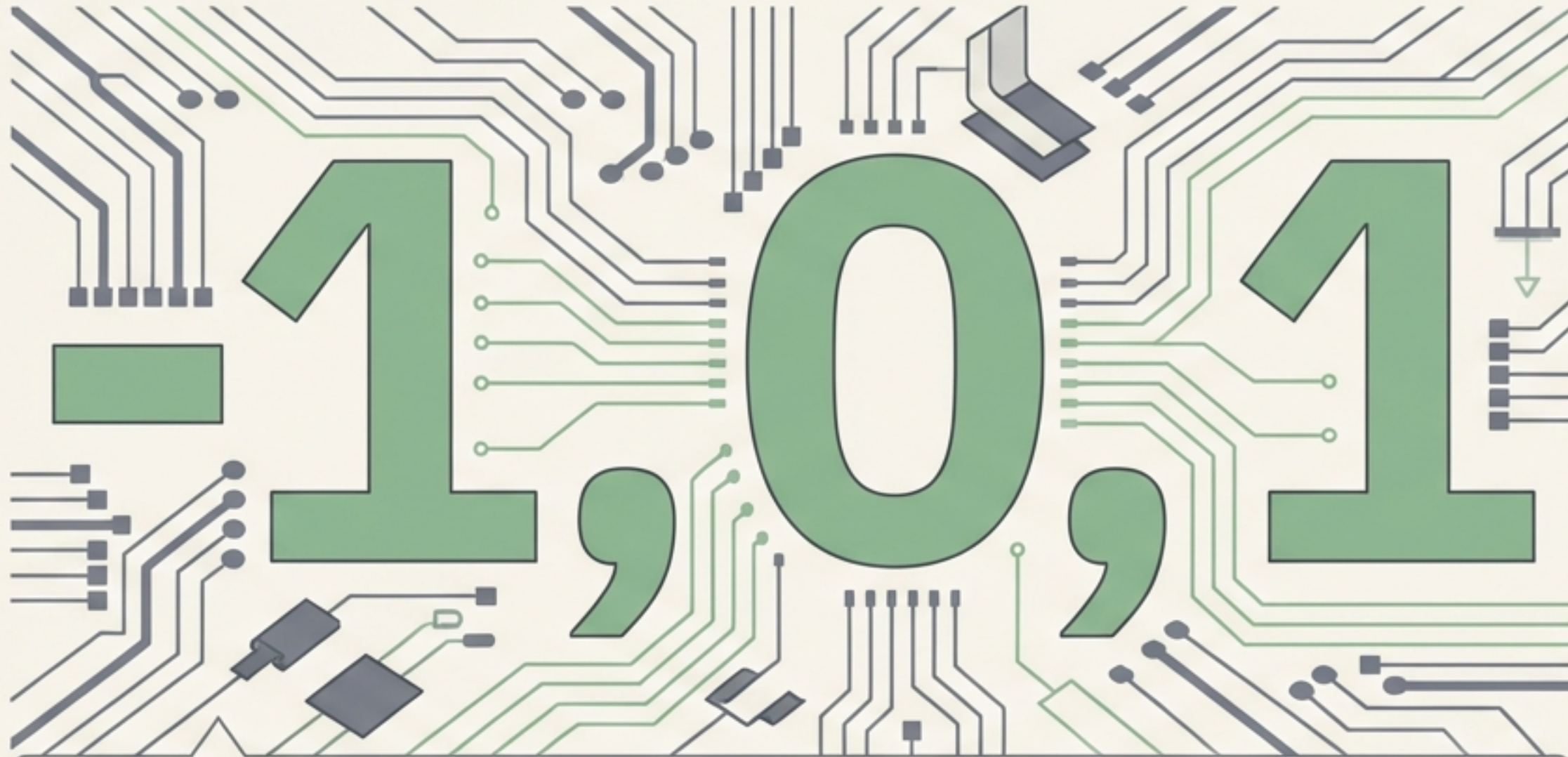
(Trillions of transistors dedicated to brute-force math)

The Heavy Cost of Floating-Point Math

Standard AI relies on 16-bit arithmetic. This compute-bound approach requires massive silicon area and enormous energy just to calculate a single token.



The Shift to Ternary Logic

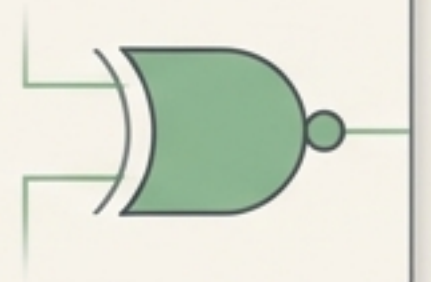


By shifting from 16-bit math to 1.58-bit Ternary Logic, complex multiplication is replaced by simple sign-flipping. When the weight is '0', the logic gate simply turns off (Sparsity), saving 100% of the energy for that cycle. [IBM Plex Sans]

1.58-bit Architecture

Replacing 65,536 possible values with just three.

three



90% Energy Reduction

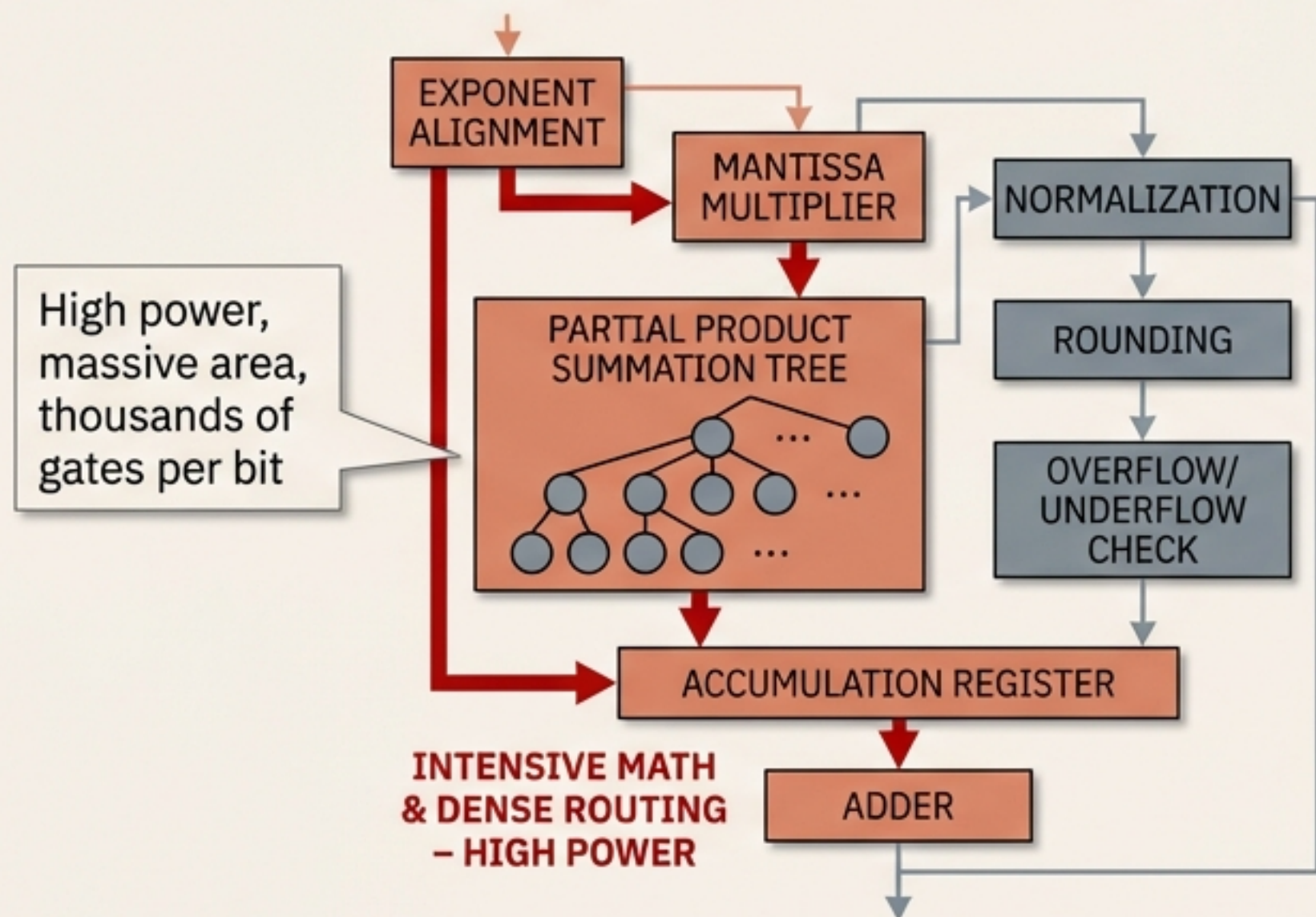
Avoiding the multiplier completely.

90%

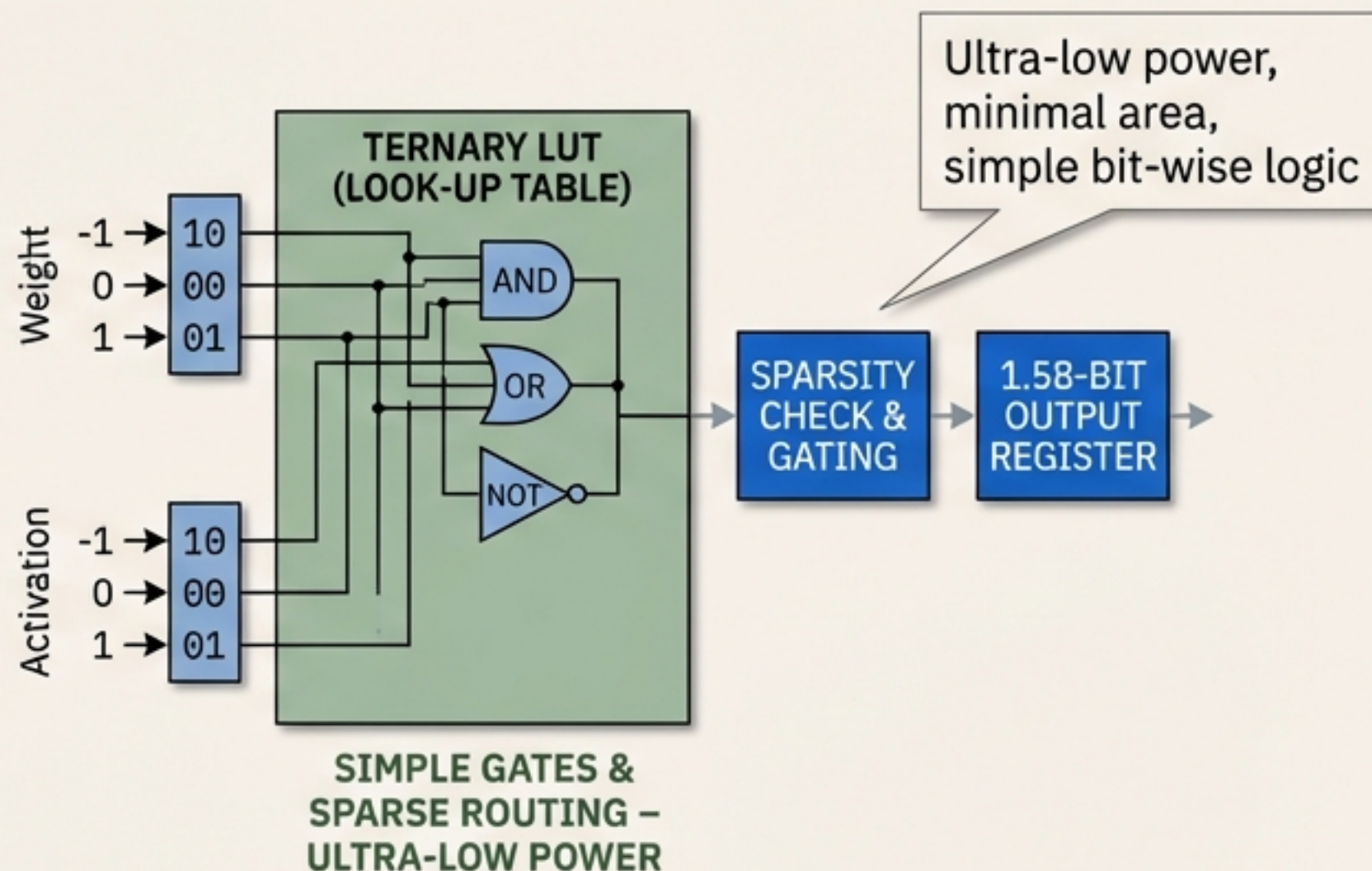


Next-Generation AI Logic: Comparing Traditional 16-Bit MAC with Ternary LUT (1.58-bit)

TRADITIONAL 16-BIT FLOATING-POINT MAC



TERNARY LOGIC GATE LUT (1.58-BIT)

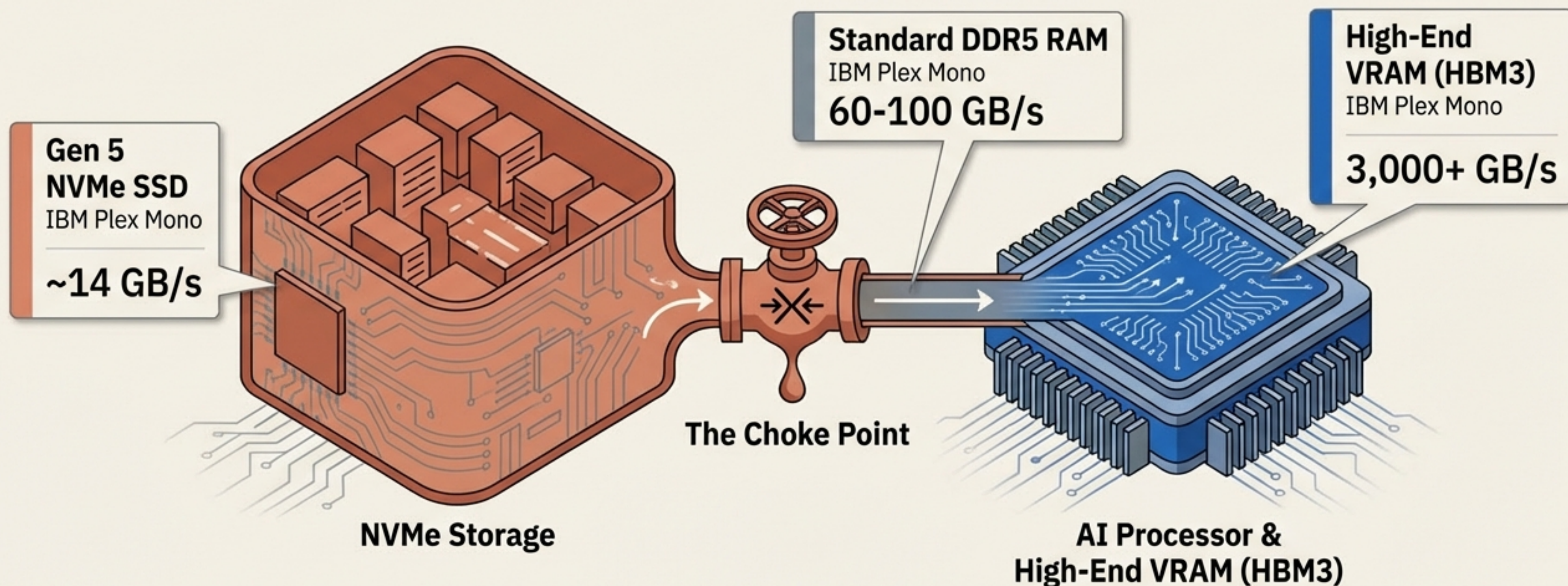


FPGAs (Field-Programmable Gate Arrays) are naturally suited for Look-Up Table (LUT) logic, replacing trillions of math transistors with thousands of programmable gates

The Great Memory Choke Point

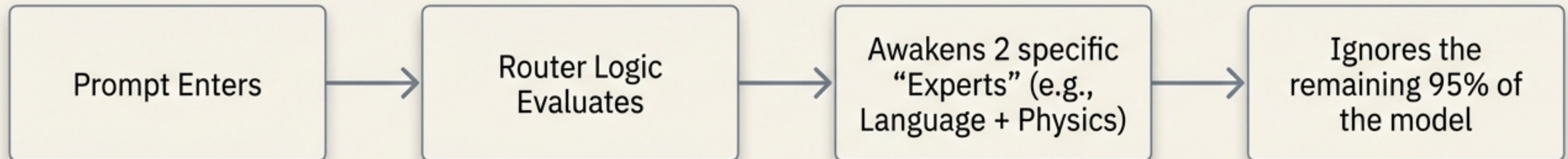
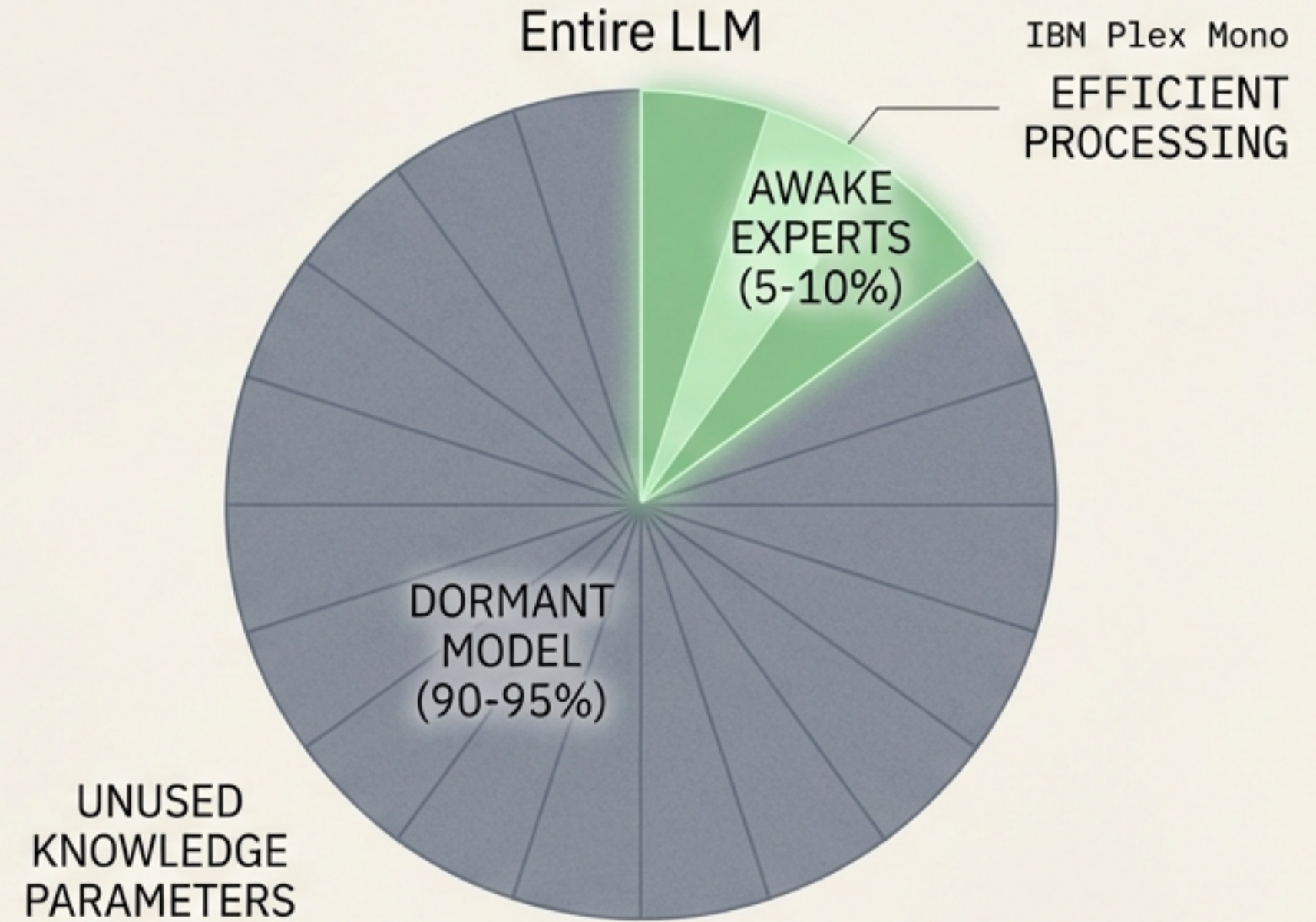
Running a massive language model directly from an NVMe drive leaves the processor waiting 99% of the time. The speed gap between storage and VRAM is the final physical barrier.

[IBM Plex Sans]



Awakening Only What is Needed

Using the Mixture of Experts (MoE) strategy, only 5% to 10% of a model is 'awake' for any given word.



IBM Plex Mono: TARGETED COMPUTE & MINIMAL RESOURCE USE

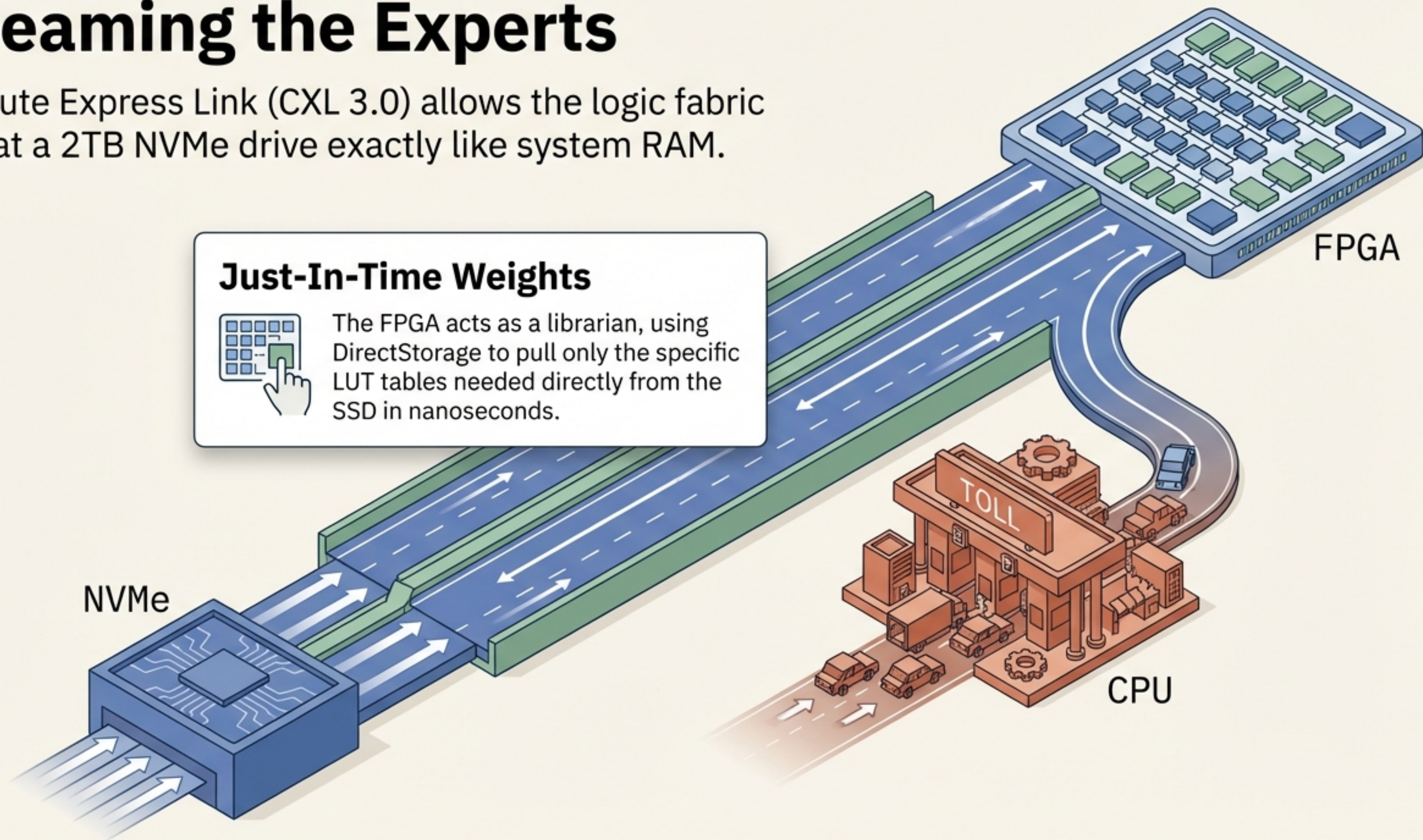
Streaming the Experts

Compute Express Link (CXL 3.0) allows the logic fabric to treat a 2TB NVMe drive exactly like system RAM.

Just-In-Time Weights



The FPGA acts as a librarian, using DirectStorage to pull only the specific LUT tables needed directly from the SSD in nanoseconds.



Deconstructing the Dream Silicon

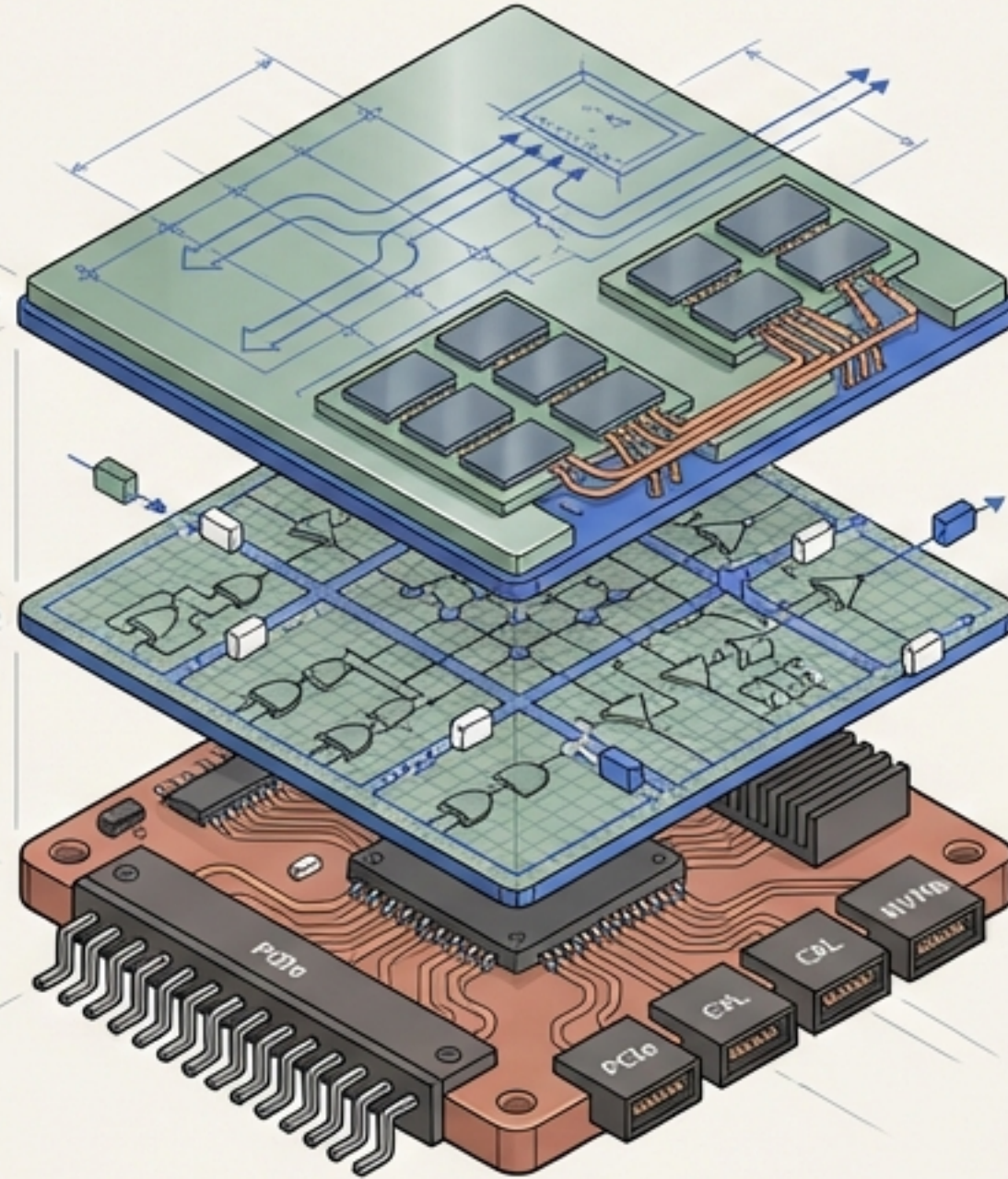
Merging memory, routing, and ternary logic into a single vertically integrated chip creates a pocket-sized supercomputer operating at under 15 Watts.

Top Layer: HBM3 (The Scratchpad)

Ultra-fast local memory for the active conversation.

Middle Layer: FPGA Fabric (The Traffic Controller)

Programmable grid routing the Mixture of Experts on the fly.



Bottom Layer: CXL Controller (The High-Speed Pipe)

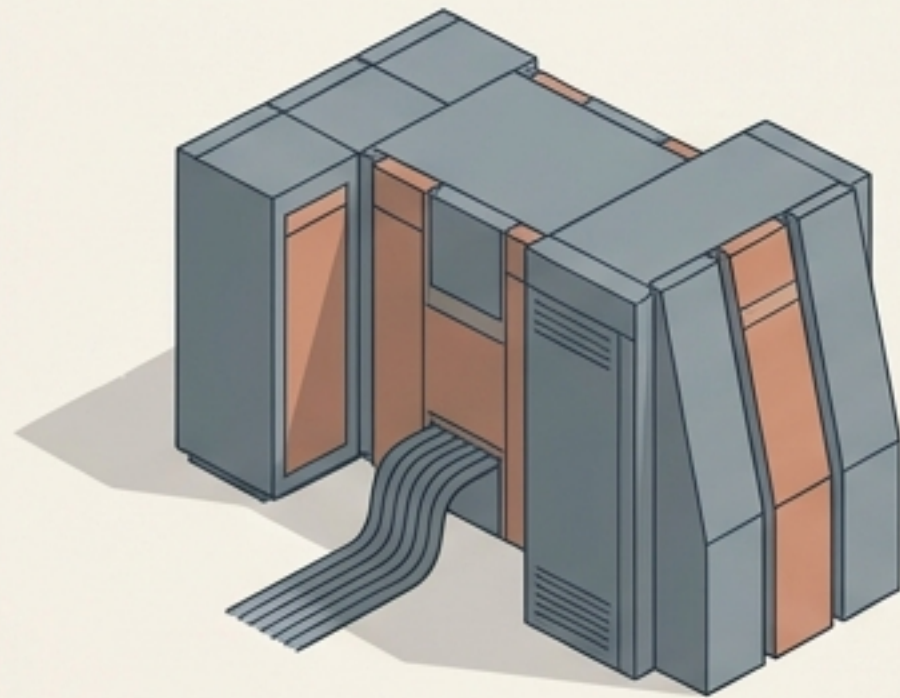
Direct, zero-latency access to massive NVMe storage.

Merging memory, routing, and ternary logic into a single vertically integrated chip creates a pocket-sized supercomputer operating at under 15 Watts.

The End of the Server Farm

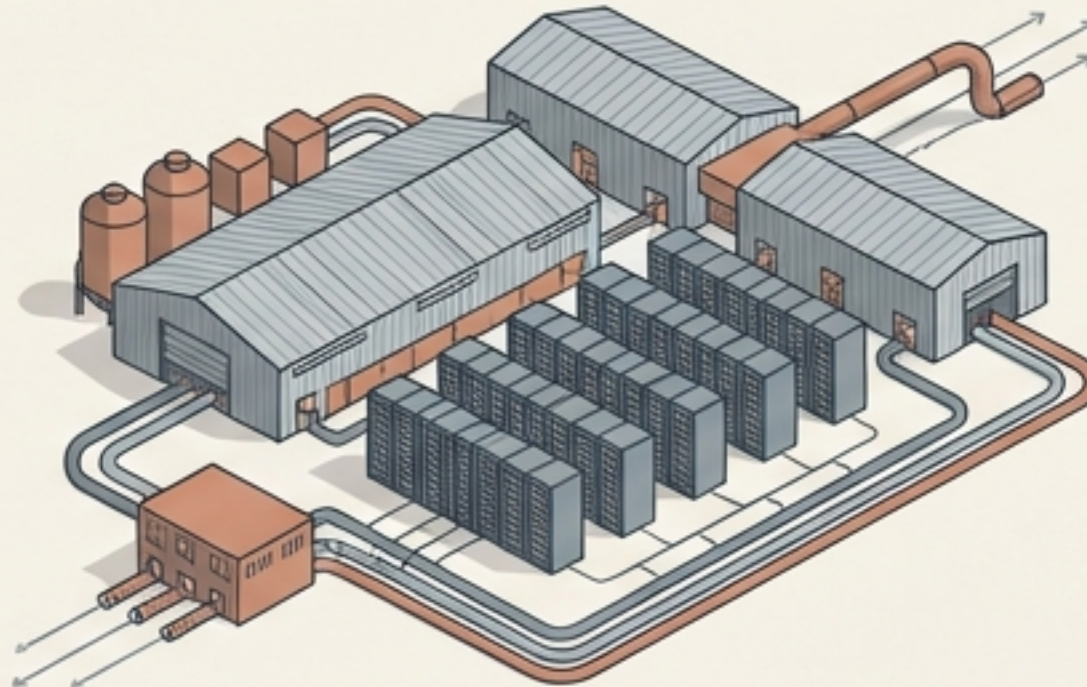
The 30-year leap from fixed supercomputers to localized, adaptive SoCs kills the centralized computing monopoly. Computing returns to the edge.

1996



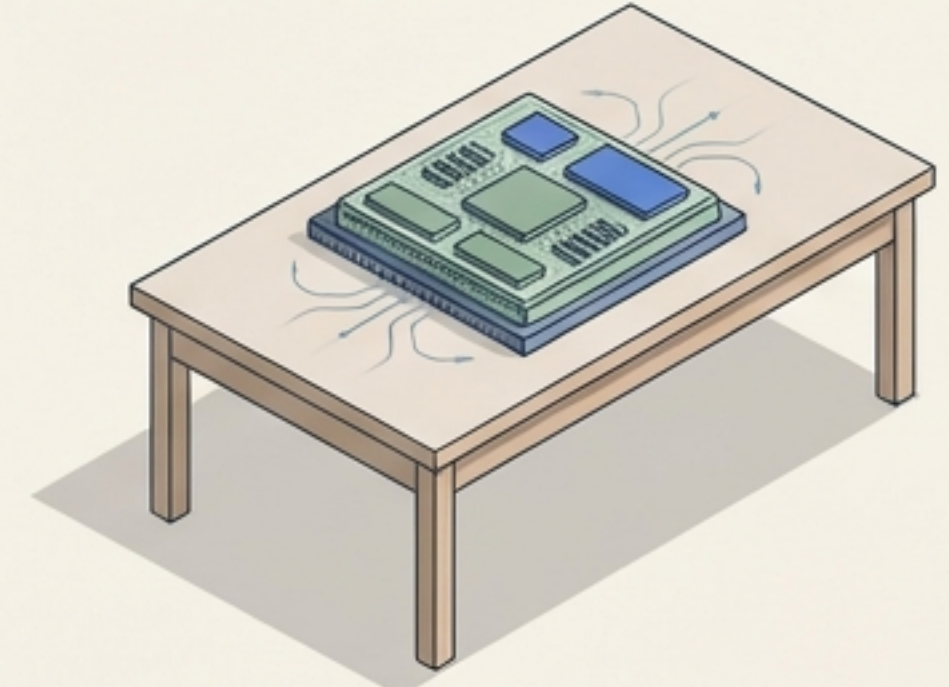
Cray T3E Supercomputer
(Megawatts of power)

2024



Cloud Server Farm
(Data leaves the house, \$20/month)

2026



"Dream Box SoC"
(Data stays on chip, 10W power,
one-time hardware cost)

Merging memory, routing, and ternary logic into a single vertically integrated chip creates a pocket-sized supercomputer operating at under 15 Watts.

The Community Brain

Decentralized intelligence distribution, leveraging localized expert models for collective, efficient computing.

The Protocol

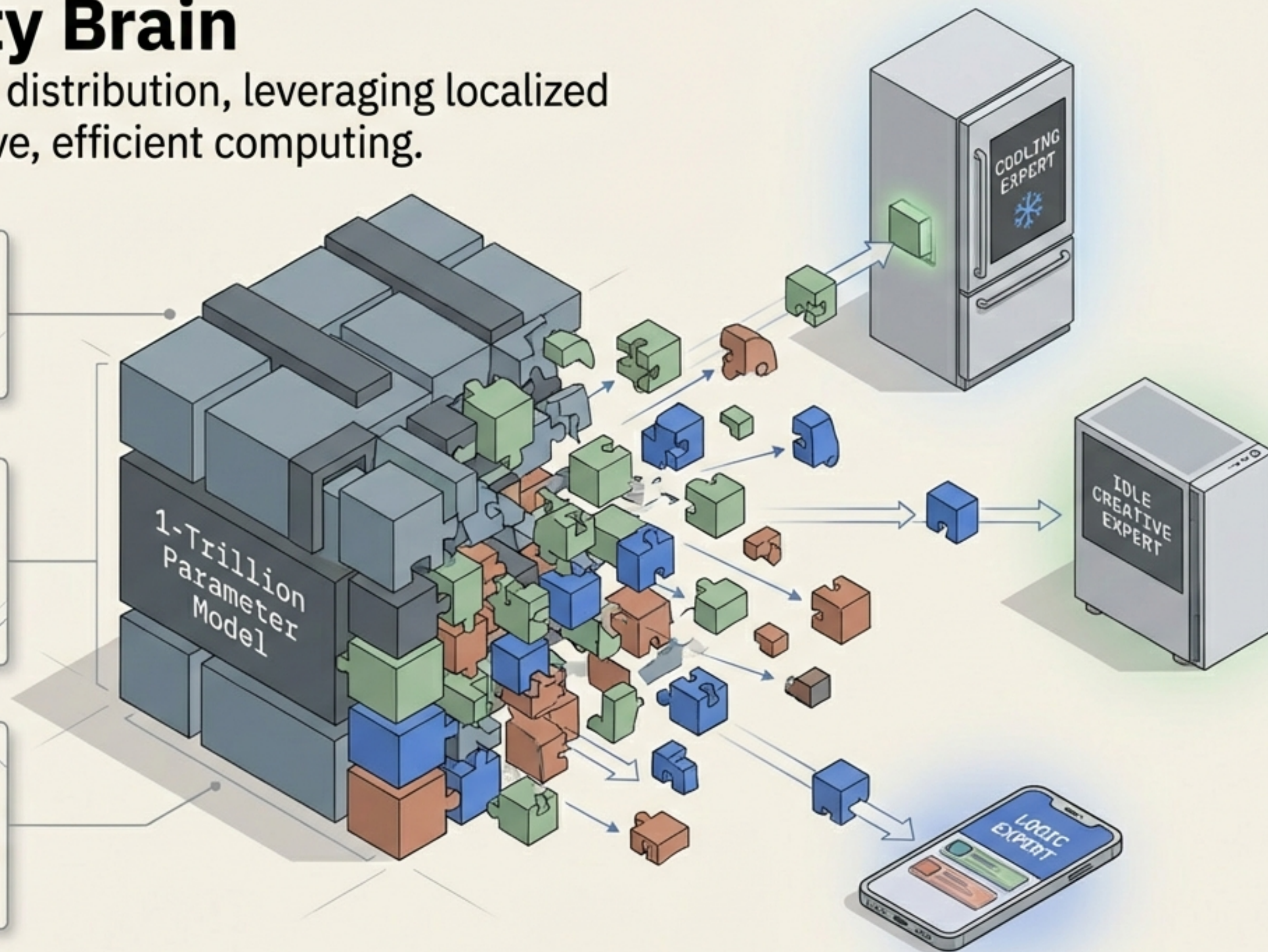
Expert P2P routing
(Swarm Inference).

The Sharding

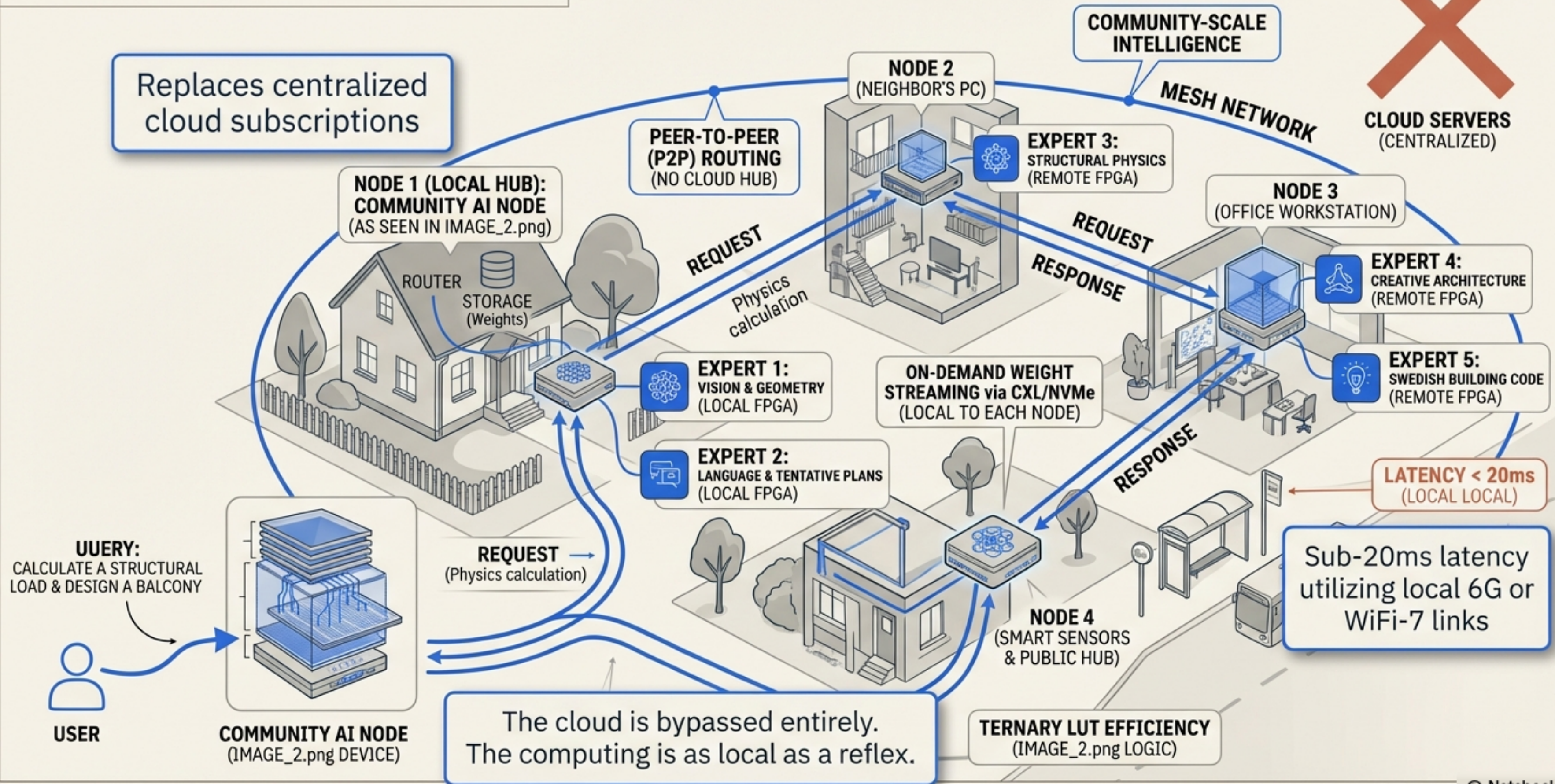
A massive model is broken down into 1,000 discrete expert blocks.

The Execution

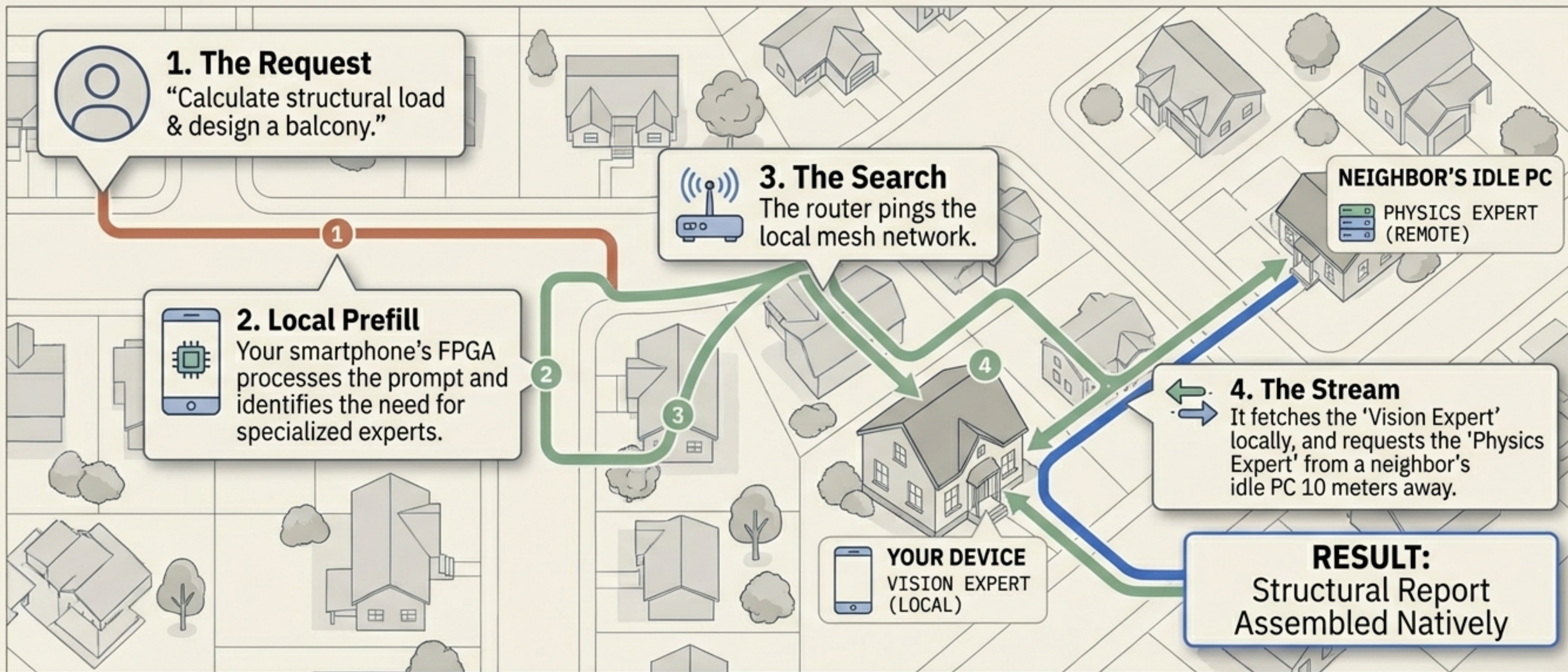
Your device holds the logic expert. Your neighbor's idle PC holds the creative expert.



Local AI Intelligence



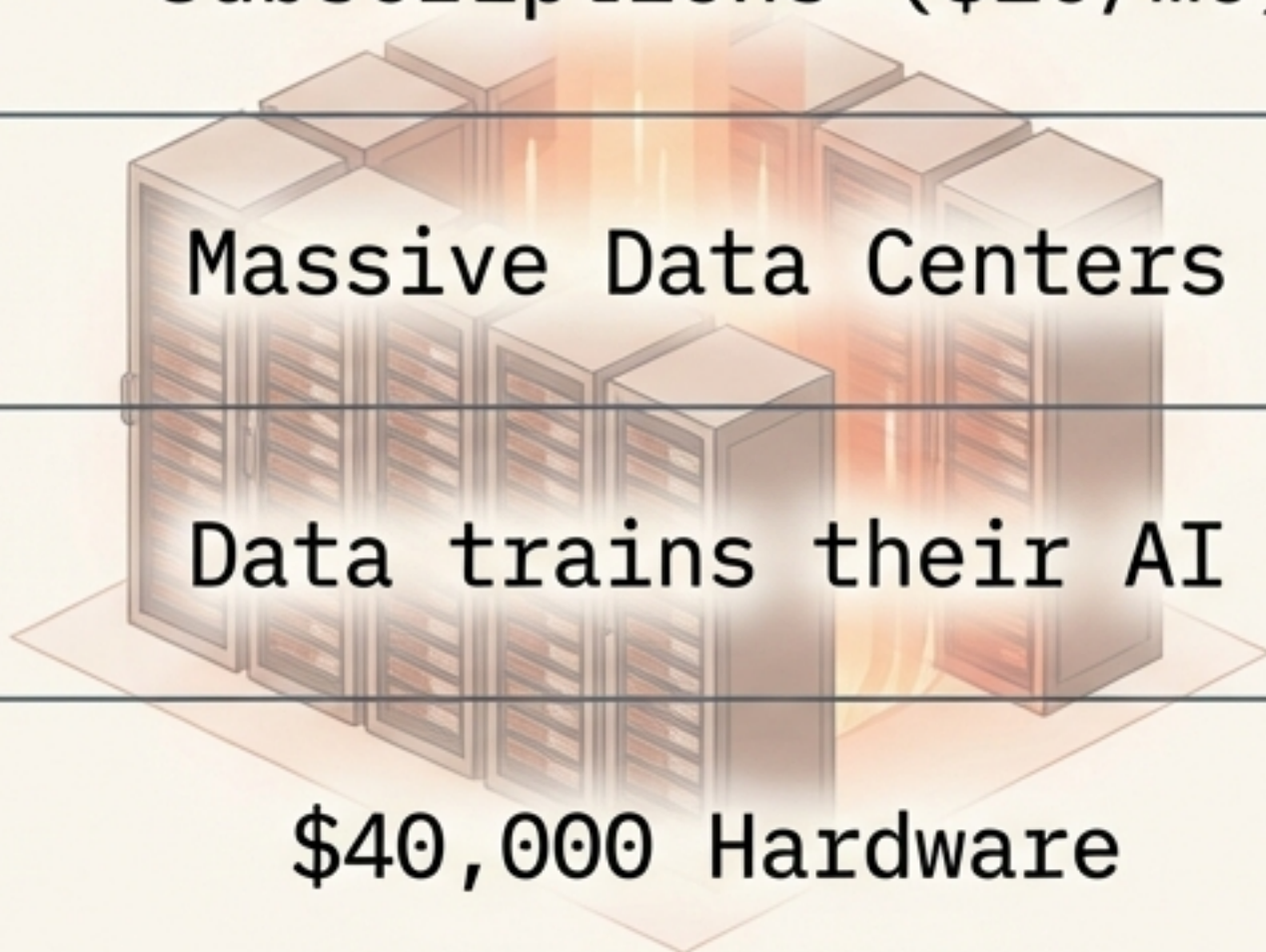
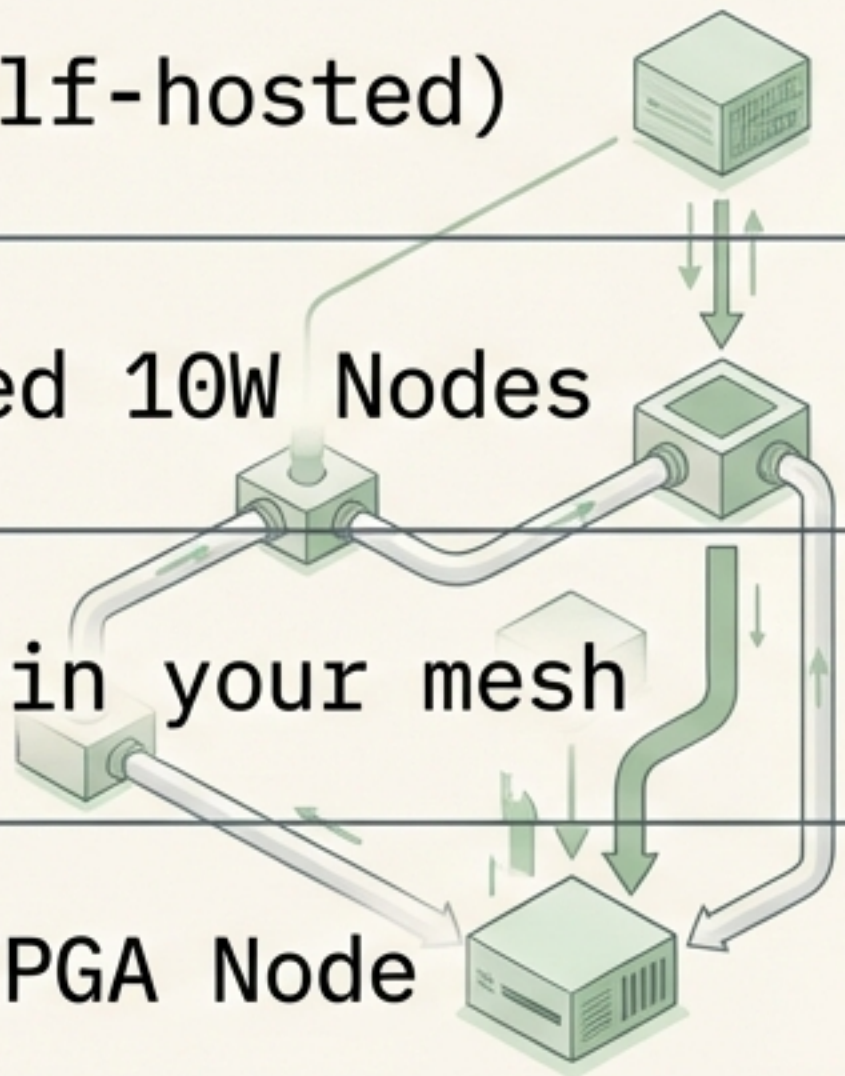
Tracing the Neighborhood Prompt



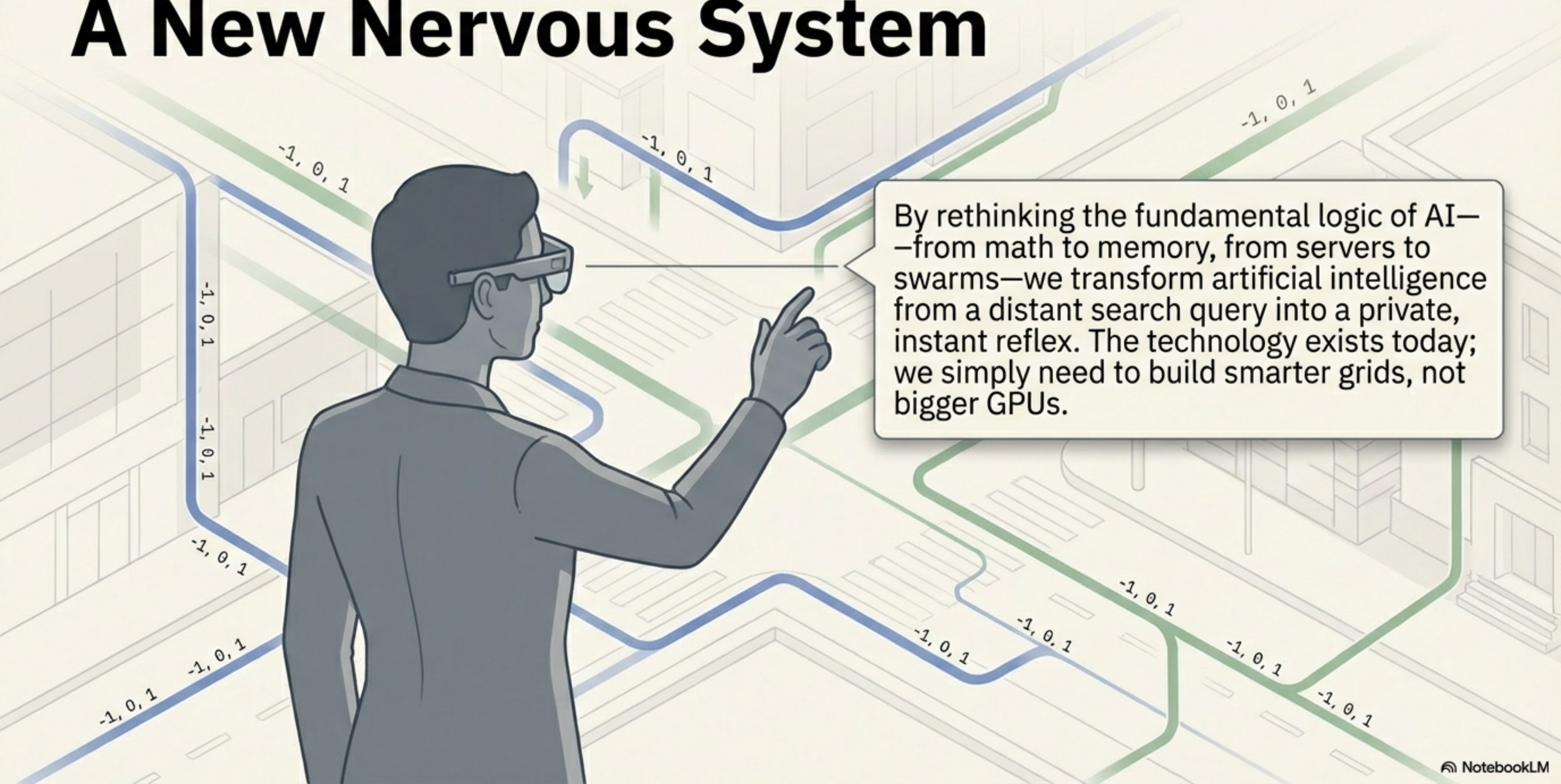
The experts work in parallel via the local mesh. The resulting structural report is assembled natively on your device. Zero data leaves your neighborhood.

The Economic Shift

The transition from compute-bound mainframes to memory-bound mesh networks democratizes AI. It ceases to be an expensive rental service and becomes a privately owned utility.

Cloud Model (2024)	Mesh Model (2026)
Subscriptions (\$20/mo)	Free (Self-hosted)
 Massive Data Centers	 Distributed 10W Nodes
Data trains their AI	Data stays in your mesh
\$40,000 Hardware	\$800 FPGA Node

A New Nervous System



By rethinking the fundamental logic of AI—from math to memory, from servers to swarms—we transform artificial intelligence from a distant search query into a private, instant reflex. The technology exists today; we simply need to build smarter grids, not bigger GPUs.