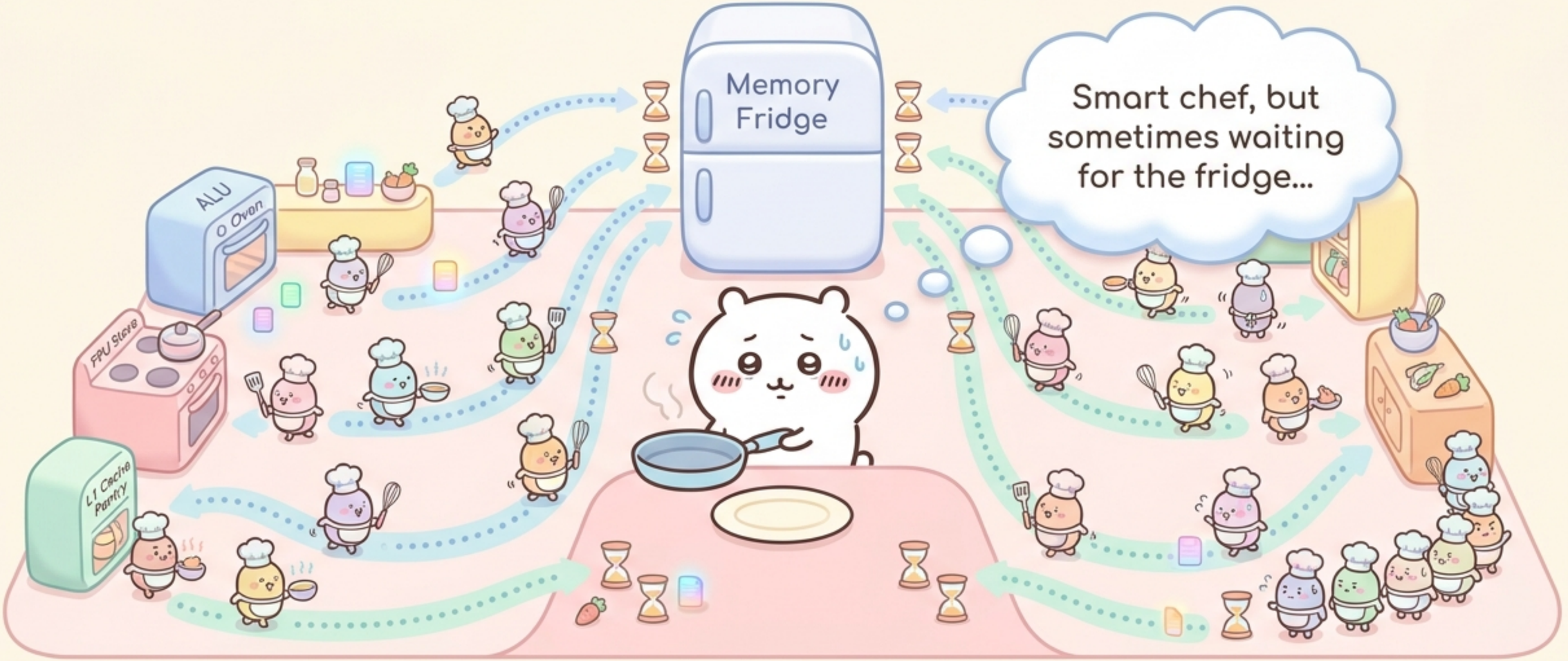
GPU Philosophy: 1,000 Independent Chefs

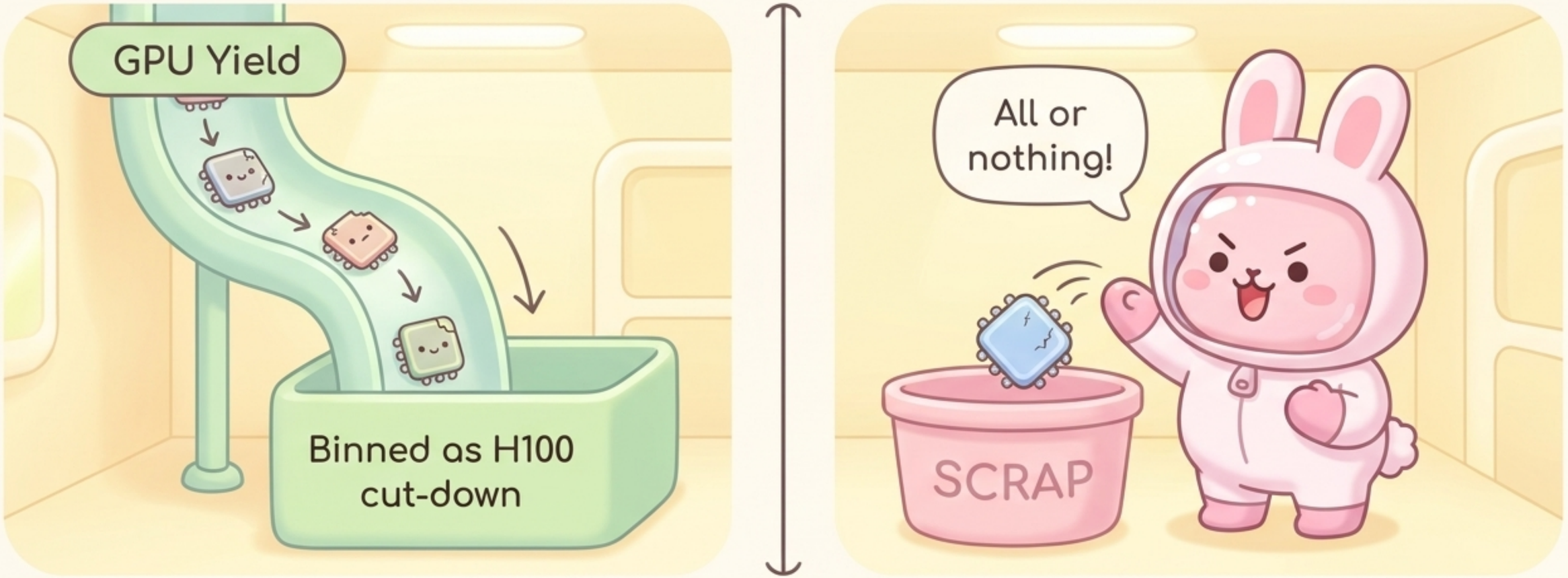


Hardware handles dynamic scheduling. Massive parallelism (SIMT), but prone to idle periods during memory round-trips.

TPU Philosophy: The Perfect Assembly Line

Hardware is simple (ASIC). The XLA Compiler is an omniscient scheduler. Data flows seamlessly without runtime prediction or idle waiting.

The Yield Bottleneck

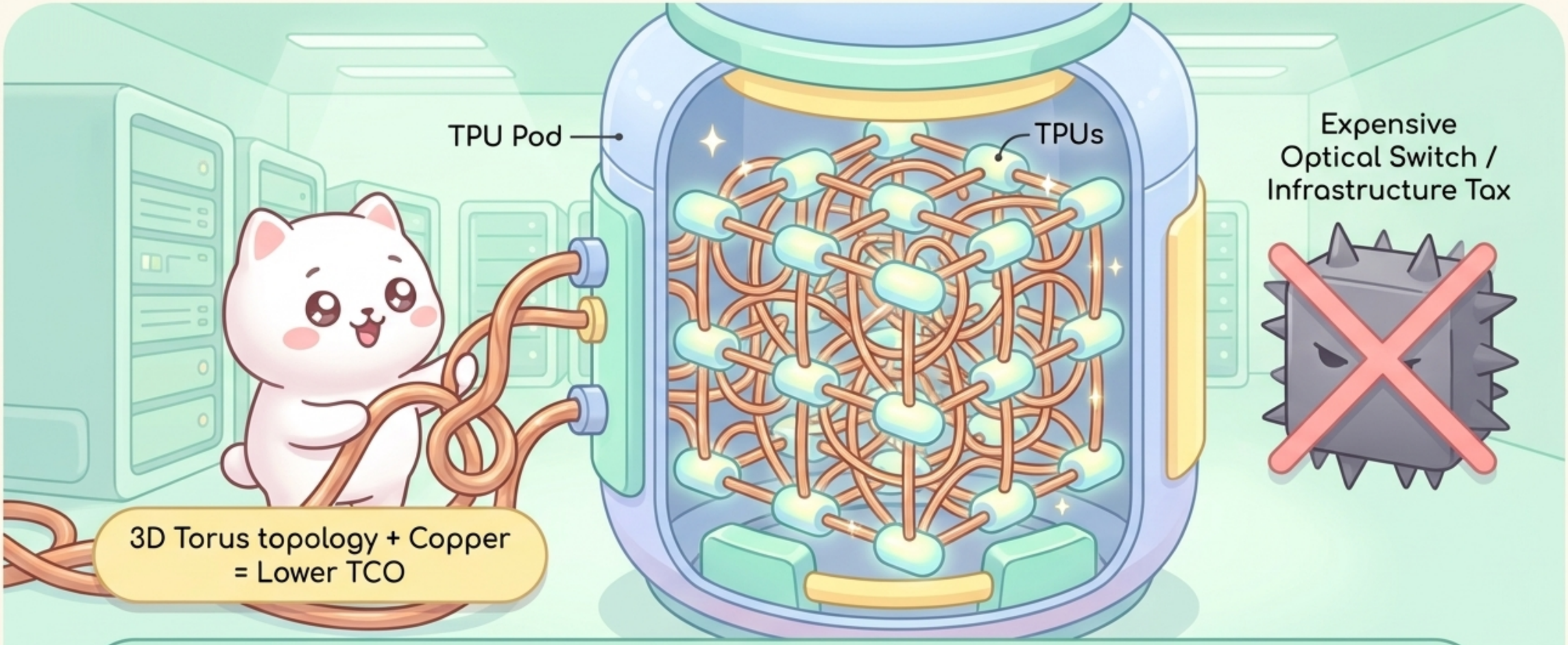


TPU Pods require strict system-level coherence. Every chip must perform identically. Unlike GPUs, defective TPUs cannot be downgraded—they are purely scrap.

The Supply Chain Reality

HBM (Memory) and TSMC CoWoS advanced packaging are severely bottlenecked.
Capacity is allocated to the highest volume buyer—Nvidia. TPU remains a secondary customer.

The TPU Pod: Network over Silicon

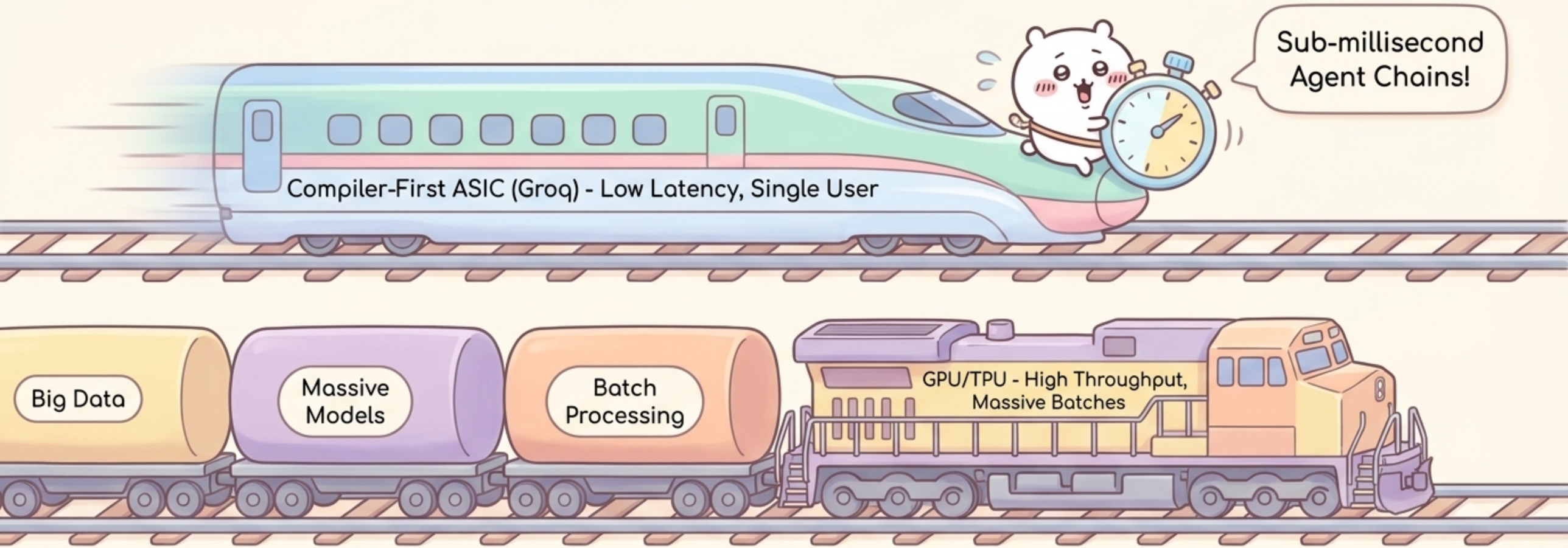


TPU's true edge isn't the single chip—it's the system. Connecting chips directly via copper interconnects eliminates expensive networking switches (the infrastructure tax).

The Invisible Moat

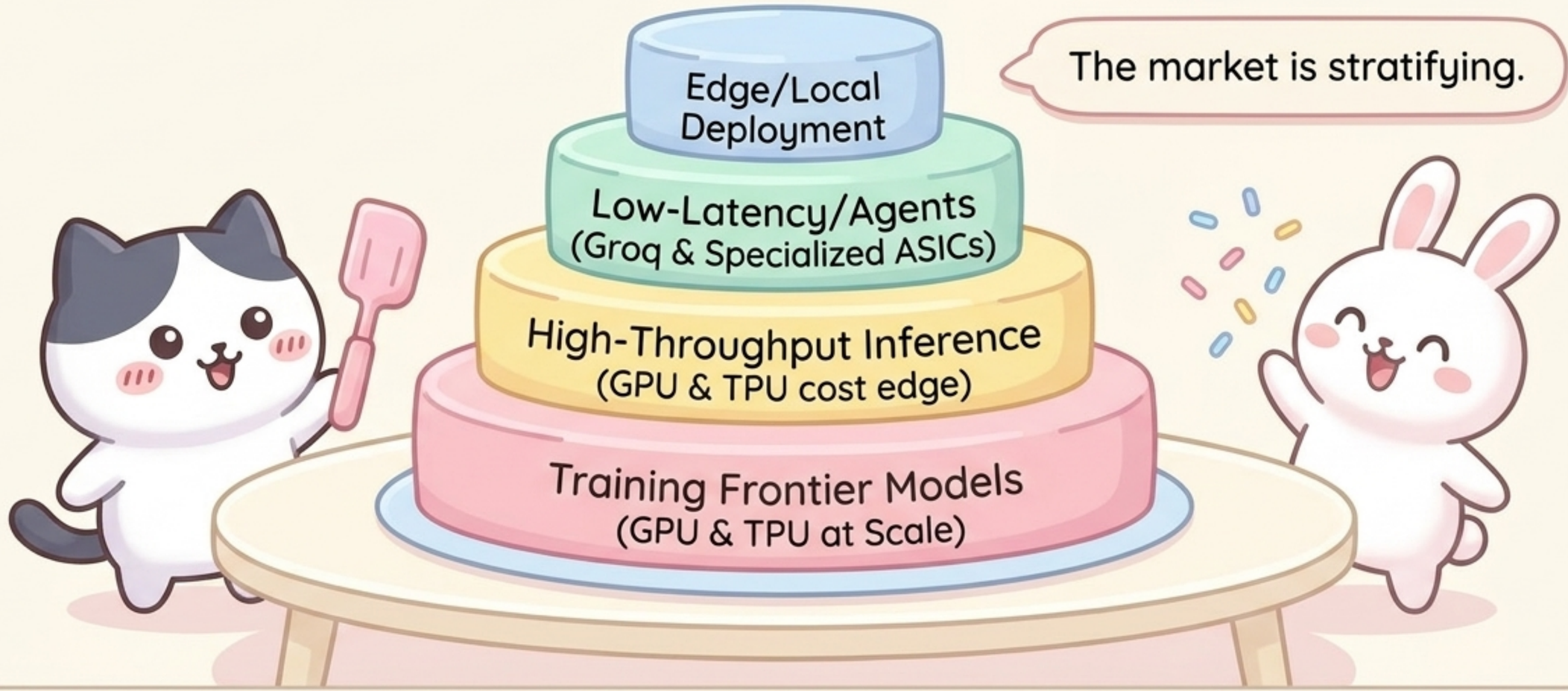
External developers cannot easily fix XLA bugs. Without deep JAX/XLA expertise (scarce human capital), TPU hardware advantages evaporate.

Groq: The Need for Speed



Built by ex-TPU compiler engineers. Groq hardware is even simpler than TPU, trading massive throughput for ultra-low per-token latency. Ideal for real-time AI agents.

The AI Infrastructure Layer Cake



Takeaway for builders: Don’t bet your architecture on a single hardware assumption. Build hardware-agnostic software to shift as economics and capabilities evolve.