Main memory and video memory parted ways long ago—CPUs stayed behind to carry water by the bucket.

1 Capacitor + 1 Transistor: DRAM constantly leaks and needs refreshing, inherently making it slower than the CPU.

50–80
Nanoseconds

To a mighty CPU core, this latency is a glacial, agonizing wait.

CPUs handle **serial work** (like 1+1, then +2). Every step needs the previous step's result before it can proceed.

Step 1 → Step 2 → Step 3

1 Mighty Core + 1 Narrow Lane (64-bit bus width) = = The CPU demands its data right now, but can only fetch it one bucket at a time.

GPUs abandon individual speed to command a lockstep legion of thousands.

If a teacher wastes 1 minute of class, 45 students wait. But they wait together. Total time wasted? Still 1 minute.

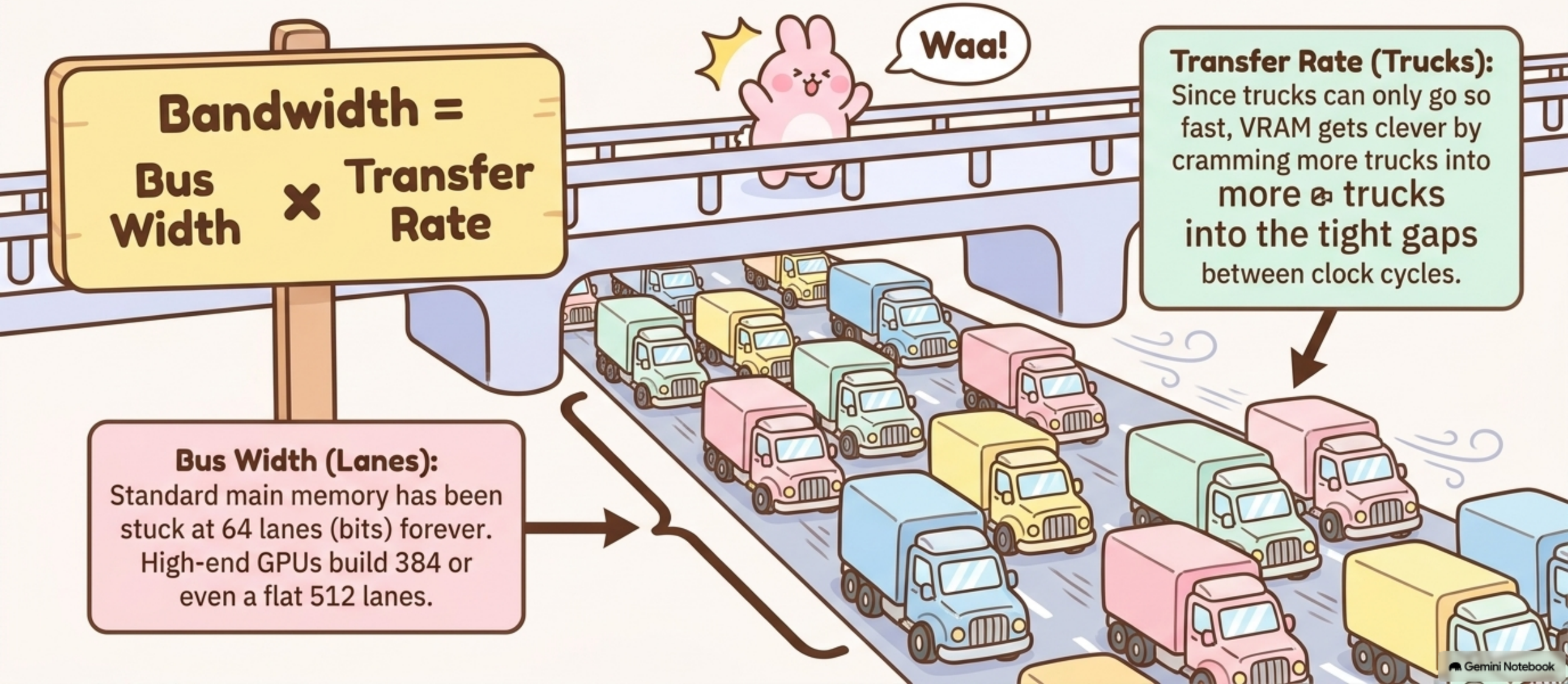
CPU Style vs. GPU Style

CPU Style	GPU Style
Few Cores Serial Tasks Hates Waiting (Low Latency required)	Thousands of Cores Parallel Tasks Doesn't Care About Waiting (High Latency acceptable)

If 10,000 cores fetch data at once, a 200-nanosecond latency is perfectly fine.

GPUs don't need fast individuals; they need massive coordinated movement.

Feeding thousands of cores simultaneously requires building massive highway systems.



Pushing transfer rates too aggressively leads to embarrassing physical limits.

The Retreat: GDDR7 safely retreated to a steadier 3-level system (PAM3) to keep its pants intact.

When highways run out of room, AI demands we bolt a skyscraper directly onto the GPU.

1 **The Breaking Point:** AI chips require a 1,024-bit bus width per stack. This means nearly 4,000 wires—too many for standard fiberglass motherboards to handle.

2 **The Interposer:** A mirror-smooth slab of pure silicon acting as the new foundation to handle the extreme wiring density.

3 **TSV (Through-Silicon Vias):** Microscopic copper pillars drilled directly through vertically stacked memory dies.

The 2 Million RMB Synthesis: AI servers cost millions because you aren't just buying the GPU—you are paying for this incredibly complex memory roof. As a side effect, this takes up standard for memory factory lines, quietly raising the price of everyone's regular RAM.