Memory System Performance

Lab tomorrow: HW2 office hours!

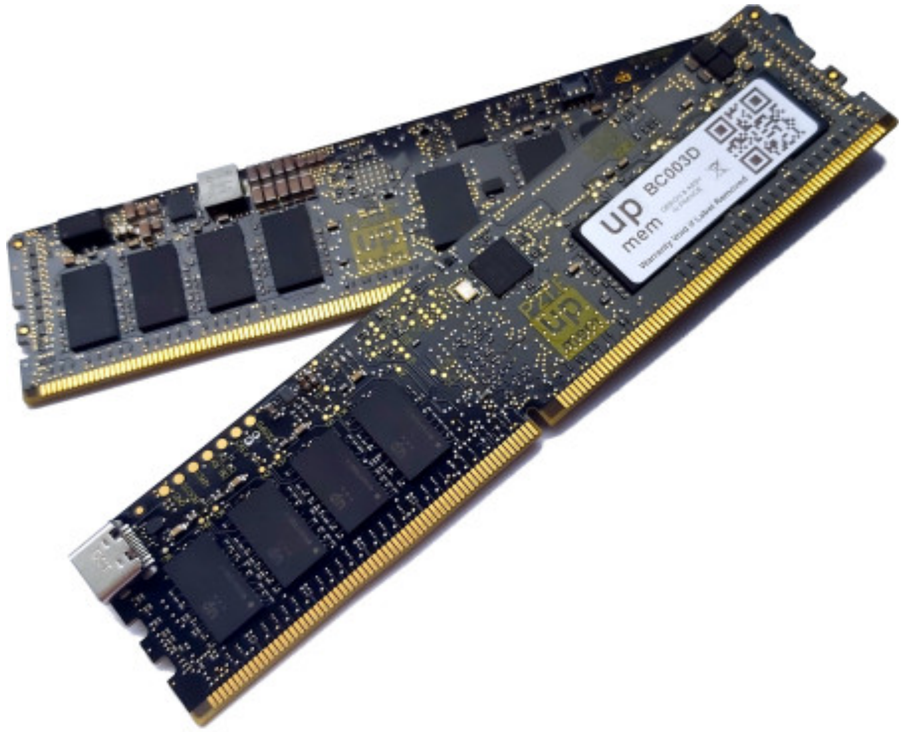


Image Credit: https://www.newswire.com/news/upmem-raises-7m-to-revolutionize-ai-and-analytics-processing-22126102

UPMEM Unleashed: Software Secrets for Speed

Krystian Chmielewski, Jarosław Ławnicki, Uladzislau Lukyanau, Tadeusz Kobus, Maciej Maciejewski
Heterogeneous Memory Software Lab
Huawei Technologies
Warsaw, Poland
name.surname@huawei.com

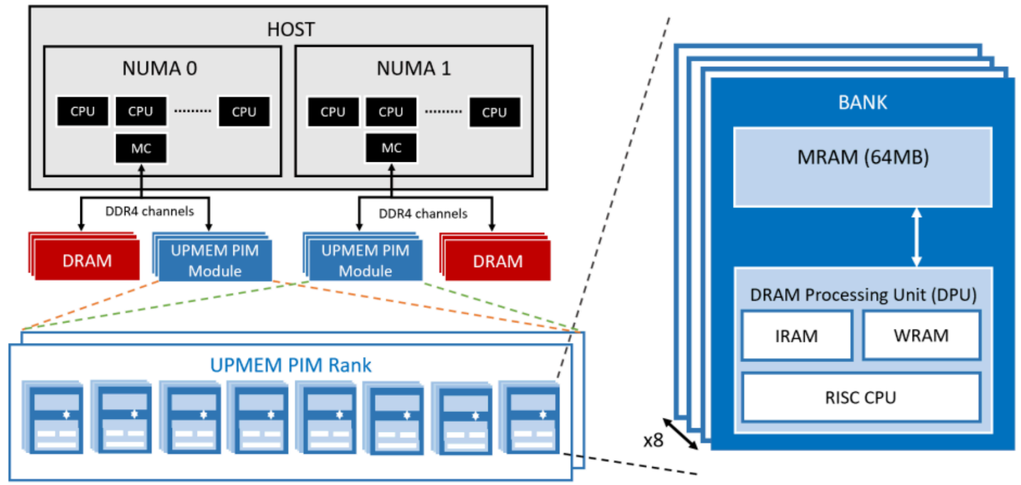


Fig. 1: Overview of the UPMEM platform architecture

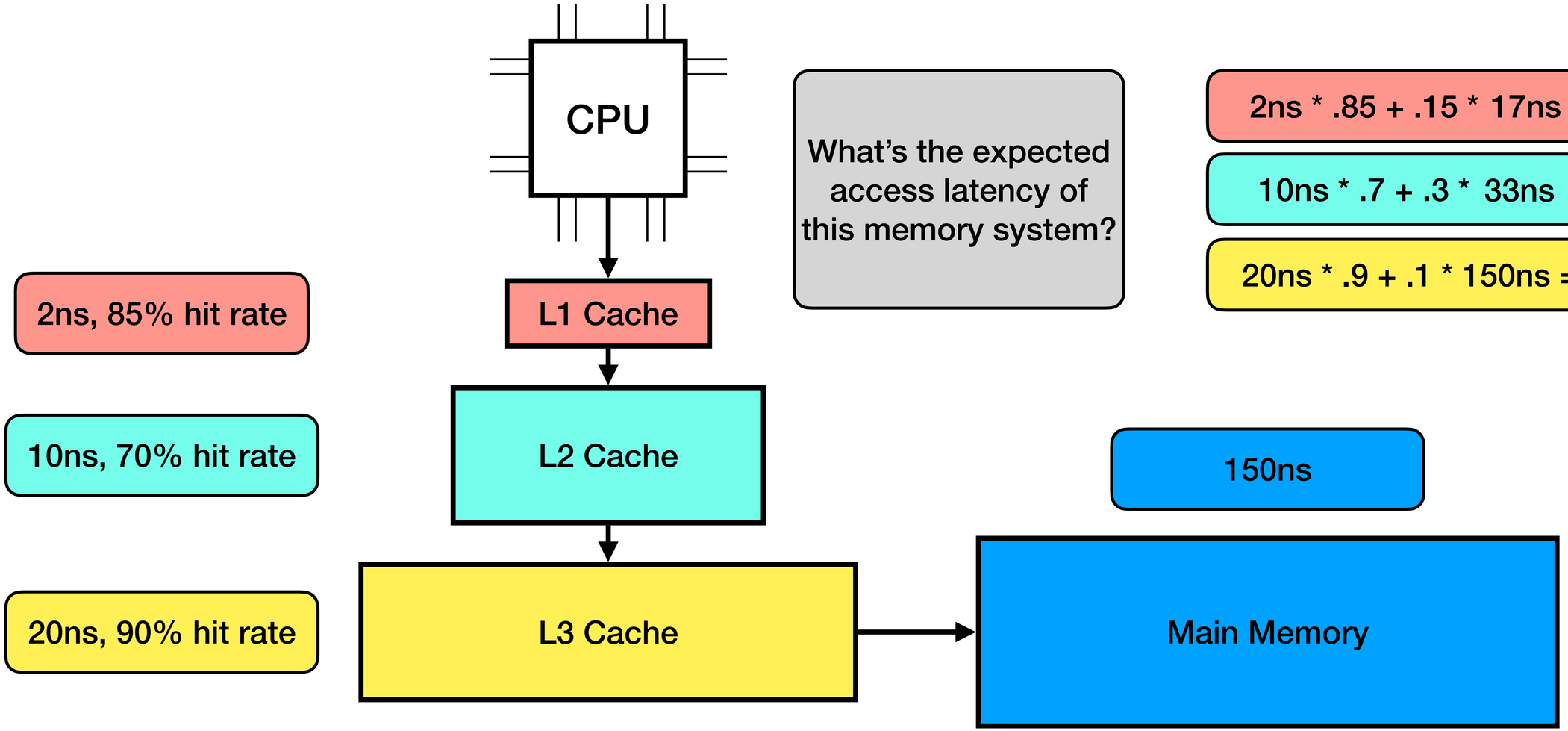
Outline

- Memory system performance metrics
- The memory wall and AI
- DRAM prices (???)

Cache Performance Metrics

- Cache hit rate: number of hits / number of accesses
- MPKI: cache misses per thousand (K) instructions
- Expected cache latency: (hit rate * hit latency) + (miss rate * miss latency)
- Miss latency is the expected cache latency of the next level below

Cache Access Latency Example



Chat with your neighbor(s)!

An average memory latency of 4ns is very tolerable for most systems. If you are a computer architect who designs memory systems, are you done? What else might need to be considered?

We also need to be able to implement devices with larger capacities

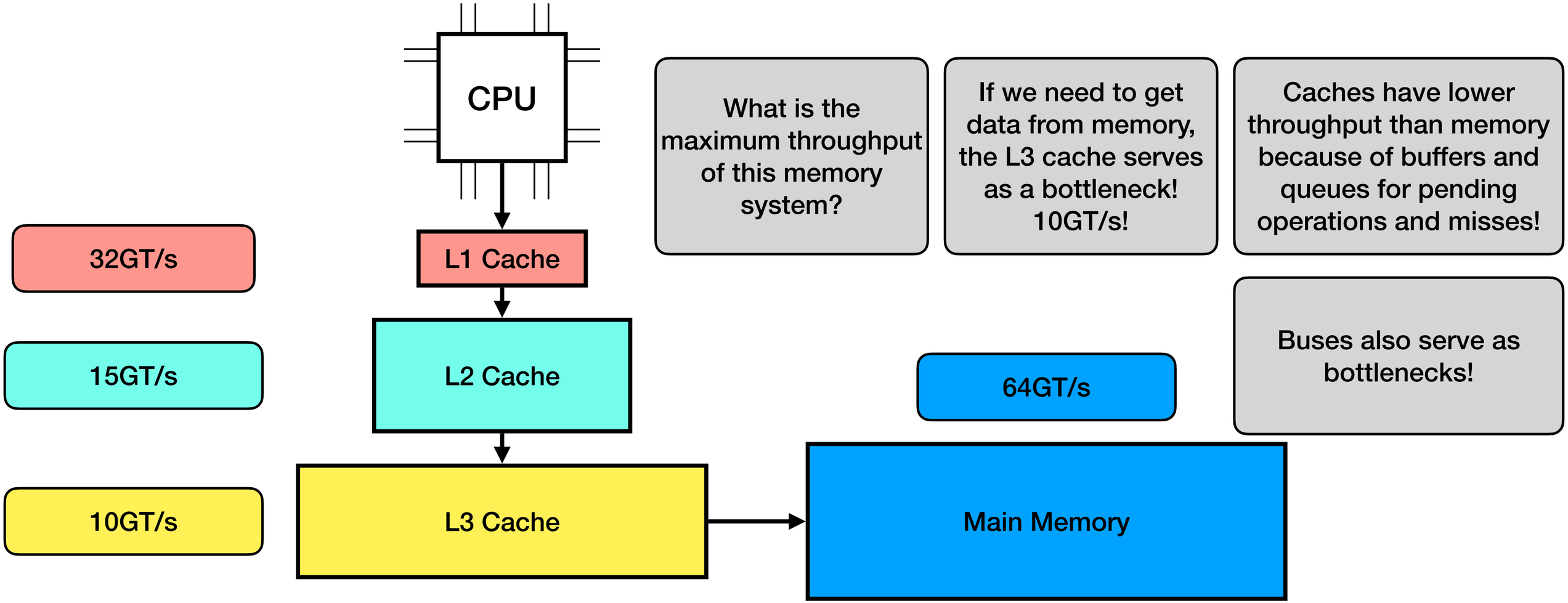
Beyond time to access (latency), we care about the work done per unit of time (throughput)!

… and the pesky security problem (we will talk about this on Friday)

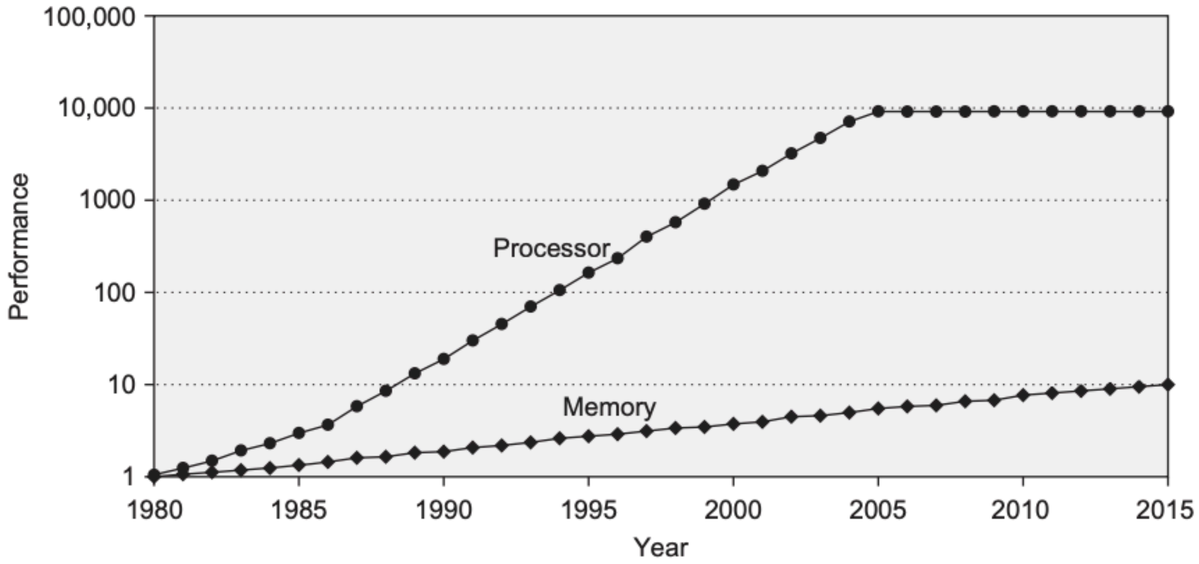
Memory System Throughput

- In order to maintain quality-of-service for an application, a memory system needs to have fast accesses and high throughput ➡️ otherwise, the application will not be able to receive from memory all of the data relevant to its execution
- In general, we can think of throughput in a memory device as the number of *transfers per unit of time* that it can achieve (e.g., 64GT/s)
- Sometimes, this is generalized by the term *bandwidth* which describes the number of *bytes transferred per unit of time* (e.g., 64GB/s)

Memory System Throughput



The Bandwidth Issue and the Memory Wall



Advances in processor designs are directly related to the increased demand for memory!

The Bandwidth Issue and the Memory Wall

AI and Memory Wall

Amir Gholami, *UC Berkeley, Berkeley, CA, 94720, USA*

Zhewei Yao, *Snowflake, Bellevue, WA, 98004, USA*

Sehoon Kim, Coleman Hooper, Michael W. Mahoney, and Kurt Keutzer, *UC Berkeley, Berkeley, CA, 94720, USA*

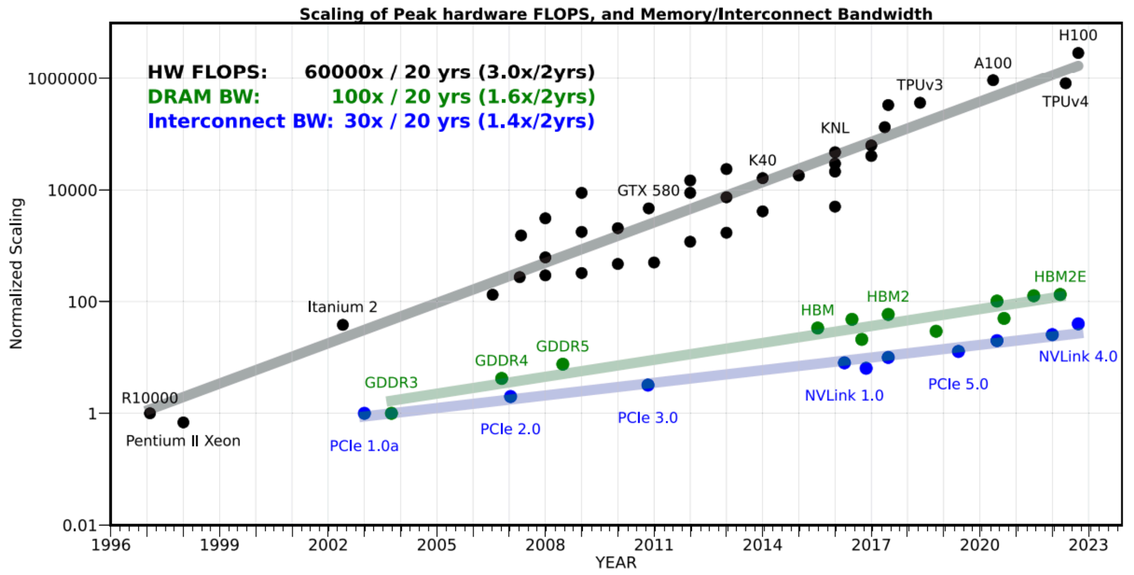
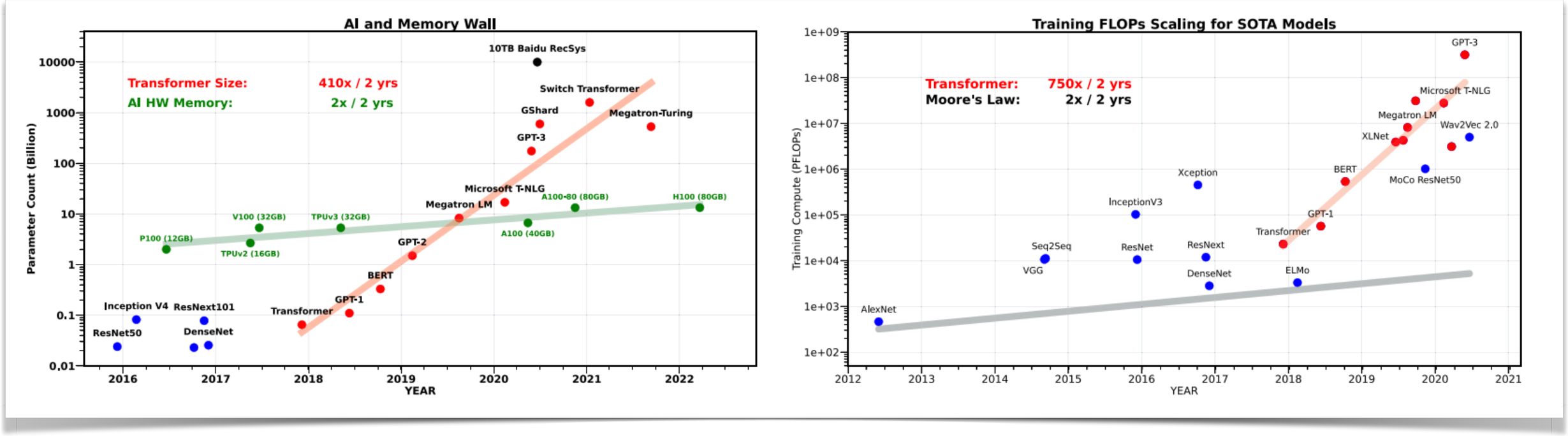


FIGURE 1. The scaling of the bandwidth of different generations of interconnections and memory as well as the peak floating-point operations per second (FLOPS). As can be seen, the bandwidth is increasing very slowly. We are normalizing hardware peak FLOPS with the R10000 system, as it was used to report the cost of training LeNet-5. DRAM: dynamic random-access memory; FLOPS: floating-point operations per second; HW: hardware; BW: bandwidth.

The Bandwidth Issue and the Memory Wall



Case Study: DRAM Prices

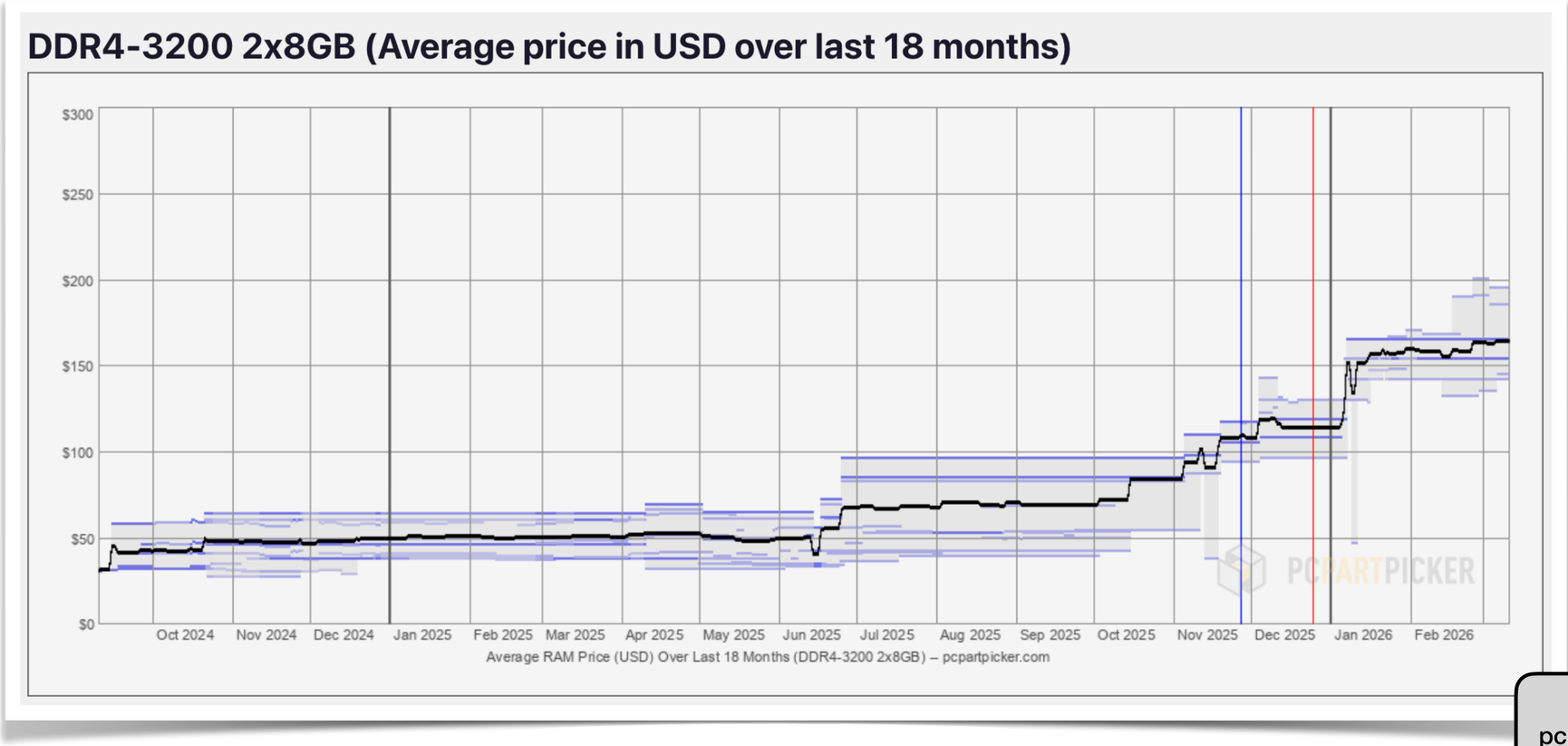


Image Credit: https://pcpartpicker.com/trends/price/memory/