

# Datacenter Memory Systems

CS59200, Fall 2026

Midhul Vuppalapati

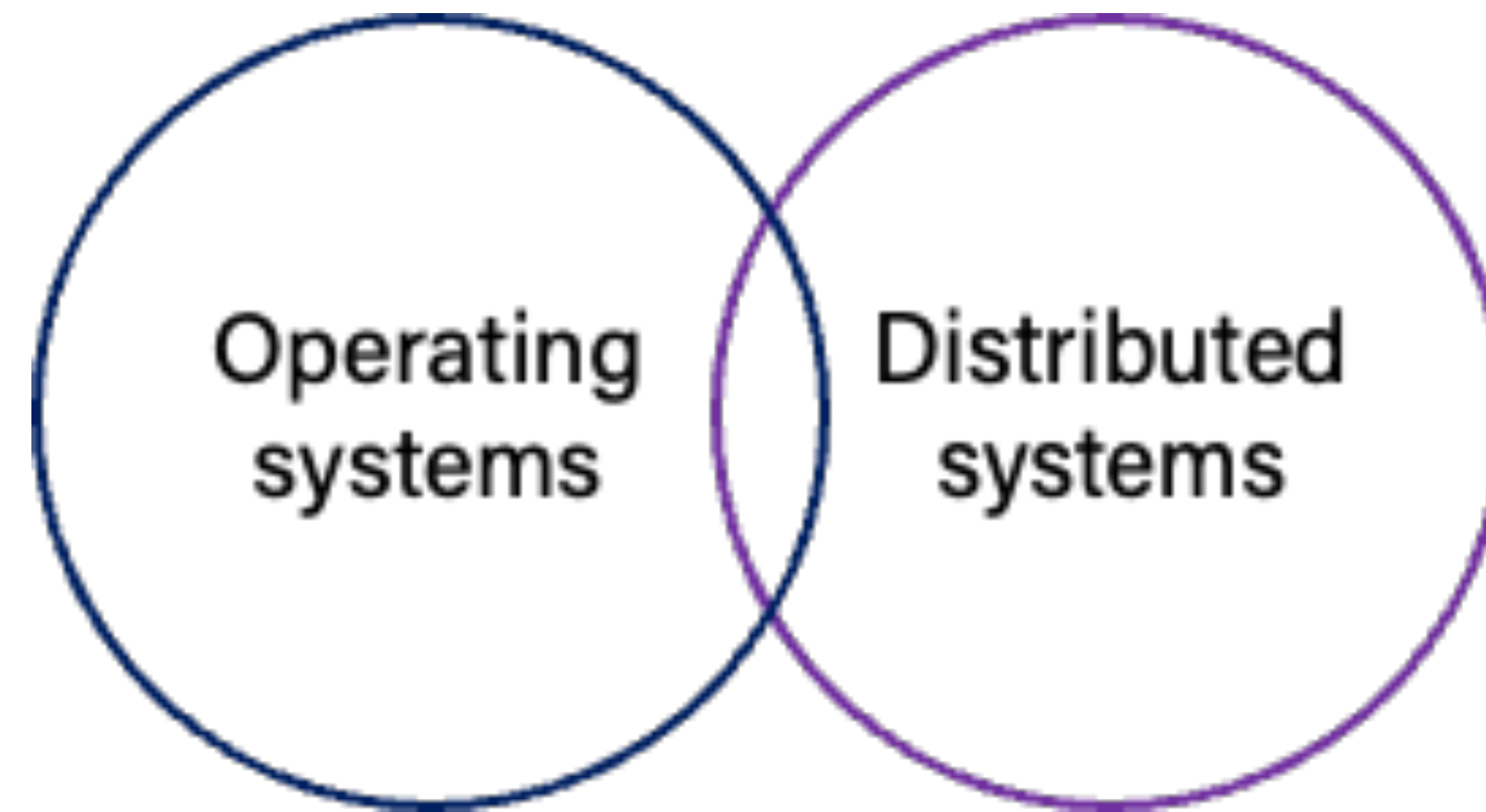
# Agenda for today

- Introductions
- Why care about datacenter memory?
- Course overview
- Course logistics
- Next class

# About me

- Midhul Vuppalapati
  - Pronunciation: me . DOOL (as in bool)
  - New Assistant Professor (officially started today)
  - PhD: Cornell University, 2026
  - Website: [midhul.com](http://midhul.com)
  - Email: [mvuppala@purdue.edu](mailto:mvuppala@purdue.edu)

# About me: Research



Key venues: SOSP, OSDI, SIGCOMM

Collabs: Security, Theory, Databases



# About me: Non-research

- Juggling (5 balls, 3 clubs, passing)
- e-Biking
- Favorite food: Biryani

# Student introductions

# A datacenter (outside)



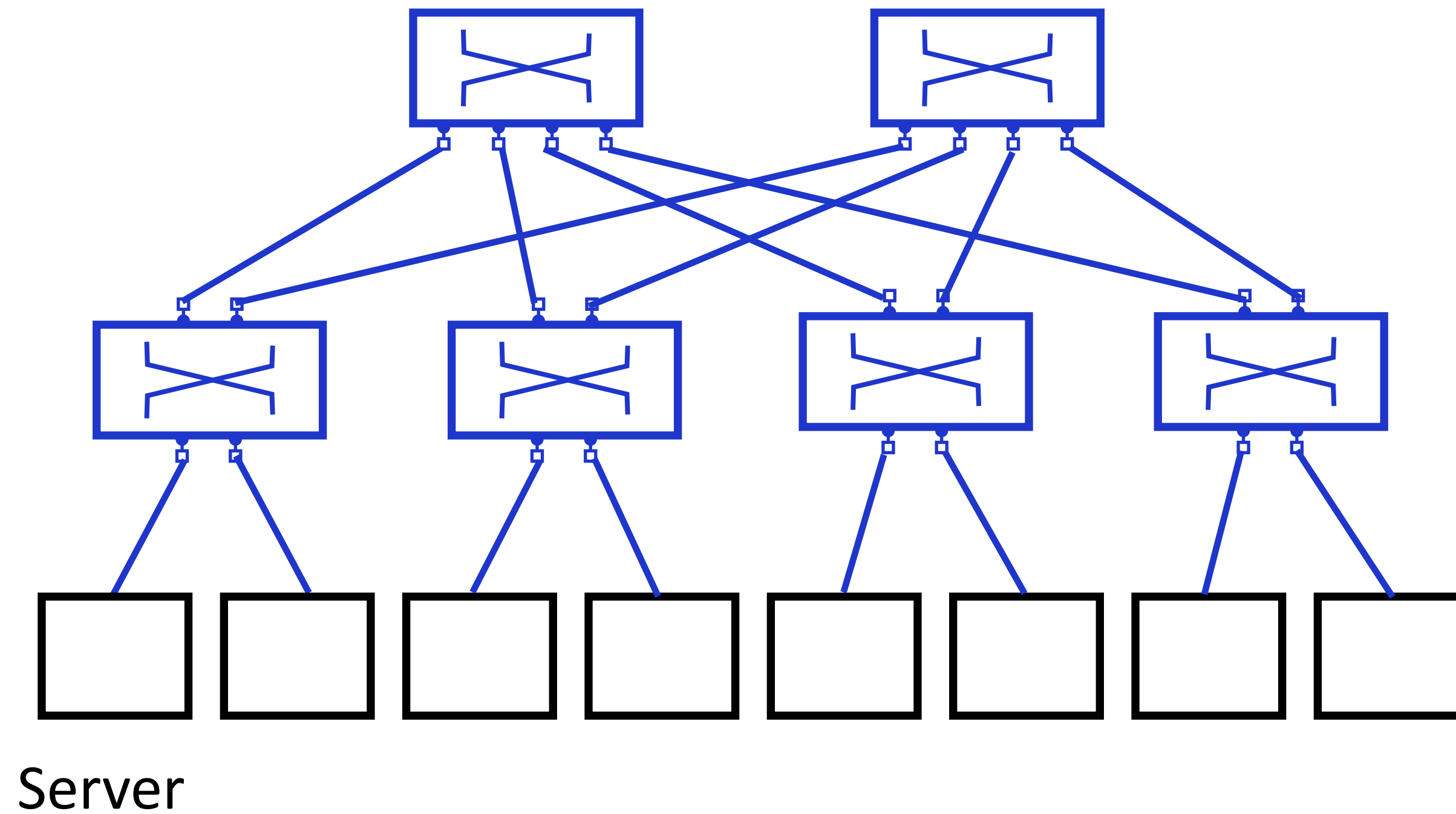
Source: <https://datacenters.google/discover-more/photo-gallery/>

# A datacenter (inside)

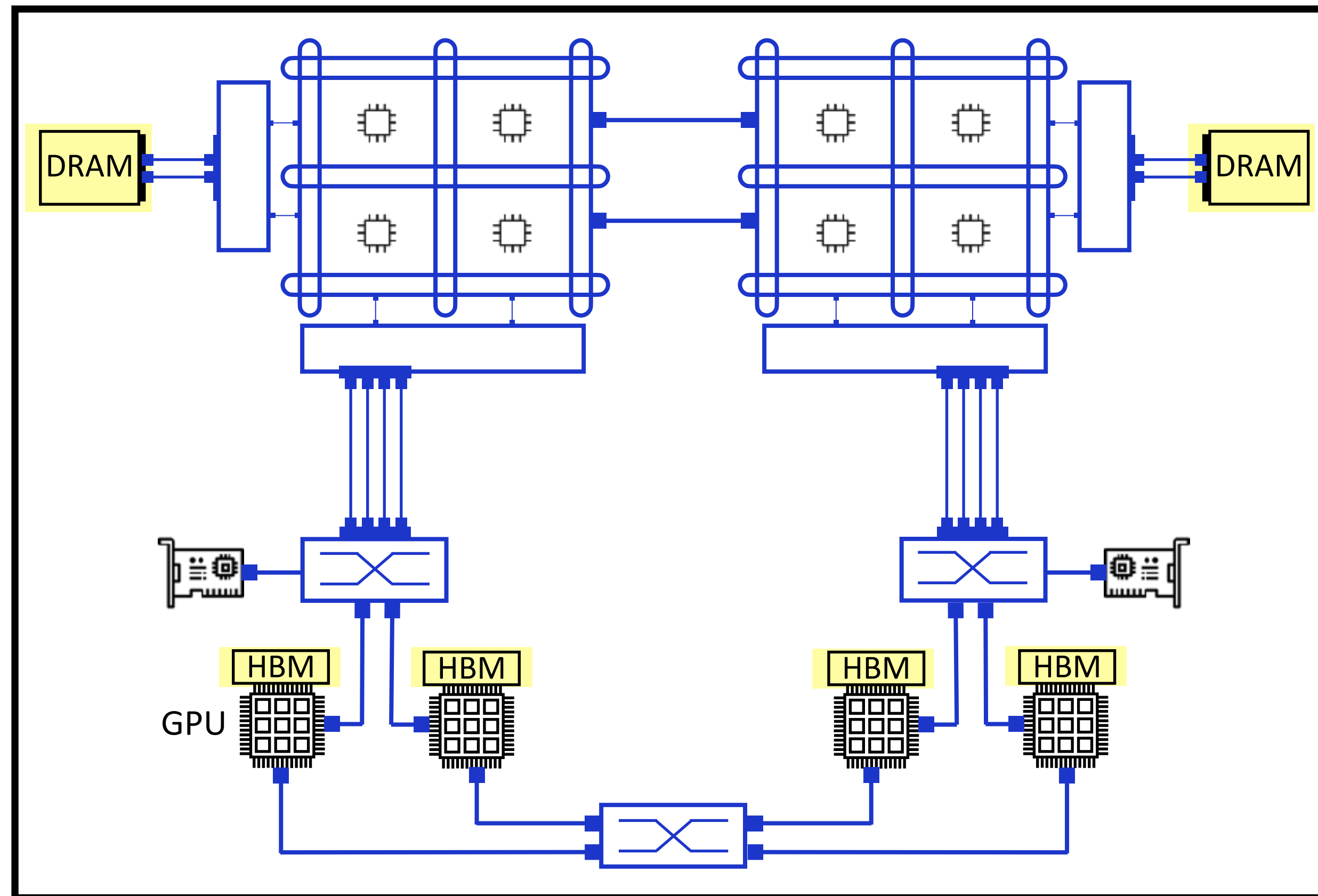


Source: <https://datacenters.google/discover-more/photo-gallery/>

# A datacenter (conceptually)



# Inside a datacenter server



# RAMageddon

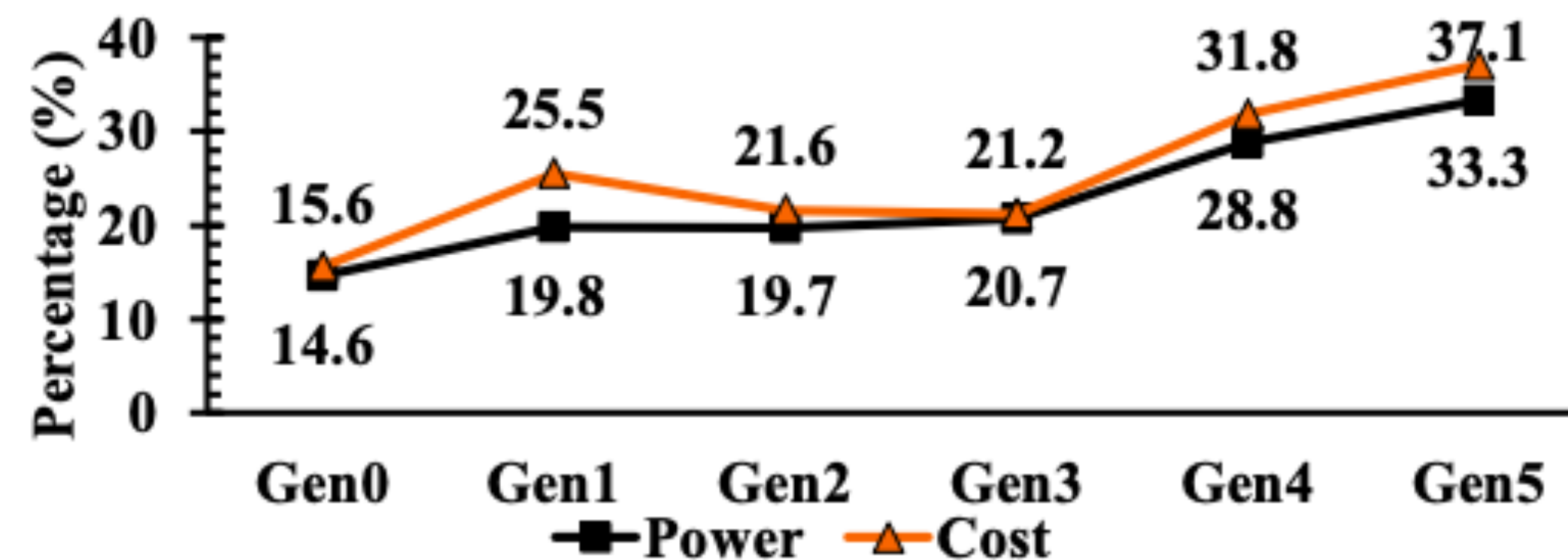
| Memory Kit Speed & Size | August 2025 Average Price | August 2026 Average Price | Change (YoY)     |
|-------------------------|---------------------------|---------------------------|------------------|
| 4800 MT/s 2×16GB        | \$90                      | \$425                     | +\$335<br>+372%  |
| 5200 MT/s 2×16GB        | \$100                     | \$480                     | +\$380<br>+380%  |
| 5600 MT/s 2×16GB        | \$116                     | \$528                     | +\$412<br>+355%  |
| 6000 MT/s 2×16GB        | \$108                     | \$572                     | +\$463<br>+429%  |
| 5600 MT/s 2×32GB        | \$191                     | \$1118                    | +\$927<br>+485%  |
| 6000 MT/s 2×32GB        | \$222                     | \$1272                    | +\$1050<br>+473% |



Source: <https://www.tomshardware.com/pc-components/ram/memory-prices-climb-500-percent-in-12-months-up-to-10x-the-lowest-ever-tracked-prices-128gb-of-ddr5-now-usd3-399>

# Datacenter memory capacity costs

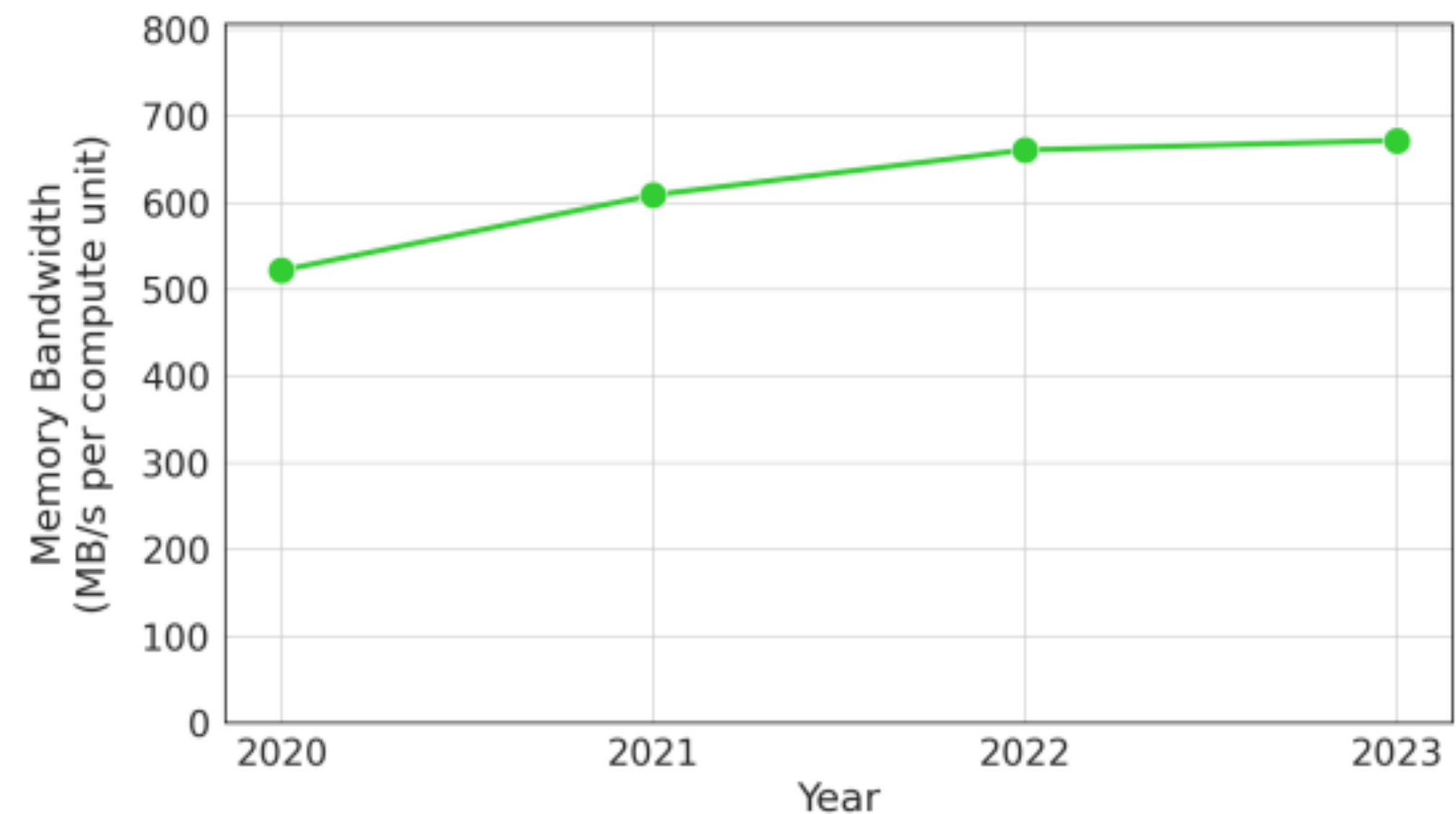
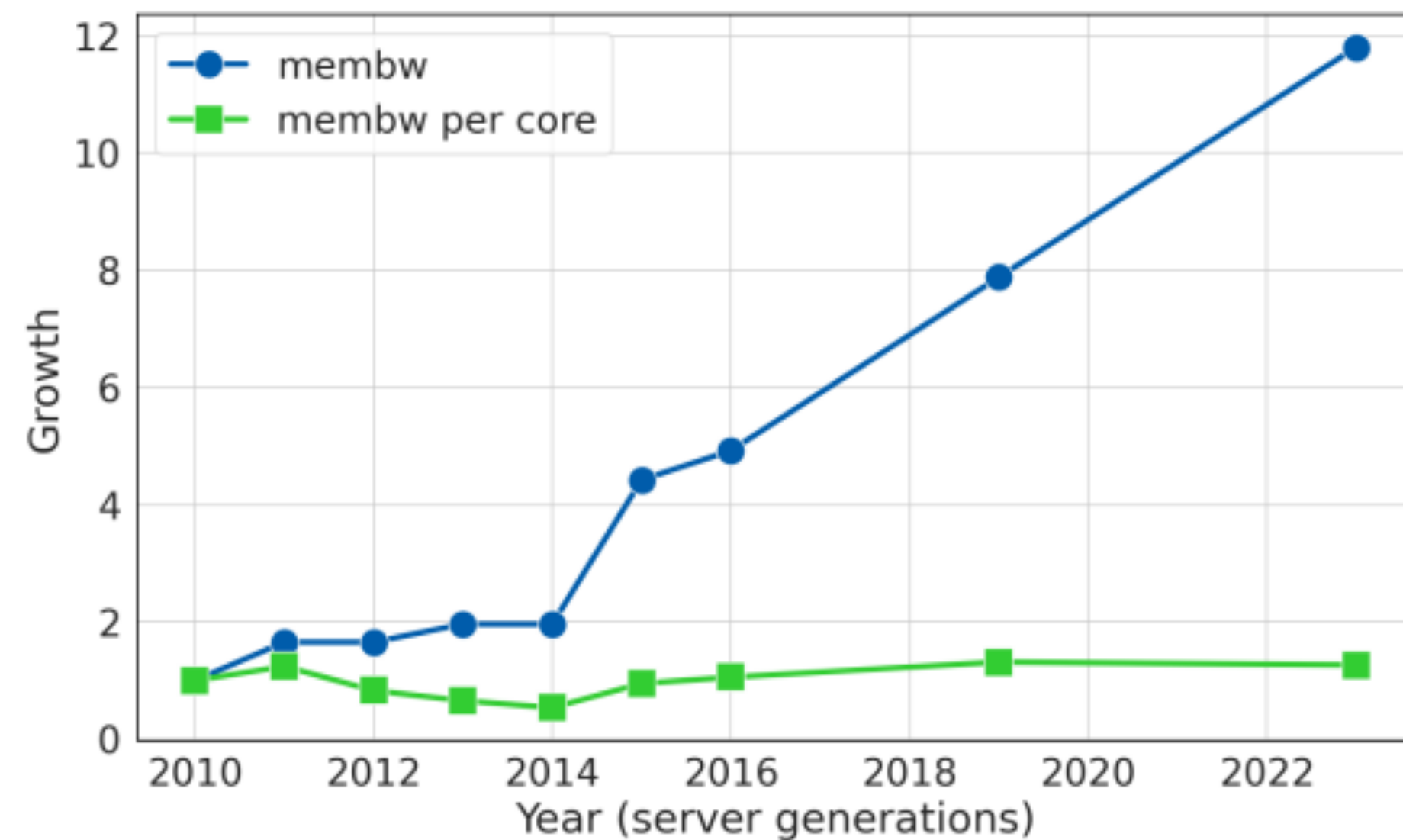
- Microsoft Azure: 50% of server cost [1]
- Meta: 37% of TCO [2]



[1] <https://www.nextplatform.com/connect/2020/04/03/cxl-and-gen-z-iron-out-a-coherent-interconnect-strategy/1644243>

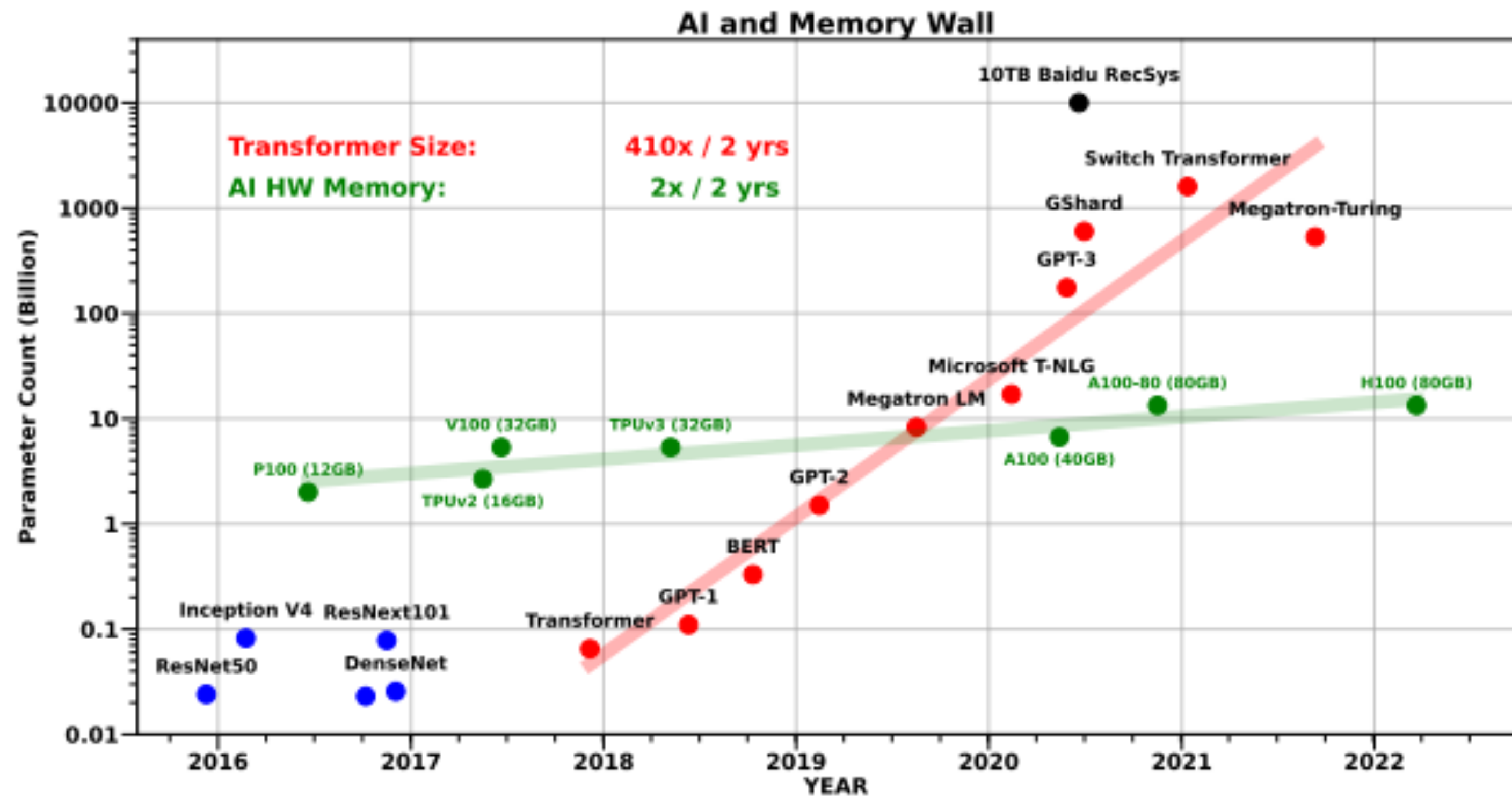
[2] TPP: Transparent Page Placement for CXL-Enabled Tiered-Memory, ASPLOS 2023

# Memory bandwidth bottlenecks



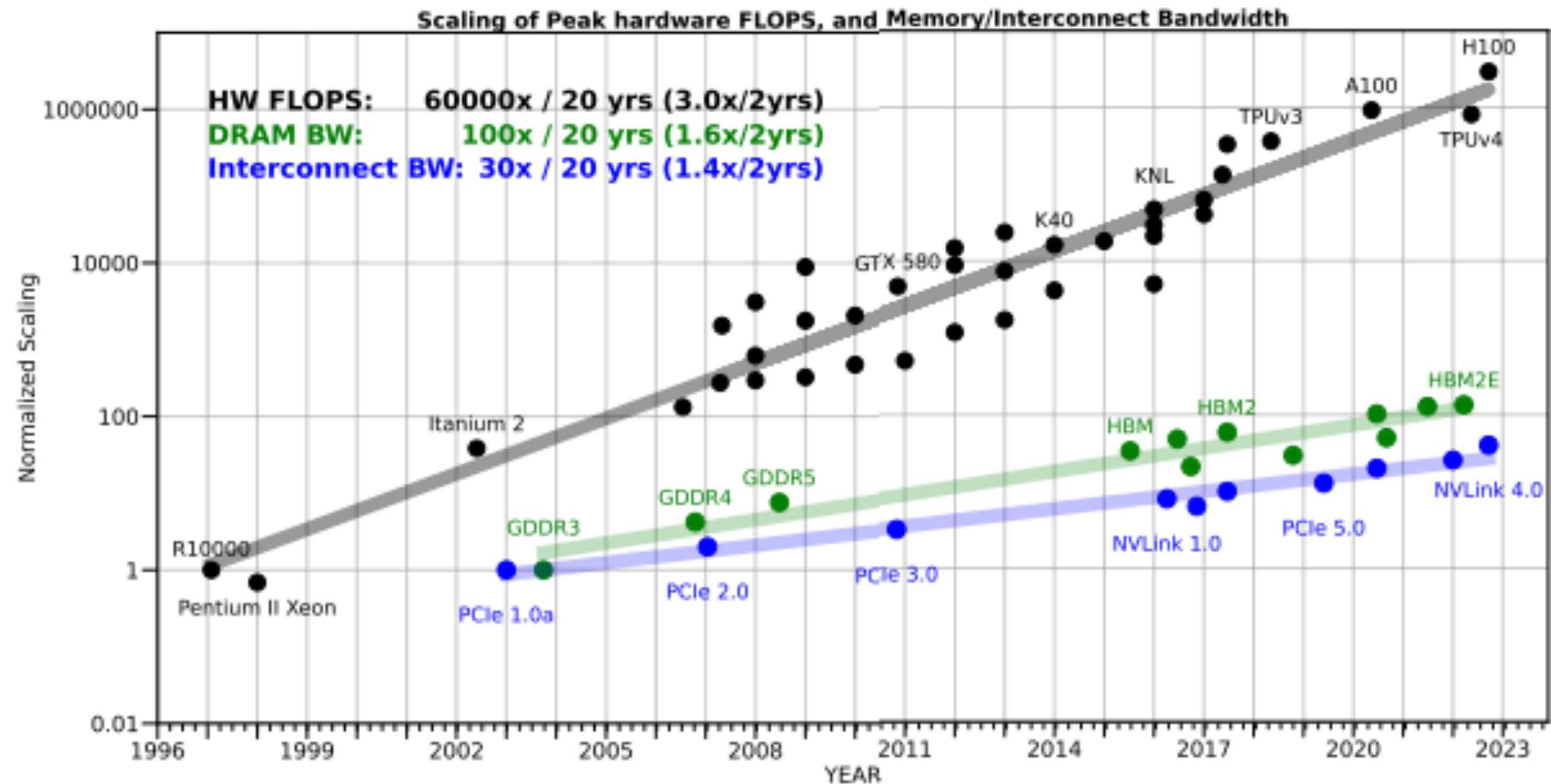
Source: Limoncello, APLOS 2024

# GPU memory capacity



Source: AI and memory wall

# GPU memory bandwidth



Source: AI and memory wall

# Takeaways

- Main memory: critical bottleneck in modern datacenters
- Unsustainable memory costs
- Severe performance impact for modern applications
- Memory bottlenecks even more acute for GPUs / AI applications

**Questions?**

# Course goal

- Explore approaches to mitigate and manage datacenter memory bottlenecks across the software and hardware stack
- Operating systems, distributed systems, computer architecture
- Research papers from top-tier conferences

# Course topics

- (Tentative, flexible, subject to change)
- OS memory management
- Memory offloading, tiering and disaggregation
- Address translation optimizations
- Memory bandwidth management
- LLM inference memory management
- New memory technologies

# Paper suggestions?

- Send me an email

# Course logistics

<https://midhul.com/courses/cs59200dms-fall26.html>

Announcements via Brightspace (also delivered via email)

# Requirements

- Read the required readings before every class (1-2 papers)
- Attend classes and participate in discussions
- One in-class paper presentation (grade: 15%)
- Two assignments (grade: 25%)
- Midterm exam (grade: 25%)
- Final project and paper (grade: 35%)

# Paper reading

- **Critically** read the required papers before each class
- Things to make note of
  - What is the **problem** being solved?
  - Why is it an **interesting** problem?
  - What are the key technical **insights**?
  - Are the claims backed by **evidence**?
  - What are the **limitations**?
  - What are potential **future directions**?

# Paper presentation

- Students will give ~45 mins presentation on 1-2 papers (single class)
- Presentation should provide necessary background
- Presentation should summarize relevant related work
  - Look at suggested readings
- **Schedule meeting with Midhul one week before presentation date**
  - Have a draft of slides ready

# Assignments

- Assignment 1: Linux kernel memory management
- Assignment 2: One of two options
  - LLM inference memory management
  - DRAM simulator

# Midterm exam

- Will be conducted during normal class session (October 20)
- Tentative plan: Oral exam
- Details will be announced shortly

# Final project

- Pick one of two options
- Option A: Research project
- Option B: Survey on a topic related to the course
- Checkpoints (dates will be announced soon)
  - Initial proposal
  - Mid-term report
  - Final report
- Final presentation

# Next class

- Deeper dive into memory technology trends
- Readings
  - Hitting the Memory Wall: Implications of the Obvious
  - AI and the Memory Wall

**The end**