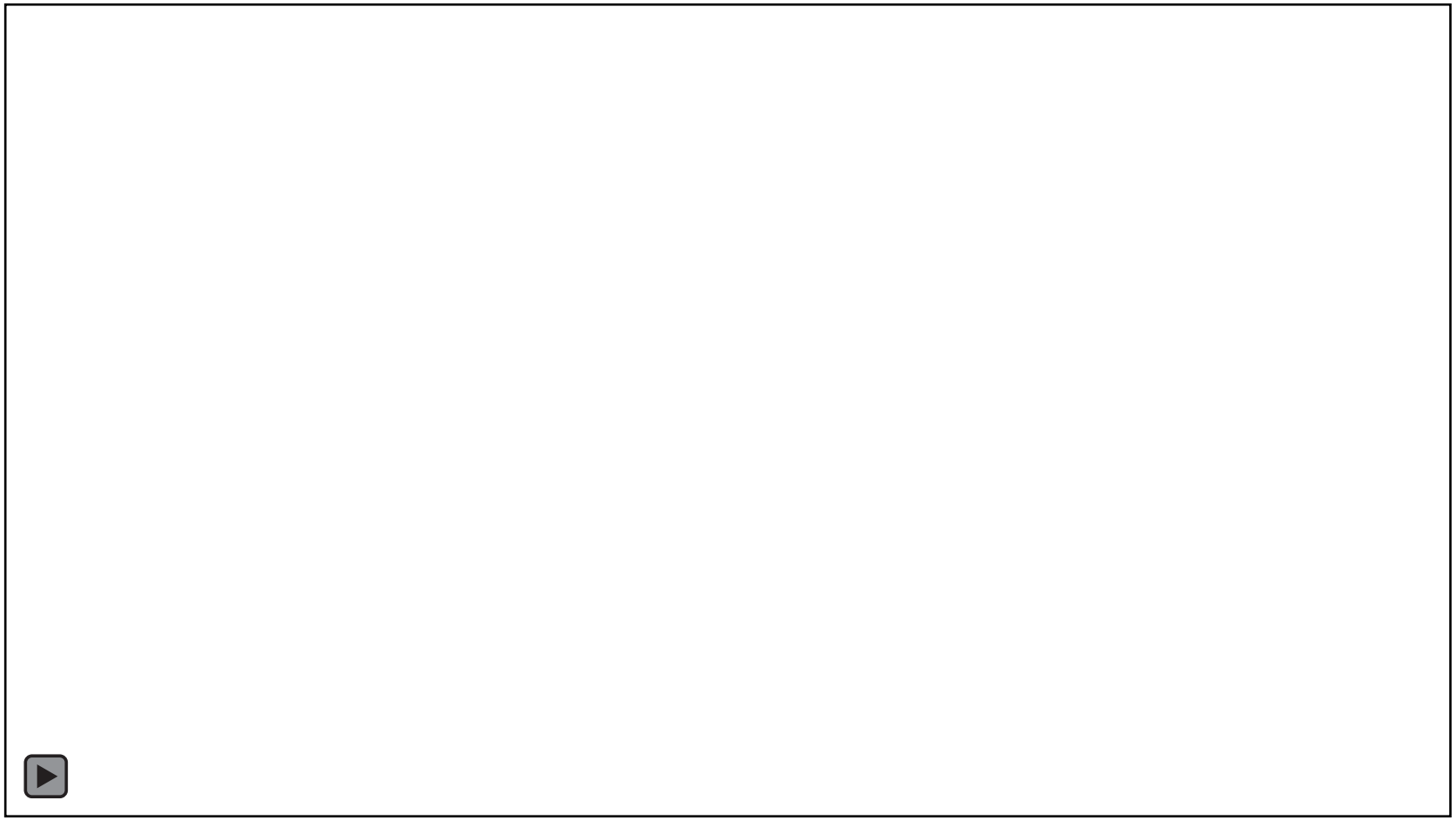
DRAM-302-1: HBM: DRAM's Bright Future

Marc Greenberg, Principal/CEO, Marc Greenberg Consulting LLC

marc@marcgreenberg.com

Summary of the conference so far



Initial prompt to ChatGPT:
Write me a prompt to generate a video showing a multitude of simple computer screens, each computer screen is drawing a line chart showing an exponential relationship between X and Y, and the screens themselves are flying around a simple empty white space.

ChatGPT generated prompt for starting image generation:
A dynamic video animation showing dozens of simple computer screens flying through an empty, white, minimalist space. Each computer screen displays a different line chart illustrating an exponential relationship between X and Y — smooth, upward-curving lines with varying growth rates. The screens are flat, rectangular, and simplistic, with minimal bezels, shown from different angles as they float, rotate, and drift in three-dimensional space. The motion of the screens is graceful and slightly slow, emphasizing the clean visual style. The charts are drawn with thin lines, primarily in blue or black, on white screen backgrounds. The overall atmosphere is calm, mathematical, and futuristic.

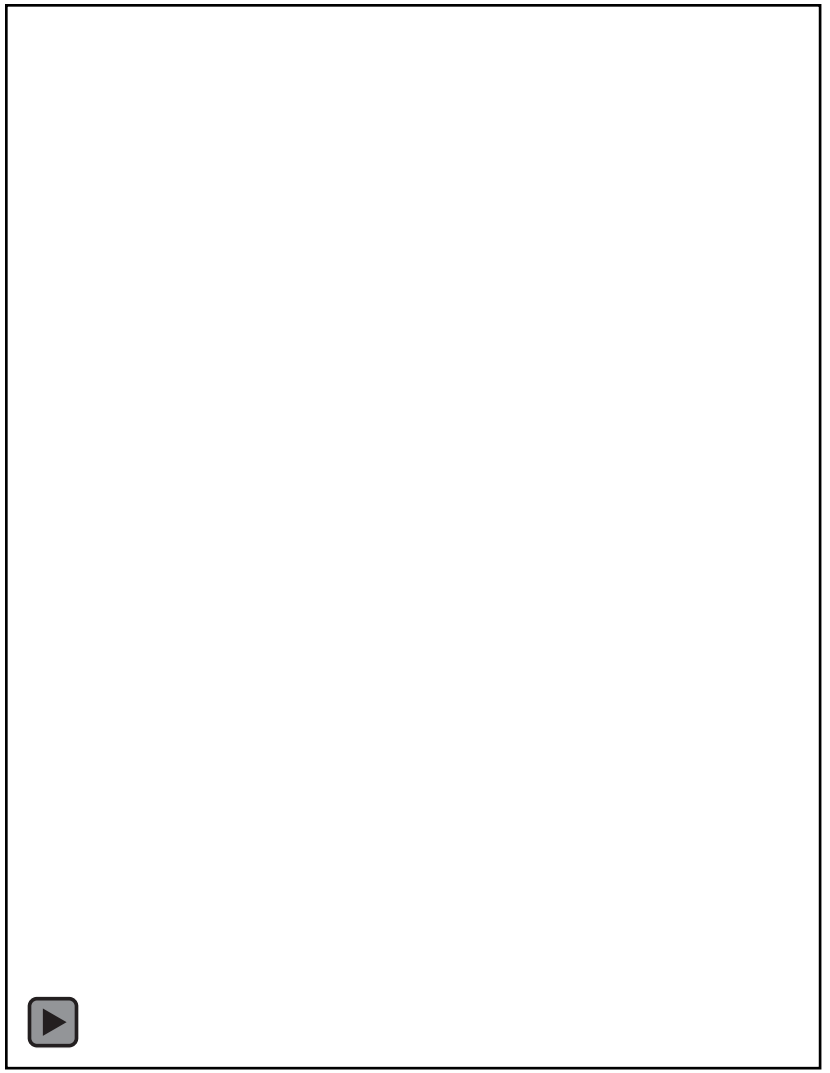
Prompt for video animation of starting image in Runwayml.com :
A dynamic video animation showing dozens of simple computer screens flying in different directions through an empty, white, minimalist space. The motion of the screens is graceful and slightly slow, emphasizing the clean visual style.

Opening statements

- AI will continue to grow in its capability for the foreseeable future

- Users will continue to demand computationally intensive AI models
  - Therefore the memory bandwidth (Gbit/s) must continue to increase
- The number of weights in tier-1 models will continue to increase
  - Therefore the the memory capacity (GBytes) must continue to increase

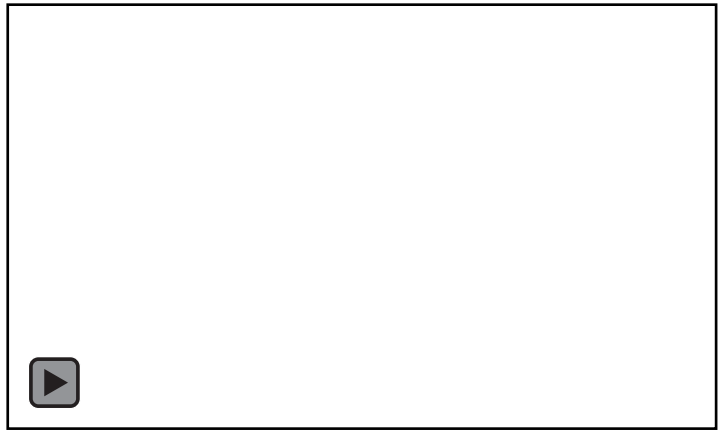
- AI TCO (Total Cost of Ownership), including cost of power, will become increasingly important as the industry matures



an image of a crystal ball in a simple room, the crystal ball contains a computer brain; turn this image into a 10 second video where the contents of the crystal ball transform into a human brain and then return to the starting image so that the video may be looped. The camera should stay static the whole time.
Runwayml.com

Know the “AI as a service” buyer

1. Cost per token
2. Initial latency to first token
3. Number of tokens per second

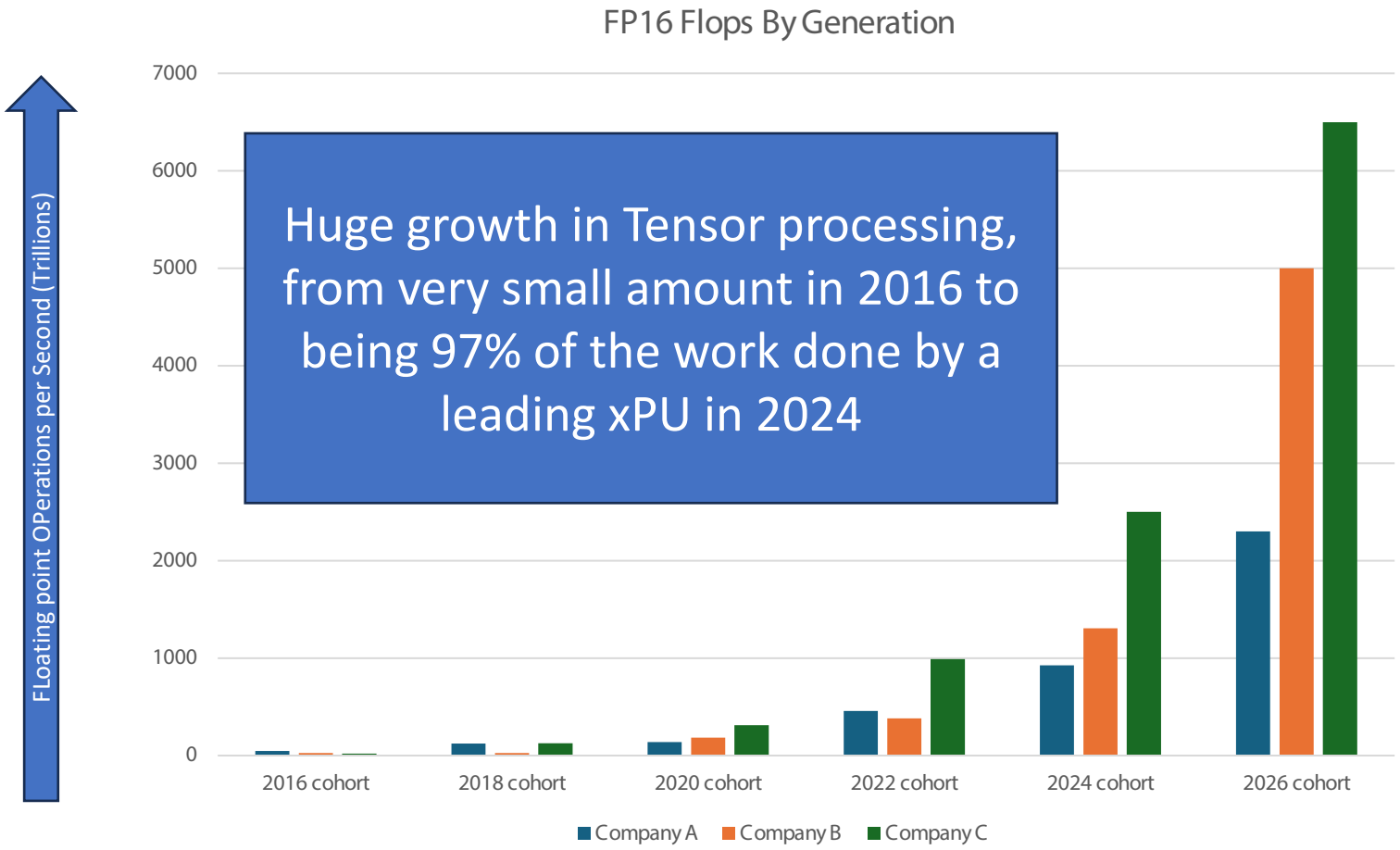


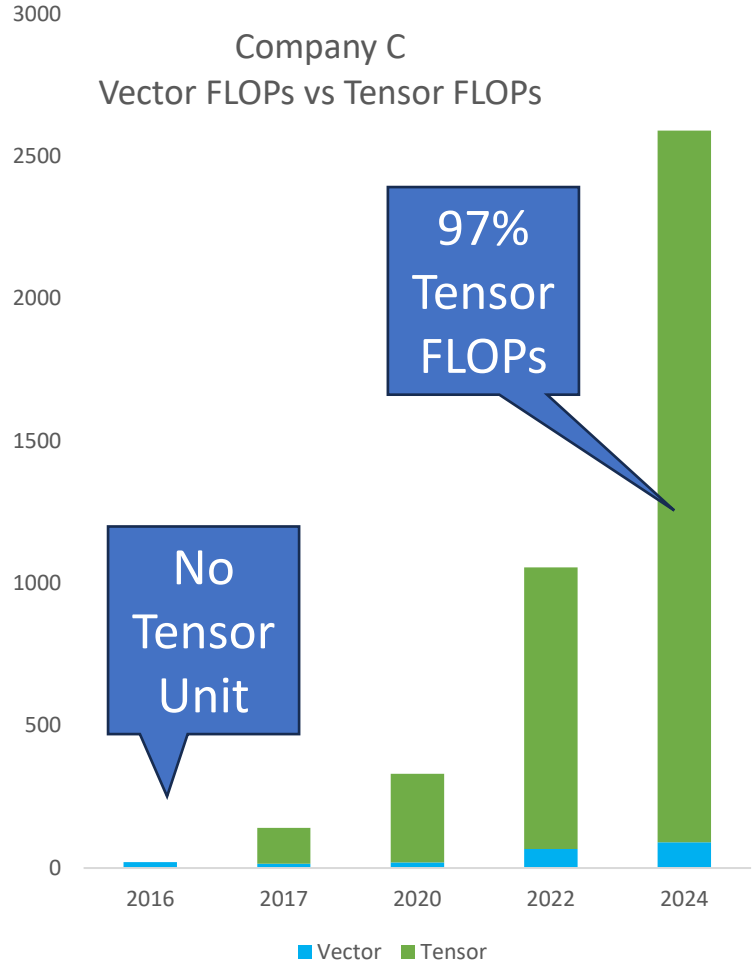
AI generated with Runwayml.com and Llama - create a text-to-video prompt to describe a close-up image of an IT manager struggling with a difficult purchasing decision

The things they are **NOT** basing their decisions on:
- What type of memory is the server using?
- What xPU architecture is being used?
- What company made the xPU?
- How much did the server cost?
- How much energy is the server using?

All of the things we care about in semiconductor!

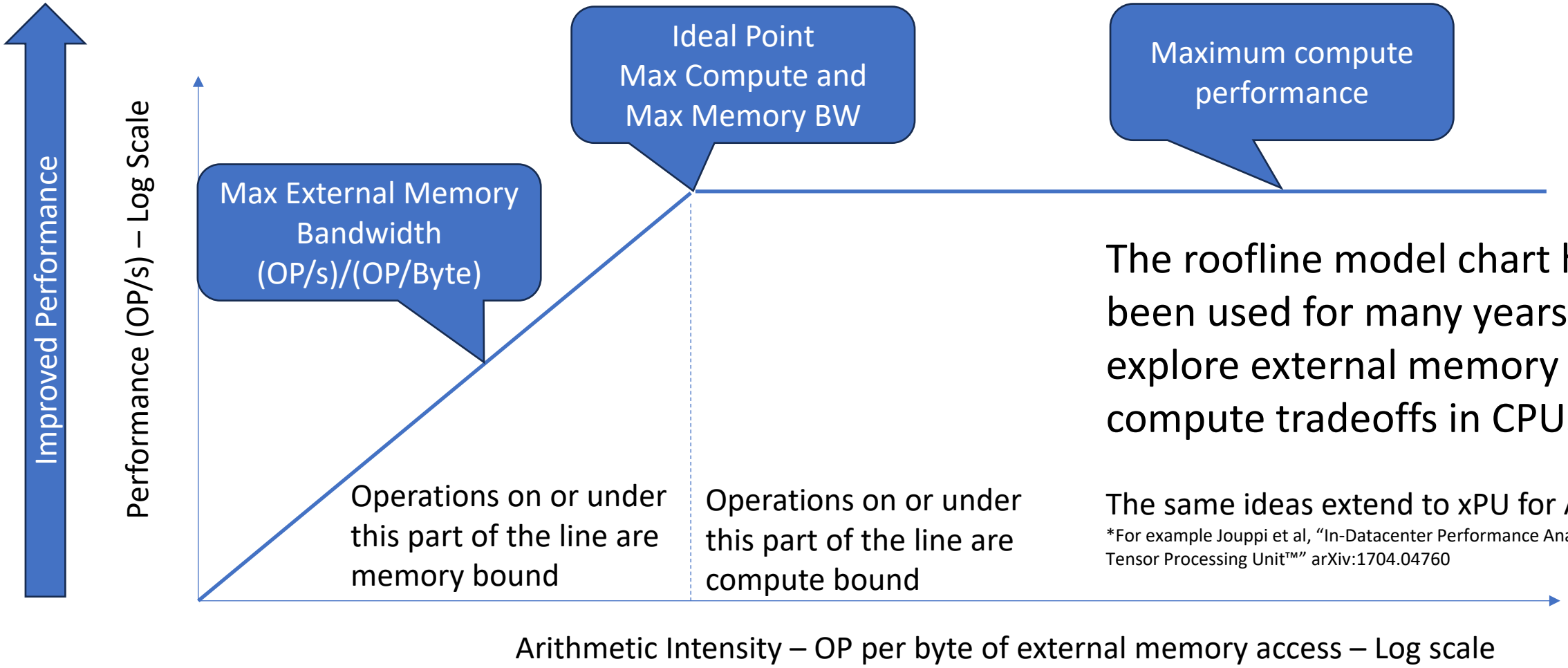
Datacenter AI Processor OP growth: ~2.7x / 2 years





Non-sparse. Per major processing element (architectures differ)
Sources: TheNextPlatform.com and their own blog (Company A),
Product datasheets and conference presentations (for 2026) (Companies B and C)

Roofline Model – How it works

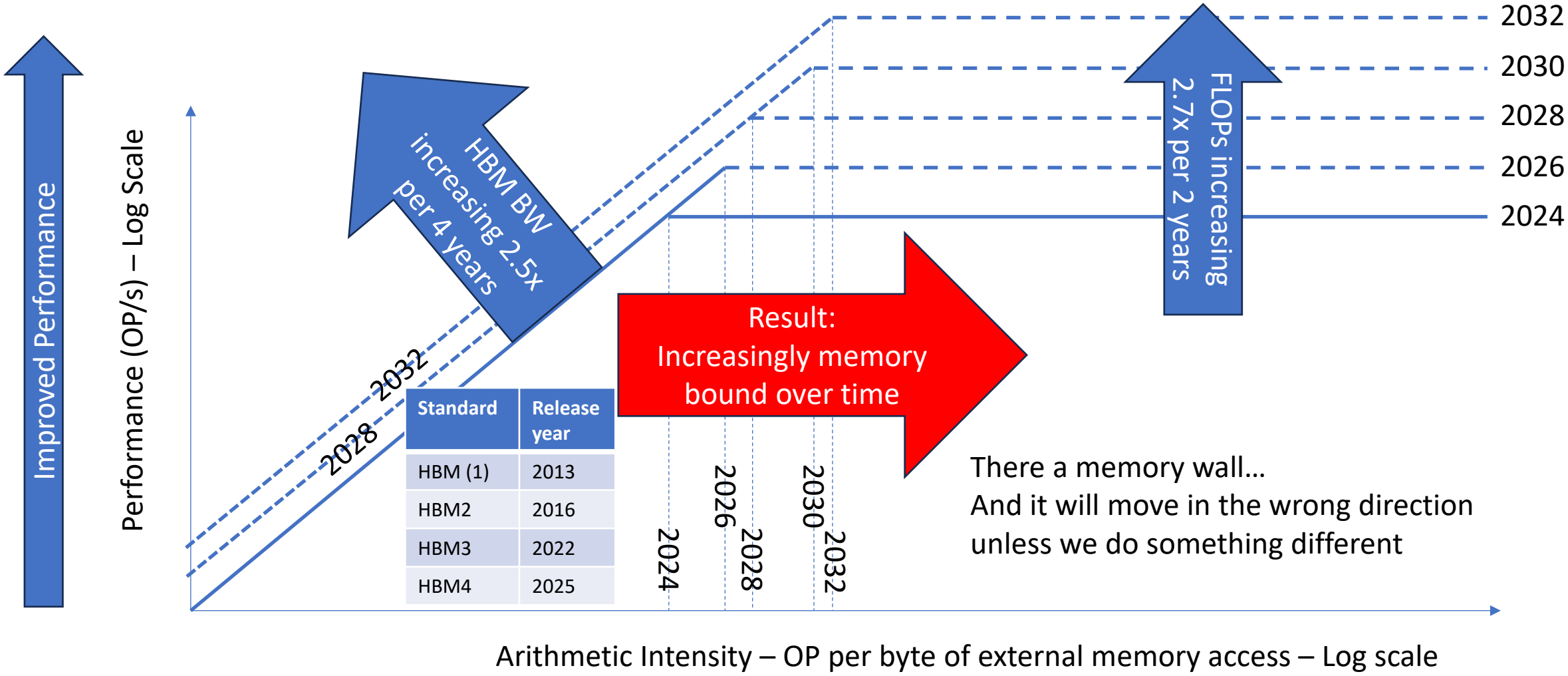


The roofline model chart has been used for many years to explore external memory vs compute tradeoffs in CPUs.

The same ideas extend to xPU for AI*

*For example Jouppi et al, "In-Datacenter Performance Analysis of a Tensor Processing Unit™" arXiv:1704.04760

Datacenter Roofline Model at the current rate



Standard	Release year
HBM (1)	2013
HBM2	2016
HBM3	2022
HBM4	2025

How did we get to here?

- Efficient MatMul units in the form of systolic arrays have been around since the 1940's (the term was invented in 1979), for CNNs & RNNs
- 2017's Transformer paper, "Attention is all you need", heavily reliant on matrix mathematics
- Early GPUs weren't designed for AI, they just happened to work because of their matrix math capabilities
- HBM wasn't designed for AI, it just happened to work because of its bandwidth and multichannel organization

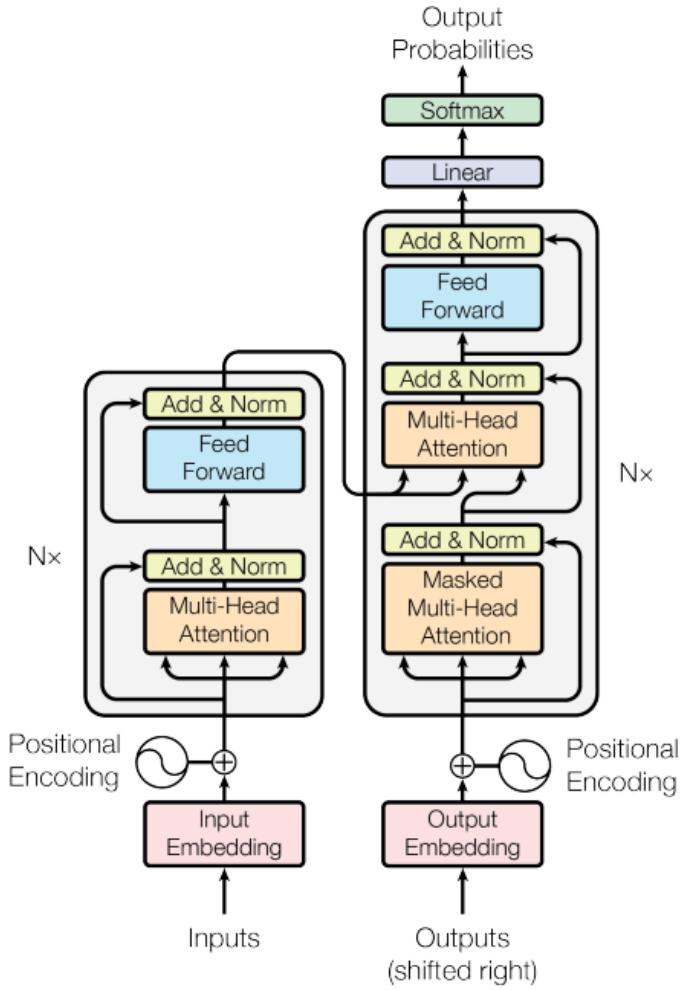


Figure 1: The Transformer - model architecture.

From "Attention is all you need" (2017) arXiv:1706.03762

HBM is awesome…

Strengths: Speed & TCO

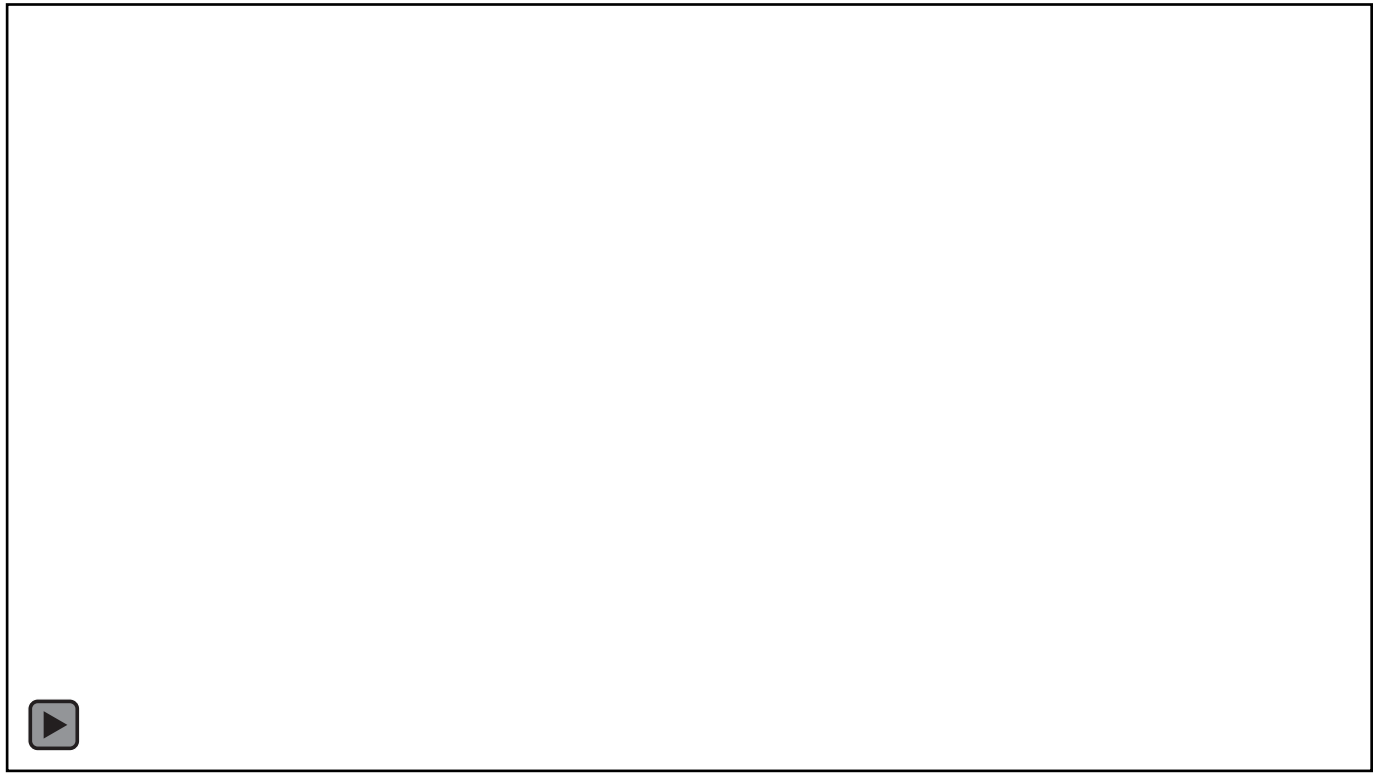
- Highest bandwidth per package/stack
- Highest bandwidth per mm of shoreline
- Lowest energy per bit
- Highest capacity per unit volume

Weaknesses: Initial Cost

- Cost per bit
- Complex manufacturing
- Almost impossible to repair or replace
  - other than on-die repair
- Requires advanced package technology
- Difficult bandwidth/capacity expansion

Capability vs Cost

- Future directions of memory for AI will be required to reduce energy per bit and increase capacity, while maintaining cost and quality of results from the AI models
- Improving the memory, and improving the xPU hardware to reduce demand on memory, are both options



Starting prompt: “A cartoonish villain named Professor Power controls electricity in all its forms - lightning bolts, sparks, electrical discharges, plasma and coronas” to get the starting image, followed by an iterative session to generate the video in Runwayml.com

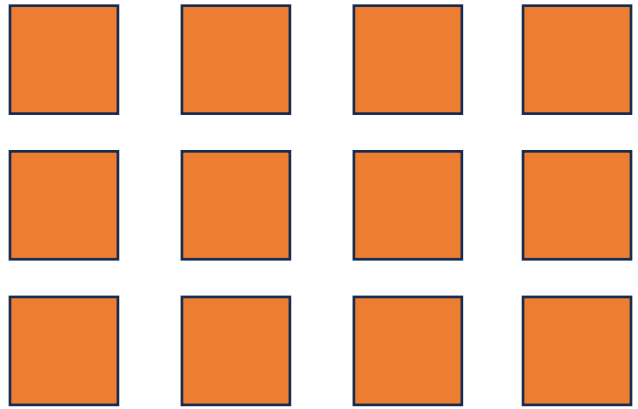
Future approaches on improved memory

- FMS2024: at every HBM session at least one person asked, "what about UCIe?" (or BoW or other D2D standards)
  - At least one vendor has announced UCIe with 20Tb/s/mm of shoreline on advanced package
    - although not specifically for memory, and there would be power and power density concerns at those speeds
    - HBM4 is about 1.5Tb/s/mm, not including commands
  - At least one vendor has proposed proprietary memory interfaces based on D2D
    - Possible application of "custom HBM" or a future standard
- Direct 3D Stacking
  - If the heat problem can be solved – DRAM degrades around 85c
- LPDDR5, LPDDR6, or GDDR7
  - Well proven on consumer devices; LPDDR5 is on the NVIDIA "Grace" module
  - But can't provide the same bandwidth as HBM
- On-die SRAM based solutions
  - Some companies using on-die SRAM successfully for datacenter xPU
    - SRAM doesn't favor large models or process shrinks

Exploiting model redundancy to help memory

Quantization

- Arbitrarily discard model data by using smaller data types (e.g., FP16→FP8)
- Approaching the fundamental limit of quantization below 4 bit
- Can’t train models with extensive quantization in xPU



Sparsity

- Selectively discard data from nodes that contribute less to the solution
- Structured sparsity is available on several xPUs, but usage is spotty
- Unstructured sparsity offers some promise, but may be difficult to implement
- Approximate computing shows promise when combined with sparsity

xPU rearchitecture to improve memory

- Processing or Compute in Memory (PIM or CIM)
  - Use memory bitcells to do computation directly
    - Antilog (log(X)+log(Y)) = X*Y
    - Difficult to get sufficient resolution and determinism
  - Implement computing structures (e.g., MatMul units) on memory die
    - Memory processes often optimized for capacitors, not transistors
- Processing near memory (PNM)
  - Use an intermediate die to compute values that are tightly coupled to RAM
    - For example, an HBM base die, if thermal issues can be managed
- Reducing caching of intermediate values in DRAM
  - Various techniques to increase arithmetic intensity
- Requires methods of mapping models on to unique hardware

Conclusion

- HBM is the best datacenter AI memory standard at present:
  - Highest Bandwidth
  - Lowest Energy per Bit (other than on-die SRAM)
- HBM suffers from cost and complexity issues
  - Expensive manufacturing for the HBM stack and advanced package substrate for XPU
  - Difficult expansion beyond max stack height
- Future iterations of datacenter AI memory are likely to further increase bandwidth while addressing cost and complexity issues