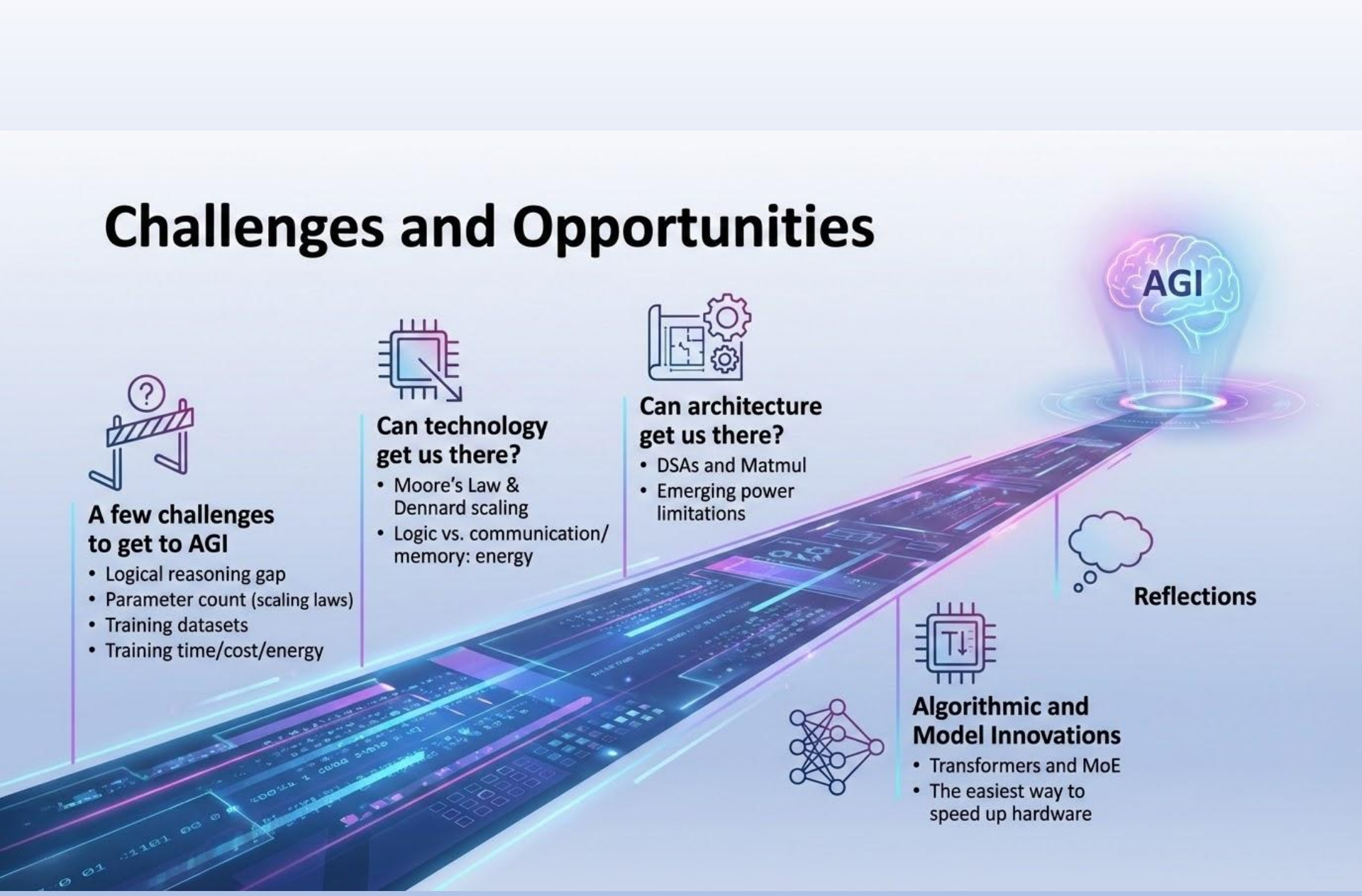
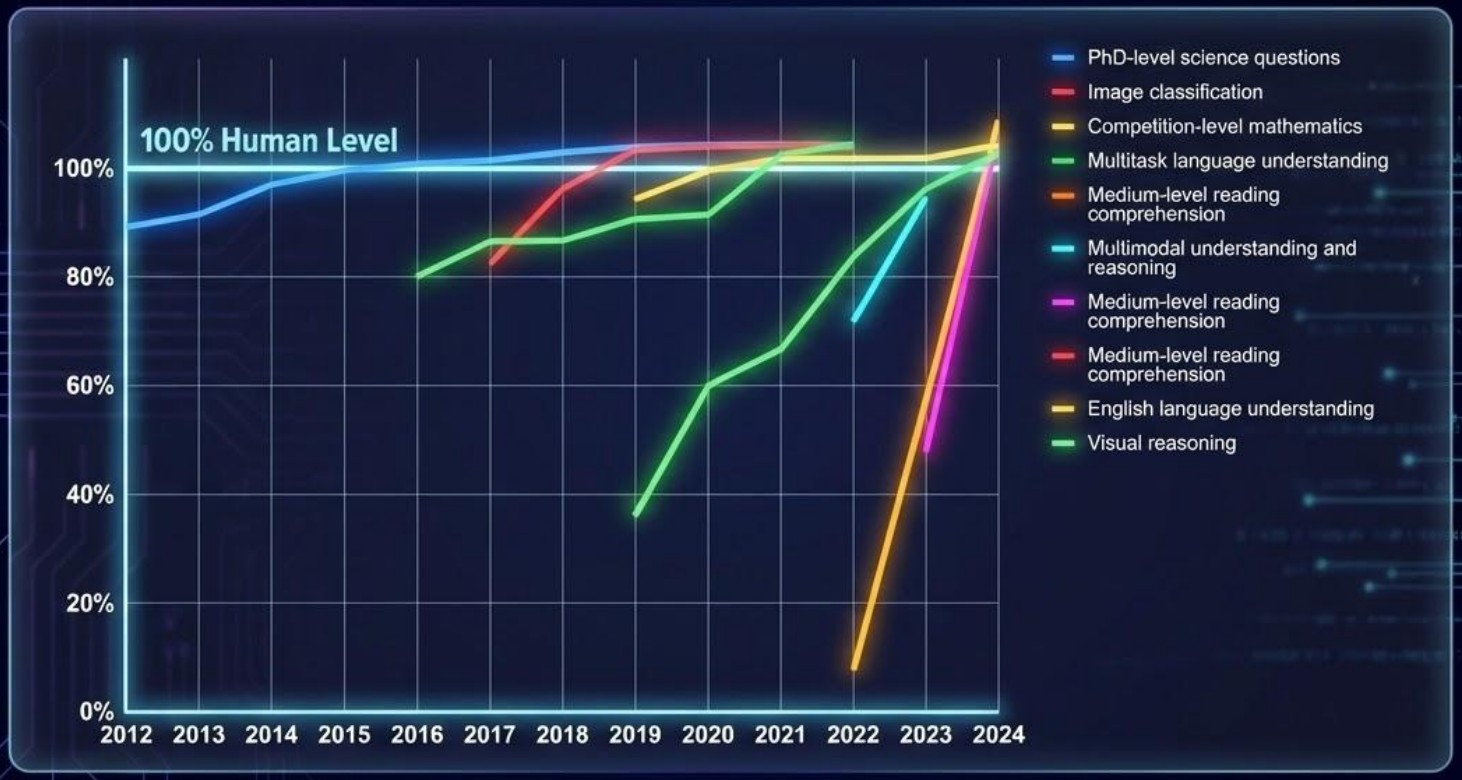

Computational Challenges
on the Road to AGI
John Hennessy | Stanford University | January 2026
AGI

# Challenges and Opportunities



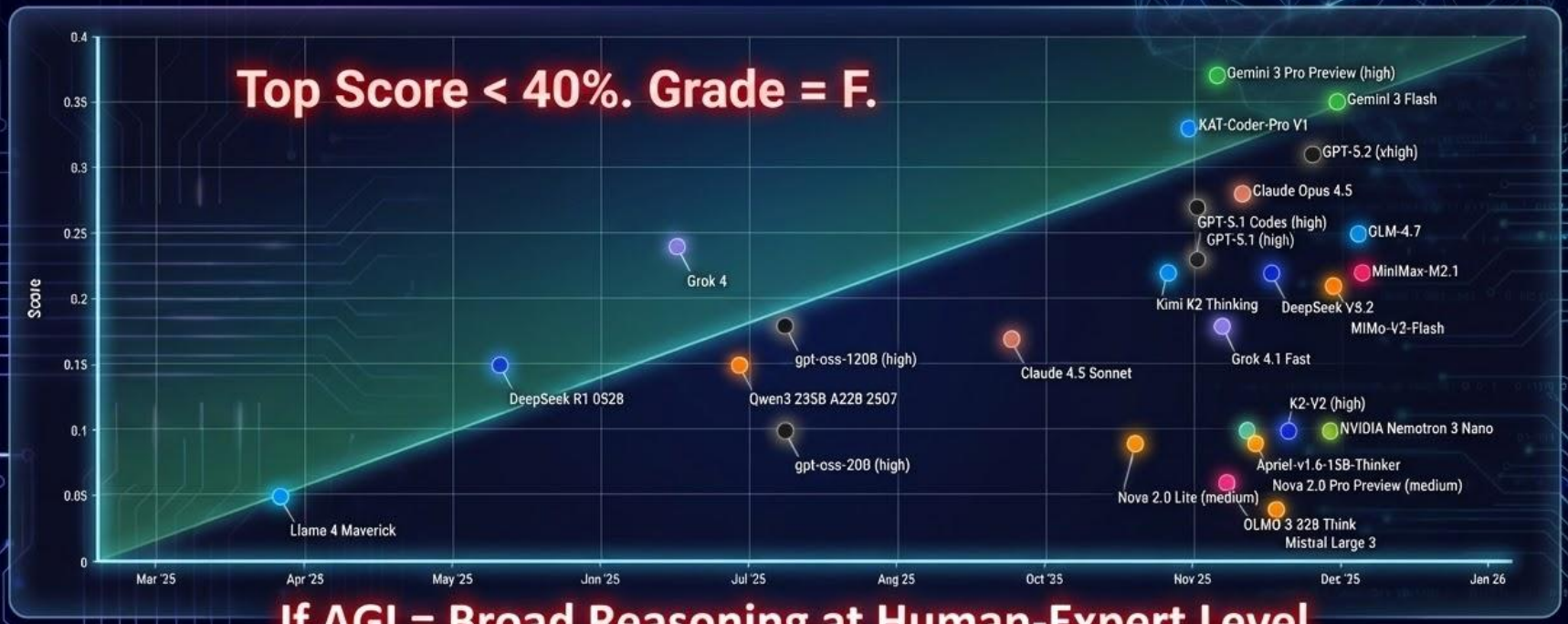
# Hurdles on the road to AGI

# Progress on Benchmarks (100% = Human Level)




Source: AI Index, 2025 | Chart: 2025 AI Index report

# Humanity's Last Exam (2,500 difficult problems)

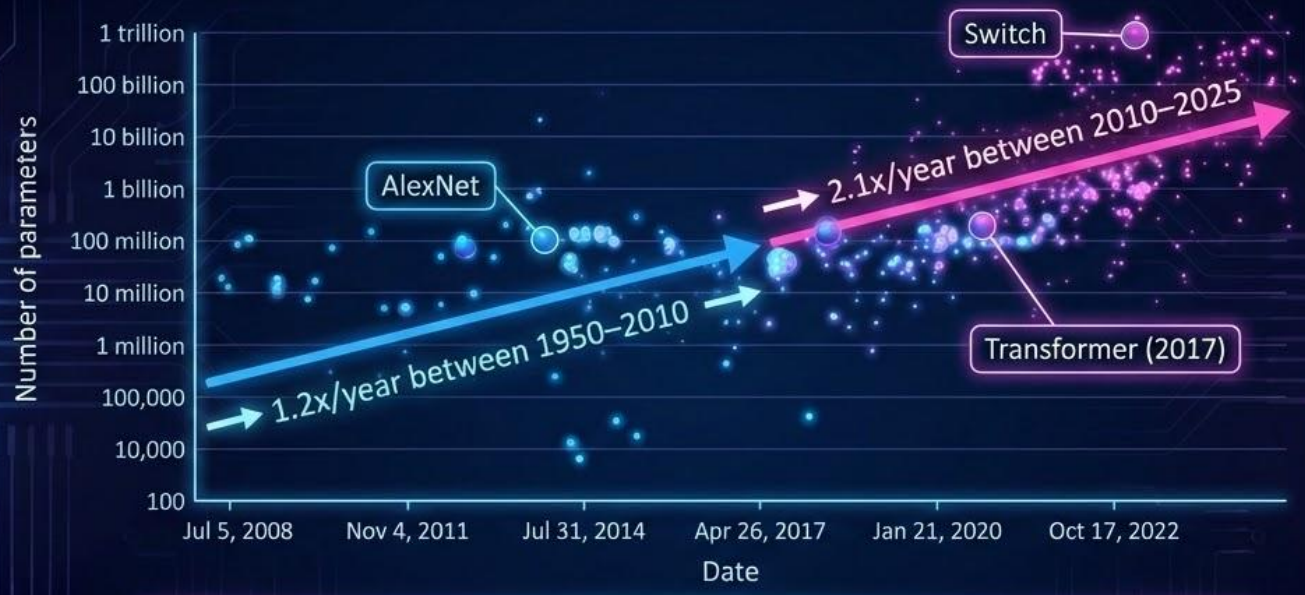



**If AGI = Broad Reasoning at Human-Expert Level**

**We still have a ways to go!**

Source: HLE Leaderboard

# Parameter Growth Rate




**Simply scaling existing models is unlikely to achieve AGI.**

Source: Epoch AI 2025

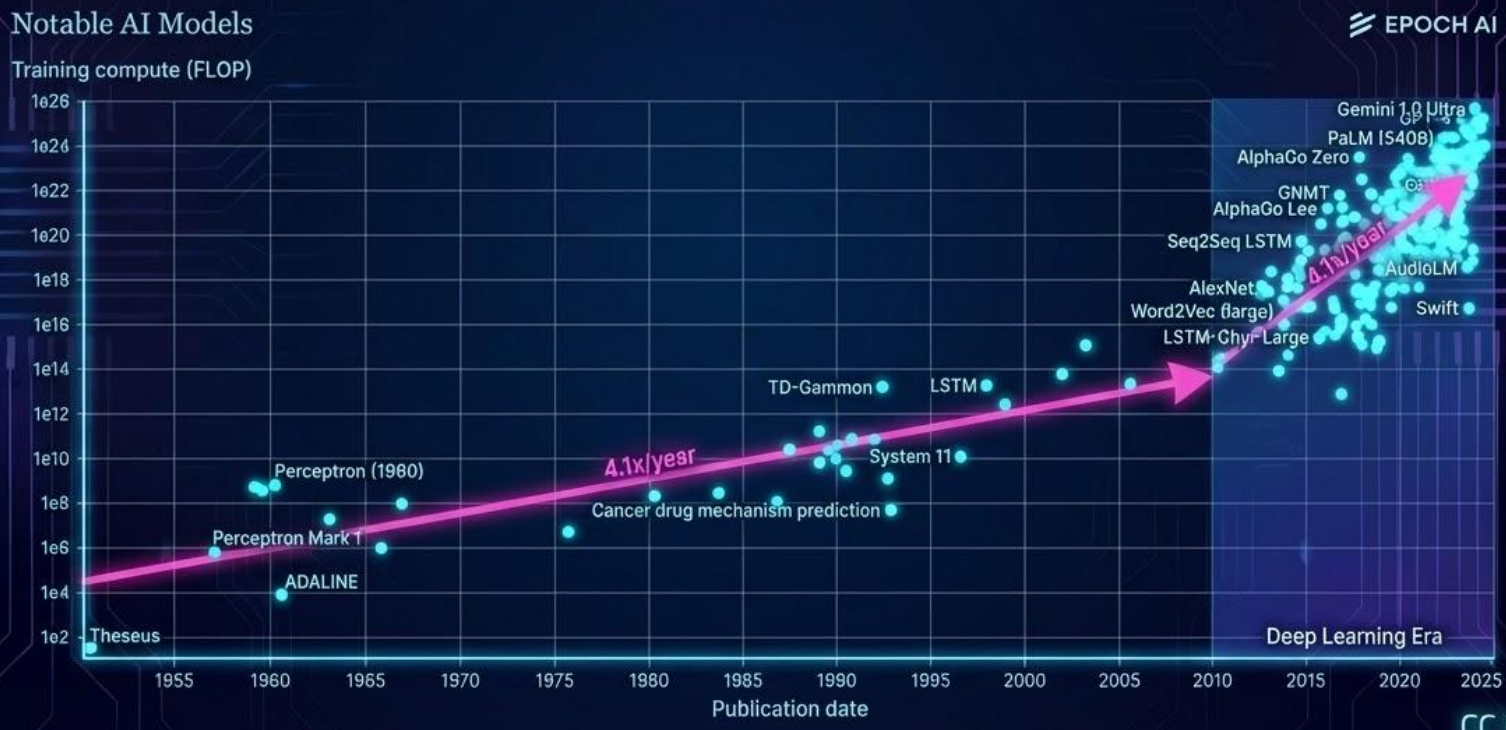
# Training Data Sets Growing Quickly


Nonetheless, it is likely that both models and (hence) training data will continue to grow rapidly.

Source: Epoch AI 2025

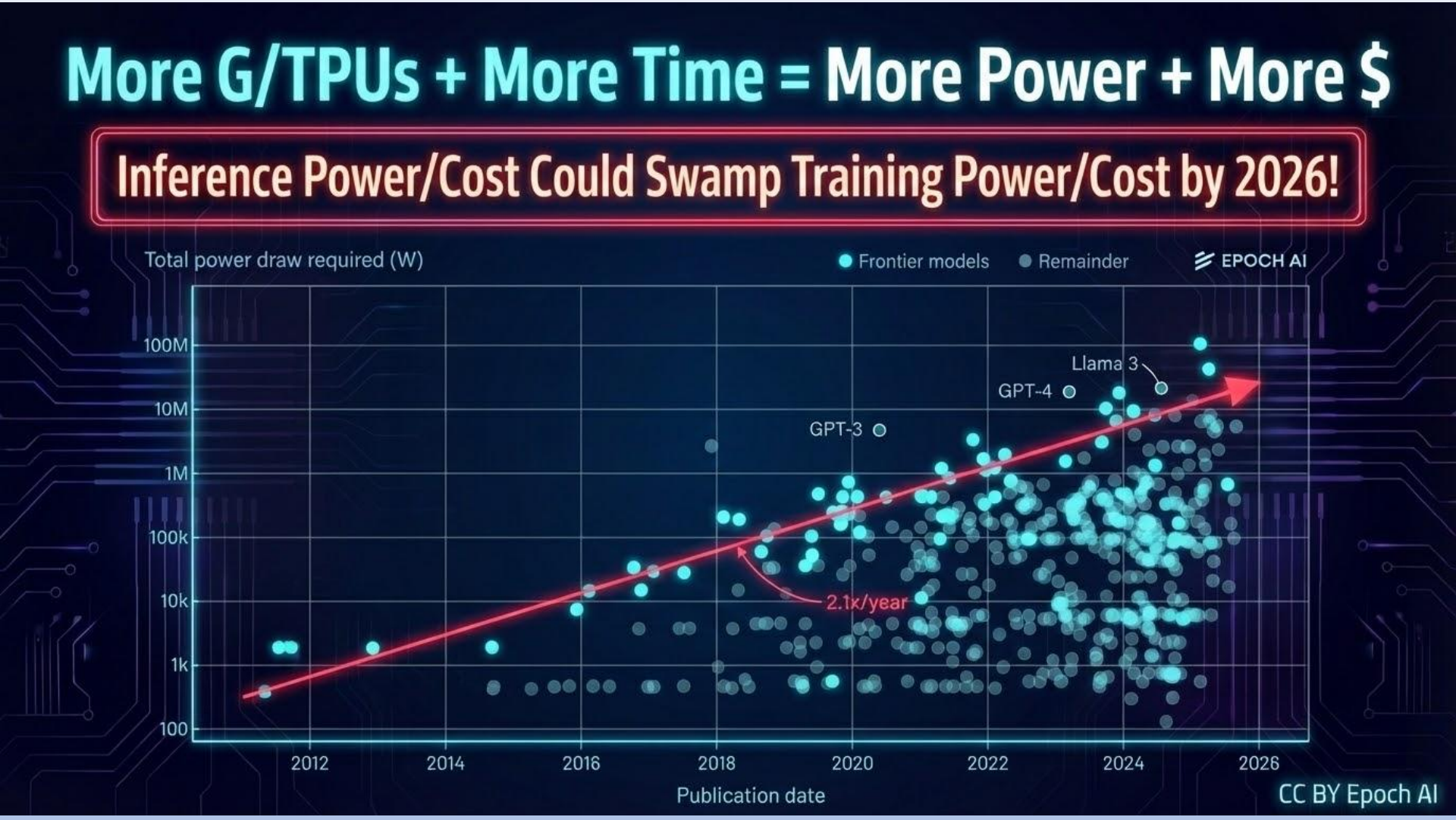
# Hence: Computational Demand Growing Fast!

## 4x annual growth is ≅2x growth rate of GPU!



# More G/TPUs + More Time = More Power + More $

## Inference Power/Cost Could Swamp Training Power/Cost by 2026!

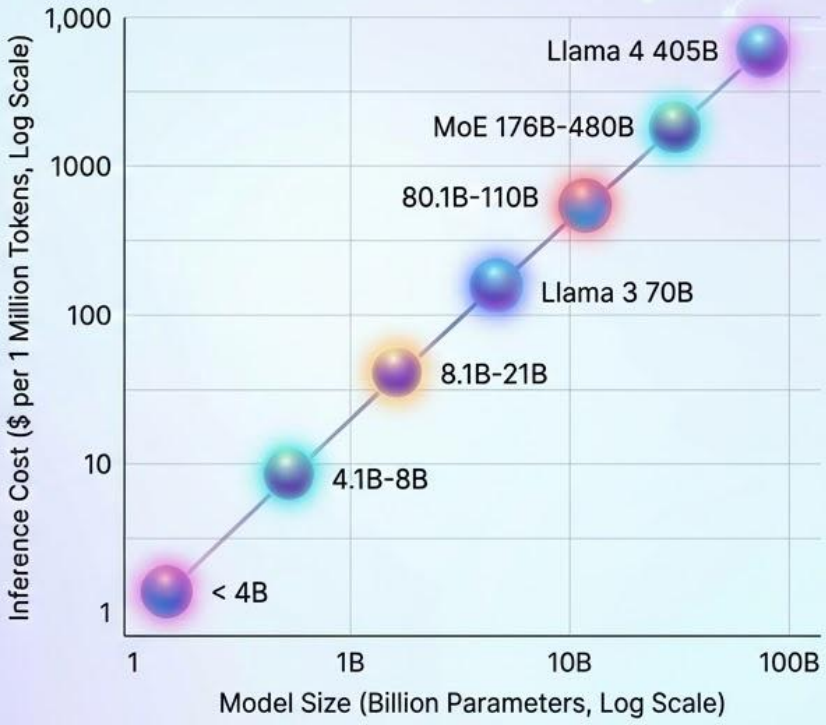


# Inference Cost vs. Model Size

Estimated Inference Cost vs. Model Size for LLMs (Together.AI)

**Hard to get consistent measures on energy + cost for inference, due to:**

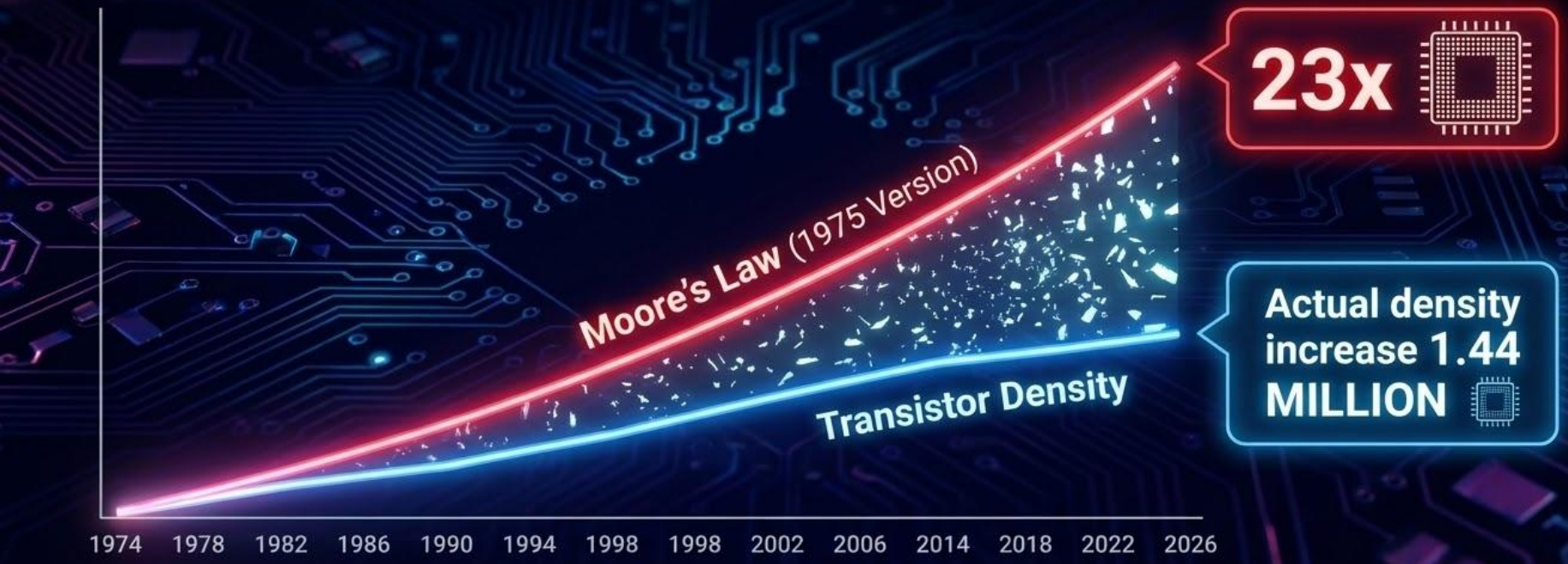
- variations in models
- variation in accuracy
- effect of batch size
- optimized models/algorithms



# Technology Hurdles

# Moore's Law: slowdown (Post Moore's Law Era)

Gap is likely to continue to increase, particularly for memory technologies.

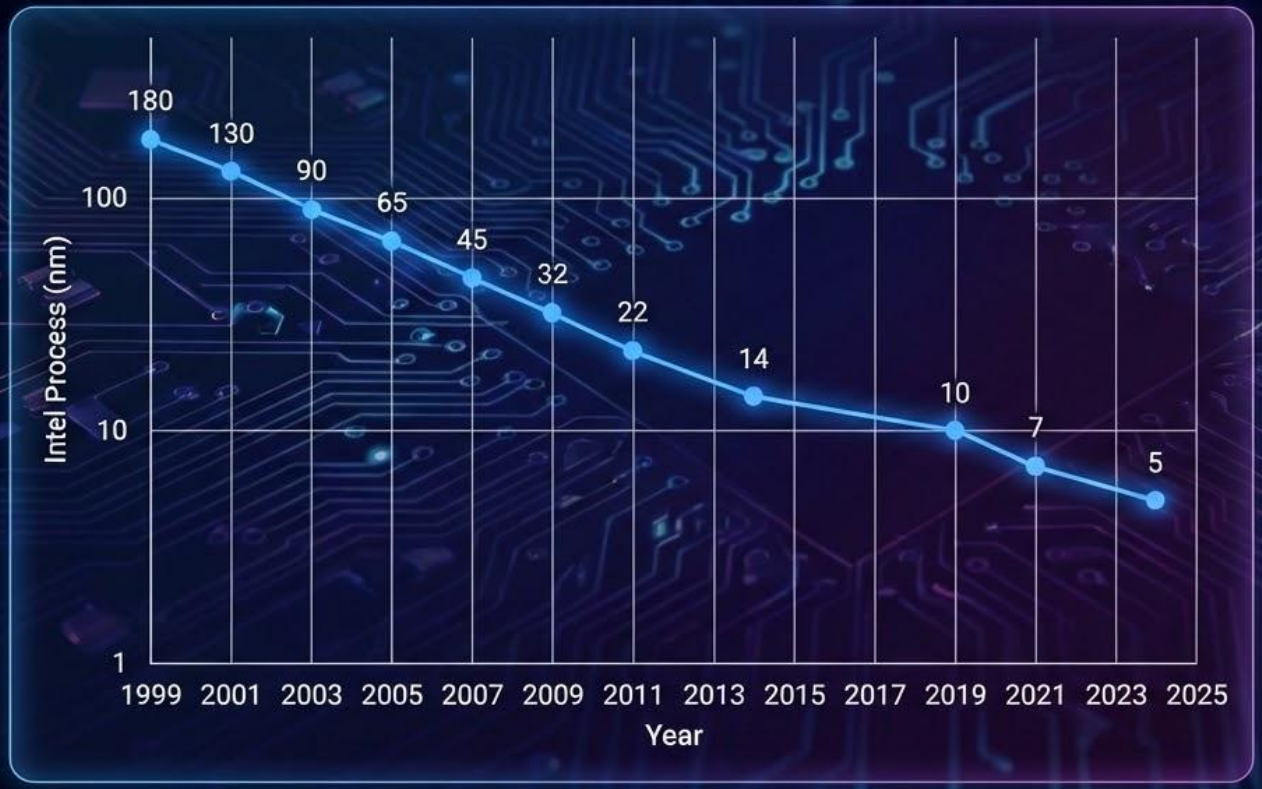



**23x**

**Actual density increase 1.44 MILLION**

⚠ **Gap is likely to continue to increase, particularly for memory technologies.**

# Moore's Law–time between nodes is increasing

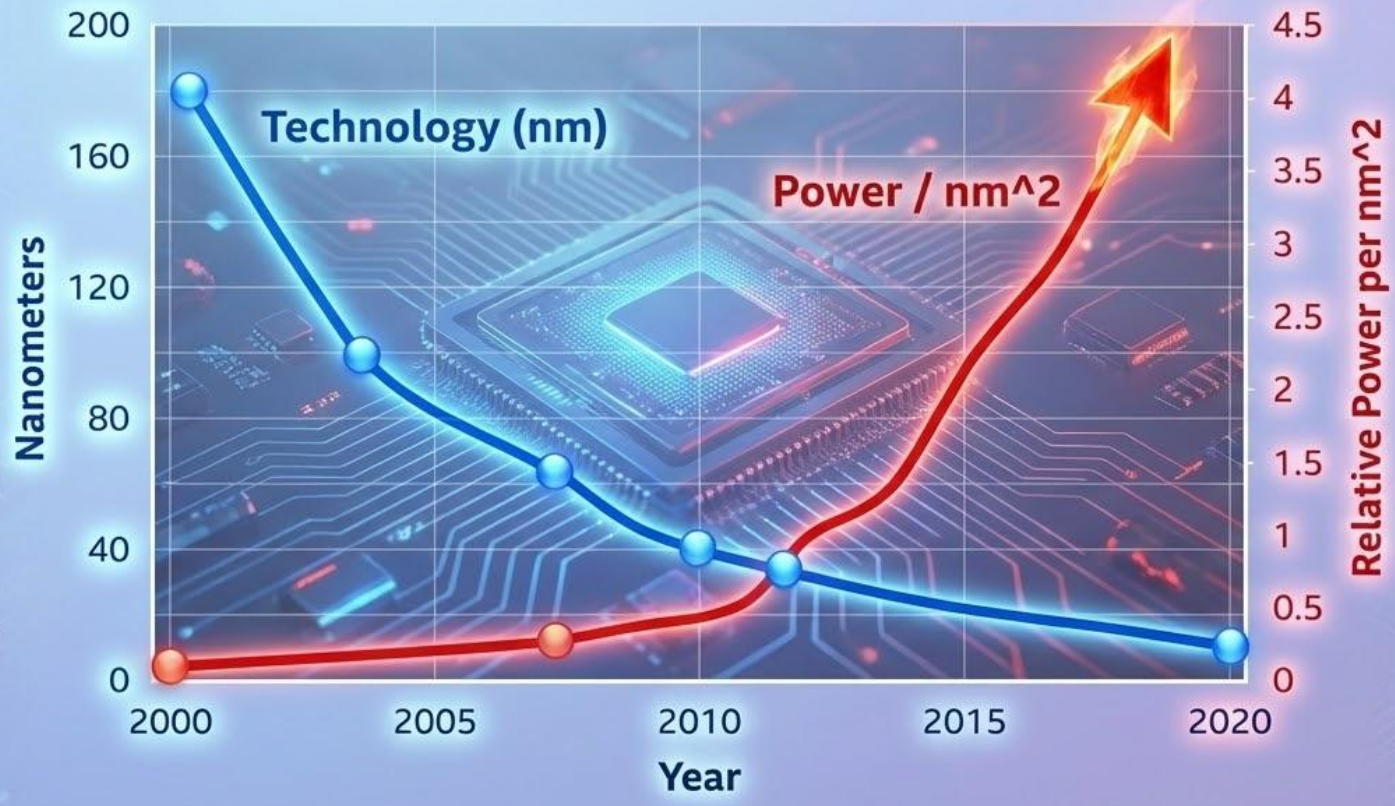



## International Roadmap for Semiconductors

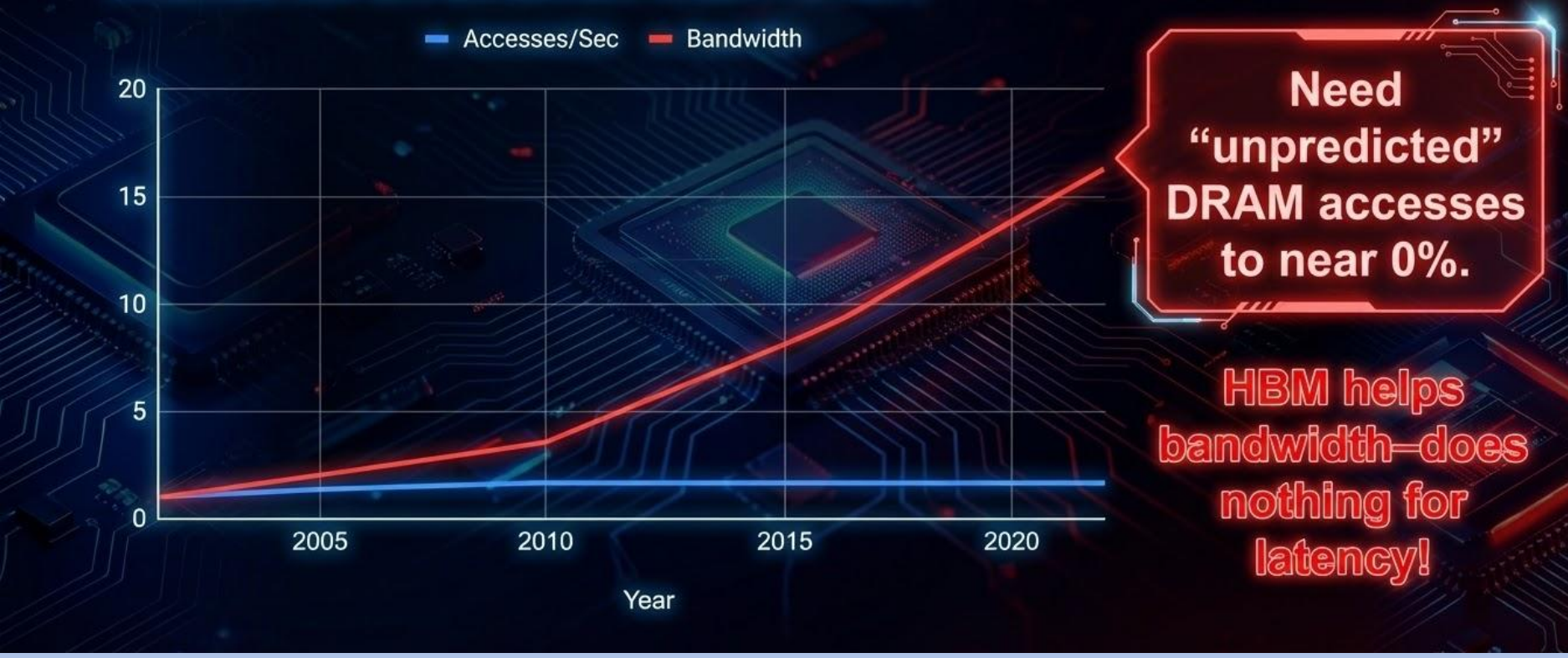
- Widely accepted plan for developing ICs.
  - Lasted 25 years.
- **Ended 10 years ago: too hard!**

# End of Dennard Scaling → Power is New Limit

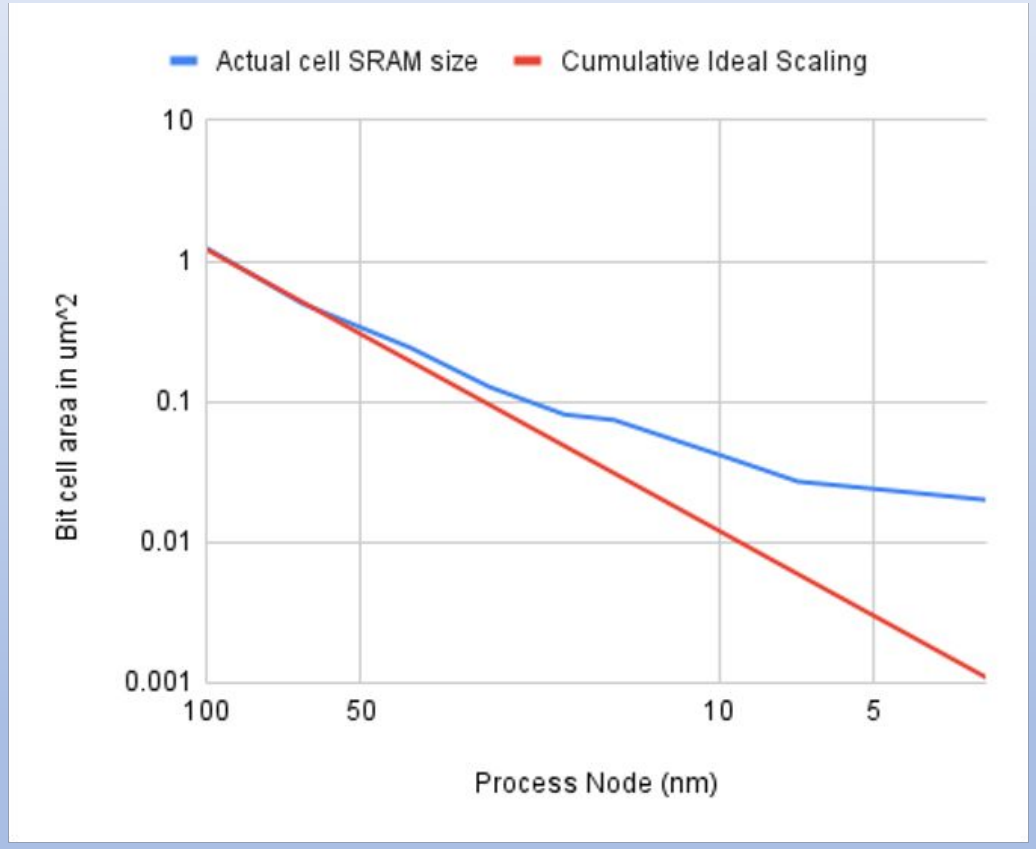
## EoML & Rise of Machine Learning

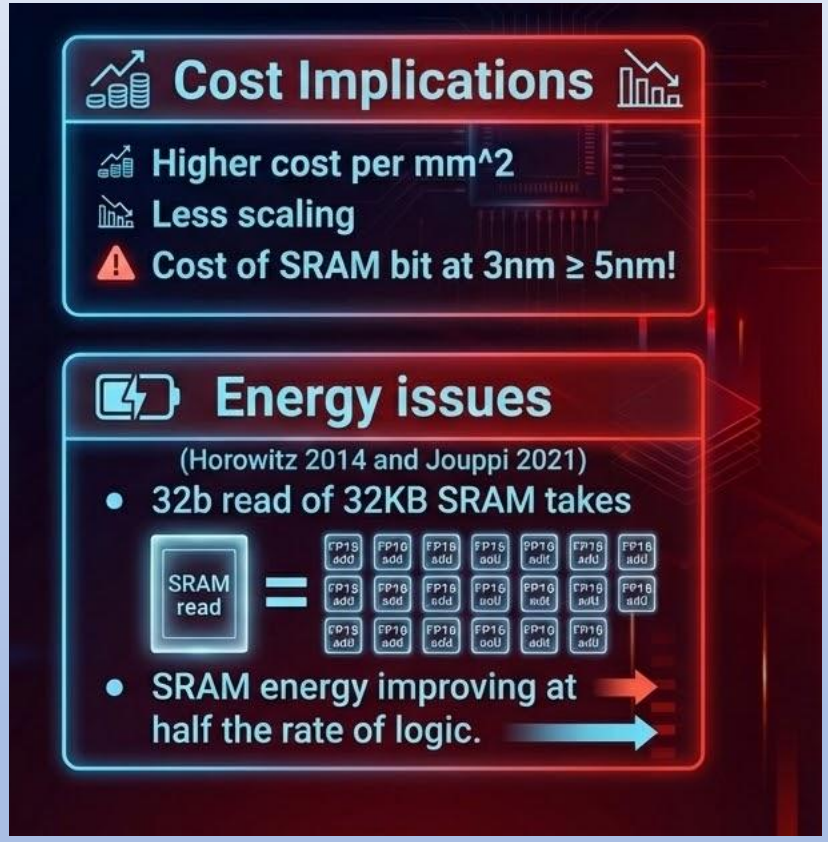


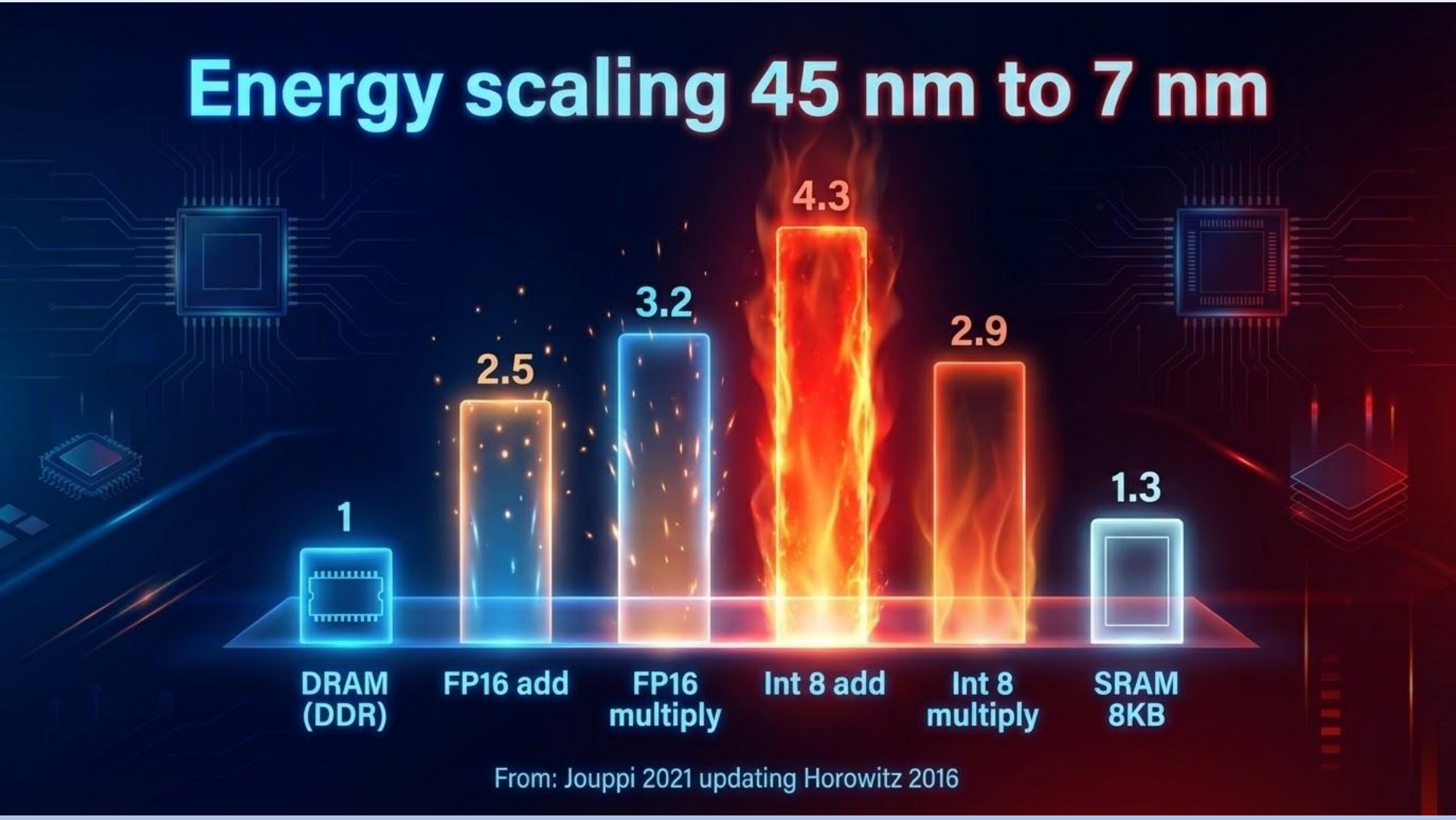
# DRAM Scaling Access Time vs. Bandwidth

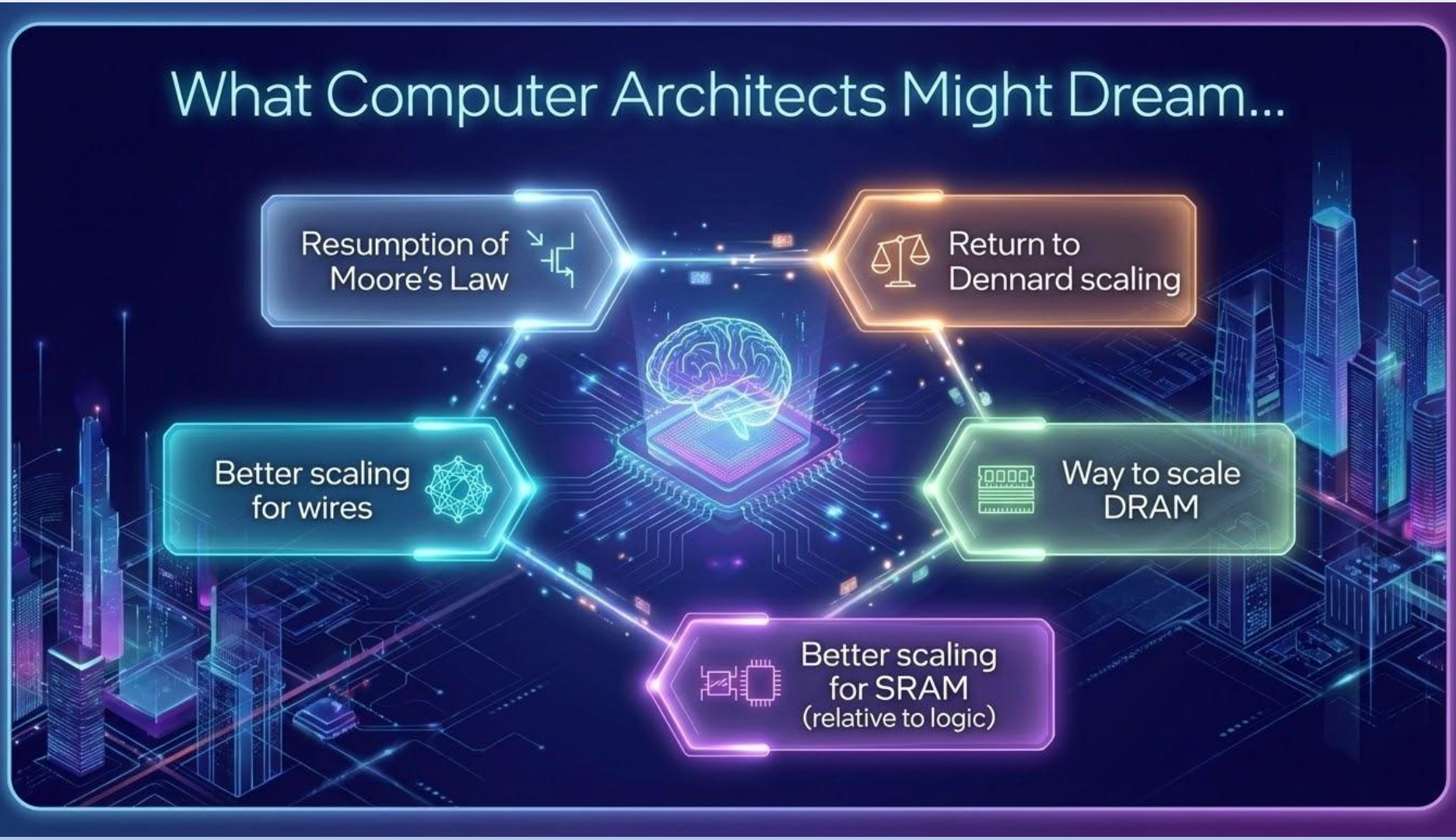


# SRAM Scaling: Falling Behind Logic




Energy scaling 45 nm to 7 nm
1
2.5
3.2
4.3
2.9
1.3
DRAM (DDR)
FP16 add
FP16 multiply
Int 8 add
Int 8 multiply
SRAM 8KB
From: Jouppi 2021 updating Horowitz 2016


What Computer Architects Might Dream...
Resumption of Moore's Law
Return to Dennard scaling
Better scaling for wires
Way to scale DRAM
Better scaling for SRAM (relative to logic)

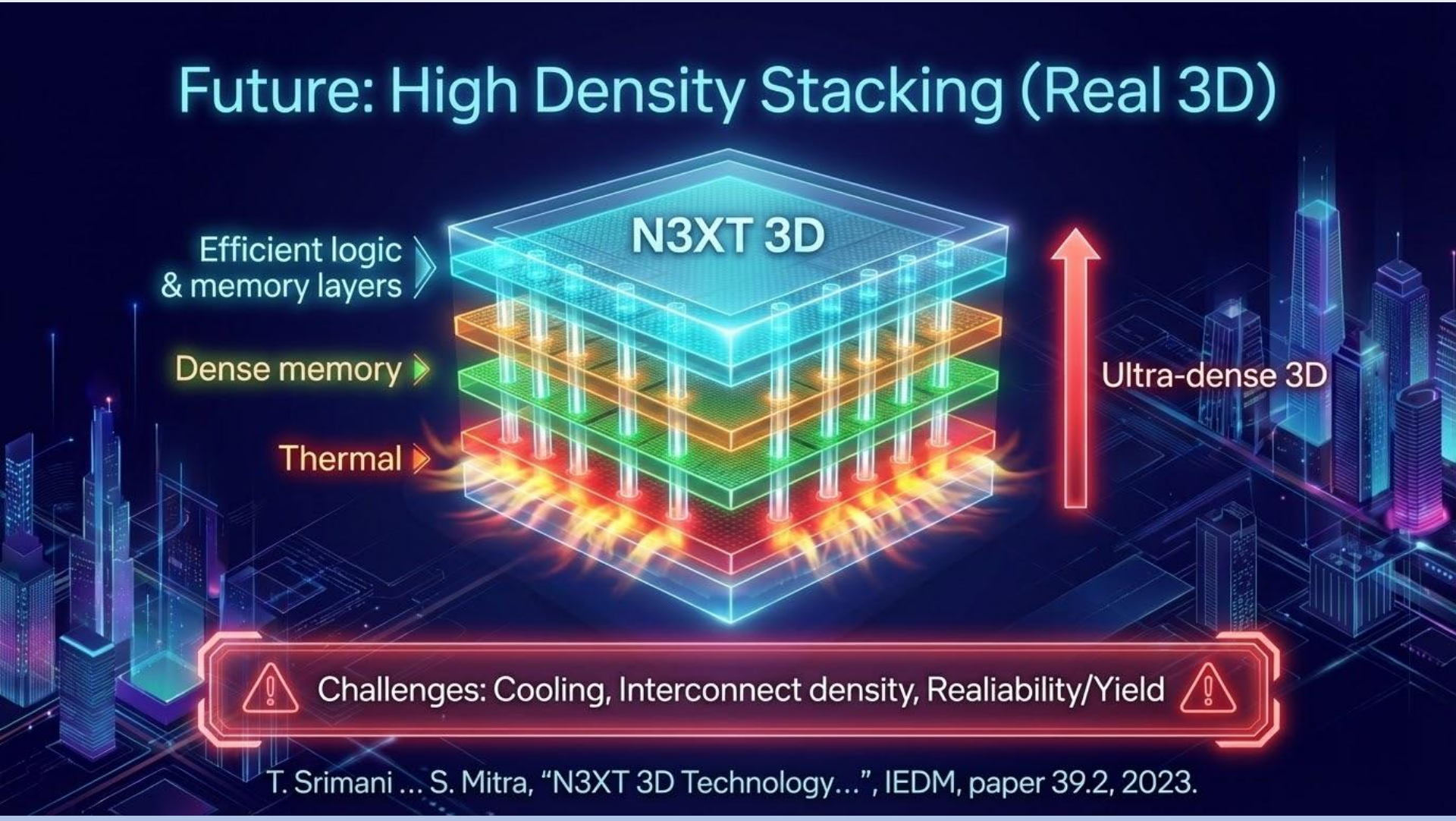
# Better Interchip Interconnect with Chiplets

**Each "chip" can use the right technology**
For example, why put large memory on die of high-end CPU or GPU? Better way to mix cores?

**Reuse of the chips lowers costs**
reduce design & mask cost: two smaller common chips replace single large chip

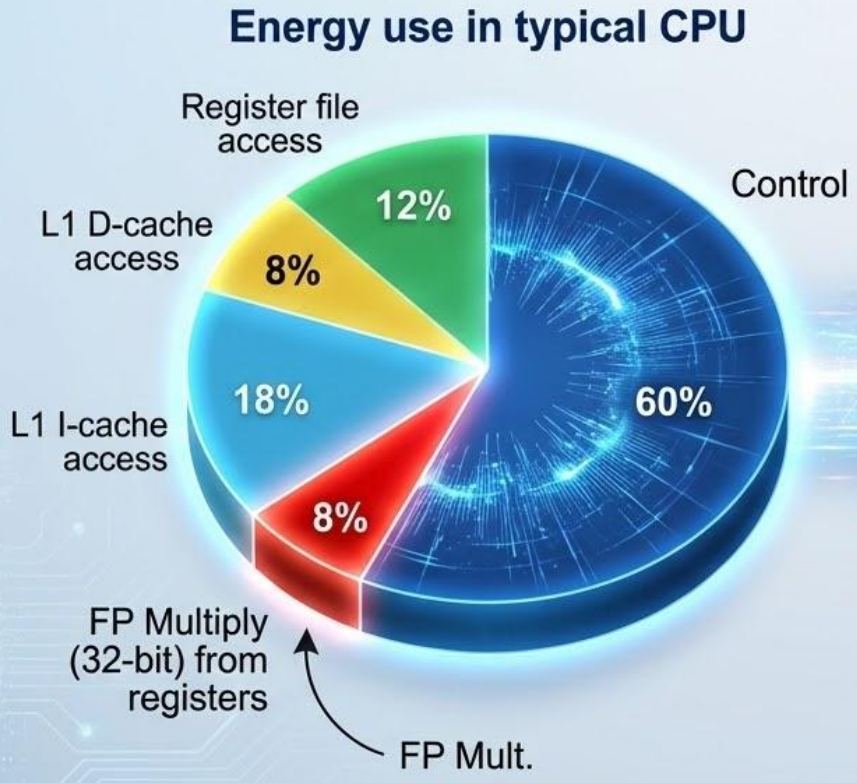

**Managing power: easier**
Both heat removal and power distribution

**Future interchip interconnect should improve quickly**
optical? chip-to-chip or chiplet?


Future: High Density Stacking (Real 3D)
N3XT 3D
Efficient logic & memory layers
Dense memory
Thermal
Ultra-dense 3D
Challenges: Cooling, Interconnect density, Realiability/Yield
T. Srimani ... S. Mitra, "N3XT 3D Technology...", IEDM, paper 39.2, 2023.

# Architecture:
# Success and Hurdles

# Why Domain Specific Architectures Won: Tailoring the Architecture




## More Efficient Parallelism

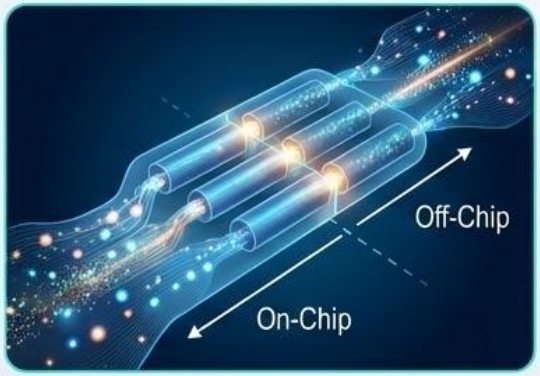
SIMD vs. MIMD:
↑FLOPs/memory access

## Less Control Overhead

VLIW/vector vs. Speculative

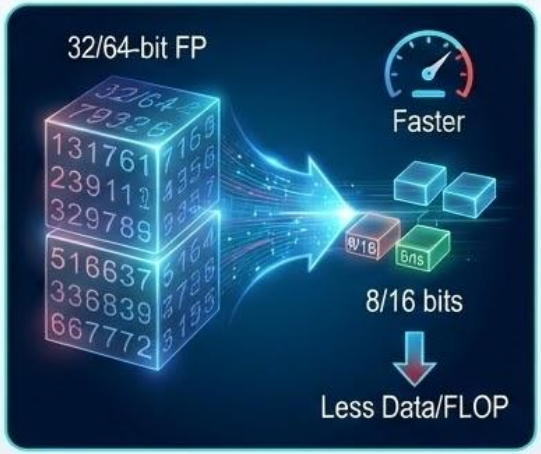
# Why Domain Specific Architectures Won: Tailor the Architecture to the Domain

## More effective use of memory bandwidth (on/off chip)



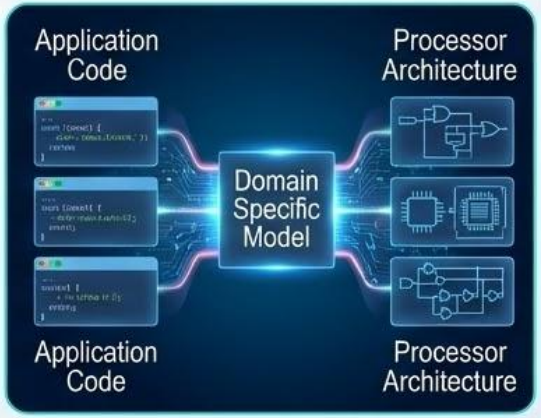

- User controlled versus caches
- Program prefetching to off-chip memory (vs. on demand)
- Massive parallelism to hide latency
- Mostly one-time enhancement

## Eliminate unneeded accuracy



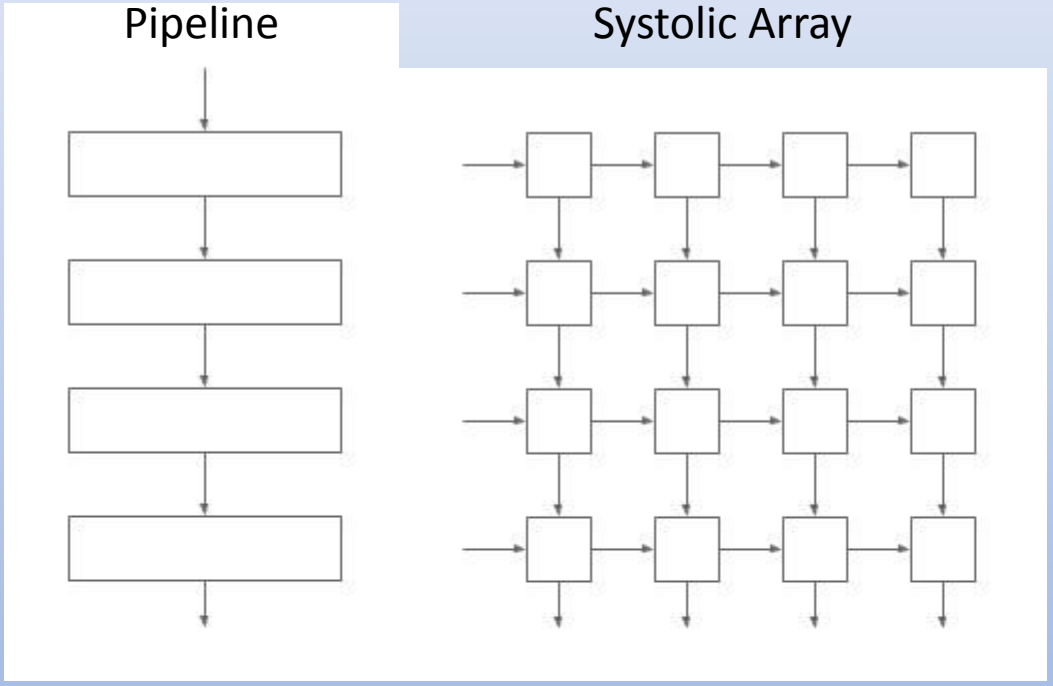

- 32-bit, 64-bit FP to 8-16 bits: faster math + less data/FLOP
- One-time improvement

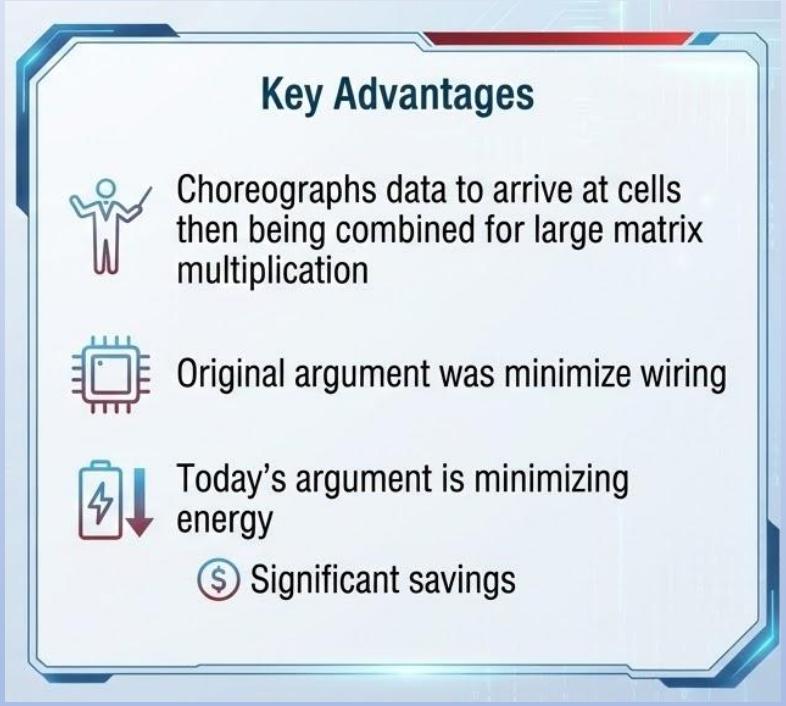
## Domain specific programming model matches application to the processor architecture




- Necessary ingredient
- Optimize the mapping from code to HW

# MXU Systolic Arrays: Two-Dimensional Pipelines



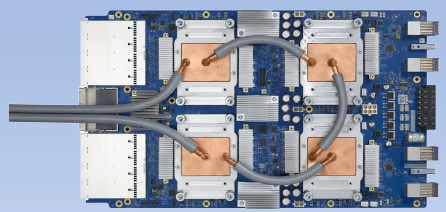



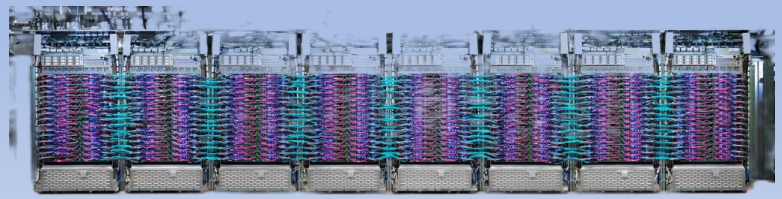



* Kung, H.T. and Leiserson, C.E., 1979. Systolic arrays (for VLSI). In Sparse Matrix Proceedings 1978 (Vol. 1, pp. 256-282). *Society for Industrial and Applied Mathematics*.

# Scale FLOPs, HBM BW, Pod Size to Meet Demand

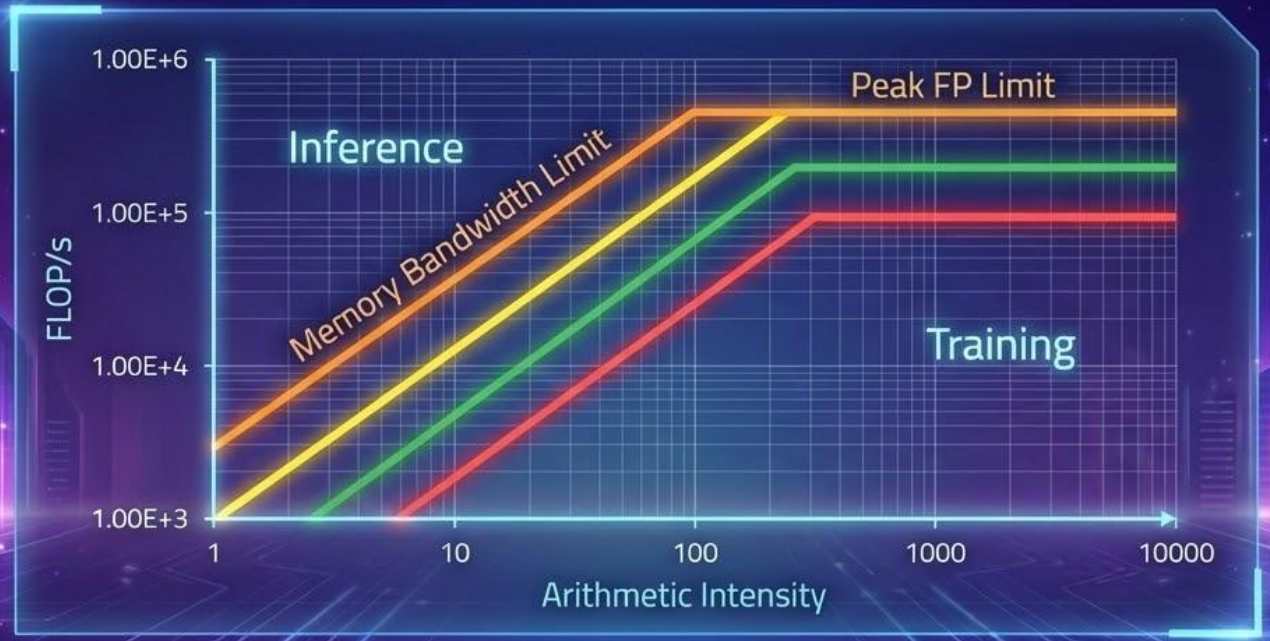
| TPU | TPU v2 | TPU v3 | TPU v4 | TPU v5p | TPU v7 |
|---|---|---|---|---|---|
| Peak BF16 (TFLOPS) | 46 | 123 | 275 | 459 | 2,300 |
| HBM capacity | 16GB | 32GB | 32GB | 96GB | 192GB |
| HBM BW (GB/s) | 600 | 900 | 1,228 | 2,765 | 7,370 |
| Chips/pod | 256 | 1024 | 4096 | 8960 | 9216 |
| Addition | HBM | H2O cooled | Optical Sparsity | Big switch | Chiplet FP8 |



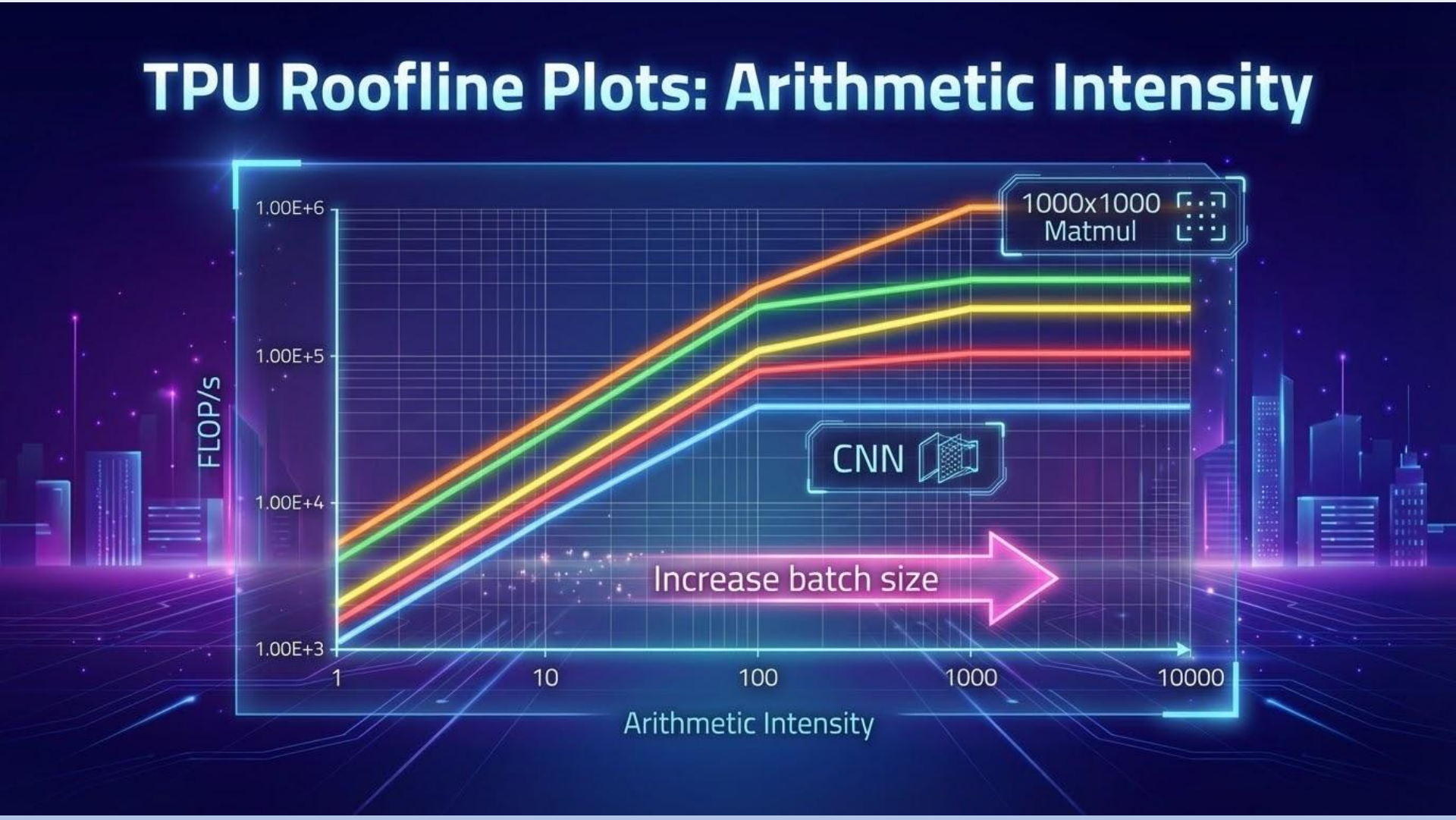


Éloïse

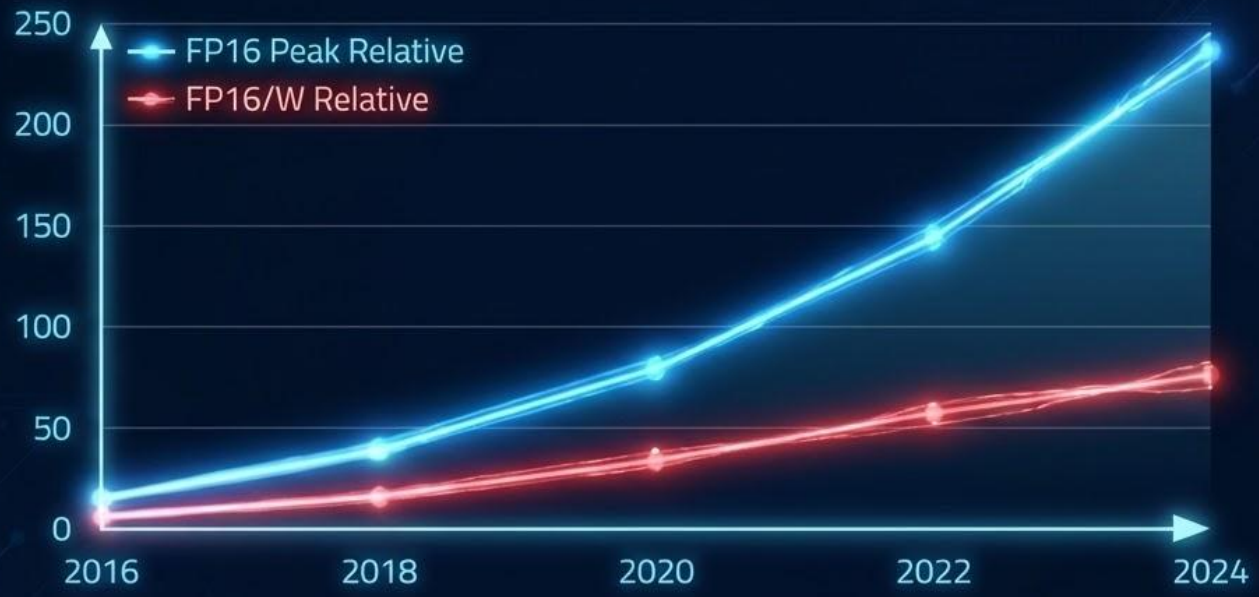
# TPU Roofline Plots




Arithmetic intensity = FLOPs per byte of memory

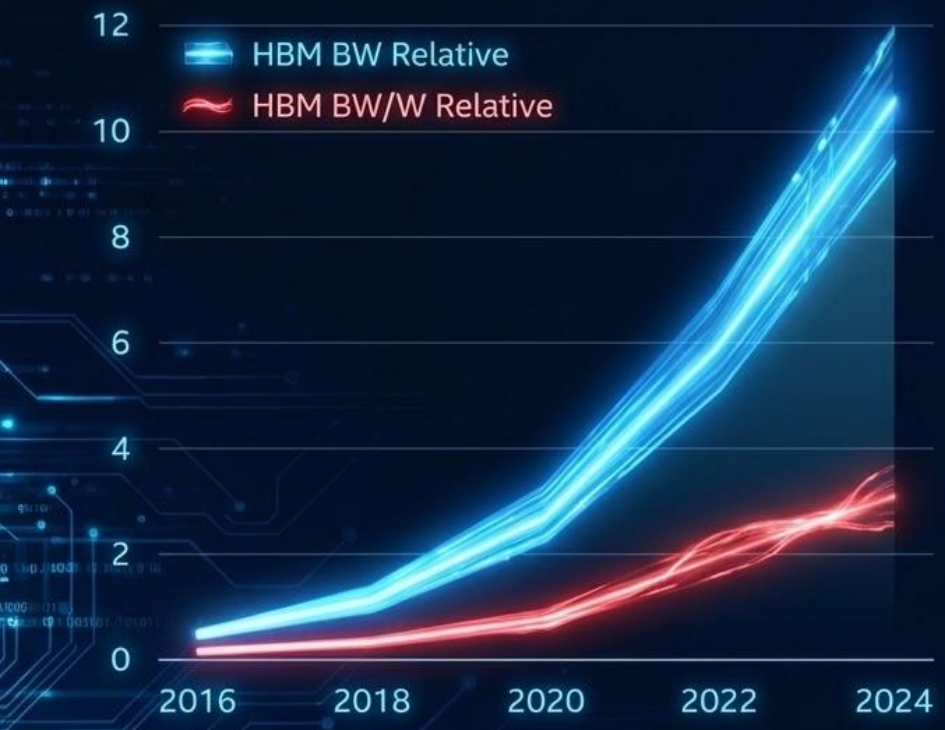

TPU Roofline Plots: Arithmetic Intensity
1.00E+6
1.00E+5
1.00E+4
1.00E+3
FLOP/s
1000x1000
Matmul
CNN
Increase batch size
1
10
100
1000
10000
Arithmetic Intensity

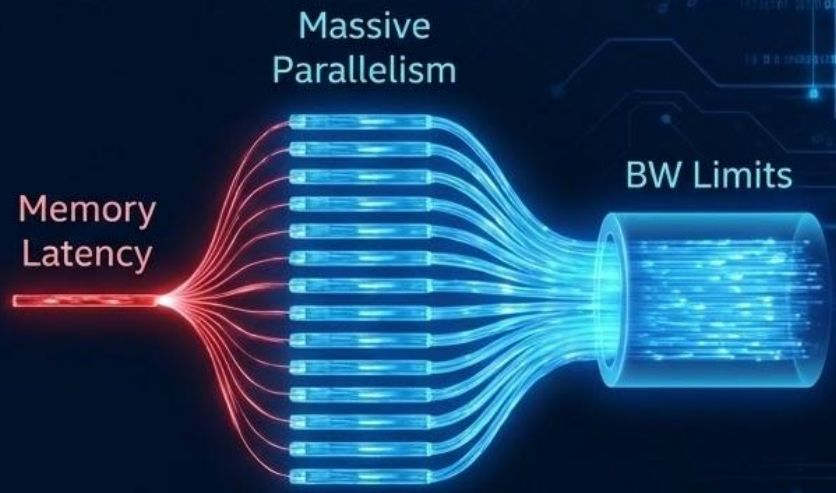
# Peak GPU FP Performance and Performance/W




FLOPS are easiest feature still scale, but lags in FLOPS/W!

# Peak HBM BW and HBM BW/W

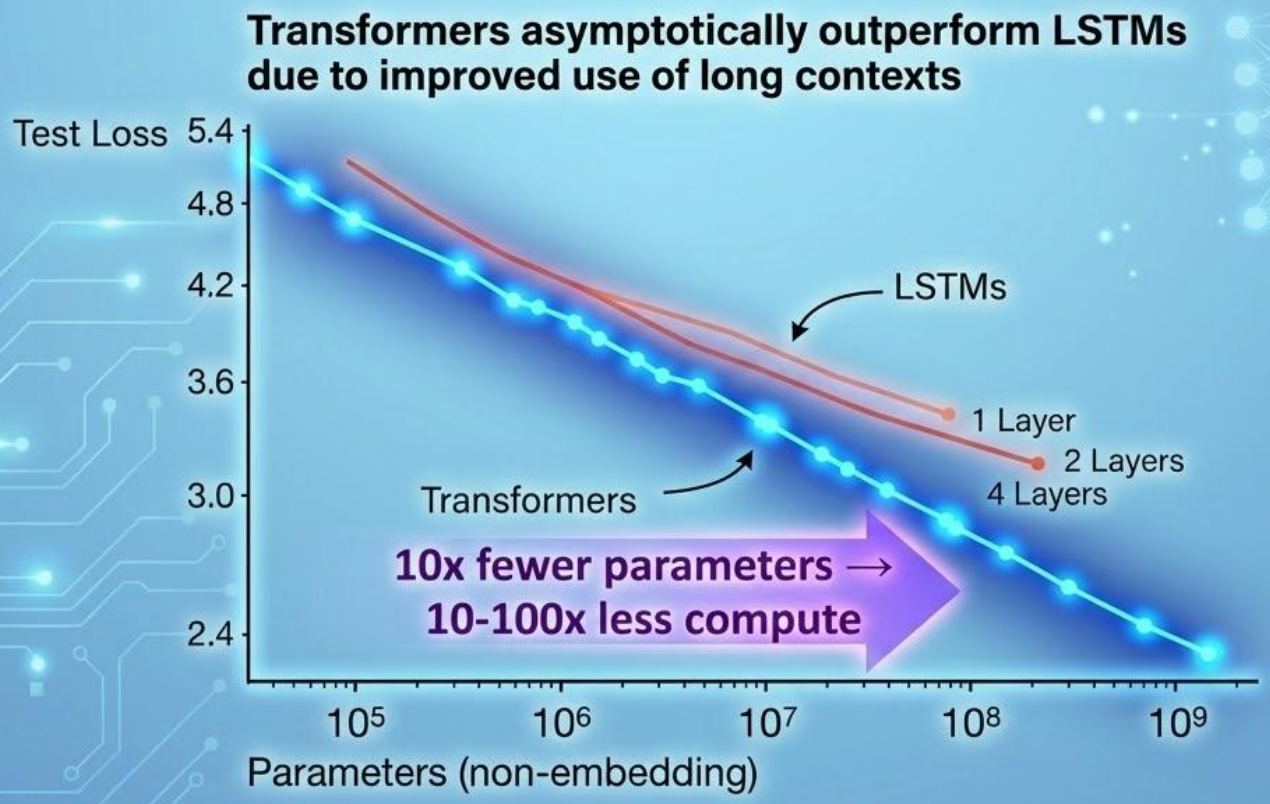



Massive parallelism hides memory latency–not BW limits

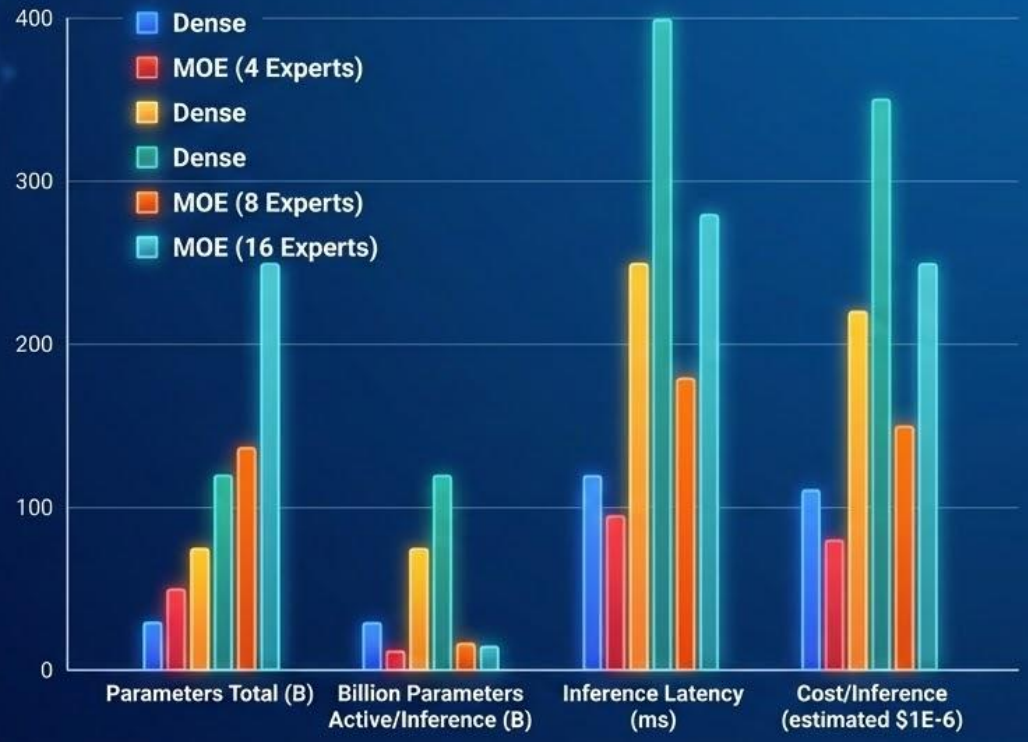
# Model Architecture & Algorithms Critical Ingredients

# The Advantage of Transformers




Data from: Scaling Laws for Neural Language Models, Kaplan, et. al.

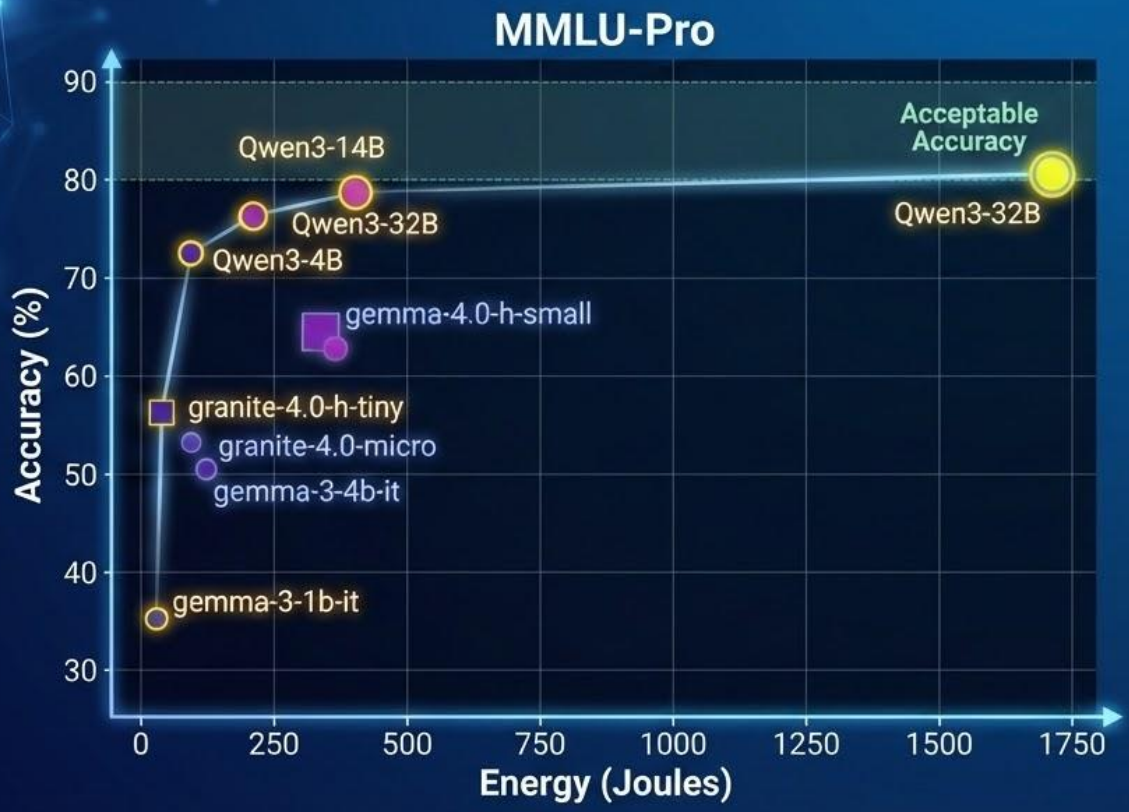
# MoE and Dense Models: Inference Cost




## Accuracy of a larger model; cost of smaller one!

Due to larger OH & sparsity, parameter reduction > latency or cost reduction.

# Choose the "Right" Model: Small is Beautiful

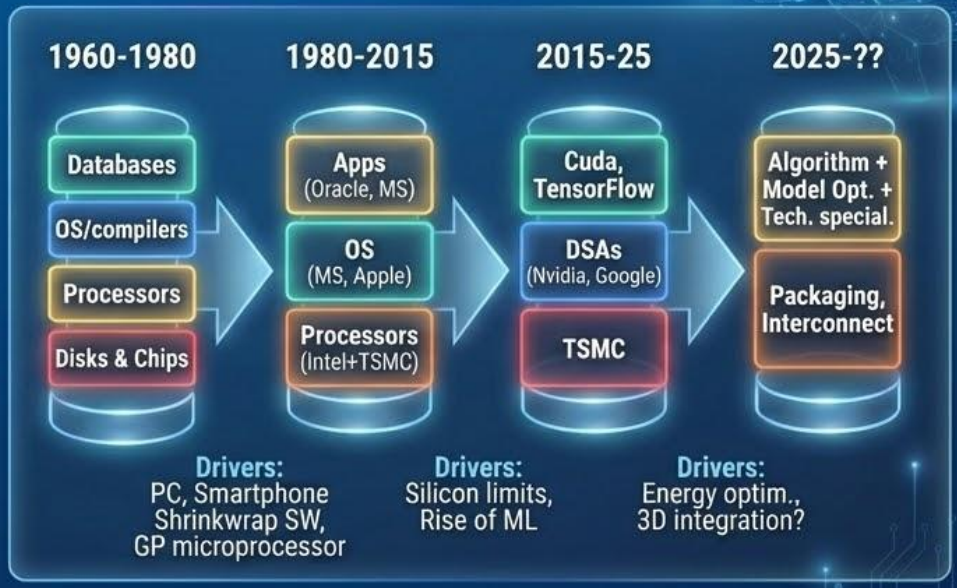



**14B-p model accuracy = 98% of 32B-p model AND uses only 20% of the energy!**

Acceptable Accuracy

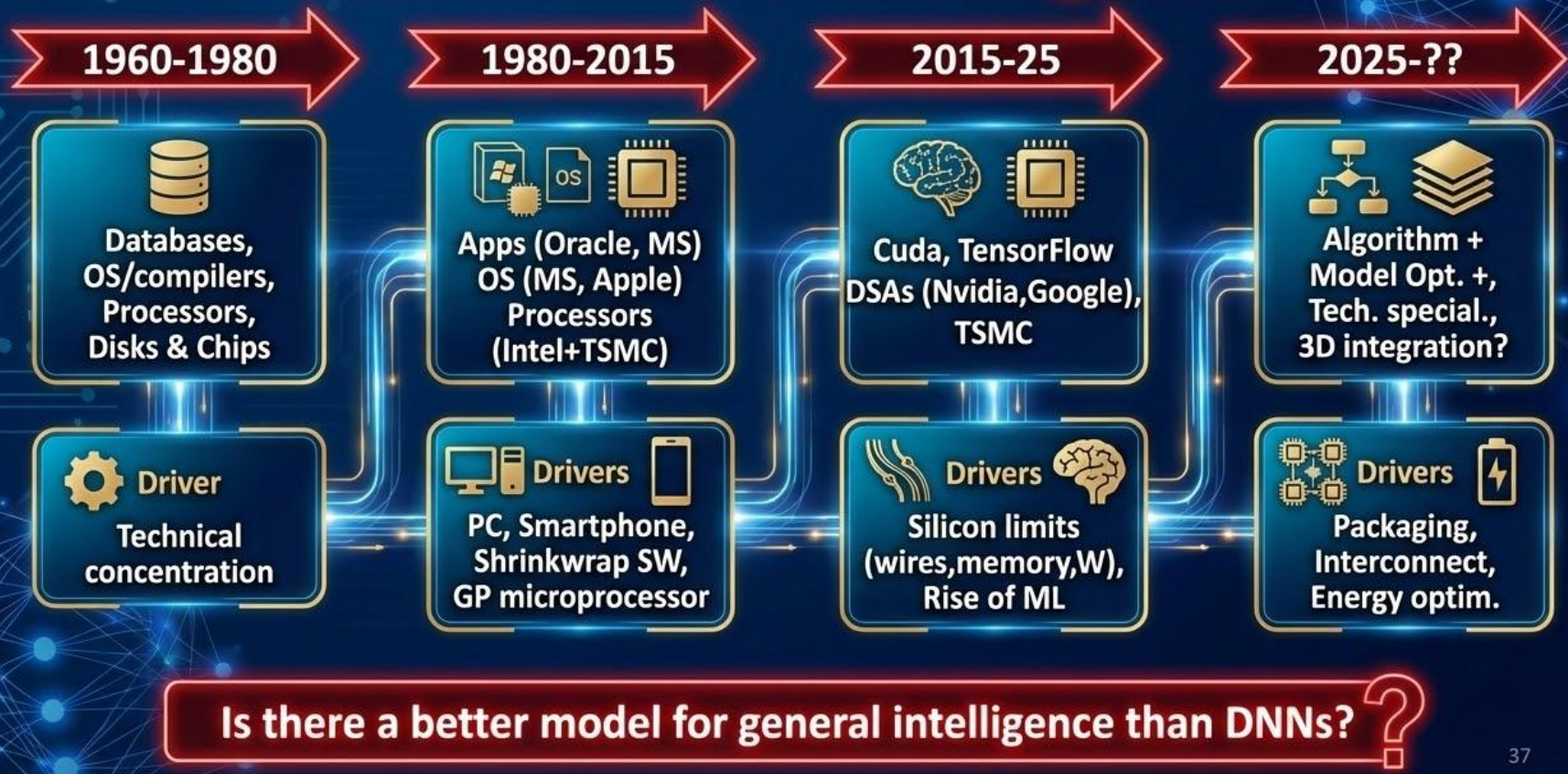
# Concluding Thoughts: Deja Vu

- In 2012±, end of Moore's Law & Dennard scaling → flatlined CPUs
- First decade of DSAs is near its end
  - Easy gains largely harvested
- Need new algorithms + new models + new architectures + enhanced technology.
- Full stack design!




Is there a better model for general intelligence than DNNs?

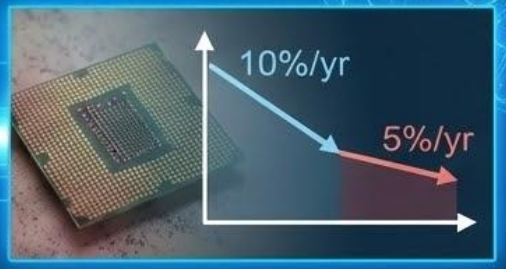
# Concluding Thoughts: Deja Vu: The Return of Vertical Integration




**Is there a better model for general intelligence than DNNs?**

# Concluding Thoughts: Deja Vu

In 2012±, end of Moore's Law & Dennard scaling → flatlined CPUs




First decade of DSAs is near its end

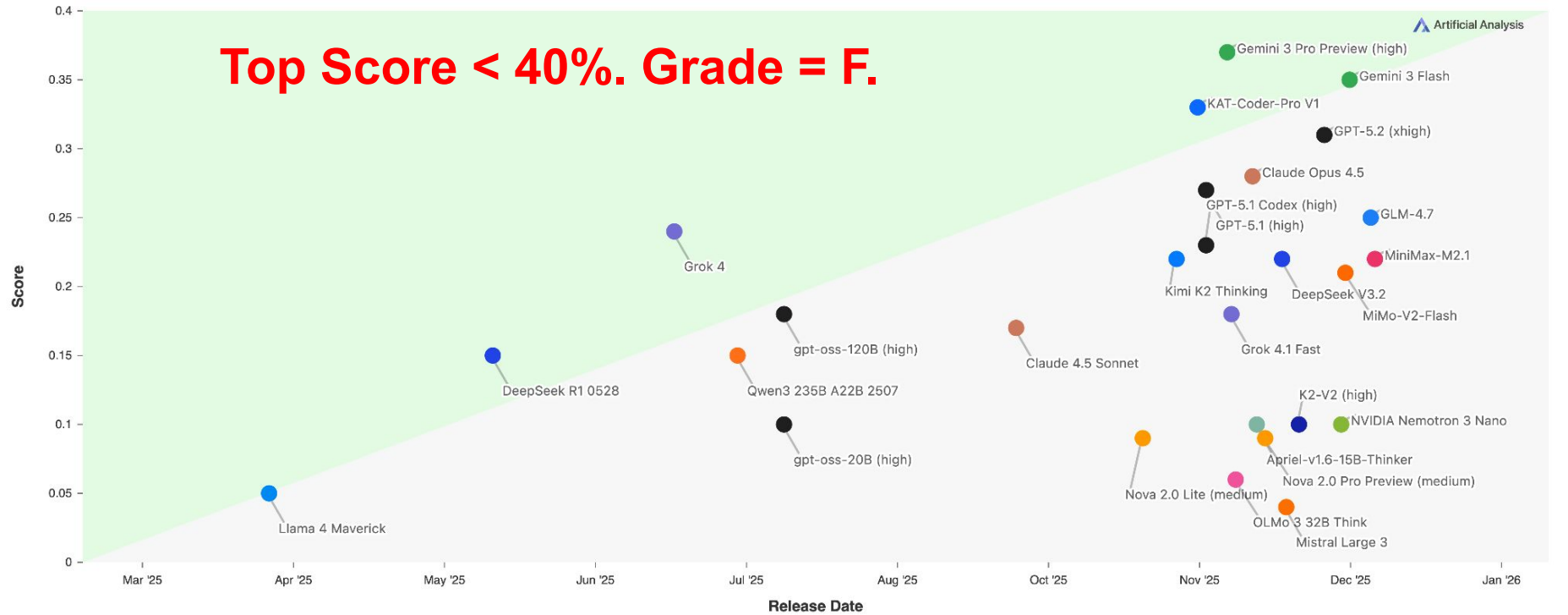
- Easy gains largely harvested


Need new algorithms + new models + new architectures + enhanced technology. Full stack design!

# THE FUTURE: ARTIFICIAL GENERAL INTELLIGENCE?


- When? 5-10 years?
  - LLMs have accelerated the crossover
  - Challenge: broad-based reasoning (MMLU, HLE!)
- DNNs are crude model for neurons (but good for silicon)
  - Natural learning (think babies)
    - Reinforcement learning growing importance
  - Learning time: e.g. top chess player
    - Human = years: 10K games?
    - Computer = hours to days: millions of games
  - Energy efficiency:
    - Brain power = 20 watts
    - Data center ≅ 1,000-10,000x
  - Can AI learn more from neuroscience?

# Humanity's Last Exam (2,500 difficult problems)



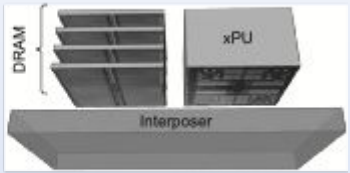

**If AGI = Broad Reasoning at Human-Expert Level**

**We still have a ways to go!**

Source: HLE Leaderboard

# Packaging: Going 3D

Vertical stacking (3D)



Interposer stacking (2.5D)

- High Bandwidth Memory (1st usage)

| | **HBM Advantage vs DDR4/5 DIMMs** |
|---|---|
| Bandwidth (per channel) | 15x higher |
| Power (per reference) | 4x lower |
| Latency | Similar or slightly worse |

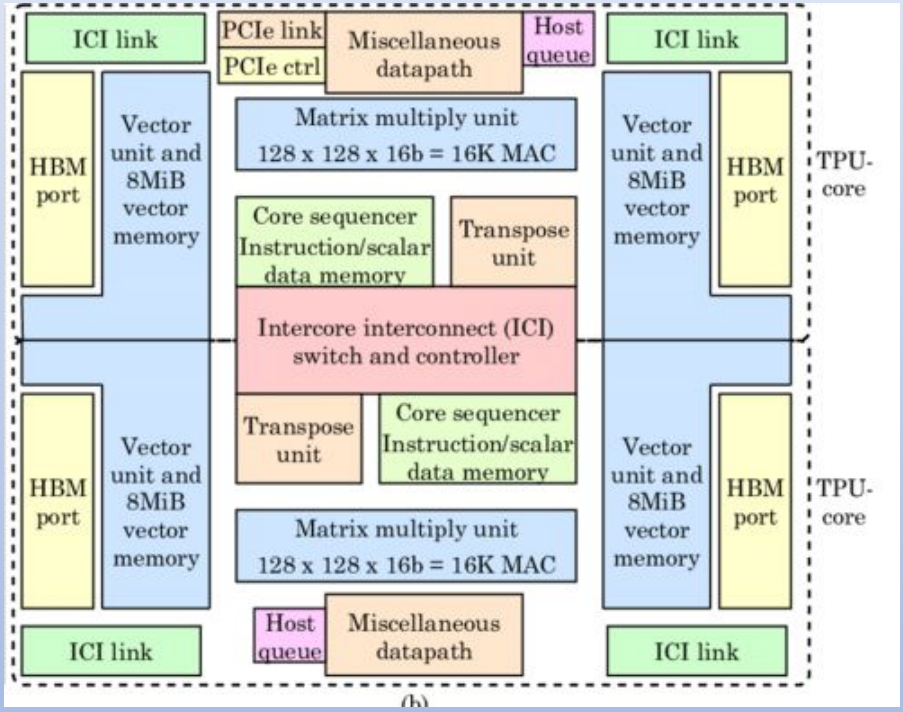
*Note: 3D in single chip isn't practical (due to processing steps).*

# Concluding Thoughts: Deja Vu
# The Return of Vertical Integration

| 1960-1980 | 1980-2015 | 2015-25 | 2025-?? |
|---|---|---|---|
| Databases<br>OS/compilers<br>Processors<br>Disks & Chips | Apps (Oracle, MS)<br>OS (MS, Apple)<br>Processors<br>(Intel+TSMC) | Cuda,TensorFlow<br>DSAs<br>(Nvidia,Google)<br>TSMC | Algorithm +<br>Model Opt. +<br>Tech. special.<br>3D integration? |
| **Driver**<br>Technical<br>concentration | **Drivers**<br>PC, Smartphone<br>Shrinkwrap SW,<br>GP microprocessor | **Drivers**<br>Silicon limits<br>(wires,memory,W)<br>Rise of ML | **Drivers**<br>Packaging,<br>Interconnect<br>Energy optim. |

**Is there a better model for general intelligence than DNNs?**

# How is Silicon Used: TPU v2 vs CPU



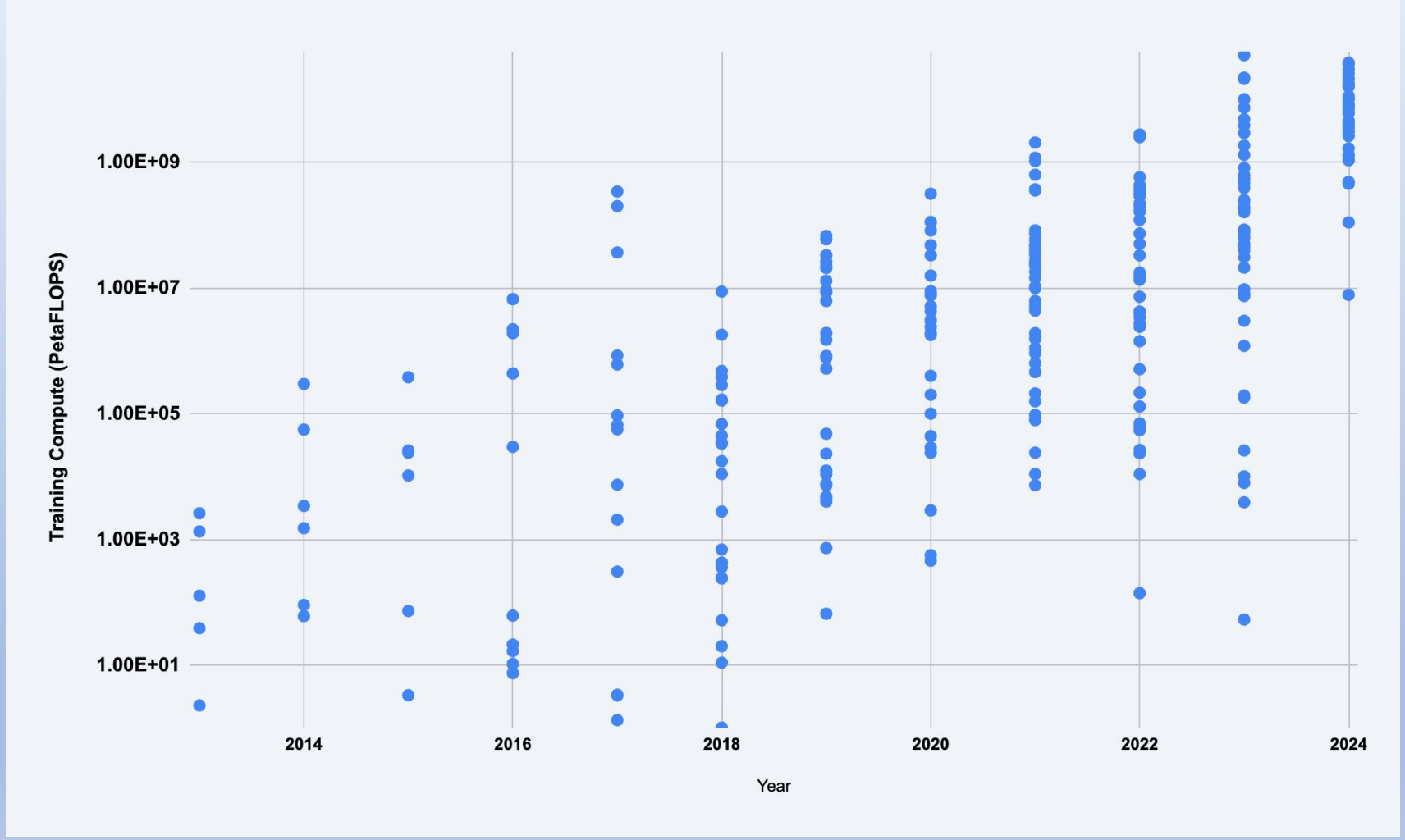

**TPU-2**

- Memory: 44%
- **Compute: 39%**
- Interface: 15%
- **Control: 2%**

CPU (Skylake core)

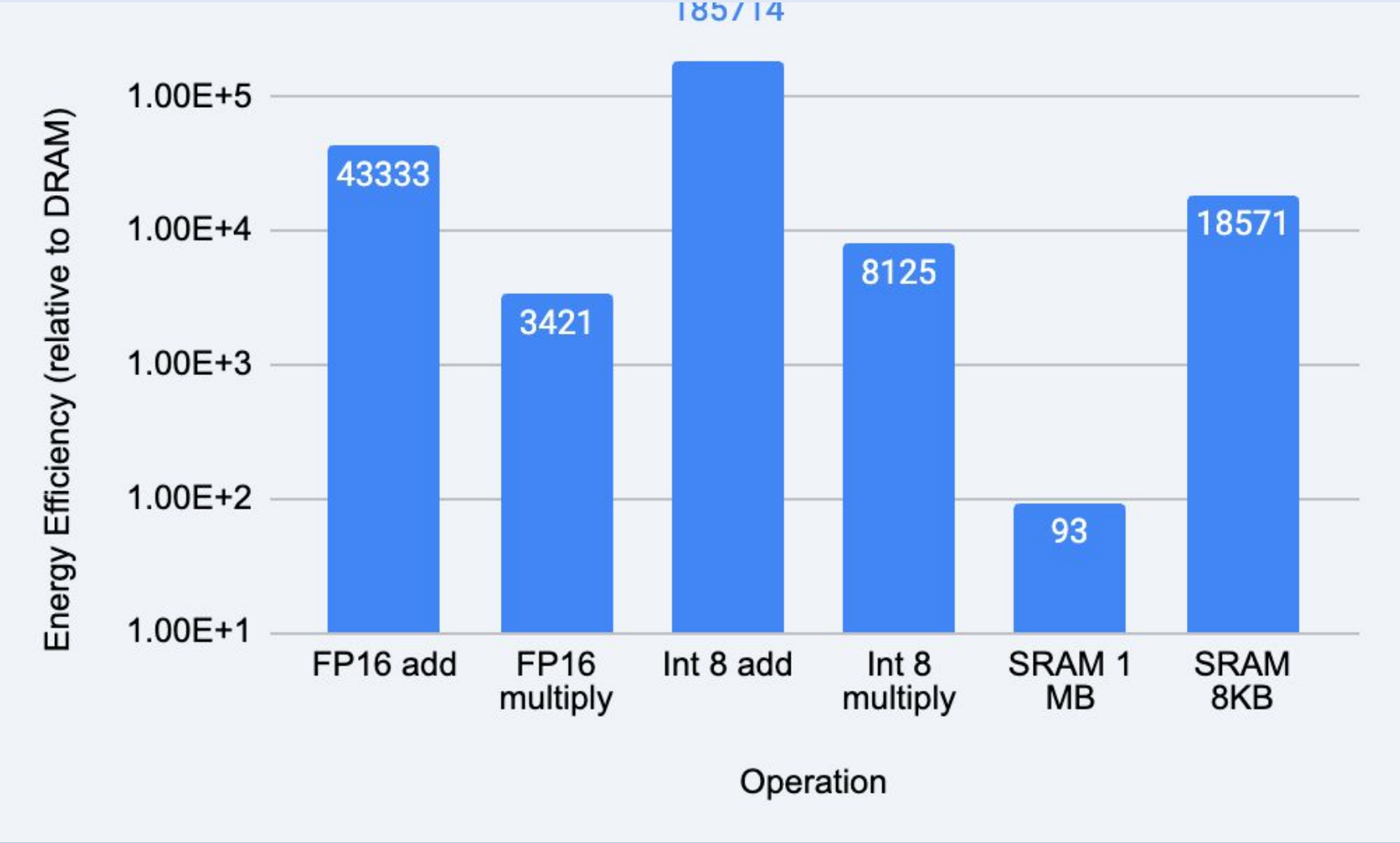
- Cache: 33%
- **Control: 30%**
- **Compute: 21%**
- Mem Man: 12%
- Misc: 4%

# Training Cost in PetaFLOPS



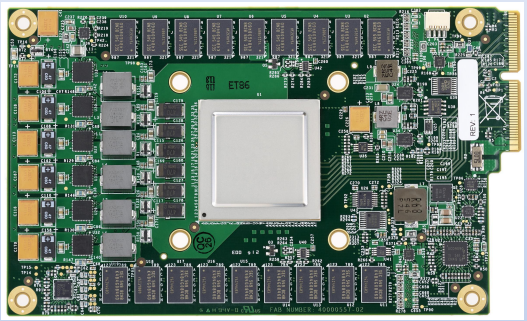

Source: Epoch AI 2025

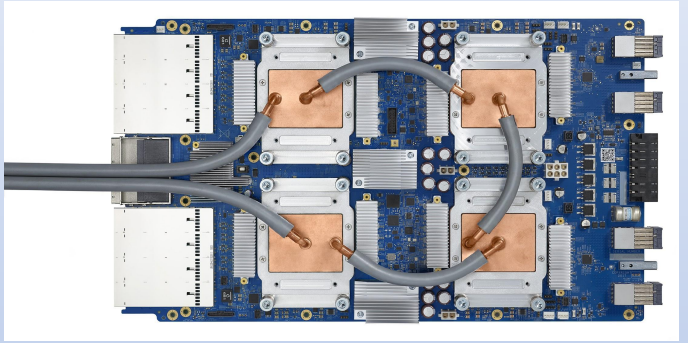
# Energy Efficiency per Operation (7 nm)

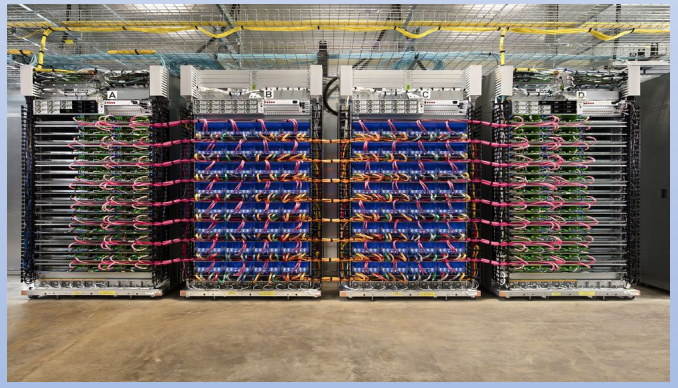


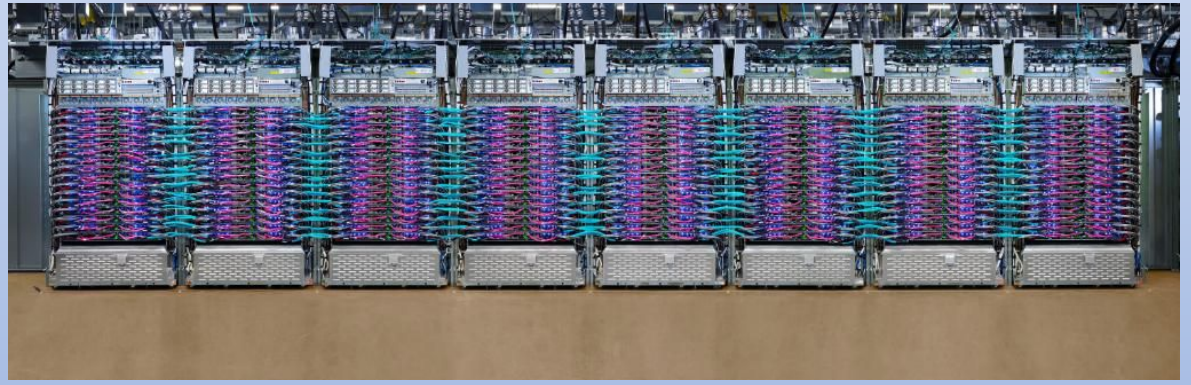

From: Jouppi 2021 updating Horowitz 2016

# **TPUv1** (2015), **TPUv2** (2017),**TPUv3** (2018)





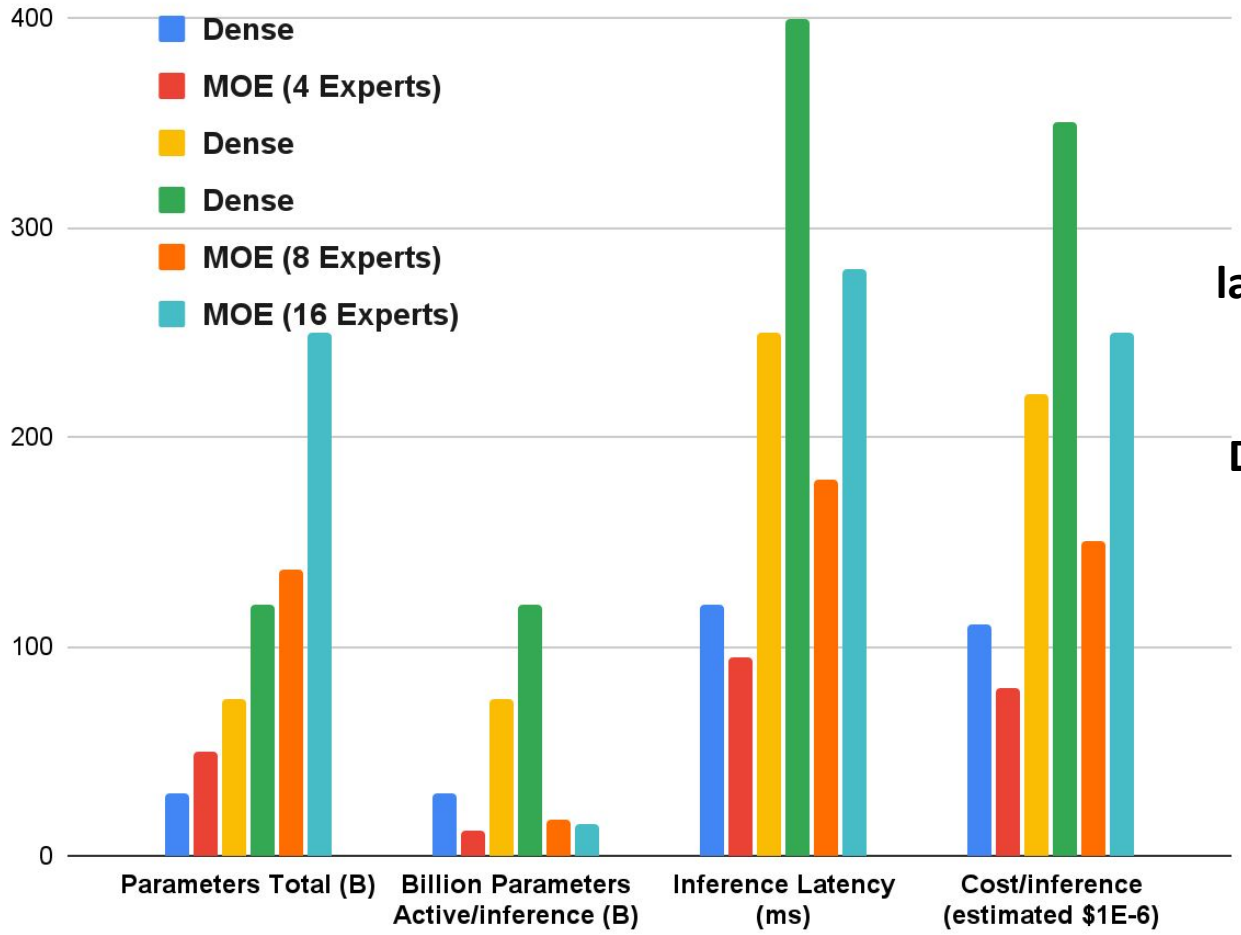




TPUv2Peak: 11 PFLOP/s

TPUv3 Peak: 100 PFLOP/s

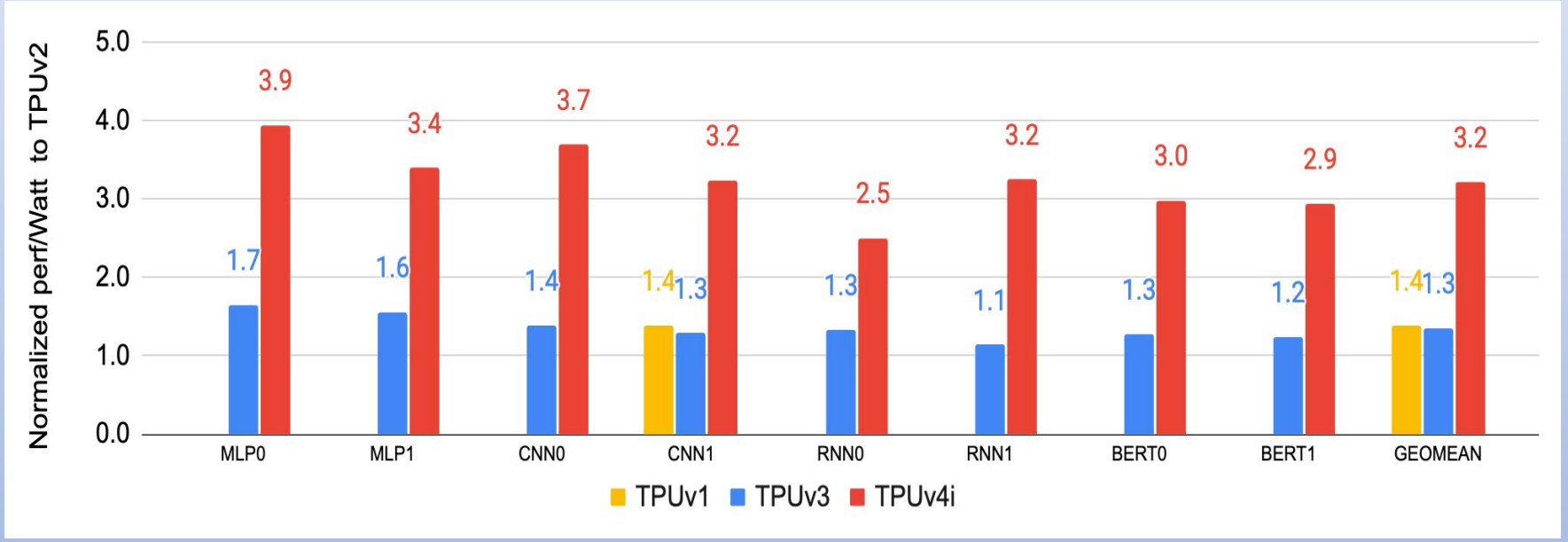
# MoE and Dense Models: Inference Cost



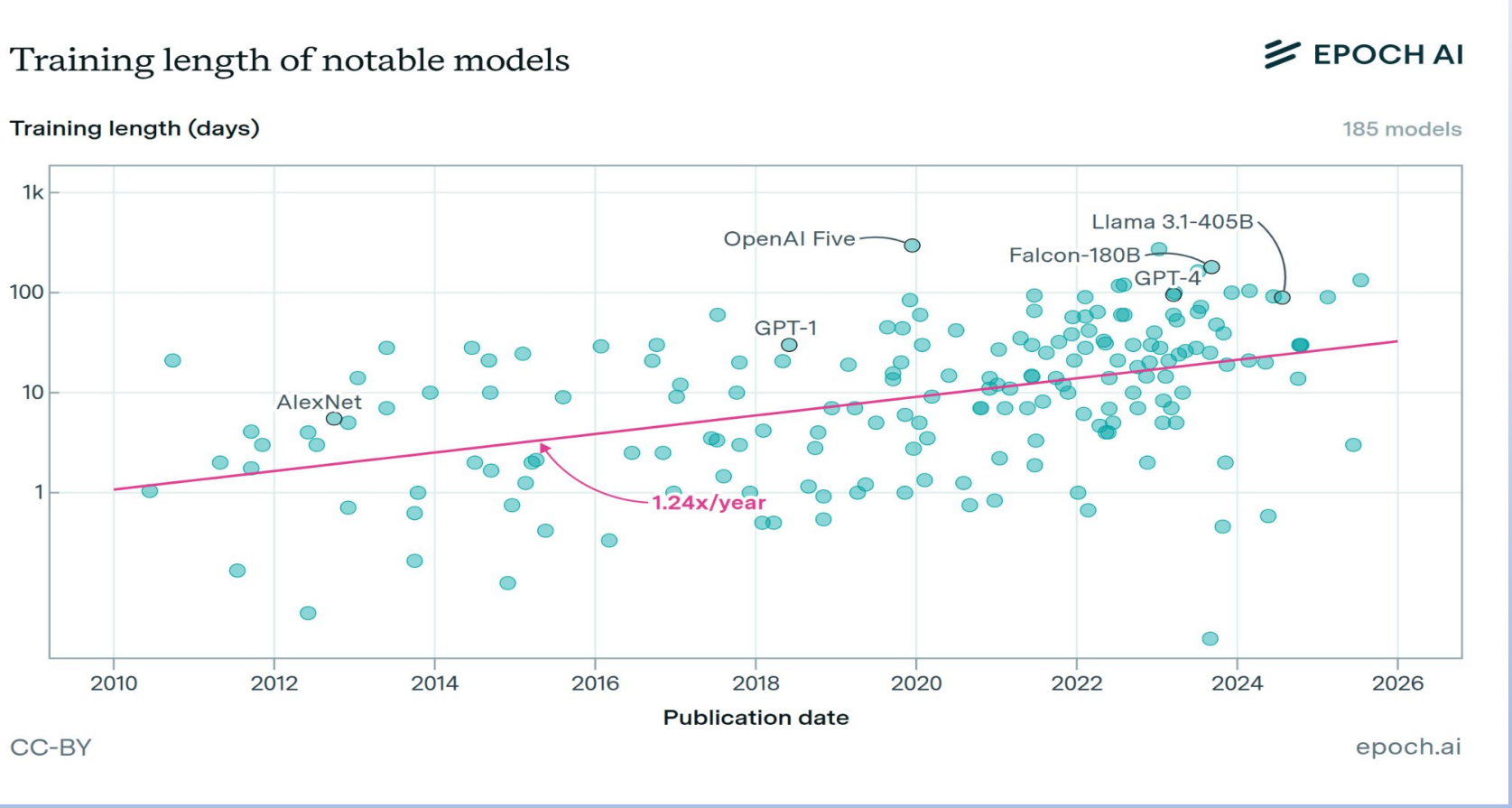

**Accuracy of a larger model; cost of smaller one!**

**Due to larger OH & sparsity, parameter reduction > latency or cost reduction.**

# Evaluation: Production Apps Perf/Watt TPUv4i vs TPUv3

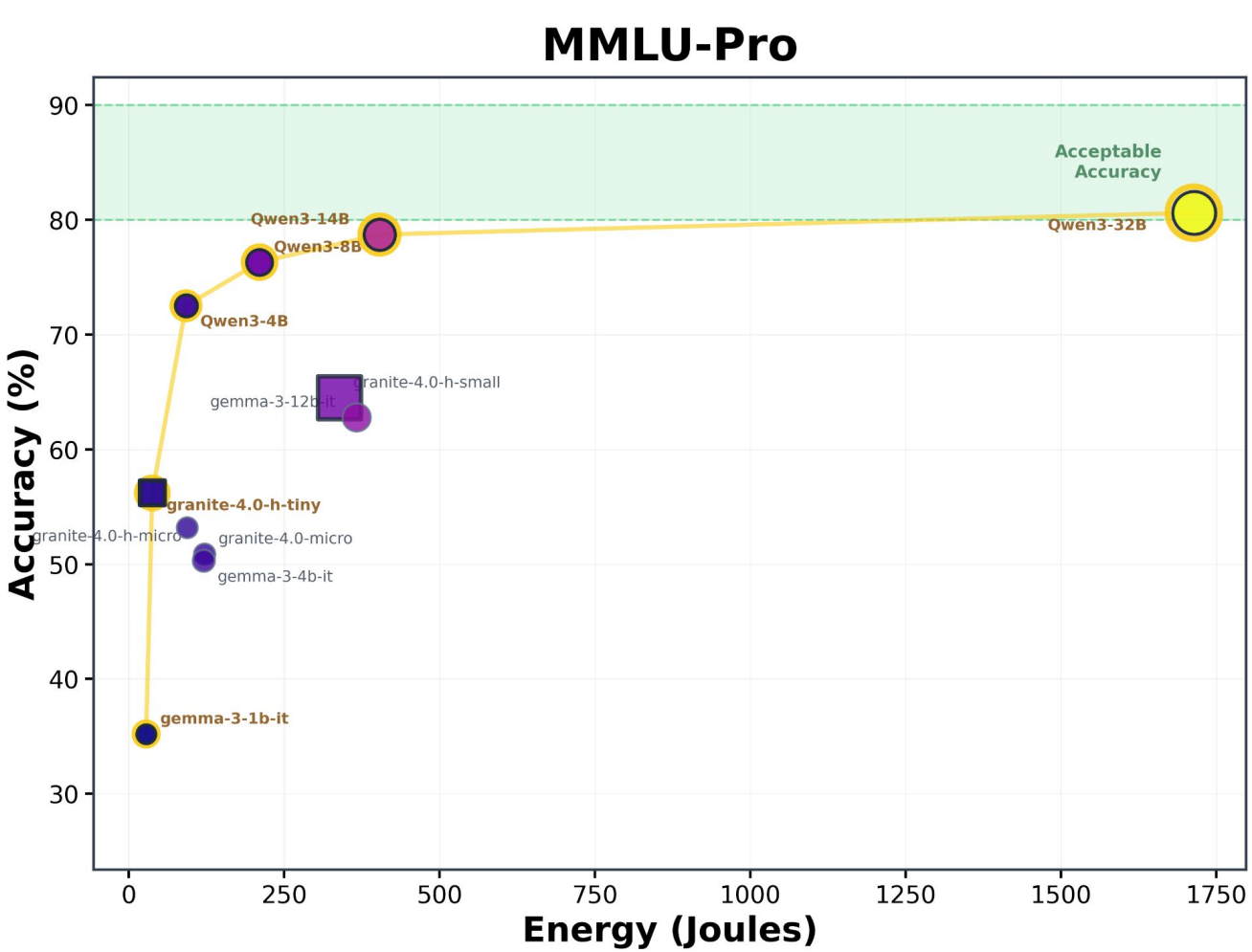


# FLOPS Needed Grow Faster than HW→Growing Training Time




To meet demand: use more TPUs/GPUs and train for longer!

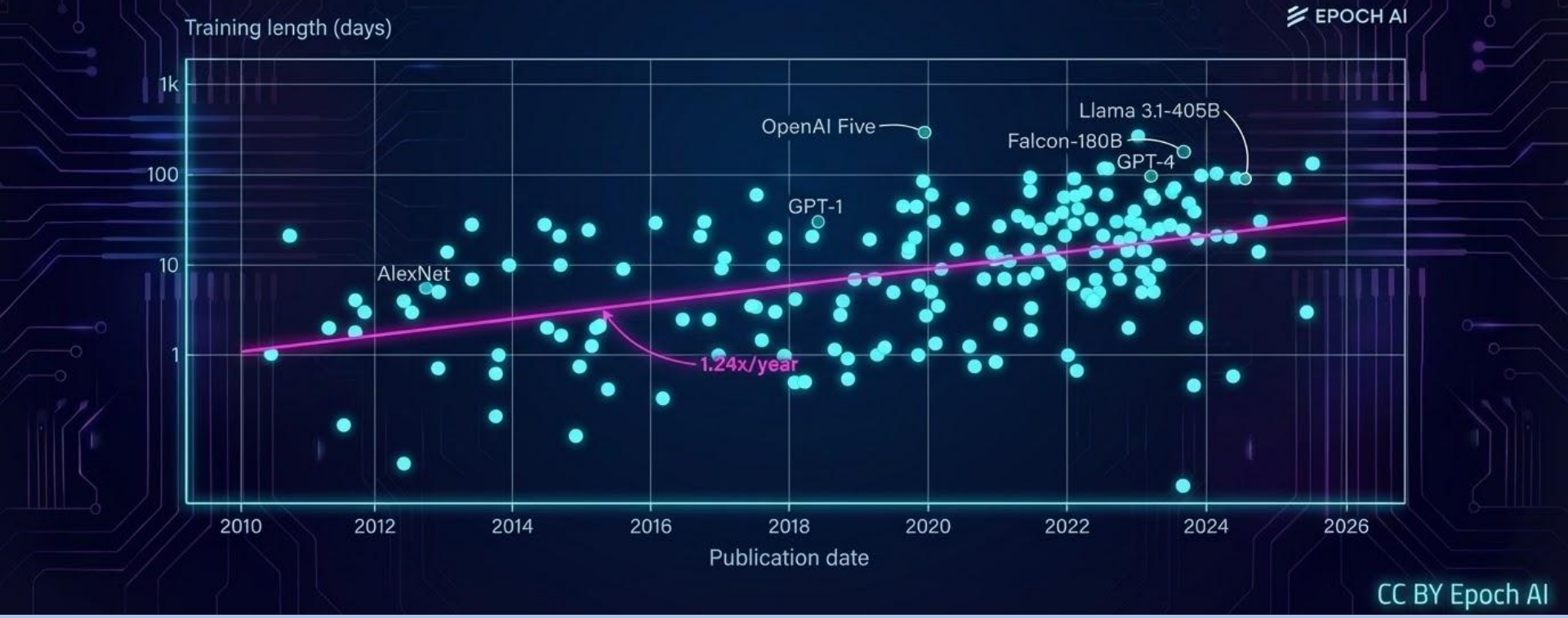
# Choose the “Right” Model: Small is Beautiful




**14B-p model accuracy =98% of 32B-p model *AND* uses only 20% of the energy!**

# FLOPS Needed Grow Faster than HW→Growing Training Time

**To meet demand: use more TPUs/GPUs and train for longer!**



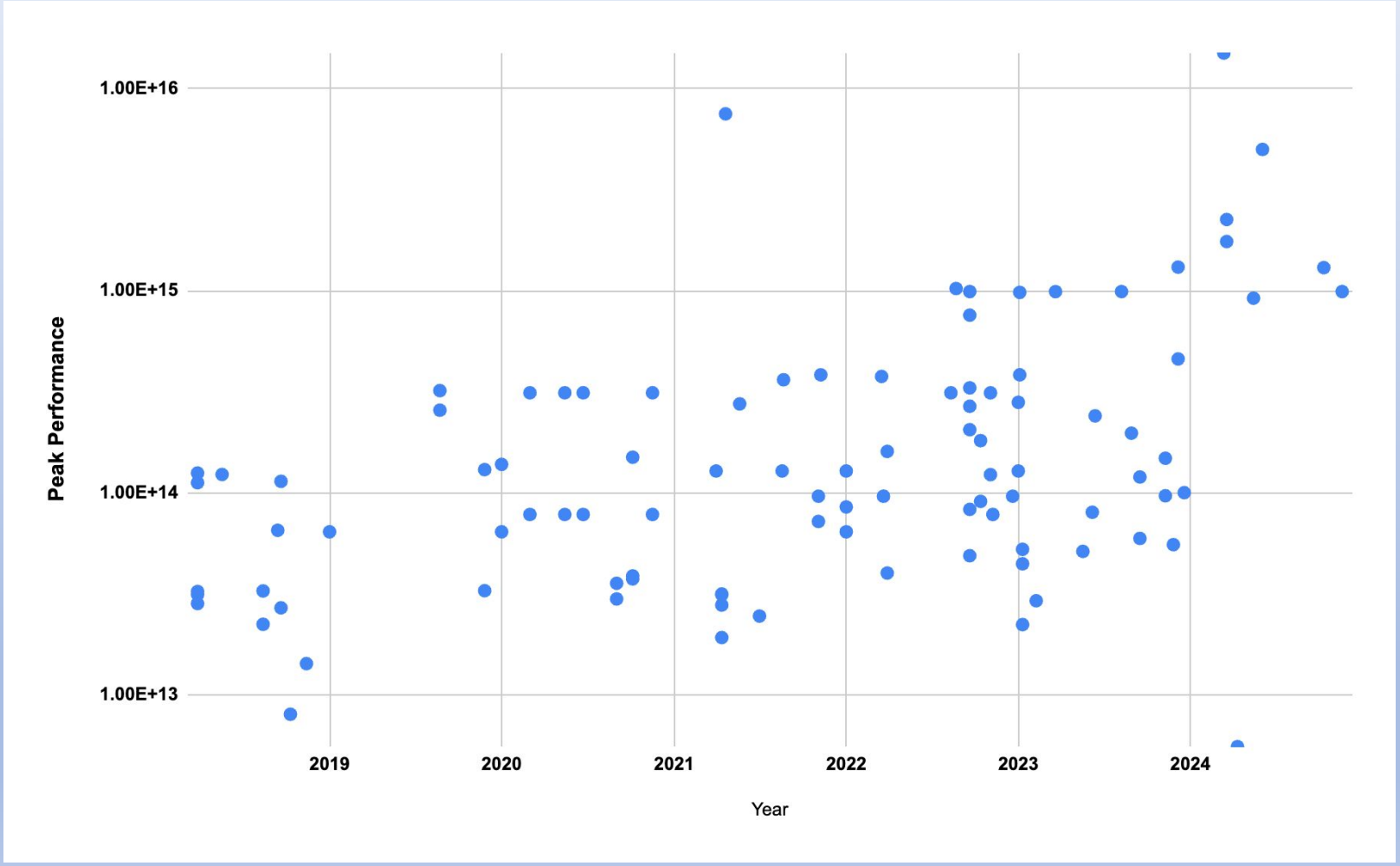
# Where does the energy go? Can DSAs do Better?

| Function | Energy in Picojoules |
|---|---|
| 8-bit add | 0.03 |
| 32-bit add | 0.1 |
| FP Multiply 16-bit | 1.1 |
| FP Multiply 32-bit | 3.7 |
| Register file access* | 6 |
| Control (per instruction, superscalar) | 20-40 |
| L1 cache access | 10 |
| L2 cache access | 20 |
| L3 cache access | 100 |
| Off-chip DRAM access | 1,300-2,600 |

* Increasing the size or number of ports, increases energy roughly proportionally.

From Horowitz [2016].

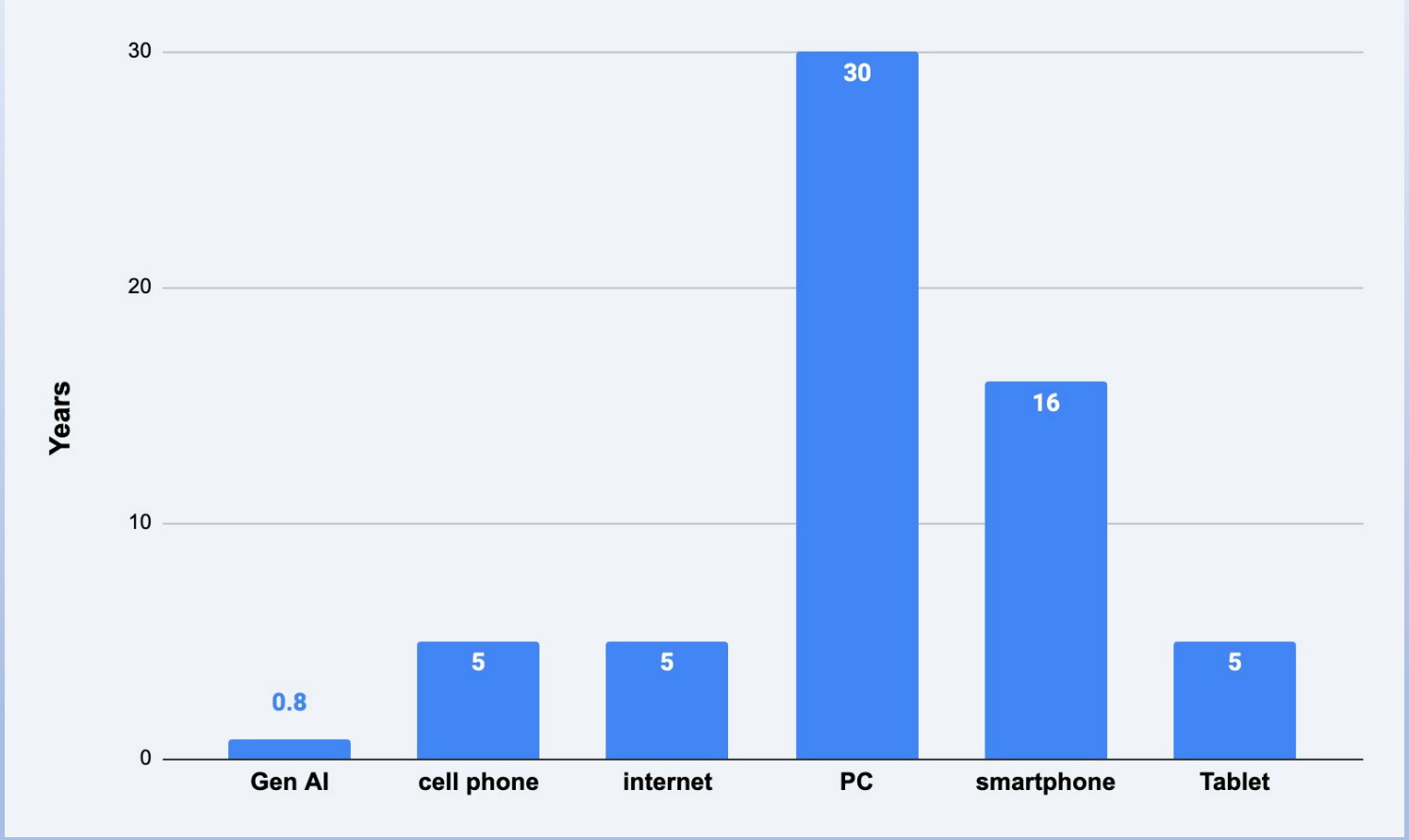
# Peak Performance 16-bit FP




Source: Epoch AI 2025

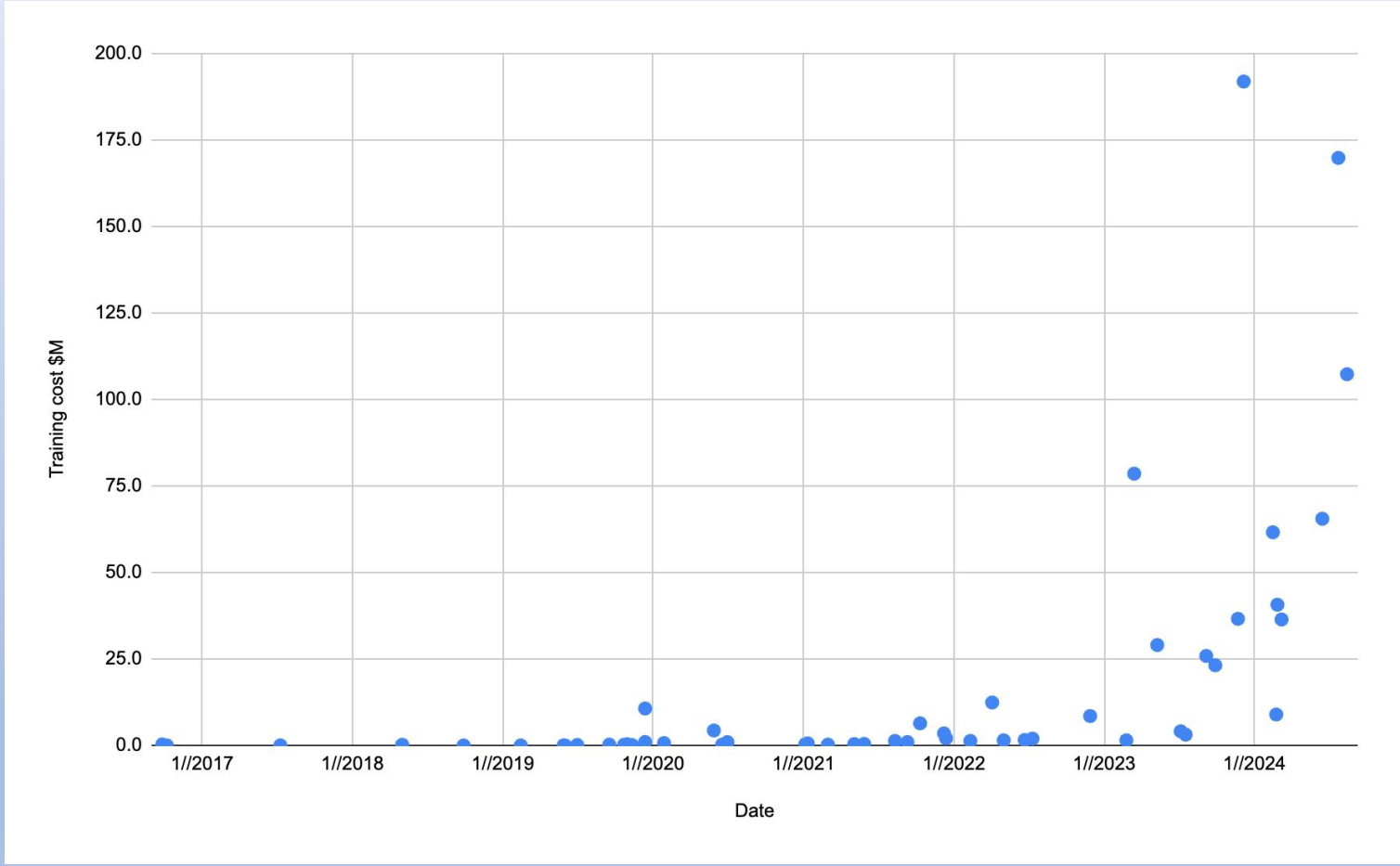
# Time to Adopt Technologies of Last 50 years

**Years to adoption by 50% of US households**




Data source: National Bureau of Economic Research

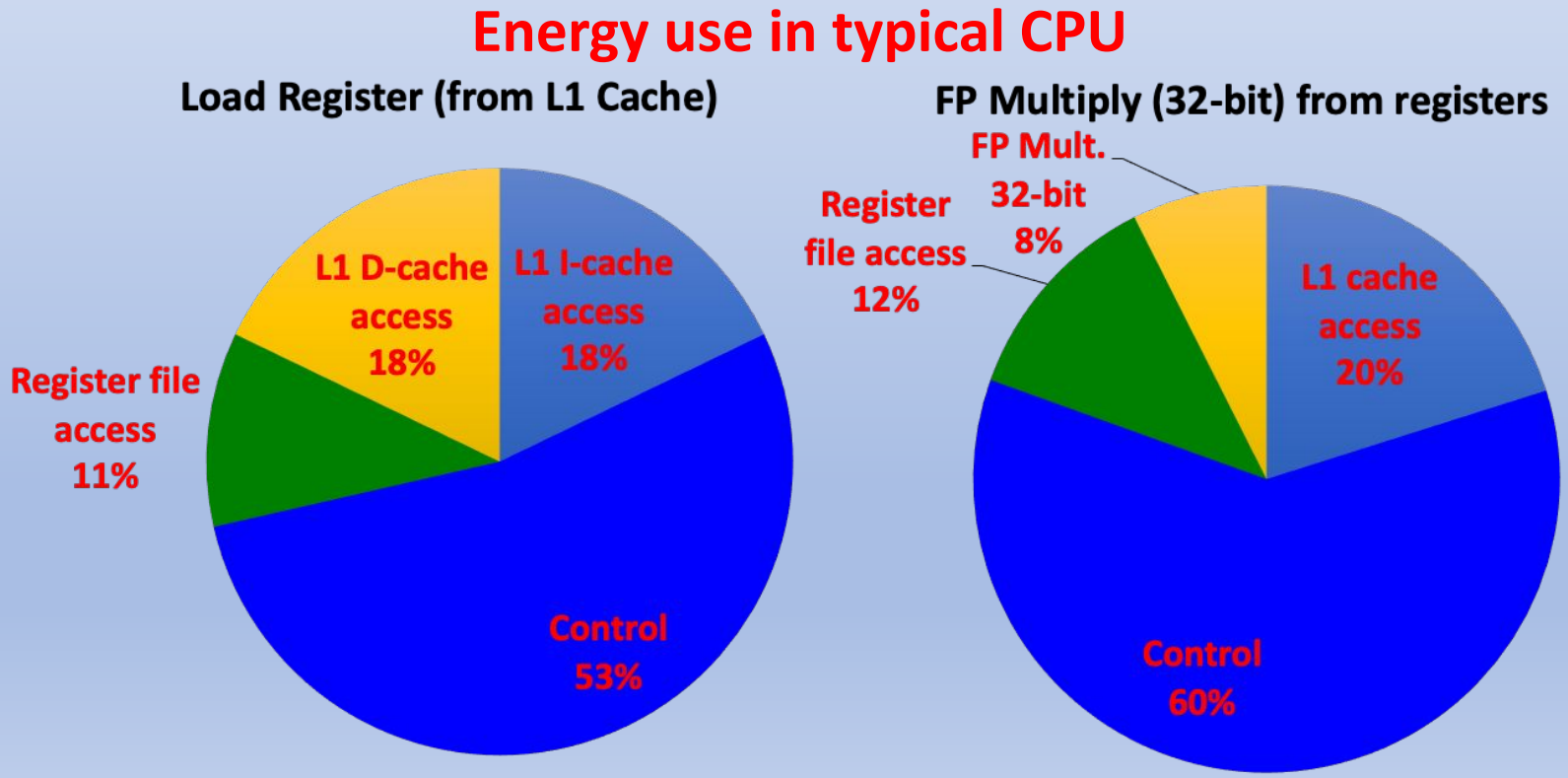
# Training Cost Rising Quickly (estimates)




Stanford HAI: AI Index 2025

# Why Domain Specific Architectures Won
# Tailor the Architecture to the Domain

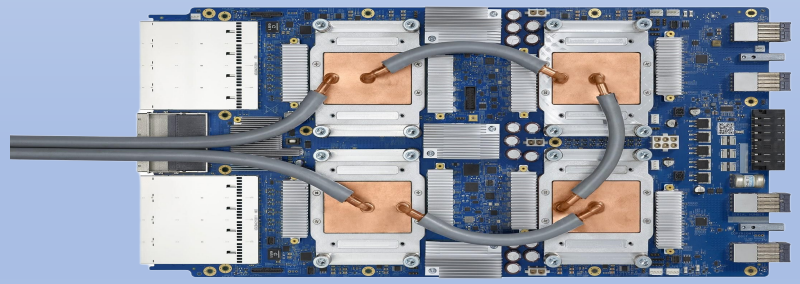
- More efficient parallelism: SIMD vs. MIMD: ↑FLOPs/memory access
- Less control overhead: VLIW/vector vs. Speculative



# Computational Challenges on the Road to AGI

John Hennessy
Stanford University

January 2026

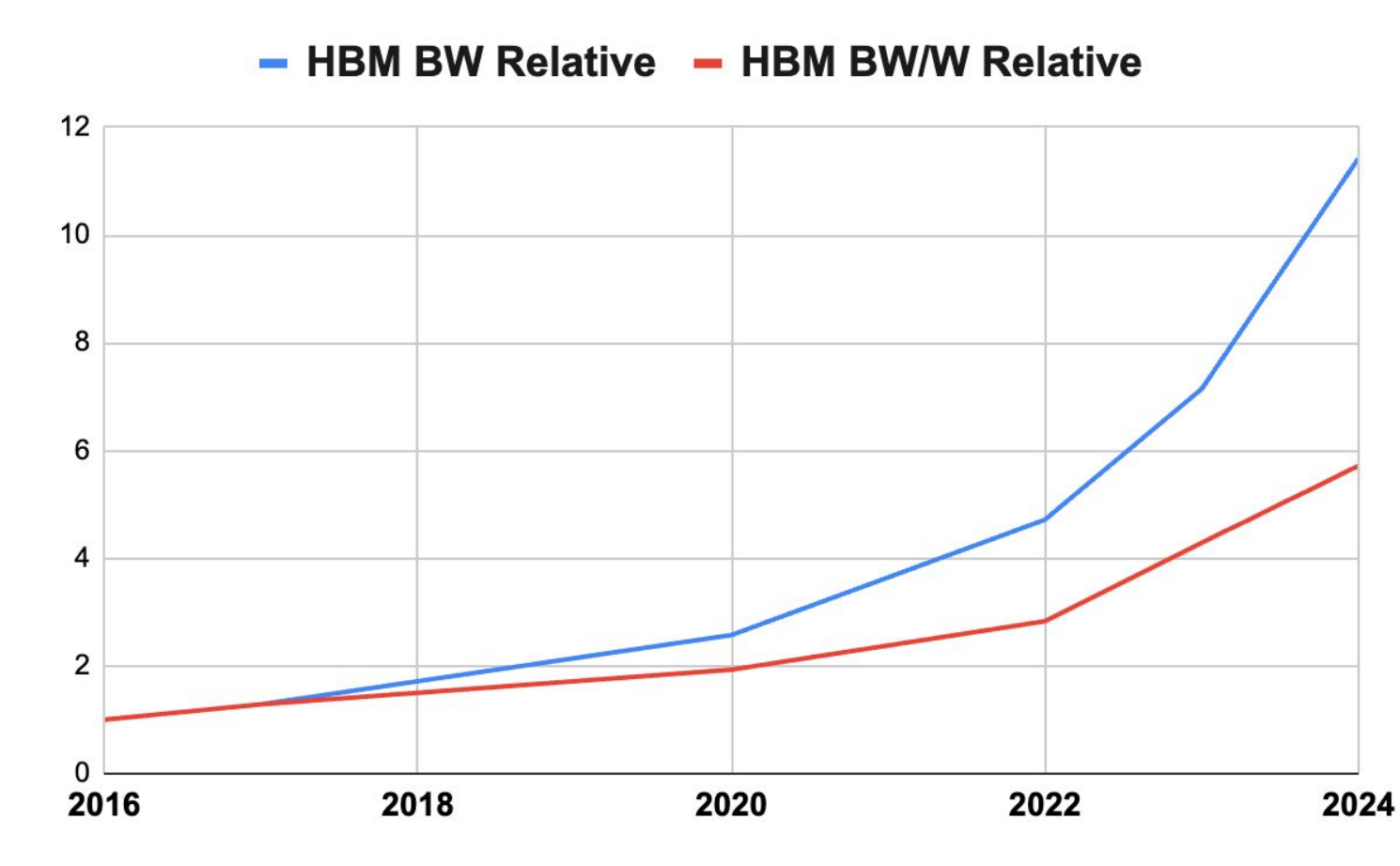


# Time to Adopt Technologies of Last 50 years

Years to adoption by 50% of US households

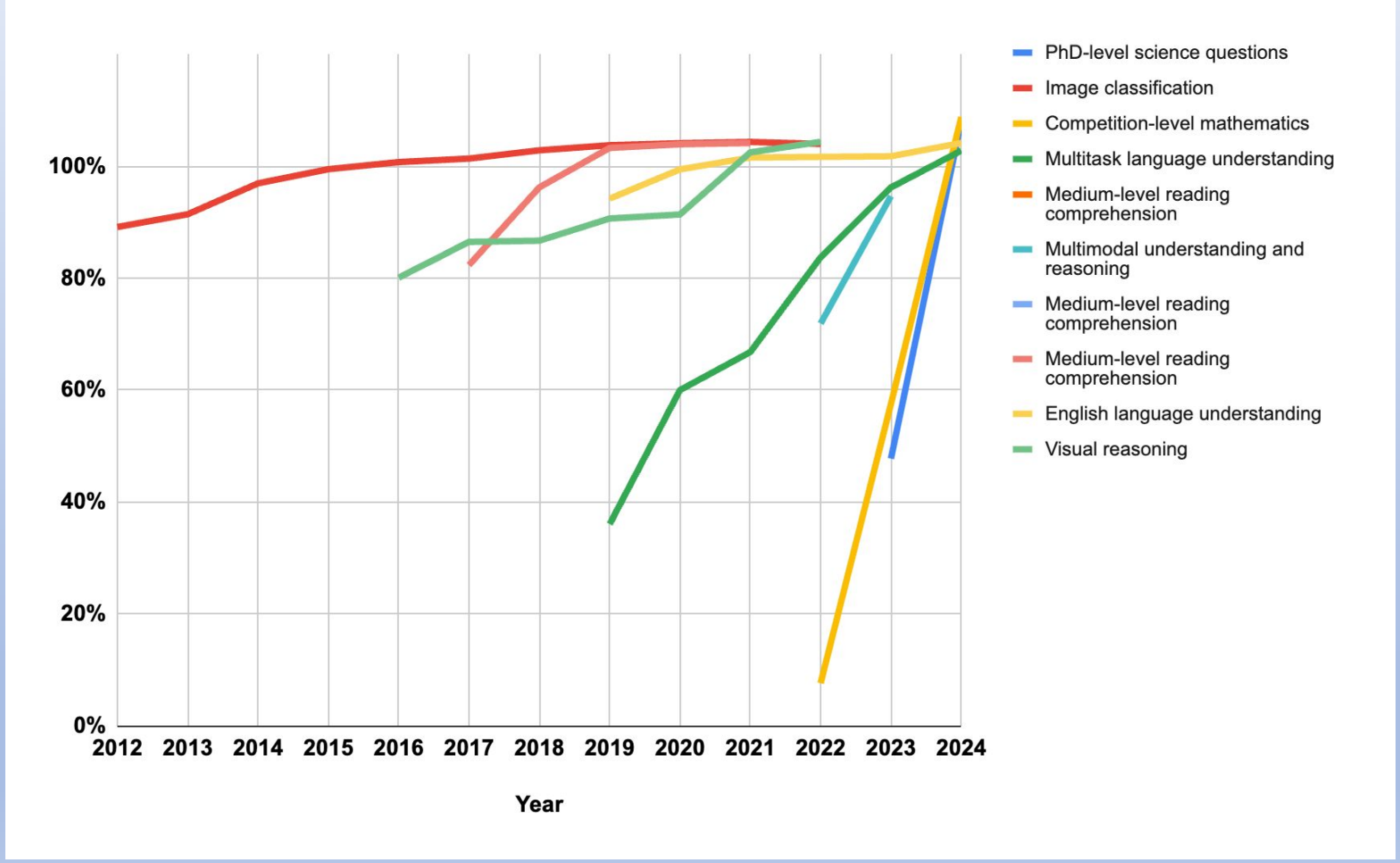

Data source: National Bureau of Economic Research

# Peak HBM BW and HBM BW/W



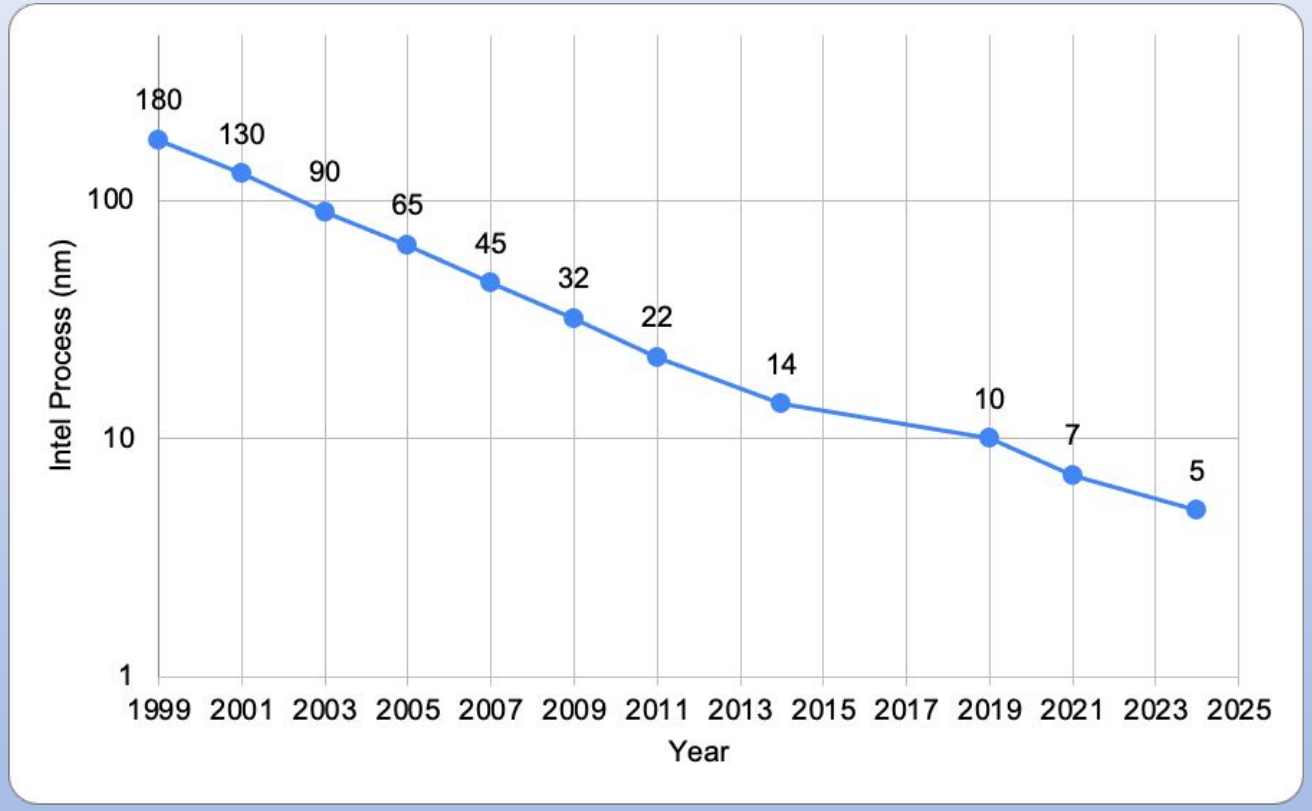

**Massive parallelism hides memory latency–not BW limits.**

# Progress on Benchmarks (100% = Human Level)




Source: AI Index, 2025 | Chart: 2025 AI Index report

# Moore's Law–time between nodes is increasing

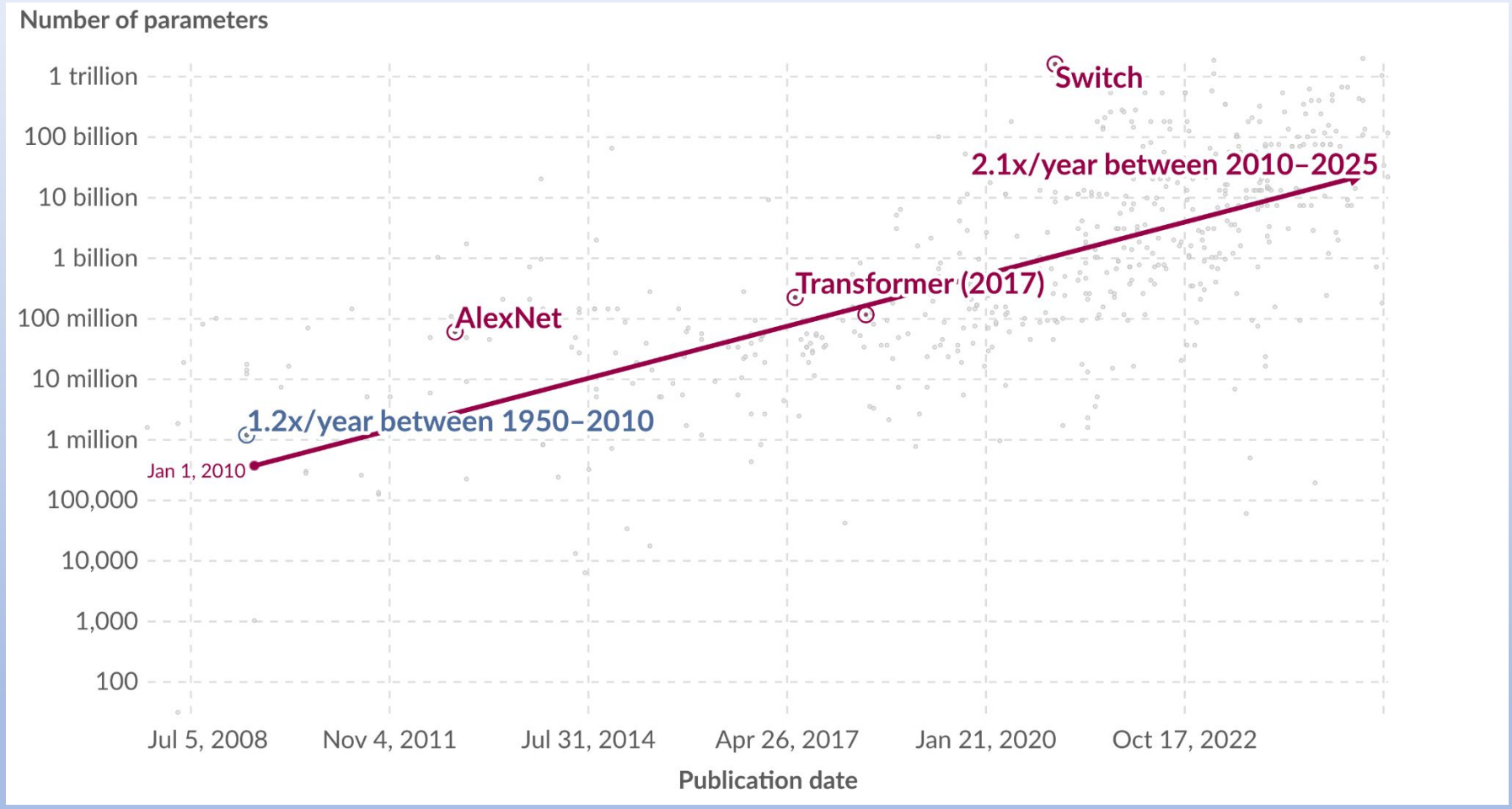



**International Roadmap for Semiconductors**

Widely accepted plan for developing ICs

- Lasted 25 years.
- Ended 10 years ago: too hard!

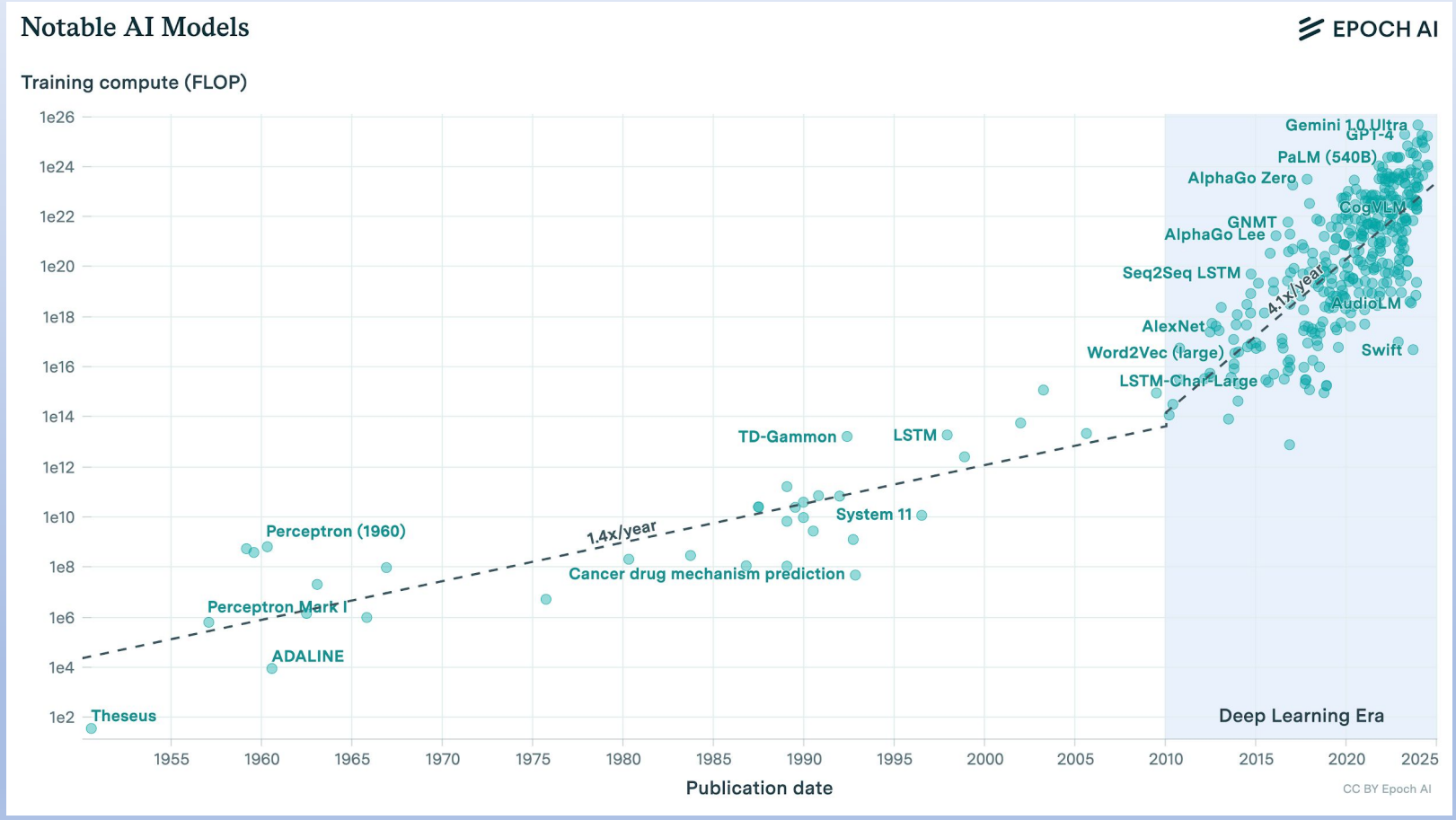
# Parameter Growth Rate




**Simply scaling existing models is unlikely to achieve AGI.**

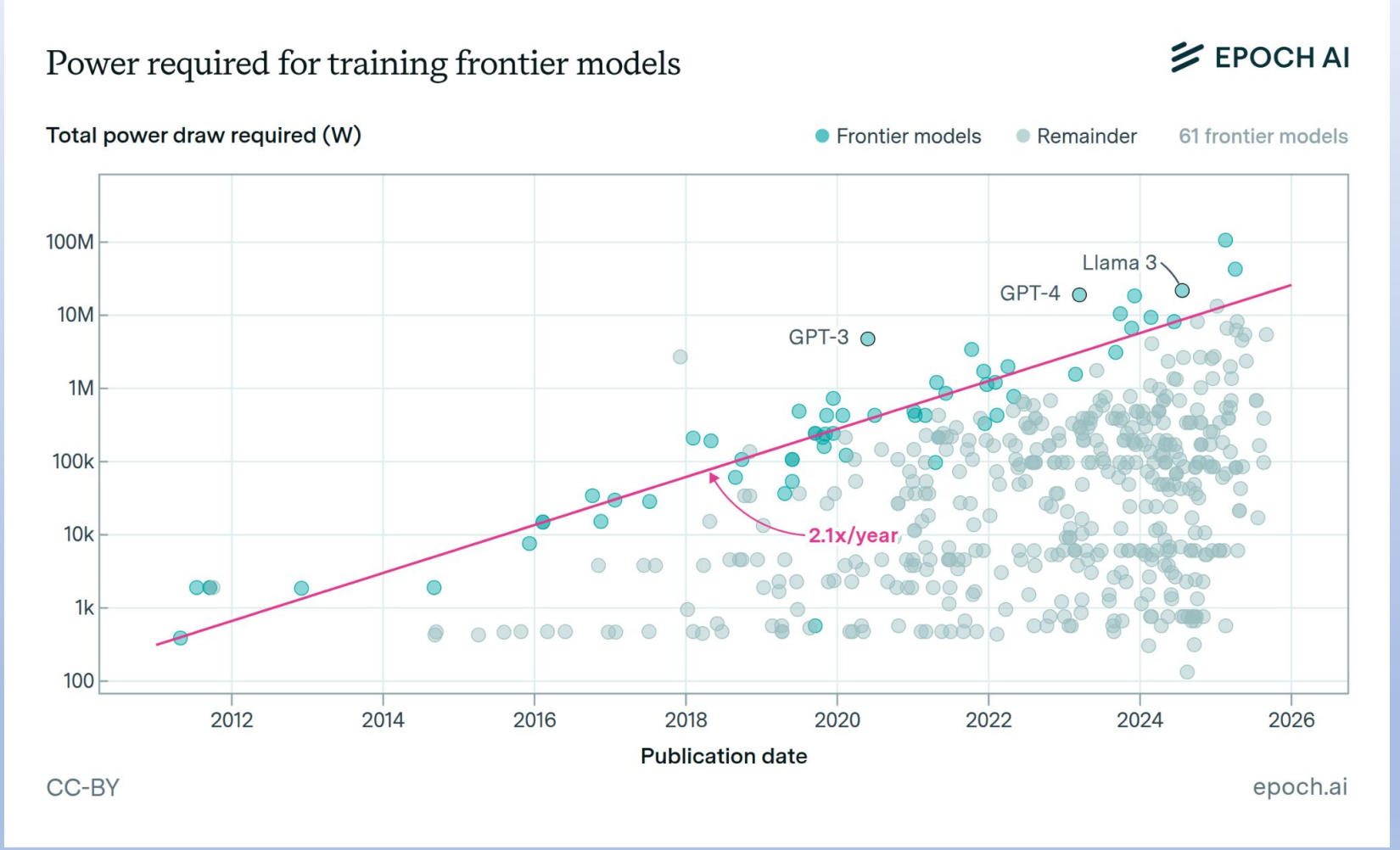
Source: Epoch AI 2025

# Hence: Computational Demand Growing Fast!



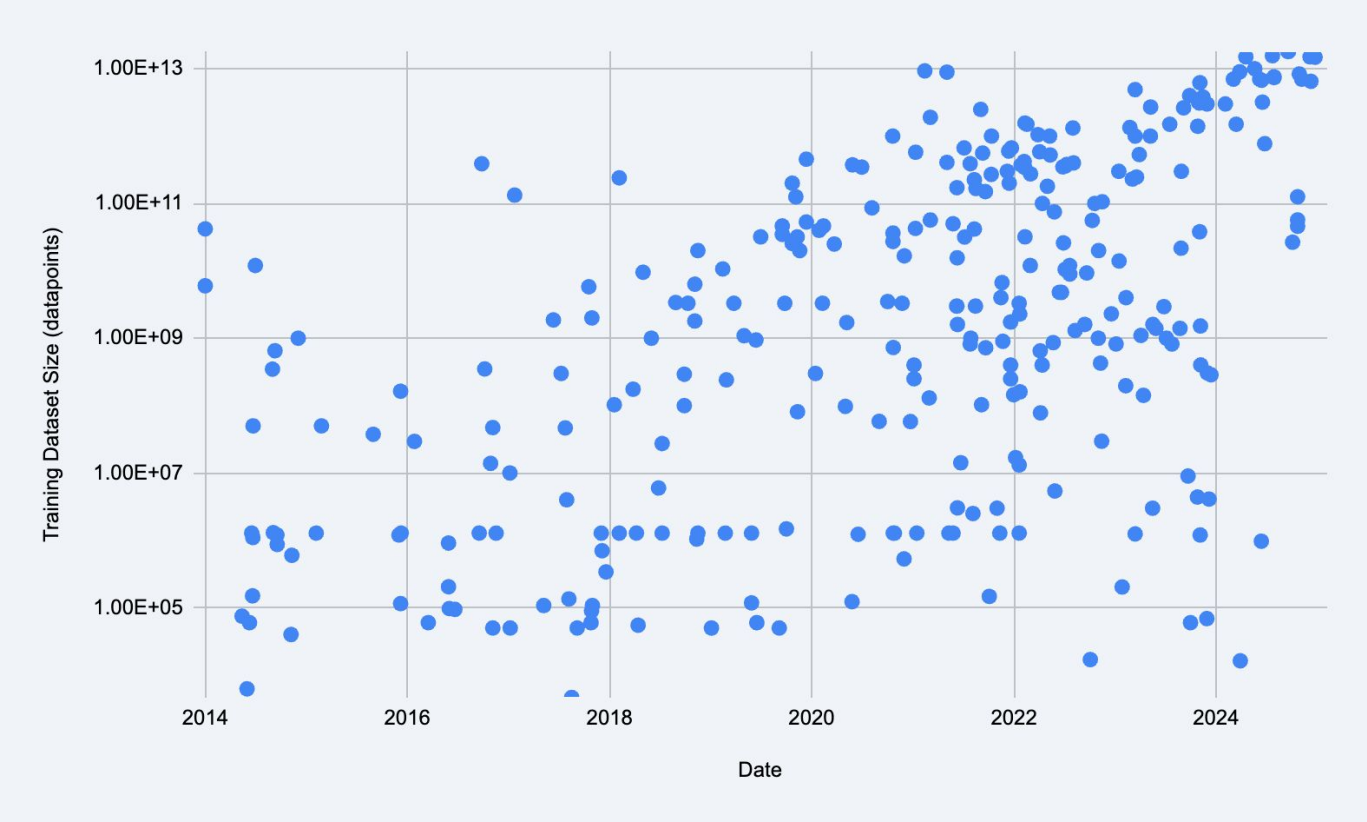

**4x annual growth is ≅2x growth rate of GPU!**

# More G/TPUs + More Time = More Power + More $



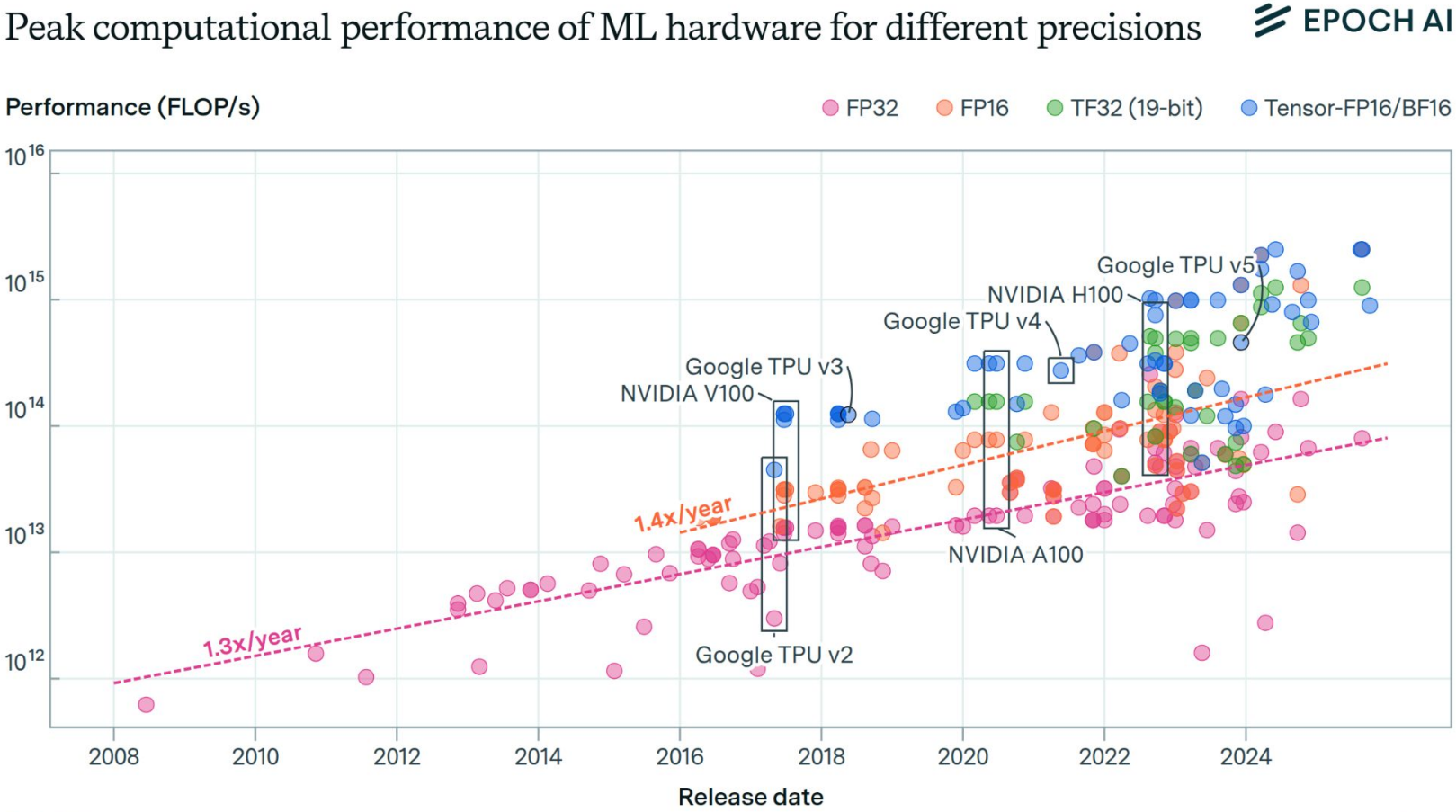

**Inference Power/Cost Could Swamp Training Power/Cost by 2026!**

# Training Data Sets Growing Quickly

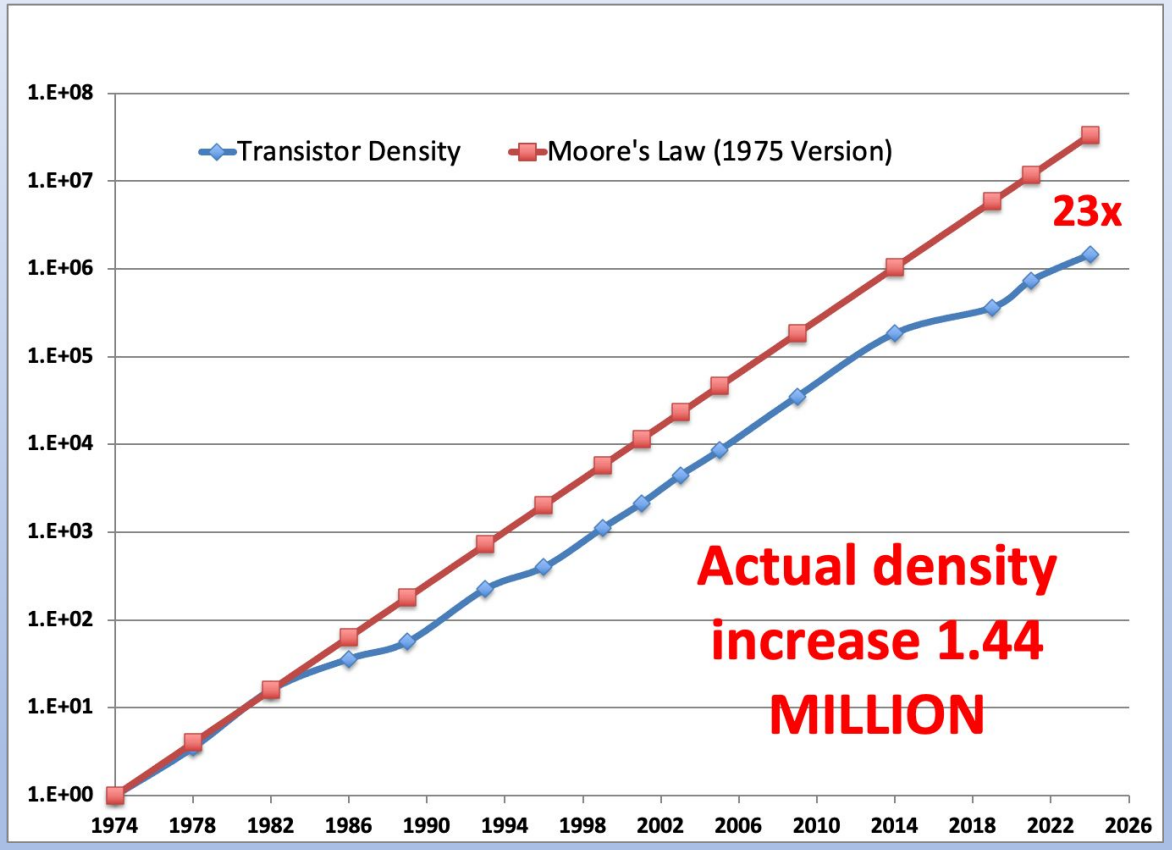



**Nonetheless, it is likely that both models and (hence) training data will continue to grow rapidly.**

# Peak Single-Chip Performance: Moore's Law Rate

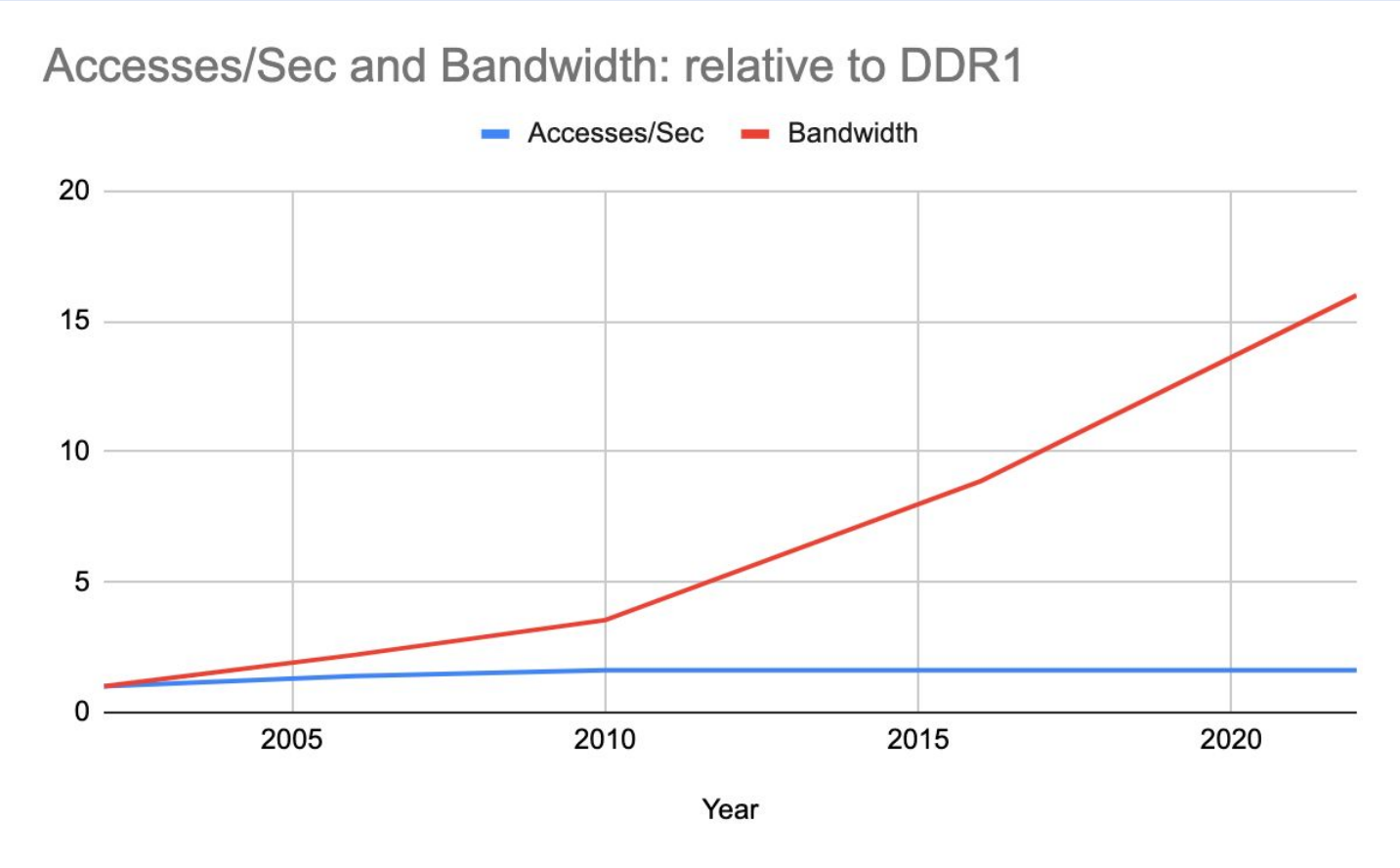


# Moore's Law: slowdown (Post Moore's Law Era)




**Gap is likely to continue to increase, particularly for memory technologies.**

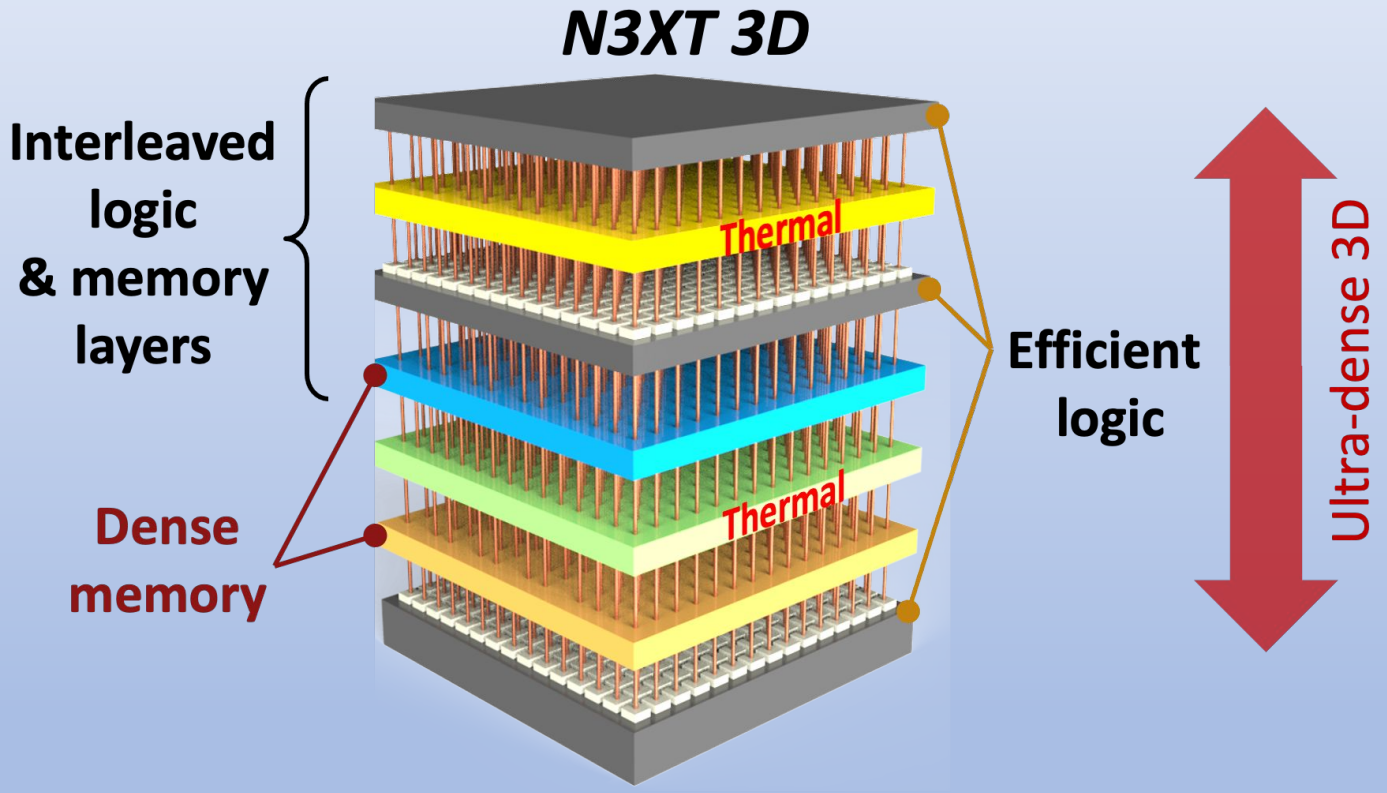
# DRAM Scaling Access Time vs. Bandwidth




**Need "unpredicted" DRAM accesses to near 0%.**

**HBM helps bandwidth–does nothing for latency!**

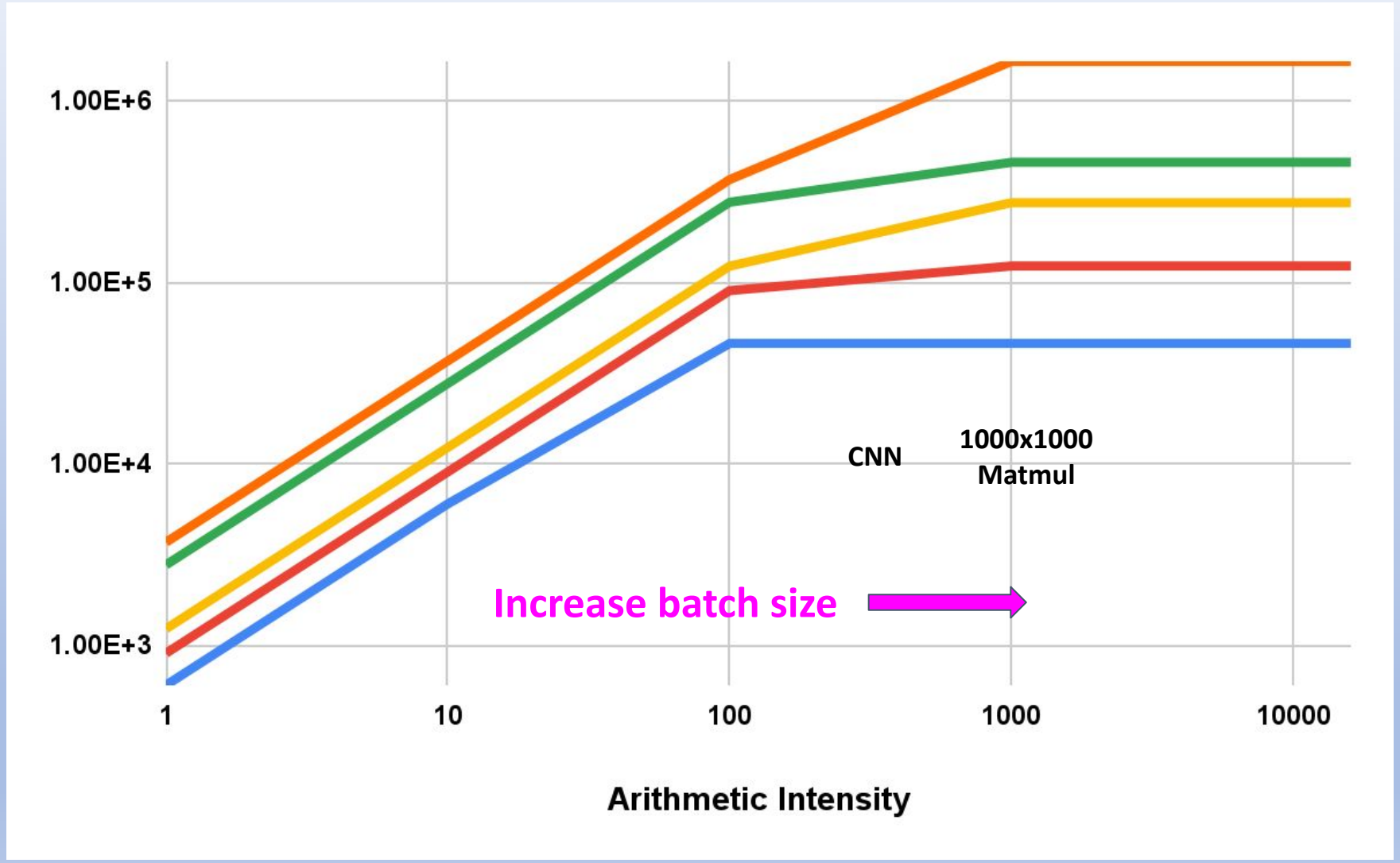
# Future: High Density Stacking (Real 3D)



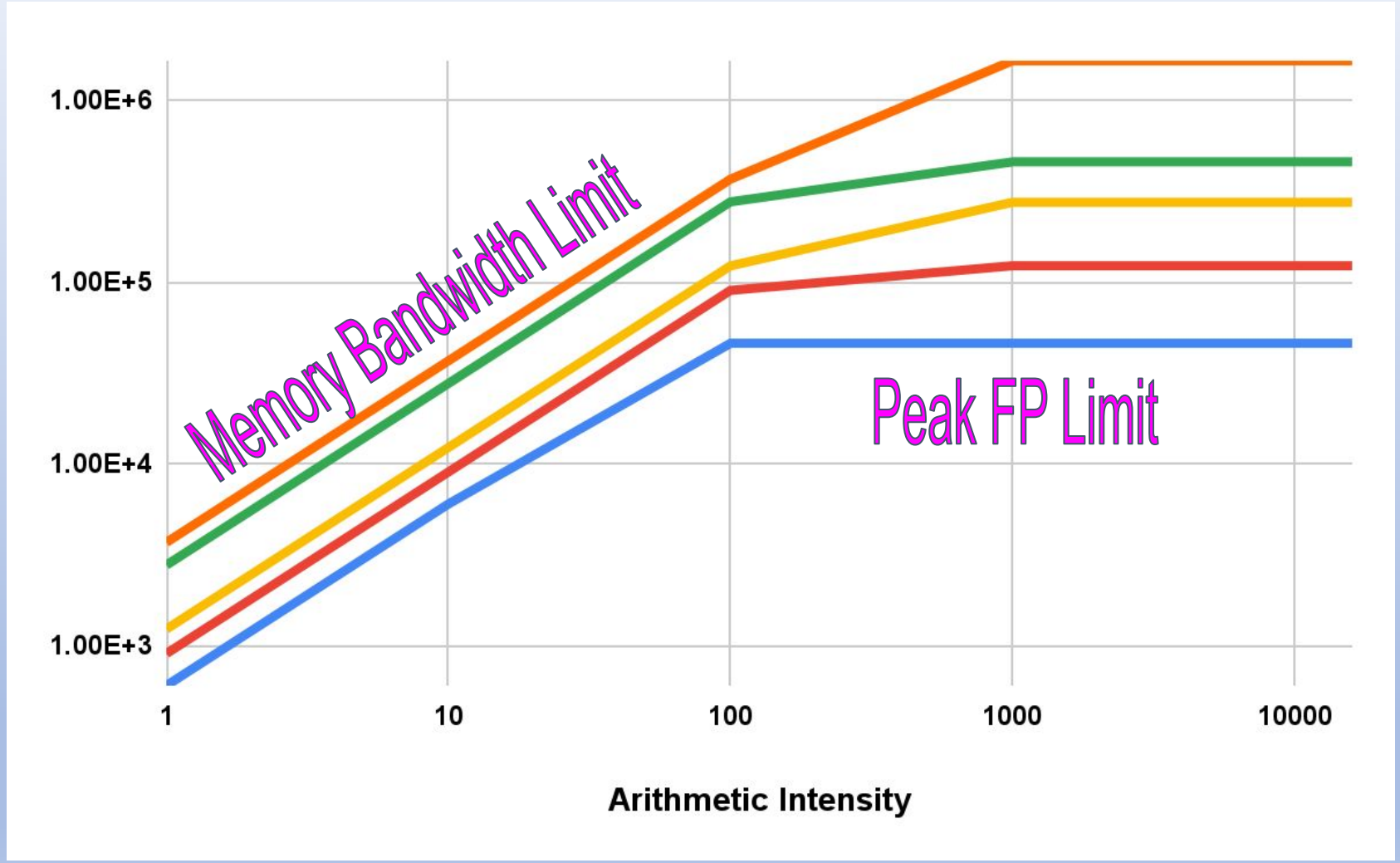

**Challenges: Cooling, Interconnect density, Realiability/Yield**

T. Srimani … S. Mitra, "N3XT 3D Technology…," *IEDM*, paper 39.2, 2023.

# TPU Roofline Plots: Arithmetic Intensity

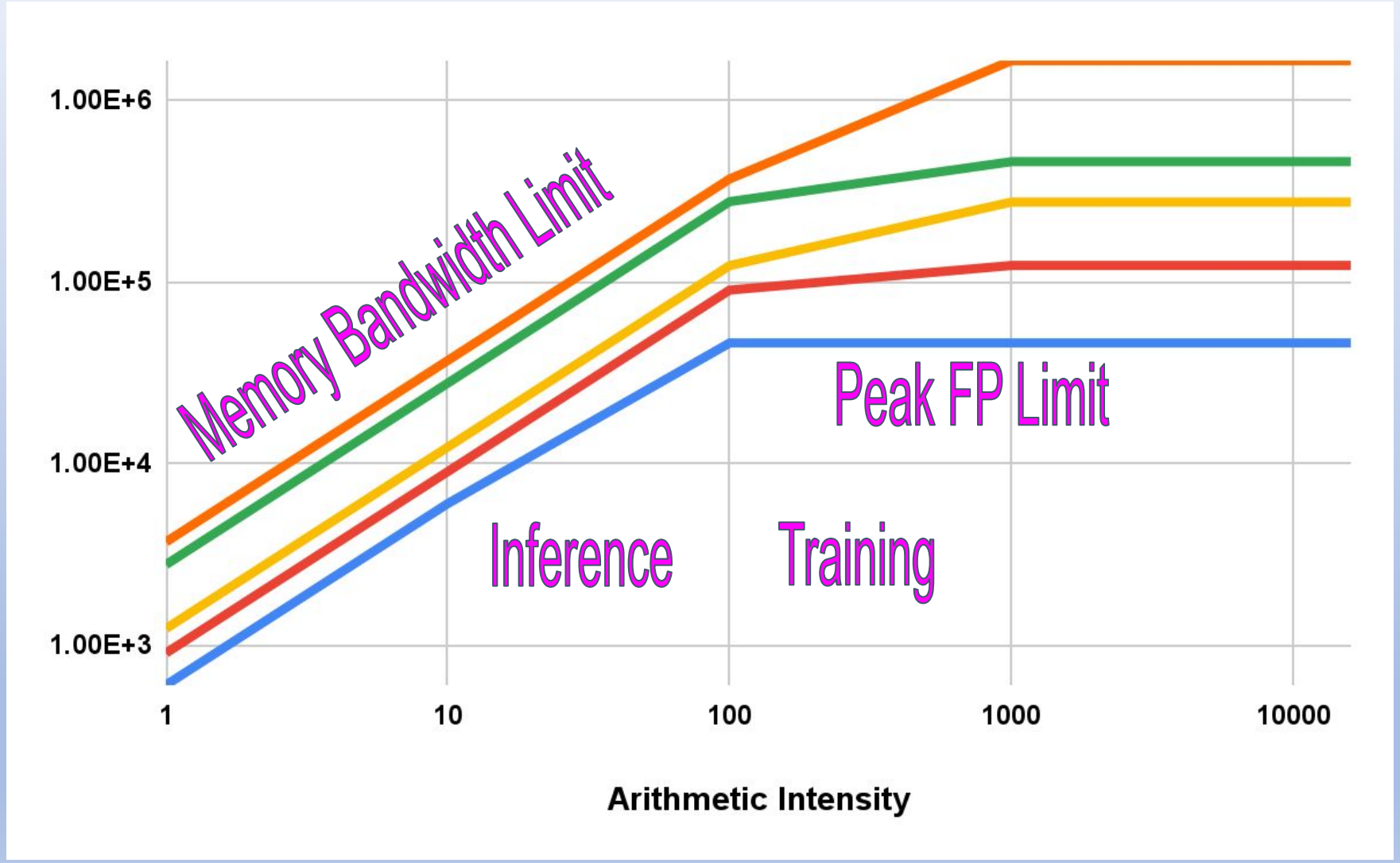


# TPU Roofline Plots

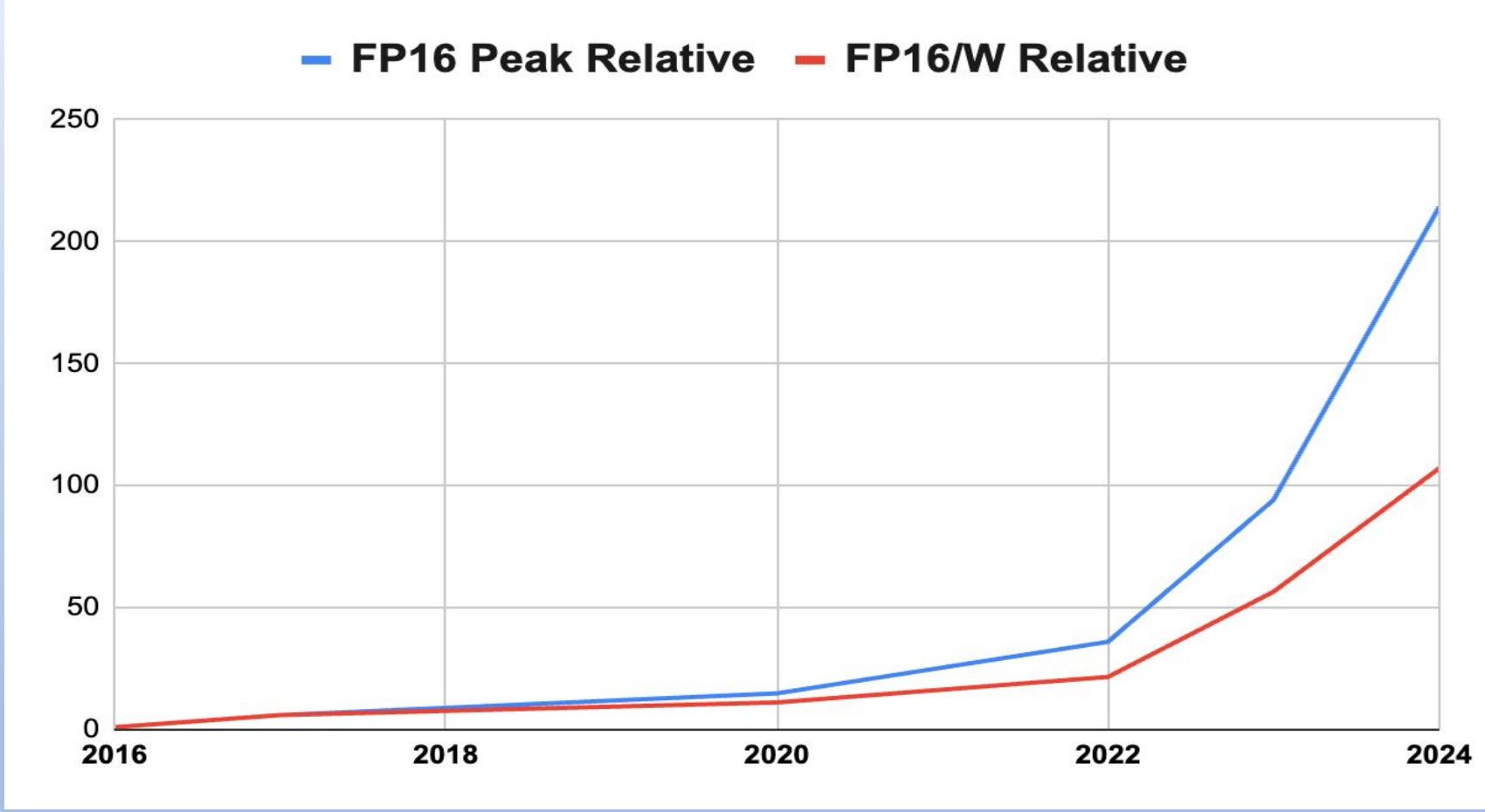



**Arithmetic intensity = FLOPs per byte of memory.**

# TPU Roofline Plots



# Peak GPU FP Performance and Performance/W




**FLOPS are easiest feature still scale, but lags in FLOPS/W!**

# Concluding Thoughts: Deja Vu

- In 2012±, end of Moore's Law & Dennard scaling→flatlined CPUs
- First decade of DSAs is near its end
  - Easy gains largely harvested
- Need new algorithms + new models + new architectures + enhanced technology.
  - Full stack design!