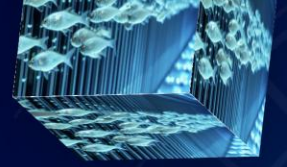
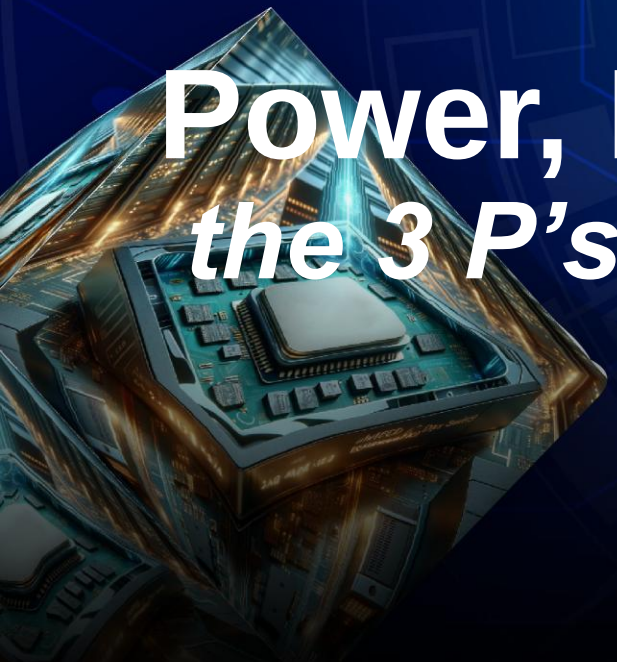




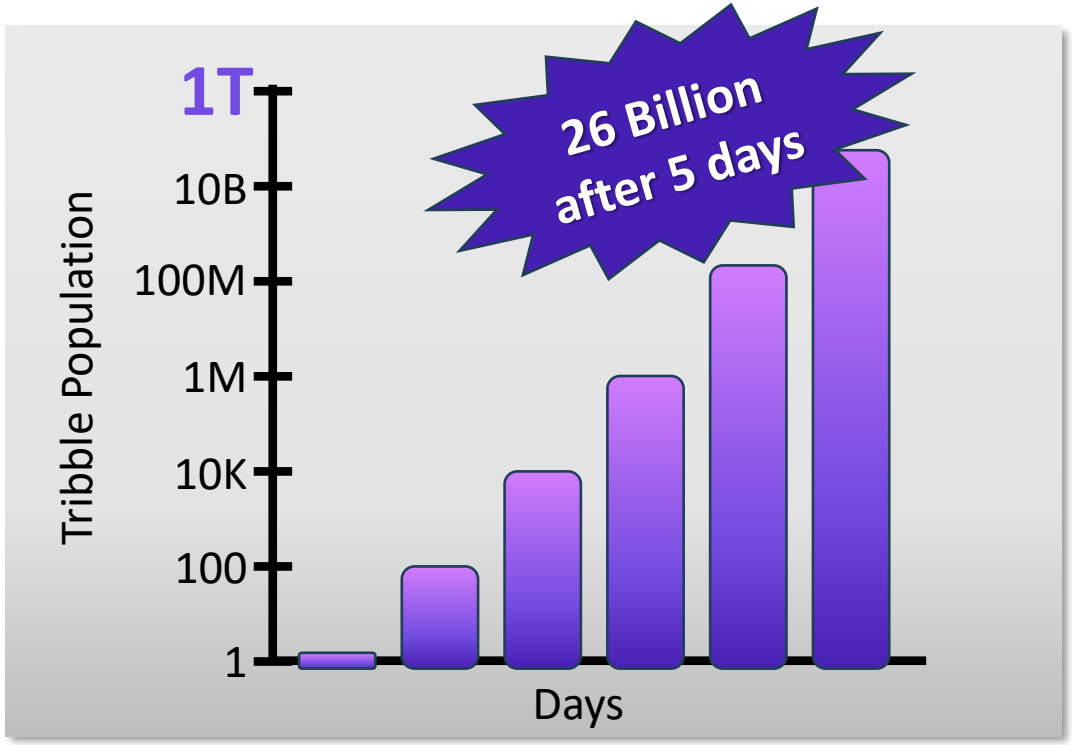
Power, Performance, and Packaging
the 3 P's Driving Infrastructure & Progress

March 2025

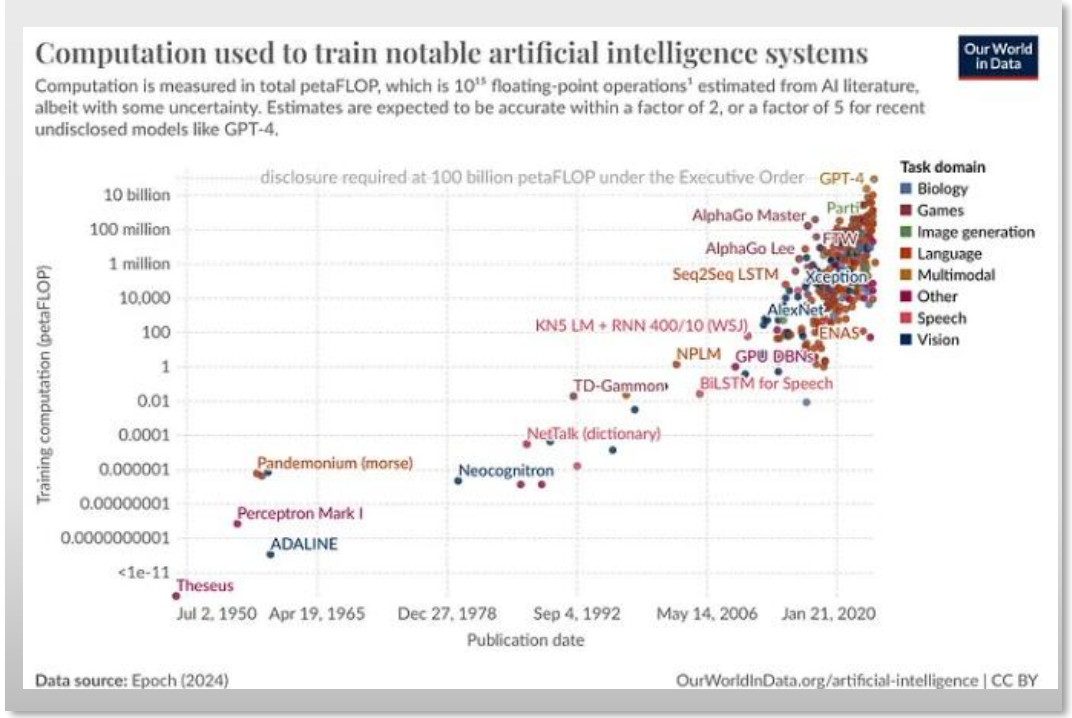
Everything we need to know can be learned from Star Trek

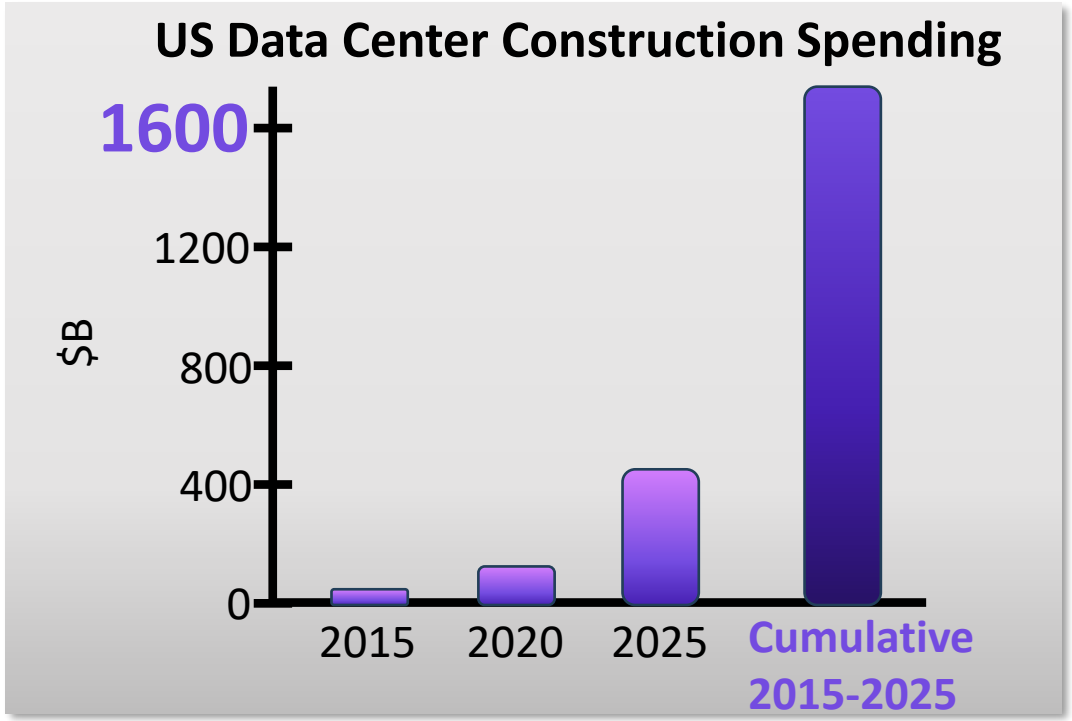
(the original series)

The Trouble with Tribbles (S2:E15)



The Trouble with Teraflops
(or Peta or Exa)





60 years to get to 1 petaflop pre-Alexnet

10 years to go from 1 to ***>10Billion***

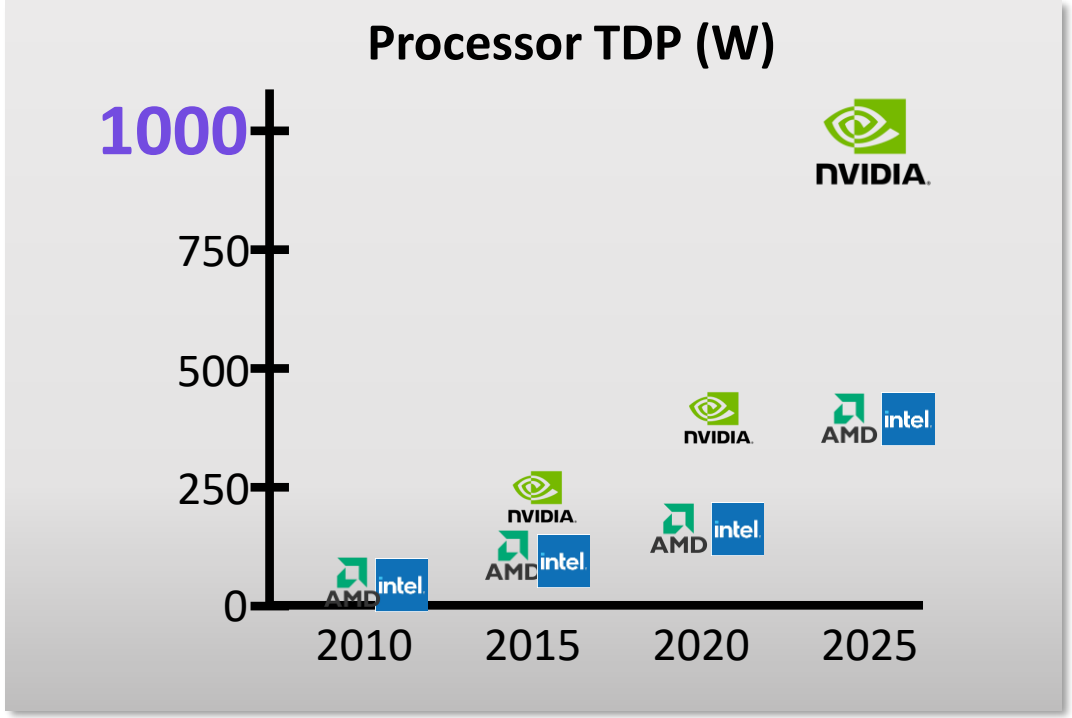
From $33B to $420B in 10 years

Over 50x cumulative… ***$1.7Trillion***

More teraflops means more….

WE NEED MORE *POWER*

Power Problems… with Chips



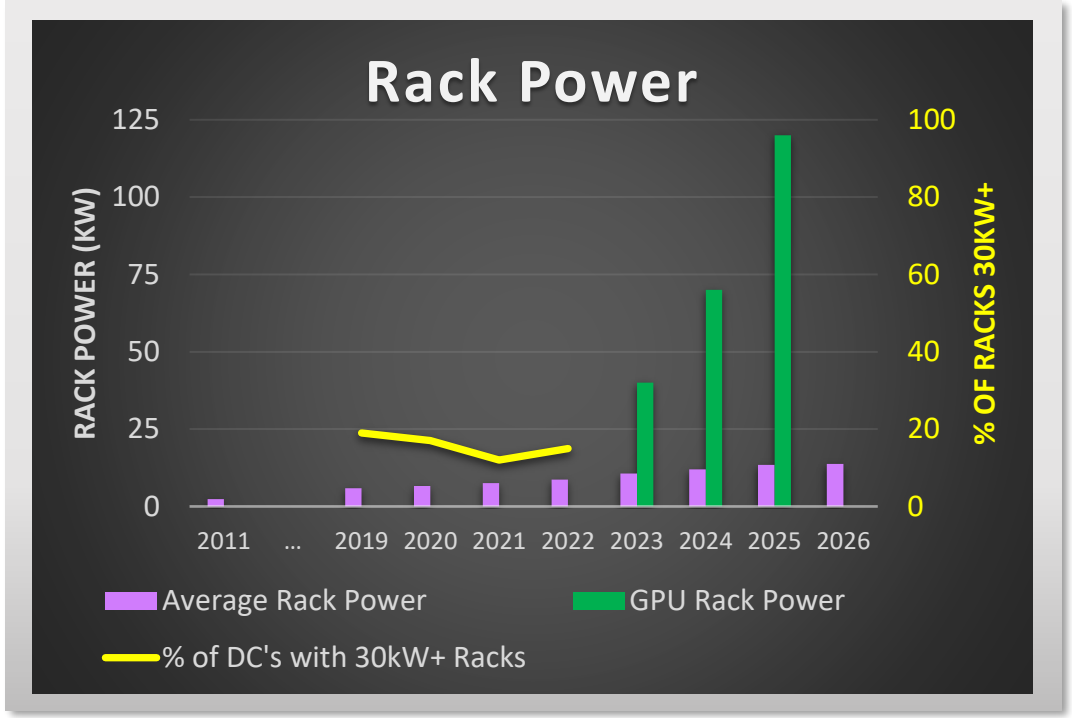
GPUs driving

10x power/chip density

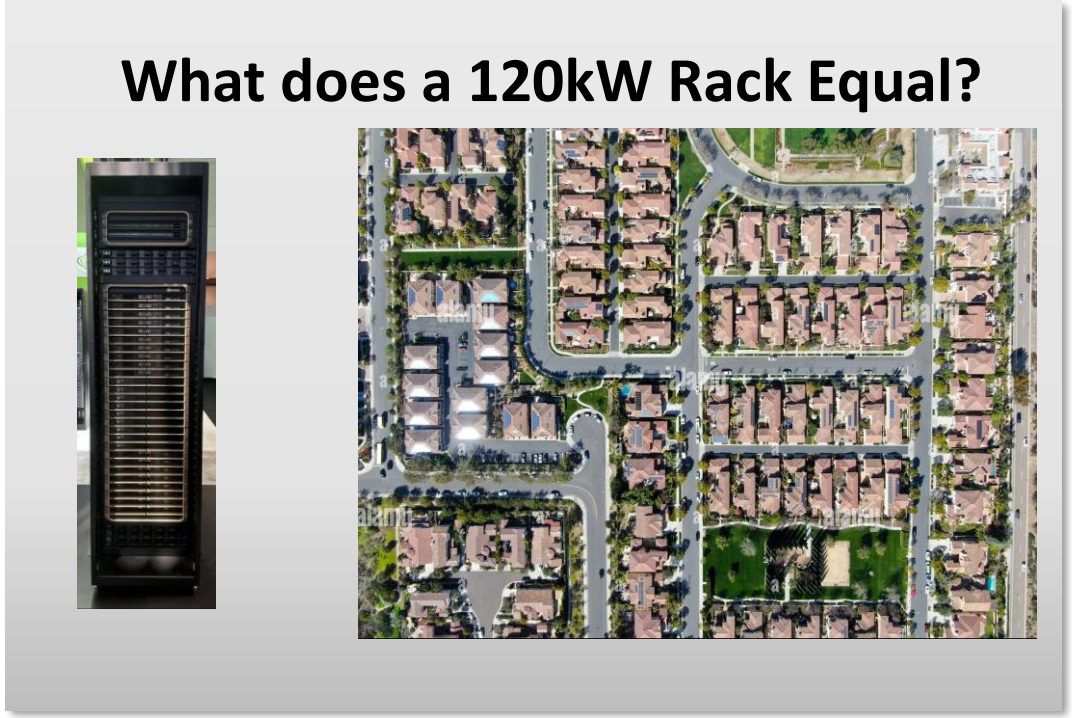
It’s getting hot

(and heavy!)

Power Problems… with Racks

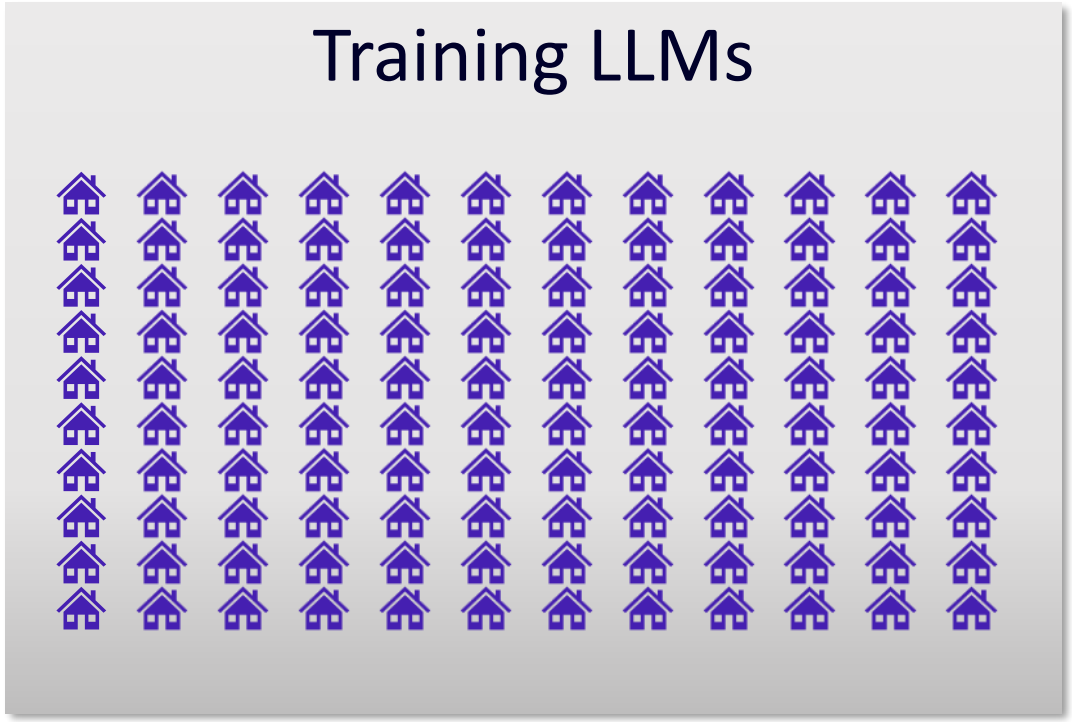


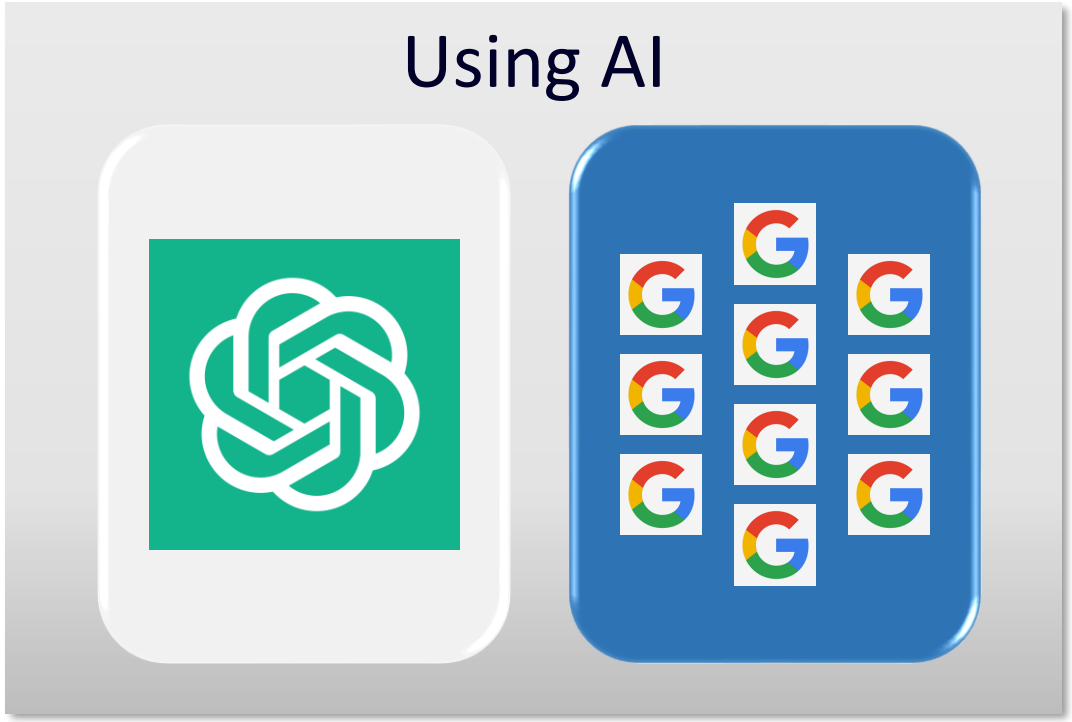
DCs not set up to handle the rack power increases



A *single* GPU Rack = *97* Homes

Power Problems…with AI



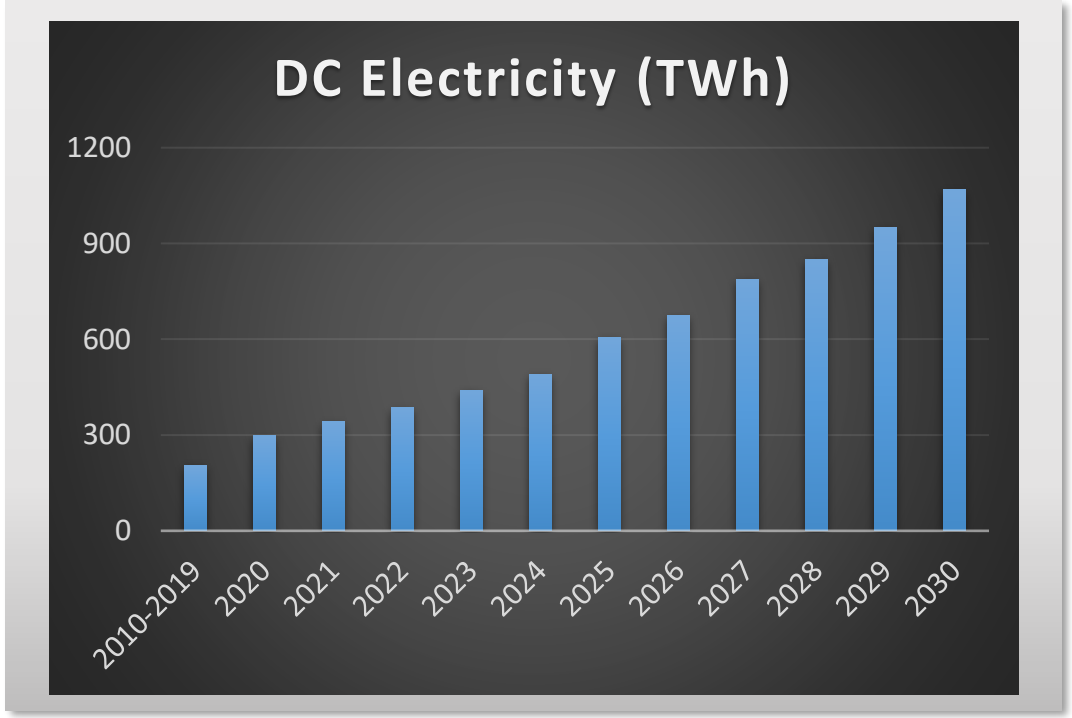


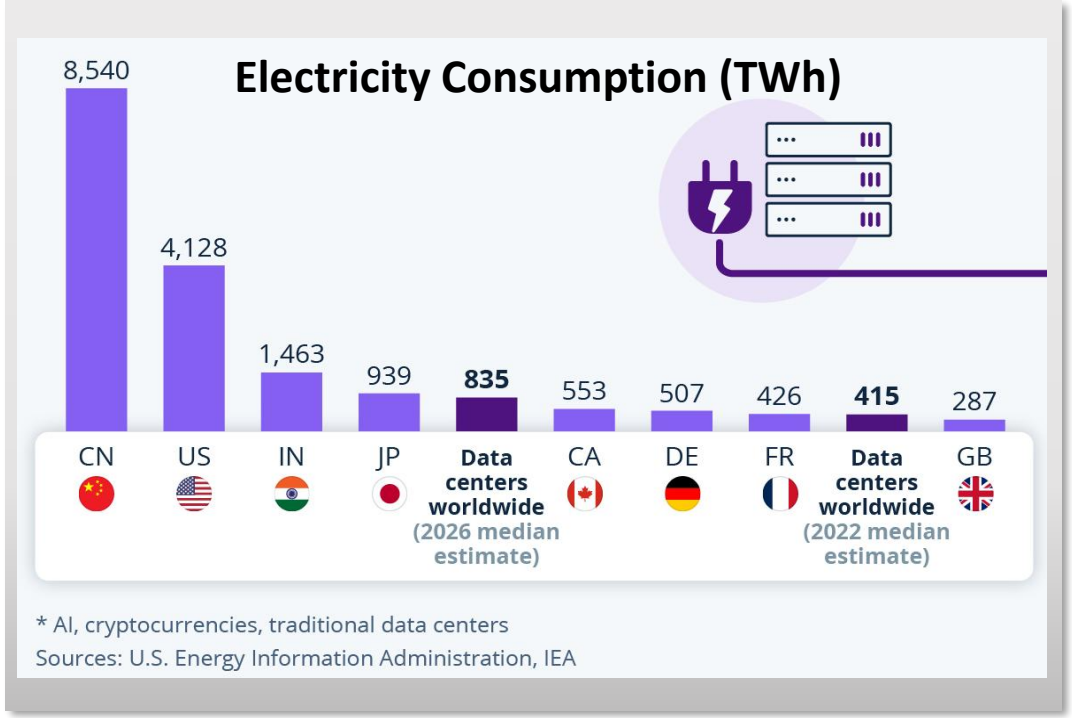
GPT-3 (1,300,000 kWh) = 120 home/yrs

GPT-4 = 160 homes

10x energy per query

Power Problems… with Data Centers





12-16% YoY growth in DC power

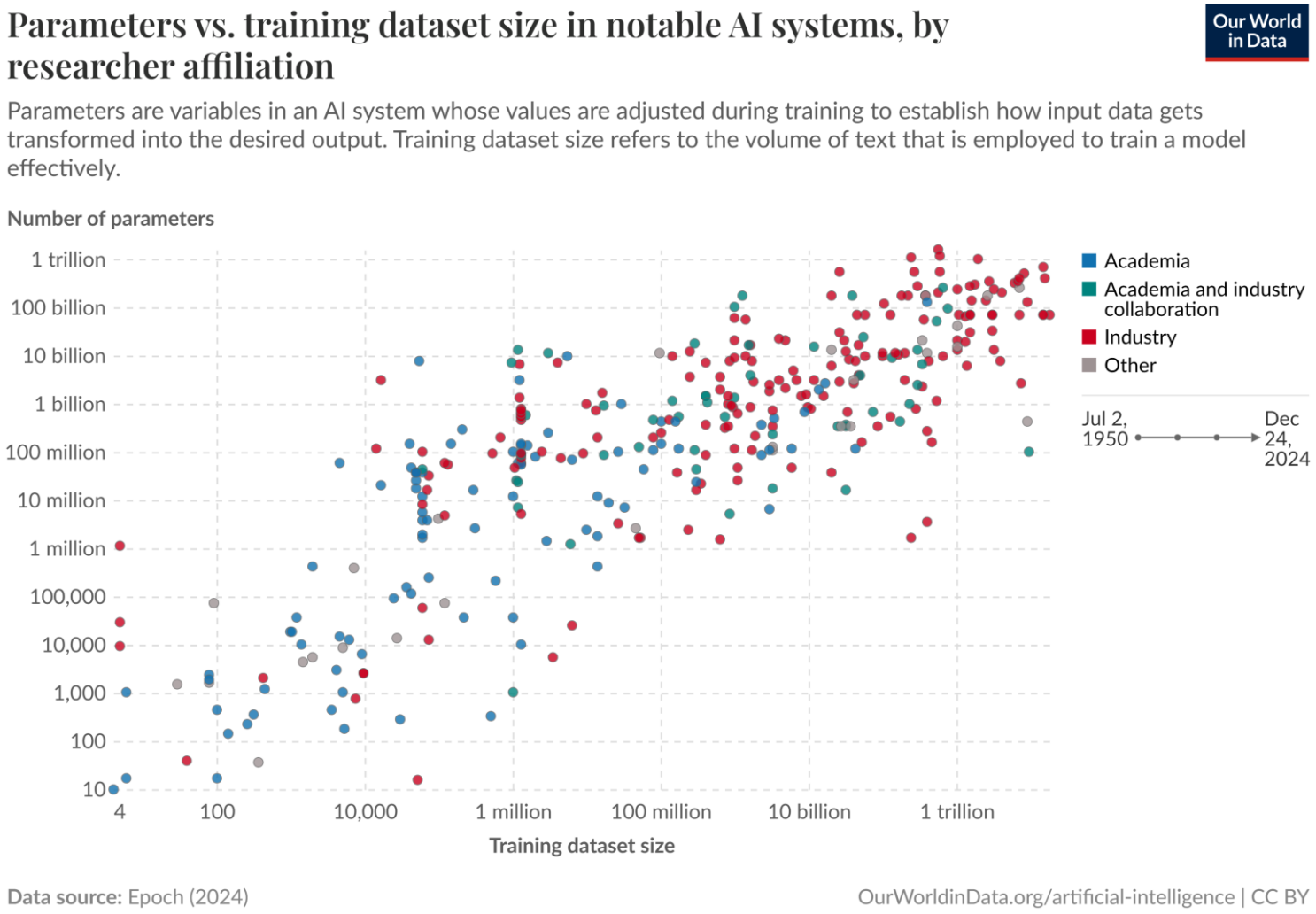
6-8x the rate for global consumption

Adding another France to the grid

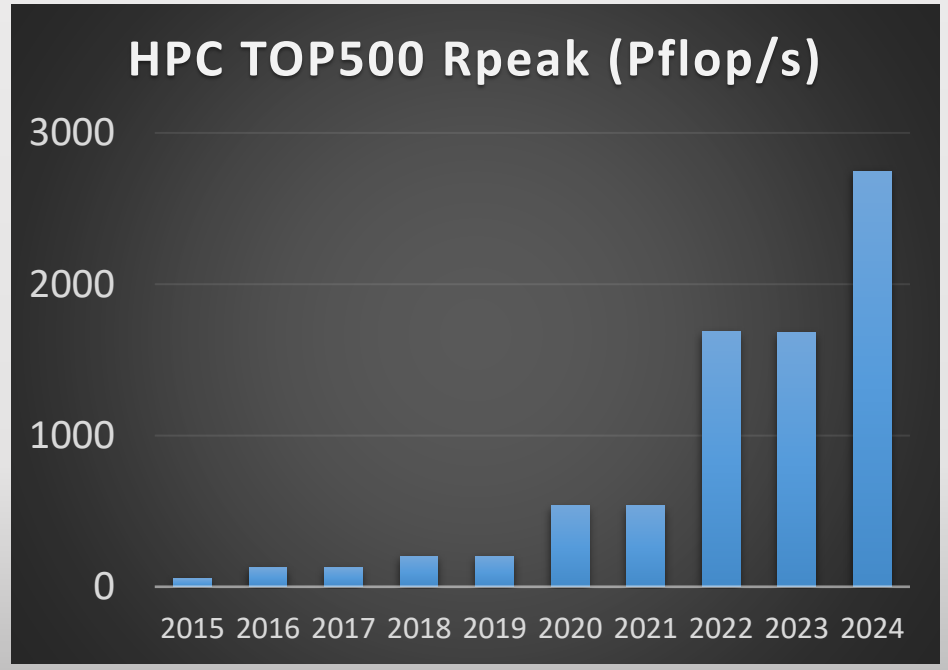
from ‘22 to ‘24 just for DCs!!

WE NEED MORE *PERFORMANCE*

Dataset size & Parameter count



System Performance



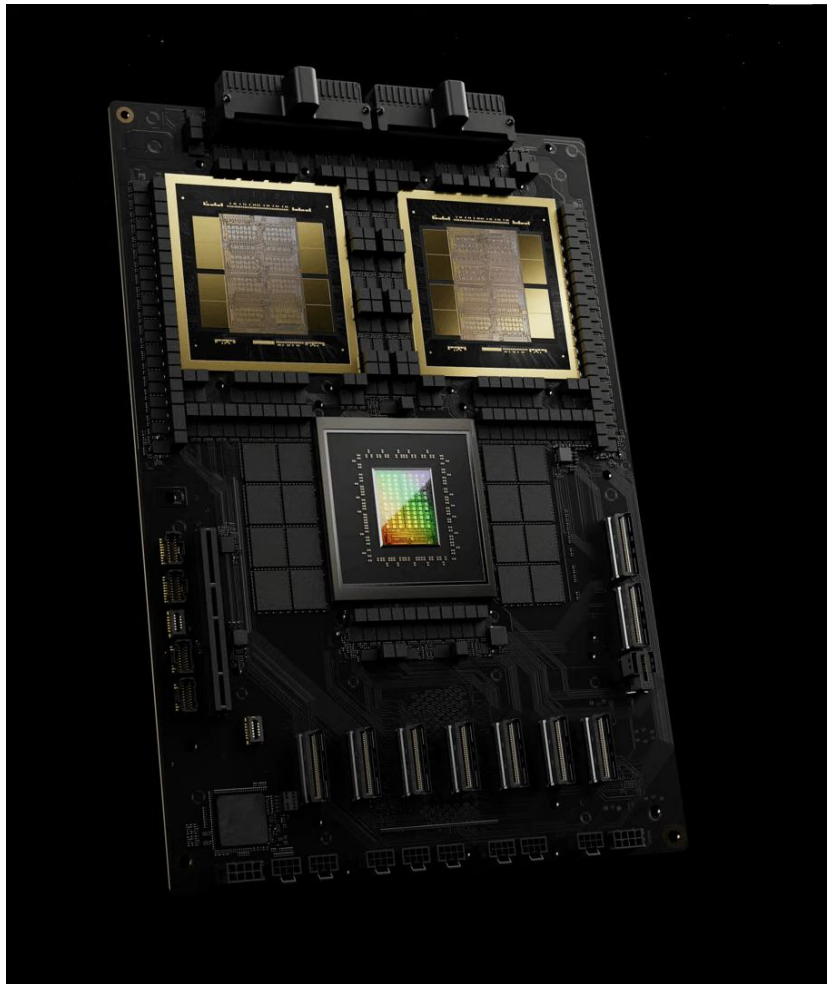
50x increase in leading system

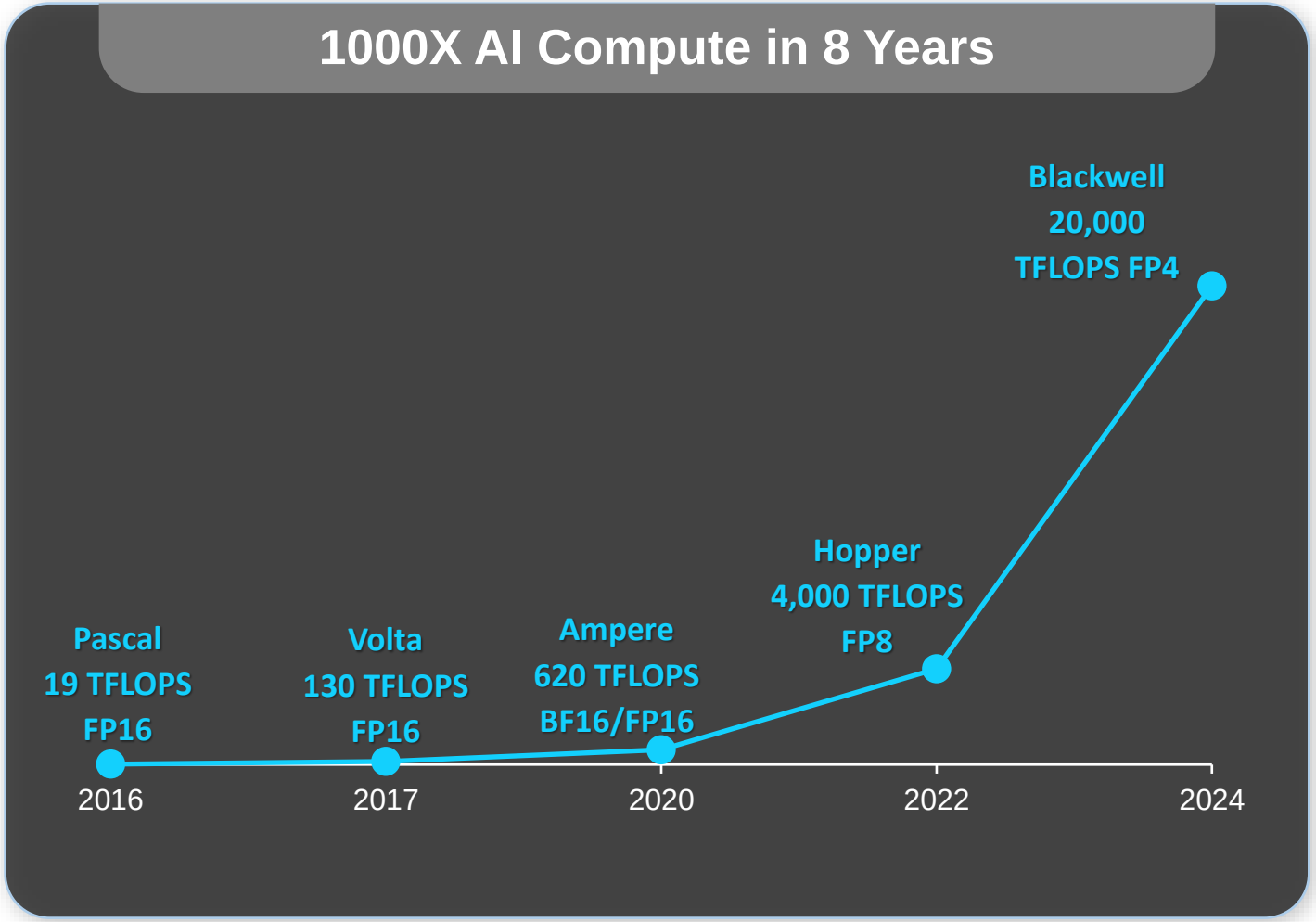
El Capitan – Lawrence Livermore

2,746 Petaflops

11M cores & 29.6MW

Processor Performance

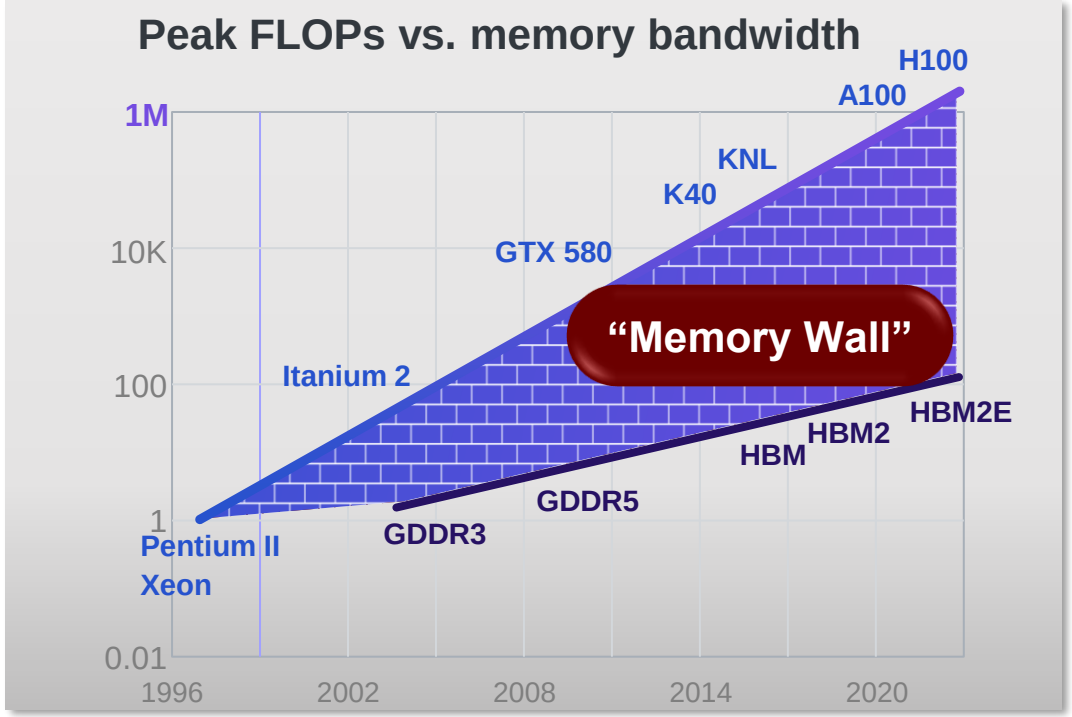


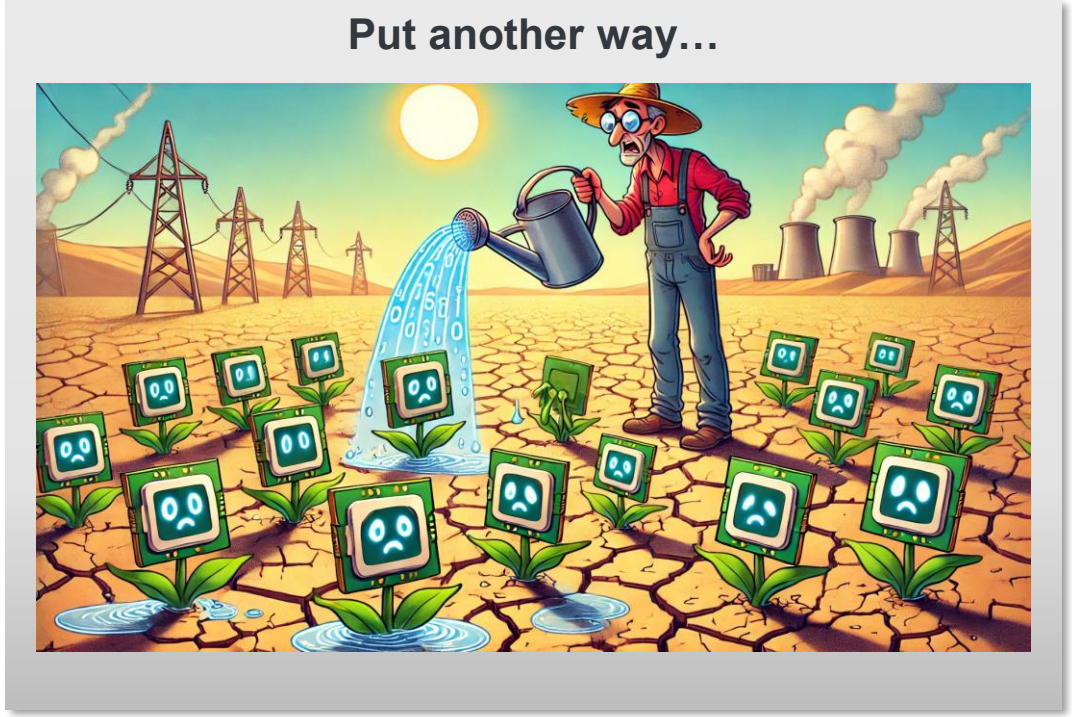


Source: Nvidia

01231

Memory Performance & the “Memory Wall”





Processor Flops: ***60,000x***

DRAM Bandwidth: ***100x***

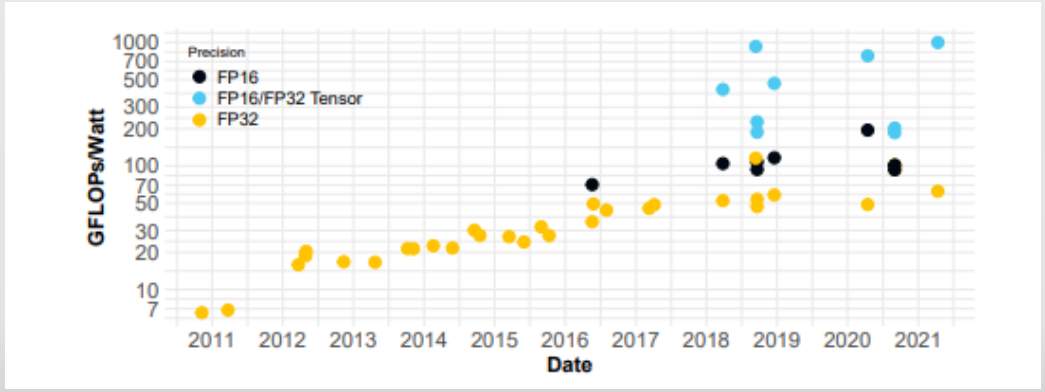
Get performance

into the data pipeline

WE NEED MORE *EFFICIENCY*

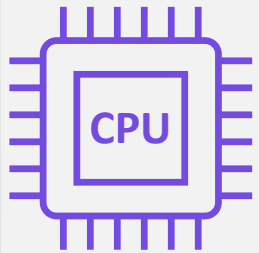
Efficiency Innovation… in the chips

Performance/Watt – *100x* over 10 years



xPUs doing their part!

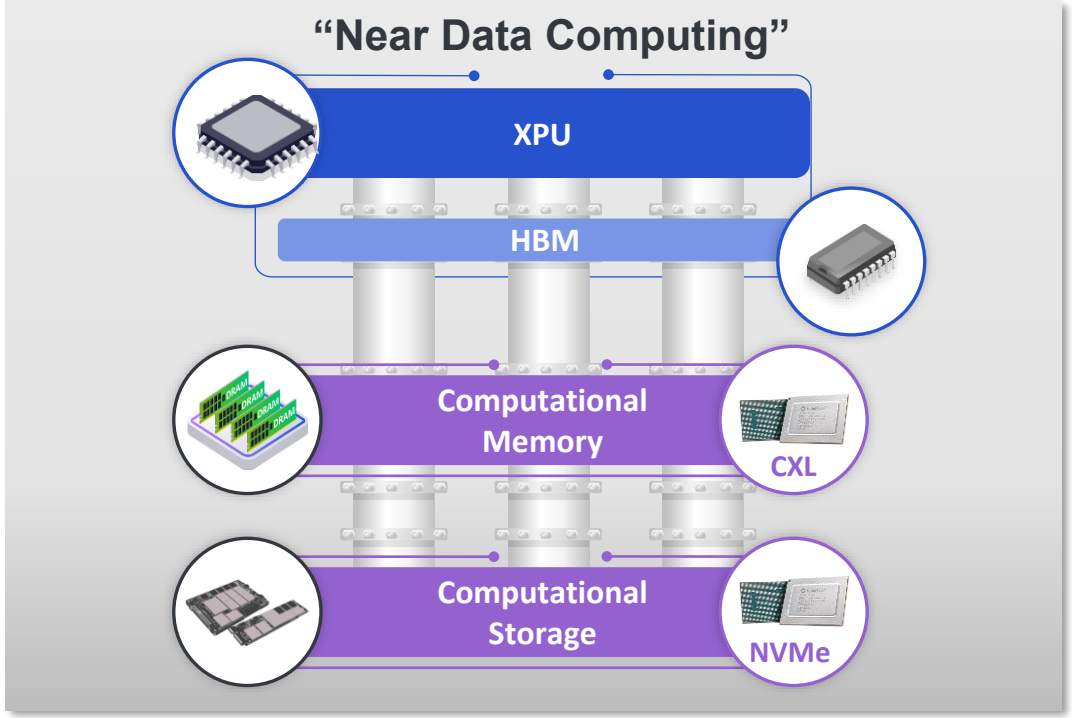
Specialization – Enter the "xPU"

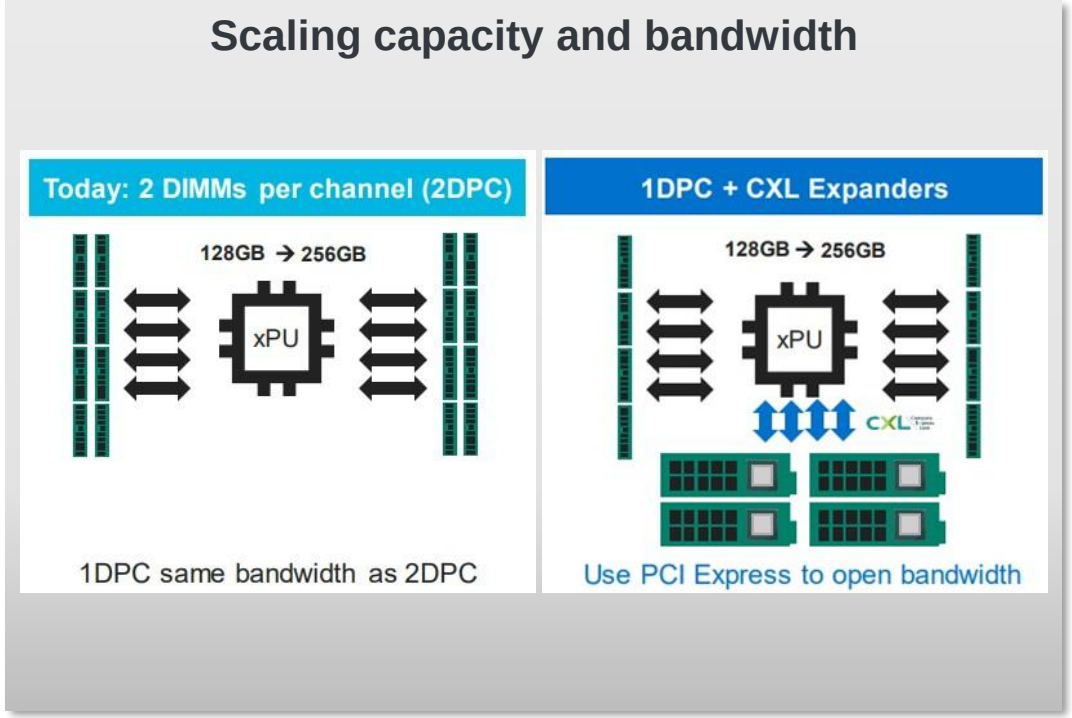


Graphics	GP GPU
Tensor	AI GPU
Network	LLMPU
Vector	Edge
Storage	IoT
Data	Security
Infrastructure	Memory

Fragmentation for optimization of processors for workloads

Efficiency Innovation… in the data pipeline

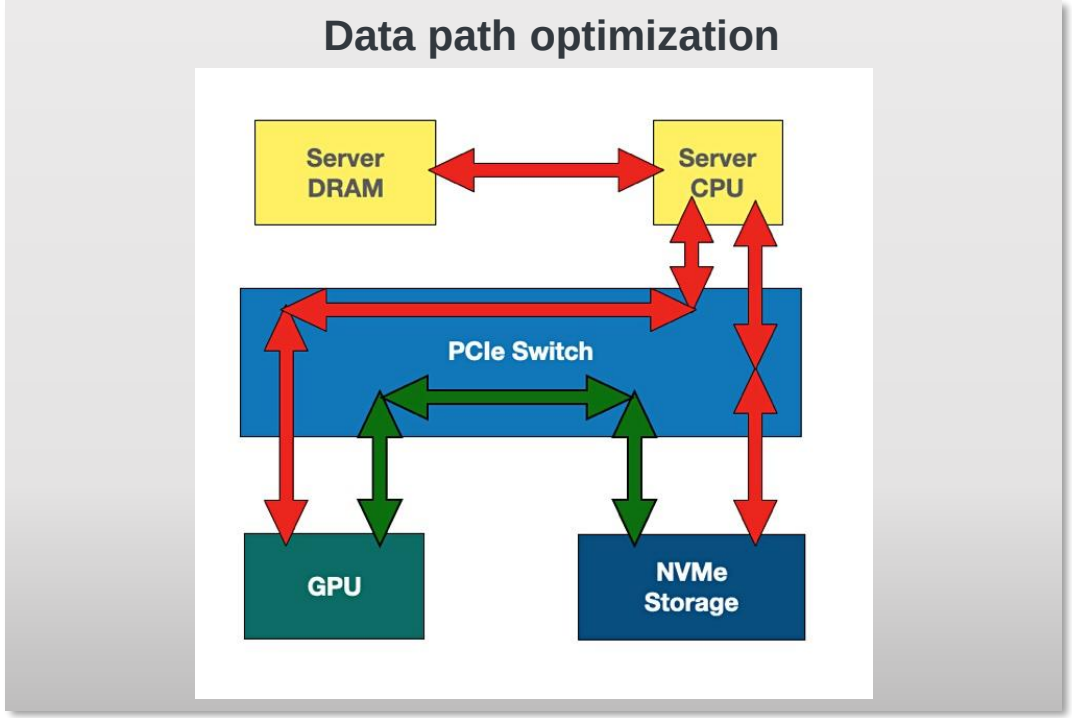




Distributed processing for optimized data movement

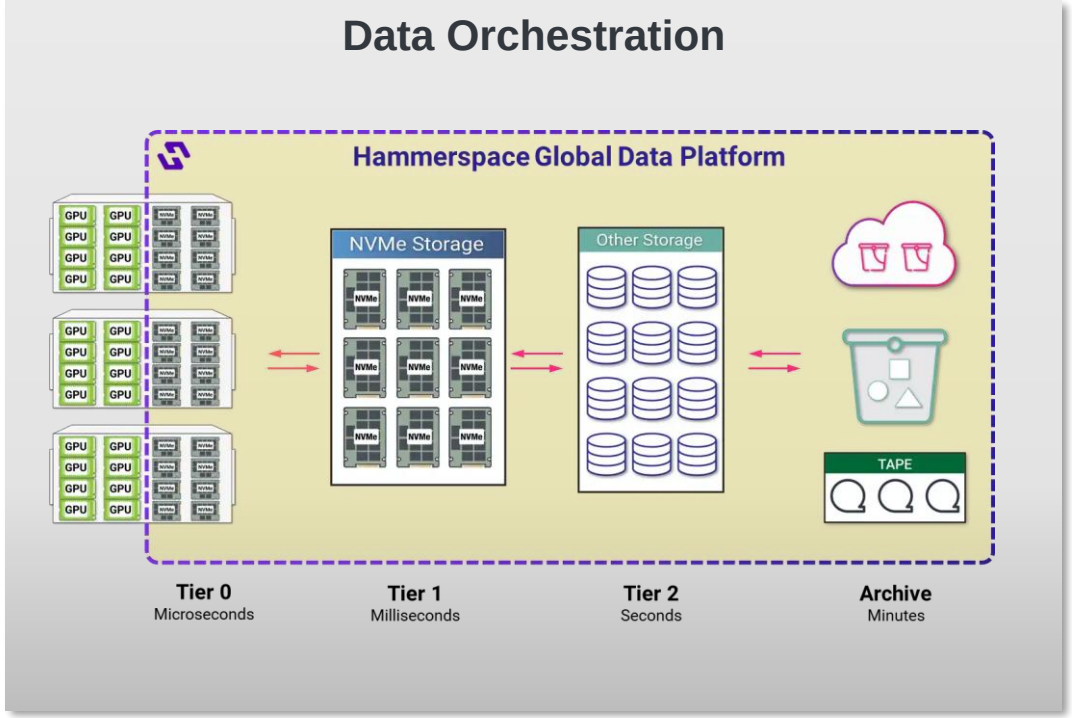
Breaking the ***"Memory Wall"*** boosting processor utilization

Efficiency Innovation… in the data pipeline



Magnum IO GPUDirect

mitigating memory traffic jams

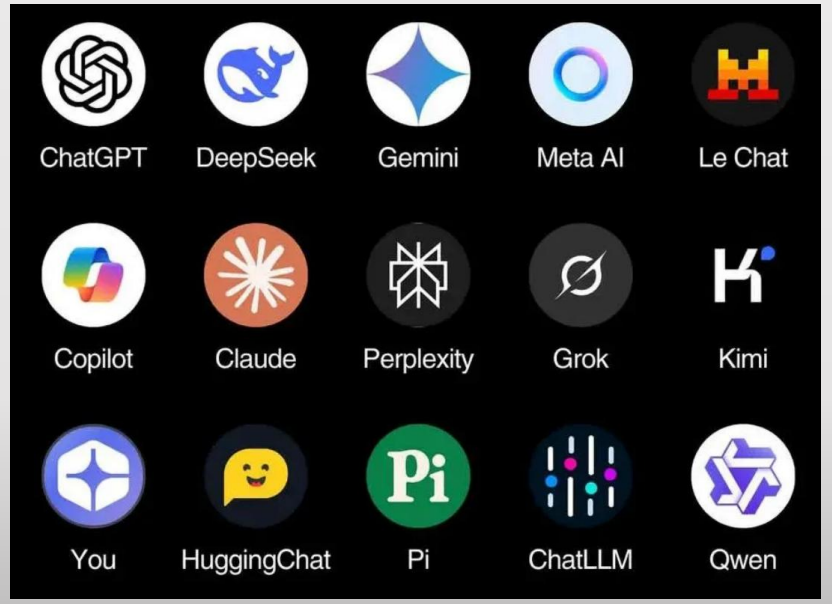


Breaking silos & separating

control from data traffic

Efficiency innovation… in the AI models

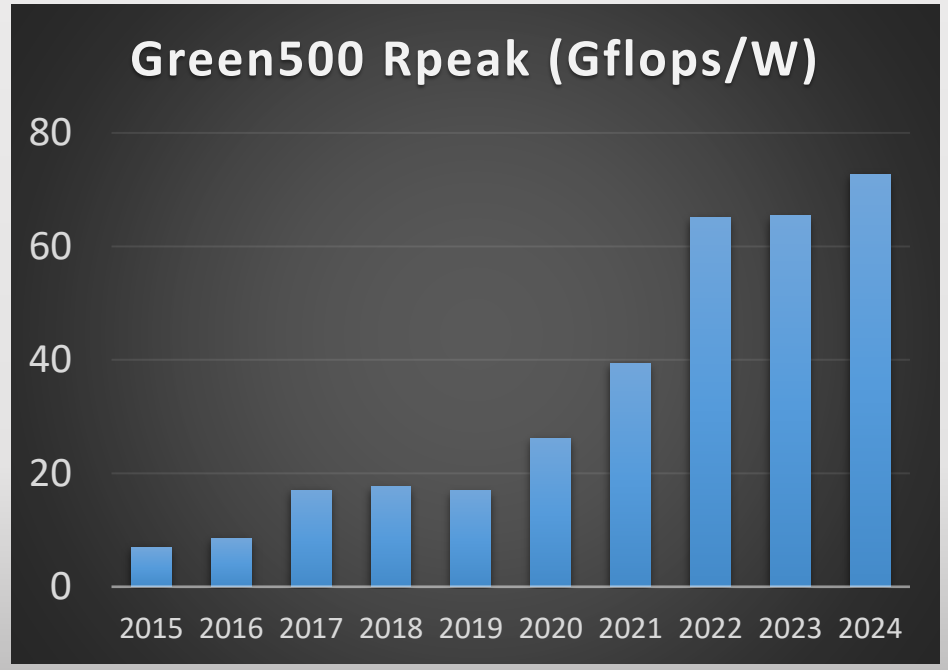
One model to rule them all??



Divergent strategies & optimizations

- Open source / Proprietary
- Mixture of Experts
- Group Relative Policy Optimization (GRPO)
- Retrieval & Cache Augmented Generation (RAG, CAG)
- Long context LLMs

Efficiency Innovation… in the systems



10x perf/W

HPC systems doing their part!

Rack & Connectivity

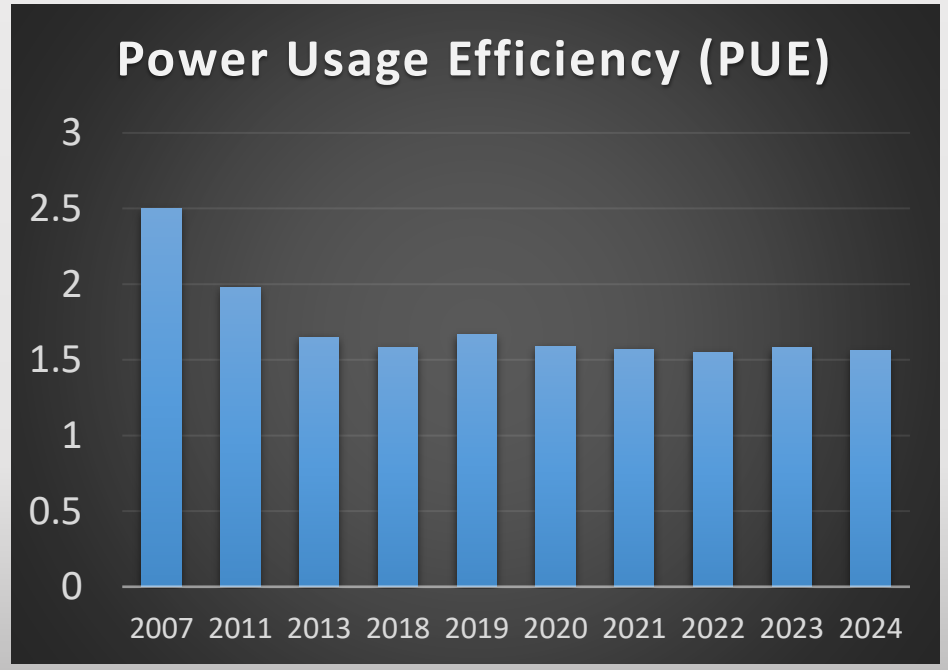
NVIDIA GB200 NVL72 – ***25x perf/W***

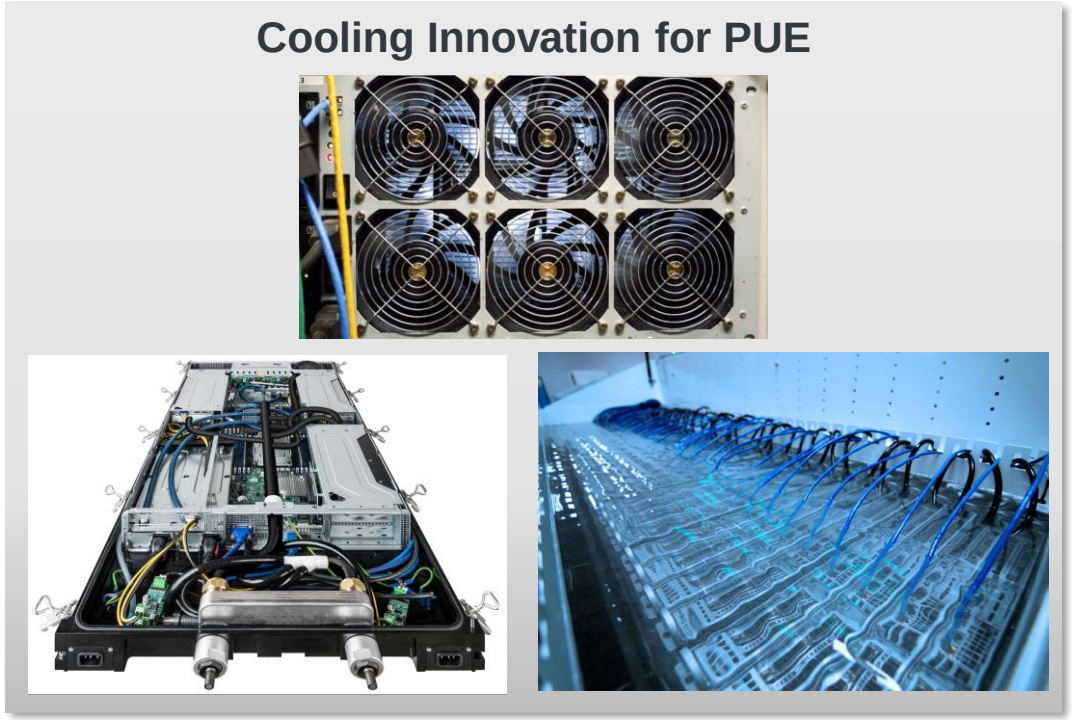
- "Rack-scale GPU"
- 30x inferencing performance
- NVLink chip-to-chip
- Liquid cooling, networking, storage, power distribution

Ultra Accelerator Link (UALink) Consortium

Expanding scope of design to ***rack-level*** optimization

Efficiency innovation… in Data centers





Average ***PUE leveling off***

After 40% reduction

Air → Direct Contact Liquid Cooling (DCLC) → Immersion

WE NEED MORE *PACKAGING*

What can you do?

Packaging for progress

- Heterogeneous integration
- Fan-Out Wafer-level packaging
- Integrated photonics
- Microfluidic cooling
- Materials innovations

“Boldly go where no one has gone before”

Thank You

900 North McCarthy Blvd
Suite #200
Milpitas, CA 95035

www.scaleflux.com
info@scaleflux.com

ScaleFlux