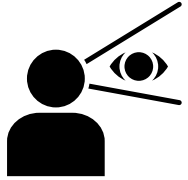


Sustainable Computing in an Era of Rising Hardware Costs and Slowing Per-Core Progress

Luca Atzori, Ian Fisk, Maria Girone

Sustainable



"Sustainable"

- Normally means discussions of how to sustain everything (reducing carbon, power, etc.). I'm going to include sustaining the mission



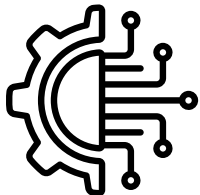
Money

- Computing budgets are assumed to be flat



Time

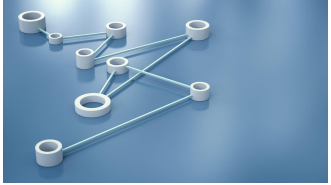
- We need to be able to calculate things in reasonable lengths of time



Technology

- We need the right technology to solve the challenges we face

Outline



Computing Evolution

- Computing processing and storage capacity and performance advances
- Following long trends and driven by technology improvements



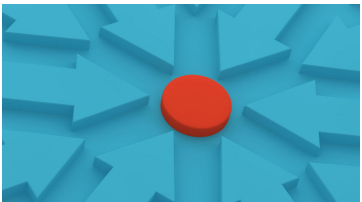
Replacement Models

- Motivation for the models we have



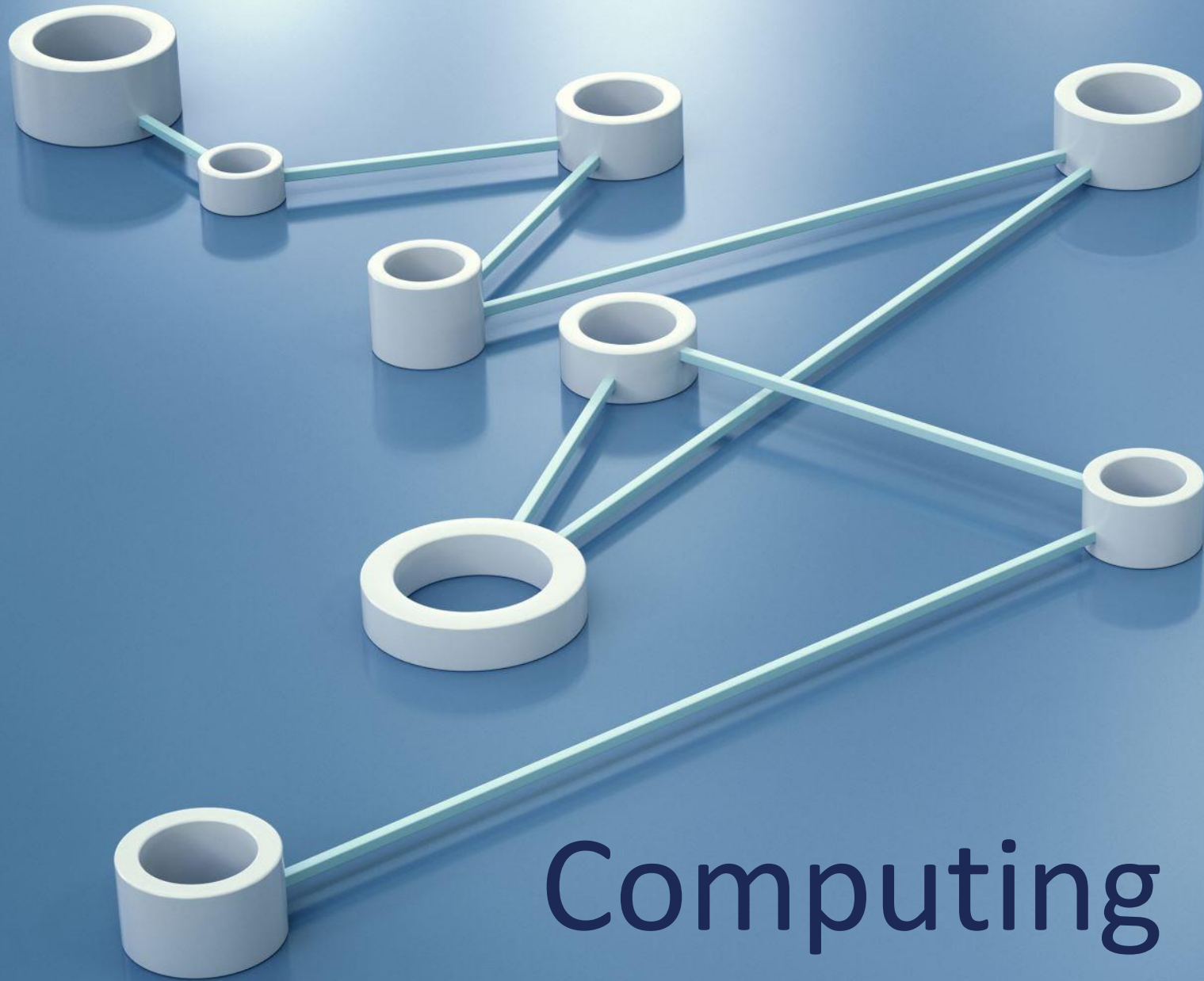
Cost Evolution

- Historically computing costs have remained stable or slowly decreased
- Not anymore



Impact

- How does this change how/when we replace computers?
- How does it change what resources we use?
- How does it change our expectations for how many resources we have?



Computing Evolution

Denard Scaling

Computer clock speed rose exponentially from 1970 to 2005

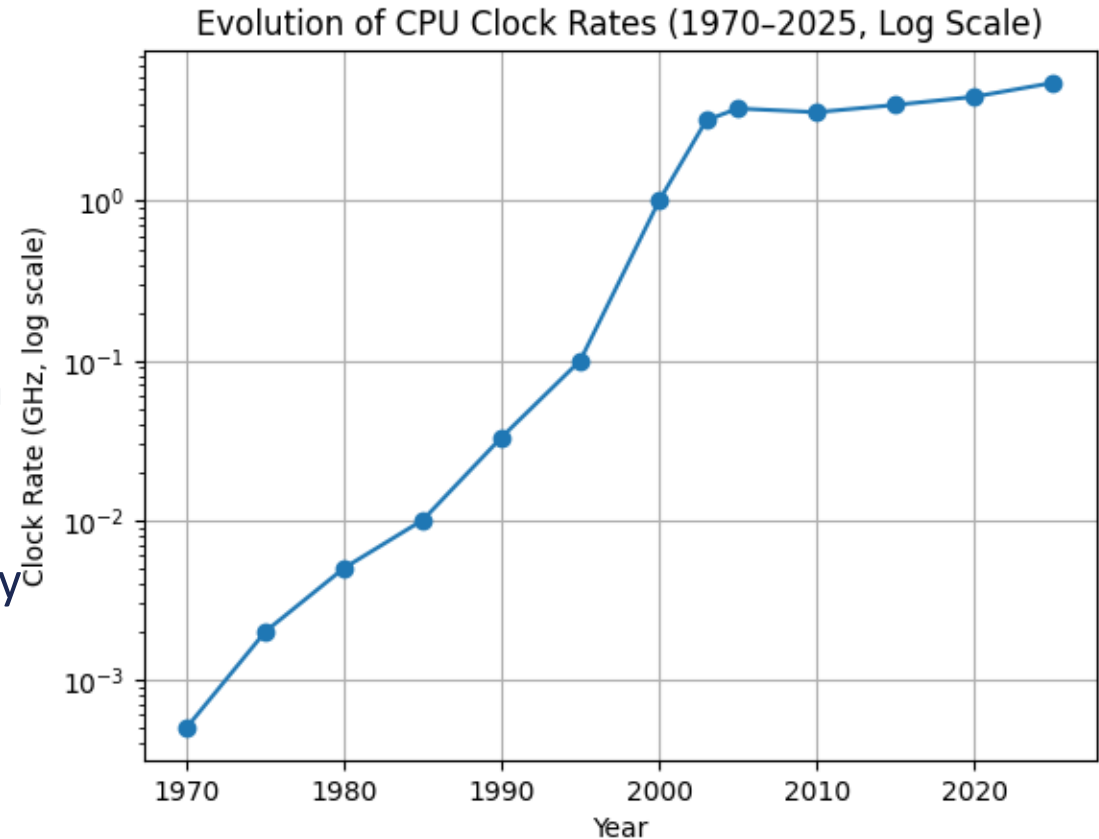
- In a 15-year period the speed of computers could increase by a factor of 100

In the last 20 years, computer clock speeds have increased by a factor of 2-3

- Modern server systems are now ~4GHz
 - Close to the 6-10GHz max possible with silicon technology and the limits of the speed of light

Performance of the individual cores increases with the wider instruction sets and more calculations per clock cycle

- Linear and not exponential improvements

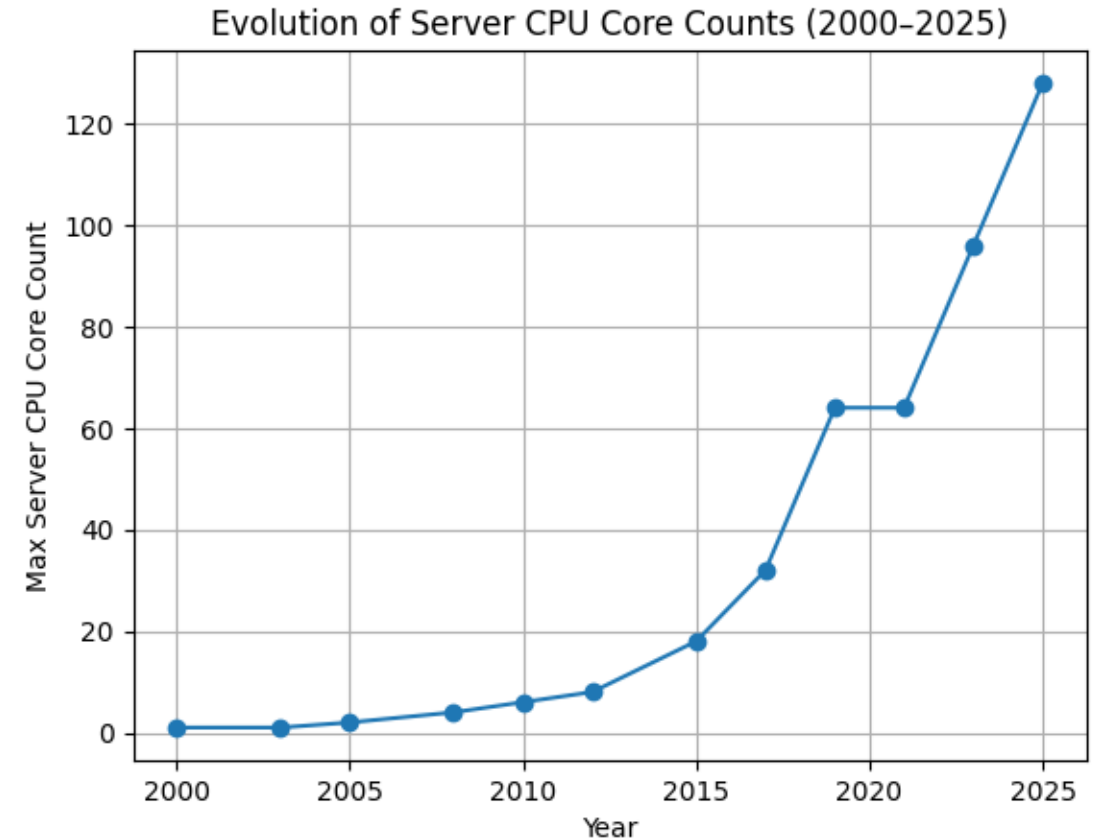


Multicore

Unable to increase the clock and still able to increase the density of transistors and size of the silicon, multi-core CPUs became standard

- Computing capacity increases with each additional core
- 256 Core CPUs will be available next year
 - Memory bandwidth per core has largely been maintained
 - 12-16 DIMM channels per socket
 - Memory per core can also be maintained with increasing DIMM size
 - IO per system also scales as 400Gb/s networks per system are available

This pushed many processes per system or highly parallel code



Watts

Increasing the density of silicon reduces the watts per core as the feature size decreases and the efficiency improves

- These improvements are flattening

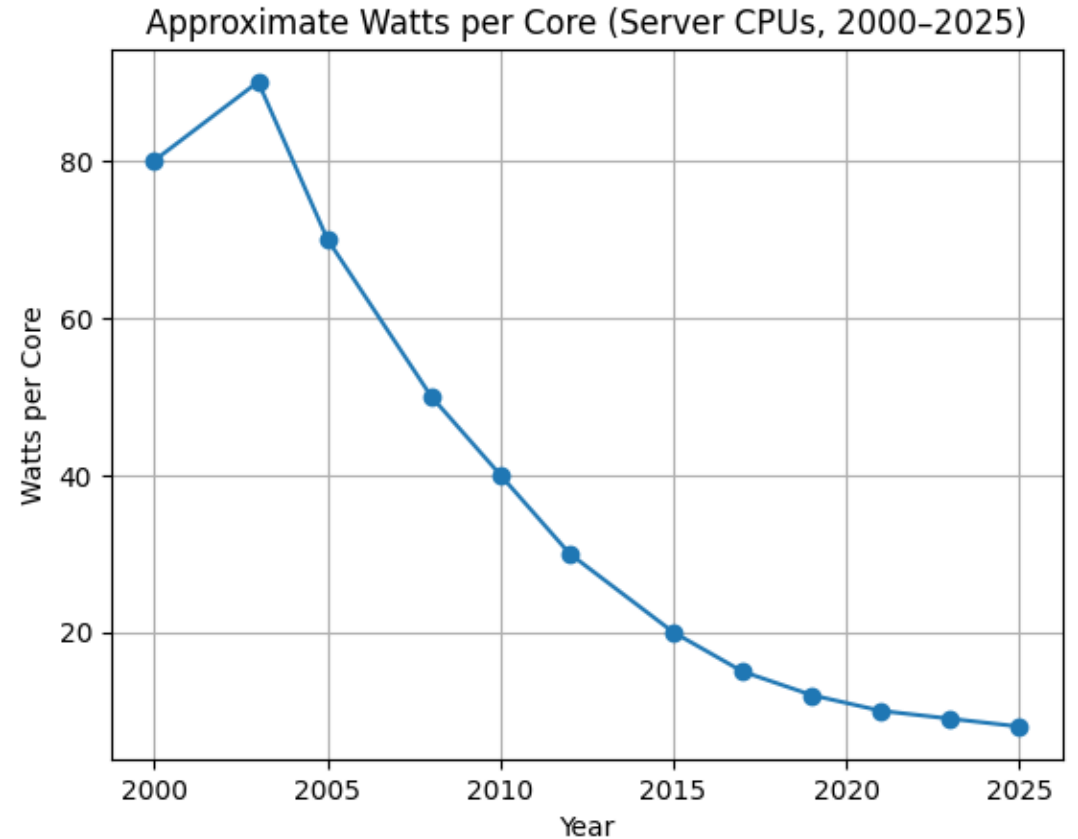
A modern CPU socket is 400W-500W

- Many layers of silicon drive the need for even more cooling

The amount of air a fan can move goes as the square of the radius

- A fan in 1U server spins 4 times as fast as a fan in the 2U server
- A 500W CPU might have 250W of power used blowing air over it.

Drives toward direct liquid cooling



Enter the GPU

First GPU used in a Top500 Supercomputer was TITAN in 2012

- Built by Oak Ridge
- 16k K20 NVIDIA GPUs
- 8MW
- Number 1 on Top500
- 17PFlop/s



GPUs were originally designed to apply rotation matrices to objects for graphics rendering

- A K20 had 2500 CUDA cores to parallelize matrix multiplication (A B200 has 18k cores)
- It is a linear algebra accelerator

GPU Performance per Watt

2013

The very first Green500 list had a GPU accelerated system at the top

- 3.2GFlops/W for NVIDIA K20
- This beats a Xeon Phi by 30%

Today

The 2025 list the top 100 slots are GPU accelerated

- 73GFlops/W Grace-Hopper
- Top CPU based systems is Fujitsu at 16GFlops/W
- Top X86 based system is 192 Core AMD at 13GFlops/W

2013

Green500 Rank	MFlops/watts	Site	System	Total Power(kW)
1	3208.8	CINECA	Eurotech Aurora HPC 10-20, Xeon E5-2687W 8C 3.100GHz, Infiniband QDR, NVIDIA K20	30.7
2	3179.9	Selex ES Chieti	Eurotech Aurora HPC 10-20, Xeon E5-2687W 8C 3.100GHz, Infiniband QDR, NVIDIA K20	31.0
3	2449.6	National Institute for Computational Sciences/University of Tennessee	Appro GreenBlade GB824M, Xeon E5-2670 8C 2.600GHz, Infiniband FDR, Intel Xeon Phi 5110P	45.1

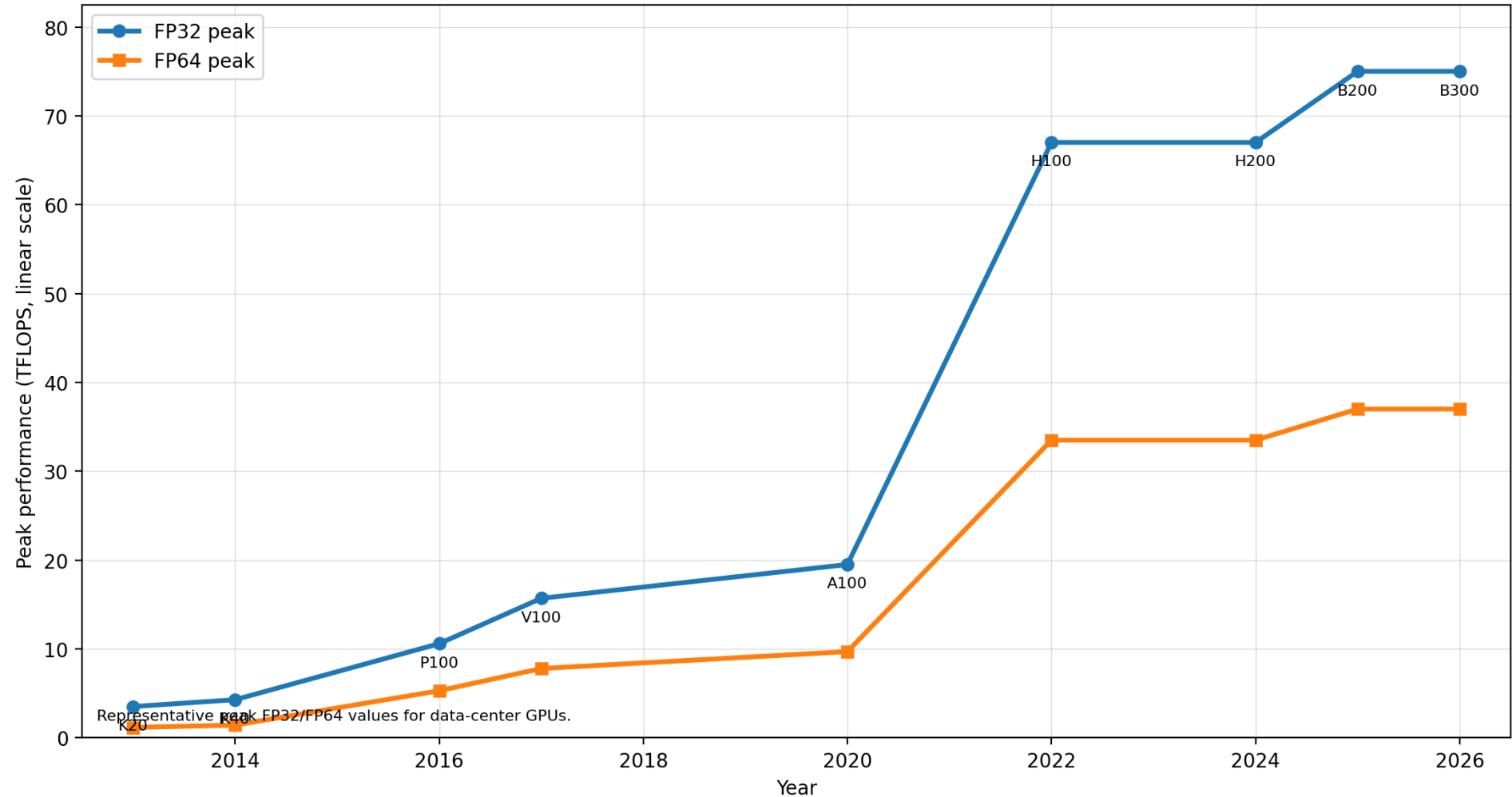
The concentration from the manufacturers has been on performance

GPU Advances

The FP32 performance of a modern GPU has increased at 25% a year for the last decade

- Outpacing CPU improvements

NVIDIA GPU FP32 vs FP64 Performance Evolution (K20 → B300)

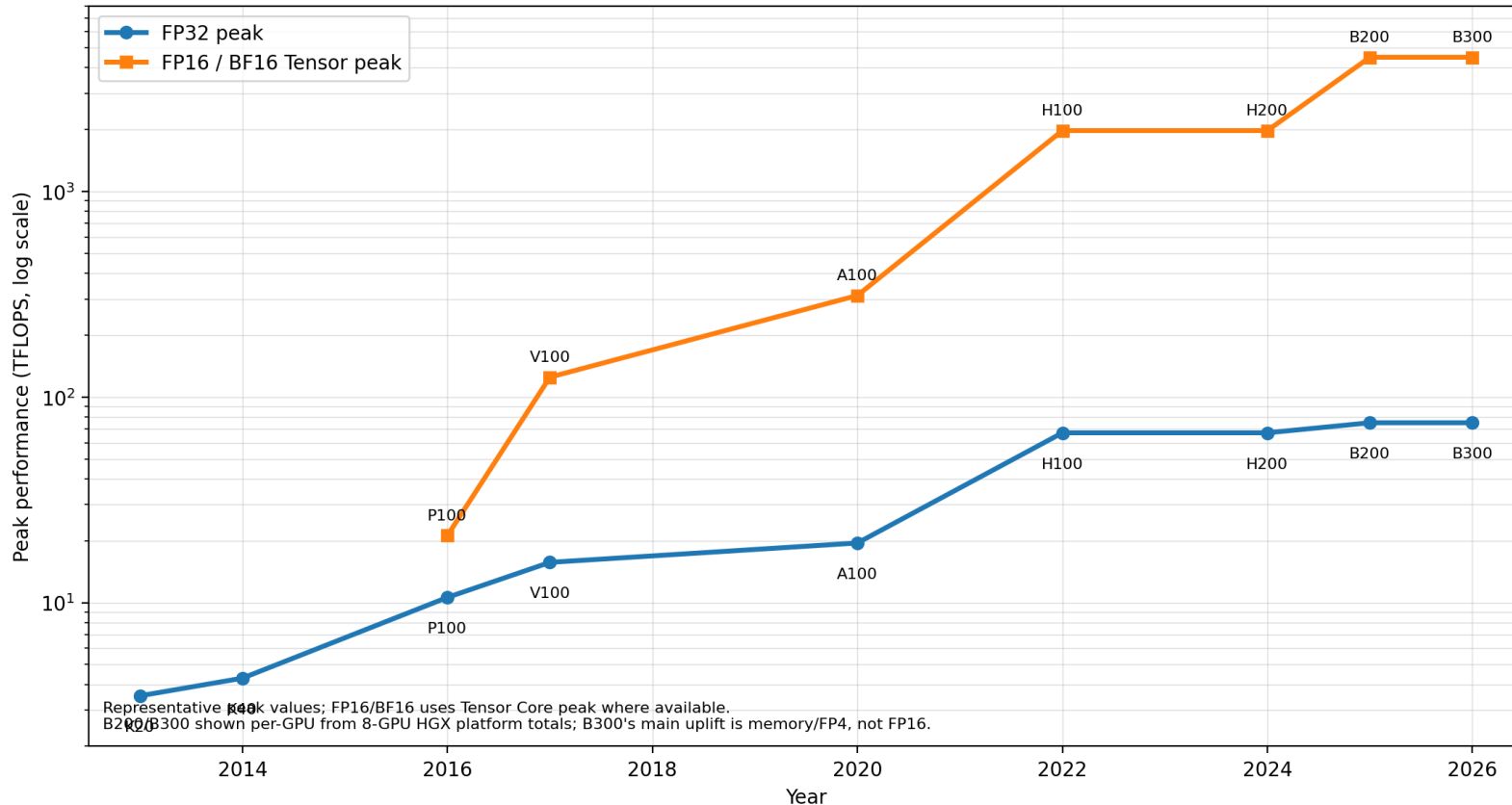


GPU Leveling

AI can benefit from low precision calculations

- There is a limit amount of space on the silicon,

NVIDIA Data-Center GPU Peak Performance Evolution: K20 to B300



H100 SXM	
FP64	34 teraFLOPS
FP64 Tensor Core	67 teraFLOPS
FP32	67 teraFLOPS
TF32 Tensor Core*	989 teraFLOPS
BFLOAT16 Tensor Core*	1,979 teraFLOPS
FP16 Tensor Core*	1,979 teraFLOPS
FP8 Tensor Core*	3,958 teraFLOPS
INT8 Tensor Core*	3,958 TOPS
GPU Memory	80GB

GPU and CPU Watts

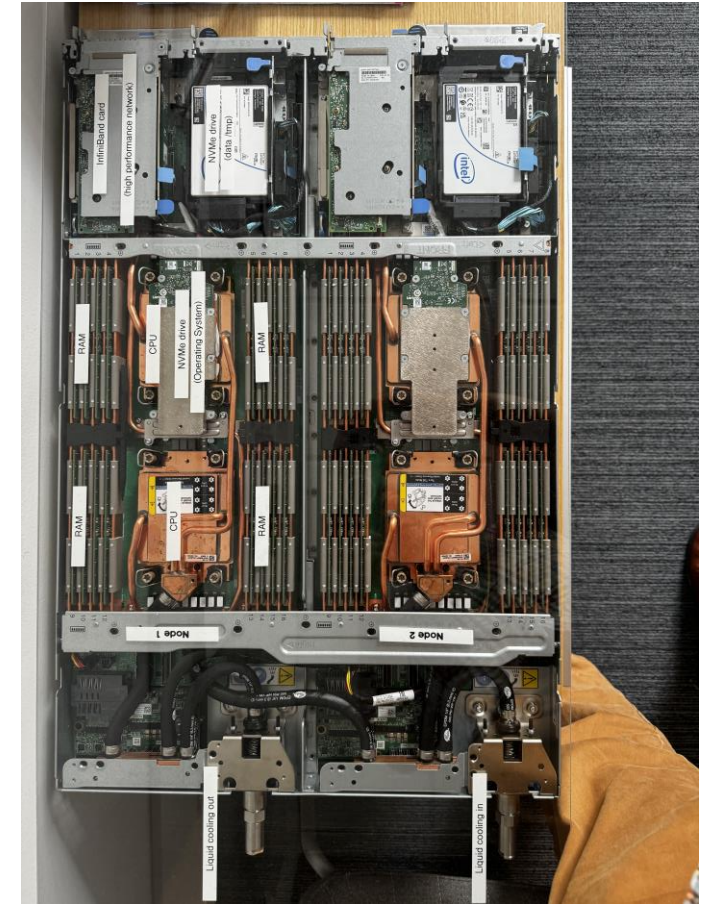
Modern processing devices use a lot of power

- 500W CPUs
- 1000W GPUs with 2000W GPUs on the roadmap

They are effectively impossible to cool with air

- Heat capacity of the water is 4000 times that of air
- Flow rate is proportional to the surface area of the fan, so fan speed increases with r^2 this uses a lot of power
- Fans are moving parts and wear out

Large scale HPC installations and AI clusters are all moving to DLC



Status

Evolution of system improvement is slowing

- 15% annually for CPUs
- 25% annually for GPUs

Efficiency in terms of Flops/W has improved, but the rate of improvement is also slowed

- GPUs are roughly 5 times as efficient as CPUs
- The focus of manufactures has been on performance
 - We have seen dramatic increases in Wattage

Our economic models are based on costs remaining constant or slowly decreasing

- When to replace
- What technology to use

Replacement Models

Replacement Models CPUs

How did we get to the 5-6 year replacement cycle for CPUs?

Historical

- When computers performance doubles every 18 months, after 5 years the old gear is 10% the performance of the new

Practical

- 5 years is about the max a company will sell a support contract for
- Air cooled systems have a lot of moving parts

Operations

- If a compute node is 800W, ~7kWh
 - At 25 centime a kWh would be 1.75kCH a year
 - At a PUE of 0.4, the cooling would be 700CH a year
- At 10k CH a node, operations costs as much as the node by year 4
- Even at 15% improvement per year, after 5 years a new system is 50% more powerful with less energy per core.

Replacement Models GPUs

GPUs are typically on a 3-4 year placement cycle

Historical

- GPUs have doubled every 18 months and GPU memory has increased at a similar rate
- Older gear is not only slower, it may be unsuitable

Practical

- 39 months is the maximum warranty NVIDIA currently offers

Cost Evolution



Current Investments in AI

US industry is expected to invest \$500B-\$700B in data centers, AI Facilities, and research this year

- This investment is concentrated in the top hyperscalers
- NVIDIA announced they have ~\$1T in orders for the next 12 months
 - Includes worldwide sales

The US government will invest roughly \$3—\$5B in AI research

It's not just that we aren't driving anymore, we're barely influencing

Impact Memory and Storage

GPUs need memory to store large and complex models

- **To make effective use of those models tremendous amounts of RAM and fast storage are needed**
- **As the focus has shifted from training to profitably using the models for inference, the demands for this storage have increased dramatically**

AI hyperscale installations are buying all the memory and fast storage

- **Exacerbated by the focus on HBM for GPUs**

Impact Memory

Micron makes memory and their stock is up nearly 700% in the last year

- Memory prices have increased by roughly a factor of 5 in the last 6 months

It is not possible to buy 128GB MR DIMMs in 2026

- All production is spoken for
- 32GB MR DIMMs will be end of life this summer as production focuses on more expensive large devices

All the Flatiron Purchase orders from the fall were cancelled by the manufacturer because they would lose too much money

- Our order placed in September was \$6.5M
- In February it was repriced at \$13M
- 3 weeks later it was priced at \$19M

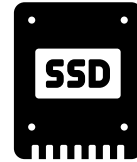
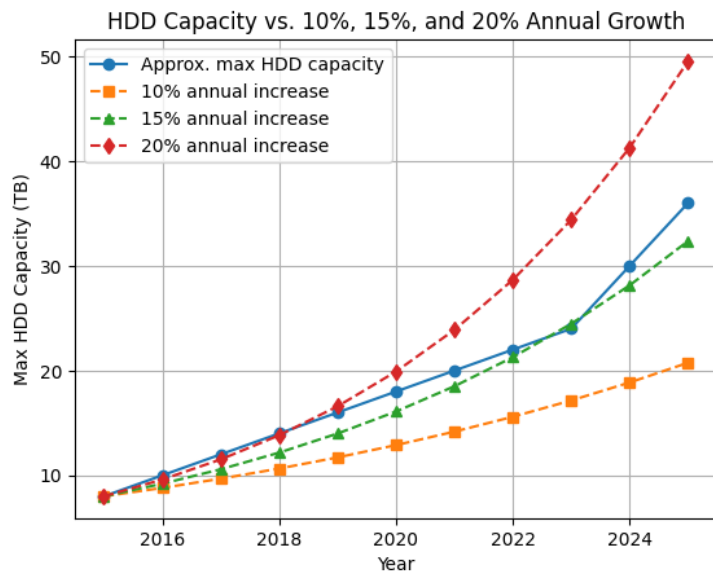


A tale of two drives



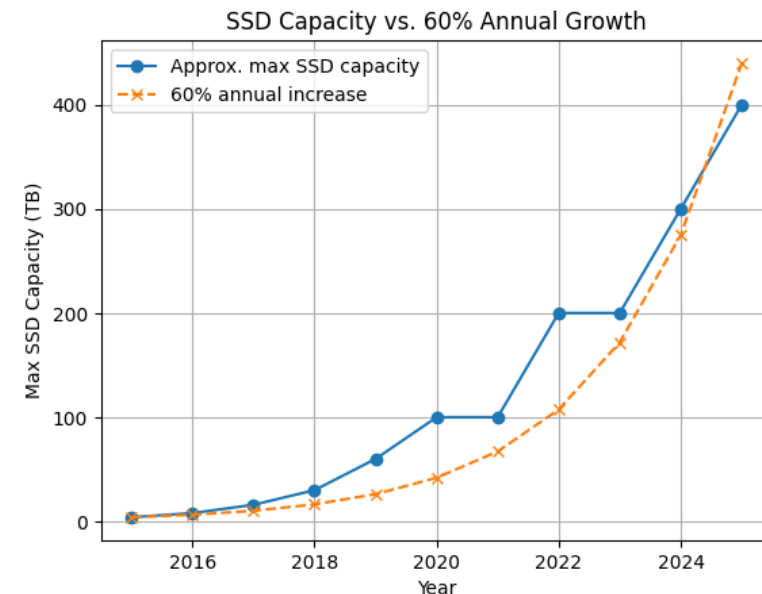
The hard disk

- The venerable tool of HEP computing
- Slowly increasing with changes in technology and number of platters
- Assembled and production capacity limited by labor and demand
- Cost has increased by about this year 10% and capacity by 15%



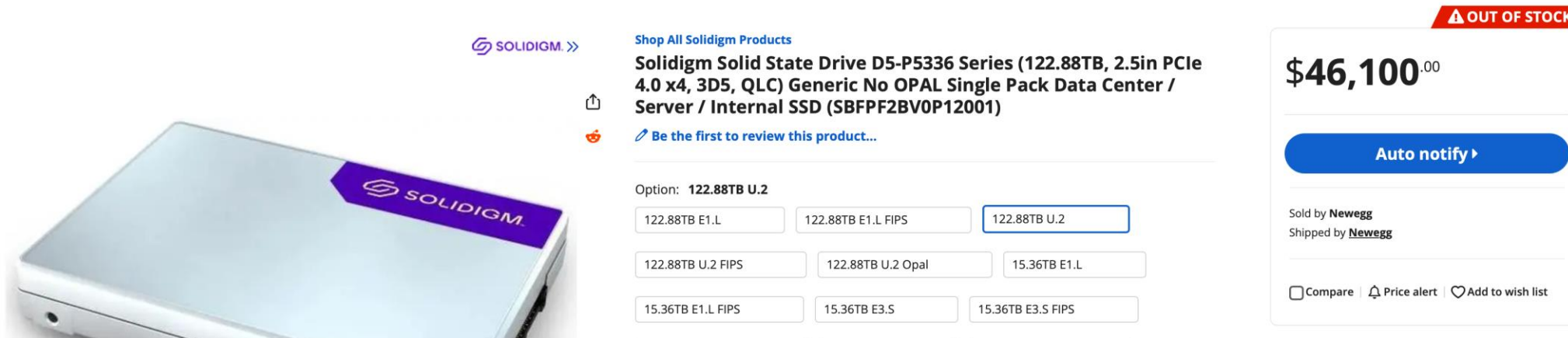
The solid-state drive

- Rapidly increasing driven by increases in density and number of layers
- Production limited by demands on fabrication facilities
- Cost has increased by a factor of 3-5 since January



Costs

- Recently, I received a quote for 200PB of SSD space
 - \$55M dollars in December 2025
 - \$96M in January of 2026
- Similar percentage increase in the components we use for processing and service nodes
 - Roughly a factor of 2 increase for the same size
- Increases are worse at the high end of the scale



SOLIDIGM >>

[Shop All Solidigm Products](#)

Solidigm Solid State Drive D5-P5336 Series (122.88TB, 2.5in PCIe 4.0 x4, 3D5, QLC) Generic No OPAL Single Pack Data Center / Server / Internal SSD (SBFPF2BV0P12001)

[Be the first to review this product...](#)

Option: **122.88TB U.2**

122.88TB E1.L	122.88TB E1.L FIPS	122.88TB U.2
122.88TB U.2 FIPS	122.88TB U.2 Opal	15.36TB E1.L
15.36TB E1.L FIPS	15.36TB E3.S	15.36TB E3.S FIPS

\$46,100^{.00}

OUT OF STOCK

[Auto notify](#)

Sold by **Newegg**
Shipped by **Newegg**

Compare | Price alert | Add to wish list

HPC/HTC Computing



8-way NVIDIA HGX Node

- \$350k (has been \$250k last year)
 - Increase of 40%
- 1 set of RAM
- Flatiron has 36 of these



HPC/HTC Node

- \$95k (had been \$32k last year)
 - Increase of 300%
- 1 set of RAM
- Flatiron wanted to buy 200 of these

It's easier and more affordable to buy AI optimized hardware than HPC/HTC

GPU Cost Calculus



Changing memory costs changes the economics of what to use



Previously

- The GPU server was ~10 the cost to buy and 10 times the power to operate of a CPU only machine
- CPU servers were much more common and general purpose
- Unless your application was 10-20 times faster on the GPU it didn't make economic sense

Now

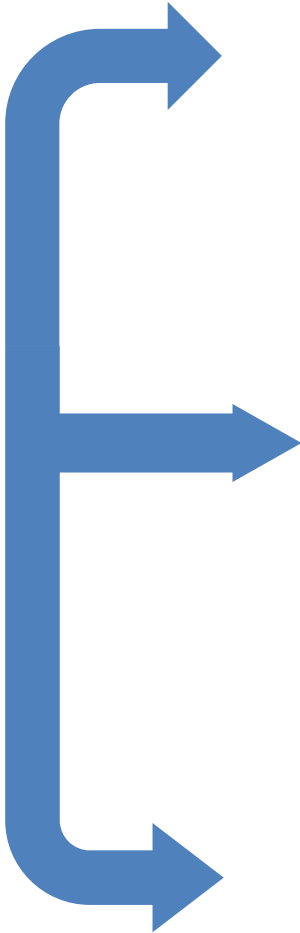
- ~4 times the cost to buy
- CPU server are still more common and general purpose but have much lower investment from industry
- Now a factor of 3-5 improvement is sufficient

Possible Evolution

It is very hard to buy computers this year

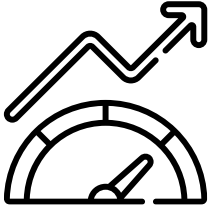
- The shutdown for HL-LHC is well timed

What will it look like moving forward



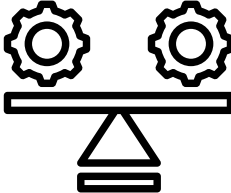
The AI investment accelerates

- Components get even more expensive



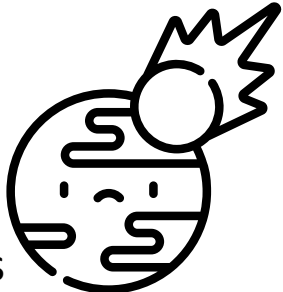
It stops getting worse

- They build new fabs
- This becomes the new pricing

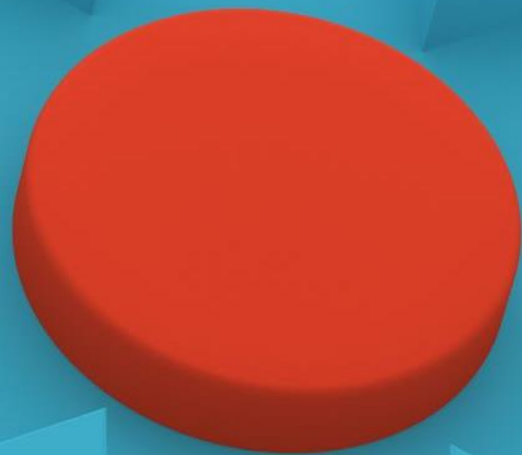


The AI bubble completely bursts

- A lot of companies go out of business
- We buy computing capacity at a steep discount



Impact

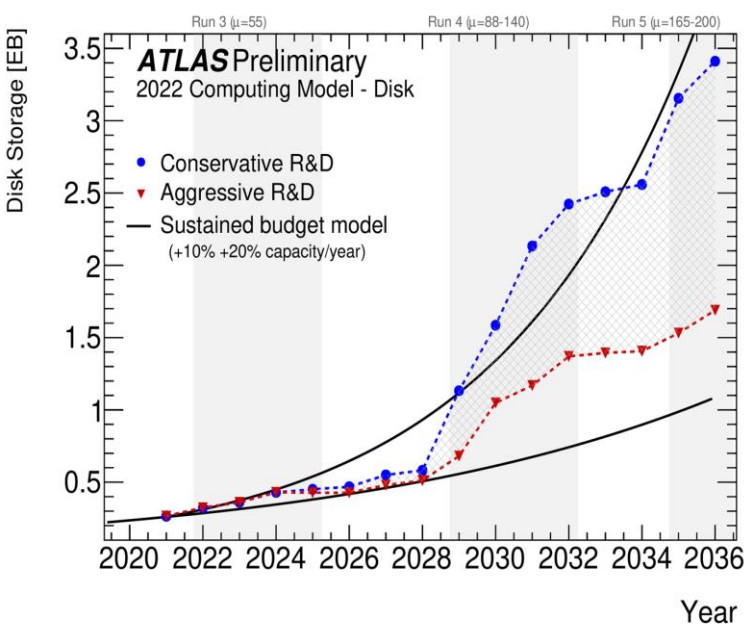
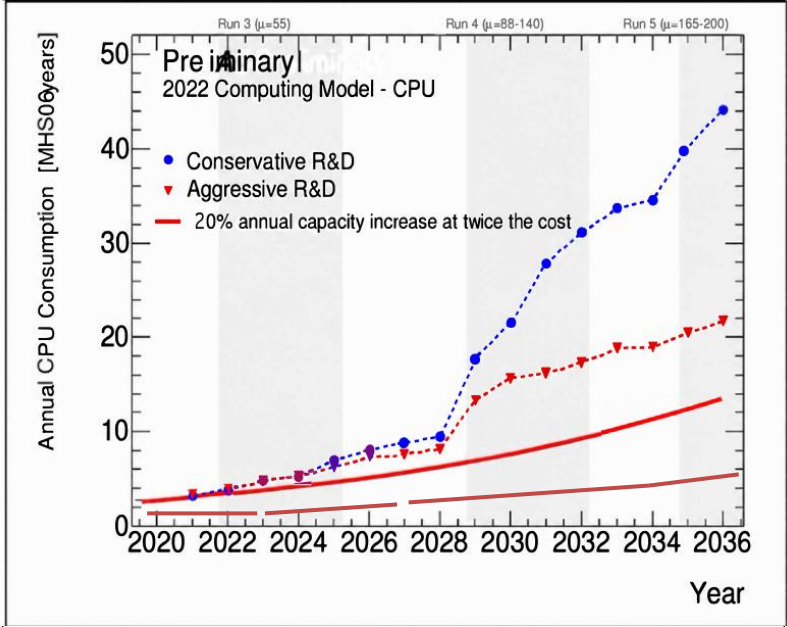
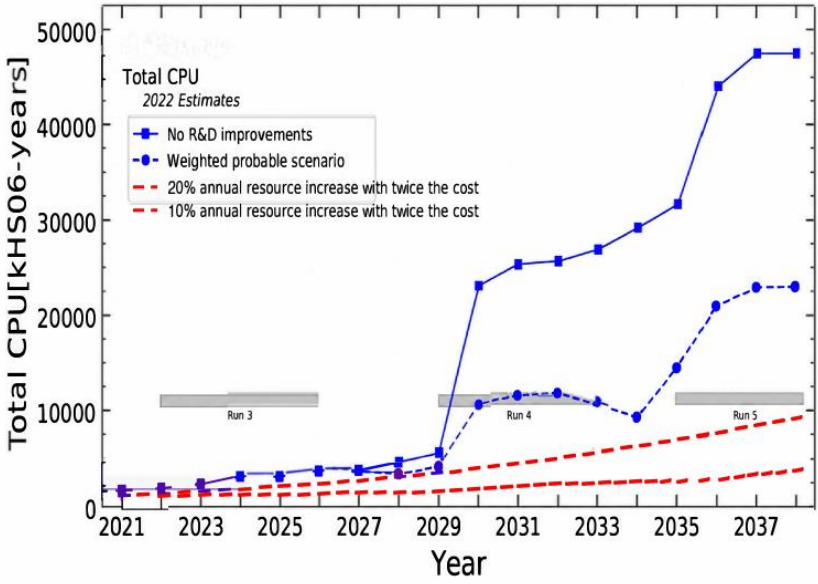


New Status

- The computers that have sustained the program for 3 decades have become twice as expensive in the last six months
 - Driving by components also needed for AI
 - HEP needs ~3GB/core
- The processors we depend on are evolving the slowest
 - Worse efficiency
 - Smaller investments
 - Slower evolution

Impact: Resources

- The experiments already have an aggressive R&D to try to fit into 10% and 20% evolution curves
 - If the computers are twice as expensive, then the lower line is maximum one would expect even with 20% evolution
 - We would not have enough resources without more money, or a change of direction



Impact: Replacement Models

With a significant increase in cost and a general slowing of improvement the model for replacement needs to be revisited

8-10 year cycles for CPU based systems

- Need to understand how to operate and maintain systems for longer
 - Migration to direct liquid cooling
 - Fewer moving parts, higher operational efficiency, lower failure rates due to more consistent temperatures
- User expectations on what is considered an old system
- Increase hosting capacity because we will be running a larger mix of old and new systems

Transitions

I am reminded of a previous time where CERN was dependent on an old and slowly evolving paradigm, which was getting expensive

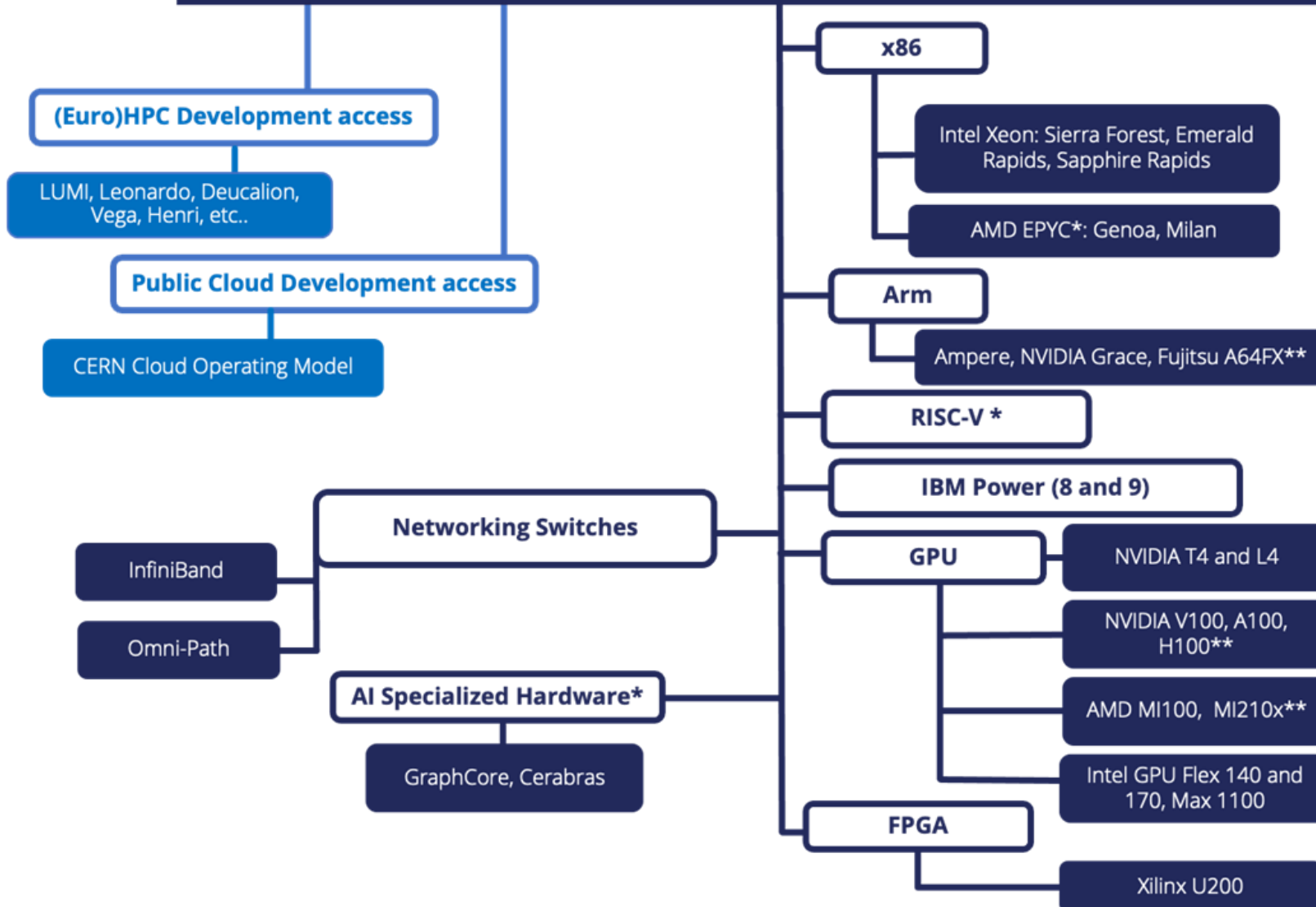
- Research projects between industry and science like CERN openlab helped facilitate the transition



Impact: Flexibility

- We could ask for more money
 - Asking for money because you don't want to evolve is not historically a winning strategy
- We need to be more flexible on the hardware architectures
 - Access to more efficient and more rapidly evolving GPUs
 - Better alignment with industry investment
 - Ability to access HPC facilities and shared computing resources
- Need broader adoptions of portability libraries
 - We need to care less about the underlying hardware
- We need to understand how many of our applications can be cast as AI applications
 - This is where the investment is

Heterogeneous Architecture Testbed: Hardware



100+ users & 290+ accounts

~95 systems, mostly bare-metal

Used by ATLAS, CMS, LHCb, QTI, openlab, CERN EP, IT, ATS

~200 tickets handled p/a

*Remote access via E4

**Remote access via Simons Foundation

Impact

The calculus for when to replace systems needs to be completely redone

- **Current model assumes gear is more capable for the same amount of money**
- **In the current environment of rapidly rising costs with small increases, our old machines will be valuable for longer**

Prices are changing rapidly. In the US vendor quotes are now valid for 2 weeks and no one will commit to a price until it ships

- **Our whole bidding and request for proposals model no longer works**

Our estimates for what hardware costs needs to be revised

The cost benefit of GPUs needs to be recalculated

We are unlikely to be able to grow as much as we would like.

Outlook (1/2)

It's a difficult time

- **No science is a driver of the computing they rely on**
 - **We don't control where investments are made**
- **We don't influence the economics**
 - **We are tiny by comparison to hyperscalers, clouds, etc.**

AI is driving, and like any disruptive technology its driving rather recklessly

Outlook (2/2)

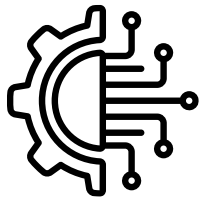


To sustainably sustain our mission something will have to give

- We will need to get more resources
 - It's not obvious that maintaining a flat budget is possible



- We will need to expand the time with fewer resources
 - Make allocations on shared resources



- We will need to expand what we can use for computing and where we work