Hardware Technology Trends in HEP Computing

Andrea Sciabà

Outline

- Introduction
- Technology tracking at CERN and in HEPiX
- WLCG evolution and technology
- Trends in semiconductor industry
- Data center infrastructure
- Processing units and memory
- Flash, disk and tape
- Network
- More on sustainability
- Conclusions

Introduction

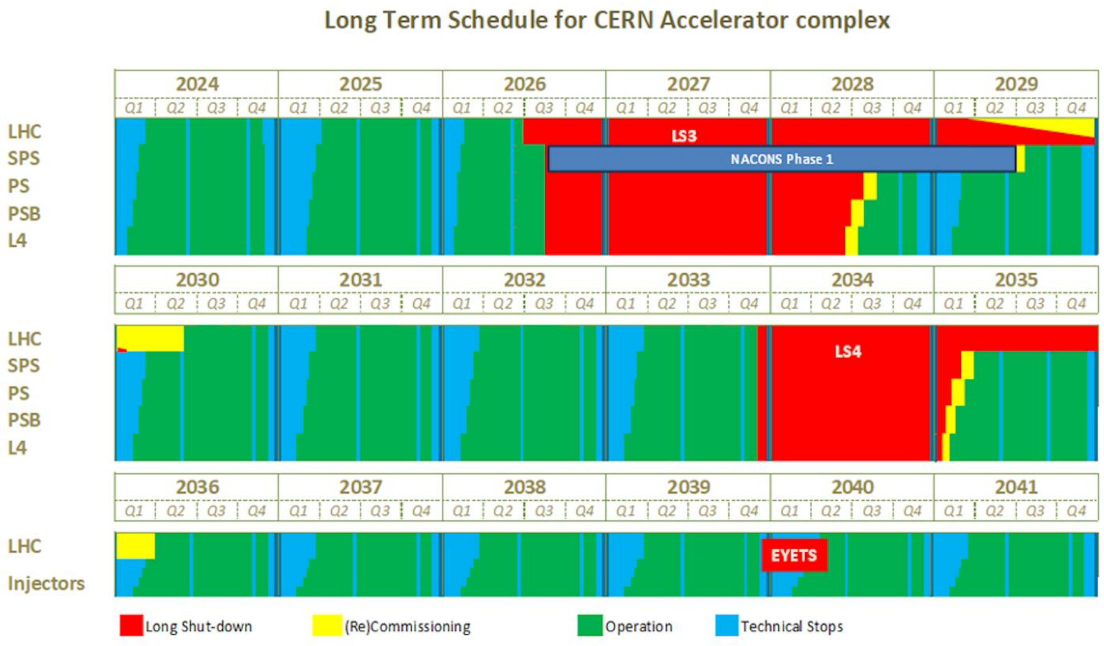
- **What technology?**
  - Hardware that matters for scientific computing (= physics experiments)
  - Changes in technology may have a profound impact on data processing and analysis
- **What scientific computing?**
  - Mostly, but not limited to HEP (e.g., gravitational waves)
  - Typical applications: event generation, simulation, reconstruction, data analysis, DAQ, trigger, etc.
  - AI algorithms increasingly used
- **Two main (and quite different) domains**
  - Offline processing: HTC workloads: lots of CPUs, some GPUs, lots of storage
    - Dedicated data centers (like WLCG sites) and HPC centers
    - Not much use for very expensive/exotic solutions
  - Online processing: CPUs, but also GPUs and FPGAs, very high bandwidth connections…
- **Some more typical HPC applications**
  - Typically for theoretical physics

Technology tracking at CERN and in HEPiX

- **CERN: the IT department, the experiments, the accelerator complex**
  - Very different communities with different needs
  - Many direct contacts with vendors (frequent NDA-covered meetings)
  - CERN openlab coordinates several joint projects
  - CERN IT CTO team follows technology evolution
- **HEPiX is a community of people operating data centers used in HEP**
  - Bi-annual workshops to share knowledge about technology choices and practical experience
  - Runs a few working groups, one being the Technology Watch
    - Participants choose the areas closest to their interests and experience
    - Delivers reports at various events (HEPiX or WLCG workshops, conferences, etc.)

WLCG evolution and technology

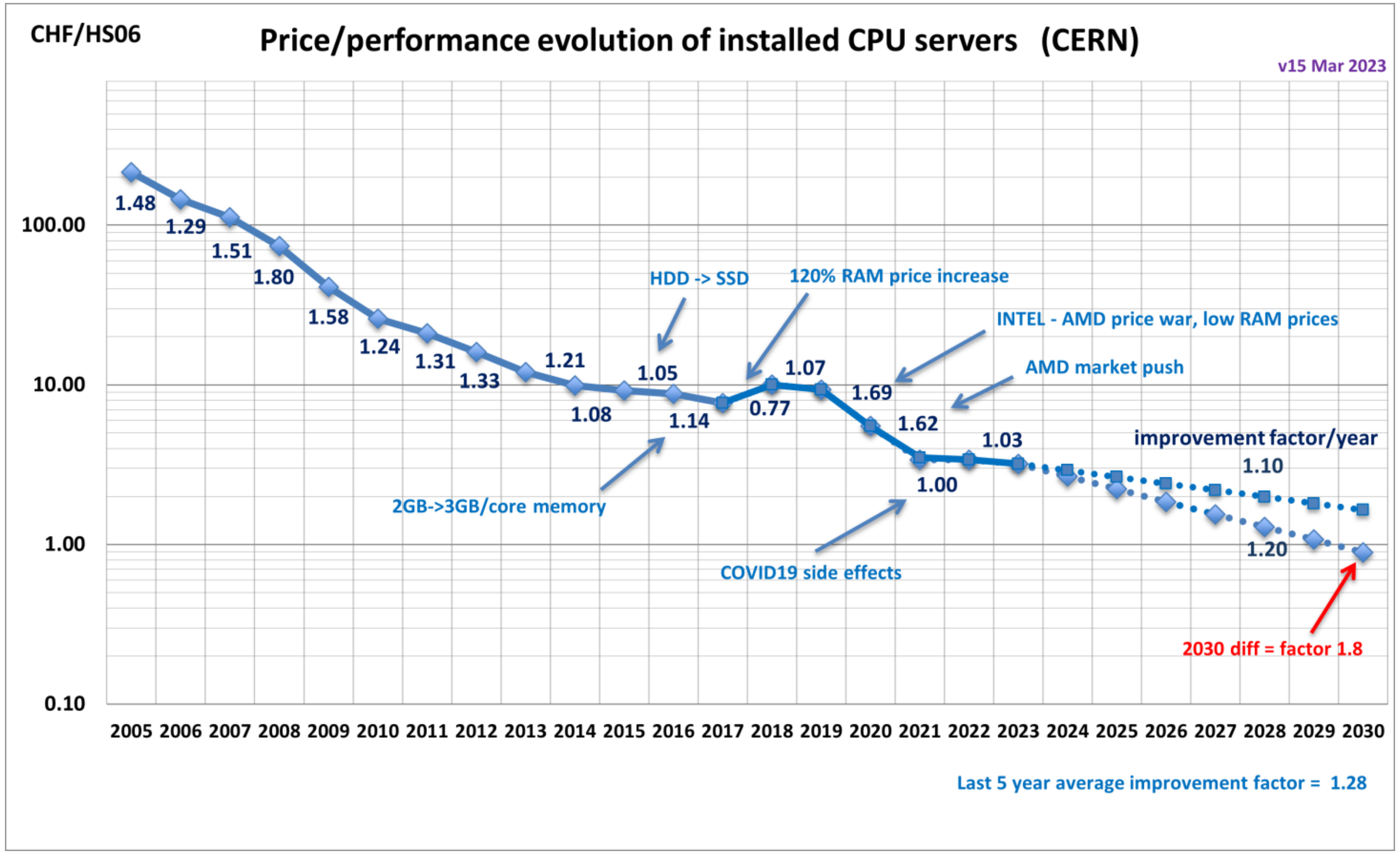
- **WLCG needs a good understanding of technology evolution for its medium-long term planning**
  - Cost of computing, operations, energy usage and efficiency
  - Given the experiment requirements, find the most cost-effective technologies fulfilling them!
- **A five years gap to prepare for HL-LHC**
  - Extremely difficult to make sensible predictions on this timescale, but we need to try



Source: CERN

CPU cost extrapolation at CERN

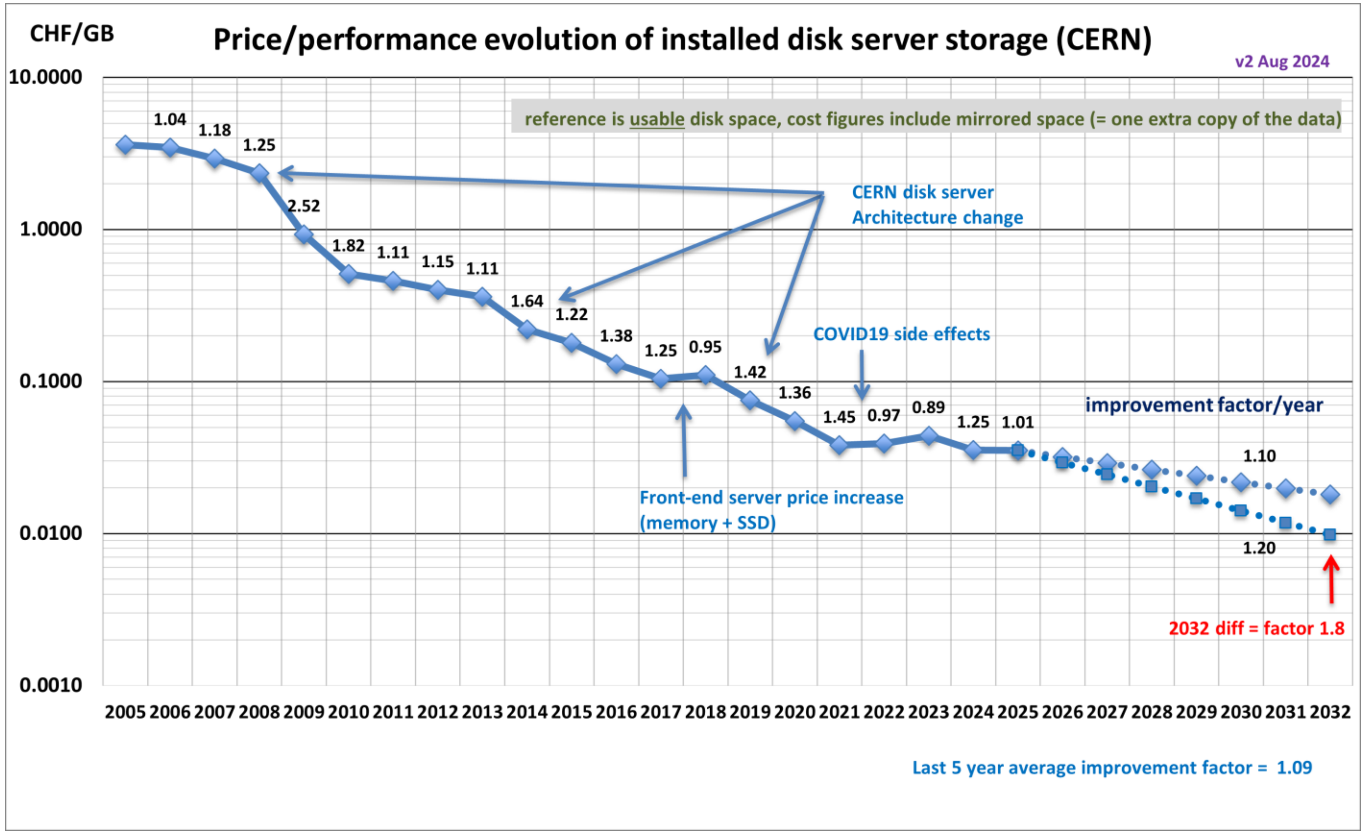
- **CPU price decrease rate relatively stable over time**
- **AMD becoming again competitive gave a boost**
- **Stagnant recently**
- **Questions:**
  - Is Intel going to be competitive again?
  - When (if) will Arm start having an impact? Or RISC-V?
  - When (if) will GPUs will make CPUs less relevant?



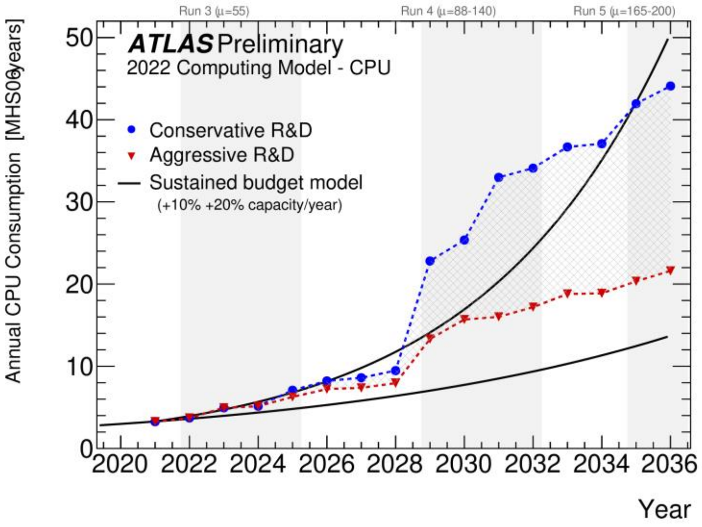
Source: B. Panzer-Steindel

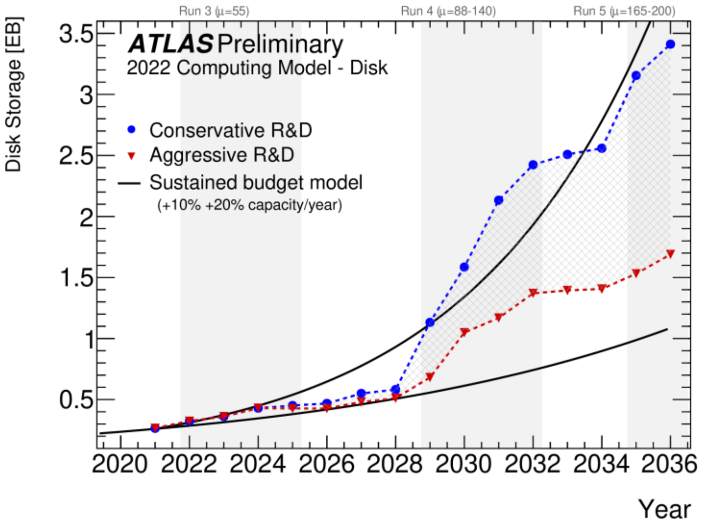
Disk cost extrapolation at CERN

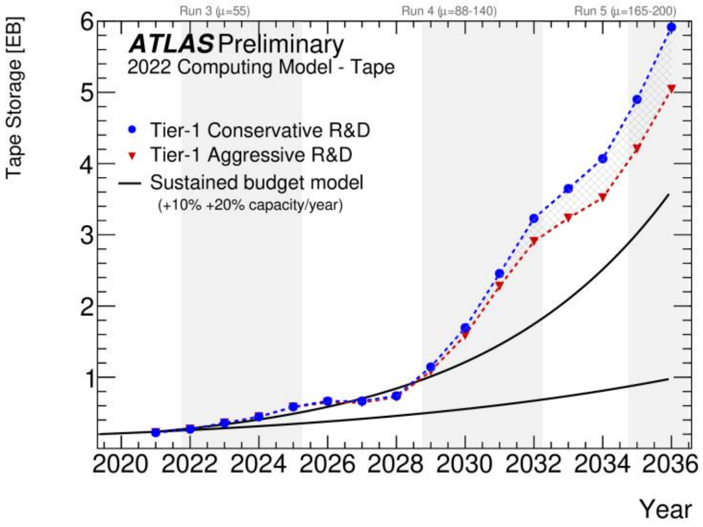
- **Stable price decrease but also flattening out recently**
- **Questions:**
  - When will energy-assisted magnetic recording HDDs will become prevalent?
  - AI demand driving up demand for all storage. Any hope to get better prices in the next 2-3 years?
- **What about SSDs?**
  - Not expected to completely replace HDDs, but usage will certainly increase and impact the overall cost of storage



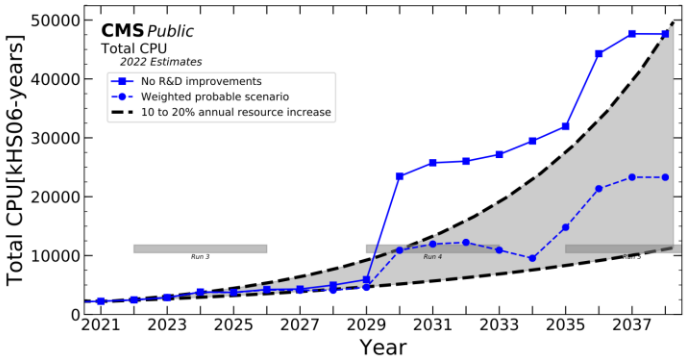
ATLAS and CMS resource needs up to HL-LHC

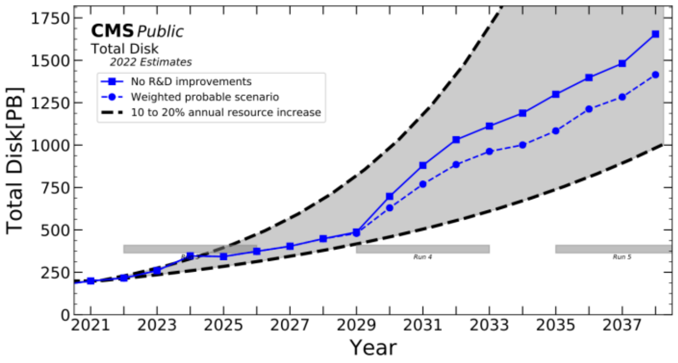


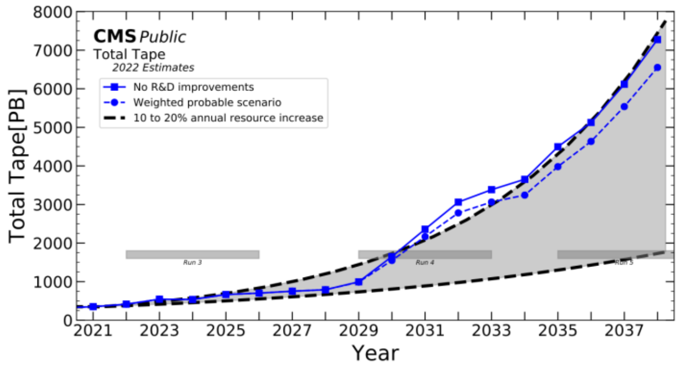




Sources: ATLAS and CMS







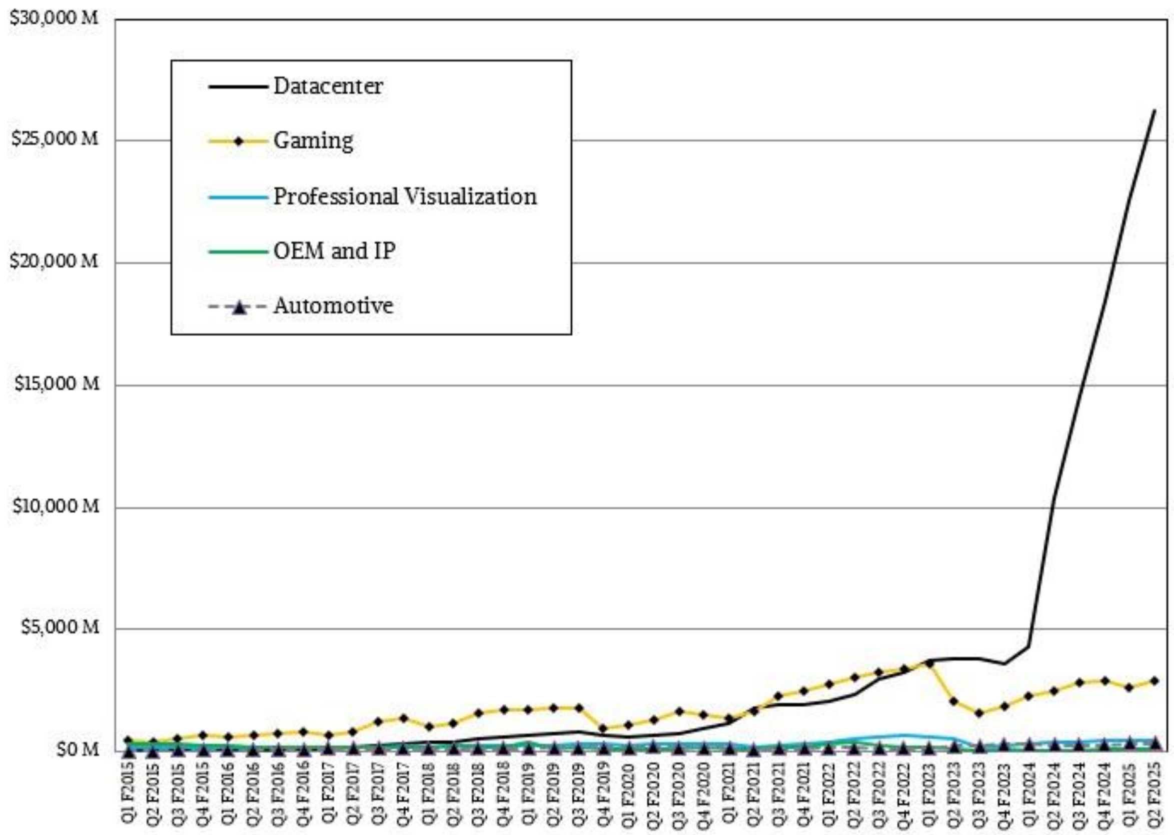
Challenges in resource requirements

- **From the experiment plots, with a flat budget, a 15%/year price reduction is required even in the most optimistic case (for CPU and disk)**
- **Situation much better than in the past, but still at the limit**
- **Stressing the importance of technology awareness**
  - i.e., how to get the most value out of the existing technology!
- **Will try to give an overview of the current state of technology relevant for HEP**
  - Just skimming the surface, as the topic is vast!

Market trends

Can we make predictions?

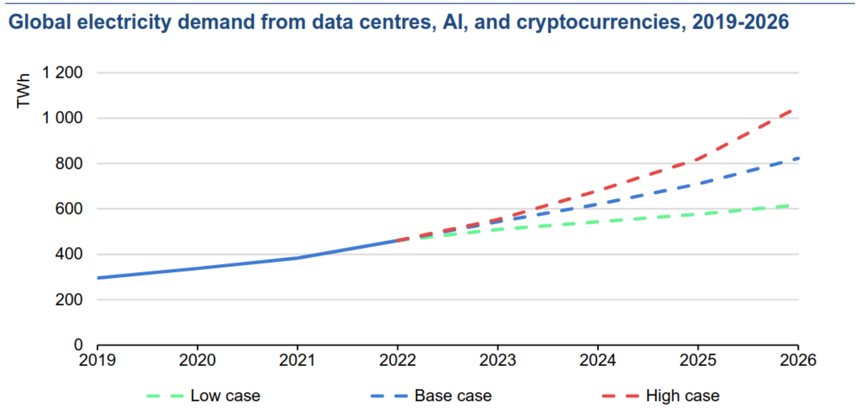
- **Extremely difficult even beyond just 1-2 years**
  - The demand can change unexpectedly. See the case of GPUs
  - E.g., will the AI bubble burst, suddenly? What if profits do not materialize early enough?
- **What are the “hottest” trends? Some examples:**
  - Sustainability and CO_2 emissions
  - Increasing memory bandwidth and latency requirements
  - Increasing power density and liquid cooling
  - Competition between spinning disks and flash memory



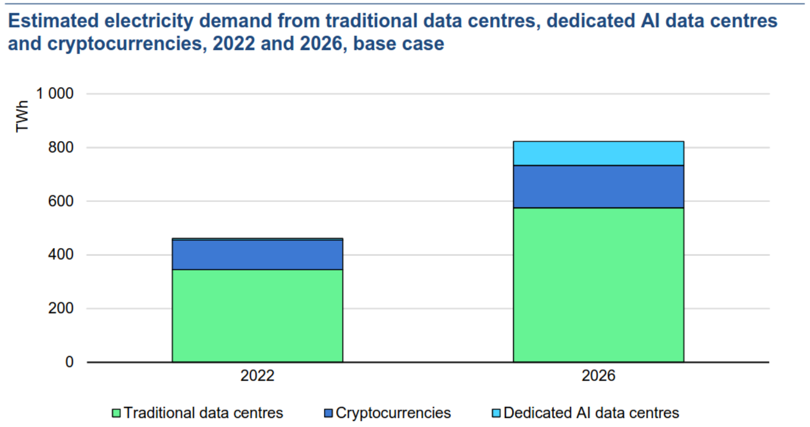
Nvidia revenues (Source: The Next Platform)

Power consumption

- **Data centers are proliferating due to the AI boom**
  - AI models increase by 1000x every three years
  - A single GPU uses ~4 MWh per year, Nvidia sold ~4 million in 2023! And soon there will be 2 kW GPUs…
  - Power infrastructure is a big constraint, strong incentive to energy efficiency
  - Data centers used ~2% of global electricity in 2022, estimated twice as much in 2026
- **The global IT market is increasingly focusing on sustainability**
  - This is becoming a hot topic also in our community
    - Dedicated WLCG workshop in December

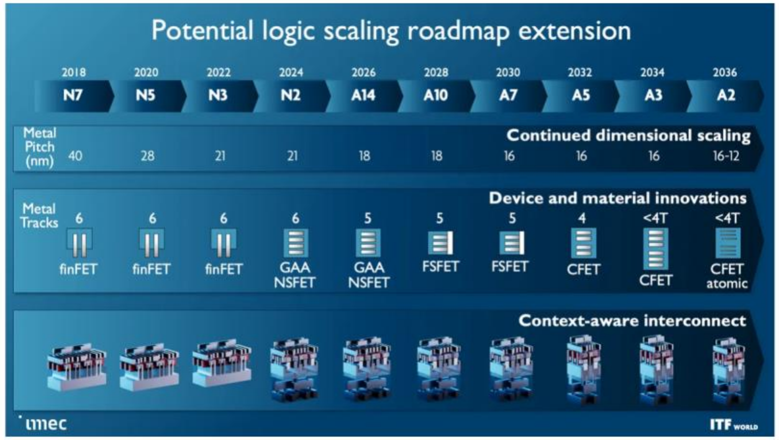


Source: IEA Report 2024

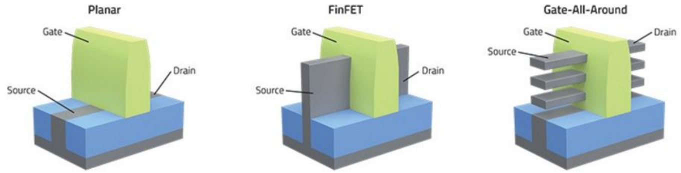


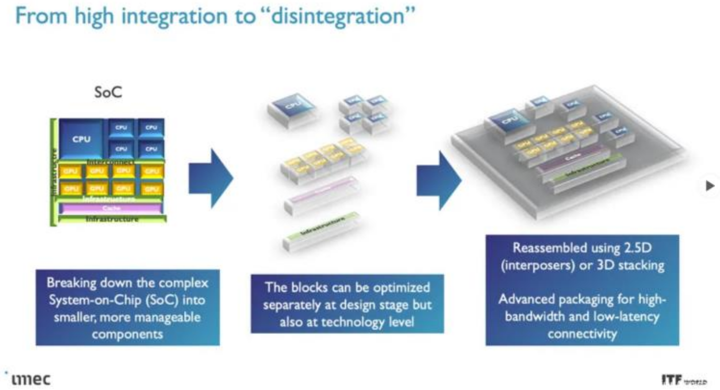
Roadmap for fabrication processes

- **Roadmap until 2036**
  - Transition from FinFET transistors to Gate All Around nanosheet designs for N2 and beyond (“Angstrom era”)
    - Less leakage, faster transistor switching
  - The current state-of the art is 3 nm, but yields are still low
- **Chiplet architectures**
  - Smaller nodes are more expensive, chiplets more economical than monolithic dies (possibly 3D stacked)
  - Voltages are not decreasing any more
- **Only three makers for leading edge chips - TSMC, Samsung, Intel**
  - Huge investments planned on fabs in diversified regions (Intel in NM, AZ, Israel, TSMC in AZ, Japan, etc.)



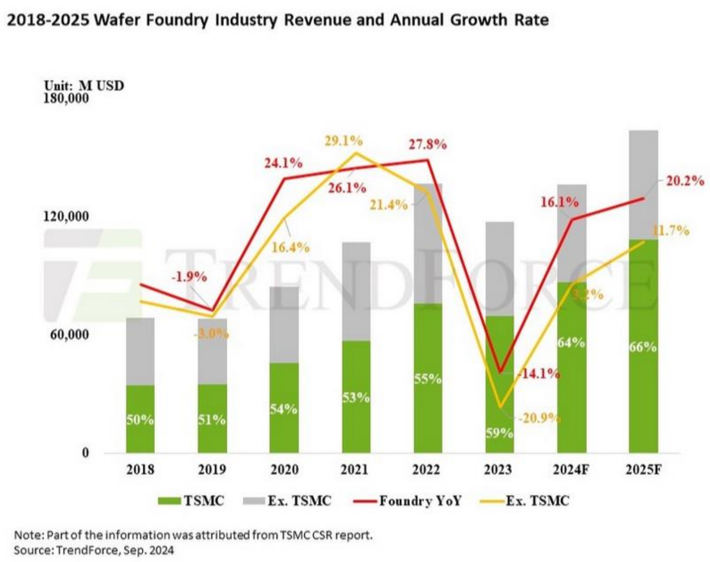
Source: Tom’s Hardware



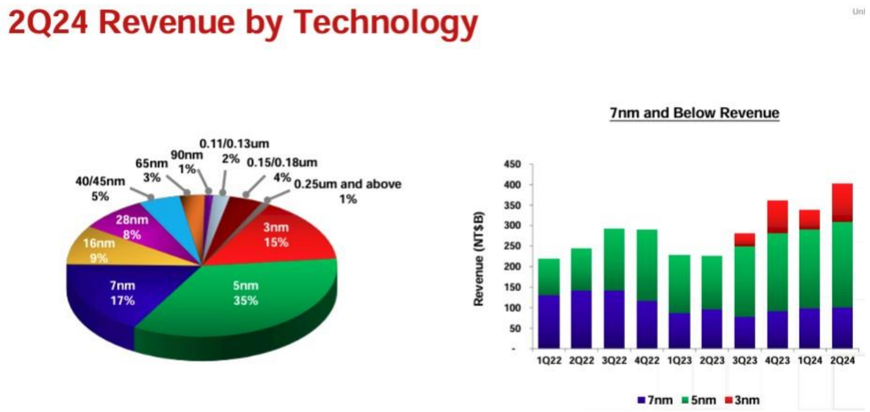


Foundry revenues

- **In 2024-2025 the market growth is primarily driven by the advanced processes**
  - TSMC is the industry leader with > 60% of the market
    - Makes all GPUs, all AMD and Apple CPUs
  - Consumer market is stagnant
- **Different nodes for different products**
  - 3 nm: high end CPUs
  - 5/4 nm: latest GPUs, ASICs
  - 7/6 nm: smartphone components



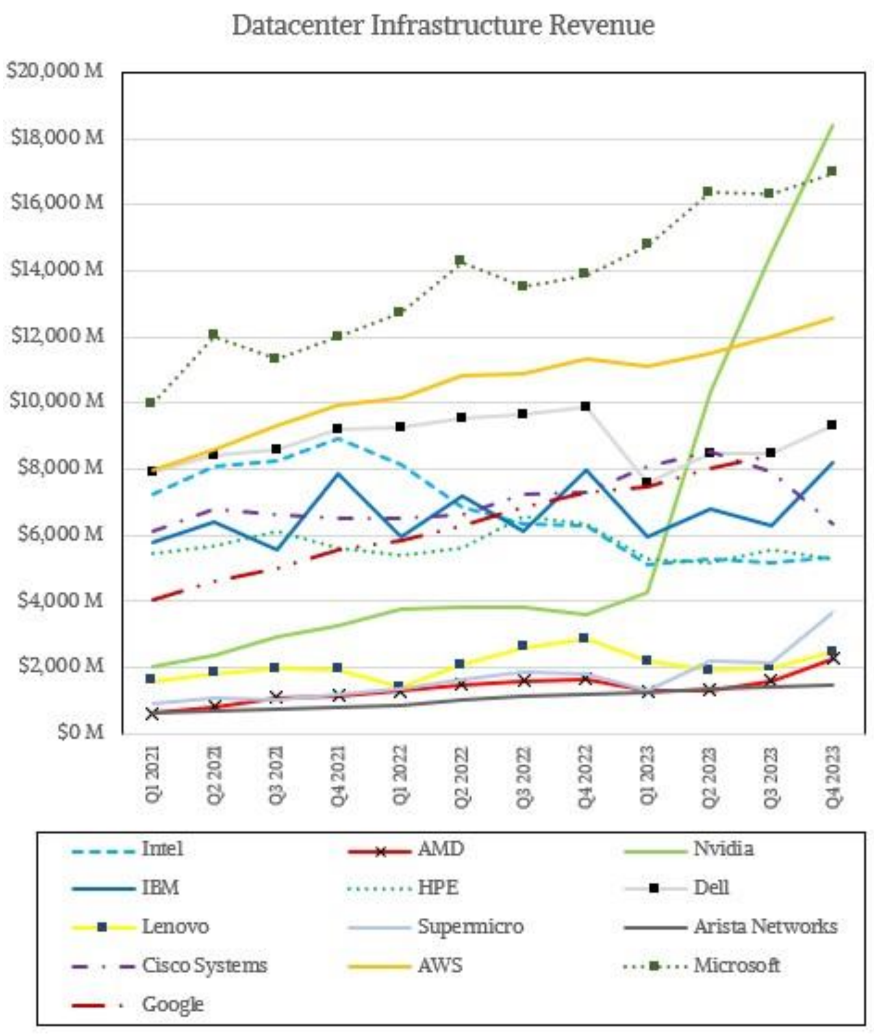
Source: Trendforce



TSMC Revenue for advanced nodes (Source: TheNextPlatform)

Comparing major players at scale

- **Revenues steadily increasing in the last few years, with a few exceptions**
  - Intel dropped quite a bit and lost most of its value, future uncertain
  - Nvidia skyrocketed in 2023
- **AI is the main driver and Nvidia has a practical monopoly**
  - AMD might increase their share, as demand is very high and have competitive products



Source: The Next Platform

Server Market

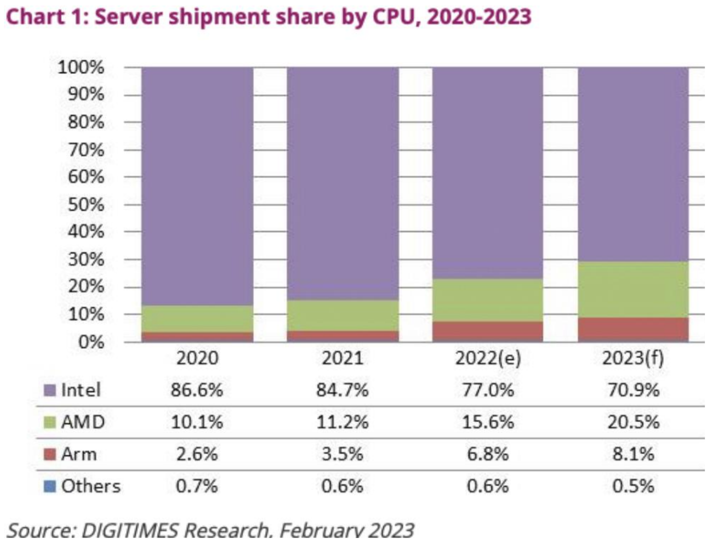
- **Server shipments are expected to slightly increase in 2024**
  - AI servers are 12% of the total and increasing at a much faster rate than general purpose servers
- **AMD quickly gaining ground**
  - 34% CPU server revenue market share in 2024
- **Arm also increasing, 3x in 3 years**
  - Best suited for hyperscalers and cloud providers
  - But Ampere's revenues are a tiny fraction of the total, and Nvidia Grace is very expensive
  - Still early for us to heavily invest on it

Share of AI servers shipped

Company	2022	2023	2024F
NVIDIA	67.6%	65.5%	63.6%
AMD (inl. Xilinx)	5.7%	7.3%	8.1%
Intel (inl. Altera)	3.1%	3.0%	2.9%
Others	23.6%	24.1%	25.3%
Total	100.0%	100.0%	100.0%

Source: TrendForce, Jul., 2024

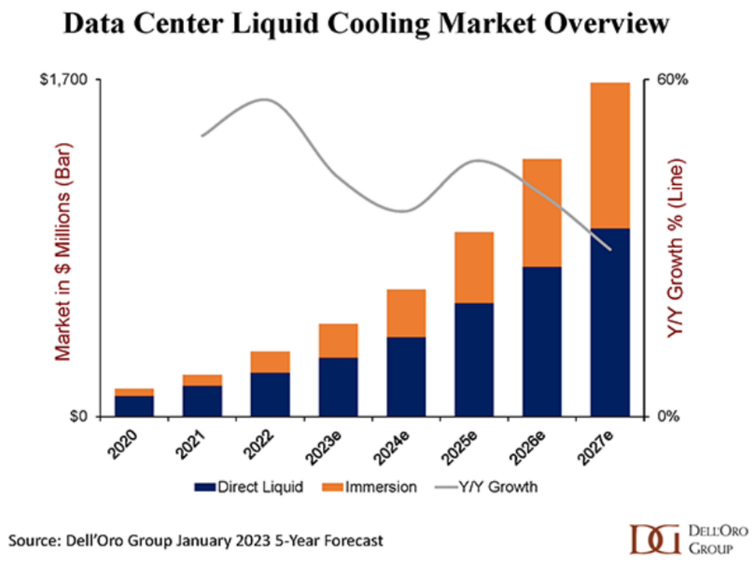
Source: Trendforce



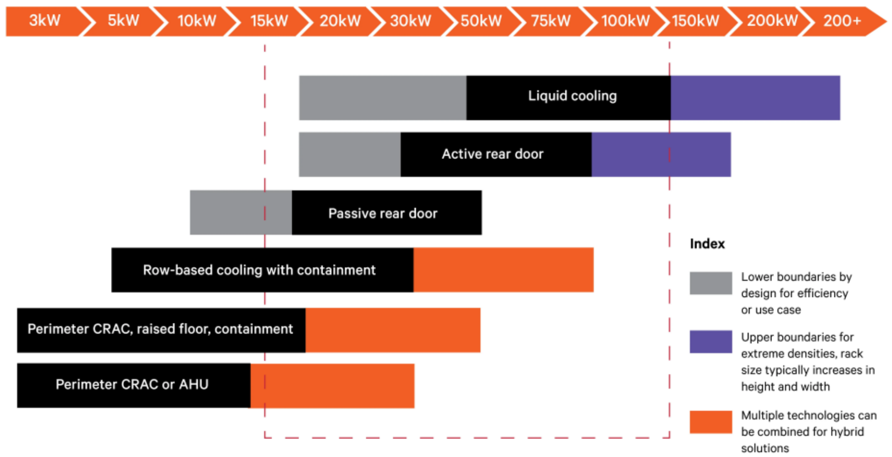
Source: Tom's Hardware

Server designs

- **High core counts make single socket servers a very interesting option**
  - Simpler, cheaper, use less power
- **Arm servers are becoming a viable alternative**
  - Better power efficiency than x86 (more on this later)
- **Liquid cooling destined to become mainstream**
  - Liquid cooling starts making sense from 30 kW per rack
  - With next-gen 500+ W CPUs, 1U systems will become rare and 2U or bigger will become the standard
  - No standard yet for liquid cooling, hopefully one will emerge in a few years
  - Some centers like NIKHEF (or experiments like LHCb) are studying liquid cooling solutions that can fit in existing air-cooled data centers



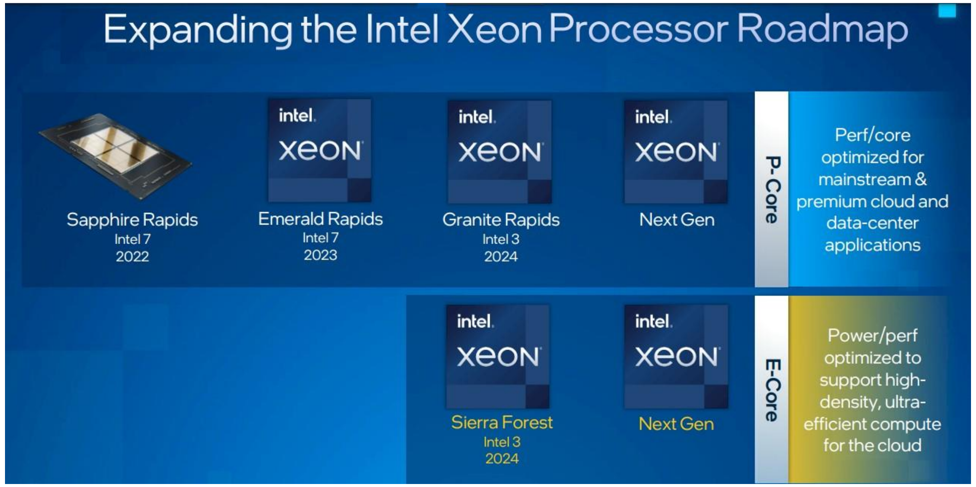
Source: Vertiv

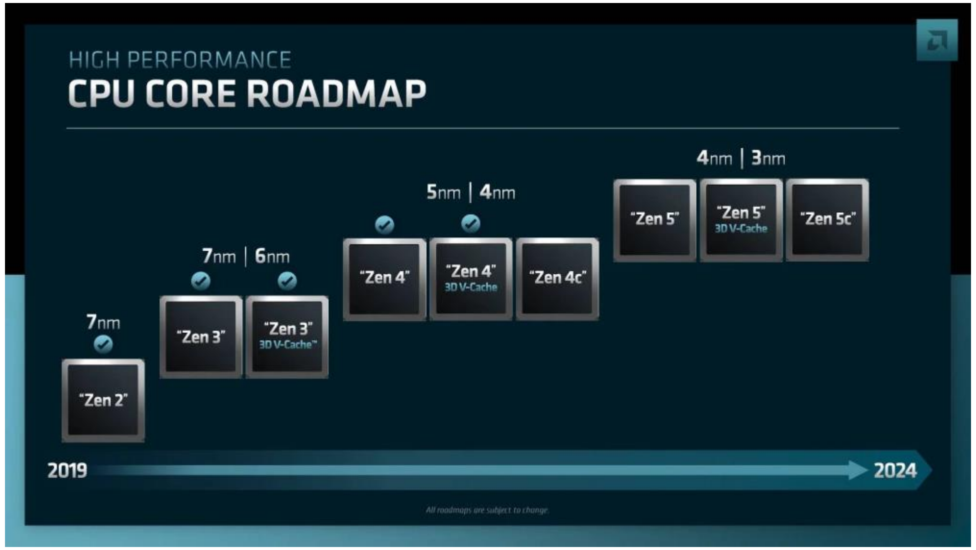


Processing units

x86 CPUs

- **CPU models being segmented in two categories**
  - **HPC**: higher frequencies, AVX512 support, multithreading
    - Intel Granite Rapids (P-cores) and AMD Zen4/Zen5
  - **Cloud**: focus on performance/Watt but lower performance
    - Intel Sierra Forest (E-cores), AMD Zen4c/Zen5c
- **The core count keeps increasing (200+)**
  - Monolithic dies not an option, chiplets to overcome scalability limits and improve yields
    - Can use different nodes for logic, cache and I/O, for cost optimization
  - TDP reached 500 W/socket this year
- **AMD leading on cost and power efficiency since a few years already**
  - More than half of the CPUs in WLCG are from AMD!
  - First information about Turin shows a large increase in performance/$
- **Intel trying to catch up with their new 2024 architectures**
  - First benchmarks indicate that their latest CPUs are more competitive





Current and upcoming x86 CPU generations

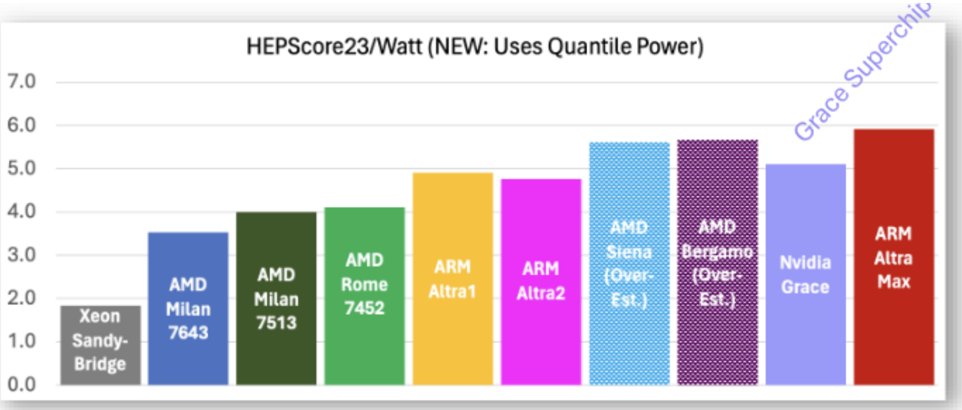
	AMD 4th gen EPYC "Genoa"	AMD 4th gen EPYC "Bergamo"	AMD 5th gen EPYC "Turin"	Intel 5th gen Xeon (Emerald Rapids)	Intel 6th gen Xeon "Sierra Forest"	Intel 6th gen Xeon 6 "Granite Rapids"	Intel "Clearwater Forest"
Launch	2022 Q4	2023 Q3	2024 Q4	2023 Q4	2024 Q2	2024 Q3	2025
Node	TSMC N5	TSMC N5	TSMC N3	Intel 7	Intel 3	Intel 3	Intel 18A
Max Cores	96 Zen4	128 Zen4c	128 Zen5 192 Zen5c	64	144 (288 next year) E-cores	128 P-cores	288?
Max L3 cache	384 MB	256 MB	384 MB	320 MB	108 MB	504 MB	?
Max TDP	360 W	400W	500 W	350 W	500 W	500 W	500 W?
Memory	12 ch DDR5 up to 4800 MHz	12 ch DDR5 up to 4800 MHz	12 ch DDR5 up to 6400 MHz + CXL 2.0	8 ch DDR5 up to 5600 MHz	Up to 12 ch DDR5-6400 CXL support	Up to 12 ch DDR5-6400 MCR-DIMM and CXL support	?
I/O	Up to 160 IO lanes of PCIe-5	Up to 160 IO lanes of PCIe-5	Up to 160 IO lanes of PCIe-5	Up to 80 IO lanes of PCIe-5	Up to 96 IO lanes of PCIe-5	Up to 96 IO lanes of PCIe-5	?

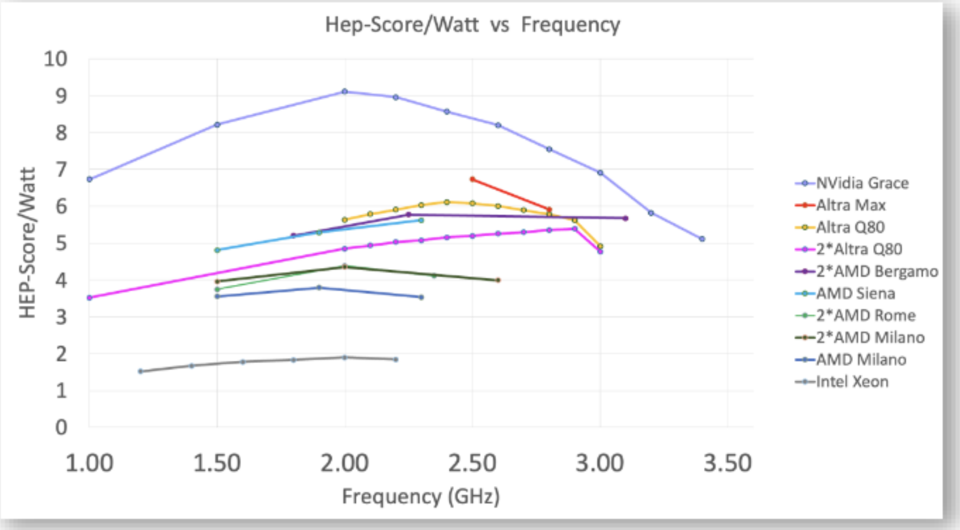
- **Choosing a specific CPU model is not easy**
  - WLCG has HEPSCORE23 as primary tool to measure performance
  - It will be interesting to determine if E-cores can be a viable option for HEP
  - Intel CPUs feature several built-in accelerators (AVX, AMX, IAA, DSA, DLB, QAT, …), probably not very useful to us

CPU	Best HS23/phys. cores
AMD Genoa	43
Intel Sapphire Rapids	40
AMD Milan	36
Nvidia Grace	32
Ampere Altra	16

What about Arm?

- **Arm entered the server market in 2018 and sells CPU designs**
  - Neoverse N1 in 2019: Ampere Altra (80 cores), AWS Graviton2
  - Neoverse N2 in 2020: Microsoft Azure Cobalt
  - Neoverse V1 in 2020: AWS Graviton3
  - Neoverse V2 in 2022: AWS Graviton4, Nvidia Grace, Google Axion
  - AmpereOne in 2024: up to 192 cores (custom design)
- **Nvidia main competitor with the Grace CPU**
  - Combined with the Hopper GPU or two CPUs in a "superchip" (144c)
- **Ampere CPUs already deployed at a few WLCG sites**
  - Notably Glasgow, published several efficiency measurements and comparisons with x86
  - Clearly better power efficiency!
- **Not quite yet a valid alternative for WLCG**
  - Not all experiment workloads are validated on Arm
  - Ampere is a very small company, Grace is very expensive
  - Both Intel and AMD are pushing strongly on power efficiency
- **And RISC-V??**
  - Not yet a viable platform for WLCG, but it is ramping up fast!





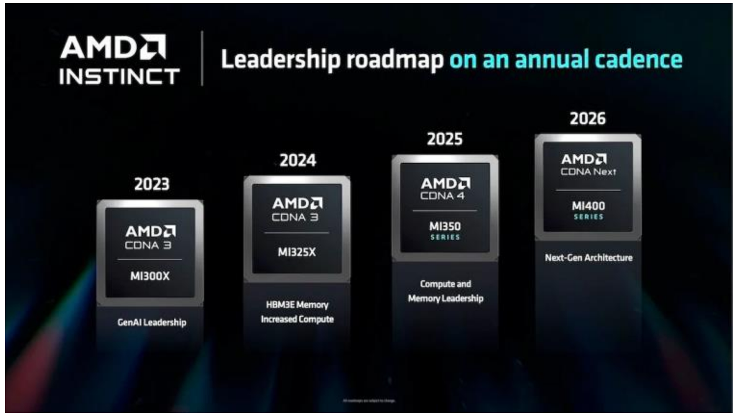
David Britton, University of Glasgow GDB, June 2024

GPUs and accelerators (1/2)

- **GPU usage in HEP quickly gaining momentum, but so far mostly on dedicated facilities (HPCs, HLT farms, analysis facilities, …) or for R&D**
  - WLCG sites should carefully clarify the needs of their experiments before buying large amounts of GPUs
- **Currently, almost an Nvidia monopoly, but AMD is gaining ground**
  - Nvidia Hopper (H100/H200), Blackwell (B200) later this year
  - AMD MI300X found to be competitive with the H100 on AI workloads, but Nvidia has better software support
  - Intel still marginal (Gaudi 2/3 are just AI accelerators, Ponte Vecchio is already old)
  - Cloud-native accelerators from AWS, Google, Microsoft, etc.

NVIDIA Data Center / AI GPU Roadmap

GPU CODENAME	X	RUBIN	BLACKWELL	HOPPER	AMPERE	VOLTA	PASCAL
GPU Family	GX200	GR100	GB200	GH200/GH100	GA100	GV100	GP100
GPU SKU	X100	R100	B100/B200	H100/H200	A100	V100	P100
Memory	HBM4e?	HBM4?	HBM3e	HBM2e/HBM3/ HBM3e	HBM2e	HBM2	HBM2
Launch	202X	2025	2024	2022-2024	2020-2022	2018	2016



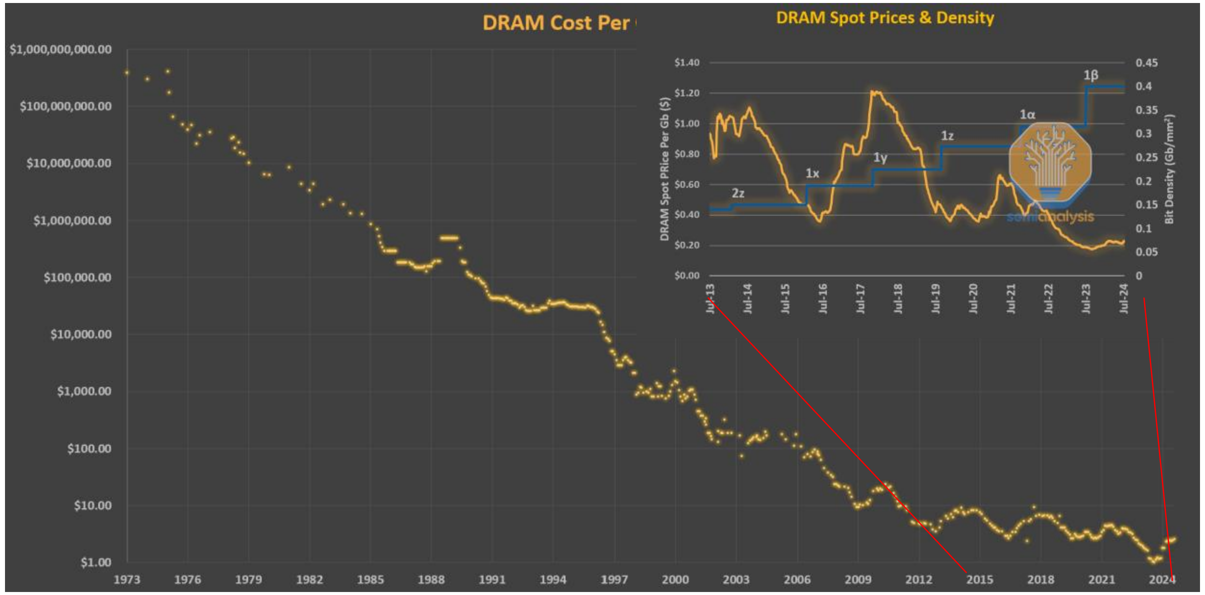
GPUs and accelerators (2/2)

- **CPU+GPU in a single package**
  - Nvidia GraceHopper
  - AMD MI300A
  - Intel Falcon Shores
  - PCIe slotted cards are much less relevant
- **Performance evolution is not going in a direction we like**
  - FP32/FP64 performance will not increase much, or at all
  - AI performance (FP16, FP8, INT8) has priority because is where most of the money is made!

	MI300X	H100 SXM	H200 SXM	Blackwell (B100 SXM)
Process	TSMC 5nm+6nm	TSMC 4N	TSMC 4N	TSMC 4NP
TBP	750 W	700 W	700 W	700 W
Memory	192 GB of HBM3	80 GB of HBM3	141 of HBM3e	192 GB of HBM3e
Memory bandwidth	5.3 TB/s	3.35 TB/s	4.8 TB/s	8 TB/s
FP64 matrix/vector	164/82 TFLOPS	67/34 TFLOPS	67/34 TFLOPS	30 TFLOPS
FP8	2615 TFLOPS	1979 TFLOPS	1979 TFLOPS	7 PFLOPS

Memory

- **DRAM is not scaling any more**
  - Bit density just doubled in the last 10 years
  - Stuck at the 10 nm node
  - Severely lagging behind logic, which improves by 30%-40% every 2 years
- **Price going down, but slower and slower and with wild fluctuations**
  - Being the technology static, prices just depend on demand and supply
- **All DRAM varieties share the same memory cell technology**
  - Differences only in packaging and circuitry
  - New memory technologies (FeRAM, MRAM, ecc.) are not a viable alternative, mainly due to cost

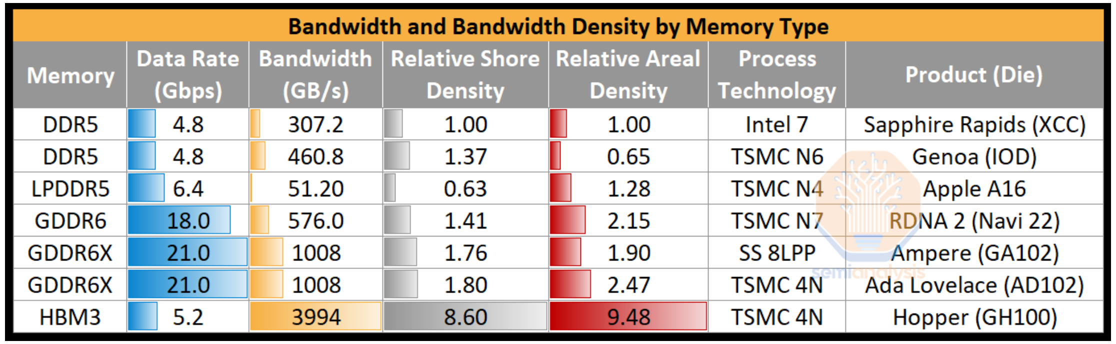


Source: Semianalysis

Memory

- **Strong push towards high bandwidth, low latency for HPC and AI**
  - Wide range of types of DRAM for different applications
- **System memory**
  - DDR5 current standard, up to 6400 MT/s
  - LPDDR5X much more power efficient, and much cheaper than HBM, but has limitations
  - MRDIMM (multi-ranked buffered DIMM) to achieve 8800 MT/s and more
    - MCR-DIMM (multiplexed combined ranks) is a similar solution, supported by Intel
  - CXL (Compute Express Link) is a protocol on top of PCIe that allows to disaggregate memory and share it with many CPUs and GPUs, combining DRAM and non-volatile storage
    - Unlikely to be very interesting for HEP offline computing, given our modest memory/core requirements, at least today, but being tested by some experiments

	Max capacity	Bandwidth	Bus width	Used on
HBM3	24 GB per stack	~820 GB/s per stack	1024 bit	Hopper, MI300
HBM3E	36 GB per stack	~1.3 TB/s per stack	1024 bit	Blackwell
GDDR7	64 Gbit per chip	160 GB/s	32 bit	RTX 50-series
MRDIMM	256 GB per module	8800 MT/s	64 bit	System

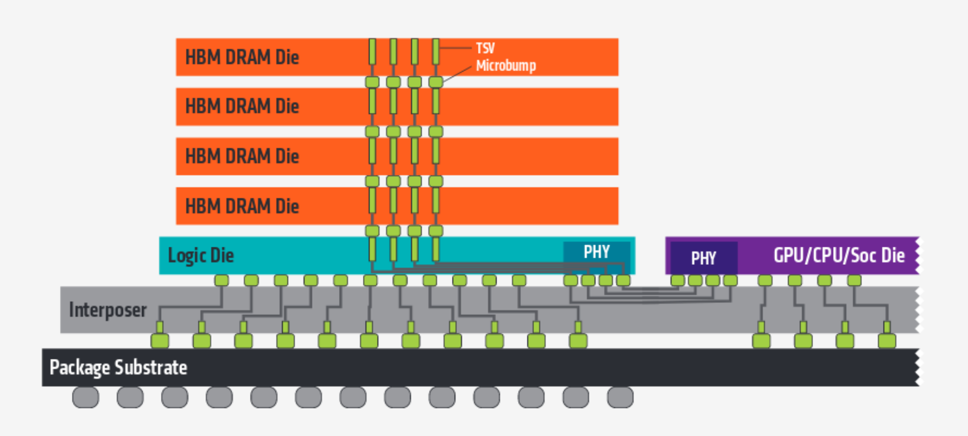

Bandwidth and Bandwidth Density by Memory Type

Memory	Data Rate (Gbps)	Bandwidth (GB/s)	Relative Shore Density	Relative Areal Density	Process Technology	Product (Die)
DDR5	4.8	307.2	1.00	1.00	Intel 7	Sapphire Rapids (XCC)
DDR5	4.8	460.8	1.37	0.65	TSMC N6	Genoa (IOD)
LPDDR5	6.4	51.20	0.63	1.28	TSMC N4	Apple A16
GDDR6	18.0	576.0	1.41	2.15	TSMC N7	RDNA 2 (Navi 22)
GDDR6X	21.0	1008	1.76	1.90	SS 8LPP	Ampere (GA102)
GDDR6X	21.0	1008	1.80	2.47	TSMC 4N	Ada Lovelace (AD102)
HBM3	5.2	3994	8.60	9.48	TSMC 4N	Hopper (GH100)

Semianalysis

High bandwidth memory

- **HBM memory in increasing demand**
  - Stacks of DRAM dies (up to 12), 1024-bit wide interface
  - Directly connected to the GPU (or CPU, FPGA)
  - Latest is HBM3e, total bandwidth per stack exceeds 1 TB/s
  - Only viable solution for large model AI accelerators
  - 3x more expensive than DDR5 (low yields due to complexity of stacking)
- **GDDR6X used for graphic cards**
  - Might also be used for HPC and AI
- **Memory market is clearly recovering after collapsing in 2022-23**
  - Memory shortages expected as HBM tends to eat up capacity at the expense of DRAM
  - Estimated to globally account for 10% of capacity (20-30% of market value) in 2025



1Q24 Global DRAM Manufacturers' Branded Memory Revenue Rankings (Unit: Million USD)

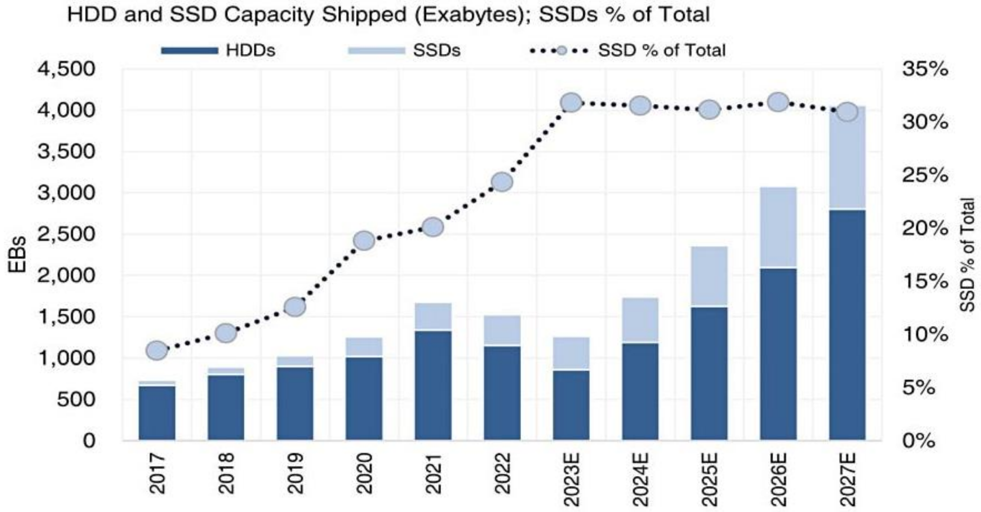
Ranking	Company	Revenue			Market Share	
		1Q24	4Q23	QoQ	1Q24	4Q23
1	Samsung	8,050	7,950	1.3%	43.9%	45.5%
2	SK hynix	5,703	5,560	2.6%	31.1%	31.8%
3	Micron	3,945	3,350	17.8%	21.5%	19.2%
4	Nanya	302	274	10.5%	1.6%	1.6%
5	Winbond	162	133	21.6%	0.9%	0.8%
6	PSMC	28	39	-28.2%	0.2%	0.2%
	Others	157	158	-0.6%	0.9%	0.9%
	Total	18,347	17,464	5.1%	100.0%	100.0%

Notes:
1. 4Q23—USD:KRW = 1:1,322; USD:TWD = 1:31.8
2. 1Q24—USD:KRW = 1:1,330; USD:TWD = 1:31.4
Source: TrendForce, Jun., 2024

Storage

Flash storage

- **Globally, shipped NAND flash capacity amounts to 30% of the total**
  - For HEP, it is much less: only as system drives and for certain high IOPS/bandwidth storage systems (data caches, tape buffers, analysis facilities…)
  - Price gap with HDDs is still 2-3x, slowly decreasing



Source: Gartner; Wells Fargo Securities, LLC. [Blocks and files]

- **Capacity increasing in two dimensions**
  - Bits/cell: SLC → MLC → TLC → QLC → PLC?, but at the expense of endurance
    - QLC for large SSD used for data serving
    - TLC and lower for high R/W rates
  - Number of layers: ~ 200-300 today, 400+ in 2025
    - But increasingly more expensive to make
  - Drive capacity soon to exceed 120 TB!

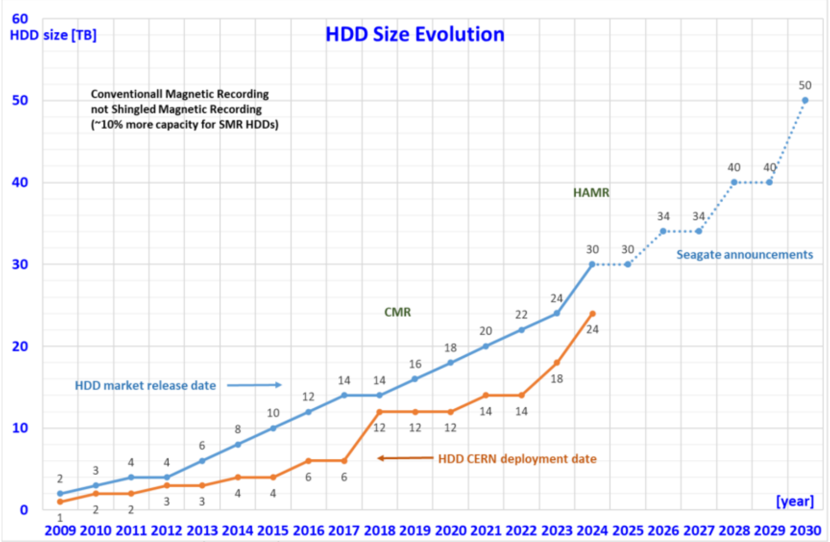
3D NAND Layer Cake

Micron		Samsung		SK hynix		SK hynix Solidigm		Western Digital/Kioxia		YMTC	
Generation	Layers	Generation	Layers	Generation	Layers	Generation	Layers	Generation	Layers	Generation	Layers
Gen 1	32	V3	48	V3	48	Gen 1	32	BiCS 2	48	Gen 1	32
Gen 2	64	V4	64	V4	72	Gen 2	64	BiCS 3	64	Xtacking 1 Gen 2	64
Gen 3	96	V5	96	V5	96	Gen 3	96	BiCS 4	96		
Gen 4	128	V6	128	V6	128	Gen 4	144	BiCS 5	112	Xtacking 2	128
Gen 5	176	V7	176	V7	176	Gen 5	192	BiCS 6 (Q1 2023)	162	2022 2H	196
Gen 6 (End 2022)	232	V8 (2022)	236	V7(2022 Q3)	238	*Gen 6?*	*?*	BiCS 7 - *skipped*	*212*	2023 Gen 4 Xtacking 3.0	232
Internal Gen 7 (2024) Branded G9 externally	276	V9 (Sep 2024)	286	*V8 (2025)*	*321*			BiCS 8 (2024)	218	*2024 H2*	*300-level*
Gen 8	*3xx?*	*V10 (2025 H2)*	*430*	*V9 (2026/7)*	*500+*			*BiCS 9 (2025)*	*300+*		
Gen 9	*4xx?*	*V11*	*>600*	*V10 (2030)*	*800+*			*BiCS 10*	*400+*		
		V12	*>800*								
		V13 (2030)	*1,000*								

[Supplier 3D NAND layer count generations]

HDD storage

- **Capacities still increasing, thanks to two technologies**
  - Shingled magnetic recording (SMR) gains 20-25% capacity but it's not transparent to software
  - Heat assisted magnetic recording (HAMR) drives have finally arrived (~50% of shipped capacity), but have limitations (e.g. very sensitive to vibrations)
    - Seagate shipping already 30+ TB drives in 2024, for now only in dedicated enclosures
  - 60 TB disks by 2028 using energy assisted magnetic recording?
- **Most of the capacity shipped is <u>nearline</u> HDDs**
  - We (HEP) still use them as our main online storage
  - Nearline drives are the last HDD holdout, but will not disappear any time soon
- **Performances <u>not</u> improving though**
  - Larger drives will exacerbate a bottleneck on IOPS and transfer rates



Source: B. Panzer-Steindel

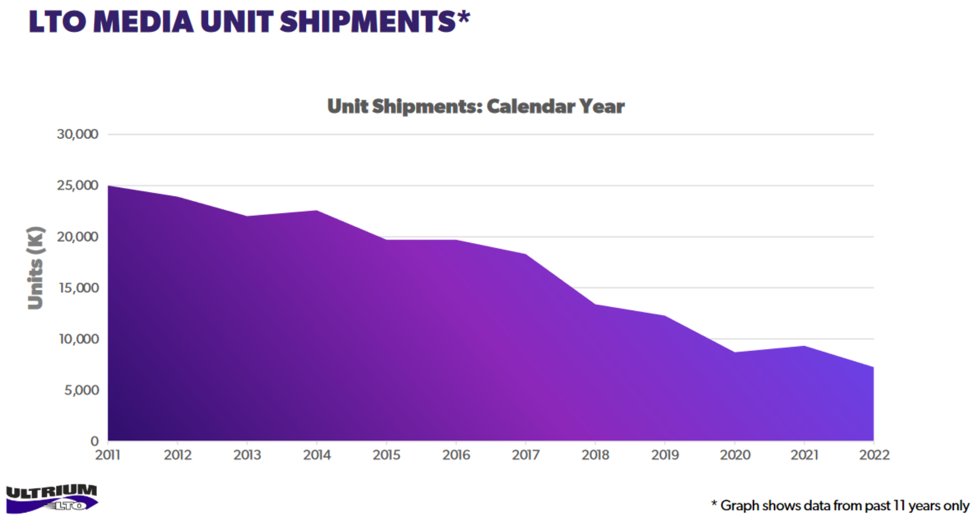
TABLE 2. Magnetic Mass Data Storage Technology Roadmap: HDD.

	Unit	2022	2025	2028	2031	2034	2037
Industry metrics							
Form factor (dominant form factor is bold)	Inches	3.5, **2.5**	3.5, **2.5**	3.5	3.5	3.5	**3.5**
Capacity	TB	1–22	2–40	6–60	7–75	8–90	10–100
Market size	Units (M)	166	173	208	249	299	359
Cost/TB (average)	$/TB	13.6	6.91	3.46	2.6	2	<2
Design/performance							
Areal density	Tb/in^2	>1	>2	>4	>6	>8	>10
Rotational latency	ms	2–12	2–12	2–12	2–12	3–12	3–12
Seek time*	ms	3–5	3–5	3–5	2–5	1.5–5	1–4
r/min		4.2–10K	4.2–10K	4.2–7.2K	4.5–7.2K	4.5–7.2K	4.5–7.2K

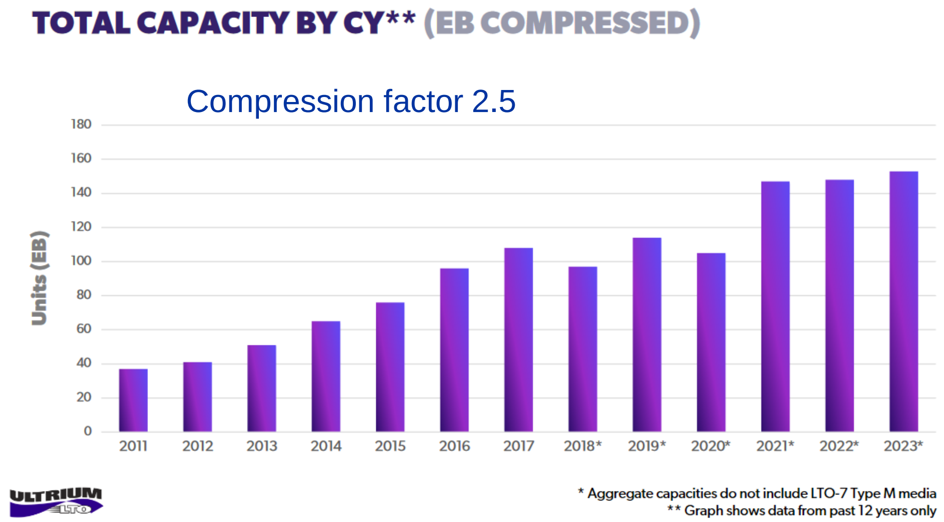
IEEE roadmap for Mass Digital Storage Technology

Archive storage

- **Magnetic Tape**
  - Still a lot of room for scaling (unlike HDD)
    - 30%-40% yearly increase in cartridge capacity
  - Lots of technology advancements in both media and drives
  - LTO the leading standard, smaller share for the IBM TS11XX format
  - Total LTO cartridges shipped has been declining, but total exabytes shipped is flat
- **Optical disk dead**
  - Panasonic and Sony discontinued Archival Disc drives and libraries
- **On the horizon**
  - Cerabyte “ceramic nano-memory” - Data etched in material via lasers
  - Glass storage (Microsoft’s project Silica), similar concept, very sparse information
  - DNA storage – Just too expensive to be practical
    - Might serve the “write once / read never-or-seldom” use case



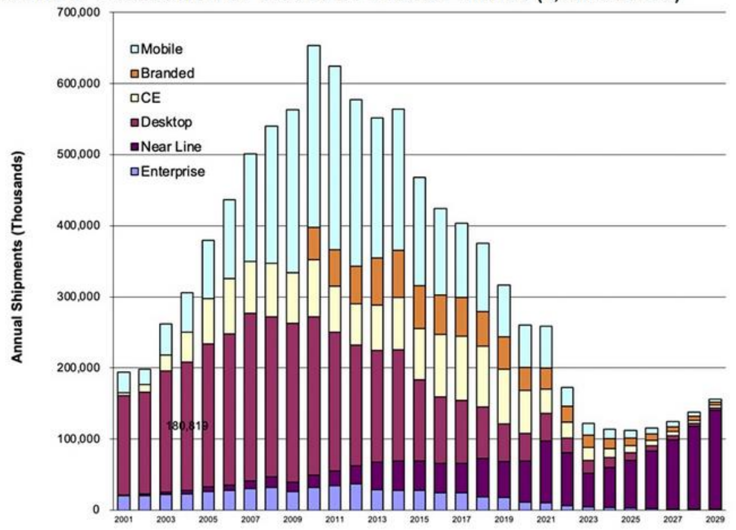
Source: LTO



Storage evolution summary

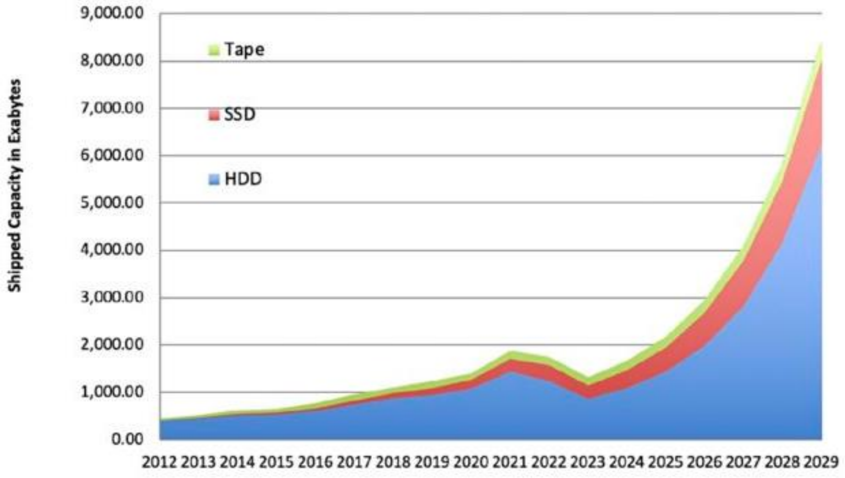
- **To summarize:**
  - AI boom drives volume increase for all types of storage
  - HDD shipments will soon be almost only nearline, but increasing in capacity
  - SSDs will not replace HDDs in data centers anytime soon
  - Our usage of SSD will probably increase to cope with the performance bottlenecks of HDDs
  - Tapes are not going anywhere either

FIGURE 7. PROJECTION OF DRIVES BY MARKET NICHES (1,000'S-UNITS)



Source: Blocks and Files

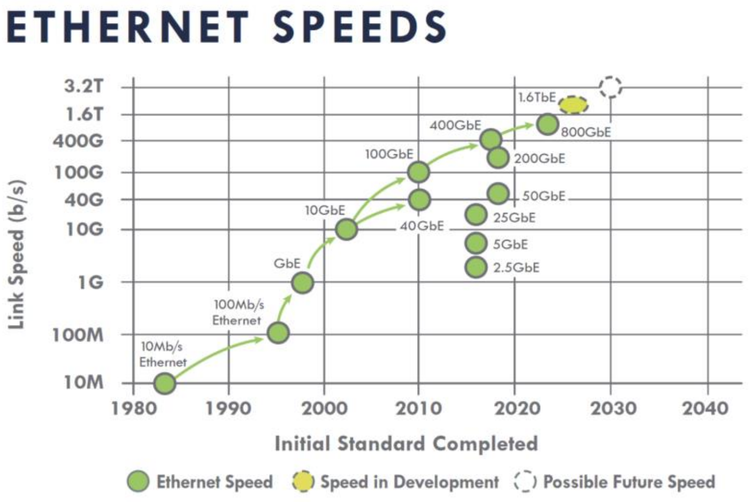
FIGURE 11. CAPACITY SHIPMENTS FOR LTO TAPE, SSDS AND HDDS



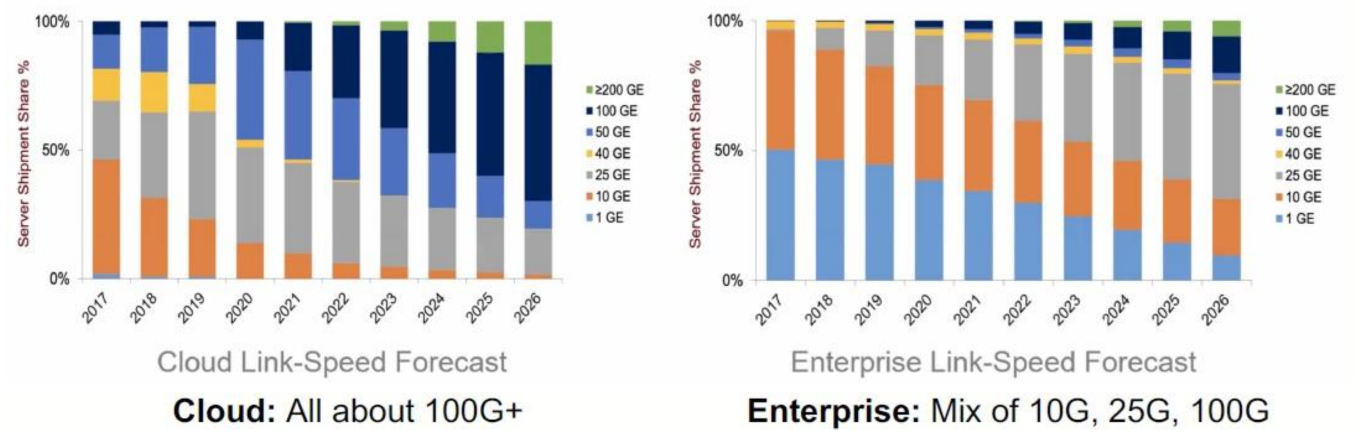
Network

LAN and interconnect technologies

- **InfiniBand provides high throughput/low latency networking**
  - Useful for AI and HPC simulations
  - Can provide Remote Direct Memory Access (RDMA)
  - Now controlled by Nvidia, cost may amount to up to 20% of an HPC cluster
  - RoCE (RDMA over Converged Ethernet) is a much cheaper alternative that works over Ethernet
- **Ultra Ethernet consortium aims at producing an alternative to InfiniBand**
  - Open standard supported by AMD, Broadcom, Cisco, HPE, Meta, Microsoft, Oracle, Linux Foundation, and many others (> 60 companies so far, even Nvidia!)
  - Improves the Ethernet protocol to allow for high bandwidth/low latency, would replace RoCE
- **Omni-Path is a competing standard originally from Intel**
  - It will be made compatible with Ultra Ethernet to stay relevant



Network interface port speeds



Paving The Way For 800 Gb/sec Ethernet In The Enterprise (nextplatform.com)

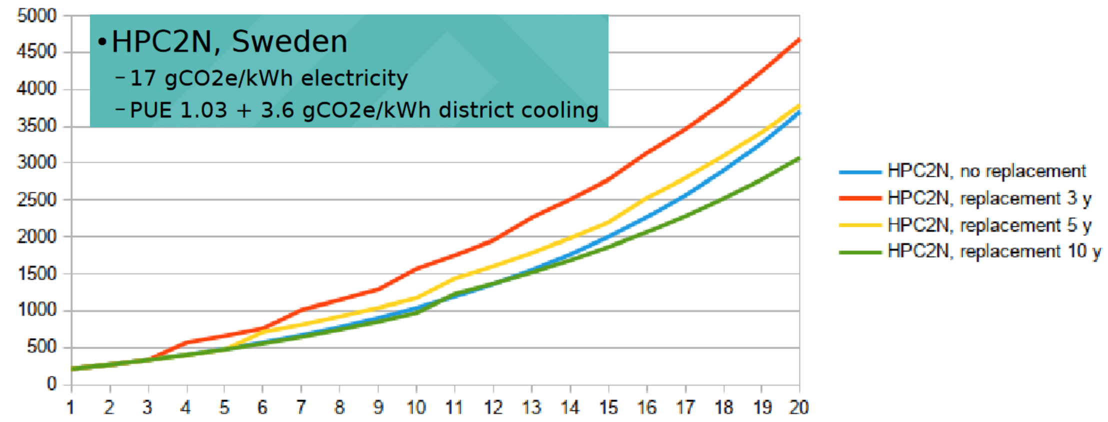
- The 200 Gpbs SerDes allows to send 200 Gbps on a single wavelength, reducing power and cost
- Co-packaged optics embed the lasers in the motherboard, potentially reducing costs

Trends on WAN connectivity

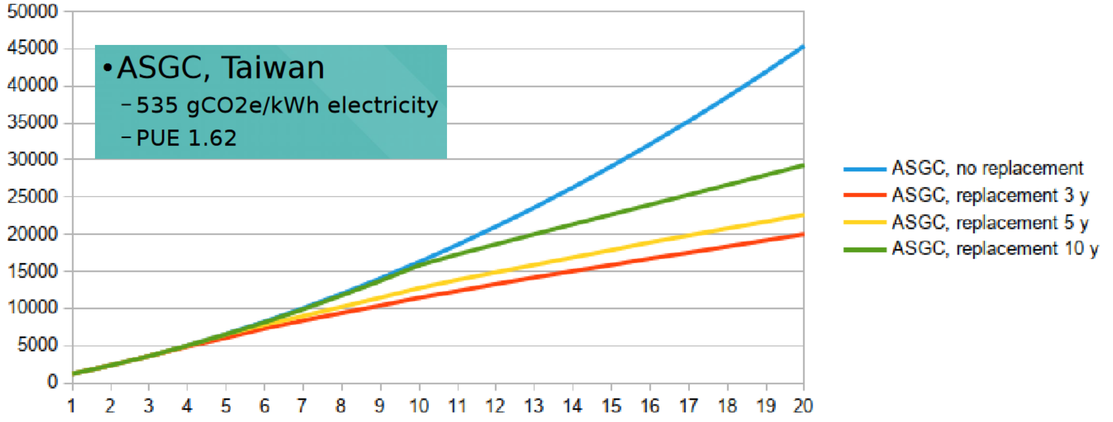
- **LHC network traffic exponentially increasing, will need Tb/s links on major routes by 2029**
  - Aggregate network traffic from ATLAS + CMS will be O(10 Tb/s)
- **R&D effort focusing on**
  - Better estimates of the required scale
  - Better models and well-defined metrics for success
  - ML for system optimization
  - Better automation (monitoring, intelligence, network OSes and tools, controllability)

More on sustainability

- **Performance/Watt is now almost as important as performance/euro**
  - Electricity prices
  - Limited cooling capacity of existing data centers
  - Need to limit CO_2 emissions
- **HDD and tape much better than SSD in terms of emissions**
  - When embedded emissions are considered!
- **Arm CPUs are attracting a lot of interest**
  - Many studies from WLCG sites comparing them to x86 CPUs in terms of efficiency
  - Still, not yet fully usable by LHC experiments (but should be soon)
  - Only two options: Ampere and Nvidia Grace
- **Many considerations enter into play**
  - “Embedded” emissions vs operational emissions: how often to replace hardware?
    - High CO_2 electricity: often! Low CO_2 electricity: less often!
  - Does it make sense to downclock (or turn off) unused nodes?
  - Cooling is critical, many new CPUs and GPUs will need liquid



Emissions (kg CO2) vs. time (y) per 1kHS23



M. Wadenstein, HEPiX Spring 2024

Conclusions

- **Technology tracking essential to make cost-efficient choices for HEP computing**
  - Done in different contexts in our community
- **Many server hardware components are rising in price due to the AI boom**
  - Memory, GPUs, flash, HDD are all affected
- **AMD, Arm and Intel show healthy competition**
  - A lot of attention to performance/Watt for many reasons
- **Evolution of GPUs is not going in a direction very useful for us**
  - FP32/64 performance not increasing in the short/medium term, to maximize AI performance
- **Shipped storage capacity increasingly driven by the global trend**
  - SSDs, HDDs and tape all still relevant and making technological progress
- **Network bandwidth correspondingly increasing on LAN and WAN**
  - To cope with increase in cores and storage/server
  - For LHC, driven by HL-LHC data rates
- **Sustainability is more important than ever**
  - CO2 emissions, liquid cooling, electricity costs and distribution

Acknowledgements

- **This work was made possible by many contributions from and discussions with**
  - The members of the HEPiX Technology Working Group
  - The members of the HEPiX Benchmarking Working Group
  - Luca Atzori, Eric Bonfillou, Bernd Panzer-Steindel, Markus Schulz

Questions?