Technology Tracking

Andrea Sciabà

INFN
NAPOLI
Istituto Nazionale di Fisica Nucleare
Sezione di Napoli

ET-PP
Preparatory Phase for the Einstein Telescope

ET EINSTEIN TELESCOPE

Workshop on Computing Infrastructures availability for ET workflows, characteristics

8-10 July 2024 - Napoli - ITALY

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 experiments
- **What scientific computing?**
  - Primarily but not limited to HEP (e.g. Virgo and DUNE are involved as associate members in WLCG)
  - Typical applications are event generation, simulation, reconstruction, data analysis, data acquisition systems, trigger, etc. with some ML used across the board
- **Two main and quite different domains in HEP**
  - Offline processing: HPC-like workloads, lots of CPUs, few GPUs, lots of storage
    - Dedicated data centers (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 HPC-like applications**
  - Typically for theoretical physics

Technology tracking at CERN and in HEPiX

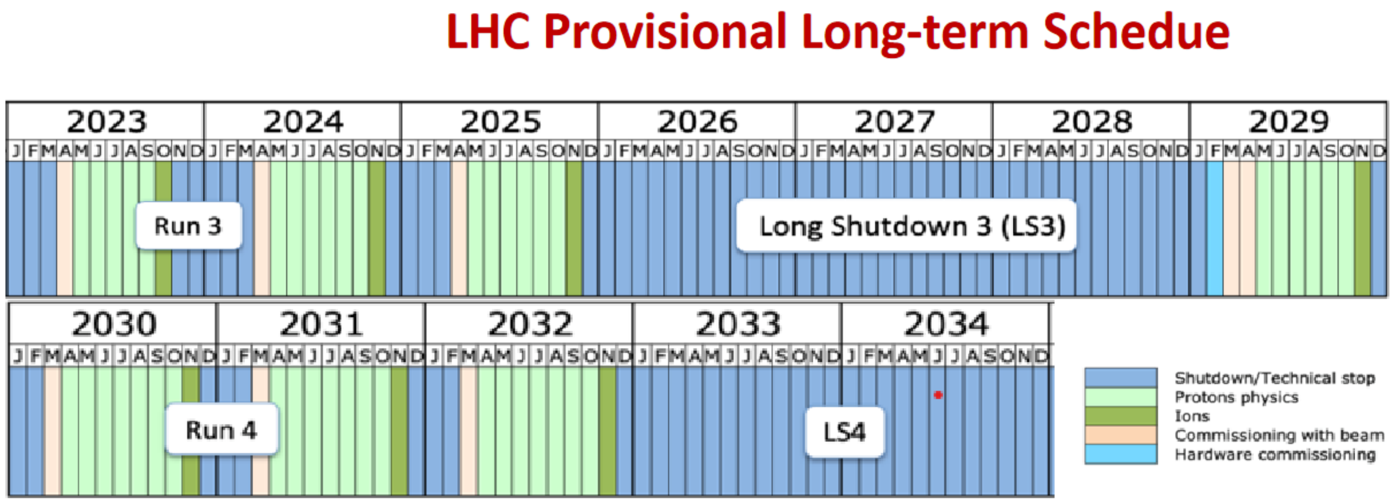
- **CERN in this context is the IT department, experiments, accelerator complex**
  - Very different communities with different needs
  - Many direct contacts with vendors (frequent NDA-covered meetings)
  - CERN openlab coordinates several joint projects
  - CERN IT CTO team closely follows technology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

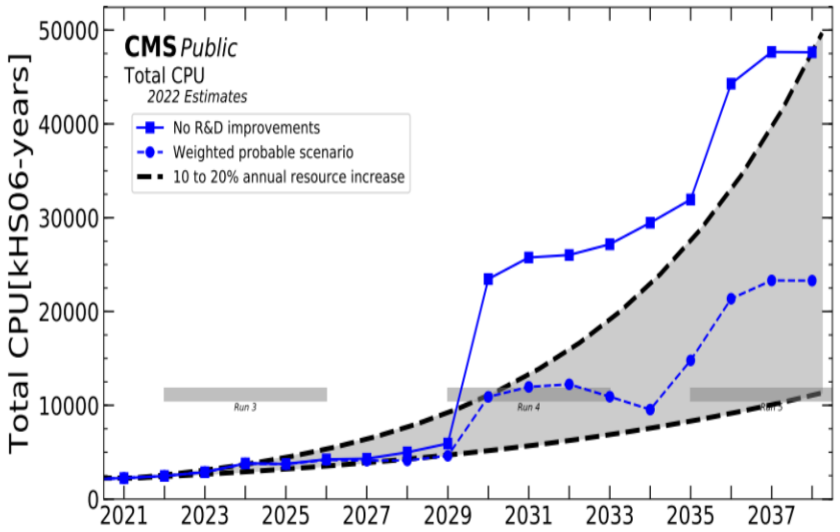
HEPiX TechWatch working group mission

- **These are the agreed goals**
  - Understand the trends and the evolution of the technology market
  - Assist in making cost predictions and optimizing investment in a sustainable way
  - Leverage the expertise of the HEPiX community
  - Inform the HEPiX board about new technologies warranting a more in-depth investigation
- **Activities**
  - Slowly restarting after a long hiatus after the pandemic
  - Mostly monthly meetings
  - Participation open to anyone!

WLCG evolution and technology

- **WLCG needs a good understanding of technology evolution for its medium-long term planning**
  - Computing cost, operations, energy usage and efficiency
  - Given the experiment requirements, find the most cost-effective technologies fulfilling them!



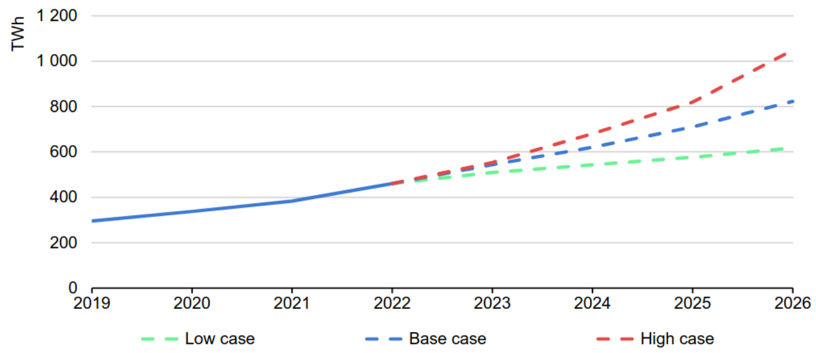


Trends in Semiconductor Industry

Power consumption

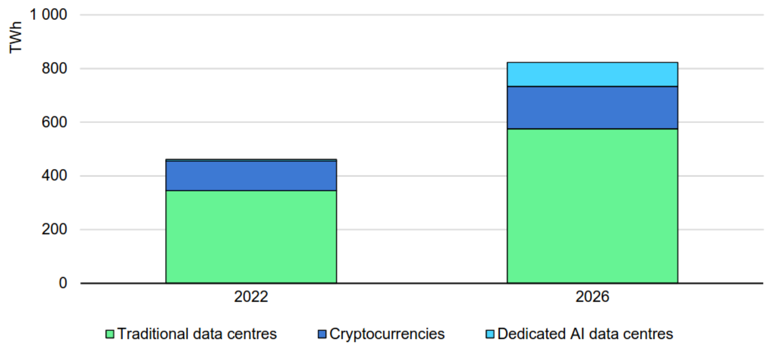
- **Data centers are proliferating more than ever due to the AI boom**
  - A single GPU uses 3.74 MWh per year, Nvidia sold 3.8 million in 2023!
  - Power infrastructure is a big constraint, strong incentive to energy efficiency
  - Data centers used ~2% of global electricity in 2022, estimated twice as much in 2026
- **The global IT market focuses more and more on sustainability**
  - This is becoming very rapidly a hot topic also in our community
    - Dedicated WLCG workshop in December

Global electricity demand from data centres, AI, and cryptocurrencies, 2019-2026



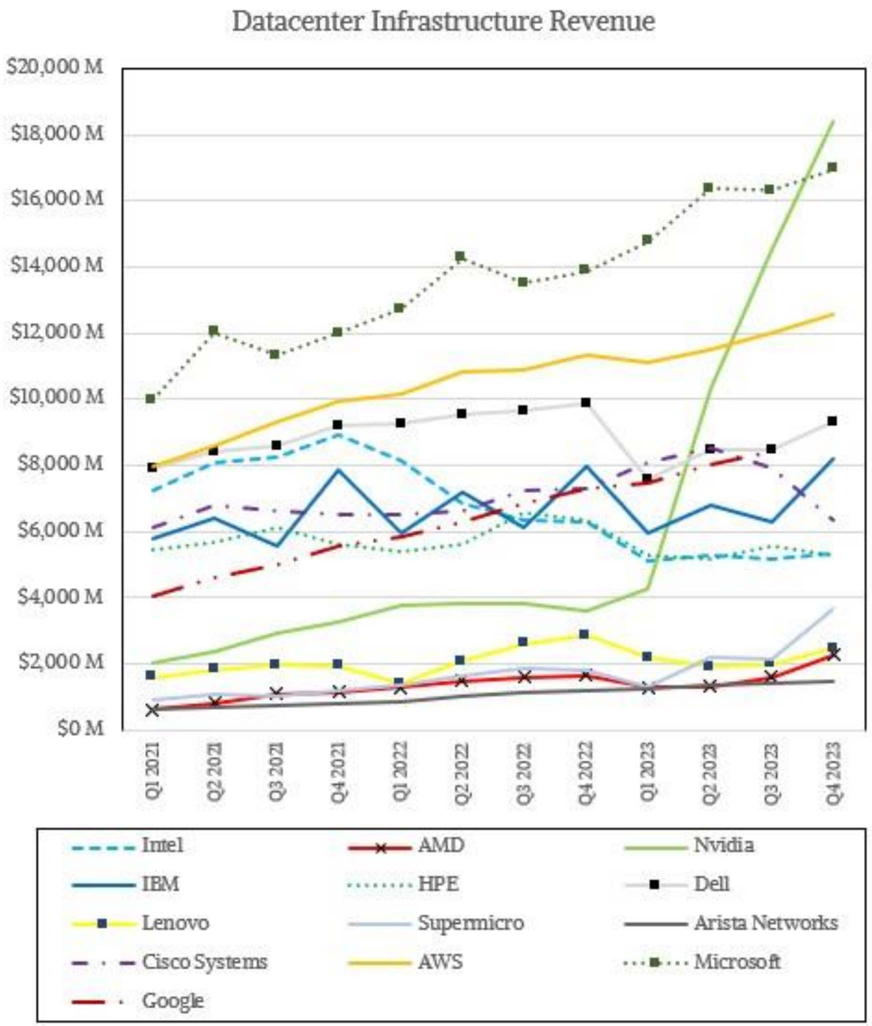
IEA Report 2024

Estimated electricity demand from traditional data centres, dedicated AI data centres and cryptocurrencies, 2022 and 2026, base case



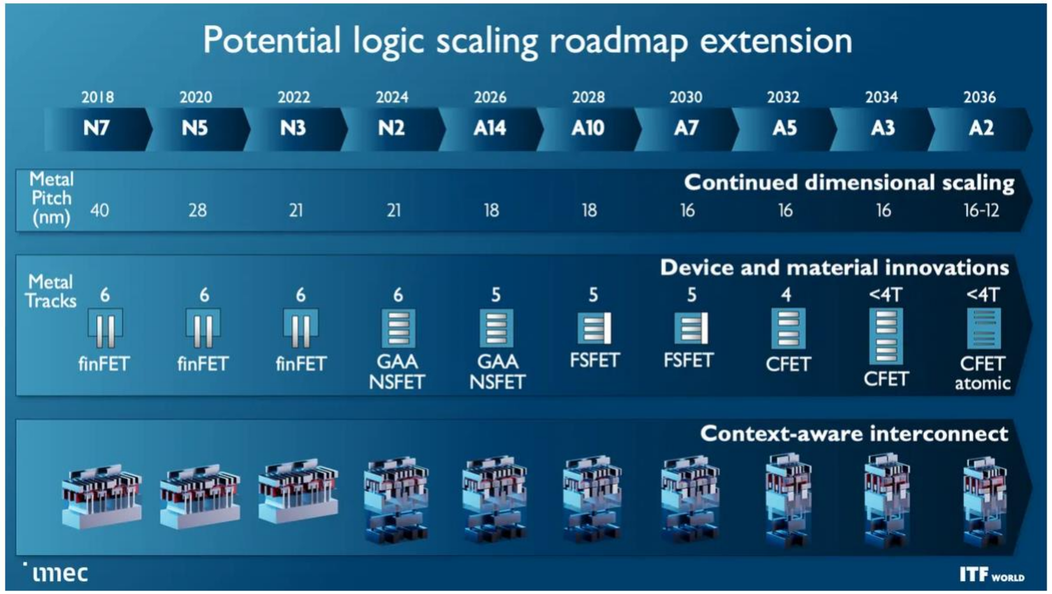
Comparing major players at scale

- **Revenues steadily increasing in the last few years, with a few exceptions**
  - Intel dropped quite a bit and is stagnant
  - Nvidia skyrocketed in 2023
- **Again, 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



Trends for fabrication processes

- **Roadmap until 2036**
  - Sub-1nm process nodes
  - Transition from FinFET transistors to Gate All Around nanosheet designs
- **“CMOS 2.0”**
  - Smaller nodes are more expensive!
  - Breaking down chips into functional units using 3D designs helps bringing down costs
- **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.)



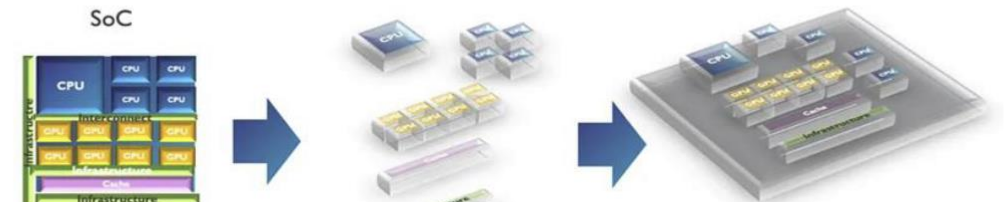
Potential logic scaling roadmap extension

	2018	2020	2022	2024	2026	2028	2030	2032	2034	2036
	N7	N5	N3	N2	A14	A10	A7	A5	A3	A2
Metal Pitch (nm)	40	28	21	21	18	18	16	16	16	16-12
Metal Tracks	6	6	6	6	5	5	5	4	<4T	<4T
	finFET	finFET	finFET	GAA NSFET	GAA NSFET	FSFET	FSFET	CFET	CFET	CFET atomic

Continued dimensional scaling

Device and material innovations

Context-aware interconnect



Breaking down the complex System-on-Chip (SoC) into smaller, more manageable components

The blocks can be optimized separately at design stage but also at technology level

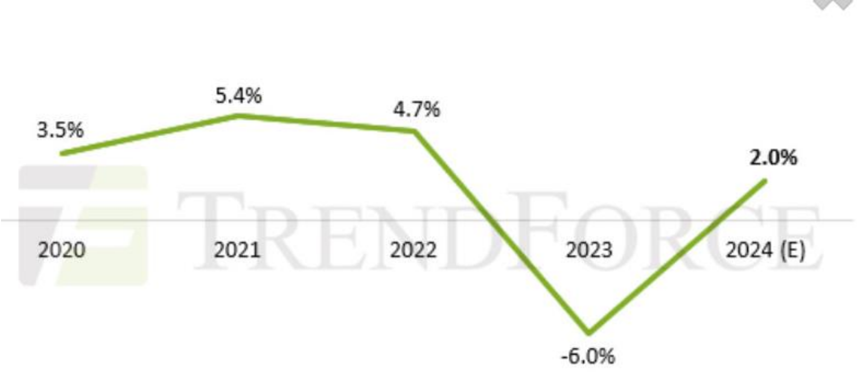
Reassembled using 2.5D (interposers) or 3D stacking

Advanced packaging for high-bandwidth and low-latency connectivity

Server Market

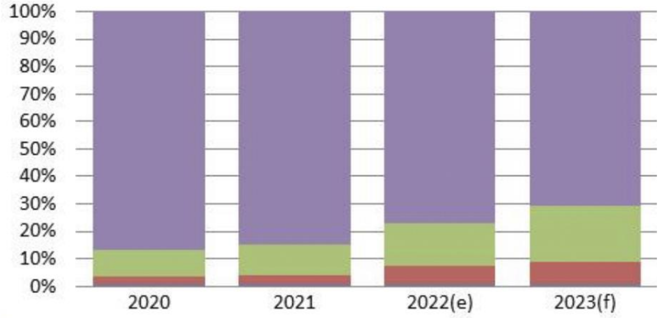
- **Server shipments are expected to slightly increase in 2024, +2% from 2023**
- **AI servers are 12% of the total**
  - Primarily driven by cloud data centers in the US
- **AMD quickly gaining ground**
  - 24% CPU server market share in 2024
- **ARM also increasing, 3x in 3 years**
  - But Ampere's revenues are a tiny fraction of the total, and Nvidia Grace is very expensive
  - Still early to heavily invest on it, for us

Global Server Shipments YoY, 2020–2024



Source: TrendForce, Feb., 2024

Chart 1: Server shipment share by CPU, 2020-2023



	2020	2021	2022(e)	2023(f)
Intel	86.6%	84.7%	77.0%	70.9%
AMD	10.1%	11.2%	15.6%	20.5%
Arm	2.6%	3.5%	6.8%	8.1%
Others	0.7%	0.6%	0.6%	0.5%

Source: DIGITIMES Research. Februarv 2023

Server and data center infrastructure

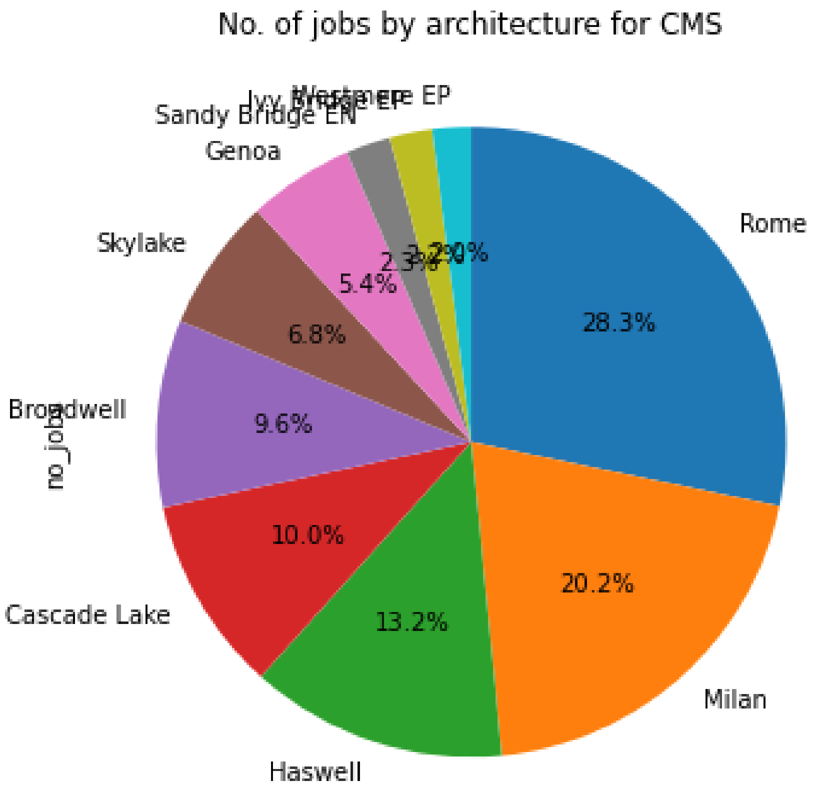
Server designs

- **High core counts make single socket servers a very viable and interesting option**
  - Simpler, cheaper, use less power
  - For the moment, only true for AMD CPUs but it will change soon
- **ARM servers are becoming a sensible alternative**
  - Better power efficiency than x86 (more on this later)
- **Liquid cooling destined to become mainstream**
  - 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

Processing units

CPUs

- **Chiplets to overcome scalability limits and improve yields**
  - Core count increasing (200+)
  - Can use different nodes for logic, cache and I/O
  - TDP will often exceed 500 W/socket this year, 1 kW in two generations
- **AMD leading the way since a few years already**
  - More than half of the CPUs in WLCG!
- **Intel catching up with their new 2024 architectures**
- **CPU models being segmented by application**
  - HPC: higher frequencies, AVX512 support, multithreading. Examples: Intel Granite Rapids and AMD Zen4/Zen5
  - Cloud: focus on performance/Watt. Examples: Intel Sierra Forest, AMD Zen4c/Zen5c



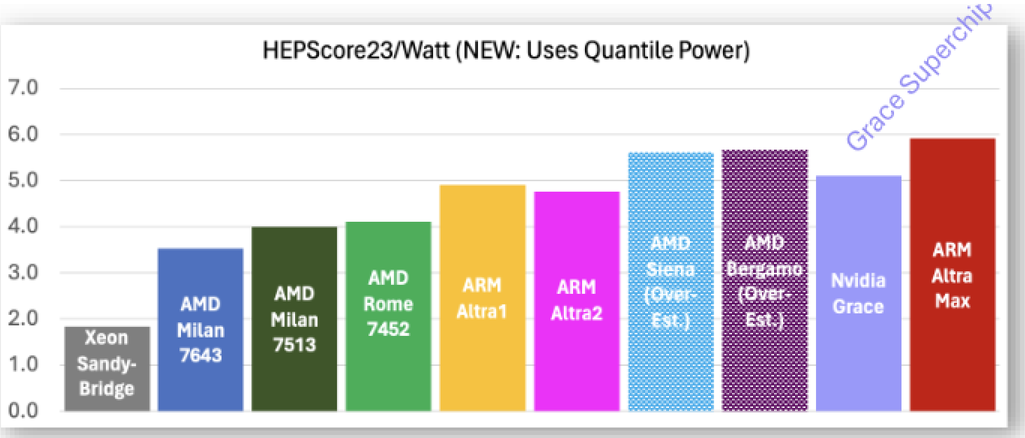
Current and next x86 CPU generations

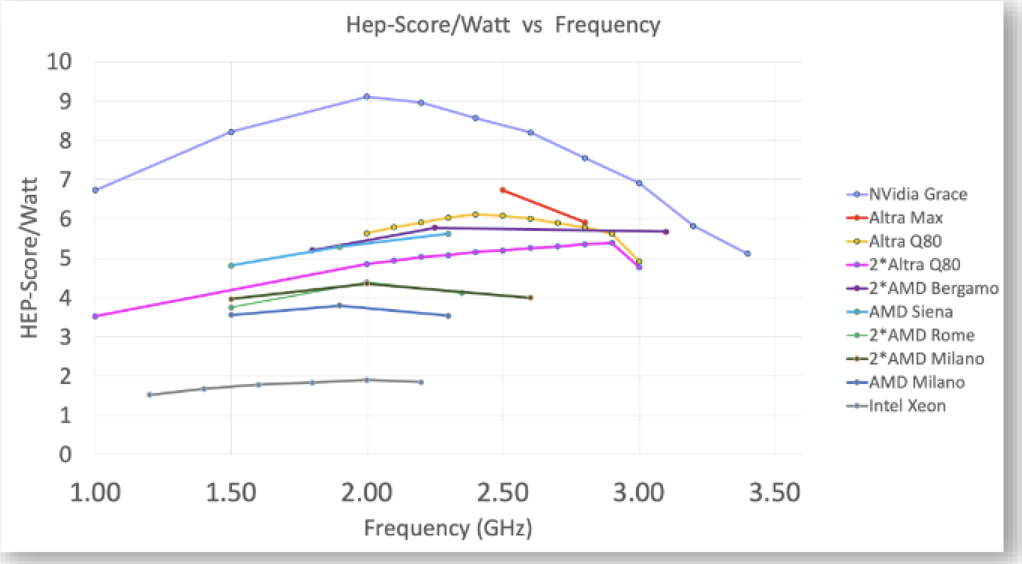
	AMD 4th gen EPYC Genoa	AMD 4th gen EPYC Bergamo	AMD 5th gen EPYC “Turin”	Intel 5th gen Xeon (Emerald Rapids)	Intel 6th gen Xeon “Sierra Forest”	Intel 6th gen Xeon 6 Granite Rapids
Launch	2022 Q4	2023 Q3	2024 H2	2023 Q4	2024 Q2	2024 Q3
Cores	Up to 96 Zen4	Up to 128 Zen4c	Up to 128 Zen5 Up to 192 Zen5c	Up to 64	Up to 144 (288 next year) E-cores	Up to 128 P-cores
Max L3 cache	384 MB	256 MB	384 MB	320 MB	108 MB (6700 series)	?
TDP	Up to 400W	Up to 400W	Up to 500 W	Up to 385 W	Up to 500 W	Up to 500 W
Memory	12 ch DDR5 up to 4800 MHz	12 ch DDR5 up to 4800 MHz	12 ch DDR5 up to 6000 MHz?	8 ch DDR5 up to 5600 MHz	Up to 12 ch DDR5 up to 6400 MHz	Up to 12 ch DDR5 up to 6400 MHz
I/O	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 heavily relies on custom benchmarks (HEPSCORE23)
  - It will be interesting to determine if E-cores can be a viable option for HEP
  - Built-in accelerators (AVX, AMX, IAA, DSA, DLB, QAT, …): not clear if and how much useful for us

ARM

- **ARM entered the server market in 2018**
  - Sells designs, not chips
  - 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 2023: up to 192 cores (custom design)
- **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
  - ARM market is still minuscule and both Intel and AMD push strongly on power efficiency… difficult to make predictions





David Britton, University of Glasgow

GDB, June 2024

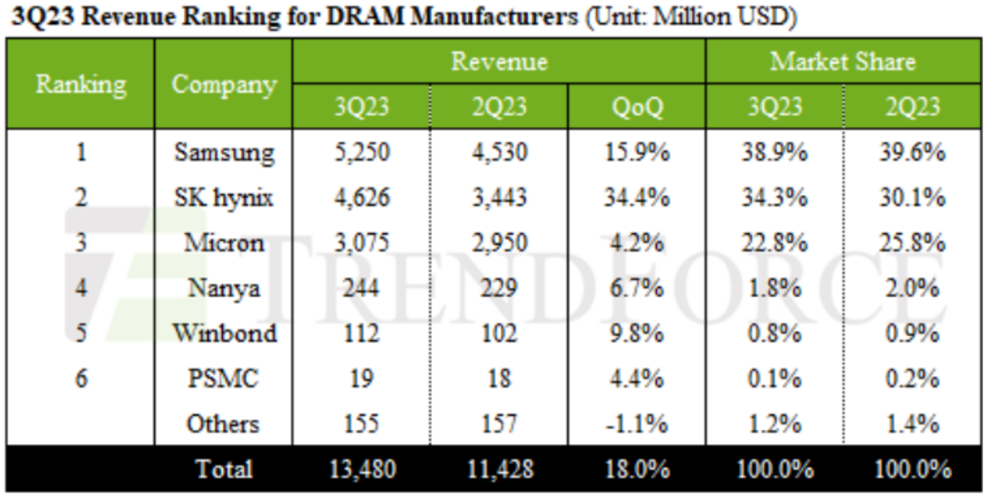
GPUs and accelerators

- **Currently, almost an Nvidia monopoly, but AMD is gaining ground**
  - Nvidia Hopper (H100/H200), Blackwell (B200) later this year
  - AMD MI300X
  - Intel still extremely marginal (Gaudi 2/3 are just AI accelerators, Ponte Vecchio is already old)
  - Cloud native accelerators from AWS, Google, Microsft, etc
- **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 soon
  - AI performance (FP16, FP8, INT8) has priority because is where most of the money is made!

	MI300X	H100 SXM	Blackwell (B100)
TBP	750 W	700 W	700 W
Memory	192 GB of HBM3	80 GB of HBM3	Up to 192 GB of HBM3e
FP64 matrix/vector	164/82 TFLOPS	67/34 TFLOPS	30 TFLOPS
FP8	2615 TFLOPS	1979 TFLOPS	7 PFLOPS

Memory

- **All modern CPUs support DDR5, DDR6 will appear in 2025**
  - Typical speed is 6.4 GT/s
  - MCR DIMM will further increase speed to 8+ GT/s
  - Nvidia Grace though uses LPDDR5X!
- **HBM memory used for data center GPUs**
  - Directly connected to the GPU
  - Latest is HBM3e, total bandwidth per stack exceeds 1 TB/s
  - Extremely useful for 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
- **All memory manufacturers are now finally using EUV lithography**

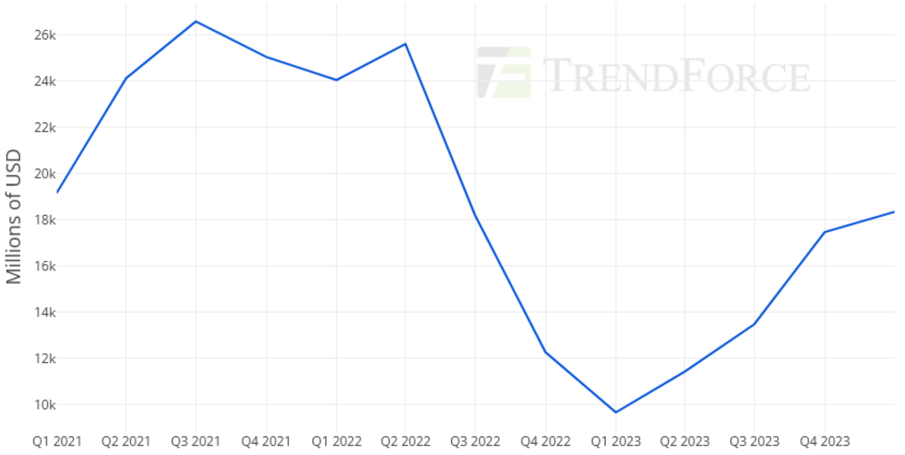


3Q23 Revenue Ranking for DRAM Manufacturers (Unit: Million USD)

Ranking	Company	Revenue			Market Share	
		3Q23	2Q23	QoQ	3Q23	2Q23
1	Samsung	5,250	4,530	15.9%	38.9%	39.6%
2	SK hynix	4,626	3,443	34.4%	34.3%	30.1%
3	Micron	3,075	2,950	4.2%	22.8%	25.8%
4	Nanya	244	229	6.7%	1.8%	2.0%
5	Winbond	112	102	9.8%	0.8%	0.9%
6	PSMC	19	18	4.4%	0.1%	0.2%
	Others	155	157	-1.1%	1.2%	1.4%
	Total	13,480	11,428	18.0%	100.0%	100.0%

Source: Trendforce

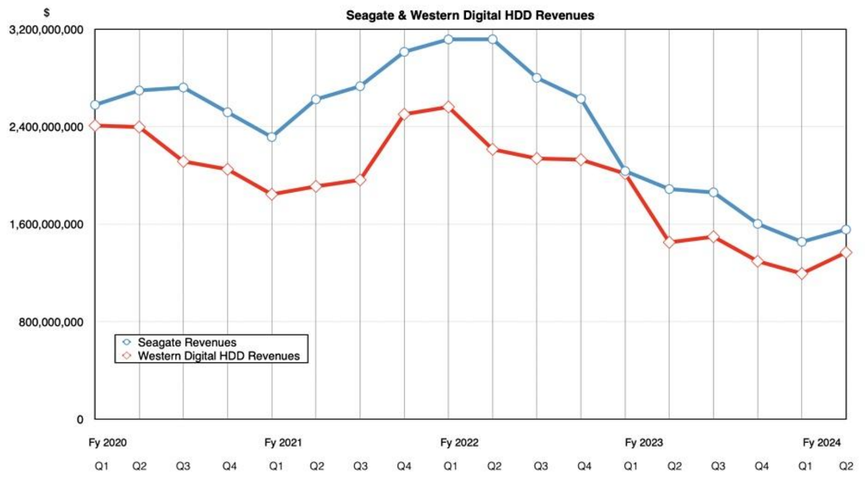
Global Branded DRAM Manufacturers' Revenue: Total



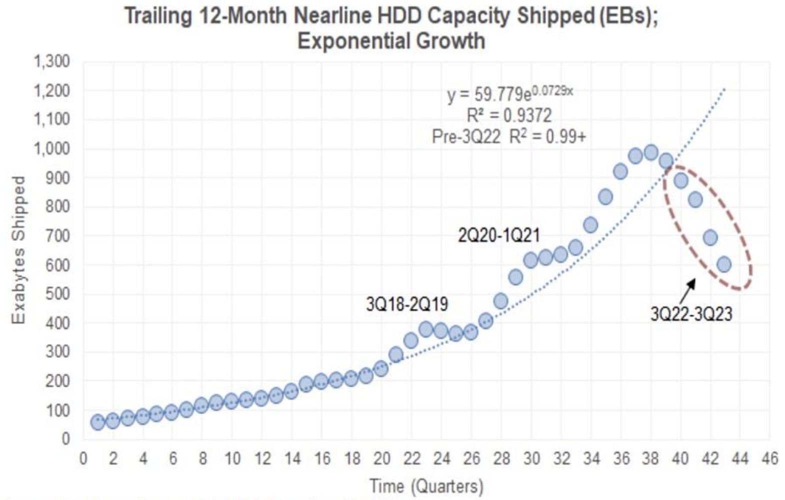
Storage

HDD storage

- **HAMR drives have finally arrived**
  - Seagate Mosaic 3+ 30TB HAMR drives shipping in Q1 2024, will eventually be adopted by a wide range of products
  - Not something that you can really buy yet: for now, only in dedicated enclosures
  - 50 TB disks by 2030?
  - Longer timescales for other players?
    - Latest WD CMR drive is 22TB
- **SMR drives give a 20% capacity increase over PMR**
  - ~50% of exabytes shipped by Western Digital are SMR drives
  - Not without issues when tested with EOS!
- **Majority of exabytes shipped and revenue are nearline HDD**
  - Market was down for 2023
  - But expected to increase by 22% (Gartner)
- **Cost/GB gap with SSD destined to decrease in the long term**
  - Now 3-5 factor in euros/TB
  - Nearline drives will be the last HDD holdout, but will not disappear any time soon



Western Digital roller coaster continues as Seagate brings down the HAMR – Blocks and Files

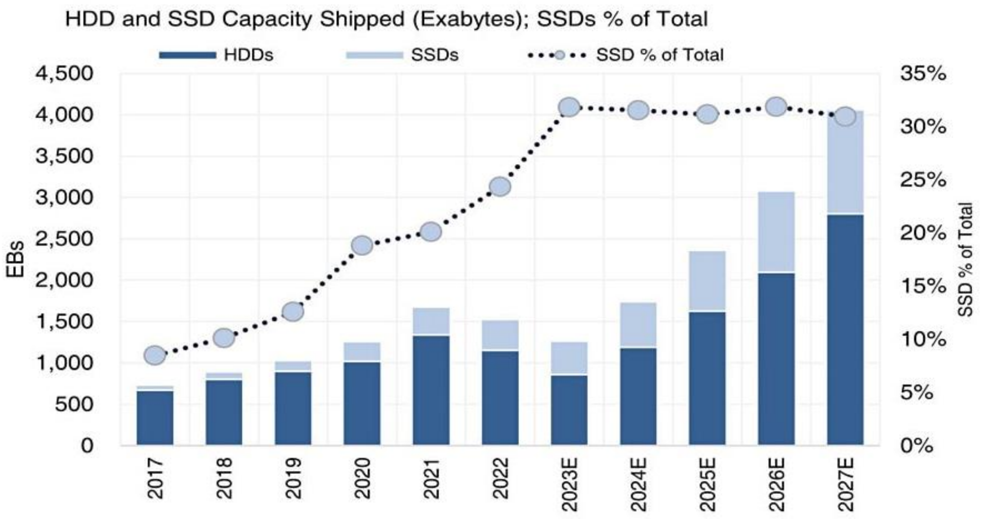


Source: TrendFocus; Company Data; Wells Fargo Securities, LLC

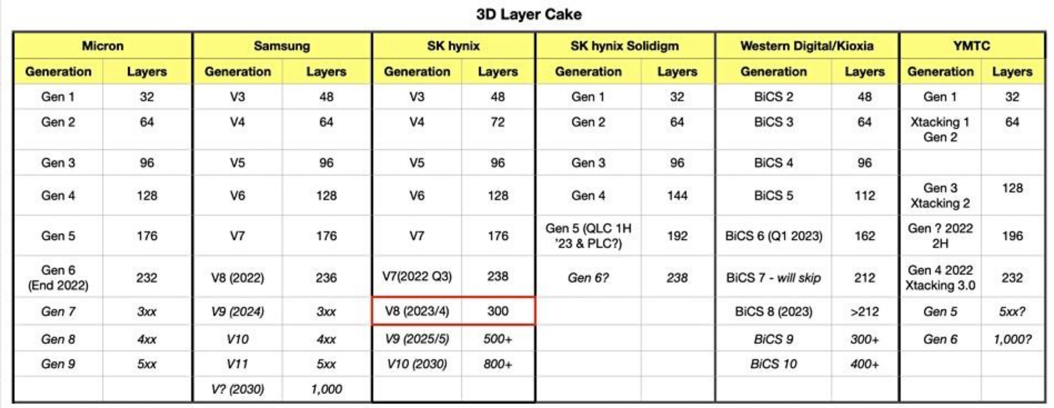
Nearline drives will be last HDD holdout by 2028 – Blocks and Files

Flash storage

- **SSD account for ~30% of storage capacity**
  - Not expected to change dramatically
- **Samsung and SK Hynix dominate**
  - Western Digital spinning off flash business
- **PCIe Gen 5 SSDs now available**
  - ~ 200+ Layer 3D NAND flash chips from all five major vendors
    - ~1000 layers by 2030?
- **Viability of penta level cells unclear**
  - Exponentially more challenging to add more bits per cell
- **Total revenues recovering from dip in late 2022/early 2023**



Source: Gartner; Wells Fargo Securities, LLC .



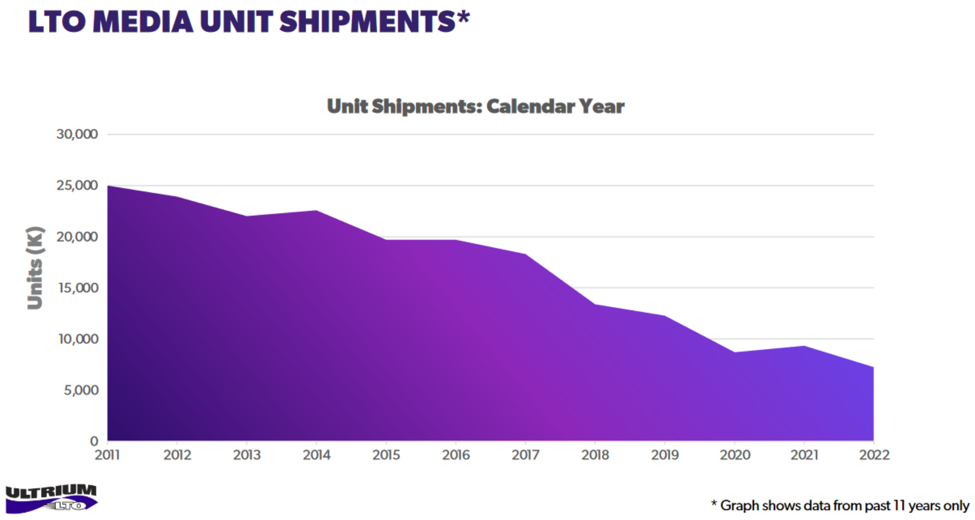
3D Layer Cake

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	Gen 3 Xtacking 2	128
Gen 5	176	V7	176	V7	176	Gen 5 (QLC 1H '23 & PLC?)	192	BiCS 6 (Q1 2023)	162	Gen ? 2022 2H	196
Gen 6 (End 2022)	232	V8 (2022)	236	V7(2022 Q3)	238	*Gen 6?*	*238*	BiCS 7 - *will skip*	212	Gen 4 2022 Xtacking 3.0	232
Gen 7	*3xx*	*V9 (2024)*	*3xx*	V8 (2023/4)	300			BiCS 8 (2023)	>212	*Gen 5*	*5xx?*
Gen 8	*4xx*	*V10*	*4xx*	*V9 (2025/5)*	*500+*			*BiCS 9*	*300+*	*Gen 6*	*1,000?*
Gen 9	*5xx*	*V11*	*5xx*	*V10 (2030)*	*800+*			*BiCS 10*	*400+*		
		V? (2030)	*1,000*								

SK hynix breezes past 300-layer 3D NAND mark – Blocks and Files

Archive storage

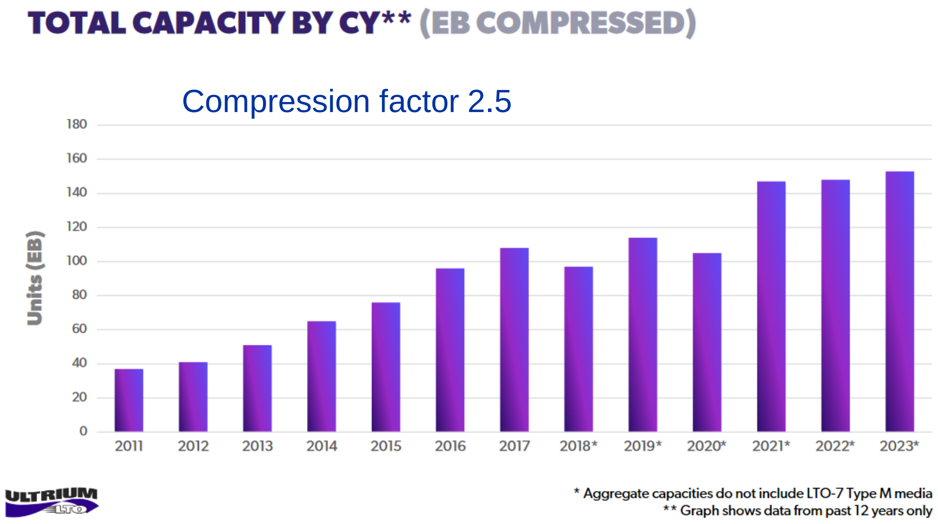
- **Magnetic Tape**
  - Still a lot of room for scaling (unlike HDD)
  - Strategy change at IBM for enterprise drives
    - TS1170 - 50TB / cartridge. No backward compatibility
  - IBM Diamondback (LTO) “library in a rack” targets cloud hyperscale and traditional enterprises
  - Total LTO cartridges 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 laser or particle beam
  - Folio Photonics - No news since 2022


LTO MEDIA UNIT SHIPMENTS*

Unit Shipments: Calendar Year

* Graph shows data from past 11 years only

https://www.lto.org/wp-content/uploads/2023/04/LTO-Ultirum-2022-Media-Shipment-Report-Slides.pdf


TOTAL CAPACITY BY CY** (EB COMPRESSED)

Compression factor 2.5

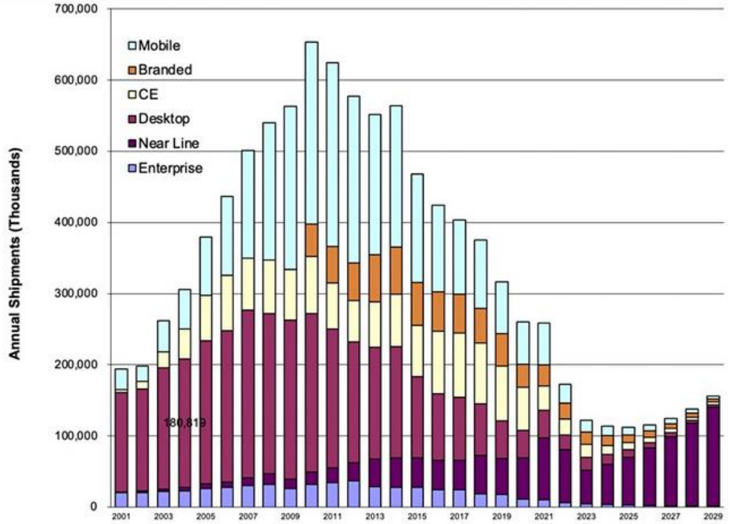
* Aggregate capacities do not include LTO-7 Type M media

** Graph shows data from past 12 years only

Storage evolution summary

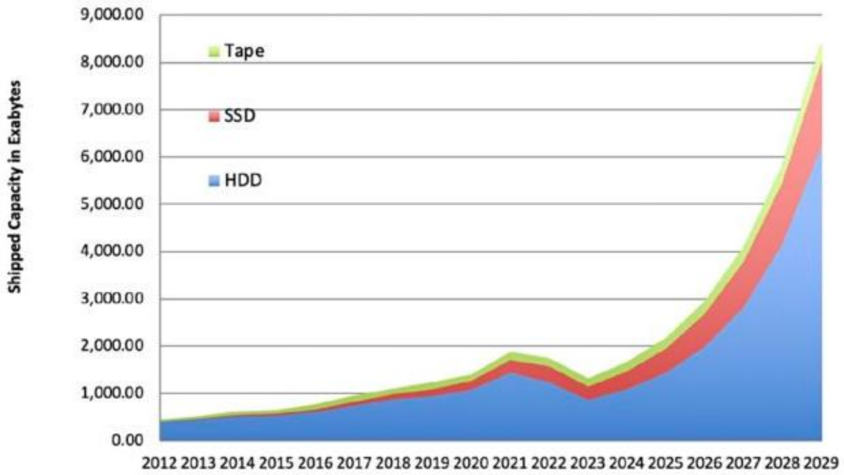
- **To summarize:**
  - AI boom drives volume increase for all types of storage
  - HDD shipments will soon be almost only nearline, but increasing
  - SSDs cannibalized HDDs but in the retail market, will not do that in the data center anytime soon
  - Tapes are not going anywhere either

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



https://blocksandfiles.com/2024/05/13/coughlin-associates-hdd/

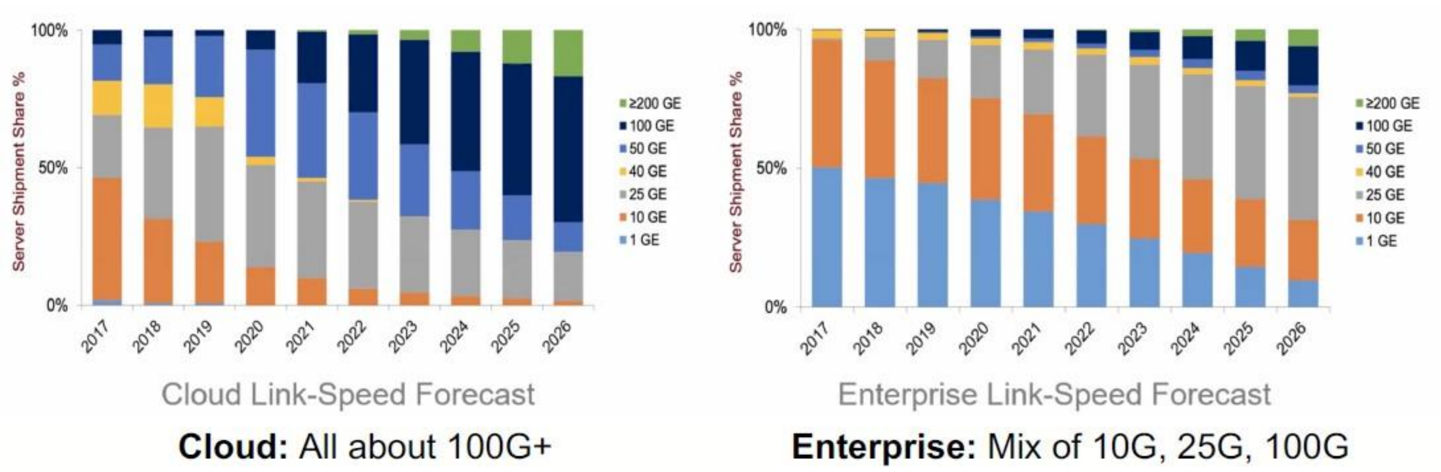
FIGURE 11. CAPACITY SHIPMENTS FOR LTO TAPE, SSDS AND HDDS



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
- **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
- **Omni-Path is a competing standard originally from Intel**
  - It will be made compatible with Ultra Ethernet to stay relevant



Cloud Link-Speed Forecast

Enterprise Link-Speed Forecast

Cloud: All about 100G+

Enterprise: Mix of 10G, 25G, 100G

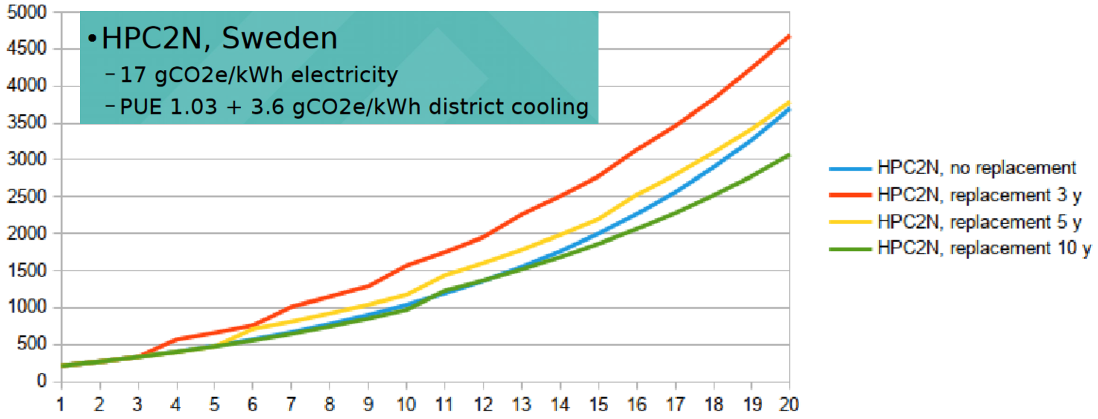
Paving The Way For 800 Gb/sec Ethernet In The Enterprise (nextplatform.com)

Trends on WAN connectivity

- **The LHC community is building on several R&D projects and the move to fully programmable ecosystems of networks and systems (SONiC P4, PolKA SRv6) and operations platforms (OSG, NRP, …)**
  - Coordination by the GNA-G, WLCG, the worldwide R&E network community
- **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 required scale
  - Better models and well-defined metrics for success
  - ML for system optimization
  - Better automation (monitoring, intelligence, network OSes and tools, controllability)

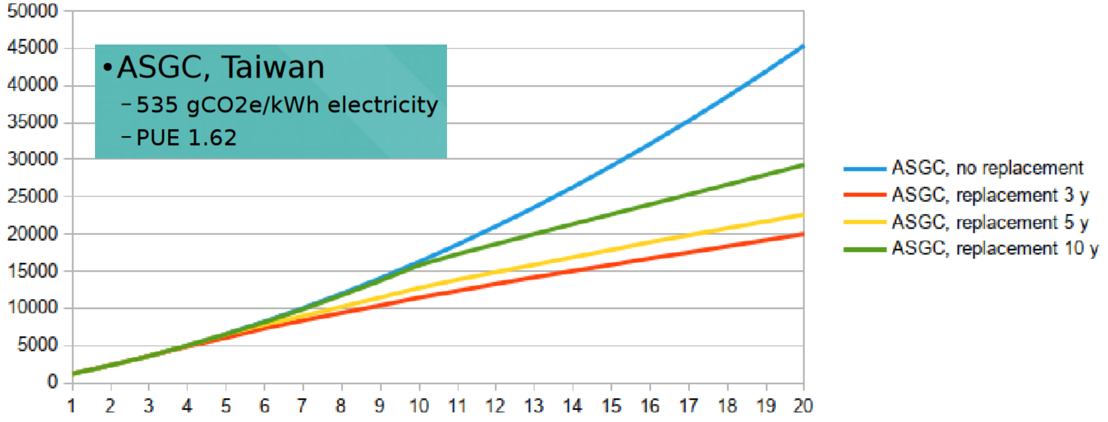
More on sustainability

- **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 taken into account
- **ARM CPUs are attracting a lot of interest**
  - Many studies from WLCG sites comparing them to x86 CPUs in terms o 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



- HPC2N, Sweden
  - 17 gCO2e/kWh electricity
  - PUE 1.03 + 3.6 gCO2e/kWh district cooling

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



- ASGC, Taiwan
  - 535 gCO2e/kWh electricity
  - PUE 1.62

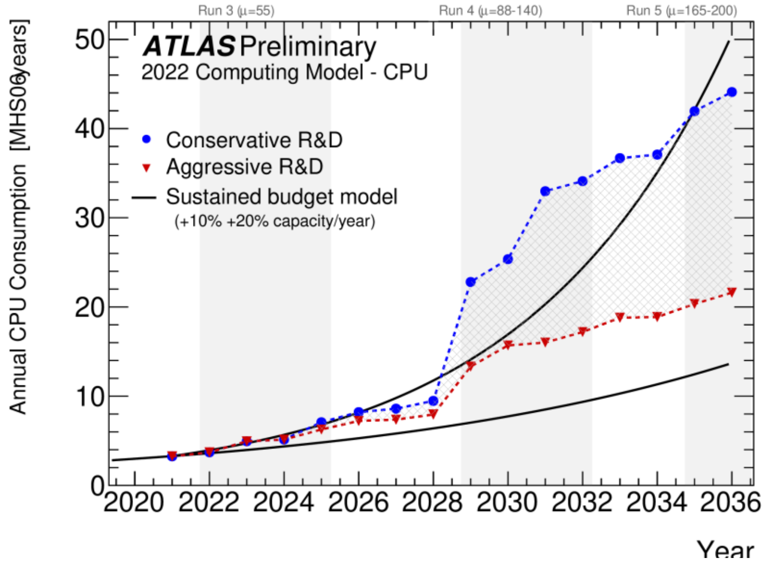
M. Wadenstein, HEPiX Spring 2024

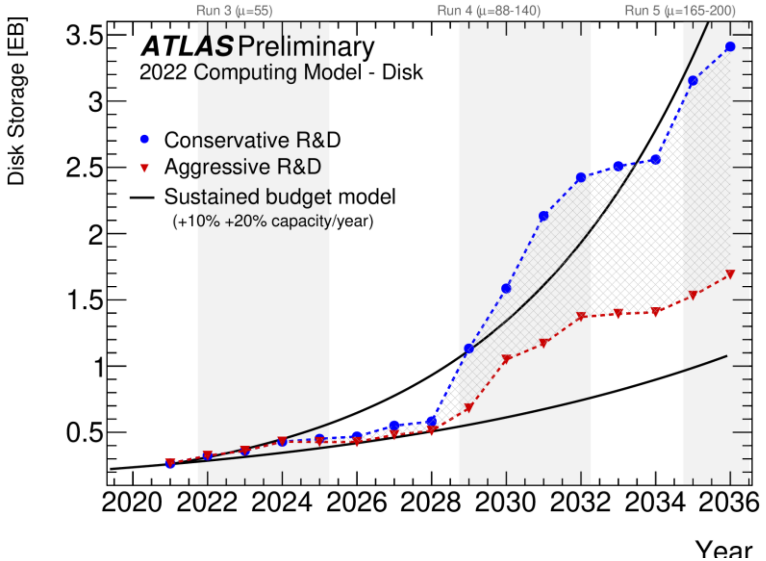
Conclusion

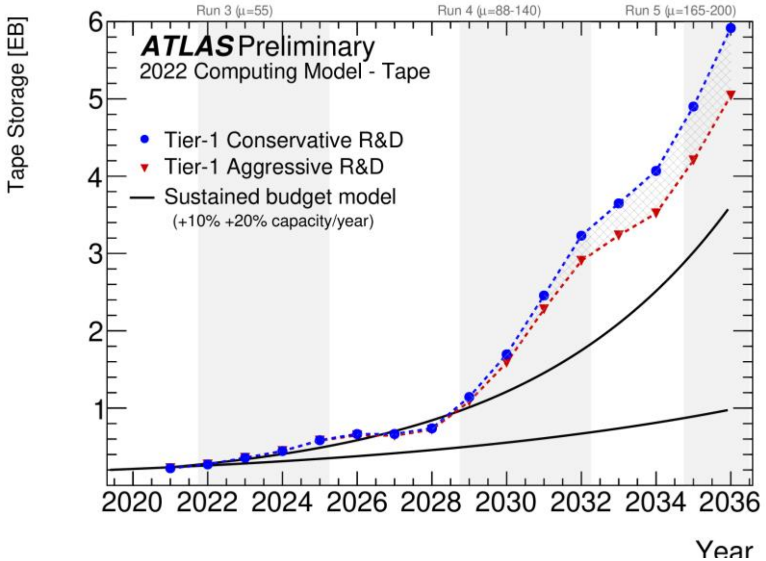
- **Technology tracking essential to make cost-efficient choices for HEP computing**
  - Done in different contexts in our community
- **Many server hardware components are rising in price due to the AI boom**
  - Memory, GPUs, flash, HDD are all affected
- **AMD, ARM and Intel show heathy competition**
  - 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 maximise AI performance
- **Shipped storage capacity increasing 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

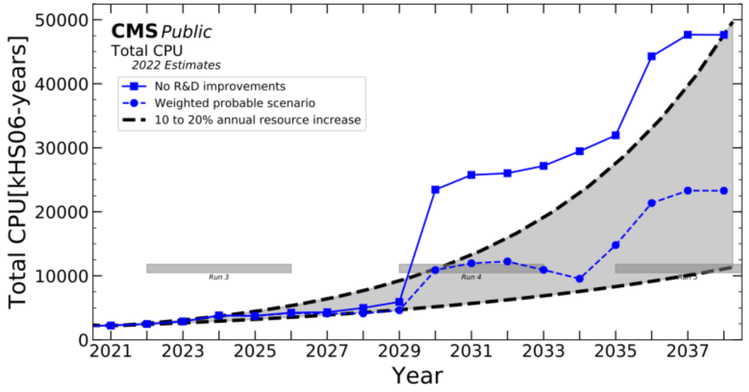
Backup slides

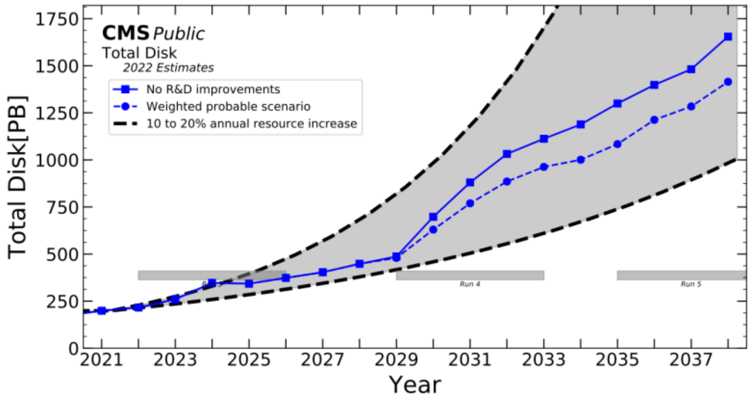
ATLAS and CMS resource needs for HL-LHC

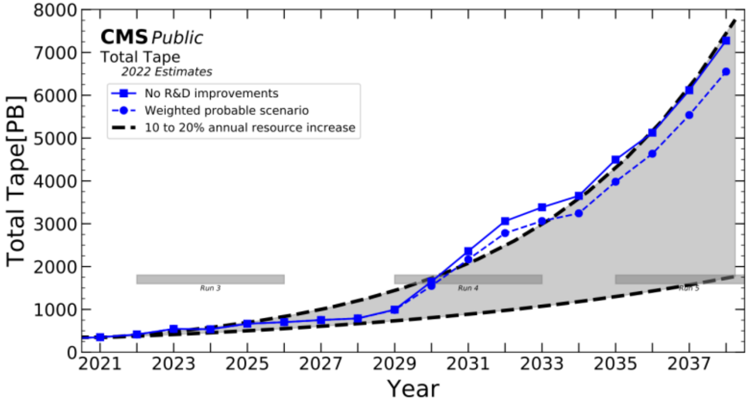




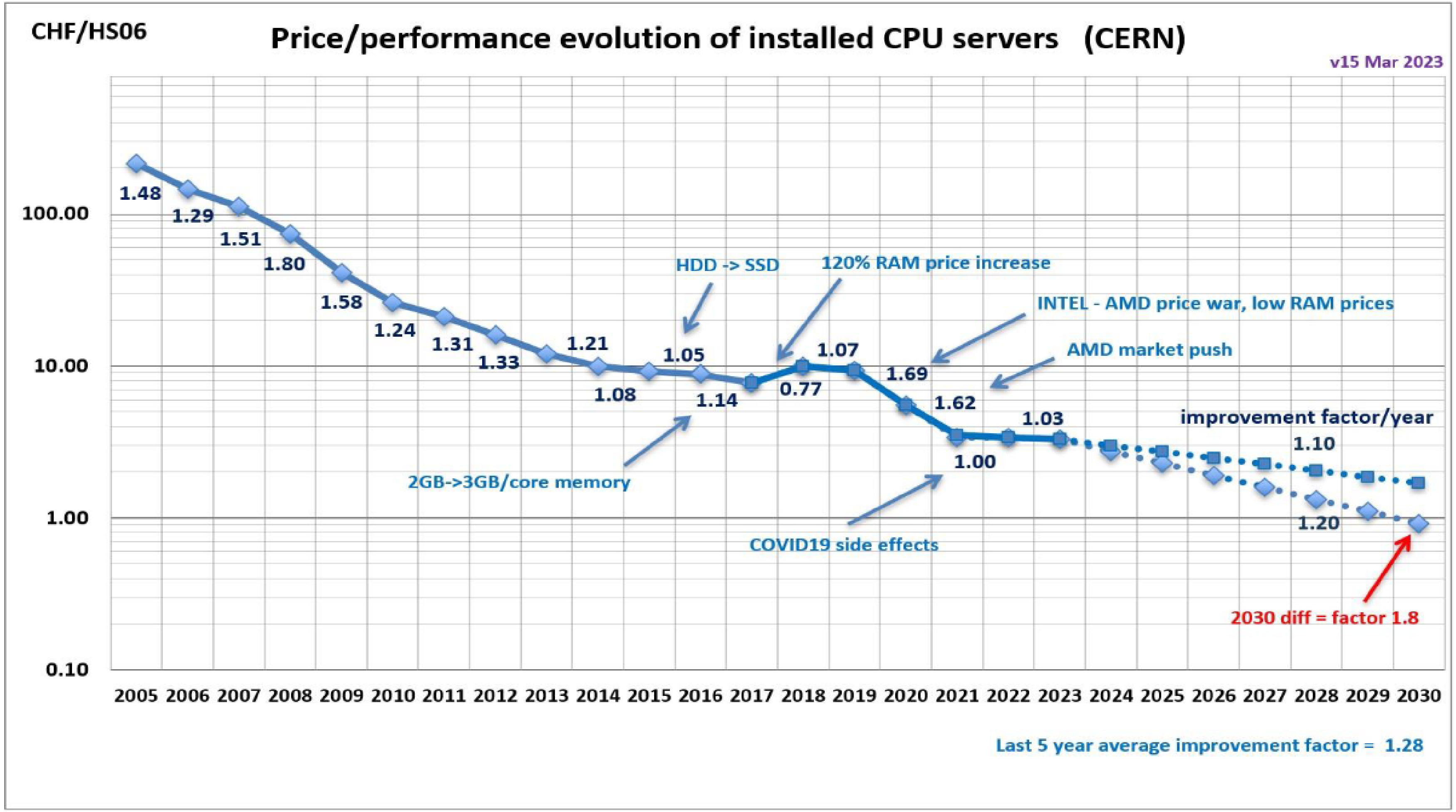






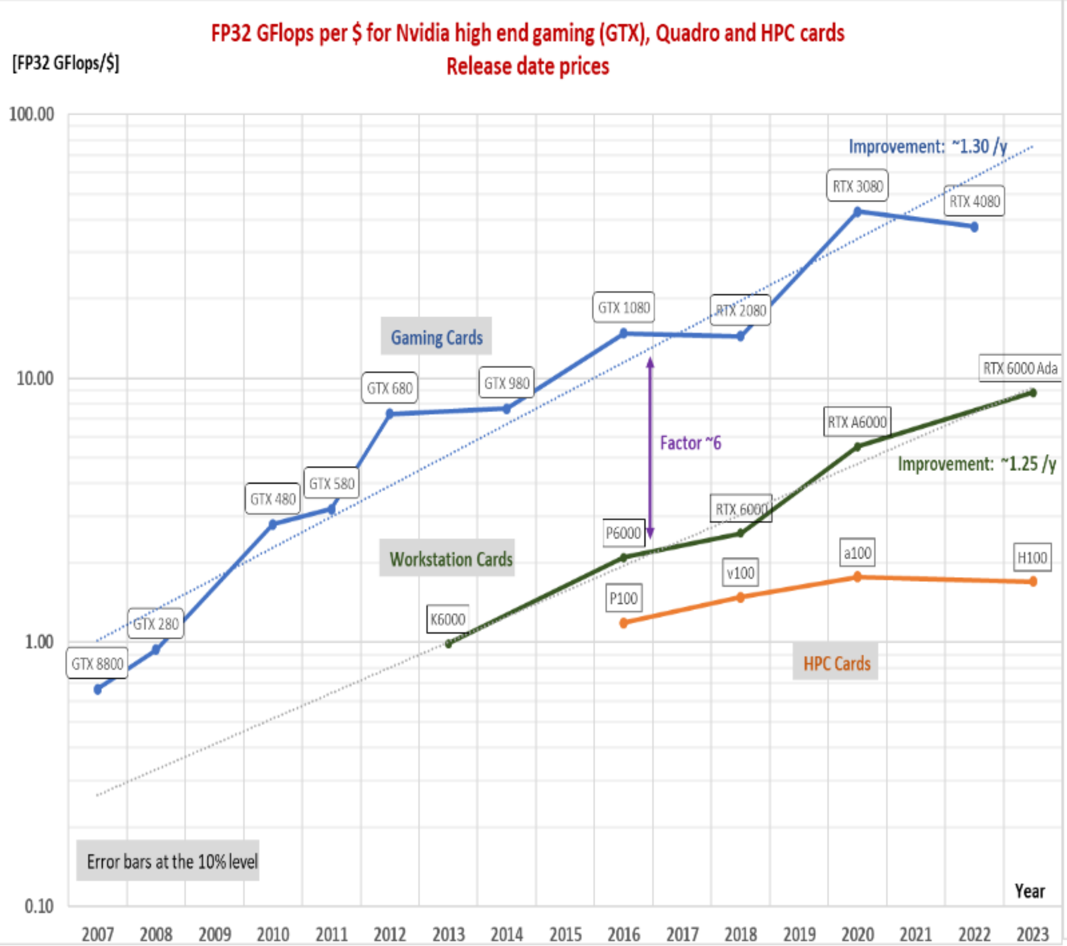


CPU processing costs

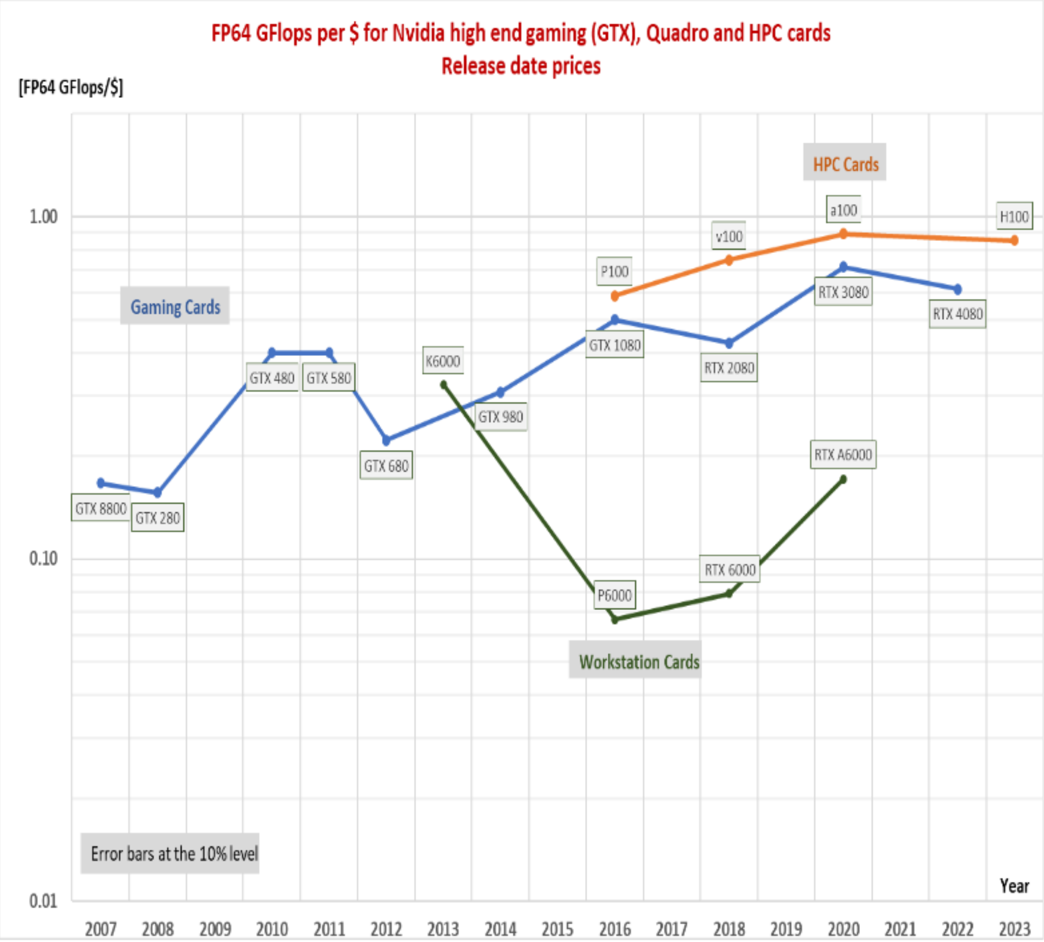


Source: B. Panzer-Steindel

GPU processing costs



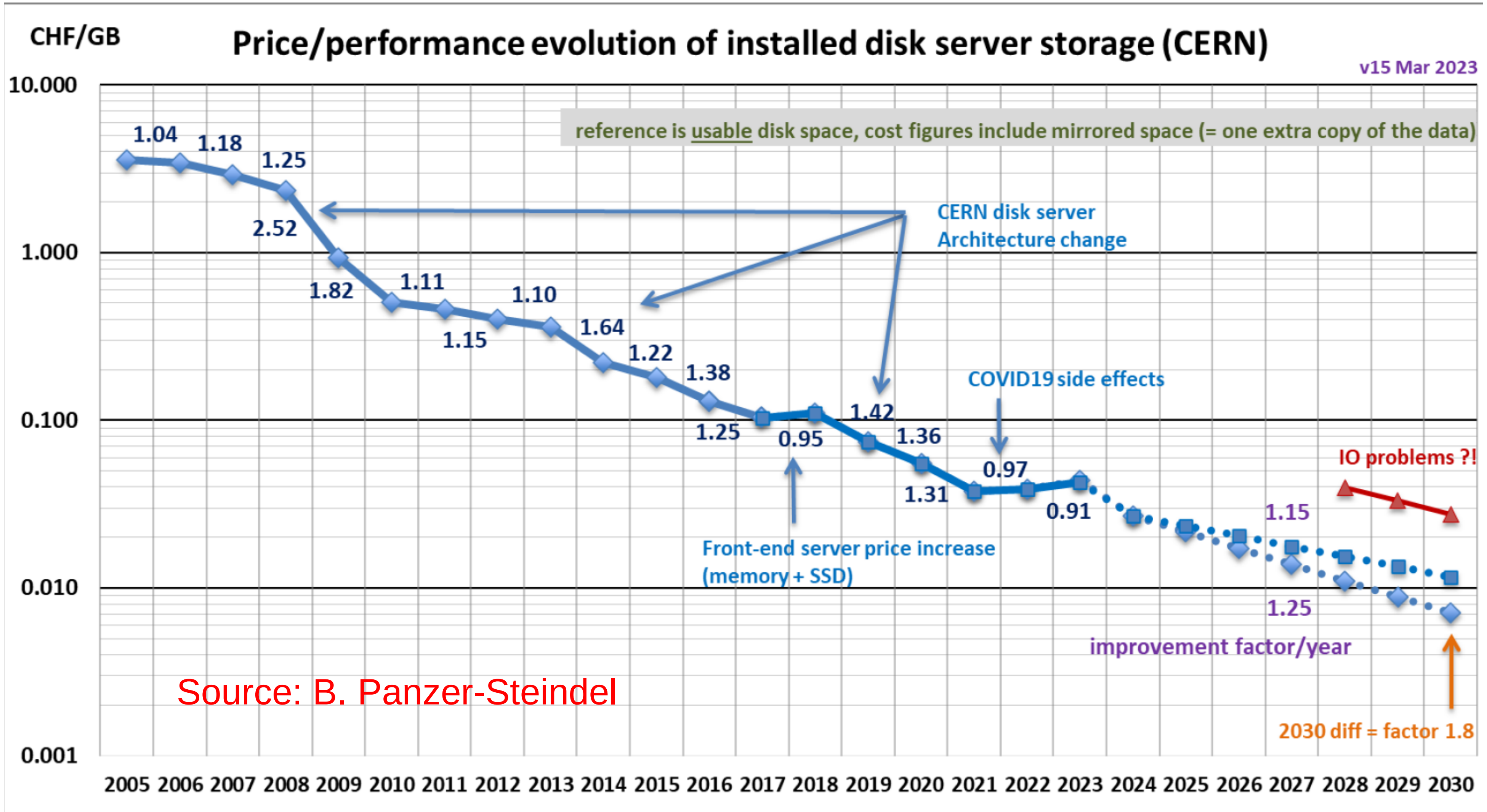
FP32 GFlops per $ for Nvidia high end gaming (GTX), Quadro and HPC cards
Release date prices



FP64 GFlops per $ for Nvidia high end gaming (GTX), Quadro and HPC cards
Release date prices

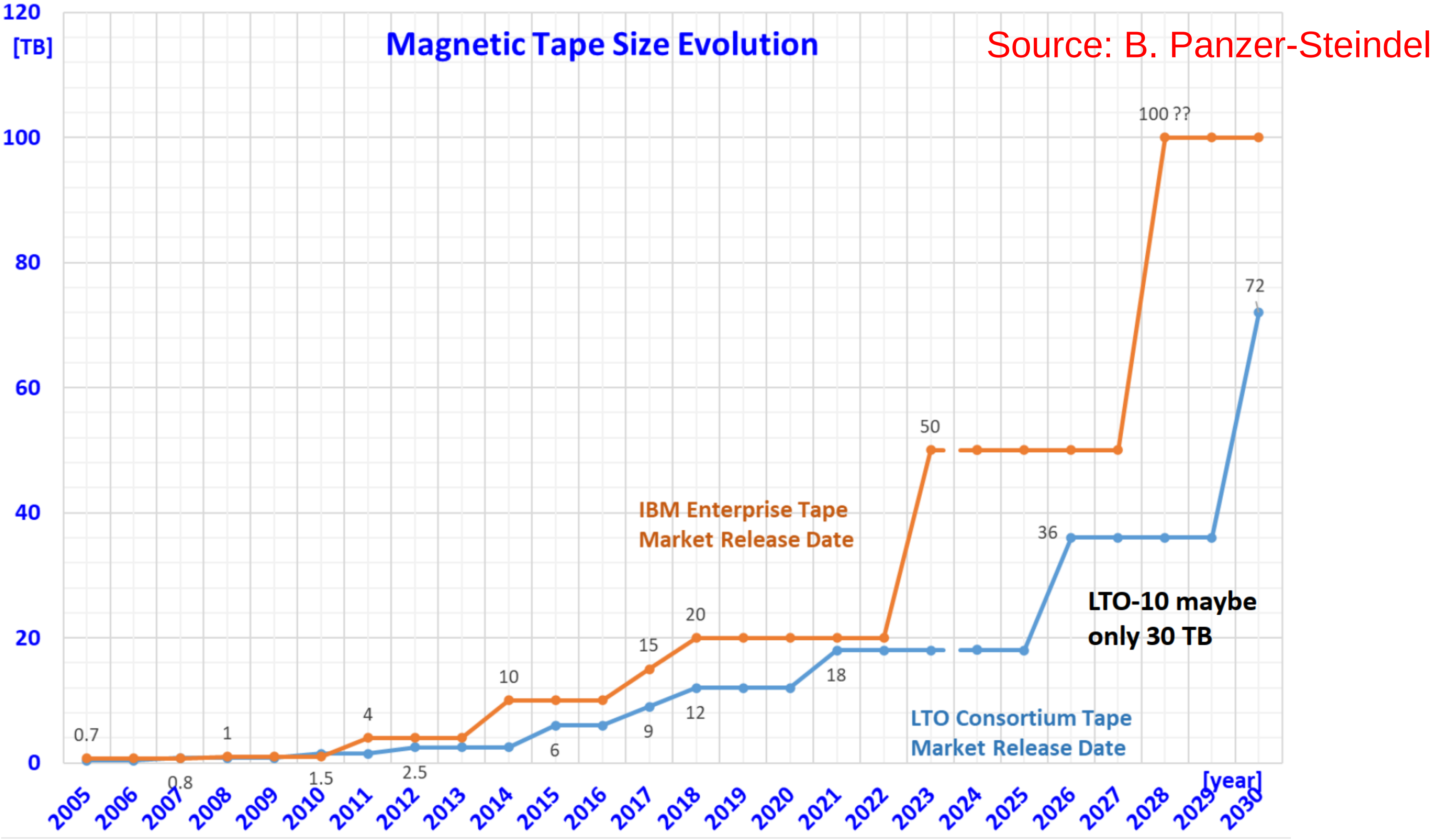
Source: B. Panzer-Steindel

Disk costs



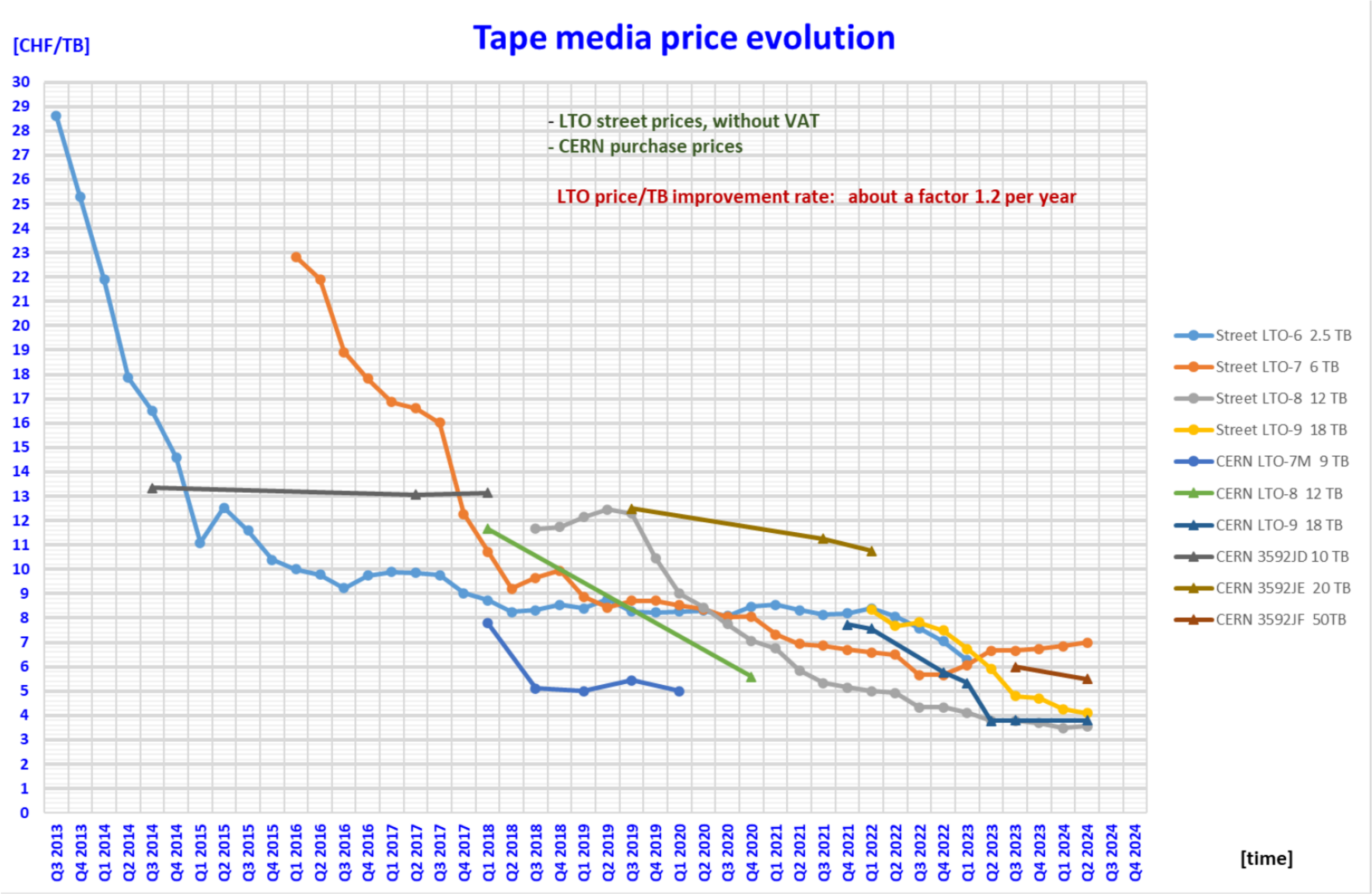
Source: B. Panzer-Steindel

Tape media evolution



Source: B. Panzer-Steindel

Tape media costs



Source: B. Panzer-Steindel