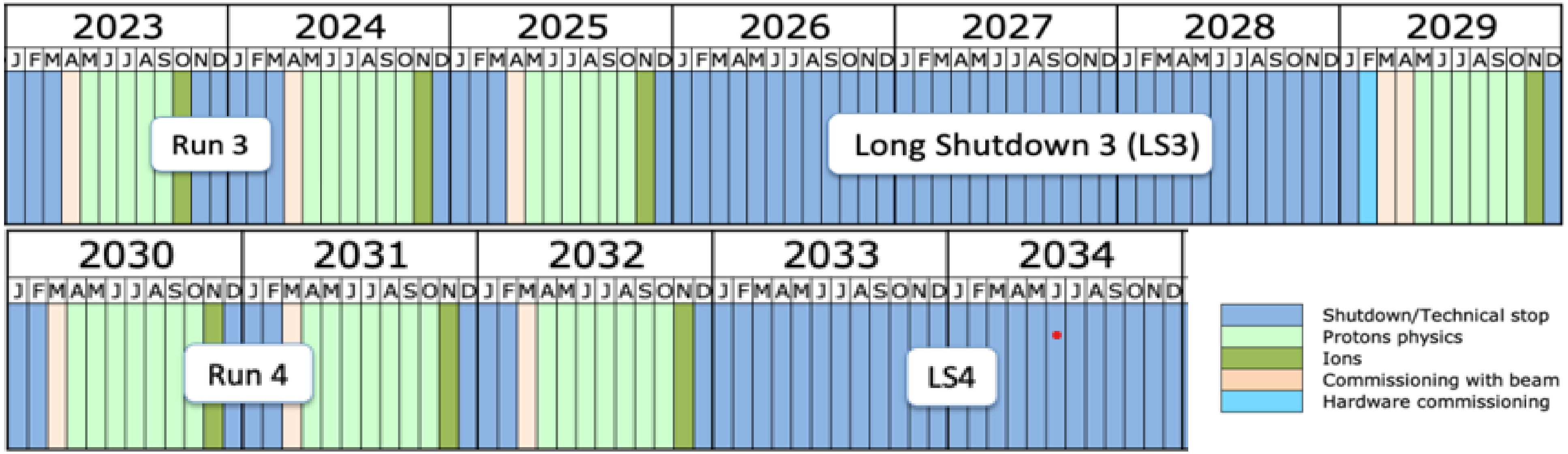
Computing Technology and Market Evolution

with a view on Run 4 (HL-LHC)

Introduction

- A bit of a random walk looking ahead to the start of Run 4
- Trying to link computing cost, energy, operations, technology and markets sprinkling some CERN T0 specific issue across the talk
- IT and experiment requirements and boundary conditions
- This is about hardware resources, not FTEs
- Large uncertainties: 20% a given, but up to a factor 2
- Certainly not a complete overview, rather a start of the discussion

LHC Provisional Long-term Schedule



Major detector upgrades/replacements for ATLAS and CMS (Phase II upgrades), while ALICE and LHCb only with minor modifications

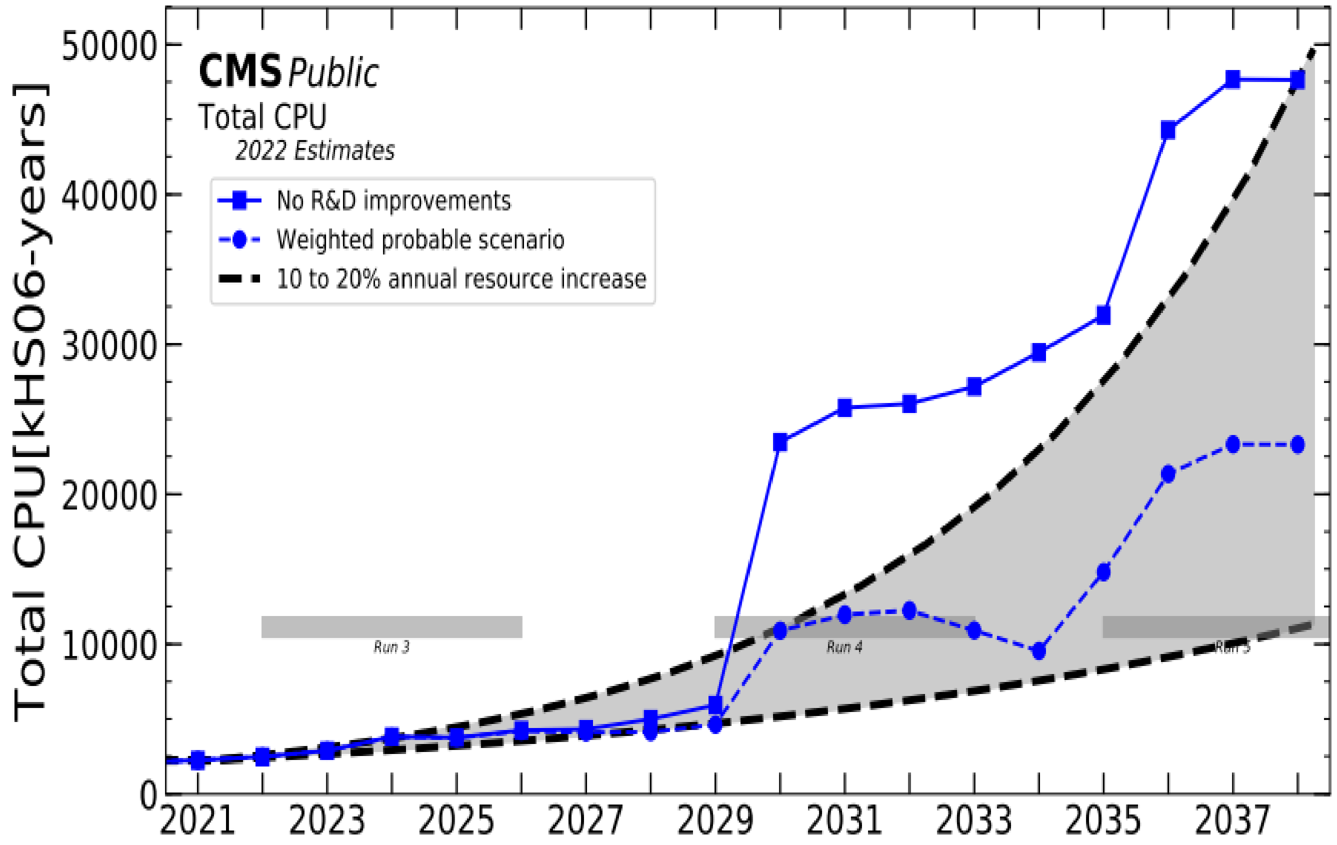
Time schedule was changed in 2020, shift by 1.5 years
Instead of a start in 2027 it is now 2029

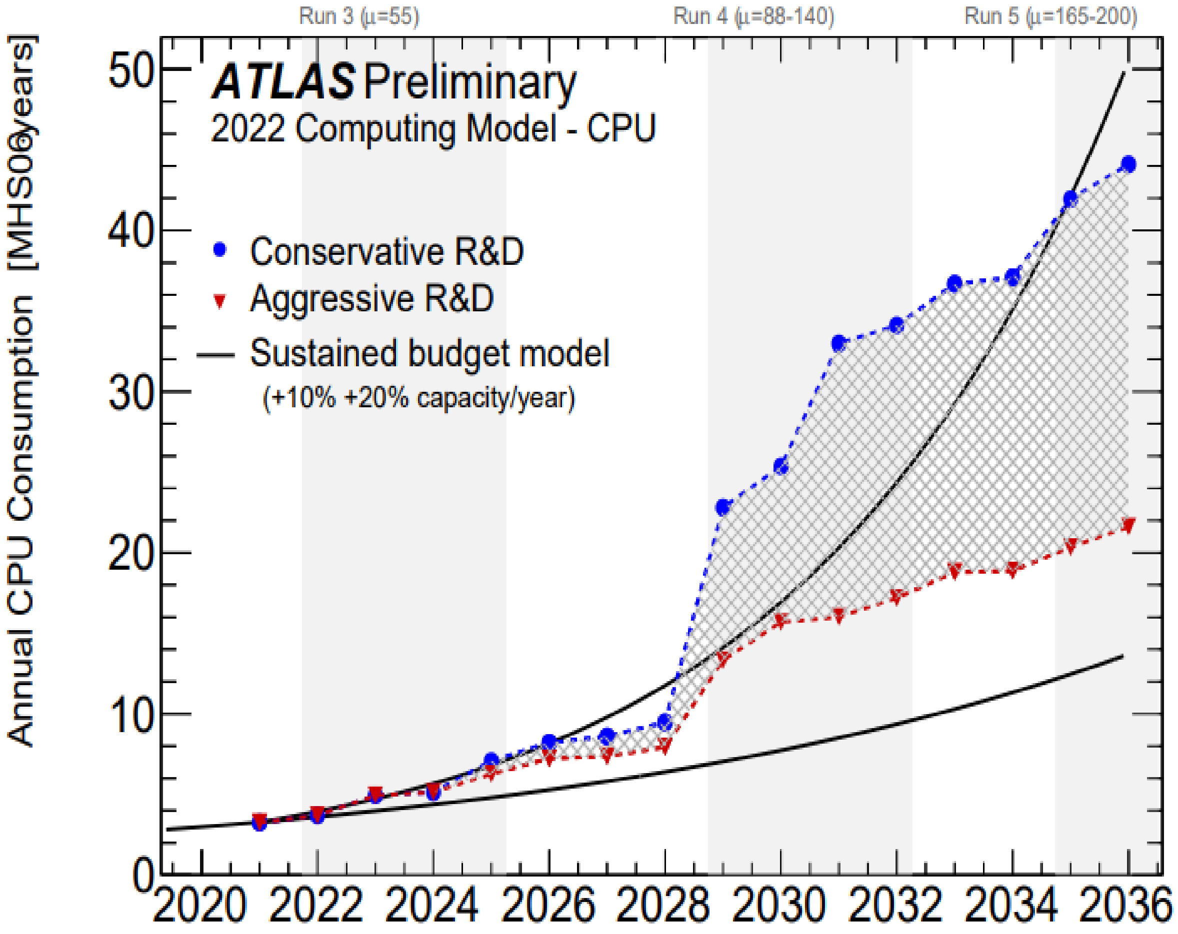
In the following estimates and calculations
it is assumed that 2029 is a 'full' running year

Experiment Predictions

Total WLCG (T0+T1+T2) resource evolution needed for HL-LHC computing, estimates from ATLAS and CMS (July 2022)

Important: two scenarios with a difference of about a factor 2

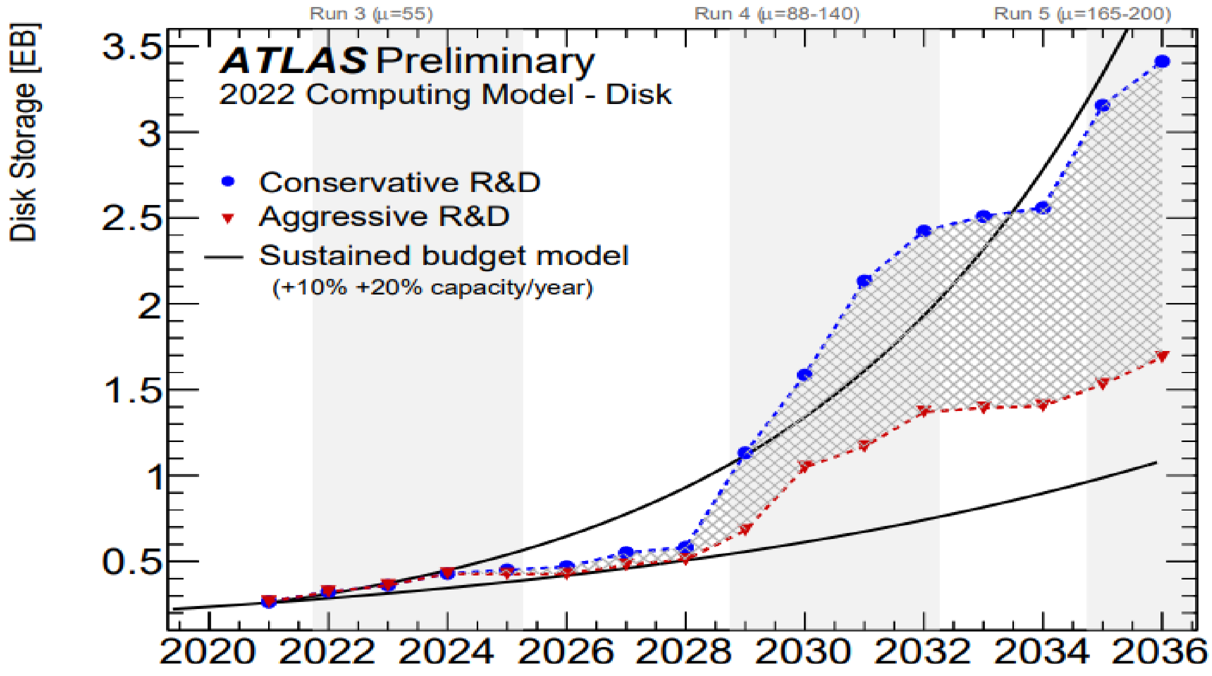


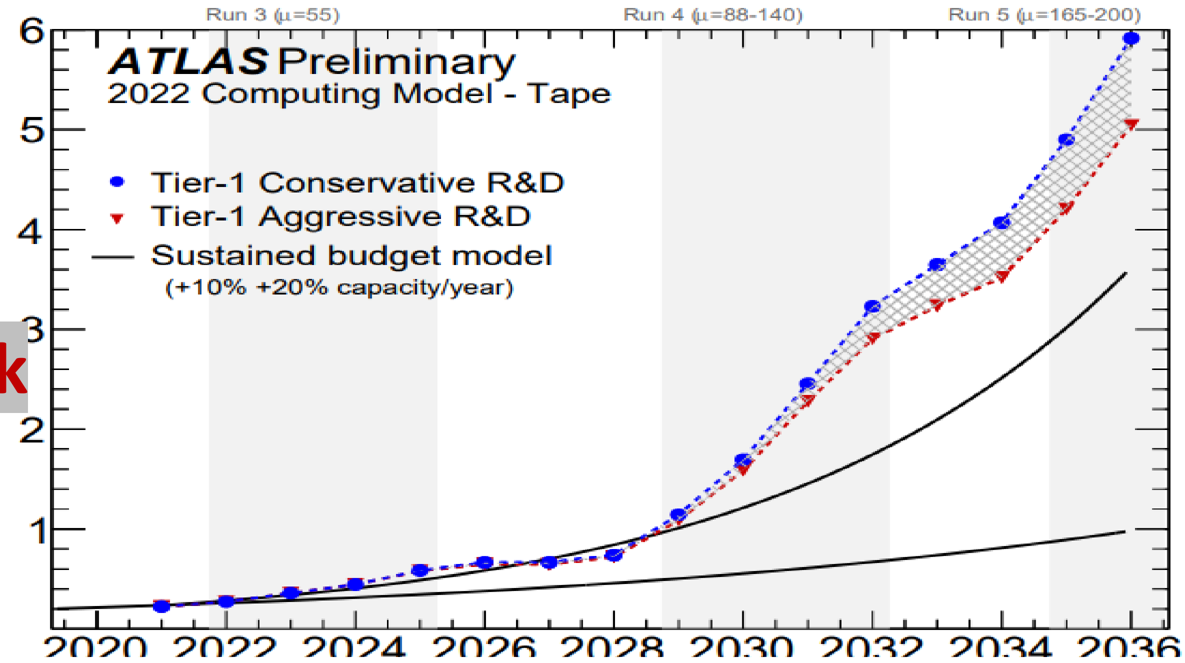


CPU processing capacity

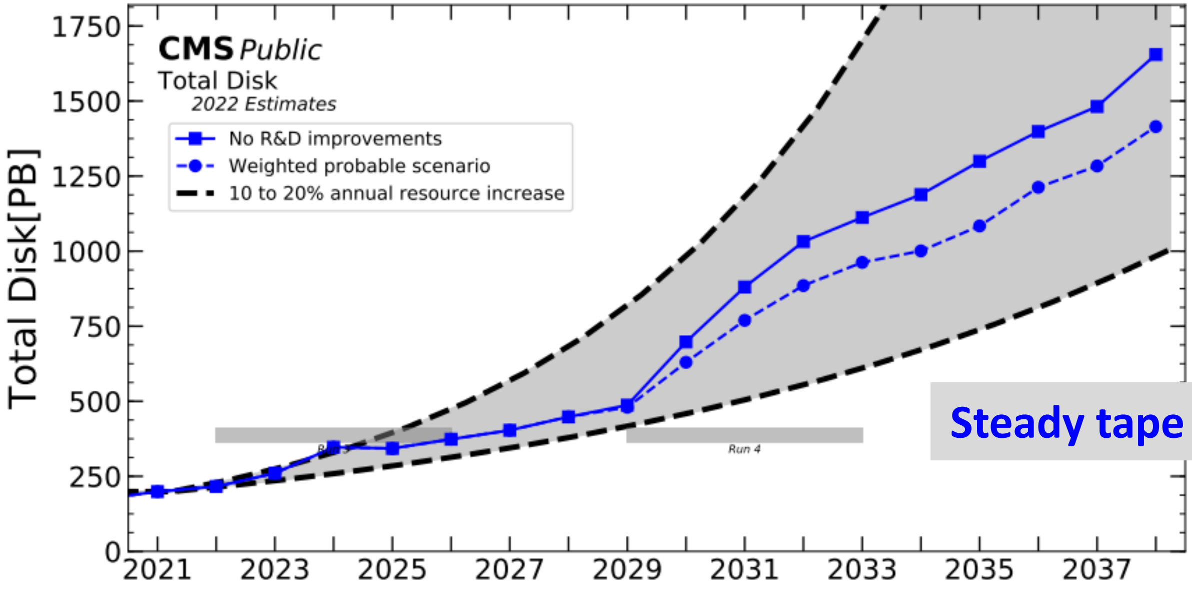
These estimates will we refined in the future and regularly discussed with the LHCC

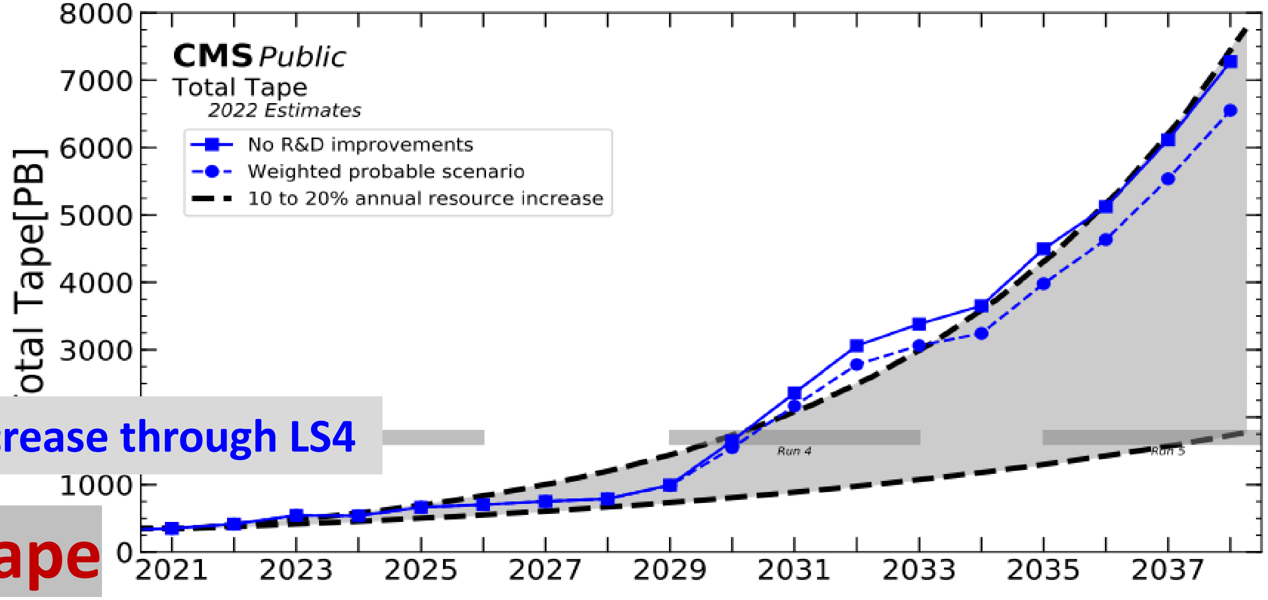
Experiment Predictions II





Disk

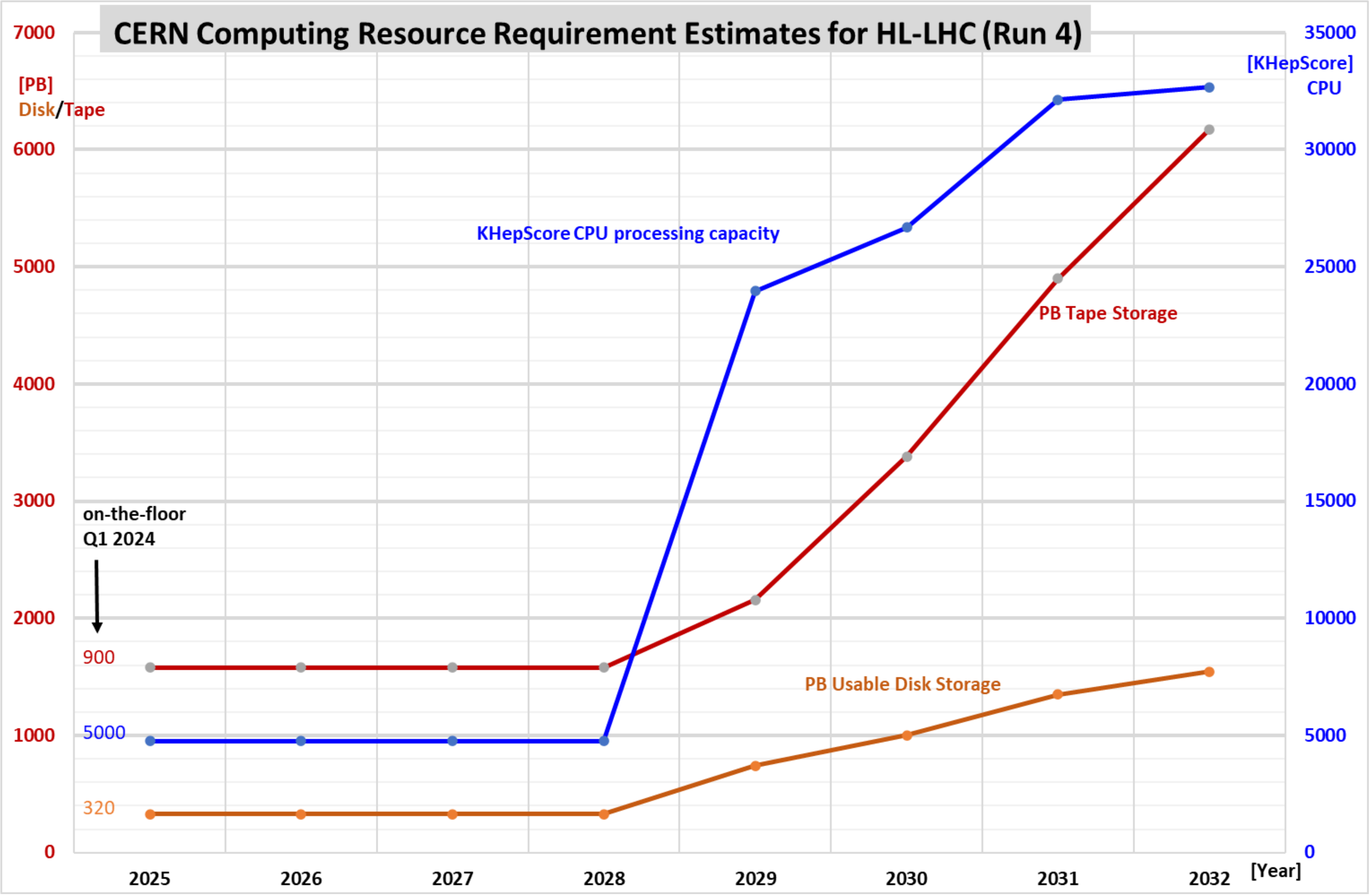




Steady tape increase through LS4

Tape

Derived CERN IT Predictions



Reference values from 2024/2025

Share of the T0 WLCG resources compared to the sum (T0+T1+T2) (average over all 4 Experiments)

CPU : 25%

Disk : 20%

Tape : 45%

CERN IT estimates are based on:

- **Averaged estimates of ATLAS+CMS**
- **ALICE+LHCb == factor 1.5 * ATLAS/CMS**
- **Assume 20% non-LHC needs**
- **Assume 15 % headroom/contingency**

The next pages are looking into the different computing areas (Technology and Markets)

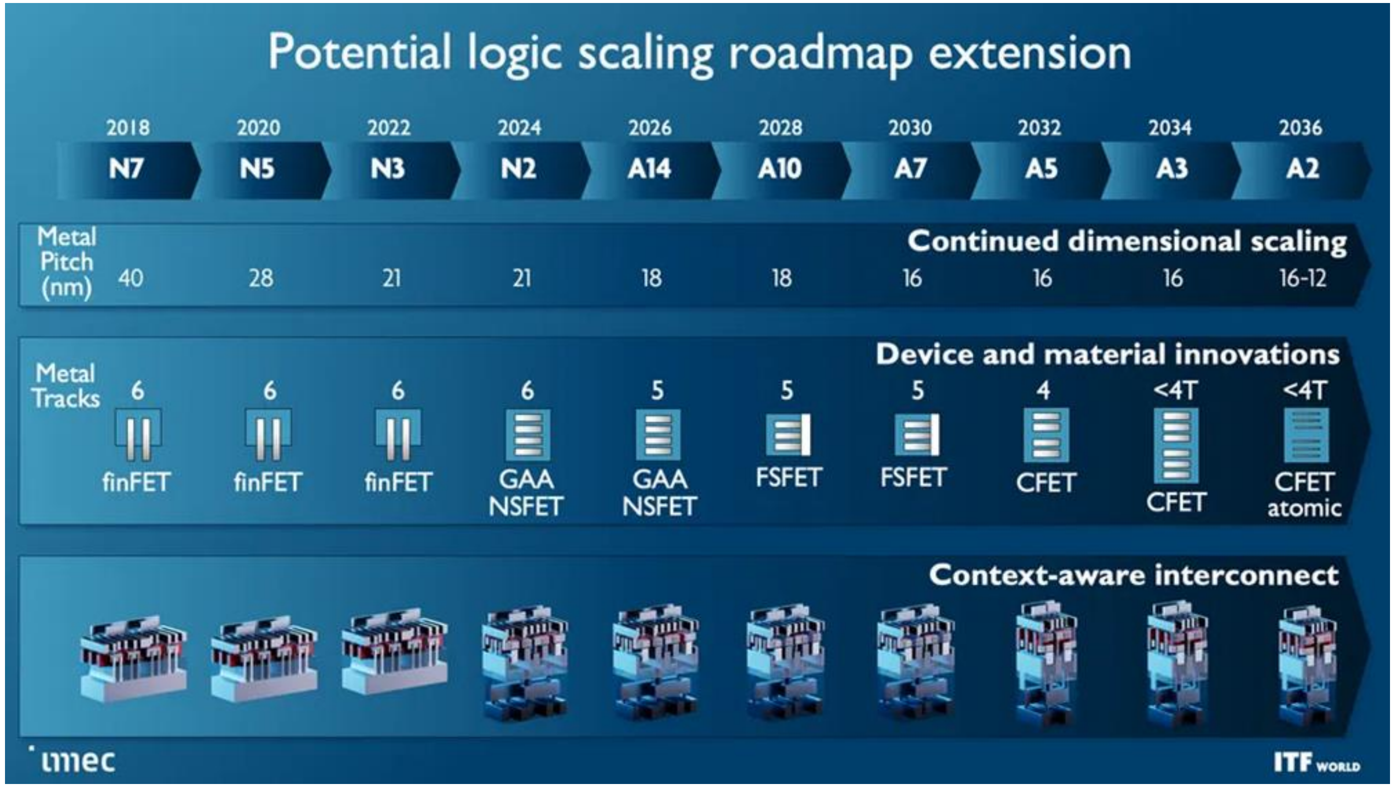
- **CPU processing**
- **Network**
- **Disk storage**
- **Tape storage**

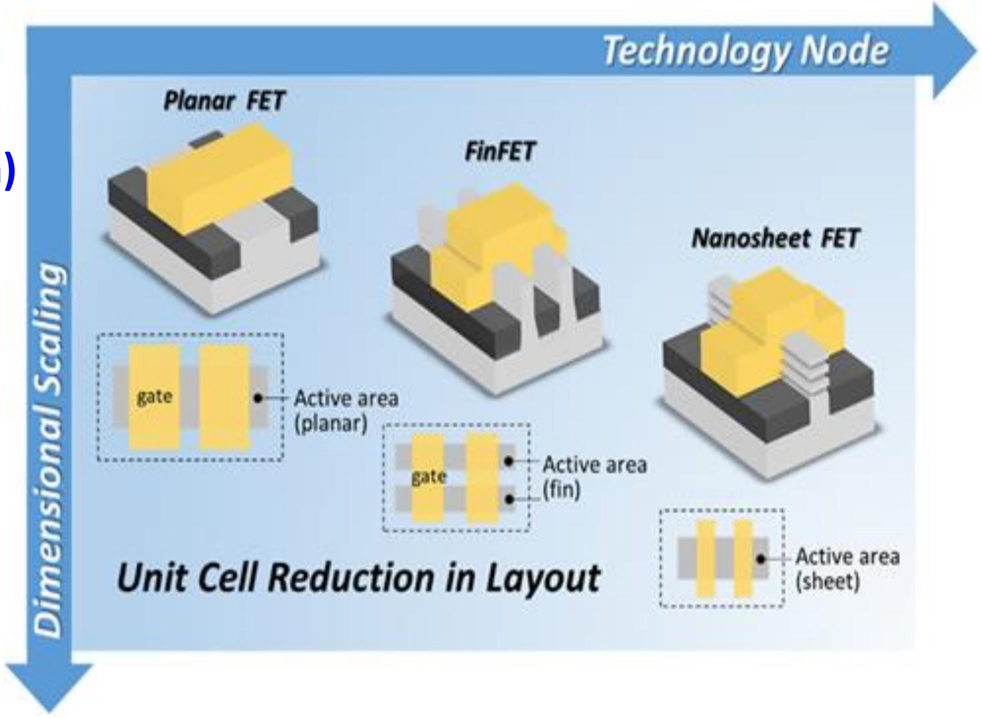
trying to identify problematic (or well working) points which might affect the cost and operations

Processor Fabrication I

Basic building block still transistor Field-Effect-Transistor (FET) made out of silicon
Sophisticated gate structure evolution enables structure scaling, faster switching, higher currents, less leakage, …….

Lithography process names
Current most advanced = N3 (‘3nm’), moving later to A14 (14 Angstrom = 1.4 nm)
Just names, nothing to do with the lithographic structure sizes on the wafer
→ on-chip 20-40 nm pitches





Detailed plan for the next ~10 years
Long term technology roadmap in good shape

Processor Fabrication II

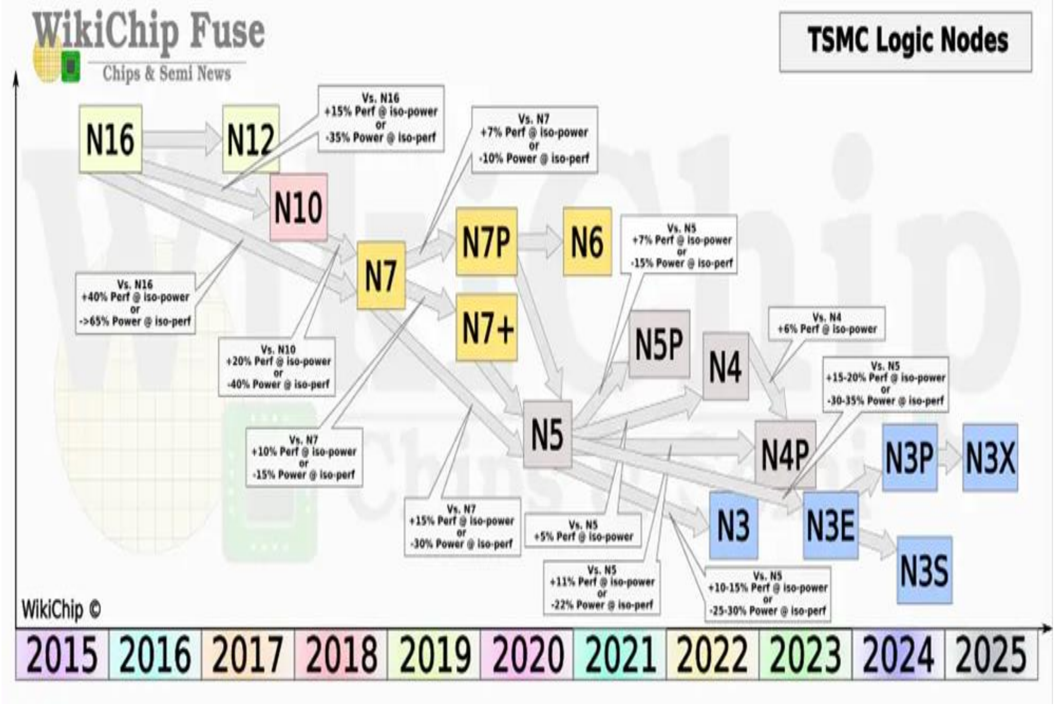
Manufacturing cost per transistor is now actually increasing with the new technology generations

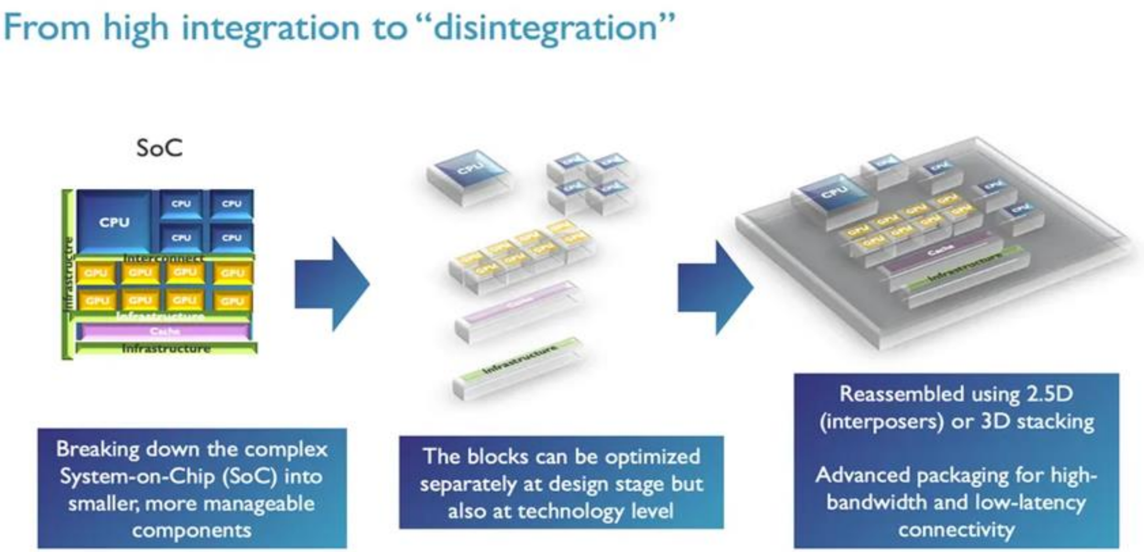
A new generation has a lot of variants and will:
- increase performance ~8-10%

<u>OR</u>

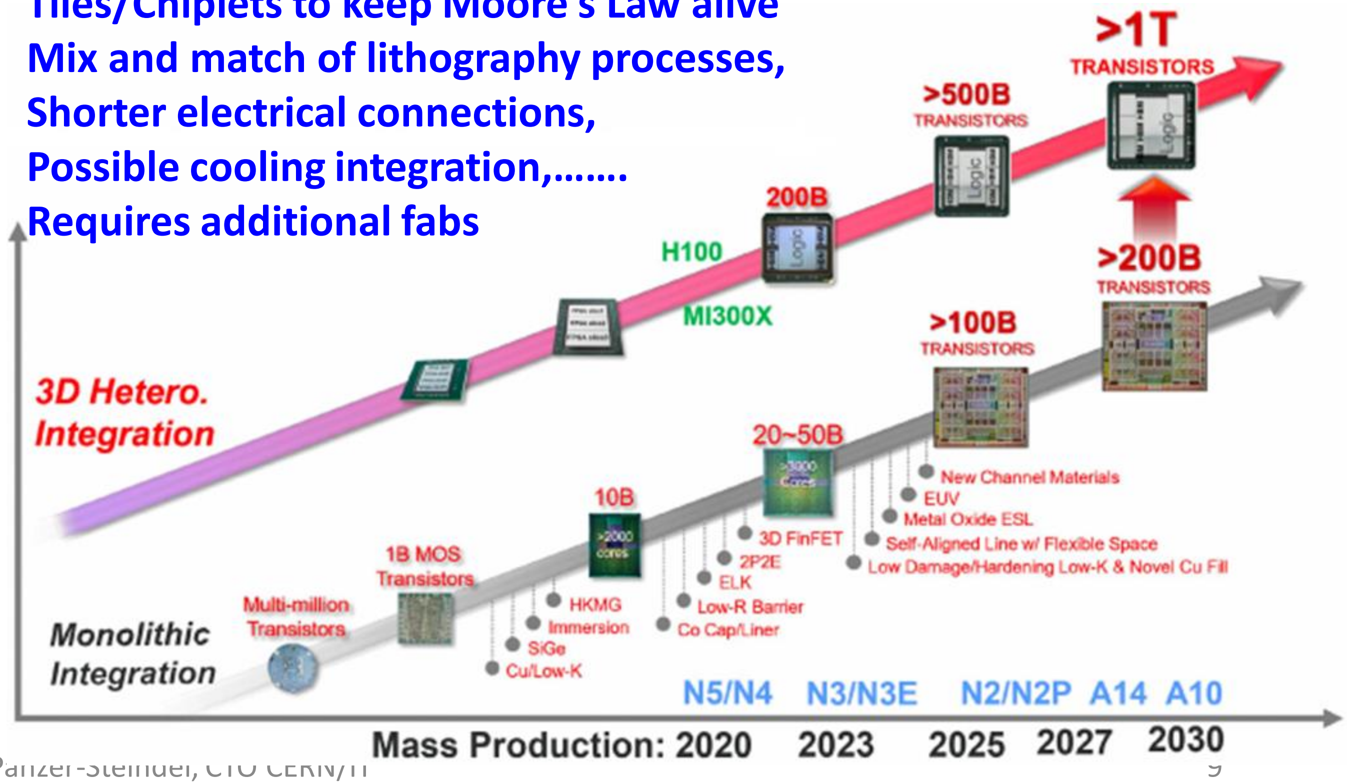
- decrease power consumption ~15-20%

TSMC A16 process announced for 2026





Tiles/Chiplets to keep Moore's Law alive
Mix and match of lithography processes,
Shorter electrical connections,
Possible cooling integration,…….
Requires additional fabs

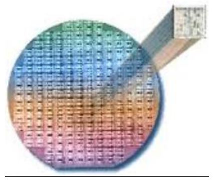


Processor Industry

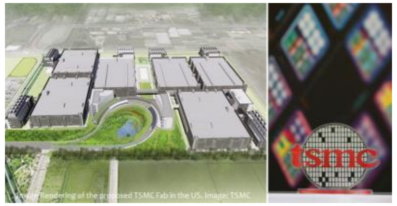
Only 3 companies in the world capable of fabricating leading-edge chips (“5nm node” or less)

		Samsung	TSMC	Intel
Revenue	:	44 B$/y	67B$/y	54 B$/y
Fabs (leading edge)	:	6	10 (5)	15 (7)
Sites	:	South Korea	Taiwan +(40B$ investment Arizona)	US, Israel, Europe
Customer (main)	:	Smartphones: ARM	AMD : all CPU and GPU; Apple : ARM Nvidia: GPU Sony, Microsoft: game console CPU+GPU some Intel processors (2024)	Intel: CPU and integrated GPU

Very complex fabrication process
A wafer stays 3 month in a fab and runs through ~1000 processing steps



- A new fab requires investments of >10B$
- All 3 companies want to invest each ~100B$ during the next years in new fabrication units

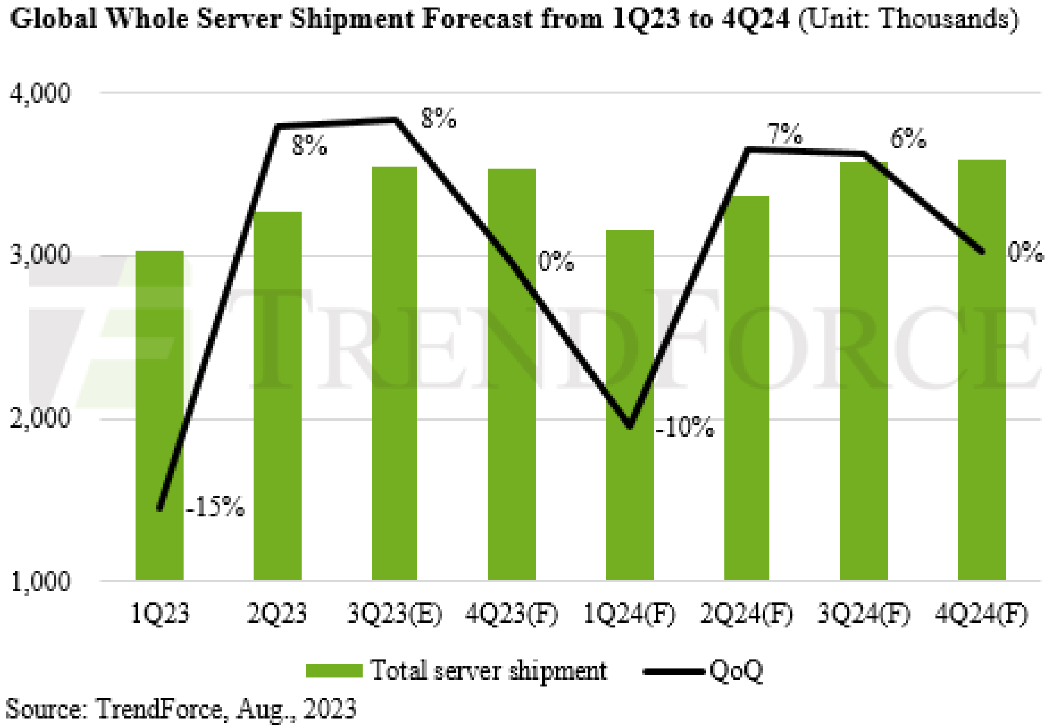


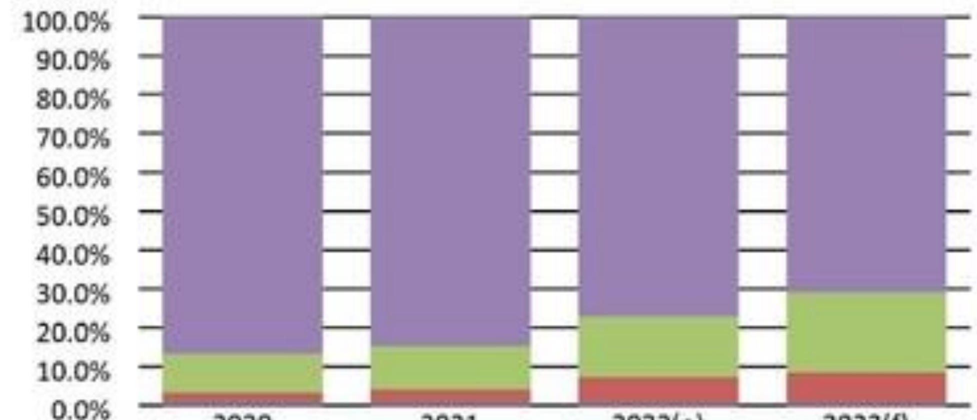
Very few companies can provide:
- Ultra-pure silicon wafers
- Special photoresist
- Precise photolithography masks
- Ultra-pure chemicals

Only one company (ASML) provides EUV Extreme Ultraviolet Lithography equipment

Monopoly, single source suppliers…….
Very sensitive to political, economical, environmental ‘hiccups’

Server Market





	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%

Growth rate is fluctuating, total number of servers sold per year is about 13.5 M units, total revenue is at the level of 110 B$

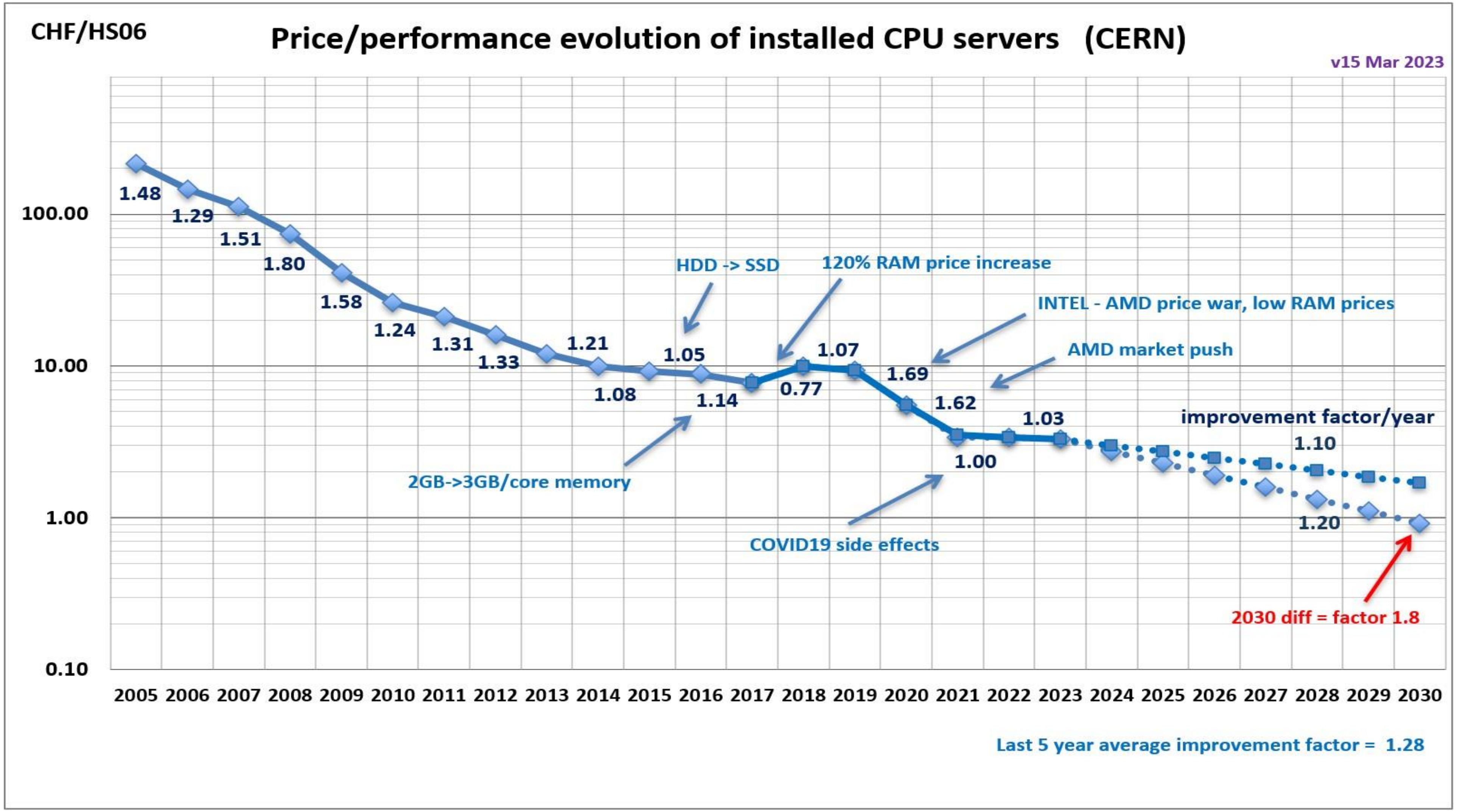
(2023 unit sales: 240 M PCs+Notebooks, ~1.2 B smartphones)

ARM server first market introduction ~10 years ago
Current ARM share in the server Market is about 8%, vast majority is Graviton from Amazon.

Ampere server <1%, revenue 0.5 B$ per year, multi-billion $ external funding, only ARM server company in the market
Amazon, Google, Microsoft are designing their own ARM systems

Need to carefully watch the ARM server evolution, still too early to invest on a larger scale

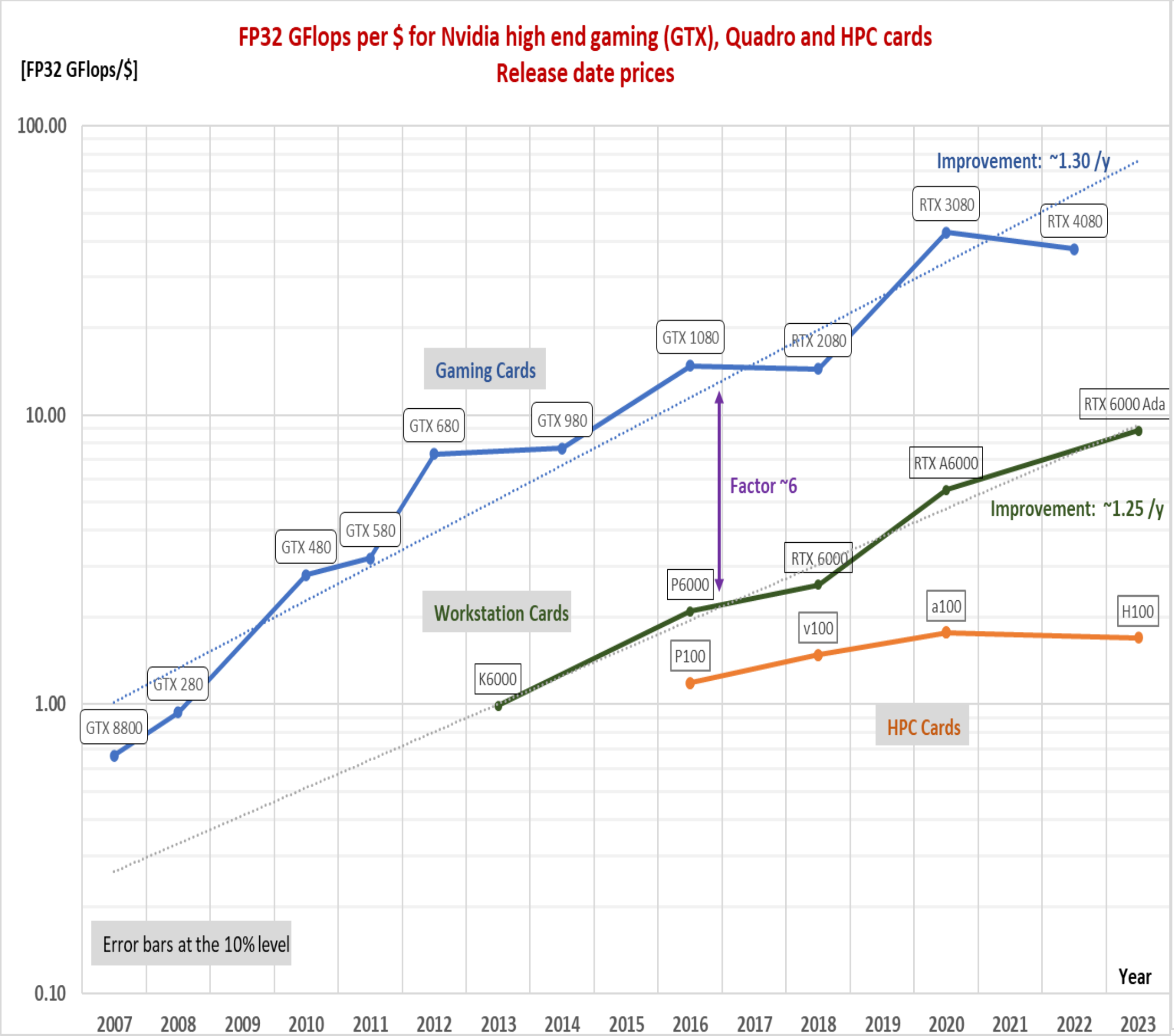
CPU processing costs

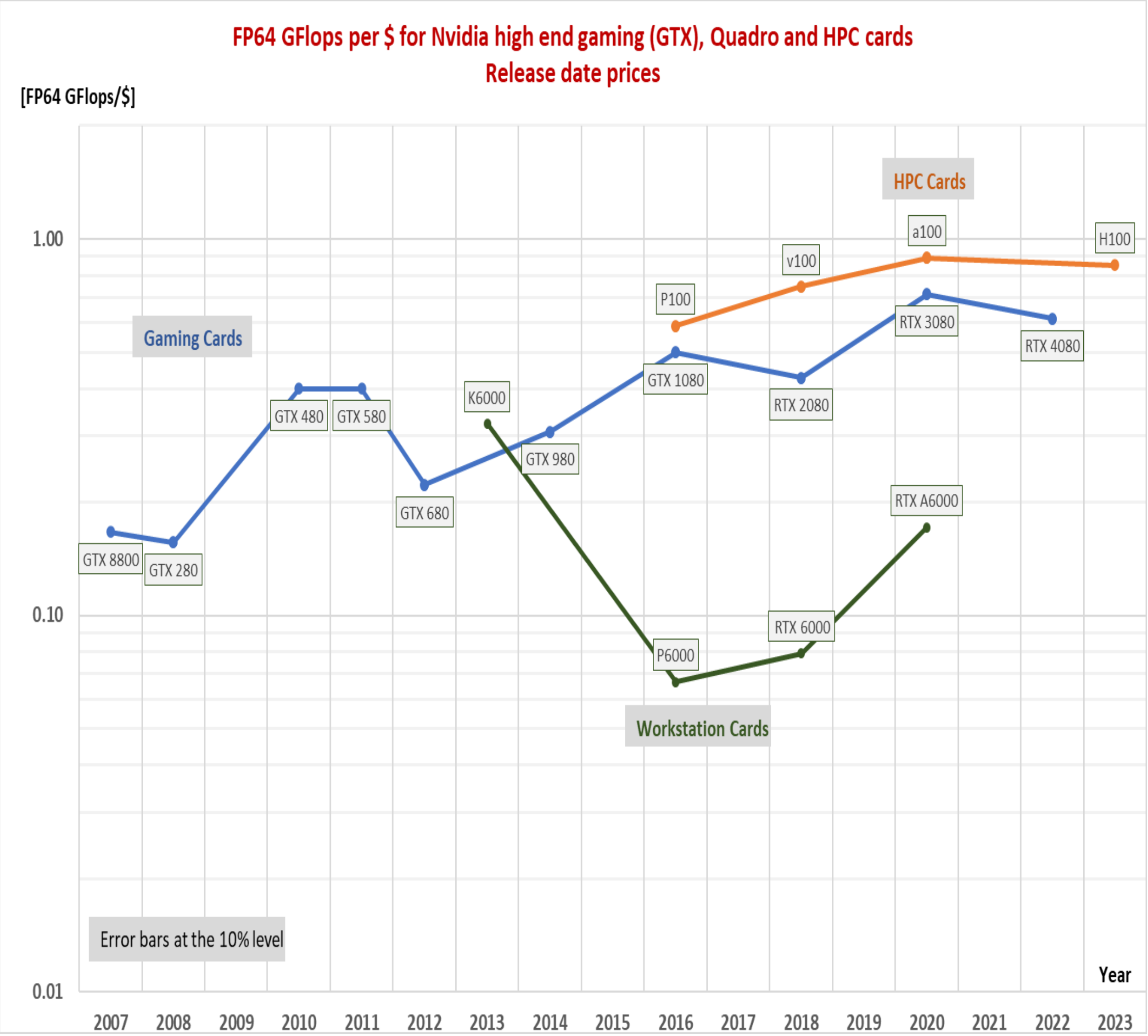


CHF per HS06/HEPScore
For a complete server
(CPU, Memory, local disks, efficient power supply, motherboard, NICs)

Average over the last 5 years still in the 20% range
But slowdown during the last 2 years -- side effects of economic and political events

GPU processing I





ML/AI/ChatGPT hype causes some market frenzy; Very volatile and high prices, will continue for the next ~2 years (NVIDIA H100 has a profit gain of ~1000%)

GPU processing II

The new Blackwell (B100, B200) HPC card from Nvidia will provide a large improvement in terms of ML performance per $, but not yet clear whether this is true for the FP64 HPC performance.
Nvidia strategy change: no single cards sales, but rather entire systems e.g. 72 GPUs + CPUs in a rack, > 50 KW, 3 M$

Site view

For GPUs 3 communities to serve, best price/performance:

- 16 bit and lower, ML → only HPC cards
- 32 bit algorithms → workstation cards
- 64 bit Engineering → only HPC cards

Online usage in ALICE, CMS and LHCb: Specific applications, partly special commercial deals → 'skewed' TCO
Not clear whether and how GPU's will play a role in the offline processing of Run 4, 32bit algorithm versus ML

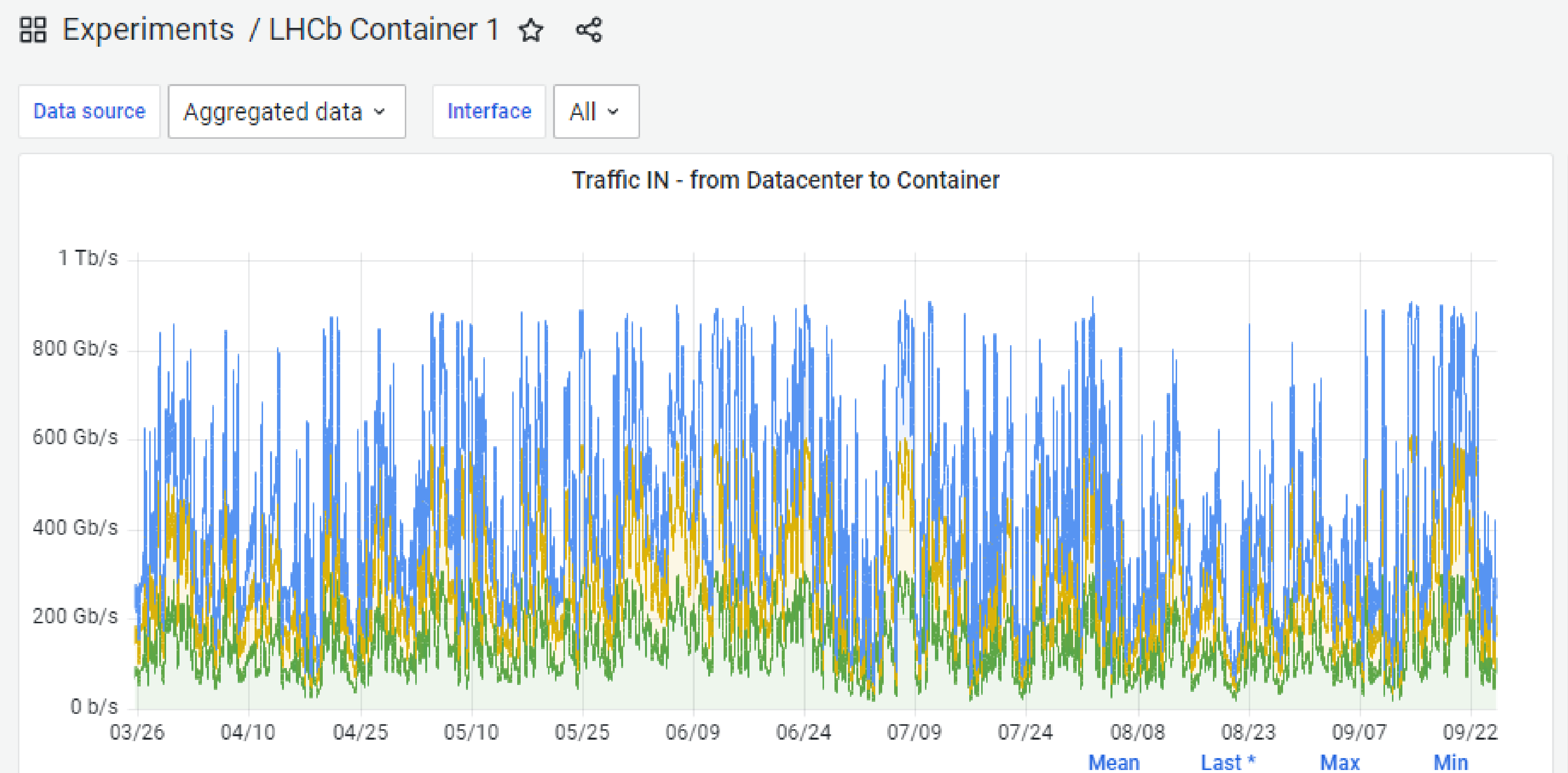
Definitely need large GPU cluster all the time for code and ML development (CERN IT has currently ~200 GPUs in production)

Need a clear TCO view by 2027 for GPUs at the latest

Operation, power side-effects, delivery times = state of the market, will need 'big-boxes' (4 GPU at least) to get the cost down (infrastructure overhead) But this will cause rack space-power Tetris complications
→ Major processing purchases starting in 2028, CPU/GPU mixture to be clarified

CPU Processing Network

CERN T0, aggregate network performance for LHCb container 1 containing ~1100 CPU server (<u>10 Gbit NIC each</u>) running a variety of jobs from processing to train-analysis



	Mean	Last *	Max	Min
Incoming l6045-c1-rjuxl-1_ae129	142 Gb/s	80.9 Gb/s	319 Gb/s	17.9 Gb/s
Incoming l6045-c1-rjuxl-1_ae28	142 Gb/s	88.7 Gb/s	313 Gb/s	17.1 Gb/s
Incoming l6045-c1-rjuxl-1_ae29	141 Gb/s	86.0 Gb/s	319 Gb/s	17.0 Gb/s

What happens if one changes the 'model', i.e. much more Analysis activities ??

Sample of the total network traffic between CERN IT buildings
(the other direction is factors lower)

The plot 'translates' into an average = 0.4 Gbit/s, peak = 0.9 Gbit/s network traffic per server
→ No network problems, sufficient headroom

Processing parameter comparison
CMS run 3 : 360 HS06/ev, 1.1 MB/ev, PU 62
CMS Run 4 : 3200 HS06/ev, 4.3 MB/ev, PU 140
→ The expected IO performance per server will actually decrease for Run 4
(depends on the per core HS06 performance and number of cores per server, etc.)

Still, expect to move to 25 Gb NICs for CPU server for HL-LHC, corresponding cost increase for the network infrastructure

Power and Cooling

Latest deliveries of CPU servers consume ~700 W
The power efficiency improvements over the last 8-10 year were on average 15%/y
This has slowed down and will slow down further due to the mentioned manufacturing issues.

Assuming 10% improvements → 5 MW needed for the CPU server in 2029
Plus, the fact that one needs to consider an overlap of old and new equipment for ~6 month
to keep the pledges stable plus some services and Business Continuity in the PCC

→ Need to upgrade the CERN PDC from 4 to 8 MW in LS3, to be ready at the beginning of 2028

Cost improvements and energy efficiency are linked via the technology evolution→ one gets a certain energy efficiency for 'free'

Where is the focus ? What is the strategy ?:

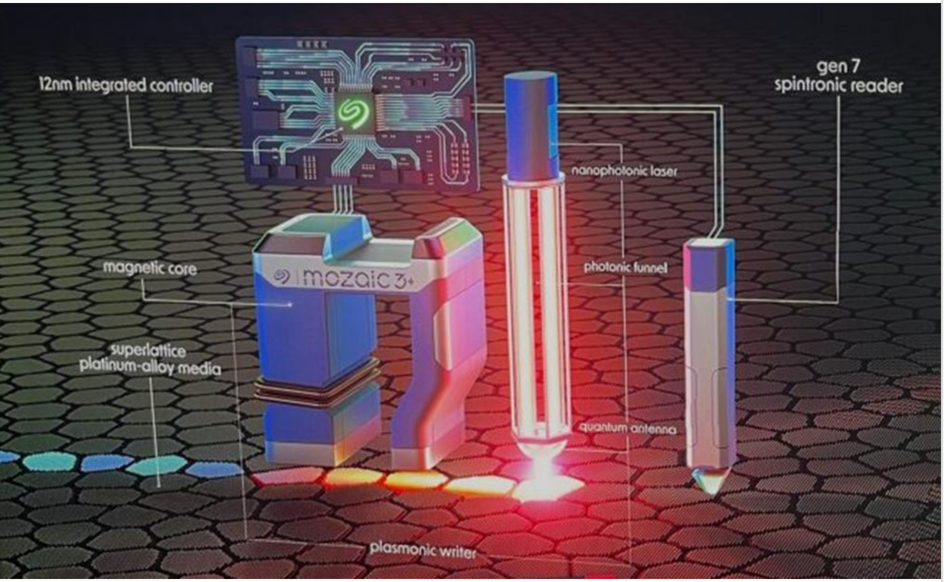
- total energy usage
- equipment energy efficiency
- sustainability
- overall cost
- CO_2 emissions

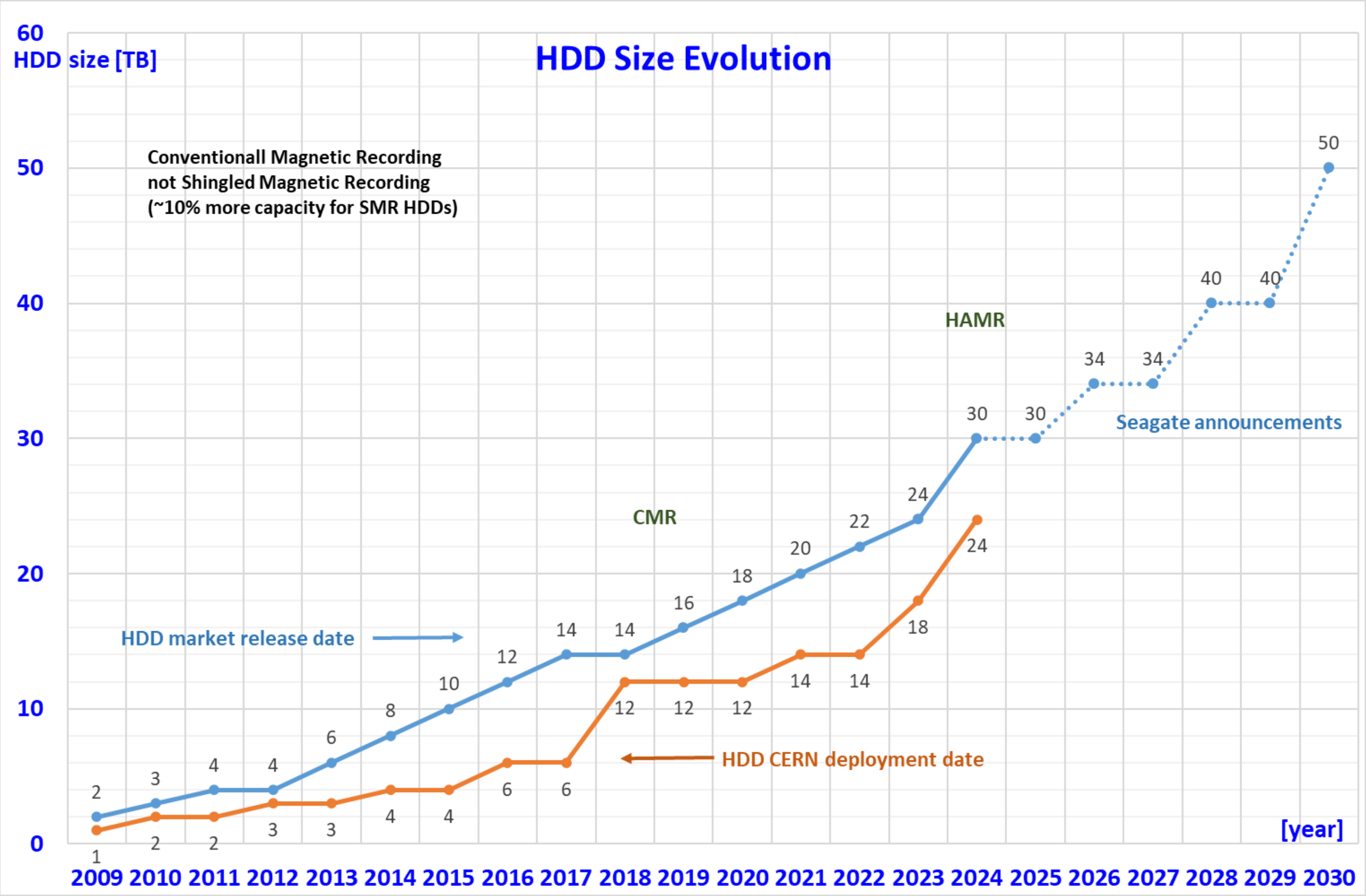
These points have partly conflicting consequences and are very strongly site dependant !!, e.g. Sustainability versus energy efficient equipment,
Today at CERN IT : to improve the total energy consumption one could do replacements with more energy efficient equipment
→ to save 100 KW one needs to invest 200 KCHF

Disk Storage Technology I

HAMR (Heat-Assisted-Magnetic-Recording) introduced in 2016, anticipated 100 TB drives in 2023/2024
Technology into market is about 7 years late, complex – expensive - low yield

First 30 TB disk (10 platter, Seagate) market availability in Q1 2024, but not for everybody plus Seagate will try to sell appliances, not single disks
Not clear how the prices will evolve
Maybe 50 TB disk earlier (2028)





very little performance improvements (MB/S, IOPS) expected in the next years

Disk Storage Technology II

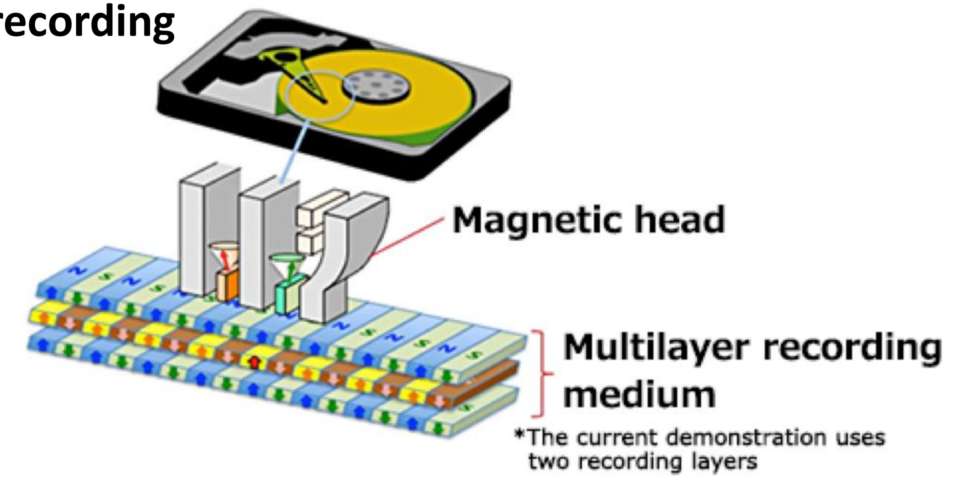
Western Digital is still trying to get the most of the existing technology step with OptiNand (integrating flash in the drive) and partly MAMR (Microwave Assisted Magnetic Recording), 18 -24 month late in adopting HAMR

MAMR should have in principle the same density curve as HAMR

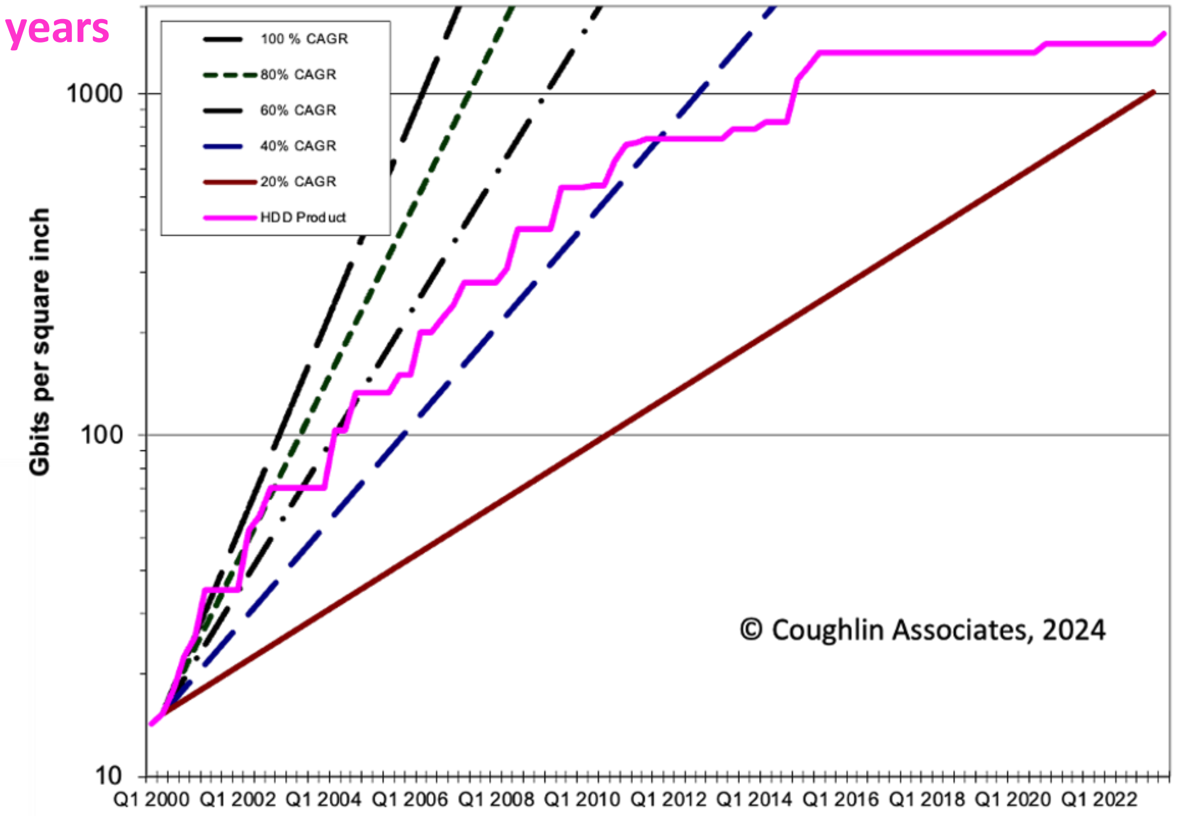
Areal density improvements have been slow during the last 8 years

PMR levels out at about	1.1 Tbit/in2
TDMR + SMR	1.4 Tbit/in2
HAMR	>1.5 Tbit/in2

Toshiba has lately revealed plans for multi-layer magnetic recording



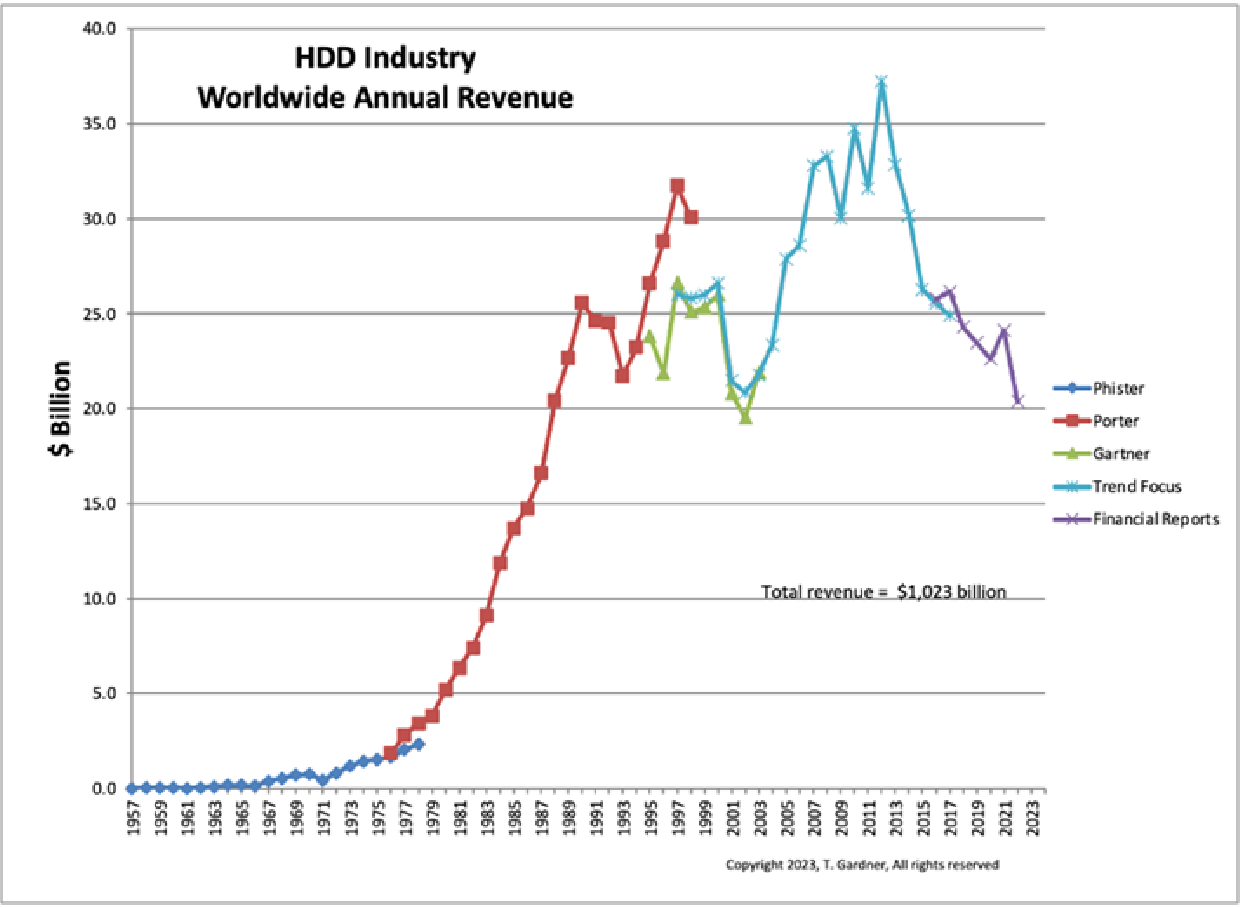
Schematic diagram of a multilayer HDD



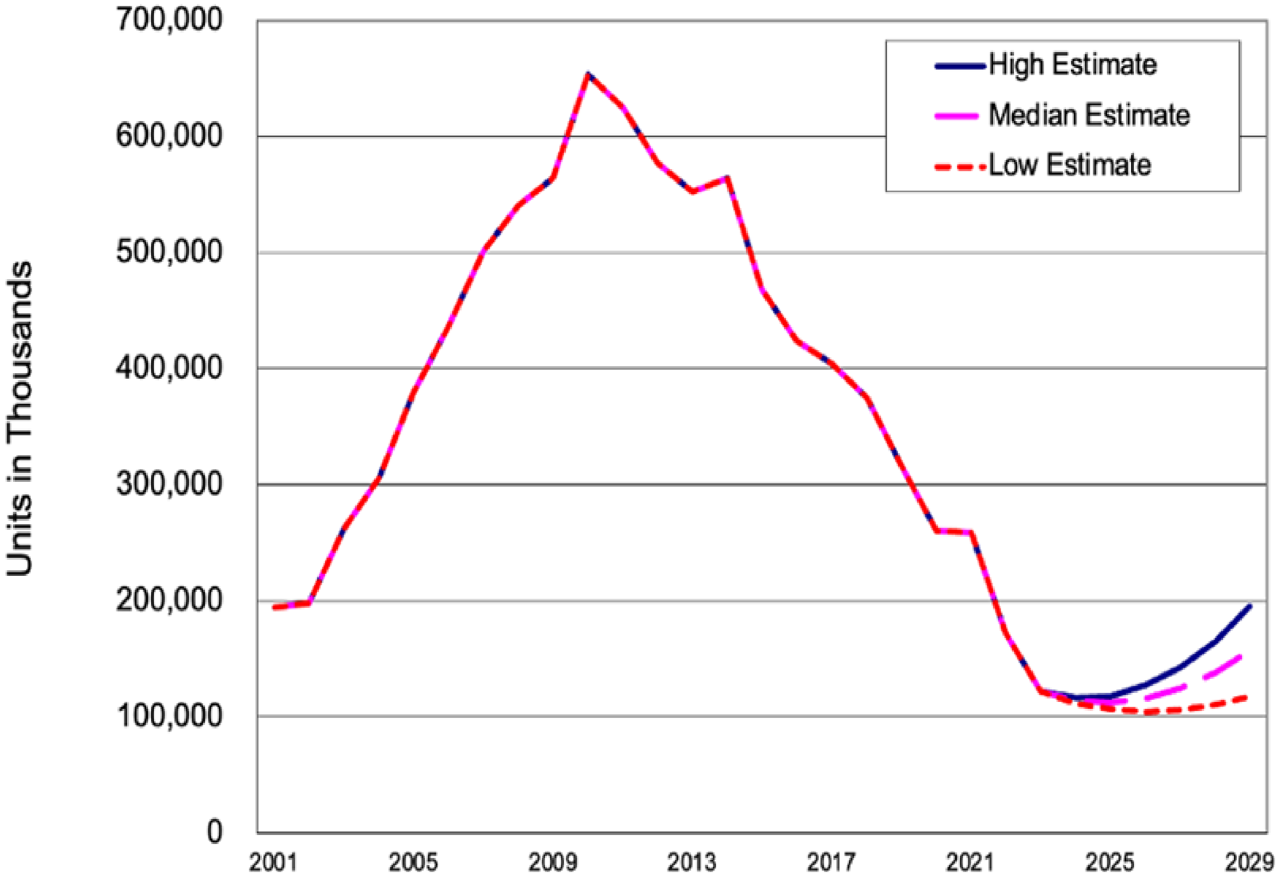
Overall, the technology roadmap is in good shape

Disk Storage Market I

HDD revenue evolution



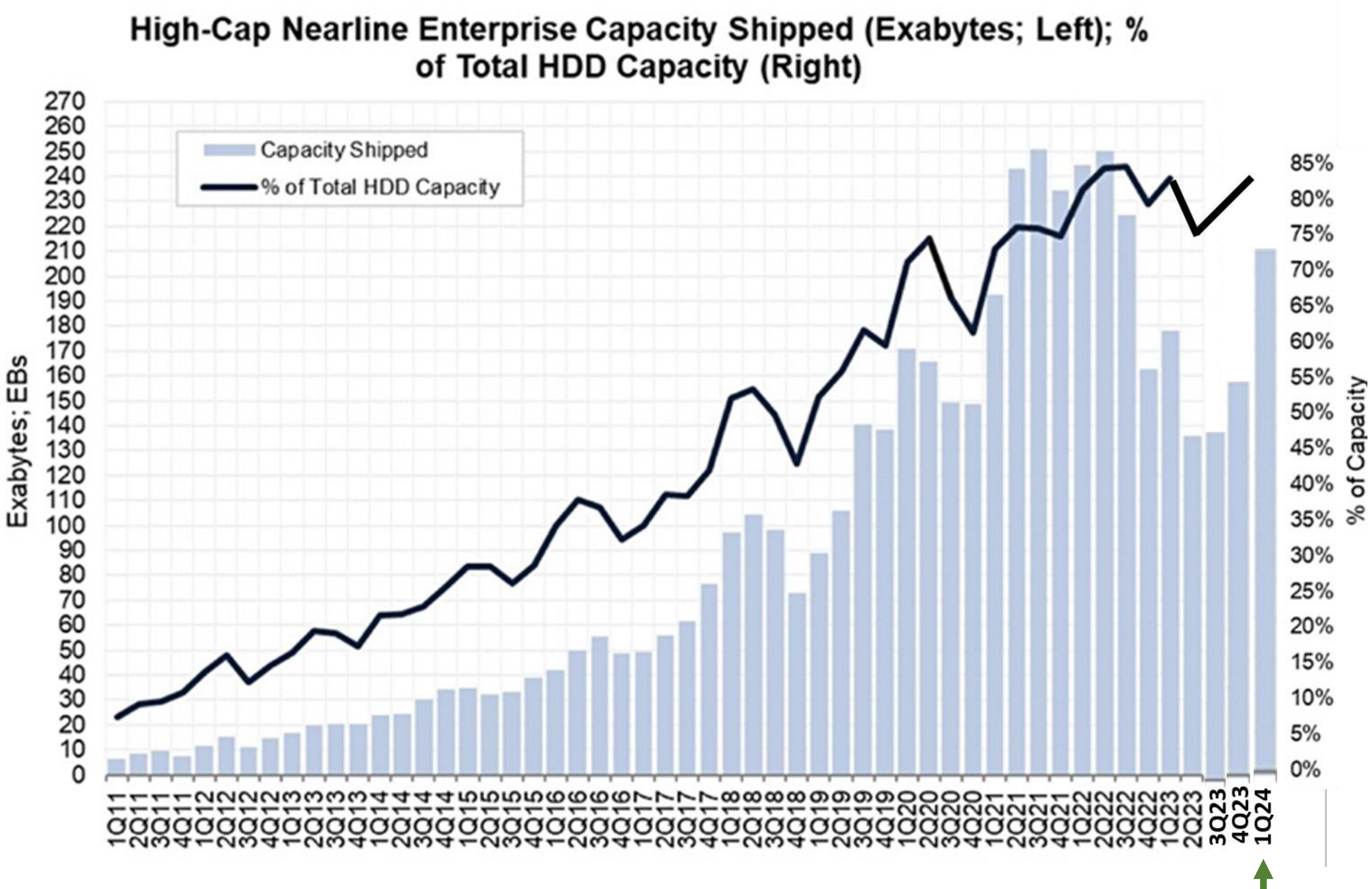
HDD unit sales per year evolution



2023: 20 B$ revenues, 130 M units, 900 EB shipped

Revenues going down and number of HDDs shipped stabilizing on a low level

Disk Storage Market II

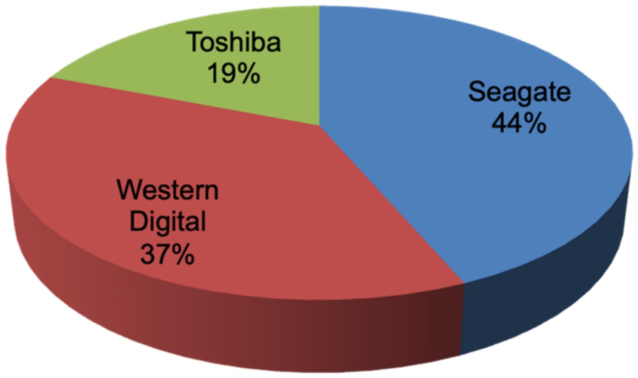


Market is expected to recover in 2024

Capacity drives (>8TB) dominating the shipments
Strong trend towards SMR drives

Notebooks are already 100% equipped with SSDs
PCs in the near future

Economic turbulences and stockpiling during COVID-19 → revenues and sales dropped in 2023



2023 HDD Market Share to Date
COUGHLIN ASSOCIATES IMAGE

Only 3 companies dominating the market

Disk Storage Performance

Currently the disk server architecture is cost/TB optimized i.e. performance comes for 'free'.
All pledges are about size and do not yet contain explicit performance values (heavily site architecture dependent)

The needed 740 PB disk storage in 2029 at the CERN T0 would result in only 150 disk server assuming:

- **Keep the current architecture: one front-end node with 120 disks**
- **50 TB disks**
- **Erasure-Coding 10+2 (like ALICE O2)**

Today we are running EOS with ~1000 disk server

1. **For the start of Run 4 one would have a factor 6 less server while the performance needs are increased by a factor >10**
2. **The performance of HDDs will only increase slowly**
3. **The total number of IO streams (== cores/jobs) will increase**

→ Expect IO problems
Thus, probably need to buy more spindles = more space = much more expensive e.g. factor 4 or more?
move to large SSDs might be as cost effective !? Today largest affordable SSDs are ~60TB → 200 TB in 2028 ?
Change the disk server architecture = small CPU server with couple of SSDs ? Mix and merge processing with storage?
Move to 400 Gb NICs plus corresponding network infrastructure !? → Overall network cost increase

SSD versus HDD Storage

SSD market: 350 M units shipped 29 B$ revenues,
260 EB shipped -- but vast majority below 2 TB capacity
(for the 260 M PC and notebooks per year)
30 million enterprise drives and less high-capacity drives

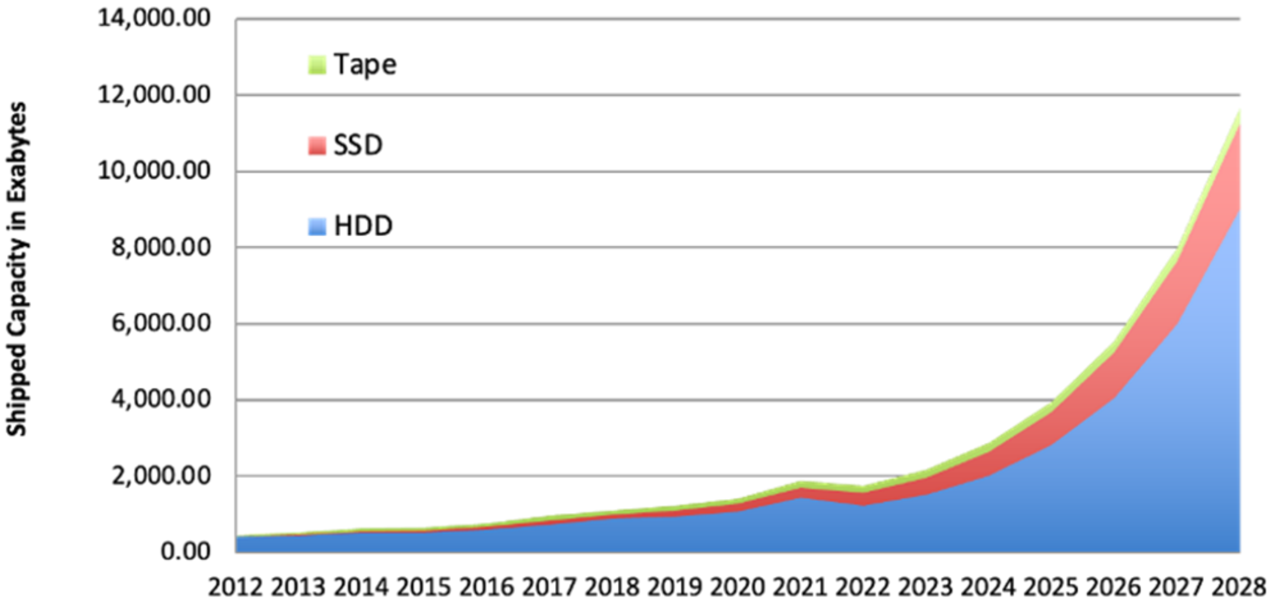
Comparison with HDDs
Choose the right metric: IO performance or capacity !

Highest SSD size today ~100 TB, affordable 61 TB

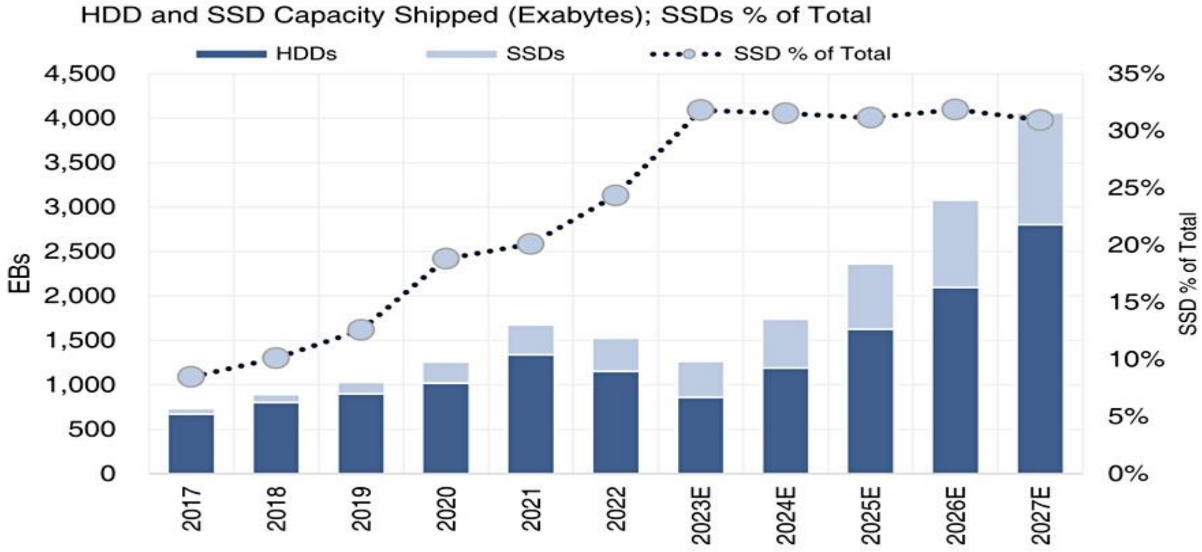
Price difference factor 3-5 in terms of CHF/TB
Idle power about the same or slightly worse

Not enough SSD manufacturing capacity to replace capacity HDDs, requires > 100B$ investments

→ SSDs will not replace capacity HDDs in the foreseeable future in terms of $/GB

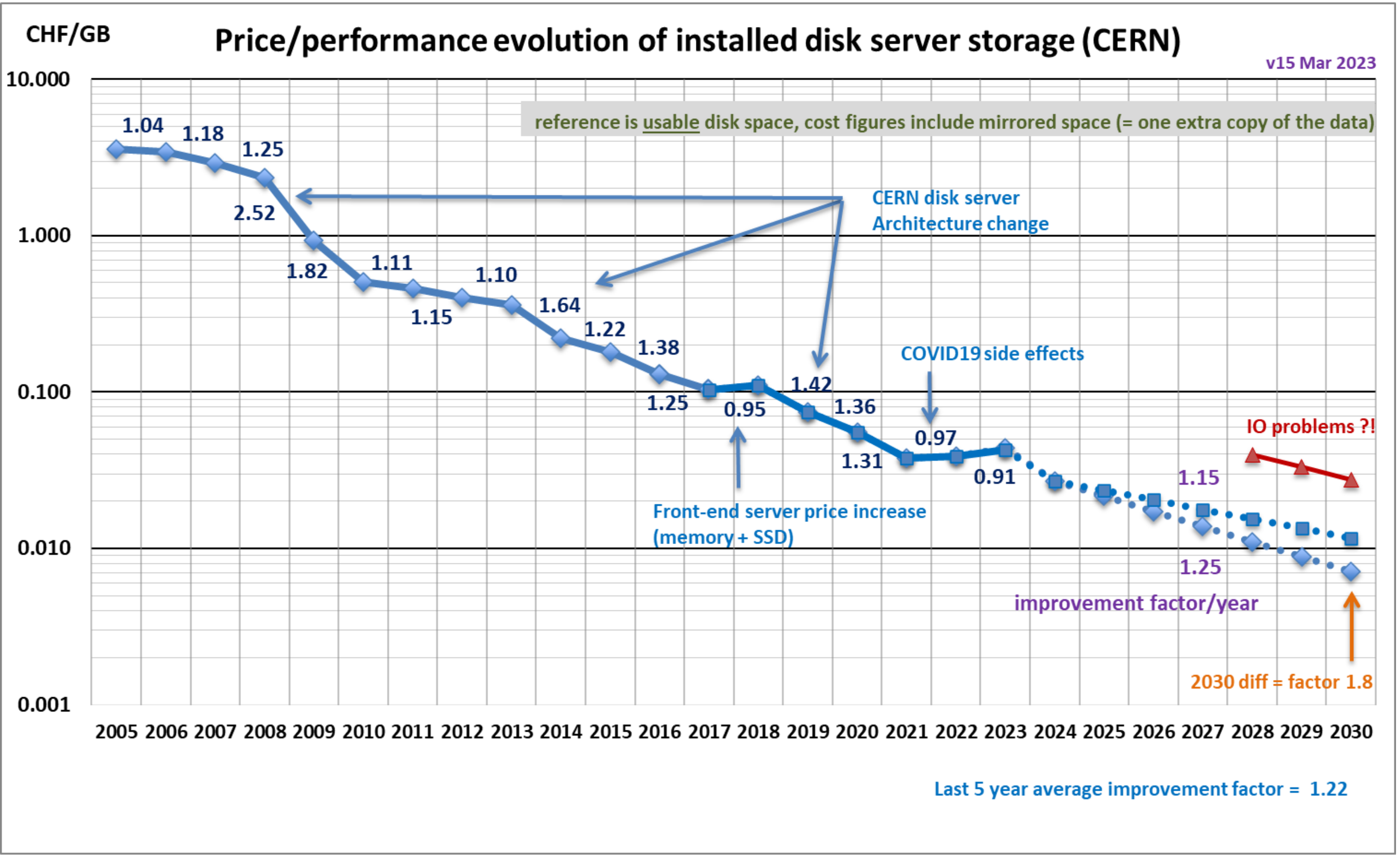


History and Projections for Digital Storage Capacity Shipments of HDDs, SSDs and Magnetic Tape
COUGHLIN ASSOCIATES CHART



Source: Gartner; Wells Fargo Securities, LLC .

Disk Storage Cost



All this assumes 'server-mirrored' space
Could have a price/TB improvement by factor ~1.5 if moving to Erasure Coding

When and what percentage of all data ?
Performance considerations

What about shingled drives ?
~10% price per GB gain
But possible software and performance caveats……

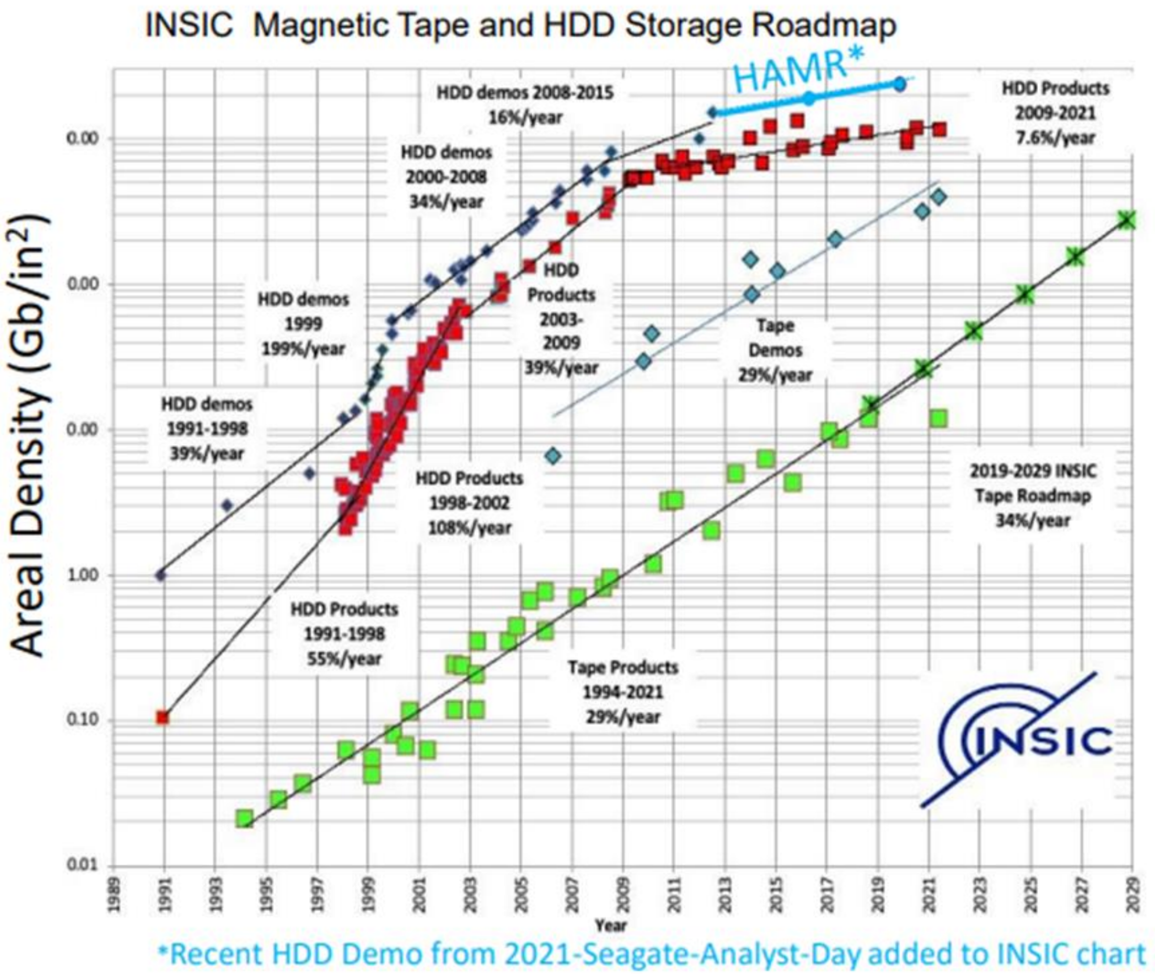
Possible cost increase in 2028 (factor 3 ?) in case of larger IO problems….

Similar development as in the CPU server case. About 20% price/performance improvements averaged over the last 5 years, but much less (actually negative trend) during the last 2 years.
How will these trends be affected by the introduction of HAMR disks ? (prices this year will increase by ~10%)

Tape Storage Technology

Tape drives from IBM, heads from WD
Media from Sony and Fujifilm

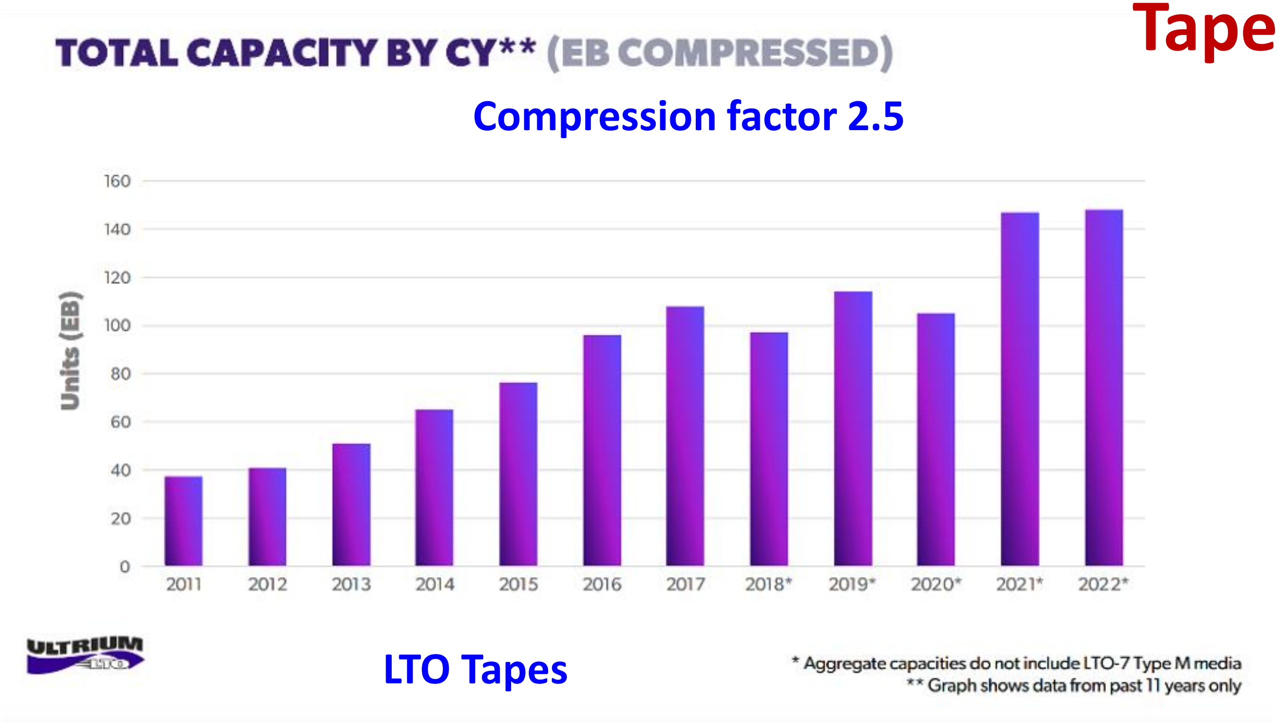
Current areal density of Barium Ferrite tapes (LTO-9) is about 12 Gb/in2
In comparison 18 TB HDDs have a density of 1022 Gb/in2

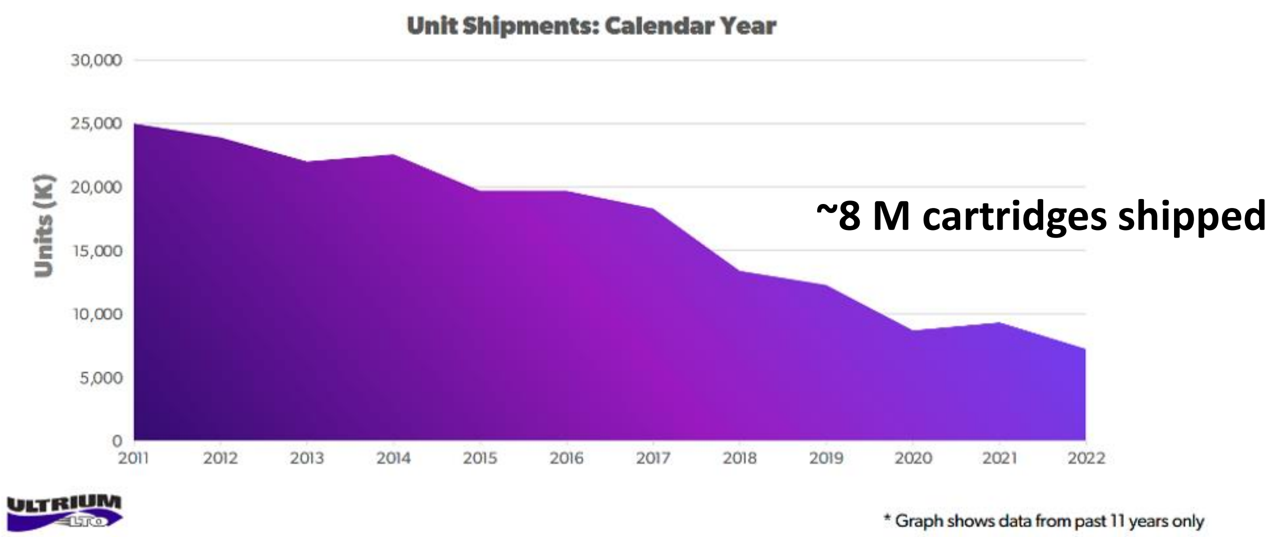


. Prototype of Strontium Ferrite media already presented in 2020 (IBM) 580 TB tape
. Latest IBM tape media for the 1170 (50 TB) cartridge is using already Strontium Ferrite

Tape technology roadmaps are in good shape

Tape Storage Market





58 EB capacity shipped in 2022, ~ 1.0 B$ media revenues
5-6 B$ total Tape market (media, drives, libraries, etc)
Market for tape media is less than 1 B$

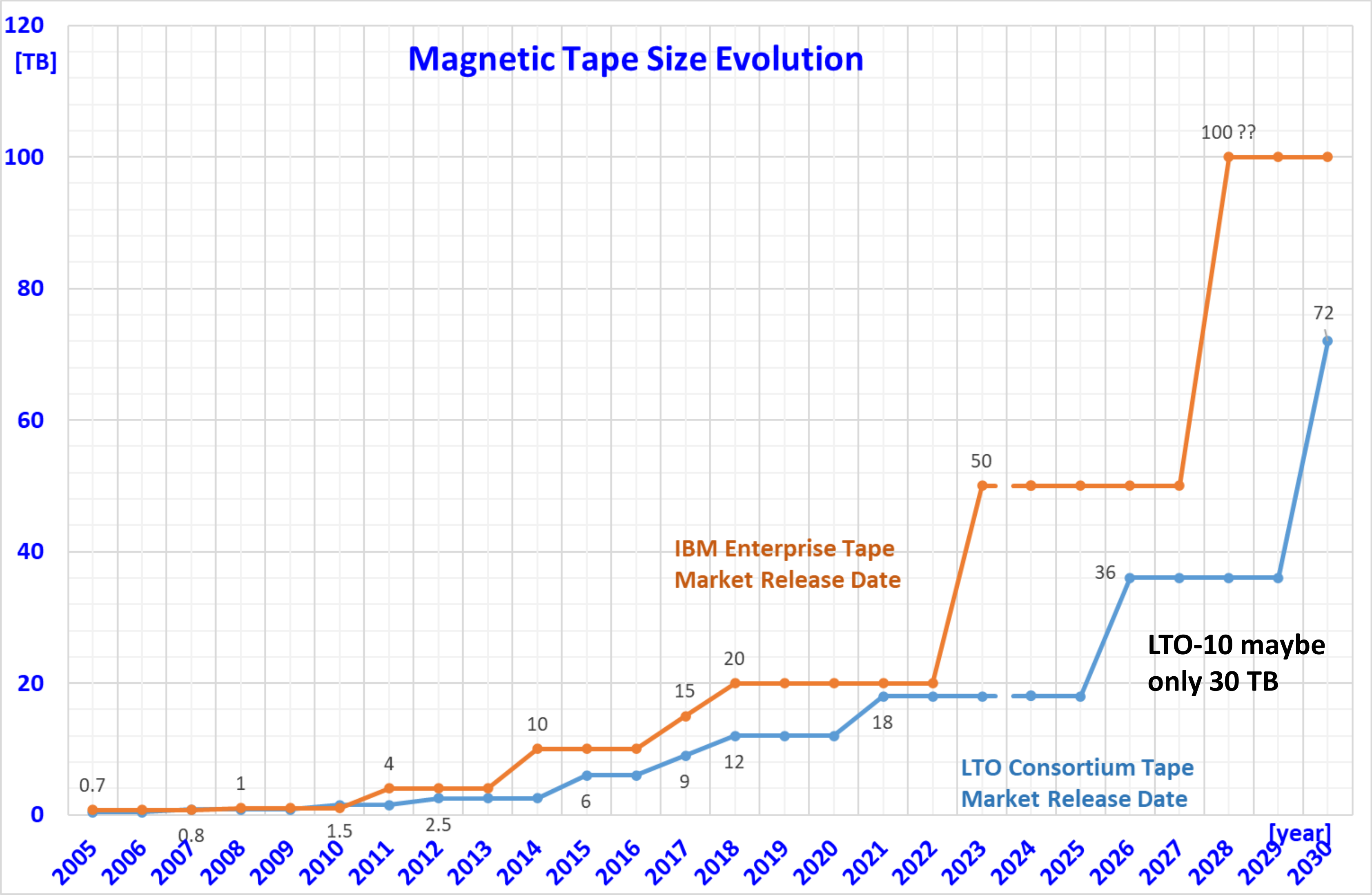
LTO tape capacity represents >85% of the market
IBM Enterprise <15%

Compared to HDD market: 20 B$ revenues
130 M HDDs shipped == 900 EB

Information about the detailed tape market is very sparse

Tape is a storage niche market

Tape Storage Infrastructure



Start of HL-LHC a mixture of 50 TB and 36 TB tapes.
CERN T0:
Already today there are ~70k LTO slots and ~30k enterprise slots in 6 tape libraries
→ ~4 EB tape space in theory

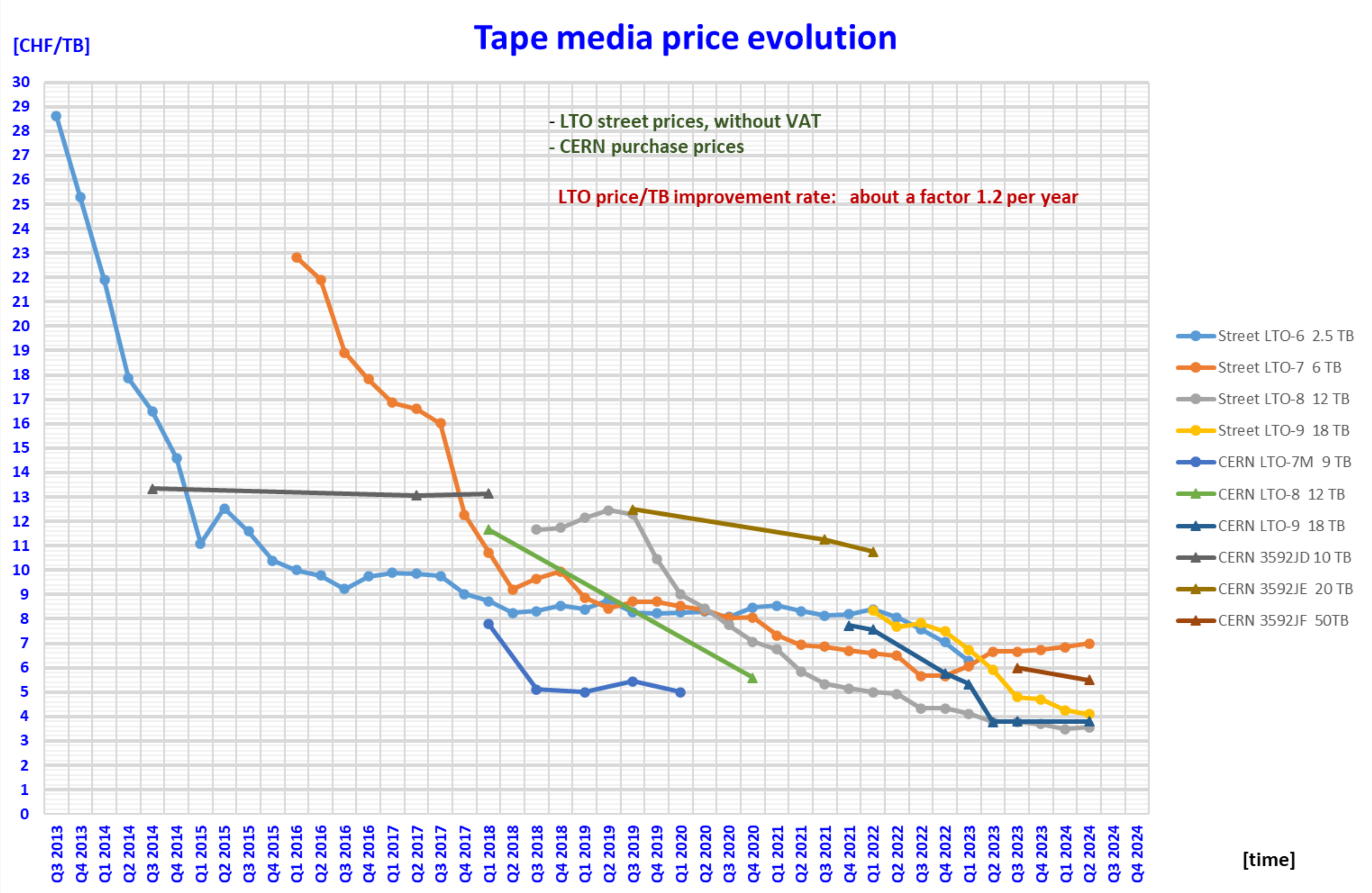
Would need in total for Run 4 about > 6EB
→ 3-4 more libraries needed

Consequence: need more physical space for tape libraries in the basement of 513

Still continue with two different types of tape media: some difference in technology, factor 2 in cost)
→ Price competition, problem mitigation e.g. ‘bad’ tapes, Fujifilm - Sony patent struggle (- LTO-8 shortage), etc.
Risk assessment versus cost difference

New technology generation every 2-3 years, now alternating between Enterprise and LTO tapes
Future strategy of IBM for Enterprise Tapes and LTO tapes still not 100% clear

Tape Storage Costs



Media cost today is about 4-5 CHF/TB.
Full tape storage cost is about 10-12 CHF/TB
(including infrastructure: libraries, server, tape drives, disk cache,..)
→ site and IO dependencies
Price increase this year will be 15% for LTO-9 and 40% for LTO-8

<u>CERN T0 investments needed for Run 4</u>

- 4 extra libraries
- 2.2 EB new tape media
- 350 new tape drives

(assume a factor 5 higher data rate, factor 2 increased tape drive performance, plus non-LHC , plus repack)

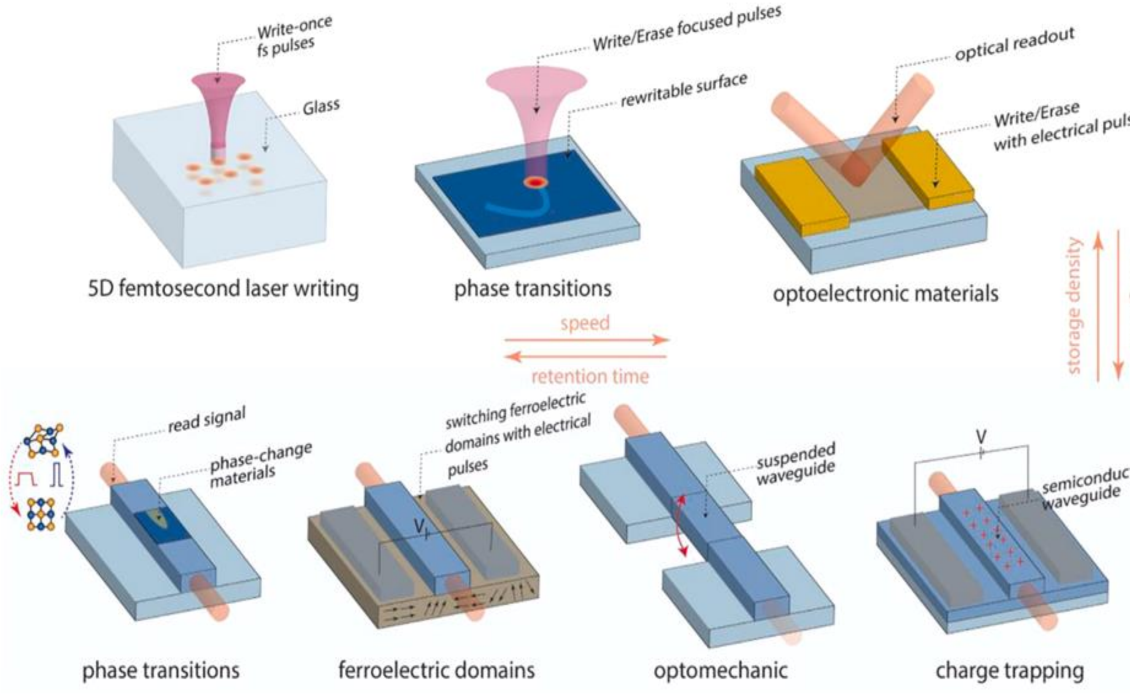
- 5-10 PB CTA instance

Realistic T0 experiment data rates need to be clear by 2027

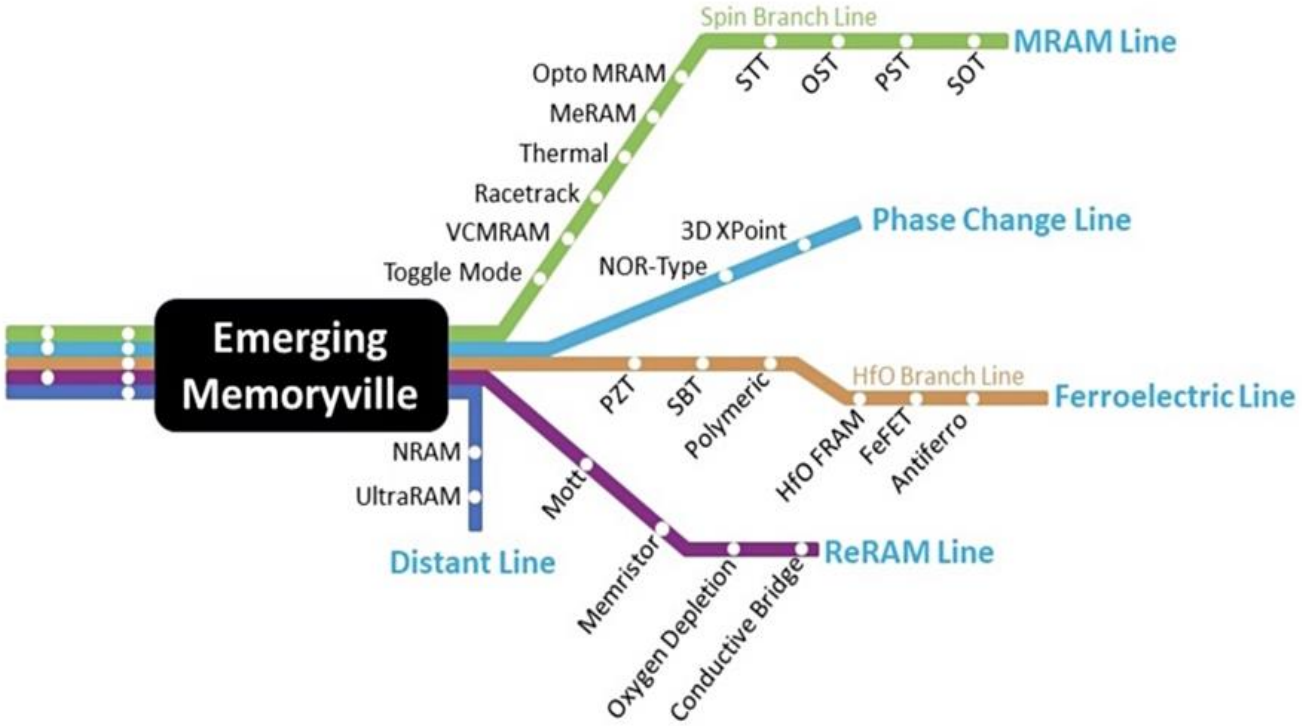
Disk storage is about a factor 3 more expensive than tape storage
→ IO performance dependent

New Storage Technologies I

Try to combine DRAM and NAND:
Non-volatile, low latency, high I/O, cheap, re-use of existing semiconductor fabrication technologies, long term durability, ………..



Taxonomy of DRAM/NAND replacement possibilities



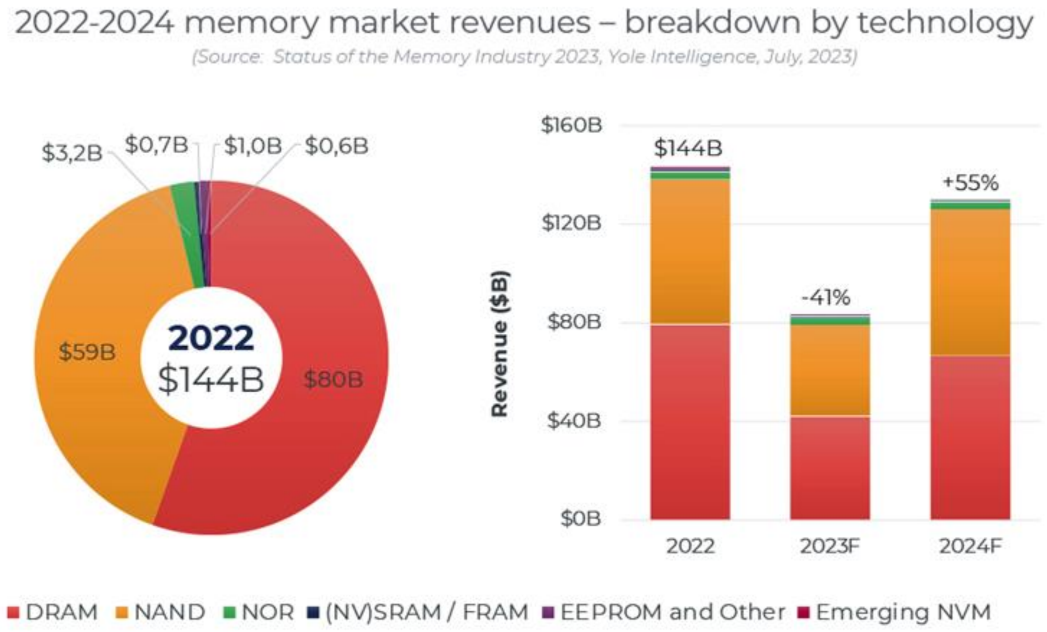
New possible storage technologies in the science news once per month

Brilliant science and technology achievements, but….

New Storage Technologies II

Everspin only MRAM supplier in the market
Revenues ~60 M$/y   compared to 144 B$ DRAM/NAND

Example: 3D Xpoint from Intel failed to make it into the market, Intel lost > 2B$

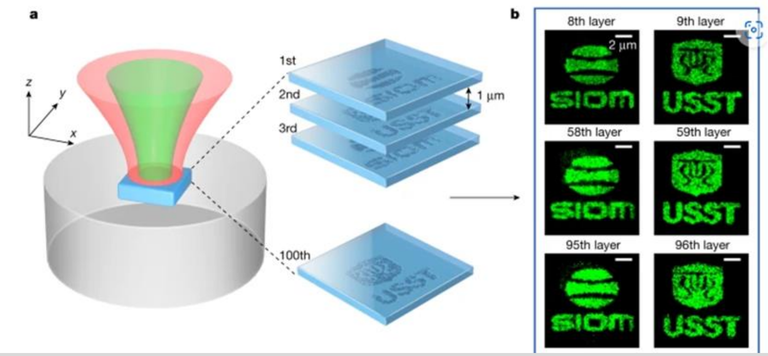


Breakthrough innovations in the storage area, black swan technology events very unlikely
Too much market entrenchments, few companies dominating, disruption of established markets requires multi-billion up-front investments and competition within a > 200B$ overall storage market

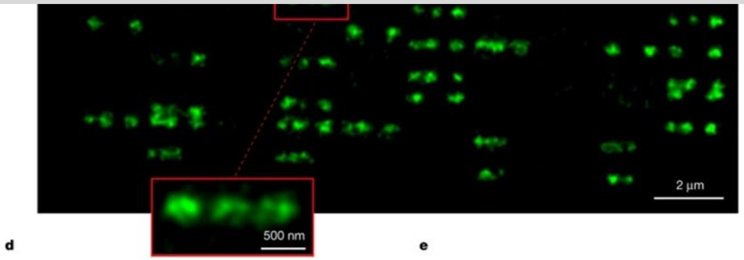
HEP computing requires 'cheap' components
→ Only mass market, not niche market

HDD/Tape replacements

Fig. 2: Demonstration of 100-layer volumetric nanoscale ODS and digital pattern encoding and decoding.



Chinese research paper in Nature, 800 Tb in 100 layer disk



DNA storage; Cerabyte
Project Silica; Folio Photonics
………..
Very unlikely to succeed and being relevant for HEP

Cost related assumptions and Uncertainties

1. Assume 2029 is a full Run 4 year → could be only 30% of a full year (cost)
2. Taken the experiment numbers based on the very little improvements scheme → full improvements factor 2 less resources
3. The trigger rates of the experiments will certainly increase more than planned for Run 4 (factor 2 ?)
4. Assumed 20% price/performance improvements in the next years → 10% leads to a difference of a factor 1.8 in 2029
   need to consider the flat budget notion
5. Assumed a WLCG T0 share for the resources as in Run 3, could change due to financial pressure from the member states
6. Electricity prices and low energy efficiency improvements might lead to budget constraints

→ The uncertainty in all the presented calculation and plans is at least a factor 2

The start of HL-LHC in 2029 will require a large investment in computing equipment for the CERN T0 O(50 MCHF)
→ Saving starts now, no major purchases during the next years
→ Extent the lifetime of existing equipment to >= 7 years (last year Microsoft increased their server lifetime from 4 to 6 years) This requires a very good understanding and monitoring of our equipment failure rates

Summary

- No technology obstacles for Run 4
- Don't expect major technology changes, still a careful watch is required. Possible adjustments in terms of the operation and architecture should be prepared early
- Need to get a grip on the GPU TCO by 2027
- Storage IO requirements need to be closely investigated
- CERN specific: save money over the next 5 years O(50M), thus extend equipment lifetime
  Need to extend the PCC from 4 to 8 MW and refurbish building 513 to host more tape libraries
- Probably lots of market instabilities during the next years (economical and political volatility), thus cost predictions will have a large error bar and maybe go in the wrong direction
- Much closer collaboration needed with the experiments during their code and data management improvements → minimize and understand hardware-software dependencies