Challenges of LHC Computing

Helge Meinhard / CERN-IT
RAPID 2018, Dortmund (Germany)
21-Nov-2018

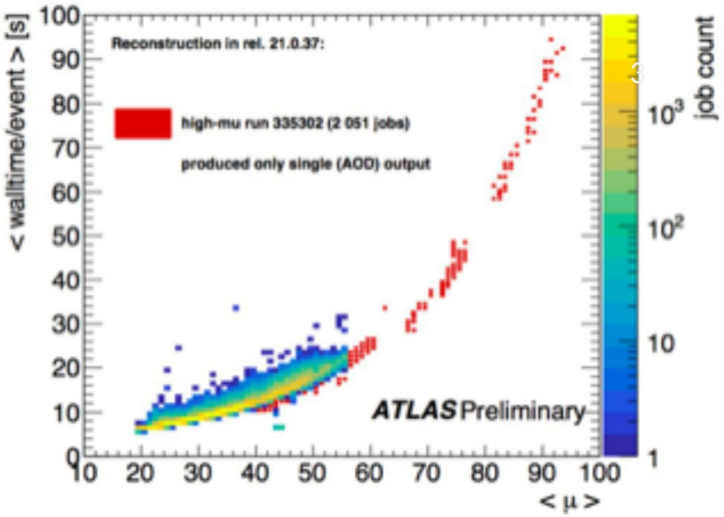
Contributions courtesy by
Simone Campana / CERN-IT
Bernd Panzer-Steindel / CERN-IT

Outline

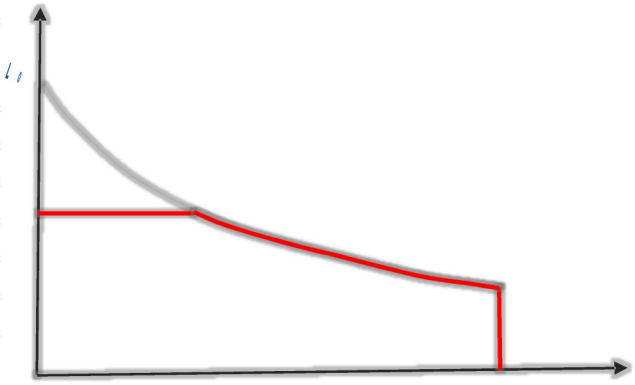
- Computing needs for Run 3 and an outlook to Run 4
- Technology and markets
- Conclusions

Run 3 for ATLAS and CMS

- Similar conditions with respect to Run 2 – with caveats
- Expect luminosity of up to 80 fb^{-1} per year (2018 was 66 fb^{-1})
- More virtual luminosity means longer leveling = more events at high pileup
- Assume 50% more computing needed

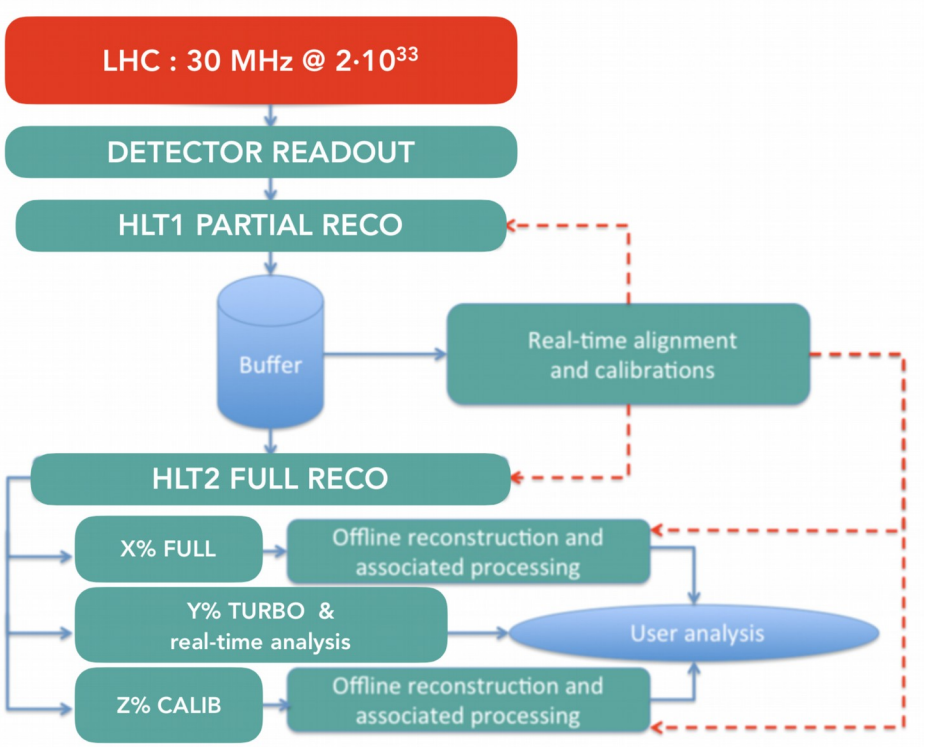


Parameter	BCMS	BCMS pushed a bit	Nominal - pushed	Comments
Energy [TeV]	7.0	7.0	7.0	
β* (1/2/5/8) [m]	0.3 / 10 / 0.3 / 3	0.3/ 10 / 0.3 / 3	0.3/ 10 / 0.3 / 3	Plus beta* levelling to 25 cm
Long-range separation [sigma] - assumed emittance	9.2 sigma - 2.5 um	9.2 sigma - 2.5 um	9.2 sigma - 2.5 um	
Initial Half X-angle (1/2/5/8) [μrad]	-160 / 120 / 160 / -150	-160 / 120 / 160 / -150	-205 / 120 / 205 / -150	Anti-levelled to 130 urad
Number of colliding bunches (1/5)	2592	2592	2748	BCMS - 240 bunches/injection from SPS
Bunch population	1.3e11	1.4e11	1.7e11*	* ruled out, initialy at least, by e-cloud heat load
Emittance into Stable Beams [μm]	2.5	2.6	3.0	
Bunch length [ns] - 4 sigma	1.1	1.1	1.1	
Virtual Luminosity (L0)	2.3e34	2.6e34	3.2e34	
Levelling time (hours)	2.0	3.8	7.9	
Luminosity per 12 hour fill (burn only)	0.65	0.7	0.8	
Luminosity lifetime (tauL) - end levelling	13 hours	14 hours	15 hours	Approx. - assuming burn only
Integrated/140 day year (fb-1)	65 - 70	70 - 75	85 - 90	NB Ballpark!



http://lhc-commissioning.web.cern.ch/lhc-commissioning/performance/Run-3-performance.htm

LHCb towards Run 3



Run 3 is a major upgrade of the detector and the computing infrastructure

- Level 0 hardware trigger to be replaced by software trigger with 30 MHz input rate
- Major re-engineering of software ongoing to cope with the increased load especially in the high level software trigger

Work towards HL-LHC will be far less demanding than Run 3 upgrade

ALICE Upgrade for Run 3 and 4

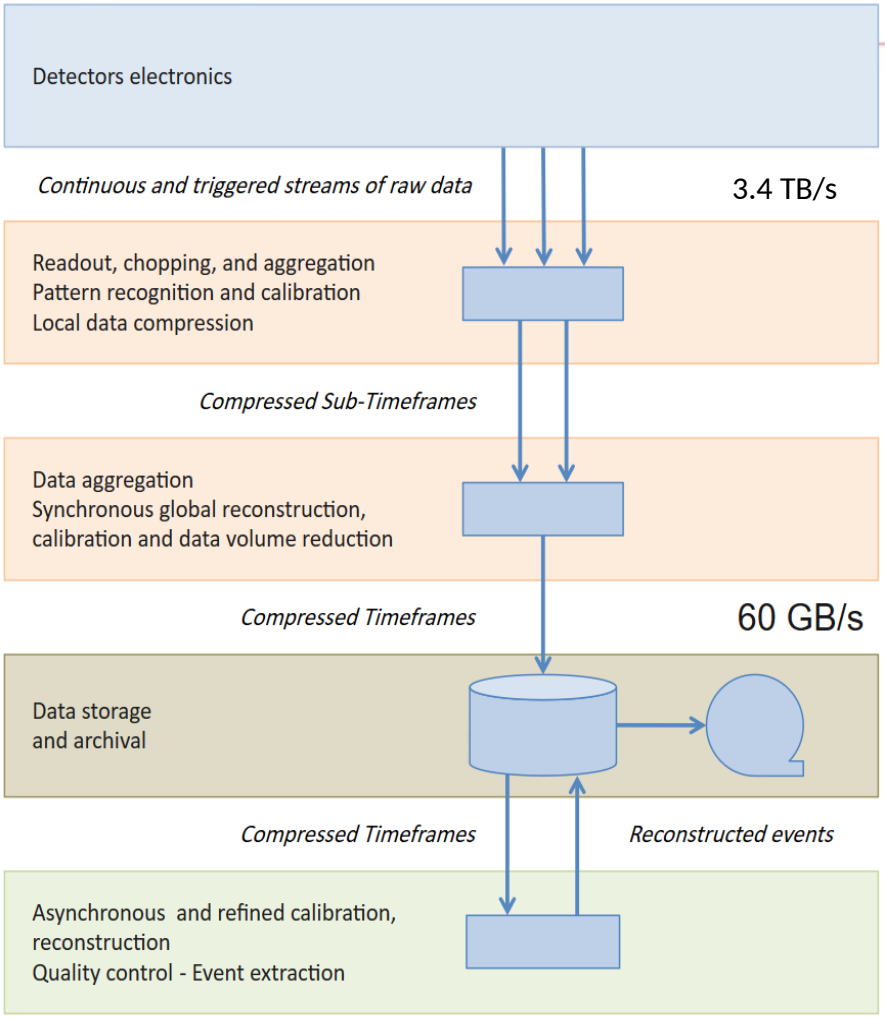
2 orders of magnitude more data from up to 500 million detector channels

Continuous (trigger-less) readout – first of its kind – with up to 3.4 TB/s from detector

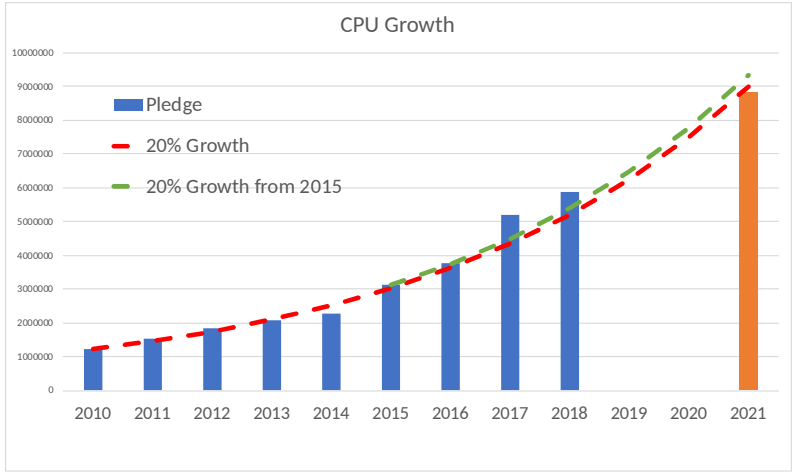
New O2 facility for data processing and compression – 1500 CPU/GPU nodes, 60 PB storage

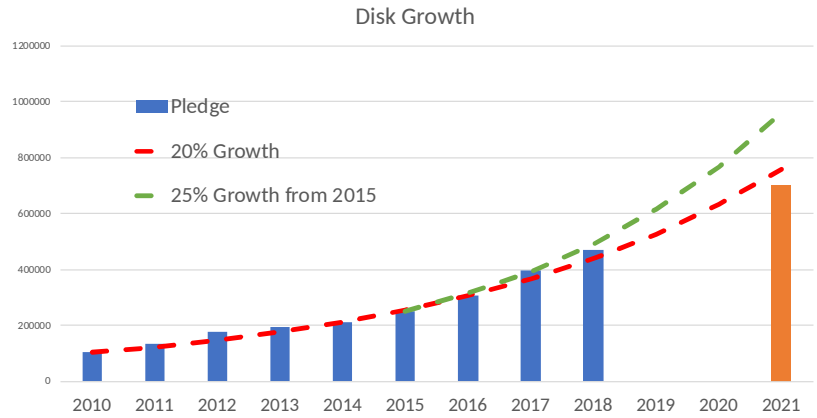
Steady progress on all O2 elements, including software framework (ALFA) and hardware components

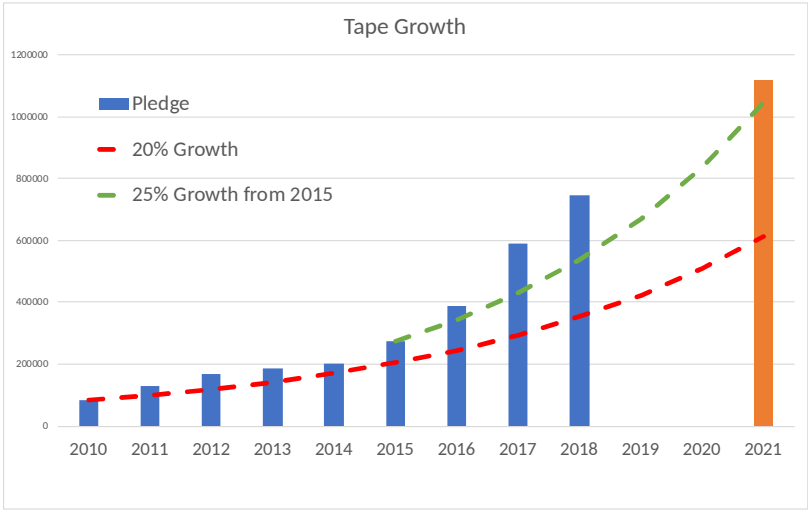
Resources growth corresponding to fixed funding (20% year on year) should be sufficient for Run 3



Run-3 Resource Evolution



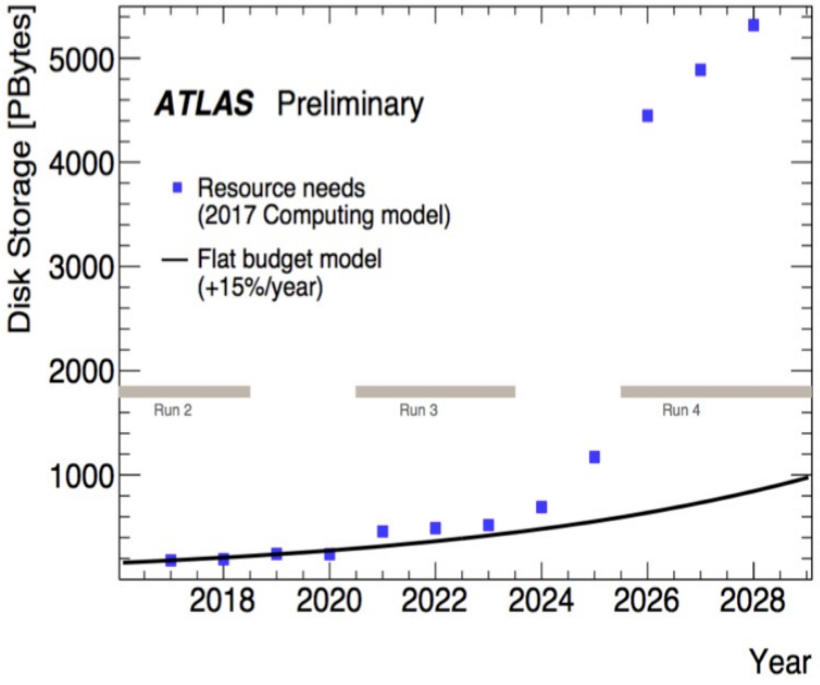


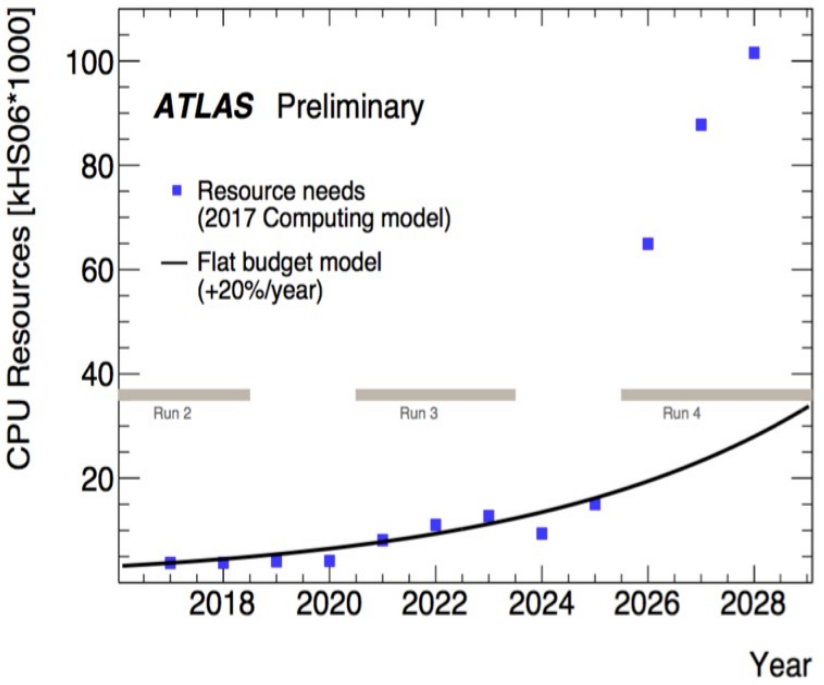


- 2010-2018 – pledges
- 2021 assume 1.5 x 2018

Overall, Run-3 resource needs look compatible with flat spending in the next years

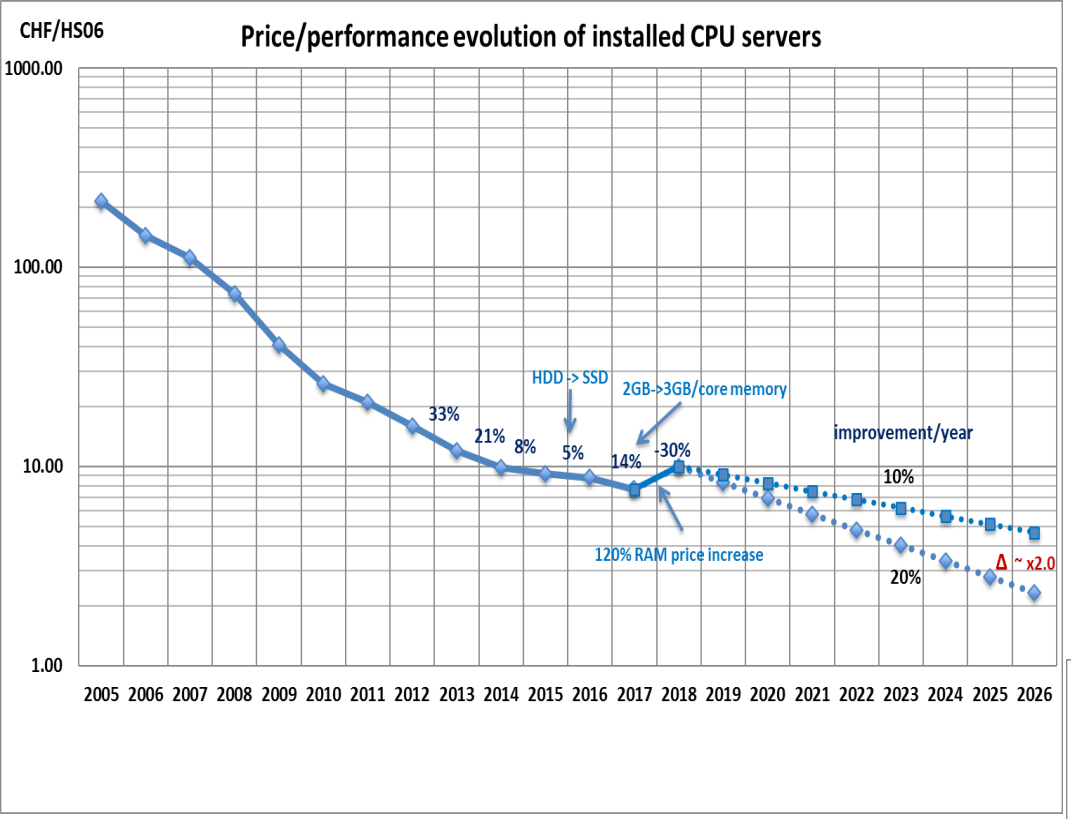
The HL-LHC Computing Challenge





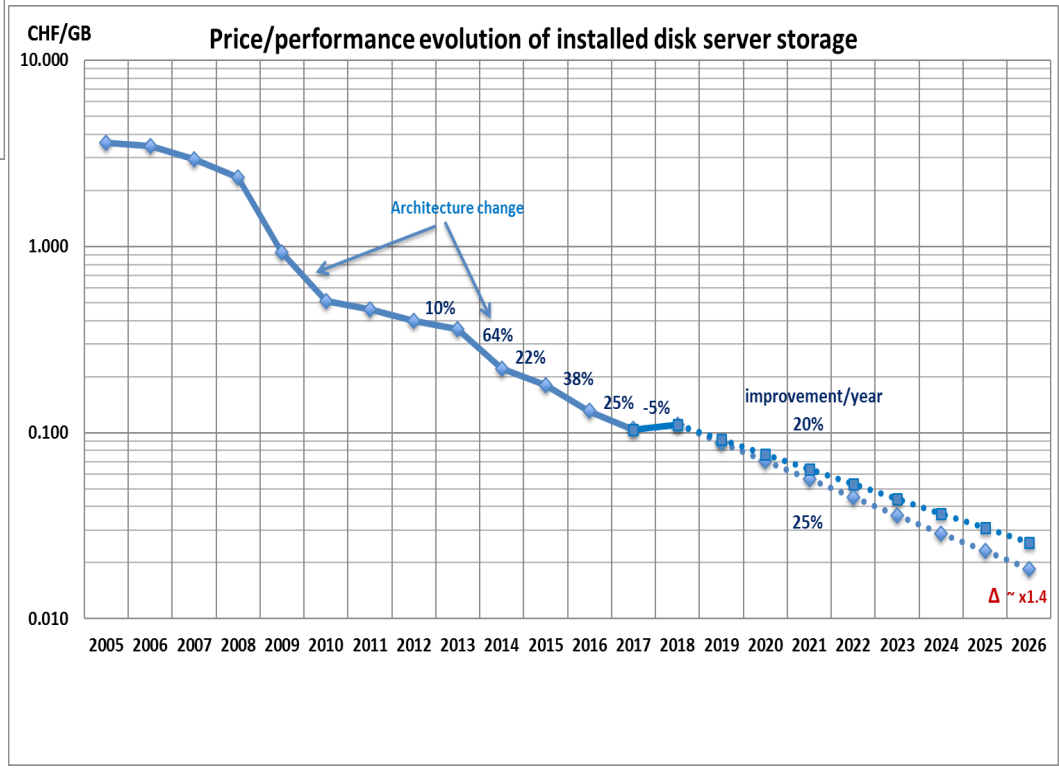
- HL-LHC needs for ATLAS and CMS are much above the expected hardware technology evolution (15% to 20%/yr) and funding (flat)
- The main challenge is storage – hence focus in the WLCG strategy

Server Cost Evolution



Disk servers: very hard to estimate real costs of HDDs

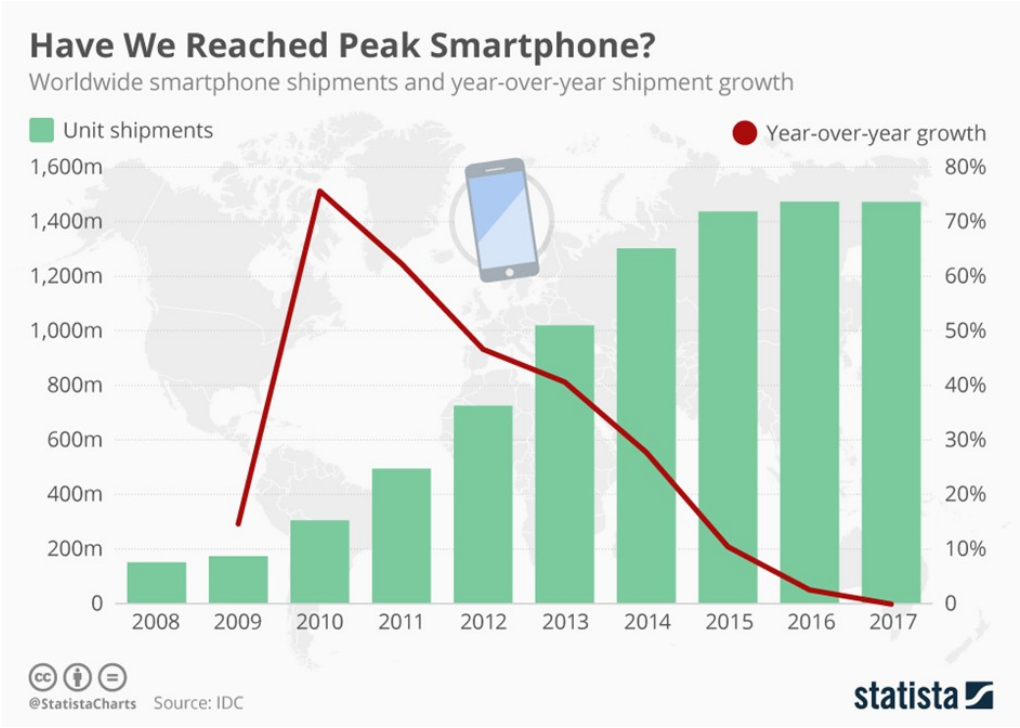
e.g. there are 70 different 6TB HDD models in the market with a price difference of a factor 2.5. At CERN we saw price differences of a factor >2 between low street prices and purchase prices; more variations between 6 TB and 8TB disks



Based on CERN procurements during the last years: current assumption:

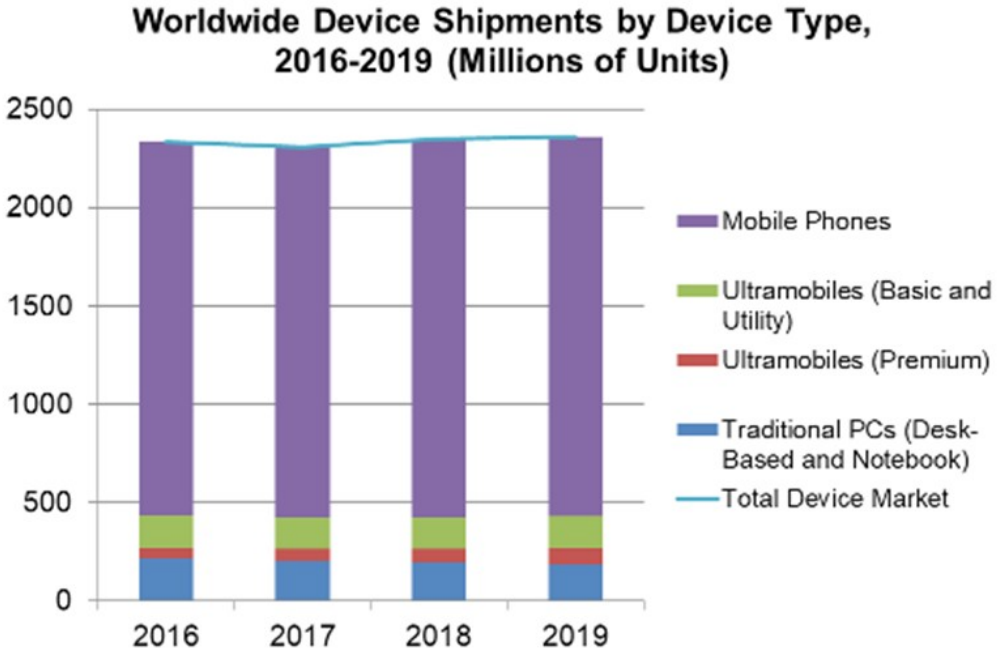
- Future CPU server price/performance improvement: 15%/year
- Future Disk server price/space improvement: 20%/year

Device Markets

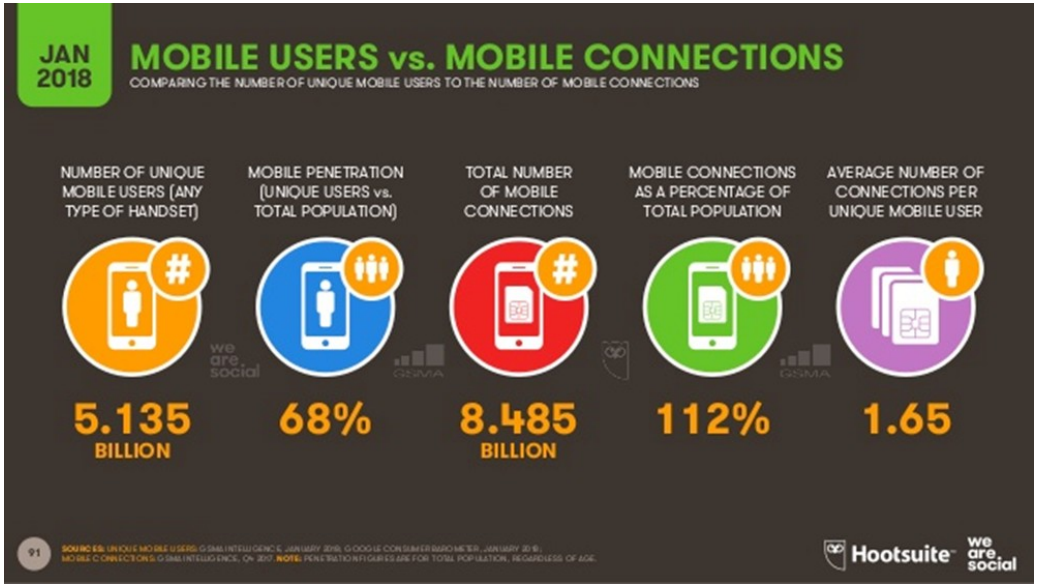


- PCs, notebooks and tablets sales declining constantly
- Smartphones sales are flat
- Attractiveness of replacement is decreasing
  - Only marginal differences between smartphone models and generations, small and little innovation
  - Consequence: Increased lifetime

- **Overall computing device market is flat**
  - Becomes replacement market



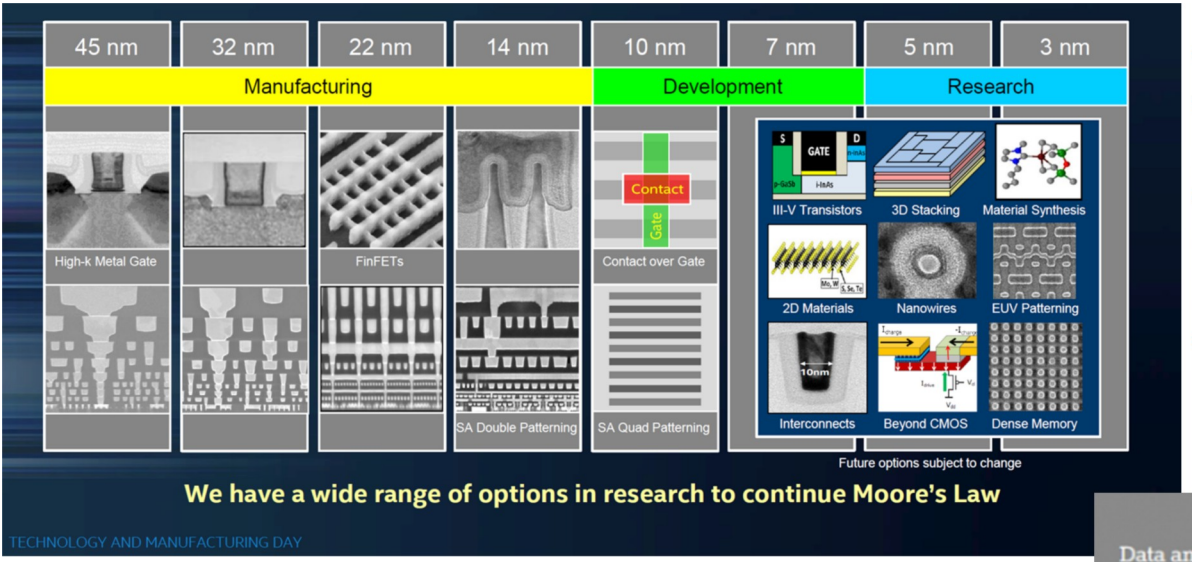
World “Internet” Population



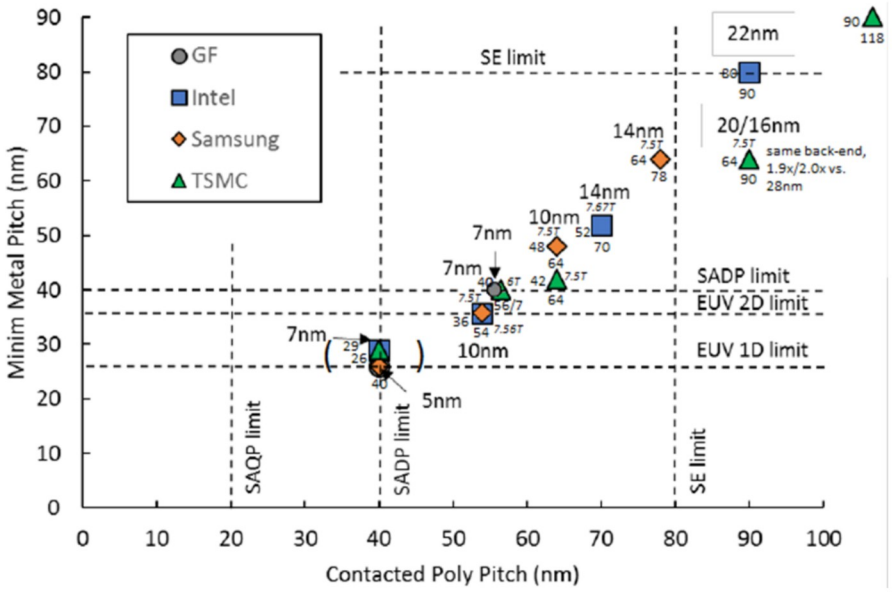
- Limited growth rates for internet devices due to limited growth rate of internet users
- Already high market penetration in the population
  - 68% of the world population have a phone
  - More mobile connections than humans

- **Strong saturation effects**

Processor Technology



- Intel has problems with their 10 nm process
- TSMC building fab 18 for their 5nm process, will be finished in 2020; 950'000 m^2 for $17B
- There is no norm for the process names: 10 nm Intel compares to a 7 nm Samsung/TSMC process
- Below 7 nm new technologies are needed (nanowires, non-silicon materials), very expensive



Industry FinFET Lithography Roadmap, HVM Start

Data announced by companies during conference calls, press briefings and in press releases

	2016	2017		2018		2019		2020		2021
		1H	2H	1H	2H	1H	2H	1H	2H	
GlobalFoundries	14LPP				7nm DUV	7nm with EUV*				
Intel	14 nm 14 nm+		14 nm++ 10 nm			10 nm+ 10 nm++				
Samsung	14LPP 14LPC	10LPE		10LPP		8LPP 10LPU	7LPP		6 nm* (?)	
SMIC	28 nm**	14 nm in development								
TSMC	CLN16FF+ CLN16FFC	CLN10FF CLN16FFC		CLN7FF CLN12FFC		CLN12FFC/ CLN12ULP	CLN7FF+		5 nm* (?)	
UMC	28 nm**	14nm		no data						

*Exact timing not announced
**Planar

New Processor Architectures

- Plethora of new processor designs, all with a focus on Machine Learning:
  - Intel: Mobileye EyeQ5 (vision processing, autonomous cars), Nervana Neural Network Processor, Movidius MyriadX VPU
  - ARM: Project Trillium, Machine Learning processor, Object Detection processor
  - Graphcore IPU (Intelligent Processing Unit)
  - Google second generation of Tensor Processing Unit TPU
  - NeuPro AI processor from CEVA
  - Neuromorphic chips from IBM (TrueNorth, 64 M neurons + 16 B synapsis) and Intel (Loihi, 130 K neurons + 130 M synapsis)
  - Nvidia enhancing their graphics cards, Titan V (110 Tflops Deep Learning), Xavier (SoC, 20 TOPS, vision accelerator)

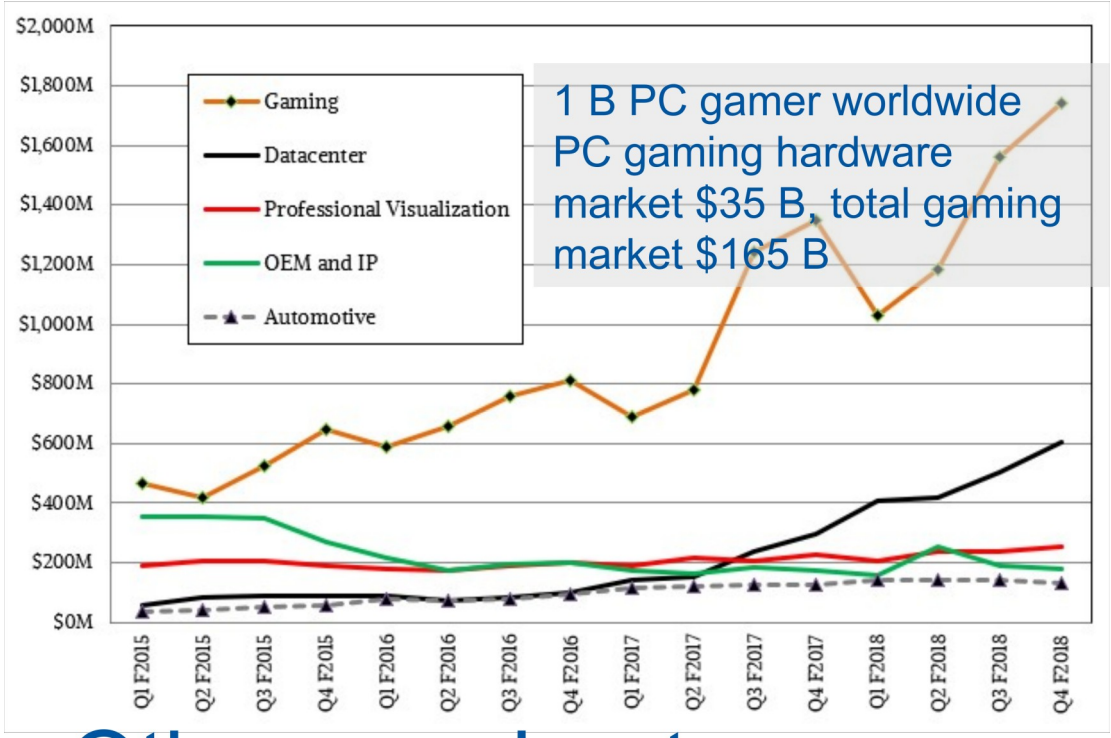
- All high-end smartphones are integrating AI chip enhancements (Qualcomm-neural processing engine, Apple- A11 Bionic chip, etc.)
  - The market for these special chips will reach $5-10 B in 2022

- The keyword is LOCAL data processing (also major impact on IoT)
  - Much less network, cloud storage and cloud processing needed

Accelerators

GPUs:

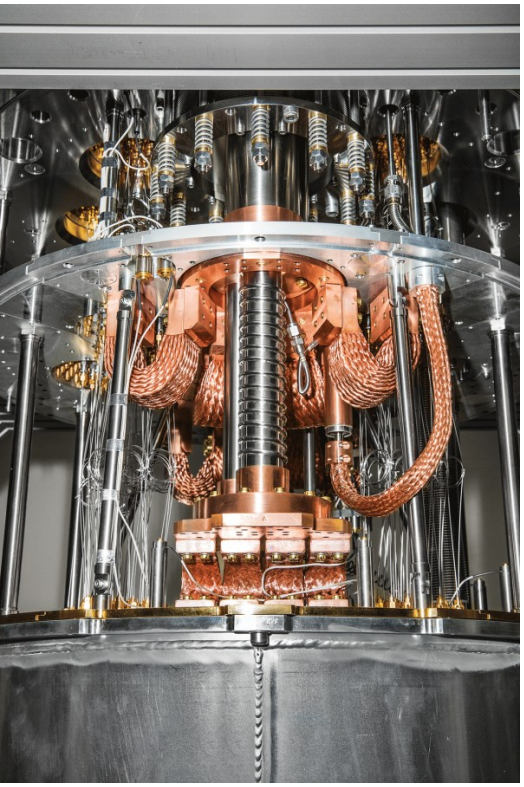
- Dedicated graphics cards market leader is Nvidia
- High end card Tesla V100 (14 TFlops SP, 110 TFlops ML, 12 nm process)
- Gaming key driver for the market (plus AI and crypto mining)
- Large price increases (up to x2), crypto mining + high memory prices
- Licence policy: no gamer cards in the data centre
  - GTX 1080 TI: $700,
    Tesla V100: $9'000, but also DP performance 20-30 higher



Other accelerators:

- Intel stopped the Xeon Phi line (Knights Mill last product)
  - No replacement in sight
- Microsoft Project brainwave, based on Intel (Altera) Stratix FPGA
- Xilinx ACAP, Project Everest many-core SoC, programmable DSPs
  - 50 B transistors, TSMC 7 nm process
- Chinese Matrix-2000 DSP accelerator for Exascale HPC
  - Current No 1 on Top 500

Quantum and Optical Computing



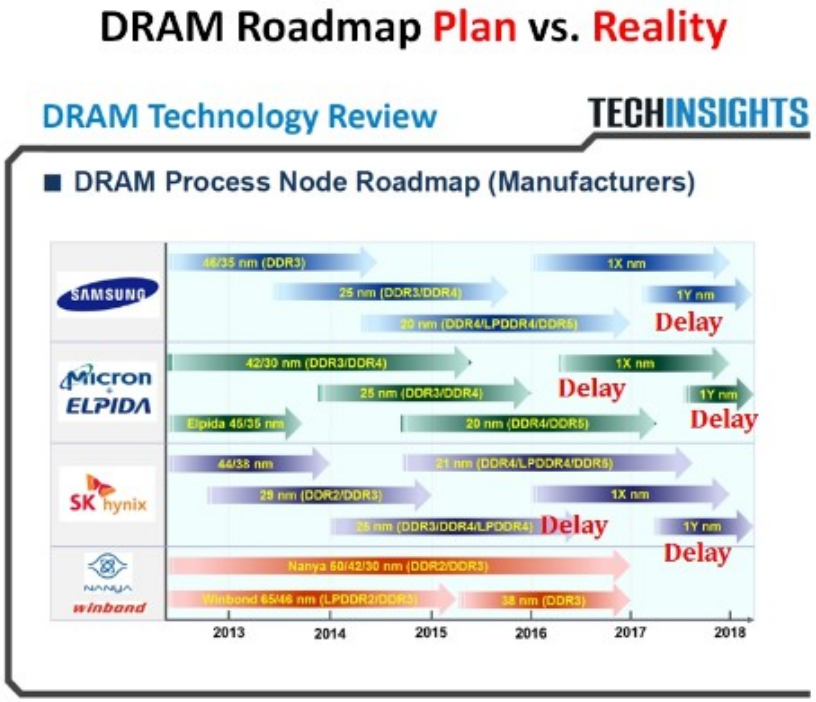
- Considerable progress during the last 2 years; number of qubits rising sharply
  - Intel 49-qubit, IBM 50, Google 72 for a quantum gate computer
  - D-wave 2000 qubits, but not a general quantum computer (e.g. no shor's algorithm, no factorization)
- Various implementations from ion traps to silicon, focus is on silicon to re-use the fabrication process of standard chips
  - Coherence time is still well below 1 ms, limits the time for quantum calculations
- Key problem is the error handling: mitigate by combining qubits
  - N physical qubits == one logical qubit, where N varies between 10 and 10'000
  - Use error correction in software, deal with approximate results
  - Machine learning algorithms
- Programming model is completely new; not clear how many algorithms can be 'converted' for a quantum computer; very, very high cost structure
- **Prognosis: Irrelevant for HL-LHC**

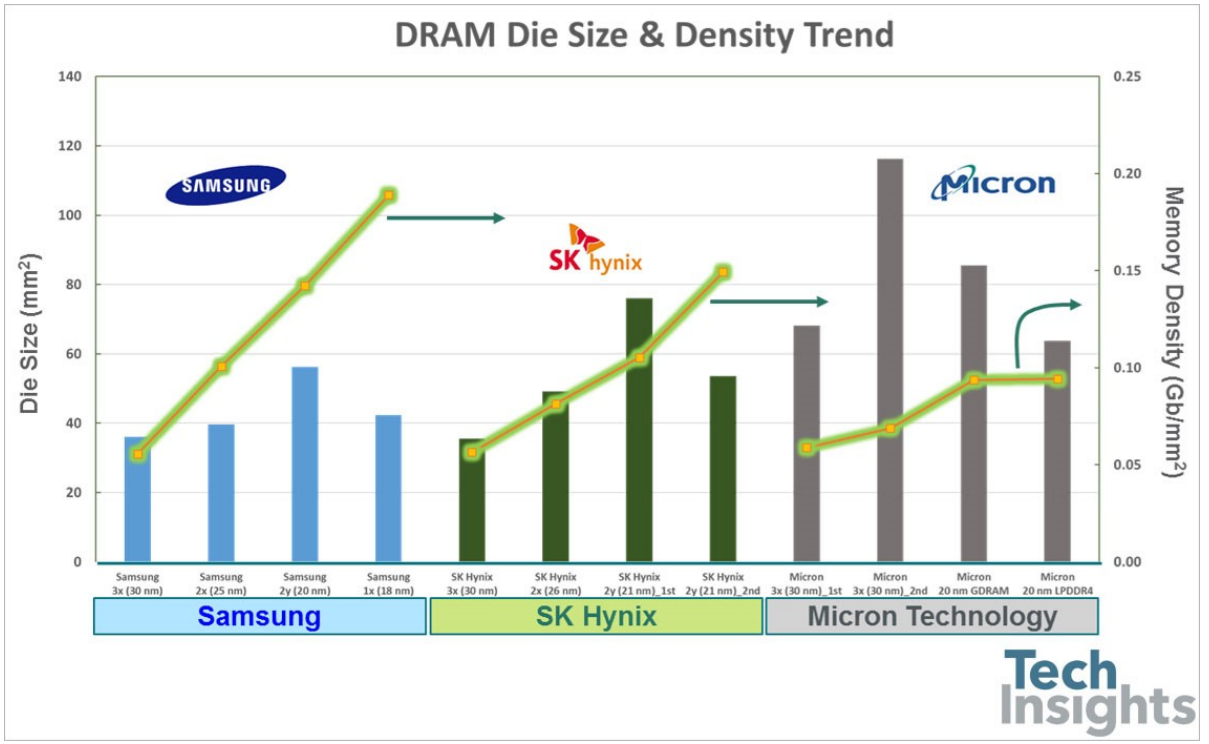
Renaissance of optical computing, this time focused on neural networks

- Optalysis: First implementation of a Convolutional Neural Network with Optical Processing Technology
- Lightelligence: Deep learning with coherent nanophotonics circuits
- Lightmatter: Photonics for AI

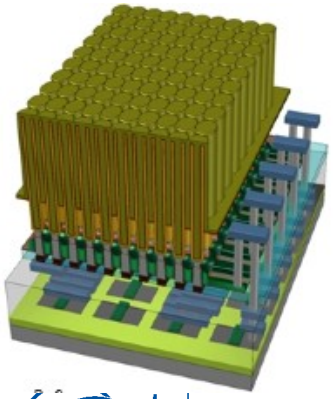
Memory: DRAM

- ~ $70 B market
- DRAM price increase during late 2016 to early 2018: ~120%





- DRAM scaling slowed down
- Capacitor aspect ratio increases exponentially with smaller cell size
- Much higher fabrication costs
- 3D DRAM not yet available



- Samsung 18nm DRAM process (36-54nm pitch), considerable density improvement
- Linear rather than exponential

New Memory Technologies

Technology Comparison

SNIA PERSISTENT MEMORY PM SUMMIT
JANUARY 24, 2018 | SAN JOSE, CA

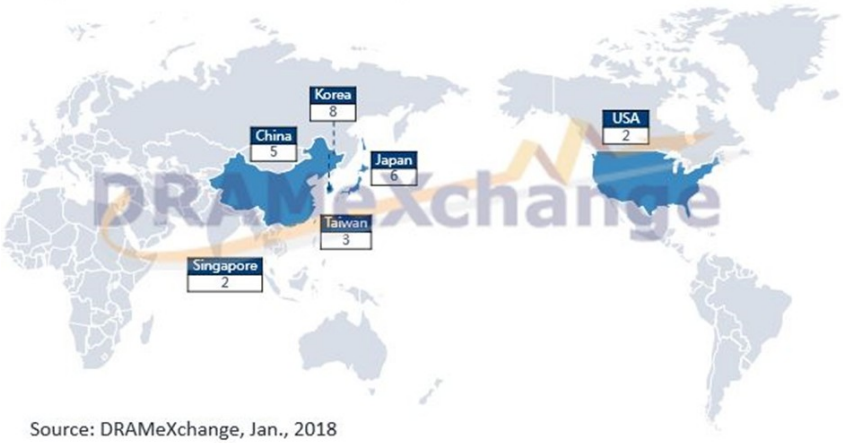
Technology	FeRAM	MRAM	ReRAM	PCM	DRAM	NAND Flash
Nonvolatile	Yes	Yes	Yes	Yes	No	Yes
Endurance	10^{12}	10^{12}	10^{6}	10^{8}	10^{15}	10^{3}
Write Time	100ns	~10ns	~50ns	~75ns	10ns	10µs
Read Time	70ns	10ns	10ns	20ns	10ns	25µs
Power Consumption	Low	Medium/Low	Low	Medium	Very High	Very High
Cell Size (f^2)	15-20	6-12	6-12	1-4	6-10	4
Cost ($/Gb)	$10/Gb	$30-70/Gb	Currently High	$0.16/Gb	$0.6/Gb	$0.03/Gb

- Several contenders for a new memory technology
  - Ideally replacing DRAM and NAND at the same time
- No cost effective solution yet

- Resistive RAM, 40nm process, Fujitsu/Panasonic
  - Aimed at Neuromorphic Computing
- Magnetic RAM, 80nm process , Everspin, first 1-2GB SSDs
- PCM Intel Optane, in production, but focus not clear
- Ferroelectric RAM, very small scale products, difficult to scale

NAND Storage

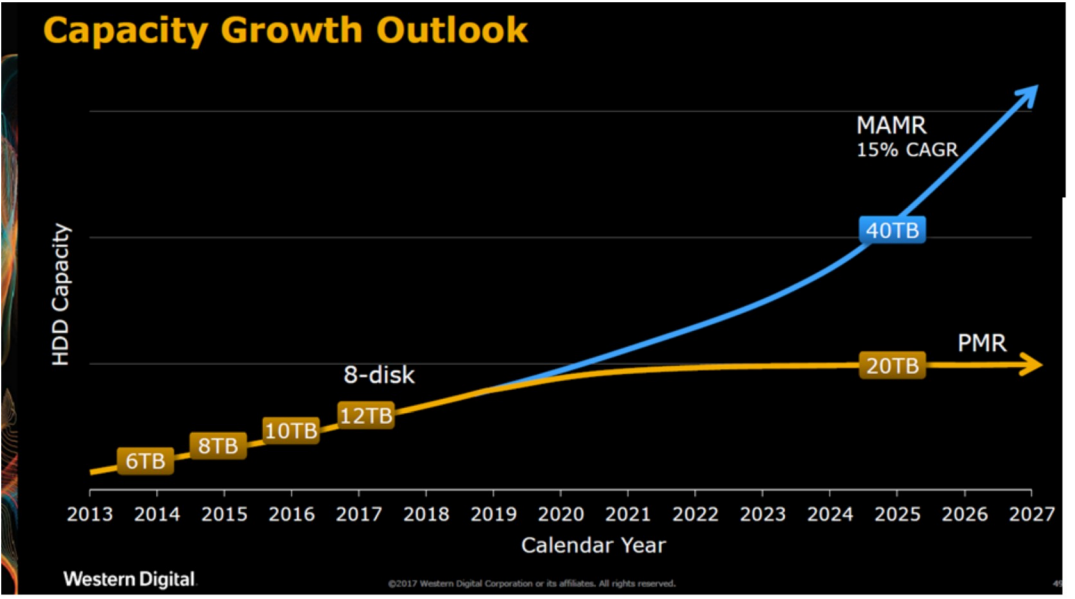
Figure: NAND Flash Factories Map in 2020

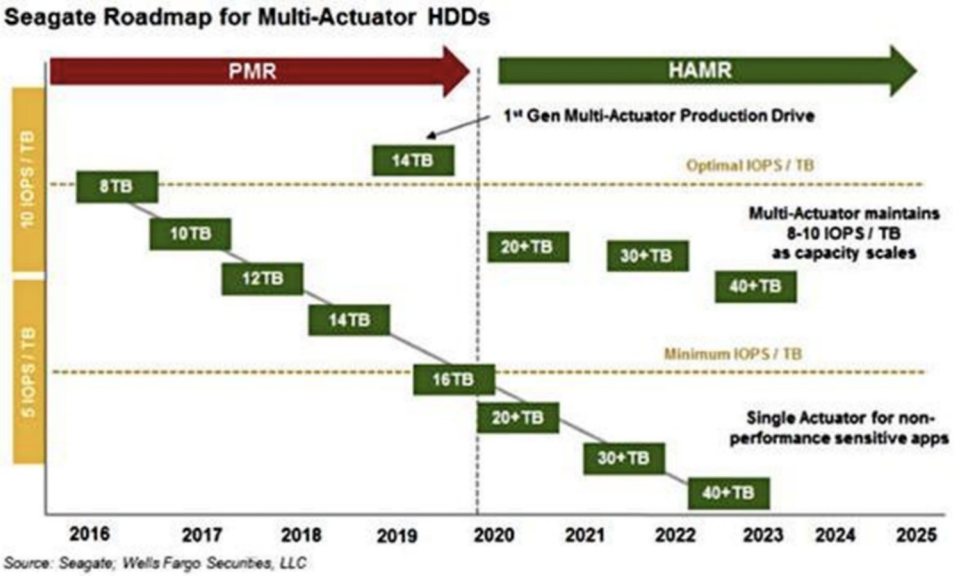


Source: DRAMeXchange, Jan., 2018

- ~ $60 B market
- Fabrication moved from 2D to 3D: 64 layers in the market, 96 layer production started, 128 layers expected for 2020
- NAND prices increased until early 2018, high request for smartphones and SSDs (Apple buys 20% of the world-wide NANDs), significant price decay since then
  - New Chinese fabs have started production
- In 2017 largest consumer of NAND chips were SSDs (surpassing smartphones)
- 4-bit cells are now feasible with 3D: ECC code easier; lab demos exist with hundreds of layers
- Investment in 3D fabrication process up to 5x higher than 2D: ~ $10 B for fabrication facility
- Technical challenges: > 64 layers show exponential scaling problems (current density, cell uniformity)
  - A wafer stays up to 3 month in the fab before the 100 defect-free layers are done
- Density improvements are now linear, adding 8/16/32 layers

Hard Disk Storage I



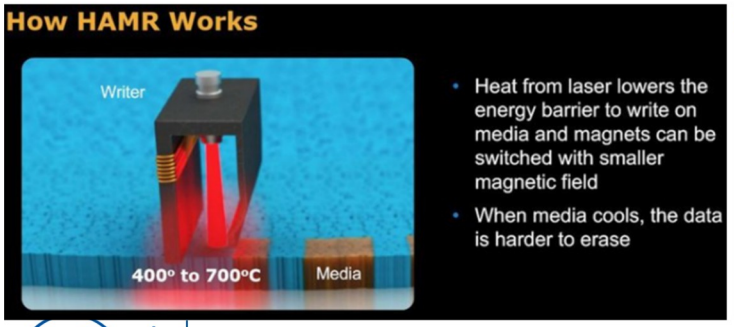


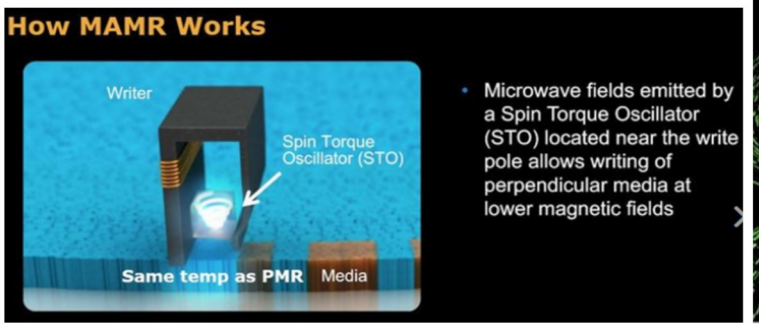
- 9 platter in one drive, 14 TB capacity today, Helium-filled
- Max with SMR is probably around 20 TB per drive

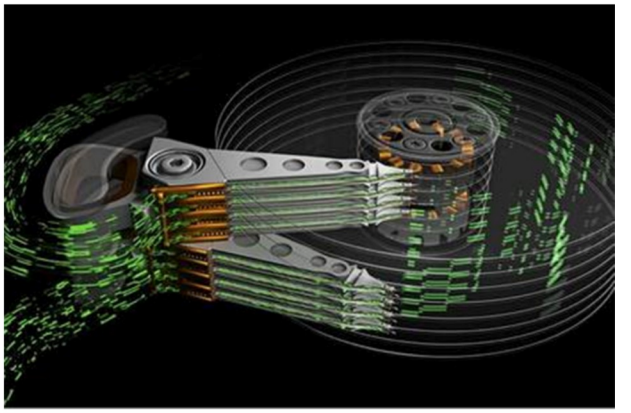
Seagate: multiple actuators per HDD to keep IOPS/TB constant

Seagate HAMR first products now in 2020

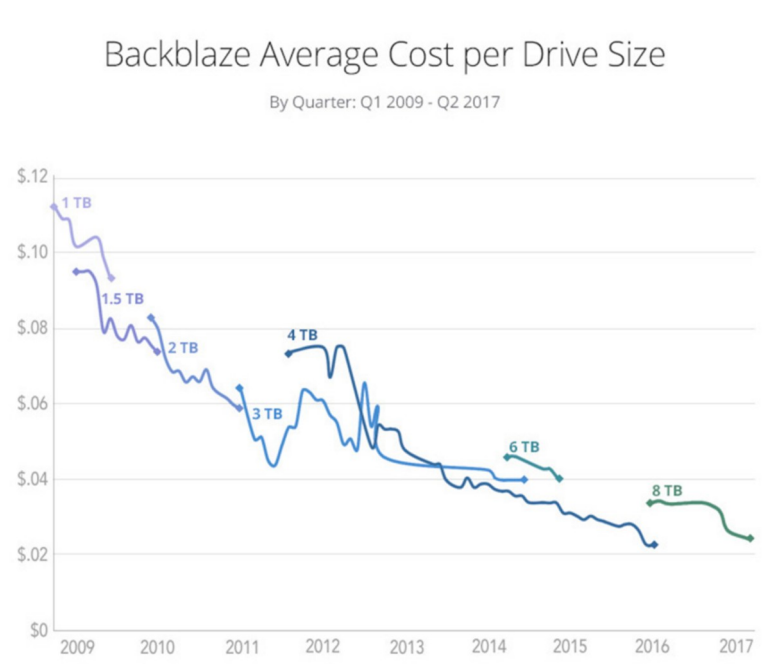
Western Digital new density approach: MAMR production in 2019

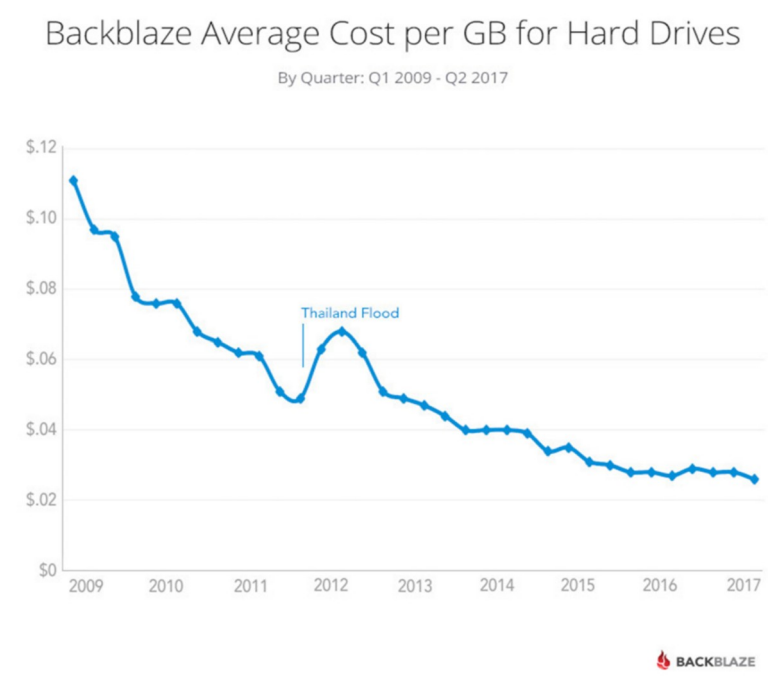


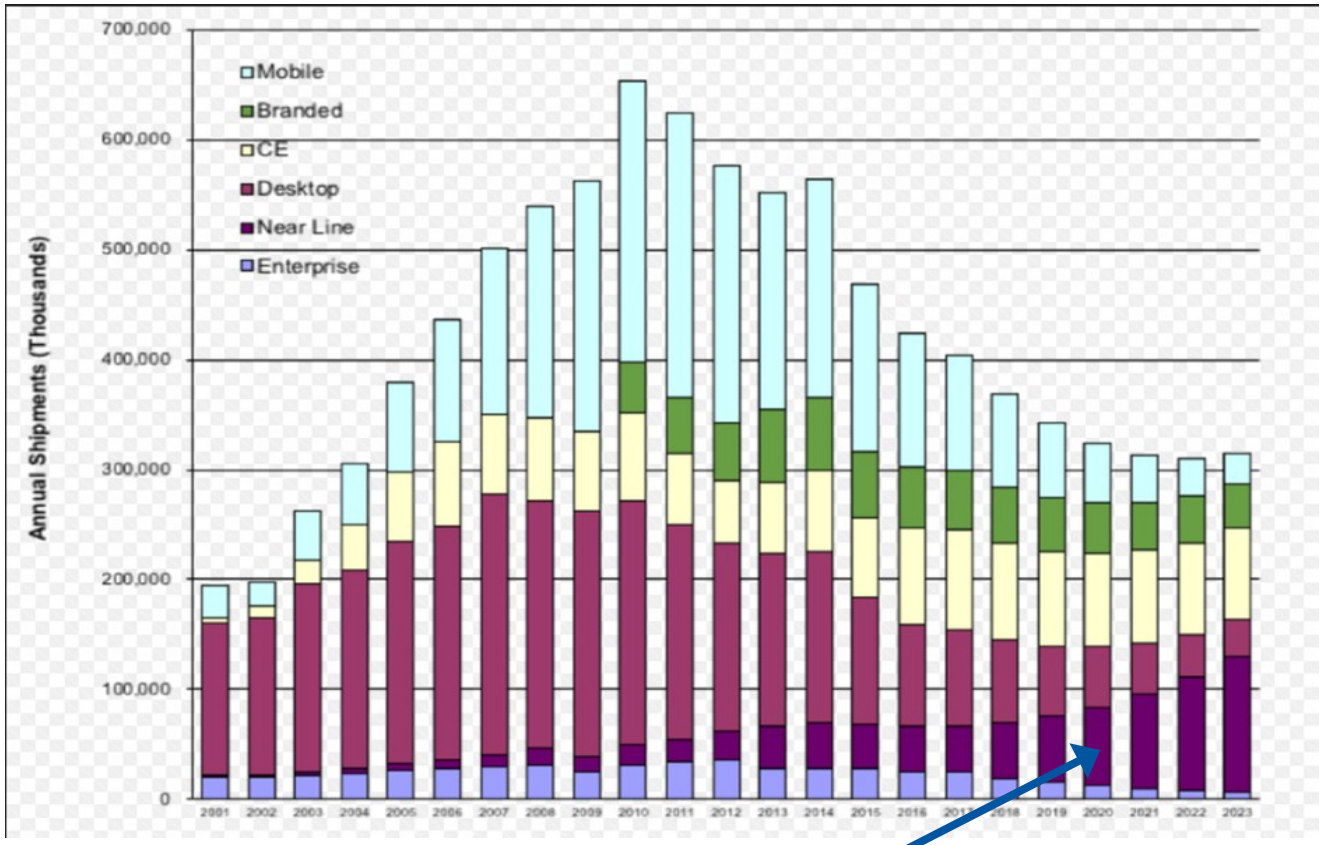




Hard Disk Storage II

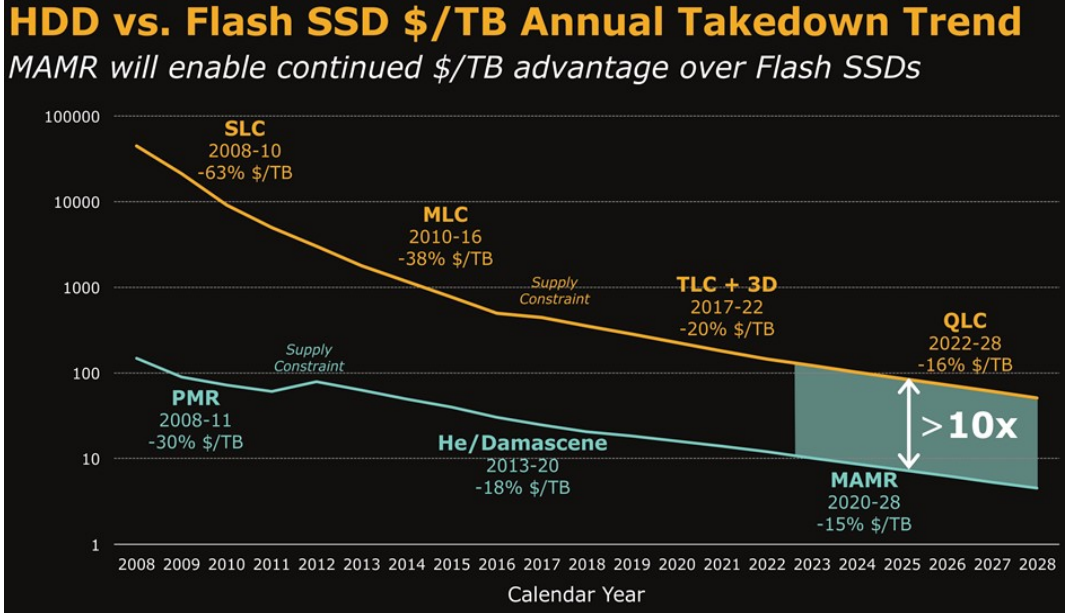


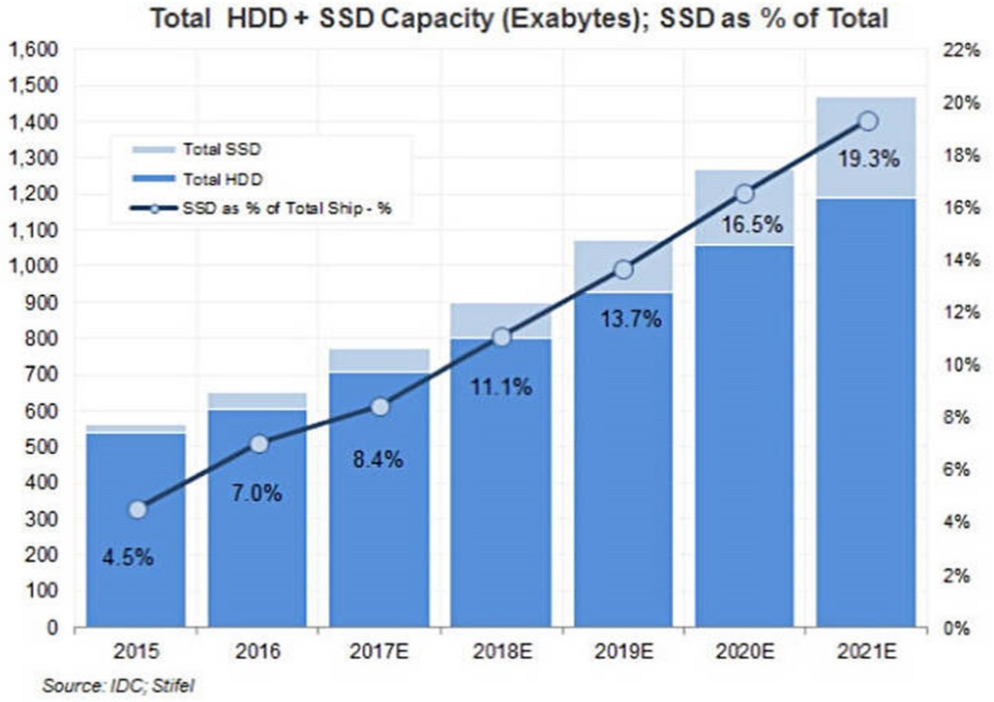


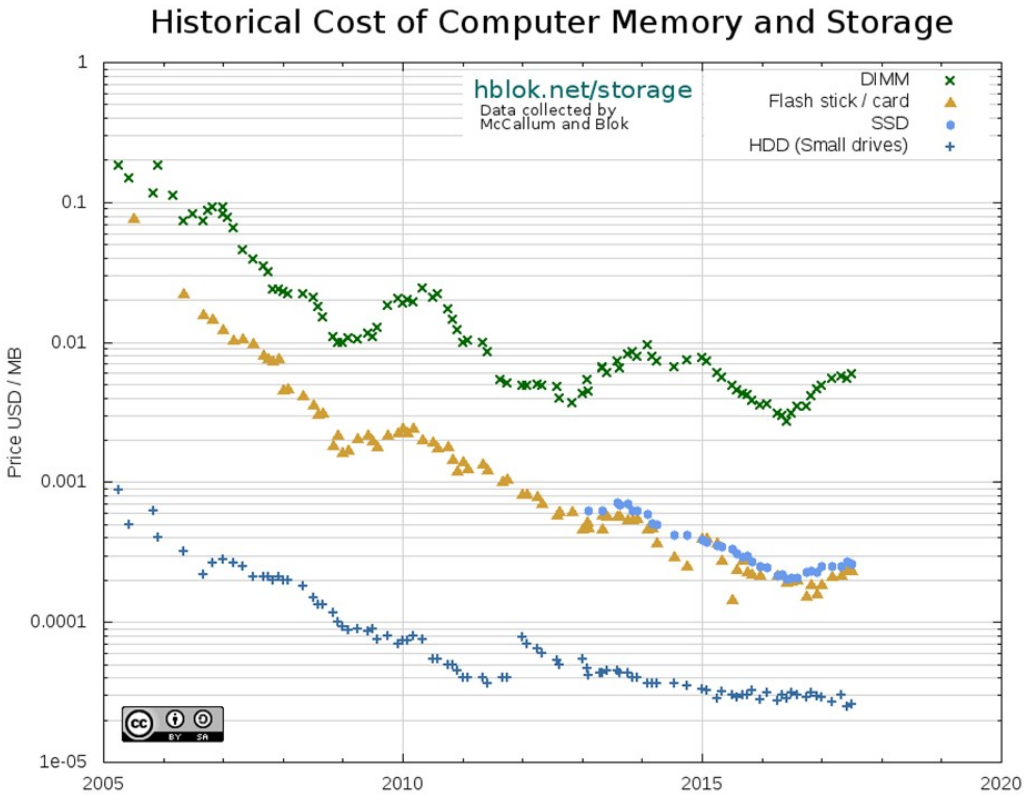


- Only growth: near-line disks (high capacity), HEP and cloud storage area
- Desktop, mobile, enterprise replaced by SSDs
- Price/space evolution flattening

Solid-state Disk Storage

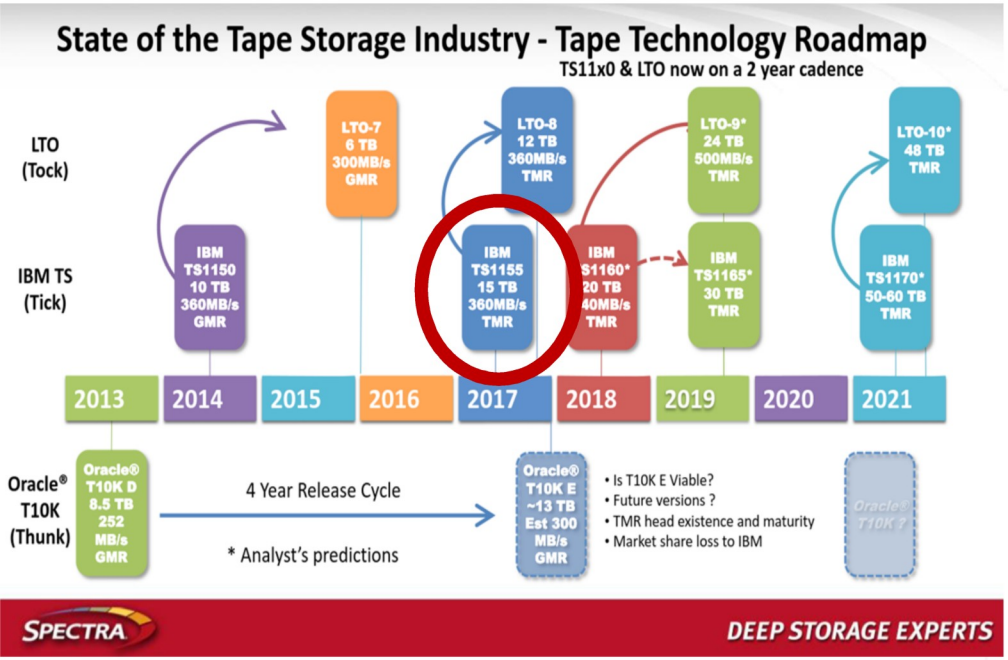






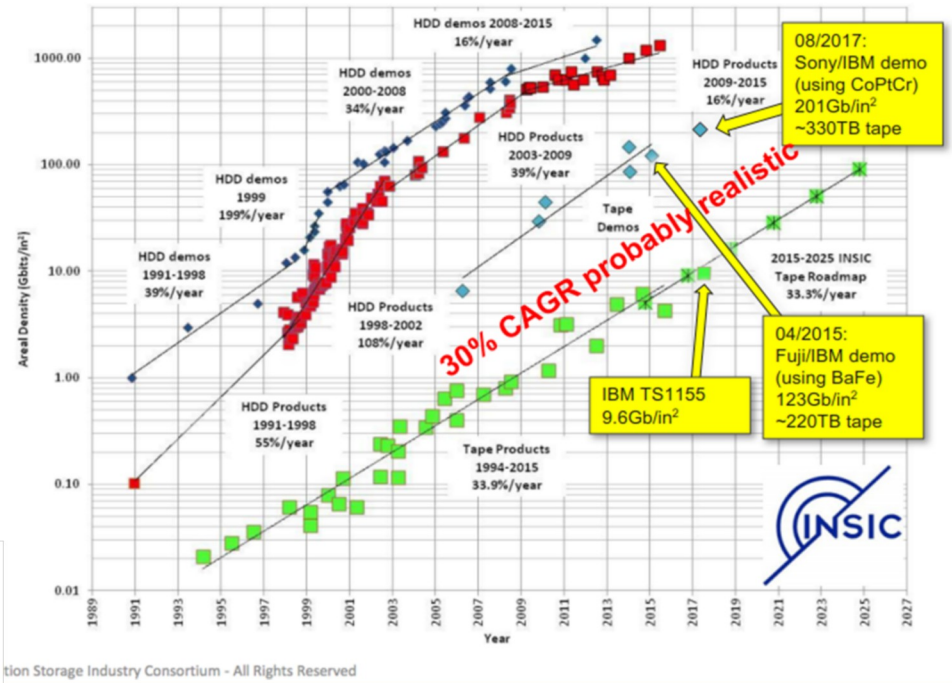
- SSD versus HDD: price difference in capacity drives will stay high for the foreseeable future
- Slowdown of yearly price improvements in all areas

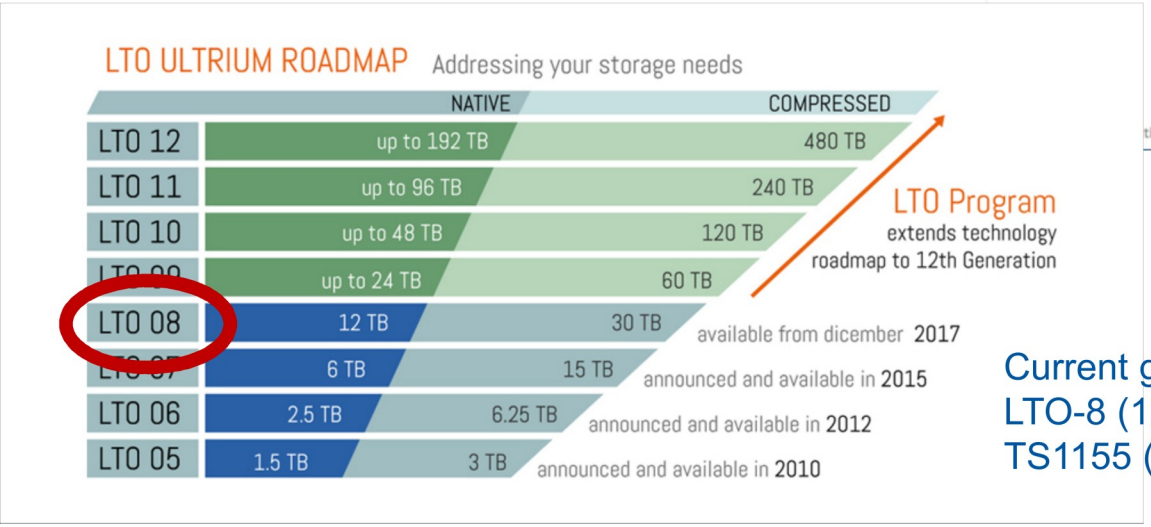
Tape Storage I



Areal Density Trends

Chart provided courtesy of the Information Storage Industry Consortium (INSIC)



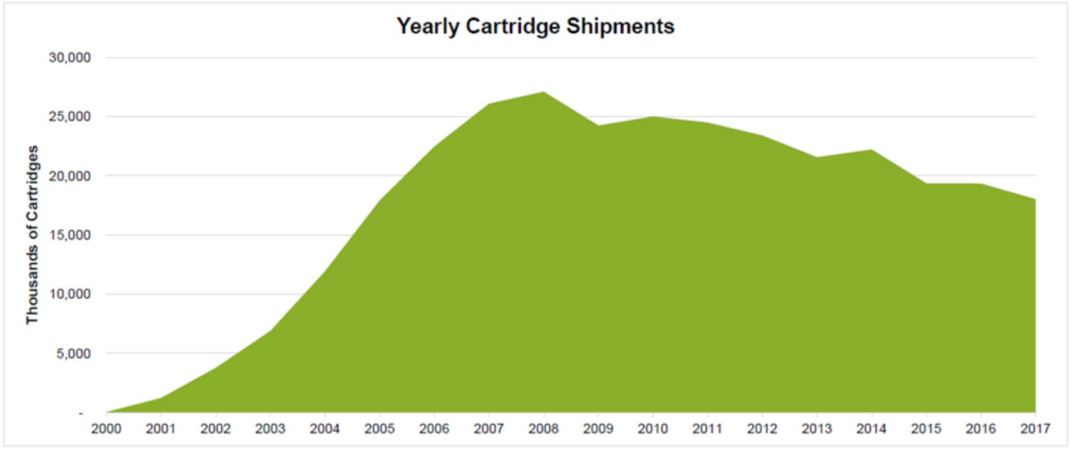


Current generation LTO-8 (12 TB) , TS1155 (15 TB)

- Technology change to Tunnel Magnetoresistive heads (used already in HDDs) for IBM TS1155 and LTO-8
- Quite some headroom for density improvements, x10 compared to HDD

Tape Storage II

Unit Shipments: Calendar Year

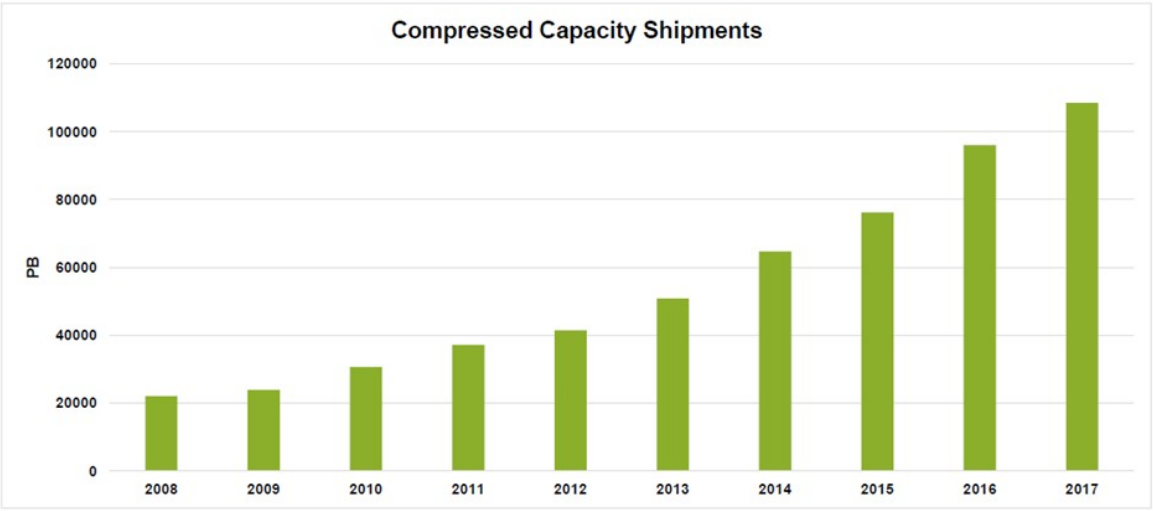


- LTO tape market domination >95%
- Enterprise tapes 4%

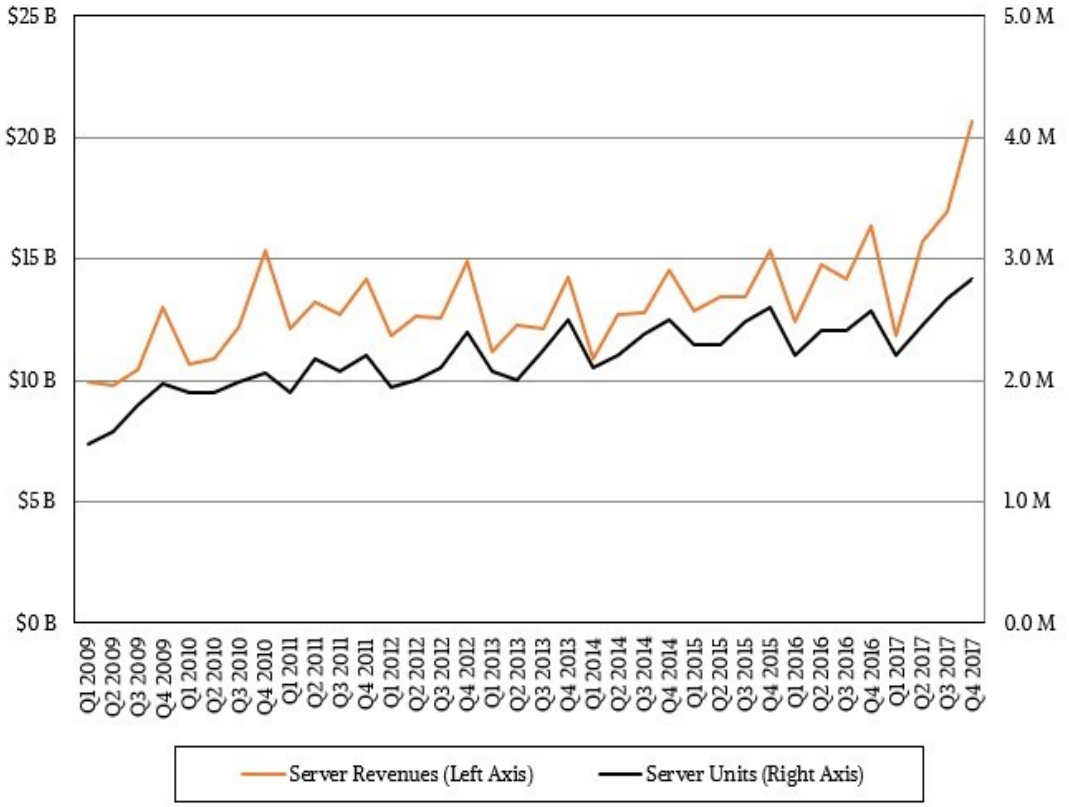
- 44 EB of tape media in 2017 compared to 750 EB HDD
- Linear increase in EB sold per year

- Declining media shipment since 10 years
- Factor 2 decrease in #drives sold over the last 4 years
- Only two suppliers of media: Fujifilm and Sony
  - Fujifilm only supplier in the US (patent 'war')
- Only IBM left for LTO and Enterprise drives

Total Capacity Shipped: Calendar Year



Server Market I



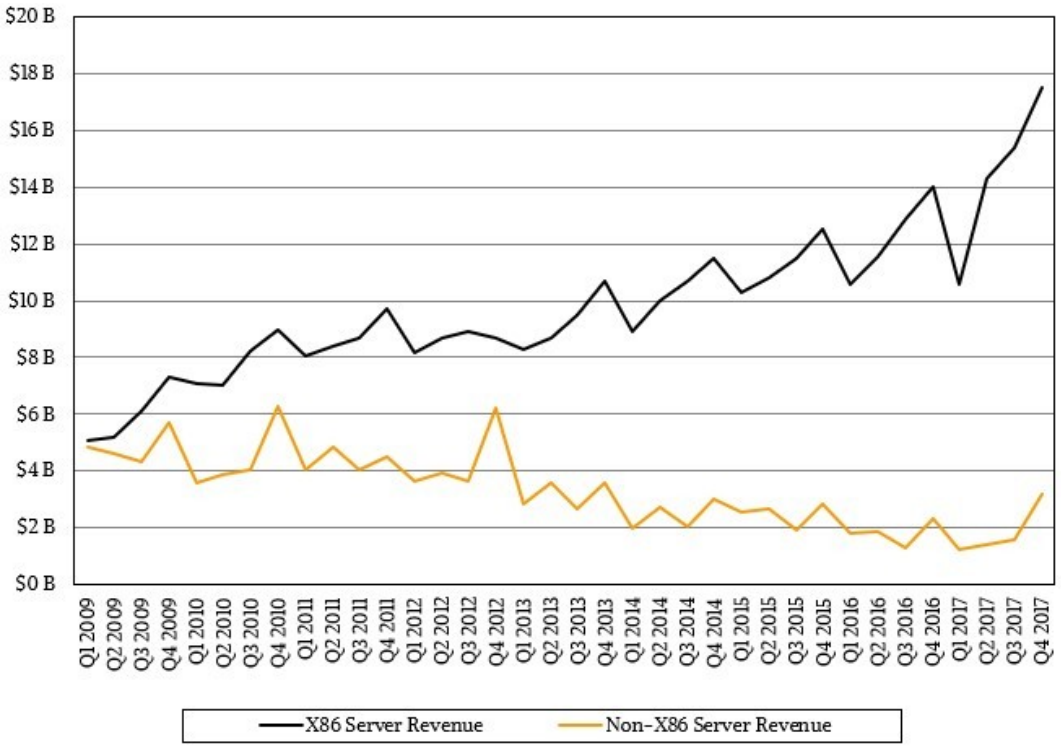
- Total server market revenues: $20.7 B in Q4 2017, 2.8 M servers shipped
- Large revenue jump: general price increases, memory price explosion, big Iron sales (z14 IBM), HPC/AI investment

- Market split into three parts based on cost per server:
  - < $25'000: $15.8 B
    (HEP buys < $5'000/server)
  - $25k-$250k: $1.9 B
  - > $250k: $2.9 B

- Intel takes 85% of the revenues and ships 99.3 % of the x86 servers processors
- The increase in units sold is due to several factors: new Skylake architecture, shift from DELL/HP/etc to cheaper ODM sellers, high demand for high-end machines with GPUs (> $25'000 per unit)
- Hyperscale data centres (> 100'000 servers) grew in 2017 from 300 to 390
- Amazon, Google, Microsoft, IBM have at least 45 centres each

Server Market II

- Intel x86 dominating server market, 99.3% of server units
- Possible contenders:
- IBM Power9
  - Aimed at high-end server and HPC/AI market (combined with Nvidia GPUs)
  - Not power efficient
  - 14 nm process, no plans for 10…7 nm
- AMD EPYC, market penetration rose in 2018, but still relatively low
- ShenWei 260-core processor (based on alpha, 6 Tflops SP)
  - China only, TaihuLight supercomputer; public market?



ARM:

- Applied Micro (now Ampere): new design 32 core 3.3 GHz end 2018
- Qualcomm Centriq: 48-core, 2.6 GHz, 10nm process, first contract with cloud gaming company
  - Doubts about Qualcomm commitment to the project
- Cavium (now Marvell): ThunderX2
  - Available in techlab, power/HS06 similar to Intel Broadwell, price/performance x2 off)

Conclusions

- Run 3 under control, at least as far as ATLAS and CMS (no longer the big experiments!) are concerned
- Run 4 remains a challenge – both for CPU and even more so for disk
- Technology progress per se is still good, but numerous obstacles ahead (CPU, RAM, NAND)
- Novel processors and accelerators difficult to exploit for HEP
- Price/performance advances are slowing down, cost of advances increases exponentially, facing stagnating demand
- Key computing markets in the hand of very few companies
- Technologies HEP relies on under pressure (e.g. HDDs, tapes)
- “Moore’s Law” at risk not because of physics or technology, but for business, financial and economic reasons!

Technology Tracking References

http://www.digitaltvnews.net/?p=30009

https://www.statista.com/chart/12798/global-smartphone-shipments/

http://www.transformingnetworkinfrastructure.com/topics/virtualization/articles/437082-almost-13-billion-cloud-waste-predicted-2018.htm

https://www.gartner.com/newsroom/id/3845563

https://www.semiconductors.org/news/2018/01/02/global_sales_report_2017/global_semiconductor_sales_increase_21.5_percent_year_to_year_in_november/

https://www.qstar.com/index.php/ltfs-linear-tape-file-system/

http://techinsights.com/about-techinsights/overview/blog/samsung-18-nm-dram-cell-integration-qpt-and-higher-uniformed-capacitor-high-k-dielectrics/

https://www.extremetech.com/computing/249075-foundry-futures-tsmc-samsung-globalfoundries-intel-gear-7nm-beyond

https://www.quora.com/When-is-Samsung-going-to-produce-7nm-and-is-it-comparable-to-Intel-10nm-in-terms-of-performance

https://www.neogaf.com/threads/glofo-7nm-details-revealed-the-most-likely-process-node-for-the-next-gen-consoles.1402083/

http://www.icinsights.com/news/bulletins/Semiconductor-Shipments-Forecast-To-Exceed-1-Trillion-Devices-In-2018/

https://www.forbes.com/sites/tomcoughlin/2018/02/05/hdd-growth-in-nearline-markets/2/#66cf02aa3e39

http://www.theregister.co.uk/2018/03/21/seagate_to_drop_multiactuator_hamr_in_2020/

https://www.theregister.co.uk/2017/12/19/seagate_disk_drive_multi_actuator/

http://www.tomshardware.com/news/seagate-wd-hamr-mamr-20tb,35821.html

https://www.extremetech.com/computing/266031-ibms-power9-dent-x86-server-market-oems-prep-new-systems-emphasize-gpu-compute

https://www.nextplatform.com/2018/03/01/server-market-booms-last/

https://www.backblaze.com/blog/hard-drive-cost-per-gigabyte

https://www.theregister.co.uk/2017/12/22/bit_price_decline_and_computestorage_closeness_to_send_enterprise_flash_use_sky_high/

https://hblok.net/blog/posts/2017/12/

https://www.businesswire.com/news/home/20180314005070/en/Record-Breaking-Amount-Total-Tape-Capacity-Shipments

https://www.dramexchange.com/WeeklyResearch/Post/5/4871.html

https://www.snia.org/sites/default/files/PM-Summit/2018/presentations/14_PM_Summit_18_Analysts_Session_Oros_Final_Post_UPDATED_R2.pdf

https://marketrealist.com/2018/01/micron-optimistic-falling-nand-prices

https://www.nextplatform.com/2018/02/09/just-large-can-nvidias-datacenter-business-grow/

https://www.eetimes.com/author.asp?section_id=36&doc_id=1331415