

Tech Watch

L. Duflot



Introduction

- ◆ **Buts:** 1/2 présentation 1/2 discussions
 - ◆ point et perspectives sur les technologies serveur (CPU, mémoire, disque, ...) qui pourrait intéresser LCG-France dans le futur
 - ◆ Pas de “market trends”, réseaux...
 - ◆ Quelles implications concrètes à court terme?
- ◆ **Sources:**
 - ◆ Talks “Tech Watch” CHEP (EGI)
 - ◆ Liste HEPiX Tech Watch
 - ◆ Sites spécialisés

Trends in Computing Technologies and Markets: The HEPiX TechWatch WG

Presented by Shigeki Misawa BNL

Contributions by Andrea Sciaba, Helge Meinhard, Germán Cancio, Martin Gasthuber, Chris Hollowell, Michele Michelotto, Edoardo Martelli, Servesh Muralidharan, Bernd Panzer-Steindel, Rolf Seuster, Peter Wegner, Eric Yen, and the TechWatch WG community.

CHEP 2019, Adelaide, AU, 04-08 November 2019



X86 CPU: AMD Challenging Intel

- Core counts increasing
 - AMD 64 cores vs Intel 28 cores
- AMD cores competitive with Intel
- Memory bandwidth increasing
 - AMD 8 DDR4 memory channels
 - Intel 6 DDR4 memory channels
- Support for more memory ($\geq 4\text{TiB}$)
- I/O bandwidth increasing
 - AMD - 128 PCI-e Gen 4 lanes
 - Intel - 48 PCI-e Gen 3 lanes
- Intel specifics
 - VNNI - neural net instructions
 - Optane Persistent Memory
- AMD specifics
 - Multi-chip “composable” CPU
 - 8 core silicon die “chiplet”
 - I/O & memory controller silicon chip
 - Harbinger of “Lego” brick customized HW?
 - Multiple high profile HPC design wins, broader adoption by server vendors



04-Nov-2019

HEPiX Techwatch Working Group

6

Pénétration augmente chez les hyperscalers

Tire les prix à la baisse

Part de marché x 2.7 en 1 an 1.6% → 4.3% (serveurs)

Choix de plusieurs futures machines HPC américains



- ◆ Un peu le serpent de mer pour ce qui est du calcul scientifique ...
- ◆ Design ARM:
 - ◆ 02/19 annonce de l'architecture **neoverse N1** dédiés aux serveurs (mais pas forcément de calcul), jusqu'à 128 coeurs
 - ◆ Repris par Amazon, Ampere, etc...
 - ◆ Ampere: 80 coeurs N1 modifiés
- ◆ Designs indépendants:
 - ◆ **Cavium Thunder X2** (X1 a été regardé au CERN TechLab) accent mis sur les perf calcul flottants: adopté par Microsoft pour une offre Azure
 - ◆ Thunder X3 en développement
 - ◆ **Fujitsu A64FX** qui vise des HPC (Japon, Espagne,...), là aussi accent sur calcul flottant et ML
- ◆ Des grands groupes comme HPE ont une offre ARM...
- ◆ Roadmap européenne...



Diversifying Landscape of CPUs for HPC

Intel Xeon

- High compute capabilities, relatively low memory bandwidth
- Dominating market position

AMD EPYC

- High compute capabilities, slightly better memory bandwidth
- Emerging market position

IBM POWER9

- Low compute capabilities, relatively good data transport capabilities
- Used for current Top500 #1 and #2, lacking HPC market uptake

Marvell ThunderX2

- Moderate compute capabilities, relatively good memory bandwidth
- Still low overall market uptake

Fujitsu A64FX

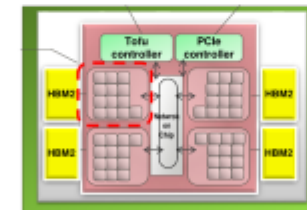
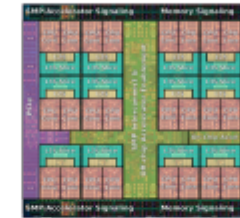
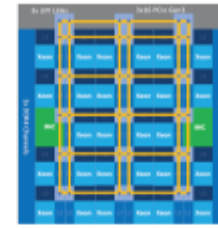
- High compute capability and memory bandwidth
- Used for Japan's exascale system

Upcoming: European Processor Initiative

Mitglied der Helmholtz-Gemeinschaft

06.11.2019

10



GPUs and AI Accelerators

- Slow improvements in CPU performance and performance/watt resulting in development of “domain specific” hardware
 - GPU currently the most popular
 - AI accelerators gaining momentum
 - Significant increase in performance metric for specific problems
- GPU and AI Accelerator Characteristics
 - Hardware architecture designed for specific data flow and compute requirements
 - Large number of simple “execution units”
 - Performance highly dependent on keeping execution units occupied
 - Execution unit occupancy depends on data re-use, locality, and layout regularity
- “Porting” applications to GPU/AI hardware requires paying close attention to keeping execution units occupied.
 - In contrast, with “general purpose” CPUs, emphasis is on structuring application to simplify solving the problem at hand.
 - Domain problem to be solved may not be able to keep execution units occupied
 - Problem exacerbated by lack of universally supported and optimized software stack.



04-Nov-2019

HEPiX Techwatch Working Group

8

Peut être pas pour la grille mais utilisé online (trigger) et analyse



Domain Specific Accelerators

- Software stack key to adoption
 - Cuda (NVIDIA), SyCL, OpenACC, OpenMP
 - Tensorflow, PyTorch, Caffe, others
- CPU to Accelerator link critical
 - High bandwidth/low latency
 - Creation of new interconnects
 - NVLink, CCIX, CXL, OpenCAPI, Gen-Z, PCI-e Gen4/5
 - New storage/memory architectures possible with interconnect.
- Available accelerators
 - NVIDIA V100 (GPU/AI) - market leader
 - AMD Radeon Instinct (GPU)
 - Google TPU v3 (AI, cloud only)
 - Habana Gaudi (AI, PCI-e)
 - Intel Nervana (AI) late 2019
 - Intel X^e (GPU) in 2020
- Drawbacks to Domain Specific hardware
 - Costly development (skill and money)
 - High risk
- FPGA Option
 - Programmable hardware
 - “IP” building blocks available
 - CPU blocks
 - AI blocks
 - DSP blocks
 - I/O blocks (PCI-e, CCIX, etc)
 - Memory controllers
 - Network connectivity
 - Build your own accelerator or standalone domain specific “CPU”
 - Chiplet implementation of selected functions
 - Available hardware
 - Xilinx “Versa”, Intel/Altera “Agilex”



04-Nov-2019

HEPiX Techwatch Working Group

9

Les usages augmentent beaucoup dans la communauté: trigger, analyse ML
Voir discussions prospectives IN2P3 et “ATLAS CAF users”



Memory: DRAM and Persistent

- Evolution of DRAM continues
 - DDR5 - 6400MT/s, 32Gb chips
 - GDDR6 - 16Gbps/pin, 16Gb chips
 - HBM2 - 2.4Gbps/pin, 12 die stack, 8Gb die
 - Multiple serial attached memory standards in development (augment or replace DDR4)
 - R&D continuing on next generation DRAM
 - Latency remains unchanged
 - DRAM market
 - DRAM prices stabilized
 - Production capacity flat
 - Revenues expected to decline 38% 2019
 - Persistent memory on the horizon
 - Significant area of innovation
 - Optane/3D NAND on memory bus
 - Two standards Intel and JEDEC
 - Different “access” models*
 - Like DRAM but larger
 - Like DRAM but persistent
 - Like disk but faster
 - Different requirements for different access models*, one or more of the following
 - CPU support
 - BIOS/OS support
 - Extensions of DDRx bus standards
 - Application support
- * SNIA NVM Programming model



04-Nov-2019

HEPiX Techwatch Working Group

10

Pas toujours vrai, très fluctuant

Coût trop élevés pour WLCG



Solid-State Storage

- 3D NAND Flash
 - Still non-volatile memory of choice
 - Capacity increases continuing
 - More layers (128 now, multiple 100's in future)
 - More bits per cell (4 bits now, 5 announced)
 - Low latency flash on the market to compete with 3D XPoint (aka Optane)
 - Rapid migration of SSD to NVMe
 - Flash revenue to shrink 32% in 2019
 - Flash oversupply remains in 2019
- Intel/Micron 3D XPoint
 - Potential challenger to 3D NAND
 - Lower latency/higher endurance compared to 3D NAND
- NVMe/NVMeof catalyzing new storage architectures
 - Removes HW/SW bottlenecks
 - Unified software stack
 - New form factors (e.g. EDSFF)
 - Network attached flash
- NVMe feature enhancements
 - NVMe v1.4 - Device capabilities
 - NVMeoF v1.1 - NVMe over networks
 - NVMe-Mi - Enclosure management



04-Nov-2019

HEPiX Techwatch Working Group

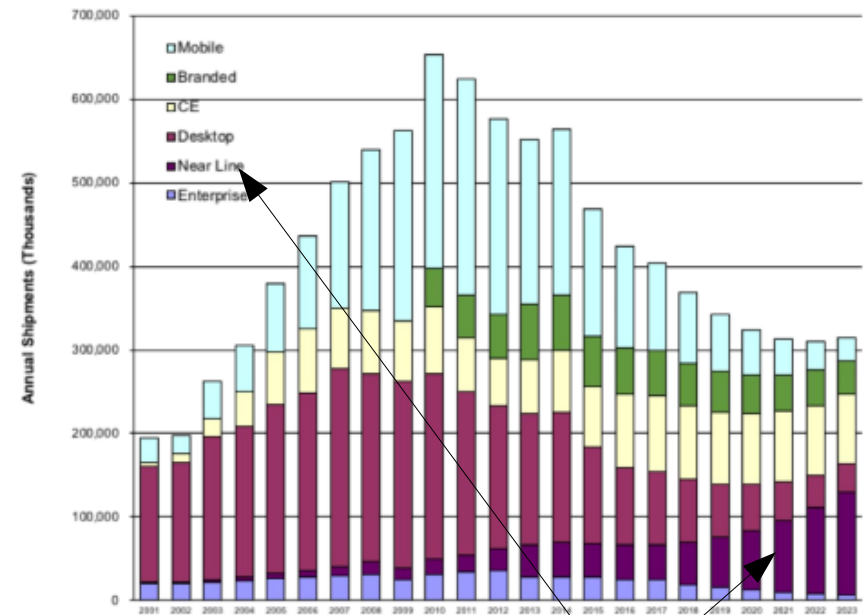
12

Ne plus acheter du SATA mais NVMe ou équivalent
Premiers produits PCIe Gen 4 sur le marchés (cf AMD EPYC "Rome")
Prix TB nu ~ 2x prix HDD environné mais reste raisonnable pour petite volumétrie



Magnetic Hard Disk Drives (HDD)

- Technology problems
 - Reaching capacity/performance limits with current technology
 - Near term solutions alter drive behavior (SMR)
 - Limited adoption
 - May require application support
 - Longer term solution still not available
 - HAMR/MAMR pushed back to 2020
 - Multi-actuator pushed back to 2020
- Total HDD market shrinking
 - Competition from SSD continuing to grow
 - Sole revenue growth is in capacity (near-line) HDD used by hyperscalers and Big Science
 - Hyperscalers purchasing ~50% of all HDDs
 - Down to 3 HDD vendors with 4 factories between the top 2 vendors
 - Technology and economic risks increasing due to shrinking market



“Notre” marché



04-Nov-2019

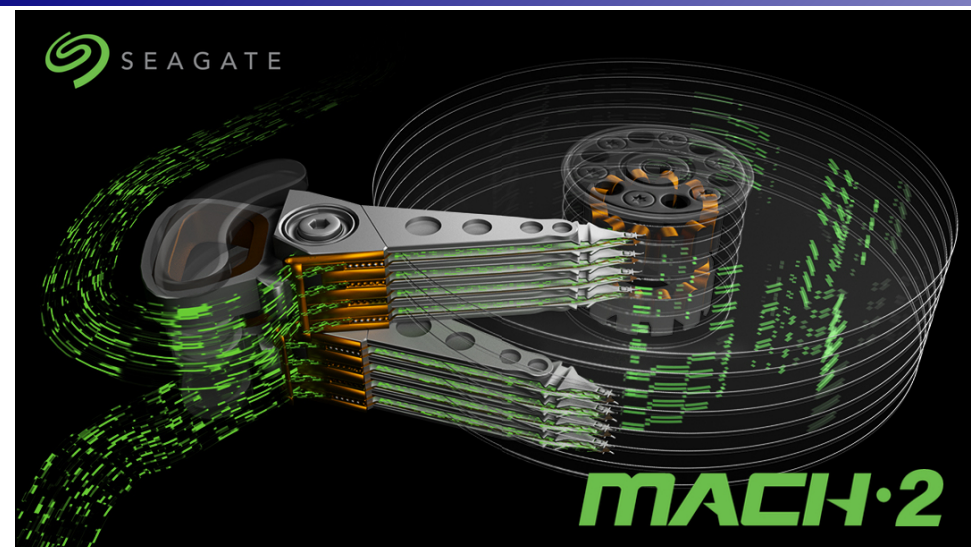
HEPiX Techwatch Working Group

11

Si le marché grand public s'effondre, le marché Nearline est en forte croissance



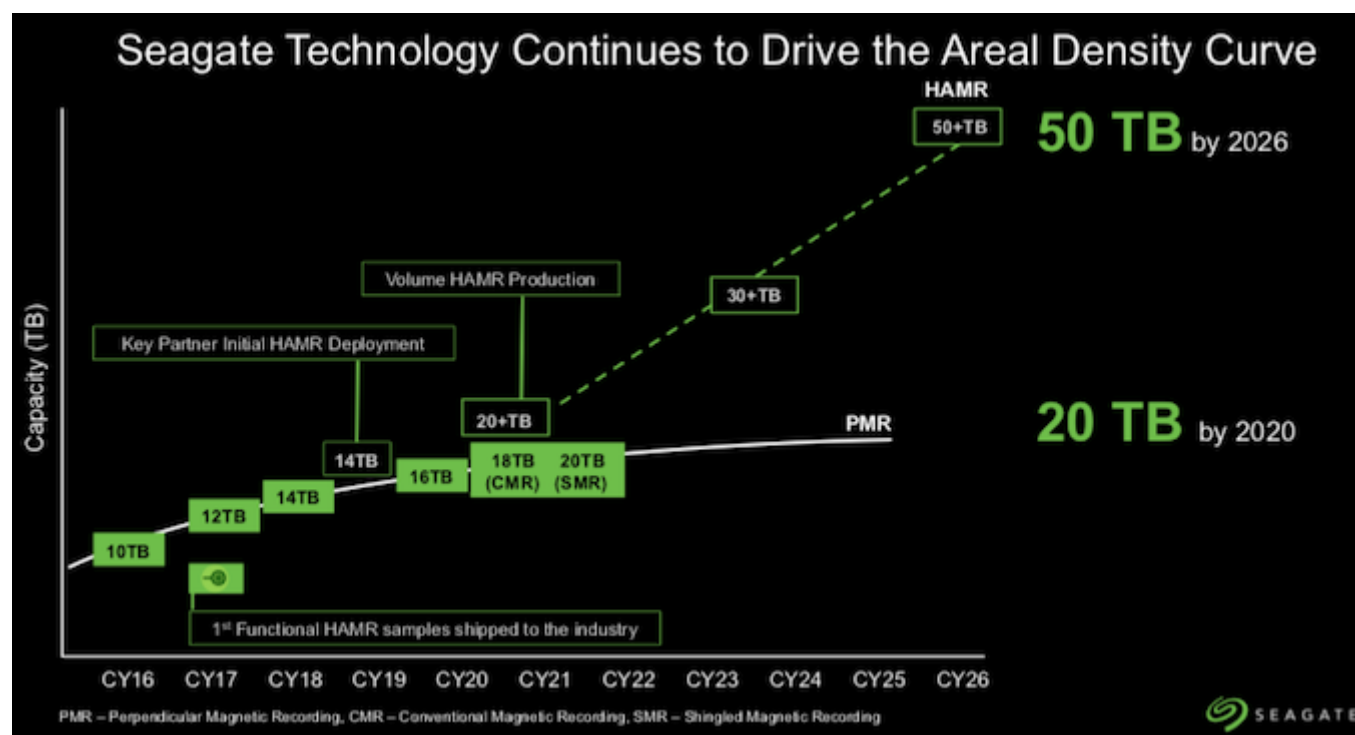
HDD: nouvelles technologies



“dual actuator”, deux bras de lecture
Pour conserver les IOPs/TB
Premier produits sur le marché

Source: anadtech

e.g H(eat)A(ssisted)MR



Open 19

- Aims at support data center design, open, economical and customizable
- Similar to Facebook OCP Open Compute Project that is more address to major internet companies
- Open19 focused on optimizing data center of any size including small edge platforms, unlike OCP
- Define a common, open standard server form factor, using rack, cages and pre-defined network and power

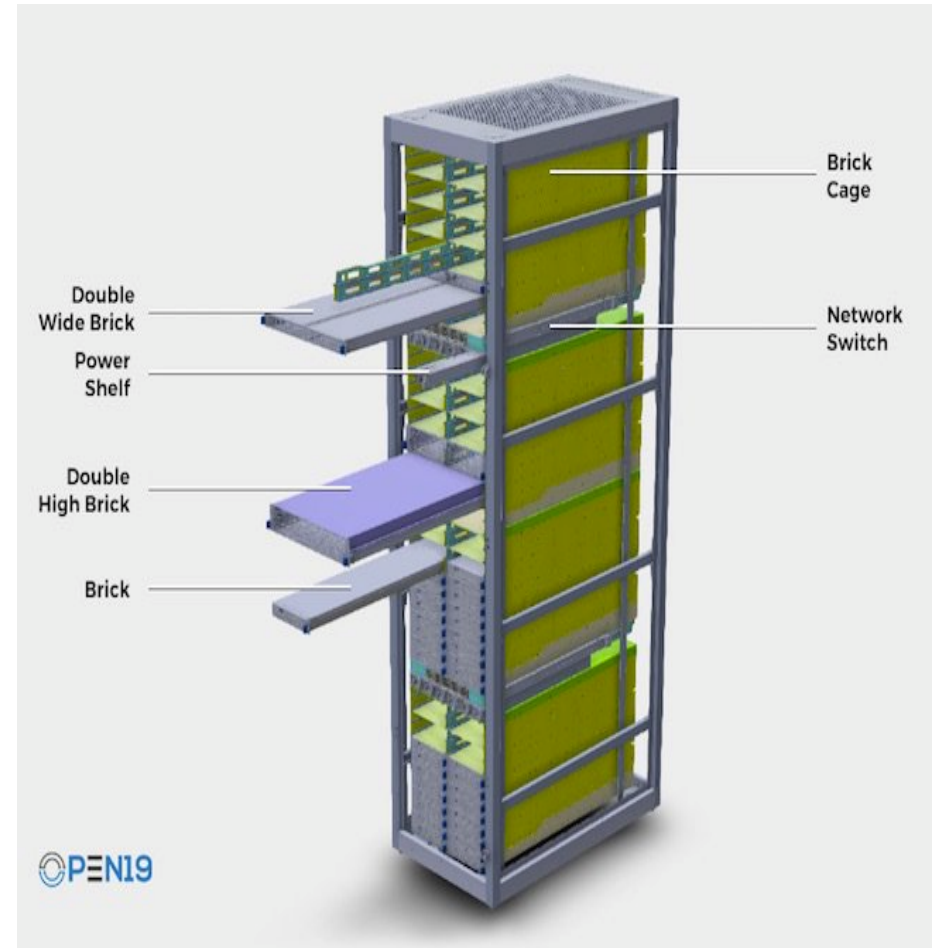


HOW Workshop
18-22 March 2019
Jefferson Lab, Newport News

Open 19

- Standard 19" 4-post rack, Brick cage
- Brick (B), Double Wide Brick (DWB), Double High Brick (DHB), Double High & Wide Brick (DHWB)
- Power shelf: 12v distribution, OTS power modules with any AC or DC inputs *12kW*
- Optional Battery Backup Unit (BBU)
- Networking switch (ToR)
- Snap-on power cables – up to 400w per brick, linear growth with size
- Snap-on data cables – up to 100G per brick, linear growth with size

HOW Workshop
18-22 March 2019
Jefferson Lab, Newport News



- **Some numbers:** Infra 40-60min per rack, 96 servers: 10 minutes per rack, Total ~100min per rack (2 technicians), 2 days for a full cluster – 1536 Servers

L. Atzorti TechWatch server WG workshop report

Et pour nous ? Discussions



- ◆ Sur le papier avantages / INTEL:
 - ◆ Bande passante mémoire (attention à bien remplir les 8 slots mémoire!)
 - ◆ PCIe Gen 4: débits vers SSD, carte RAID, etc... si Gen 4 est supporté par ce matériel
 - ◆ Premiers SSD NVME PCI gen 4 sur le marché
- ◆ Spec: voir présentation de M. Puel
- ◆ Coûts? A priori le CPU lui-même a un prix très agressif / INTEL
 - ◆ Il n'y a pas que DELL....
- ◆ À regarder très rapidement !
- ◆



- ◆ Utilité encore incertaine pour le calcul scientifique
- ◆ Serait bien que CERN TechLab continue d'avoir des machines pour faire des benchmarks, quid de Barcelone (Mare Nostrum)?



- ◆ Les prix continuent de baisser mais restent trop importants pour envisager les SSD comme remplacement des HDD sur les serveurs de stockage
- ◆ ... mais la différence de prix permet d'envisager des usages de niche sur les serveurs spécifiques:
 - ◆ Tape/dCache buffer? Déjà le cas?
 - ◆ Machines dédiées analyse IO-bound (ntuple) ?
- ◆ Sur un serveur CPU “type”: HDD → SSD surcoût de ~5%?
 - ◆ Quel gain ?
 - ◆ Analyse en copy-to-disk?
 - ◆ ATLAS: pileup overlay (mais va changer), merging
 - ◆ Si nous avons une machine nous pourrions utiliser les benchmark containerisés pour mesurer



- ◆ Encore pas mal de marge à moyen terme en IO (“dual actuator”) et en capacités
- ◆ 50TB /disque en 2026
 - ◆ Encore possible de faire du RAID (temps de reconstruction)?
 - ◆ 1 PB dans un serveur de stockage....
 - ◆ Effet des indisponibilités
 - ◆ Quelle bande passante réseau?
 - ◆ CEPH+EC: temps de reconstruction?



Autre point de vue – mêmes questions

- ◆ Quel serveur disque standard en 2026?
 - ◆ Volumétrie (HDD)
 - ◆ Quand le SSD entrera-t-il dans le jeu?
 - ◆ Réseau ? 40G/s?
- ◆ Quel serveur CPU en 2020 ?
 - ◆ AMD vs INTEL
 - ◆ SSD?
 - ◆ Réseau 10-20G/s?
- ◆ Machines spécifiques?
 - ◆ E.g. IO machines pour labo ou ferme d'analyse
 - ◆ GPU



Backups

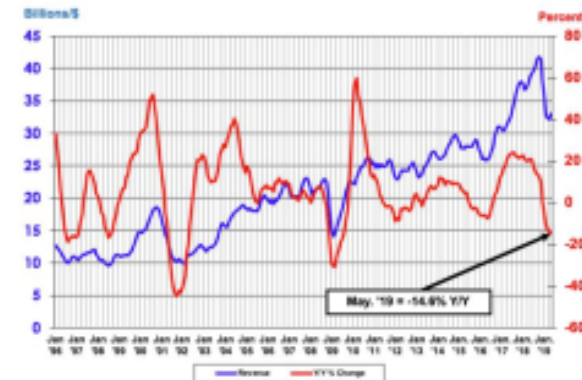


Semiconductor Device Market and Trends

- Global demand for semiconductors in CY18 topped 1 trillion units shipped for the first time
- Global semiconductor sales dropped 14% in 1H19 relative to 1H18 (semiconductors.org)
- Revenue drop driven by IC sales, particularly with memory. (wsts.org)
- Markets expected to recover in 2020
- China/US trade war disrupting “supply chain”
- Long-term outlook remains promising, due to the ever-increasing semiconductor content in a range of consumer products
- Foundries migrating to 7nm and EUV lithography. 5nm in development
- Down to 3 foundries at the leading edge
- Technical and economic factors driving increased innovation in IC's unrelated to node shrinks..

Worldwide Semiconductor Revenues

Year-to-Year Percent Change



Source: WSTS

Tracking Semiconductor Unit Growth



Source: IC Insights

HEPIX Techwatch Working Group

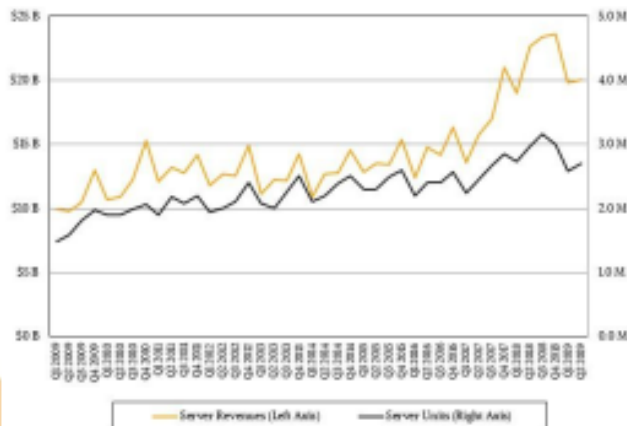
04-Nov-2019

4

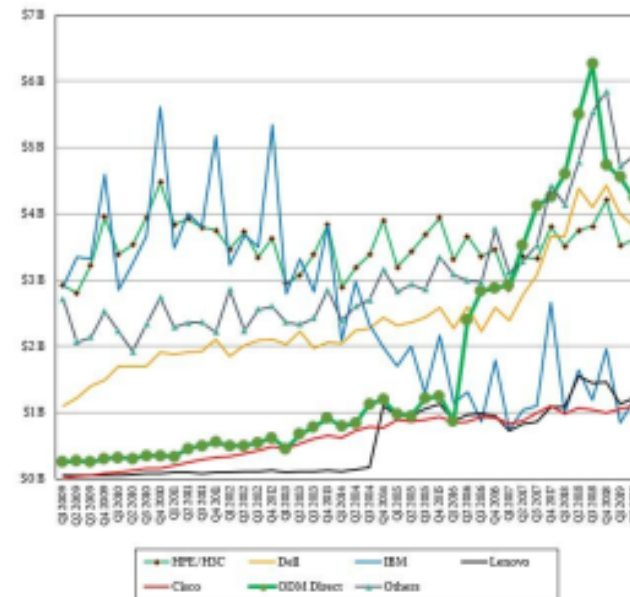


Server Market

- Global server market revenue down 11.6% YoY in 2Q19
 - \$20 billion in 2Q 2019 with 2.7 million servers shipped
 - Slow down in all areas
- Server ASP stable
 - Resurgent AMD affecting CPU prices
 - DRAM/Flash drop due to decreased demand and increased supply

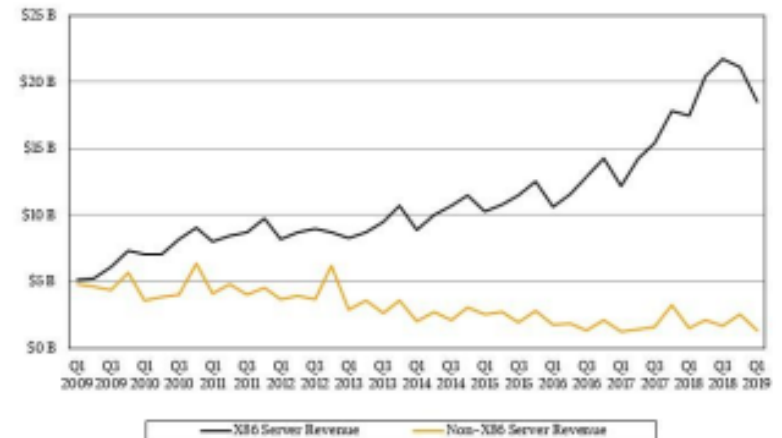


- In FY2018, Hyperscalers purchased 35% of Intel's processor
- ODMs have larger market share than any single Tier-1 brand (Dell, HPE, Lenovo)



Non-x86 CPU

- IBM Power 9 - Differentiation
 - SMT4/SMT8
 - Early support for PCI-e Gen4
 - High speed access to accelerators and I/O devices
 - Coherent access via CAPI/OpenCAPI
 - CPU support for NVLink connectivity to Nvidia GPUs
 - Buffered memory
 - Open Memory Interface (OMI) serial attached memory
- ARM - Slow progress in the data center
 - Next generation in development
 - Marvell ThunderX3, Ampere "QuickSilver"
 - Fujitsu A64FX ("Post-K" HPC) in 2021
 - ARM Neoverse (CPU microarchitecture for servers)
 - In EU EPI and China HPC roadmaps
 - Leverages existing ARM ecosystem



- RISC-V - Open source ISA
 - Following path taken by ARM
 - Currently not in data center, but growing presence in embedded/SoC market
 - In EU EPI accelerator roadmap



04-Nov-2019

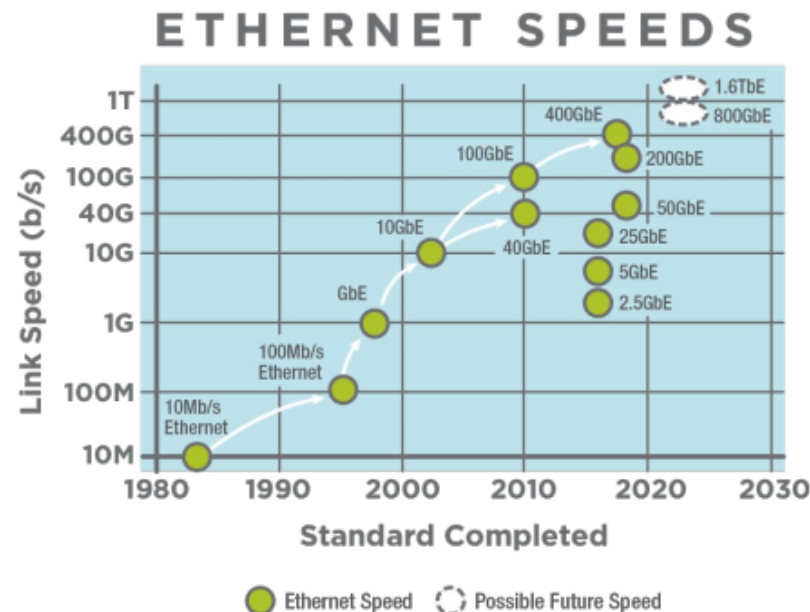
HEPIX Techwatch Working Group

7



Network: Rapid Change

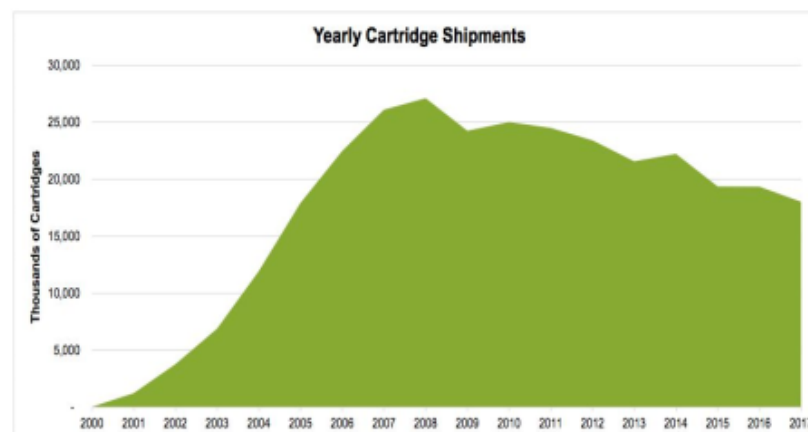
- Evolution of single lane bit rate
 - 25Gbs -> 50Gbs -> 100Gbs
- Changes in available networks
 - 400Gb Ethernet industry standard
 - 200Gb SlingShot “HPC” Ethernet (Cray)
 - 200Gb HDR Infiniband (NVIDIA/Mellanox)
 - OmniPath dead (Intel)
- Ethernet in Flux
 - Shorter technology lifecycles driven by hyperscalers
 - Higher bandwidth at lower costs yesterday
 - Dizzying array of non-interoperable optical options (some IEEE std and others industry standard based)
 - Being partially driven by hyperscalers
 - Pace of change exceeding IEEE standards process.



Magnetic Tape

- Tape technology still improving
 - LTO-Gen 8 - 360MB/s, 12TB media
 - IBM TS1160 - 400MB/s, 20TB media
 - LTO-9 expected in 2020
 - Order of magnitude increase in capacity demonstrated in lab
- Tape Market
 - Market in decline past 10 years
 - Revenue estimates at 0.7B USD
 - IBM is the leading drive manufacturer and the dominant driver of tape R&D.
 - Two remaining media vendors only recently settled legal dispute
 - Can market support 4 tape library vendors ?

Unit Shipments: Calendar Year



- Risk exposure with Tape
 - Drop in \$/TB critical for Big Science
 - R&D and drive manufacturing effectively controlled by a single company
 - Hyperscaler influence in tape market is unclear



04-Nov-2019

HEPiX Techwatch Working Group

13

