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

Challenges

- Huge computing demands of Big Science projects
  - LHC Run 4 / phase II
  - DUNE
  - SKA
  - Other Big Sciences
- Understanding better where technology and markets go becomes indispensable
- Wide area of topics to be covered in a regular and systematic manner

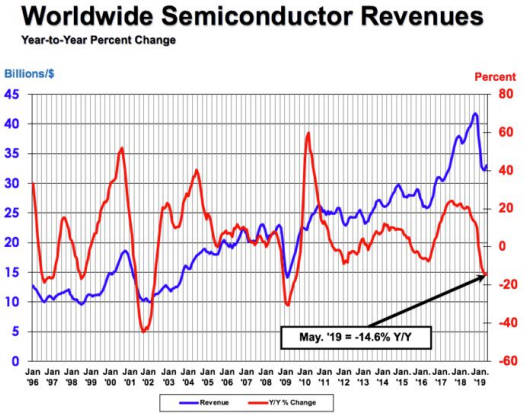
HEPiX Techwatch Working Group

- Leverage expertise/interest in community
- Follow technology and market trends in computing
- Forum for interacting with other interested parties
- Advise planners, architects, and developers on direction of computing technology
  - Presentations at meetings, workshops, and conferences
  - Web pages/documents

Semiconductor Device Market and Trends

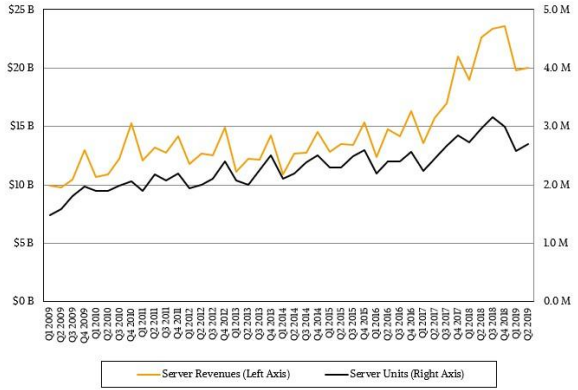
- Global demand for semiconductors in CY18 topped 1 trillion units shipped for the first time
- Global semiconductor sales dropped 14% in 1H19 relative to 1H18 (semiconductors.org)
- Revenue drop driven by IC sales, particularly with memory. (wsts.org)
- Markets expected to recover in 2020
- China/US 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..

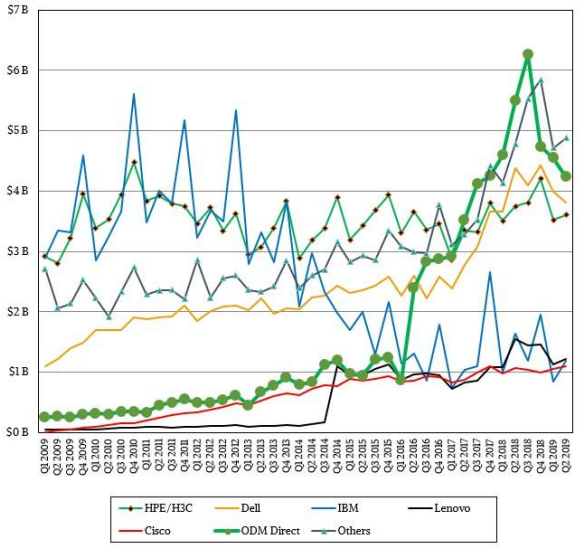


Server Market

- 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)



X86 CPU: AMD Challenging Intel

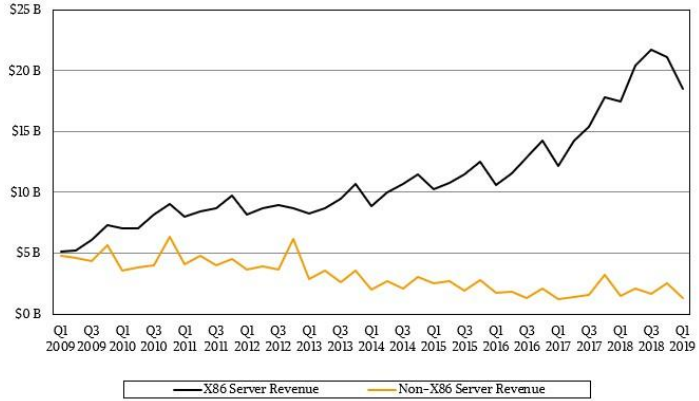
- Core counts increasing
  - AMD 64 cores vs Intel 28 cores
- AMD cores competitive with Intel
- Memory bandwidth increasing
  - AMD 8 DDR4 memory channels
  - Intel 6 DDR4 memory channels
- Support for more memory (≥4TiB)
- I/O bandwidth increasing
  - AMD - 128 PCI-e Gen 4 lanes
  - Intel - 48 PCI-e 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

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

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 dependents 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.

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
    - Xilinix “Versa”, Intel/Altera “Agilex”

Memory: DRAM and Persistent

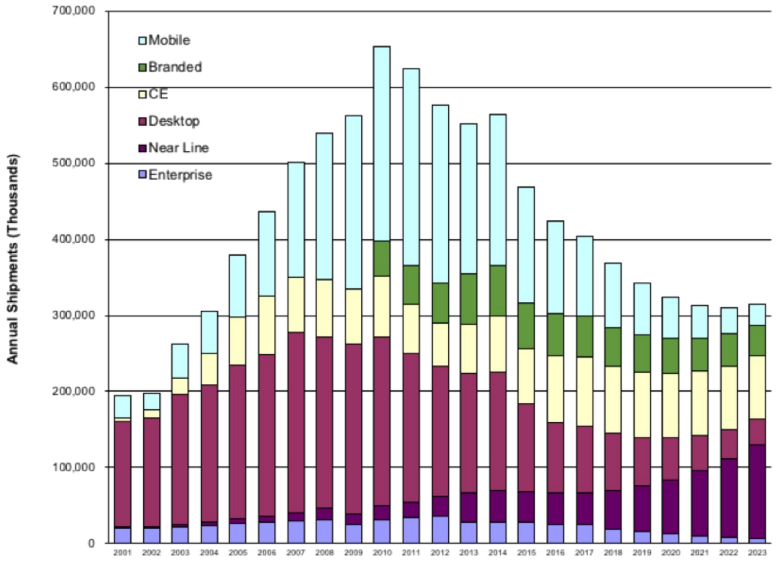
- Evolution of DRAM continues
  - DDR5 - 6400MT/s, 32Gb chips
  - GDDR6 - 16Gbps/pin, 16Gb chips
  - HBM2 - 2.4Gbs/pin, 12 die stack, 8Gb die
  - Multiple serial attached memory standards in development (augment or replace DDR4)
  - R&D continuing on next generation DRAM
- Latency remains unchanged
- 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

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



Solid-State Storage

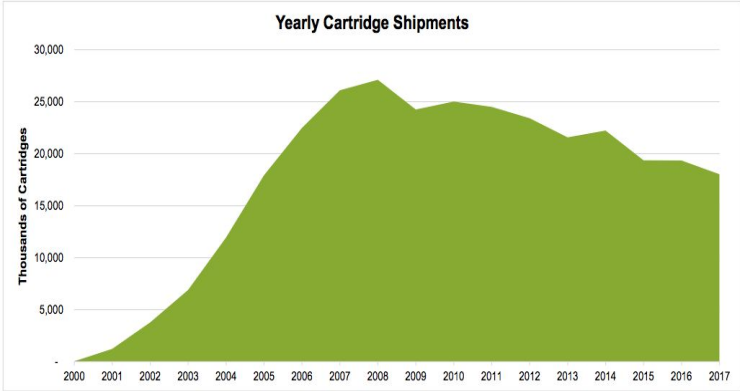
- 3D NAND Flash
  - Still non-volatile memory of choice
  - Capacity increases continuing
    - More layers (128 now, multiple 100's in future)
    - More bits per cell (4 bits now, 5 announced)
  - Low latency flash on the market to compete with 3D XPoint (aka Optane)
  - Rapid migration of SSD to NVMe
  - 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

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 ?

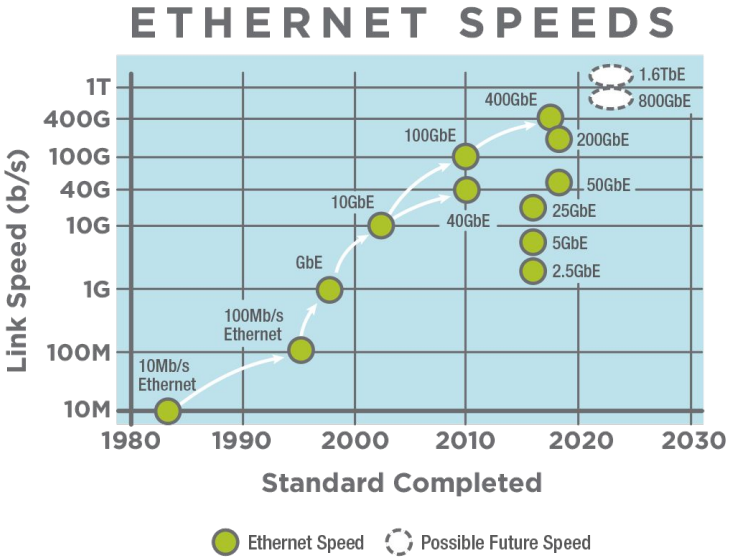
Unit Shipments: Calendar Year



- 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

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.



Conclusions (1)

- Technology progress per se remains good, but obstacles ahead
- Key computing markets in the hand of very few companies
- Hyperscalers influencing direction of technology
  - Risk of divergence from needs of HEP
- Price/performance advances are slowing down
- Accelerators economics
  - GPU availability driving application porting, driving more GPU development
  - TPU availability (enabled by Google), driving similar feedback loop for ML hardware
  - DARPA attempting to “kickstart” similar feedback loop for graph analytics (HIVE program)
  - Can FGPA’s change economics to democratize availability of other domain specific hardware?
  - Can HEP effectively utilize ?
- FPGA/chiplets/interconnect enable novel compute/storage architectures
  - Any applicability to HEP problems ?

Conclusions (2)

- Storage markets are a concern as data volumes increase
  - SSDs not cost effective for HEP near-line storage
  - HDD capacity appears to be stalled, market financials a concern
  - Market financials for tape a big concern.
- Meta-conclusions
  - Previously believed physics or technology would limit progress
  - Business economics now appears to be another limitation, if not more important one.
- Meta-meta-conclusions
  - Tech watch is becoming essential for the Big Science

Final Remarks

- Thanks to all WG members and the conveners for their dedication and contributions so far
- Interested? Never too late to join – contact conveners or Helge Meinhard
- Ideas for opportunities (in particular presentations)? Contact conveners or Helge Meinhard

Techwatch WG

- Chairs: Helge Meinhard, Bernd Panzer-Steindel
- Six sub-groups (with conveners):
  - General market trends, semiconductor markets, unit sales (Servesh Muralidharan, Peter Wegner)
  - Server markets (Chris Hollowell, Michele Michelotto)
  - CPUs and accelerators (Andrea Sciaba, Eric Yen)
  - Memories (Shigeki Misawa)
  - Storage (German Cancio, Martin Gasthuber)
  - Network (Edoardo Martelli)
- 59 persons subscribed to mailing list

References

- Presentations:
  - HOW2019 workshop: https://indico.cern.ch/event/759388/sessions/295048/#20190319
  - HEPiX spring 2019 workshop: https://indico.cern.ch/event/765497/sessions/303981/#20190328
- Techwatch WG internal Web site: http://w3.hepix.org/techwatch/
- Subgroup documents:
  - CPUs, GPUs and accelerators: https://docs.google.com/document/d/1dUcuuSubLO4WeFzF0mlsA8mPgHV-46z6Rga7wCt0kFk/edit?usp=sharing
  - Memory: https://docs.google.com/document/d/1a8K_BA8ipy5l0NvcNjrOLqDsN_RjJXDnX4LHRrfrNE0/edit?usp=sharing
  - Storage: https://docs.google.com/document/d/1IS4_raw7PE0wVTNWDJmUGmneV1zpg9-29vz7XP4ChRA/edit
  - Networking: https://twiki.cern.ch/twiki/bin/view/HEPIX/TechwatchNetwork/WebHome