Technology Shifts

HEPiX Fall 2024 Workshop
November 4, 2024

Presenter: Shigeki Misawa
Scientific Data and Computing Center (SDCC)
Brookhaven National Laboratory

HEPiX Techwatch Working Group

- Advances in compute, network and storage technologies are an expected and essential part of the lifecycle of HEP and NP experiments
- Advances are assumed to make everything better, faster, and cheaper.
- However, changes in requirements and advances in technology can trigger the need to changes to existing practices
- Monitoring technology and requirement trends is critical in identifying areas where existing practices may be disrupted
- The HEPiX Techwatch Working Group, was created in 2018 to monitor trends in technology and provide some early warning of these events

Working Group Charge

The Working Group is tasked with the following duties:

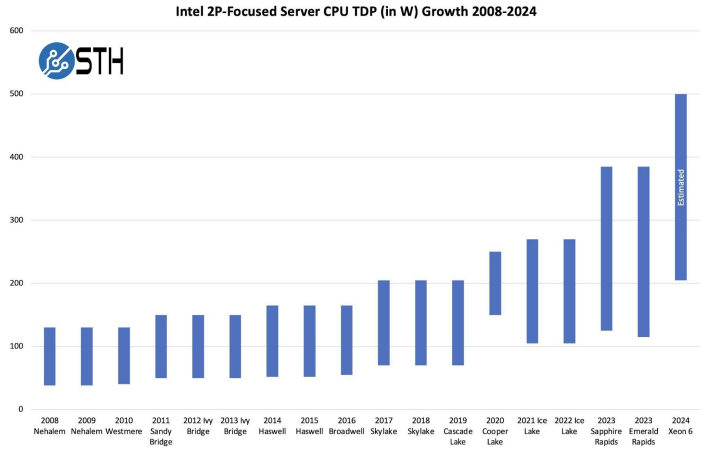
- Understand the trends and the direction of the technology markets using publicly available sources
- Assist in making cost predictions and optimizing investments, taking also into account sustainability
- Provide technical, and where possible financial, risk assessments for technologies
- Leverage the expertise of the HEPiX community
- Inform the HEPiX board about technologies that may warrant a more in-depth investigation

Techwatch Areas of Interest

- General market trends
- Server and data center infrastructure
- Processing units, memory, buses and interconnects
  - Semiconductor process technology
  - Chiplets and die interconnect
  - x86, ARM and Power CPUs
  - GPU
  - AI/ML processors
- Storage
  - Disk
  - Flash (solid state)
  - Tape
- Network
  - WAN
  - LAN

Trends in CPUs: It's All About Power (Consumption)

- CPU TDP increasing [1]
  - 2U server power approaching 1KW
- GPU TDP even higher
  - 700 W NVIDIA H100 TDP
  - 400 W previous generate NVIDIA A100 TDP
- NVIDIA DGX H100
  - Up to 10.2KW in 8U
- Driving investigation of direct chip and RDHx cooling solutions



CPU TDP Growth over time from ServetheHome.com [1]

- Numbers to Note
  - $917K / year - Electricity cost for 1MW of IT load @ US average $0.0872/KWH and a PUE of 1.2
  - $917/year - Cost of electricity for 1KW server
  - How much of the operational budget be taken by power costs ?

[1] Why Servers Are Using So Much Power TDP Growth Over Time, ServetheHome.com

CPU Market Segmentation

- High Performance
  - Intel “P” Cores/ AMD Zen5 - Target performance
- “Efficiency”
  - Intel “E” Cores/AMD Zen5C/AmpereOne - Target performance per watt
  - Differing engineering tradeoffs made by Intel/AMD/Ampere potentially affects performance of cores on HEP workloads, e.g. single threaded vs multi-threaded cores
    - Are they targeting the same market or different markets?
- “Other”
  - Large L3 cache (AMD 3D V-Cache)
- Determining “Best” CPU
  - Requires workload specific benchmarks
  - Answer may differ based on data center power and cooling constraints.

Trends in Nearline Storage

- HDD capacity continues to grow
  - In Jan ‘24 Seagate announced “Regular” and Shingled HAMR drives @ 30TB and 32TB
    - Still not generally available
  - In Oct ‘24 Western Digital announced CMR and SMR drives @ 26TB and 32TB respectively
- Issues to track
  - HDD IOPs/TB decreasing over time
    - Multi-actuator drives available from Seagate but viability is an issue
  - SMR drives ~50% of drives shipped [1]
    - Limited use in HEP/NP due to lack of software to make them usable
    - Commercial cloud appear to be the primary consumers

[1] Western Digital Begins Volume Shipments of 24TB CMR HDDS; Industry Adoption of SMR Strengthens as 28TB SMR HDD Ramps | Western Digital

Effects on Nearline Storage (Example)

- Fixed storage budget yields roughly fixed # JBOD chassis in the data center (cost per chassis over generations is relatively stable)
  - Total IOPS remains relatively stable, while storage capacity increase
- Failed disk rebuilds problematic as capacity increases
  - Time and impact on performance
- System level mitigations
  - IOPS limitations - Use of more complex, tiered storage capabilities to split out small files and meta-data to separate storage tier (e.g. SSD) from “bulk” data (HDD) at the node level
  - Rebuild Issues - Use of dRAID and similar technologies to reducing time spent in degraded mode and reduce operational impact caused by rebuild load
  - Consider system level tiered storage to compensate
- Effectiveness and cost impact of these changes TBD

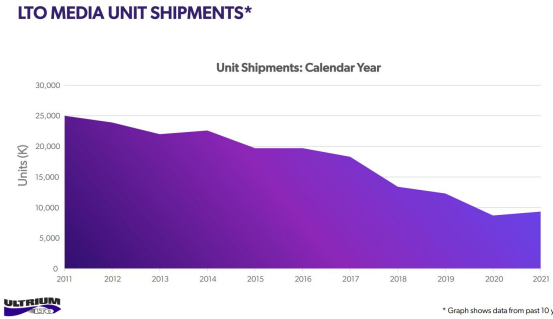
High Performance Buffering Systems

- Increasing number of HDD spindle in data buffering systems are required
  - To keep up with increasing DAQ data rates
  - To keep up tape drives streaming
- High I/O performance of flash suggest an alternative, but some open questions remain
  - SSD endurance - are 3 DWPD SSD required/sufficient ? Can 1 DWPD “read optimized” work ?
  - What fraction of the product “glossy” performance numbers are actually achievable in real world 24x7 sustained write environments ?
  - What is the overhead and performance of software RAID? Are “exotic” HW RAID solutions better ?
- Do the economics work out ?

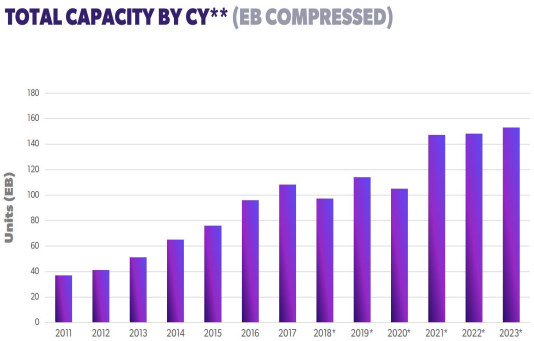
Trends in Tape Storage

- Events in the tape market
  - IBM TS1170 - 50TB / cartridge. No backward compatibility
    - Strategic shift or technology limitation ?
  - IBM Diamondback “library in a rack” targets cloud and **R**edundant **A**rray of **I**ndependent **L**ibraries (RAIL) archives
    - Cloud tape archives expected to exceed Enterprise archives, in exabytes, by CY2025 [1]
    - Divergence of cloud and enterprise markets ?
  - Total LTO cartridges shipped has been declining, total exabytes shipped is flat
- What is the end game if demand growth continues to slow ?

[1] ”Tape Technology Update” presentation by Rich Gadomski (Fujifilm) at the Library of Congress Designing Storage Architectures for Digital Collections meeting April 2024



https://www.lto.org/wp-content/uploads/2022/04/LTO-Ultrium-2021-Media-Shipment-Report-Slides_FINAL-1.pdf



https://www.lto.org/wp-content/uploads/2024/05/LTOUltrium_2023_MediaShipmentReport.pdf

Call for Participants

- We need people to volunteer to lead subgroups or contribute information.
- Meeting times and agenda are posted in CERN indico.
  - https://indico.cern.ch/category/10621/
- Techwatch mailing is available for communication.
  - hepix-techwatch-wg@hepix.org