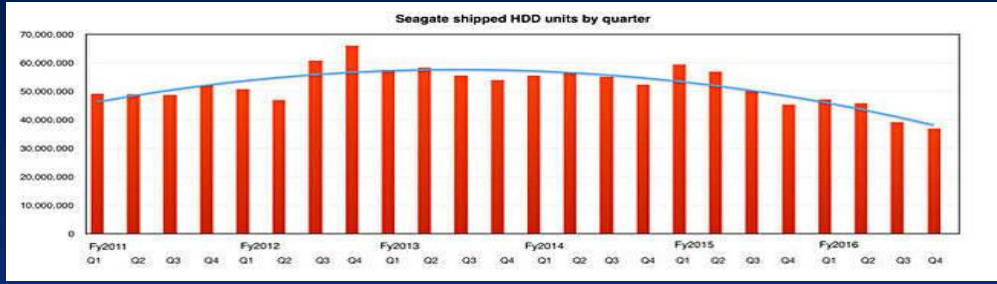
Networks in the Time of Flash

Jim O'Reilly
President, Volanto
A Consulting Company

Santa Clara, CA
August 2016

Storage is at a Crossroad

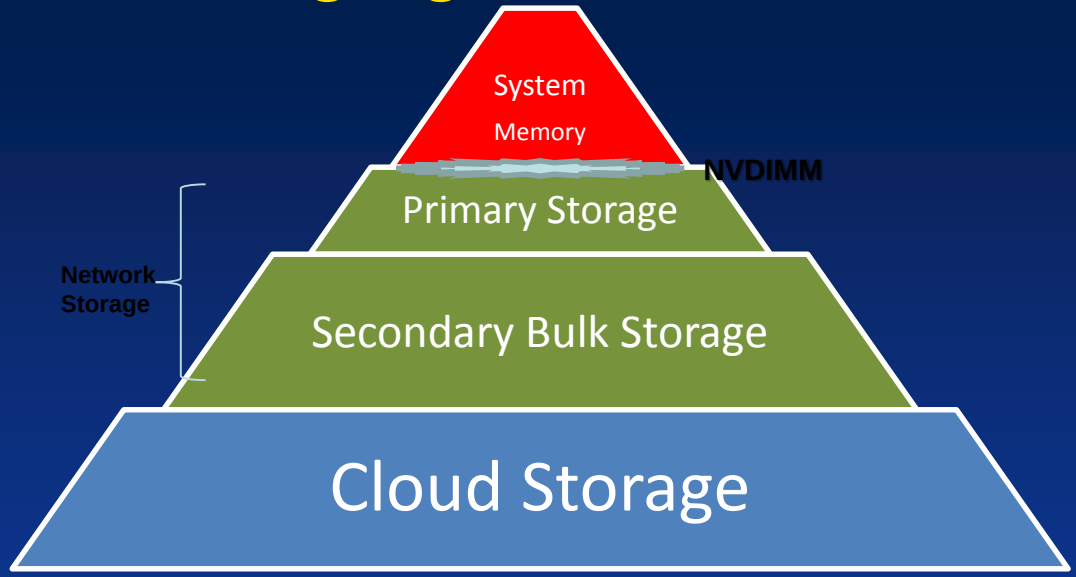


From TheRegister.com

- We've passed "Peak disk"
- New speeds and capacities
- New connectivities
- New ways to use drives
- Consumer of IO is changing rapidly, too

The Tiers are Changing

- New tiering
- Solid-state primary storage
  - NVDIMM
  - SSD



- Bandwidth per drive 5x (SATA) to 50x (NVMe)
- IOPS – 400x to 10,000x

Systems IO-Starved

- Today, 3,000 Docker containers per server
- HDD would give 1/20th IOPS per container!
- We NEED flash solutions!

- We are moving towards much higher core counts and the use of GPUs
- Add HMC or its variants and core memory moving into terabytes will need feeding

Data Protection Model Changing

- Appliance-level redundancy is replacing RAID
  - Erasure coding and replication
- Hyper-convergence model
  - Remote write needed
  - Sharing between server nodes
- Appliance model also needs remote write
- Flash raises the bar for networks in either case

The Storage LAN

- We've passed peak SAN, too

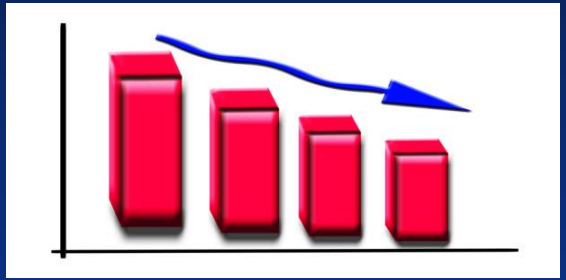
- The future of storage networks is Ethernet S-LAN
  - Horsepower, technology investment, common fabric
- Ethernet is evolving rapidly
- You should be using 10GbE now
  - 25GbE is coming this year
  - 50 GbE will arrive in two years
- Use quad-links to connect storage to TOR switches

Is Speed Enough?

- High-speed Ethernet traditionally means high CPU overhead
- RDMA partially solves this
  - Latencies are much lower
  - Traffic is reduced on the network
  - CPU overhead drops
- Traditional file stack still slows things down

NVMe Shrinks I/O Stack

- Queue-based solution with interrupt consolidation
- Will be available over RDMA Ethernet
- Lowest latency, highest bandwidth
- Currently needs a special NIC

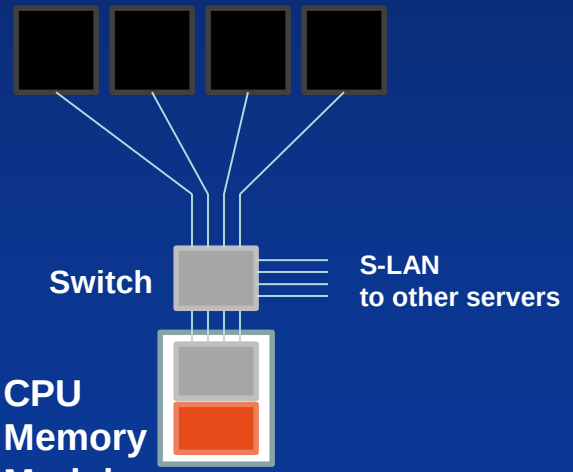


- Cost offset: Performance gains will reduce server count needed in many use cases

Should Drives Connect to S-LAN?

- With NVMe over Fabrics, should the fastest drives talk directly to the S-LAN?
  - Simplifies remote access
  - No conversion layers

- NO SAS or SATA



Should the Secondary Tier Drives be Ethernet?

- WD-Labs demo
  - 504 Ethernet drives
  - Ceph OSD per drive
- Fits SDS brilliantly
  - Auto-orchestrate
  - Services in servers

From Ceph Blog

Are There Alternative Fabrics?

- InfiniBand – Lower Latency, but higher price

- PCI-e – Maybe for small clusters
- Omnipath – May be faster for a while, but little infrastructure and high CPU overhead
- Fibre Channel – NVMe over RDMA over Fibre Channel?

- All miss out on the “One fabric=less complexity” benefit of Ethernet!

But What of File Systems?

There's one more step:

- Layered file systems are passé
- Alternatives:
  - Flat storage space
  - Key/data
  - Direct addressing