Enterprise Flash Storage Annual Update

Or how the data center is replacing spinning rust with solid state

Howard Marks
Chief Scientist
DeepStorage.net

Santa Clara, CA
August 2017

Your not so Humble Speaker

- 30+ years of consulting
  & writing for trade press
- Occasional blogger at TechTarget
- Chief Scientist DeepStorage, LLC.
  - Independent test lab and analyst firm
- Cohost Greybeards on Storage podcast
- @DeepStorageNet on Twitter
- Email:Hmarks@DeepStorage.Net

Agenda

- A quick update on flash and enterprise SSDs
- PCIe/NVMe rising
- NVMe over fabrics and the new tier 0
- Post Flash and persistent memories

Flash Has Won

- Over 85% of VNX/FAS have some flash
- Even SMB solutions are now flash driven
  - Nexsan's Unity hybrid only
- AFA market $1.7 billion Q4 2016
  - External storage down 6.7%
  - AFA up 61.2%
- Hybrids $2.5 billion (38% market share)

Source: IDC Worldwide Quarterly Enterprise Storage Systems Tracker

Evolution of Enterprise Flash

2010

- 100K+ IOPS
- Consistent sub-millsec latency
- Go fast for special cases

2012

- Still a point solution
- Becoming cost effective
- Limited data services
- Data reduction

2016

- Flash is mainstream
- Full data services & data reduction
- Cost effective for most applications
- New solutions for new applications

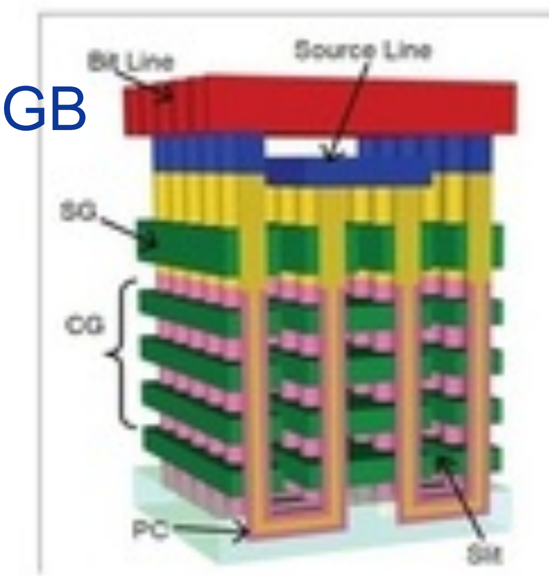
The All Flash Data Center?

- All flash is inevitable
- Facebook…
- Murphy’s law
- Growing our TAM

- Flash cheaper than disk, really?
  - No enterprise SSD 25X cost/ GB of 8TB disk
- Kryder’s law

Flash ~~Goes~~ Went 3D

- Smaller cells were denser, cheaper, crappier
  - Beyond 15nm untenable
- All 4 foundries
  - Samsung, Toshiba/WD 64L 512GB
  - Intel/Micron 32L (64L, 256GB)
  - Hynix 48L 256GB (72L, 256GB)
- 3D allows larger cells
  - Makes TLC useable
    - Faster write, higher endurance
  - QLC even



The Great Flash Shortage of 2016-7

- 2008-2015 SSD $/GB −30%/yr
- 2016
  - Fabs stumble on 3D conversion
  - Prices flat to +30% from the usual suspects
  - SSD lead times up to 120 days
  - Vendors shift from client to enterprise
- Note: DRAM prices also up (50% or more)
  - Fabs can switch back and forth (w/limits)
- Relief to come late 2018/19

Enterprise SSD Evolution

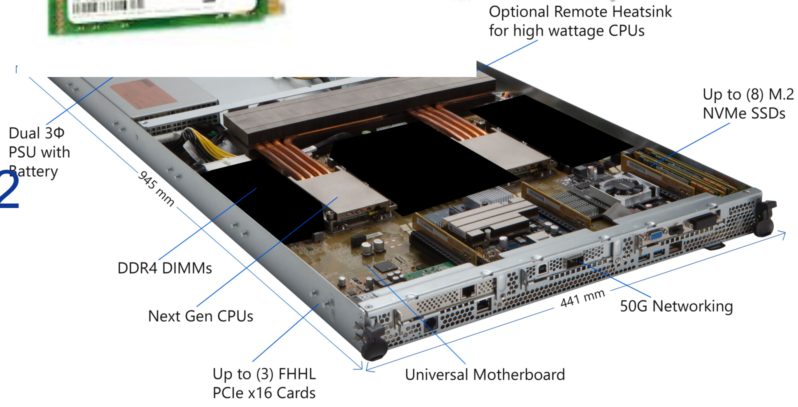
- Media
  - 3D TLC now standard
- Density - Today’s largest devices
  - SAS - 15.8TB SATA – 4TB
  - PCIe – 64TB AIC, 7.6TB U.2, 8TB M.2
- Interfaces
  - Last year U.2 was the big thing
  - Dual port U.2
  - Server support from most vendors

U.2/SFF-8639 PCIe for 2.5” SSDs

- Adds x4 PCIe 3.0 lanes to SAS/SATA connector
  - Dual ports to x2
- In servers from all the major players
  - Making PCIe/NVMe SSDs hot swappable
- Moving into storage arrays
  - Tegile
  - Pure FlashArray (in-house)

M.2 Goes Enterprise

- 4x PCIe or SATA channels
  - Same as U.2
- “Wrigley” form factor
- Server vendors replacing SD with M.2 for boot devices
- Plug in std slot w/low cost adapter
- Seagate’s 64TB SSD is 8x8TB M.2 SSDs and a PCIe switch chip
- Eight M.2 Slots on Microsoft Project Olympus server (Azure)



The Viking 50TB SAS SSD

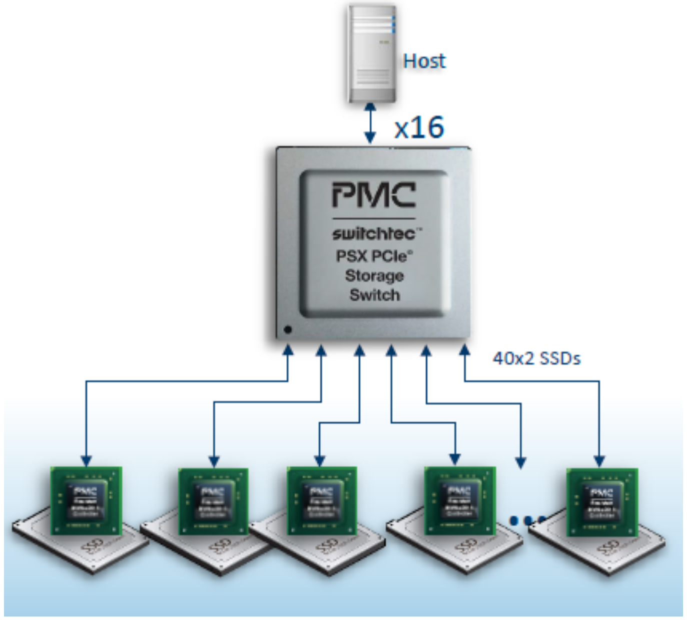
- 3.5 large form factor
- 6Gbps not 12Gbps SAS
- 1 Drive Write per Day endurance
  - But 1.7 days to fill at spec'd 500MB/s
- Different use model:
  - Hyperscaler's long tail
  - Long term retention

SDD Advances

- Field configurable SSDs
  - SSD has xGB flash
    - User chooses balance between useable capacity & endurance
  - Sophisticated users, like webscalers
- Host Managed SSDs
  - Give host system control of garbage collection
  - System can avoid writes to drive in GC
  - More consistent latency

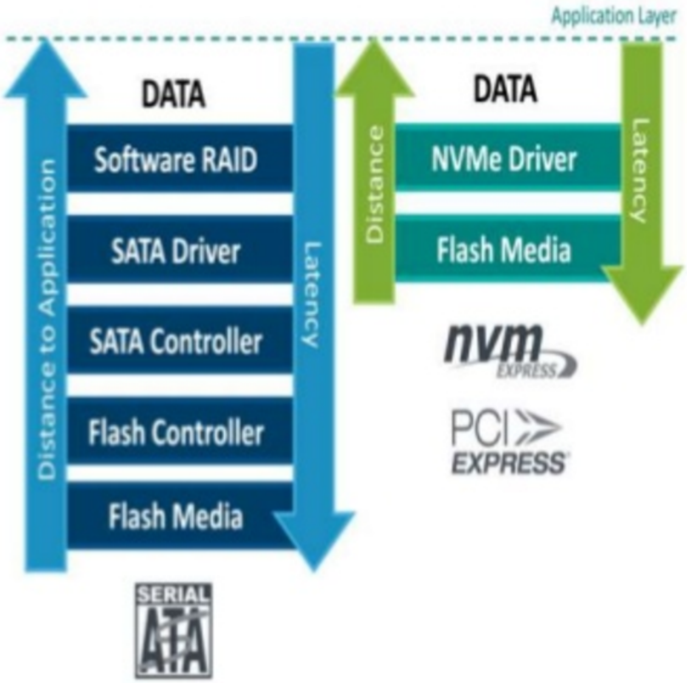
(Last year) The Future is PCIe

- PCIe offers:
  - Low latency, high bandwith, RDMA
- PCIe Switch chips
  - PLX and PMC – 96 lane
- Use for:
  - Controller to controller link
  - U.2 SSDs in storage system
  - ~~Rack scale switched system (DSSD)~~
  - ~~External PCI standards exist~~

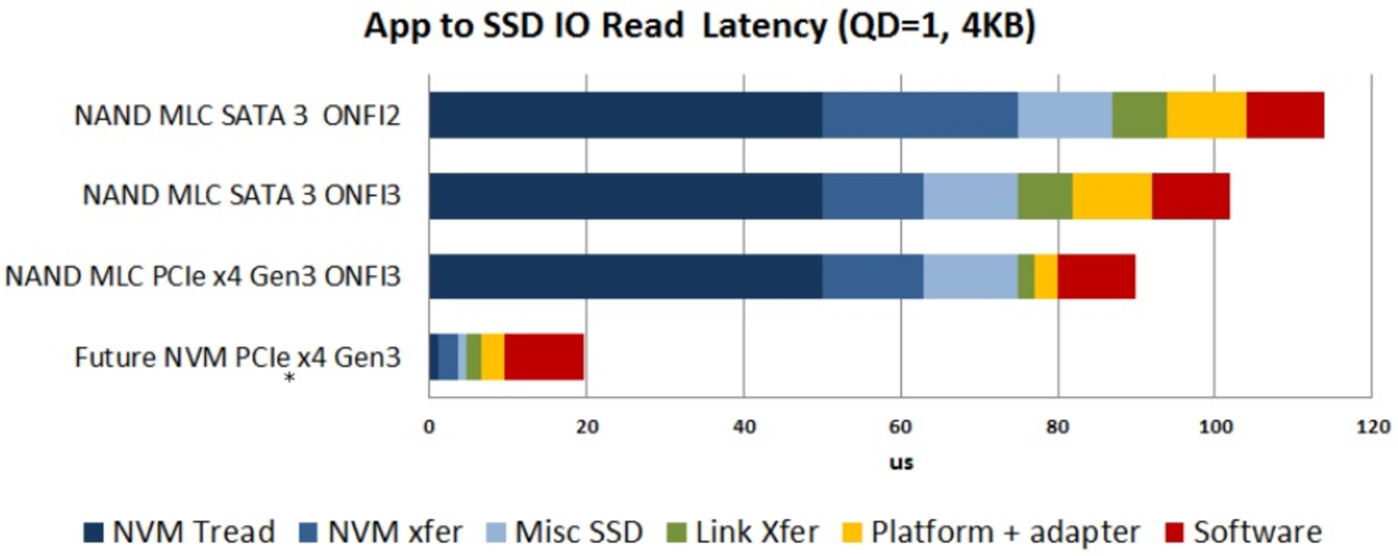


The Near Future is NVMe

- Gen1 and 2 PCI SSDs
  - ACHI (SATA command set)
  - Propreatary (Fusion-IO, Verident) with heavy software
- Enter NVM Express
  - A new software protocol for non-volatile memory access
- Lower compute overhead than SCSI
- 64K queues of 64K entries vs SCSI 1 queue of 32 entries



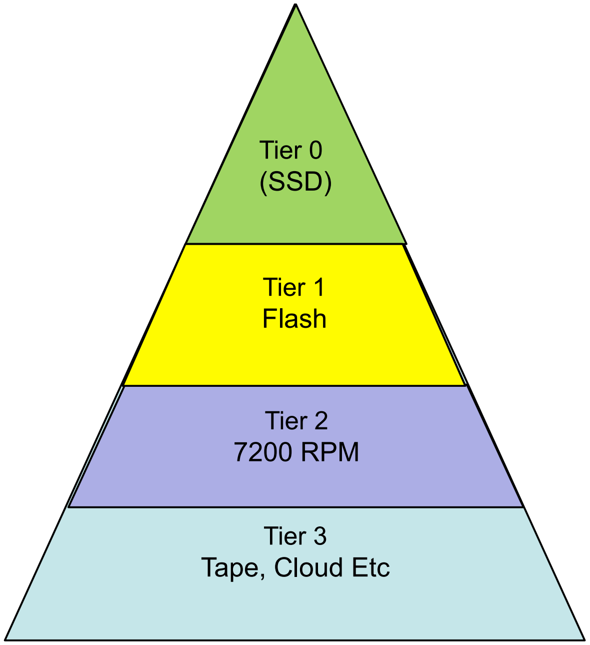
NVMe = Lower Overhead & Latency



- In 2016 NVMe is leading from desktop M.2 to the datacenter

A New Tier 0

- We've redefined tier 1
  - 100,000s of IOPS
  - Sub-millisecond latency
  - 100s TB useable capacity
  - Rich data services
  - Back to $/GB
- A new tier 0 emerges
  - 1,000,000s of IOPS
  - Latency under 100µsec
  - Application resiliency
- NVMe over network

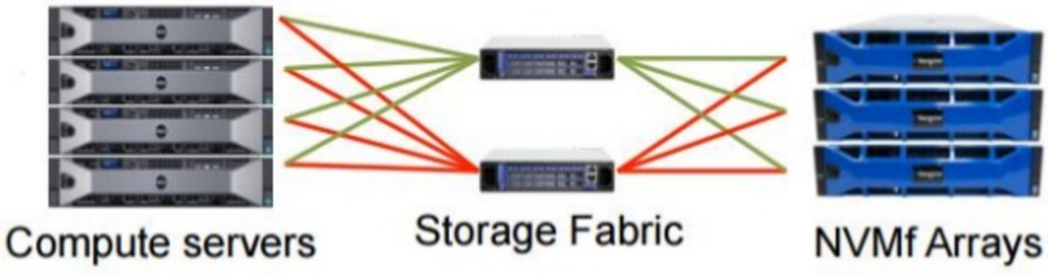

Tier 0
(SSD)
Tier 1
Flash
Tier 2
7200 RPM
Tier 3
Tape, Cloud Etc

New Tier 0 (2016)

- EMC DSSD
  - Rack scale (48 hosts) switched PCIe
  - Very custom hardware
  - Block, key-value, direct memory APIs
- Several NVMe over Network startups
  - Apeiron – 40Gbps Ethernet switch in JBOF
  - E8 – Dual controller array – basic services
  - Mangstor – x86 NVMEoF target
  - Excellero – Low CPU SDS, RDMA

NVMe Over Fabrics (NVMEoF)

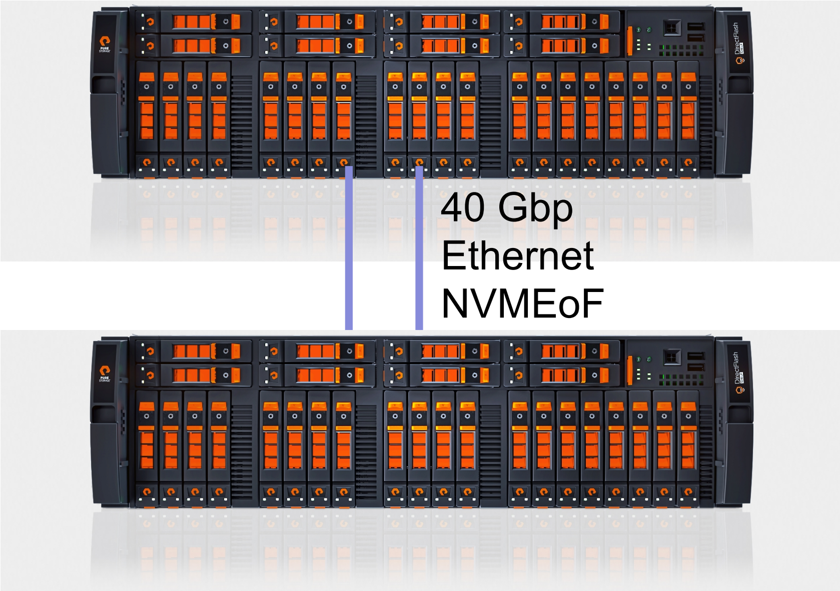
- Extends/encapsulates NVMe semantics over
  - Ethernet with RMDA
    - ROCE – RDMA over Converged Ethernet
    - iWARP – RDMA over TCP
  - Fibre Channel
  - Infiniband (no products yet announced)
- Adds name spaces and discovery
- 10-50µsec protocol and network overhead



The New Tier 0 (2017)

- NVMe over Fabrics standard announced at FMS
  - Drivers now in all major OSes/Hypervisors
  - Intel SPDK high performance requestor and target
- DellEMC cancels DSSD
  - Promises tech will live on in other products

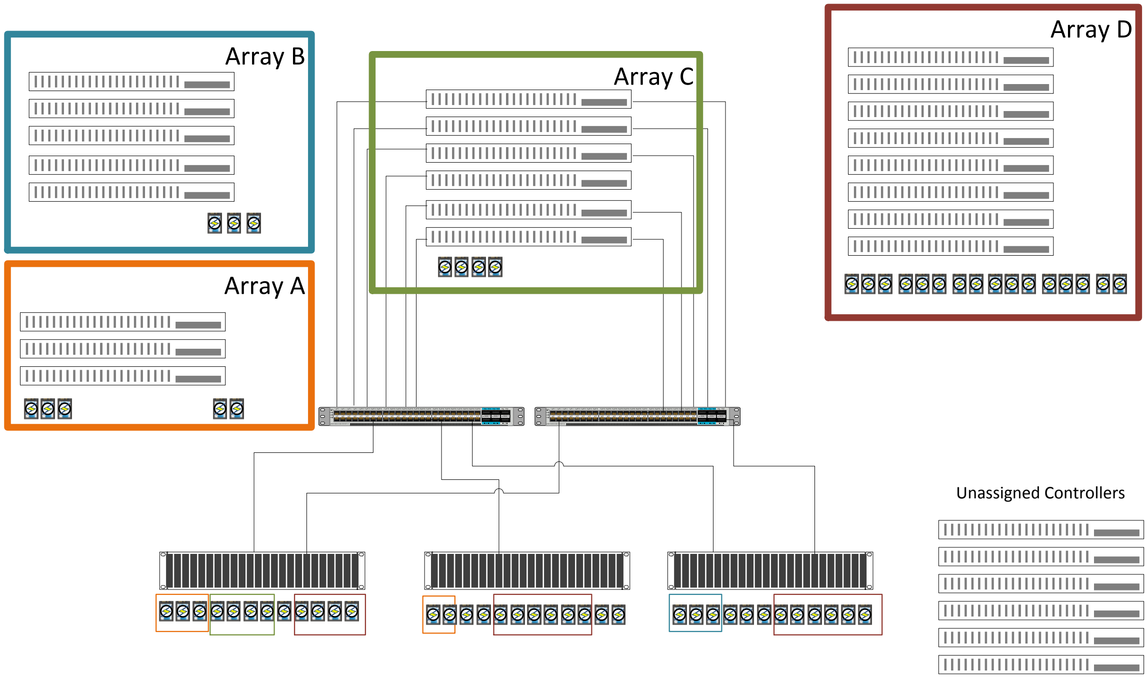
Pure FlashArray//x



40 Gbp
Ethernet
NVMEoF

- Replaces //m SAS SSDs with NVMe flashmodules
- Expansion via SAS or NVMEoF JBOF
- NVMEoF target on 40Gbps Ethernet
- Full services

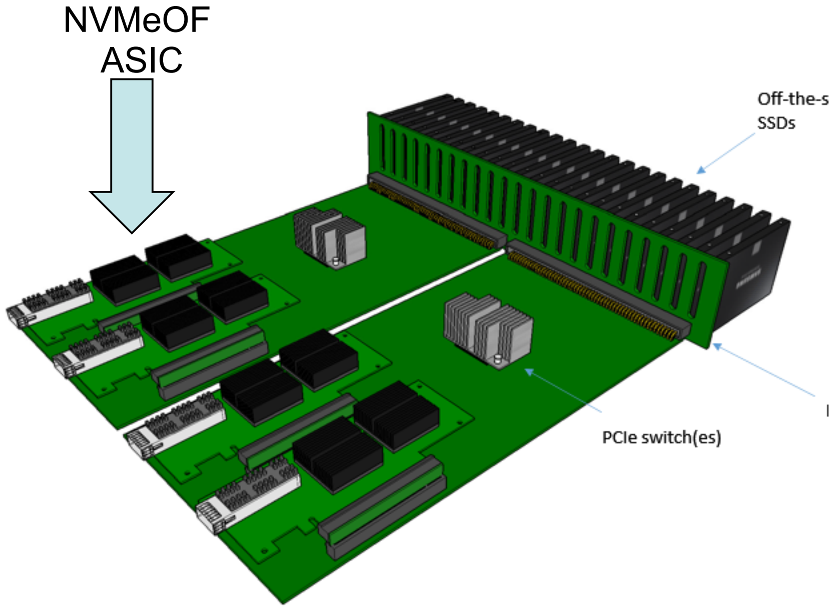
Kaminario K2 Composeable



- NVMEoF
  - Controller to JBOF
  - Host to array (opt)
- Dynamically assign controllers and flash to virt array

NVMe JBOFs Emerge

- Today’s JBOFs are x86 servers
  - Dual servers needed for HA
  - High flexibility
  - High cost
- NVMEoF ASICs
  - Vastly reduce costs
  - Sampling from
    - SolarFlare Xilinx
    - Kazan Networks



NVMeOF
ASIC

Or The Future is Persistent Memory

- Scaleable Xeon servers come with NV-DIMM support
  - Good for software delivered storage
  - Small (8GB)
- Large persistent memory the next big thing
- Today’s In-Memory database must log writes
  - Fast storage still required
- Tomorrow RDMA into another node’s NVmem

Diablo Puts Flash on the Memory Bus

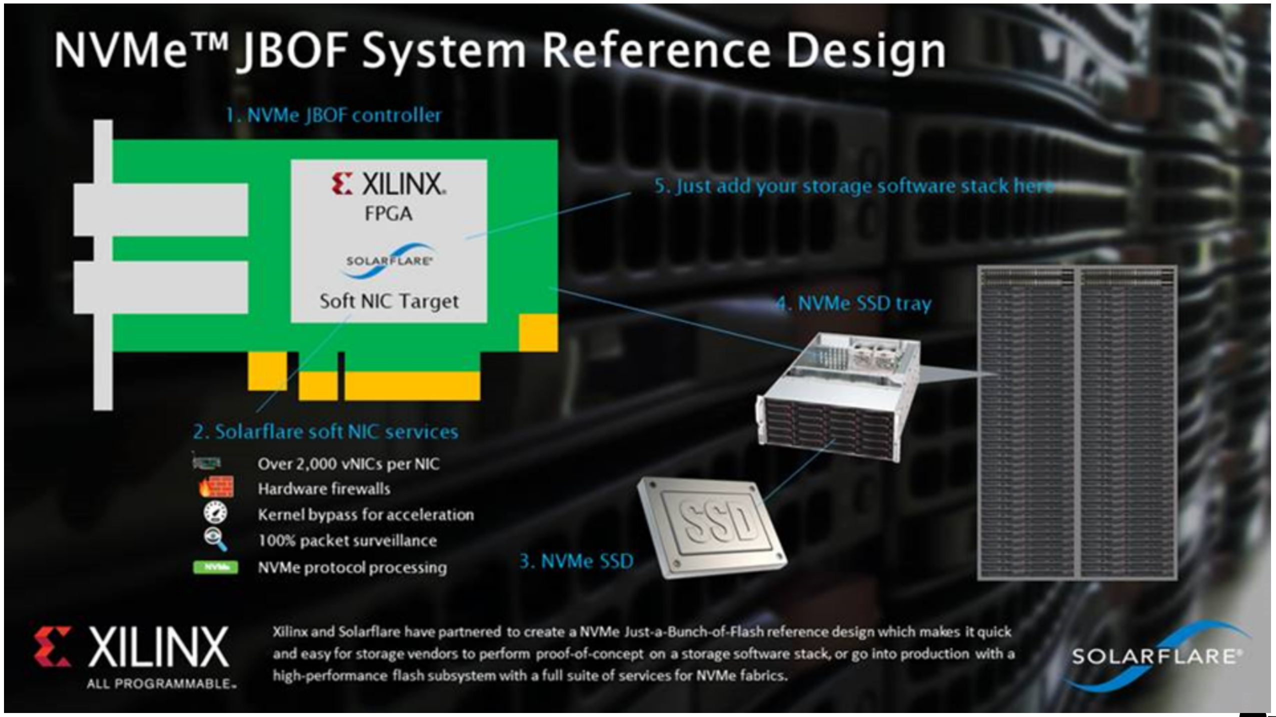
- Memory Channel Flash
  (SanDisk UltraDIMM)
  - Block storage or direct memory
  - Write latency as low as 3µsec
  - Requires BIOS support
- Memory1
  - 400GB/DIMM
  - No BIOS/OS Support
  - Volatile

The Future

- All ~~PCIe~~ NVMe storage systems
  - As conventional storage
  - With memory interfaces
- Next-gen memory (PCM, 3d Xpoint, Etc)
  - First as write cache in SSD
  - Later as memory
  - Taking a bit longer than expected
- More persistent memory as memory
  - Needs application support ala SAP Hana

Q
A

Flash Memory SUMMIT



NVMe™ JBOF System Reference Design

- Over 2,000 vNICs per NIC
- Hardware firewalls
- Kernel bypass for acceleration
- 100% packet surveillance
- NVMe protocol processing

Xilinx and Solarflare have partnered to create a NVMe Just-a-Bunch-of-Flash reference design which makes it quick and easy for storage vendors to perform proof-of-concept on a storage software stack, or go into production with a high-performance flash subsystem with a full suite of services for NVMe fabrics.