Storage Interfaces

Introduction

Tom Gardner

IEEE Silicon Valley Technology History Committee

- Part of IEEE Santa Clara Valley Section
- Webinars on history of Silicon Valley history

www.SiliconValleyHistory.com

Storage Interfaces Agenda

Speaker	Subject
Tom Gardner	Introduction & History
Grant Saviers	Evolution of Storage Interfaces
Jai Menon	Evolution of Block Storage System Interfaces
Amber Huffman	Fast & Efficient Interfaces for the SSD Era
Questions & Answers	

Presentations are in two files
Gardner and Saviers: Storage Interfaces 20210511 Part 1.pdf
Menon and Huffman: Storage Interfaces 20210511 Part 2.pdf

Jai Menon

Evolution of Block Storage System Interfaces

Evolution of Block Storage System Interfaces

(From mainframes to industry-standard servers)

Jai Menon

5/11/21

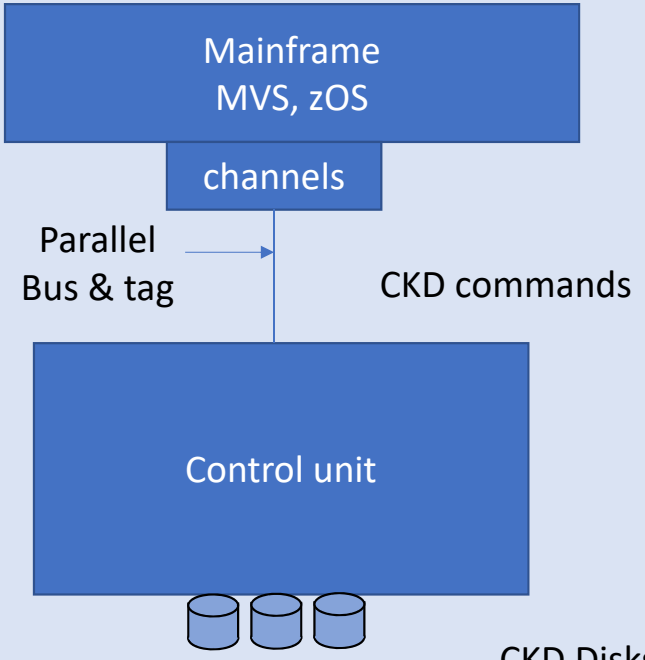
Talk Outline

- Mainframe block storage system interfaces
  - CKD, ECKD
  - CKD-Emulation on FB disks
- Block Storage systems interfaces for industry-standard servers
  - FC, iSCSI, NVMeOF

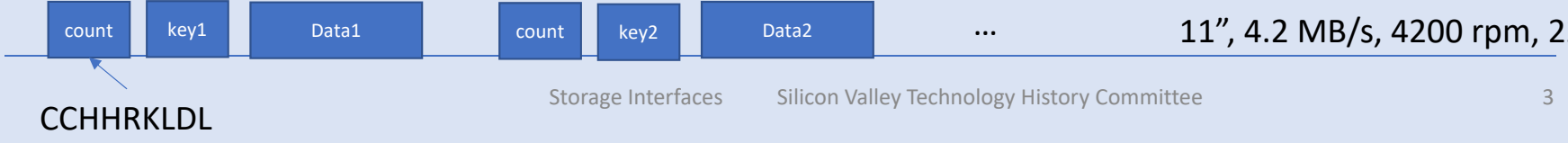
No coverage of file or object interfaces

Mainframes – CKD

CKD refers to both disk format and commands (CCWs)



Seek
Set sector
Search Key Equal
Tic *-8
Read Data



Introduced in 1964 with S/360

Synchronous operation

Gaps between fields must be large enough to allow channel to read, match, and turn around to issue next command

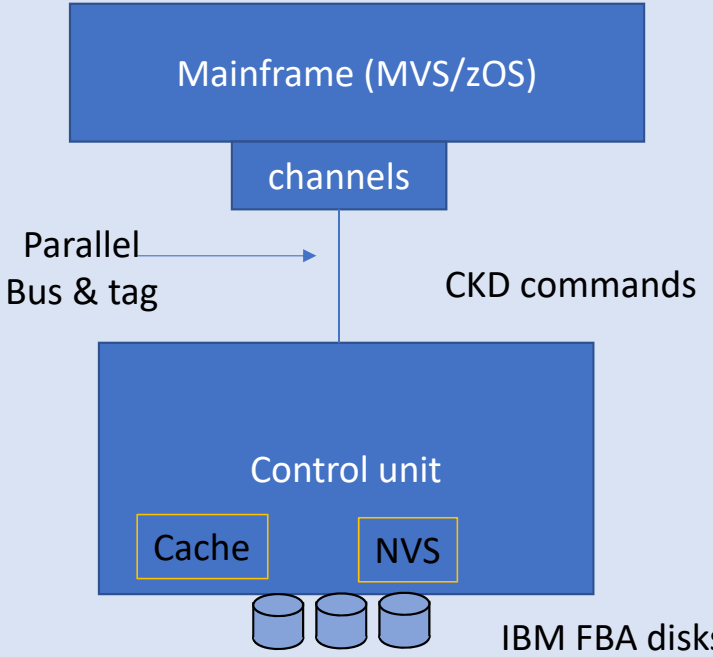
As disk track density increased, more and more space wasted in gaps

Each CKD disk cost IBM approx. $100M to develop

Last CDK disk 3390-9 in 1993
11”, 4.2 MB/s, 4200 rpm, 22.7 GB

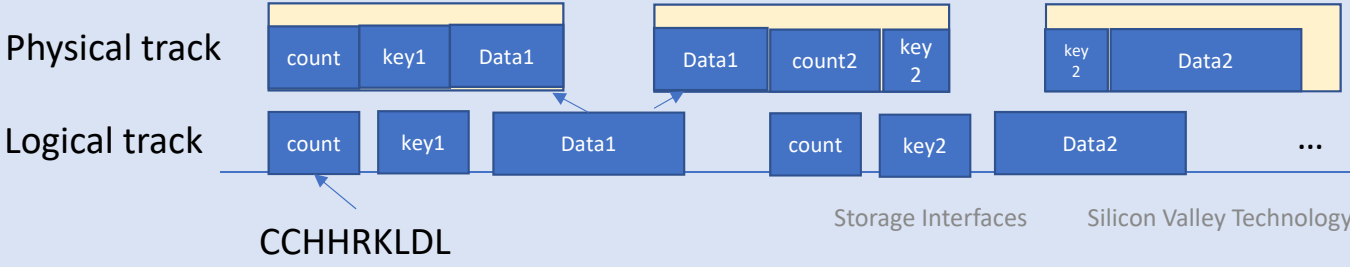
Mainframes – CKD-Emulation

Possible because processors and memory became cheaper



Seek
Set sector
Search Key
Equal
Tic *-8
Read Data

IBM FBA disks first
Industry standard FB disks later



Key challenges:

- Performance
  - Preserving sector positioning
  - RMW for CKD writes (PTT)
- Reliability
  - Data fields span FB sectors

Emulation Approaches:

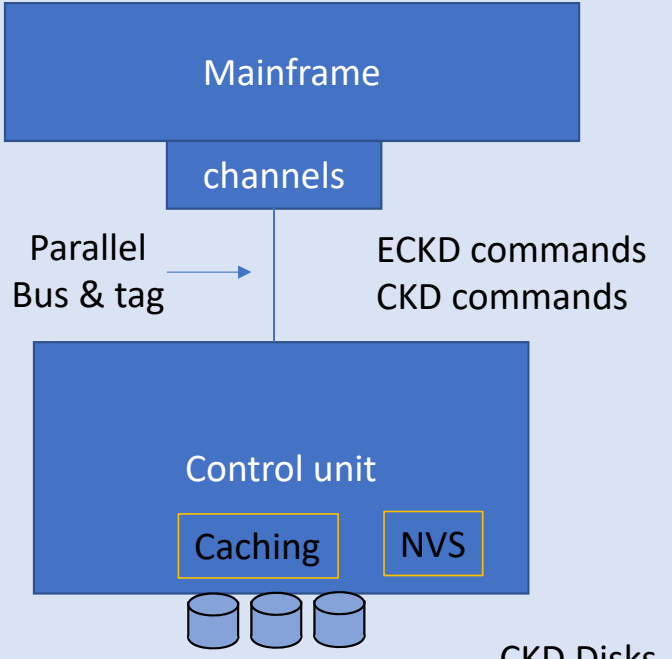
- Channel to CU interface
- CU to disk interface

History:

- First done on 9370 (1987)
- EMC (1990)
- IBM RAMAC Array (1994)
  - Emulates 3390 disk
- Still done to this day (DS8900)

Mainframes – ECKD

processors and memory became cheaper



Define Extent
Locate
Read Data

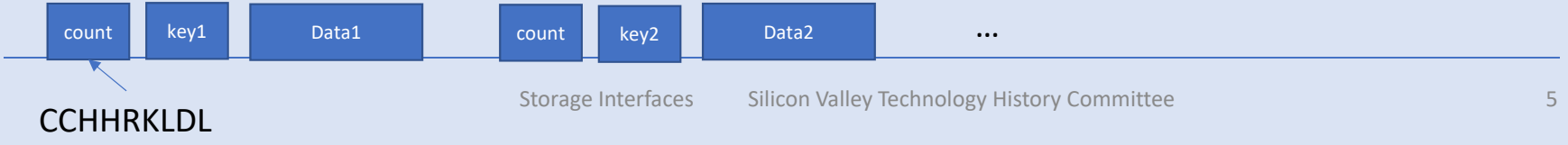
Introduced in 1985 with IBM 3880 Control Unit

Channel and CU/disk operate asynchronously

Longer distance possible from mainframe to disks

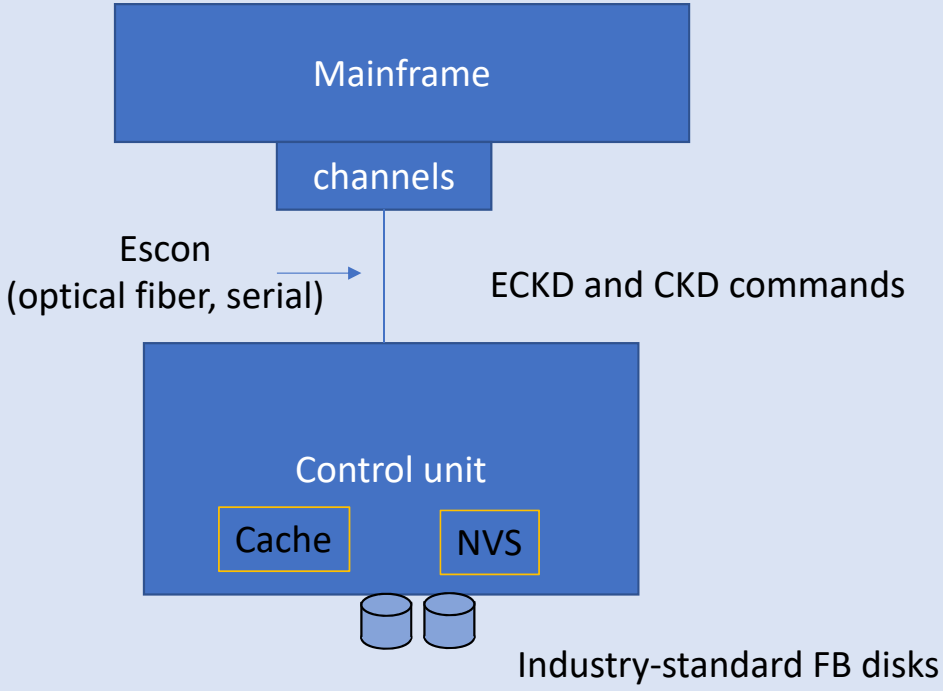
Better performance on hits (with caching)

Some access methods still used CKD



Mainframes – Escon

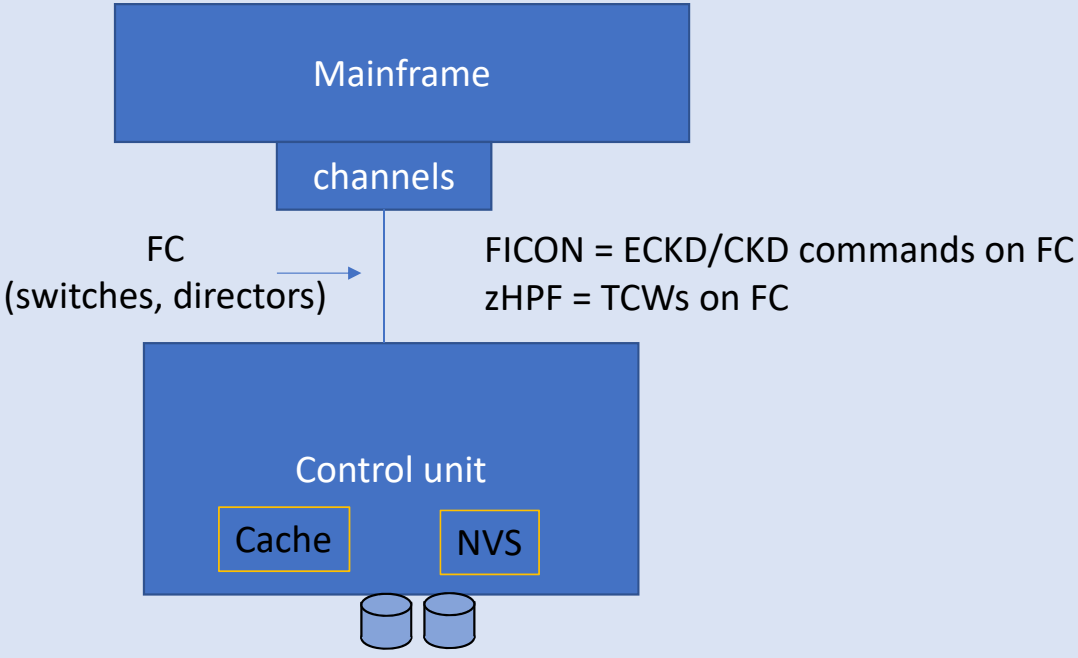
Introduced in 1990 as part of S/390



Operated originally at 10 MB/sec
Allowed greater distance
Allowed switching
Single CU to 8 mainframes

Mainframes – FICON, zHPF

FICON Introduced in 1998 as part of 5th gen S/390

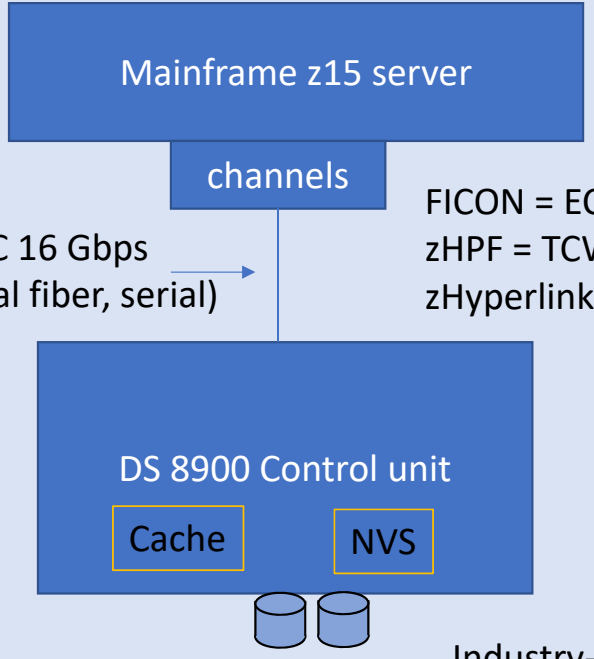


Industry-standard FB disks

FICON
- 1, 2, 4, 8, 16 Gbps
- FC switches and Directors
- Up to 100 kms

zHPF (high Performance FICON)
- Introduced in 2009
- Use FCP approach to transfer cmds, data & status
- Optional feature
- TCWs (Transfer Control Words) instead of CCWs
- 90% typical perf improvement
- Fewer channels needed

Mainframes today – DS8900 controller



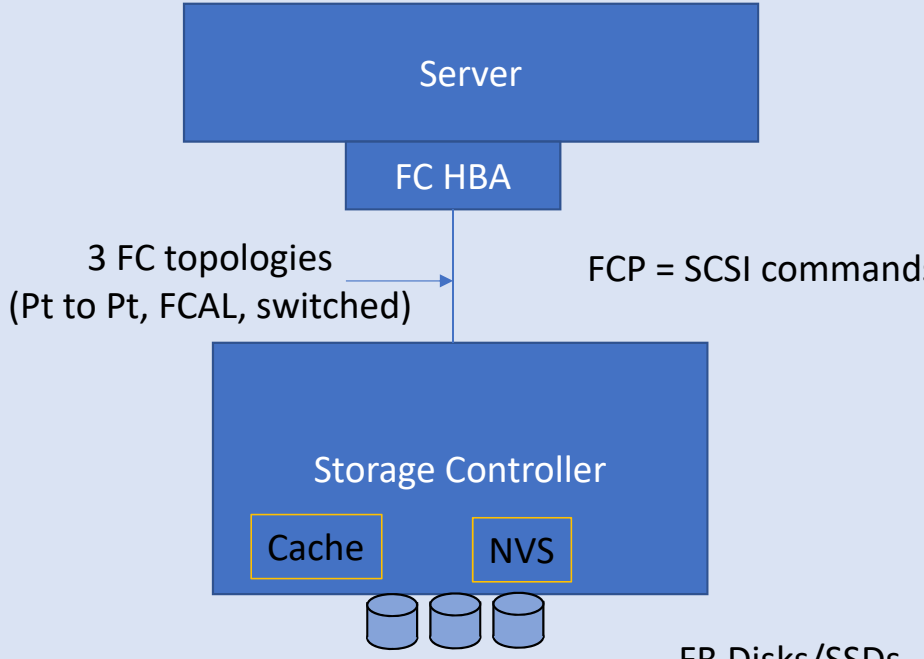
FICON = ECKD/CKD commands on FC
zHPF = TCWs on FC
zHyperlink (PCIe connection)

Industry-standard FB disks/SSDs

Still does CKD-Emulation
Emulates 3390 and 3380 CKD disks
Supports FICON and zHPF
zHyperlink announced in 2017 with z14 & DS8880

- Upto 32 ports per z14
- Synchronous I/O, point to point
- PCIe based
- 18 usecs response time with zHyperlink
  point to point <150m

Industry-standard servers – SCSI over FC

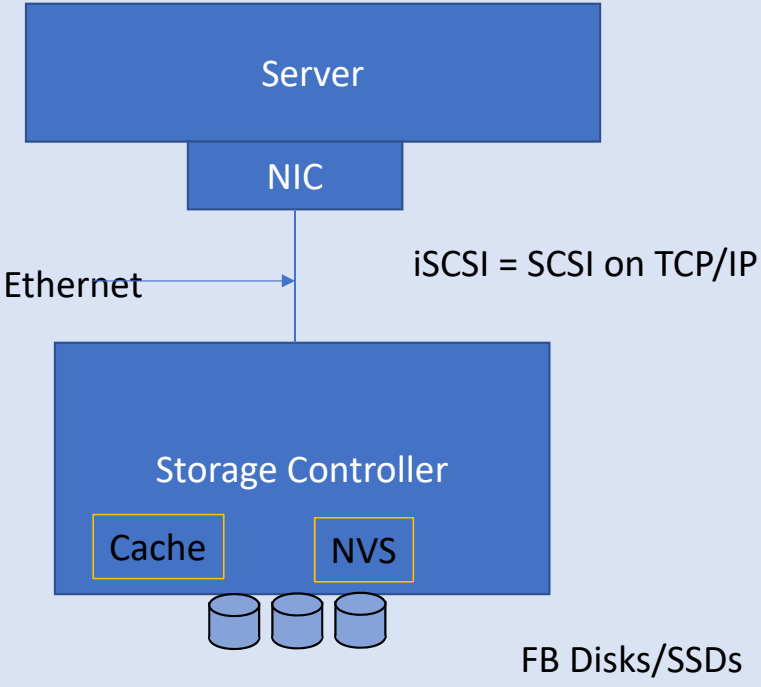


FCP = SCSI commands on FC (ANSI standard approval in 1994)

1st FCP controller - the Sun SPARCStorageArray (March 1994)

- FC point-point to the Sun server and SCSI disks
- 31.5 GB RAID Array, RAID 0, 1, 0+1, 5
- 25 MB/s interface
- 2000 iops
- $1.62/MB

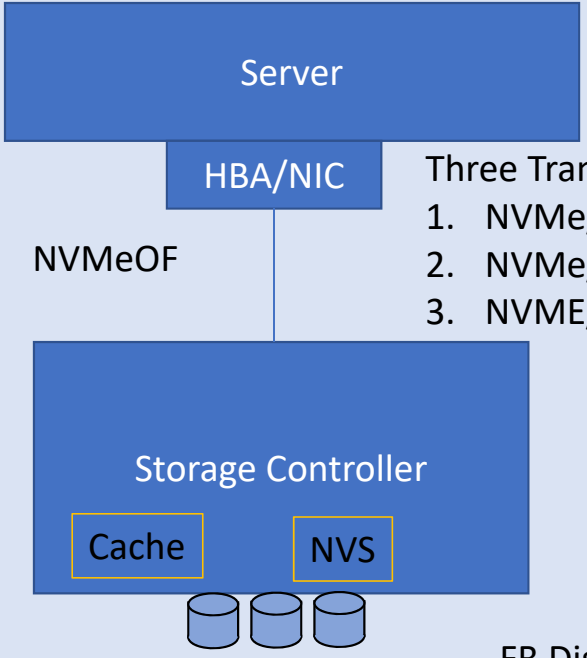
Industry-standard servers – iSCSI = SCSI on TCP/IP on Ethernet



- 1st draft IBM/Cisco 2000; ratified in 2003
- Lower cost than FC
- Longer distance than FC
- Runs on existing network infrastructure
- IBM Total Storage IP Storage 200i, which shipped in 2001, is the earliest iSCSI product in the industry.

Industry-standard servers – NVMe over Fabrics (NVMeoF) Focus on All Flash Storage Systems

(Extends NVMe over PCIe to other transports)



Three Transports

1. NVMe/TCP
2. NVMe/FC (FC-NVMe)
3. NVME/RDMA (IB, RoCE, iWarp)

- Designed to support SSD storage
- Higher performance (64K queues, 64K cmds per queue)
- Lower complexity

Examples

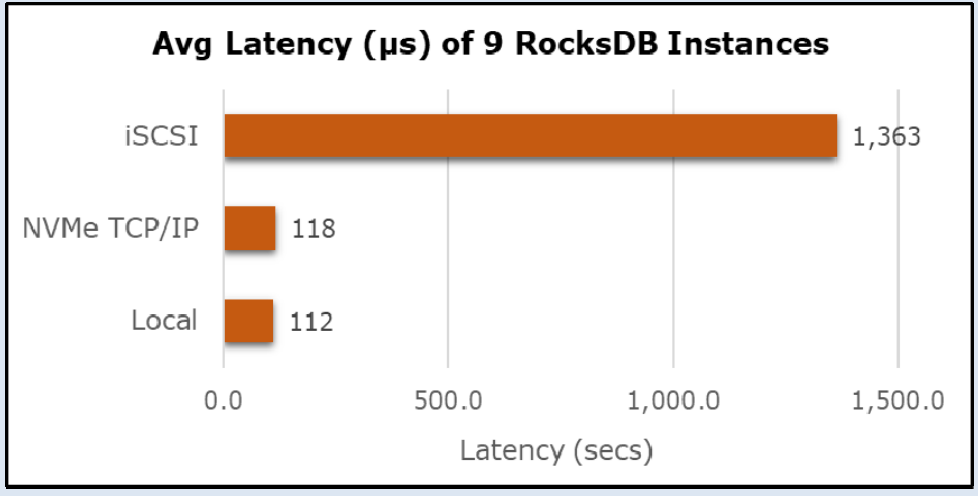
- Pure Storage (NVMe/RoCE) – 2019
- NVMe/TCP
  - Fungible Storage cluster (NVMe/TCP) - 2020
    - 2U, 24 SSDs
    - 15M IOPS in 2U, 120 usecs latency
    - 8x100 Gbps ports
  - Lightbits Labs - 2020

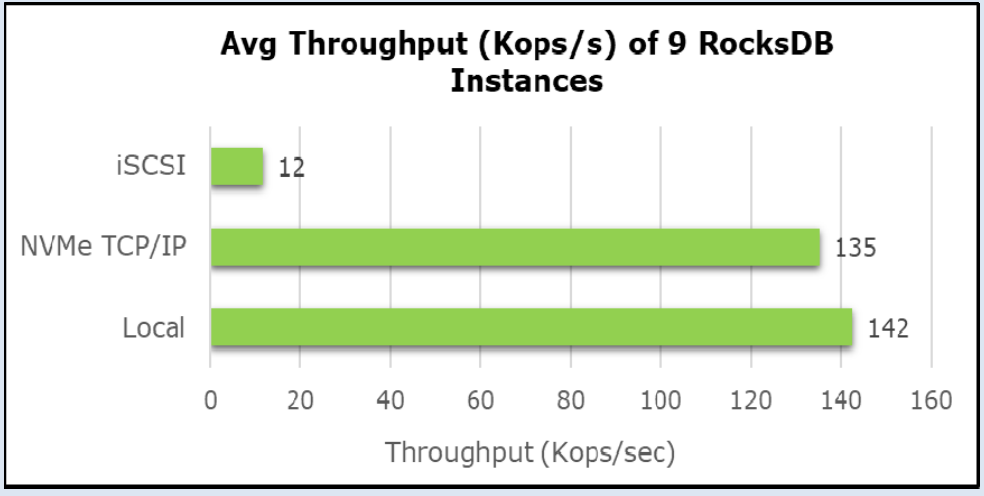
Performance Comparison: Random Read

STORAGE

iSCSI and NVMe/TCP AFA targets built from Intel Gold Servers

(Manoj Wadekar, FB and Anjaneya Chagam, Intel - 2019)





Minimal performance overhead with NVMe over TCP/IP

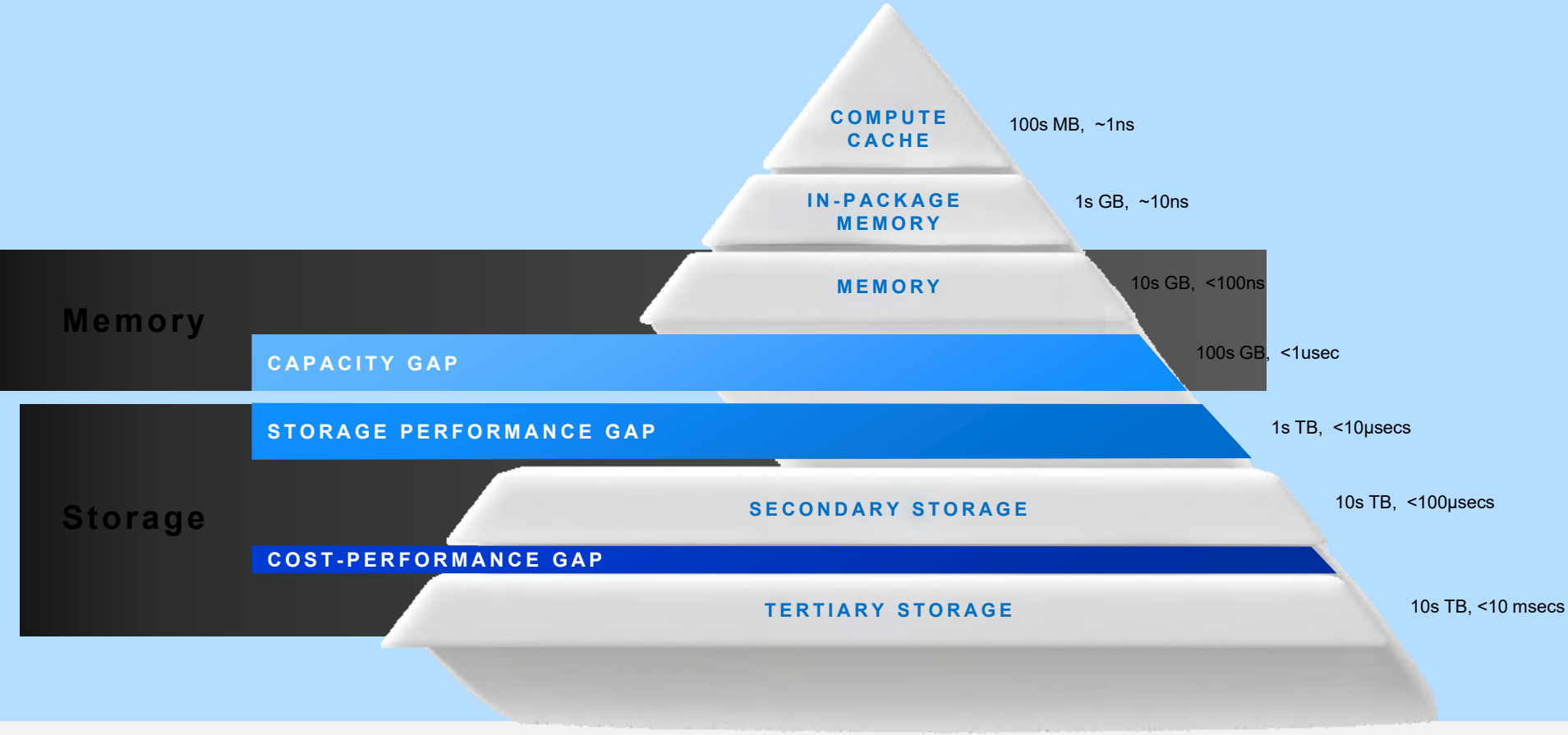
The Next Speaker

- Amber Huffman
- Fast and efficient interfaces for the SSD Era

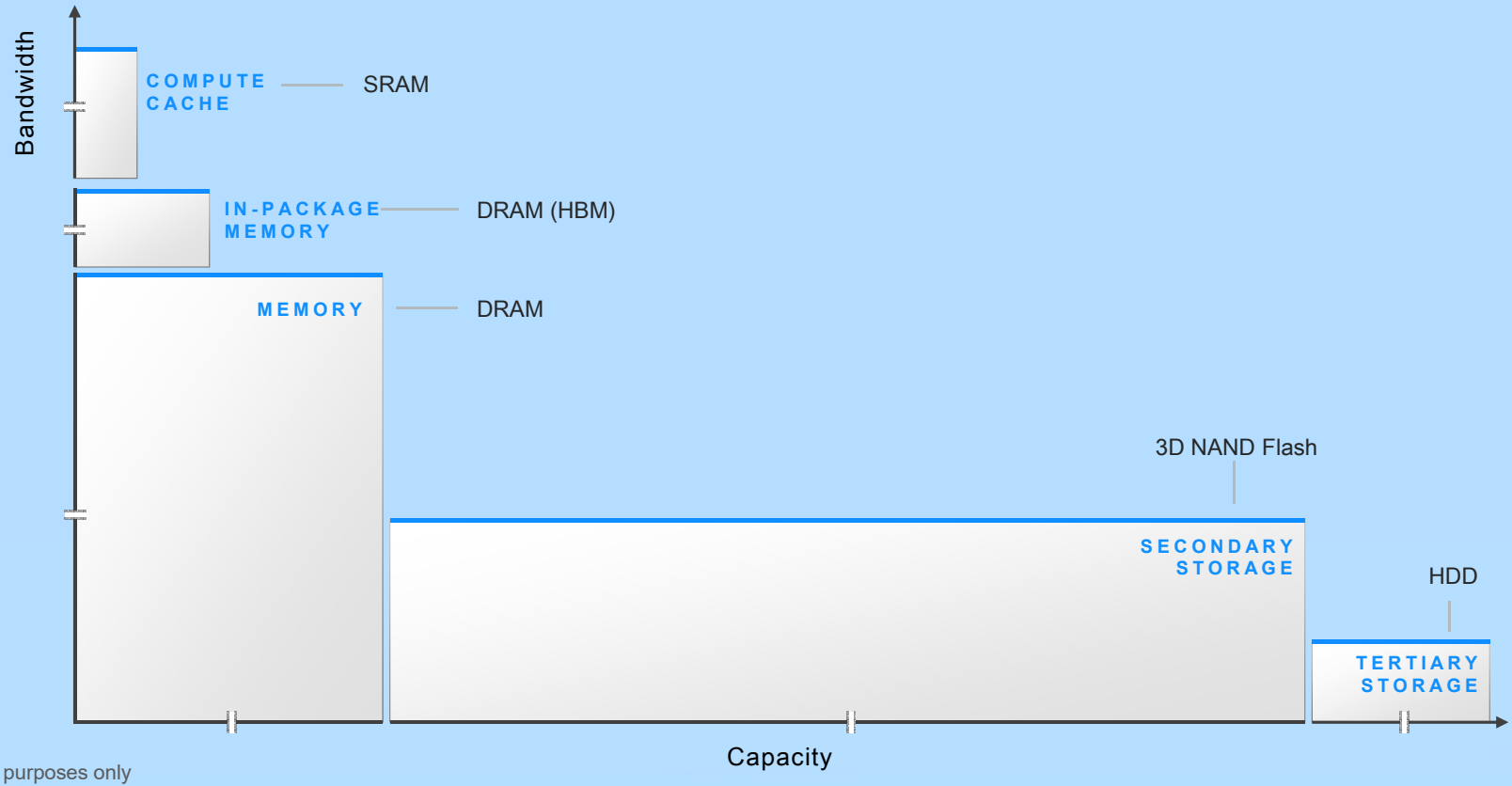
NVMe® Technology
Fast & Efficient Interfaces for the SSD Era

Amber Huffman
Fellow & Chief Technologist of IP Engineering Group, Intel Corporation
President, NVM Express, Inc.

Memory and Storage Hierarchy Gaps

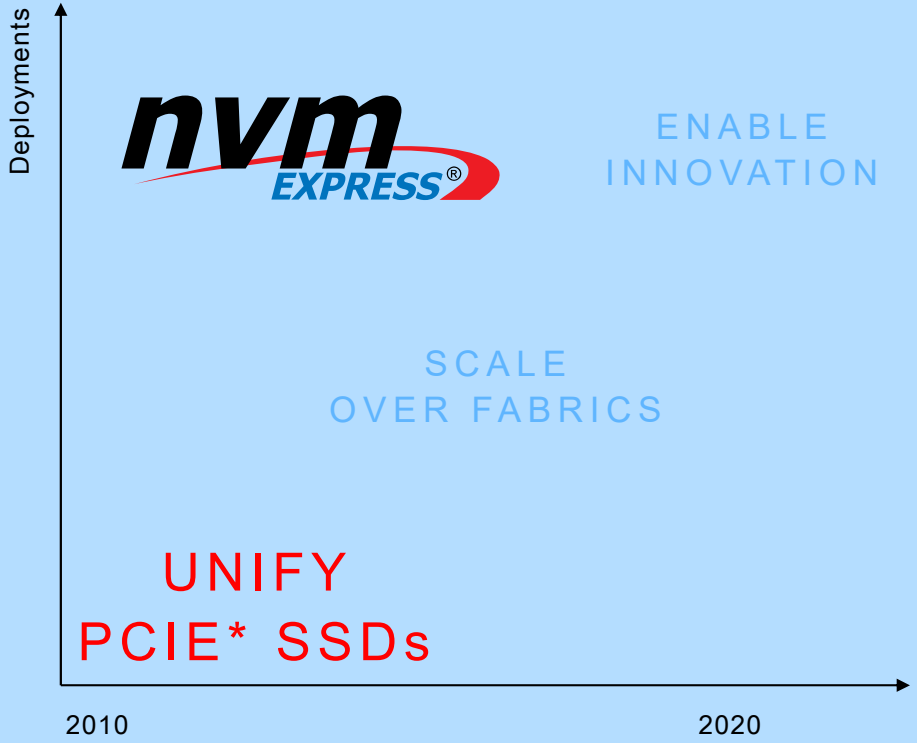


Memory and Storage Hierarchy Gaps



For illustrative purposes only

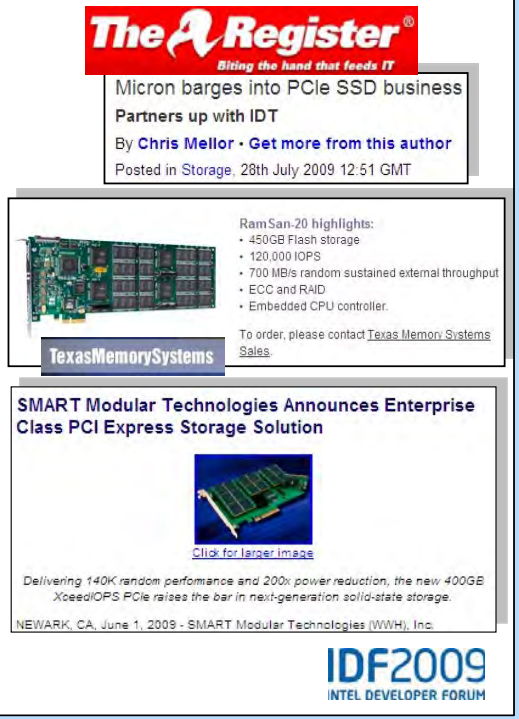
The Evolution of NVMe® Technology



Framing the Need
Why Standard Interface for PCIe* SSDs

Enabling Broad Adoption of PCIe SSDs: Enterprise NVMHCI

- PCIe SSDs are attractive and popping out of the woodwork
  - Eliminates SAS infrastructure, plenty of PCIe lanes, bandwidth can be concentrated, lower latency
- PCIe SSDs **lack** standard OS support & driver infrastructure, since there is no standard host controller interface
- Impact of no standard infrastructure:
  - Requires SSDs vendor to provide drivers for every OS
  - OEM features, like error logs, are implemented in an inconsistent fashion
  - OEMs have to validate multiple drivers
- Enable broad adoption by extending NVMHCI (Non-Volatile Memory Host Controller Interface) to Enterprise
  - Leverage the client NVMHCI i/f, software infrastructure, and Workgroup to fill gap quickly with streamlined solution
  - http://www.intel.com/standards/nvmhci



Building a Coalition

Pull Together Key Stakeholders to Drive Change

Deliver the Baseline Specification
And Select a Much Better Name!

NVM Express* Overview

- NVM Express is a scalable host controller interface designed for Enterprise and Client systems that use PCI Express* SSDs
  - Includes optimized register interface and command set
- NVMe was developed by industry consortium of 80+ members and is directed by a 10 company Promoter Group

CISCO, DELL, EMC², NetApp, ORACLE, STEC, IDT, Micron, intel, SandForce

- NVMe 1.0 published on March 1st, available at *nvmexpress.org*

NVMe standardization effort is complete & stable
Product introductions coming in 2012

IDF2011 INTEL DEVELOPER FORUM

9

Paint the Future
Why It's Worth Crossing the Adoption Chasm

NVMe*: Efficient SSD Performance

	AHCI[1]	nvm express
Uncacheable Register Reads Each consumes 2000 CPU cycles	4 per command 8000 cycles, ~ 2.5 µs	**0** per command
MSI-X and Interrupt Steering Ensures one core not IOPs bottleneck	No	Yes
Parallelism & Multiple Threads Ensures one core not IOPs bottleneck	Requires synchronization lock to issue command	No locking, doorbell register per Queue
Maximum Queue Depth Ensures one core not IOPs bottleneck	32	64K Queues 64K Commands per Q
Efficiency for 4KB Commands 4KB critical in Client and Enterprise	Command parameters require two serialized host DRAM fetches	Command parameters in one 64B fetch

NVMe designed for high parallelism and low latency

10 [1]AHCI: Serial ATA programming interface. See http://www.intel.com/technology/serialata/ahci.htm

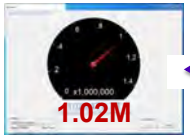
IDF2011 INTEL DEVELOPER FORUM

Scalability for Future NVM

- NVMe* is defined to scale for future NVM
  - Host controller standards live for 10+ years
  - Future NVM may have sub microsecond latencies
- 1M IOPS needs highly efficient driver approach
  - Benefits from removing OS queues, IO scheduler, and SCSI layer while optimizing for NVMe
- Block layer attach reduces overhead > 50%
  - Block layer: 2.8 µs, 9100 cycles
  - Traditional: 6.0 µs, 19500 cycles

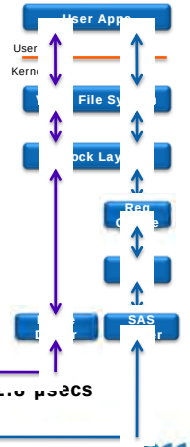
Chatham NVMe Prototype

Prototype Measured IOPS



Cores Used for 1M IOPs

Linux * Storage Stack



15 Measurement taken on Intel® Core™ i5-2500K 3.3GHz 6MB L3 Cache Quad-Core Desktop Processor using Linux RedHat EL6.0 2.6.32-71 Kernel

IDF2011 INTEL DEVELOPER FORUM

Provide Physical Infrastructure
Optimized Form Factors for SSDs

SSD Form Factor Working Group
Driving Enterprise SSD Infrastructure

Form Factor

Benefit from current 2.5" HDD form factor

Expand power envelope

Connector

Support Multiple Protocols
- PCI Express* 3.0, SAS 3.0, SATA 3.0

Management Bus

Dual port (PCIe)

Multi-lane capability (PCIe/SAS)

SAS & SATA Backward Compatibility

Hot-Plug

Hot-Plug Connector

Identify desired drive behavior

Define required system behavior

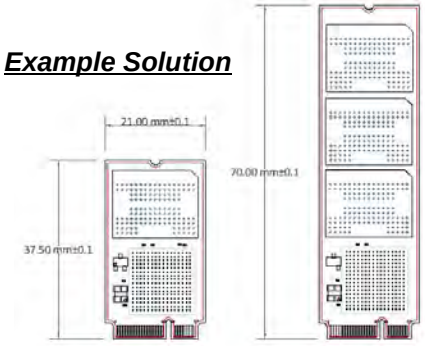
IDF2011 INTEL DEVELOPER FORUM

23

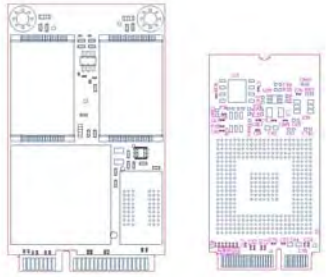
An Optimized Caseless Form Factor

- Attributes to strive for in a new standardized caseless SSD FF:
  - Scalable from small to large capacity points
  - Support SATA Gen3 and two lanes of PCI Express* Gen3
  - Optimize for Z height (e.g. board edge connector, reduce PCB thickness)
  - Mounting strategy will limit board area and reduce fasteners
  - Optimize board size based on BGA NAND package & ensure efficient tiling

Example Solution



mSATA vs Example



Watch for new standard caseless form factor effort

IDF2011 INTEL DEVELOPER FORUM

29

Enabling the Ecosystem
Drivers & BIOS Enabling Required

Reference Drivers for Key OSs

- Linux*
  - Already accepted into the mainline kernel on kernel.org
  - Open source with GPL license
  - Refer to http://git.infradead.org/users/willy/linux-nvme.git
- Windows*
  - Baseline developed in collaboration by IDT*, Intel, and LSI*
  - Open source with BSD license
  - Maintenance is collaboration by NVMe WG and Open Fabrics Alliance
  - Refer to https://www.openfabrics.org/resources/developer-tools/nvme-windows-development.html
- VMware*
  - Initial driver developed by Intel
  - Based on VMware advice, “vmk linux” driver based on Linux version
  - NVMe WG will collaborate with VMware on delivery/maintenance

IDF2012 INTEL DEVELOPER FORUM

23 NVMe = NVM Express

Reference Drivers for Key OSs (cont.)

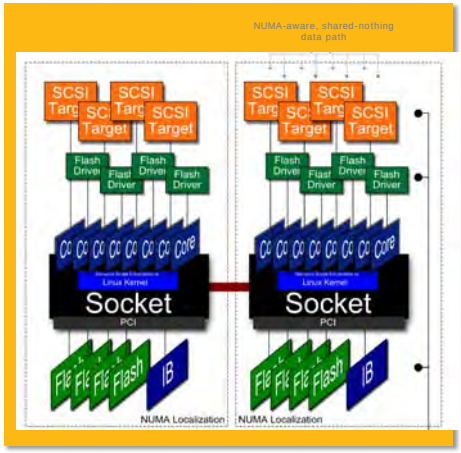
- Solaris*
  - There is a working driver prototype
  - Planned features include:
    - Fully implement and conform to 1.0c spec
    - Efficient block interfaces bypassing complex SCSI code path
    - NUMA optimized queue/interrupt allocation
    - Reliable with error detect and recovery fitting into Solaris* FMA
    - Build ZFS with multiple sector sizes (512B, 1KB, 2KB, 4KB) on namespaces
    - Fit into all Solaris disk utilities and fwflash(1M) for firmware
    - Boot & install on SPARC and X86
    - Surprise removal support
  - Plan to validate against Oracle* SSD partners
  - Plan to integration into S12 and a future S11 Update Release
- UEFI
  - The driver is under development
  - Plan to open source the driver in Q1 ‘13, including bug/patch process
  - Beta quality in Q1’13, production quality Q2’13

IDF2012 INTEL DEVELOPER FORUM

24

Industry Thought Leaders
Show What's Possible…

Designing with the Right Pieces



multi-core parallelism
+ PCI Express* Gen3 bandwidth
+ NUMA-aware software
+ NVMe flash
= millions of IOPs...

...at microsecond latencies

new math for storage platforms

40 NVMe = NVM Express

Proof Points in EMC* Products

Project Thunder: A networked non-volatile memory appliance

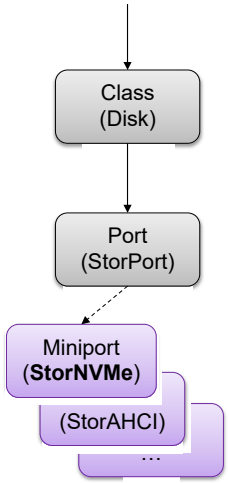
- Building block design center:
  - 10-20TB of flash capacity
  - Consistent 2.5M IOP throughput @ 150us
- NVMe ready:
  - Today: improved latency, reduced processor overhead
  - Tomorrow: ready for nano-second class NVM
  - On the showcase floor until **2pm**

41 NVMe = NVM Express

Industry Thought Leaders
Inbox Driver Support by Microsoft

Microsoft's Support of NVM Express

- The Natural Progression from SATA for NVM
  - Standardized PCI Express* Storage
  - First devices are enterprise-class
  - High-Density / High-Performance
  - Closing the latency gap with RAM
- Windows* Inbox Driver (StorNVMe.sys)
  - Windows Server 2012 R2 (enterprise)
  - Windows 8.1 (client)
  - Stable Base Driver
- The Storport Model
  - Reduced development cost
    - Offloads Basics: PnP, Power, Setup, Crash, Boot*
  - Mature / Hardened Model
  - Storport optimized for performance
  - RAM-backed NVMe device
    - > 1 million IOPS with < 20µs latencies



IDF13

38

StorNVMe Delivers a Great Solution

- StorNVMe Implementation Highlights
  - Uses hardened Enterprise Storage Stack
  - Strives for 1:1 mapping of queues to processors
  - NUMA optimized
  - Asynchronous not
  - Interrupt coalesci
  - Rigorous testing
  - Firmware Update/
- With great IOPs
- And low latency

System Configuration: 2 Socket Romley-based Server, 16 physical processors (32), Random Read Workload, Tool: IOmeter

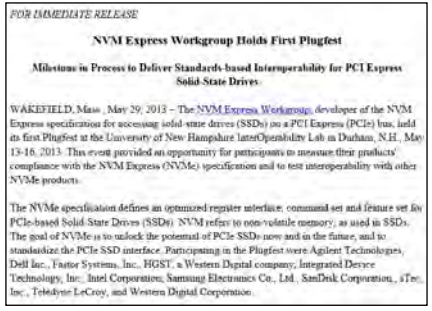
IDF13

41

Crossing the Chasm
Requires Patience and Perseverance

NVM Express* Deployment is Starting

- First plugfest held May 2013 with 11 companies participating
  - Three devices on Integrator's List
  - Next plugfest planned for Q4
- Samsung announced first NVM Express* (NVMe) product in July



FOR IMMEDIATE RELEASE

NVM Express Workgroup Holds First Plugfest

Milestone in Process to Deliver Standards-based Interoperability for PCI Express Solid-State Drives

WAKEFIELD, Mass., May 29, 2013 – The NVM Express Workgroup, developer of the NVM Express specification for accessing solid-state drives (SSDs) on a PCI Express (PCIe) bus, held its first Plugfest at the University of New Hampshire InterOperability Lab in Durham, N.H., May 13-16, 2013. This event provided an opportunity for participants to measure their products' compliance with the NVM Express (NVMe) specification and to test interoperability with other NVMe products.

The NVMe specification defines an optimized register interface, command set and feature set for PCIe-based Solid-State Drives (SSDs). NVM refers to non-volatile memory, as used in SSDs. The goal of NVMe is to unlock the potential of PCIe SSDs now and in the future, and to standardize the PCIe SSD interface. Participating in the Plugfest were Agilent Technologies, Dell Inc., Fastor Systems, Inc., HGST, a Western Digital company, Integrated Device Technology, Inc., Intel Corporation, Samsung Electronics Co., Ltd., SanDisk Corporation, , PCIe, Inc., Teledyne LeCroy, and Western Digital Corporation.

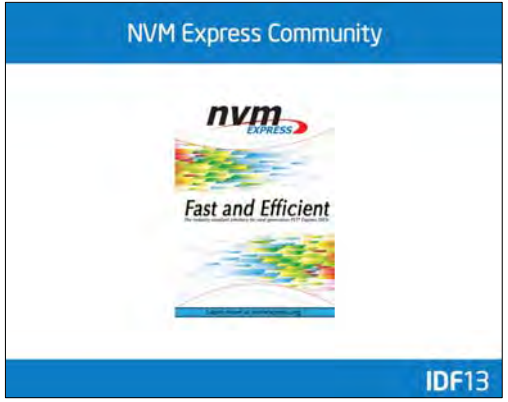
NVMe products targeting Datacenter shipping this year

IDF13

44

Learn More in the NVMe Community

Check out the NVMe Community in the Showcase to see NVMe products in action



Company	Booth #
Dell	726
Intel	727 & 734
EMC	728
Micron	729
SanDisk	730
LSI	731
SNIA	732
PMC-Sierra	733
Agilent	735
Western Digital	736
Teledyne LeCroy	737
Viking Technology	738
Tektronix	739

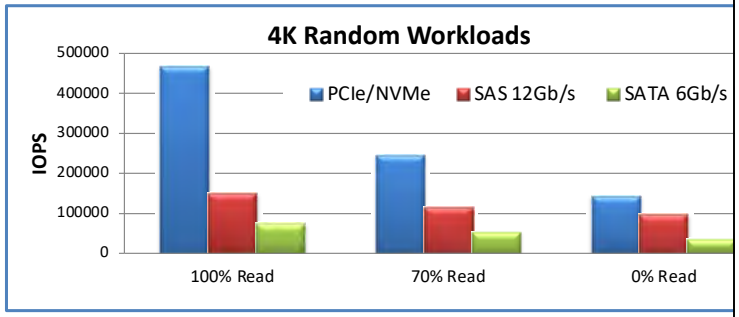
IDF13

55

Evangelize the Value Proposition
Fast, Efficient, Simple

NVM Express* (NVMe) Delivers Best in Class IOPs

- 100% random reads: NVMe has >3X better IOPs than SAS 12Gbps
- 70% random reads: NVMe has >2X better IOPs than SAS 12Gbps
- 100% random writes: NVMe has ~ 1.5X better IOPs than SAS

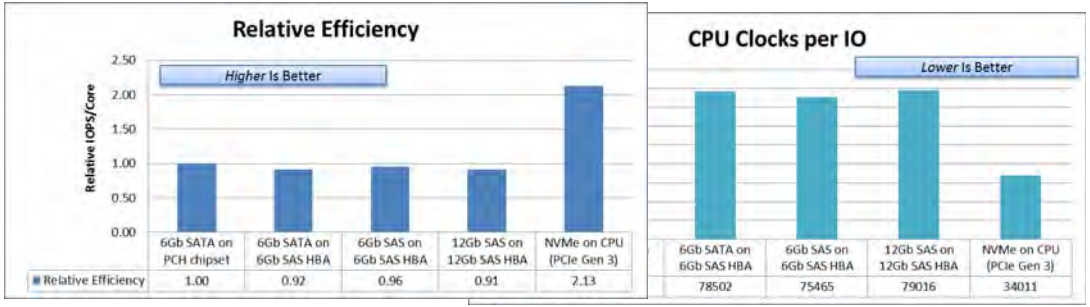


Note: PCI Express* (PCIe*)/NVM Express* (NVMe) Measurements made on Intel®Core™i7-3770S system @ 3.1GHz and 4GB Mem running Windows* Server 2012 Standard O/S, Intel PCIe/NVMe SSDs, da IOmeter* tool. PCIe/NVMe SSD is under development. SAS Measurements from HGST Ultrastar* SSD800M/1000M (SAS) Solid State Drive Specification. SATA Measurements from Intel Solid State Drive Product Specification. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMar using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performar in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

14

The Efficiency of NVM Express* (NVMe)

- CPU cycles in a Data Center are precious
- And, each CPU cycle required for an IO adds latency
- NVM Express* (NVMe) takes less than half the CPU cycles per IO as SAS



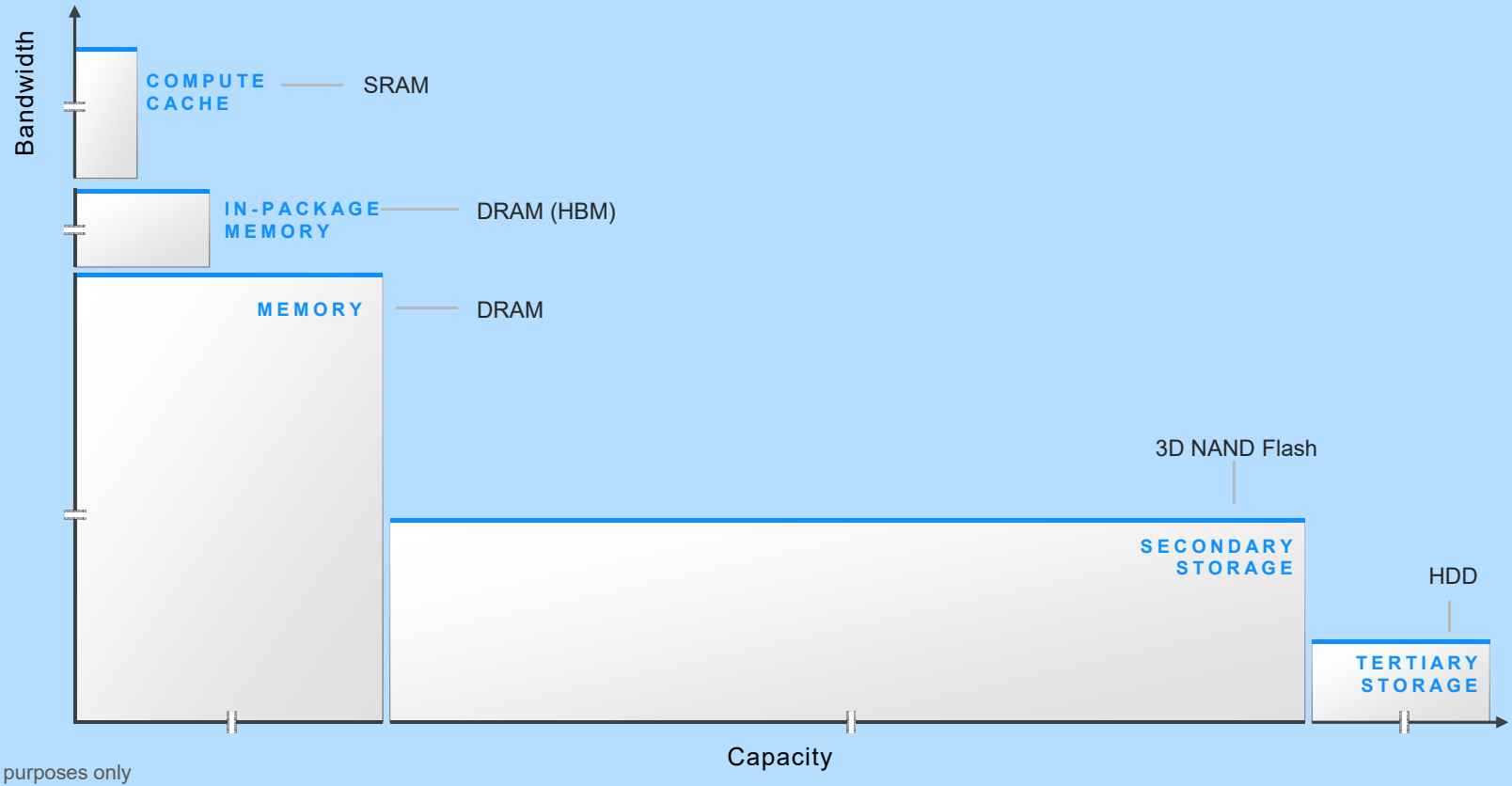
With equivalent CPU cycles, NVMe delivers over 2X the IOPs of SAS!

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For detailed configuration information, refer to "Setup for Efficiency and Latency Analysis" foil in Backup.

17

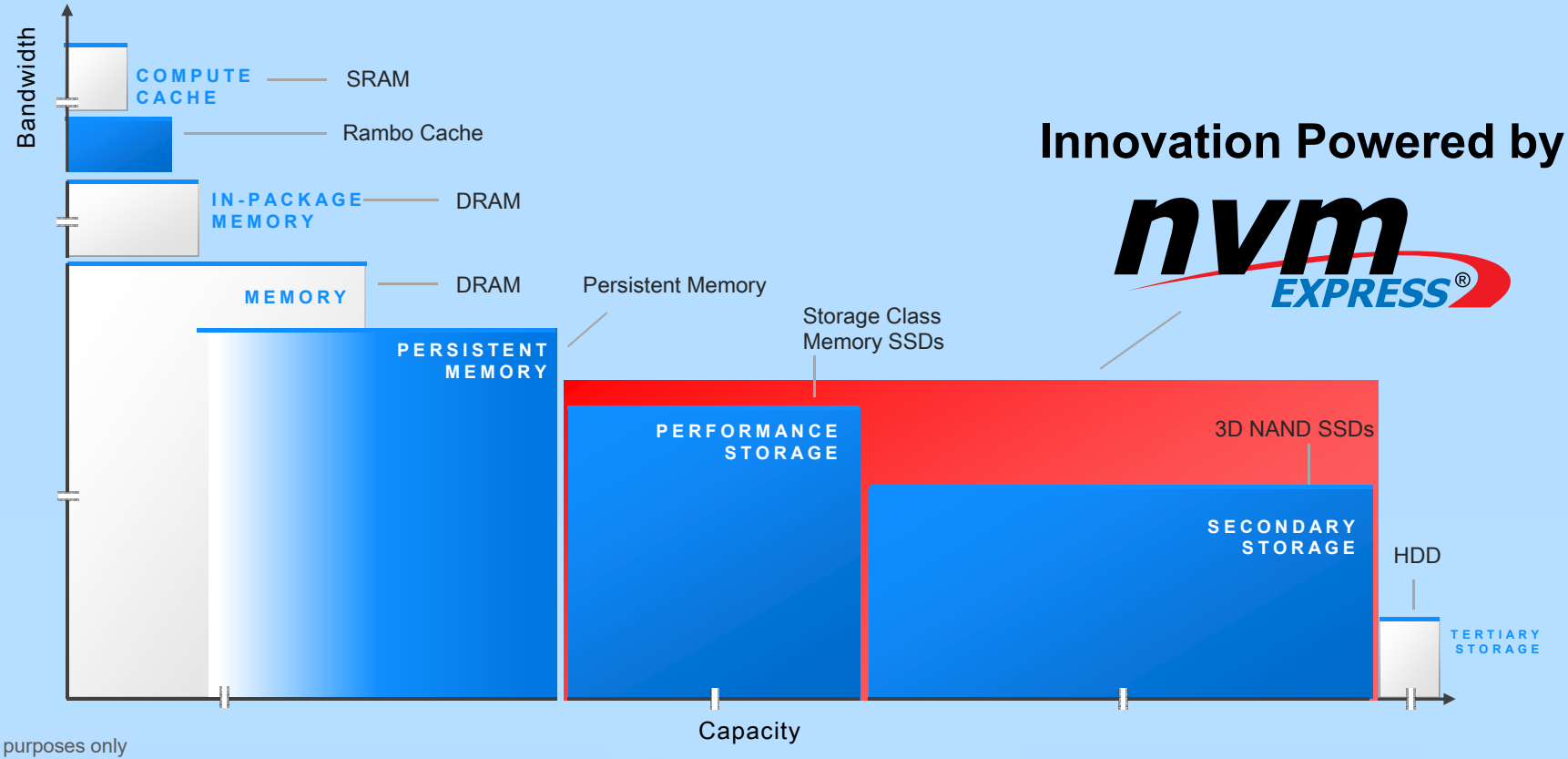
IDF14

Memory and Storage Hierarchy Gaps



For illustrative purposes only

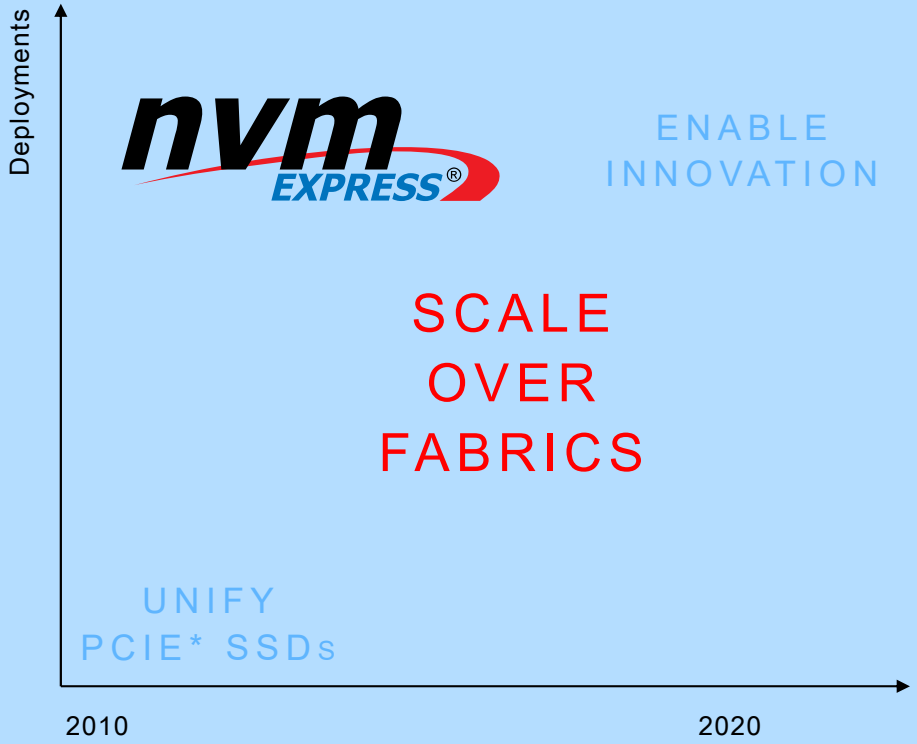
Memory and Storage Hierarchy ~~Gaps~~ Solutions



Innovation Powered by nvm EXPRESS®

For illustrative purposes only

The Evolution of NVMe® Technology



Enabling More Use Cases
NVMe* over Fabrics

Why NVM Express* (NVMe) over Fabrics?

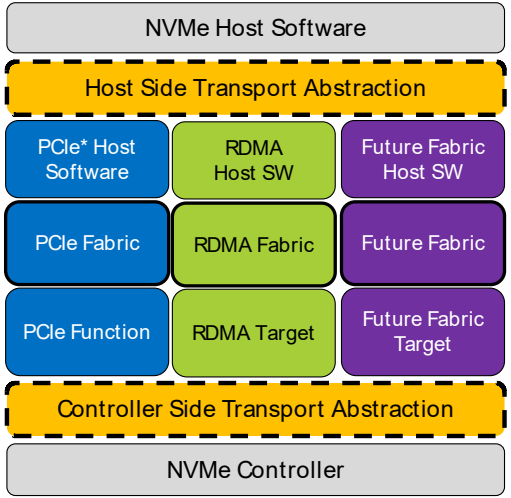
- Simplicity, Efficiency and End-to-End NVM Express* (NVMe) Model
  - NVMe has a single Admin queue pair with 10 required commands
  - NVMe supports up to 64K I/O Queues with 3 required commands
  - Simplicity of protocol enables hardware automated I/O Queues – transport bridge
  - No translation to or from another protocol like SCSI (in firmware/software)
  - Inherent parallelism of multiple I/O Queues is exposed
  - NVMe commands and structures are transferred end-to-end

Goal: Make remote NVMe equivalent to local NVMe, within ~ 10 µs latency.

31

Architectural Approach

- The NVM Express* (NVMe) Workgroup is starting the definition of NVMe over Fabrics
- The first fabric definition is the RDMA protocol – used with Ethernet and InfiniBand™
- A flexible transport abstraction layer is under definition, useful for many different fabrics



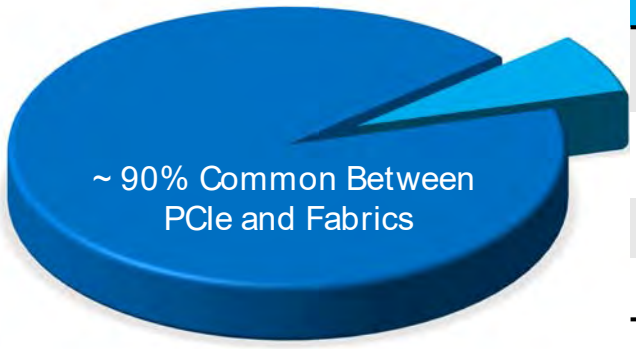
32

IDF14

Enabling More Use Cases
NVMe* over Fabrics

Commonality Between PCI Express®and Fabrics

- The vast majority of NVM Express™(NVMe) is leveraged as-is for Fabrics
  - NVM Subsystem, Namespaces, Commands, Registers/Properties, Power States, Asynchronous Events, Reservations, etc.
- Primary differences reside in enumeration and queuing mechanism



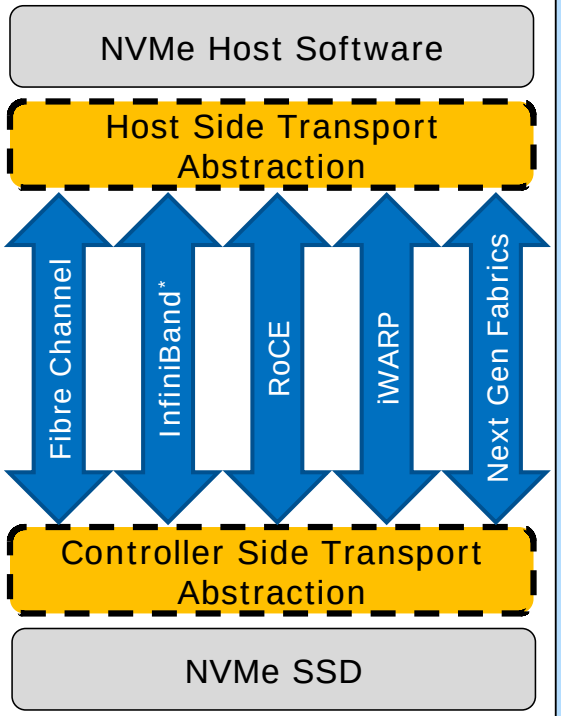
Differences	PCI Express®(PCIe)	Fabrics
Identifier	Bus/Device/ Function	NVMe Qualified Name (NQN)
Discovery	Bus Enumeration	Discovery and Connect commands
Queuing	Memory-based	Message-based
Data Transfers	PRPs or SGLs	SGLs only, added Key

IDF15 INTEL DEVELOPER FORUM

Enabling More Use Cases
NVMe* over Fabrics

NVMe* over Fabrics

- Use NVMe* end-to-end to get the simplicity, efficiency and low latency benefits
- NVMe over Fabrics is a thin encapsulation of the base NVMe protocol across a fabric
  - No translation to another protocol (e.g., SCSI)
- NVMe over Fabrics 1.0 includes RDMA binding enabling Ethernet and InfiniBand™
  - INCITS T11 defining Fibre Channel binding

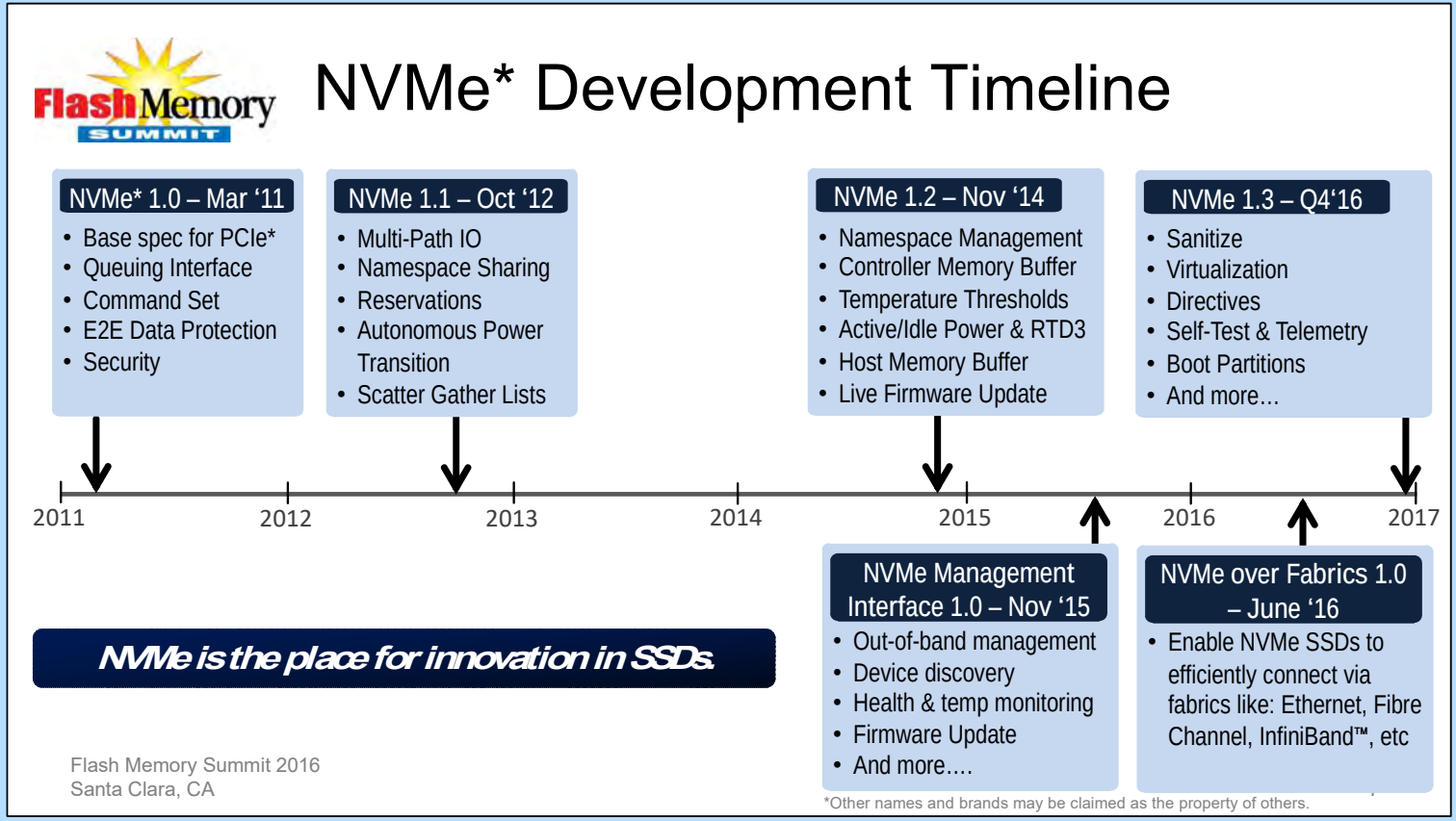


Flash Memory Summit 2016
Santa Clara, CA

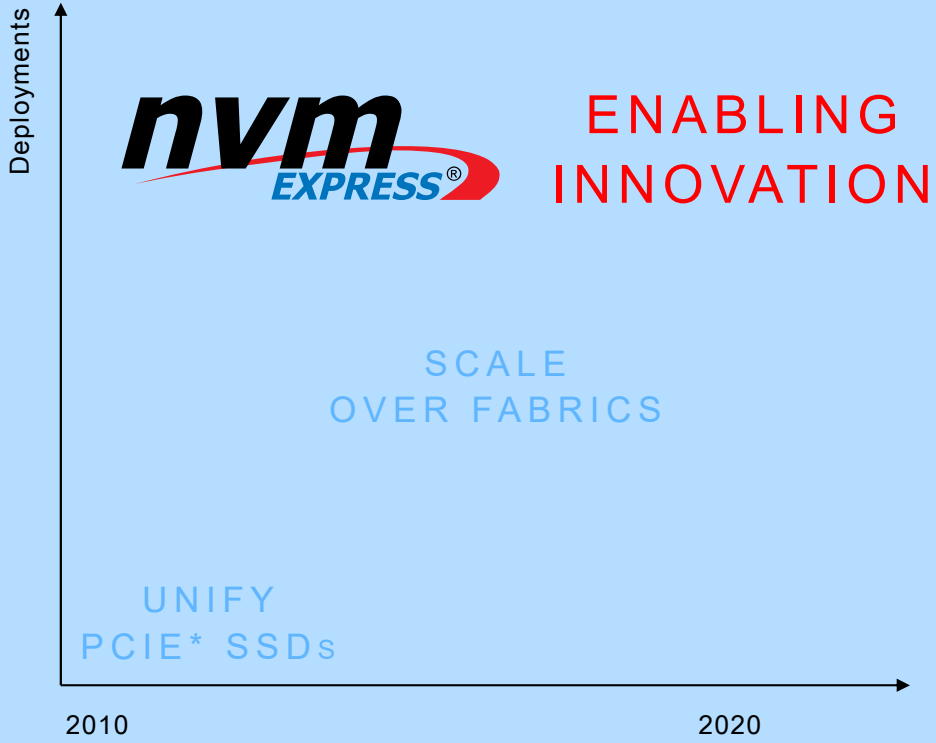
*Other names and brands may be claimed as the property of others.

16

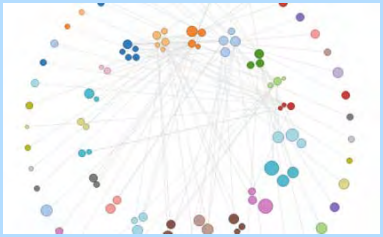
NVMe® Technology
Roadmap Evolution



The Evolution of NVMe® Technology

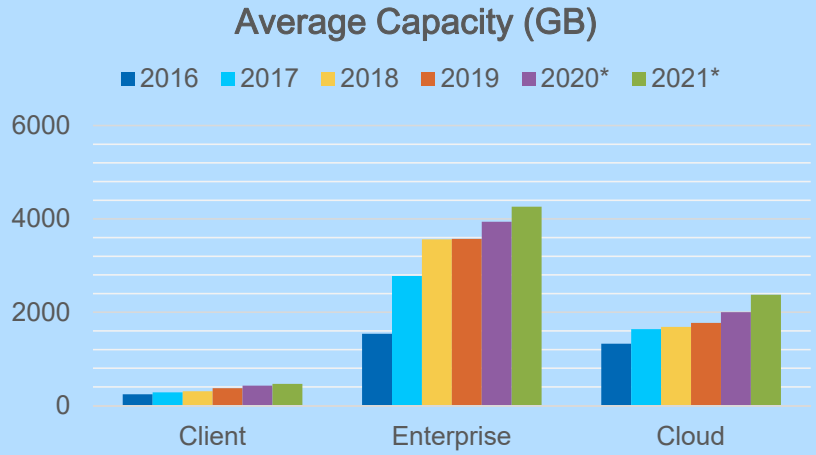


NVMe® Technology Powers the Connected Universe



Units (Ku)	2016	2017	2018	2019	2020*	2021*
Enterprise	364	749	1,069	2,045	4,067	5,554
Cloud	2,051	3,861	10,369	12,276	18,982	21,999
Client	33,128	48,951	82,587	143,236	202,348	258,791

* Projections provided by Forward Insights Q2'20

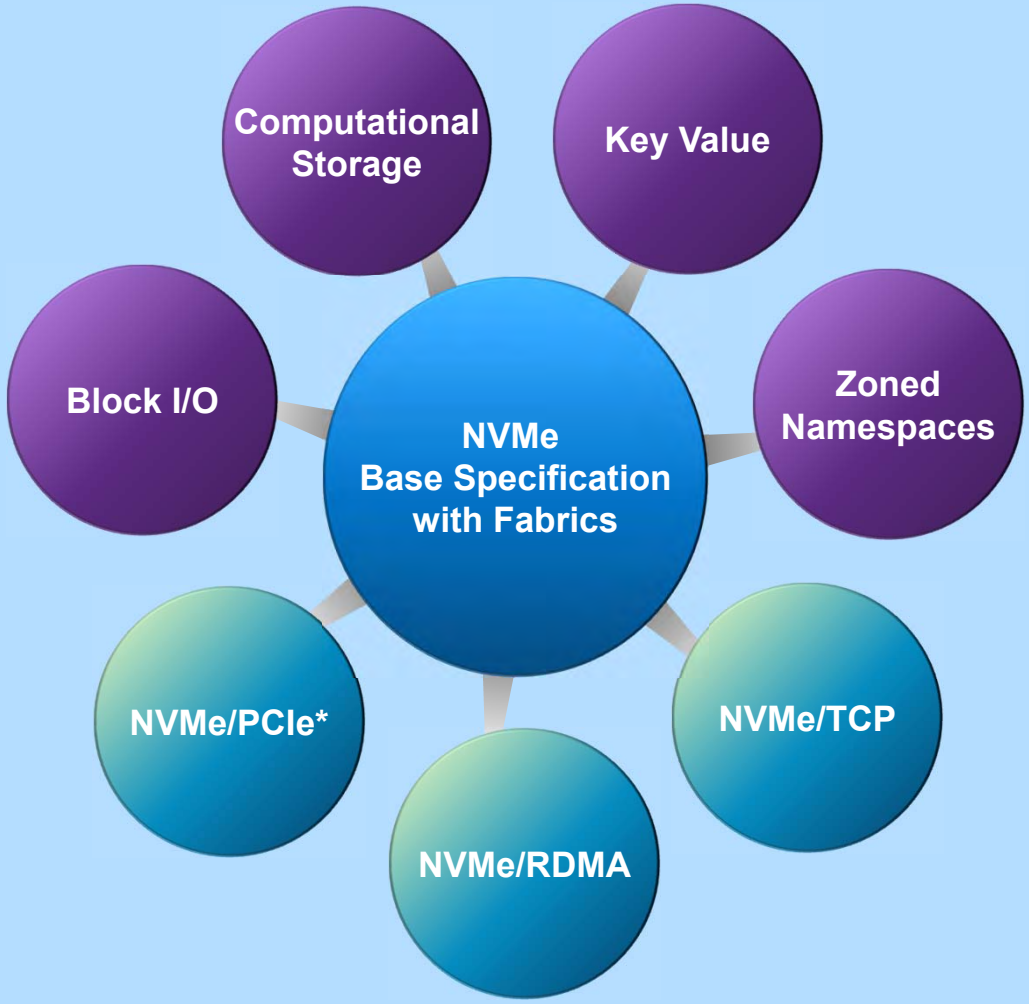


- NVMe technology grew from 3 Petabytes to 29 PB shipped per year from 2016 to 2019
- For 2020, the projection is 54 PB
- NVMe technology demand projected to remain strong in a post COVID world

Driving Simplicity in a World of Complexity

Driving Simplicity in a World of Complexity

- Focused on core values… Fast, Simple, Scalable
- Foster areas of innovation AND avoid impact to broadly deployed solutions
- Create an extensible infrastructure that will take us through the next decade of growth



Specification Families

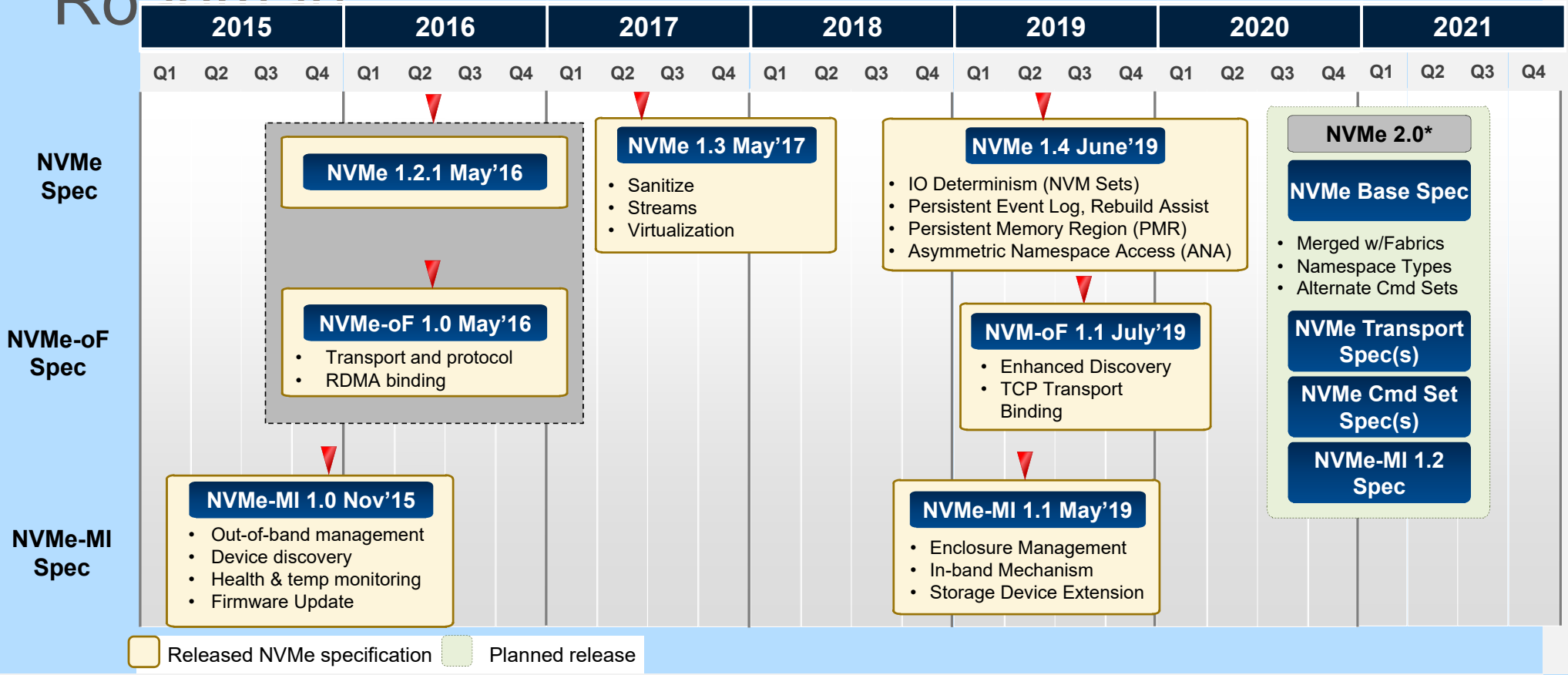
THE CORE

COMMAND SETS

TRANSPORT LAYERS

- The core of NVMe and NVMe over Fabrics integrated into a base specification
- Modular command set specifications (Block, Zoned Namespaces, Key Value, etc)
- Modular transport layer specifications (PCI Express*, RDMA, TCP)
- Maintain Management Interface as separate modular specification

NVM Express Technology Specification Roadmap



Q&A answers by:

I Dal Allan

Tom Gardner

Amber Huffman

Jai Menon

Grant Saviers