

NVMe devices & IBM i

A closer look...

COMMON Luxembourg

20/02/2020

Fabian Michel

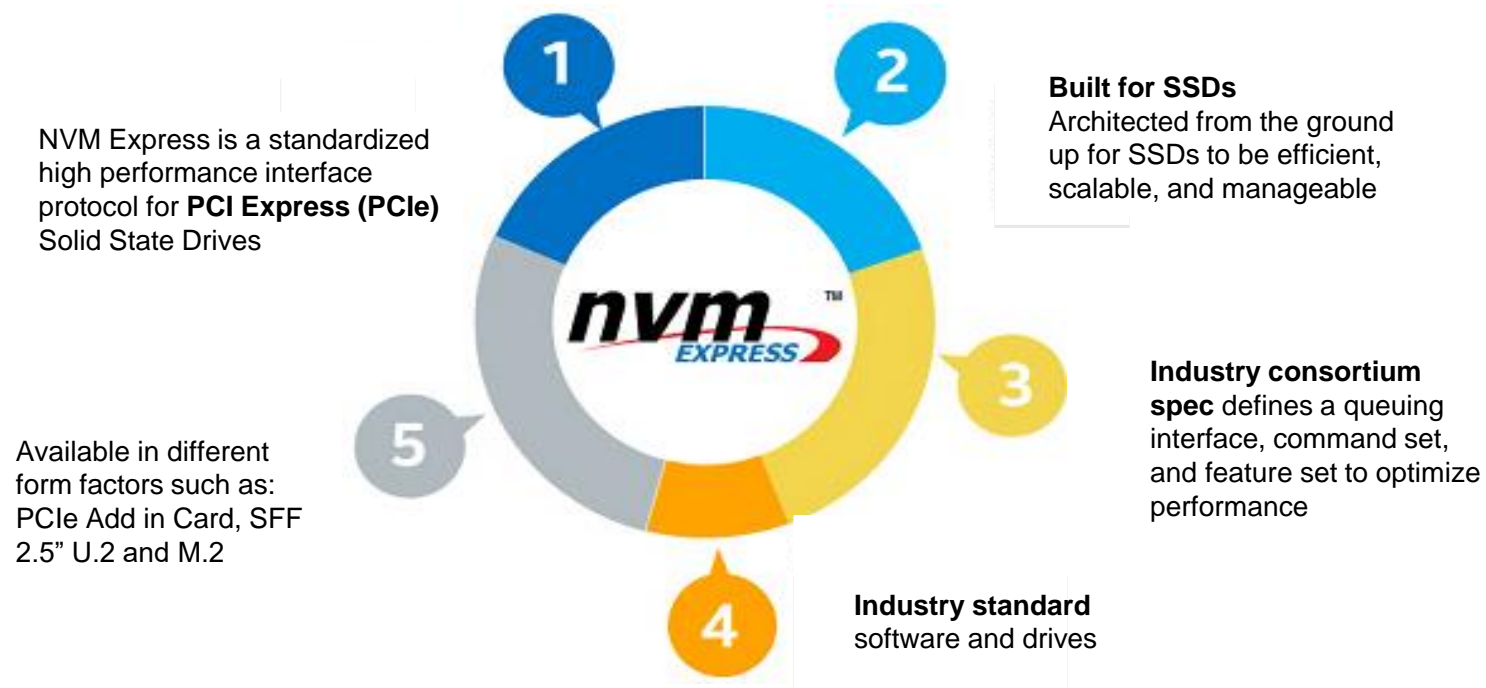
Client Technical Architect



What is NVMe?

NVMe - Non-Volatile Memory express

<http://www.nvmexpress.org/>

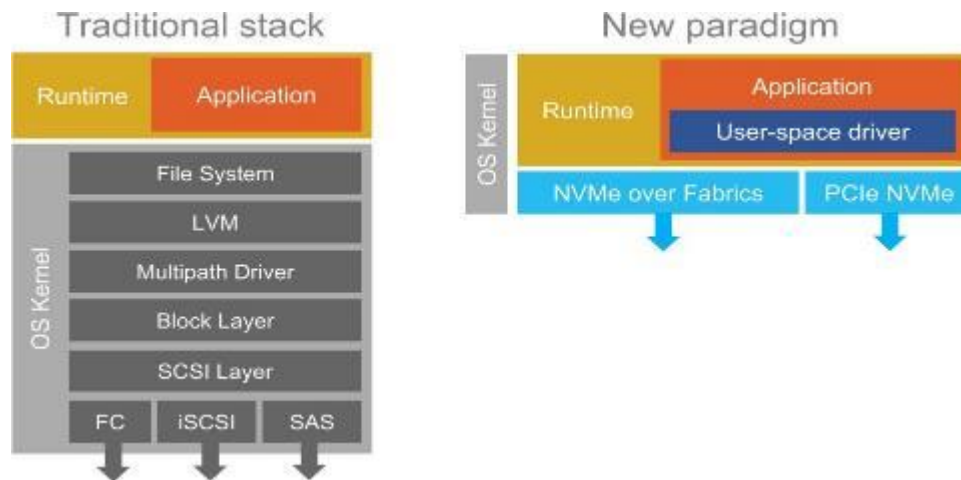


Direct and Network attached low latency, high bandwidth, cost effective Flash solutions

What is NVMe?

SSDs are fast. **So fast in fact, their limiting factor is not their own hardware, but rather the SAS or SATA connection that hard drives** have traditionally used.

NVMe - “Non-Volatile Memory Express,” NVMe is an open standard developed to allow modern SSDs to operate at the read/write speeds their flash memory is capable of. Essentially, it allows flash memory to operate as an SSD directly through the PCIe interface rather than going through SATA and being limited by the slower SATA speeds.



Storage Latency and Command / Data Throughput

Latency	Bus	Media	Read Lat. (us)	Write Lat. (us)	Read (IOPs)	Write (IOPs)	Read Tp (GB/s)	Write Tp (GB/s)	Cost
	Memory (in CEC)	DRAM/DIMM	<1	<1	Not a Persistent Storage				
		SCM	<1	<1	Persistent Storage				
	PCIe (NVMe)	SCM	<10	<10	550K	500K	2.4	2.0	
		LL Flash	<15	<15	750K	180K	3.2	3.0	
		Flash	<90	<25	750K	180K	3.5	3.0	
	SAS	Flash	150	60	420K	50K	2.2	1.6	
	SATA	Flash	1.8ms	3.6ms	93K	25K	0.5	0.5	
	SAS / SATA	HDD	>ms	>ms	200	200	0.15	0.15	
		TAPE	"secs"	"secs"	"slow"	"slow"	"slow"	"slow"	

SCM: 3DXP from Intel/Micron. Bytes addressable in DIMM (Apache Pass) and Block addressable(M.2/U.2/AIC..) in NVMe interface.

NVMe/SCM: Performance numbers are of Intel's Optane PCIe Gen 3 x4 Add in Card. Endurance 30 DWPD.

NVMe/LL Flash: Performance numbers are of Samsung's zSSD projections.

NVMe: Intel, Samsung, WD, Micron adapters are PCIe Gen 3 x 4. Performance limited by the controller.

SAS SSD: Assumes 12G dual port active/active. Performance of single port operation (typical) expected to be lower.

IOPs and Latencies: Normally measured on a random 4K ops. * <1us for 1K transfer utilizing Persistent Log Buffer feature

Data throughput: Normally measured on a large sequential 256KB ops

POWER9 PCIe Add In Card NVMe Device



Hardware Features

- NVMe Specs. 1.2.1 Compliant
- NVMe Over Fabrics 1.0 Capable
- PCIe Gen 3 x 8
- Multiple Name Space (32)
NS Granularity 16GB
- Half Height Half Length (HH-HL)
- Power ≤ 25W
- Block Size 4096(Default), 512, **4160 (IBM i)**
- End-To-End Protection: T10 DIF & DIX
- Non Volatile Write Buffer
- **Endurance 5 DWPD** for 1.6/3.2/6.4TB
- PCIe Vendor VPD Support (IBM Provides content)
- Boot: Option ROM BAR 128KB (IBM Provides content)
- Warranty ≥ 5 years
- Hot Plug capable
- ECC ≥ 100 bits per 4KB
- RAID: Tolerant of single flash die failures
- MTBF ≥ 2 million hours
- End Of Life Data Retention ≥ 3 months
- EEH Support
- Live Firmware update
- NVMe-MI (Optional)
- **Non-TCG SED**
- **No support for MEX Drawer**

	PCIe3 NVMe Flash Adapter		
	1.6TB	3.2TB	6.4TB
FC (LP/FH)	EC5G / EC5B	EC5C / EC5D	EC5E / EC5F
IBM i FC (LP/FH)	EC6U / EC6V	EC6W / EC6X	EC6Y / EC6Z
Workload	Target (1.6 TB)	Target (3.2/6.4)	
Read (IOPS)	700K	910K	
Write (IOPS)	100K	170K	
Mixed R/W (70/30)	250K	320K	
Read Data Tp (GB/s)	4.7	6.0	
Write Data Tp (GB/s)	1.9	3.0	
Read Latency (us)	110	110	
Write Latency (us)	30	30	

Notes:

1. IOPs and Latency #'s on random 4K
2. Data throughput #'s are on sequential 256KB work load

Software Support

- Linux
 - Power VM:** RHEL 7.5LE, SLES 12 SP3 LE
 - Ubuntu 18.04
 - Power NV:** RHEL 7.5LE, Ubuntu 18.04
- AIX (7.1Z & 7.2F), VIOS (2.2.6)
- **IBM i (7.4 TR1)**
- Load Source
- Software RAID 0, 1, 5 & 6 (Linux)
- OS Mirroring (AIX, IBM i)
- DIAG Support
- NVMe Over Fabrics (Linux Only)

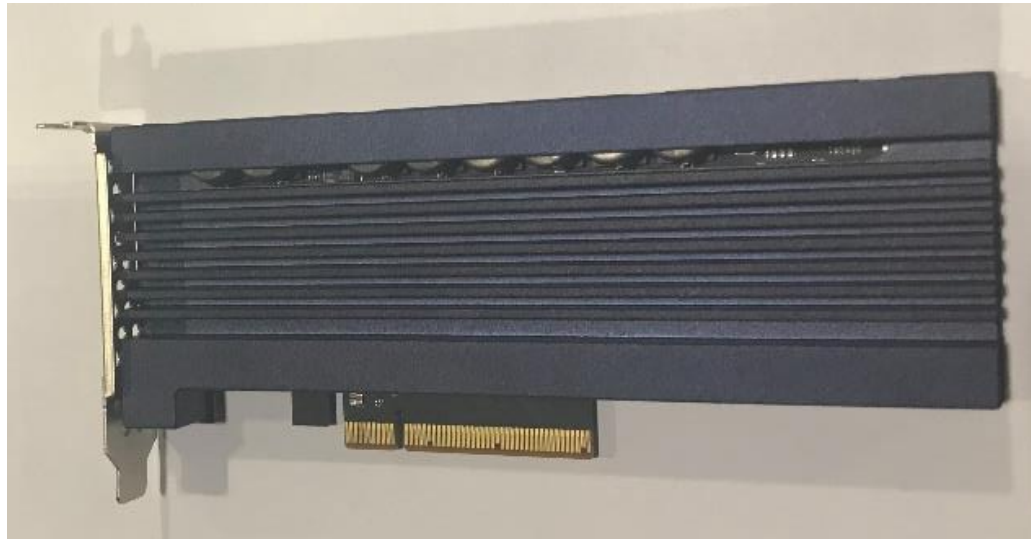


NVMe and IBM i

- NVMe is able to provide **higher performance than SSDs**
- NVMe will provide **additional virtualization** capabilities on Power
 - Every device is a PCIe endpoint that can be dedicated to a partition/LPAR
- At least **one identical NVMe adapter pair is required**; subsequent NVMe adapter pairs can be different than the first pair. After an identical pair is on the order, one NVMe adapter of different capacity is allowed.
- NVMe devices require **IBM i operating system mirroring** as there isn't hardware RAID support. Mirrored pairs must be on different physical devices. NVMe can only mirrored to NVMe and SAS drives can only be mirrored to SAS drives.
- IBM i supports virtualized NVMe via VIOS
- NVMe devices are planned to be included as **supported direct attached devices for IBM Db2 Mirror for i (SOD)**

NVMe and IBM i

- NVMe is **only supported in the system unit**. Not supported in a PCIe Gen3 I/O drawer.
- **S914 supports up to 3 NVMe. S924 supports up to 5 NVMe.**
- **S914 4-core P05 system is limited to 2 x 1.6TB devices only.** Mixing NVMe and SAS drives is not allowed (ten maximum of SAS drives or two maximum of NVMe).
- **E980 supports up to 8 NVMe per CEC** (6 first drawer, 8 each drawer 2, 3, and 4 for a maximum total of 30)

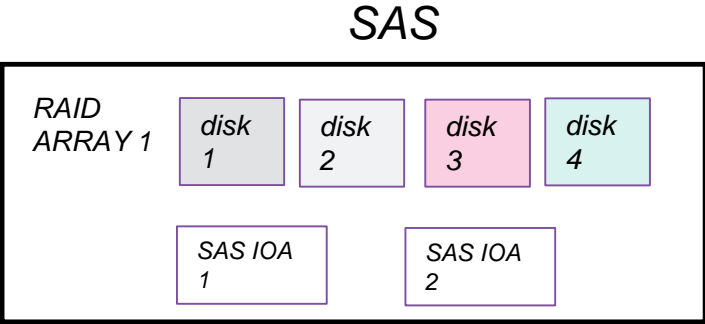


NVMe Namespaces and IBM i

- **NVMe uses namespaces** which is a collection of logical blocks whose logical block addresses range from 0 to the size of the namespace. A namespace ID (NSID) is an identifier used by a controller to provide access to a namespace.
- **With NVMe, an 'arm' (logical unit) is a namespace.** A namespace is a logical chunk of a physical NVMe device and multiple namespaces are allowed on one NVMe device.
- **IBM i is the management interface** used by a customer to create and manage namespaces
- **IBM i's use of NVMe architected multiple namespaces provides for many 'arms' on a small number of high capacity NVMe physical devices**
- IBM i can use a NVMe device (up to 16 TB) with only a single namespace for the whole device. However, for almost all customers, this will cause sub-optimum performance since more (and smaller) 'arms' (logical units) are better than fewer and larger.

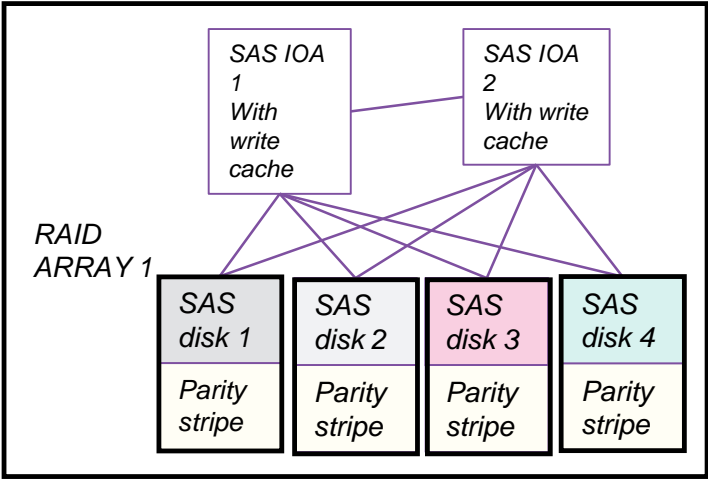
SAS Versus NVMe Storage with IBM i

Customer and operating system view

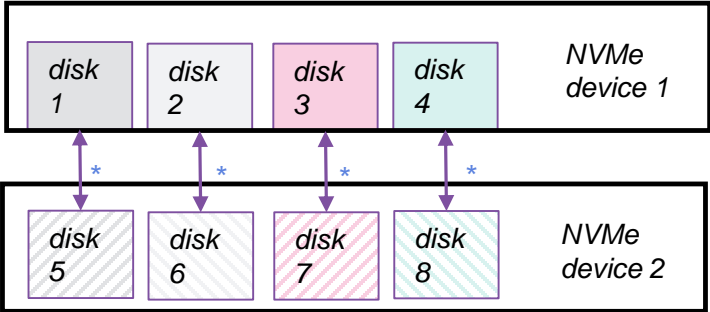


OS is aware of hardware RAID

Device physical view

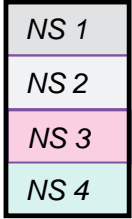


NVMe



**OS mirroring*

NVMe device 1



NVMe device 2



Recommended NVMe Namespace Sizes

- First generation NVMe devices have a hardware **boundary of 16 GB for name spaces**. Device capacity can be wasted/lost if name spaces are not multiples of 16. The **maximum number of namespaces on a device is 32**.
- IBM i screens show Capacities in 'GB' (1000**3 (GB), not 1024**3 (GiB))
- IBM recommends **namespaces of 188 GB or 393 GB**
- Some possible configuration:
 - S914 4 cores : **max 2 x 1.6TB = 1448 GB** net (8 namespaces)
 - S914 6/8 cores : **max 3 x 6.4TB = 9432 GB** net (16 namespaces)
 - S924 : **max 5 x 6.4TB = 15720 GB** net (16 namespaces)
 - E980 : **max 30 x 6.4TB = 94320 GB** net (16 namespaces)

Device Nominal Size	Device Actual Size	Number of Namespaces	Namespace Size	Total User Capacity Used By Namespace	Remaining Space on the Device (unallocated)
1.6TB	1575	8	188	1448	87
3.2TB	3151	16	188	2977	174
6.4TB	6364	32	188	6016	348
1.6TB	1575	4	393	1556	19
3.2TB	3151	8	393	3112	38
6.4TB	6364	16	393	6288	76

Work with Disk Units

Select one of the following:

1. Save load source disk unit data
2. Copy load source disk unit data
3. Display/change page data
4. Analyze disk unit surface
5. Initialize and format disk unit
6. Reclaim I/O cache storage
7. Stop device parity protection
8. Update system vital product data
9. Start device parity protection - RAID 5
10. Start device parity protection - RAID 6
11. Start device parity protection - RAID 5 with hot spare
12. Start device parity protection - RAID 6 with hot spare
13. Start device parity protection - RAID 10
14. Start device parity protection - RAID 10 with hot spare
15. Work with NVM Devices

Selection

15

F3=Exit

F12=Cancel

Create NVM Namespaces

Device	Serial Number	Resource Name	Type	Logical Address
NVM	Y0YACBYCB06Z	DC04	58FD	U78D2.001.WZS0067-P1-C8

NVM Configuration	-----Capacity in GB-----			---Namespaces---		
	Used	Available	Total	Used	Avail	Total
Current :	0	3151	3151	0	32	32

Type choices, press Enter.

Quantity of namespaces to create :	<u>16</u>	(1 - 32)
Capacity of each namespace :	<u>188</u>	(64 - 3151) GB

F3=Exit

F5=Refresh

F12=Cancel

Confirm Create NVM Namespaces

Device	Serial Number	Resource Name	Type	Logical Address
NVM	Y0YACBYCB06Z	DC04	58FD	U78D2.001.WZS0067-P1-C8

Quantity of namespaces to create : 16

Capacity of each namespace to create : 188 GB

-----Capacity in GB-----				---Namespaces---		
NVM Configuration				Used	Avail	Total
Current :				0	3151	3151
Projected :				3008	143	3151

Note: Each namespace will be shown as a non-configured disk unit when the create operation completes.

Press F10 to confirm the choice to create namespaces.
Press F12 to return to change your choice.

F10=Confirm F12=Cancel

Display NVM Namespaces

NVM Device	ASP	Unit	Serial Number	Type	Model	Resource Name	Namespace Capacity in GB
1			Y0YACBYCB06Z	58FD		DC04	
	*	*	YKXN3RU6X288	6B7D	205	DD001	188
	*	*	YQARSJVDPW9S	6B7D	205	DD002	188
	*	*	YAPM3HNKEDTW	6B7D	205	DD003	188
	*	*	YW7PU8J4AJ5C	6B7D	205	DD004	188
	*	*	YE6949WFVM5C	6B7D	205	DD005	188
	*	*	YDBURU4M8QBK	6B7D	205	DD006	188
	*	*	YTUVKJKRFNMA	6B7D	205	DD007	188
	*	*	YFH2NN8VPFAX	6B7D	205	DD008	188
	*	*	YBAHV8X4AZL2	6B7D	205	DD009	188
	*	*	YBJK8AAXQXSA	6B7D	205	DD010	188
	*	*	YWGJJ5EBZXRv	6B7D	205	DD011	188
	*	*	YJBHLHLJYZ4U	6B7D	205	DD012	188
	*	*	YMNDKGCDL9U5	6B7D	205	DD013	188
	*	*	Y8FR686Z4J8F	6B7D	205	DD014	188
	*	*	YPZG7P9TQD5Y	6B7D	205	DD015	188
							More...
* - Non-configured disk unit							
F3=Exit		F5=Refresh		F12=Cancel			

Display NVM Namespaces							
NVM Device	ASP	Unit	Serial Number	Type	Model	Resource Name	Namespace Capacity in GB
2	*	*	YTZFM7DGE5FY	6B7D	205	DD016	188
			Y0YACBYCB071	58FD		DC05	
	*	*	YMDM4NWZC9A3	6B7D	205	DD017	188
	*	*	YWBCDJ8RHNF3	6B7D	205	DD018	188
	*	*	YQPM2CPACRBU	6B7D	205	DD019	188
	*	*	YKEPEJSLUM8E	6B7D	205	DD020	188
	*	*	Y2982PQE93M6	6B7D	205	DD021	188
	*	*	Y65294CC2D8V	6B7D	205	DD022	188
	*	*	YJQDCS3PAAT5	6B7D	205	DD023	188
	*	*	YAX6BKRPKEK8	6B7D	205	DD024	188
	*	*	YGR2LA9PAEUW	6B7D	205	DD025	188
	*	*	YG3MPFT7254D	6B7D	205	DD026	188
	*	*	YLT2V9V7PGKH	6B7D	205	DD027	188
	*	*	YQ9VQMAUCFKW	6B7D	205	DD028	188
	*	*	YLZSKL7L8MDN	6B7D	205	DD029	188
	*	*	YAN7MYFWAUAA	6B7D	205	DD030	188
							Bottom
* - Non-configured disk unit							
F3=Exit		F5=Refresh		F12=Cancel			

Confirm Start Mirrored Protection

Press Enter to confirm your choice to start mirrored protection. During this process the partition will be IPLed. You will return to the DST main menu after the IPL is complete. The ASP will have the displayed protection.

Press F12 to return to change your choice.

ASP	Unit	Serial Number	Type	Model	Resource Name	Protection	Hot Spare Protection
1						Mirrored	
	1	YKXN3RU6X288	6B7D	205	DD001	planar	N
	1	YMDM4NWZC9A3	6B7D	205	DD017	planar	N
	2	YQARSJVDPW9S	6B7D	205	DD002	planar	N
	2	YWBCDJ8RHNF3	6B7D	205	DD018	planar	N
	3	YAPM3HNKEDTW	6B7D	205	DD003	planar	N
	3	YQPM2CPACRBU	6B7D	205	DD019	planar	N
	4	YW7PU8J4AJ5C	6B7D	205	DD004	planar	N
	4	YKEPEJSLUM8E	6B7D	205	DD020	planar	N

More...

F12=Cancel

Work with Disk Status

H10002A0

10/02/19 05:10:22 CDT

Elapsed time: 00:00:00

Unit	Type	Size (G)	% Used	I/O Rqs	Request Size (K)	Read Rqs	Write Rqs	Read (K)	Write (K)	% Busy
1	6B7D	188	4.6	.0	.0	.0	.0	.0	.0	0
1	6B7D	188	4.6	.0	.0	.0	.0	.0	.0	0
2	6B7D	188	.1	.0	.0	.0	.0	.0	.0	0
2	6B7D	188	.1	.0	.0	.0	.0	.0	.0	0
3	6B7D	188	.1	.0	.0	.0	.0	.0	.0	0
3	6B7D	188	.1	.0	.0	.0	.0	.0	.0	0
4	6B7D	188	.1	.0	.0	.0	.0	.0	.0	0
4	6B7D	188	.1	.0	.0	.0	.0	.0	.0	0
5	6B7D	188	.1	.0	.0	.0	.0	.0	.0	0
5	6B7D	188	.1	.0	.0	.0	.0	.0	.0	0
6	6B7D	188	.1	.0	.0	.0	.0	.0	.0	0
6	6B7D	188	.1	.0	.0	.0	.0	.0	.0	0
7	6B7D	188	.1	.0	.0	.0	.0	.0	.0	0

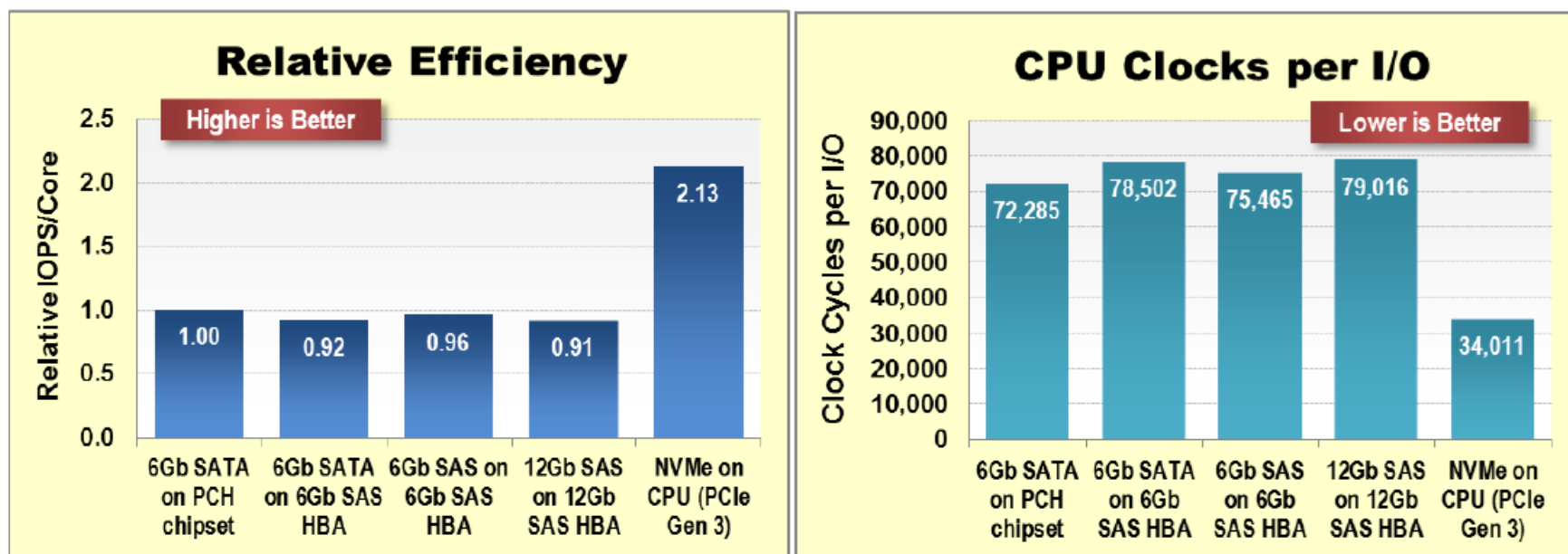
More...

Command

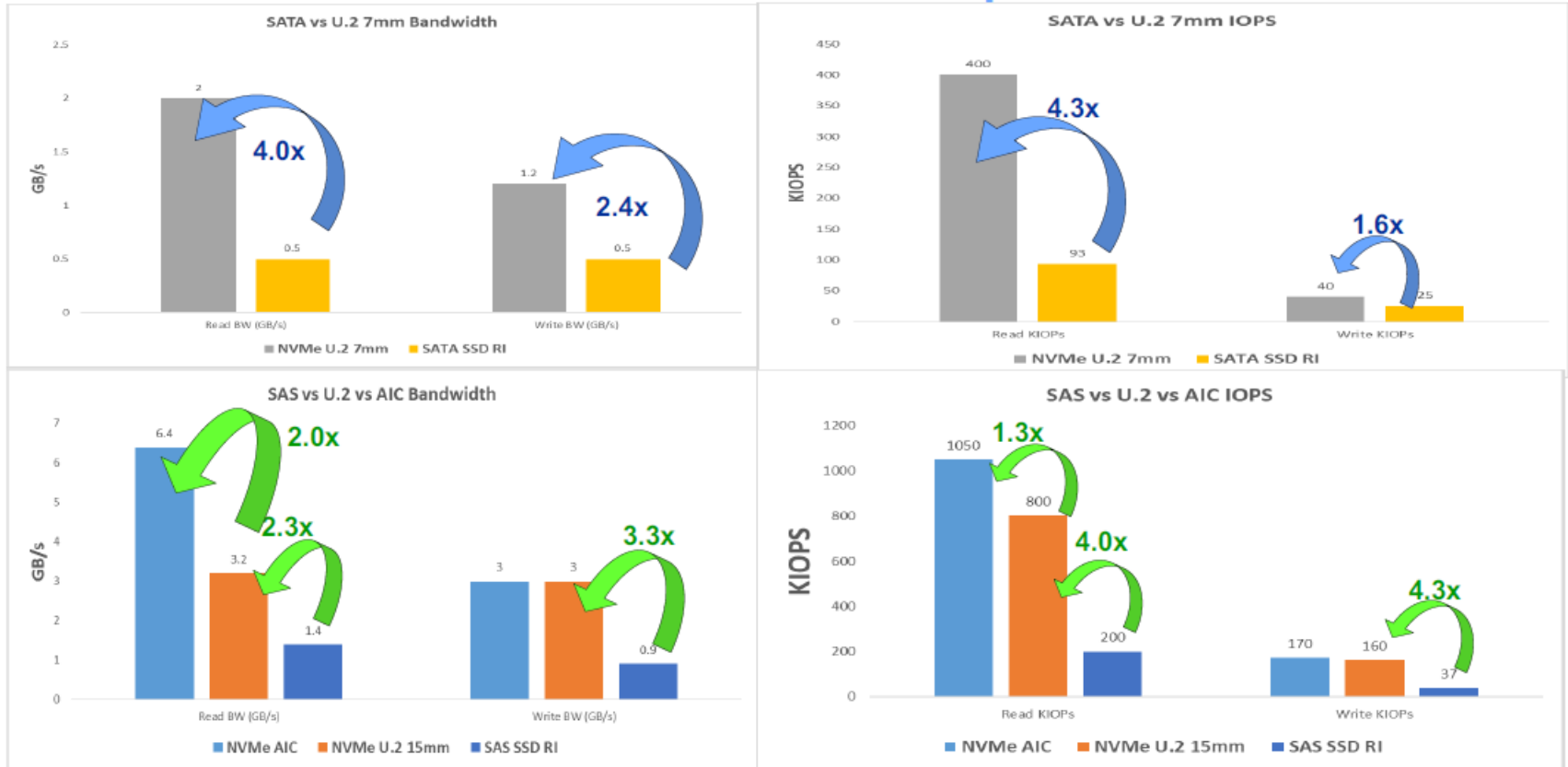
==>

F3=Exit F5=Refresh F12=Cancel F24=More keys

Is NVMe Fast?



SATA vs SAS vs NVMe SSD Comparison



What Does NVMe Cost vs SAS Storage

	HDD – 283G	SSD – 387G	Controller	HS Controller	Total
Simple RAID with 4 HDD	4		1		1132GB @ \$3299
Simple RAID with 4 SSD		4	1		1549G @ \$9799
High Speed RAID 4 SSD		4		1	1548GB @ \$12799

	Total
Dual 1.6TB NVMe	1.6TB @ \$6198
Dual 3.2TB NVMe	3.2TB @ \$12198
Dual 6.4TB NVMe	6.4TB @ \$24198

