Exploring the world of NVMe and IBM i

Douglas Gibbs
Offering Manager – Power Systems IO
douglasg@ca.ibm.com
@DouglasGibbsIBM

Dan Sundt
Partner Technical Advocate
dansundt@us.ibm.com

Today's Agenda

- What is an NVMe Device
- What are the Enterprise Feature Codes
- How to Configure and Use Them
  - Review on IBM i
- Performance Comparisons
  - SAS RAID-5 vs. NVMe Mirrored
  - SAS RAID-10 vs. NVMe Mirrored
- Conclusion

What is NVMe

Non-Volatile Memory Express

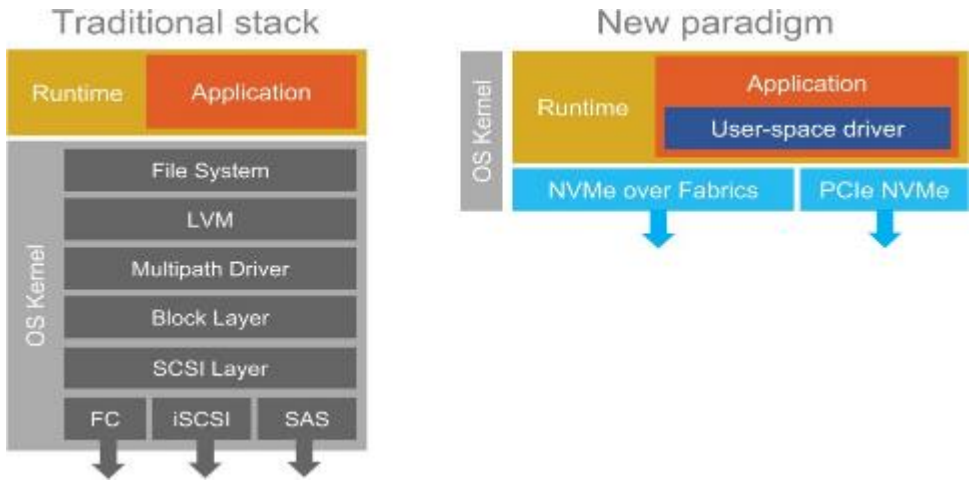
NVM Express® (NVMeTM) is an **optimized, high-performance scalable host controller interface** designed to address the needs of Enterprise and Client systems that utilize PCI Express®-based solid-state storage. Designed to move beyond the dark ages of hard disk drive technology, NVMe is built from the ground up for non-volatile memory (NVM) technologies. NVMe is **designed to provide efficient access to storage** devices built with non-volatile memory, from today's NAND flash technology to future, higher-performing, persistent memory technologies.

NVMe Industry Consortium: 13 Promoters elected for 2 years term and 100s of Contributor and Adopter member companies

https://nvmexpress.org

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" 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.



Note that the diagram above is just a representation. Operating systems such as IBM i have a machine interface and single level storage, thus does not let applications write directly to hardware.

What is NVMe

NVMe Controller

PCIe Attached
Parallel Architecture (Multi Q, Q pairs …)
Low Latency Design
Fabrics Attach Friendly
Self Encryption & Sanitize
Virtualization (Multiple Namespace, SR-IOV)
IO Determinism
Zoned Namespace
Management Interface Support (in & out of band)
Computational Storage
… More Innovations

Media

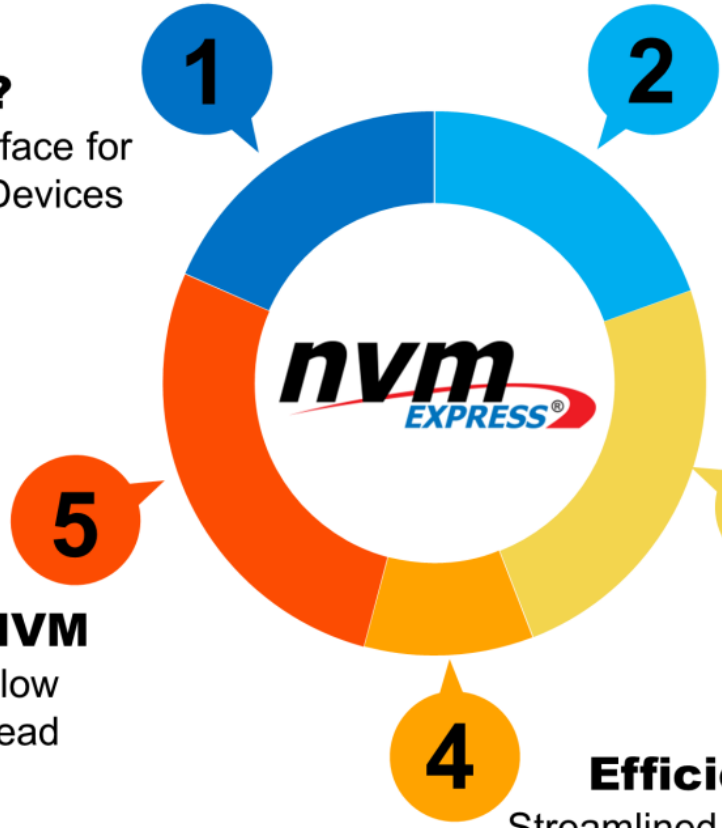
NAND TLC (most widely used)
Optane 3DXP
LL NAND
NAND QLC
DRAM – Flash backed
MRAM
.. More to come

Form Factors: M.2, U.2 (2.5” thin & thick), EDSFF (E1.S, E1.L, E3.S, E3.L thin & thick), Add in card

Software Infrastructure: Investment protected by re-using on multiple vendor devices

What is NVMe

What is NVMe?
THE industry standard interface for
PCI Express® Solid-State Devices



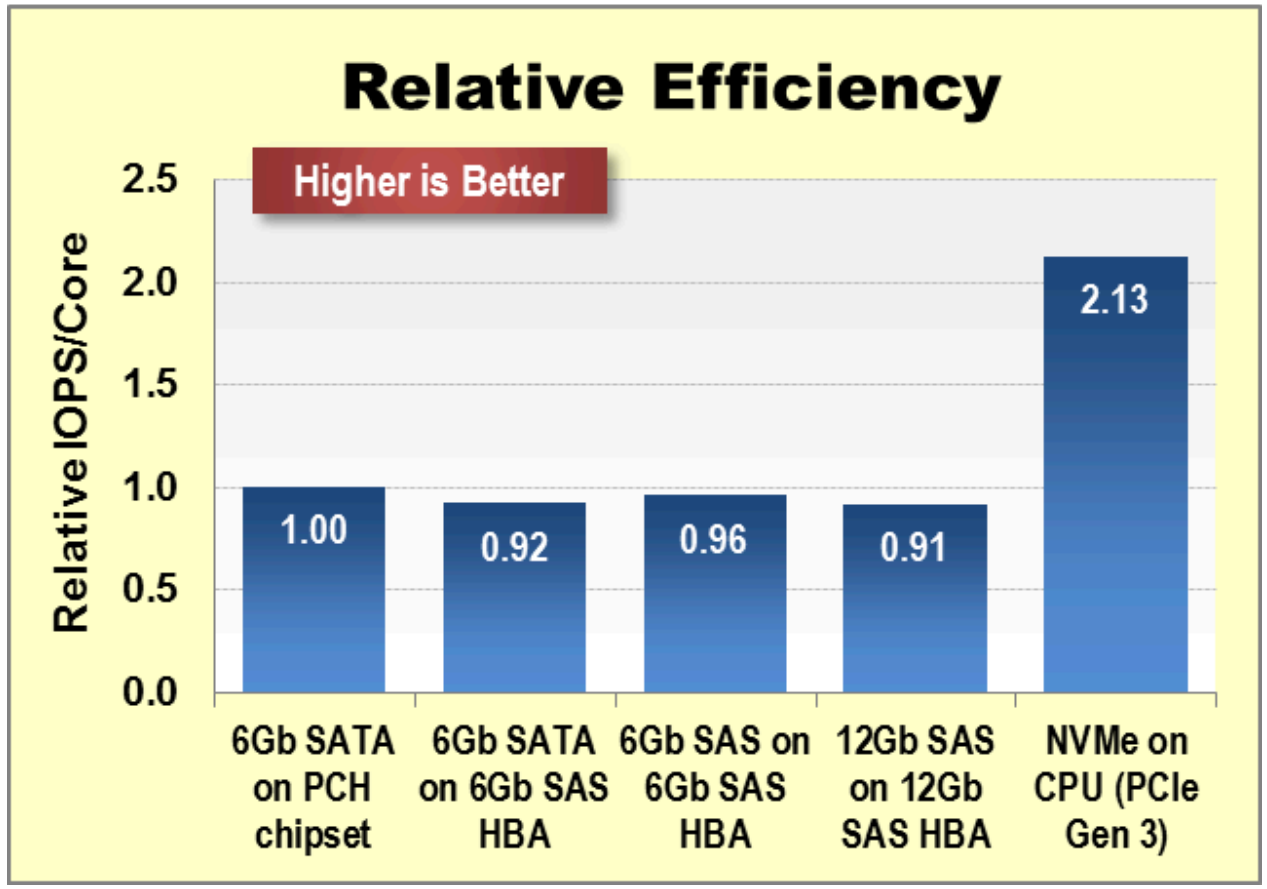
Architected for NVM
Eliminates hard drive legacy that
SATA and SAS are based upon

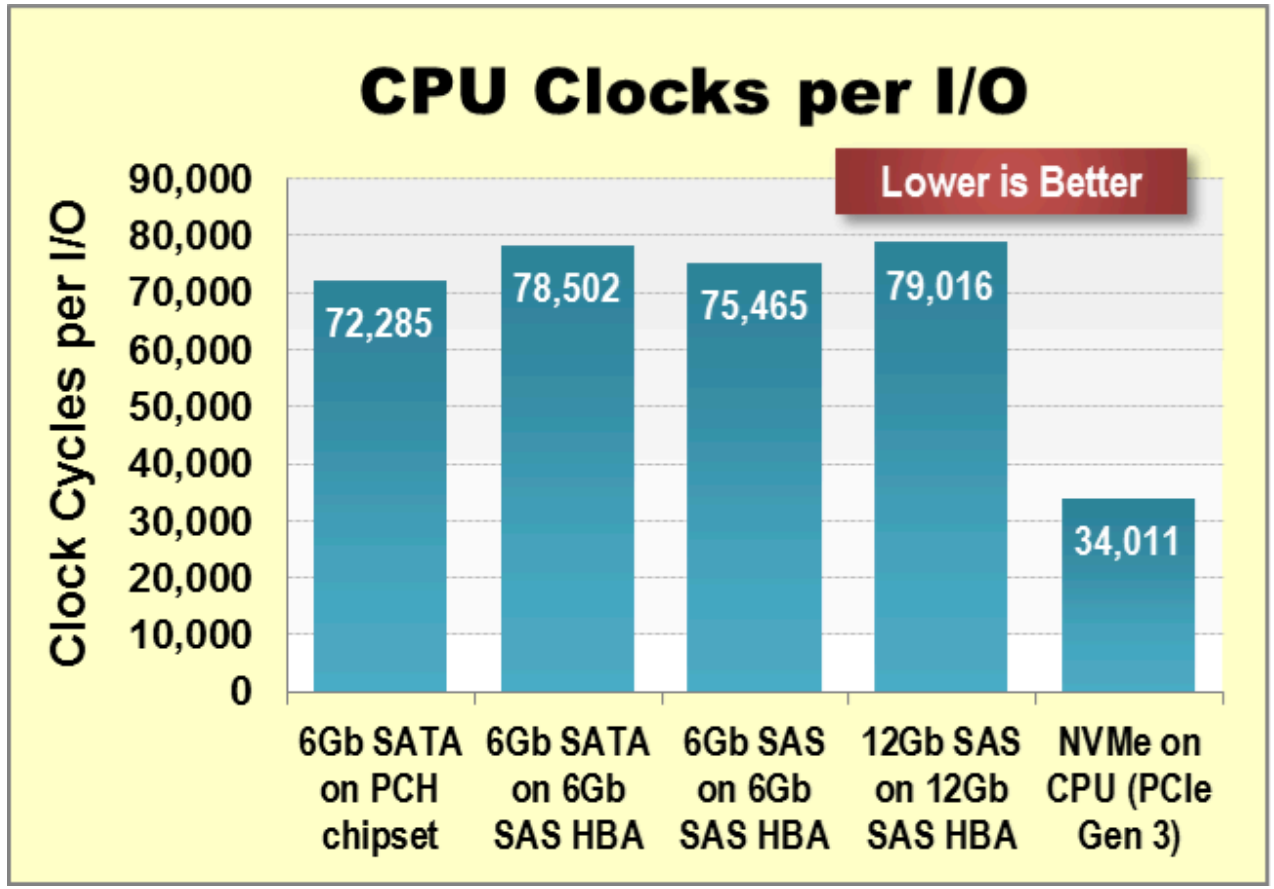
Works Out of the Box
Drivers for all major
operating environments

Efficient & Scalable
Streamlined protocol & command set,
fewer clock cycles per I/O

Optimized for
Next Generation NVM
New storage stack with low
latency and small overhead

Is NVMe Fast?



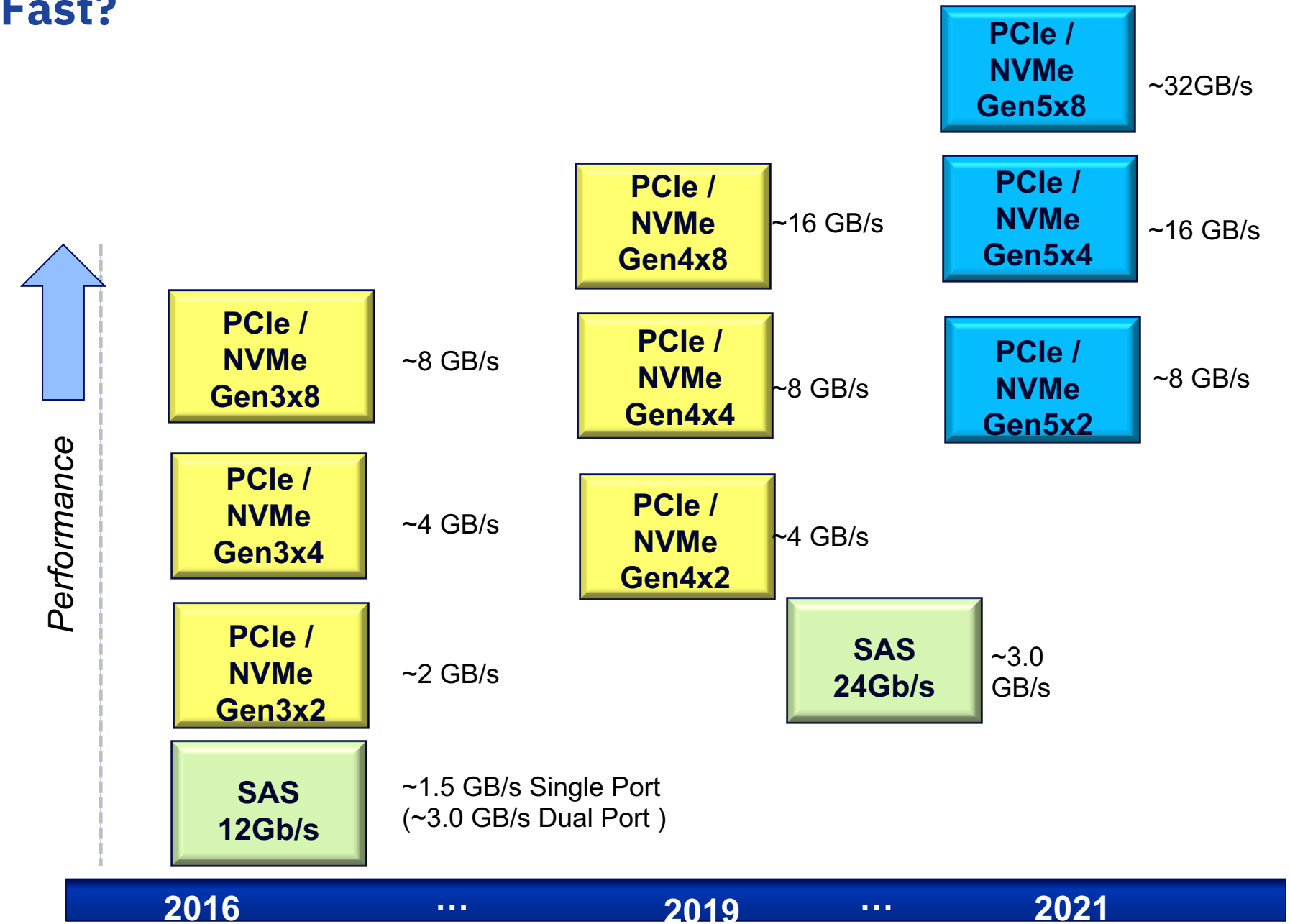


Is NVMe Fast?

Latency	Bus	Media	Read Lat. (us)	Write Lat. (us)	Read (IOPs)	Write (IOPs)	Read Tp (GB/s)	Write Tp (GB/s)	Approx. $ Scale (2Q '19)	Cost
	Memory (in CEC)	DRAM	<1	<1	Not a Persistent Storage				20x	
		3DXP	<1	<1	Persistent Storage				15x	
		LL Flash (HMS)	TBD	TBD						
	PCIe (NVMe)	3DXP	<10	<10	550K	500K	2.4	2.0	10x	
		LL Flash	<20	<20	800K	240K	3.5	3.4	1.1x	
		Flash	<90	<25	1500K*	250K	6.4	3.8	1x	
		QLC Flash	>150	>60	80K	25K	2.0	0.6	0.5x	
	SAS	Flash	150	60	420K	50K	2.2	1.6	0.7 to 1x	
	SATA	Flash	1.8ms	3.6ms	93K	25K	0.5	0.5		
	NL-SAS / SATA	HDD	>ms	>ms	200	200	0.15	0.15	0.1x	
		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, Gen 3 x4. NVMe QLC Client Flash are based on Intel 660p M.2, Gen 3 x4.
NVMe/Flash*: PCIe Gen 4 x4 U.2 devices. Gen 4 x8 Add In Card will have additional bandwidth and IOPS performance.
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

Is NVMe Fast?

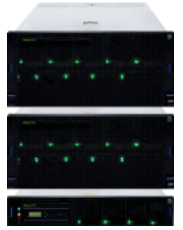


NVMe on IBM POWER9 Systems Hardware

- NVMe is able to provide higher performance than SSDs. NVMe technology can provide significantly more read or write IOPS and significantly larger throughput (GB/sec) compared to SAS/SATA SSDs. Actual system or application performance differences will vary by customer and workload.
- NVMe provides additional virtualization capabilities since 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. Best practice would be to only order in pairs to make mirroring the most straightforward.
- 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 be mirrored to NVMe and SAS drives can only be mirrored to SAS drives.
- Hot spare is not supported in IBM i, however an extra NVMe could be on the system as a cold spare to speed up the repair process, and it is only a spare in the fact that a customer doesn't have to order/plug it in. IBM i development is aware of the desire for something more capable than a cold spare.

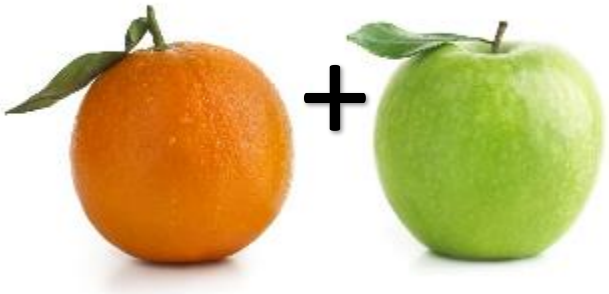
NVMe on IBM POWER9 Systems Hardware

- NVMe are only supported in the system unit of POWER9 servers. They are not supported in a PCIe Gen3 I/O drawer due to bandwidth and potential performance issues.
- S914 supports up to 3 NVMe. S924 or S922 supports up to 5 NVMe (3 with only one processor module).
- S914 4-core P05 system is limited to 2 x 1.6TB devices only. No mixing of NVMe and SAS drives is allowed on the 4-core (ten maximum of SAS drives or two maximum of NVMe).
- E950 supports up to 9 NVMe PCIe addin cards and up to 4 U.2 7mm NVMe devices per server
- E980 supports up to 8 NVMe PCIe addin cards per drawer (6 first drawer, 8 each drawer 2, 3, and 4 for a maximum total of 30) and up to 4 U.2 7mm NVMe devices per drawer
- Note that using multiple PCIe card NVMe will affect available I/O slots which can affect system configurations with multiple partitions and using external drawers may affect the maximum amount of PCIe NVMe.



NVMe on IBM POWER9 Systems Hardware

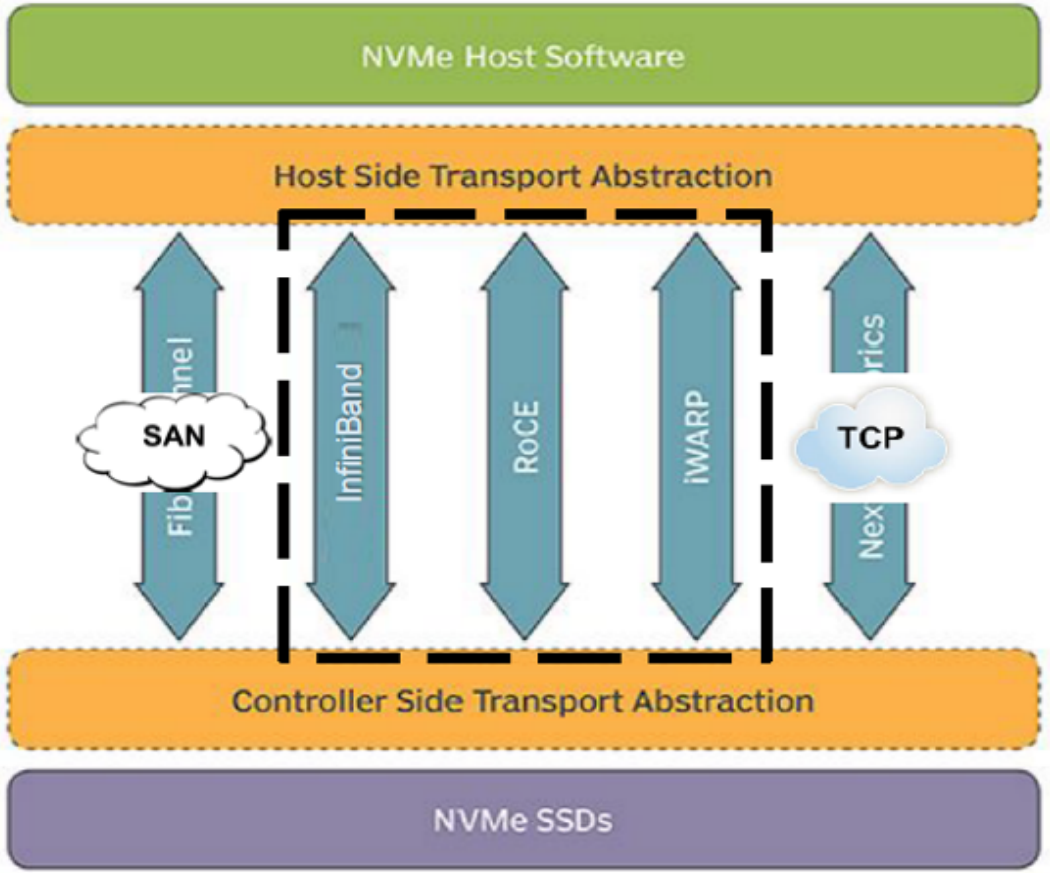
- Internal HDDs/SSDs are allowed in the system unit “with” PCIe card NVMe (except the S914 4-core with IBM i).



- A storage backplane is “not” required when PCIe card NVMe adapters but may be nice to have for future storage expansion.

What is NVMe Over Fabrics? (NVMeoF)

•Reliable, credit-based flow control and delivery mechanisms.
•An NVMe-optimized client.
•A low-latency fabric.
•Reduced latency and CPU utilization adapters or interface cards.
•Fabric scaling.
•Multi-Host support.
•Multi-port support.
•Multi-path support.



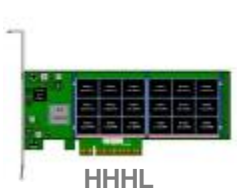
NVMe over Fabrics is poised to extend the low-latency efficient NVMe block storage protocol over fabrics to provide large-scale sharing of storage over distance.

NOTE: We support NVMe devices within IBM Storage and other brand's external devices that use Fibre Channel as the interconnect

NVMe Product Form Types used in POWER9 Systems

Add-in Card (AIC)

- x4 / x8 PCIe Gen3
- High performance and power (25-50 Watts)
- Focus on high capacity - up to 6.4TB currently
- Best performance and latency


HHHL


FHHL

Feature Code	PCIe3 Card NVMe Flash Adapter		
	1.6TB	3.2TB	6.4TB
AIX/Linux (LP/FH)	#EC5G / #EC5B	#EC5C / #EC5D	#EC5E / #EC5F
IBM i (LP/FH)	#EC6U / #EC6V	#EC6W / #EC6X	#EC6Y / #EC6Z

U.2

- x2 / x4 PCIe Gen3
- Higher power required to achieve max SSD performance
- Expect to be direct attached to CPU vs through a HBA

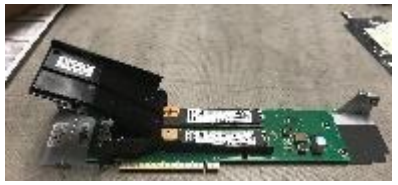

2.5” 7mm

E950/E980
NVMe
800GB - #EC5J
1.6TB - #EC5K
3.2TB - #EC5L

M.2

- Withdrawn from marketing 3/24/2020, announcement 920-065
- x2 / x4 PCIe Gen3
- Different physical sizes
- Power less than 9 watts
- Limitations: Minimal Connector Plug Capability (<10), No hot swap
- Targeted as a replacement for SATA SSDs

Card #EC59 with 1-2 #ES14



S9xx
NVMe
400GB #ES14

What Are My Datacenter NVMe Options

Hardware Features

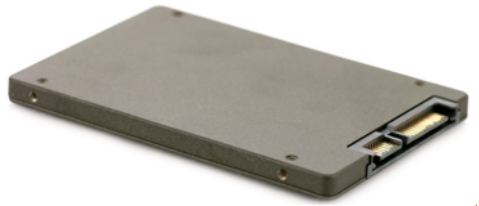
- NVMe Specs. 1.2.1 Compliant
- PCIe Gen 3 x 4
- SFF U.2 (2.5”) 7mm z-Height
- Power < 12W
- Block Size 4096(Default), 512
- Non Volatile Write Buffer
- Endurance 2.4 DWPD
- Power Loss Protection
- RAIF: Tolerant of single flash die failures
- PCIe Vendor VPD Support
- Boot: Option ROM BAR 128KB
- Warranty 5 years
- Hot Plug capable
- MTBF ≥ 2 million hours
- End Of Life Data Retention ≥ 3 months
- EEH Support
- Live Firmware Update
- NVMe-MI Support

	PCIe3 800GB NVMe U.2 Slim SSD		
	800GB	1.6TB	3.2TB
FC	EC5J	EC5K	EC5L

Workload	Target (800GB)	Target (1.6/3.2TB)
Read (IOPS)	380K	500K
Write (IOPS)	60K	80K
Mixed R/W (70/30)	135K	150K
Read Data Tp (GB/s)	1.6	2.6/3.2
Write Data Tp (GB/s)	1.05	1.9
Read Latency (us)	80	80
Write Latency (us)	20	20

Software Support

- AIX, VIOS
- Linux



Supported POWER9 Platforms

- E950
- E980

Notes:
1. IOPs and Latency #’s on random 4K
2. Data throughput #’s are on sequential 256KB workload

What Are My Enterprise NVMe Options

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
- RAIF: 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 1.6TB NVMe Flash Adapter III x8		
	1.6TB	3.2TB	6.4TB
FC (LP/FH)	EC5G / EC5B	EC5C / EC5D	EC5E / EC5F
IBM i FCs (LP/FH)	EC6U / EC6V	EC6W / EC6X	EC6Y / EC6Z

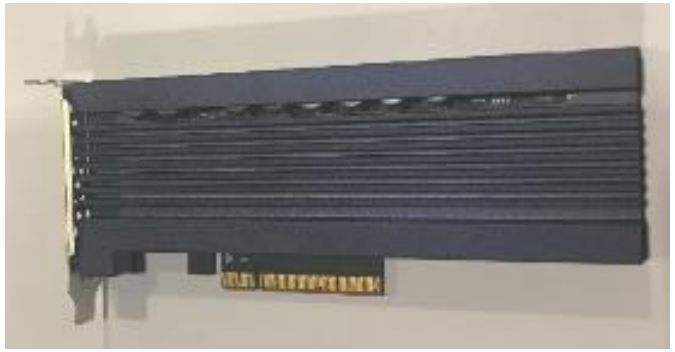
Performance

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)

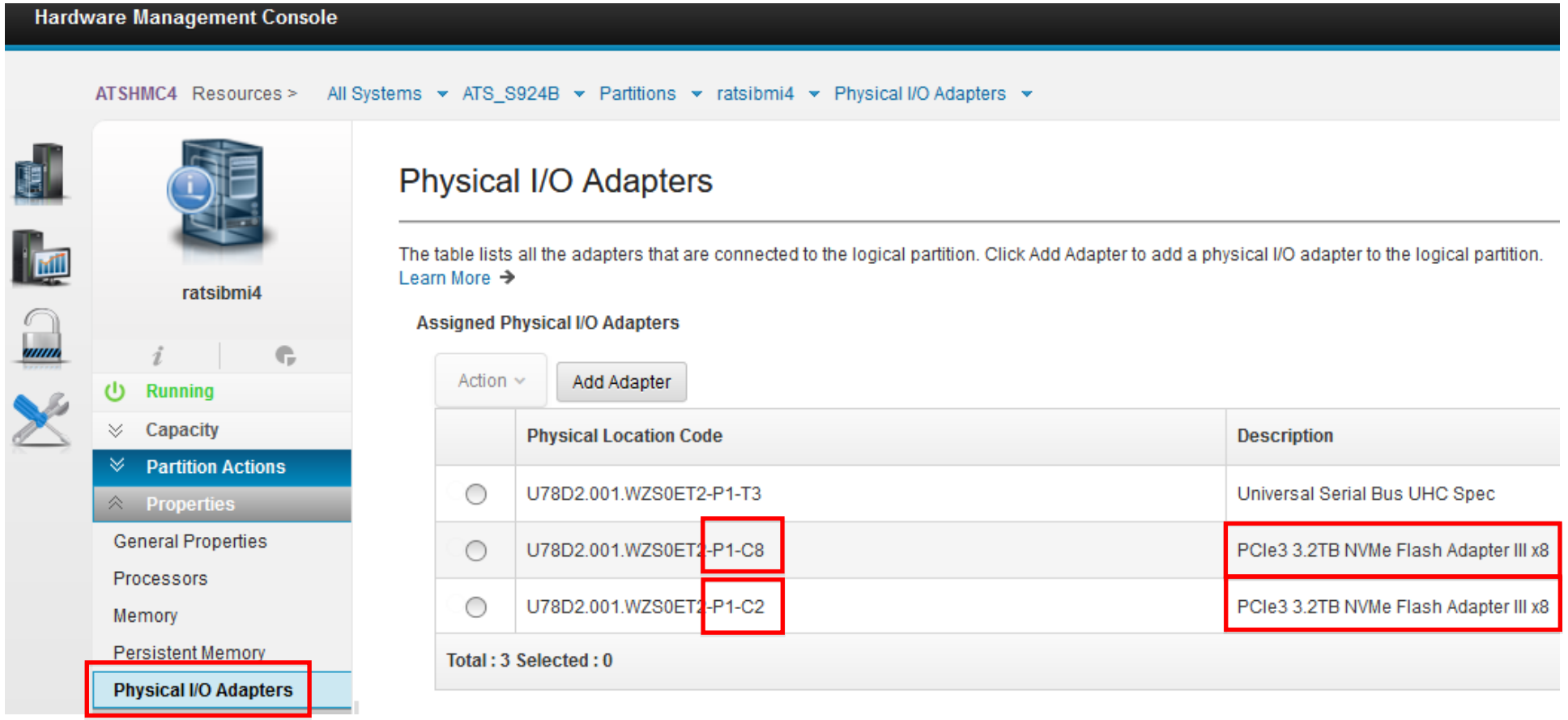


PCIe Card NVMe and IBM i

- IBM i supports virtualized NVMe (only for the PCIe Add In Card) via VIOS and requires the use of the VIOS LVM (Logical Volume Manager). This supports IBM i 7.2, 7.3 or 7.4. Since this has more "layers" between IBM i and the storage, it will not perform the same as native NVMe, thus would not be the recommended option if the best/most performance is required.
- PCIe card NVMe service/repair is similar to other PCIe card slot concurrent maintenance, but with extra steps such as described below:
  https://www.ibm.com/support/knowledgecenter/ssw_ibm_i_74/rzaly/rzalynvmefailure.htm
- Encryption is not supported today. IBM i development is aware of the desire for hardware encryption support.
- IBM i treats NVMe as the same tier as SSDs, so currently there isn't a way to tier (say using the Trace ASP Balance (TRCASPBAL) command) between them.
- NVMe devices are now supported (4/14/20 announce) as direct attached devices for IBM Db2 Mirror for i.

PCIe Card NVMe in an S9x4 Adapter Slot

PCIe Card NVMe IBM i HMC Screenshot

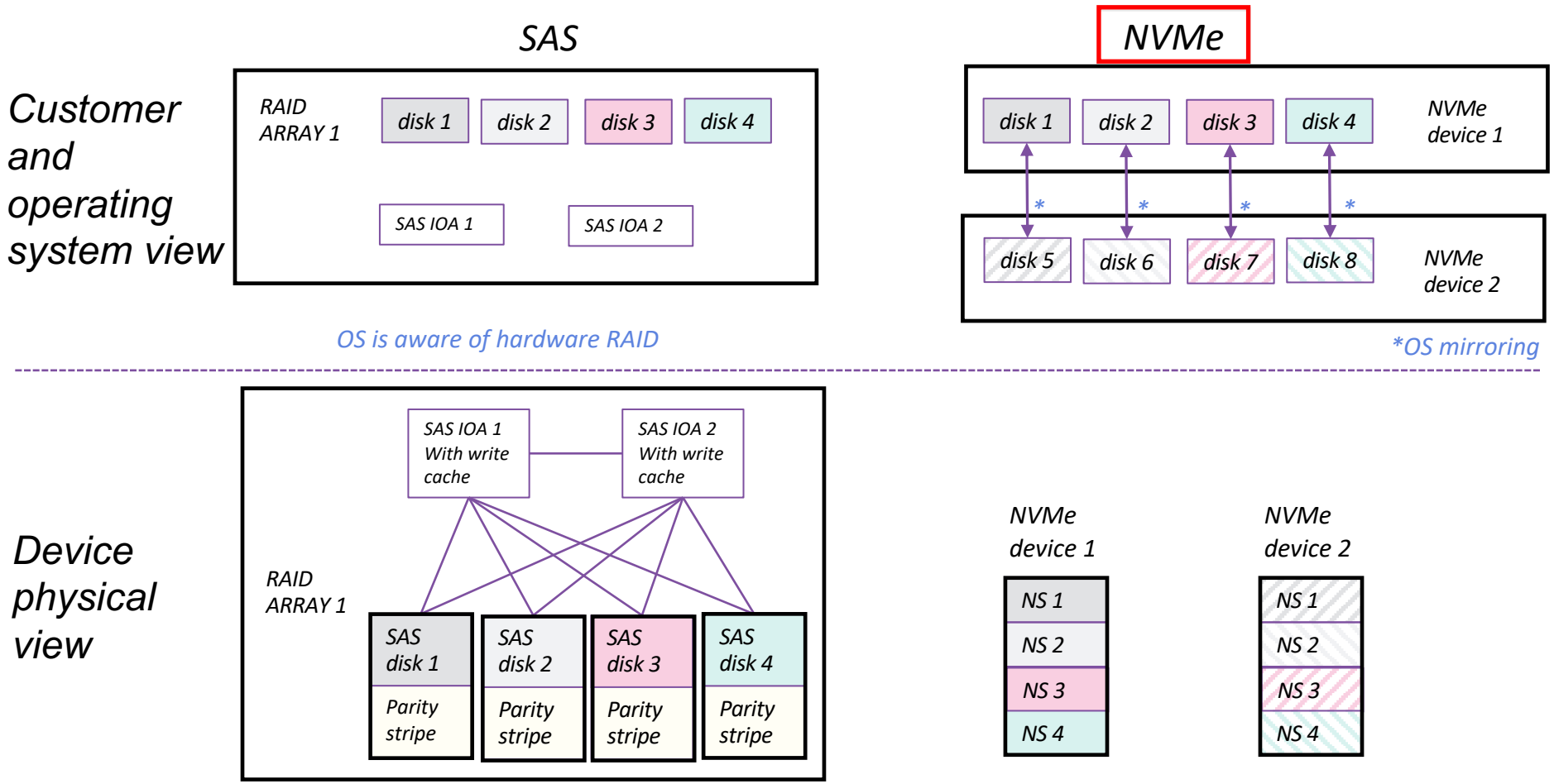


NOTE: A HMC server is not required to work with NVMe devices

PCIe Card NVMe Namespaces and IBM i

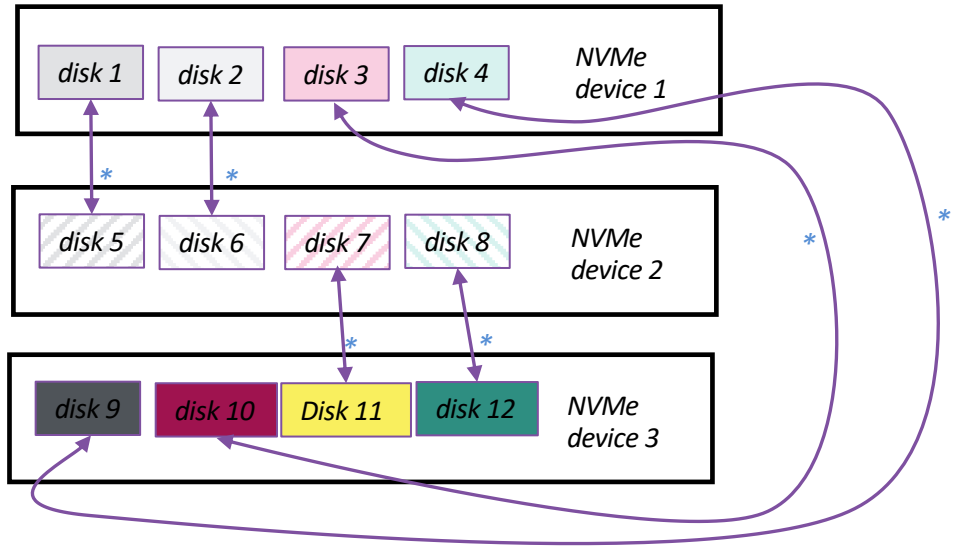
- NVMe architecture can use namespaces which is a collection of logical blocks whose logical block addresses range from 0 to the size of the namespace. A namespace ID (NSID) is an identifier used by a controller to provide access to a namespace.
- With NVMe, an ‘arm’ (logical unit) is a namespace. A 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.
- Note that the word “namespace” is used in the industry and by IBM i in different ways and in different contexts, so for example, a NVMe namespace should not be confused with and has nothing to do with ASP namespaces.

SAS Versus PCIe Card NVMe Storage with IBM i



IBM i PCIe Card NVMe Mirroring (3 NVMe Option 1)

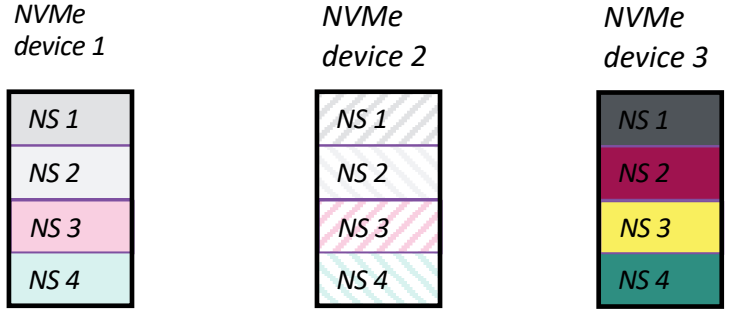
Customer and operating system view



All mirrored name spaces are the same size on all devices

**OS mirroring*

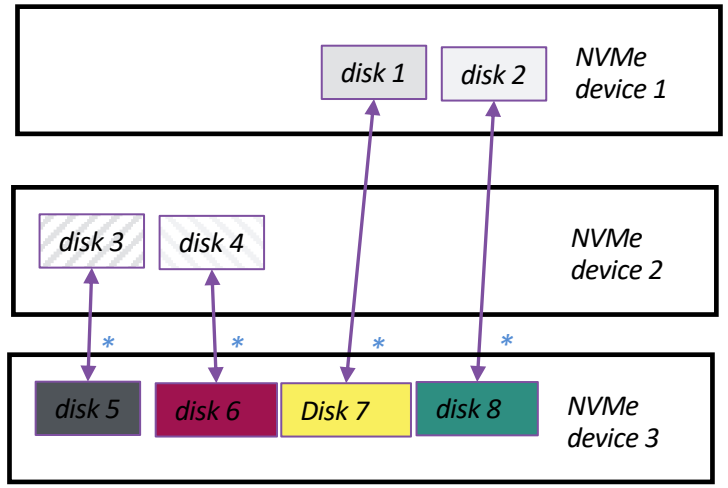
Device physical view



Device physical capacity does not have to be the same size

IBM i PCIe Card NVMe Mirroring (3 NVMe Option 2)

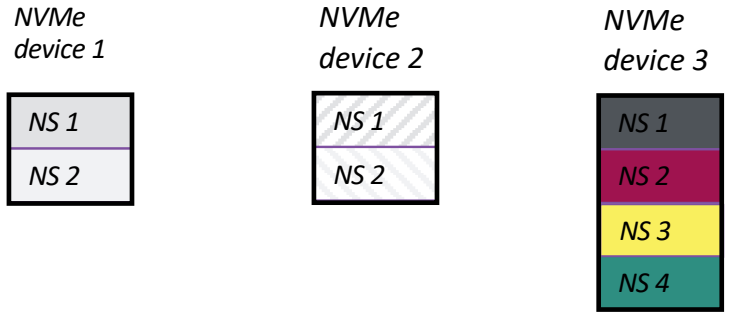
Customer and operating system view



All mirrored name spaces are the same size on all devices

**OS mirroring*

Device physical view



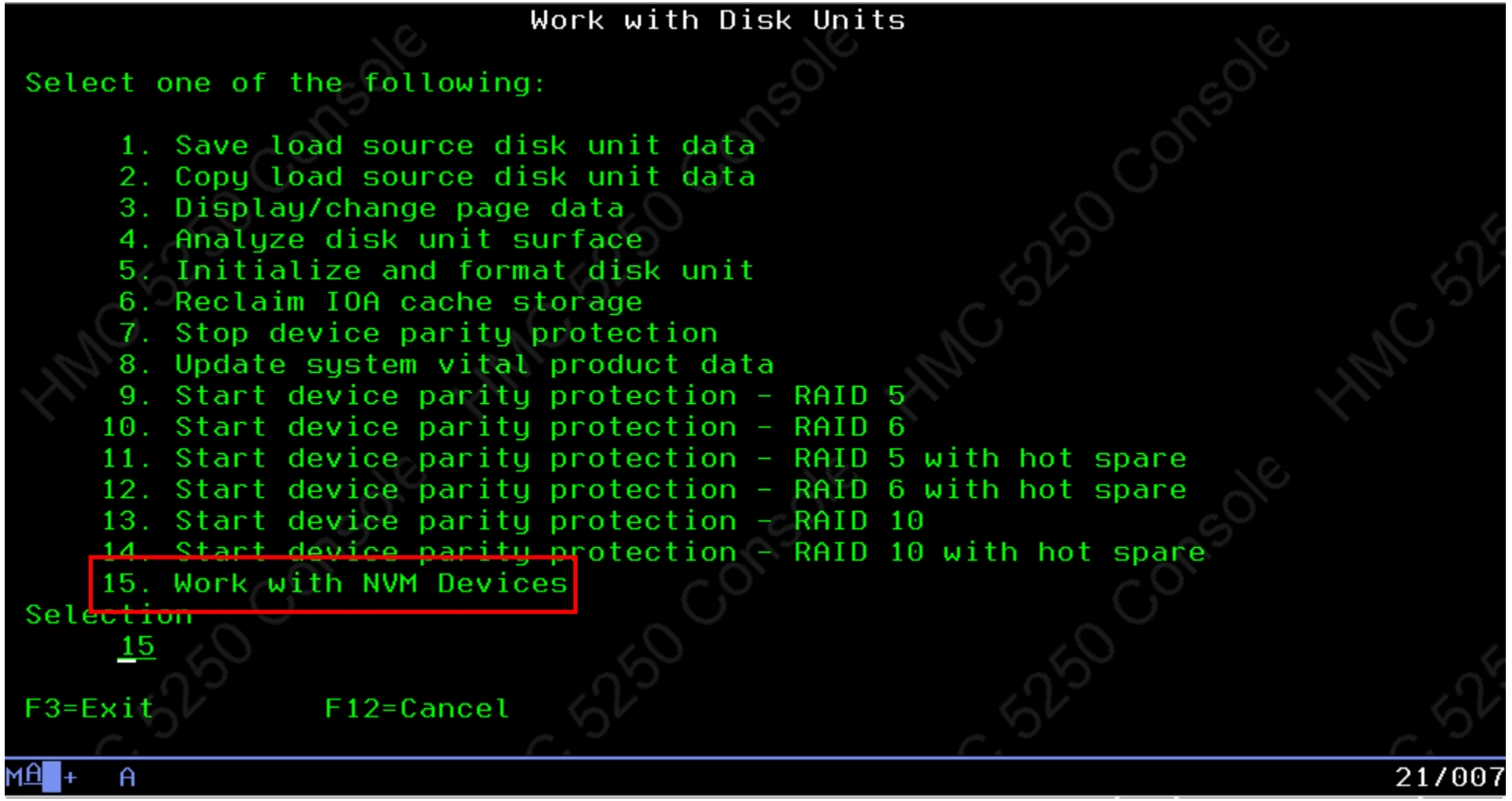
Device physical capacity does not have to be the same size

Recommended PCIe Card NVMe Namespace Sizes

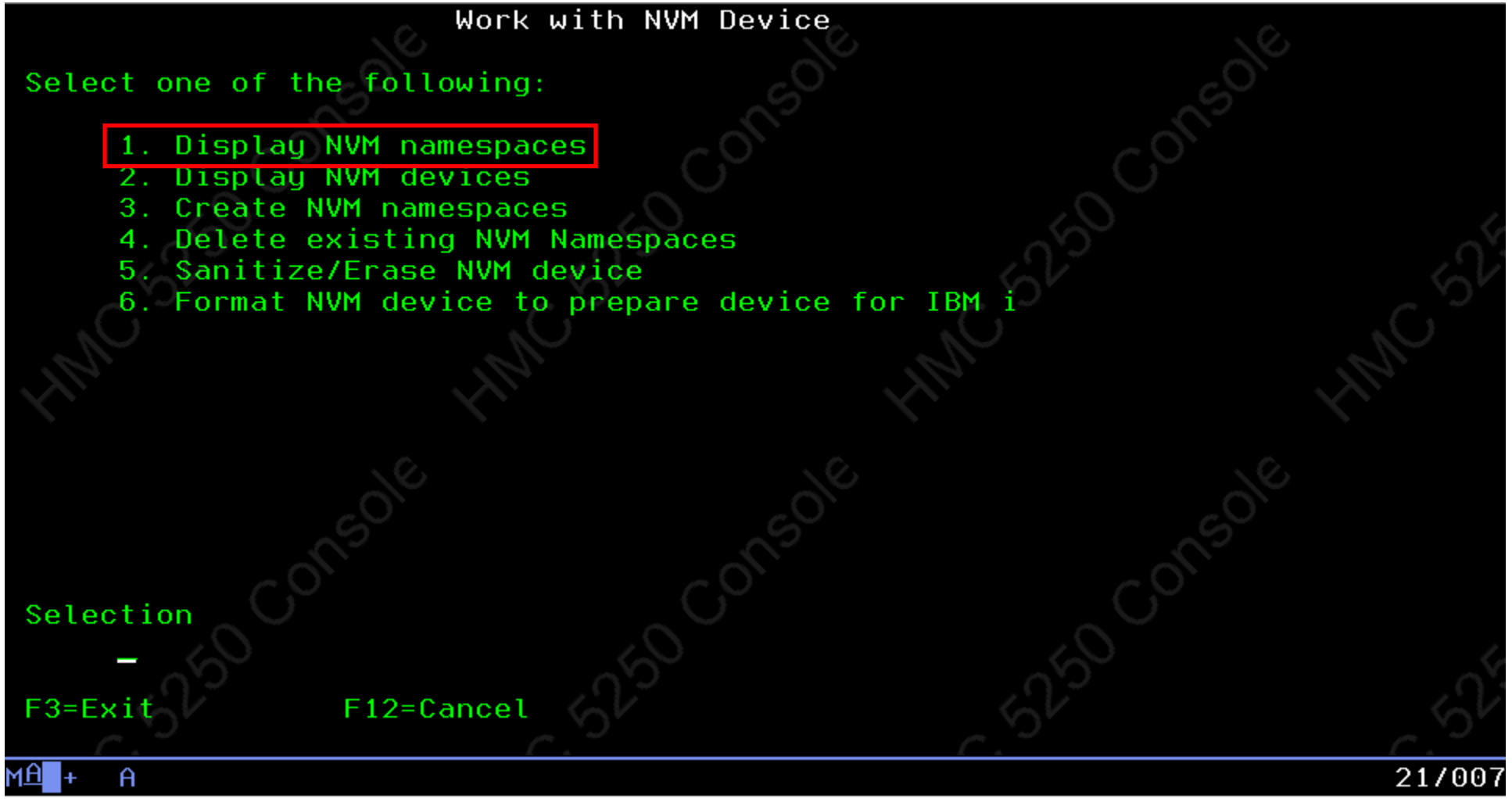
- First generation NVMe devices have a hardware boundary of 16 GB for name spaces. Device capacity can be wasted/lost if name spaces are not multiples of 16. The maximum number of namespaces on a device is 32.
- IBM i screens show Capacities in 'GB' (1000**3 (GB), not 1024**3 (GiB))
- IBM recommends namespaces of 188 GB or 393 GB (the e-Config default)
- Consider only using 393 GB. On the 6.4 TB device there is a lot of unused space when using 188 because the maximum number of namespaces is 32. The choice of namespace size is a balance between number of "arms" for (storage management) performance and other individual customer factors such as scaling/growth of the system.

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

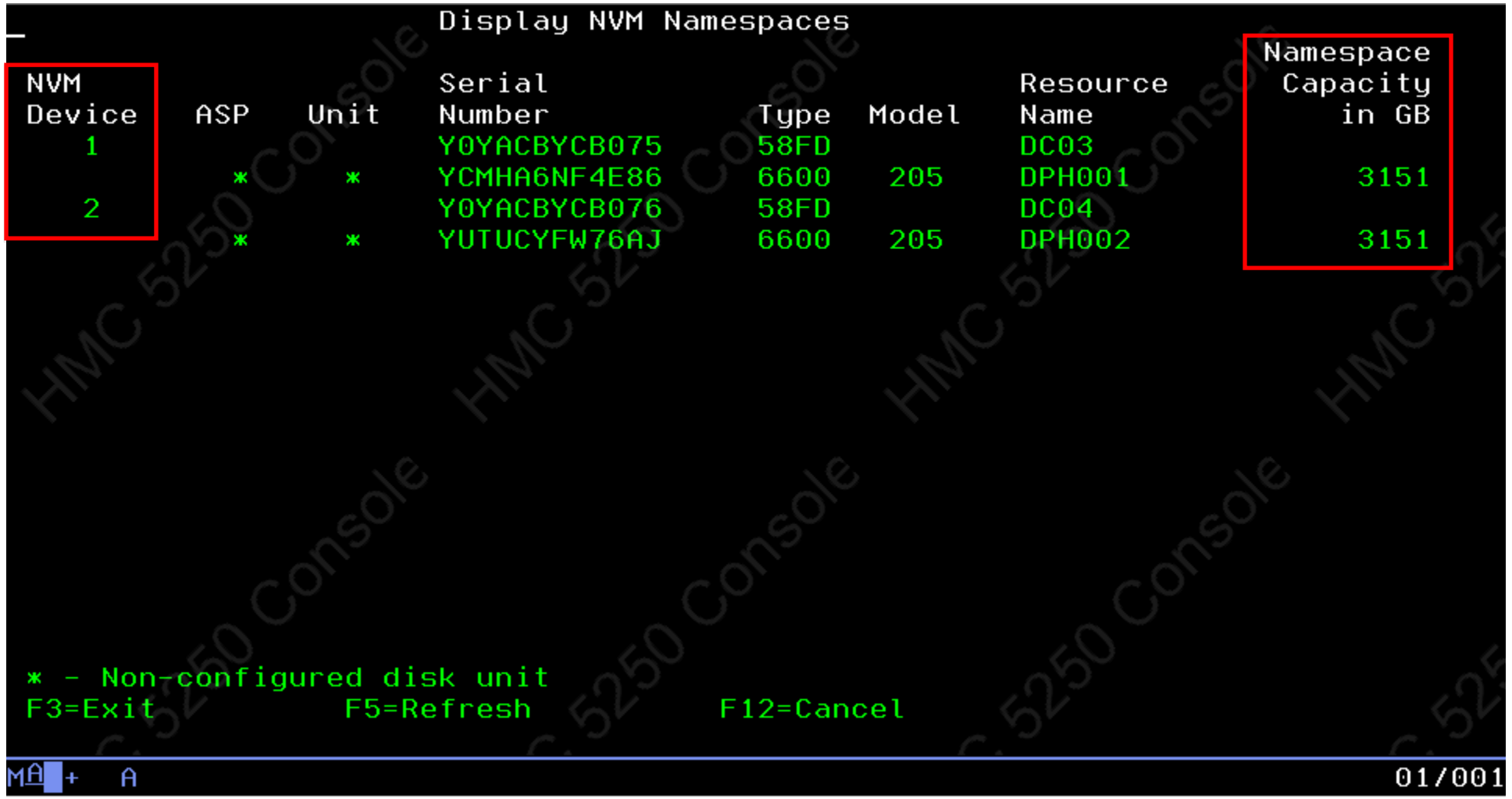
PCIe Card NVMe IBM i DST/SST Screenshots



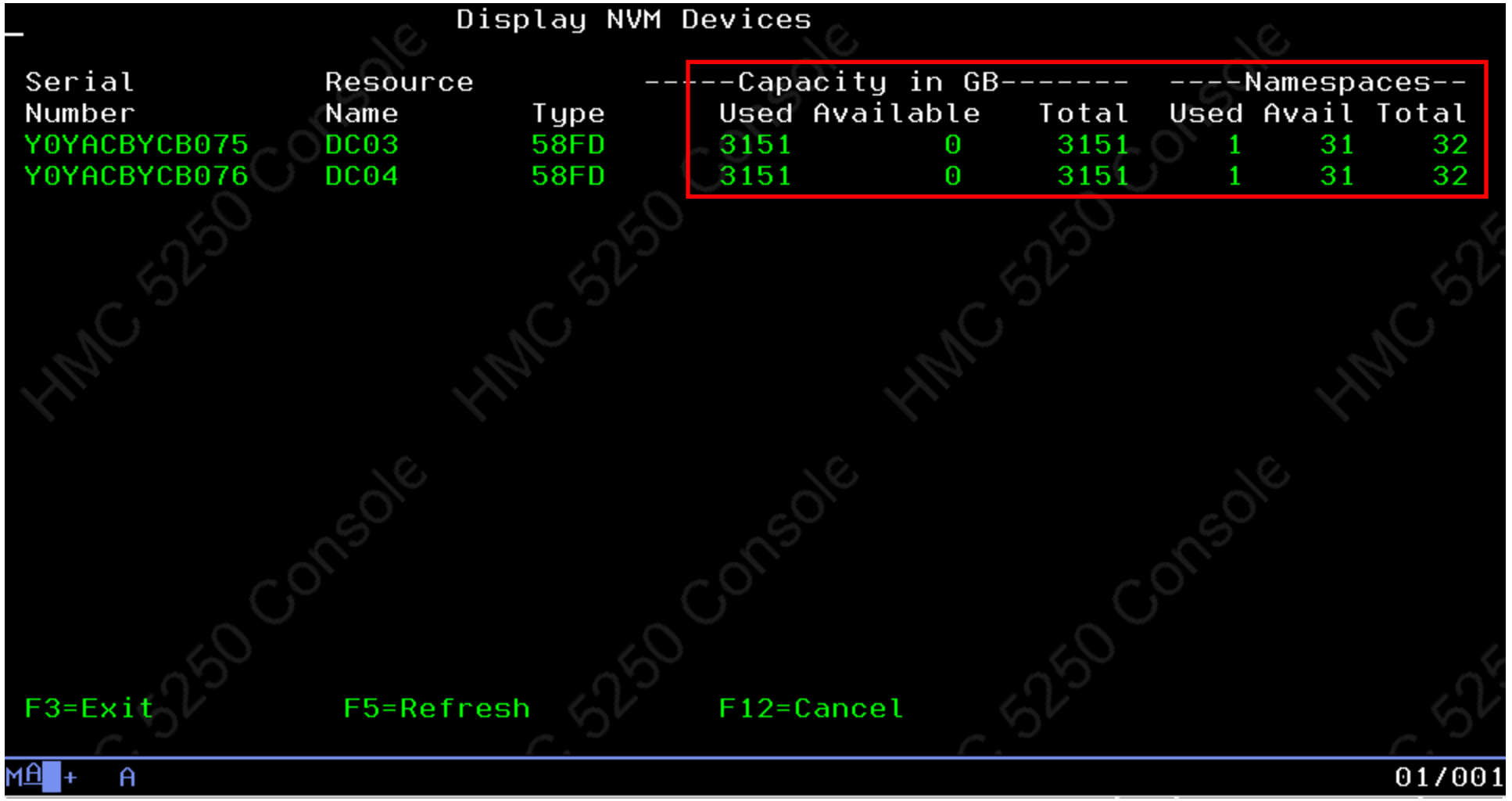
PCIe Card NVMe IBM i DST/SST Screenshots



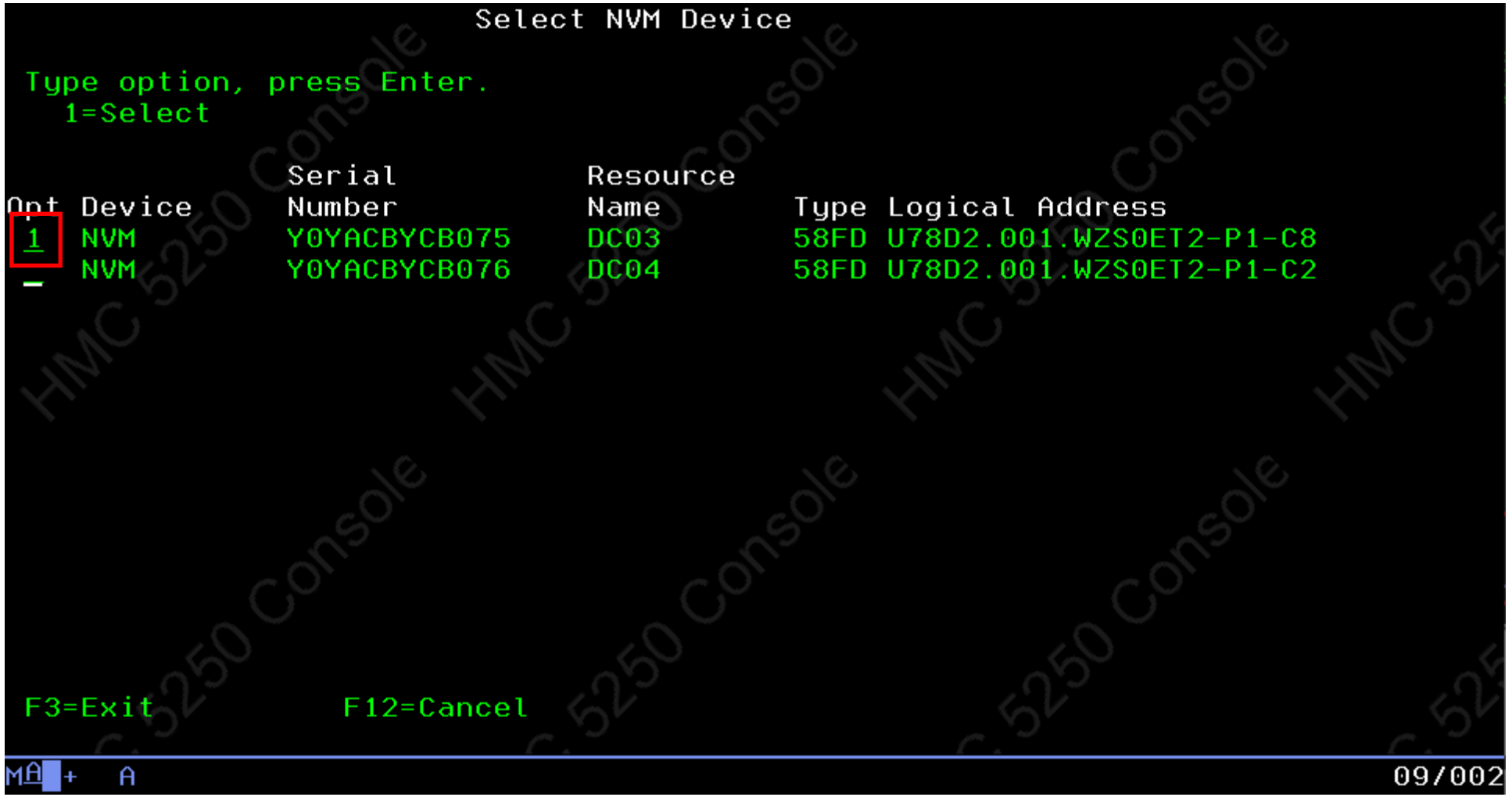
PCIe Card NVMe IBM i DST/SST Screenshots



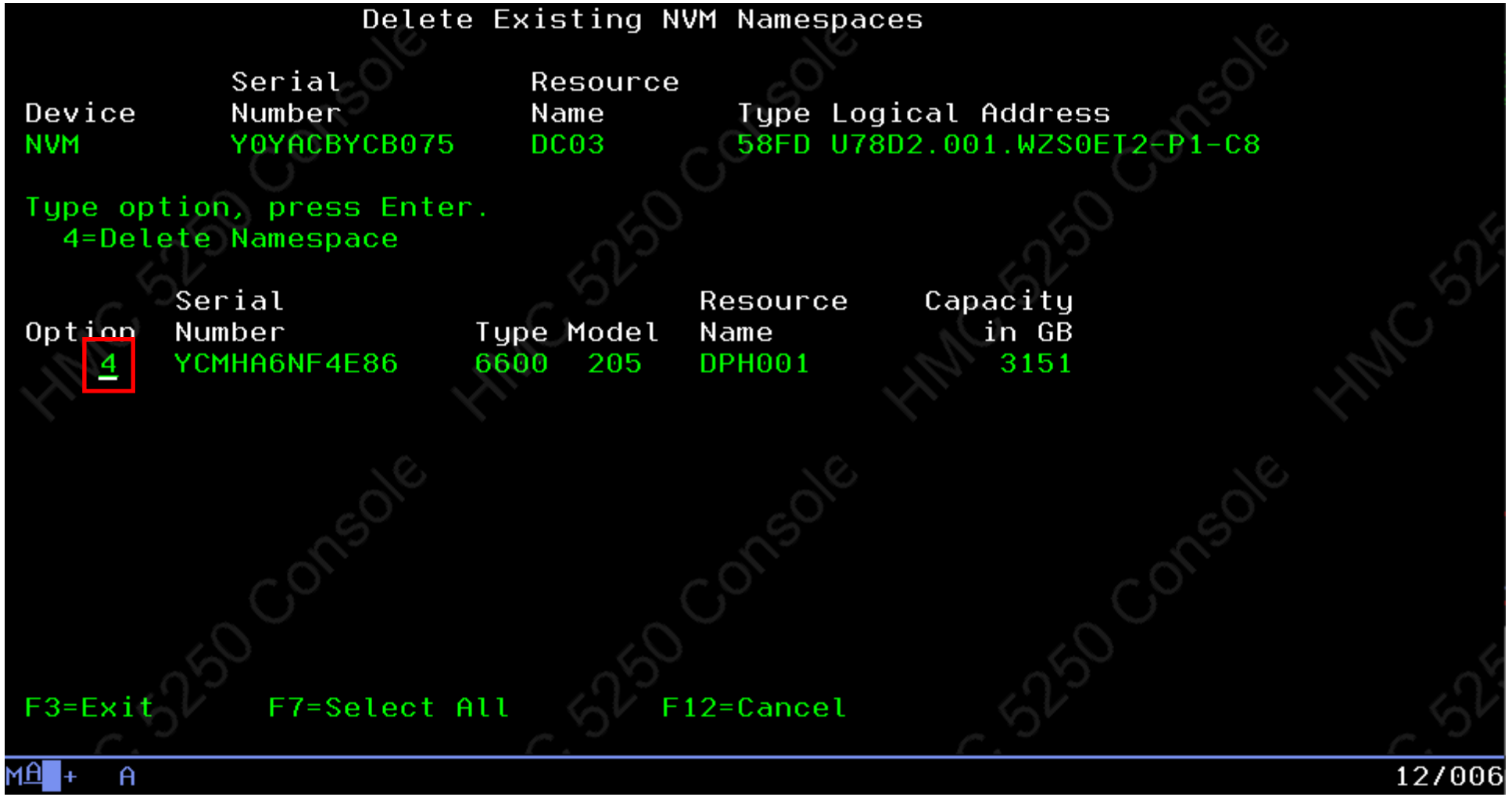
PCIe Card NVMe IBM i DST/SST Screenshots



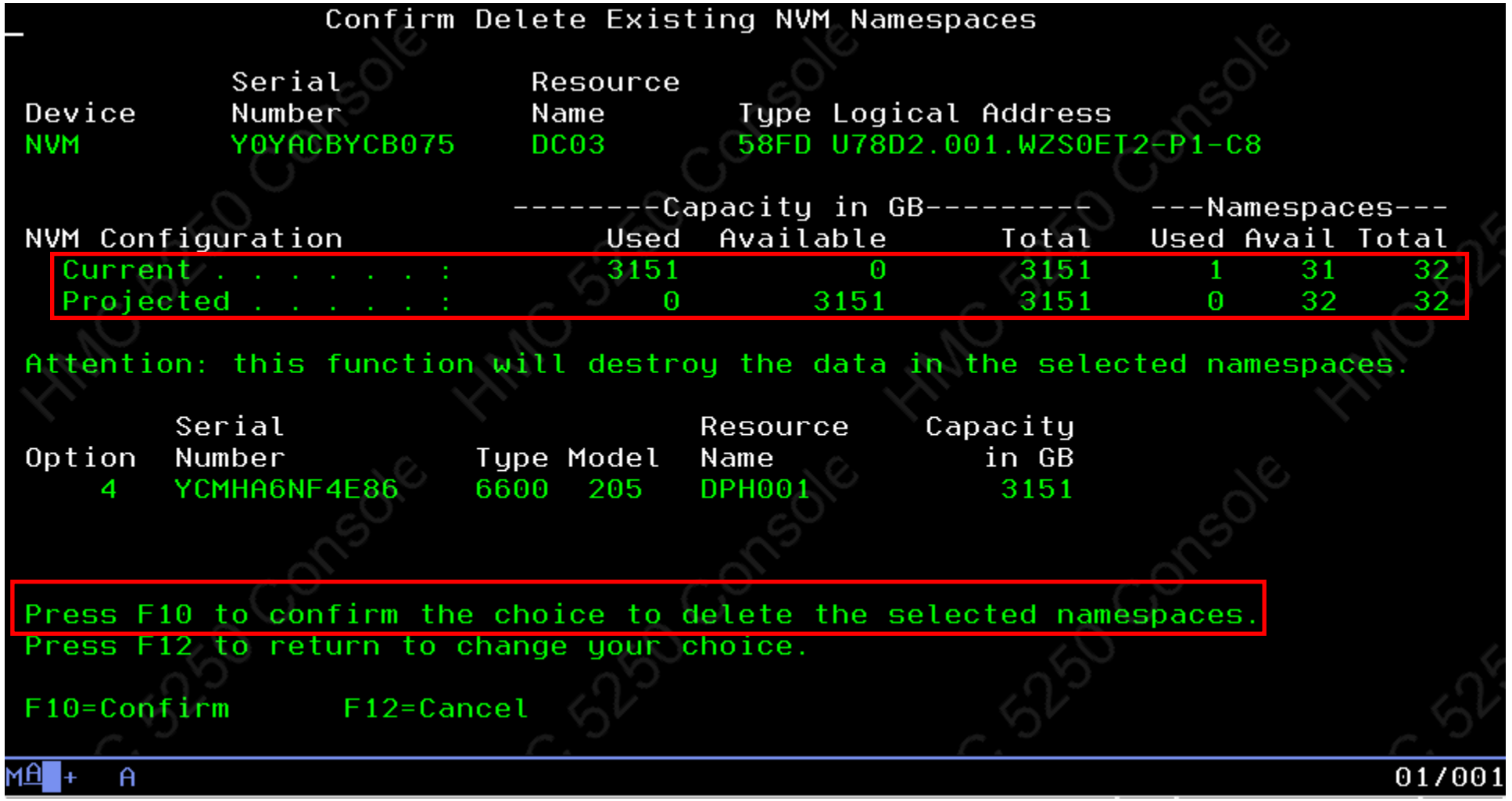
PCIe Card NVMe IBM i DST/SST Screenshots



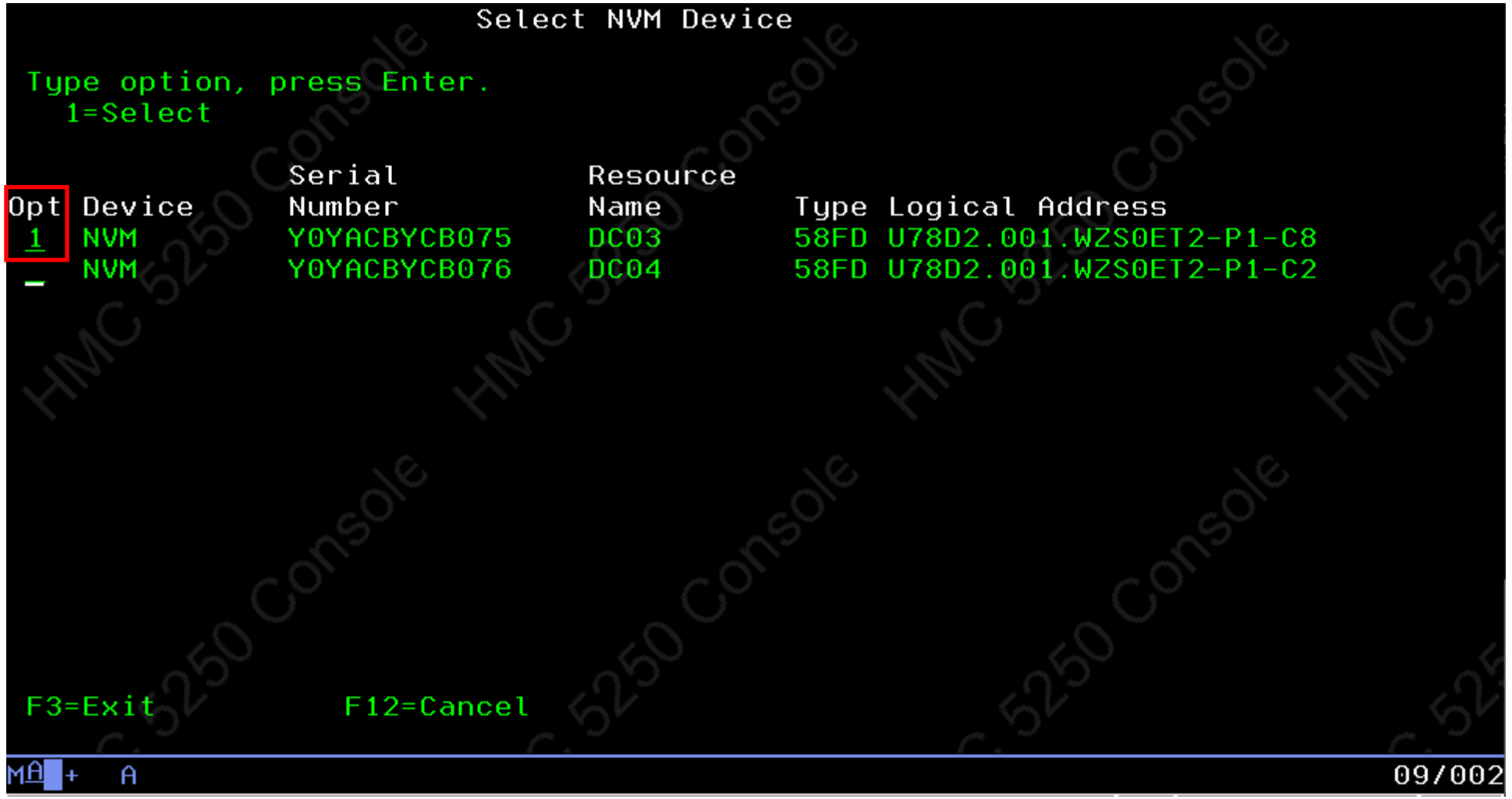
PCIe Card NVMe IBM i DST/SST Screenshots



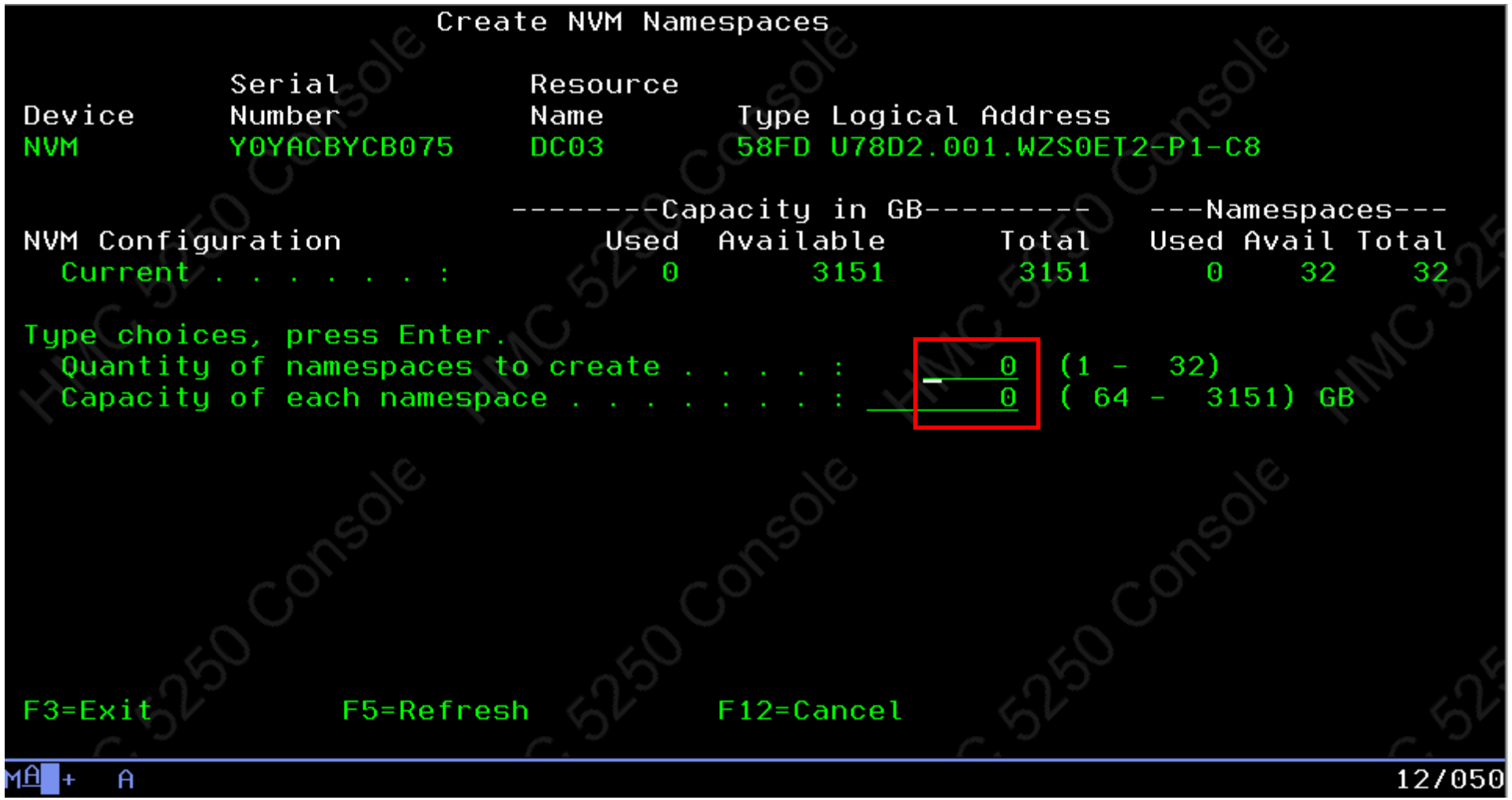
PCIe Card NVMe IBM i DST/SST Screenshots



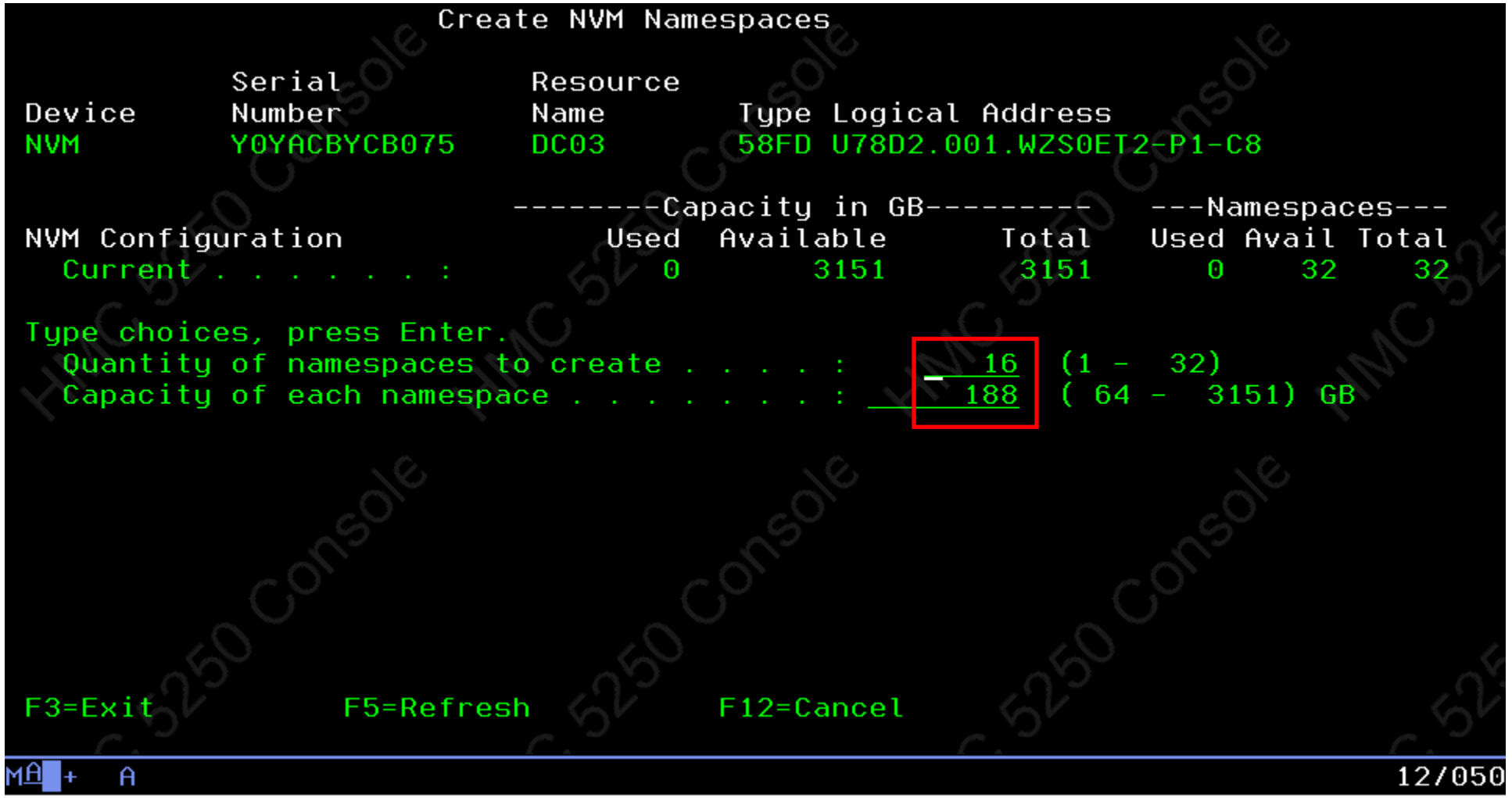
PCIe Card NVMe IBM i DST/SST Screenshots



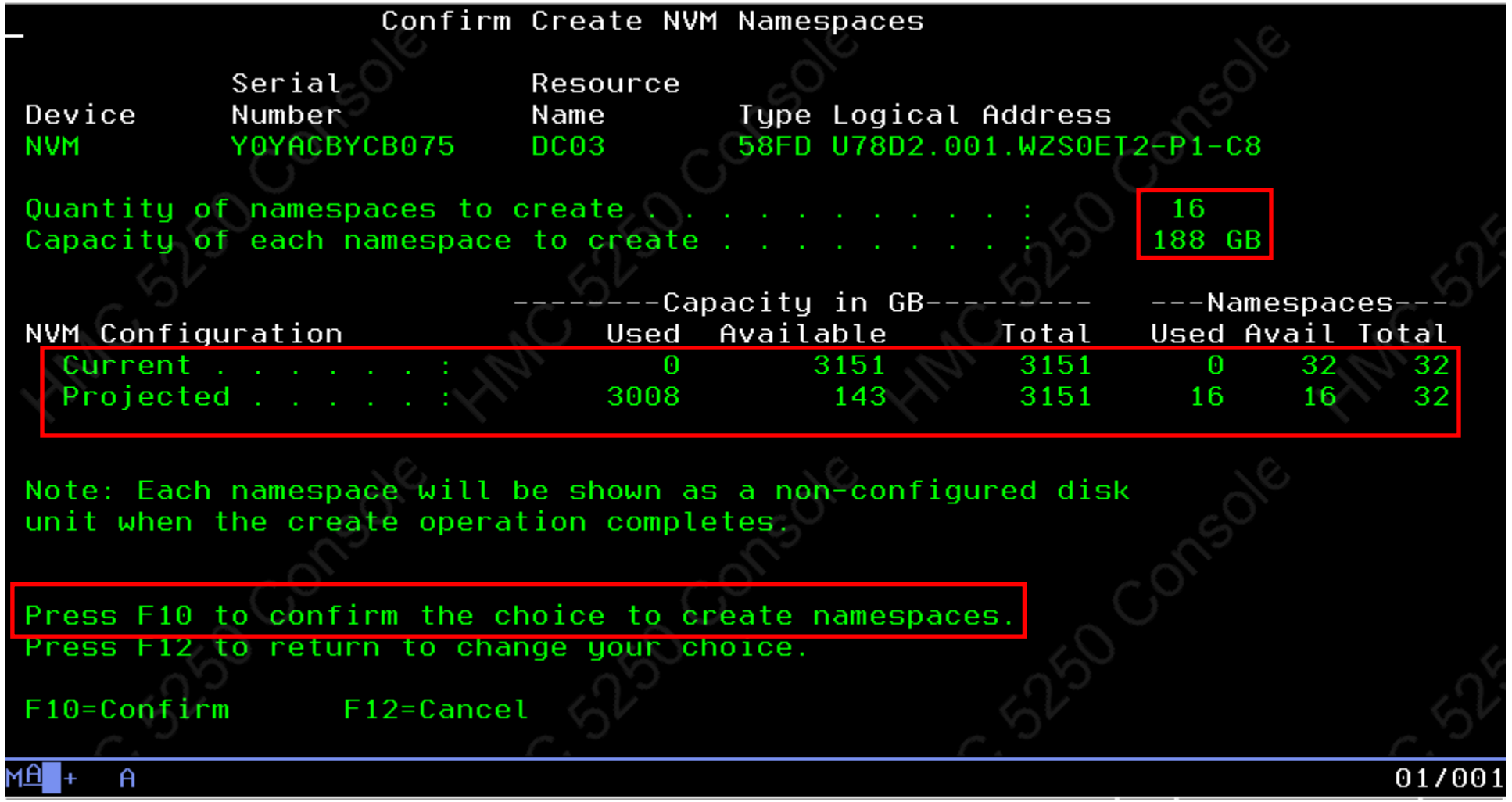
PCIe Card NVMe IBM i DST/SST Screenshots



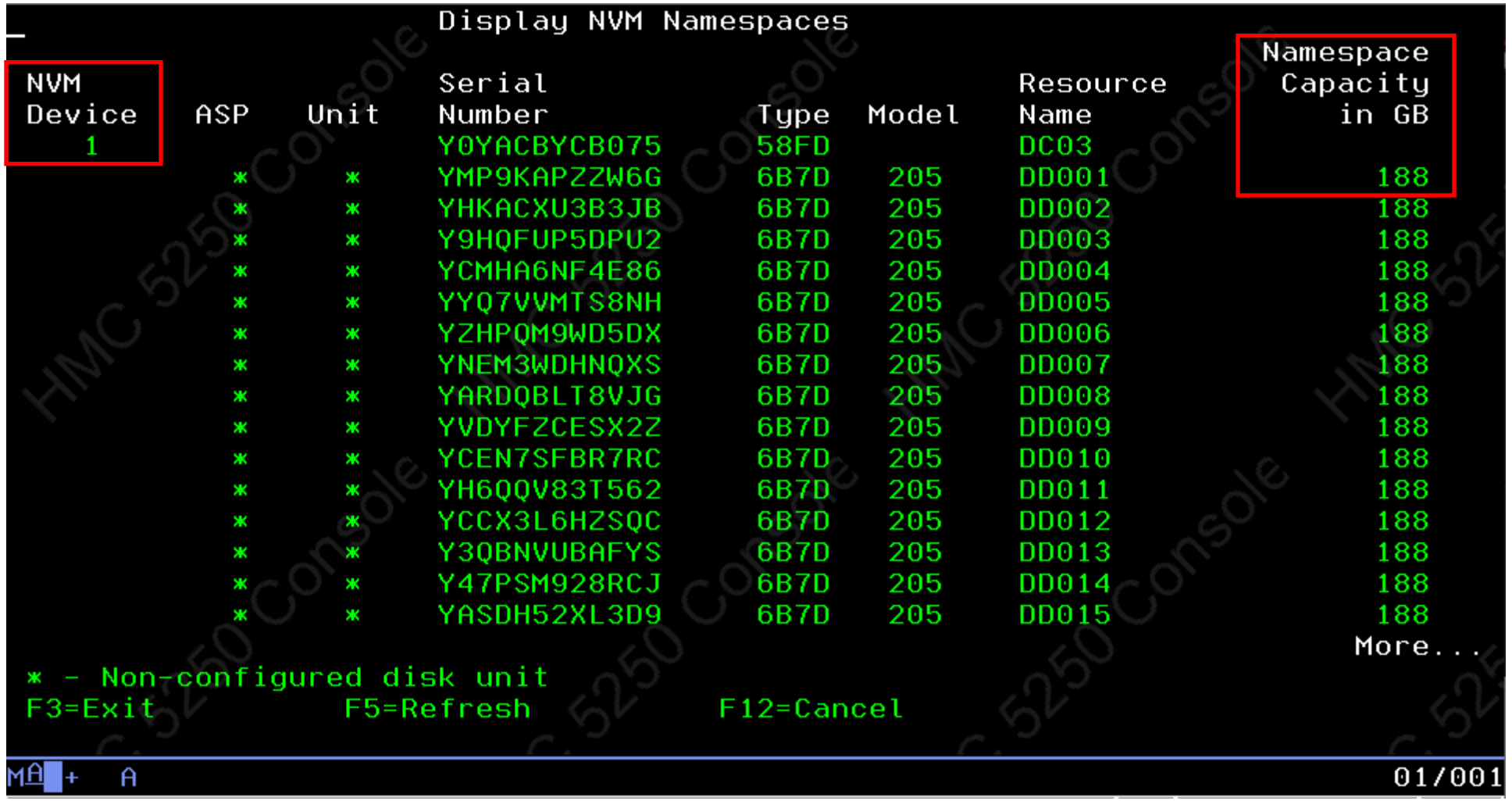
PCIe Card NVMe IBM i DST/SST Screenshots



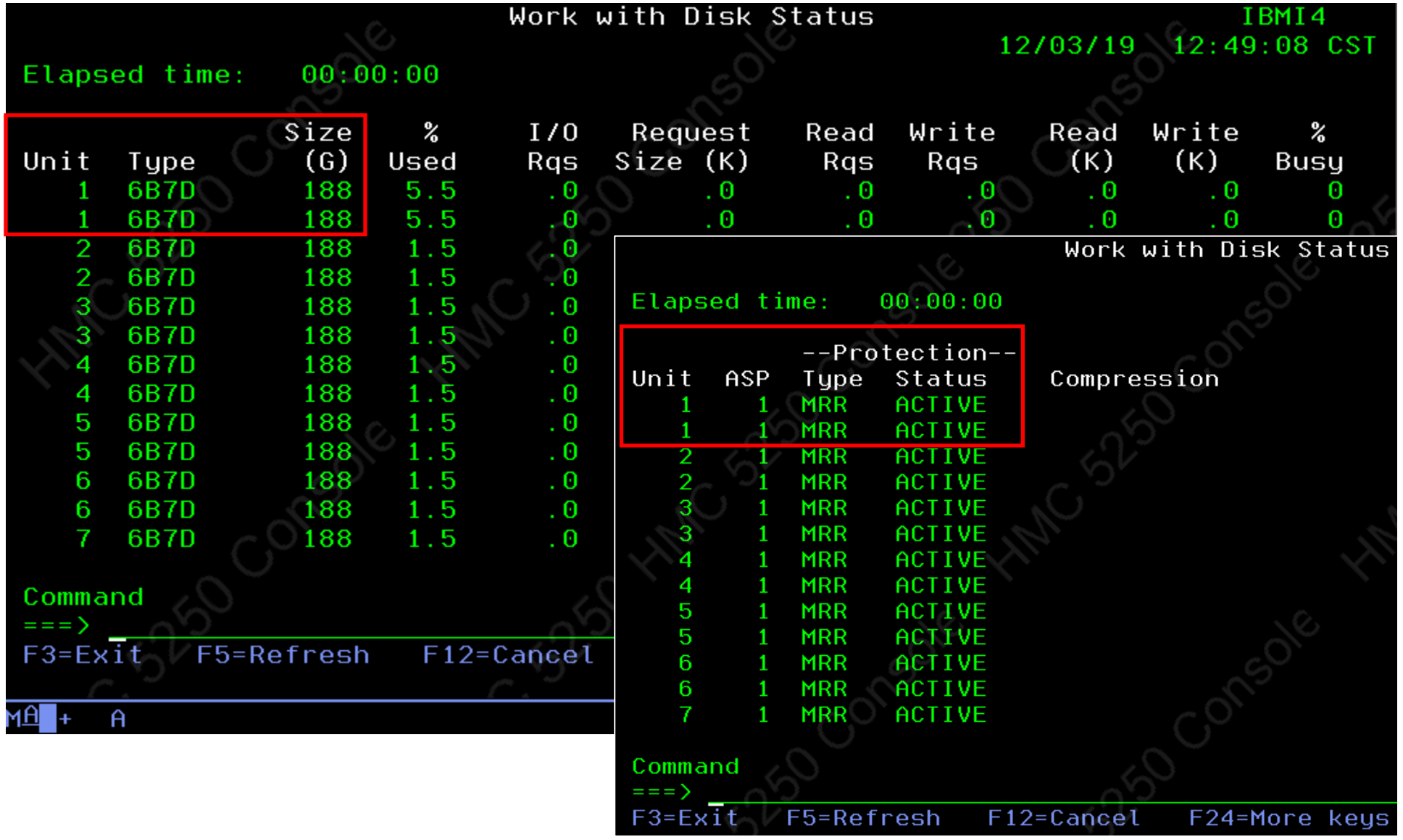
PCIe Card NVMe IBM i DST/SST Screenshots



PCIe Card NVMe IBM i DST Screenshots



PCIe Card NVMe IBM i WRKDSKSTS Screenshots



```
                                Work with Disk Status                   IBMI4
                                                          12/03/19  12:49:08 CST
Elapsed time:   00:00:00

                          Size    %      I/O    Request    Read   Write    Read   Write     %
Unit   Type               (G)  Used     Rqs  Size (K)     Rqs    Rqs     (K)    (K)   Busy
   1   6B7D               188   5.5      .0        .0      .0     .0      .0     .0      0
   1   6B7D               188   5.5      .0        .0      .0     .0      .0     .0      0
   2   6B7D               188   1.5      .0
   2   6B7D               188   1.5      .0
   3   6B7D               188   1.5      .0
   3   6B7D               188   1.5      .0
   4   6B7D               188   1.5      .0
   4   6B7D               188   1.5      .0
   5   6B7D               188   1.5      .0
   5   6B7D               188   1.5      .0
   6   6B7D               188   1.5      .0
   6   6B7D               188   1.5      .0
   7   6B7D               188   1.5      .0

Command
===>
F3=Exit    F5=Refresh    F12=Cancel

MA  +   A
```

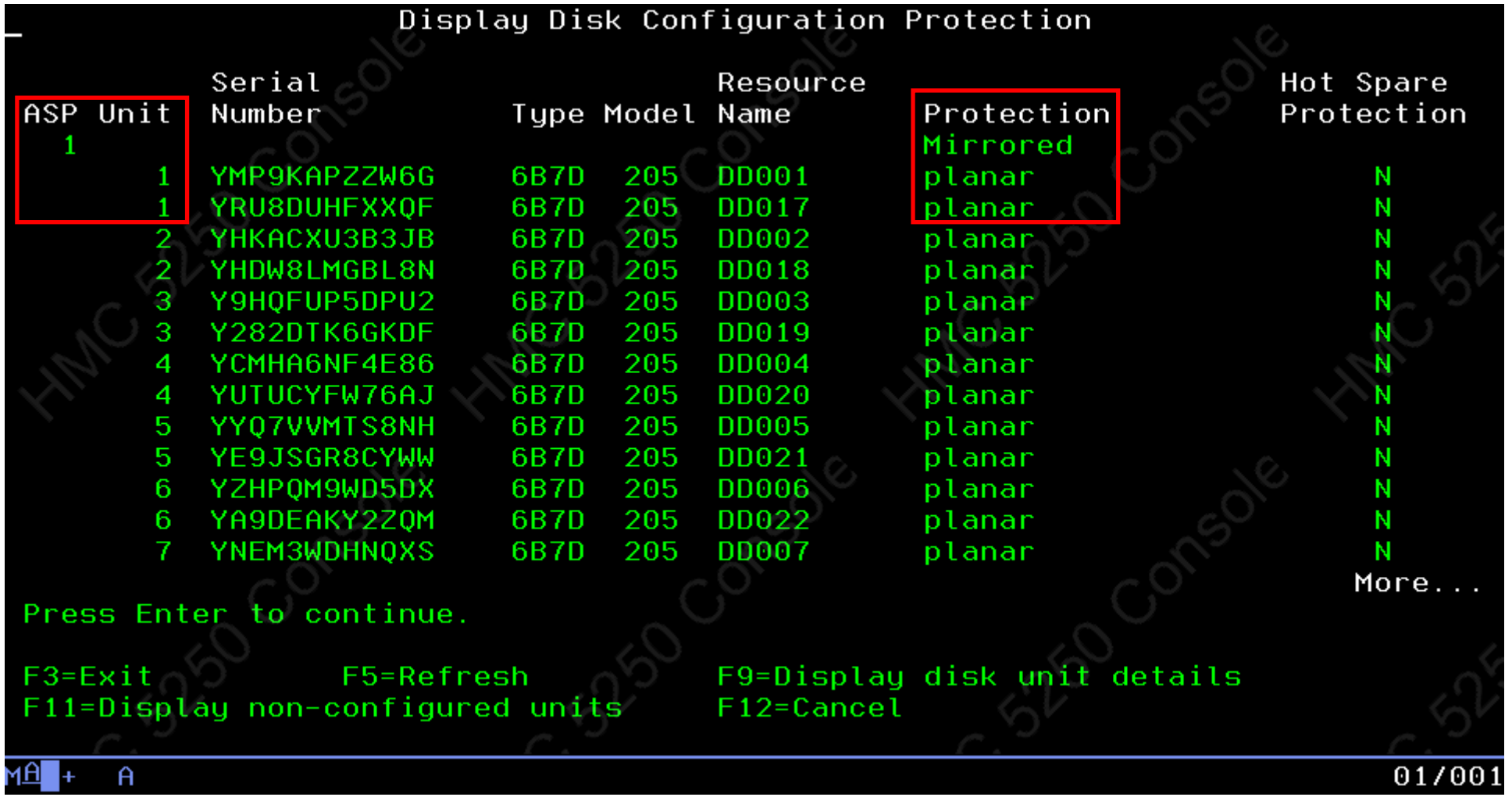
```
                                Work with Disk Status

Elapsed time:   00:00:00

                          --Protection--
Unit   ASP   Type   Status     Compression
   1     1   MRR    ACTIVE
   1     1   MRR    ACTIVE
   2     1   MRR    ACTIVE
   2     1   MRR    ACTIVE
   3     1   MRR    ACTIVE
   3     1   MRR    ACTIVE
   4     1   MRR    ACTIVE
   4     1   MRR    ACTIVE
   5     1   MRR    ACTIVE
   5     1   MRR    ACTIVE
   6     1   MRR    ACTIVE
   6     1   MRR    ACTIVE
   7     1   MRR    ACTIVE

Command
===>
F3=Exit    F5=Refresh    F12=Cancel    F24=More keys
```

PCIe Card NVMe IBM i DST/SST Screenshots



NVMe, IBM i, and SQL (new 2Q2020)

- The QSYS2.SYSDISKSTAT view contains information about disks
- The view is enhanced to recognize NVMe devices

```
select disk_model, disk_type,
       case when unit_type = 1 then 'SSD' else
       'Spinning'  end as "Type of disk",
       case when unit_nvme = 1 then 'NVMe' else 'Not NVMe'
       end as "NVMe indicator", percent_used from sysdiskstat;
```

- Mixed SSD and NVMe example

DISK_MODEL	DISK_TYPE	Type of disk	NVMe indicator	PERCENT_USED
0099	198E	Spinning	Not NVMe	7.848
0050	198C	Spinning	Not NVMe	0.006
0205	6B7D	SSD	NVMe	0.006
0205	6B7D	SSD	NVMe	0.006
0205	6B7D	SSD	NVMe	0.006

NVMe, IBM i, and SQL (new 2Q2020)

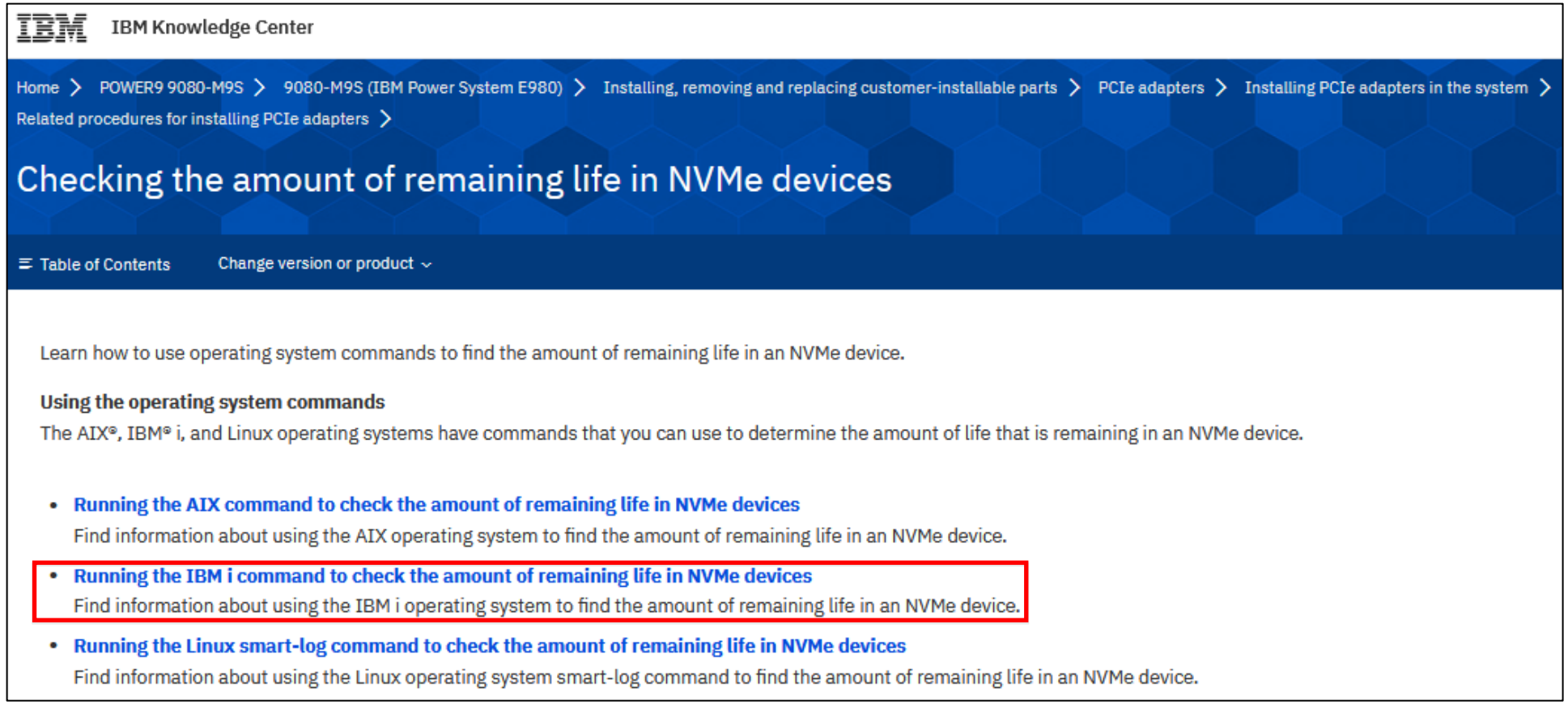
- All NVMe system example (using ACS Run SQL Scripts)

```
1  select disk_model, disk_type,
2  case when unit_type = 1 then 'SSD' else 'Spinning'
3  end as "Type of disk",
4  case when unit_nvme = 1 then 'NVMe' else 'Not NVMe'
5  end as "NVMe indicator", percent_used from qsys2.sysdiskstat;
```

DISK_MODEL	DISK_TYPE	Type of disk	NVMe indicator	PERCENT_USED
0205	6B7D	SSD	NVMe	4.902
0205	6B7D	SSD	NVMe	1.852
0205	6B7D	SSD	NVMe	1.855
0205	6B7D	SSD	NVMe	1.853
0205	6B7D	SSD	NVMe	1.852
0205	6B7D	SSD	NVMe	1.852
0205	6B7D	SSD	NVMe	1.852
0205	6B7D	SSD	NVMe	1.852
0205	6B7D	SSD	NVMe	1.852
0205	6B7D	SSD	NVMe	1.852
0205	6B7D	SSD	NVMe	1.852
0205	6B7D	SSD	NVMe	1.854
0205	6B7D	SSD	NVMe	1.853
0205	6B7D	SSD	NVMe	1.852
0205	6B7D	SSD	NVMe	1.852
0205	6B7D	SSD	NVMe	1.852
0205	6B7D	SSD	NVMe	1.852

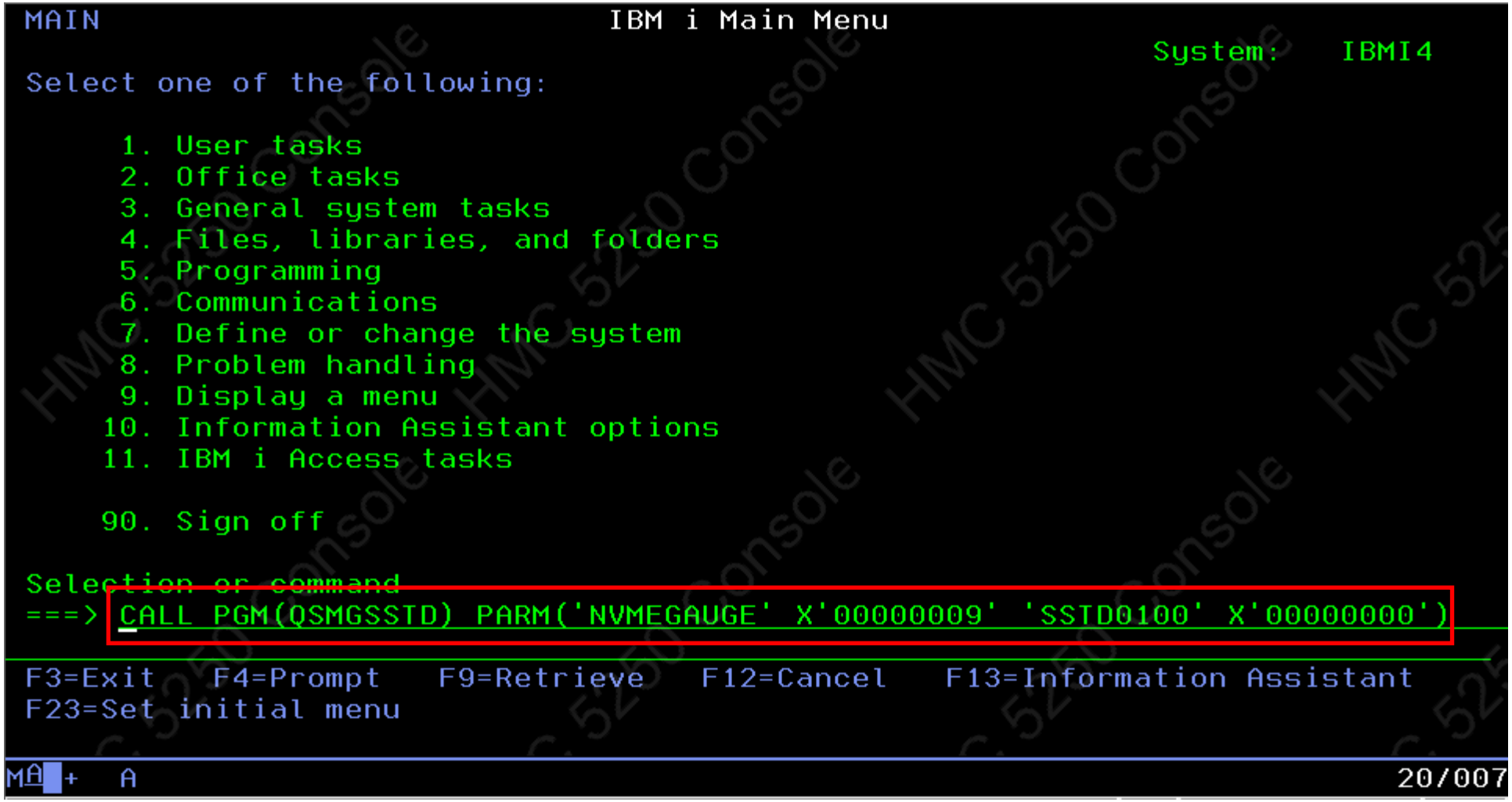
Done: 16 rows retrieved.

NVMe Fuel Gauge

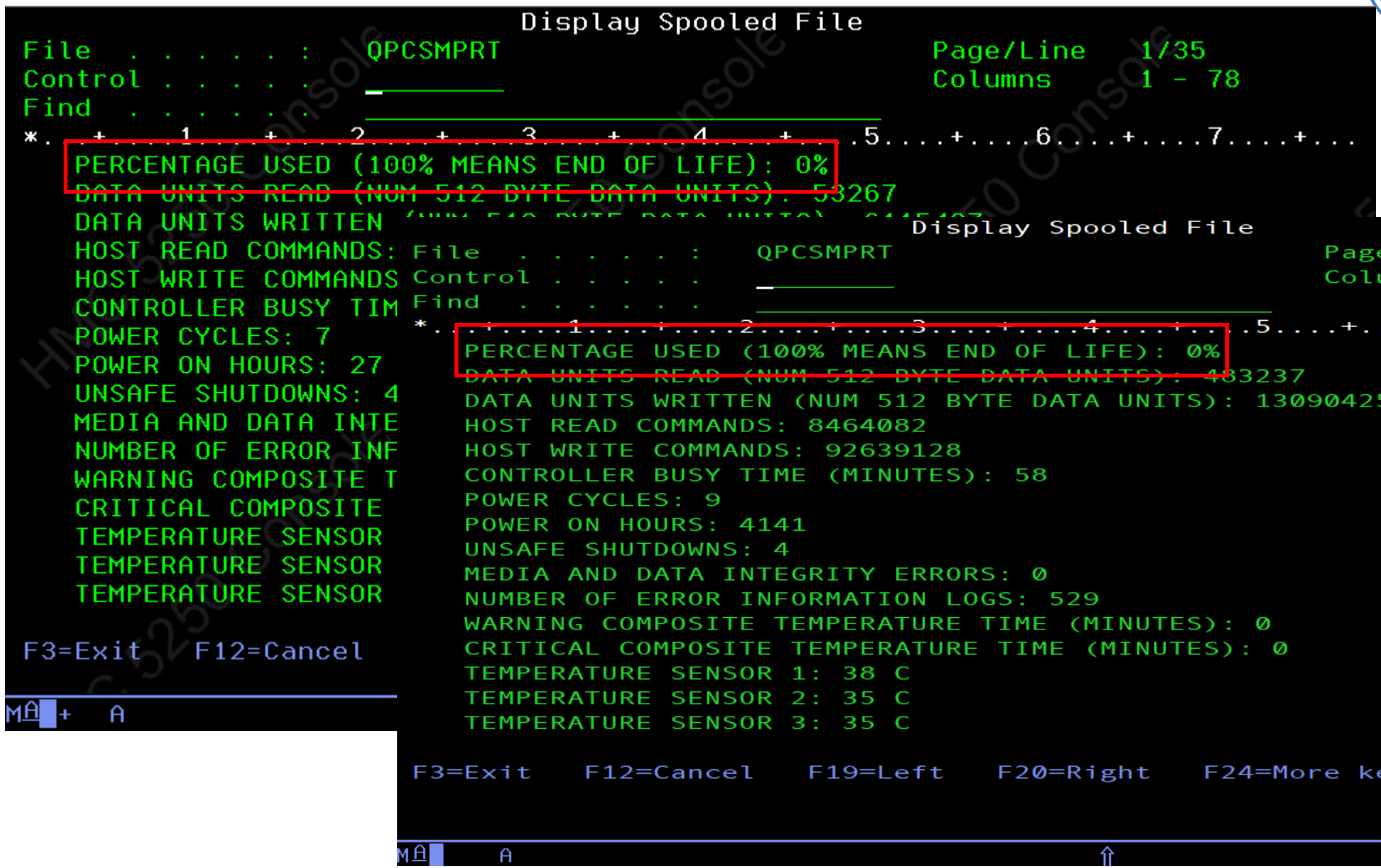


https://www.ibm.com/support/knowledgecenter/en/9080-M9S/p9hak/pxhak_nvme_remaining_life_kickoff.htm

IBM i NVMe Fuel Gauge



IBM i NVMe Fuel Gauge – Page 3



Performance Testing Workload Generator

Generator Used : BatchWL

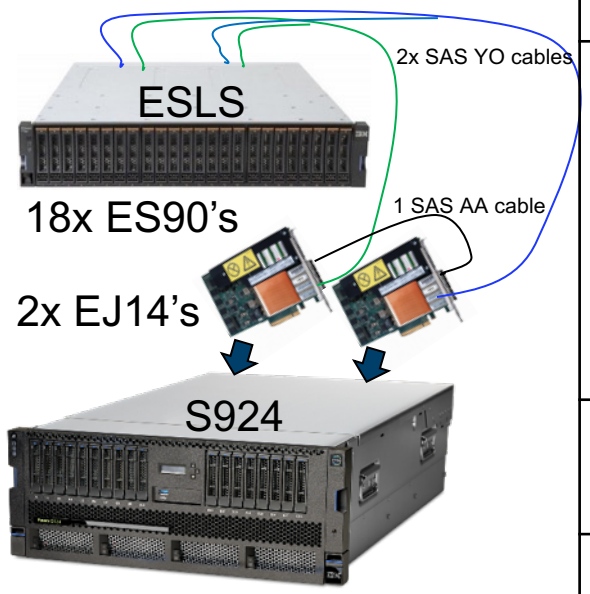
- Db2 or native I/O database queries
- The databases used:
  - Keyed files
  - Small databases used 256 byte records
  - Large databases used 32K byte records
  - Files are unique to each run thread to eliminate locking
  - Files are each over 3GB in size to reduce caching effects
- Each run thread does an I/O operation on its unique file
- Sequential I/O operations access each record by incremental key
- Random I/O operations access each record by a random key
- Performance data is gathered per thread and as a total to report the workload results

Workloads Used (6)

Extreme “4 corners” to stress software and hardware,
via small random ops and large sequential ops with 100% reads or 100% writes
+ small length random updates (50/50 read/write) and large sequential updates

Configurations Tested

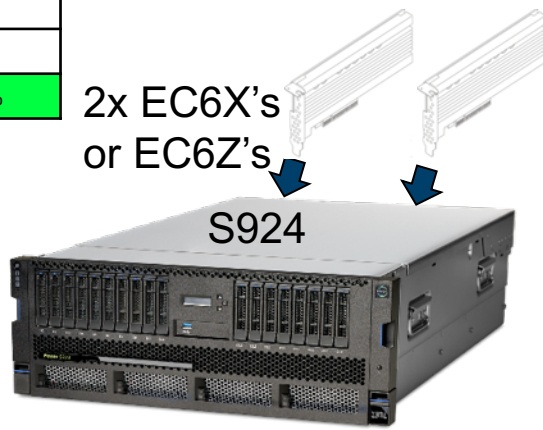
RAID-5 vs. Mirrored | Mirrored vs. Mirrored



SAS versus NVMe Comparison								
	Parameters	**Units**	**SAS**	**NVMe**	**Difference**	**SAS**	**NVMe**	**Difference**
Protection	RAID Type	0,1,5,6,10	RAID 5	OS Mirror		RAID 10	OS Mirror	
Endurance	DWPD		10	5		10	5	
Bill Of Materials	Adapters	#	2	2		2	2	
		FC / CCIN	EJ14/57B1	EC6Z/58FE		EJ14/57B1	EC6X/58FE	
	Signal Cables	#	4	0		4	0	
		FC	Various	N/A		Various	N/A	
	Enclosure	#	1	0		1	0	
		FC	ESLS	N/A		ESLS	N/A	
	Drives	#	18	2		18	2	
		FC	ES91	EC6Z		ES91	EC6X	
Capacity	Per Drive	GB	387	6400		387	3200	
	# of Arrays	#	2	1		2	1	
	Total	TB	6.192	6.4	3.4%	3.483	3.2	-8.1%
Real Estate	PCIe Slots	#	2	2		2	2	
	Rack Space	Rack Us	2	0		2	0	
	Total	Us	2.2	0.2	-90.9%	2	0.2	-90.9%

Color Key	Meaning
Green	Advantage NVMe
Yellow	Within +/-10%
Red	Advantage SAS

POWER9



List pricing as of 12/6/2019 assumes MTM = 9009-42A

All testing was done with native IBM i device driver.

Capacity and Performance Comparison

RAID-5 vs. Mirrored | Mirrored vs. Mirrored

SAS vs NVMe Performance Comparisons

	Parameter	Units	SAS	NVMe	Difference	SAS	NVMe	Difference
Protection	**RAID Type**	**0,1,5,6,10**	**RAID 5**	**OS Mirror**		**RAID 10**	**OS Mirror**	
Application Speed	Random 255B Reads	Trans/s	273,000	390,000	42.0%	276,185	387,284	40.2%
	Random 255B Writes	Trans/s	2,898,000	5,824,000	101.0%	6,674,419	5,785,859	-13.3%
	Sequential 32KB Reads	GB/s	3.51	8.22	134.2%	3.33	8.22	146.8%
	Sequential 32KB Writes	GB/s	1.01	2.46	143.6%	1.79	2.48	38.5%
	Random 4KB Reads	io/s	218,000	368,000	68.8%	217,629	385,410	77.1%
	Random 4KB Writes	io/s	167,000	690,000	313.2%	385,088	666,441	73.1%
Storage Subsystem	Sequential 32KB Reads	GB/s	3.48	8.05	131.3%	3.22	7.588	135.4%
	Sequential 32KB Writes	GB/s	1.00	4.56	356.0%	1.80	4.969	175.4%
	Random 4KB Reads	us/io	123	99	-19.5%	133	99	-25.6%
	Random 4KB Writes	us/io	59	24	-59.3%	34	28	-17.6%

- Conclusions: Lots of green for NVMe indicates consistent better performance
- Random 255B Writes see a slower NVMe performance because the Mirror code is offloaded to the CPUs instead of the SAS Controllers

Color Key	Meaning
Green	Advantage NVMe
Yellow	Within +/-10%
Red	Advantage SAS

How affordable are NVMe Devices again?

1.6TB Comparison

Feature Code	Description	US List Price
ESB9	387GB Enterprise SAS 4k SFF-3 SSD for IBM i (x4 devices)	$7,796
EJ1M	Expanded Function Storage Backplane 12 SFF-3 Bays/RDX Bay/Opt Ext SAS Port	$4,099
Total		$11,895

Feature Code	Description	US List Price
EC6V	PCIe3 x8 1.6 TB NVMe Flash Adapter for IBM i (x2 devices)	$5,198
	Total Savings	**$6,697**

3.2TB Comparison

Feature Code	Description	US List Price
ESB9	387GB Enterprise SAS 4k SFF-3 SSD for IBM i (x8 devices	$15592
EJ1M	Expanded Function Storage Backplane 12 SFF-3 Bays/RDX Bay/Opt Ext SAS Port	$4,099
Total		$19,691

Feature Code	Description	US List Price
EC6X	PCIe3 x8 3.2 TB NVMe Flash Adapter for IBM i (x2 devices)	$9,198
	Total Savings	**$10,493**

S922/S924 US List Price Comparisons
List prices are subject to change

What about compared to HDD?

1.6TB

Feature Code	Description	US List Price
ESB9	283GB 15K RPM SAS SFF-3 4k Block Cached Disk Drive (x6)	$3,750
EJ1M	Expanded Function Storage Backplane 12 SFF-3 Bays/RDX Bay/Opt Ext SAS Port	$4,099
Total		$7,849

Feature Code	Description	US List Price
EC6V	PCIe3 x8 1.6 TB NVMe Flash Adapter for IBM i (x2 devices)	$5,198
	Total Savings	**$2,651**

3.2TB

Feature Code	Description	US List Price
ESB9	283GB 15K RPM SAS SFF-3 4k Block Cached Disk Drive (x12)	$7,500
EJ1M	Expanded Function Storage Backplane 12 SFF-3 Bays/RDX Bay/Opt Ext SAS Port	$4,099
Total		$11,599

Feature Code	Description	US List Price
EC6X	PCIe3 x8 3.2 TB NVMe Flash Adapter for IBM i (x2 devices)	$9,198
	Total Savings	**$2,401**

S922/S924 US List Price Comparisons
List prices are subject to change

Conclusions...

- Multiple NVMe devices are available on IBM POWER9 servers
- NVMe devices are supported on AIX, Linux, IBM i and VIOS
- NVMe devices are easy to install and configure
- NVMe devices are fast


THE END

IBM i I/O Support Summary

IBM i I/O Support	Type of Configuration (Native, VIOS, iVirt, All)	IBM i 7.4	IBM i 7.3	IBM i 7.2
Enhancements from Nov 2019				
#EC6V, #EC6U - PCIe3 x8 NVMe 1.6 TB SSD NVMe Flash Adapter - IBM i	All	Tech Refresh 1	Base (VIOS, iVirt)	Base (VIOS, iVirt)
#EC6X, #EC6W- PCIe3 x8 NVMe 3.2 TB SSD NVMe Flash Adapter - IBM i	All	Tech Refresh 1	Base (VIOS, iVirt)	Base (VIOS, iVirt)
#EC6Z, #EC6Y - PCIe3 x8 NVMe 6.4 TB SSD NVMe Flash Adapter for IBM i	All	Tech Refresh 1	Base (VIOS, iVirt)	Base (VIOS, iVirt)
#ESB9 - 387GB Enterprise SAS 4k SFF-3 SSD for IBM i	Native, iVirt	Base	Tech Refresh 3	Tech Refresh 7
#ESBB - 387GB Enterprise SAS 4k SFF-2 SSD for IBM i	Native, iVirt	Base	Tech Refresh 3	Tech Refresh 7
#ESBF - 775GB Enterprise SAS 4k SFF-3 SSD for IBM i	Native, iVirt	Base	Tech Refresh 3	Tech Refresh 7
#ESBH - 775GB Enterprise SAS 4k SFF-2 SSD for IBM i	Native, iVirt	Base	Tech Refresh 3	Tech Refresh 7
#ESBK - 1.55TB Enterprise SAS 4k SFF-3 SSD for IBM i	Native, iVirt	Base	Tech Refresh 3	Tech Refresh 7
#ESBM - 1.55TB Enterprise SAS 4k SFF-2 SSD for IBM i	Native, iVirt	Base	Tech Refresh 3	Tech Refresh 7
#ESB8 - 387GB Enterprise SAS 4k SFF-3 SSD for AIX/Linux	VIOS	Base	Tech Refresh 3	Tech Refresh 7
#ESBA - 387GB Enterprise SAS 4k SFF-2 SSD for AIX/Linux	VIOS	Base	Tech Refresh 3	Tech Refresh 7
#ESBE - 775GB Enterprise SAS 4k SFF-3 SSD for AIX/Linux	VIOS	Base	Tech Refresh 3	Tech Refresh 7

https://www.ibm.com/support/pages/node/1135378

IBM i I/O Support Details

November 2019 - IBM i 7.4 Technology Refresh 1 and IBM i 7.3 Technology Refresh 7 and IBM i 7.2

For more details on the new Power I/O features and enhancements listed below, see the October 8, 2019, announcement letter IBM Power Systems enhancements.

Dedicated and VIOS support for selected PCIe3 x8 SSD NVMe adapters - IBM i 7.4 TR 1, IBM i 7.3, IBM i 7.2

IBM i 7.4 TR1 natively supports PCIe3 x8 SSD NVMe adapters that provide multiple capacity points for enterprise workloads on selected Power servers with POWER9™ technology. Support is for selected dedicated and VIOS VSCSI attached NVMe devices. These low latency devices can be used as IBM i load sources and are able to provide a high number of IOPS and enhanced virtualization capabilities. Mirroring is required. Pairs of these storage devices can be added to LPAR configurations as dedicated PCIe devices.

Each device is a partitionable endpoint that can be dedicated to an LPAR, meaning that multiple partitions may be configured for a single system unit without the need to use virtualization.

IBM i 7.4, 7.3, and 7.2 configurations with VIOS are also supported as virtual SCSI drives that are backed by NVMe devices. Other previously announced devices can be used for the VIOS boot disk but should not be virtualized to IBM i. For VIOS configurations, support is for VIOS VSCSI LVM client only, so block size is 4096.

	IBM i	IBM i with VIOS
1.6 TB	#EC6V, #EC6U	#EC5B, #EC5G
3.2 TB	#EC6X #EC6W	#EC5D, #EC5C
6.4 TB	#EC6Z, #EC6Y	#EC5F, #EC5E

Additional code levels required:

FW940, or later

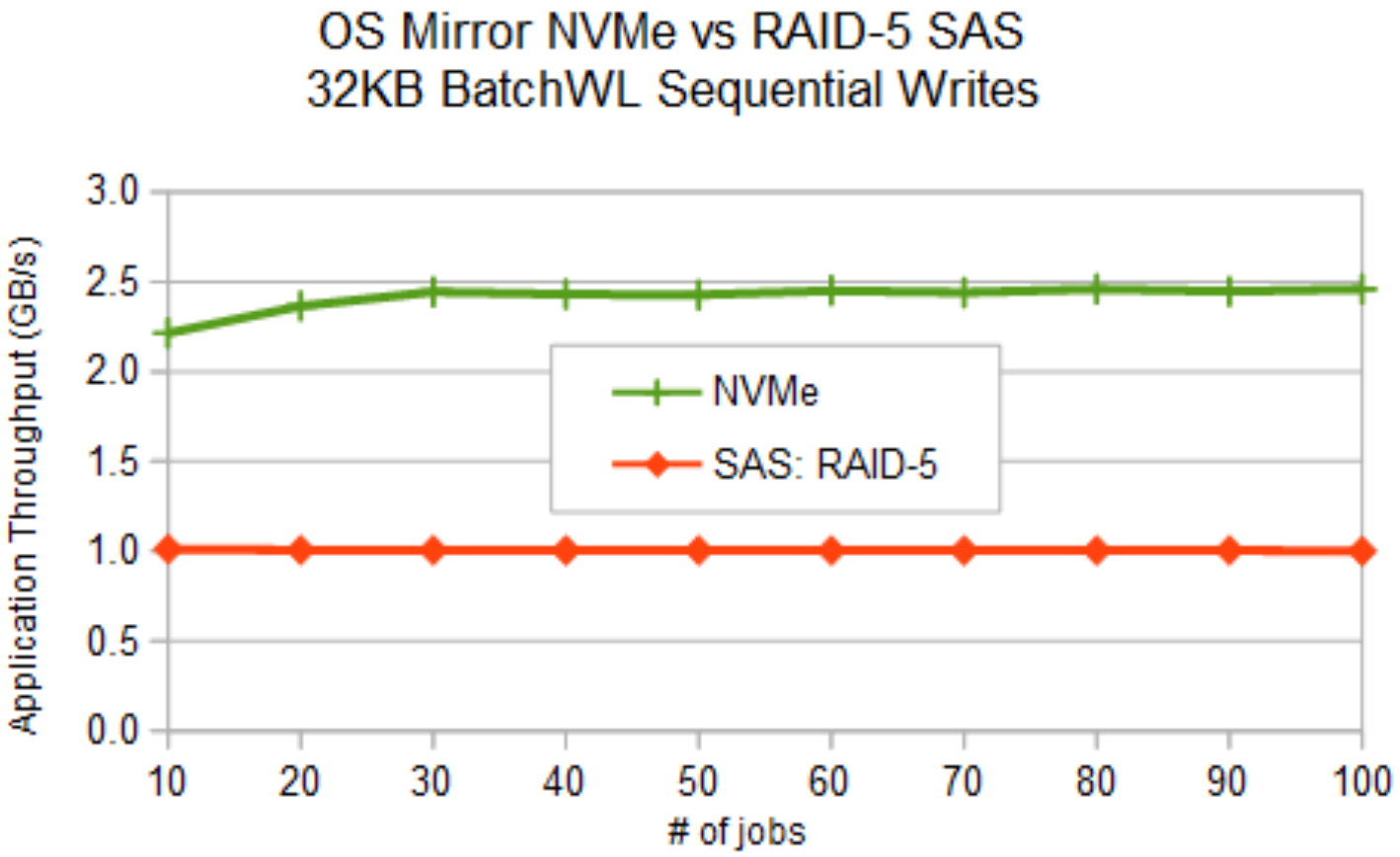
VIOS 3.1.0 (for VIOS configurations)

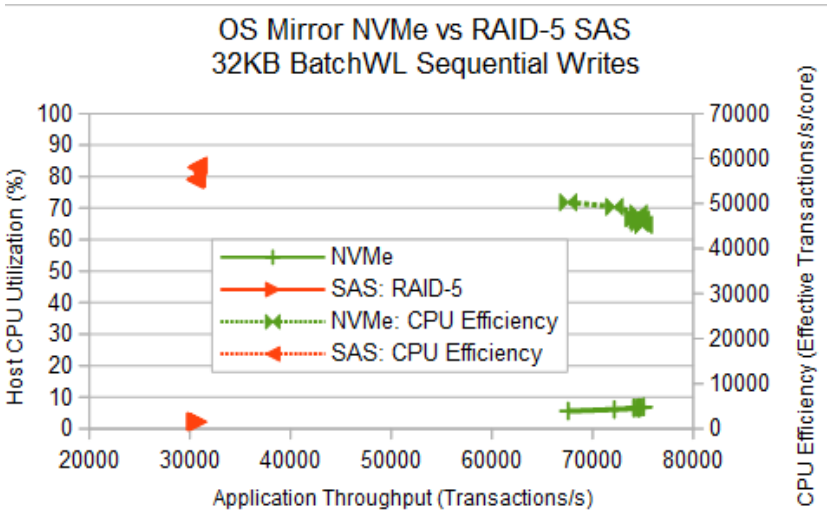
New Enterprise SSDs – 387 GB, 775 GB, 1.55 TB - IBM i 7.4, IBM i 7.3 TR 3, and IBM i 7.2 TR 7

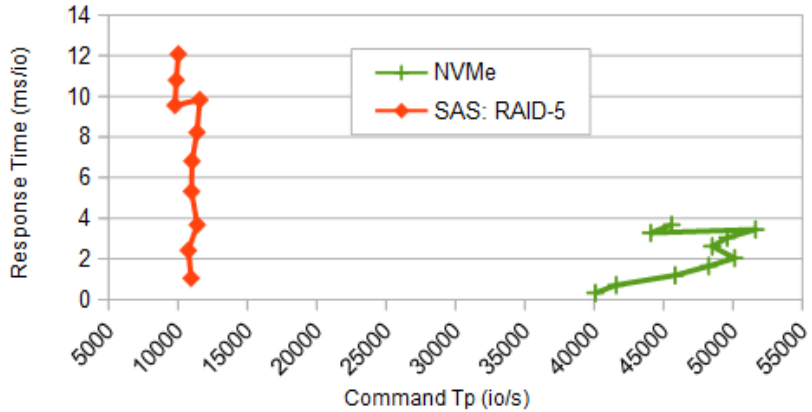
A new generation of enterprise 2.5-inch solid-state drives (SSD) improves enterprise-class reliability, endurance, and capacity characteristics.

https://www.ibm.com/support/pages/node/1137508

Large Sequential Writes: Mirror vs RAID-5



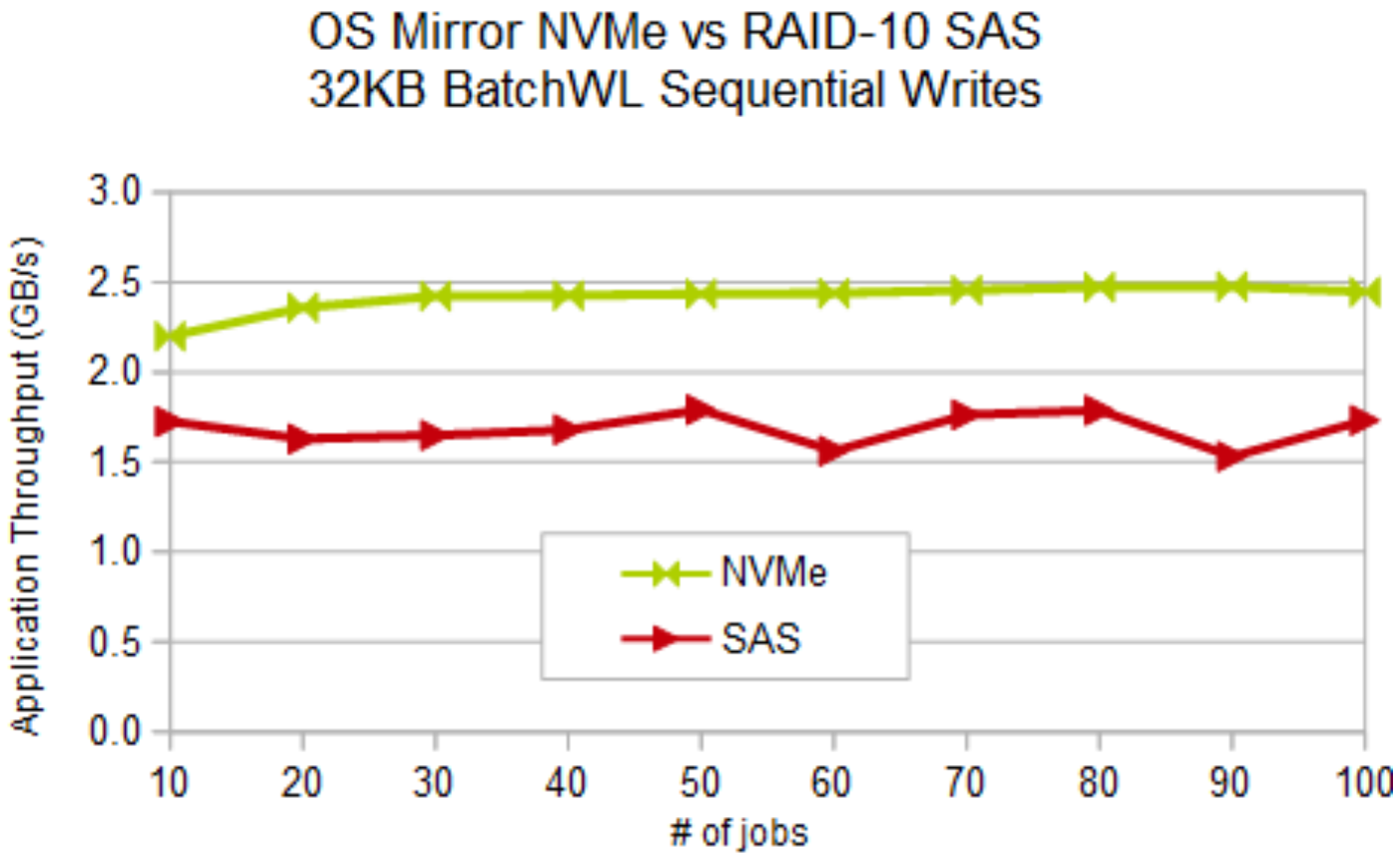


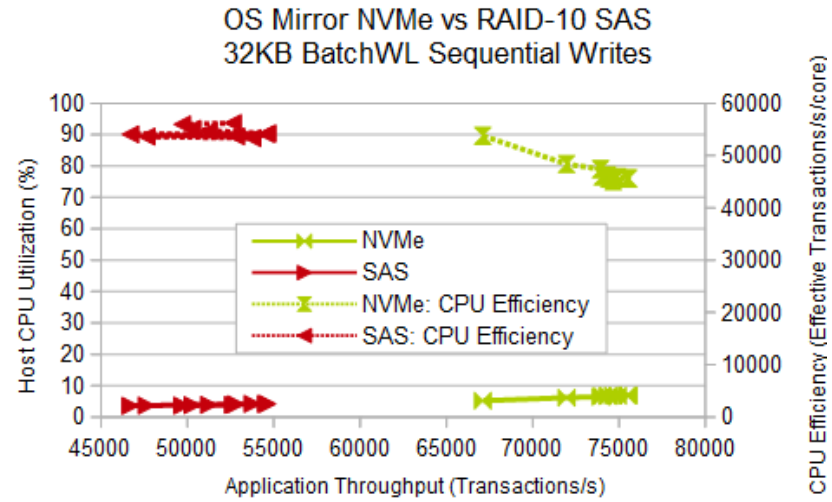


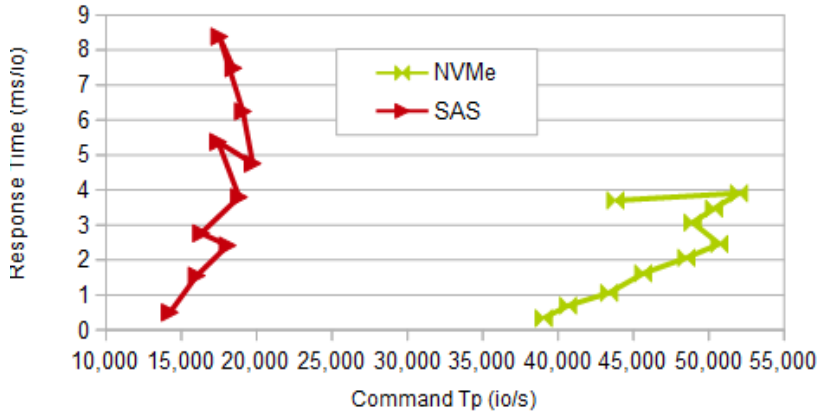
Take-aways:

1. Writing to a mirror is typically 2x faster than to RAID-5/6 (not new news).
2. There is a CPU utilization efficiency (amount of CPU per IO) cost for doing RAID (mirroring) function in the host CPUs.
3. NVMe write response times compare nicely with writes to the SAS IOA's DRAM cache.

Large Sequential Writes : Mirror vs Mirror



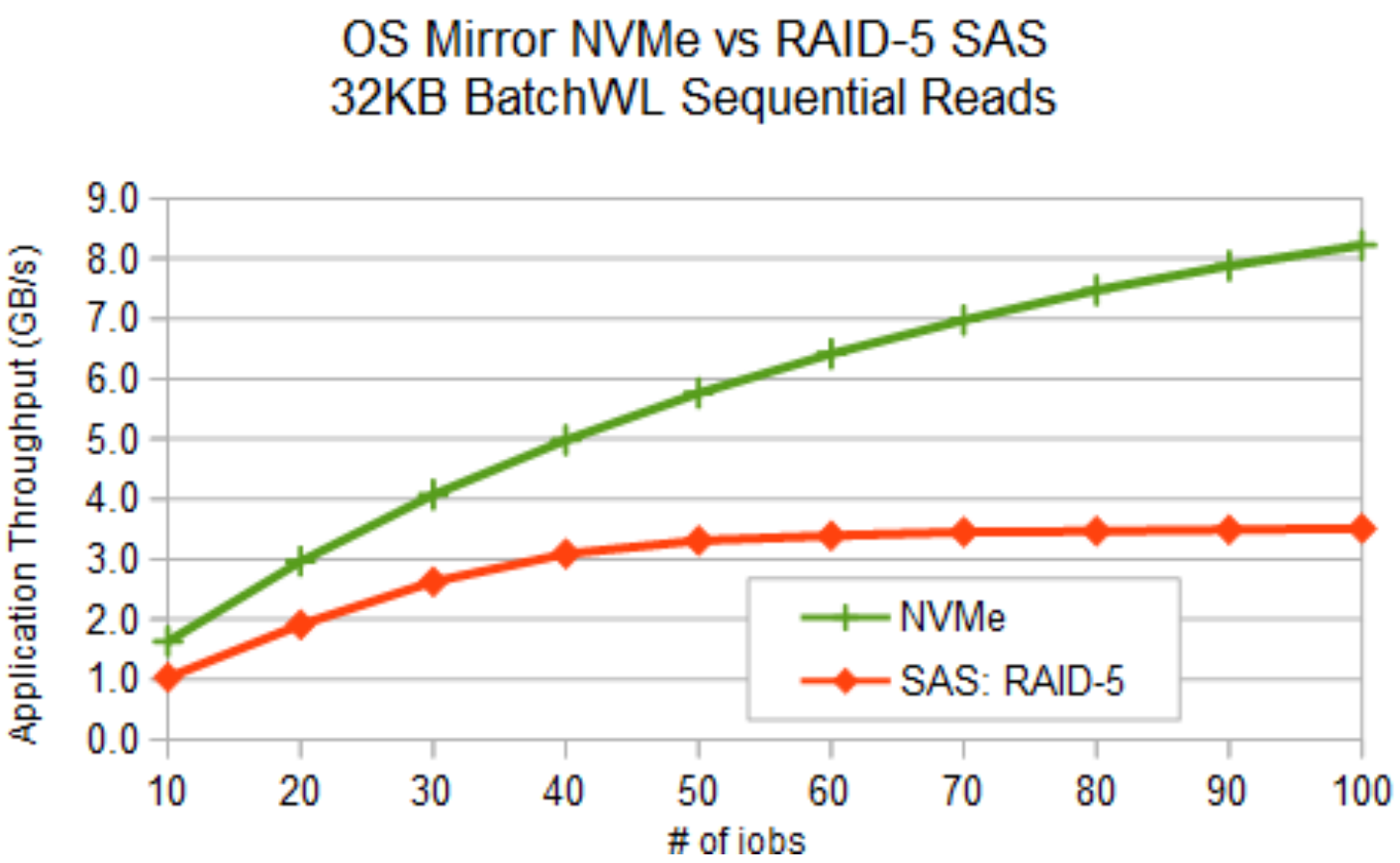


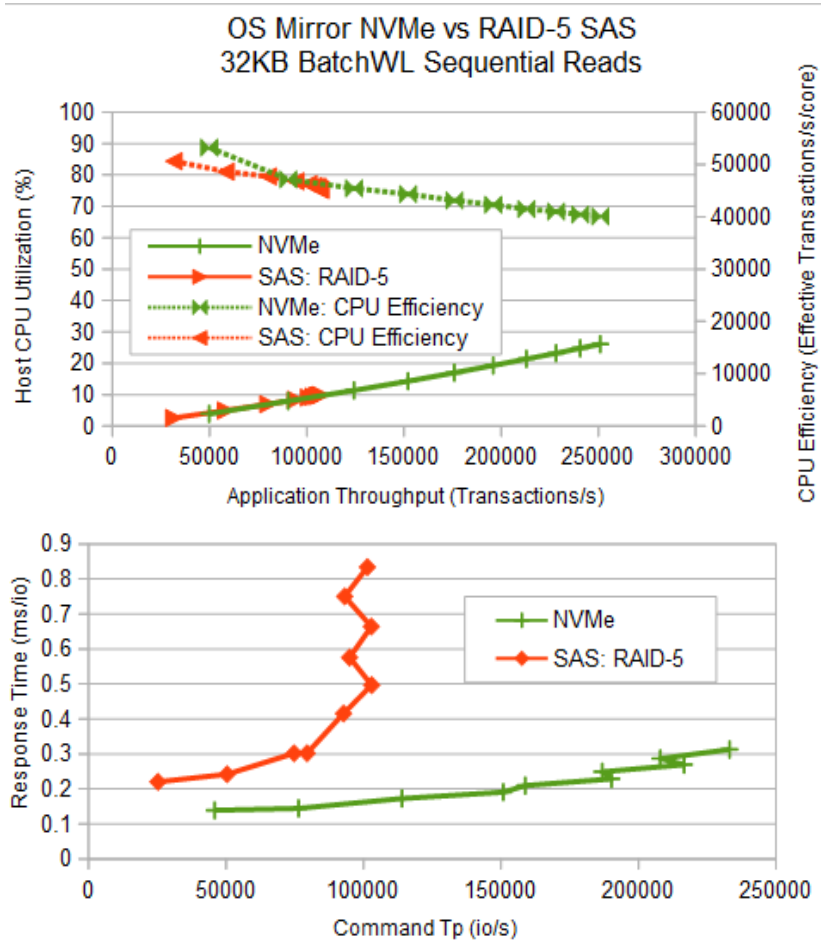


Take-aways:

1. One can make the SAS rig faster by throwing more $ at it by adding another ESLS drawer (more SAS fabric bandwidth).
2. There is a CPU utilization efficiency cost for doing RAID (mirror) function in the host CPUs.
3. NVMe write response times compare nicely with writes to the SAS IOA's DRAM cache.

Large Sequential Reads : Mirror vs RAID-5

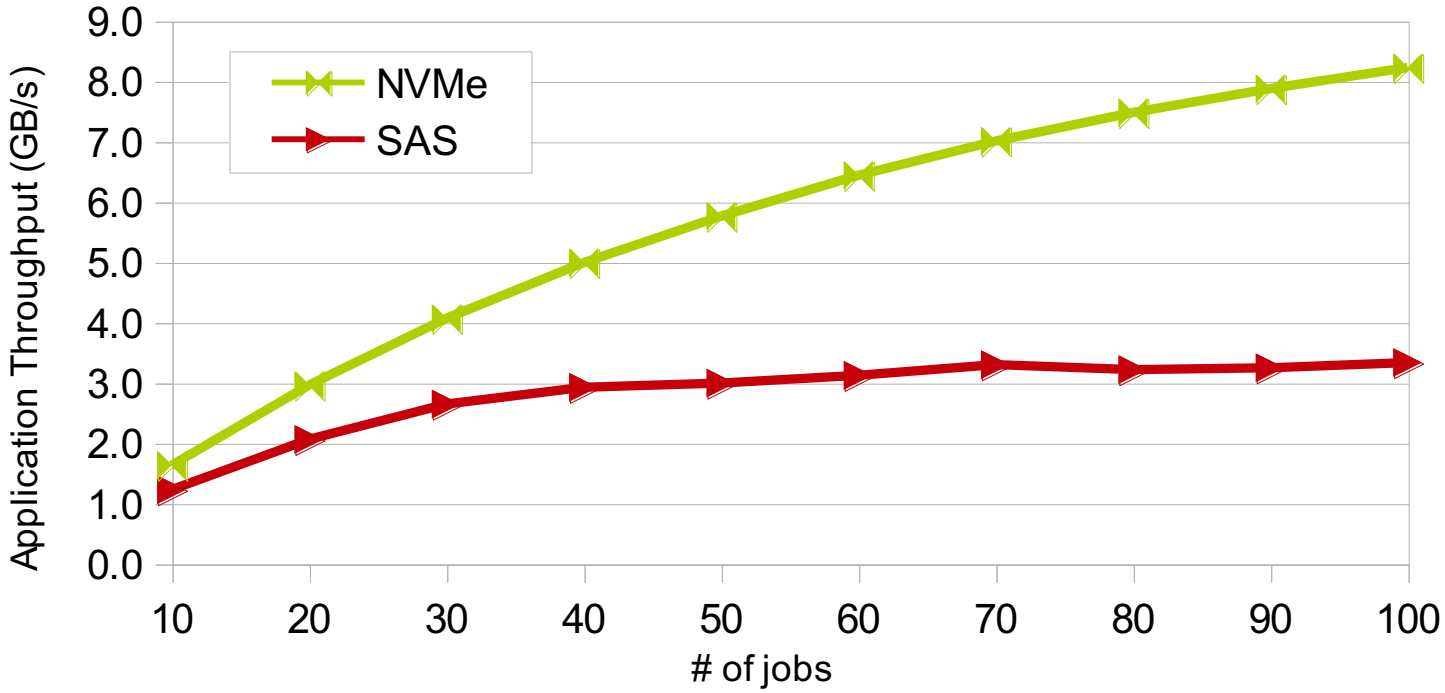


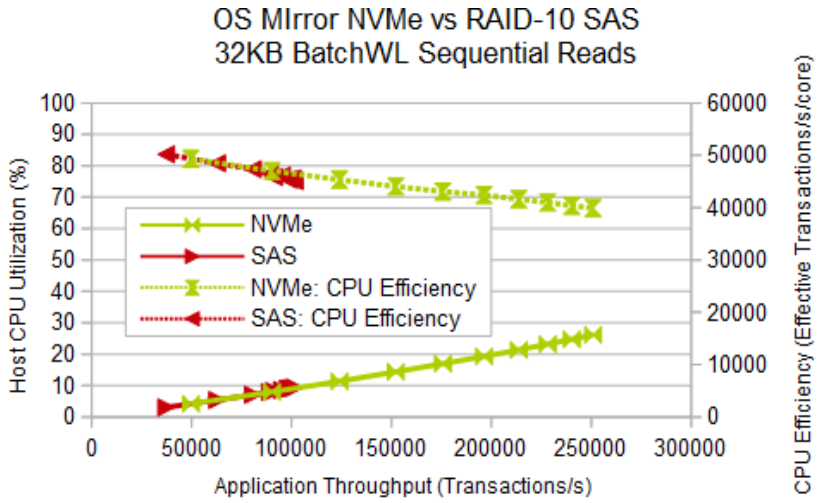


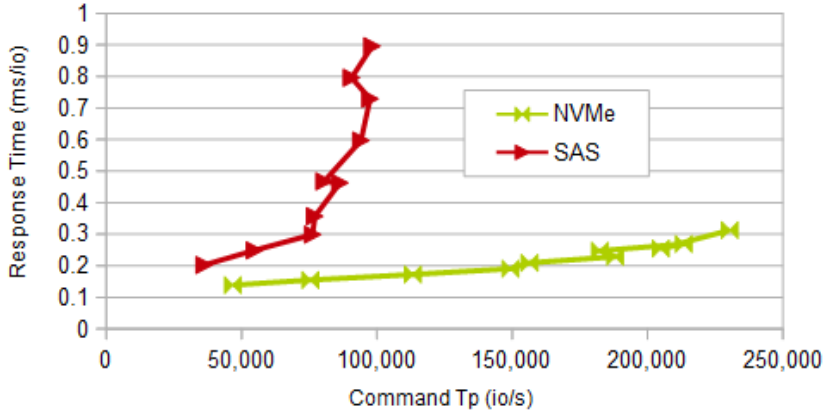
Take-aways:

1. The NVMe drive's PCIe bus is a clear winner over the SAS fabric.
   - Like the big writes, one could double the SAS throughput by adding a 2nd ESLS.
2. NVMe read response times are clear winners. (Thank you PCIe bus !)
3. There is still a slight CPU utilization efficiency cost for doing RAID function in the host CPUs.

Large Sequential Reads : Mirror vs Mirror



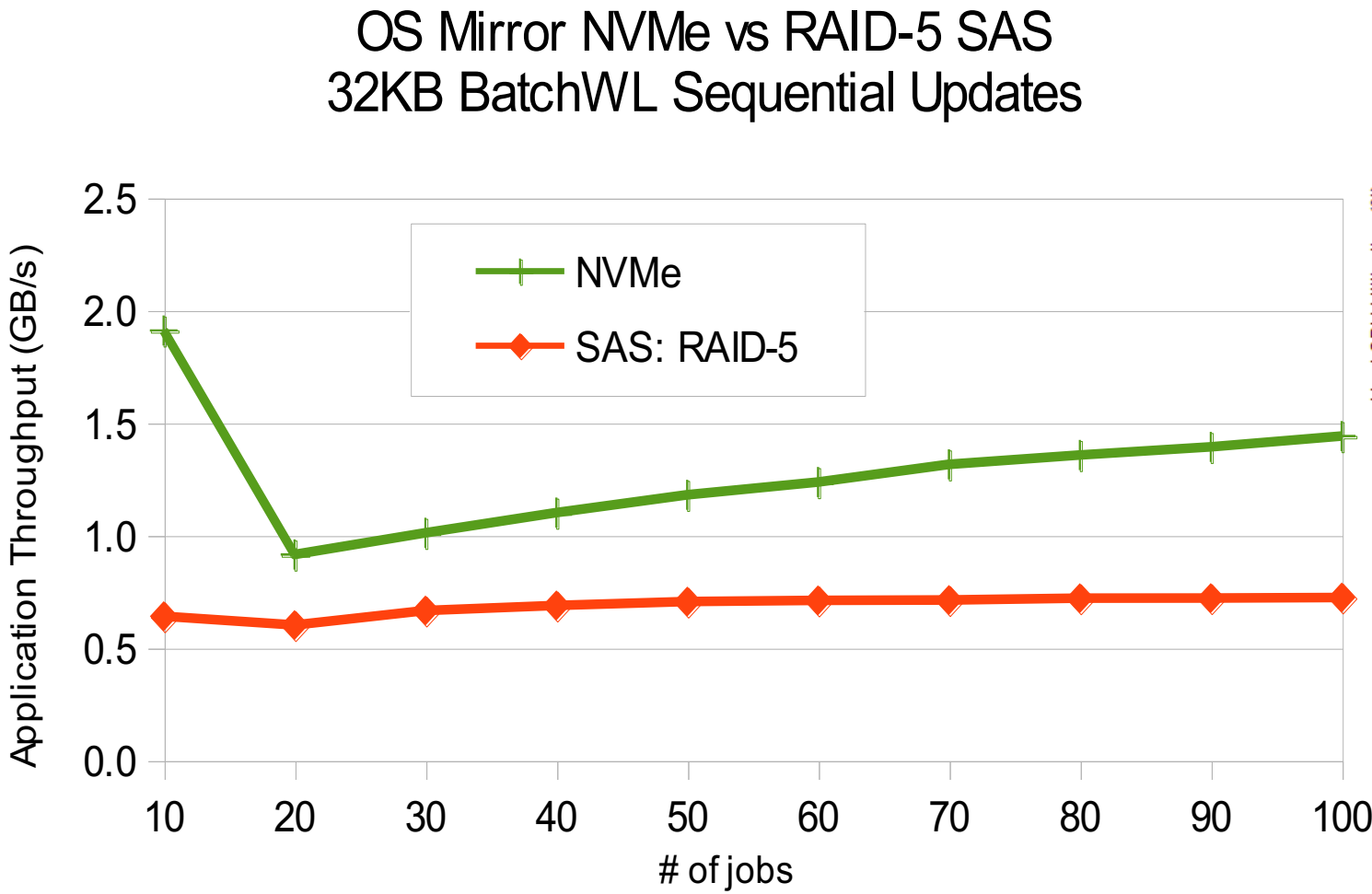


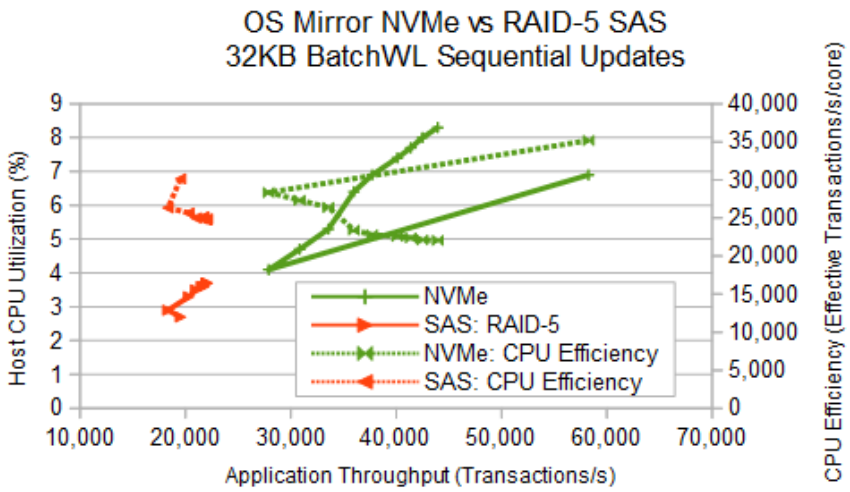


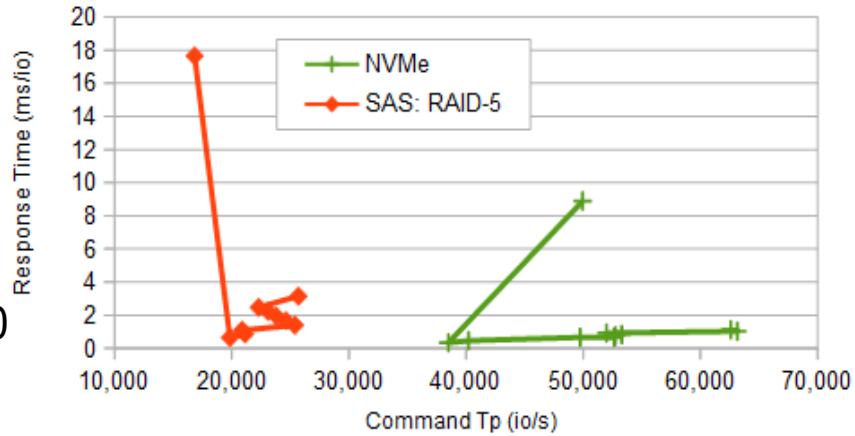
Take-aways:

1. The NVMe drive's PCIe bus is a clear winner over the SAS fabric.
   - Like the big writes, one could double the SAS throughput by adding a 2nd ESLS.
2. NVMe read response times are clear winners. (Thank you PCIe bus !)
3. There is still a slight CPU utilization efficiency cost for doing RAID function in the host CPUs.

Large Sequential Updates : Mirror vs RAID-5







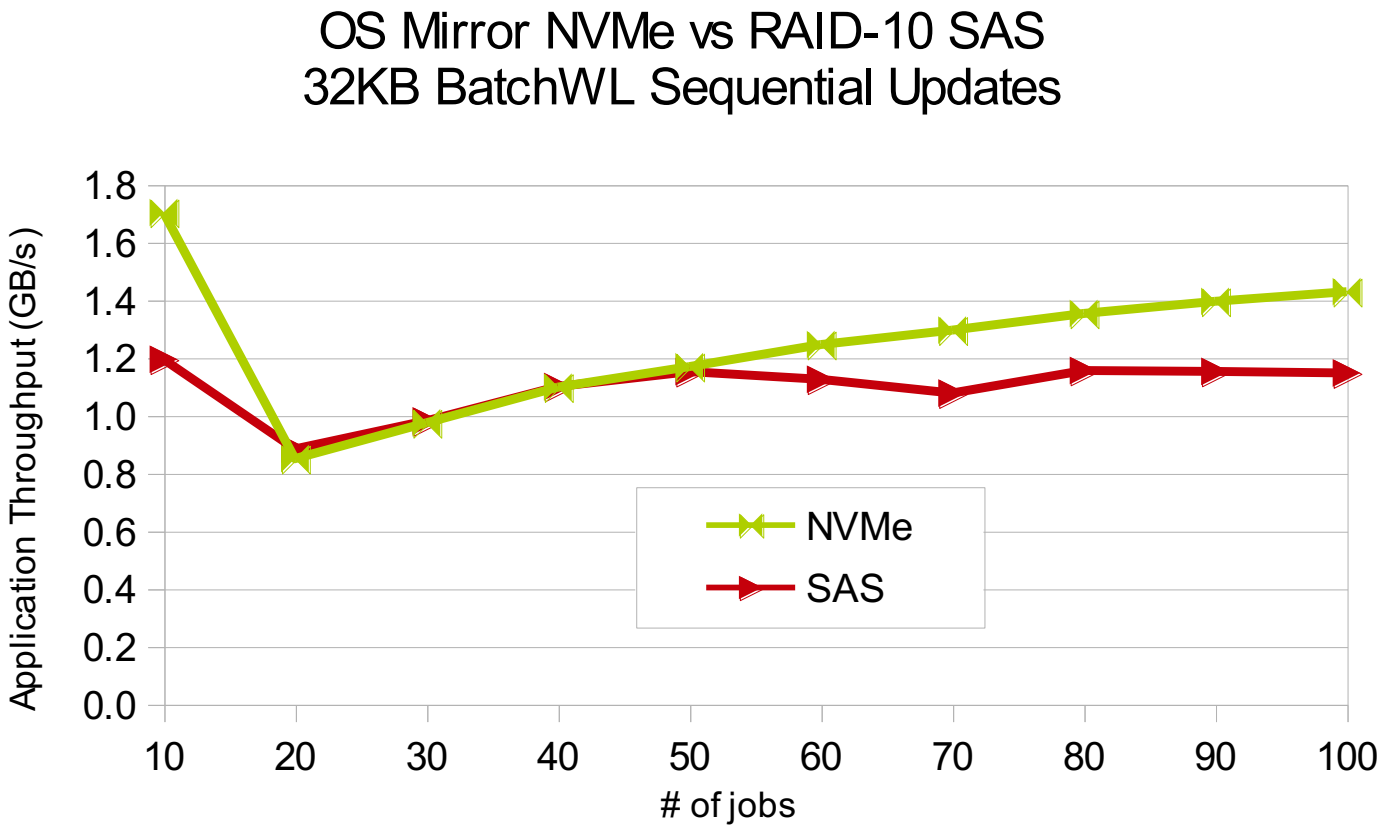
Take-aways:

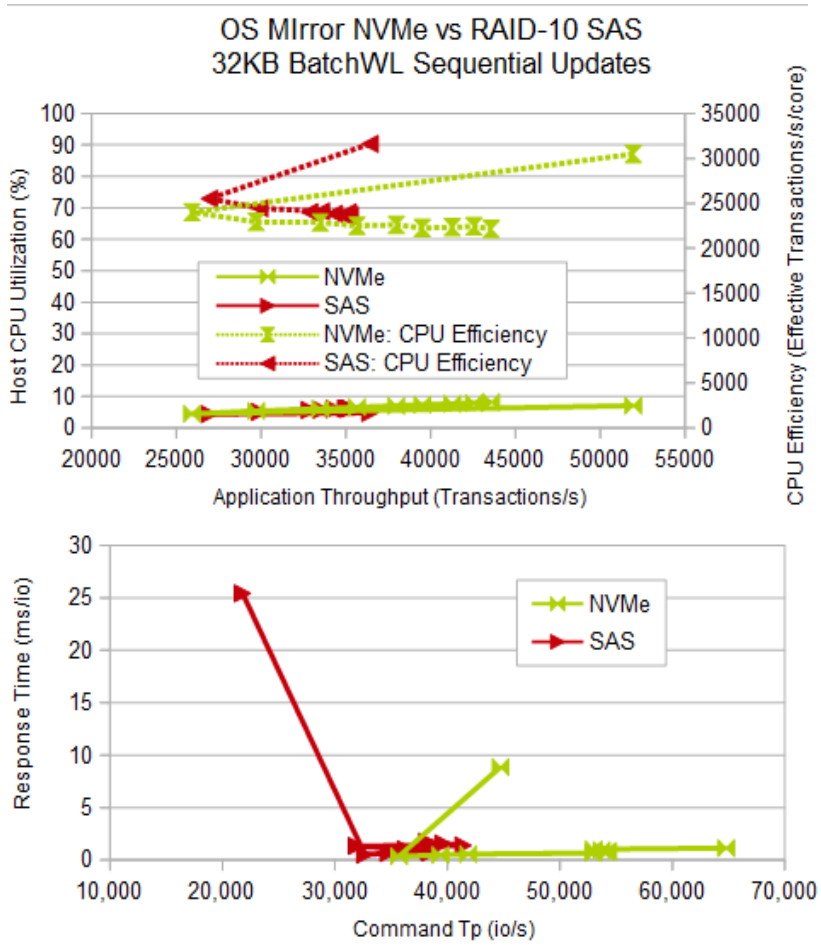
(1) The write portion of the update procedure benefits mirroring over RAID-5 (and -6).

(2) Root cause for the low load (10 jobs) NVMe advantage has yet to be isolated.

- We will study that closer when gen4 NVMe drives arrive on the scene.

Large Sequential Updates : Mirror vs Mirror



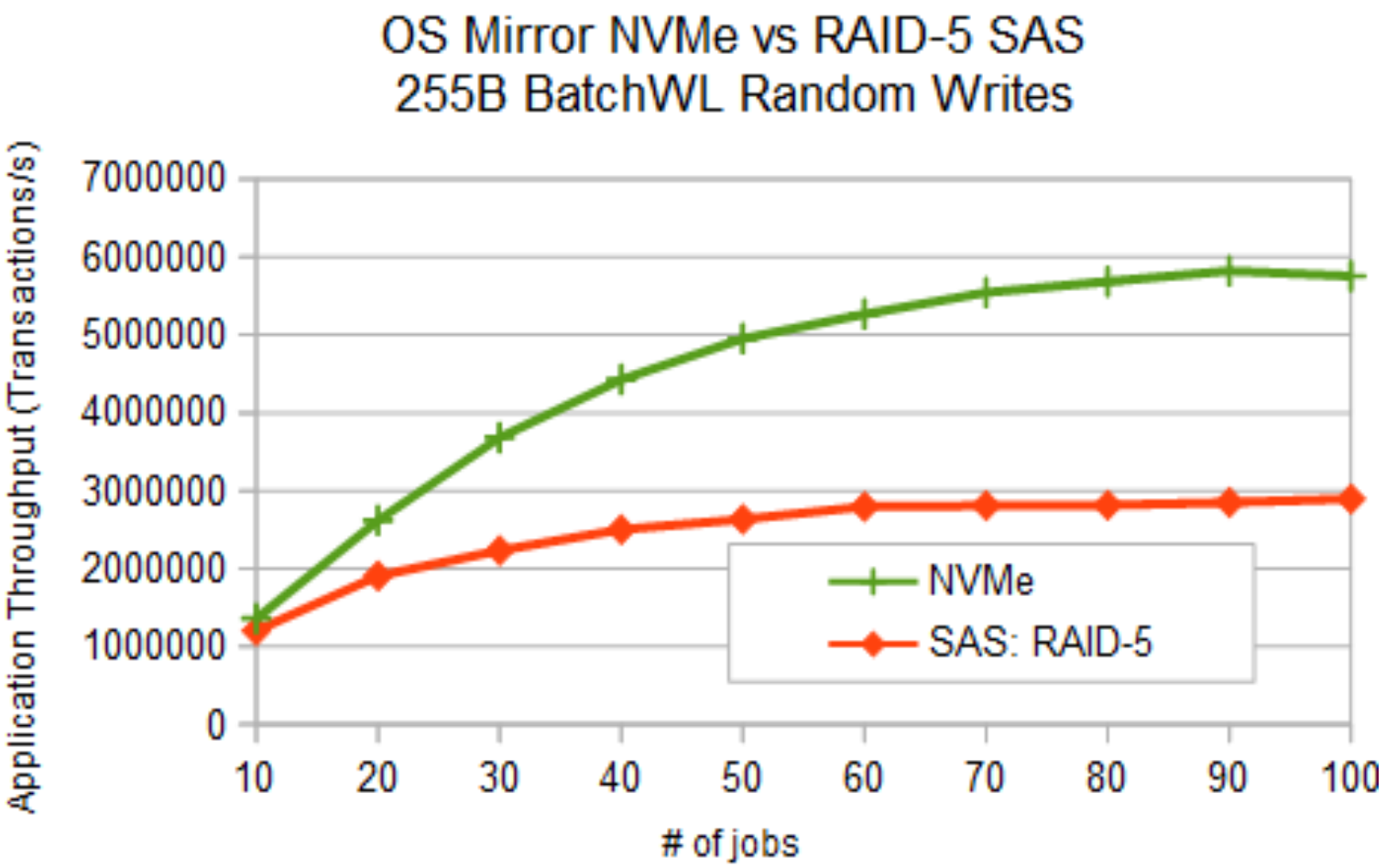


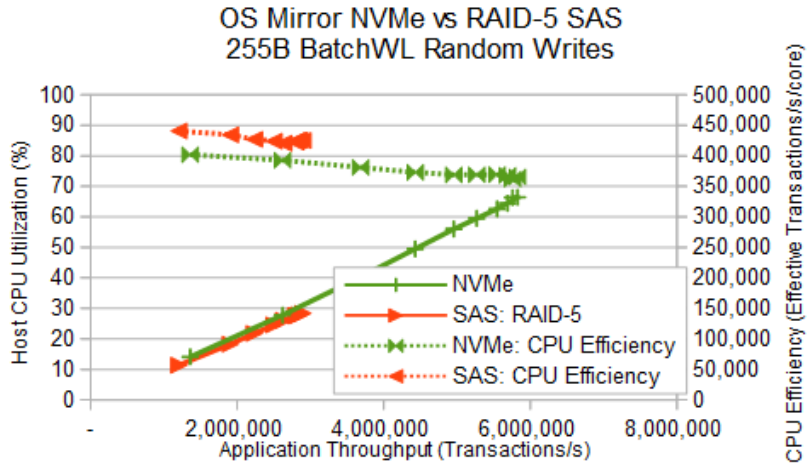
Take-aways:

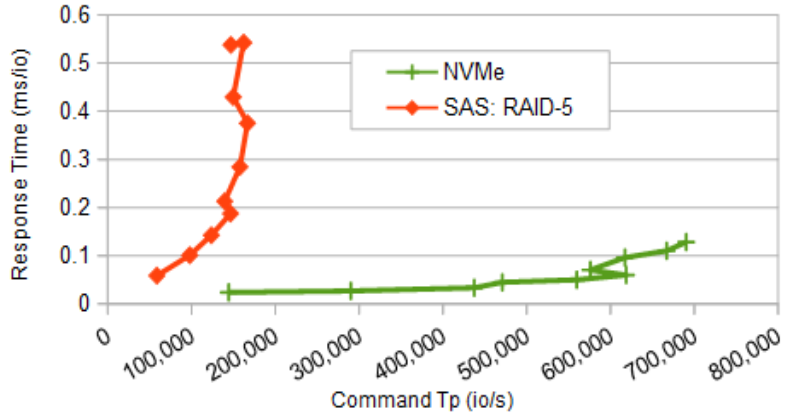
(1) A bit cloudier picture than for pure reads and pure writes, and versus RAID-5.

(2) Yet the 50/50 read/write workload shows a clear advantage for NVMe over SAS in terms of throughput and response time.

Small Op Random Writes : Mirror vs RAID-5



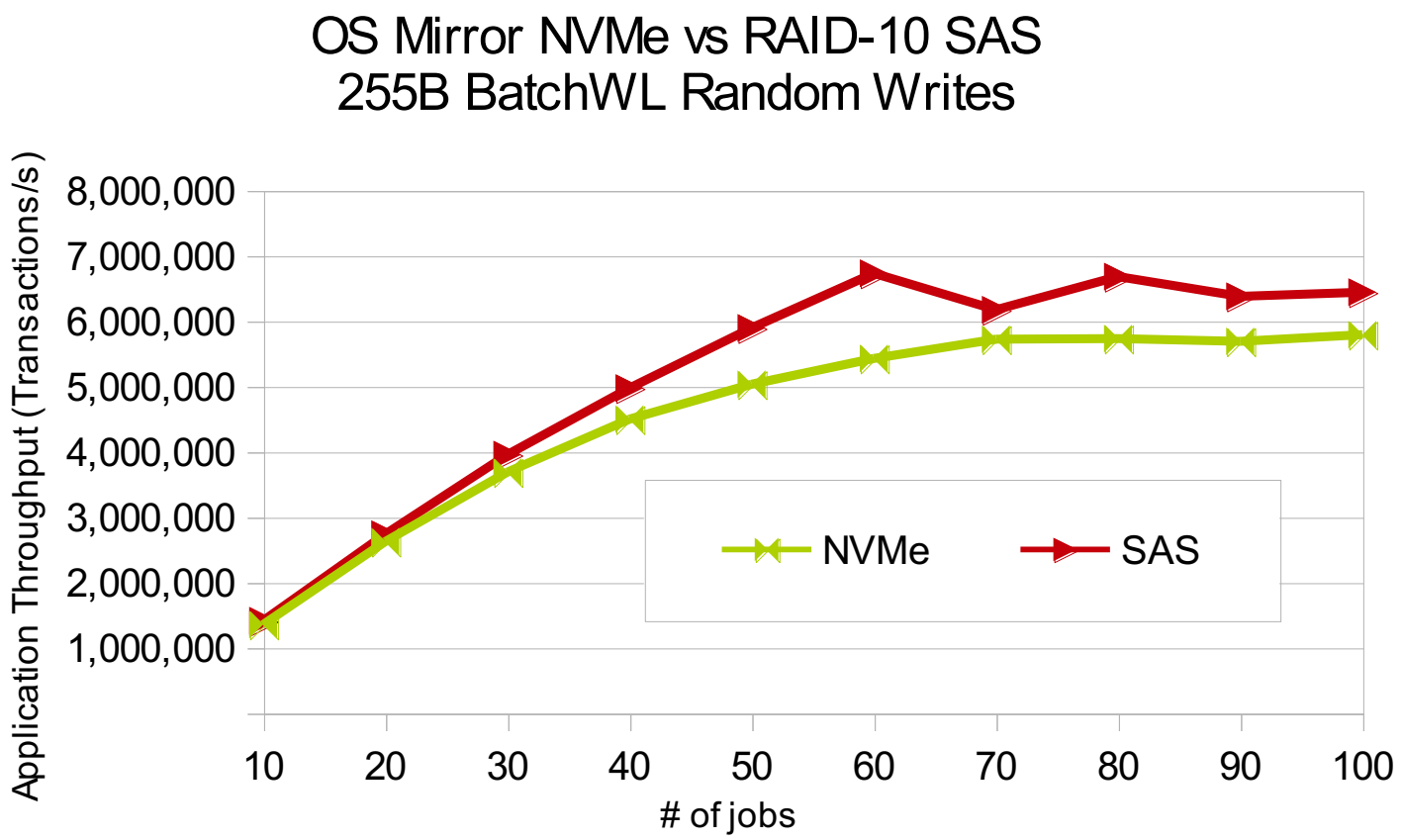


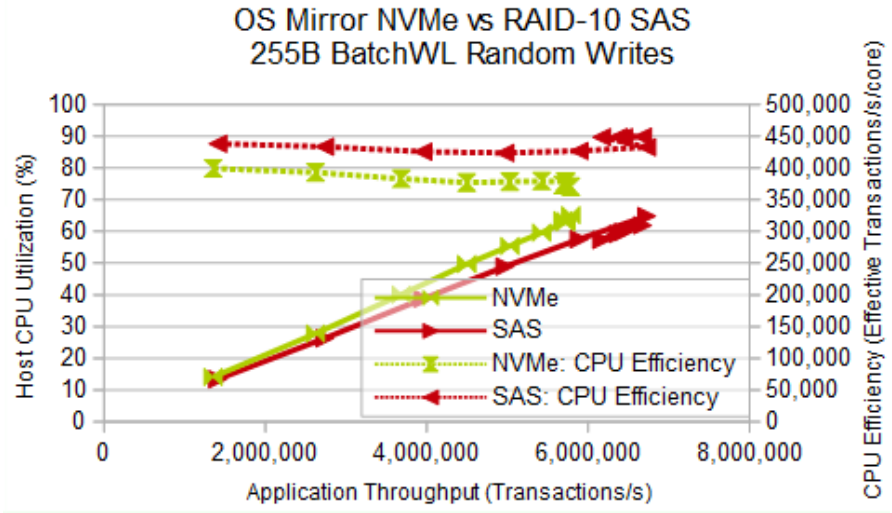


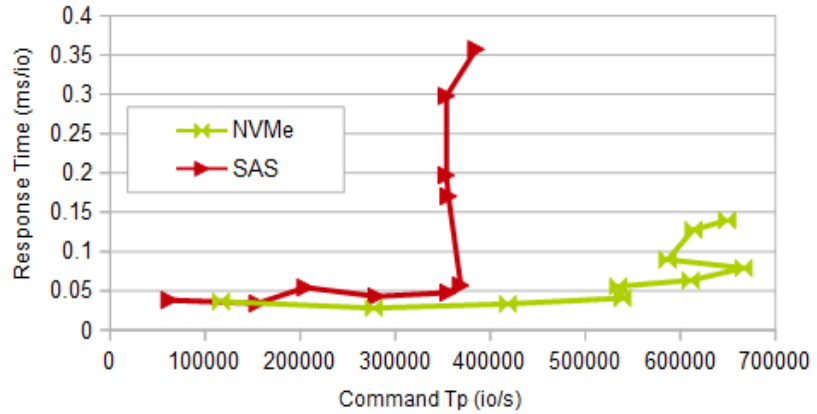
Take-aways:

1. Writing to a mirror is still faster than to RAID-5/6 ☺.
2. CPU utilization efficiency cost for doing RAID function in the host CPUs does allow SAS to use less CPU/io than NVMe.
3. Yet response times for NVMe writes still look A-OK. (i.e. lower than SAS.)

Small Op Random Writes : Mirror vs Mirror



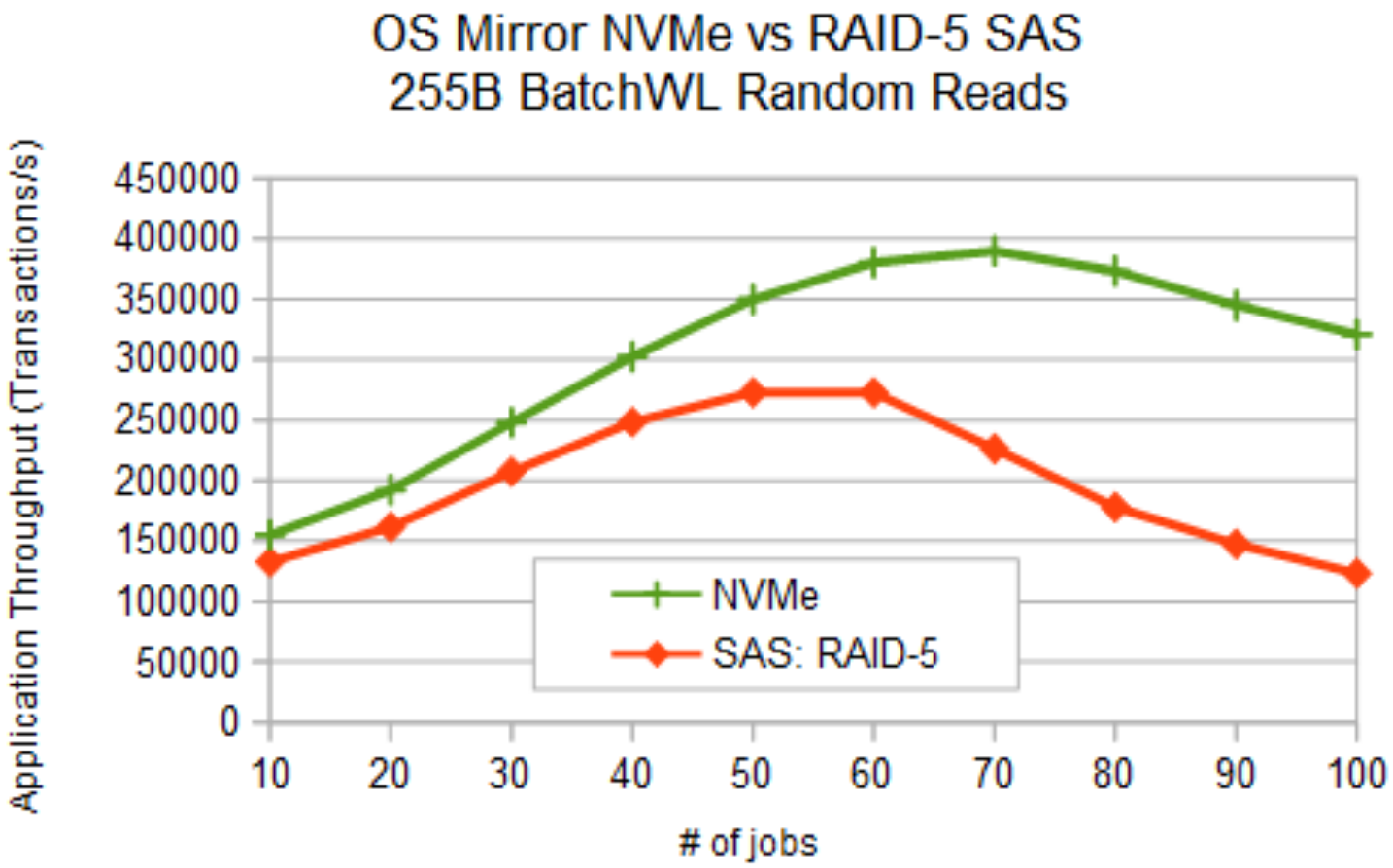


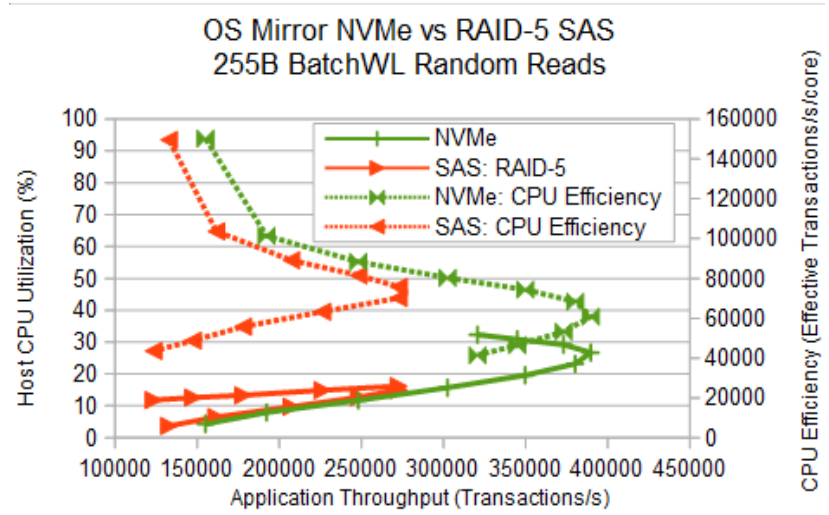


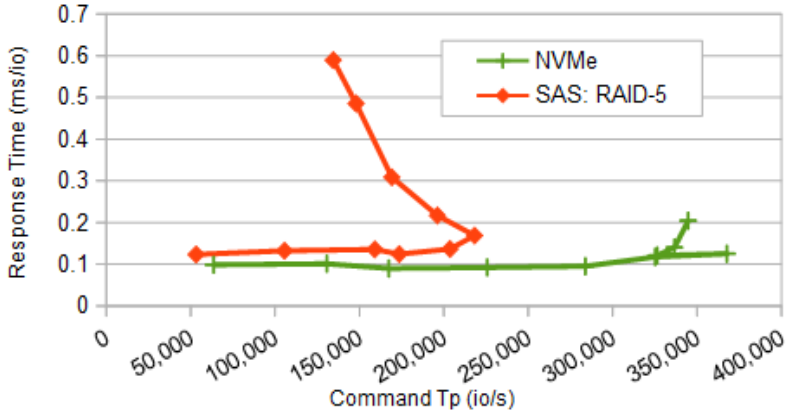
Take-aways:

1. Due to IBM i's Storage Management's excellent data placement (use of buffers), the high transaction rate stresses CPU.
2. CPU utilization efficiency cost for doing RAID function in the host CPUs does allow SAS to use less CPU/io than NVMe.
3. Yet response times for NVMe writes still look A-OK. (i.e. lower than SAS.)

Small Op Random Reads : Mirror vs RAID-5



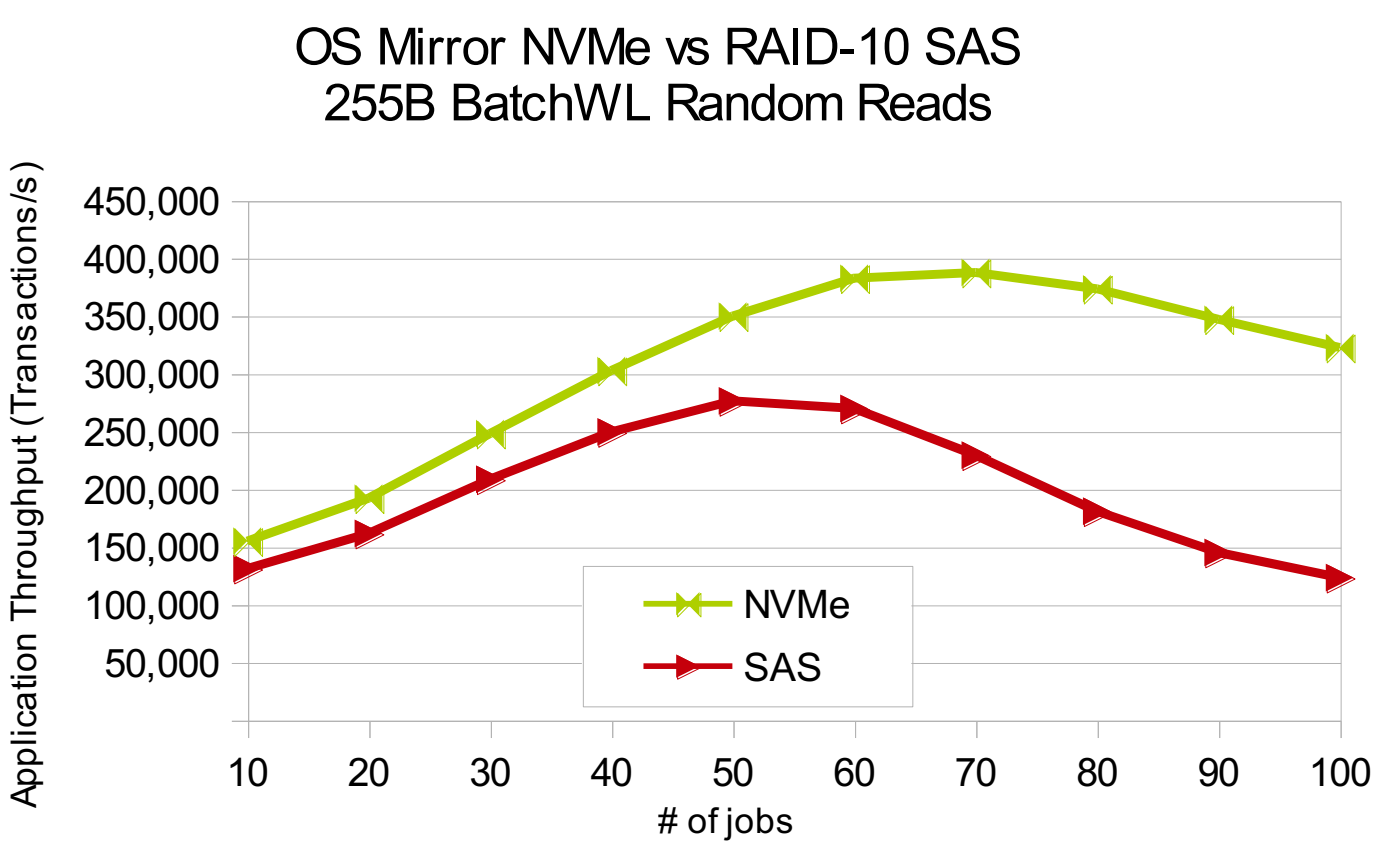


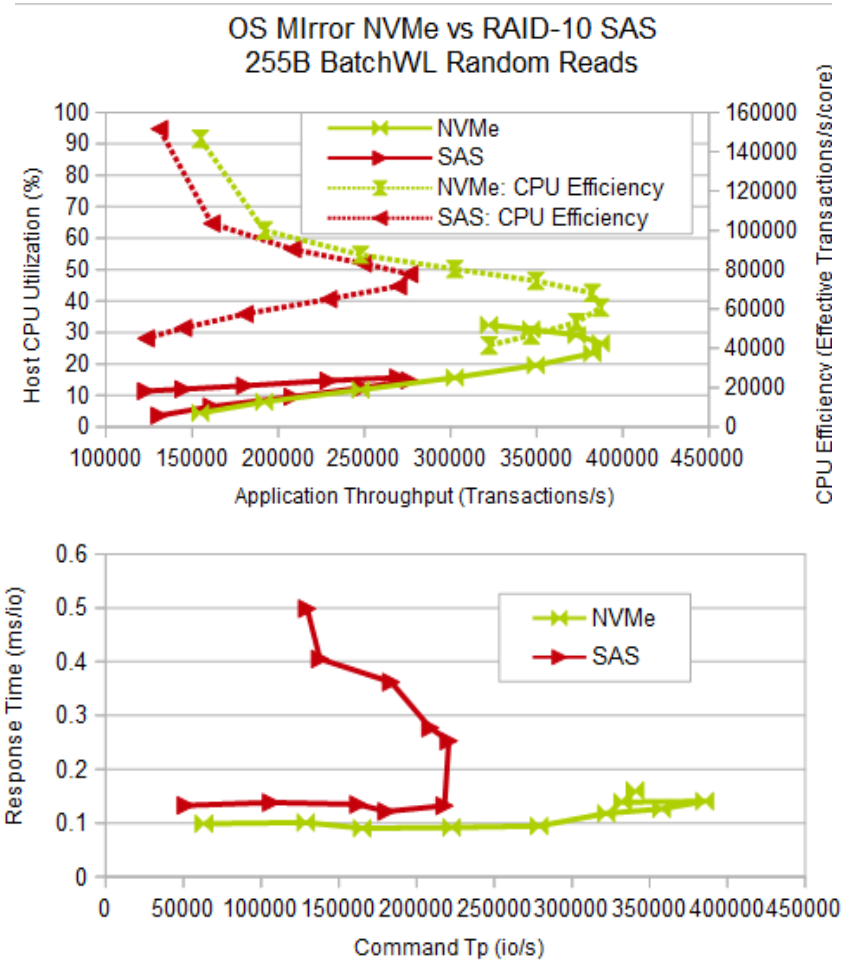


Take-aways:

1. High correlation between application and subsystem command throughput confirms that caches do not help truly random reads. Thus the raw capabilities of the 2 NVMe drives versus the 18 SAS drives (and 2 IOAs) shines through.
2. Eliminating the SAS adapter chip and bus lowers response times (latency).
3. Nice to see NVMe host CPU mirroring code impact is actually lower !

Small Op Random Reads : Mirror vs Mirror

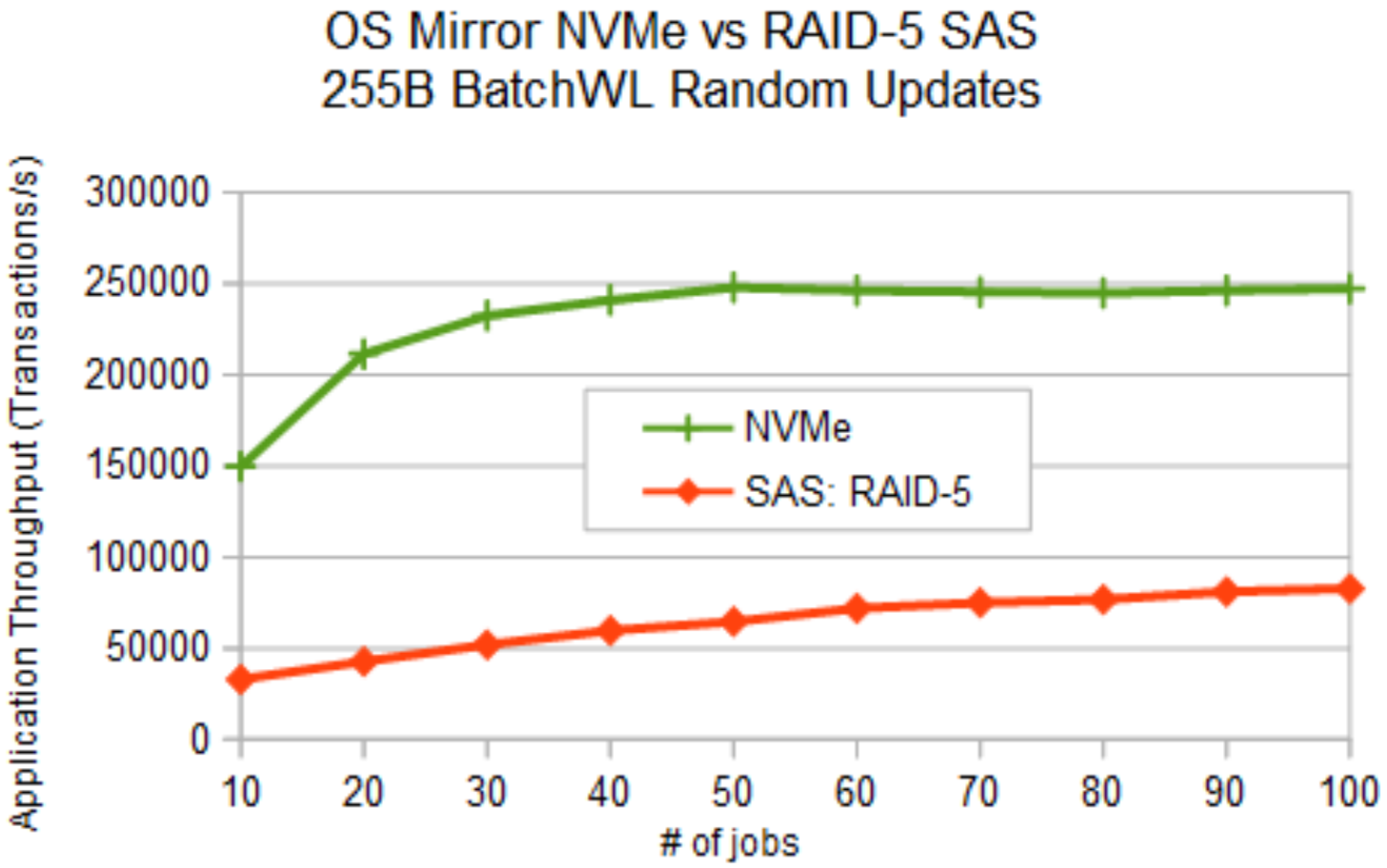


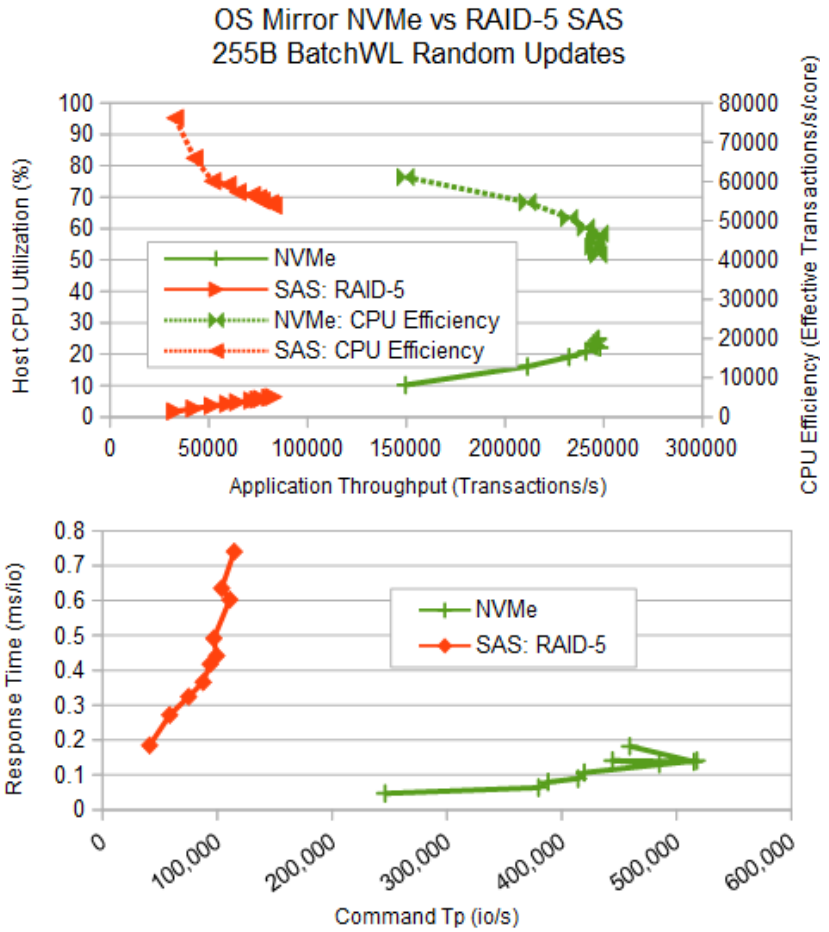


Take-aways:

1. High correlation between application and subsystem command throughput confirms that caches do not help truly random reads. Thus the raw capabilities of the 2 NVMe drives versus the 18 SAS drives (and 2 IOAs) shines through.
2. Eliminating the SAS adapter chip and bus lowers response times (latency).
3. Nice to see NVMe host CPU mirroring code impact is actually lower !

Small Op Random Updates : Mirror vs RAID-5

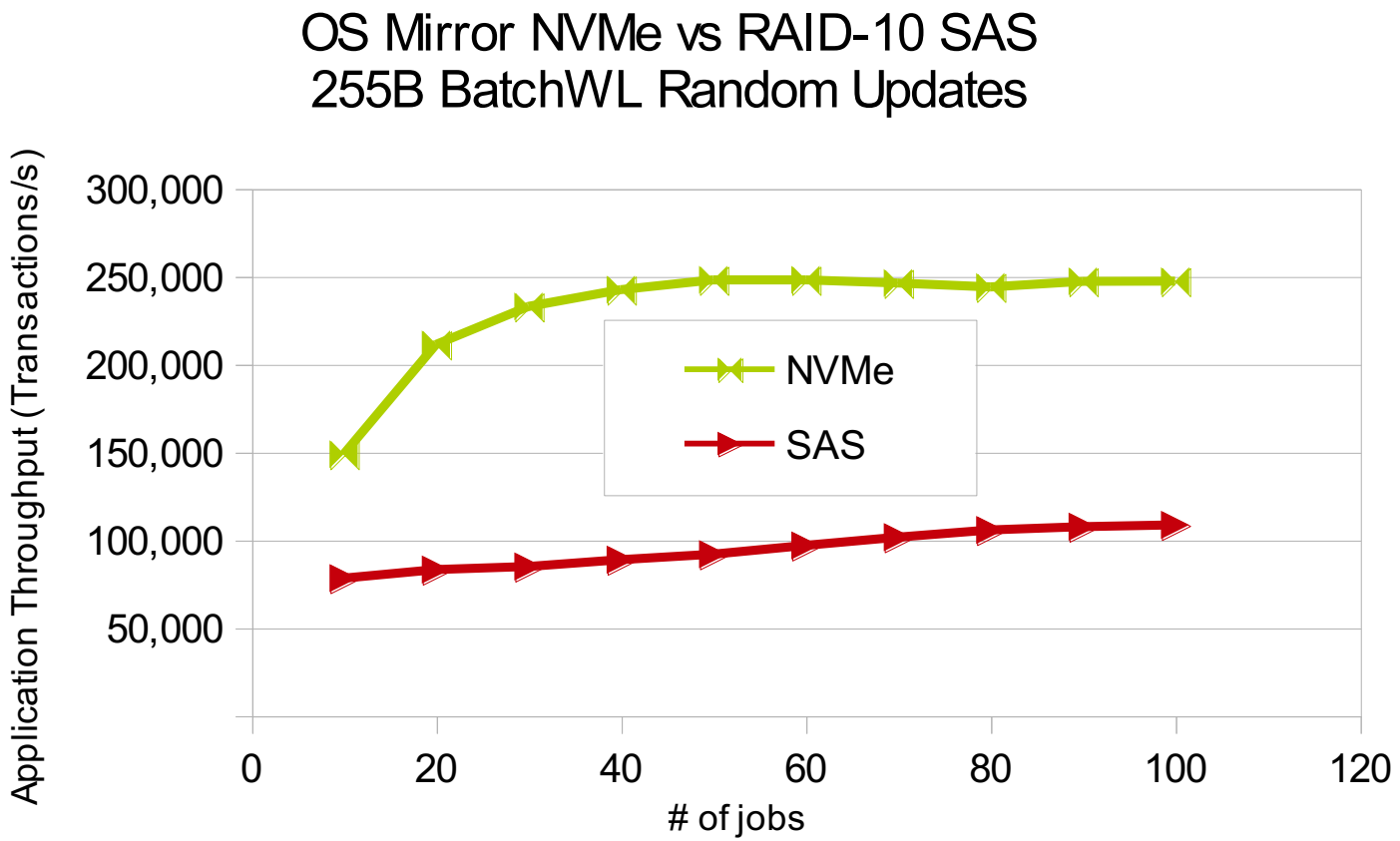


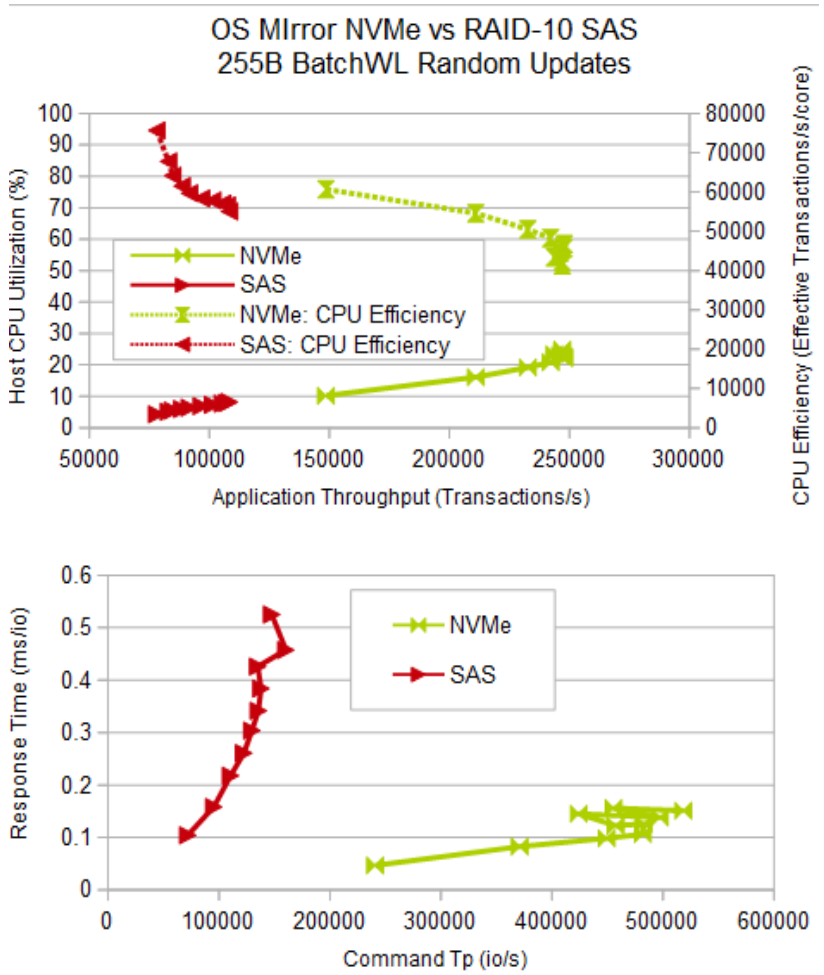


Take-aways:

1. Coupling the RAID-5 write cost and raw NVMe subsystem latency reductions, migrating from SAS RAID-5/6 to NVMe mirrored subsystem will provide LOTS of headroom for application throughput growth.
2. CPU Efficiency at a given throughput is fairly close despite the fact that NVMe uses CPU cycles to manage protection while neither adapter level protection does.

Small Op Random Updates : Mirror vs Mirror





Take-aways:

1. Running SAS in a mirror does allow it to support more throughput than in RAID-5 mode, but still not nearly enough to keep up with the mirrored NVMe config.
2. CPU Efficiency at a given throughput is fairly close despite the fact that NVMe uses CPU cycles to manage protection while neither adapter level protection does.