# NVMe devices & IBM i

## A closer look...

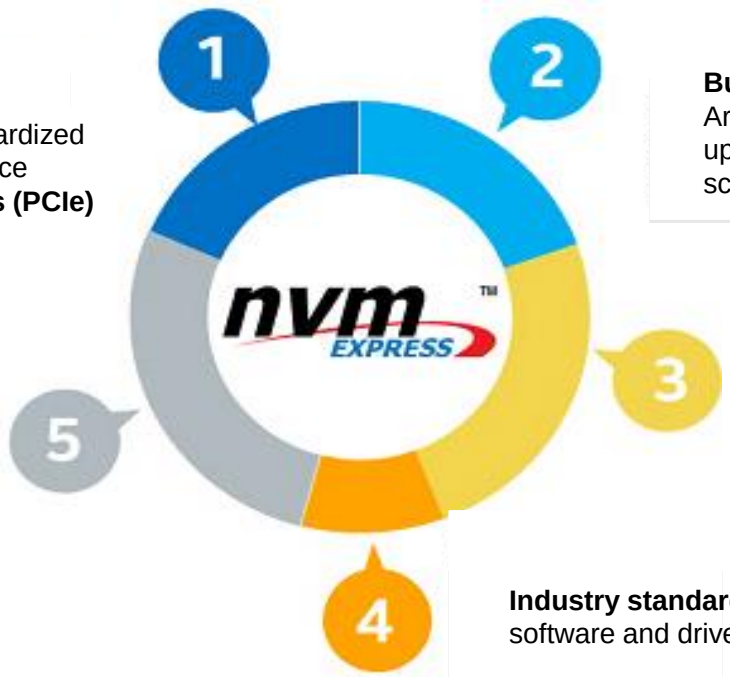
*COMMON Luxembourg*
*20/02/2020*

*Fabian Michel*
*Client Technical Architect*

# What is NVMe?

NVMe - Non-Volatile Memory express
http://www.nvmexpress.org/

NVM Express is a standardized high performance interface protocol for **PCI Express (PCIe)** Solid State Drives




**Built for SSDs**
Architected from the ground up for SSDs to be efficient, scalable, and manageable

**Industry consortium spec** defines a queuing interface, command set, and feature set to optimize performance

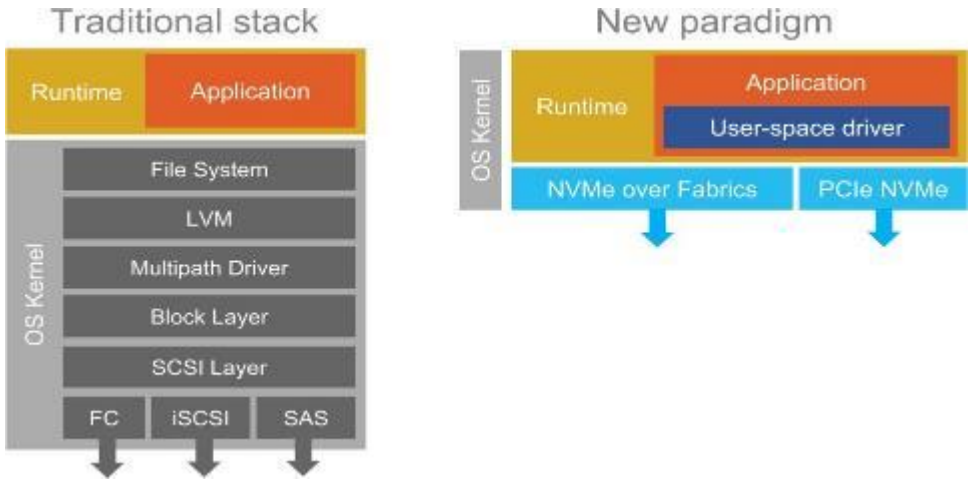
Available in different form factors such as: PCIe Add in Card, SFF 2.5” U.2 and M.2

**Industry standard** software and drives

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) | DRAMDIMM | <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

NEW

## 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**

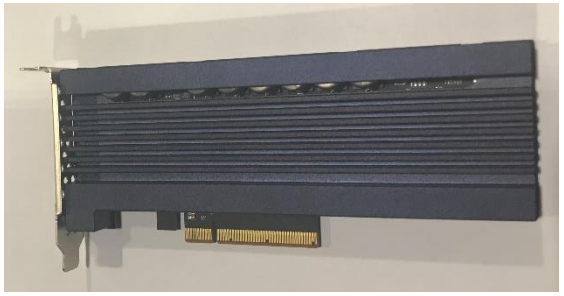
| | PCIe3 NVMe Flash Adapter | | |
|---|---|---|---|
| | **1.6TB** | **3.2TB** | **6.4TB** |
| FC (LP/FH) | EC5G / EC5B | EC5C / EC5D | EC5E / EC5F |
| **IBM i FC (LP/FH)** | **EC6U / EC6V** | **EC6W / EC6X** | **EC6Y / EC6Z** |

| 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)



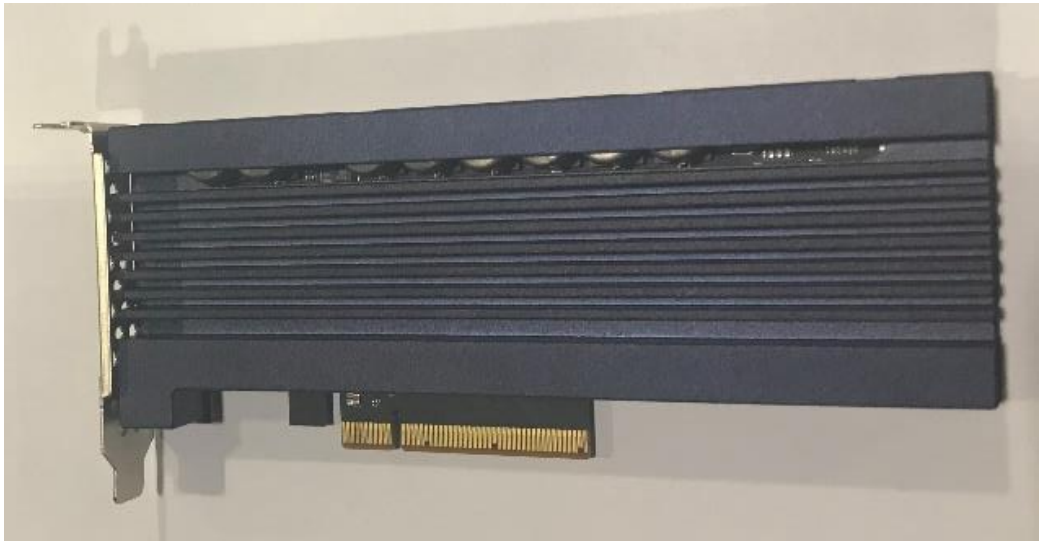
nvm EXPRESS

# 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)

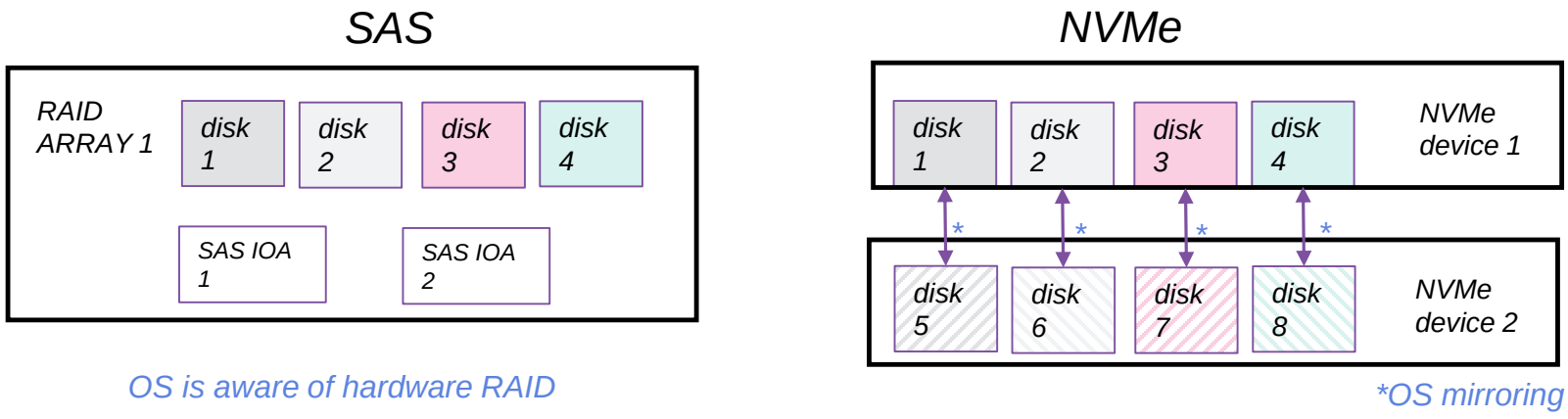


# NVMe Namespaces and IBM i

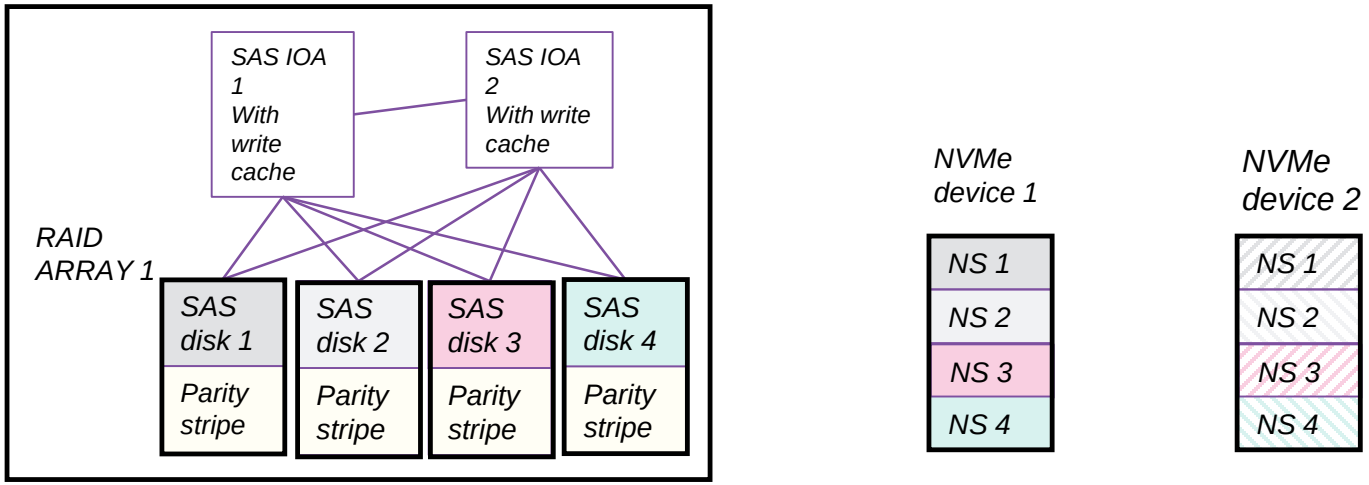
- **NVMe uses namespaces** which is a collection of logical blocks whose logical block addresses range from 0 to the size of the namespace. A namespace ID (NSID) is an identifier used by a controller to provide access to a namespace.
- **With NVMe, an 'arm' (logical unit) is a namespace**. A namespace 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*



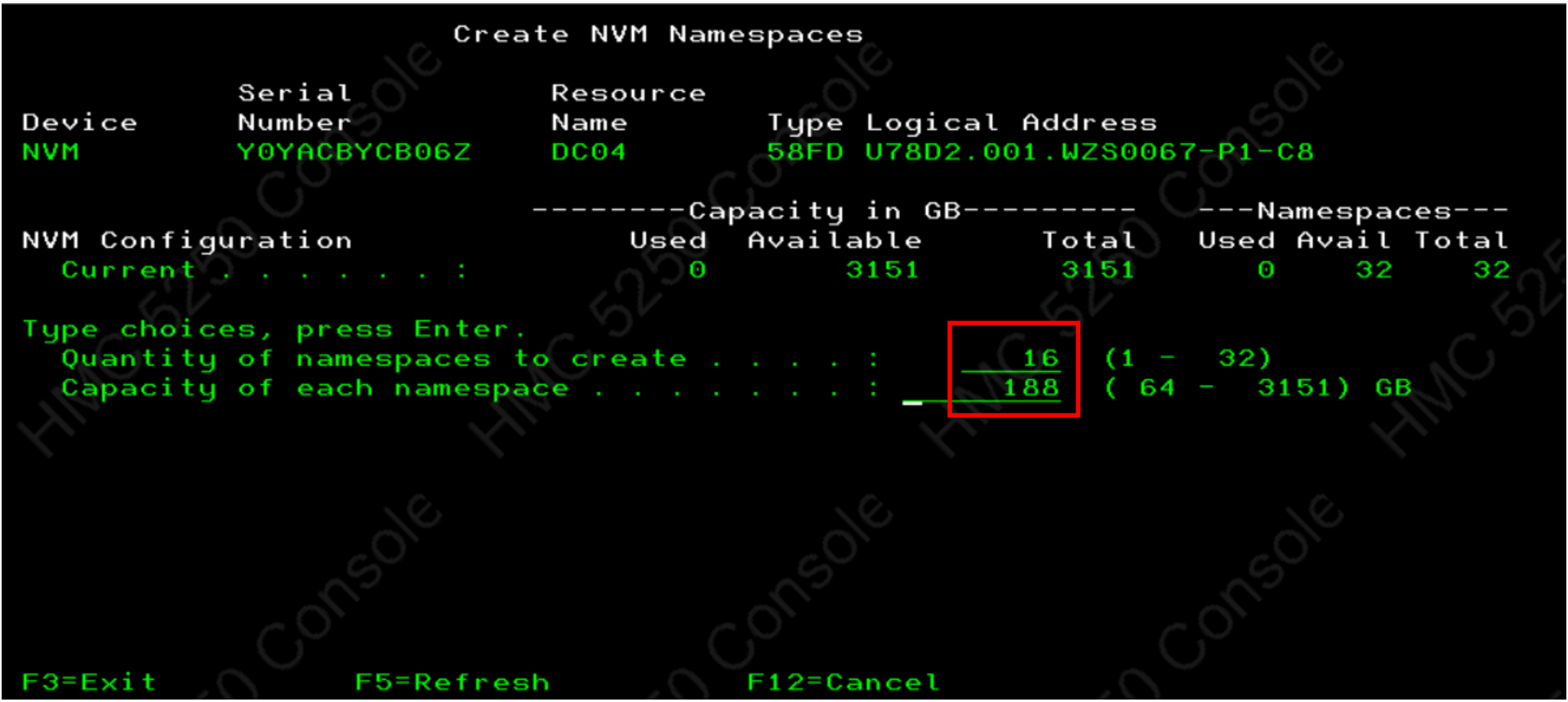

*Device physical view*

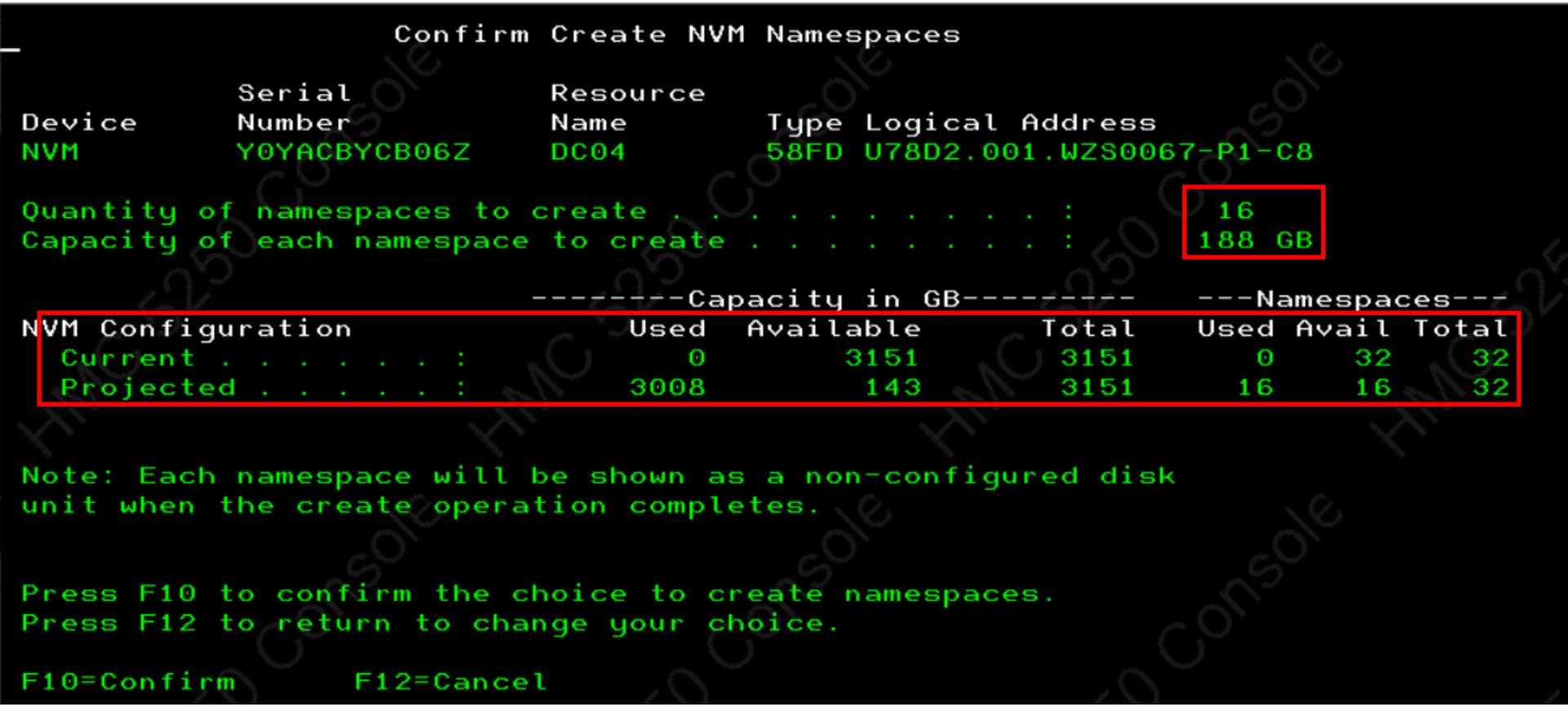


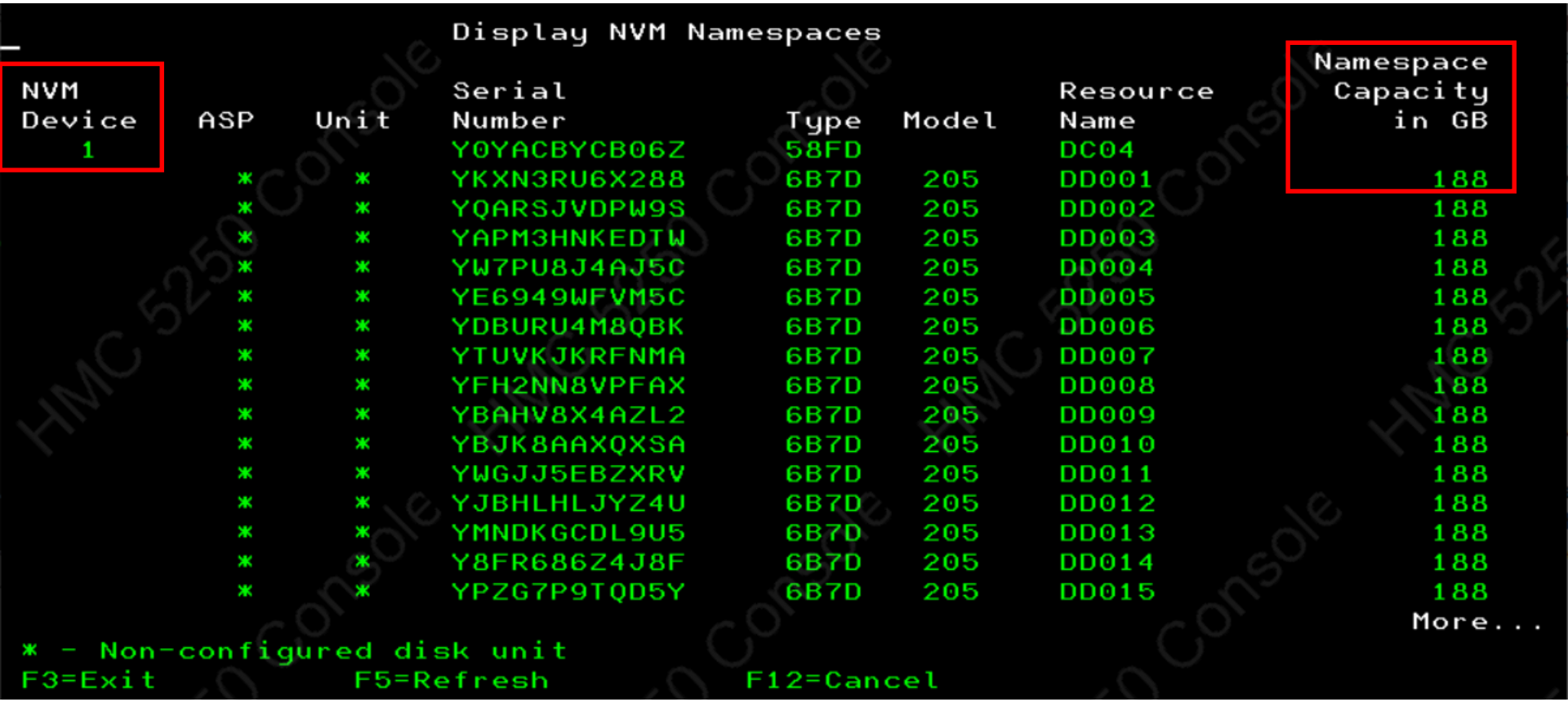
# Recommended NVMe Namespace Sizes

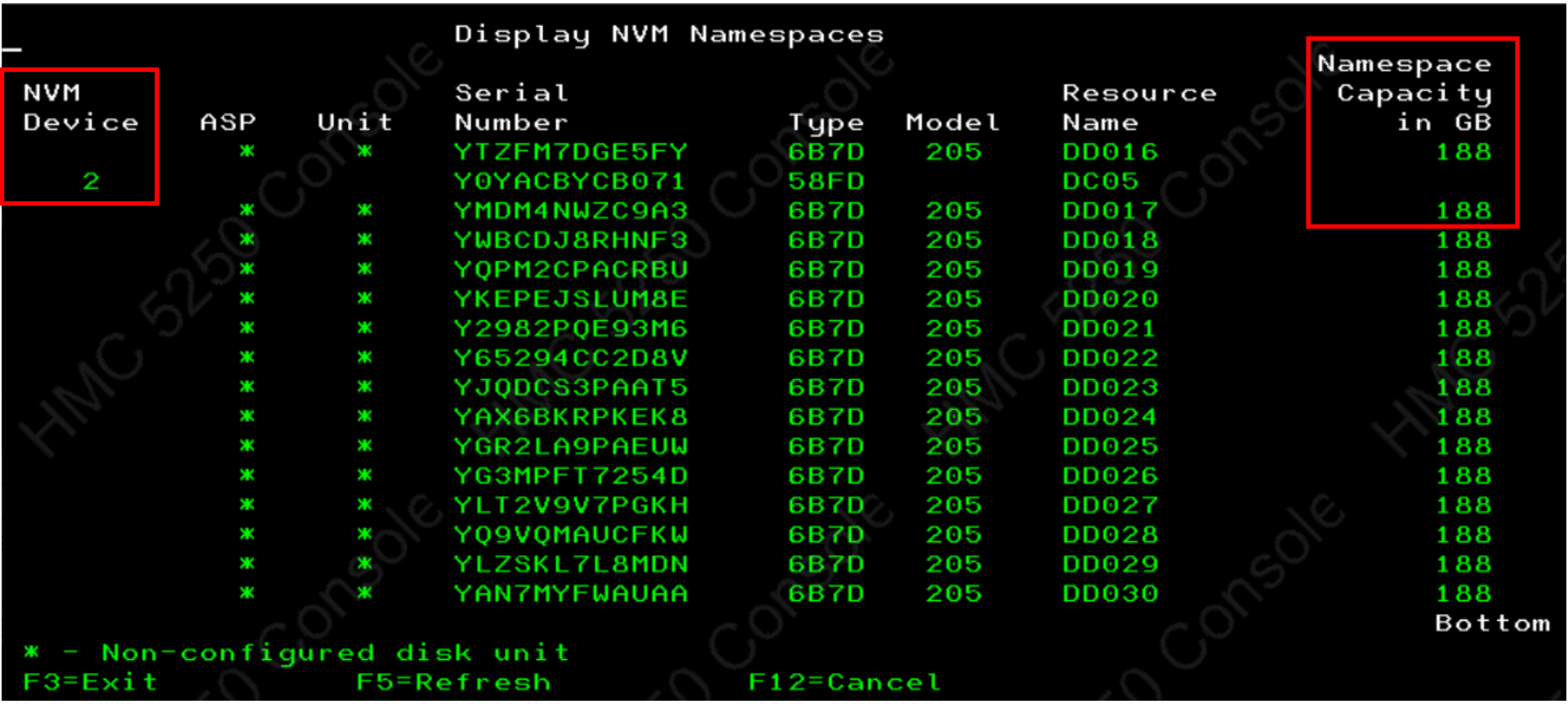
- First generation NVMe devices have a hardware **boundary of 16 GB for name spaces**. Device capacity can be wasted/lost if name spaces are not multiples of 16. The **maximum number of namespaces on a device is 32**.
- IBM i screens show Capacities in 'GB' (1000**3 (GB), not 1024**3 (GiB))
- IBM recommends **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)

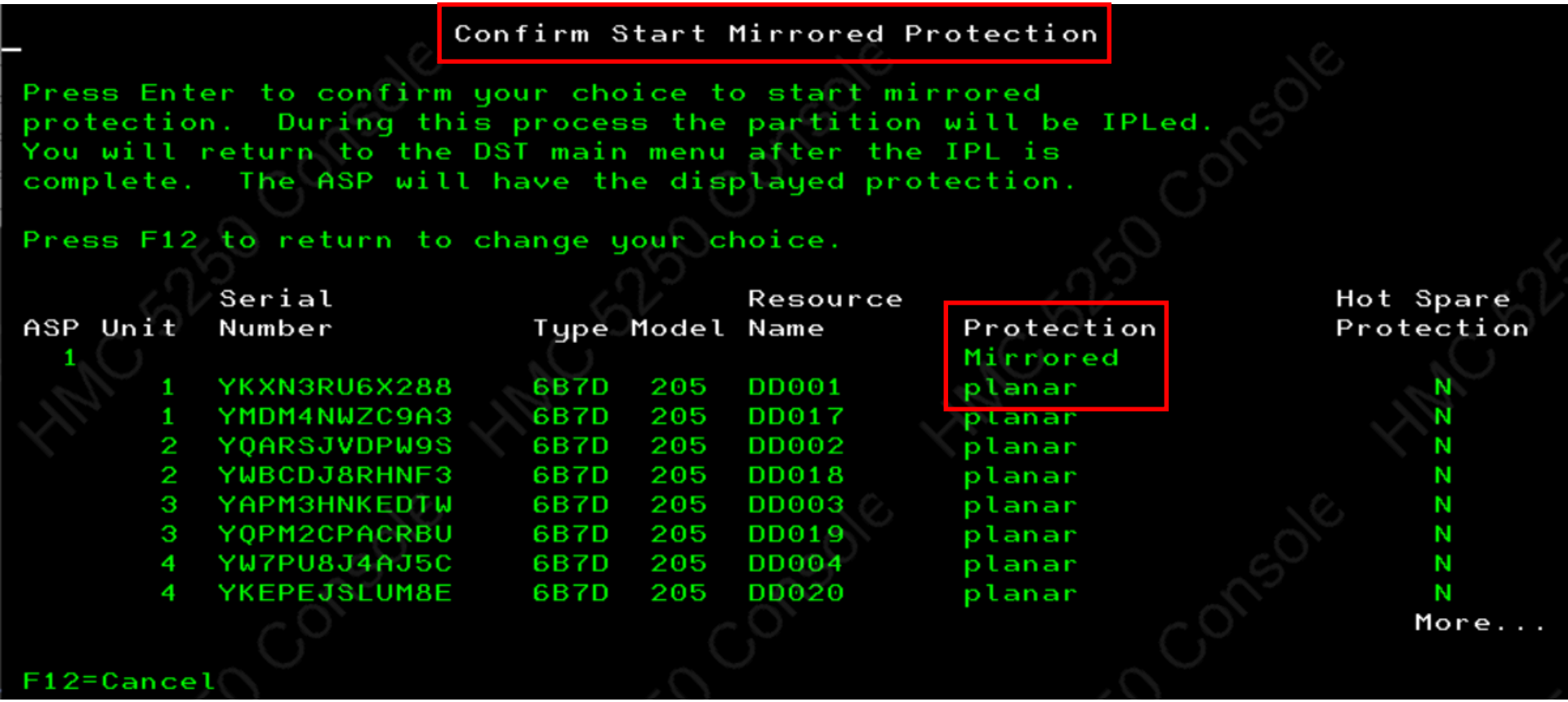
| Device Nominal Size | Device Actual Size | Number of Namespaces | Namespace Size | Total User Capacity Used By Namespace | Remaining Space on the Device (unallocated) |
|---|---|---|---|---|---|
| 1.6TB | 1575 | 8 | 188 | 1448 | 87 |
| 3.2TB | 3151 | 16 | 188 | 2977 | 174 |
| 6.4TB | 6364 | 32 | 188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

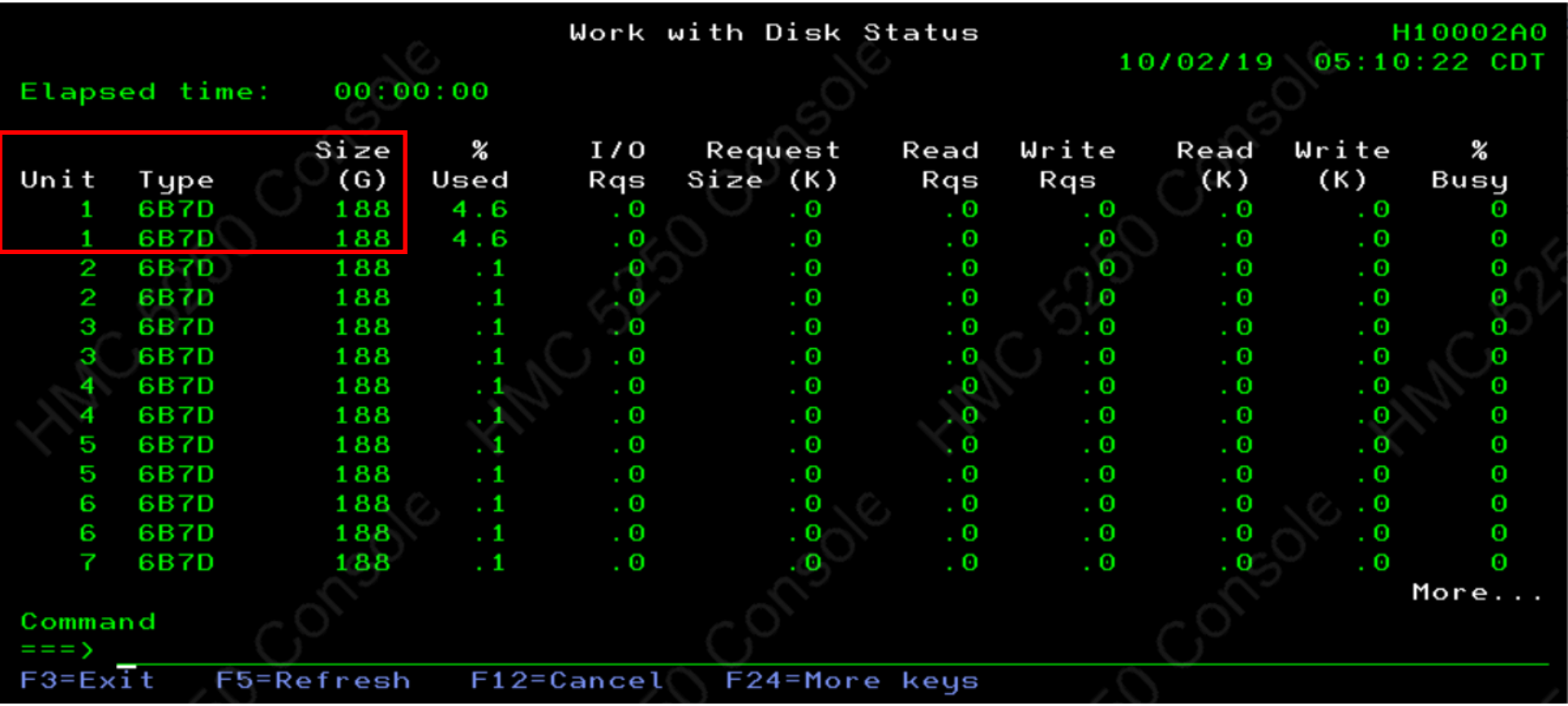




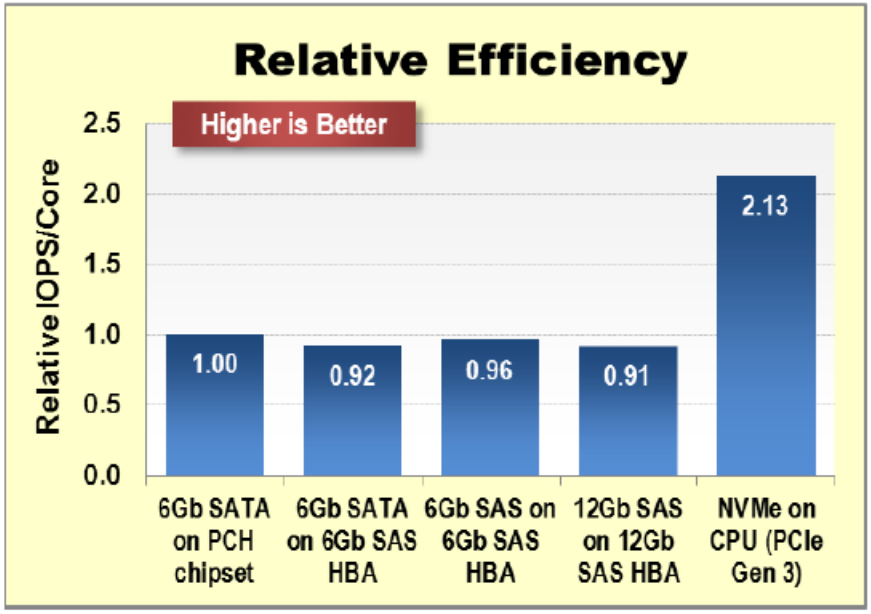


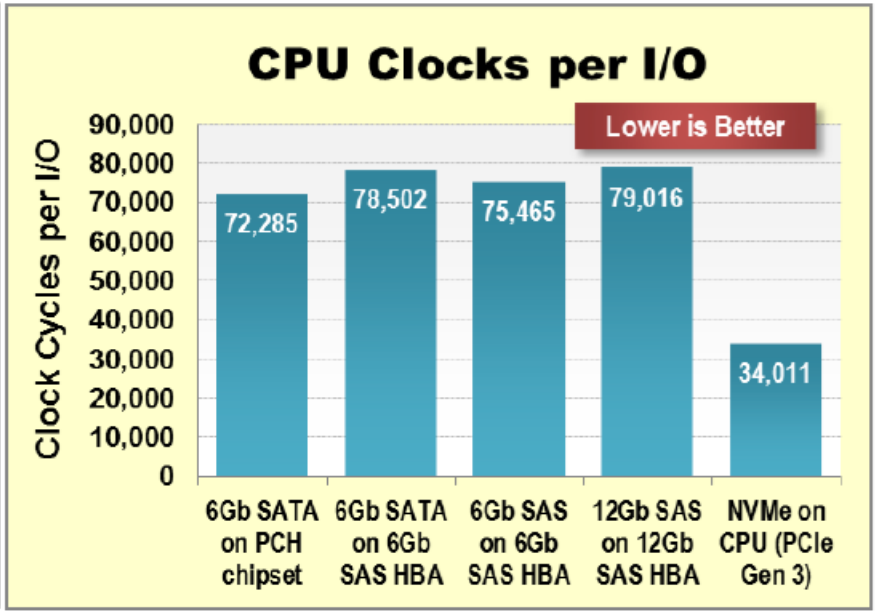




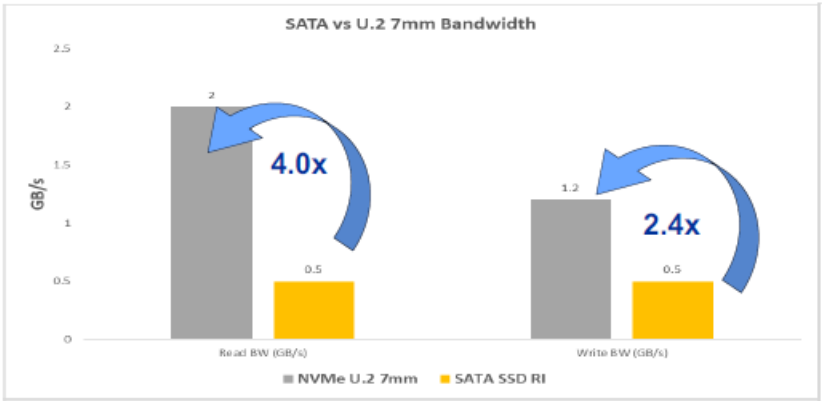


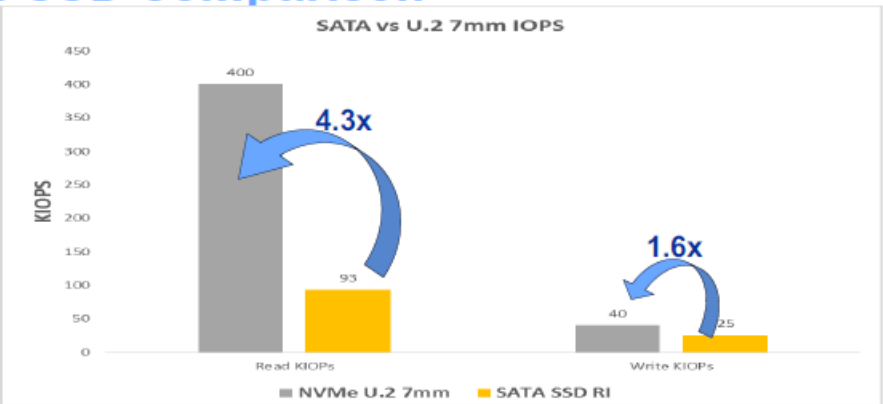
## Is NVMe Fast?

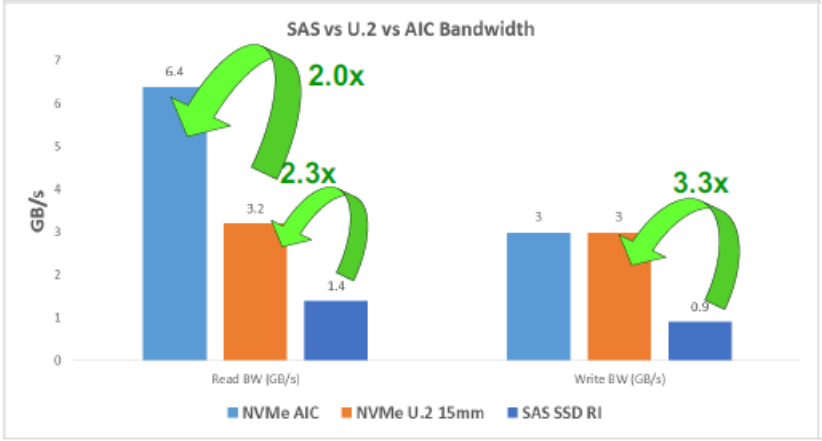


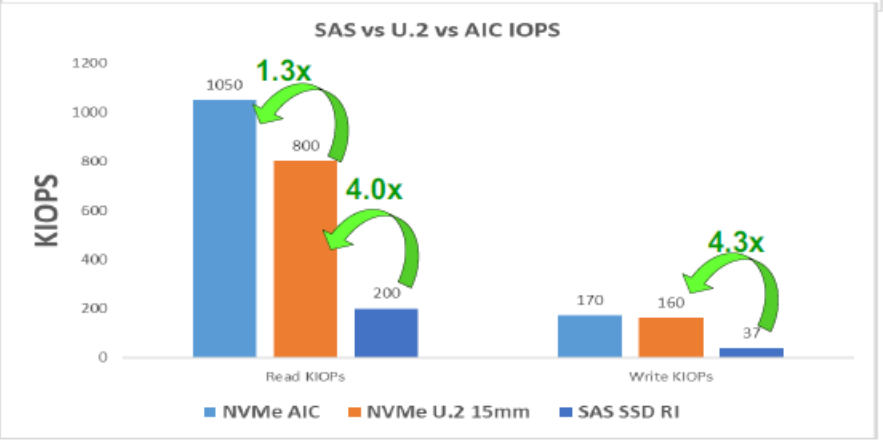



# SATA vs SAS vs NVMe SSD Comparison



# 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 |