The future of storage as we know it.

Non-Volatile Memory Express (NVMe) is specifically designed to deliver high bandwidth, low latency access to storage devices over a PCI-Express interface.

In partnership with Supermicro, We're proud to have an extensive range of NVMe based platforms which can suit a variety of different applications.

Originally released as a version 1 specification by the NVM Express Work Group in March 2011 and through several minor revisions since, NVMe has gone from conception to maturity and is now tipped to be one of the most important storage technologies for both server and client computing in the coming decade and beyond.

A collaboration between key industry players including Cisco, Dell, EMC, Intel, Micron, Microsemi, Microsoft, NetApp, Oracle, Samsung, Seagate and Western Digital, NVMe as a technology has had both the input and the backing necessary to make it the de-facto high performance storage interface of today, with great foundations of becoming used as mainstream storage for future deployments.

WHAT EXACTLY IS NVMe?

Non-Volatile Memory Express (or NVMe) is a collection of open standards which can be used to design and implement non-volatile memory products.

It's been specifically designed to deliver high bandwidth, low latency access to storage devices over a PCI-Express interface.

That's not all, however, the standard defines both the interface, command set and goes further to define management and implementation of NVMe over scalable fabrics, giving potential to finally rival Fibre Channel, or SCSI /SAS protocol in a SAN environment.

The three definitions are as below:

- NVM Express – The register interface and command set for PCI Express attached storage with industry standard software available for numerous operating systems. NVMe is widely considered the defacto industry standard for PCIe SSDs

- NVMe Management Interface – The command set and architecture for out of band management of NVM Express storage (e.g., discovering, monitoring and updating NVMe devices using a BMC).

- NVMe over Fabrics – The extension to NVM Express that enables tunnelling the NVM Express command set over additional transports beyond PCIe. NVMe over Fabrics extends the benefits of efficient storage architecture at scale in the world's largest data centres by allowing the same protocol to extend over various networked interfaces.

These standards are important to help unify the industry in their efforts, avoiding the generation of multiple different and possibly propriety designs becoming common, fragmenting the market and causing consumer confusion.

Additionally, having a standard NVMe I/O stack, means that most operating systems can include an NVMe driver which is compatible with drives from a variety of different vendors directly in their distribution distribution – in fact Windows Server 2012 R2, Windows Server 2016 and Linux distributions with a kernel of 3.3 or greater are NVMe ready today. These operating systems should provide at least the basic operating level of functionality without having to install a vendors' own third party code, which was previously necessary for practically all PCI-E based flash devices that came before the NVMe standard.

WHAT DOES IT LOOK LIKE?

Today's implementations of NVMe are available as low profile x4 or x8 PCI-Express based SSDs, PCI-Express x4 U.2 (2.5" / SFF) devices or as M.2 embedded devices.

The standard PCI-Express x4 or x8 add in card form factor's main advantage is that it has compatibility with legacy systems which don't include a U.2 form factor slot and enables them to take advantage of NVMe's enhanced pefornmance and standardised stack design. It works just as well with regards to performance, and has the potential to be faster when PCI-Express x8 is utilised, but it's really U.2 that really makes NVMe interesting for server deployment.

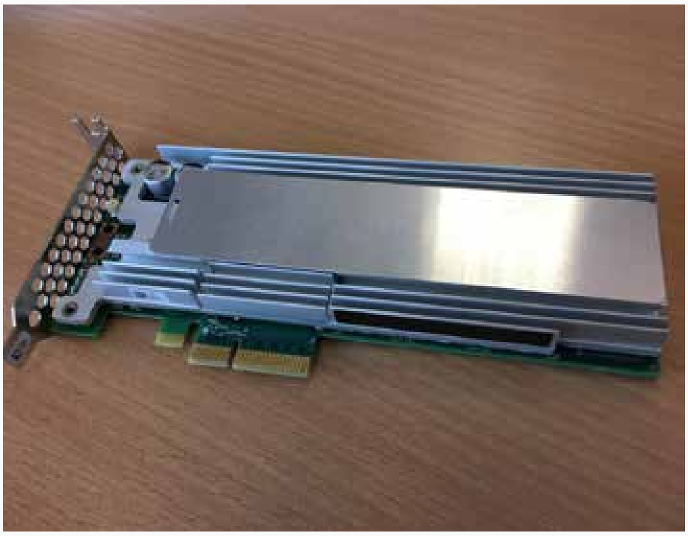


Figure 1:Intel P3700 PCI-Express x4 NVMe device

U.2/2.5" SFF PCI-EXPRESS x4 FORM FACTOR

Firstly, you'll notice that U.2 is actually using the same 2.5" form factor and a similar connector style as existing SAS / SATA devices. This is very useful for system designers as it means that they can build in "hybrid" 2.5" bays to support both types of device depending on the controller being used.

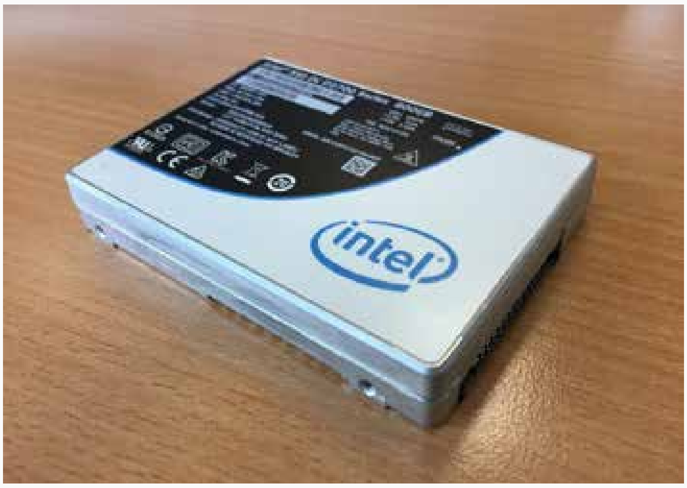


Figure 2:Intel D3700 U.2 Form Factor NVMe device

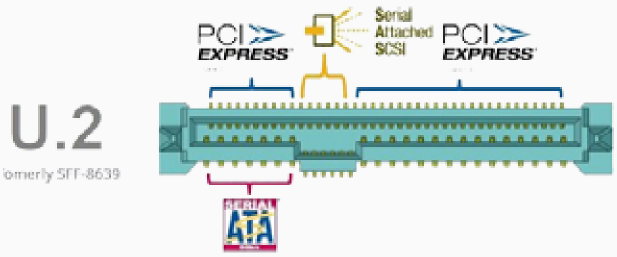


Figure 3:The anatomy of a U.2 connector

Supermicro for example have systems with 10 x SAS/S-ATA bays, with 2 of those bays designated as NVMe ready, denoted by the orange release tab.

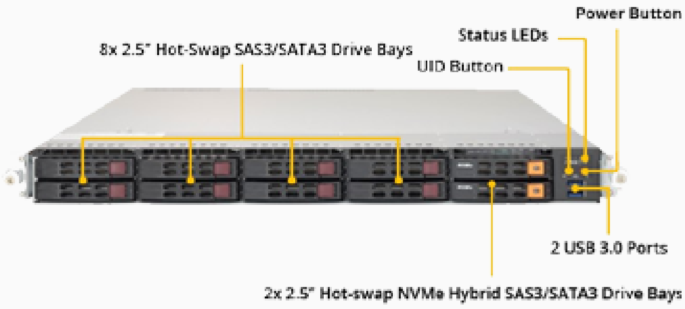

Power Button
Status LEDs
8x 2.5" Hot-Swap SAS3/SATA3 Drive Bays
UID Button
2 USB 3.0 Ports
2x 2.5" Hot-swap NVMe Hybrid SAS3/SATA3 Drive Bays

Figure 4 :Supermicro 1U Ultra server with hybrid NVMe U.2 slots

Additionally, U.2 devices support simple hot plug support when coupled with the right controller and operating system, but note that not all implementations support hot plug this so be cautious before pulling drives without notice. Standard internal PCI-Express x4 add-in cards don't typically support hot plug at all, although it is technically possible when carefully designed and executed. With the clumsy add-in card installation and traditional screw to fix them in place, add in cards certainly don't lend themselves to this application nearly as well as U.2 does.

Finally, U.2 also supports splitting the PCI-Express x4 connector on the device to two PCI-Express x2 connectors, effectively making a dual port device. This requires explicit support from the drive though, and this is currently limited to just a few models on the market which have been designed with dual port in mind.

Dual port is useful for applications where drives could be shared between systems for the purposes of failover. This kind of implementation is already common with Fibre Channel and SAS devices, and is used where high availability is important for the system design.

Supermicro have a storage centric platform based around this concept using the storage bridge bay (SBB) platform of two x86 based servers, in one enclosure, sharing an NVMe dual port hot plug backplane.

Figure 5:Supermicro's storage Bridge Bay

Such types of shared disk implementations work well with failover clustering configurations which use software disk pooling, such as Windows Storage Spaces or ZFS. Using these technologies together, a high availability, yet high performance all flash storage array, or hyper converged storage and virtualisation application becomes a real possibility. You will find implementations based on this exact architecture in our upcoming Igloo generation of SAN and NAS devices.

M.2 SFF/ EMBEDDED FORM FACTOR

The M.2 form factor packs everything that the standard add-in card implementation does, but delivers it in a compact device which is perfect for boot and embedded computing

This type of connector is already commonly found on desktop or gaming motherboards as a high end storage connector, but more recently has been incorporated into more enterprise / server grade platforms to be used as boot or caching devices.

Drives are available in a variety of different widths and lengths, but generally most are 22mm wide and between 32 and 100mm long. Note though, it's important to ensure your motherboard or M.2 adaptor card can support the model you wish to use before making a purchase.

Figure 6:Intel P3100 M.2 NVMe device

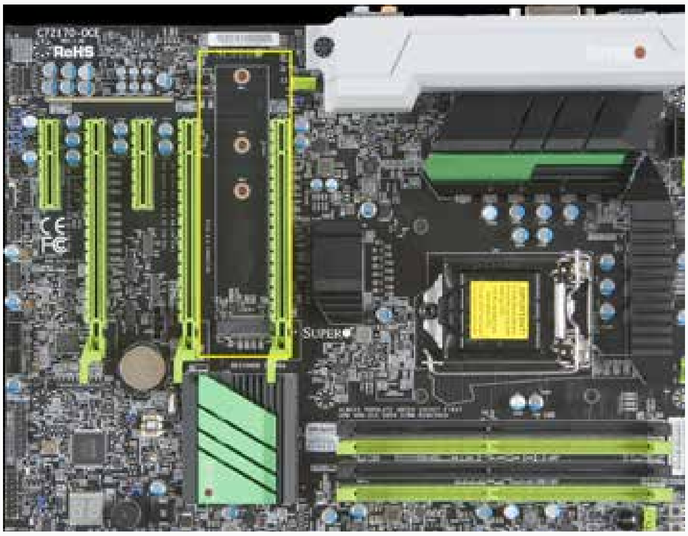


Figure 7:Supermicro C7Z170-OCE motherboard supporting M.2

The above Supermicro motherboard for example, supports M.2 devices of either 60, 80 or 110mm long, as highlighted by the yellow box. There are supporting screw holes at these distances to enable the M.2 device to be fixed into place, any other length could not be properly fitted as a result.

Another consideration with M.2, is that there are three different implementations of M.2, and they are not all compatible. These are, legacy S-ATA, PCI-Express using AHCI and PCI-Express using NVMe.

It is essential to pair your device and slot accordingly, as installing NVMe devices into S-ATA based M.2 connectors or vice versa is not a functional combination, so be careful!

PERFORMANCE

As you would imagine, being designed with flash storage as a primary consideration, NVMe is a great improvement over the commonly used SAS and S-ATA interfaces which it looks to replace.

At face value, a single NVMe interface for example has more than double the bandwidth of a single SAS 3 device, and more than 5 times that of a Single S-ATA 3 port.

NVMe	S-ATA 3	SAS 3
400MB/s	750MB/s	1500MB/s

Table 1: Theoretical maximum throughput of NVMe, SATA and SAS today

Comparing the typical specifications of a S-ATA or SAS SSD's to NVMe, we can see that NVMe drives generally promise much greater bandwidth and hundreds of thousands of IOPs more. The table on the next page compares different SSD devices that are all generally considered to be in the higher performance category for their interface.

There is a considerable increase in throughput from S-ATA to SAS and from SAS in turn to NVMe. I/O per second metrics also take large jumps too. This will be in part due to the bandwidth improvements of the bus and protocol efficiency, however it is also partly due to faster controllers and flash devices being used which can take advantage of this.

DEVICE	Intel P3700 U.2 2TB	Intel S3700 2.5" 800GB	Seagate 1200.2 2.5" 1600GB
INFERFACE	PCI-EXpress x4 Gen 3	S-ATA 6Gbit/s	SAS 12Gbit/s
SEQUENTIAL READ MAX	2800MB/s	500MB/s	1900/1100MB/s (Dual / Single Port)
SEQUENTIAL WRITE MAX	1900MB/s	460MB/s	850MB/s
RANDOM READ	450,000 IOPS	75,000 IOPS	200,000 IOPS
RANDOM WRITE	175,000 IOPS	32,000 IOPS	80,000 IOPS
LATENCY READ/WRITE	20uS/20uS	50us/65uS	115uS Average

Table 2: SDD device comparison

But that's not all, overall access latency on NVMe is lower by as much as 200 micro seconds or 50% in some cases. This improvement is the result of as S-ATA and SAS devices having to be connected via a HBA or the platform's I/O chipset, instead of being direct on the PCI-Express bus. The HBA or chipset itself in turn connects to the PCI-Express bus in the processor, adding an additional step and incurring extra latency.

DATA PROTECTION / RAID

In the storage industry, it's become an accepted inevitability that one disk device isn't enough to satisfy an applications performance or availability requirements. As a result, the concept of RAID was developed to group drives together and to handle some of the technical limitations of traditional spinning hard disks.

Whilst NVMe solves some of the performance shortfalls and has a lower fail rate than mechanical disks, applications will always push the performance limits of technology and it is still susceptible to failures. In turn, RAID has become a hot topic for NVMe, as there are additional complexities to consider when combining the two technologies.

Firstly, with PCI-Express being the native interface for NVMe, most hardware RAID controllers (themselves PCI-E x8 or x16 devices) are not well equipped to deal with their bandwidth requirements. Aggregating 8 x PCI-Express x4 NVMe devices with RAID would require a PCI-Express x32 interfaced RAID controller to be truly non-blocking for example.

Secondly, with NVMe devices having such high performance I/O and throughput characteristics, your typical hardware based RAID chip will struggle handle the throughput of several devices in parallel, let alone the parity calculations necessary for a RAID 5 or 6 implementation.

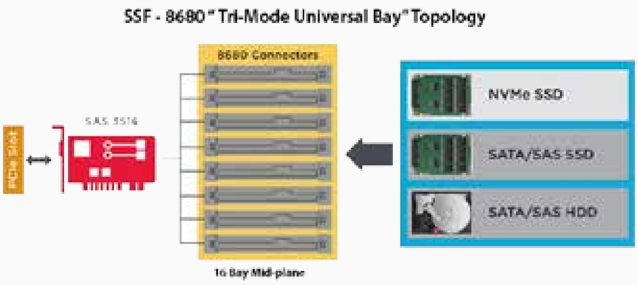
This has led to a new generation of RAID implementations which look to work around these limitations to get the best from NVMe.

HARDWARE RAID

Instead of being dedicated to NVMe devices, the hardware based RAID controllers which have been launched with NVMe support so far have focused combining drive technologies to get the best features of each.

Broadcom's Tri-Mode models for example, combine NVMe drives with traditional SAS and S-ATA devices in a hybrid deployment.

The controller supports a maximum of up to 16 directly connected devices of all three interface types, NVMe, SAS3 or S-ATA3, but more can be connected with expanders or switches.



Strictly, with a x8 PCI-Express connector on the RAID controller card for connectivity to the host, you can install up to two x4 NVMe drives before you saturate the available bandwidth. You can continue to add as many as 8 NVMe devices directly to the controller, which will share the available bandwidth, but it's likely that the card's own link will become a bottleneck. Support increases to as many as 16 devices when used with an additional PCI-Express based switch on the backplane, but naturally there will be a serious limitation to the performance of those devices, should they all be active at the same time.

As a result, it's useful if you needed to install a number of NVMe devices with a small number of PCI-Express lanes available in your system, or to build RAID volumes with protection and maintain a degree of performance too. Broadcom state that the chip is capable of over 1 million IOPS, which is certainly more than enough for most applications.

An optimum application for this controller, however, is actually for hybrid storage use, where NVMe drives are paired with traditional S-ATA / SAS spinning disks, allowing the benefits of both to be realised. With up to 10TB HDD drives for capacity on the market today, combined with NVMe drives used as a cache or hot data tier, greatly cost effective yet viable solution for most applications is available.

SOFTWARE RAID / DISK POOLING

Rather than travel down the hardware path for NVMe RAID, many of our partners are deciding to concentrate their efforts on software based solutions. This is becoming common in the industry in general for all kinds of storage applications too, with disk pooling and software defined everything trending from datacentres right through to the enterprise.

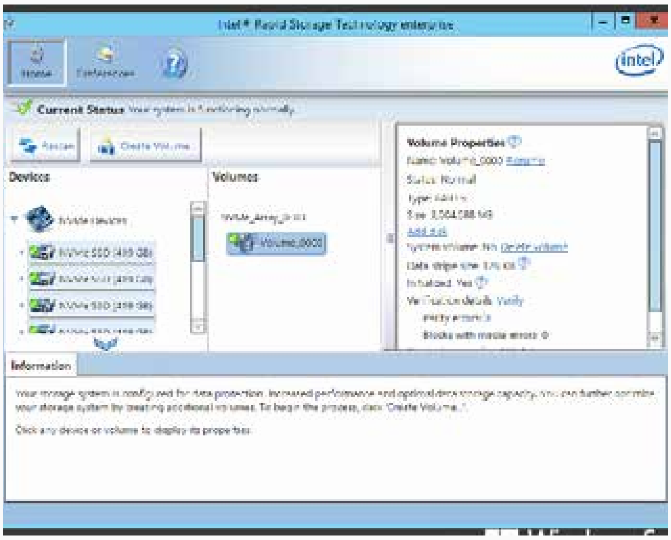
The concept is simple; your system processor is incredibly powerful and consists of more cores than are generally utilised by modern applications. Why not take the CPU cycles that might be wasted and have them take care of RAID or I/O that might otherwise require costly dedicated hardware?

There are several options for disk pooling or software RAID, many of which can be considered for both hard disks and SSD's alike; each has their own specific advantages.

Here are some of those that we've worked with.

INTEL RSTe FOR NVMe

Available for both Windows and Linux systems alike, RSTe is a simple and free to use (for Intel-NVMe drives) RAID engine.



Current Status Your system is functioning normally.

Volume Properties
Name: Volume_0000 Rename
Status: Normal
Type: RAID 5
Size: 3,564,088 MB
Add disk
System volume: No Delete volume
Data stripe size: 128 KB
Initialized: Yes
Verification details Verify
Parity errors: 0
Blocks with media errors: 0

Your storage system is configured for data protection, increased performance and optimal data storage capacity. You can further optimize your storage system by creating additional volumes. To begin the process, click 'Create Volume.'

Click any device or volume to display its properties.

It supports RAID levels 0,1,5 and 10 and is driven by an easy to follow a GUI under Windows, with an optional CLI for more advanced administrators.

Under Linux, it uses the familiar mdadm built in software RAID layer utility for configuration via a CLI.

Performance is generally good and volume can be configured which deliver hundreds of thousands of IOPs.

Currently, it's not possible to boot from volumes built by RSTe, so a boot device is still required.

WINDOWS STORAGE SPACES

Microsoft provide the built in Windows Storage Spaces disk pooling technology built into most of it's operating systems as standard since Windows Server 2012 in the enterprise space or Windows 8 on the desktop.

Storage spaces allows any type of drive from any vendor to be pooled and presented as a virtual disk with stripe, mirror, or parity data layout. This gives similar performance and data protection capabilities as a RAID 0,1 and 5 configurations respectively.

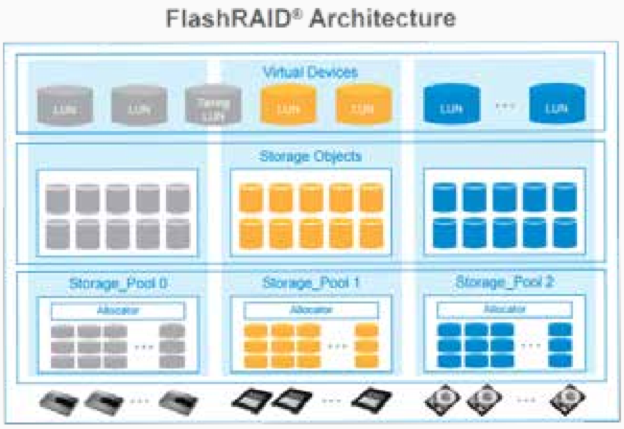
Storage spaces also supports thin provisioning and the use of NVMe or other SSD drives as both a write back cache and a fast tier for a pool of traditional hard disk drives in a hybrid configuration.

As with RSTe, it's not actually possible to boot from a storage spaces volume so a separate boot device is essential.

MEMBLAZE FLASH RAID

Supporting all vendors NVMe drives, but available at a discount for use with Memblaze devices, FlashRAID offers more advanced RAID functionality than those previously mentioned in this article.

Standard RAID levels, and nested RAID levels are possible, including 0,1,5,6,10,01,50 and 51, plus it also has support for Linux and Windows (in coming releases). Automated tiering with different layers of storage from NVMe to S-ATA / SAS HDD's is also supported for hybrid configurations that balance the performance of flash with the capacity of traditional hard disk drives.



What's unique about FlashRAID is that it has been designed with NVMe as the primary use case, and is optimised to work with flash memory from ground up. For example, it includes wear levelling optimisation to improve SSD performance, and to avoid multiple SSD's failing simultaneously as a result of wearing at the same rate; something that could easily cause problems with flash pools towards the end of their life span.

Advanced reconstruction features too improve the time to RAID recovery by prioritising and scheduling I/O to deliver consistent application performance but minimise the time data is at risk. Part of what makes it so quick to rebuilt is its advanced RAID 2.0 engine, which enables reconstruction performance to scale linearly with the number of drives in a pool.

As with RSTe and storage spaces, FlashRAID also doesn't support bootable volumes just yet, so a separate boot device is necessary to utilise it.

NVMe BASED SERVERS FROM SUPERMICRO

In partnership with Supermicro, We're proud to have an extensive range of NVMe based platforms which can suit a variety of different applications.

In a 1U enclosure, we offer both hybrid 3.5" SAS/SATA & NVMe bays, or a full complement of up to 10 x 2.5" U.2 bays in our dual Xeon processor based servers.

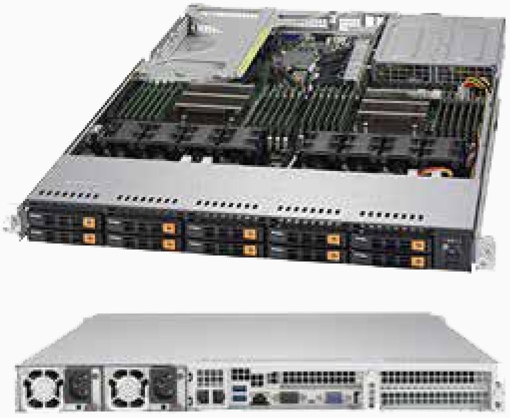


Figure 11: Supermicro's 1028U-TN10RT+ with 10 x NVMe bays in a 1U chassis

SPECIFICATIONS

- *Virtualisation Hosting, Cloud Computing, Data Center*
- *Dual Socket R3 (LGA 2011) support: Intel® Xeon® processor E5-2600 v4/v3 family; QPI up to 9.6GT/s*
- *10x 2.5" Hot-swap bays; 6 NVMe + 4 NVMe/SAS3 Hybrid ports*
- *24 DIMM slots; up to 3TB ECC 3DS LRDIMM or RDIMM DDR4-2400MHz*
- *3 PCI-E 3.0 x8 slots (2 FH 10.5" L, 1 LP)*
- *2x 10GBase-T, 1 Video, 1 COM/Serial, 5 USB 3.0*
- *1000W Redundant Titanium Level (96%) Power Supplies*

With increased overall dimensions, 2U servers increase the drive count to up to 24 x 2.5" / U.2 bays with front accessible hot plug trays, again with dual Xeon processor support.

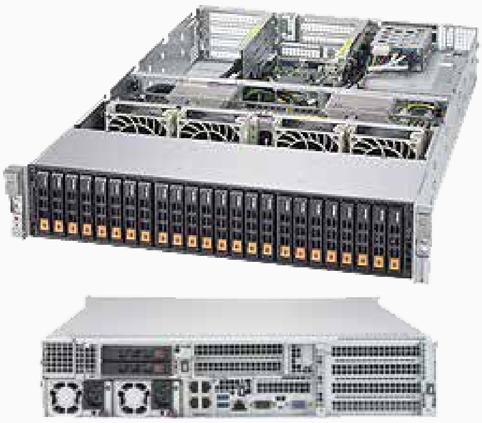


Figure 12: Supermicro's 2028U-TN24R4T+ with 24 x NVMe bays in a 2U chassis

SPECIFICATIONS

- *Virtualisation Hosting, Cloud Computing, Data Center*
- *Dual Socket R3 (LGA 2011) support: Intel® Xeon® processor E5-2600 v4/v3 family; QPI up to 9.6GT/s*
- *24x 2.5" Hot-swap bays; 20 NVMe + 4 NVMe/SAS3 Hybrid ports*
- *24 DIMM slots; up to 3TB ECC 3DS LRDIMM or RDIMM DDR4-2400MHz*
- *3 PCI-E 3.0 x8 slots (2 FH 10.5" L, 1 LP)*
- *2x 10GBase-T, 1 Video, 1 COM/Serial, 5 USB 3.0*
- *System management: Built-in Server management tool*
- *8 heavy duty fans w/ optimal fan speed control*
- *1000W Redundant Titanium Level (96%) Power Supplies*

That's not all however, there's also the simply double range, which includes a second row of 24 x 2.5" U.2 NVMe drive bays behind the first on a swivelling enclosure for hot plug support. This additional row brings the total to an incredibly dense 48 NVMe devices in a 2U enclosure.

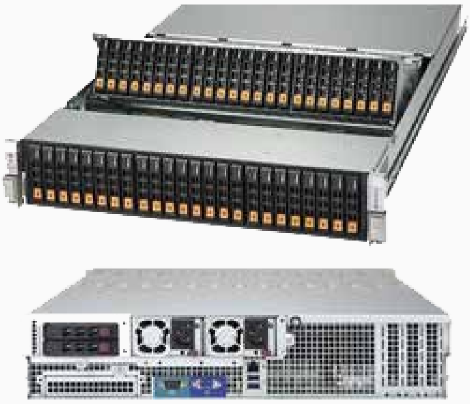


Figure 13: Supermicro's 2028R-NR48N with 48 x NVMe bays in a 2U chassis.

SPECIFICATIONS

- *2U Storage Server - Expansion / Performance / Flexibility*
- *IOPS Intensive Storage Applications, Database Applications (MySQL, Casandra), Virtual Storage Environments etc.*
- *Dual Socket R3 (LGA 2011) support: Intel® Xeon® processor E5-2600 v4/v3 family; QPI up to 9.6GT/s*
- *48x 2.5" Hot-swap NVMe bays; 2x 2.5" Hot-swap drive bays (rear)*
- *Cable-arm for second row drives' hot-swap access*
- *24 DIMM slots; up to 3TB ECC 3DS LRDIMM or RDIMM DDR4-2400MHz*
- *1 PCI-E 3.0 x16 and 1 PCI-E 3.0 x8 slots*
- *1600W Redundant Titanium Level (96%) Power Supplies*

Further still, there is the "Storage Bridge Bay" / SBB, dual controller storage server from Supermicro in the 2U form factor. This not only houses 40 x 2.5" U.2 NVMe devices in a 2U, but also has 2, hot pluggable dual x86 server nodes connected to the backplane via dual ports. It's perfect for high availability storage applications, such as Windows Storage Spaces or ZFS based disk pools.

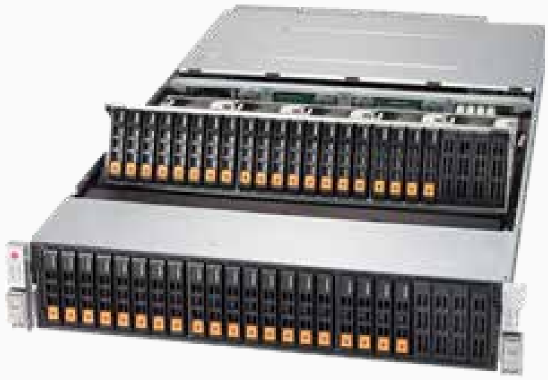


Figure 14: Supermicro's SSG-2028R-DN2R40L with 40 x NVMe and dual x86 systems in a 2U

SPECIFICATIONS

- *2U Mission critical storage server*
- *Lowest latency:*
  - *- 40x Dual port NVMe*
  - *- Omni-Path SIOM support*
- *Robust Data Availability:*
  - *- Shared storage between redundant controllers*
  - *- Node to node heartbeat/10G connectivity*
  - *- Redundant power supply and cooling system*
- *Greener Storage:*
  - *- Titanium Level (96%+) efficiency 2000W redundant power supplies*
  - *- 90%+ VRM power efficiency*

Not only are there NVMe dedicated systems like those highlighted, there are also a number of hybrid systems available which mix standard 3.5", or 2.5" SAS / SATA drive bays. One such example is the Big Twin™, a unique multi node server which includes 2 or 4 dual processor systems in one modular 2U enclosure.

Supporting 6 x SAS 3, or 4 x NVMe devices per node, the Big Twin™ is extremely flexible, and offers configurations to support all kinds of workload, from compute to I/O intensive and beyond.

SPECIFICATIONS

- *4 DP nodes in 2U; depth 28.75"*
- *Dual Socket R3 (LGA 2011) support: Intel® Xeon® processor E5-2600 v4/v3 family*
- *24 DIMM slots; up to 3TB ECC 3DS LRDIMM or RDIMM DDR4-2400MHz*
- *2 PCI-E 3.0 x16 Low-profile slots, 1 Networking SIOM card support*
- *6 SAS3 or 4 NVMe + 2 SAS3 Hot-swap 2.5" drive bays per node*
- *SAS3 support via LSI 3008; IT mode*
- *2200W Redundant Titanium Level (96%) High-efficiency Power Supplies*

Boston has a variety of server solutions too featuring NVMe, such as the Fenway 2224-0T, a great hybrid offering supporting up to 24 SAS / SATA drives, with 4 of those supporting NVMe for a mixed pool. Our friends at IT Pro took the time to review it, find out what they had to say about it here - http://bit.ly/2mWx132

SPECIFICATIONS

- *2U Rackmount*
- *2x Intel Xeon E5-2600 v4 processors*
- *Up to 1.5TB LRDIMM support*
- *7x PCI-E 3.0 x8 add in cards supported*
- *24 x SAS / SATA or 20x SAS / SATA + 4 NVMe 2.5" hot swap bays*

Boston also has hyper converged and pure storage solutions based around VMWare VSAN, and Microsoft Windows Server 2016 with S2D (Storage Spaces Direct). These take advantage of NVMe as both a high performance tier and as a cache for high capacity traditional disks.

Our sales team are standing by ready to tell you more about these, and to prepare a customised server solution quote for you, drop us a line for a no obligation consultation.

Whatever your application, there's an NVMe based solution for you and Boston are on hand to support you.

WITH THANKS TO OUR PARTNERS

ABOUT BOSTON

Boston Limited has been providing cutting edge technology since 1992 using Supermicro® building blocks. Our high performance, mission-critical server and storage solutions can be tailored for each specific client, helping you to create your ideal solution. From the initial specification, solution design and even full custom branding – we can help you solve your toughest business challenges simply and effectively.

CONTACT

+44 (0) 1727 876 100
sales@boston.co.uk
www.boston.co.uk

REFERENCES

"SAS3516 Tri-Mode ROC," March, 2016
https://docs.broadcom.com/docs/BC00-0442EN

"Supermicro NVMe Platforms" March, 2017
https://www.supermicro.nl/products/nfo/NVMe.cfm

"About NVMe", March, 2016
http://www.nvmexpress.org/about/

https://en.wikipedia.org/wiki/NVM_Express

BOSTON
Servers | Storage | Solutions