Celestica™

Next Gen E3 Storage Controllers

Evaluating system design and performance considerations

Authors:
Scott Markinson, Director, System Architect
Rob Harvey, Director, Sales Engineering
Farid Yavari, Director, Business Development

Table of Contents

Introduction 3

Comparing U.2 and E3.S Drives 4

System Balance: U.2 Drives 4

System Balance: E3.S Drives 5

System Design: Drive Count and Performance Evaluation 6

System Design Considerations 8

Additional Considerations: Power Consumption 9

Looking Ahead: PCIe Gen 6 9

Summary 10

Figures

Figure 1: E3 SSD Modules 3

Figure 2: Block Diagram of 1 of 2 Nodes in a 24x U.2 HA Controller System 4

Figure 3: Block Diagram of 1 of 2 Nodes in a 32x E3 HA Controller System 5

Tables

Table 1: SFF-TA-1008 E3 Media Device Form Factor Specification 3

Table 2: Kioxia CM7 SSD Performance Data 6

Table 3: System Performance Calculations of U.2 and E3 Systems 7

Introduction

The Enterprise/Data Center Small Form-Factor (EDSFF) drive from-factor is an emerging form-factor compared to the more classic U.2 (also known as 2.5") SSD form-factor. There are several versions of EDSFF drives including E1.S, E1.L, E3.S-1T, E3.L-1T, E3.S-2T, E3.L-2T as outlined in SNIA specifications:

- SFF-TA-1006 Enterprise and Data Center 1U Short SSD Form Factor
- SFF-TA-1007 Enterprise and Data Center 1U Long SSD Form Factor
- SFF-TA-1008 E3 Media Device Form Factor Specification (ie. E3 specification)

There are four form factors defined for E3. These form factors are shown in Figure 1, with the respective device implementations (starting from right to left) listed in Table 1.

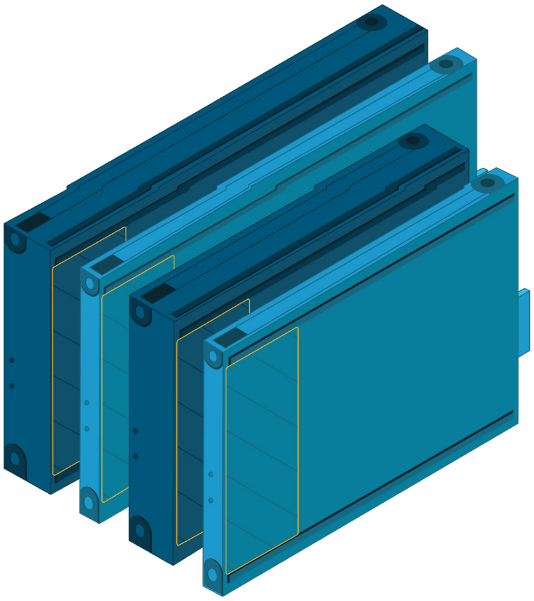


Figure 1: E3 SSD Modules

Module	Height	Length	Thickness
E3 short single thickness device (E3.S, also referenced as E3.S-1T)	76 mm	112.75 mm	7.5 mm
E3 short double thickness device (E3.S-2T)	76 mm	112.75 mm	16.8 mm
E3 long single thickness device (E3.L, also referenced as E3.L-1T	76 mm	142.2 mm	7.5 mm
E3 long double thickness device (E3.L-2T)	76 mm	142.2 mm	16.8 mm

Table 1: SFF-TA-1008 E3 Media Device Form Factor Specification

This white paper will focus on design considerations for systems using EDSFF form factor drives, specifically the most popular and available size E3 SSD module, "E3.S".

Comparing U.2 and E3.S Drives

Both U.2 and E3 modules are available with the latest technology NAND chips and the PCIe Gen 5 bus interface; however, the thinner E3.S SSD allows for more SSDs in the same physical space.

- U.2 thickness = 15 mm
- E3.S thickness = 7.5 mm

However, an E3.S drive has less internal volume for NAND chips in comparison to a U.2 drive. U.2 SSDs are available with ~30TB of storage, but due to its smaller real estate with a single PCB, E3.S drives using the same NAND chips only allow ~15TB of storage.

The next generation of NAND shows a doubling of both of these numbers, but an E3.S drive will continue to contain about one-half the capacity of a U.2 drive.

System Balance: U.2 Drives

For high availability (HA) storage systems, in addition to selecting the right system architecture for the intended application, it is also important to balance performance to minimize bottlenecks. An HA system will have two PCIe lanes (x2) from each node to each Dual-Port SSD. As the read performance of an SSD is close to the PCIe-Gen 5 bandwidth (BW), this paper will simplify the discussion of the SSD BW by reviewing PCIe lane connections and their BW.

Figure 2 shows a block diagram of a 24-drive U.2 system. There are 48 PCIe Gen 5 lanes of bandwidth available between the CPU and the drives, aggregating to ~32Gb/s * 48 lanes = 1536 Gb/s (~4GB/s * 48 = 192GB/s). When looking at the external IO connections on the right side of the CPU, if 400Gb/s NIC cards (such as the Nvidia CX7), are used in the 2 x16 slots for uplink to network servers, there would be 800Gb/s (100GB/s) of bandwidth in/out of each controller canister.

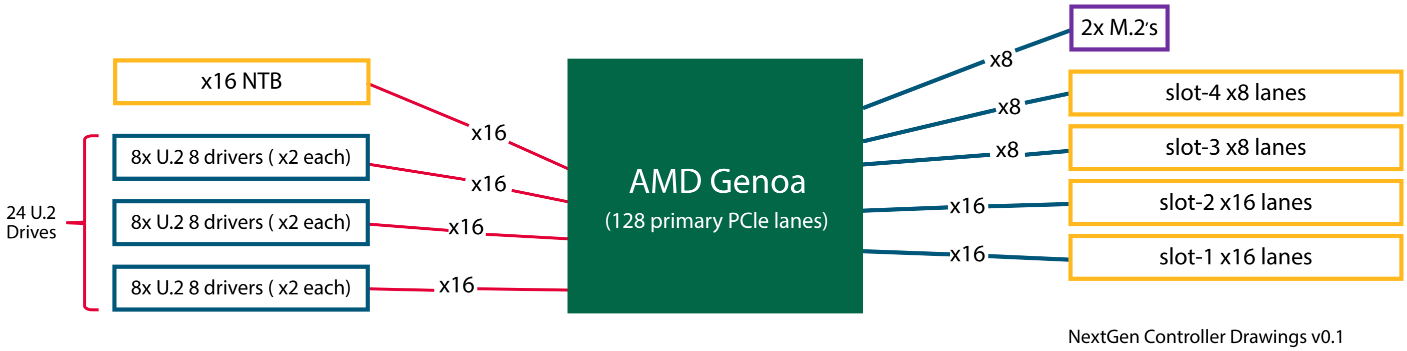
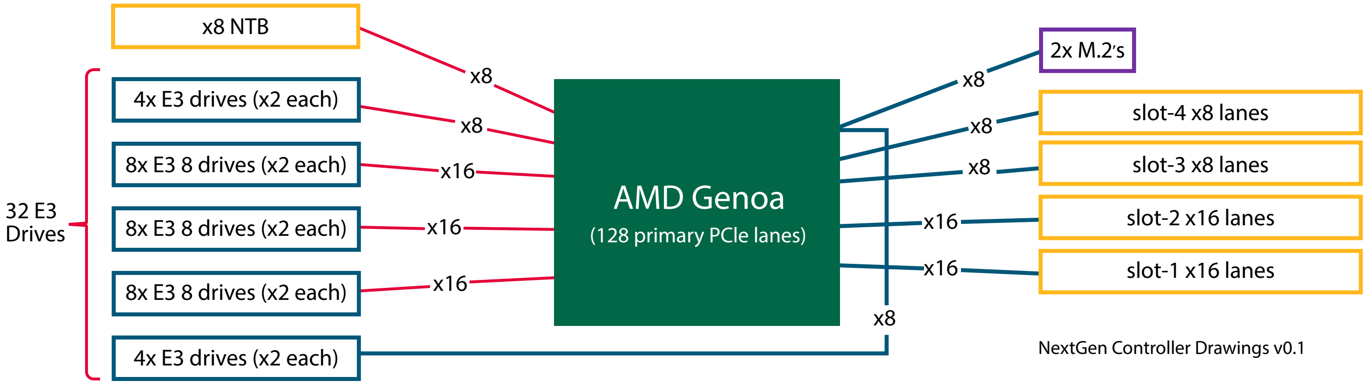


Figure 2: Block diagram of 1 of 2 nodes in a 24x U.2 HA controller system

System Balance: E3.S Drives

Figure 3 shows a system with SSDs direct to the CPU. On the left side of the CPU there are 64 PCIe-Gen 5 lanes of bandwidth available between the CPU and the drives, ~32Gb/s * 64 lanes = 2048 Gb/s (~4GB/s * 64 = 256GB/s).

As with the U.2 example shown earlier in Figure 2, there is 800Gb/s (100GB/s) of bandwidth in/out of each controller canister.



NextGen Controller Drawings v0.1

Figure 3: Block Diagram of 1 of 2 nodes in a 32x E3 HA Controller system

To fully utilize the SSDs' performance, the Genoa CPU would need to perform operations on local data, and not just service incoming read/write requests, as the aggregate bandwidth to the drives is now even faster than the uplink Bandwidth to the network connections.

Some of this extra bandwidth could be leveraged based on the managing software in use. Depending on how the data files are manipulated, write-amplification occurs using this extra local bandwidth. An example would be if software writes the raw data as it arrives, then later does background compression on the files, and needs to write this now smaller file with the same data.

System Design: Drive Count and Performance Evaluation

Most front-load storage controller systems have space for 24 U.2 SSDs, but the thinner E3.S SSDs allow for 32 (and possibly more) SSDs across the front of the controller system.

Supporting large numbers of SSDs on the PCIe bus often requires the use of PCIe switches to expand the limited PCIe lanes to ensure the CPU has enough lanes to support the drives. These switches limit the drive performance down to the bus speed & lane-width connection between the CPU(s) and the PCIe switches.

For example, HA systems such as the Celestica SC6100 (U.2) & SC6110 (E3) dual redundant node HA storage controller systems use dual-port SSDs to allow both nodes access to all drives. Both nodes can operate as active-active on the same data at the same time, or independently on different data sets as needed by system software.

Both the SC6100 & SC6110 use the AMD Genoa CPU, single-socket per node, which allows the SSDs a direct connection to the PCI lanes of the CPU without the use of a switch device, maximizing overall bandwidth.

Table 2 provides a view of overall bandwidth, from Kioxia, one of the major SSD industry providers of PCIe Gen 5 SSDs in both U.2 and E3 formats.

Kioxia CM7 SSDs Performance information:

SeqR U.2	14,000 MB/s (128KB data size)
SeqR E3	13,000 MB/s (128KB data size)
Seq W U.2	7,000 MB/s (128KB data size)
Seq W E3	6,300 MB/s (128KB data size)
Rand R U.2	2,400 KIOPs @4 KB=> 9,600 MB/s
Rand R E3	2,000 KIOPs @4 KB => 8,000 MB/s
Rand W U.2	550 KIOPs @4 KB => 2,200 MB/s
Rand W E3	470 KIOPs @4 KB =>1,880 MB/s

Table 2: Kioxia CM7 SSD performance data

Source: https://www.kioxia.com/en-jp/business/ssd/enterprise-ssd.html)

Table 3 shows the performance of both a 24 SSD U.2 system (Celestica SC6100) and a 32 Drive E3 system (Celestica SC6110), in which all SSDs:

- Connect without a PCIe switch to the CPU
- Contain two Nvidia CX7 (or similar 400GbE NIC cards), also without a switch to each of the two compute nodes.

Note: actual measured system data will be lower than shown to account for various real-world communications on each of the buses.

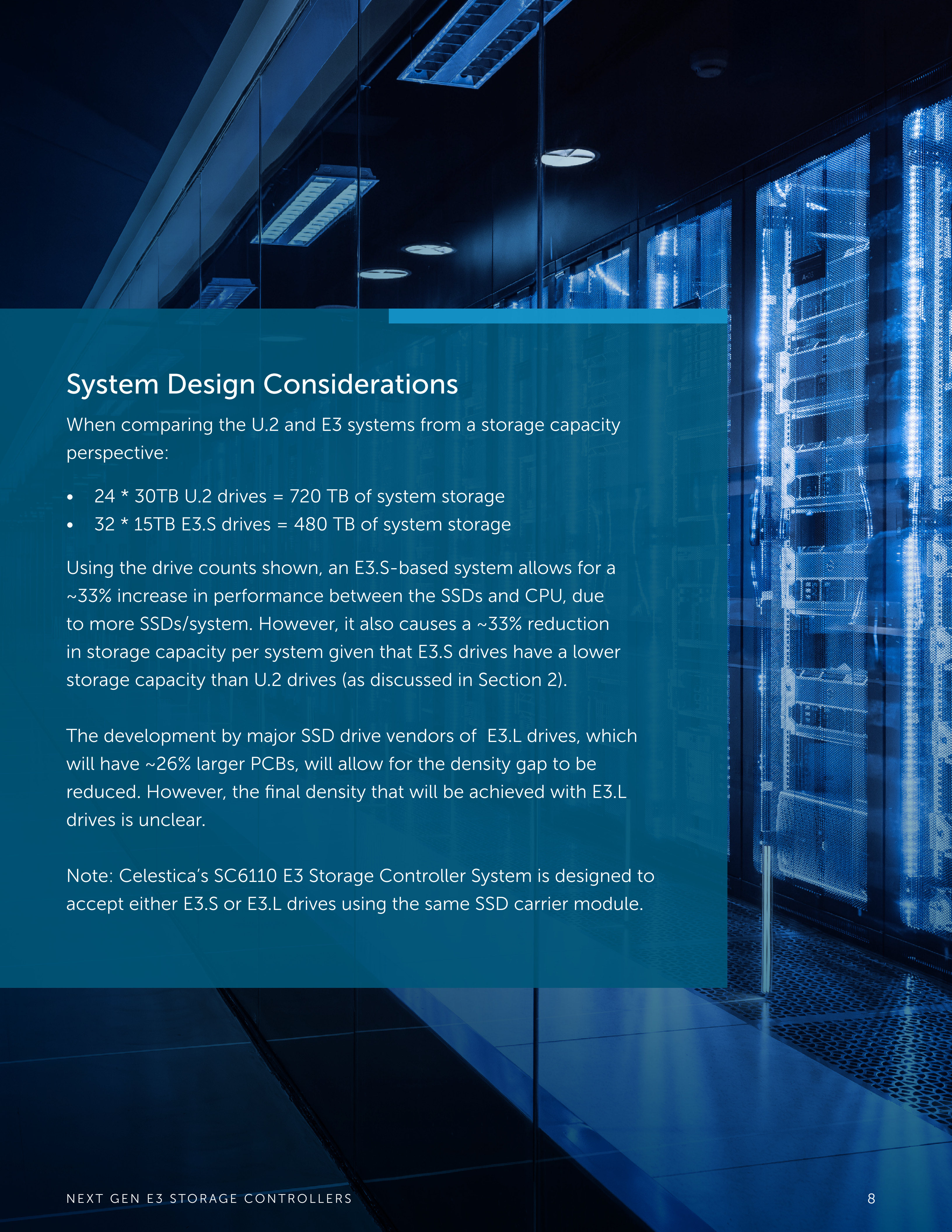
Operation	Count	Performance	%SSD	NIC Cards (4x 400Gb/s NIC) for System	
SeqR U.2	24	336,000 MB/s => 2,688 Gb/s	168%	1,600 Gb/s	UPLINK is LIMITING SYS BW
SeqR E3	32	416,000 MB/s => 3,328 Gb/s	208%	1,600 Gb/s	UPLINK is LIMITING SYS BW
Seq W U.2	24	168,000 MB/s => 1,344 Gb/s	84%	1,600 Gb/s	
Seq W E3	32	201,600 MB/s => 1,613 Gb/s	101%	1,600 Gb/s	UPLINK is LIMITING SYS BW
Rand R U.2	24	230,400 MB/s => 1,843 Gb/s	115%	1,600 Gb/s	UPLINK is LIMITING SYS BW
Rand R E3	32	256,000 MB/s => 2,048 Gb/s	128%	1,600 Gb/s	UPLINK is LIMITING SYS BW
Rand W U.2	24	52,800 MB/s => 422 Gb/s	26%	1,600 Gb/s	
Rand W E3	32	60,160 MB/s => 481 Gb/s	30%	1,600 Gb/s	

Table 3: System Performance Calculations of U.2 and E3 Systems

** Total System Performance assumes both nodes are fully accessing data from/to the drives

This data illustrates that if these systems are used solely for data storage and retrieval, the networking cards become the system's performance limiter in most operations, regardless of the drive counts. Systems using E3 drives would best utilize their aggregate SSD bandwidth if they also perform local data operations within the same system.

Table 3 also illustrates that for systems where less data is accessed by the CPU for local computing, a 24-drive U.2 system can also provide full bandwidth to a pair of 400Gb/s NIC cards. As such, U. 2-based systems may be more appropriate when "data capacity (TB/system)" per system is important.



System Design Considerations

When comparing the U.2 and E3 systems from a storage capacity perspective:

- 24 * 30TB U.2 drives = 720 TB of system storage
- 32 * 15TB E3.S drives = 480 TB of system storage

Using the drive counts shown, an E3.S-based system allows for a ~33% increase in performance between the SSDs and CPU, due to more SSDs/system. However, it also causes a ~33% reduction in storage capacity per system given that E3.S drives have a lower storage capacity than U.2 drives (as discussed in Section 2).

The development by major SSD drive vendors of E3.L drives, which will have ~26% larger PCBs, will allow for the density gap to be reduced. However, the final density that will be achieved with E3.L drives is unclear.

Note: Celestica's SC6110 E3 Storage Controller System is designed to accept either E3.S or E3.L drives using the same SSD carrier module.

Additional Considerations: Power Consumption

It is also important to note that the system's power consumption will increase with the move from 24 U.2 drives to systems with 32 E3.S drives. Since both U.2 and E3.S drives have the same maximum power specification of 25W, a system based on 32 E3.S drives will consume ~33% more power than if the system were using 24 U.2 drives.

- U.2: 24 * 25W = 600W
- E3.S: 32 * 25W = 800W

This is set to increase with E3.L and its 40W power specification; however, preliminary data from vendors is indicative of expected power consumption at 30W to 35W per drive:

- E3.L: 32 * 40W = 1,280W
- E3.L: 32 * 35W = 1,120W
- E3.L: 32 * 30W = 960W

Looking Ahead: PCIe Gen 6

The discussion of U.2 and E3 is expected to become less important as systems move to PCIe Gen 6 in the near future, as the U.2 form-factor SSDs device is not expected to be available at PCIe Gen 6 speed, and the industry will fully move to E3 form-factors for signal-integrity reasons.

Summary

EDSFF E3-based systems offer improved performance due to higher SSD module density, making them ideal for applications requiring fast data throughput and high drive counts. However, these benefits come with trade-offs in storage capacity and higher power consumption. E3.S SSDs typically have 50% of the storage capacity of the same generation U.2 version. As a result, even though there are more SSDs in an E3 system, there is still ~33% less storage capacity compared to U.2 systems. This may lead to additional costs if more storage is required.

The increased number of drives in E3.S systems results in higher overall power consumption, leading to higher energy costs over the system's lifespan. When considering Total Cost of Ownership, factors such as power usage, drive density, and the cost of additional infrastructure (such as PCIe switches or cooling systems) must be accounted for.

While the transition to PCIe Gen 6 is expected to make E3 form-factors the industry standard, the need for frequent upgrades to keep pace with technology might make U.2 systems a more cost-effective choice in the short term. Therefore, organizations need to carefully balance performance needs, storage requirements, and long-term operational costs to make the most cost-effective decision.

Celestica™

North America: (Toll Free) +1 888 899 9998

Global: +1 416 448 5800 | Europe: +49 21 6257 88031

contactus@celestica.com

celestica.com

in Celestica | f @CelesticaInc

080425