NVMe SSD Adoption Consideration in China Data Center

Taile Zhang

VP, Memblaze

The SSD Adoption Survey

- 15 top Data center Customers/OEMs be asked

- 12 key factors we think are important to SSD adoption

- Each customer pick top 3 they cared the most
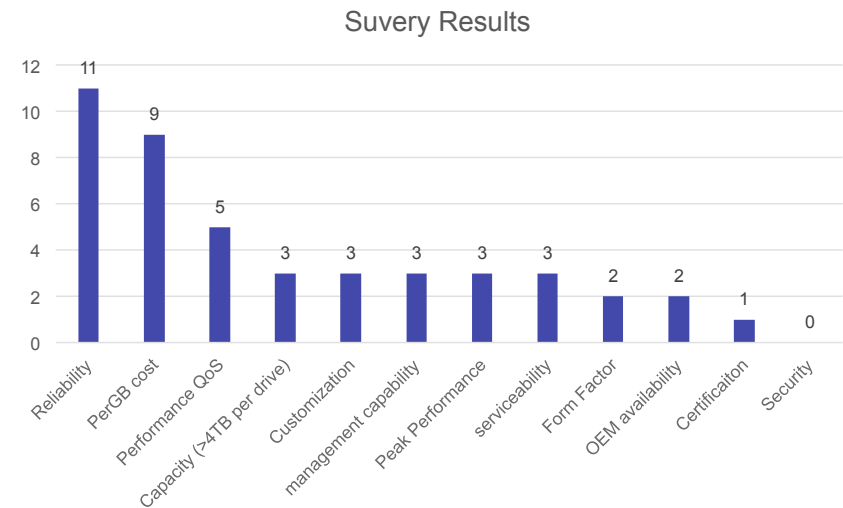
The 12 Options

items	Questions
Capacity	Need >4TB per Drive? Is the density important for you ?
Certification	Do you care PCIe-SIG, vSAN, IOL NVMe?
Customization	User specific commands and features?
Form Factor	Are you flexible for form factor choices, like U.2, AIC?
Management	Management tools, UI, OOB management , monitoring?
OEM availability	Do you need OEM certification before test?
Peak Performance	How many IOPS and bandwidth do you need?
Performance QoS	Are you care guaranteed the performance bottom line? QoS for each application on mixed workload?
Per GB Price	More lower more better
Reliability	No Hang, No crash, and No data loss
Security	Encryption, Trust computing.
Serviceability	On site support, RMA friendly , Response speed.

The Survey Results and learns

- No Surprise, Reliability is most important thing
- Cost is one of the most important factors in china market
- Data Center Customers are more care about Performance QoS rather than Peak Performance
- Management, customization , serviceability is much more important for Data center application
- Security is still less be considered on disk level. System level solutions are more preferred

Suvery Results

Category	Value
Reliability	11
PerGB cost	9
Performance QoS	5
Capacity (>4TB per drive)	3
Customization	3
management capability	3
Peak Performance	3
serviceability	3
Form Factor	2
OEM availability	2
Certificaiton	1
Security	0

3 Top challenges

- Reliability
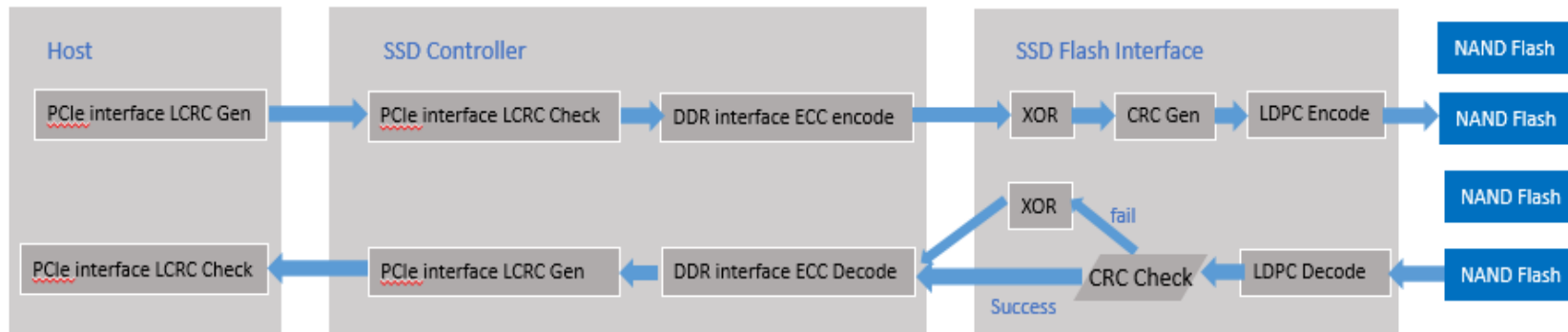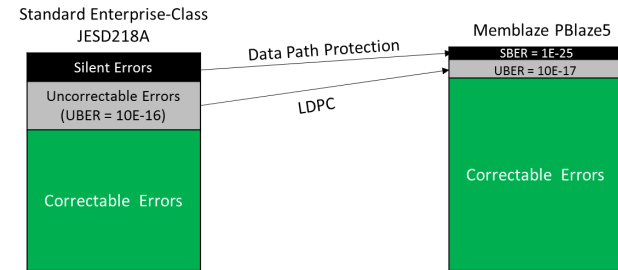- Performance QoS
- Cost Saving by 3D NAND

Full Data Path Protection

Full Data Path Protection high reliability design

- Data Path Protection include data transfer and storage
- Host to controller PCIe interface with LCRC Gen/Check
- SSD controller DDR with ECC Encode/Decode
- SSD Flash interface with XOR
- SSD Flash interface with CRC Gen/Check
- SSD Flash interface with LDPC Encode/Decode

Standard Enterprise-Class JESD218A

| Silent Errors |
| Uncorrectable Errors (UBER = 10E-16) |
| Correctable Errors |

Data Path Protection

LDPC

Memblaze PBlaze5

| SBER = 1E-25 |
| UBER = 10E-17 |
| Correctable Errors |

Host
- PCIe interface LCRC Gen
- PCIe interface LCRC Check

SSD Controller
- PCIe interface LCRC Check
- DDR interface ECC encode
- PCIe interface LCRC Gen
- DDR interface ECC Decode

SSD Flash Interface
- XOR
- CRC Gen
- LDPC Encode
- XOR
- fail
- CRC Check
- LDPC Decode
- Success

NAND Flash
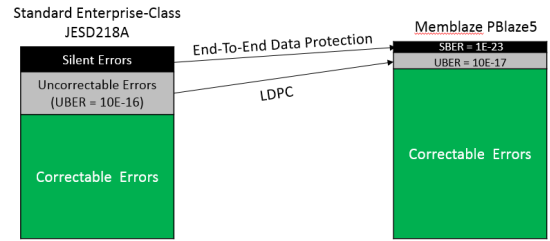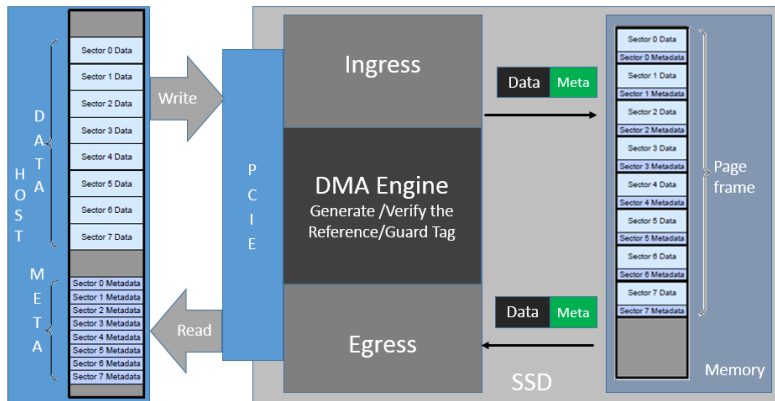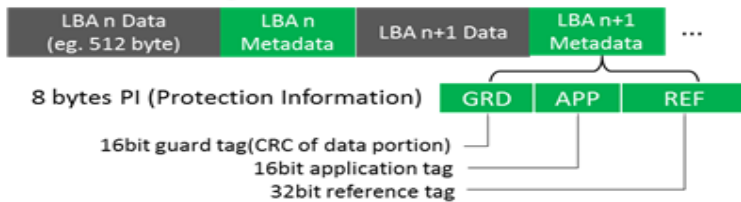NAND Flash
NAND Flash
NAND Flash

Reduce Silent Data Error

Silent Data Error

Bit flips elsewhere, such as controller, DRAM, even caused by cosmic rays created by exploding stars(Supernovas)

Standard Enterprise-Class JESD218A

| Silent Errors |
| Uncorrectable Errors (UBER = 10E-16) |
| Correctable Errors |

End-To-End Data Protection

LDPC

Memblaze PBlaze5

| SBER = 1E-23 |
| UBER = 10E-17 |
| Correctable Errors |

Reduced Silent Bit Error Rate
SBER: 1E-23

Metadata contiguous with LBA

8 bytes PI (Protection Information) | GRD | APP | REF

16bit guard tag(CRC of data portion)
16bit application tag
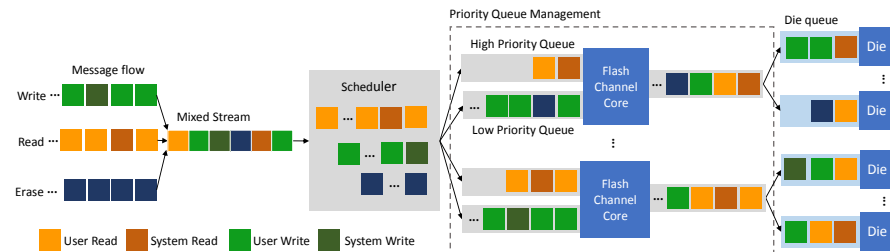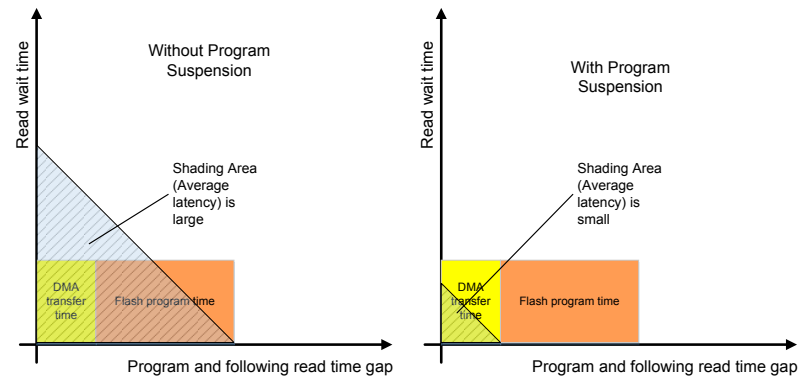32bit reference tag

The core technology in End to End Protection is PI (Protection Information) generation and checking.

- The DMA engine is the key component to perform data Protection Information (PI) generation and checking.

- 8 bytes PI consists of guard tag (a CRC16 computed over the contents of corresponding user data sector), application tag and reference tag.

- Received inbound sector data on writes (eg. 512bytes) will be appended with PI via ingress DMA engine.

- On a read, the checksum in the PI is compared to the value computed over the outbound sector data by Egress DMA engine. If the two do not match, then the data with an incorrect mark and send to host.

- Silent errors can be recovered via re-read, RAID protection or other possible error correction ways at host-end.

- According by the NVMe protocol parameters, the PI info could be generated in the Host or NVMe Device.

MEMBLAZE

7
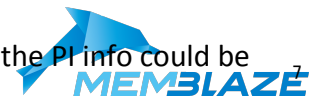
Optimize Read Latency

- NAND flash data low latency access commands.
    - NAND flash can be read in fast mode (Such as SNAP read)
    - Program suspension can lower read latency significantly. This should be done with policy so that suspension does not hurt flash data integrity.

- Priority queue management.
    - With priority queue management, received read requests can be re-ordered to high priority queue.
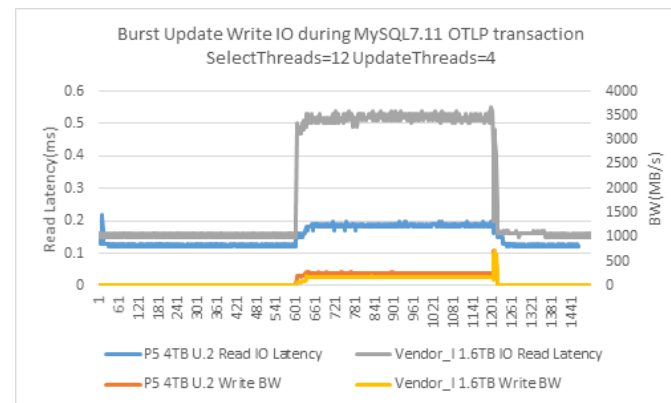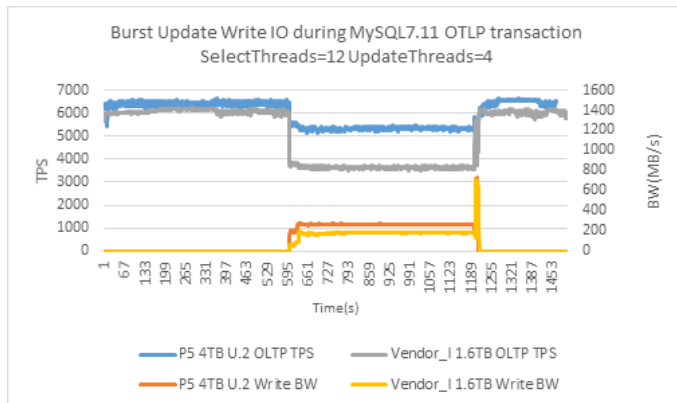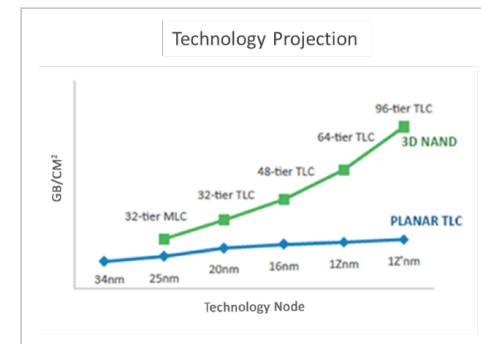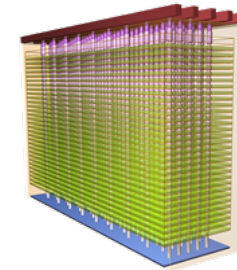
Optimize Read Latency (Cont'd)

- Random read throughput and latency with and without QoS enable.

Burst Update Write IO during MySQL7.11 OTLP transaction
SelectThreads=12 UpdateThreads=4

P5 4TB U.2 OLTP TPS   Vendor_I 1.6TB OLTP TPS
P5 4TB U.2 Write BW   Vendor_I 1.6TB Write BW

Burst Update Write IO during MySQL7.11 OTLP transaction
SelectThreads=12 UpdateThreads=4

P5 4TB U.2 Read IO Latency   Vendor_I 1.6TB IO Read Latency
P5 4TB U.2 Write BW   Vendor_I 1.6TB Write BW

3D TLC NAND enabling

PBlaze Series Evolutions	PBlaze4	PBlaze5
Controller	Princeton	Belmar
NAND Type	2D MLC	3D NAND
Bits per Cell	2 (MLC)	3 (TLC)
NAND Process	15nm	32nm
NAND Layers	1	32
NAND Density(Gb)	128	384
Maximum User Capacity	3.2T	11T
Performance	750K	1M
ECC Capability	BCH (100bits/4KB)	LDPC (550bit/4KB)
Price per GB	$$	$

Technology Projection

GB/CM²

96-tier TLC
64-tier TLC 3D NAND
48-tier TLC
32-tier TLC
32-tier MLC

PLANAR TLC

34nm 25nm 20nm 16nm 12nm 1Z'nm

Technology Node

MEMBLAZE

Summary

- The Survey identify top 3 customers value points
  - Reliability, Cost, & Performance QoS
- PBlaze SSD enhancements examples

Please visit our booth at #523.

Taile Zhang

VP, Memblaze