Impact of High Capacity and QLC SSD

Hyung-Seuk Kim
Samsung Electronics

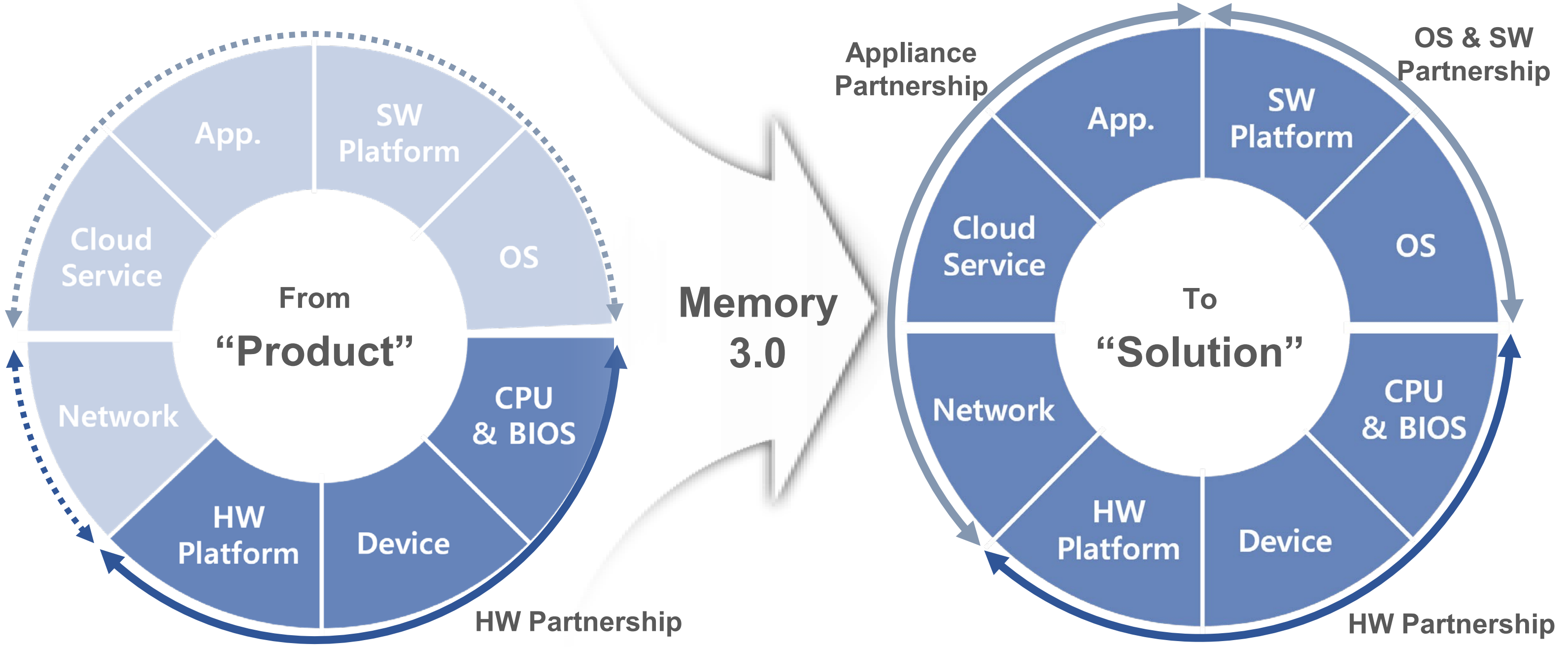
Contents

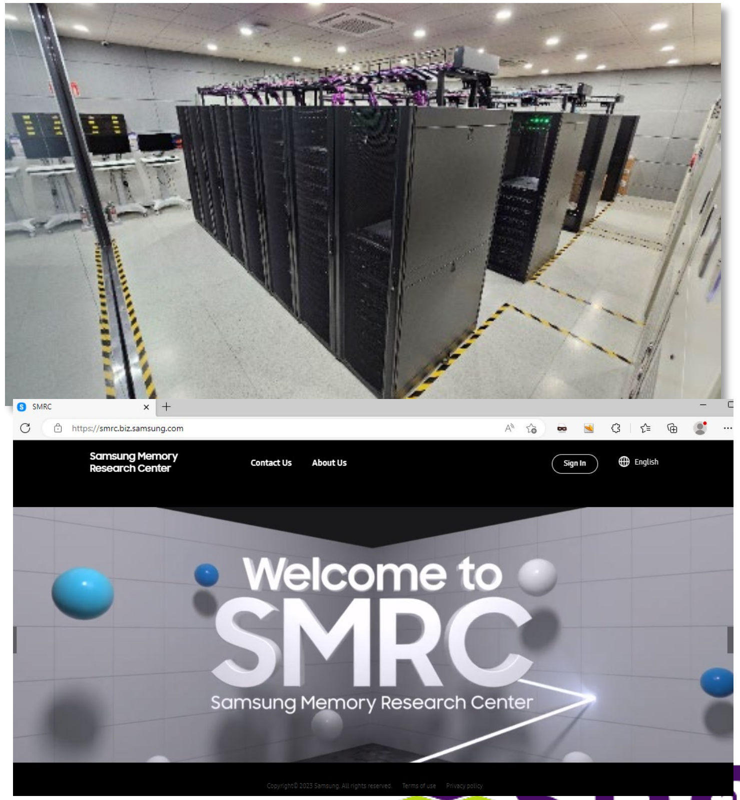
- **Introduction**
- **Storage system and SSD performance**
- **Storage cluster performance test**
- **High capacity QLC SSD limitation and its mitigation**
- **Conclusion**

SMRC - The Division I am part of

Samsung Memory Research Center "SMRC" is an open collaborative space for customers and partners. Our mission is to:

- Accelerate the next evolution through innovative technological collaboration to achieve optimal solutions.
- Contribute to the IT ecosystem with innovative solutions.
- Develop reference architectures for next-generation system solutions.





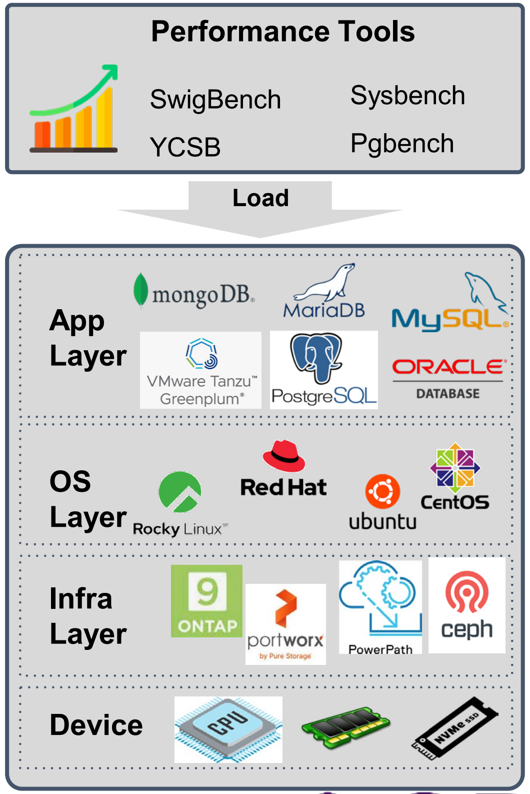
Samsung proprietary

SMRC – Resources and Infra

SMRC offers various hardware configuration settings tailored for different environments

Processors	DRAM	SSD	Accelerator	Network
Intel Icelake	DDR4 - RDIMM	SATA	GPGPU	400GbE
Intel Sapphire Rapids	DDR5 - RDIMM	PCIe Gen4	FPGA	200GbE
AMD Genoa	CXL 1.1	PCIe Gen5	DPU	100GbE
ARM - Ampere	CXL 2.0	QLC NVMe	NPU	10GbE
	HBM3	FDP NVMe	GPU Server	InfiniBand 400GbE
	MRDIMM	Smart SSD		InfiniBand 200GbE
	PNM - AxDIMM			

- Red Hat OpenStack
- Red Hat OpenShift
- Bare Metal
- VMWare vCenter
- VMWare Tanzu



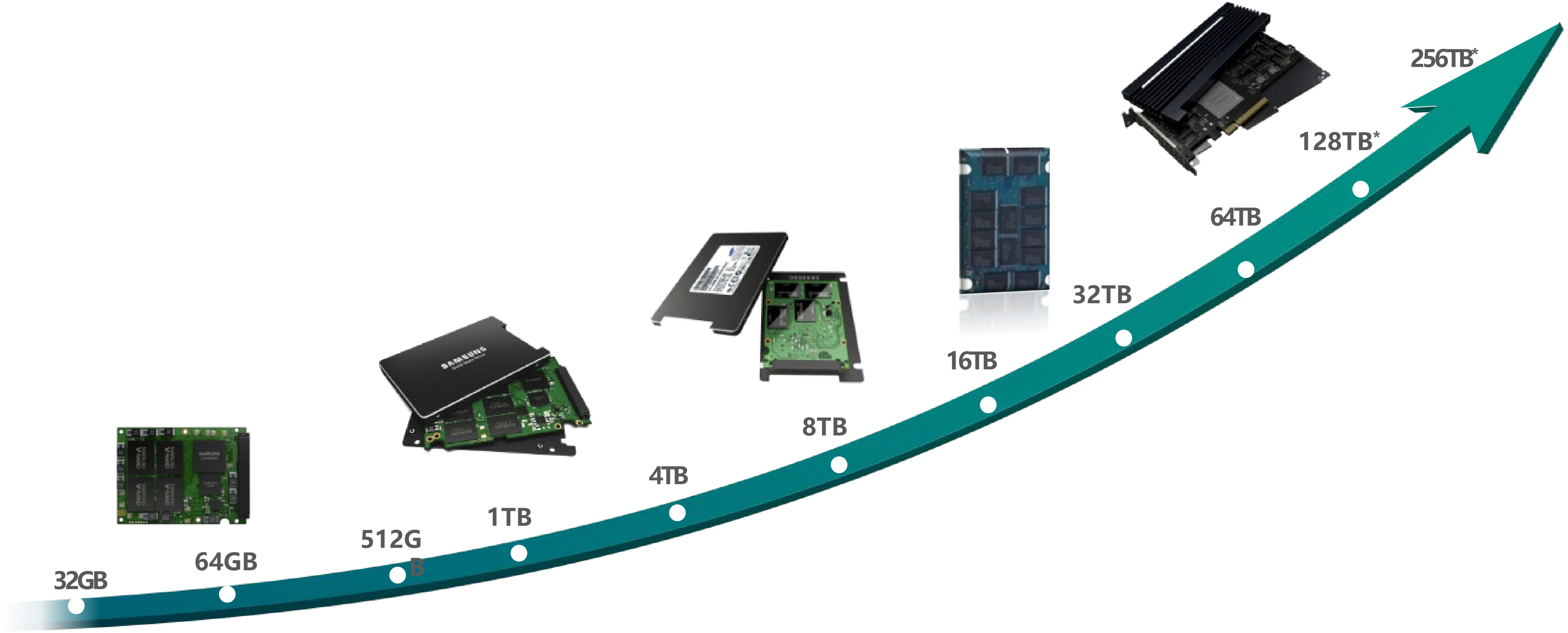
 |

Samsung proprietary

SDC 24

SSD Capacity Trend

- SSD capacity has been growing and will continue

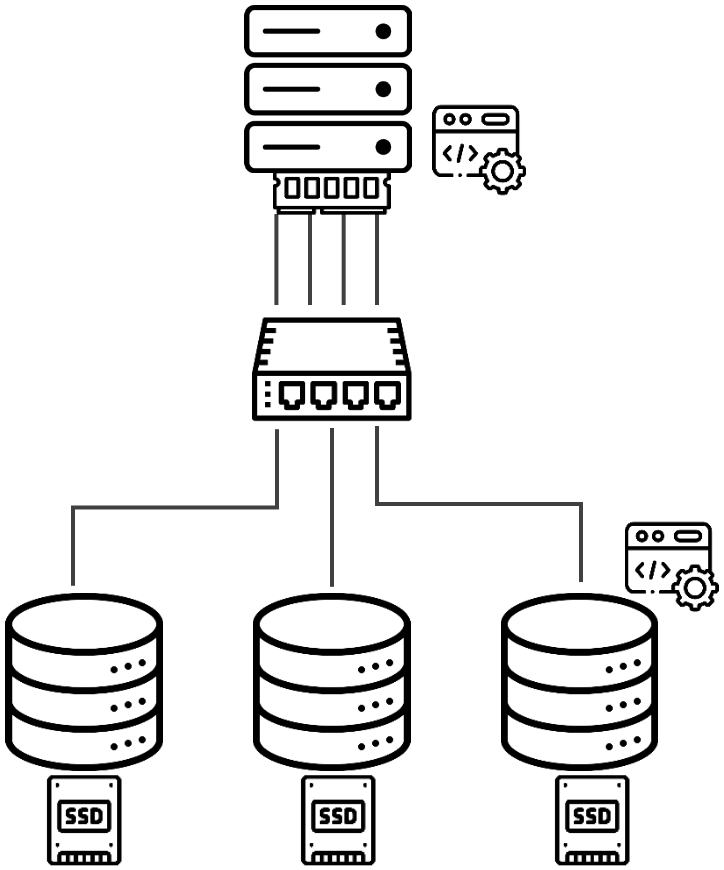


Samsung proprietary

Contents

Storage System Performance

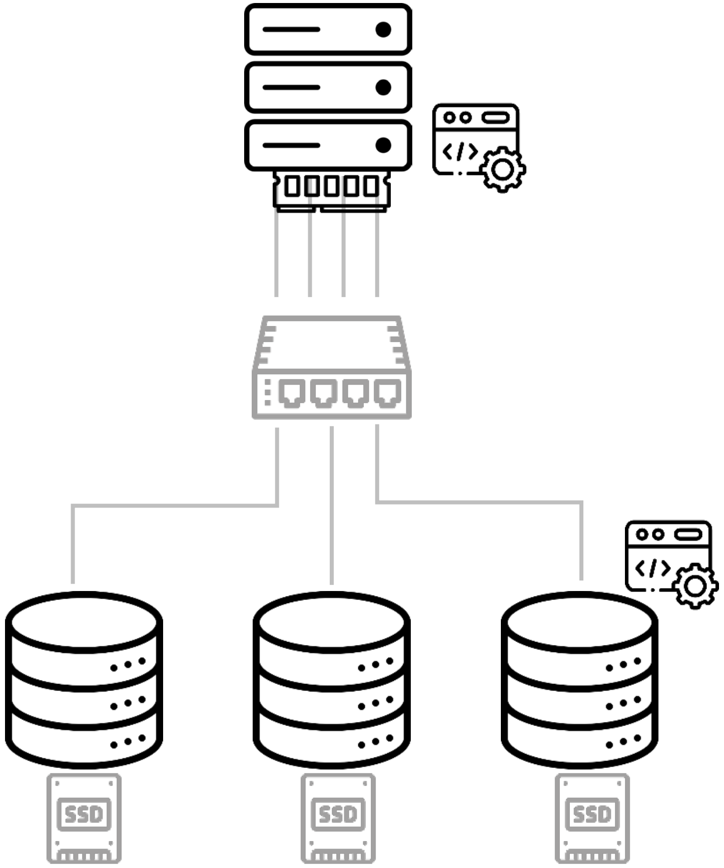
- Influential factors on system performance



- Host Processing Capability
- Buffer Memory
- Network Bandwidth
- Software Overhead & Limitation
- Storage Processing Capability
- Storage Device Performance

Storage System Performance

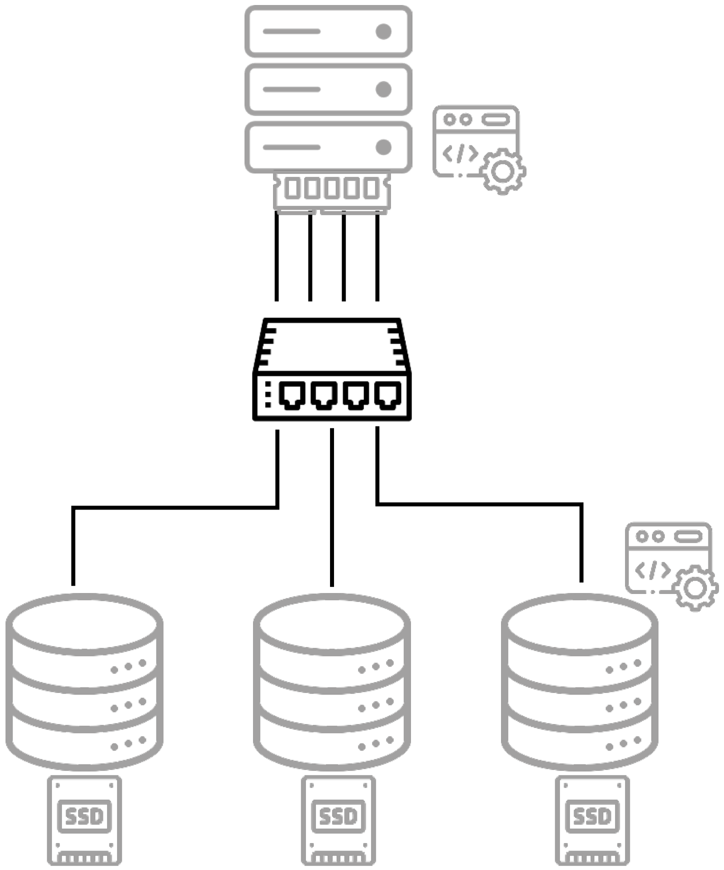
- Processing capability and S/W are defined by system product



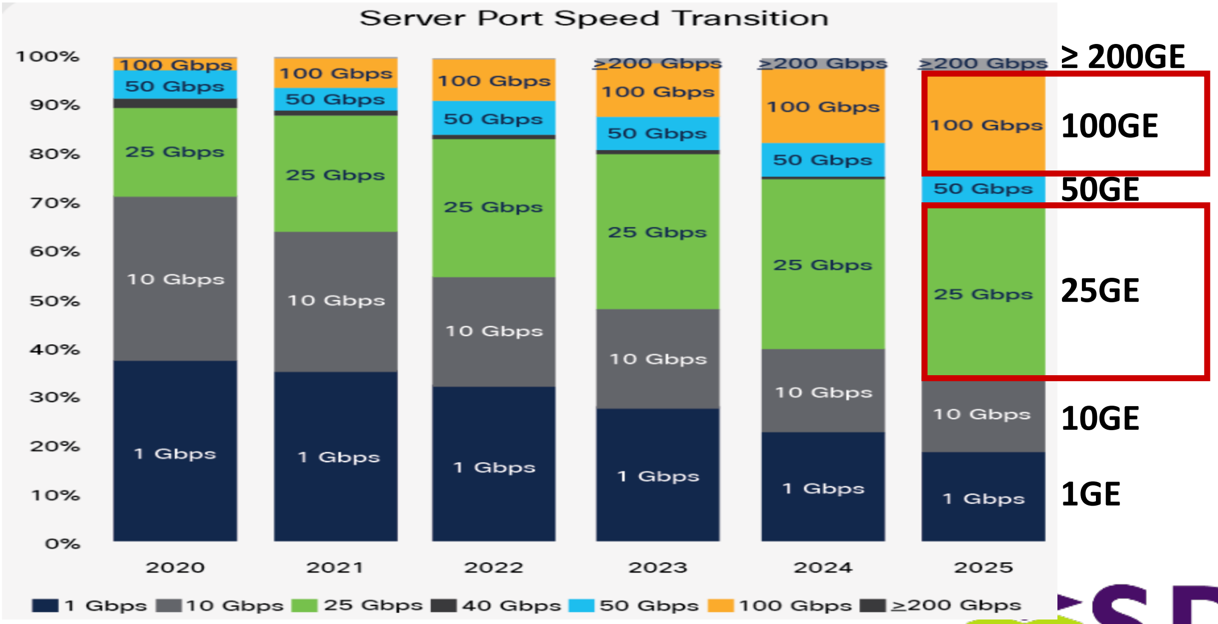
- Host & Storage Processing Capability
  - Determinants are CPU and DRAM
  - They are design choices by the users
  - It is recommended to be sufficient not to be performance bottleneck
- Software
  - Software overhead or limiting the allocation of resources
- Buffer Memory
  - Density and I/O bandwidth may affect load query speed
  - It is configured by storage S/W

Storage System Performance

- 25 Gbps and 100 Gbps are dominant network



- Network Bandwidth
  - One of the bottleneck of storage system
  - 25 and 100GE are expected to be dominant in near future
    · 25 and 100GE are used in this presentation

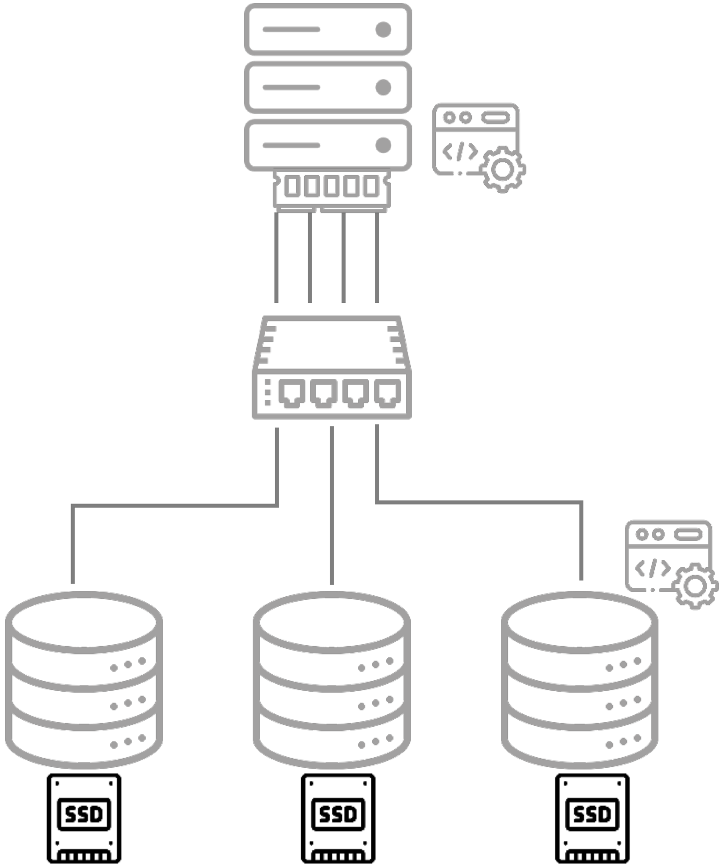


Source: Dell'Oro Group

Samsung proprietary

Storage System Performance

- Storage device performance always affects the system performance



- Storage Device Performance
  - For a slow storage device, SSD performance significantly impacts system performance
  - Even if system does not utilize full SSD bandwidth due to network or S/W bottleneck, SSD latency is a factor especially for short queries

→ TLC vs QLC on later part of this presentation

QLC SSD – Misconception

① Can QLC core speed match TLC as NAND technology evolves?

② Will faster SSD interface (PCIe Gen5/Gen6) improve QLC SSD performance?

③ The impact of advancements in NAND interface (Toggle 5.0, SCA) on QLC SSD

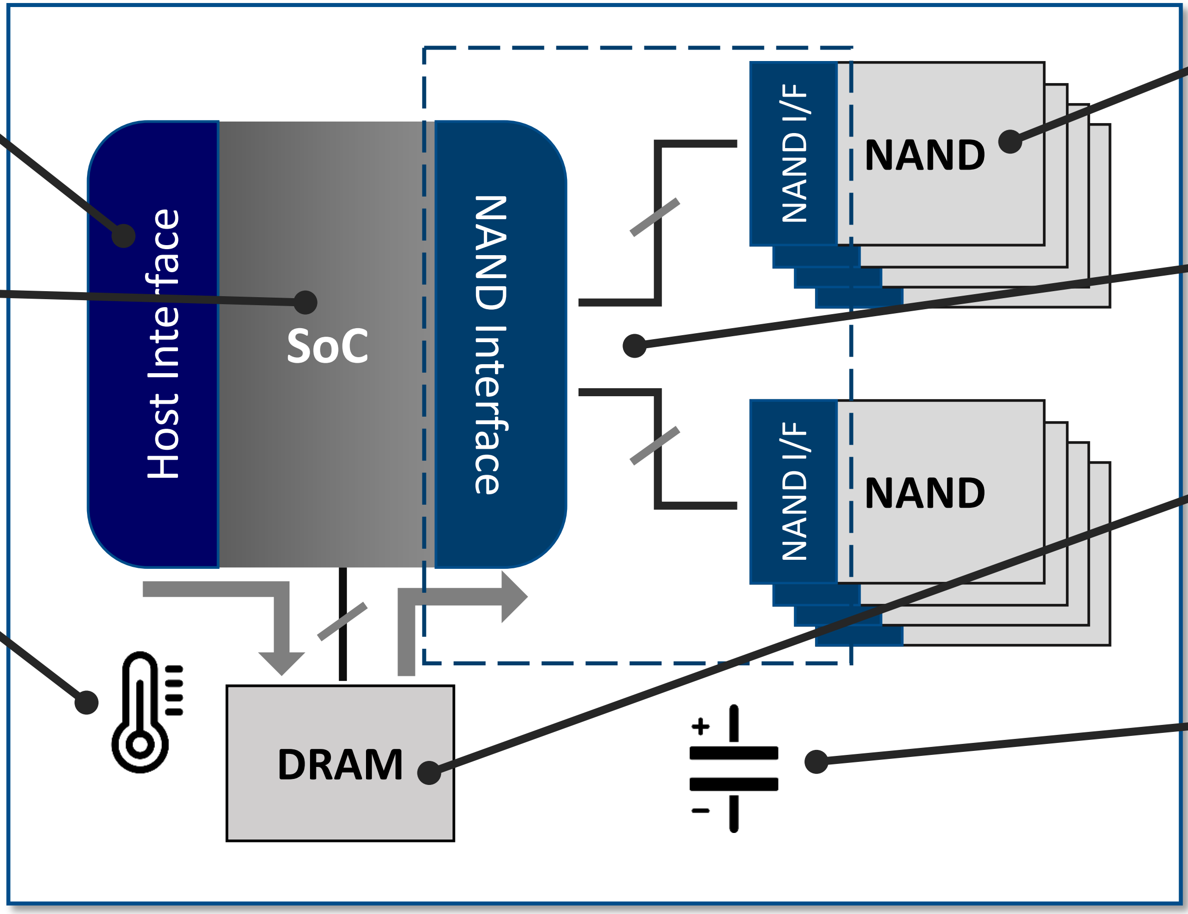
④ Will a powerful SSD SoC and high-end DRAM improve QLC SSD performance?

⑤ SSD performance scales as SSD capacity increases

SSD Performance Factors

- Influential factors on SSD performance

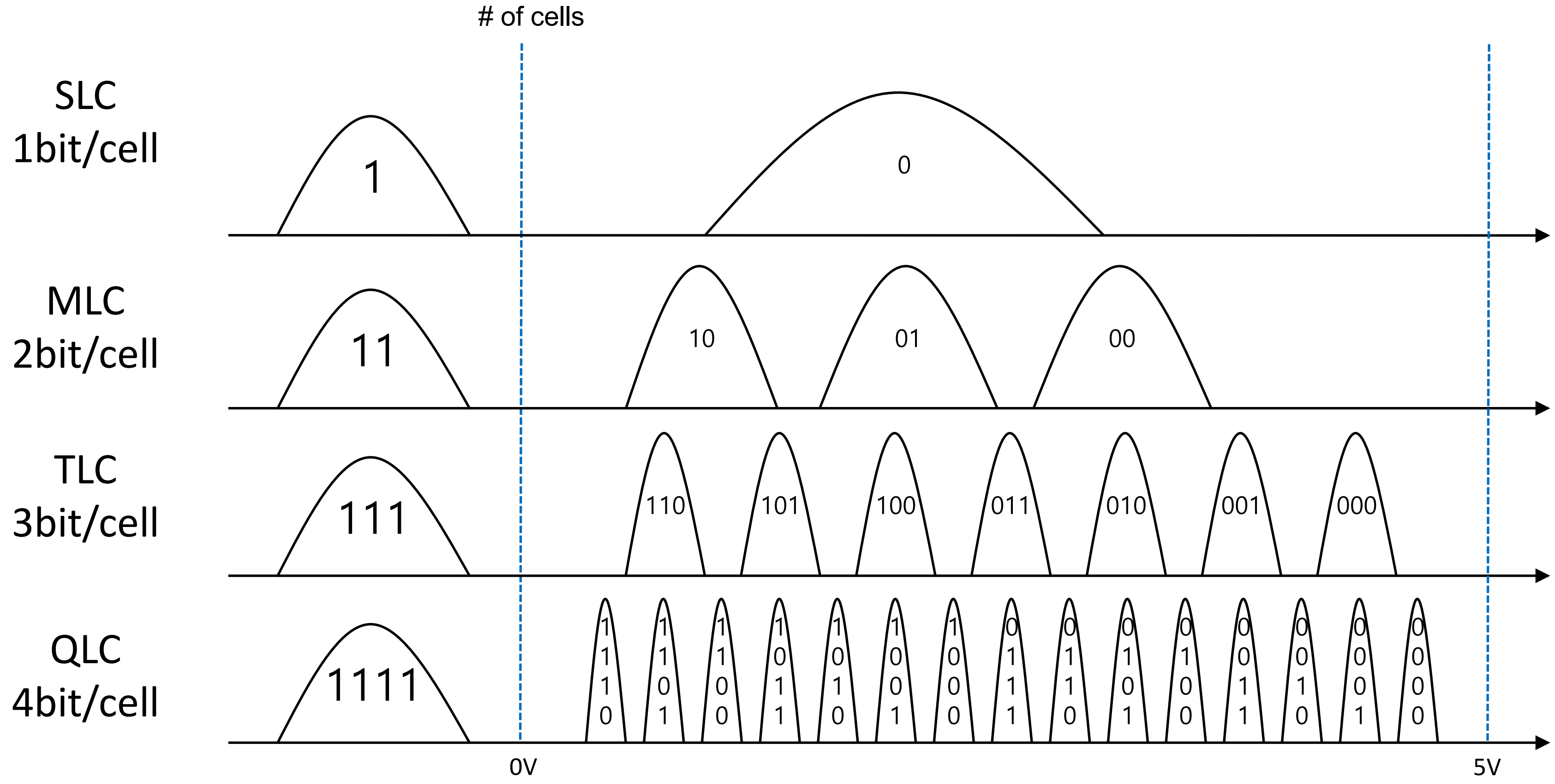
- SATA → PCIe Gen3 → ***Gen4 → Gen5***
- SoC computing power for random access
- Temperature limit for high performance SSD.



- NAND speed tPROG, tR, ways
- NAND I/F: channels Toggle 4.0 → 5.0 non-SCA → SCA
- DRAM bandwidth (***DDR5)*** and buffer size for write speed
- Capacitor for write buffer size

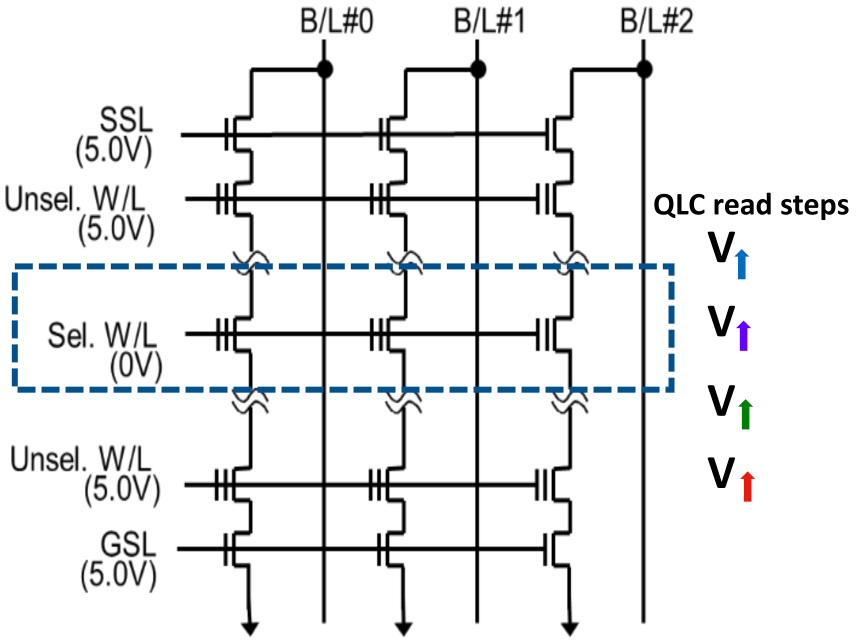
QLC Device – Variation Distribution

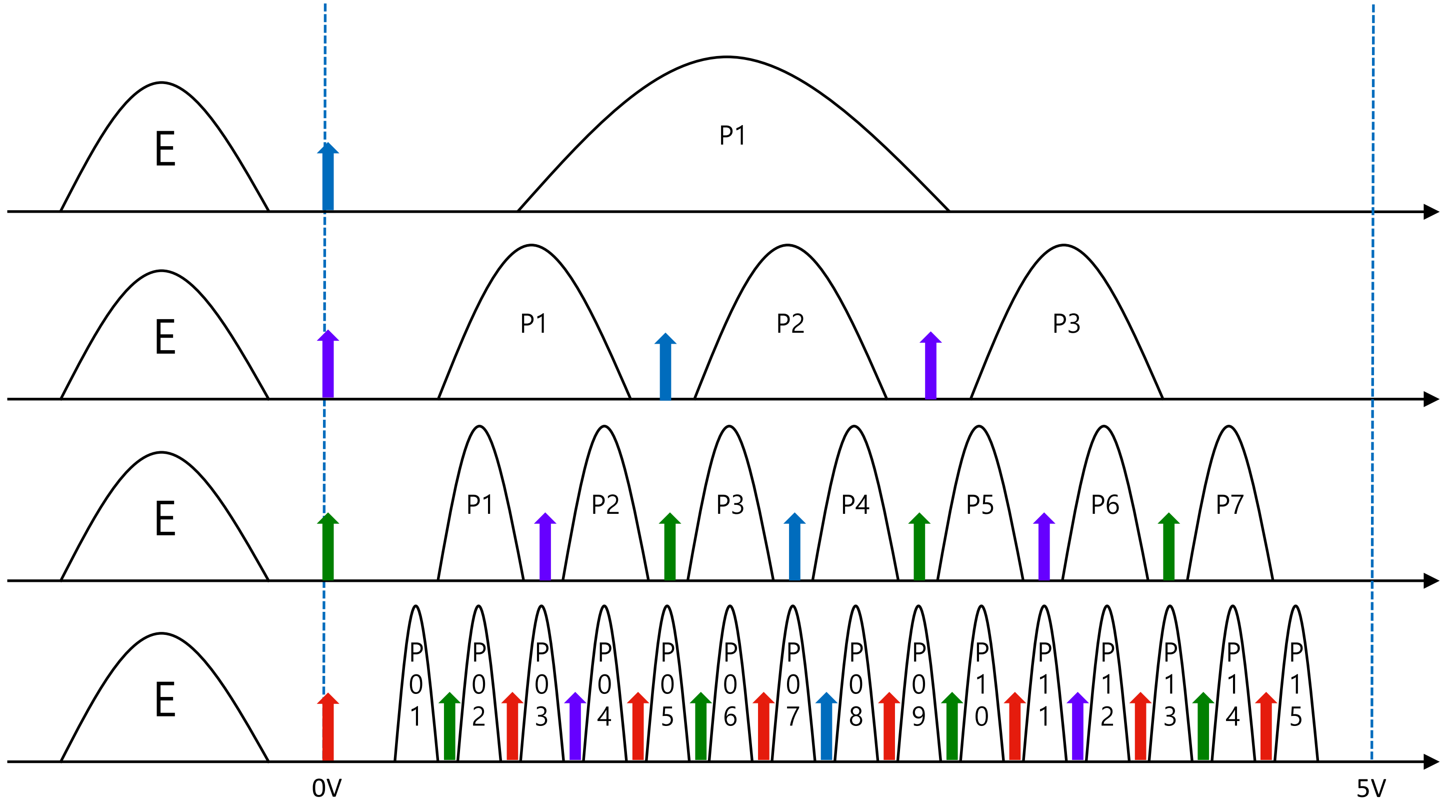
- QLC distribution is narrower and shorter distance to adjacent level



QLC Device – Read Time

- QLC requires a longer read time due to additional steps
  - Example: additional sensing with different *Selected W/L*





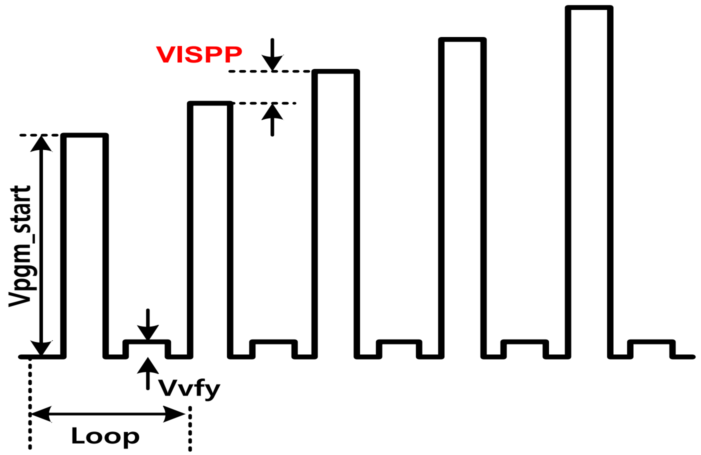
 |

Samsung proprietary

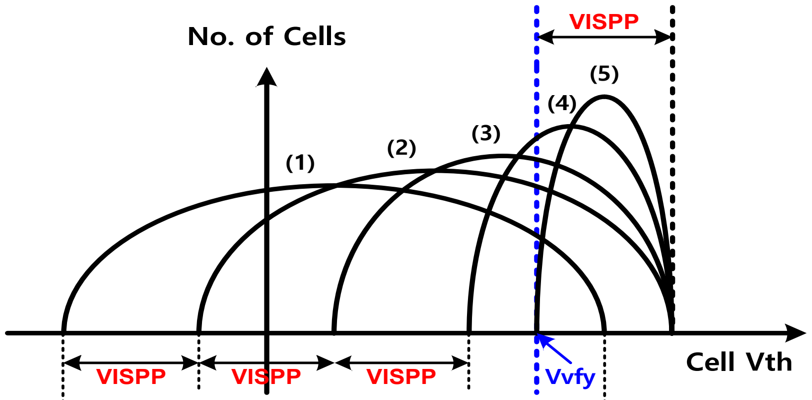
QLC Device – Program Time

- QLC requires a longer program time due to complex write algorithm
  - Example: ISPP (Incremental Step Pulse Programming) for sharper distribution

ISPP voltage steps



Effect of ISPP

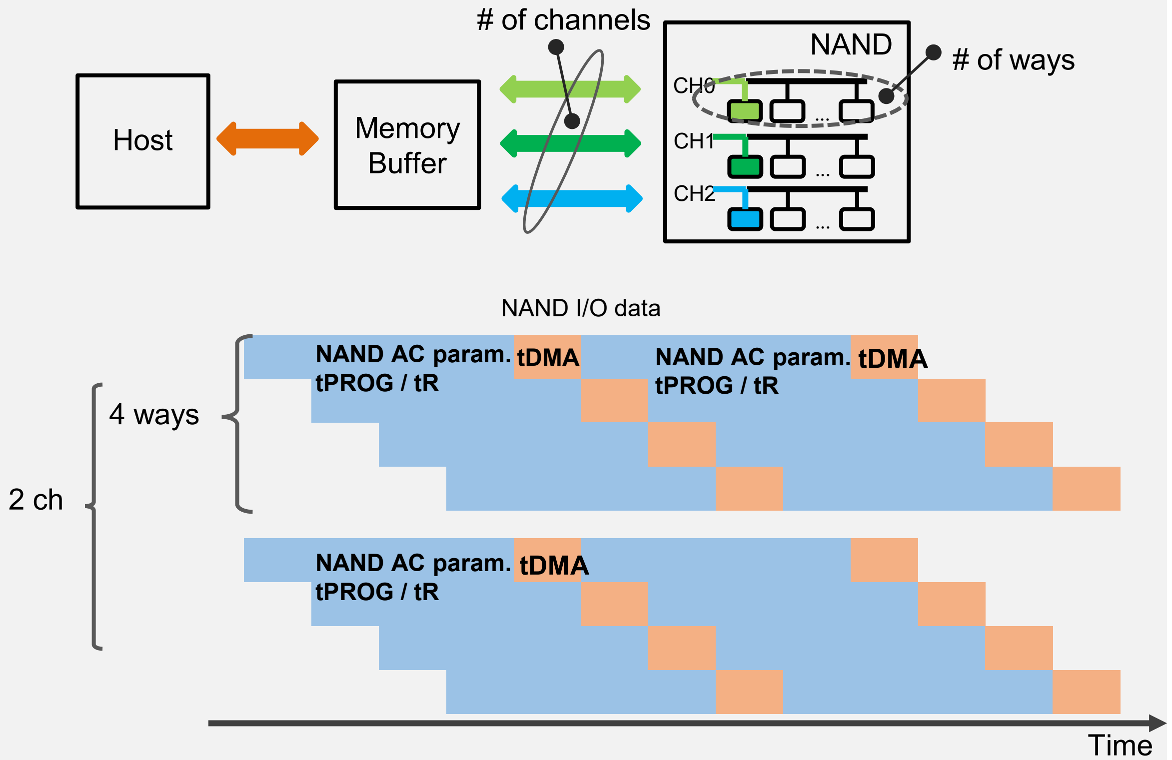


SSD Performance Factors

- Impact of channels, ways, AC parameter, tDMA varies depending on cases.

SSD Ways and Channels

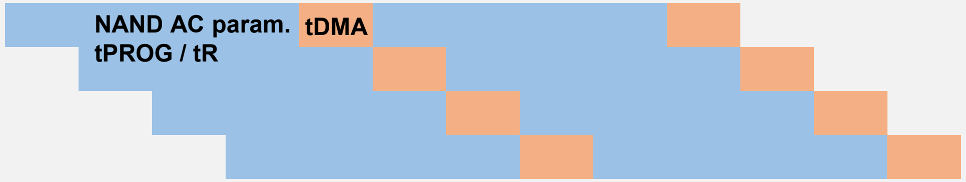
- Ways and channels increase performance with parallelism



NAND Package Level Performance Bound

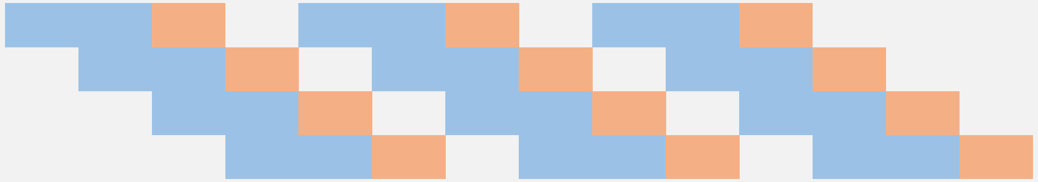
- Performance factors differ based on the bound case

< Bounded by NAND AC parameter >



✓ Dominant factors are AC parameter (tR, tPROG) and # of ways
✓ As AC parameter increases, the impact of tDMA decreases

< Bounded by tDMA >



✓ Dominant factor is tDMA (NAND clock speed) and CMD/ADD efficiency
✓ Performance is independent of # of ways and AC parameter

High Capacity QLC SSD Performance Factors

- # of ways and NAND AC parameters are dominant performance factors for high capacity QLC SSD

	TLC SSD	QLC SSD
Sequential Read	Host interface bandwidth (PCIe)	Host interface bandwidth (PCIe)
Random Read	tDMA	# of ways and tR
Sequential Write	# of ways and tPROG tDMA	# of ways and tPROG
Random Write	# of ways and tPROG tDMA	# of ways and tPROG

Contents

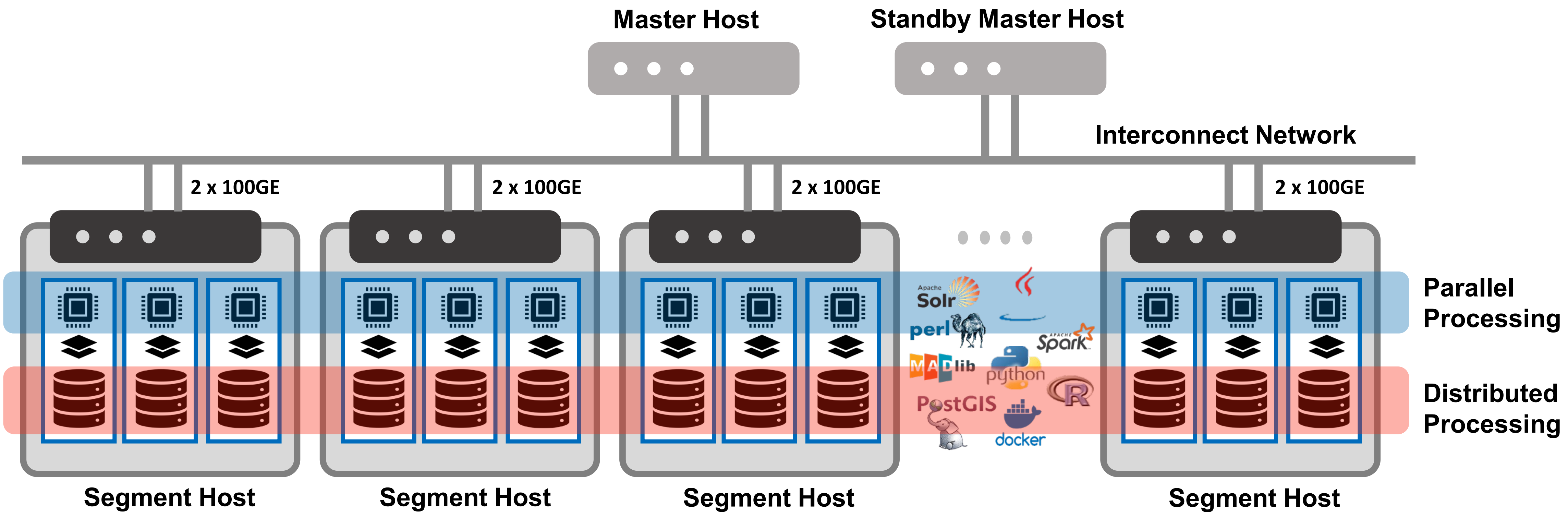
SSD Comparison

- Storage performance was evaluated with commercially available TLC and QLC SSDs

	16TB TLC SSD PCIe Gen5 NVMe	16TB QLC SSD PCIe Gen3 NVMe
Sequential Read	14,000 MB/s	3,200 MB/s
Sequential Write	7,000 MB/s	1,000 MB/s
Random Read	2,500 KIOPS	400 KIOPS
Random Write	360 KIOPS	36 KIOPS

Case 1: Massively Parallel Processing Database

- Massively parallel processing (MPP) database is exhibit high storage I/O demands.
- For this experiment, VMware Greenplum Database (GPDB) is used to test performance result of TLC and QLC SSD.

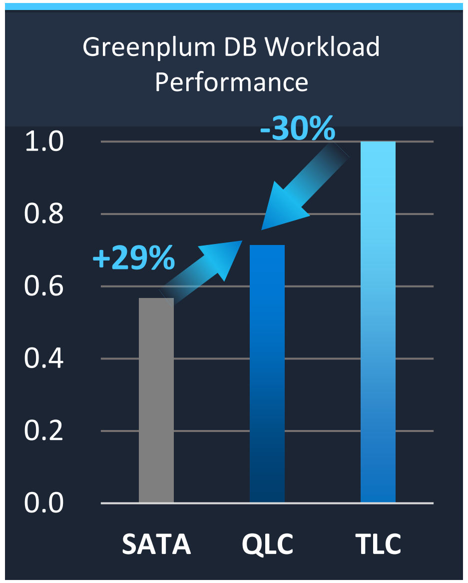


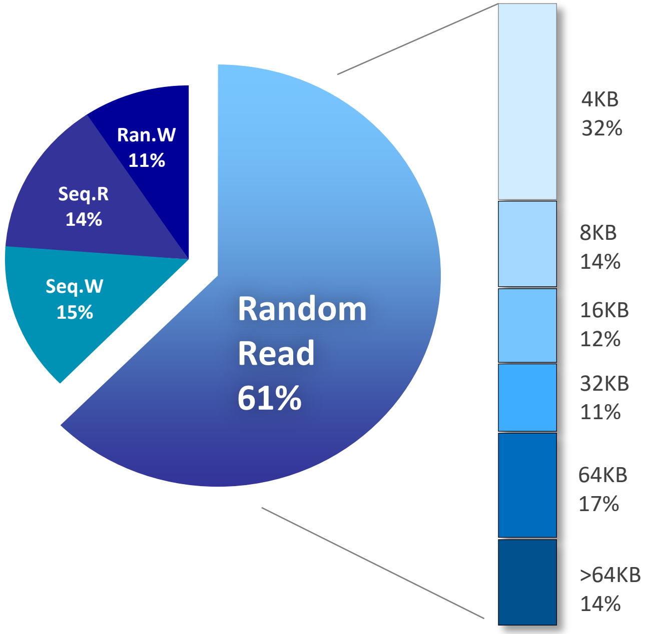
SDC 24

Case 1: Massively Parallel Processing Database

- 30% of performance drop has been observed. Dominant SSD I/O is random read with small chunk size.
- However, the performance with QLC storage is 29% higher than typical GPDB appliance with SATA SSD.

	SATA	TLC	QLC
# of node (server)	4	4	4
CPU Core	48	48	48
Memory per nod e	512 GB	384 GB	384 GB
SSD type	SATA	PCIe Gen5 NVMe	PCIe Gen3 NVMe
SSD Density	3.84TB	15.36TB	15.36TB
# of SSD per node	16	4	4



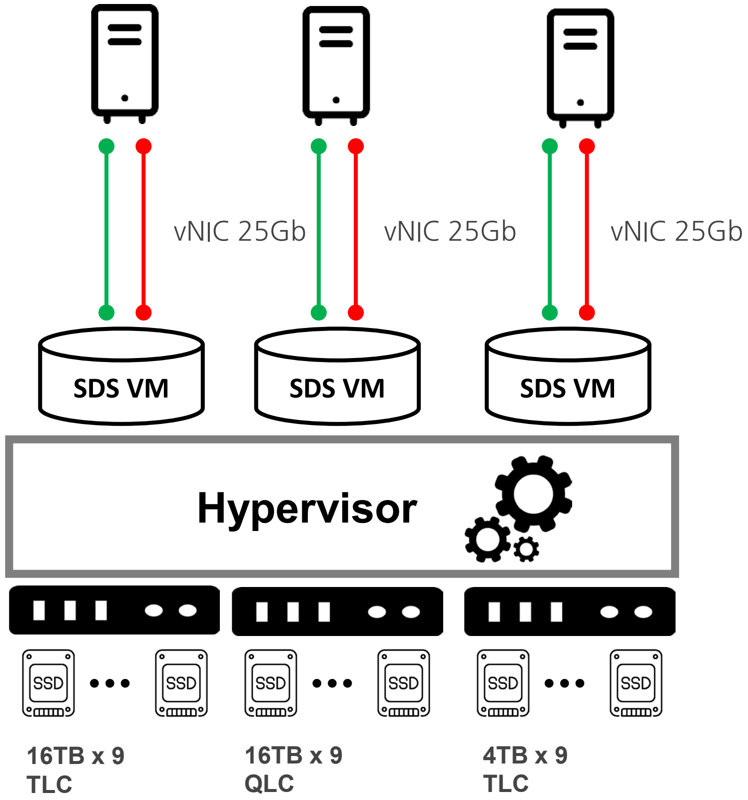


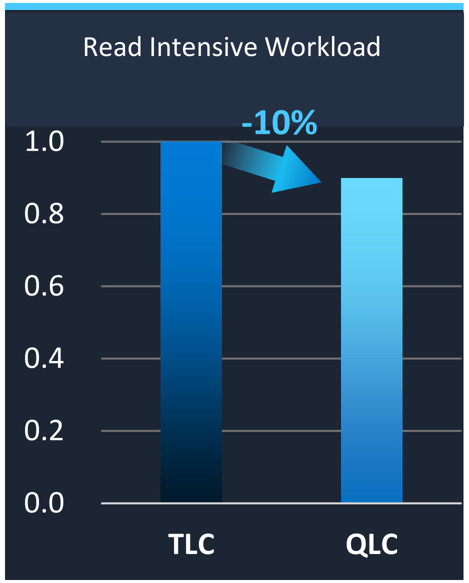
Workload is provided and validated by "VMware by Broadcom"

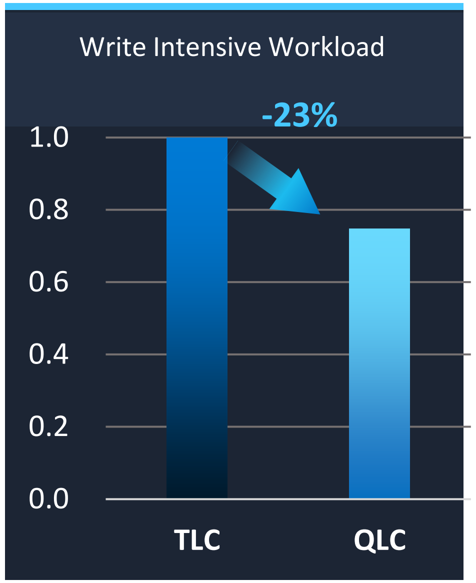
Samsung proprietary

Case 2: NAS File based Storage

- QLC SSD performance has a greater impact on write intensive workloads.

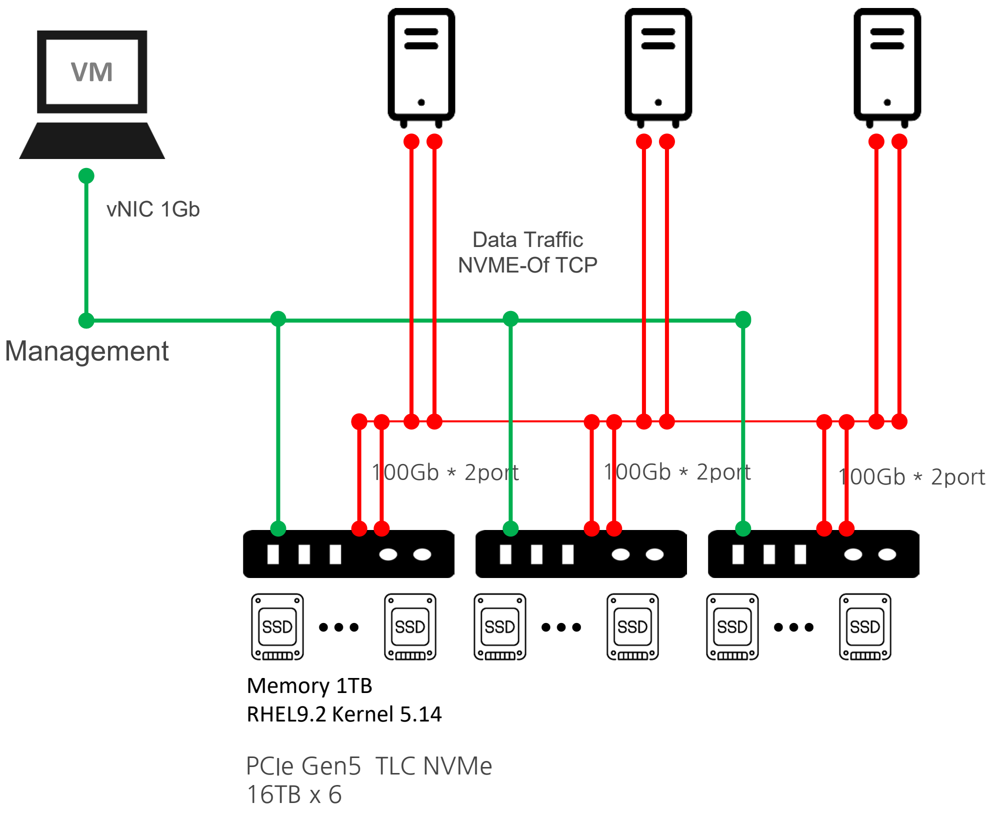


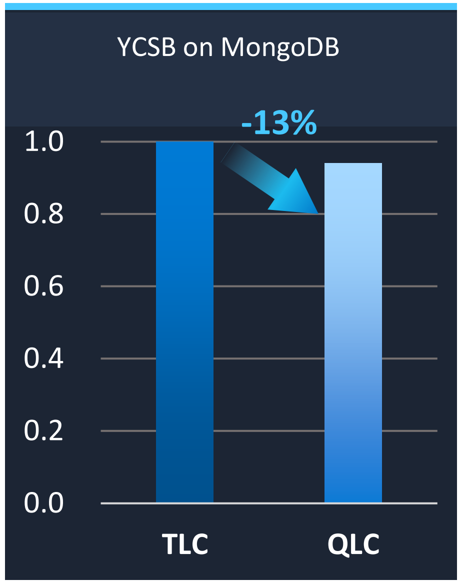


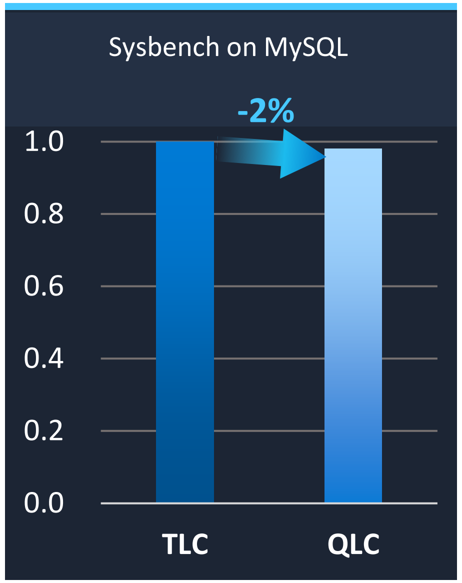


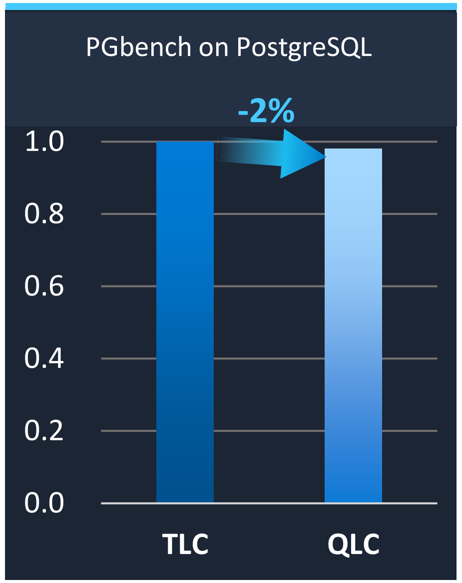
Case 3: Block Storage NVMeoF

- QLC SSD performance has less impact on workload where cache hit is high



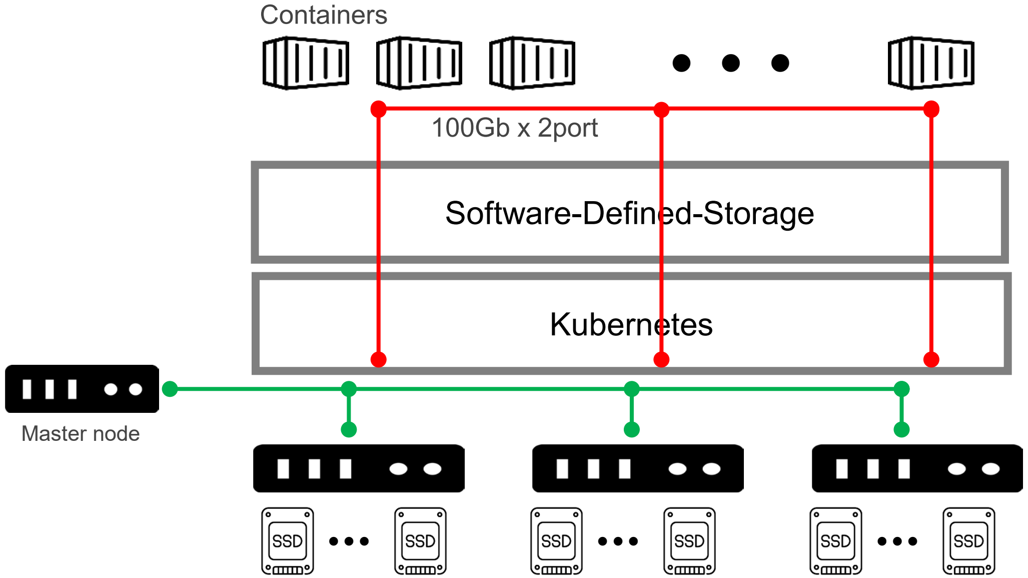


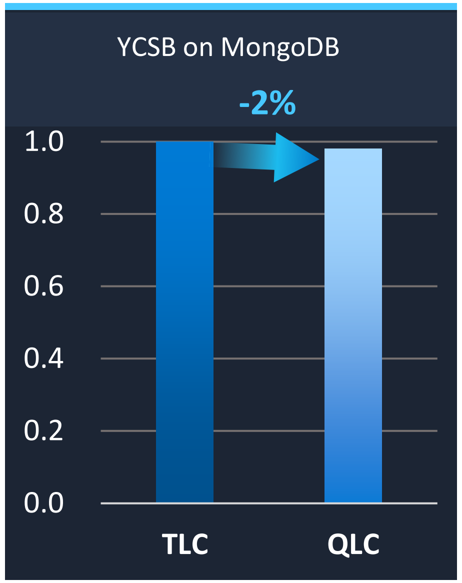


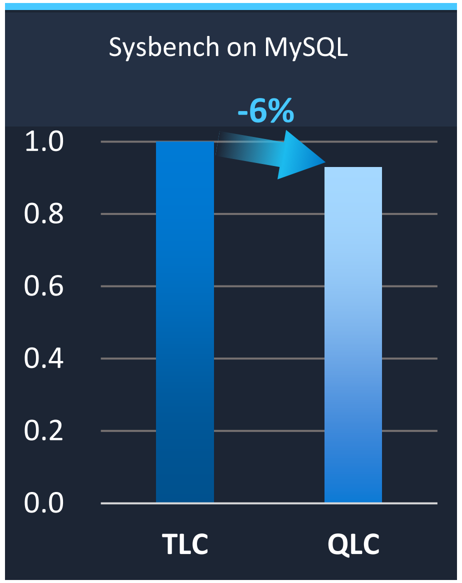


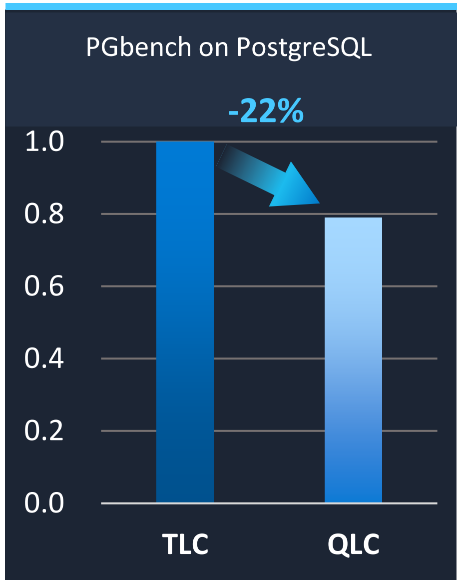
Case 4: Object Storage on Kubernetes

- QLC SSD performance has lower impact on workload where cache hit is high



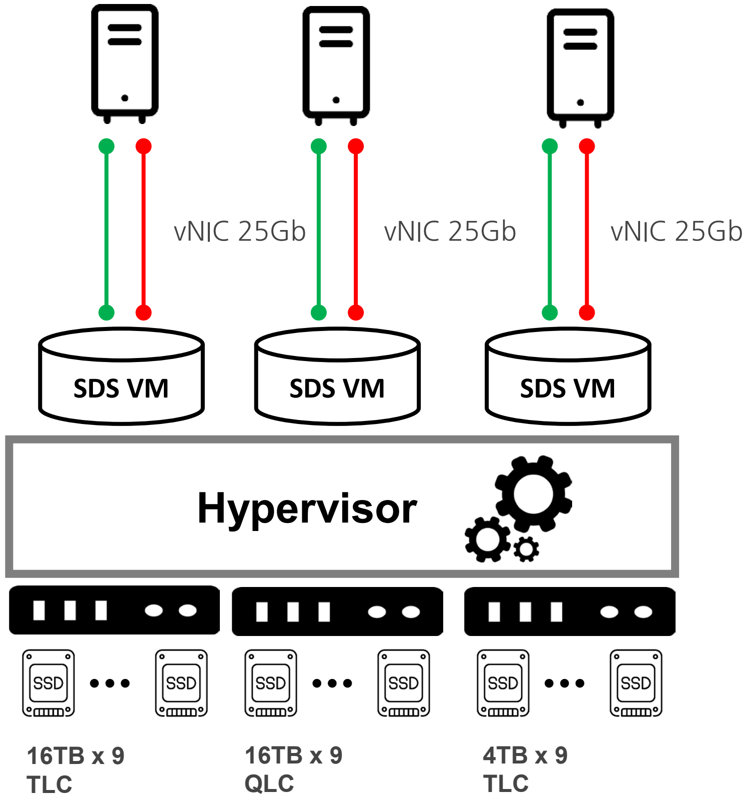


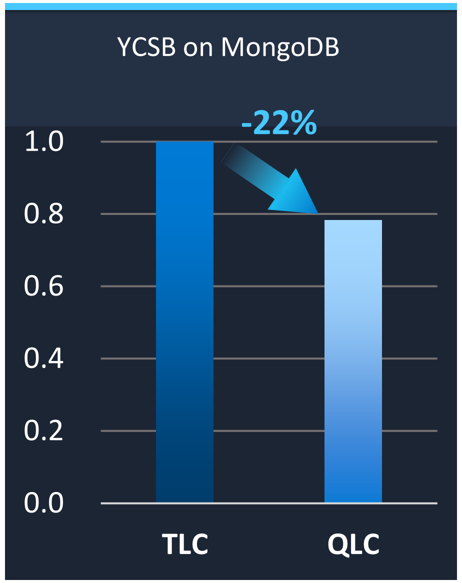


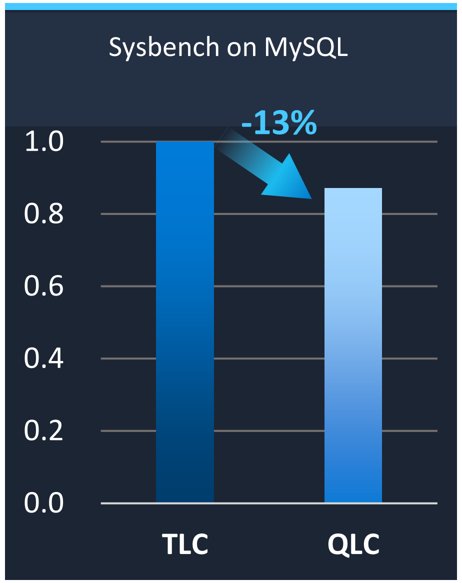


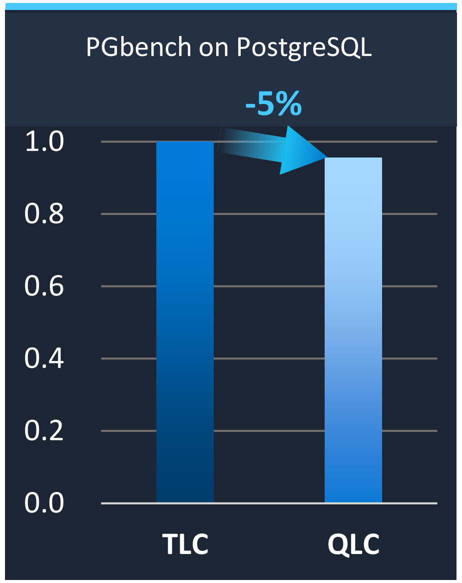
Case 5: Block Storage iSCSI

- QLC SSD performance has less impact on workload where cache hit is high









Case 6: Service Level Agreement of Availability

- Service Level Agreement, SLA, is an agreement between a service provider and a customer.
- SLA of availability can be determined by RTO (recovery time objective), where storage device RAID or erasure coding reconstruction time can be a limited factor for RTO

$$SLA = \frac{Total\ Time\ - Downtime}{Total\ Time}, \qquad where\ RTOs < Downtime$$

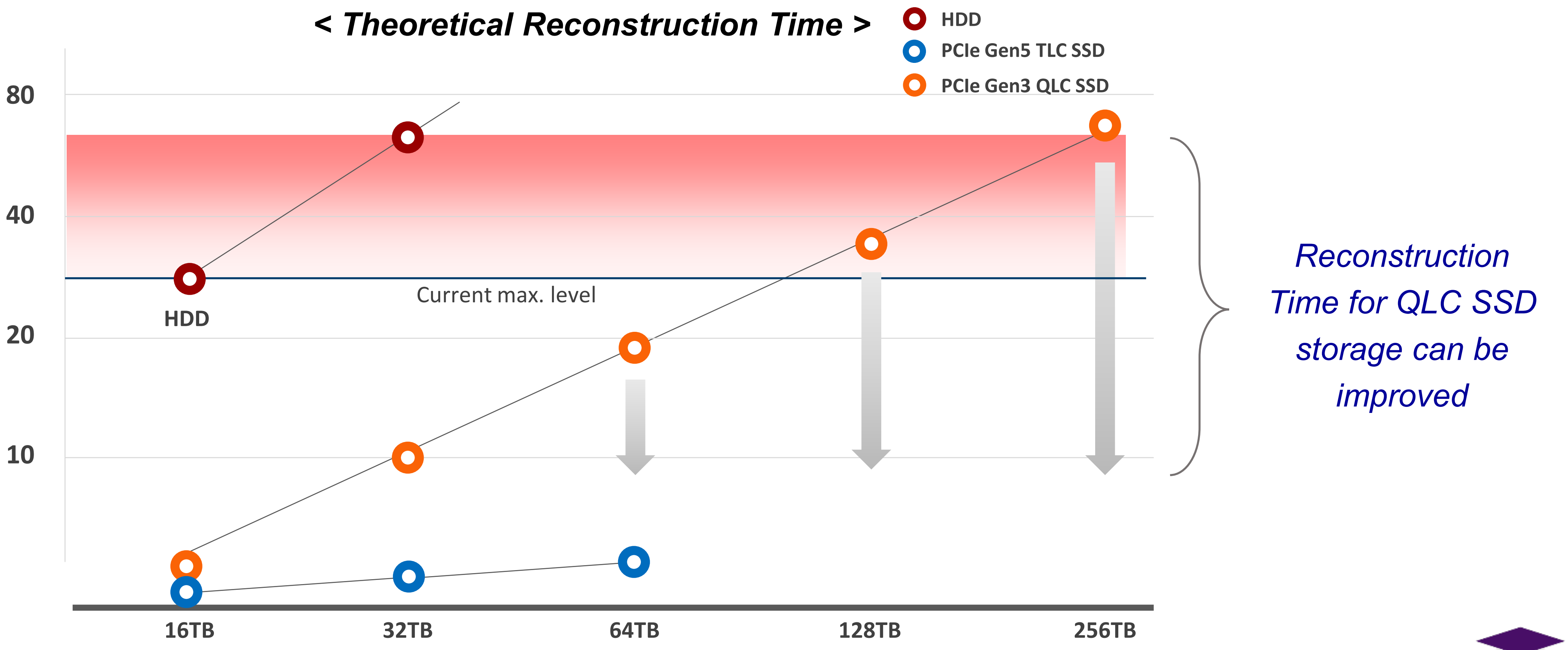
$$RAID\ Recovery\ Time = T_{read} + T_{write} + T_C\ + T_{Ready}$$

Assume disk failure is the only reason of service downtime

	1 disk failure / year	**2 disk failures / year**
Storage Configuration	HDD 24TB RAID 2:1 parity	HDD 24TB RAID 2:1 parity
RTO	44hrs 30min	89hrs
SLA	99.4920%	98.9840%

Case 6: Service Level Agreement of Availability

- Reconstruction time for QLC SSD storage will increase with respect to capacity.
- However, SLA will be improved with higher performance QLC SSD,



(reference: www.seagate.com, www.solidigm.com, www.samsung.com)

Samsung proprietary

Summary of Test Results

- Results from replacing high performance TLC with lower performance QLC

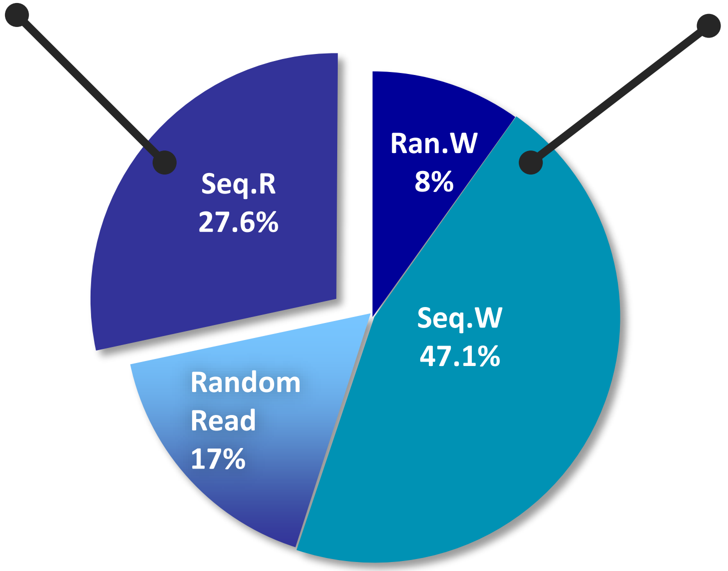
		Results
Performance Drops	Massively Parallel Processing DB on Bare Metal	30%
	NAS Filed based Storage	10 ~ 25%
	Block Storage iSCSI	5 ~ 22%
	Block Storage NVMeoF	2~13%
	Object Storage on Kubernetes	2~21%
SLA	Reconstruction Time	**+ 370%**

Contents

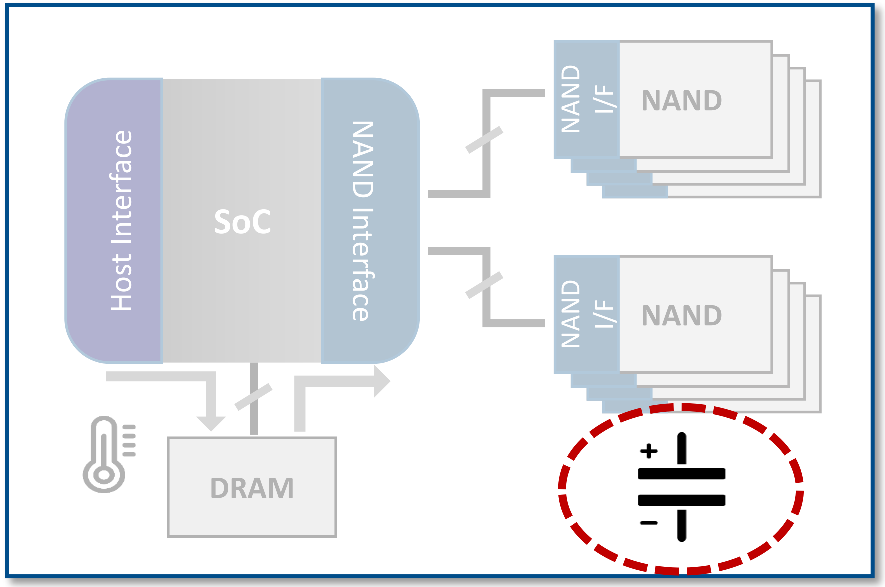
Service Level Agreement – Solution

- Reconstruction time of high capacity SSD will be improved for following reasons:

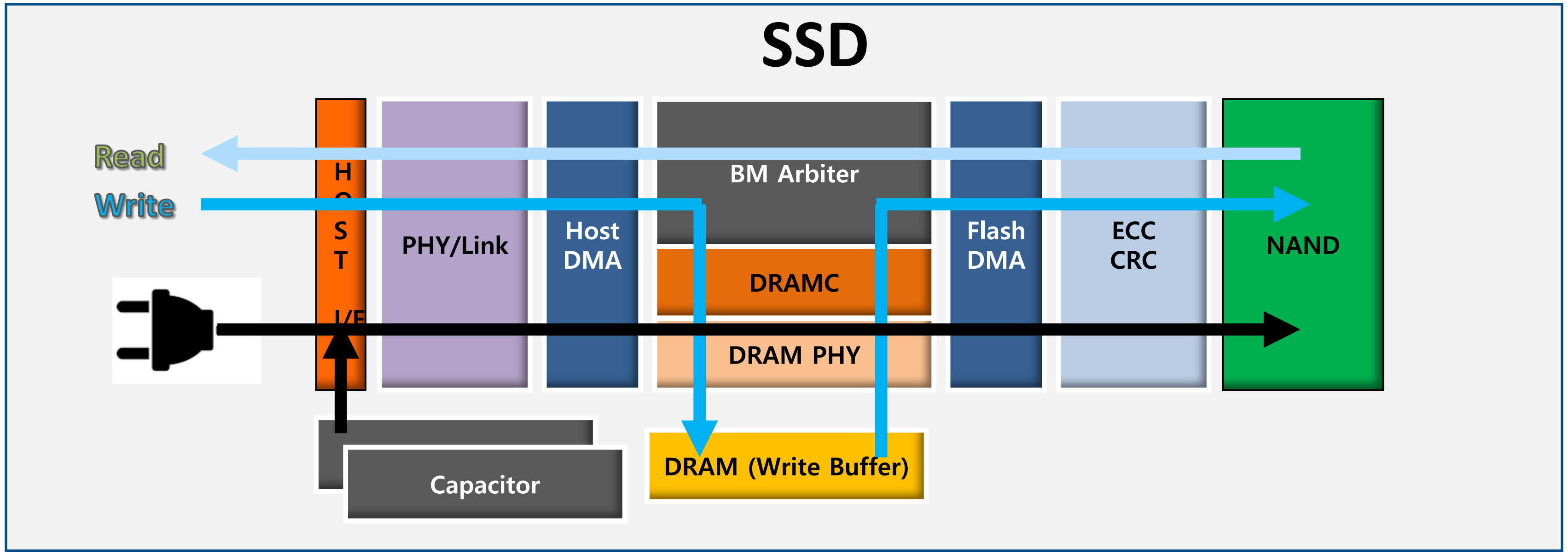
  - SSD interface speed
    PCIe Gen3 → Gen4/5
  - High capacity SSD will have more ways and planes
  - tDMA and AC parameter will be improved



However, CAPACITOR is an issue now



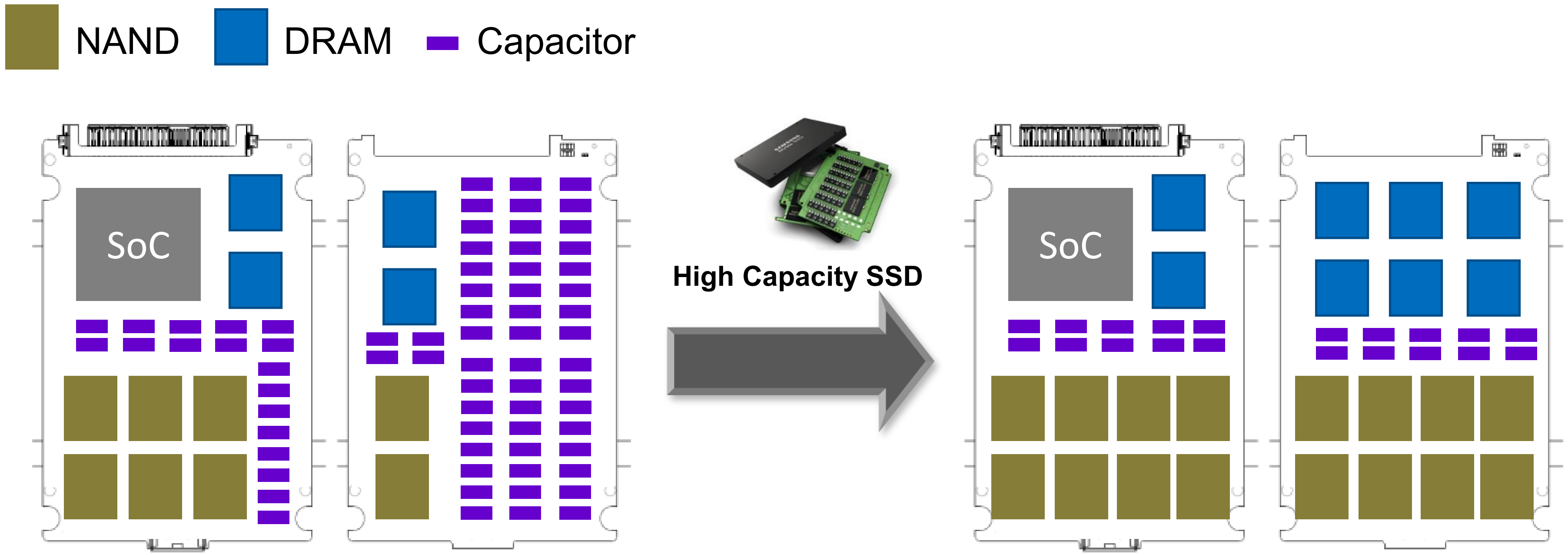
High Capacity SSD – Power Loss Protection



SPOR **(Sudden power off recovery)**

1. CTRL detects a drop in the input power below the threshold
2. Start flushing the data in-flight and the data present in the DRAM quickly to the NAND flash
3. With the power failure protect capacitor

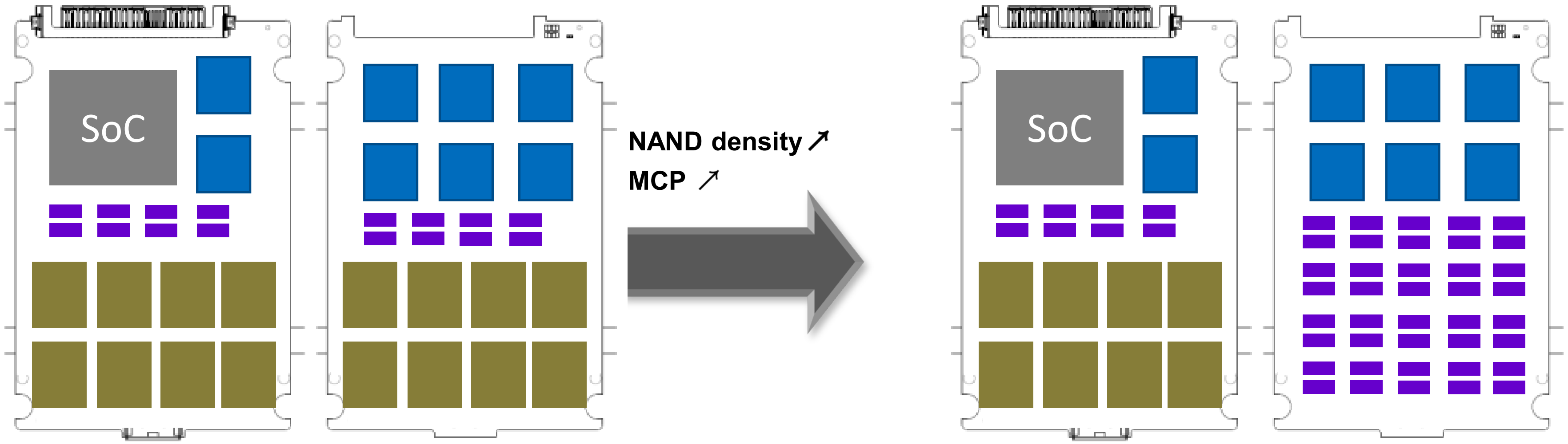
High Capacity SSD – Physical Space



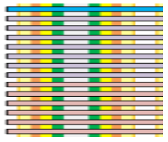
- High capacity SSD requires more NAND and DRAM packages.
- Space for capacitor will be diminished.
  ➔ Less buffer memory, lower write performance

High Capacity SSD – Physical Space

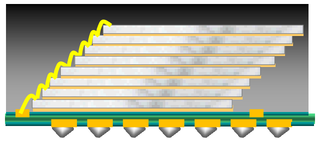
NAND DRAM Capacitor



- NAND density will continue to increase
  - More cell layers and vertical/horizontal scaling
  - Number of die on Multi-Chip Packaging (MCP) will be increased

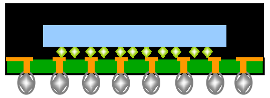


Vertical scaling



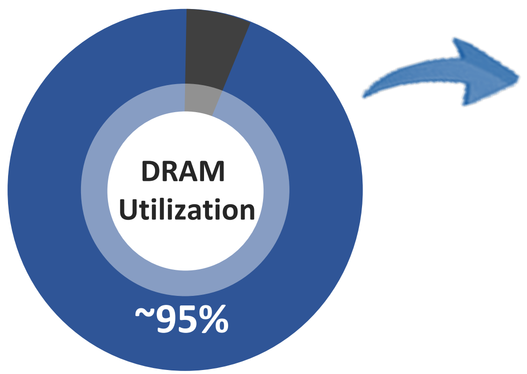
Multi die wire bonding

- DRAM density does not scale as NAND
  - DRAM die density does not increase w.r.t. storage roadmap, and DRAM package cannot have many die due to I/O speed

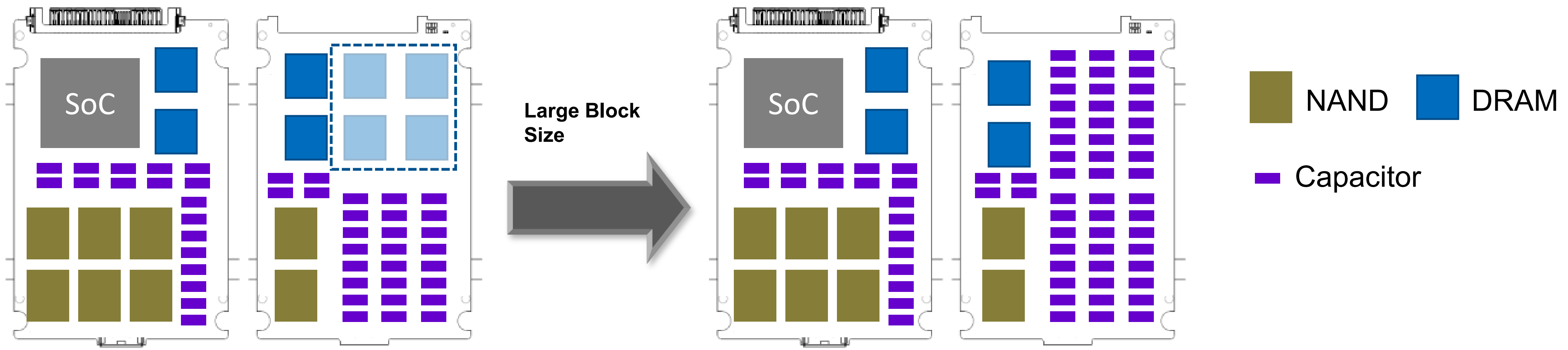


Flip chip on PCB

SLA of High Density QLC SSD – Solution: LBS

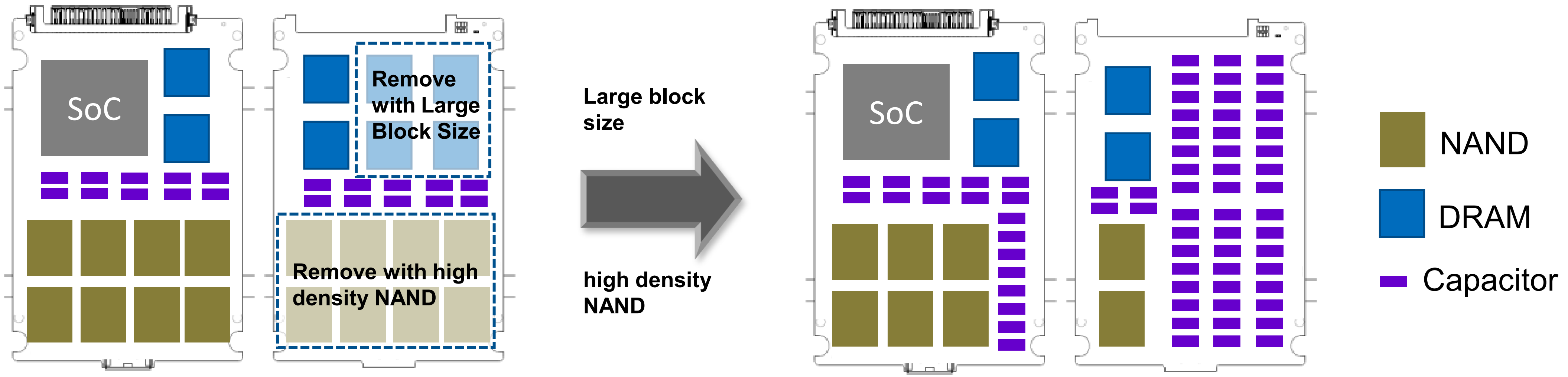


- Most of DRAM utilization is to store L2P mapping table
  - L2P is Logical-to-Physical table that translates logical block address to physical address
  - Size of current logical block address is 4KB
- DRAM capacity decreases with larger logical block size
  - SSD access size 4KB → 16/32KB results 1/4, 1/8 DRAM capacity
  - The solution to have more capacitor, hence higher write speed



SLA of High Density QLC SSD – Solution: LBS

- Improve sequential write and random read performance of SSD
- Allocate space in SSD for capacitors



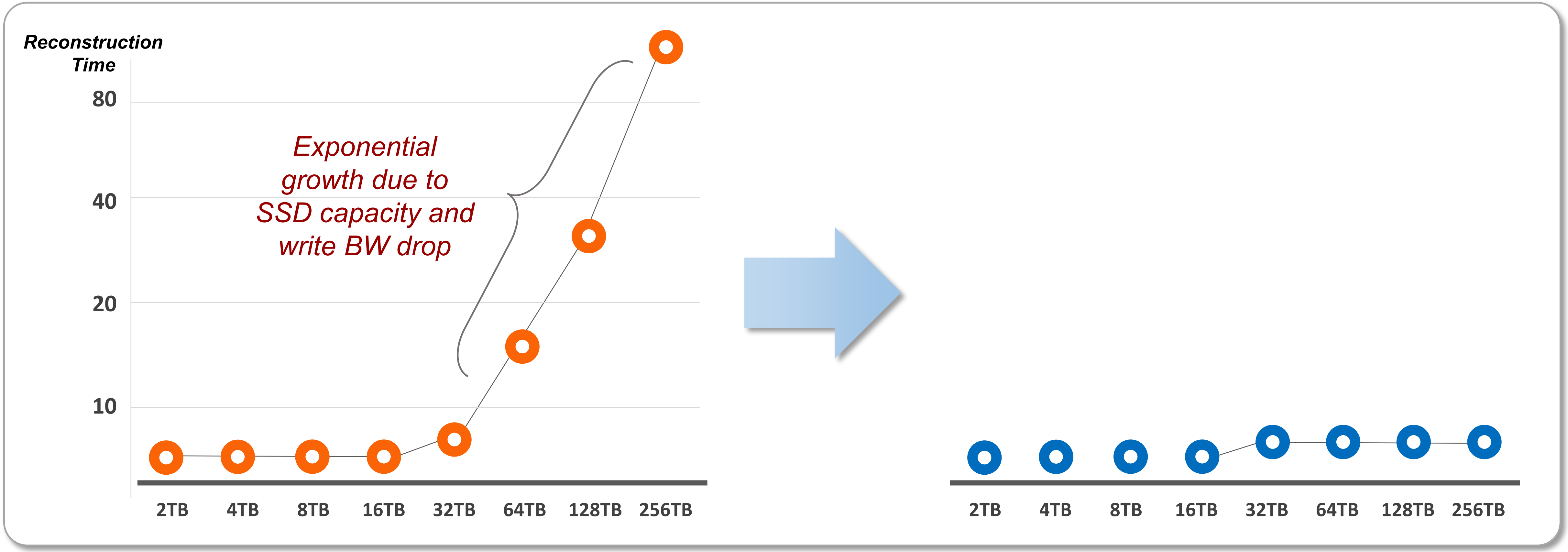
NAND device makers are developing high density NAND, but

→ **Industry need to collaborate to establish LBS ecosystem.**

(SSD, storage S/W, operating system, platform, hypervisor, etc)

SLA Forecast

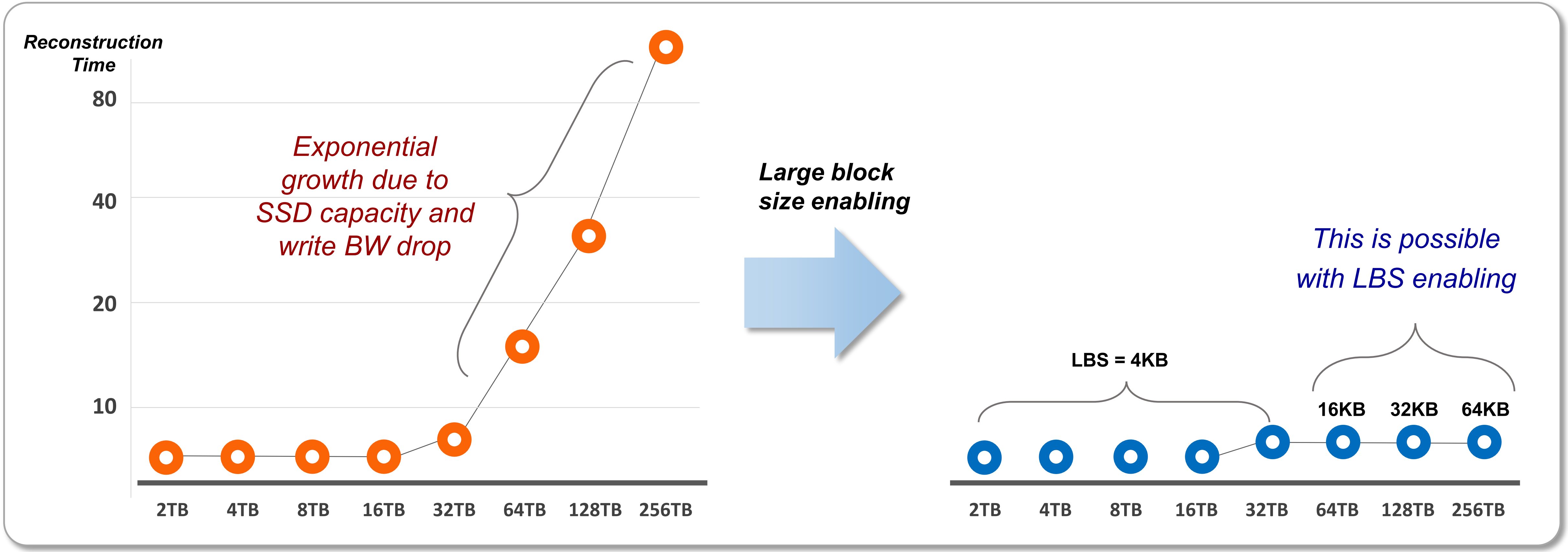
- SLA will remain at the current SSD storage level with the implementation of ____?



Samsung proprietary

SLA Forecast

- SLA will remain at the current SSD storage level with the implementation of LBS



Conclusion

- We request collaboration to build an ecosystem optimized for high-capacity SSD and LBS
  - Work together with Storage S/W, O/S, SSD device maker
  - Optimize the entire system, including file system, metadata, snapshot, compression, deduplication, compaction, RAID, erasure coding
- Samsung is working on LBS Eco system, and device level tests are available
- SMRC(Samsung Memory Research Center) can build a cluster of storage with LBS for such collaboration since various configuration, environment and Samsung PoC devices are available

- For more information about LBS and technical requirement, attend session:

"SSD Architecture Challenges with DRAM" by Dan Helmick.