Cloud Storage Acceleration Layer FDP NVMe Technology Implementation and Scaling with PCIe Gen5 Cache SSD

Presenter: Mariusz Barczak, Principal Engineer, Solidigm

August 2024

Authors: Mariusz Barczak, Wojciech Malikowski, Mateusz Kozlowski

Solidigm Disclaimers

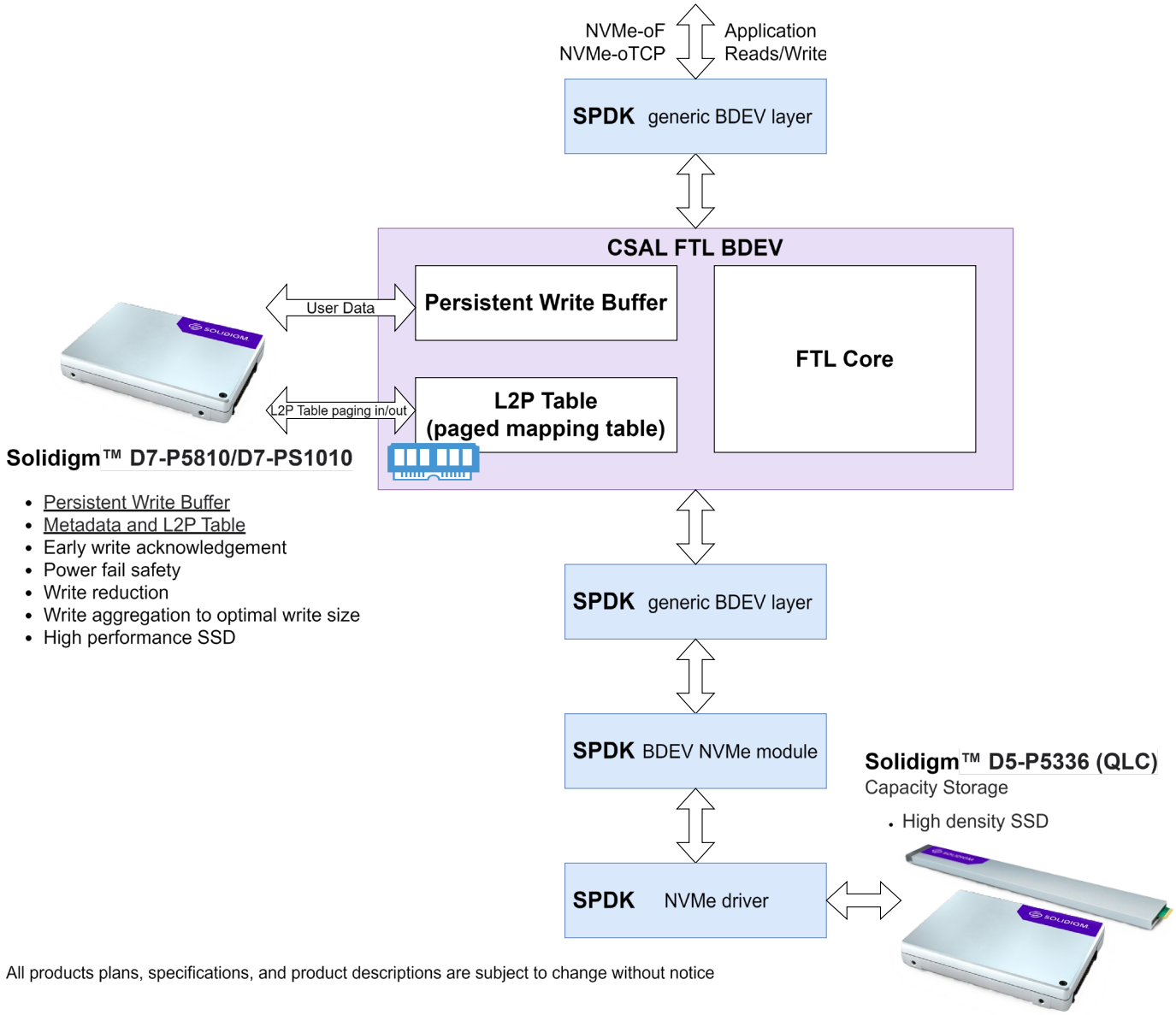
All information provided here is subject to change without notice.

- The products described in this document may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Solidigm technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer.
- Solidigm disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.
- Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.
- Cost reduction scenarios described are intended as examples of how a given Solidigm-based product, in the specified circumstances and configurations, may affect future costs and provide cost-savings. Circumstances will vary. Solidigm does not guarantee any costs or cost reduction.
- Solidigm does not control or audit the design or implementation of third-party benchmark data or Web sites referenced in this document. Solidigm encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase.
-

What's CSAL?

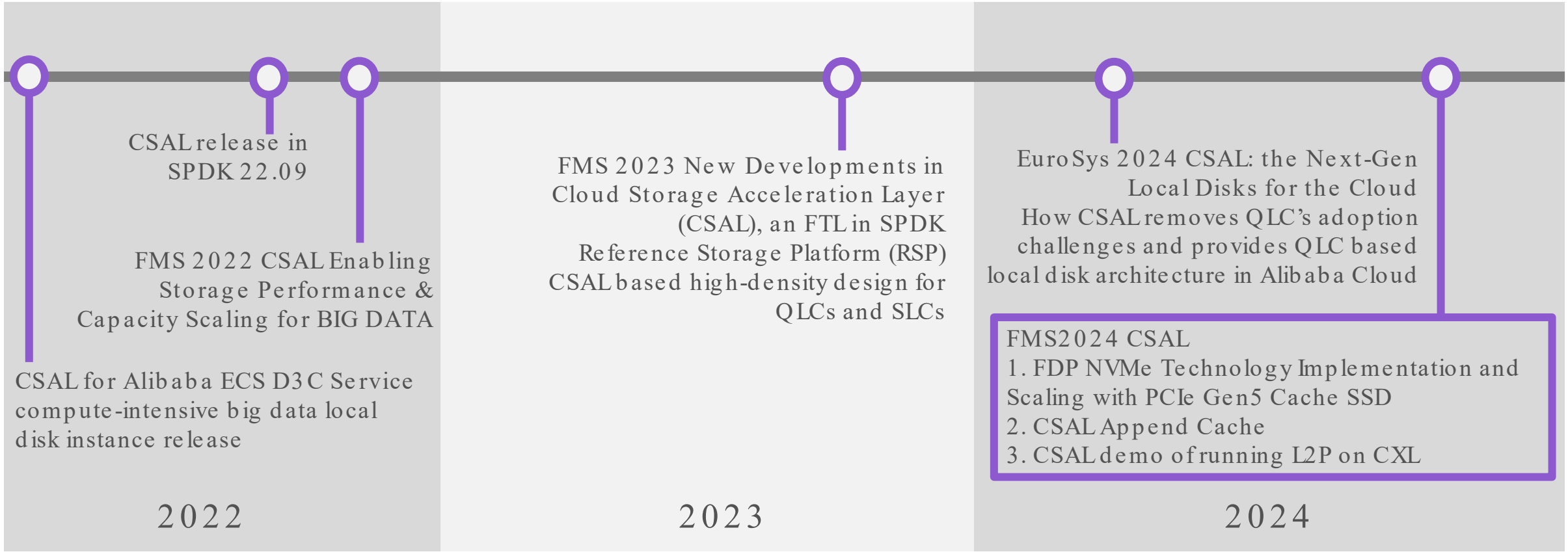
- Open-source cloud-scale shared-nothing Flash Translation Layer (FTL bdev) in Storage Performance Development Kit (SPDK)
- CSAL provides transparent block device to the upper application
- Ultra-fast cache and write shaping tier to improve performance and endurance to scale QLC value
- Consistent performance in multi-tenant environment
- Flexible scaling of NAND performance and capacity to the user/workload needs

Cloud Storage Acceleration Layer (CSAL)



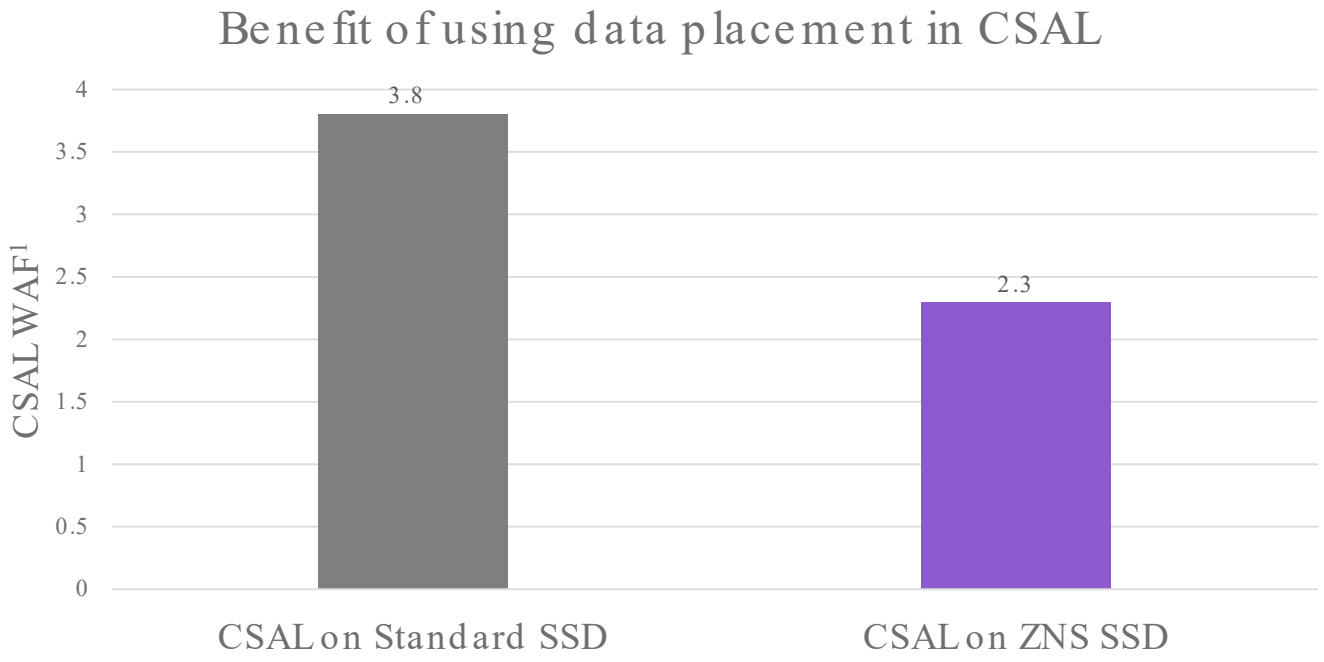
All products plans, specifications, and product descriptions are subject to change without notice

CSAL Evolution



Is data placement needed for CSAL?

- Handling heterogenous streams without sorting them increases write amplification factor (WAF)
- When placing data separately it reduces data movement and decreases WAF
- CSAL proved this method works with ZNS drives
- The same effect can be achieved when CSAL uses Flexible Data Placement (FDP) drives



Tenants:

Job	Block Size [KiB]	Pattern	Queue Depth
1 writer	4	sequential	128
1 writer	4	random	128
1 writer	4	zipf 0.8	128
1 writer	4	zipf 1.2	128

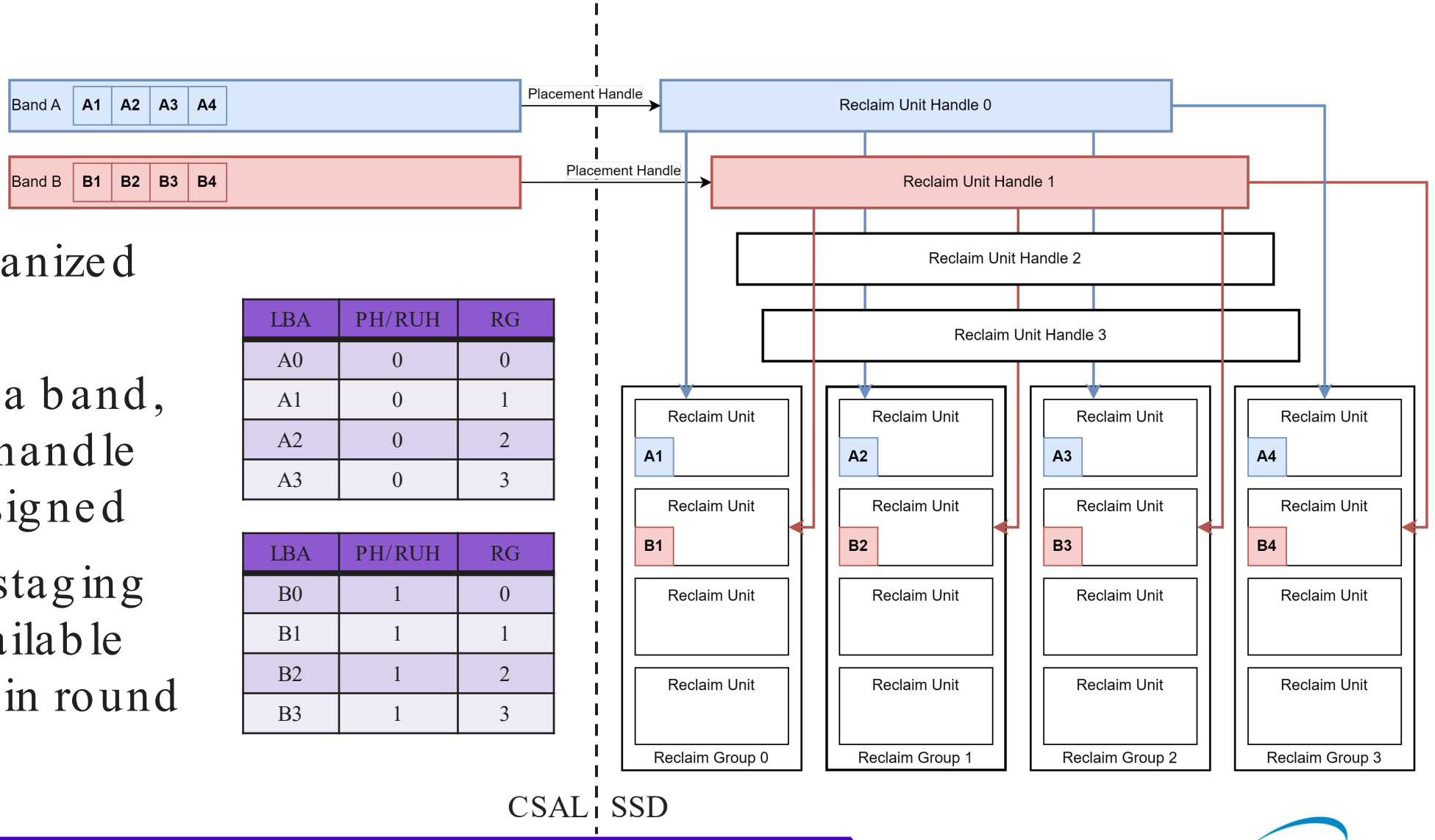
Yes, data placement helps to reduce system WAF

FDP implementation in CSAL

- CSAL is still organized using bands
- When opening a band, the placement handle needs to be assigned
- CSAL can start staging writes using available reclaim groups in round robin order

LBA	PH/RUH	RG
A0	0	0
A1	0	1
A2	0	2
A3	0	3

LBA	PH/RUH	RG
B0	1	0
B1	1	1
B2	1	2
B3	1	3

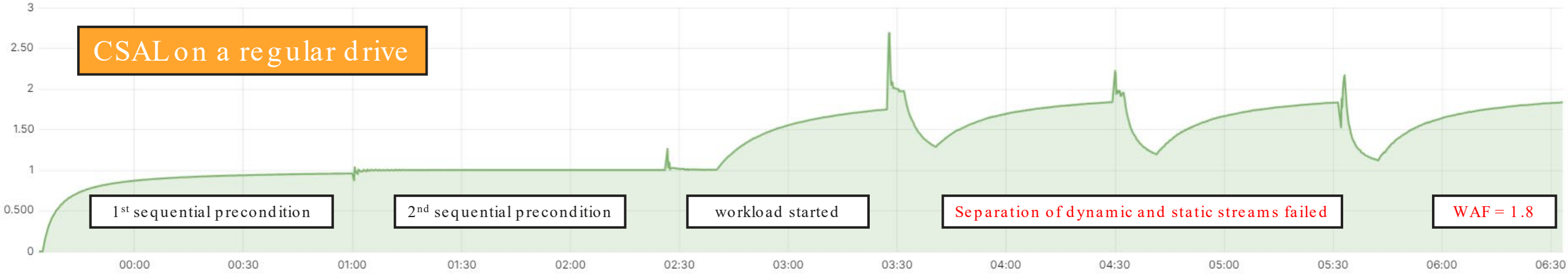


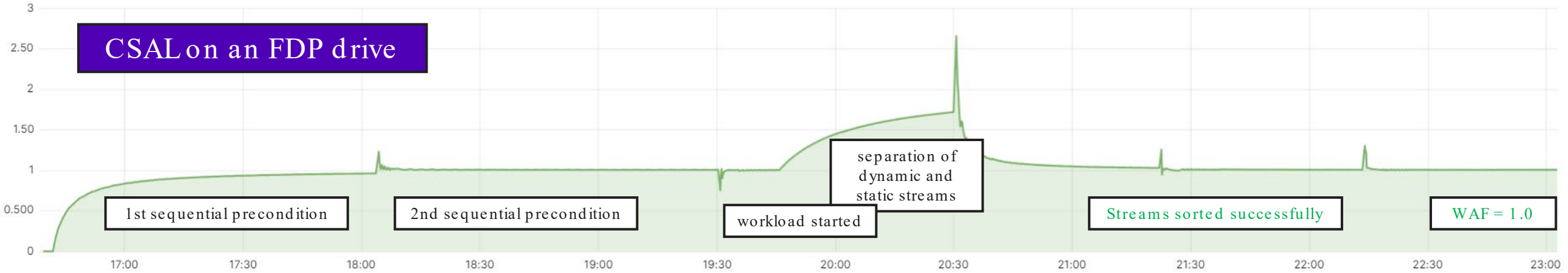
CSAL+FDP Evaluation

Workload and setup:

- Software: CSAL on FDP which separates internal streams of compaction and GC
- QEMU based environment to emulate an FDP drive
- Precondition:
  - fill all partitions executing sequential write
- Heterogenous streams workload example:
  - 8 jobs simulating independent tenants or streams
    - 4 jobs: 64k sequential writes → dynamic stream
    - 4 jobs: 64k random reads → static stream

CSAL Write Amplification Comparison



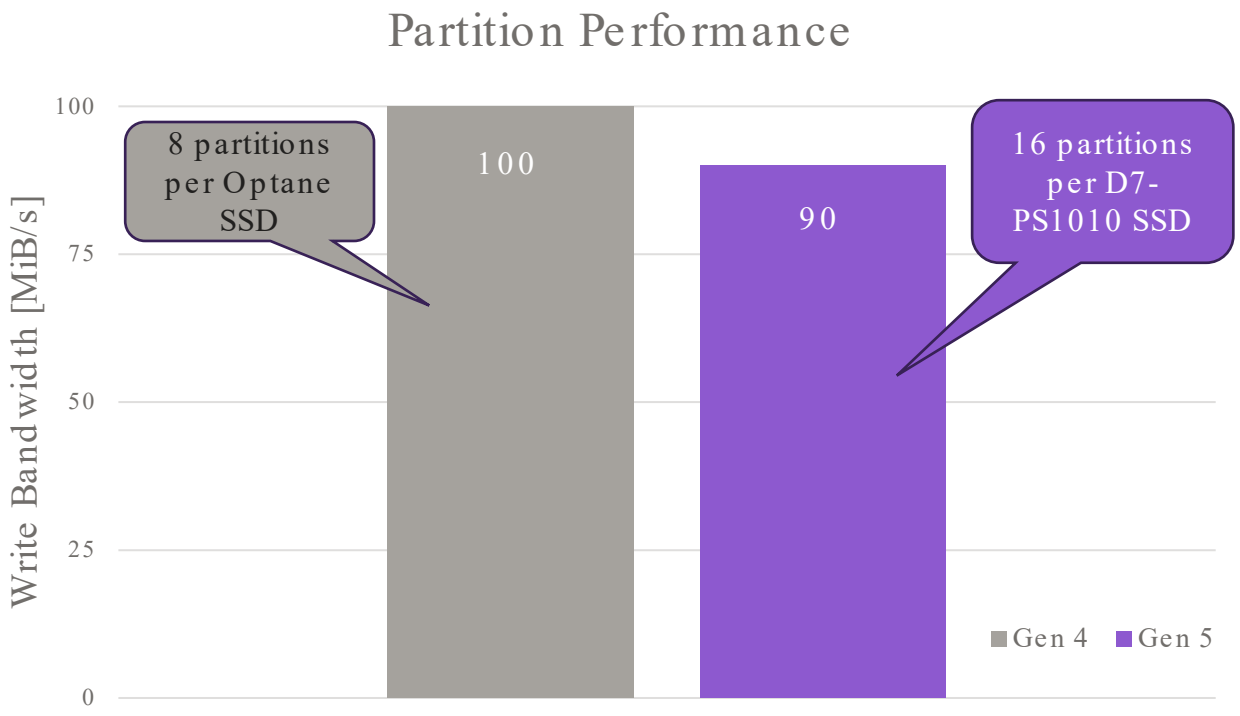


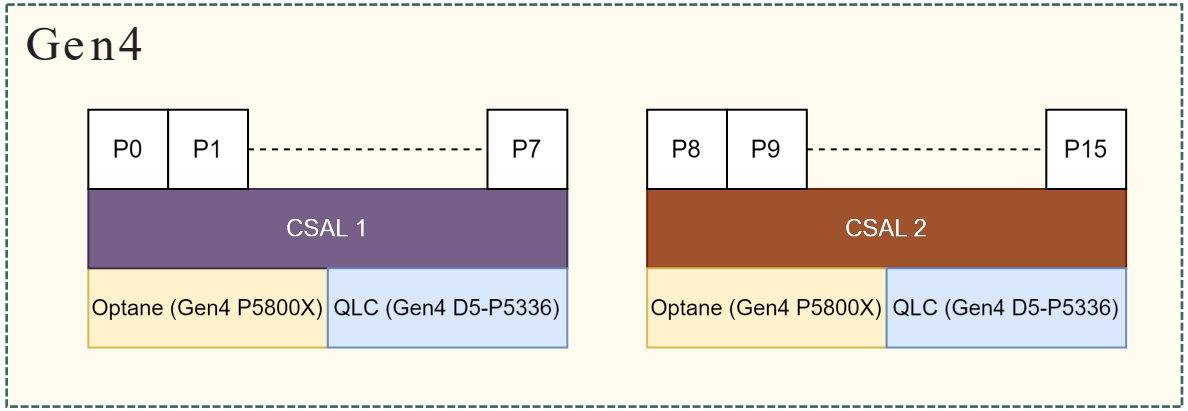
CSAL using FDP reduces Write Amplification Factor

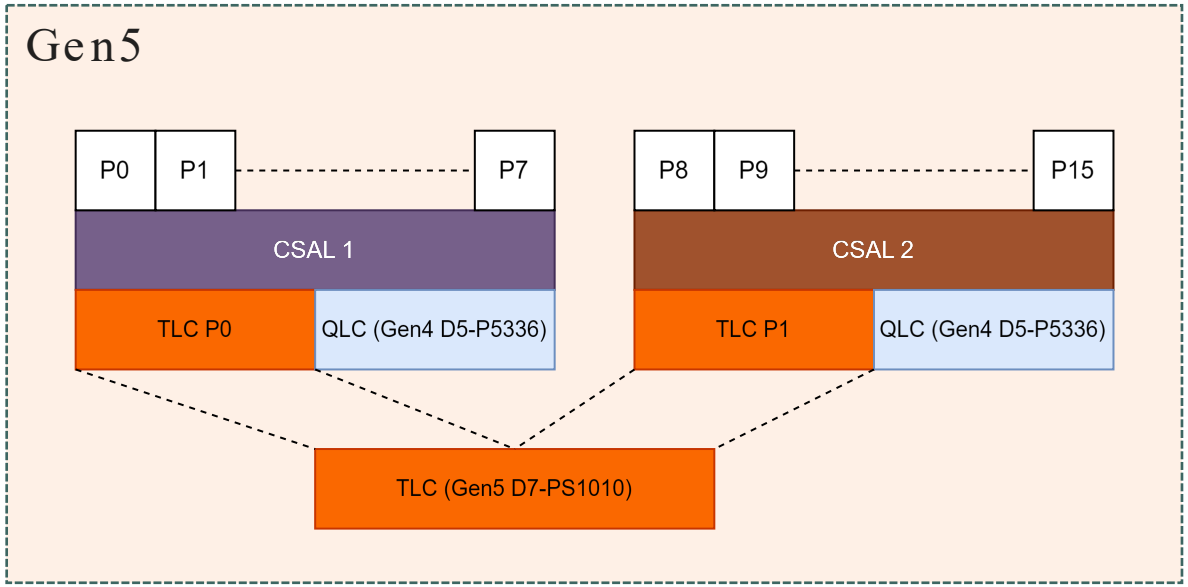
CSAL – PCIe Gen5 Cache SSD Evaluation

- PCIe Gen5 SSDs almost double the bandwidth
- Hypothesis
  - Can a single Gen5 SSD support 2x FTLs as compared to Gen4 without compromising performance?
- Test setup and config:
  - Software: CSAL
  - Drives: QLC Solidigm™ D5-P5336 15TB, TLC Solidigm™ D7-PS1010 3.84TB, Intel®Optane™ SSD DC P5800X 800GB,
  - Workload: 16 jobs, uniform random writes, block size: 64kiB
  - Server: Intel®Xeon®Gold 6426Y
  - Operating system: Fedora 39, kernel version: 6.8.7-200.fc39.x86_64

Scaling Number of FTLs with Gen5







NVMe PCIe Gen5 drive improves platform utilization while delivering comparable performance

Gen5 and Gen4 Cache Drive Impact

	Gen4	Gen5
PCIe Slots	4	3
Job Performance [MiB/s]	100	90
Relative SSD Cost[2]	100%	~75%
Active Power	2x 21W (Optane) + 2x 24W (D5-P5336)	1x 25W (D7-PS1010) + 2x 24W (D5-P5336)
Idle Power	4x 5W	3x 5W

Pros

Cons

Mitigation

Better Performance

Not affecting failure points

Higher Cost

More PCIe slots

Less PCIe slots

Cost Reduction

Extended Point of Failure on cache

Cache WAF might increase

RAID1

FDP

Using a Mixed Media Drive

Lower (Acceptable) Performance

Conclusion

- CSAL demonstrates WAF reduction when handling various streams/tenants with FDP
- CSAL is now Flexible Data Placement ready
- Gen5 SSDs can be deployed with CSAL reducing total cost of ownership

Next Steps

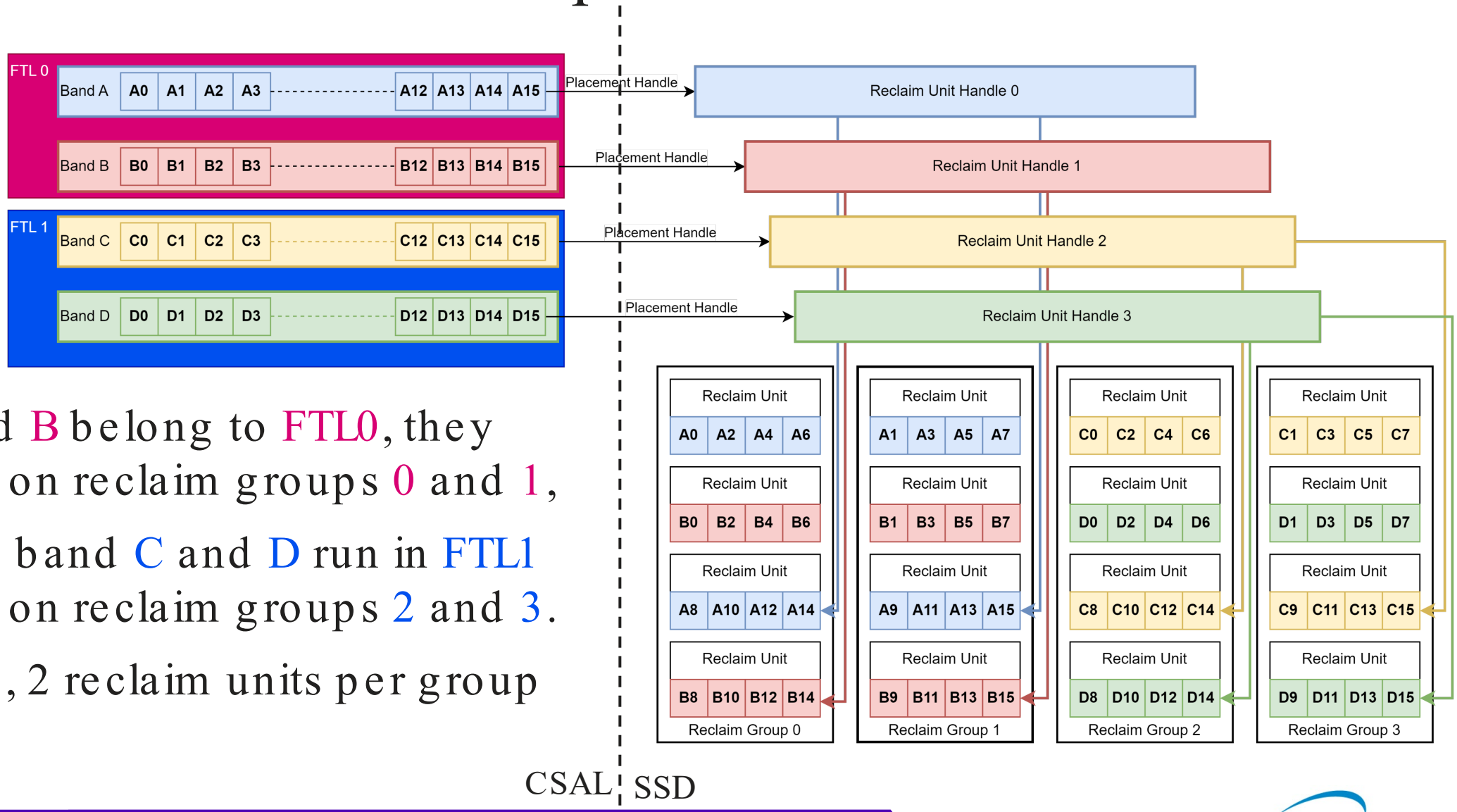
- CSAL is going to enhance usage of FDP technology (e.g., full tenant isolation) and scale with PCIe Gen5 bandwidth.
- Evaluation with NAND based FDP SSD.

Q&A
Cloud Storage Acceleration Layer

Contact us:
dl_csal@solidigm.com

Backup

CSAL FDP Multi-tenant implementation



- Band A and B belong to FTL0, they are placed on reclaim groups 0 and 1,
- Meanwhile band C and D run in FTL1 are placed on reclaim groups 2 and 3.
- To fill band, 2 reclaim units per group are utilized

[1] CSAL manages SSD WAF level of 1.0 thus CSAL WAF is equivalent to the system WAF

[2] Approximate Gen5 and Gen4 SSD Cost Relation Calculation

	$/GB	Capacity	Drive Cost
Optane	0.75	800	600
TLC	0.15	3840	576
QLC	0.10	15360	1536
Solution Cost	**Backend Drive Cost**	**Cache Drive Cost**	**Solution Cost**
PCIe Gen4	1536	1200	2736
PCIe Gen5	1536	576	2112
	Cost Relation		
PCIe Gen4	**100%**		
PCIe Gen5	**77%**		