Performance Analysis of 3D XPoint SSDs in Virtualized and non-Virtualized Environments

ICPADS 2018

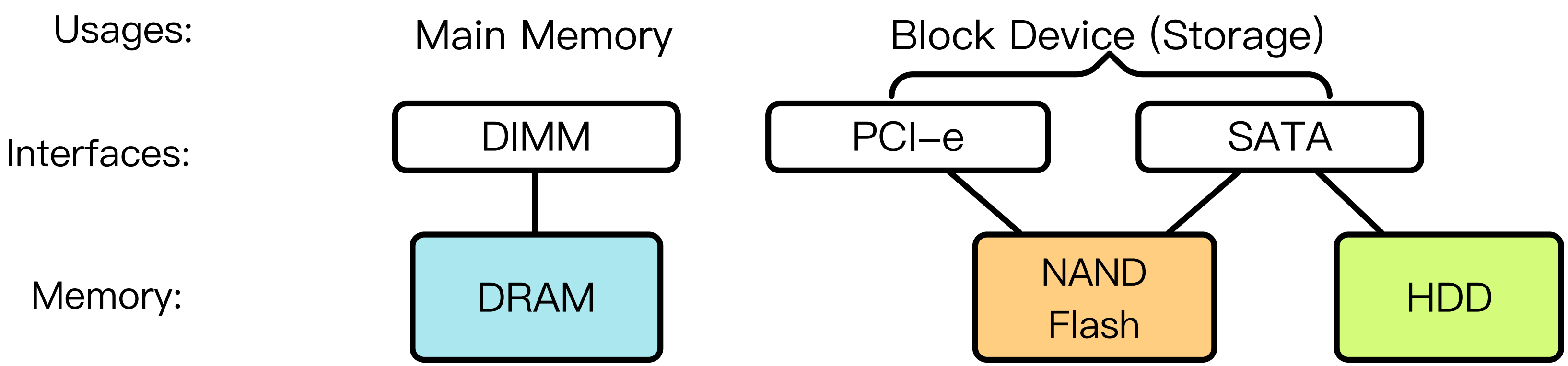
Jiachen Zhang, Peng Li, Bo Liu, Trent G. Marbach, Xiaoguang Liu, Gang Wang

Nankai - Baidu Joint Lab, Nankai University, China

Parallel and Distributed Software Technology Lab

Nankai - Baidu Joint Laboratory

Memory and Storage are Separated before …

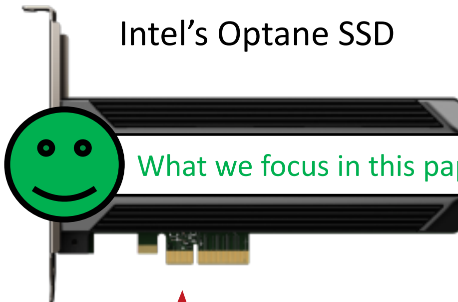


Memory and Storage are Separated before ...

The Non-Volatile Memory

Intel's Optane DC Persistent Memory

Not broad available now.



Intel's Optane SSD

What we focus in this paper.

Usages:

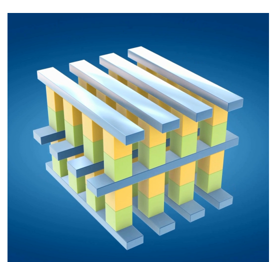
Interfaces: DIMM PCI–e SATA

Memory: DRAM NVM (3D XPoint) NAND Flash HDD

Byte-addressable

Byte-addressable & non-volatile

Non-volatile



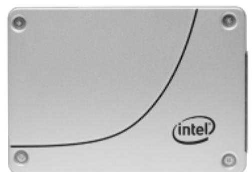
3D XPoint
Non-volatile memory (NVM)

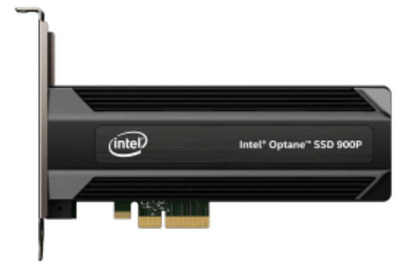
Comparison Method

NAND Flash SATA SSD
(Intel S3510)

3D XPoint Optane SSD
(Intel Optane 900P)

Non-Virtualized
Environment
(Linux Host)





Virtualized
Environment
(QEMU VM)

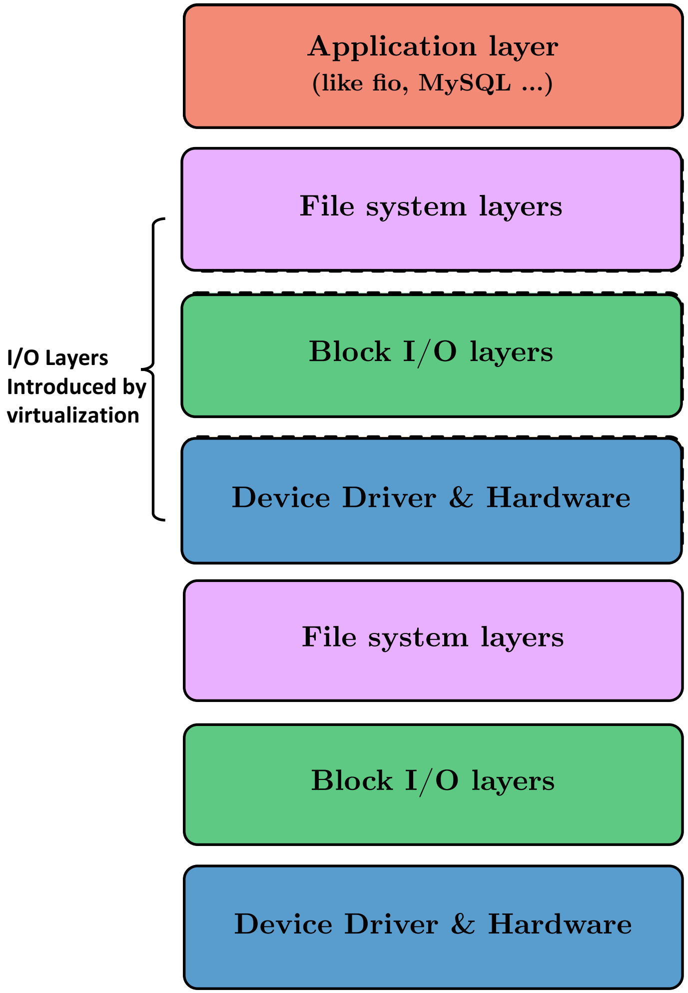
VM

VM

Agenda

- Impacts of Storage Stacks
- Micro-benchmarks
- Impacts on Storage Systems
- Tests in Database (MySQL)

Storage Stack is Complex



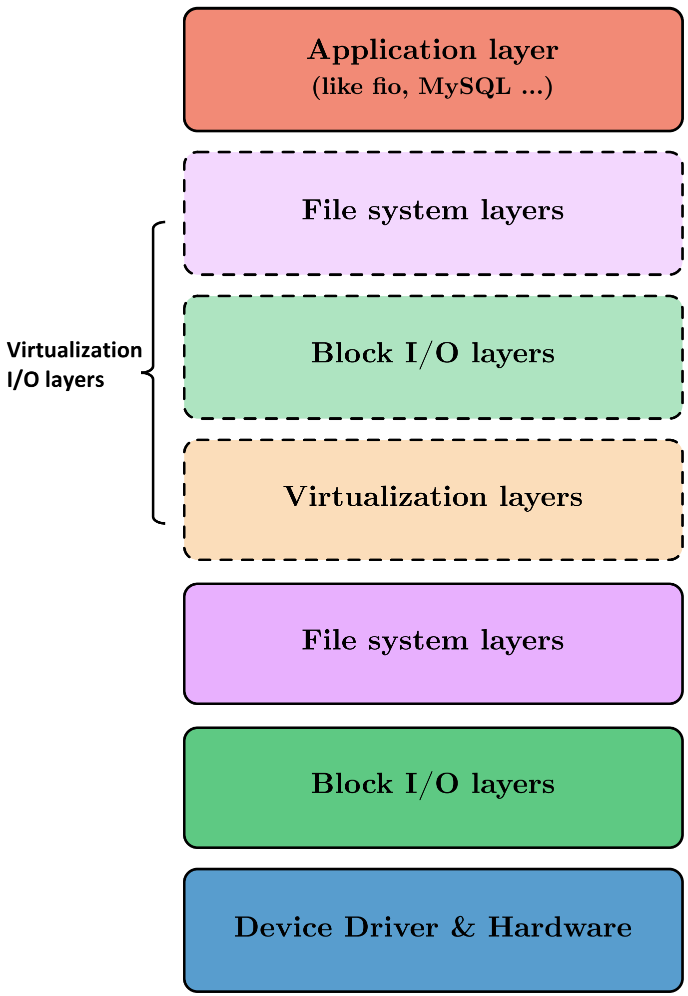
Operating system's storage stack is complex.

- I/O requests will go through application, file system layers, block I/O layers, device driver and hardware.

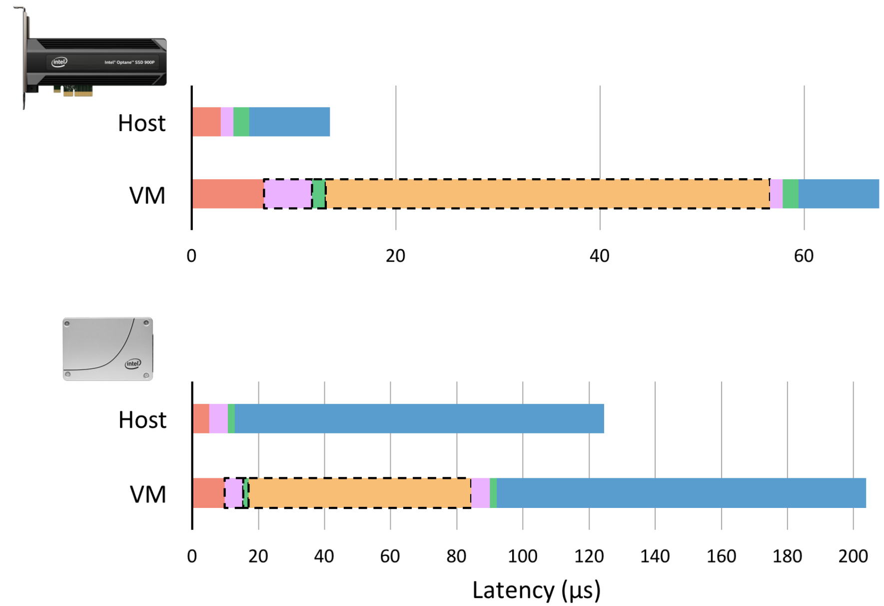
I/O path in virtualized environments is doubled.

- Virtual machine hypervisors (like QEMU) introduce many I/O virtualization layers.
- Guest OS also introduces filesystem and block I/O layers.

Storage Stack is Complex



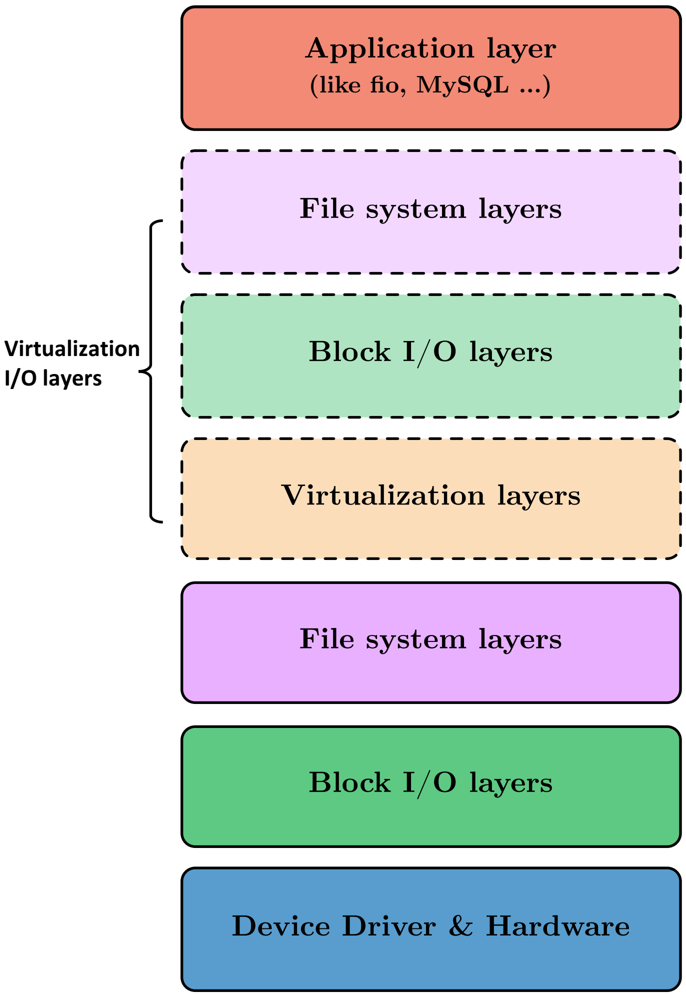
Latency breakdown: (Test env. : Fio 4K read, ext4, Linux, QEMU)

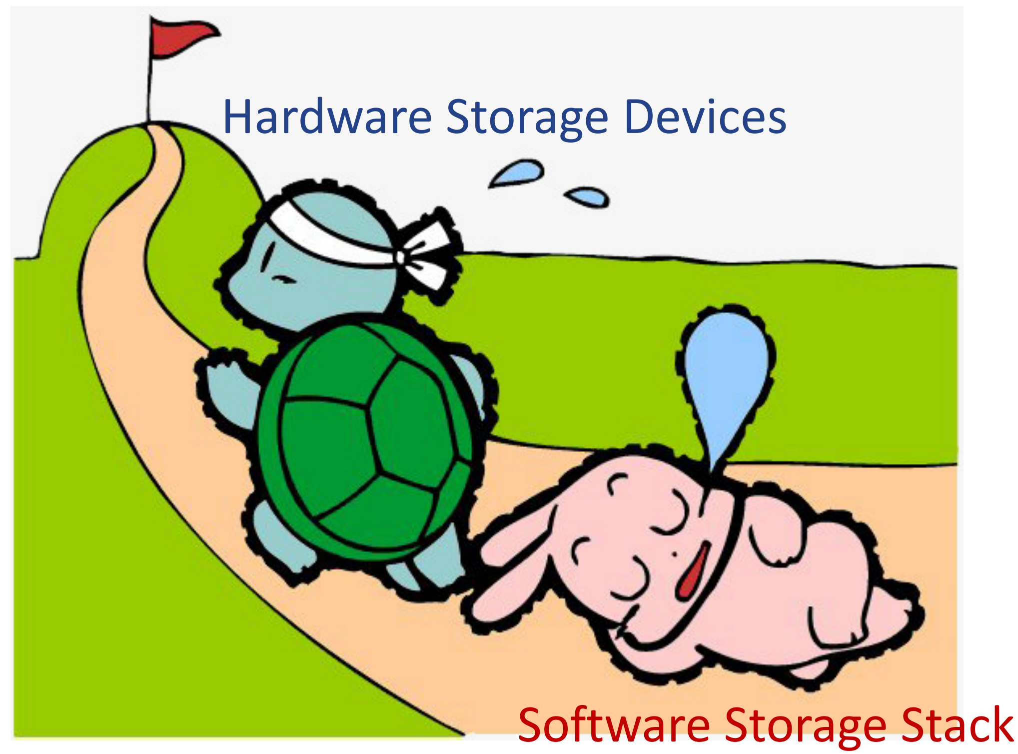


For Optane SSD:

- Hardware latency no longer dominate. (blue part)
- Overhead of virtualization layers is the largest. (dotted box)

Storage Stack is Complex



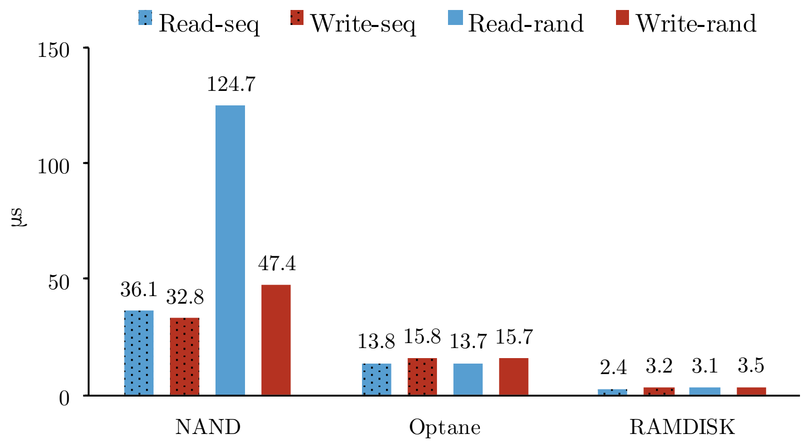


Stop sleeping, hardware is catching up!

Agenda

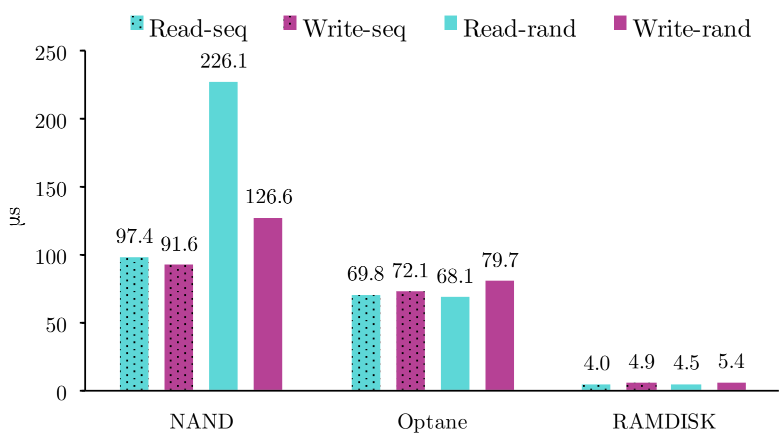
- Impacts of Storage Stacks
- Micro-benchmarks
  - Latency
  - Bandwidth
  - IOPS
- Impacts on Storage Systems
- Tests in Database (MySQL)

Micro-benchmarks --- Latency



Optane in host:

- Write is as fast as read.
- Random is as fast as sequential.

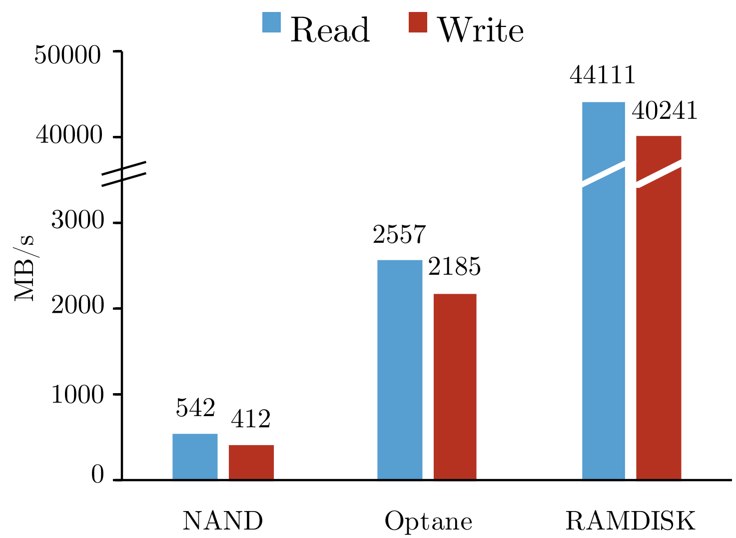


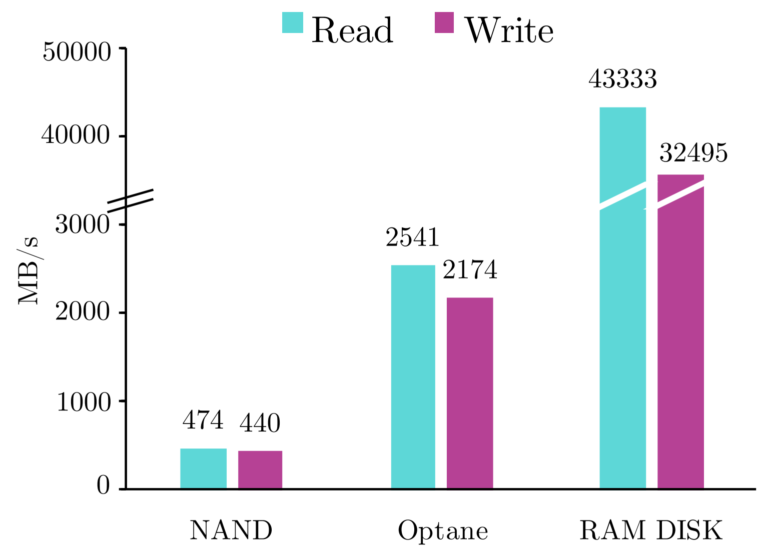
Optane in virtualized env.:

- Write is as fast as read.
- Random is as fast as sequential.
- Performance significantly drops.

Optane is better for latency-sensitive workload in non-virtualized environment.

Micro-benchmarks --- Bandwidth



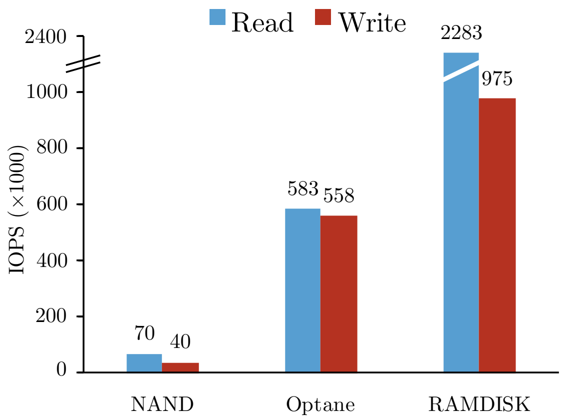


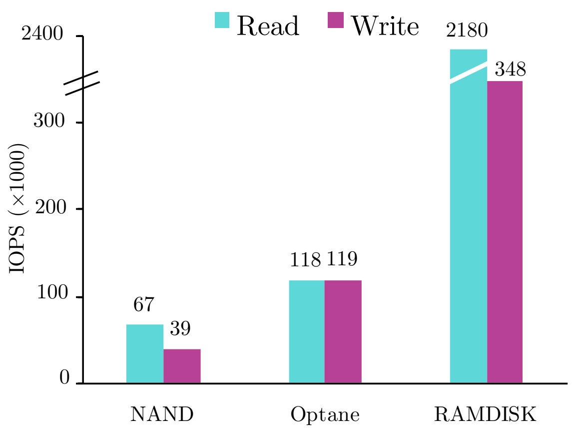
Optane's bandwidth is about 5 times better than NAND.

Virtualized environment's bandwidth performance is good.

Optane is better for high I/O off-line tasks in both environments.

Micro-benchmarks --- 4K IOPS



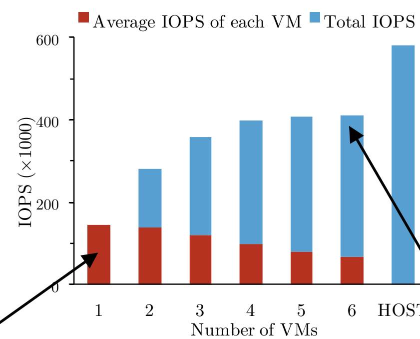


For Optane:

- No gap between read and write.
- Bad IOPS performance in virtualized env.

Optane is better for high concurrency workload in non-virtualized environment.

A way to use Optane in virtualized env. better: multiple VMs!

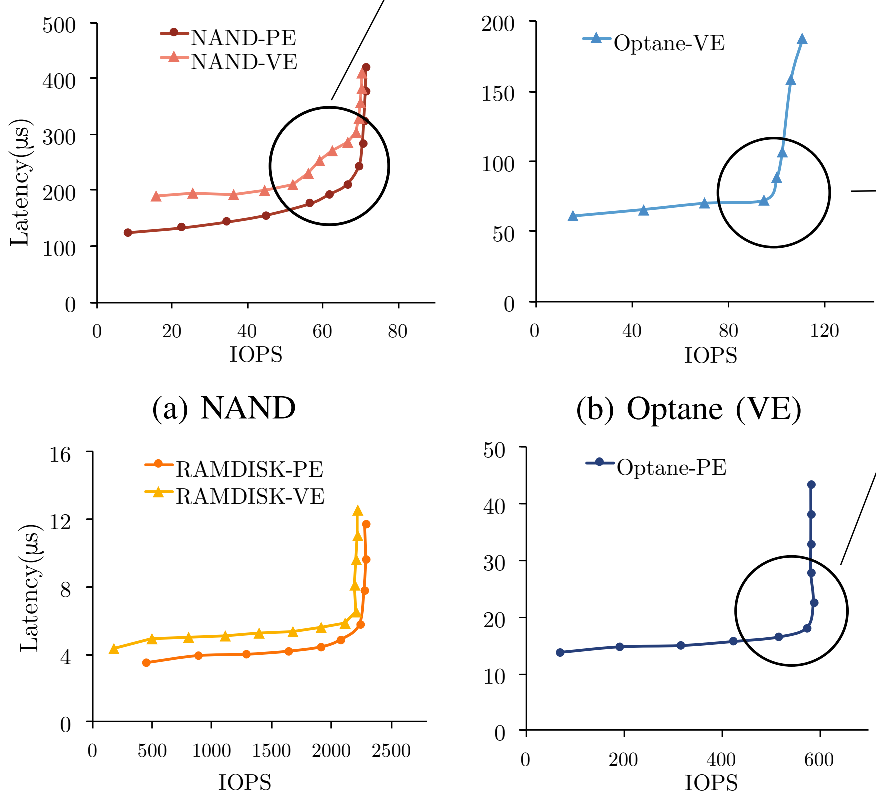


One VM is bad.

Multiple VMs is better.

Micro-benchmarks --- IOPS-latency Curve

NAND SSD's curves are flat.



(a) NAND

(b) Optane (VE)

(c) RAMDISK

(d) Optane (PE)

Optane's curve grows quickly when achieving the maximum IOPS.

When achieving 95% of maximum IOPSs, the latency increase:

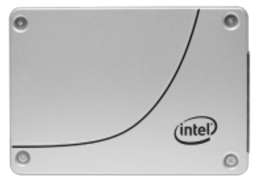
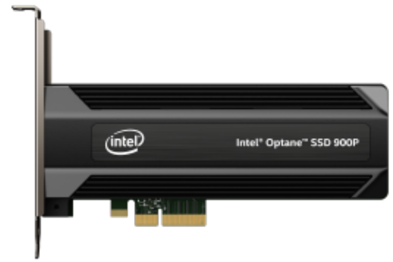
25% (for Optane),

54% (for RAMDISK),

80% (for SSD).

Optane is also better for high concurrency & latency-sensitive workload.

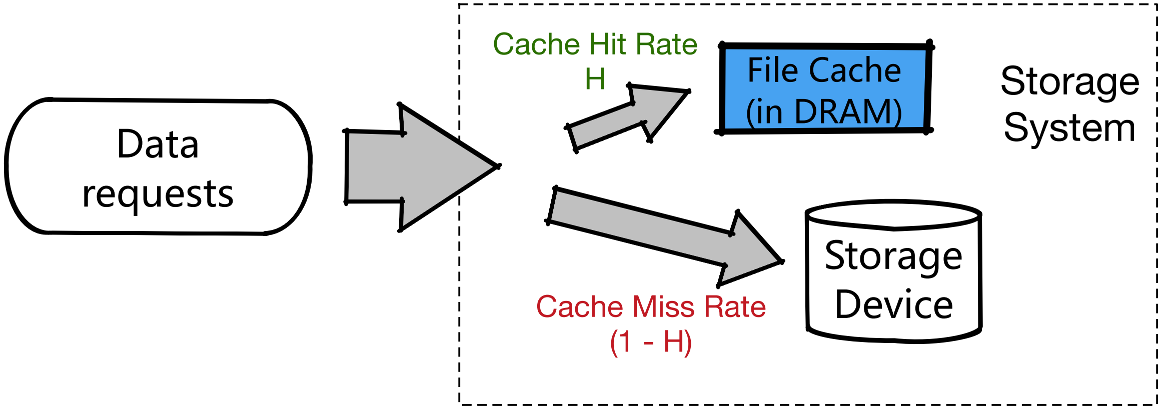
Comparison between Devices

	 NAND Flash SATA SSD (Intel S3510)	 3D XPoint Optane SSD (Intel Optane 900P)	 RAM DISK (Micron DDR4 emulated)
Latency	~50 us	~14 us	~3 us
Latency (VM)	~100 us	~70 us	~5 us
Bandwidth	~500 MB/s	~2500 MB/s	~40000 MB/s
Bandwidth (VM)	~450 MB/s	~2500 MB/s	~40000 MB/s
IOPS (4 KB)	~50k	~600k	~2000k
IOPS (4 KB) (VM)	~50k	~100k	~1000k
Dollars per GB	0.625	1.25	8

Agenda

- Impacts of Storage Stacks
- Micro-benchmarks
- Impacts on Storage Systems
  - File Cache
  - I/O Granularity
  - Data Compression
- Tests in Database (MySQL)

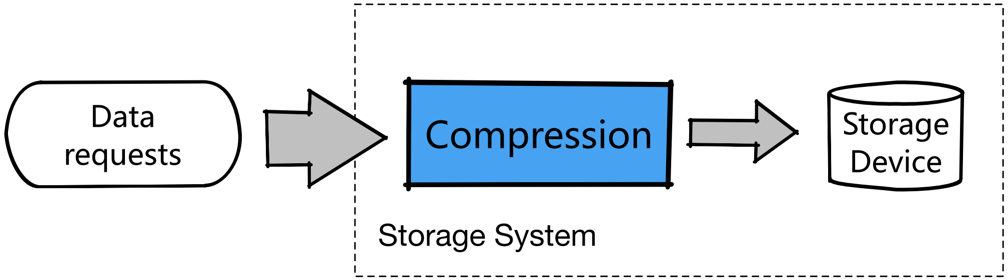
Impacts on Storage System --- File Cache



$$\text{Latency} = t_{\text{I/O}} \times (1 - H) + t_{\text{load}} \times H$$

File I/O benefits less from DRAM cache when using Optane.

Impacts on Storage System --- Data Compression

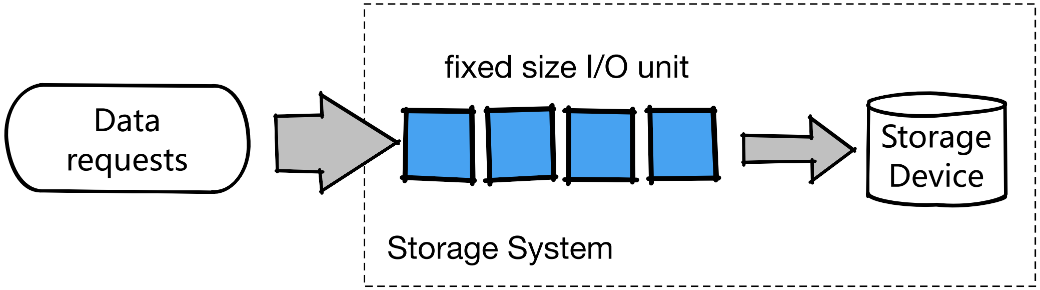


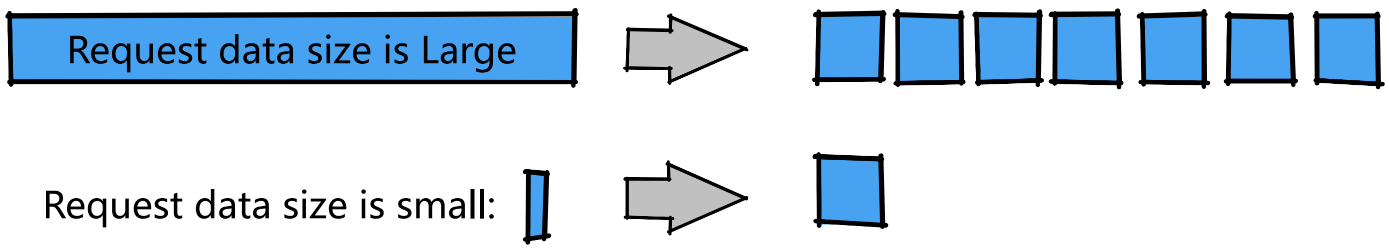
I/O Devices	Read (MB/s)	Write (MB/s)
NAND Flash SSD	542	412
Optane SSD	2557	2185

Algorithms	Decoding (MB/s)	Encoding (MB/s)
LZ4	2013	356
Snappy	915	269
Zlib defalte	133	23

Data compression will cause great performance degradation.

Impacts on Storage System --- I/O Granularity

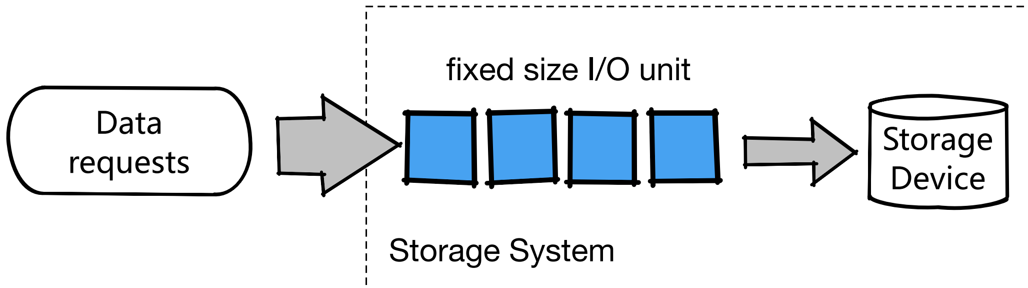




Common experience:

Faster devices benefit from smaller I/O granularity.

Impacts on Storage System --- I/O Granularity



Software	t_{app}	Application latency
	t_{stk}	Storage stack latency
Hardware	t_{seek}	Hardware I/O latency
	b	Hardware I/O bandwidth
	$\bar{d}$	Average range I/O size
	d_a	Best app. I/O Granularity
	d_s	OS I/O Granularity
	m	Point I/O access number
	n	Range I/O access number

$$d_a = \sqrt{\frac{n(t_{\text{app}} + t_{\text{seek}})\bar{d}}{m(t_{\text{stk}}/d_s + 1/b)}}.$$

- For slow devices, t_{seek} and $1/b$ dominate the best choice of I/O granularity.
- For high speed Optane, t_{app} and t_{stk} matters more.

~~Faster devices benefit from smaller I/O granularity.~~

More analysis are needed to chooce the best I/O granularity.

Impacts on Storage System --- Suggestions

File I/O benefits less from DRAM cache when using Optane.

Data compression will cause great performance degradation.

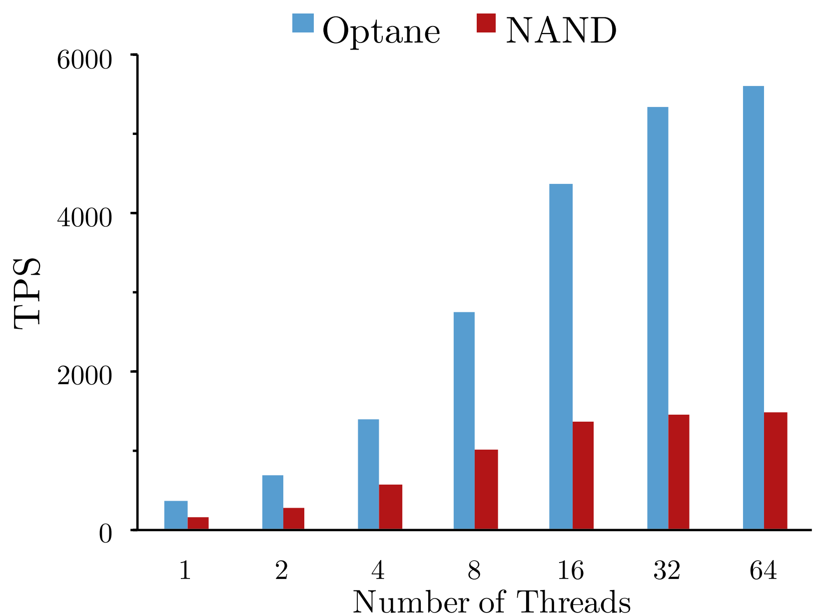
More analysis are needed to choose the best I/O granularity.

Agenda

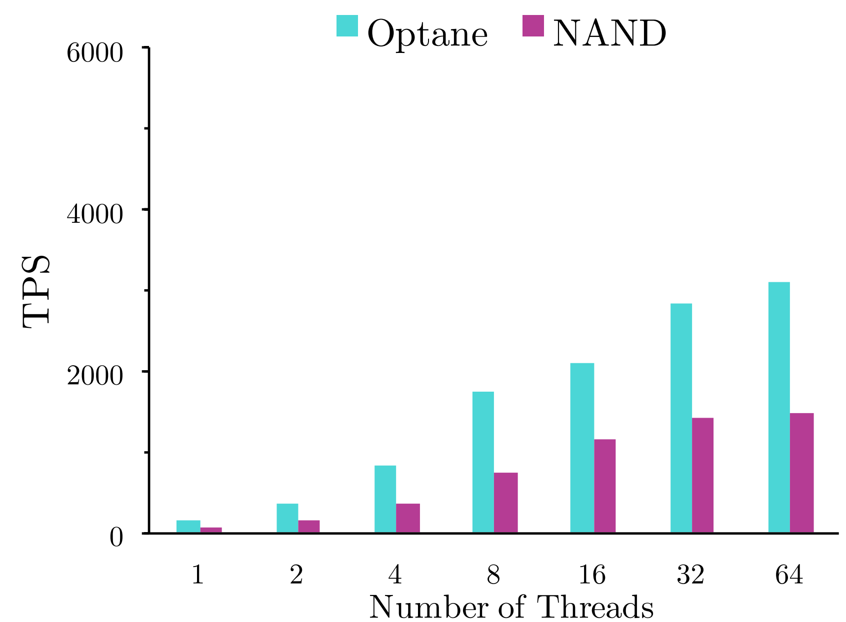
- Impacts of Storage Stacks
- Micro-benchmarks
- Impacts on Storage Systems
- Tests in Database (MySQL)
  - File Cache
  - I/O Granularity
  - Transparent Compression

Tests in Database (MySQL) --- Scalability

(Sysbench OLTP benchmark, Gaussian distribution, read)



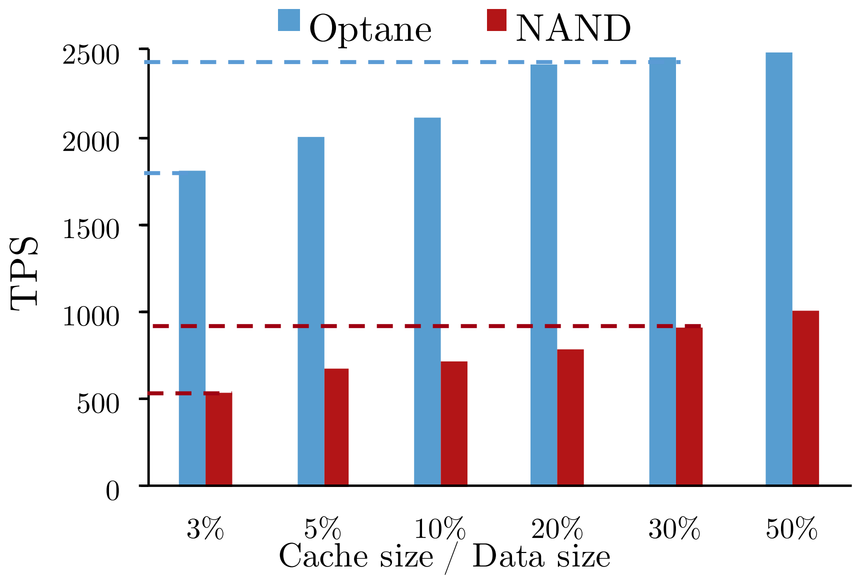
Host



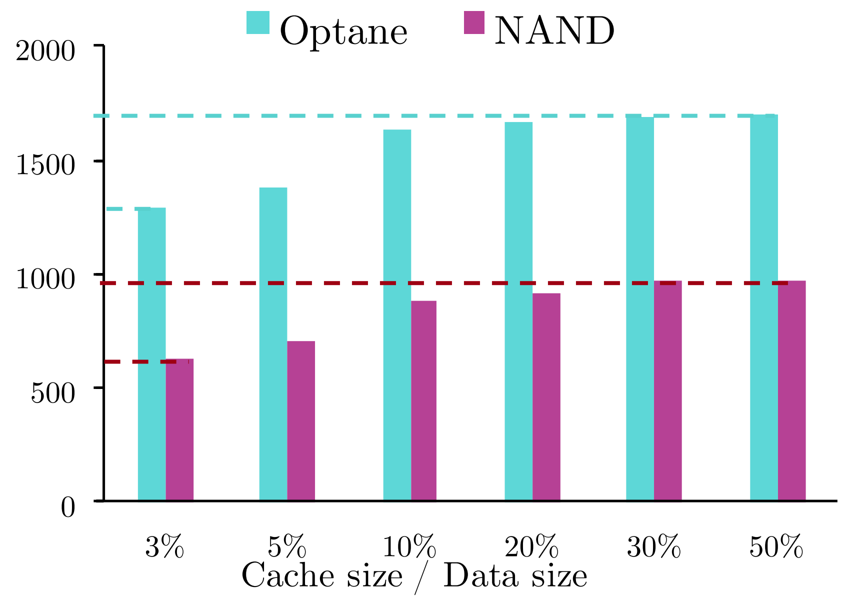
VM

Tests in Database (MySQL) --- File Cache

(Sysbench OLTP benchmark, Gaussian distribution, r/w:)



Host



VM

Cache - Data ratio 3% -> 50%
TPS improvement:

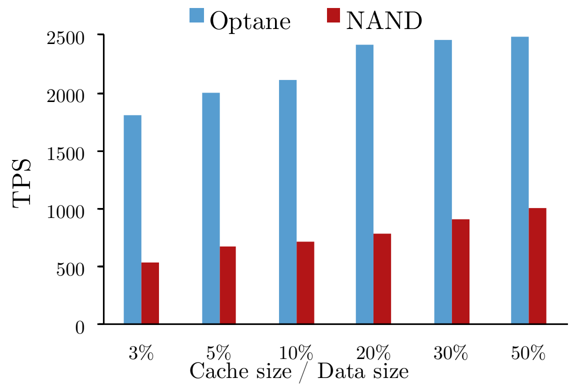
NAND : 90%
Optane : 40%

Cache - Data ratio 3% -> 50%
TPS improvement:

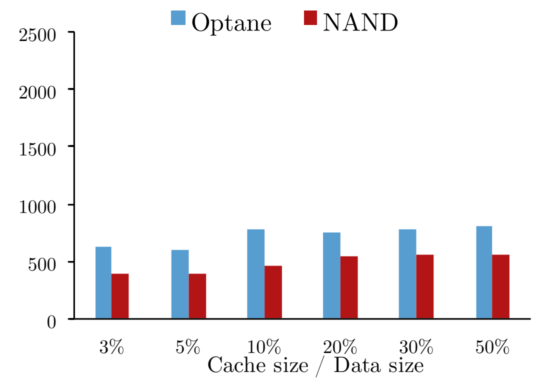
NAND : 40%
Optane : 30%

File I/O benefits less from DRAM cache when using Optane.

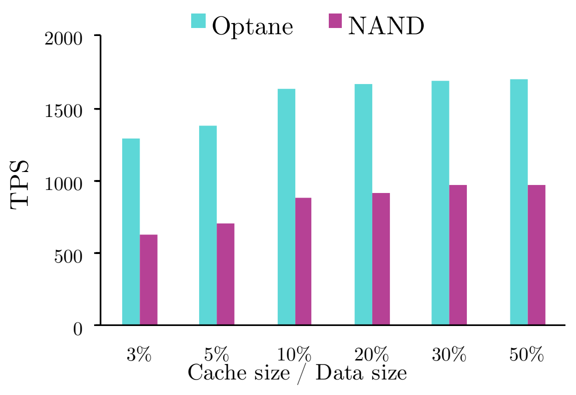
Tests in Database (MySQL) --- Compression



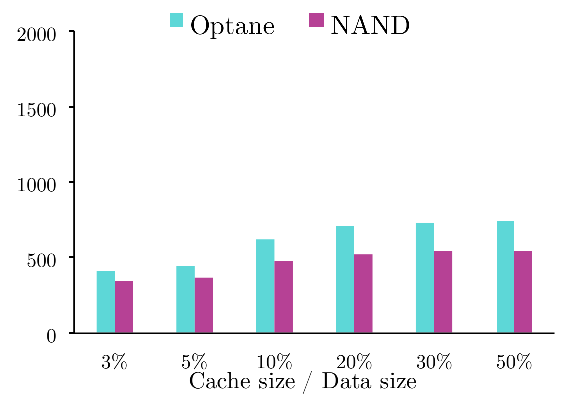
Host, compression disabled



Host, compression enabled



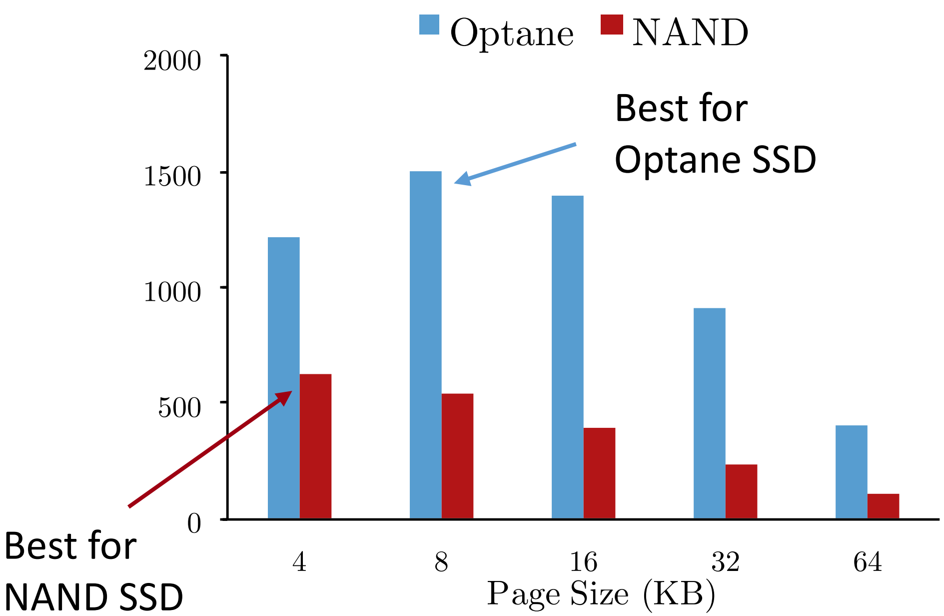
VM, compression disabled



VM, compression enabled

Data compression will cause great performance degradation.

Tests in Database (MySQL) --- I/O Granularity



Device	Read	Mixed R&W	Write
Optane	16	8	8
Optane (VE)	16	4	4
NAND	8	4	4
NAND (VE)	8	8	4

Best page sizes

~~Faster devices benefit from smaller I/O granularity.~~

More analysis are needed to chooce the best I/O granularity.

Summary

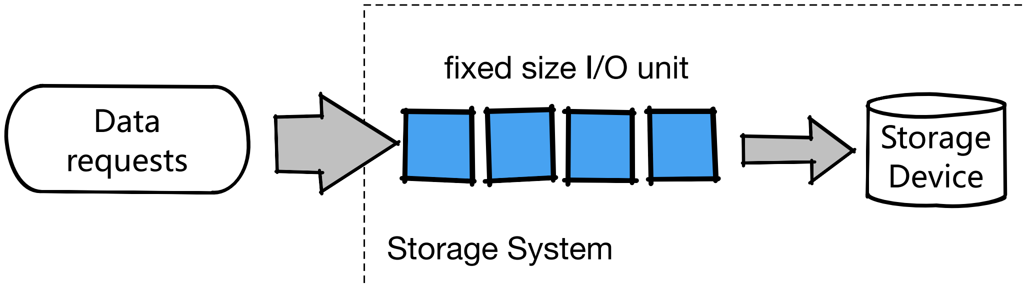
- We analysis the impacts of storage stacks on Optane's performance.
- We test the basic metrics of Optane and make comparisons with NAND SSDs.
- We analysis the impacts of Optane on the common storage systems.
- We give suggestions on storage system optimization and verified in MySQL.

Any questions?

Nankai - Baidu Joint Lab, Nankai University: http://nbjl.nankai.edu.cn

Thanks!

Impact on Storage System --- I/O Granularity



t_{app}	App. latency
t_{stk}	OS latency
t_{seek}	Hardware I/O latency
b	Hardware I/O bandwidth
$\bar{d}$	Average range I/O size
d_a	Best app. I/O Granularity
d_s	OS I/O Granularity
m	Point I/O access number
n	Range I/O access number

$$T = T_{\text{app}} + T_{\text{stk}} + T_{\text{dev}}.$$

$$\begin{aligned} T_D = & m(t_{\text{app}} + t_{\text{stk}}\frac{d_a}{d_s} + t_{\text{seek}} + \frac{d_a}{b}) \\ & + n(t_{\text{app}} + t_{\text{stk}}\frac{d_a}{d_s} + t_{\text{seek}} + \frac{d_a}{b})\frac{\bar{d}}{d_a} \end{aligned}$$

$$d_a = \sqrt{\frac{n(t_{\text{app}} + t_{\text{seek}})\bar{d}}{m(t_{\text{stk}}/d_s + 1/b)}}.$$