Gdevops

全球敏捷运维峰会

基于K8S的高性能RDS实践

演讲人：金戈

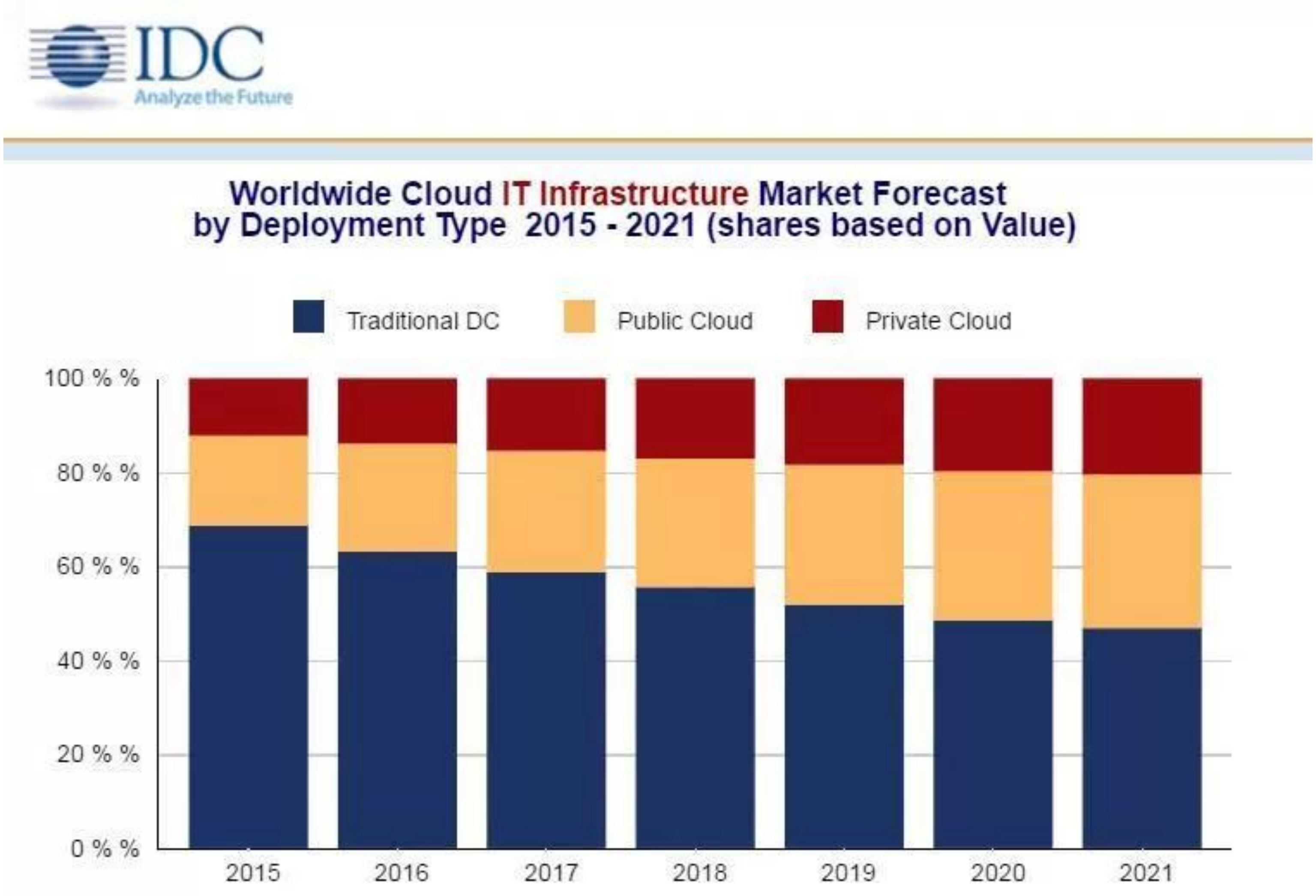
Agenda

The story of our RDS

The performance

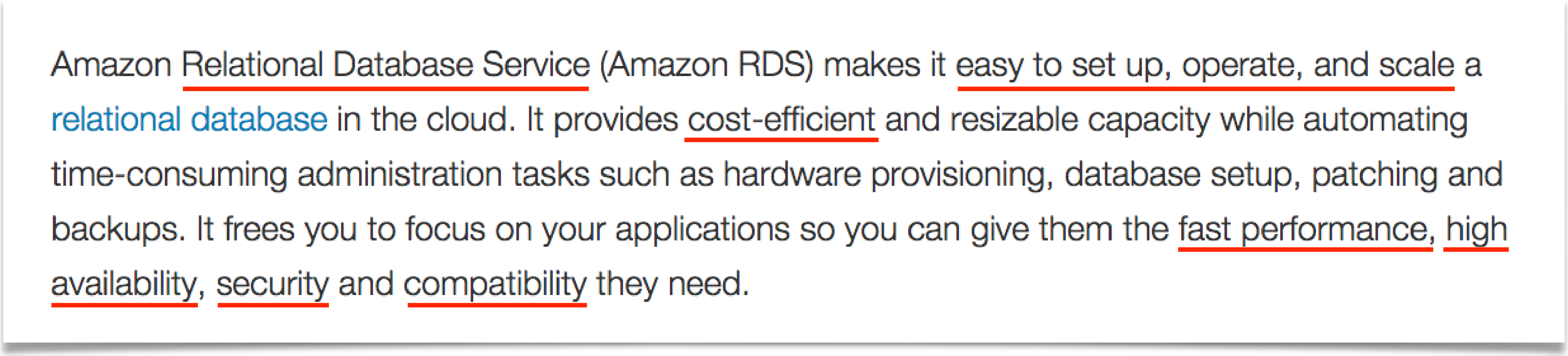
The service by k8s

Why RDS



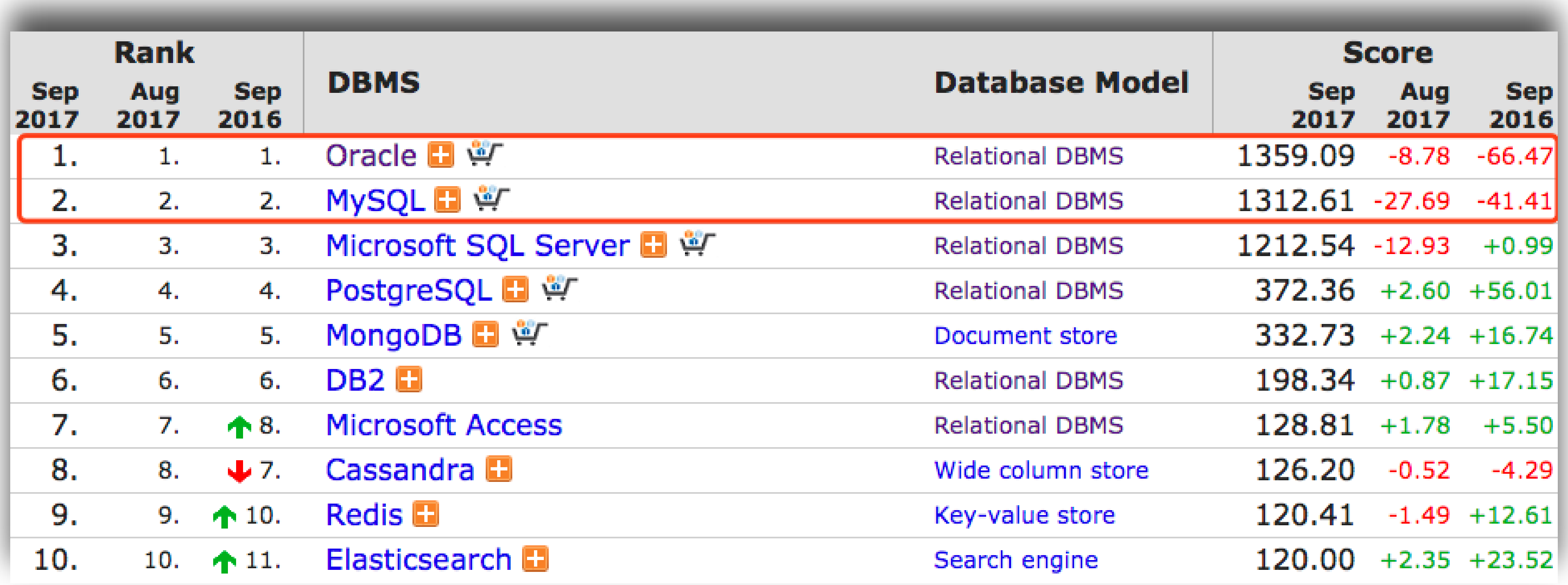
Source : Worldwide Quarterly Cloud IT Infrastructure Tracker, Q4 2016

Relational Database Service



Amazon Relational Database Service (Amazon RDS) makes it easy to set up, operate, and scale a relational database in the cloud. It provides cost-efficient and resizable capacity while automating time-consuming administration tasks such as hardware provisioning, database setup, patching and backups. It frees you to focus on your applications so you can give them the fast performance, high availability, security and compatibility they need.

Relational Database Service



Rank			DBMS	Database Model	Score		
Sep 2017	Aug 2017	Sep 2016			Sep 2017	Aug 2017	Sep 2016
1.	1.	1.	Oracle	Relational DBMS	1359.09	-8.78	-66.47
2.	2.	2.	MySQL	Relational DBMS	1312.61	-27.69	-41.41
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1212.54	-12.93	+0.99
4.	4.	4.	PostgreSQL	Relational DBMS	372.36	+2.60	+56.01
5.	5.	5.	MongoDB	Document store	332.73	+2.24	+16.74
6.	6.	6.	DB2	Relational DBMS	198.34	+0.87	+17.15
7.	7.	↑ 8.	Microsoft Access	Relational DBMS	128.81	+1.78	+5.50
8.	8.	↓ 7.	Cassandra	Wide column store	126.20	-0.52	-4.29
9.	9.	↑ 10.	Redis	Key-value store	120.41	-1.49	+12.61
10.	10.	↑ 11.	Elasticsearch	Search engine	120.00	+2.35	+23.52

Relational Database Service

- fast performance
- cost-efficient
- high availability
- easy to set up, operate and scale

fast performance

导致数据库性能问题 **应用、Schema、Index、SQL、执行计划、CPU、内存……**

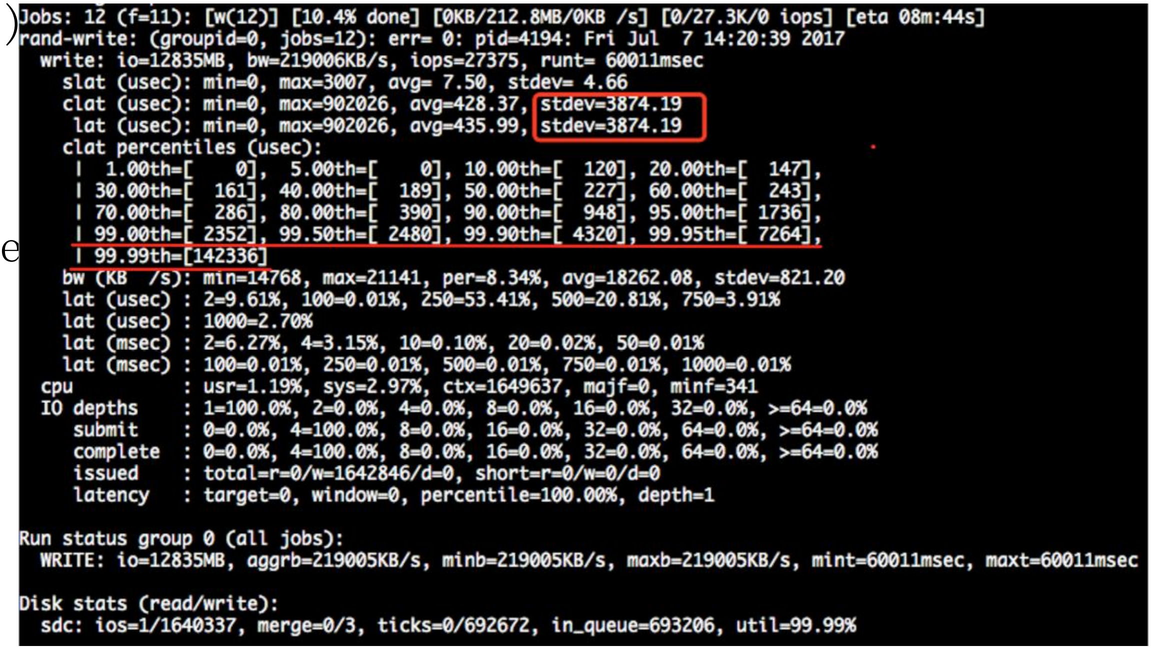
：

IO模型：（以online redo日志为例）

- WAL：Write-ahead logging
- Direct、sync、连续、512byte

对存储的要求：

- IOPS
- Latency：QoS、Jitter



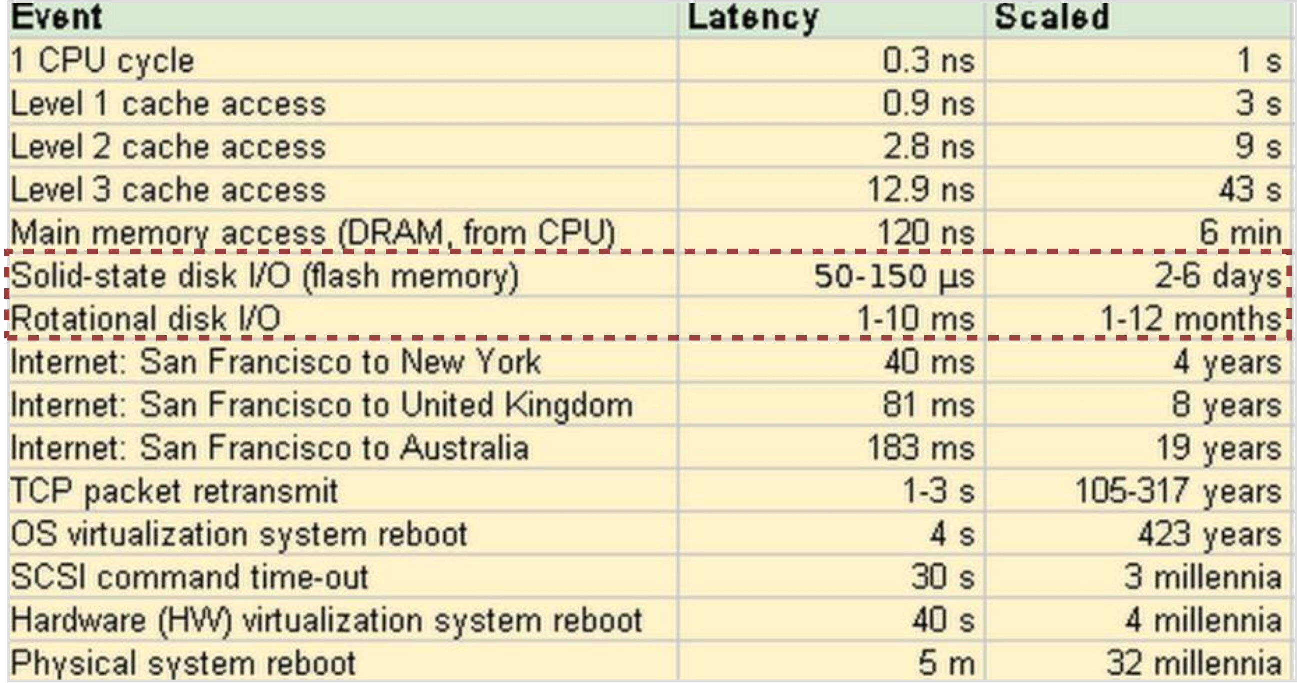
```
Jobs: 12 (f=11): [w(12)] [10.4% done] [0KB/212.8MB/0KB /s] [0/27.3K/0 iops] [eta 08m:44s]
rand-write: (groupid=0, jobs=12): err= 0: pid=4194: Fri Jul  7 14:20:39 2017
  write: io=12835MB, bw=219006KB/s, iops=27375, runt= 60011msec
    slat (usec): min=0, max=3007, avg= 7.50, stdev= 4.66
    clat (usec): min=0, max=902026, avg=428.37, stdev=3874.19
     lat (usec): min=0, max=902026, avg=435.99, stdev=3874.19
    clat percentiles (usec):
     |  1.00th=[    0],  5.00th=[    0], 10.00th=[  120], 20.00th=[  147],
     | 30.00th=[  161], 40.00th=[  189], 50.00th=[  227], 60.00th=[  243],
     | 70.00th=[  286], 80.00th=[  390], 90.00th=[  948], 95.00th=[ 1736],
     | 99.00th=[ 2352], 99.50th=[ 2480], 99.90th=[ 4320], 99.95th=[ 7264],
     | 99.99th=[142336]
    bw (KB  /s): min=14768, max=21141, per=8.34%, avg=18262.08, stdev=821.20
    lat (usec) : 2=9.61%, 100=0.01%, 250=53.41%, 500=20.81%, 750=3.91%
    lat (usec) : 1000=2.70%
    lat (msec) : 2=6.27%, 4=3.15%, 10=0.10%, 20=0.02%, 50=0.01%
    lat (msec) : 100=0.01%, 250=0.01%, 500=0.01%, 750=0.01%, 1000=0.01%
  cpu          : usr=1.19%, sys=2.97%, ctx=1649637, majf=0, minf=341
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     issued    : total=r=0/w=1642846/d=0, short=r=0/w=0/d=0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
  WRITE: io=12835MB, aggrb=219005KB/s, minb=219005KB/s, maxb=219005KB/s, mint=60011msec, maxt=60011msec

Disk stats (read/write):
  sdc: ios=1/1640337, merge=0/3, ticks=0/692672, in_queue=693206, util=99.99%
```

fast performance — 存储介质

- Principle of Locality
- Shaving x off latency at every layer in the stack

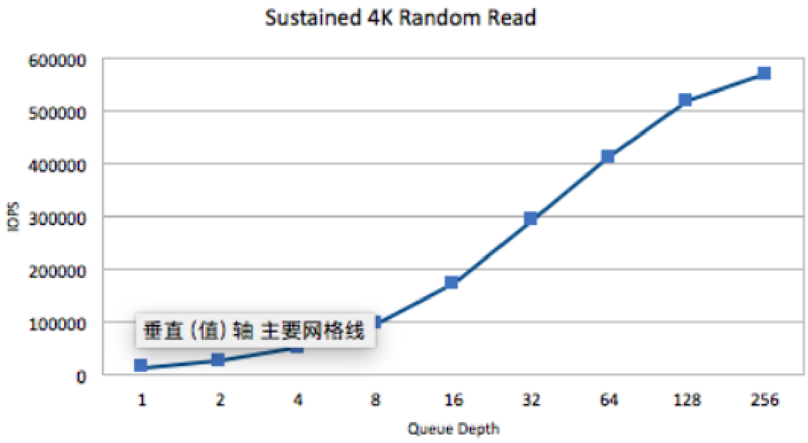


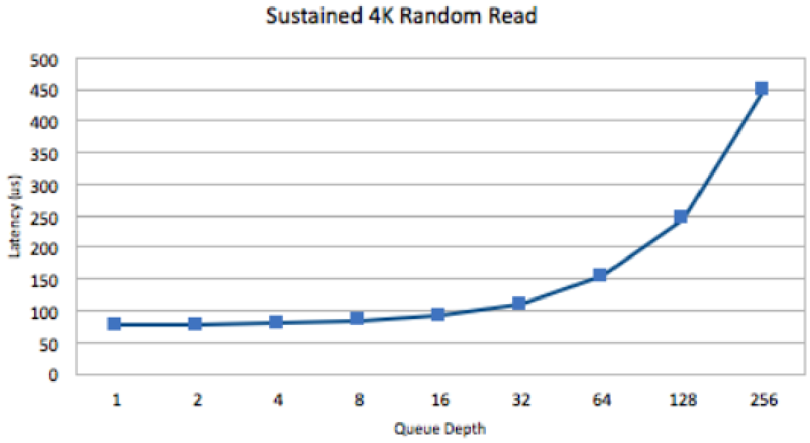
Event	Latency	Scaled
1 CPU cycle	0.3 ns	1 s
Level 1 cache access	0.9 ns	3 s
Level 2 cache access	2.8 ns	9 s
Level 3 cache access	12.9 ns	43 s
Main memory access (DRAM, from CPU)	120 ns	6 min
Solid-state disk I/O (flash memory)	50-150 μs	2-6 days
Rotational disk I/O	1-10 ms	1-12 months
Internet: San Francisco to New York	40 ms	4 years
Internet: San Francisco to United Kingdom	81 ms	8 years
Internet: San Francisco to Australia	183 ms	19 years
TCP packet retransmit	1-3 s	105-317 years
OS virtualization system reboot	4 s	423 years
SCSI command time-out	30 s	3 millennia
Hardware (HW) virtualization system reboot	40 s	4 millennia
Physical system reboot	5 m	32 millennia

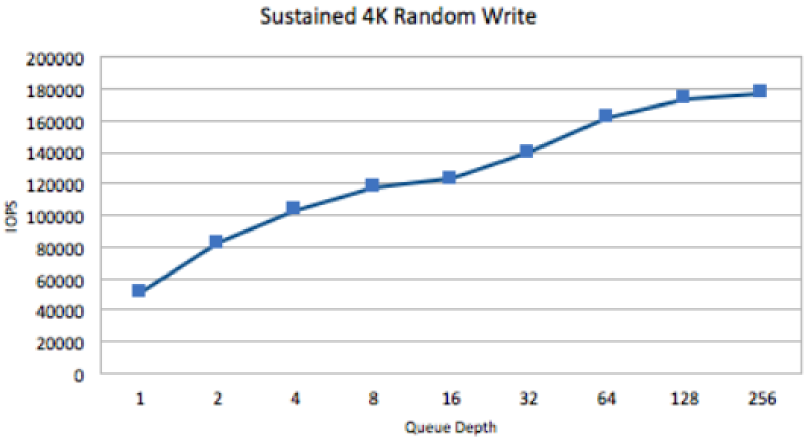
SSD 解救 DBA ？

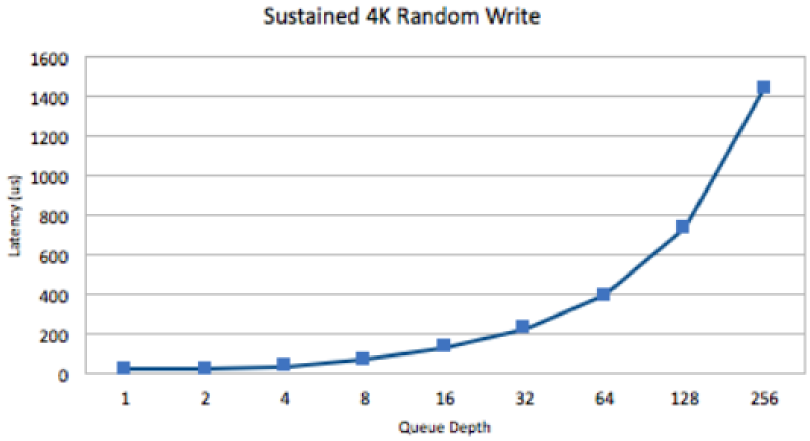
fast performance — 存储介质

NAND SSD / Flash 可以解决所有问题吗?


Sustained 4K Random Read

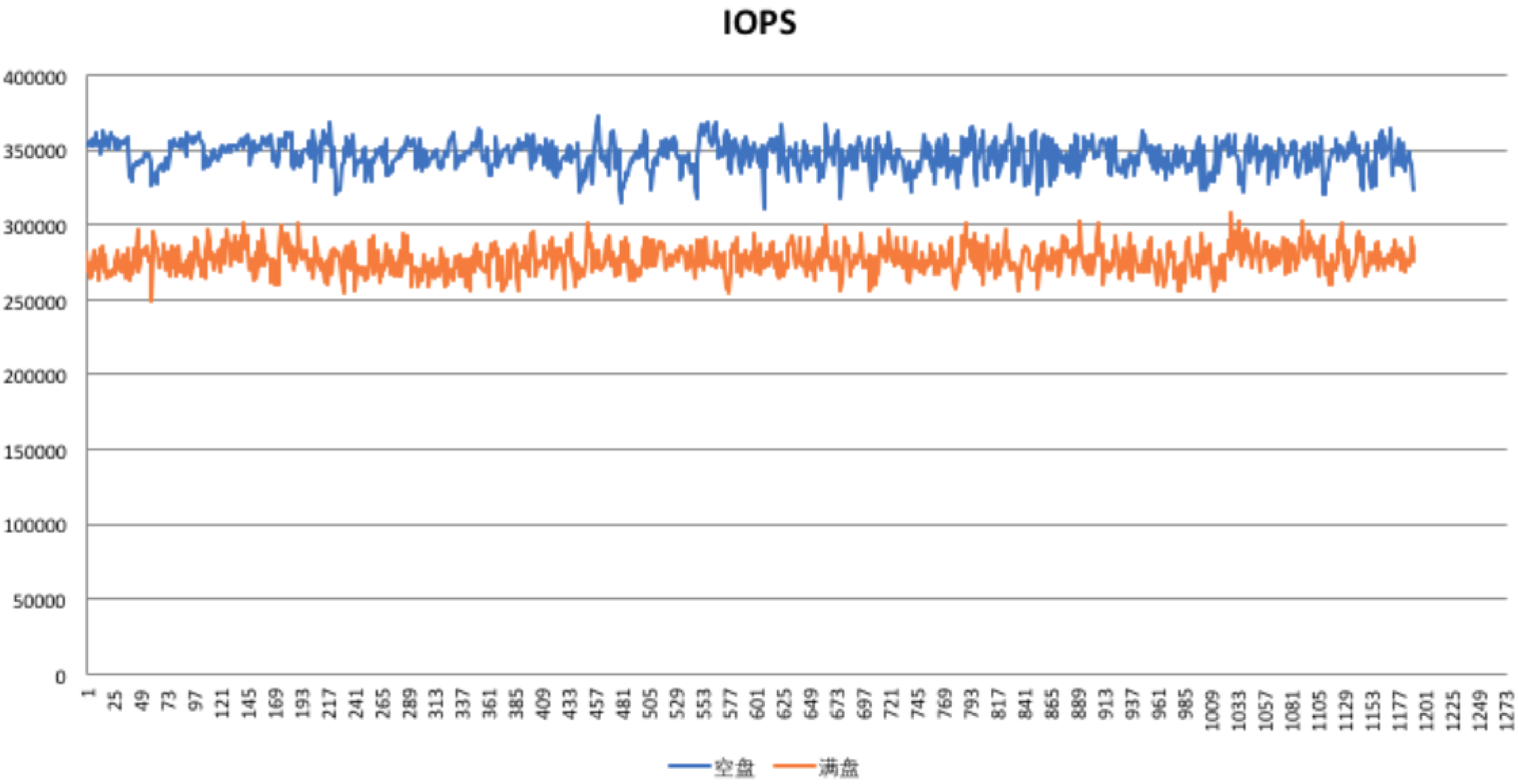

Sustained 4K Random Read


Sustained 4K Random Write


Sustained 4K Random Write

fast performance — 存储介质

NAND SSD / Flash 可以解决所有问题吗?



fast performance — 存储介质

NAND SSD / Flash 可以解决所有问题吗?

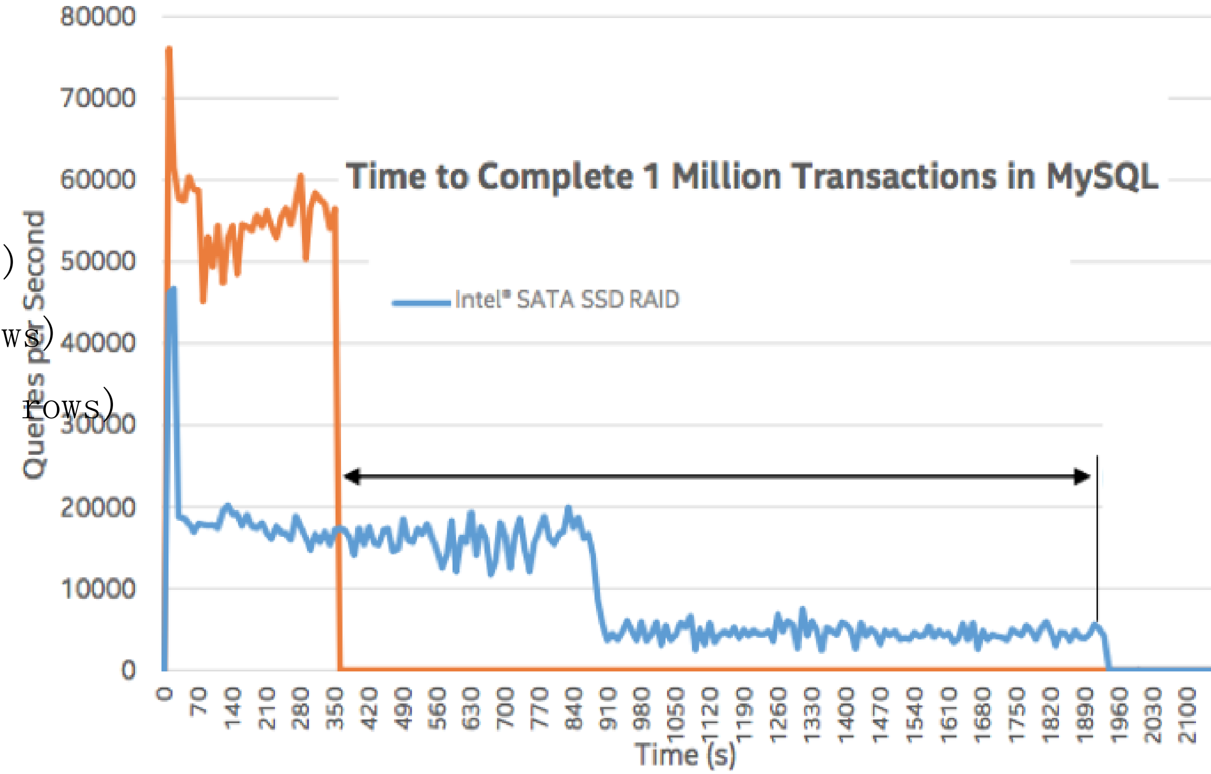
- Write amplification
- Garbage Collection
- IO Queue Depth
- 读/写
- 空盘/满盘
- 抖动

fast performance — 存储介质

关注蓝线

测试模型

- point selects (single row)
- range selects (multiple rows)
- sum range selects (multiple rows)
- order range selects (multiple rows)
- distinct range selects (multiple rows)
- row updates/deletions/insertions

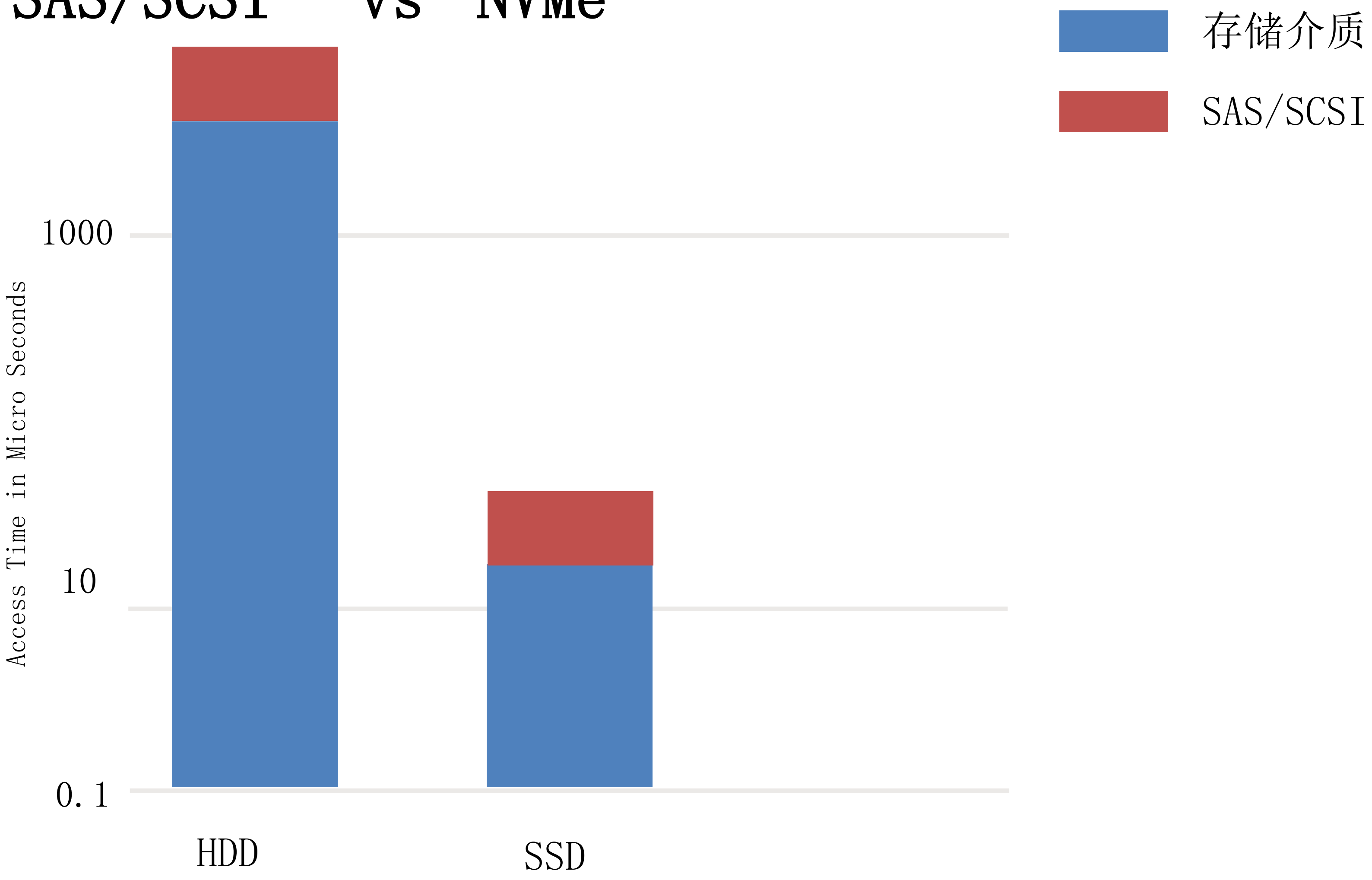


问题：

蓝线为什么有两次下降？

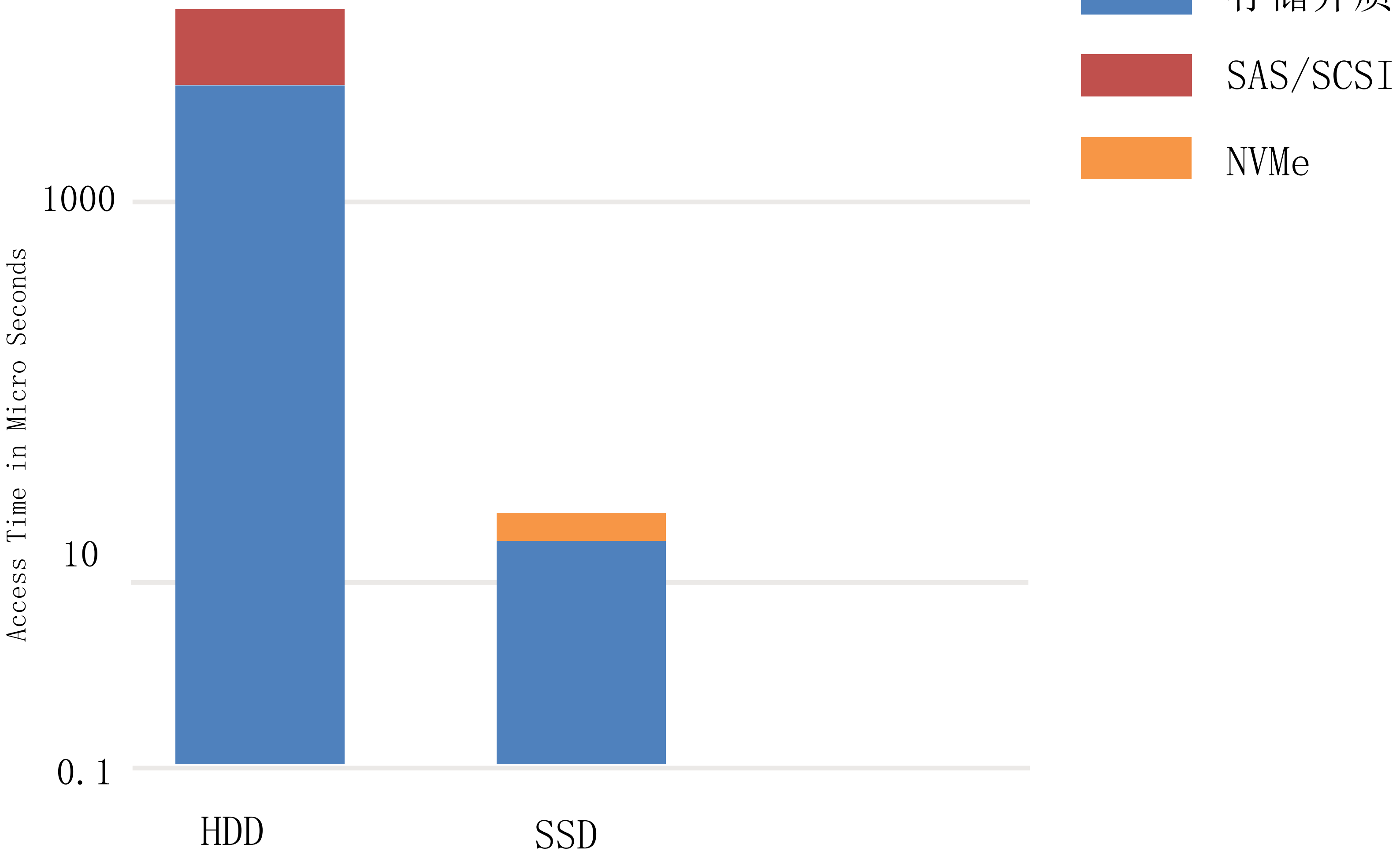
fast performance — 存储协议

SAS/SCSI vs NVMe



fast performance — 存储协议

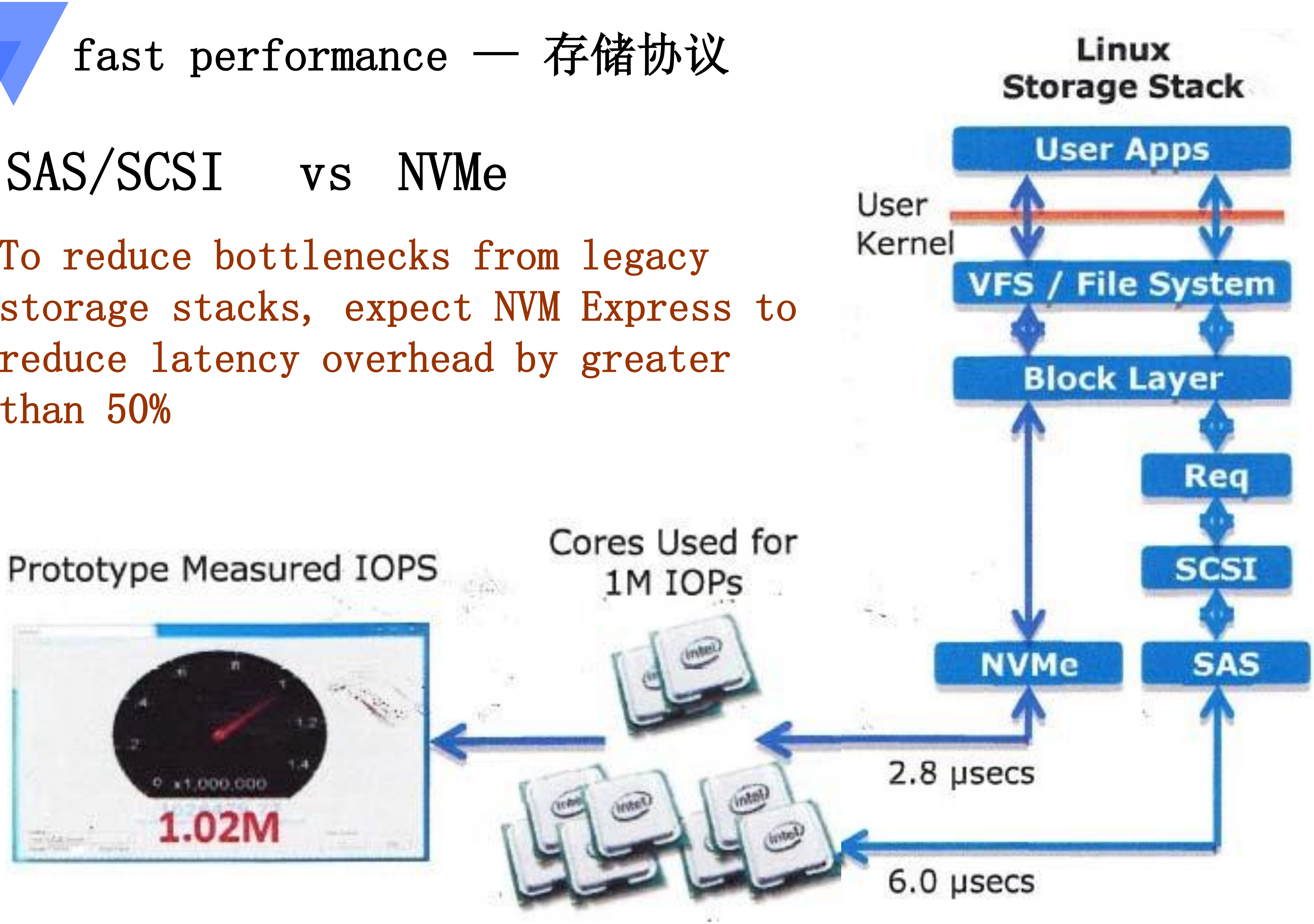
SAS/SCSI vs NVMe



fast performance — 存储协议

SAS/SCSI vs NVMe

To reduce bottlenecks from legacy storage stacks, expect NVM Express to reduce latency overhead by greater than 50%



Linux Storage Stack

User Apps

User

Kernel

VFS / File System

Block Layer

Req

SCSI

NVMe

SAS

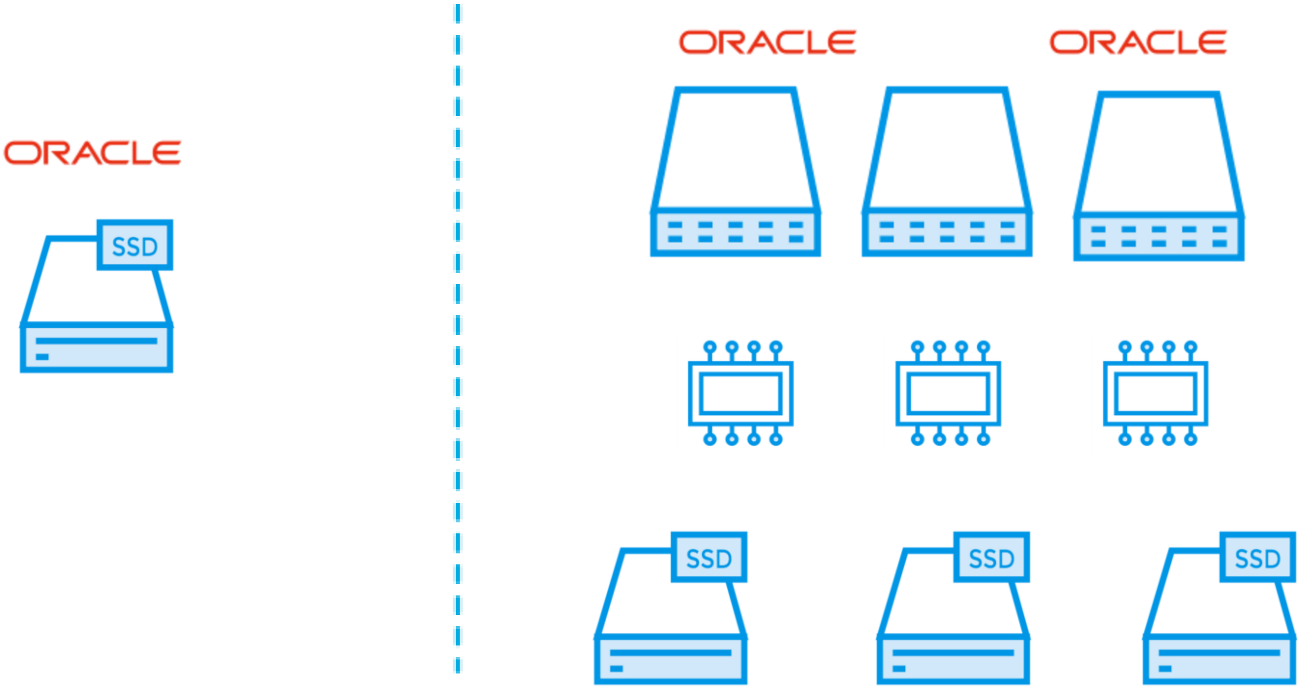
2.8 μsecs

6.0 μsecs

Cores Used for 1M IOPs

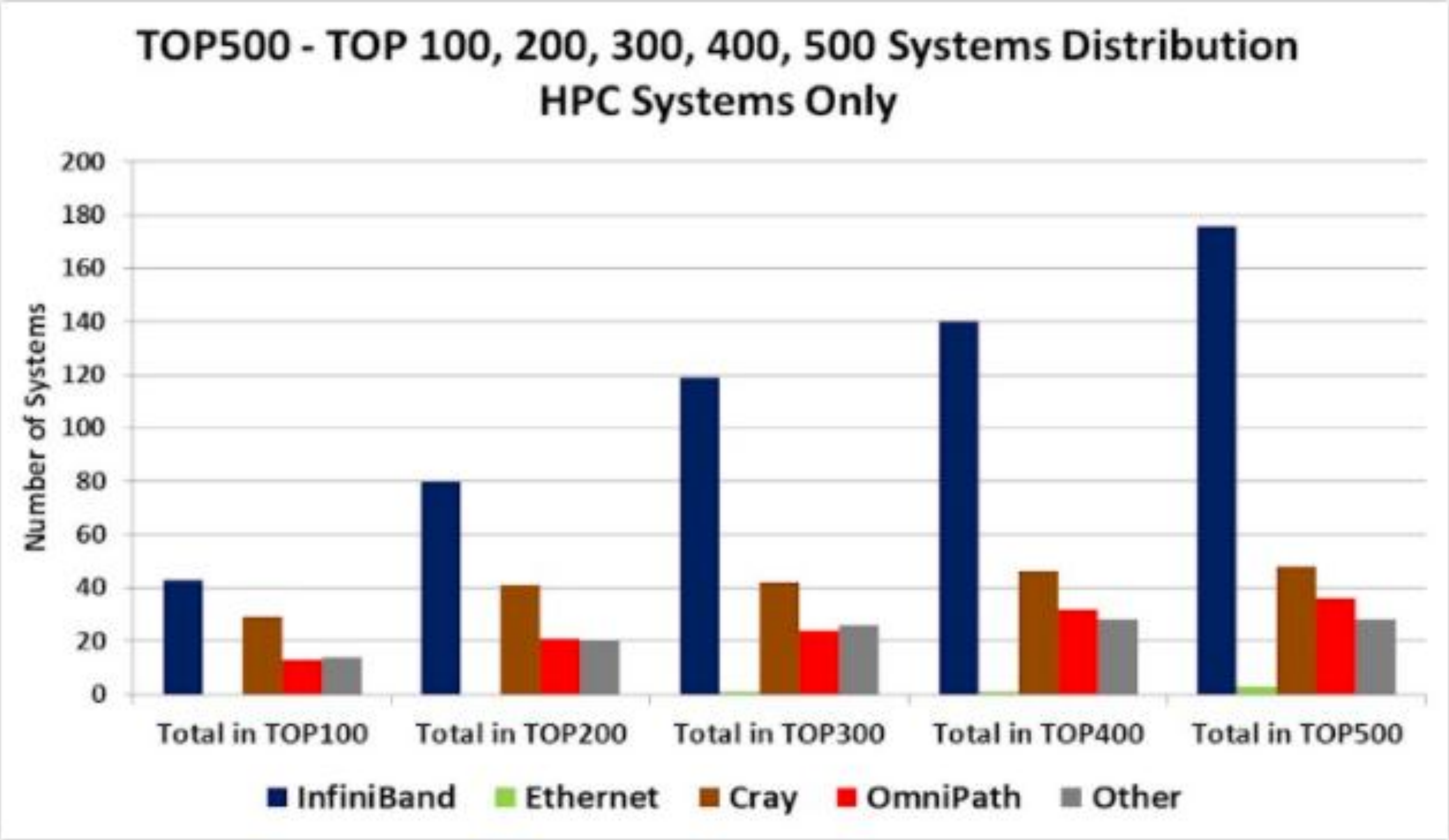
Prototype Measured IOPS

fast performance — 存储网络

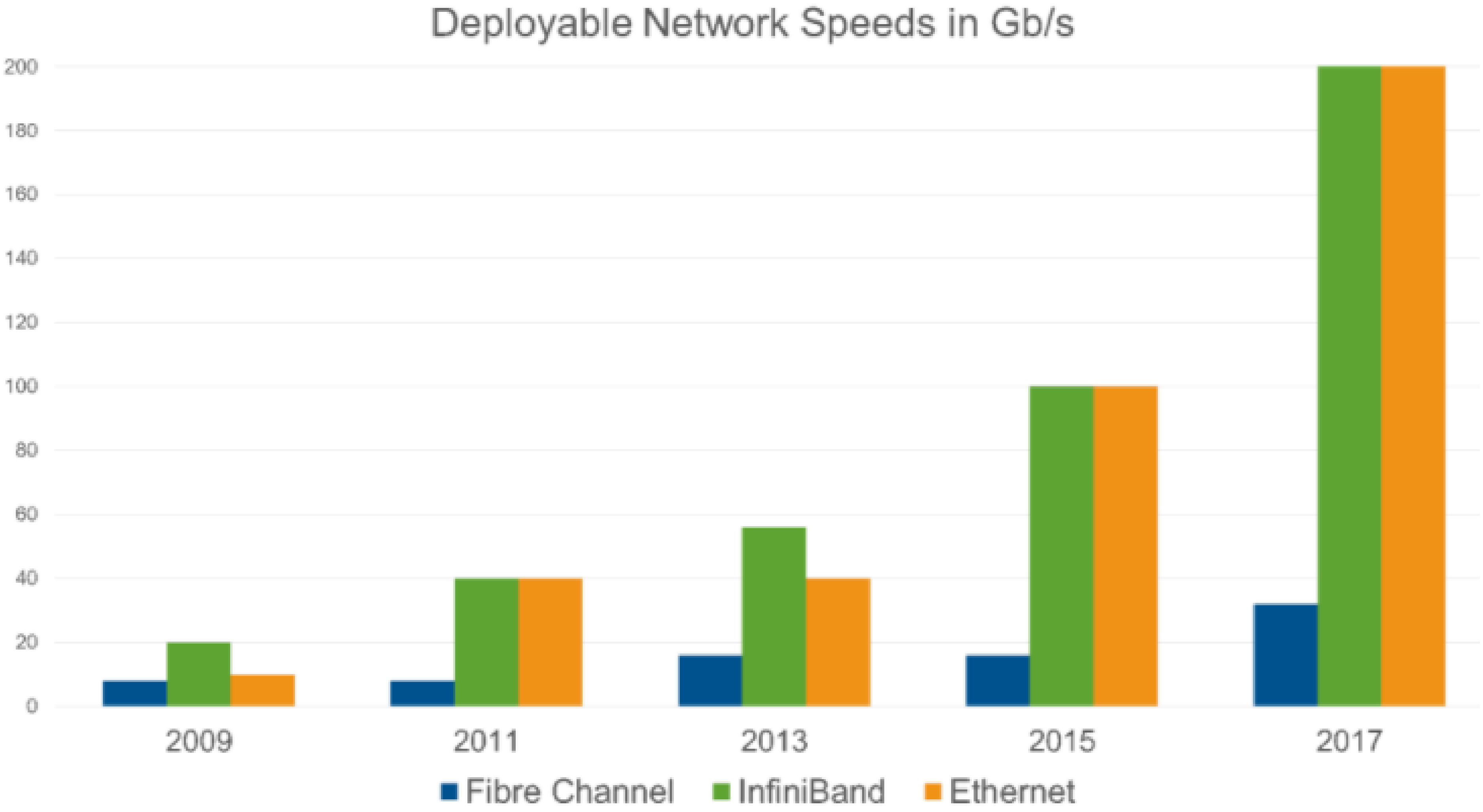


Faster Storage Needs a Faster Network

fast performance — 存储网络

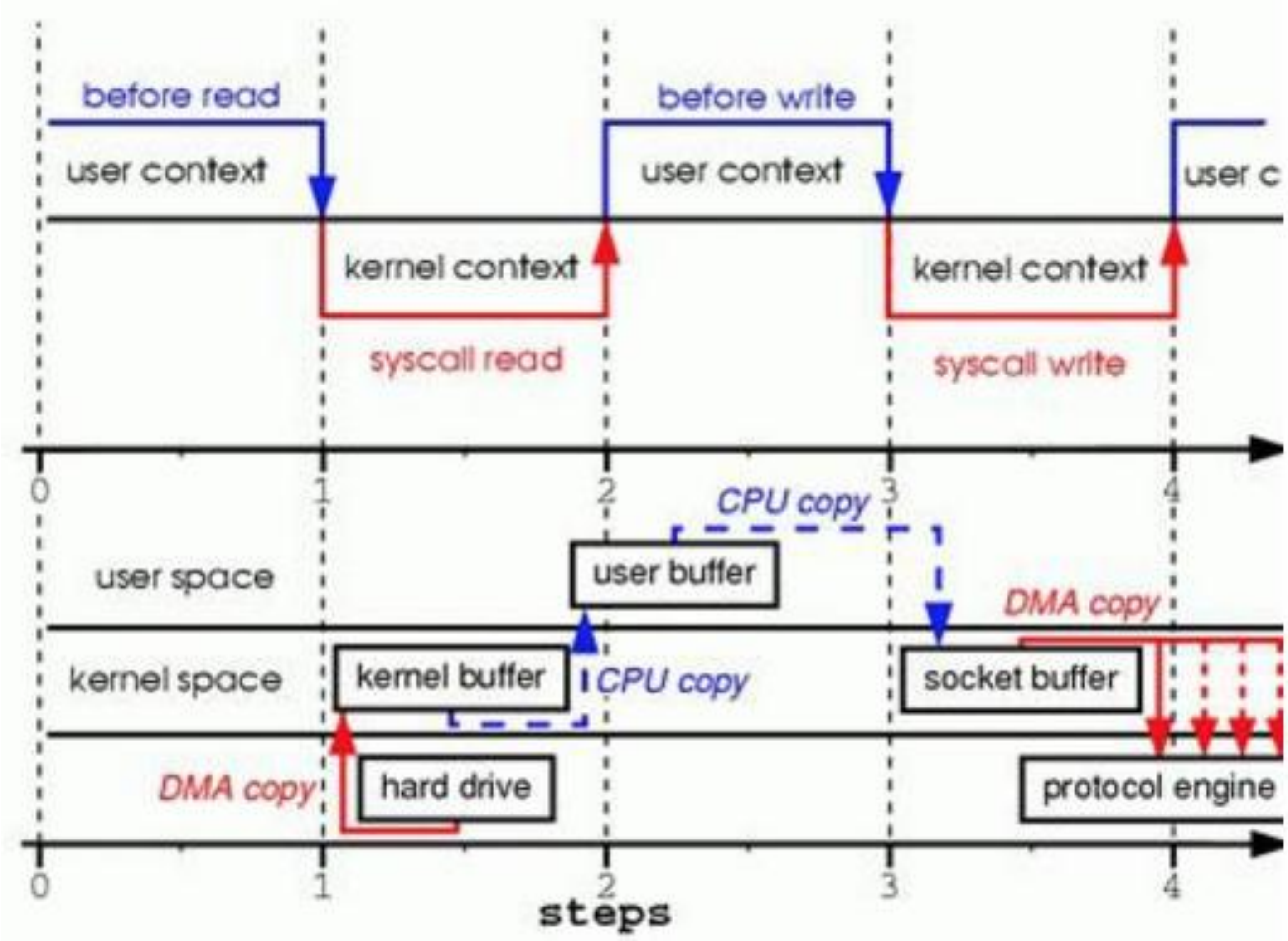


fast performance — 存储网络



fast performance — 存储网络

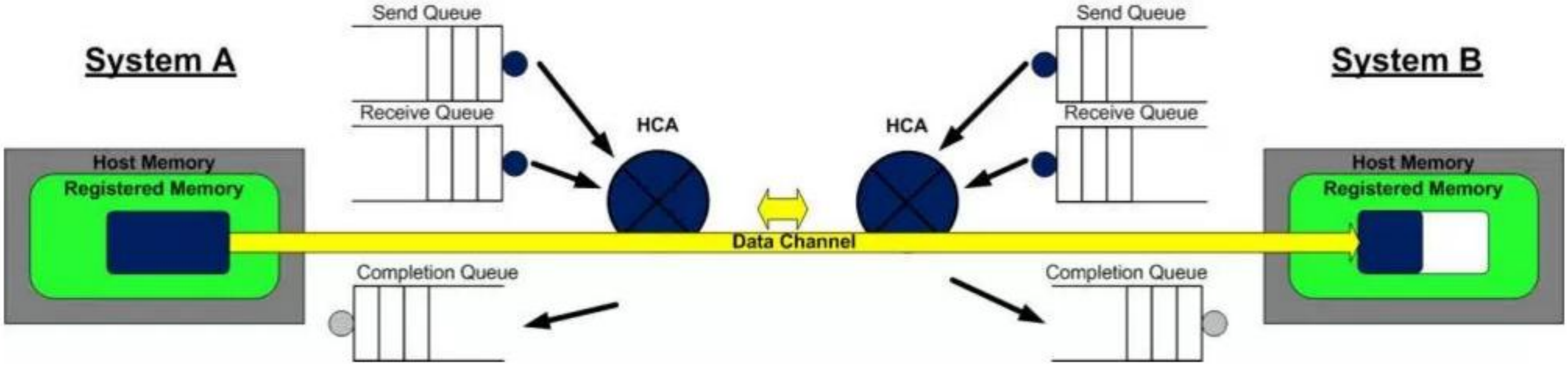
一个标准的数据传输操作



fast performance — 存储网络

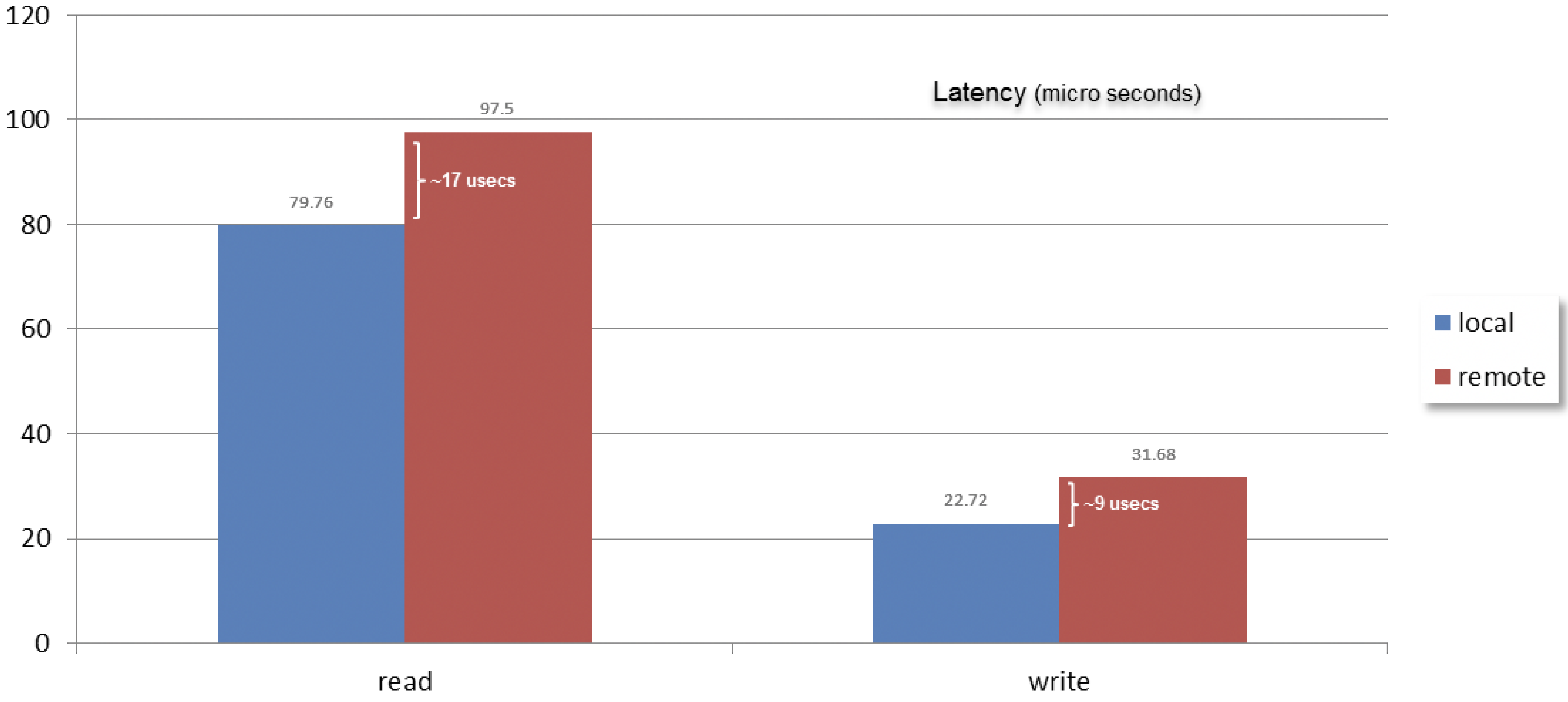
NVMf : allows the new high performance SSD interface, Non-Volatile Memory Express (NVMe), to be connected across RDMA-capable networks.

- Zero-Copy
- Kernel bypass
- No CPU involvement



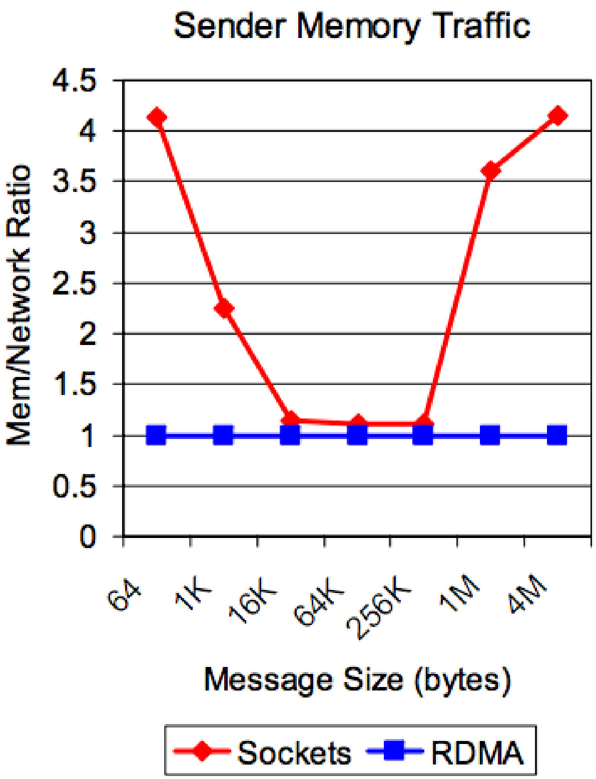
fast performance — 存储网络

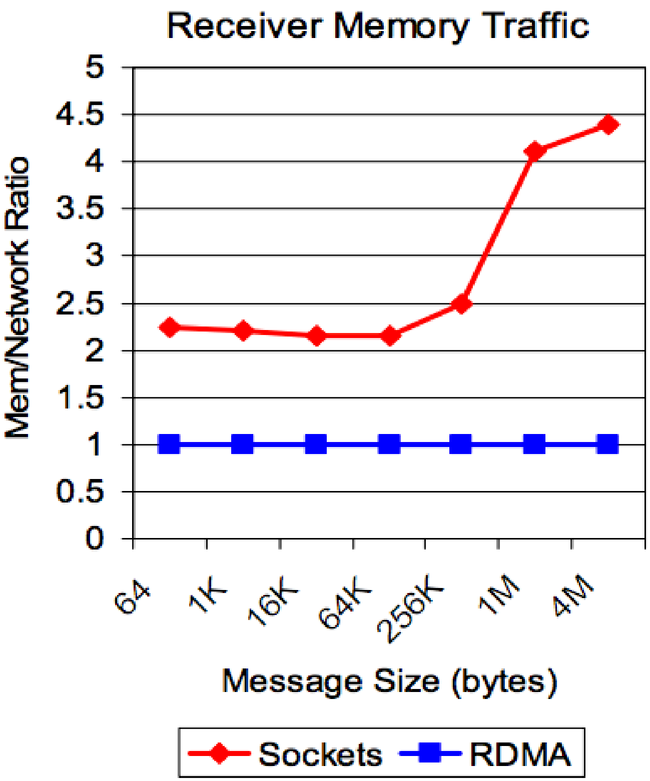
NVMf : allows the new high performance SSD interface, Non-Volatile Memory Express (NVMe), to be connected across RDMA-capable networks.



fast performance — 存储网络

NVMf : allows the new high performance SSD interface, Non-Volatile Memory Express (NVMe), to be connected across RDMA-capable networks.



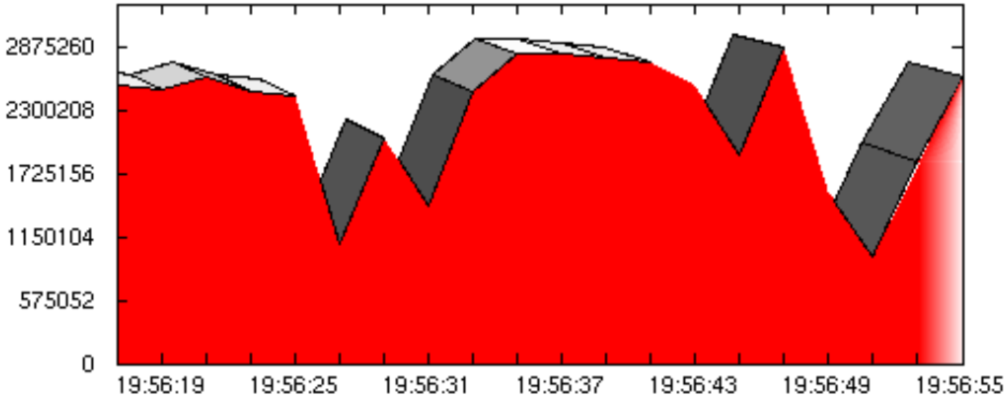


- Sockets can force up to 4 times more memory traffic compared to the network traffic
- RDMA allows has a ratio of 1 !!

fast performance — NVMf

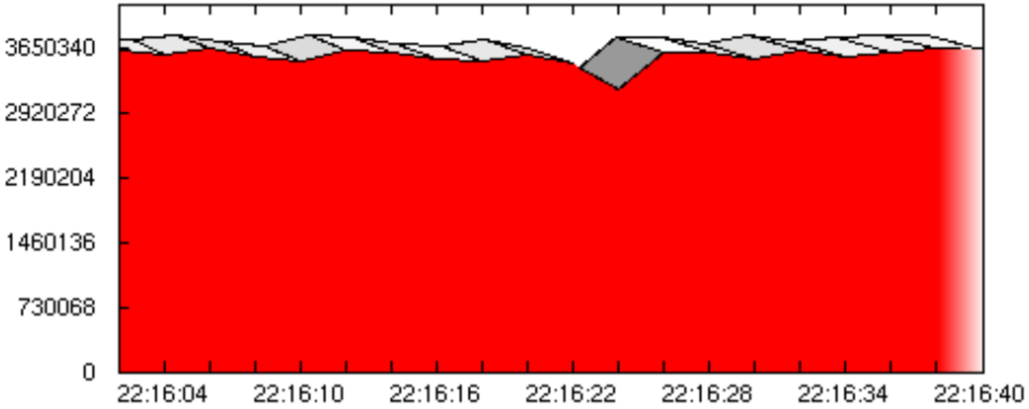
- **iSER** (iSCSI Extensions for RDMA)

  iSCSI + RDMA + Infiniband



- **NVMf** (NVMe Over Fabric)

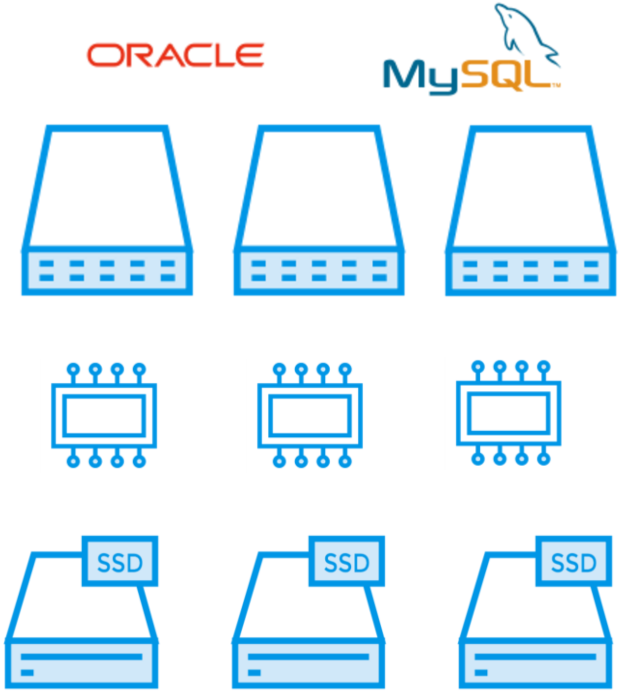
  NVMe + RDMA + Infiniband



fast performance

分布式存储

- Better utilization
  - capacity
  - rack
  - space
  - power
- Better scalability
- Management
- Fault isolation



fast performance — 分布式存储

易用

- 支持容量透明的 scale up/out

数据安全

- 支持多种冗余模式：mirror， raid

易维护

- 完善的 FA 机制
- Online rebuild / Online increment rebuild
- 可控制的 rebuild power

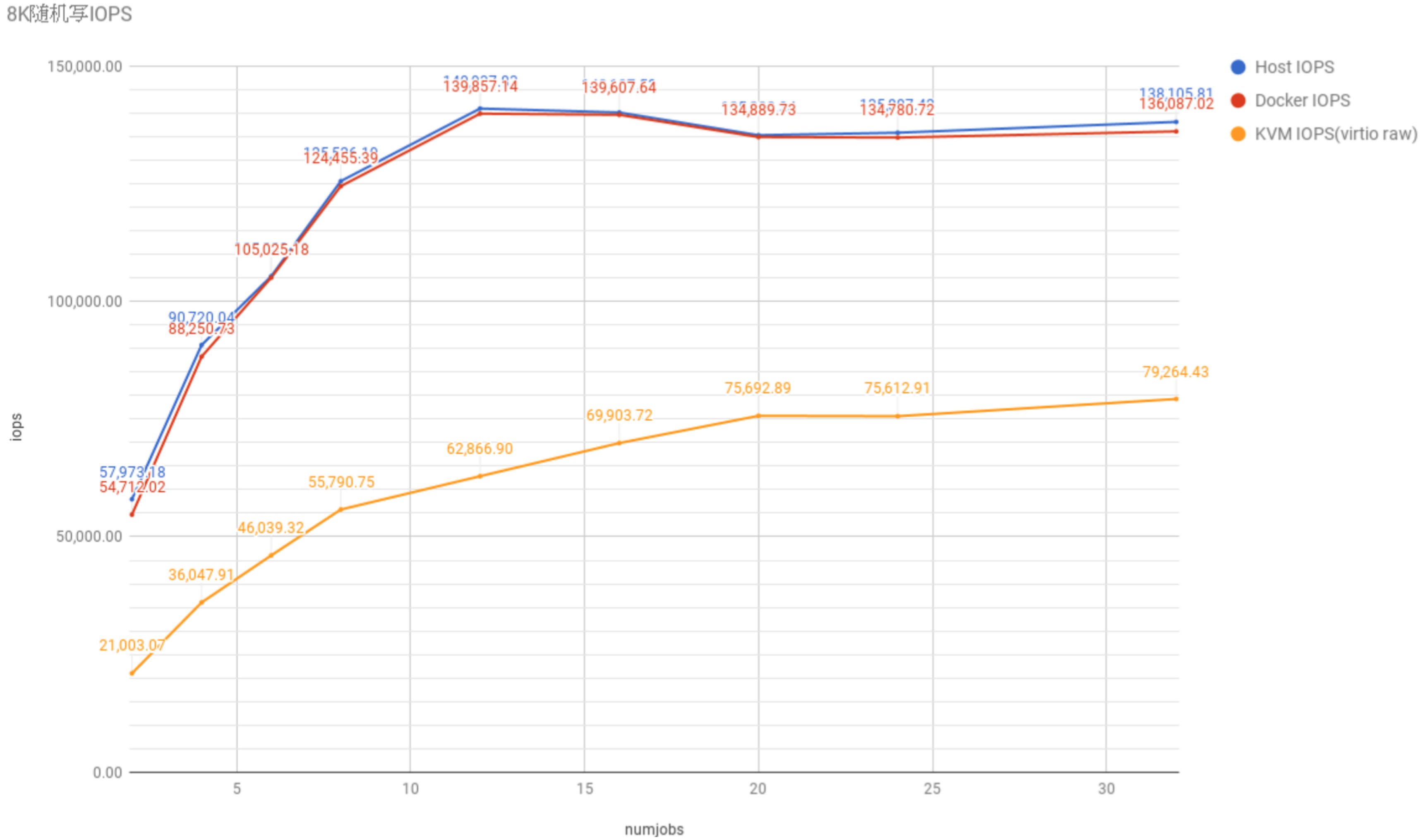
优化

- snapshot， compression
- 基于最新存储技术进行优化

- fast performance
- cost-efficient
- high availability
- security
- easy to set up, operate and scale

cost-efficient

Host/KVM/Docker

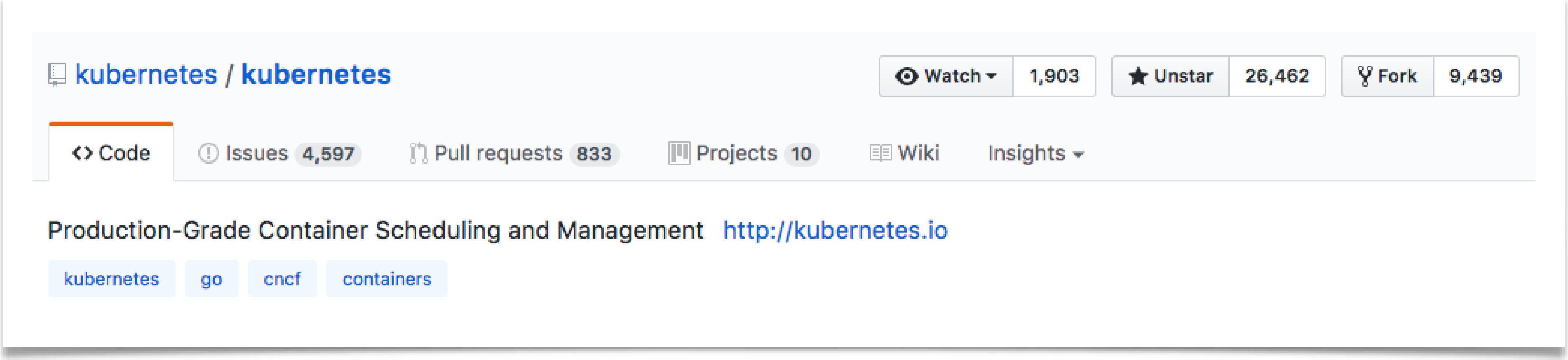


- fast performance— NVMe+Infiniband
- cost-efficient— Docker
- high availability
- security
- easy to set up, operate and scale

Kubernetes

Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications.

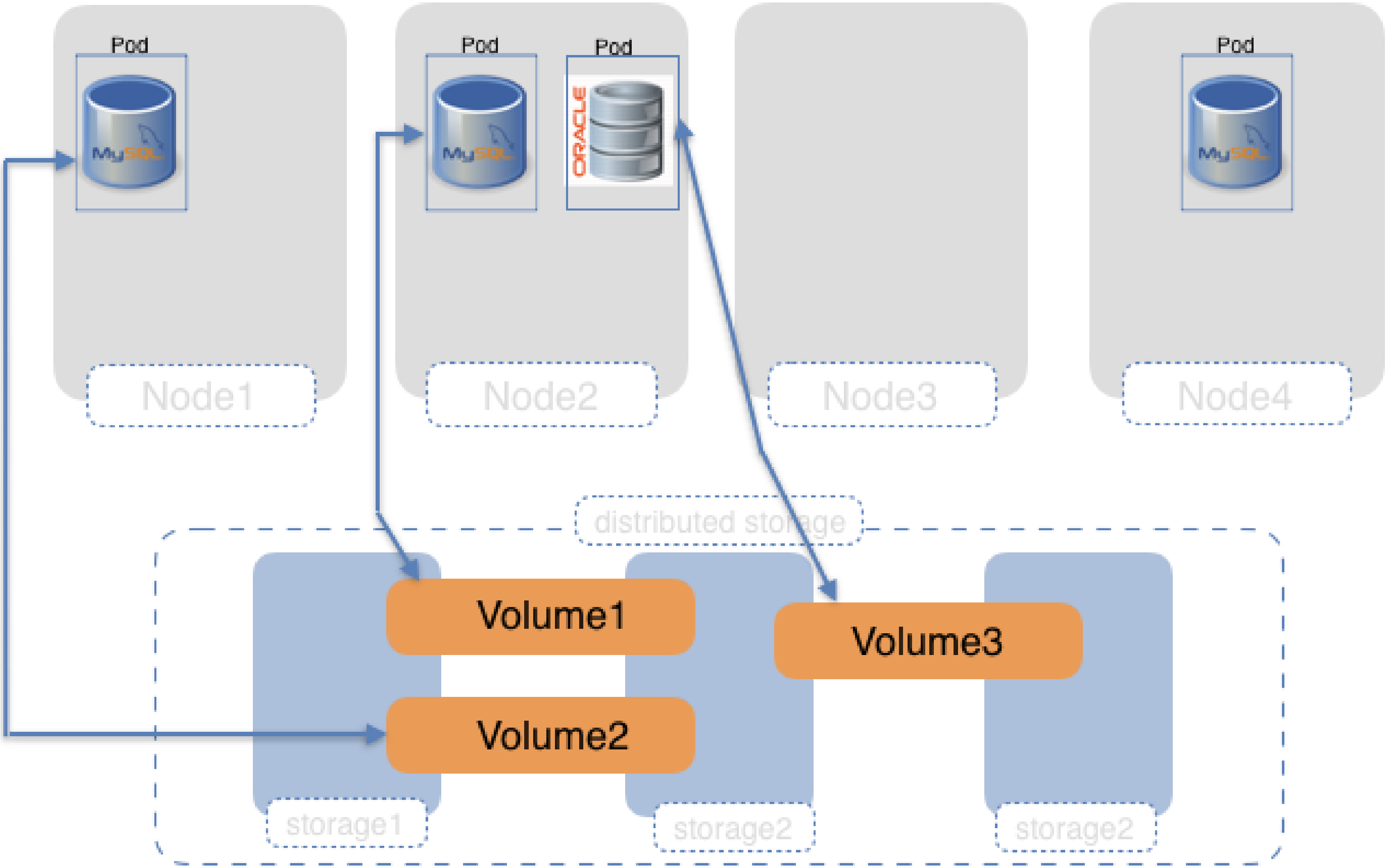
It groups containers that make up an application into logical units for easy management and discovery. Kubernetes builds upon 15 years of experience of running production workloads at Google, combined with best-of-breed ideas and practices from the community.



kubernetes / kubernetes

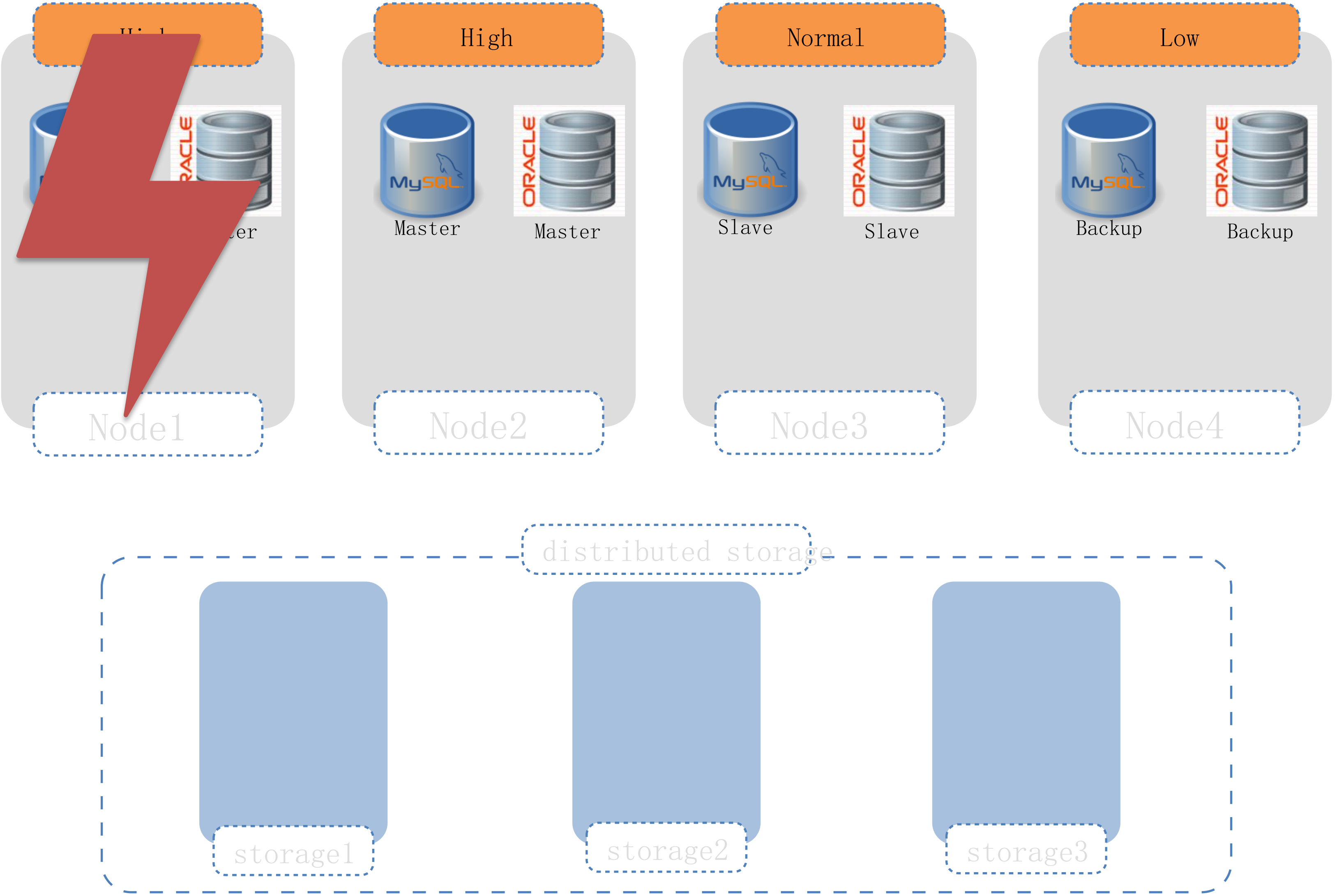
Production-Grade Container Scheduling and Management http://kubernetes.io

Kubernetes


Pod
Pod
Pod
Pod
MySQL
MySQL
ORACLE
MySQL
Node1
Node2
Node3
Node4
distributed storage
Volume1
Volume2
Volume3
storage1
storage2
storage2

- fast performance
- cost-efficient
- high availability
- security
- easy to set up, operate and scale

high availability



Gdevops

全球敏捷运维峰会

THANK YOU！