高性能RDBMS及基于K8S的RDS尝试

杭州沃趣科技股份有限公司
Hangzhou WOQU Technology Co., Ltd.

Let Data Drive!

Performance

Infiniband

SSD IOPS

Flash Oracle

Kubernetes

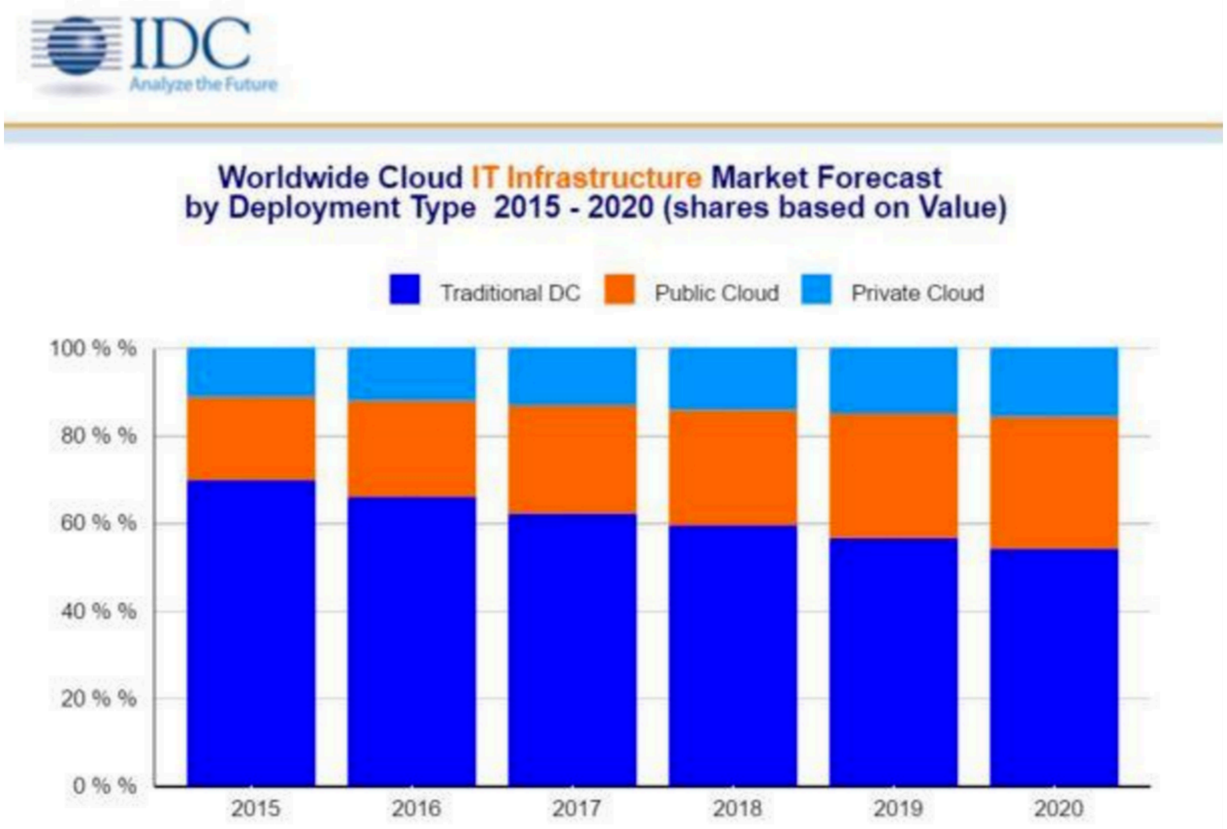
RDBMS Cloud NVMe

Availability Latency

RDS DB-Engine

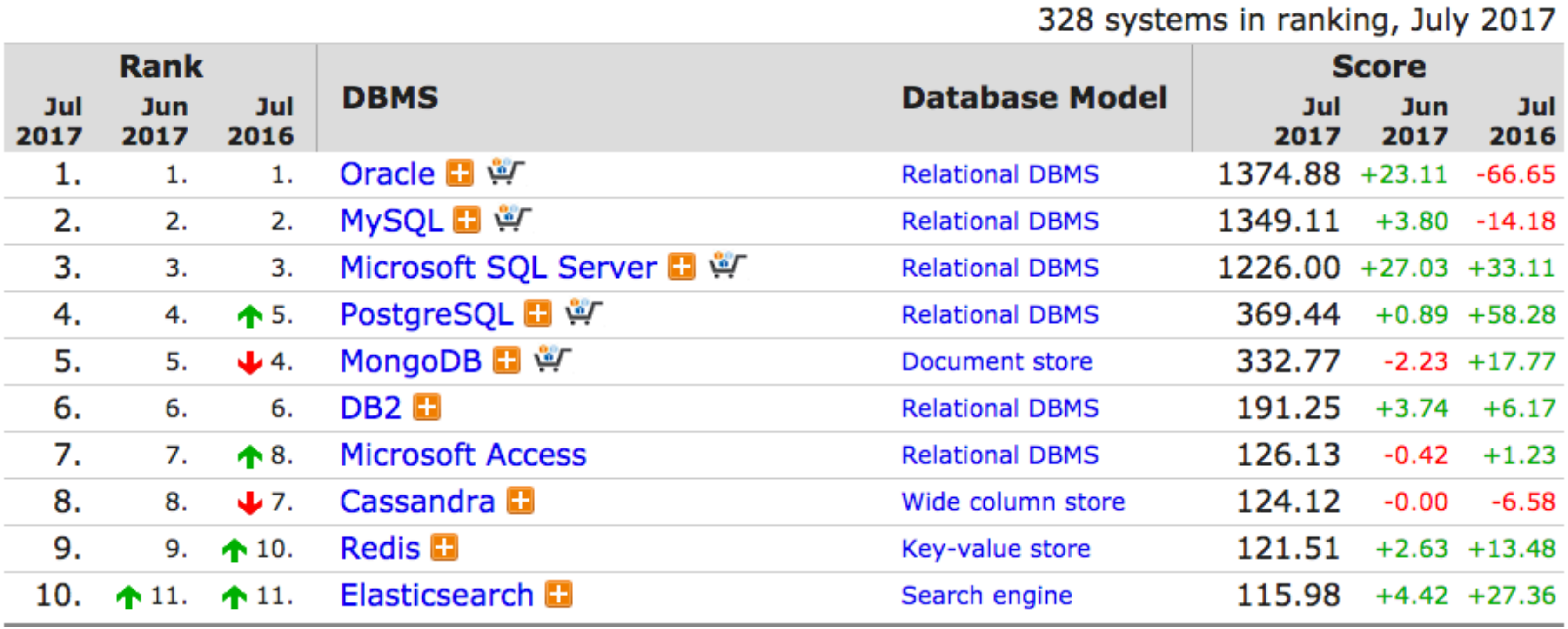
RDMA MySQL

Private Cloud



Relational Database

Graph DBMS , Key-value stores , Relational DBMS , Multivalue DBMS, Object oriented DBMS



328 systems in ranking, July 2017

Rank			DBMS	Database Model	Score		
Jul 2017	Jun 2017	Jul 2016			Jul 2017	Jun 2017	Jul 2016
1.	1.	1.	Oracle	Relational DBMS	1374.88	+23.11	-66.65
2.	2.	2.	MySQL	Relational DBMS	1349.11	+3.80	-14.18
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1226.00	+27.03	+33.11
4.	4.	↑ 5.	PostgreSQL	Relational DBMS	369.44	+0.89	+58.28
5.	5.	↓ 4.	MongoDB	Document store	332.77	-2.23	+17.77
6.	6.	6.	DB2	Relational DBMS	191.25	+3.74	+6.17
7.	7.	↑ 8.	Microsoft Access	Relational DBMS	126.13	-0.42	+1.23
8.	8.	↓ 7.	Cassandra	Wide column store	124.12	-0.00	-6.58
9.	9.	↑ 10.	Redis	Key-value store	121.51	+2.63	+13.48
10.	↑ 11.	↑ 11.	Elasticsearch	Search engine	115.98	+4.42	+27.36

Relational Database Service

Service : AWS RDS

- It provides **cost-efficient** and resizable capacity while automating time-consuming administration tasks such as hardware provisioning, **database setup, patching and backups**
- **the fast performance, high availability, security** and compatibility they need

- ✓ fast performance
- ✓ cost-efficient
- ✓ services
  - ✓ high availability
  - ✓ security

fast performance

导致数据库性能问题：应用,Schema,Index ,SQL,执行计划,CPU, 内存......

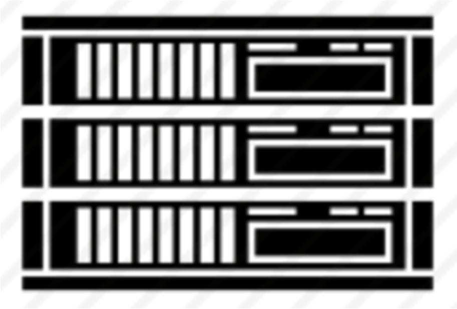
IO 模型：(仅以online redo日志为例)

- WAL：Write-ahead logging
- direct,sync,连续,512byte

对存储的要求是：

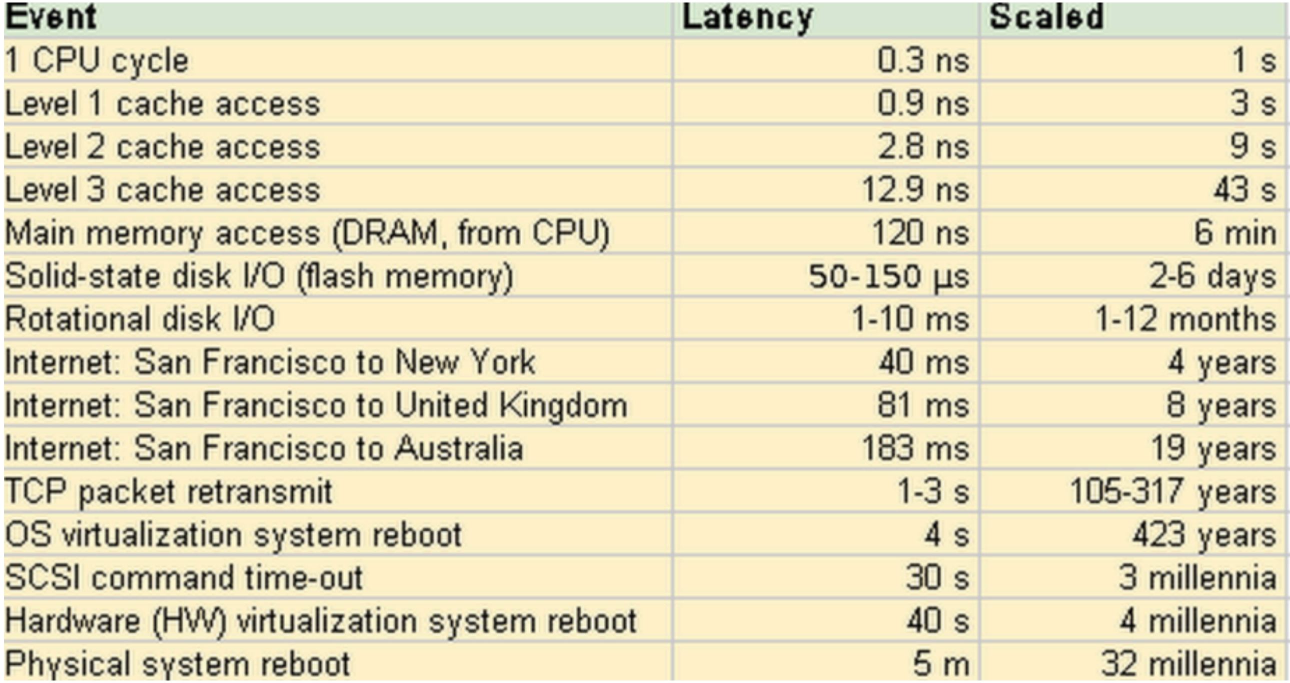
- IOPS
- 延时：QoS,Jitter

大多数时候



fast performance : 存储介质

- **Principle of Locality**
- **Shaving x off lantency 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

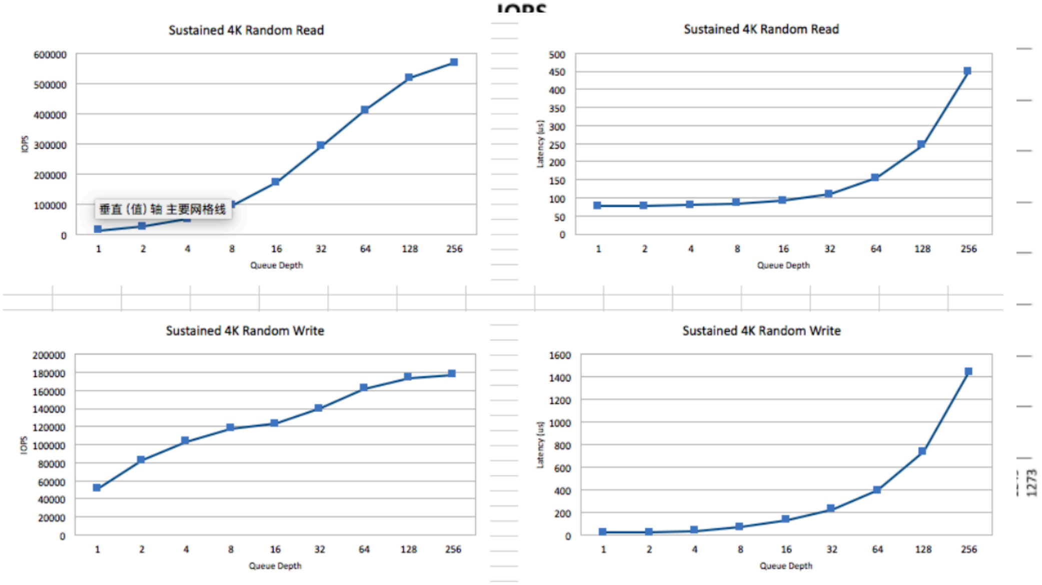
SSD 解救 DBA

fast performance :存储介质

NAND SSD / Flash 可以解决所有问题吗?

Write amplification
Garbage Collection

- IO Queue Depth
- 读/写
- 空盘/满盘
- 抖动



fast performance：存储介质

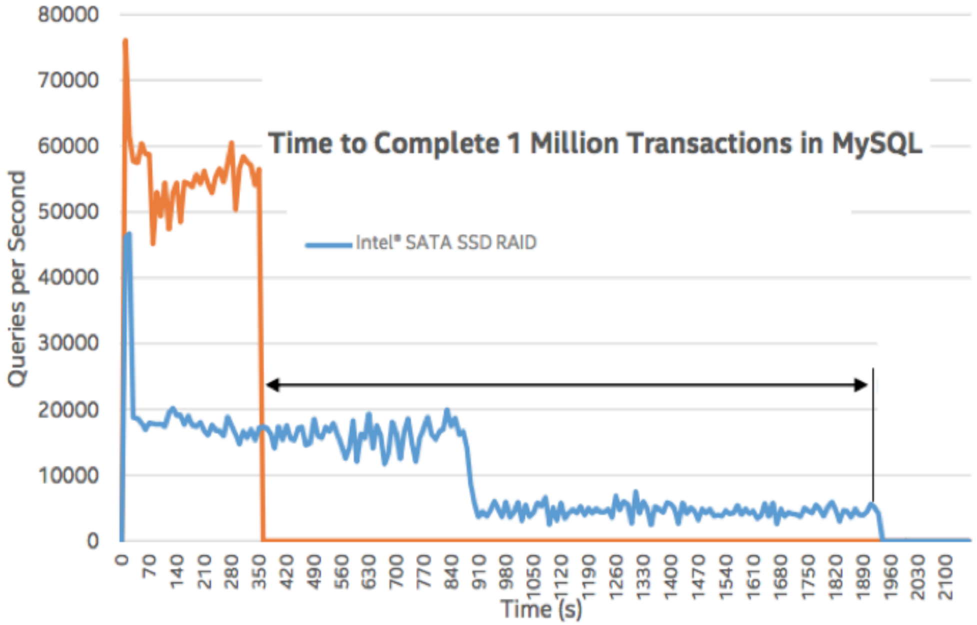
看蓝线

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**

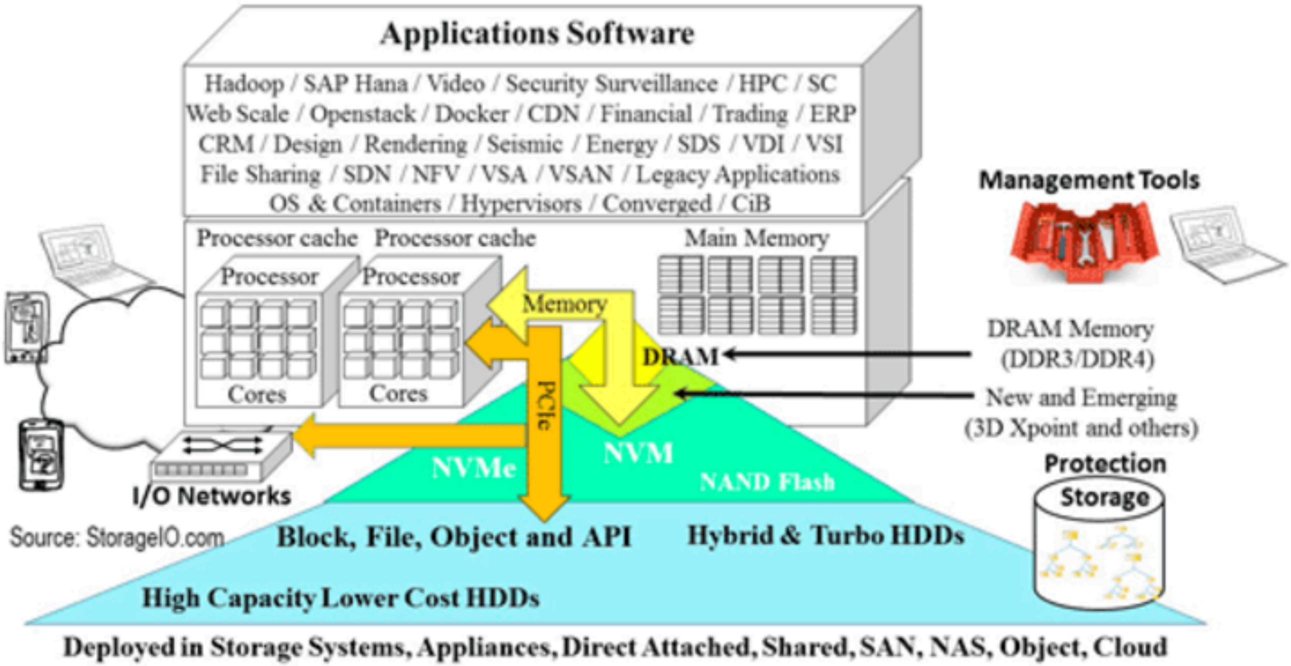
问题：

- **蓝线有两次下降**



fast performance :存储协议

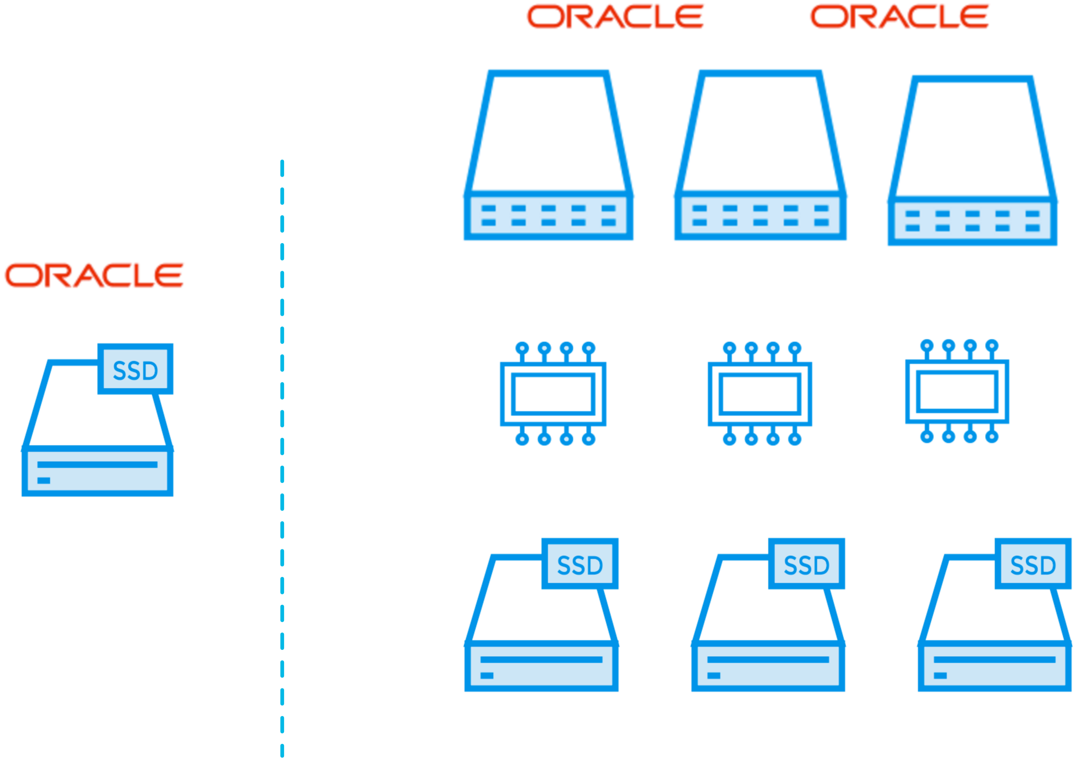
- **SAS/SCSI**
- **NVMe**


Applications Software
Hadoop / SAP Hana / Video / Security Surveillance / HPC / SC
Web Scale / Openstack / Docker / CDN / Financial / Trading / ERP
CRM / Design / Rendering / Seismic / Energy / SDS / VDI / VSI
File Sharing / SDN / NFV / VSA / VSAN / Legacy Applications
OS & Containers / Hypervisors / Converged / CiB
Processor cache
Processor cache
Processor
Processor
Cores
Cores
Memory
Main Memory
PCIe
DRAM
NVM
NVMe
NAND Flash
I/O Networks
Management Tools
DRAM Memory
(DDR3/DDR4)
New and Emerging
(3D Xpoint and others)
Protection
Storage
Block, File, Object and API
Hybrid & Turbo HDDs
High Capacity Lower Cost HDDs
Deployed in Storage Systems, Appliances, Direct Attached, Shared, SAN, NAS, Object, Cloud

Source: StorageIO.com

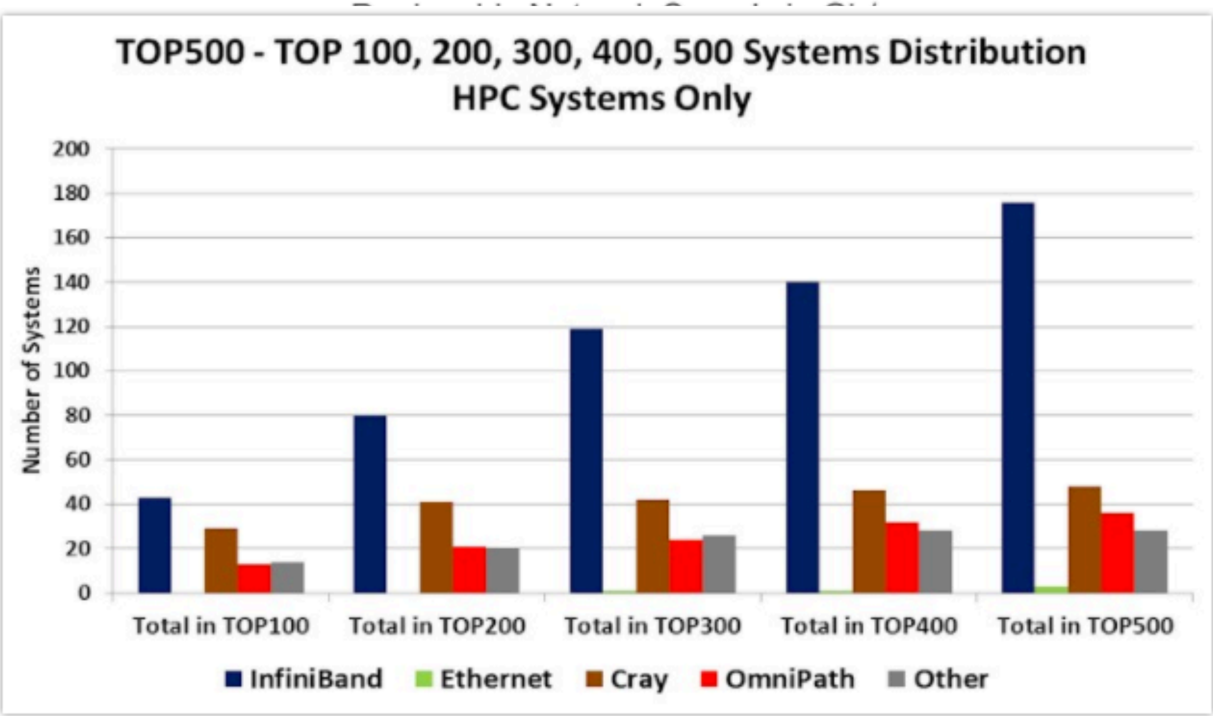
NVM-Express protocol, which allows compute and memory complexes to talk directly to flash storage rather than have the flash emulate a disk and go through the SCSI device driver stack

fast performance :存储网络



fast performance :存储网络

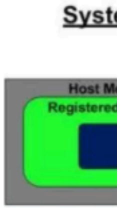
硬件层面

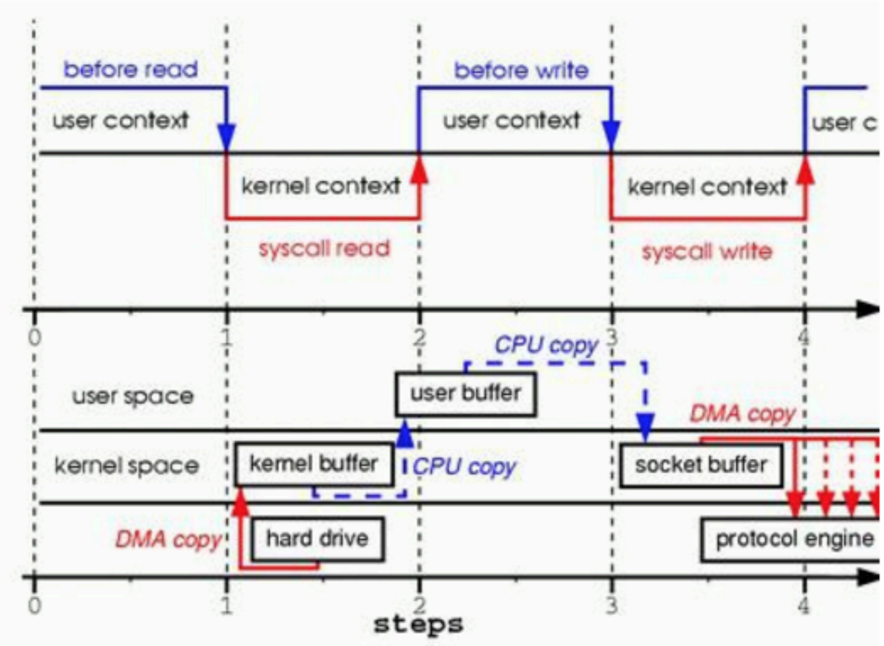


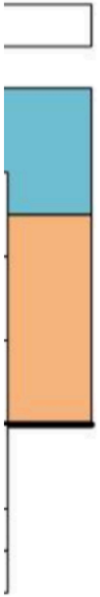
fast performance :存储网络协议

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







• RDMA allows has a ratio of 1 !!

fast performance : NVMf

iSer

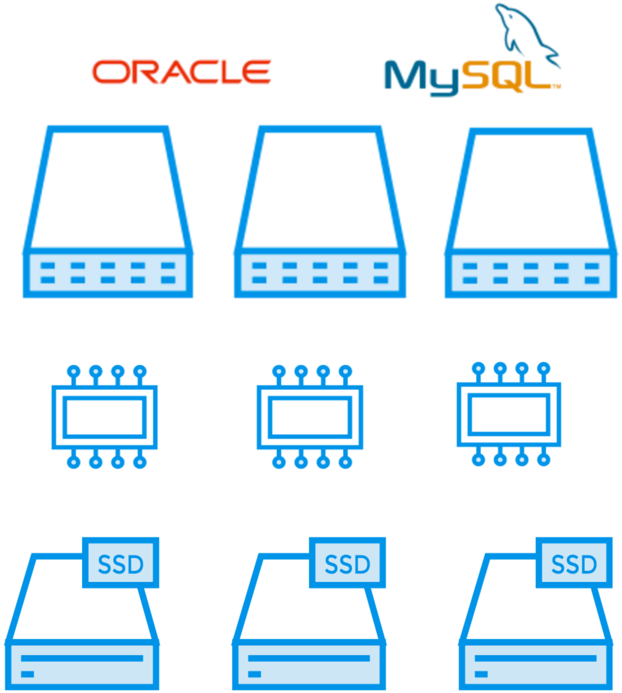
iscsi + rdma +infiniband

NVMf

NVMe+ rdma +infiniband

测试模型需要继续优化

fast performance



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 increament 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 :分布式存储

- ✓ fast performance
- ✓ cost-efficient
- ✓ services
    - ✓ high availability
    - ✓ security

cost-efficient :分布式存储

Host/KVM/Docker

- ✓ fast performance
- ✓ cost-efficient
- ✓ Services
  - ✓ high availability
  - ✓ security

services

high availability :

- Oracle Rac /MySQL Galera

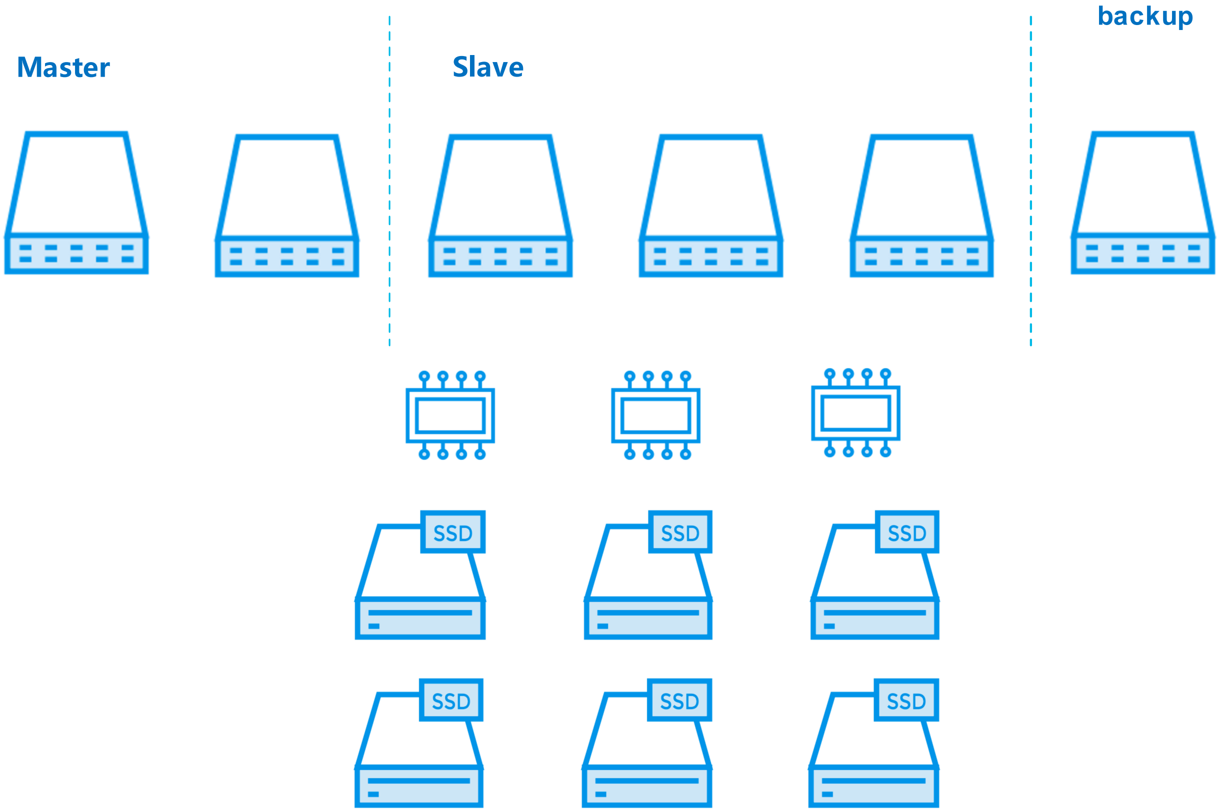
用户需要什么:

- 故障检测机制
- 免干预的切换流程
- 60s 内完成切换过程
- 应用透明

services:

- 备库水平扩展,逻辑/物理备份
- 安全

services


Master
Slave
backup
SSD
SSD
SSD
SSD
SSD
SSD