南京大学
NANJING UNIVERSITY

# 英特尔®傲腾™在存储中的应用

姚 舸

南京大学

人工微结构科学与技术协同创新中心

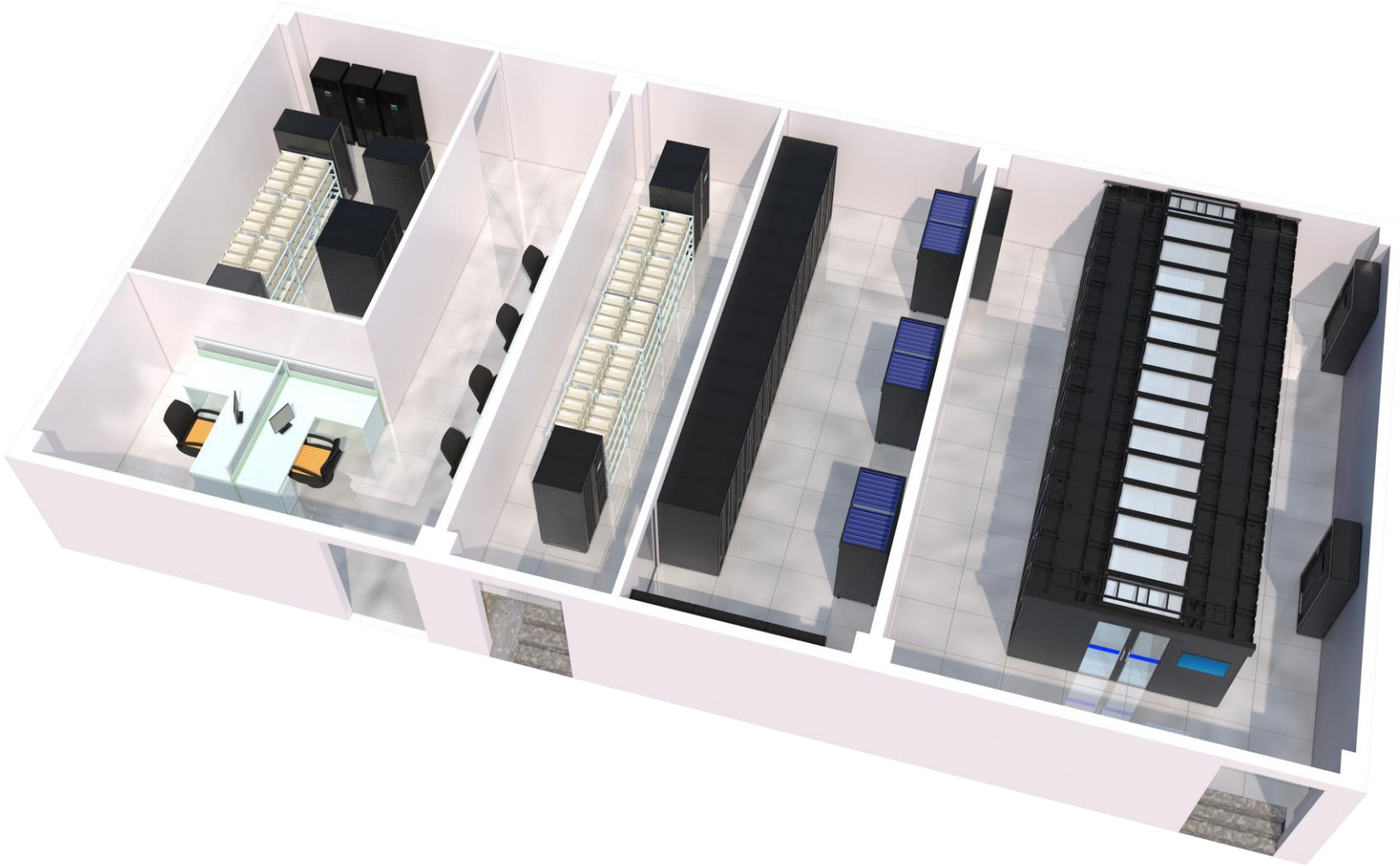
yaoge@nju.edu.cn my@yaoge123.com

# e-Science 中心

- 服务教学科研
  - 高性能计算
  - 云盘
  - 开源镜像
  - Git
  - 云平台
  - 站群
  - 信息安全
  - 实训平台
  - 仪器管理
  - 信息发布
  - 门禁
  - 财务
  - 资产

# 基础设施

- 低密度机房
  - 高业务连续性
  - 云平台系统
  - HPC虚拟化
- 高密度机房
  - HPC计算与IO



# HPC管理

- 统一机房课题组自行采购管理
  - 采购竞争不充分，硬件与机房环境可能不匹配
  - 学生管理为主，精力不足、更替传承问题
  - 资源使用不均衡，运行效率低

# HPC管理

- 统一机房课题组自行采购管理
  - 采购竞争不充分，硬件与机房环境可能不匹配
  - 学生管理为主，精力不足、更替传承问题
  - 资源使用不均衡，运行效率低
- 各个课题组集群融合统一管理
  - 全生命周期管理，协助选型参与购置
  - 自主设计、安装、部署、运维
  - 统一管理、单一集群资源共享
  - 培训讲座，QQ群、微信群、公众号在线服务

# 品牌多样化

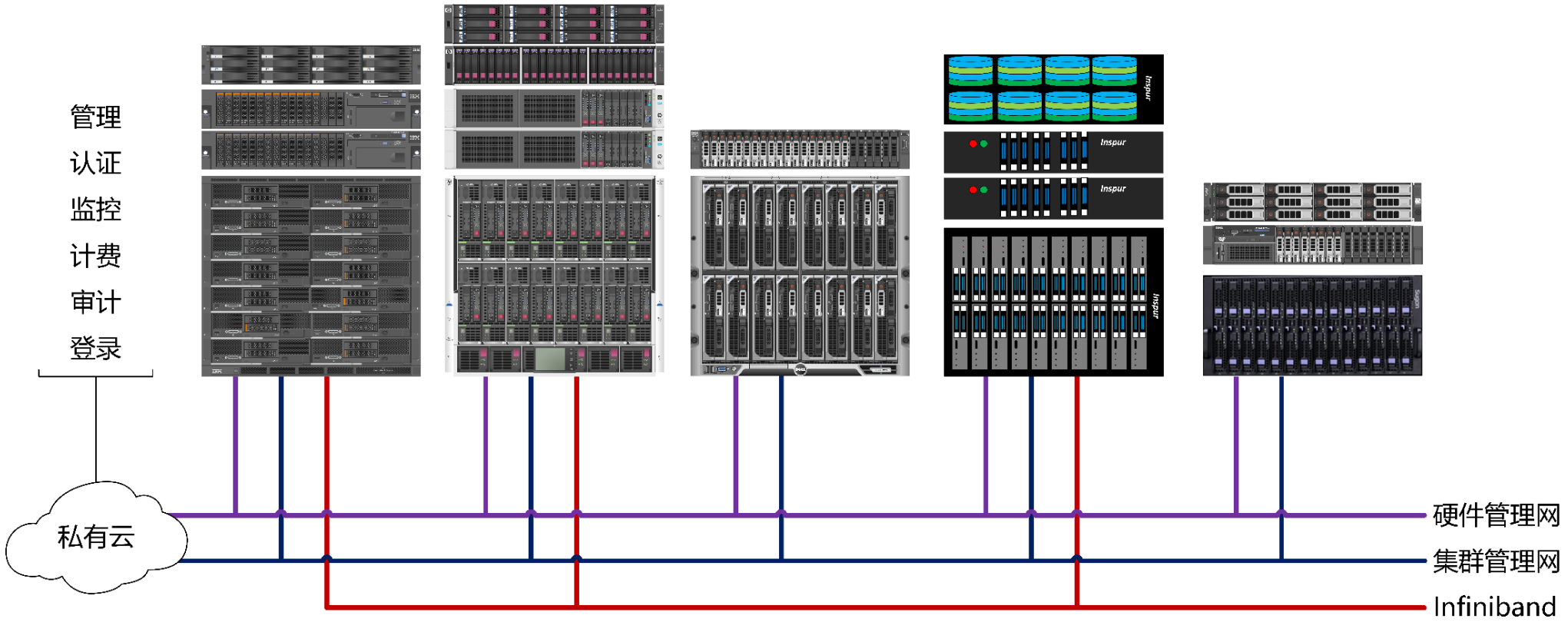

计算
IBM
Hewlett Packard Enterprise
DELL
intel
Lenovo
SUPERMICRO
H3C
ASUS
中科曙光
Sugon
inspur
GIGABYTE
HUAWEI
存储
IBM
Hewlett Packard Enterprise
DELL
DDN
AI · BIG DATA · HPC
HUAWEI
紫光西部数据
UNIS WDC
inspur
网络
cisco
Hillstone NETWORKS
H3C
BLADE NETWORK TECHNOLOGIES
DELL
RUCKUS an ARRIS company
Voltaire
Mellanox TECHNOLOGIES
HUAWEI
intel
Extreme

# 配置多样化

- CPU
  - Intel Xeon ~~5150~~, ~~E5430~~, ~~X5550~~, X7542, ~~E5620~~, E5645, X5650, E5-2643, E5-2660, E5-2692v2, E5-2680v3, E5-2630v4, E5-2640v4, E5-2650v4, E5-2660v4, E5-2680v4, E5-2682v4, E7-4850v4, Gold 6140, Gold 6148, Gold 6248, Platinum 9242, Gold 5220R, Gold 6226R.
  - AMD EPYC 7302P, 7702.
- GPU
  - Nvidia Tesla K40, P100, V100(PCIe), V100(SXM2), GeForce RTX 2080 Ti, TITAN RTX.
- NET
  - 1/10/25/40/100Gb Ethernet, 10Gb iWARP Ethernet, 20/40/56/100/200Gb InfiniBand, 100Gb Omni-Path Architecture.

# 架构多样化

- 计算节点：619　　理论峰值：661 TFLOPS + 411 TFLOPS
- IO+BB: 9+4 台；虚机: 12 个；存储: 4 套；交换机: 16 ETH、4 IB、2 OPA



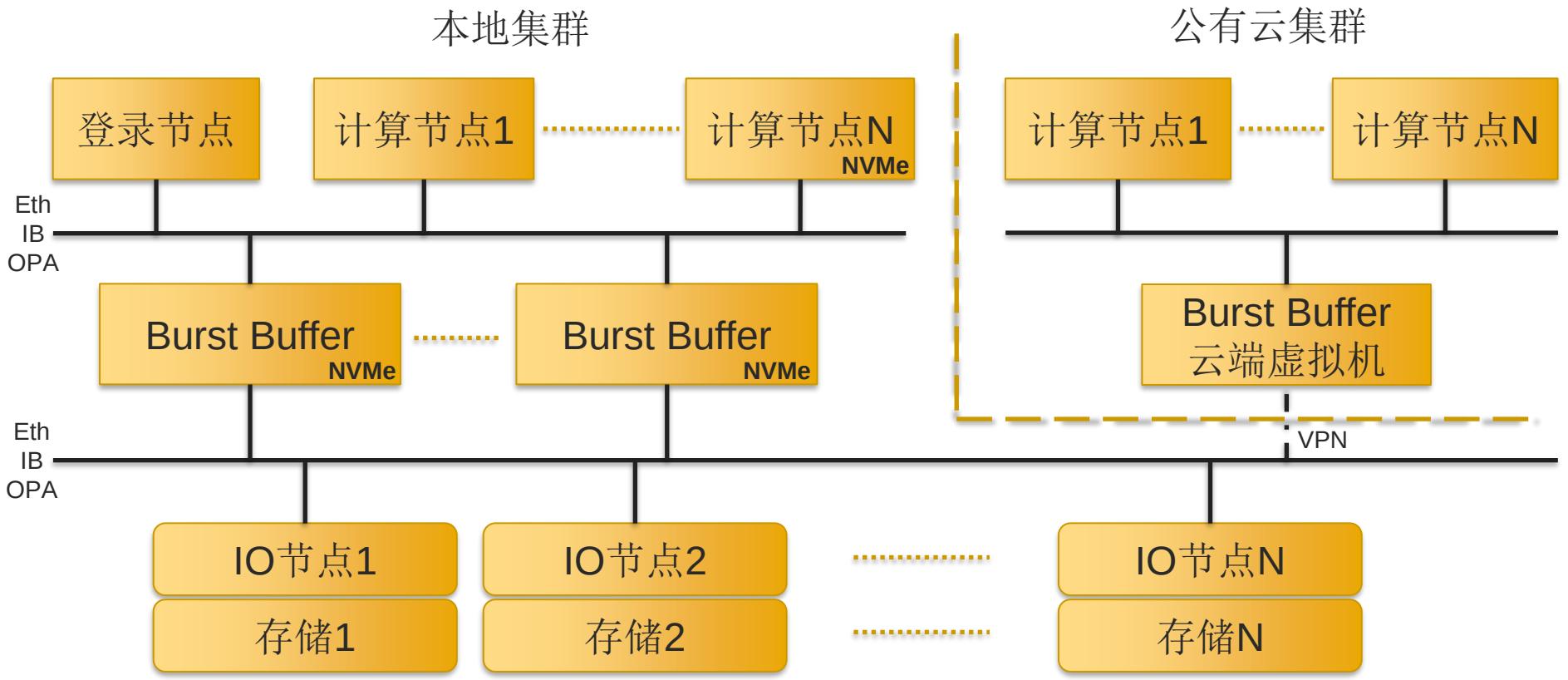
# 安全运营


无安全防范，访问入口不统一
NGFW+WAF
校内校外统一访问入口
重保时期持续校外访问
弱密码，密码共享，否认登录
禁用Key，静态+动态密码
TOTP 30s有效，一次一密
集中认证
账号共享，攻击集群，否认操作
一人一账号
统一认证同步
完整操作审计

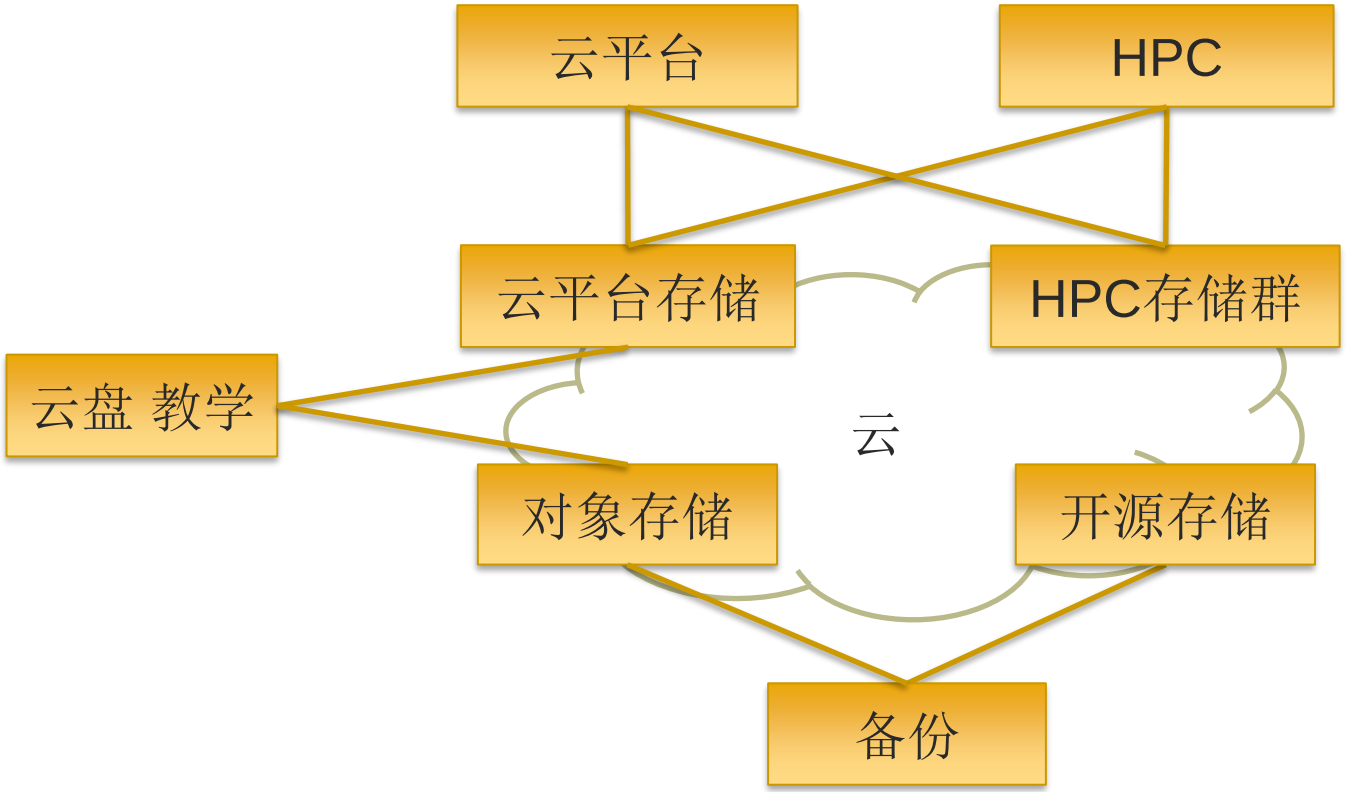
# 安全运营


无安全防范，访问入口不统一
NGFW+WAF
校内校外统一访问入口
重保时期持续校外访问
弱密码，密码共享，否认登录
禁用Key，静态+动态密码
TOTP 30s有效，一次一密
集中认证
账号共享，攻击集群，否认操作
一人一账号
统一认证同步
完整操作审计

# HPC 存储群


本地集群
公有云集群
登录节点
计算节点1
计算节点N
NVMe
计算节点1
计算节点N
Eth
IB
OPA
Burst Buffer
NVMe
Burst Buffer
NVMe
Burst Buffer
云端虚拟机
VPN
Eth
IB
OPA
IO节点1
IO节点2
IO节点N
存储1
存储2
存储N

# 存储总览



# 开源存储 OpenZFS

高性价比，高数据持久性，业务连续性要求不高

- 集成了逻辑卷管理的文件系统
- 代替存储或RAID卡提供数据保护

# 开源存储 OpenZFS

## 提升性能

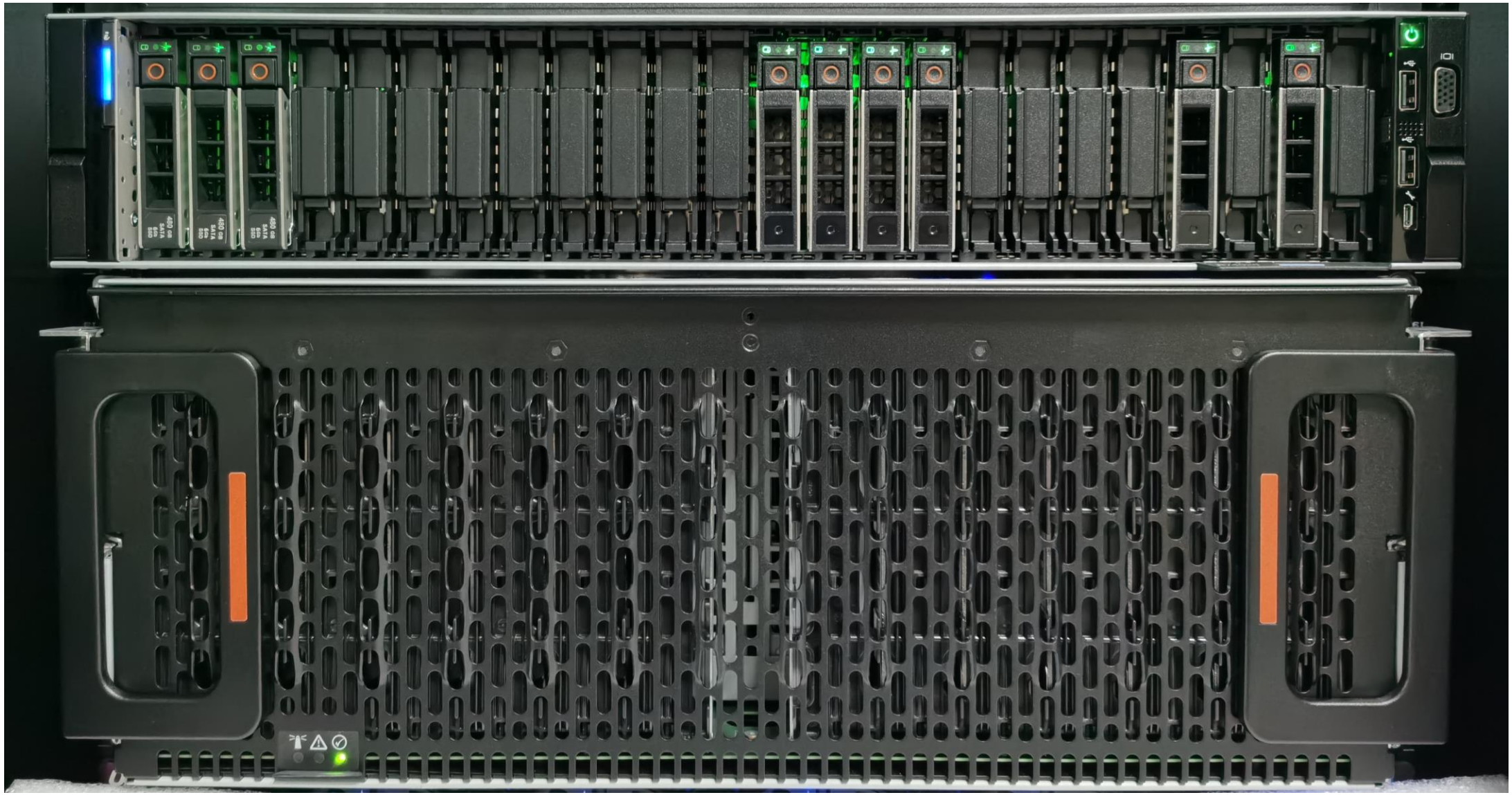
- ARC：内存中的缓存
- L2ARC：ARC的扩展，一般采用SSD
- ZIL：持久性写缓存，用于意外恢复

# 开源存储 OpenZFS

## 通用硬件

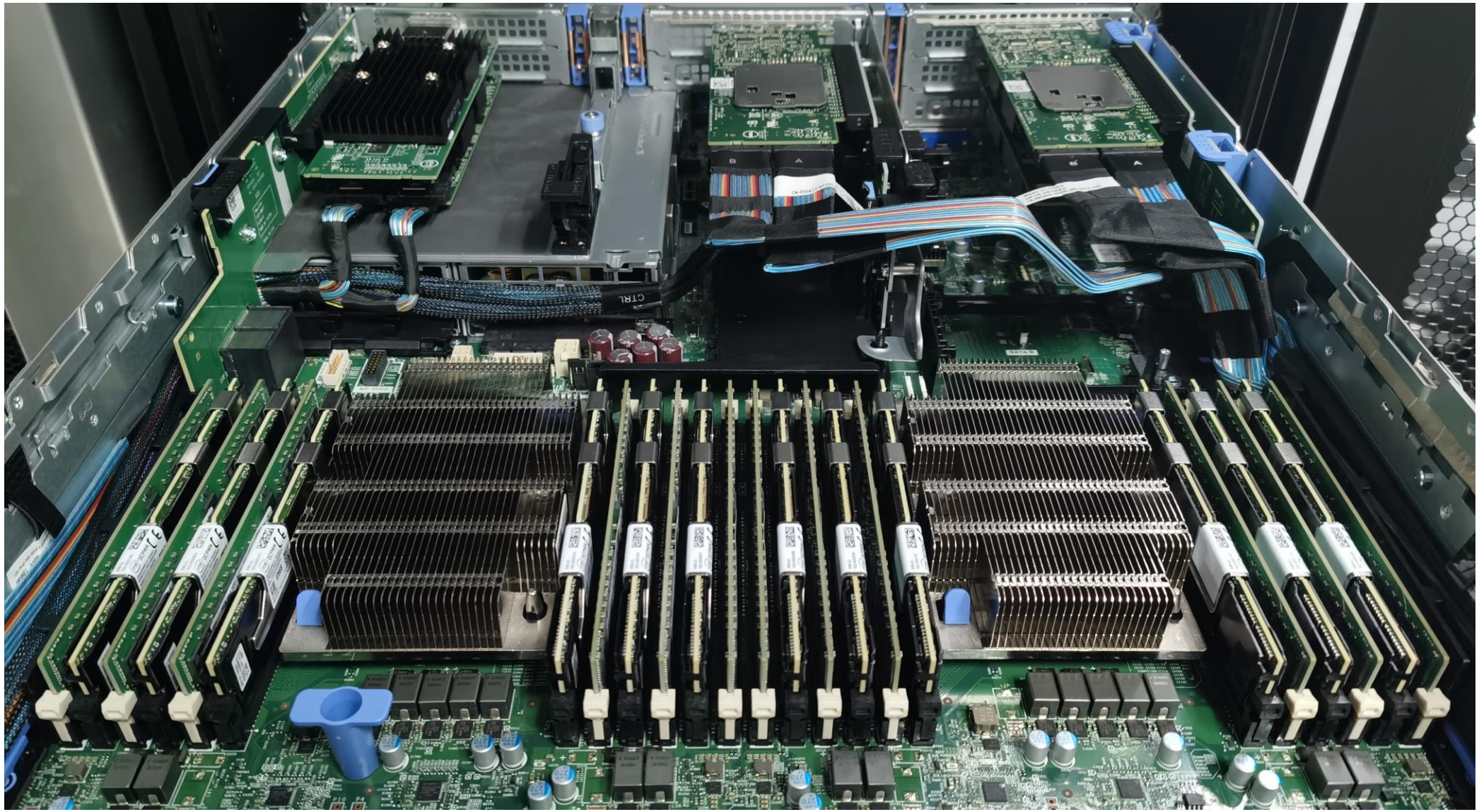
- X86服务器：Dell R740xd, 2*Intel Gold 5218, 12*32GB RAM, Broadcom Dual 25Gb Ethernet, 480GB SSD RAID1, 2*Dell 12Gbps HBA(SAS3508)连接JBOD
- JBOD：HGST Ultrastar Data60 H4060-J，双控制器，60*14TB NL-SAS
- 4*Intel SSD P4510 2TB U.2
- 2*Intel Optane SSD P4800X 375GB U.2
- 12*Intel Optane Persistent Memory 128GB

# 开源存储 OpenZFS



# 开源存储 OpenZFS

# 开源存储 OpenZFS



# Optane Persistent Memory

MM+AD 混合模式

- 95%设置为内存模式

  ipmctl create -goal MemoryMode=95

- 创建两个内存盘：

  ndctl create-namespace -m raw

  ndctl create-namespace -m raw

# 开源存储 OpenZFS

- 调整ARC容量

  /sys/module/zfs/parameters/zfs_arc_max

- 添加L2ARC

  zpool add *tank* cache nvme0n1 nvme1n1 nvme2n1 nvme3n1

- 添加ZIL使用镜像保护

  zpool add *tank* log mirror pmem0 pmem1

- 异步写也落盘（可选）

  zfs set sync=always *tank*

# 测试环境

- HDD
  - ARC: 247 GB DDR4 RAM
  - ZIL: 5* Raidz3 (8+3) 14TB NL-SAS

- NVMe SSD
  - ARC: 247 GB DDR4 RAM
  - L2ARC: 2* 2TB Intel SSD P4510
  - ZIL: 2* 2TB Intel SSD P4510 Mirror

# 测试环境

- NVMe Optane
  - ARC: 1.3 TB Optane MM
  - L2ARC: 4* 2TB Intel SSD P4510
  - ZIL: 2* 375GB Intel Optane SSD P4800X Mirror

- DIMM Optane
  - ARC: 1.3 TB Optane MM
  - L2ARC: 4* 2TB Intel SSD P4510
  - ZIL: 2* 36GB Optane AD Mirror

# 写入寿命

| Product | Capacity | Lifetime Writes |
|---|---|---|
| Intel SSD P4510 | 1 TB | 1.92 PBW |
| Intel SSD P4510 | 2 TB | 2.61 PBW |
| Intel SSD P4510 | 4 TB | 6.3 PBW |
| Intel SSD P4510 | 8 TB | 13.88 PBW |
| Intel SSD P4610 | 1.6 TB | 12.25 PBW |
| Intel SSD P4610 | 3.2 TB | 21.85 PBW |
| Intel SSD P4610 | 6.4 TB | 36.54 PBW |
| Intel Optane SSD P4801X | 100 GB | 10.9 PBW |
| Intel Optane SSD P4800X | 375 GB | 41.0 PBW |

写入速度：100MB/s

五年写入：14.69PB

# 异步顺序写测试

| Block Size | Process | HDD | NVMe SSD | NVMe Optane | DIMM Optane | |
|---|---|---|---|---|---|---|
| 128k | 1 | 1138<br>109 | 1123<br>111 | 963<br>136 | 1001<br>124 | Bandwidth (MiB/s)<br>Total latency (us) |
| 128k | 64 | 1424<br>5616 | 1531<br>5220 | 2073<br>4042 | 1660<br>4815 | Bandwidth (MiB/s)<br>Total latency (us) |
| 4k | 1 | 149k<br>6.43 | 151k<br>6.36 | 146k<br>6.57 | 148k<br>6.48 | IOPS<br>Total latency (us) |
| 4k | 64 | 347k<br>183 | 345k<br>185 | 371k<br>172 | 375k<br>170 | IOPS<br>Total latency (us) |

Ubuntu 20.04.1 LTS, 5.4.0-42-generic, OpenZFS 0.8.3-1ubuntu12.3, fio-3.16
fio ioengine=psync iodepth=1 direct=1 rw=write -sync=0

# 异步随机写测试

| Block Size | Process | HDD | NVMe SSD | NVMe Optane | DIMM Optane | |
|---|---|---|---|---|---|---|
| 128k | 1 | 1153<br>108 | 1033<br>120 | 1234<br>101 | 1380<br>90 | Bandwidth (MiB/s)<br>Total latency (us) |
| 128k | 64 | 378<br>21145 | 348<br>22964 | 345<br>23159 | 501<br>15968 | Bandwidth (MiB/s)<br>Total latency (us) |
| 4k | 1 | 7930<br>125 | 7203<br>138 | 8103<br>122 | 10200<br>97 | IOPS<br>Total latency (us) |
| 4k | 64 | 3177<br>20099 | 3181<br>20103 | 3634<br>17601 | 4708<br>13586 | IOPS<br>Total latency (us) |

Ubuntu 20.04.1 LTS, 5.4.0-42-generic, OpenZFS 0.8.3-1ubuntu12.3, fio-3.16
fio ioengine=psync iodepth=1 direct=1 rw=randwrite -sync=0

# 同步顺序写测试

| Block Size | Process | HDD | NVMe SSD | NVMe Optane | DIMM Optane | |
|---|---|---|---|---|---|---|
| 128k | 1 | 6.72<br>18590 | 473<br>263 | 407<br>306 | 250<br>499 | Bandwidth (MiB/s)<br>Total latency (us) |
| 128k | 64 | 209<br>38240 | 1384<br>5777 | 1360<br>5877 | 974<br>8213 | Bandwidth (MiB/s)<br>Total latency (us) |
| 4k | 1 | 77<br>12900 | 10.3k<br>96 | 9.66k<br>103 | 10.1k<br>98 | IOPS<br>Total latency (us) |
| 4k | 64 | 4089<br>15640 | 143k<br>447 | 135k<br>472 | 90.8k<br>704 | IOPS<br>Total latency (us) |

Ubuntu 20.04.1 LTS, 5.4.0-42-generic, OpenZFS 0.8.3-1ubuntu12.3, fio-3.16
fio ioengine=psync iodepth=1 direct=1 rw=write -sync=1

# 同步随机写测试

| Block Size | Process | HDD | NVMe SSD | NVMe Optane | DIMM Optane | |
|---|---|---|---|---|---|---|
| 128k | 1 | 5.17<br>24160 | 443<br>281 | 442<br>281 | 253<br>492 | Bandwidth (MiB/s)<br>Total latency (us) |
| 128k | 64 | 71.5<br>111 | 436<br>18351 | 513<br>15591 | 601<br>13300 | Bandwidth (MiB/s)<br>Total latency (us) |
| 4k | 1 | 68<br>14580 | 5540<br>179 | 5960<br>166 | 5636<br>176 | IOPS<br>Total latency (us) |
| 4k | 64 | 971<br>65820 | 2980<br>21467 | 3605<br>17741 | 3800<br>16833 | IOPS<br>Total latency (us) |

Ubuntu 20.04.1 LTS, 5.4.0-42-generic, OpenZFS 0.8.3-1ubuntu12.3, fio-3.16
fio ioengine=psync iodepth=1 direct=1 rw=randwrite -sync=1

# 顺序读测试

| Block Size | Process | ARC 2GB HDD | ARC 247GB HDD | DIMM Optane | |
|---|---|---|---|---|---|
| 128k | 1 | 127<br>982 | 2622<br>47.42 | 2516<br>49.41 | Bandwidth (MiB/s)<br>Total latency (us) |
| 128k | 64 | 851<br>9383 | 3314<br>2384 | 3539<br>2230 | Bandwidth (MiB/s)<br>Total latency (us) |
| 4k | 1 | 32.4k<br>30.34 | 287k<br>3.29 | 288k<br>3.29 | IOPS<br>Total latency (us) |
| 4k | 64 | 218k<br>29.32 | 856k<br>74.08 | 879k<br>71.94 | IOPS<br>Total latency (us) |

Ubuntu 20.04.1 LTS, 5.4.0-42-generic, OpenZFS 0.8.3-1ubuntu12.3, fio-3.16
fio ioengine=psync iodepth=1 direct=1 rw=read

# 随机读测试

| Block Size | Process | ARC 2GB HDD | ARC 247GB HDD | DIMM Optane | |
|---|---|---|---|---|---|
| 128k | 1 | 11.8<br>10622 | 2528<br>49.04 | 2345<br>52.91 | Bandwidth (MiB/s)<br>Total latency (us) |
| 128k | 64 | 119<br>67150 | 1290<br>6119 | 1548<br>5113 | Bandwidth (MiB/s)<br>Total latency (us) |
| 4k | 1 | 97<br>10252 | 24.4k<br>40.62 | 33.8k<br>29.26 | IOPS<br>Total latency (us) |
| 4k | 64 | 1209<br>52904 | 27.9k<br>2294 | 35.0k<br>1777 | IOPS<br>Total latency (us) |

Ubuntu 20.04.1 LTS, 5.4.0-42-generic, OpenZFS 0.8.3-1ubuntu12.3, fio-3.16
fio ioengine=psync iodepth=1 direct=1 rw=randread

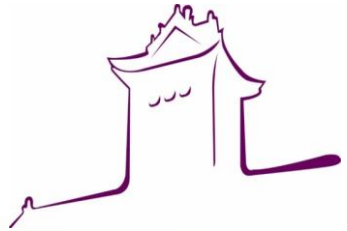
# ZFS延迟（生产环境）

| Type | HDD | NVMe Optane | DIMM Optane |
|---|---|---|---|
| Total IO time - Read | 27 ms | 9 ms | 10 ms |
| Total IO time - Write | 16 ms | 3 ms | 3 ms |
| Disk IO time - Read | 19 ms | 8 ms | 8 ms |
| Disk IO time - Write | 6 ms | 0.77 ms | 0.73 ms |
| Sync queue time - Read | 8 ms | 0.36 ms | 0.77 ms |
| Sync queue time - Write | 1 ms | 0.003 ms | 0.003 ms |
| Async queue time - Read | 7 ms | 1 ms | 1 ms |
| Async queue time - Write | 18 ms | 3 ms | 3 ms |

# Optane Persistent Memory

优势

- 比内存更低的成本
- 比NAND更长的写入寿命
- MM+AD混合模式下节约用于ZIL的SSD




南京大學