颠覆性存储技术，助力云计算发展

毛磊
SSD方案架构师
2014年10月21日

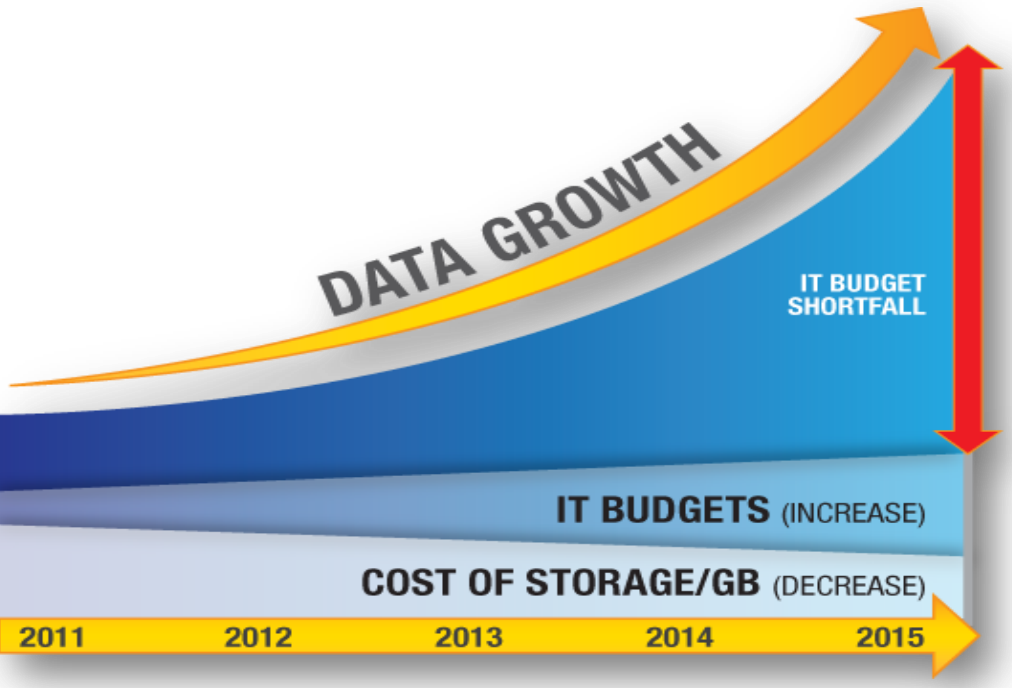
议程

SSD在分布式存储场景的使用

SSD概述

SSD在各个应用场景下的解决方案概述

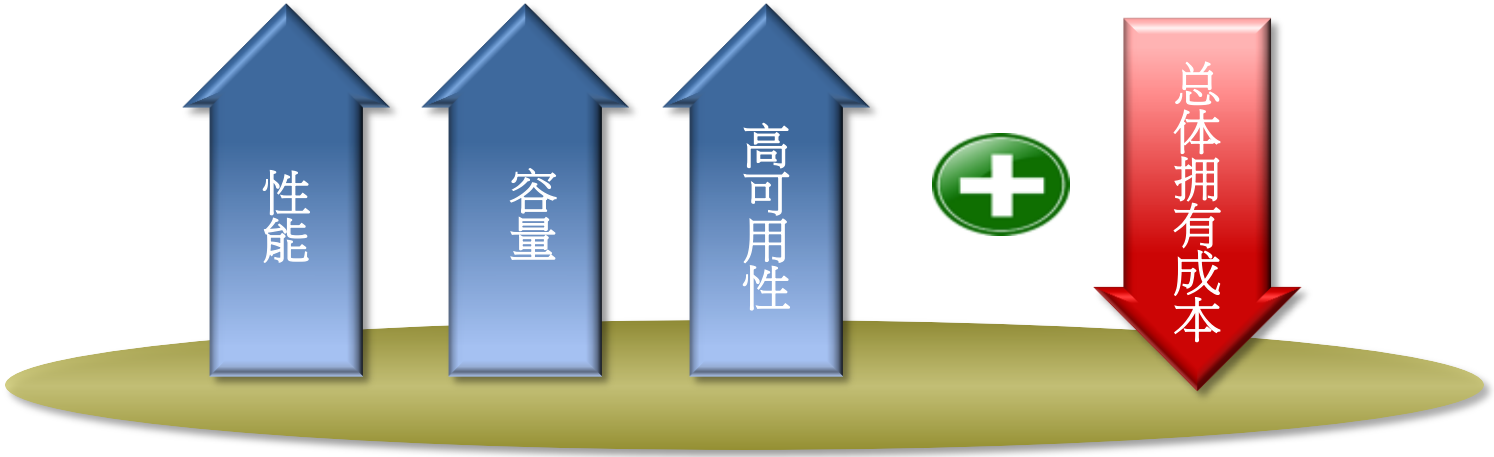
数据量不断增长，IT预算却年年控制



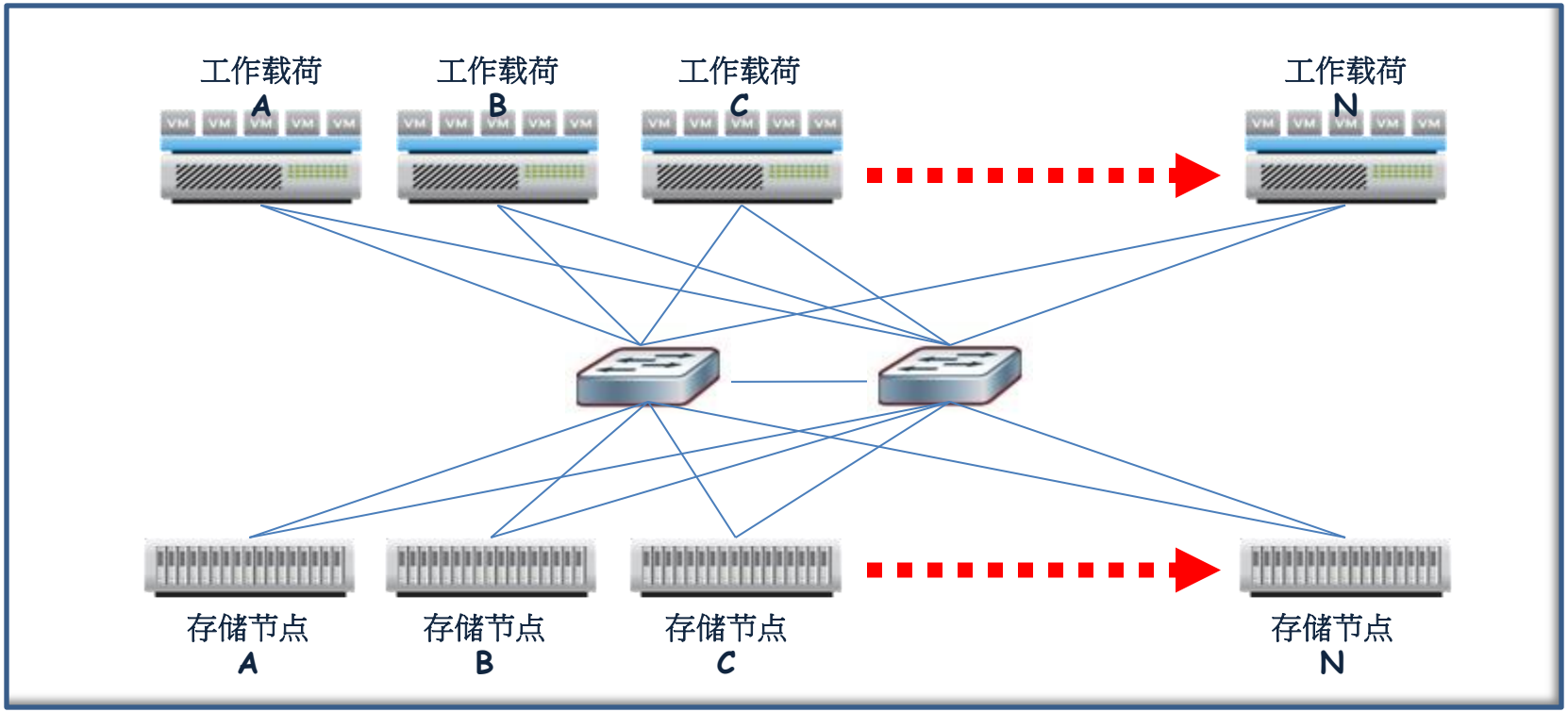
- 企业数据量每年以平均超过50%的速度增长。
- 对于存储的需求平均每年成长接近40%。
- 存储的成本每年减少25%。
- IT预算仅以每年3%~6%的速度增长。

数据存储解决方案的新规范

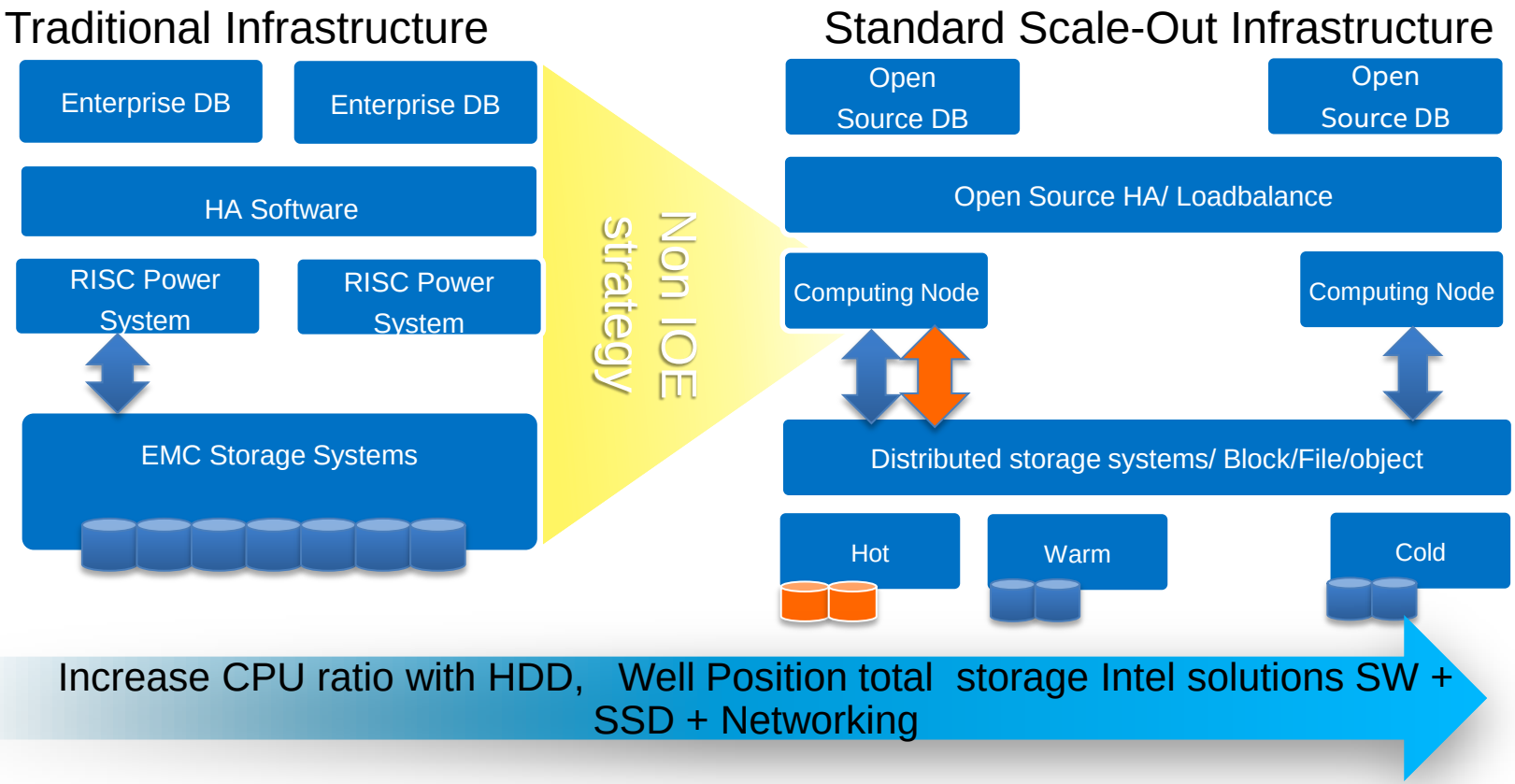
如何用非破坏性的方式来提高系统的存储容量和性能，而同时却能降低系统的总体拥有成本(TCO)?



答案：水平扩展(Scale-Out)的存储方案



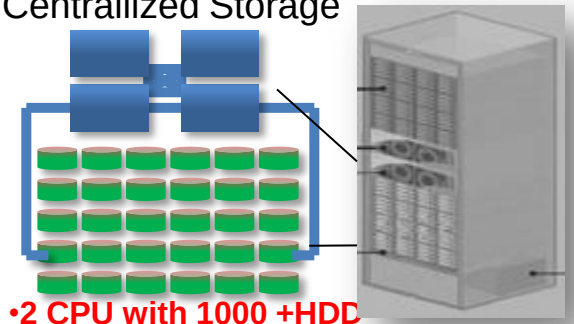
Re-Arch Datacenters – Non IOE Strategy



Storage Segmentation

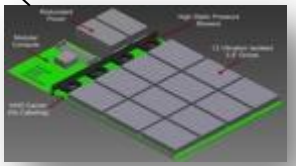
Storage Segmentations

Centrailized Storage



•2 CPU with 1000 +HDD
•CPU: HDD 1: ~100

Scale-Out Storage



•1-2U Storage Server with 8-24HDD
•CPU: HDD 1: ~10

20~300TB
IOPS ~50K

Storage Segmetation

~100+PB

Commercial Solution

Open Source Solution

采用分层的方法来降低存储的成本

热数据
经常存取，
最低延时响应

服务器层

冷数据
不经常存取，
允许较长的延时响应

数据生命周期

Intel® Cache Acceleration Software (Intel® CAS)

服务器层需要按照应用的需求提供最高的数据服务能力

Intel® Cache Acceleration Software (Intel® CAS) 的介绍

- 基于服务器的软件将 *最活跃* 的数据保存在固态盘或 PCIe* 闪存卡上

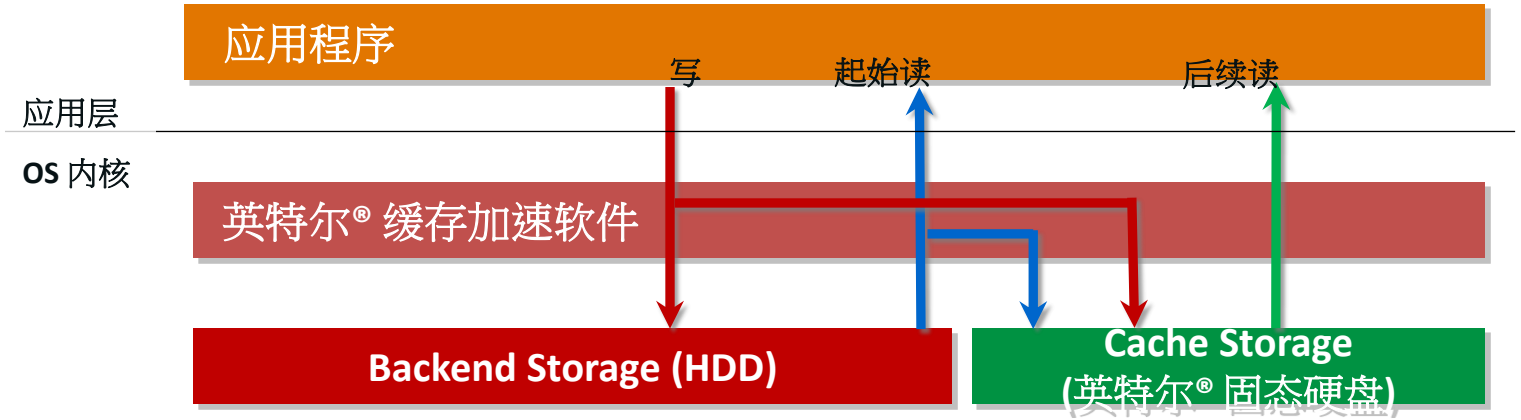
- 性能可以几乎达到将全部硬盘替换成固态盘 但是成本仍保持极大的优势[3]

*Other names and brands may be claimed as the property of others.

英特尔® 缓存加速软件逻辑结构

在服务器端，英特尔® 固态硬盘上缓存重要的和“热”的数据

- 英特尔® 固态硬盘的好处:
- 100-1000倍的 IOPS
- 10-100倍的时延减少



Intel Cache Acceleration Software (CAS) & Ceph

SSD Caching of Ceph OSD drives

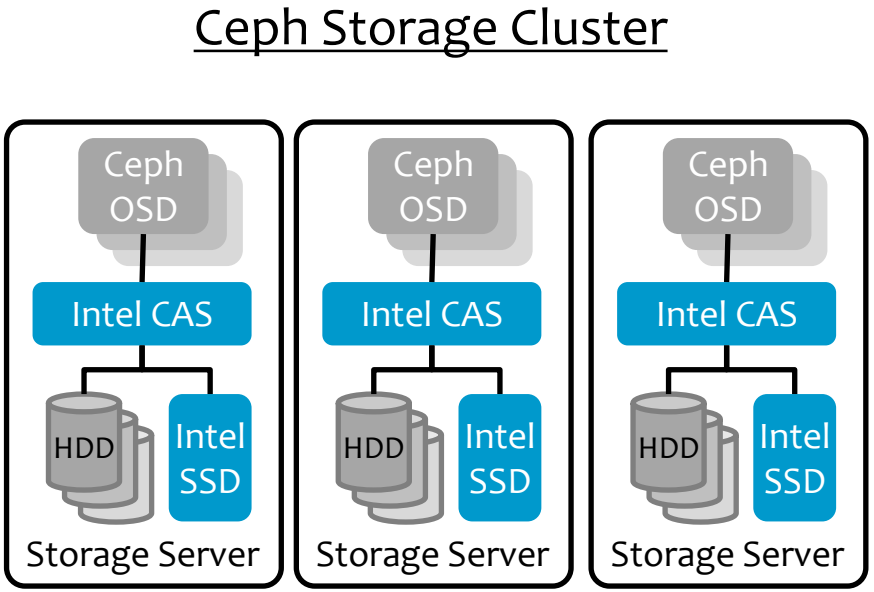
- Intel CAS caches data disks in the Ceph cluster
- CAS places "hot" OSD data on SSD

Boost Performance

- For compatible workloads (1), CAS increases the total available read IOPS and lowers the cost of storage per attached VM

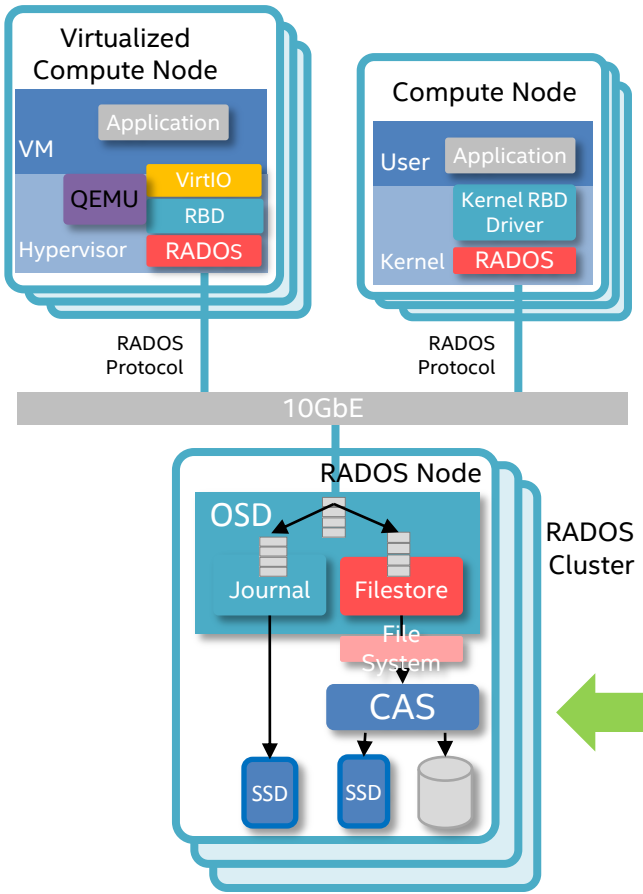
Maintain high availability

- Caching in storage nodes eliminates SSD cache as single point of failure (2)



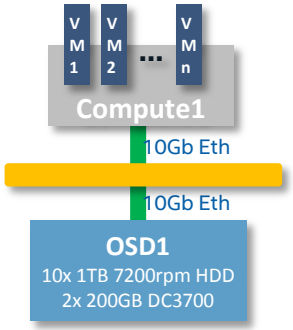
(1) Compatible workloads are aggregate workloads for which a majority of hot data fits in available SSD cache
(2) To maintain data integrity in the event of power failure, CAS operates as a write-through cache

Storage Node Caching with CAS



- HDDs in Ceph storage nodes cached by Intel Cache Acceleration Software (CAS) to Intel SSDs
- Provides large performance gains for compatible workloads (1)
- Caching in storage nodes eliminates cache SSDs as single point of failure
- Write-through cache reduces opportunity for data corruption in event of system or power failure
- SSDs continue to be used as Ceph journals, accelerating write performance
- Increases the total available IOPS and lowers the cost of storage per attached VM

HDD+ SSD Cache Cluster Test Configuration



- 10GbE Network
- One Xeon E3 servers for Ceph cluster
  - 16Gb memory (each node)
  - 10x 1TB SATA HDD for data through LSI2808 HBA (JBOD), each parted into one partition for OSD daemon
  - One journal SSD directly connected with SATA controller, 20GB for each OSD
  - Two cache SSDs – Five HDDs per cache SSD
- Single client node
  - Dual Xeon E5

Ceph Performance
HDD Array vs HDD + SSD Cache Array
180 GB Cache - 40GB Span per FIO VM – 4K Random Read

14X performance gain at 100% hit rate

5X performance gain at 90% hit rate

议程

SSD在分布式存储场景的使用

SSD概述

SSD在各个应用场景下的解决方案概述

非易失性存储是突破性的技术！

跨越太平洋
1700's ~180 天

~速度加快
1000倍

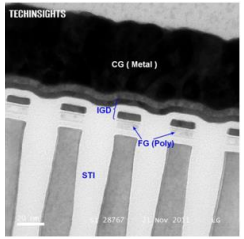
跨越太平洋
Now ~10 小时

~速度加
快500倍

穿越大西洋
光 < 30秒

硬盘

nvm EXPRESS 固态盘

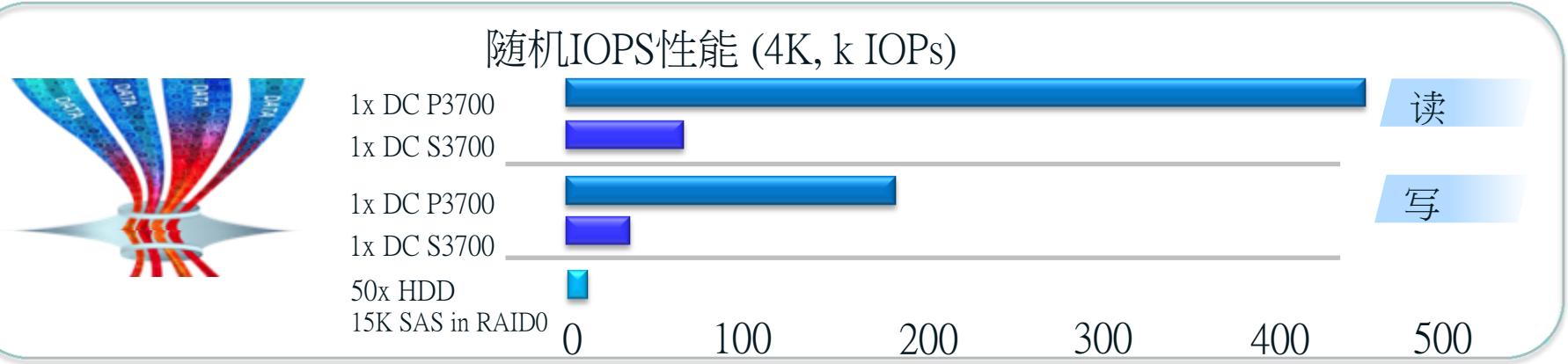


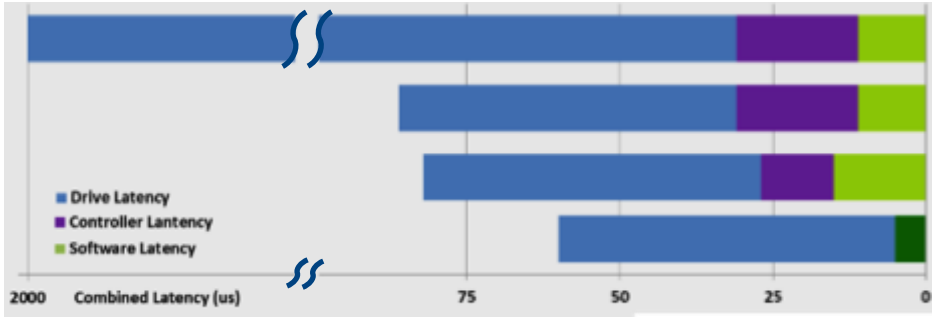
下一代非易失性存储

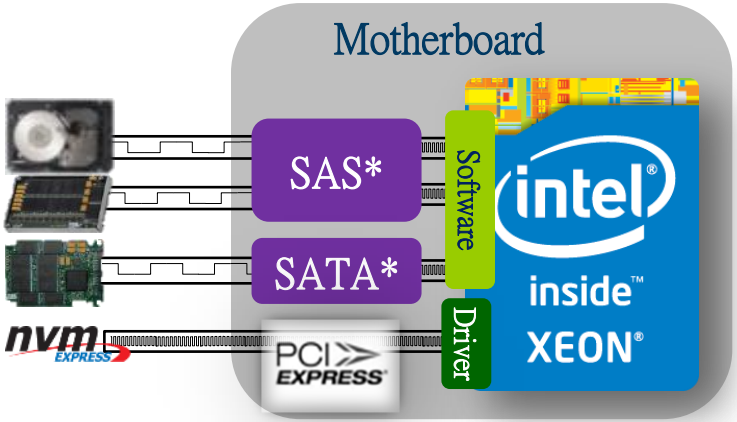
当前用于数据中心的 Intel® SSDs家族…

新一代SSD具备PCIe3.0接口和NVMe协议的非易失性存储

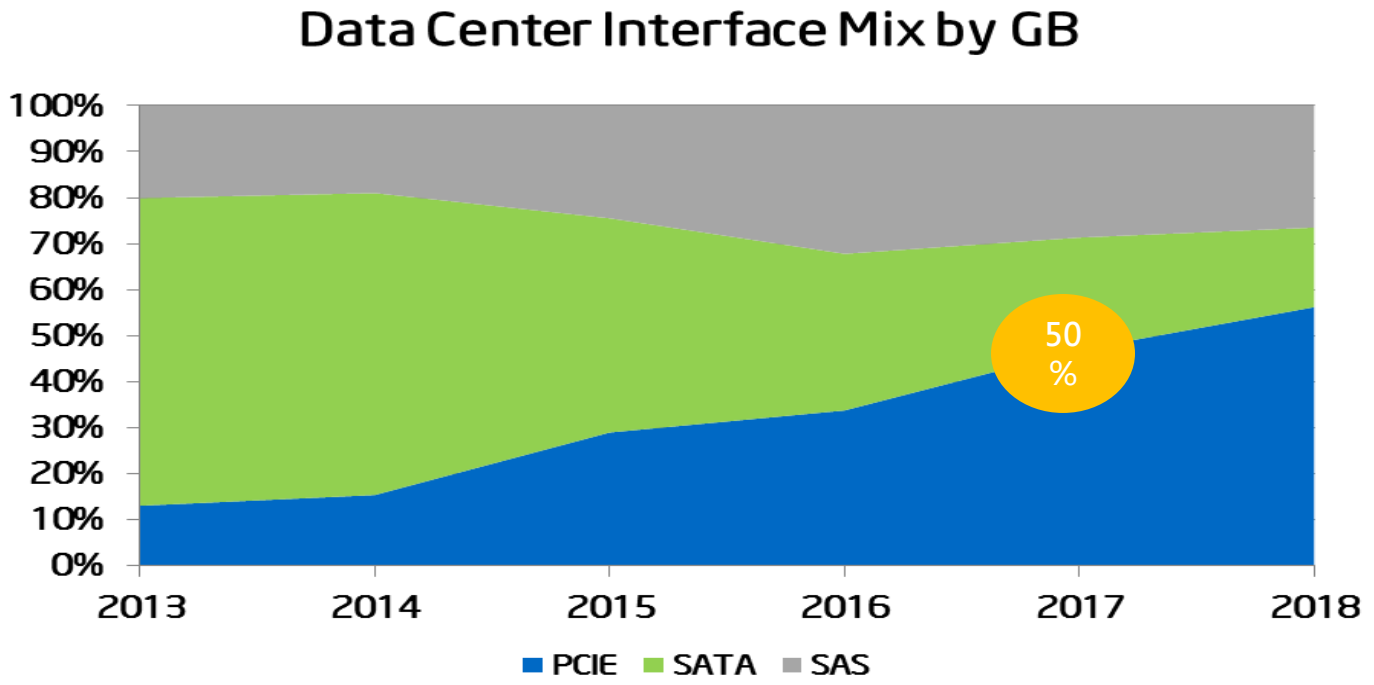
性能比较 – P3700 vs. S3700 vs. HDD







NVMe– PCIe 是下一代突破性接口



英特尔预测 ~2016年底PCIe 使用率达到50%

Source: Intel Market Forecast

议程

SSD在分布式存储场景的使用

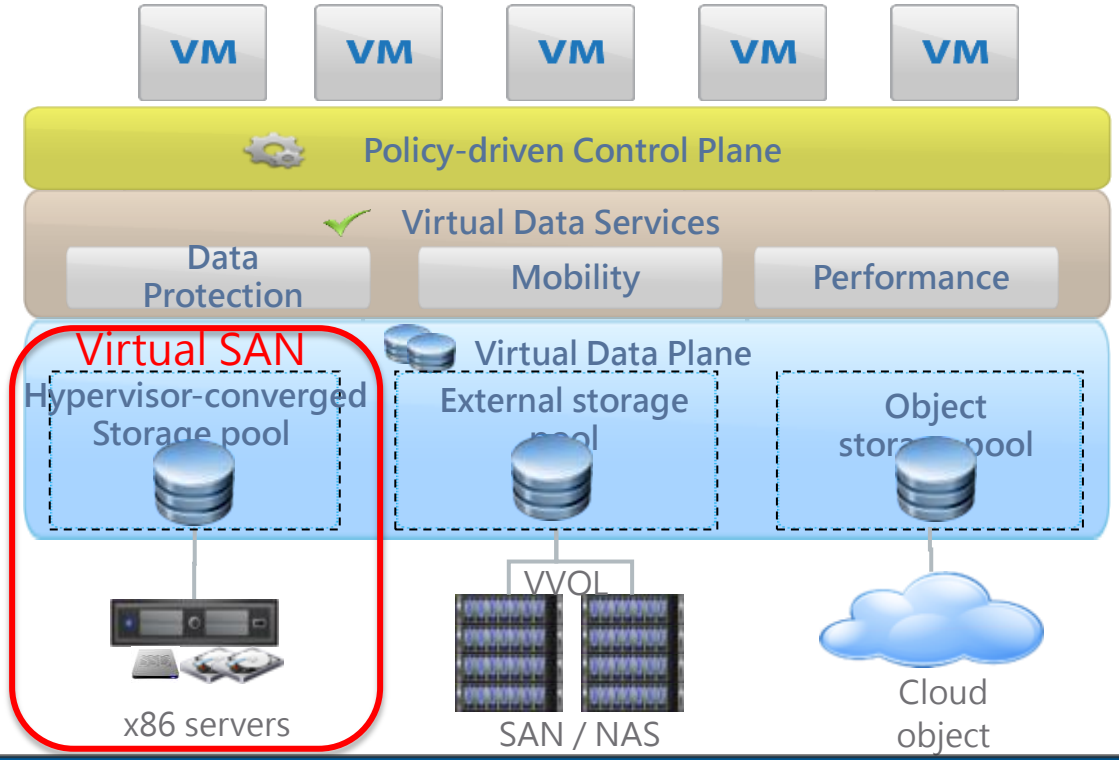
SSD概述

SSD在各个应用场景下的解决方案概述

针对各种应用场景的存储使用

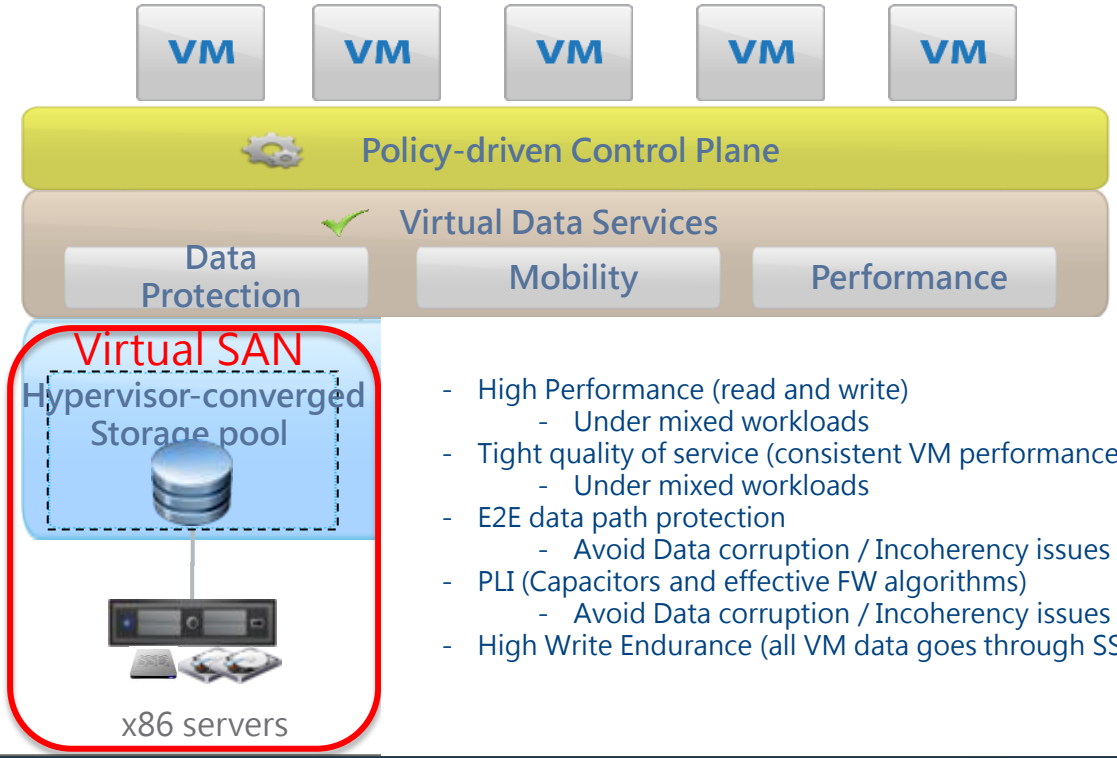
Cloud&SDS（云&软件定义存储）

SSDs in Software Defined Storage (SDS)

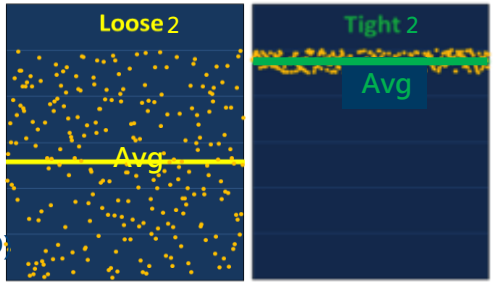


Remove expensive/slow SAN/NAS

SSDs in Software Defined Storage (SDS)



- High Performance (read and write)
  - Under mixed workloads
- Tight quality of service (consistent VM performance)
  - Under mixed workloads
- E2E data path protection
  - Avoid Data corruption / Incoherency issues
- PLI (Capacitors and effective FW algorithms)
  - Avoid Data corruption / Incoherency issues
- High Write Endurance (all VM data goes through SSD)



Valuable SSD attributes (DC S3700 supports ALL)

SSD in VMware & Virtualization（虚拟化）…

Cloud, VMware*, SW Defined Storage（软件定义存储）

- Symantec* Storage Foundations + Intel® SSD DC S3700 http://bit.ly/KKHAQd
- Openfiler* + Intel® SSD DC S3700 https://iref.intel.com/GetDoc.aspx?RefLibObjectID=0902007c8002080e
- Nexenta* + Intel® SSD DC S3500 https://iref.intel.com/GetDoc.aspx?RefLibObjectID=0902007c8002392f
- VMware VSAN* + Intel® SSD DC S3700 & Intel® SSD DC S3500
  https://iref.intel.com/GetDoc.aspx?RefLibObjectID=0902007c800207c7
- VMware VFRC* & “Swap to Host Cache”
  http://www.vmware.com/files/pdf/vSphere_55_Flash_Read_Cache_Whats_New_WP.pdf
- KVM* & Xen* Hypervisors – Ephemeral (local) Storage – Caching in Shared Block Storage

Case: Delivering SDS: Storage Foundation* + Intel® SSD

Symantec™
Storage Foundation*
Clustered File System 6.1*

+

Intel® SSD DC S3700

Results:

- 4x performance over EMC* VMAX 20k
- 90% Oracle* log transactions <1ms
- >80% cost reduction + high availability
- CFS 6.1* & TPC-like benchmark

Symantec Storage Foundation High Availability solutions provide end-to-end visibility and optimization across physical, virtual and heterogeneous environments. Symantec solutions are trusted by the world's leading commercial banks, government agencies, financial data services, computer software, and teleco companies. *symantec.com/storage-foundation*

Intel® SSD DC S3700 provides consistently amazing latency, power-loss protection, low power consumption, validated data integrity and reliability for up to 10 drive writes per day. *intel.com/ssd*

Clustered file system breaks the lock-in of SAN

*Other names and brands may be claimed as the property of others.

http://www.symantec.com/content/en/us/enterprise/white_papers/b-storage-foundation-intel-joint-whitepaper-WP-21327636.en-us.pdf

Delivering SDS: VMware* VSAN + Intel® SSD

vmware®

Virtual SAN (VSAN*)

+

Intel® SSD DC S3700

Intel® SSD DC P3700

Results:

- Up to 15k IOPS w/SSD(4) + HDD(16)
- Up to 45k IOPS w/SSD(4) + SSD(16)
- 70/30 R/W shared-nothing scale-out
- Integrated w/VMware vCenter* & ESXi*

VMware, Inc. is an American software company that provides cloud and virtualization software and services, and was the first to successfully virtualize the x86 architecture.
vmware.com

VMware VSAN is a scale out Software Defined Storage (SDS) solution that utilizes local HDDs for storage and local SSDs for cache to create fault tolerant shared-nothing VM datastores for ESXi 5.5 & greater.

Intel® SSD DC S3700 provides consistently amazing latency, power-loss protection, low power consumption, validated data integrity and reliability for up to 10 drive writes per day.
intel.com/ssd

Scale out Software Defined Storage (SDS)

intel

Acceleration Options VMware* + Intel® SSD

vmware® + Intel® SSD DC S3700

Intel® SSD DC P3700

VFRC – Virtual Flash Read Cache
'Swap to Host Cache'

VMware, Inc. is an American software company that provides cloud and virtualization software and services, and was the first to successfully virtualize the x86 architecture. *vmware.com*

VFRC and 'Swap to Host Cache are two features in ESXi 5.1+ that allow specific IO to be migrated from shared storage (SAN) to local SSDs in order to speed VM IO operations.

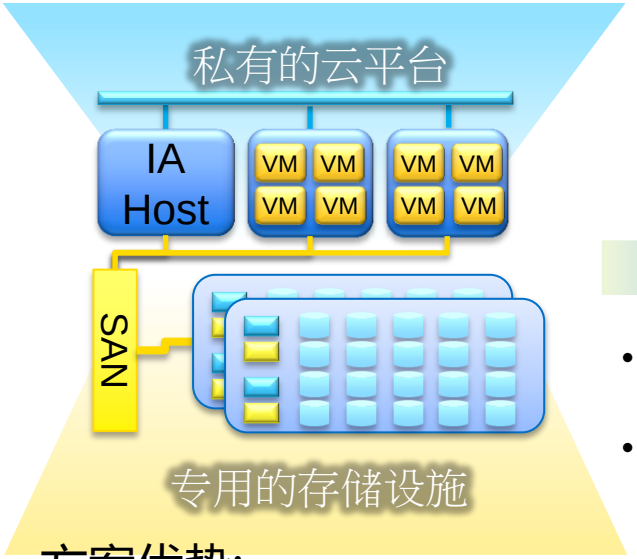
Intel® SSD DC P3700 provides the same consistently amazing features as the Intel® SSD DC S3700 with a PCIe interface, NVMe protocol for superior latency and up to >6x more performance than SATA based SSDs. *intel.com/ssd*

Use Case:

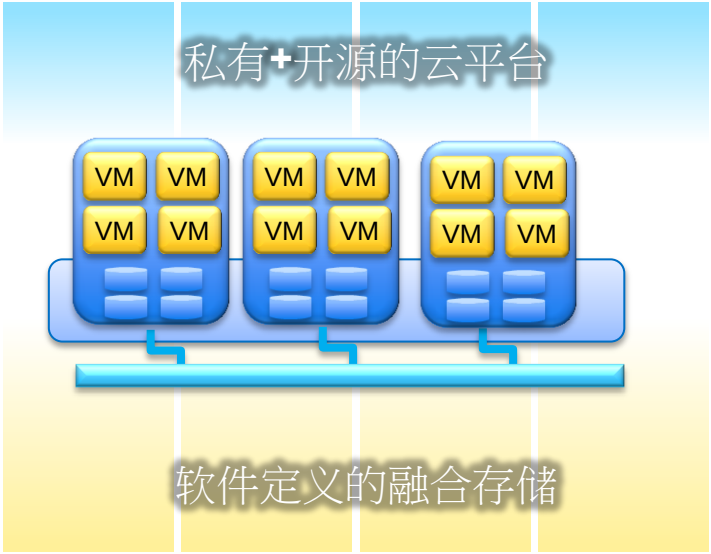
- VFRC acts as local high speed read cache for VMs – frees up SAN
- 'Swap to Host Cache' - local high speed storage for .vswap file – frees up SAN

Augment VM IO with Intel® SSDs – Works with vMotion*

一体化私有云解决方案



- 计算与存储同时虚拟化
- 一种类型的标准服务器堆叠扩展



方案优势:

- 免除专用的SAN存储设备，借鉴互联网成功经验，利用Intel平台的基础构件深度调优，为最终用户极大节省成本
- 消除了计算和存储间的带宽瓶颈，更好的可管理性和水平扩展特性
- 一站式交付，集成Intel平台的最先进技术，省电省空间，易于集成 “应用一体机”

“怡德数码的虚拟桌面架 构在引入英特尔® 固态盘之 后，相对于其他固态盘 产品，并发处理能力提高1/3 以上1，有力地应对教学 过程中使用虚拟桌面基础 架构应用时可能产生的高IO 冲击；整体解决方案配合英特尔固态盘带来的卓越的 稳定性也大大减轻了我们的部署和维护成本。我们的教师和学生对全新解决方案带来的体验都十分满意.

信息中心部主任
崔然

使用传统方案面临的挑战

中国科学技术大学

- 使用机械盘所支持的虚拟用户数量有限，每块盘建议只配置5个虚拟用户，
- 每个虚拟用户常规分配15-20个IO, 而在启动风暴时,对IO 的需求一般在300个IO 左右，严重影响了系统启动及运行
- 中科大某班级65名学生，启动Photoshop需要5分钟

解决方案

每个刀片部署800G 英特尔®固态盘（2块S3500 480G）*
其中，每个节点四个刀片，可根据客户需求灵活增添英特尔®固态盘数量

Servers + Storage + Network = 42U
- Power Consumption = 6,800W
- Max VDI users per 2U = 5个

2 Complete Blocks + Network = 5U
- Power Consumption = 2, 000W
- Max VDI users per 2U = 320个

山东科技大学

- 稳定的写性能使得减少固态盘的投入（减少50%）
- 满足启动风暴和程序编译等高IO需求，提升 60%
- 降低磁盘阵列规模，简化存储架构（1:125替换传统磁盘）
- 上述场景启动时间30秒

方案	主要设备	高度	电源	重量	数量	高度合计	支持虚拟用户	重量合计	电源合计	1年电费（万元）	3年电费（万元）
传统虚拟化	VNX 5500	4U	1650w	51.5kg	1	14U	35个	206kg	7050w	**4.18**	12.53
	x3850	4U	1350w	38.5kg	4						
基于英特尔固态盘的虚拟化	NX3450	2U	1400w	45.8kg	1	2U	320个	45.8kb	1400w	**0.83**	2.50

针对各种应用场景的存储使用

Bigdata&Enterprise App

SSD in Big Data & Enterprise Applications…

Big Data & HPC

- Hadoop* Acceleration w/SSD https://iref.intel.com/GetDoc.aspx?RefLibObjectID=0902007c800207d4
- Intel® SSG-DRD CRT Datacenter – Scratch 100% SSD + Lustre @ 40GB/Sec
- Intel® /Livermore* “Catalyst” Supercomputer http://1.usa.gov/1awPWou
- Intel® IT Design Compute Acceleration http://intel.ly/1gZLnER

Accelerating Big Data: Hadoop* + SSD

Use Case:

- Testing at NSG/DCG/CMG in progress
- SSD as temp/intermediate data storage can speed Hadoop jobs
- Add 1x SSD/Host – modify .xml file
  - mapred.local.dir in mapred-default.xml config

Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware. Hadoop is an Apache top-level project being built and used by a global community of contributors and users.*. *projects.apache.org - hadoop*

Intel® SSD DC P3700 provides the same consistently amazing features as the Intel® SSD DC S3700 with a PCIe interface, NVMe protocol for superior latency and up to >6x more performance than SATA based SSDs. *intel.com/ssd*

Accelerating BigData architectures with Intel® SSDs

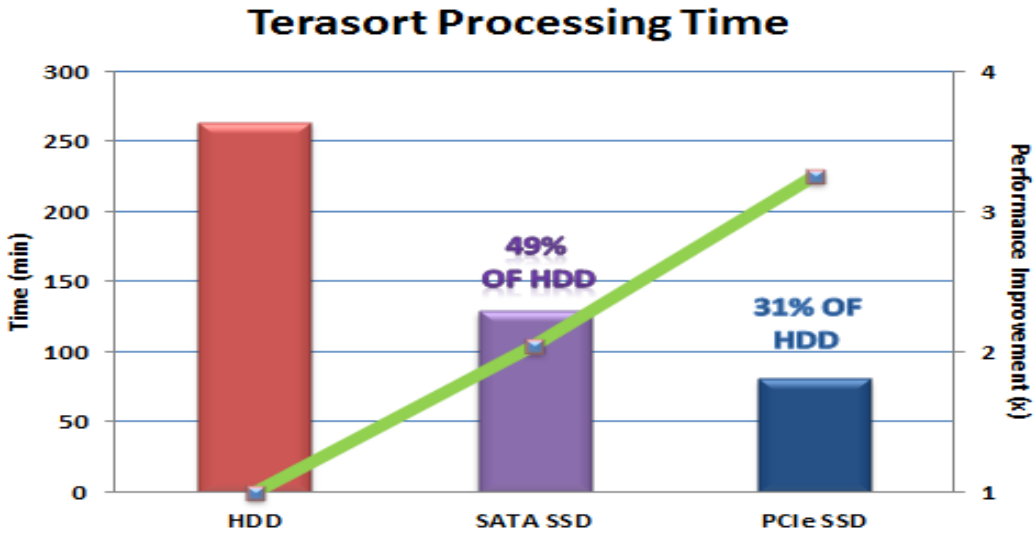
Hadoop基准测试数据

处理时间的节省

- 49%，采用SATA SSD
- 31%，采用PCIe SSD

性能提升

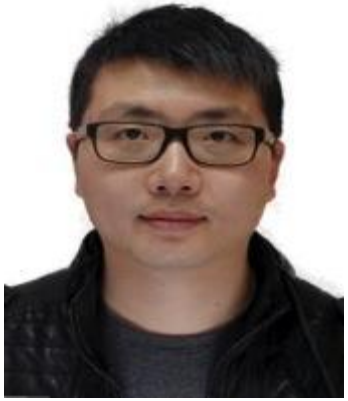
- 2x，采用SATA SSD
- 3x，采用PCIe SSD



每个host添加一个SSD，存储YARN临时数据(Temp)

Test Configuration: 3 data nodes, each has:
Ivy Bridge, 128GB RAM, 10 GbE. 1x 500GB HHD or 1x 800GB P3700 or 1x 800 GB S3700 drive as temp/intermediate data in the cluster

浙江省疾病预防控制中心



终选择的方案经过 我们谨慎的评估和测试。 方案部署后，中心的系统性 能和工作效率有显著提升。 一方面，查询数据的速度 较原有方案相比提升了几十 倍，另一方面，能够将省市 间跨平台数据实现很好的同 步。新的方案对我们在疾病 研究工作方面助力非常大，也极大地改善了中心工作 人员的工作体验。”

浙江省疾病控制中心
大数据平台项目负责人
叶飞

挑战

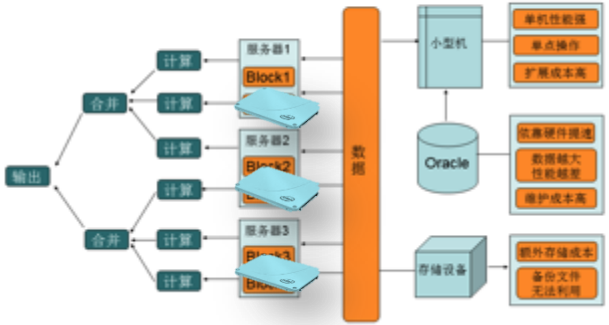
•传统方案性能差，一次多条件模糊查询需等待十几分钟，生成一份报告需要长达一周时间！
•传统方案无法承受大量访问和频繁检索，**普通固态盘稳定性难以保证**
•无法实现多个省市平台跨数据库实时同步

解决方案

基于英特尔® 固态盘的Hadoop 方案

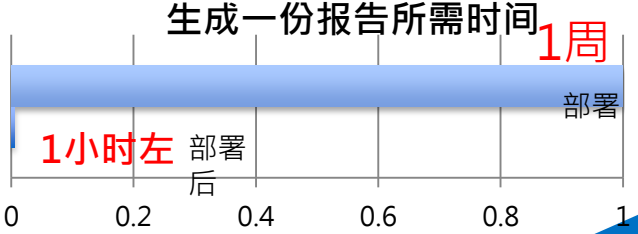
1. Hadoop**管理节点** 2台，Intel S3500系列 SSD2块
2. Hadoop**数据节点** 8台，Intel S3500系列 SSD 8块

Hadoop技术架构　　传统计算结构图



收益

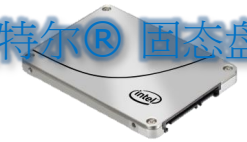
1. 千万级数据库、多条件模糊查询时间由**十几分钟**缩短至**十几秒**, 如果在数据节点增加SSD配置，则在**一秒以内。**
2. 生成一份典型研究报告时间由**1周**缩短至**1小时**
3. 多个跨数据中心**实时**同步，真正进入大数据时代



超级计算的重大突破：Lustre* + 固态盘

Lustre

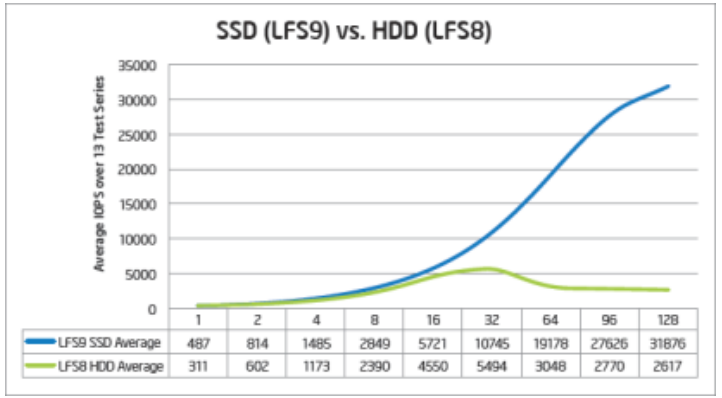
英特尔® 固态盘 DC S3500



Lustre* 是一个并行分布文件系统，主要用于大规模集群计算。针对 Lustre* 的英特尔® 企业版软件对针对 Lustre* 的英特尔® 管理器的安装、配置和监控特性进行了简化，后者是一款专为 Lustre* 打造的管理解决方案。
intel.com/lustre

结果：

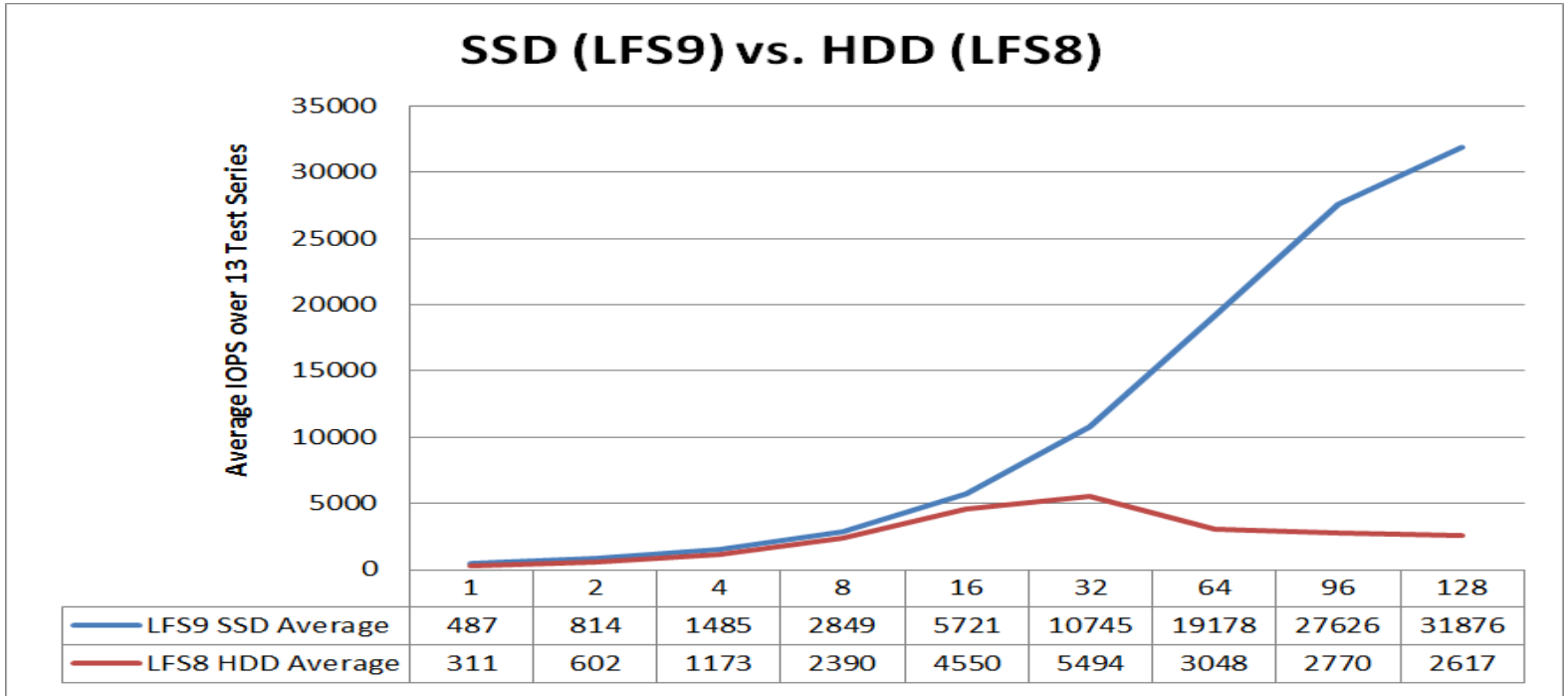
- SSD 实现 4 倍的吞吐量 @44GBps
- 基本成本较商用 HDD 解决方案低 4 倍
- 在 32x 客户端后持续交叉扩展 – Iozone*



	1	2	4	8	16	32	64	96	128
LFS9 SSD Average	487	814	1485	2849	5721	10745	19178	27626	31876
LFS8 HDD Average	311	602	1173	2390	4550	5494	3048	2770	2617

借助英特尔® SSD 在大型 HPC 集群中加速 scratch/temp

案例分析 – SSD vs. HDD方案结果比较



	1	2	4	8	16	32	64	96	128
LFS9 SSD Average	487	814	1485	2849	5721	10745	19178	27626	31876
LFS8 HDD Average	311	602	1173	2390	4550	5494	3048	2770	2617

针对各种应用场景的存储使用

Database application

SSD in Database（SSD在数据库方面的使用）...

Database

- IBM* TEM (Security management) MSSQ DB and Intel® IT http://intel.ly/1aOFEMZ
- Intel® SSD DC S3700 for Oracle* Log Writer http://intel.ly/1fkobDJ
- Oracle* TimesTen + Intel® SSD DC S3700 http://intel.ly/1ewyjEm
- Aerospike* Acceleration w/Intel® SSD http://bit.ly/1hF8p6w
- Top SSD use cases（主要的SSD使用场景）:
  - Log Writer（写日志操作） – Traditional & In Memory
- Cache/B-Cache/CAS & TempDB (Sort)：利用一些cache软件配合使用
- SQL – Buffer Pool Extension (Classic SQL)：2014版本的MS SQL具备的功能
  - Pure SSD for IO intensive DB：纯SSD用于对IO十分敏感的数据库
- #1 Industry Trend = Move to in-memory DB：行业趋势是更多使用内存数据库

Accelerating DB: Transaction Commit

Results:

- Millisecond to Microsecond log writes
- Order of magnitude faster response

Intel® SSD DC P3700 provides the same consistently amazing features as the Intel® SSD DC S3700 with a PCIe interface, NVMe protocol for superior latency and up to >6x more performance than SATA based SSDs. *intel.com/ssd*

Accelerating traditional DB with Intel® SSDs

Accelerating DB Startup: In-Memory DB

+

TimesTen is an in-memory, relational database management system with persistence and recoverability. Acquired by Oracle Corporation in 2005. *oracle.com/timesten*

Intel® SSD DC S3700 provides consistently amazing latency, power-loss protection, low power consumption, validated data integrity and reliability for up to 10 drive writes per day. *intel.com/ssd*

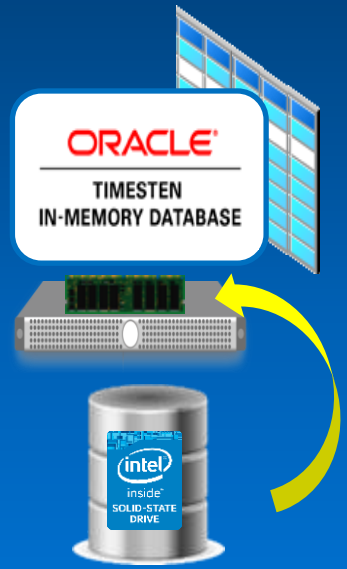
Intel® SSD DC P3700 provides the same consistently amazing features as the Intel® SSD DC S3700 with a PCIe interface, NVMe protocol for superior latency and up to >6x more performance than SATA based SSDs. *intel.com/ssd*

Results:

- Parallelized IO across 16-threads
- >4x better than SATA SSD
- Loads 100GB Database in 48sec
- Enable memory to block ops in parallel

Accelerating In-Memory DB with NVMe/PCIe Intel® SSDs

PCIe 固态盘节省时间，“时间就是金钱”



Oracle TimesTen checkpoint file loaded to memory

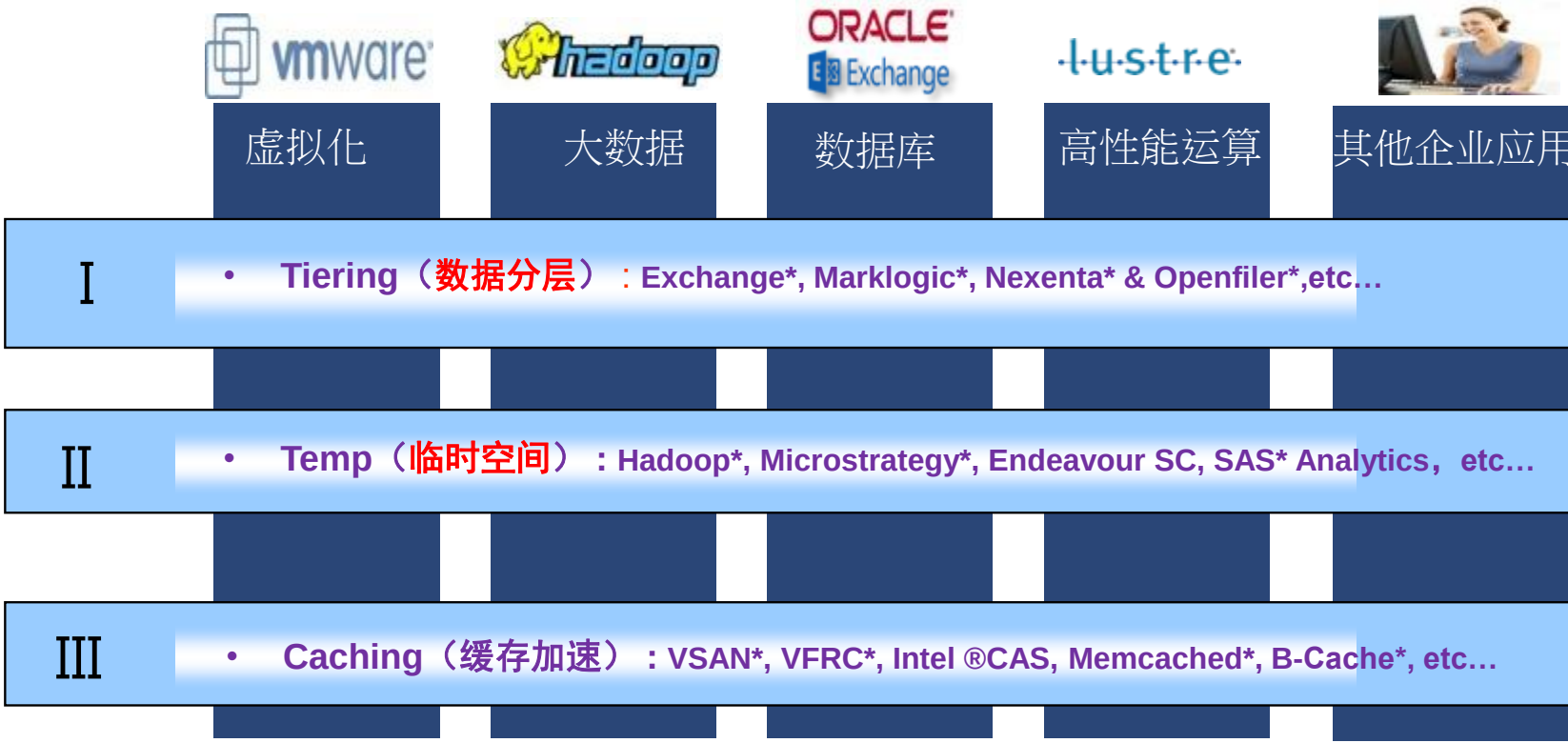
转移100GB文件的时间

31 分钟

< 1 分钟

使用英特尔®固态盘DC P3700系列使速度加快超过30倍

SSDs更多的实用案例正在越来越多部署

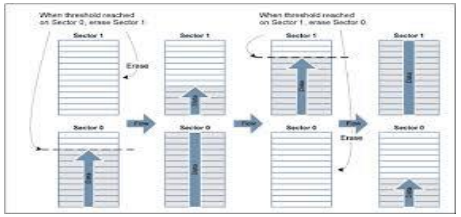


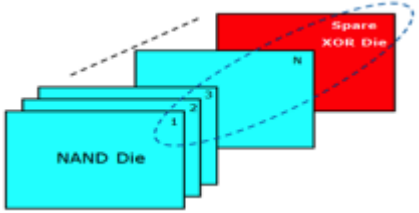
Why Choose Intel SSD?

并行优化

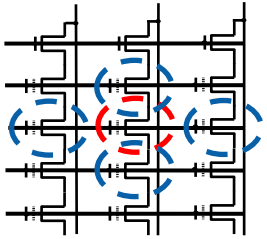
测试

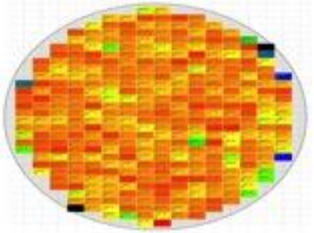
系统





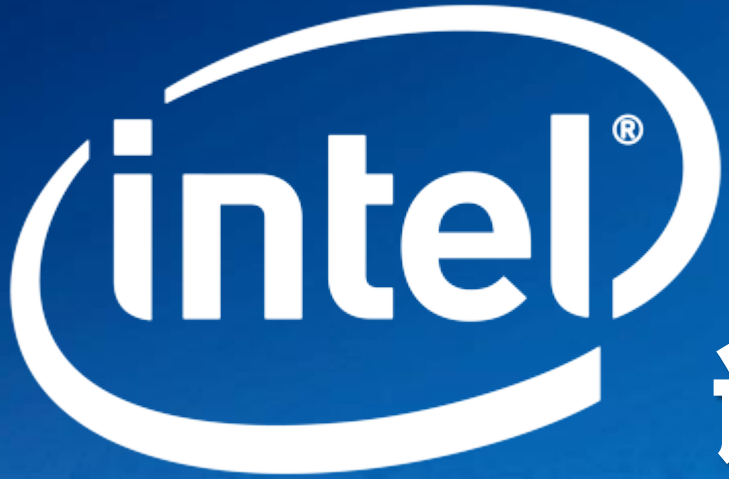
硅技术





World-class Intel® SSD Solutions

不是所有的SSD都叫数据中心级SSD



谢谢您！

英特尔®固态盘在奇虎360的应用

- **主流SSD**
  - 代表：P3500、S3500等；
  - 特点：性能偏向读为主、写持久性较弱；
  - 应用场景：搜索、云杀毒、以及缓存类应用等；
- **高性能/高寿命SSD**
  - 代表：P3700, S3700等；
  - 特点：处理高并发的读写请求、日均写入量较大、写持久性强；
  - 应用场景：CDN等；
- **注重读写平衡的SSD**
  - 代表：P3600等；
  - 特点：处理读写混合请求，性能、寿命、价格均衡；
  - 应用场景：MySQL、NoSQL等；

Intel® SSD DC S3700 & S3500 Series

	Intel® SSD DC S3700 Series	Intel® SSD DC S3500 Series
Form Factor	2.5” & 1.8”	2.5”, 1.8” & M.2 80mm
Capacity	2.5” – 100, 200, 400 & 800GB 1.8” – 100, 200 & 400 GB	2.5” – 80, 1- 800GB, 1.2TB & 1.6TB 1.8” – 80, 240, 400 & 800GB M.2 – 80, 120, 340GB
Performance	500/460MBps R/W Sequential 75k/36k IOPS 4K R/W Random	500/400MBps R/W Sequential 75k /11k IOPS R/W Random
Power Max/Idle	2.5” – 6W / 650 mW Typical 1.8” – 5.3W / 650 mW Typical	2.5” – 5.0W / 650 mW Typical 1.8” – 5.2W / 650 mW Typical
Security	End-to-End Data Protection Power Loss Data Protection AES 256b encryption	End-to-End Data Protection Power Loss Data Protection AES 256b encryption
Endurance	10 drive writes Per Day for 5 years	.3 drive writes Per Day for 5 years
Warranty	5 year	5 year

Today

Today

*HET

Q3’14

Intel® SSD DC P3700, P3600 & P3500 Series

	Intel® SSD DC P3700 Series	Intel® SSD DC P3600 Series	Intel® SSD DC P3500 Series
Form Factor	2.5" / AIC	2.5" / AIC	2.5" / AIC
Capacity	2.5": 400, 800GB 1.2TB, 1.6 and 2.0TB AIC: 400, 800GB 1.2TB, 1.6 and 2.0TB	2.5": 400, 800GB 1.2TB, 1.6 and 2.0TB AIC: 400, 800GB 1.2TB, 1.6 and 2.0TB	2.5": 250, 500GB, 1TB & 2TB AIC: 250, 500GB, 1TB & 2TB
Performance	2.8/1.7GBps R/W Sequential 450/150k IOPS R/W 4K Random	2.8/.85GBps R/W Sequential 450/70k IOPS R/W 4K Random	2.8/.65GBps R/W Sequential 450/40k IOPS R/W 4K Random
Power Max/Idle	25W / 10W Typical	25W / 10W Typical	25W / 10W Typical
Security	End-to-End Data Protection Power Loss Data Protection AES 256b encryption	End-to-End Data Protection Power Loss Data Protection AES 256b encryption	End-to-End Data Protection Power Loss Data Protection AES 256b encryption
Endurance	10 drive writes Per Day for 5 yrs	3 drive writes Per Day for 5 yrs	0.3 drive writes Per Day for 5 yrs
Warranty	5 years	5 years	5 years

Q3'14

*HET