HUAWEI ENTERPRISE ICT SOLUTIONS A BETTER WAY

华为分布式存储技术与应用实践

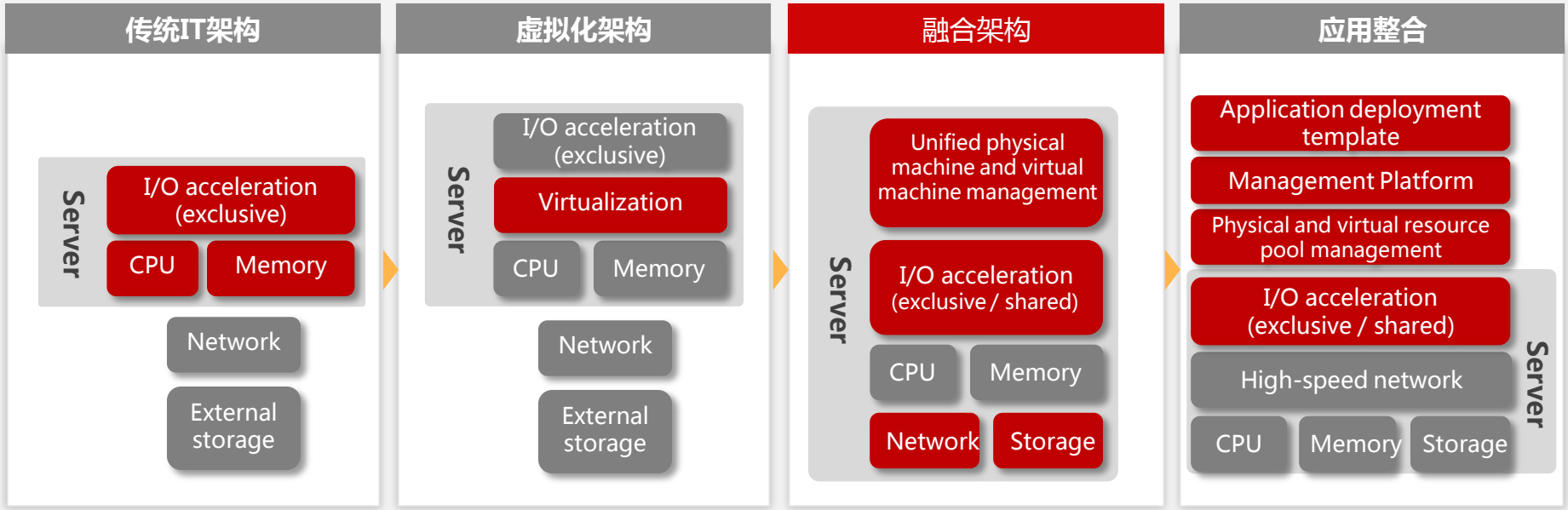
Author: 陈坚
Version: V1.0(201404)

enterprise.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.

HUAWEI

Content

IT 架构演进趋势



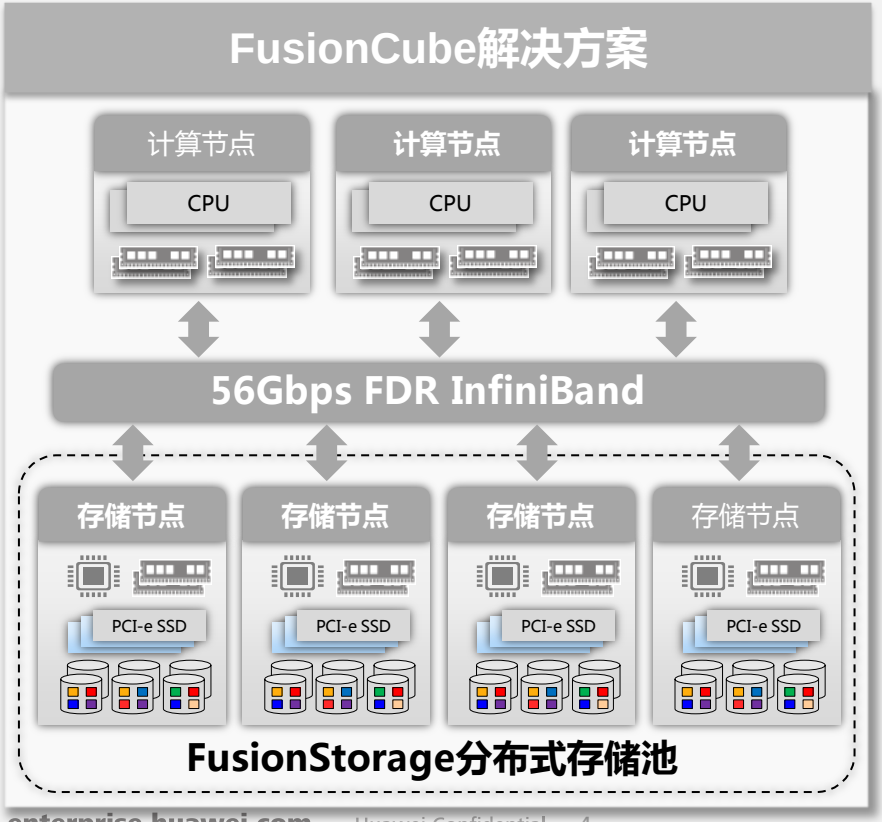
IT架构演进方向：计算、存储架构融合；资源统一管理；业务按需部署

HUAWEI

创新公司不断涌现，传统IT软硬件厂商也纷纷加入计算存储融合阵营

HUAWEI

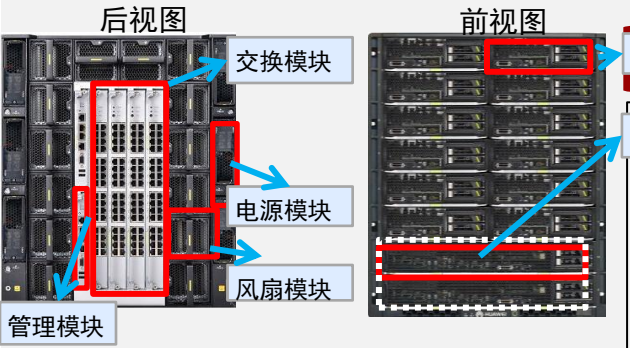
华为解决之道：计算存储融合架构一体机FusionCube



- **融合架构**：计算网络存储融合设计，计算刀片和存储刀片灵活配置，大内存，内置GE/10GE/IB多协议交换板
- **FusionStorage**：Scale-Out架构，计算存储深度融合，分布式存储解决集中式机头的瓶颈
- **无阻塞IB交换**：高速互联，降低时延，提高带宽，提高数据库多节点横向扩展能力
- **PCI-E SSD**：作为主存，提升随机读写IO能力，优化数据读写模型

HUAWEI

FusionCube：业界领先的计算、存储、交换组件



- 单刀片：未来四代CPU；756G~1.5T内存；15块硬盘；4 PCIE标准扩展卡；
- 网络：GE/10GE/IB 40G/IB 56G交换；15.6Tbps无源背板；
- 存储： 无须外置San存储；3~5倍 IOPS;
- 单机框64颗cpu

计算节点


CH121
计算节点


CH221
IO扩展型计算节点


CH223
IO扩展型计算节点


CH222
存储扩展型计算节点


CH240
计算节点

交换模块


CX110
GE交换模块


CX116
GE直通模块


CX310
10GE交换模块


CX311
10GE/FCoE/FC融合交换模块


CX317
10GE直通模块


CX610
Infiniband QDR/ FDR融合交换模块


CX911
10GE/FC多平面交换模块

HUAWEI

Content

数据仓库性能瓶颈在于IO吞吐

传统架构

可视化图表

OLAP过程

读写瓶颈

写瓶颈

时延瓶颈

内存限制

读瓶颈

Avg/Sum

Group

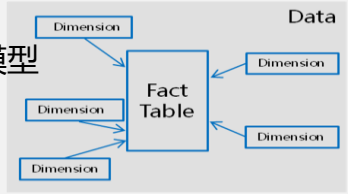
Join

需要从磁盘读取大量的临时数据，，瓶颈在于网络和磁盘IO

由于内存大小限制，需要将Group操作的临时数据写入磁盘，瓶颈在于网络和磁盘IO。

全表扫描，瓶颈在于磁盘IO

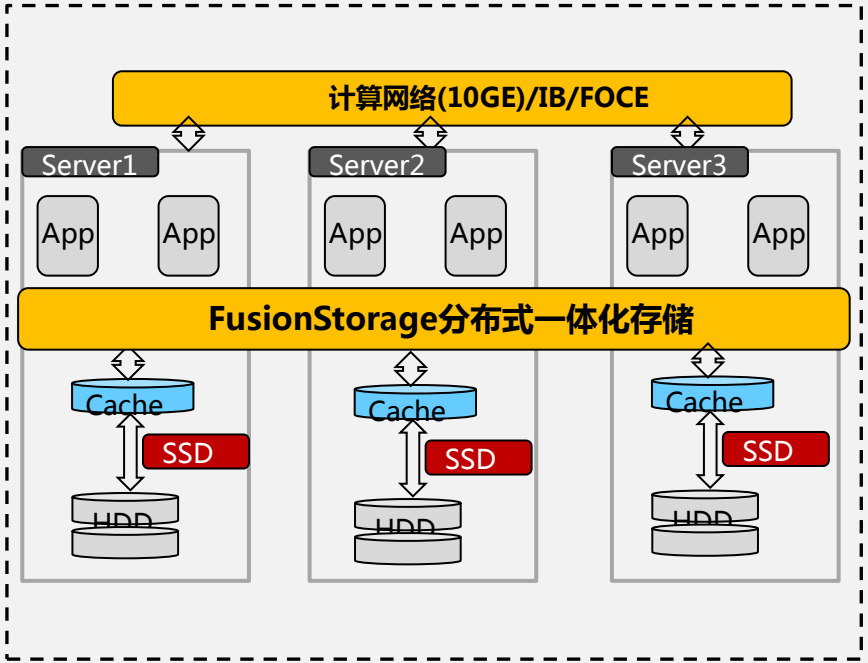
星型/雪花型模型



数据仓库主要的瓶颈是计算和存储节点间的网络IO和主存的磁盘IO!

HUAWEI

华为分布式存储FusionStorage主要特点

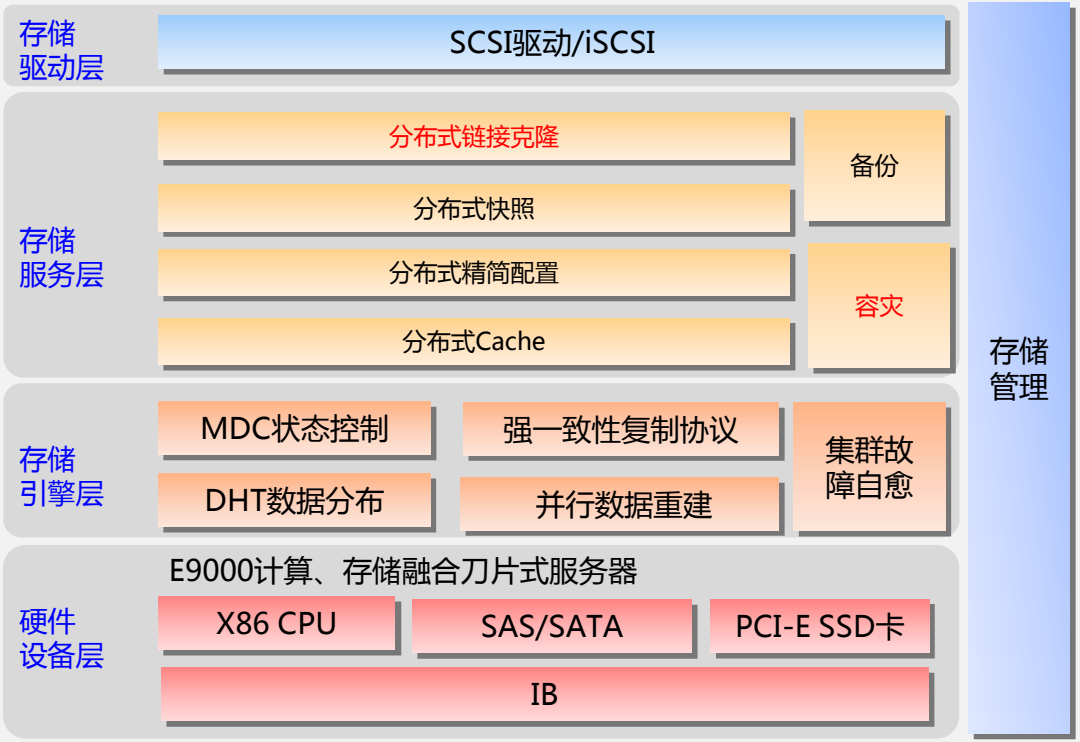


FusionStorage 分布式存储系统

主要特点

- **水平扩展、超大容量：**分布式系统，无管理机头瓶颈，容量几乎不受限制
- **高IOPS：**应用大容量分布式Cache技术，提升IOPS
- **低时延：**应用程序通过Cache/SSD直达存储，时延更低
- **数据重建快：**并行重建，重建数据量小
- **管理简单：**结构简单带来管理简单

FusionStorage 总体架构



存储接口层：

通过SCSI驱动接口向操作系统、数据库提供卷

存储服务层：提供各种存储高级特性，如快照、链接克隆、精简配置、分布式cache、容灾备份等

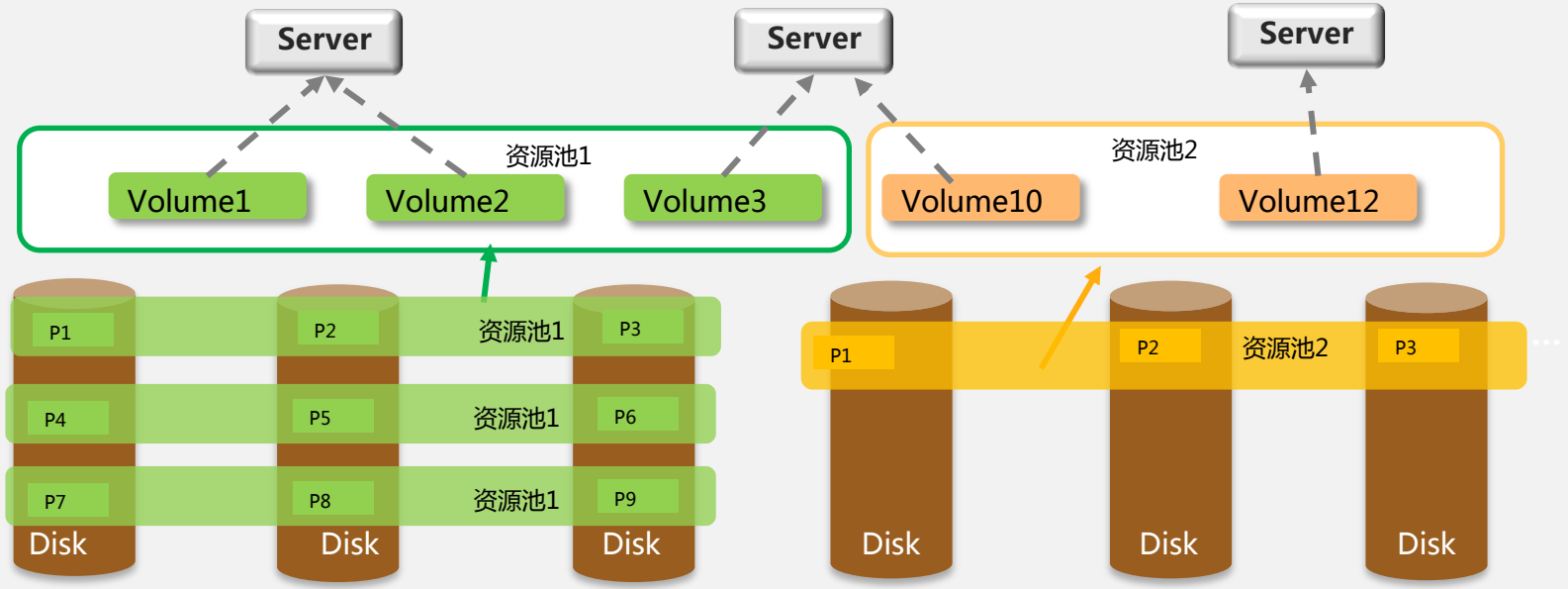
存储引擎层：

FusionStorage存储基本功能，包括MDC总控集群、DHT数据路由、分布系统、强一致性复制协议；及在单节点故障时，集群故障自愈与并行数据重建子系统

硬件设备层：

基于E9000计算、存储融合刀片式服务器，无需外置SAN，支持IB高速交换、PCI-E SSD卡

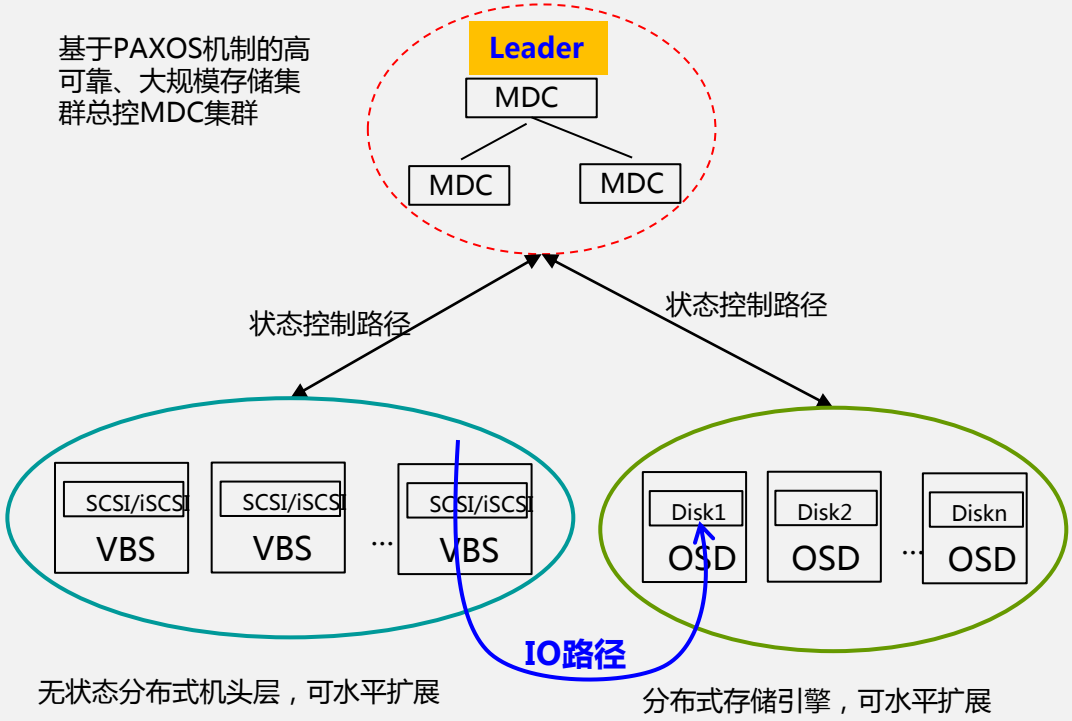
FusionStorage 基本原理 - 卷映射



资源池: 类似于SAN的RAID组概念，与RAID相比，其优点是：

- **大容量**： 最大96块盘，提升超大存储空间，避免高IO应用导致热点瓶颈
- **动态热备**：所有硬盘都可用作资源池的热备盘
- **简单结构**：资源池、Volume二层结构，没有LUN结构，服务器直接看到Volume

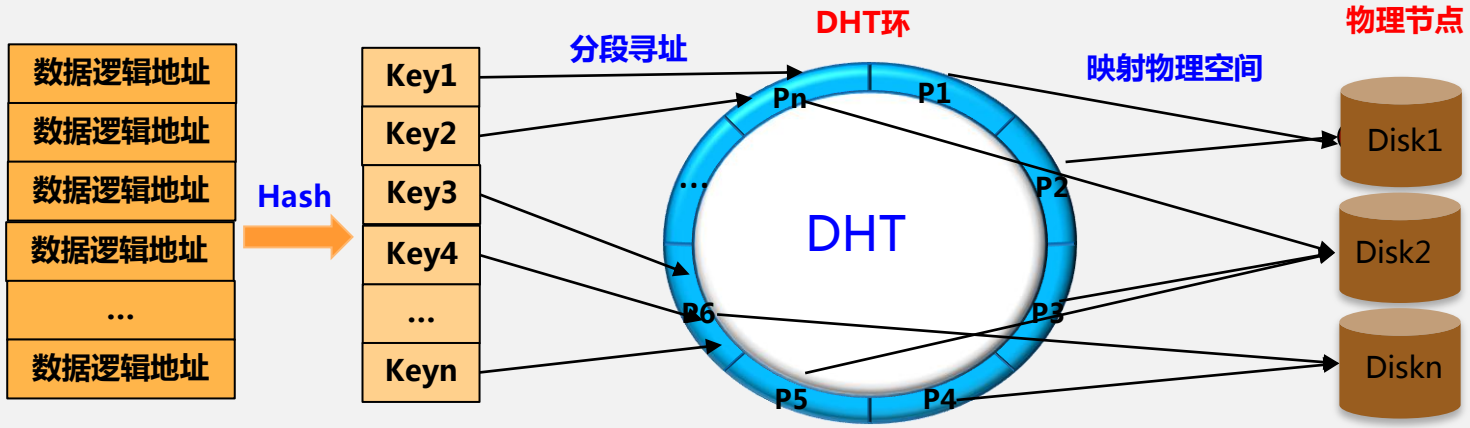
FusionStorage 分布式软件架构



- **全分布式架构，水平扩展**：无状态机头层，每个机头可以平滑添加与减少；
- **无状态分布式存储引擎：** 可以水平扩展单板、磁盘
- **计算、存储全融合架构，超高性能**：Cache更大，不再受到传统SAN机头限制
- **高吞吐量，不再有机头瓶颈：** IB/10GE/FOCE并发支持，带宽是传统SAN的10倍以上

HUAWEI

FusionStorage 基本原理 - DHT寻址与水平扩展技术



DHT（Distributed Hash Table)

DHT环：2^{32}超大虚拟节点构成的环形空间

Partition：将DHT环空间划分为N等份，每一等份是一个分区

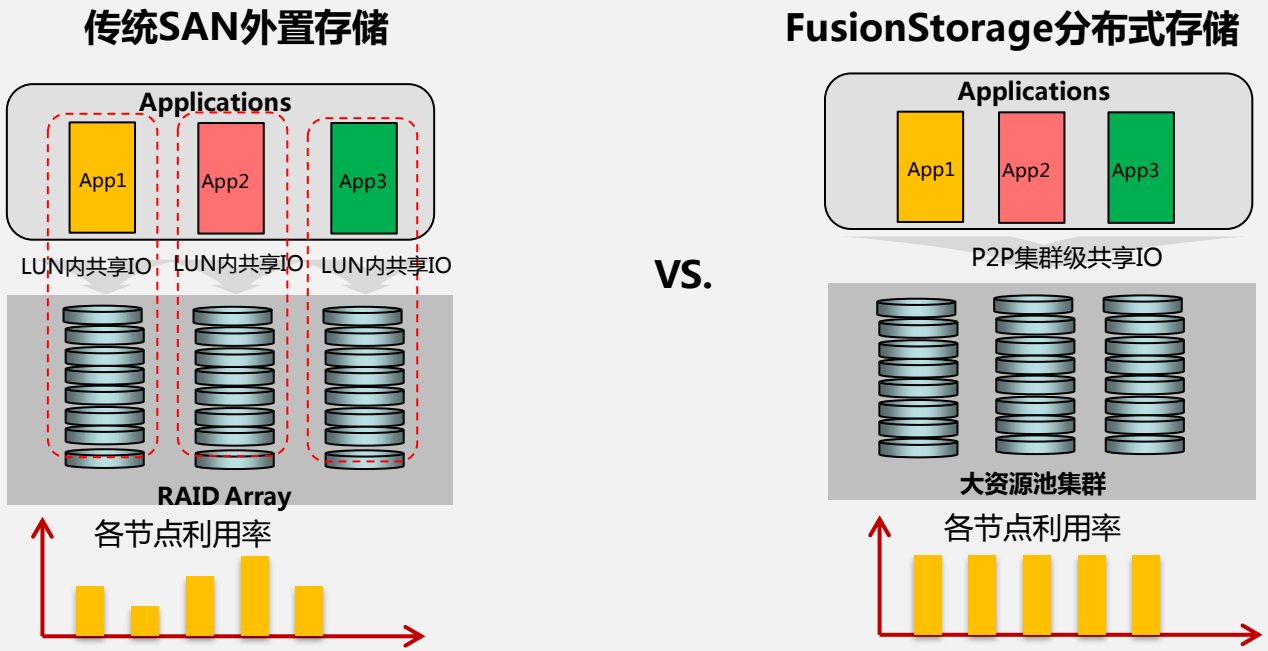
物理节点：即一个DISK， 与Partition分区对应

优点:

水平扩展速度快： 新物理节点加入时， 只需要搬移部分数据（partition），并达到负载均衡

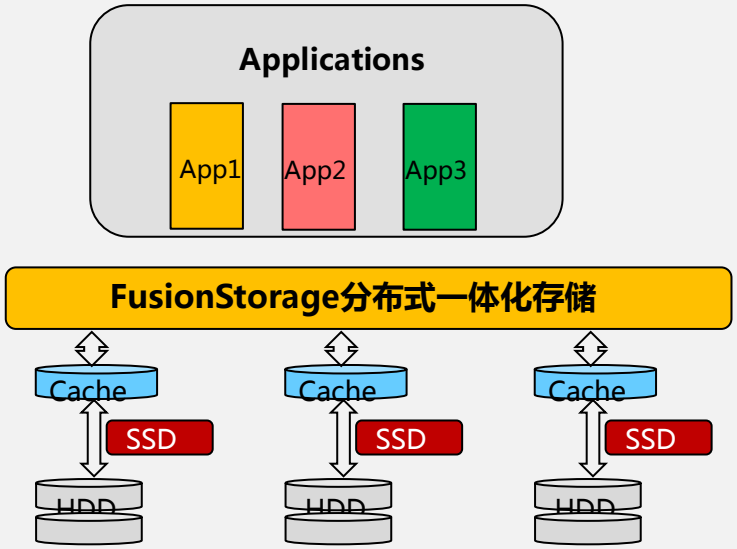
数据可靠性高：可灵活配置的分区分配算法，避免2个副本位于同一个Disk、同一块板、同一个机柜

高性能——DHT并行IO读写



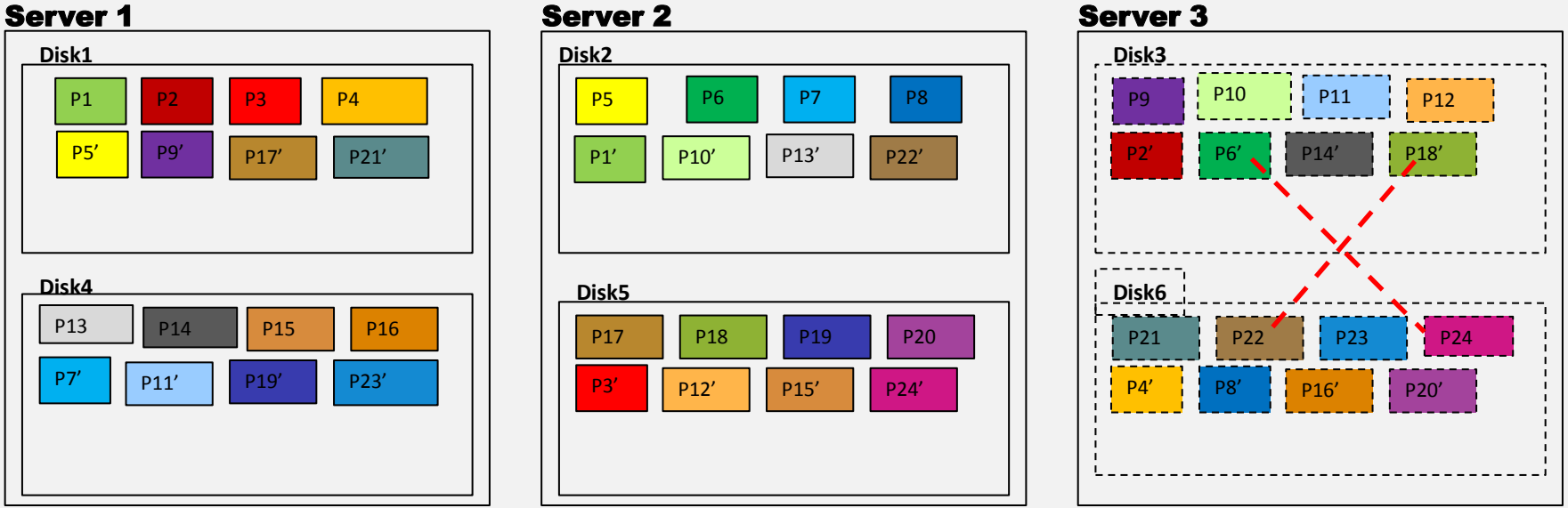
- 分布式存储架构（FusionStorage)及基于ETH/IB的点对点互联网络，不再有带宽瓶颈
- 更多硬盘在扁平P2P架构下实现为同一App实例或VM提供并发读写服务，使得突发MBPS提升3-5倍以上；
- 更大资源池， 负载均衡，利用率更高

高可靠——多重数据安全保障机制



- **多副本备份：**根据安全级别可灵活配置1副本（相当于RAID10）或多副本（3副本情况下，数据可用性达到7个9以上)；
- **NVDIMM Cache技术：**读写速度快，掉电数据不丢失；
- **强一致性复制协议：**应用程序写入一份数据时，如果成功，后端的一份或多份副本必然一致，再次读时，无论从哪个副本都可读到正确的数据；
- **数据高可用：**可以跨服务器或跨机柜分布数据，不会因某个服务器、或者某个机柜故障导致数据不可访问；

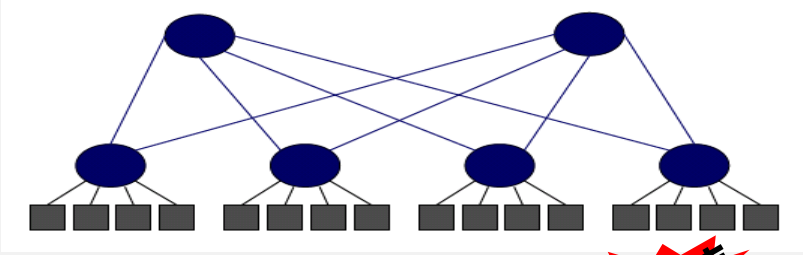
高可靠——并行、快速数据重建

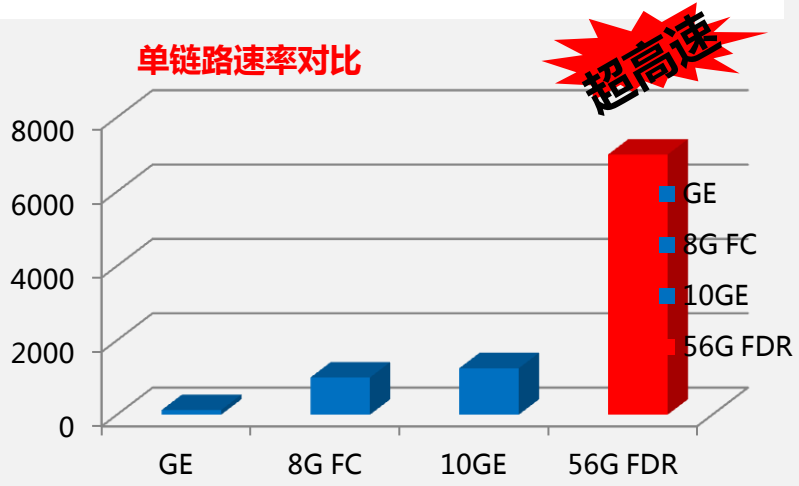


- 数据分布可以跨服务器或跨机柜，不会因某个服务器、机柜故障导致数据不可访问
- 数据分片在资源池内打散，硬盘故障后，可在全资源池范围内自动并行重建，仅重建实际数据，无需热备盘；

重建1TB数据时间 < 30分钟(传统IPSAN 重建1TB数据需要12小时)

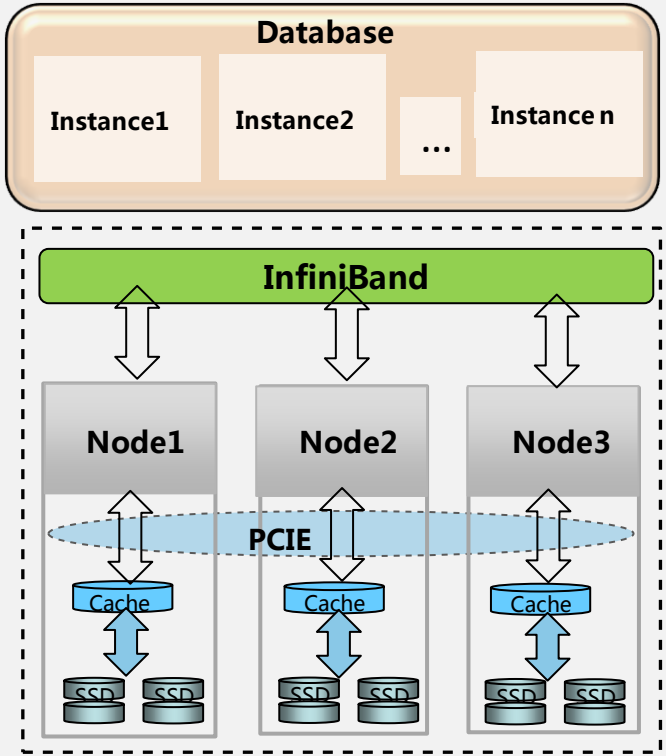
高速Infiniband网络互联，计算、存储交换无瓶颈





- 56Gbps FDR InfiniBand，超高速互联
- P2P无阻塞通信网络，数据交换无瓶颈
- ns级通信时延，计算存储信息及时传递

高性能、低时延—支持全SSD 存储



- **高IO**：整柜IOPS达240万
- **低时延**：读时延49us，写时延8us，仅为传统SAS盘的1/100~1/1000
- **高吞吐**：整柜带宽达120 GB/s

分布式SSD存储系统，主要用于数据仓库一体机场景

Content

华为FusionCube数据仓库加速解决方案

高性能、低成本的基础设施平台，混搭架构，可灵活应对不同应用负载

大数据和MPP DB

FusionInsight

GBASE

Greenplum

- 海量数据非结构化
- 高并发数据分析处理
- CEP流处理

内存数据库

SAP HANA

- 实时商业洞察
- 性能快100-100,000倍

传统数据仓库

ORACLE 12c DATABASE

Sybase IQ

IBM DB2

达梦数据库

Microsoft SQL Server

- 主数据仓库
- 关系型结构化数据

数据库整合

Oracle

SQL Server

- 减少企业数据库实例
- DBaaS 服务提供

ETL、建模、分析

SAP BusinessObjects

ORACLE BUSINESS INTELLIGENCE

DataStage

- 多维建模分析工具
- ETL、报表展现

HUAWEI

案例：财经数据仓库库外集市：FusionCube for Oracle RAC

- **客户的主数据仓库使用Oracle Exadata，还承担了数据集市的功能。高并发情况下，存储性能上存在瓶颈，CAPEX和OPEX都很高，而且扩容困难**
  - Exadata在多任务时存在资源竞争，导致报表计算能力不足。
  - 并发用户数达不到业务要求
  - Exadata扩容成本高
- **实施效果：保护已有的数据仓库投资，支持业务平滑扩容。**
  - 报表加速：报表查询时间达到业务承诺要求
  - 扩容：满足未来的2年数据存储180TB要求
  - 满足业务峰值并发用户1千

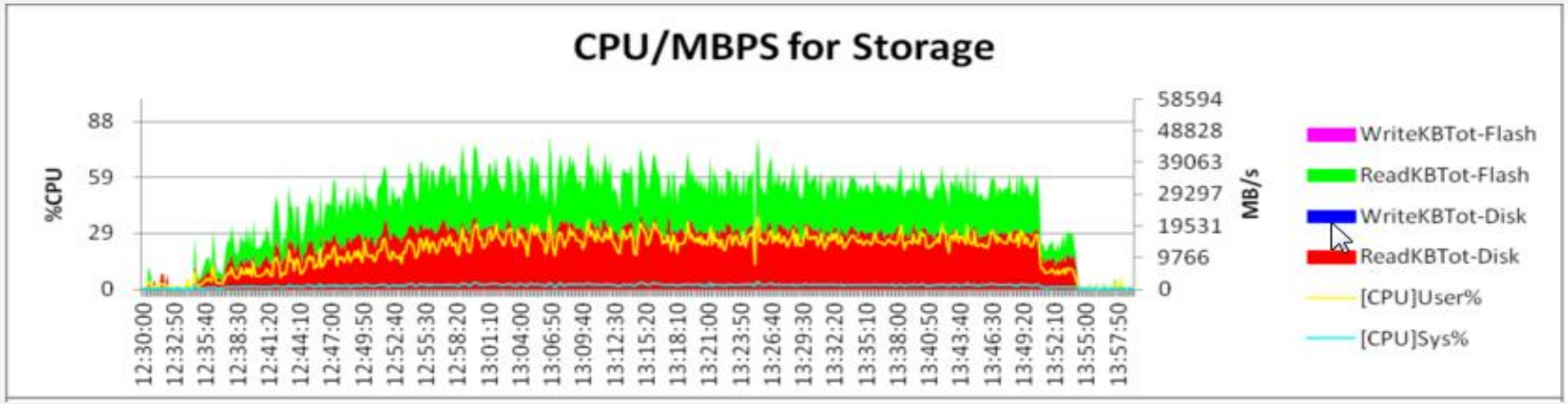
实施方案：Exadata仅作为主数据仓库，财经报表的数据集市由FusionCube + Oracle RAC承担

支持从库内集市到库外集市，到分布式数据仓库架构演进

FusionCube 2.0专为数据仓库负载进行了深度优化，性价比是Exadata的4倍

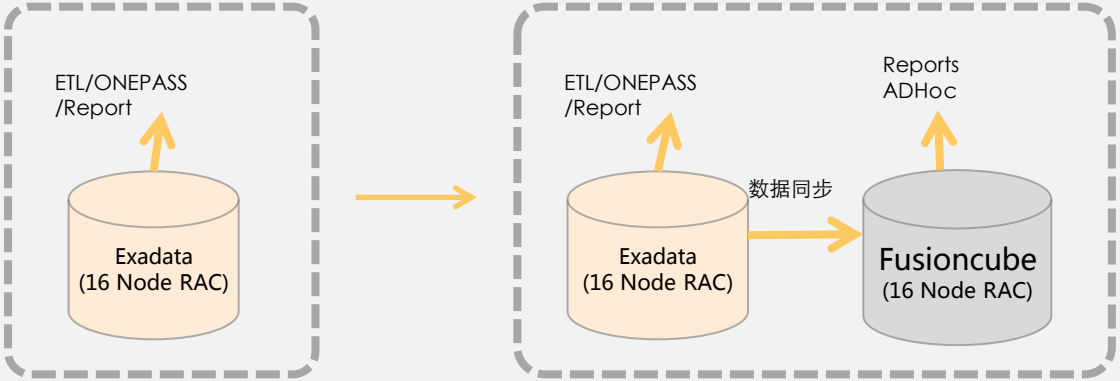
HUAWEI

原Oracle Exadata系统面临的升级扩容



- 400并发时平均IO流量已经达到34GB/s（接近一个机架的理论最大值-50GB/s）
- **不能满足业务1000并发用户数量的需求**
- 为保障相关业务顺利完成，不能开放用户查询
- Exadata扩容成本高：2013年总数据量超过90TB。每T综合成本40-50万人民币

引入FusionCube：16节点Oracle RAC



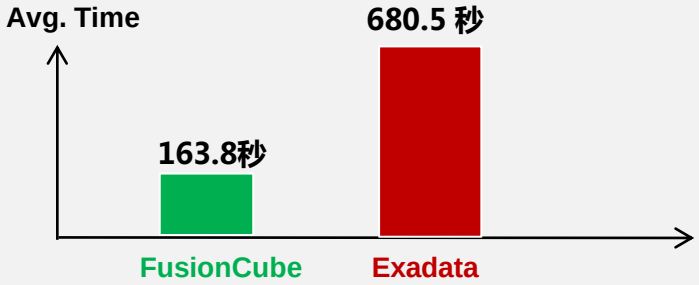
保证稳定和性能的前提下，逐步迁移，发挥华为FusionCube的性能优势

- FusionCube性能卓越，灵活扩容，满足未来5年业务发展需要
- 依然是Oracle数据库，业务平滑扩容，保护已有投资
- 高速数据同步，ETL数据同步，同步DM和明细报表数据，同步数据量达1-2T/H
- 2年数据>180TB，业务峰值并发用户**从400扩大到1500**
- 已采购250存储节点

FusionCube for Oracle RAC方案实施效果

标准报表

	EXA(s)	FU2(s)	FU2提升
订货明细	66	26	154%
应收回款考核调整明细	4601	493	833%
中国地区部明细	406	319	27%
经营概览钻取	333	60	455%
运营商服务成本明细	9180	2940	212%
平均提升	3.3倍		



✓FusionCube1柜相对Exadata**2柜**标准报表性能平均**提升300%以上**
✓越复杂的查询，提升越明显

环境	配置	压力	CPU利用率
FU2	11个计算节点 +16个存储节点	800 OBIEE + 100 MCA	平均30%，峰值50%
EXA	16个计算节点 +28个存储节点	800 OBIEE + 100 MCA	平均30%，峰值65%

案例：华为FusionCube SQL Server数据库整合

- **客户概况**
  - 某海关是国家海关总署直属数据中心，承担地区海关通关业务，其核心业务系统主要采用SQL Server数据库
- **现状与挑战**
  - 业务系统独立； 数据分散；业务、数据库和业务垂直部署（大于100个）
  - 风险数据库平台，日常查询和报表缓慢，影响正常通关
  - 客户级别的实时查询达不到要求； TB级历史数据查询缓慢
  - 数据中心的利用率和运维管理急需提高
- **解决方案**
  - 把100+ 个数据库和20+个左右的应用系统类 (5000+公司客户业务)实施整合
    - 配置FusionCube 数据库
    - 采用SSD作为主存，性能提升100+倍

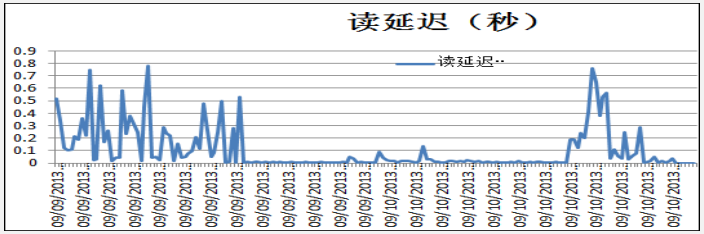
方案特点：

□全部数据库和应用整合到2个FusionCube
□采用SSD作为主存，性能提升100+倍
□统一管理接口，实现DB全程生命周期管理

HUAWEI

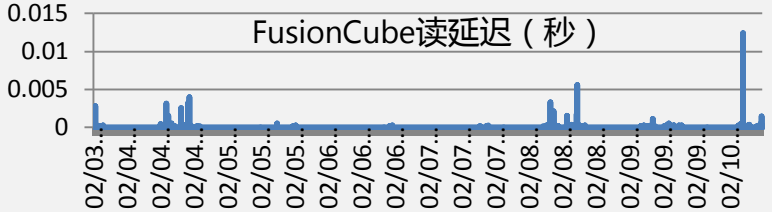
华为FusionCube SQL Server数据库整合

现业务系统测试

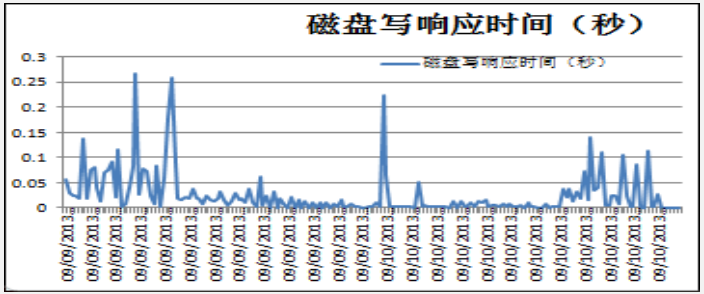


s→ms

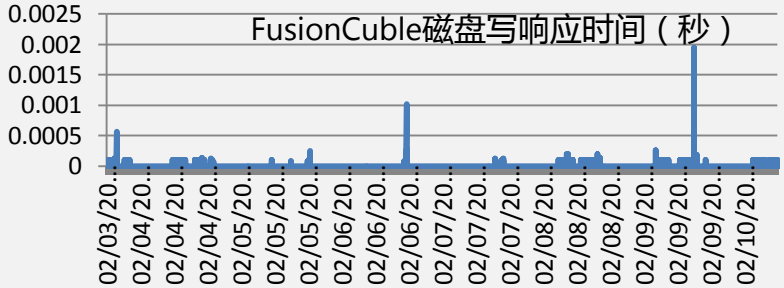
华为方案系统测试



分析结论1：读延迟最大12ms，平均值<1ms；读时延性能提升67倍



ms → us



分析结论2：写延迟最大2ms，平均值<0.2ms；写时延性能提升130倍

性能测评结果：

- 华为系统读性能提升67倍；写性能提升130倍；业务峰值IO无任何阻塞
- 华为FusionCube方案可以满足业务高突发和低时延要求

HUAWEI

华为一体机全球参考案例

金融 & 零售	制造业	运营商	公共部门	教育
Spanish Exchanges	SAP Labs China	Asia Pacific Telecom	Spanish Xanit hospital	Peking University
Hong Kong Infocast	Germany Helipark	Telefonica	Jeddah Municipality	Beijing Jiaotong University
Citic Trust	Spanish GAMO	China Mobile	Senegal E-government	Shanghai Maritime University
Italian Retail	Huawei	China Telecom	Zimbabwe Revenue Authority	TAWAZUN RABDAN

HUAWEI ENTERPRISE ICT SOLUTIONS A BETTER WAY