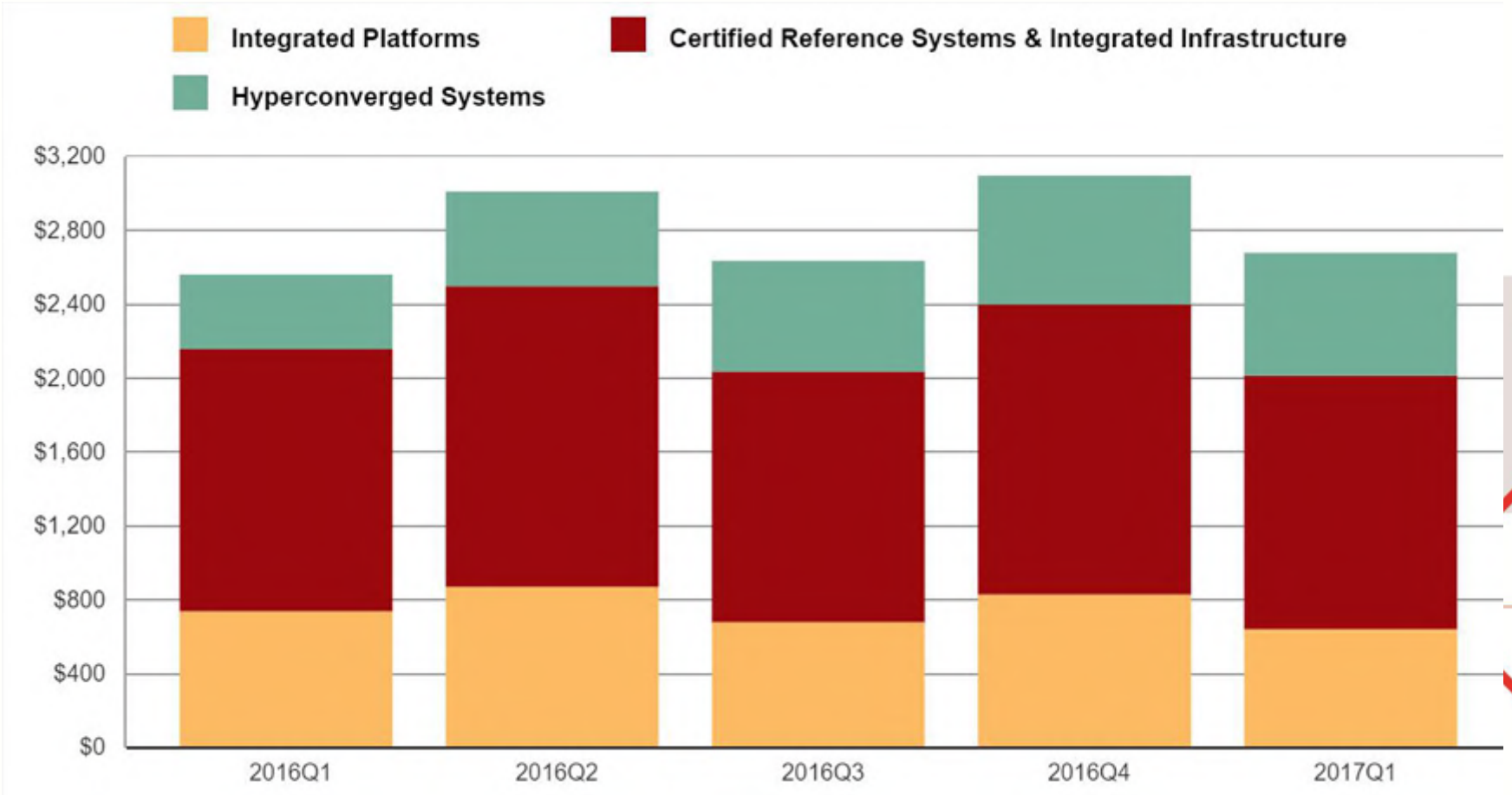
2017 可信云大会
可信云标准新一代
客户满意是未来
Cloud Computing Summit 2017.7.25-26

超融合架构及其发展

企事录技术服务公司 创始人　　张广彬

超融合架构(HCI)简析

融合系统与超融合系统



来源：IDC全球融合系统追踪，2017Q1

融合系统的形态演变

软件定义存储：从独立的系统走向（x86服务器上）共享的资源

集成系统

起步：Rack级

专用存储设备（SAN）

及其网络（FC）

虚拟化

数据库

SAN存储

（超）集成系统

起步：Rack级

基于x86的存储角色

Server (as) SAN

虚拟化

数据库

存储型服务器（存储节点）

超融合系统

起步：Chassis/Box级

无专用存储角色

Server (also as) SAN

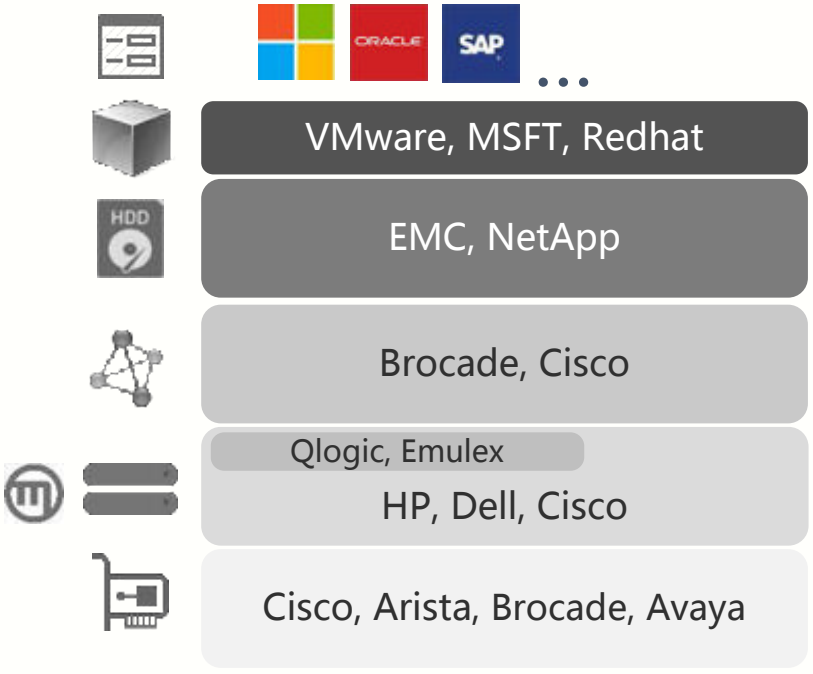
虚拟化

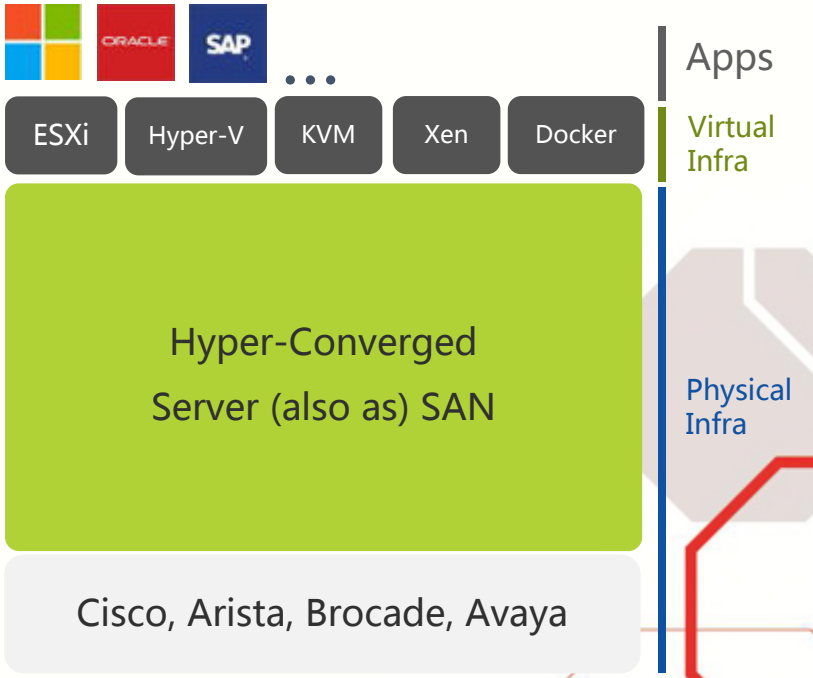
多节点服务器

（计算&存储）

?

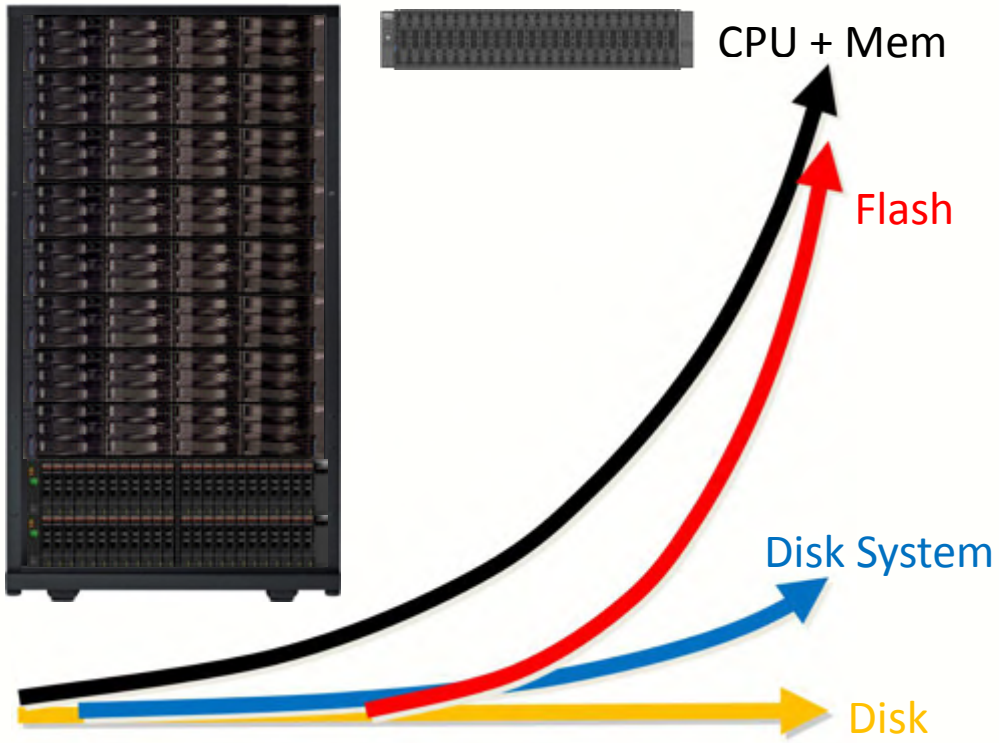
计算与存储的(超)融合，网络...？





Server SAN 即“用Server做SAN”（以Server取代SAN），属于软件定义存储（SDS）的一类实现
超融合架构在 Server SAN 之上加入（用户应用的）计算任务，实现了计算与存储的一体化

变革根源：闪存替代硬盘



为什么会有SAN？

- 计算靠摩尔定律
- 存储靠增加盘的数量
- 存储网络带宽是潜在瓶颈

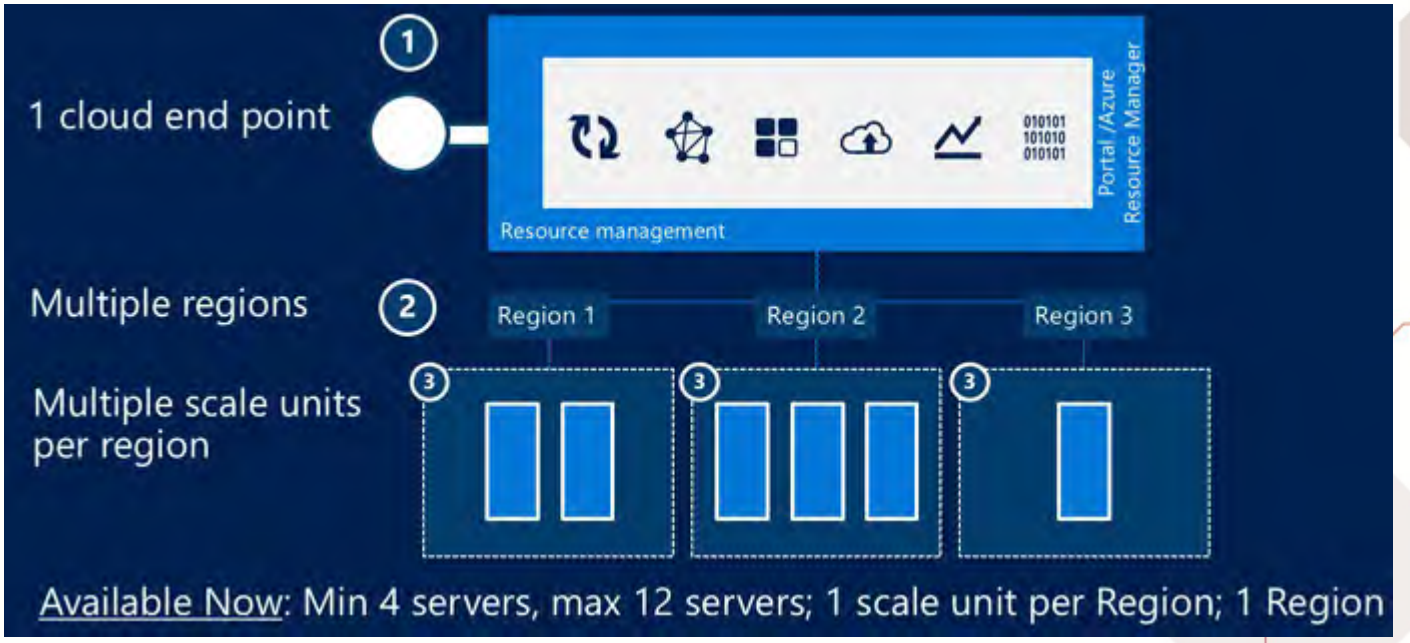
Server SAN & HCI

- 闪存也遵循摩尔定律
- Server I/O 潜力可比 SAN
- 对计算和网络要求较高

HCI: HyperConverged Infrastructure
（超融合架构）

构建私有云的基石

- 超融合架构（HCI）是传统企业（私有）应用与互联网（云）技术的结合
- 基于超融合系统部署私有云，比基于 SAN 的三层架构更具优势
  - ✓ 国际：微软 Azure Stack 选择超融合作为其混合云部署的私有云基础
  - ✓ 国内：青云QingCloud 选择超融合部署私有云



超融合系统的商业模式

类似其他融合系统，超融合系统主要是一种商业模式上的创新：

Server SAN + hypervisor + 管理，解决存储是关键

- 合作模式：（超融合）一体机，或（认证）参考架构
- 一体机：X OEM Y，品牌（通常）属于强势方

软件提供商	服务器厂商	模式	实例
独立	弱势品牌	软件品牌一体机	Nutanix
	强势品牌	硬件品牌一体机	联想、Dell、H3C…
独立	各种品牌	认证参考架构	Maxta MaxDeploy
半独立	母公司	母公司品牌一体机	Dell EMC VxRail
	各种品牌	认证参考架构	vSAN ReadyNode
非独立	母公司	母公司品牌一体机	华为FusionCube

延伸：HPE SimpliVity？OmniCube & OmniStack

小结：HCI & Server SAN

① 存储是各种融合系统（集成系统到超融合系统）的核心

② 超融合用 Server SAN 替换了 SAN

③ 超融合架构 ≈ 计算虚拟化 + 软件定义存储（Server SAN）

④ Server SAN是“服务器 即 存储” （Server as Storage）

⑤ 超融合是“服务器 亦 存储” （Server also Storage）

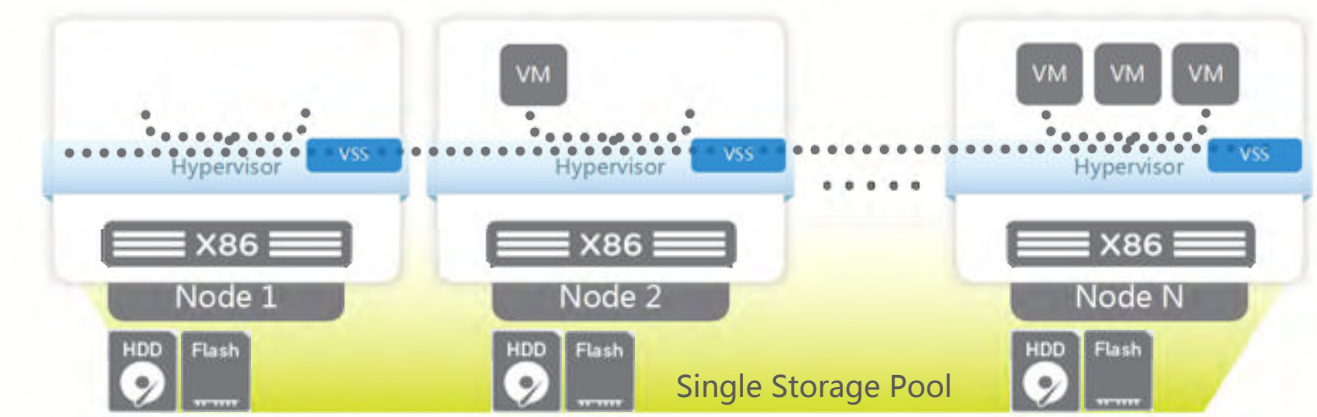
⑥ 集成系统是硬盘时代走向尾声的产物

⑦ 超融合系统伴随着闪存存储的崛起

超融合系统的“边界”

HCI & Server SAN：模糊的边界

只需1个用户VM，Server SAN 就可“升级”为超融合



用户应用

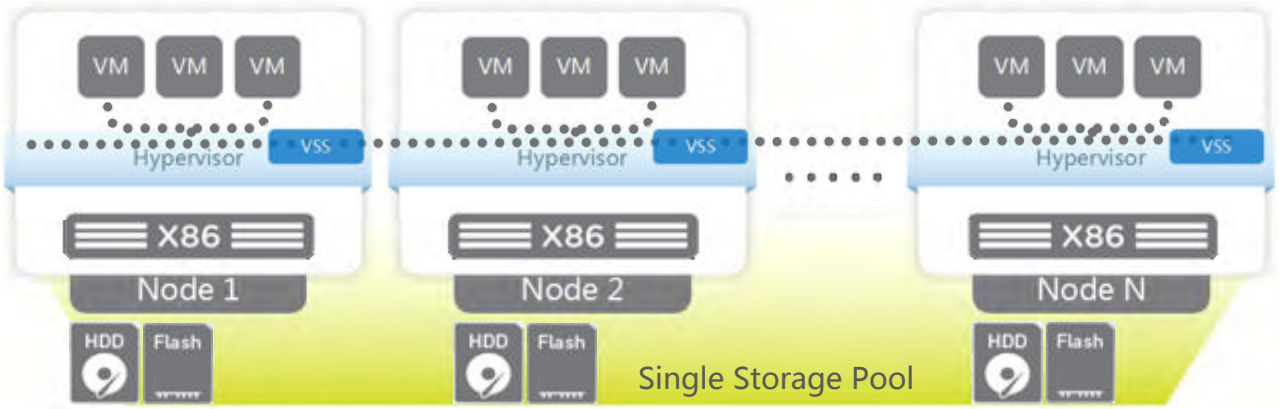
存储资源

VSS (Virtual Storage Service) 是分布在各个节点上的存储任务，可运行在hypervisor/OS级或虚拟机（CVM）中。作用有二：

- 接受本节点运行应用（如用户VM）的存储请求；
- 将本节点拥有的物理资源贡献到统一的存储资源池。

节点级：超融合的“常态”

每个节点都同时运行存储任务和用户应用



用户应用

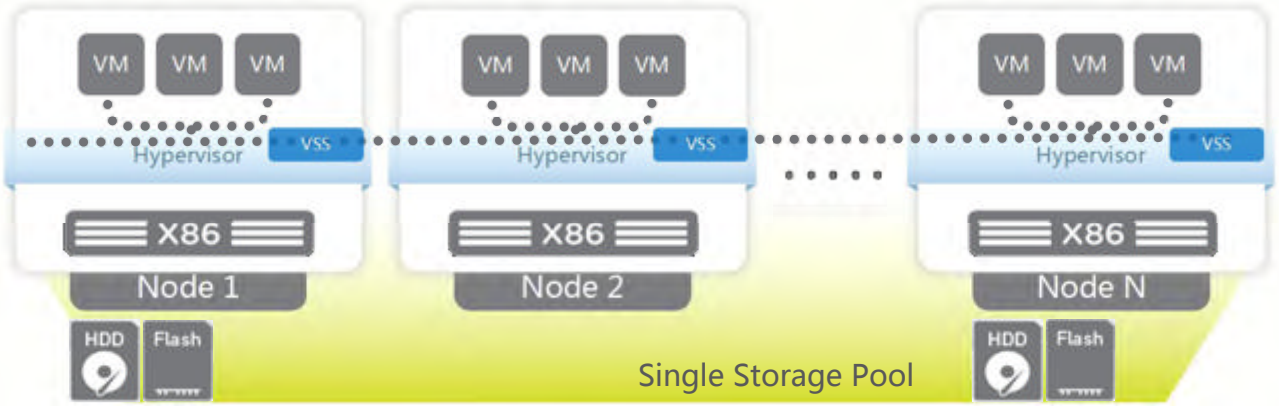
存储资源

存储资源与计算资源：

- 提供存储资源的节点至少要有1-2个SSD（纯硬盘性能太差）；
- SSD性能越高、数量越多，对计算资源要求越高。

“部分”超融合(1)：仅计算的节点

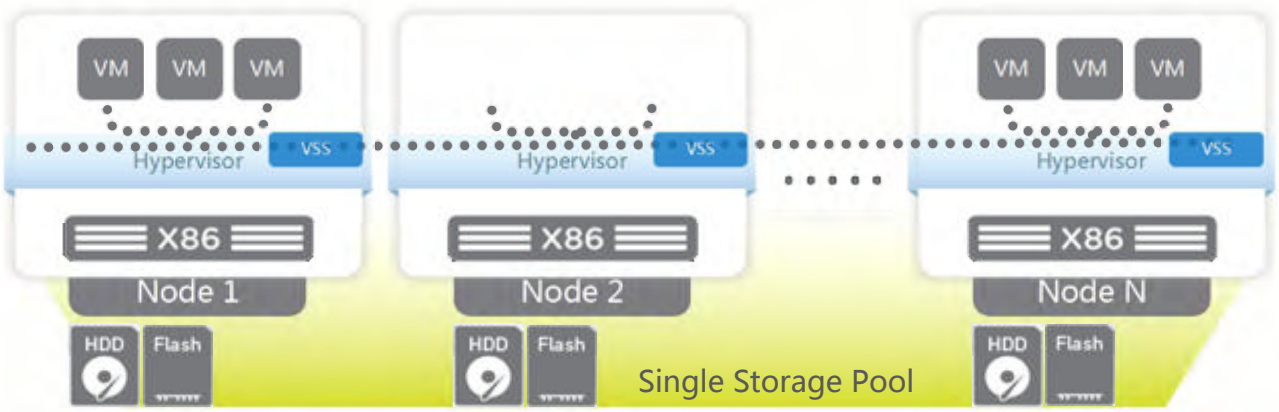
部分节点不贡献存储资源



部分节点因缺乏存储硬件（SSD、硬盘、HBA）支持等原因而只作为计算节点，运行用户应用。实际上，存储应用（VSS）也在运行，为该节点上的用户应用提供存储资源，但不能对其他节点贡献存储资源。

“部分”超融合(2)：仅存储的节点

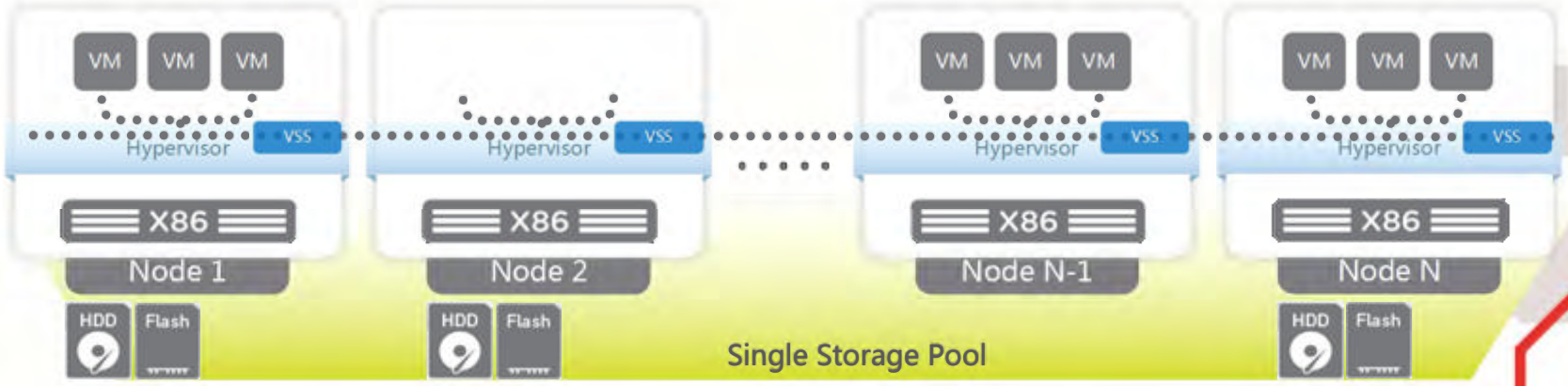
部分节点不运行用户应用



部分节点因计算能力（CPU、内存）不足而只作为存储节点，运行存储任务（相当于Server SAN），向集群中的其他节点贡献存储资源，而不承载用户应用。

系统级：超融合的“变态”

计算+存储、计算、存储等不同类型节点并存



不同类型的节点各有侧重，但是超融合“系统”的边界在哪里？

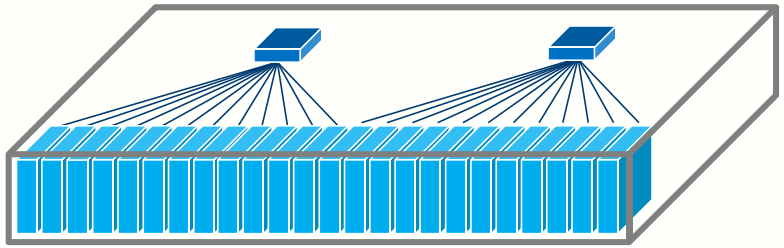
如果不考虑纯存储节点（Server SAN，无hypervisor）的情况，那么，加入物理机（裸机）作为纯计算节点呢？

NVMe、网络及分离部署

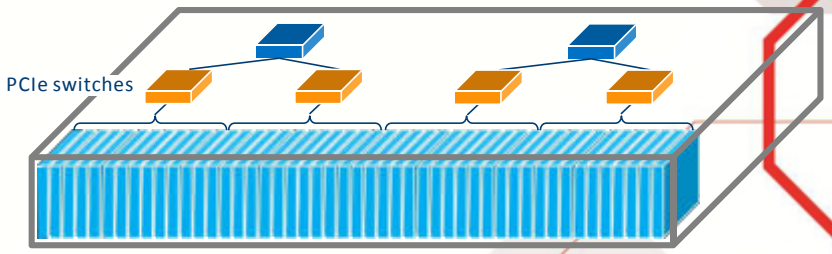
新硬件平台的福音

新一代服务器 CPU 拥有更多 PCIe 信道，可支持更多 NVMe SSD

- ✓Intel Xeon Scalable：40→48，双路=96
- ✓AMD EPYC：单双路均为128



每台2 RU 服务器 24 个 **U.2 15mm**
无 PCIe Switch（CPU直连）



每台 2 RU 服务器 48 个 **U.2 7mm**
通过 PCIe Switch

（全）闪存的挑战

- 硬件配置：大量高性能NVMe SSD对CPU和网络提出了更高的要求
- 软件堆栈：需要针对性的优化，以充分发挥 NVMe、3D Xpoint 等新型存储介质的性能

节点类型	特征	节点类型	CPU	核数	内存	（数据）存储	网络
高端虚拟化	计算+存储	超融合	2 × E5-2658A v3	2 × 12	256 GB	10 × 2TB SATA 800GB SSD	2 × 10GbE
Oracle存储	全闪存	Server SAN	2 × E5-2660 v3	2 × 10	160 GB	6 × 3.2TB SSD	2 × 10GbE 4 × 56G IB
虚拟化	计算+存储	超融合	2 × E5-2630 v3	2 × 8	128 GB	6 × 2TB SATA 600GB SSD	2 × 10GE
HANA存储	硬盘+缓存	Server SAN	2 × E5-2620 v3	2 × 6	64 GB	12×900GB SAS 1.2TB SSD	2 × 10GbE 2 × 56G IB

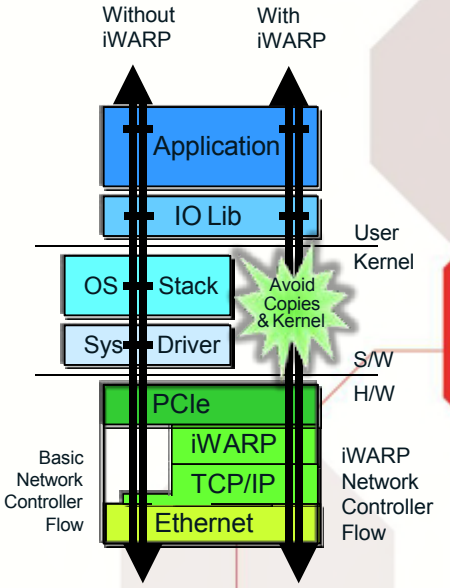
示例：高性能闪存对计算和网络资源的消耗，未必逊于用户应用

网络：以升级促融合

存储决定超融合的下限，网络决定超融合的上限

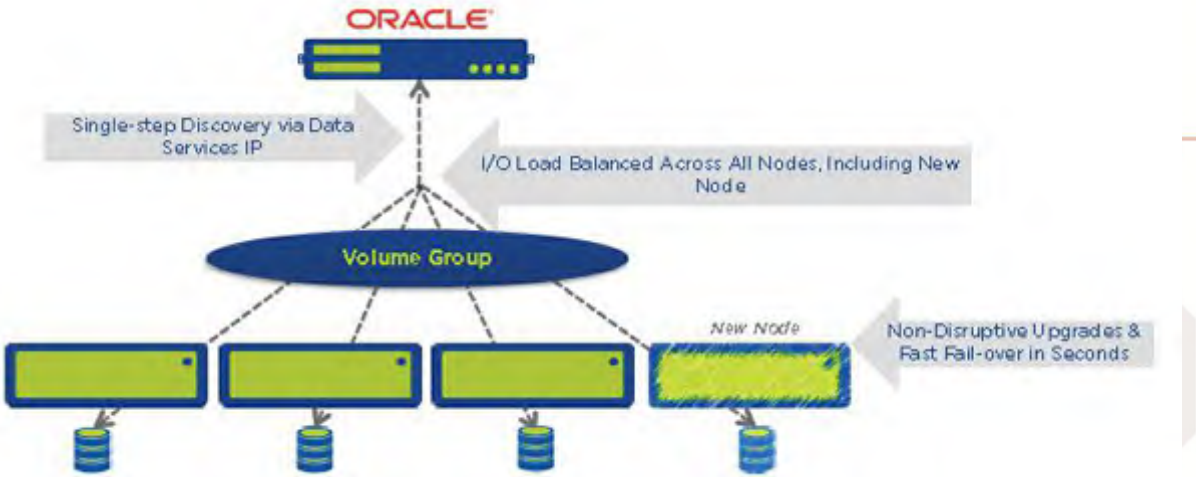
1. 性能需求：带宽和延迟对分离部署尤为重要
   - 提高带宽：40GbE 或 25GbE？
   - 降低延迟：RDMA/iWARP ……
2. 功能需求：规模、灵活、易用
   - SDN & NFV
   - 自动化

25GbE < PCIe 3.0 x4 < 4x 10GbE (40GbE)



Server SAN 是大势，超融合看场景

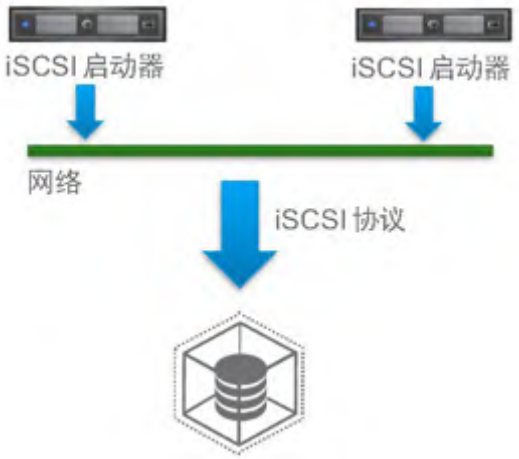
1. 基于 SAN 存储的集成系统支持两种应用场景：
   - 虚拟化
   - 数据库
2. 现有基于虚拟化的超融合架构：
   - 天生适用虚拟化场景
   - 不能很好满足要求裸机安装的数据库等场景
3. 可将超融合的存储（Server SAN）与外部计算资源（如物理机）结合，形成一种计算与存储分离部署的局面

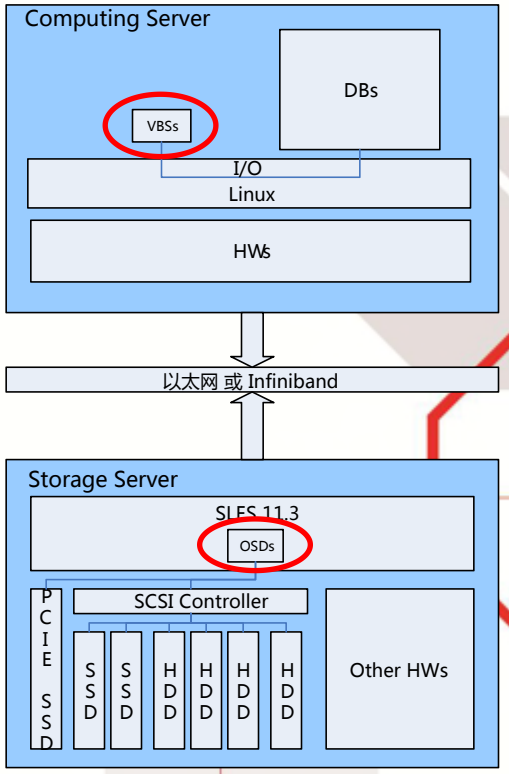


分离部署：HCI as Server SAN

类似 Server SAN，超融合系统可以用两种方式对外输出存储：

1. 专有接口（计算端需安装代理）
2. 标准接口，如 iSCSI
   - SAN 提供块存储接口，Server SAN 亦当如此；
   - 几乎必须是 iSCSI；
   - 操作系统广泛支持，简化计算节点部署





总 结

1. 超融合架构的核心是 Server SAN
   - HCI 与 Server SAN 的边界很容易模糊
   - x86服务器是 Server SAN/超融合的基础
   - SSD 是 Server SAN/超融合的助推器和试金石
   - 网络正在走向“超”融合阶段
2. 基于 Server SAN 的新一代融合架构
   - 超融合系统的存储属性（如支持 iSCSI）
   - 计算与存储根据需要（超）融合或分离部署
   - 资源搭配更灵活，适合更多应用场景
   - 节点级超融合与系统级超融合

Thanks

北京企事录技术服务公司