WHITE PAPER

借助InfiniBand网络搭建可扩展存储系统

当前面临的挑战

对于当下数据中心的工作人员来说，数据中心的资金和营运费用中，构建和维护存储系统所占比例越来越大，这一点已不足为奇。而增加数据中心存储能力和性能的需求，其因素来自各个方面。计算能力的提高，新软件模式的开发，使我们可对浩瀚的数据进行有用的分析。由于存储每千兆字节的存储硬件成本有所降低，所以企业机构可以存储更多更精细的数据，并且能够将数据保存更长的时间。

上图所示为结构化数据的范例（沃尔玛500TB数据库）和非结构化数据范例（交通和安全摄像头每天8Tb的数据）。随着我们对非结构化数据的处理（如进行分布式计算），对动态产生的数据，如网络日志、环境传感器数据以及邮件等进行保存所带来的用处就越来越大，因此存储的需求也就随着暴涨。

对生产商来说，要跟上现代系统所需要的计算性能，试图仅仅建立一个具备所需容量和性能的节点已经不再可行。因此，存储的发展方向是逐渐横向扩展而不是按比例增加。为保持与数据库、分析或应用的发展同步性，这一横向扩展方法对高速存储网络。

除此之外，不管是非结构化数据应用程序还是传统结构数据库，也都会移动至分布式处理系统（最明显的是分布式计算Hadoop和Oracle RAC）。为了支持可扩展性更高的应用程序，存储本身必须能够扩展。

存储系统不仅须具备移动大量数据的能力，还须具备快速发现并且存取单个数据的能力。这一属性通常称为存取时间。在分布式数据库和分析模型中，延迟对查询整体性能的重要程度与带宽相同。

数据库和应用程序的分布式虽然还并非趋势，但必定是未来发展的方向。解决较大数据量的能力，仍将来自配置的横向扩展。是否具备支持横向扩展设计的网络互连能力，将会是区别“可扩展大数据”领域的关键。

传统解决方案及其固有弊端

目前市场上存在多种存储技术和协议，其中包括SATA、SAS、SSD、iSCSI、InfiniBand、FCoE、以及传统的Fiber Channel。现代数据中心很有可能采用其中多种技术和协议，以便在存储性能、成本和可用性方面取得平衡。

业界公认Fiber Channel的发展已经趋于停滞。Fiber Channel的发展已经跟不上当今应用领域所需的性能需求。性能只有8Gb/s，Fiber Channel的性能已经比不上以太网。在网络之外单独针对存储系统提供一个专用的SAN网络会增加固定资本和营运成本，也无法进行扩展。当前产业的发展方向是多网融合的网络架构解决方案，即网络和存储系统采用一个相同的网络互连。

尽管发展趋势如此，但Fiber Chanel仍有其优势。Fiber Channel的设计是一种无损光纤通道网络架构的专用通道。与不太可靠的以太网相比，通过Fiber Channel传递的数据由于无损且不会重传，所以其存取时间较为一致。其次，由于 Fiber Channel为专用基础架构，所以不会因其他网络流量类别而导致网络拥塞。最后，它不受因共同分享同一个物理网络的路由协议产生的网络事件的影响。这些都是以太网难以复制的重要特性，而对存储系统进行扩展时这些特性尤为重要。

对存储系统进行扩展既要求互连具备更高的网络互连性能，也要求互连具备一定的复杂度，能够对复制、多重路径和高可用性提供支持。当然，需要使用好的自动配置软件来对这些解决方案的复杂度加以控制，以降低错误操作的风险，防止出现数据丢失或者系统故障。

在设计即能满足这些要求，又能够保持使用简便性的存储系统时，大部分销售商采取的是整体封装设备方案。整体封装存储设备背板采用为快速和可靠的网络互连（如SAS、Fiber Channel或InfiniBand等），但前端提供的是网络连接（最为常见的是以太网或InfiniBand）。采用这种方法最为知名的实例有：Oracle Exadata、EMC Symmetrix、Isilon、Panasas以及Data Direct Network。

公司名称	背板	网络
Oracle Exadata	InfiniBand	InfiniBand
EMC Symmetrix	Fiber Channel	以太网
Isilon	InfiniBand	以太网
Panasas	SAS	以太网, InfiniBand
Data Direct Networks	InfiniBand, SAS	InfiniBand

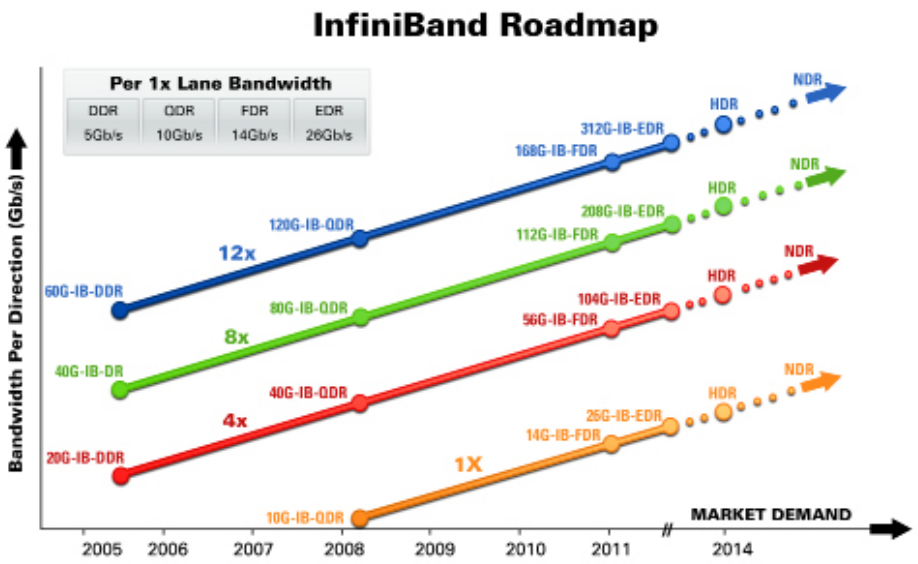
此类设备中，每个设备都采用一种网络互连在背板中提供存储通道，另一种网络提供应用程序节点间网络连接。注意，InfiniBand是唯一一种既用作存储通道也用作网络互连的互连方式。

关键优势之InfiniBand

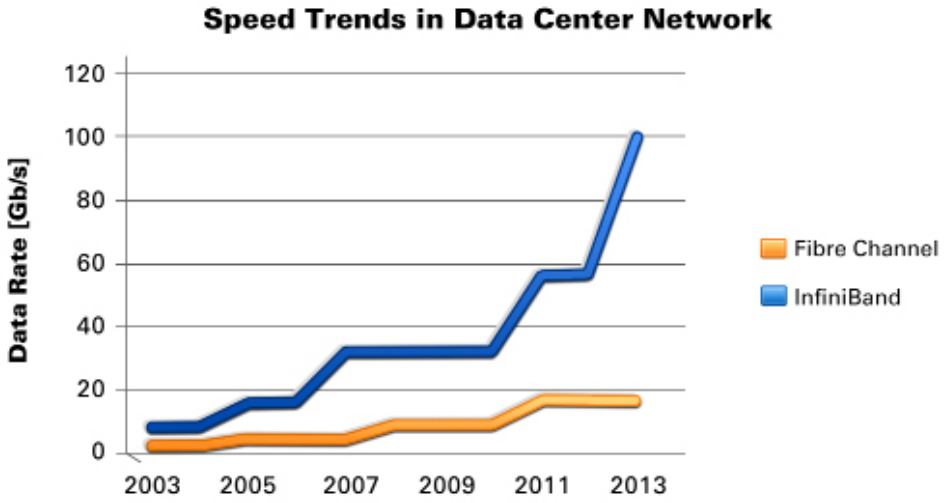
那么，什么是InfiniBand，为什么越来越多的存储系统销售商不管是背板还是网络连接都要用到此连接？InfiniBand是一种在2000年左右出现的，基于标准的网络协议。InfiniBand整合了NGIO和Future I/O(PCI总线替换技术的竞争技术)这两种技术。从设计上来说，InfiniBand具有总线技术的特点，但实际上，PCI Express——最终产生的PCI替换技术，从概述上来说是InfiniBand的一个子集。

InfiniBand与其他网络的核心区别有两个方面。首先，其采用的是一种基于信用的流量控制系统。即在接收对象未保证充足的缓冲之前，不会发送数据。这样，就使得InfiniBand成为像无损光纤通道网络架构那样的光纤通道。其次，InfiniBand支持远程直接内存访问（RDMA），具备在完全卸载CPU和操作系统的方式下，在两个远程系统的存储区域移动数据的能力。作为原始总线设计遗留下来的的理念，如要对分布式系统进行扩展，RDMA是关键。有RDMA的InfiniBand具备多种关键优势。

InfiniBand的物理信号技术一直超前于其他网络技术，使得它都具备比其他任何网络协议都大的带宽。目前以56Gb/s运行的InfiniBand，其发展路线预计达到EDR(100Gb/s)的时间是一年半左右。



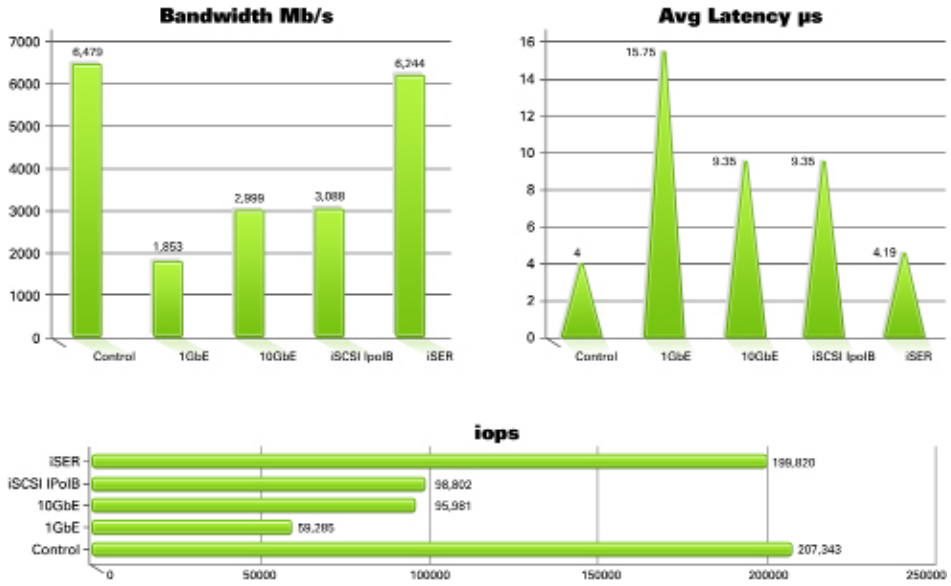
InfiniBand这一名称本身即说明了其无限的带宽发展前景。InfiniBand路线图设计的目的就是要保证单个链路的带宽能够保持在大于PCIExpress(PCIe)总线数据速率的水平。这样，系统就能够以其可产生的最快速度，在网络间移动数据，并且不会因出现因网络限制而导致的备份。这样，就可让 InfiniBand具备无限带宽。



虽然带宽高可能是InfiniBand最为显著的特点，但实际上，RDMA所具备的优点能够给大多数存储应用带来更大的性能提升。InfiniBand借助RDMA的使用，能够绕过操作系统和CPU，从而能够让数据移动传输效率更高。系统内的所有资源，包括CPU和IO设备的访问均由操作系统负责管理。像TCP、UDPO、NFS和iSCSI这样的协议，其数据路径通常需要与其他应用程序和系统进程一起排队，等待CPU处理。这样不仅会降低其所用网络的速度，而且也会占用本可以用来加速任务处理速度的系统资源。

RDMA无需如此，能够让InfiniBand的数据路径跳过这一排队过程。数据收到后，不受CPU负载决定的不确定延迟影响，可立即放到存储器之中。这一特点具备3个好处。首先是无需等待，这样交易的延迟就非常低。RDMA 1/2 RTT延迟小于1微秒。在下图中，您可看到iSER通信协议下（能够进行RDMA的iSCSI），存储应用程序运行时的存取时间。注意，这些存取时间与本地存取的存取时间非常接近。其次，由于不需争用资源，所以延迟较为一致。最后，由于RDMA的使用跳过了操作系统（OS），所以会大幅节约CPU资源。若系统效率更高，则这些节约的CPU周期可用来对应用程序的性能进行加速。

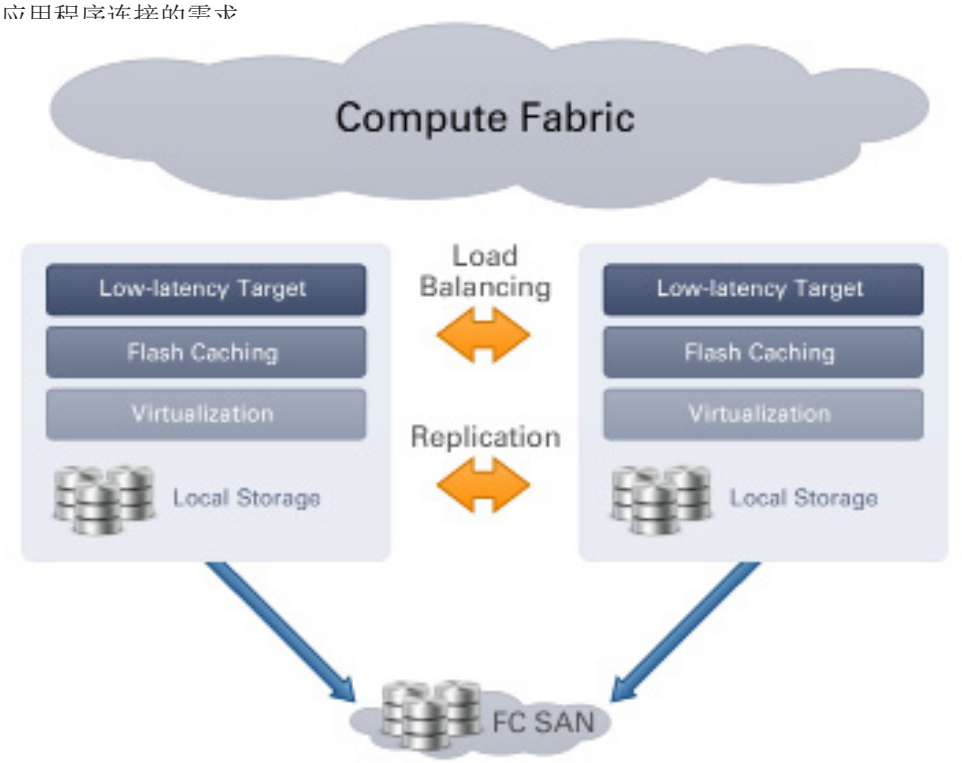
通过以下试验，可了解到RDMA对存储性能的影响。在这一实验室试验中，使用InfiniBand交换机将发起与目标服务器相连（对于试验试验中的以太网部分，则不通过交换机，采用直连）。首先在存储目标系统上进行本地测控试验。此测控试验旨在确定目标系统预期最佳性能的基准。然后再通过多个传输方式进行同样的试验，其中包括千兆以太网的iSCSI传输，10GigE的iSCSI传输，IPoIB的iSCSI传输（非RDMA的InfiniBand传输）以及可进行基于RDMA的iSCSI传输iSER。



这一简单试验的结果非常具有说服力。万兆以太网相比千兆以太网的iSCSI传输性能有大幅提升。然而从万兆以太网至4万兆InfiniBand，我们未看到较大变化。其原因是目标系统无法使整个InfiniBand传输能力饱和。网络并非瓶颈，瓶颈是CPU。处理TCP与iSCSI操作需要占用多个CPU周期，然而，在InfiniBand链路使用RDMA(iSER)协议时，我们能够利用带宽增加这一优势。使用iSER协议时，可实现的性能水平为存储系统能力的96%。

VSA通过核心技术构建解决方案

基于InfiniBand的性能，对于其逐渐成为存储设备的互联选择也就不足为奇了。对可扩展应用的需求，如其分布数据库、分布式计算、云和HPC等，产生了将InfiniBand直接与应用程序连接的需求



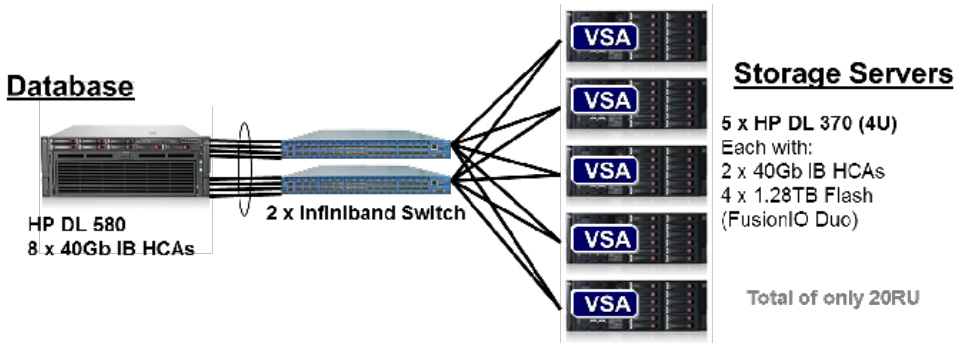
Mellanox Technologies是全世界领先的InfiniBand厂商，坚信InfiniBand技术是适合此类应用领域的存储互连技术。Mellanox向其合作伙伴和系统集成商提供一种存储加软件（名称为VSA），此种软件是围绕iSER技术构建的一种软件平台。VSA简化了存储客户和设备生产商采用InfiniBand连接存储的进程。

VSA的设计适用于多种存储架构。VSA是一种轻量级的块存储（block storage）软件层，提供高性能iSER和iSCSI目标的存储。此外，VSA能够对诸如多模式自动配置、监测、复制以及高可靠性等特点进行集中管理。除其他特点外，VSA还具备以下特点：

- 对某集群式储存器进行集中管理
- 高性能iSER/iSCSI目标
- 支持闪存或SSD作为缓存层使用
- 通过RDMA复制
- 平衡负载
- 高可靠性 （High Availability）
- Fiber Channel桥接

数据库

使用基于VSA的设备，通过“扇入”（Fan-in）方法，提高集群式数据库的性能。诸如HPDL580等高端SMP机器，其处理交易的能力要大于本地存储器或者专用SAN处理交易的能力。其中一种解决方案是对5个存储服务器的硬盘提供虚拟RAID。在本示例中，InfiniBand能够让DL580的IO总线达到饱和。



在此配置中，随机IO操作为每秒2.5M次，其性能为23Gb/s。如使用Fiber Channel，则同样的存储配置需要50条FC线缆！

数据仓库Petabyte级存储

web2.0应用程序、云存储以及大型数据应用程序对数据量的要求，以及统计和数据可追溯性标准都是驱动数据仓储应用程序提高能力需求的因素。数据仓储应用中考虑的主要度量单位为每千兆字节需要的花费美元数。目前已有VSA设备，不仅能够以最低价格提供最大容量，而且其性能也是无以伦比。VSA的使用使得可扩展存储设计能够提供线性性能扩展，也能提供中央管理。

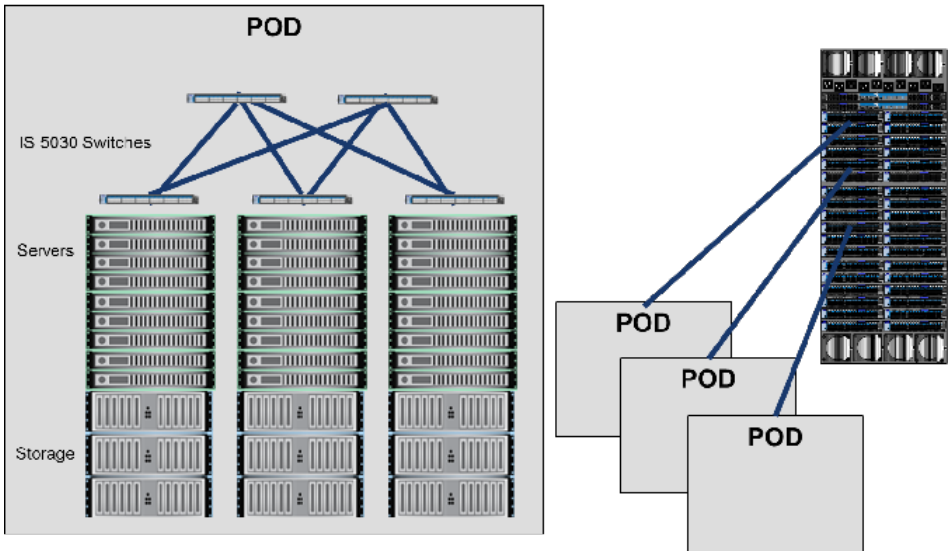
云计算解决方案

基于可扩展性和弹性设计原则的云才是好的设计。将存储与云架构中的管理程序（Hypervisor）分开，实现自动配置的速度提升，从而提高灵活性，更好地对资源进行利用，并且能够快速进行动态迁移。传统的SAN/网络云架构需要为存储、网络、管理和动态迁移提供单独的专用适配器。而InfiniBand强大吞吐量，可以使得存储、动态迁移、管理以及标准网络数据流共用同一个网络。

这样，管理节点（hypervisor node）在实现专用存储SAN所具备优点的同时，可用一个网络进行连接。借助PXE的支持从InfiniBand进行启动后，VSA可将存储与无盘瘦管理节点分开。借助

RDMA实现CPU负载卸载，每个管理程序（hypervisor）可创建多个虚拟机。

InfiniBand具备在单个2层子网上扩展到成千数万节点的能力，从而简化大型云应用的配置。



小结

采用Mellanox解决方案，在不增加数据成本的同时，建立高带宽、低延迟以及高IOPs的经济型存储网络架构，可能满足企业不断提升的存储需求。Mellanox为客户提供基于InfiniBand的存储连接解决方案，能够在节约数据中心功耗和空间的同时，将客户基础网络架构整合到选择的单一网络架构内。

Mellanox VSA，基于InfiniBand和以太网的存储解决方案，以低廉的价格为您提供卓越的性能。将您的投入实实在在地转化为全球范围内的客户优势，如最大程序提高服务器利用率，提高应用程序性能，减少备份时间，强化系统整合，降低功耗以及降低购置成本（TCO）。

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com