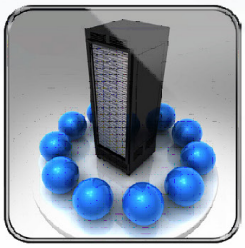
解决方案简介

通过 vSphere vMotion 和 Mellanox 端到端 40GbE 互连解决方案加速虚拟机迁移



简介

大型虚拟化环境（包括云基础架构中的虚拟化环境）要求虚拟机间通信和 Hypervisor 服务（如虚拟机实时迁移）具有高 I/O 性能。借助性能更高且更有效的网络，数据中心经理可以更快地将虚拟机迁移到不同的物理服务器，从而以更低的总体拥有成本 (TCO) 满足严格的服务水平协议 (SLA)。更快的实时虚拟机迁移可以最大限度地减少服务器的"非活动"时间，从而支持每秒运行更多作业，这可以最大限度地提高虚拟化基础架构的效力。

Mellanox 端到端 40GbE 解决方案包括 SX1012 40GbE 交换机、ConnectX-4 Lx 网卡和 LinkX 线缆，能够以创纪录的时间迁移单个大型虚拟机（占用大量内存）或同时迁移大量虚拟机。

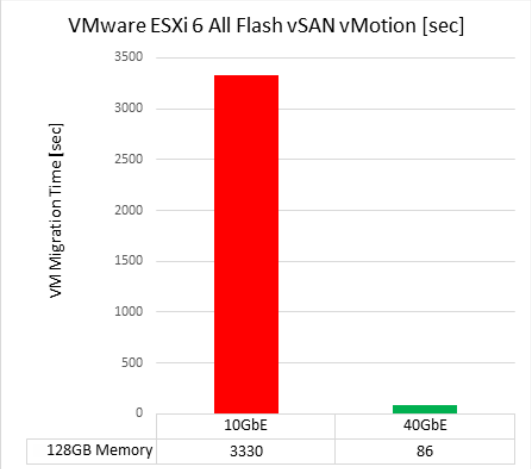


***图 1.** 在运行含 Microsoft FileIO 工作负载的虚拟机的 Linux 来宾操作系统上实现 38 倍 vMotion 加速*

结果显示，在 40GbE 上运行 vSphere 6.0 vMotion 与 10GbE 相比，在 Linux 来宾操作系统上运行虚拟机时可让迁移速度提高 38 倍，在 Windows 来宾操作系统上运行虚拟机时可提高 5 倍。这两种情况都显著提高了虚拟化基础架构的效率。

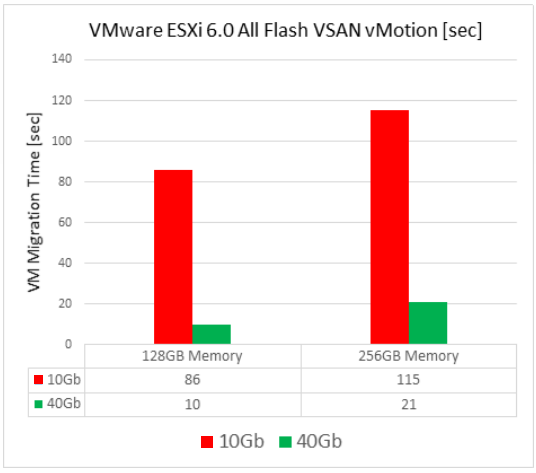


***图 2.** 在运行含 IOmeter 工作负载的虚拟机的 Windows 来宾操作系统上实现 5.5 倍 vMotion 加速*

测试设置

目标是在两台 ESX 6.0 服务器之间实时迁移一个虚拟机，比较通过 40GbE 和 10GbE 进行迁移所需的时间。测试设置包括具有以下配置的四台服务器：VMware ESX 6.0 Hypervisor、Mellanox 收件箱 10/40GbE 驱动程序（ESX 6.0 服务器）和 ConnectX-4 Lx 10/25/40/50GbE 网卡。ESX 6.0 服务器之间通过两台基于 SwitchX-2 的 SX1012 40GigE 交换机和 LinkX 铜缆连接。第五台服务器在 Microsoft Windows Server 2012 R2 64 位操作系统上运行 VMware vCenter。使用 VSAN 6.0 全闪存作为存储，在四台 ESX 6.0 服务器中各包含 1 个 PCIe 800GB SSD 和 6 个 800GB SSD。

ESX 6.0 服务器具有两种内存配置：

1. 384GB 总内存，1 个虚拟机配置为 128GB
2. 384GB 总内存，1 个虚拟机配置为 256GB

对于每种内存配置，执行 vMotion 测试，并将整个内存分配给虚拟机。在现实生活中，举例来说，这种情况模拟了活动虚拟机将实时接收的数据记录连续写入内存的情况。

每个测试都执行了几次迭代，以验证结果的稳定性。

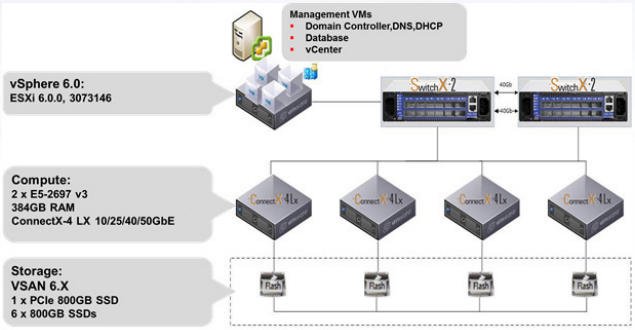


图 3. *设置框图*

基准测试和结果

由于使用 VSAN 作为存储，只需要迁移虚拟机内存。为了执行该作业，vMotion 需要执行以下任务[1]：

1. 在目标主机上创建一个影子虚拟机。
2. 通过 vMotion 网络将每个内存页从源服务器复制到目标服务器。此阶段称为 preCopy。
3. 对虚拟机内存执行另一轮复制，复制在上一次 preCopy 迭代期间发生改变（或变"脏"）的任何页。
4. 继续此迭代内存复制过程，直到没有发生改变的页（待复制的页）。
5. 在源服务器上眩晕 (stun) 虚拟机，并在目标服务器上恢复该虚拟机。

当每次完成 preCopy 迭代的时间比上一次 preCopy 迭代短时，此过程进行得很顺利。在这种情况下，虚拟机实时迁移将会收敛。但是，如果活动虚拟机修改（或"污染"）包含的内存量比迁移的内存量大，则迁移过程将不会收敛。为了避免这种情况，vMotion 使用发送页时眩晕 (SDPS) 操作，这会降低虚拟机的运行速度。激活 SDPS 将确保内存修改速率比 preCopy 传输速率慢，并保证实时迁移过程的收敛。

我们的基准测试表明，在比较通过不同速度的网络实时迁移持续污染内存的活动虚拟机时，迁移时间的差异要大于网络速度之间的比率。图 1 和图 2 很好地展示了这一点，在 Linux 来宾操作系统上运行时，40GbE 比 10GbE 快 38 倍，在 Windows 来宾操作系统上运行时快 5.5 倍（二者都采用活动 SDPS）。

摘要

使用 Mellanox 端到端 40GbE 解决方案获得的性能结果表明迁移速度获得大幅提升，这是因为与 10GbE 相比提供了更高的 I/O 带宽。预计在使用 ConnectX-4 Lx 50GbE 的 50GbE 上和使用 ConnectX-4 100GbE 的 100GbE 上将实现更高的效率（通过 Mellanox Spectrum 10/25/40/50/100GbE 交换机和 LinkX 线缆连接）。

可以通过使用 Mellanox 端到端的高性能互连解决方案来克服云和 Web 2.0 超大规模计算中不断增加的 I/O 需求，该解决方案可以提高云服务提供商对具有新虚拟机和应用程序需求的新用户进行预配置的速度，并且可以在不影响现有用户服务的情况下满足其 SLA。模式从跨物理服务器批量实时迁移虚拟机以进行灾难恢复转变为用于维护或负载平衡的计划内迁移，通过提高迁移速度来实现更高的 SLA 成果，这样可以提高基础架构效率和最大限度地提升投资回报率。

参考资料：

[1] http://blogs.vmware.com/vsphere/2011/02/vmotion-whats-going-on-under-the-covers.html

[2] http://sparrowangelstechnology.blogspot.com/2012/11/vsphere-51-vmotion-best-practices.html

北京市朝阳区望京东园七区保利国际广场 T1 15 层
Tel: 010-5789 2000
www.mellanox.com