IBM 和 Mellanox

为复杂的建模、研究和分析启用高可用性的弹性存储

执行摘要

在当今全新的计算时代，处理复杂计算的应对能力已经无法满足对快速分析的需求。多核处理器、集群横向扩展计算和价格低廉的 RAM 已经对提供快速有效地运行大型和复杂模拟所需的强大基础架构作出了回答。

目前已经具备解决这些类型的计算所需的计算能力。但管理这些计算产生的不断增长的非结构化数据，已成为现今面临的新挑战。随着各种研究学科对超级计算机能力的需求不断增长，包括气候变化建模、物理和生命科学研究以及高频率交易的研究集群，大量非结构化数据的交付正在飞速发展。这种数据通常被称为"临时数据"，在模拟运行期间使用，但并不总需要存储以供长期访问使用。存储基础架构的最重要标准是它必须与企业可靠性的能力以及超级计算的高带宽和容量相符。重要的是选出适合本页内容的带宽，因为我们将要讨论的是 56 Gb 以太网基础架构而不是 40 Gb 以太网的使用。关于这些环境中的存储问题，性能通常与网络绑定在一起，因此可实现的性能主要取决于所使用的网络技术及其扩展能力。

Mellanox 56Gb 以太网与 40Gb 以太网成本相同，但成效更佳

- 吞吐量高出 40%
- 处理能力高出 2.5 倍
- 恢复时间缩短
- 亚微秒级延迟
- 信息速率最高
- 线性可扩展性

IBM 弹性存储服务器

IBM 弹性存储服务器 (ESS) 是高性能、通用并行文件系统 (GPFS) 网络共享磁盘解决方案，非常适合提供对高级集群服务器数据的快速、可靠访问。在集群服务器上运行的应用程序可以使用标准文件系统接口轻松访问文件，还可以通过多个服务器或协议同时访问同一个文件。IBM ESS 旨在使用一个或多个构建块，其中一个构建块是一对带有共享磁盘附件的服务器。GPFS 软件作为存储服务器的基础文件系统，它借助"分簇"RAID 技术提供出色的吞吐量、极高的数据完整性，同时增强数据保护并加快重建时间。这种组合缩短了计算结果的交付时间。与在超级计算世界一样，不同于关注数据可用性、数据保护和安全归档等数据问题的传统企业存储，它关乎的是交付结果。使用已存储数据的能力更多在于以有帮助的方式和更高的频率在极端的时限内回答问题。

为了解 IBM ESS 如何满足企业和超级计算的存储要求，可将其与传统存储阵列进行对比，您会发现它更倾向于是一种以强大的服务器前端作为接口的集群超级计算环境。主要构建块是带有两个 10 核 POWER8 处理器（已经过进一步加速且加载了 RAM 和 SSD）和扩展柜（有许多存储插槽，可容纳各种存储设备，包括小格式和大格式的 HDD 与 SSD）的高性能电源系统服务器。由于其提供低延迟、高性能以太网或 InfiniBand 连接和极高数据完整性的能力，ESS 具有最快的重建时间为了向非结构化数据提供无与伦比的 I/O 性能，在多个磁盘和多个节点上进行数据剥离，以启用高性能元数据扫描，从而帮助实现最快的答复时间。

吞吐量要求

数据访问可通过具有多种网络选项的 TCP/IP 或 InfiniBand 连接完成，包括 10 Gb 以太网、40 Gb 以太网、FDR 和 EDR InfiniBand。但很少有人知道 Mellanox 的 56 Gb 以太网接口。这使 IBM ESS 以与 40 Gb 以太网相同的成本提供较 40 Gb 以太网高出 40% 的吞吐量。在计算和分析集群中，需要对庞大的数据存储库进行可靠、快速的访问。存储基础架构中的带宽越多，系统应对大量非结构化数据存储需求的能力就会越高。毕竟，如果您注重的基本要求是答复或结果的交付时间，那么高性能解决方案将是适合您的选择。在 56 Gb 以太网上，可轻松超过光纤通道存储实现的最大速度（使用 32 Gb 光纤通道），但考虑到编码、帧间间隔和帧头，仅实现了 28.05 Gb/秒传输速率的线路速率。以 56Gbps 运行 IBM ESS，使其成为可用的最快生产存储设备。

延迟和消息速率

交易平台，特别是那些支持算法交易的平台，需要小于 5 微秒的延迟，并且具有极低的数据包丢失水平。Mellanox 是全球领先的高速以太网和低延迟网络解决方案提供商，率先推出 56GbE 的适配器速度。使速度从 10 或 40 跳到 56GbE，将使消息递送速率呈指数增长。这确保了网络和与服务器的连接不会产生瓶颈，阻碍未来解决方案最佳出现时机。低至几微秒的超低适配器延迟和不到 300 纳秒的交换机延迟，相当于接近 300 万 PPS 的每秒数据包 (PPS) 超高速率。56GbE 的更快速度和更低延迟实现了更快的执行速率，并允许进行更多在线处理。

复杂研究

自创建以来，GPFS 文件系统就已在各种类型和规模的集群中部署，其中更大集群为大学院校和国家研究实验室的科学计算需求提供服务，中小型和较大集群为高性能计算 (HPC) 应用提供服务。分布式锁定体系架构可与这些可扩展的通用文件服务应用程序良好匹配，特别是当工作负载包括访问不同文件集的不同系统的大型集合时。通过 GPFS 进行文件访问与本地文件系统一样高效，但它具有独特的扩展能力，满足不断增长的带宽和容量需求。这使得 IBM ESS 解决方案具有出色的性能，并可与 56 Gb 以太网兼容，从而使研究人员能够进行复杂的研究或建模，以设计出最高性能的解决方案。

缩短恢复时间

利用超级计算机进行分析和提供最快响应时间的研究领域为了跟上发展步伐，需要更大的存储环境，较小却非常重要的硬盘驱动器可能每周都会出现故障。存储的兆兆字节数据越多，磁盘越多，发生故障的可能性越大。在这些环境中，磁盘出现故障已成为常态，而不是罕见的特例。在典型的 RAID 方案中，当磁盘出现故障时，所有剩余磁盘将成为重建过程的关键。但在 GPFS 使用的分簇阵列方案中，只有很小一部分磁盘对重建至关重要。仅重建那些关键带区，只需几分钟便可使系统恢复正常。仅当系统上没有用户 I/O 活动时，才可重建驱动器其余的非关键区域。这避免了对用户应用程序带来任何性能影响，并且是高性能、大容量文件系统（如 IBM ESS）中的一个重要功能。

在常规 RAID 保护的存储阵列中，重建单个 TB 驱动器可能需要长达 12 小时的时间，并且重建期间所需的 I/O 和处理能力可能会对子系统的处理能力、性能和对可用带宽的整体活动带来不利影响。IBM ESS 解决方案支持在一小时内完成 TB 级磁盘快速重建。由于 GPFS 本机 RAID (GNR) 是一种高级备用空间磁盘布局方案，它通过集成到 ESS I/O 服务器中的软件 RAID 算法来统一扩展或 "拆分" 数据，因此这是可行的。它消除了当使用在大多数典型存储阵列中使用的标准不对称 RAID 算法时出现的缺点。通过使用精确的数据算法并在更大量的操作磁盘上扩展用户数据、冗余信息和备用空间，以及利用 56 GbE 速度，实现最大可能的带宽并显著缩短从磁盘故障恢复所需的时间。

结论

首先，对于那些希望缩短必须通过一组庞大的数据池交付结果的时间的组织，具有 GPFS 和 GNR 的 IBM ESS 是适合其需要的高性能存储解决方案。IBM 和 Mellanox 之间的成功合作带来了几个互连解决方案，包括 10 GbE、40 GbE、FDR 和 EDR InfiniBand，确保适合任何环境。然而，对于在大数据 MapReduce 分析、基因测序、数字媒体和可扩展文件服务中，借助 IBM DB2® 等数据库，为业务分析和从气候建模到龙卷风模拟的高性能计算应用提供支持的集群，可使用 56 GbE 以实现 40% 的带宽增加且不会产生额外费用。这是迄今为止您可以在当今市场上找到的最快响应速度的解决方案。

北京迈络思科技有限公司
咨询电话：+86-10-57892000
销售咨询：china_sales@mellanox.com
市场合作：marketing_cn@mellanox.com
*欲了解更多欢迎登陆www.mellanox.com

微信