案例研究

在 GPFS NSD 内池化 NVMe 实现空前的突发缓冲区带宽级别

背景

SciNet 是加拿大的最大超级计算机中心，为加拿大研究人员提供必要的计算资源和专业知识，以便大规模地开展研究。该中心帮助推动从生物医药科学和航空航天工程到天体物理和气候科学等各方面的研究工作。

新的基于 InfiniBand 的 SciNet 超级计算机位于多伦多大学，需要达到高级别的可用性，才能确保其用户得到很高的投资回报。利用 EDR InfiniBand 技术，该超级计算机提供多项技术革新和网络加速，具有世界级的应用程序性能。此外，该系统还利用 Dragonfly+ 拓扑，随着计算和存储需求的增长而实现无缝扩展。另一项有趣的技术革新是使用突发缓冲区，实现智能、快速的检查点创建。

挑战和解决方案

高性能计算应用包含可能运行数周的复杂进程集合。中断任何进程均可能破坏整个计算作业的结果。随着超级计算机变得越来越强大，该问题愈加明显，想象一下加拿大最大超级计算机的挑战。因此，并行计算应用程序在中断情况下使用检查点重启 - 允许从最近保存的检查点重新启动计算作业的技术。

检查点通常保存在共享的并行文件系统中；SciNet 选择了 GPFS。但是，随着集群变得越来越大和每个节点的内存用量增加，每个检查点也变得更大，要么需要更多时间完成，要么需要高性能文件系统。当系统正在建立检查点时，它不执行计算任务，从而减少了系统的可用性得分。为缩短这些不可用的时刻，SciNet 决定利用 Mellanox 互连加速和 Excelero 的 NVMesh 实现基于 InfiniBand 的突发缓冲区。

联合解决方案将 Excelero 的 NVMesh 与 Mellanox 全球领先的端到端 InfiniBand 网络解决方案相结合，使 SciNet 能够构建 PB 级规模的分布式高性能 NVMe 统一池，作为用来创建检查点的突发缓冲区。NVMe 池提供 230GB/s 的吞吐量和远超过 20M 随机 4k IOPS，使 SciNet 能够满足其可用性 SLA。

亮点

使用案例

大规模建模、仿真、分析和可视化

挑战

在 15 分钟内完成检查点以符合可用性 SLA

解决方案

NVMesh 支持 PB 级规模的分布式高性能 NVMe 闪存统一池，作为用来创建检查点的突发缓冲区

结果

- 80 个池化的 NVMe 设备
- 148 GB/s 的写数据突发（受设备限制）
- 230GB/s 读数据吞吐量（受网络限制）
- 远超过 20M 随机 4k IOPS

优势

- 满足 15 分钟检查点窗口
- 极具成本效益
- 空前的突发缓冲区带宽

"对于 SciNet 来说，NVMesh 是达到前所未有的突发缓冲区带宽的极具成本效益的方法。"

首席技术官
Daniel Gruner 博士
SciNet 高性能计算联合会

关于突发缓冲区

突发缓冲区是计算节点非持久内存与永久存储（并行文件系统）之间的快速中间存储层。这一层被配置为以非常高的速率接收大量的写IO。在突发（检查点）完成后，使用 GPFS 策略引擎将写入的数据"排放"到并行文件系统。这使得检查点能够迅速完成，以便系统满足可用性 SLA. 当使用闪存作为突发缓冲池时，增加了有利于（需要时）更快重新启动的优势，因为检查点重启通常会对底层存储形成非常大的随机读取负载。考虑到这些好处，将 SciNet 突发缓冲区的大小设定为容纳两个检查点，以便最近完成的检查点能够用于重启。为最大限度地提高性能和缩小检查点窗口，SciNet 决定利用更高性能的 NVMe SSD。

达到空前的 NVMESH® 突发缓冲区带宽

"对于 SciNet 来说，NVMesh 是达到前所未有的突发缓冲区带宽的极具成本效益的方法。"SciNet 高性能计算联合会首席技术官 Daniel Gruner 博士说，"通过向计算节点以及为超级计算机本身提供的低延迟网络架构添加商用闪存驱动器和 NVMesh 软件，NVMesh 在不影响目标 CPU 的情况下提供了冗余。这让标准服务器能够在充当数据块目标方面超过其正常作用–服务器现在还能充当文件服务器。"

Excelero 能够让客户构建高性能突发缓冲区，而不需要其他专用阵列。NVMesh 客户能够在其应用程序服务器中使用标准 NVMe 驱动器来构建本地突发缓冲区，或构建聚合文件系统/数据块服务器装置。

这种方法具有增加冗余与集中化管理的优势，同时为应用程序自身保留所有计算资源。

NVMesh 及其建立在 Mellanox 远程直接内存访问 (RDMA) 之上且获得专利的 Remote Direct Drive Access (RDDA) 技术，允许客户将计算节点中的 NVMe 驱动器与 CPU 资源在逻辑上分离开来。这样一来，远程计算节点能够使用本地 NVMe 驱动器，而不消耗本地 CPU。因此，每个计算节点的本地 NVMe 驱动器进行了池化，以供集群使用。在过分简单化的方式下，每个驱动器的一半可以用作本地突发缓冲区，另一半继续保留为对等的冗余副本。因此，当一个节点出现故障时，其暂存数据得以保留，可供其他节点（架构上的任一节点）访问。

"在超级计算中，任何不可用性都会浪费时间，减少系统的可用性得分，阻碍科学探索的进度。我们很高兴为 SciNet 及其研究人员提供重要存储功能，以明显降低的价格实现目前行业内最高的性能 – 同时确保至关重要的科学研究能迅速得到发展。"Excelero 共同创立人兼首席执行官 Lior Gal 说。

MELLANOX 的端到端 EDR INFINIBAND

Mellanox 的端到端 EDR InfiniBand 采用世界最快的存储网络技术，支持任何主要网络架构的最高带宽 (EDR 100Gb/s) 和最低延迟（小于 90 纳秒端口到端口）。

Mellanox EDR 解决方案为高性能数据中心简化了高速网络的部署和管理，优化要求最苛刻应用工作负载的总体性能、功率和密度。提供最高的数据速度和性能增强的 CPU 卸载，Mellanox InfiniBand (IB) 解决方案最大限度地提高数据中心投资回报，降低拥有成本。

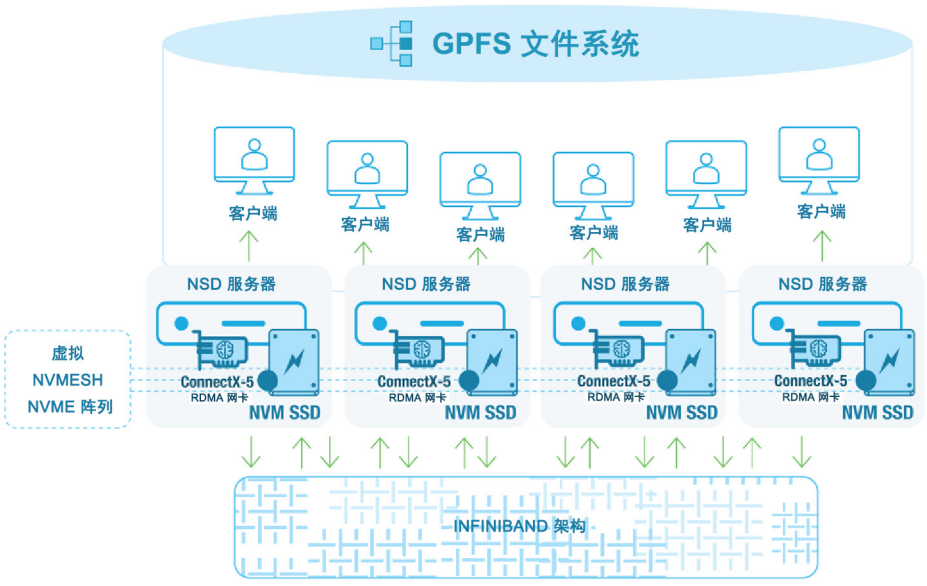

图 1. Excelero-Mellanox 联合解决方案，提供高速度、低延迟、共享存储，具有服务器内的闪存性能

全球第一智能网络交换机

Mellanox 行业领先的智能 Switch-IB™ 2 是世界第一台智能网络交换机，针对超低延迟无损网络架构进行了优化。为支持 SciNet 和其他苛求性能的数据中心而构建，智能 Switch-IB™ 2 是面向超融合基础架构部署的理想架顶式交换机，可达到目前市场上最高的架构性能。

Mellanox 支持 RDMA 的 InfiniBand ConnectX®-5 网络适配器卡，搭配智能 ASIC，提供高级 NVMe over Fabric (NVMe-oF) 目标卸载功能，提高 NVMe 存储访问效率级别，无需任何 CPU 介入，达到小于 600 纳秒的延迟级别。由于可绕过 CPU，RDMA 为存储和超级计算任务腾出了 CPU 资源，尤其是在计算密集的环境下，从而实现更高的可扩展性和数据中心效率级别。

"Mellanox 的智能、可扩展 NVMe 加速让用户能够最大限度地提高其存储性能和效率。"Mellanox Technologies 市场营销副总裁 Gilad Shainer 说，"利用 InfiniBand 的优势，Excelero 提供全球领先的 NVMe 平台，对下一代超级计算机进行加速。"

协同设计的创新 SHARP™ 技术

Mellanox 行业领先的 100Gb/s InfiniBand 智能交换机通过协同设计的 SHARP™ 技术，使 SciNet 具有网络内计算的能力，实现最高 7.2Tb/s 的无阻塞带宽以及 90 纳秒端口到端口延迟。由 Mellanox 自有的 ASIC 驱动，36 端口无阻塞 EDR 100Gb/s IB 智能交换机在 40/56/100 Gb/s 线速下具有直通交换能力，无数据包丢失。动态共享的交换机缓冲区提供最佳的突发流量采用，因而让存储解决方案能够提供有保证的吞吐量和延迟。Mellanox 交换机与更快的拥塞通知相结合，形成一种网络架构，充分发挥整个网络内 NVMe 存储池的最大力量。

联合解决方案

Mellanox-Excelero 联合解决方案提供远程、高速度、低延迟共享存储，具有服务器内闪存性能。在现有超级计算应用集群上部署该解决方案，将其转化成具有极高级别的计算、存储和应用性能的融合基础架构。

Excelero 的 NVMesh RDDA 技术使苛求性能的超级计算应用程序能够享有底层服务器和存储设备的全部性能、容量和处理能力，Mellanox 的 InfiniBand RDMA 技术利用智能加速来推进联合解决方案，启用网络内计算，确保各种应用程序工作负载得到更快的数据处理、更高的性能和效率。联合解决方案最大限度地提高总体系统性能，使客户能够以最佳性价比达到最高的性能级别。

SCINET 的 NVMESH 突发缓冲区实现

利用 Excelero 的 NVMesh，SciNet 能够创建 PB 级规模的分布式高性能 NVMe 闪存统一池，保留直连式介质的速度和延迟。在只有 10 台 NSD 服务器中包含 80 个 NVMe 设备的 NVMe 池，提供大约 148GB/s 的写数据突发（受设备限制）和 230GB/s（受网络限制）的读数据吞吐量以及远超过 20M 的随机 4k IOPS。此配置满足 15 分钟检查点窗口绰绰有余，这也是满足为新超级计算机定义的可用性 SLA 所必需的。

对于 SciNet 来说，通过向计算节点和低延迟网络架构添加商用闪存驱动器和 NVMesh 软件，NVMesh 已经能够达到空前的突发缓冲区带宽。NVMesh 提供冗余，而不影响目标 CPU，使标准服务器不仅能充当数据块目标，而且能充当文件服务器。而且，NVMesh 看起来像一个简单的块设备，所以，与 SciNet 的并行文件系统集成显得非常直接。

NVMESH 用于突发缓冲区的好处

- PB 级规模的高性能闪存的统一池，保留直连式介质的速度和延迟。
- 支持大规模建模、仿真、分析和可视化。
- 在数百个计算节点上可视化超级计算机仿真数据。
- 实现快速检查点创建和计算机作业重启。
- 以最低的价格达到最高的性能。
- 在网络上大规模利用 NVMe SSD 的全部性能。
- 线性地扩展您的性能和容量。
- 容易管理和监控，减少维护总体拥有成本。
- 可使用任何服务器、存储设备和网络供应商的硬件。不受供应商限制。

关于 SciNet

SciNet 是加拿大的最大超级计算机中心，为加拿大研究人员提供必要的计算资源和专业知识，以加拿大前所未有的规模开展研究。SciNet 推动从生物医药科学和航空航天工程到天体物理和气候科学等各方面的研究工作。SciNet 是利用超级计算驱动创新的国家基础架构 Compute Canada 的一部分，其创立者是 CFI、NSERC、安大略政府、Fed Dev Ontario 和多伦多大学。了解详细信息：www.scinethpc.ca/about-scinet

关于 Mellanox

Mellanox Technologies 是针对服务器和存储的端到端InfiniBand及以太网互连解决方案和服务的领先提供商。Mellanox 互连解决方案可提供最高吞吐量和最低延迟，更快地向应用程序传递数据并充分发挥系统性能，从而提高数据中心效率。Mellanox 提供一系列快速互连产品：适配器、交换机、软件、线缆和芯片，它们可针对广泛的市场（包括高性能计算、企业数据中心、Web 2.0、云、存储和金融服务）加快应用程序运行时间并最大程度实现业务成果。了解详细信息：www.mellanox.com

关于 Excelero

使用 Excelero，企业和服务提供商能够利用标准服务器和高性能闪存设计横向扩展存储基础架构。它于 2014 年由一批存储领域的资深人士创立，受到科技巨头的 Web 规模应用程序无共享体系架构的启发，该公司设计了一种软件定义的块存储解决方案，可满足最大的 Web 规模和企业应用程序的性能和可扩展性要求。了解详细信息：www.excelero.com

关于 NVMesh

使用 Excelero 的 NVMesh，客户能够为混合应用程序工作负载构建分布式、高性能服务器 SAN。客户受益于本地闪存的性能，具有集中式存储的方便性，同时避免受专有硬件的限制，并且减少了总体存储拥有成本。

已经为超大规模工业物联网服务、机器学习应用程序和大规模仿真可视化部署了该解决方案。了解详细信息：www.excelero.com/product/nvmesh

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

060077CS Rev1.0