IBM Spectrum Scale 5.0.0 的 IO 性能

Silverton Consulting, Inc. StorInt™ 简报

RAYONSTORAGE.COM |GREYBEARDSONSTORAGE.COM
SILVERTONCONSULTING.COM | TWITTER.COM/RAYLUCCHESI

前言

高性能计算（HPC）和科学计算都处在持续演变期。人工智能（AI）、机器学习（ML）和深度学习（DL）都属于变革性的开发成果，都曾经历过长期的孕育期，但现在都已经成为主流。AI/ML/DL 以及大数据/数据分析及数学建模已成为科学计算的支柱，而且在未来的十年及更长时间内均是如此。

IBM Spectrum Scale™ 存储系统已发展成为一款领先的存储解决方案，在某些情况下，可供科学计算和 HPC 社区使用。在该解决方案的最新版 5.0.0 中，IBM® 引入了一些增强功能，旨在应对 AI/ML/DL 活动的一些新的 IO 特点。具体来说，IBM 提升了这些新工作负载所需的小文件 IO 性能。

下面我们将探讨 IBM Spectrum Scale 5.0.0 的一些性能增强，并介绍为何这些增强功能非常适于美国的一个实验室计划，该计划已经开发出了全球最快的 HPC 超级计算环境，旨在应对新的工作负载及未来的工作负载。

IBM Spectrum Scale 5.0.0 的增强功能

小文件的 IO 在元数据和小数据块方面非常密集。若要在这些限制条件下确保性能，存储系统必须执行元数据状态并实现超快速的更新，同时非常快速地读写小数据块。

Spectrum Scale 5.0.0 显著改善了小文件的 IO 性能。一些专门针对小文件 IO 性能的增强功能包括：

- **增强了远程直接内存访问（RDMA）支持**：有助于加速集群间的通讯，而且能够在确保降低开销的情况下执行节点间的数据传输。
- **新的"无锁定"读取路径**：有助于提升并行性，并缩短读取串行化的等待时间。
- **新的多层写入缓存**：使用 NVMe 存储来加速小文件块的写入。
- **可变的 SubBlock（子数据块）大小**：有助于改善小空间的效率和系统 IO 性能。

对于大文件而言，大多数的 IO 都耗费在读取或写入大数据块上面。尽管吞吐量非常重要而且元数据也需要扩展，但大文件的元数据其实并不会经常访问。相比而言，小文件的元数据不仅要扩展到更多文件，而且访问得更加频繁，这样才能确保性能。此外，系统还需要能够更快地读取和写入小数据块。

若要实现更快的元数据更新，就需要确保节点间的通讯响应更快速且响应时间可预测。若要确保系统完整性及元数据的有效性，必须以确保原子性、一致性、隔离性和持久性 (ACID) 的方式完成更新。若要规模化地执行此类更新，需要使用高效的跨节点通讯或集群间通讯一致地在节点之间更新一个类似数据库的大型结构。在 Spectrum Scale 5.0.0 中，IBM 利用 RDMA 等协议技术提升了集群间的通讯速度。通过这种方式，元数据访问和集群间的小数据块通讯就不会出现性能瓶颈。

不过，一致的元数据更新通常需要串行化，而串行化会拖慢 IO 性能。在此情况下，Spectrum Scale 5.0.0 降低了元数据访问及小数据块读取活动的串行化需求，进而改善了小文件的性能。

近期，存储行业引入了新的 NVMe 固态硬盘 (SSD)，以改善 IO 性能。NVMe 协议从整体上了改变了系统的接口驱动构造，能够显著减少驱动器的 IO 响应时间/延迟。IBM Spectrum Scale 5.0.0 利用速度更快的新 NVMe 存储技术，引入了一个多层小文件块写入缓存，可大幅提升小文件的写入性能。

在之前各代的 Spectrum Scale 产品中，IBM 采用了被称作 SubBlock 的结构，同时采用固定分区，以将文件数据传输映射到存储系统的物理 IO。这种 SubBlock 结构代表的是 IO 性能与空间效率之间的权衡。在 Spectrum Scale 5.0.0 中，IBM 引入了一种可变的子数据块结构，该结构能够按照系统所用文件的大小来优化系统 IO。通过这种方式，无论是大文件还是小文件，均可使用根据它们的需求而定制的最优系统 IO 大小。

通过这些增强功能，使得 IBM Spectrum Scale 5.0.0 的小文件 IO 性能得到了大幅提升。此外，通过提升集群间的通讯速度、加快元数据的访问速度、减少读取串行化等方式，也提升了大文件的 IO 性能。

举例来说，无论是小文件传输还是大文件传输，节点间的数据传输都非常普遍。借助在节点间数据传输方面的增强功能，无论是小文件还是大文件，它们的读/写 IO 均可执行更多的传输，同时还能够降低开销，进而从实质上改善了大文件的 IO 性能。

除了这些性能增强功能之外，IBM Spectrum Scale 5.0.0 的操作和部署也变得更加轻松。在最新版的 5.0.0 发行版中，许多之前需要手动指派的默认参数设置都实现了自动化，这样便可大幅提高 IBM Spectrum Scale 的开箱即用部署速度。此外，IBM Spectrum Scale 5.0.0 中还增加了一些新的安全和合规功能，包括新文件事件审计日志记录、文件不可变性以及动态数据和静态数据安全功能等。

新的 AI/ML/DL IO 需求

机器学习（ML）、深度学习（DL）和人工智能（AI）领域的最新趋势均涉及到了神经网络的多阶段分类计算。这些阶段包括（1）数据摄入；(2) 数据清理和转换；(3) 神经网络探索和架构设计；(4) 神经网络训练；及 (5) 推理生成。尽管神经网络探索和架构设计阶段涉及的 IO 不太多，但其他阶段都需要在各种数据集之间进行一道或多道 IO。

通过当今的 AI/MI/DL 算法而积累的一个关键洞察力是：用于训练神经网络所用的数据越多，结果的质量就越高。换言之，采用更大规模的训练数据集能够改善神经网络分类的准确性。因此，AL/ML/DL 数据集的规模越大，这些算法的性能就越高。

在任何 AI/ML/DL 算法开发中，第一个阶段都是**数据摄入**。该阶段是指将数据摄入到 ML/DL 框架中。一般而言，该阶段的数据通常都是大文件，而且从 ML/DL 的角度来说，它们均属于读写顺序 IO。在该阶段结束时，ML/DL 框架中的所有数据都已准备就绪，可用于后续处理。

接下来是**数据清理和转换**阶段，这是一个冗长的阶段，主要是对数据进行清理并将其从原始格式转换为可用于训练 ML/DL 神经网络的格式。从本质上来说，该流程一般是属于读/写顺序，读取大文件进来，然后写出经正确分类的小文件。

再然后是**神经网络探索和架构设计阶段**，该阶段主要是构建 ML/DL 算法所需的神经网络架构模型，之后便可开始时间训练。作为**神经网络训练**各个阶段的一部分，所有经过清理的小训练文件会由 ML/DL 框架进行随机读取，再使用当前的神经网络进行实验评估，然后对神经网络进行修改，以减少（分类）错误。在该流程中，需要针对训练数据集中的大量小文件进行多次随机读取。（神经网络训练主要通过 GPU 完成。）

在完成 AI/ML/DL 神经网络训练之后，即为**推理生成**阶段。在该阶段，会使用 ML/DL 神经网络算法针对新的输入数据作出推理。在推理阶段，IO 活动一般都是以实时的方式按顺序读取小传感器、视频、音频帧。

总而言之，AI/ML/DL 在各个阶段的 IO 需求都不相同，但基本上都会处理大量的数据。在这些数据中，一些是按顺序读取和写入的大文件，但在数据清理及后续流程中所发生的大多数 IO 都涉及到了大量小文件的随机读取和写入。

某高性能计算实验室在存储系统方面的新需求

某高性能计算实验室的超级计算机正在逐渐被淘汰，其目的在于提供一台新的世界级超级计算机，以支持最新的计算与存储服务。这种新的第四代超级计算环境能够支持其研究社区成员所需的混合 CPU-GPU AI/ML/DL 计算工作负载及其他科学计算。

该超级计算机将由约 4,600 个混合（2 个 IBM Power9 CPU 和 6 个 NVIDIA Volta V100 GPU）计算节点构成，每个节点包含 0.5 TB 的 DRAM 及 1.6 TB 的永久存储。这些节点将会通过双轨 EDR InfiniBand 网络连接，可提供 23GB/秒的节点注入带宽，还包括 250 PB 的文件存储容量。

这个为多个高性能计算实验室构建的下一代的超级计算机能力强大。接下来我们将会讨论它的存储系统的 IO 性能需求，以及 IBM Spectrum Scale 5.0.0 如何满足或超越所有这些规格需求。

超级计算机 IO 子系统的需求如下：

- 每秒至少完成 50,000 次文件访问；
- 1MB 顺序读/写总计带宽最低达到 1TB/秒；
- 顺序读/写总计峰值带宽达到 2.5TB/秒；且
- 每秒总计完成 260 万次 32 Kb 文件（小文件）的访问。

IBM 为超级计算机配置的存储系统

IBM 的 IBM Spectrum Scale 存储平台参与了超级计算机存储系统的投标。参与投标的系统由一个包含 77 个节点的 IBM Storage™ Server (ESS) 集群组成，该集群在 IBM Spectrum Scale 5.0.0 上运行。IBM ESS 的节点使用两台双插槽 IBM POWER9 存储服务器配置为 ½ (20U) 机架构建块，每个服务器配备 1 TB 的内存及 4 个 4U/106 驱动机箱（每个机箱配备 104 个磁盘及 2 个 NVMe SSD），每个节点的原始磁盘存储容量可达到 4 PB。每个 IBM ESS 节点通过一个 4X EDR InfiniBand 网络与集群中的其他节点相连接，该网络最高可实现 90GB/秒的网络带宽。

IBM 存储系统的测试结果

正如之前所讨论的，超级计算机具有许多独特的 IO 子系统性能目标，而 IBM ESS 集群必须满足这些目标。

	超级计算机的 IO 需求	IBM ESS 构建块的性能	IBM ESS（77 个节点）的集群性能
每秒的文件创建量	50,000/秒	~57,000/秒	
1MB 读/写顺序带宽	1 TB/秒	客户端 IOR 至 ESS：23 GB/秒 TQOSPERF：16 GB/秒	TQOSPERF：1.2 TB/秒
峰值顺序读/写带宽	2.5 TB/秒	读取：43.4GB/秒 写入：36.3GB/秒	读取：3.3 TB/秒 写入：2.8 TB/秒
32K（字节）文件创建量/秒	2.6 M/秒	~56,000/秒	4.3 M/秒

首先，超级计算机存储系统的小文件 IO 性能必须达到或超过每秒 50,000 个小文件的创建量。IBM ESS 解决方面每秒大约可创建 57,000 个文件。在此次测试中，IBM ESS 使用的是 1 Kb 的小文件，并通过 23 个客户端节点完成测试。该测试在 23 个节点中使用了小文件 IO 和一个共享式目录。IBM Spectrum Scale 5.0.0 的许多性能增强推动了这次测试的顺利完成，尤其是在节点间通讯速度方面的提升。

其次，超级计算机要求 1MB 顺序读/写总计带宽最低要达到 1TB/秒。为此，我们对 IBM Spectrum Scale 5.0.0 ESS 进行了两次测试：一次是使用 Lustre IOR，另一次是使用 GPFS TQOSPERF。两次测试均针对单个客户端/单个文件顺序写入工作负载进行配置。IOR 测试结果显示，每个 ESS 节点的顺序写入性能是 23GB/秒，而 TQOSPERF 的基准测试结果显示，每个 ESS 节点的顺序写入性能是 16GB/秒。当扩展到 77 个 ESS 节点后，两个测试的顺序写入带宽均超过了 1TB/秒。在这一方面，新的多层小数据块写入缓存起到了非常大的作用。

第三，该超级计算机要求顺序读/写总计峰值带宽达到 2.5TB/秒。在针对约 49TB 的文件使用 16 MB 的数据传输规模时，IBM ESS 存储系统的峰值顺序写入性能是 36.2GB/秒，峰值顺序读取性能是 43.4GB/秒。当扩展到 77 个 ESS 节点之后，顺序读取和写入性能可轻松超过 2.5TB/秒。

第四，要求每秒总计完成 260 万次 32 Kb 文件的创建。IBM ESS 与 IBM Spectrum Scale 5.0.0 的组合可通过 32 KB 小文件的非共享目录实现这一要求，而且每个 ESS 节点也可实现每秒 5.6 万个文件的创建量。当扩展到 77 个 ESS 节点之后，IBM ESS 每秒将能够实现超过 430 万个文件的创建量。

额外性能成效

在元数据活动方面，IBM ESS 与 IBM Spectrum Scale 5.0.0 的组合执行了单个线程（从客户端到服务器再到设备）的小数据块随机读取和写入操作，分别实现了平均 80 µs 和 200 µs 的响应时间。

此外，在多线程小数据块读取访问方面，之前一代的 IBM Spectrum Scale (4.2.2) 在使用 4 Kb 传输规模的情况下，3 到 4 个线程时可实现的峰值性能不到 40 万次操作/秒；而借助 IBM Spectrum Scale 5.0.0，在使用 4 Kb 传输规模的情况下，16 个线程时可实现的峰值性能大约可达到 270 万次操作/秒，**在小文件读取 IO 性能方面提升了 700%。**

最后，之所以能满足上述需求，很大程度上有赖于 IBM ESS 能够将单个节点的性能扩展到 77 个节点的能力。IBM 能够在多个节点中提供经验证的 ESS 节点性能。举例来说，IBM ESS 与 IBM Spectrum Scal 5.0.0 的组合在 12 个节点中，每秒可完成超过 500 万次远程程序调用 (PRC)。

总结

IBM ESS 与 IBM Spectrum Scale 5.0.0 的组合在各个方面都满足或超出了超级计算机的 IO 子系统性能需求。IBM Spectrum Scale 5.0.0 包含有大量设计和代码方面的增强功能，可显著提升小文件的 IO 性能，在单个用例中，可将小文件块的读取 IO 性能提升 700%。

对于当今的多工作负载 HPC/科学计算环境而言，它们要求存储系统可实现多个 TB/秒的大文件带宽、1TB/秒或以上的小文件带宽，以及每秒数百万个文件的创建量，这样才能支持容量可能超过 200PB 的存储需求。在这个环境中，IBM ESS 与 IBM Spectrum Scale 5.0.0 的组合可提供其他产品难以企及的世界级性能。

Silverton Consulting, Inc. 是美国的一家存储、战略和系统咨询功能，致力于为数据存储社区提供产品和服务。

免责声明：本文档的编写获得了 International Business Machines Corporation (IBM) 的资金支持。尽管本文档使用了 IBM 等各方公开发表的资料，但这并不一定反映了他们对本文档中所述问题的立场。