解决方案简介

分布式异步对象存储 (DAOS)
英特尔® 傲腾™ 技术

DAOS：借助英特尔® 傲腾™ 技术推动高性能存储变革

随着数据呈指数级增长，分布式存储系统不仅成为了数据中心的核心，同时也成了其主要的瓶颈。数据访问延迟高、可扩展性差、管理大型数据集难度大、缺乏查询功能，这些都是常常会遇到的阻碍。传统存储系统是针对旋转介质和 POSIX* 输入/输出 (I/O) 所设计的。这类存储系统出现了巨大的性能瓶颈，很难通过升级来支持新的数据模型和下一代工作流程。

高性能计算、大数据和人工智能的融合

存储需求不断发展，而需要处理的数据集也在持续增加，这使得消除数据与计算之间的障碍变得愈发迫切。存储不再由具有大量流写入的传统工作负载（例如检查点/重启）主导，而是越来越多地受到新主流存储的复杂 I/O 模式的支配。高性能数据分析工作负载正在生成大量随机读取和写入。人工智能 (AI) 工作负载的读取需求远超传统高性能计算 (HPC) 工作负载。从设备流向高性能计算集群的数据需要更高的服务质量 (QoS) 以避免数据丢失。现在，数据访问速度变得与写入带宽同等重要。数据集的查询、分析、过滤和转换有赖于新的存储语义。所以，能够允许全新工作流程将高性能计算、大数据和人工智能相结合以进行数据交换和通信的单一存储平台非常重要。

DAOS 软件堆栈

英特尔长期致力于为以数据为中心的计算构建完全开源的软件生态系统，并针对英特尔® 架构和非易失性存储器 (NVM) 技术（包括英特尔® 傲腾™ 持久内存和英特尔® 傲腾™ 固态盘）进行了全面优化。分布式异步对象存储 (DAOS) 是英特尔构建的百亿亿次级 (exascale) 存储堆栈的基础。DAOS 是一种开源软件定义横向扩展对象存储，可为高性能计算应用提供高带宽、低延迟和高 IOPS 的存储容器。下一代以数据为中心的工作流程将结合仿真、数据分析和人工智能，而 DAOS 能够为其提供支持。

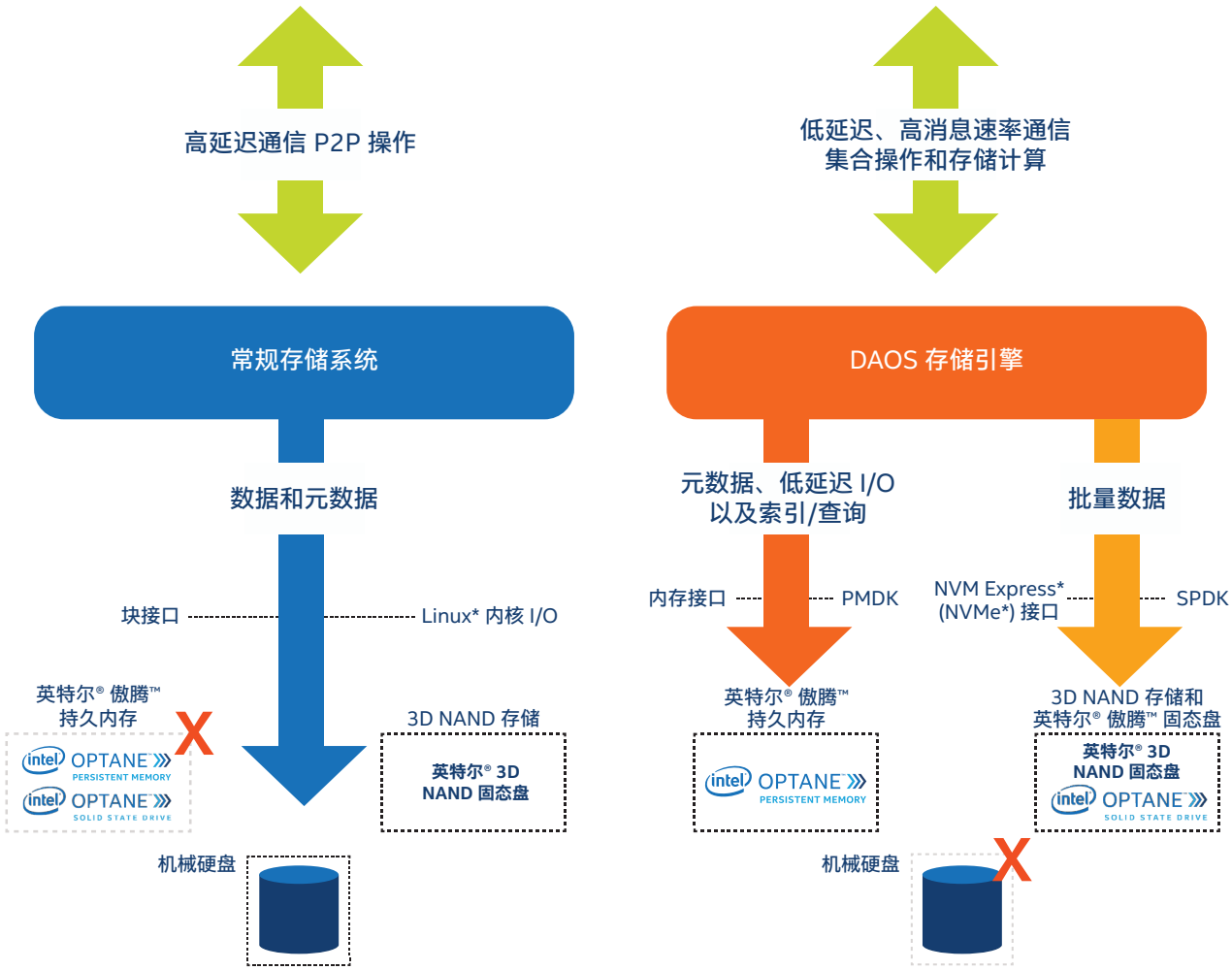

图 1. DAOS 架构与传统存储系统的对比

与主要针对旋转介质设计的传统存储堆栈不同，DAOS 针对全新 NVM 技术进行了重新构建。此外，DAOS 还是一套轻量级的系统，可在用户空间中端对端地运行，并能完全绕开操作系统。DAOS 没有延续针对高延迟、块存储的 I/O 模型，而是选择了为访问高细粒度数据提供原生支持的 I/O 模型，以此释放下一代存储技术的性能。图 1 对比了 DAOS 架构与现有的存储系统。

现有的分布式存储系统使用高延迟的点对点通信，而 DAOS 使用能够绕过操作系统的低延迟、高消息速率用户空间通信。当下，大多数存储系统都是针对块 I/O 设计的，所有 I/O 操作都通过块接口在 Linux* 内核中进行。为了优化对于块设备的访问，业界已经付出了例如合并、缓冲和聚合等方面的许多努力。但是，所有这些优化都无法适用于英特尔着力发展的下一代存储设备，因为使用这些优化可能会产生不必要的开销。而 DAOS 专为优化对英特尔® 傲腾™ 持久内存和 NVM Express* (NVMe*) 固态盘 (SSD) 的访问而设计，它规避了这些不必要的开销。

DAOS 服务器将其元数据保存在持久内存中，而将批量数据直接保存在 NVMe 固态盘中。此外，少量 I/O 操作在聚合之前就会被吸收到持久内存中，然后再迁移到大容量闪存。DAOS 使用持久内存开发套件 (PMDK) 提供对于持久内存的事务访问，并使用存储性能开发套件 (SPDK) 为 NVMe 设备提供用户空间 I/O[1,2]。**这种架构的数据访问速度比现有存储系统快好几个数量级（从毫秒 [ms] 级加快到微秒 [µs] 级）。**

DAOS 软件堆栈提供:

- 超高细粒度、低延迟和真正零拷贝的 I/O
- 非阻塞型数据和元数据操作，以支持 I/O 和计算重叠
- 先进的数据放置，以解决故障域
- 由软件管理冗余，可通过在线重建，支持复制和擦除代码
- 端到端 (E2E) 数据完整性
- 可扩展的分布式事务，提供可靠的数据一致性和自动恢复功能
- 数据集快照功能
- 安全框架，用于管理存储池的访问控制
- 软件定义存储管理，用于调配、配置、修改和监控存储池
- 通过 DAOS 数据模型和 API，为 I/O 中间件库（例如 HDF5*、MPI-IO* 和 POSIX）提供原生支持。应用无需移植代码，即可直接使用 DAOS API
- Apache Spark* 集成
- 使用发布/订阅 API，实现原生生产者/消费者工作流程
- 数据索引和查询功能
- 存储内计算，以减少存储和计算节点之间的数据移动
- 容灾工具
- 与 Lustre* 并行文件系统无缝集成，并能扩展到其他并行文件系统，从而为跨多个存储层的数据访问提供统一的命名空间
- 数据搬运器，用于在 DAOS 池之间迁移数据集，将数据集从并行文件系统迁移到 DAOS，反之亦然

如图 2 所示，DAOS 软件堆栈依赖于客户端-服务器模型。I/O 操作将在与应用直接连接的 DAOS 库中处理，并由在 DAOS 服务器节点 (DN) 上的用户空间中运行的存储服务提供支持。

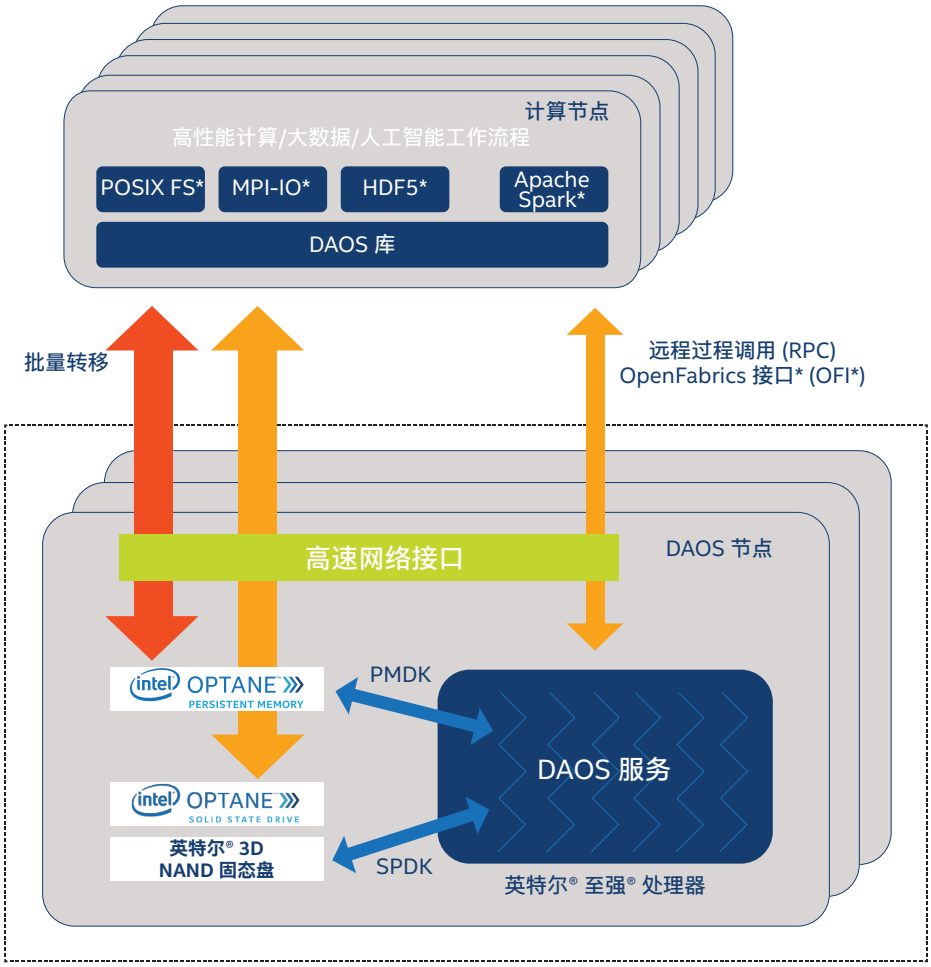

图 2. DAOS 软件堆栈

应用接口和 I/O 中间件集成

DAOS 客户端库的占位极小，可尽可能减少计算节点上的噪声并支持显示进度的非阻塞型操作。利用 libfabrics* 和 OpenFabric 接口* (OFI*)，就可将 DAOS 操作交付给 DAOS 存储服务器，以充分利用架构中的任何远程直接内存访问 (RDMA) 功能。

在这种新的存储范例中，POSIX 不再是新数据模型的基础，而是像其他任何 I/O 中间件一样，POSIX 接口将构建为 DAOS 后端 API 顶部的库。POSIX 命名空间可以封装在容器中，并由应用挂载到其文件系统树中。对于成功打开容器的应用，其中的任何任务都可以访问此应用专有的命名空间。用于解析已封装命名空间的工具也是其所提供的一部分。对于已封装的 POSIX 文件系统，数据和元数据都将以渐进式布局完全分布在所有可用存储中，以帮助确保性能和弹性。此外，POSIX 仿真还具有以下功能：可扩展的目录操作、可扩展的共享文件 I/O、可扩展的文件/进程 I/O以及用于从故障或损坏的存储中恢复的自我修复功能。

尽管大多数高性能计算 I/O 中间件可以通过 POSIX 仿真层在 DAOS 后端透明地运行，但通过迁移 I/O 中间件库实现对 DAOS API 的原生支持后，就能利用 DAOS 丰富的 API 和先进功能。图 3 显示了设想中的 DAOS 生态系统。

DAOS 容器通过多个 I/O 中间件库向应用公开，从而在提供平滑迁移路径的同时，尽可能减少应用代码的修改量（甚至无需修改）。在 DAOS 库顶部运行的中间件 I/O 库包括：

- **POSIX FS：** DAOS 提供两种支持 POSIX 的操作模式。第一种模式是针对生成无冲突操作以支持高并发性的“行为良好”的应用。第二种模式是面向对一致性要求更为严苛但可牺牲部分性能的应用。
- **MPI-I/O：** ROMIO* 驱动程序在 DAOS 顶部为 MPI-I/O 提供支持。凡是使用 MPI-I/O 作为 I/O 后端的应用或中间件 I/O 库都可以在 DAOS 顶部无缝使用此驱动程序。这一驱动程序已推送到上游的 MPICH* 存储库，并且可移植到其他使用 ROMIO 作为 MPI-IO 标准 I/O 实现的 MPI 实施方案中。DAOS MPI-IO 驱动程序直接基于 DAOS API 构建。
- **HDF5：** HDF5 虚拟对象层 (VOL) 连接器使用 DAOS 来实现 HDF5 数据模型。通过 VOL 插件，使用 HDF5 表示和访问数据的应用只需进行少量修改（甚至无需修改）现有 HDF5 API 的代码，就可以利用 DAOS 容器替换 POSIX 文件中的传统二进制格式。该连接器通过原生 DAOS 后端实现官方 HDF5 API。在内部，HDF5 库管理 DAOS 事务并提供从 H5Fopen() 到 H5Fflush()/H5Fclose() 的一致性。另外，还通过 API 扩展提供异步 I/O、快照和查询/索引等新功能。

Silo*、MDHIM* 和 Dataspaces* 等其他高性能计算 I/O 中间件也能从 DAOS API 的原生端口中受益。英特尔与其他企业和机构（例如天气预报机构）以及行业先锋（例如娱乐业、云服务以及石油和天然气行业）密切合作，通过 DAOS 支持新的数据模型。

此外，英特尔正在探索如何在大数据和数据分析框架中实现 DAOS。更具体地说，就是如何为 Apache Arrow* 提供 DAOS 后端。Apache Arrow 标准定义了要存储在列向量中的数据，以支持数据分析用例。此标准的目的是为其他数据分析系统（例如 Apache Spark、Apache Thrift* 和 Apache Avro*）定义标准。目前，这些系统的格式各异，而在使用通用的 Apache Arrow 格式后，就无需在这些系统之间将共享数据序列化/反序列化。Apache Arrow 旨在作为紧密集成其他大数据和数据分析系统的组件。Apache Arrow 还提供 I/O API，以将文件存储在磁盘上。目前，此 API 适用于 Apache Hadoop 生态系统中的 Apache Hadoop* 分布式文件系统 (HDFS*)。面向 Apache Arrow 的 DAOS 插件可在内存中将 Apache Arrow 格式转换为 DAOS 容器，从而可让更多应用适合高性能计算系统。

图 3. DAOS 中间件生态系统

DAOS 部署和路线图

DAOS 根据 Apache 2.0* 许可在 GitHub* (https://github.com/daos-stack/) 上提供。DAOS 管理指南提供有关如何安装、配置和管理 DAOS 安装的说明（详见 http://daos.io/doc）。

计划每六个月发布一个新的 DAOS 版本；更多信息，请查阅 DAOS 路线图 (http://daos.io/roadmap)。如果遇到问题，应通过 https://jira.hpdd.intel.com 报告。如有可能，请同时提供能反映该问题的相关信息。

社区邮件列表也可从 https://daos.groups.io 获取。

结论

英特尔® 傲腾™ 持久内存凭借其超低延迟和对于持久存储的高细粒度访问为推动行业转型、克服当下数据中心存储系统的局限性提供了大好时机。英特尔® 傲腾™ 固态盘进一步增强了此解决方案，在提供高 IOPS 和并发处理读写的同时，依然保持出色性能。但是，现有的分布式存储软件并不是针对这些新技术而构建的，因而可能会限制新技术发挥价值。因此需要重构软件存储堆栈以设计全新的解决方案，从而摆脱针对磁盘设计的无关优化、通过丰富的存储语义支持高细粒度数据和低延迟存储访问并挖掘这些革命性分布式存储技术的潜力。

[1] “Persistent Memory Programming”（持久内存编程），http://pmem.io/pmdk/。

[2] “Storage Performance Development Kit”（存储性能开发套件），spdk.io/。