大数据负载特性及基于内存技术的优化

刘根贤
(清华大学 计算机科学与技术系,北京 100084)

摘 要:涌现于社交网络、电子商务中的超大规模非结构化数据标志着大数据时代的到来。大数据的多样性、超大规模和可扩展性等特征对运行平台产生新的要求。随着大数据的产生和发展,形成了具有代表性的信息体系结构,包括编程模型、虚拟化和分布式文件系统等。随着对大数据研究的深入,通过对大数据负载特性的分析,发现制约大数据的并不是计算能力,而是 I/O 延迟,采用基于内存的分布式文件系统,用于存储和处理大规模分布式文件系统查询的索引,可以有效降低 I/O 延迟,提高应用性能。

关键词:大数据;负载特征;内存系统;系统结构
中图分类号:TP311.5 **文献标识码**:A **文章编号**:1674-7720(2014)02-0015-03

Workload character of big data and optimization based on memory technology

Liu Genxian
(Department of Computer Science & Technology, Tsinghua University, Beijing 100084, China)

Abstract: As the advent of social network and e-commerce, the amount of unstructured data grows rapidly. The 4Vs of big data (Volume, Velocity, Variety and Veracity) motivate the architecture design of new computing system, including programming model, virtualisation technology and distributed file system. According to the analysis on the big data workloads, I/O latency is one of the dominate performance bottleneck. Techniques that create and store index with memory-based distributed file system are proposed, which are able to significantly reduce I/O latency and thus improve system performance.

Key words: big data; workload characteristic; memory system; system architecture

随着电子技术的发展,计算成本降低,内存容量增加,大部分平台都可以用于高性能计算,可以处理比以往更多的数据信息,此外大规模集群技术的成熟,促成了多样性、超大规模和可扩展的多种典型应用,即大数据应用。大数据是一种数据分析技术,用于从超大规模的多源信息中快速获得有价值数据[1]。

在大数据应用的发展进程中,逐步形成了典型信息处理架构。硬件架构而言,其为大规模可扩展但不稳定的运算和存储基础设施;软件架构而言,其涵盖虚拟化、分布式数据库系统、数据挖掘和机器学习等。

随着大数据应用的扩展,对其研究的深入,提出了一些更有效的方法,这里基于大数据负载特性,分析以传统用于计算的内存作为存储介质,以加速制约大数据应用性能的延迟问题[2]。

1 大数据应用典型架构

以大数据应用为典型架构,从开发搜索引擎的需要出发,Google 提出了 Map/Reduce 架构,此后开源社区根据该思想实现了 Hadoop 系统,并得到广泛应用,使得大数据研究迅速成为热点。

大数据应用与高性能计算有典型区别,大数据应用中基础节点失效是正常的,运算向存储节点迁移,非结构化文件的操作被优化为追加操作。大数据典型结构如图 1 所示,在底层服务器节点之上是由存储文件系统元数据的主节点和存储文件实际数据的子节点组成的 GFS 或 HDFS 之类的分布式文件系统[3]。文件系统之上是分布式数据库系统,代表性的有 BigTable 和 Hbase。数据库节点同样也分为两类:Master 节点管理元数据并处理客户端元数据请求,Tablet 节点存储数据并处理数据请求。

开源分布式文件系统 HDFS 由大量普通计算机组成,任何时候任何节点都有可能出现故障,因此在 HDFS 的核心架构设计时,基础节点的出错检测和自动恢复是

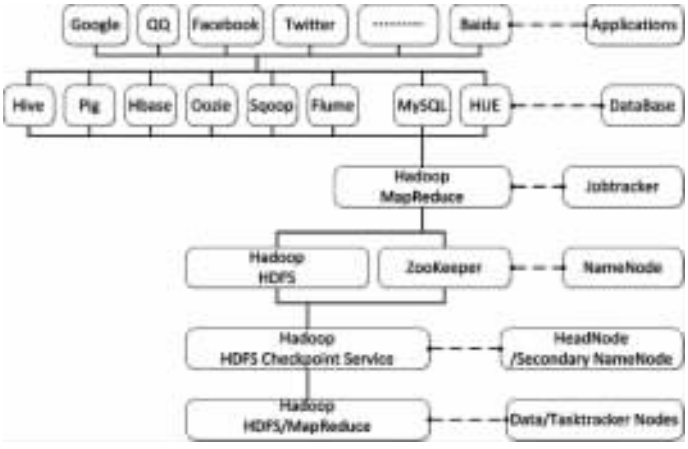

图 1 大数据典型应用

关键目标之一。HDFS 系统采用主/从架构，包括目录服务器节点和数据节点。系统中文件实际上划分为数据块冗余存储在多个数据节点里，数据节点在目录节点的控制下执行文件数据块的操作，存储管理本节点上的文件数据块。目录节点执行文件系统的目录空间操作，同时决定文件数据块到具体数据节点的映射，目录节点管理文件系统的目录和客户端对文件数据的访问。

HDFS 系统中分布式文件系统之上是数据库系统。利用数据库系统，用户可以按照类似数据库范式进行数据处理。大数据应用中为非结构化数据，即 NoSQL 数据库系统，它支持简单查询操作，而将复杂查询交给应用层处理（例如基于 Map/Reduce 框架实现大规模数据分析）。BigTable 和 Hbase 数据库是典型的主/从结构的键/值存储 NoSQL 数据库系统。

HDFS 系统结构中的顶层即为大数据应用，典型的应用例如搜索引擎和社交网络等。其中许多大数据应用将 Hadoop 系统当作数据存储设施，在此存储设施上进一步挖掘获取或分析其中的有价值信息[4]。

2 大数据负载特征

基于分布式文件系统以及其上的分布式数据库系统的大数据应用，随着应用规模的扩大，分布式系统中数据访问延迟将极大地影响应用的性能体验。Map/Reduce 架构通过将计算迁移到数据节点，同时使用冗余请求有效降低延迟。尽管如此，大数据应用的访问延迟仍然占到较大比例，其中最大负担是硬盘访问延迟，因此系统发展趋向于使用 SSD 硬盘替换磁介质硬盘，并将频繁随机访问的数据都放到节点各自内存中，而将顺序操作数据存储到硬盘。

网络延迟对分布式系统的影响也是一个关键问题。在高性能计算架构中采用 InfiniBand、Myrinet 和 Arista 等高性能互联技术可实现跨数据中心微秒级延迟通信，而大数据应用中服务器节点之间广泛采用的 TCP/IP 以太网的延迟达到数百微秒。因此网络延迟的优化是影响大数据应用性能的关键因素之一。

3 基于内存的延迟优化技术

内存是计算机系统中处理器之外访问延迟最小的部件，因此利用内存技术优化分布式系统数据访问延迟是一个关键策略。以搜索引擎应用为例，典型的搜索引擎已经完全将网页索引数据全部存储在内存中，有的系统甚至把所有网页快照都全部存储在分布式服务器内存中[1]。

利用内存的访问延迟优势，基于内存的数据库可以支持对数据进行实时处理。系统配置的内存越多，对数据的处理速度也就越快。计算机系统中不同部件的数据访问速度如图 2 所示。目前典型数据库系统都支持多核处理器平台，基于内存的数据库技术，利用更多的内存资源，克服目前分布式系统服务延迟性能瓶颈。其中应用较多的典型内存数据库有 Memcached 和 Redis 等。

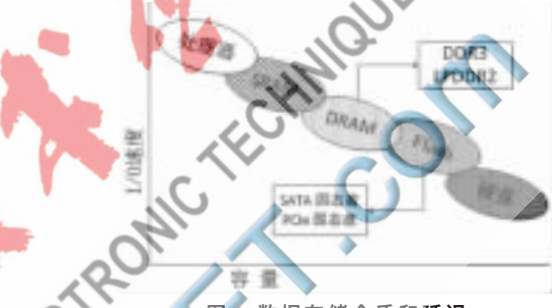

图 2 数据存储介质和延迟

Memcached 是一个结构简洁的高性能内存数据库，常用于网络应用以减轻数据访问负载。它通过在内存中缓存数据来减少系统访问次数，从而提高基于数据库网络应用的响应速度。未使用内存数据缓存时，大的数据记录在进行读写访问时需要较长时间响应，尤其是并发访问频繁时，严重影响系统性能。而 Memcached 通过使用简单的键值对存储访问数据，可以较好地提高应用性能。与 Memcached 相似的 Redis 是一个采用 hash 结构来做键值对存储的基于内存的 NoSQL 数据库[5]。

4 基于内存数据库元数据节点

大数据应用中元数据的查询管理操作直接影响整个系统性能，因此利用基于内存技术的 NoSQL 数据库管理元数据是一个较好的优化方法。

Memcached 用作分布式内存数据缓存服务器，将键值数据对存储于节点主内存中，其瓶颈在于需要处理存储的键值数据对与后台数据库服务器之间的一致性，需要刷新缓存值以更新数据库，因此需要对具体应用进行管理，增加了应用开发的复杂性。这类非关系型 NoSQL 数据库以键值对方式存储数据，每一个元组可以有不一样的字段，结构不固定，可以根据需要增加一些键值数据对，不局限于固定结构，可以有效提高系统性能，但其后台访问速度[6]依然是瓶颈。

利用内存技术，特别是分布式系统中大量节点的内存存储数据以优化访问速度，可以有效提高系统性能，

典型例子就是 RAMCloud 内存云技术。这是利用数据中心或集群系统的大量服务器的内存来存储所有应用数据的存储结构。传统保存在磁盘上的所有数据都可以保存在 RAMCloud 内存存储中。RAMCloud 可提供比磁盘存储低数百倍延迟和比磁盘存储高近千倍的吞吐量。利用内存的访存特点以及成熟的分布式系统技术，RAMCloud 具有优异的性能体验和良好的可扩展性，使之可以成为大数据应用中性能优化的关键技术。

RAMCloud 的原理是基于分布式节点的内存提供一个通用的存储系统，提供一个简单易用的存储模型，具有良好的扩展性。开发人员不需要采取特殊的方式对待 RAMCloud 数据存储，原有应用程序不需要做架构上的改变就可以迁移到 RAMCloud 平台。

基于分布式内存的 RAMCloud 的访问延迟可低至微秒级别。这比传统磁盘快近千倍，比基于半导体闪存器件的 SSD 要快数倍。RAMCloud 的低延迟特性对于对响应要求苛刻的网络应用和频繁访问数据为瓶颈的一些应用(例如高性能计算)来说极为重要。

4.1 RAMCloud 模型

基于内存技术的 RAMCloud 的低延迟和可扩展属性，便于大规模部署，消除了大数据应用所面临的性能和扩展性问题，可以处理比目前多数百倍的数据。RAMCloud 技术的可扩展性可以支持各个级别规模的应用，并可在小型应用扩展为大型应用时确保顺利进行，不涉及额外的存储结构。基于 RAMCloud 模型的应用系统框图如图 3 所示。

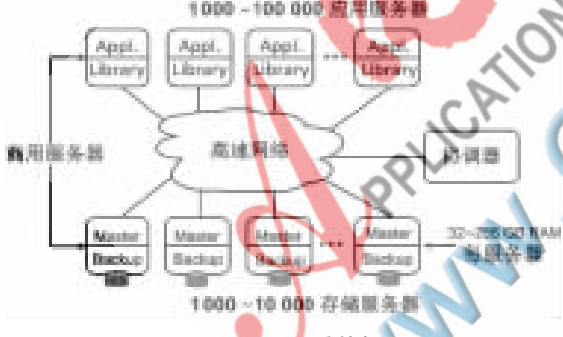

图 3 RAMCloud 系统框图

基于内存技术的 RAMCloud 代表存储服务的一种新存储模型。RAMCloud 与传统存储系统的区别在于，首先所有应用数据在任何时候都存储在构成 RAMCloud 存储系统的分布式内存中；其次 RAMCloud 必须建立在一定数量的服务器上，并实现节点的出错检测和自动恢复。与传统存储系统一样，存储在 RAMCloud 系统的数据就像存储在磁盘上那样是持久的。单一节点的存储出现故障后不会造成数据丢失或数据不可用的状况。

与典型大数据应用结构一致，基于内存技术的分布式内存存储系统也可以支持将计算迁移到数据节点，同时使用冗余请求进一步降低延迟。利用分布式内存存储技术，在应用服务器上运行的进程访问数据的延迟有可能降低到微秒级别，而基于传统磁介质存储系统通常为近毫秒级别。

4.2 分布式内存存储分析

基于分布式内存技术的大数据应用新架构，传统的架构是应用程序的代码和本地局部数据被加载到计算机主存储中，需要时访问本地或远程存储节点。图 4 显示各种存储方式下数据的访问延迟。传统应用的性能瓶颈是显而易见的，不同数据的频繁访问操作、应用程序的并发访问、规模大小都可能造成系统性能瓶颈。

图 4 内存云存储技术及延迟

基于分布式内存存储技术代替传统的存储系统，采用基于轻量低功耗处理器的微服务器，将在线应用数据的主要存储中心从传统存储迁移到分布式内存上，利用成熟的集群技术，构建可扩展的基于内存的存储系统，利用 Map/Reduce 框架实现大数据应用。

基于内存技术的 RAMCloud 架构原理在于将所有应用的数据信息存储在分布式内存上，并使用大量服务器构建可扩展的大型存储系统。利用内存的访存延迟极低的特性，存储在内存上的数据的延迟要比存储在基于传统存储系统上低近千倍，而吞吐量则会高数百倍。

大数据应用主要延迟来自数据访问延迟，对处理器计算能力的需求远低于处理器所能提供的性能。采用基于轻量低功耗处理器的微服务器，将应用数据从传统存储迁移到分布式内存上。内存存储充分结合了内存的低延迟和集群的规模化优势，保持应用可扩展性的同时降低了数据访问延迟。这种基于分布式内存存储的大数据可以同时实现大规模和低延迟的优势，有效加速大数据应用。

参考文献

[1] KAI H, GEOFFREY C F, JACK J D. Distributed and cloud computing: from parallel processing to the internet of things[M]. Massachusetts: Morgan Kaufmann Publishers, 2012.

[2] Jia Zhen, Wang Lei, Zhan Jianfeng, et al. Characterizing data analysis workloads in data centers[C]. In Workload Characterization (IISWC), 2013 IEEE International Symposium on. IEEE. 2013.

[3] 吴朱华. 云计算核心技术剖析[M]. 北京：人民邮电出版

社,2011.
[4] 王鹏.云计算的关键技术与应用实例[M]. 北京:人民邮电出版社,2010.
[5] 曾超宇,李金香.Redis 在高速缓存系统中的应用[J].微型机与应用, 2013,32(12):11-13.
[6] 张青凤, 张凤琴, 王磊.多数据中心的数据同步模型研究与设计[J]. 微型机与应用, 2013,32(12):60-62.

(收稿日期:2013-10-10)

作者简介:
刘根贤,男,1974 年生,博士研究生,主要研究方向:计算机体系结构 CPU 设计。