哈爾濱工業大學

毕业设计（论文）开题报告

题 目：基于 **RDMA** 与 **NVM** 的低延迟分布式存储系统

专　　业　计算机科学与技术

学　　生　张婉茹

学　　号　**1170500116**

指导教师　王宏志

日　　期　**2020** 年 **12** 月 **02** 日

哈尔滨工业大学教务处制

1．课题来源及研究的目的和意义

随着移动互联网的快速发展，人们的生活与数据有着密切的联系，每个人每天都产生着大量的数据，使得数据量呈现爆发式的增长，海量数据的存储和处理能力也面临前所未有的压力。传统计算机体系架构中以“计算为中心”的模式,难以应对数据规模与数据处理能力之间日益突出的矛盾,现存的计算机体系架构面临着严峻的挑战和生存瓶颈。

近年来,随着新型非易失性内存介质技术的兴起,打破传统的体系架构,构建大容量,非易失,高可靠的内存系统,将数据大量或整体存放于内存中,形成以内存为主体的新型内存计算模式,从而完成计算模式从以“计算为中心”向以“数据为中心”的转变,成为学术界和工业界关注的热点。

非易失性内存具有持久性、字节寻址、密度高和低能耗等优点，这对于内存计算、内存存储和数据密集型的场景都具有重要的意义。然而，传统的 I/O 软件系统结构并不适用于非易失性内存，如将其直接部署在非易失性内存之上并不能充分发挥非易失性内存自身的优势。比如在硬件方面，基于磁盘的存储系统中，主要设计思路是通过减少磁盘磁头的移动时间来提升系统的性能，而非易失性内存具有字节寻址的特性，随机读写的性能强；在物理特性方面，非易失性内存的耐久性较弱，传统的存储系统软件针对磁盘设计，磁盘并不存在耐久度的问题，所以现有的存储系统软件需要重新考虑耐久性的问题，减少非易失性内存的写入磨损；在软件方面，传统的存储系统软件会针对磁盘而优化，比面向分布式非易失性内存的新型存储系统的设计与实现 如，顺序读写，追加写，而非易失性内存随机读写和顺序读写性能相差不多，但存在读写不对称，读性能远高于写性能。以上问题均表明现有的存储系统不适合于非易失性内存，针对磁盘的 I/O 栈优化在非易失性内存存储系统上需要重新考虑。

所以我想做基于非易失性内存的分布式存储系统，通过将数据存储在可字节寻址的非易失性内存中来提供低延迟存储服务。同时，由于非易失性内存读写速度快的特点，系统的延迟、性能瓶颈将转移到网络 IO。远程直接内存访问（RDMA）相对于传统以太网具有高吞吐和低延迟的特性，因此，利用 RDMA 直接读写 NVM 中的数据。该技术旁路了系统网络协议栈，降低延迟的同时，也减少了数据的拷贝与 CPU 时间。而两者结的过程中会产生新的问题，比如 RDMA 访问服务器端内存的一致性问题，本地非易失性内存管理低效的问题等，都有待于优化和解决。

2．国内外在该方向的研究现状及分析

分布式存储系统是大规模网络服务的关键基础设施。通过提供统一的应用编程接口（Application Programming Interface，API），能够屏蔽数据分布在成千上万服务器节点的事实，使得客户端能够像操作单机数据库般操作分布式存储系统中的数据。传统的分布式数据库随着数据规模的增长，愈来愈遇到延迟、性能的瓶颈。而新的非易失性内存和远程直接内存存取技术的出现，将助力分布式存储系统克服这些瓶颈。本节将从下面两个方面来

介绍分布式存储、非易失性内存和远程直接内存存取技术今年的研究现状：非易失性内存存储系统和 RDMA 技术的研究与应用。目前，国内外鲜有将非易失性内存应用到分布式存储系统的研究。

2.1 非易失性内存存储系统

由于接近随机存取内存的读写速度和可字节寻址特性，非易失性内存可被直接连接到计算机系统的 PCI 总线，作为第一级主存储设备。近年来涌现了许多将非易失性内存用于持久化对象、数据存储的研究。其中一些为针对非易失性内存的内存文件系统。以 PMFS 为代表的非易失性内存文件系统采用日志结构来存储数据，来达到非易失性内存的读写均衡。而 BPFS 则通过在非易失性内存系统中运用一种称为 short-circuit shadow paging 的技术来提供原子的、细粒度的更新和持久化存储。NOVA 则结合了随机存取内存和非易失性内存的优点，设计了应用于混合内存架构的文件系统。

另一方面，许多研究致力于应用非易失性内存构建结构化持久对象存储。如 Shivaram Venkataraman 的研究提出了一致性的持久数据结构（Consistent and Durable Data Structures，CDDS）。其通过 versioning 技术在无需写日志的前提下向用户提供一致性的持久数据存储。而 Amirsaman Memaripour 则研究了非易失性内存上的事务处理。

其通过维持数据的额外副本来达到事务级的原子更新操作。综合来看，非易失性内存存储系统的研究均归结于如何充分发挥非易失性内存持久化和可字节寻址的特性，向应用层提供兼容的，更高性能的持久化存储服务。

2.2 RDMA 技术的研究与应用

直接内存存取（RDMA）技术已被广泛应用于高性能计算机中。通过 RDMA 技术，应用程序能够直接读写远程计算机的内存数据。RDMA 技术旁路了系统网络协议栈，且 RDMA 读写无需对端 CPU 参与，这大大降低了应用程序读写远程计算机内存的延迟。

利用 RDMA 技术构建高性能的分布式存储系统也得到了广泛的研究。例如 Christopher Mitchell 发表的 Pilaf，通过单边 RDMA 读来构建快速、高效的分布式键值存储系统。而 Anuj Kalia 的 HERD 则通过结合 RDMA 写和 RDMA 消息 verb 达到了更高的性能。

不过 Pilaf 和 HERD 系统均面临着可扩展性差的问题。因为，两个系统都应用了面向连接的 RDMA 读写 verb，这使得 RDMA 性能在大规模集群中下降。为了同时解决可扩展性问题，FaSST 应用基于报文的 RDMA 消息 verb 构建了高性能的分布式事务。而 FaRM 则通过 RDMA 技术将多个服务器节点的内存构建为统一的共享内存空间。应用可以无差别得读写本地和远程的内存地址空间。得益于 RDMA 的高带宽低延迟特性，FaRM 提供了十倍于基于 TCP/IP 协议的同类系统。

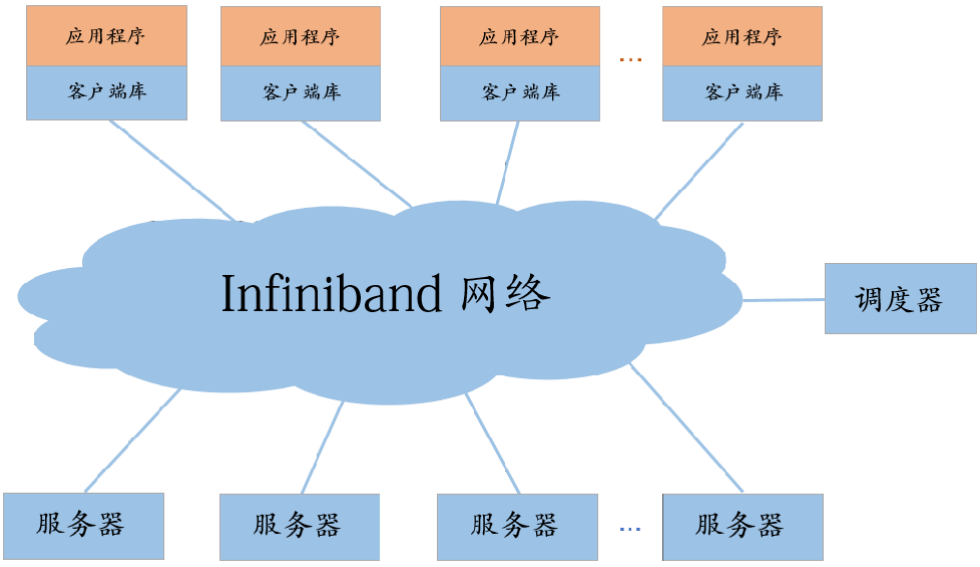
3. 主要研究内容

基于 RDMA 和 NVM 分布式存储系统的设计：该设计以降低请求延迟为主要目标。同时，充分发挥非易失性内存的持久化、可字节寻址等特性，降低系统成本和资源开销。目前还没想好是做键值存储系统还是文件存储系统。

同时针对一致性问题，理解前人设计的算法（Raft,Paxos 等），并理解这些算法在 RDMA 和 NVM 上会遇到的问题，理解有文章修改了协议流程中对一致性成功的判定方式以及领导者选举和日志恢复中的部分流程，利用 RDMA Write 操作的绕过目标服务器 CPU 的特性、NVM 的持久性和可字节寻址的特性，使得修改后的协议具有低时延、高吞吐量和低 CPU 消耗等特点。并尝试提出解决更好的方案。

4. 研究方案

如图 3－1 所示为系统整体架构。集群中的服务器节点，调度器，客户端均通过支持 RDMA 的 Infiniband 网络连接。对于一些实时性和带宽要求不高的请求，我们使用 TCP/IP 协议进行通信，如服务器节点向调度器请求加入，以及调度器向服务器节点广播集群配置和索引信息等。



- 调度器

调度器在所设计的分布式系统中扮演着至关重要的角色。为了防止调度器成为系统的性能瓶颈，调度器主要负责服务器、客户端连接管理等工作。而负载较大的数据查询、更新等请求，有客户端直接发往对应的服务器节点。调度器管理服务器节点的加入和删除。

- 数据模型

有可能是用 KV 存储也有可能是文件系统。

- 空间管理

  有可能是用基于日志的内存空间管理方法，也有可能是基于空闲块的内存分配器进行空间管理。下面是基于空闲块的内存分配器进行空间管理的一些想法。

  由于 NVM 与 DRAM 巨大的特性差异,当前主流的内存分配器都无法胜任 NVM 内存空间的分配。NVM 内存分配器必须支持 malloc_at(addr, size)分配操作,即既能指定分配的大小,又能指定分配的地址,以使得用户能够根据之前写入数据的地址重新访问到持久化在 NVM 上的数据,并将该数据由 NVM 内存分配器保护起来,不再被分配出去。设计动态内存分配器，设计非易失性内存分配的分配释放的算法，更好的适应于数据模型，也使得 malloc()、malloc_at()和 free()操作都有良好的性能。

- 远程过程调用

  通过对 RDMA 原语的应用，设计高效、紧凑的远程过程调用的方法。为了设计实现出低延迟的存储系统，我们在服务器节点中分配了专门的线程用于轮询客户端的消息。

- 一致性问题

  为了防止单点故障，我们需要把数据进行复制和备份，而为了追求低延迟，可能只有一个机器的非易失性内存上存好了数据，我们立即返回应答给用户的请求，而此时如果断电，数据还没来得及备份到其他机器上，就会出现不一致的问题，通过 log 和其他算法来更好解决一致性问题。

5．进度安排，预期达到的目标

11 月-12 月末：阅读各类相关论文，对前人对 RDMA 和 NVM 的研究有一个全面的认知。同时争取把 6.824 课程的学完，对分布式系统可靠性，一致性，容错性等问题有更深的了解。

1 月初-2 月末：在集群上着手搭建自己的系统，完成各个模块。

3 月初-4 月末：对一致性问题进行深入的研究和思考。

5 月初-6 月：和前人搭建的系统作对比实验，撰写毕业论文。

6．课题已具备和所需的条件、经费

实验室已具备条件：分布式的集群，集群可以通过网卡通信，达到一定带宽，集群机器的内存是 DDR4，由于还没有可以量产的可字节寻址非易失性内存，因此我们使用 DRAM 来模拟非易失性内存，以进行存储实验。如 PMFS 所提到的，由于非易失性内存的速度与 DRAM 接近，因而在一定程度上是合理的。

7．研究过程中可能遇到的困难和问题，解决的措施

整个系统需要考虑的问题都比较多和繁杂，所以只能是多学多看来解决问题。

8．主要参考文献

[1]董康平. 基于非易失性内存和 RDMA 的低延迟分布式键值存储系统的设计与实现[D].上海交通大学,2018.

[2] 刘志祥. 基于 RDMA 的非易失性内存文件系统设计与实现[D].重庆大学,2018.

[3]陈波. 面向分布式非易失性内存的新型存储系统的设计与实现[D].江苏大学,2019.

[4]刘昊. 面向非易失性内存的系统软件若干问题的研究[D].上海交通大学,2018.

[5]陈游旻,陆游游,罗圣美,舒继武.基于 RDMA 的分布式存储系统研究综述[J].计算机研究与发展,2019,56(02):227-239.

[6] 周坚石. 基于非易失性存储器（NVM）的内存分配器的设计与实现[D].南京大学,2018.

[7] Youyou Lu, Jiwu Shu, Youmin Chen, and Tao Li. 2017. Octopus: an RDMA-enabled Distributed Persistent Memory File System. In 2017 USENIX Annual Technical Conference (ATC '17). Santa Clara, CA, USA.

[8] WU H, CHEN K, WU Y W, ZHENG W M.Research on the consensus of big data systems based on RDMA and NVM. Big Data Research[J], 2019, 5(4):89-99

[9]GlusterFS on RDMA."https://gluster.readthedocs.io/en/latest/AdministratorGuide/RDMATransport/".

[10]Accelio."http://www.accelio.org", 2013.