基于高性能计算机内存构建分布式文件系统的网络性能优化方法

武春佳　刘光明　刘欣

(国防科学技术大学计算机学院，长沙 410073)

(通信作者邮箱：wucj@nscc-tj.gov.cn)

摘　要

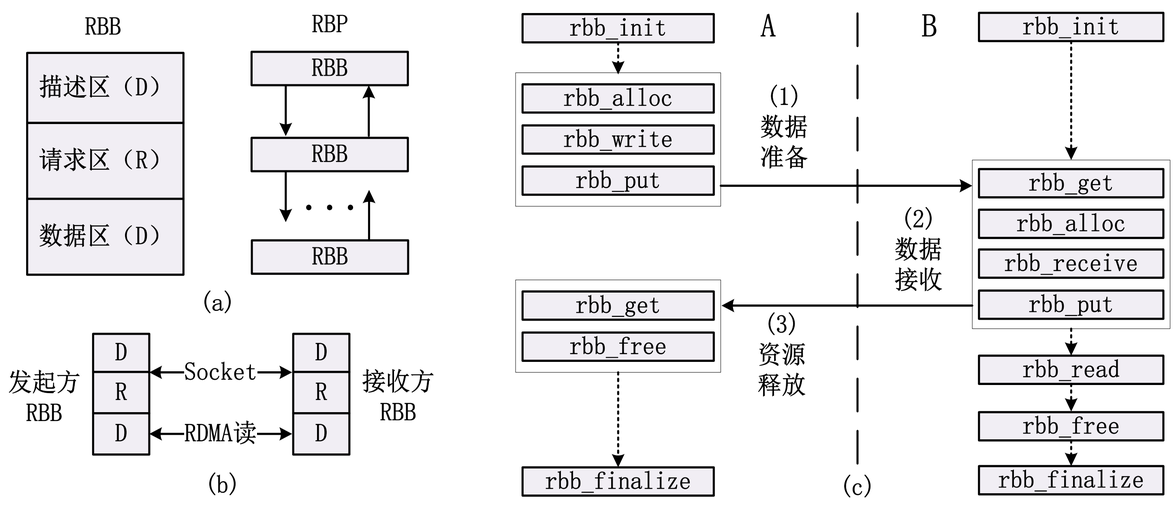
高性能计算机在发展过程中，计算节点的内存容量在不断提升。与此同时，应用对节点内存的不均衡需求，又导致不同节点的内存利用率往往差异较大。于是，为充分利用系统的整体内存资源，产生了基于计算节点空闲内存，构建分布式文件系统的需求。当传统分布式文件系统的底层存储介质从磁盘变为内存，原有的基于Socket的通信方式已无法充分发挥内存级存储带来的性能优势，从而导致网络通信性能成为制约系统整体性能的主要瓶颈。鉴于RDMA通信在带宽和延迟方面的良好特性，同时结合当前大多数高性能计算机支持RDMA通信的固有优势，本文提出利用RDMA通信改进传统分布式文件系统网络性能的解决思路。

在很多分布式文件系统与RDMA通信相结合的研究中，都首先对分布式系统的结构组成、IO特点、通信模式等进行分析，以此为依据设计引入RDMA通信的环节和方式。此外，引入RDMA通信所增加的控制复杂度和额外开销也是不能忽视的重要因素。本文的研究中，选取MooseFS作为系统框架，通过对MooseFS的读/写流程进行分析得知，在处理读/写请求的过程中，有2个环节涉及到实际的数据操作：一是数据服务器对本地磁盘进行IO操作；二是客户端程序与数据服务器之间通过Socket传输数据。基于磁盘文件和基于内存块的对比实验表明，基于内存块的存储形式可以使数据服务器的写性能提高数倍，但对系统整体写性能的提升却非常有限。因而本文研究重点放在如何利用RDMA通信改善IO请求数据的传输上。为此，提出一种基于RDMA的高速缓冲池RBP（RDMA Buffer Pool）机制。将基于RBP改进前后的MooseFS系统先后部署在TH-1A系统上，使用应用广泛的IO性能测试工具IOR，在相同的部署环境和系统配置下进行了对比测试。测试结果表明，改进后系统客户端的顺序读写速度和服务端在顺序读写时的聚合带宽均有明显提高，单客户端顺序读、写速度最大可达到原系统的2.0、2.6倍，单服务端在顺序读、写时的聚合带宽最大可达到原系统的2.0、2.4倍。

关键设计

RBP机制的工作原理是通过预先注册一块或多块支持RDMA操作的内存区，按照系统需求将这片区域划分成多个尺寸规格的缓冲块RBB（RDMA Buffer Block）。再根据不同用途，将同样尺寸的RBB组织成不同的缓冲池RBP，配合一套专用的API，以RBB为基本单位提供高性能的数据传输服务。RBB由描述区、请求区、数据区3个部分组成：描述区负责提供RDMA通信内存区的描述信息，请求区负责提供控制传输和请求信息，数据区负责提供实际存储空间。RBP是组织和管理RBB的主体，其主要功能是维护一个由RBB作为元素的双向链表

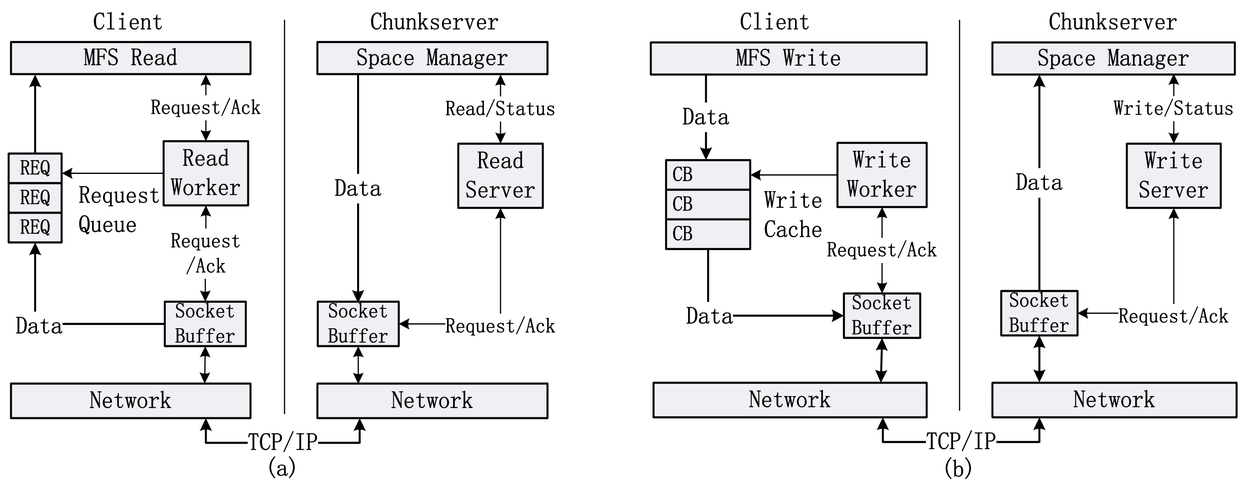
RBP机制基于TH-1A的通信软件GLEX实现RDMA通信。由于GLEX的RDMA单边写仅保证将数据成功写入到网络，需要引入远程事件机制才能检测数据的写入状态，会增加控制消息的复杂度。而RDMA单边读通过本地事件机制即可检测读取状态，因此RBP的RDMA通信操作仅支持单边读，根据数据流的方向确定操作发起方。在利用RBB进行RDMA通信时，RBB需要在通信两端成对使用。一般流程是由操作发起方通过Socket获取远端RBB的描述信息，再根据数据请求信息读取远端RBB数据区内的数据，完成后通过Socket返回操作状态。



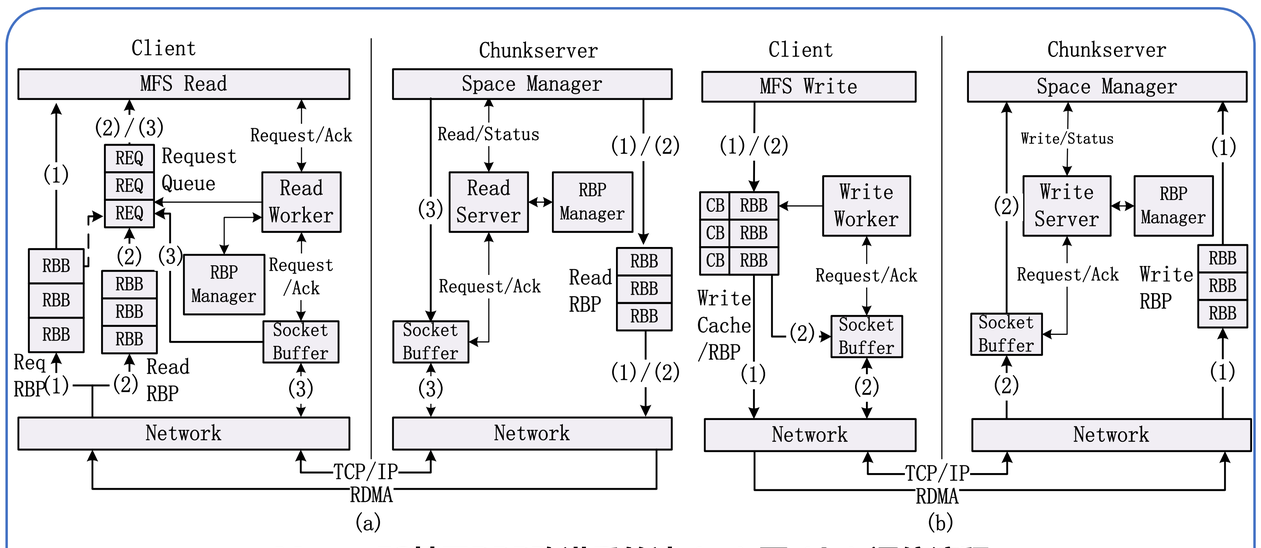
RBP组织结构（a）（b）与一般工作流程（c）

优化方法

为支持更多的应用场景，RBP提供显式和隐式两种使用方式，显式使用是指使用者在RBP创建后就分配得到全部的RBB，此后由使用者自行管理，适用于用途明确且管理简单的情形；隐式使用是指使用者在需要时从RBP分配RBB，使用后再将RBB释放，由专门的RBP管理模块进行管理，RBB的分配与释放对使用者是透明的，适用于用作临时用途的情形。由于RBP并不改变系统原有读/写控制流程，只是在文件数据传输时代替Socket，因此可以非常灵活地配置和使用。对读的优化包括在客户端程序增加少量读大块数据的专用RBP和读一般数据的临时RBP，在数据服务器增加配合远程读数据的临时RBP、引入连续读流水线和保留原Socket数据通道用于小块数据传输的多通道设计等。对写的优化包括将客户端程序写缓存与专用RBP进行合并，在数据服务器增加配合远程写数据的临时RBP和保留Socket数据传输通道用于小块数据传输的多通道设计等。



MooseFS基于Socket的读（a）写（b）通信流程

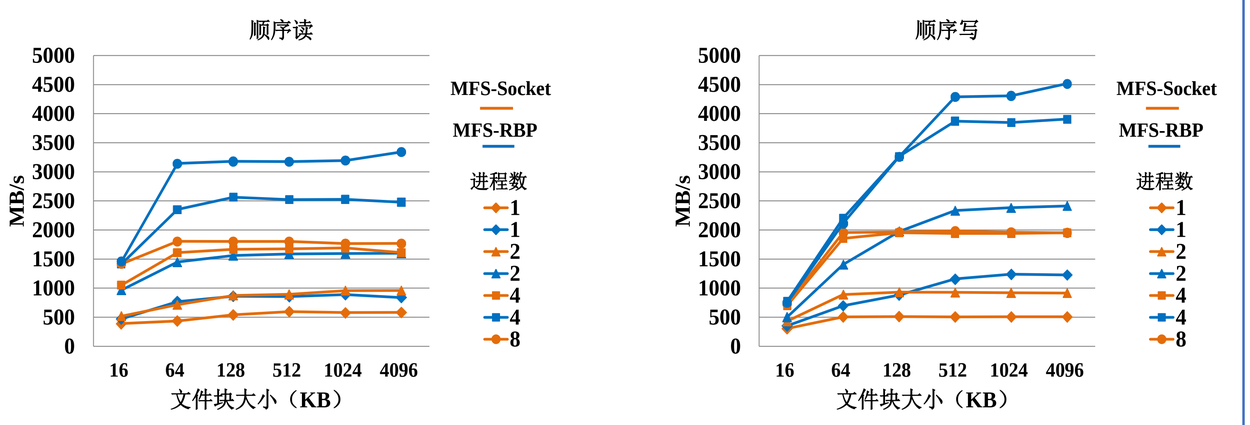


MooseFS基于RBP改进后的读（a）写（b）通信流程

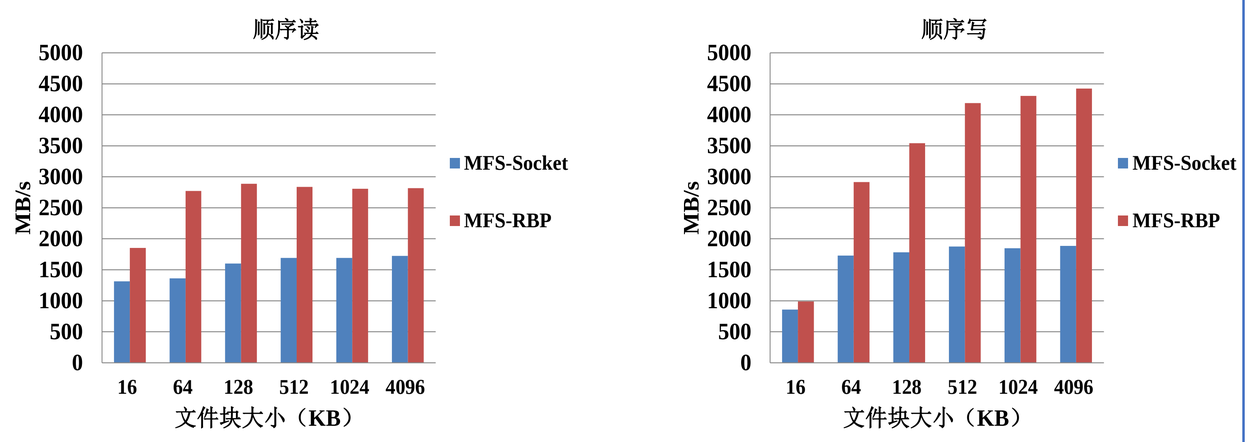
性能评测

客户端对比测试在1个Client下进行，分别采用1、2、4、8个进程进行并行读写，以测试单个客户端的整体性能。，在相同文件块大小和相同进程数时，改进后系统的顺序读写速度全面优于原系统。读速度最大可达到原系统的2.02倍；写速度最大可达到原系统的2.63倍。

服务端对比测试在1个Chunkserver下进行，采用4个Client，每个Client采用单进程进行并发读写，以测试单个服务端在顺序读写时提供的聚合带宽，测试结果如图6所示。改进后系统的单个服务端在顺序读时，向4个Client提供的带宽最大可达到原系统的2.04倍；顺序写时的带宽最大可以达到原系统的2.35倍。而且顺序写时的带宽最大值为4.42GB/s，占到计算节点之间RDMA通信最大单向带宽的接近70%。



系统改进前后单客户端顺序读（a）写（b）性能对比



系统改进前后单服务端顺序读（a）写（b）带宽对比

总　结

基于高性能计算机构建分布式内存文件系统时，传统基于Socket的通信方式无法充分发挥内存的性能优势。为此，本文提出一种基于RDMA的高速缓冲池RBP。RBP以分布式文件系统原有的控制流程为基础，极大地降低了系统改造的难度。通过采用多种切实有效的设计将RBP引入MooseFS这一典型架构的分布式文件系统中。测试结果表明，改进后系统的网络性能在整体上得到大幅提升，但对小数据和多进程的支持还存在很大的改进空间。下一步考虑结合数据预取、写合并、最小匹配等技术，使RBP具有更全面的性能表现和更广泛的应用前景。

参考文献

[1] Ghemawat S. The Google file system[J]. Acm Sigops Operating Systems Review, 2003, 37(5):29-43

[2] Yang Xuejun, Liao Xiangke, Lu Kai, et al. The TianHe-1A Supercomputer: Its Hardware and Software[J]. Journal of Computer Science & Technology, 2011, 26(3):344-351

[3] Xie Min, Lu Yutong, Liu Lu, et al. Implementation and Evaluation of Network Interface and Message Passing Services for TianHe-1A Supercomputer[C] //Proc of the 2011 IEEE Symposisum on High Performance Interconnects. Piscataway, NJ: IEEE, 2011:78-86

[4] Islam N S, Rahman M W, Jose J, et al. High performance RDMA-based design of HDFS over InfiniBand[C] //Proc of the 2012 International Conference for High Performance Computing, Networking, Storage and Analysis. Piscataway, NJ: IEEE Computer Society, 2012:1-12

[5] Mitchell C, Geng Y, Li J. Using one-sided RDMA reads to build a fast, CPU-efficient key-value store[C] //Proc of USENIX ATC '13. Berkeley, CA: USENIX, 2013:103-114

[6] Xiong Wen, Yu Zhibin, Xu Chengzhong: A Characterization and Analysis of Distributed File Systems[J]. Journal of Integration Technology, 2012, 1(4): 58-62