CS 5204 Project Final Report
A Feasible Study of MPI-IO on Top of HDFS

Luna Xu (xuluna@cs.vt.edu) Adam Binford (adamq@vt.edu)

December 1, 2014

1 Team member

Luna Xu (xuluna)
Adam Binford (adamq)

2 Introduction

MapReduce [2] and its most popular implementation Hadoop [1] have become the dominant distributed processing framework for big data analytics. Despite of the ease-of-use and scalibility of Hadoop, researchers also found the limitations of Hadoop lie in for example, inter-process communication. For such limitations, the well established Message Passing Interface (MPI) [14] is more suitable due to its ability to support any communication pattern. X. Lu et al [19] find that the message latency of MPI is about 100 times less than Hadoop primitives. The average peak bandwidth of MPI is about 100 times higher than Hadoop RPC – the fundamental communication mechanism in Hadoop. Moreover, there exists data analytics workflows such as Metagenomics [18] that consist both compute- and communication-intensive computations. To better conduct such workflows and to avoid data movement between clusters [20], resource coordination platforms such as Mesos [16], Omega [26], YARN [27] enable different programming paradigms including MPI and MapReduce to co-exist in the same cluster. Though not realized yet, hosting MPI and Hadoop in the same cluster is highly promising as YARN claims to embrace MPI as a first class citizen.

One of the biggest challenges of co-hosting MPI and Hadoop is to decide the underlying shared file system. As a big feature of MPI-2 standard [14], MPI-IO provides parallel IO support to MPI programs and enables MPI to process data-intensive workloads as well. Currently, such support requires an underlying network/parallel file system such as NFS [9], PVFS [22], Lustre [25], GPFS [24] to achieve the best performance. However, these file systems are focused on optimization for MPI-IO [13, 17, 21] and have a big network overhead on hosting Hadoop with the absence of data locality [23]. IBM's GPFS is originally designed as a SAN file system as the data is striped and placed in a round-robin fashion [24], which prevents it from being used in Hadoop. With a support of File Placement Optimization (FPO), GPFS-FPO makes it possible to efficiently support Hadoop. However, GPFS is shipped with IBM SP system and is not available as opensource as Lustre. Moreover, IBM tailors MPI-IO according to GPFS in their own MPI implementation [21], which is not supported in more wildly used implementations such as MPICH [7] and OpenMPI [10].

Another solution is to support MPI-IO on top of the distributed file systems used by MapReduce such as GFS [15] and HDFS [11]. HDFS is integrated inherently in Hadoop releases and is the default file system used in Hadoop community. By bringing MPI to HDFS, it is possible to keep existing applications in Hadoop ecosystem without any changes. C. Cranor et al [12] explore the performance of MPI-IO on HDFS using PLFS. However, HDFS is supported as a component of PLFS and no data locality is achieved for MPI jobs. As far as we know, there is no such work on supporting MPI on top of HDFS directly. This project focuses on exploring the feasibility and performance of enabling MPI-IO on top of HDFS using existing technologies. This study is based on the observation that MPI-IO provides great flexibility that it is possible for users to decide the process-to-block mapping. In this work, we explore and evaluate the existing methods including FuseDFS [6], Native Library [5], HDFS-NFS-Proxy [4]. We also develop our own MPI-IO hook using the native library provided by HDFS. In order to evaluate the performance, we also developed a MPI-IO benchmark. This report reflects the difficulties we encountered, whether each method can

be adopted by MPI-IO, and how they performs.

During our study and development, we encounter several challenges including ones that are still not solved yet.

- Existing methods such as FuseDFS and HDFS-NFS-Proxy are originally designed to ease the file maintenance instead of parallel file accessing, thus only sequential write operations are supported. Often failure is the case when we try to perform parallel writes (Section **??**).
- MPI-IO accesses underlying file systems using standard POSIX I/O system calls, while HDFS is not designed to be a mountable POSIX file system and must be accessed through its own API. The native library provided by HDFS gives a C/C++ compatible library that can be adopted for MPI files. However, MPI-IO does not have the corresponding support to access through the library. We develop a hook which enables MPI-IO to access HDFS files using the native library transparently so that the only thing the user needs to provide is the HFDS URL to locate the file (Section **??**).
- During our development, we found that HDFS does not support what writes? parallel writes? please describe here.
- HDFS is an append-only file system, while sometimes MPI jobs want to modify the data in the file. To achieve that, one must rewrite the entire file and replace the old file. This can cause a huge overhead on small file updates. However, this problem is not addressed in this report because our study and benchmark focus on parallel reads and writes throughput without consideration of file updates.

3 Design and Implementation

4 Experiment

5 Conclusion

6 Project Progress

We have finished investigating the possible ways to mount HDFS as a regular file system that can be interacted with by any file I/O. We got the throughput for manual copy, native NFS as the ideal performance, fuse-dfs throughput for parallel read. However we could not perform parallel write using fuse-dfs. Table 1 shows the errors we encountered during our tries. We tried using `MPI_File_write_at` where each process holds an individual file pointer, as well as `MPI_File_write_shared` where all processes hold a shared file pointer. We open the file using different mode and with the combinations we get mainly two errors. The error we get from the `APPEND` mode is reported in the MPI program side, others are shown in the fuse-dfs side. Another method that we explored is Native HDFS Fuse [8], which utilizes only protobuf to communicate with Namenode directly. Hence no fuse or native lib is involved. However, the program dumped a segmentation fault when we tried to run. HDFS-NFS solution is also not successful nor desirable because either it has requirements for specific (2.3.0) Hadoop version [3] or it only supports the cloudera distribution of Hadoop [4].

Function	Mode	Error
MPI_File_write_at	CREATE|RDWR	cannot open an hdfs file in O_RDWR mode
MPI_File_write_at	CREATE|WRONLY	cannot open an hdfs file in O_RDWR mode
MPI_File_write_at	WRONLY	cannot open an hdfs file in O_RDWR mode
MPI_File_write_shared	WRONLY	cannot open an hdfs file in O_RDWR mode
MPI_File_write_shared	APPEND	file open. code: 201388309

Table 1: Error codes for parallel writes on fuse-dfs.

We are now focusing on creating a library to hook MPI [7] I/O function calls to use the HDFS native library to interact with HDFS. Our goal is to allow unmodified MPI applications to interact with HDFS by simply loading our library at runtime. So far we have successfully hooked MPI functions at runtime, and verified our functions were being called. Additionally, we have read from and written to files in our running HDFS using the HDFS native library. When reading a single file from multiple processes, we have observed an increase in bandwidth when increasing processes. This confirms that multiple processes can read from the same file at once using the native library. To complement this work, we have developed scripts to compile and run these HDFS native library applications easily.

7 Future work

The final steps we have to do are implementing the necessary MPI I/O functions in our hooking library to use the HDFS native library as the file I/O method. We have already done each of these pieces individually, hooking and using the native library, we simply must combine them. The hooking functions need to be able to use the parameters they are given to seamlessly work with HDFS without the MPI program knowing anything is different.

Additionally, we must find out if it is possible to implement some support for writing to HDFS through the hooked MPI routines. The native library only allows appending to a file, and only one thread can access a file for writing at one time. We must either modify the behavior of the I/O of the MPI program or set restrictions on what MPI programs running on HDFS are allowed to do. Finally, we must simplify the scripts required for our solution to work to put as small of a burden on the user as possible.

References

[1] Apache. Hadoop. http://hadoop.apache.org/.

[2] Hadoop MapReduce. http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html.

[3] Hdfs nfs gateway. http://hadoop.apache.org/docs/r2.3.0/hadoop-project-dist/hadoop-hdfs/HdfsNfsGateway.html.

[4] hdfs-nfs-proxy. https://github.com/cloudera/hdfs-nfs-proxy.

[5] Libhdfs. http://wiki.apache.org/hadoop/LibHDFS.

[6] Mountablehdfs. https://wiki.apache.org/hadoop/MountableHDFS.

[7] Mpich. www.mpich.org.

[8] Native hdfs fuse. https://github.com/remis-thoughts/native-hdfs-fuse.

[9] Network File System. http://en.wikipedia.org/wiki/Network_File_System.

[10] Open mpi. http://www.open-mpi.org/. Accessed: 2014-08-11.

[11] Hadoop Distributed File System (HDFS). http://hortonworks.com/hadoop/hdfs/, 2014.

[12] Chuck Cranor, Milo Polte, G. G. Hpc computation on hadoop storage with plfs. Tech. Rep. CMU-PDL-12-115, Parallel Data Laboratory, Carnegie Mellon University, Pittsburgh, PA, November 2012.

[13] Dickens, P., and Logan, J. Towards a high performance implementation of mpi-io on the lustre file system. In *On the Move to Meaningful Internet Systems: OTM 2008*. Springer, 2008, pp. 870–885.

[14] Forum, M. P. I. Mpi: A message-passing interface standard. Tech. rep., Message Passing Interface Forum, September 2012.

[15] Ghemawat, S., Gobioff, H., and Leung, S.-T. The google file system. In *Proc. of the 19th ACM SOSP* (2003).

[16] Hindman, B., Konwinski, A., Zaharia, M., Ghodsi, A., Joseph, A. D., Katz, R. H., Shenker, S., and Stoica, I. Mesos: A platform for fine-grained resource sharing in the data center. In *NSDI* (2011), vol. 11, pp. 22–22.

[17] Ilroy, J., Randriamaro, C., and Utard, G. Improving mpi-i/o performance on pvfs. In *Euro-Par 2001 Parallel Processing*. Springer, 2001, pp. 911–915.

[18] Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews 72*, 4 (2008), 557–578.

[19] Lu, X., Wang, B., Zha, L., and Xu, Z. Can mpi benefit hadoop and mapreduce applications? In *Parallel Processing Workshops (ICPPW), 2011 40th International Conference on* (Sept 2011), pp. 371–379.

[20] Monti, H. M., Butt, A. R., and Vazhkudai, S. S. Catch: A cloud-based adaptive data transfer service for hpc. In *Proc. IPDPS* (2011).

[21] Prost, J.-P., Treumann, R., Hedges, R., Jia, B., and Koniges, A. Mpi-io/gpfs, an optimized implementation of mpi-io on top of gpfs. In *Supercomputing, ACM/IEEE 2001 Conference* (2001), IEEE, pp. 58–58.

[22] Ross, R. B., Thakur, R., et al. Pvfs: A parallel file system for linux clusters. In *in Proceedings of the 4th Annual Linux Showcase and Conference* (2000), pp. 391–430.

[23] Rutman, N. Map/reduce on lustre-hadoop performance in hpc environments. *Langstone Road, Havant, Hampshire, P09 ISA, England, Tech. Rep* (2011).

[24] Schmuck, F. B., and Haskin, R. L. Gpfs: A shared-disk file system for large computing clusters. In *FAST* (2002), vol. 2, p. 19.

[25] Schwan, P. Lustre: Building a file system for 1000-node clusters. In *Proceedings of the 2003 Linux Symposium* (2003), vol. 2003.

[26] Schwarzkopf, M., Konwinski, A., Abd-El-Malek, M., and Wilkes, J. Omega: flexible, scalable schedulers for large compute clusters. In *SIGOPS European Conference on Computer Systems (EuroSys)* (Prague, Czech Republic, 2013), pp. 351–364.

[27] Vavilapalli, V. K., Murthy, A. C., Douglas, C., Agarwal, S., Konar, M., Evans, R., Graves, T., Lowe, J., Shah, H., Seth, S., et al. Apache hadoop yarn: Yet another resource negotiator. In *Proceedings of the 4th annual Symposium on Cloud Computing* (2013), ACM, p. 5.