Software Researches for Big Data and Extreme-Scale Computing

Gfarm/BB – Gfarm File System for Node-local burst buffer

http://oss-tsukuba.org/en/software/gfarm

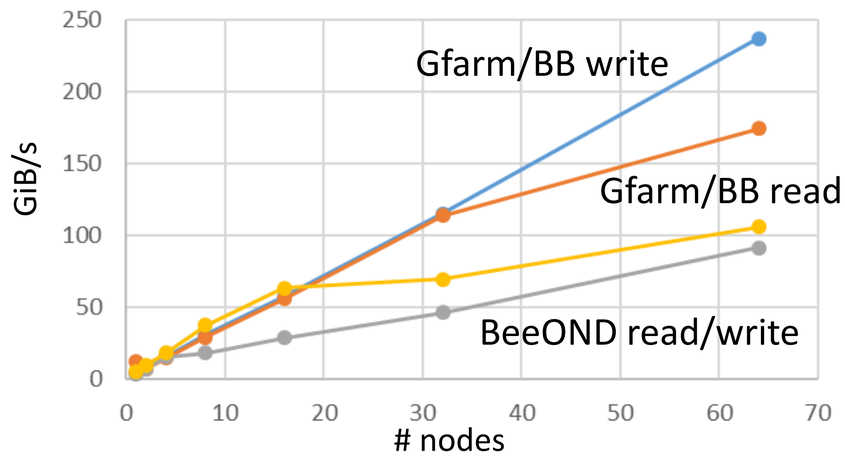


Fig. 1: IOR file-per-process read/write performance on Cygnus supercomputer

```
gfarmbb –h hostfile –m mount_point start
…
gfarmbb –h hostfile stop
```

Features include

- Open source
- Exploit local storage, and data locality for scalable I/O performance
- InfiniBand support
- Data integrity is supported for silent data corruption
- 22,000 downloads since March 2007
- Production systems: 8PB JLDG, 100PB HPCI Storage, etc.

Locality-aware MPI-IO: Scalable I/O access in N-1 pattern for existent MPI-IO applications [1]

To achieve highly-scalable storage by increasing compute nodes, the node-local storage is a key component. However, it is not obvious to enable the application to utilize the benefit of node-local storage because of its access pattern problem (e.g. N-1 pattern). We convert N-1 to N-N automatically in the MPI-IO library-level to enable that. Our implementation changes MPI-IO behavior implicitly so that the developer does not need to modify existent HPC application codes.

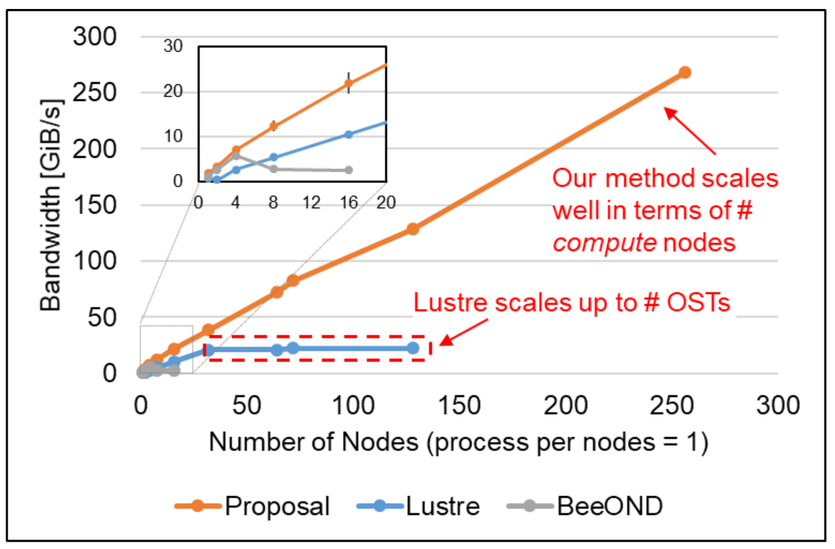


Fig. 2: IOR runs in single-shared-mode (emulates N-1 access). Each IOR process writes 10 GiB in MPI-IO non-collective access. The experiment was conducted on TSUBAME 3.0 at Tokyo Tech. Lustre has 68 OSTs. Our method successfully demonstrates scalable bandwidth.

Design of Ceph Storage Connector for Apache Spark [2]

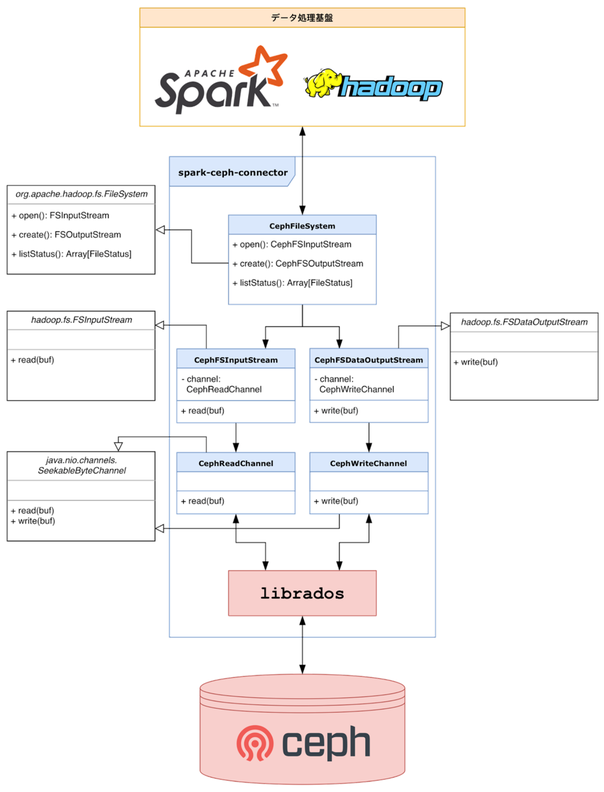


Fig. 3: Design of spark-ceph-connector implementing Hadoop Filesystem API

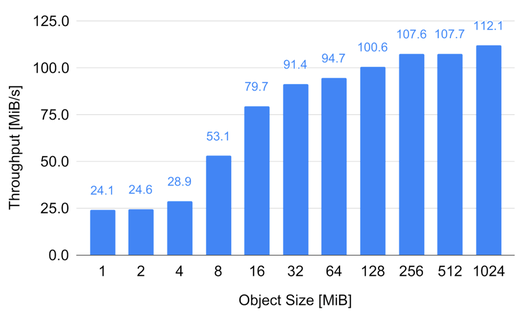


Fig. 3: Read performance

Fig. 4: Read performance

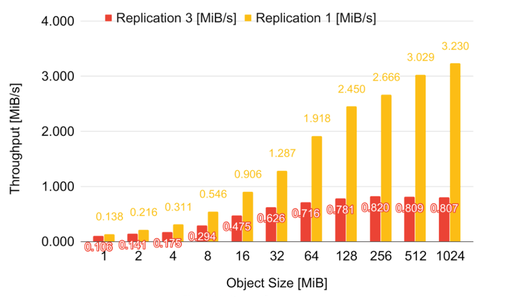


Fig. 4: Write performance

Fig. 5: : Write performance

We design and implement the storage connector of Ceph for Apache Spark & Apache Hadoop. It exploits the scalability of the Ceph distributed object store to analyze large-scale data.

The connector implements the Hadoop Filesystem API-compatible interface in Scala. It also can be used by any HDFS-compatible data processing applications.

We are currently developing the connector to improve its write performance and testing its scalability on the environment at large scale.

Construction/Destruction of Docker Cluster with User Privileges [3]

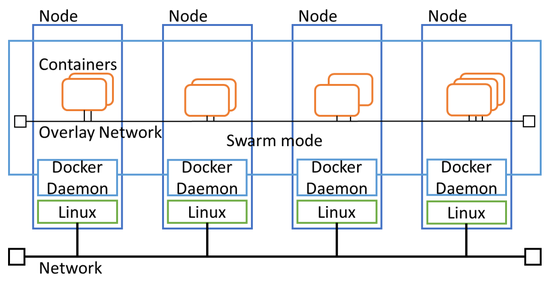
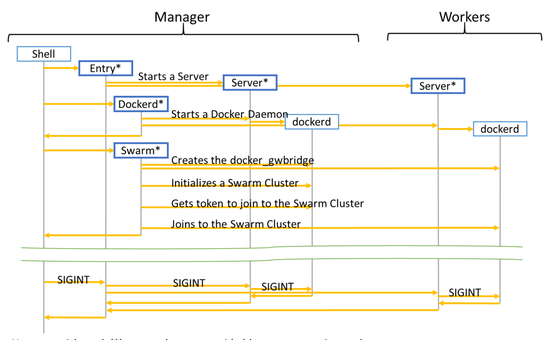


Fig. 6: Summary of an orchestration



Note: asterisk mark (*) means that are provided by our automation tools.

Fig. 7: Automation of construction/destruction a swarm cluster

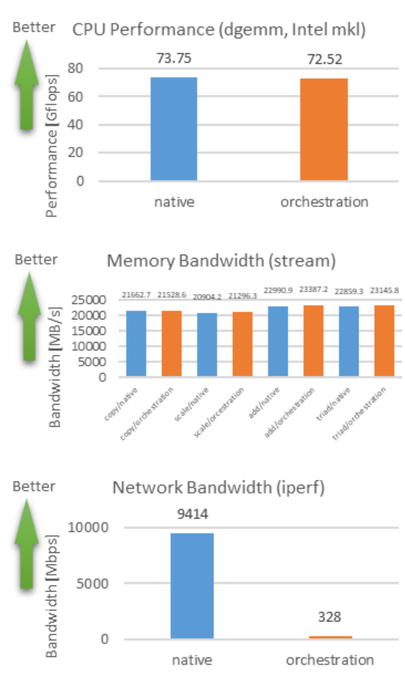


Fig. 8: Performances of a swarm cluster

In HPC, it is progressing to use container technology. But there are problems to use multi-container; how to determine on which node the containers will run, or how to resolve names.

An Orchestration using Docker Rootless mode (introduced in v1 903) will solve these problems. We are working on automation of construction/destruction of the cluster, and evaluation of orchestration.

Reference

[1] Kohei Sugihara, Osamu Tatebe, "Designing MPI-IO for Node-Local Burst Buffer", IPSJ SIG Technical Reports, Vol. 2019-HPC-168, No. 22, pp. 1-7, 2019 (In Japanese/Kanji)

[2] TAKAHASHI Shuuji*, TATEBE Osamu: “Design of Ceph storage connector for Apache Spark”, IPSJ SIG Technical Reports, Vol. 2019-HPC-171, No. 1, pp. 1-8, 2019 (In Japanese/Kanji)

[3] Tomoyuki Hatanaka, Osamu Tatebe: ”Construction/Destruction of Docker Cluster with User Privileges”, IPSJ SIG Technical Reports, Vol. 2019-HPC-171, No. 3, pp. 1-6, 2019 (In Japanese/Kanji)

Acknowledgment

This work is partially supported by the JST CREST Grant Numbers JPMJCR1414, JSPS KAKENHI Grant Number JP17H01748, New Energy and Industrial Technology Development Organization (NEDO), and Fujitsu Laboratories Ltd.

contact address: pr@ccs.tsukuba.ac.jp

https://www.ccs.tsukuba.ac.jp/