# Mig-drive – a High Performance HDF5 Driver with Meta-data Migrate

Cao Liqiang, Shen Weichao
Institute of Applied Physics and Computational Mathematics, Beijing, 100094, China

HDF5 is a data format and I/O library used by many scientific computing applications. Data in HDF5 are described and organized by meta-data in file. Both data and meta-data are aligned in logical storage space management by HDF5's file driver. In most of applications, data may take up to 95% or more file space while meta-data may take less than 5%. Most of HDF5's file driver, such as default Sec2 and Mpio, write data and meta-data to one file, while Split driver writes data and meta-data in files with different suffix.

We develop a new HDF5 file driver called Mig-driver which move HDF5's meta-data file between memory file system and storage file system in Linux. We keep HDF5 file's consistency by meta-data file migrate when HDF5 create, open and close. While HDF5 creates file, Mig-driver create a data file in working directory and a shadow meta-data file in Tmpfs. All meta-data I/O is directed to shadow file until file closed, and then shadow file will be move to working directory. While HDF5 open file in working directory, we create a shadow meta-data file in Tmpfs by copying meta-data to it firstly, and then all meta-data I/O is direct to Tmpfs till HDF5 file closed. If meta-data file changed, meta-data in working directory will be updated.

Scientific computing applications may invoke tens of thousands HDF5 calls in minutes. Most of HDF5 calls, such as type related (H5T) and space related (H5S) is read/write meta-data only, and others is mixed meta-data read/write and data read/write. Compared with storage file system, Tmpfs reduces I/O latency up to 6 orders of magnitude, which improves HDF5's meta-data I/O performance substantially even meta-data migrate may take some time.

We test I/O bandwidth of HDF5 with Sec2 driver and Mig-driver with a serial I/O benchmark and parallel data processing tools called TeraVAP. In serial write test, we write 5000 data blocks one by one. In read test, we read 1250 blocks back with a random choice from 5000. TeraVAP is a parallel data processing tools developed by IAPCM(Institute of Applied Physics and Computational Mathematics). It read and processing HDF5 data in parallel mode. We compared the read bandwidth of HDF5 with two different drives in TeraVAP with 56GB HDF5 file.

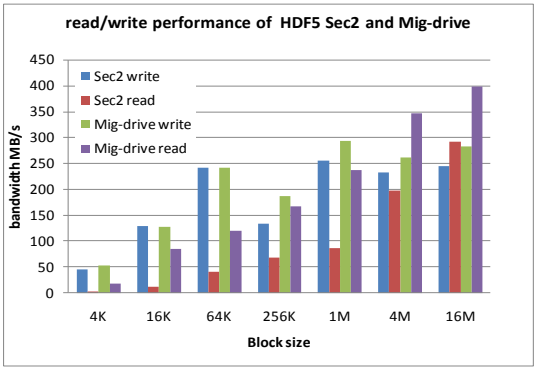


Fig.1 Serial test of HDF5 Mig-drive and Sec2

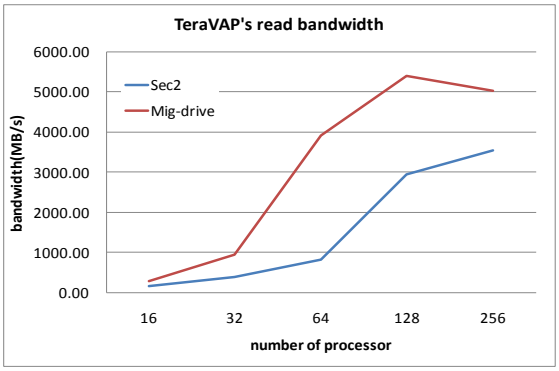


Fig. 2 read Performance with TeraVAP

As Fig. 1 shows, write bandwidth of Mig-drive is minor faster than Sec2. That's because of the file system cache improve Sec2's meta-data write performance. Without cache hit in reading, Mig-drive's bandwidth improves to 3.5 times on average. Fig.2 is parallel reading test of TeraVAP. From 16 processors to 256 processors, Mig-drive improves parallel read performance up to 2.4 times on average.