High Performance Computing Systems (CMSC714)

Lecture 18: Parallel I/O

Abhinav Bhatele, Department of Computer Science

UNIVERSITY OF MARYLAND

Summary of last lecture

- Task mapping can be used to optimize the placement of MPI processes within a job allocation
- Can reduce inter-node communication volume and optimize it
- Heuristic-based approaches
- Metrics: hop-count, hop-bytes

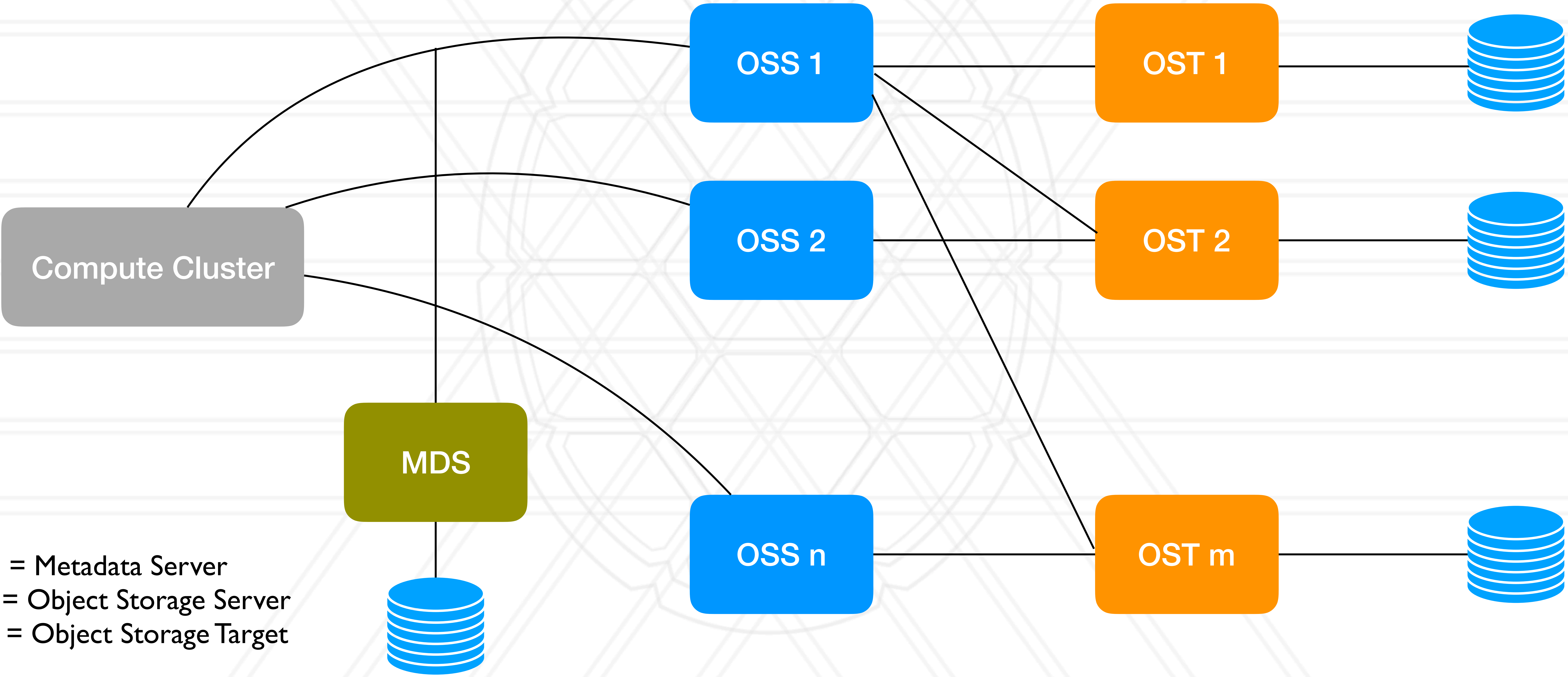
When do parallel programs perform I/O?

- Reading input datasets
- Writing numerical output
- Writing checkpoints

Non-parallel I/O

- Designated process does I/O
- All processes send data to/receive data from that one process
- Not scalable

Parallel File System



MDS = Metadata Server
OSS = Object Storage Server
OST = Object Storage Target

Parallel I/O: One file per process

- Does not scale to large process counts
- Large number of files leads to bottlenecks in metadata operations
- Simultaneous disk accesses lead to contention for file system resources

Parallel I/O: Shared file

- All processes write to the same file
- Simultaneous disk accesses lead to contention for file system resources

Parallel I/O: Hybrid approach

- Subsets of processes aggregate data to a designated process
- Some designated processes perform I/O
- Leads to less contention during metadata operations
- Less contention for file system resources

Available systems/tools

- Parallel file system: Lustre, GPFS, PVFS
- Middleware: MPI-IO
- Higher-level libraries: HDF5, netCDF, Adios

Questions

PVFS: A Parallel File System for Linux Clusters

- Are all I/O daemons necessarily of same size?
- Can you illustrate a little bit about section 3.3 about how ROMIO works? What small set of functions it will use? Data sieving?
- Is PVFS also sensitive to cache problem as Lustre and GPFS? Do Lustre and GPFS have trapping calls?
- Is PVFS still popular, or is it replaced by other alternatives?
- What is Myrinet in more detail, what made it so fast?

Questions

Parallel I/O and the Metadata Wall

- What is a node aggregator?
- In figure 5 or other figures, what is directory stat? What is stat operation? Why it scales poorly?
- What is a balanced storage area network design looks like? Any example?
- Besides goods calling reason, what else reason in the end the author mentioned about using high-level interfaces? What are the ways to tune it to take advantage of the bandwidth in the file system?
- What is the current situation on this topic? Did this paper push Lustre or IBM to improve metadata operations?
- From my understanding metadata is very small compared to the actual data, is it not possible to keep metadata in some kind of memory, instead of the disk? and write metadata to disk rarely.

Questions?

UNIVERSITY OF
MARYLAND

Abhinav Bhatele

5218 Brendan Iribe Center (IRB) / College Park, MD 20742

phone: 301.405.4507 / e-mail: bhatele@cs.umd.edu