Parallel I/O Analysis

Instructor: Dr. Preeti Malakar

- P1: Aditya Rohan (160053)
- P2: Anshul Vijayvergiya (150113)

Goals

- Profile and trace the I/O performance of cse cluster and HPC 2010
- Study how the topology is affecting the performance

- Suggest semi-topology aware optimizations

Parallel IO

- Multiple processes reading/writing to storage simultaneously
- Sequential I/O is too slow for data of order of TBs
- Parallel I/O needs a parallel FS

Profiling vs Tracing

- Profile: Details about the execution time of different program entities and performance events; ignores the chronological order
  - Helps identify sources of contention
- Trace: Collection of time-stamped sequence of events, data increases with longer exec times
  - Helps identify cause of contention

IOPin

- On modern parallel machines, the I/O software consists of several layers, including high-level libraries such as Parallel netCDF and HDF, middleware such as MPI-IO, and low-level POSIX interface supported by the file systems.

- Pin is a software system that performs runtime binary instrumentation of Linux and Window applications.

DXT: Darshan Extended Tracing

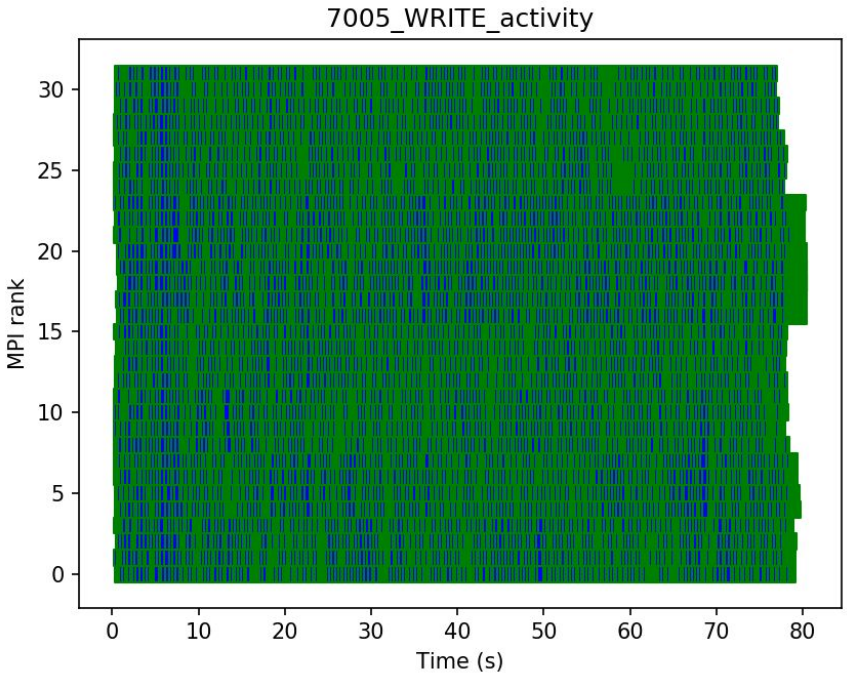
- Can be inserted at runtime(dynamic exec.) or linktime(static exec.)
- export DXT_ENABLE_IO_TRACE=1
- Overhead introduced by DXT is less than 1%
- Produces I/O activity summary for each job,
    - Counters for file operations
    - Timestamped access and execution times

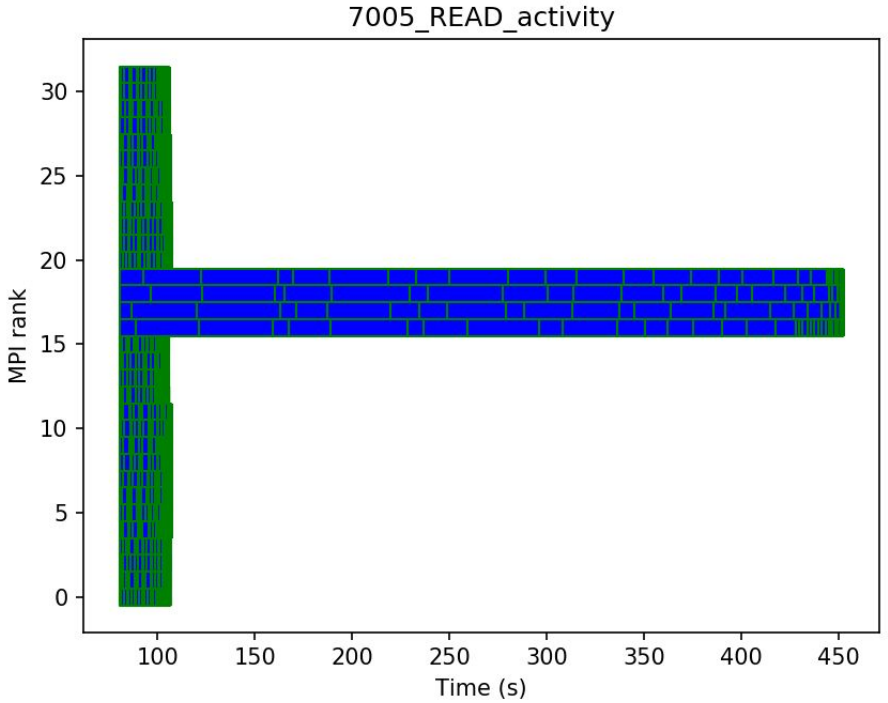
Experiment general details

- IOR, **I**nterleaved **o**r **R**andom is a parallel IO benchmark.
- Other benchmarks to be tried: S3D-IO, PnetCDF
- blockSize - size (in bytes) of a contiguous chunk of data accessed by a single process
- transferSize - size (in bytes) of a single data buffer to be transferred in a single I/O call

CSE Cluster(NFS)

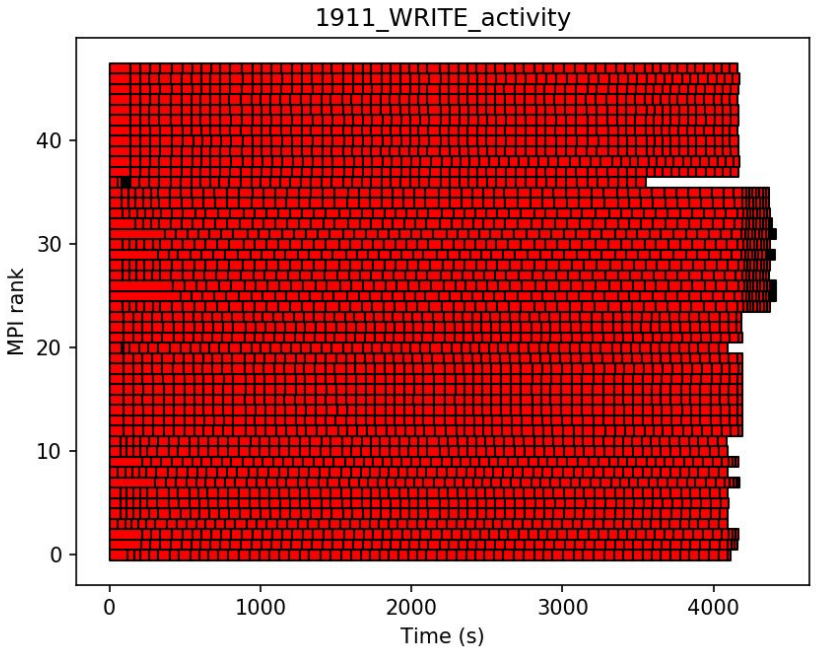
- Nodes: 1, 3, 4, 5, 6, 7, 8, 9
- 4ppn, blockSize: 16M, transferSize: 1M

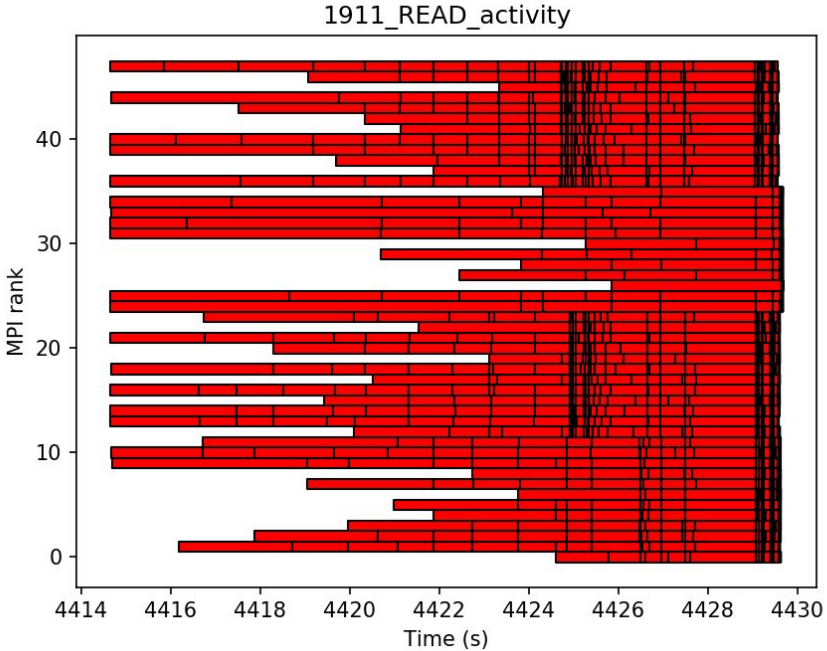




CSE Cluster(NFS)

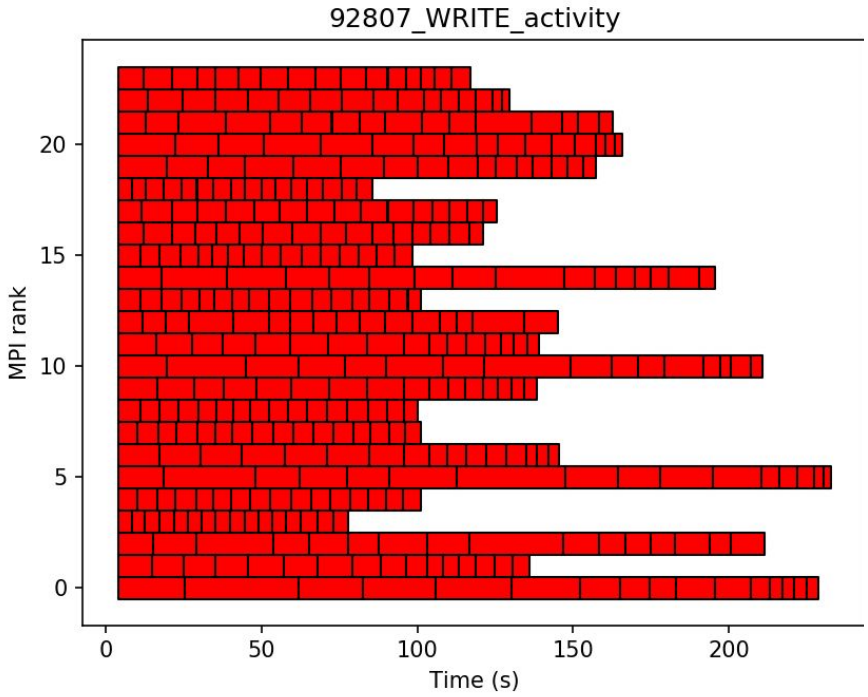
- Nodes: 18,19,20,21
- 8ppn, blockSize: 512M, transferSize: 128M

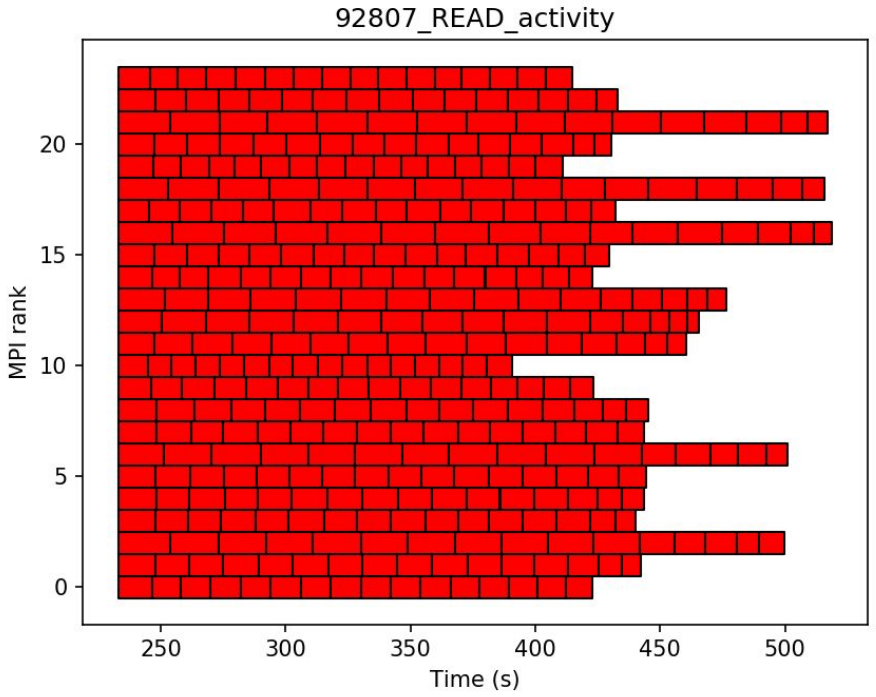




HPC - Lustre(MPIIO)

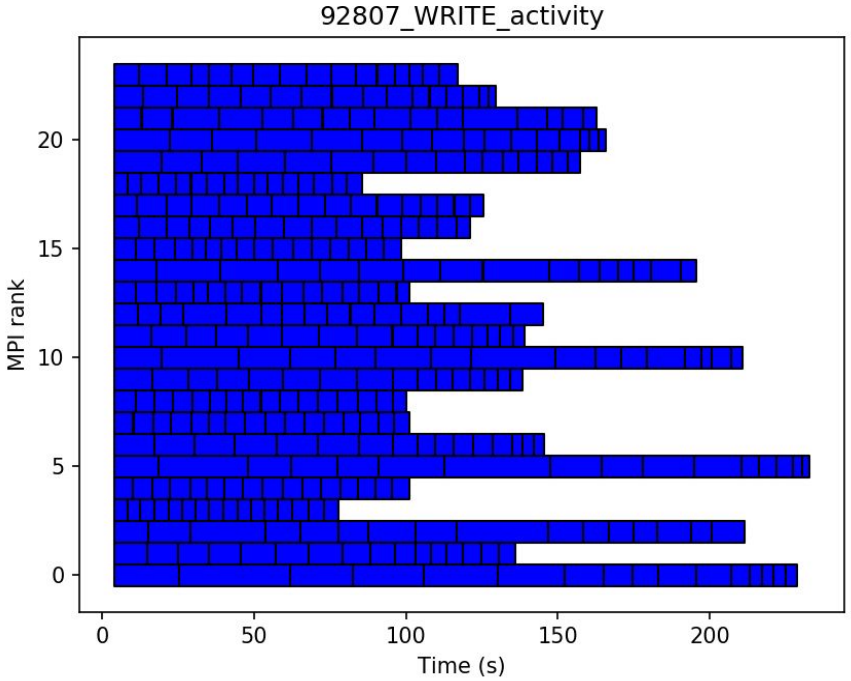
- 3 nodes, 8 ppn, blockSize:1G, transferSize: 1G

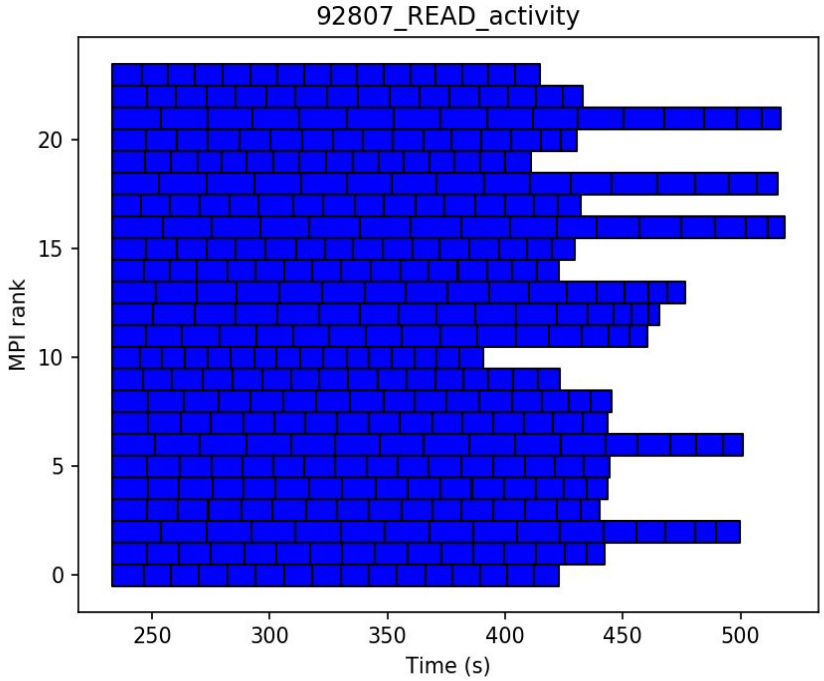




HPC - Lustre(POSIX)

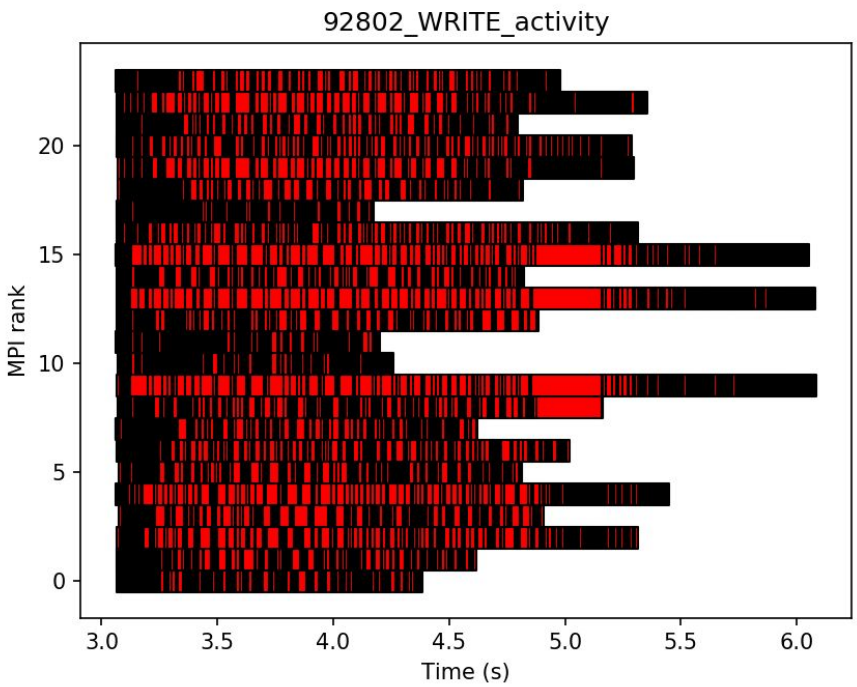
- 3 nodes, 8 ppn, blockSize:1G, transferSize: 1G

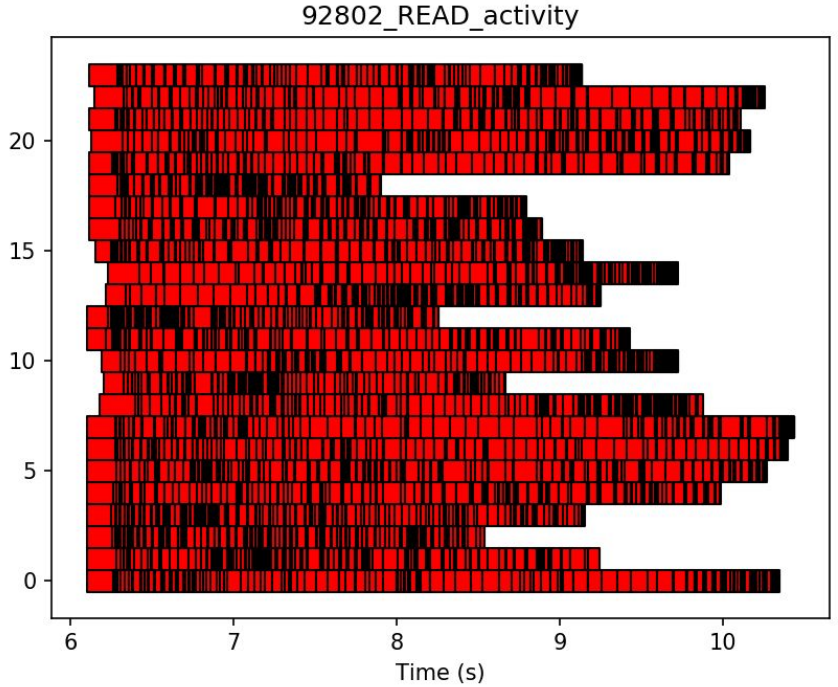




HPC - Lustre(POSIX)

- 3 nodes, 8 ppn, blockSize:16M, transferSize: 1M






Thank You!
It ain't much, but it's honest work

References

- https://media.readthedocs.org/pdf/ior/latest/ior.pdf
- https://cug.org/proceedings/cug2017_proceedings/includes/files/pap105s2-file1.pdf
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6495796
- https://www.slideshare.net/insideHPC/hpc-io-for-computational-scientists
- http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-Parallel-IO.pdf