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
- Test effect of Darshan and DXT on application performance

Parallel IO

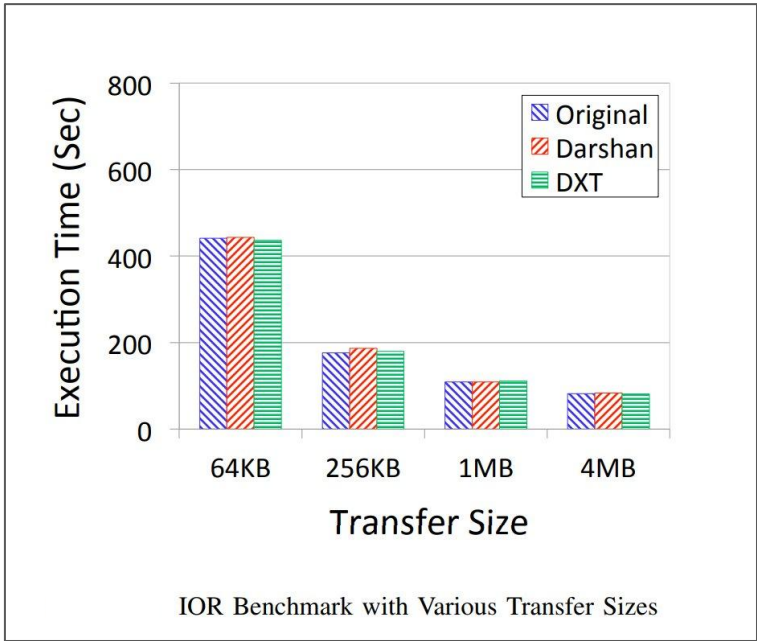
- Multiple processes reading/writing to storage simultaneously
- Sequential I/O is too slow for data of order of TBs
- Parallel I/O needs a parallel FS

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

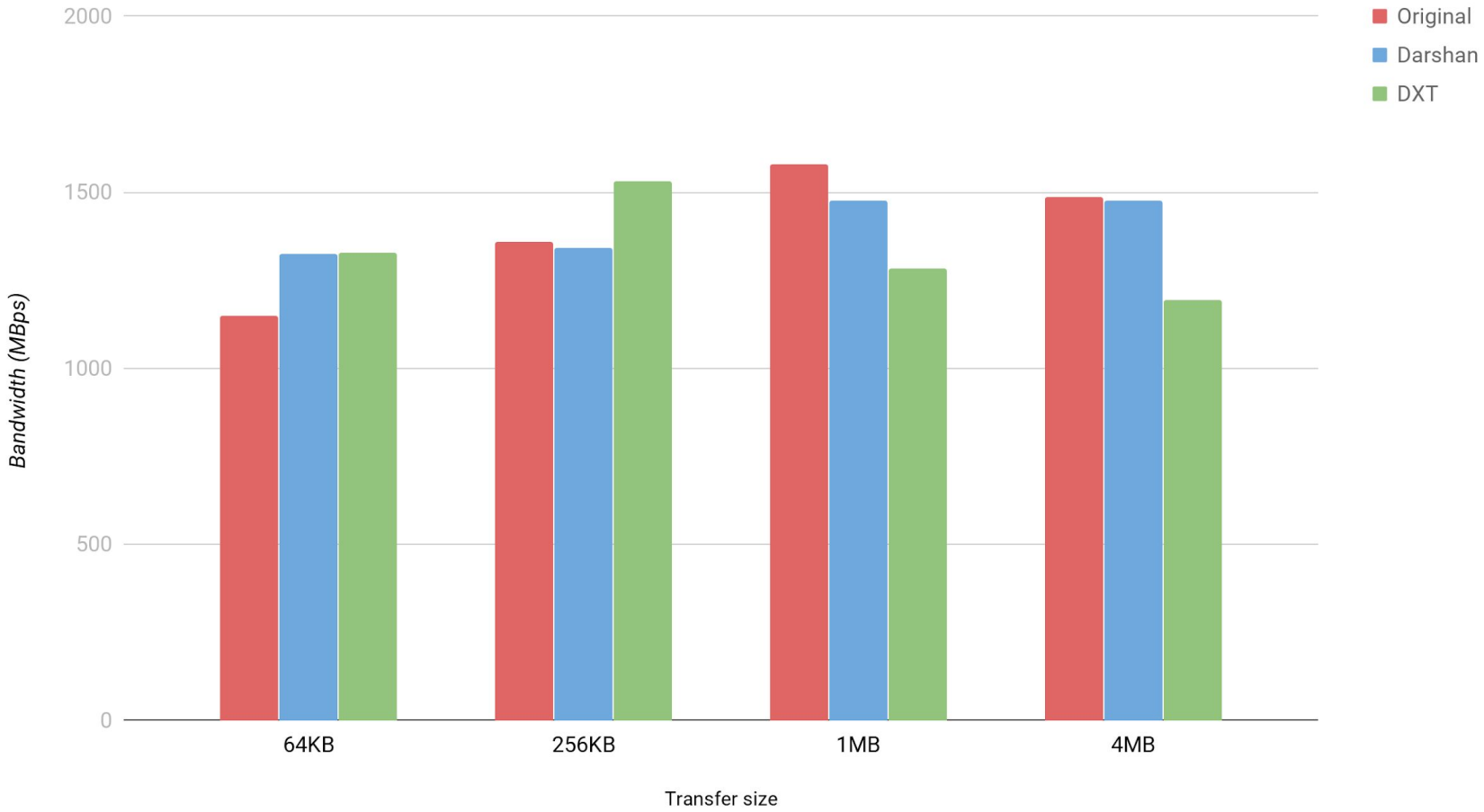
DXT: Darshan Extended Tracing

- 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

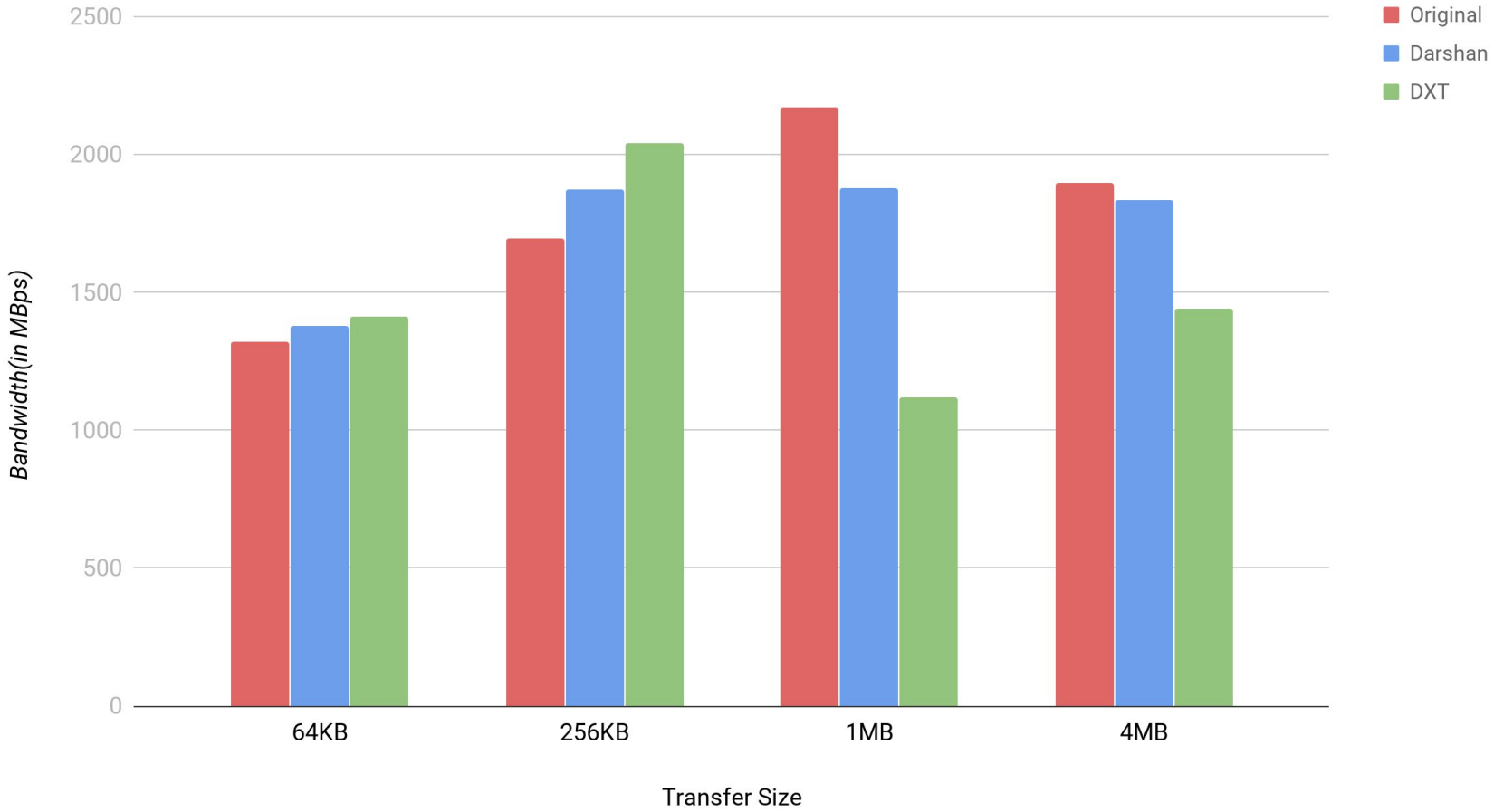


IOR Benchmark with Various Transfer Sizes

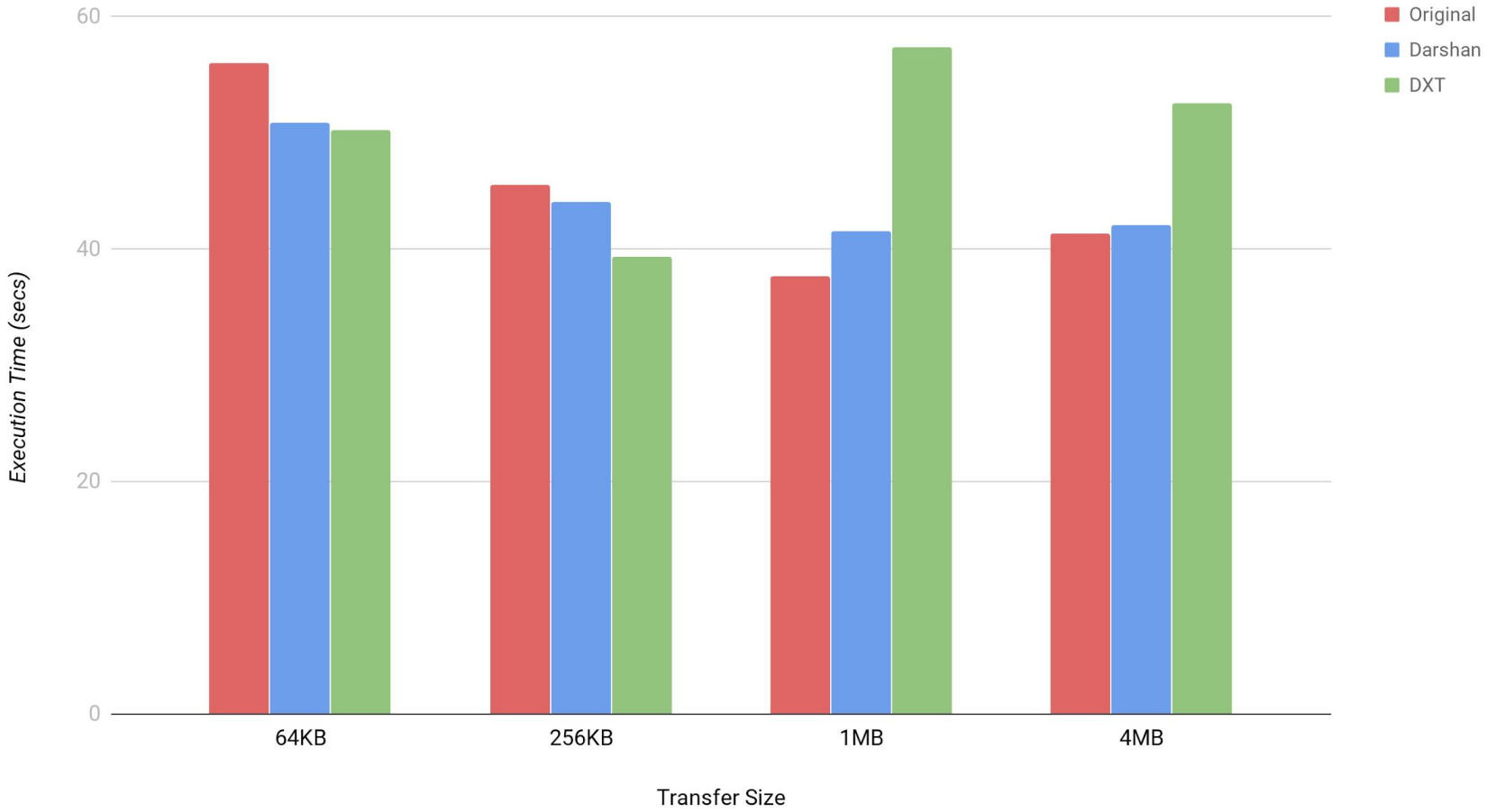
Read Bandwidth 16 processes



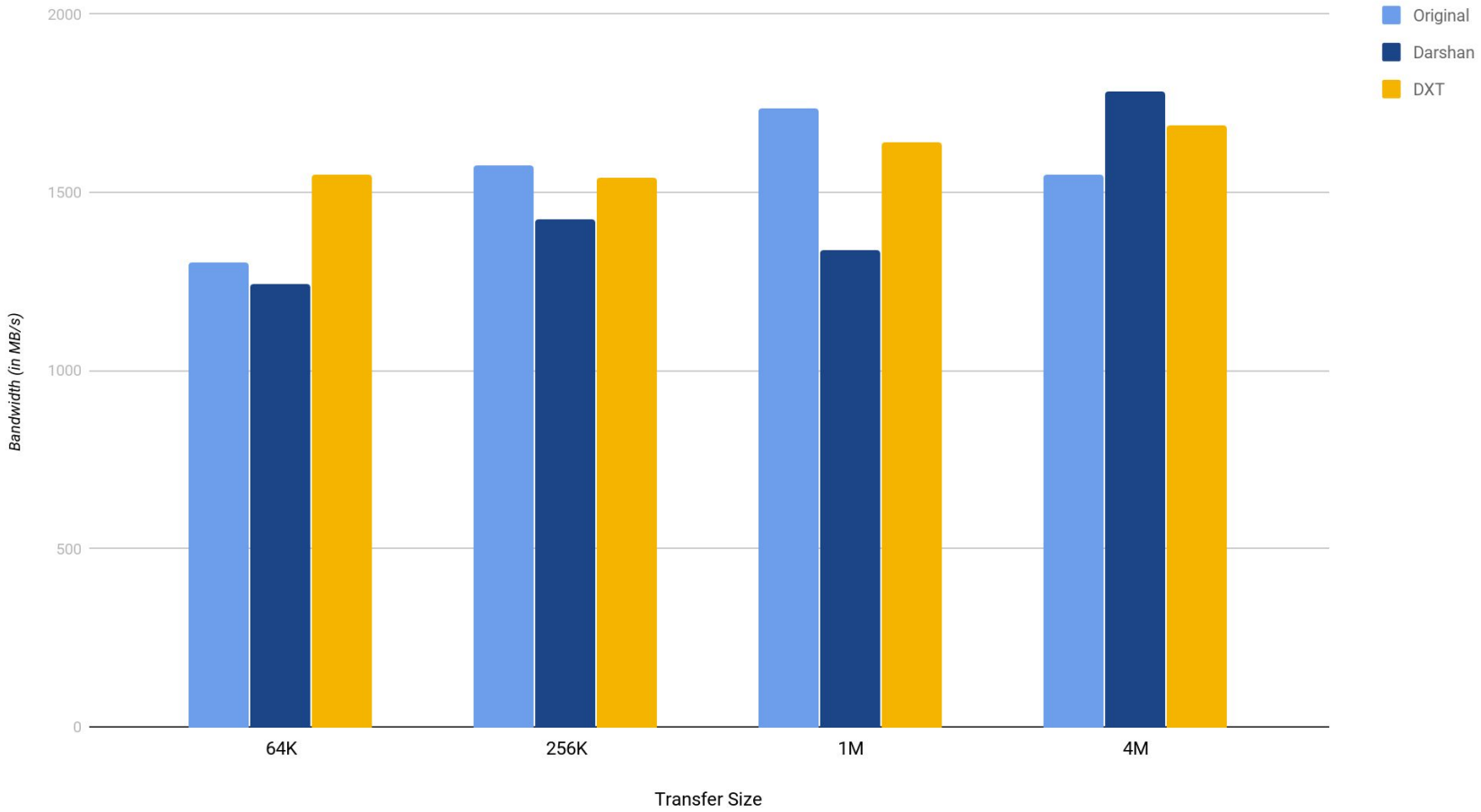
Write bandwidth 16 processes



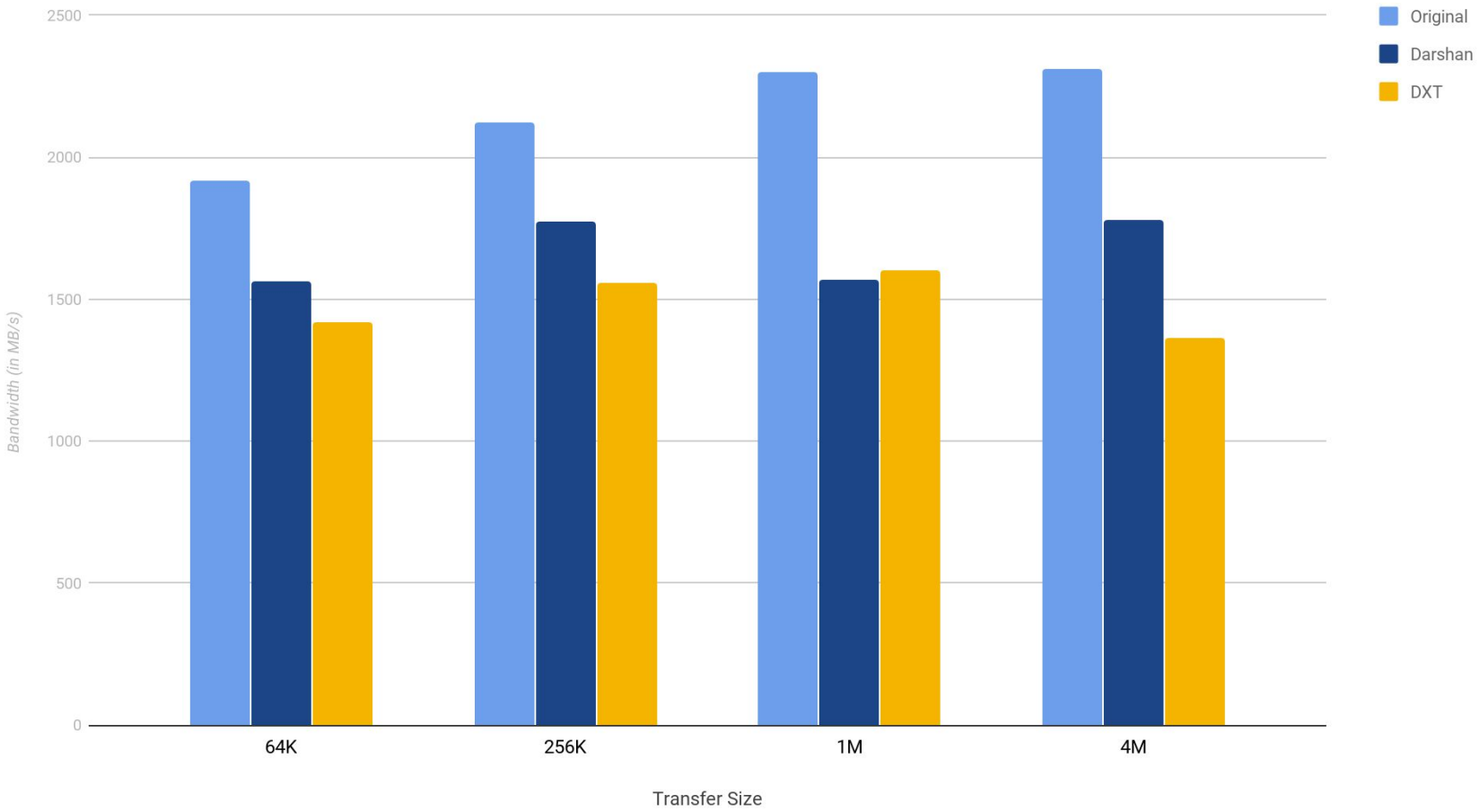
Execution Time for 16 processes



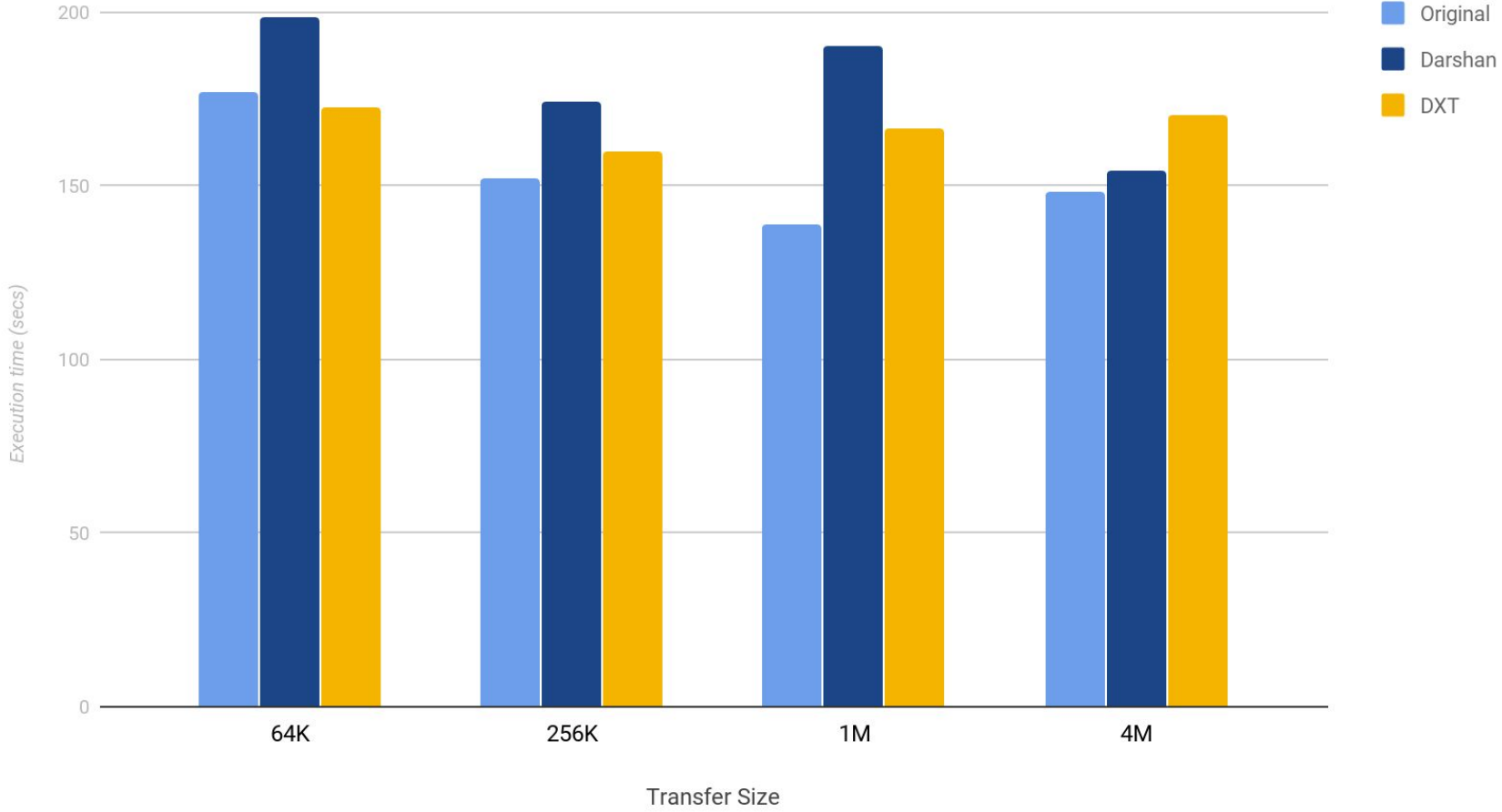
Read Bandwidth for 64 processes



Write Bandwidth for 64 processes



Execution Time for 64 processes



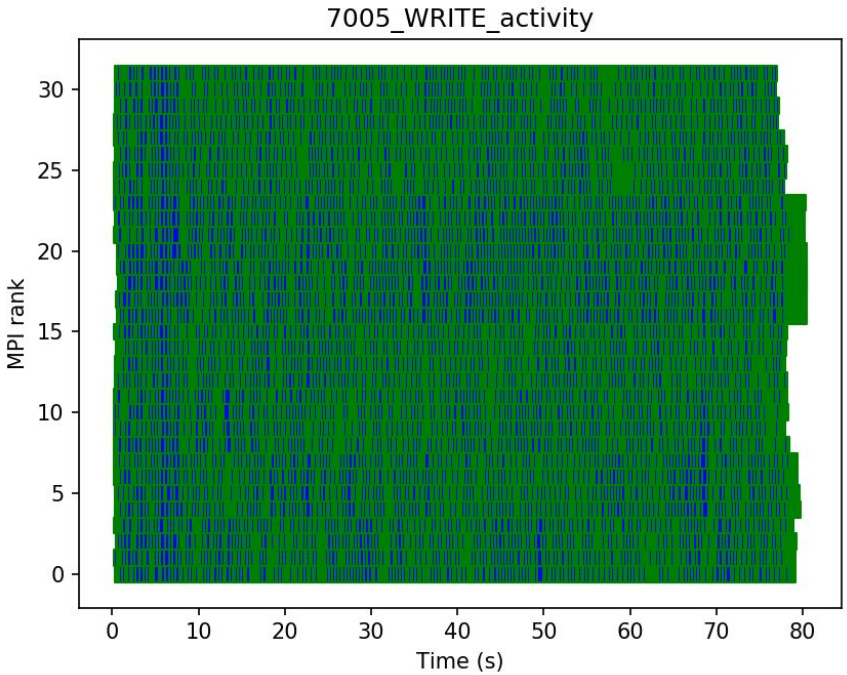
I/O Profiling Experiments

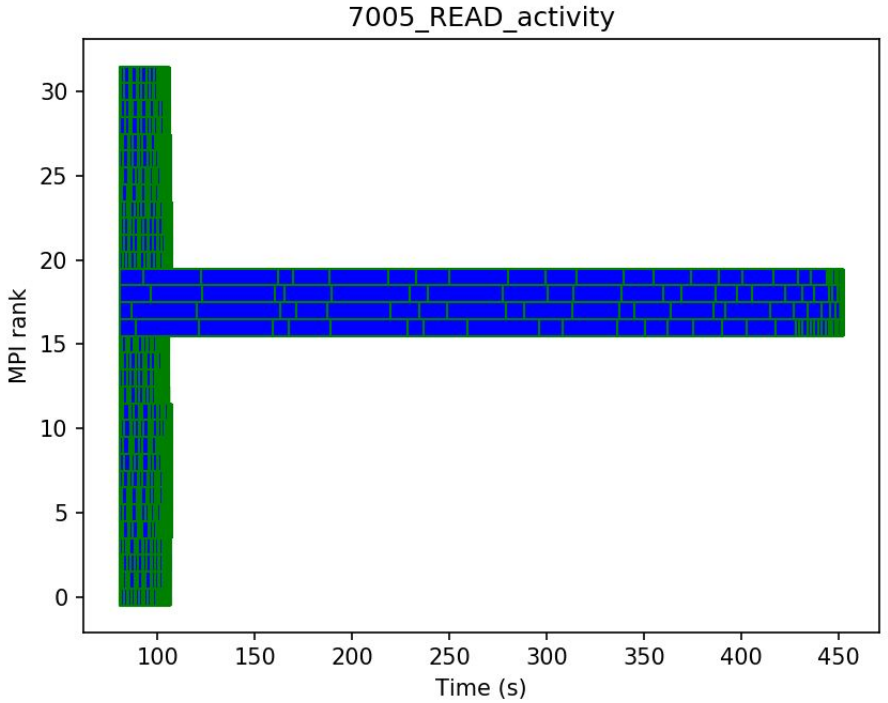
Experiment general details

- IOR, **I**nterleaved **o**r **R**andom is a parallel IO benchmark.
- S3D-IO
- 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)

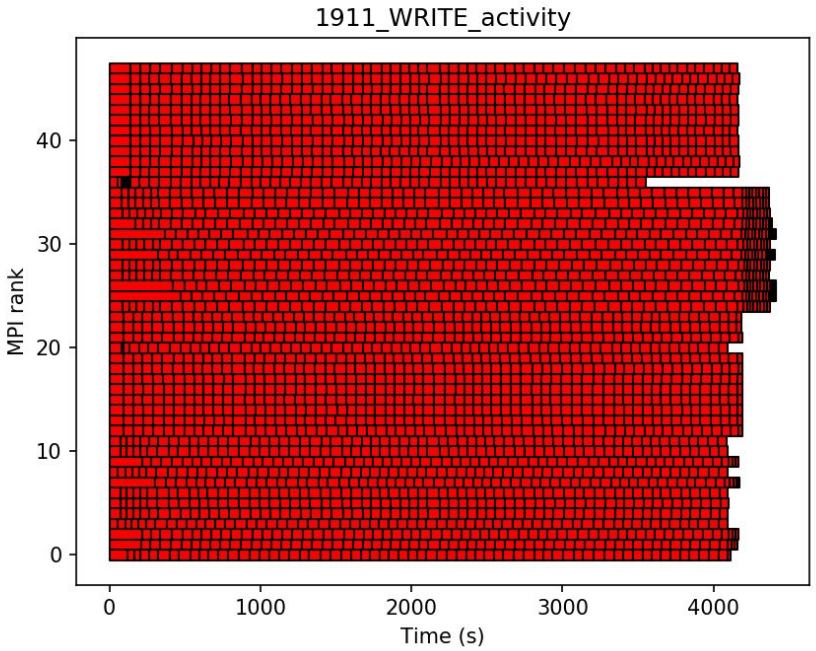
- Nodes: 1, 3, 4, 5, 6, 7, 8, 9
- 4ppn, blockSize: 16M, transferSize: 1M

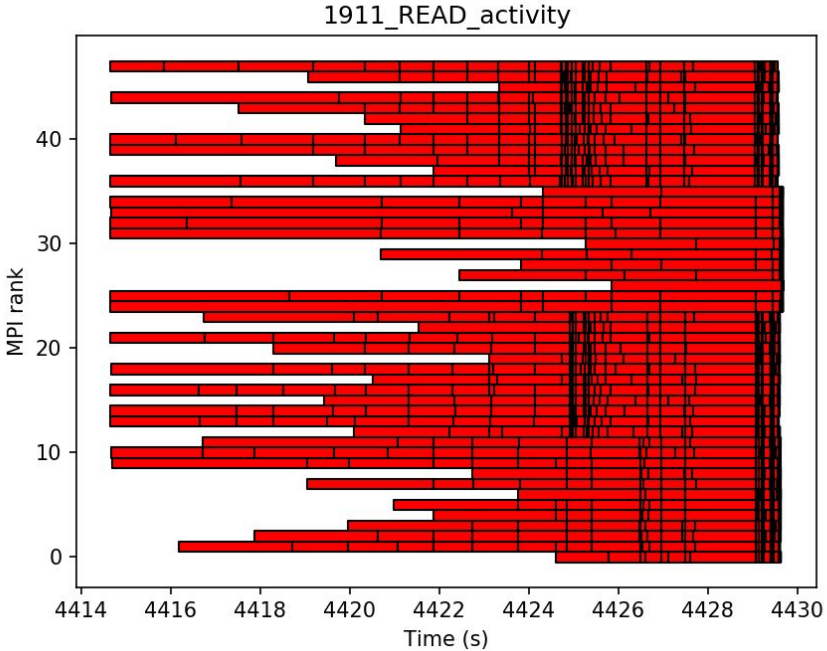




CSE Cluster(NFS)

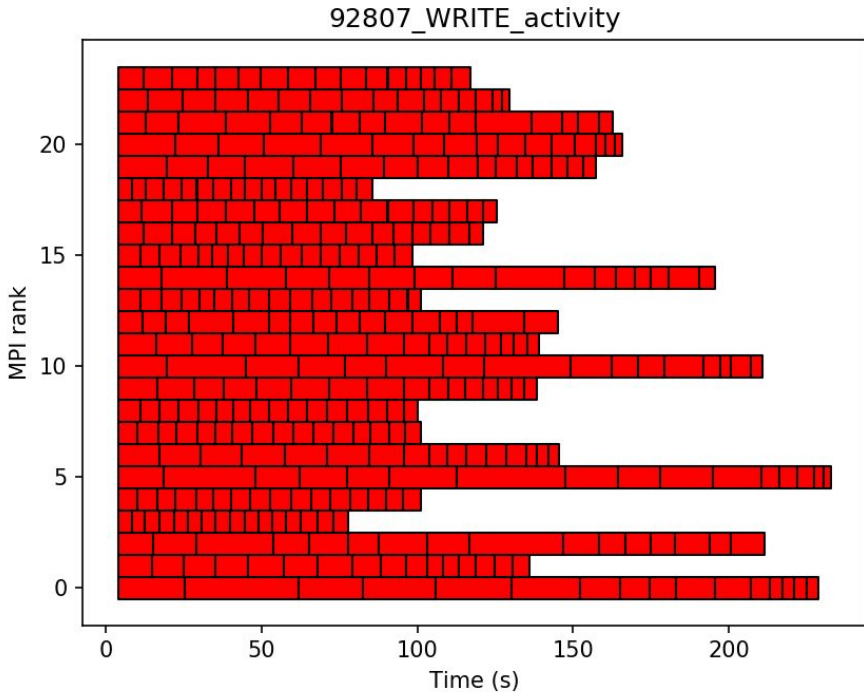
- Nodes: 18,19,20,21
- 8ppn, blockSize: 512M, transferSize: 128M

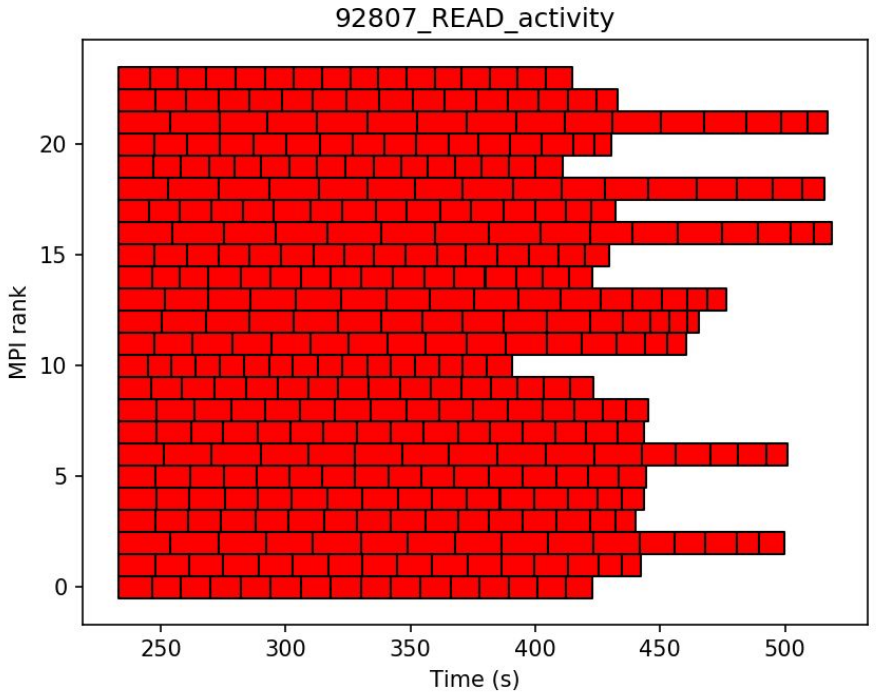




HPC - Lustre(MPIIO)

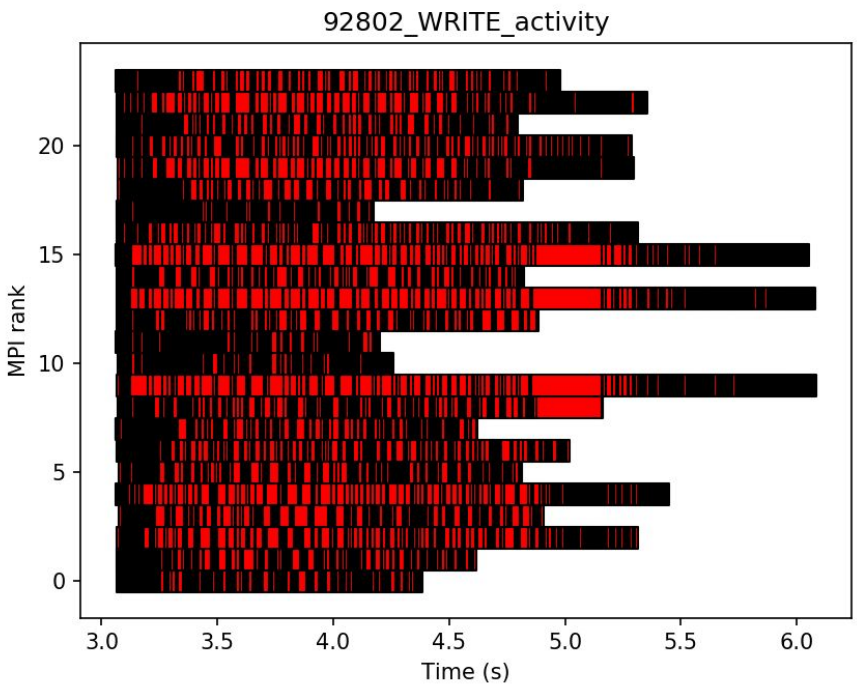
- 3 nodes, 8 ppn, blockSize:1G, transferSize: 1G

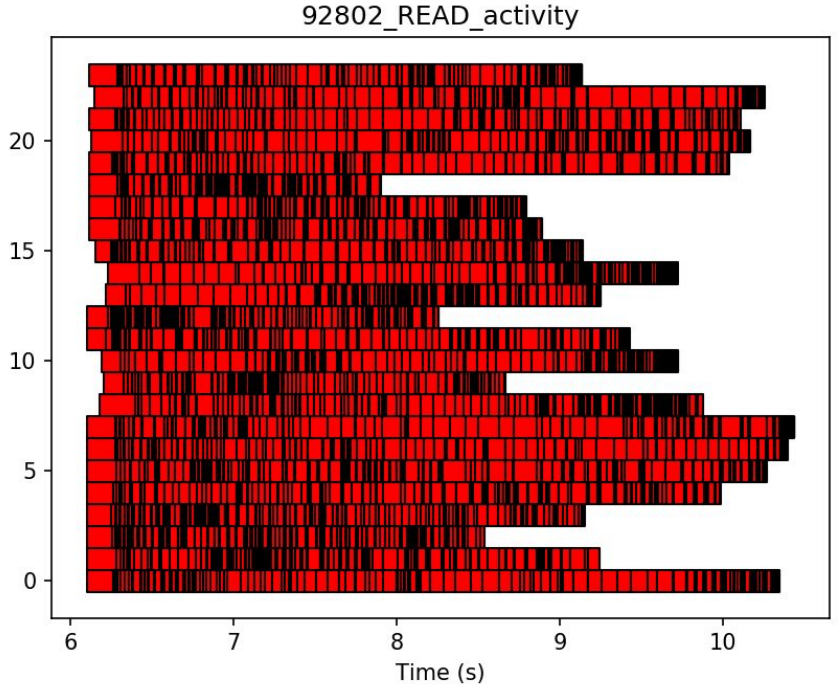




HPC - Lustre(POSIX)

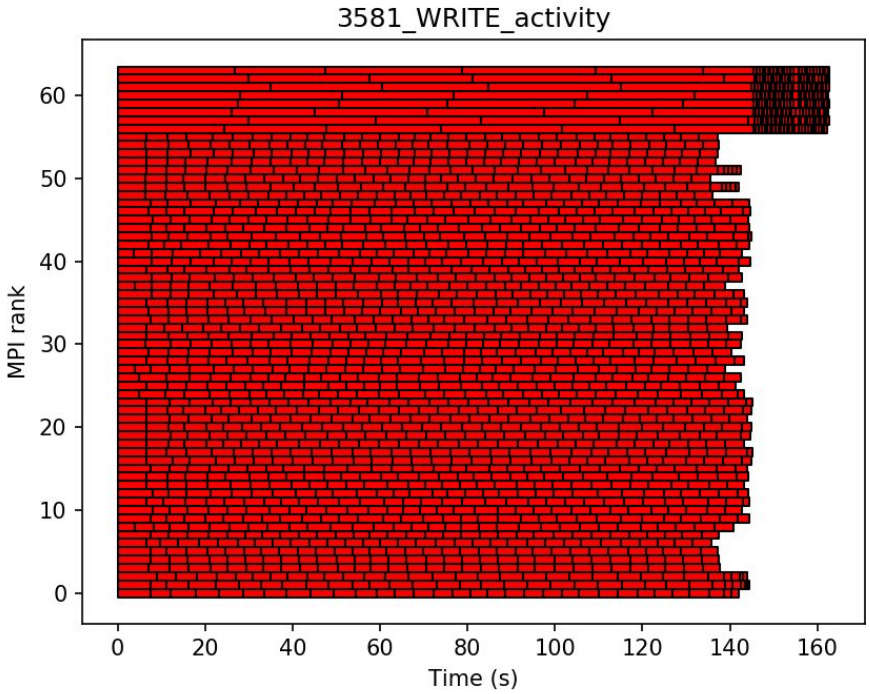
- 3 nodes, 8 ppn, blockSize:16M, transferSize: 1M

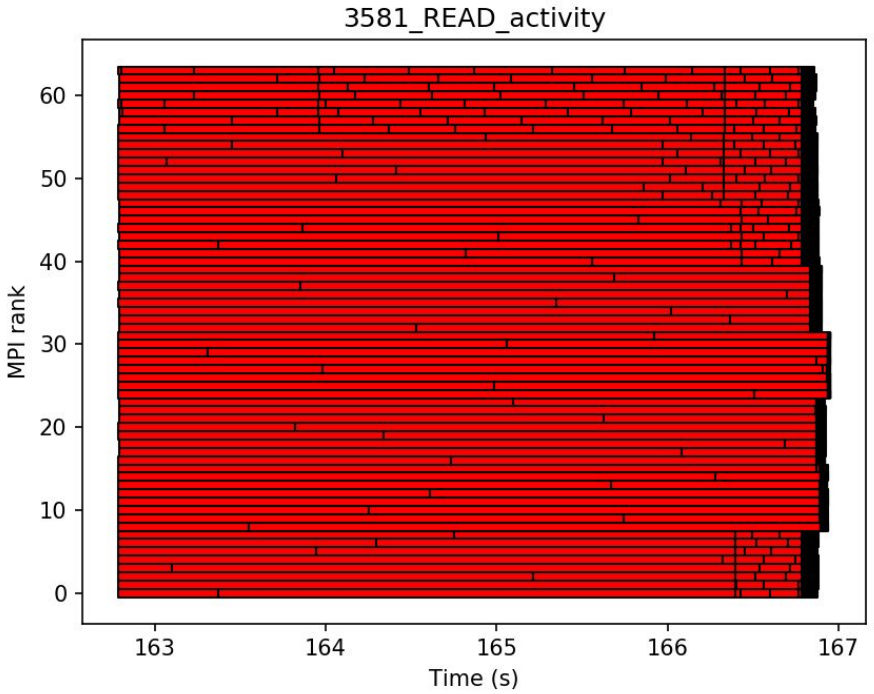




CSE Cluster (MPIIO)

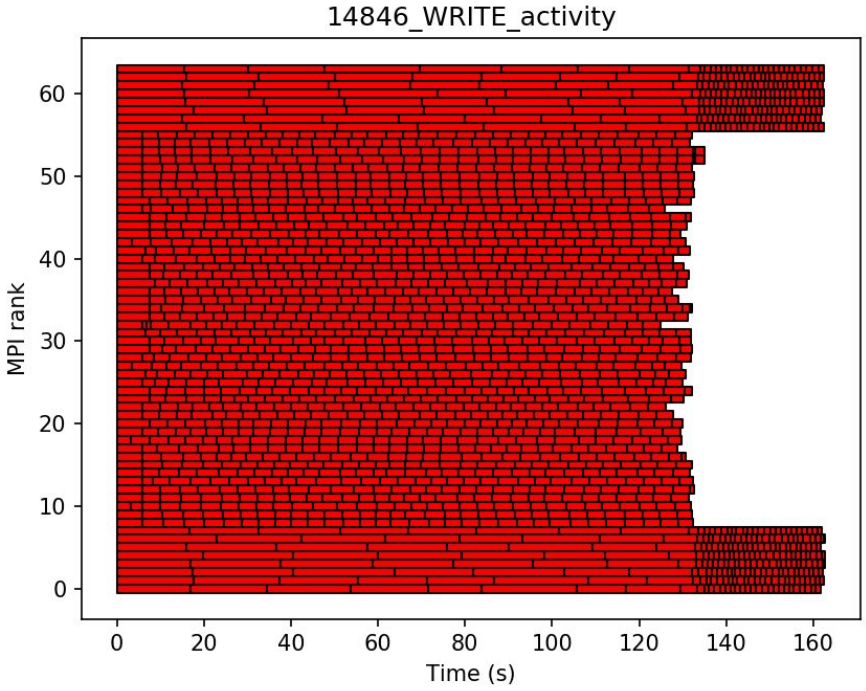
1234567,23 | 8ppn | 4m:tr | 16m:bl

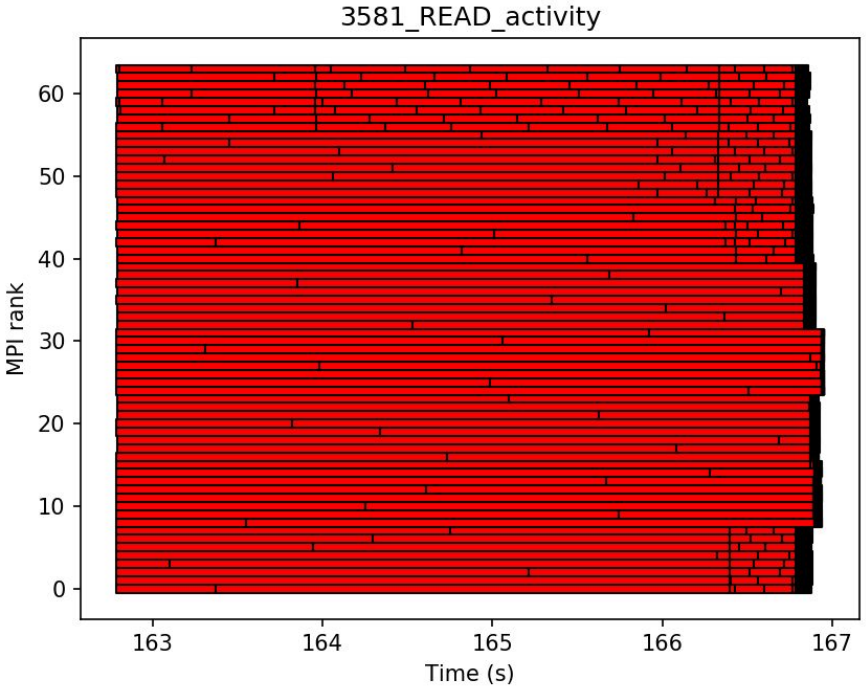




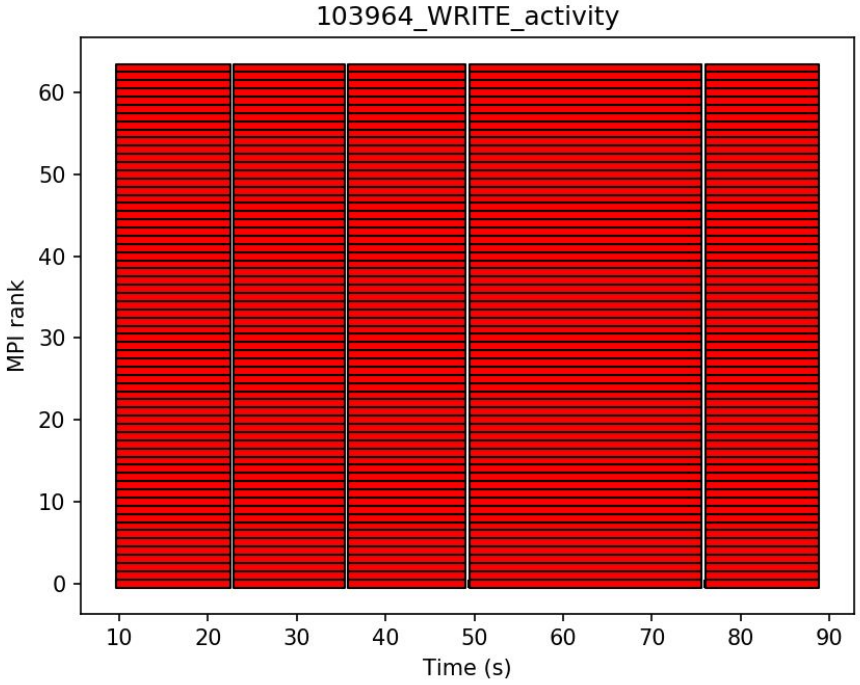
CSE Cluster (MPIIO)

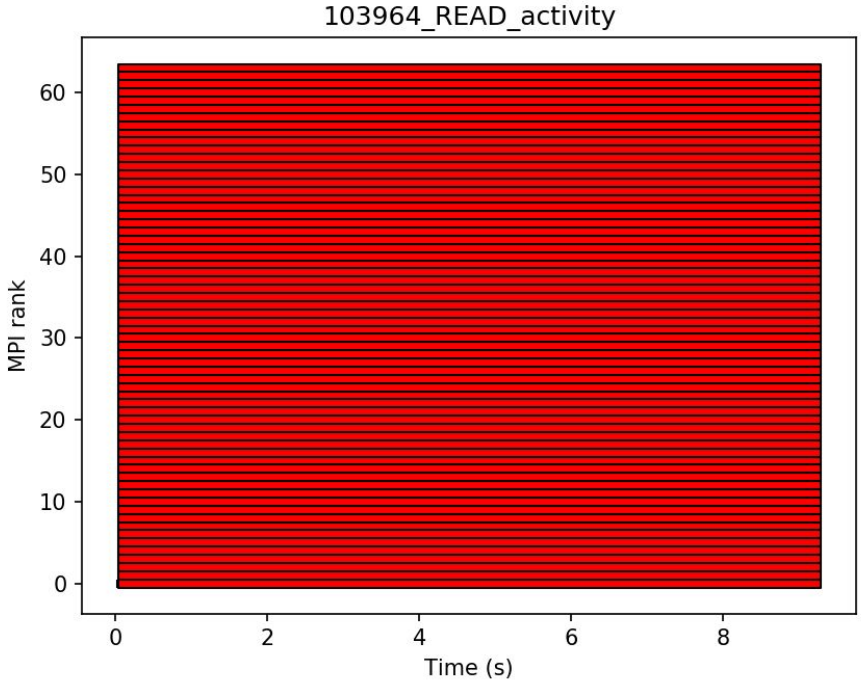
27,2,3,4,5,6,7,23 | 8ppn | 8m:tr | 16m:bl

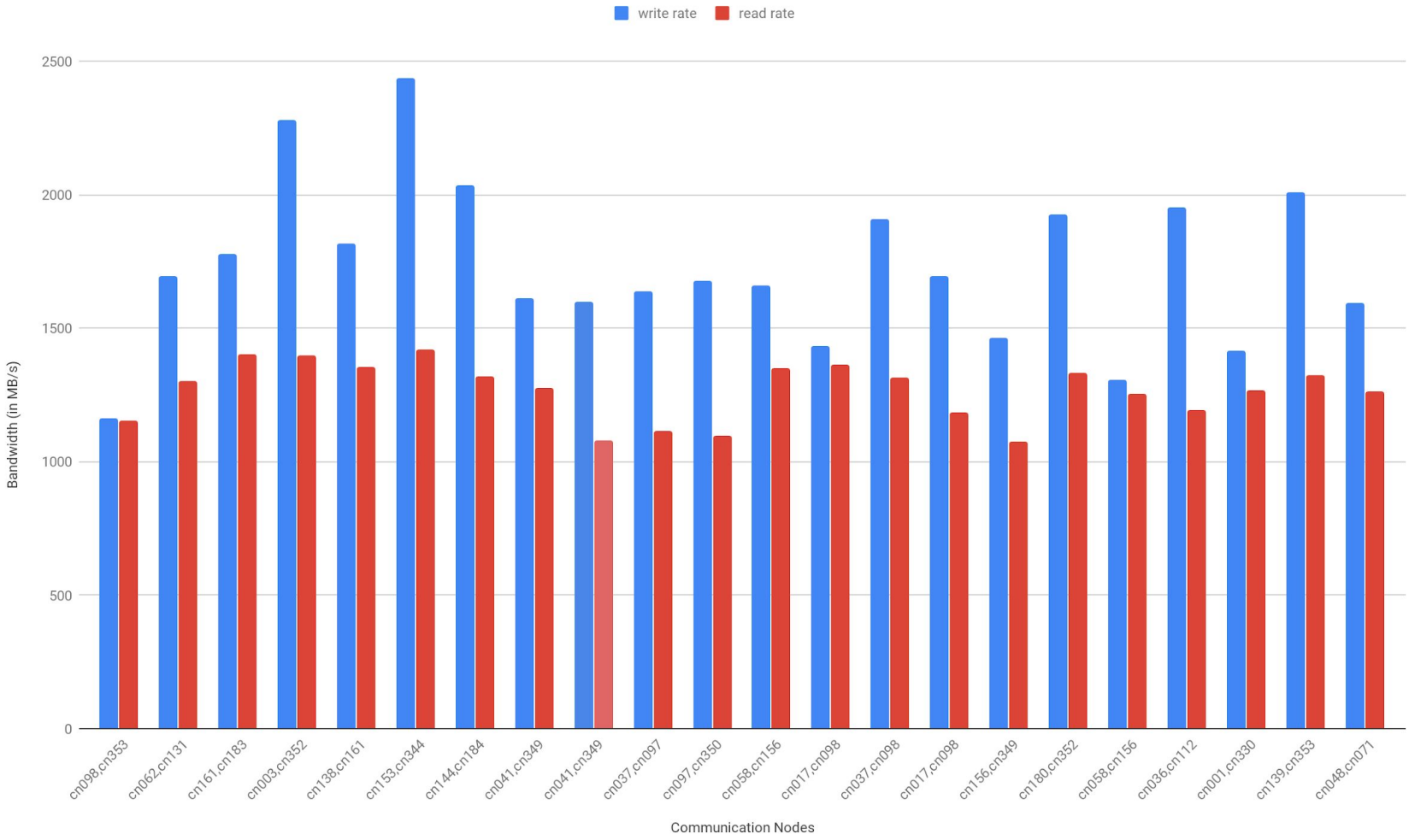


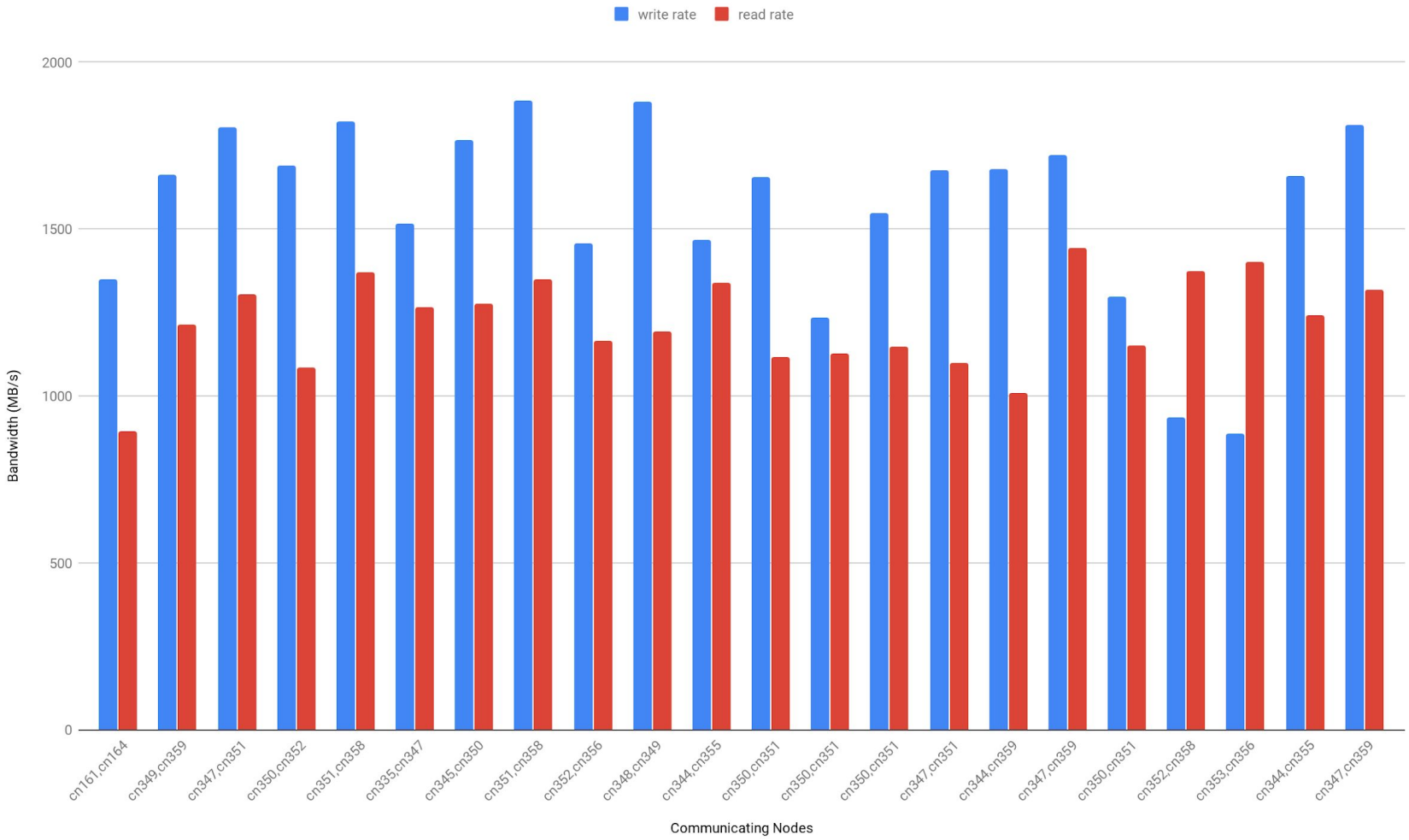


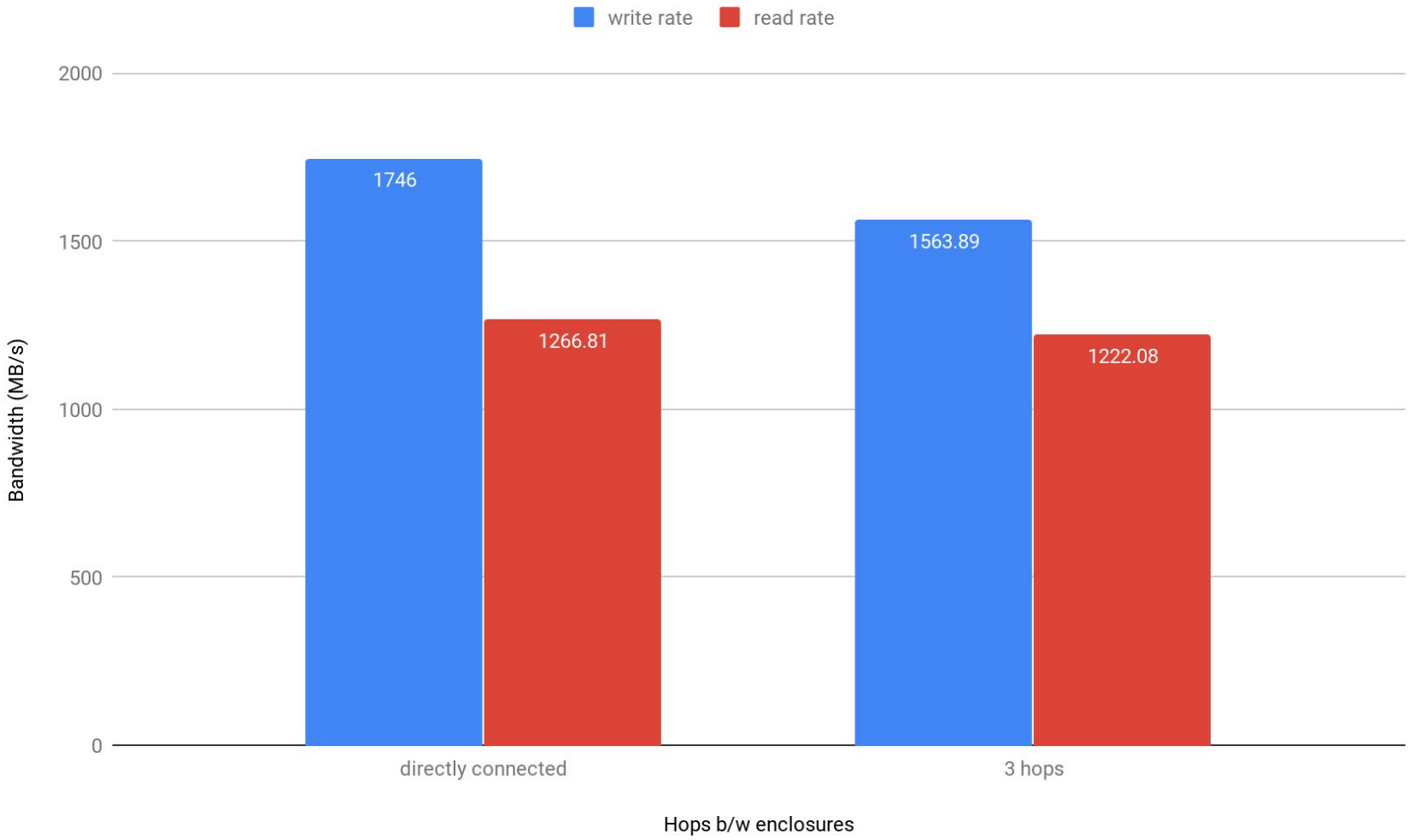
HPC - Lustre (S3D-IO)






write rate
read rate
2500
2000
1500
1000
500
0
Bandwidth (in MB/s)
cn098,cn353
cn062,cn131
cn161,cn183
cn003,cn352
cn138,cn161
cn153,cn344
cn144,cn184
cn041,cn349
cn041,cn349
cn037,cn097
cn097,cn350
cn058,cn156
cn017,cn098
cn037,cn098
cn017,cn098
cn156,cn349
cn180,cn352
cn058,cn156
cn036,cn112
cn001,cn330
cn139,cn353
cn048,cn071
Communication Nodes


write rate
read rate
2000
1500
1000
500
0
Bandwidth (MB/s)
cn161,cn164
cn349,cn359
cn347,cn351
cn350,cn352
cn351,cn358
cn335,cn347
cn345,cn350
cn351,cn358
cn352,cn356
cn348,cn349
cn344,cn355
cn350,cn351
cn350,cn351
cn350,cn351
cn347,cn351
cn344,cn359
cn347,cn359
cn350,cn351
cn352,cn358
cn353,cn356
cn344,cn355
cn347,cn359
Communicating Nodes


write rate
read rate
2000
1500
1000
500
0
Bandwidth (MB/s)
1746
1266.81
1563.89
1222.08
directly connected
3 hops
Hops b/w enclosures

Conclusion

- DXT and Darshan’s effect on I/O performance is more significant than we earlier believed it to be.

- I/O between nodes on different enclosures isn’t affected by number of hops between two nodes but by the network congestion.

Thank You!

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