Explanation of the third parallel WRF-HDF5 performance report at SDSC TG

According to the last report:

"From the dashed line (Fig. 1), we can see that with the increasing of number of nodes, the wall clock time of sequential IO case decreases. We know that the only difference between parallel HDF5 IO module and sequential NetCDF IO module is the time to output the data. So the unexpected increasing of wall clock time used parallel HDF5 when the number of node is greater than 8 is in parallel IO layer.

Three factors may affect the performance in parallel IO layer:

1. Parallel HDF5
2. MPI-IO
3. Configuration of GPFS

We cannot test the third factor. We are still in the process of investigating the first and the second factors."

We successfully ran MPI-IO performance tests at SDSC TG. h5perf, a benchmark to measure parallel IO performance inside HDF5 library, is used. We ran the same benchmark with the number of node at 1, 2, 4, 8, 16, 32 and 64. We made three runs for each case. There are five iterations at each run and we obtained the best throughput for each case. The result can be observed from Fig. 2. The file size for all runs is 4GB. We used throughput (MB/Sec) as the benchmark. We are testing MPI collective IO. Transfer buffer size is 1MB. We can observe that with the increasing of number of IO nodes, the accumulated MPI-IO throughput increases until the number of IO nodes becomes 8. After the number of IO nodes exceeds 8, the accumulated MPI-IO throughput becomes flat and worse when the number of IO node is beyond 32. This can partially explain why the wall-clock time of WRF-Parallel HDF5 increases after the number of IO node is greater than 8 from Fig. 1. The overhead caused by Parallel HDF5 can also make the performance deteriorated.

Fig. 1: WRF IO performance comparision Parallel HDF5 VS Sequential NetCDF
data size: 38.6 GB, 2 processors per node
SDSC Teragrid Linux 2.4, Intel Itanium 2, compiler icc, ifc 7.1,
mpich version: mpich-gm-1.2.5..10-intel-r1

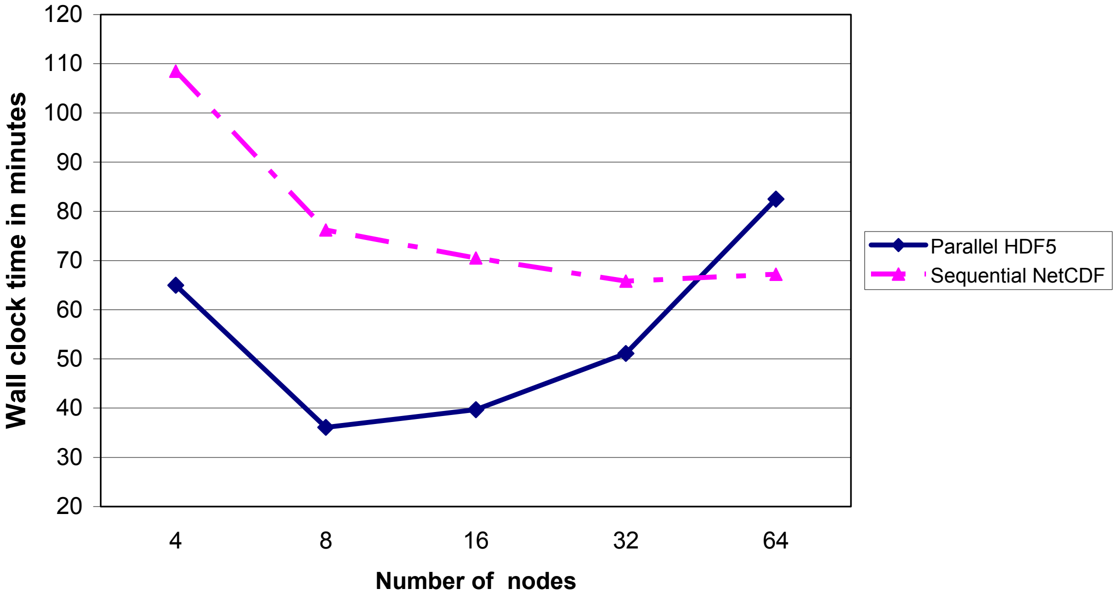
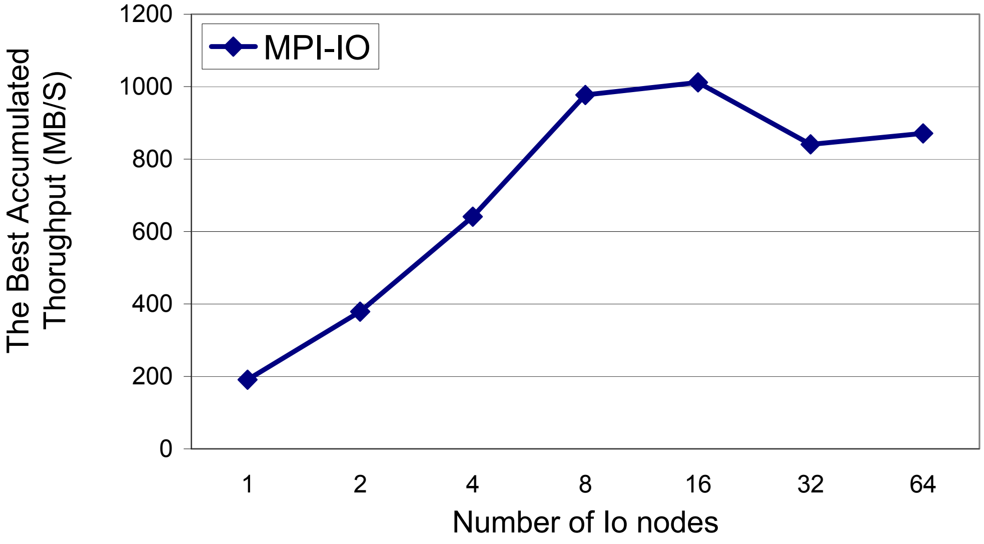


Fig.2: Performance of MPI-IO throughput at SDSC TG
Collective IO
File size = 4 GB
Number of iterations=5*3,
Transfer buffer size = 1MB
Block size = 128 KB


The Best Accumulated Thorughput (MB/S)
MPI-IO
Number of Io nodes