Performance Analysis and Optimization of Parallel I/O in a large scale groundwater application on the Cray XT5

Vamsi Sripathi[1], Glenn E. Hammond[2], G. (Kumar) Mahinthakumar[1], Richard T. Mills[3], Patrick H. Worley[3] and Peter C. Lichtner[4]

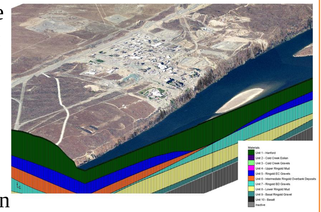
[1]{North Carolina State University}, [2]{Pacific Northwest National Laboratory}, [3]{Oak Ridge National Laboratory}, [4]{Los Alamos National Laboratory}

Introduction

• The Cray XT5 (JaguarPF) is ranked as the second fastest supercomputer according to the June 2009 compilation of the top500 list. It has a total of 150,152 AMD 2.3 GHz compute cores, giving a peak performance of 1.38 petaflop per second.

• JaguarPF uses the Lustre file system (lfs) to facilitate high bandwidth I/O. Lustre is an object based parallel file system that has 3 main components: Object Storage Servers (OSS's), Meta Data Servers (MDS) and File system clients. Each OSS can serve multiple Object Storage Targets (OST's) which act as I/O servers, the MDS manage the names and directories for the file system and the file system clients are the Cray XT5 compute nodes.

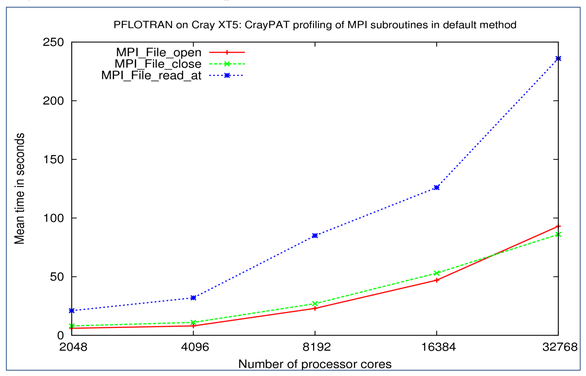
• PFLOTRAN is a highly scalable groundwater simulation code that solves multi-phase groundwater flow and multicomponent reactive transport in three-dimensional porous media. It uses the Portable, Extensible Toolkit for Scientific Computation (PETSc) for numerical equation solving and the parallel Hierarchical Data Format 5 (HDF5) library for I/O operations.

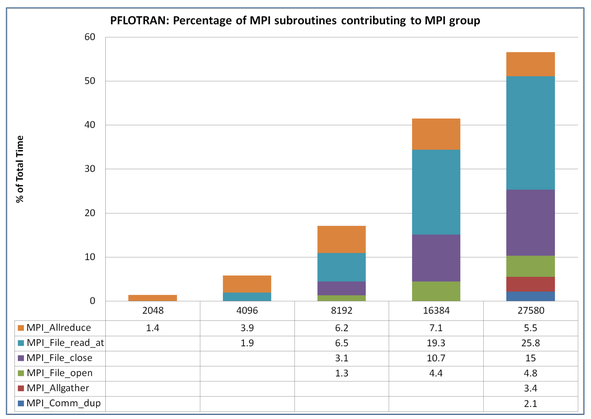


Performance Analysis

• Strong scaling study with a test problem (270 million degrees of freedom, 30 time steps) of PFLOTRAN from 2048 to 65,536 compute cores of Cray XT5.

• Cray Performance Analysis Tool (PAT) is used to obtain detailed analysis of PFLOTRAN up to 32,768 cores.





	2048	4096	8192	16384	27580
MPI_Allreduce	1.4	3.9	6.2	7.1	5.5
MPI_File_read_at		1.9	6.5	19.3	25.8
MPI_File_close			3.1	10.7	15
MPI_File_open			1.3	4.4	4.8
MPI_Allgather					3.4
MPI_Comm_dup					2.1

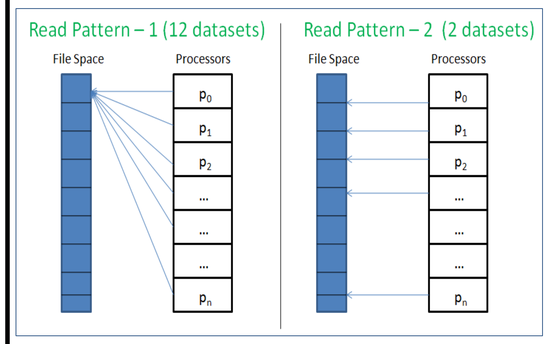
Default I/O Method

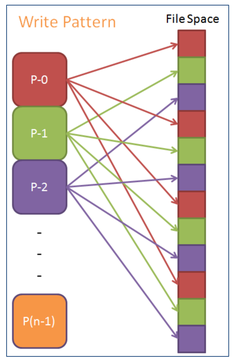
• All processes participate in parallel I/O operations involving a single HDF5 file.

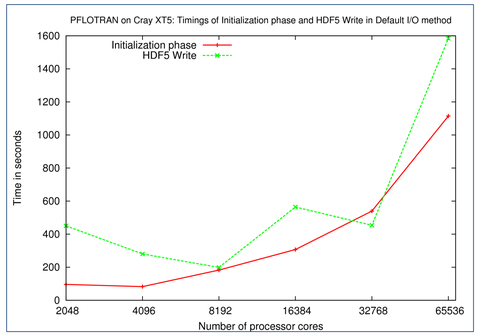
• Two different read access patterns in which all processes (1) read the same contiguous region and (2) read a different contiguous region of a dataset.

• Output is written to a single HDF5 file for every time step.

• Each process writes to a non-contiguous region of the file and all such writes are interleaved between processes.







Performance Bottlenecks

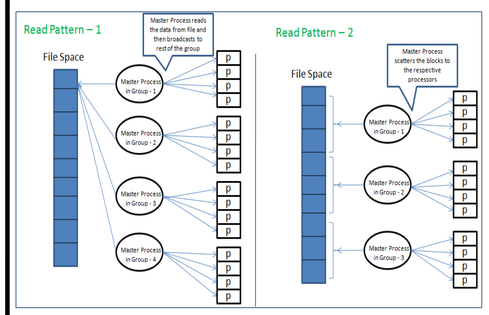
• Need to combine multiple small I/O disk requests into one large disk request to decrease the number of I/O disk accesses at higher processor counts.

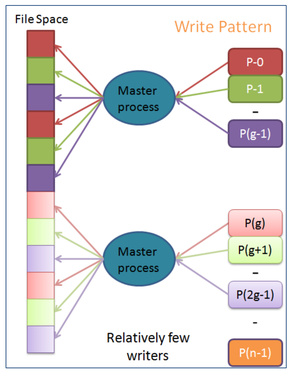
• All processors cannot afford to access the file simultaneously because of the limitation of the Lustre file system (1 Meta Data Server).

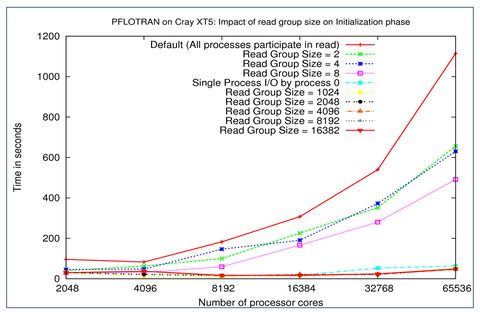
Performance Optimization

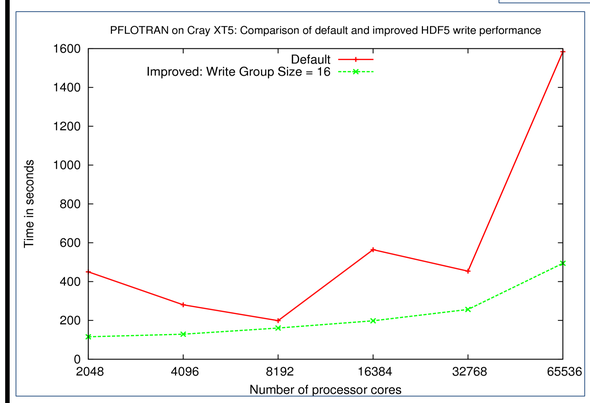
• Implemented a two-phase I/O (communication phase and I/O phase) approach at the application level by splitting the MPI global communicator into multiple sub-communicators.

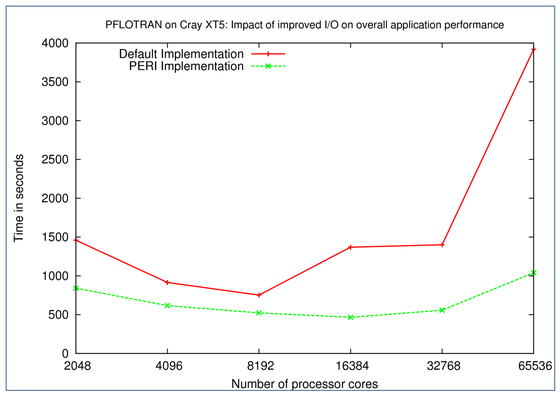
• The root process in each sub-communicator is responsible for performing the I/O operations for the entire group and then distributing the data to rest of the group.



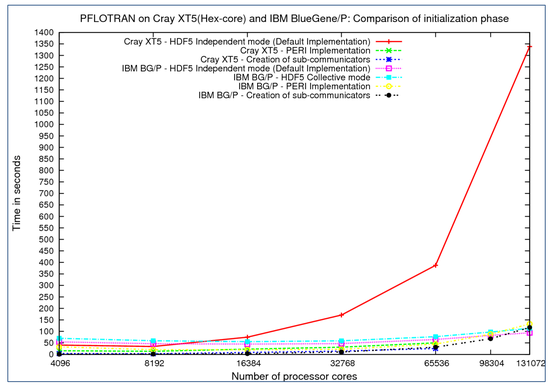


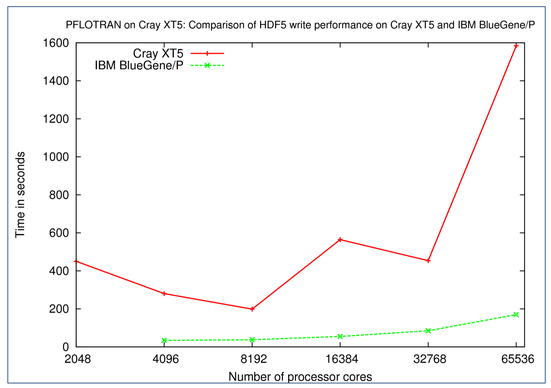






Comparison between Cray XT5 and IBM BlueGene/P





Conclusions

• Performance improvement of 25X in initialization phase and 3X in HDF5 file write at 65,536 cores on Cray XT5.

• Overall application improvement of 5X at 65,536 cores of Cray XT5.

Acknowledgements

• This work is part of Performance Engineering Research Institute (PERI) and is funded by the U.S. Department of Energy Office of Science, as part of its second Scientific Discovery through Advanced Computing (SciDAC-2) research program.

• This research used resources of the National Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-0OR22725.

• This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357.