A Lightweight, High-performance I/O Management Package for Data-intensive Computing

DOE Early Career Principal Investigator Program 2005-2008
PI: *Jun Wang*
Computer Science and Engineering Department
University of Nebraska Lincoln
DOE Collaborators: *Rob Ross* and *Rajeev Thakur* from Argonne National Lab

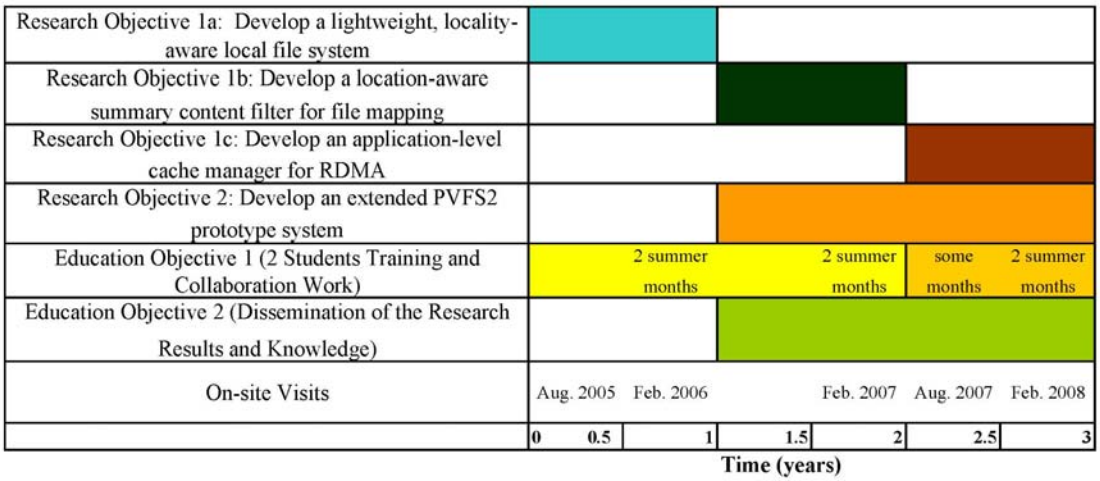
File storage systems are playing an increasingly important role in high-performance computing as the performance gap between CPU and disk increases. Researchers have proposed several solutions to make I/O faster in high-performance computing, such as MPI-IO. Recently, the Department of Energy (DOE) and its collaborators have been developing new parallel file systems such as the parallel virtual file system—PVFS2, Lustre and IBM GPFS.

The use of SciDAC and other emerging data-intensive applications in the Department of Energy, such as high-energy physics and climate modeling, bring new technology challenges today. It could take a long time to develop an entire system from scratch. Solutions will have to be built as *extensions* to existing systems. If new portable, customized software components are plugged into these systems, better sustained high I/O performance and higher scalability will be achieved, and the development cycle of next-generation of parallel file systems will be shortened.

The *overall research objective* of this ECPI development plan aims to develop a lightweight, customized, high-performance I/O management package named LightI/O to extend and leverage current parallel file systems used by DOE. We will develop three novel components in LightI/O and prototype them into PVFS2 to deliver better performance and reliability to users. These three components are as follows:

1: A lightweight, locality-aware, segment-structured local file system (LL-SFS) to effectively and efficiently handle both small and large file I/Os;
2: A scalable distributed file mapping management component employing Location-aware Summary Content Filters or Hierarchical Bloom filter Arrays to balance the scalability and high-performance;
3: An application-level, RDMA-based cache manager (CacheRDMA) to facilitate the deployment of the intra-cluster RDMA data communication scheme;

The phased timeline for the research activities over three years is illustrated as below:



During the course of the project, we will deliver a prototype version of LightI/O to the high-performance cluster computing community. We will build, simulate and prototype three LightI/O components into PVFS2, and evaluate the resultant prototype—extended PVFS2 system on data-intensive applications. Furthermore, we will deploy the product on large-scale clusters such as ANL's Chiba City and UNL's PrairieFire, to gauge its effectiveness at this scale. The broader impact of this project is expected to advance the state-of-the-art of the development of high-performance file storage system for high end computing, be training and career preparation of undergraduate and graduate students, dissemination of research knowledge and the enhancement of research and education infrastructure at DOE, UNL and the State of Nebraska.