

PVFS2: The Parallel File System

Scalability and high performance, by design



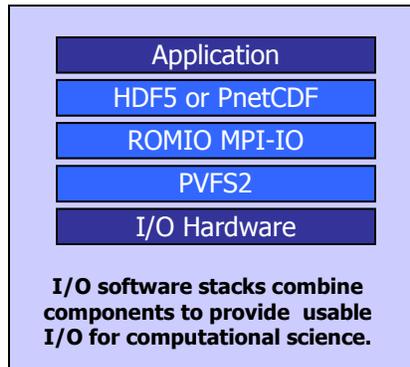
PVFS2 at a Glance

- **Scalable architecture** – no single bottlenecks in I/O path, optional distributed metadata, and no locking subsystem to cause contention between clients
- **Optimized MPI-IO support** – tuned ROMIO MPI-IO implementation leverages unique PVFS2 scalability features
- **Hardware independence** – operates on IA32, IA64, Opteron, PowerPC, and Alpha based Linux systems with TCP/IP, Myrinet, or InfiniBand networks and local or shared storage
- **Fault tolerance** – stateless system compatible with commodity HA software such as the heartbeat package for configuring fault tolerant installations
- **Proven development team** – brings together experts in parallel I/O, message passing, and networking software
- **Freely available, open source** – all collaborators contribute to a single, publicly accessible CVS tree. Code is GPL/LGPL licensed

PVFS2 brings state-of-the-art parallel I/O concepts to production parallel systems.

Developed by a multi-institution team of parallel I/O and networking experts, PVFS2 embodies the expertise of designers who have worked for nearly a decade in the field of parallel I/O.

Rather than compromising performance at scale to create a general-purpose solution, the designers carefully focused their efforts on creating a solution that provides scalability to tens of thousands of clients and hundreds or thousands of servers for computational science applications. At the same time, familiar UNIX tools such as cp and ls work seamlessly with PVFS2.



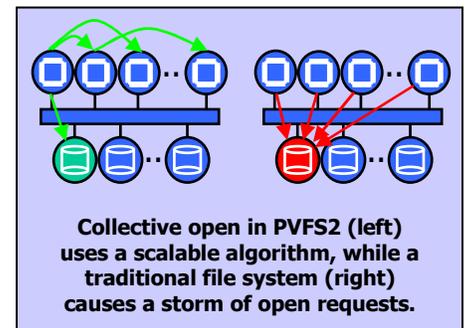
I/O Software Stacks

Parallel file systems are one component in a collection of software layers that provide the parallel I/O environment used by scientific applications. This collection includes high-level I/O libraries such as HDF5 and Parallel netCDF and I/O middleware such as the ROMIO MPI-IO implementation.

These additional software layers provide opportunities for optimizations that cannot be implemented on top of the traditional file system interface. For this reason PVFS2 breaks with the traditional file system interface.

Built for Science

PVFS2 provides a rich, stateless file system interface that is tailored to best fit with the I/O software layers above the file system, such as the MPI-IO layer. This leads to optimizations for a wide variety of access types, from collective I/O down to the process of opening files, that are simply not possible with traditional APIs.



For example, when a file is collectively opened in PVFS2 through MPI-IO, two steps are performed. First, a single process obtains a handle allowing access to the file. Second, this process broadcasts the handle to all other processes using a scalable algorithm.

In traditional file systems, the API forces each process to interact with the file system. This can cause an overwhelming *storm* of requests to the file system, temporarily stalling file system operations.

Likewise, the PVFS2 structured I/O API allows for key optimizations in reads and writes. This capability eliminates both the overhead of excess data transfer and the added latency of multiple file system operations that are seen when using traditional APIs to perform I/O for computational science applications.

PVFS2 Details and Requirements

Development and Support

PVFS2 is a joint project between Argonne National Laboratory and Clemson University with close collaboration from the Ohio Supercomputer Center, all of which contribute directly to the core PVFS2 implementation and help with support issues. This group has a proven track record of delivering and supporting quality software over the past six years.

Additional collaborations with Northwestern University, Ohio State University, and Pennsylvania State University focus primarily on future directions for the project, including aggressive server and client-side caching schemes and mechanisms for enhanced consistency semantics with low overhead.

PVFS2 user-space code is released under the LGPL license to allow for linking to libraries under alternative licenses, while kernel code is released under the GPL. All collaborators contribute to a single CVS source tree made available via anonymous access to anyone. This open access guarantees that users can obtain the most up-to-date code at any time.

Support is via mailing lists and IRC, where users and administrators have direct access to the developers.

System Requirements

Hardware Platforms

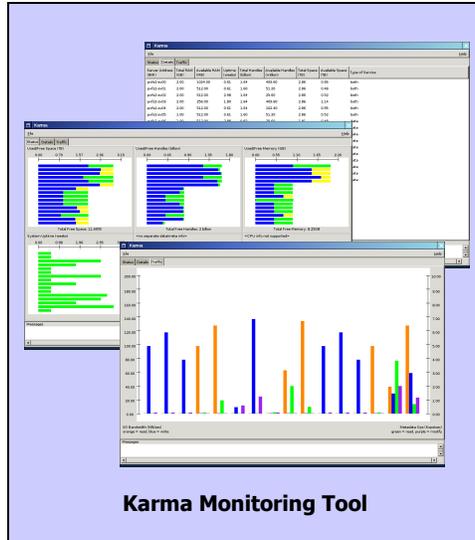
Any mix of IA-32, IA-64, Opteron, PowerPC, and Alpha systems

Operating Systems

Linux 2.4.19+ and Linux 2.6.0+

Networking

TCP/IP, Myrinet, and InfiniBand networks, including support for multi-homed servers



POSIX Considered Harmful

PVFS2 explicitly rejects the POSIX I/O consistency semantics - a standard that forces sequential consistency of file system operations and, in doing so, severely limits the design options available to parallel file system architects.

Instead PVFS2 provides an alternative semantic, termed *nonconflicting writes*, that best matches the needs of computational science applications while retaining the most opportunities for parallelism. A similar semantic has been used successfully in PVFS1 for the past five years. As a side-effect of this semantic, PVFS2 does not need a locking subsystem, eliminating one of the most complicated and stateful components of a file system.

PVFS2 thus can offer higher performance, greater scalability, greater robustness, and a simpler design than would be possible otherwise.

File System Management

PVFS2 includes many tools to aid in management of the file system:

- karma, a GUI monitoring tool
- command line tools for monitoring, common maintenance operations, and generating configuration files
- pvfs2-fsck, a tool for salvaging damaged file systems
- documentation on configuration, adding servers, and setting up high availability configurations

PVFS2 uses simple, readable, text-based configuration files, and no additional system services are necessary for managing configuration.

For More Information...

PVFS2 Web Site

<http://www.pvfs.org/pvfs2/>

PVFS2 Software Downloads

<ftp://ftp.parl.clemson.edu/pub/pvfs2/>

PVFS2 Mailing Lists

<http://www.pvfs.org/pvfs2/lists.html>

PVFS2 IRC (Internet Relay Chat)

#pvfs2 on irc.freenode.net

E-Mail Contacts

Rob Ross <rross@mcs.anl.gov>

Walt Ligon <>walt@clemson.edu>

Pete Wyckoff <pw@osc.edu>

Phil Carns <pcarns@parl.clemson.edu>

Neill Miller <neillm@mcs.anl.gov>

Rob Latham <robl@mcs.anl.gov>