DAFS Storage for High Performance Computing using MPI-I/O: Design and Experience

Vijay Velusamy, Anthony Skjellum
MPI Software Technology, Inc.
Email: {vijay, tony}@mpi-softtech.com

Arkady Kanevsky*†, Peter Corbett
Network Appliance
Phone: (781) 768-5395
Fax: (781) 895-1195
Email: {arkady, pcorbett}@netapp.com

A key goal of this effort is to demonstrate and develop heterogeneous and distributed computing technologies that are applicable to DoD and scientific communities, while maintaining and benefiting from industry standards that could be applied to high performance computing.

High performance computing systems based on clusters of compute nodes, connected via a high speed interconnect, are becoming popular for large-scale parallel applications, forming a highly scalable infrastructure. Most large-scale scientific applications are highly I/O-centric and have a tremendous need for a similarly scalable file I/O subsystem. DAFS is a well-defined high-performance protocol for file access across a network as well as set of APIs, uDAFS, for user level OS-bypassing application programming, designed to take advantage of RDMA-based transports, such as Virtual Interface Architecture (VIA), InfiniBand Architecture[1], and iWARP.

Although DAFS was not originally designed for parallel I/O, this effort aims to demonstrate that DAFS can easily used directly or extended for the creation of a parallel file system. Most parallel file systems are designed to use metadata for parallel files, stored as objects (often themselves files) separate from the data of those files, and the data of the individual files is striped across a number of different server nodes. It is envisioned that this effort would demonstrate the adoption of DAFS and its parallel features to high performance computing environments, strengthening the technology base, and providing an opportunity for the adoption of widely used technology to the largest high performance computing programs such as ASCI.

This effort utilizes ChaMPIon/Pro, an efficient commercial implementation of the MPI-1.2 standard for message passing interfaces [3], to demonstrate the applicability of DAFS for scalable I/O. The MPI I/O implementation in ChaMPIon/Pro is designed to support both parallelism, and portability [2]. Parallelism is achieved in the MPI I/O layer, by implementing the MPI I/O APIs, while the BAFS layer provides portability (Figure 1). BAFS consists of a set of APIs that provide the necessary functionality for parallel I/O. BAFS provides a thin abstract I/O interface between MPI I/O and DAFS. It provides a non-collective I/O interface that function independently for each MPI process created. MPI I/O operates on the communicator level, involving communication between multiple processes. The MPI I/O layer manages collective I/O and shared file pointers. This layer maintains data consistency across different processes and delivers file atomicity semantics.

The BAFS implementation benefits from early binding and persistency for repeated file access patterns that use the same data structure. In this case, the data structure is allocated and initialized only once. The user can then call the non-blocking data access routines to fetch the data repeatedly making use of the same data structure.

---

* Corresponding Author
† Presenting Author

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
20 AUG 2004	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
DAFS Storage for High Performance Computing using MPI-I/O: Design and Experience	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **MPI Software Technology, Inc.**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001694, HPEC-6-Vol 1 ESC-TR-2003-081; High Performance Embedded Computing (HPEC) Workshop (7th)., The original document contains color images.

14. ABSTRACT

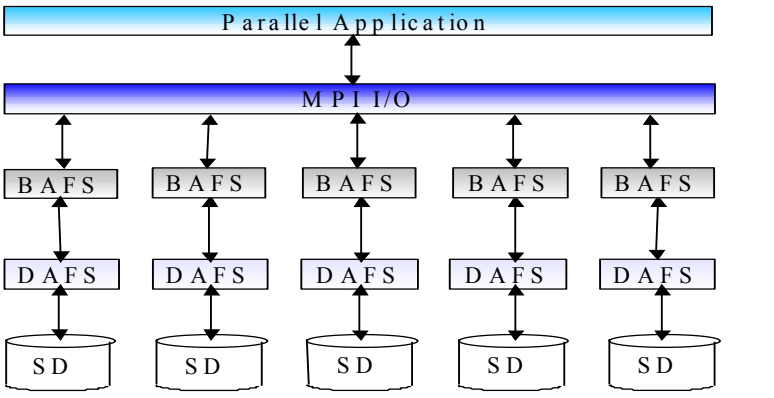
15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**14**	

In parallel file systems, files are generally striped across the server nodes, with each server node managing its own pool of disk space. File data is distributed by dividing the file into subfiles, referred to as cells [4]. Cells are distributed among the server nodes, so that each cell resides entirely on one server node, and zero, one or more cells of a file may reside on each of the server nodes in the parallel file server. The cell usually has an inode, including a blocklist, and the server node allocates space to each cell that it owns from its own pool of disk. In this effort for adapting DAFS for parallel I/O, a cell specific decomposed view of the parallel file is presented, so that the I/O library can perform the mapping of data to cells directly. Also file metadata servers are not separated from file data servers. In other words, the external metadata of a file, that is visible to the clients, including access control lists, cells locations, file attributes, are stored in a single metadata object, called metanode of the file. The details of the metadata need not be known to any application that needs to access the file.

The design of MPI I/O for DAFS allows the cache coherence and token/lock based read and write access control that are generally present in cluster file systems to be eliminated. There is no contention among the nodes of a parallel application for temporary access rights to data once the parallel program gains access rights to the entire file. This avoids any bottlenecks to restrict data flow, enhances the scalability of the system. Since DAFS is designed to benefit from high bandwidth low-latency RDMA access to file data, and because it can offload the client almost completely from performing file system and I/O tasks, the client CPU utilization for I/O is reduced tremendously, allowing the overlap of computation and I/O and better utilization of these extra CPU cycles.

It is expected that performance numbers for this effort would be available in September.



References:

[1] DAFS Collaborative, "Direct Access File System API Specification v1.0," http://www.dafscollaborative.org

[2] Kumaran Rajaram, Anthony Skjellum, Rossen P. Dimitrov, Purushotham V. Bangalore, Vijay Velusamy, and David Leimbach, "Design, Implementation, and Evaluation of a High Performance Portable Implementation of the MPI-2 I/O Standard API," submitted to Parallel Computing, November 2002.

[3] MPI Forum, "**MPI** - The Message Passing Interface Standard,"

http://www-unix.mcs.anl.gov/**mpi**/

[4] Peter Corbett, "DAFS Extensions for Parallelism," Network Appliance, August 2002.

DAFS Storage for High Performance Computing using MPI-I/O: Design and Experience

Arkady Kanevsky & Peter Corbett
Network Appliance

Vijay Velusamy & Anthony Skjellum
MPI Software Technology

Why DAFS?

- **DAFS for Parallel I/O**
  - **Industry Storage Standards for HPC Community**
  - **Performance, Performance, Performance**
  - **Reduce TCO (total cost of ownership)**
- **Performance of RDMA based File System**
  - **Bandwidth ↑**
  - **Latency ↓**
  - **CPU overhead ↓**
- **Transport independence**
  - **Virtual Interface (VI) Architecture**
  - **InfiniBand Architecture**
  - **iWARP**
- **Network Appliance filer as DAFS server for transport independence, performance and multi-protocol support**

MPI Software Technology

CHAMPION/PRO

NetworkAppliance®

What is DAFS?

- Direct Access File System protocol
- A file access protocol designed specifically for high-performance data center file sharing
  - Optimized for high performance
  - Semantics for clustered file sharing environment
- A fundamentally new way for high-performance and cluster applications to access file storage
  - Provides direct application access to transport resources
  - Avoids Operating System overhead

MPI Software Technology

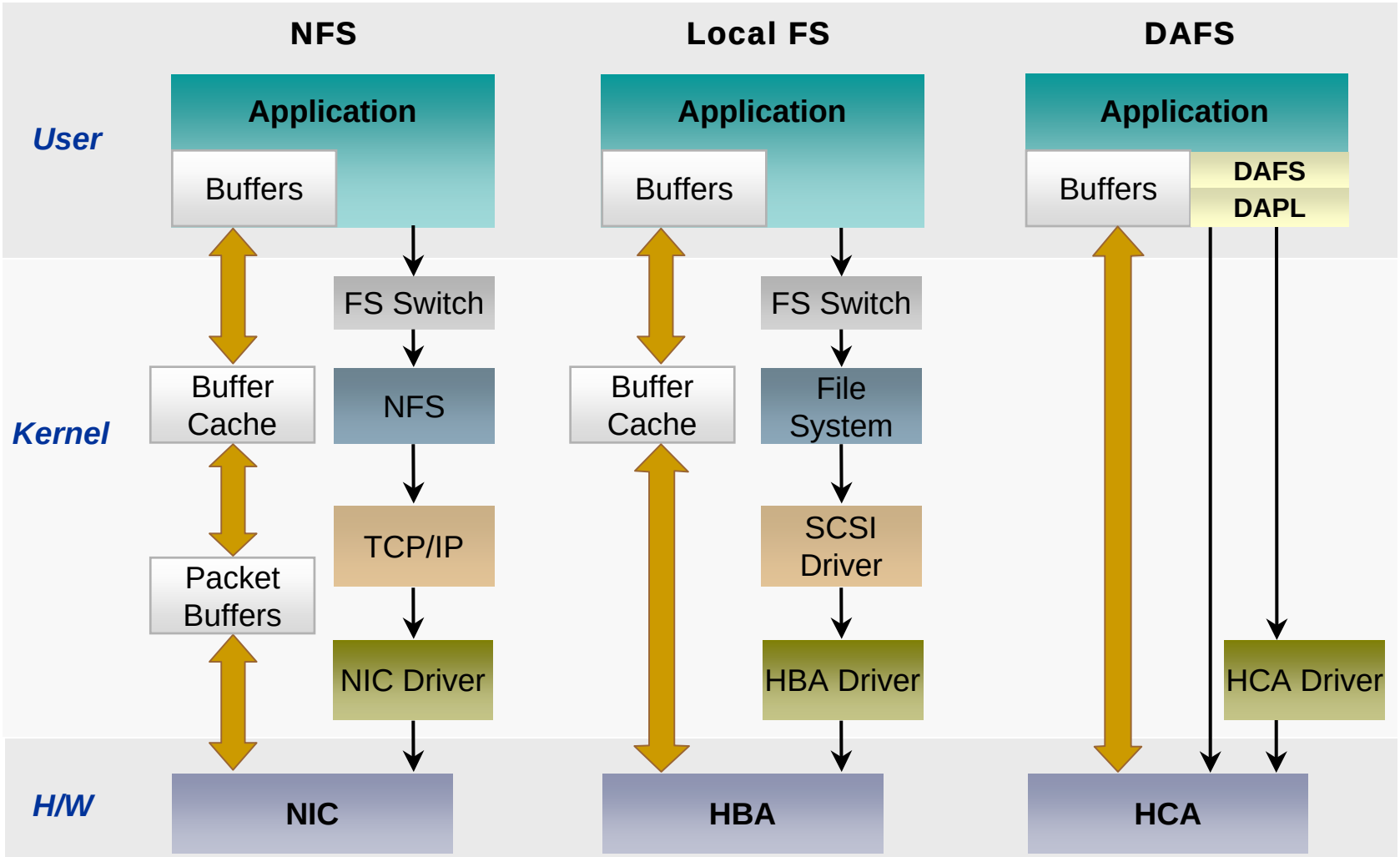
CHAMPION/PRO

NetworkAppliance®

What Does DAFS Do?

- File access protocol providing all the features of NFS v3
- Includes NFS v4 features
  - File locking, CIFS sharing, security, etc
- Adds data sharing features for clusters
  - Clustered apps
  - Graceful fail-over of clustered file servers
  - High volume, optimized I/O applications
  - Efficient multi-client sharing
- Designed for Direct Access (RDMA) Transports
  - Optimal use of transport capabilities
  - Transport independent

File Access Methods



MPI Software Technology

CHAMPION/PRO

NetworkAppliance®

Direct Access Performance Benefits

- No data packet fragmentation or reassembly
  - Benefit similar to IP Jumbo Frames, but with larger packets
    - less transmission overhead, fewer interrupts
    - no ordering and space management issues
    - no data copying to recreate contiguous buffers
- No realignment of data copies
  - Protocol headers and data buffers transmitted separately
  - Allows data alignment to be preserved
- No user/kernel boundary crossing
  - Less system call overhead
- No user/kernel data copies
  - Data transferred directly to application buffers

DAFS Performance

	App Server CPU µsec / op
Direct-attached (FC) storage w/ volume manager	113
Direct-attached (FC) storage w/ local FS (ufs)	89
Raw access to direct-attached (FC) storage	76
DAFS kernel device driver (w/ VI/IP HBA)	70
User-level DAFS client (w/ VI/IP HBA)	28
User-level DAFS client (w/ 4X IB HCA) - estimated	<20

- Sun E3500 (400MHz) w/ Solaris 2.8
- OLTP workload – 66% reads
- 4kB transfers; async I/O

MPI Software Technology

CHAMPION/PRO

NetworkAppliance®

Why MPI-IO?

- Parallel File System API
- Combined API for I/O and Communication
- File I/O and direct storage semantic support
- File Info for file “partitioning”
- Memory Registration for both I/O and Communication
- ChaMPIon/Pro for parallelism and portability
  - first commercial MPI-2.1 version
  - Scaling to thousands and tens of thousands of processors and beyond
  - Multi-device support (including InfiniBand Architecture)
  - Topology awareness
  - Thread safety
  - Optimized collective operations
  - Efficient memory (and NIC resource) usage
  - Integration with debuggers and profilers
  - Optimized MPI-IO
    - Early binding
    - Persistency
    - Layering blocking MPI calls on asynchronous transport operations

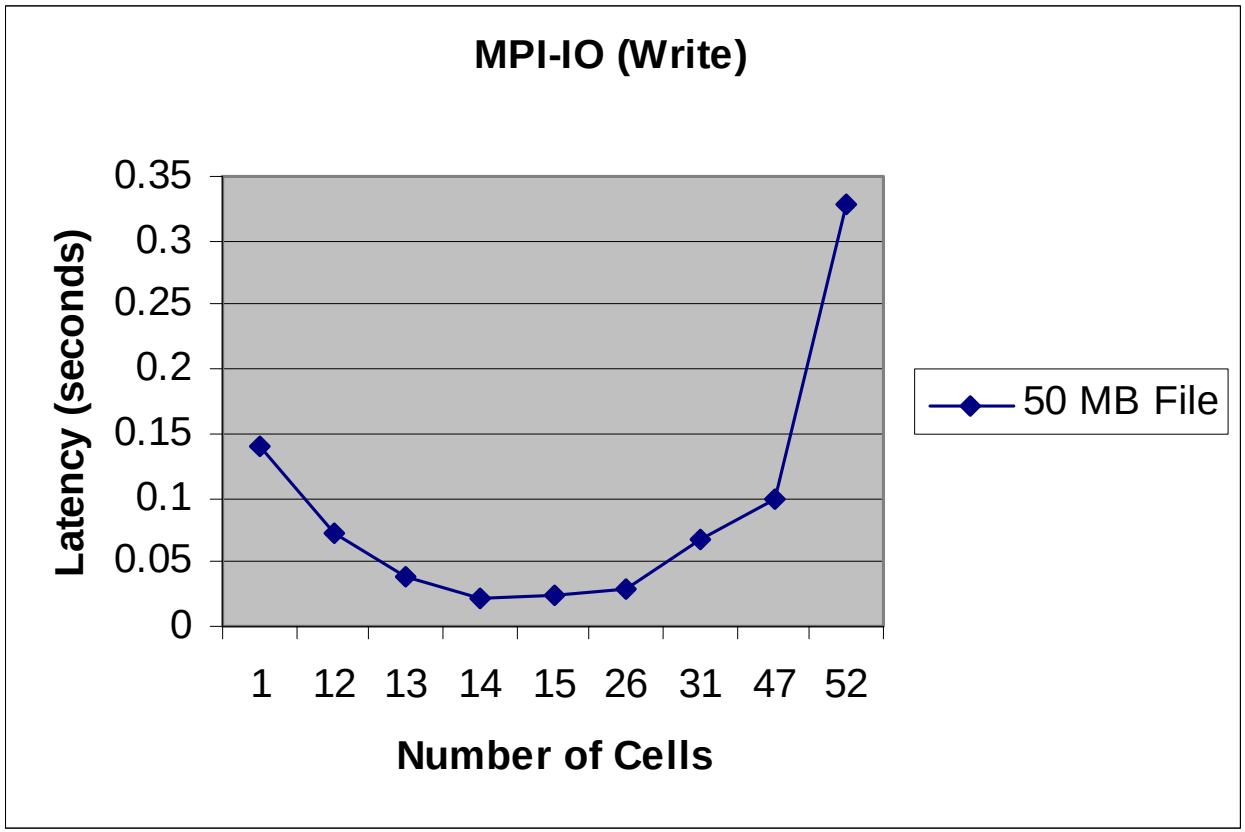
Design overview

- MPI-IO partitions user file according to MPI_FILE_INFO into cells
- Uses uDAFS API on the client to reduce CPU overhead and improve other performance measures
- Each cell is a separate file stored on DAFS server (NetApp filer)
  - Distribute cells across multiple DAFS servers
  - Multiple cells can be stored on the same DAFS server
- Metadata per file
  - Metadata is stored as a file and accessed on File Open, Close, or attributes changes
- MPI file accesses: Read, Write - directly accesses cells with no or minimal conflict.
  - No Metadata accesses
  - DAFS supports locking for conflict resolution.

Metadata & File Virtualization

- **Metadata contains:**
  - File ACL
  - File attributes
  - Cell list
    - Cell number
    - Cell file name
- **Cell file convention**
  - “file_name”_ “unique_identifier”_ “cell_number”
- **Metadata Files**
    - Separate volume on DAFS server
    - Volume mirrored between 2 DAFS servers
- **Cells are in a separate volume on each DAFS server**
- **Security**
  - Metadata ACL determines access to all of its cells

Early Experience - I



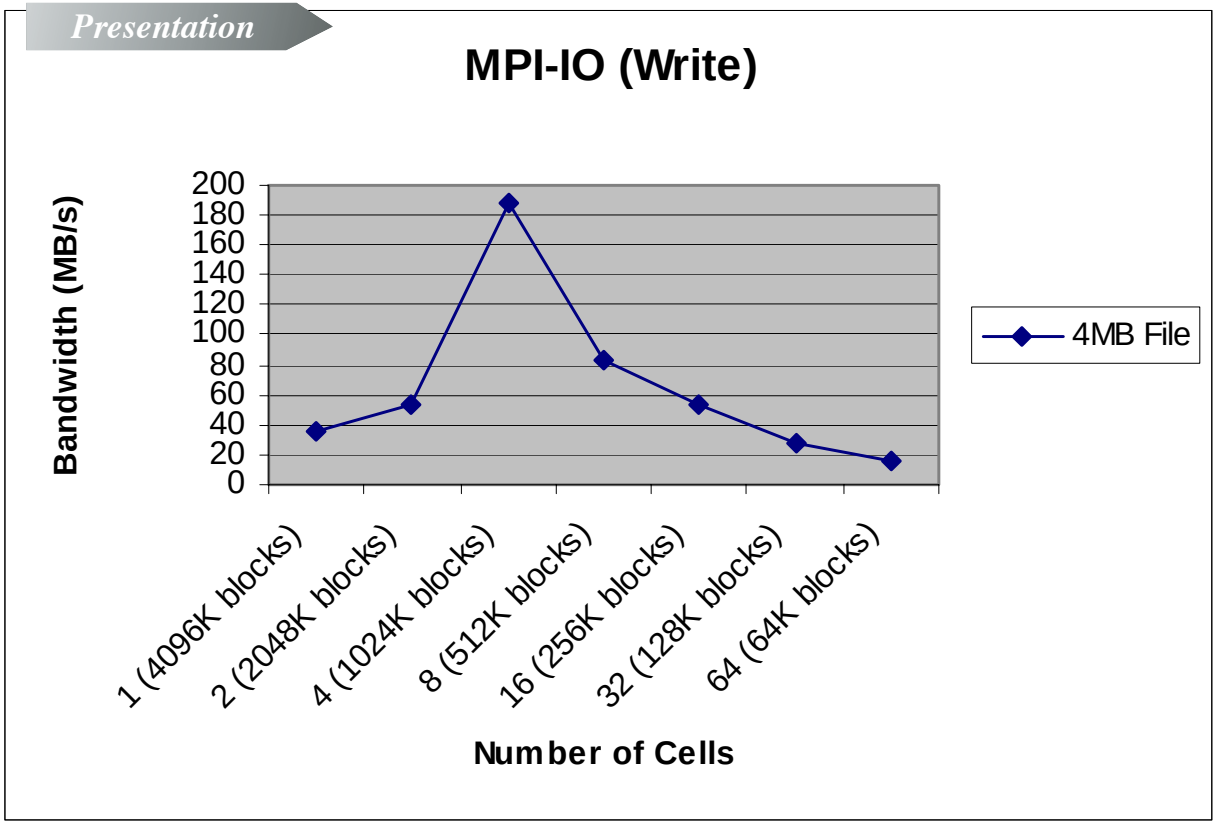
- 2 DAFS servers, 2 MPI client
- Clients on Sun Ultra 30 with 296 MHz processors
- Network – 1Gb VI-IP

MPI Software Technology

CHAMPION/PRO

NetworkAppliance®

Early Experience - II



- 2 DAFS servers, 2 MPI client
- Clients on Sun Ultra 30 with 296 MHz processors
- Network – 1Gb VI-IP

MPI Software Technology

NetworkAppliance®