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

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**

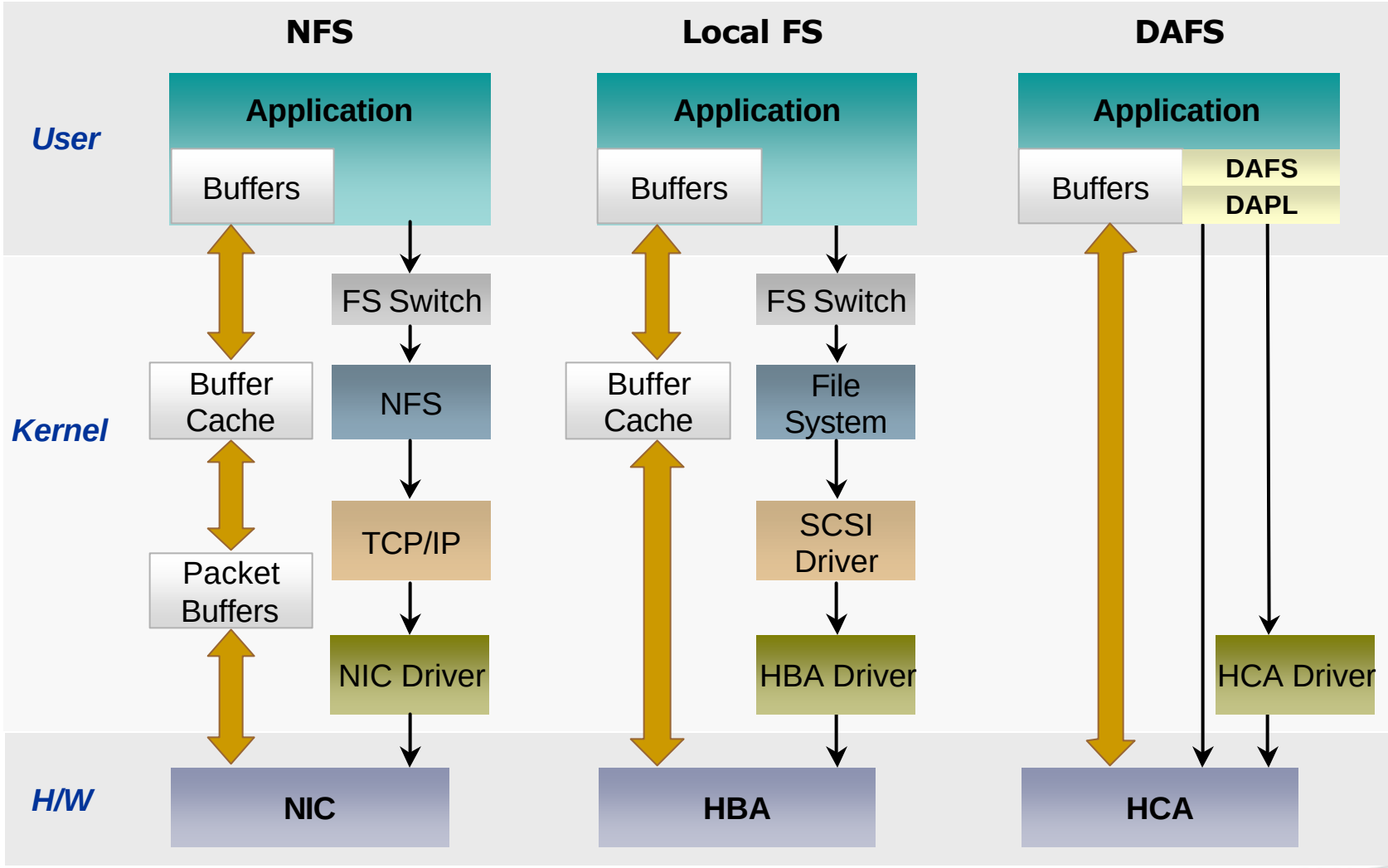
MPI Software Technology

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

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 ?sec/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

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.

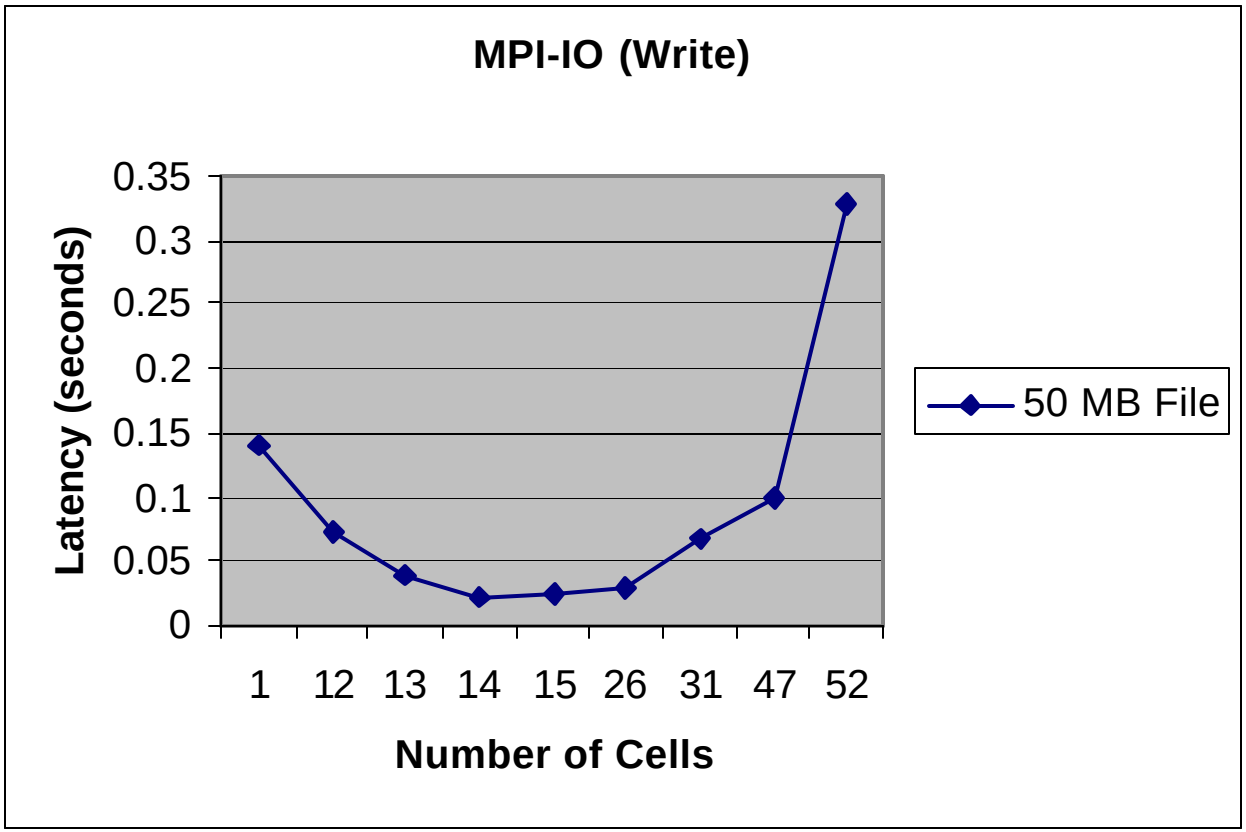
Metadata & File Virtualization

- **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

MPI Software Technology

NetworkAppliance®

Early Experience - I

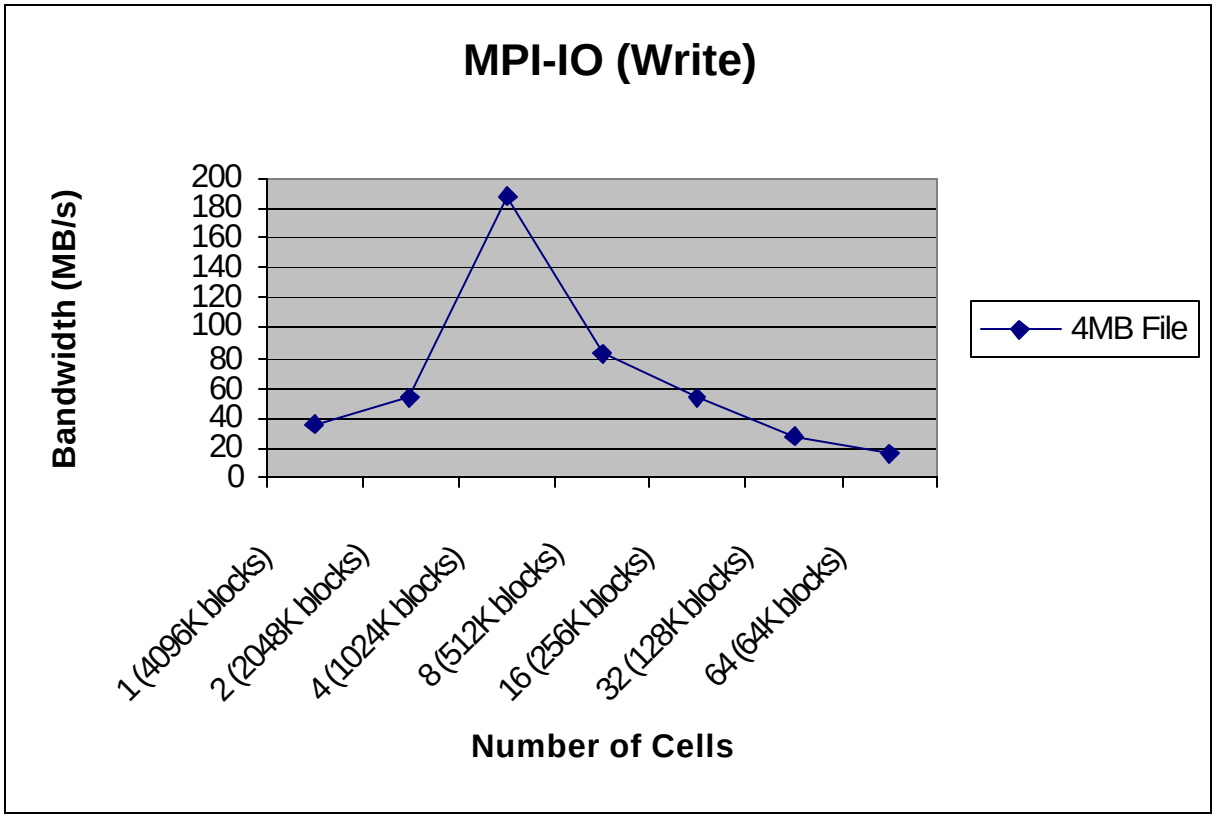


- 2 DAFS servers, 2 MPI client
- Clients on Sun Ultra 30 with 296 MHz processors
- Network – 1Gb VI-IP

MPI Software Technology

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®