Evaluating the Influence of File System Interfaces and Semantics on I/O Throughput in High Performance Computing

Christina Janssen, Michael Kuhn, Thomas Ludwig

Scientific Computing
Department of Informatics
University of Hamburg

2012-02-17

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

informatik die zukunft

Introduction & Motivation

- I/O performance is influenced by several factors
  - File systems
  - File system's interface & semantics
  - Application's implementation
- Focus is on I/O interfaces and semantics
- Evaluation of performance
- Suggestion of a new file system prototype for I/O optimization

File Systems

- Might limit performance of the accessing application
- Often specialized on recurring access patterns
- Important factors to consider for best performance:
  - Concurrent file accesses
  - Keeping data consistent
  - Managing metadata efficiently

I/O Interfaces

- Connect file system and accessing application
- Might be a bottleneck with regards to performance
- Interfaces examined:
  - POSIX
  - MPI-I/O
  - ADIOS

I/O Interfaces

- POSIX:
  - De-facto standard interface for file systems
  - Guarantees portability to most architectures
  - Developed for traditional local file systems
- MPI-I/O:
  - Part of the MPI-2 specification
  - Designed for high performance parallel I/O
  - ROMIO provides portability and high performance
- ADIOS (ADaptable I/O System):
  - I/O configuration in a separate XML file
  - Supports different backends, like MPI-I/O and POSIX
  - Can drastically increase performance, especially with iterative applications

Semantics

- Define the file system's behaviour
- Strict vs. relaxed file access semantics
  - Decision depends on the access pattern of the accessing application
- Have a strong impact on performance

Evaluation of Interfaces and File Systems

- Interfaces: POSIX, MPI-I/O, ADIOS
- File Systems: NFS, PVFS
- Iterative benchmark, run on a cluster consisting of one master and ten compute nodes
- I/O by writing a checkpoint (20.71 MB) every other iteration
- Total I/O size for each test case: 5177.5 MB after 500 iterations

Evaluation Results

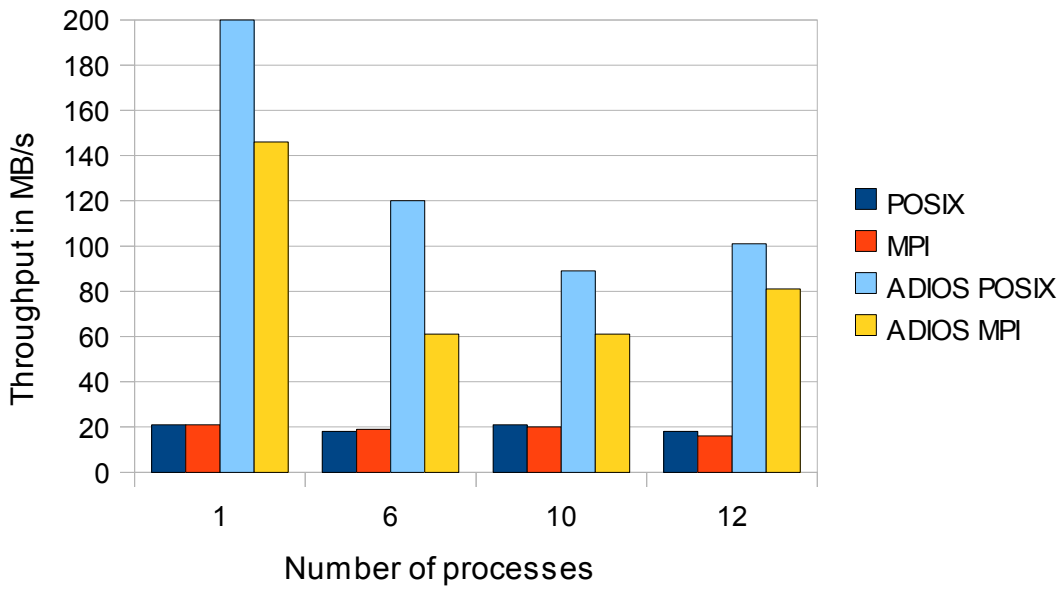

Figure: Performance of NFS on one node

- Not much difference between POSIX and MPI
- ADIOS performs best in all cases
- ADIOS with POSIX even better than with MPI

Evaluation Results

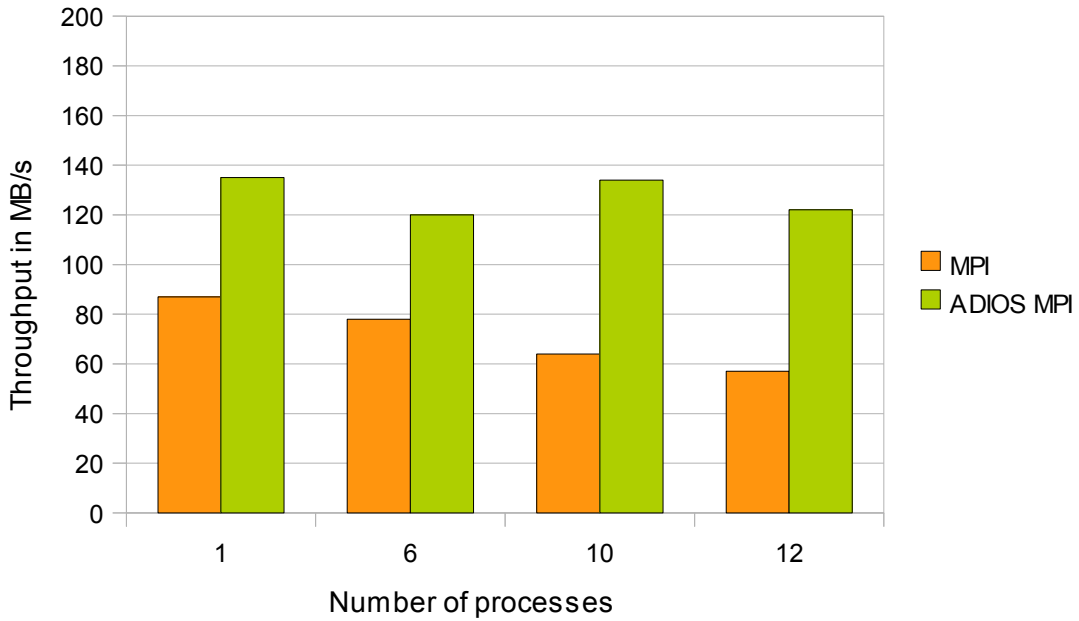

Figure: Performance of PVFS on one node

- MPI's performance decreases with increasing number of processes
- ADIOS constantly high performance

Evaluation Results

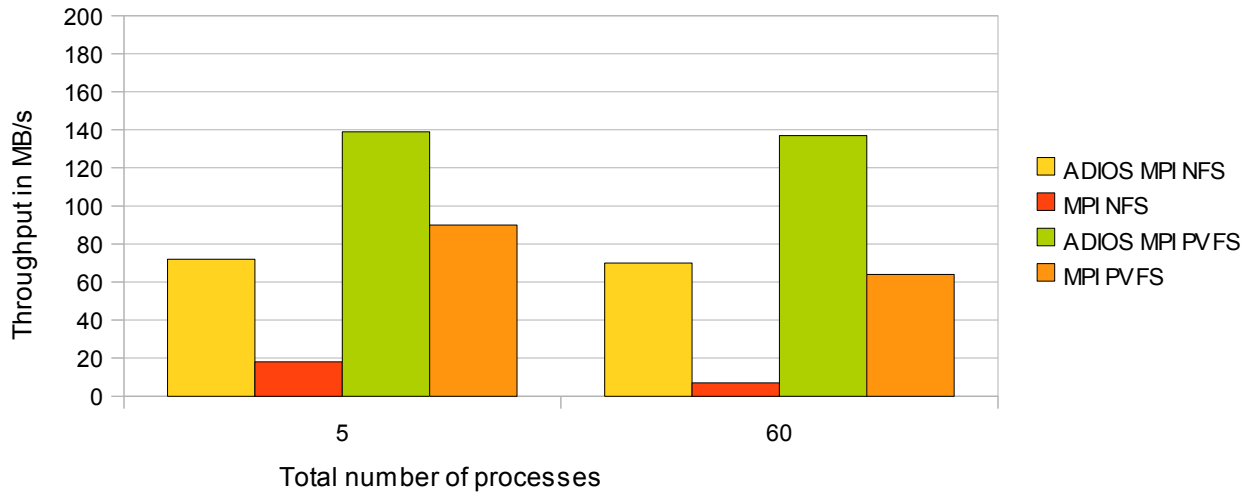

Figure: Performances of NFS and PVFS on five nodes

- PVFS overall better performance
- NFS cannot use MPI's optimization potential
- File system's and interface's influence on performance clearly visible

Work in Progress

- Developing a new file system prototype
  - Interface specifically suited for HPC requirements
  - Very limited file system hierarchy
  - Reduced set of metadata
  - File access via *operations*
  - Possibility to specify the semantics at runtime on a per-operation basis
    - For example: consistency, persistency, concurrency and security