

# THE **EXPAND** PARALLEL FILE SYSTEM

A FILE SYSTEM FOR CLUSTER AND GRID COMPUTING



José Daniel García Sánchez  
ARCOS Group – University Carlos III of Madrid



# Contents

2

- **The ARCOS Group.**
- Expand motivation.
- Expand design.
- Expand evaluation.
- Conclusions.
- Ongoing Work.

# University Carlos III of Madrid

3

- Founded in 1989
- Three faculties:
  - ▣ Faculty of Social Sciences and Law.
  - ▣ Faculty of Humanities, Documentation and Communication.
  - ▣ **Higher Technical School.**



The Expand Parallel File System

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 University of Modena

# The ARCOS Group

4

- The **Computer Architecture, Communications and Systems Group** is part of the Department of Computer Science.
  
- 20 full time members
  - ▣ 9 PhD's (2 full professors + 4 associate professors + 3 visiting professors).
  - ▣ 11 PhD students

# Research lines

5

- Data management on Grid environments.
- Parallel file systems.
- Optimization of irregular applications.
- OS for Wireless Sensor Networks.
- Real-time systems.

# Some products

6

- Expand: A parallel file system for cluster and grid environment.
- WinPFS: Windows Parallel File System.
- MiMPI: MPI implementation for heterogeneous cluster environments

# Contents

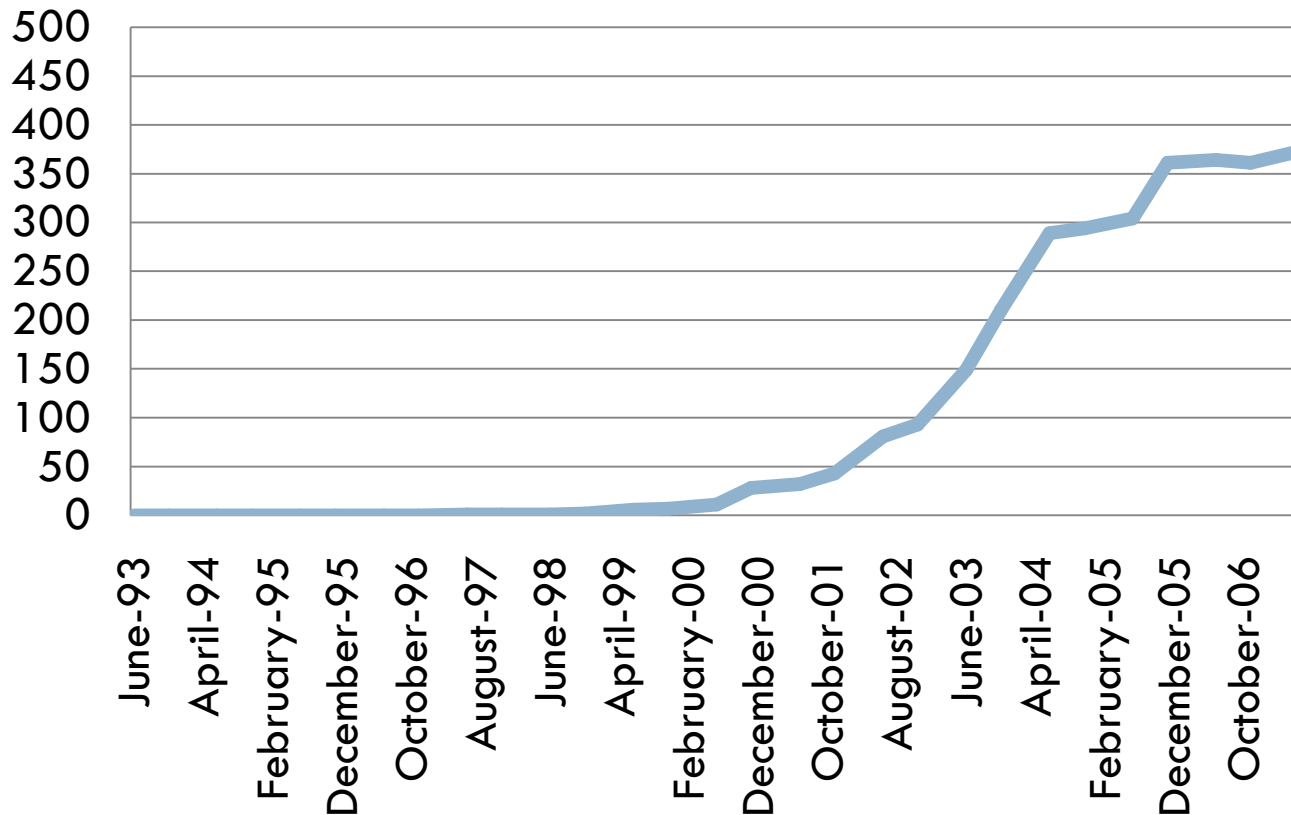
7

- The ARCOS Group.
- **Expand motivation.**
- Expand design.
- Expand evaluation.
- Conclusions.
- Ongoing Work.

# Trends in the supercomputing environment

8

## Clusters in top500.org



**75 % of supercomputers in top500 are clusters.**

The Expand Parallel File System

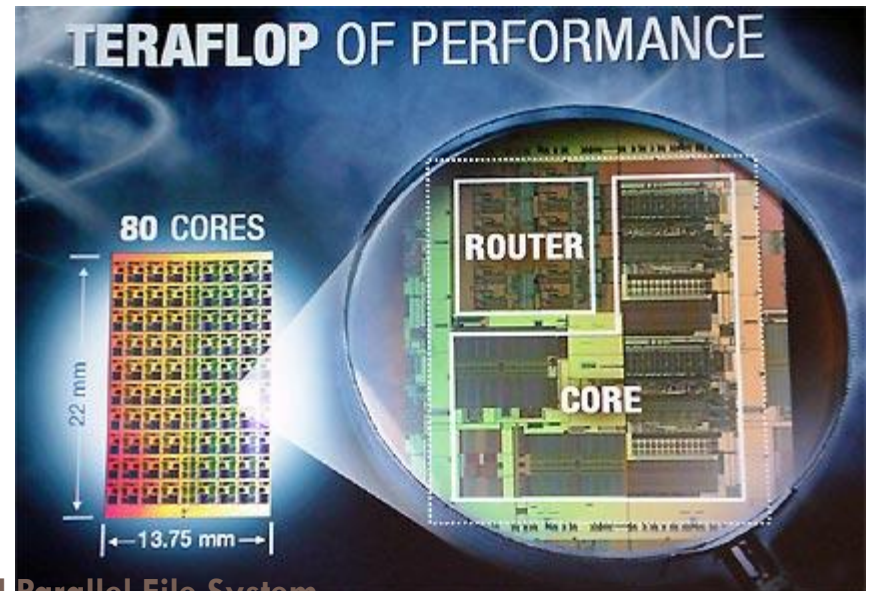
José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

# Trends in the supercomputing environment

9

- Number of transistors per chip still **doubling** every 1.5 years.
  - ▣ Does not mean doubling frequency, performance, ...
  - ▣ More space → more cores per chip.

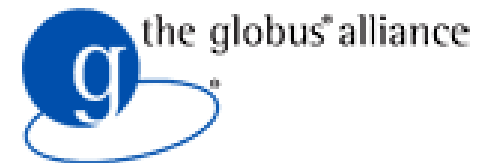
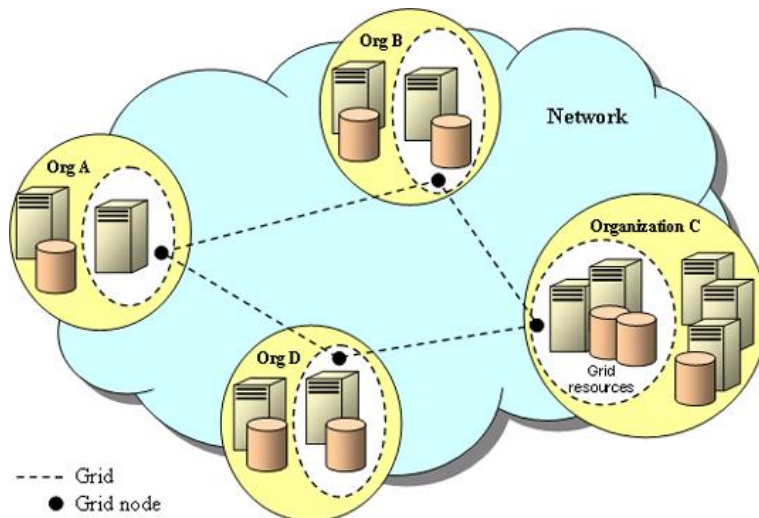


The Expand Parallel File System

# Trends in the supercomputing environment

10

- **Grid Computing:** Interconnecting supercomputers to aggregate geographically distributed resources.
  - ▣ Applications are deployed somewhere in the grid.
  - ▣ Applications read input data and produce output data.



**The Expand Parallel File System**

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

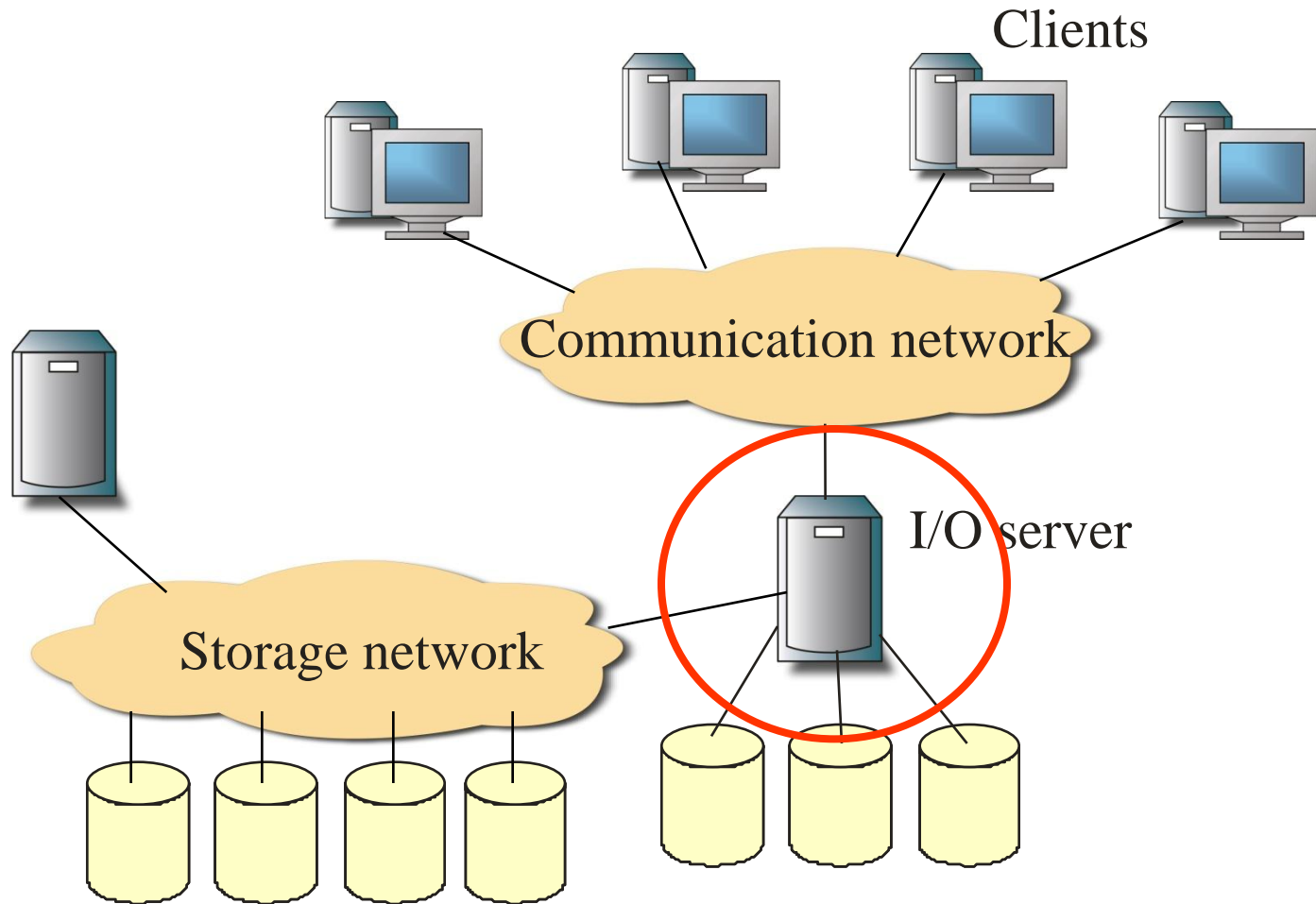
# Trends in the supercomputing environment

11

- Clusters becoming the preferred option for supercomputing.
- Processors with increasing capacity.
- Grid computing using clusters as a building block.
  
- I/O will remain as a major bottleneck.

# Storage system typical architecture

12

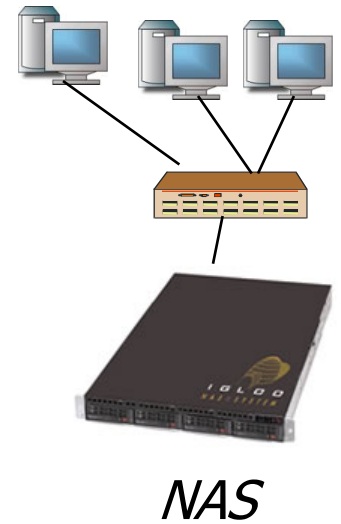
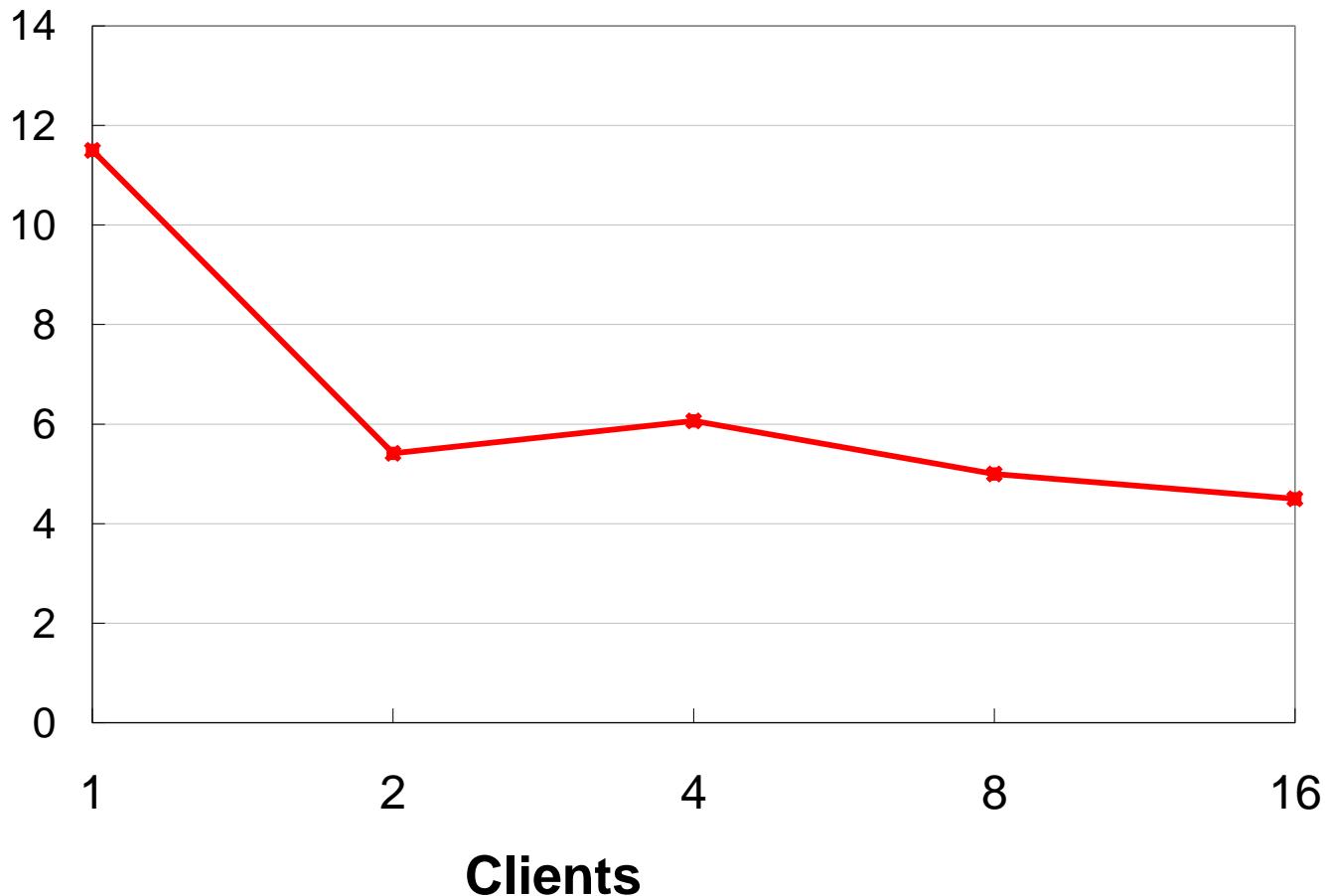


**The Expand Parallel File System**

# Problems with storage architectures

13

Aggregated bandwidth (MB/s)



The Expand Parallel File System

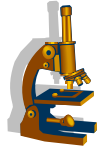
José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

# Solution: Parallelism

14

Parallel applications



Parallel computers



Parallel file systems



Parallel devices



**Exploit parallelism at multiple layers**



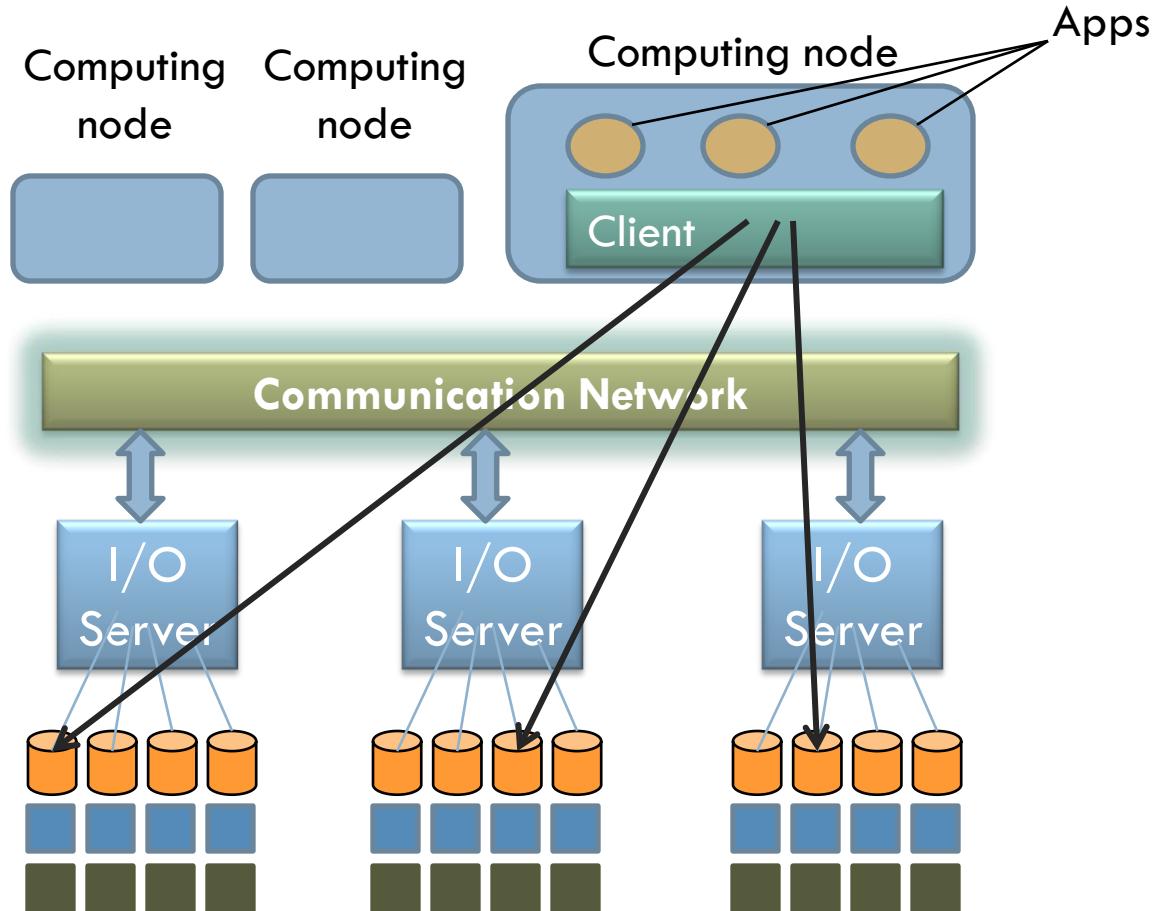
The Expand Parallel File System

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

# Parallel File System Architecture

15



**PVFS**  
PARALLEL VIRTUAL FILE SYSTEM

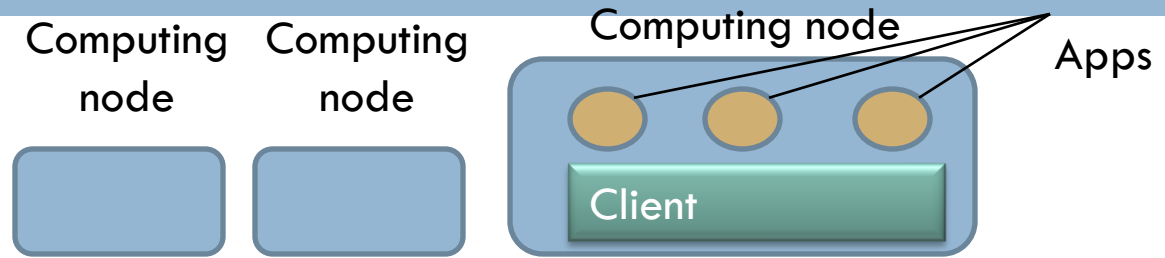
The Expand Parallel File System

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

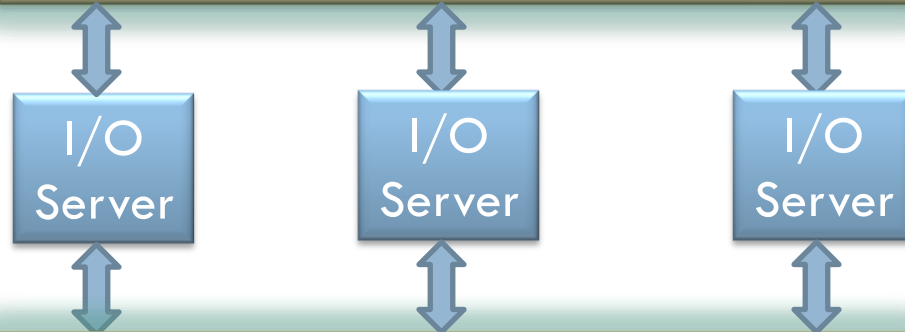
July 2007 - University of Modena

# Parallel File System Architecture

16



GPFS



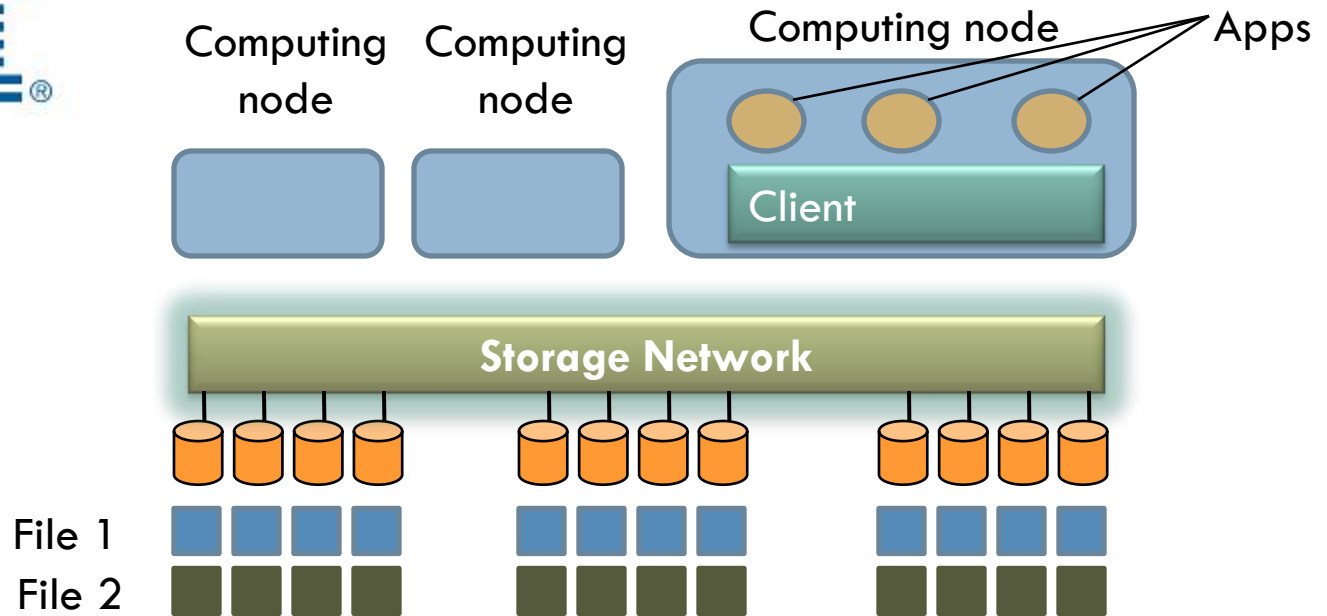
The Expand Parallel File System

# Parallel File System Architecture

17



GPFS



The Expand Parallel File System

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

# Expand Parallel File System: Motivation

18

- Provide a high performance storage system by using standard protocols and servers.
  - ▣ Easy integration of heterogeneous systems.
  - ▣ Reuse and aggregation of existing resources.
  - ▣ Parallel data access.

# Why Expand?

19

- A standard data server already includes almost all the needed functionality.
  - ▣ Reuse.
- Standard protocols and servers make resources universally available.
  - ▣ Easy to deploy.
- Independence of the underlying storage infrastructure.
  - ▣ Portability.

# Objective

20

- Offer a new approach to build PFS for cluster and grid environments by using standard data servers.
  
- Advantages:
  - **No server change needed.**
    - Operations at client side.
  - **Independence of client and server OS's.**
    - Operations through standard protocols.
  - **Simplified PFS construction.**
    - Take advantage of already implemented server high performance mechanisms.
  - **Allows mixing servers with different platforms and OS's.**
  - **Easy installation and configuration.**

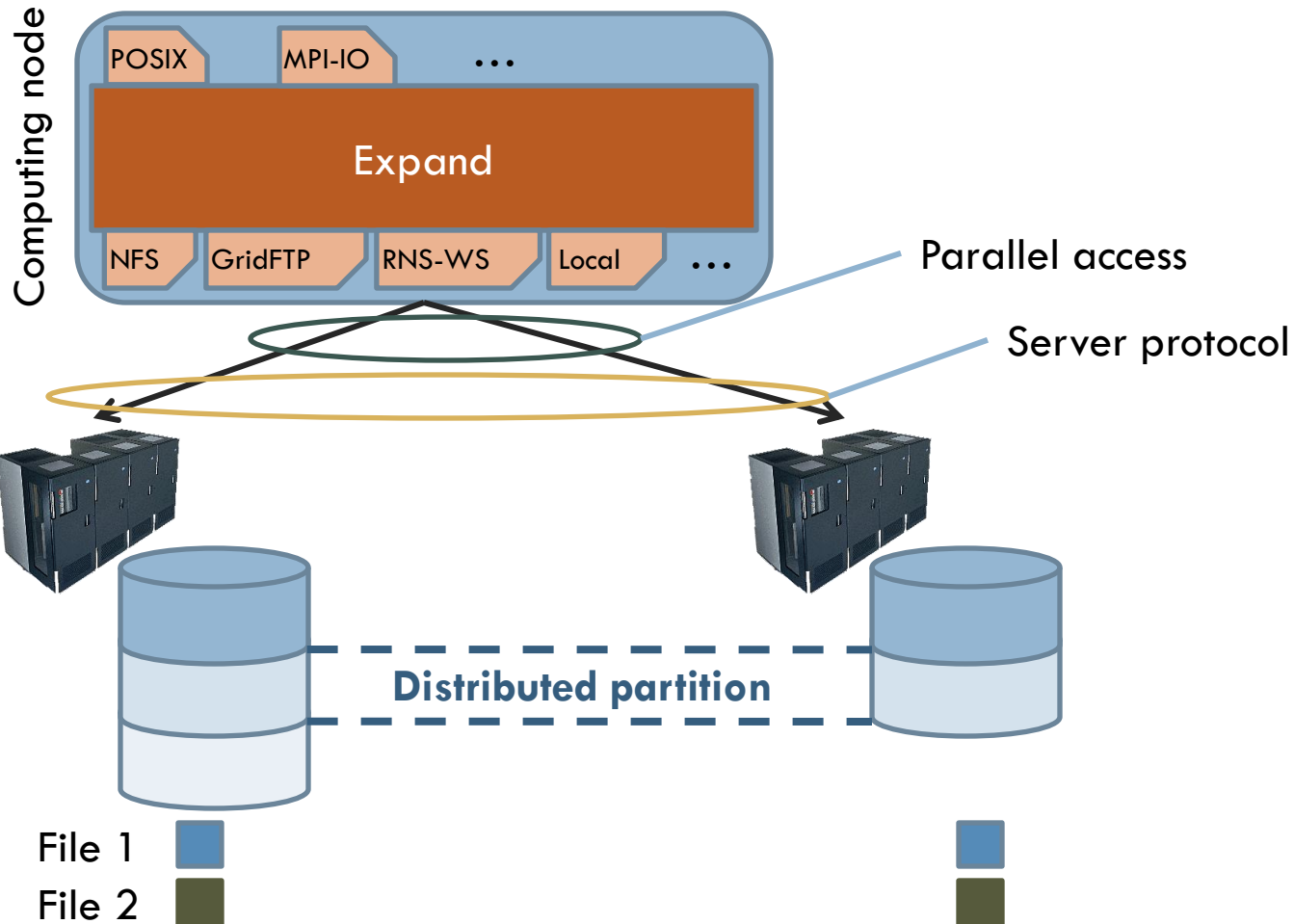
# Contents

21

- The ARCOS Group.
- Expand motivation.
- **Expand design.**
- Expand evaluation.
- Conclusions.
- Ongoing Work.

# Architecture

22



**The Expand Parallel File System**

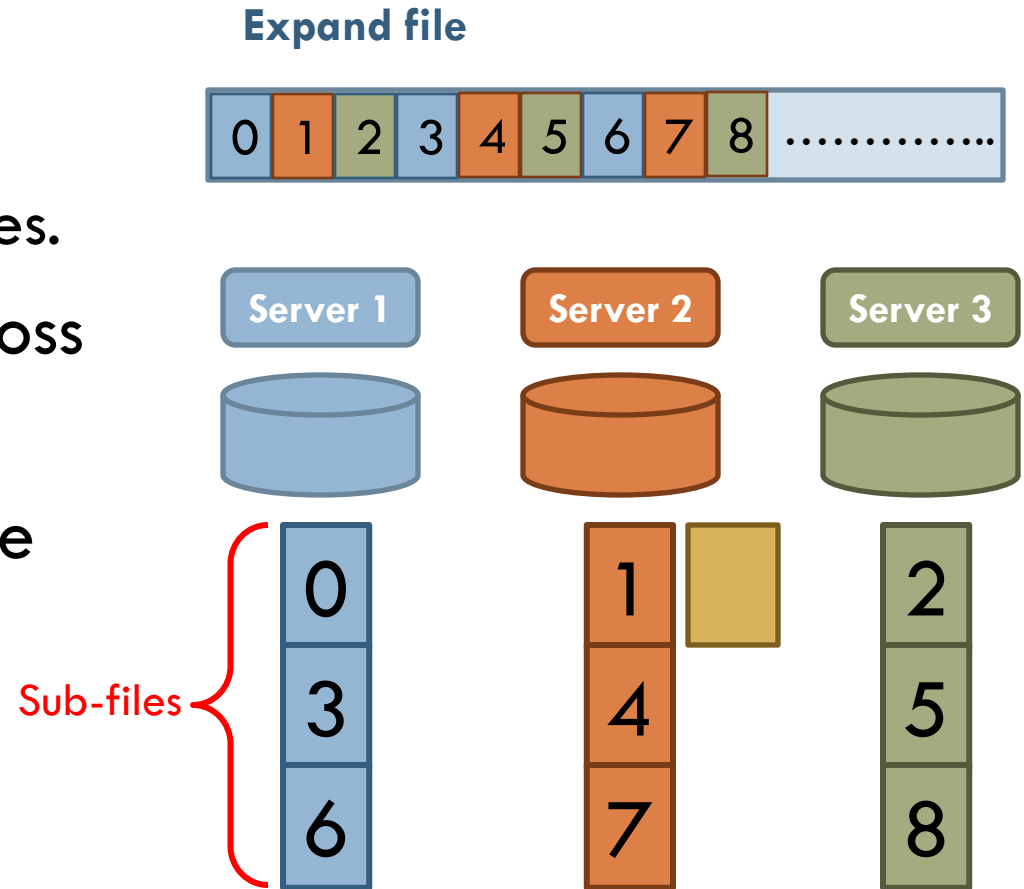
José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

# File structure

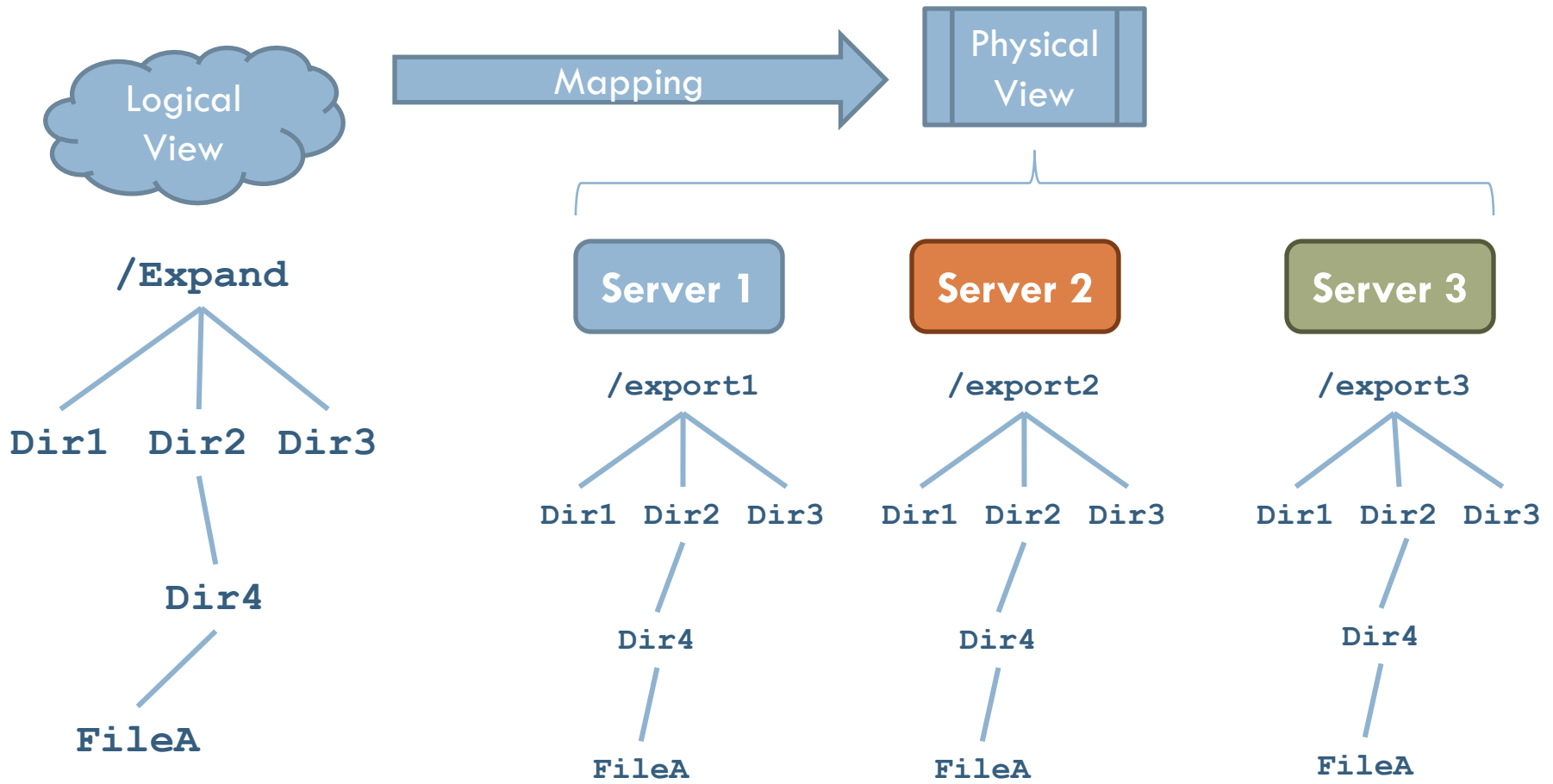
23

- Expand file:
  - ▣ Metadata sub-file.
  - ▣ Several data sub-files.
- Data distributed across several servers.
- File-to-server flexible mapping policy.



# Directory structure

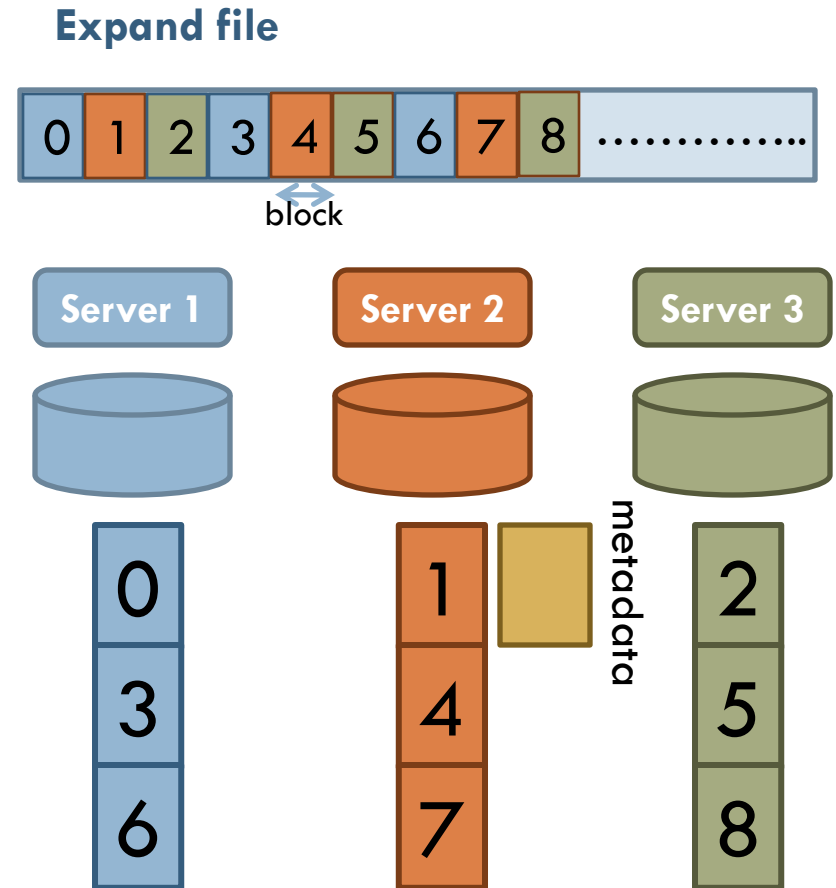
24



# Metadata management

25

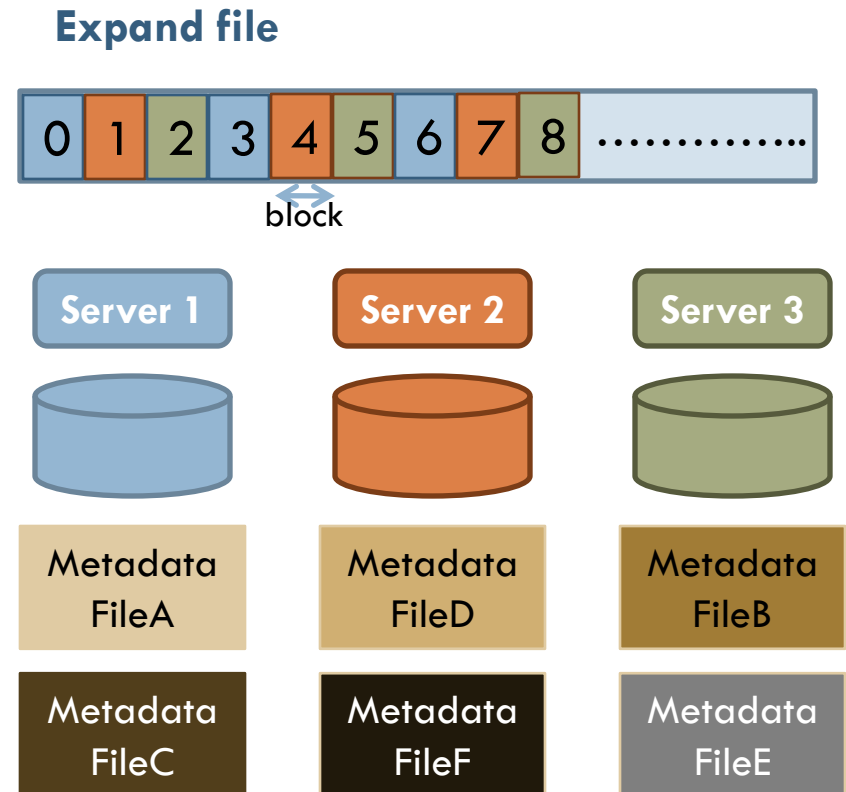
- Metadata distributed management.
  - Two levels.
  - Without locking.
  - No metadata manager.
- Metadata distributed across servers.
  - Master node.
  - Hashing on name.
  - Load balancing.



# Metadata management

26

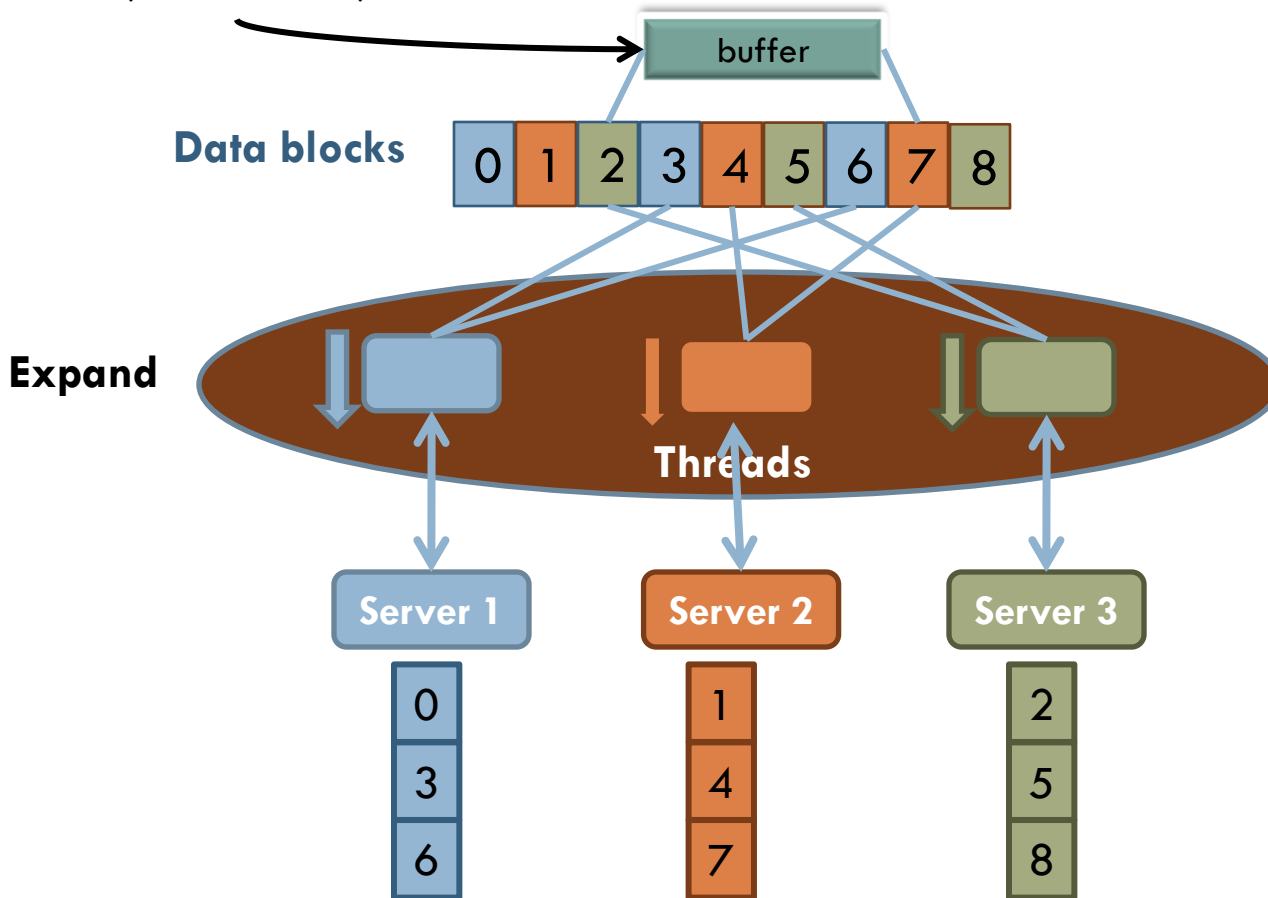
- Metadata distributed management.
  - Two levels.
  - Without locking.
  - No metadata manager.
- Metadata distributed across servers.
  - Master node.
  - Hashing on name.
  - Load balancing.



# Parallel access

27

`read(fd, buffer, count)`



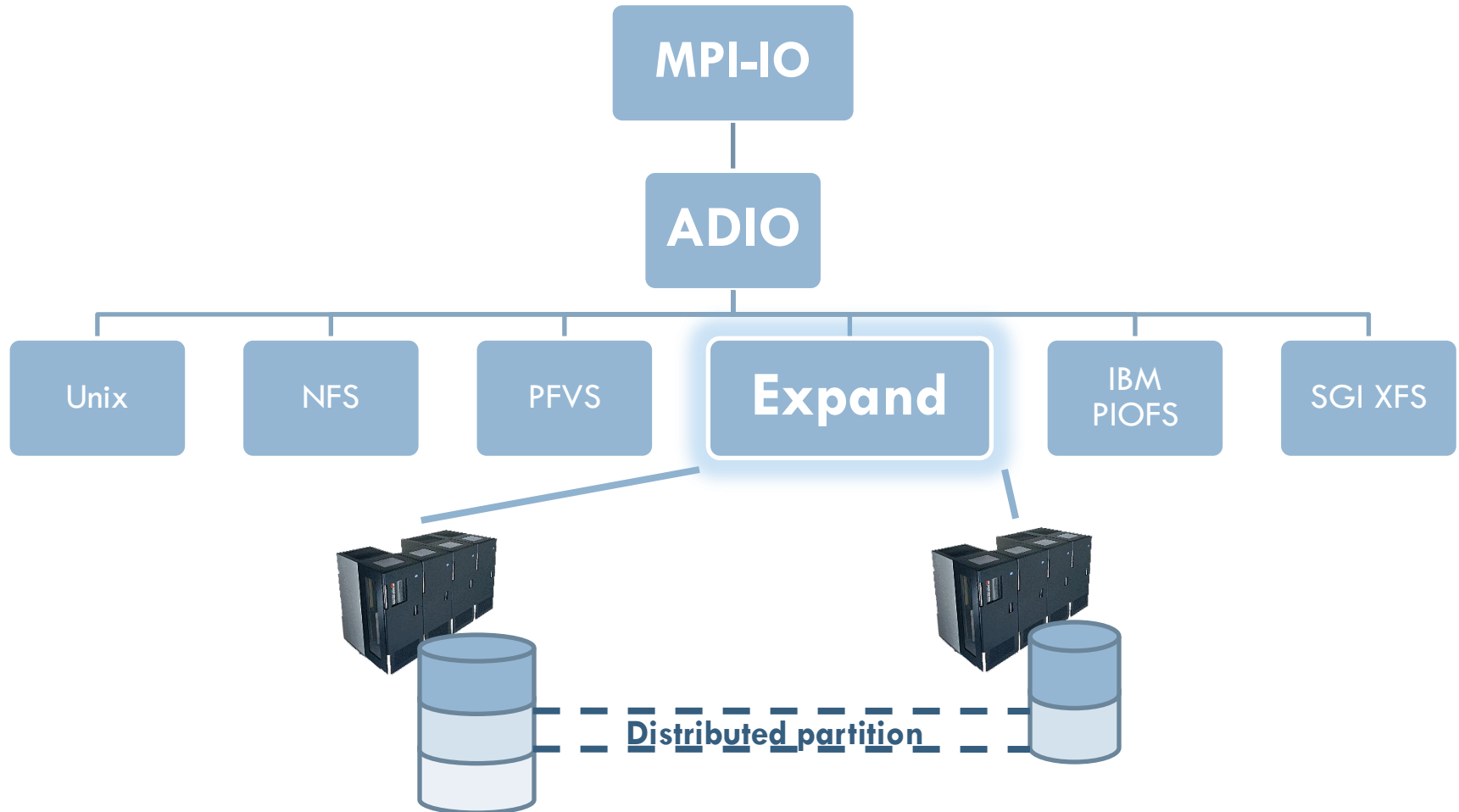
The Expand Parallel File System

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

# MPI-IO interface using ROMIO

28



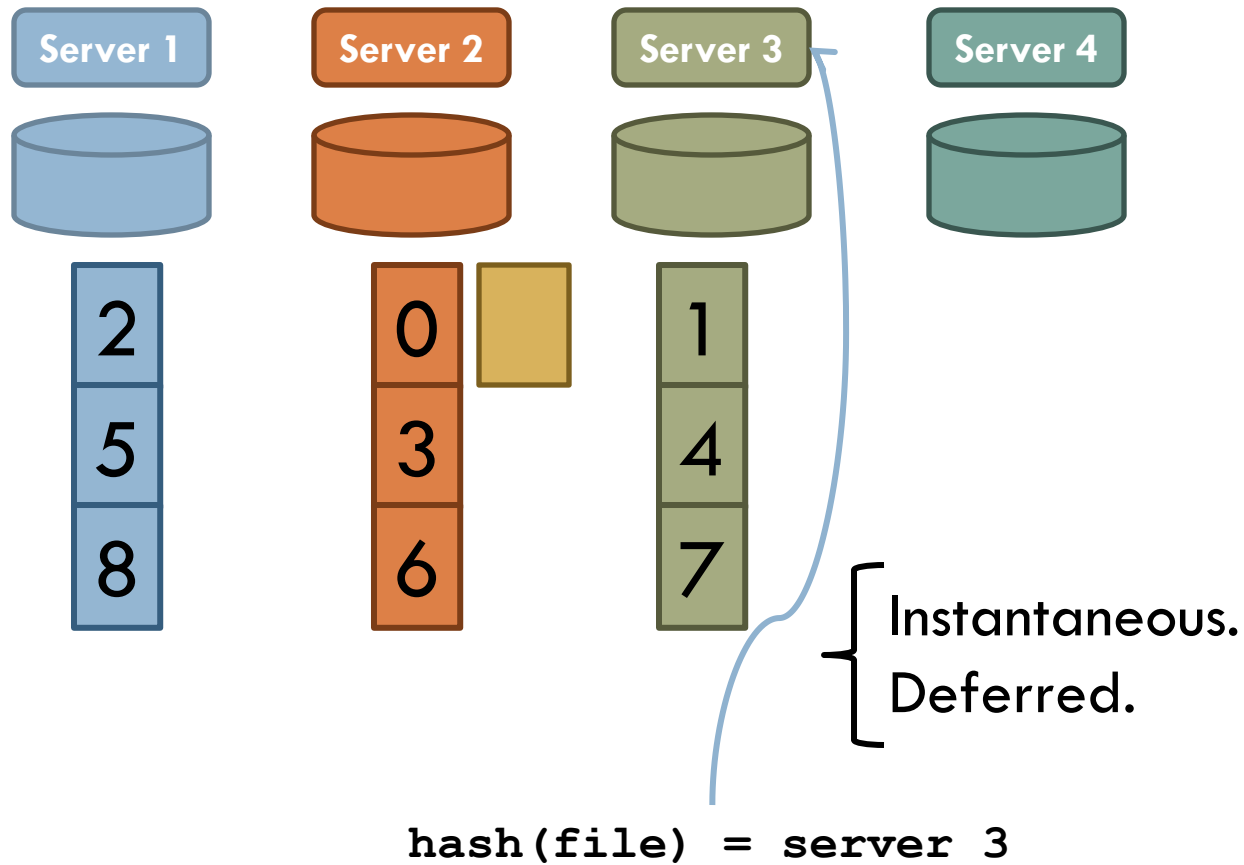
**The Expand Parallel File System**

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

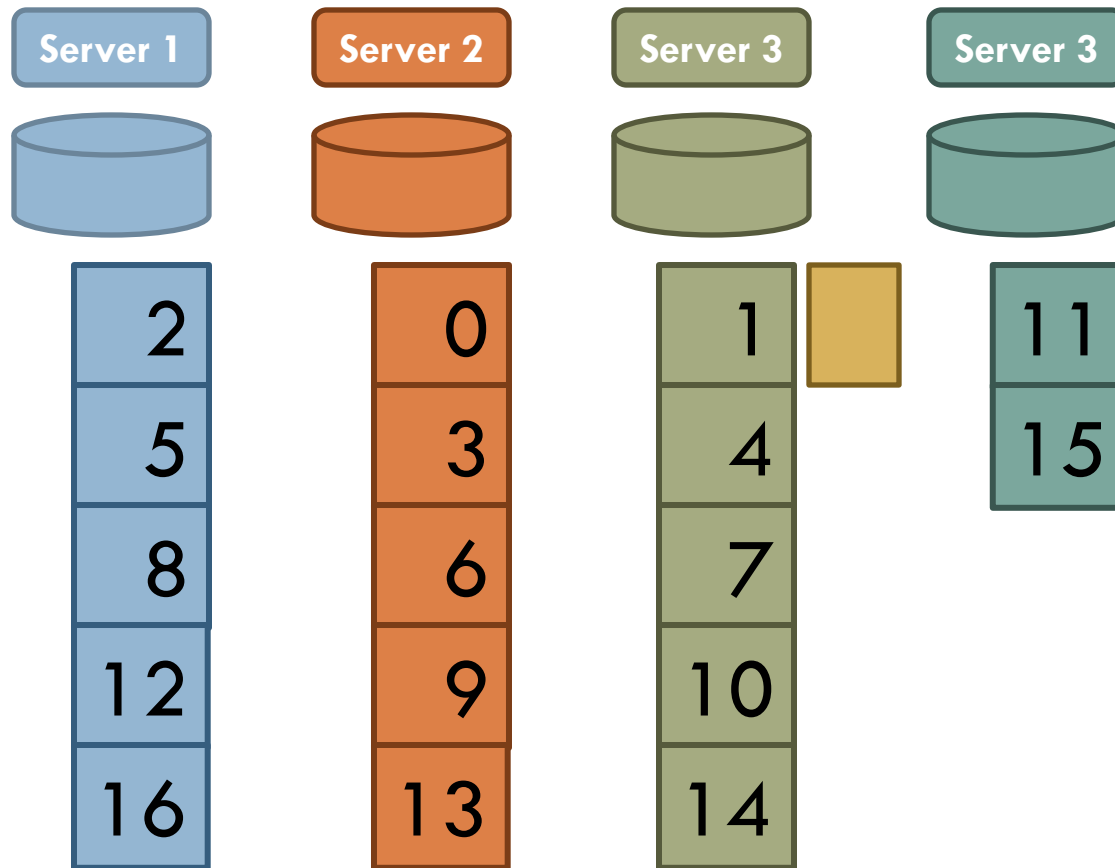
# Dynamic partition reconfiguration

29



# Dynamic partition reconfiguration

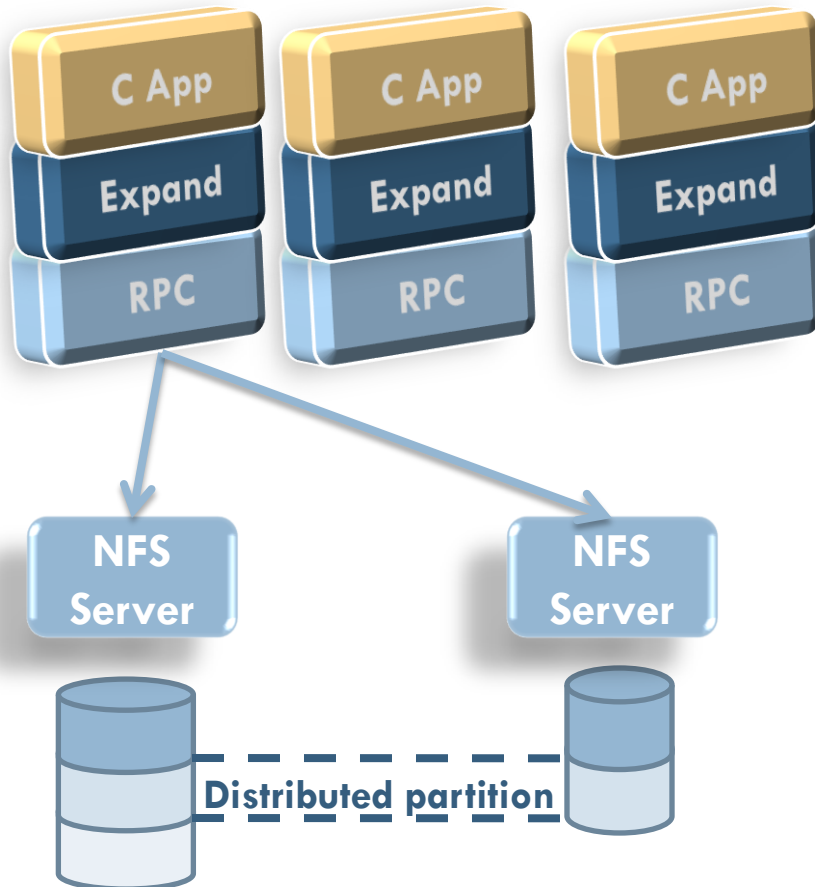
30



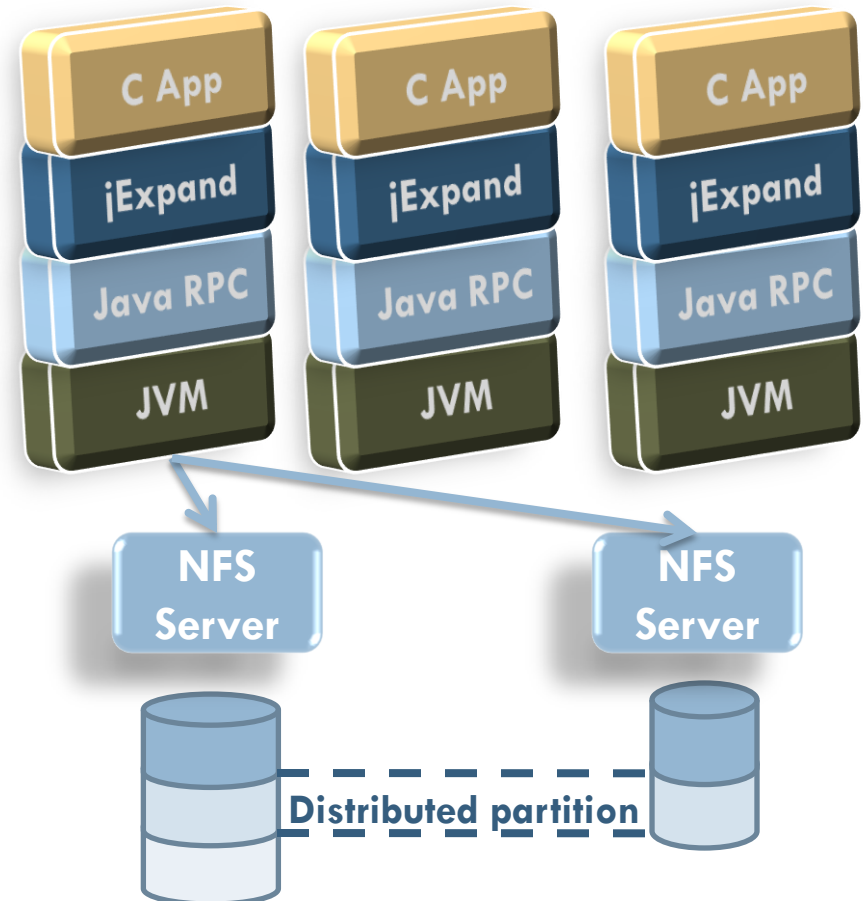
# Expand cluster versions

31

## Linux/NFS



## Java/NFS



The Expand Parallel File System

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

# Contents

32

- The ARCOS Group.
- Expand motivation.
- **Expand design.**
  - ▣ **Expand adaptation for Grid Computing.**
- Expand evaluation.
- Conclusions.
- Ongoing Work.

# Requirements for a Grid File System

33

- **Hierarchical logical space name.**
  - Resource Namespace Service (RNS).
- **Standard access interface.**
  - POSIX and MPI-IO.
- **Data access.**
  - GridFTP.
- **Security.**
  - Grid Security Infrastructure (GSI).
- **Performance optimization and improvement.**
  - Paralle I/O.



# Contents

35

- The ARCOS Group.
- Expand motivation.
- Expand design.
- **Expand evaluation.**
- Conclusions.
- Ongoing Work.

# Evaluation

36

- How does Expand behaves compared to other existing solutions?
  - Cluster
    - PFVS.
    - GPFS.
  - Grid
    - Globus Grid services.

# Cluster environment

37

- 8 biprocessors (Pentium VI, 3.2 GHz).
- 2 GB RAM per node.
- Network: Gigabit ethernet.
  - ▣ Expand.
  - ▣ PVFS.
  - ▣ GPFS.

# Cluster benchmarking

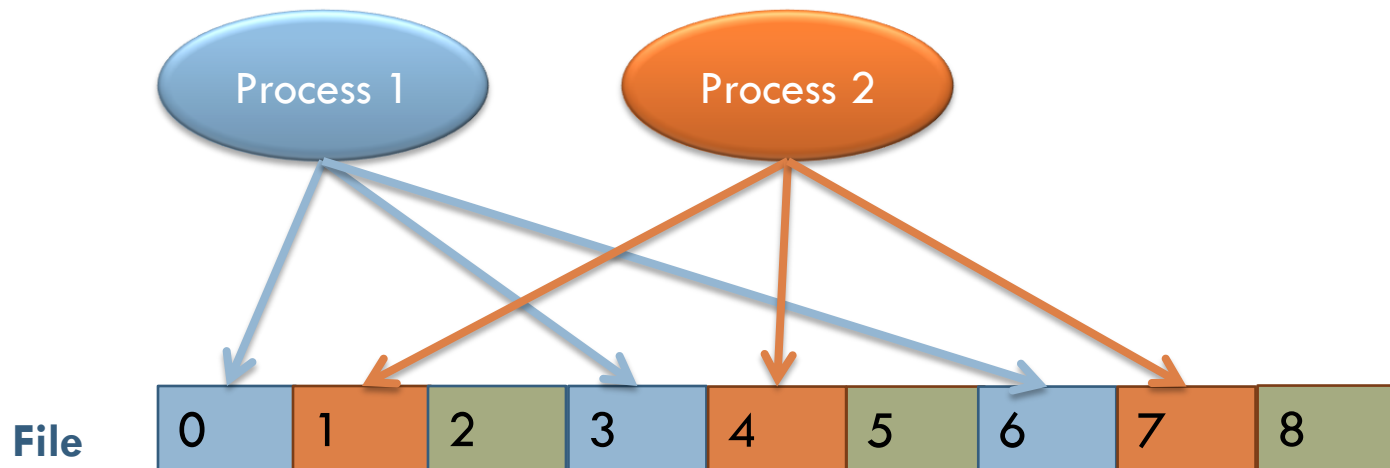
38

- High performance.
  - ▣ Parallel access to a file: IOR benchmark.
  - ▣ FLASH I/O benchmark.
- Metadata operations.
- High throughput.
  - ▣ Image processing.
- Dynamic partition reconfiguration.

# High performance: Parallel access to a file

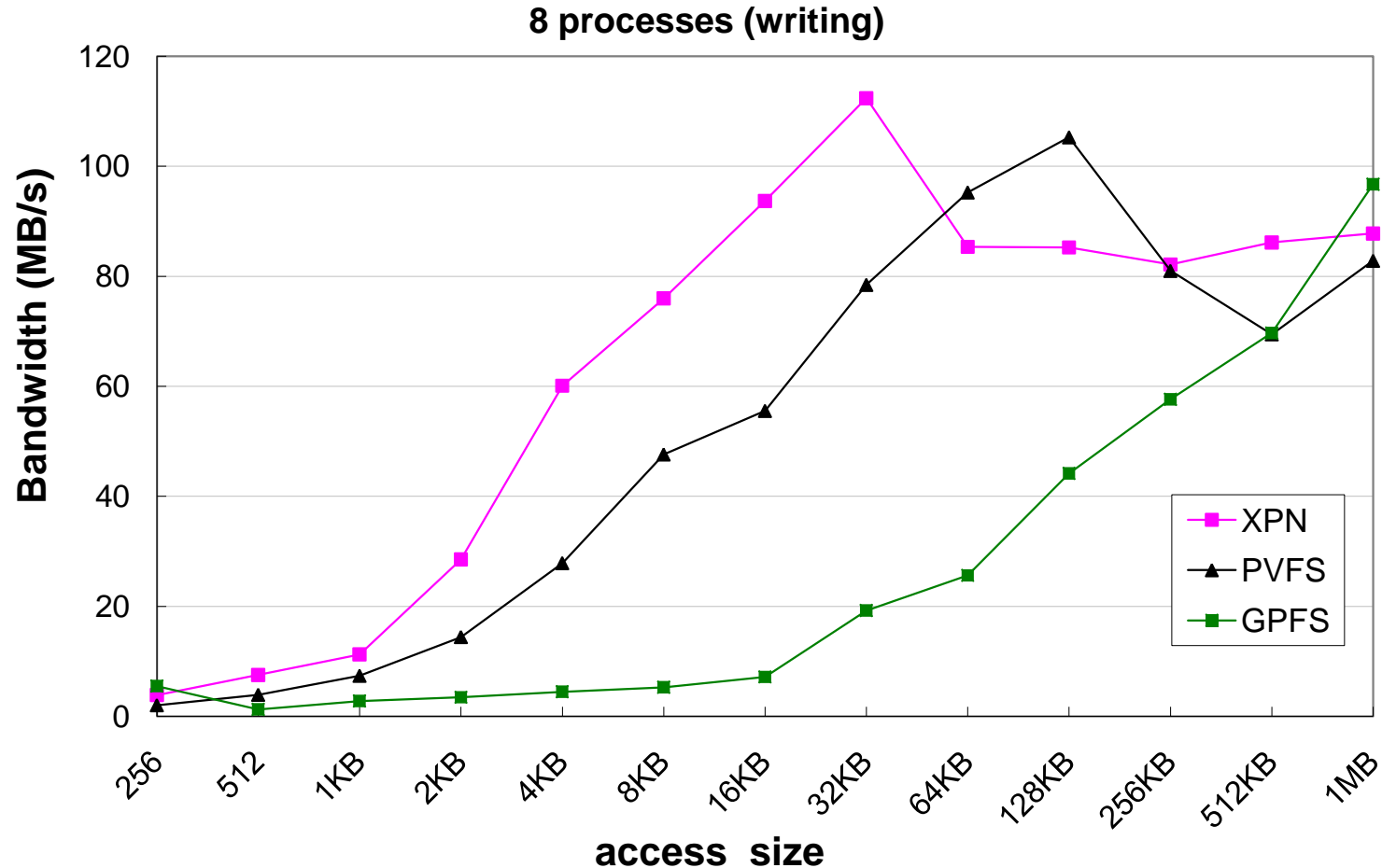
39

- Parallel program (I/O) making interleaved writes and reads to a single file with different access sizes.
- MPI-IO interface



# High performance: Parallel access to a file for writing

40



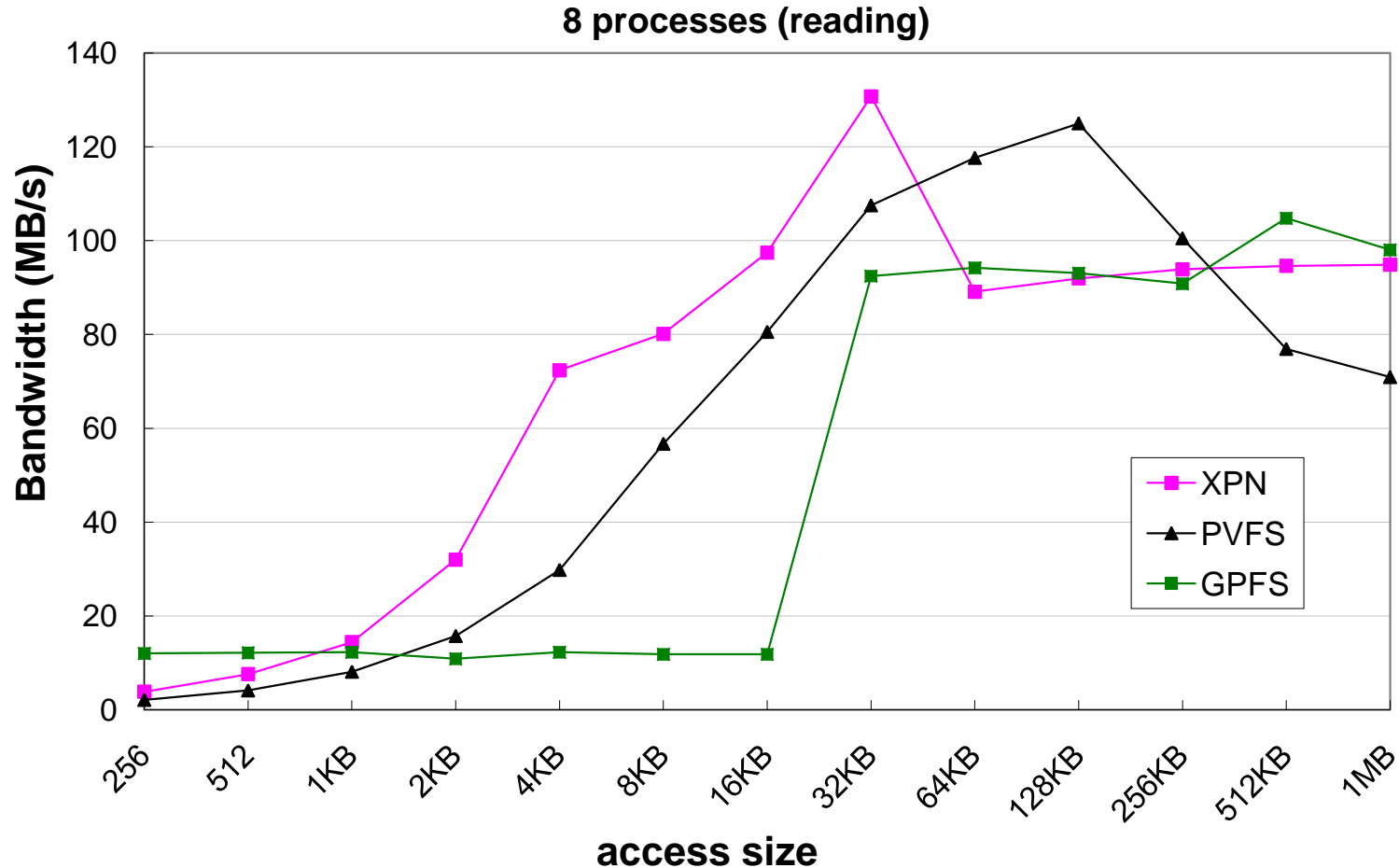
The Expand Parallel File System

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

# High performance: Parallel access to a file for reading

41



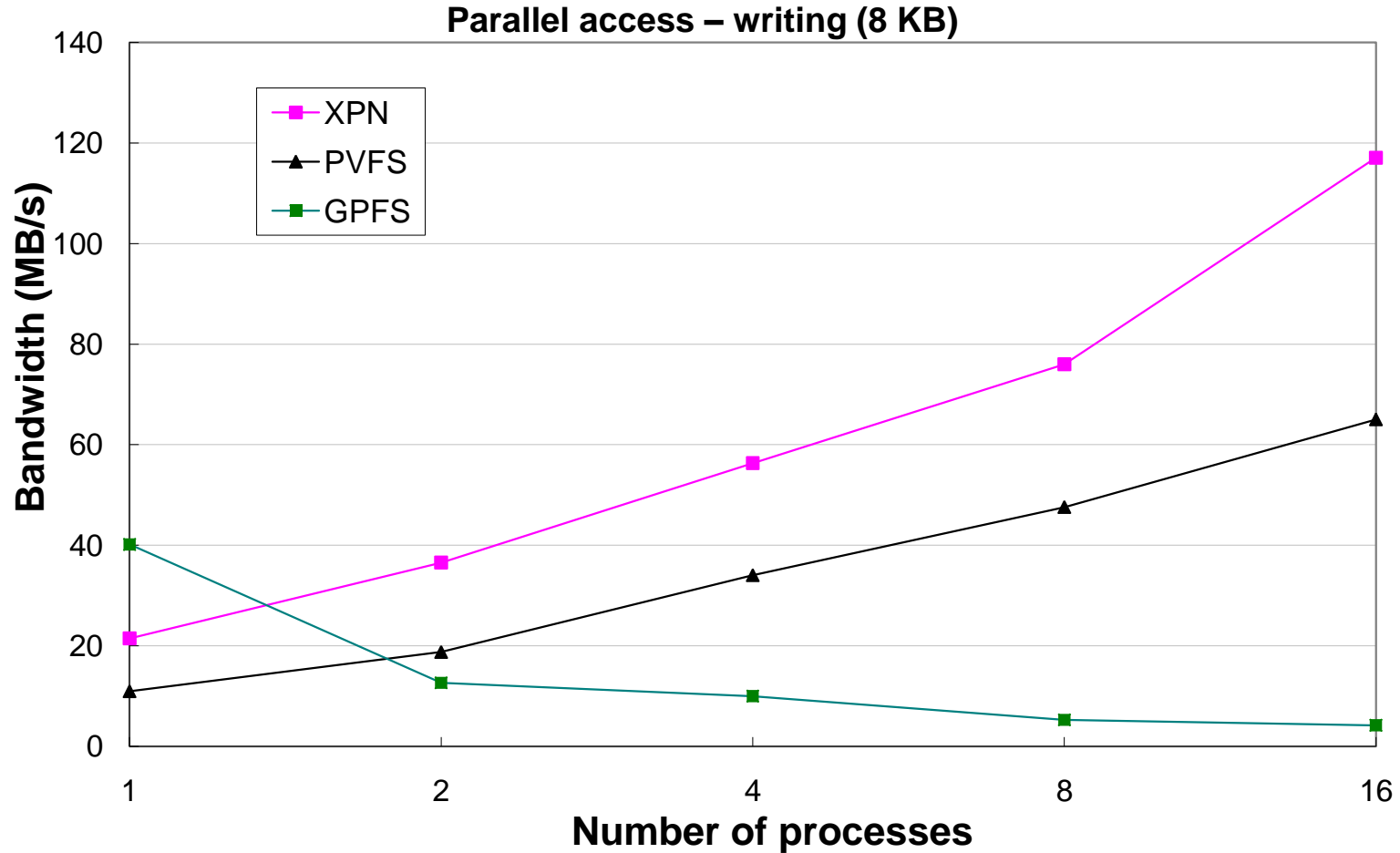
The Expand Parallel File System

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

# High performance: Parallel access to a file for writing

42



The Expand Parallel File System

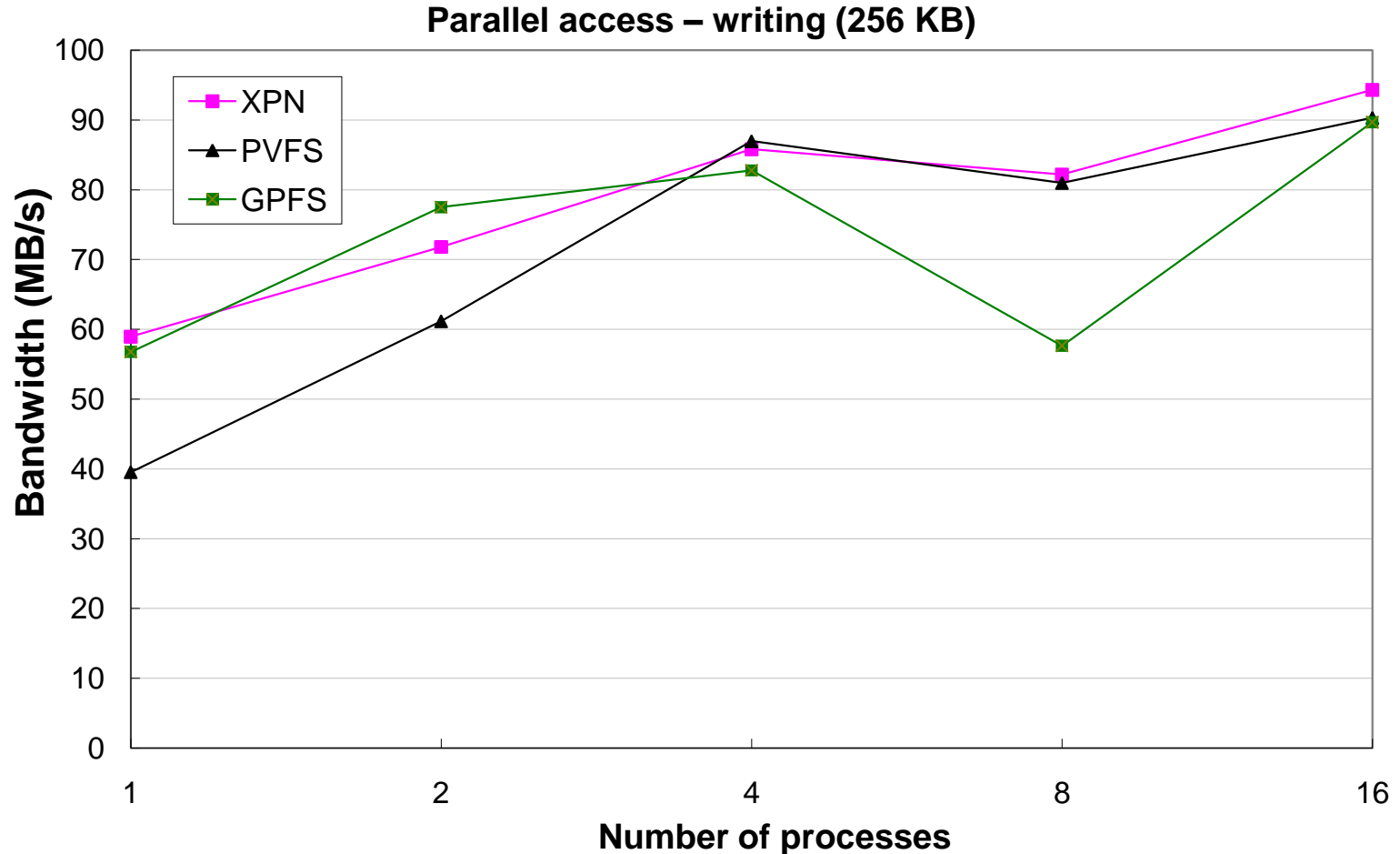
José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

# High performance.

## Parallel access to a file for writing

43



The Expand Parallel File System

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

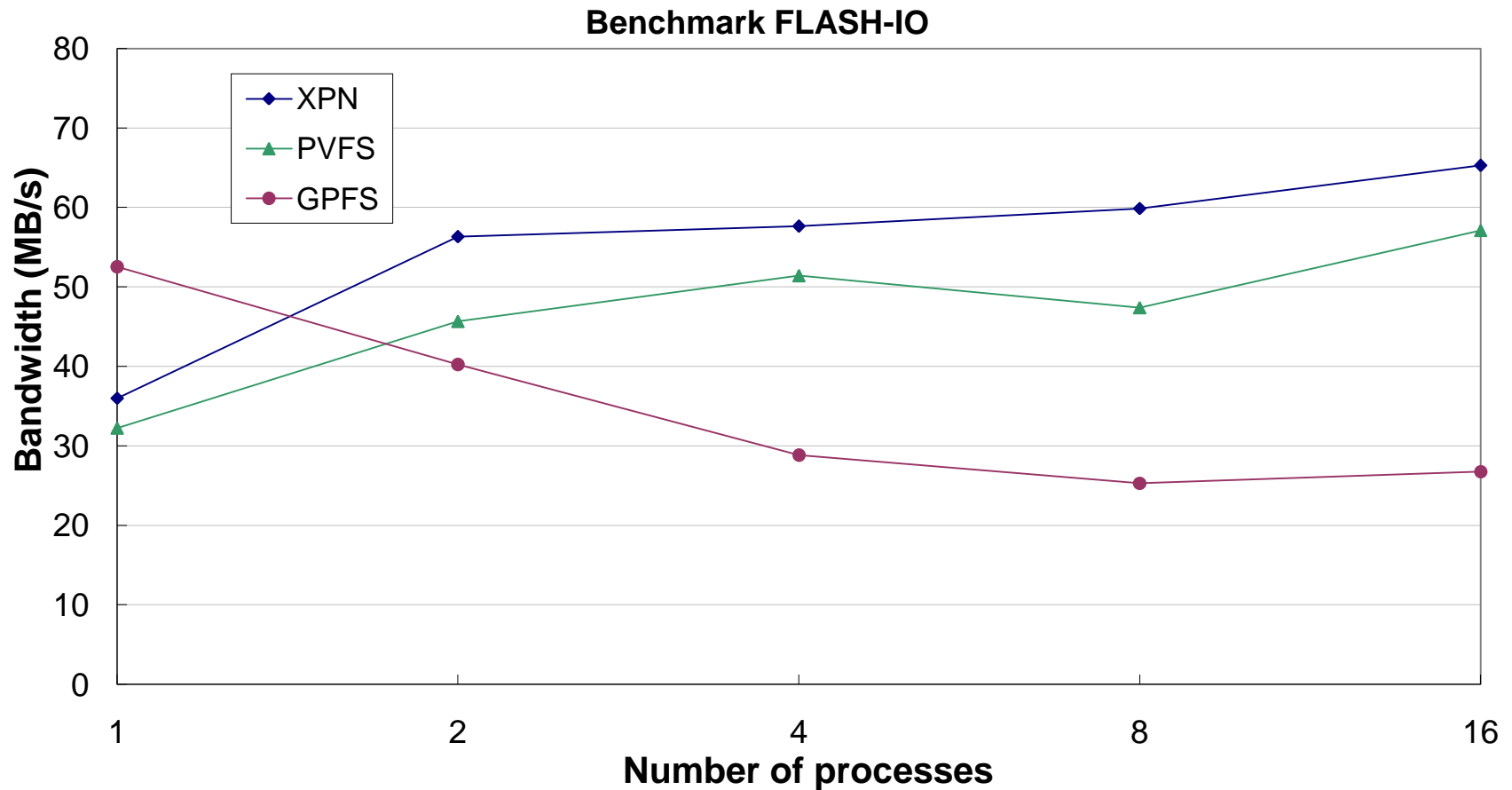
# High Performance: FLASH-IO

44

- FLASH is a parallel application simulating thermonuclear flashes.
- FLASH-IO simulates I/O operations performed by FLASH.
- Data size is proportional to number of running processes.
  - ▣ 1 process → 73.53 MB
  - ▣ 16 processes → 1.16 GB

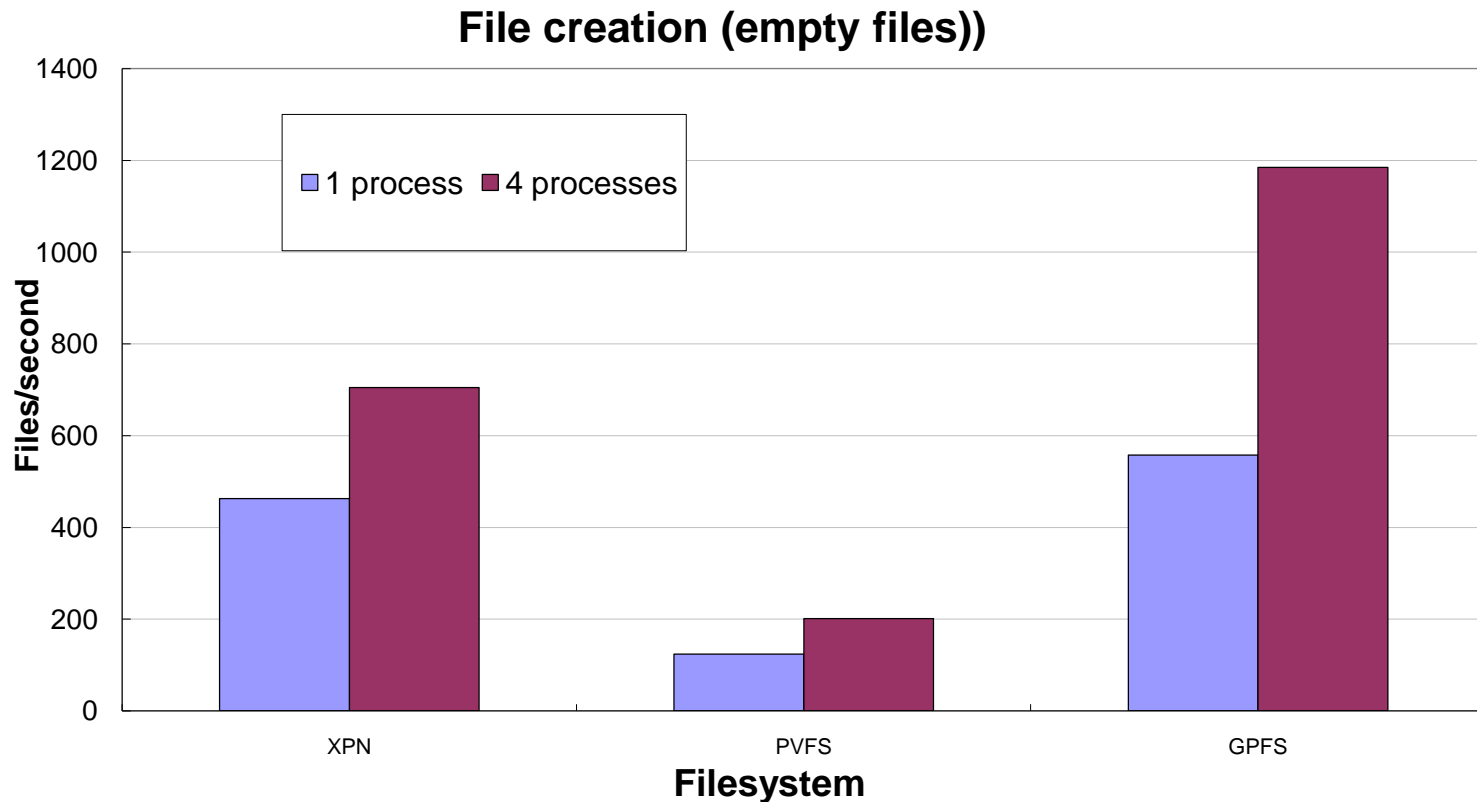
# High Performance: FLASH-IO

45



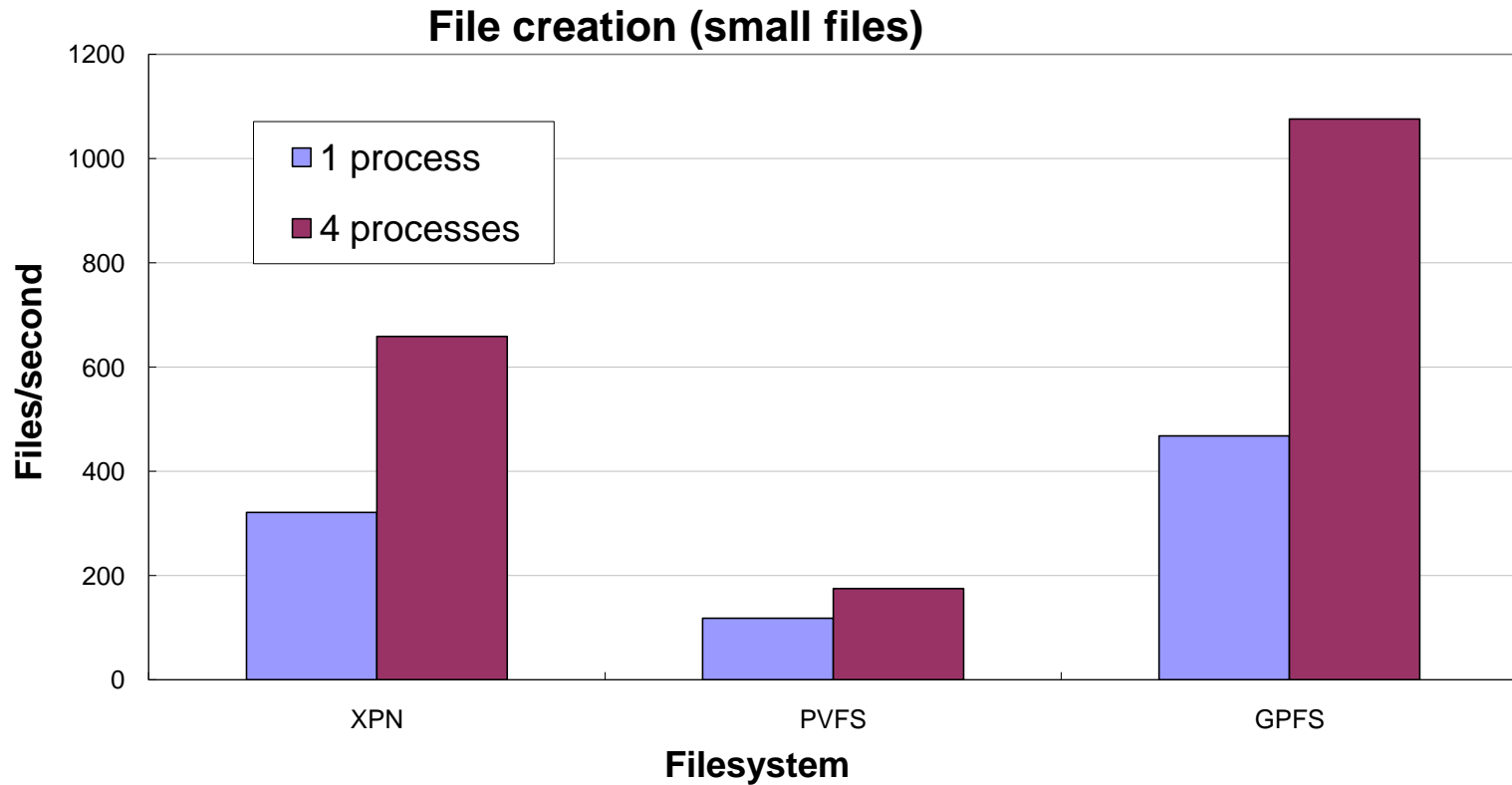
# Metadata: Creating empty files

46



# Metadata: Creating small files

47



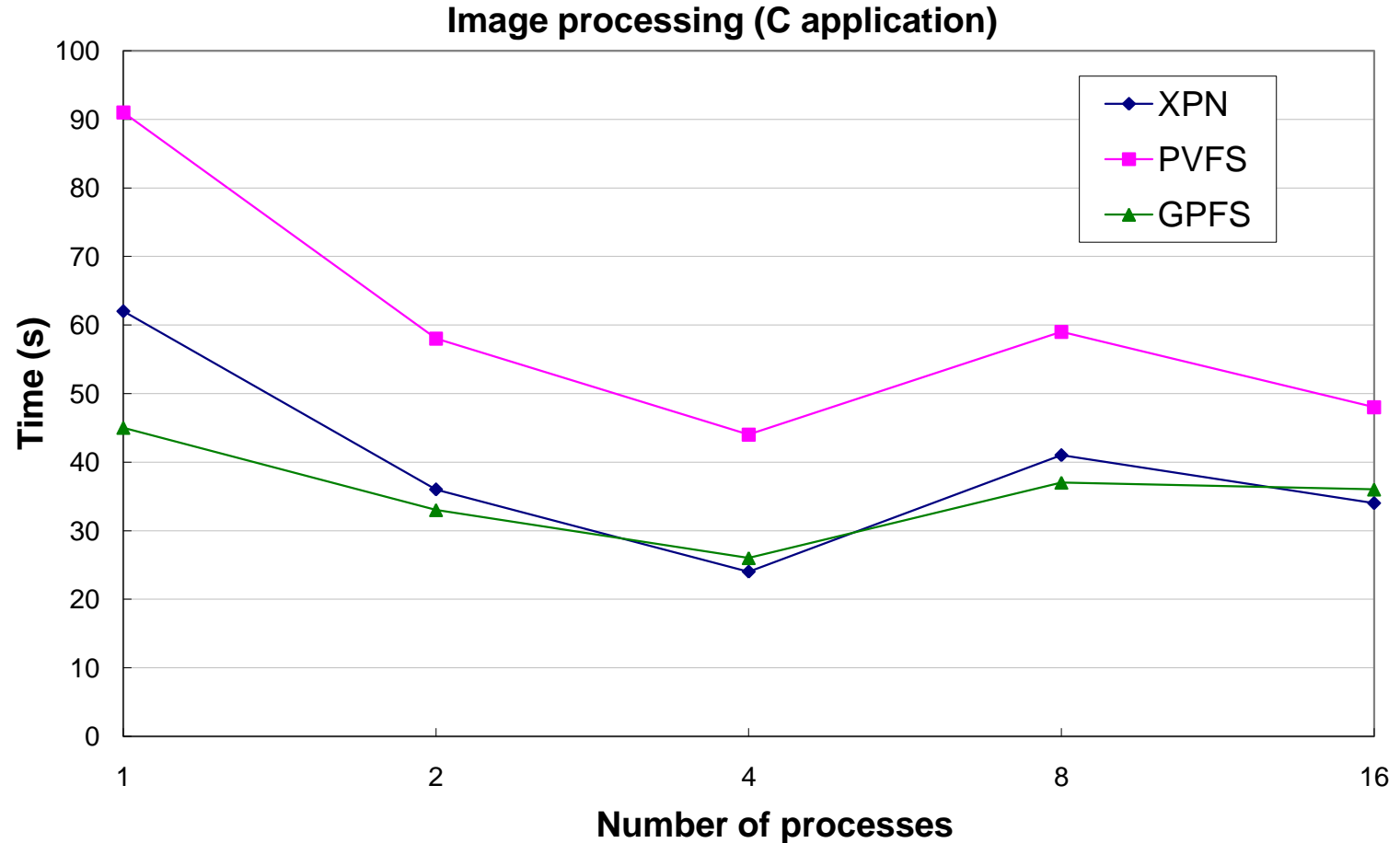
# High throughput

48

- Parallel application processing a set of 256 images.
  - ▣ Each process works on a subset of images independently.
  - ▣ No concurrent access to a file.
- Sizes:
  - ▣ Image file → 5 MB.
  - ▣ Full dataset → 2.5 GB.
- The process applies to each image file a fixed bitmask to generate a new image file.

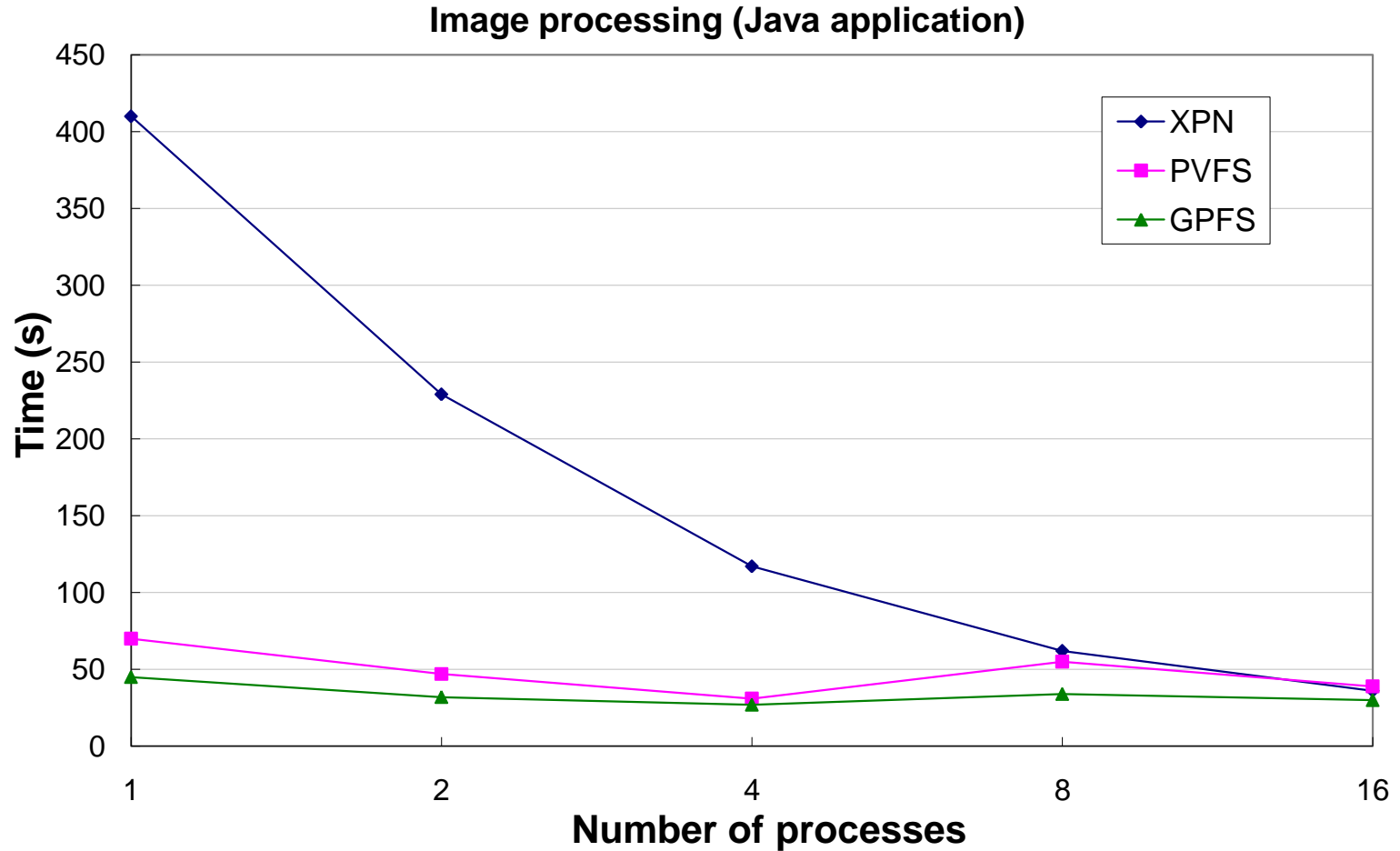
# High throughput: Image processing in C

49



# High throughput: Image processing in Java

50



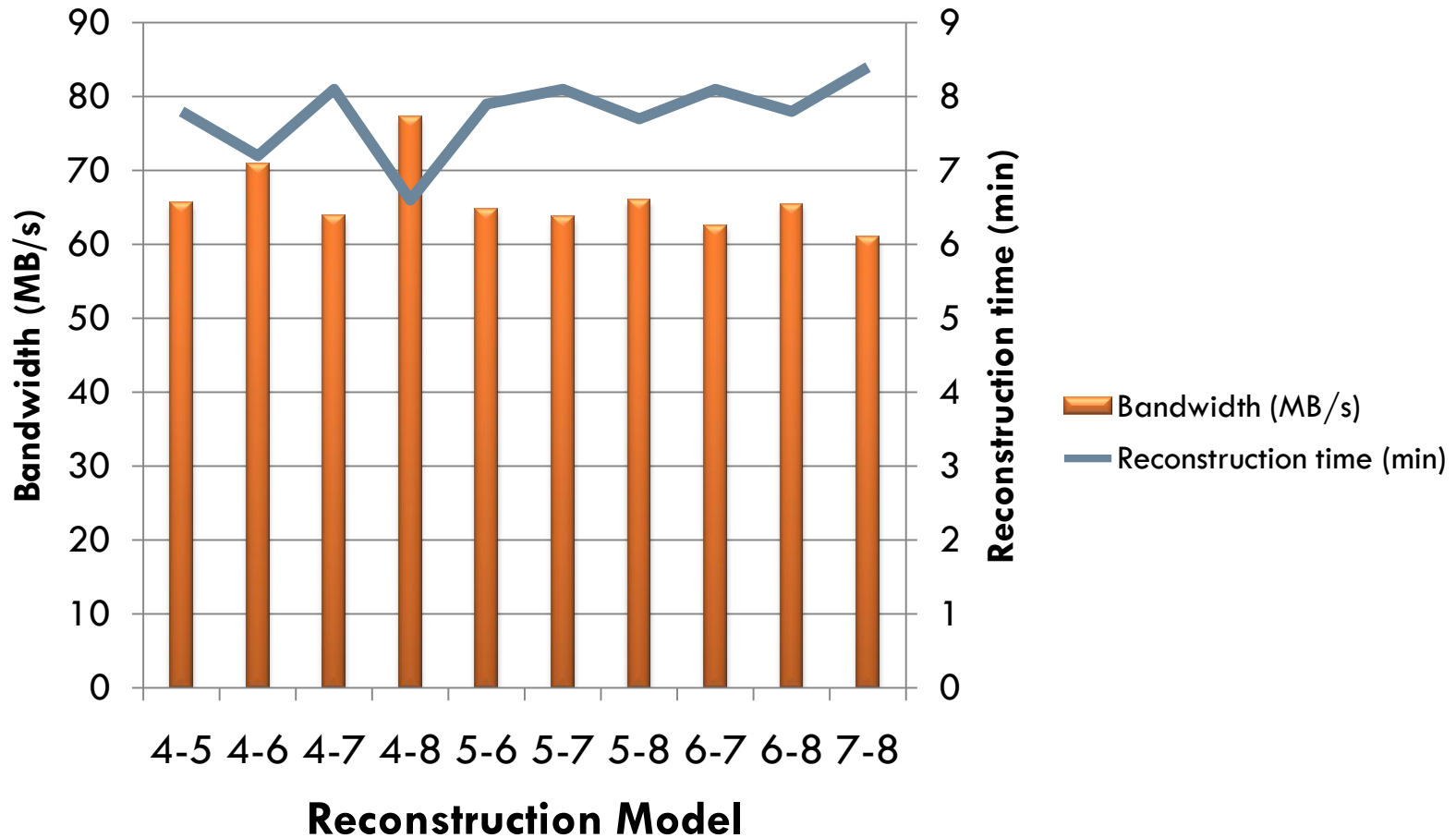
The Expand Parallel File System

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

# Dynamic partition reconfiguration: Adding new nodes

51



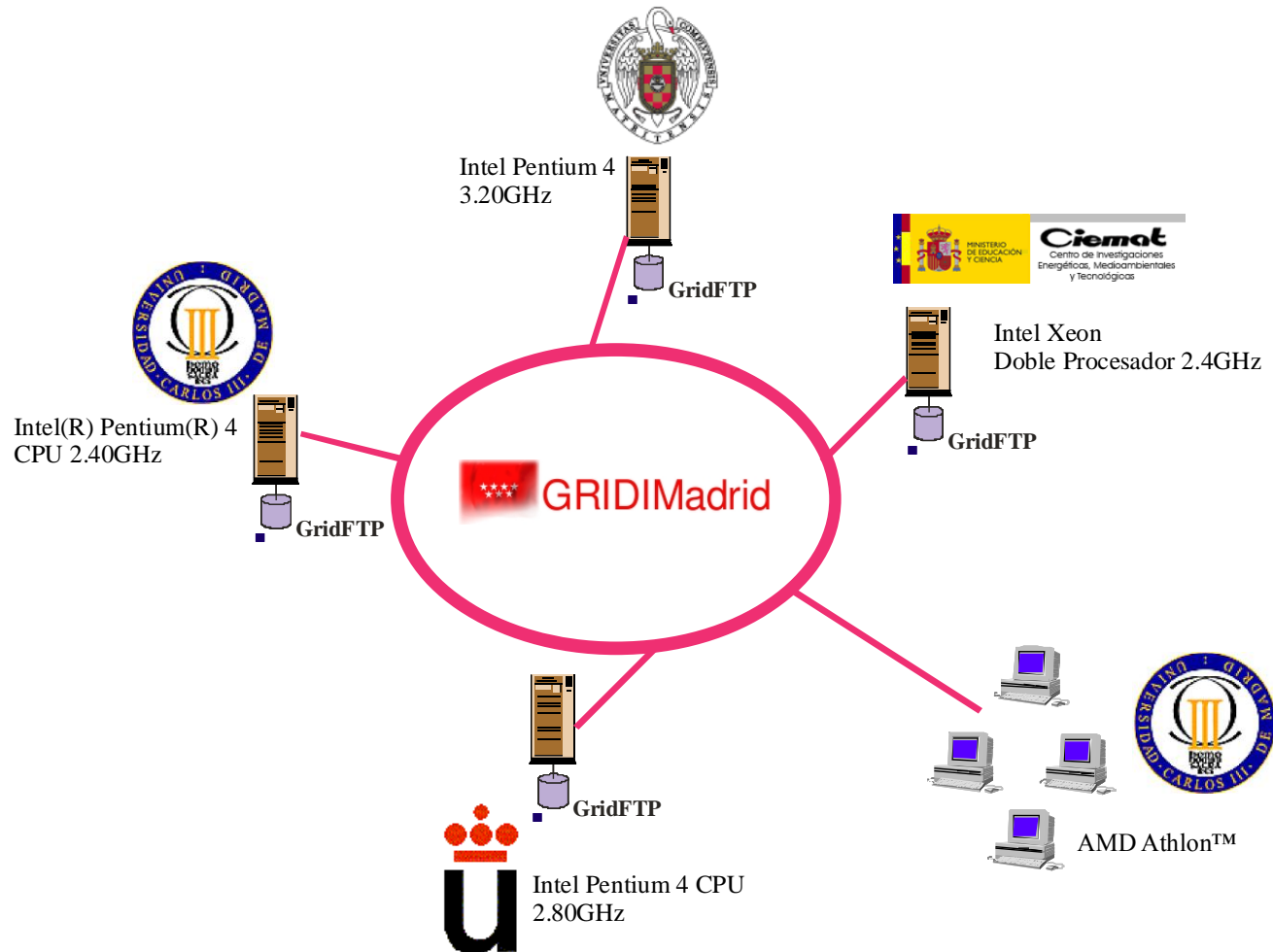
# Grid evaluation environment

52

- Evaluation for high throughput.
  - ▣ Perform 500 jobs.
  - ▣ Each job selects randomly a file (among 200) to access.
  - ▣ File size is 200 MB.

# Testbed environment

53



The Expand Parallel File System

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

# Evaluated scenarios

54

## Typical Grid

- ❑ Completely transfer file to local node.
- ❑ Processing starts after transfer finishes.
- ❑ Globus services for transfer.
  - ❑ `globus-url-copy`

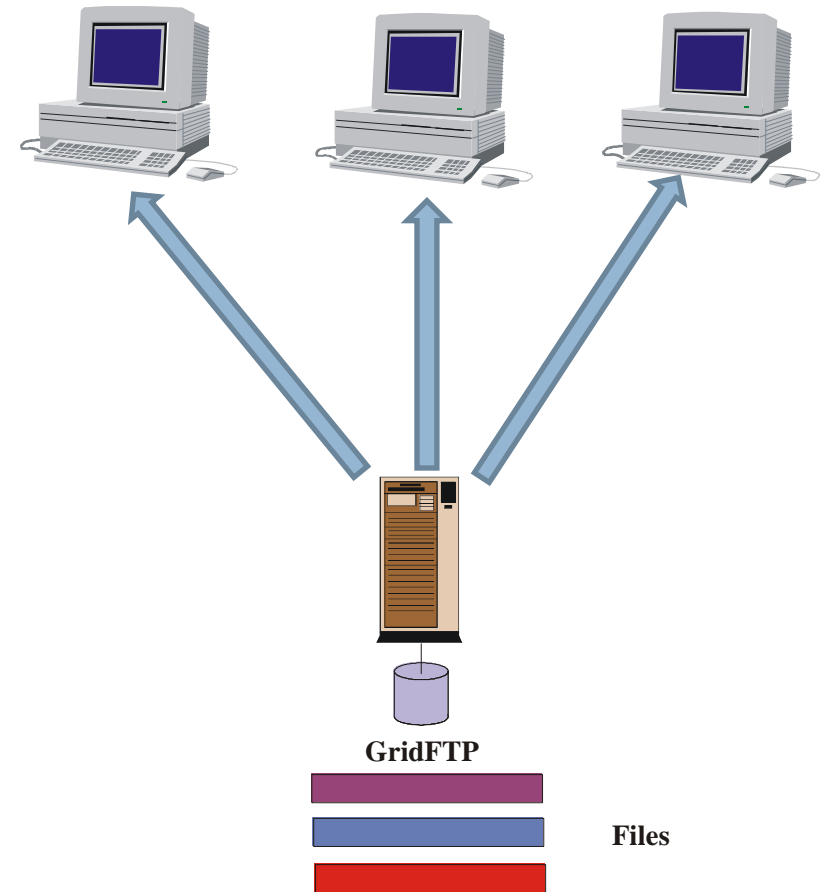
## Expand

- ❑ Direct remote access to file.
- ❑ **No previous transfer to node needed!**

# Model 1 / Scenario 1

55

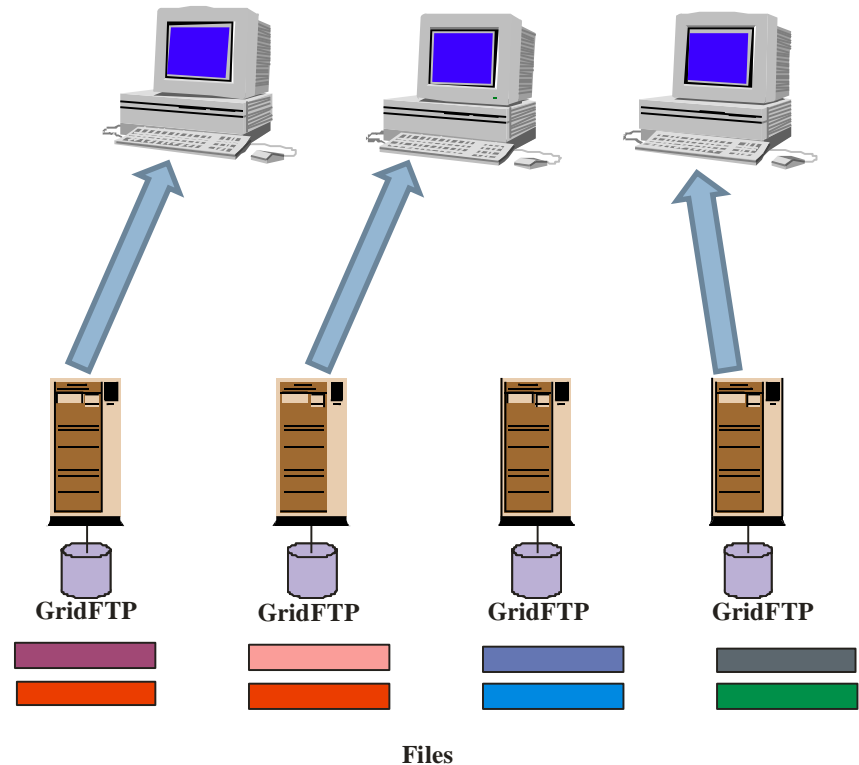
- 1 server
- Complete transfer to local node.
- Application access local copy.



# Model 2 / Scenario 1

56

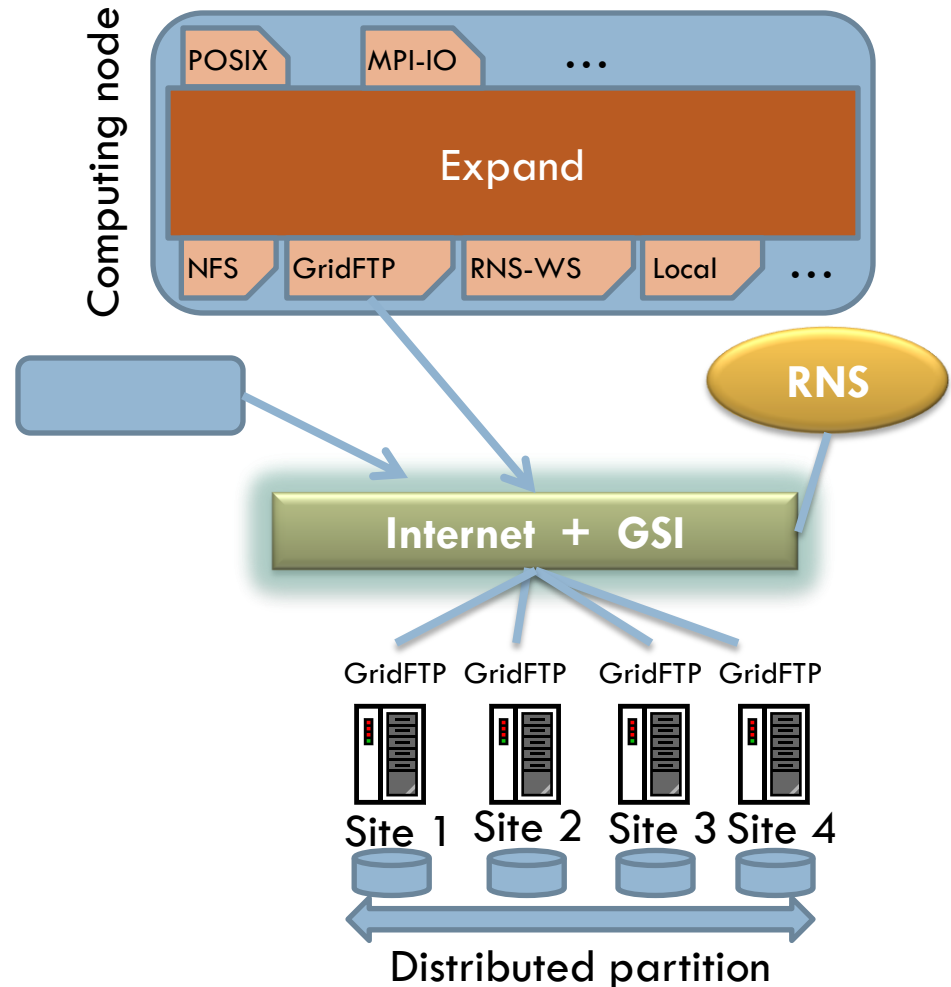
- 4 servers.
- Distributed files.
  - ▣ Each server stores 50 files.
- Complete transfer to local node.
- Application accesses local copy.



# Scenario 2 (Expand)

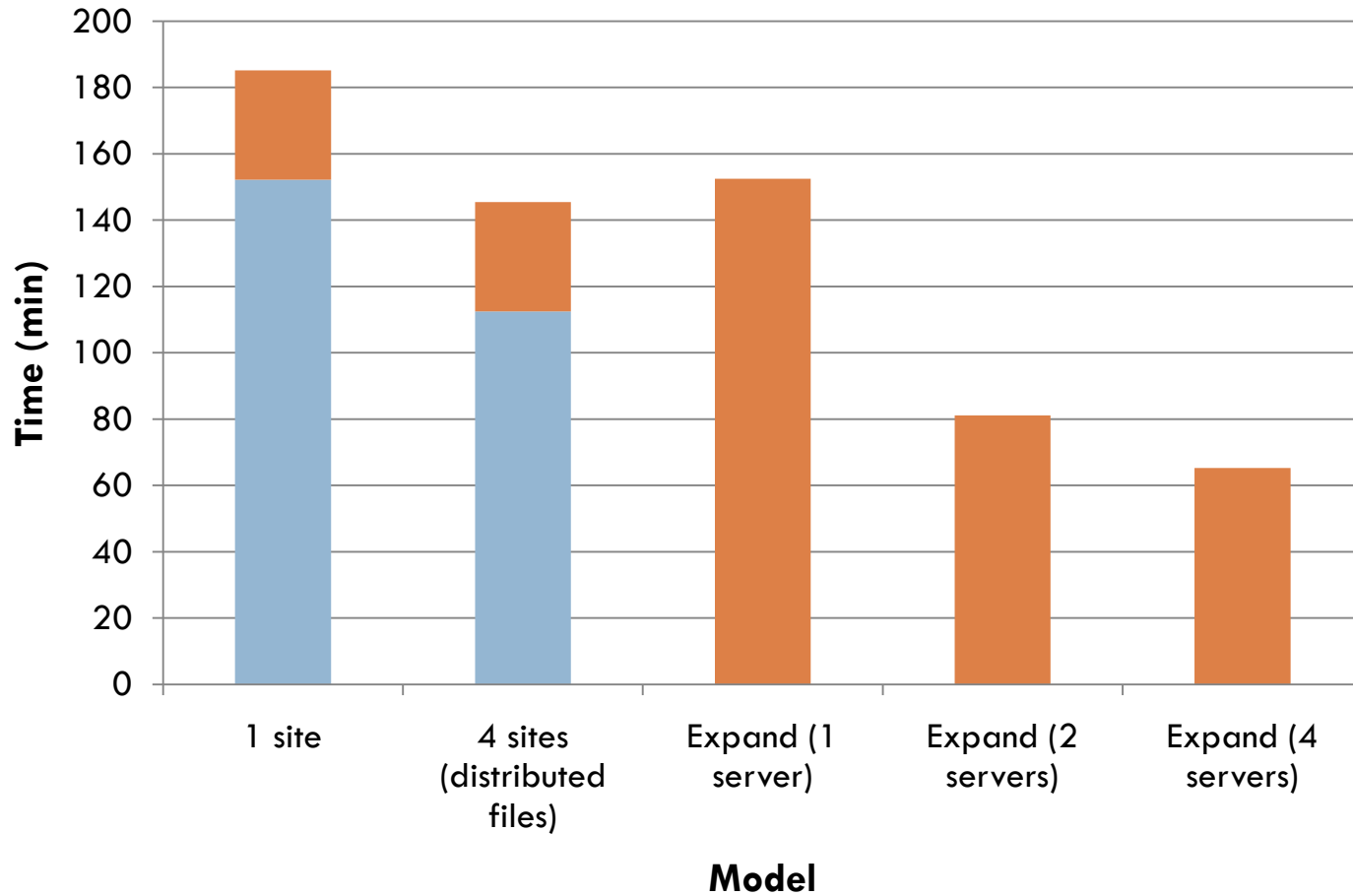
57

- Expand with 1, 2 and 4 servers.
- Local node accesses remotely needed data.
  - No previous transfer needed.



# Grid Evaluation

58



**The Expand Parallel File System**

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena

# Contents

59

- The ARCOS Group.
- Expand motivation.
- Expand design.
- Expand evaluation.
- **Conclusions.**
- Ongoing Work.

# Conclusions

60

- It is feasible to build parallel file system by using standard protocols and servers.
- Our solution is easily adaptable to different environments/situations (cluster and grid are examples).
- Performance results are comparable to other solutions (even comercial).

# Contents

61

- The ARCOS Group.
- Expand motivation.
- Expand design.
- Expand evaluation.
- Conclusions.
- **Ongoing Work.**

# Ongoing work

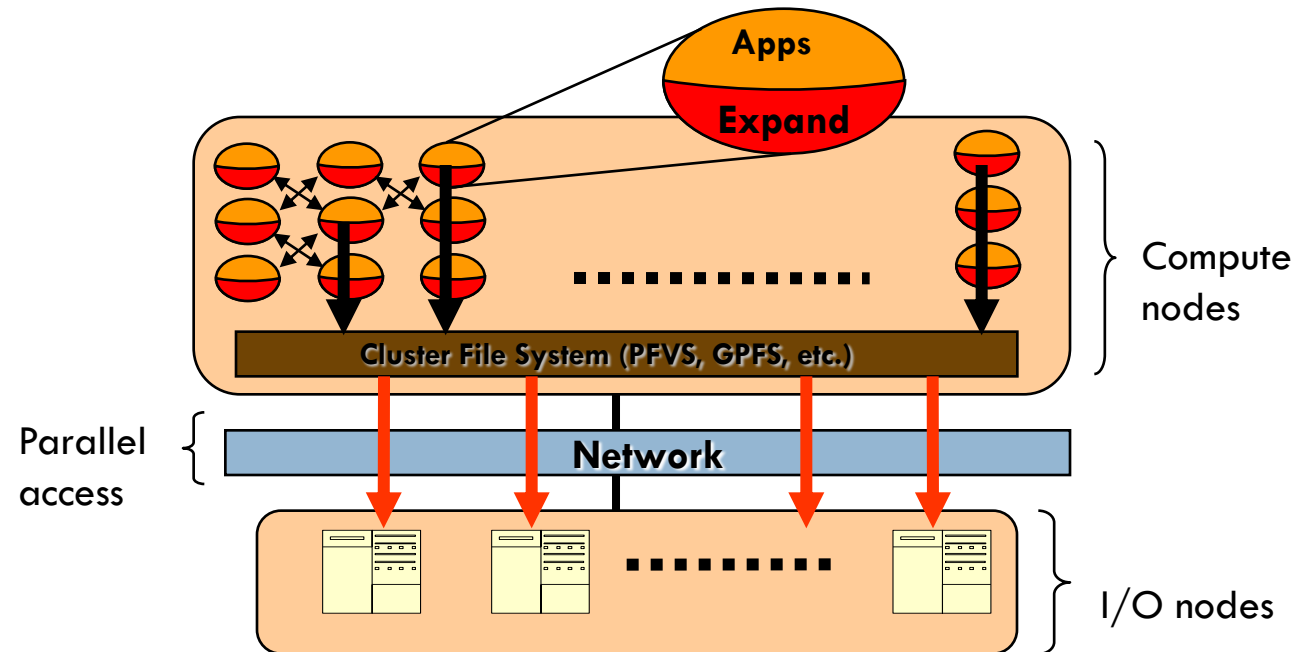
62

- Add new protocols (e.g. Web Services)
- Evaluation in large clusters and grid environments.
- Use Expand to improve performance when accessing replicated data.

# Ongoing work

63

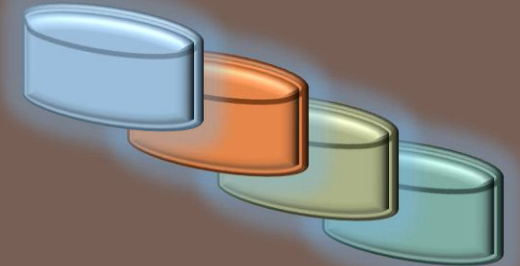
- Use Expand as intermediate file system in large clusters.



**The Expand Parallel File System**

José Daniel García Sánchez – ARCOS Group – University Carlos III of Madrid

July 2007 - University of Modena



# THE **EXPAND** PARALLEL FILE SYSTEM

A FILE SYSTEM FOR CLUSTER AND GRID COMPUTING



José Daniel García Sánchez  
ARCOS Group – University Carlos III of Madrid

