Universitat Autònoma de Barcelona

Computer Architecture & Operating Systems

Methodology for Performance Evaluation of the Input/Output System on Computer Clusters

Speaker: Carlos Núñez

Authors: Sandra Méndez, Dolores Rexachs, Emilio Luque
Contact: sandra.mendez@uab.es

September 30, 2011

IEEE CLUSTER 2011

INTERNATIONAL CONFERENCE ON CLUSTER COMPUTING
AUSTIN, TEXAS - SEPTEMBER 26-30, 2011

Outline

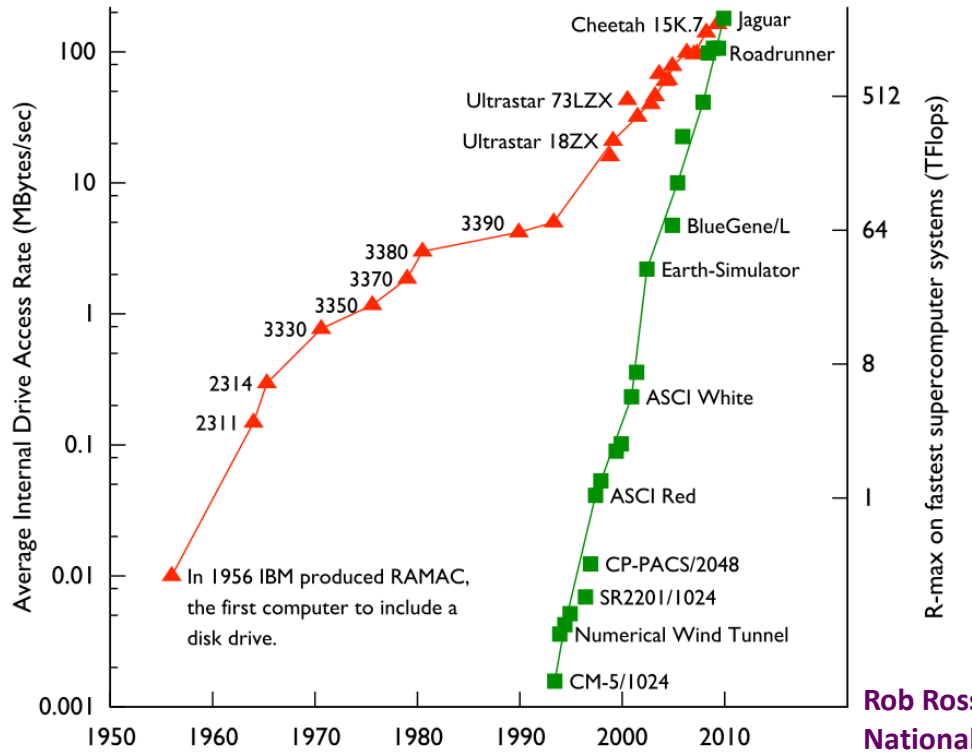
- Motivation
- Objective
- Proposed Methodology
- Experimentation
- Conclusions and Future Work

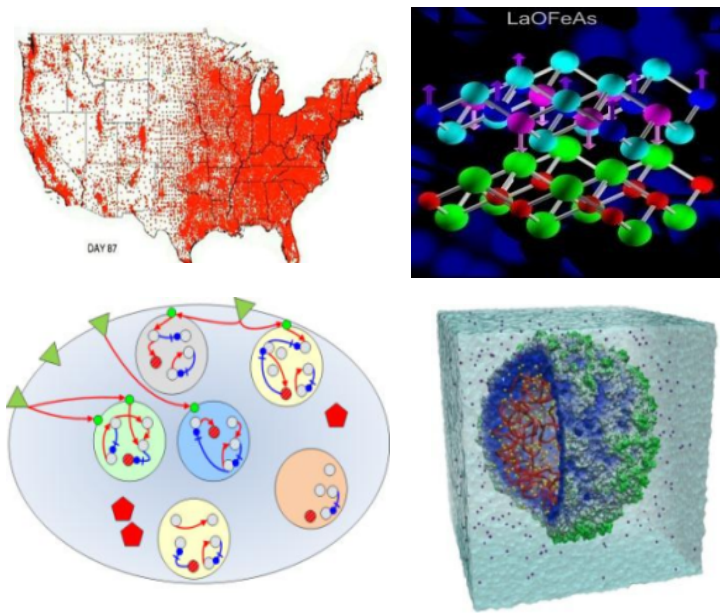
Motivation

- **Motivation**
- Objetivo
- Proposed Methodology
- Experimentation
- Conclusions and Future Work

Motivation

- The historical gap the computing and I/O performance
- Increase in processing units
- Advance in speed and computational power
- The increasing complexity of scientific applications

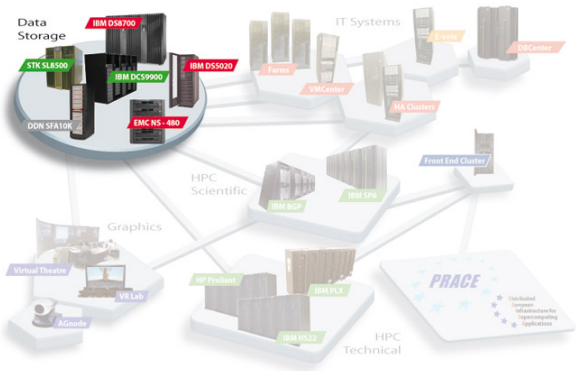


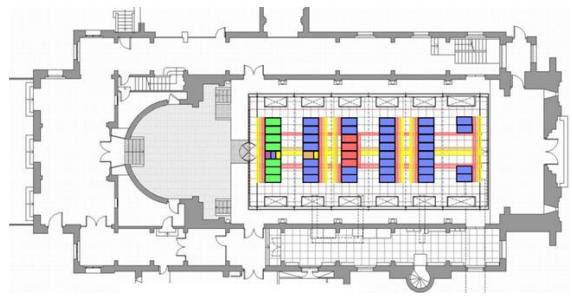


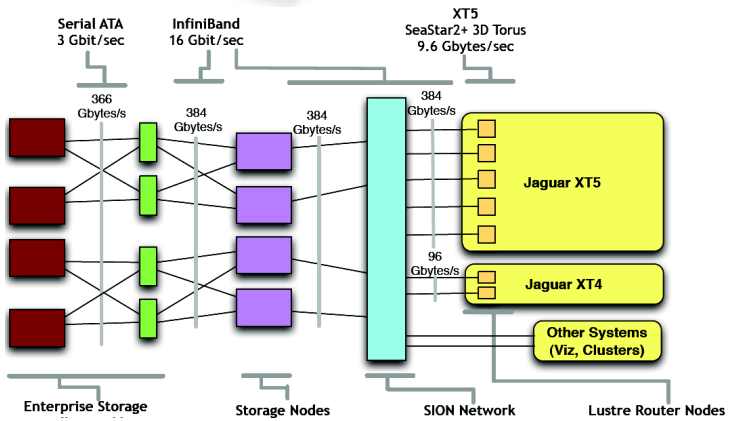
Rob Ross; Mathematics and Computer Science Division; Argonne National Laboratory

Motivation

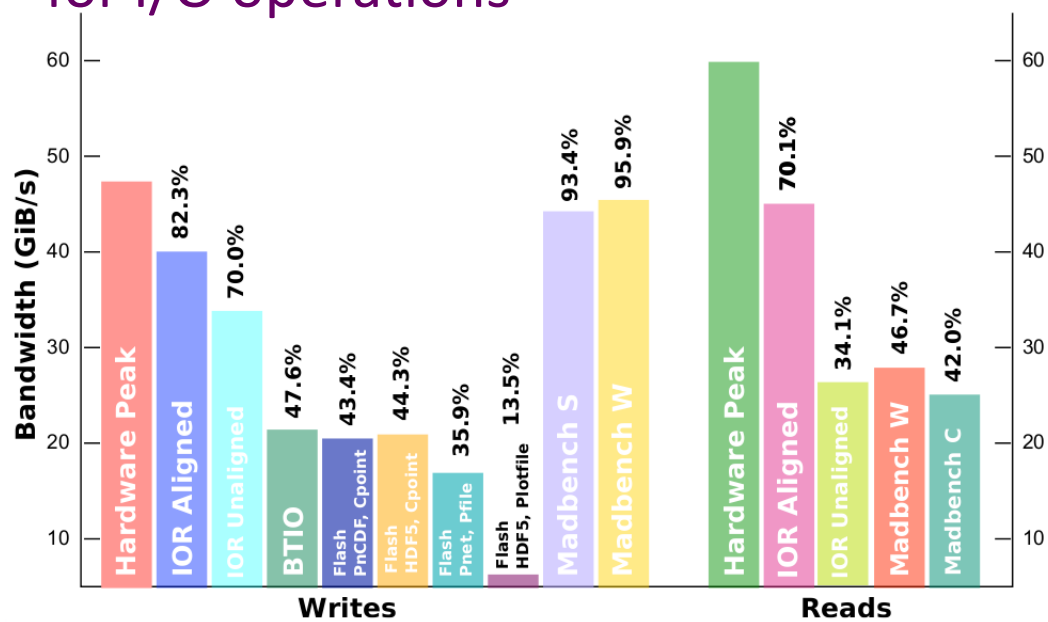
- Different Input/Output Architectures







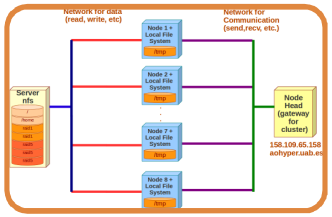
- Different Input/Output performance for I/O operations



Objective

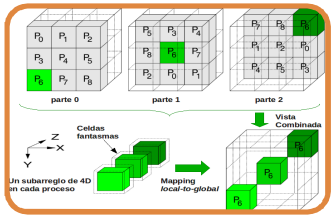
- Motivation
- **Objective**
- Proposed Methodology
- Experimentation
- Conclusions and Future Work

Objective

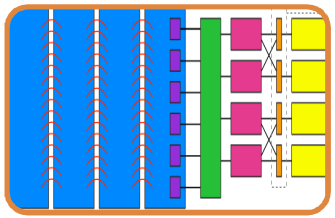


Propose a methodology to evaluate and analyze of the Input/Output system performance of computer clusters.

Context



I/O intensive applications



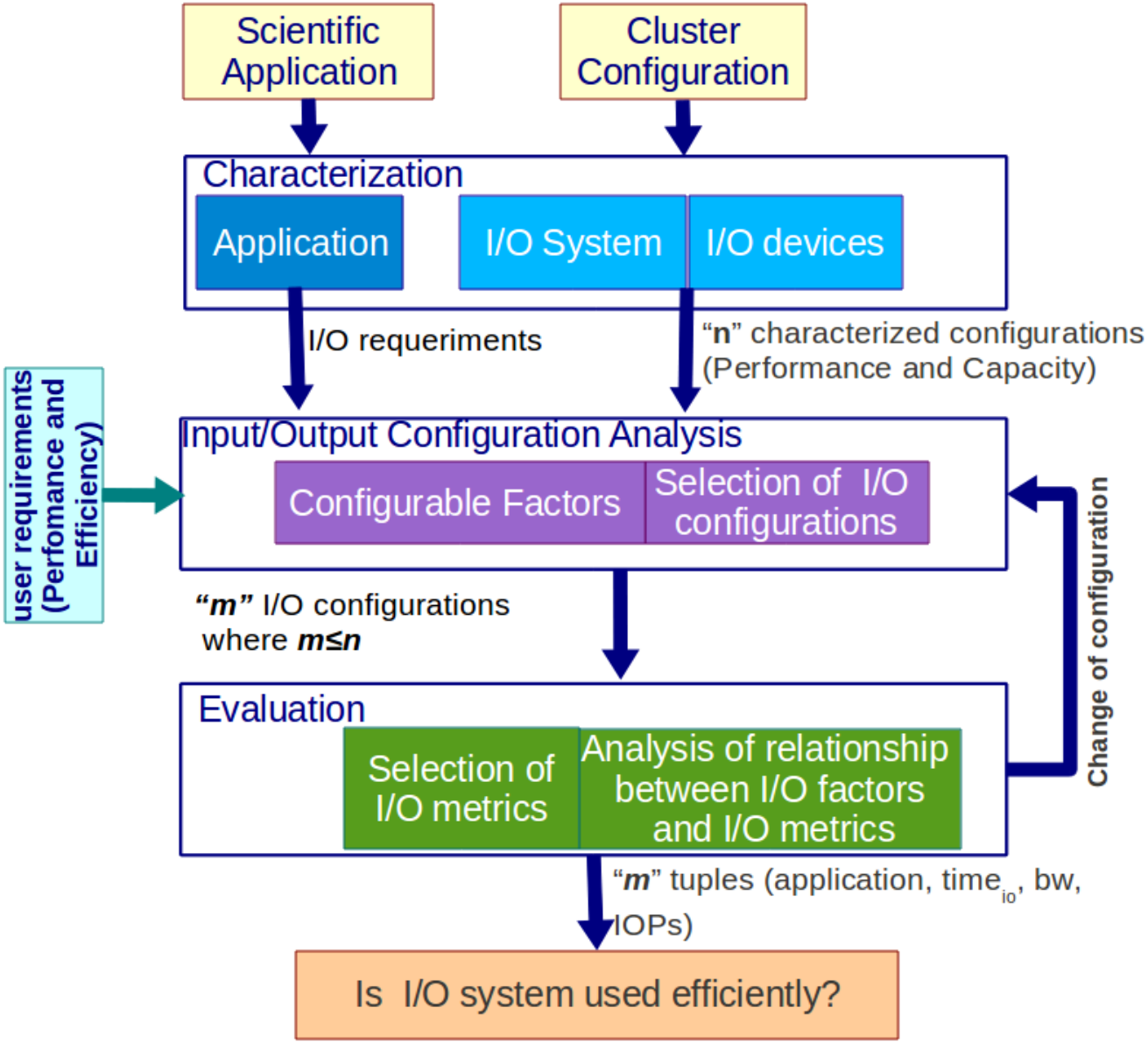
Focused on I/O architecture:

Filesystem, Network, RAID Level, I/O node, Buffering/cache.

Proposed Methodology

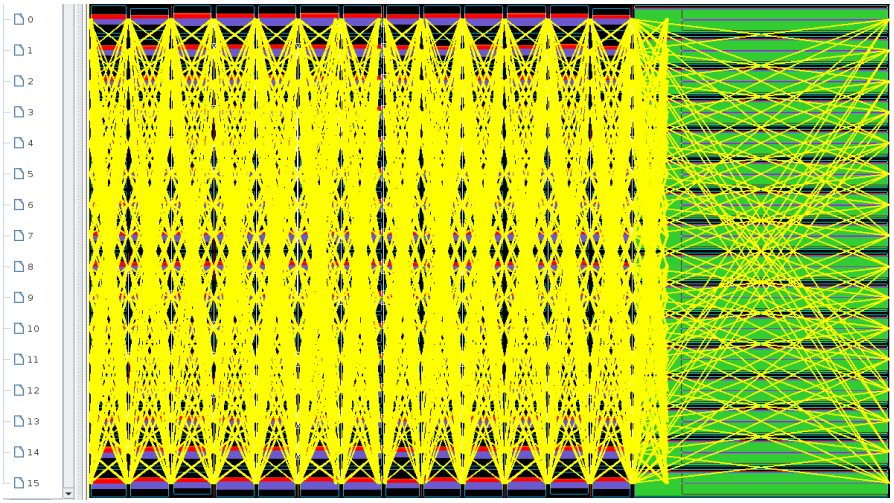
- Motivation
- Objetivo
- **Proposed Methodology**
- Experimentation
- Conclusions and Future Work

Proposed Methodology



Proposed Methodology

Scientific Application | Characterization



What?
- Access Patterns
- I/O Phases
- Workload
- Number of files
- Number of processes
- Data mapping
- I/O library

How?
- Tracing Tools: mpe, DARSHAN, PAS2P
- Linux System tools (vmstat)
- Application Documentation

```
Rank MPI-IO #OP …
0 MPI_File_delete 0 1 0 0.468036 0.161379 0
0 MPI_File_open 0 2 9 0.479119 0.000062 0
0 MPI_File_write_at_all 265302 269 10612080 12.825201 0.218666 0
0 MPI_File_read_at_all 265302 4872 10612080 224.882046 0.230365 0
```

Monitored Information

Direct
- Number of files
- Number of writing operations
- Block size for writing operations
- Number of reading operations
- Block size for reading operations
- Number of open call

Indirect
- I/O Phases
- Number of processes in I/O operations
- Access Type
- Access Mode

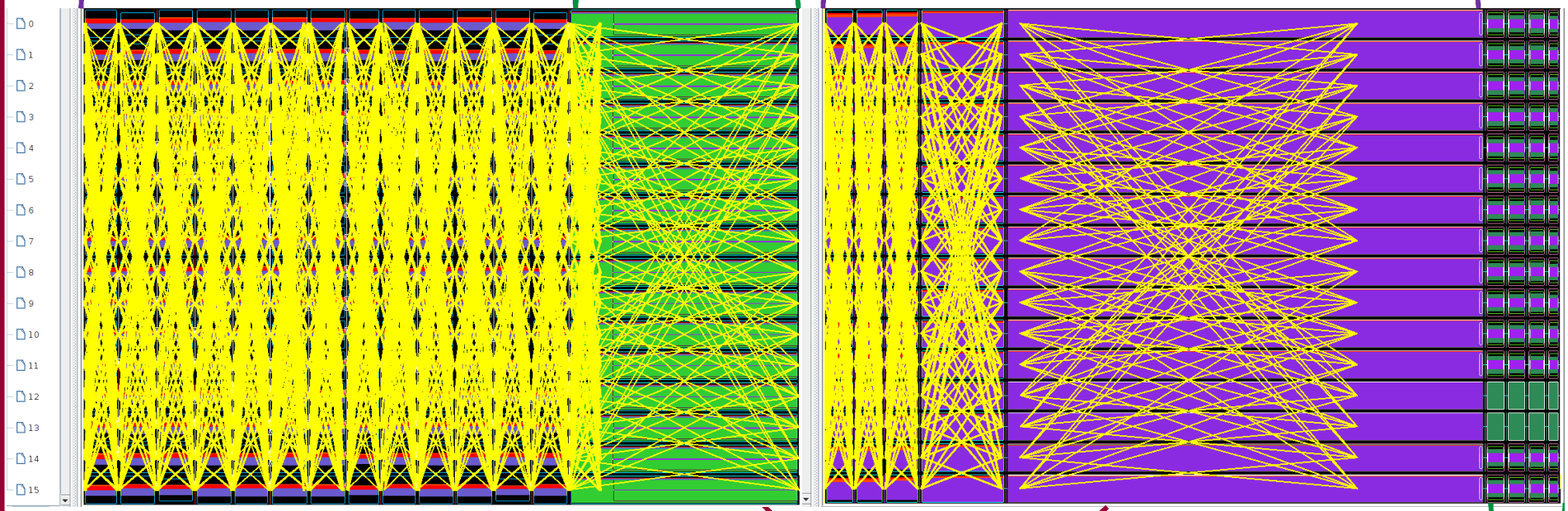
Proposed Methodology

Scientific Application | Characterization

Write phase + compute + communication

Read phase + communication

Write phase + compute + communication

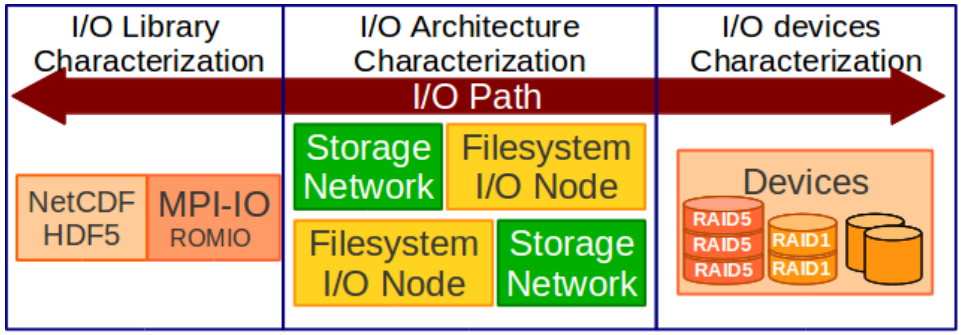


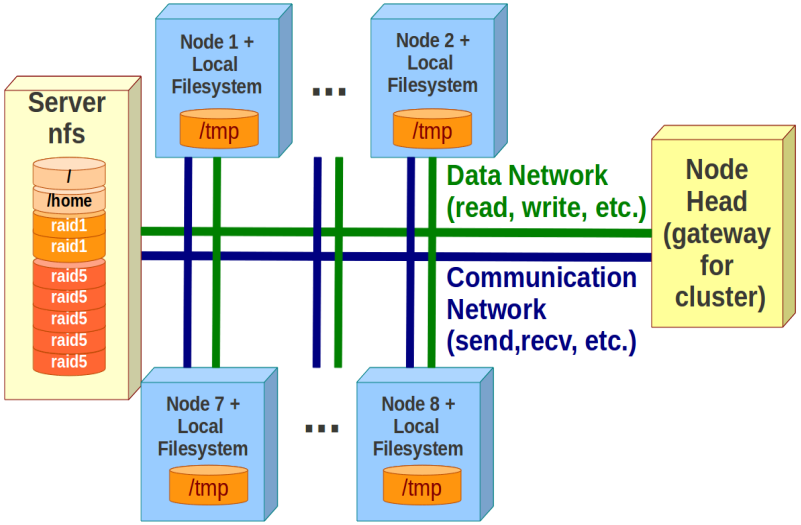
Parameters	full 16p	simple 16p
$numFiles$	1	1
$numIO_{read}$	640	2,073,600 and 2,125,440
$numIO_{write}$	640	2,073,600 and 2,125,440
bk_{read}	10 MB	1.56KB and 1.6KB
bk_{write}	10 MB	1.56KB and 1.6KB
$numIO_{open}$	32	32
$accessType$	Global	Global
$accessMode$	Sequential	Sequential
$numProcesos$	16	16

Read phase + Write phase + communication

Proposed Methodology

Cluster Configuration | Characterization

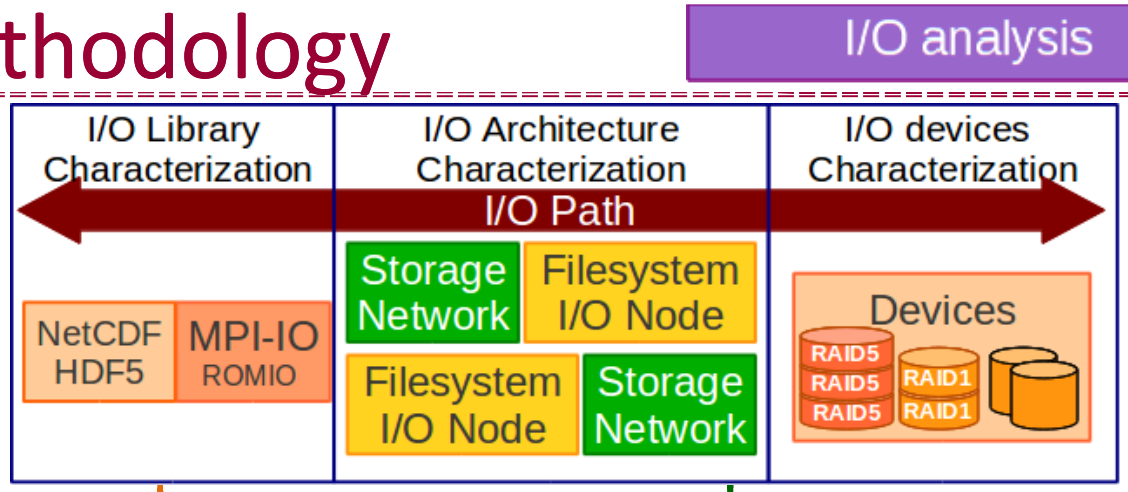




I/O Architecture

- **Connection:** I/O-compute node, I/O node I/O network, shared network.
- **Management:** centralized, distributed, hybrid (depending of filesystem)
- **Placement:** the position of the I/O nodes or devices in the network topology (DAS, NAS, NASD, SAN)
- **Buffering/Cache:** placement, size and state
- **Availability:** data level (RAID) and service level (replication)

Proposed Methodology



- I/O library benchmark (b_eff_io, IOR)
- Filesystem and disk Benchmarks (bonnie++, iozone)
- Linux tools (iostat, hdparm)

Selectable or configurable factors

- Number of filesystem
- Filesystem Type
- Number of networks
- Network Type
- Buffer/Cache State
- Buffer/Cache placement
- Number of I/O node
- I/O node placement
- RAID level
- Number of devices on RAID

1st I/O configuration

On each I/O path level:

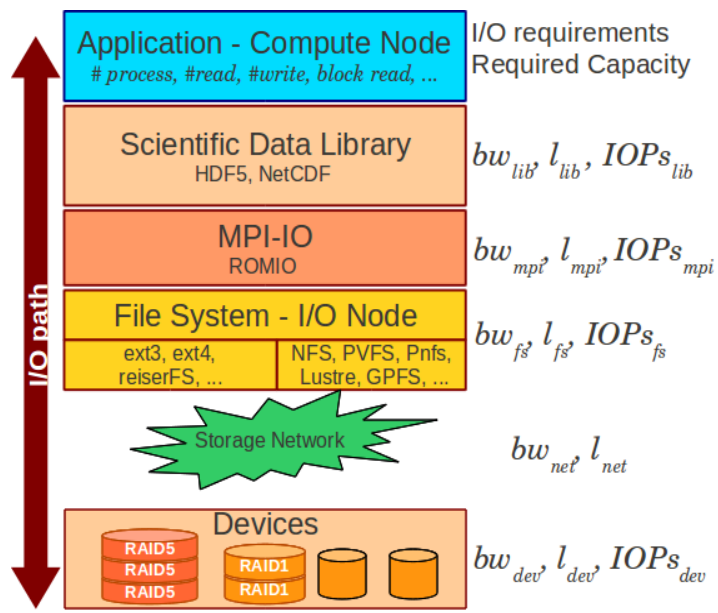
- Performance:
  - Bandwidth
  - IOPs
  - Latency
- Capacity:
  - Type devices
  - Storage (Gbytes)

2nd I/O configuration

.

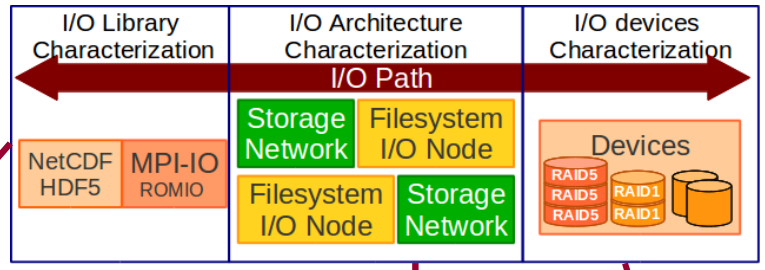
.

N-th I/O configuration



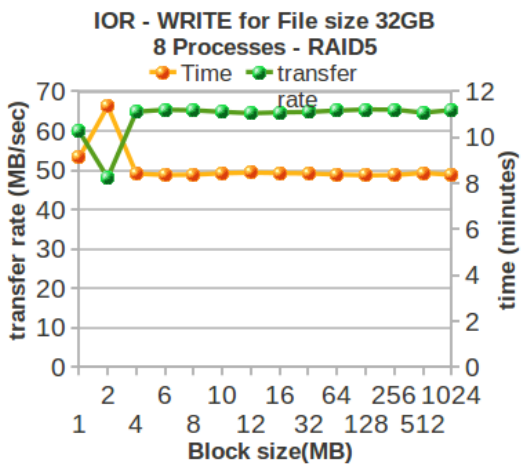
Proposed Methodology

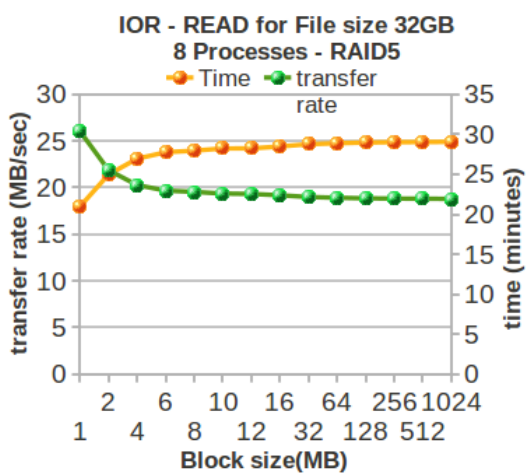
I/O analysis

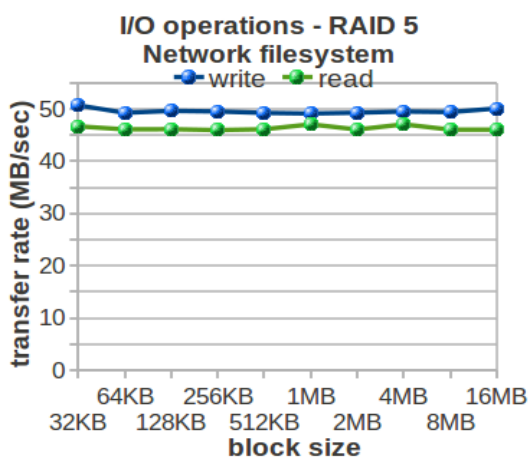


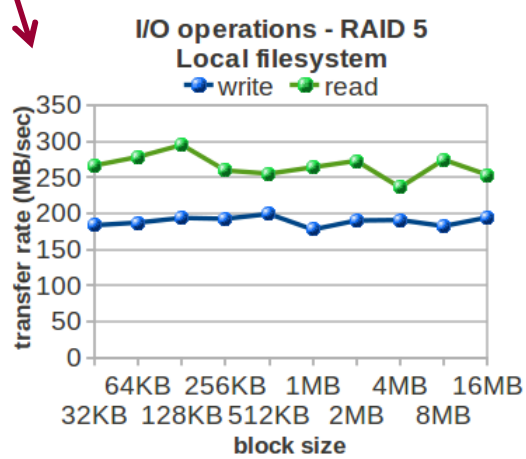
Data structure of I/O system performance

- Operation Type (enumerate {0 (read), 1 (write)})
- Block size (double (MBytes))
- Access Type (enumerate {0 (Local), 1 (Global)})
- Accesses Mode (enumerate {0 (Sequential), 1 (Strided), 2 (Random)})
- **transfer Rate** (double (MBytes/second))
- Latency (double (microsecond))
- IOPs (integer)









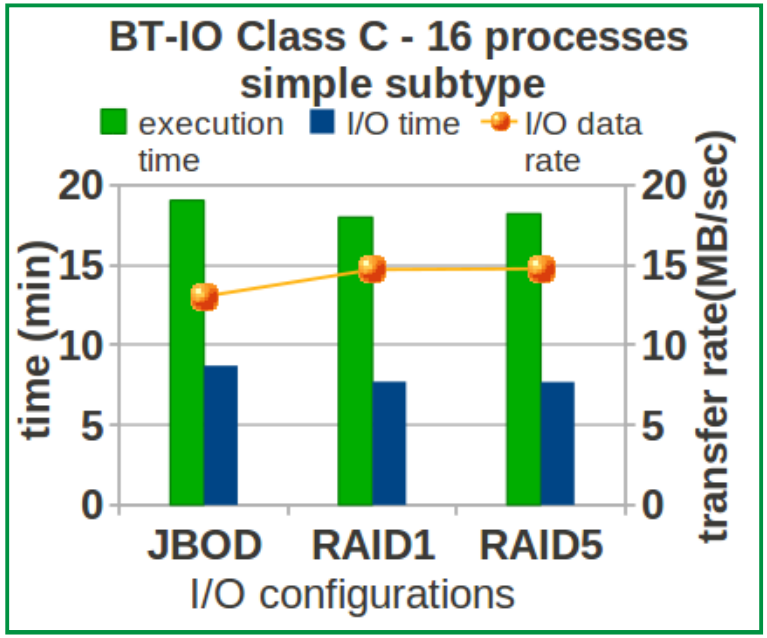
Proposed Methodology

Evaluation

Metrics

- Run time
- I/O time
- Input/output operation per second
- throughput (megabytes per second of data transfers)

I/O level	Read	Write	Read	Write
Library (MB/sec)	28	48	29	48
Global Filesystem (MB/sec)	48	50	47	50
Local Filesystem (MB/sec)	133	113	273	193



Characterization Values (Bandwidth, IOPs and Latency)

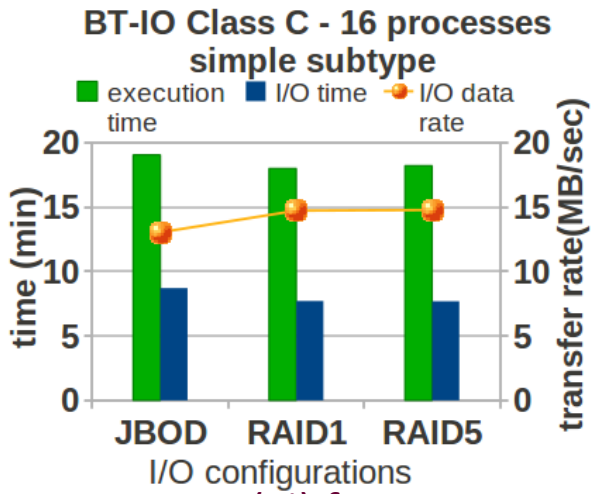
- File System
- Network Interconnection
- I/O Library
- Devices

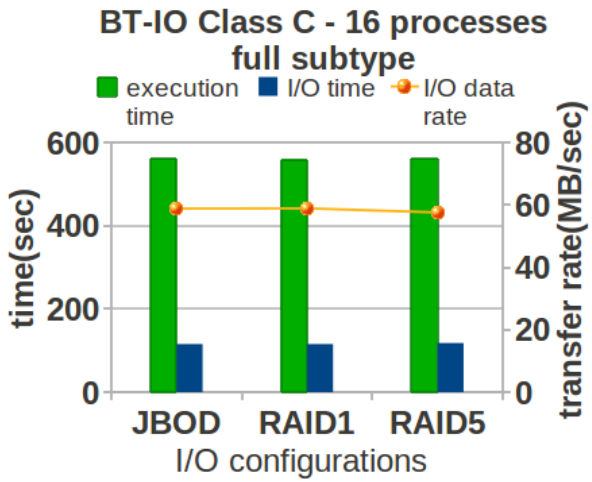
$$\%\ IO\ Level\ Use = \frac{bandwidth(Application)}{BW(IO\ Level)} * 100$$

A table of the I/O system use for each I/O configuration

Proposed Methodology

Evaluation





I/O system use (%) for writing operations

I/O configuration	I/O Lib	NFS	Local FS	SUBTYPE
JBOD	101.47	117.70	78.00	FULL
RAID1	140.24	120.20	54.04	FULL
RAID5	88.60	115.18	29.69	FULL
JBOD	25.06	26.06	15.33	SIMPLE
RAID1	27.75	30.65	13.37	SIMPLE
RAID5	24.60	29.52	8.07	SIMPLE

I/O system use (%) for reading operations

I/O configuration	I/O Lib	NFS	Local FS	SUBTYPE
JBOD	309.74	127.93	60.00	FULL
RAID1	310.00	128.04	43.63	FULL
RAID5	303.11	125.20	22.76	FULL
JBOD	54.29	28.96	18.61	SIMPLE
RAID1	54.48	31.98	12.68	SIMPLE
RAID5	56.77	31.40	5.55	SIMPLE

Parameters	full 16p	simple 16p
$numFiles$	1	1
$numIO_{read}$	640	2,073,600 and 2,125,440
$numIO_{write}$	640	2,073,600 and 2,125,440
bk_{read}	10 MB	1.56KB and 1.6KB
bk_{write}	10 MB	1.56KB and 1.6KB
$numIO_{open}$	32	32
$accessType$	Global	Global
$accessMode$	Sequential	Sequential
$numProcesos$	16	16

Experimentation

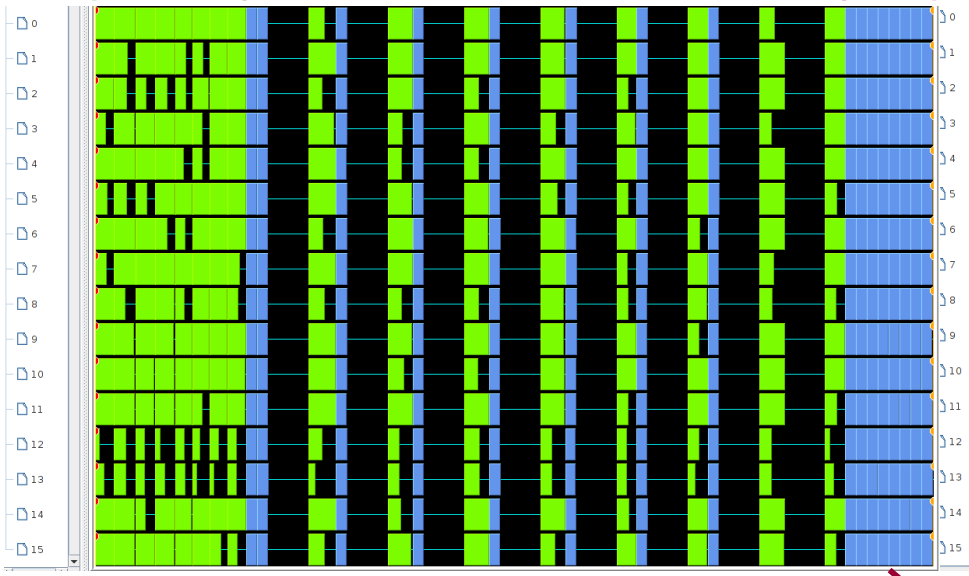
- Motivation
- Objetivo
- Proposed Methodology
- **Experimentation**
- Conclusions and Future Work

Experimentation

MadBench2 | Scientific Application | Characterization

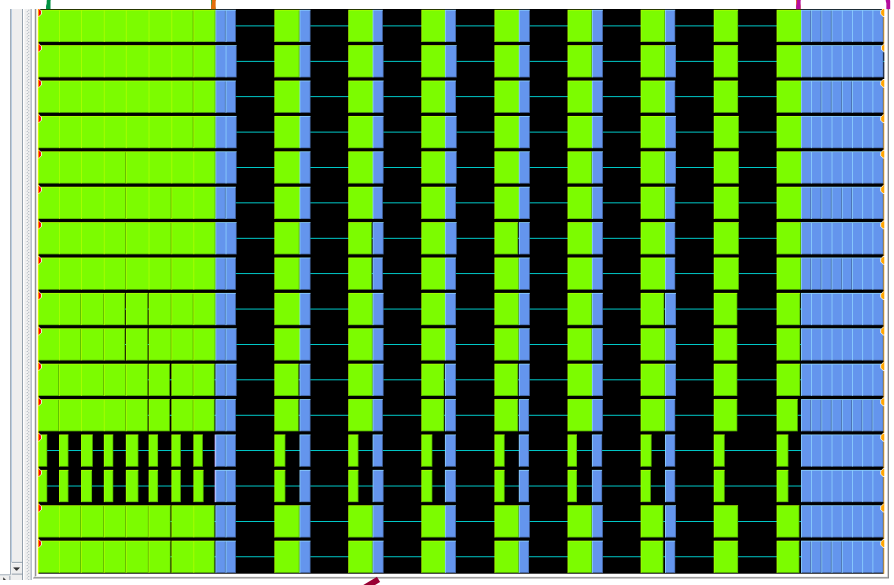
UNIQUE filetype

S_w (Write phase) | W_r (Read phase) | W_w (Write phase) | C_r (Read phase)



SHARED filetype

S_w (Write phase) | W_r (Read phase) | W_w (Write phase) | C_r (Read phase)



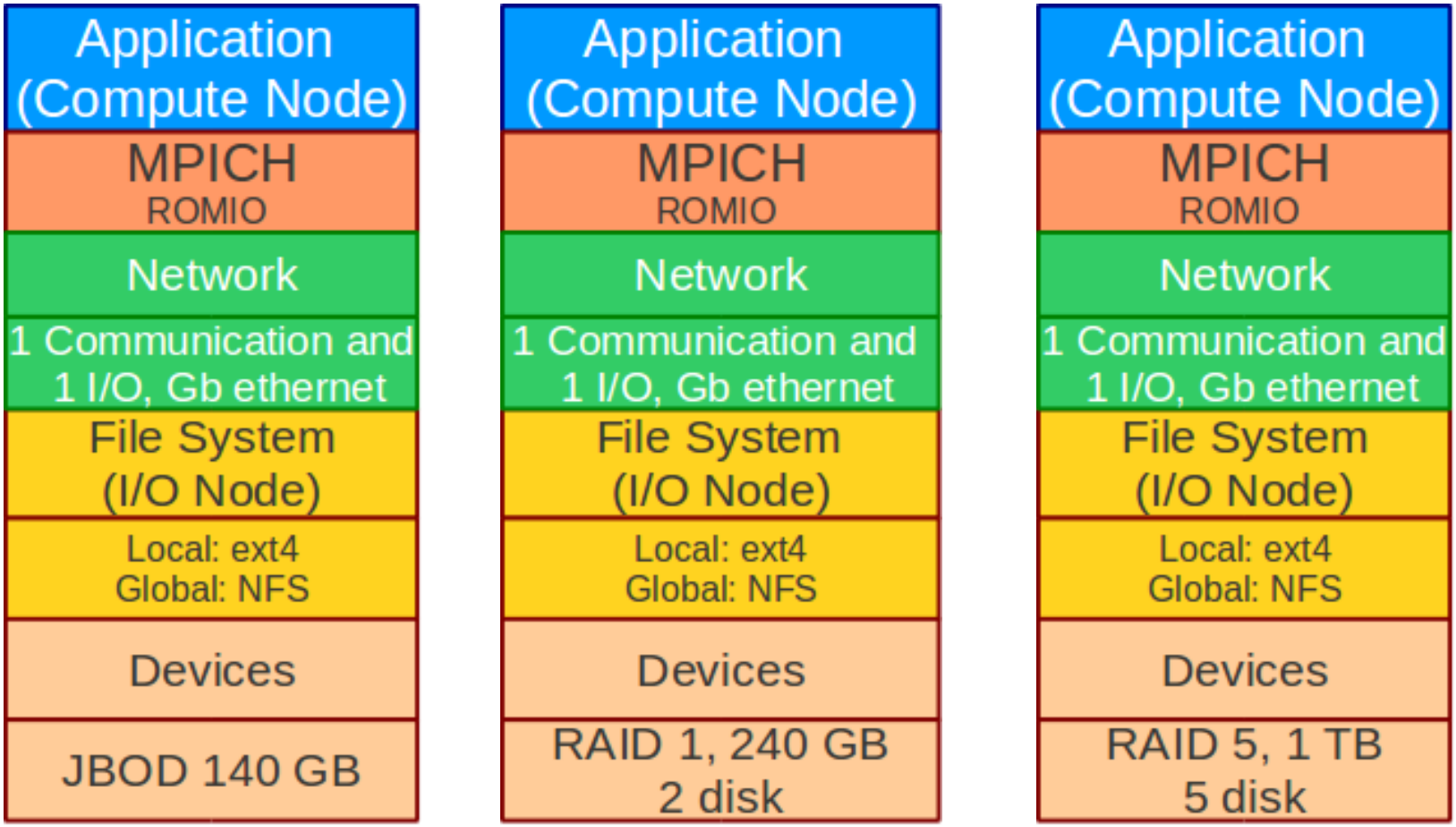
Parameters	UNIQUE	SHARED
$numFiles$	16	1
$numIO_{read}$	16 x file	256
$numIO_{write}$	16 x file	256
bk_{read}	162 MB	162 MB
bk_{write}	162 MB	162 MB
$numIO_{open}$	16	16
$numIO_{close}$	16	16
$numProcesos$	16	16

Experimentation

Cluster Aohyper

Characterization

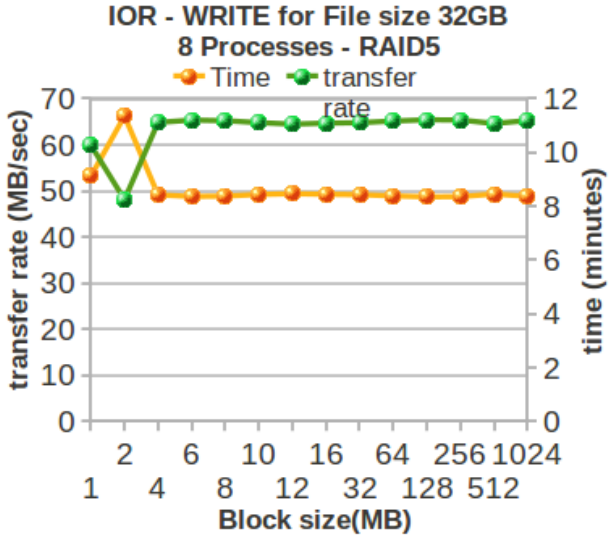
•8 nodes AMD Athlon(tm) 64 X2 Dual Core Processor 3800+, 2GB RAM memory, 150GB local disk. Local filesystem is linux ext4 and global filesystem is NFS.

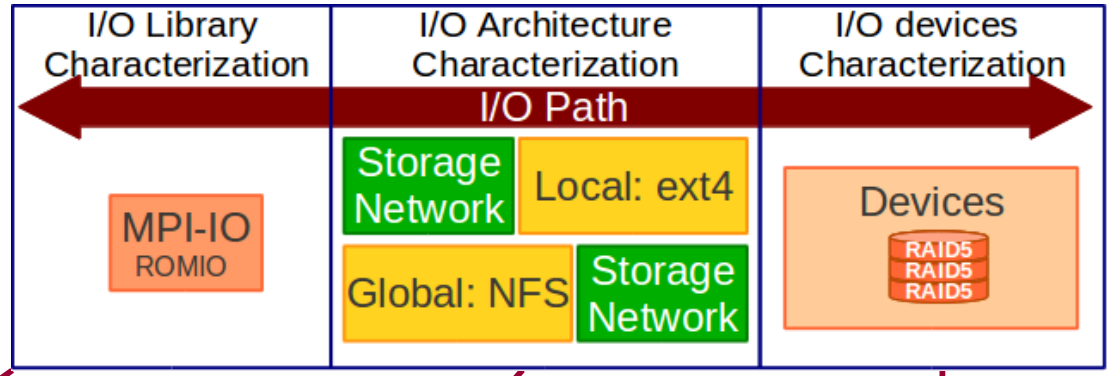


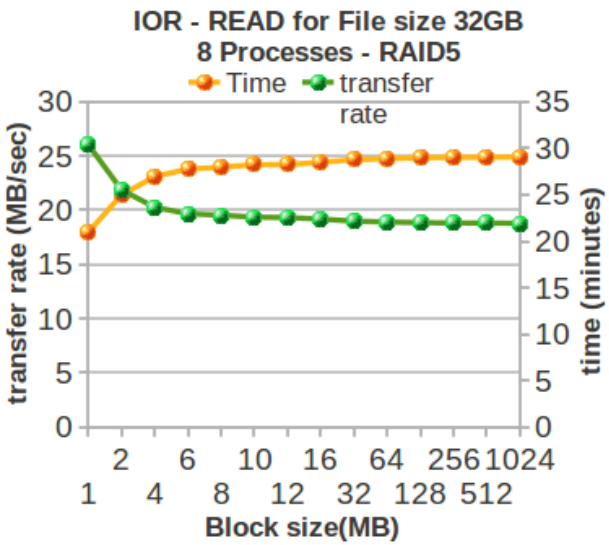
Experimentation

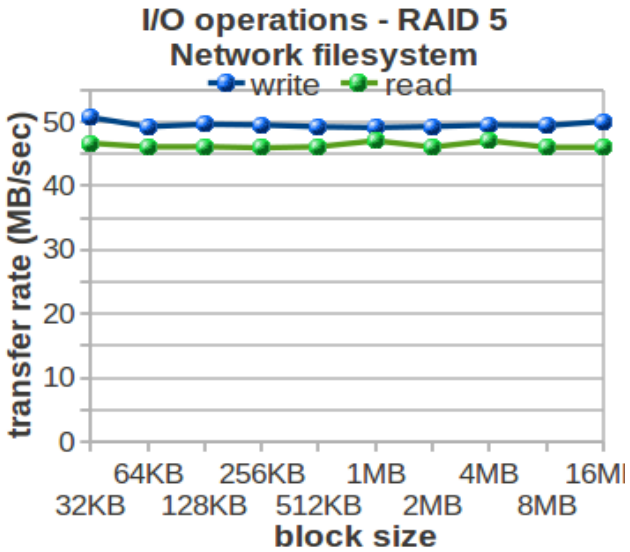
Aohyper – RAID5

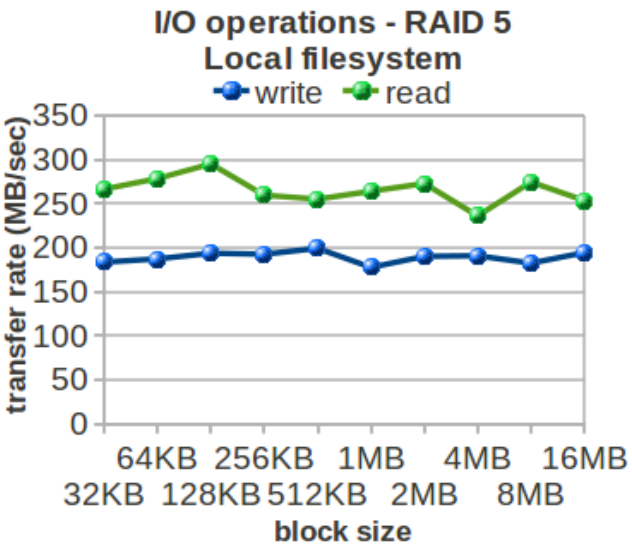
I/O analysis







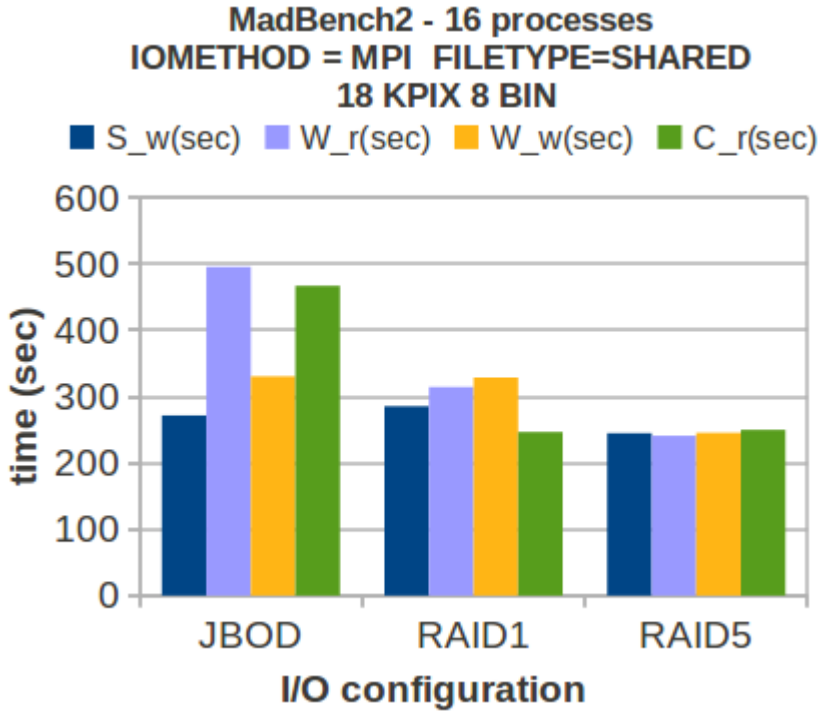


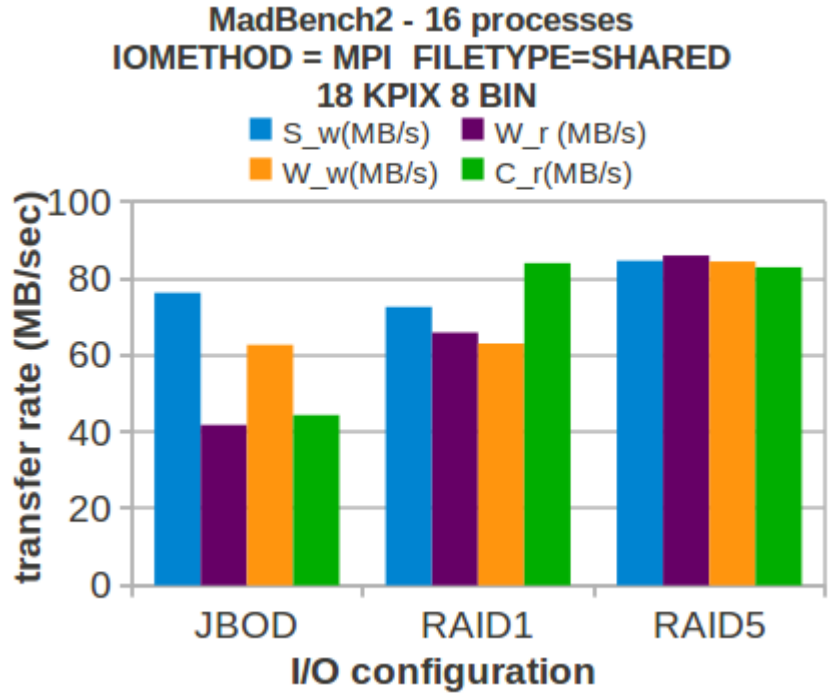


Experimentation

MadBench2 SHARED - Aohyper

Evaluation





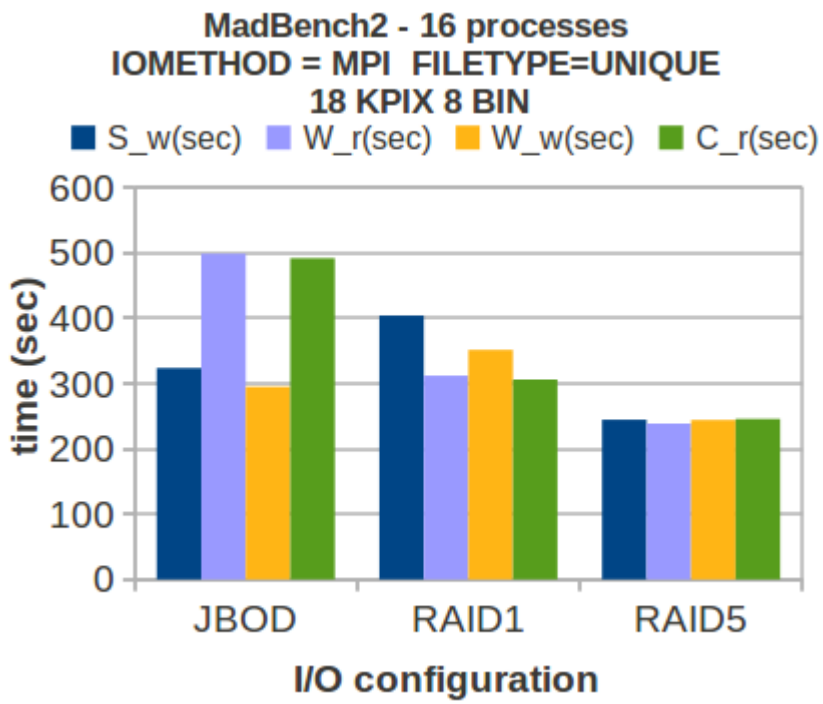
PERCENTAGE (%) OF USE FOR MADBENCH2 ON LOCAL FILESYSTEM

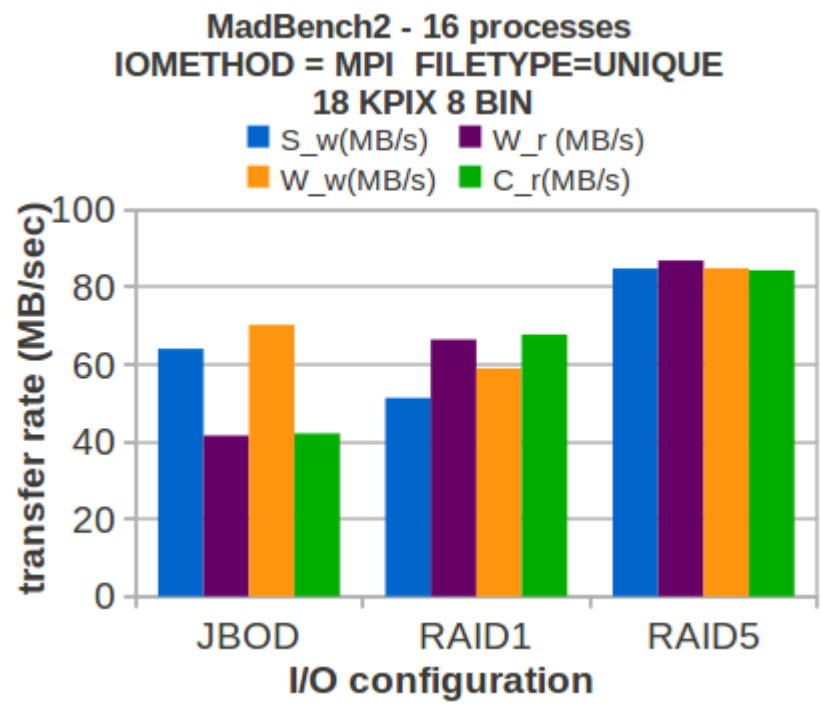
I/O configuration	W_r	C_r	S_w	W_w	FILETYPE
JBOD	67.44	71.59	95.36	78.39	SHARED
RAID1	44.53	56.78	49.06	42.61	SHARED
RAID5	33.09	31.91	32.57	32.49	SHARED
JBOD	67.08	67.98	80.08	87.84	UNIQUE
RAID1	44.90	45.76	34.69	39.87	UNIQUE
RAID5	33.42	32.45	32.64	32.65	UNIQUE

Experimentation

MadBench2 UNIQUE - Aohyper

Evaluation





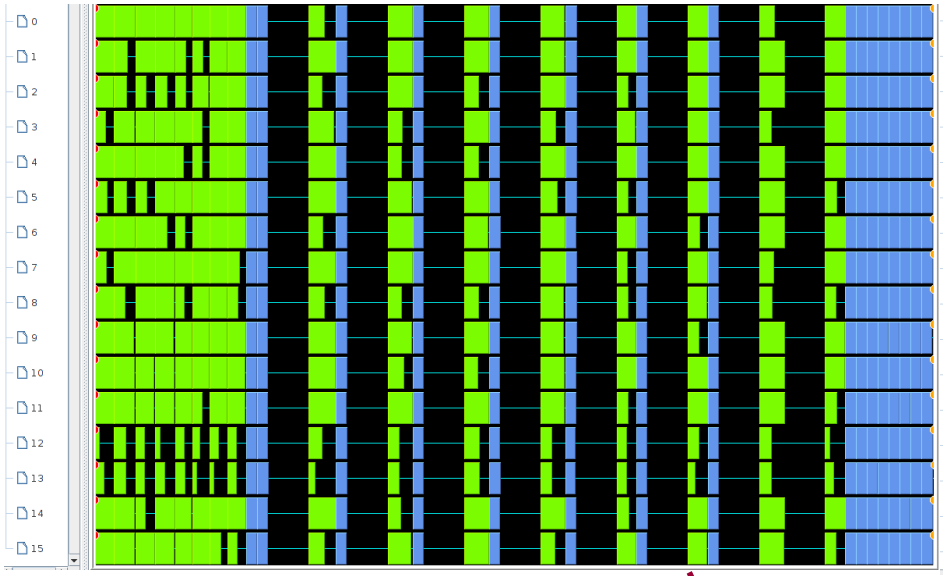
PERCENTAGE (%) OF USE FOR MADBENCH2 ON LOCAL FILESYSTEM

I/O configuration	W_r	C_r	S_w	W_w	FILETYPE
JBOD	67.44	71.59	95.36	78.39	SHARED
RAID1	44.53	56.78	49.06	42.61	SHARED
RAID5	33.09	31.91	32.57	32.49	SHARED
JBOD	67.08	67.98	80.08	87.84	UNIQUE
RAID1	44.90	45.76	34.69	39.87	UNIQUE
RAID5	33.42	32.45	32.64	32.65	UNIQUE

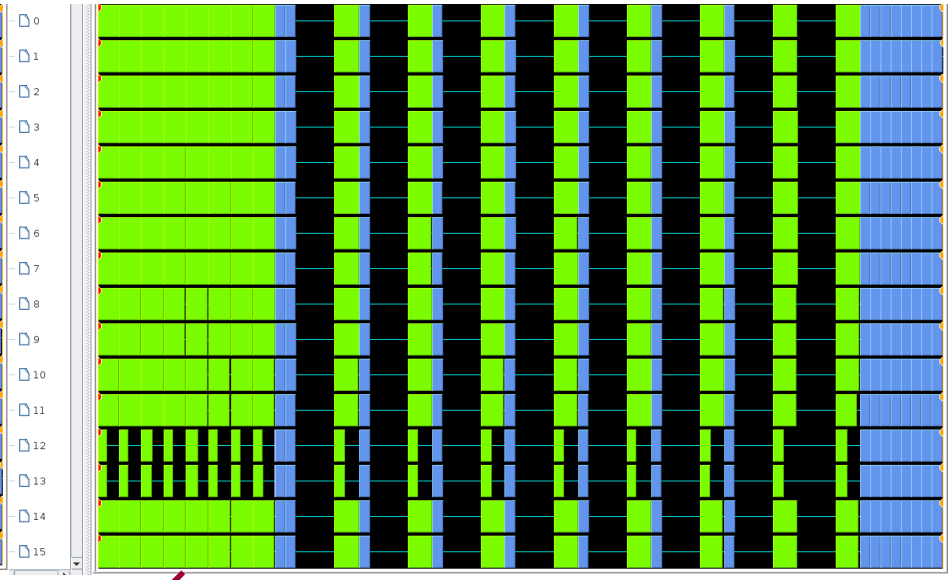
Experimentation

MadBench2 | Scientific Application | Characterization

UNIQUE filetype



SHARED filetype



Parameters	UNIQUE	SHARED
$numFiles$	16	1
$numIO_{read}$	16 x file	256
$numIO_{write}$	16 x file	256
bk_{read}	162 MB	162 MB
bk_{write}	162 MB	162 MB
$numIO_{open}$	16	16
$numIO_{close}$	16	16
$numProcesos$	16	16

64 processes

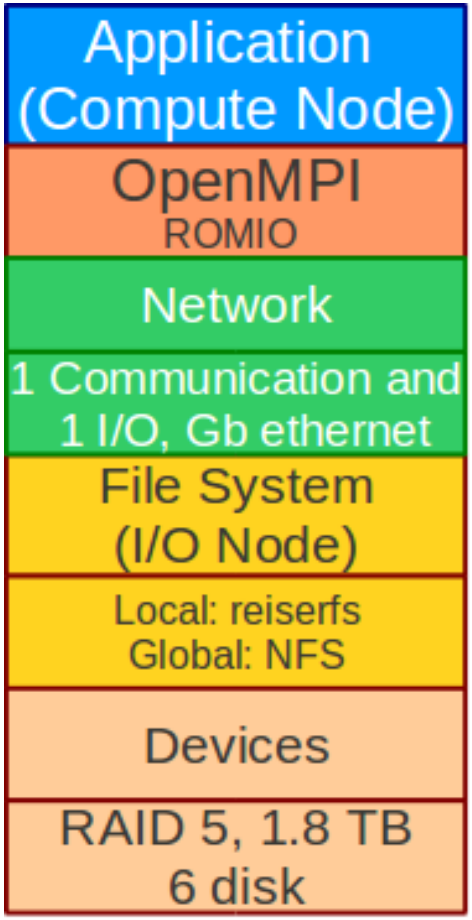
UNIQUE	SHARED
64	1
16 x file	1024
16 x file	1024
40.5 MB	40.5 MB
40.5 MB	40.5 MB
64	64
64	64
64	64

Experimentation

Cluster "A"

Characterization

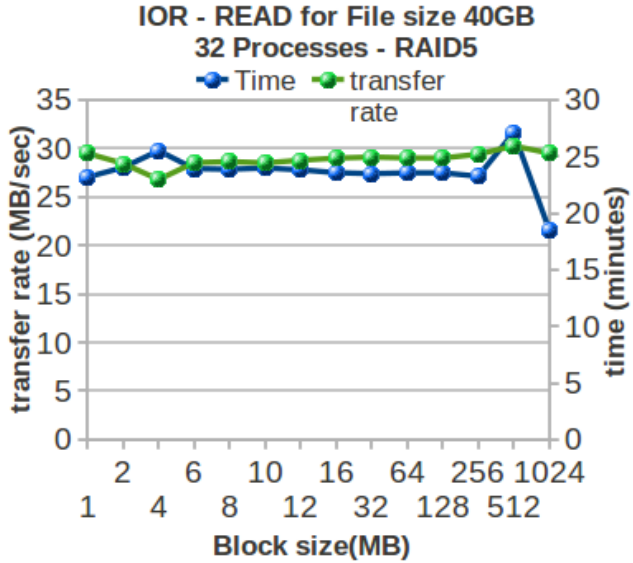
- 32 compute nodes (2 x Dual-Core Intel (R) Xeon (R) 3.00GHz, 12 GB of RAM, and 160 GB SATA disk Dual Gigabit Ethernet)
- A front-end node as NFS server: Dual-Core Intel (R) Xeon (R) 2.66GHz, 8 GB of RAM.

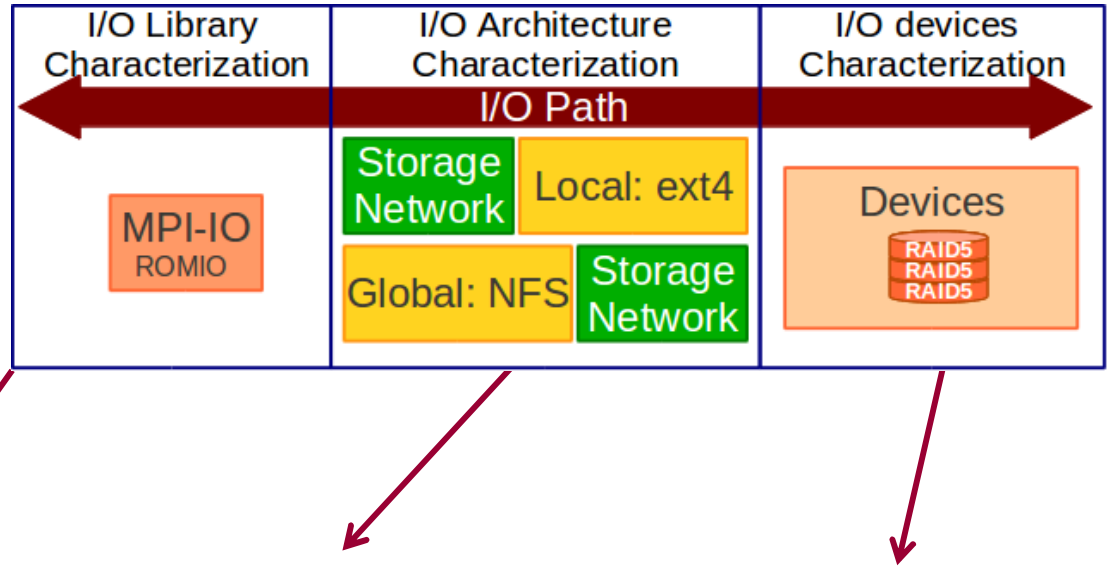


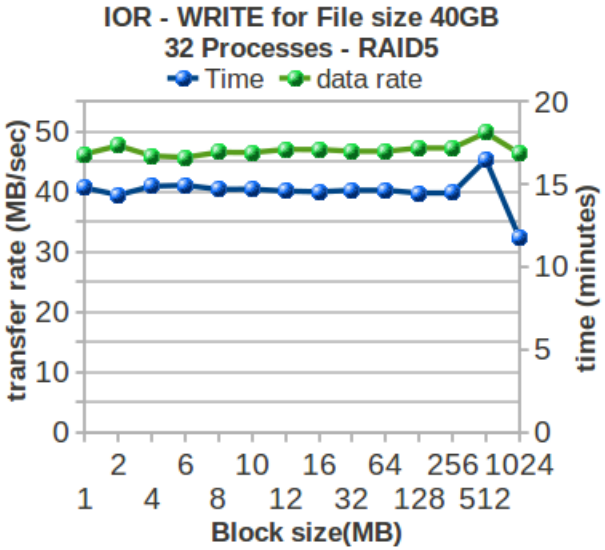
Experimentation

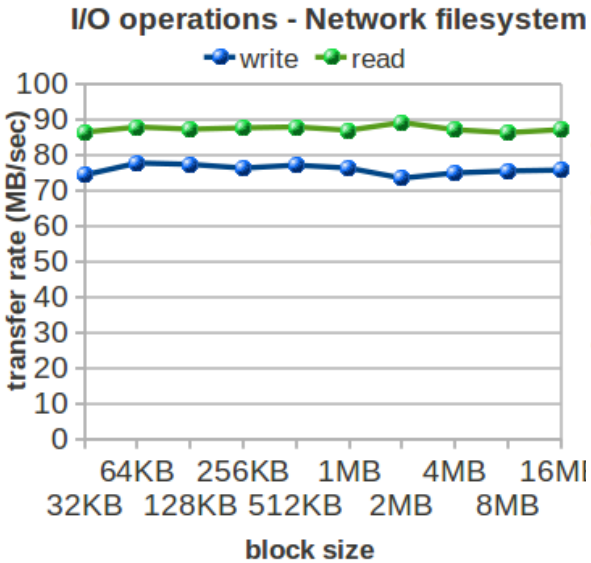
Cluster "A"

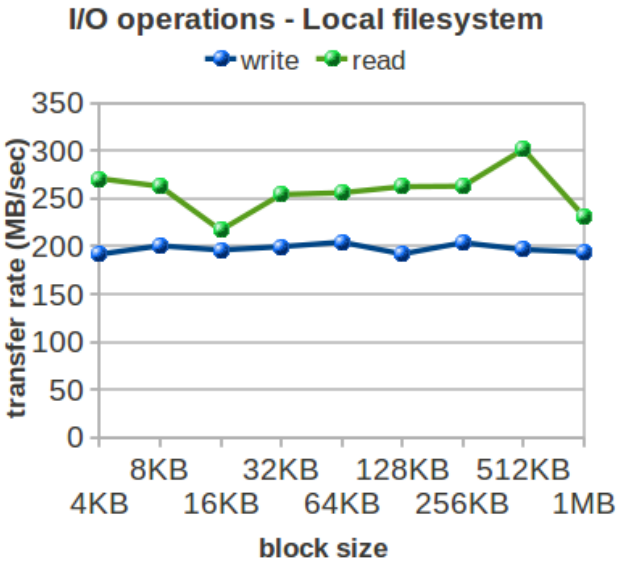
I/O analysis







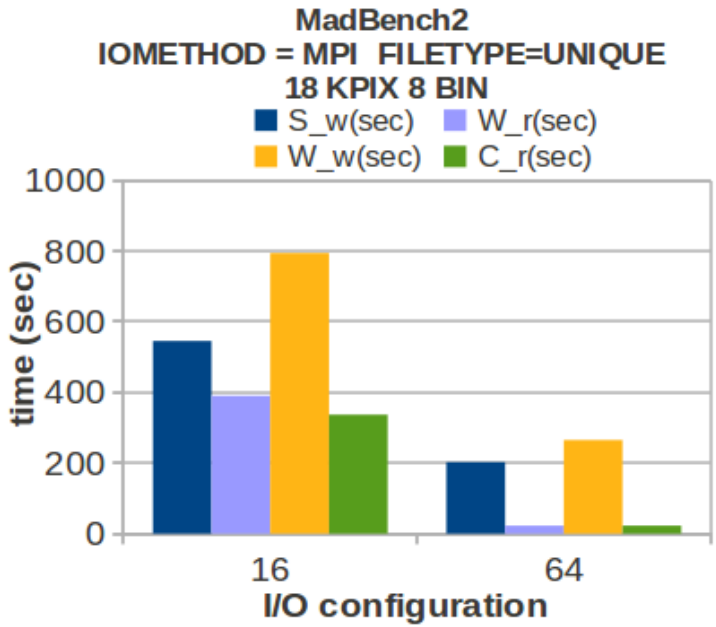


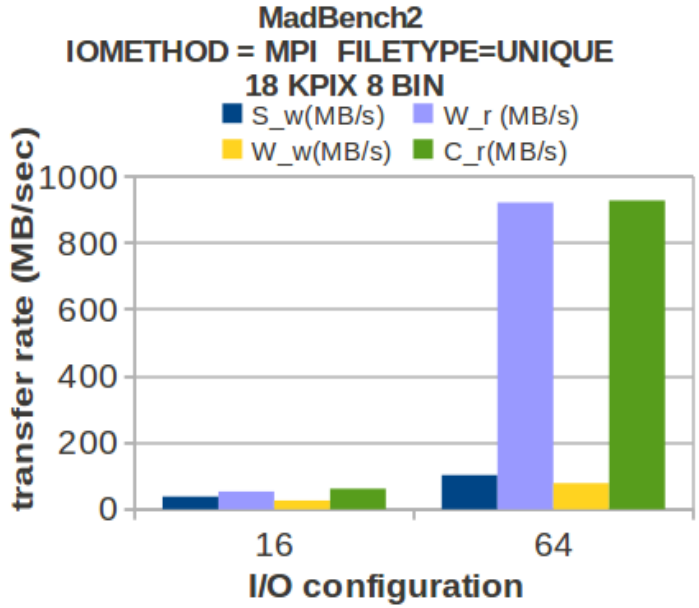


Experimentation

MadBench UNIQUE – "A"

Evaluation





USED PERCENTAGE (%) BY MADBENCH2 ON NETWORK FILESYSTEM

I/O configuration	W_r	C_r	S_w	W_w	FILETYPE
16	32.09	33.61	63.03	42.75	SHARED
64	75.48	78.85	96.41	76.89	SHARED
16	59.01	68.39	54.29	37.28	UNIQUE
64	1024.91	1031.80	146.42	111.79	UNIQUE

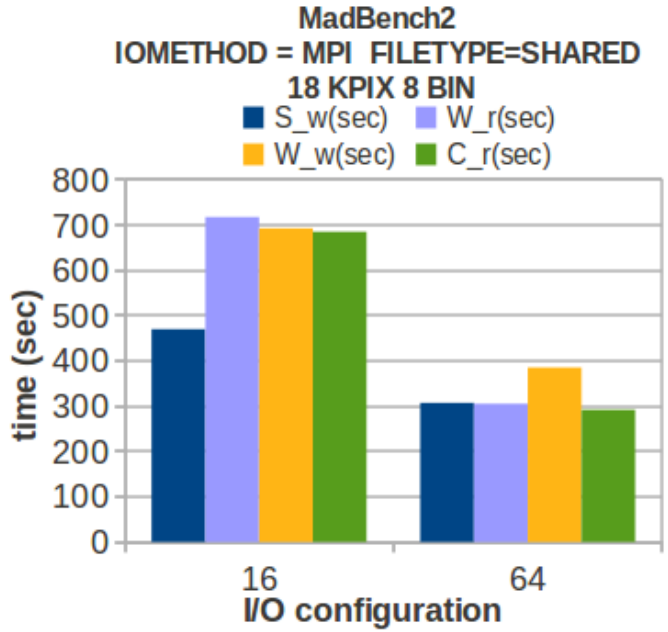
USED PERCENTAGE (%) BY MADBENCH2 ON LOCAL FILESYSTEM

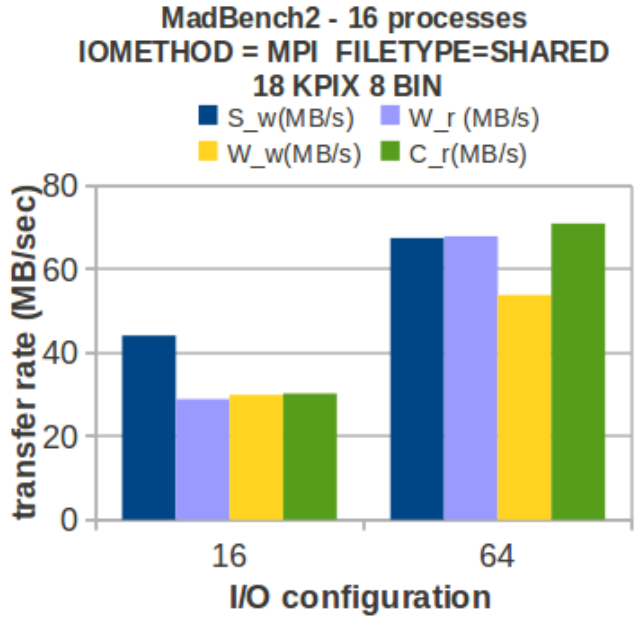
I/O configuration	W_r	C_r	S_w	W_w	FILETYPE
16	12.03	12.60	22.06	14.96	SHARED
64	28.30	29.57	33.74	26.91	SHARED
16	22.13	25.65	19.00	13.05	UNIQUE
64	384.34	386.92	51.25	39.13	UNIQUE

Experimentation

MadBench SHARED – "A"

Evaluation





USED PERCENTAGE (%) BY MADBENCH2 ON NETWORK FILESYSTEM

I/O configuration	W_r	C_r	S_w	W_w	FILETYPE
16	32.09	33.61	63.03	42.75	SHARED
64	75.48	78.85	96.41	76.89	SHARED
16	59.01	68.39	54.29	37.28	UNIQUE
64	1024.91	1031.80	146.42	111.79	UNIQUE

USED PERCENTAGE (%) BY MADBENCH2 ON LOCAL FILESYSTEM

I/O configuration	W_r	C_r	S_w	W_w	FILETYPE
16	12.03	12.60	22.06	14.96	SHARED
64	28.30	29.57	33.74	26.91	SHARED
16	22.13	25.65	19.00	13.05	UNIQUE
64	384.34	386.92	51.25	39.13	UNIQUE

Conclusions and Future Work

- Motivation
- Objetivo
- Proposed Methodology
- Experimentation
- **Conclusions and Future Work**

Conclusions and Future Work

- A methodology to analyze I/O performance of computer clusters was shown.
- The characterization of the I/O system at different levels: devices, I/O system and application.
- Analyzed and evaluated configuration of different elements that impact on performance by considering the application and the I/O architecture.
- Applied in two different clusters for the NAS BT-IO and MadBench2 benchmark.
- I/O systems features are shown, as well as their impact on the performance of the application.

Conclusions and Future Work

- An application I/O model is being defined to give support to the evaluation and design of configurations.
- The Model is to determine which configuration of I/O meets the performance requirements of the user, taking into account the behavior of the application I/O in a given system.
- For evaluation of other configurations, we are analyzing the simulation framework SIMCAN. We plan to use this tool to model full I/O architectures.

Universitat Autònoma de Barcelona

Computer Architecture & Operating Systems

Methodology for Performance Evaluation of the Input/Output System on Computer Clusters

Speaker: Carlos Núñez

Sandra Méndez, Dolores Rexachs, Emilio Luque
Contact: sandra.mendez@uab.es

Thanks!

September 30, 2011

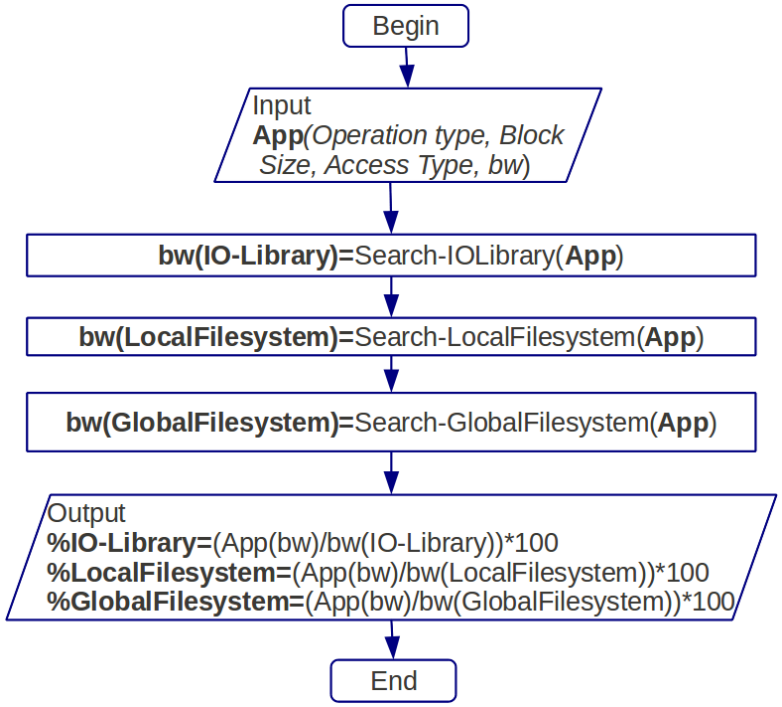
IEEE CLUSTER 2011

INTERNATIONAL CONFERENCE ON CLUSTER COMPUTING
AUSTIN, TEXAS - SEPTEMBER 26-30, 2011

Proposed Methodology

Evaluation

Generation of tables of the I/O system use for each I/O configuration



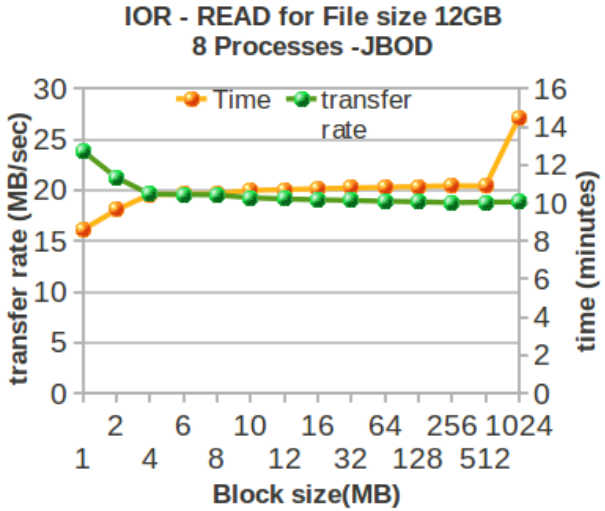
Searching of the bandwidth on each I/O configuration

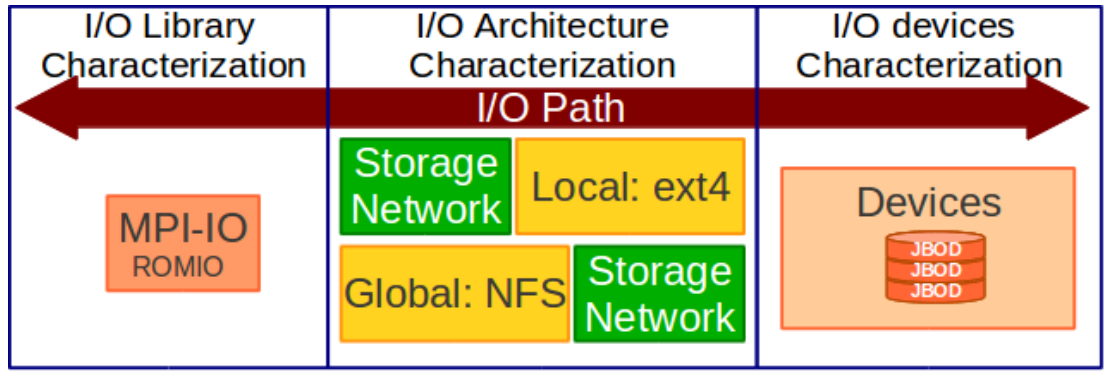
a) Opening the table of performance and setting the variable "found" to stop the searching when the values are found.
b) If the operation type, access mode, and access type is equal to a value in the performance table, and the block size of the operation is:
   1. less than minimum block size of the performance table then it selects the transfer rate corresponding to minimum block size.
   2. greater than maximum block size of performance table then, it selects the transfer rate corresponding to the maximum block size.
   3. equal to a block size of the performance table then it selects the transfer rate corresponding to such block size.
   4. a value between the characterized values then it selects the closest upper value to the searched value.
c) When the search finishes then the performance table is closed and the transfer rate is returned.

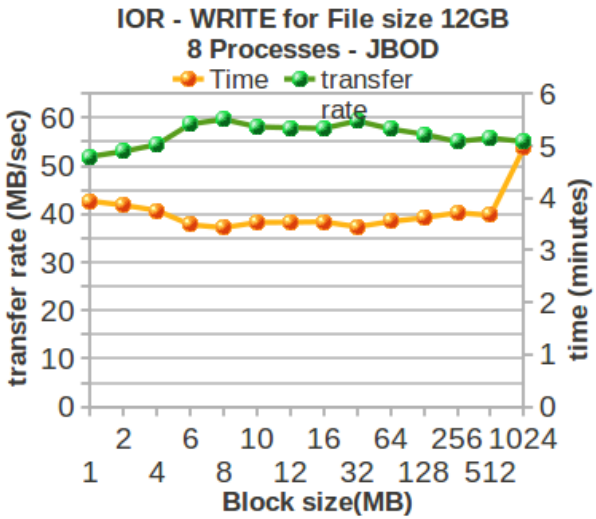
Experimentation

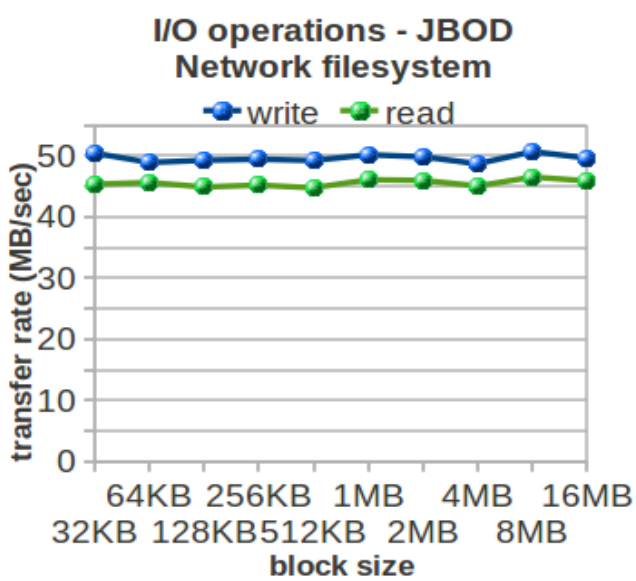
Aohyper - JBOD

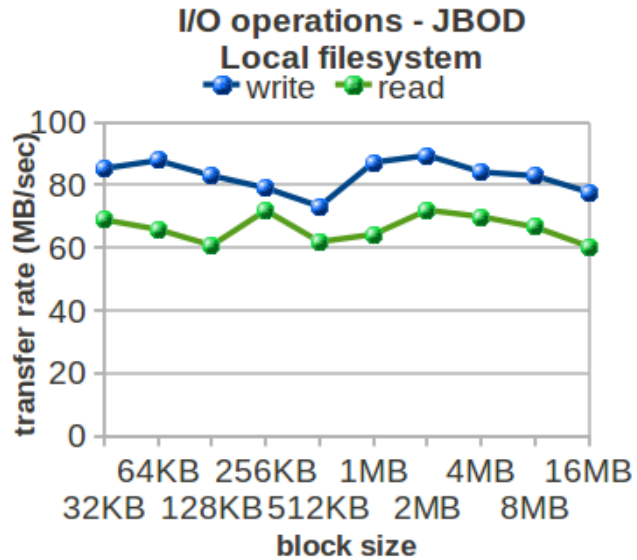
I/O analysis



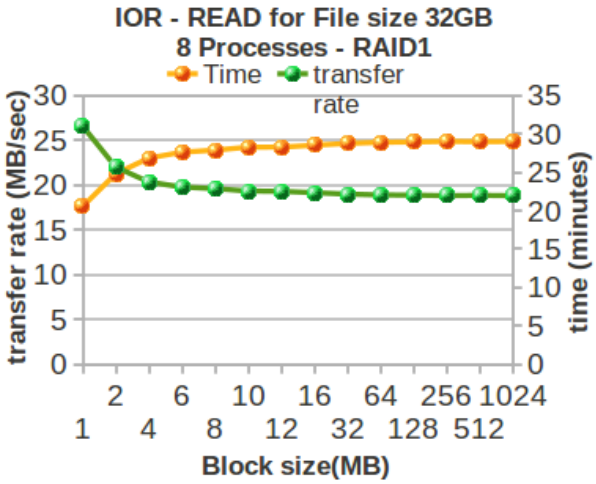
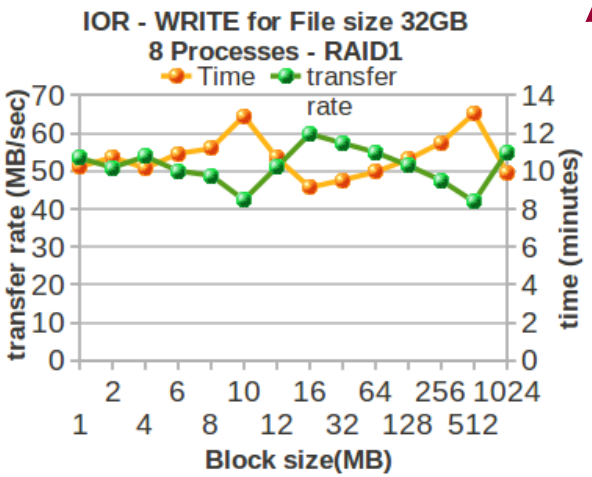
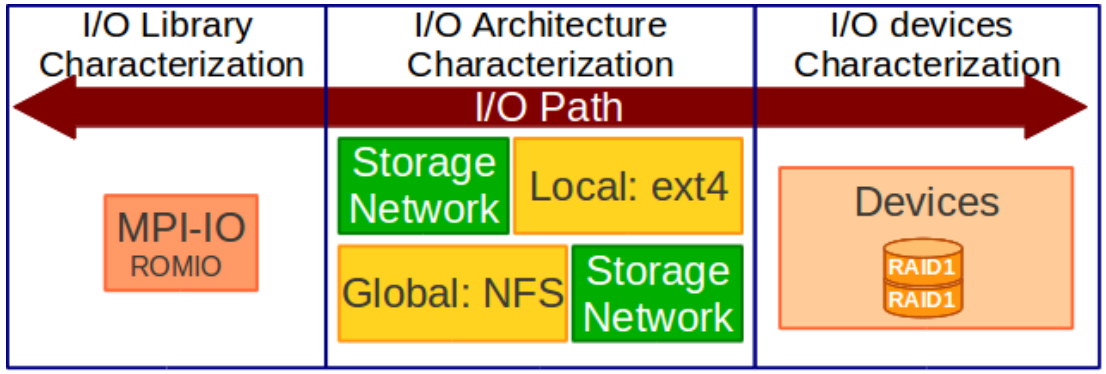
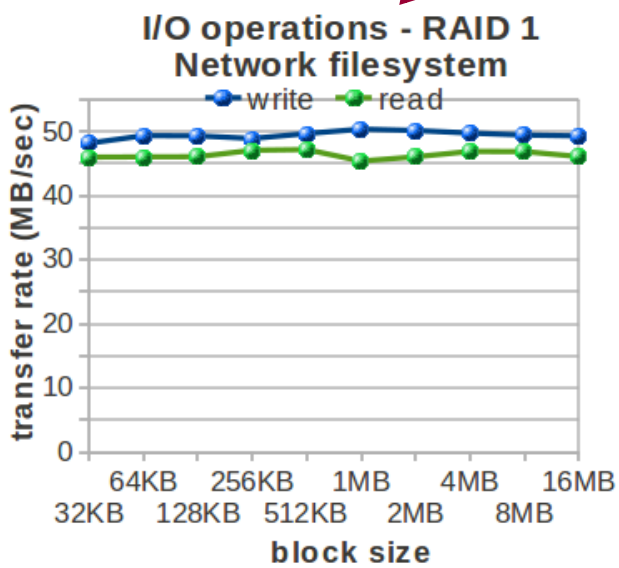
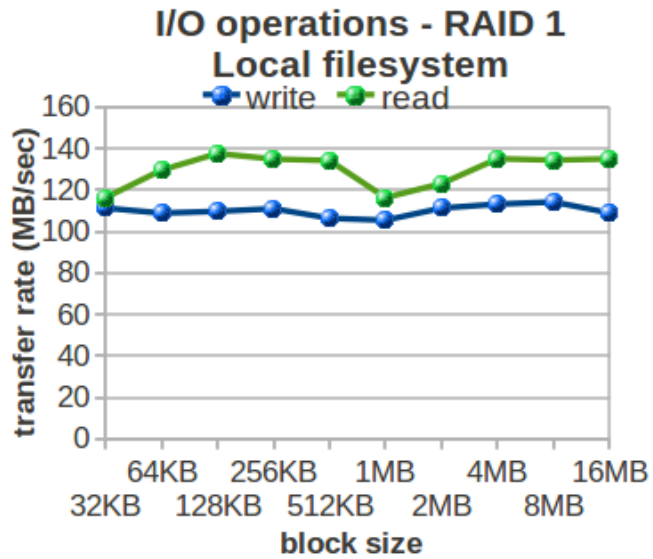








Experimentation

Aohyper – RAID1

I/O analysis