
# Analyzing the Parallel I/O Severity of MPI Applications

Authors: Sandra Mendez, Dolores Rexachs, Emilio Luque

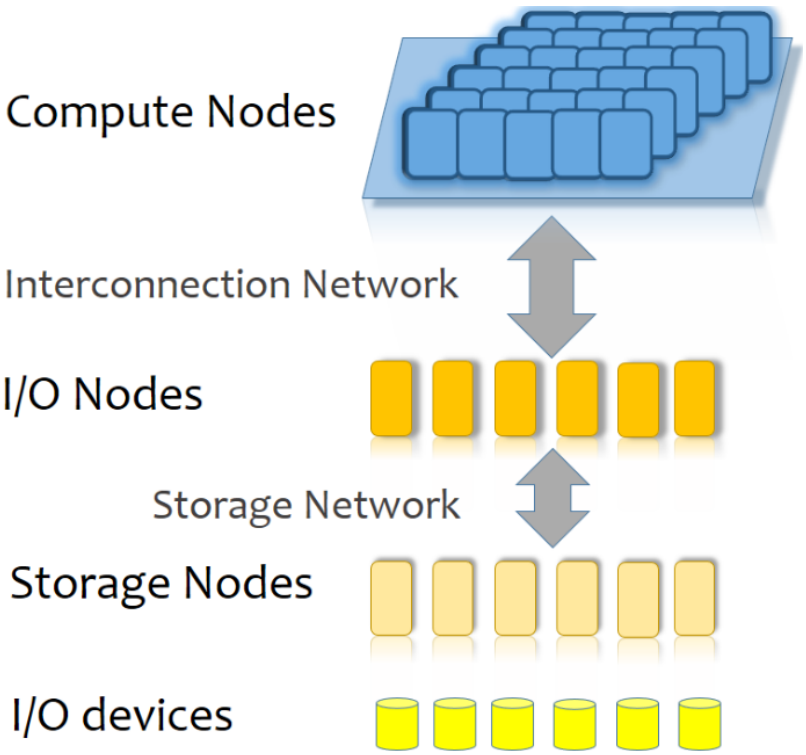
**Speaker: Sandra Mendez, PhD.**

*Research Associate, HPC Group, LRZ.*

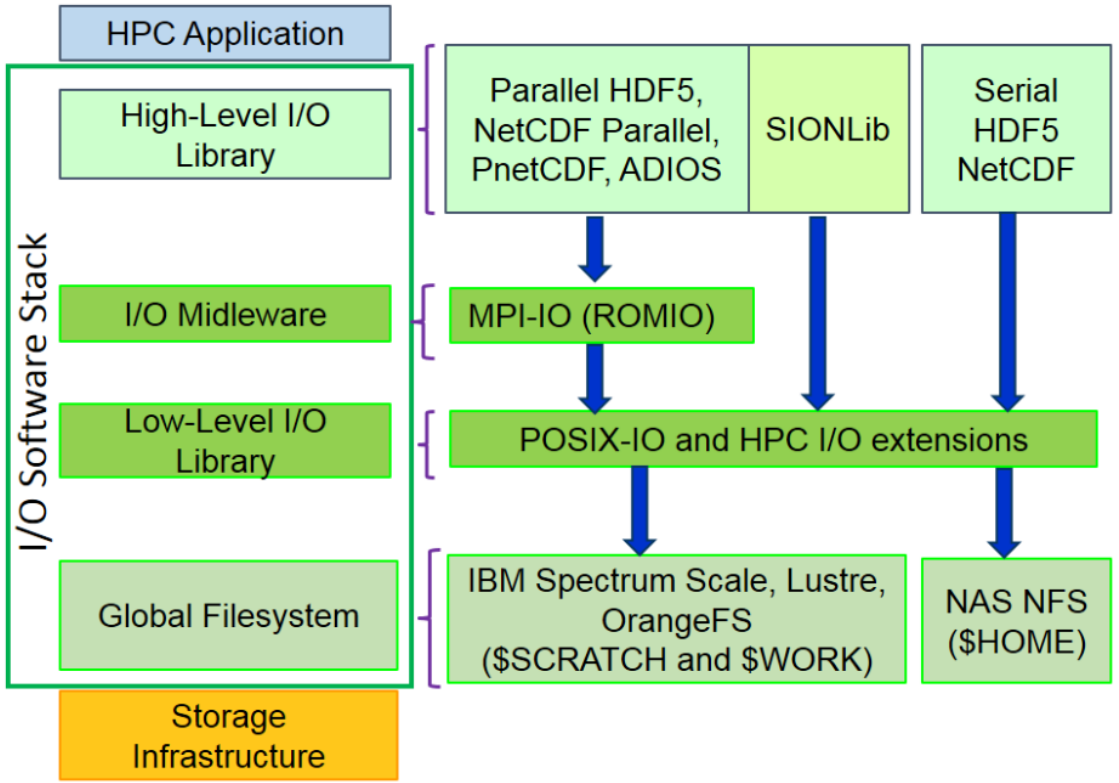
*External Researcher, HPC4EAS Research Group, UAB.*

# Outline

# The I/O System

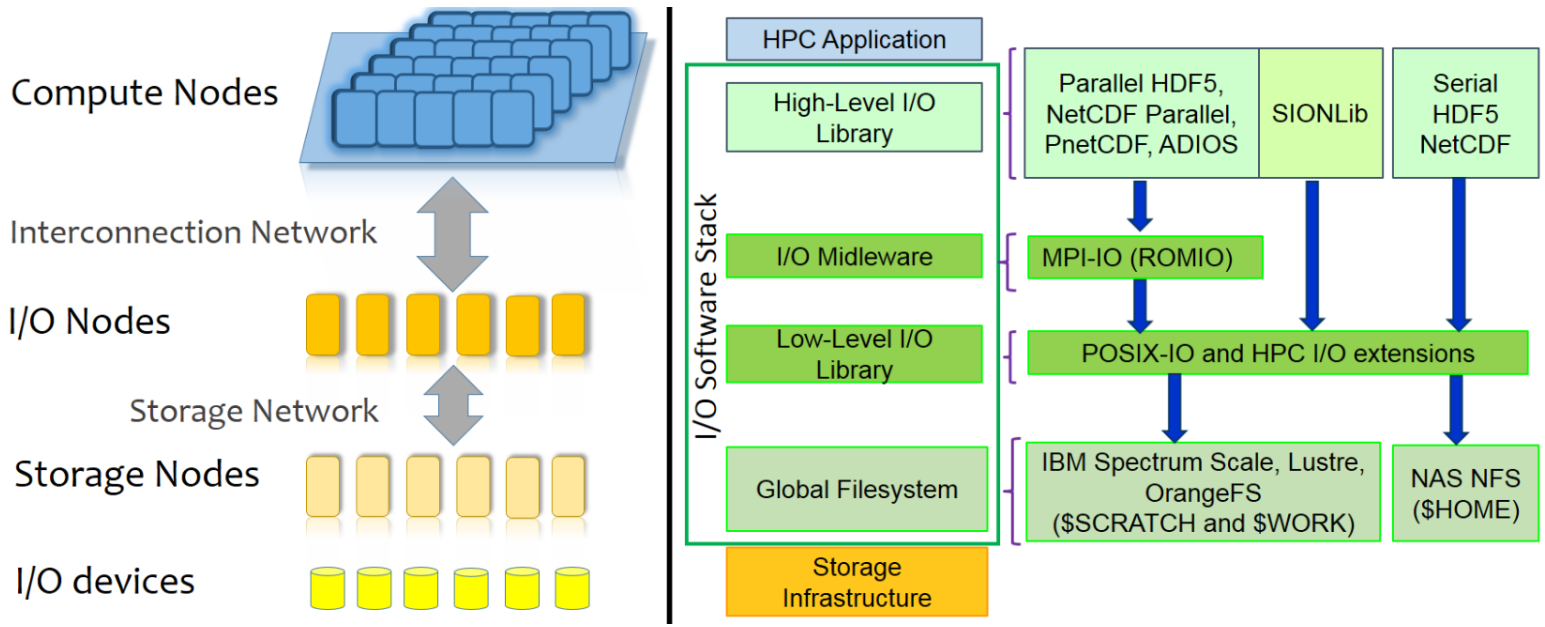



I/O Hardware



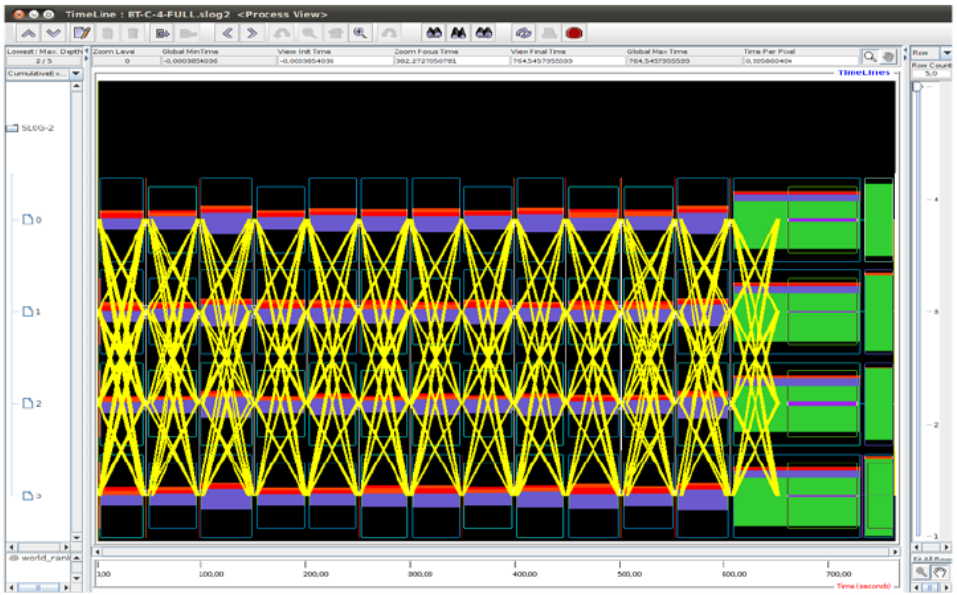

I/O Software

# I/O Performance Evaluation

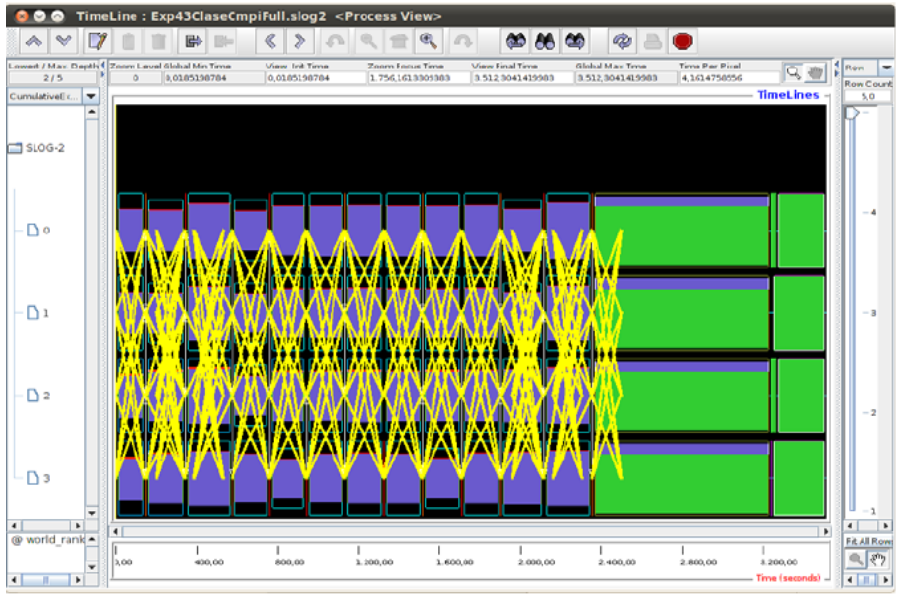



- Parallel I/O performance evaluation is a non-trivial task. It depends on the I/O pattern of application, the I/O software stack and the I/O system infrastructure.
- Depending on the I/O patterns and utilized resources, the parallel application I/O performance can vary considerably.
- How we know if the data transfer rate is appropriate for a parallel application?

System A




System B




compute
communication
write
read

## Severity

The impact degree on I/O performance based on application I/O requirements (features) and system configuration.

## Severity

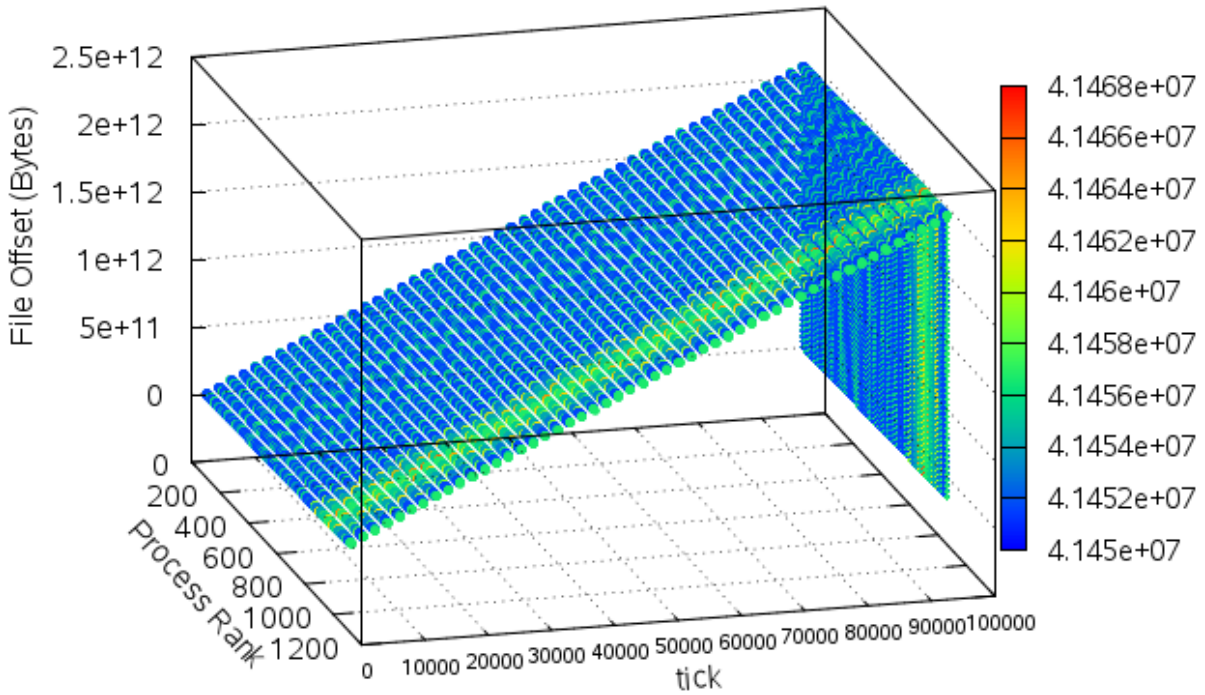
The impact degree on I/O performance based on application I/O requirements (features) and system configuration.

- Represent a parallel application through its key I/O characteristics.
- Define the I/O requirements based on the characteristics and parameters of the HPC system.
- Define degrees of severity by using the I/O requirements.

# Characteristics

Parallel application:

- $NP$ the number of MPI processes of the application
- $FI = \{F_1, F_2, ..., F_n\}$ a set of files of the application
- $ST_{app}$ the storage capacity required by the application
- $DT_{app}$ data transferred to file system by the application




FI = F1; NP = 1024

$ST_{app}$ = 2 TiB

$DT_{app}$ = 4 TiB

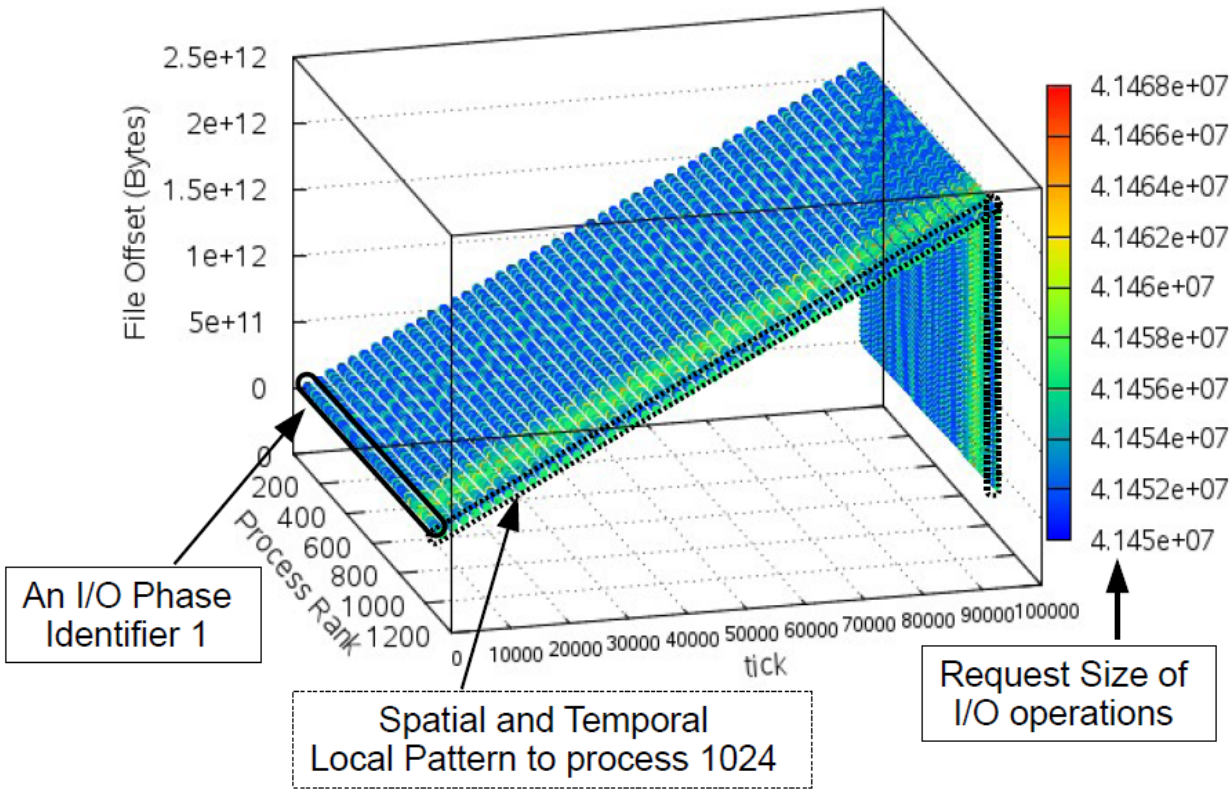
# File Characteristics

The metadata information of each $F_i$ comprises:

- $NP_{io}$ the maximum number of I/O processes
- $Fi_{size}$ the file size
- $AM$ the access mode (strided or sequential)
- $AT$ the access type, which can be Unique or Shared
- the `access data type` that can be write-only, read-only or read-write.
- $Fi_{data}$ the amount of data to be transferred of the file $F_i$.
- $\#PhIO_{F_i}$ the count of I/O phases of the file $F_i$

$NP_{io}$=1024; $Fi_{size}$=2TiB; Strided AM; Shared AT; Write/Read;
$PhIO_{F_i}$= (ph1, .., ph51)and $\#PhIO_{F_i}$= 51

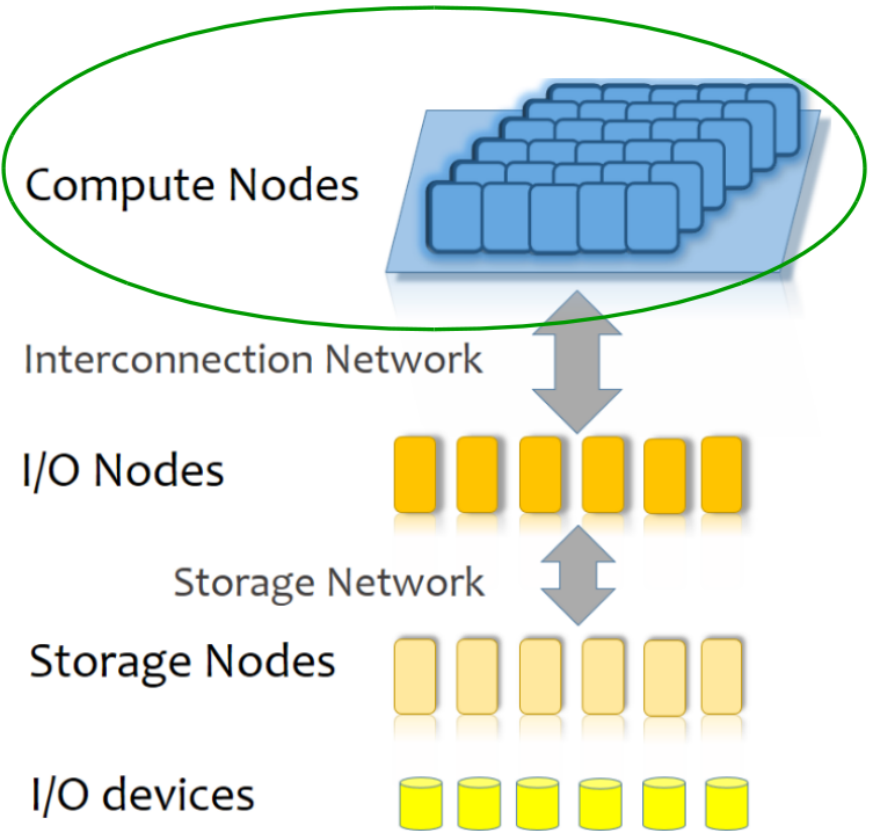
# Characteristics - Phases




*ph*1 to *ph*50: 1 `MPI_File_write_at_all` per MPI process. $P_{iop} = 1$, rs(MiB)=40.96, $P_{data}$=40.96MiB

*ph*51: 50 `MPI_File_read_at_all` operations. $P_{iop} = 50$, rs(MiB)=40.96, $P_{data} = 40.96MiB * P_{iop}$

# Requirements



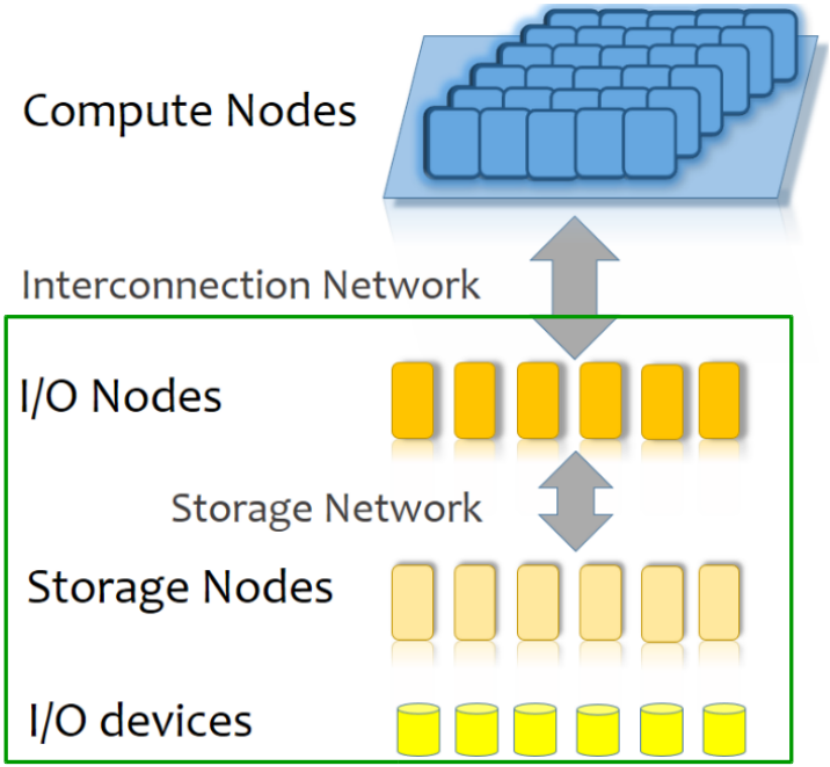

- Percentage of data required by I/O phases per compute node

$$\%Data_{CN} = \frac{(np_{CN} \times P\_data) \times 100}{RAM_{CN}}$$

- Parallel I/O in a compute node

$$PIO_{CN} = \frac{rs \times np_{CN}}{BW_{CN}}$$

# Requirements




- Percentage of the storage capacity required the application

$$\%ST_{Req} = \frac{ST_{app}}{FS_{size} - FS_{used}} \times 100$$

- Weight of the I/O operations

$$Ideal_{IOP} = \frac{Fi_{data}}{Stripe_{size}}$$

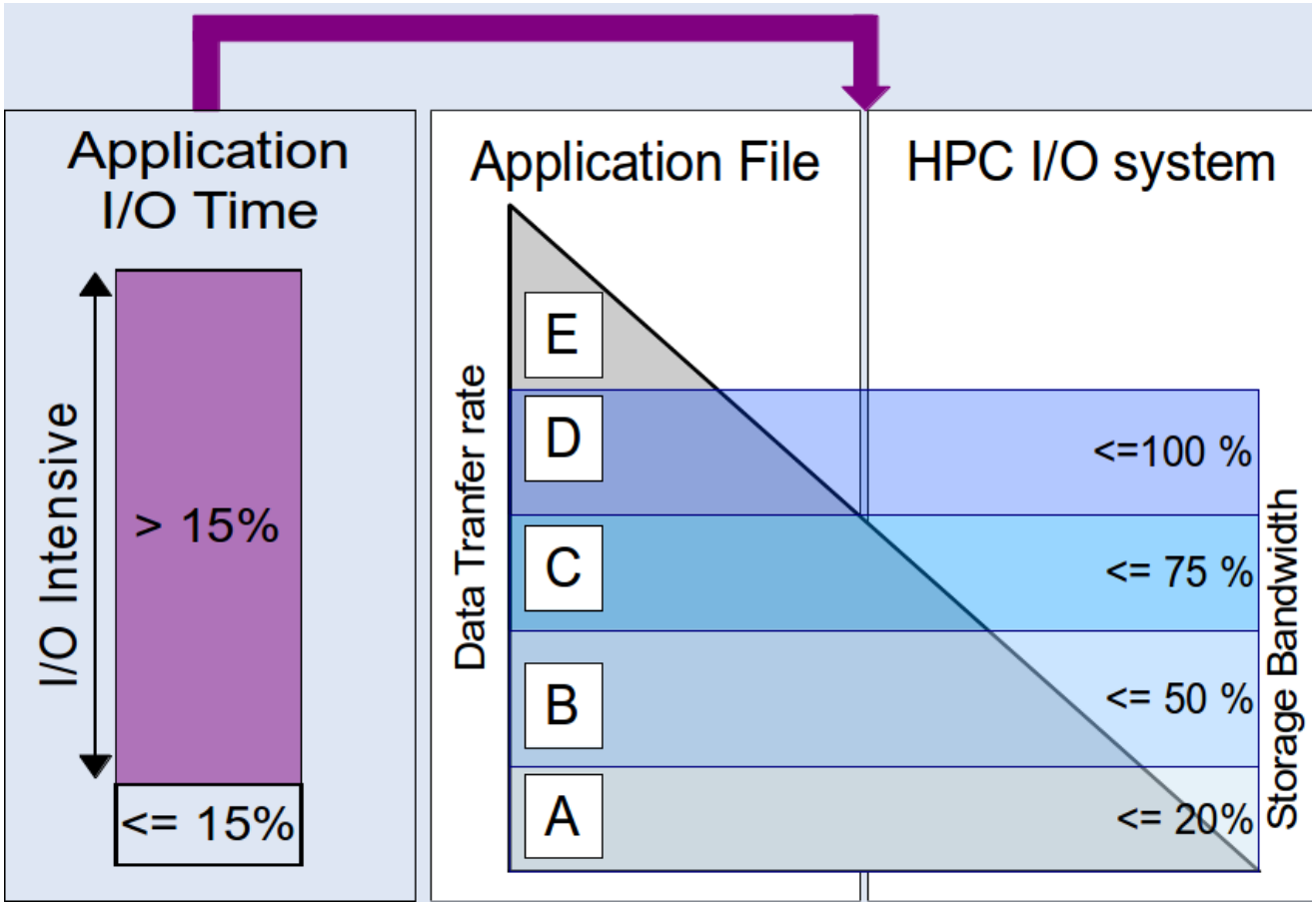
$$WIOP_{Fi} = \frac{\sum_{ph=1}^{\#PhIO_{Fi}} np_{io}(ph) \times P\_\#iop(ph)}{Ideal_{IOP}}$$

# Five Severity Degrees:

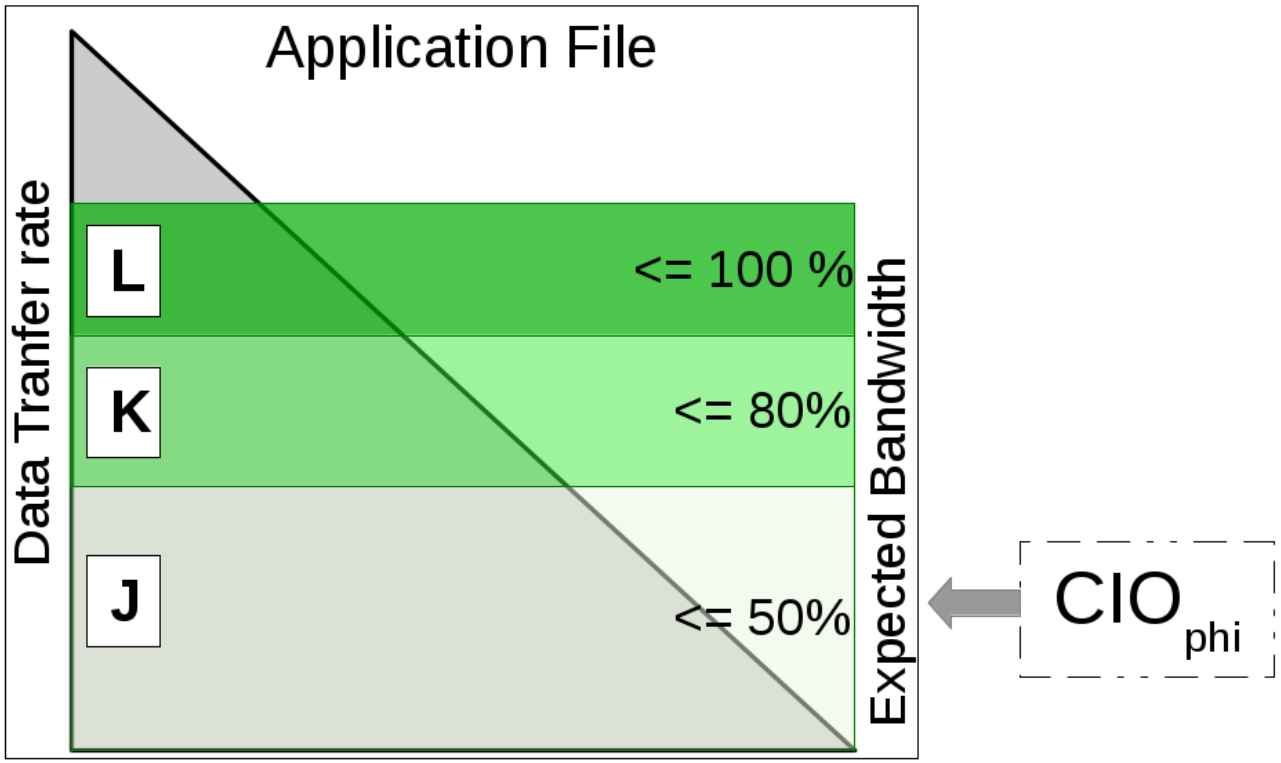
- No Severe (NS)
- Low (Lo)
- Medium (M)
- High (H)
- Very High (VH)

| **Severity** | **Requirement Evaluation** | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\%Data_{CN}$ | | $\%ST_{Req}$ | | $WIOP$ | | $\%PIO_{CN}$ |
| NS | $(< 10)$ | **and** | $(< 20)$ | **and** | $(\leq 1)$ | **and** | $(65 < and \leq 100)$ |
| Lo | $(10 \leq and < 20)$ | **and** | $(20 \leq and < 30)$ | **and** | $(1 < and < 2)$ | **and** | $(50 < and \leq 65)$ |
| M | $(20 \leq and < 30)$ | **or** | $(30 \leq and < 40)$ | **or** | $(2 \leq and < 3)$ | **or** | $(35 < and \leq 50)$ |
| H | $(30 \leq and < 70)$ | **or** | $(40 \leq and < 80)$ | **or** | $(3 \leq and \leq 4)$ | **or** | $(15 < and \leq 35)$ |
| VH | $(\geq 70)$ | **or** | $(\geq 80)$ | **or** | $(\geq 5)$ | **or** | $(\leq 15$ or $> 100)$ |

# Methodology for the I/O performance evaluation (1)

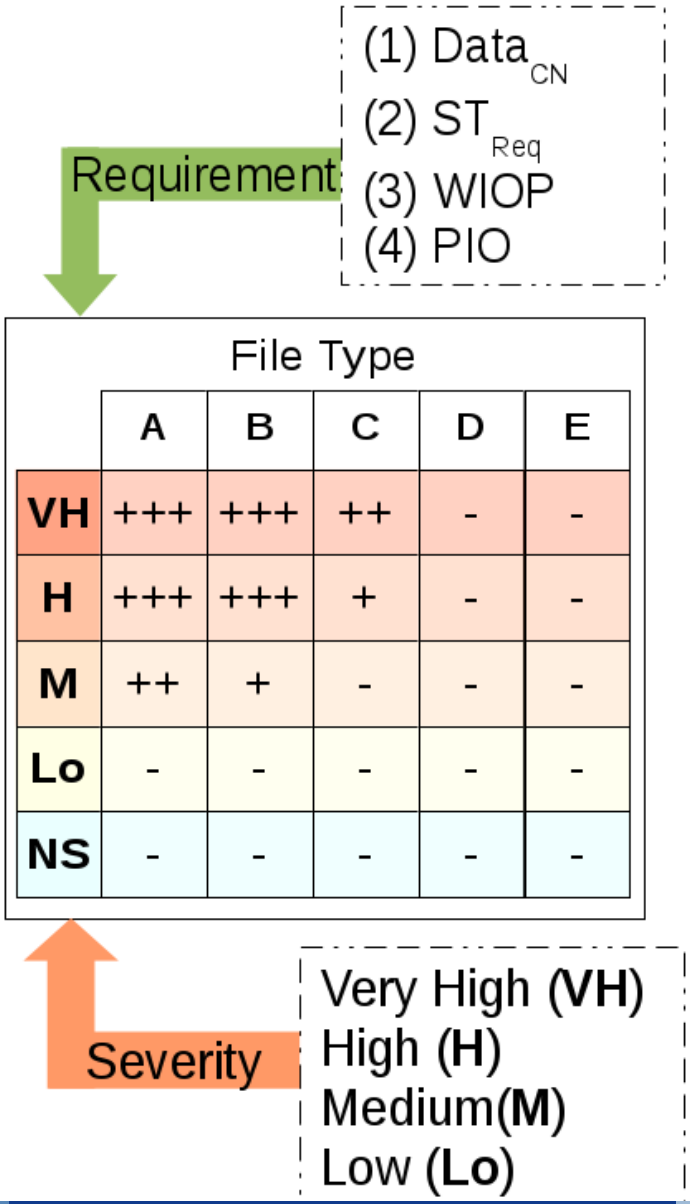


# Methodology for the I/O performance evaluation (2)

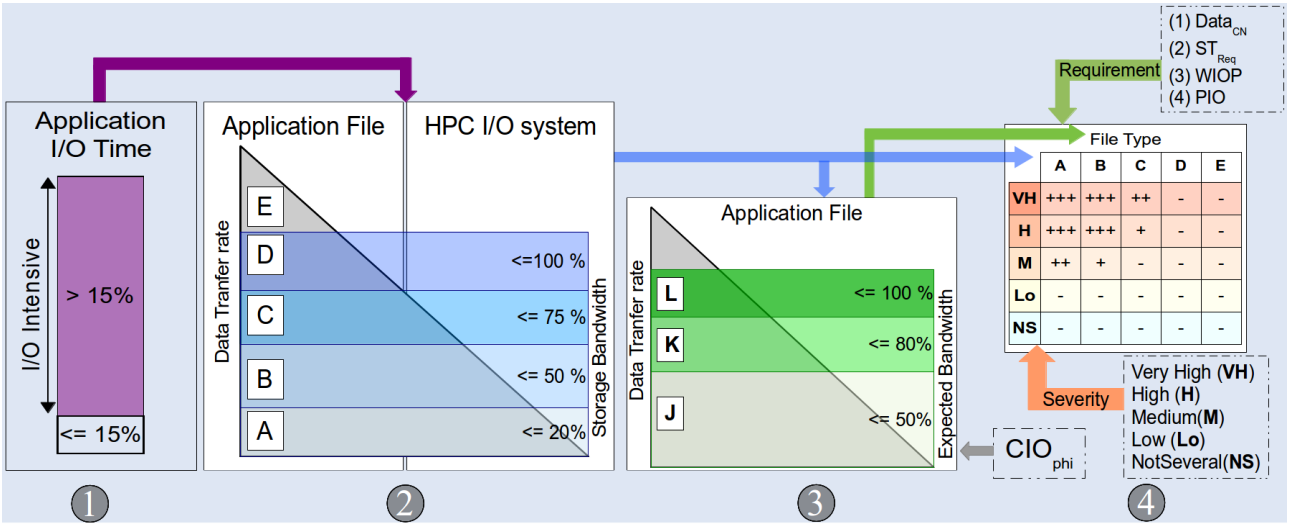



$$CIO_{ph_i} = (rs \times np_{CN} \times CN)/sec$$

# Methodology for the I/O performance evaluation (3)



# Analyzing the I/O requirements, severity and application characteristics




| | A | B | C |
|---|---|---|---|
| J | VH, H ⟹ I/O Bound +++ <br> M ⟹ I/O could improve ++ | VH, H ⟹ I/O Bound +++ <br> M ⟹ appropriate I/O + | VH, H ⟹ I/O could improve ++ |
| K | VH, H ⟹ I/O could improve ++ <br> M ⟹ appropriate I/O + | VH, H ⟹ I/O could improve ++ <br> M ⟹ appropriate I/O + | VH, H, M ⟹ appropriate I/O + |
| L | VH, H, M ⟹ appropriate I/O + | | |

# Experimental Environment

Table: SuperMUC supercomputer

| **Compute System** | **Description** | |
|---|---|---|
| Number of nodes | 9216 | |
| Sockets per Node | 2 | |
| Cores per Node | 16 | |
| Memory per node (GByte) | 32 | |
| **I/O System** | FS1 | FS2 |
| MPI library | IBM MPI | IBM MPI |
| Communication Network | FDR10 IB | FDR14 IB |
| Storage Network | FDR10 IB | FDR14 IB |
| Servers-Devices Connection | FDR10 IB | 12Gbps SAS |
| Parallel Filesystem | GPFS | GPFS |
| Data Server<br>Metadata Server | 80 NSD | 16 NSD |
| Stripe/Block Size | 8MiB | 8MiB |
| Level of Redundancy | RAID 6 | GPFS Native Raid |
| Number of I/O Devices | 10 $\times$ 564 HDDs | 8 $\times$ 348 HDDs |
| Device Capacity | 3TiB | 4TiB |
| Filesystem Capacity | 12 PiB | 5.2 PiB |
| **Max. I/O Performance** | | |
| Write | $\approx$ 180 GiB/sec | $\approx$ 130 GiB/sec |
| Read | $\approx$ 200 GiB/sec | $\approx$ 150 GiB/sec |
| Compute Node | $\approx$ 4.5 GiB/sec | |

# S3D-IO: Workload 1200x1200x1200 - 16 MPI proc. per CN

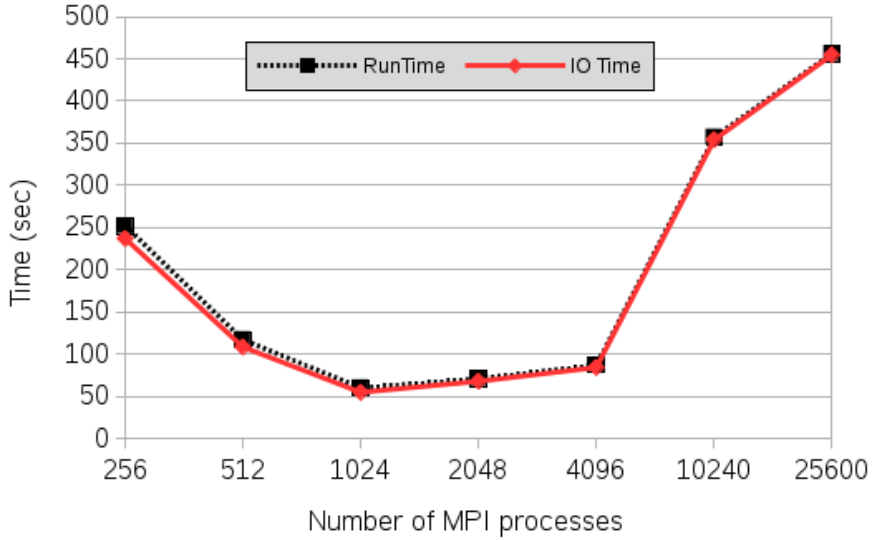
- NP = {256, 512, 1024, 2048, 4096, 10240, 25600}
- FI = {F1, F2, F3, F4, F5} $\forall$ np $\in NP$, where the number assigned to the files indicates the order in which these are written.
- $ST_{app}$ = 1030 GiB; $Data_{app}$= 1030 GiB

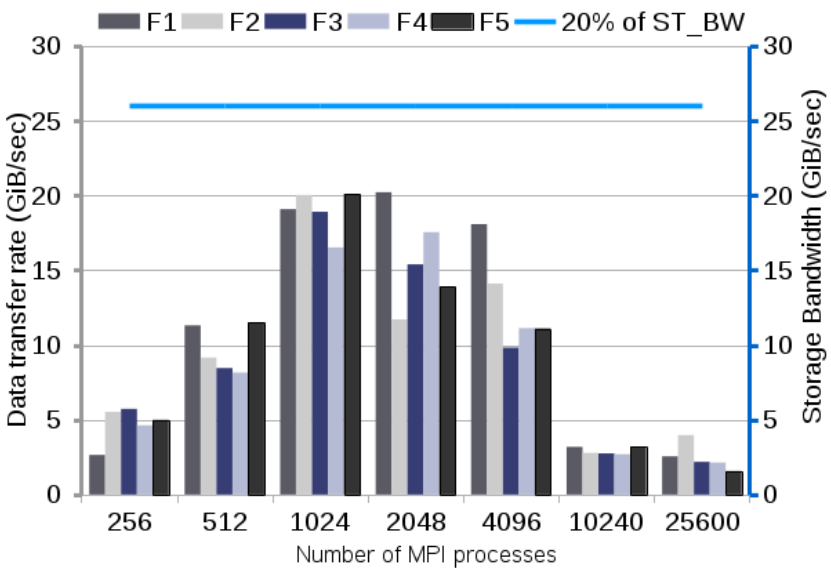
FI is described as:

- $NP_{io}$ = {256, 512, 1024, 2048, 4096, 10240, 25600}
- $Fi_{size}$ (GiB) = {206, 206, 206, 206, 206}
- $Fi_{data}$ (GiB) = {206, 206, 206, 206, 206}
- For F1 to F5: Strided AM; Shared AT; the access data type is write-only.
- $\forall\ Fi_i \in FI$, $PhIO_{Fi_i} = (ph_1)$ and $\#PhIO_{Fi_i} = 1$

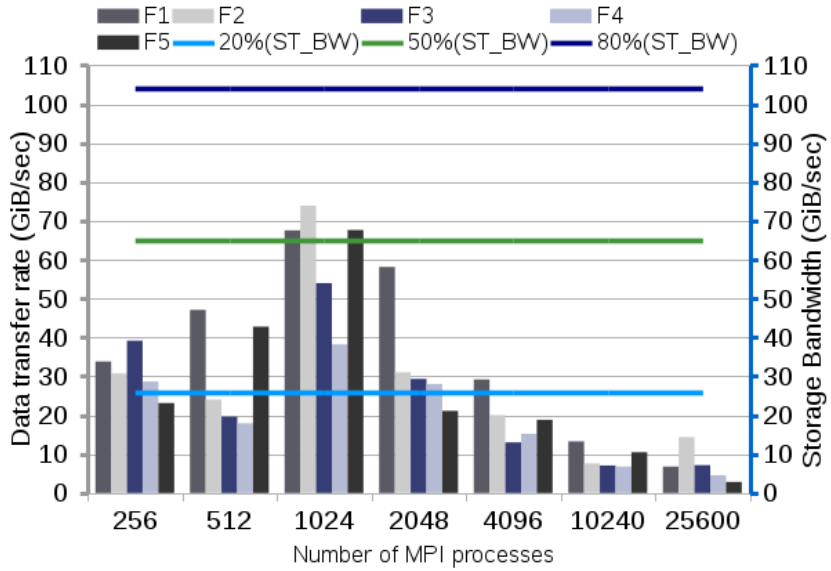
| Runing Parameters | | | APP Characteristics | | | Requirements | | | | | Severity |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $np$ | $np_{CN}$ | CN | P_#iop | $rs$ (MiB) | P_data (MiB) | $PIO_{CN}$ (GiB) | $\%PIO_{CN}$ | $Data_{CN}$ | $WIOP$ | $ST_{Req}$ | Degree |
| 256 | 16 | 16 | 1 | 824 | 824 | 12.87 | 286% | 40.23% | 0.01 | 0.02% | H |
| 512 | 16 | 32 | 1 | 412 | 412 | 6.44 | 143% | 20.12% | 0.02 | 0.02% | M |
| 1024 | 16 | 64 | 1 | 206 | 206 | 3.22 | 72% | 10.06% | 0.04 | 0.02% | Lo |
| 2048 | 16 | 128 | 1 | 103 | 103 | 1.61 | 36% | 5.03% | 0.08 | 0.02% | M |
| 4096 | 16 | 256 | 1 | 51 | 51 | 0.80 | 18% | 2.51% | 0.16 | 0.02% | H |
| 10240 | 16 | 640 | 1 | 21 | 21 | 0.32 | 7% | 1.01% | 0.39 | 0.02% | VH |
| 25600 | 16 | 1600 | 1 | 8.2 | 8.2 | 0.13 | 3% | 0.40% | 0.97 | 0.02% | VH |

# S3D-IO Kernel




(a) MPI-IO




(b) POSIX-IO

# Conclusions and Future Work

## Conclusions

- Methodology for the I/O performance evaluation for parallel scientific applications based on the I/O characteristics of the application, requirements and severity degree
- Five severity degrees were defined considering the I/O requirement of parallel applications and parameters of the HPC system.
- Severity and requirement provide useful information to reduce the complexity of the I/O performance evaluation.

## Future Work

- Apply the methodology in other HPC systems and
- Evaluate the definition of a new requirement based on the meta-data operations.

Thank you for your attention!