Investigating I/O approaches to improve performance and scalability of the Ocean-Land-Atmosphere Model

Rodrigo Virote Kassick[1 2], **Francieli Zanon Boito**[1 2], Philippe Navaux[1], Yves Denneulin[2]

[1] GPPD – II – **Federal University of Rio Grande do Sul** (UFRGS), Brazil

[2] INRIA – LIG – **Grenoble University**, France

The Seventh Workshop of the INRIA-Illinois Joint Laboratory on Petascale Computing

June 13th 2012

L I G

INSTITUTO DE INFORMÁTICA UFRGS

Investigating I/O approaches to improve performance and scalability of the **Ocean-Land-Atmosphere Model**

Rodrigo Virote Kassick[1 2], Francieli Zanon Boito[1 2], Philippe Navaux[1], Yves Denneulin[2]

[1] GPPD – II – Federal University of Rio Grande do Sul (UFRGS), Brazil

[2] INRIA – LIG – Grenoble University, France

The Seventh Workshop of the INRIA-Illinois Joint Laboratory on Petascale Computing

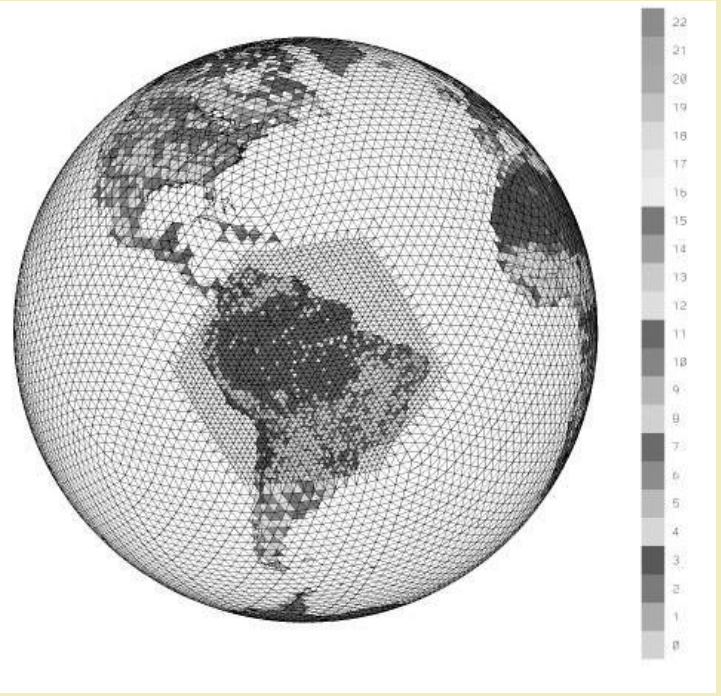
June 13th 2012

L I G

INSTITUTO DE INFORMÁTICA UFRGS

Investigating I/O approaches to improve performance and scalability of the **Ocean-Land-Atmosphere Model**

Weather/climate forecast model



L I G

INSTITUTO DE INFORMÁTICA UFRGS

Investigating I/O approaches to improve performance and scalability of the **Ocean-Land-Atmosphere Model**

- Federal University of Rio Grande do Sul (**UFRGS**)
- Center for Weather Forecast and Climate Studies (**CPTEC**)
- National Laboratory of Scientific Computing (**LNCC**)
  - University of Miami (USA)

L I G

Laboratório Nacional de Computação Científica

CPTEC

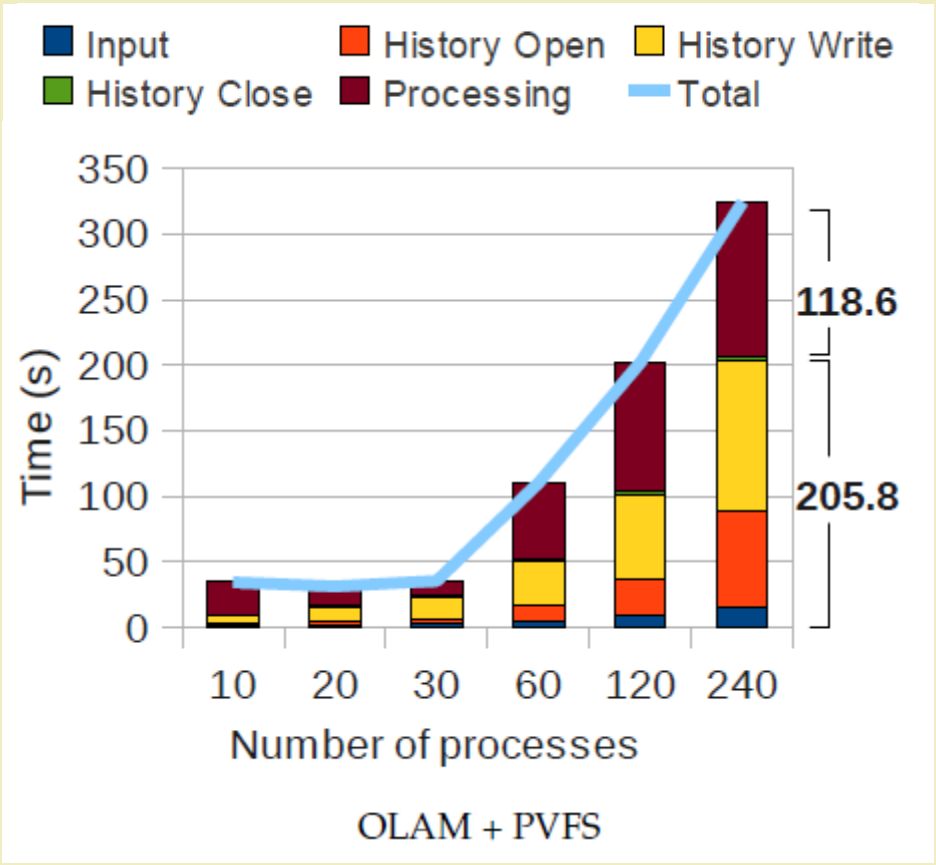
.inf INSTITUTO DE INFORMÁTICA UFRGS

Investigating I/O approaches to **improve performance and scalability** of the Ocean-Land-Atmosphere Model

L I G

INSTITUTO DE INFORMÁTICA UFRGS

Investigating I/O approaches to **improve performance and scalability** of the Ocean-Land-Atmosphere Model



L I G

INSTITUTO DE INFORMÁTICA UFRGS

Investigating I/O approaches to improve performance and scalability of the Ocean-Land-Atmosphere Model

Investigating I/O approaches to improve performance and scalability of the Ocean-Land-Atmosphere Model

Damaris

L I G

INSTITUTO DE INFORMÁTICA UFRGS

Investigating I/O approaches to improve performance and scalability of the Ocean-Land-Atmosphere Model

Investigating I/O approaches to improve performance and scalability of the Ocean-Land-Atmosphere Model

Initial results

Agenda

OLAM and its Performance Problem

OLAM + Damaris

Performance Results

Conclusions

Future Work

Agenda

OLAM and its Performance Problem

OLAM + Damaris

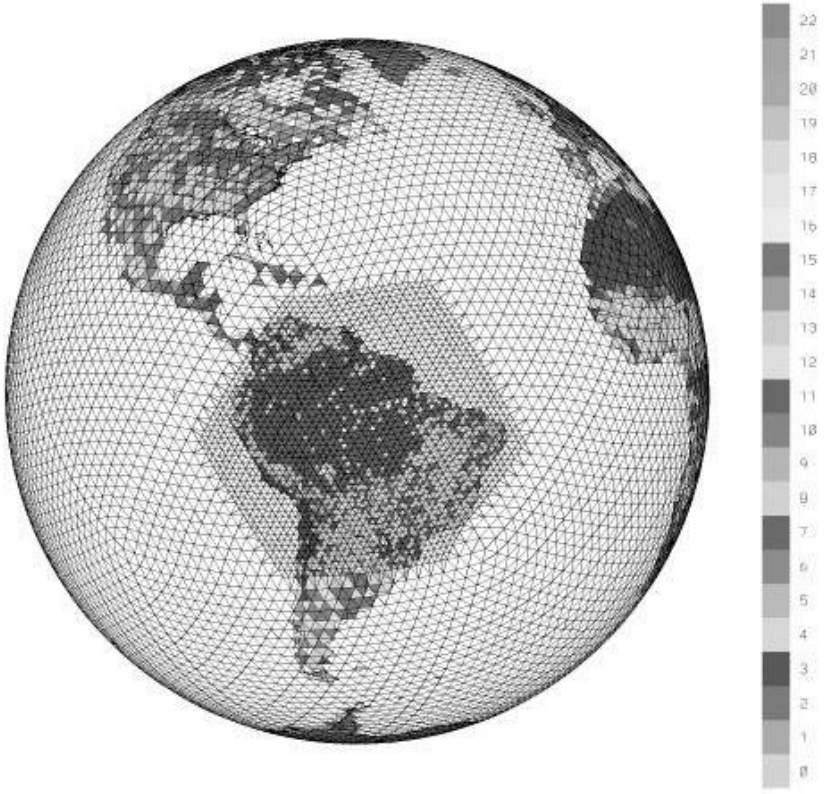
Performance Results

Conclusions

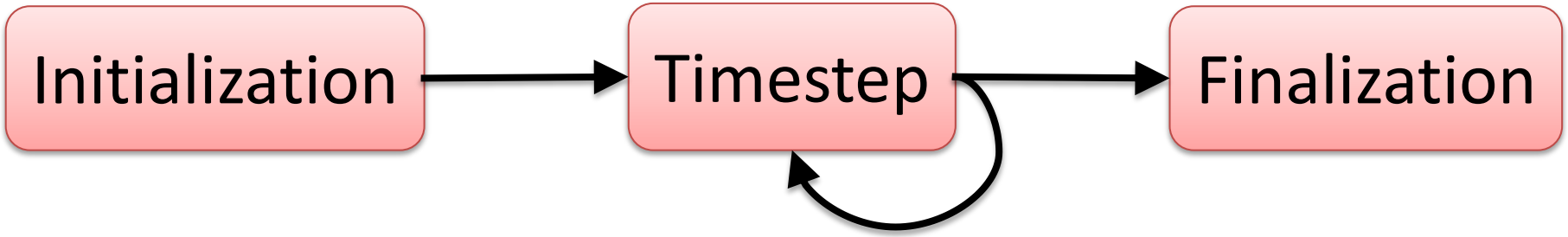
Future Work

Ocean-Land-Atmosphere Model

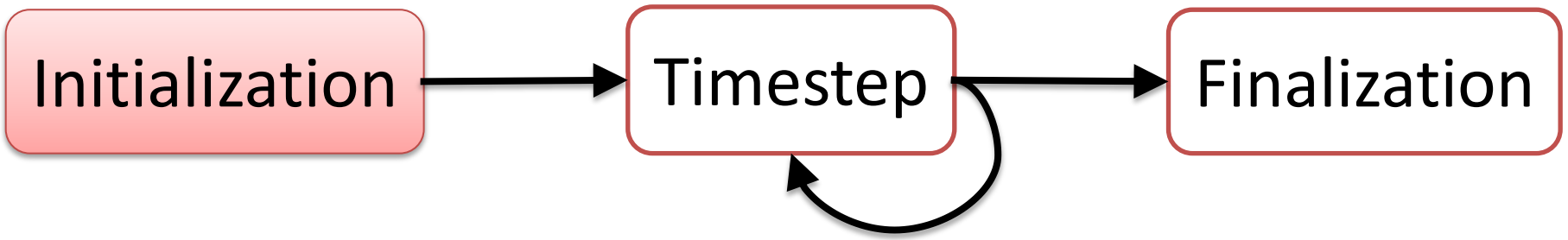
- Fortran 90 + **MPI**
- Developed at Duke University
- Global grid with local refinements



Ocean-Land-Atmosphere Model

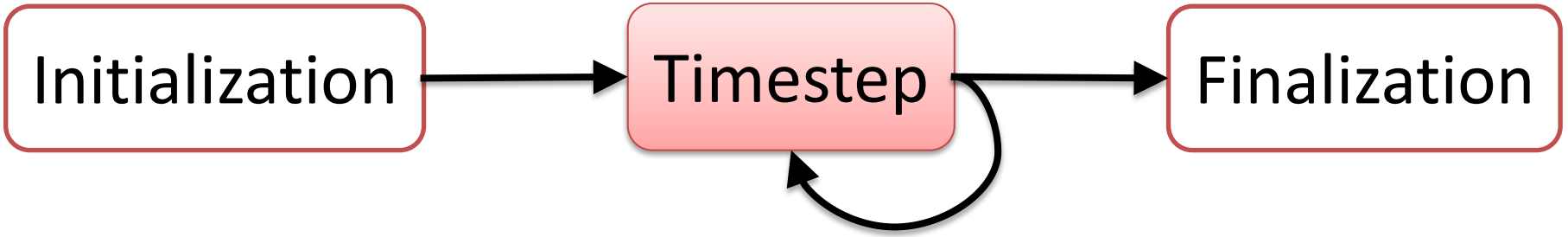


Ocean-Land-Atmosphere Model



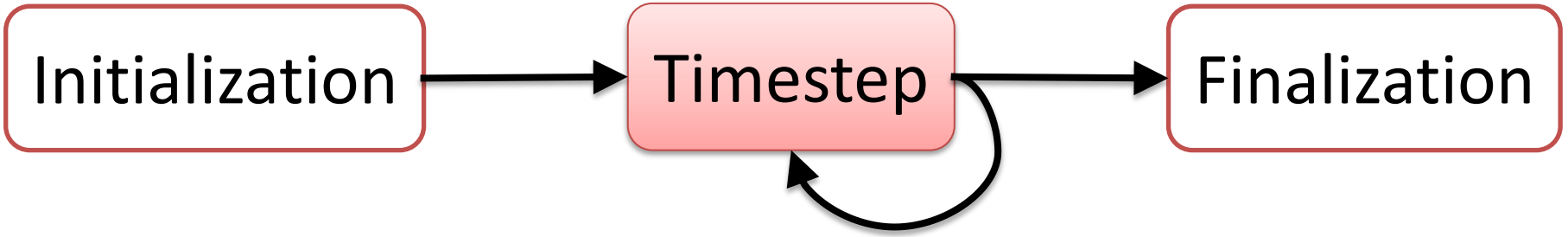
Read input files
(~600MB)
All the processes

Ocean-Land-Atmosphere Model



Every N timesteps,
write output files
(file per process)
with HDF5

Ocean-Land-Atmosphere Model



Each file:
~500KB

Every N timesteps, write output files (file per process) with HDF5

Ocean-Land-Atmosphere Model

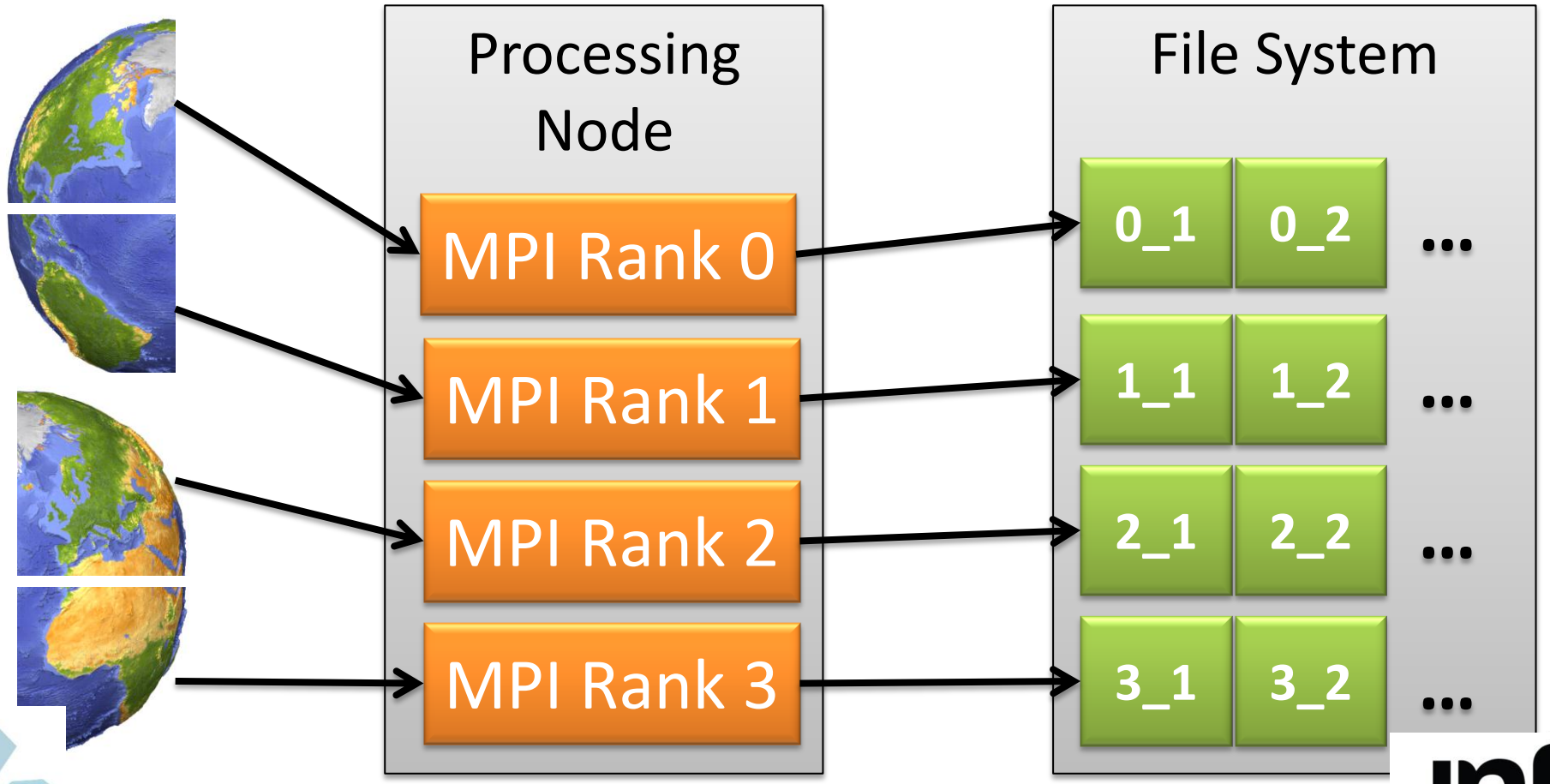
- "**Large number of small files**" access pattern
  - Poor I/O performance
  - Overhead of file creation + small requests

- **Intra-node concurrency**

Ocean-Land-Atmosphere Model

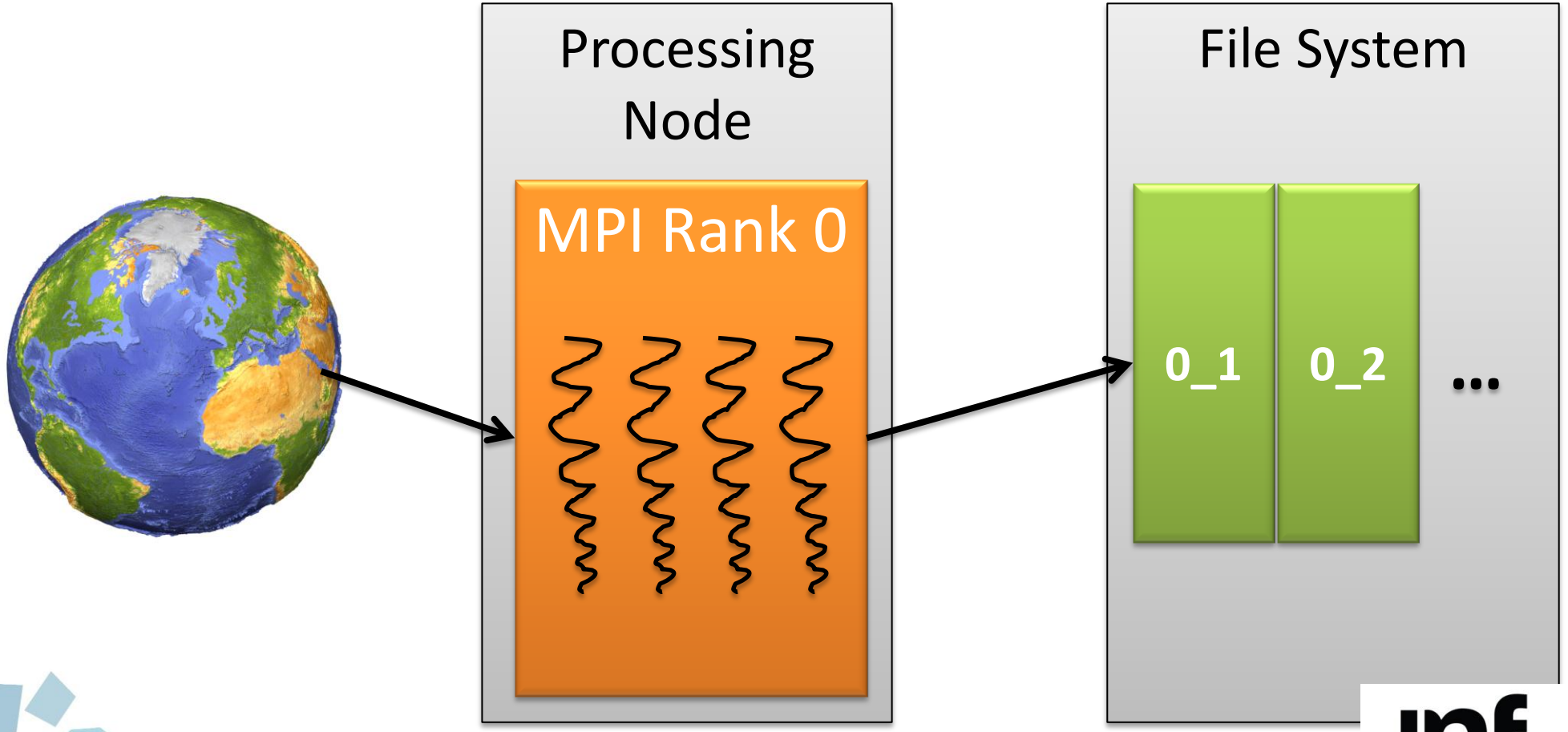
- Approach 1: **OLAM MPI+OpenMP**

Ocean-Land-Atmosphere Model



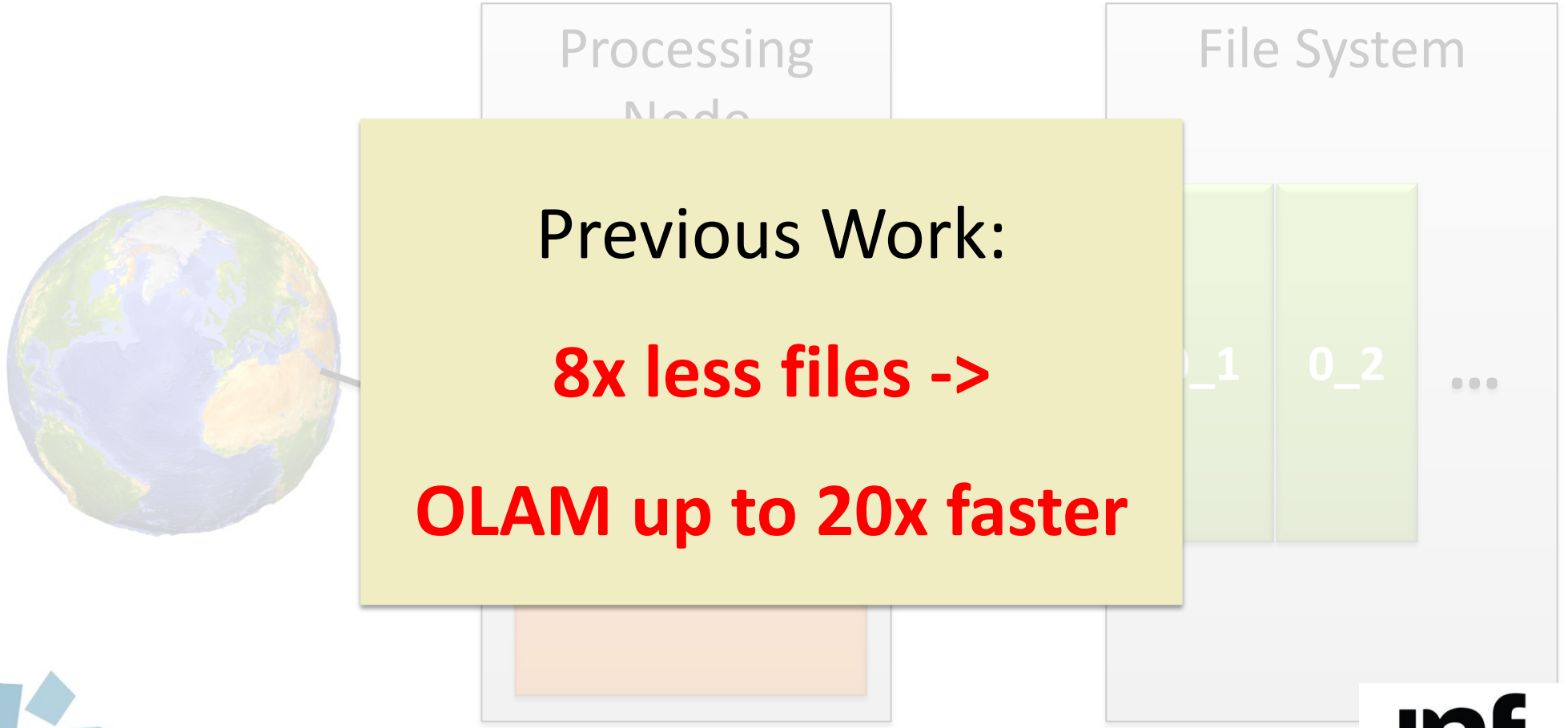
Ocean-Land-Atmosphere Model

- Approach 1: **OLAM MPI+OpenMP**



Ocean-Land-Atmosphere Model

- Approach 1: **OLAM MPI+OpenMP**



Previous Work:

8x less files ->

OLAM up to 20x faster

Ocean-Land-Atmosphere Model

- Approach 1: **OLAM MPI+OpenMP**

Processing Node

0_1

0_2

...

Previous Work:

8x less files ->

OLAM up to 20x faster

Not always...

Ocean-Land-Atmosphere Model

- Approach 2: **OLAM + Damaris**

Agenda

OLAM and its Performance Problem

OLAM + Damaris

Performance Results

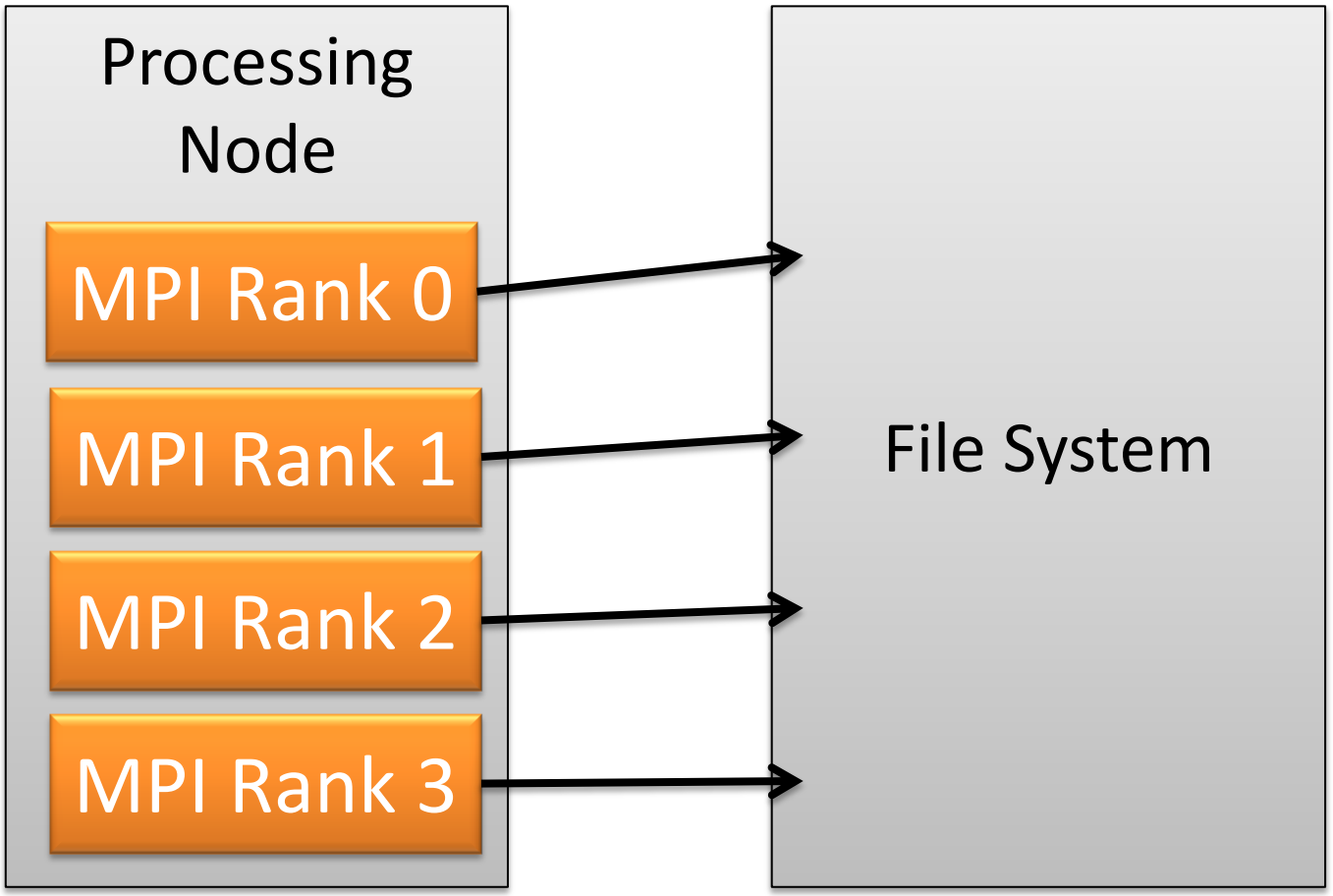
Conclusions

Future Work

OLAM + Damaris

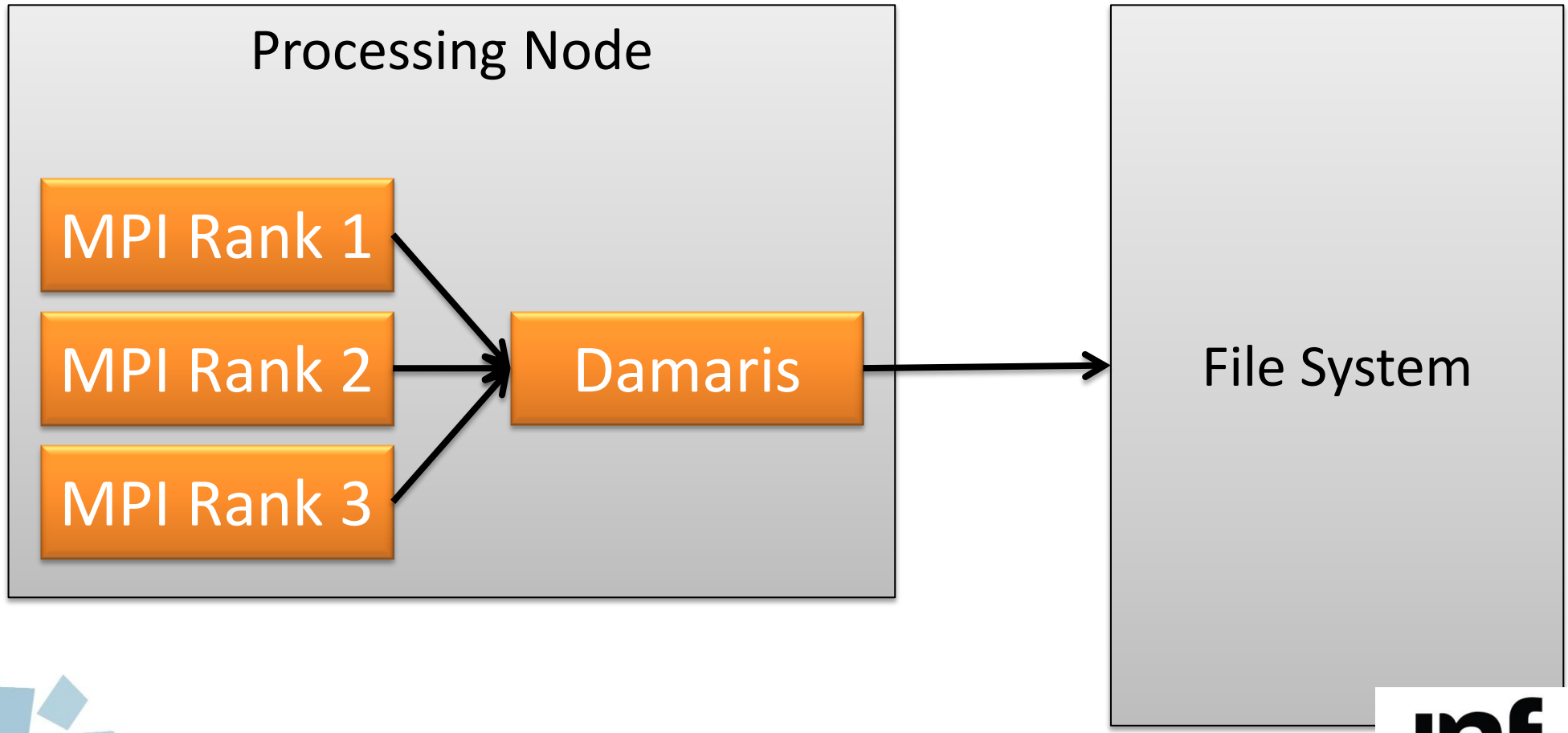
- **Damaris**
  - Library/infrastructure to help **coordinate I/O**
  - **Dedicate one core** of each node to I/O
  - Does I/O on behalf of the application

OLAM + Damaris



OLAM + Damaris

- Approach 2: **OLAM + Damaris**



OLAM + Damaris

Output phase:

1. Create file
2. Write set of variables to file
    1. Create HDF5 Dataset (H5Dcreate)
    2. Write HDF5 Dataset (H5Dwrite)
3. Close file

OLAM + Damaris

Output phase **with Damaris:**

1. ~~Create file~~
2. Write set of variables to file
    1. Create HDF5 Dataset (H5Dcreate)
    2. Write HDF5 Dataset (H5Dwrite)
3. ~~Close file~~

OLAM + Damaris

Output phase **with Damaris:**

1. ~~Create file~~
2. Write set of variables ~~to file~~ **to Damaris**
   1. ~~Create HDF5 Dataset~~ **df_chunk_set**
   2. ~~Write HDF5 Dataset~~ **df_chunk_write**
3. ~~Close file~~

OLAM + Damaris

Output phase **with Damaris:**

1. ~~Create file~~
2. Write set of variables ~~to file~~ **to Damaris**
   1. ~~Create HDF5 Dataset~~ **df_chunk_set**
   2. ~~Write HDF5 Dataset~~ **df_chunk_write**
3. ~~Close file~~
4. **df_signal**

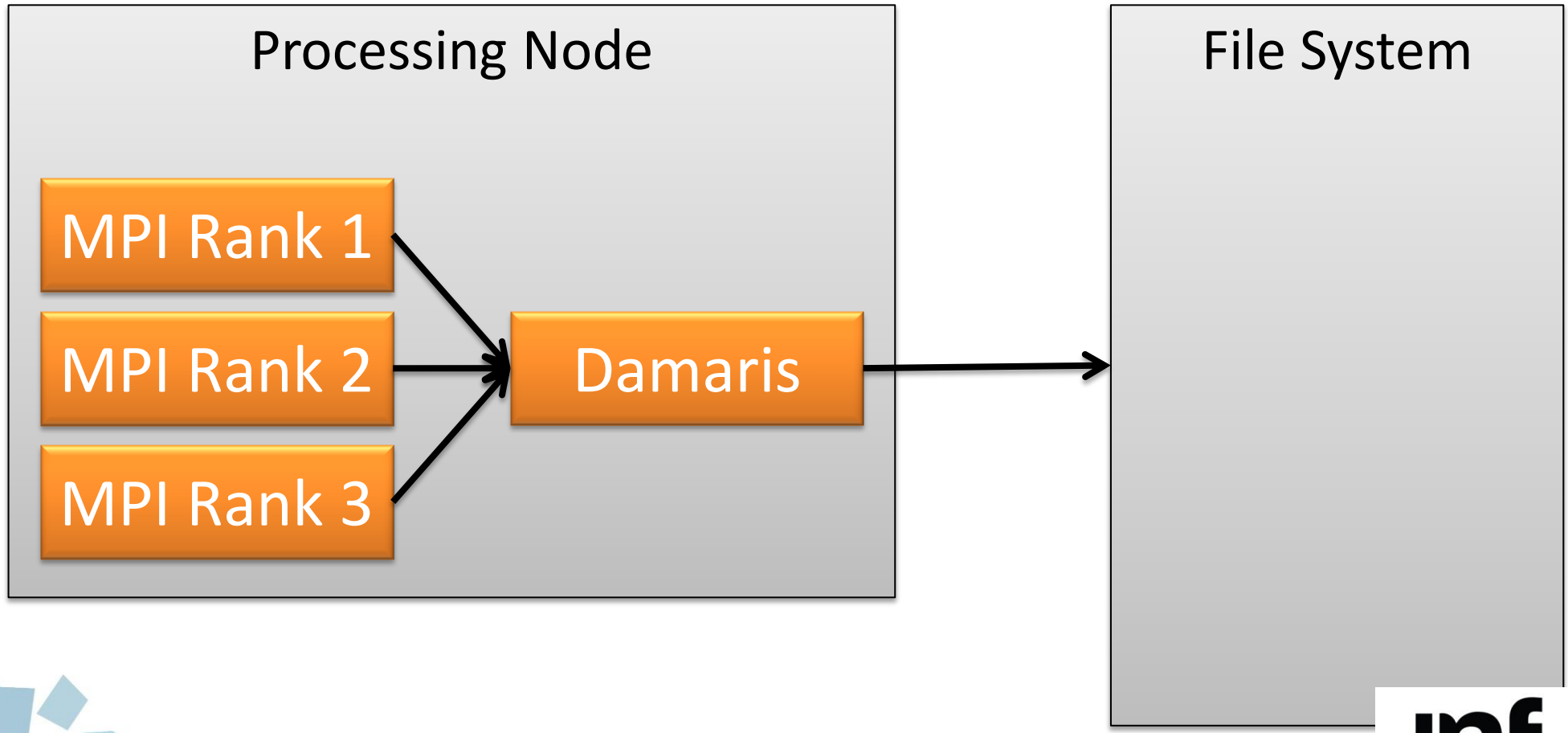
OLAM + Damaris

- **+ Plugin** to do the actual I/O
  - When the application calls **df_signal**
  - Create, write to **HDF5 files** and close

- **+ XML file**
  - List name and type of all variables

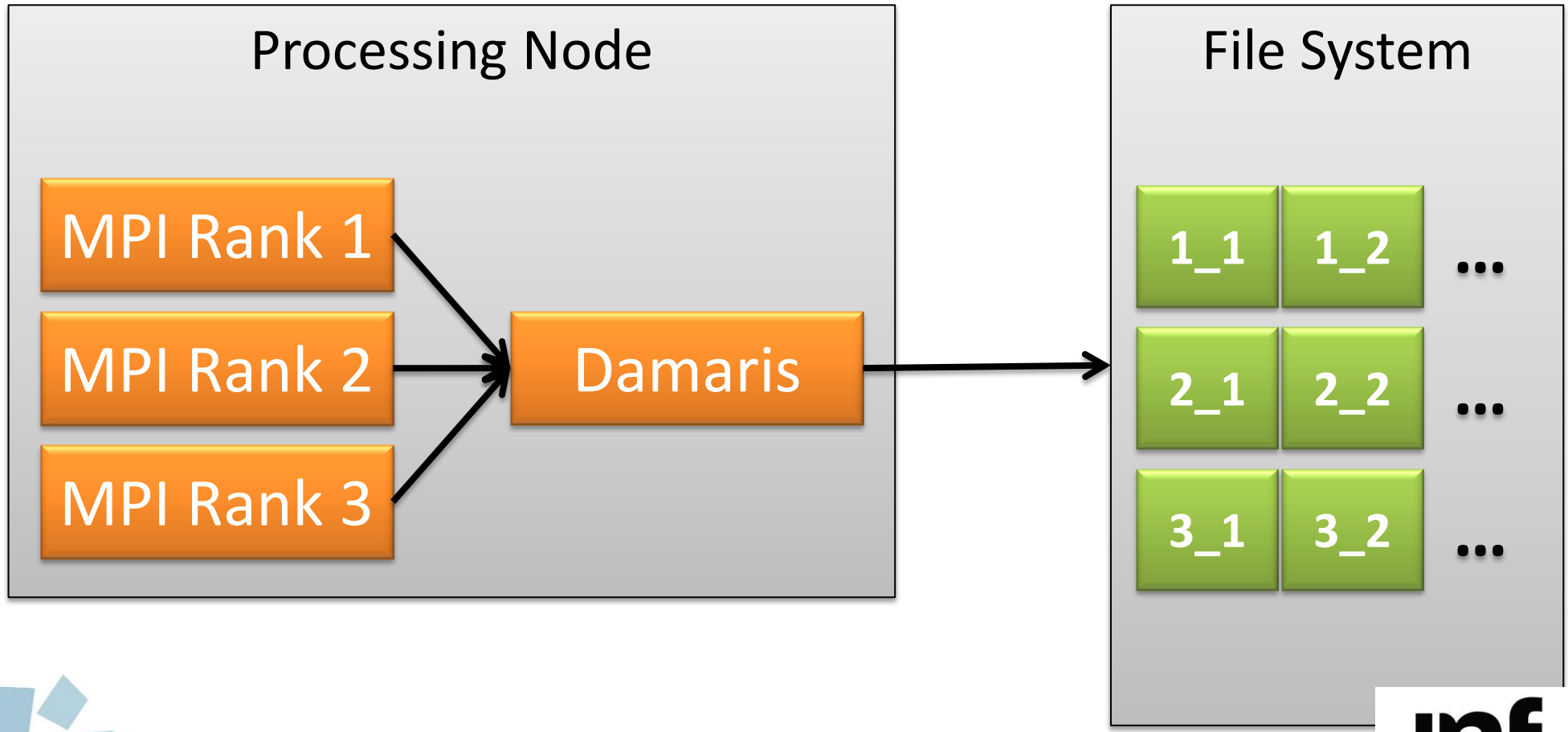
OLAM + Damaris

- **Plugin 1: File-per-process**



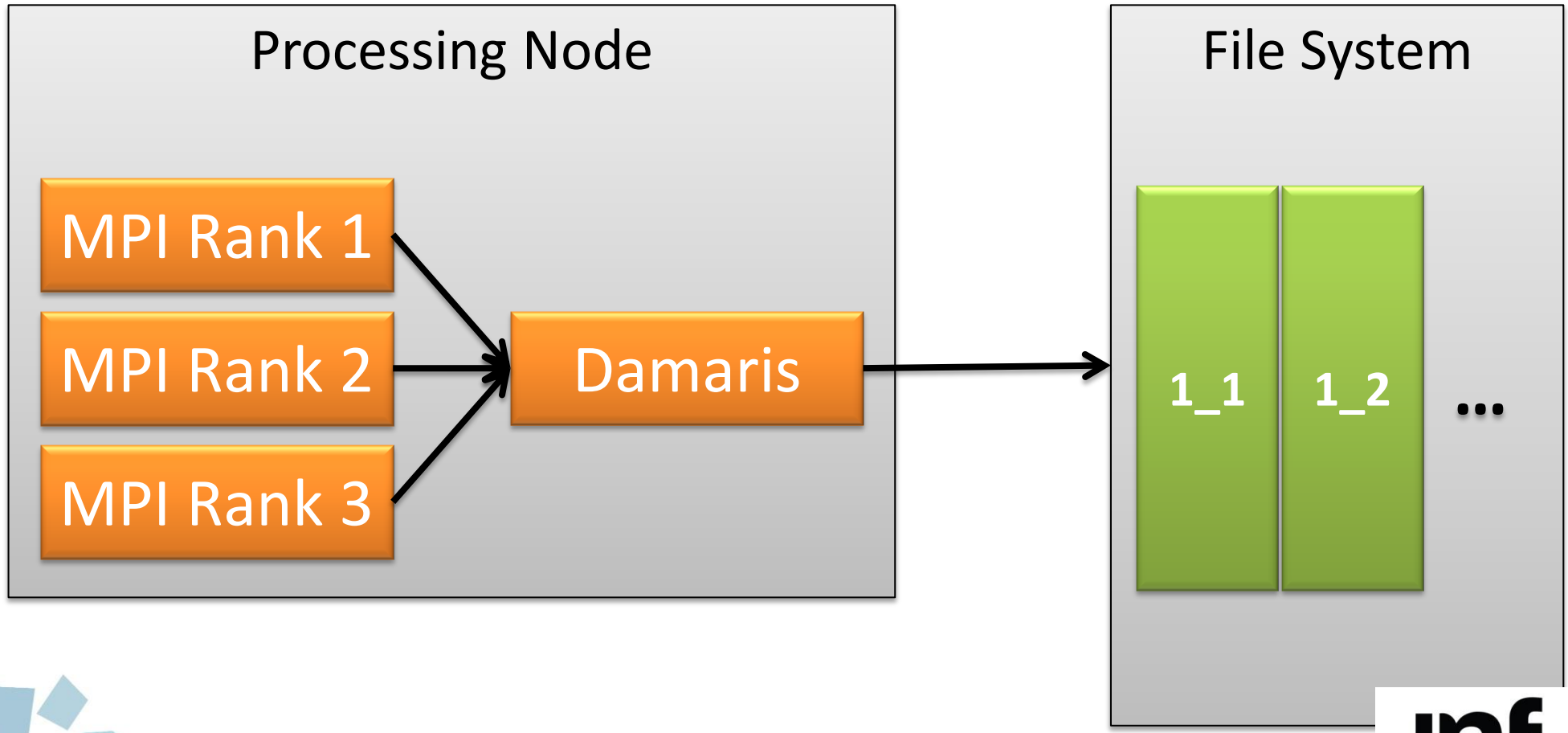
OLAM + Damaris

- **Plugin 1: File-per-process**



OLAM + Damaris

- **Plugin 2: Merger**



Agenda

OLAM and its Performance Problem

OLAM + Damaris

Performance Results

Conclusions

Future Work

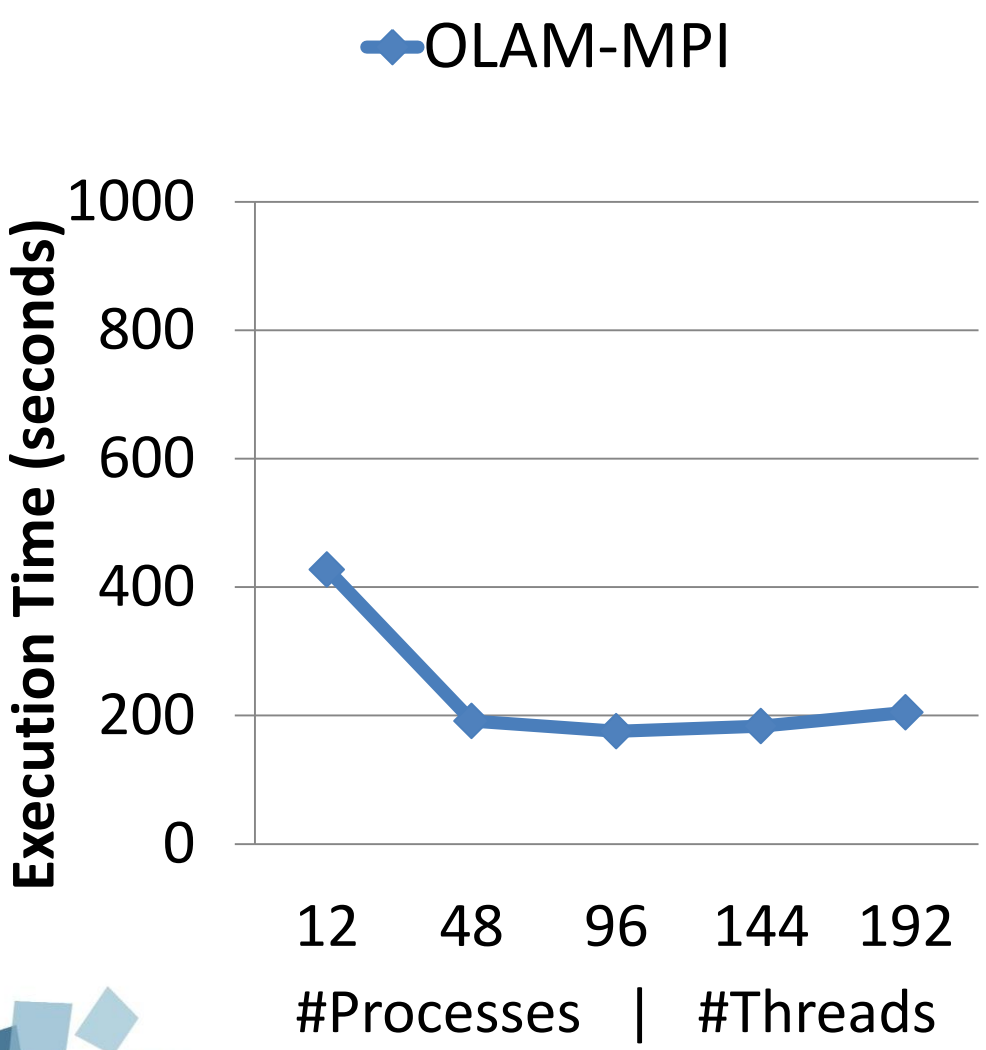
Performance Evaluation

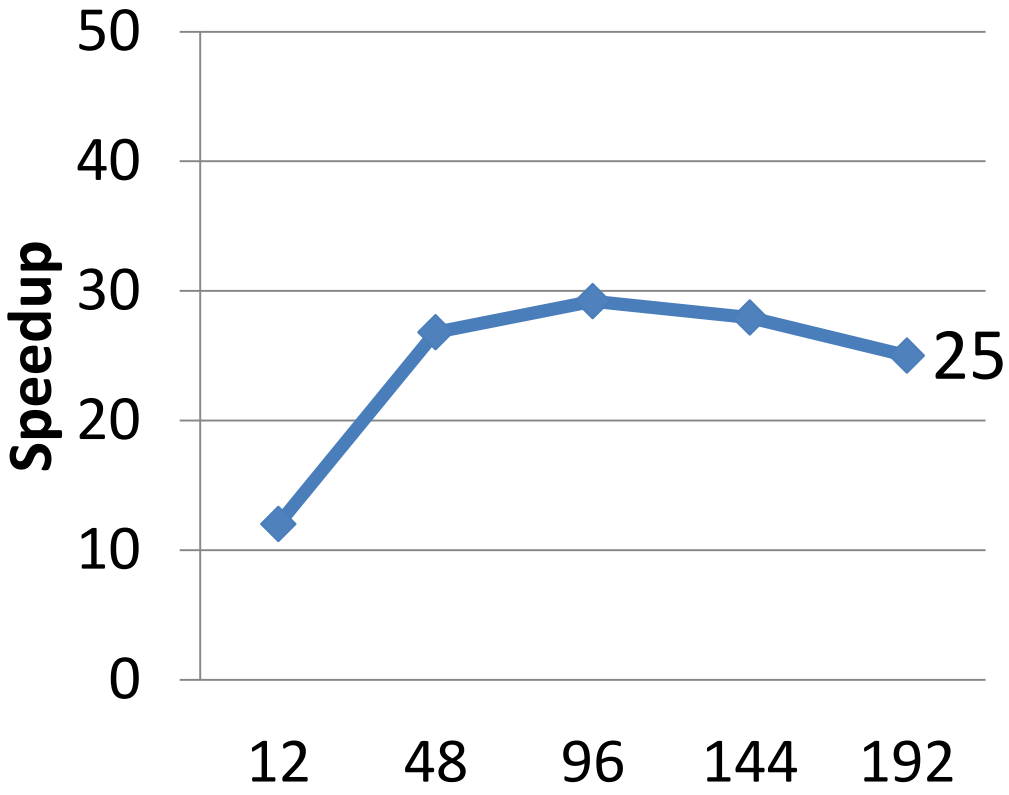
1. OLAM – MPI
2. OLAM – MPI + OpenMP
3. OLAM + Damaris – File-per-process plugin
4. OLAM + Damaris – Merger plugin

Performance Evaluation

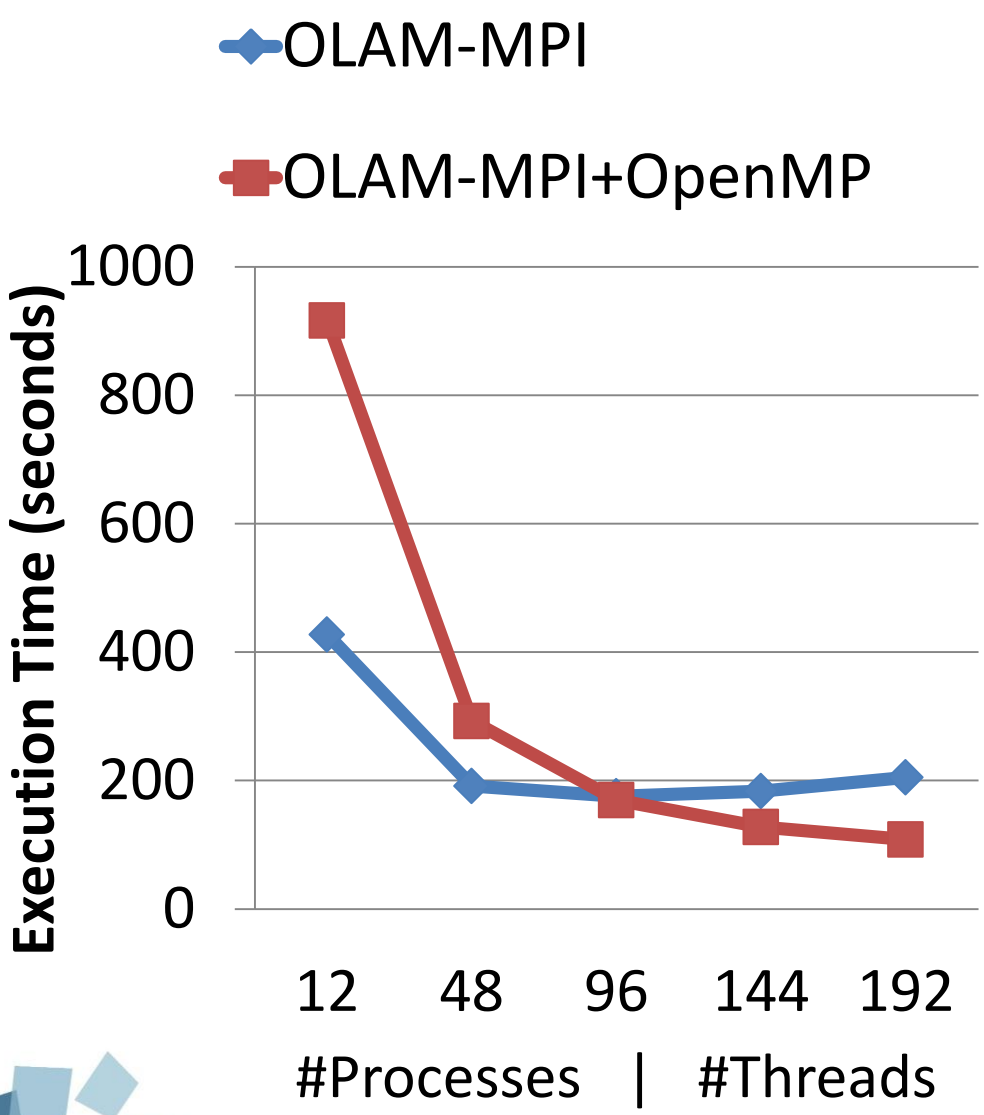
- Clusters @ Rennes . Grid5000
  - 16 nodes from **Parapluie** (processing nodes)
    - **12 cores**
  - 10 nodes from Parapide (FS nodes)
- OrangeFS (**PVFS**)
  - 10 servers (meta and data servers)

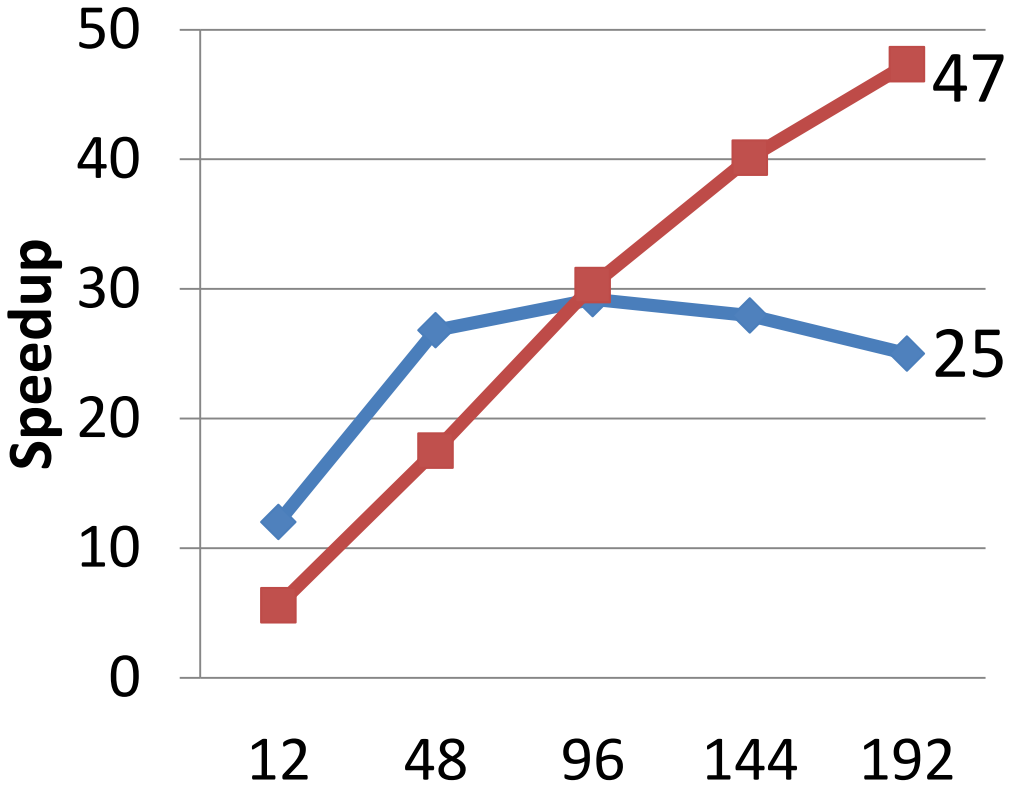
Performance Evaluation



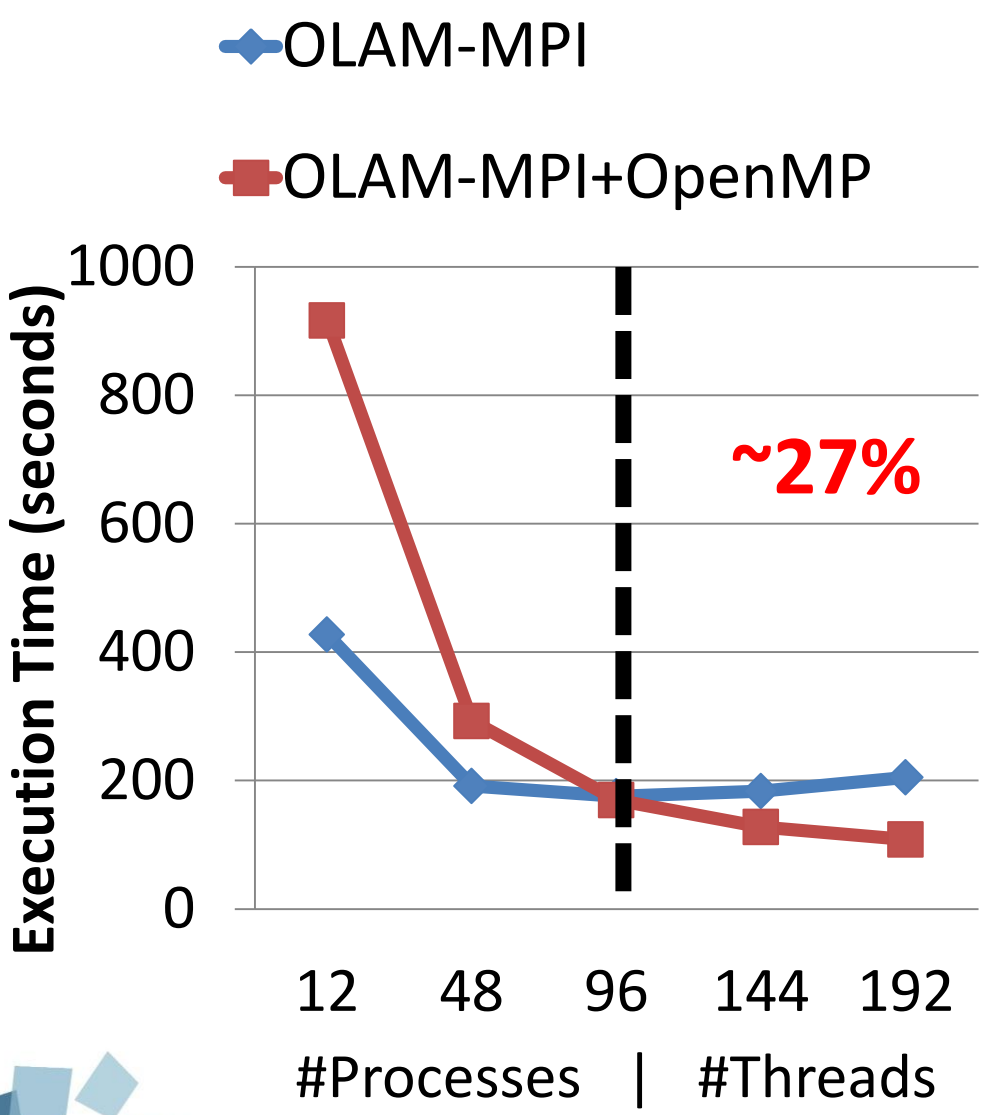


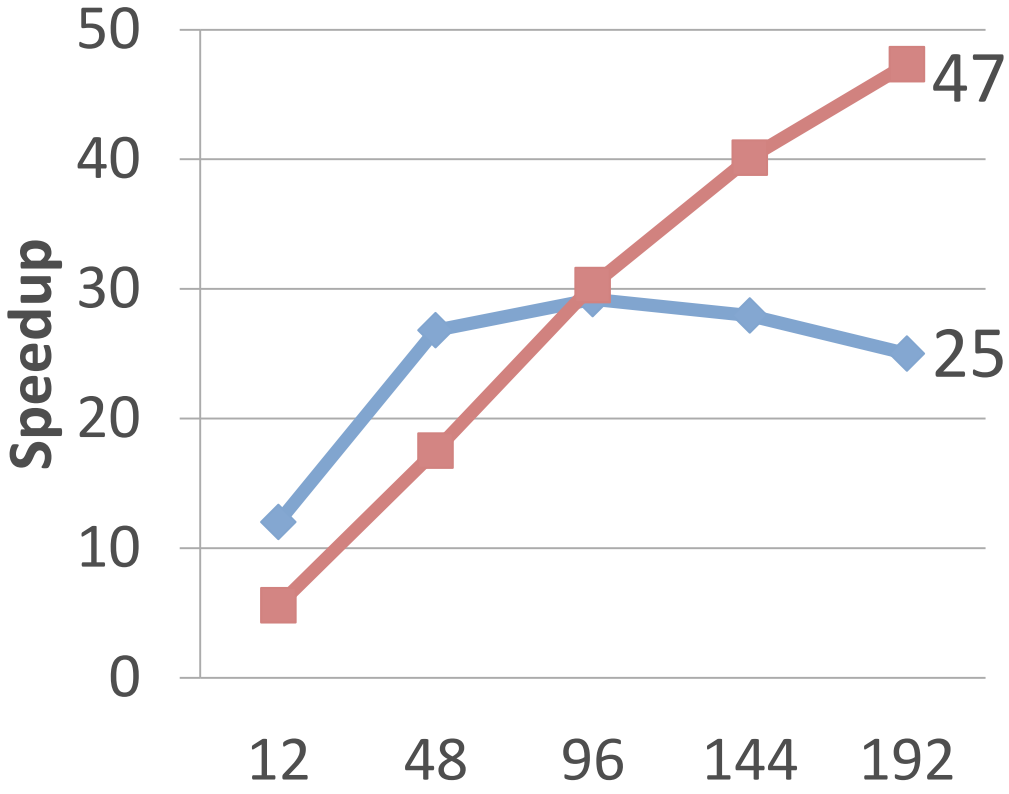
Performance Evaluation



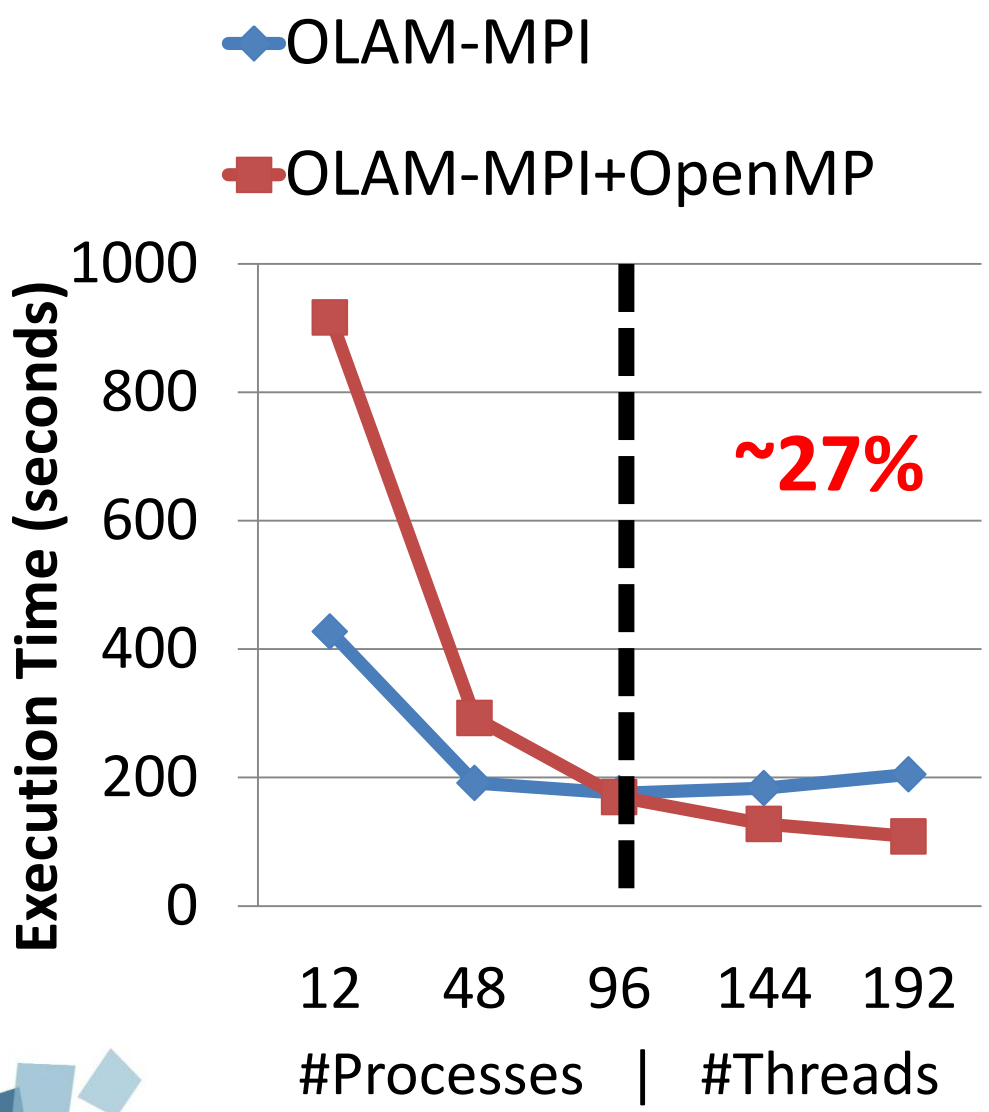


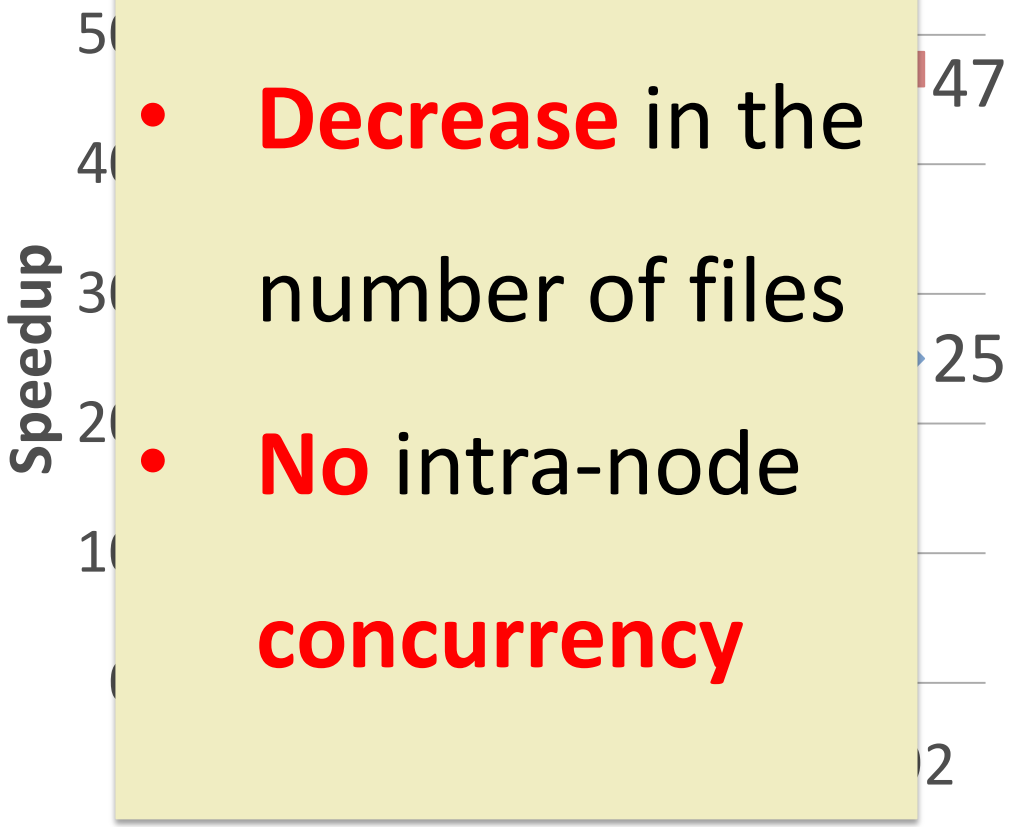
Performance Evaluation





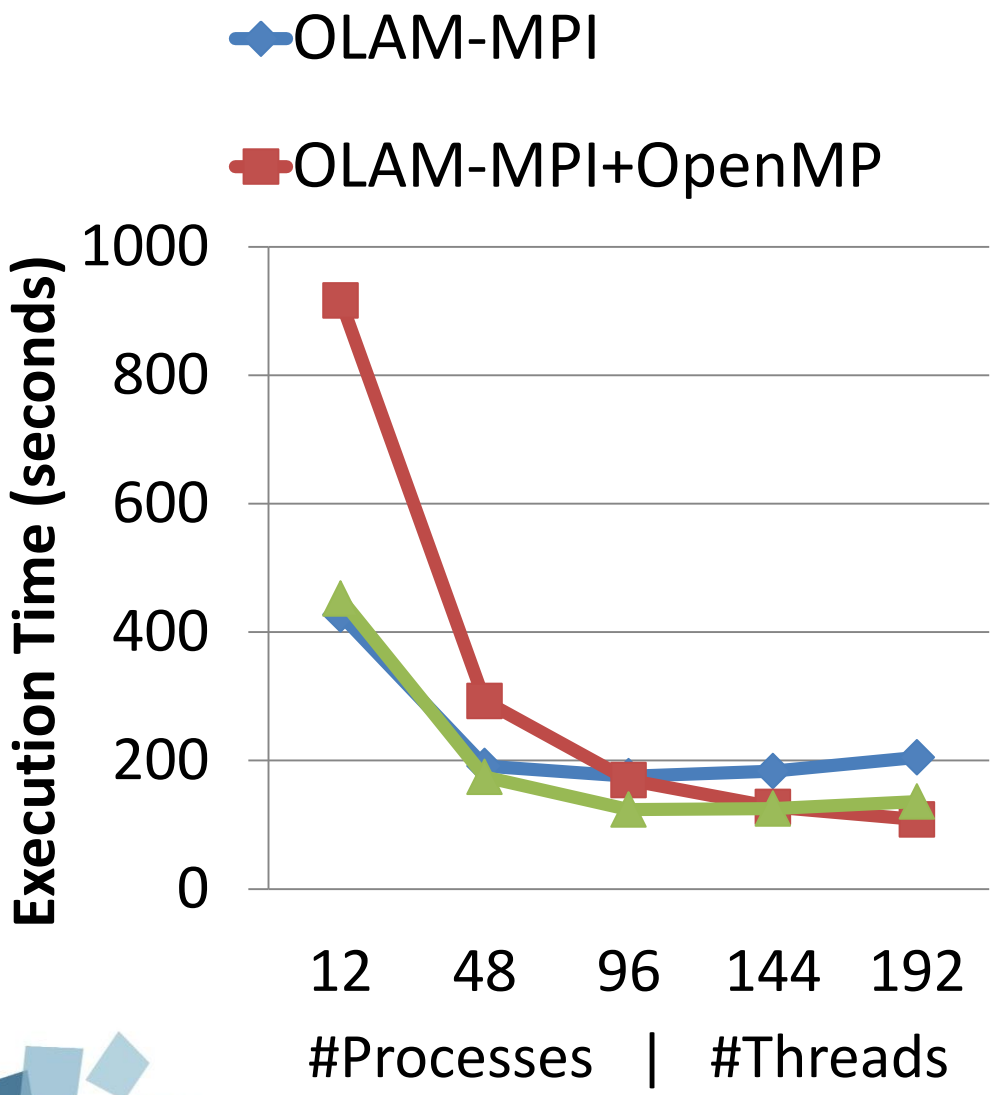
Performance Evaluation

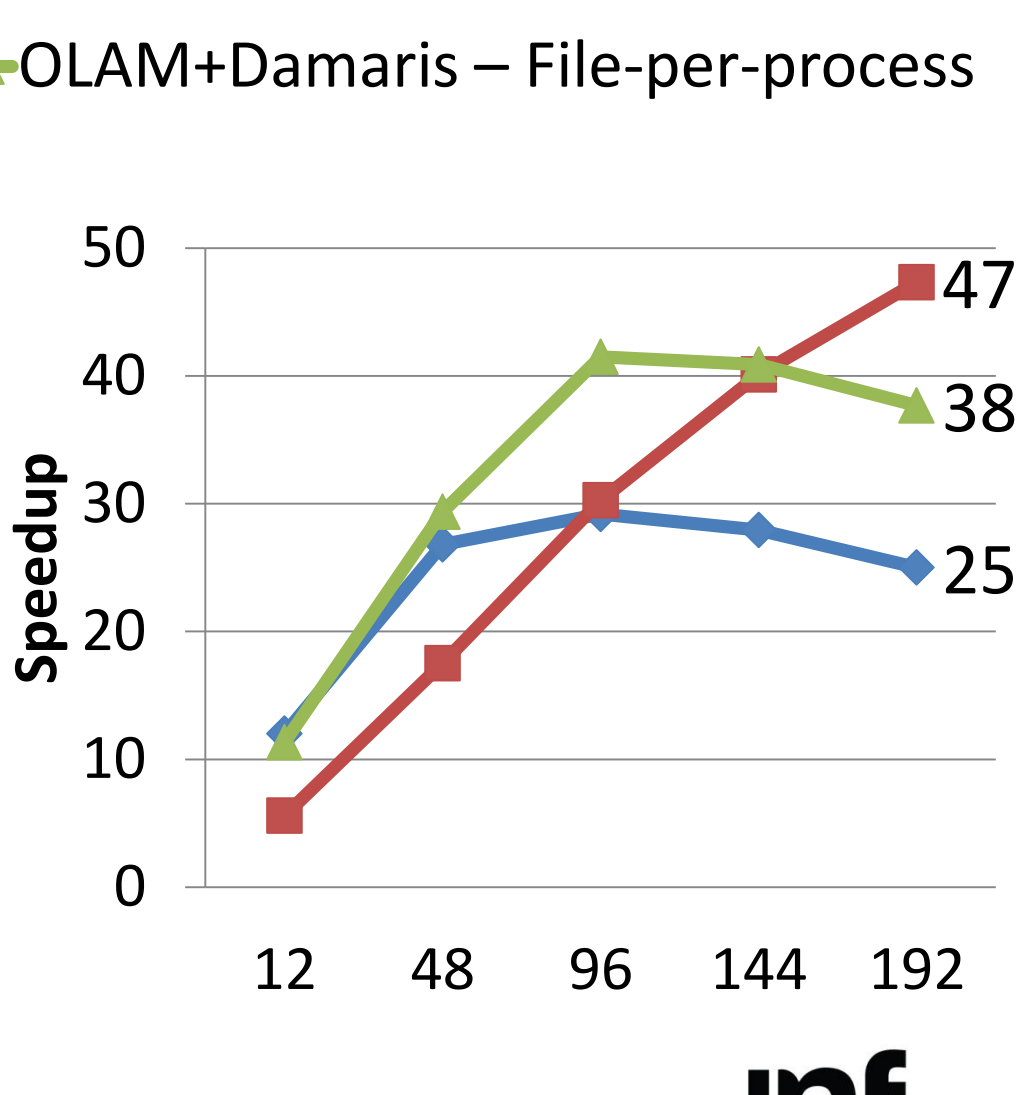




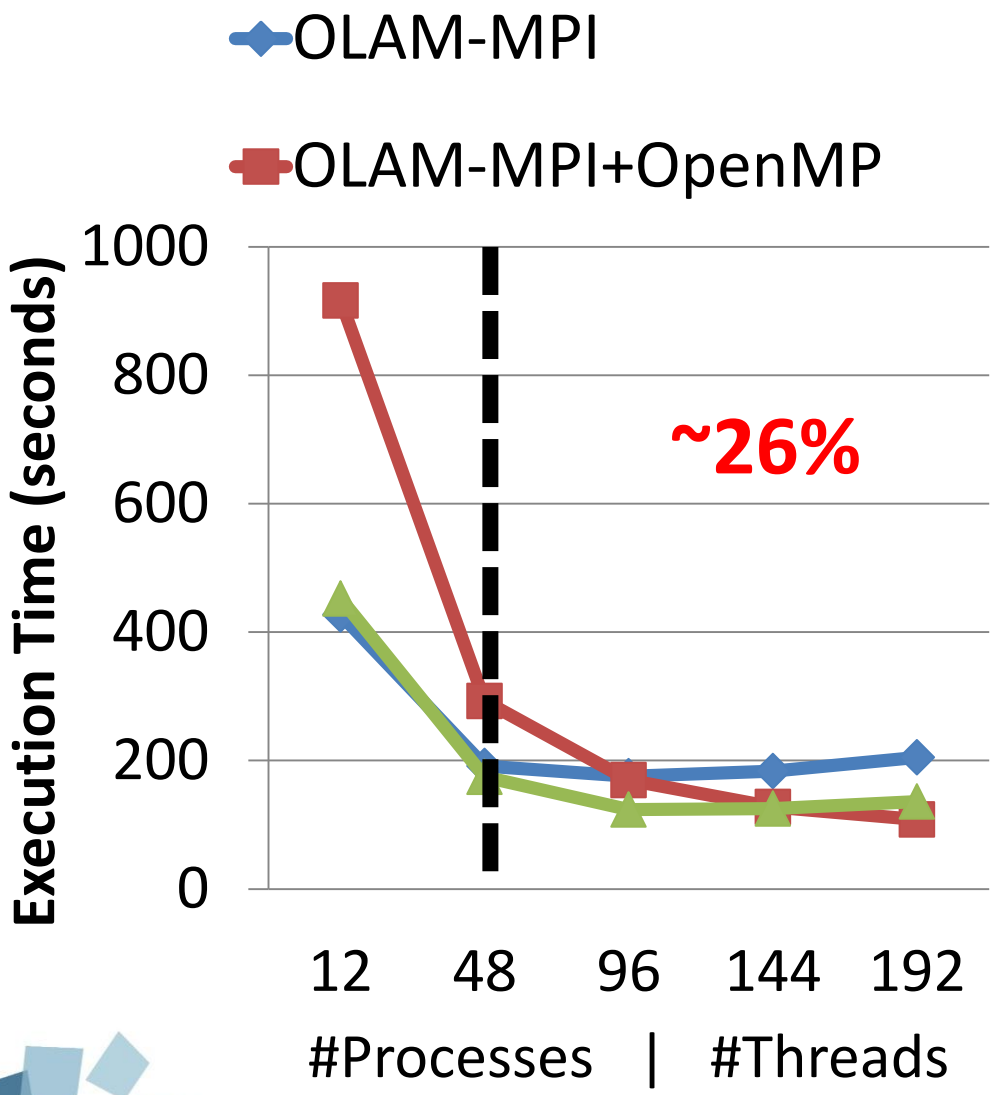
- **Decrease** in the number of files
- **No** intra-node **concurrency**

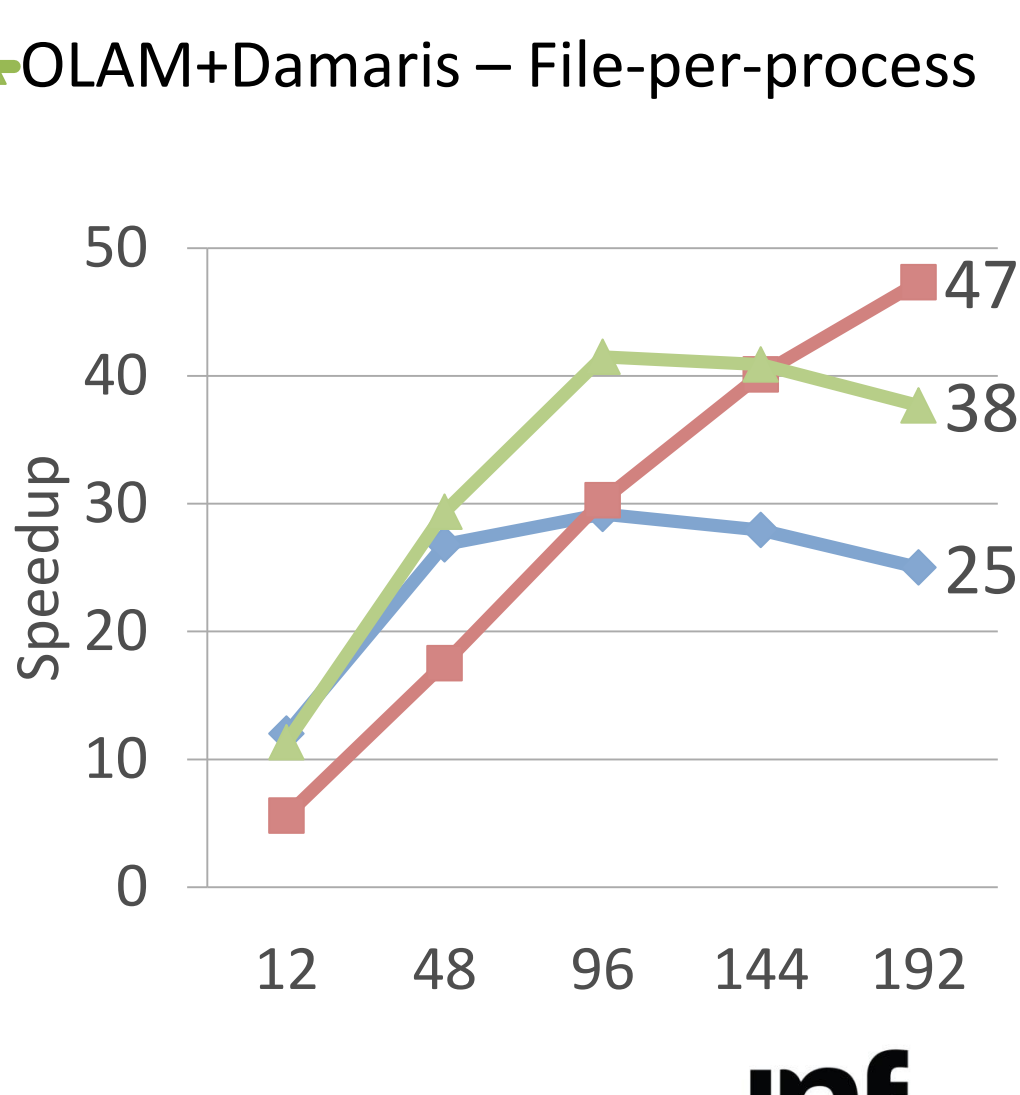
Performance Evaluation



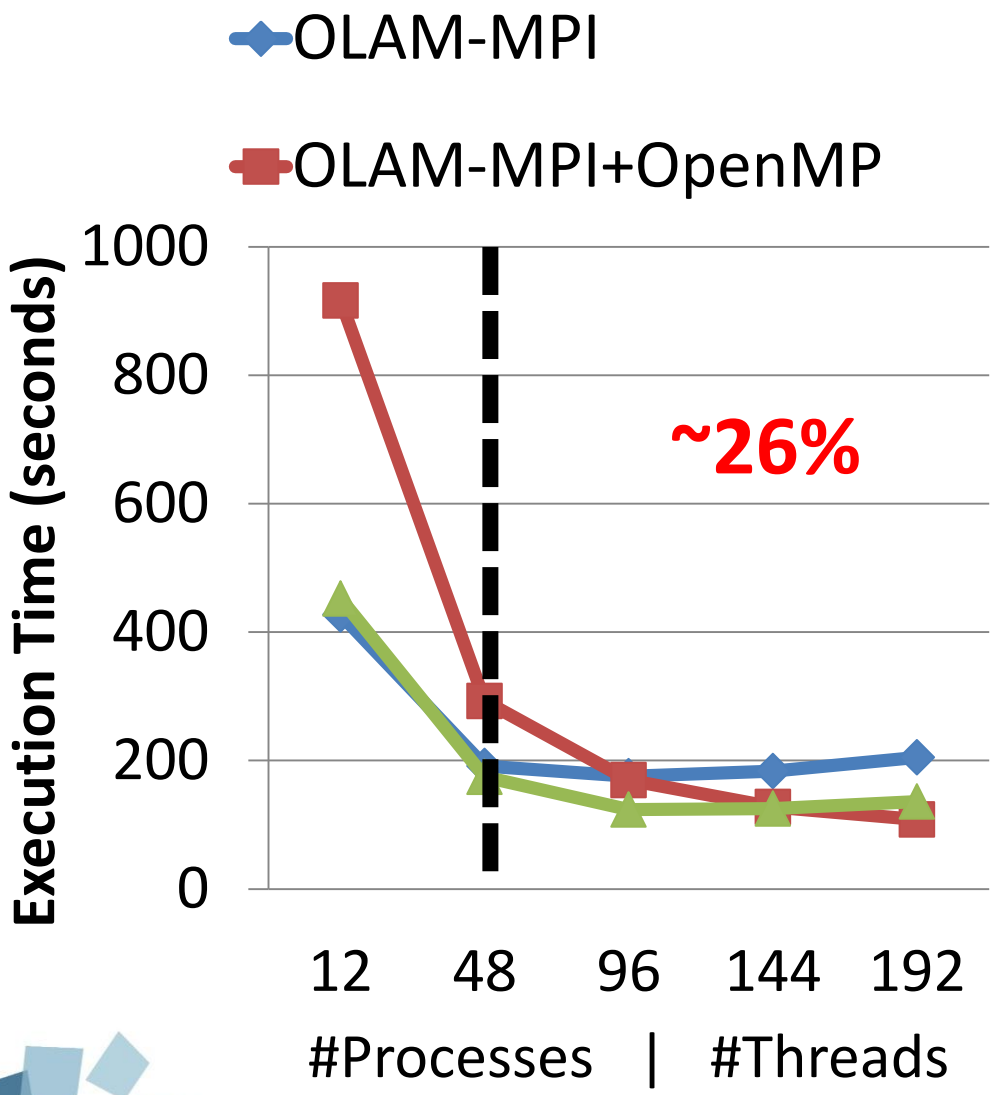


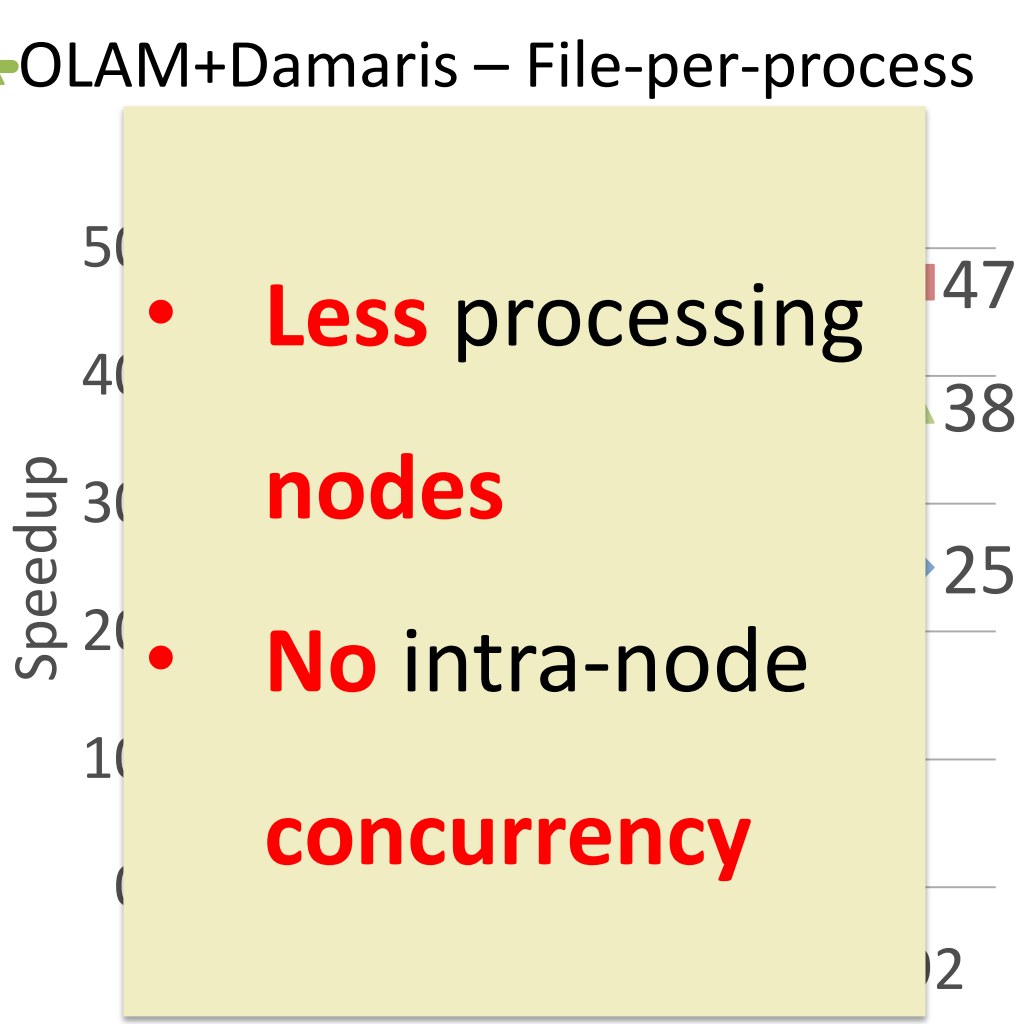
Performance Evaluation





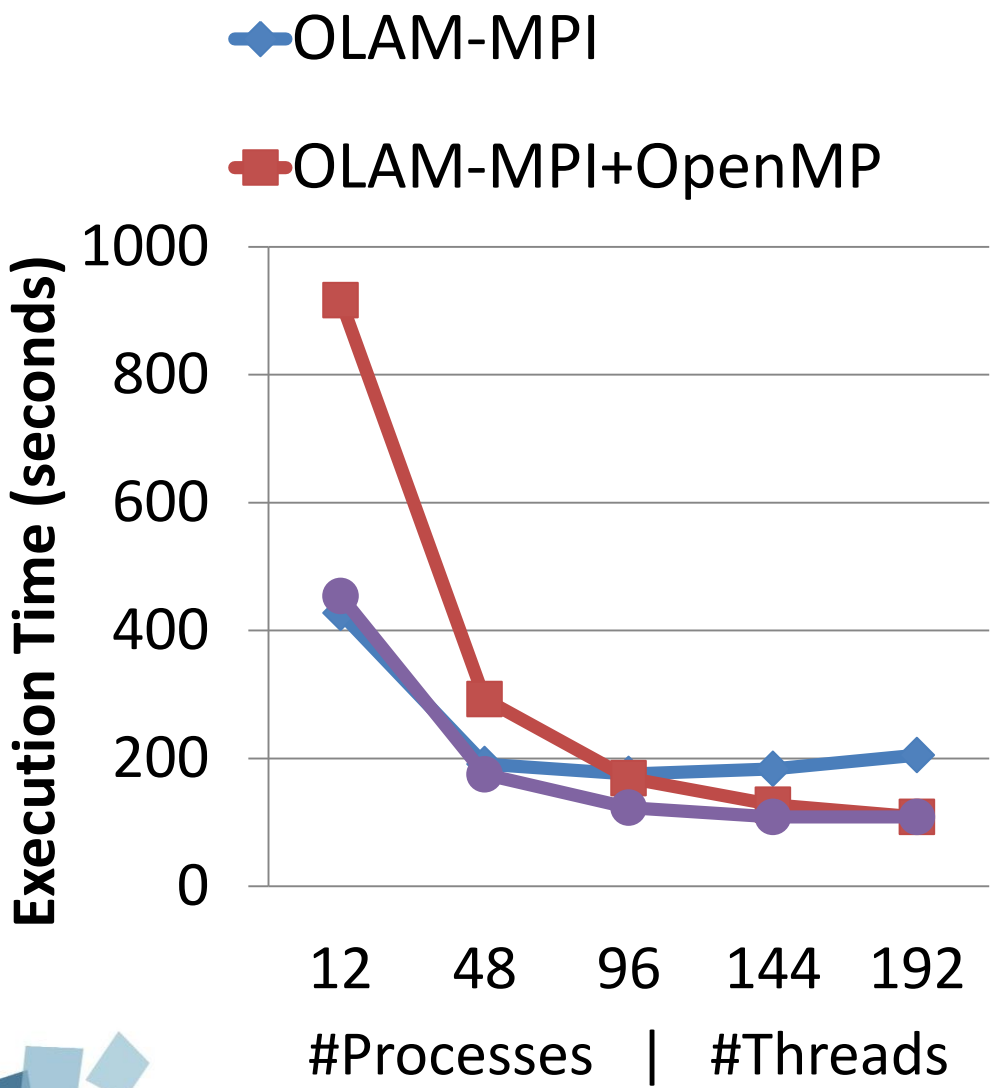
Performance Evaluation

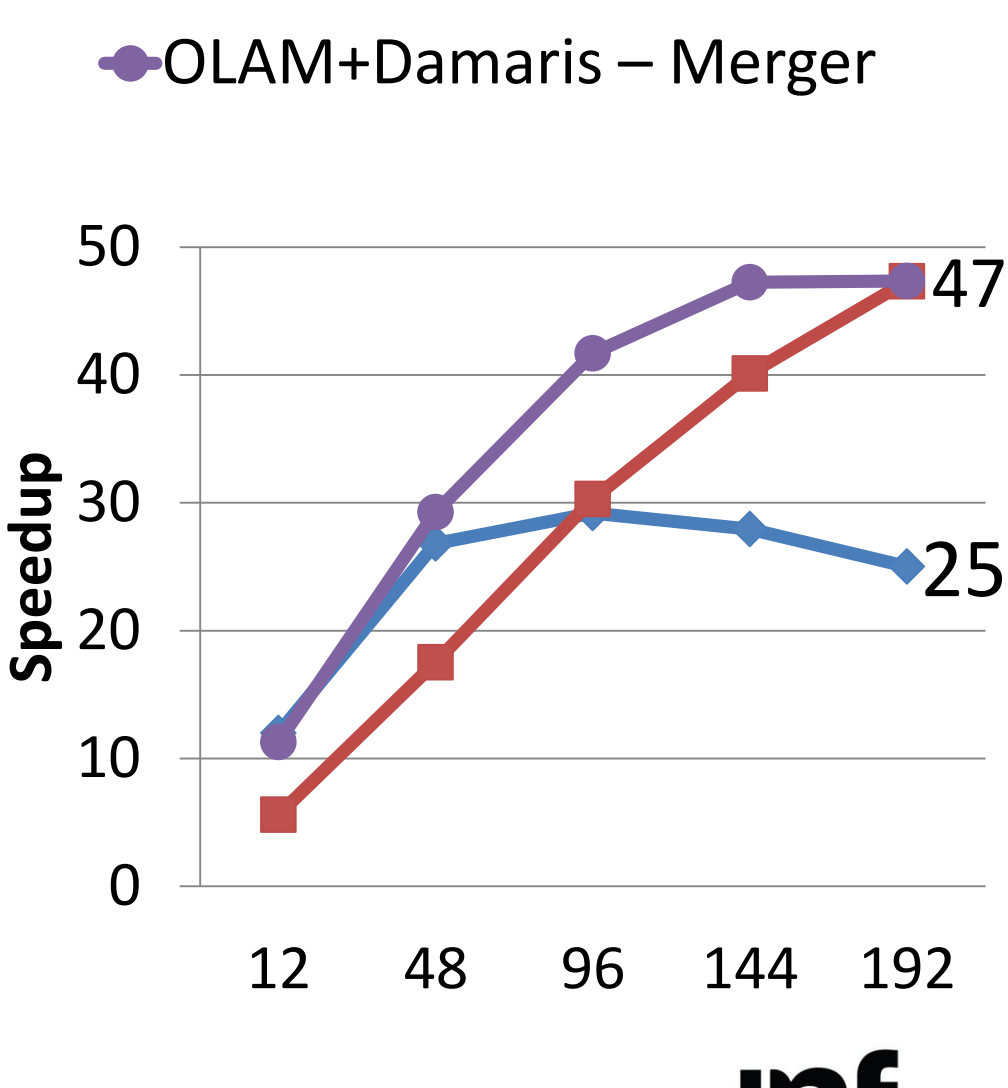




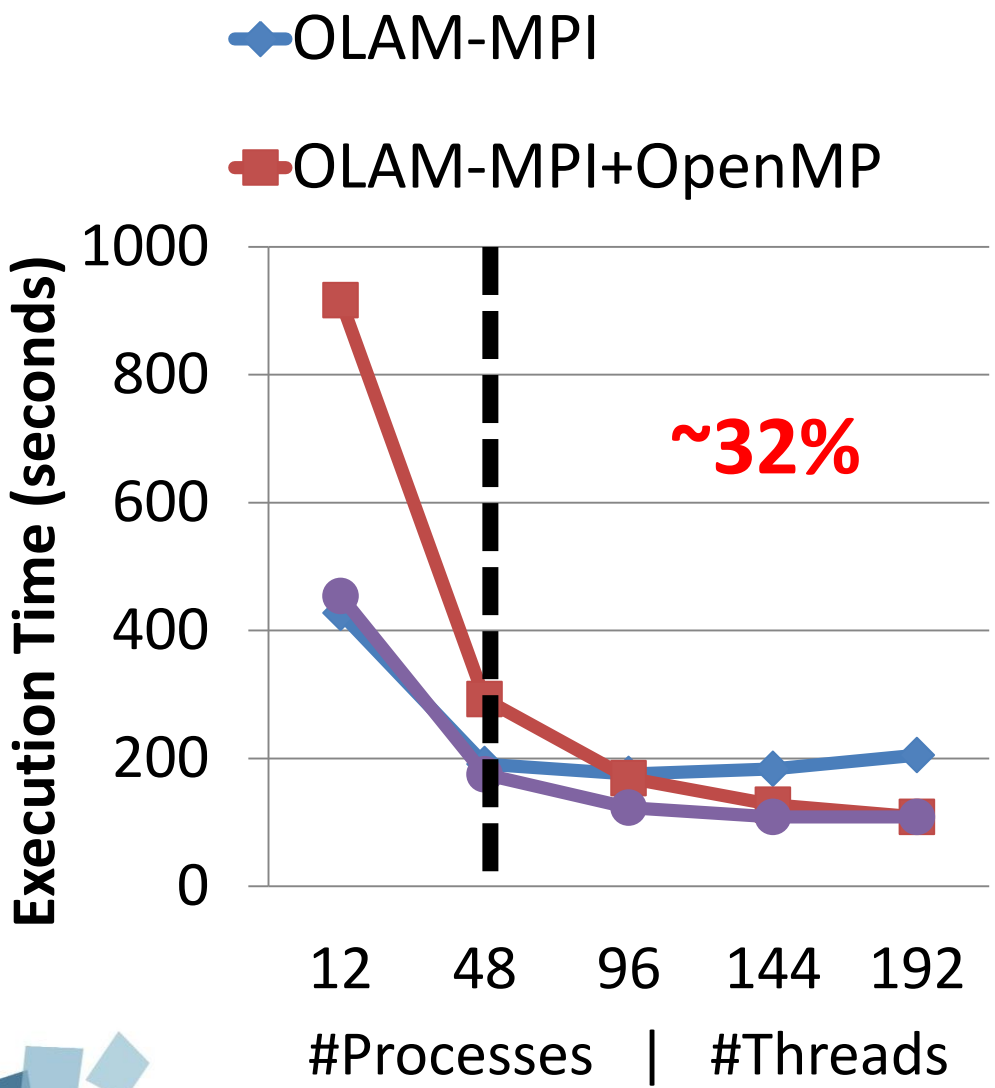
- **Less** processing **nodes**
- **No** intra-node **concurrency**

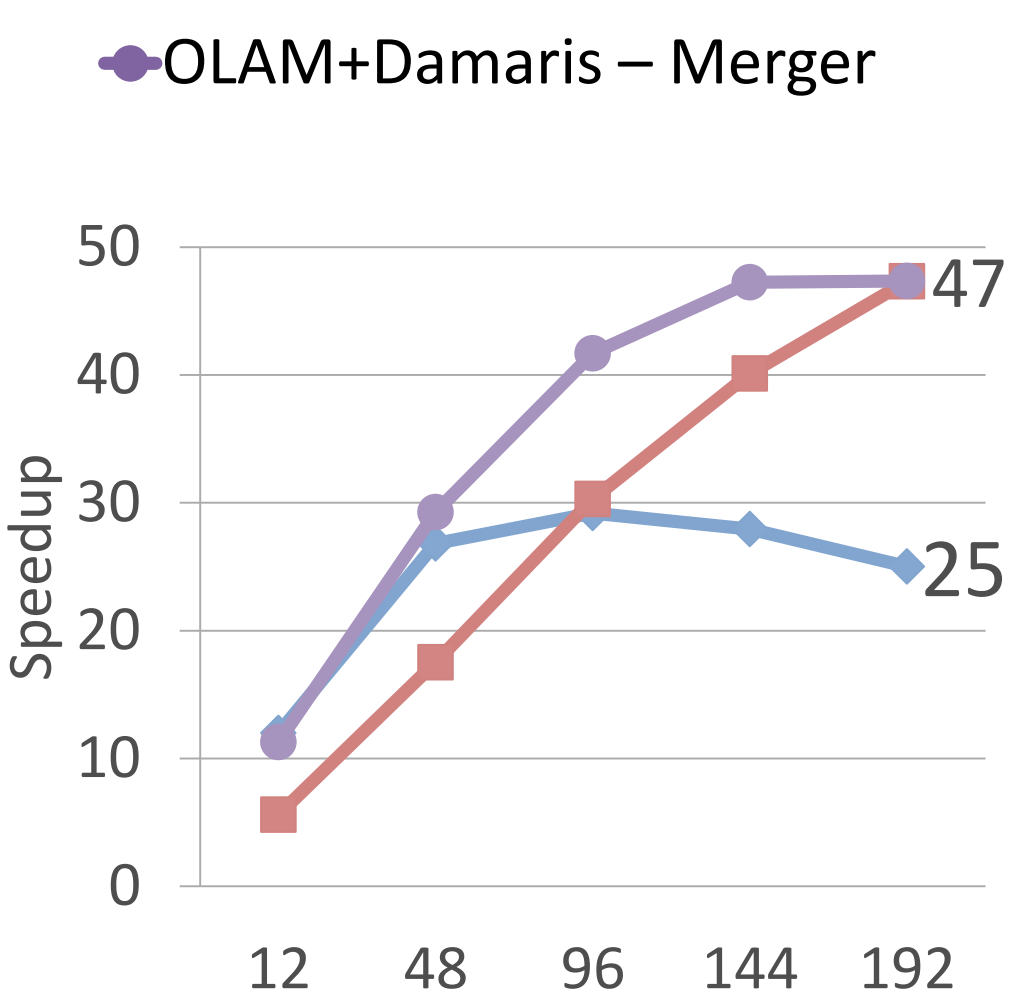
Performance Evaluation



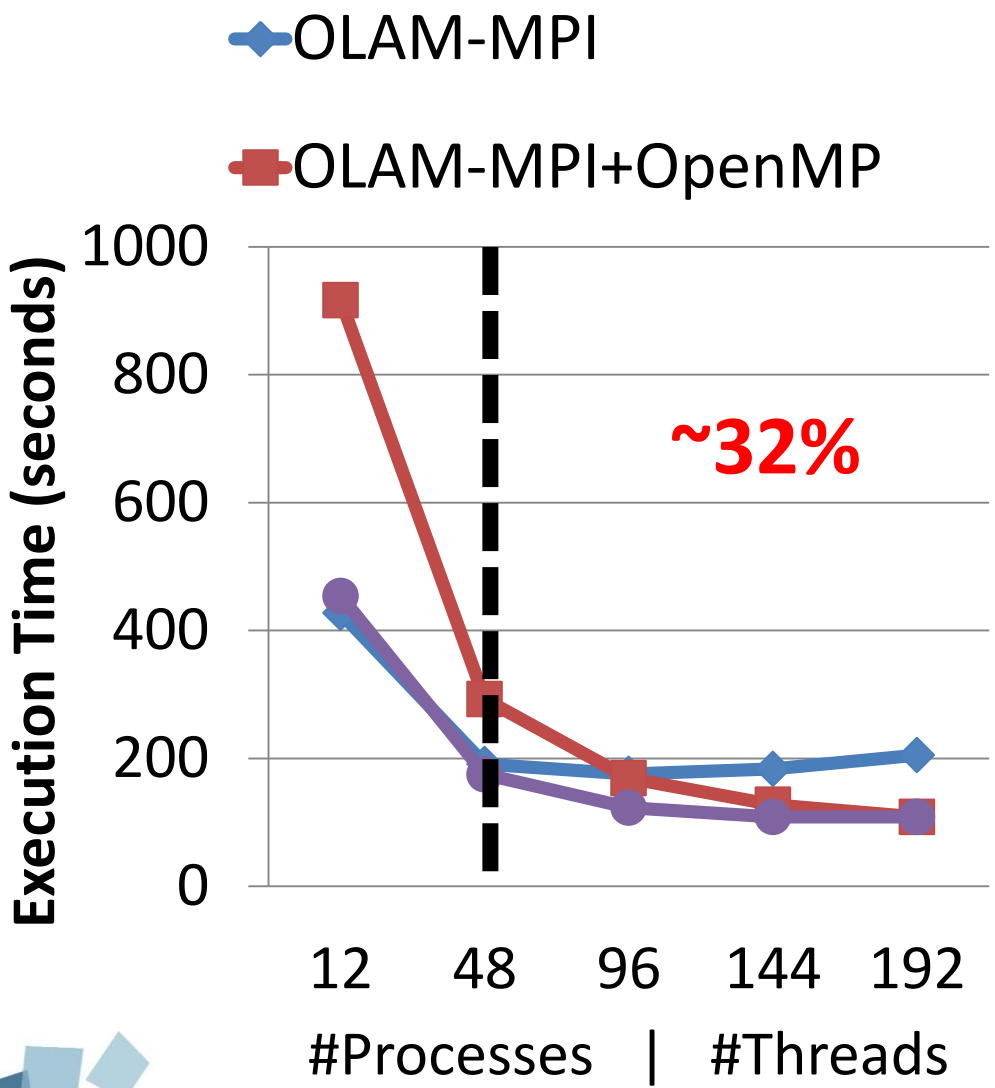


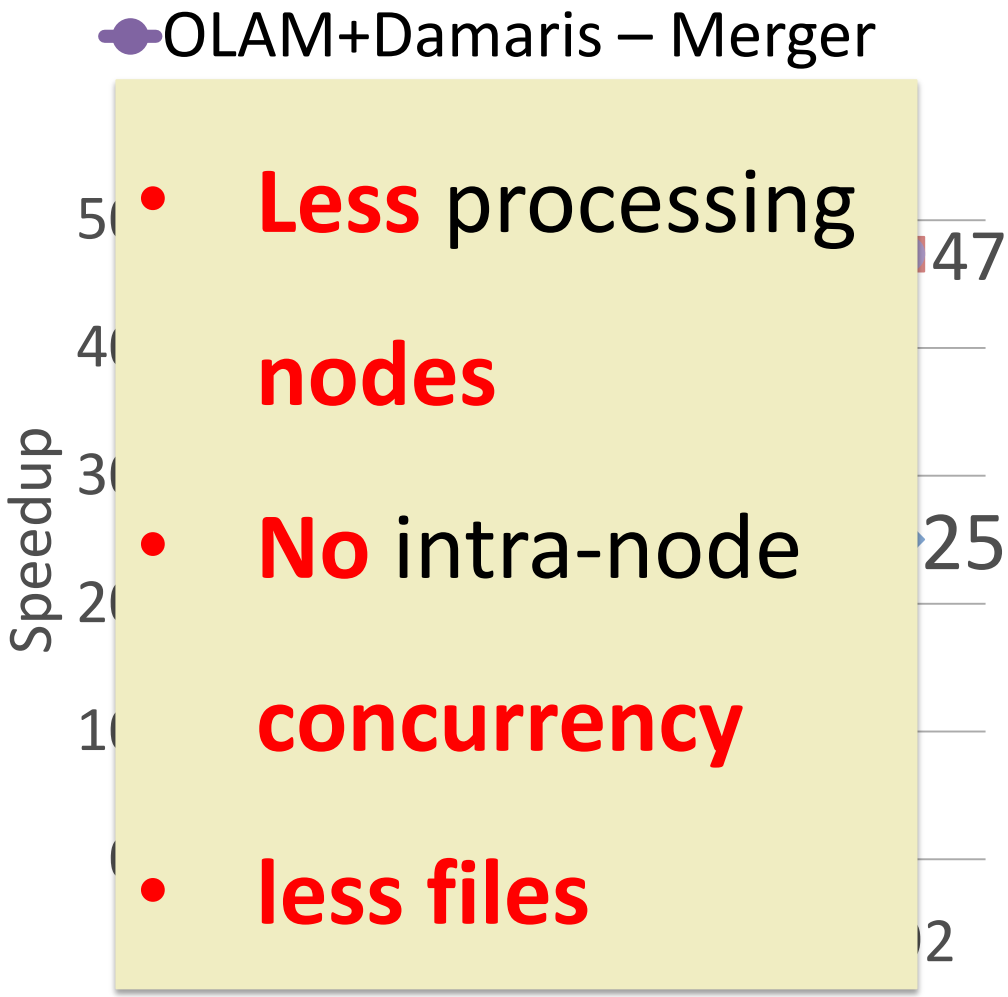
Performance Evaluation





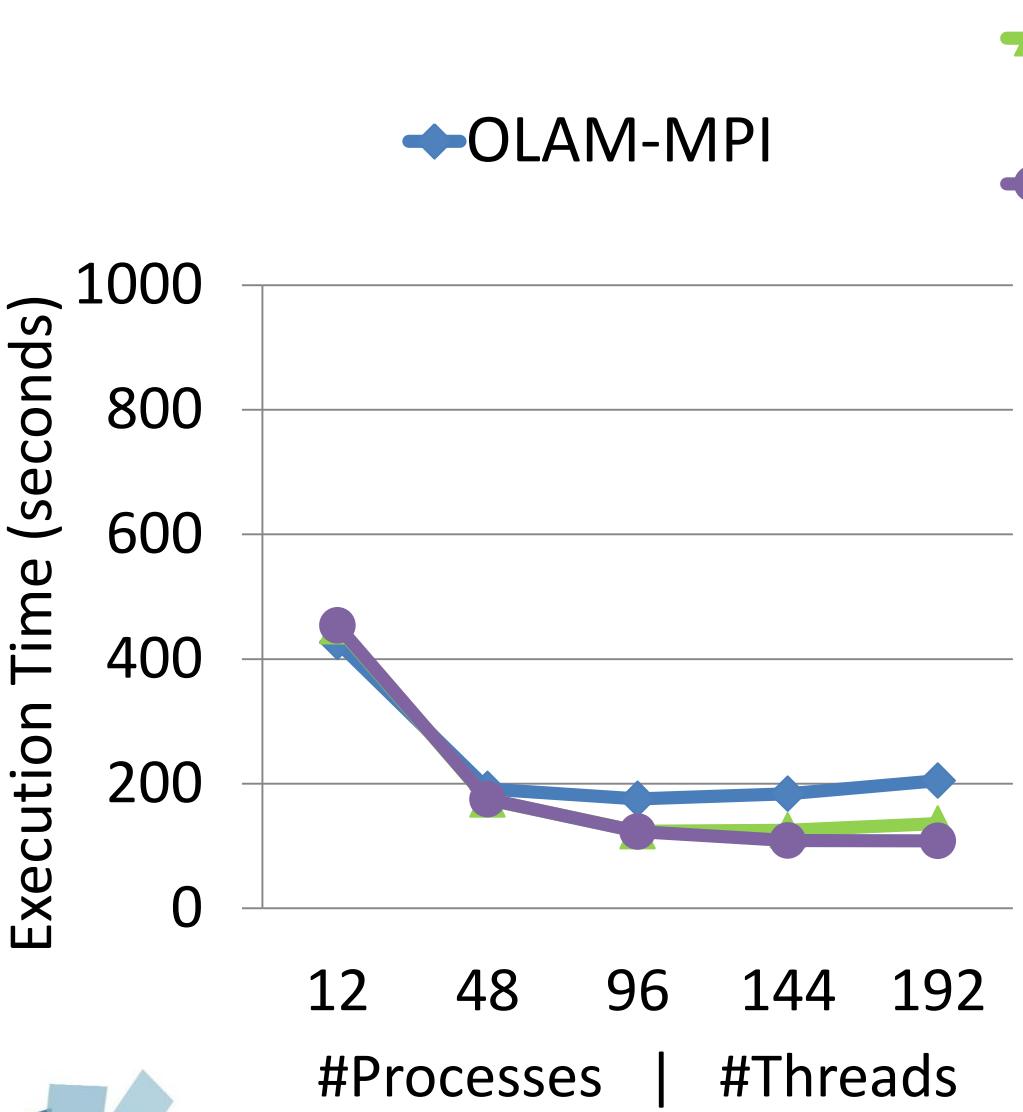
Performance Evaluation

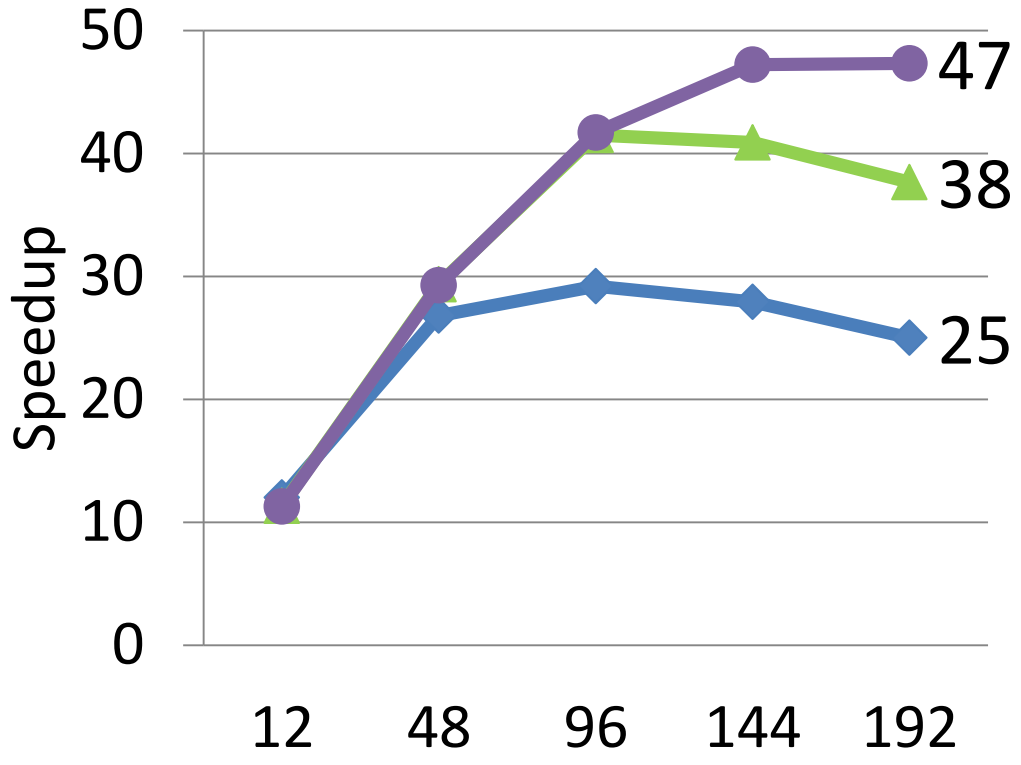




- **Less** processing **nodes**
- **No** intra-node **concurrency**
- **less files**

Performance Evaluation





Agenda

OLAM and its Performance Problem

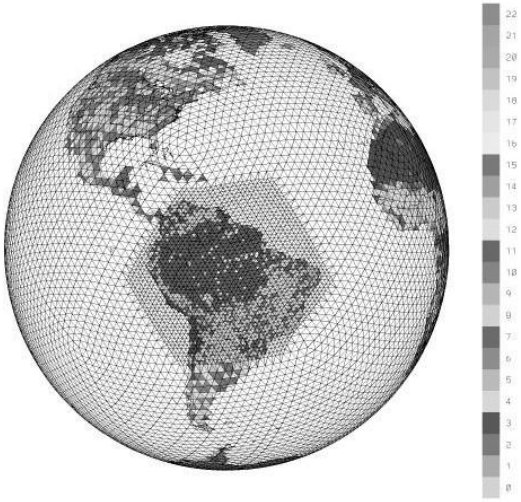
OLAM + Damaris

Performance Results

Conclusions

Future Work

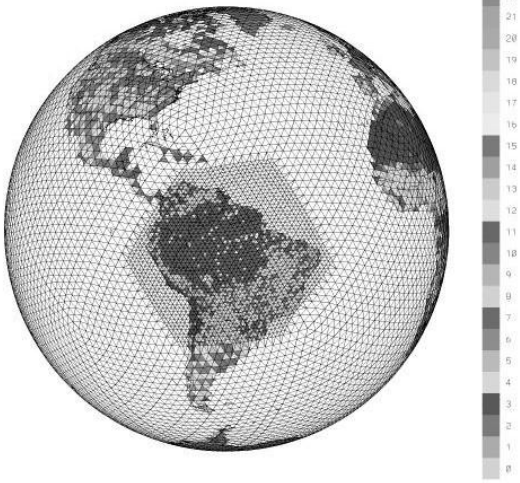
Conclusions



Aplication (**OLAM**)

Poor **I/O** performance

Conclusions



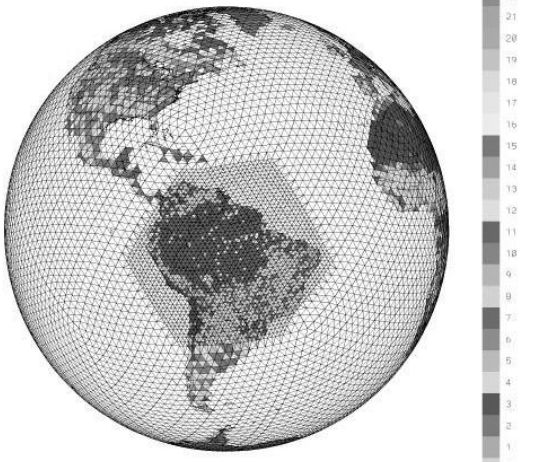
Aplication (**OLAM**)

Poor **I/O** performance

OpenMP

Damaris

Conclusions



Aplication (**OLAM**)

Poor **I/O** performance

OpenMP → **~27%**

Damaris → **~26%**

Damaris → **~32%**

Conclusions

- **OLAM+Damaris** was simple
  - Results better than the MPI+OpenMP version
- It's easy to **test new I/O approaches**
- **Visualization** tool

Agenda

OLAM and its Performance Problem

OLAM + Damaris

Performance Results

Conclusions

Future Work

Future Work

- **OLAM+Damaris**
  - Keep investigating!
  - Can we use it to input too?

- Damaris
  - How much information can we get about the **applications' access pattern**? (Francieli)
  - Can it help **pipeline-style scientific workflow**? (Rodrigo Kassick)

Investigating I/O approaches to improve performance and scalability of the Ocean-Land-Atmosphere Model

Thank you!

Rodrigo Virote Kassick[1 2], Francieli Zanon Boito[1 2], Philippe Navaux[1], Yves Denneulin[2]

[1] GPPD – II – **Federal University of Rio Grande do Sul** (UFRGS), Brazil

[2] INRIA – LIG – **Grenoble University**, France

The Seventh Workshop of the INRIA-Illinois Joint Laboratory on Petascale Computing

June 13th 2012