Towards On-Demand I/O Forwarding in HPC Platforms

Jean Luca Bez, Francieli Zanon Boito, Ramon Nou,
Alberto Miranda, Toni Cortes, and Philippe O. A. Navaux

jean.bez@inf.ufrgs.br

PDSW 2020 – International Parallel Data Systems Workshop

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Inria informatiques mathématiques

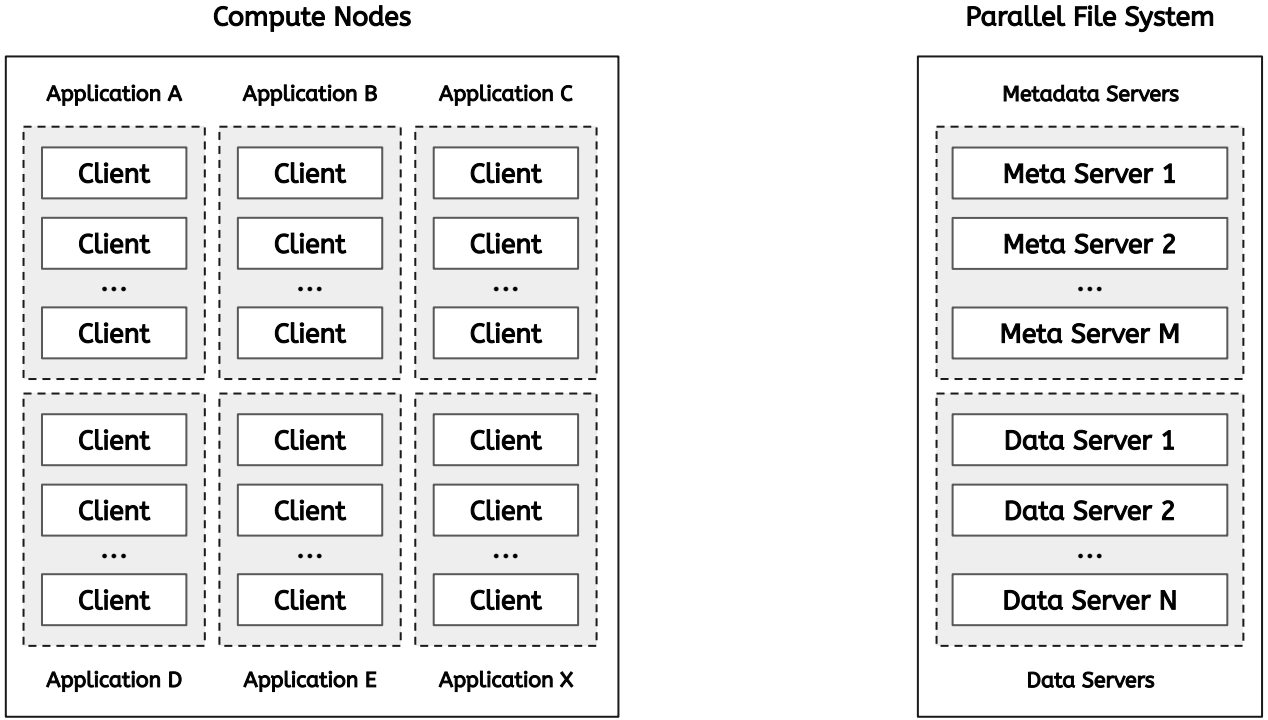
UPC

SC20
Everywhere we are | more than hpc.

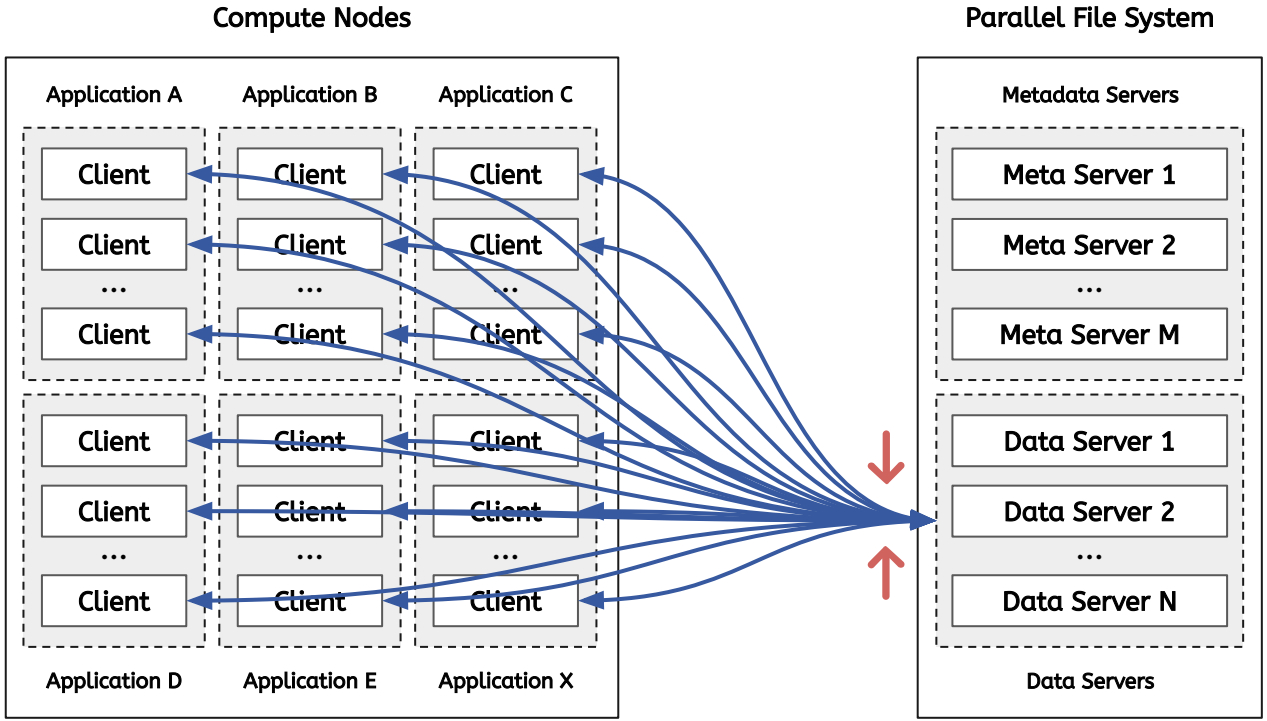
Agenda

- The I/O Forwarding Layer
- Motivation
- **FORGE** The I/O Forwarding Explorer
- Forwarding in MareNostrum 4
- Forwarding in SDumont
- Conclusion

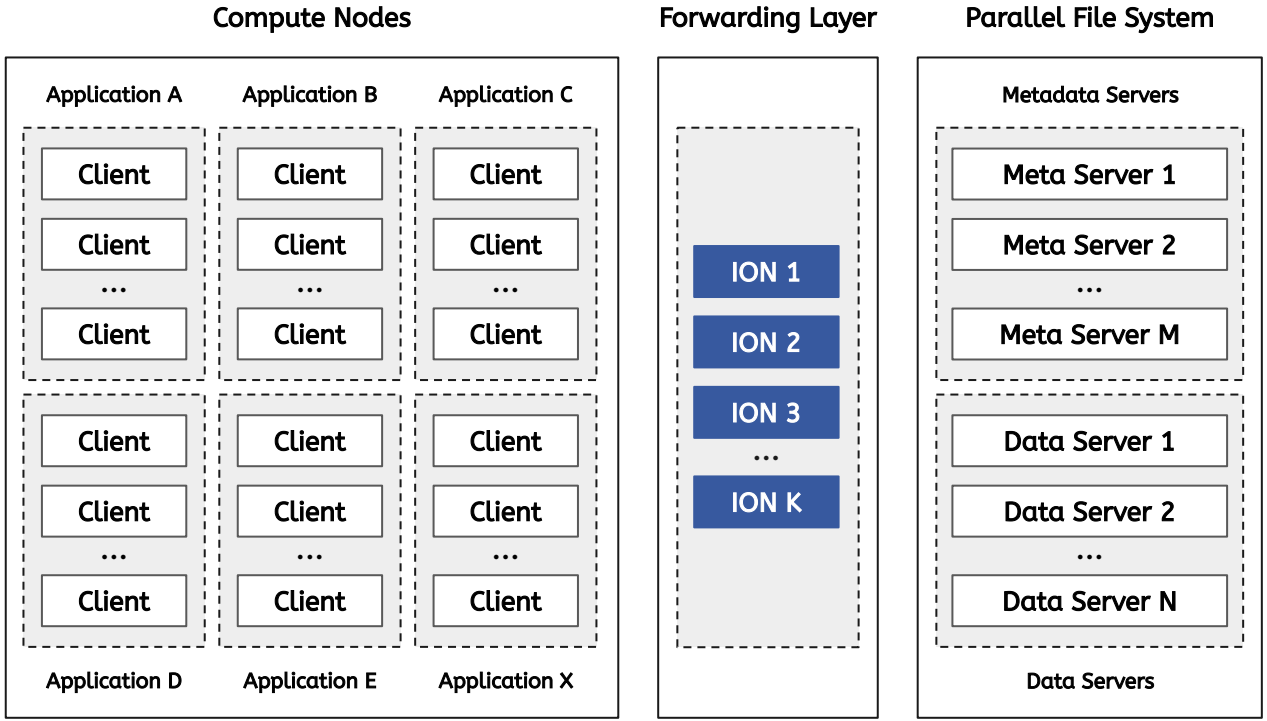
The I/O Forwarding Layer



The I/O Forwarding Layer

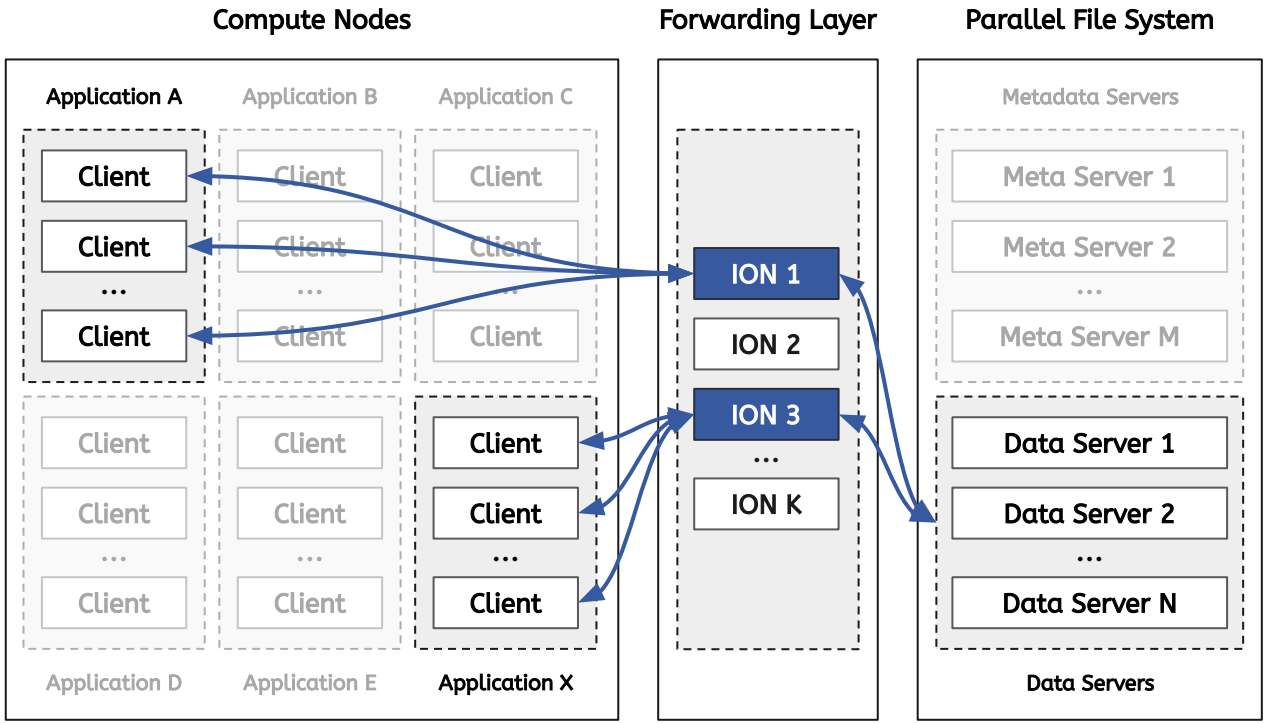


The I/O Forwarding Layer



SC20
Everywhere we are | more than hpc.

The I/O Forwarding Layer



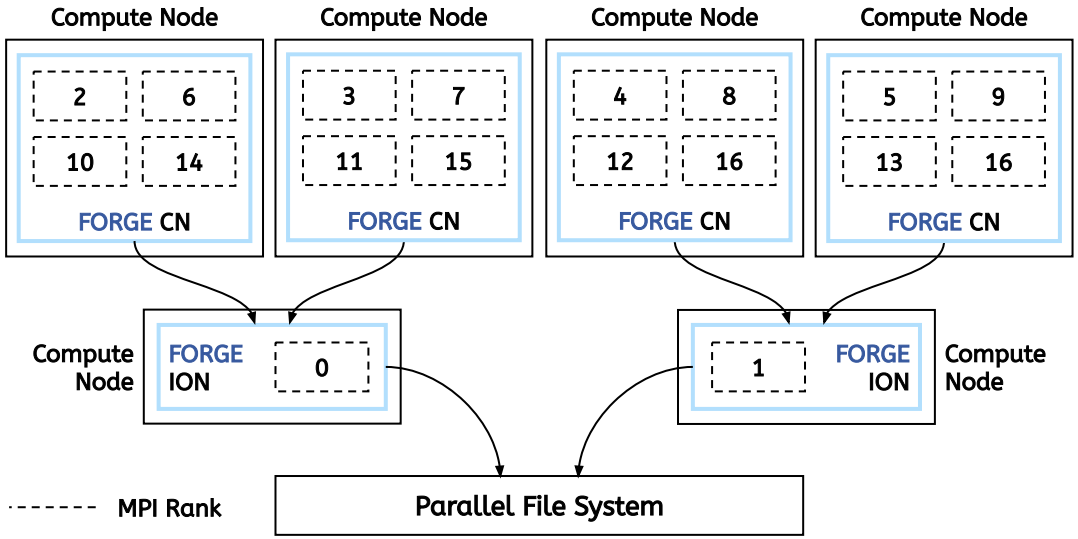
SC20
Everywhere we are | more than hpc.

Motivation

- Investigate the **impact of I/O forwarding** on performance
- Take into account the **application's access pattern**
- Most machines cannot be easily reconfigured
- End-users are not allowed to change this layer
- We need a **research/exploration** alternative!

- When forwarding is the best choice?
- How many I/O nodes should an application use?

FORGE: The I/O FORwardinG Explorer



SC20
Everywhere we are | more than hpc.

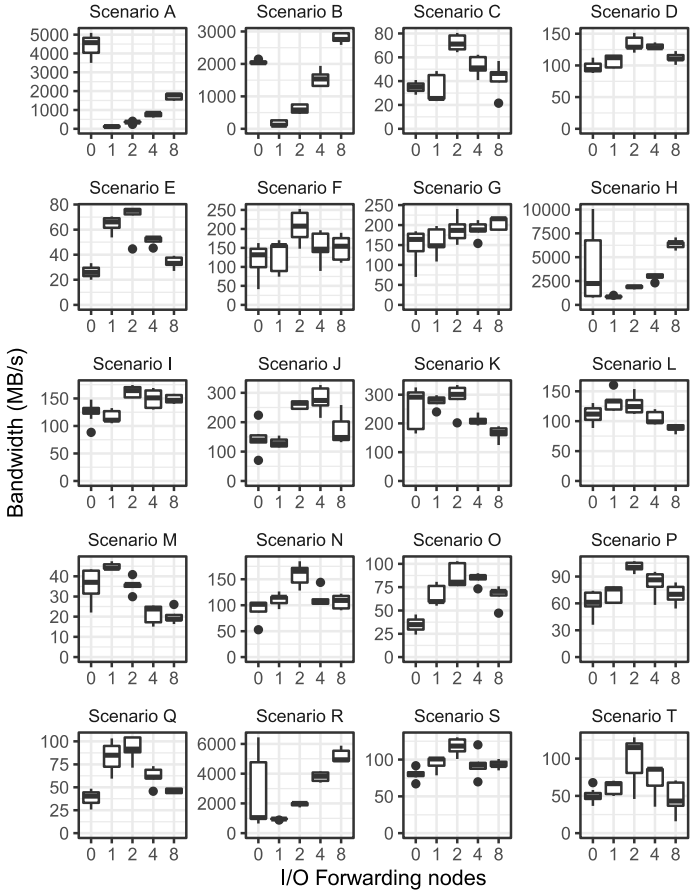
EXPERIMENTS

- **MareNostrum 4** (Spain) and **Santos Dumont** (Brazil) supercomputers
- **189 distinct scenarios** (access patterns and deployments):
  - Compute nodes: 8, 16, and 32
  - Client processes per compute node: 12, 24, and 48

    (96, 192, 384, 768, and 1536 processes in total)
  - File layout: file-per-process or shared file
  - Spatiality: contiguous or 1D-strided
  - Operation: WRITE
  - Request sizes: 32KB, 128KB, 512KB, 1MB, 4MB, 6MB, and 8MB
  - Stonewall: one second

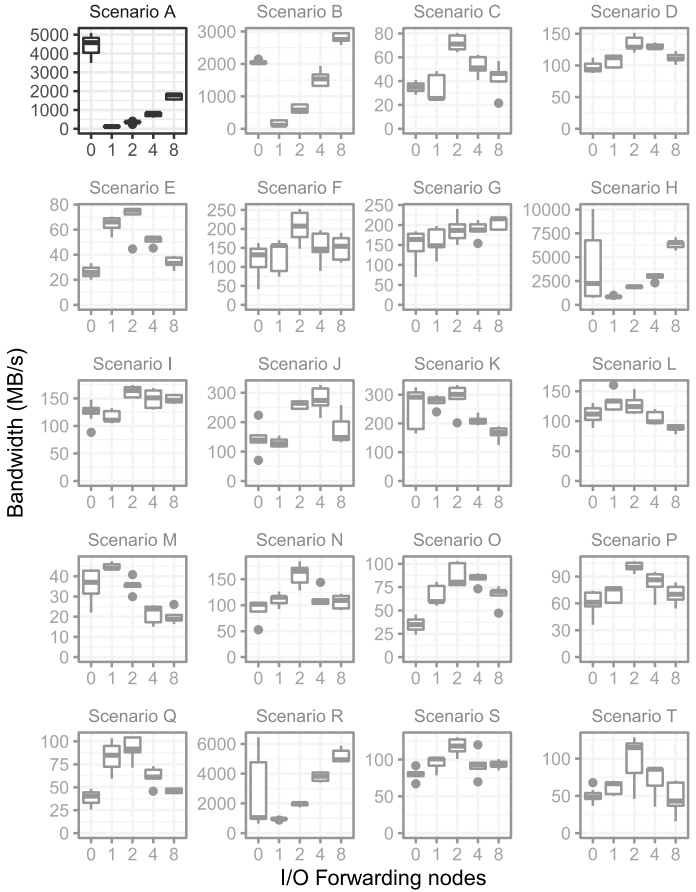
MareNostrum 4

- Bandwidth at client-side
- 5 repetitions for each
- Different days and periods



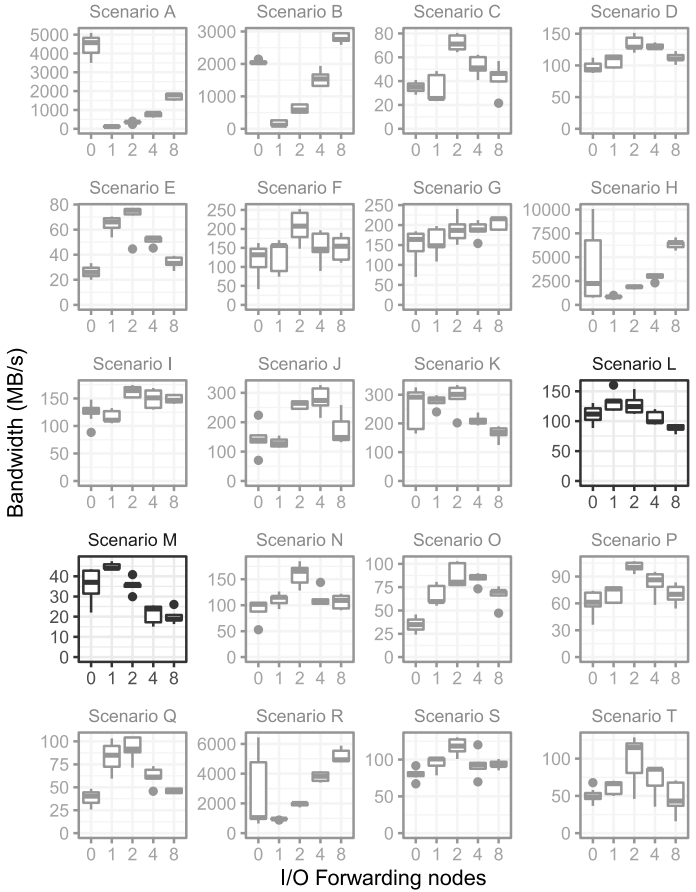
MareNostrum 4

- Bandwidth at client-side
- 5 repetitions for each
- Different days and periods



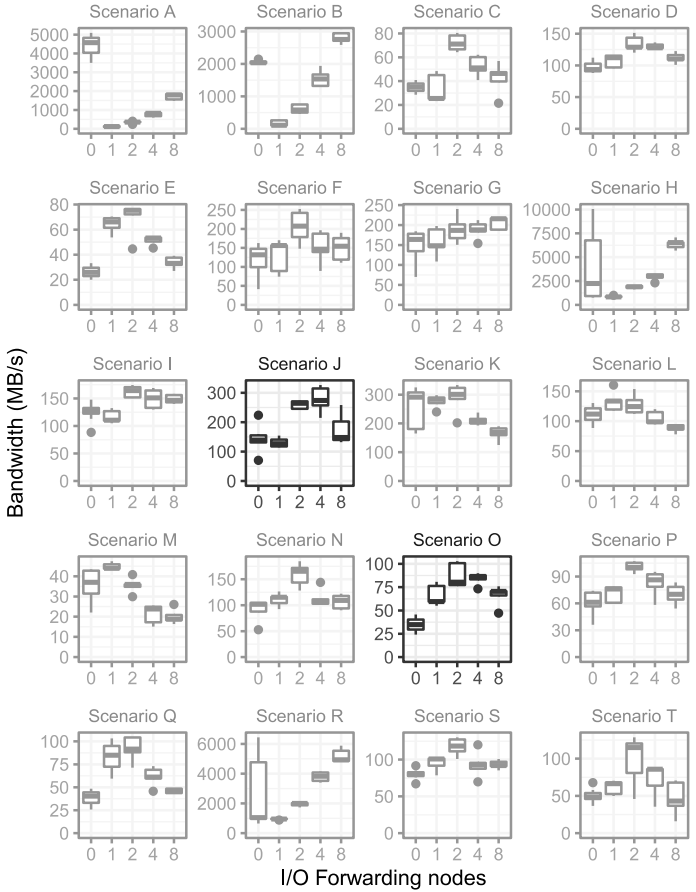
MareNostrum 4

- Bandwidth at client-side
- 5 repetitions for each
- Different days and periods



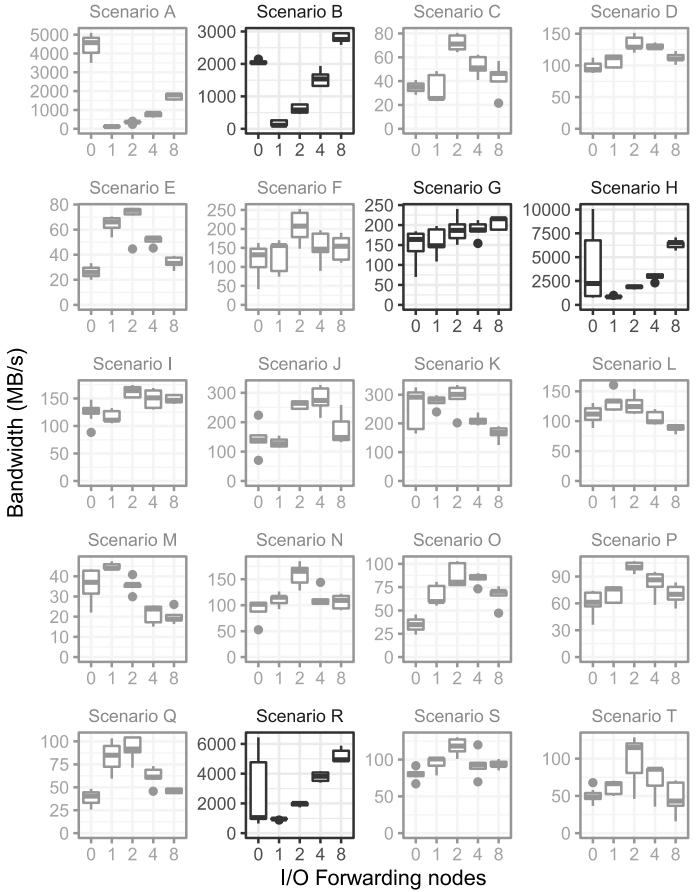
MareNostrum 4

- Bandwidth at client-side
- 5 repetitions for each
- Different days and periods



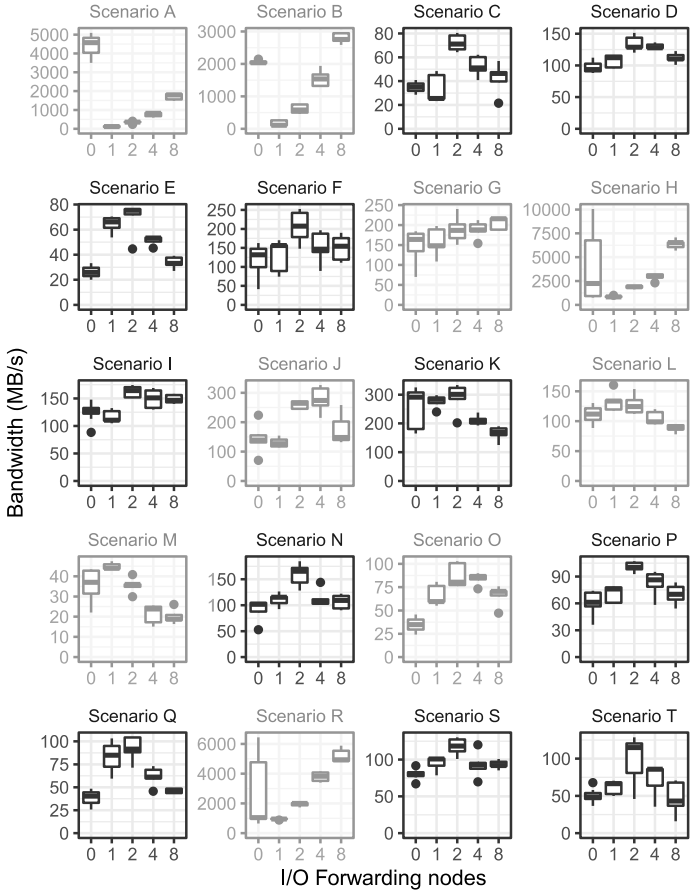
MareNostrum 4

- Bandwidth at client-side
- 5 repetitions for each
- Different days and periods



MareNostrum 4

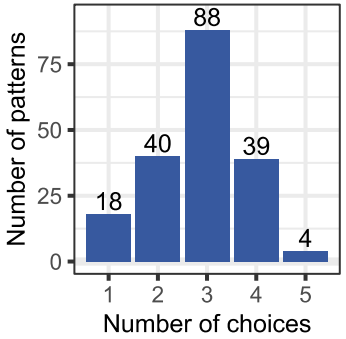
- Bandwidth at client-side
- 5 repetitions for each
- Different days and periods

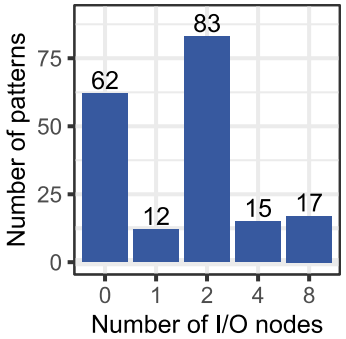


MareNostrum 4

- How many **choices** do we have to consider?
- Dunn's nonparametric test
- 3 choices impact performance

  46% patterns (88 out of 189)
- What is the **best number** of I/O nodes?
- **No simple rule to fit all**

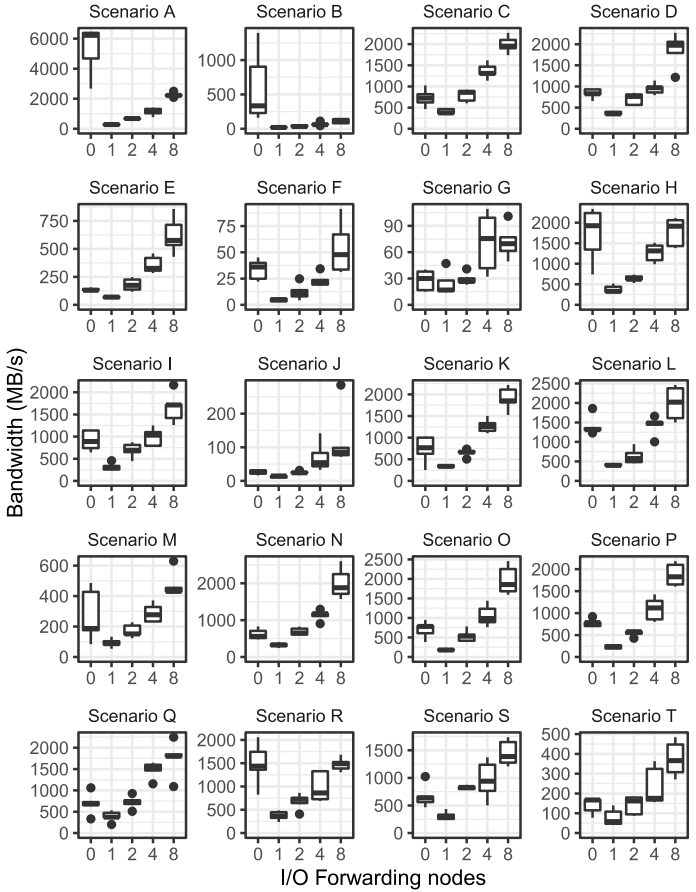




SC20
Everywhere we are | more than hpc.

Santos Dumont

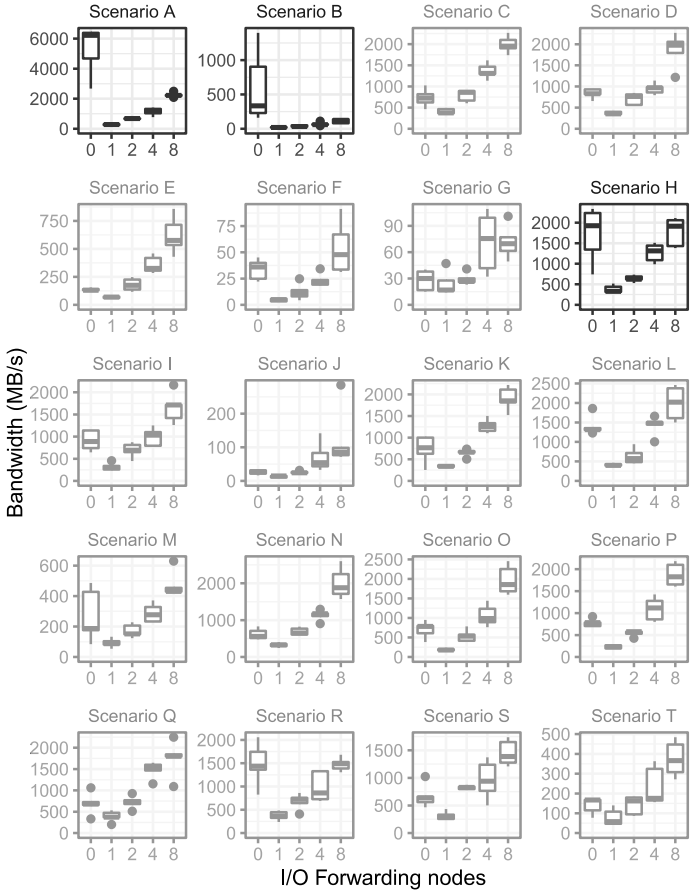
- Forwarding **impact is different**!
- The **more** I/O nodes, the **better**
- Not forwarding is an option



I/O FORWARDING

Santos Dumont

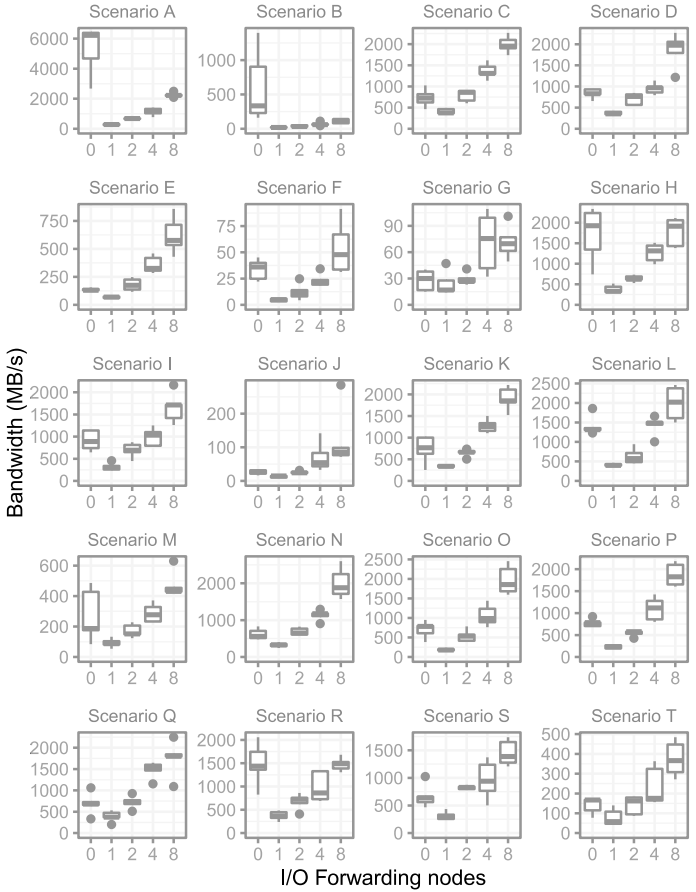
- Forwarding **impact is different**!
- The **more** I/O nodes, the **better**
- Not forwarding is an option



SC20
Everywhere we are | more than hpc.

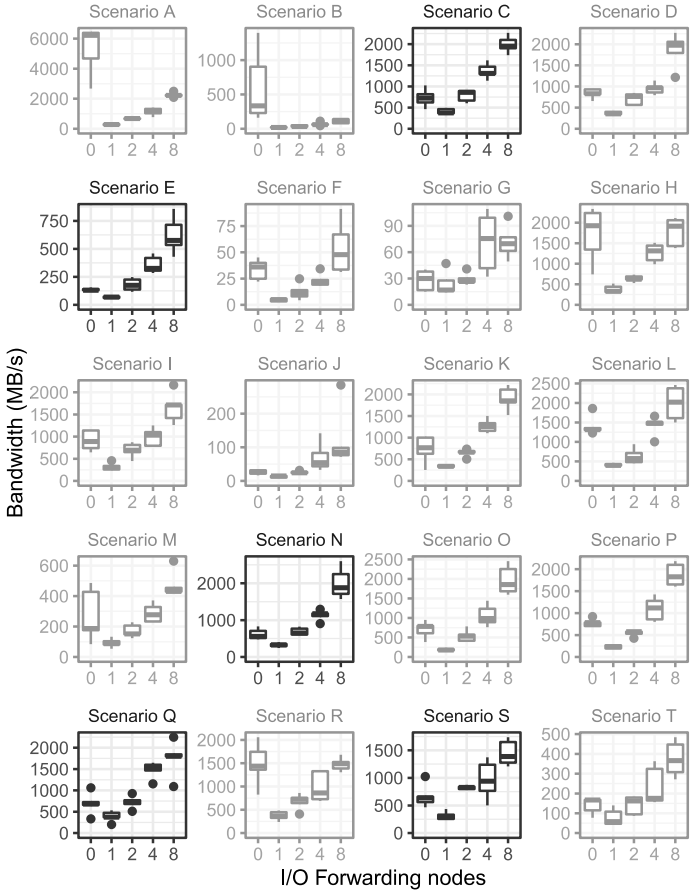
Santos Dumont

- Forwarding **impact is different**!
- The **more** I/O nodes, the **better**
- Not forwarding is an option



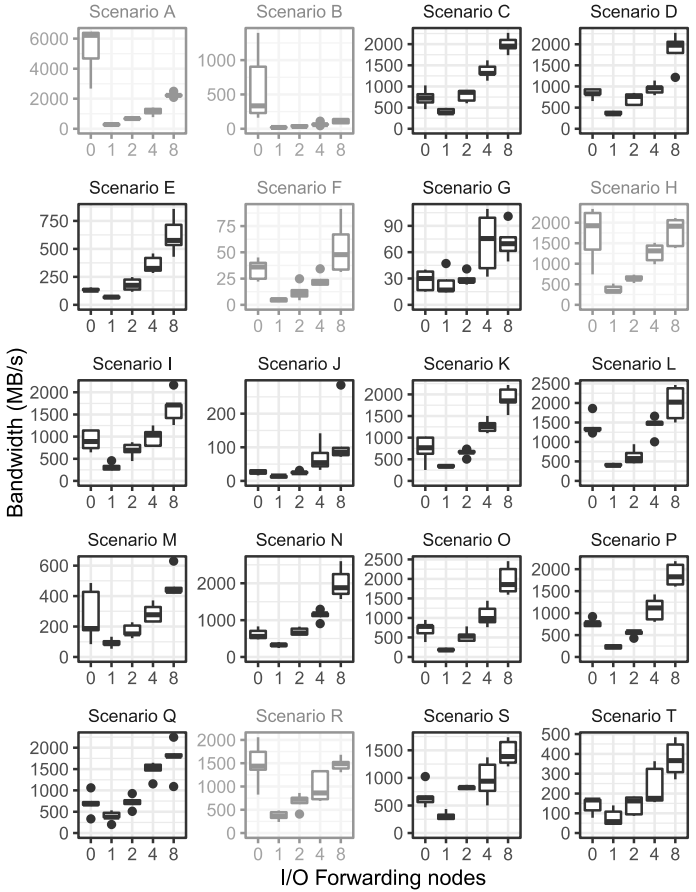
Santos Dumont

- Forwarding **impact is different**!
- The **more** I/O nodes, the **better**
- Not forwarding is an option



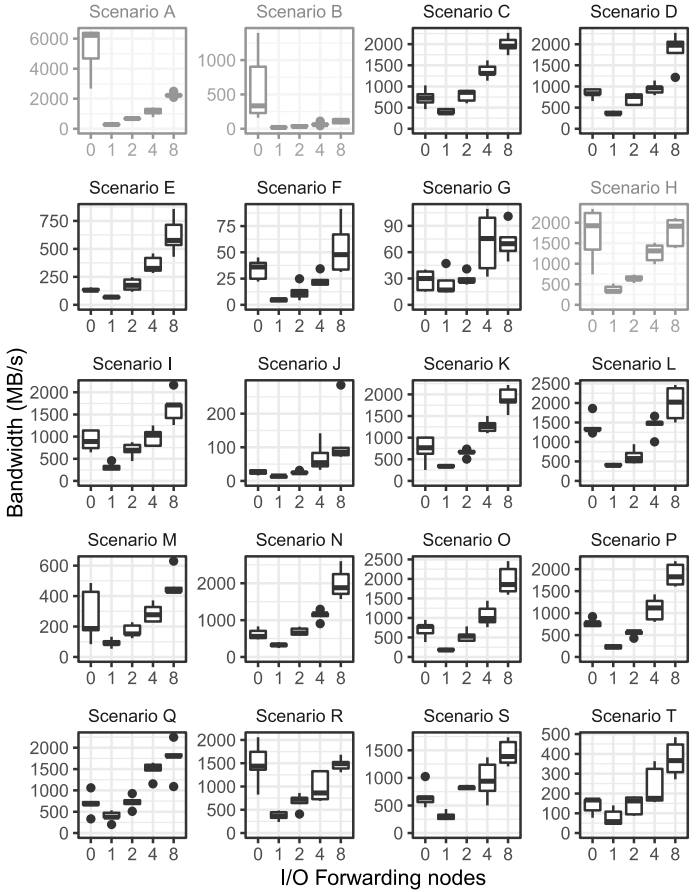
Santos Dumont

- Forwarding **impact is different**!
- The **more** I/O nodes, the **better**
- Not forwarding is an option



Santos Dumont

- Forwarding **impact is different**!
- The **more** I/O nodes, the **better**
- Not forwarding is an option



Discussion

- Increasing heterogeneous applications
- Shift from **must-use** to **on-demand** I/O forwarding layer
- Transparently **reshape** the flow of requests
- Towards a **dynamic allocation of I/O nodes**
- Idle or reserved set of compute nodes could act as I/O nodes
- Interference on I/O could not be reduced or eliminated

Conclusion

- I/O forwarding is an established and widely-adopted technique
- Not always possible to **explore its advantages** under different setups
- Impact or disrupt production systems
- **FORGE**: a lightweight **forwarding layer** in user-space
- Understand the impact of forwarding different **access patterns**
- Evaluation in **MareNostrum 4** and **Santos Dumont** supercomputers
- Shift from **must-use** to **on-demand** I/O forwarding layer

ACKNOWLEDGMENTS

This study was financed by the **Coordenação de Aperfeiçoamento de Pessoal de Nível Superior** - Brasil (CAPES) - Finance Code 001. It has also received support from the **Conselho Nacional de Desenvolvimento Científico e Tecnológico** (CNPq), Brazil; It is also partially supported by the **Spanish Ministry of Economy and Competitiveness** (MINECO) under grants PID2019-107255GB; and the **Generalitat de Catalunya** under contract 2014—SGR—1051. The author thankfully acknowledges the computer resources, technical expertise and assistance provided by the **Barcelona Supercomputing Center** - Centro Nacional de Supercomputación. The authors acknowledge the **National Laboratory for Scientific Computing** (LNCC/MCTI, Brazil) for providing HPC resources of the SDumont supercomputer, which have contributed to the research results reported within this paper. URL: http://sdumont.lncc.br.

SC20
Everywhere we are | more than hpc.

Towards On-Demand I/O Forwarding in HPC Platforms

Jean Luca Bez, Francieli Zanon Boito, Ramon Nou,
Alberto Miranda, Toni Cortes, and Philippe O. A. Navaux

jean.bez@inf.ufrgs.br

PDSW 2020 – International Parallel Data Systems Workshop

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Inria informatiques mathématiques

UPC

SC20
Everywhere we are | more than hpc.