WSPPD 2016 - XIV Workshop de Processamento Paralelo e Distribuído

Coordinating Data Access at I/O Forwarding Nodes

Jean Luca Bez, Francieli Zanon Boito, Lucas Mello Schnorr, Philippe O. A. Navaux
{jean.bez,schnorr,navaux}@inf.ufrgs.br, francieli.boito@posgrad.ufsc.br

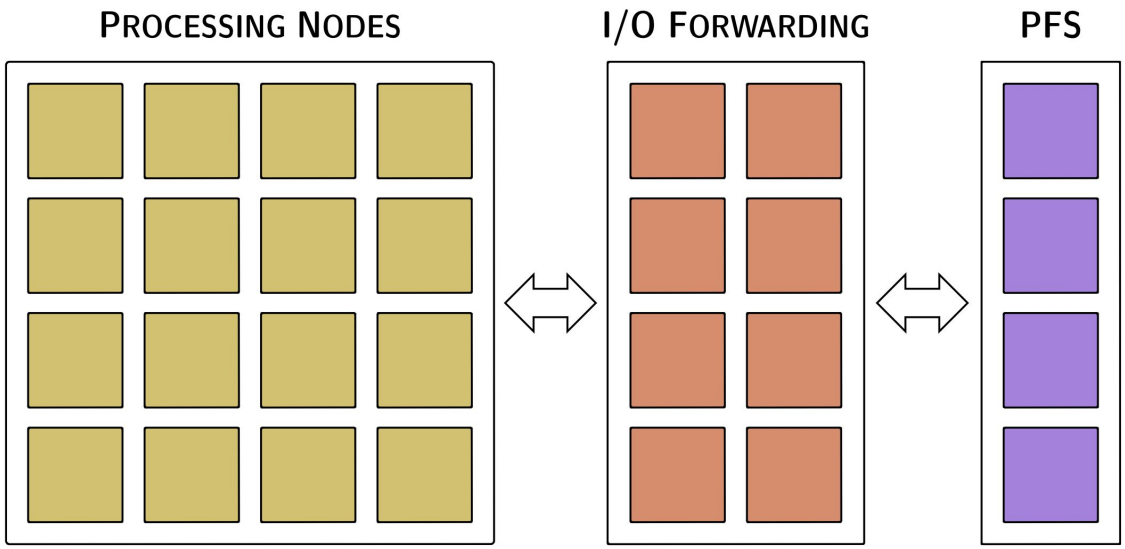
Summary

- Introduction
- Background and Related Work
  - The Forwarding Layer
  - AGIOS
- Coordinating Server Access with TWINS
- Experimental Results
- Conclusions

Introduction

- Scientific applications demand increasing performance
- Shared storage infrastructure over a dedicated set of nodes
- Parallel File System deployment
- I/O forwarding to reduce concurrency
- Read requests represents significant time (ARGONNE, 2015)

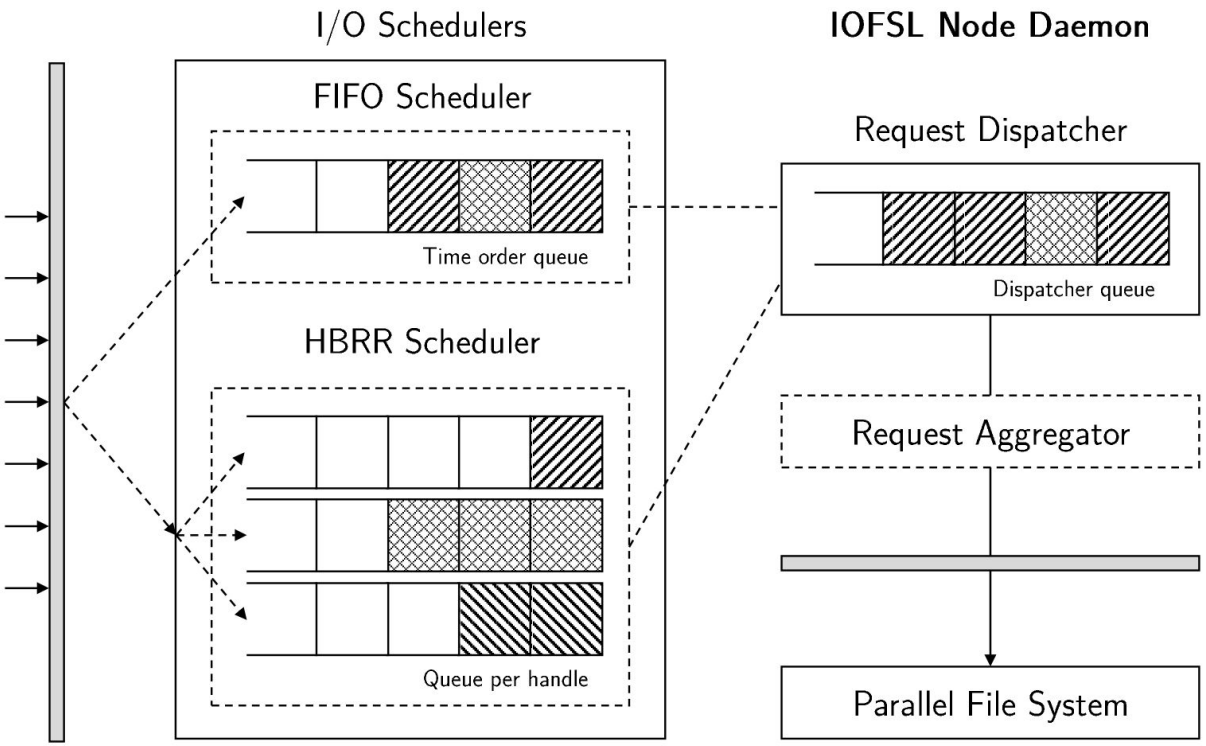
The Forwarding Layer



The Forwarding Layer

- IOFSL - *I/O Forwarding Scalability Layer*
- Open-souce forwarding framework
- Tested on large scale clusters and supercomputer
- Two schedulers
  - FIFO (First In, First Out) (OHTA, 2010)
  - HBRR (Handle-Based Round Robin) (OHTA, 2010)

The Forwarding Layer



AGIOS

- AGIOS - *Application Guided I/O Scheduler* (BOITO, 2013)
- Library to manage requests at file level (file offsets)
- Can be used at clients, I/O nodes, PFS data servers, PFS metada servers
- Tested on PFS servers
- API to implement new schedulers

Coordinating Server Access with TWINS

- *Time WINdows Scheduler*
- Coordinate I/O nodes accesses to the file system so that:
  1. an I/O node is focusing **all its accesses on one server**;
  2. **different I/O nodes** are focusing on **different servers**.
- One request queue per data server
- Iterates in a Round-Robin fashion
- Respecting the time window dedicated to each server
- Additional waiting times

Coordinating Server Access with TWINS

- Requires to know the location of the first stripe
- Information is collected when creating or opening a file
- Little overhead is expected
- Data layout is available in several PFS
- Additional translation step to focus the requests on different servers
- I/O node uses the N^{th} permutation of the data servers as a translation rule

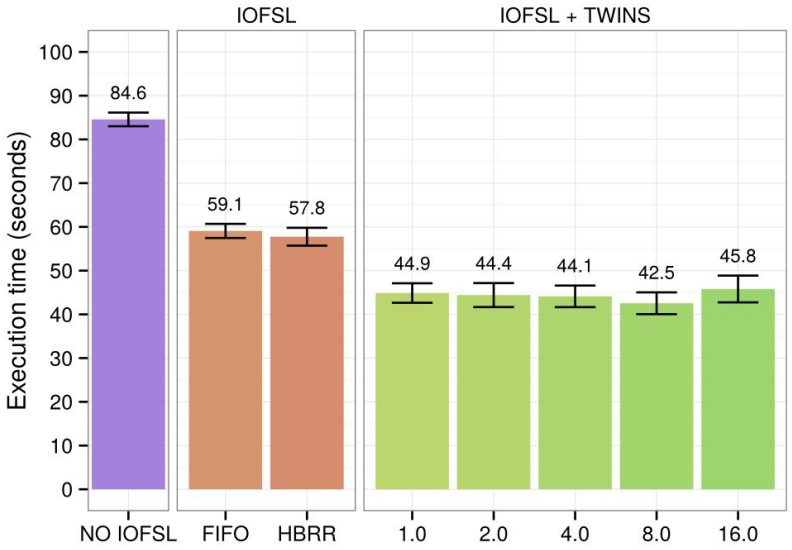
Experimental Testbed

- Nancy @ Grid5000
- Grimoire
  - 4 PVFS servers (metadata + data)
- Grisou
  - 4 IOFSL servers
  - 32 clients
  - Clients are equally distributed among I/O nodes
- Both clusters were completely reserved for the experiments

Experimental Setup

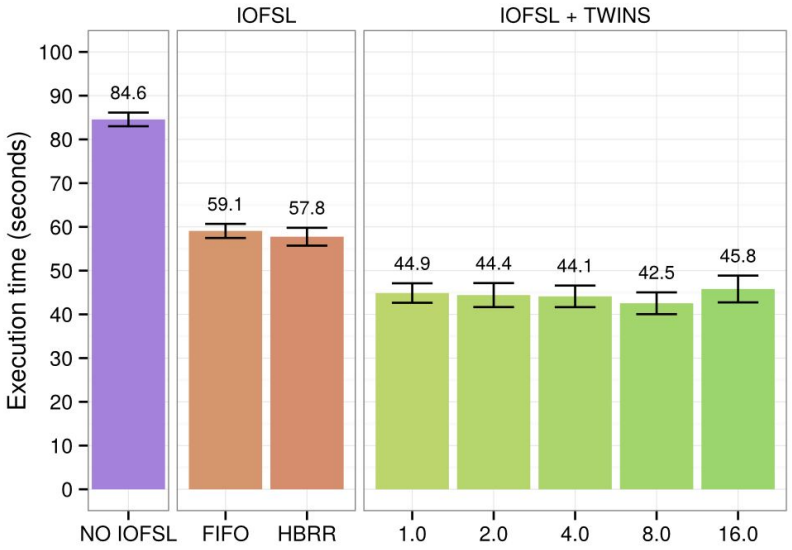
- MPI-IO Benchmark
  - 128 processes (4 per node)
  - 32MB per process (1024 requests of 32KB)
  - 4GB per experiment
- 1D-strided access pattern
- Makespan time (slowest process to complete)
- 8 repetitions in random order
- 99.7% confidence interval

Experimental Results



- Improvements of 30% over HBRR
- Improvements of 51% over not using IOFSL

Experimental Results



- Different window sizes
  - Small window does not hold requests for long to be aggregated
  - Large window implies in additional waiting times

Conclusions

- Requests reordering and aggregation are partially effective
- TWINS is able to coordinate accesses, reducing contention
- Window size of 8ms presented the best results
- Improvements over FIFO and HBRR
- Future work:
  - Additional benchmarks
  - Different access patterns
  - Automatic mecanism to tune the window size

Coordinating Data Access at I/O Forwarding Nodes

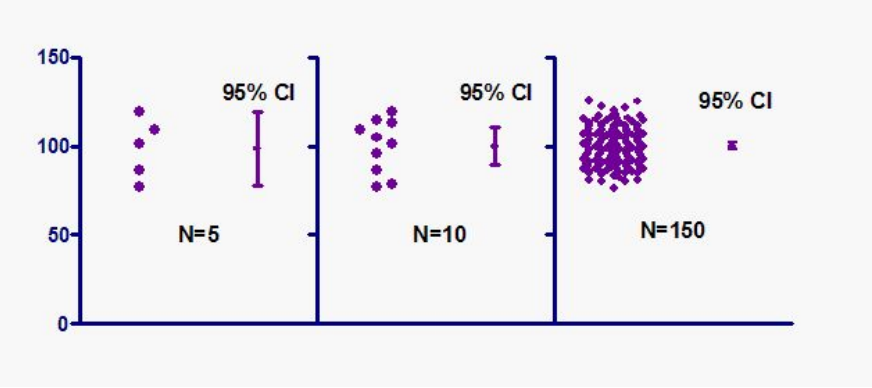
Thank you!

Jean Luca Bez, Francieli Zanon Boito, Lucas Mello Schnorr, Philippe O. A. Navaux
{jlbez,schnorr,navaux}@inf.ufrgs.br, francieli.boito@posgrad.ufsc.br

Confidence Interval

- A confidence interval **does not** quantify **variability**

"A 95% confidence interval is a range of values that you can be 95% certain contains the true mean of the population. This is not the same as a range that contains 95% of the values. The graph below emphasizes this distinction."



- It is correct to say that there is a 95% **chance** that the **confidence interval** you calculated **contains** the **true population mean**.