Application-aware arbitration of I/O resources in HPC machines

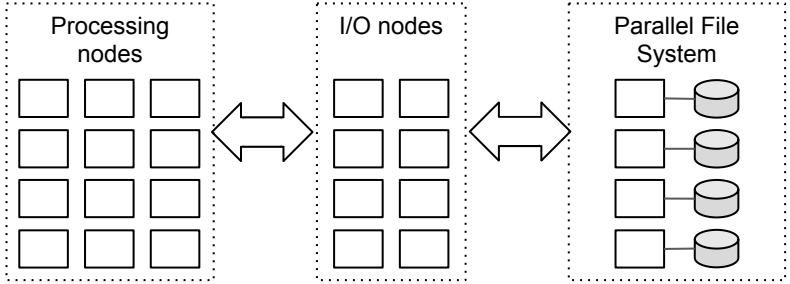
Francieli Zanon Boito

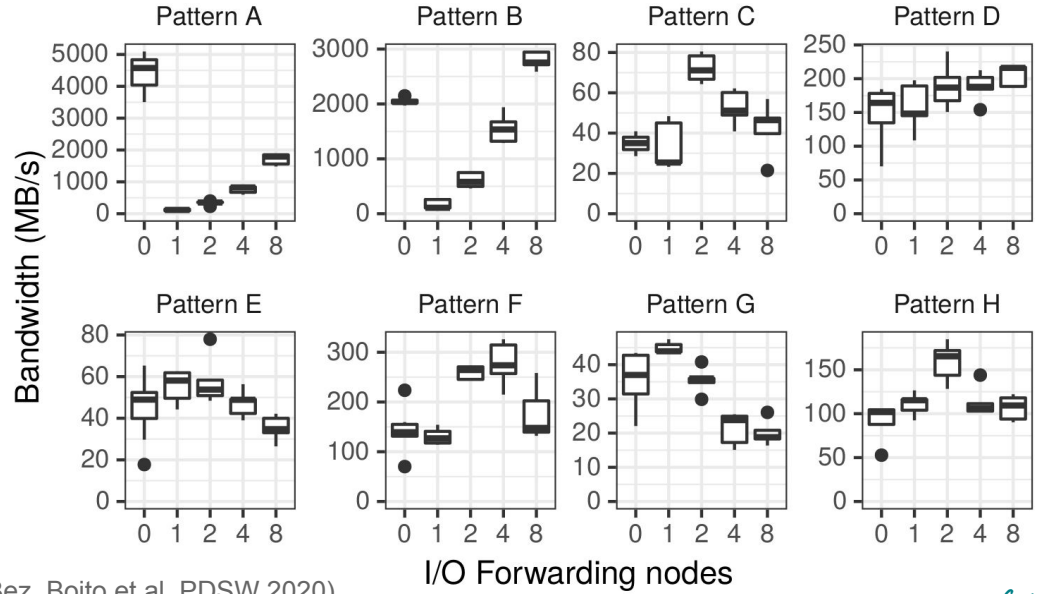
TADaaM seminar - March 2021

université de BORDEAUX LaBRI Inria

Motivation

- **HPC resources are arbitrated according to processing power**
  - exclusive access to cores/nodes, shared I/O infrastructure
- The number of I/O nodes is usually static
  - N compute nodes per I/O node, it depends on the placement
  - But it has a strong impact on performance





Graph from (Bez, Boito et al. PDSW 2020)

Can we do better?

The MCKP allocation policy

Results

} IPDPS 2021 paper

Estimation of application performance ← ongoing work

Future work

Arbitration policies for on-demand user-level I/O forwarding on HPC platforms

Jean Bez, Alberto Miranda, Ramon Nou, Francieli Zanon Boito,

Toni Cortes, Philippe Navaux

IPDPS 2021

https://hal.inria.fr/hal-03149582

The problem

- A set of applications (with known "performance curves")
- A number of I/O nodes (homogeneous)
- Goal: to maximize the **global bandwidth**
- Multiple-Choice Knapsack Problem (MCKP)
  - dynamic programming pseudo-polynomial solution in $O(W\sum N_i)$

$$\text{maximize} \sum_{i=1}^{k} \sum_{j \in N_i} p_{ij} x_{ij}$$

$$\text{subject to} \sum_{i=1}^{k} \sum_{j \in N_i} w_{ij} x_{ij} \leq W,$$

$$\sum_{j \in N_i} x_{ij} = 1, \forall i \in \{1, ..., k\}$$

$$x_{ij} \in \{0, 1\}, \forall i \in \{1, ..., k\}, \forall j \in N_i.$$

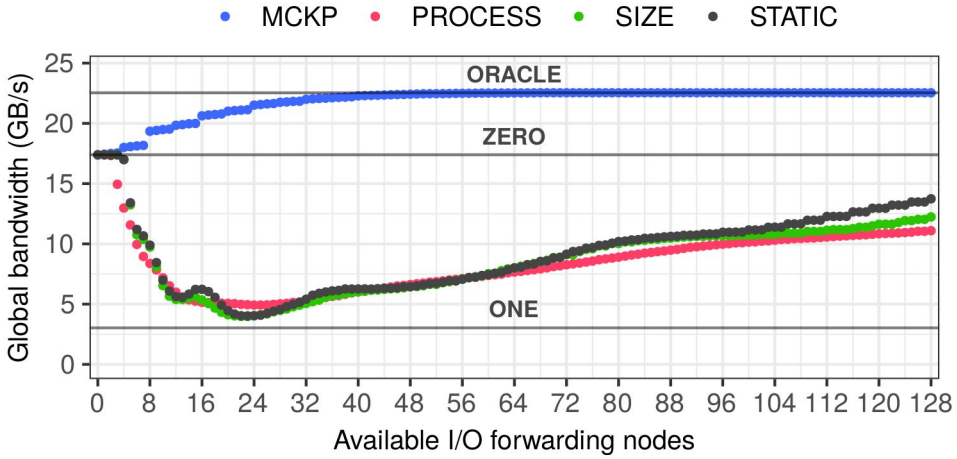
The MCKP allocation policy

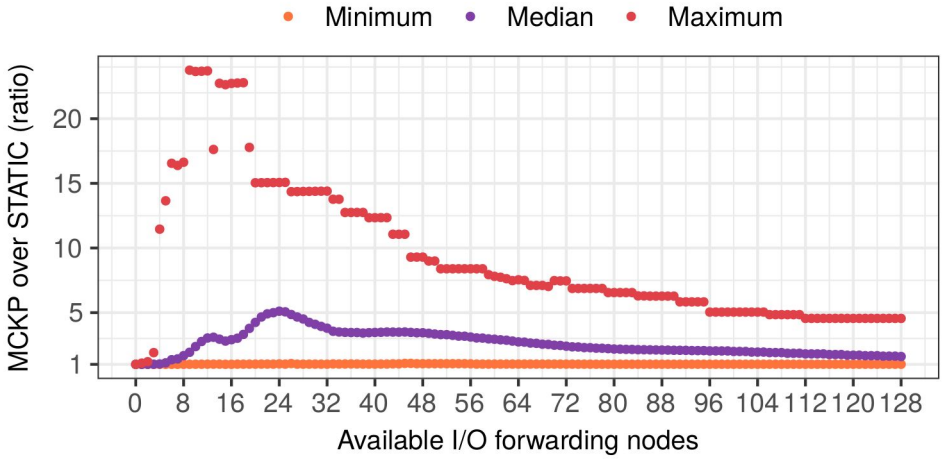
- Each application is a class
  - the options in the class are the numbers of I/O nodes
  - ION in [1,CN] if CN % ION == 0 (load balancing)
  - it is a good idea to decrease the number of options
- Unknown application curve: give it the static number
- If number of I/O nodes >= number of applications, optimal solution
- if we must share, add this option and arbitrate N-1 I/O nodes
  - **we avoid sharing as much as possible**
  - shared option: bandwidth with 1 I/O node / number of applications (pessimistic)
- **Dynamic:** every time we change the set of running applications

Static evaluation I

Benchmarks (with FORGE https://github.com/jeanbez/forge/) on MareNostrum4 (@BSC)

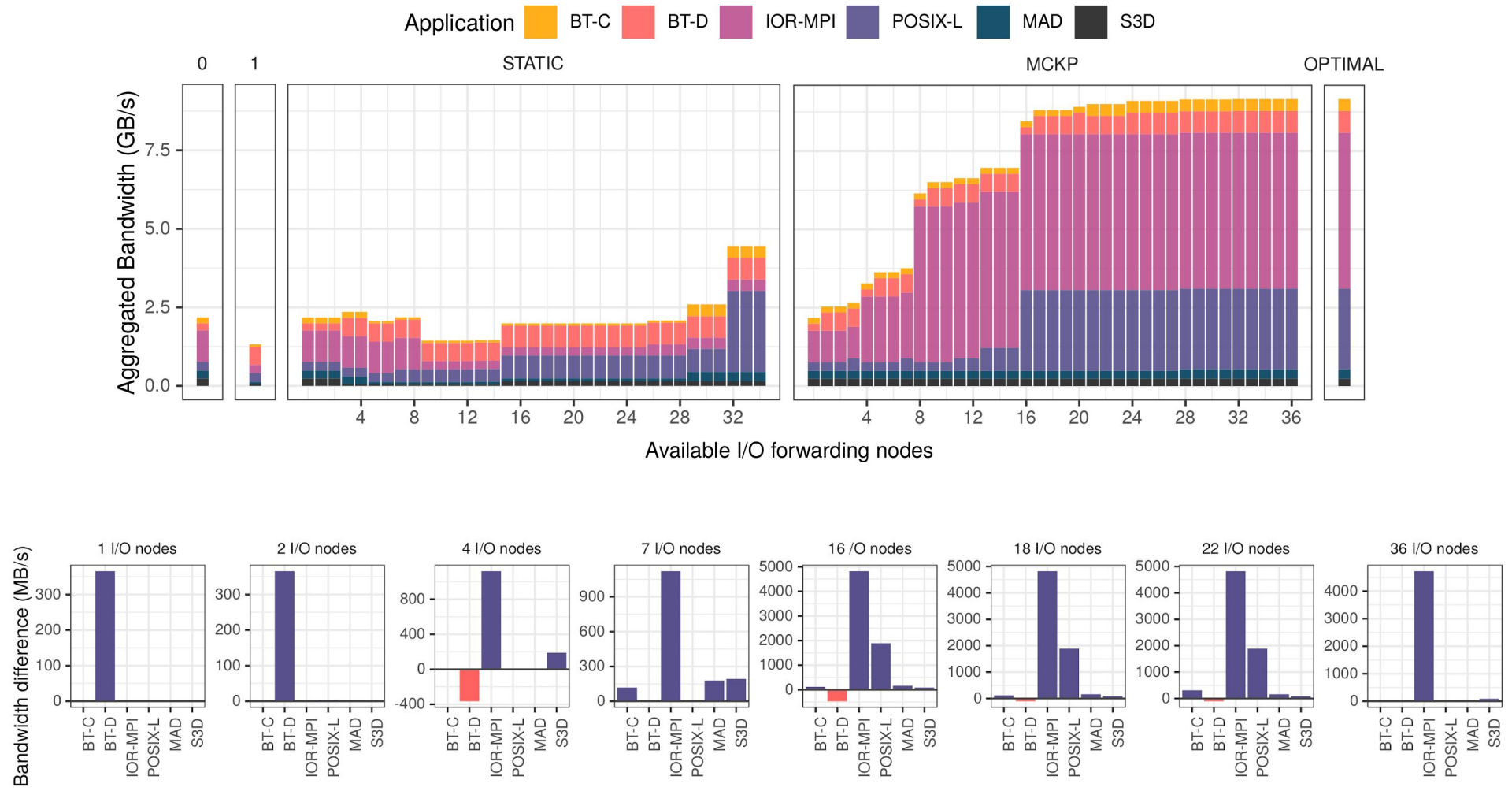
189 access patterns -> 10k random sets of 16 benchmarks (simulation)





Static evaluation II

- GekkoFWD - a forwarding mode to the GekkoFS temporary file system
  - https://github.com/bsc-ssrg/gekkofs
  - integrated with the AGIOS I/O request scheduling library https://github.com/francielizanon/agios
- Grid'5000 - 6 applications on 72 compute nodes


Application
BT-C
BT-D
IOR-MPI
POSIX-L
MAD
S3D
0
1
STATIC
MCKP
OPTIMAL
Aggregated Bandwidth (GB/s)
0.0
2.5
5.0
7.5
4
8
12
16
20
24
28
32
36
Available I/O forwarding nodes
Bandwidth difference (MB/s)
1 I/O nodes
2 I/O nodes
4 I/O nodes
7 I/O nodes
16 /O nodes
18 I/O nodes
22 I/O nodes
36 I/O nodes
BT-C
BT-D
IOR-MPI
POSIX-L
MAD
S3D

Final remarks

- Dynamic evaluation on the paper
  - MCKP was **up to 85% better than STATIC**
- 2.7s to arbitrate 256 I/O nodes to 512 applications
  - Asynchronous notification of all compute nodes
- Application-aware arbitration of I/O nodes is a good idea

Estimation of the impact of I/O forwarding on application performance

Francieli Zanon Boito

ongoing work

Preliminary research report at https://hal.inria.fr/hal-02969780

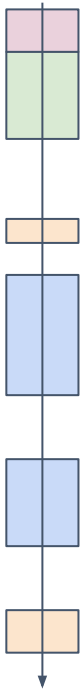
Motivation

- MCKP needs to know the "application curves"
  - we can profile the most important applications
  - we can learn them over time
  - the technique will still work with partial information
- **What if we want to avoid profiling all applications?**
  - approximate application performance by benchmarks

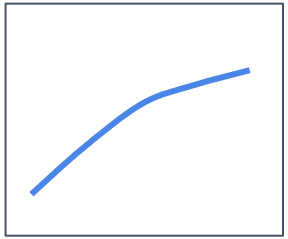
Darshan trace

estimate

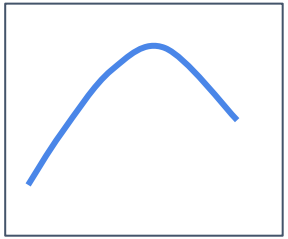
Application I/O phases



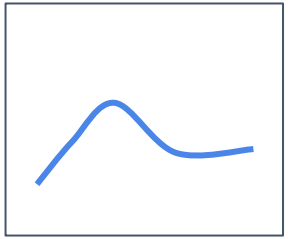
Pattern 1



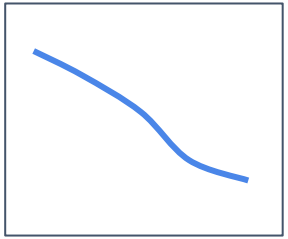
Pattern 2



Pattern 3



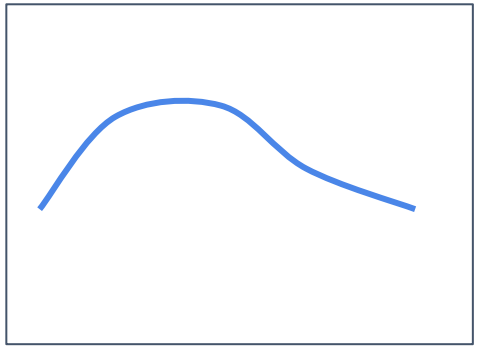
Pattern 4



+

=

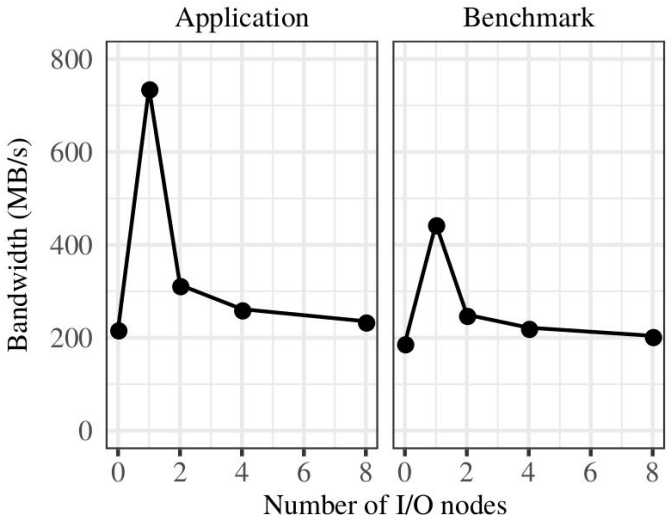
Application I/O performance



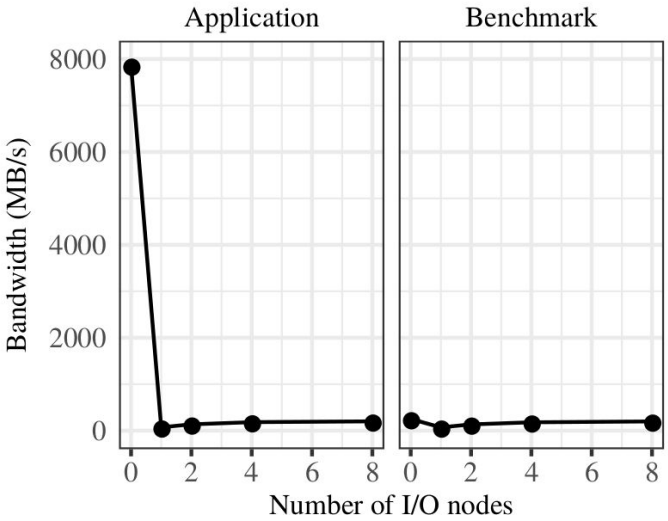
Number of I/O nodes

Periodic applications

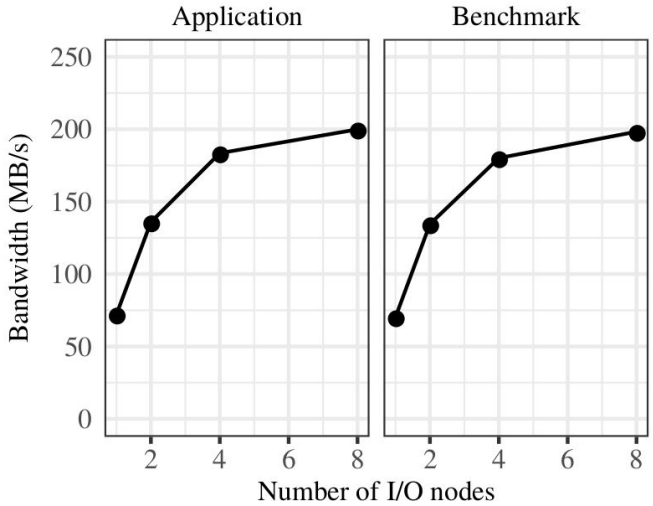
Median error of up to 20%



Periodic-A



Periodic-B



Periodic-B with ION > 0

Final remarks

- Other case studies in the research report
- **Ongoing work**
- Preliminary conclusion: the idea has some success
  - We must be ready to deal with error
  - Limitation of using an aggregated trace
- Next step: approximate benchmark parameters

Future work perspectives

a.k.a. "call for collaboration"

Application-aware arbitration of I/O resources in HPC machines

Francieli Zanon Boito

TADaaM seminar - March 2021

université de BORDEAUX LaBRI Inria