Inria

ADMIRE

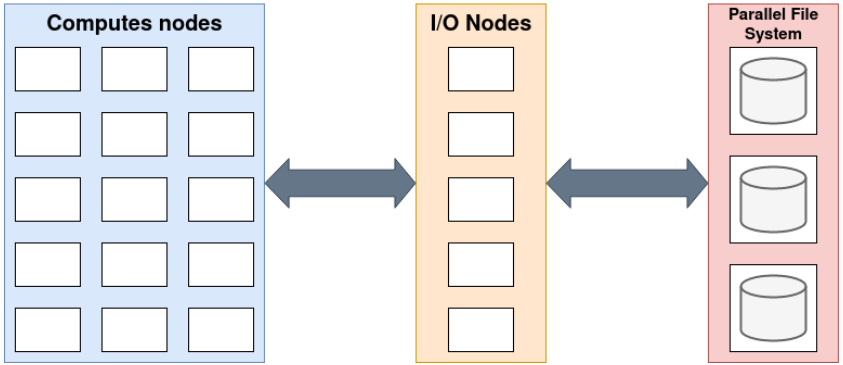
malleable data solutions for HPC

ADAPTIVE MULTI-TIER INTELLIGENT DATA MANAGER FOR EXASCALE

# I/O nodes sharing between applications

Alexis Bandet, Francieli Boito & Guillaume Pallez
Inria Bordeaux Sud-Ouest, France

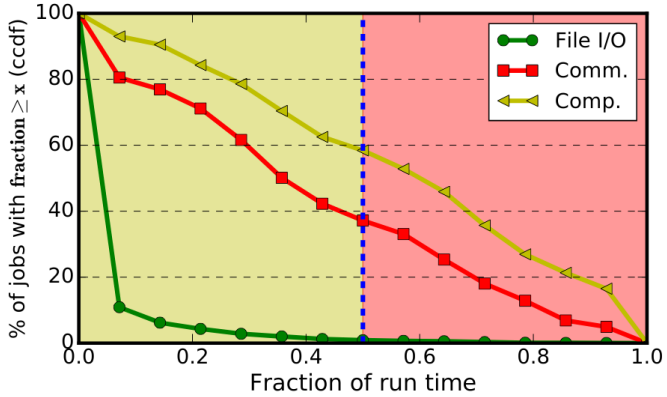
Per3s workshop – 13th june 2022
IMT Saclay




I/O forwarding layer

- Intermediate layer between File System and Computation machines
- #I/O Nodes based on #Computes node (static method [1])
- MCKP scheduling for optimal bandwidth optimization [2]
- Exclusive allocation

Source : Liu, Z et al.[3]




Complementary cumulative distribution functions of time spent on file I/O, MPI communication and computation, expressed as a fraction of the total runtime[3]

- HPC applications have relatively limited I/O time
- 95% of application spent less than 20% of their time doing I/O
- Exclusive policy may lead to waste of ressources

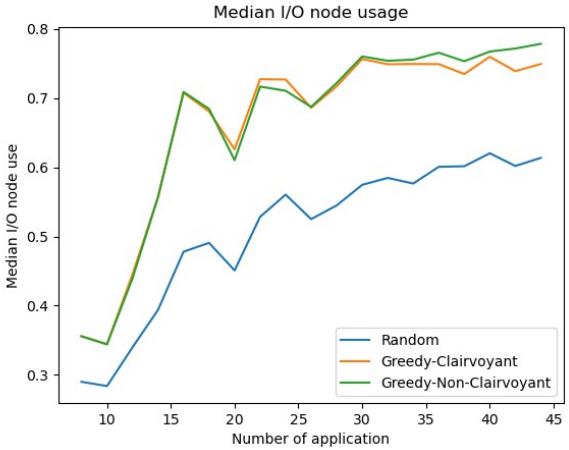
# How to share I/O nodes between applications ?

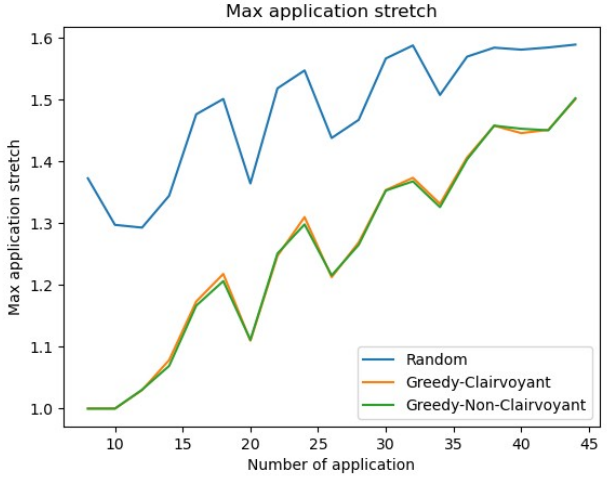
Model :

- Set of K applications $\{ A_1,...,A_k \}$ , each defined by a number of I/O nodes $n_i$ and a ratio of time spend doing I/O operations $r_i$
- N I/O nodes

Two algorithms :

- **Greedy-Clairvoyant** tries to balance the load (from the $r_i$) across the I/O nodes.
- **Greedy-Non-Clairvoyant** tries to balance the number of applications across the I/O nodes.




- Use represents fraction of time I/O nodes spent in I/O state
- Shows sharing efficiency at machine level
- No benefit from I/O ratio information for scheduling

- Stretch acknowledge sharing penalty at user level
- Once again no benefit with I/O ratio data

## Conclusion :

- It is possible to share I/O nodes in an efficient way
- This maximize ressource utilization
- Limited impact on application

## Future work

- Refining application model
- Compare to states of the art forwarding scheduling technics
- Evaluate the combination of placement policies with heuristics for the selection of the number of I/O nodes
- Test the sharing efficiency outside simulation

# Thanks for your attention !

**Références** :
[1] Xu, W(.2014). Hybrid hierarchy storage system in MilkyWay-2 supercomputer. Frontiers of Computer Science, 8(3), 367-377.
[2] Bez, J. L.,(2021, May). Arbitration Policies for On-Demand User-Level I/O Forwarding on HPC Platforms. In 2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS) (pp. 577-586). IEEE.
[3] Liu, Z.(2020, June). Characterization and identification of HPC applications at leadership computing facility. In Proceedings of the 34th ACM International Conference on Supercomputing (pp. 1-12).