Evaluation of scheduling algorithm using realistic simulation

Adrien Faure[1,2], Millian Poquet[1], Olivier Richard[1]

DATAMOVE Team, LIG

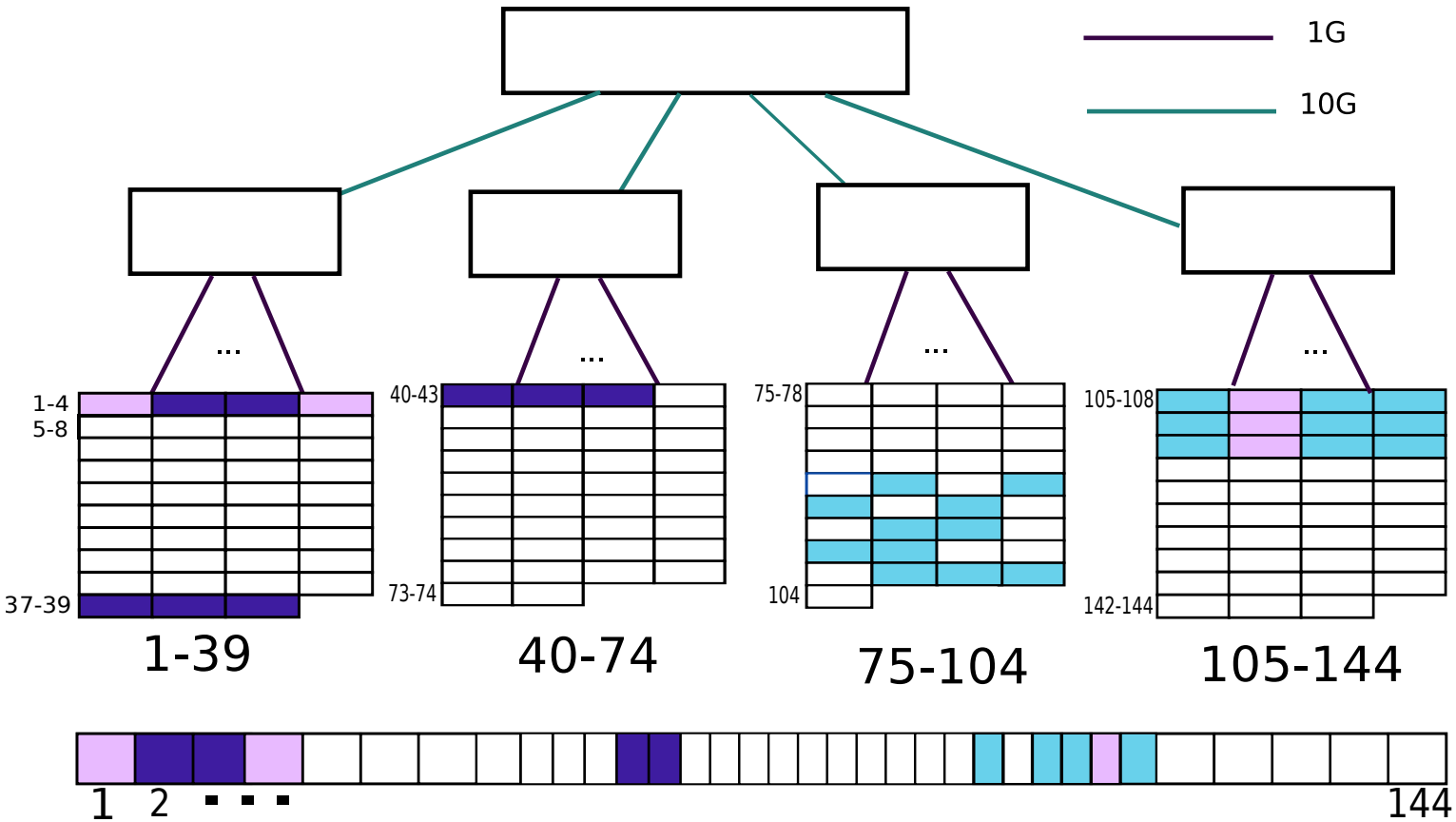
1

1

2

Compas Juillet 2018 Toulouse

Bad placement degrades application performances



heterogeneous platform (nodes, network)

HPC cluster management

Resources and
Jobs Management Systems (RJMS)

- AKA batch scheduler
- Orchestrates
  resources on HPC clusters
  - Implements scheduling policies
  - Manages parallel jobs
- Examples:
  Slurm, OAR, TORQUE, PBS...

RJMS Facts

- Large scale: from 100 to 100 000 nodes
- Heterogeneous nodes with gpgpu, nvram

Objectives

Questions

How study and improve the scheduler on HPC systems?

We need to experiment on the RJMS but...

Production systems are not available for testing RJMS

- They are already full of users jobs!
- Energy/time cost of experiments is not affordable

State of the art

DIY

- most papers
- publish and perish?

Long-term

- Examples: Alea, Batsim, AccaSim
- Maintained?

Challenges

- Assessed against reality?
- Intra/Inter job interferences?

State of the art

DIY

- most papers
- publish and perish?

Long-term

- Examples: Alea, **Batsim**, AccaSim
- Maintained?

Challenges

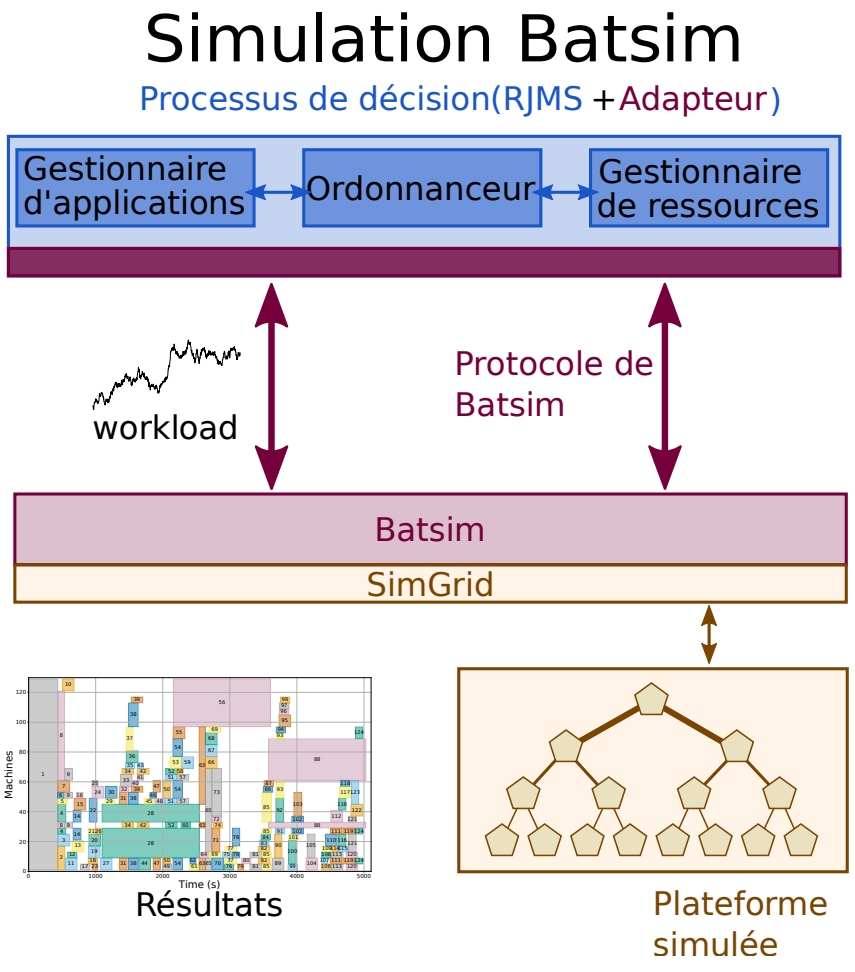
- Assessed against reality?
- Intra/Inter job interferences?

Outline

Batsim Overview

Infrastructure simulator: Study scheduling algorithms

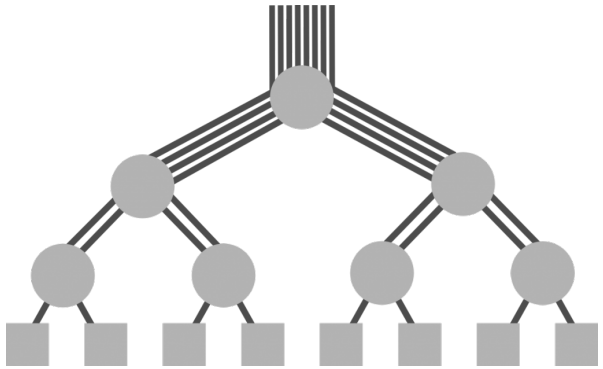
- Based on SimGrid
  - Reliable: 15+ years, strong community
  - Topology-aware validated network models
- Modular
- $\simeq$ 9k C++ LOC
- Packaged with Nix



Batsim inputs

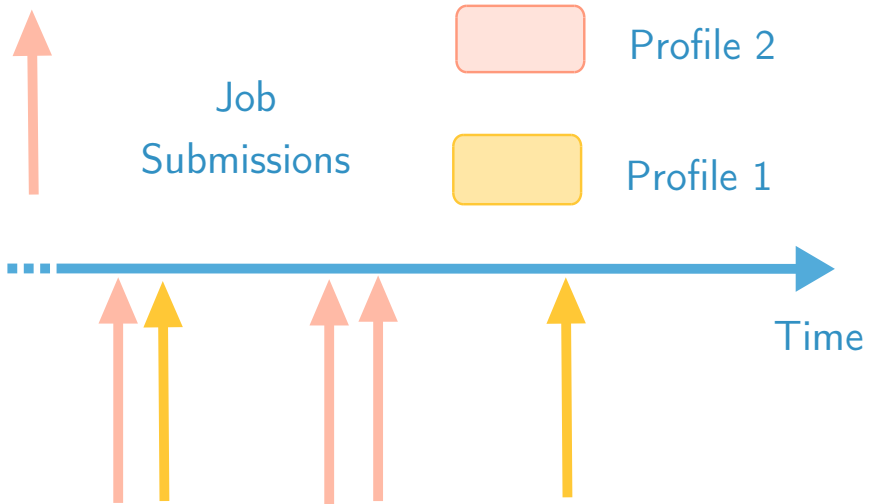
What is a Batsim platform?

- Batsim platform $\simeq$ SimGrid platform



What is a Batsim workload?

- List of jobs
  - Submit time
  - Walltime (user-given maximum run time)
  - Required resources
- Each job is associated to a profile



Job Profile types

Delay
- Fixed amount of time

MSG
- A computation vector (1D matrix)
- A communication 2D matrix

Sequence
- A sequence of profiles
- Repeated *n* times
- à la BSP[1]

[1] Bulk Synchronous Parallel model

Job Profile types

Delay
- Fixed amount of time

MSG
- A computation vector (1D matrix)
- A communication 2D matrix

Sequence
- A sequence of profiles
- Repeated *n* times
- à la BSP[1]

SMPI
- Replay of time-independent MPI traces

[1] Bulk Synchronous Parallel model

Experimentation Design

Algorithms

- Very Simple Scheduling Algorithm
- Different Allocation Policies
  - Contiguous allocation
  - Not Contiguous allocation

Workload

- Generated workload
- 512 jobs
- 8, 16, 32 nodes

Profiles

- We use Time-Independant SMPI Traces
- NAS Parallel Benchmarks

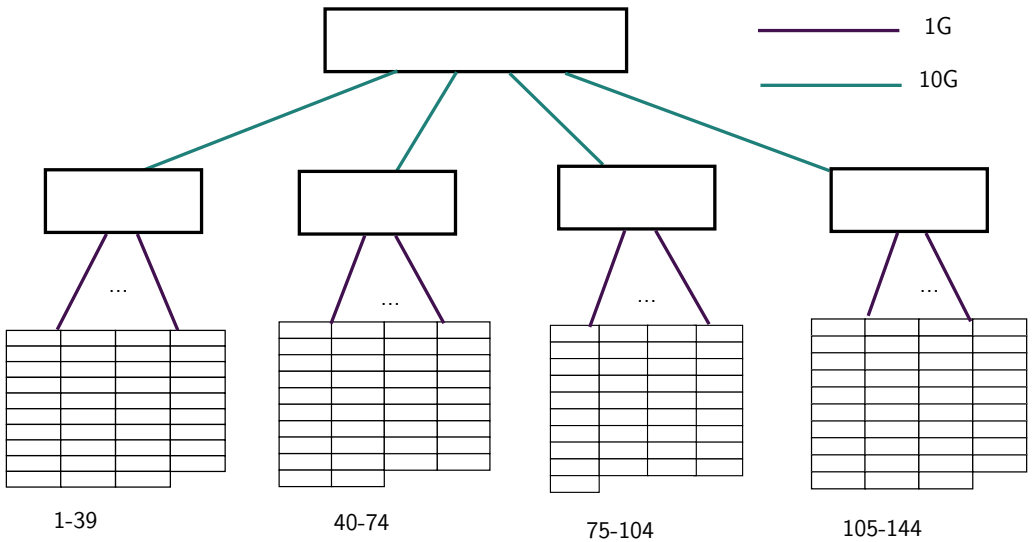
Platform Modeling

Graphene

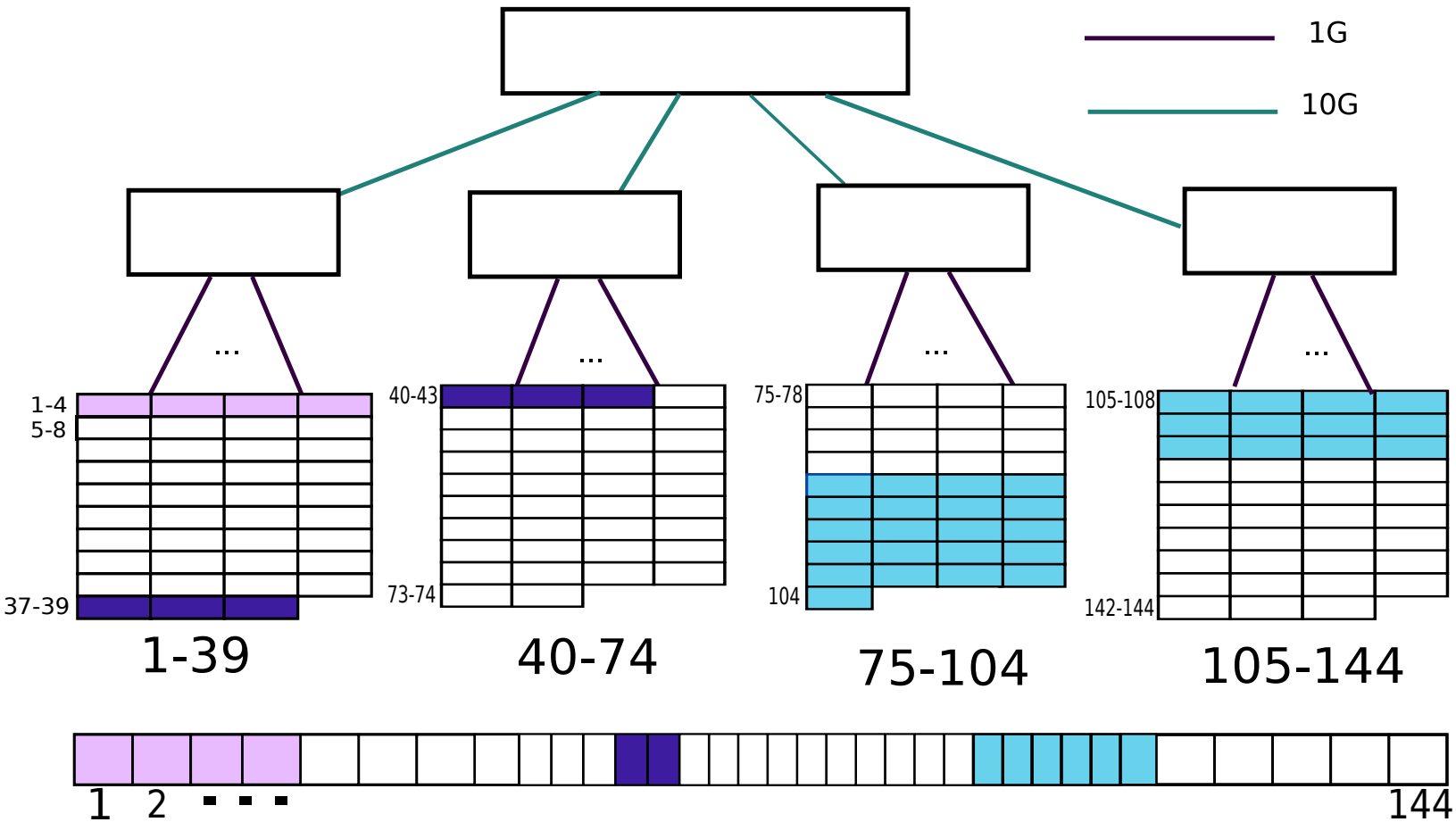
- Grid 5000 at Nancy
- 144 nodes
- 4 irregular cabinets
- tcp network

Contention points

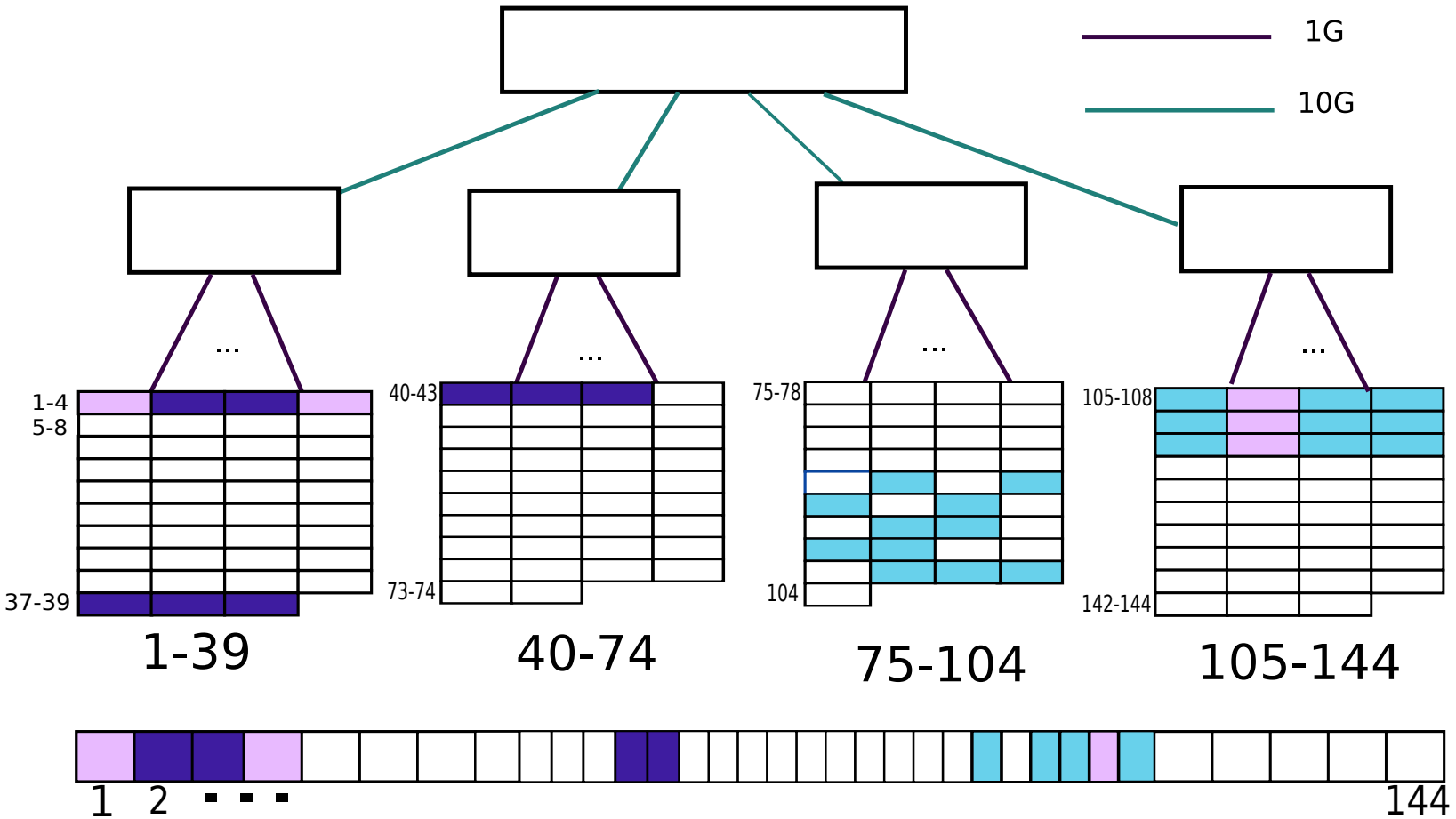
- At nodes level
- Inside a cabinet
- Between cabinets



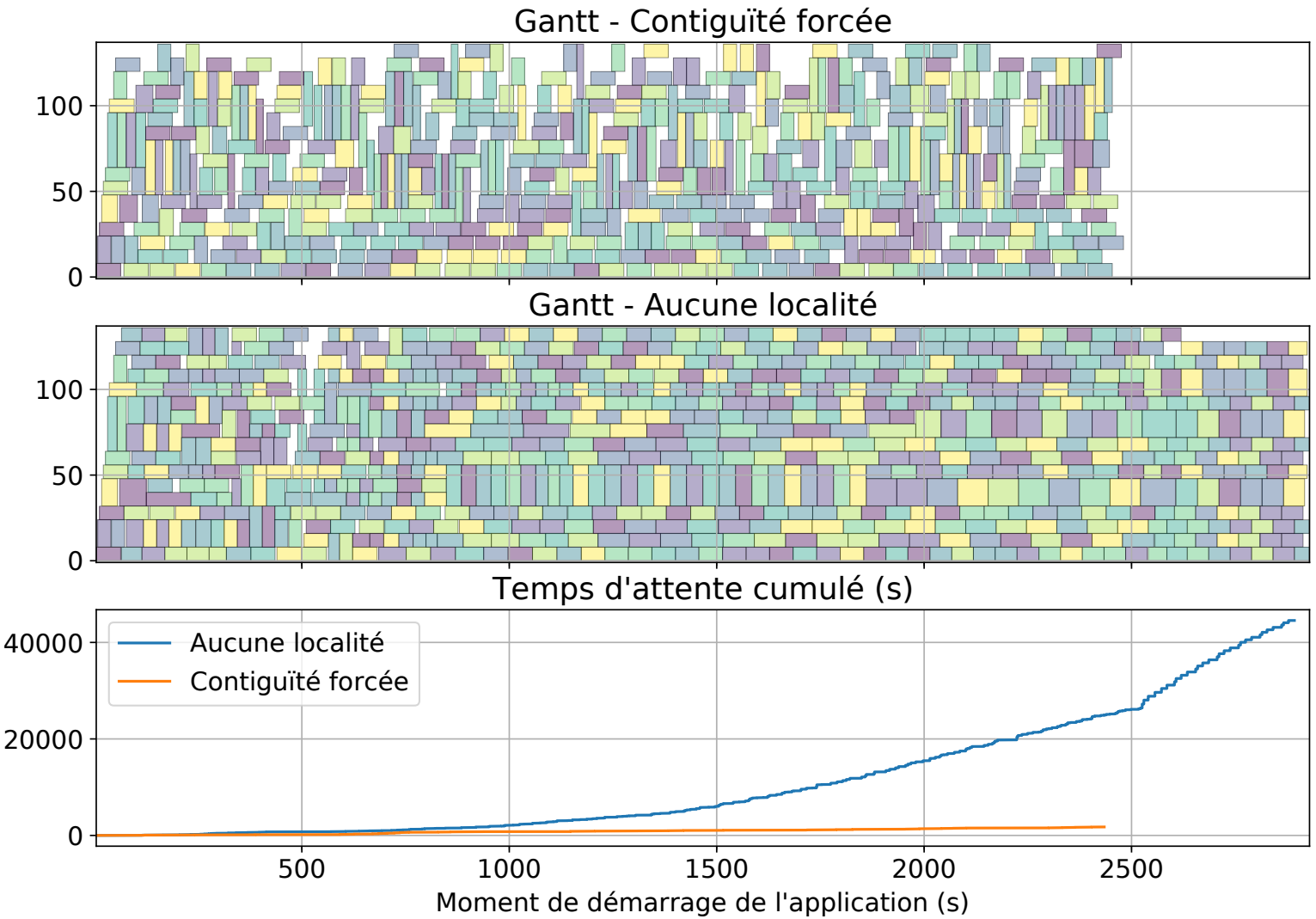
Contiguous Allocation Policy



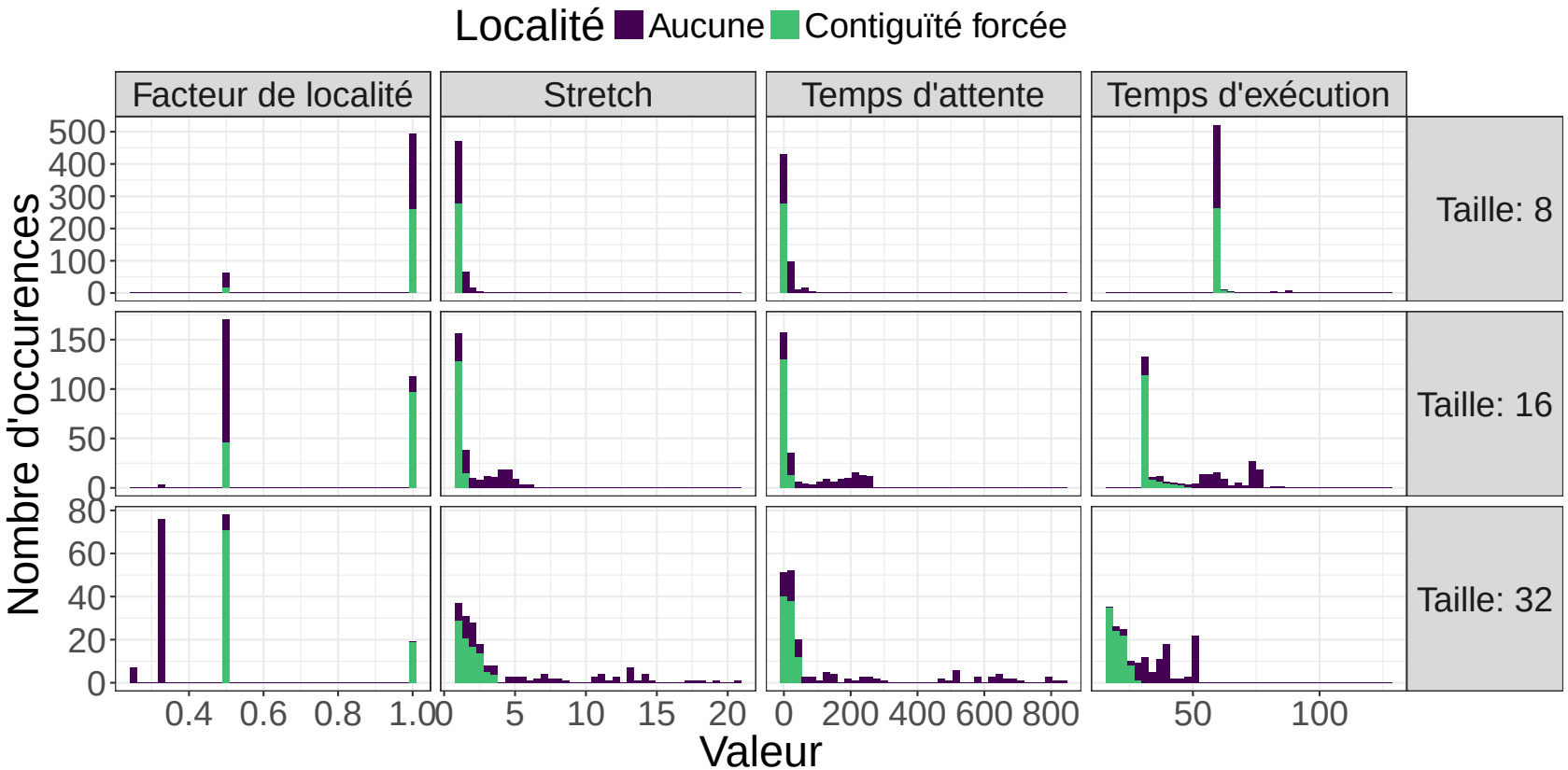
Not Contiguous Allocation Policy



Gantt


Gantt - Contiguïté forcée
100
50
0
Gantt - Aucune localité
100
50
0
Temps d'attente cumulé (s)
40000
20000
0
Aucune localité
Contiguïté forcée
500
1000
1500
2000
2500
Moment de démarrage de l'application (s)

Metrics



Locality Factor $\frac{MinimalNumberOfSwitch_i}{NumberOfSwitch_i}$

Conclusion

- SMPI for realistic simulation

Future Works

- Validation of simulation for Batch Scheduler
- Applications behavior
  - Can they be regrouped in category
  - Detect phases (computation, communication, I/O)

Thanks!

Batsim:
https://github.com/oar-team/batsim

Contacts

- Email: adrien.faure@inria.fr
- Mattermost: https://framateam.org/batsim

References:

- Dalibor Klusáček, Hana Rudová. **Alea 2 - Job Scheduling Simulator.** In proceedings of the 3rd International ICST Conference on Simulation Tools and Techniques (SIMUTools 2010), ICST, 2010.
- Jose A. Pascual, Jose Miguel-Alonso, Jose A. Lozano. **Locality-aware policies to improve job scheduling on 3D tori.** The Journal of Supercomputing, 2015, vol. 71, no 3, p. 966-994.

Acknowledgments

I'd like to thanks to Michael Mercier that gladly let me use his slides from https://github.com/oar-team/batsim/blob/master/publications/Batsim_JSSPP_2016.pdf.