Modeling Large Scale Systems and Validating their Simulators

Arnaud Legrand (MESCAL)
Martin Quinson (ALGORILLE/VERIDIS)

Hemera evaluation, December 17, 2014

Simulation of Parallel/Distributed Systems

Network Protocols: Standards emerged: GTNetS, DaSSF, OmNet++, NS3

- Grid Computing — OptorSim, ChicagoSim, GridSim, JFreeSim, ...
- Peer-to-peer — P2Psim, SimP2P, PeerSim, OverSim, ...
- Volunteer Computing — SimBA, EmBOINC, SimBOINC, ...
- HPC/MPI — Dimemas, PSinS, BigSim, LogGoPSim, XSim, SST, ...
- Cloud Computing — CloudSim, GroudSim, iCanCloud, GreenCloud, ...

This raises severe **methodological/reproducibility** issues:

- Short-lived, badly supported (software QA), sparse validity assessment

Simulation of Parallel/Distributed Systems

Network Protocols: Standards emerged: GTNetS, DaSSF, OmNet++, NS3

- Grid Computing — OptorSim, ChicagoSim, GridSim, JFreeSim, ...
- Peer-to-peer — P2Psim, SimP2P, PeerSim, OverSim, ...
- Volunteer Computing — SimBA, EmBOINC, SimBOINC, ...
- HPC/MPI — Dimemas, PSinS, BigSim, LogGoPSim, XSim, SST, ...
- Cloud Computing — CloudSim, GroudSim, iCanCloud, GreenCloud, ...

This raises severe **methodological/reproducibility** issues:

- Short-lived, badly supported (software QA), sparse validity assessment

SimGrid: a 15 years old joint project

- Versatile: Grid, P2P, Clouds, HPC, Volunteer
- Collaborative: (CNRS, Univ., Inria) Open Source:, active community
- Widely used: 150 publications by 120 individuals, 30 contributors

http://simgrid.gforge.inria.fr

SimGrid Key Features: Fluid Network Model

- Packet level models: Full net stack. Inherently slow, hard to instantiate
- Simple models: Delay-based, distribution, coordinates
  Very scalable, but no topology, *no network congestion*

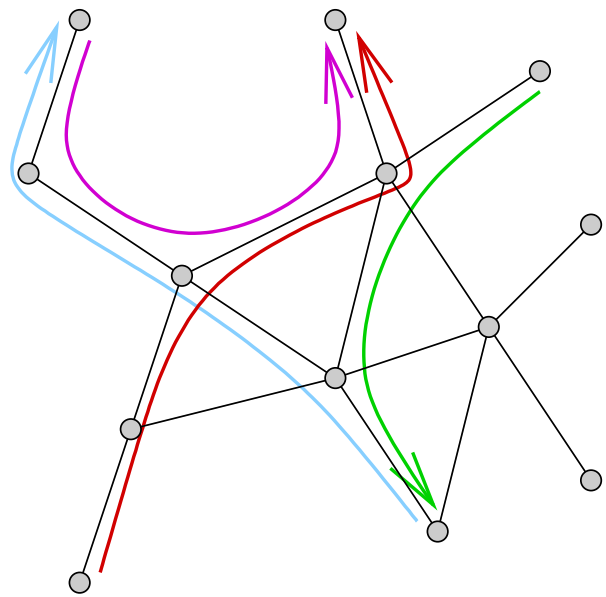
SimGrid Key Features: Fluid Network Model

- Packet level models: Full net stack. Inherently slow, hard to instantiate
- Simple models: Delay-based, distribution, coordinates
  Very scalable, but no topology, *no network congestion*
- Fluid models: **Share bandwidth between flows** on macroscopic evts

(bandwidth) Sharing as an optimization problem

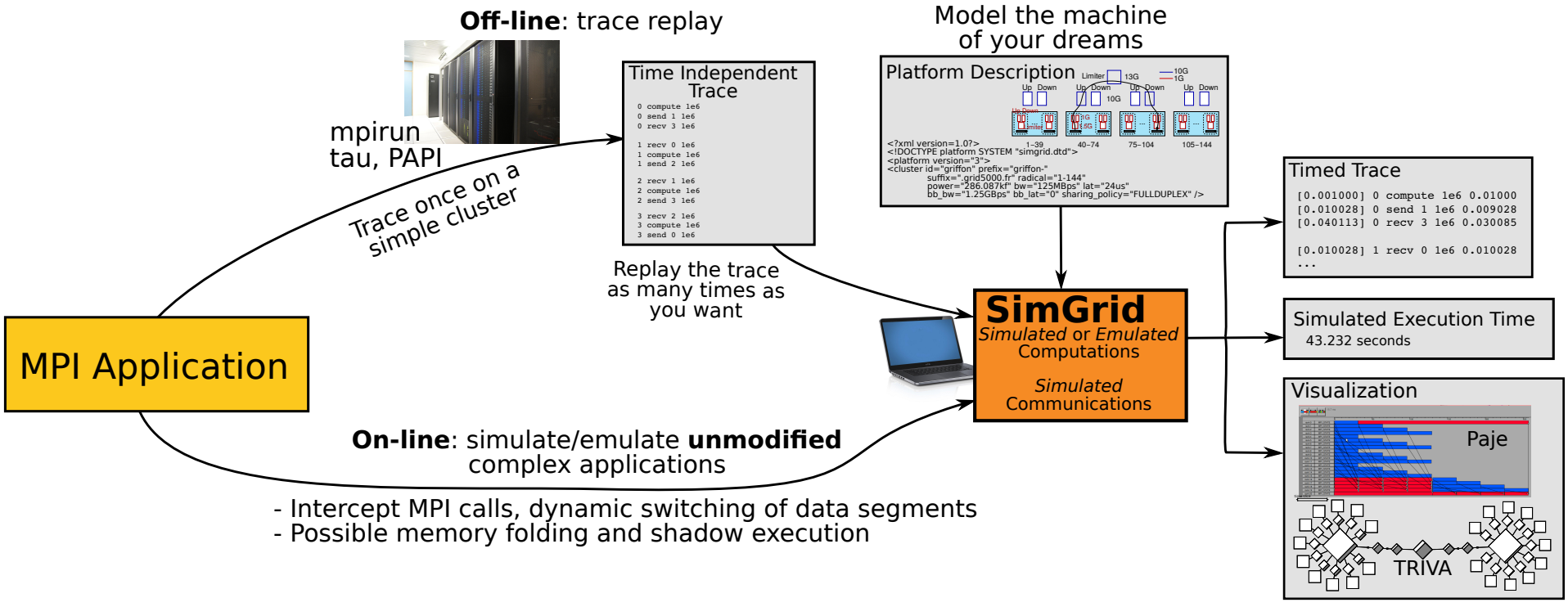
$$\sum_{\text{if flow i uses link j}} \rho_i \leqslant C_j$$

- Max-Min objective function: $\max\left(\min\left(\rho_i\right)\right)$
- Reno fairness: $\max\left(\sum \arctan\left(\rho_i\right)\right)$
- Vegas fairness: $\max\left(\sum \log\left(\rho_i\right)\right)$



We implemented, (in)validated and optimized these models

SimGrid Key Features: *mulation



Offline Simulation

Most tools use this approach

- Large traces are a pain
- Extrapolation?
- Adaptive applications?

Online Simulation

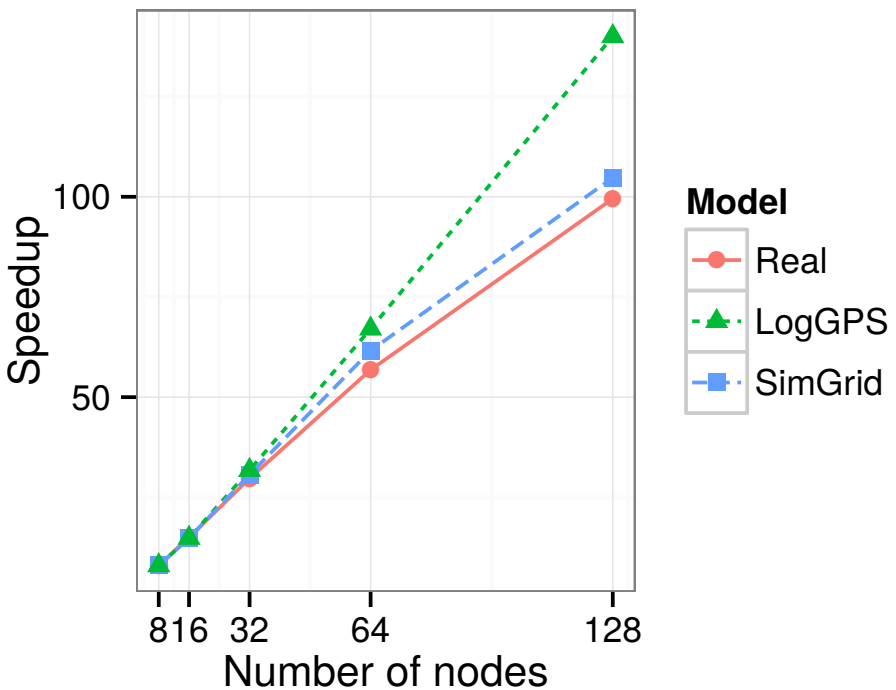
Works out of the box with NAS PB, SpecFEM3D, Ondes3D, BigDFT

- Annotations allow scaling

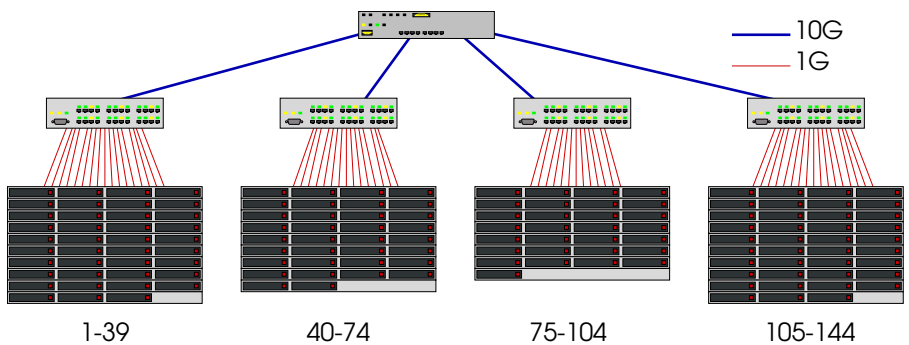
Success Stories I

BigDFT on a prototype ARM-based cluster from BSC (Mont-Blanc)

Key modeling aspects to obtain such results:

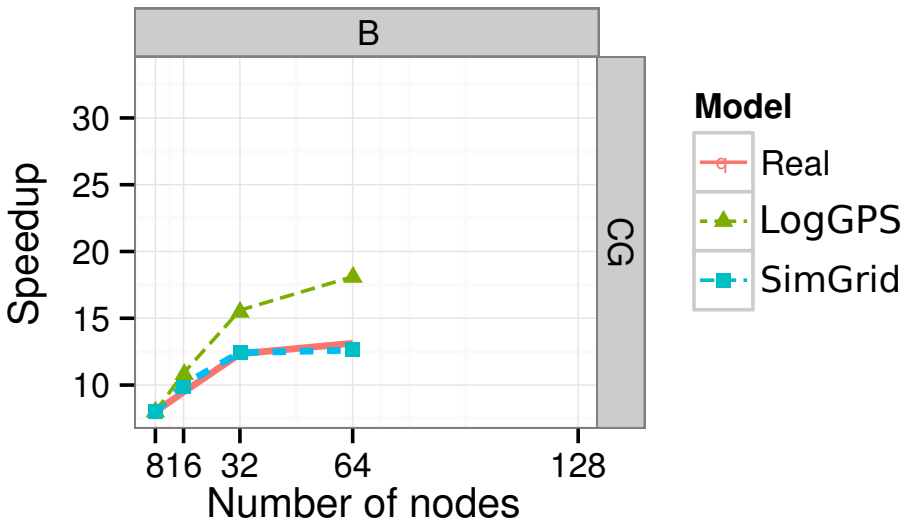


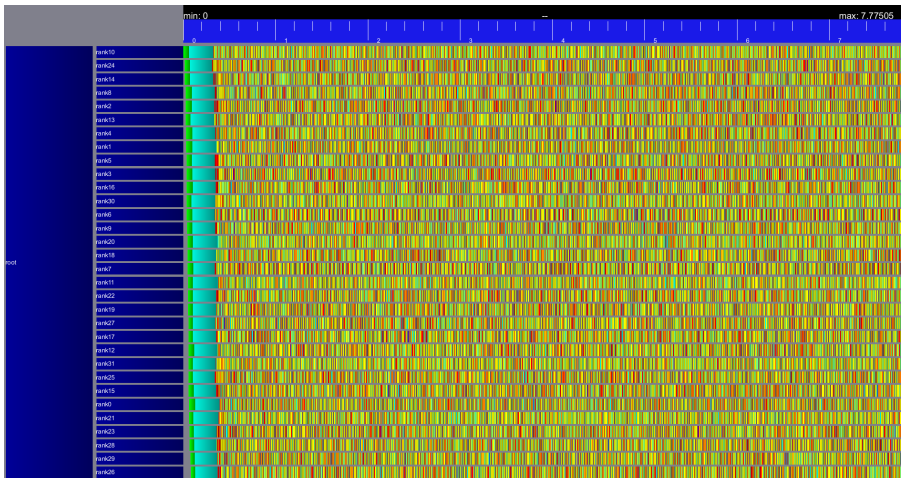
- Topology and contention...
- Collective operations
  Stolen from real implems
- Correct platform description
  Matching *effects*, not HW doc ;)



Success Stories II

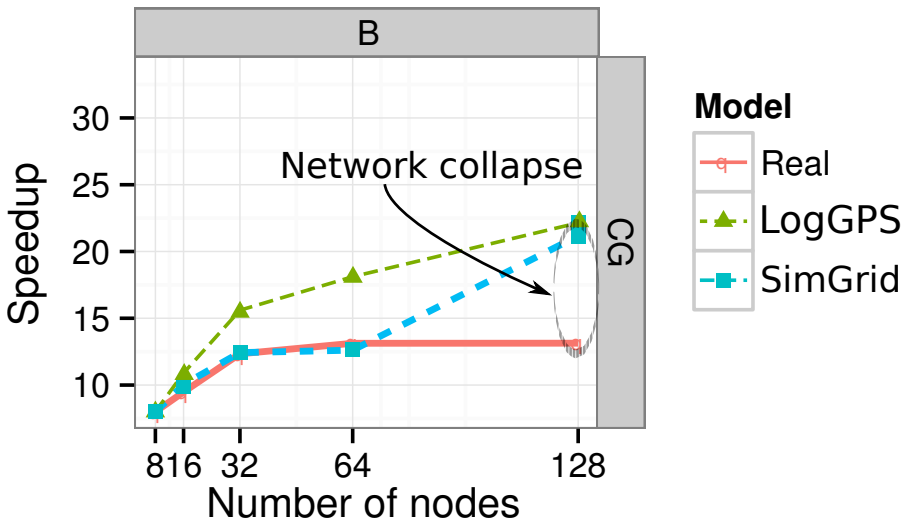
NAS CG on a TCP/Ethernet cluster (Grid5000)

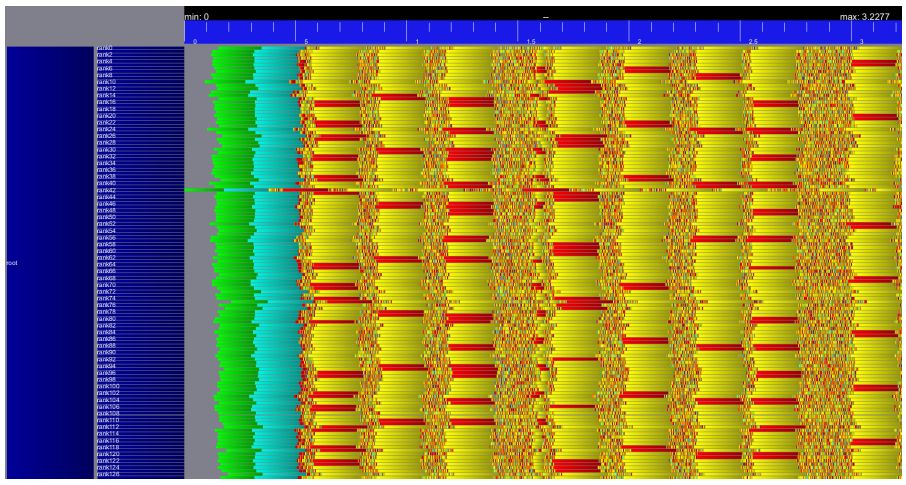




Success Stories II

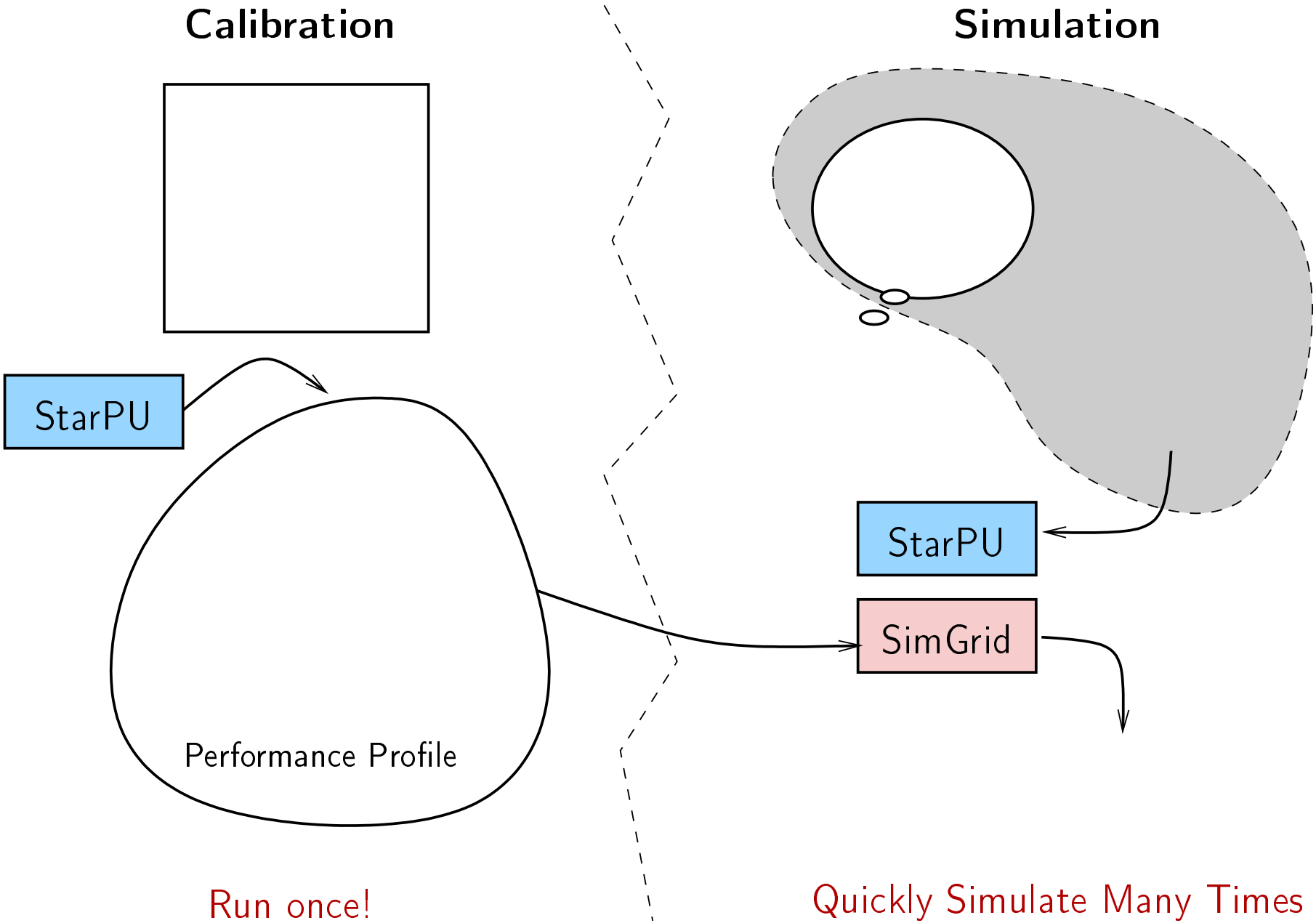
NAS CG on a TCP/Ethernet cluster (Grid5000)



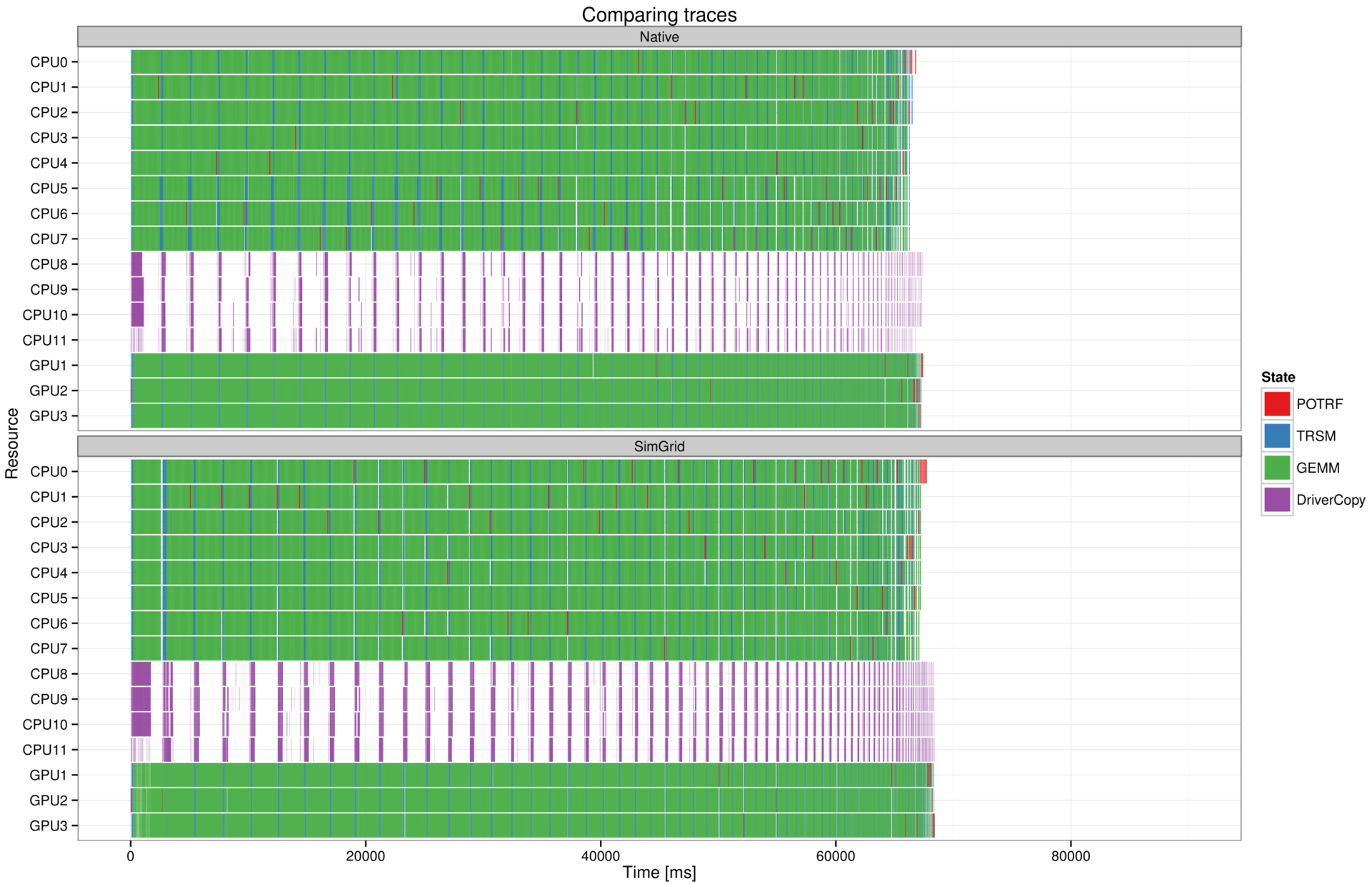


- Massive switch packet drops lead to 200ms timeouts in TCP!
- Tightly coupled: the whole application hangs until timeout
- Noise easy to model in the simulator, but useless for that very study
- Our prediction performance is more interesting to detect the real issue

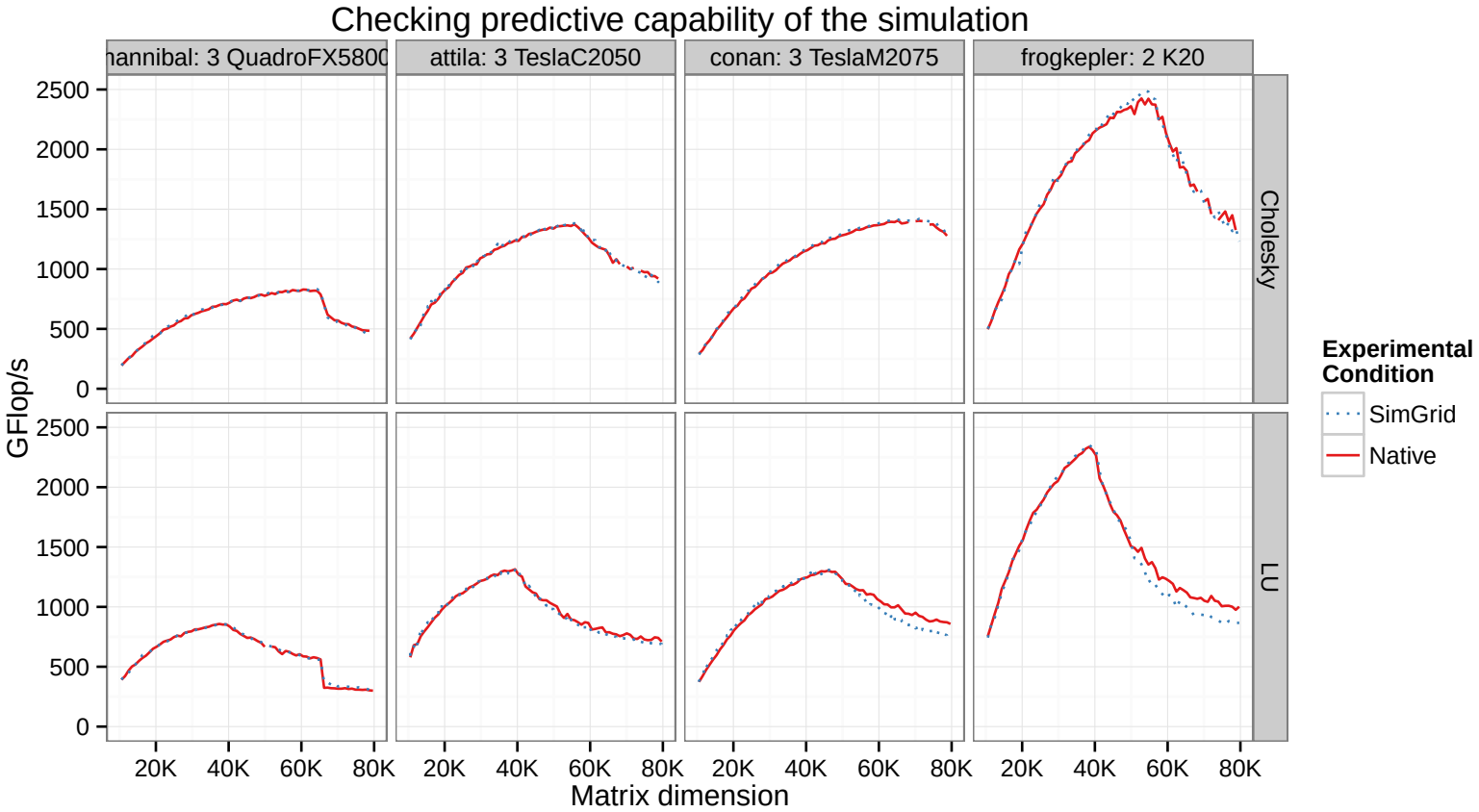
StarPU SimGrid



StarPU SimGrid



StarPU SimGrid



Key aspects to obtain such results:

- Model heterogeneity and contention (on comms) and Platform model
- Applicative model: virtualize and model the right functions and kernels
- Open Science to trust the results of these tedious (in-)validation
- And now, these results (calibration+modeling) are reused by others

SimGrid, the Hemera project and beyond

Initial proposal

- Challenge 1: Scientific instrument for the domain specialists
  - Scalable simulation kernel for large scale systems assessed on Grid'5000
  - Validated models, using Grid'5000 for (in-)validations
- Challenge 2: Trustable description of the Grid'5000 platform

Mid-term evaluation

- SimGrid evolved from a prototyping tool toy to a scientific instrument
- Presented research efforts enabled by Grid'5000 (experimental quality)

Today

- SimGrid quickens development of HPC systems: Debug your SW & HW
- Research efforts enabled by the Hemera community (beyond our ANR)

SimGrid, the Hemera project and beyond

Initial proposal

- Challenge 1: Scientific instrument for the domain specialists
  - Scalable simulation kernel for large scale systems assessed on Grid'5000
  - Validated models, using Grid'5000 for (in-)validations
- Challenge 2: Trustable description of the Grid'5000 platform

Mid-term evaluation

- SimGrid evolved from a prototyping tool toy to a scientific instrument
- Presented research efforts enabled by Grid'5000 (experimental quality)

Today

- SimGrid quickens development of HPC systems: Debug your SW & HW
- Research efforts enabled by the Hemera community (beyond our ANR)

Our Future Work on SimGrid

- Focus on HPC; Detect SW/HW defects; Formal assessment; Teaching

SimGrid, the Hemera project and beyond

Initial proposal

- Challenge 1: Scientific instrument for the domain specialists
  - Scalable simulation kernel for large scale systems assessed on Grid'5000
  - Validated models, using Grid'5000 for (in-)validations
- Challenge 2: Trustable description of the Grid'5000 platform

Mid-term evaluation

- SimGrid evolved from a prototyping tool toy to a scientific instrument
- Presented research efforts enabled by Grid'5000 (experimental quality)

Today

- SimGrid quickens development of HPC systems: Debug your SW & HW
- Research efforts enabled by the Hemera community (beyond our ANR)

Our Future Work on SimGrid

- Focus on HPC; Detect SW/HW defects; Formal assessment; Teaching
- **Concern: we need an instrument providing $\frac{\text{Hemera}}{\text{Grid'5000}}$'s stack**