Instituto de computação

MUSIC – Kikoff Meeting

Cloud-based Scientific Workflow Management

Prof. D.Sc. Daniel de Oliveira
danielcmo@ic.uff.br

HPC Scientific Cloud Scenario

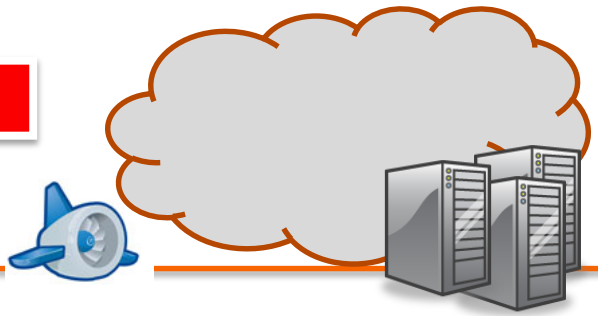
Some initiatives propose solutions for managing parallel executions workflows in cloud environments

1. Data is generated and collected

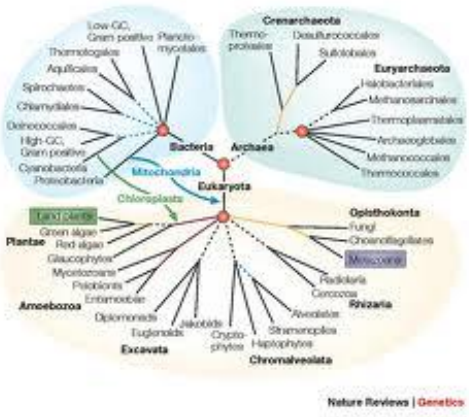
2. Data is first analyzed by programs MAFFT, Muscle...

3. Large volume of data are produced...

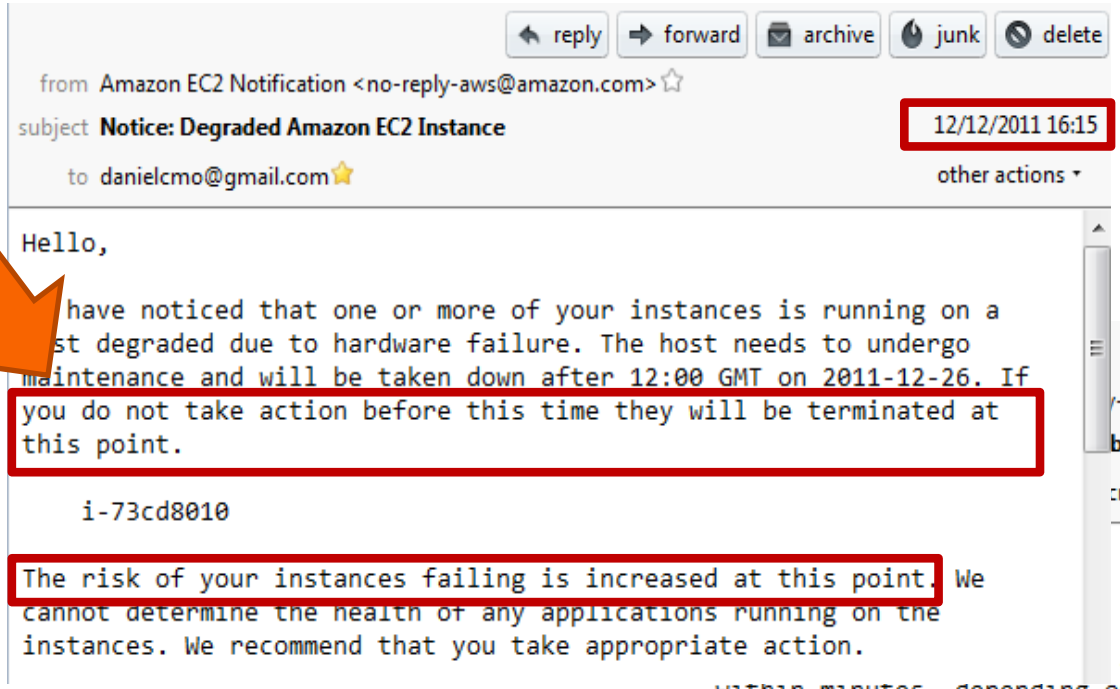
4. ...which need to be processed ay a virtual cluster using MG and RAxML



5. Final results are analyzed



Challenges....



reply forward archive junk delete

from Amazon EC2 Notification <no-reply-aws@amazon.com>

subject **Notice: Degraded Amazon EC2 Instance** 12/12/2011 16:15

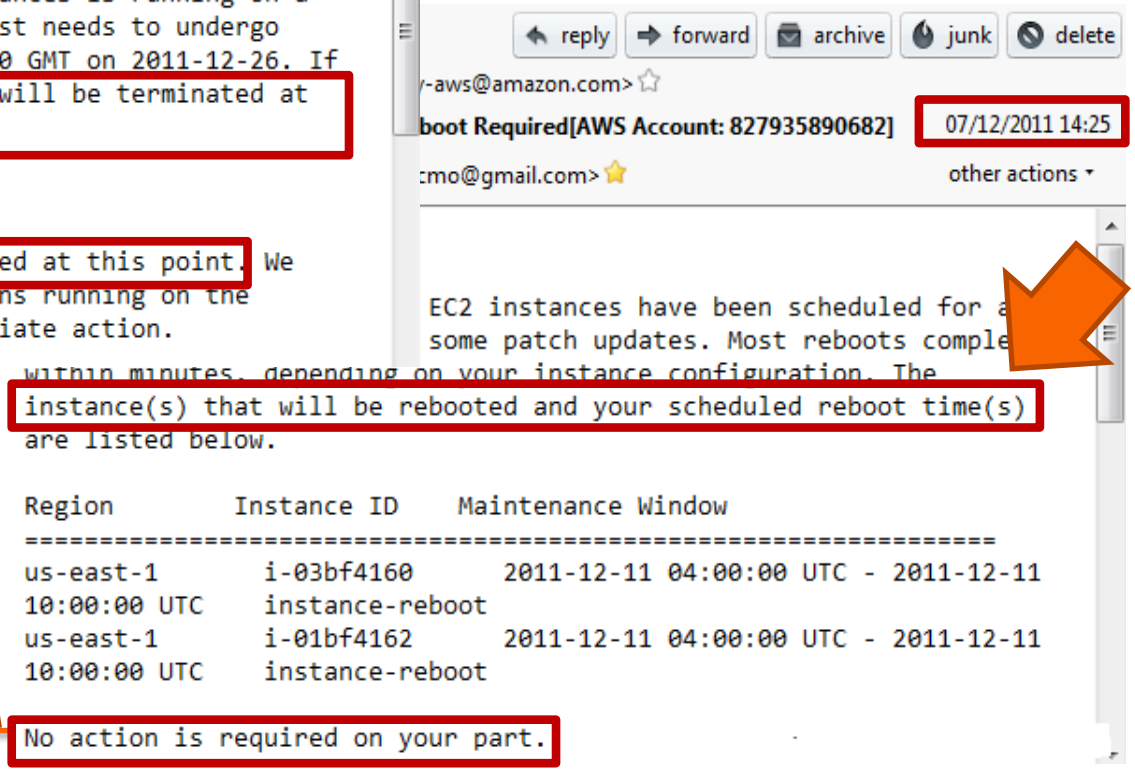
to danielcmo@gmail.com other actions

Hello,

I have noticed that one or more of your instances is running on a host degraded due to hardware failure. The host needs to undergo maintenance and will be taken down after 12:00 GMT on 2011-12-26. If you do not take action before this time they will be terminated at this point.

i-73cd8010

The risk of your instances failing is increased at this point. We cannot determine the health of any applications running on the instances. We recommend that you take appropriate action.



reply forward archive junk delete

-aws@amazon.com>

boot Required[AWS Account: 827935890682] 07/12/2011 14:25

cmo@gmail.com> other actions

EC2 instances have been scheduled for a some patch updates. Most reboots comple within minutes, depending on your instance configuration. The instance(s) that will be rebooted and your scheduled reboot time(s) are listed below.

Region	Instance ID	Maintenance Window
us-east-1	i-03bf4160	2011-12-11 04:00:00 UTC - 2011-12-11 10:00:00 UTC
	instance-reboot	
us-east-1	i-01bf4162	2011-12-11 04:00:00 UTC - 2011-12-11 10:00:00 UTC
	instance-reboot	

No action is required on your part.

SciCumulus Cloud Engine

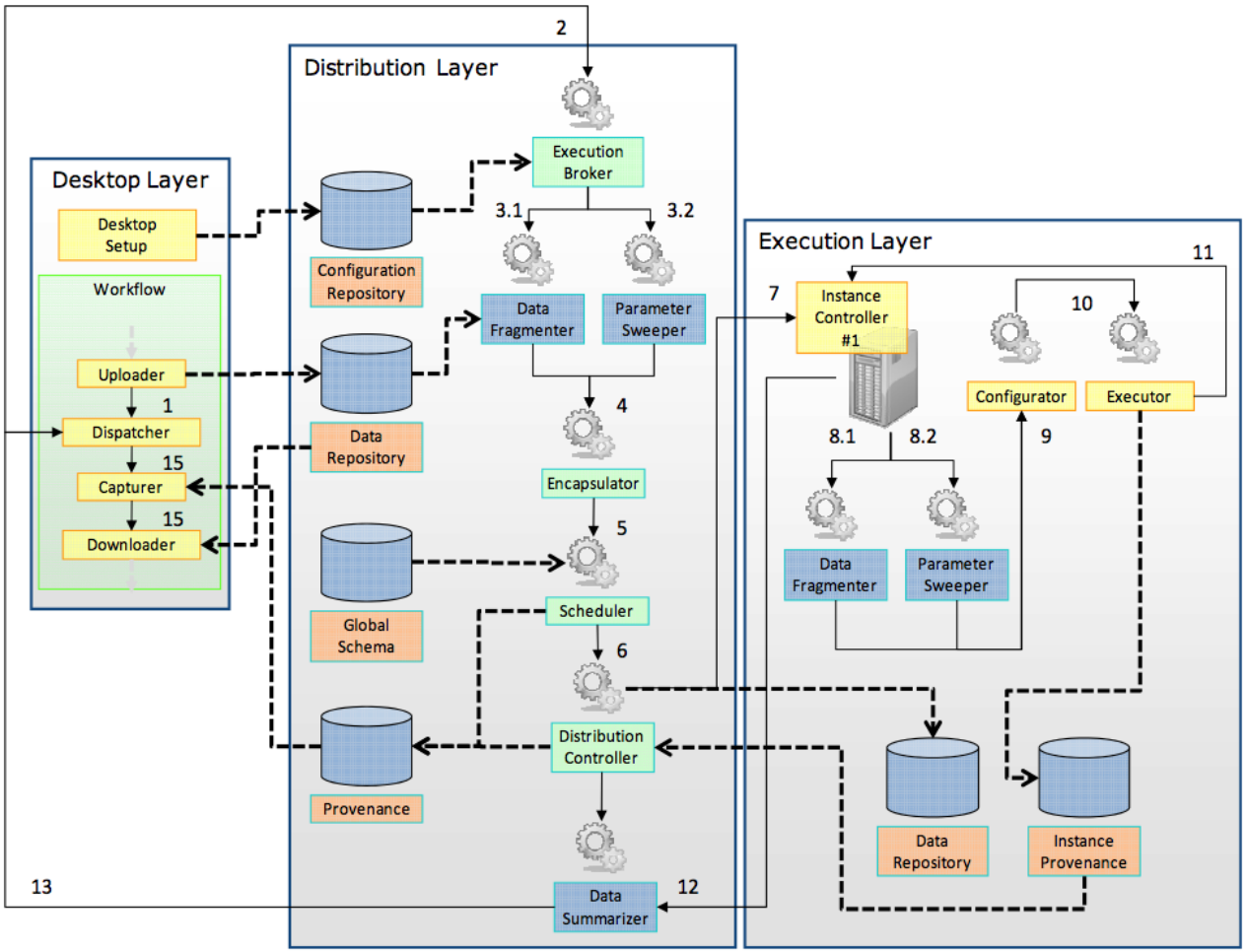
- SciCumulus is a parallel workflow engine for cloud environments
- Implements a provenance-based adaptive scheduling heuristic
- Presents a 3-objective weighted cost model considering **total execution time**, **financial cost** and **reliability**

SciCumulus Cloud Engine

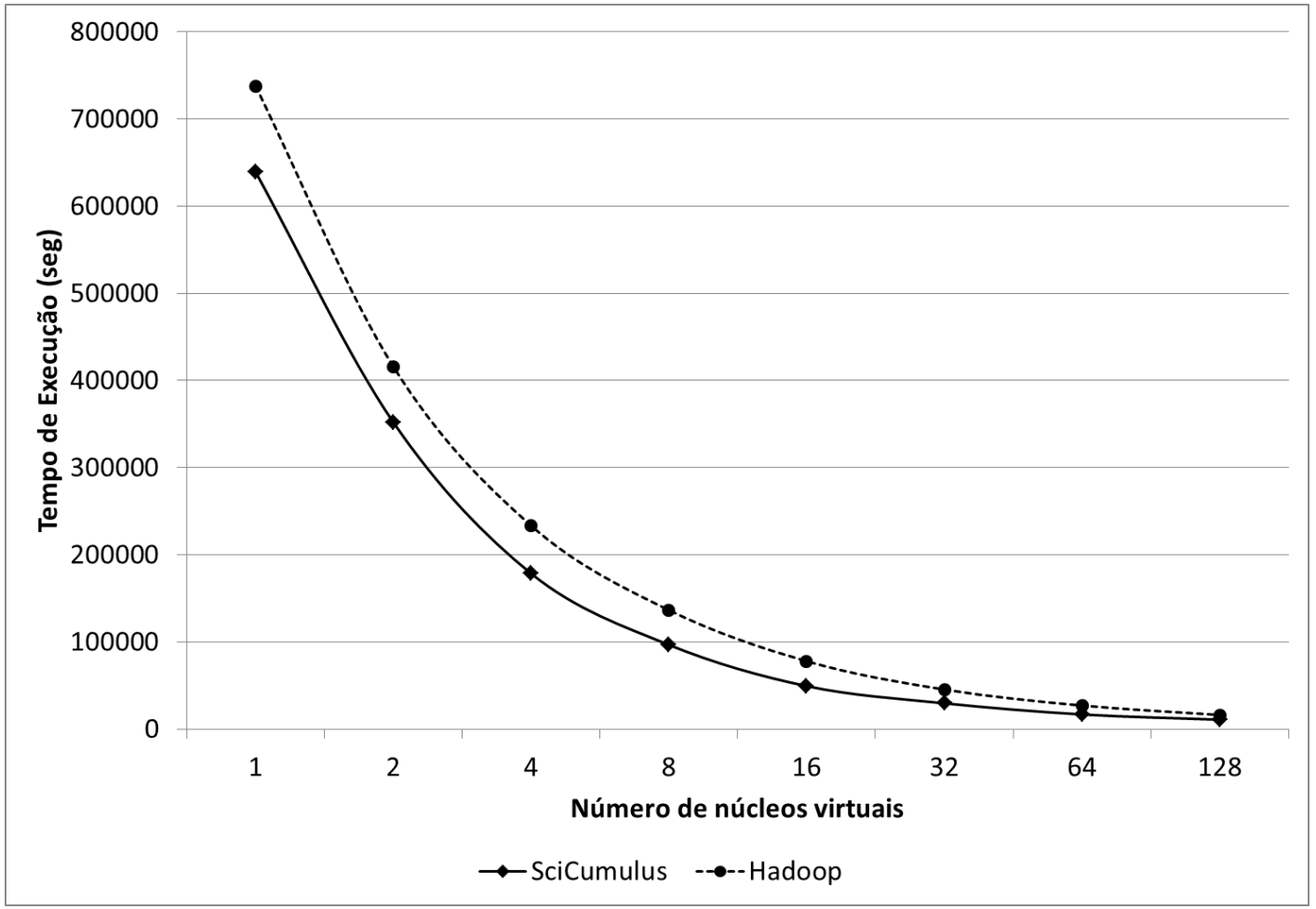
- SciCumulus is a parallel workflow engine for cloud environments
- Implements a provenance-based adaptive scheduling heuristic
- Presents a 3-objective weighted cost model considering **total execution time**, **financial cost** and **reliability**

OLIVEIRA, D. ; OGASAWARA, E. ; BAIAO, F. ; MATTOSO, M. L. Q. . SciCumulus: A Lightweight Cloud Middleware to Explore Many Task Computing Paradigm in Scientific Workflows. In: The 3rd IEEE CLOUD, 2010. p. 378-385.

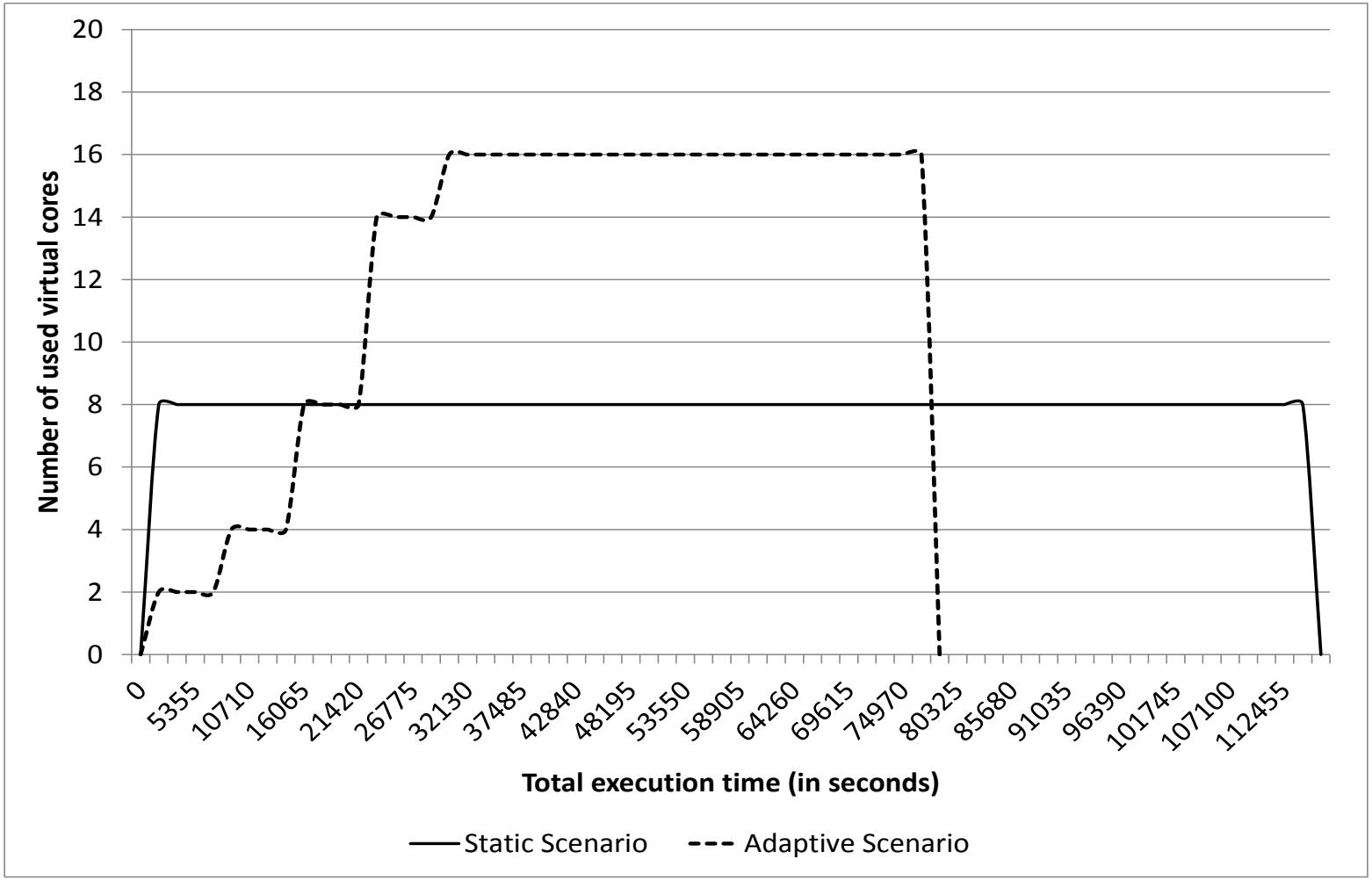
SciCumulus Architecture



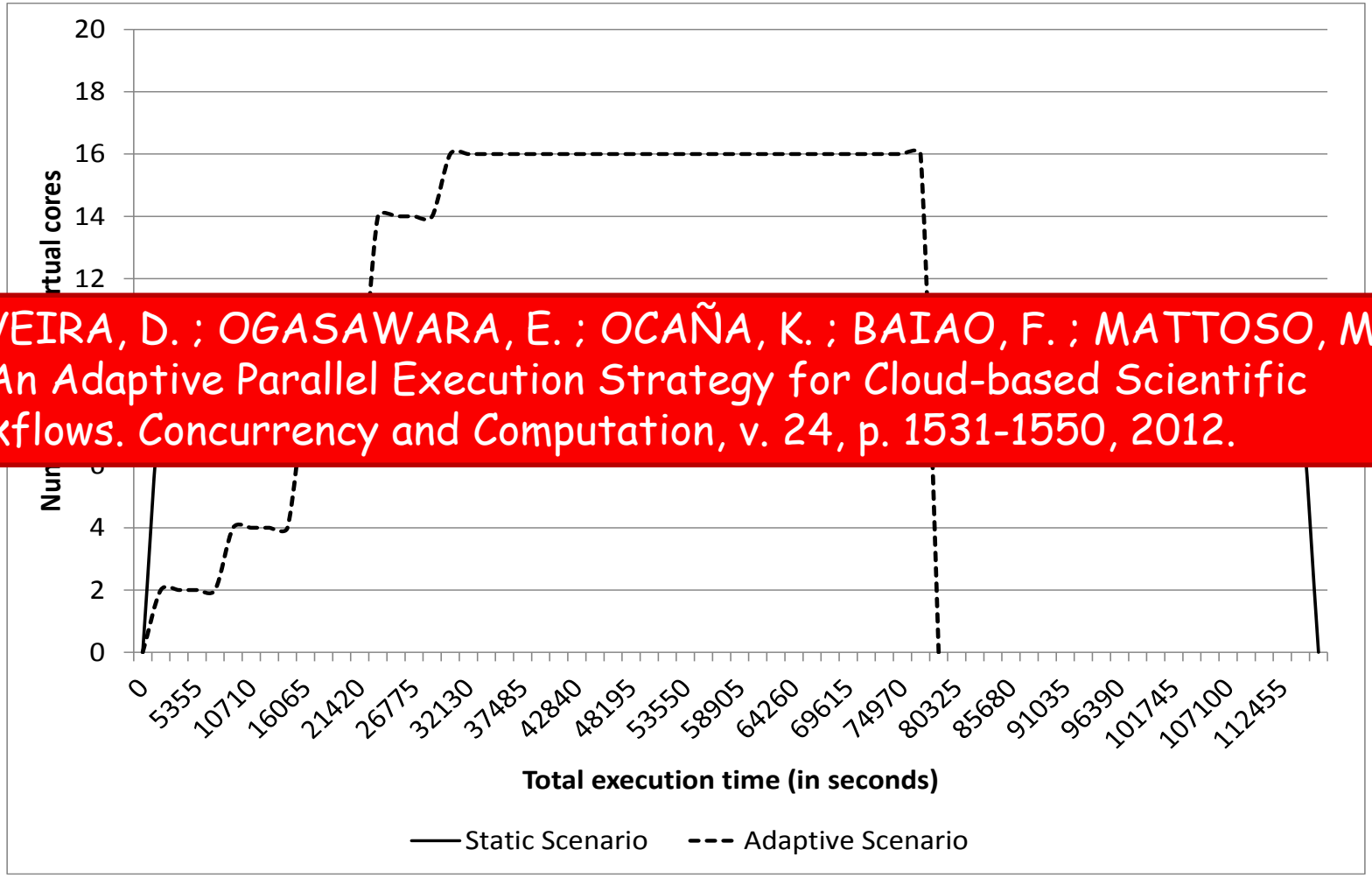
Static Execution



Analysis of the Adaptive Execution

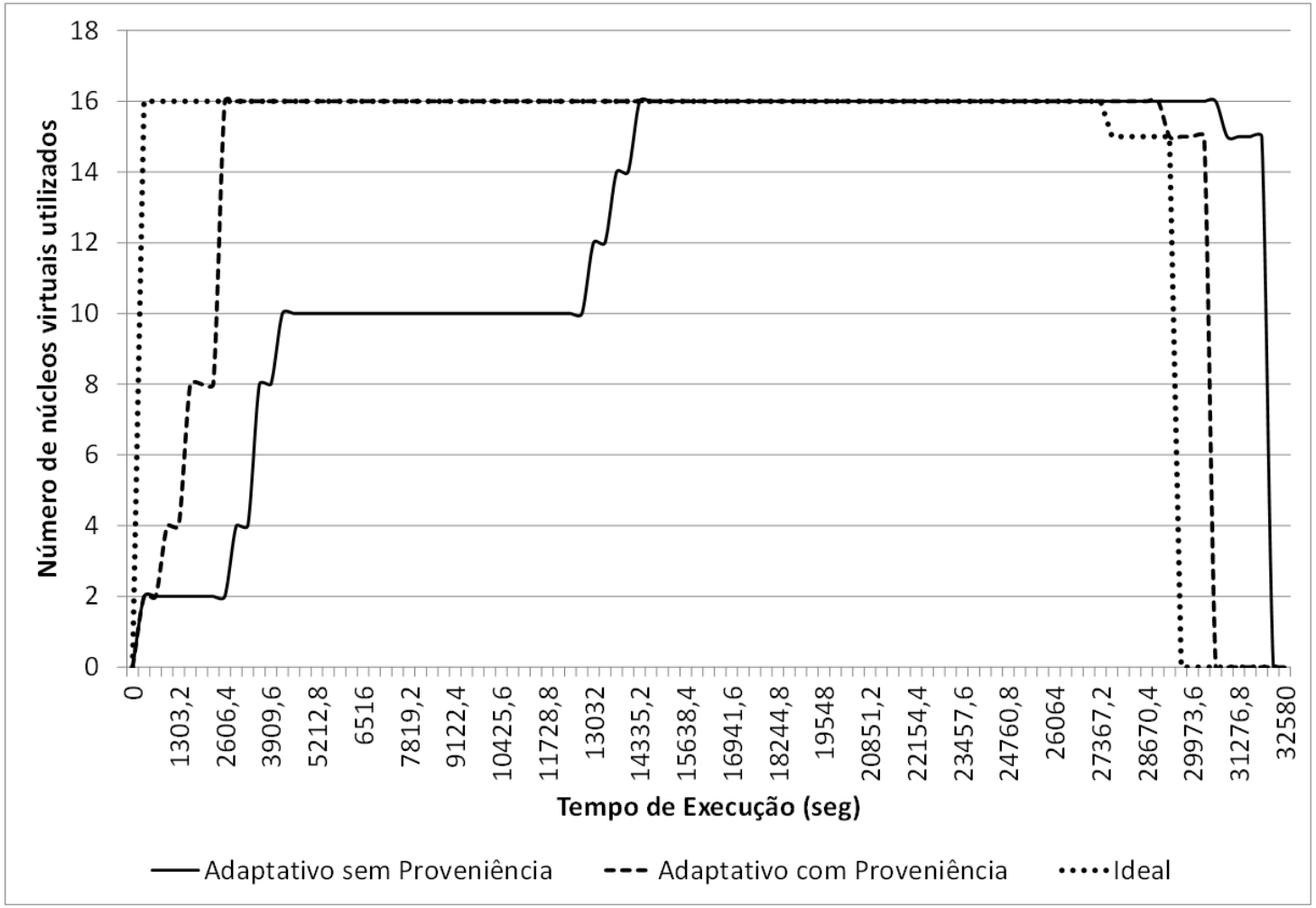


Analysis of the Adaptive Execution

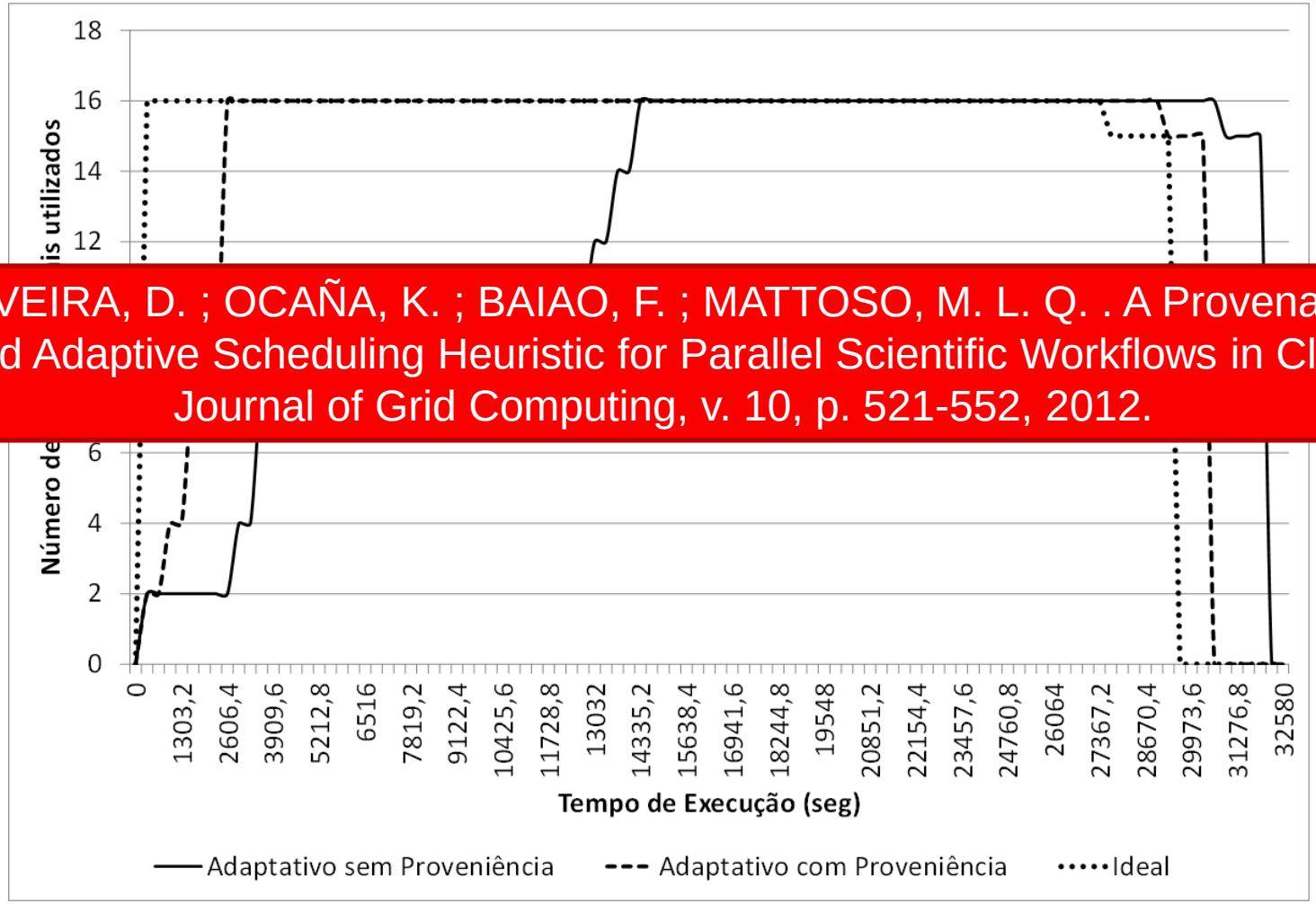


OLIVEIRA, D. ; OGASAWARA, E. ; OCAÑA, K. ; BAIAO, F. ; MATTOSO, M. L. Q. . An Adaptive Parallel Execution Strategy for Cloud-based Scientific Workflows. Concurrency and Computation, v. 24, p. 1531-1550, 2012.

Adaptive Execution

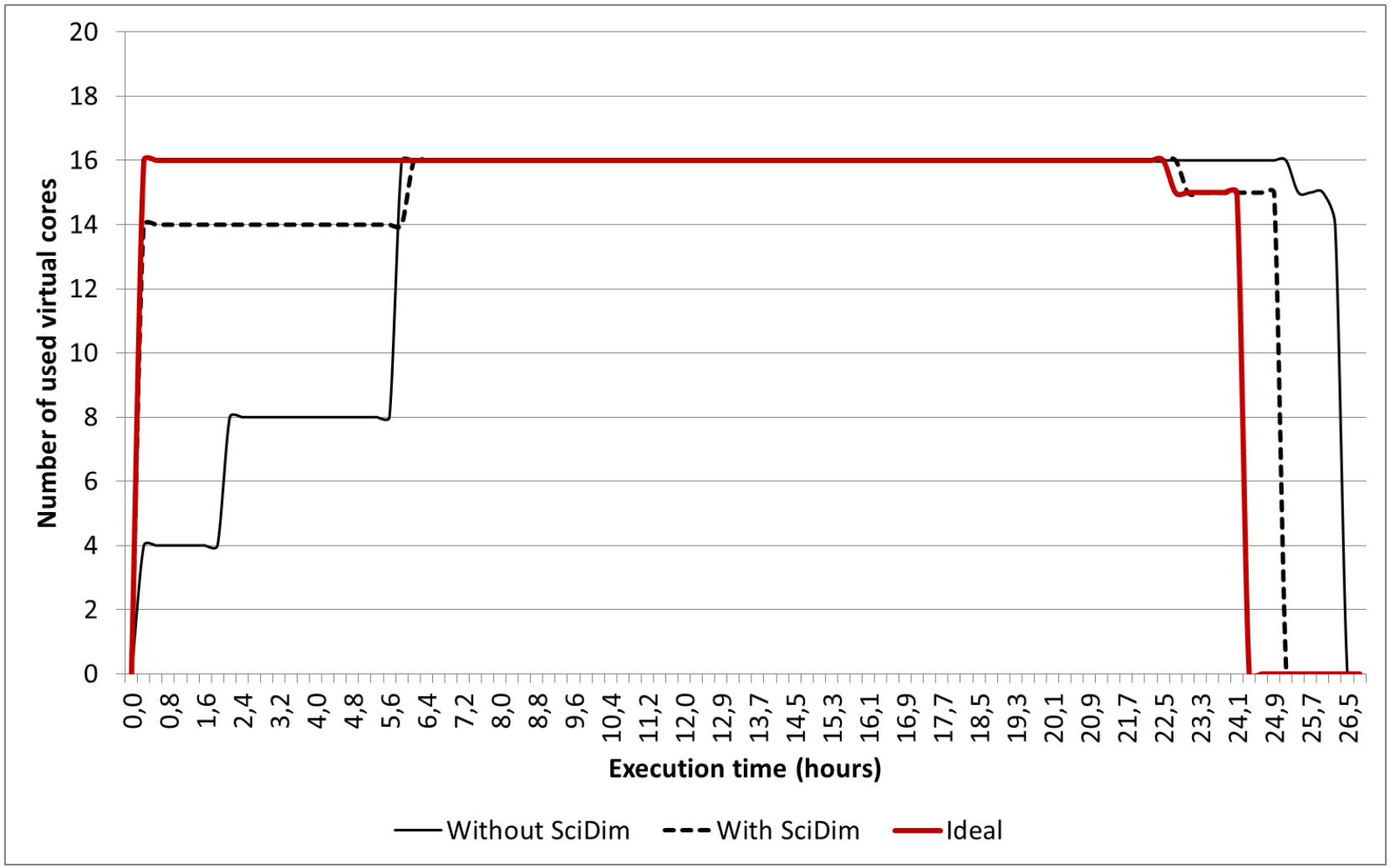


Adaptive Execution

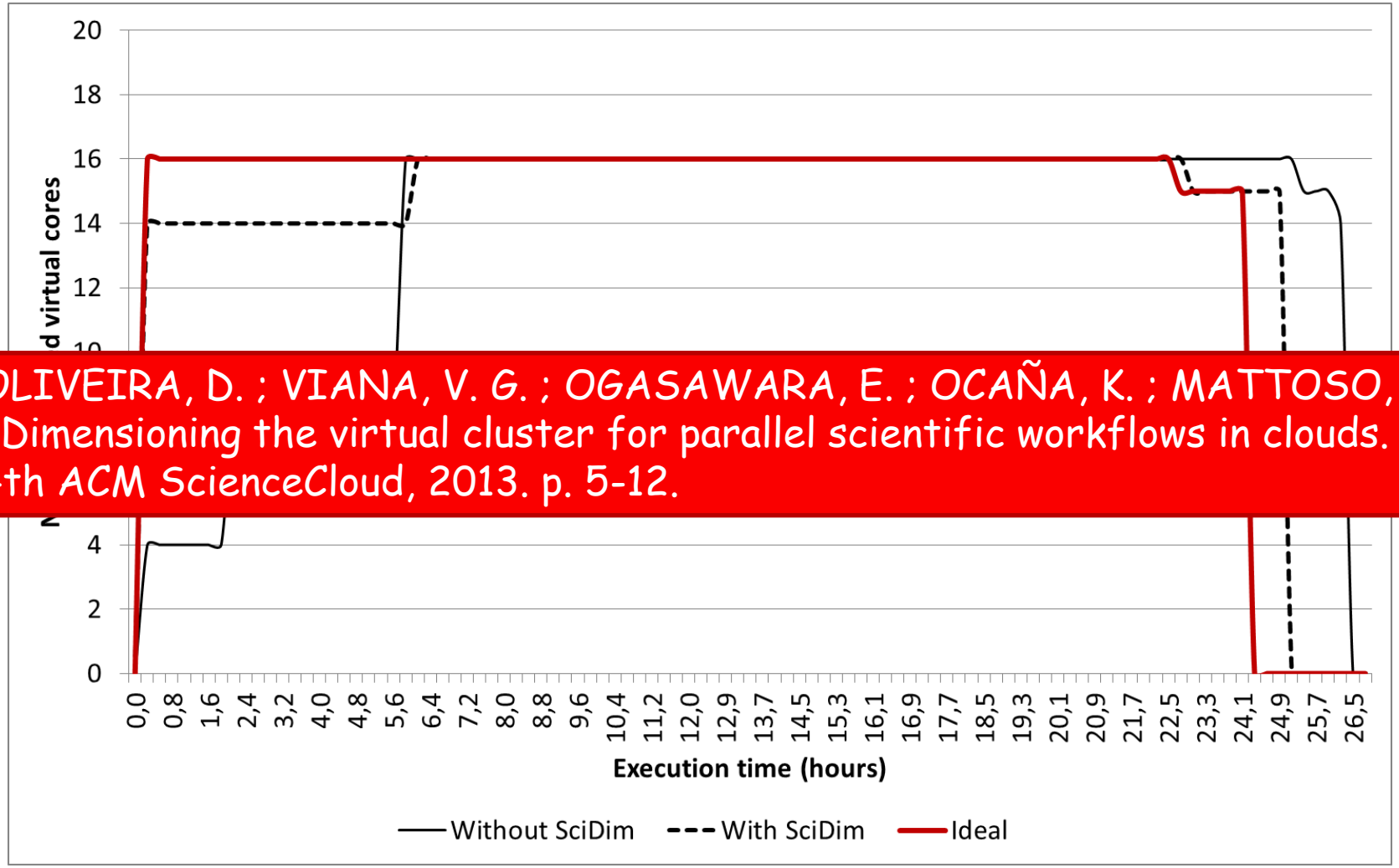


OLIVEIRA, D. ; OCAÑA, K. ; BAIAO, F. ; MATTOSO, M. L. Q. . A Provenance-based Adaptive Scheduling Heuristic for Parallel Scientific Workflows in Clouds. Journal of Grid Computing, v. 10, p. 521-552, 2012.

Dimensioned Adaptive Execution

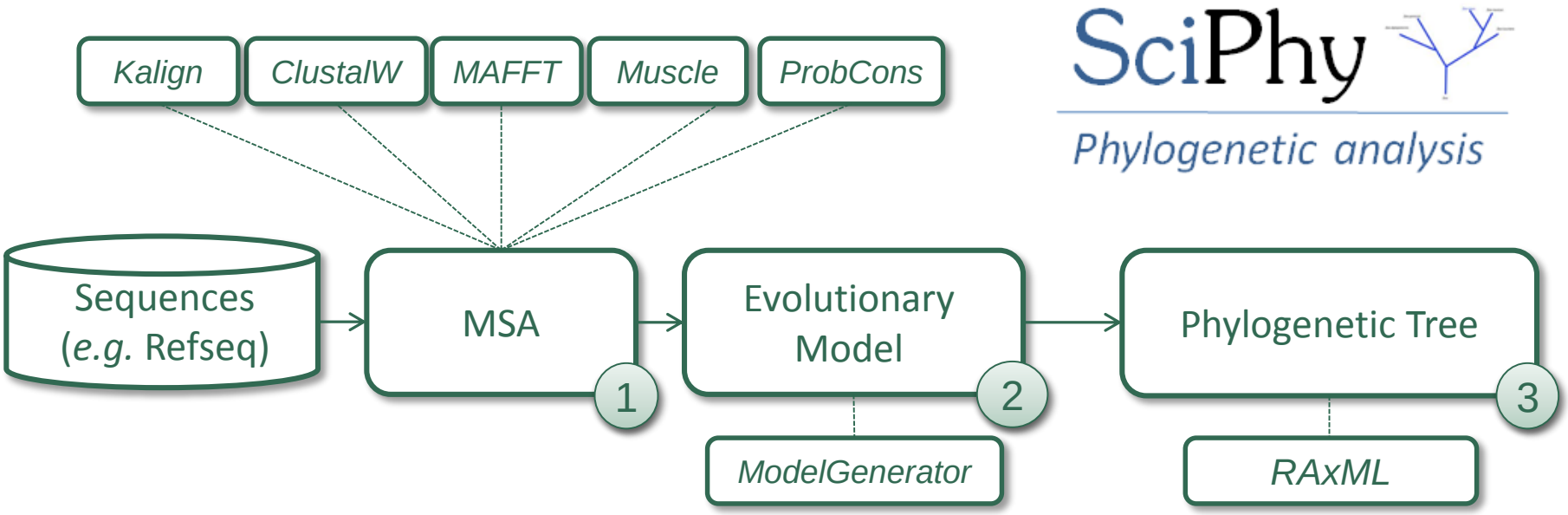


Dimensioned Adaptive Execution



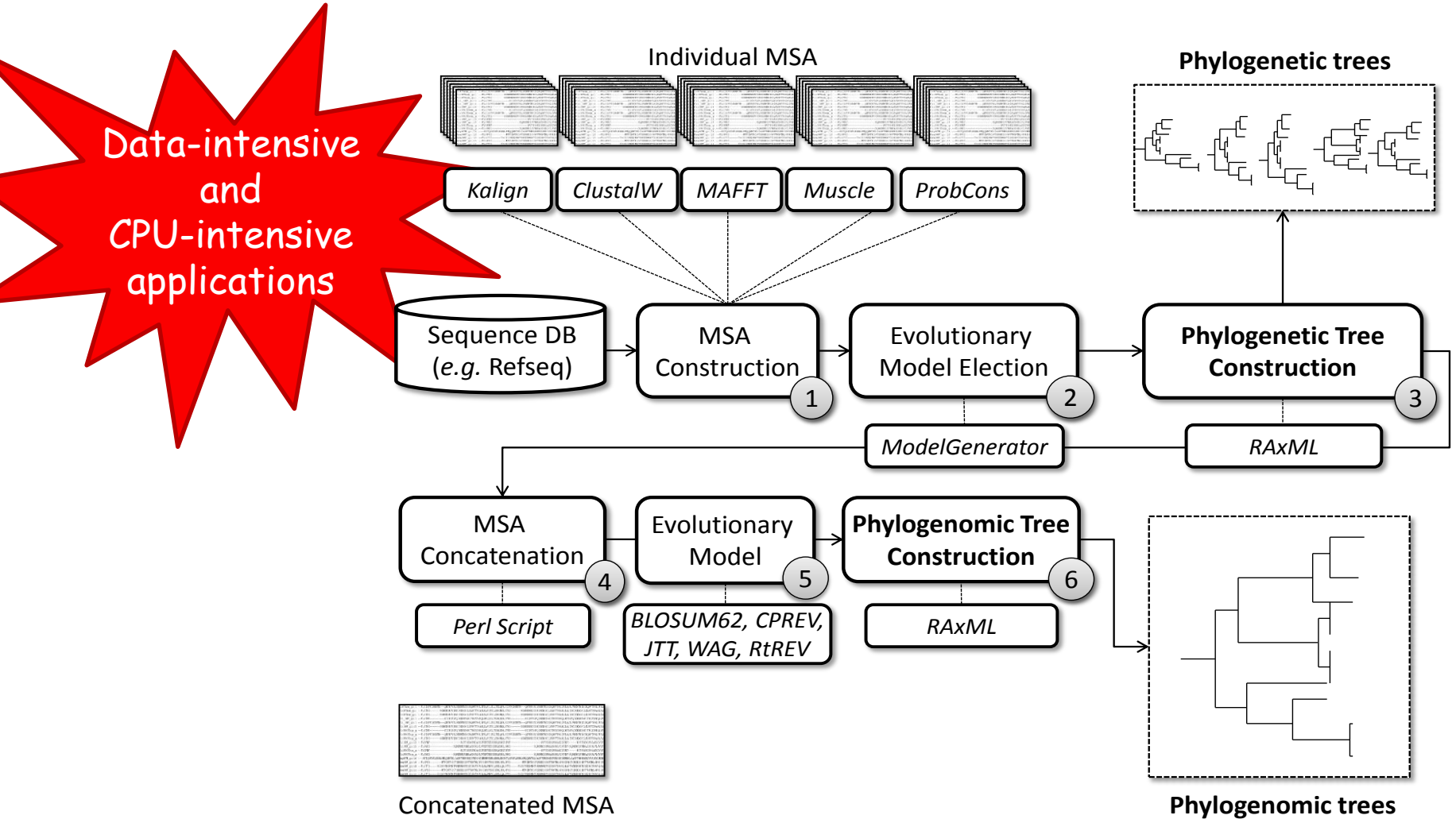
OLIVEIRA, D. ; VIANA, V. G. ; OGASAWARA, E. ; OCAÑA, K. ; MATTOSO, M. . Dimensioning the virtual cluster for parallel scientific workflows in clouds. In: 4th ACM ScienceCloud, 2013. p. 5-12.

Phylogenetic Analysis Workflow

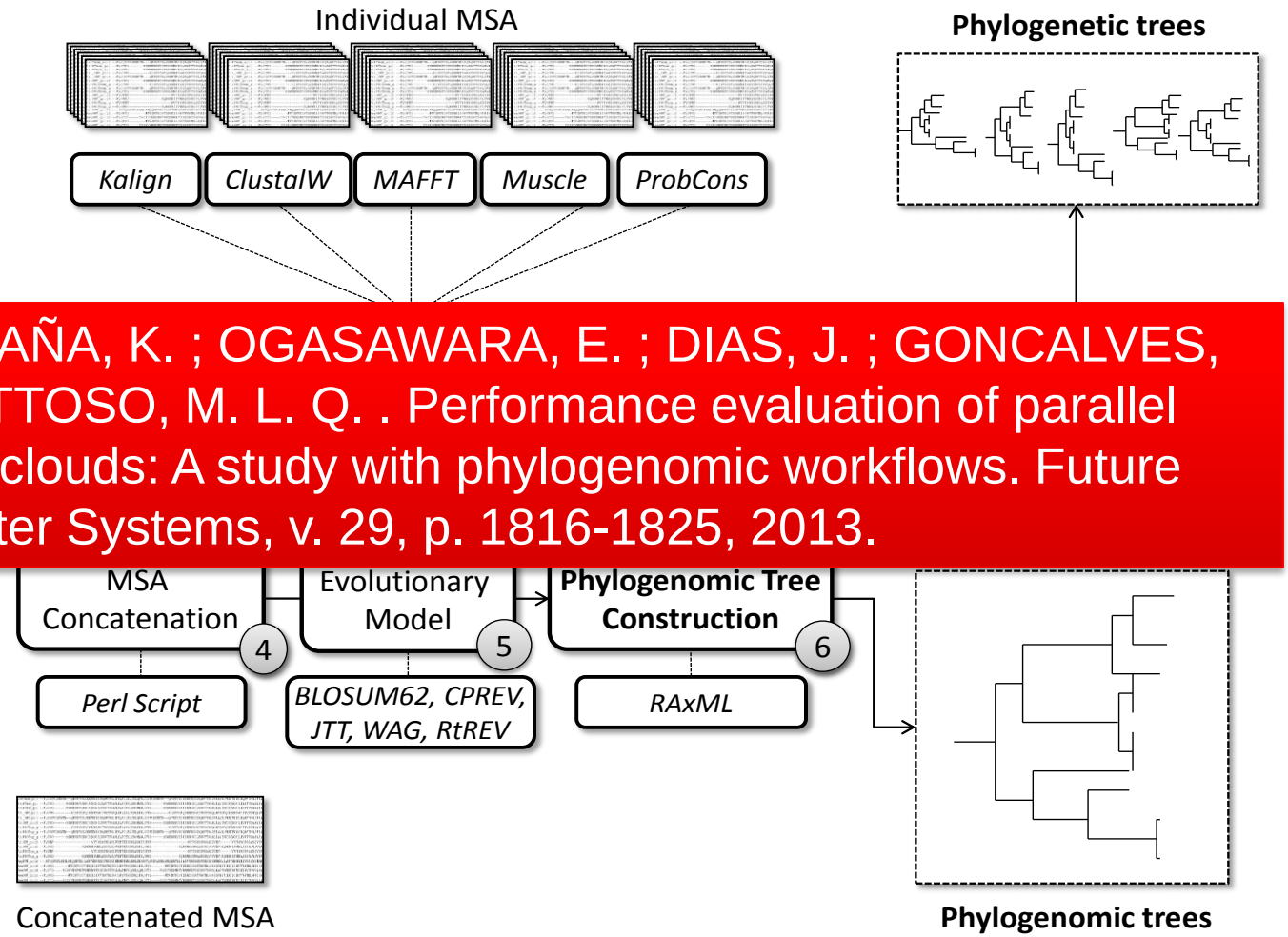


Ocaña K.A.C.S., Oliveira D., Ogasawara E., Dávila A.M.R., Lima A.A.B., and Mattoso M. **SciPhy: A Cloud-Based Workflow for Phylogenetic Analysis of Drug Targets in Protozoan Genomes**. Springer Berlin Heidelberg, pp. 66-70, 2011.

Phylogenomic Analysis Workflow



Phylogenomic Analysis Workflow



OLIVEIRA, D. ; OCAÑA, K. ; OGASAWARA, E. ; DIAS, J. ; GONCALVES, J. ; BAIAO, F. ; MATTOSO, M. L. Q. . Performance evaluation of parallel strategies in public clouds: A study with phylogenomic workflows. Future Generation Computer Systems, v. 29, p. 1816-1825, 2013.

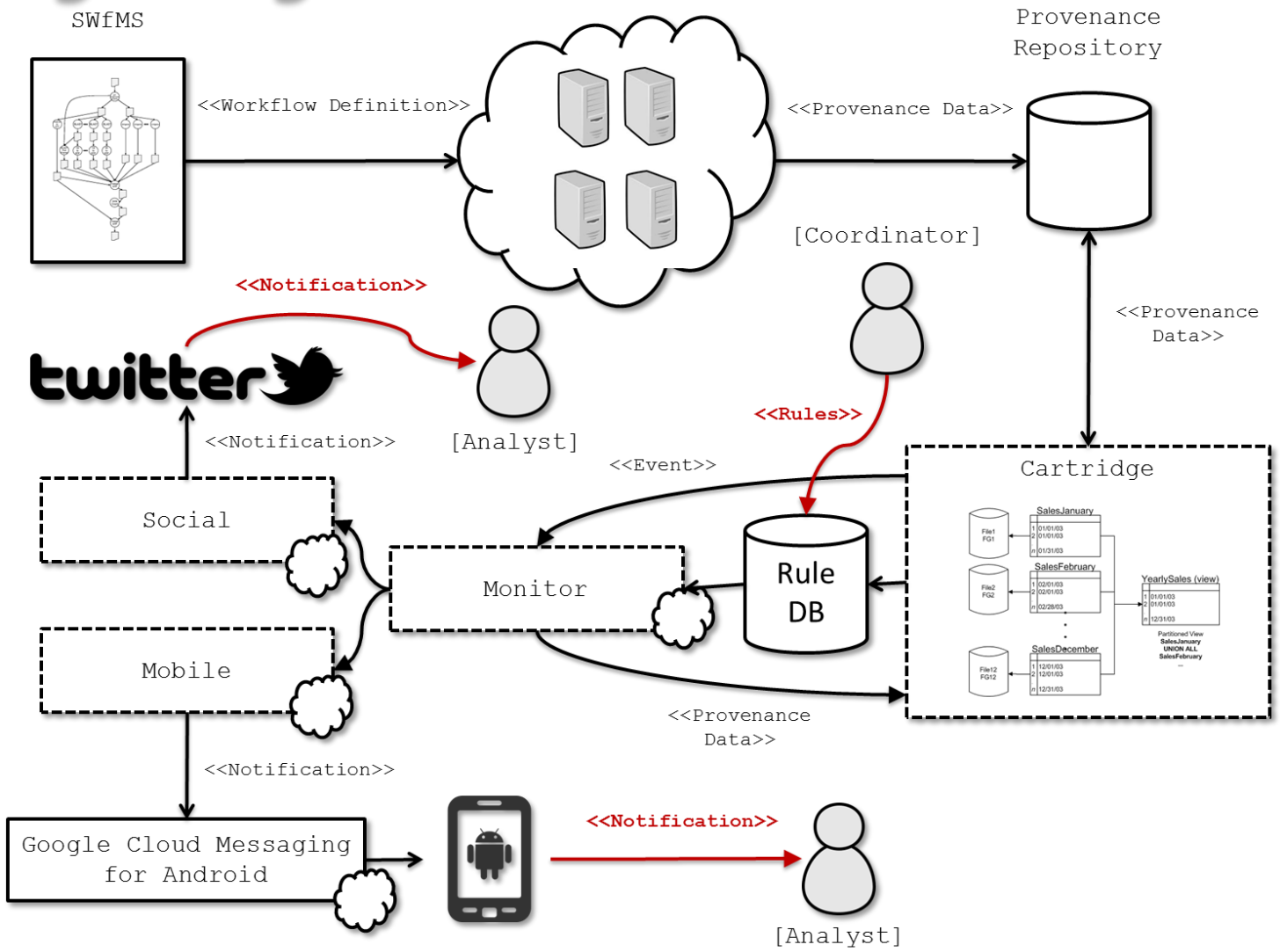
SciLightning

- Provides monitoring information for SWfMS
- A non-intrusive approach
- No change in the SWfMS is required
  - Provenance data has to be provided at runtime
- It is based on event notification
- Its components are distributed in the cloud (in several virtual machines) to monitor different scientific workflows on demand.

SciLightning

- Provides monitoring information for SWfMS
- A non-intrusive approach
- PINTAS, J. ; DE OLIVEIRA, D. ; OCAÑA, K. ; OGASAWARA, E. ; MATTOSO, M. L. Q. . SciLightning: a Cloud Provenance-based Event Notification for Parallel Workflows. In: 3rd International Workshop on Cloud Computing and Scientific Applications (CCSA), 2013, Berlim. Service-Oriented Computing - ICSOC 2013 Workshops, 2013. p. 1-6.
- It is based on event notification
- Its components are distributed in the cloud (in several virtual machines) to monitor different scientific workflows on demand.

SciLightning Architecture



Dimensioning

- Depending on the cloud provider, there may be several possible VM type combinations to choose and
  - it can be tedious, overpriced and error-prone to be performed manually.
- Use of genetic algorithms to find the best possible amount of VMs for a specific execution

OLIVEIRA, D. ; VIANA, V. G. ; OGASAWARA, E. ; OCAÑA, K. ; MATTOSO, M. . Dimensioning the virtual cluster for parallel scientific workflows in clouds. In: 4th ACM ScienceCloud, 2013. p. 5-12.

Opportunities

- Model new cloud-based scientific workflows for different experiments (including bioinformaticians in France)
- Extend the proposed approaches for multi-site cloud execution
  - Scheduling
  - Adaptive Execution (auto-scaling)
  - Monitoring
  - Dimensioning

Thank you!