POLITECNICO DI MILANO





Analysis of the influence of application deployment on energy consumption

M. Gribaudo, Nguyen T.T. Ho, B. Pernici, G. Serazzi

Dip. Elettronica, Informazione e Bioingegneria

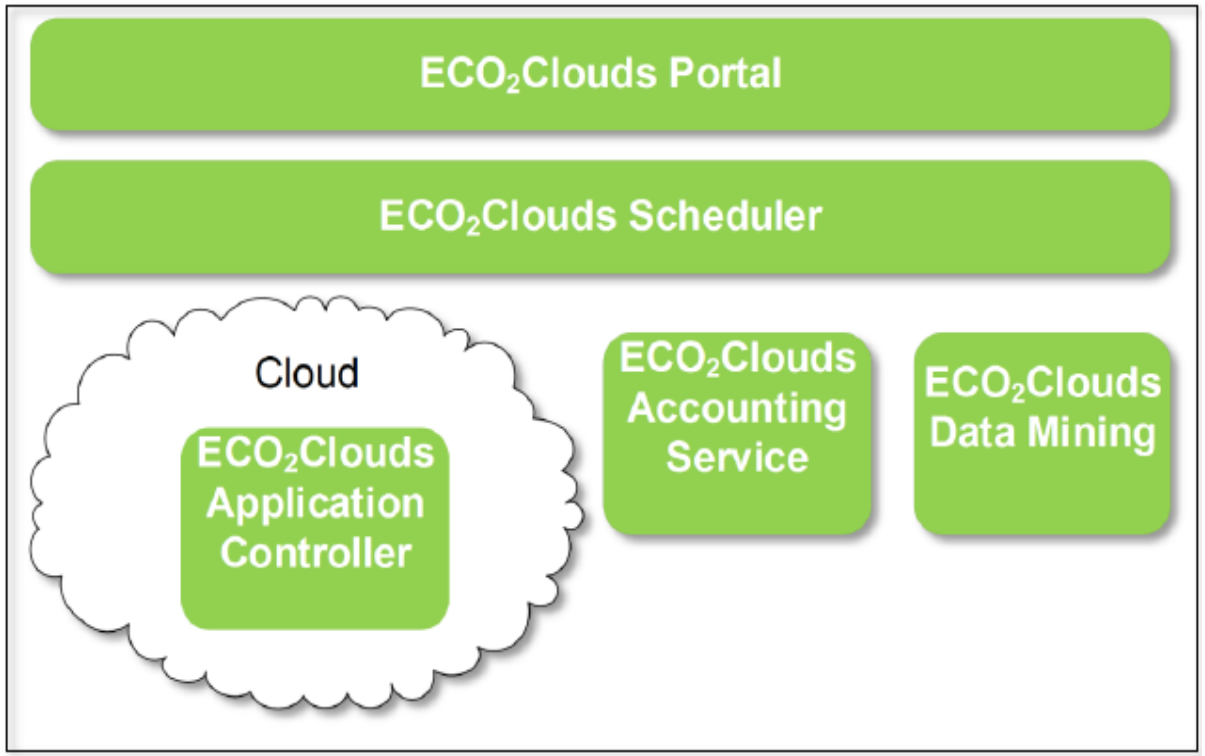
Politecnico di Milano

Motivation

- Data centers in clouds are the dominant contributor to CO_2 footprint
- Impact of application profile
  - Response time
  - CPU utilization
  - Memory usage
- Understand the influence of application deployment on energy consumption in cloud environments

ECO$_2$Clouds project

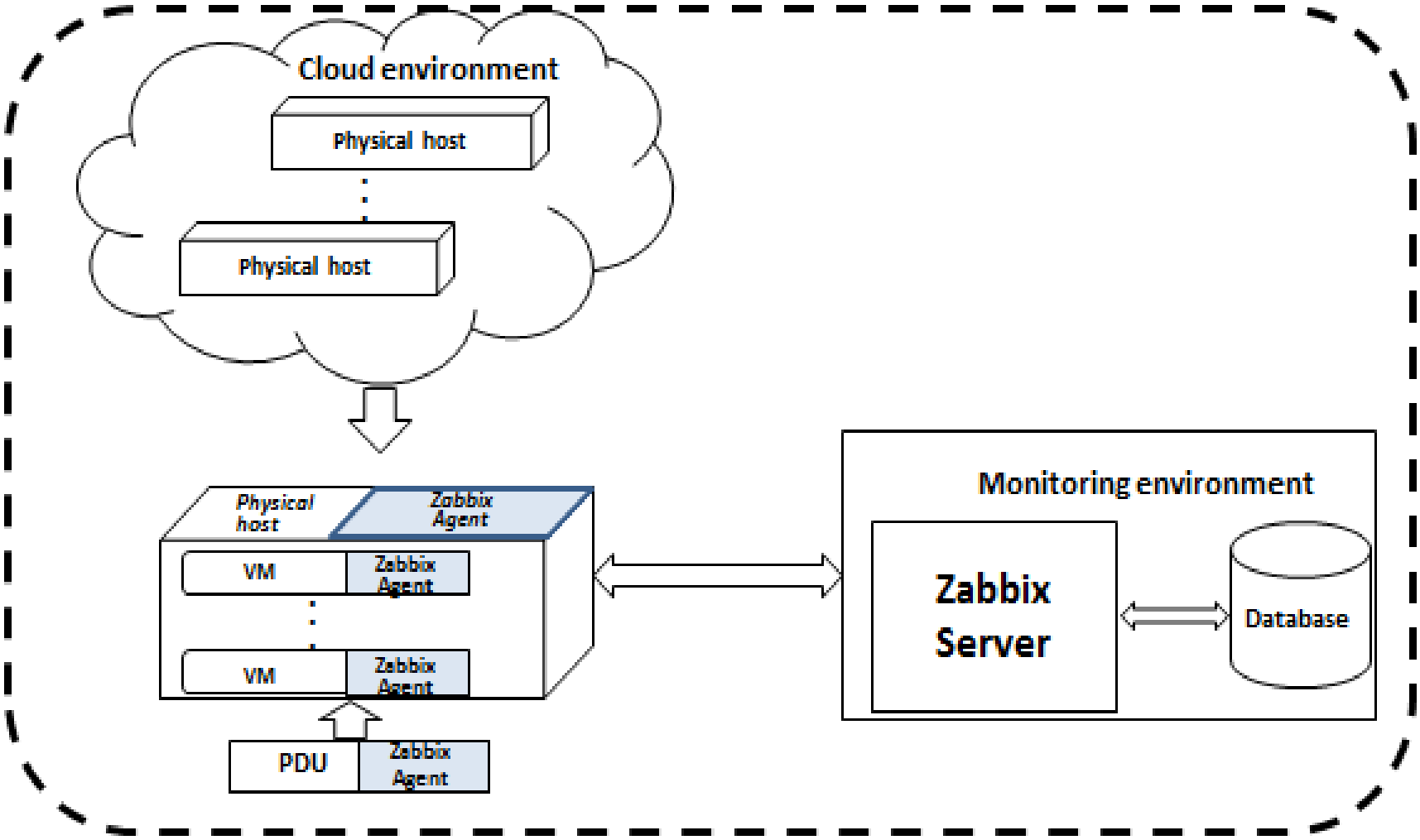
- European project (http://eco2clouds.eu)
- Develop energy efficient solutions for deployment of workloads on Cloud infrastructures



3 Data Centers:
- EPCC - UK
- HLRS - Germany
- INRIA - France

ECO$_2$Clouds architecture

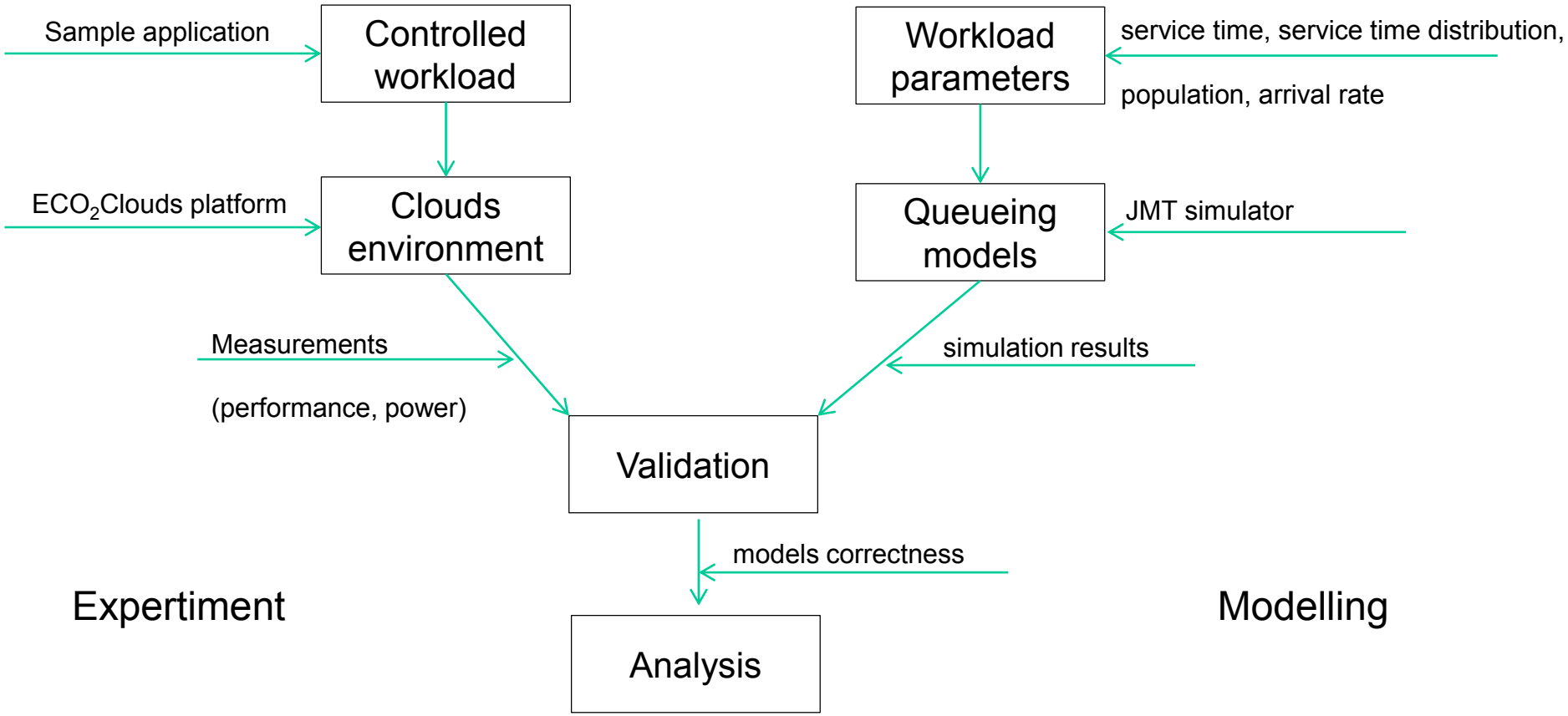
ECO$_2$Clouds project



Eco$_2$Clouds monitoring environment

What is our approach?

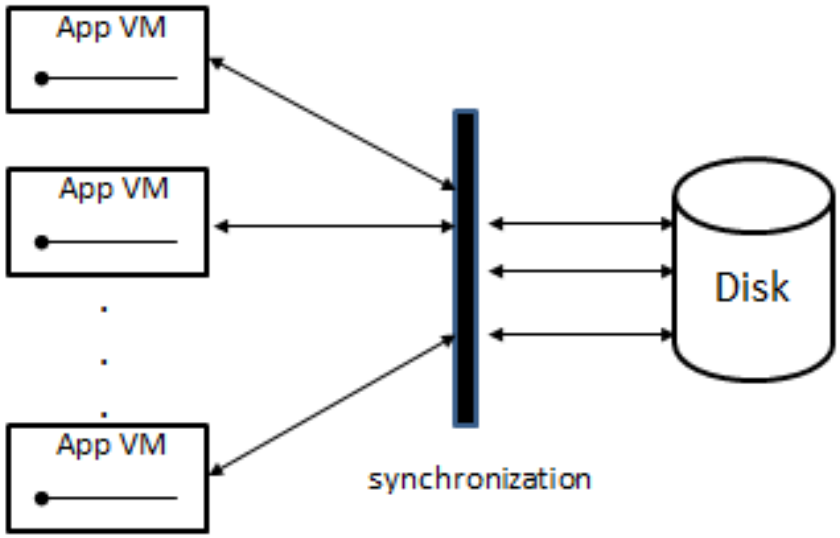
- Investigate different ways to deploy an application in clouds, analyze simultaneously energy consumption and system performances for each deployment configuration



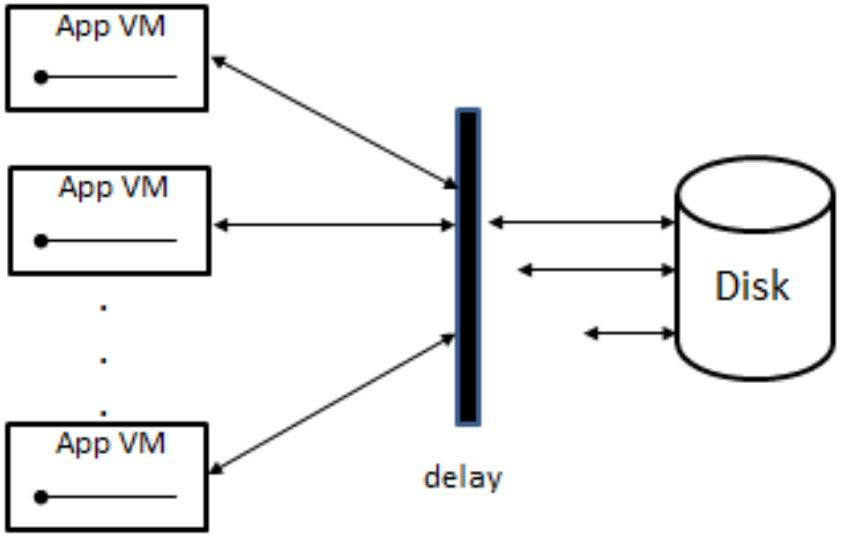
Application profile and experimental platform

- Sample application profile
  - Data loading: 3 mins
  - Data processing: 30 mins
- System characteristics
  - One class workload
  - One bottleneck
  - Bottleneck can migrate depending on number of application instances, or access pattern
- Cloud environment
  - ECO_2Clouds platform, Zabbix monitoring system
- Modeling technique
  - Queueing networks
  - JMT tools

Different deployment strategies



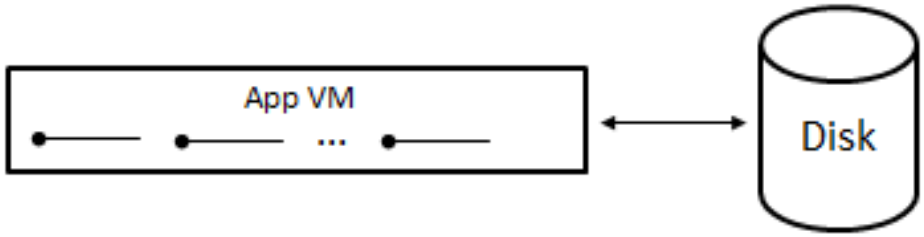
Configuration 1



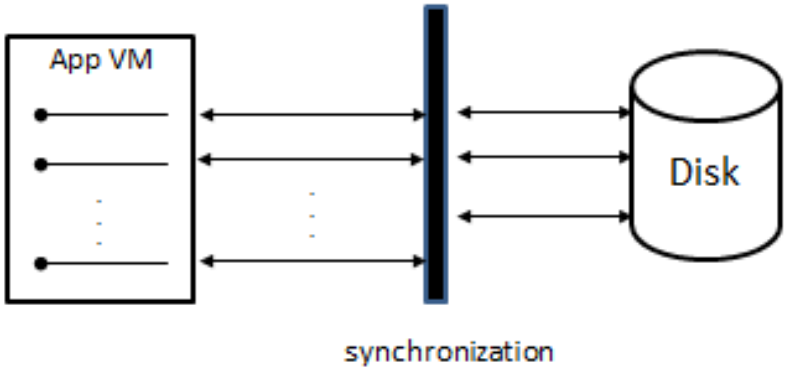
Configuration 2

Synchronous and Asynchronous parallel execution

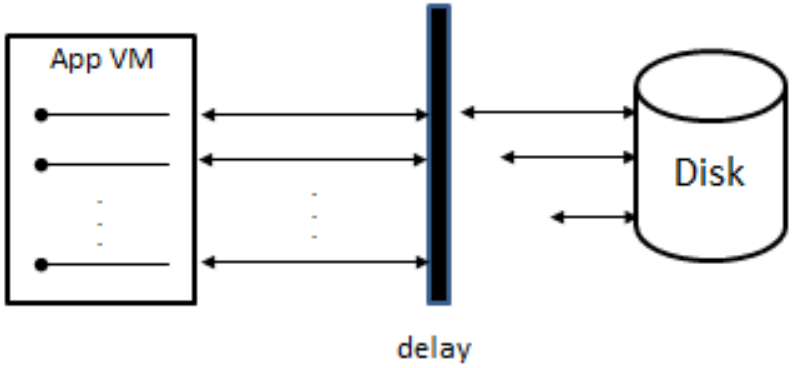
Different deployment strategies



Configuration 3: Sequential execution



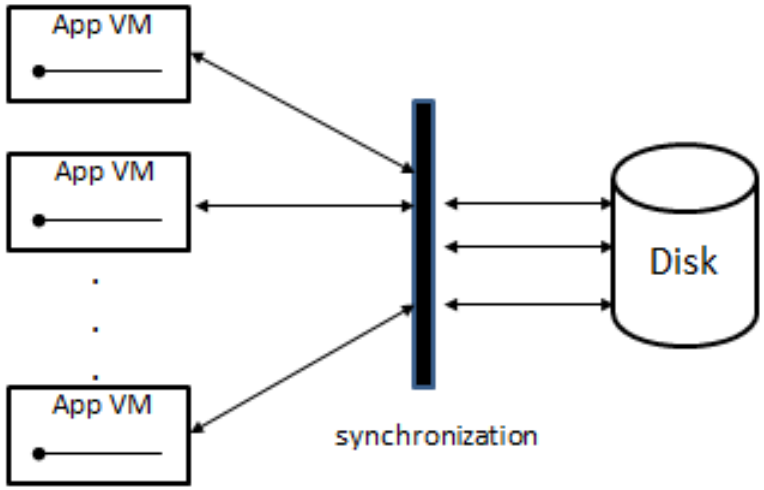
Configuration 4

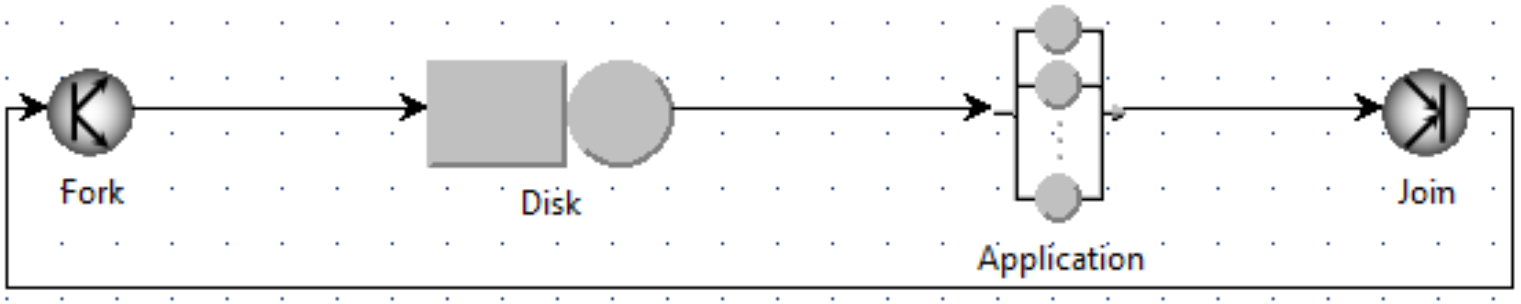


Configuration 5

Synchronous and Asynchronous parallel execution with minimal resources

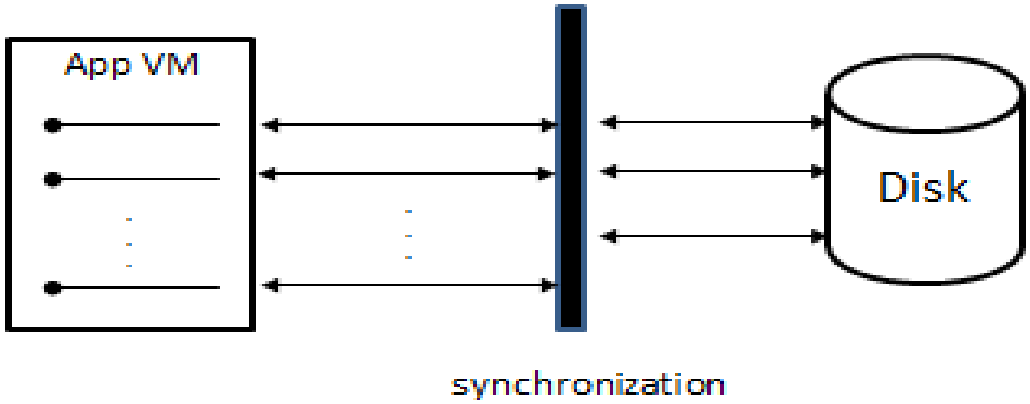
Implemented models using queueing networks





Configuration 1 – Synchronous parallel execution

Implemented models using queueing networks



Configuration 4 – Synchronous parallel execution with minimal resources

Power model

- Simple power model [Fan et al.]:

$$P(u) = P_{idle} + (P_{busy} - P_{idle}) * u \qquad \text{(eq. 1)}$$

- Power model using multiple physical hosts:

$$P(u) = P_{idle} * \#hosts + (P_{busy} - P_{idle}) * u * N \qquad \text{(eq. 2)}$$

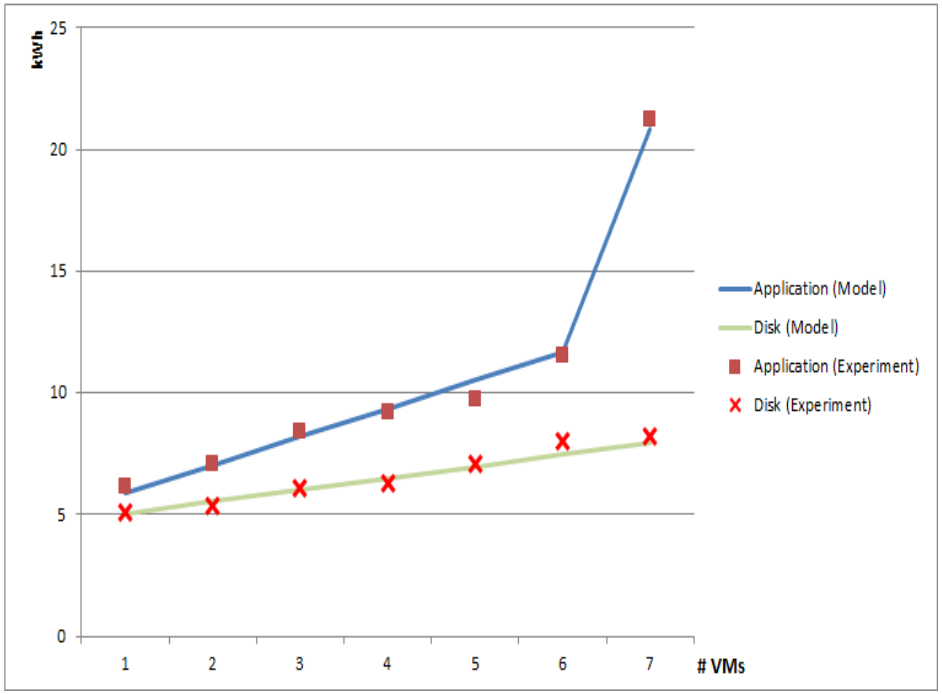
where #hosts = ceil(N/MaxVM)

- Energy model:

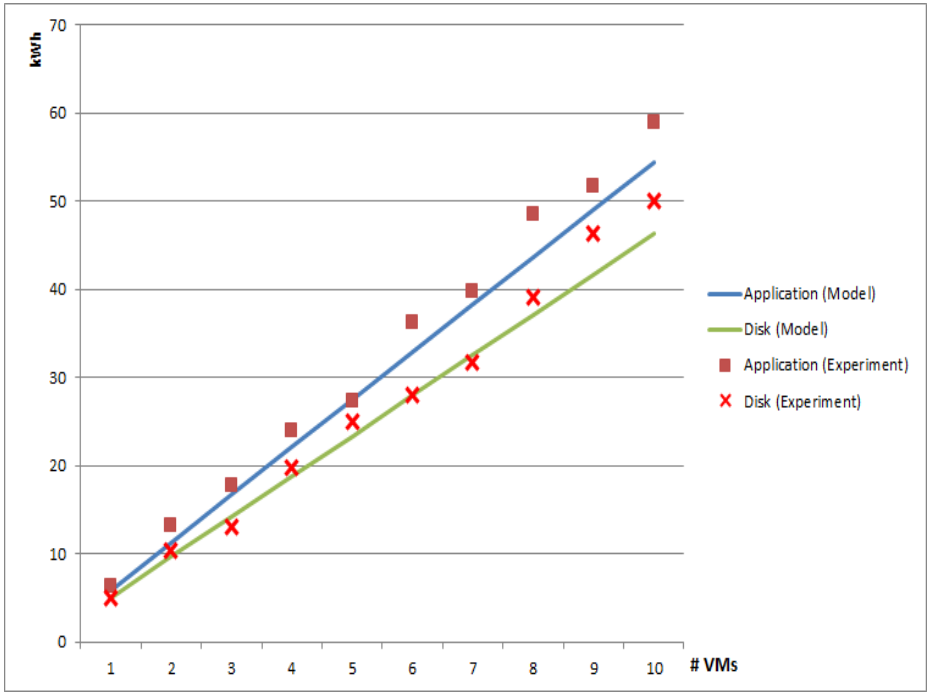
$$E = P(u) * R \qquad \text{(eq. 3)}$$

Validation

- Validate Configuration 1 and Configuration 4

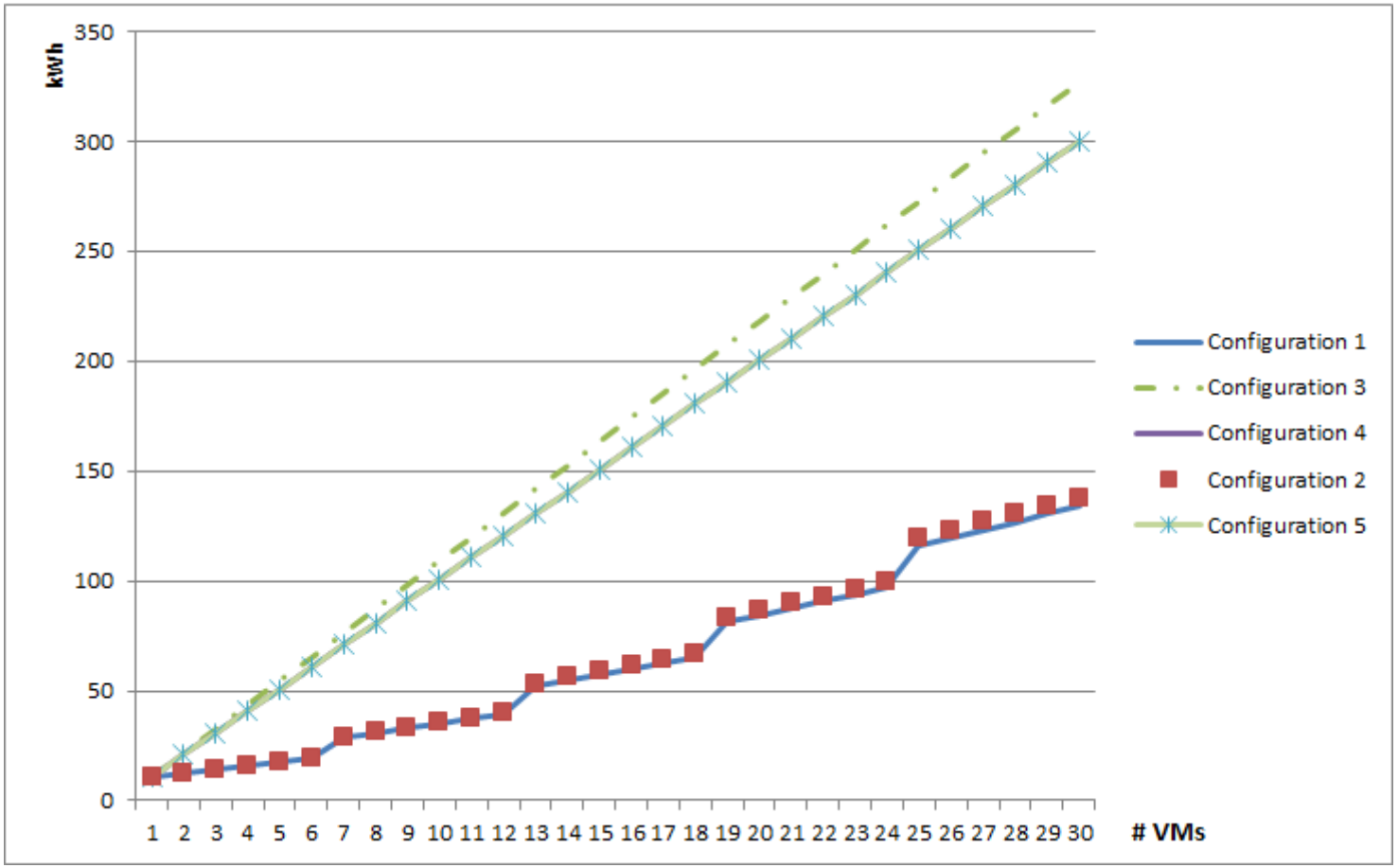


Configuration 1



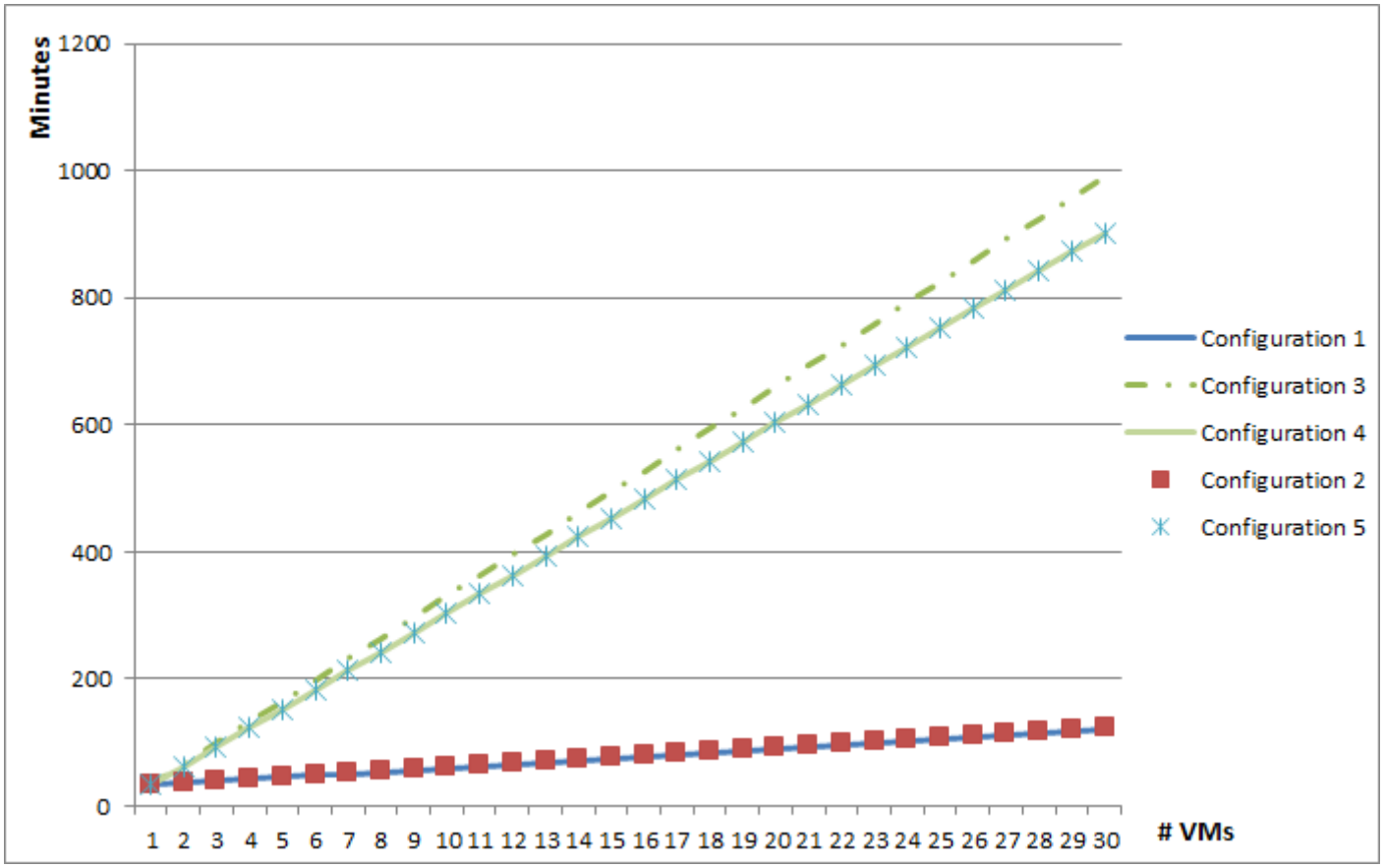
Configuration 4

Further analysis



Energy consumption of each configuration

Further analysis



System response time of each configuration

Exploitation and use of the modeling approach

- Examine different deployment configurations of specific application profile on ECO_2Clouds platform
- Use queueing models to model each configuration
- Validate models correctness
- Use models for predictions and suggest optimal deployment strategy

Future work

- Use the work at different scales (application instances, task instances)
- Extend to other types of application such as web services
- Extend to two-classes workload and find optimal mixed workload considering saving energy consumption
- Extend the work to consider adaptation at runtime

Thank you
Q & A

References

1. Global e-Sustainability Initiative (GeSI). SMART 2020: Enabling the Low Carbon Economy in the Information Age. 2008

2. Saurabh Kumar Garg and Rajkumar Buyya: Green Cloud Computing and Environmental Sustainability, in Harnessing Green IT: Principles and Practices, 315-340 pp, S. Murugesan and G. Gangadharan (eds), Wiley Press, UK, October 2012.

3. Mayo, R. N. and Ranganathan P., 2005. Energy Consumption in Mobile Devices: Why Future Systems Need Requirements-Aware Energy Scale-Down. Proceedings of 3rd International Workshop on Power-Aware Computer Systems, San Diego, CA, USA.

4. M. Vitali and B. Pernici: A Survey on Energy Eciency in Information Systems, Journal on Cooperative Information Systems, March 2014, http://www.worldscientic.com/doi/abs/10.1142/S0218843014500014

5. P. Melia, M. Schiavina, M. Gatto, L. Bonaventura, S. Masina, R. Casagrande: Integrating Field Data into Individual-based Models of the Migration of European Eel Larvae. Marine Ecology Progress Series. Vol. 487: 135149, 2013

6. Anton Beloglazov, Rajkumar Buyya, Young Choon Lee, and Albert Zomaya: Taxonomy and Survey of Energy-Efficient Data Centers and Cloud Computing Systems, Advances in Computers, Volume 82, Vol. 2, 47-111 pp, Elsevier, Amsterdam, The Netherlands, March 2011.

7. Nowak, A., Leymann, F., Schleicher, D., Schumm, D., Wagner, S.: Green Business Process Patterns. In: Proceedings of the 18th Conference on Pattern Languages of Programs, ACM (2011)

8. Ying Song, Yuzhong Sun, Weisong Shi: A Two-Tiered On-Demand Resource Allocation Mechanism for VM-Based Data Centers, IEEE Transactions on Services Computing, Vol. 6:1, pp. 116-129, 2013

9. Xiaobo Fan, Wolf-Dietrich Weber, Luiz Andre Barroso: Power Provisioning for a Warehouse-sized Computer. In Proceedings of the ACM International Symposium on Computer Architecture, San Diego, CA, June 2007

10. Cinzia Cappiello, Sumit Datre, Maria Grazia Fugini, Paco Melia, Barbara Pernici, Pierluigi Plebani, Michael Gienger, Axel Tenschert: Monitoring and Assessing Energy Consumption and CO2 Emissions in Cloud-based Systems. Proc. IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2013

11. M. Bertoli, G. Casale, G. Serazzi: JMT: Performance Engineering Tools for System Modeling. ACM SIGMETRICS Performance Evaluation Review, Volume 36 Issue 4, New York, US, March 2009, 10-15, ACM press.

12. B. Pernici and U. Wajid, Assessment of the Environmental Impact of Applications in Federated Clouds. SmartGreens 2014, Barcelona, April 2014

Implemented models using queueing networks

- Input params

  N=1,

  $D_{storage}$= 3 mins

  D_{app} = 30 mins

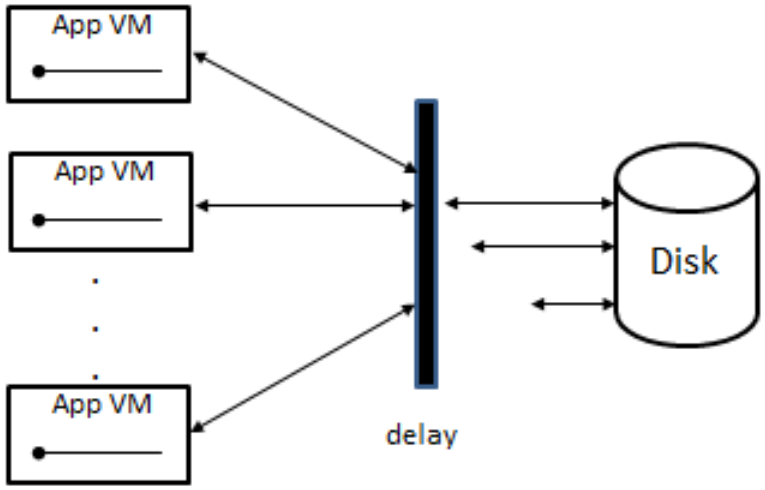
- Performance indices

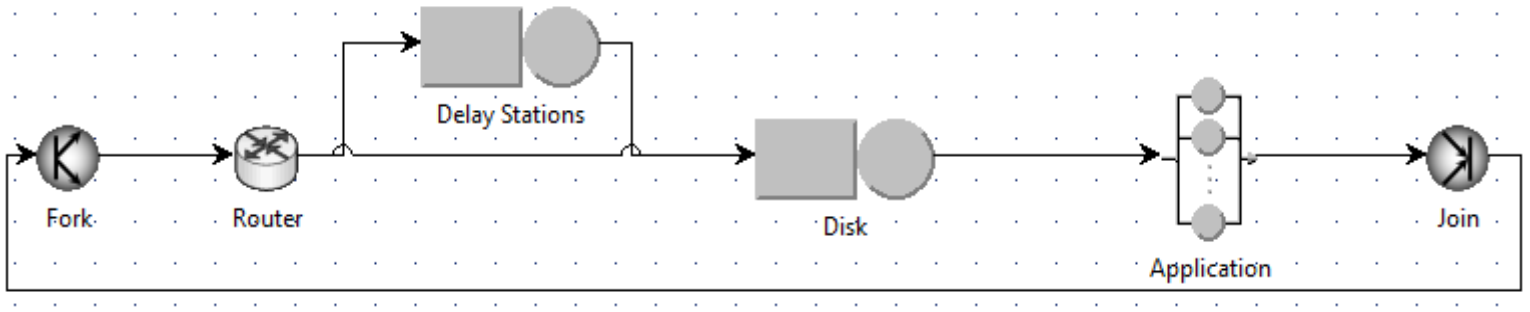
  $U_{storage}$ = 3/(3+30)= 0,091

  U_{app} = 30/(3+30)= 0,909

  R = 3 + 30 = 33 mins

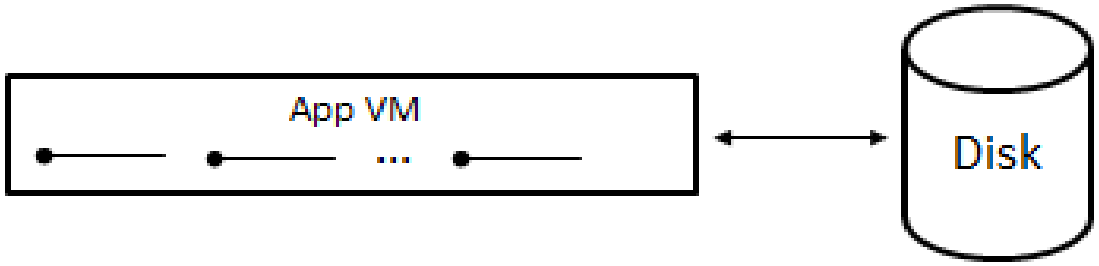
Implemented models using queueing networks

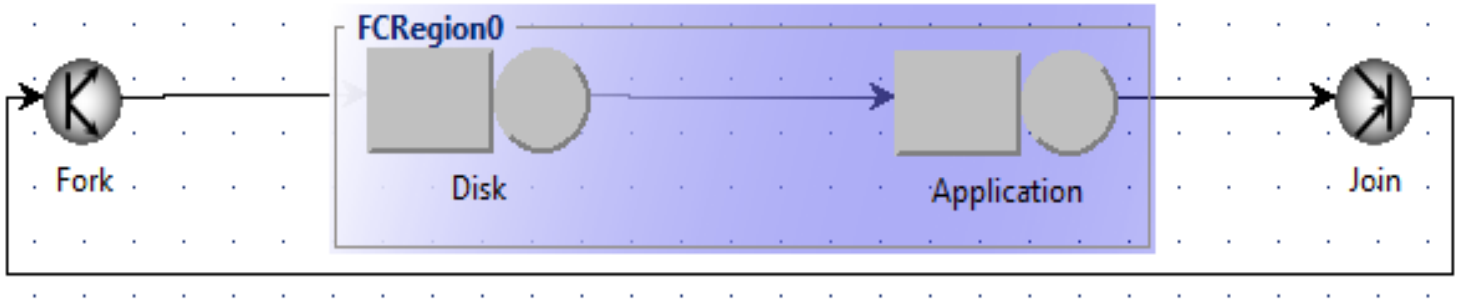




Configuration 2 – Asynchronous parallel execution

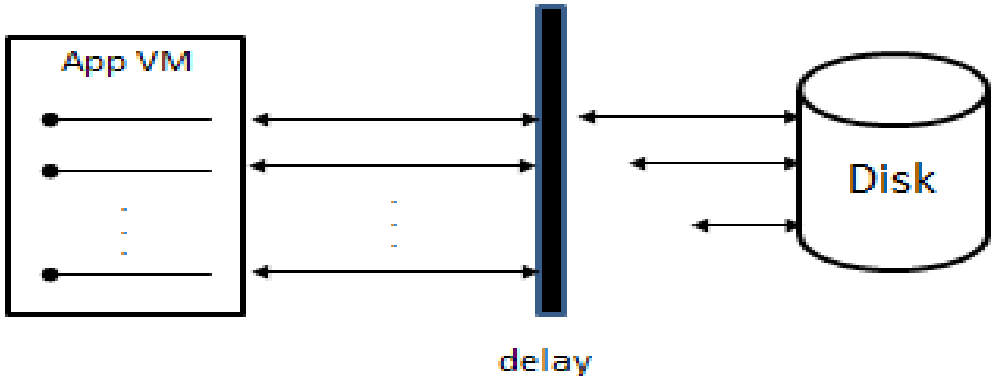
Implemented models using queueing networks

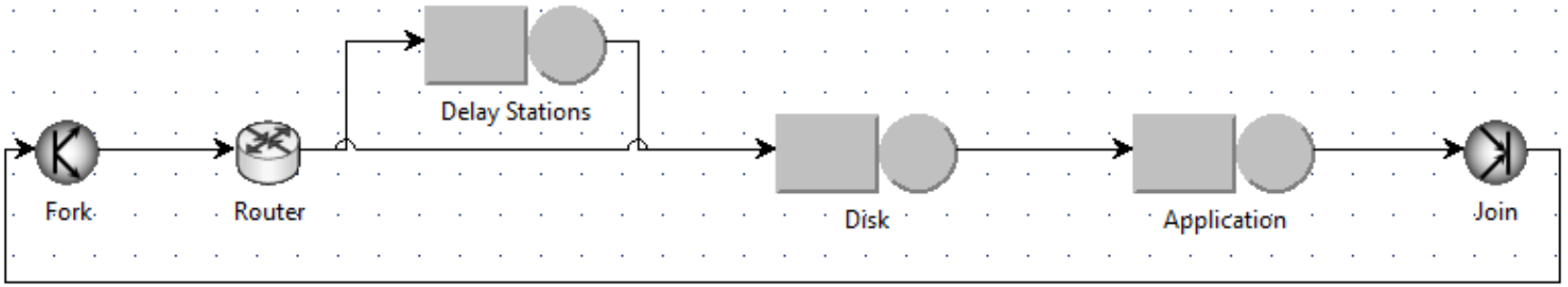




Configuration 3 – sequential execution

Implemented models using queueing networks

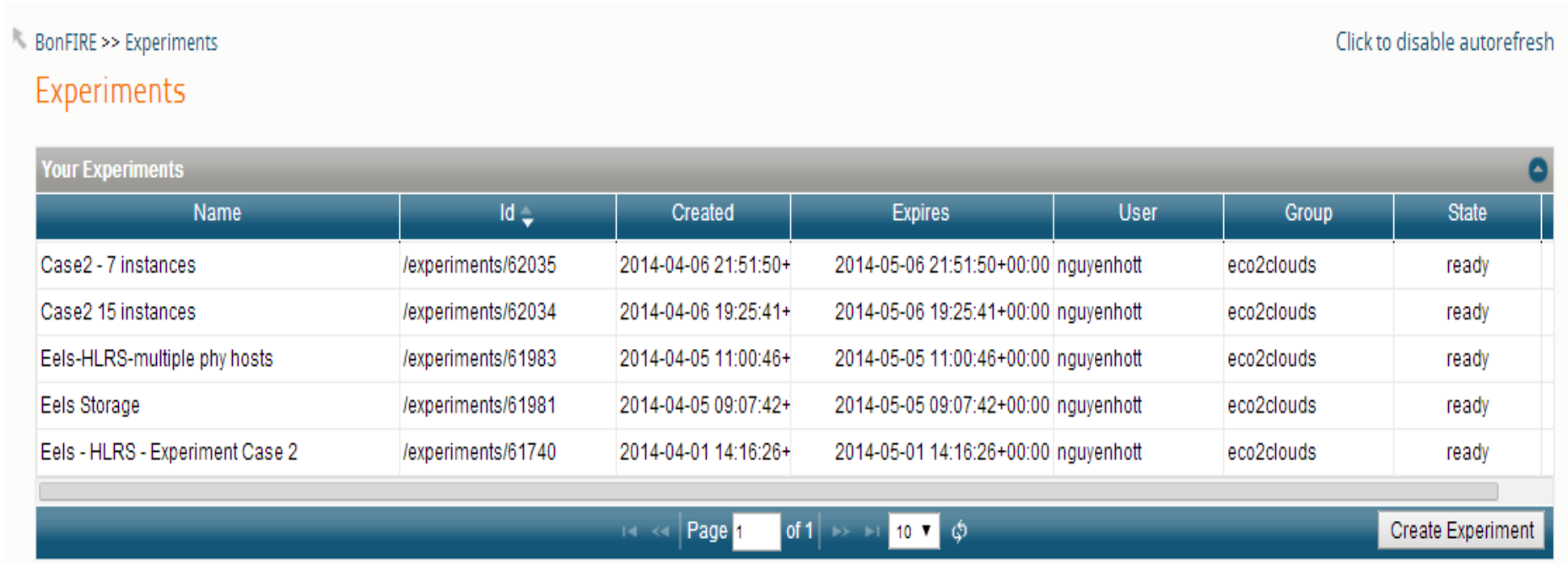




Configuration 5 – Asynchronous parallel execution with minimal resources

Experiments

- Infrastructure configurations
  - Site: HLRS
  - Physical node: 2 x QuadCore Intel Xeon @ 2.83 GHz, 32 GB RAM
  - Storage VM: Medium size (CPU = 1; Mem = 2048 MB)
  - App VM: Custom (CPU = 1; Mem = 4096 MB)



Experiments

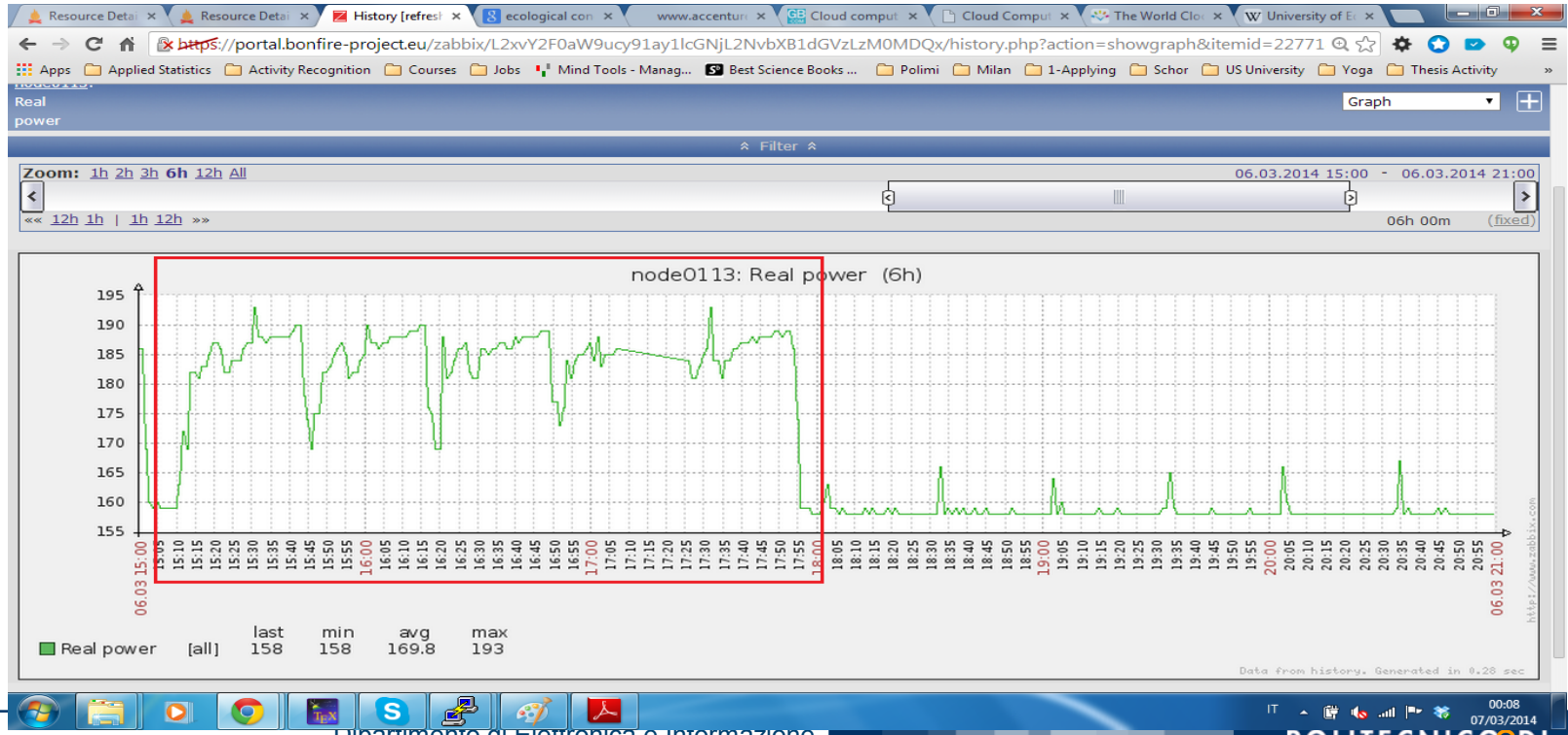
- Modify the Eels application
  - Allow 3 different running modes: simutaneous, delay and sequential
  - Data are loaded into different folders
  - Allow writing logs to record time to load data and time to execute the application

How many experiments?

- Two different configurations
  - Configuration 1 and 4
- 1 physical host
  - 6 different experiments with #VMs = 1, ..., 6
- Multiple physical hosts
  - #VMs = 7, 12, 15

Experiments – Monitoring power

- Import energy templates
- Collect power measures (of the application and storage) between the execution period of the application



Experiments – Problems that I encountered

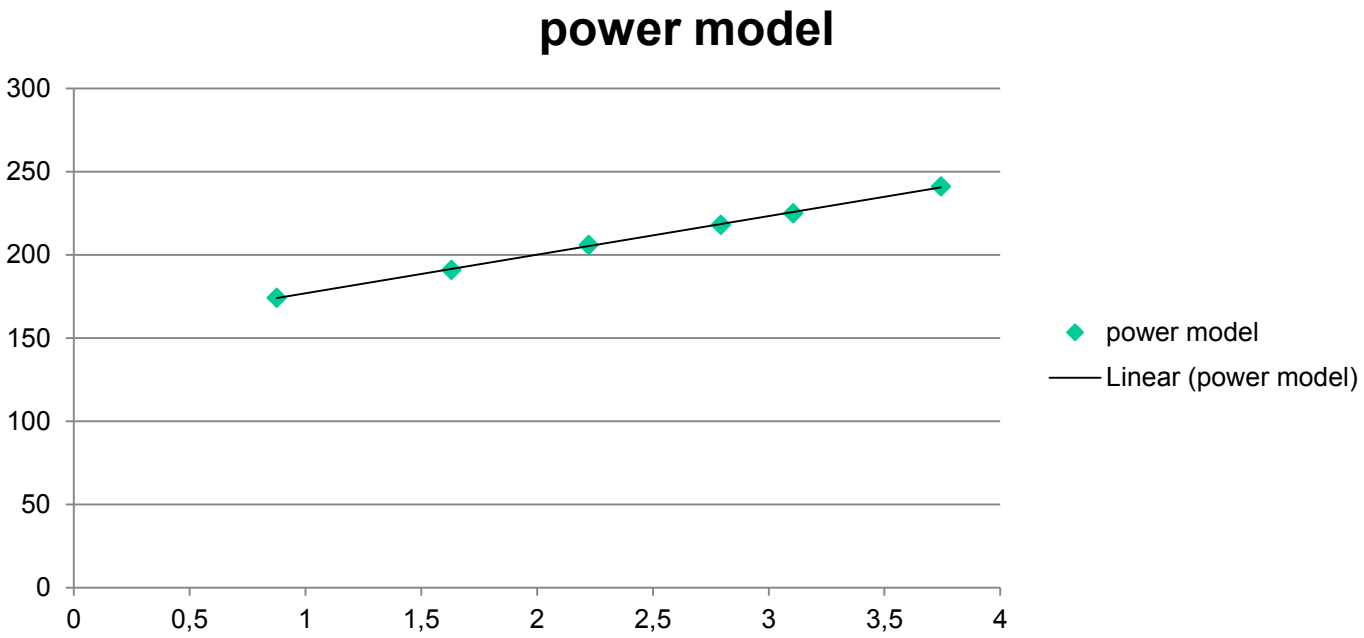
- Modify the Eels applications
- Prepare running environment on HLRS: VM images, Oceanographic data
- Understand different parameters in Zabbix monitoring system
- Unstable running environment when updates occur during the experiments

Power model – identify parameters

$$P(u) = P_{idle} + (P_{busy} - P_{idle}) * u * N$$

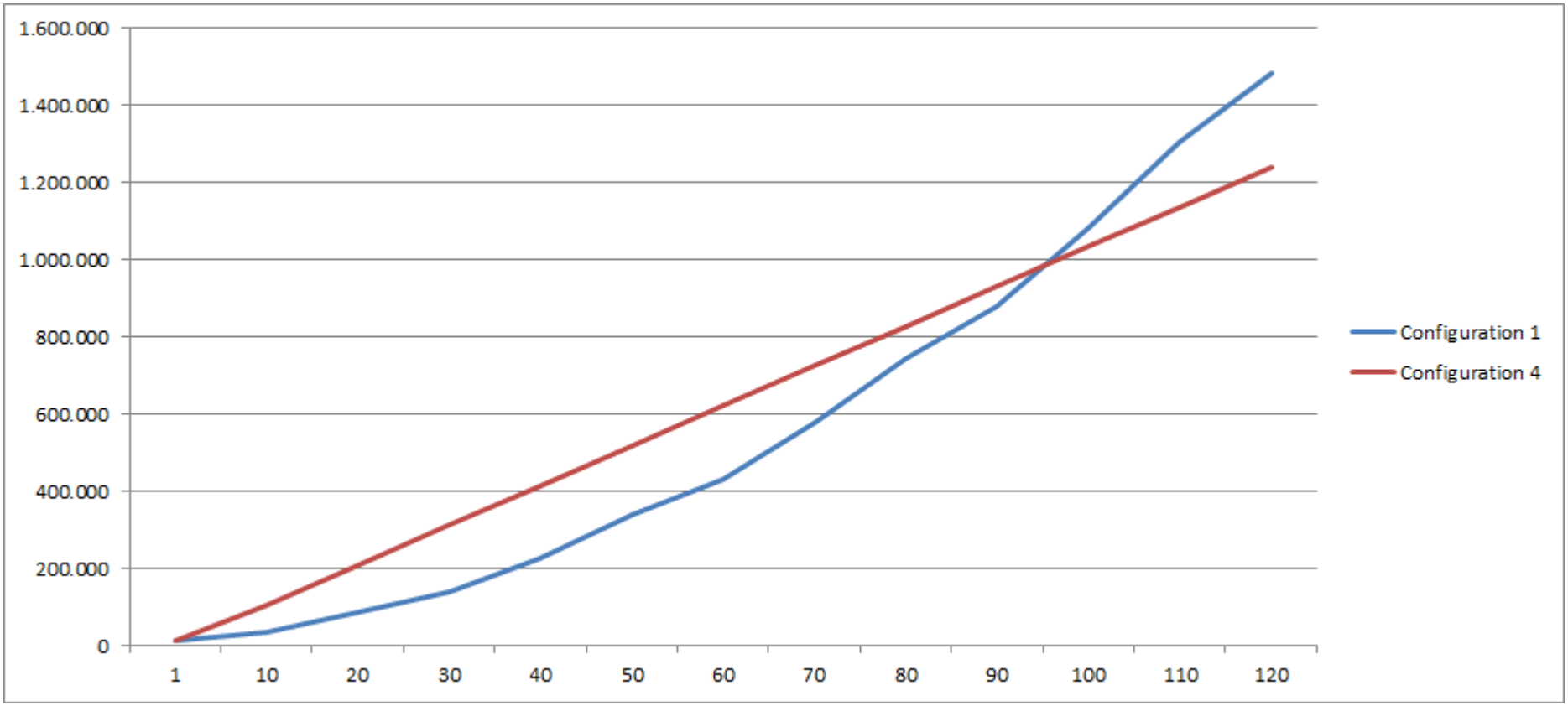
	VM	Mean CPU User Use	Ref. Mean CPU	Mean Power	U x #VMs
1 instance	1	0,876335307	0,909	174,3529412	0,876335307
2 instances	2	0,815200795	0,831	191,0924855	1,630401591
3 instances	3	0,741120261	0,767666667	205,9794872	2,223360782
4 instances	4	0,698368315	0,71325	218,1512195	2,793473258
5 instances	5	0,621102453	0,6658	225,0610329	3,105512265
6 instances	6	0,624	0,624	241,1578947	3,744

Power model – identify parameters



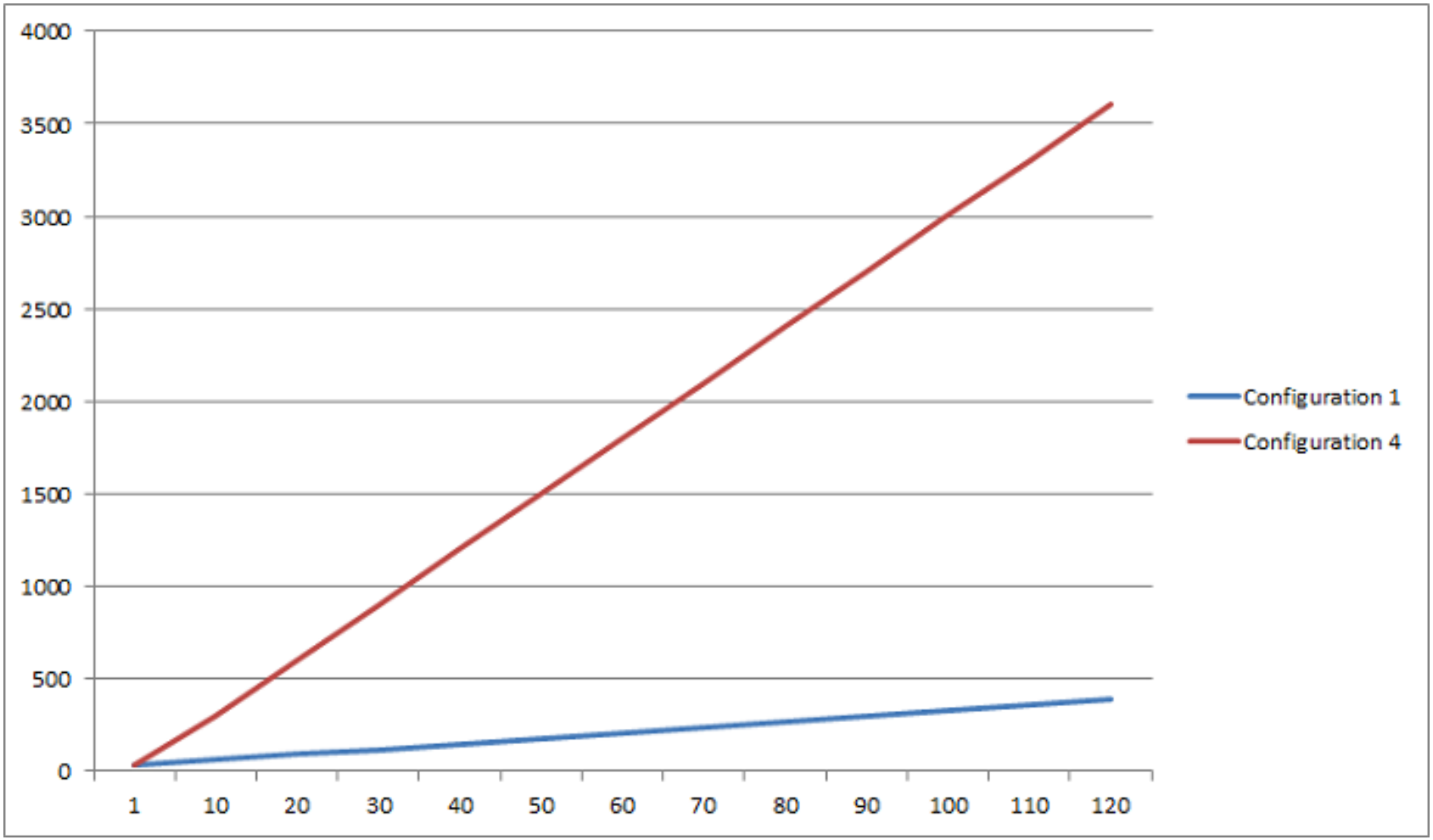
Slope	23,18098801
Intercept	153,7687986

Exploitation



Switching energy consumption

Exploitation



System response time