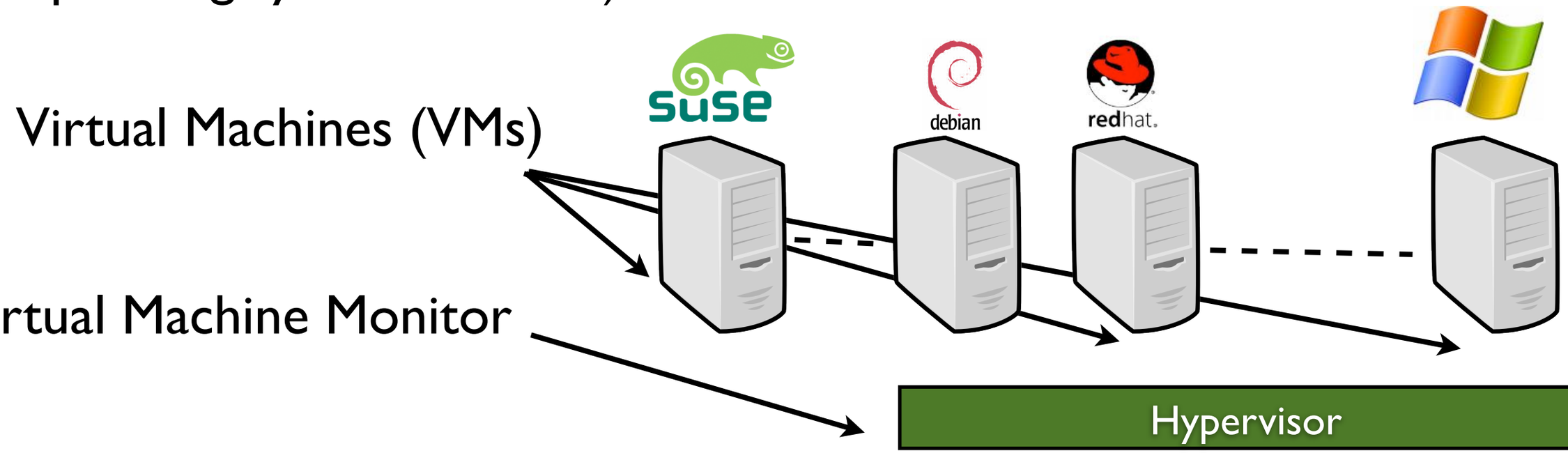
Large Scale Management of Virtual Machines

Cooperative and Reactive Scheduling in Large-Scale Virtualized Platforms

Adrien Lèbre
EPI ASCOLA / HEMERA
Flavien Quesnel, Phd Candidate
February 2013

System Virtualization

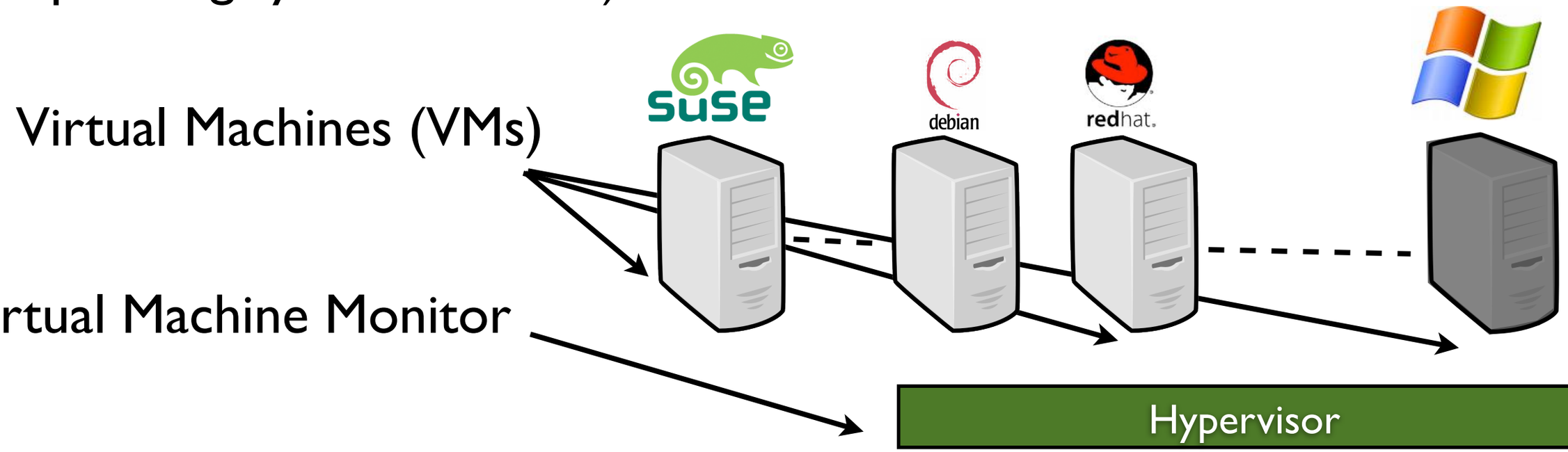
- One to multiple OSes on a physical node thanks to a hypervisor (an operating system of OSes)



"A ***virtual machine*** (VM) provides a faithful implementation of a physical processor's hardware running in a protected and isolated environment.
Virtual machines are created by a software layer called the ***virtual machine monitor*** (VMM) that runs as a privileged task on a physical processor."

System Virtualization

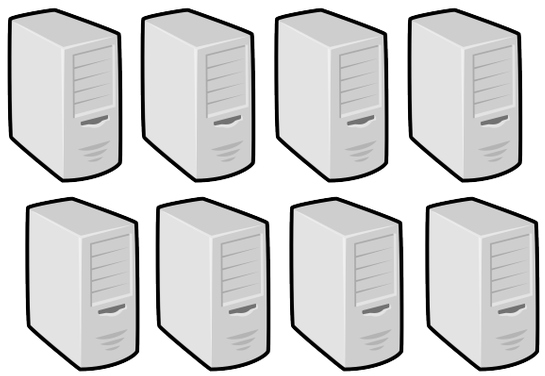
- One to multiple OSes on a physical node thanks to a hypervisor (an operating system of OSes)



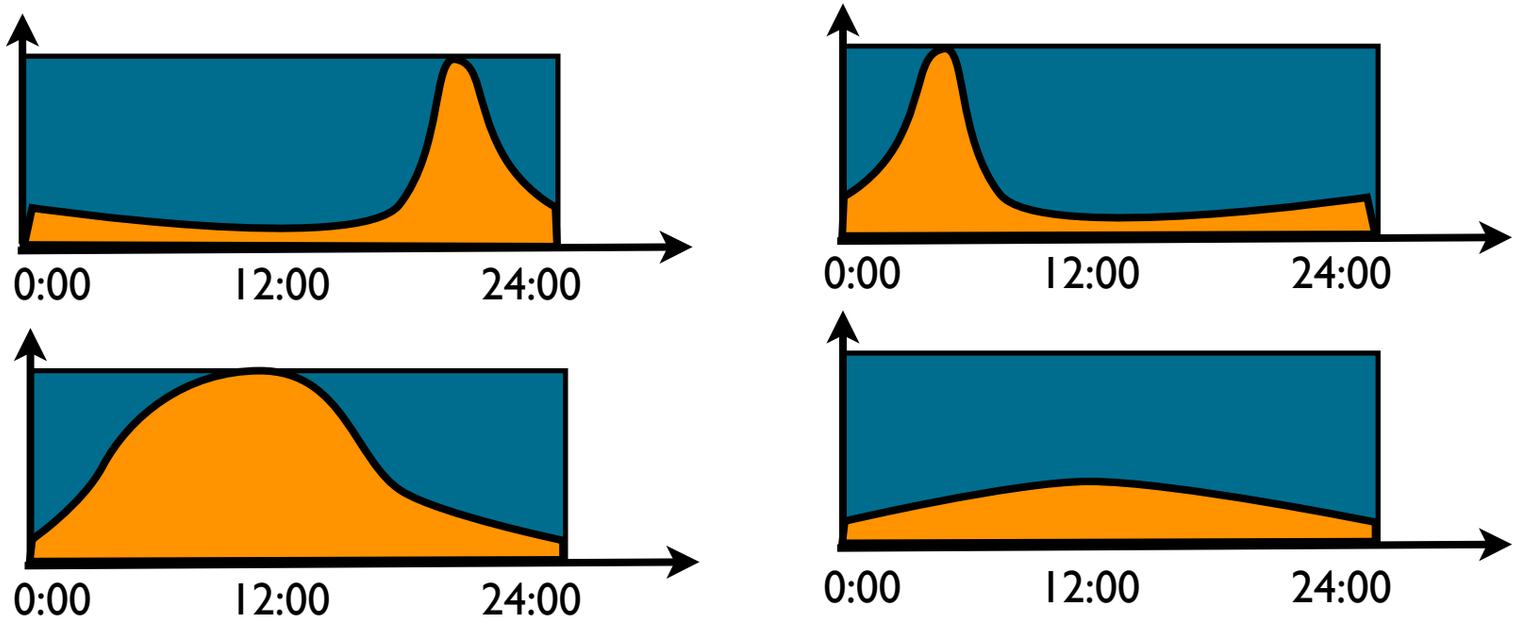
"A ***virtual machine*** (VM) provides a faithful implementation of a physical processor's hardware running in a protected and isolated environment.
Virtual machines are created by a software layer called the ***virtual machine monitor*** (VMM) that runs as a privileged task on a physical processor."

Context

- Efficient use of resources in data centers

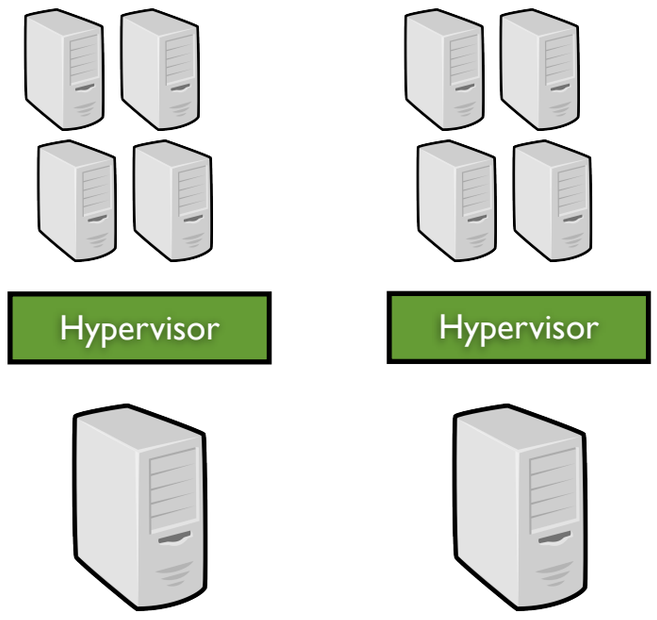


Physical Machines



Virtual Machines

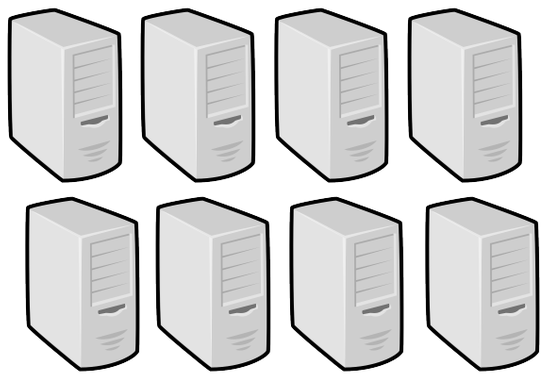
Case 1: Static Consolidation



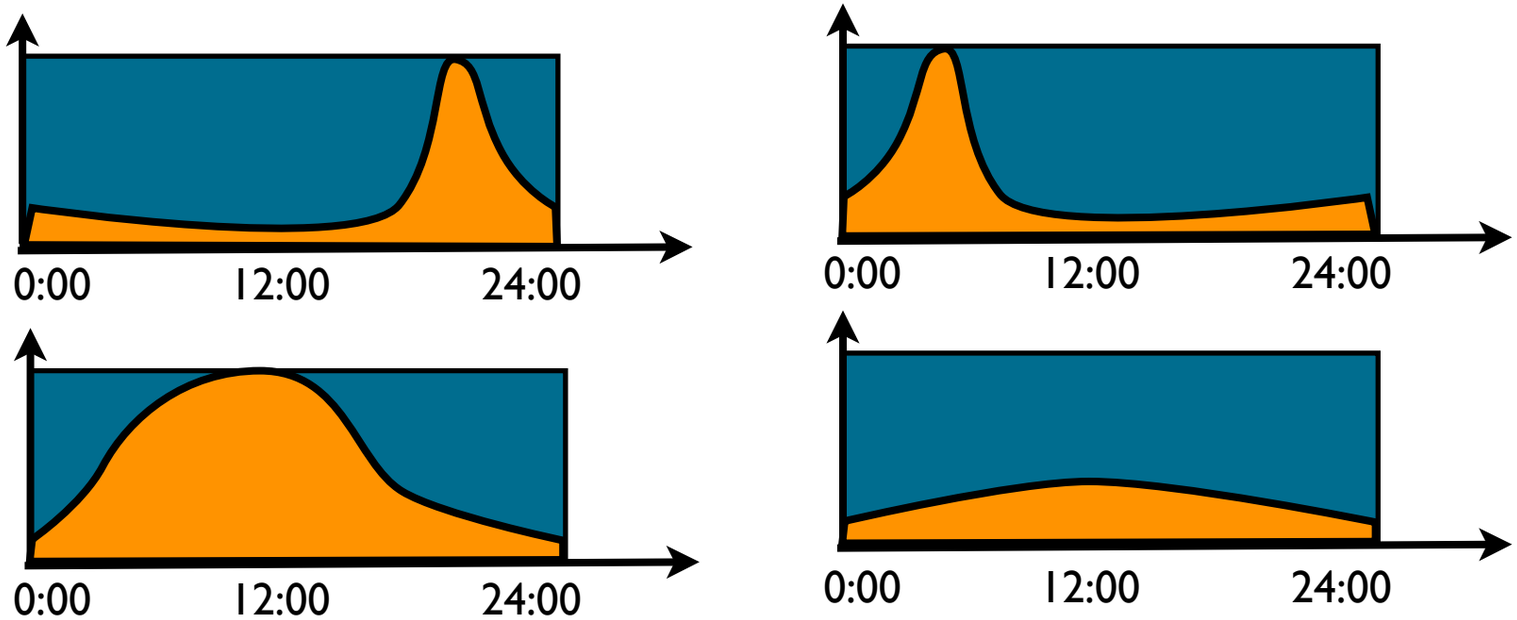
Physical Machines

Context

- Efficient use of resources in data centers

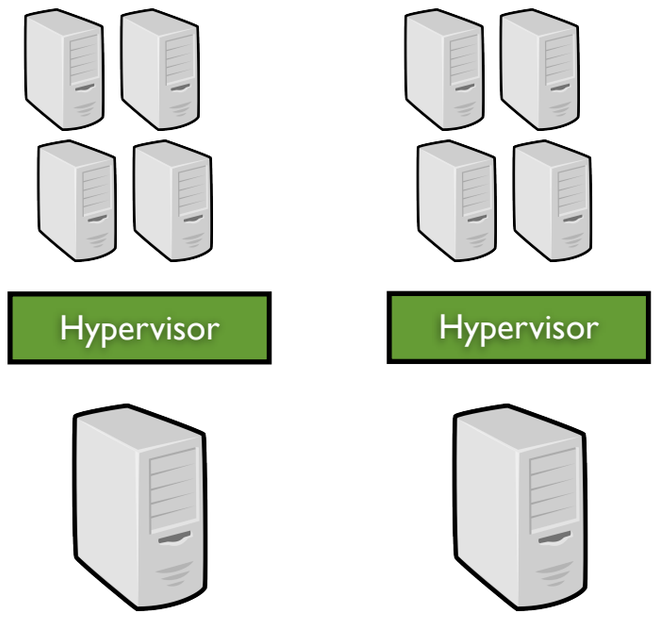


Physical Machines



Virtual Machines

Case 1: Static Consolidation



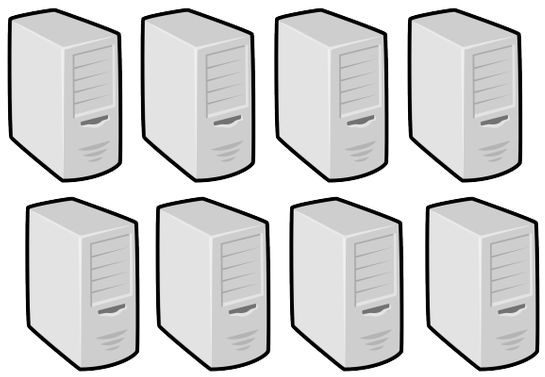
Physical Machines

⇒ Resources may not be fully used

Context

- Efficient use of resources data centers

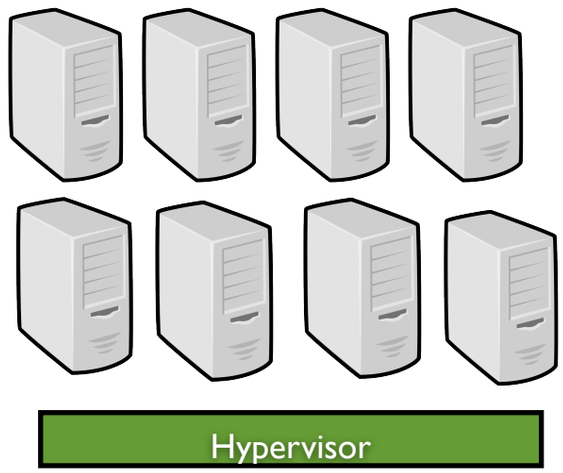
  ⇒ Schedule VMs according to their effective usage



Physical Machines

Case 2: Dynamic Consolidation

Virtual Machines

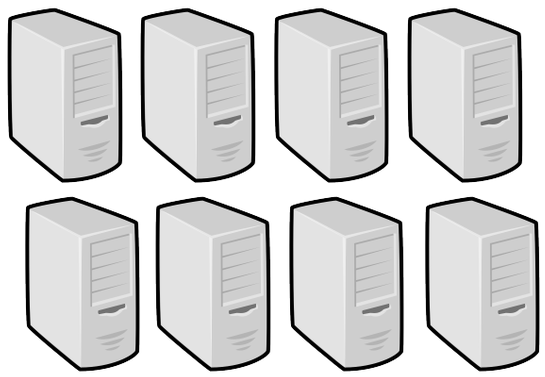


Physical Machine

Context

- Efficient use of resources data centers

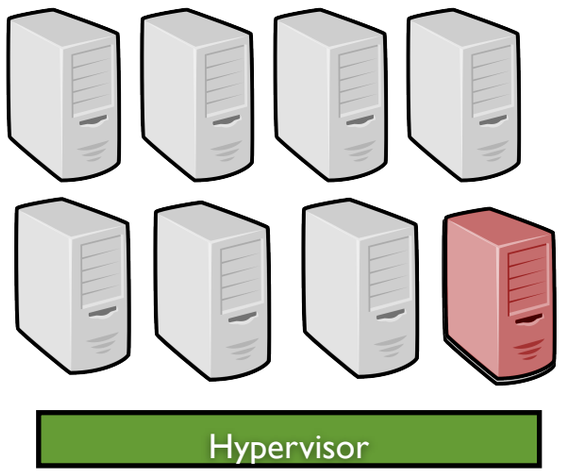
  ⇒ Schedule VMs according to their effective usage



Physical Machines

Case 2: Dynamic Consolidation

Virtual Machines

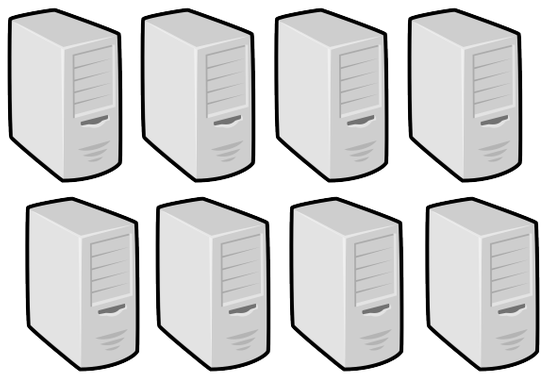


Physical Machine

Context

- Efficient use of resources data centers

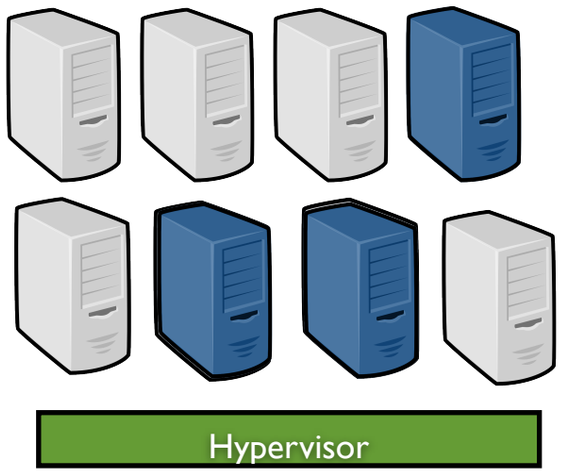
  ⇒ Schedule VMs according to their effective usage



Physical Machines

Case 2: Dynamic Consolidation

Virtual Machines

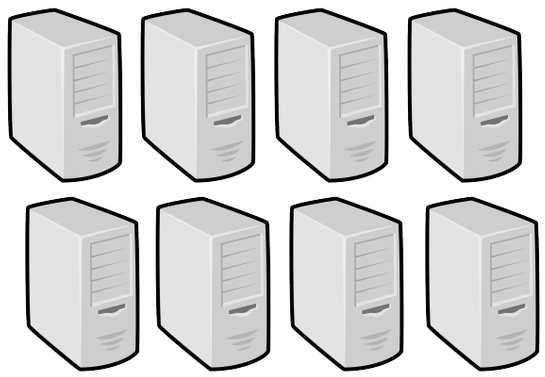


Physical Machine

Context

- Efficient use of resources data centers

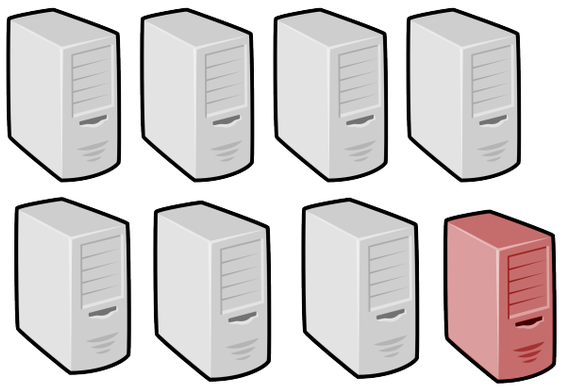
  ⇒ Schedule VMs according to their effective usage



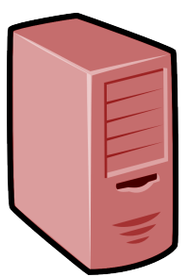
Physical Machines

Case 2: Dynamic Consolidation

Virtual Machines



Hypervisor



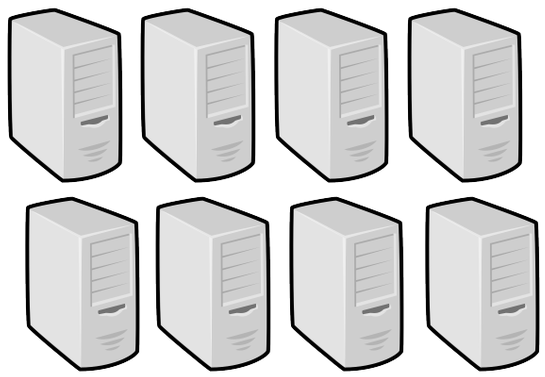
Physical Machine

Hypervisor

Context

- Efficient use of resources data centers

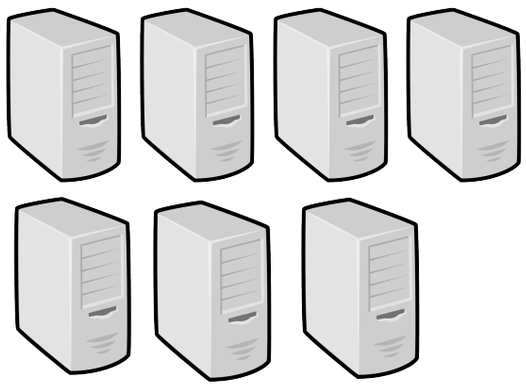
  ⇒ Schedule VMs according to their effective usage



Physical Machines

Case 2: Dynamic Consolidation

Virtual Machines



Hypervisor

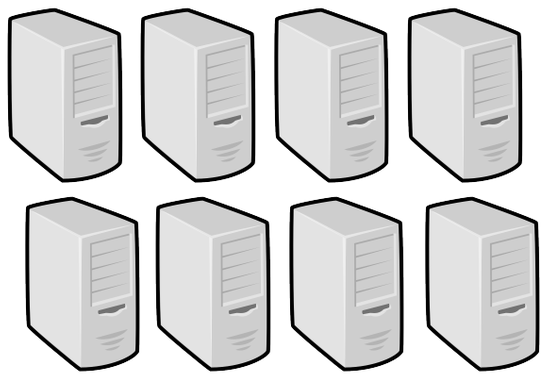
Physical Machine

Hypervisor

Context

- Efficient use of resources data centers

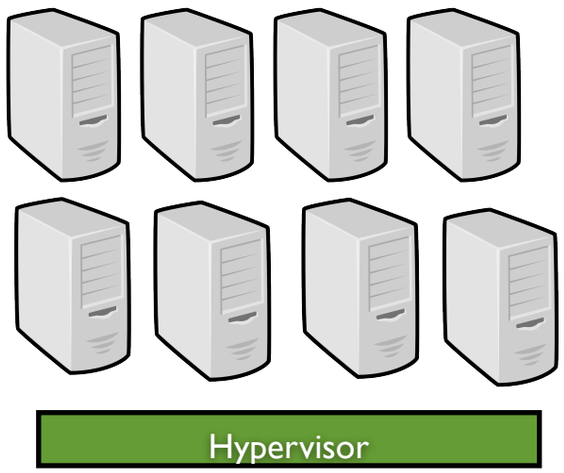
  ⇒ Schedule VMs according to their effective usage



Physical Machines

Case 2: Dynamic Consolidation

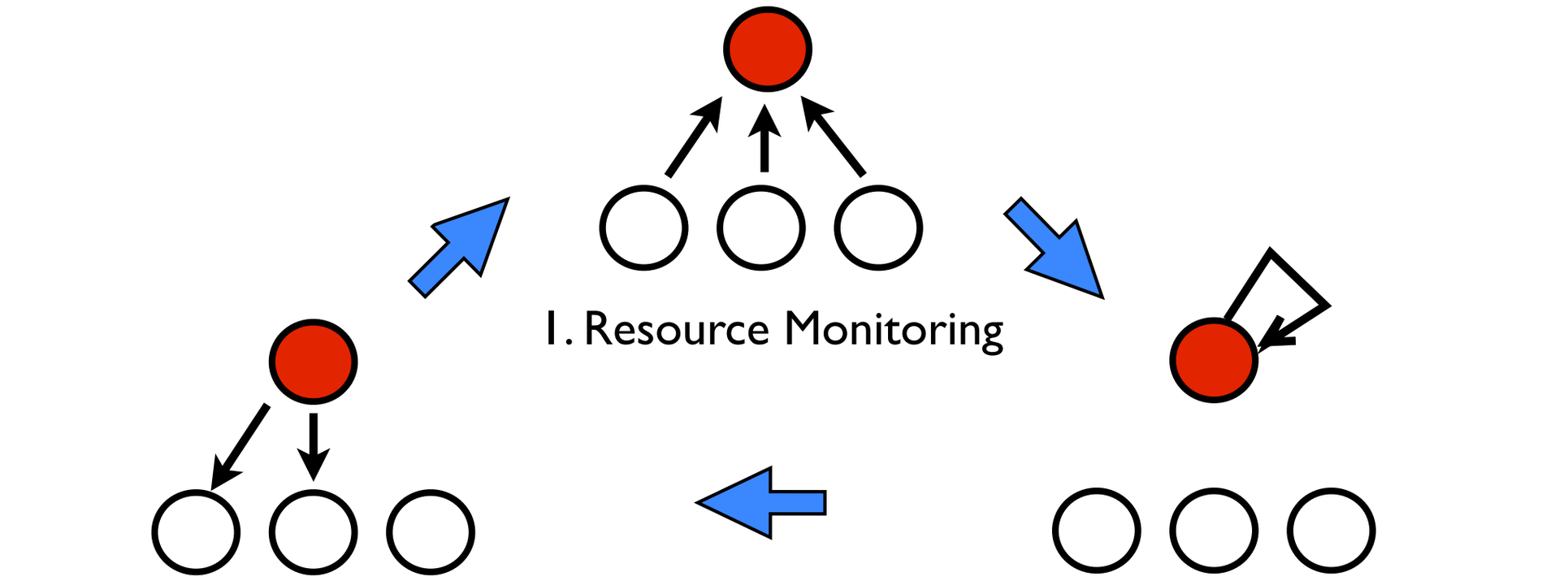
Virtual Machines



Physical Machine

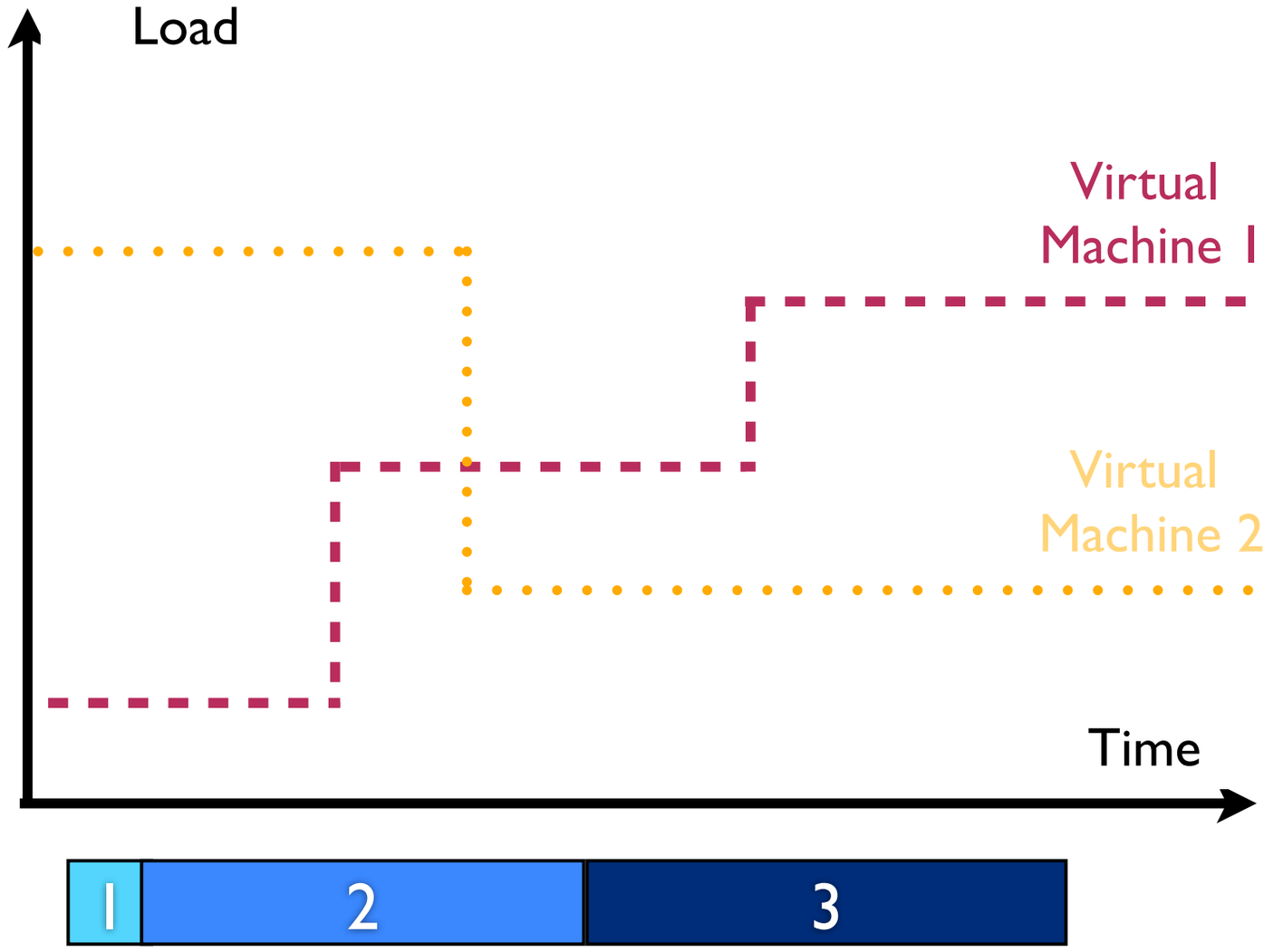
First solutions...

- Centralized approaches



First solutions...

- Centralized approaches

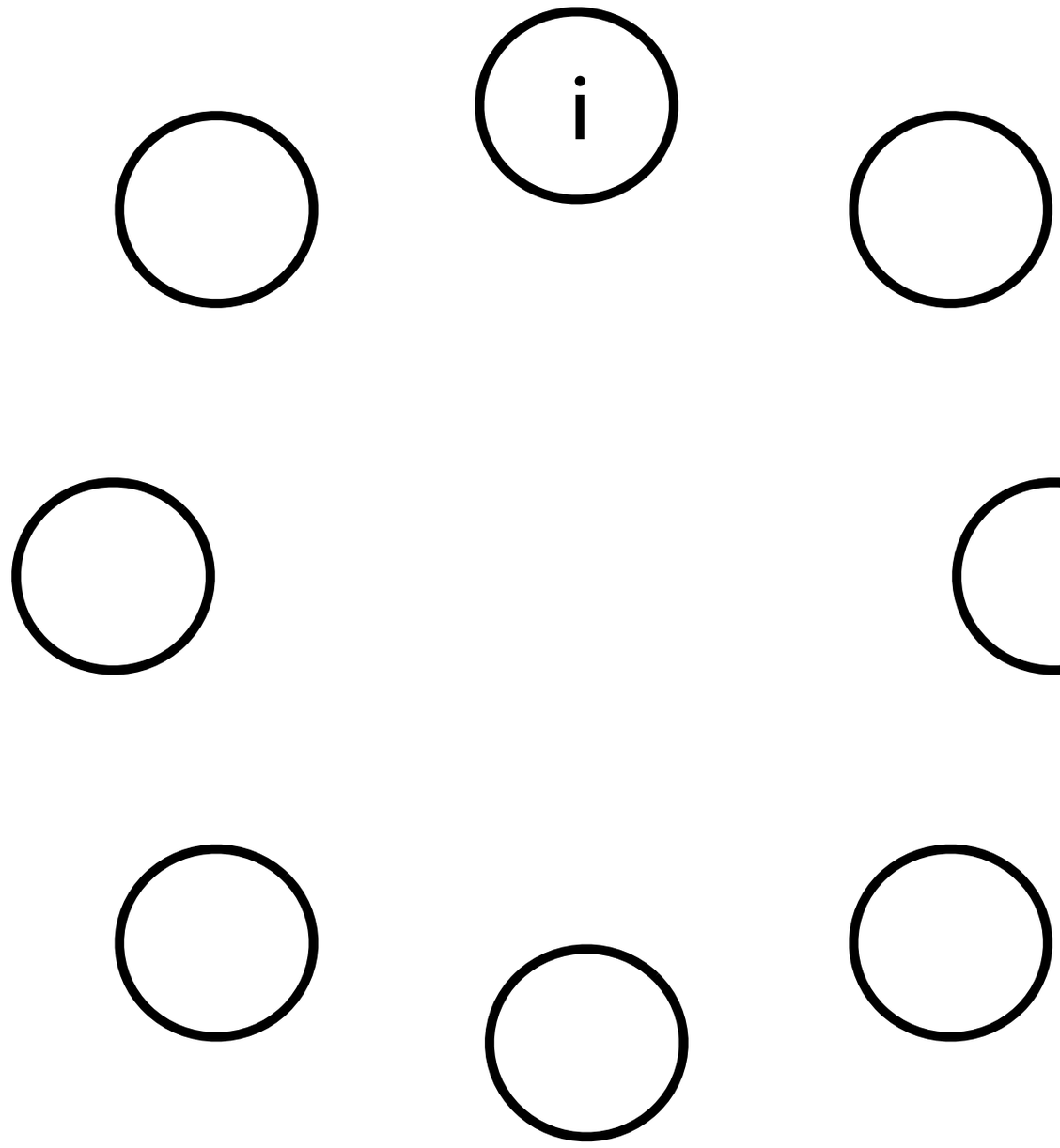


⇒ Scalability/Reactivity concerns

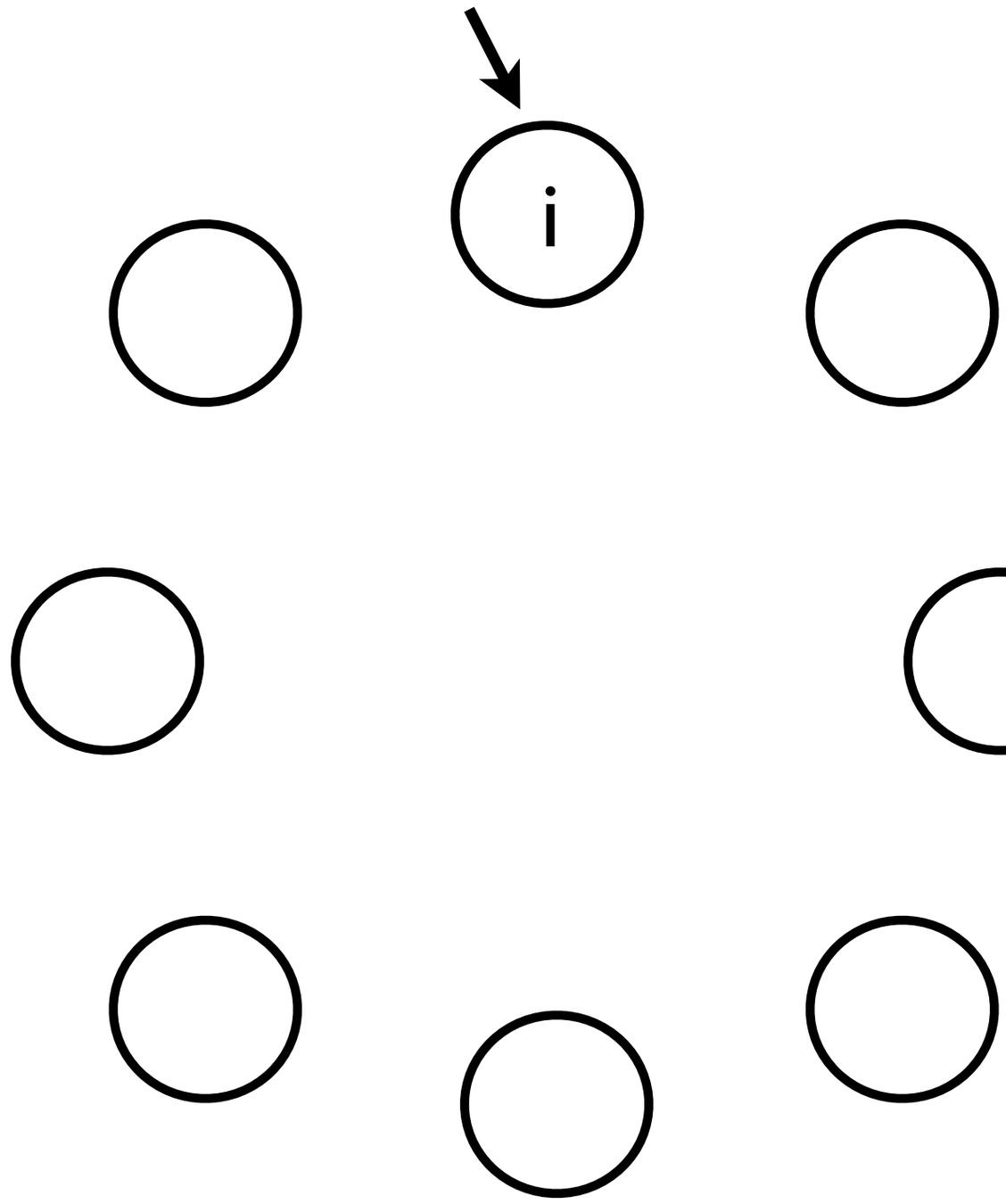
Proposal overview

- Main characteristics
  - Event driven
  - Peer to peer, no service node
  - Local interactions between nodes
    - Monitoring
    - Scheduling

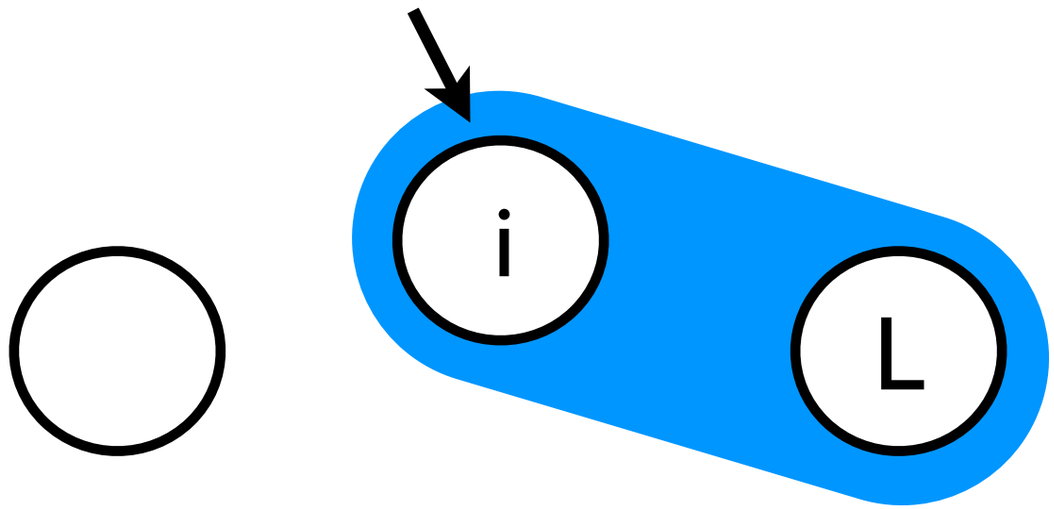
DVMS Algorithm



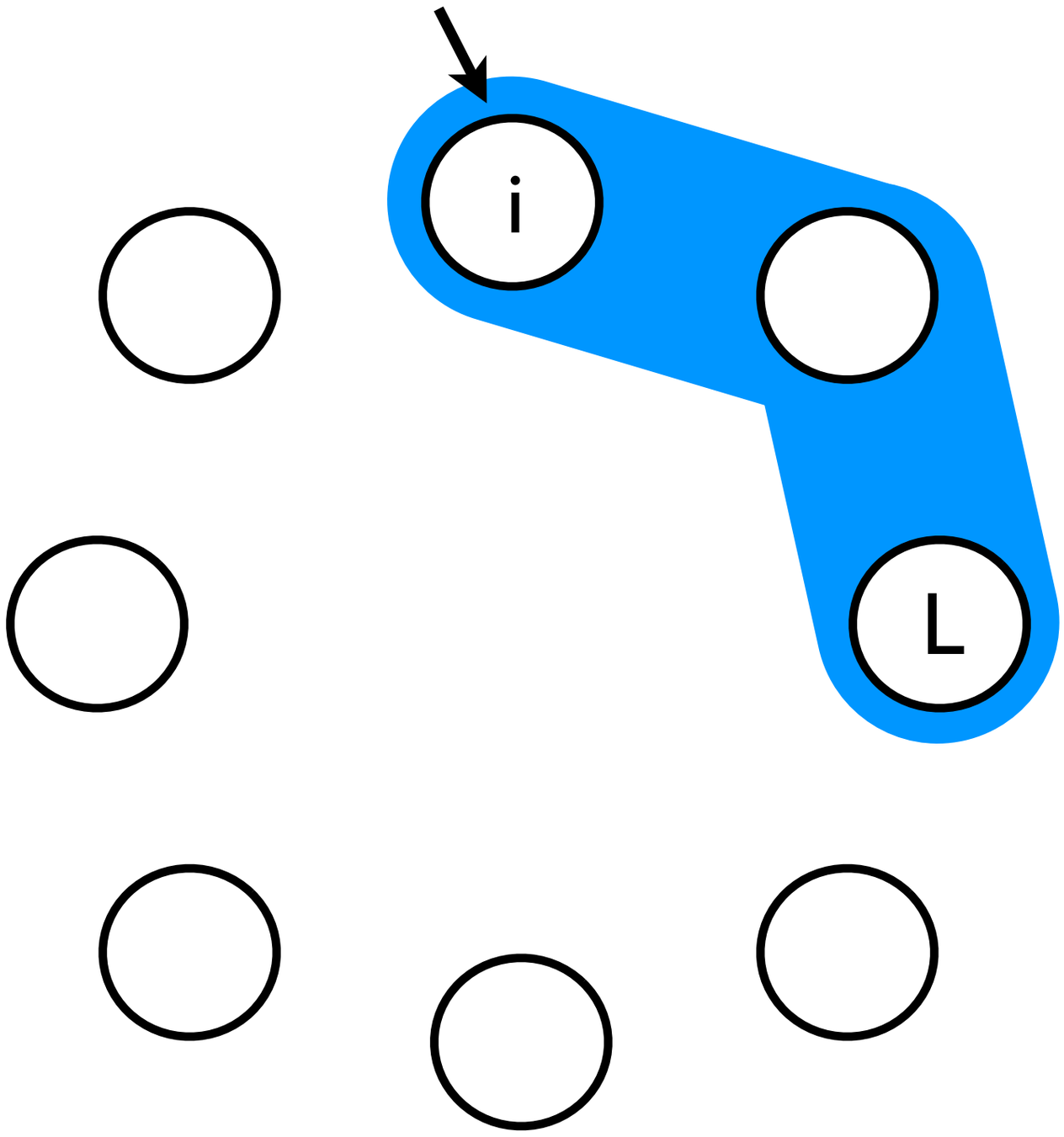
DVMS Algorithm



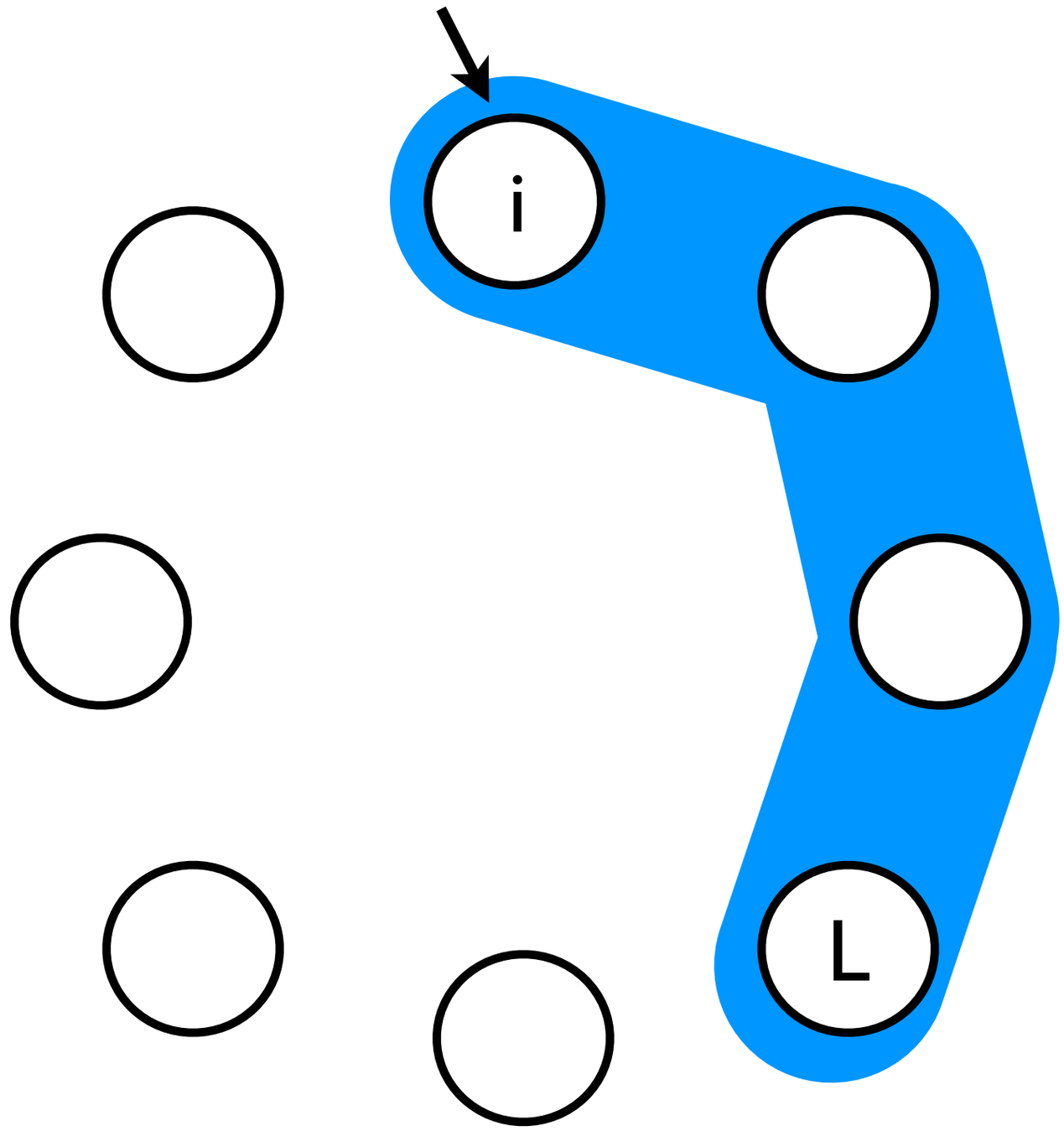
DVMS Algorithm



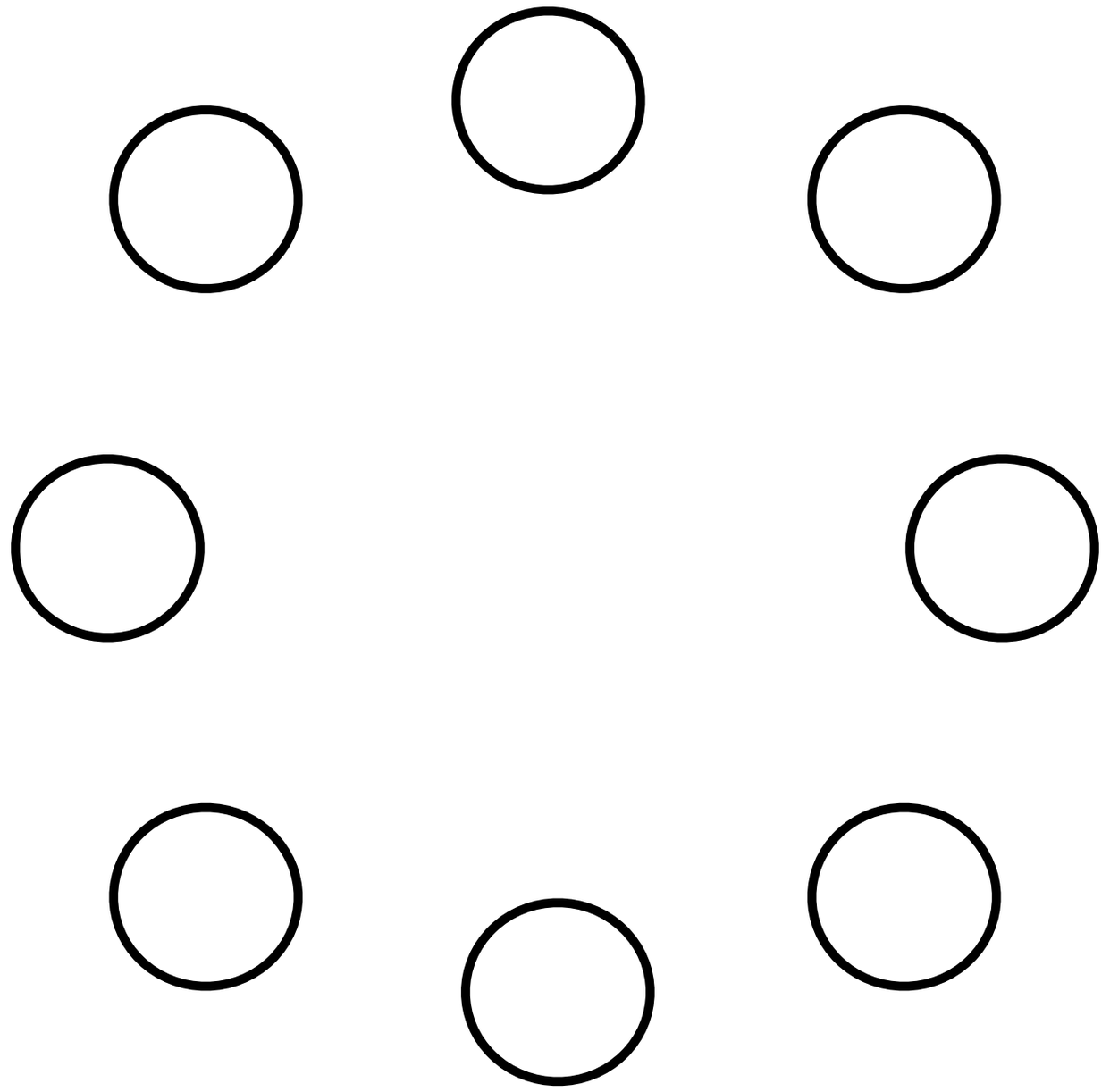
DVMS Algorithm



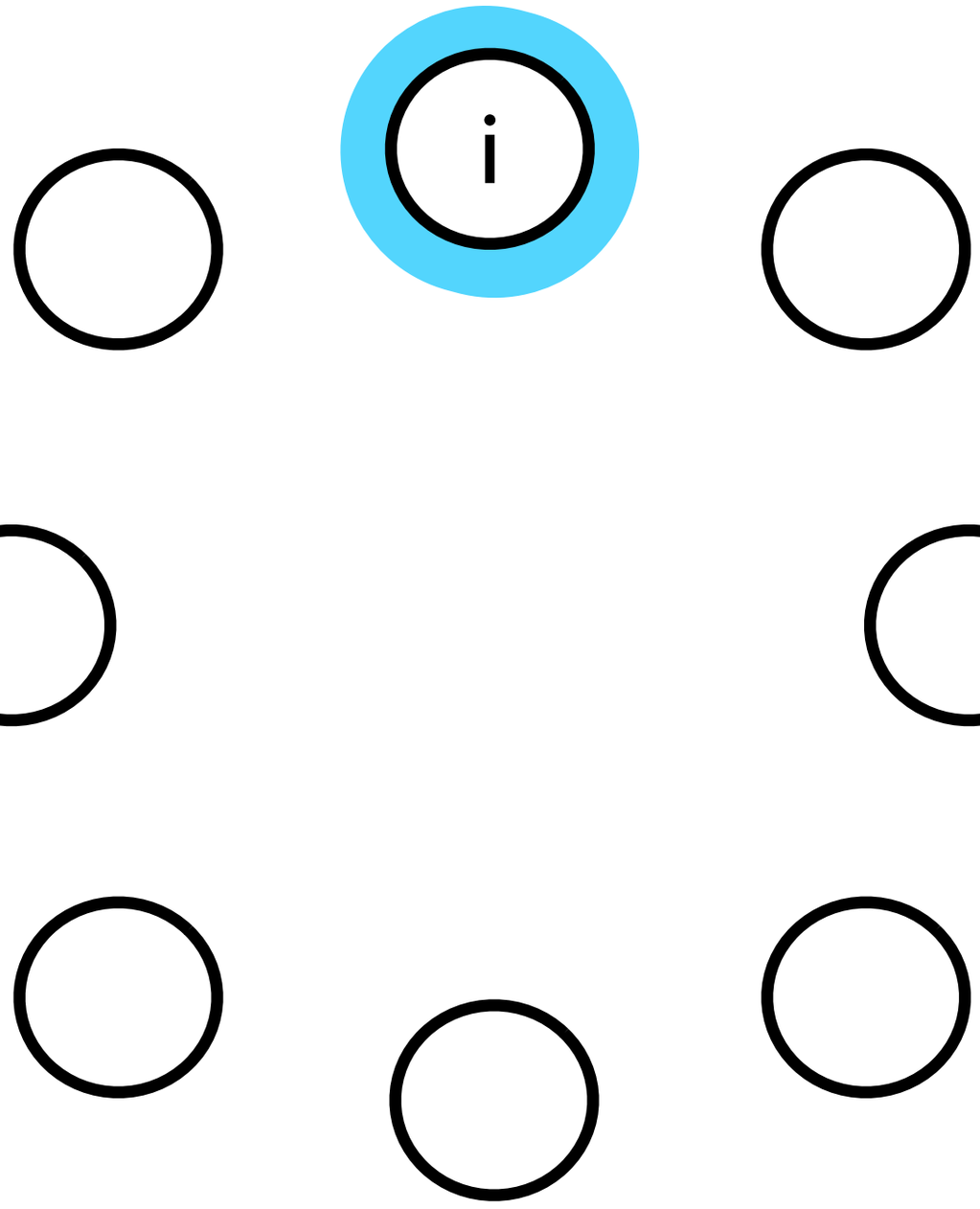
DVMS Algorithm



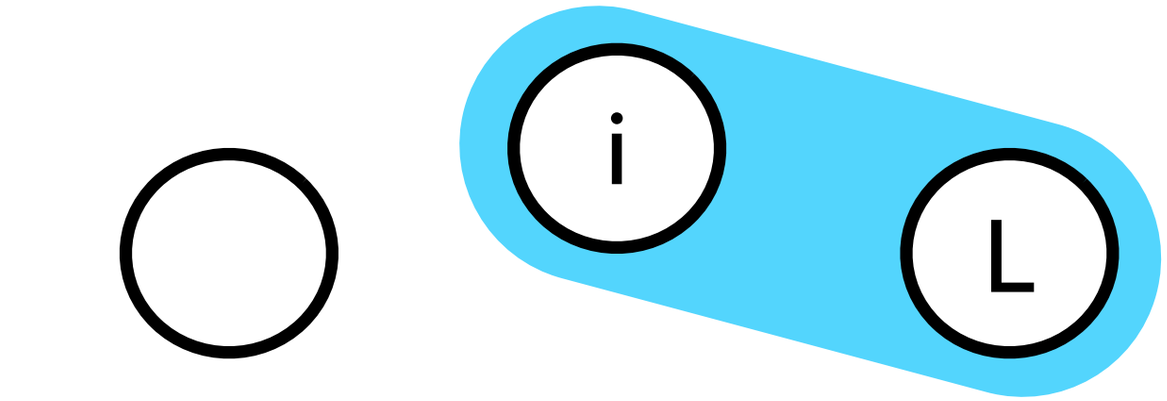
DVMS Algorithm

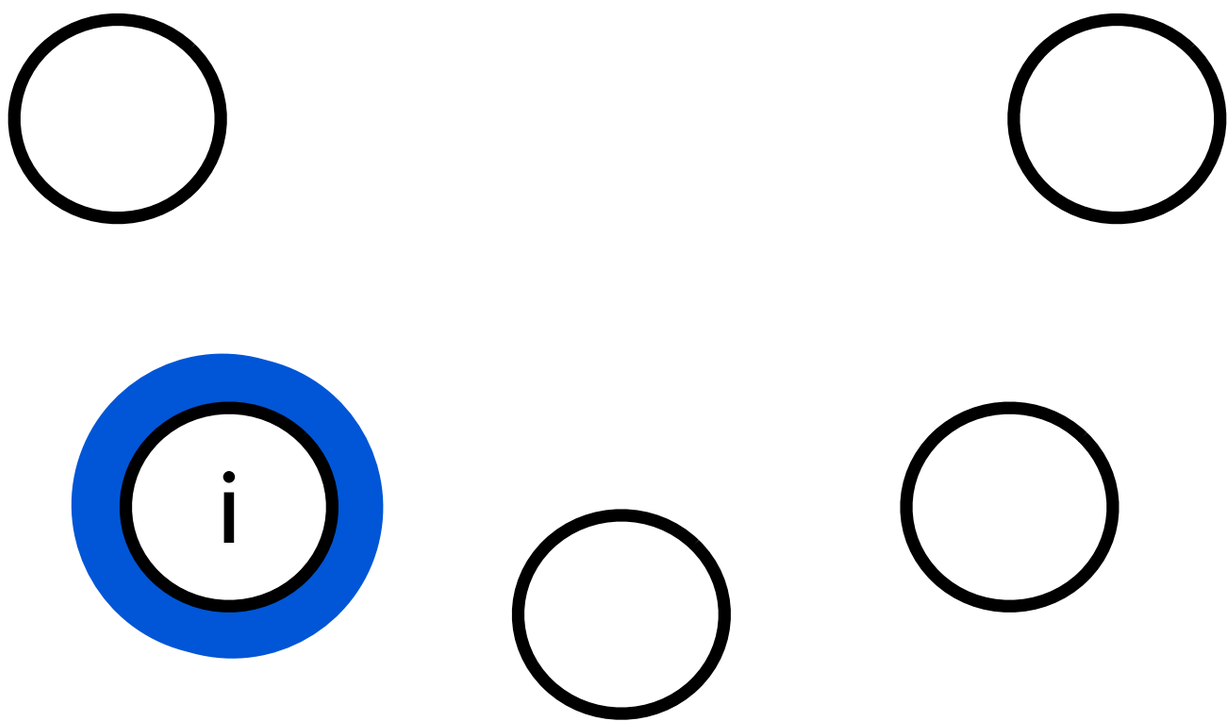


DVMS Algorithm

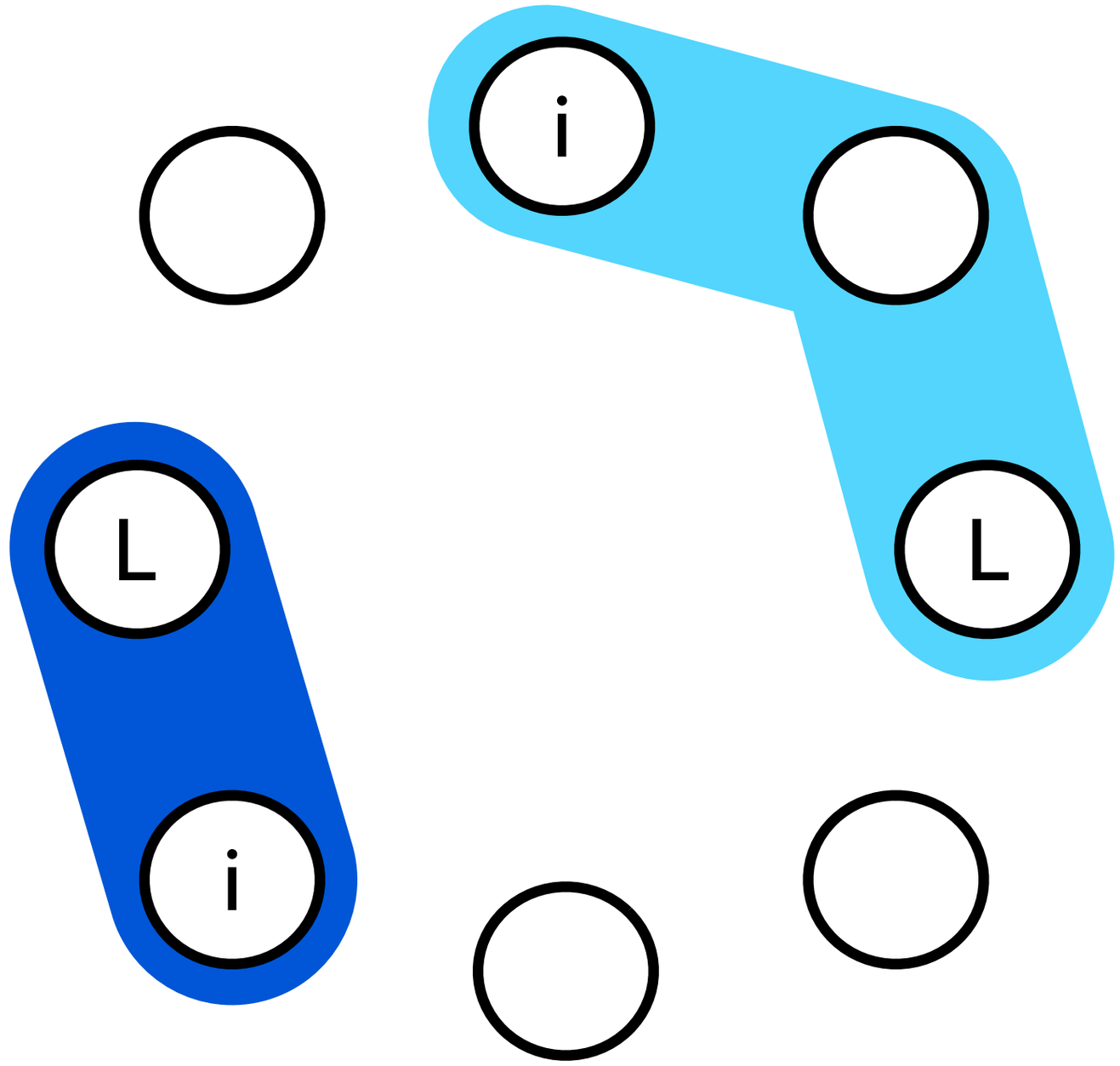


DVMS Algorithm

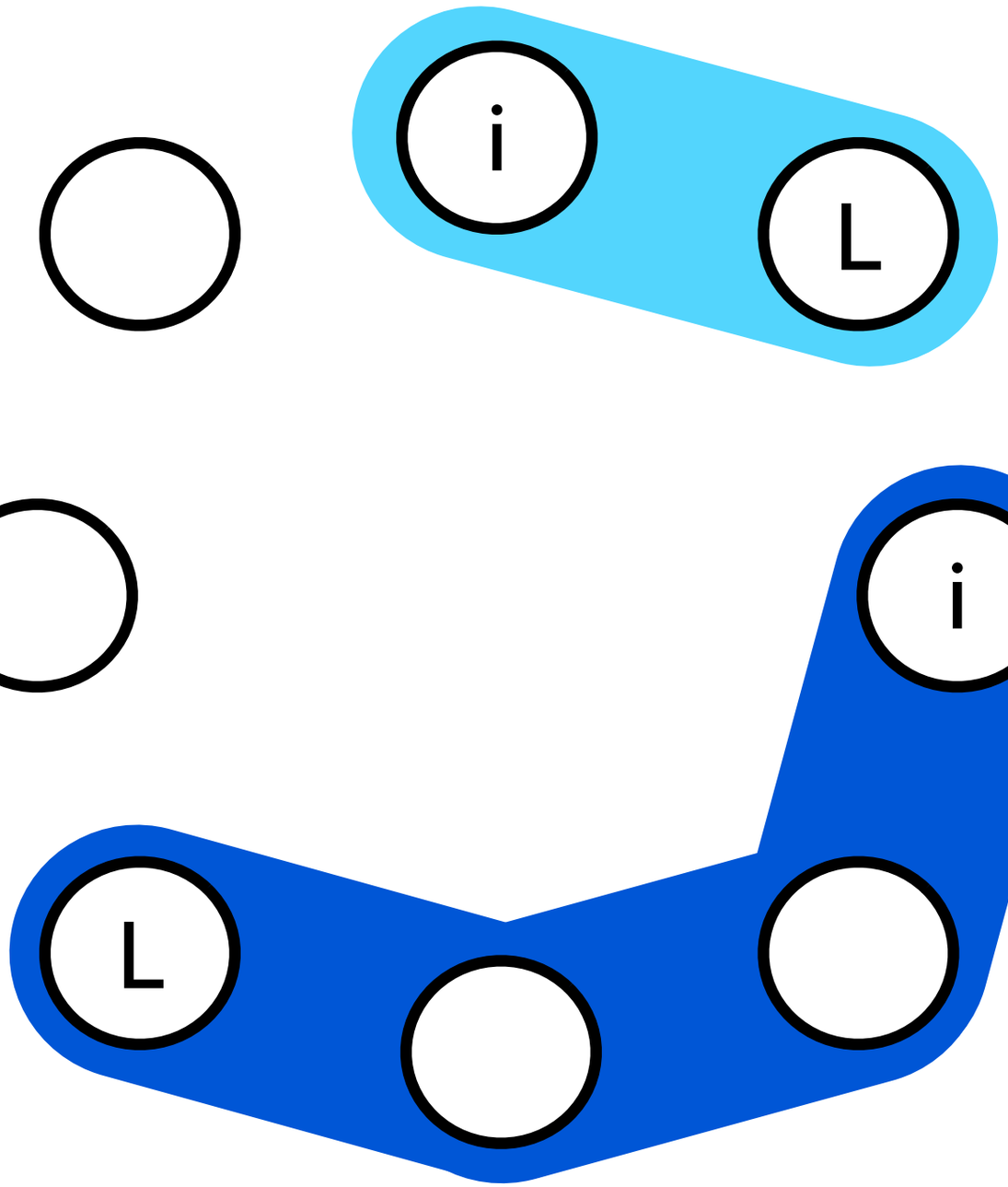




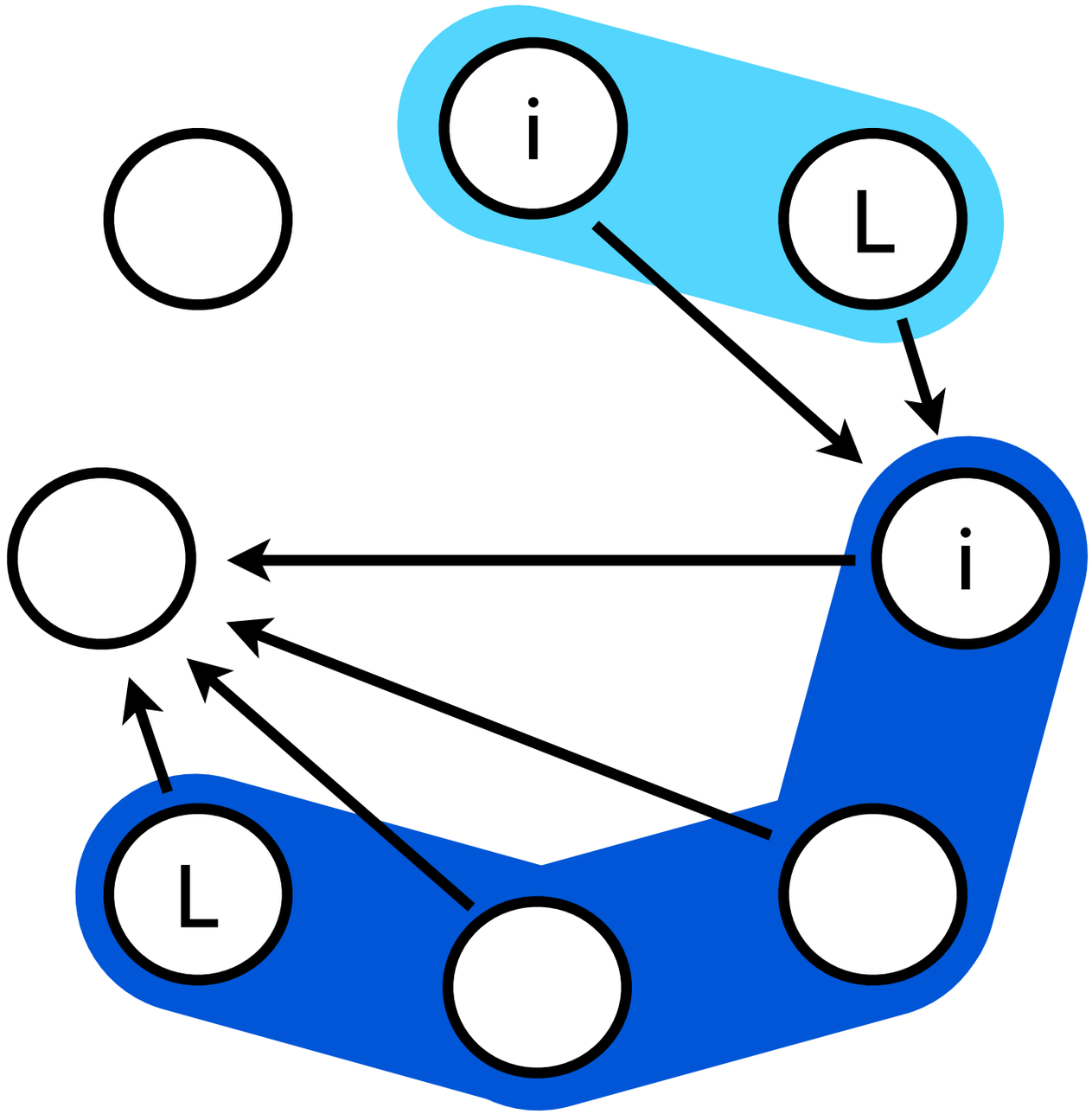
DVMS Algorithm



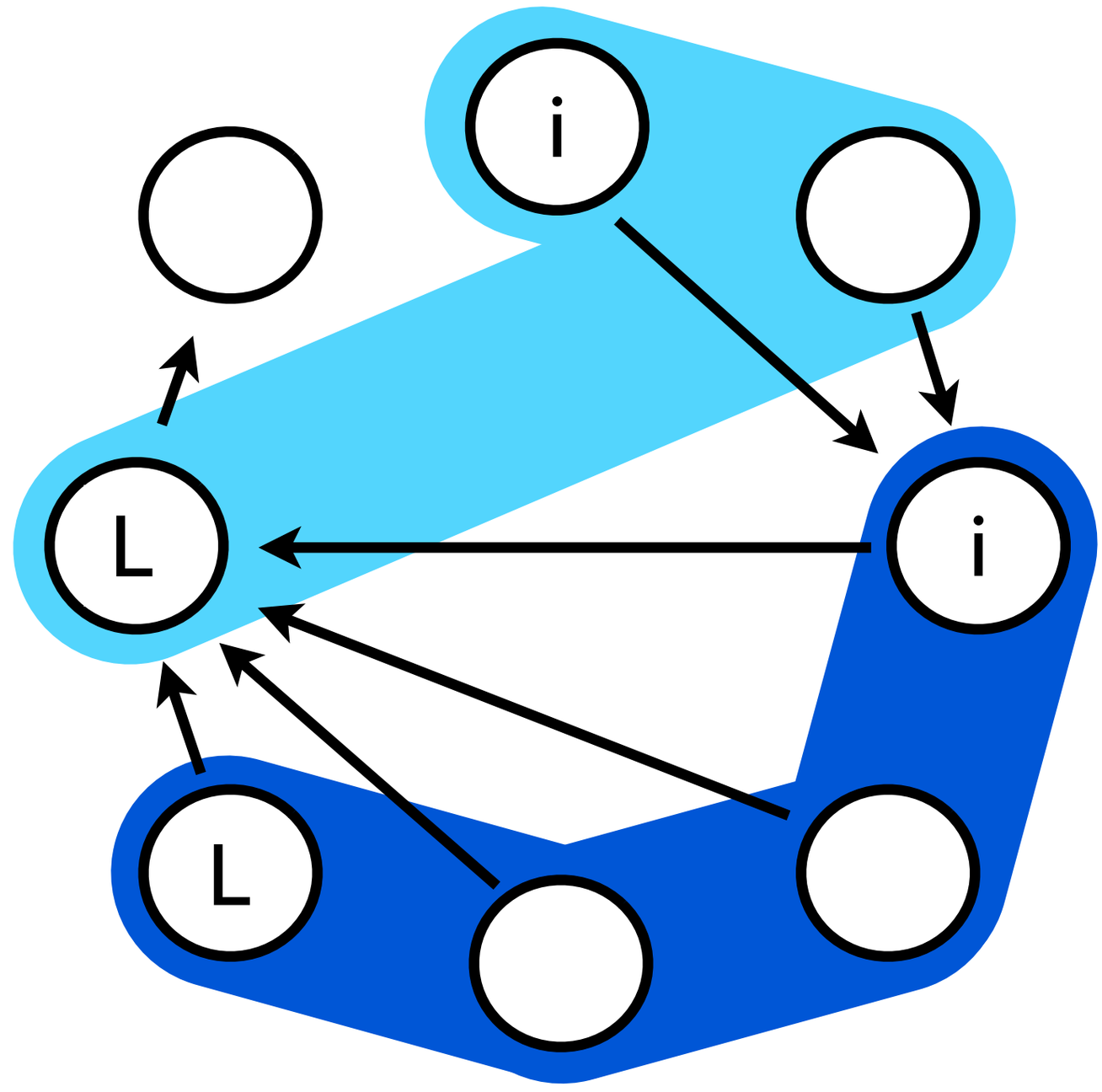
DVMS Algorithm: Shortcuts



DVMS Algorithm: Shortcuts



DVMS Algorithm: Shortcuts



Pros

- Reactivity/scalability
  - Scheduling started when an event is generated
  - Few nodes considered for scheduling

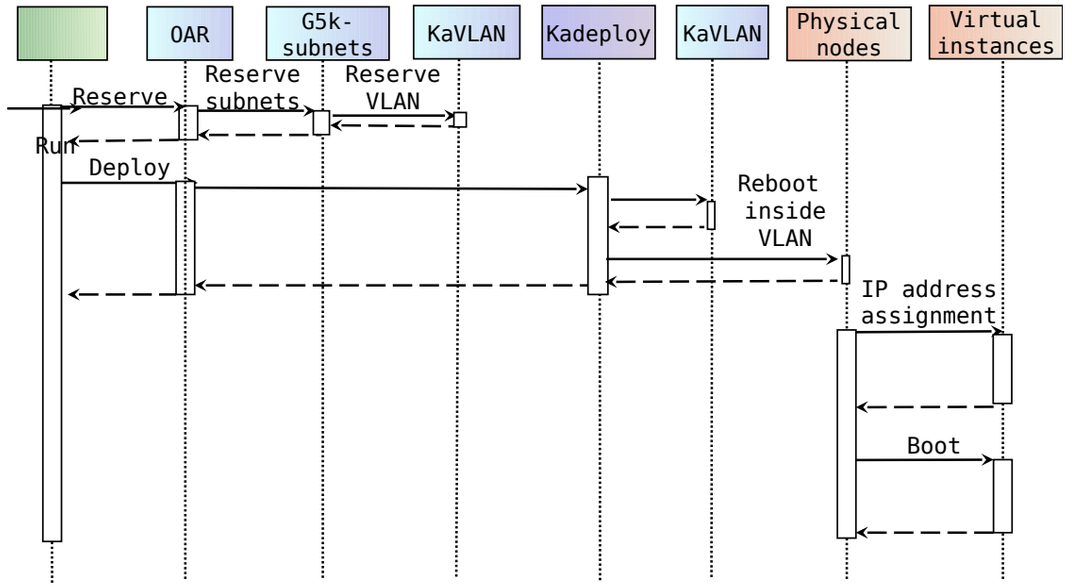
    ⇒ much faster computation
- Resources are partitioned in time and space
  - Several events can be processed simultaneously and independently
- First validation: SimGrid 100K VMs/10K PMs
- Second validation: "in vivo" experiments ?

Large Scale Mamangent of VMs

- Investigating VM concerns implies to ...

  Deploy the template
  Configure/Start each instance
  Control the execution

  ... before conducting experiments



- Performing such a task on
  - Few VMs on one node
  - Hundred of VMs on one site
  - Thousands of VMs on distinct sites

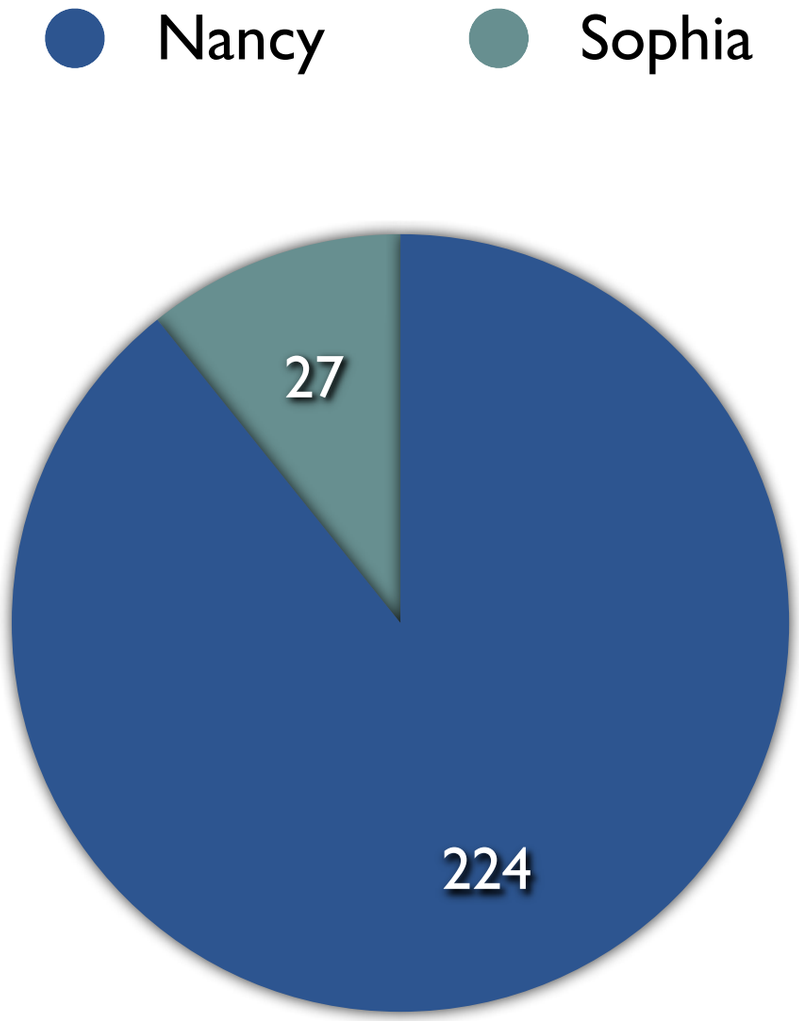


⇒ Designing/Implementing tools to make the study and the investigation of virtualization concerns at large scale easier

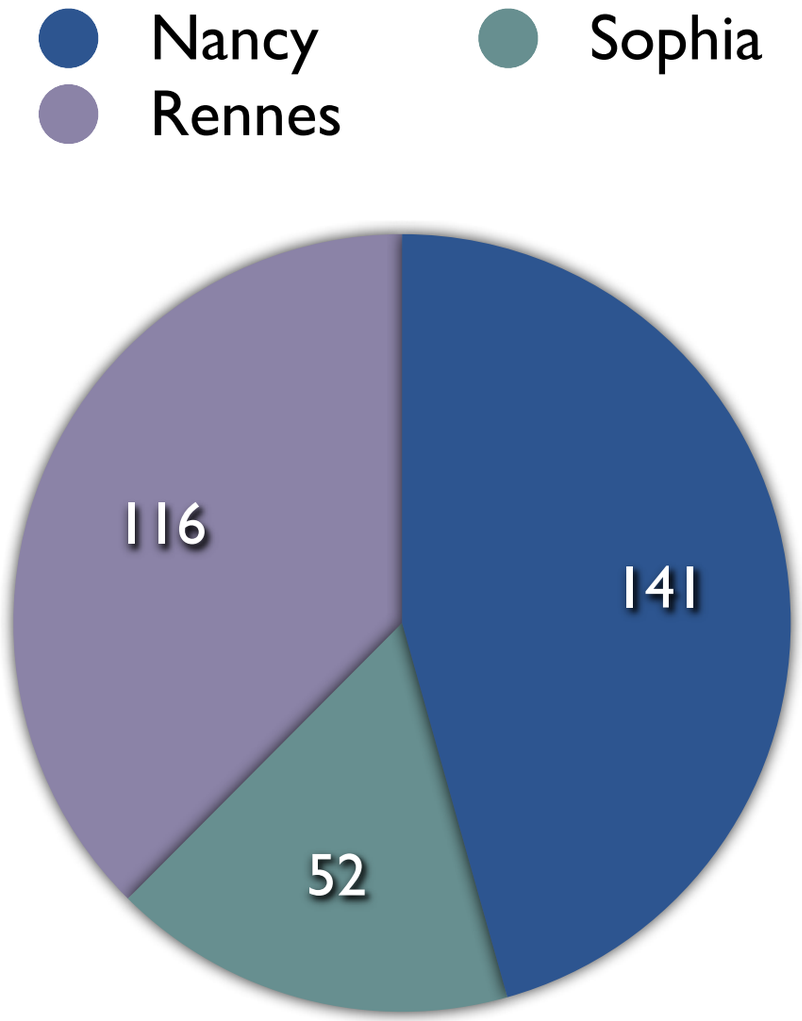
Experiments with Flauncher

- DVMS prototype
  - Language: Java
  - Events activated: node overload
- Comparison between Entropy and DVMS
  - Several non-viable configurations generated
  - VMs started with Flauncher
  - How Entropy and DVMS solve problems for each configuration?

Nodes reserved

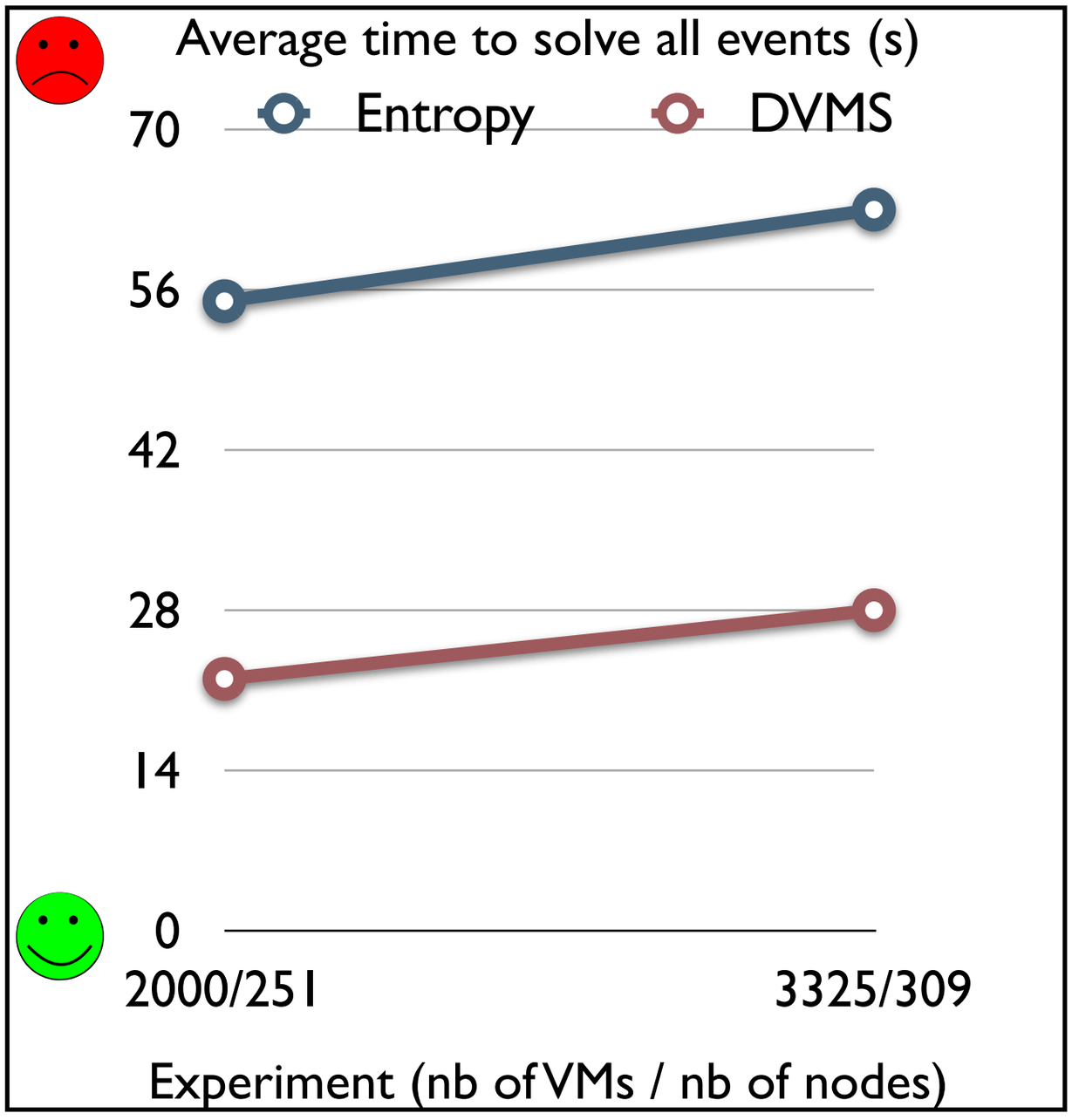


2000 VMs, 251 nodes

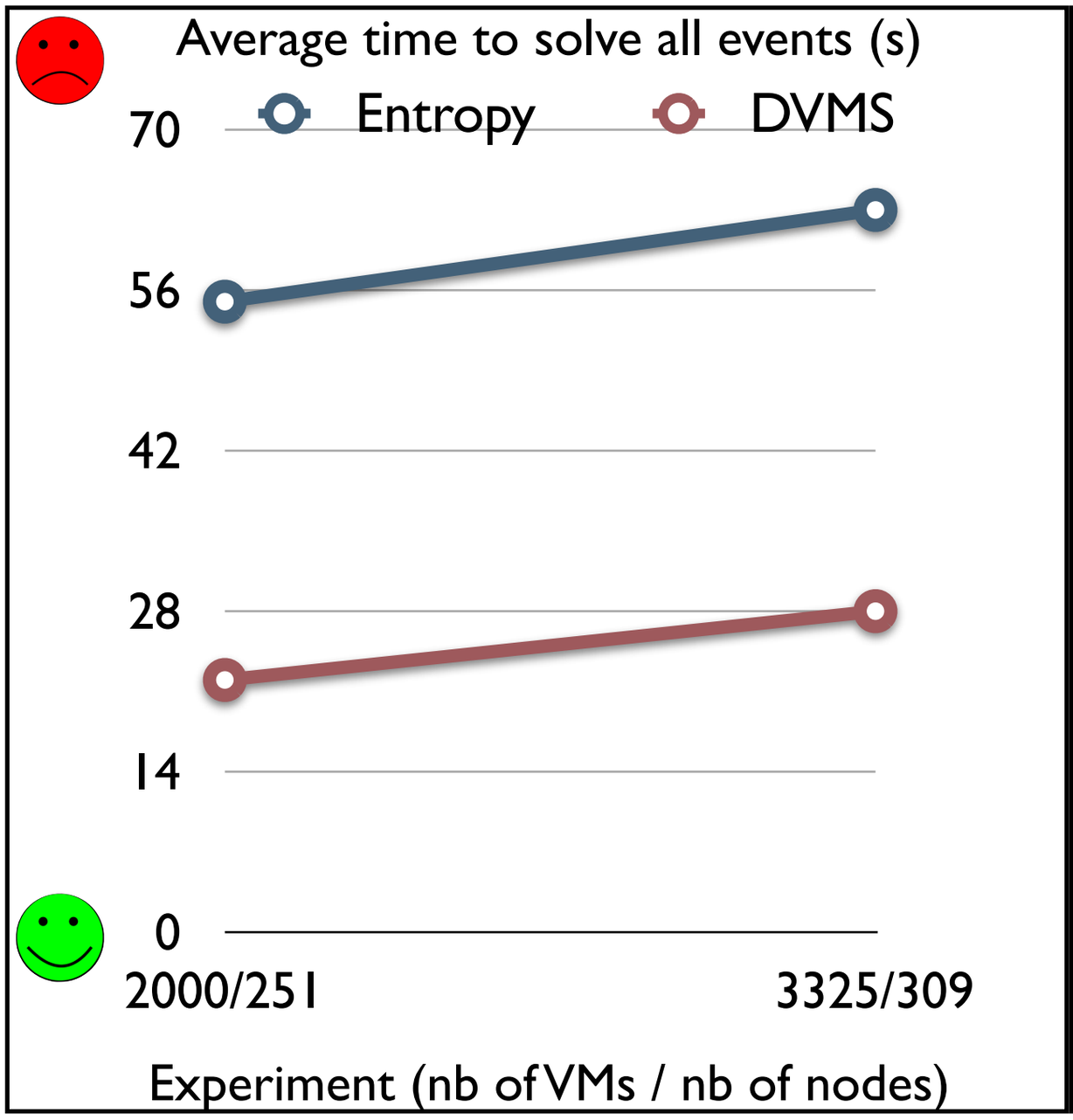


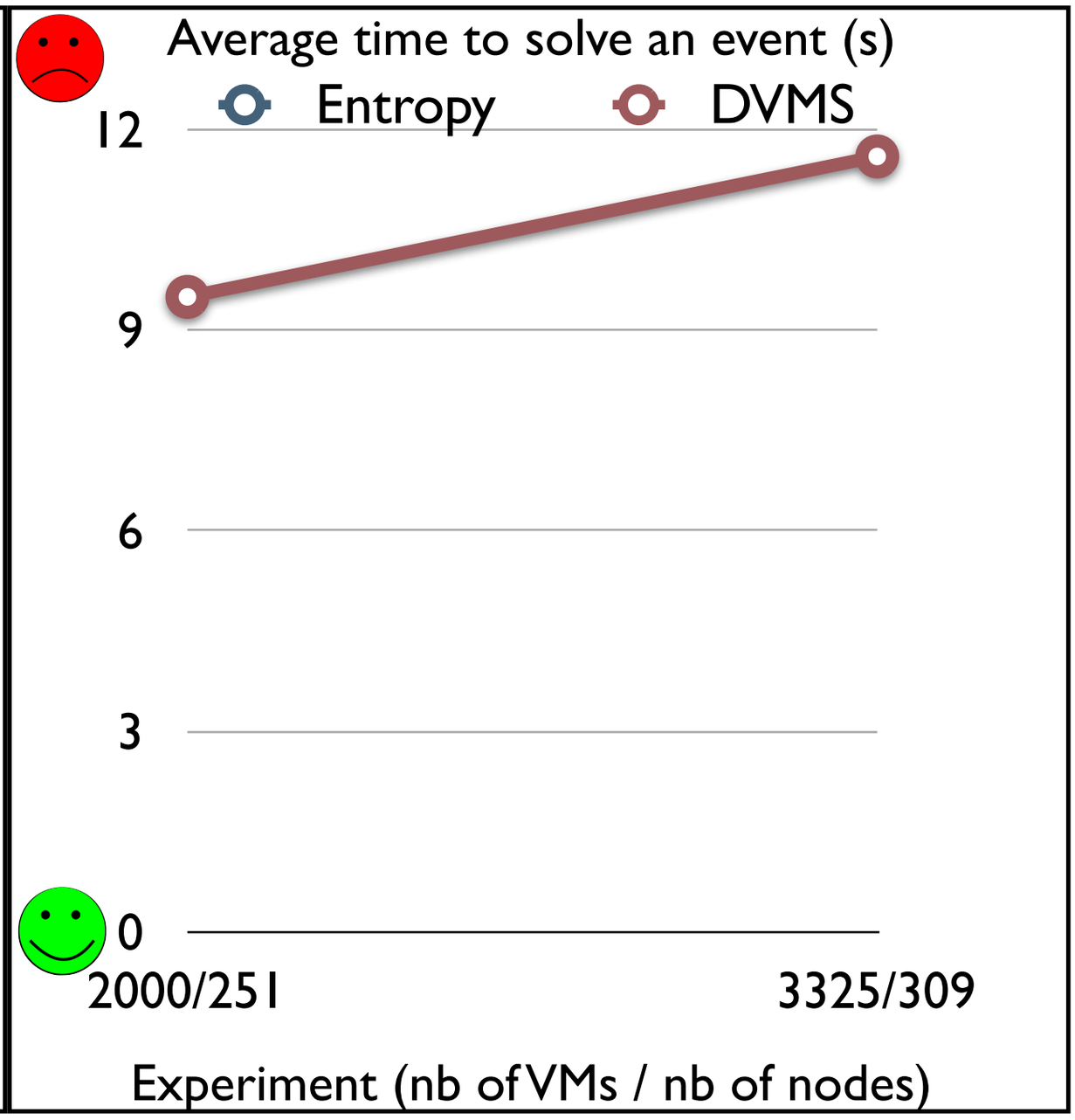
3325 VMs, 309 nodes

Results

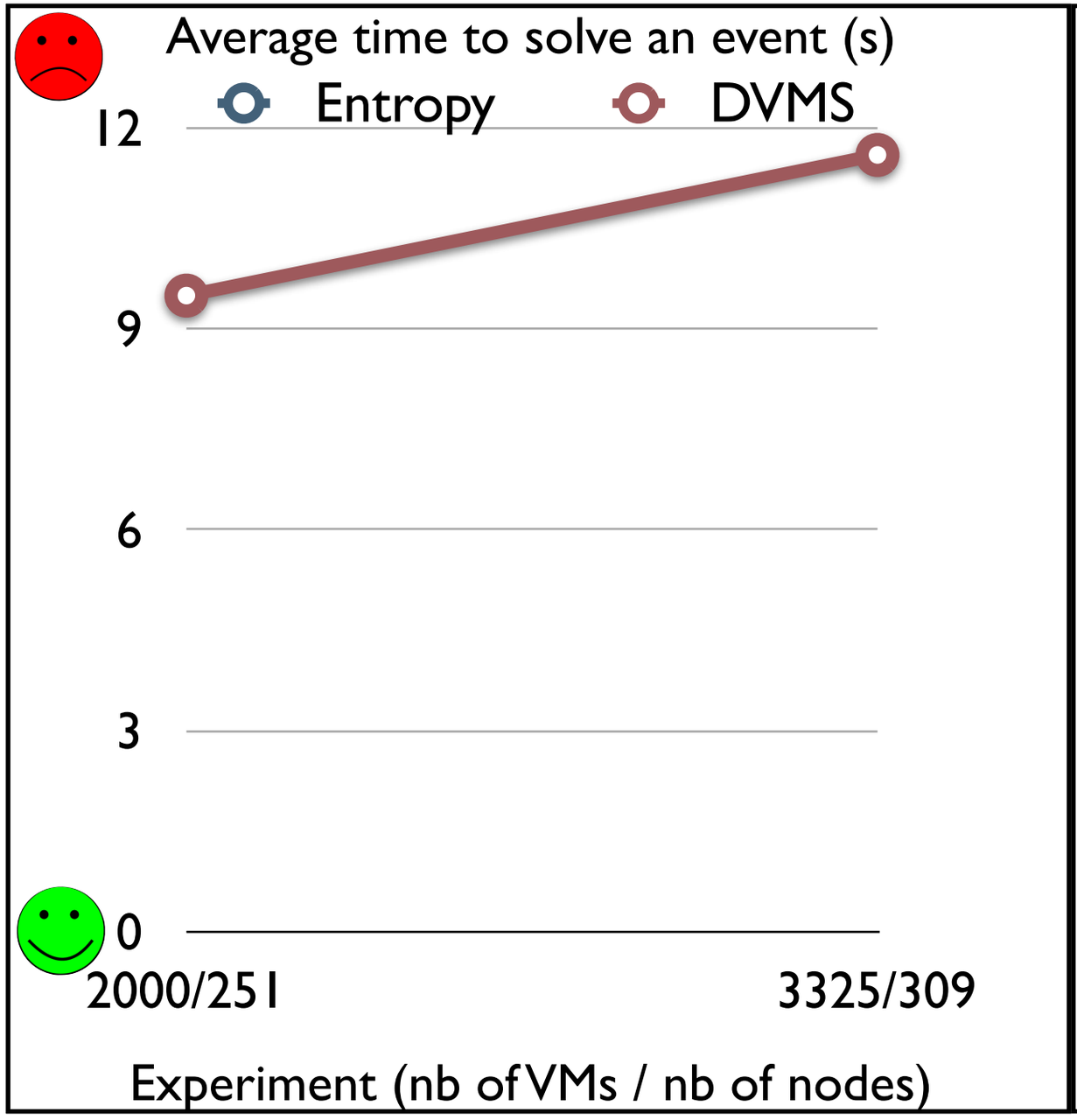

Average time to solve all events (s)
Entropy
DVMS
70
56
42
28
14
0
2000/251
3325/309
Experiment (nb of VMs / nb of nodes)

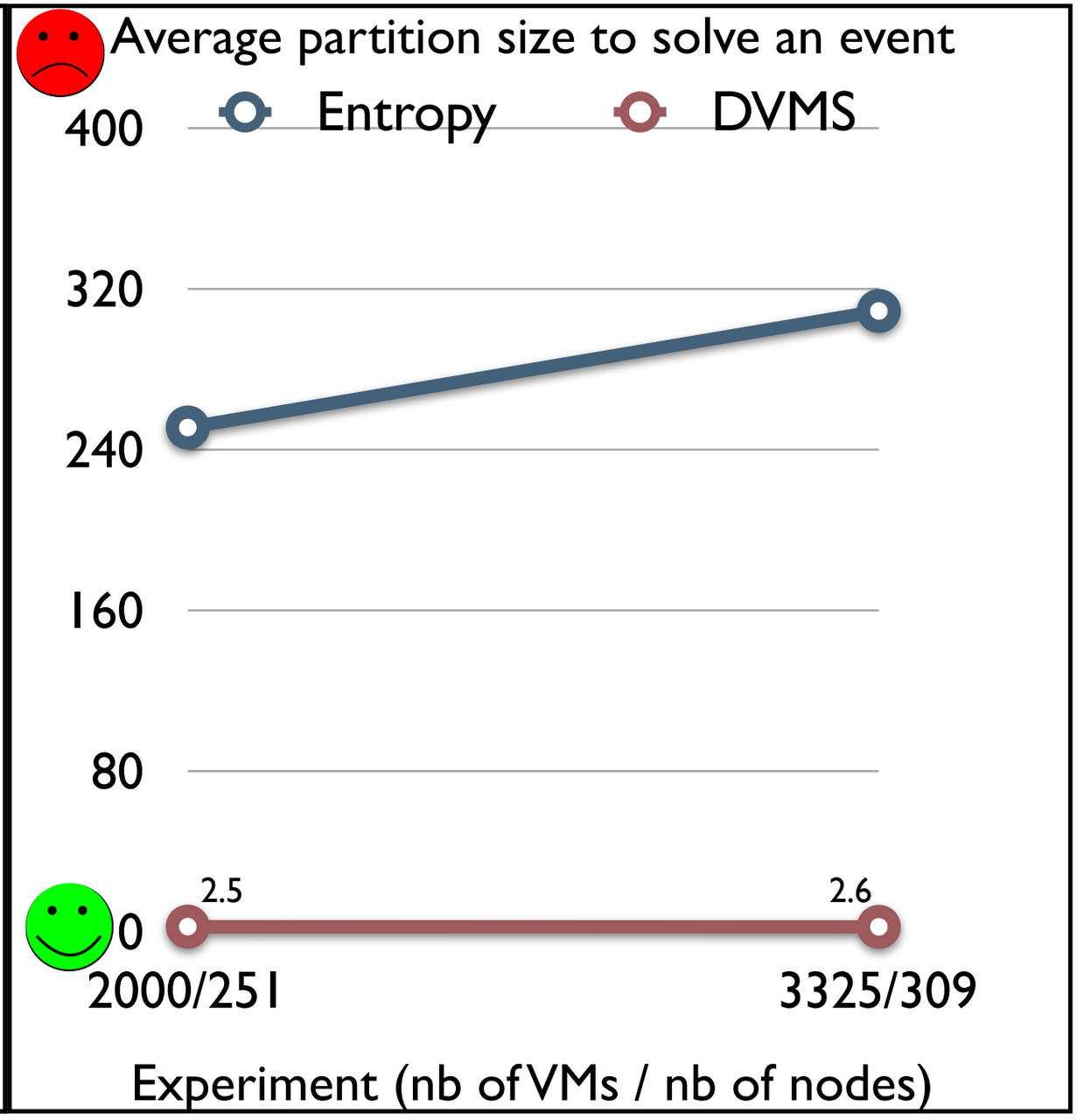
Results


Average time to solve all events (s)
Entropy
DVMS
70
56
42
28
14
0
2000/251
3325/309
Experiment (nb of VMs / nb of nodes)


Average time to solve an event (s)
Entropy
DVMS
12
9
6
3
0
2000/251
3325/309
Experiment (nb of VMs / nb of nodes)

Results


Average time to solve an event (s)
Entropy
DVMS
12
9
6
3
0
2000/251
3325/309
Experiment (nb of VMs / nb of nodes)


Average partition size to solve an event
Entropy
DVMS
400
320
240
160
80
0
2.5
2.6
2000/251
3325/309
Experiment (nb of VMs / nb of nodes)

Lessons learnt

- Nodes not correctly deployed
  - Redeploy, katapult
- Nodes that did not reboot correctly
  - Kareboot
- Routing communications between sites
  - KaVLAN
- Communications with more than 1024 machines
  - Increase the size of ARP table
- Migration errors with libvirt
  - Ensure that each node has a unique id
- ...

Conclusion

- Flauncher, a set of scripts
  - To create, stress and migrate a huge number of VMs distributed on several sites of Grid'5000 (10K VMs so far throughout 5 sites)
  - That can be leveraged to study virtualization and live migration in large scale infrastructures (presented during the Grid'5000 winter school in Nantes, Dec 2012, [RR-8026])
- Next
  - Extension to execute advanced workflows (stress VMs, migrations, ...)
  - Collect./Analyze/Run real traces from cloud providers

Conclusion

- DVMS
  - A distributed solution to schedule VMs dynamically in large scale infrastructures
  - [VHPC 2011], [CCPE 2012] and an ongoing submission.
- Next
  - Mid term, resiliency aspects ("Robustness of Large System of High Churn" challenge)
  - Long term, a building block of a proposal aiming at designing a P2P like Cloud OS (a co-supervised Phd between AVALON and ASCOLA)

Tutorials

- Flauncher
  - https://www.grid5000.fr/mediawiki/index.php/Booting_and_Using_Virtual_Machines_on_Grid'5000
- DVMS
  - http://www.emn.fr/z-info/ascola/doku.php?id=internet:members:fquesnel:deploysimulatorg5kv3

Thank you

References (1/2)

- [Cornabas 10] Jonathan Rouzaud Cornabas. A distributed and collaborative dynamic load balancer for virtual machine. In Euro-Par 2010 workshops , Ischia, Naples Italy, August 2010. Springer.

- [Feller et al. 11] Eugen Feller, Louis Rilling, and Christine Morin. Snooze: A Scalable and Autonomic Virtual Machine Management Framework for Private Clouds. Research report, INRIA Rennes, Rennes, France, December 2011.

- [Hermenier et al. 09] Fabien Hermenier, Xavier Lorca, Jean M. Menaud, Gilles Muller, and Julia Lawall. Entropy: a consolidation manager for clusters. In VEE '09: Proceedings of the 2009 ACM SIGPLAN/SIGOPS international conference on Virtual execution environments, pages 41–50, New York, NY, USA, March 2009. ACM.

- [Hoffa et al. 08] On the Use of Cloud Computing for Scientific Workflows, Hoffa, C., G. Mehta, T. Freeman, E. Deelman, K. Keahey, B. Berriman, J. Good. SWBES 2008, Indianapolis, IN. December 2008

- [Lèbre et al. 11] DISCOVERY, Beyond the Clouds, Lebre, A. and Anedda, P. and Gaggero, M. and Quesnel, F. In Euro-Par 2011 workshops, Bordeaux, France, August 2011. Springer.

References (2/2)

- [Lowe 09] Scott Lowe. Introducing VMware vSphere 4. Wiley Publishing Inc., Indianapolis, Indiana, first edition, September 2009.

- [Mastroiani et al. 11] Carlo Mastroianni, Michela Meo and Giuseppe Papuzzo. Self-economy in cloud data centers: statistical assignment and migration of virtual machines. In Euro-Par'11: Proceedings of the 17th international conference on Parallel processing, pages 407-418, Bordeaux, France, August 2011. Springer.

- [Nurmi et al. 09] Daniel Nurmi, Rich Wolski, Chris Grzegorczyk, Graziano Obertelli, Sunil Soman, Lamia Youseff, and Dmitrii Zagorodnov. The eucalyptus open-source cloud-computing system. In CCGRID ’09: Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, volume 0, pages 124–131, Washington, DC, USA, May 2009. IEEE Computer Society.

- [Sotomayor et al. 09] Borja Sotomayor, Rubén S. Montero, Ignacio M. Llorente, and Ian Foster. Virtual infrastructure management in private and hybrid clouds. IEEE Internet Computing, 13(5):14–22, September 2009.

- [Yagiz et al. 10] Yagiz O. Yazir, Chris Matthews, Roozbeh Farahbod, Stephen Neville, Adel Guitouni, Sudhakar Ganti, and Yvonne Coady. Dynamic resource allocation in computing clouds using distributed multiple criteria decision analysis. In Cloud ’10: IEEE 3rd International Conference on Cloud Computing, pages 91–98, Los Alamitos, CA, USA, July 2010. IEEE Computer Society.