Resource Management
for
Virtual Clusters

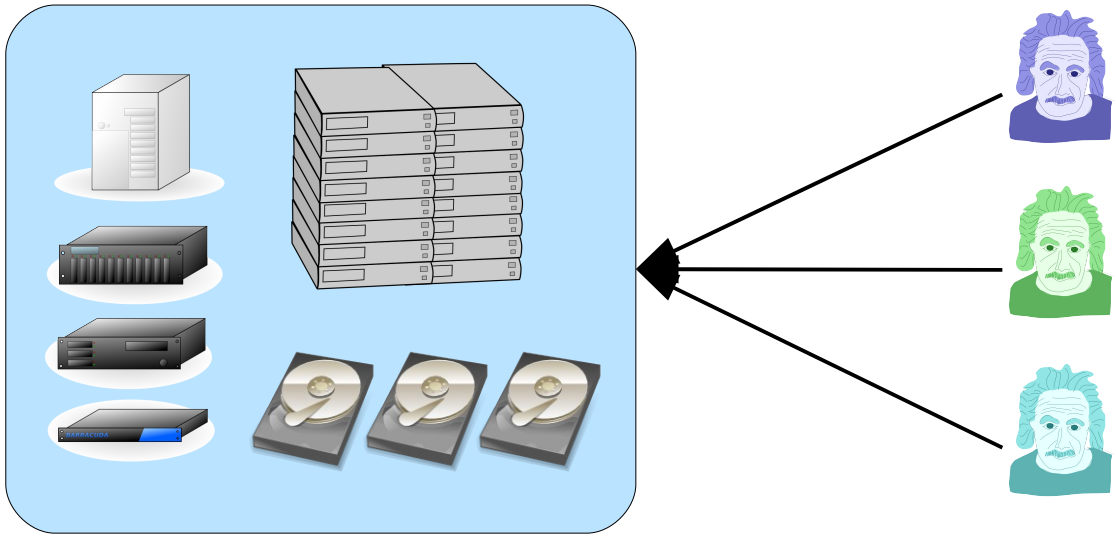
Borja Sotomayor
DSL Workshop
06-02-2006

Index

- Problem and Status
- Scheduling Virtual Workspaces
- Roadmap

Index

- Problem and Status
- Scheduling Virtual Workspaces
- Roadmap



Resource provider

Provides computational, storage, and network resources

Resource consumers

Want to run experiments on the resources, but they each have different software and hardware requirements

Has to provide resources to several users at once	Wants as many resources as possible
Has to balance the software needs of multiple users	Wants to use certain software packages
Has to provide a limited execution environment for security reasons	Wants as much control as possible over resources

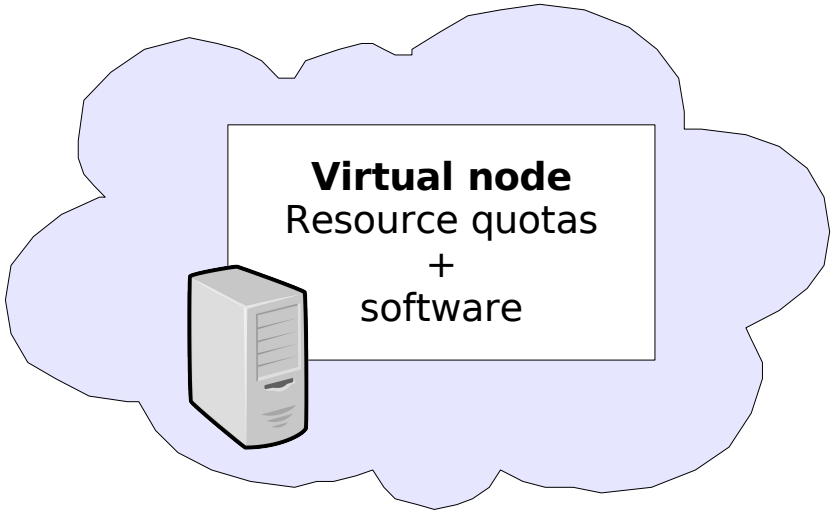
Problem (I)

- Current solution: Impose restrictions on resource consumers.
  - Widespread abstraction: *job*
- Ideally, we want to eliminate these conflicts.
- Possible solution: *virtual workspaces*

Workspace Refresher (I)

- Let's take a look at how virtual workspaces work.

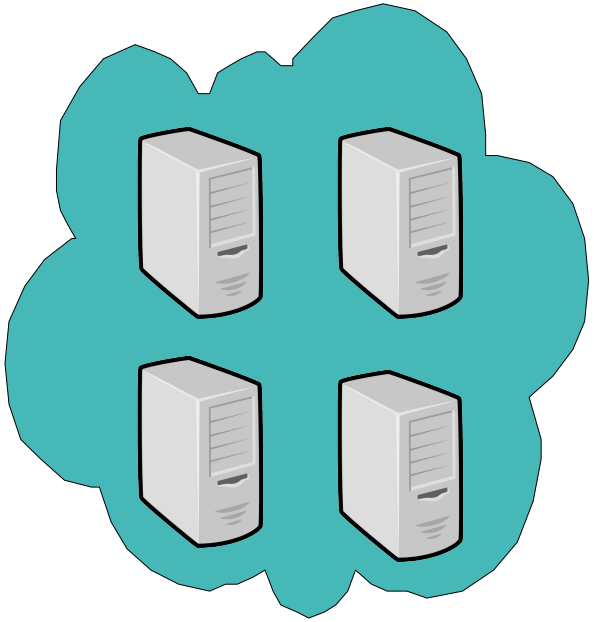
Workspace Refresher (II)



Virtual workspace

- A virtual workspace includes...
  - Resource allocation (disk, CPU, memory, ...)
  - Software (encapsulated in a VM)

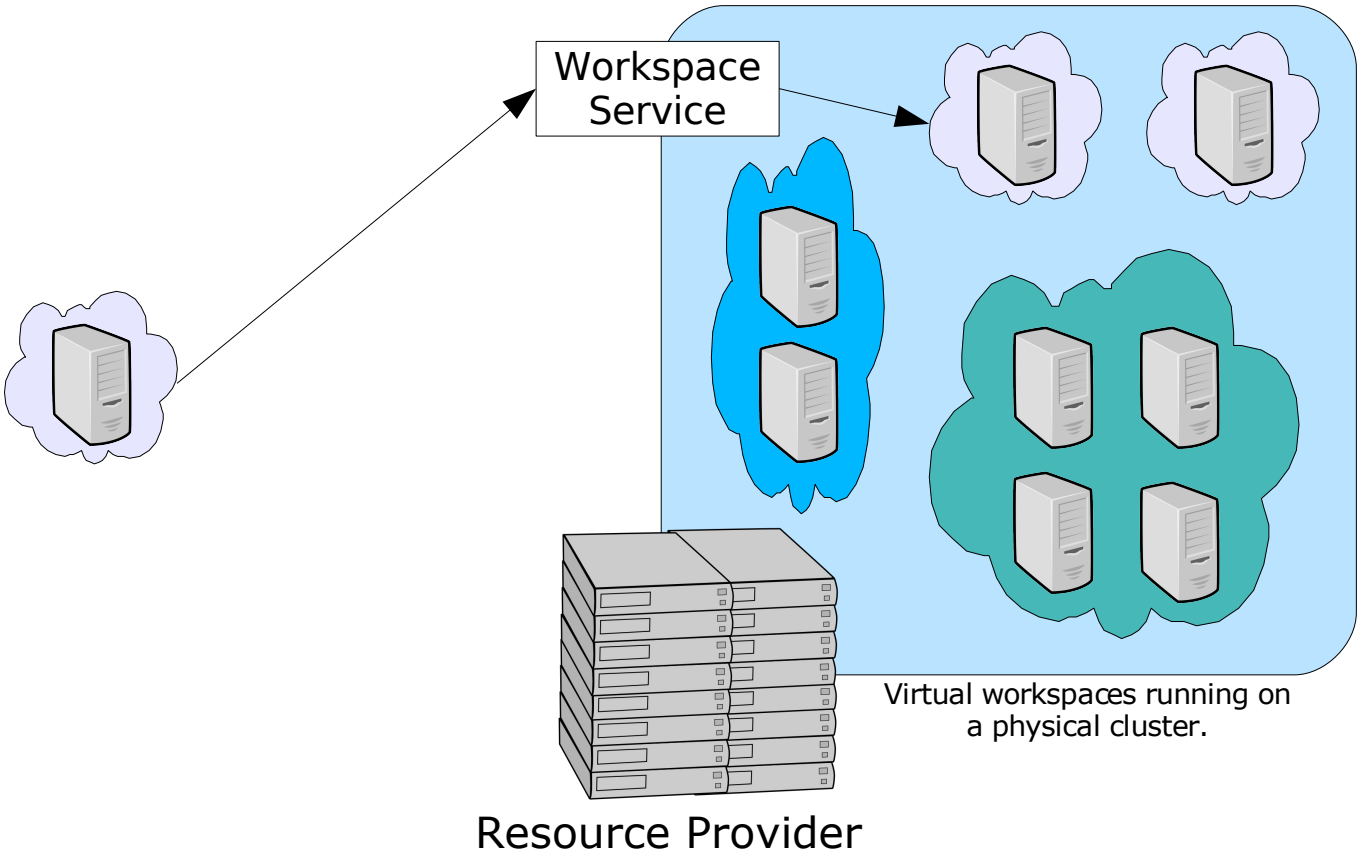
Workspace Refresher (III)



Virtual workspace

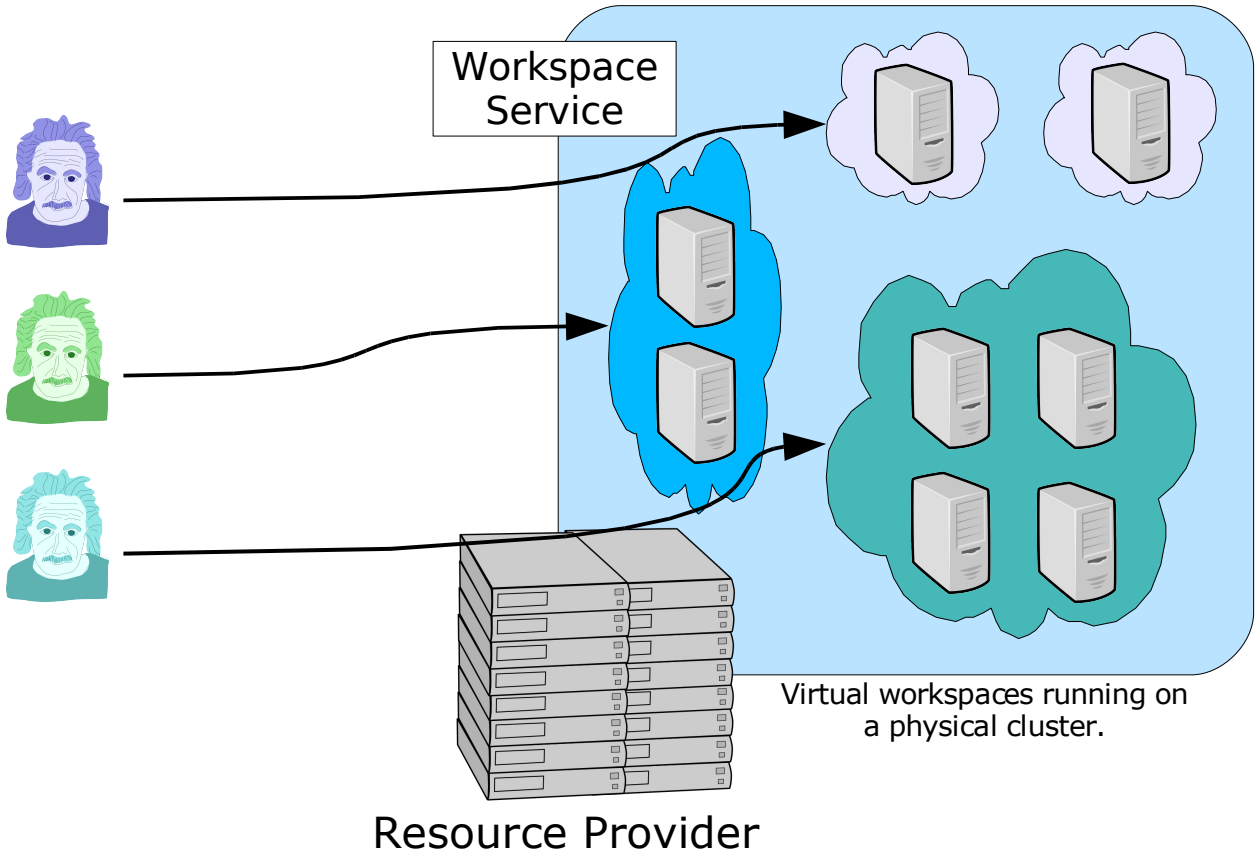
- A virtual workspace can have multiple nodes (*aggregate workspace* or *virtual cluster*)

Workspace Refresher (IV)



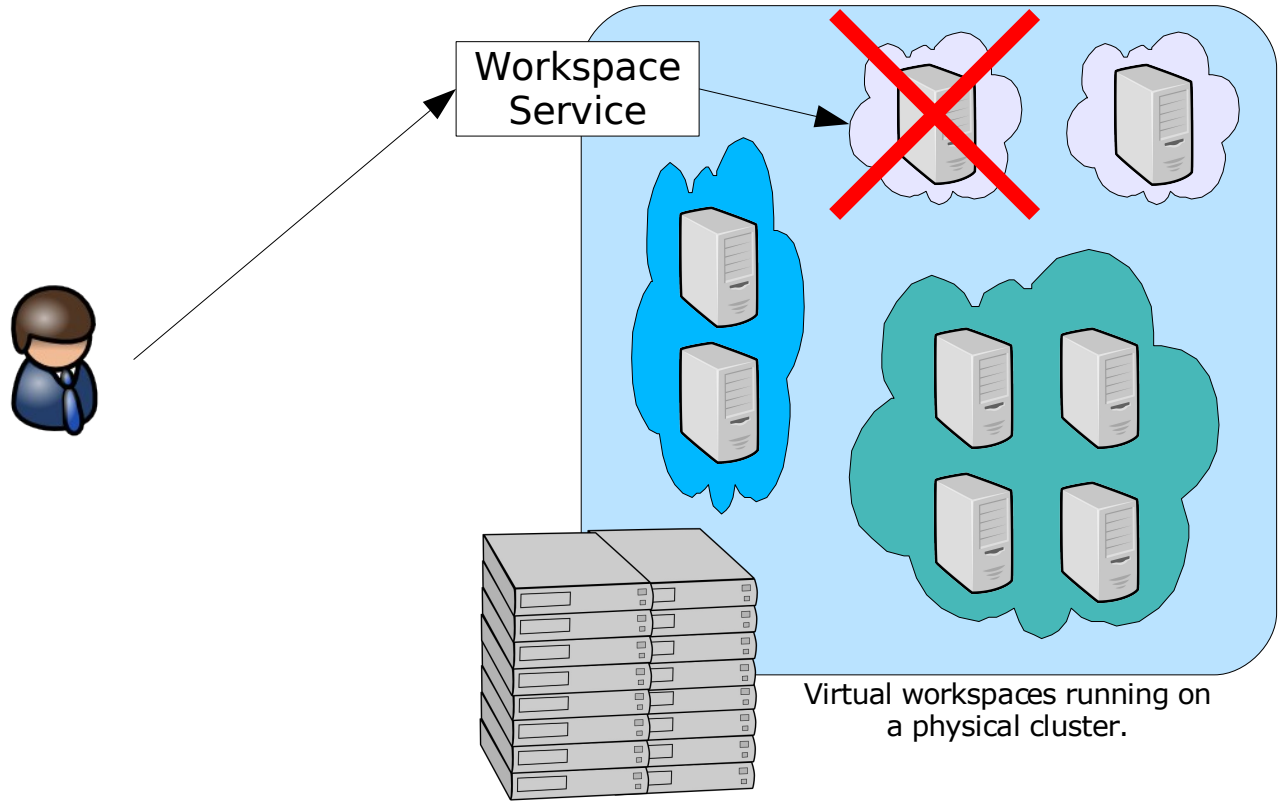
- A virtual workspace is deployed into a resource provider using the workspace service. The workspaces are VMs running on the resource provider's nodes (which must be VM-enabled)

Workspace Refresher (V)



- Users interact with the workspaces as if they were just another physical resource.

Workspace Refresher (VI)



- The workspace's creator can manage it through the Workspace Service (pause, destroy, etc.)

Problem (II)

- Use cases
  - Educational
    - Virtual labs
    - Homework
    - Virtual servers
  - Scientific
    - Interactive experiments
    - Batch jobs
    - Event-driven jobs

Problem (III)

- General scenarios
  - Advance Reservation (AR)
    - Typically, but not necessarily, interactive workloads
  - Batch
    - Generally preemptible
  - Event-driven
    - High priority

Status (I)

- Unfortunately, there's still a lot of work to be done in virtual workspaces!
- Several groups are working on Virtual Workspaces, including Globus.
  - VIOLIN + VioCluster
  - Virtuoso
  - In-VIGO
  - Cluster-On-Demand
- Generally geared towards batch workloads, assuming 1 job/workspace.
- No advance reservation, and no scheduler that can deal with the three workloads simultaneously.

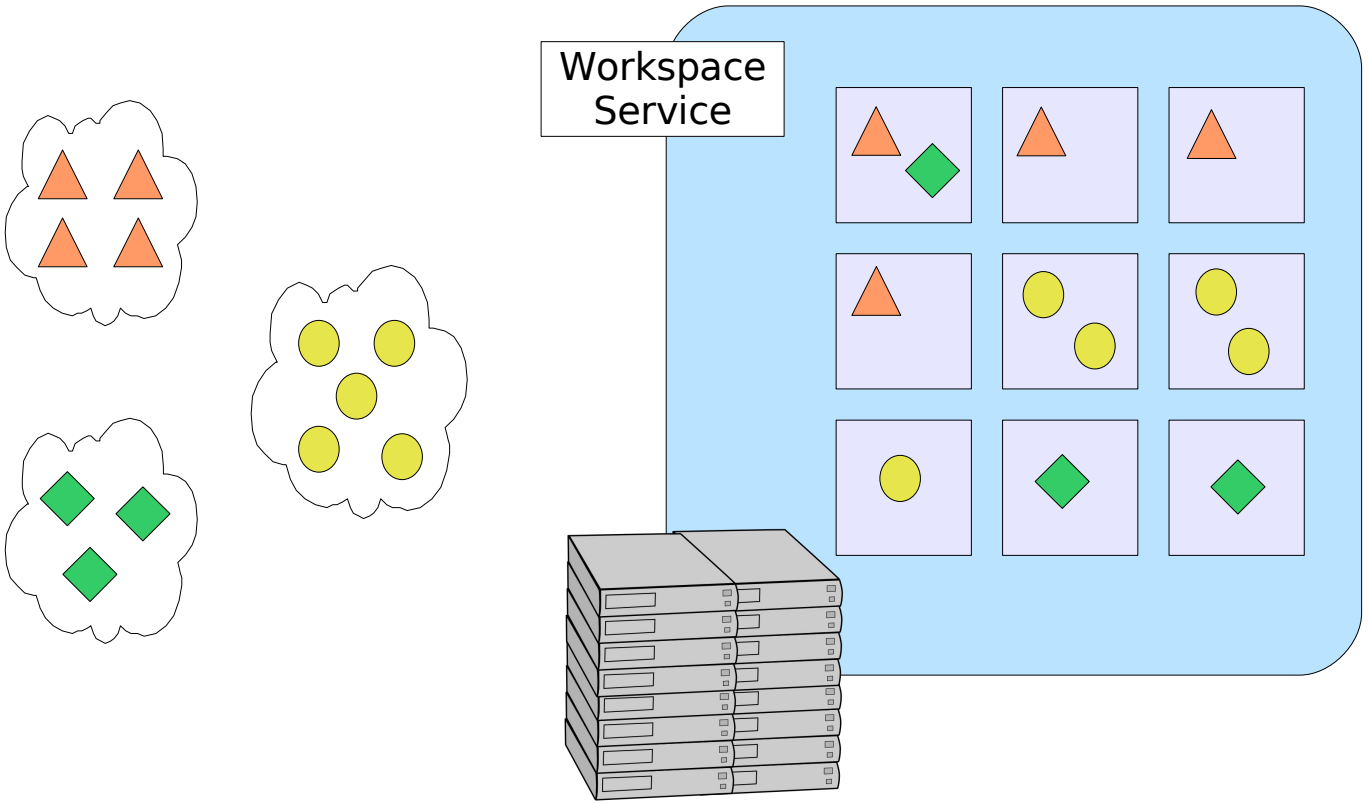
Status (II)

- GT4 Virtual Workspaces
  - http://workspace.globus.org/
  - Technology Preview 1.1 includes support for atomic virtual workspaces.
  - We're working on supporting virtual clusters.
- The main challenge is developing a virtual cluster scheduler.

Index

- Problem and Status
- Scheduling Virtual Workspaces
- Future work

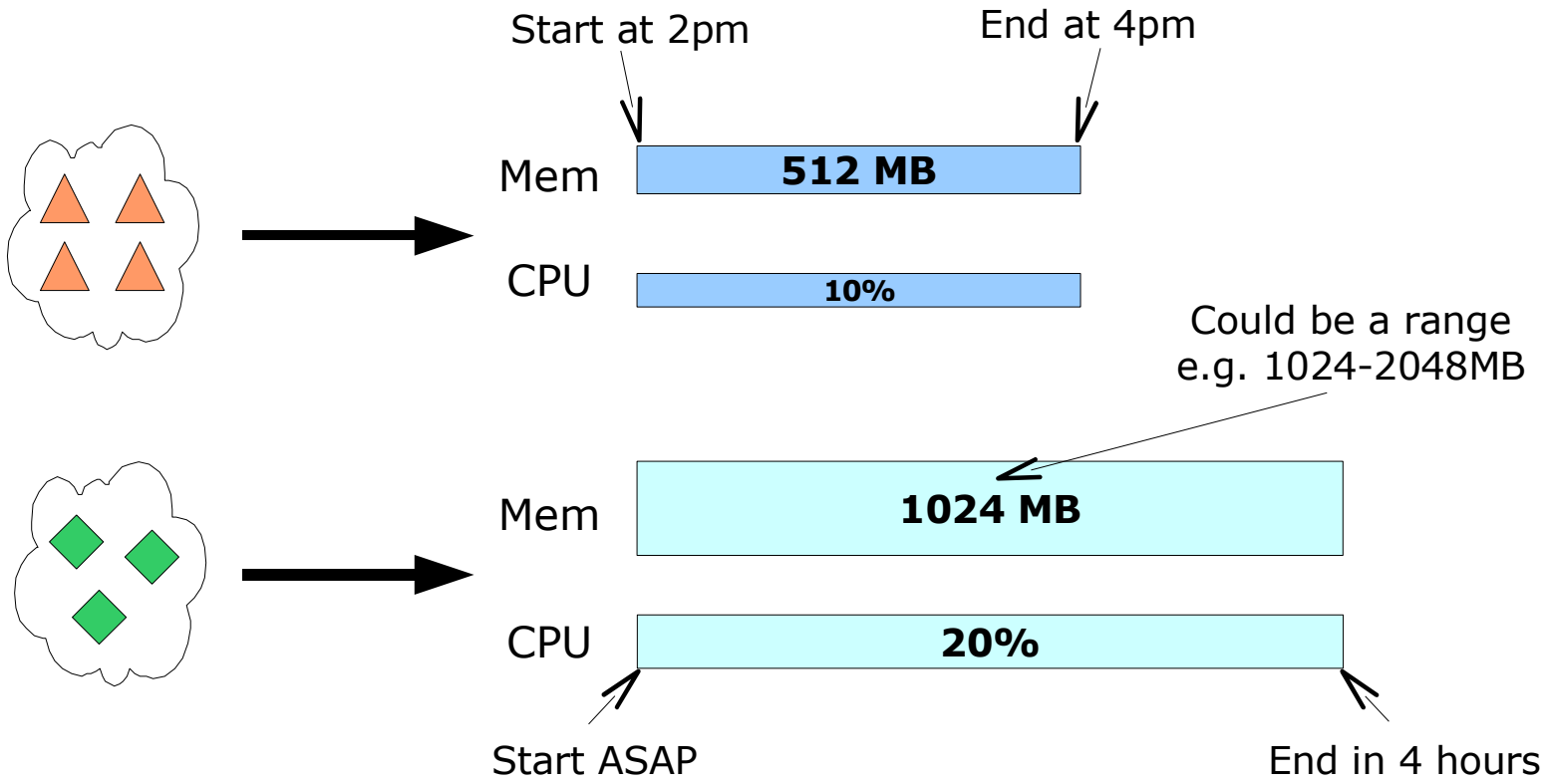
Easier said than done!



- How do we map virtual resources to physical resources? A lot of variables to consider!
  - Advance reservation? Preemptible? Resource allocation? Overhead?

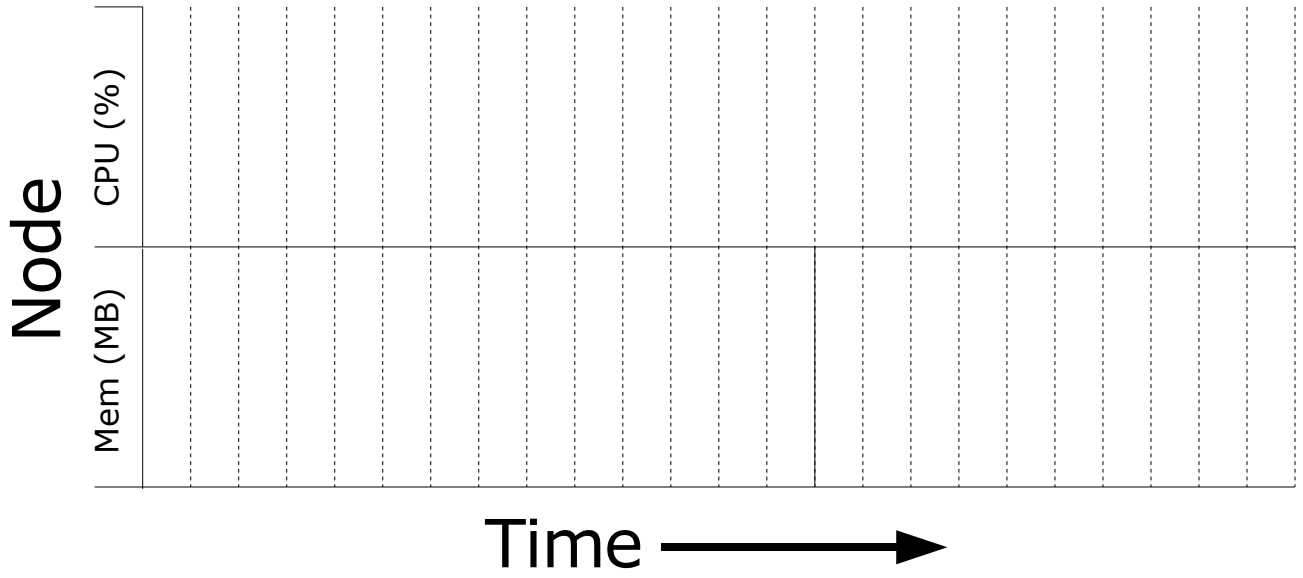
Model (I)

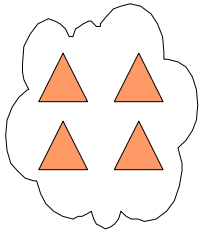
- We propose a model where the virtual resources are seen as *resource slots*.



Model (II)

- Physical nodes are empty resource slots where the virtual resources are mapped to.





Mem: 512 MB

CPU: 10%

Mem: 512 MB

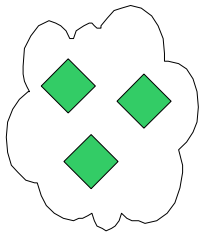
CPU: 10%

Mem: 512 MB

CPU: 10%

Mem: 512 MB

CPU: 10%



Mem: 1024 MB

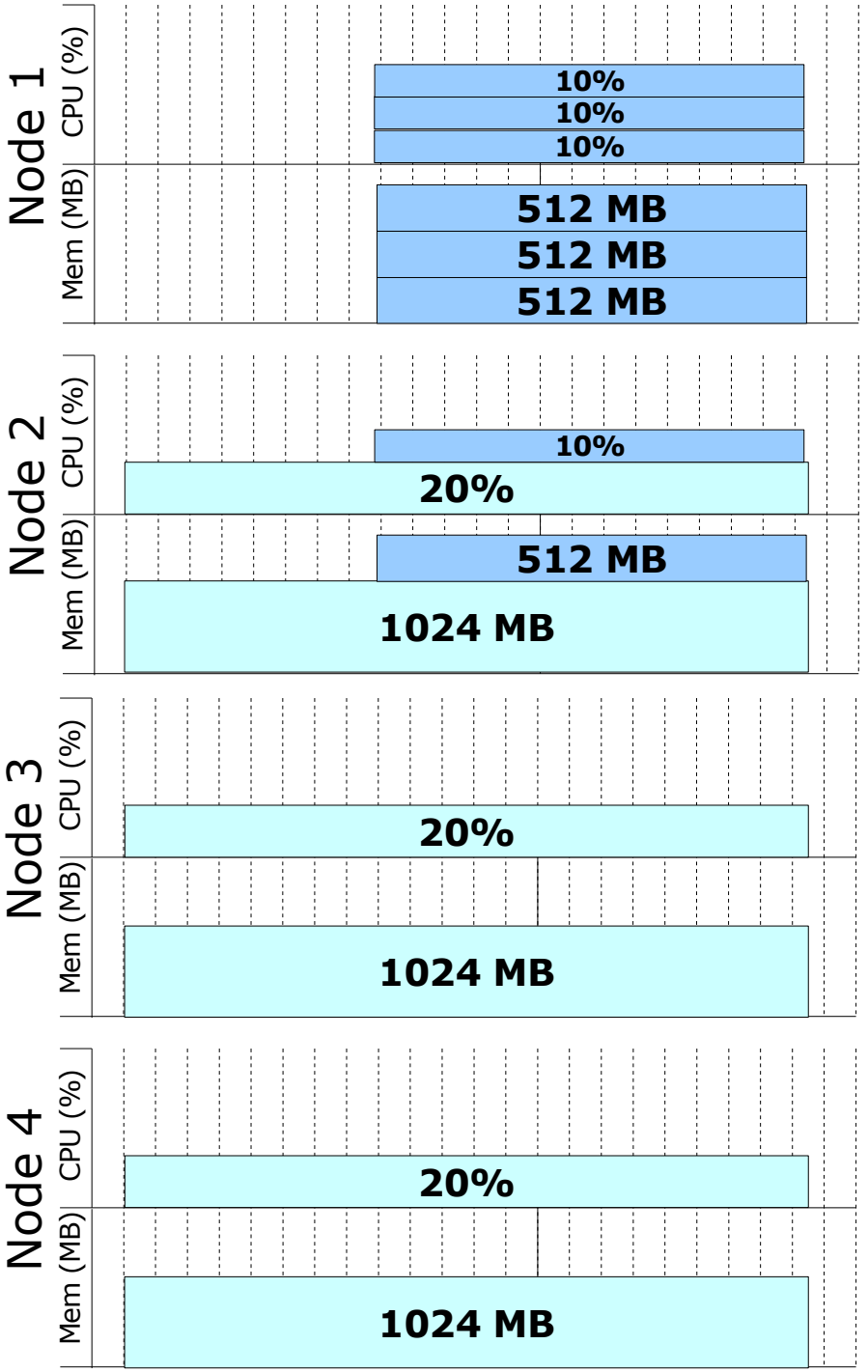
CPU: 20%

Mem: 1024 MB

CPU: 20%

Mem: 1024 MB

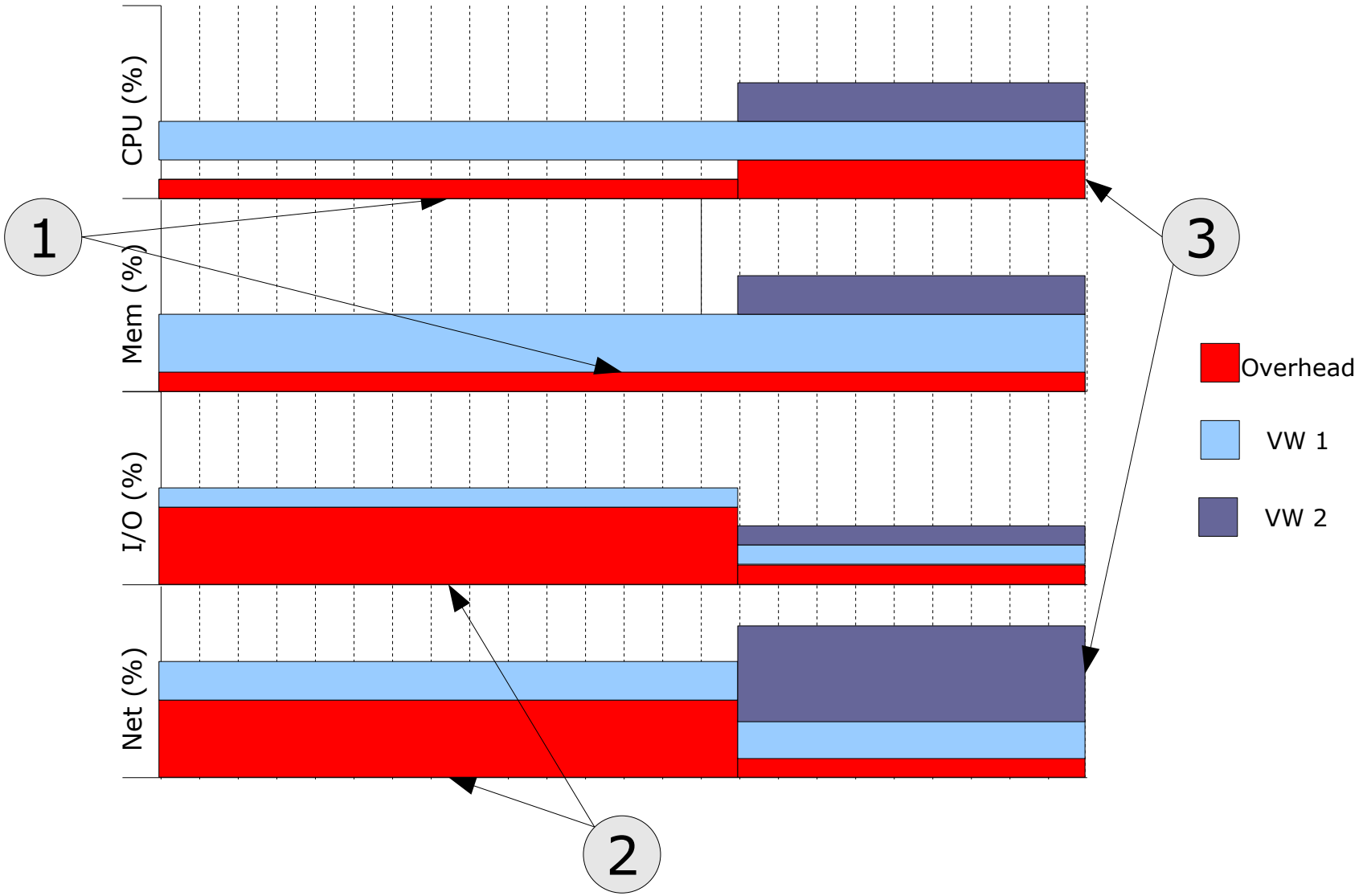
CPU: 20%



Model (III)

- However, this model doesn't account for overhead.
  - Other virtual workspace implementations downplay the importance of overhead.
  - We hold that an adequate overhead management can result in higher performance.
- Two types of overhead:
  - VM Hypervisor overhead
  - Scheduling activities

Model (IV)



Scheduling (I)

- The centerpiece of our system will be the scheduler.
- Scheduler must:
  - Perform admission control
    - Policies
    - Is request feasible?
  - Map virtual resources to physical resources
  - Manage execution
  - React to changes
    - Resource allocation renegotiations
    - Failures

Scheduling (II)

- The main challenges in designing and developing this scheduler are:
  - Managing overhead
  - Mapping virtual resources to physical resources
  - Handling changes in the system

Scheduling (III)

- Building the entire scheduler is a huge undertaking.
- We are currently focusing on specific problems, and making certain assumptions.
  - We will gradually deal with all scheduling scenarios, with as few assumptions as possible.

Index

- Problem and Status
- Scheduling Virtual Workspaces
- Roadmap

Roadmap

- What we're working on right now
  - Overhead: Staging VW images to the nodes where they're needed.
  - Scheduler that only considers CPU and memory as apportionable resources.
  - Experiments
- What we'll work on next
  - More powerful scheduler (capable of allocating network and disk bandwidth)
  - Resource allocation renegotiation
  - Leveraging live migration of VMs to perform load balancing.

Questions?

Borja Sotomayor
University of Chicago
Department of Computer Science
borja@cs.uchicago.edu

Bibliography

- Cluster-On-Demand + Shirako (Duke University)
  - "*Sharing Networked Resources with Brokered Leases*", David Irwin, Jeff Chase, Laura Grit, Aydan Yumerefendi, David Becker, and Ken Yocum, USENIX Technical Conference, June 2006, Boston, Massachusetts.
  - "*Adaptive Virtual Machine Hosting with Brokers*". Laura Grit, Jeff Chase, David Irwin, Aydan Yumerefendi. Submitted to Supercomputing'06.
  - "*Toward a Doctrine of Containment: Grid Hosting with Adaptive Resource Control*". Lavanya Ramakrishnan, Laura Grit, Anda Iamnitchi, David Irwin, Aydan Yumerefendi, Jeff Chase. Submitted to OSDI (OS Design and Implementation) '06.
  - http://www.cs.duke.edu/nicl/cod/
  - http://www.cs.duke.edu/nicl/cereus/shirako.html

Bibliography

- Violin + VioCluster (Purdue)
  - Paul Ruth, Junghwan Rhee, Dongyan Xu, Rick Kennell, Sebastien Goasguen, "*Autonomic Live Adaptation of Virtual Computational Environments in a Multi-Domain Infrastructure*", ICAC'06
  - Paul Ruth, Phil McGachey, Dongyan Xu, "*VioCluster: Virtualization for Dynamic Computational Domains*", Proceedings of the IEEE International Conference on Cluster Computing (Cluster'05), Boston, MA, September 2005.
  - Paul Ruth, Xuxian Jiang, Dongyan Xu, Sebastien Goasguen, "*Virtual Distributed Environments in a Shared Infrastructure*", IEEE Computer, Special Issue on Virtualization Technologies, May 2005.
  - http://www.cs.purdue.edu/homes/ruth/violin/index.html

Bibliography

- In-VIGO (UFlorida)
  - Adabala, Sumalatha; Chadha, Vineet; Chawla, Puneet; Figueiredo, Renato; Fortes, Jose; Krsul, Ivan; Matsunaga, Andrea; Tsugawa, Mauricio; Zhang, Jian; Zhao, Ming; Zhu, Liping; Zhu, Xiaomin '*From Virtualized Resources to Virtual Computing Grids: The In-VIGO System*'. In Future Generation Computer Systems, vol 21, no. 6, April, 2005. DOI:10.1016/j.future.2003.12.021.
  - Matsunaga, Andrea , M. Tsugawa, S. Adabala, R. Figueiredo, H. Lam and J. Fortes '*Science gateways made easy: the In-VIGO approach*'. In Workshop on Science Gateways, Global Grid Forum, 06/2005
  - https://www.acis.ufl.edu/~acis/ivwiki/index.php/Main_Page

Bibliography

- Virtuoso (Northwestern)
  - R. Figueiredo, P. Dinda, J. Fortes, *Resource Virtualization Renaissance*, (Guest Editors' Introduction to the IEEE Computer Special Issue On Resource Virtualization), May, 2005.
  - B. Lin, and P. Dinda, *VSched: Mixing Batch and Interactive Virtual Machines Using Periodic Real-time Scheduling*, Proceedings of ACM/IEEE SC 2005 (Supercomputing), November, 2005.
  - A. Sundararaj, M. Sanghi, J. Lange, P. Dinda, *An Optimization Problem in Adaptive Virtual Environments*, Proceedings of the Seventh Workshop on Mathematical Performance Modeling and Analysis (MAMA 2005).
  - Zhao, Ming , J. Zhang, R. Figueiredo '*Distributed File System Virtualization Techniques Supporting On-Demand Virtual Machine Environments for Grid Computing*'. In Cluster Computing Journal, 9(1) (to appear), 01/2006

Bibliography

- Maestro-VC (UIUC)
  - Nadir Kiyanclar, Gregory A. Koenig, William Yurcik. *Maestro-VC: A paravirtualized Execution Environment for Secure On-Demand Cluster Computing*. CCGrid'06.
- Fine-grained resource allocation enforcement
  - Gupta, Diwaker; Cherkasova, Ludmila; Gardner, Rob; Vahdat, Amin, *Enforcing Performance Isolation Across Virtual Machines in Xen*. Hewlett-Packard Tech Report HPL-2006-77.

Bibliography

- Virtual Workspaces (Globus)
  - *Virtual Clusters for Grid Communities.* I.Foster, T.Freeman, K.Keahey, D.Scheftner, B.Sotomayor, X.Zhang. CCGrid 2006
  - *Division of Labor: Tools for Growth and Scalability of Grids*. I.Foster, T.Freeman, K.Keahey, A.Rana, B.Sotomayor, F.Wuerthwein. ANL/MCS-P1316-0106
  - *Virtual Workspaces: Achieving Quality of Service and Quality of Life in the Grid*, Keahey, K., I. Foster, T. Freeman, and X. Zhang. Accepted for publication in the Scientific Programming Journal, 2006
  - *Virtual Workspaces in the Grid*, Keahey, K., I. Foster, T. Freeman, X. Zhang, D. Galron. Europar 2005, Lisbon, Portugal, September, 2005.
  - *From Sandbox to Playground: Dynamic Virtual Environments in the Grid*, Keahey, K., K. Doering, and I. Foster. 5th International Workshop in Grid Computing (Grid 2004), Pittsburgh, PA, November 2004.