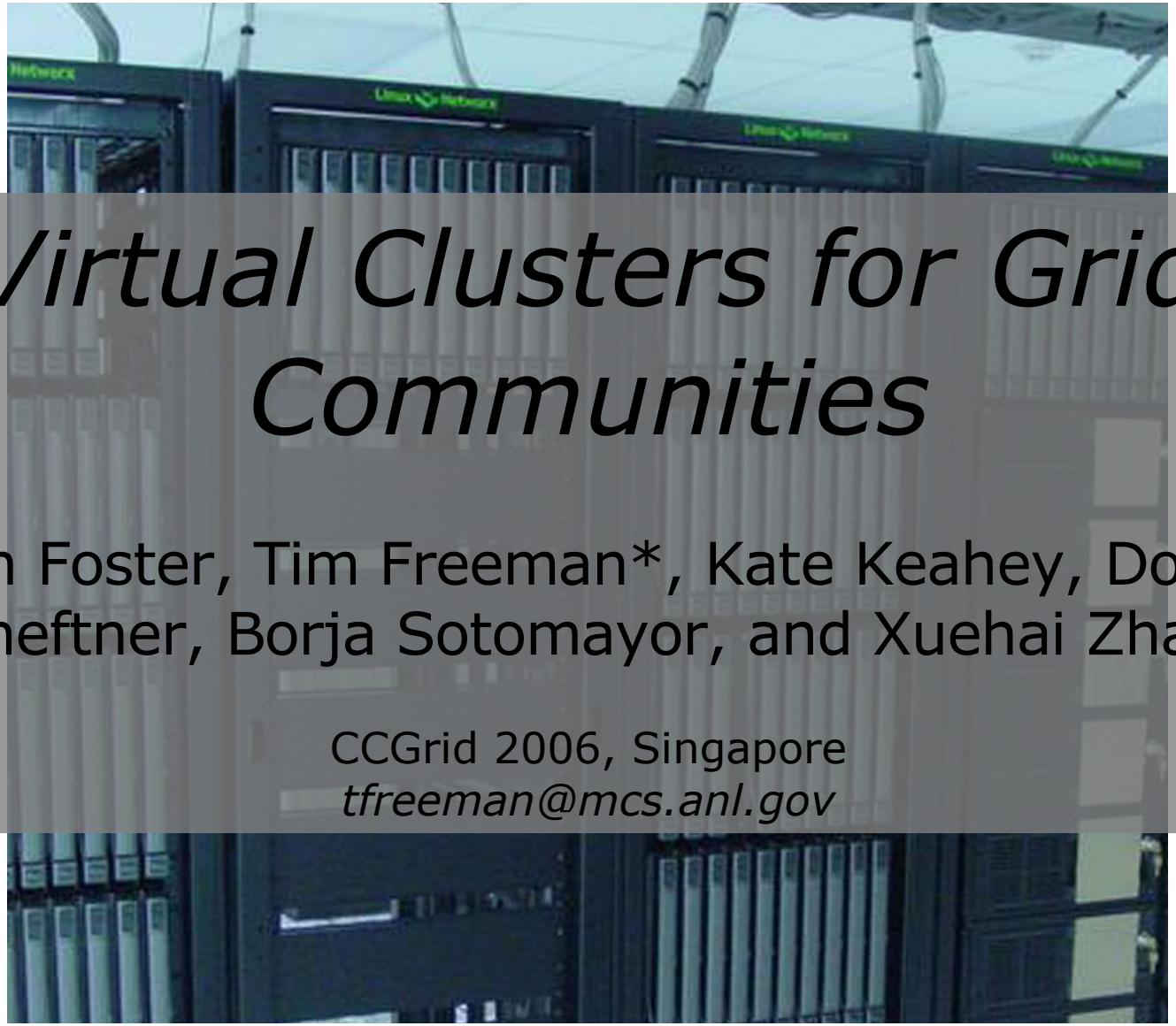
the globus alliance
www.globus.org



Virtual Clusters for Grid Communities

Ian Foster, Tim Freeman*, Kate Keahey, Doug Scheftner, Borja Sotomayor, and Xuehai Zhang

CCGrid 2006, Singapore

tfreeman@mcs.anl.gov

- Introduction & Motivation
- Workspace Basics
- Virtual Machine Implementation
- Virtual Cluster Workspaces
  - Problem Statement
  - Workspace Deployment: Metadata/Allocation
  - Aggregate Metadata
  - Aggregate Resource Allocations
  - Experimental Results
- Analysis
- Ongoing and Future Work

Workspaces: Motivation

Required Environments

- Diverse client environment requirements
  - Library versions
  - Application versions
  - Custom applications (with possibly complex installs)
  - OS type, version, modules

Required Environments

- Diverse client environment requirements
  - Library versions
  - Application versions
  - Custom applications (with possibly complex installs)
  - OS type, version, modules

vs.

- Provider constraints
  - Security policies
  - Administrator time

the globus alliance
www.globus.org

TERAGRID

Applications Software

Home > View by Category

The following is a list of installed software by category. Click
collapse the menu tree. You may view detailed information ab
name.

Expand to 2nd Level | Expand All | Collapse All

- Applications - Scientific and Engineering
- Benchmark and Example Programs
- Data Analysis and Visualization
- Distributed Processing Tools
- Mathematics and Statistics
- Numerical Programs and Routines
  - ARPACK
  - ATLAS
  - ESSL
  - gmp
  - GNU Scientific Library (GSL)
  - GOTO
  - gsl
  - LAPACK
  - MKL - Math Kernel Library
  - PESSL
  - Graph and Mesh Partitioning
  - Linear Algebra
  - Miscellaneous
- Parallel Processing Tools
- Performance Evaluation

Open Science Grid

Operating Systems

The instructions assume you're using one of the following Linux distributions:

- Red Hat 7.x
- Red Hat 9.0
- Red Hat Enterprise Linux 3
- Fedora Core 3
- Debian Linux 3.1 (Sarge)

Installation may be successful with other Linux distributions, but they have not been tested
"binary-compatible" distributions such as Scientific Linux Fermi 3.0.x (x=3,4) and Rocks 3.x
OSG-ITB treating them as Red Hat Enterprise Linux 3 but no support is implied.

eGee
Enabling Grids
for E-sciencE

DOCUMENTS

Requirements related documents:

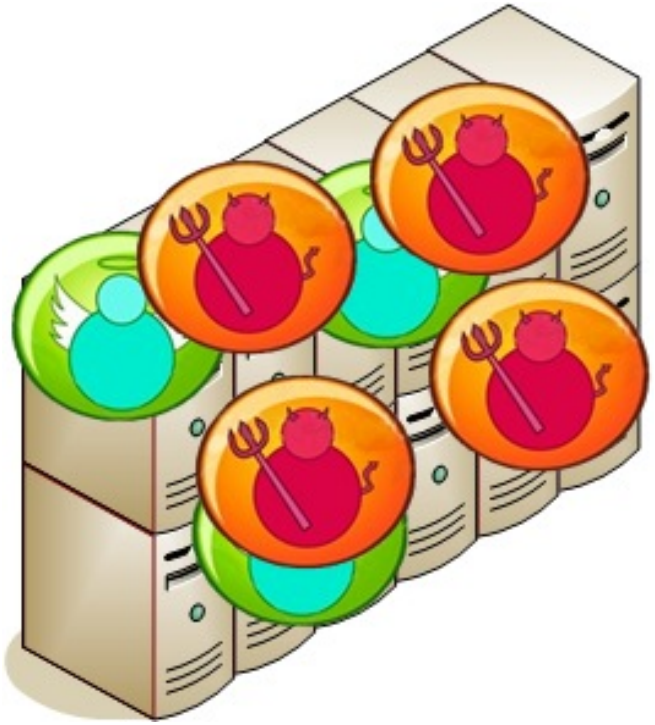
- EDG Application Working Group documents: **Joint list of usecases** and **recommandations**.
- **Usecases for HEP Common Application Layer** (HEPCAL) document.
- EDG WP10 (biomedical) **requirements** (see section 6, page 42) and **key improvements needed** (see section 5.4, page 46).
- EDG WP9 (earth observation) **deliverable on EDG testbed evaluation** (see section 5.3 p. 67 and section 5.4 p. 73) and **generic applications questionaire**.
- **Biomedical application requirements**

Isolation, Trust and Accounting

"demo":

Isolation, Trust and Accounting

"demo":



Isolation, Trust and Accounting

Use Cases

- Scientific Gateways
- Educational resources
- **Example:** Flex cluster
  - Simulation code
    - Runs for weeks
    - OK to preempt
  - Pulse data analysis
    - Runs for minutes
    - Time critical

Workspaces

Workspaces

- **Sandbox:** isolates clients/providers from one another
- **Execution environment** is captured in a workspace
  - Physical workspaces
  - Virtual machine workspaces
  - Pre-deployed: dynamic accounts
- **Resource allocation**
  - Client and provider enter into an *agreement*
- **Dynamic**
  - Client deploys workspace "into" resource allocation
  - Provider allows workspace management/inspection

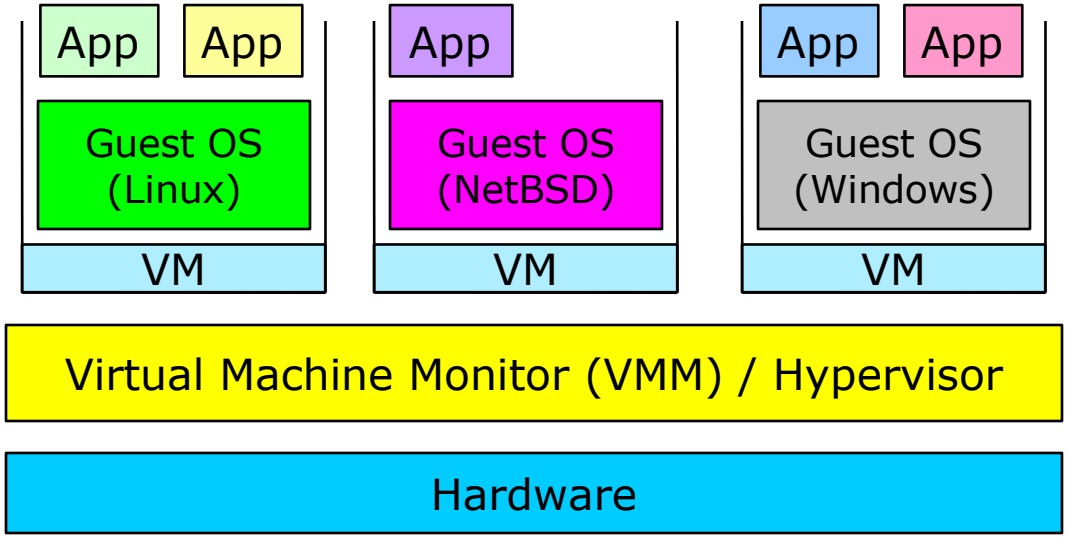
Workspaces

- **Sandbox:** isolates clients/providers from one another
- **Dynamic**
  - Client deploys workspace "into" resource allocation
  - Provider allows workspace management/inspection

A provisioned computing "capsule" whose internals can be managed by the client

Workspaces: VM implementation

Virtual Machine Basics



- A VM can serialize all of its state (including RAM)
- A VM image is simply a collection of files
  - Disk partitions, RAM, configuration information
  - Image can be easily moved (*migrated*) between hypervisors of the same type
  - Image can also be saved and used for rollbacks

Virtual Machines

- **Isolation**
  - Security enforced at hypervisor layer
  - Fine grain (alterable) resource allocations
- Flexible **control** and accounting for site
- Customization: **any software** (including legacy)
- Client can have administrator privileges
- Site software requirements reduced to VMM
- **Performance** overhead is becoming acceptable
  - Currently support Xen (studies: *within 5%*)
  - Experimented with VMware in the past

Virtual Cluster Workspaces

Problem Statement

- Building virtual clusters
  - Can we automate configuring cluster topologies, networking patterns, and sharing mechanisms?
  - How can we optimize virtual cluster deployment?

Problem Statement

- Building virtual clusters
  - Can we automate configuring cluster topologies, networking patterns, and sharing mechanisms?
  - How can we optimize virtual cluster deployment?

What are the logistics and cost of virtual cluster deployment?

Problem Statement

- Building virtual clusters
  - Can we automate configuring cluster topologies, networking patterns, and sharing mechanisms?
  - How can we optimize virtual cluster deployment?

What are the logistics and cost of virtual cluster deployment?

- Using virtual clusters
  - What is the overhead of running applications of different profiles on a virtual cluster?
  - When is this cost acceptable?

Problem Statement

- Building virtual clusters
  - Can we automate configuring cluster topologies, networking patterns, and sharing mechanisms?
  - How can we optimize virtual cluster deployment?

What are the logistics and cost of virtual cluster deployment?

- Using virtual clusters
  - What is the overhead of running applications of different profiles on a virtual cluster?
  - When is this cost acceptable?

Can applications use virtual cluster efficiently?

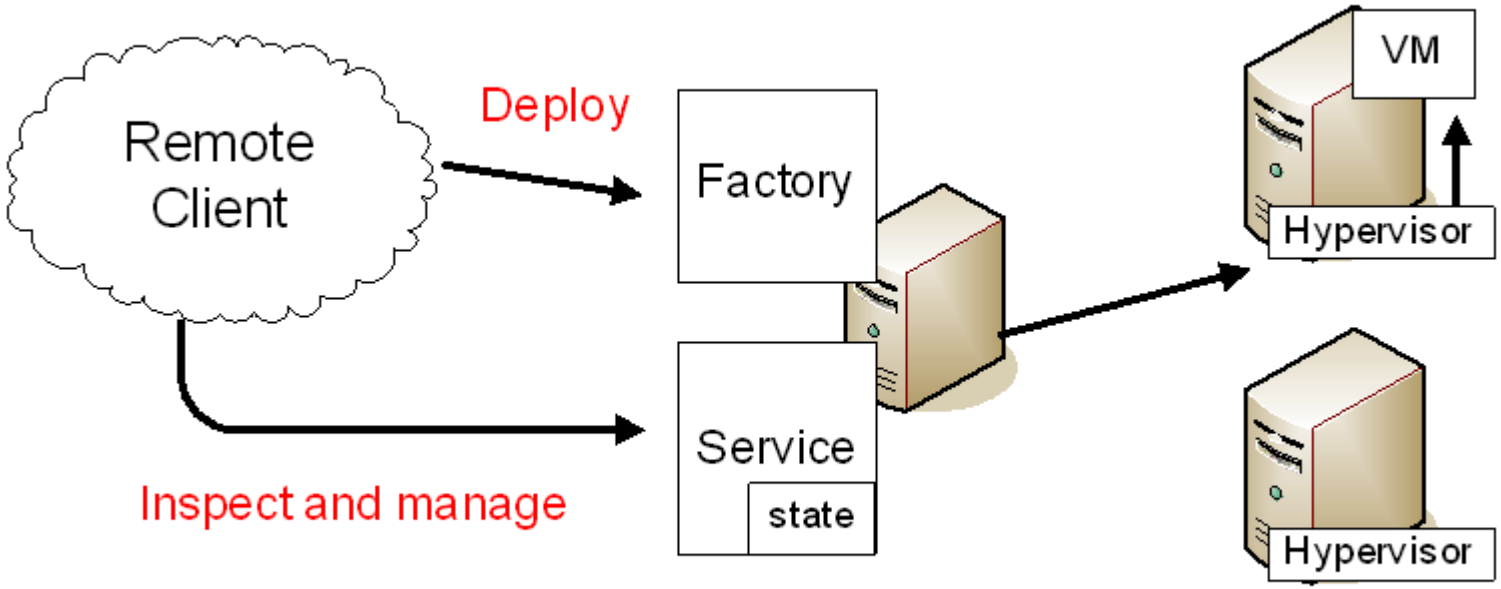
Problem Statement

- Building virtual clusters
  - Can we automate configuring cluster topologies, networking patterns, and sharing mechanisms?
  - [illegible]
- Us
  - [illegible]
  di
  - When is this cost acceptable?

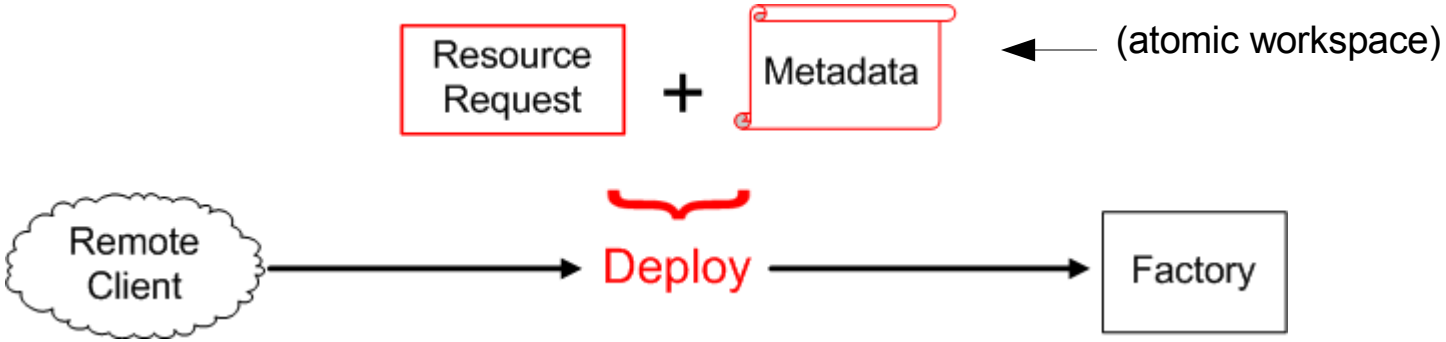
What is the best way of using virtual clusters given application types and problem sizes?

Can applications use virtual cluster efficiently?

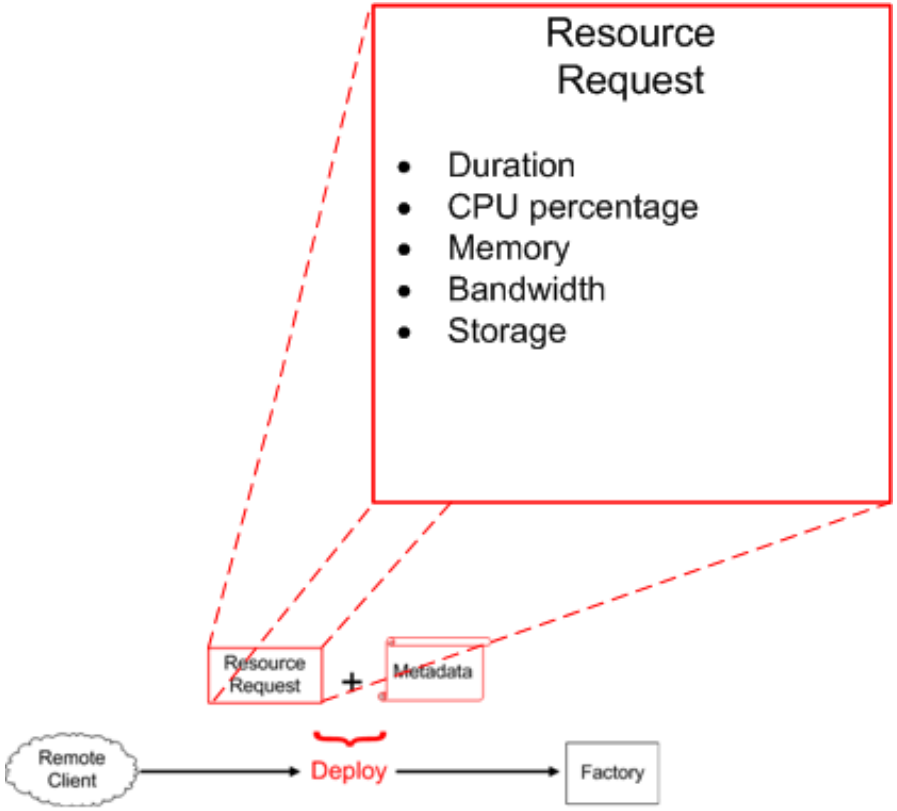
Workspace Service



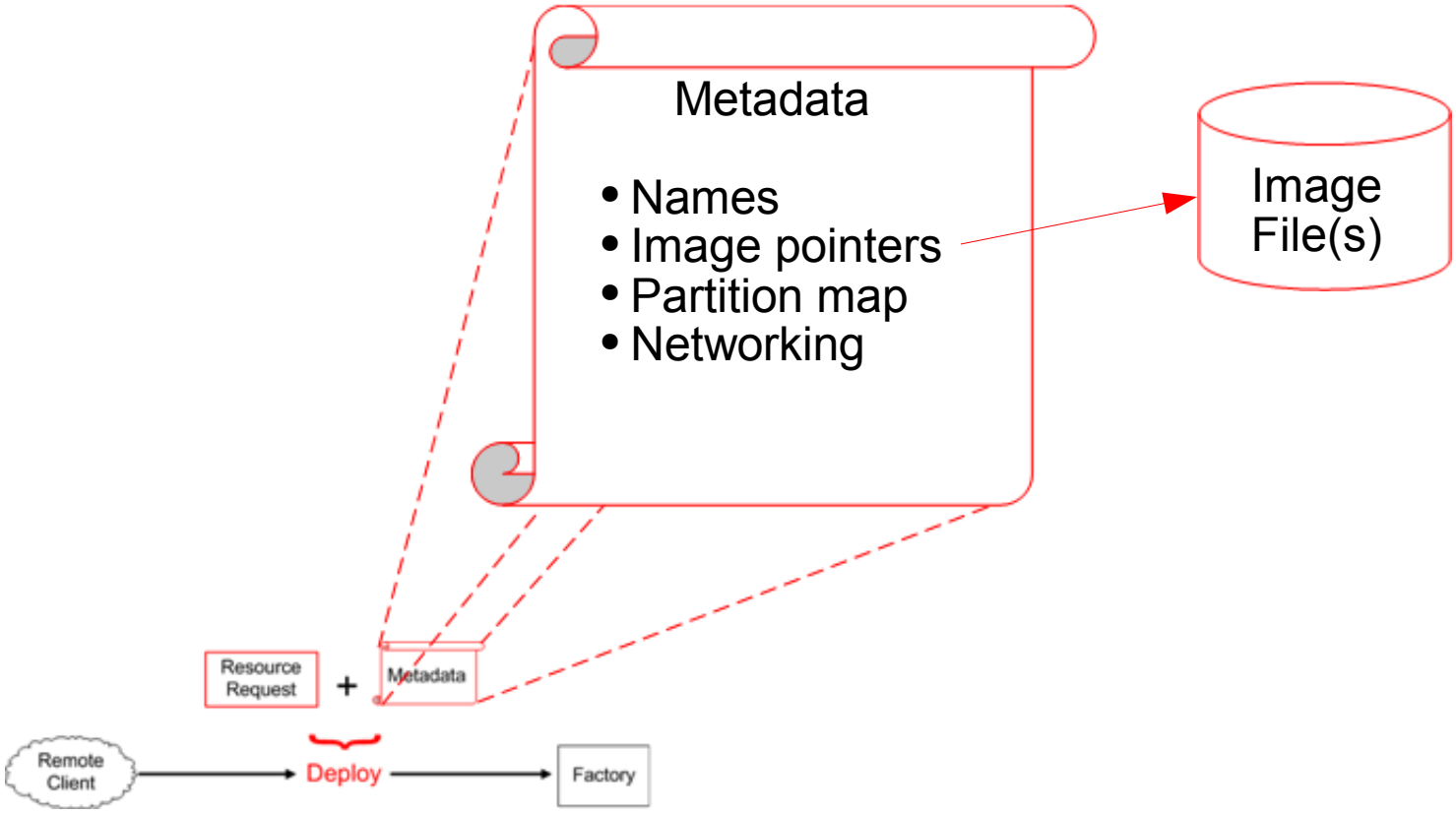
Workspace Service



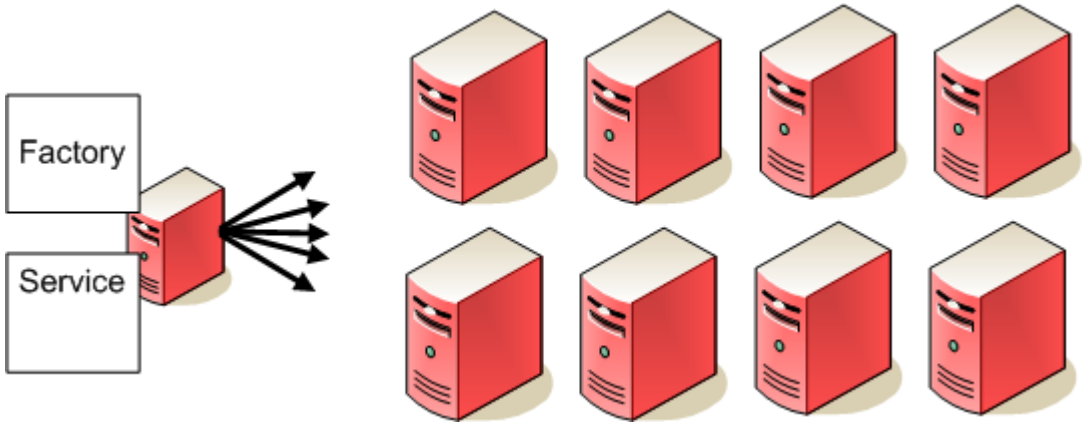
Workspace Service



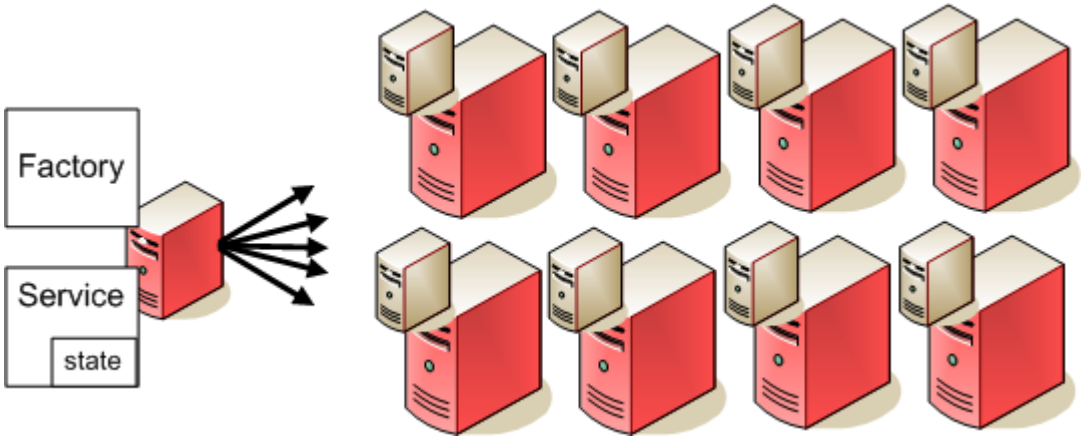
Workspace Service



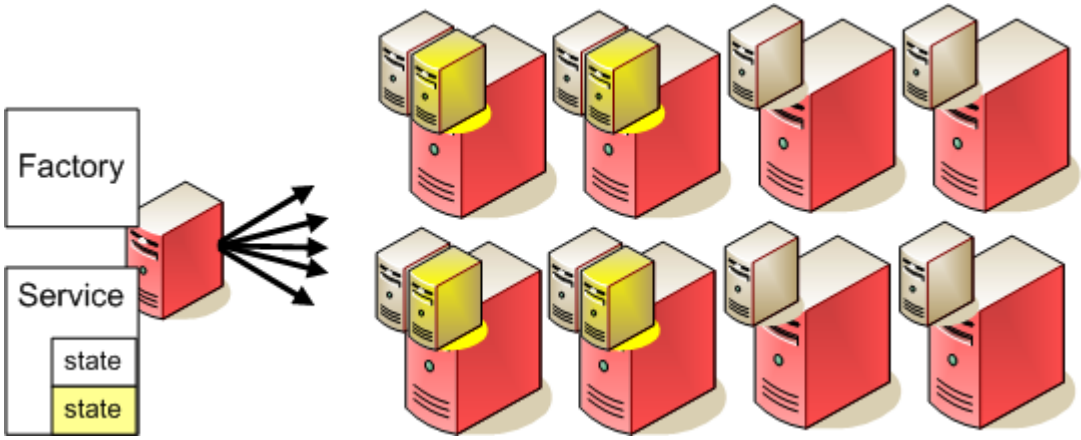
Virtual Clusters 'Demo'



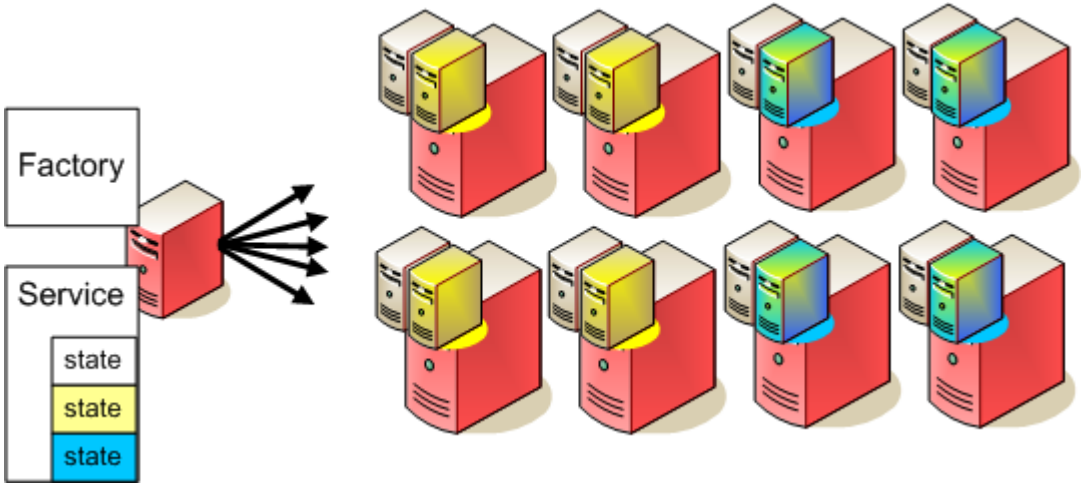
Virtual Clusters 'Demo'



Virtual Clusters 'Demo'



Virtual Clusters 'Demo'



How should we represent clusters?

- Deployed *as a whole*
- Issues:
  - Disk per compute node would be costly
    - Image sharing
  - Network coherence
  - Configurations for service coherence
  - Efficient deployment mechanisms

How should we represent clusters?

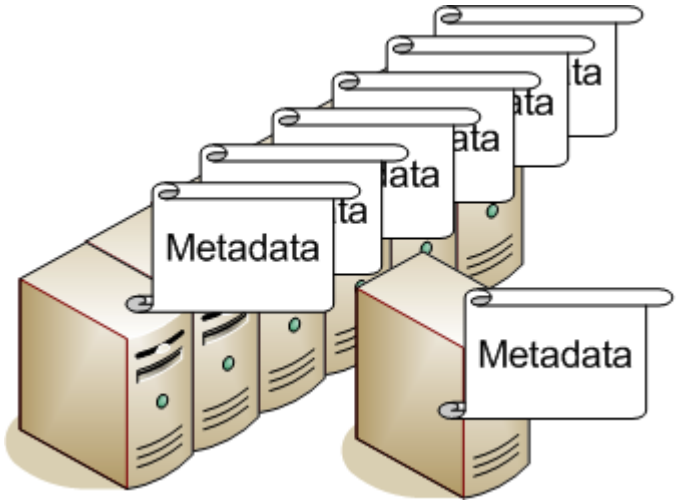
A simple, common virtual cluster

Virtual compute nodes

Virtual head node

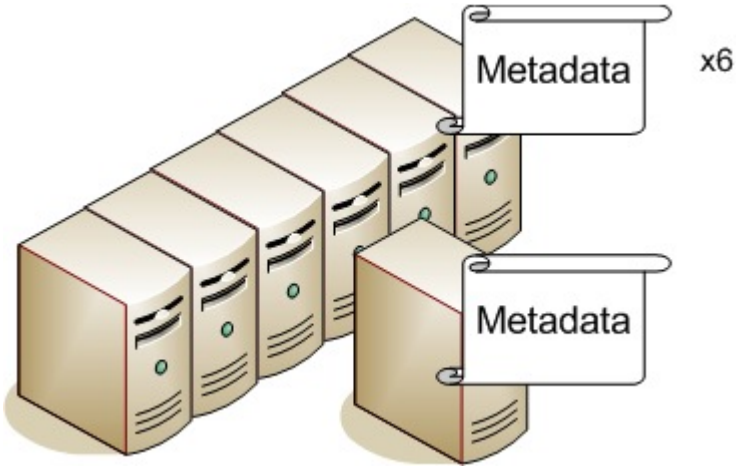
How should we represent clusters?

Explicitly?



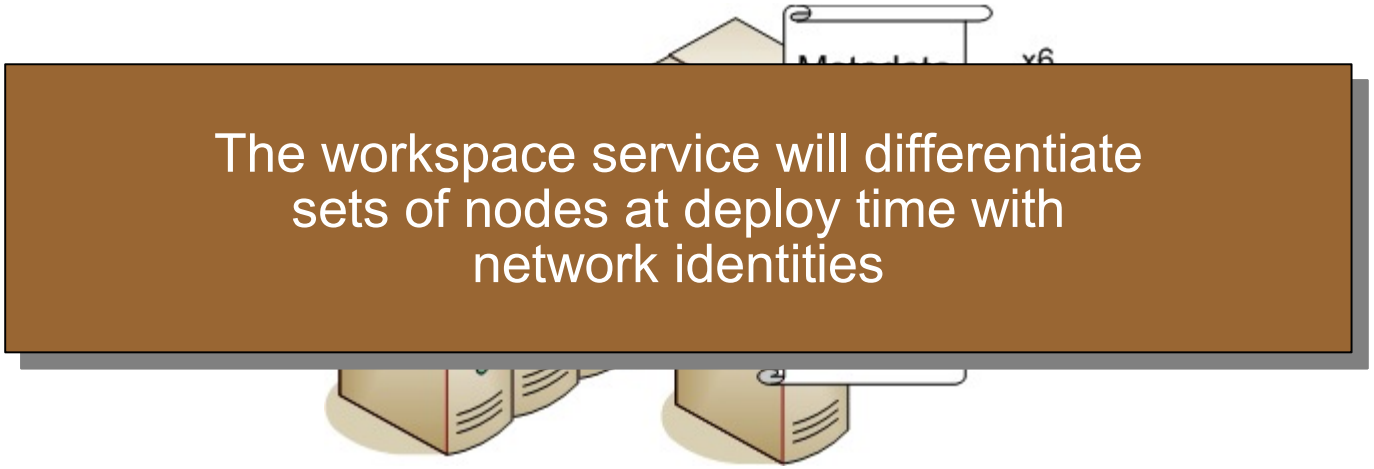
How should we represent clusters?

Consolidation into sets



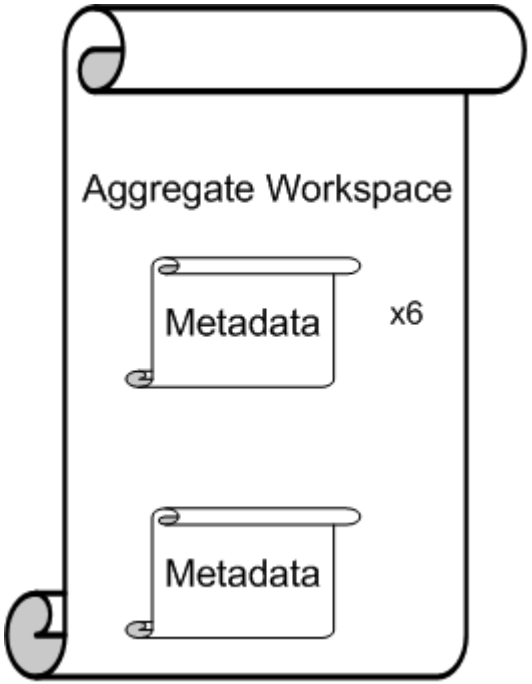
How should we represent clusters?

Consolidation into sets



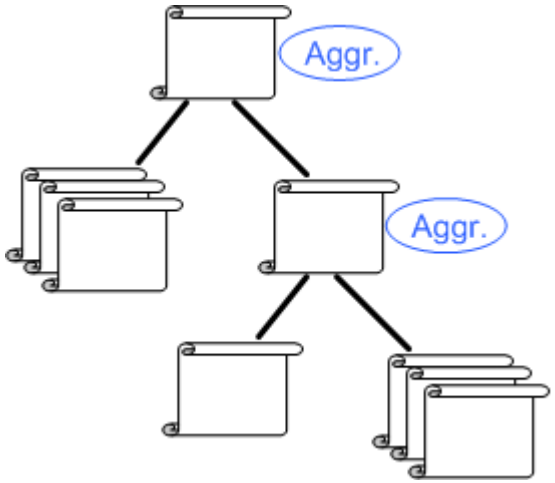
Aggregate Workspaces

- Composition of atomics
- Atomic is set of one



Aggregate Workspaces

- Composition of atomics
- Atomic is set of one
- Aggregates can contain other aggregates
- A **tree** structure



Aggregate Workspaces

Recalling the issues

- Deployed *as a whole*
- Network coherence is possible
- Configurations for service coherence performed by the workspace service
- Well defined and shared image parts allow for efficient deployment mechanisms

How should we map clusters to resources?

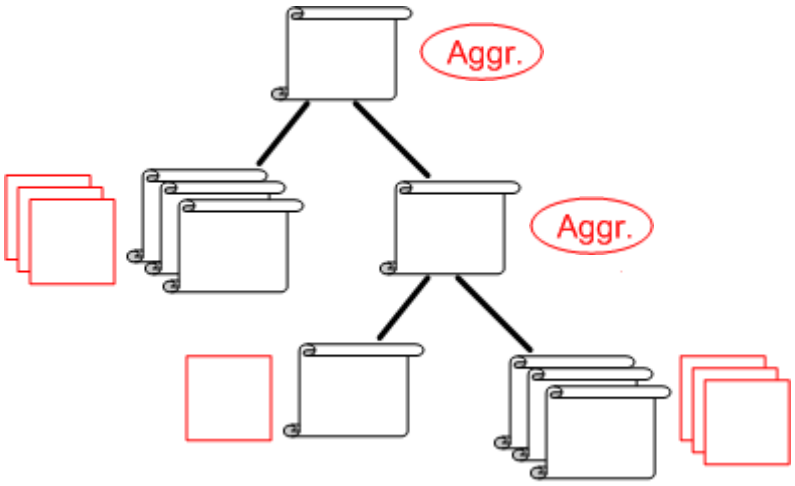
- Problem is tied to representation
- Issues:
  - Some nodes may need different allocations
  - Many nodes will need identical allocations
  - Entire allocation must be dealt with *as a whole*

Aggregate Resource Allocation

- Similar to aggregate workspace
- A **tree** structure
  - Does **not need to match metadata topology**
  - Heterogeneous
  - Aggregate allocation can be changed, signed, pointed to (e.g., for WS-Agreement) a*s a whole*

Aggregate Resource Allocation

One atomic R.A. per atomic workspace is possible ...

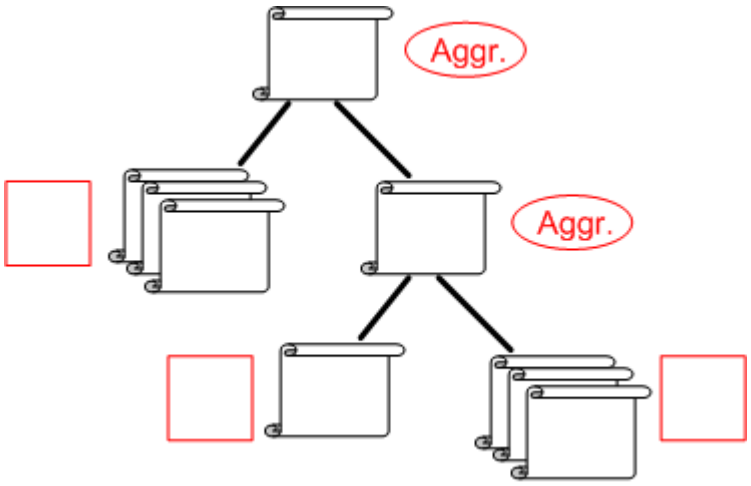


Aggregate Resource Allocation

One atomic R.A. per atomic workspace is possible ...

... but not required.

Heterogeneous configurations do not imply R.A. also needs to be heterogeneous.

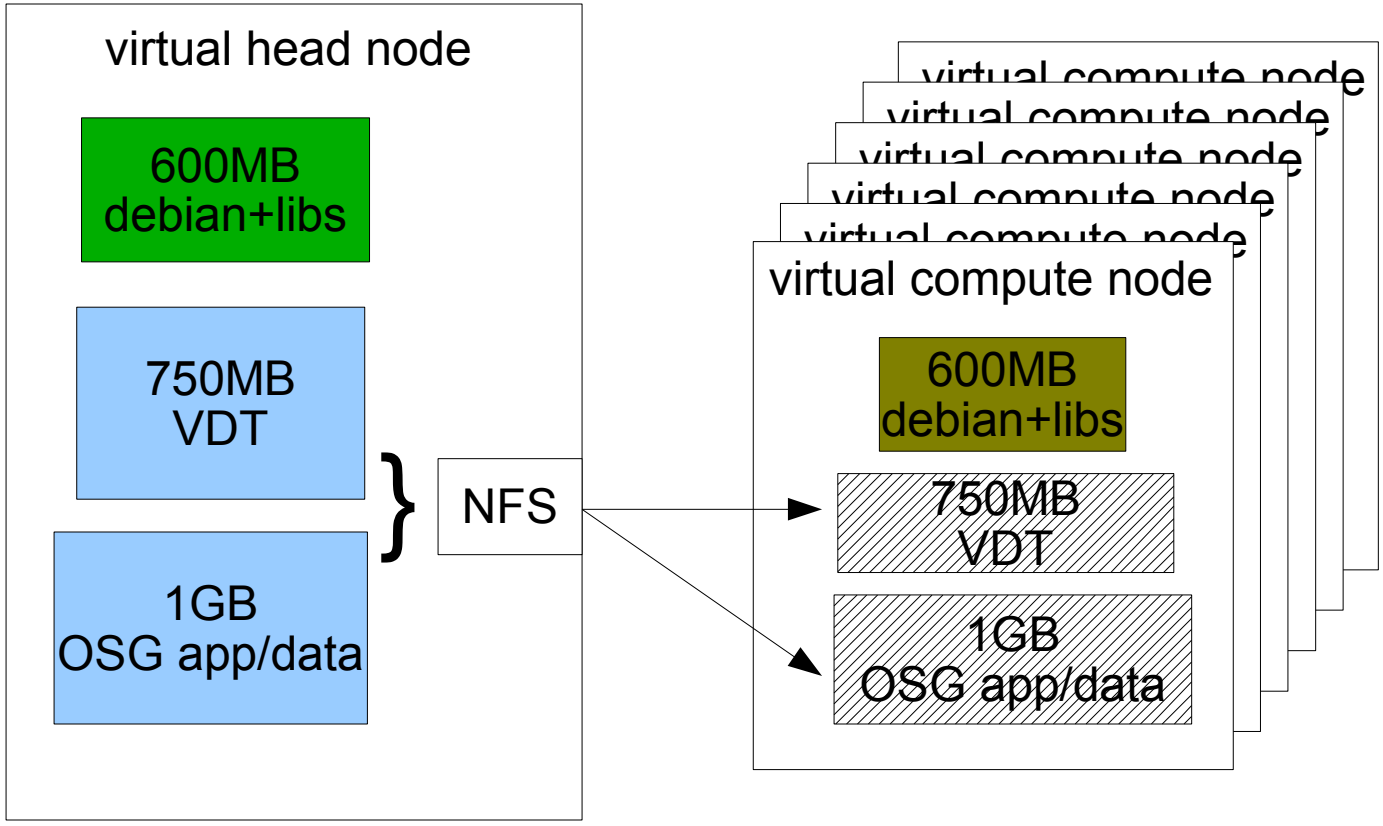


Experiments

A Virtual OSG Cluster

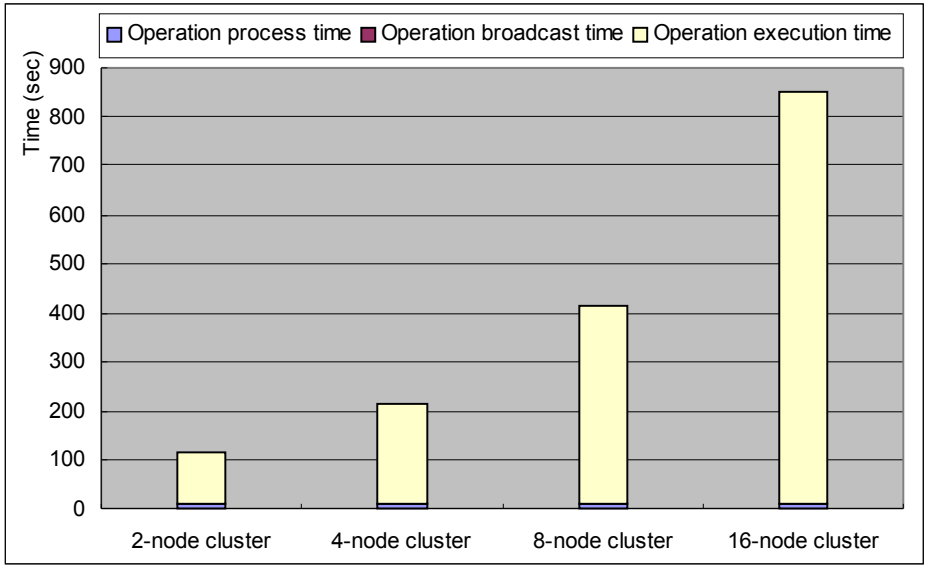
- Experimented with a real world example
- OSG requirements
  - Debian Linux 3.1 (Sarge)
  - A local batch scheduler, such as Condor, PBS, LSF or SGE
  - All service and compute nodes have access to NFS
  - Grid infrastructure, typically GRAM and GridFTP
  - Submit host (not part of virtual cluster)
  - VDS: Pegasus, DAGMan, and Condor-G

A Virtual OSG Cluster

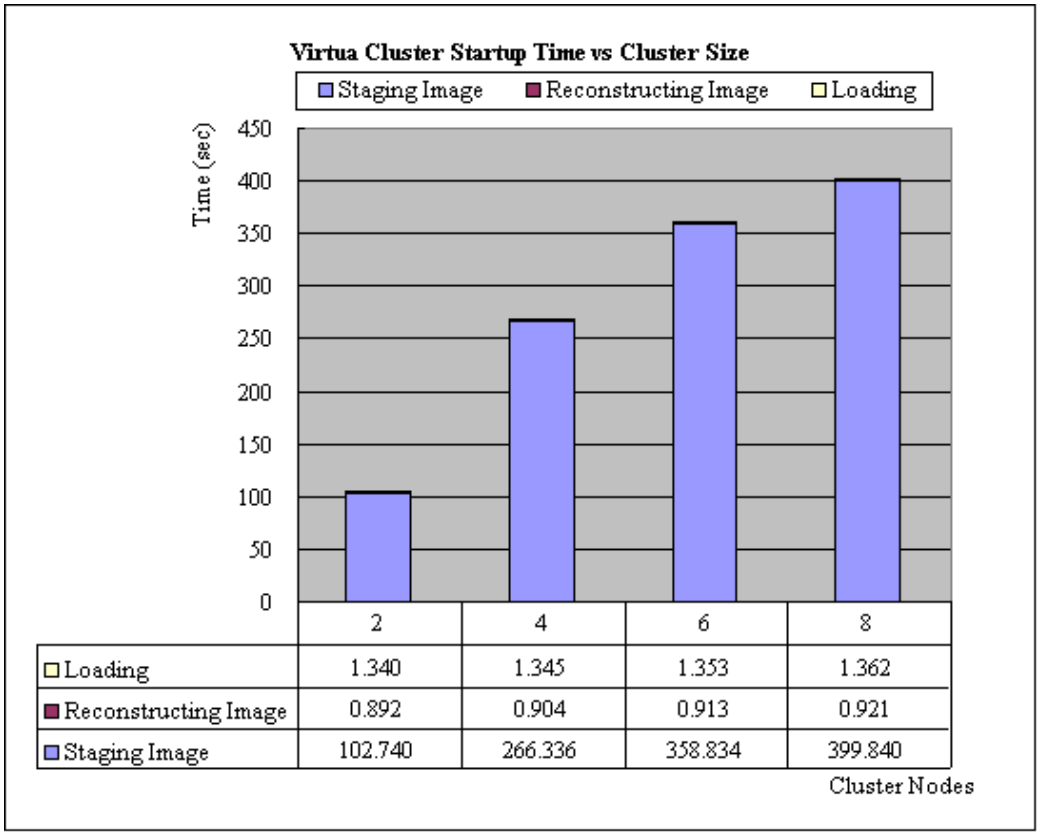


Propagation

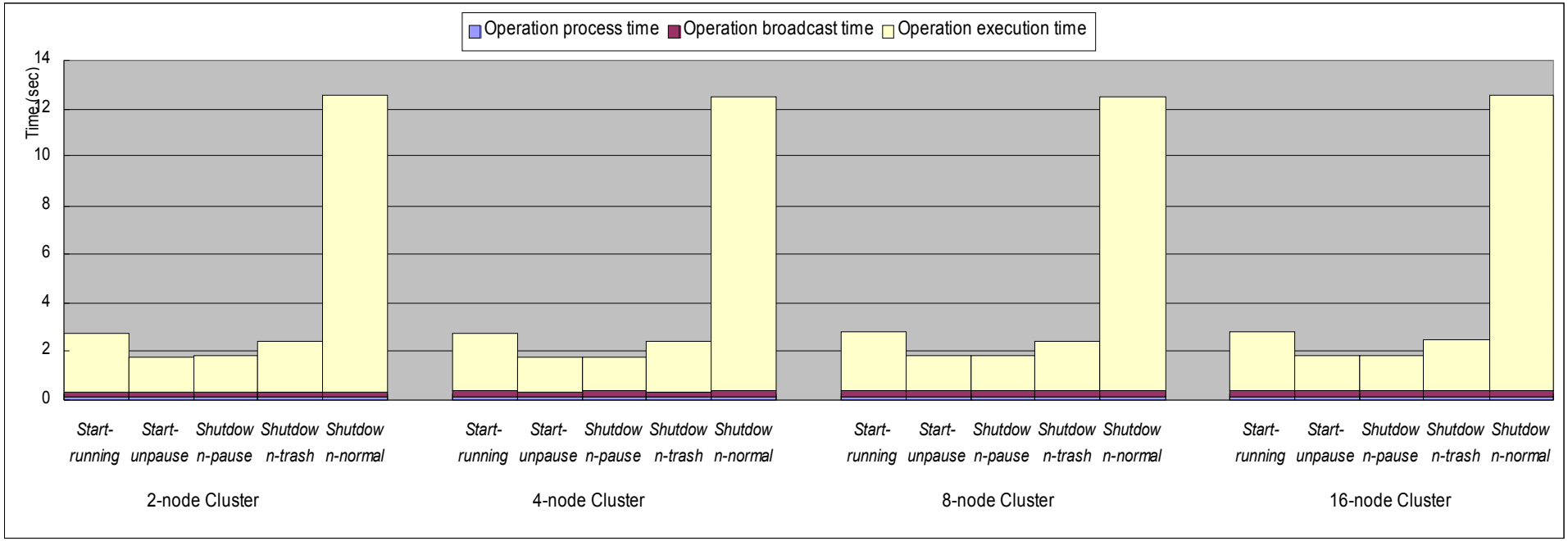
- Images are staged to physical cluster nodes
  - (The GigE effect)
  - Any transport method possible, “nfs copy” data:



Customization

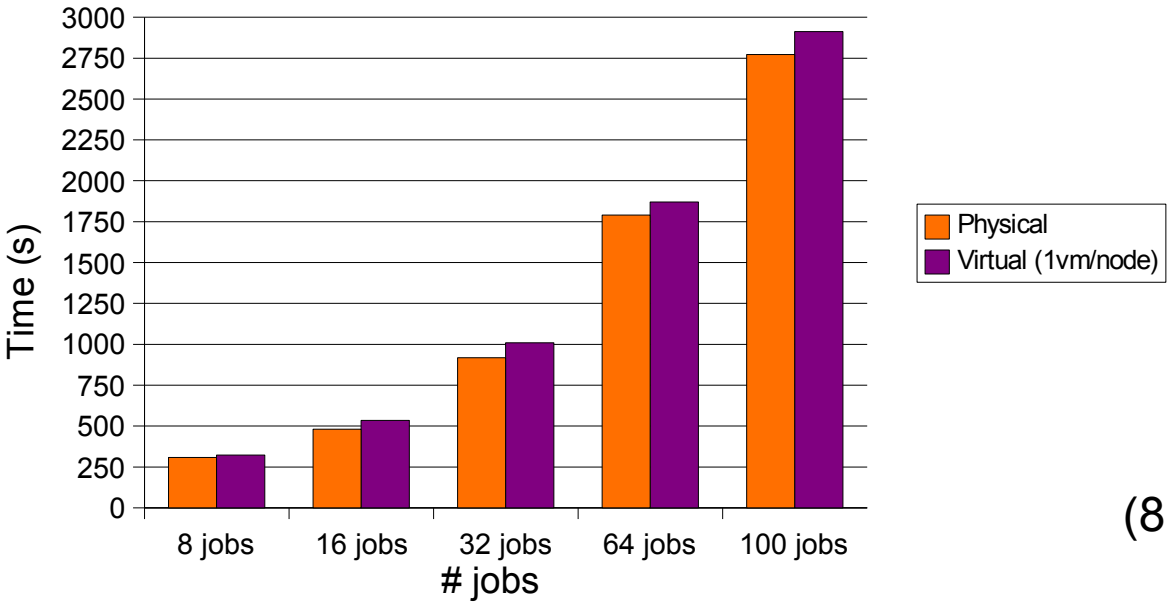


Management


Operation process time
Operation broadcast time
Operation execution time
Time (sec)
14
12
10
8
6
4
2
0
Start-running
Start-unpause
Shutdown-pause
Shutdown-trash
Shutdown-normal
2-node Cluster
4-node Cluster
8-node Cluster
16-node Cluster

Application Performance

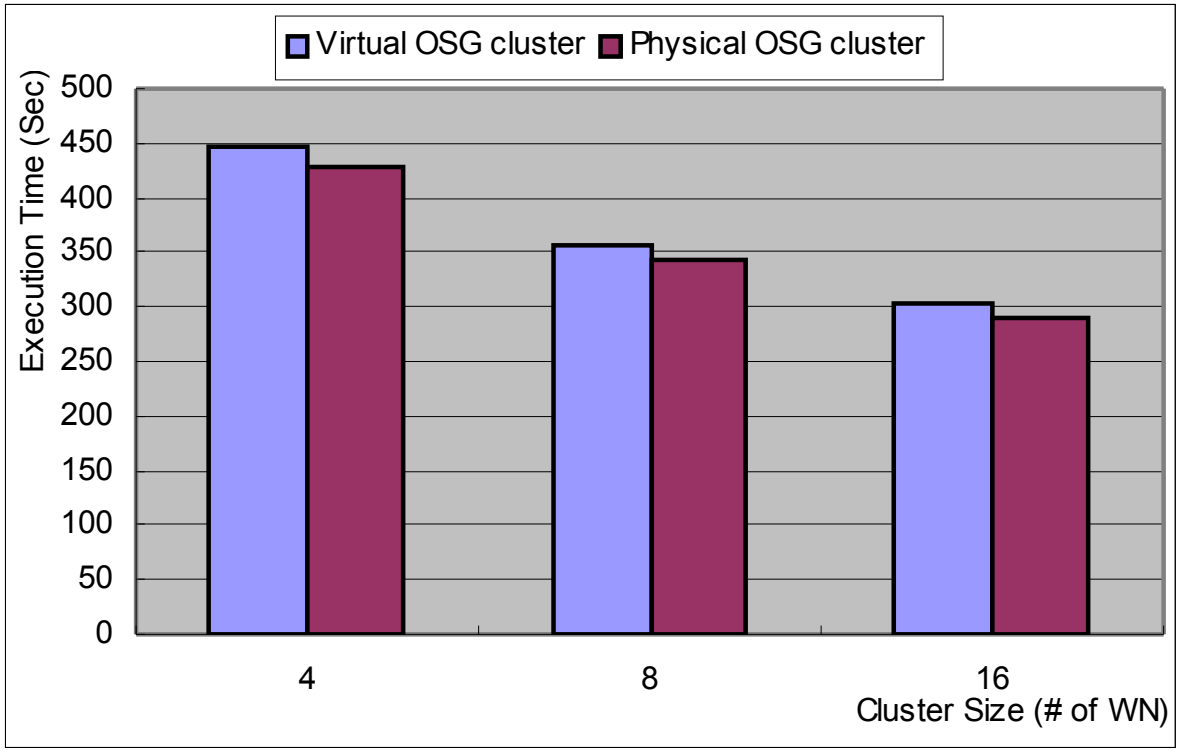
- BLAST - Embarrassingly parallel



(8 nodes)

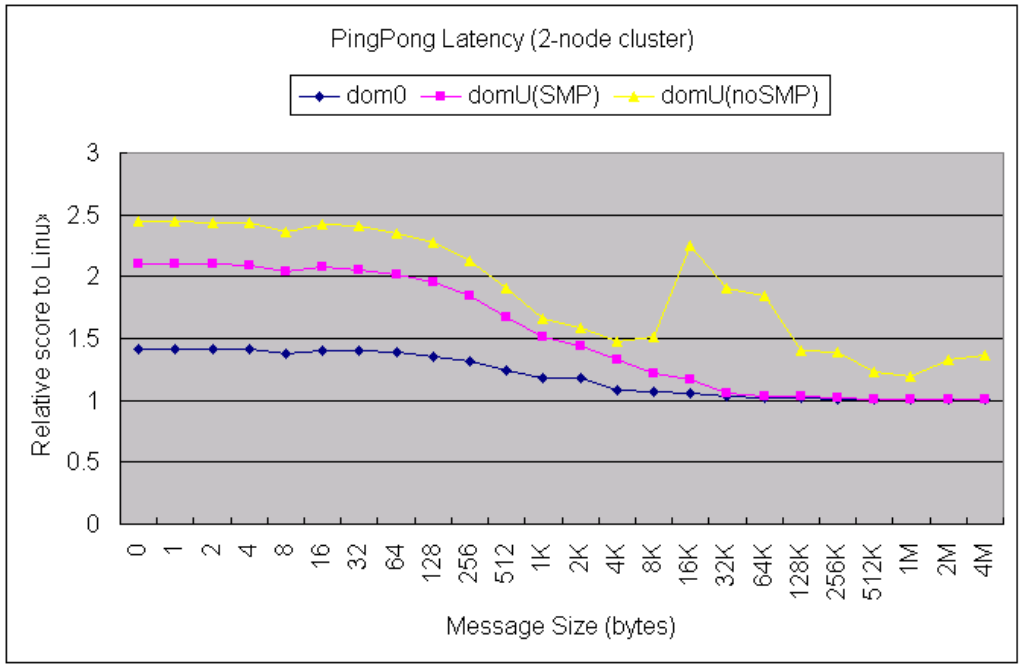
Application Performance

- FOAM: MPI but communication roughly 10%



Application Performance

- MPI study: http://people.cs.uchicago.edu/~hai/vcluster/PMB/
- More dominate communication patterns show problems such as latency issues from interrupt queueing



Analysis

- File staging can be expensive
  - Optimizations discussed earlier
- Network latency may be an issue for some HPC applications
- Aggregate workspace abstraction can handle flexible topologies and resource requirements
- Workspace service handles network and other coherence issues

Analysis

- Management overhead can be offset with longer running or shared virtual clusters
- Both management and performance expenses offset by the inherent advantages of workspaces:
  - Hosting several applications at once
  - Quality of service, isolation
  - Ease of contributing nodes to a grid
  - Flexibility

Ongoing and Future Work

- Resource management issues
  - Fine grained resource allocation
  - Complex scheduling use cases
  - WS-Agreement
  - Economic modelling
- Deploying VMs securely
  - Identity/Networking issues
- Building and deploying entire ***virtual grids***

Thankyou

http://workspace.globus.org

- » Code
- » Documentation
- » Support (mailing lists)
- » Publications