

On-demand Virtualization and Grid/Cloud Integration

D. Salomoni, INFN Tier-1 Computing Manager

`Davide.Salomoni@cnaif.infn.it`

INFN-CNAF

EGEE '09, 21/9/2009



- 1 Motivation and Architectural Design
 - Virtualization of a Farm Cluster
 - Extensions of the WNoD architecture
 - Architectural Design
- 2 Current Status and Perspectives
 - Farm Virtualization
 - Distributed Submission
 - Cloud Services
 - Future Work

Outline

1 Motivation and Architectural Design

- Virtualization of a Farm Cluster
- Extensions of the WNoD architecture
- Architectural Design

2 Current Status and Perspectives

- Farm Virtualization
- Distributed Submission
- Cloud Services
- Future Work

Some Common Requests

- Operating System:
 - "I want SLC3". *"Hey, no, my application [only runs / is only certified] on Ubuntu 9.04."* "What? Forget it! I need `afs` and SL5."
 - "Would you please upgrade all your worker nodes to a 64-bit O/S *by this week?*"
- Applications:
 - "I *absolutely* need you to install application X.Y version Z on all your nodes."
 - "Please don't change *that* system library!" ("and don't you dare to upgrade the kernel!")
- Intra-VO requirements may also apply, e.g. different sets of users belonging to the same VO may have different needs.
- The INFN Tier-1 currently supports ~ 20 Virtual Organizations.
 - Optimize resource usage, e.g. avoid static allocations, try not to waste CPU cycles, do not [buy | set-up] dedicated infrastructures to fix requirements/problems.
 - Maintain **full control of the site**.
 - Do not change established workflows.

Some Common Requests

- Operating System:
 - "I want SLC3". *"Hey, no, my application [only runs | is only certified] on Ubuntu 9.04."* "What? Forget it! I need `afs` and SL5."
 - "Would you please upgrade all your worker nodes to a 64-bit O/S *by this week?*"
- Applications:
 - "I *absolutely* need you to install application X.Y version Z on all your nodes."
 - "Please don't change *that* system library!" ("and don't you dare to upgrade the kernel!")
- Intra-VO requirements may also apply, e.g. different sets of users belonging to the same VO may have different needs.
- The INFN Tier-1 currently supports ~20 Virtual Organizations.
 - Optimize resource usage, e.g. avoid static allocations, try not to waste CPU cycles, do not [buy | set-up] dedicated infrastructures to fix requirements/problems.
 - Maintain **full control of the site**.
 - Do not change established workflows.

Some Common Requests

- Operating System:
 - "I want SLC3". *"Hey, no, my application [only runs | is only certified] on Ubuntu 9.04."* "What? Forget it! I need `afs` and SL5."
 - "Would you please upgrade all your worker nodes to a 64-bit O/S *by this week?*"
- Applications:
 - "I *absolutely* need you to install application X.Y version Z on all your nodes."
 - "Please don't change *that* system library!" ("and don't you dare to upgrade the kernel!")
- Intra-VO requirements may also apply, e.g. different sets of users belonging to the same VO may have different needs.
- The INFN Tier-1 currently supports ~20 Virtual Organizations.
 - Optimize resource usage, e.g. avoid static allocations, try not to waste CPU cycles, do not [buy | set-up] dedicated infrastructures to fix requirements/problems.
 - Maintain **full control of the site**.
 - Do not change established workflows.

Some Common Requests

- Operating System:
 - "I want SLC3". *"Hey, no, my application [only runs | is only certified] on Ubuntu 9.04."* "What? Forget it! I need `afs` and SL5."
 - "Would you please upgrade all your worker nodes to a 64-bit O/S *by this week?*"
- Applications:
 - "I *absolutely* need you to install application X.Y version Z on all your nodes."
 - "Please don't change *that* system library!" ("and don't you dare to upgrade the kernel!")
- Intra-VO requirements may also apply, e.g. different sets of users belonging to the same VO may have different needs.
- The INFN Tier-1 currently supports ~ 20 Virtual Organizations.
 - Optimize resource usage, e.g. avoid static allocations, try not to waste CPU cycles, do not [buy | set-up] dedicated infrastructures to fix requirements/problems.
 - Maintain **full control of the site**.
 - Do not change established workflows.

Use Virtual Machines as Compute Nodes

Making them available *on-demand*, i.e. when needed

The main use case initially considered was to provide **dynamic virtual execution environments**: virtual *Worker Nodes on Demand* (WNoD).

- Decouple installed software from physical hardware.
- VMs can run different images (e.g., different O/S, or same O/S but with different package sets) for different (sets of) users.
- VMs are created *on-demand*, i.e. at job execution time, so resources are always dynamically shared.
- VM images are initially retrieved from shared storage (gpfs at the INFN Tier-1) and then stored locally.
- VM images are **never** obtained directly by users, and are always under the control of the resource provider.
- WNoD *does not alter* user procedures for job submission.

Use Virtual Machines as Compute Nodes

Making them available *on-demand*, i.e. when needed

The main use case initially considered was to provide **dynamic virtual execution environments**: virtual *Worker Nodes on Demand* (WNoD).

- Decouple installed software from physical hardware.
- VMs can run different images (e.g., different O/S, or same O/S but with different package sets) for different (sets of) users.
- VMs are created *on-demand*, i.e. at job execution time, so resources are always dynamically shared.
- VM images are initially retrieved from shared storage (gpfs at the INFN Tier-1) and then stored locally.
- VM images are **never** obtained directly by users, and are always under the control of the resource provider.
- WNoD *does not alter* user procedures for job submission.

Use Virtual Machines as Compute Nodes

Making them available *on-demand*, i.e. when needed

The main use case initially considered was to provide **dynamic virtual execution environments**: virtual *Worker Nodes on Demand* (WNoD).

- Decouple installed software from physical hardware.
- VMs can run different images (e.g., different O/S, or same O/S but with different package sets) for different (sets of) users.
- VMs are created *on-demand*, i.e. at job execution time, so resources are always dynamically shared.
- VM images are initially retrieved from shared storage (*gpfs* at the INFN Tier-1) and then stored locally.
- VM images are **never** obtained directly by users, and are always under the control of the resource provider.
- WNoD *does not alter* user procedures for job submission.

Use Virtual Machines as Compute Nodes

Making them available *on-demand*, i.e. when needed

The main use case initially considered was to provide **dynamic virtual execution environments**: virtual *Worker Nodes on Demand* (WNoD).

- Decouple installed software from physical hardware.
- VMs can run different images (e.g., different O/S, or same O/S but with different package sets) for different (sets of) users.
- VMs are created *on-demand*, i.e. at job execution time, so resources are always dynamically shared.
- VM images are initially retrieved from shared storage (`gpfs` at the INFN Tier-1) and then stored locally.
- VM images are **never** obtained directly by users, and are always under the control of the resource provider.
- WNoD *does not alter* user procedures for job submission.

Use Virtual Machines as Compute Nodes

Making them available *on-demand*, i.e. when needed

The main use case initially considered was to provide **dynamic virtual execution environments**: virtual *Worker Nodes on Demand* (WNoD).

- Decouple installed software from physical hardware.
- VMs can run different images (e.g., different O/S, or same O/S but with different package sets) for different (sets of) users.
- VMs are created *on-demand*, i.e. at job execution time, so resources are always dynamically shared.
- VM images are initially retrieved from shared storage (`gpfs` at the INFN Tier-1) and then stored locally.
- VM images are **never** obtained directly by users, and are always under the control of the resource provider.
- WNoD *does not alter* user procedures for job submission.

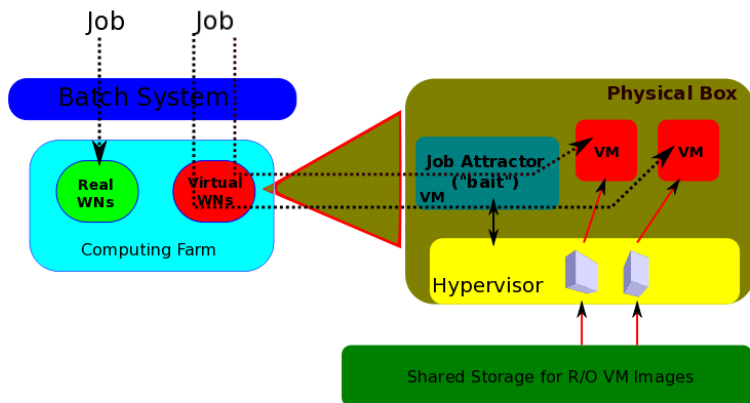
Use Virtual Machines as Compute Nodes

Making them available *on-demand*, i.e. when needed

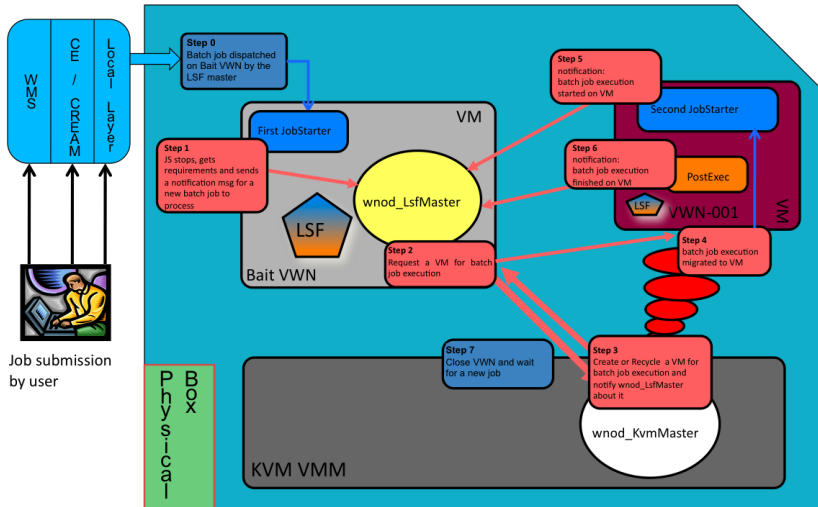
The main use case initially considered was to provide **dynamic virtual execution environments**: virtual *Worker Nodes on Demand* (WNoD).

- Decouple installed software from physical hardware.
- VMs can run different images (e.g., different O/S, or same O/S but with different package sets) for different (sets of) users.
- VMs are created *on-demand*, i.e. at job execution time, so resources are always dynamically shared.
- VM images are initially retrieved from shared storage (`gpfs` at the INFN Tier-1) and then stored locally.
- VM images are **never** obtained directly by users, and are always under the control of the resource provider.
- WNoD *does not alter* user procedures for job submission.

Worker Nodes on Demand High-Level Overview



The Worker Nodes on Demand Process Flow



Outline

1 Motivation and Architectural Design

- Virtualization of a Farm Cluster
- **Extensions of the WNoD architecture**
- Architectural Design

2 Current Status and Perspectives

- Farm Virtualization
- Distributed Submission
- Cloud Services
- Future Work

Dynamic Selection of Services beyond Local Site Resources

WNoD provides a solution to dynamically manage computing environments for local or grid jobs in a strictly site-specific fashion. It is interesting to extend this mechanism to:

- Selection of VM images through **standard Grid jobs** to select custom (e.g., per-VO, per-role) resources.
- Provisioning of **cloud-like services**, e.g. of standard computing cores accessible via regular `ssh`.

Dynamic Selection of Services beyond Local Site Resources

WNoD provides a solution to dynamically manage computing environments for local or grid jobs in a strictly site-specific fashion. It is interesting to extend this mechanism to:

- Selection of VM images through **standard Grid jobs** to select custom (e.g., per-VO, per-role) resources.
- Provisioning of **cloud-like services**, e.g. of standard computing cores accessible via regular `ssh`.

Outline

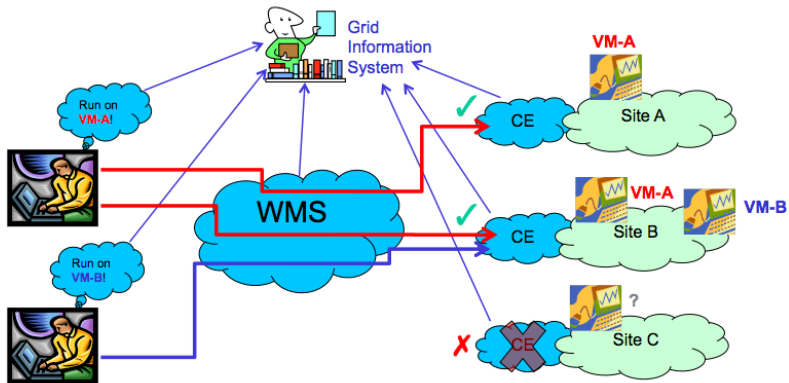
1 Motivation and Architectural Design

- Virtualization of a Farm Cluster
- Extensions of the WNoD architecture
- **Architectural Design**

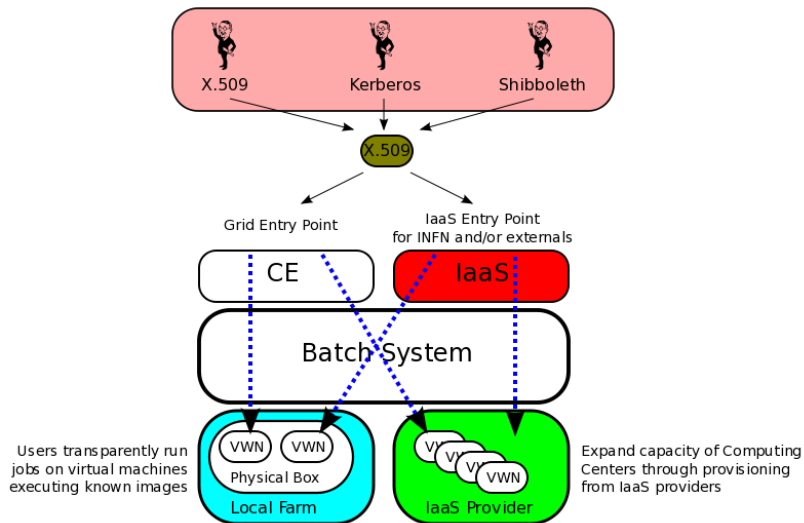
2 Current Status and Perspectives

- Farm Virtualization
- Distributed Submission
- Cloud Services
- Future Work

Allow Dynamic Selection of Resources through Standard Grid Jobs



Make Grid and Cloud Services Converge around Dynamically Provisioned Resources



Outline

1 Motivation and Architectural Design

- Virtualization of a Farm Cluster
- Extensions of the WNoD architecture
- Architectural Design

2 Current Status and Perspectives

- **Farm Virtualization**
- Distributed Submission
- Cloud Services
- Future Work

WNoD Deployment

WNoD are being **deployed in production** at the INFN Tier-1:

- More than 400k test jobs run. Currently ca. 130 VMs.
- Regular SAM tests running successfully on virtual WN, on SL5.3.
- ETICS running on WNoD only, on custom (i.e. ETICS-specific) SL5.3 image.
- ATLAS and CDF transparently running on both traditional and virtual WN, on SLC4.4.
- CMS are testing SL5.3 on WNoD right now. ATLAS, Alice and LHCb will start tests of SL5.3 shortly.
- Thanks to WNoD, migration to SL5 is as flexible as it can be: tell us when you're ready, and we'll direct you to the images you want, without static allocations.
- Monitoring and accounting already supporting WNoD.

WNoD Deployment

WNoD are being **deployed in production** at the INFN Tier-1:

- More than 400k test jobs run. Currently ca. 130 VMs.
- Regular SAM tests running successfully on virtual WN, on SL5.3.
- ETICS running on WNoD only, on custom (i.e. ETICS-specific) SL5.3 image.
- ATLAS and CDF transparently running on both traditional and virtual WN, on SLC4.4.
- CMS are testing SL5.3 on WNoD right now. ATLAS, Alice and LHCb will start tests of SL5.3 shortly.
- Thanks to WNoD, migration to SL5 is as flexible as it can be: tell us when you're ready, and we'll direct you to the images you want, without static allocations.
- Monitoring and accounting already supporting WNoD.

WNoD Deployment

WNoD are being **deployed in production** at the INFN Tier-1:

- More than 400k test jobs run. Currently ca. 130 VMs.
- Regular SAM tests running successfully on virtual WN, on SL5.3.
- ETICS running on WNoD only, on custom (i.e. ETICS-specific) SL5.3 image.
- ATLAS and CDF transparently running on both traditional and virtual WN, on SLC4.4.
- CMS are testing SL5.3 on WNoD right now. ATLAS, Alice and LHCb will start tests of SL5.3 shortly.
- Thanks to WNoD, migration to SL5 is as flexible as it can be: tell us when you're ready, and we'll direct you to the images you want, without static allocations.
- Monitoring and accounting already supporting WNoD.

WNoD Deployment

WNoD are being **deployed in production** at the INFN Tier-1:

- More than 400k test jobs run. Currently ca. 130 VMs.
- Regular SAM tests running successfully on virtual WN, on SL5.3.
- ETICS running on WNoD only, on custom (i.e. ETICS-specific) SL5.3 image.
- ATLAS and CDF transparently running on both traditional and virtual WN, on SLC4.4.
- CMS are testing SL5.3 on WNoD right now. ATLAS, Alice and LHCb will start tests of SL5.3 shortly.
- Thanks to WNoD, migration to SL5 is as flexible as it can be: tell us when you're ready, and we'll direct you to the images you want, without static allocations.
- Monitoring and accounting already supporting WNoD.

WNoD Deployment

WNoD are being **deployed in production** at the INFN Tier-1:

- More than 400k test jobs run. Currently ca. 130 VMs.
- Regular SAM tests running successfully on virtual WN, on SL5.3.
- ETICS running on WNoD only, on custom (i.e. ETICS-specific) SL5.3 image.
- ATLAS and CDF transparently running on both traditional and virtual WN, on SLC4.4.
- CMS are testing SL5.3 on WNoD right now. ATLAS, Alice and LHCb will start tests of SL5.3 shortly.
- Thanks to WNoD, migration to SL5 is as flexible as it can be: tell us when you're ready, and we'll direct you to the images you want, without static allocations.
- Monitoring and accounting already supporting WNoD.

WNoD Deployment

WNoD are being **deployed in production** at the INFN Tier-1:

- More than 400k test jobs run. Currently ca. 130 VMs.
- Regular SAM tests running successfully on virtual WN, on SL5.3.
- ETICS running on WNoD only, on custom (i.e. ETICS-specific) SL5.3 image.
- ATLAS and CDF transparently running on both traditional and virtual WN, on SLC4.4.
- CMS are testing SL5.3 on WNoD right now. ATLAS, Alice and LHCb will start tests of SL5.3 shortly.
- Thanks to WNoD, migration to SL5 is as flexible as it can be: tell us when you're ready, and we'll direct you to the images you want, without static allocations.
- Monitoring and accounting already supporting WNoD.

WNoD Deployment

WNoD are being **deployed in production** at the INFN Tier-1:

- More than 400k test jobs run. Currently ca. 130 VMs.
- Regular SAM tests running successfully on virtual WN, on SL5.3.
- ETICS running on WNoD only, on custom (i.e. ETICS-specific) SL5.3 image.
- ATLAS and CDF transparently running on both traditional and virtual WN, on SLC4.4.
- CMS are testing SL5.3 on WNoD right now. ATLAS, Alice and LHCb will start tests of SL5.3 shortly.
- Thanks to WNoD, migration to SL5 is as flexible as it can be: tell us when you're ready, and we'll direct you to the images you want, without static allocations.
- Monitoring and accounting already supporting WNoD.

Outline

1 Motivation and Architectural Design

- Virtualization of a Farm Cluster
- Extensions of the WNoD architecture
- Architectural Design

2 Current Status and Perspectives

- Farm Virtualization
- **Distributed Submission**
- Cloud Services
- Future Work

Integration with gLite Job Submission

- Available VM images are published in the Grid Information System using the Glue attribute `SoftwareRunTimeEnvironment`.
- Selection of VM images is done by users via standard JDL statements, e.g.

```
CeRequirements = "(Member(\"vm_etics_slc44\",  
other.GlueHostSoftwareApplicationRuntimeEnvironment))"
```

- Users (or sets thereof) must be authorized by resource providers to locally use the selected VM images through proper WNoD configuration.
- This **works today** with the latest CREAM and WMS builds.

Integration with gLite Job Submission

- Available VM images are published in the Grid Information System using the Glue attribute `SoftwareRunTimeEnvironment`.
- Selection of VM images is done by users via standard JDL statements, e.g.

```
CeRequirements = "(Member(\"vm_etics_slc44\",  
  other.GlueHostSoftwareApplicationRuntimeEnvironment))"
```

- Users (or sets thereof) must be authorized by resource providers to locally use the selected VM images through proper WNoD configuration.
- This **works today** with the latest CREAM and WMS builds.

Integration with gLite Job Submission

- Available VM images are published in the Grid Information System using the Glue attribute `SoftwareRunTimeEnvironment`.
- Selection of VM images is done by users via standard JDL statements, e.g.

```
CeRequirements = "(Member(\"vm_etics_slc44\",  
other.GlueHostSoftwareApplicationRuntimeEnvironment))"
```

- Users (or sets thereof) must be authorized by resource providers to locally use the selected VM images through proper WNoD configuration.
- This **works today** with the latest CREAM and WMS builds.

Integration with gLite Job Submission

- Available VM images are published in the Grid Information System using the Glue attribute `SoftwareRunTimeEnvironment`.
- Selection of VM images is done by users via standard JDL statements, e.g.

```
CeRequirements = "(Member(\"vm_etics_slc44\",  
  other.GlueHostSoftwareApplicationRuntimeEnvironment))"
```

- Users (or sets thereof) must be authorized by resource providers to locally use the selected VM images through proper WNoD configuration.
- This **works today** with the latest CREAM and WMS builds.

Outline

1 Motivation and Architectural Design

- Virtualization of a Farm Cluster
- Extensions of the WNoD architecture
- Architectural Design

2 Current Status and Perspectives

- Farm Virtualization
- Distributed Submission
- **Cloud Services**
- Future Work

Providing `ssh` Access to Compute Nodes

- We are writing a layer to interface WNoD with cloud APIs (like OCCl, or Amazon EC2).
- You request a computing resource, get an IP address back, `ssh` to the resource, do your work, and then dispose of the resource.
- This **works today** with a very basic user interface. It dynamically instantiates a VM taken from the standard set of Tier-1 resources and allocates it to the requester through WNoD and the LRMS.

Providing `ssh` Access to Compute Nodes

- We are writing a layer to interface WNoD with cloud APIs (like OCCI, or Amazon EC2).
- You request a computing resource, get an IP address back, `ssh` to the resource, do your work, and then dispose of the resource.
- This **works today** with a very basic user interface. It dynamically instantiates a VM taken from the standard set of Tier-1 resources and allocates it to the requester through WNoD and the LRMS.

Providing `ssh` Access to Compute Nodes

- We are writing a layer to interface WNoD with cloud APIs (like OCCl, or Amazon EC2).
- You request a computing resource, get an IP address back, `ssh` to the resource, do your work, and then dispose of the resource.
- This **works today** with a very basic user interface. It dynamically instantiates a VM taken from the standard set of Tier-1 resources and allocates it to the requester through WNoD and the LRMS.

Outline

1 Motivation and Architectural Design

- Virtualization of a Farm Cluster
- Extensions of the WNoD architecture
- Architectural Design

2 Current Status and Perspectives

- Farm Virtualization
- Distributed Submission
- Cloud Services
- **Future Work**

Many Things Still to Come...

- Full deployment of WNoD at the INFN Tier-1.
- System- and user-level documentation.
- Support for MPI/multicore jobs.
- Grid/Glue schema standardization efforts and integration of the latest CREAM and WMS builds into the gLite release.
- Cloud UI improvements and compliance to OCCl (Open Cloud Computing Interface) standards.
- Security considerations wrt the cloud interface.
- Billing for cloud usage.
- (Transparent) Use of external cloud providers for flash requests.
- And others...

Many Things Still to Come...

- Full deployment of WNoD at the INFN Tier-1.
- System- and user-level documentation.
- Support for MPI/multicore jobs.
- Grid/Glue schema standardization efforts and integration of the latest CREAM and WMS builds into the gLite release.
- Cloud UI improvements and compliance to OCCl (Open Cloud Computing Interface) standards.
- Security considerations wrt the cloud interface.
- Billing for cloud usage.
- (Transparent) Use of external cloud providers for flash requests.
- And others...

Many Things Still to Come...

- Full deployment of WNoD at the INFN Tier-1.
- System- and user-level documentation.
- Support for MPI/multicore jobs.
- Grid/Glue schema standardization efforts and integration of the latest CREAM and WMS builds into the gLite release.
- Cloud UI improvements and compliance to OCCl (Open Cloud Computing Interface) standards.
- Security considerations wrt the cloud interface.
- Billing for cloud usage.
- (Transparent) Use of external cloud providers for flash requests.
- And others...

Many Things Still to Come...

- Full deployment of WNoD at the INFN Tier-1.
- System- and user-level documentation.
- Support for MPI/multicore jobs.
- Grid/Glue schema standardization efforts and integration of the latest CREAM and WMS builds into the gLite release.
- Cloud UI improvements and compliance to OCCl (Open Cloud Computing Interface) standards.
- Security considerations wrt the cloud interface.
- Billing for cloud usage.
- (Transparent) Use of external cloud providers for flash requests.
- And others...

Many Things Still to Come...

- Full deployment of WNoD at the INFN Tier-1.
- System- and user-level documentation.
- Support for MPI/multicore jobs.
- Grid/Glue schema standardization efforts and integration of the latest CREAM and WMS builds into the gLite release.
- Cloud UI improvements and compliance to OCCl (Open Cloud Computing Interface) standards.
- Security considerations wrt the cloud interface.
- Billing for cloud usage.
- (Transparent) Use of external cloud providers for flash requests.
- And others...

Many Things Still to Come...

- Full deployment of WNoD at the INFN Tier-1.
- System- and user-level documentation.
- Support for MPI/multicore jobs.
- Grid/Glue schema standardization efforts and integration of the latest CREAM and WMS builds into the gLite release.
- Cloud UI improvements and compliance to OCCl (Open Cloud Computing Interface) standards.
- Security considerations wrt the cloud interface.
- Billing for cloud usage.
- (Transparent) Use of external cloud providers for flash requests.
- And others...

Many Things Still to Come...

- Full deployment of WNoD at the INFN Tier-1.
- System- and user-level documentation.
- Support for MPI/multicore jobs.
- Grid/Glue schema standardization efforts and integration of the latest CREAM and WMS builds into the gLite release.
- Cloud UI improvements and compliance to OCCl (Open Cloud Computing Interface) standards.
- Security considerations wrt the cloud interface.
- Billing for cloud usage.
- (Transparent) Use of external cloud providers for flash requests.
- And others...

Many Things Still to Come...

- Full deployment of WNoD at the INFN Tier-1.
- System- and user-level documentation.
- Support for MPI/multicore jobs.
- Grid/Glue schema standardization efforts and integration of the latest CREAM and WMS builds into the gLite release.
- Cloud UI improvements and compliance to OCCl (Open Cloud Computing Interface) standards.
- Security considerations wrt the cloud interface.
- Billing for cloud usage.
- (Transparent) Use of external cloud providers for flash requests.
- And others...

Many Things Still to Come...

- Full deployment of WNoD at the INFN Tier-1.
- System- and user-level documentation.
- Support for MPI/multicore jobs.
- Grid/Glue schema standardization efforts and integration of the latest CREAM and WMS builds into the gLite release.
- Cloud UI improvements and compliance to OCCl (Open Cloud Computing Interface) standards.
- Security considerations wrt the cloud interface.
- Billing for cloud usage.
- (Transparent) Use of external cloud providers for flash requests.
- And others...

Summary

- **WNoD works today** at the INFN Tier-1 and is being used to solve real use cases, like migration to a new O/S (SL5.3) or provisioning of custom environments.
- **Distributed selection of custom execution environments** in EGEE has been demonstrated.
- Dynamic integration of **cloud and grid services using the same set of resources works today** and can be exploited to provide new services.




Summary

- **WNoD works today** at the INFN Tier-1 and is being used to solve real use cases, like migration to a new O/S (SL5.3) or provisioning of custom environments.
- **Distributed selection of custom execution environments** in EGEE has been demonstrated.
- Dynamic integration of **cloud and grid services using the same set of resources works today** and can be exploited to provide new services.

Summary

- **WNoD works today** at the INFN Tier-1 and is being used to solve real use cases, like migration to a new O/S (SL5.3) or provisioning of custom environments.
- **Distributed selection of custom execution environments** in EGEE has been demonstrated.
- Dynamic integration of **cloud and grid services using the same set of resources works today** and can be exploited to provide new services.

References I

-  M. Cecchi, A. Ghiselli, A. Italiano, M. Orrù, D. Rebatto, D. Salomoni, V. Venturi, L. Zangrando
The INFN approach to Grid and Cloud Integration
Presentation to the CCR and INFN-Grid Workshop, Palau 2009.
-  A. Chierici, D. Salomoni, R. Veraldi
Measuring Performance of Linux Hypervisors
Il Nuovo Cimento, accepted, C033, 2009.
-  M. Cecchi, A. Ghiselli, A. Italiano, M. Orrù, D. Rebatto, D. Salomoni, V. Venturi, L. Zangrando
Distributed Job Submission in Dynamic Virtual Execution Environments
Journal of Grid Computing, submitted, 2009.