Connecting Multiple Clouds and Mixing Real and Virtual Resources via the Open Source WNoDeS Framework

Davide Salomoni, INFN

on behalf of the WNoDeS team

CHEP - NYC, May 22, 2012

Agenda

- The Open Source WNoDeS stack
- Connecting cloud services
  - Interconnection of Grids and Clouds, federation of Cloud providers
  - Proxy services for resources (in particular, jobs)
- Exploiting virtualization without disrupting computing centers: WNoDeS Mixed Mode
- Concluding remarks

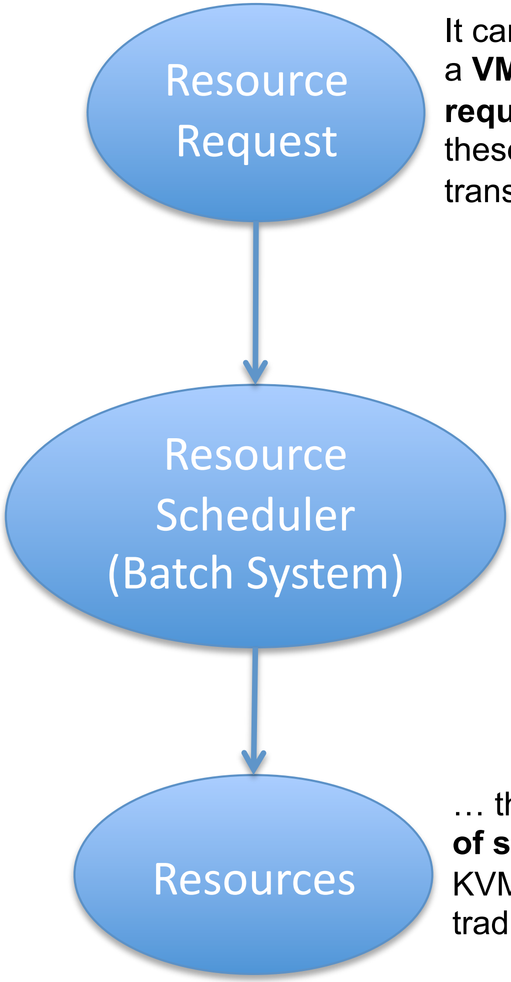
Agenda

- The Open Source WNoDeS stack
- Connecting cloud services
  - Interconnection of Grids and Clouds, federation of Cloud providers
  - Proxy services for resources (in particular, jobs)
- Exploiting virtualization without disrupting computing centers: WNoDeS Mixed Mode
- Concluding remarks

WNoDeS

- WNoDeS → **W**orker **N**odes **o**n **De**mand **S**ervice
  - D. Salomoni et al., 2011 J. Phys.: Conf. Ser. 331 052017 (CHEP 2010)
- A software framework created in 2008 by INFN to integrate Grid and Cloud provisioning through virtualization
  - Key feature: all resources (presented via Grid, Cloud, or else) are taken from a **common pool** to avoid static partitioning.
  - Key feature: resource matchmaking policies are handled by an LRMS.
- Scalable and reliable – it is in production at several Italian centers, including the INFN Tier-1 since November 2009
  - Dynamically managed up to 2,000 on-demand Virtual Machines (VMs) there.
- Transparent for local and Grid Computing Center services
  - Batch jobs can be directed to run on VMs based on LRMS policies.
- Leveraging proven open source software technologies like Linux KVM, Torque/Maui (Platform LSF also supported), EMI gLite middleware
- WNoDeS version 2 is part of EMI-2, released on May 21, 2012

WNoDeS, Architectural Overview

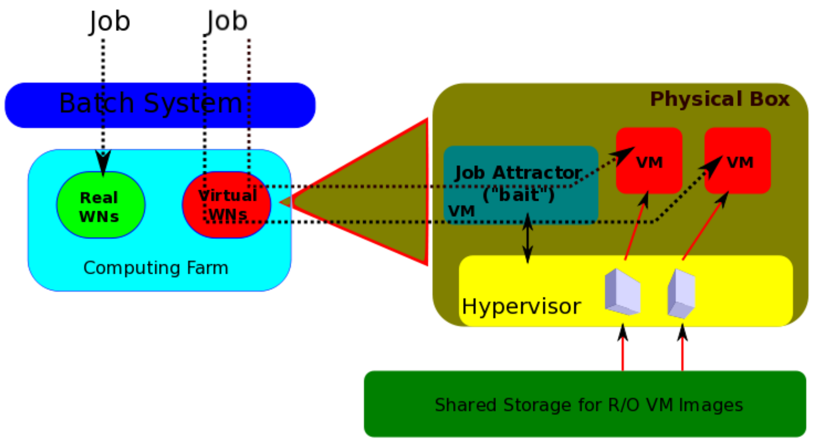


It can be a **local job**, a **Grid job**, a **VM instantiation request**, a **request for a Cloud service**. All these requests get transparently translated into "jobs"…

… that are handled **by an LRMS** (a batch system). LRMS-based policies allow flexible scheduling and scalable access to **resources**…

… that can be based on a **mix of systems**: some capable of KVM-based virtualization, some traditional, non-virtual resources.

General schema to handle a VM/ service/job instantiation request



Every physical system runs a special process called "bait". Its purpose is to manage local resources and arbitrate access to dynamically created local VMs.

Agenda

- The Open Source WNoDeS stack
- Connecting cloud services
  - Interconnection of Grids and Clouds, federation of Cloud providers
  - Proxy services for resources (in particular, jobs)
- Exploiting virtualization without disrupting computing centers: WNoDeS Mixed Mode
- Concluding remarks

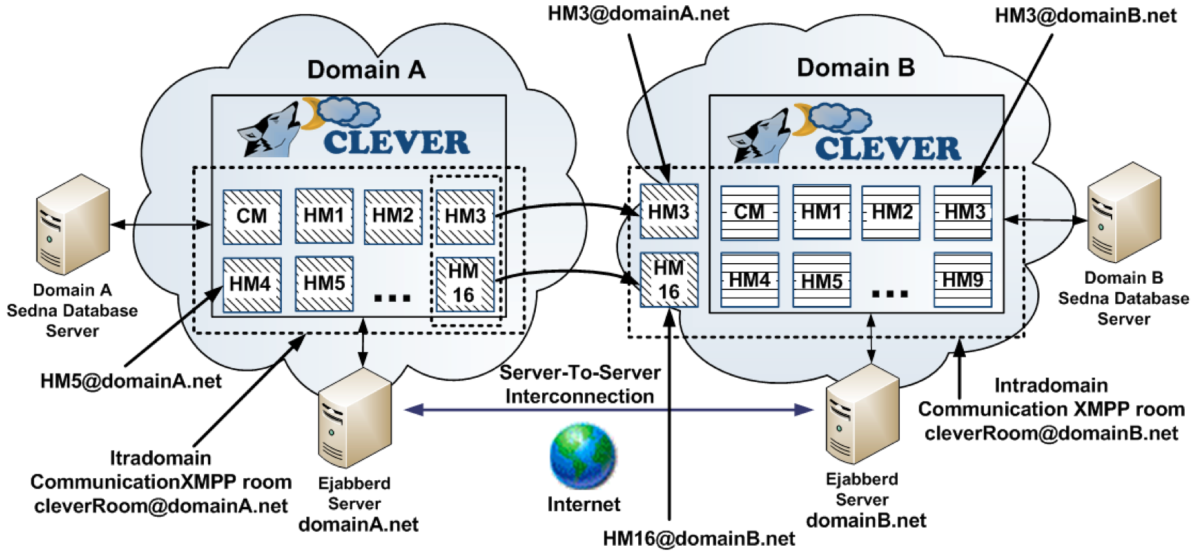
Integration of Clouds and Grids

- A general goal is for us to offer an **integrated, federated computing and storage infrastructure** providing:
  - Grid Computing services
    - Accessible also to users who are not part of a VO, and who do not use X.509 digital certificates.
  - Cloud Computing services, targeted to both "traditional" Grid users and to other, "cloud native" users
    - Using existing credentials: X.509 certificates (+ VOMS), federated identities (Shibboleth), or other common authentication methods (e.g. OpenID or OAuth)
    - For implementation examples, see e.g. "A General Purpose Grid and Cloud Portal to Simplify Scientific Communities Integration into Distributed Computing Infrastructures", EGI CF, March 2012, http://goo.gl/w2tvt
  - Interconnection of cooperating Cloud Computing resource centers
    - Integrating with Virtual Infrastructure Management services and with Cloud Manager services provided by the **CLEVER** research project.
    - See also the EGI Federated Clouds Task Force, http://goo.gl/7o2Eo, for discussions about e.g. cloud brokering options
- Another important goal is to offer a **bursting service**, acting as transparent proxy for jobs or in general service requests
  - Mediating access to other resources on behalf of the user

CLEVER

- CLEVER: CLoud-Enabled Virtual EnviRonment
  - A project developed by University of Messina, cf. 2010 IEEE Symposium on Computers and Communications (ISCC), pp. 477-482, DOI 10.1109/ISCC.2010.5546555
  - Simplification of access management of private/hybrid clouds
  - Provisioning of simple interfaces to interact with different "interconnected" clouds
    - With VM deployment and load balancing through migration

Inter-Domain Communication



HM3@domainA.net
HM3@domainB.net
Domain A
Domain B
CLEVER
CLEVER
CM
HM1
HM2
HM3
HM4
HM5
...
HM 16
HM3
HM 16
CM
HM1
HM2
HM3
HM4
HM5
...
HM9
Domain A Sedna Database Server
Domain B Sedna Database Server
HM5@domainA.net
Server-To-Server Interconnection
Intradomain Communication XMPP room cleverRoom@domainB.net
Itradomain CommunicationXMPP room cleverRoom@domainA.net
Ejabberd Server domainA.net
Internet
Ejabberd Server domainB.net
HM16@domainB.net

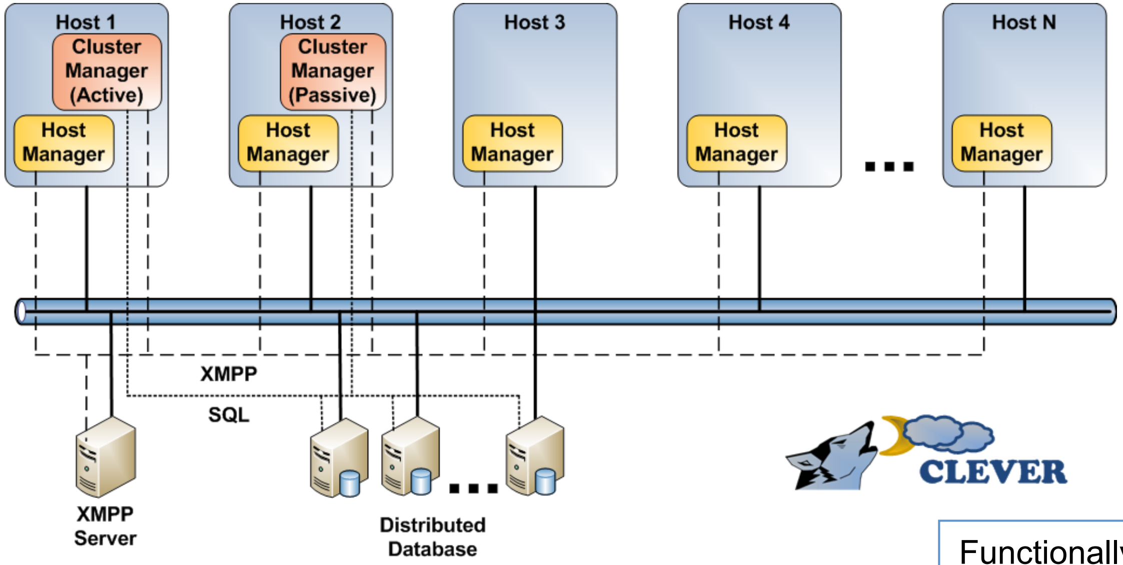
- Communication protocol over XMPP (IETF's Extensible Messaging and Presence Protocol)
  - Building on the expertise gained with past projects such as the RESERVOIR framework

- Main characteristics:
  - Open source protocol
  - Peer-to-peer approach
  - Auto-setup through DHCP discovery
  - In-band registration
  - Firewall pass-through

Communication Architecture

✓We are including parts of CLEVER (host and cluster manager) in WNoDeS to support XMPP registration and communications

- This gives us a foundation for scalable, resilient, global provisioning and monitoring system for Grid and Cloud integration for WNoDeS sites



Functionally overlapping with the WNoDeS HV and bait services

Functionally overlapping with the WNoDeS Name Server

- Host Management Layer: Host Manager (HM)
  - Performs physical resources monitoring and VM allocation
- Cluster Management Layer: Cluster Manager (CM)
  - Monitors the overall state of the cluster, coordinates HMs
- External components: XMPP Server and Distributed Database
- XMPP advantages: host presence, open standard
  - No central failure point: fault tolerance mechanism with multiple CM instances

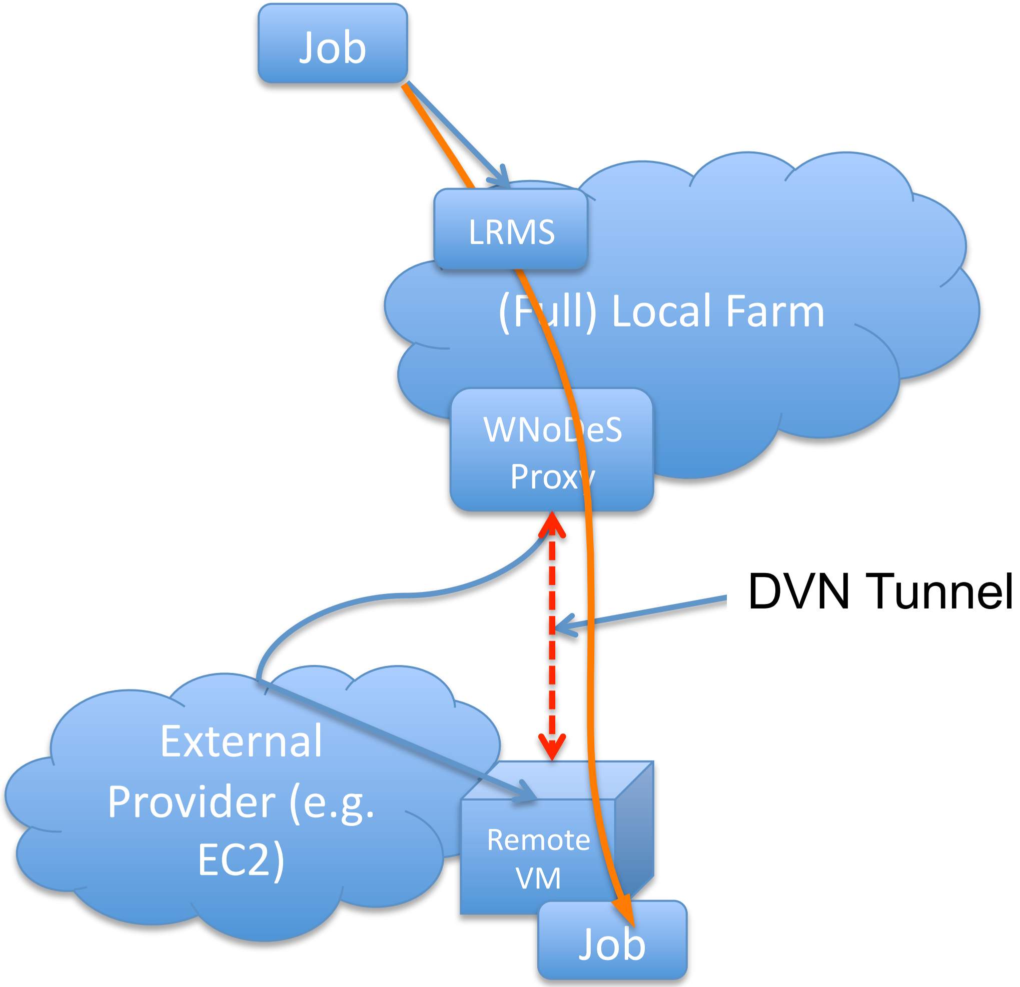
Agenda

- The Open Source WNoDeS stack
- Connecting cloud services
  - Interconnection of Grids and Clouds, federation of Cloud providers
  - Proxy services for resources (in particular, jobs)
- Exploiting virtualization without disrupting computing centers: WNoDeS Mixed Mode
- Concluding remarks

Transparent Job Proxy

- Use case: temporary job bursting (e.g. flash requests) and transparent acquisition of resources from an external provider
  - Not a federated infrastructure: a resource provider simply acts as transparent proxy to offer additional (computing) resources to customers willing to pay for this service
- Uses the WNoDeS Dynamic Virtual Networks (DVN) module to create a NAT'd, transparent GRE tunnel to an external resource – see poster [508] at this conference
  - The job is then moved to the remote resource using the established WNoDeS architecture and executed as if the remote resource was local.

Proxy Architecture



- Remarks:
  - Best suitable for CPU-intensive workloads. However, access to storage, software areas, etc. is transparent (bandwidth/latency aside…)
  - Access through Amazon EC2 is only an example. Possibly appealing, esp. for a funding agency, is also access to a national, cooperating Tier-x with unused capacity.

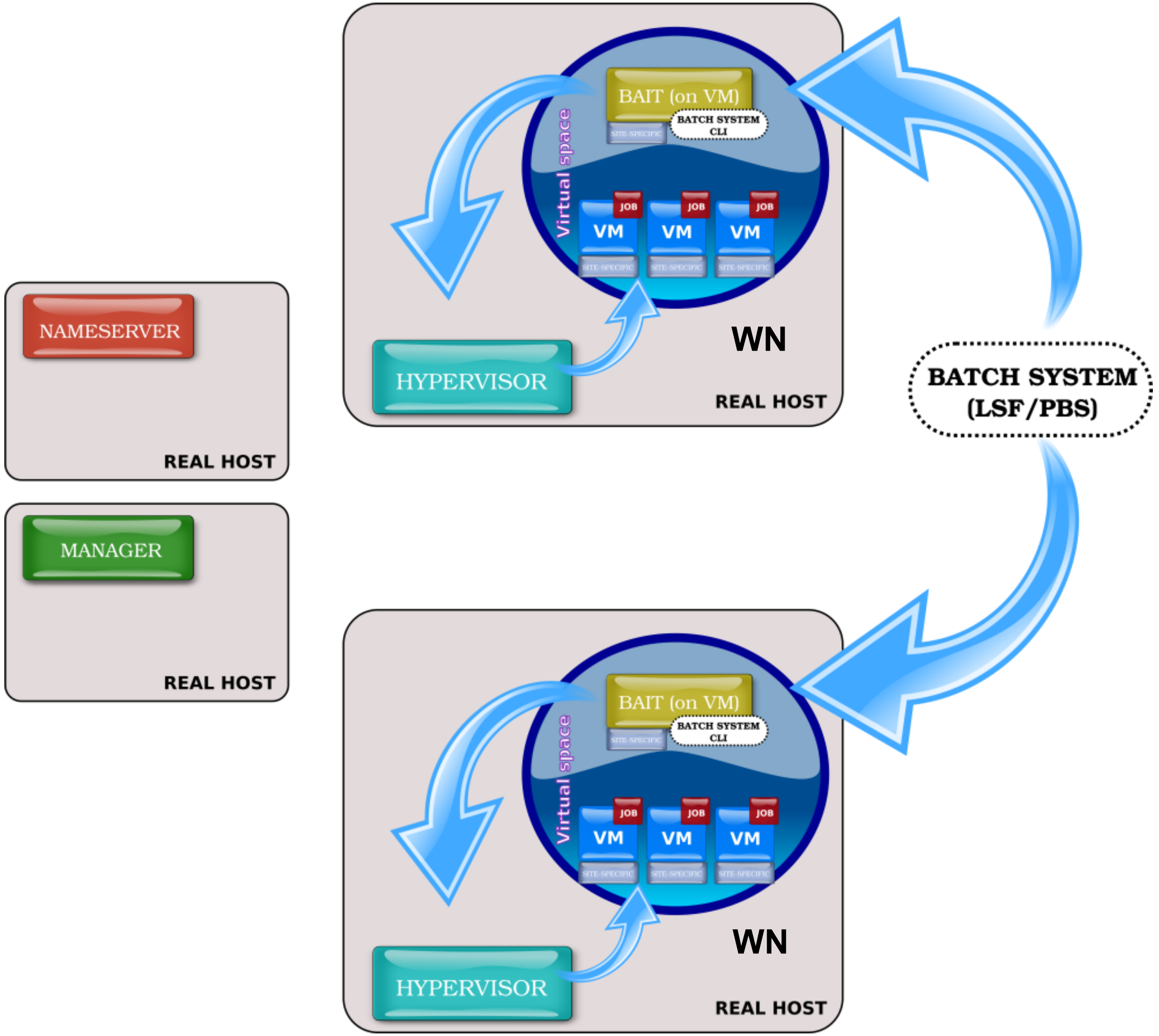
Agenda

- The Open Source WNoDeS stack
- Connecting cloud services
  - Interconnection of Grids and Clouds, federation of Cloud providers
  - Proxy services for resources (in particular, jobs)
- Exploiting virtualization without disrupting computing centers: WNoDeS Mixed Mode
- Concluding remarks

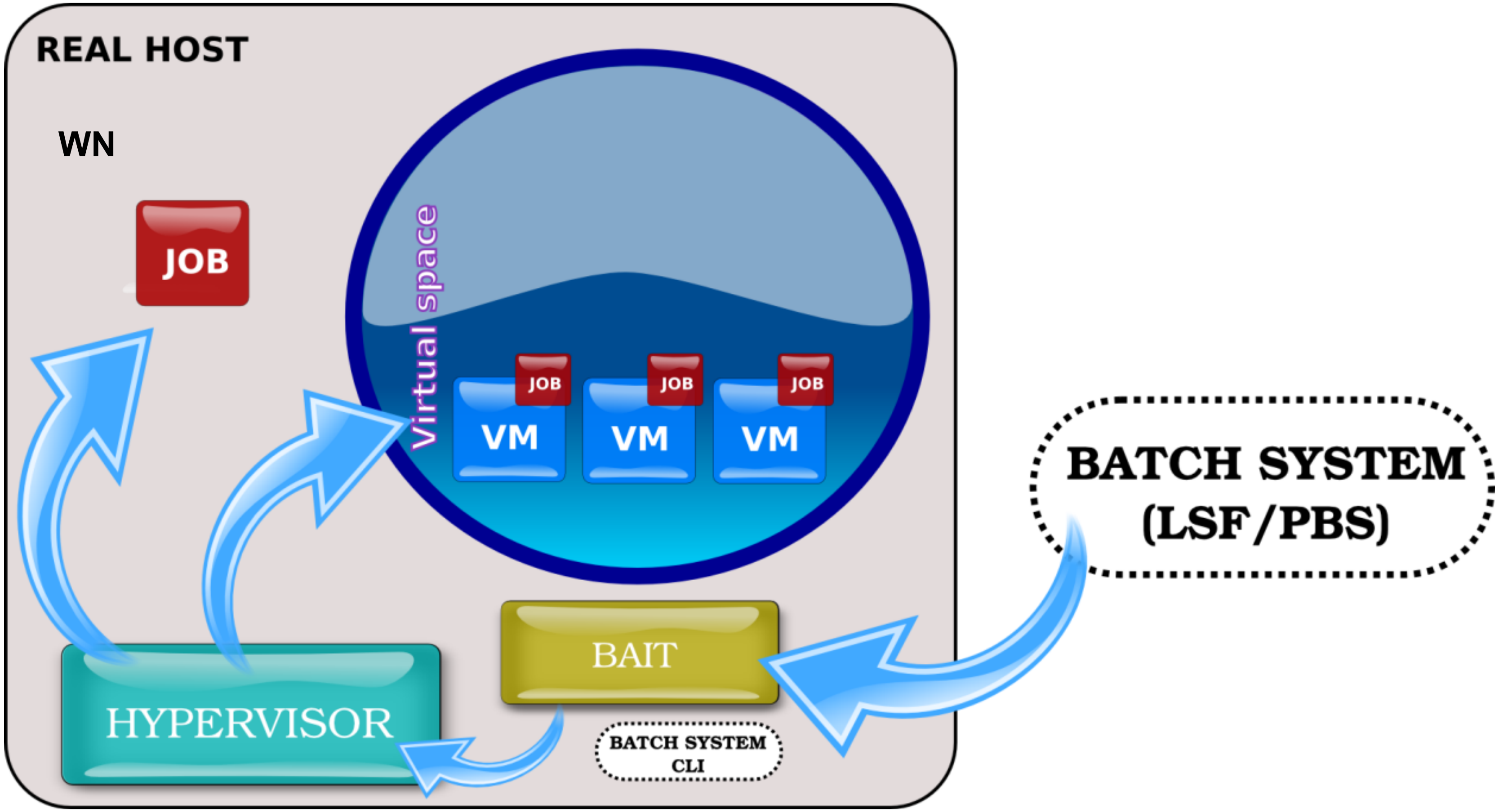
WNoDeS Mixed Mode

- **What**
  - A WNoDeS configuration option allowing the use of physical resources as both traditional batch nodes **and** as hypervisors for the instantiation of virtual machines – on the same hardware, at the same time.
  - VMs can be used to also run batch jobs, or to provide cloud services.
- **Why**
  - Some tasks are not suitable to being executed on virtual nodes – for example, jobs requiring GPGPU resources, or jobs with high I/O requirements → run them on physical nodes.
  - On the same physical nodes, one can also offer virtualized services for those users requiring them → no need to set aside nodes for virtualized services.
- **Where**
  - Mixed mode is included in the WNoDeS version released with EMI-2 and can be administratively turned on or off.

WNoDeS, Mixed Mode **off**



WNoDeS, Mixed Mode **on**



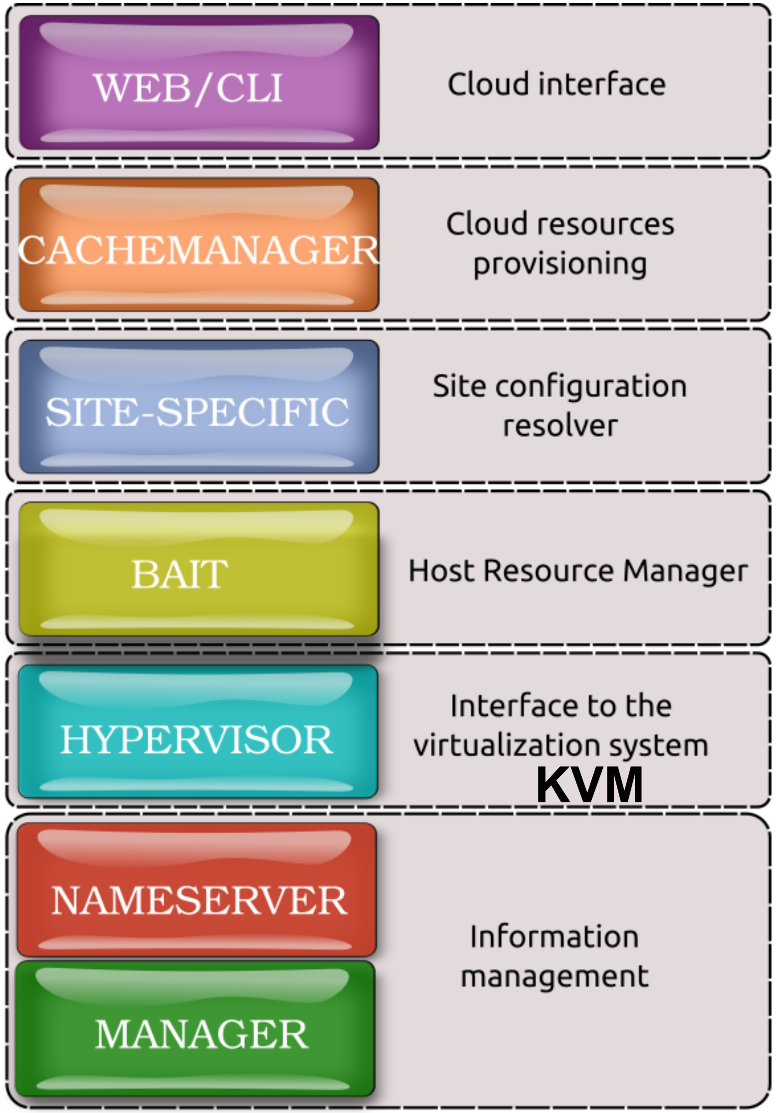
Mixed Mode Pros and Cons

- **Pros**
  - Progressively install WNoDeS in a farm without first having to decide which nodes will support virtualization and which not.
  - Add support e.g. for Cloud computing, interactive usage on custom VMs etc. in a traditional farm.
  - Direct jobs to VMs or to real hardware using LRMS policies and a simple pre-exec/prologue script (a template is supplied with the WNoDeS distribution). One can differentiate real vs. virtual requests/jobs e.g. based on queues, users, requirements, Grid VOs, etc.
- **Cons**
  - In a purely virtual farm set up, physical systems are only used as hypervisors, so they can be put e.g. in private address space. With mixed mode, they can also be used (like in a traditional farm) to run jobs and may need public access.
  - With mixed mode, a physical system is part of the LRMS cluster and may use up LRMS licenses proportionally to the number of its cores. If the same physical system is then also used to create VMs that become part of the LRMS cluster (e.g. to run batch or grid jobs), these VMs will also use up LRMS licenses and the total number of LRMS licenses used by a physical system may be O(2*cores). This can be a problem with some sites using commercial LRMS.

Mixed Mode Performance

- Mixed mode is supported on systems running either SL5.x or SL6.x
  - SL6.x introduces `cgroups`, allowing one to assign fine-grained priorities to processes.
    - Jobs on the HV are processes, and KVM VM's are also processes. It is therefore possible to alter the relative priorities of real vs. virtual jobs.
    - However, WLCG jobs are still not supported on SL6.x, so for a WLCG-type mixed mode scenario the system (hypervisor) should be running SL5.x.
- Performance measurements were done in order to evaluate the impact of having jobs on the hypervisor and VM's also running jobs at the same time, on the same system.
  - Baseline is the time it takes to execute a job (either real or on a VM) without any interference.
  - Then we added a variety of other jobs, either real or on a VM, doing either I/O or CPU-intensive tasks.
  - Results show that in general resources are shared fairly between jobs running on the hypervisor and jobs running on VMs. (no cgroups)

WNoDeS, Status

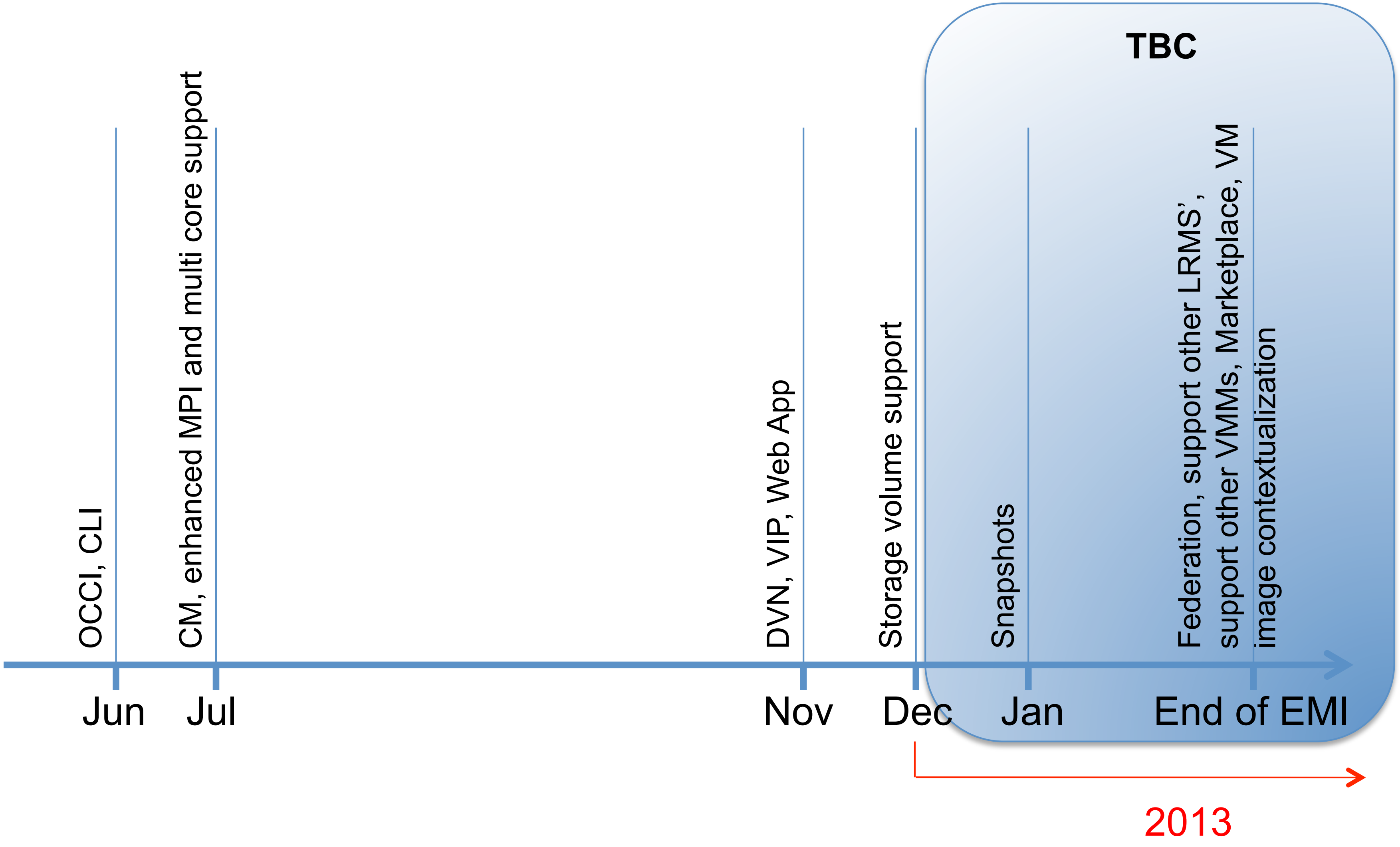


EMI updates
(e.g. DVN, federation)

EMI-2 – now
(includes mixed mode)

EMI INFSO-RI-261611

Current WNoDeS Timeline



Concluding Remarks

- For what regards interconnection of clouds, rather than trying to solve a general problem, we are focusing on connecting peer-to-peer cooperating IaaS providers with simple brokering services.
- Job proxying is an option for the use case of flash requests. It can be applied to public cloud services or to resource centers with which there is a mutual agreement.
- WNoDeS Mixed Mode lets a resource center to progressively introduce virtualized services without disrupting existing set-ups and maximizing resource utilization.
- WNoDeS itself is now released as part of the EMI stack and currently supports Platform LSF and Torque/Maui as LRMS.

Thanks

- Related contributions at this conference:
  - A. Chierici et al., "Increasing performance in KVM virtualization within a Tier-1 environment" – poster [326]
  - D. Andreotti et al., "The WNoDeS Cache Manager, an efficient method to self-allocate virtual resources" – poster [500]
  - M. Caberletti et al., "Creating Dynamic Virtual Networks for network isolation to support Cloud computing and virtualization in large computing centers" – poster [508]
- Acknowledgments:
  - we gratefully acknowledge support for this work by the INFN Tier-1, the CLEVER team and the Italian Grid Infrastructure (IGI).
- More info:
  - http://web.infn.it/wnodes
  - wnodes-support@lists.infn.it