WNoDeS
Grid / Cloud Integration Framework

INFN
Istituto Nazionale di Fisica Nucleare

WNoDeS, a flexible and scalable Grid / Cloud virtualization system

Davide Salomoni (INFN-CNAF)
for the WNoDeS Project
http://web.infn.it/wnodes

2nd ASPERA Workshop
Barcelona, 30-31 May 2011

Background to this work

- INFN is the Italian National Institute for Nuclear Physics. It is engaged in many international physics experiments *and* in developing, delivering and supporting storage and computing services for them.
- I am the manager of computing services at the INFN National Computing Center (CNAF), located in Bologna, Italy.
- CNAF actually hosts about 8,500 computing cores (10,500 by September 2011), 9 PB of disk space and 10 PB of tape space.
  - Each day, about 40,000 jobs get executed at CNAF.
  - It supports 20 international scientific experiments.
    - About 14% of the CNAF computing resources are pledged for astro-particle physics experiments (e.g., AMS2, Argo, Auger, Fermi/Glast, Magic, Pamela, Virgo).
  - It is the Italian Tier-1 for CERN-based LHC experiments and Tier-0/1 for several others.

The **W**orker **N**odes **o**n **De**mand **S**ervice, a.k.a. **WNoDeS**, in one slide

- A software framework created by INFN to integrate Grid and Cloud provisioning through virtualization
  - All resources (Grid, Cloud, or else) are taken from a common pool
- Scalable and reliable – it is in production at several Italian centers, including the INFN Tier-1 (CNAF, Bologna) since November 2009
  - Currently managing about 2000 on-demand Virtual Machines (VMs) there
- Totally transparent for Grid users and for users of traditional Computing Centers batch systems
- Supports a native Cloud interface
  - OCCI (Open Cloud Computing Interface) compliant
  - A Cloud Web portal
- Integrates authentication, policy and accounting
- It is not required to convert an *entire* farm to WNoDeS: it can coexist and share resources with a traditional compute cluster not using virtual machines

WNoDeS use case #1: Local Users

- **Jobs** submitted by users of a traditional computing center
  - Users log on to a front-end system of a resource center and submit jobs to a computer farm
  - WNoDeS can associate jobs belonging to a user or a set of users to Virtual Machines specifically created for them
  - This is completely transparent for users
    - No need to change anything from the users' point of view

WNoDeS use case #2: Grid Users

- **Jobs** submitted by users of a Grid-based distributed infrastructure. Two possibilities here:
  - All jobs beloning to certain Virtual Organizations (VOs) can be directed to pre-packaged VMs. This is completely transparent for users.
  - Grid users can specify which VM they want their jobs to run on
    - Using standard EMI (European Middleware Initiative) gLite job management tools.
- CNAF is part of the WLCG infrastructure, granting access to distributed computing and storage
  - Access to distributed compute nodes via the EMI CREAM Compute Elements.
  - For distributed access to storage, see the talk *GEMSS, the Grid Enabled Mass Storage System* by L. dell'Agnello.

WNoDeS use case #3: Cloud Computing

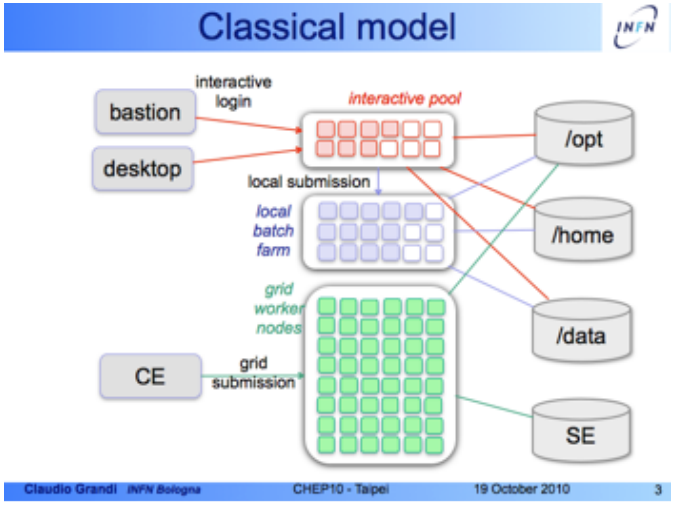
- **Self-allocation of compute resources** (Cloud Computing). This can happen via:
  - A standard API called the Open Cloud Computing Interface (OCCI), developed by the Open Grid Forum (OGF)
    - Rarely employed directly by users.
  - A Cloud Web portal
    - A user-friendly way to self-allocate resources;
    - Which will be integrated into a generic resource allocation and management portal, usable for both Cloud and Grid computing.

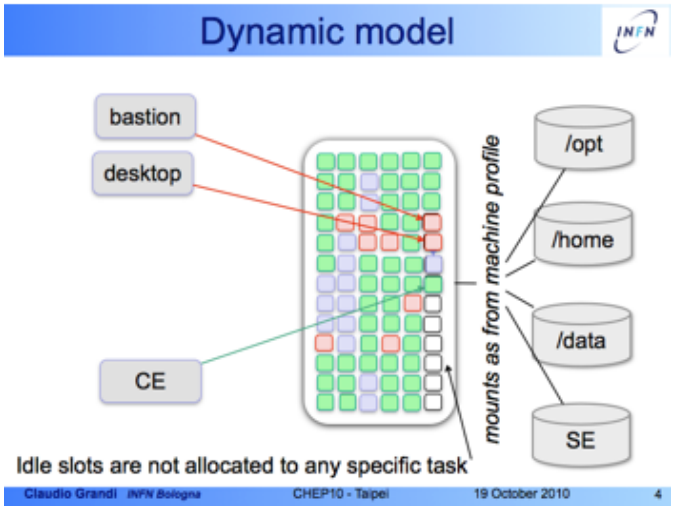
WNoDeS Cloud access

WNoDeS use case #4:
Virtual Interactive Pools (VIP)

- **Self-allocation of systems** by users of a traditional computing center
  - Systems are provisioned so that users can log on to them with their local account (no root access).
  - Users may specify characteristics such as VM image, number of CPUs, amout of RAM, local filesystems to be mounted.
  - These systems can be employed by users for instance to create pools of machines for interactive analysis or to instantiate ad-hoc services.

This is a kind of cloud computing applied to a traditional computing center designed to efficiently offer new services, without incurring the overhead to dedicate resources for this purpose.





WNoDeS Key Advantages

- Use of a **common pool of resources**
  - There is no need to dedicate resources to "user interfaces", "Grid computing", "Cloud computing", "local users" – all types of resources are taken from a common pool, resulting in overall better utilization of resources.
- **Integrated support of old and new use cases** (local access, Grid computing, Cloud computing)
- **Re-use of ten years of worldwide development, expertise and resources** brought about by **Grid Computing**
  - Applied to the key areas of Authentication, Authorization, Accounting, Information Systems, Brokering.
    - This will make it possible for example to **inter-connect Clouds** without starting from scratch in these areas again.
- **Flexibility and scalability**
  - At the core of WNoDeS there is a standard batch system used for resource provisioning and policing – a mature, stable piece of software found in any sizeable resource center. No need to rewrite this key and complex part.
- Possibility to access **resources also from external providers**
  - Acting as a transparent broker for users – for example, to serve extra load-requests.

WNoDeS status

- WNoDeS 1 in production at the INFN Tier-1 and at other INFN sites in Italy
  - Focusing on providing virtualization services for local and Grid access.
- WNoDeS 2 planned for release in September 2011
  - Introducing (among other things) the Cloud portal and the VIP interface and support for multiple batch systems.
- The WNoDeS development program focuses in particular on:
  - Dynamic network virtualization;
  - VM image abstraction;
  - Efficient access to large storage systems;
  - Inter-cloud connectivity.

A WNoDeS adoption case: Auger

- **Auger**: a 3000 m^2 cosmic ray observatory located in Argentina, studying ultra-high energy cosmic rays.
- Need read-only access to a mysql-based condition database to perform detector simulation from hundreds of compute nodes concurrently. Two solutions using WNoDeS were tested for this:
  1. A VM, dedicated to Auger and including mysqld and the condition DB, was created. Each compute node (each VM) is completely independent. This works well, but requires rebuilding the VM image every time the condition DB changes.
  2. A VM, dedicated to Auger and including mysqld but accessing the condition DB over a networked file system (at CNAF, this is GPFS), was created. All compute nodes see the same mysql tables. When the condition DB needs to be updated, the tables stored on GPFS are changed. This works well, but puts some strain on GPFS metadata; the storage subsystem then needs to be adequately architectured. This is our current solution.
- In both cases, WNoDeS provided a flexibility to Auger that would have not been possible with a classical set-up.

Auger/WNoDeS statistics

- Auger uses both real and virtual compute nodes (only the latter are managed by WNoDeS), and both are part of the same cluster. Auger may also submit both Grid and local jobs.
  - Naturally, only WNoDeS-managed nodes may run simulations involving the Auger condition DB since this requires a special environment (achieved through a VM).
  - CNAF is the only Auger site where offline reconstruction with full condition DB is performed via Grid.
- The use of other WNoDeS characteristics is being planned:
  - A second virtual machine image, without the DB; the proper VM image should be requested depending on the job type.
  - A VIP-based dynamically created machine (or machines) to run or test build and installations processes.
- Since February 2011, Auger has run about 81,300 jobs at INFN CNAF, for a total of about 325,000 CPU hours and 101,000 HEP-SPEC06. This is an average of ≈ 880 HS06 / day
  - The pledged resources for Auger at CNAF amount to 800 HS06.
  - In May alone, Auger has executed more than 46,000 jobs and used an average of more than 1,400 HS06 per day.

Future improvements to access the Auger Condition DB

- To further improve scalability of access to the current mysql-based condition DB, we are testing two alternatives:
  - Use a FroNTier-based type of access
    - A web service providing HTTP access to a central database service
    - Extensively used by CMS and ATLAS already
  - Distribute the condition DB on the WNoDeS hypervisors, letting VMs access the DB on their own hypervisor
    - This removes the need to regenerate the VMs every time the condition DB changes
- If the use of mysql were not a constrain, we could also consider serving files stored on a CernVM-FS
  - A network file system based on HTTP, already deployed at INFN CNAF and used by LHC experiments to access their software area

Thanks

For further information and questions:

Davide Salomoni, INFN-CNAF

Davide.Salomoni@cnaf.infn.it

The WNoDeS web site: http://web.infn.it/wnodes

Acknoledgements for this talk:

K.Calabrese, A.Italiano, G.Zizzi (INFN CNAF),
G.Cataldi, D.Martello (INFN Lecce)