int.eu.grid

Interactive European Grid

An study and implementation of virtual servers for a parallel grid

Carlos Fernández Sánchez, Lino García Tarrés, César Veiga García, Javier López Cacheiro, Isabel Campos, Sven Stork, Marcin Plociennik.

CESGA

IberGrid, Santiago de Compostela, 15 May 2007

Outline

What is Virtualization?

Using Xen Virtualization to Implement a Grid Infrastructure

Full Virtualization & Paravirtualization

Hardware & Software configuration

CESGA Int.eu.grig site

Supporting Parallel Applications in the Grid

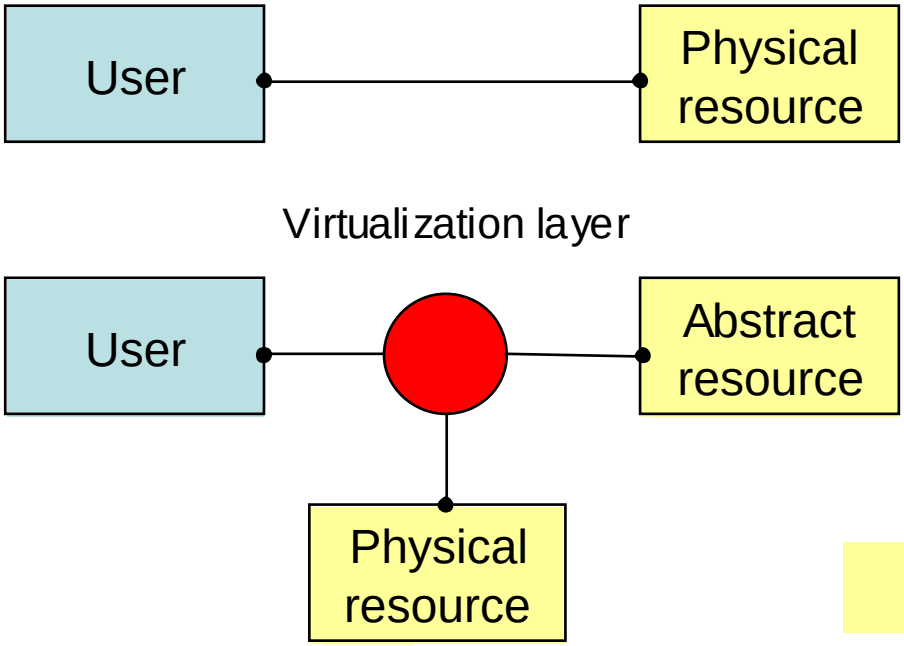
Benchmarking the virtual grid

Discussion of the results obtained

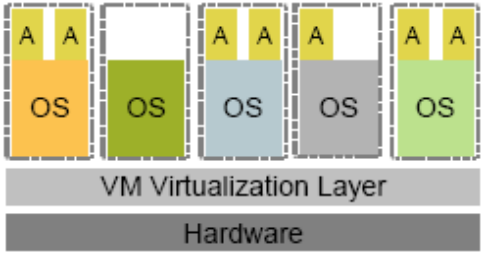
Conclusions

What is Virtualization?

Process of replacing a direct interface linking a resource (often hardware) and its user with an indirect, **software-mediated connection**.



Using Xen Virtualization to Implement a grid Infrastructure

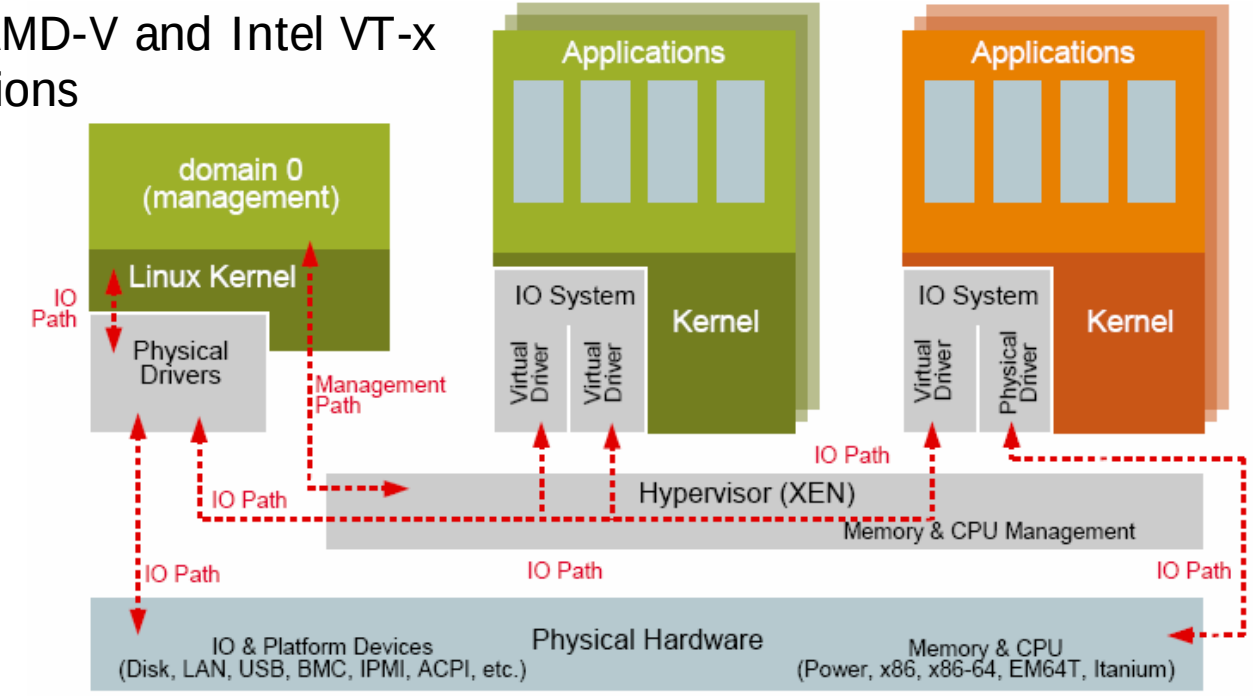


Virtualization advantages in grid environment:

- **Provides** the use of **new hardware** architectures not directly supported by gLite midleware.
- **Hardware shared** between different grid infrastructures.
- **Fault tolerant** and checkpoint & restart.
- Server provisioning **on demand**, e.g. deploy working nodes as needed.
- Better **isolation**. An application can only hang a “virtual server” that could be easily restarted.

Full Virtualization & Paravirtualization

- Full virtualization (hardware)
  - No modifications in guest O.S.
  - Complete OS isolation
  - Performance penalty
  - Uses AMD-V and Intel VT-x extensions
- Para Virtualization (user mode Linux)
  - Modifies the guest OS kernel.
  - Insecurities in O.S. cached data
  - Near-native performance



Hardware & Software Configuration

SOFTWARE:

- CESGA has selected to use **full virtualization** to take advantage of Intel Virtual technology (VT).
- Host O.S.: Linux **OpenSUSE 10.2 Xen 3.0.3-0** release.
- Virtual Machine: one virtual CPU and 512 MB RAM.

  Scientific Linux.

HARDWARE:

- System: Dell Power Edge 1955.
- CPU: two VT-enabled **quad-core** 1.6 GHz Intel **Xeon** 5310 CPUs (**total eight cores**) and 4 GB of RAM installed.
- IO: dualport **1Gbps Ethernet** adapter and one 73GB SAS disk drive.

CESGA Int.eu.grid site

Cesga Int.eu.grid virtual site

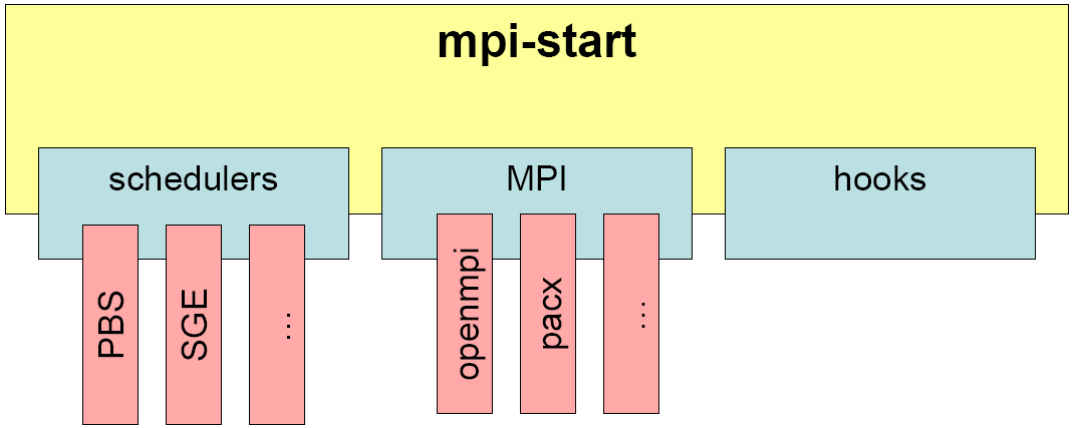
Computing element	mon	Storage element	User interface		
Rgma server	Monitoring Dev.	Monitoring Prod.	wn001	wn002	
wn003	...			wn007	
wn008	...			wn012	Accounting Portal
8 cores paravirtual wn013. SMP 					

8 cores Blade. Xen Domain 0

1 Core. Xen Full Virtual machine

Supporting Parallel Applications in the Grid

- **Remove** all the “Message Passing Interface” (**MPI**) **hard coded** implementation.
- Use a generic interface called *mpi-start:*
  - mpi-start supports **different schedulers and different MPI implementations**.
  - Supports **simple file distribution** by using scripts to be inserted in the job definition language.
  - **Hides** from the user the **particularities of the site**.



Benchmarking the virtual grid

- **Linpack**: numerical linear algebra. measures the time to solve a dense n by n systems of linear equations by Gaussian elimination with partial pivoting.

  Results: 3% less performance. **Acceptable in our grid**.

- **Bonnie++**: tests hard drive and file system performance.

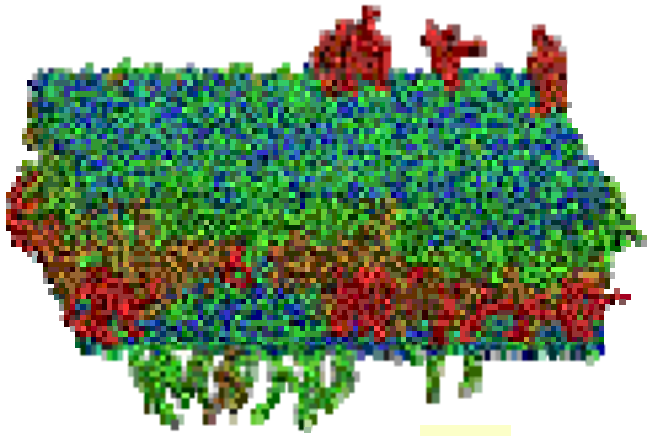
  Results: performance 5x-10x times slower. **Negative impact** on the performance of the system.

- **Iperf**: measures TCP bandwidth, delay jitter and datagram loss.

  Results: 801 Mbits/second vs. 37.7 Mbits/sec. **Enough for typical applications**.

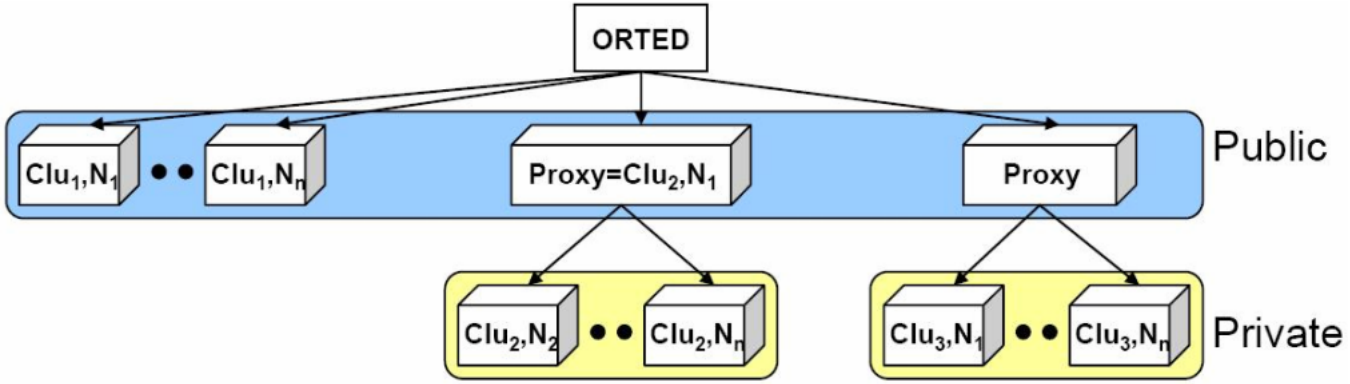
“Real-life” application benchmark

- **Gromacs**: Is a package to simulate the motion of molecular systems by computing the Newtonian equations of motion of its atoms. It can be run in parallel, using standard MPI communication. The results are:
  - Serial (single processor systems):
    - 2.8% slower when using a virtual server. **Very good.**
  - Parallel (symmetric multi processing systems SMP):
    - Physical worker nodes: 71% of peak scalability. **Very good.**
    - Virtual worker nodes: execution 3x times longer. **Very bad**.



Virtual network latency

- **Network latency results**: latency 100x times bigger when the information is transferred from a Xen host to one of his virtual working nodes.
- **Solutions:**
  - Group CPU's on a single virtual **SMP Working node** (solution adopted by CESGA).
  - Create **alternative** virtual **communication channels** (to be evaluated).
  - New resources: wait for Intel new **VT-d** direct I/O CPUs ?



Conclusions

- **Network and disk performance issues** that can represent a big difference to some applications, specially in parallel computing.
- **Network and disk virtualization in Xen are not mature** enough by now to be implemented in a production environment but are more a quick response to implement hardware virtualization with some issues not yet resolved.
- This technology **has to be implemented in** other high-end hardware as **low latency networks like Infiniband or Myrinet** to be really useful in high performance computing.

Questions

Questions are welcomed