

# Virtualization

Physics at Terascale Workshop Karlsruhe

21.02.08

**Volker Buege<sup>1),2)</sup>, Marcel Kunze<sup>1)</sup>, Jens Milke<sup>1)</sup>, Oliver Oberst<sup>1),2)</sup>, Guenter Quast<sup>2)</sup>**

**1) IWR – Forschungszentrum Karlsruhe (FZK)**

**2) IEKP – University of Karlsruhe**



# Summary

- Computing infrastructures at universities
- Virtualization
- XEN / VMWare Esx
- Server Consolidation / HA
- Virtualization in HPC
  - Dynamic cluster partitioning



Universität Karlsruhe (TH)  
Forschungsuniversität • gegründet 1825



Forschungszentrum Karlsruhe  
in der Helmholtz-Gemeinschaft

# High Performance Computing infrastructure at universities

Two possibilities to run a high performance computing infrastructure:

- Departments or user groups are running their own computing infrastructure.

- **Pros:**

- Setup will cope all their needs in hardware and software.

- **Cons:**

- Overhead in administration, administration has to be done by each group
- No load balancing between the isolated (“private”) computing clusters

- Department or user groups run a shared computing infrastructure:

- **Pros:**

- Administration can be centralised (e.g.at the Computation Centre of a University)
- Shared funding may lead to a favourable hardware price
- Loadbalancing

- **Cons:**

- Setup has to be a compromise!

# High Performance Computing infrastructure at universities II

- Compromise is not desirable/possible in some cases:
  - **Incompatibilities** between **software and operating systems (OS)** within the needs of the different groups.
  - Some groups want to participate in a Grid environment (may lead to point above). **The Grid environment should be isolated** from the local users (security...)

- Department or user groups run a shared computing infrastructure:
  - **Pros:**
    - Administration can be centralised (e.g. at the Computation Centre of a University)
    - Shared funding may lead to a favourable hardware price
    - Loadbalancing
  - **Cons:**
    - Setup has to be a compromise!



**Virtualization**

# Server Infrastructure at university departments

- Server load is often less than 20%
- Historically grown server infrastructure consumes:
  - energy, climate and space
  - Manpower in administration:
    - Hardware/Software failover
    - etc.



**Load Balancing**



**Consolidation**



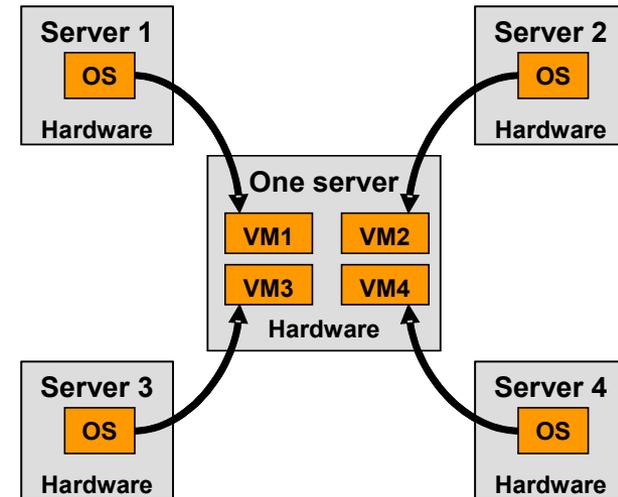
**High Availability etc.**



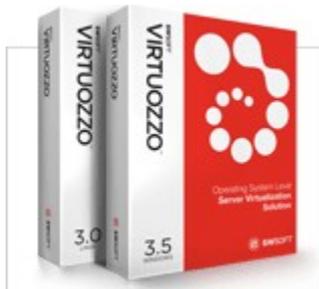
**Virtualization**

# Virtualization

- Possible Definition:
  - sharing resources of one physical machine between different **independent** operating systems (OS) in Virtual Machines (VM)
  
- Requirements:
  - Support multiple OS like Linux and Windows on commodity hardware
  - Virtual machines have to be isolated
  - Acceptable performance overhead



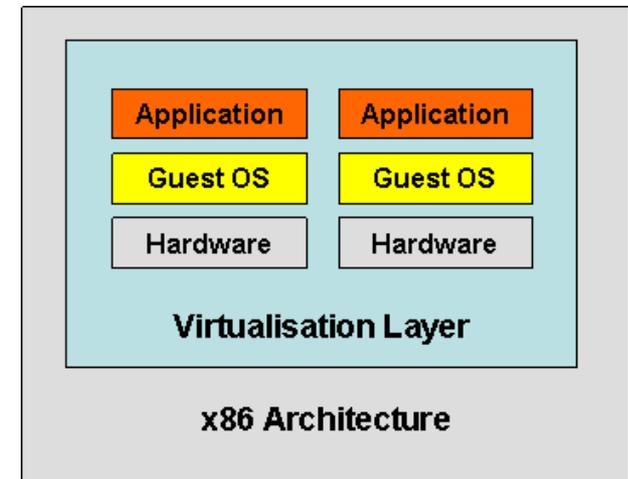
# Virtualization - Products



and many more ...

# VMware ESX

- Full Virtualization
- Virtualization layer is directly installed on the hardware host
- Optimized for certified hardware
- Provides advanced administration tools
- Near native performance while emulating hardware components
- Some Features:
  - Memory ballooning
  - Over-commitment of RAM
  - Live migration of VMs



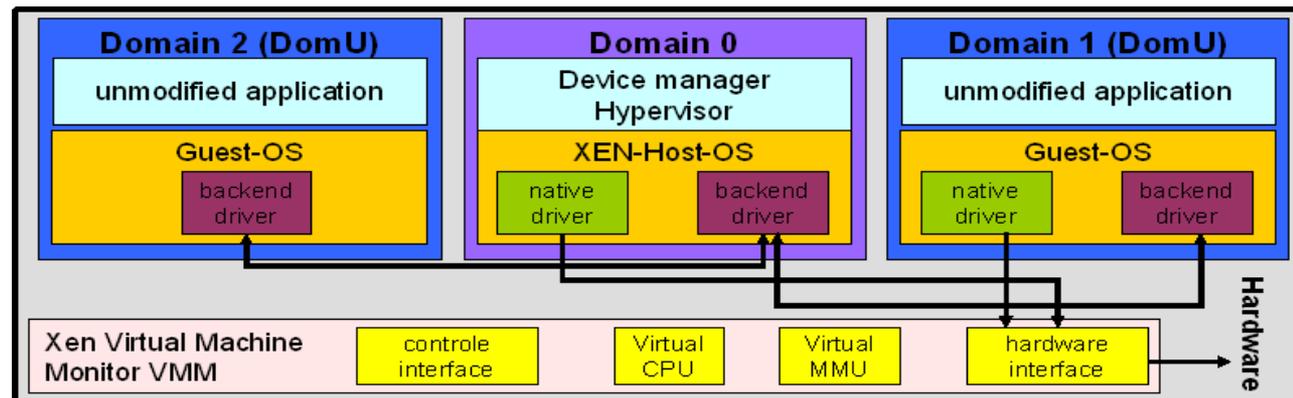
Schematic overview of  
VMware ESX-Server

# XEN (Open Source)

- Paravirtualization (or full virtualization – CPU support needed)
  - Hardware is not fully emulated → Small performance loss
- Layout:
  - Hypervisor runs on the privileged host system (dom0)
  - VMs (domUs) work cooperatively
- Host and Guest Kernels have to be adopted. But most of common Linux distributions provide XEN packages (XEN-kernel / XEN tools)

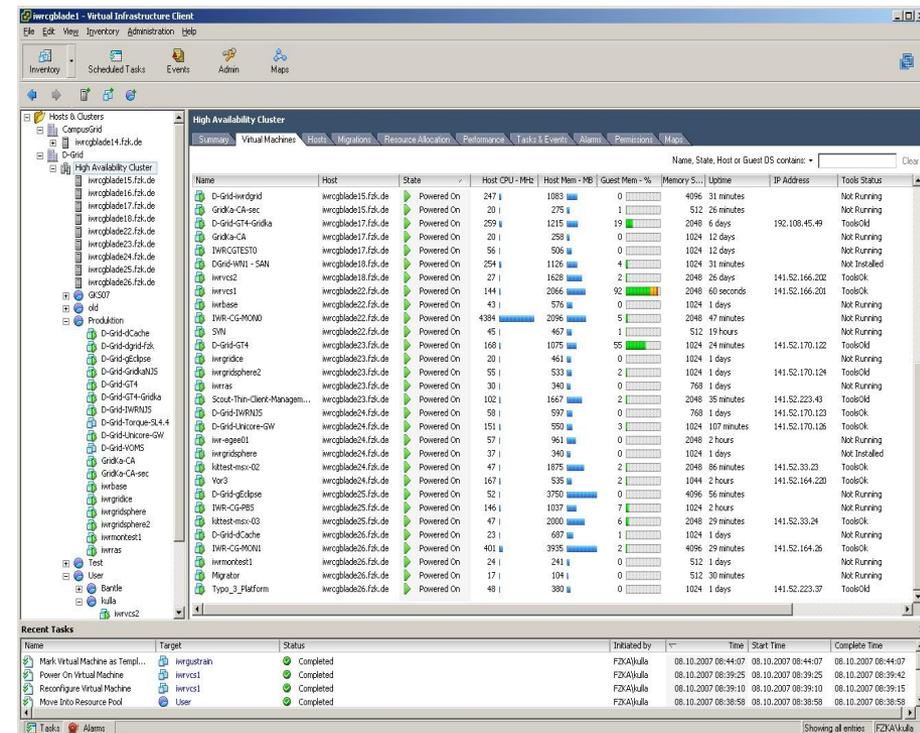
- Some Features:

- Memory ballooning
- Live-migration



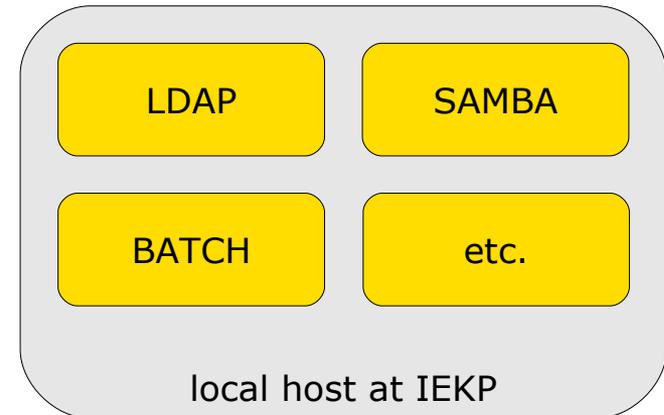
# Virtualization at IWR (FZK) – VMWare ESX

- Two ESX Environments:
  - Production:
    - 10 hosts (Blades) used
    - 30 VMs running D-Grid servers
    - 50 VMs others
  - Test:
    - 4 hosts used
    - 40 VMs
- ESX @ Gridka-School 07
  - ~50 VM for the workshops
    - gLite Introduction Course (UIs)
    - Unicore
    - ...



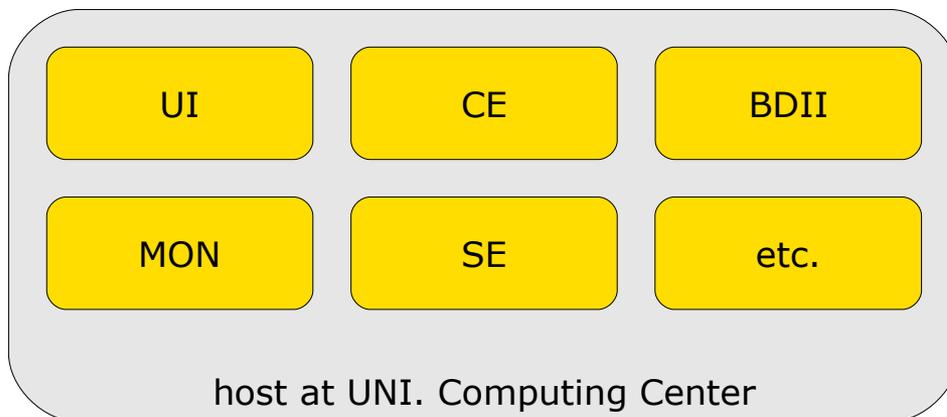
# Virtualization at IEKP (UNI) – XEN

- Two main server infrastructures:
  - local services (ldap, cups, samba, local batch system, .... )
  - gLite grid services of the UNI-KARLSRUHE Tier 3 site
    - moved to Computing Center of the University test cluster from local IEKP cluster



## ■ Virtualization Hardware:

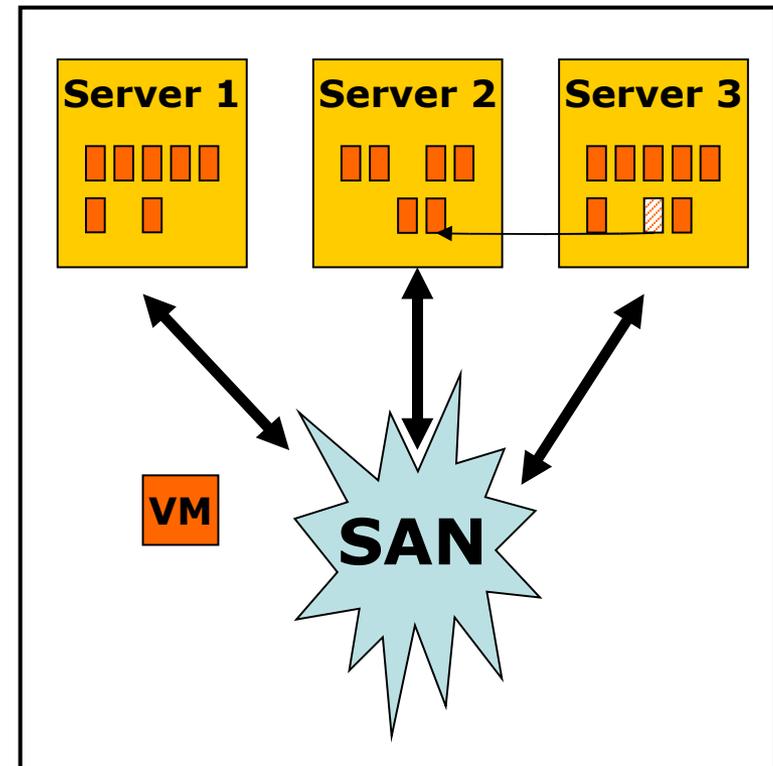
- Two hosts (local IEKP):
  - AMD Athlon 64 X2 4200+
  - 6 GB RAM
  - 400 GB Raid10 disk space for VMs
- Virtualization Portal at Uni. KA computing center:
  - 2x Dual-Core AMD Opteron
  - 8GB RAM
  - 400GB Disk Space



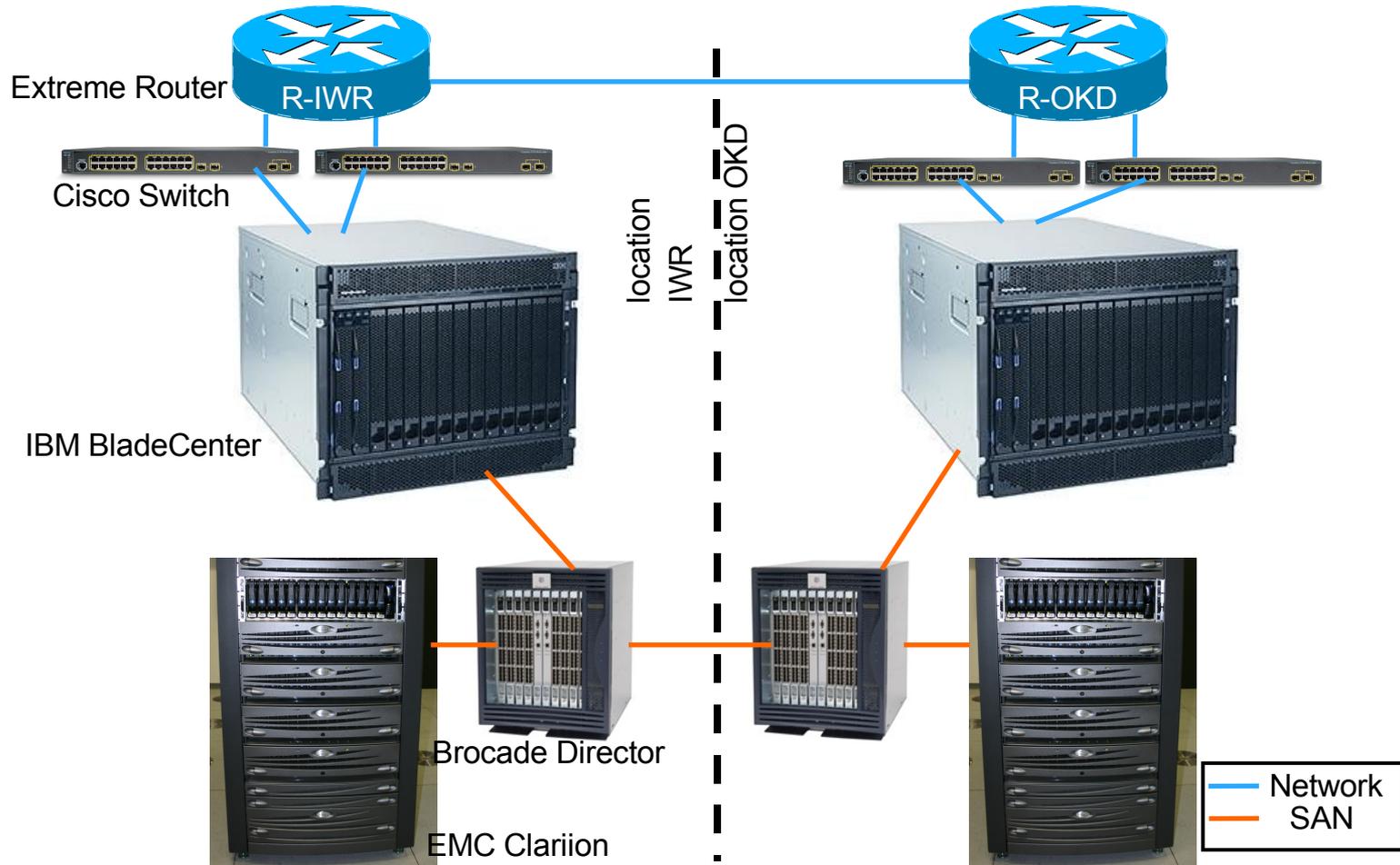
# High Availability I

One approach:

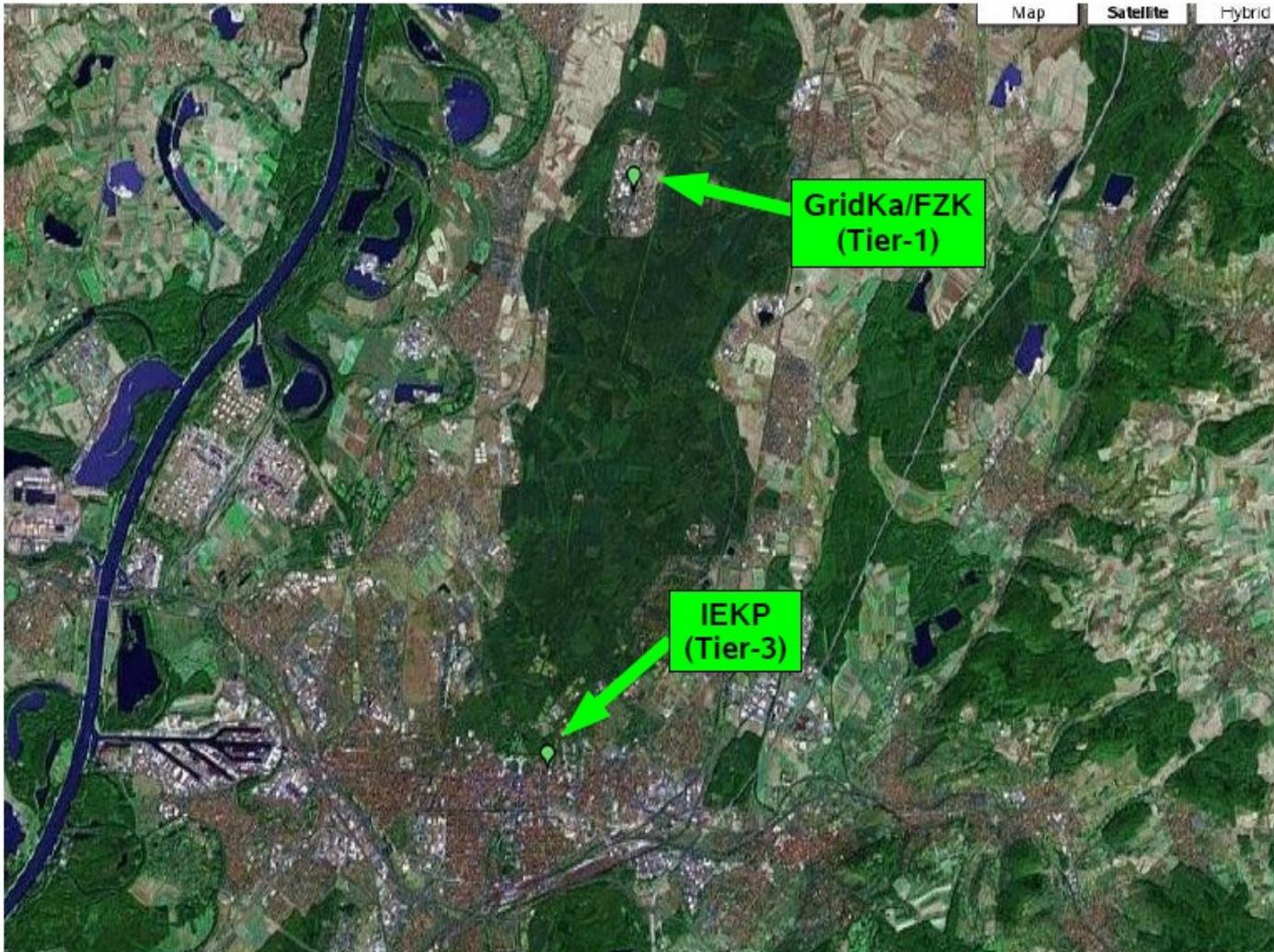
- Storage of the **VM file system** in a highly available and redundant **SAN**
  - Use **host systems** with **redundant LAN, SAN and power connections**
- **Live migration** in case of hardware problems or maintenance of a hardware box
  - **Load balancing** between the hardware server
  - Automated tools for both exist e.g. for the VMware ESX server.
  - **All services can be offered without** or only a short **interruption**



# High Availability II – Example IWR-FZK



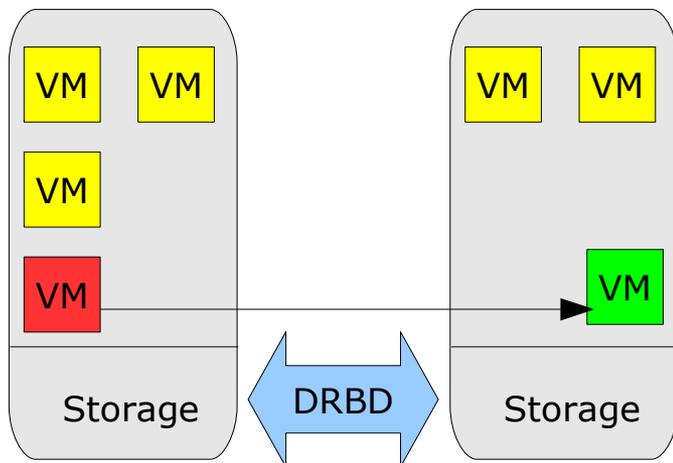
by Fabian Kulla



# High Availability III

Second Approach (“Poor Man's Solution”):

- Combination of spare machines and SAN is an overkill if only a few critical services are hosted (example: IEKP)
- Solution should be without too much hardware overhead
- Possibility: Use two powerful host machines with same architecture in combination with a *Distributed Replicated Block Device* (DRBD) to mirror disk space between the machines (Raid 1 over Ethernet) for the VM images



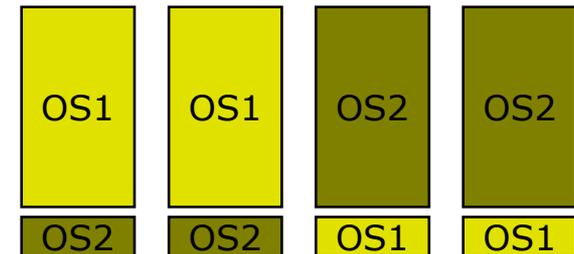
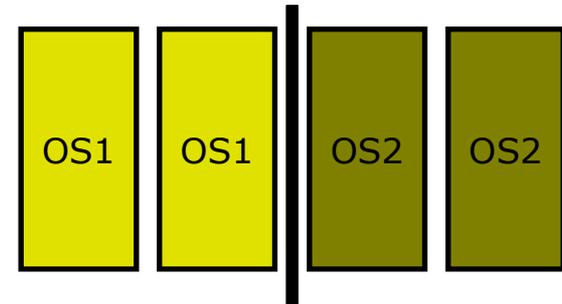
- In case of hardware problems or high load the machines can easily be migrated
- Not yet implemented:
  - Heartbeat: in case of complete hardware breakdown the machines will be restarted on the other host

# Dynamic Cluster Partitioning Using Virtualization

- Motivation:
  - Shared Cluster between several groups with different needs (OS, architecture)
    - Example: New shared cluster at the University of Karlsruhe computing center (in March 2007)
      - 200 worker nodes:
        - » CPU: 2x Intel Xeon quad core
        - » RAM: 32 GB
        - » Network: Infiniband / Gigabit Ethernet
      - ~200 TB Storage:
        - » File system: Lustre
      - OS: Suse Linux Enterprise 5
      - Shared between 7 different university institutes

# Dynamic Cluster Partitioning Using Virtualization

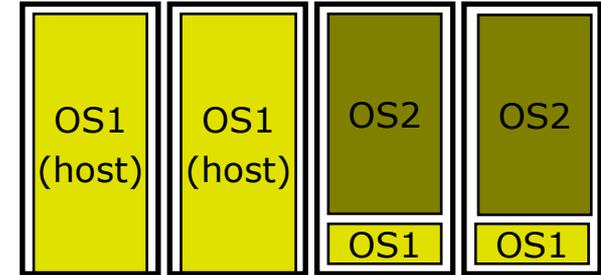
- Static partitioned cluster:
  - No load balancing between the partitions
  - changing the partitions is time consuming
  
- Dynamic partitioned cluster:
  - First approach (tested on IEKP local production cluster):
    - Using XEN to host the virtualized worker nodes
    - All needed VMs are running simultaneously. Minimum memory is assigned to the not needed VM
    - Managed by additional software daemon controlling batch system and VMs
    - Tests were run for several weeks on local IEKP cluster



# Dynamic Cluster Partitioning Using Virtualization

## ■ New Approach:

- Pre-configured VM Images
- “wrap jobs” start the VM on the host worker node and pass the original job to the booted VM
- Finishing jobs stop the VM after job output is passed out
- Job cancels simply kills the VM instantly



## ■ Main Advantages:

- “Bad” grid jobs which may leave bad processes in memory are intrinsically stopped and modified VMs are removed after job
- No software is needed everything is done by the batch system
- VM Images could be deployed by the VO with tested software installation!!

## ■ Performance:

- measured a performance loss of about 3-5% with experiment software (CMSSW)
- VM boot time: about 45s at the test cluster (old hardware)
- the possibility to participate within the shared cluster makes that acceptable

# Conclusion/Outlook

- Building up a shared computing cluster together with other institutes has several advantages
- The problems of such shared clusters can be solved using virtualisation techniques
  - Virtualisation of the batch queuing system offers all groups the needed OS but keeps the possibility of opportunistic use
  - It allows the consolidation of many servers on few host machines ...
- ... but we need concepts in case of failures of such a host machine
  - Spare machines and high available SAN for larger computing centres
  - DRBD can be used to build up a high availability infrastructures for a limited number of VMs
  - Concept currently under investigation at the IEKP

Questions?

**[Oliver.Oberst@iwr.fzk.de](mailto:Oliver.Oberst@iwr.fzk.de)**

**[Jens.Milke@iwr.fzk.de](mailto:Jens.Milke@iwr.fzk.de)**