Virtualisation

Owen Synge (DESY HH)

Owen Synge
Virtualisation
GridKa School, 5/9/2011

HELMHOLTZ ASSOCIATION

DESY

Overview

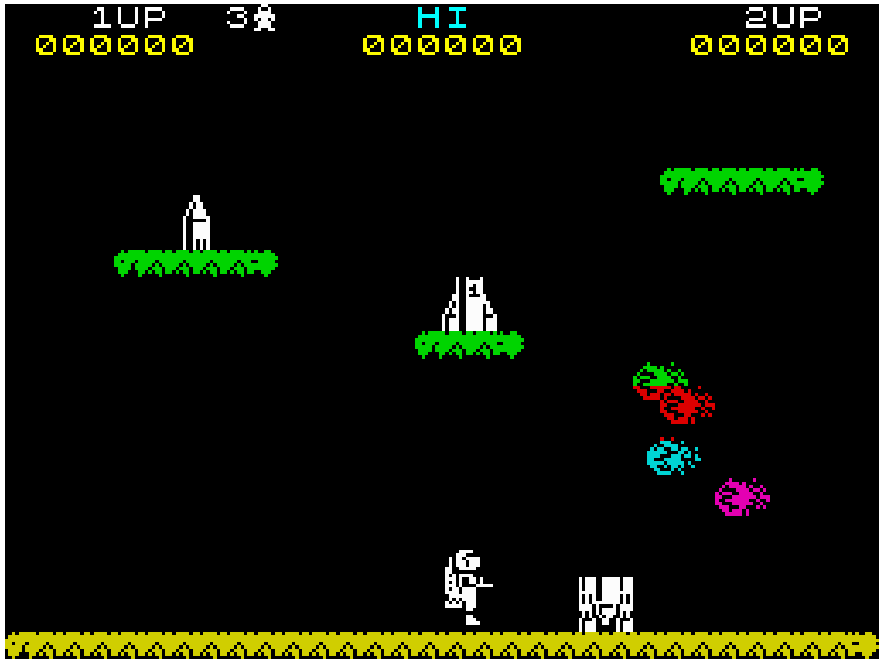
- Background to Virtualisation
- Why use it?
- Who is using it?
- What is going to be done with it?
- What about Clouds?

My first computer!

- What was yours?
  - ZX Spectrum
  - Comadore 64
  - Amega
  - IBM PC
- I wanted to play a spectrum game.
  - My spectrum was not available.
  - The software was available.
  - I had an Intel 286 based computer.

Emulators.

- Emulators.
  - An piece of software or hardware that “Emulates another computer”
  - Software can allow me to play my old computer game.
  - Intel/AMD CPU's Emulate the 8086 CPU.
- This is a from of Virtualisation!
  - We are not going to talk about old games!



History of Virtualisation.

- IBM operating systems in the 1960's did not support multiple users.
  - Failed project to make a modern time sharing operating system.
    - Lets Change the Hardware its easier ?
  - CP/CMS with virutalised computers reaches production in 1967.
  - VM/370 released in 1972 IBM supports virtualisation as center of mainframe computer.
- Operating system level virtualisation
  - The first “chroot”
  - Solaris - Containers/Zones
  - BSD - Jails, with FreeBSD 4.0 (2000)
  - AIX - “workload partitions”
  - Linux – OpenVZ / Virtuozzo
- Modern Full and Para virtualisation
  - Vmware, Xen, Kvm and many many more.

Emulators

- Definition
  - Application or hardware that behaves like another type of hardware.
- Advantages
  - You don’t need the old hardware (like a Commodore 64)
  - Support many old CPU's
    - I was taught assembly on a PDP 7 at university.
      - No I am not old enough to have used a real one
  - Executing application does not know its in an Emulator.
- Disadvantages
  - Slow, Inefficient, Resource intensive.
    - Intel 386 was to slow to emulate my ZX spectrum without tricks such as frame skipping.
  - Complex to implement.
    - Need full understanding of original hardware.
      - Amiga Emulation took a long time to get it working.
- Summary
  - No place in a high throughput compute cluster.
  - Useful for cross platform testing, and development.

Operating System Level Virtualisation.

- Definition
  - Operating systems provide environment for applications.
  - Multitasking OS's can run more than one application at same time.
  - Why not run multiple environments and application at the same time?
- Advantages.
  - Native OS performance.
  - OS ensures applications cant effect one and other.
- Disadvantages.
  - Only one OS can run at a time.
  - OS is providing application environment isolation.
    - UNIX is not good at application isolation
      - Ever seen a fork bomb?
- Summary.
  - Useful in many environments when performance is critical.
  - Consolidating servers.
  - Improved isolation of applications.
    - Running Ubuntu on an Android phone anyone?

Hardware Virtualisation.

- Definition : Popek and Goldberg virtualization requirements (1974)
  - Equivalence / Fidelity
    - A program running under the VMM should exhibit a behavior essentially identical to that demonstrated when running on an equivalent machine directly.
  - Resource control / Safety
    - The VMM must be in complete control of the virtualized resources.
  - Efficiency / Performance
    - A statistically dominant fraction of machine instructions must be executed without VMM intervention.
- Some Hardware capable of doing this
  - System/370 (Main frame)
  - Power PC (Main frame)
  - SPARC (Unix punks)
  - IA-64 (Unix Punks)
  - Amd64/IA32 with either AMD-V or Intel VT-x extensions (Commodity hardware)
    - Now a normal desktop or laptop can do virtualisation.
    - This is exciting, virtualisations not just for big Iron.

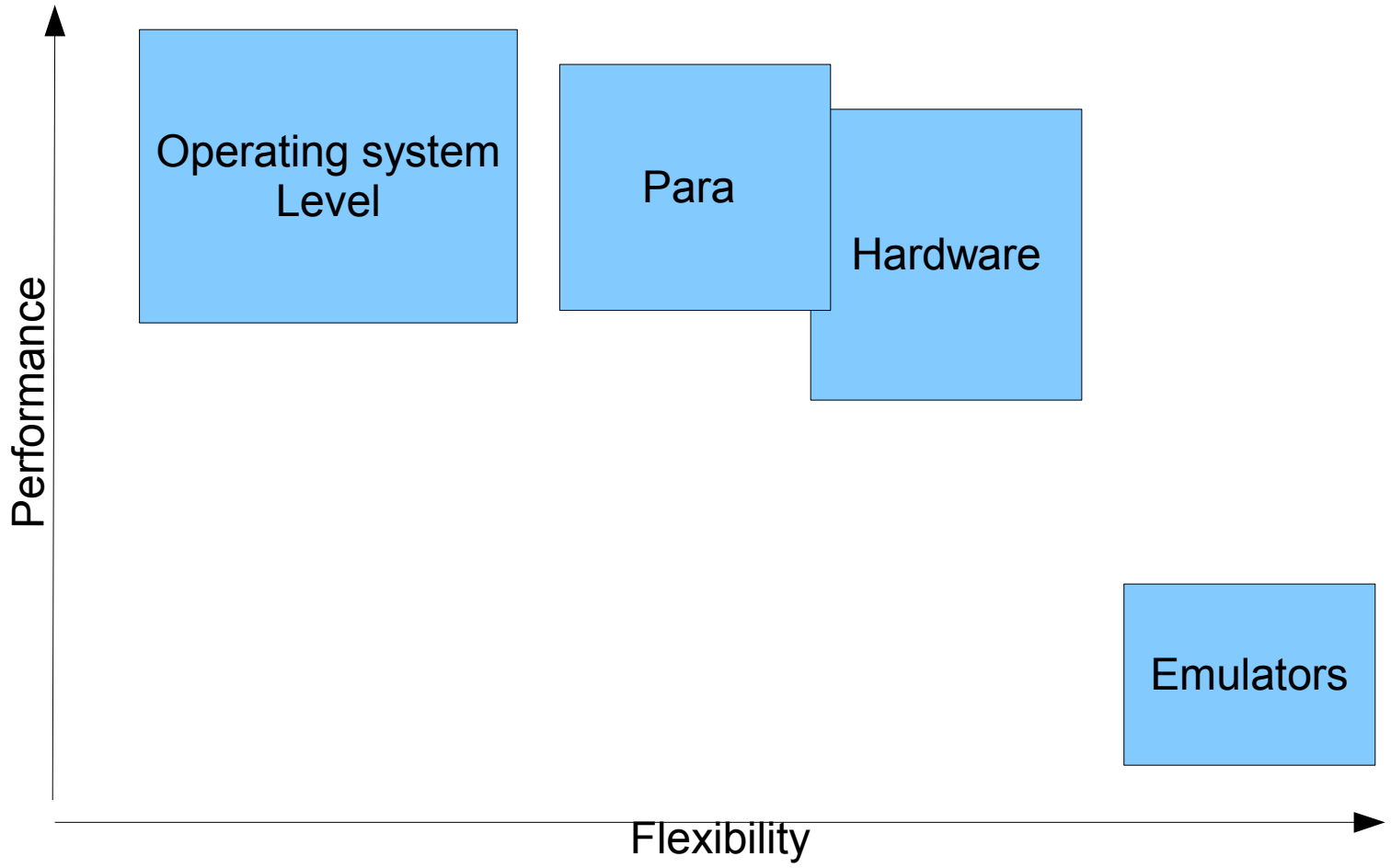
Hardware Virtualisation Pros and Cons.

- Advantages.
  - Can run different Operating systems on same hardware.
    - eg. Linux running Windows VM's is not an issue.
  - Hardware provides Isolation between operating systems.
  - Decoupling of VM and VMM operating system is complete.
    - VM crash should not effect VMM layer.
- Disadvantages.
  - Performance is effected by having multiple levels of scheduling by multiple OS.
  - VM Hardware must match Physical Hardware (drivers can isolate details).
  - Performance on accessing resources accessed by multiple OS's can suffer greatly.
    - Intel and AMD are working on networking and Disk performance.
- Summary.
  - Very useful for running applications that must run on a foreign OS.
  - Great for consolidating services.
  - Great for OS portability testing.

Para Virtualisation

- Definition.
  - Hybrid between OS level Virtualisation and Hardware Virtualisation.
    - Typically using 'drivers to communicate between operating systems
- Advantages
  - Performance can get closer to OS level virtualisation performance.
  - Isolation is better than just OS level virtualisation.
- Disadvantages
  - Isolation is typically closer to OS level performance than Hardware level isolation.
    - So VM may effect VMM layer.
  - Coupling between OS of VM and OS of VMM. (Some kernels work together some don't)
    - Need to support this in VM and VMM.
- Summary.
  - Faster than hardware virtualisation.
    - KVM and XEN are usually used para virtualised.
    - 4% CPU and Network overhead is possible.
  - Para virtualisation is suitable for Worker node virtualisaiton.
  - You cant run Windows 95 on a Linux box using para Virtualisation.
    - OS support required.

Forms of Virtualisation



Whats the difference between a VM and a Real Machine

- Not much (see definition on previous slide).
- Easy to snapshot. (so you can roll back changes)
- Potential for High availability. (moving OS across machines)
- Can share hardware so reduce energy demands.
  - Even RAM can be shared.
- Hardware can be reassigned while running.
  - Adding a CPU to a running system.
- Higher latency.
- Poor latency.
  - This is getting better.
- Slow disk access.
- Failures can be bigger.

What should we use VM's for?

- Software testing.
  - 30 seconds to restore a VM to its original image.
    - For me with vmimagemanager
  - Can be easily scripted on the VM host.
  - Is used by Me, Etics, EGI certification testbed.
- Consolidation of resources.
  - Most servers spend most of their time doing nothing.
  - Ideally services with low disk IO.
- Long term application environments
  - Like reusing my old ZX Spectrum games.
    - LTDA= Long term data Analysis?
- Worker node flexability.
  - Migrating all users to same OS at same time is not easy.

Virtualisation for testing.

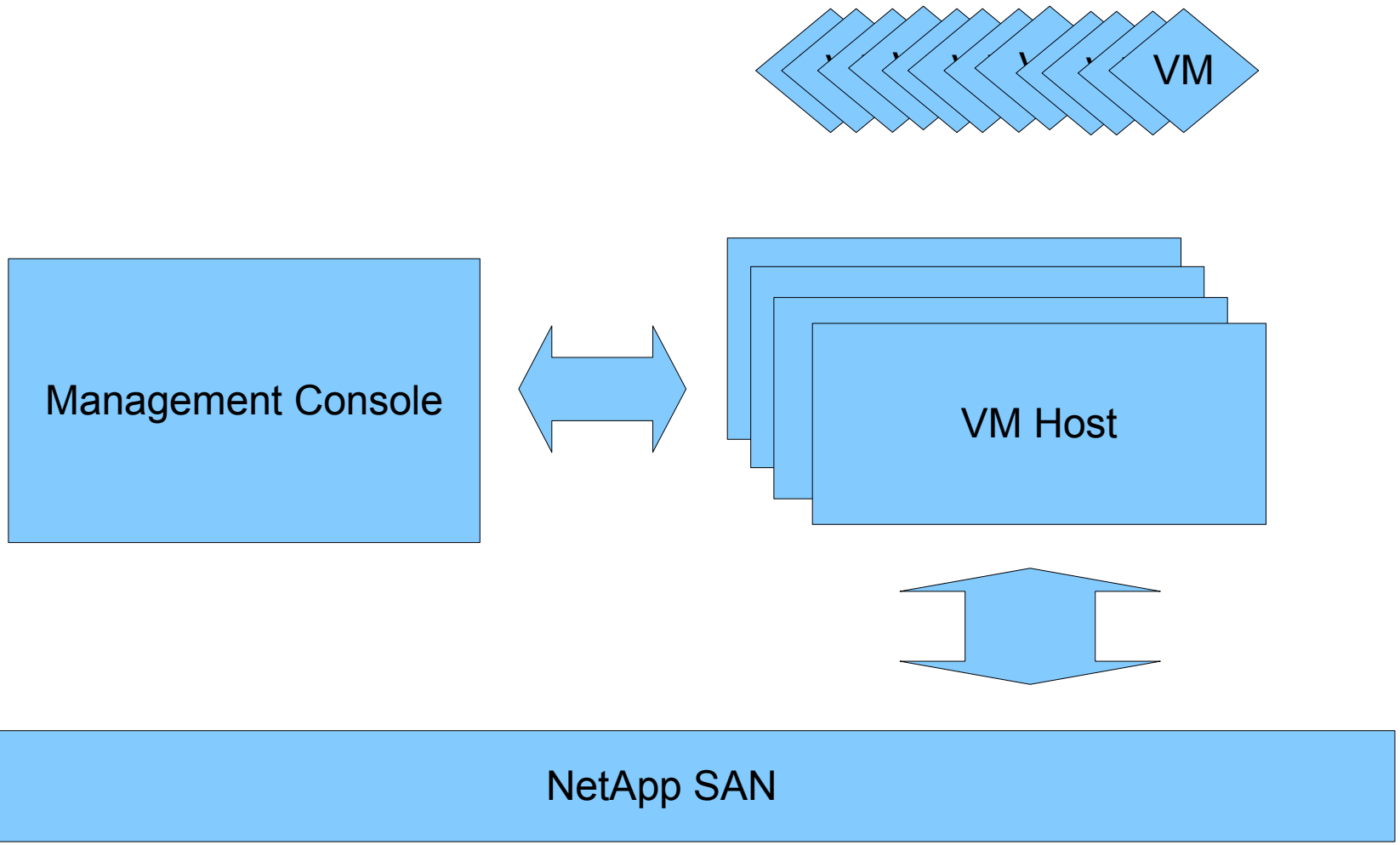
- Common for deployment testing.
  - Grid Irland, CERN, and my self been doing this for more than 5 years.
  - Quattor, Puppet, YAIM configuration management,
  - Great benefits in speed of resetting machines.
- Common for dependency testing.
  - All dependencies are installed from a base image.
  - Trap dependency changes in a nightly build
  - Etics, and myself been doing this for more than 4 years.
  - Developers have a nasty habit of adding dependencies
    - I do it myself.
- Testing large clusters.
  - If not performance critical this can be useful.

DESY

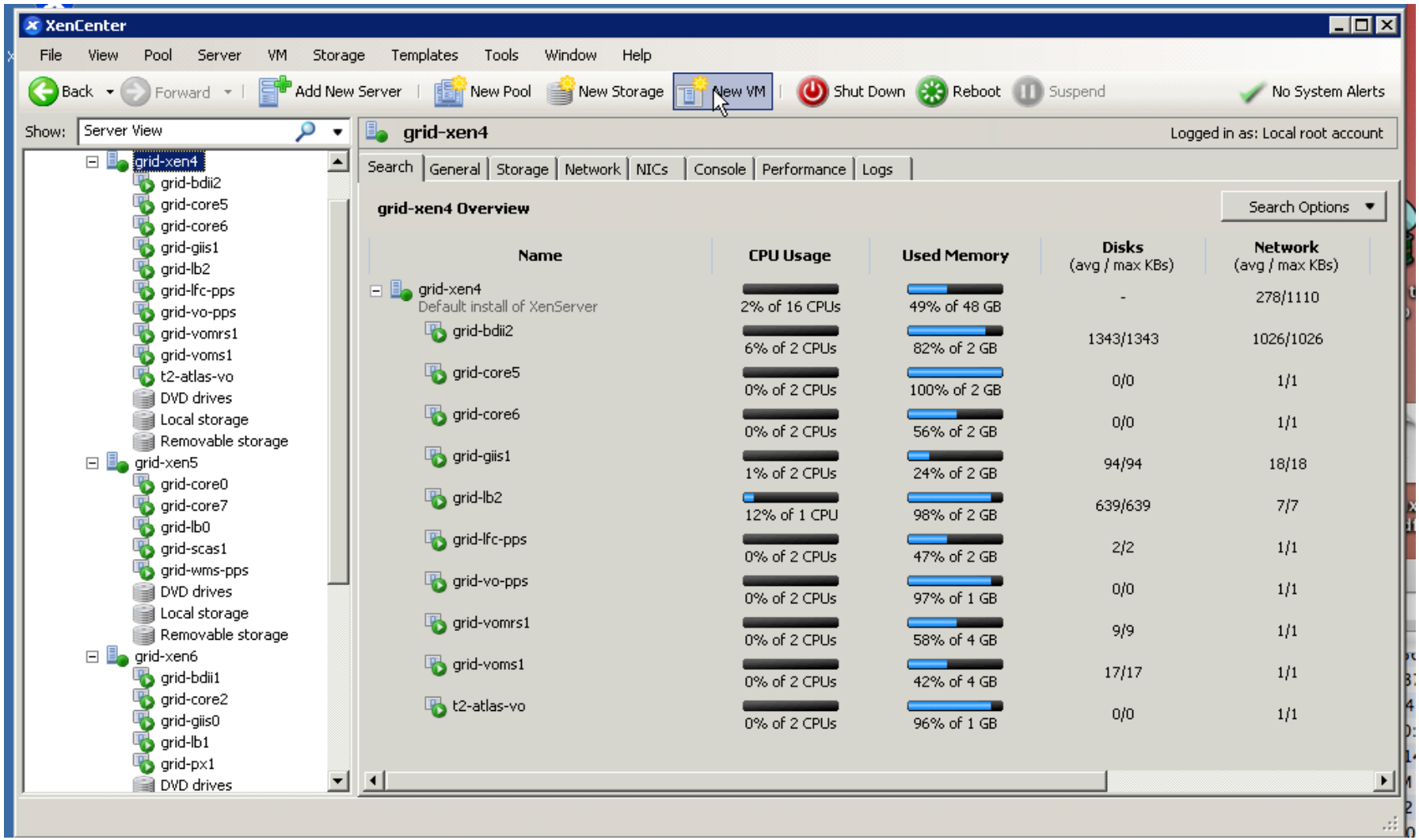
Desy Xen Cloud : Consolidation - Hardware and numbers.

- Many Available solutions from Vmware, Redhat, OpenStack, M$, etc.
- 300VMs with different OSs
  - Windows 2003 2008 XP 7 and Debian, SL, Oracle,Ubuntu, Solaris
- 8 DELL R815 48Cores AMD MagnyCours with 128GB RAM
- 5 R610 12Core Intel Gulftown 96GB RAM
- 1 Netapp FAS6040 with 20TB over Fiber Channel Brocade Fabric
  - ISCSI was a disaster
    - Imagine al l300 VM's loosing write access to their disc.
  - IO Performance is still an issue.
    - Periodic latency spikes in Disc latency.
- Although we had problems this is a good solution for consolidation.
  - We would do this again if we had not done this at DESY.

Xen Service layout



Desy Xen Cloud : Management console.



Open Stack

- Controler Node
  - Runs Certificate Authority for security.
  - Runs message queue.
- Compute Node
  - Runs VM's and requests them from Object store.
- Object store
  - Stores snapshots of images.
- Image service.
  - Registers images for creating or installing on VM's

A validation system for data analysis in HEP using virtualization

- motivation
- concepts and design
- walk through the implementation
- summary and outlook

Yves Kemp (DESY IT), Marco Strutz (HTW Berlin)

Fifth Workshop on Data Preservation and Long Term Analysis in HEP
Fermilab, 05/16/2011

… but first some thoughts about “Pizza Preservation”

How to preserve a pizza?

- Couple of days
  - Fridge
- Couple of month
  - Deep freezer
- Couple of years???
  - Preserve the recipe
  - Practice it often: You will not forget the recipe and you can detect variations in external dependencies

Putting software in the fridge or in the deep freezer

- How? Ranges from just “saving the source code” to build complex cloud-like virtualization production frameworks
- Pro’s and con’s have been discussed at many occasions … personal summary
- Pro’s:
  - Easy to do (manpower), easy to do (time)
- Con’s:
  - Runability of the software and correctness of results not guaranteed
  - Changes if needed will become more difficult the longer SW is frozen
- Freezing SW OK if timeline and scope reduced
  - E.g. makes perfectly sense for BaBar SW and analysis
- … but this is probably not the case for HERA: No successor experiment foreseen
  - So, cook the same recipe ever and ever again, and validate the output - automatically

DESY

• Bar Bar and the Big Freezer : Design Requirements

- Assume the back versioned OS are compromised
  - The LTDA system shall not be able to harm other systems at SLAC or outside
    - Isolation of compromised components
  - The LTDA system shall prevent accidental modification or deletion of data
  - Nearly impossible to protect against intentional acts
  - Maintain user identity for access to old OS; it can be done in simple ways (LRM, ssh)
  - Detect all compromised elements
- Directly affects the network architecture
  - Isolation of back versioned components
  - Physical hosts centrally managed by SLAC CD
  - Firewall rules

BaBar and the big Freezer : BaBar's Conclusions

- LTDA is progressing quickly
  - Prototype infrastructure ready and working
  - BaBar Framework running
- DOE in general very supportive for the LTDA project
- Other activities going on as part of the LTDA
  - Documentation and Outreach
  - Next big step: finalize the design and get ready to purchase the first half of the LTDA before the end of FY11

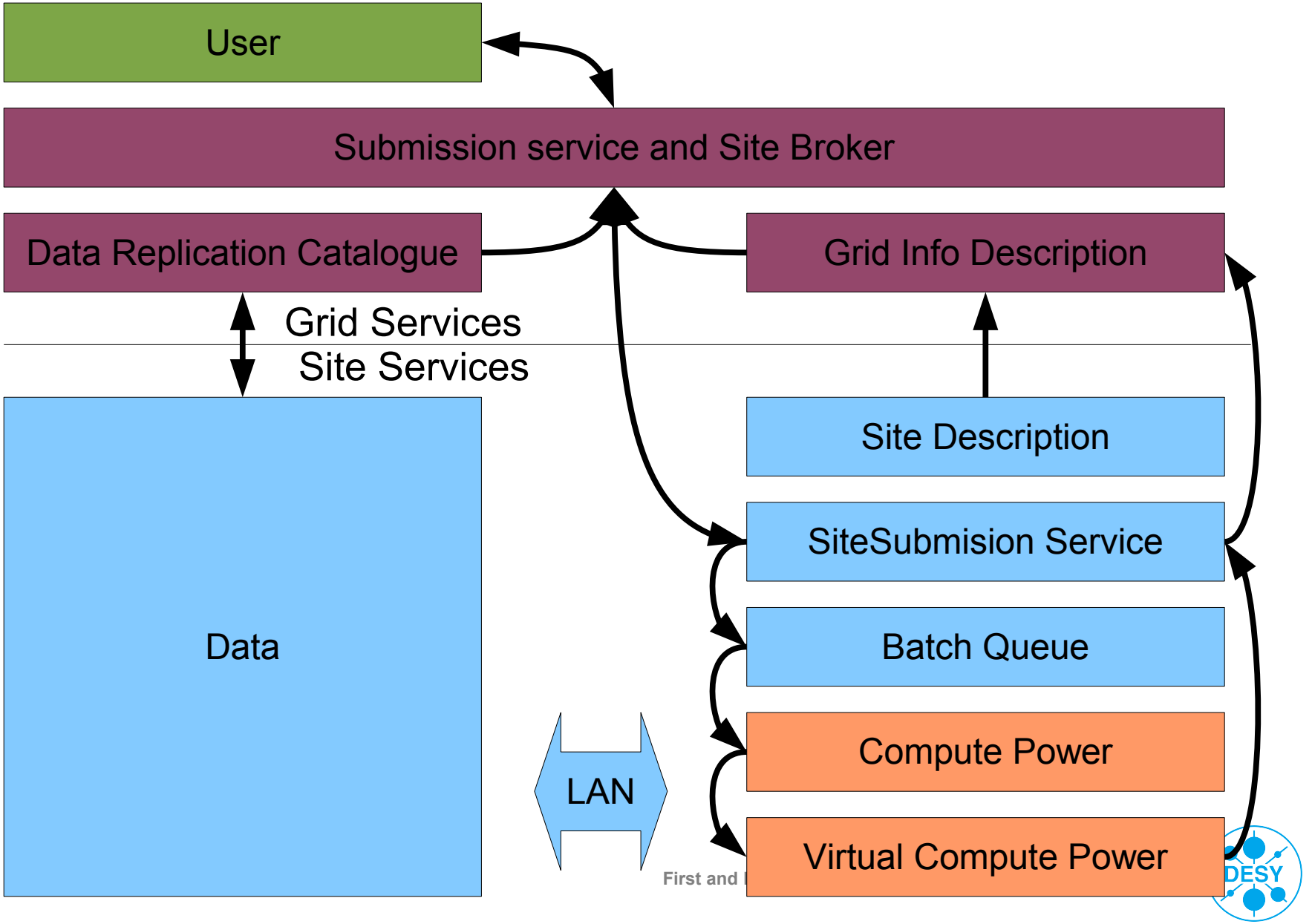
- Notes taken from Archive by Tina Cartaro (SLAC)
  - On behalf of BaBar LTDA Group

High Throughput Virtualisation some comments.

- What do HEP users do with a cloud?
  - In Canada first thing they do is install a batch queue.
- Why HEP uses batch queues.
  - To maximize through put.
  - Users don't always submit jobs when resources are available.
  - Node can always busy!
    - Small sites wont have 90% occupancy.
      - But if reliable site they will get closer to this level of use.
- Why putting Batch queue on a Cloud is silly! (at the moment)
  - Batch Queue Fair share allows resources to be scavenged.
    - One group can use another unused resources.
  - Clouds allocate resources before the VM is started.
    - Batch queues don't like their size to change frequently.
    - So efficiency at site level goes down.
  - Budgets are limited you cant just buy more hardware!
    - So Clouds will fill up, no one seems to know what to do then.

DESY

Virtualising the Worker Node.



5 Models of worker node Virtualization

- Defined at DESY virtualization workshop.*
  1) Worker node running one persistent virtual machine with a single OS image.
  2) Worker node running multiple/2 persistent virtual machines with multiple/2 OS images.
  3) Worker node running non persistent virtual machine images.
  4) Worker node running non persistent virtual machine image from a library of OS images.
  5) Worker Node running non persistent virtual machines and using user defined images.
- Models 1,2 and 3 in production in 2007 at some sites.
- Model 5 blocked by data access concerns in 2007.
  - Virtualised Networks overcome this, but what about storage access?
- Model 4 Seems acceptable to sites running HEP jobs.
  - On presenting to HEPIX in Umea 2009

16-17 January 2007
https://indico.desy.de/conferenceDisplay.py?confId=155

Images and the Issues involved

- Software has security bugs.
  - When these are discovered they must be patched fast.
  - How do we manage this?
  - How do we manage this in many sites?
  - Do we care if its securely wrapped up on a Virtual network and a Virtual PC?
    - What about storage access?
- How do we deploy images at all the sites in a Grid comunity?
  - In amazon / Rackspace this is easy as you only use one site.
- Configuring your cloud.
  - Suddenly users have to manage their cloud.
    - Cfengine/Puppet/Quattor?
  - Skills need to move from data center to experiments.
    - This is not trivial work.
    - This is a LOT of work

Image transfer Objective

- How to transfer images securely.
  - We know who made the image (Endorser)
  - We know the image is unmodified after endorsement.
  - We know the endorser cant repudiate their image list.
- Privileged images on sites must be authorized by administrator.
  - Can subscribe to an image from an image list.
  - Have minimal work for a site admin.
- Site must be able to revoke Images.
  - An image, an endorser or a set of image subscription.

Stratus Lab : Model

- Market place of images.
- RDF store of image metadata.
- Uses simple SQL like Query language for finding images.
- Images can be instantiated directly to Open Nebular Clouds.
- Currently in development.

HEPIX VWG : Publish Subscribe Image list model.

- Endorser signs an Image list.
  - Image expires if not in image list.
  - Endorser is responsible for Image.
- The sites VMIC subscribes to Endorsement list.

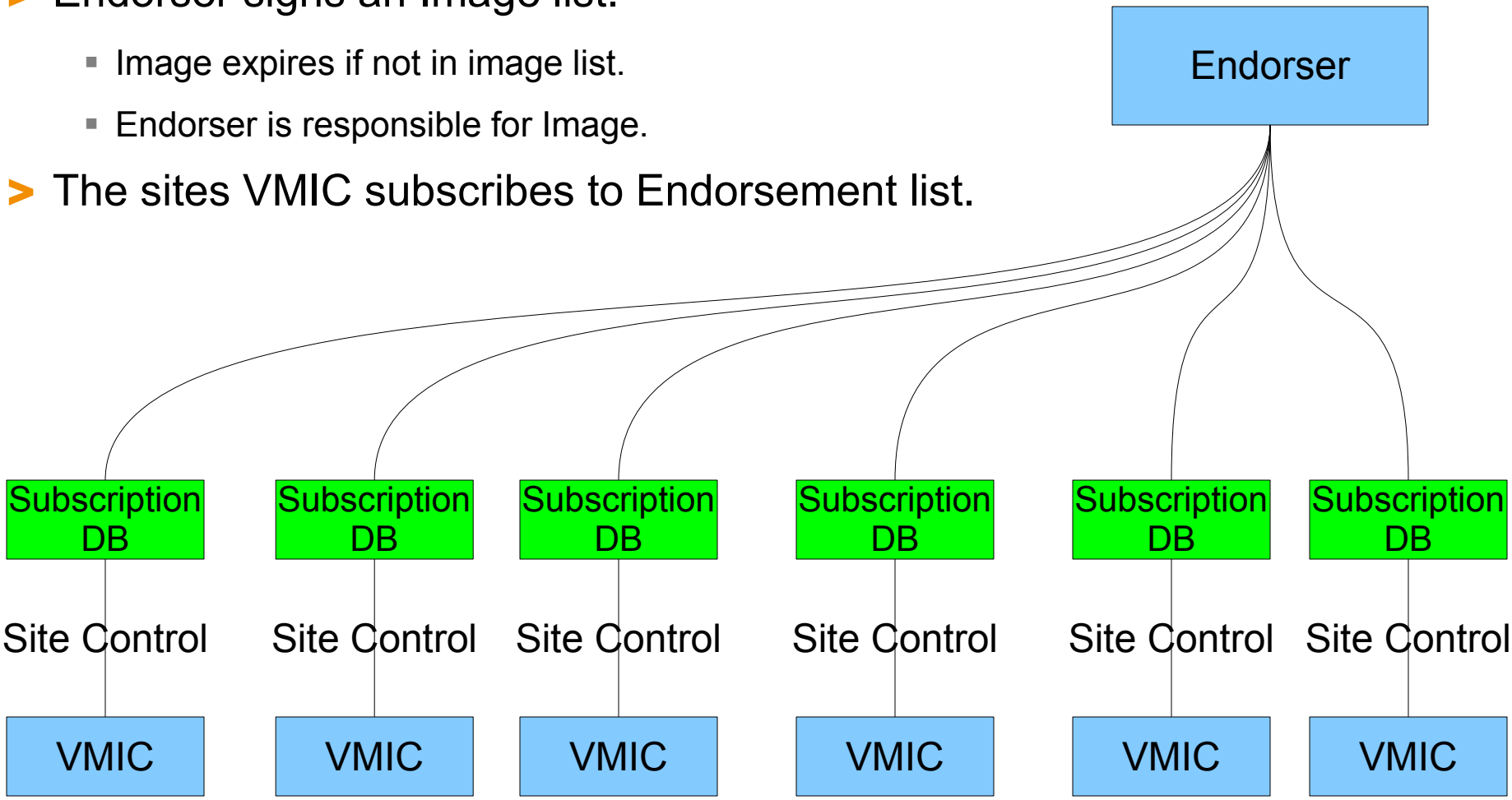


Image to Meta data Binding (Hepix VWG and StratusLab)

- Image to Meta data binding.
  - Cryptographic hashes.
    - It is easy to compute the hash value for any given data.
    - It is infeasible to generate a message that has a given hash.
    - It is infeasible to modify a message without hash being changed.
    - It is infeasible to find two different messages with the same hash.
  - Chose to use sha512 and file size to validate data.
    - HEPIX VWG Following Stratus Lab's recommendation.
  - Other hashes can be added.
    - If sha512 and size are later found to be too week.
  - URI to retrieve image.
    - Can be cached locally.
  - Each image has a UUID
    - So we know which image is expired and which is upgraded.

Signed messages. (HEPIX VWG and StratusLab)

- Meta-data authenticity.
  - X509 + signatures. (SMIME or XML signatures)
    - Gives non repudiation, and confidence in who endorsed.
    - Give tamper proof message.
    - Signature can be checked by all clients,
    - Allows checking of historic meta-data changes.
  - Version number.
    - Prevents man in middle attacks.
    - Man In Middle attempts to return an old list blocked by this.
  - UUID on Image (and Image list for HVWG)
    - Allows messages to be identified.
    - So messages cannot effect each other.
    - So images can be expired and updated.

CERNvm and CERN VMFS

- Aims to provide the single image for all wLCG computing.
  - Automatically caches latest experimental software.
- Simple image with a striped down OS.
  - Same image repackaged for many image formats.
    - Vmware, Virtualbox, Xen, KVM images all available.
  - Designed for your laptop.
    - So scientists can debug their code.
  - Designed for your data center.
    - So scientists can use their code.
- You can subscribe to their image list and always have the latest version.

Summary

- Virtualisation comes in three flavors in our data centers.
  - OS level, Para virtualisation and Hardware.
  - All are useful but for different tasks.
- Virtual machines are like real machines
  - But allows us some new flexibility (Dynamic RAM/CPU).
  - Performance overhead is now down to 3-5% for CPU and network.
  - Performance overhead of 40% for disk is not unusual.
    - We hope this reduces soon.
  - Latency is still an issue.
  - But dont use them for main storage, or RDBMS server.
- Consolidation of resources is a great thing.
  - Greatly reduces unused hardware.
- Cloud and Virtual Worker Nodes are going to be standard.
  - Image distribution is being dealt with.
  - People are publishing images today.

References

- A brief history of virtualisation.
  - http://www.theregister.co.uk/2011/07/14/brief_history_of_virtualisation_part_2/
- IBM **VM (operating system)**
  - http://en.wikipedia.org/wiki/VM_(operating_system)
- Popek and Goldberg virtualization requirements
  - http://en.wikipedia.org/wiki/Popek_and_Goldberg_virtualization_requirements