

Report on the HEPiX Virtualisation Working Group

Thomas Finnern

Owen Synge

(DESY/IT)



The Arts of Virtualization

- > Operating System Virtualization
 - Core component of today's IT infrastructure
- > Application Server Virtualization
 - Virtual server front end for load balancing and high availability
- > Application Virtualization
 - Application GUI runs e.g. on local thin client connecting to central service
- > Management Virtualization
 - Role based segmented administration model on virtual (overlapping) resources
- > Network Virtualization
 - Virtual network adapters, VLAN's, virtual routing tables
- > Hardware Virtualization
 - Similar to OS virtualization, slicing hardware to functional pieces with resource allocation
- > Storage Virtualization
 - Block and file virtualization in SAN and NAS, also RAID and iSCSI
- > Virtual People and Organizations
 - As in Second Live or robotics and as VO's for GRID
- > ...
- > Service Virtualization
 - Towards cloud computing: Simply Access Service by GUI, API or appliance



Grid and Clouds

> Grid

- Wikipedia: Grid computing is the combination of computer resources from multiple administrative domains for a common goal. Grid computing (or the use of a computational grid) is applying the resources of many computers in a network to a single problem at the same time - usually to a scientific or technical problem that requires a great number of computer processing cycles or access to large amounts of data.
- Virtual Organisations
- Middleware



> Clouds

- Wikipedia: Cloud computing is a style of computing in which dynamically **scalable** and often **virtualized** resources are provided **as a service** over the Internet. Users need not have knowledge of, expertise in, or control over the technology infrastructure "in the cloud" that supports them.
- Promises

Unlimited Computing Resources available on Demand
No up-front Commitment by the User
Pay for use of computing resources only as needed

- Main Components:

Local Internet Device
Computing
Storage
Communication (Network)





> Startup

- At the Fall 2009 Berkeley meeting

> Chair

- Tony Cass (CERN)

> Objective

- Enable virtual machine images created at one site to be used at other HEPiX (and WLCG) sites.
- Working assumptions

Images are generated by some authorised or trusted process

Images are “contextualised” to connect to local site workload management system

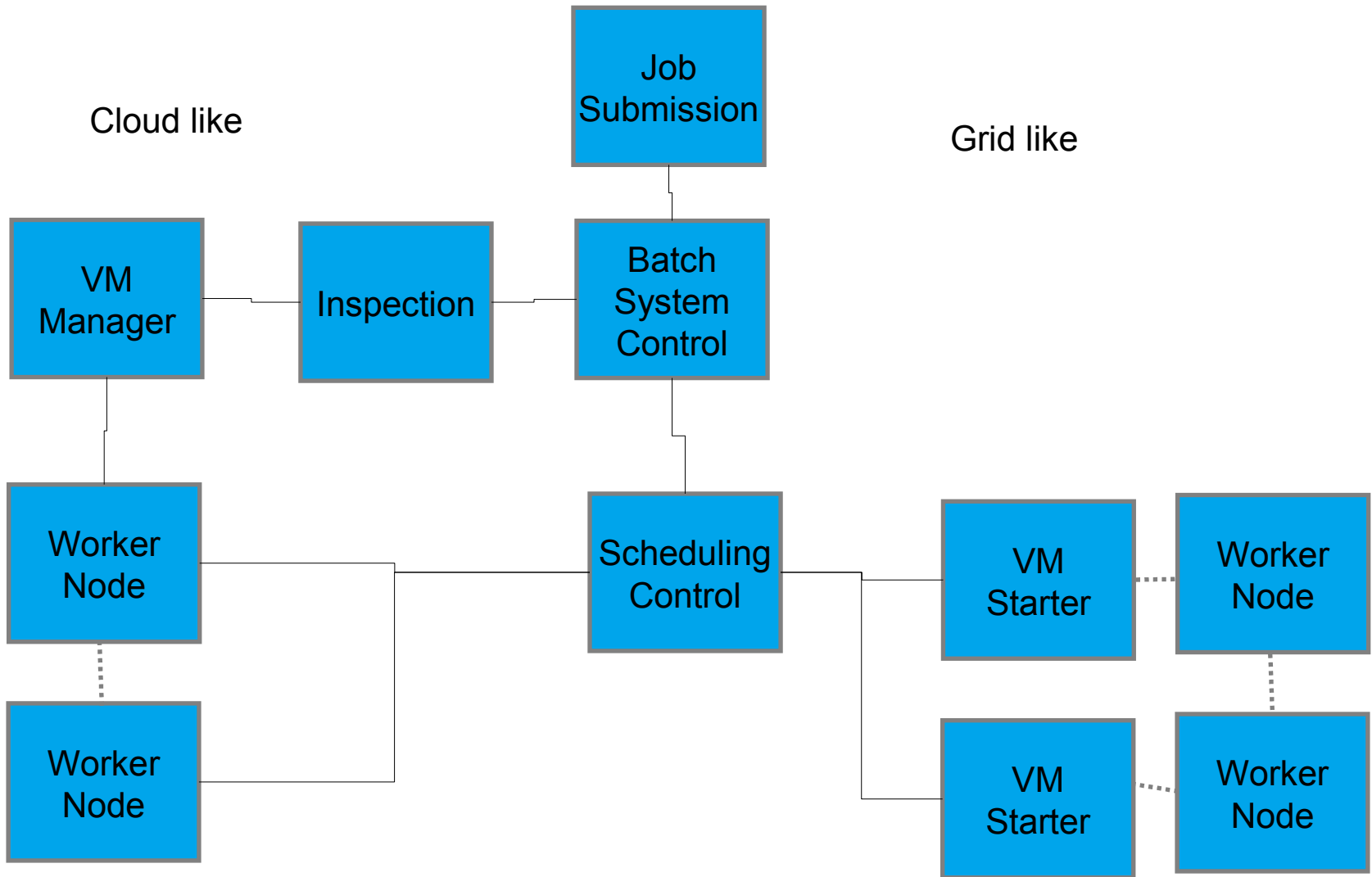
> Results

- Expected until end 2010 ...

Image Life Cycle



Batch System Integration



Background – Grids

> Grids.

- Homogeneous OS.
- Distributed job submission.
- 1000's of sites.
- Accountancy based on Job scheduler.
- Problems.
 - > Inconstancies in deployments.
 - > Synchronizing updates between VO and site.
- Benefits
 - > Experiments dont need to mange OS.
 - > Run in trusted environments (eg DESY)

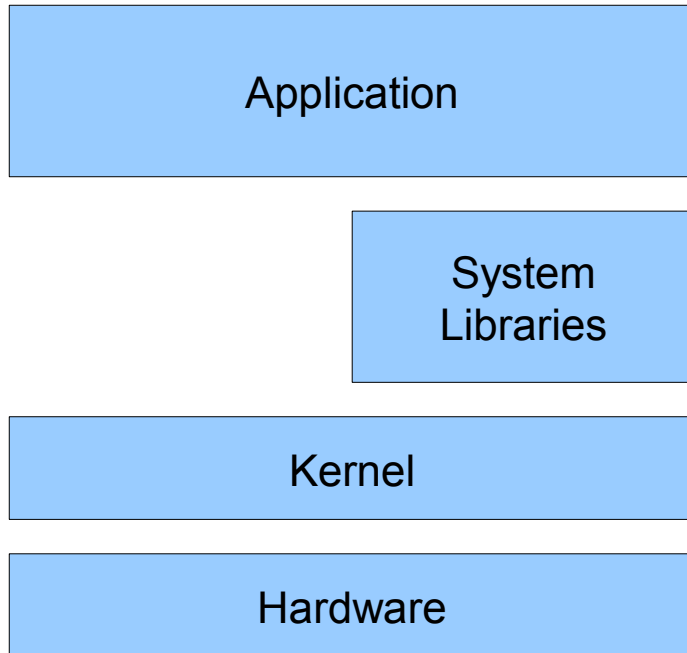


Background - Clouds

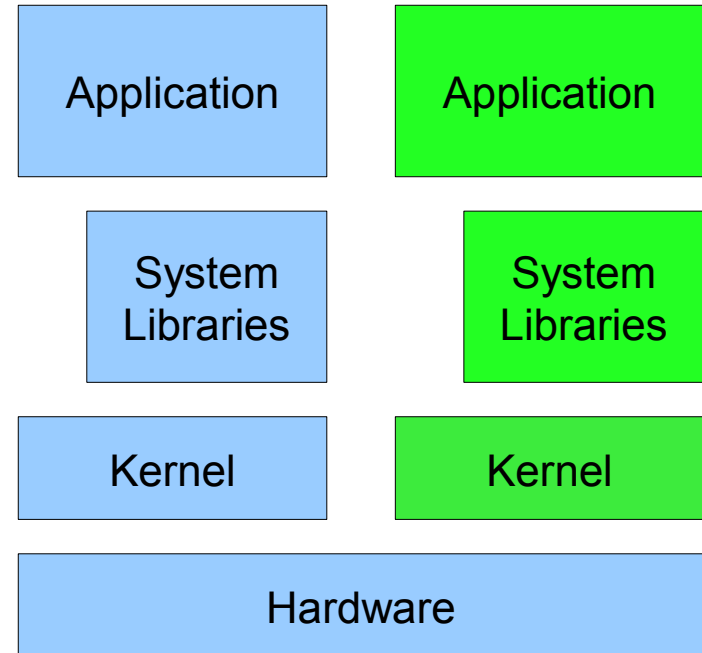
- > Shown to work for HEP in CHEP 2009.
- > Commercial system.
 - Cost based on CPU/network/time.
- > Based on Virtualization.
 - Upload OS to host, Host runs it.
- > Problems.
 - Cost of network/storage (CPU is cheap).
 - Experiment manages OS.
- > Benefits.
 - Get the OS you want.



Traditional OS



Virtual Applications



Current Virtualization use in Europe.

> Run grid jobs on virtual machines (VM).

- Currently being done at 2 big sites.

> CERN.

- Runs local cloud infrastructure.
- VM == traditional Worker node.
- VM reinstalled once a night.

> PIC.

- WN batch queue connected to Host VM.
- VM created for job, destroyed after.



Goals

- Improve Grid Computing.
 - For the VO.
- Consistent run time experiences.
 - Across sites, Across Countries and Continents.
- Common infrastructure and approach.
 - Trust building.
- Share images across sites/continents.



HEPIX assumptions.

- > Sites will have their own Virtualization systems.
 - Site Image deployment is Sites business.
- > Sites need to control which images run.
 - Security implications.
 - Audit trails. (concept of endorser)
- > Sites will not update images.
 - Potential cause of inconsistency.
 - Images will have security issues/updates.
 - > Then expire the “bad” image.
- > Jobs run as unprivileged user.
 - Some concerns of letting VO have root. (eg NFS 3)

Recipient site controls how
“payload” ends up in the image

No root access by end user
during image generation.



> Targets

- Same Image run at many sites.
- End of 2010 target for first results.

> Work on going.

- Policies and tools to make this happen.
- VO's now encouraged to join in.
 - > Mail Tony.Cass@cern.ch to join
 - > hepix-virtualisation@cern.ch



Whats missing! Before virtualisation is Grid wide.

- > Images must be trusted.
 - Who made an image?
- > Images must be generated.
 - How to make them universal?
 - How to manage life cycle?
- > Images must be shared across sites.
 - Security of images / expiring images, auditing?
- > Images must run at many sites.
 - How to integrate images?



Working group areas & Status

> Generation

- Led by Dave Kelsey & Keith Chadwick
- Likely to produce

Policy proposal for image generation process. If sites can demonstrate they meet the requirements of the policy then their images should be trusted for execution at remote sites
Recommendations for hypervisor configuration to ensure maximum security.

> Transmission

> Expiry & Revocation

> Contextualisation

> Support for multiple Hypervisors

Sites anyway expected to follow best practices.

Current discussion is around roles and endorsers for the different components ("base" operating system and VO software) and about who can be trusted.

HEPIX virtualisation 4 Groups

> Image Generation.

- Policy proposal for image generation.
 - > http://www.jspg.org/wiki/Policy_Trusted_Virtual_Machines
 - > Give sites trust in images.
- Best practice of images generation.
 - > Audit trail.

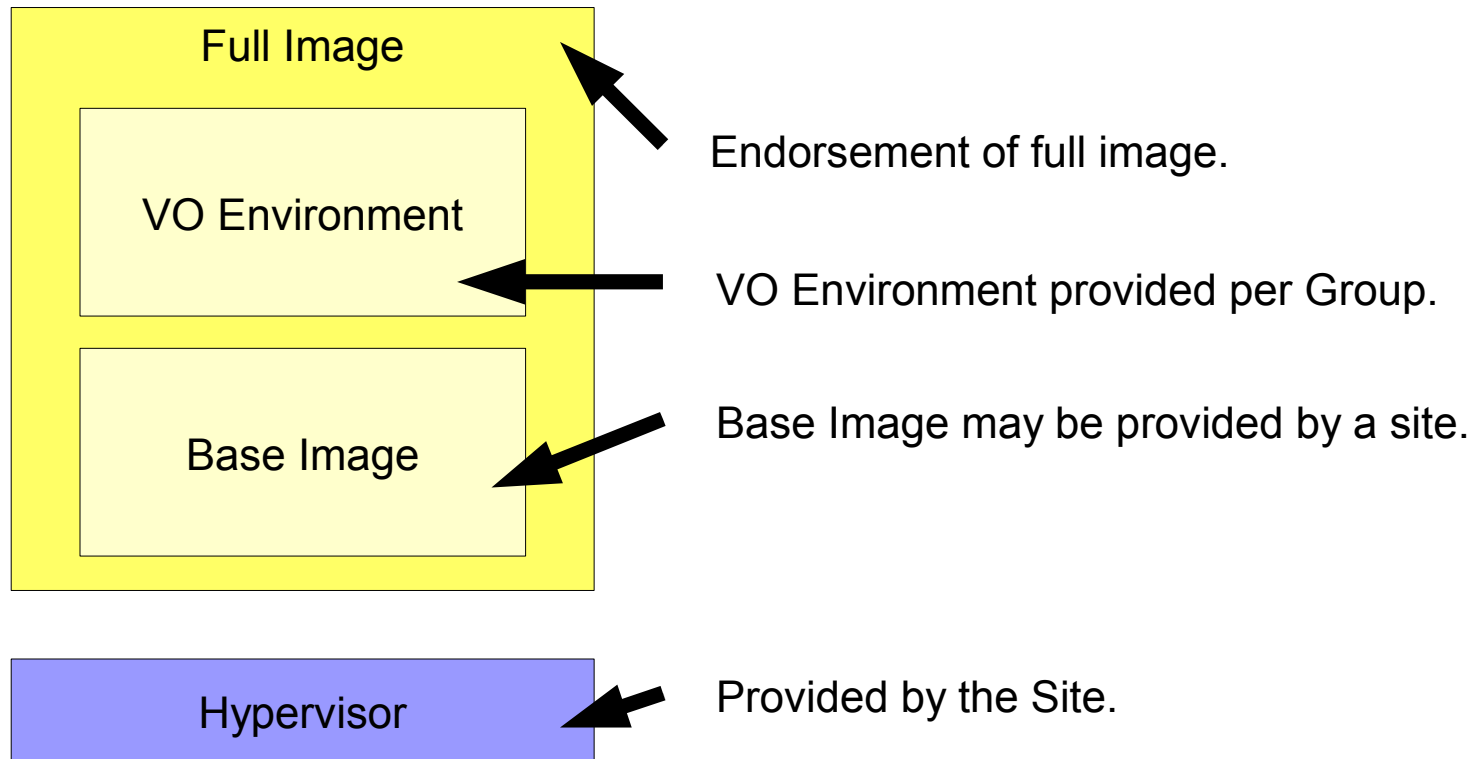
> Image Transmission.

> Image Contextualisation.

> Image + Hypervisors.



Generation of Images.



Working group areas & Status

> Generation

> Transmission

- Led by Owen Syngde
- Likely to produce

Recommendation for basic transport protocol(s) to be supported

- Prescriptive for sites wishing to generate images

Proposal for optional protocols to improve transmission efficiency

- E.g. transmission of only differences w.r.t. a reference image
- Status of “interesting” protocols such as bitTorrent likely to be an issue.

- Unlikely to comment on intra-site image transmission

> Expiry & Revocation

> Contextualisation

Will not

> Support for multiple Hypervisors

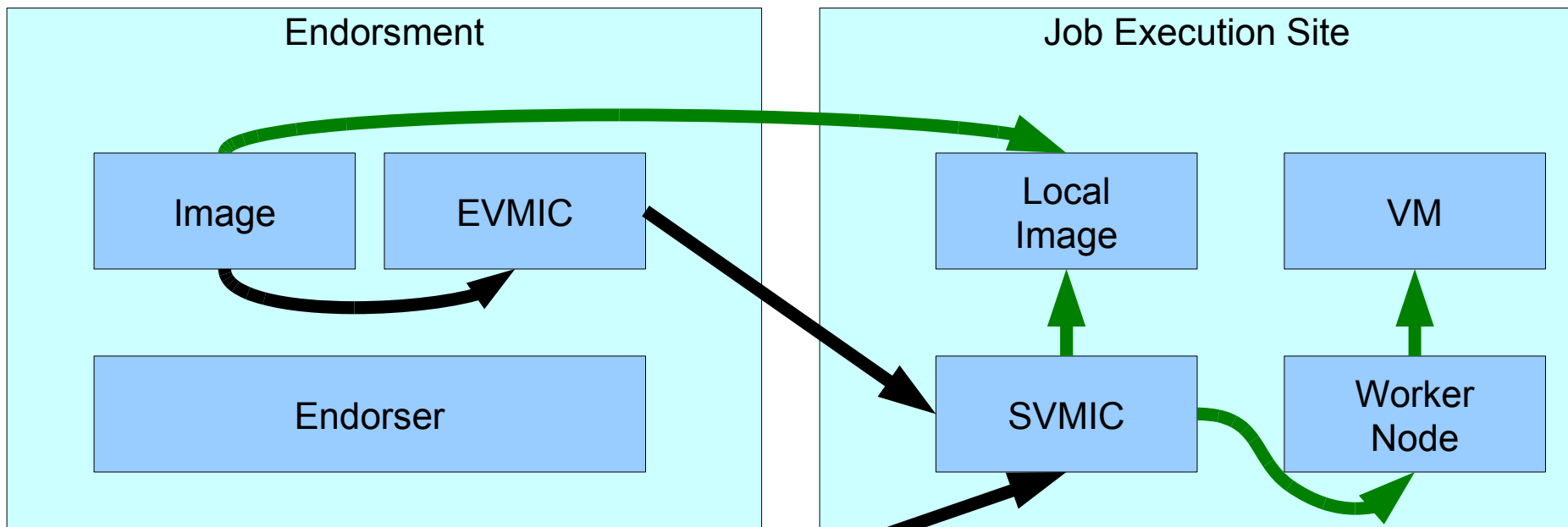
Current model is tagged images distributed in manner akin to mechanism used for VO software today.

HEPIX virtualisation 4 Groups

- > Image Generation.
- > Image Transmission.
 - Will not comment on intra site transmission.
 - Managing image trust.
 - > Validation/Revocation of images. (Like CRL)
 - Current model of VO software style distribution.
 - > Bittorrent may be controversial
- > Image Contextualisation.
- > Image + Hypervisors.



Images Endorsers and Sites.



Producer Actions.
VO Lists Endorsers.
Endorsers list Images
EVMIC holds meta data.
Expiry
UUID
Tag
VO

Site Actions.
Validate Image
Trusted chain of Endorser VO etc.
Add and remove
Image
Endorser
VO
Audit images, and see history.

Working group areas & Status

> Generation

> Transmission

> Expiry & Revocation

> Contextualisation

- Led by Sebastien Goasguen

- Likely to produce

 - Proposal for mechanism allowing site to configure image

 - File system mounted at image instantiation and automated invocation of scripts on the file system during the initialisation.

 - Final job/payload will not execute as root

 - Restrictions on aspects sites are allowed to configure

 - No changes to C compiler, perl, python, ... to be allowed

> Support for multiple Hypervisors

Only basic discussions so far.
Contentious issue is kernel patching.
Group conclusion is that this is not allowed; sites who have security concerns with an image must refuse to run this and must notify the endorser to allow wider revocation. This ensures that all sites are protected.

HEPIX virtualisation 4 Groups

- > Image Generation.
- > Image Transmission.
- > Image Contextualisation.
- > Image + Hypervisors.
 - Identified KVM and XEN as of 1' interest.
 - Ideally one image for all hypervisors.
 - Recommendations / Recipes
 - > Work with Hypervisors.
 - > Performance with Hypervisors.



Summary

- > (Nearly) Identical Virtual worker nodes across sites coming soon.
 - Maybe very soon
 - Lots of work still to do
- > Trust Issue
 - Getting sites to trust VM images is hard
- > Being an endorser is a big responsibility
 - Is the cost too high ?
- > VO and site feedback needed now.
 - Will VO's want to manage OS level details (It won't be managed by Magic)
 - Base OS/VO environment issues hidden from this workflow.
 - > Can VO and Base Image generators work this out together ?

