The batch virtualization project at CERN

Tony Cass, Sebastien Goasguen, Ewan Roche, Ulrich Schwickerath

EGEE09 conference, Barcelona

See also:

- Virtualization vision, Tony Cass, GDB 9/9/2009
- Virtualisation of Batch Services at CERN, Sebastien Goasguen, *http://indico.cern.ch/conferenceDisplay.py?confId=56353*

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Batch virtualization: Why ?

User perspective:

- Customization of images for specific use cases
- Allows skipping of initial consistency checks (context: pilot jobs)
- Possibility to run old code for longer time

Operations perspective:

- Strict encapsulation of user jobs in their sand box
- Decoupling of applications from underlying hardware
  - Recent hardware only works with recent OS versions
  - Complex applications are difficult to port to new OS versions
- Easy transition from one OS to another
- Dynamic change of worker node types dependent on requirements
  - Better use of resources
- Easier handling of intrusive updates and operations
- Roll-out of important security updates in a rolling way

NOTE: service consolidation is a different project with different use cases!

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Requirements

Operational requirements:

- Backward compatible approach
  - No change for traditional users
- Start small and grow
- Seamless integration into the existing system
- No (or minimal) additional work load on the service managers
- Less work load after deployment
- Possibility to mix virtual and real resources

Infrastructure:

- No changes to the infrastructure required
- Full integration into the existing systems and databases

Vision:

Be backward compatible while opening the door to explore new computing models

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Basic concepts, proof-of-concept

First step: adding virtual batch worker nodes to lxbatch at CERN

Hypervisors:

- Unique set of machines, organized in a new cluster “lxcloud”
- Fully Quattor managed and Lemon monitored
- Very restricted access (no user access!)
- Minimal software setup
- XEN kernel (later move to KVM)

Golden Nodes:

- Fully Quattor managed VM with worker node setup
- Running AFS and LSF services
- Don't run jobs but get updated
- Serve as templates for the creation of images
- Snapshotted once per day

Worker nodes:

- Virtual machines with public IPs
- Derived from Golden nodes
- Not quattor managed
  - Dynamically join the existing LSF cluster and accept jobs
- Limited life time (draining after 24h and shutdown after that)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Initial phase: supported images

Supported images are 1:1 snapshots of current worker nodes

Step 1: fully backward compatible approach
- SLC4/64bit with 32bit compatibility
- SLC5/64bit with 32bit compatibility
- 2GB RAM + some work space
- 1CPU only

Step 2: support also specialized images
- More CPUs
- More Memory
- e.g. 2 CPU/4GB, 4 CPUs/8GB etc
- On user demand

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Dealing with user requirements

We don't allow user provided images.

So how can the user tell what he needs ?

- When a new (virtual) worker node reports it's capabilities to LSF
- The user can specify resource requirements at submission time
- LSF will pick only candidate nodes which match the user requirements

Special case: OS system selection

If the user does not specify the OS system version, this information is inherited from the submission host OS system type(*)

(*) more precise: the LSF type, what LSF thinks the type is

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Image selection and VM placing

Image selection:

- Driven by user demand
  - Requirements of pending jobs in the queues
  - Job priority
- Requires coupling between VM management software and batch system
- Can be tricky to implement

VM placing:

- Simple load driven resource allocation manager is fine
- We rely on existing solutions
- Both commercial and free solutions are available
- Looked at **Platform VMO (ISF)** and **OpenNebula**
- Both have their pros and cons

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

More on image selection ...

Need a **meta-scheduler** which connects the batch system with the image deployment mechanism.

Requirements:

- Optimization goal must be efficient use of resources
- Respecting fair share and queue priorities
- Light-weight
- Minimizing number of queries to the batch system

Status:

- Existing prototype in VMO/ISF, developed by Platform with/for CERN
- Based on batch system queries, not really integrated into LSF
- Simple prototype for OpenNebula done by Sebastien Goasguen

Remarks/Ideas for improvements/implementation:

- More sophisticated approach is needed, current system may not scale well
- Could be **slow-control** and run entirely offline
- Respect boundaries: offer a minimum and maximum number of specific VMs
- Use historical data instead of online requests ?
- Get it right on average within boundaries, no short term big changes

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

VMO versus OpenNebula

VMO (Virtual Machine Orchestrator):

- Uses EGO as resource manager (same as LSF version 7.X)
- Existing prototype for coupling with LSF
- Demonstrated to work at the multi-threading work shop at CERN
- Commercial product
- Some issues due to networking boundary conditions at CERN

OpenNebula:

- Intuitive user interface
- Naturally no integration with any batch system
- Used in the prototype at CERN as well
- Free software, public sources so easy to debug
- Easier to integrate into the existing infrastructure at CERN

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

GRID integration

Prototype setup:

- Up to 18 nodes acting as hypervisors
- 1 LSF master node
- 1 CREAM CE
- 1 management host (for OpenNebula)

CREAM CE modifications:

- Request to execute external plugin to pass user requirements to the batch system always if if it exists (default: it's only executed if there are actual requirements)

- **glite-blah-local-submit-attributes-lsf** new version
  - Always sets the submission host type (so that it can be consumed by an external coupling but also as a work around for a misbehaving LSF scheduler when dynamic hosts are present)
  - Add support to change the OS type

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

GRID integration

Direct job submissions through the CREAM CE with :

CERequirements = "(Member(\"SLC4_64\", other.GlueHostOperatingSystemName))";

gets translated into user job requirements (LSF syntax)

-R”select[type==SLC4_64]”

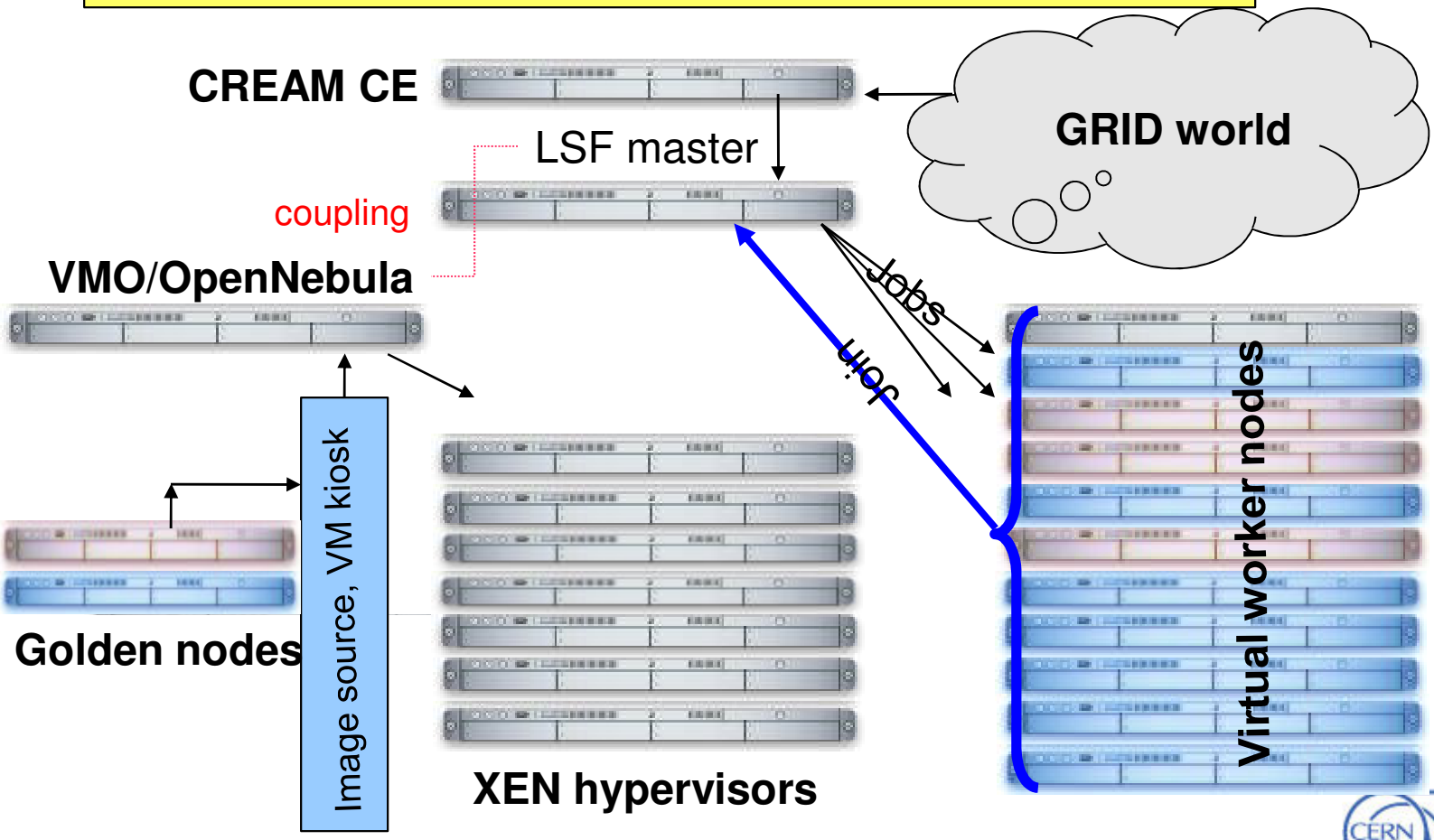
- For each supported OS type the CE needs to publish a subcluster (!)
- All CE nodes can be made equal
  (right now: one set of CEs per OS)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Current Status: proof of concept

Phase 1 **proof of concept** is done, with some additions:

- Successful launch of CERNVM images as well
- Testing of OpenNebula Cloud interface just started



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Towards prototype and production …

From **proof of concept** to **prototype** and **production** issues:

- Hardware selection:
  - A 16GB RAM hypervisor cannot run 8 2GB VMs (!)
  - KVM may help
- Current images are huge: the temporary space is part of the VMs
- Still a weak point: the virtualization kiosk (=image repository)
  - HTTP based, with SQUID buffers ?
  - Idea: move images asynchronously to the hypervisors if changed
  - Scalability yet unclear
- IP adresses:
  - We definitely need outbound connectivity
  - CERN can use public IP addresses, NAT is not needed.
- May need a fairly clever placing mechanism
- Image selection mechanism can become tricky
  - Respect shares and priorities

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

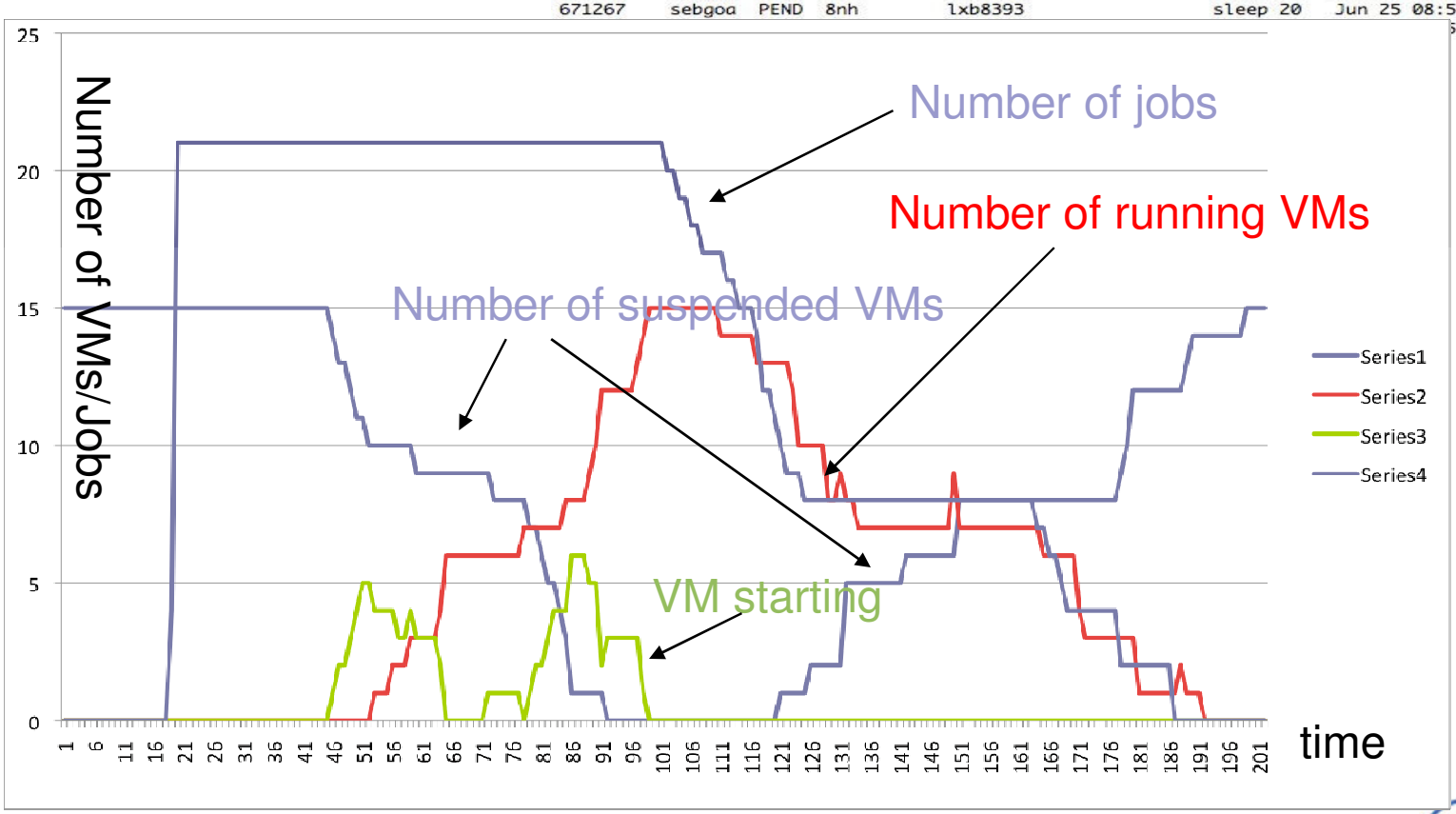
Basic Functionality Tests

- VMO LSF coupling tests:

Default

Every 2.0s: bjobs — Thu Jun 25 08:59:40 2009

JOBID	USER	STAT	QUEUE	FROM_HOST	EXEC_HOST	JOB_NAME	SUBMIT_TIME
671251	sebgoa	RUN	8nh	lxb8393	eeng101	sleep 20	Jun 25 08:54
671252	sebgoa	RUN	8nh	lxb8393	eeng102	sleep 20	Jun 25 08:54
671259	sebgoa	RUN	8nh	lxb8393	eeng081	sleep 20	Jun 25 08:54
671261	sebgoa	RUN	8nh	lxb8393	eeng031	sleep 20	Jun 25 08:54
671262	sebgoa	RUN	8nh	lxb8393	eeng071	sleep 20	Jun 25 08:54
671263	sebgoa	RUN	8nh	lxb8393	eeng022	sleep 20	Jun 25 08:54
671264	sebgoa	RUN	8nh	lxb8393	eeng082	sleep 20	Jun 25 08:54
671265	sebgoa	PEND	8nh	lxb8393		sleep 20	Jun 25 08:54
671266	sebgoa	PEND	8nh	lxb8393		sleep 20	Jun 25 08:54
671267	sebgoa	PEND	8nh	lxb8393		sleep 20	Jun 25 08:54



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

GRID integration tests

CREAM job submission: basic tests

- Direct submission through CREAM CE, hello world
- Attribute passing through the CREAM CE

CREAM CE: real payload

- Asked ALICE to submit jobs
- Need outbound connectivity
- Success with pilots and real payload
- Success rate to be quantified and compared to physical boxes

Beyond phase 1 … VM Visions

- **Reusing images outside CERN (phase 2)**
  - Phase 1 images are highly CERN customized
  - Requires trusted portable base image ...
  - … plus a more complex contextualization phase at the site
  - A possibility for the site to add site specific software without compromising the image integrity
- **Experiment specific images (phase 3)**
  - As before but this time the software environment gets customized for a specific experiment
  - Removes the need for pilot jobs to check the environment which is already correct by construction
  - Requires a good control of the experiment software stack to avoid compromising the image integrity
- User defined images / CERN-VM
- **Use of resources in a cloud like operation mode (phase 4)**
- **Images join different batch farms at startup (phase 5)**
  - Controlled by experiment
  - Spread across sites
  - Possibly replace current pilot job frame work

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Summary, conclusion and plans

- **CERN has developed a concept for a migration of batch services to virtual machines**
  - Start small then grow
  - Backward compatible
  - Opens the doors for new computing models
- **A proof of concept has been demonstrated with success**
- **Not production ready (yet)**
- **A proposal for a prototype installation is in progress**

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it