WLCG et la Virtualisation

Michel Jouvin
LAL, Orsay
jouvin@lal.in2p3.fr
http://grif.fr

LCG France, Marseille
24 Juin2009

Agenda

- Le contexte
- La vue des expériences et des sites
- Les images virtuelles
- Le Virtualisation WG d'HEPiX
- Quelques exemples
- Autres projets complémentaires : CERNVM, StratusLab
- Conclusions

Credits

- Cette présentation s'appuie abondamment sur des présentations données par T. Cass et P. Mato
  - Tony Cass (CERN IT) : chairman du HEPiX Virtualization working group
  - P. Mato (CERN PH) : leader du projet CERNVM
- Elle fait écho à de nombreuses discussions au sein de HEPiX et de son virtualisation WG

Le Contexte

- Forte focalisation sur le contexte d'exécution des applications utilisateurs
- Le middleware de grille impose des contraintes fortes sur la version d'OS des WNs
  - Parfois contradictoires avec les besoins des SW des expériences (ex: version Python ou glibc)
- L'upgrade du MW impose souvent une mise à jour de l'OS
  - Ex: Migration gLite 3.2
  - SW des expériences pas nécessairement prêt
  - Impose que toutes les expériences soient prêtes en même temps
- SL n'est pas forcément le Linux de référence de toutes les communautés
  - Eventuellement des requirements conflictuels entre VOs
- Les besoins des VOs doivent être implémentés par les sites
  - Potentielle lourdeur, manque de réactivité

Virtualisation hors WLCG

- Beaucoup d'autres raisons de s'intéresser à la virtualisation
  - Consolidation : plusieurs services sur le même hardware
  - Haute disponibilité : facilité de réallouer une machine virtuelle à un nouveau hardware
  - Machine de référence en backup
  - Infrastructure de tests déployable à la demande ou d'infrastructure type T3 avec peu de manpower
- Cloud : allocation à la demande en fonction des besoins des ressources HW
  - Elasticité
  - Contrôle de l'environnement par l'utilisateur
- Ces autres besoins ne sont pas couvert par cette présentation centrée sur les WNs grille (gLite)
  - Pas moins légitimes ou bien fondés

Vue des Expériences...

- Pas vraiment de vue unifiée des 4 expériences mais un projet R&D au CERN : CERNVM
  - Leader : Pere Mato
- 2 workshops dédiés à la virtualisation (et au support multi-core) dans WLCG
  - 2ème : lundi et mardi dernier
- Plutôt intéressées par la virtualisation, sauf CMS
  - Atlas le plus moteur
- Performances
  - CPU : pas de pénalité (sensible)
  - I/O : le problème mais conviction que cela diminuera sur les nouvelles générations de machine/CPU

… Vue des Expériences

- Souhait de pouvoir disposer d'une machine HW dédiée à la VM
  - Indépendant de la virtualisation : souhaite disposer de tous les cores d'une machine
- Souhaiteraient pouvoir utiliser une VM plusieurs jours après son démarrage
  - Efficacité vs. Turnaround : // avec les multi-user pilot jobs
  - Accounting CPU-based difficilement acceptable
- Distribution du SW : intégré à l'image ou bien téléchargé à la demande (CVMFS)
  - CVMFS : file system à base de Squid + FUSE développé dans le cadre de CERNVM
- Besoins en cluster interactif (PROOF) à la demande
  - Plutot une problématique cloud

Vue des Sites

- Pas de position formelle mais en général plutôt réticents
  - Perte de contrôle sur l'utilisation des ressources
  - Problèmes potentiels de sécurité
  - Utilisation sous optimale des ressources HW
  - Complexité d'administration
- Souhaitent généralement contrôler les images
  - Expériences inintéressées si chaque site produit son image
    - Restera le besoin de valider chaque site

Les Images

- Le cœur du problème de la virtualisation…
- Qui les produit ?
  - N'ont d'intérêt que si elles sont adaptées/spécifiques aux besoin d'une expérience/VO
    - Potentiellement même plusieurs images par VO suivant les applis
  - Doivent être produites centralement et utilisables par tous les sites
    - Indispensable pour la généralisation sur la grille
- Qui les contrôle ?
  - Rôle indispensable pour permettre leur adoption par des sites qui ne les ont pas produites
  - Doit permettre la révocation
- Quelles règles ?
  - Pour les produire
  - Pour les customiser

HEPiX WG

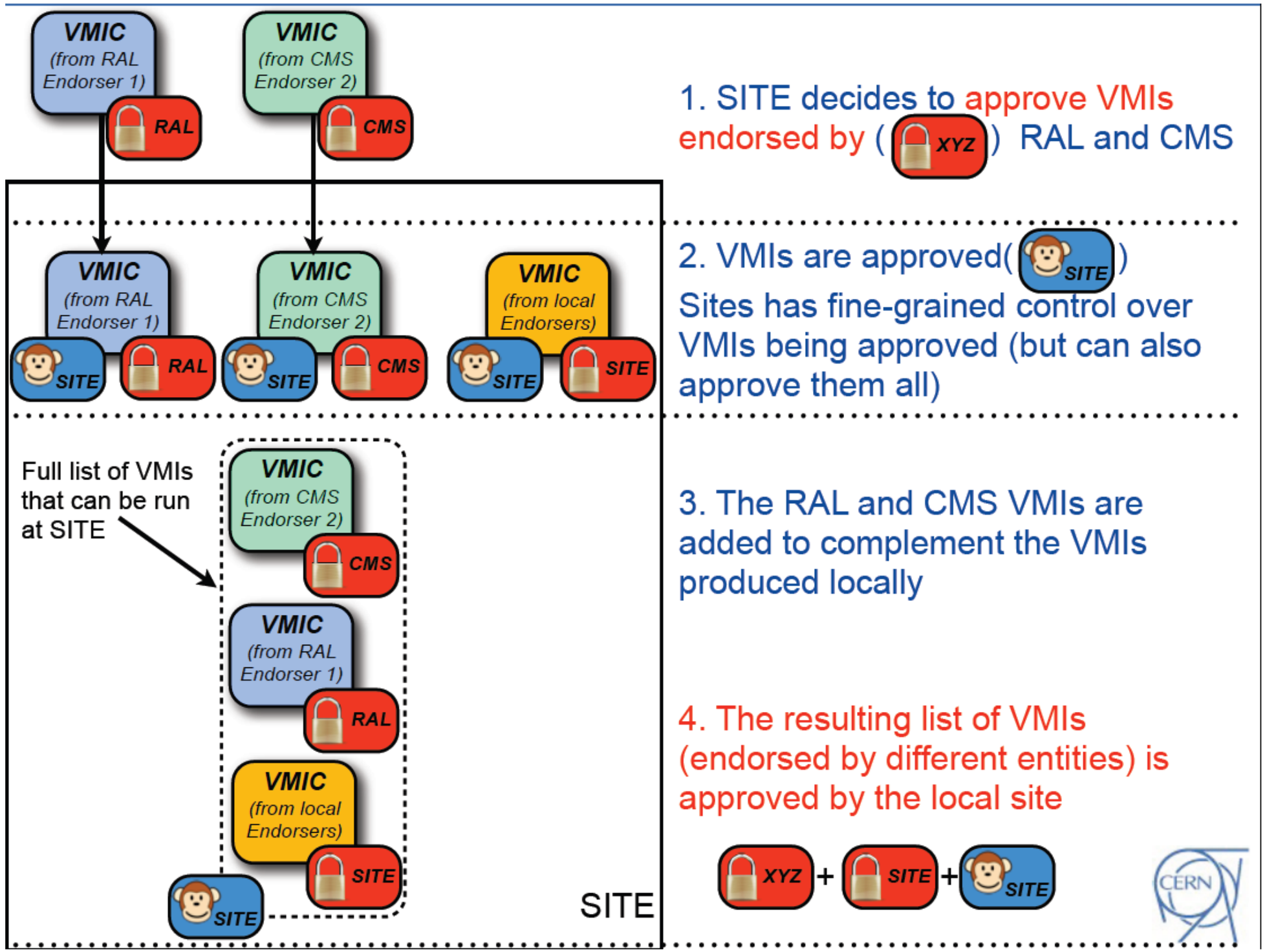
- But : produire les conditions d'une adoption large de la virtualisation des WNs par les sites
  - Créé lors du HEPiX Berkeley (Oct. 2009)
  - Chairman : Tony Cass (CERN IT)
- Essentiellement des représentants de site
  - Pas de lien direct avec CERNVM
    - CERNVM n'est pas nécessairement le modèle d'image proposé pour la phase 1
  - Coordination avec les expériences prévues ultérieurement
- Approche pragmatique : construire sur les succès
  - Offrir un environnement dans la VM aussi proche que possible de celui d'un WN
    - Conserver le paradigme WN connecté à un batch system
    - Ne pas interdire des expérimentations plus cloud-oriented
  - Se servir des expériences en cours pour créer de la confiance

WG : Axes de Travail...

- Politique de génération d'image « trustable »
  - Définie avec le JSPG
  - Pas d'accès root aux VMs pour les utilisateurs
  - Rôle de l'endorser : celui qui certifie la validité d'une image, responsable de la révocation des images en cas de pb
  - Focus sur le process de création de l'image plutôt que son contenu
  - http://www.jspg.org/wiki/Policy_Trusted_Virtual_Machines
- Transmission et instanciation
  - Transmission : uniquement inter-site, protocole http
  - Instanciation : vérification de la validité d'une image au moyen d'un VM Image Catalog (VMIC)
    - Utilisé par l'endorser pour « publier » ses images et par le site pour définir les images qu'il accepte
    - Utilisé lors de l'instanciation d'une image pour vérifier sa validité
    - Développement prévu par le groupe, concertation avec d'autres projets

VMIC Workflow

Credits to R. Wartel and U. Schwickerath



… WG : Axes de Travail

- Contextualisation
  - Action de customisation d'une image par un site
  - Pas de modification de l'image
    - Sinon elle devient non vérifiable
  - Des « hooks » dans le startup permettant l'adaptation à la configuration locale : réseau, batch system…
    - Aucun client de batch dans les images
    - Interdiction de changer les modules kernel, Python, Perl, glibc…
- Support d'hyperviseurs multiples
  - Recommendation pour construire des images pouvant être utilisées par plusieurs hyperviseurs
    - Focus principal : Xen, KVM
  - Principale approche : image fully-virtualized avec utilisation des drivers virtio/libvirt
    - Vérification en cours de l'impact sur les performances
  - Modules kernel spécifiques pour la para-virtualisation
    - Plusieurs entrées Grub sélectionnables lors du boot (pygrub)

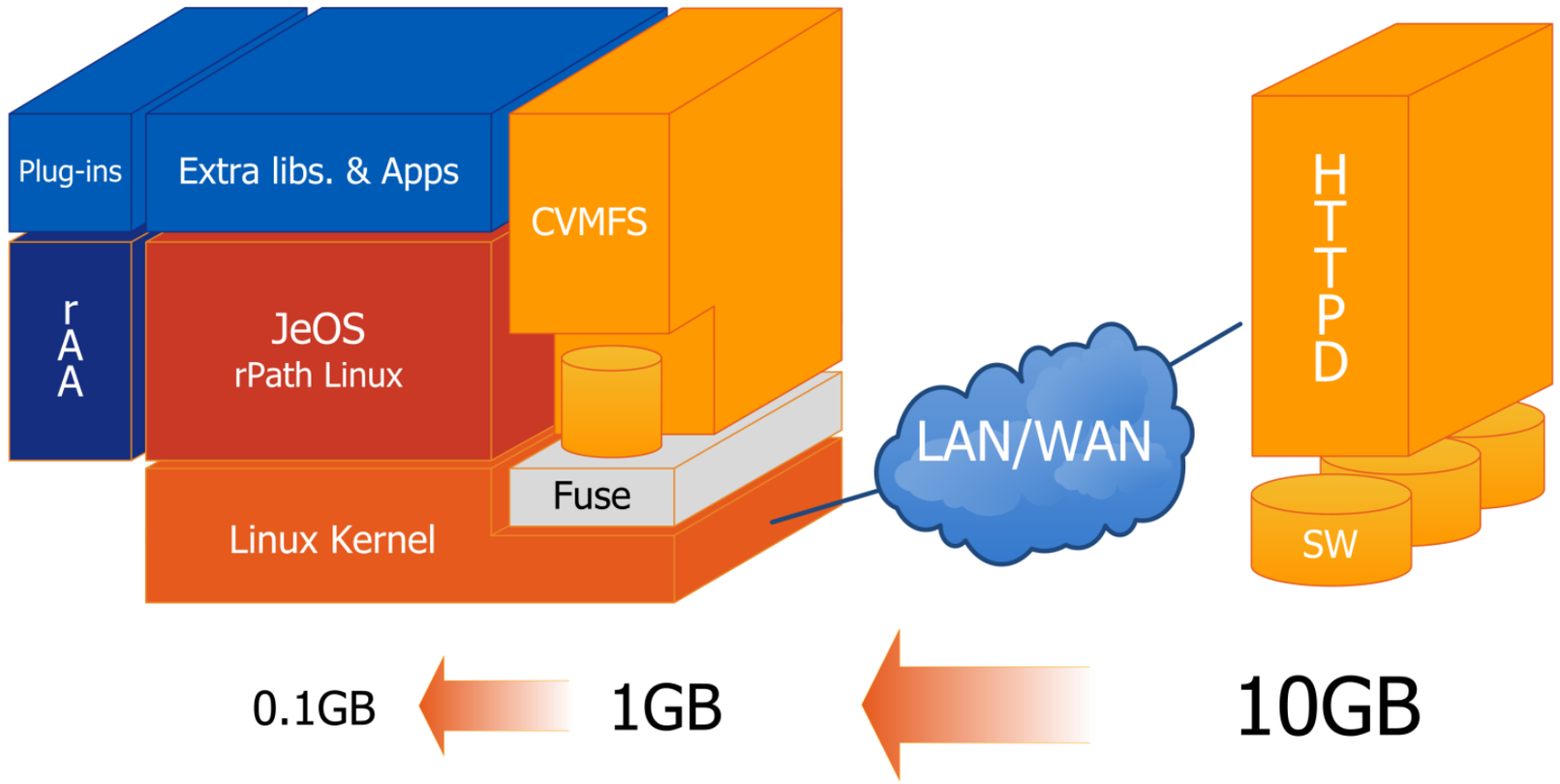
Des Exemples

- CNAF : virtualisation en cours de l'ensemble des WNs
  - Démarré il y a 1 an, 2/3 des WNs virtualisés
  - Basé sur KVM
  - Hyperviseur est un client LSF et crée 1 VM par job accepté
  - Images produites par CNAF
- CERN
  - Basé sur lxcloud : ensemble de VMs gérées par LSF VM Orchestrator ou openNebula
  - VM produite par le CERN et sélectionnable par job
    - Chaque VM est configurée comme un client LSF
    - 1 VM peut traiter plusieurs jobs
- UVic
  - VMs (Nimbus?) connectées à Condor
  - Reprend/utilise le principe des Condor glide-ins
- Et d'autres...
  - Voir dernier HEPiX (Lisbonne)

CERNVM

- 1 image + 1 service
  - Implémenté au CERN principalement
- Image : OS minimum + chargement à la demande du soft des expériences et de leur dépendance
  - Basé sur CVMFS : file system « web » basé sur le concept de Content Distribution Network
    - FUSE + squid + serveurs http
  - Devrait être générique pour toutes les expériences
  - Possibilité de créer des snapshots « customisés »
- Très attirant mais…
  - Contenu dynamique des images rend difficile leur « endorsement »
  - Très adapté à l'analyse end-user mais pas forcément efficace pour la production
  - Serveurs CVMFS : quel politique de déploiement/maintenance du contenu

CERNVM Architecture



StratusLab

- Projet FP7 autour de l'intégration grid/cloud
  - Démarré le 1/6, project leader : Cal
  - Utile OpenNebula comme gestionnaire de cloud
    - Cloud de type *Infrastructure as a Service* (IaaS)
  - LAL impliqué dans la mise en place du testbed (WP5)
- Pas restreint à la virtualisation des WNs
  - Testbed : 1 site « complet » implémenté dans un cloud
    - Un prototype déjà effectué il y a 2 ans dans Amazon EC2
- Développer une API utilisateur pour l'accès au cloud
  - Fonctionnellement équivalente à l'API Amazon mais open source
  - Permettre l'accès direct aux ressources du cloud par une application, sans passer par les APIs grille
- Valider l'impact et l'intégration avec les outils de déploiements
  - LAL particulièrement impliqué pour Quattor

Conclusions

- La virtualisation est une technologie mature qui peut aider à séparer environnement utilisateur et contraintes des sites sur les WNs
  - Sites : machines déployées en fonction des contraintes du MW et des autres contraintes du site
  - Expériences produisent les images adaptées à leurs besoins applicatifs
- Production « centralisée » d'images « trustable » est le principal challenge pour une large adoption
  - Problème politique : définition de règles
  - Outil d'implémentation de la chaine de trust
  - Sujet principal traité par le HEPiX Virtualisation WG
- D'autres approches possibles et complémentaires
  - CERNVM : image « dynamique »
  - StratusLab : gestion cloud-like (*elastique*) des ressources

Liens Utiles

- Agenda HEPiX WG (documents attachés)
  - http://indico.cern.ch/categoryDisplay.py?categId=2738
  - Report at HEPiX meetings. Last report (HEPiX Lisbon): http://indico.cern.ch/getFile.py/access?contribId=59&sessionId=29&resId=1&materialId=slides&confId=73181
- CERNVM
  - http://cernvm.cern.ch/cernvm/
- StratusLab
  - http://stratuslab.eu
- 2nd WLCG Workshop on Virtualization
  - http://indico.cern.ch/conferenceDisplay.py?confId=89681