Virtualisation des nœuds de calcul

Mattieu Puel / groupe système

Plan

Groupe HEPIX Virtualization

Quoi ? Qui ?
Motivations
Réalisations

Implémentation au CC

Et demain ?

HEPIX Virtualization working group

But du groupe

Virtualisation des nœuds de calcul de la grille

Orienté WLCG

Motivations :

Souplesse :

- Faciliter le déploiement des softs utilisateurs
- Choix de softs personnalisé

Support d'environnements multiples

HEPIX Virtualization working group

Qui ?

CERN
- Tony Cass
- Sebastien Goasguen
- Ulrich Schwickerath
- Romain Wartel

LAL
- Michel Jouvin

DESY
- Owen Synge
- Thomas Finnern

UVIC
- Ian Gable

RAL
- Ian Collier
- John Gordon
- David Kelsey
- Martin Bly

INFN
- Andrea Chierici

FNAL
- Keith Chadwick

HEPIX Virtualization working group

Périmètre:

- Génération distante d'images
- Échange des images entre expériences / sites
- Inter-opérabilité des images entre sites
- Permettre une configuration spécifique des sites

Génération distante d' images

- Document de référence sur la politique de génération des VMs
- Support hyperviseurs (Xen, KVM, ...)
- Méta-données : contenu et format
- Signatures (SHA512 + X509)

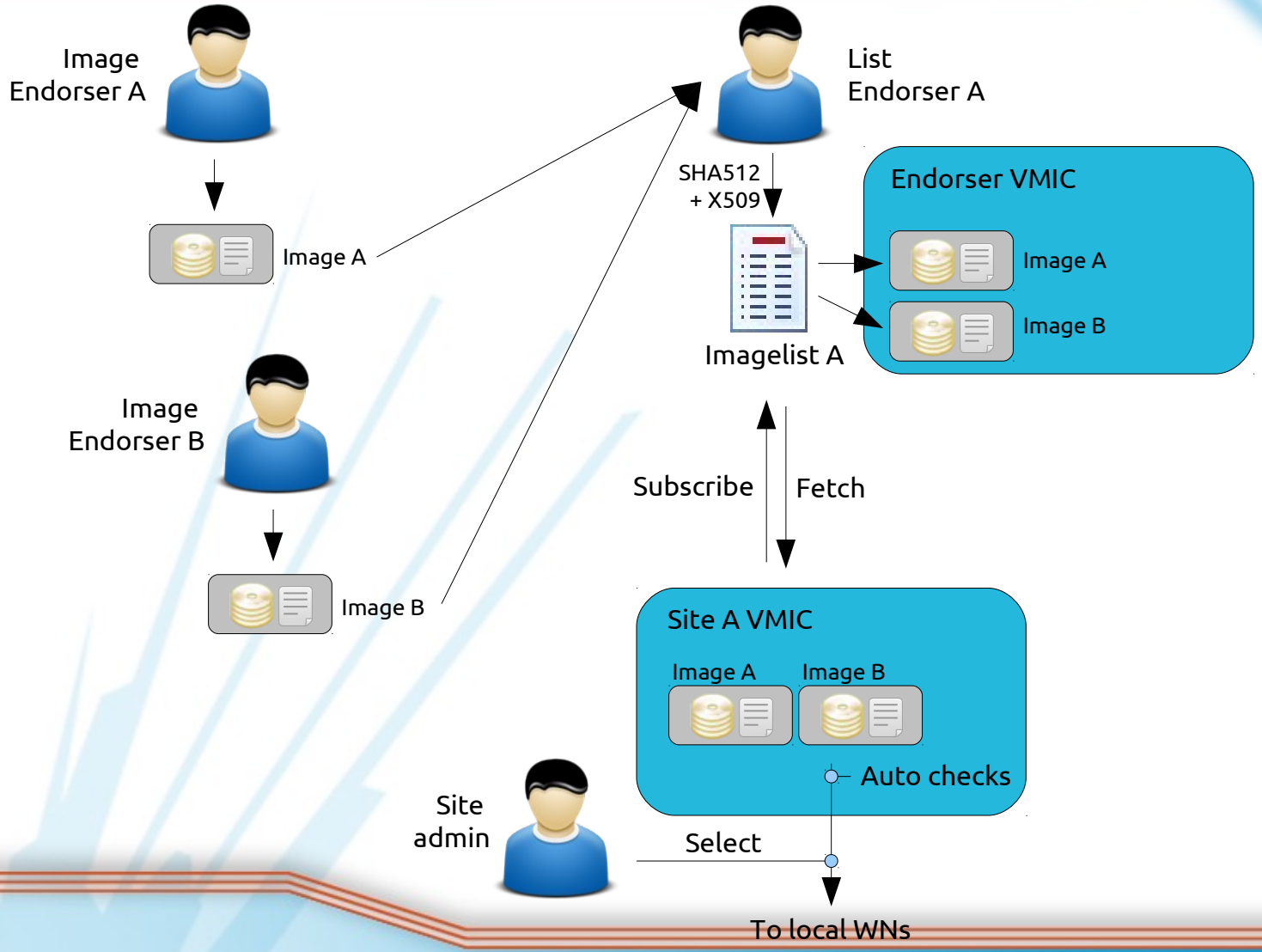
Échange des images entre expériences / sites

Définition des rôles :

Expérience	**Image endorser** : génere les images **List endorser** : endosse des listes d'images pour l'expérience
Site	**Site admin** : souscrit aux listes d'images

Catalogue endorser / site

HEPIX Virtualization working group



Permettre une configuration spécifique des sites

Contextualisation:

- Moyen technique de personnalisation de l'instance
- Montage d' un FS iso (cdrom) et exécution de scripts

Exemple d'utilisation :

- Récupération de logs
- Aménagement d'accès pour les administrateurs du site

HEPIX Virtualization working group

État de l'art

Formalisation des besoins / moyens (depuis dec 2009)

Implémentations (depuis juin 2010)

- Signature/vérification d'images, de listes (O. Synge)
- Catalogues d'images VMIC (O. Synge/U. Schwikerath)

Début de tests d'échanges de d'images et d'instanciations

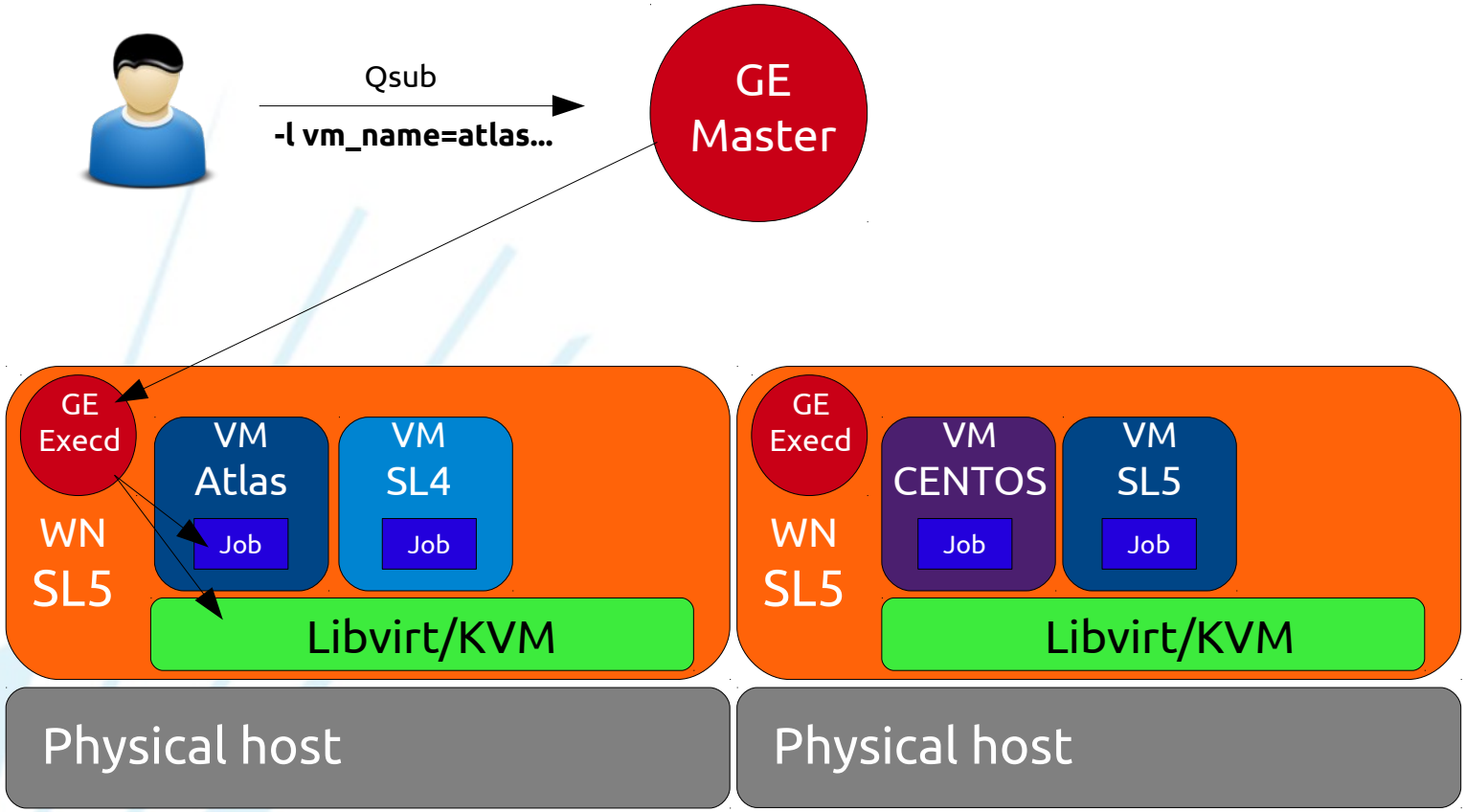
Evolution du groupe... vers le CLOUD

- StratusLab Intégration des outils de StratusLab (Marketplace)

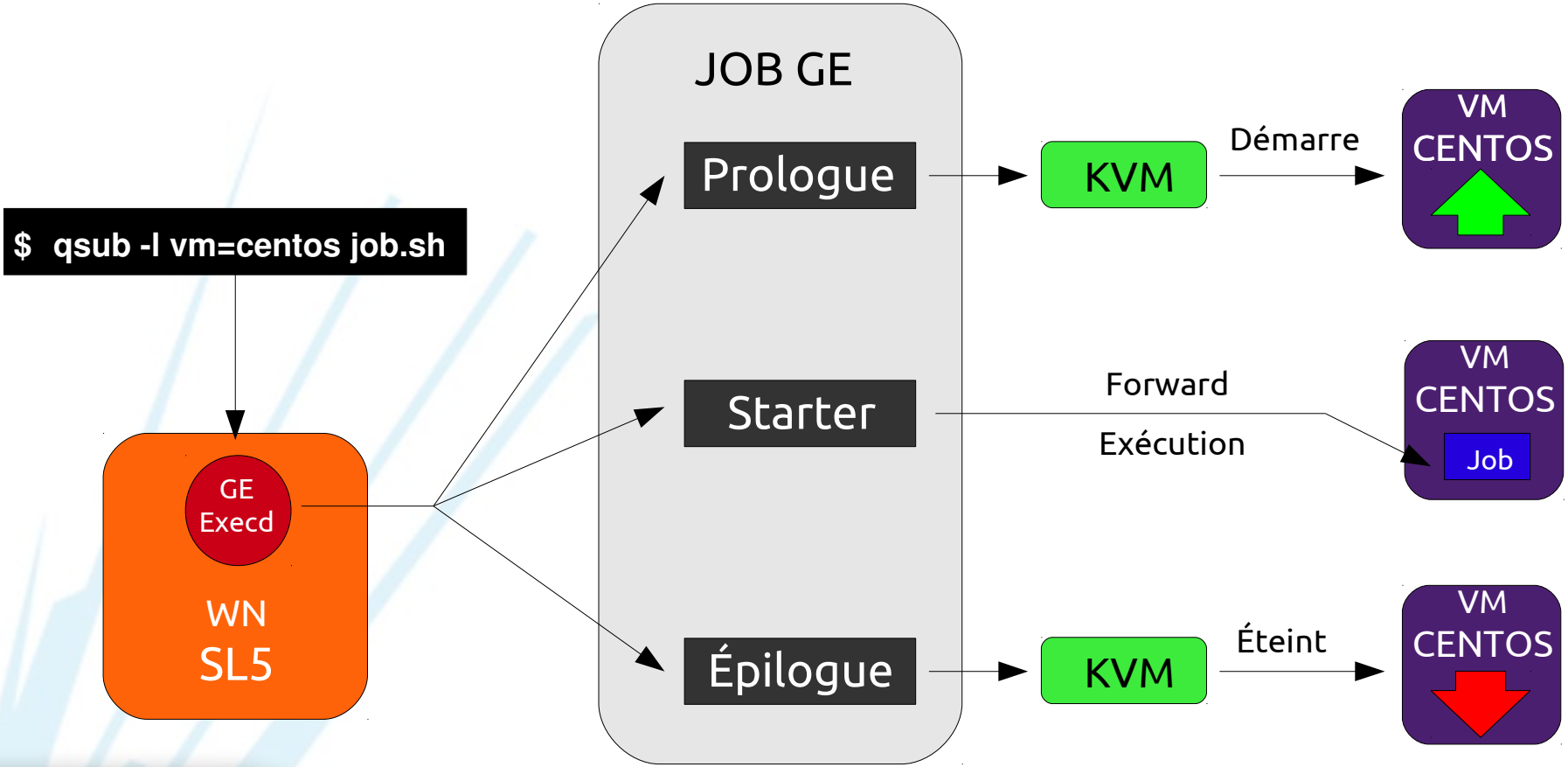
Implémentation au CC
(prototype)

- Hyperviseur
- Réseau
- Stockage
- Todos

Cluster GE virtualisé (prototype)



Déroulement du job



Implémentation au CC

Technos de virtualisation, hyperviseur

- Xen
- KVM

Implémentation au CC

Réseau

Paramètres réseau Qemu :

- Virtuel vs **paravirtuel**
- Réseau publique vs **privé**

Overhead : <3% paravirtuel, ~10% virtuel

Implémentation au CC

Stockage (système, scratch)

Leviers principaux (non exhaustif) :

- **Block** vs file
- type d'image: thin vs **thick provisionning**
- Périphérique **virtuel scsi** vs virtio

Overhead : 10% dans les meilleurs conditions

Implémentation au CC

Stockage (système, scratch)

Snapshots LVM :

+ instanciation/recyclage direct
+ mutualisation du stockage

Pre-staging :

+ limitation des transferts réseau
- gestion de cache

Implémentation au CC

Todos

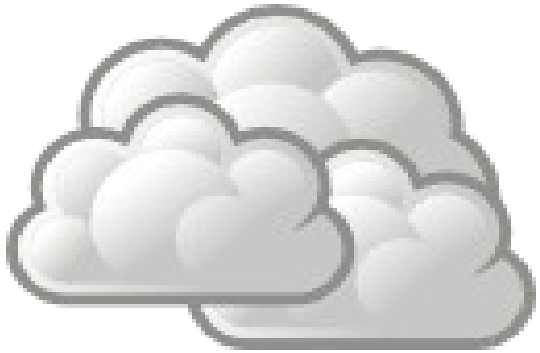
- Intégration VMIC et branchement au cache local
- Déploiement du cache local vers les Wns (bittorent, scp-tsunami...)
- Proof of concept OK, et les tests ?

  production LHC (vm CC + soft expérience)

  VM générée par les utilisateurs

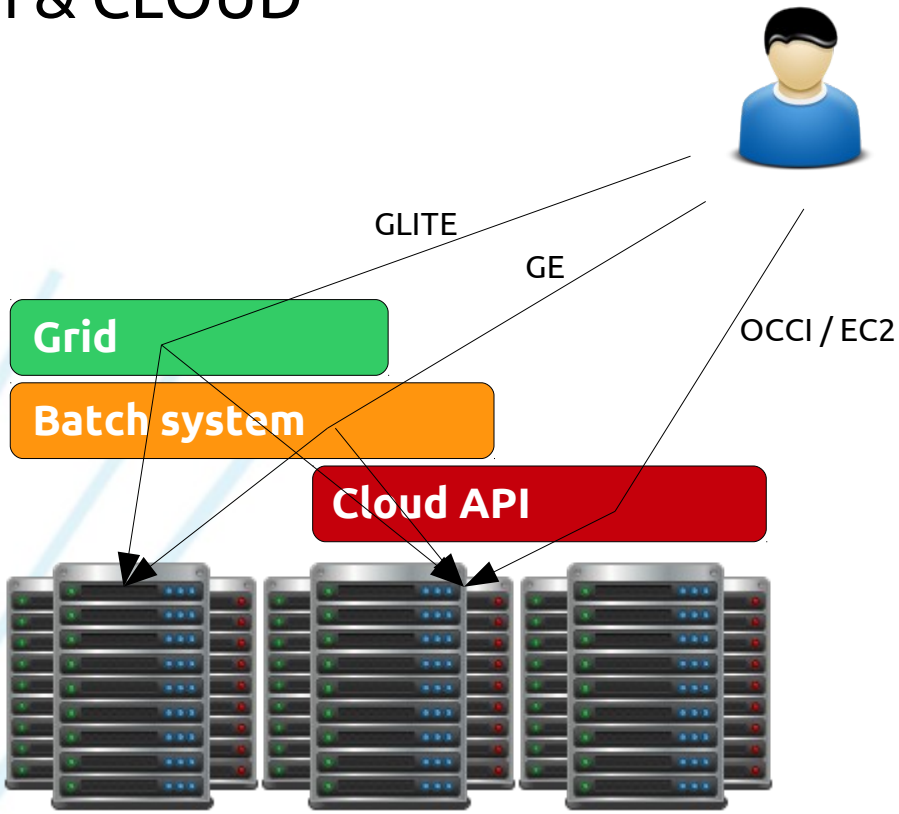
Prospectons : et demain ?

Vers...



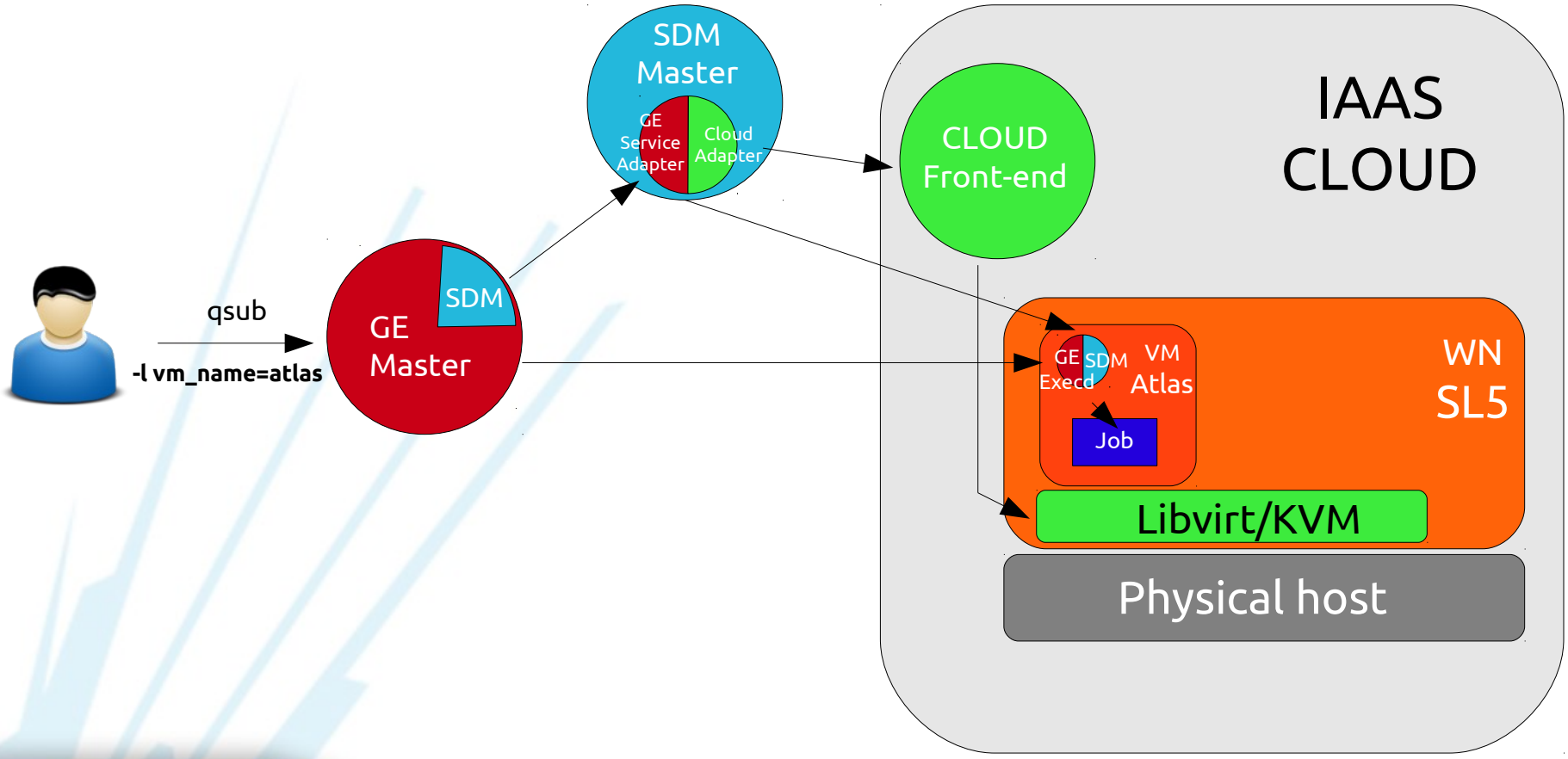
Et demain ?

Grid, batch & CLOUD



Et demain ?

GE over CLOUD ?



Et demain ?

CLOUD : Intérêts croisés

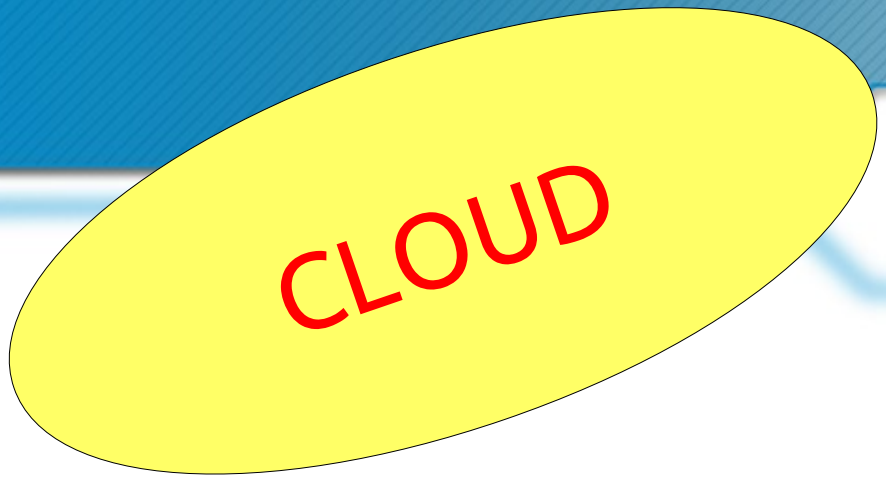
EGI (User Virtualisation Workshop may 2011)

WLCG

Les prochains défis du CC :

- Capri (Appel à projet équipements d'excellence)
- Organisation des journées CLOUD au CC les 20/21 oct
- Mise en place d'un groupe de travail au CC

La slide BONUS



Maturité des technologies

Prinicipalement pour les grandes expériences

Comprendre le modèle :

contradiction avec la volonté de ne pas être un fournisseur de puissance
sécurité (fonctionnement proche de l'hébergement)

Des nouveaux usages (self-service de... service)

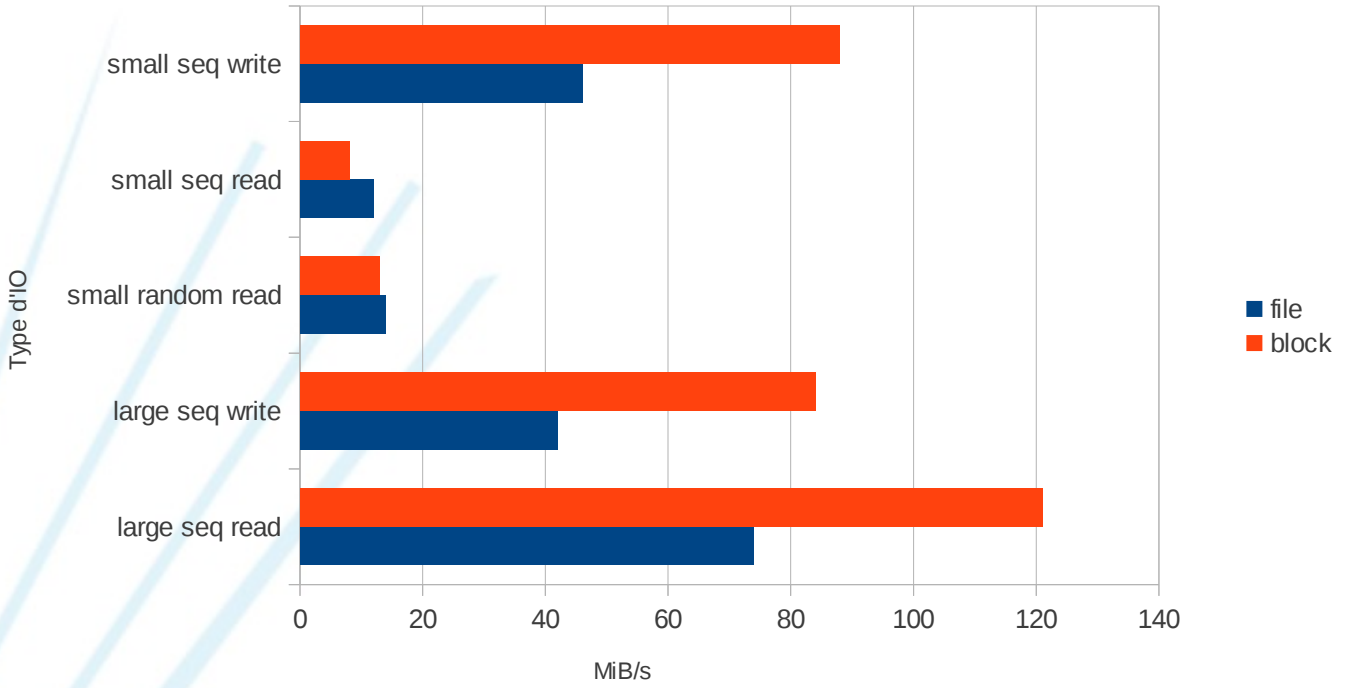
Questions ?

Annexe A

Benchs de technologies disque

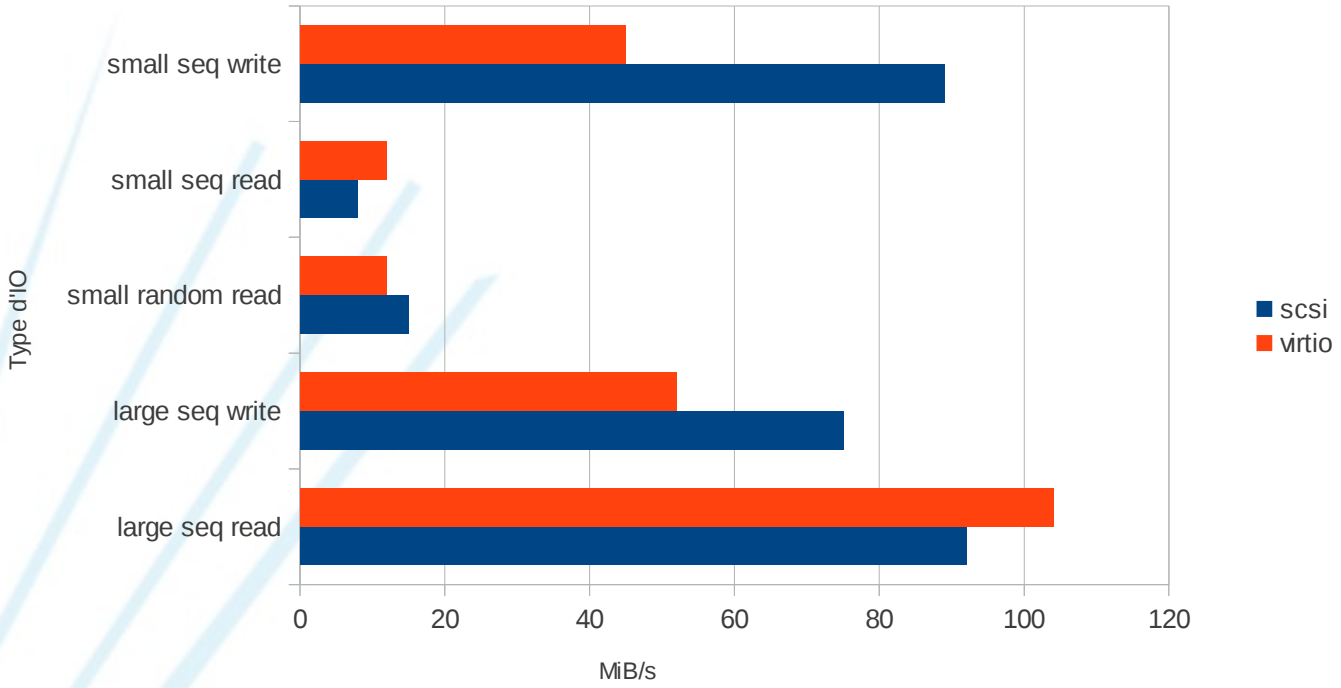
Implémentation au CC
Annexe A

File versus block image type (Xio@LT)



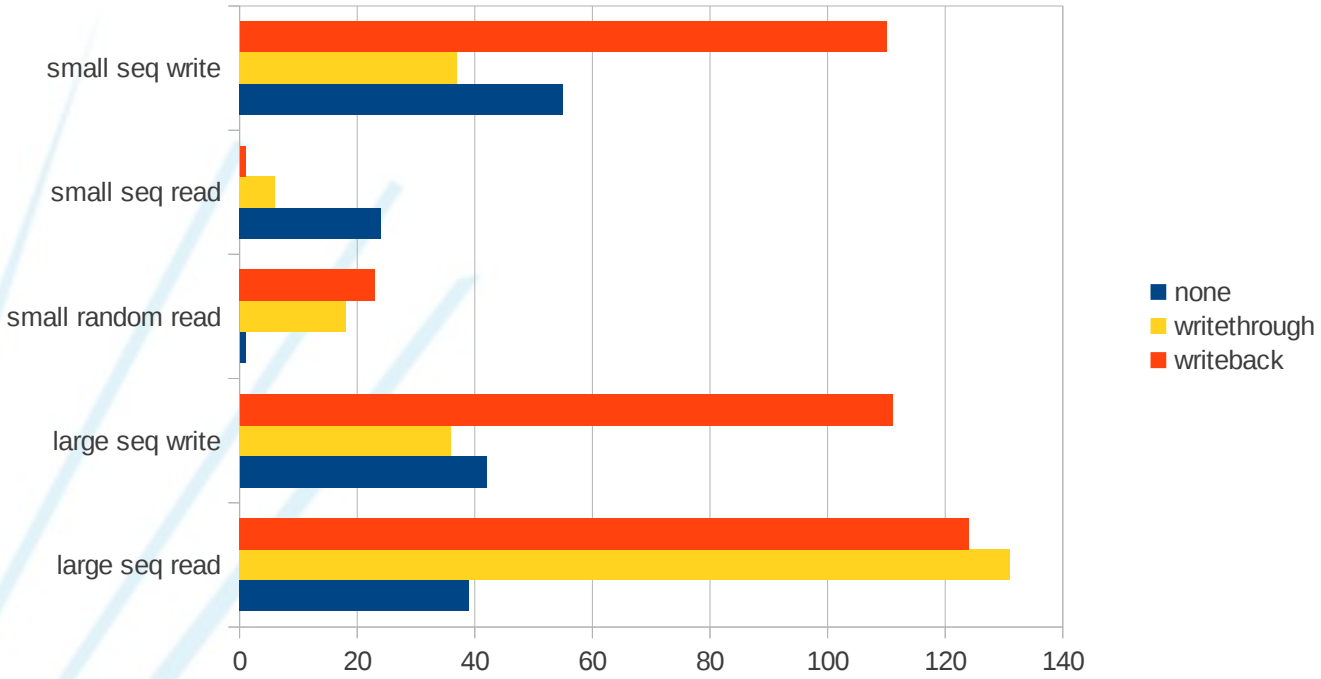
Implémentation au CC
Annexe A

Full virt vs paravirt (Xio@LT)



Implémentation au CC
Annexe A

Cache (Xio@LT)



Annexe B
Exemples d'implémentation de sites HEP

CERN

Université Victoria

CNAF

Exemples d'implémentation de sites HEP
Annexe B

CERN

CLOUD soft : OpenNebula et Platform ISF, vers OpenStack ?

LSF batch system

Snapshots LVM, pré-staging

Réseau publique

Déploiement vers les Wns : bittorrent

96 nœuds (12 hyperviseurs) depuis dec 2010

Exemples d'implémentation de sites
Annexe B

UVIC

Pour Babar

CLOUD soft : Nimbus

Condor batch system, client pré-installé dans la VM

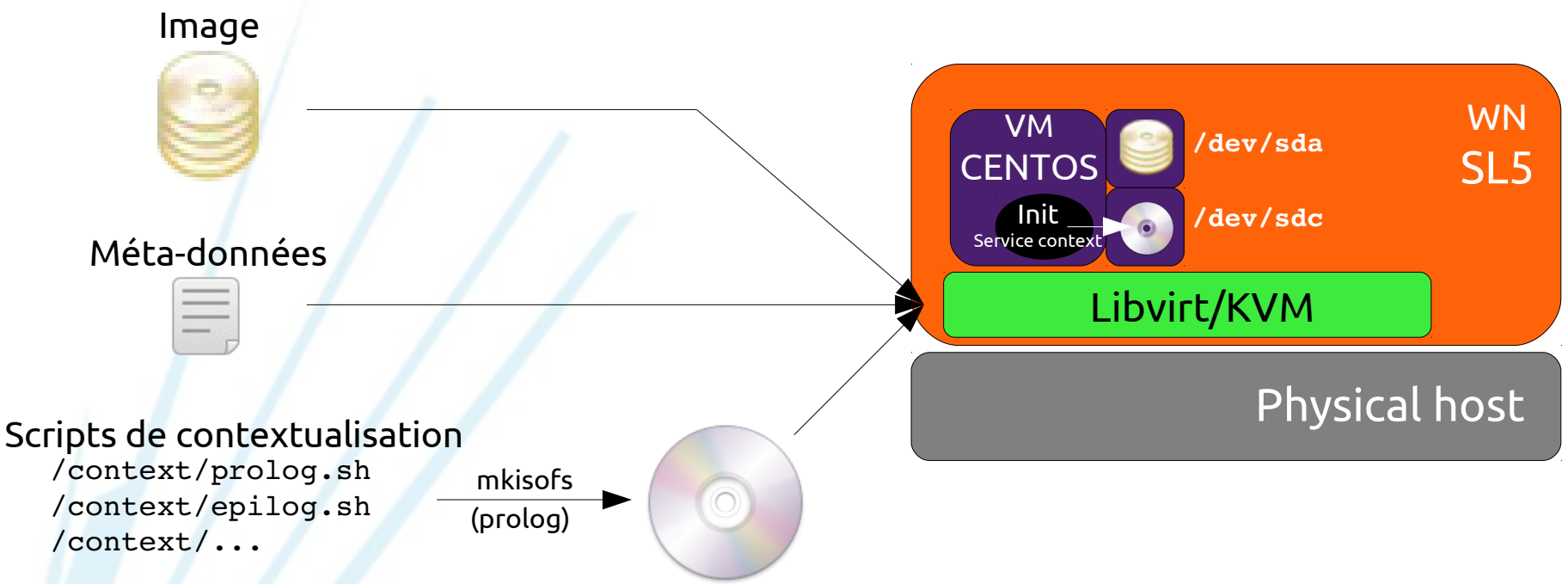
Implémentation du composant interface Batch / plusieurs CLOUDs

Exemples d'implémentation de sites
Annexe B

CNAF

- Branchement de LSF sur QEMU/KVM directement
- Client de batch sur l'hyperviseur
- Déploiement des images via GPFS et/ou HTTP
- Pré-staging
- Stockage local de l'image dans un fichier (pas de snaps LVM)

Contextualisation



Implémentation au CC
Annexe D : La démo sans effet démo

```
zsh)puel.cctest15 []% qsub -q virt -l vmem=2g -ac vm_name=ccin2p3 <<EOF
cat /proc/cpuinfo
free
EOF
Your job 36 ("STDIN") has been submitted


zsh)puel.cctest15 []% cat STDIN.o36
processor       : 0
vendor_id       : GenuineIntel
cpu family      : 6
model           : 6
model name      : QEMU Virtual CPU version 0.9.1
stepping        : 3
cpu MHz         : 2327.496
[...]

             total       used       free     shared    buffers     cached
Mem:       1962048     174868    1787180          0       9876     131600
-/+ buffers/cache:      33392    1928656
Swap:       506036          0     506036
```