Virtualization in ATLAS
experiences, feedback, requirements

Flavia Donno, Yushu Yao, (Paolo Calafiura)
Jun 21 2010

Overview

- Virtualizing ATLAS central services

  (Flavia)
- Virtualization Experiences in ATLAS
  - CloudCRV and Virtual Cluster Appliances

    (Yushu)
- Summary of ATLAS feedback/requirements

- Talk about ATLAS Tier3 tomorrow

  (Doug Benjamin)

Virtualizing ATLAS Central Services

Flavia Donno (CERN)

Disclamer

- What follows is based on
  - ATLAS current experience with the VOBOX service provided by IT
  - ATLAS assumptions on what the VOBOX virtualization project will provide
- ATLAS feels that dedicated "understand/study" meetings, accurate preparation and planning are necessary before virtualizing ATLAS central services at CERN

ATLAS Central Services

- ATLAS counts on about 50 ATLAS specific services running on VOBOXes at CERN
- ATLAS VOBOXes are managed by the ATLAS Central Services Operations Team (ATLAS VO Contact – VOC) following the recommendations of CERN IT and the CERN Security Team.

ATLAS service criticality

- The services are divided into 3 categories, according to their "criticality":
  - **Very high**: interruption of these services affects online data-taking operations or stops any offline operations. Service downtime or reduced availability should be solved within (max) 4 hours
  - **High**: interruption of these services perturbs seriously offline computing operations. Service downtime or reduced availability should be solved within 12 hours (any time)
  - **Moderate**: interruption of these services perturbs software development and part of computing operations. Service interruption or reduced availability should be solved within 2 working days.
- Service criticality review is a continuous process.

ATLAS Service Management

- All Central Services machines are managed through quattor.
- ATLAS Services are delivered through rpms stored in the ATLAS rpm quattor repository.
- Configuration is being automated through a quattor component that execute a configuration script shipped through the service rpm.
- In most cases, configuration files are still shipped and installed manually, sometime via svn.
- Experiment MOD and expert on call might need to be able to change the service configuration.
- Configuration files with sensitive information are most of the time created/installed by hand.
- The knowledge about configuration is sometimes still in the hands of the service providers.

Services Feedback

- Some services are very demanding in terms of CPUs, Memory, I/O, Network, Storage
  - Hardware sharing for these services can be difficult
- Special needs
  - Reliable and fast connection to an external filesystem
  - Backup
  - MultiGbit network connections
- Very high availability (max 1 hour downtime)
  - If DNS load balancing is used, must be done using different hardware
  - DNS load balancing not always feasible given the stateful nature of the services
  - Need for available hot spare to be kept in sync with master copy (in terms of OS and software configuration)
- Dependency on machine IP
  - Some IPs are registered in pit firewall
- Sensitive to re-installation
  - Some services count on scp through scripting

Wish list for Service Virtualization

- Reduced time to recover the service in the case of hardware failure
- Share hardware resources wherever possible offering service insulation.
- Service level equal of better to the current one
- Completely quattorized setup
- It should be possible to use DNS load balancing with VM and PM in a mixed configuration.
- Special needs should be clearly specified in and arranged through the HW request form.

Wish list for Service Virtualization

- It should be possible to reconfigure VM hardware setup after the machine has been assigned, depending on the specific needs.
- Console access to VM
- VOC controllable VM HW migration
- TSM available for VM
- Gigabit network link in VM
- Overbooking of hw resources for VM.
- VOCs should be able to manage VM without opening support tickets. Tickets should only be needed at VM creation.
- Support for VM snapshots (virtual images) is desirable.

Virtualization Experiences in ATLAS

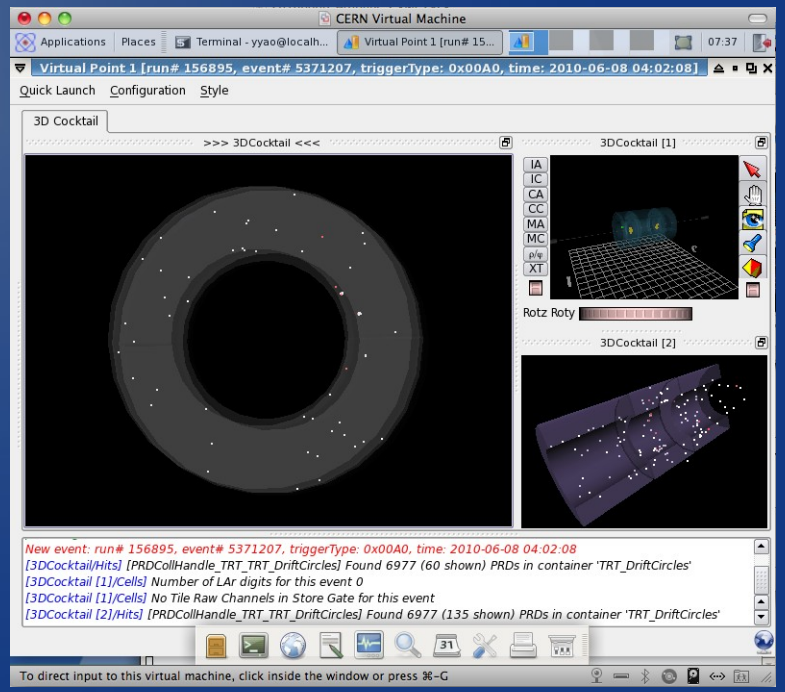
Yushu Yao (LBNL)

CernVM

SoFTware Appliance

The simplest yet fully functional Tier3-workstation.

- All ATLAS SW, Grid Job Submission
- 1-click VP1 Live
- Tutorial: https://twiki.cern.ch/twiki/bin/view/Atlas/CernVMTutorialHead
- Many Users, and increasing
- ATLAS wants CernVM to be supported like SLC at CERN.



ATLAS Tier3 Efforts

Work group March - June

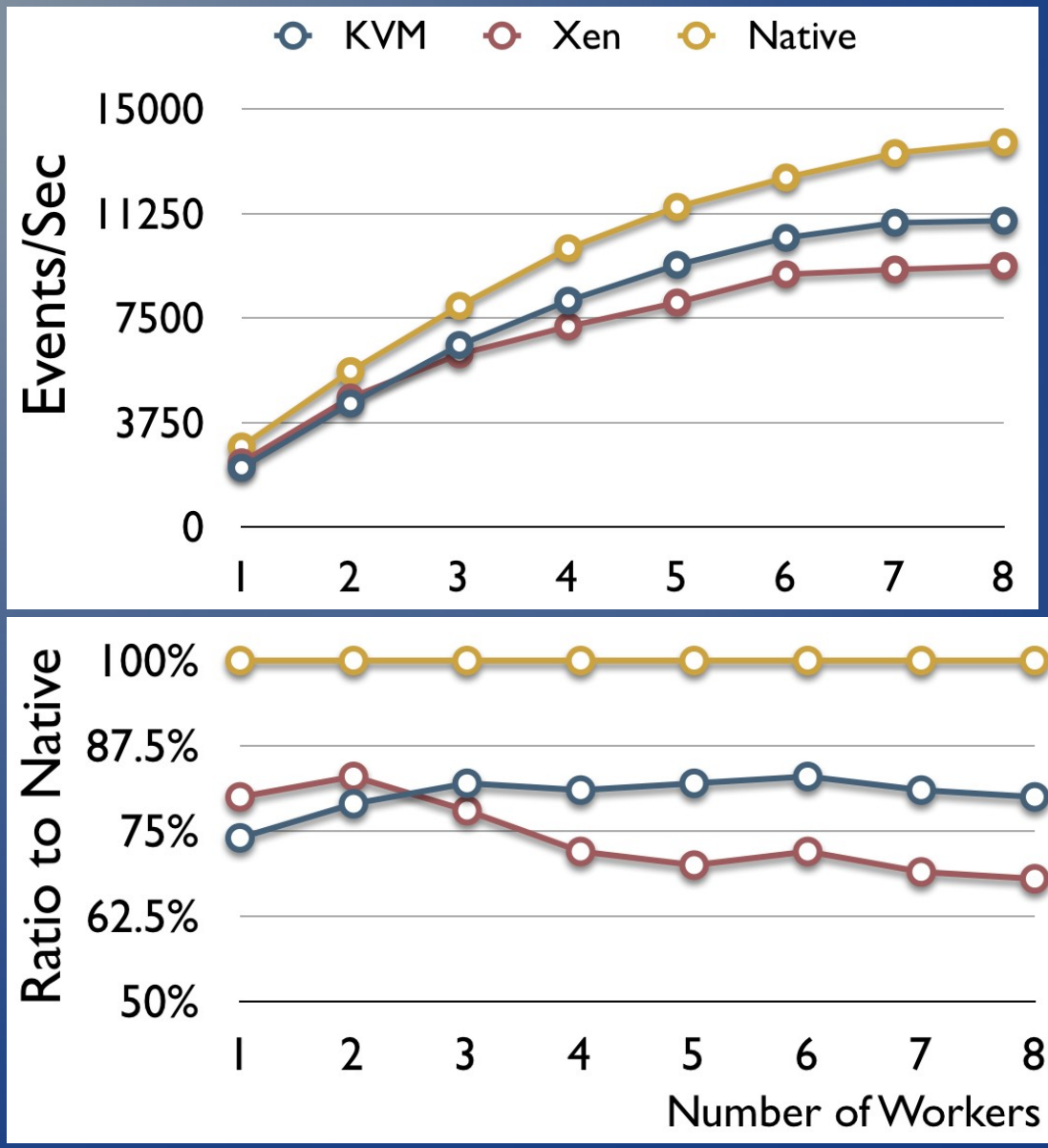
- Cluster Deployment/Configuration Automation (Simplification)
- Performance Tests (Services and Workers)

- Use Puppet+VM in Tier3's to simplify configuration
- CernVM-FS
  - ATLAS offline SW releases are officially distributed by CernVM-FS (Need long term support from IT)
- Details in the talk of Doug Benjamin.

Performance Testing

- Basic tools: nbench, iperf, bonnie++, ..
- MjMon/MpMon (ATLAS-specific Mous Tatarkhanov):
  - fire up N ATLAS jobs on a cluster
  - measure the total throughput as well as CPU/Mem/Disk usage statistics
  - + advanced features like NUMA monitoring, parallel processing, etc
- Test result in a nutshell (Single VM vs. PM)
  - CPU: several percent penalty
  - Disk I/O: 10-30% penalty
  - Network I/O: non-detectable penalty
- Cluster tests in progress (performance hit expected in VM clusters)
- Details http://vmstudy.blogspot.com/
- So? Good for service consolidation, no good to run jobs
  - Until disk I/O performance problem solved

A proof-lite job (I/O intensive)



CloudCRV & Virtual Cluster Appliance

Yushu Yao (LBNL)

Virtual Cluster Appliances

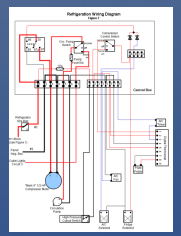
Virtual Appliance

OS and Application Stack
e.g. CernVM

What if my tasks need more than one VM to perform? e.g. A condor cluster

Combine Multiple Virtual Appliances
+
How they work together

Virtual Cluster Appliance



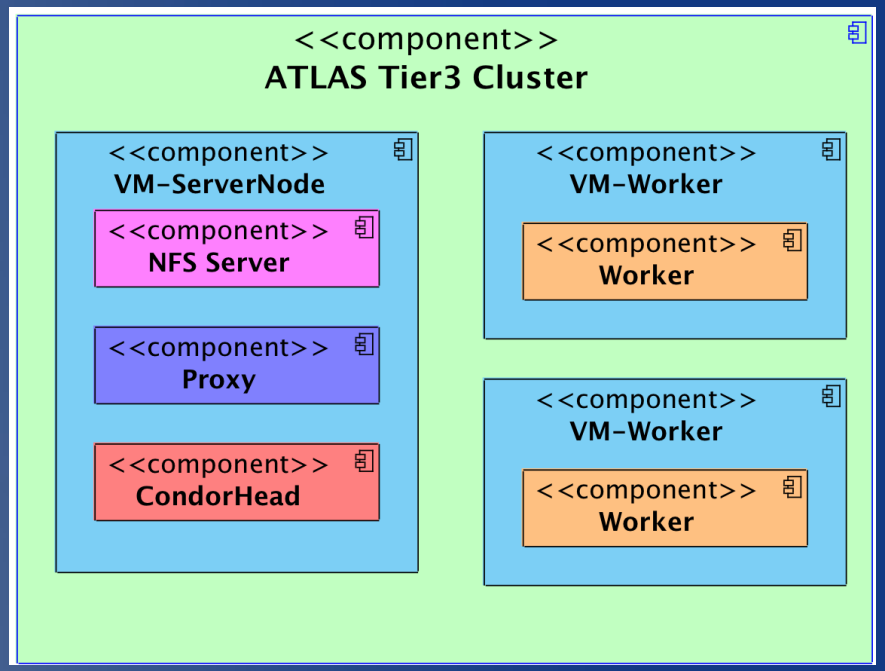
Manufacture

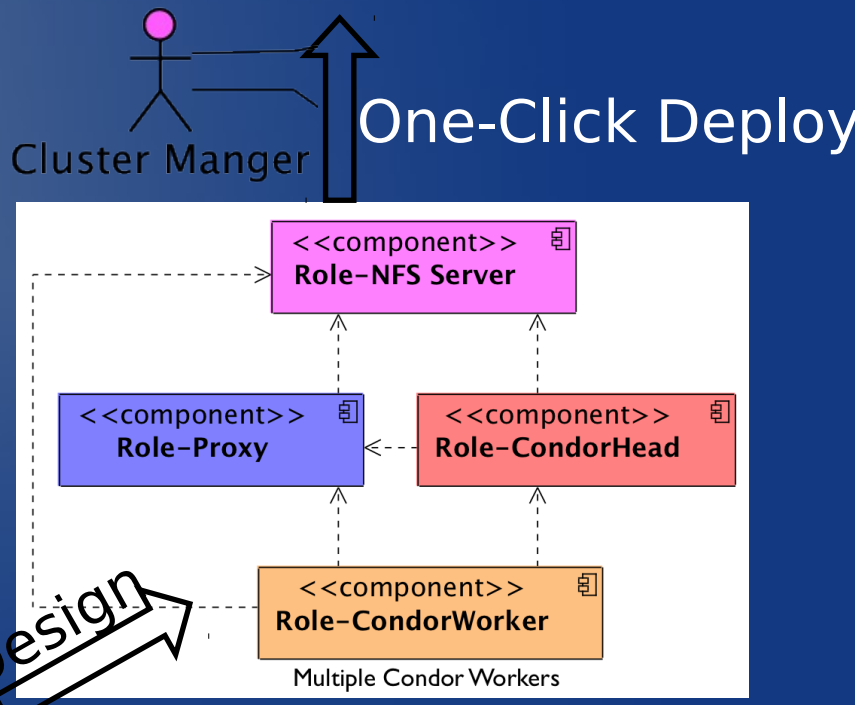
Ship entire Cluster as a product instead of VMs. Most people prefer to buy a fridge rather than assemble it from parts... Need little IT knowledge to deploy it, and need little effort to maintain.

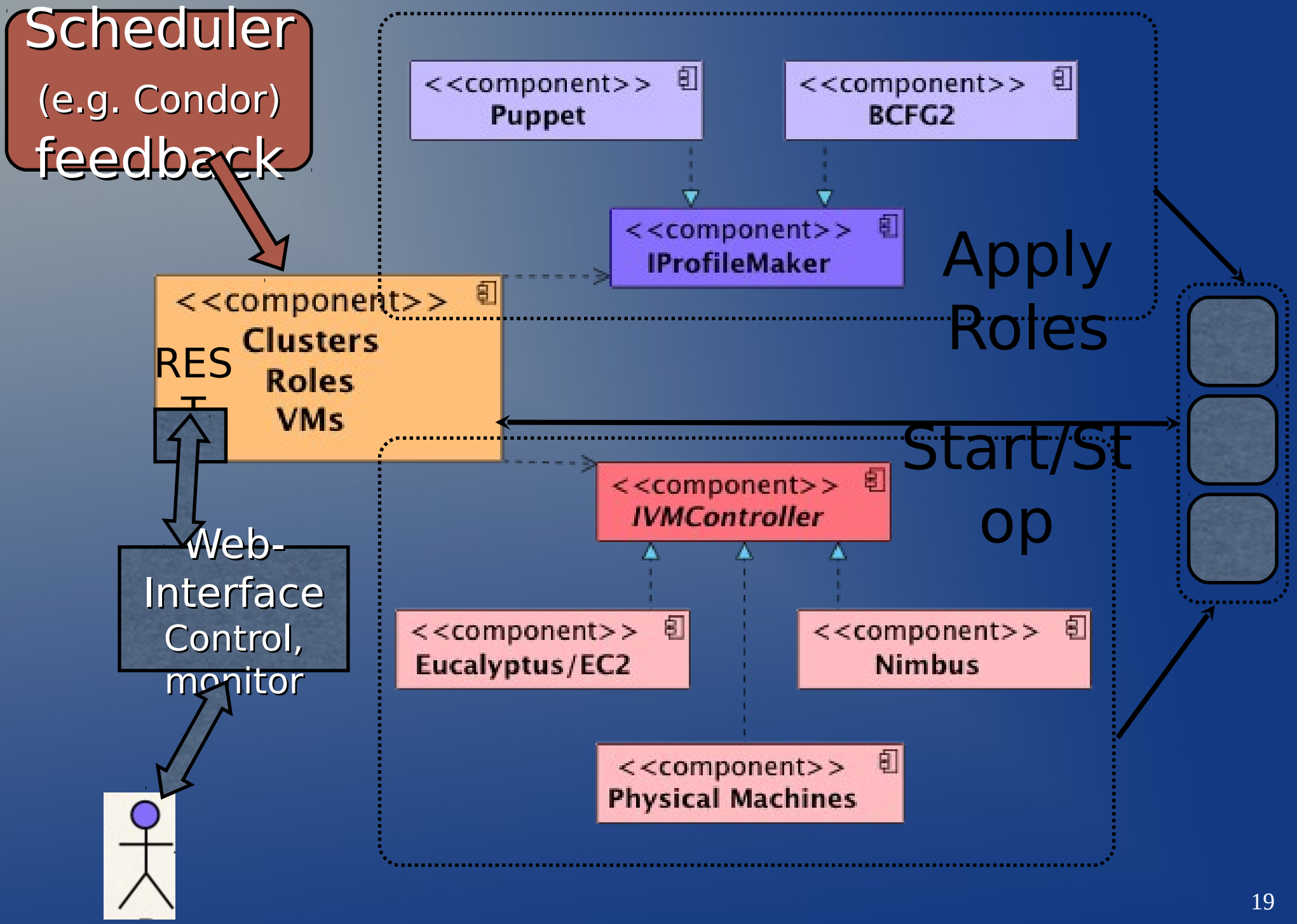
CloudCRV "Cluster-Role-VM"

deploy ("contextualize") set of *Roles* on PMs or VMs

- *Cluster Designers* design *abstract clusters:* a set of Roles and their dependencies ready to be deployed at sites
- *Cluster Managers* deploy the cluster with a click of a button
  - all Roles realized, and dependencies configured automatically with CM tools like Puppet
- The cluster can be deployed to resources like physical, virtual clusters and cloud.

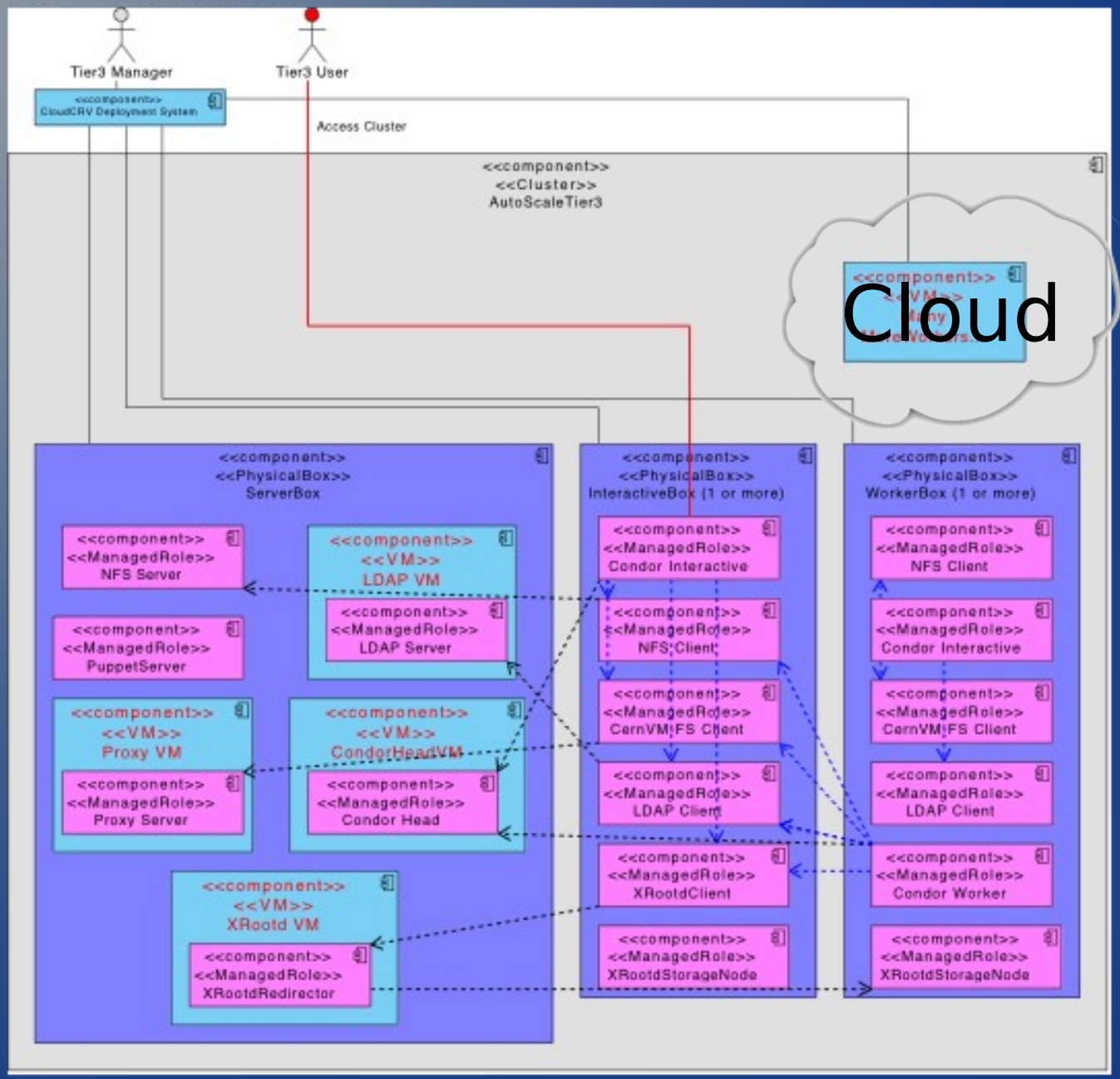





Scheduler
(e.g. Condor)
feedback
<<component>>
Puppet
<<component>>
BCFG2
<<component>>
IProfileMaker
Apply
Roles
<<component>>
Clusters
Roles
VMs
RES
Start/St
op
<<component>>
IVMController
Web-
Interface
Control,
monitor
<<component>>
Eucalyptus/EC2
<<component>>
Nimbus
<<component>>
Physical Machines

19

CloudCRV Today



Conclusions

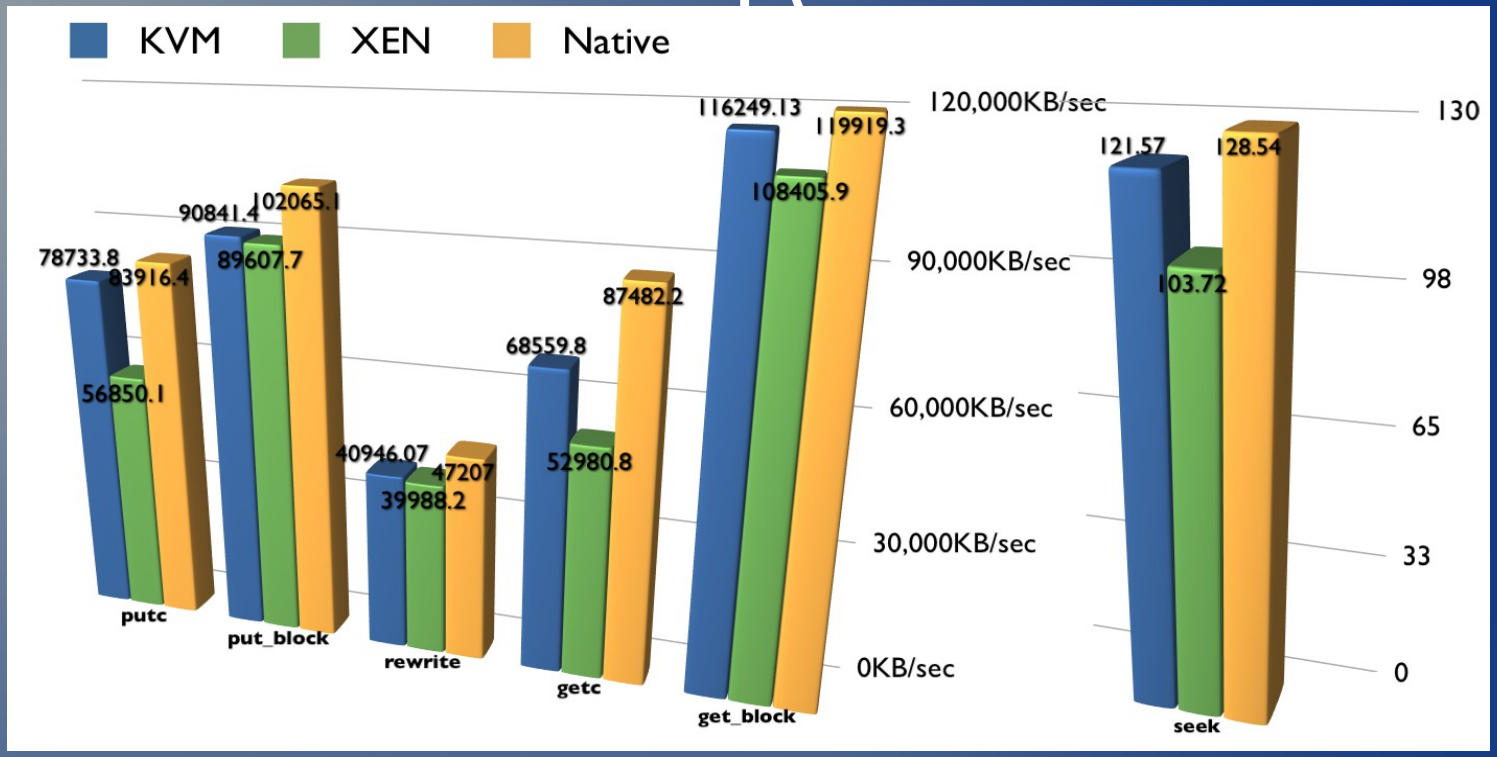
- VMs should provide the current VOBOX service level, with better availability and lighter, more automatic operations
- CernVM, cvmfs are priorities for ATLAS distributed analysis, and should be supported
- Virtualization is no replacement for proper software distribution and configuration management, but it can make these tasks much easier and faster.

Backup

ATLAS Service Types

- Specialized
  - Developed by ATLAS according to their computing model
  - This is the case for most of the highly critical services
- Web services
  - Developed using various frameworks (Django, mod_python, php, cherrypy, etc.)
  - Normally behind a web application firewall
  - Very critical services can follow in this category

Disk Access Speed in VM (bonnie+ ...)



- Test Machine Setup:
  - Nehalem dual-socket 8 core (total physical), 24GB Memory, SATA 3.0Gbps Drives
  - SL5.4 64-bit (Guest/Host)
- Using XEN & KVM included in SL
- In KVM passing a physical partition to guest via virtio (nocache)
- In XEN passing a physical partition to guest via xvd (slow might due to non-optimal configuration)
- Tested multiple configurations (including image based, best result shown above)