A ground-up approach to High-Throughput Cloud Computing in High-Energy Physics

from bare metal to end-user analysis

Dario Berzano

Supervisors: M.Masera, G.Ganis, S.Bagnasco

Università di Torino - Doctoral School of Sciences and Innovative Technologies

Ph.D. Seminar - Torino, 18 feb 2013

Overview

Institutions and projects

INFN Torino
Centro di Calcolo

Università di Torino

CERN PH-SFT
PROOF and CernVM

The ALICE Experiment
@ LHC

Computing in High-Energy Physics

ALICE: 80 TB raw data collected last week (pA)

High Throughput Computing: limited by I/O more than CPU

- Data processing is event-wise
  → *embarassingly parallel at order 0*

	RAW	ESD
p-p	0.41	0.23
Pb-Pb	5.59	5.26

ALICE event size [MB]

- In ALICE, reconstruction (RAW → ESD) does not reduce data size:
  - disks → store and locate large amounts of data
  - memory → event data structures consume RAM during processing, exposing performances to the risk of swap
  - network → massive data transfers in LAN and WAN

Virtualization and Cloud Computing

Virtualization

hardware resources do not correspond to exposed resources

Partitioning resources

export virtual small disks or CPUs not corresponding to actual hardware resources (e.g. Dropbox)

Emulating resources

run different OSes on top of each other, emulate old hardware and different architectures

Cloud computing

leverage virtualization to turn computing resources to services

Software as a Service

software is not installed but runs "somewhere" and expose a remote interface (e.g. Gmail)

Infrastructure as a Service

virtual machines run "somewhere" and they can be accessed as if they were physical nodes

Current computing center stack in Torino

End-user services

Services | Tier-2 Grid | Temporary farms

Maps services to virtual resources

OpenNebula IaaS cloud controller

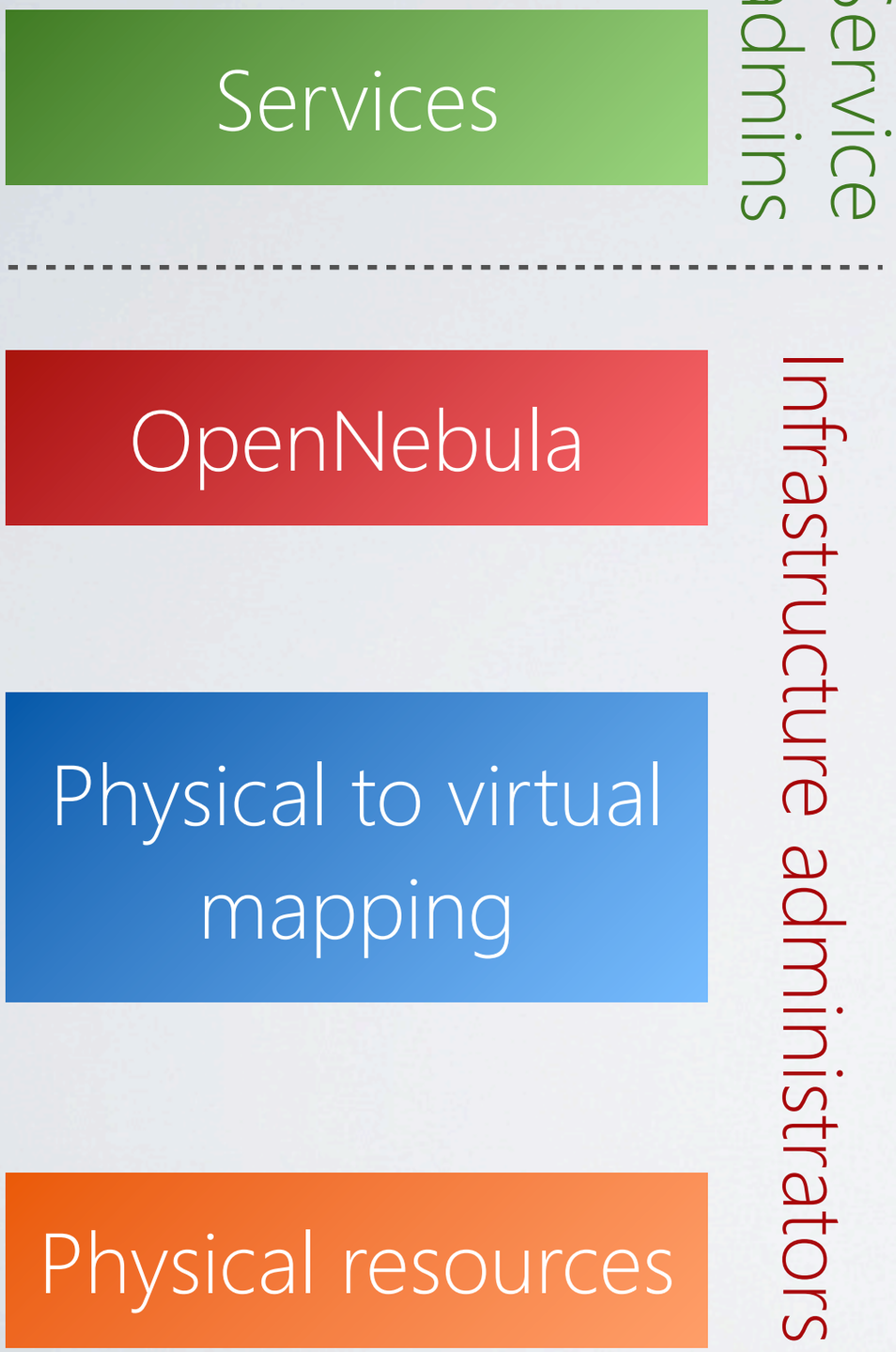
Map virtual resources to hardware resources

OpenWRT Virtual Router | QEMU-KVM hypervisor | GlusterFS shared and distributed

Bare metal

Network | Hosts | Storage

System administration in the cloud



- Clear separation of tasks between system and service administrator
- System administrator does not take care of running applications
- Service administrator is not concerned of hardware installation and failures
- Even our small computer center can now quickly support diverse (non-Grid) computing needs

Turning Torino's computing center into a private cloud provider

Design of our cloud architecture

- Limited amount of resources (*i.e.* money)
- Seamlessly integrate with existing services → smooth transition

Classes of hypervisors

Service hypervisors

aimed to provide high availability for critical services

- Virtual machines run from a disk shared between hypervisors
- Live migration in case of failure
  → *no service interruption*
- At the expenses of throughput
  → *only non-demanding head nodes*

Worker hypervisors

aimed to provide maximum throughput for data crunching

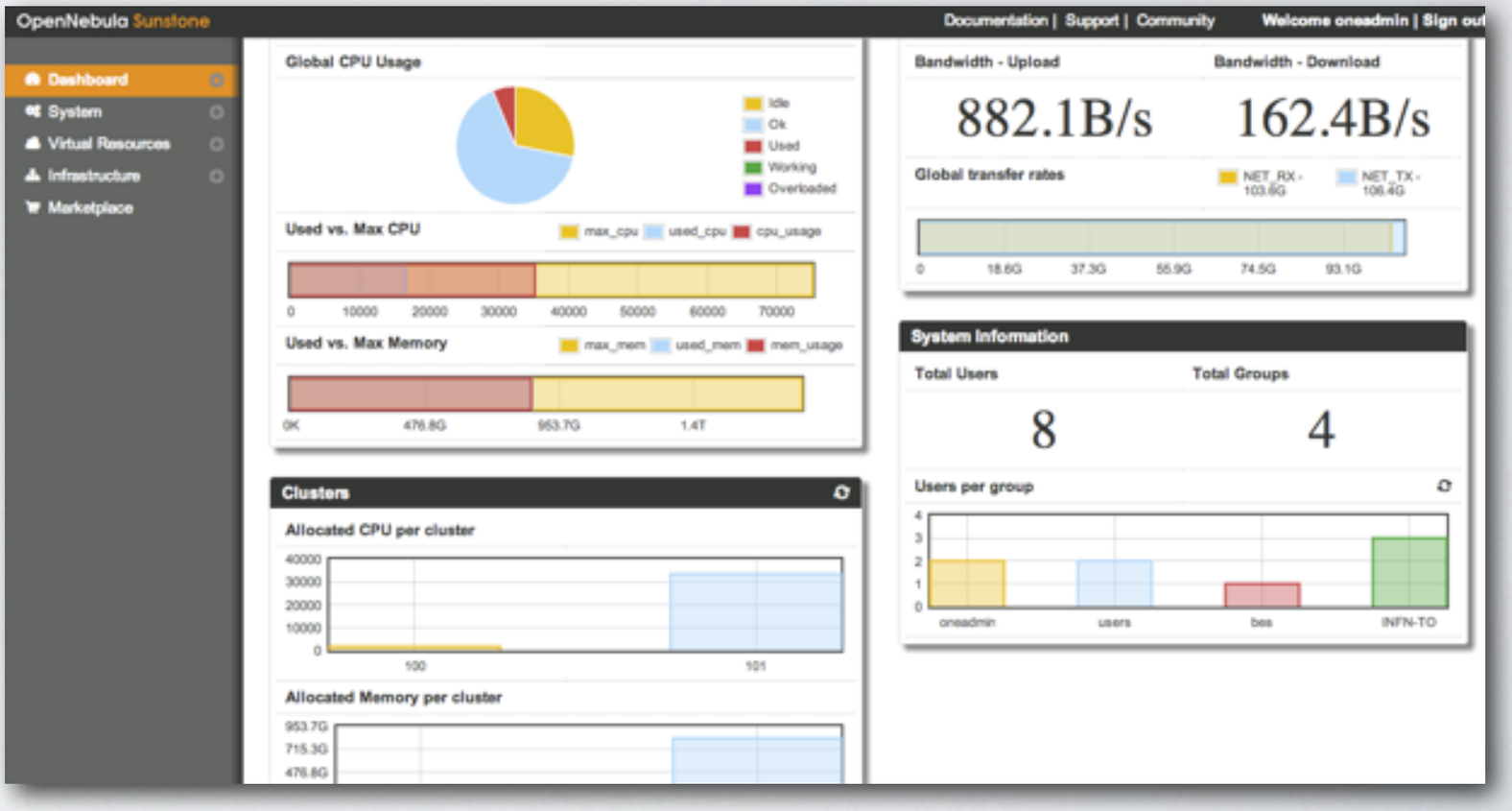
- Virtual machines run from their local disk
- No live migration possible, but maximum throughput (local I/O)
- Failures acceptable
  → *run restartable jobs (Grid)*

OpenNebula

OpenNebula: open source tool to orchestrate private clouds

- Controls and monitors hypervisor classes
- Manages lifecycle of virtual machines and virtual farms
- Repository of base images and marketplace for external images

- Web interface with virtual screen
- Clear codebase
  → *easy to customize*



OpenNebula.org

OpenNebula customizations

VM drain mode
for Grid over cloud integration

- Shutting down a Grid node does not have immediate effect
- Tell VM not to accept new jobs
- Shut down VM automatically when no jobs left
- Automatism very convenient with hundreds of VMs!

LVM caching+snapshotting
faster creation of LVM blockdevs

- When creating a VM disk over Logical Volume Manager
- Many copies of the same VM
- Base image cached on a LVM volume, transferred only once
- Copy on write: VM disks only store the differences wrt cache

Those features were patched to our OpenNebula installation

Fast deployment techniques

qcow2 images

- empty space does not count in image size
- slow when writing → attach scratch disks for output

Fast image transfer

- Commonly used images are stored in a separate OpenNebula repository
- Program to sync repository on hosts
  → *new images cached, old deleted*
- rsync in parallel, based on scp-wave: each dest node becomes source
  → *logarithmic speedup wrt n. hosts*

Quick scratch disk creation

- Creating ext3/4 filesystem is slow
- XFS for scratch disks
  → *mkfs immediate, size-independent*
- ext3/4 fast creation *(lazy_itable_init)*
- Scratch disks created on hypervisors with ext4 filesystem *(fallocate)*

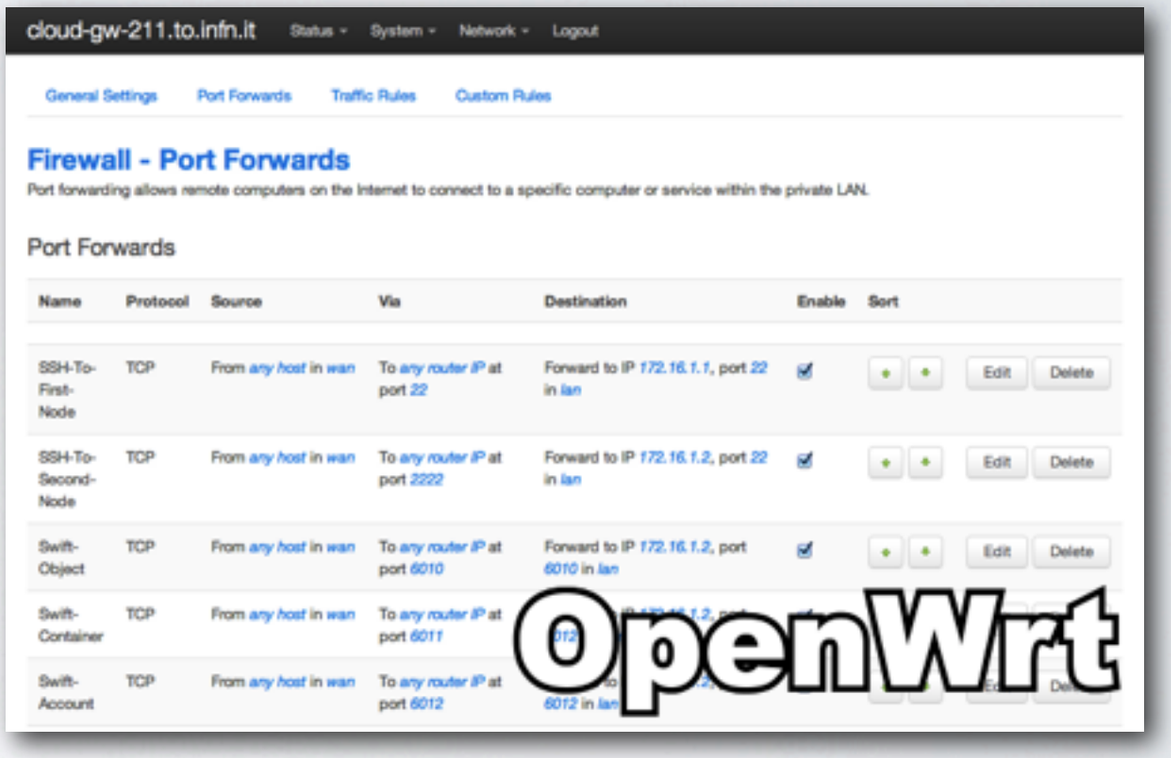
Single node boots in less than 15 seconds
40 nodes boot in less than 2 minutes

OpenWRT Virtual Router

Provide network connectivity and isolation for virtual farms

routers on VMs with functionalities very close to domestic routers

- Linux embedded distribution for routers with web interface: OpenWRT
  → *we tuned and compiled it to work on virtualized infrastructures*
- Low resources needed:
  → *1 CPU, < 200 MB RAM*
- MAC address net isolation
  → *ebtables*
- NAT, DNS forwarder, DHCP
- Firewall + forward to internal ports
  → *iptables*



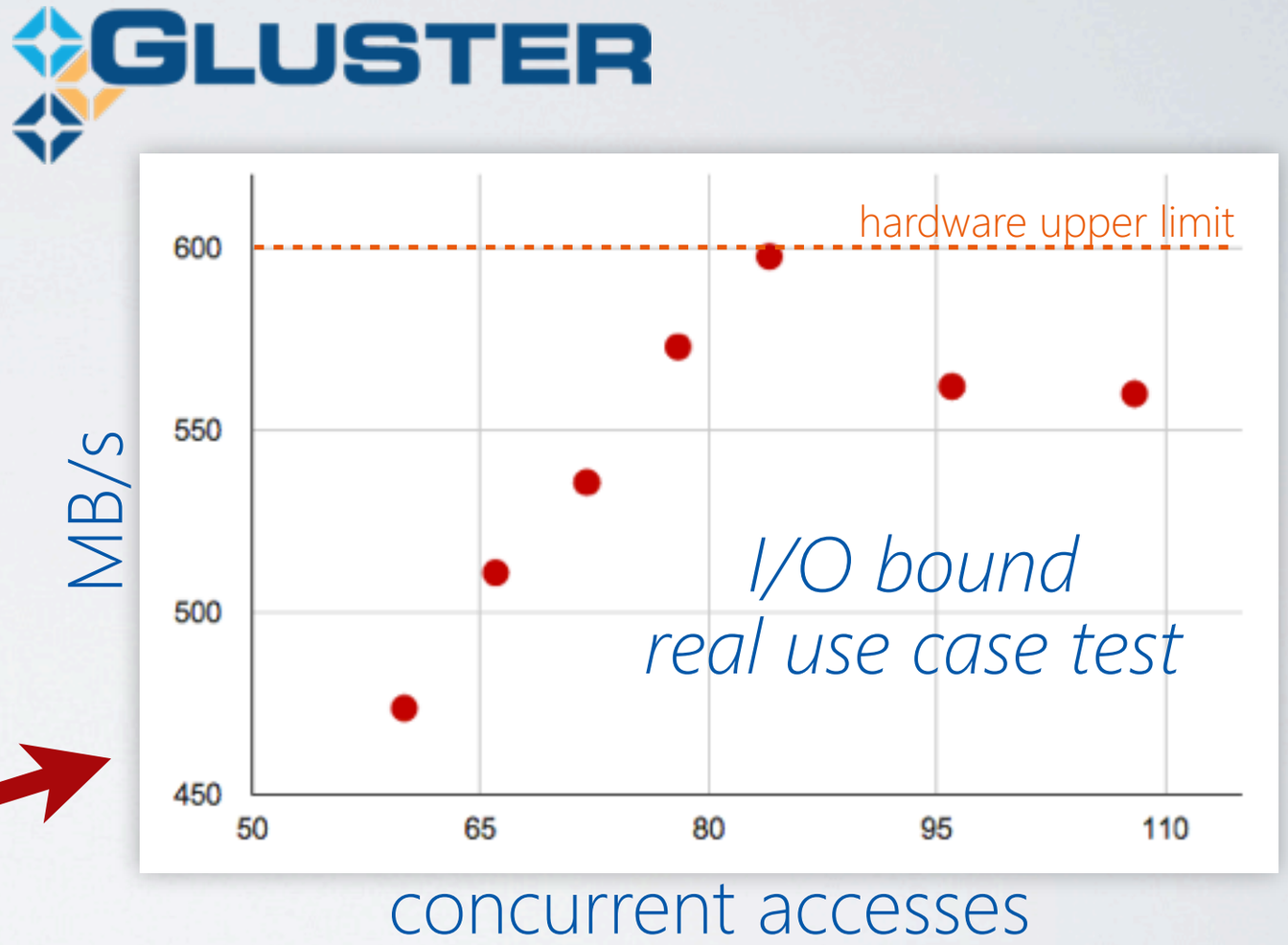
Cross-cloud geodistributed farms bridged via OpenVPN

ongoing testbed with Roma: first results promising

GlusterFS shared filesystem

Shared and distributed horizontally-scalable filesystem

- Mountable fs with FUSE
- RAID-alike functionalities
- We use it for:
  - image repository
  - live virtual machines
    → *two frontend servers with replicated disks for failover*
  - data repository
    → *fast concurrent access*

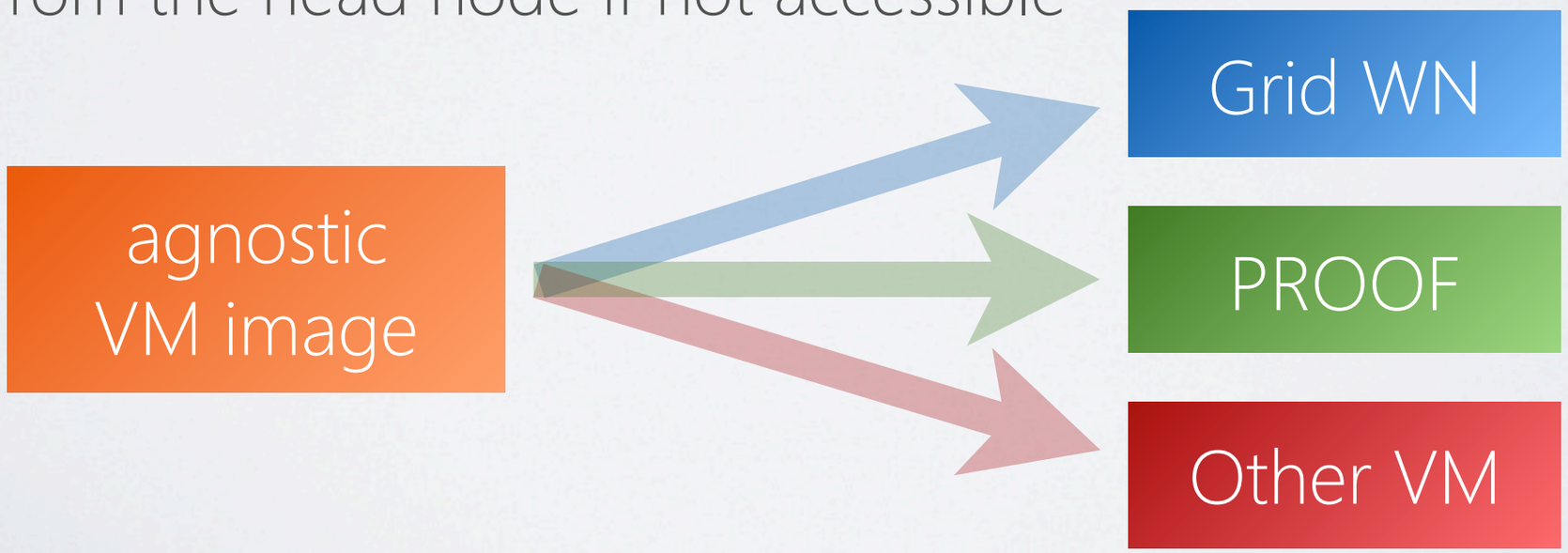


- Thanks to M.Concas for helping in testing and setup *(B.Sc. thesis)*
- Soon testing v3.4 (now in alpha) with direct QEMU support
  → *should provide speedup for running virtual machines*

Contextualization and elasticity

We maintain only one VM image for all major services

- OpenNebula: run contextualization script (HEPiX) inside a virtual machine
  → *context defines the fate of a virtual machine*
- Move as much configuration as possible in context script
  → *only one image to maintain (fresh software, security updates)*
- Elasticity: major services *(e.g. Grid)* tweaked to deal with dynamic removal and addition of computing nodes
- Elasticity policies: node self-addition and self-removal to its cluster, removal enforced from the head node if not accessible



Virtual services
running on our cloud

Sky computing or Grid over cloud

Seamlessly virtualize existing Grid worker nodes

- Legacy existing PBS/TORQUE local batch system to support the Grid
  → *doesn't play well with dynamic nodes but needs to be supported*
- Several workarounds to add dynamicity to PBS
  → *nodes self addition and removal through SSH*
- New virtual nodes mixed with old physical nodes
- Puppet configuration and Zabbix monitoring
- Rolling updates: major VM updates are performed by draining all running VMs and gradually replace them with fresh VMs
  → *this procedure is fully automatic*

Since Dec 2011: 85% of worker nodes are now virtual

Since Nov 2012: all critical Grid head nodes are virtual as well

M5L-CAD automatic lung CT analysis

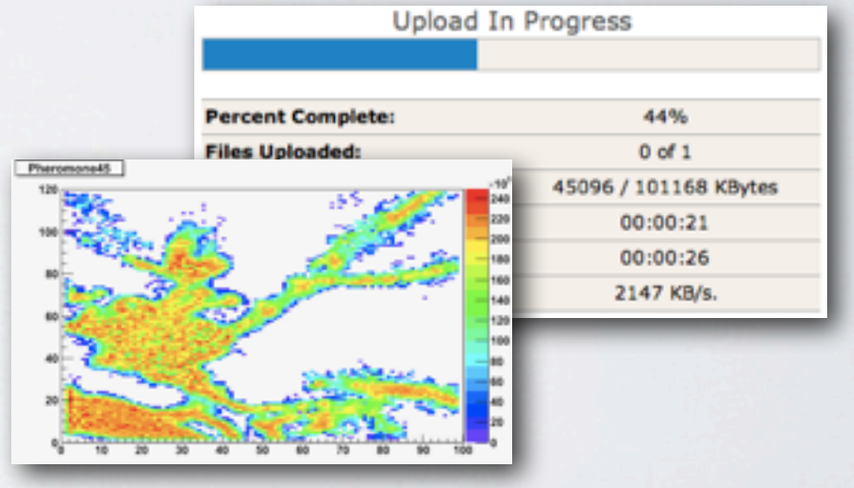
Fight lung cancer by massive pre-screening and cloud analysis

- Identification algorithms: RGVP, CAM, VBNA
  → *results and probabilities are then combined*
- SaaS Interface: WIDEN (Web-based image and diagnosis exchange network) manages the workflow
  → *the physician requires only a browser*
- Positives are signalled through email and SMS
  → *number of false positives is low*
- Provide opportunistic IaaS cloud resources
  → *VMs started to compute and halted when idle*

WIDEN®

dixit

MAGIC5



Poster presented at IEEE NSS MIC RTSD 2012

Other notable customers

Our private cloud lets us effortlessly support spot customers

we give them temporary VMs, they do the rest

- Solid state group (J.Forneris, E.Vittone)
  → *virtual machines for IBIC monte carlos using Open MPI*

- Theoretical physics (G.Passarino)
  → *lattice simulations requiring very large amount of memory*

- BESIII (M.Maggiora)
  → *ongoing whole computing model definition based on the cloud*

No hardware installation but temporary virtual resources

resources are claimed back to the Grid WNs when not needed

A Virtual Analysis Facility based on PROOF and CernVM

PROOF: the Parallel ROOT Facility

Using parallel resources interactively with ROOT

not a batch system: dynamic workload management and data merging

- Main advantage: it is based on ROOT, as all LHC HEP frameworks
- Work done for PROOF (integrated and distributed with ROOT):
  - The dataset stager: daemon that resiliently orchestrates data transfer and integrity from a remote site to the local PROOF facility. Supports dynamic reconfiguration without stopping current transfers
  - Interface to the native ALICE AliEn File Catalog that automatically discovers which files are staged locally
- Ongoing work:
  - Exploiting interactivity with a results collector that avoids merging phase

Issue: PROOF itself lacks a native user scheduler

users might ask and obtain too many unavailable resources

PROOF on Demand over HTCondor

PoD runs PROOF over a batch system

- Batch system to schedule PROOF processing
  → *users queue a resource allocation request (i.e., "I want X nodes")*
- HTCondor: deals optimally with dynamic addition/ removal of nodes
- Once requests are satisfied, user can reuse them "as many times as he/she wants" (modulo a job lifetime limit)
- Computing nodes (including head node) are HTCondor VMs
- User does not need to use a shell on the remote endpoint
  → *control from their laptop with pod-remote*

PoD
PROOF on Demand

HTCondor
High Throughput Computing

We don't preallocate PoD computing VMs

we monitor the queue and launch enough VMs to meet the demand

The CernVM ecosystem

Reference VM for LHC experiments and meta cloud controller

- Our base virtual machines are CernVM images
  → *ATLAS+CMS using as Grid nodes, ALICE next*
- Environment where your software is guaranteed to work
  → *dependencies always satisfied, low error rate on Grid*
- VM fate is entirely decided in context through CernVM Online
  → *contexts can be paired at VM boot interactively (or not) via web*
- Work done for CernVM and PROOF:
  - Provide a context def independent from clouds and experiments
  - Cloud Scheduler ideas to bring elasticity to CernVM Cloud *(ongoing)*

Preliminary CernVM Virtual Analysis Facility running in Torino

based on another VAF (non-CernVM) we presented at ACAT 2011
full presentation at ROOT Users Workshop 2013 in March

Users authentication

sshcertauth: use existing Grid credentials to access PoD

- PoD encrypts all communications through simple SSH
- Grid users already have a RSA private key and X.509 certificate
  → *the private key can be used directly to access SSH*

Source (PHP+BASH):
https://github.com/dberzano/sshcertauth

- Problem: the server does not recognize the public key
  → *public key is authorized by connecting via SSL auth to HTTPS*
- For ALICE: Unix users on VMs are the same user names on the Grid
  → *association between X.509 certificate and username done via LDAP*
- Very secure, works with any standard browser and normal SSH client

Work done for CernVM: embedded sshcertauth in CernVM

Notable analysis currently running on PROOF

Λ hypernuclei in ALICE through invariant mass analysis

- Searching Λ hypernuclei (hypertriton and anti-hypertriton) in heavy ion collisions through invariant mass analysis
- Rare events: requires large amount of data to analyze
  → *ideal for storage testing: ~20 TB for a lead-lead LHC 2010 run period*
- Many cuts plus topological cuts to improve signal (also TPC+ITS refit!)
  → *CPU efficiency above 80%*
- *Particle identification selection through OADB access*
  → *accessed for each run: network also heavily stressed*
- Light output data: ~80 histograms
- Analysis with S.Bufalino, E.Botta

Well-balanced between CPU, network and disk usage

perfect use case for PROOF and ideal testbed for our cloud deployment

Presentations and publications

Computing presentations and publications

- Advanced Computing and Analysis Techniques 2011
  Berzano D, Bagnasco S, Brunetti R, Lusso S "PROOF on the Cloud for ALICE using PoD and OpenNebula" 2012 J. Phys.: Conf. Ser. 368 012019

- sshcertauth
  Berzano D "SSH authentication using Grid credentials" INFN-12-20/TO

- Workshop congiunto INFN CCR-GARR 2012
  Berzano D, Bagnasco S, Brunetti R, Lusso S "Consolidating and extending Tier-2 resources to serve a larger user community"

- IEEE NSS MIC RTSD 2012
  Berzano D, Bagnasco S, Brunetti R, N Camarlinghi et al. "On-Demand Lung CT Analysis with the M5L-CAD via the WIDEN Front-End Web Interface and an OpenNebula-Based Cloud Back-End" IEEE NSS MIC RTSD 2012 Conference Record N14-73

- Workshop GARR - Calcolo e Storage distribuito 2012
  Berzano D, Bagnasco S, Brunetti R, Concas M, Lusso S "Gestione di applicazioni di calcolo eterogenee con strumenti di Cloud Computing: l'esperienza di INFN Torino"

ALICE Collaboration publications/1

- Measurement of charm production at central rapidity in proton-proton collisions at sqrt(s) = 7 TeV
  JHEP 01 (2012) 128

- J/psi polarization in pp collisions at sqrt(s)=7 TeV
  Phys.Rev.Lett. 108 (2012) 082001

- Heavy flavour decay muon production at forward rapidity in proton-proton collisions at sqrt(s) = 7 TeV
  Phys. Lett. B 708 (2012) 265

- Measurement of Event Background Fluctuations for Charged Particle Jet Reconstruction in Pb-Pb collisions at sqrt(s_NN) = 2.76 TeV
  JHEP 03 (2012) 053

- Light vector meson production in pp collisions at sqrt(s) = 7 TeV
  Phys. Lett. B 710 (2012), pp. 557-568

- J/psi Production as a Function of Charged Particle Multiplicity in pp Collisions at sqrt(s) = 7 TeV
  Phys. Lett. B712 (2012) 165-175

ALICE Collaboration publications/2

- Multi-strange baryon production in pp collisions at sqrt(s) = 7 TeV with ALICE
  Phys. Lett. B 712 (2012) 309
- Underlying Event measurements in pp collisions at sqrt(s) = 0.9 and 7 TeV with the ALICE experiment at the LHC
  JHEP 1207 (2012) 116
- Measurement of charm production at central rapidity in proton proton collisions at sqrt(s)=2.76 TeV
  JHEP 1207 (2012) 191
- J/psi suppression at forward rapidity in Pb-Pb collisions at sqrt(s_NN) = 2.76 TeV
  Phys. Rev. Lett. 109, 072301 (2012)
- Transverse sphericity of primary charged particles in minimum bias proton-proton collisions at sqrt(s)=0.9, 2.76 and 7 TeV
  Eur. Phys. J. C (2012) 72:2124

ALICE Collaboration publications/3

- Production of muons from heavy flavour decays at forward rapidity in pp and Pb-Pb collisions at sqrt(s_NN) = 2.76 TeV
  Phys. Rev. Lett. 109, 112301 (2012)
- Suppression of high transverse momentum prompt D mesons in central Pb-Pb collisions at sqrt(s_NN)=2.76 TeV
  JHEP 9 (2012) 112
- Neutral pion and **η** meson production in proton-proton collisions at sqrt(s) = 0.9 TeV and 7 TeV
  Phys. Lett. B 717 (2012), pp. 162-172
- K0s-K0s correlations in 7 TeV pp collisions from the LHC ALICE experiment
  Phys. Lett. B 717 (2012) pp. 151-161
- Production of K*(892)0 and phi(1020) in pp collisions at sqrt(s)=7 TeV
  Eur. Phys. J. C72(2012)2183

ALICE Collaboration publications/4

- Measurement of prompt J/psi and beauty hadron production cross sections at mid-rapidity in pp collisions at sqrt(s)=7 TeV
  JHEP11(2012)065
- Coherent J/Psi photoproduction in ultra-peripheral Pb-Pb collisions at sqrt(s_NN)=2.76 TeV
  Phys. Lett. B 718 (2013), pp. 1273-1283
- Ds meson production at central rapidity in proton-proton collisions at sqrt(s) = 7 TeV
  Phys. Lett. B 718 (2012), pp. 279-294
- Inclusive J/psi production in pp collisions at sqrt(s)=2.76 TeV
  Phys. Lett. B718 (2012) 295
- Pion, Kaon, and Proton Production in Central Pb-Pb Collisions at sqrt(sNN) = 2.76 TeV
  Phys. Rev. Lett. 109, 252301 (2012)

ALICE Collaboration publications/5

- Measurement of the Cross Section for Electromagnetic Dissociation with Neutron Emission in Pb-Pb Collisions at sqrt(s_NN) = 2.76 TeV
  Phys. Rev. Lett. 109, 252302 (2012)

- Measurement of electrons from semileptonic heavy-flavour hadron decays in pp collisions at sqrt(s) = 7 TeV
  Phys. Rev. D 86, 112007 (2012)

- Charge separation relative to the reaction plane in Pb-Pb collisions at sqrt(s_NN)= 2.76 TeV
  Phys. Rev. Lett. 110 (2013) 012301

- Long-range angular correlations on the near and away side in p--Pb collisions at sqrt(s) = 5.02 TeV
  Phys. Lett. B 719 (2013), pp. 29-41

- Pseudorapidity density of charged particles in p-Pb collisions at sqrt(s) = 5.02 TeV
  Phys. Rev. Lett. 110, 032301 (2013)

Submitted and forthcoming submissions

- ROOT Workshop 2013
  "Distributed data analysis with PROOF and PoD" (accepted)

- Advanced Computing and Analysis Techniques 2013
  Berzano D, Bagnasco S, Brunetti R, Lusso S "Managing a Tier-2 Computer Centre with a Private Cloud Infrastructure" (abstract submitted)

- IEEE CLOUD 2013 *(full paper submission - forthcoming)*

- CHEP 2013 *(two abstract submissions - forthcoming)*