CHEP 2013 – Track 3A Summary "Distributed Processing and Data Handling" on "Infrastructure, Sites and Virtualization"

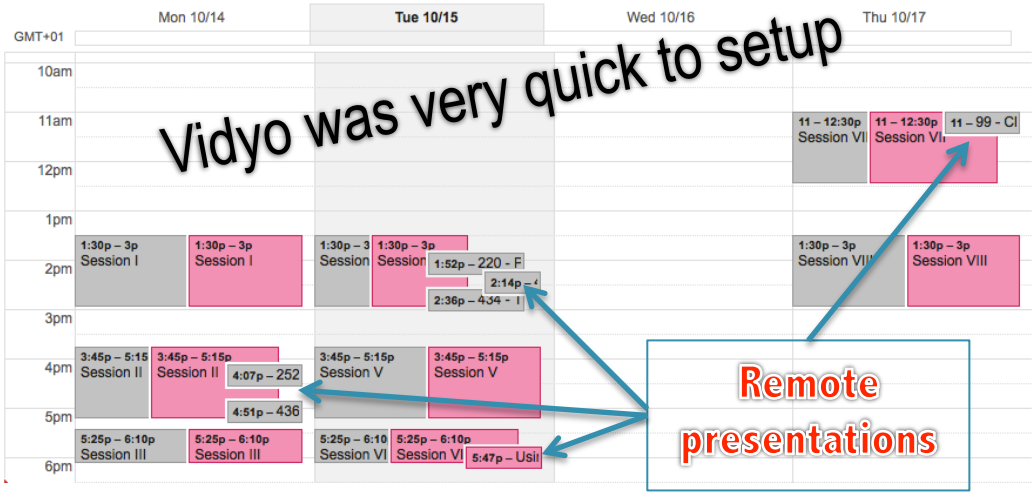
Stefan Roiser (CERN),
Davide Salomoni (INFN)

Track3 why A?

- The “Distributed Computing” track had double amount (142 accepted) of submissions to the next “busiest” track. Therefore we split
  - Track 3A “Infrastructure, Sites and Virtualization”
  - Track 3B “Experiment Data Processing, Data Handling and Computing Models”
- 4 (actually 3) conveners for both

US government shutdown

- Mostly in 3B many remote presentations, b/c of government shutdown ☹
  - Several other presentations had to change speaker

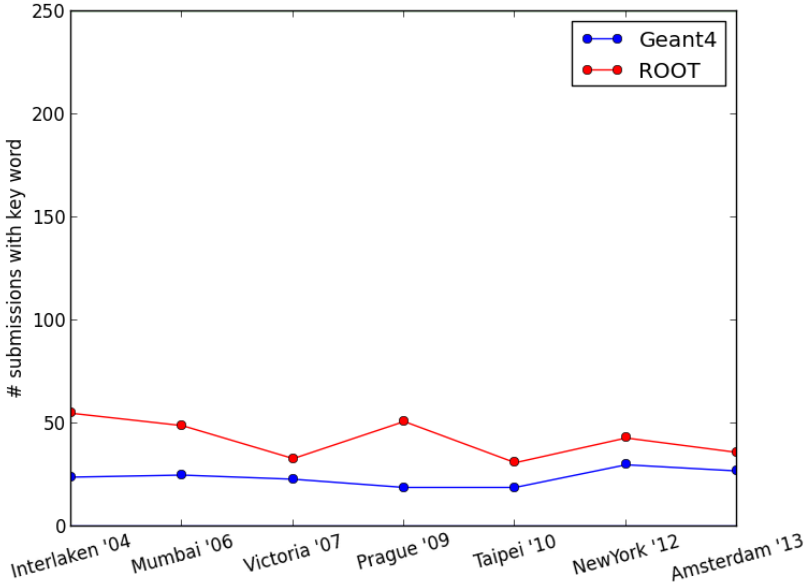


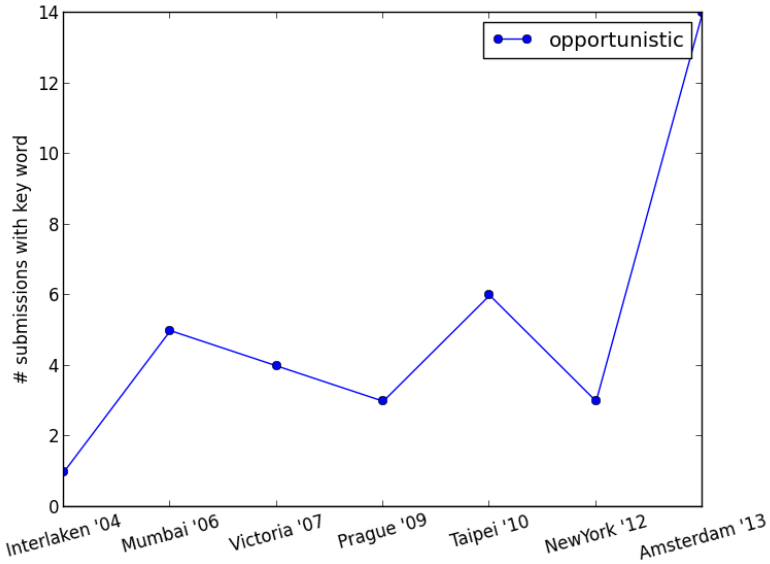
Track 3A Statistics

- As for other parallel tracks it contained 8 sessions with 28 talks
  - Opportunistic Computing
  - Grid & Cloud
  - (Common) Tools
- Sessions were very well attended
  - For some talks there was too little time for questions / discussion

Trends?

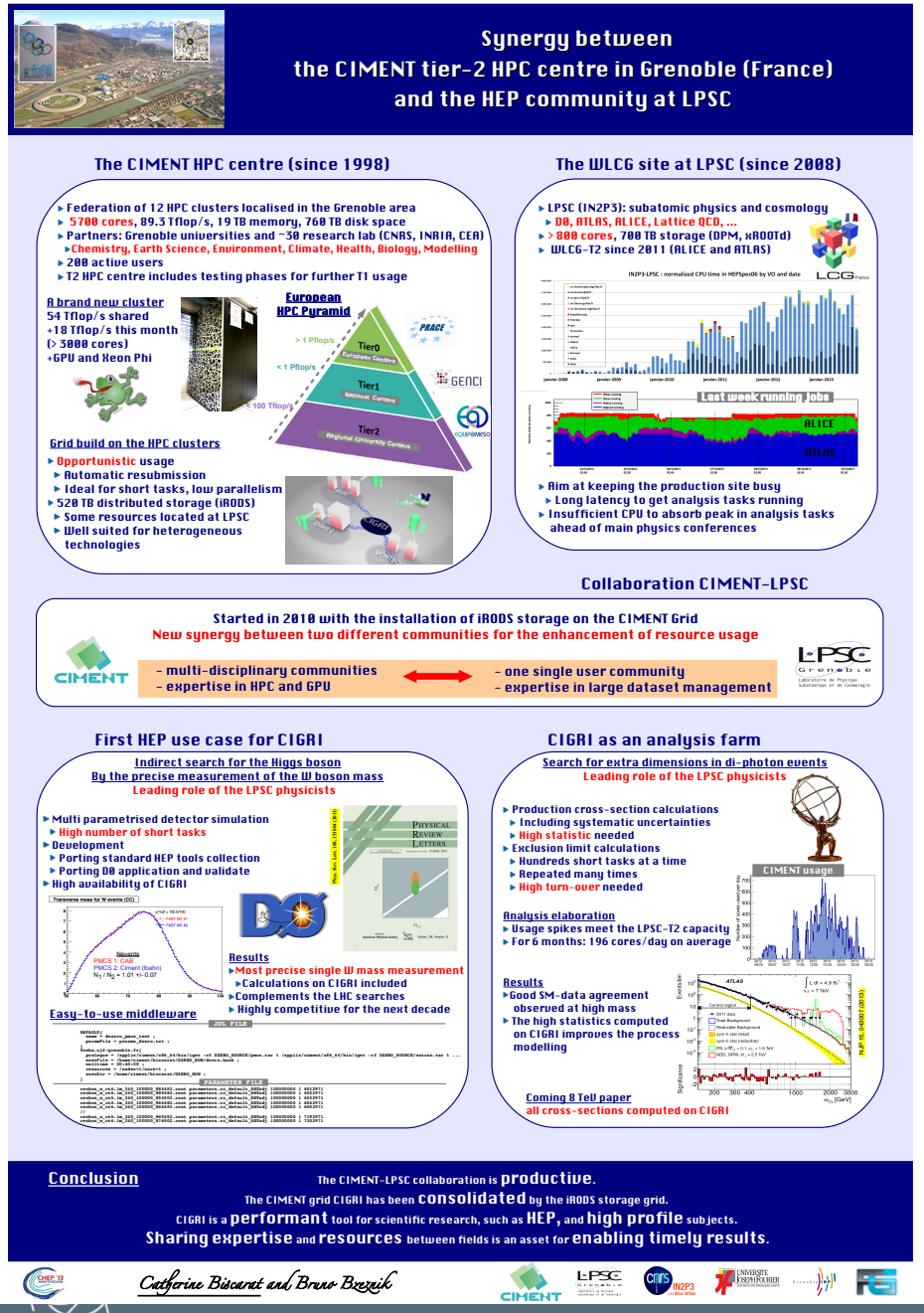
- Idea to look for “keywords” in the last couple of CHEP “book of abstracts” (since Interlaken 2004)
  - Plots show how many abstracts contain a given “<keyword>”



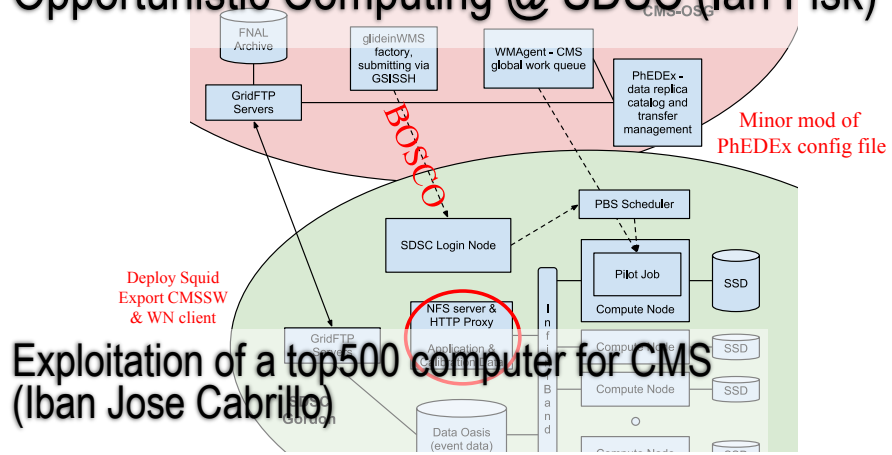


OPPORTUNISTIC COMPUTING

Opportunistic Computing large scale



Opportunistic Computing @ SDSC (Ian Fisk)



Exploitation of a top500 computer for CMS (Iban Jose Cabrillo)

Ex1: Needed to process "parked data" with extra resources @ SDSC

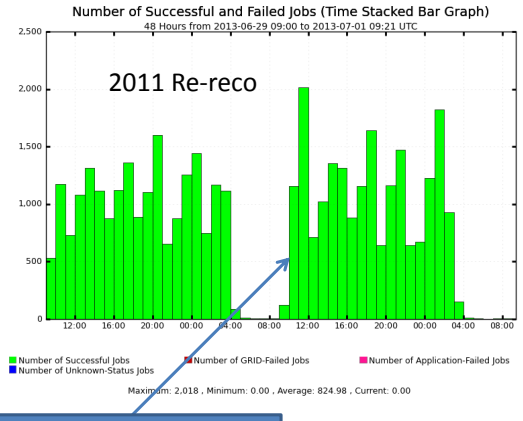
Ex2: Integration of Altamira supercomputer for peak loads

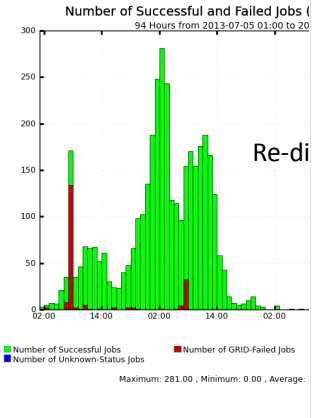
- Poster #226 on BOSCO (used to interface to HPC cluster)
- Poster: Leveraging HPC resources for High Energy Physics

Usage of HLT farms

Usage of the CMS HLT Farm as Cloud Resource (David Colling)

- used as a production resource





June and July

~6000 Jobs running
If only 1000 finishing/hour

Two workflows that were run

CMS usage of HLT farm, setup openstack to allow disentangle HTL/offline. Used in production during June/July

Sim@P1 PROJECT

DESIGN AND PERFORMANCE OF THE VIRTUALIZATION PLATFORM FOR OFFLINE COMPUTING ON THE ATLAS TDAQ FARM

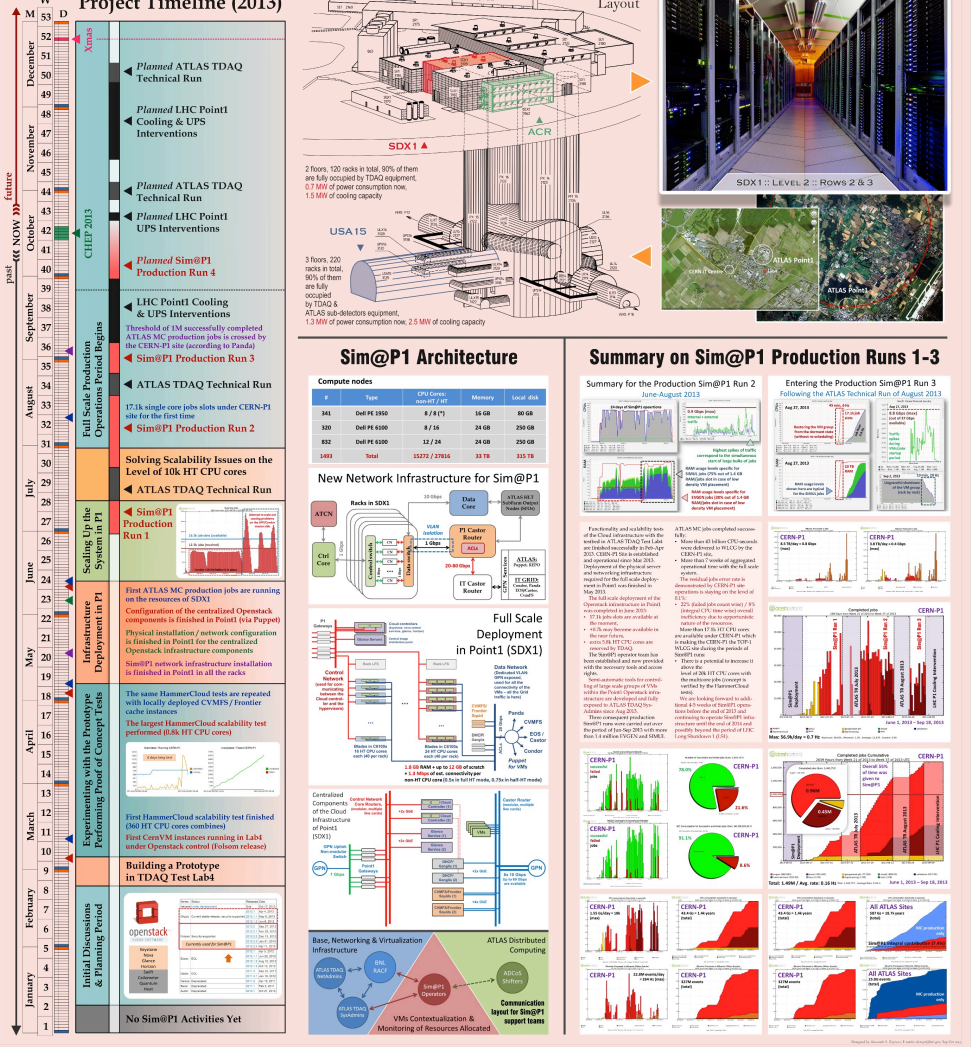
ATLAS TDAQ SysAdmins Group: S. Ballestrero, S.M. Batraneanu, F. Brasolin, C. Contescu, A. Di Girolamo, C.J. Lee, M.E. Pozo Astigarraga, D.A. Scannicchio, M.S. Twomey

ATLAS TDAQ NetAdmins Group: A. Zaytsev

ATLAS Distributed Computing Group

RHIC & ATLAS Computing Facility (RACF)

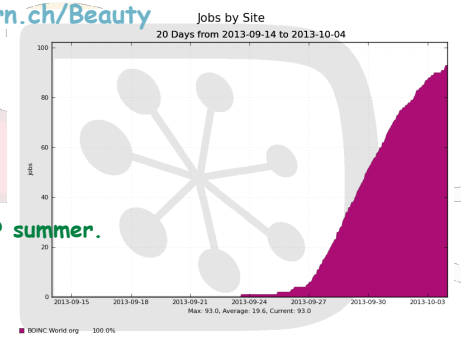
CERN, Switzerland
University of Johannesburg, South Africa
University of California, Irvine, USA
Istituto Nazionale di Fisica Nucleare Sezione di Bologna, Italy
Polytechnic University of Bucharest, Romania
CERN IT-SDC-OL
University of Washington Department of Physics, USA
Brookhaven National Laboratory (BNL), USA



Opportunistic computing small scale

Integration .. Cloud .. LHCb

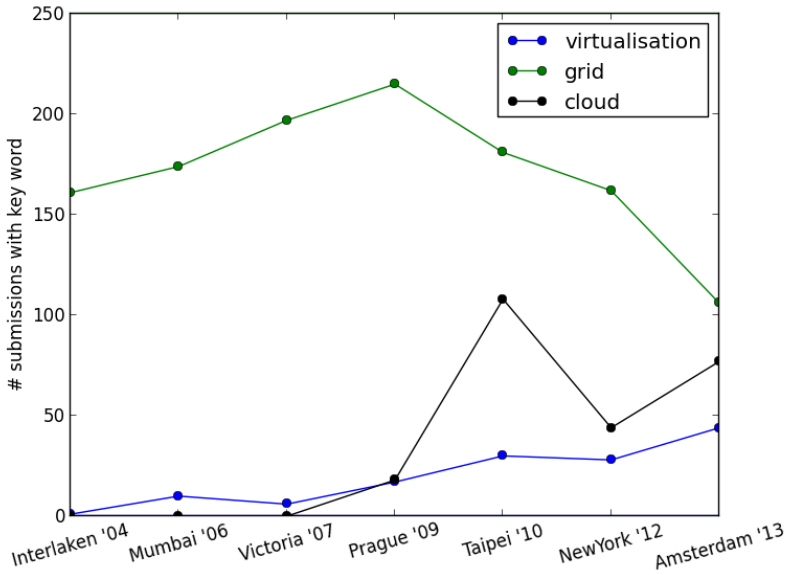
- Thanks B.Cabarrou for your CPU Cycles !
- If you want to play:
  - http://lhcbathome.web.cern.ch/Beauty
  - lhcb-boinc@cern.ch
- Different contributions
  - BOINC sites:
    - Building 2 @ CERN,
    - University Student Labs @ summer.
  - Single users.
    - There are power users !



Do you leave your desktop ON during the night ?

Integration of cloud resources (Mario Ubeda). LHCb use of BOINC is ready in production and used on "corridor PCs" and laptops

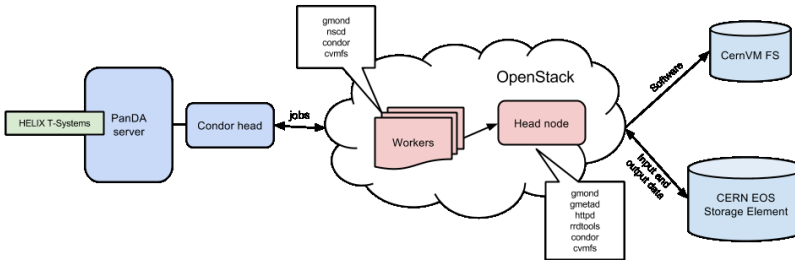
GRID & CLOUD


virtualisation
grid
cloud
submissions with key word
250
200
150
100
50
0
Interlaken '04
Mumbai '06
Victoria '07
Prague '09
Taipei '10
NewYork '12
Amsterdam '13

VO Interaction with IaaS Infrastructures

(Ramon Medrano Llamas)

Deployment



CERN AI infrastructure, currently 15k cores, scale up to 300k by 2015

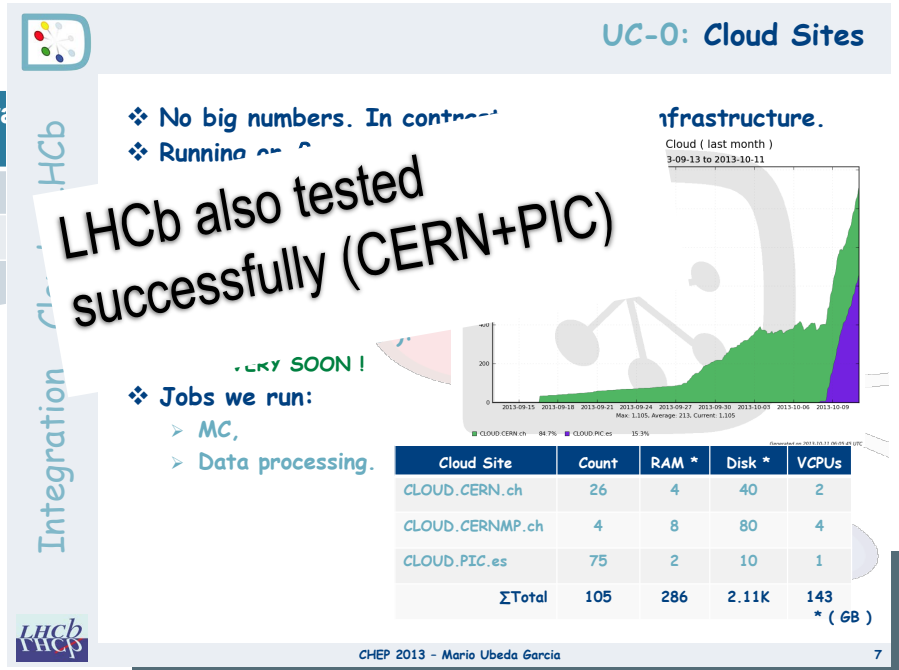
Tested with total of 690k ATLAS and 337k CMS jobs (on 770/200 cores) with relatively short jobs

Performance testing: ATLAS, Gr

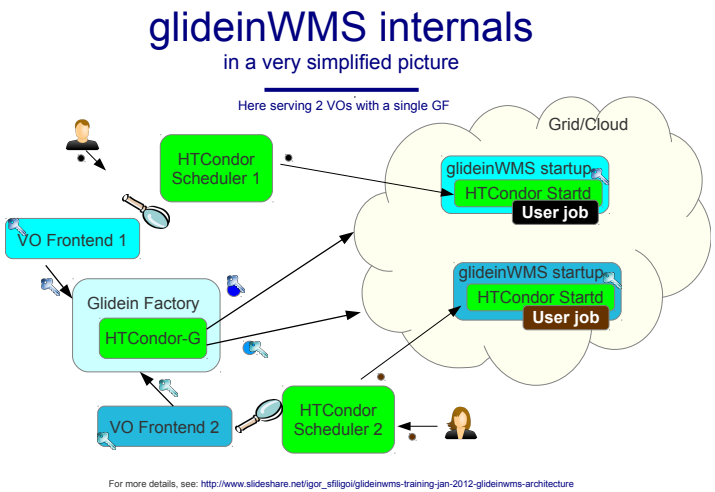
Site	Wallclock (s)	CPU efficiency(%)	Failure r (%)
OPENSTACK_CLOUD	1,827	82.3	13.7
BNL_CLOUD	1,960	69.9	-
IAAS	1,417	67.5	-
CERN-PROD	1,499	82.3	
BNL_CVMFS_1	1,611		

- Test
- G
  - A

Job success rate and efficiency comparable to bare metal batch systems



VOs Interfacing to Cloud Resources



Cloud Bursting with glideinWMS (Igor Sfiligoi), enabling HTCondor/glideinWMS to interface to grid resources

Cloud flexibility using DIRAC Interware

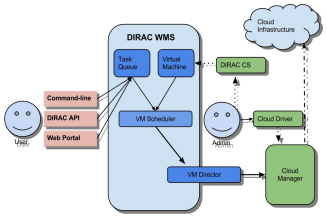
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

DIRAC THE INTERWARE

VMDIRAC Multi-Platform

The DIRAC Admin have to upload the images to the Cloud Manager using the corresponding Cloud Driver, and set Cloud specific values on the DIRAC Configuration.

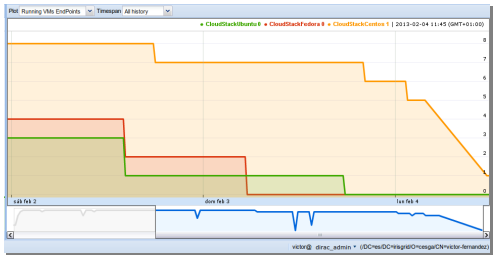
The VM Scheduler starts a full VM with DIRAC pre-installed and configured to execute the Job Agent, together with a VM Monitor Agent.



VMDIRAC can interacts with generic Cloud Managers to virtual machine submission. The Cloud Managers can be CloudStack, OpenStack, Amazon EC2 or OpenNebula.

VMDIRAC was designed with a Multi-Endpoint concept. DIRAC provides flexibility with the support of Grid, BOINC, Clusters and Clouds.

Results

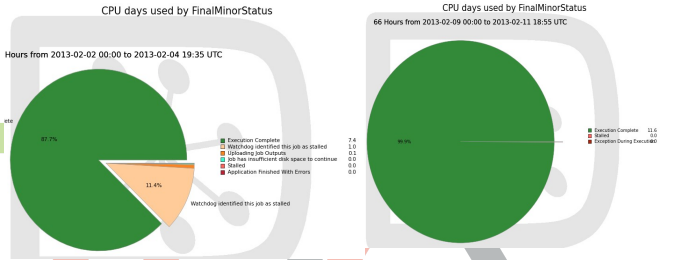


- 500 short jobs, with a time execution of 20 minutes

- 50 long jobs, of around 8 hours of execution time.

- Each user has been configured in a specific VO, and each VO has been assigned to a unique Platform.

- CernVM-FS was tested successfully in Centos, Ubuntu and Fedora.



Testbed structure

Simulations of Multi-VO were executed in different Platforms:

- Fedora 12 - Centos 5.5 - Ubuntu 9.04

The test infrastructure was using KVM hypervisor with 4 nodes

(IntelXeon X5355 @ 2.66GHz, 16 GB RAM)

(1) DIRAC admin is in charge of adding the Cloud settings in the DIRAC CS, taking care of the different preconfigured images of the Cloud manager.

(2) Job submission, with the 3 ways to submit the job.

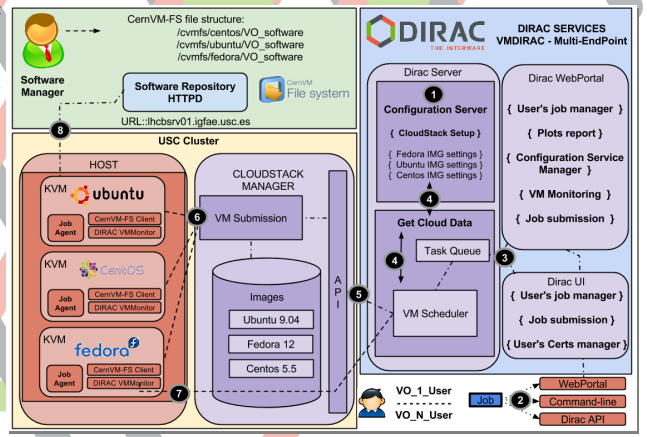
(3) Cloud information is obtained from the DIRAC CS according to the user credentials.

(4) The VM Scheduler component sends the specific EndPoint command to the CloudStack Server API.

(5)The Cloud Manager submits the specific image, which in this case correspond to Ubuntu, Centos and Fedora.

(6) VM Scheduler that is running in the DIRAC get a started message Notification ("Up status") from the Virtual Machine.

(7) CernVM-FS client connects to the USC CernVM-FS repository, which is hosted in USC TIER-2 and provides the software.



V. Fernandez-Albor, [1]M. Seco-Miguelez, [1]T. Fernandez-Pena, [2]R. Graciani-Diaz, [3]V. Mendez-Muñoz, [1]J. Saborido-Silva
[1]Universidade de Santiago de Compostela (USC), A Coruña, Spain
[2]University of Barcelona (UB), Barcelona, Spain
[3]Port d'Informació Científica (PIC), Universitat Autònoma de Barcelona, Spain

CHEP '13 AMSTERDAM

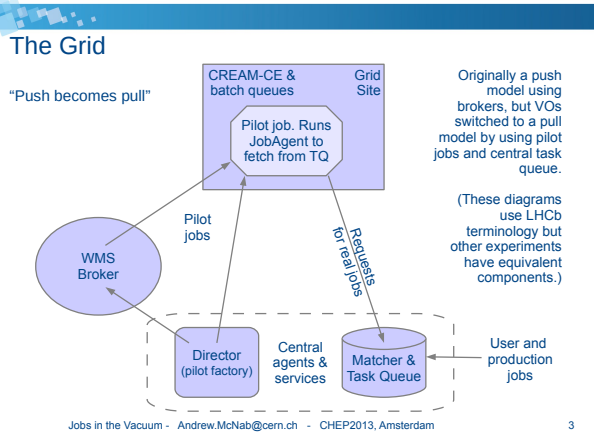
Universitat de Barcelona

PIC port d'informació científica

VAC model

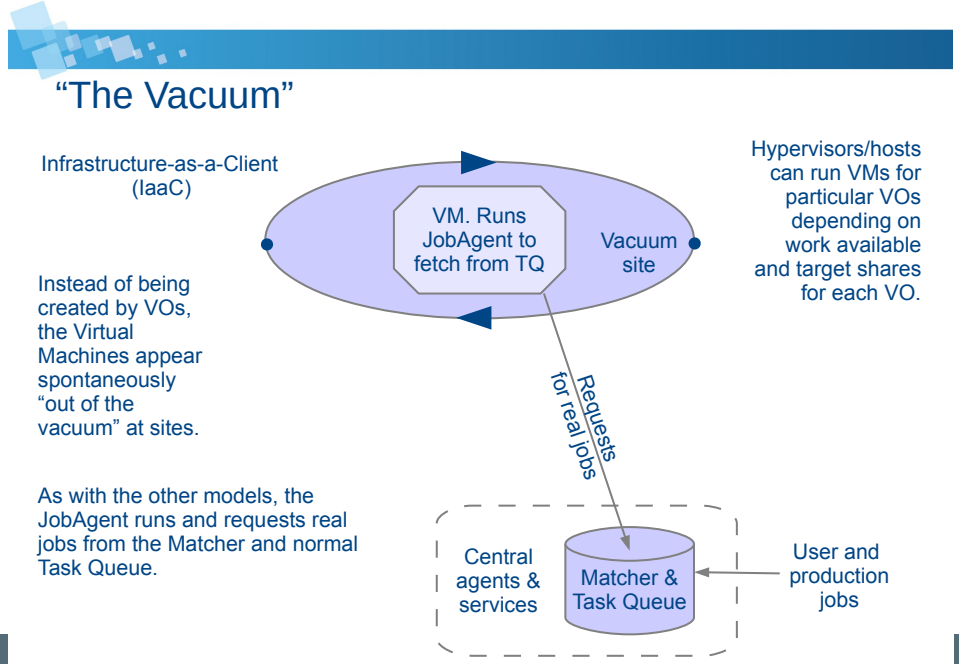
Running jobs in the Vacuum (Andrew McNab)

Move from “classic” grid model to a distributed management of VMs.



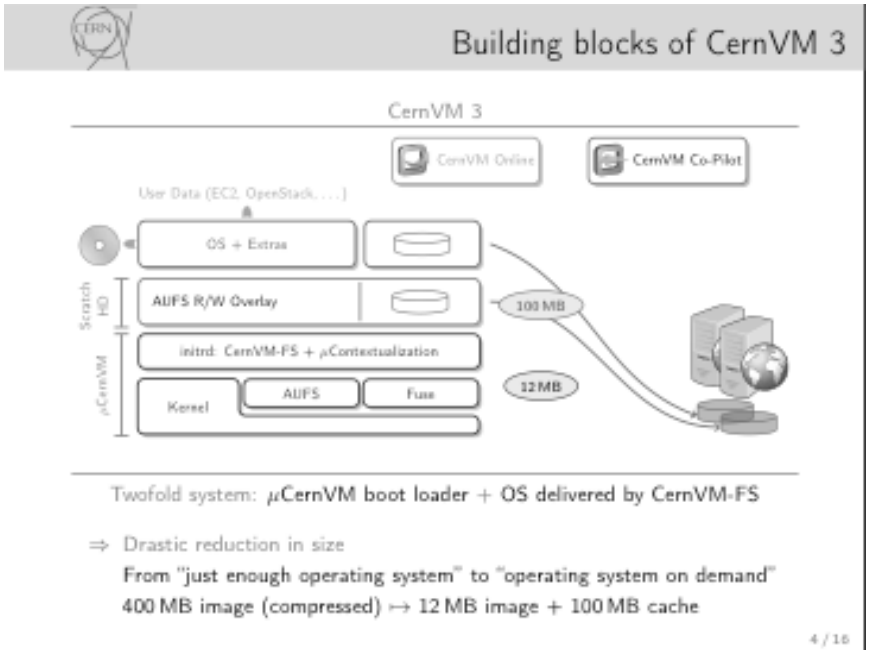
VMs communicate via UDP to each other. Managing up/down scaling, shares, etc.

Running 3k jobs across 3 UK sites with this model, now scaling up

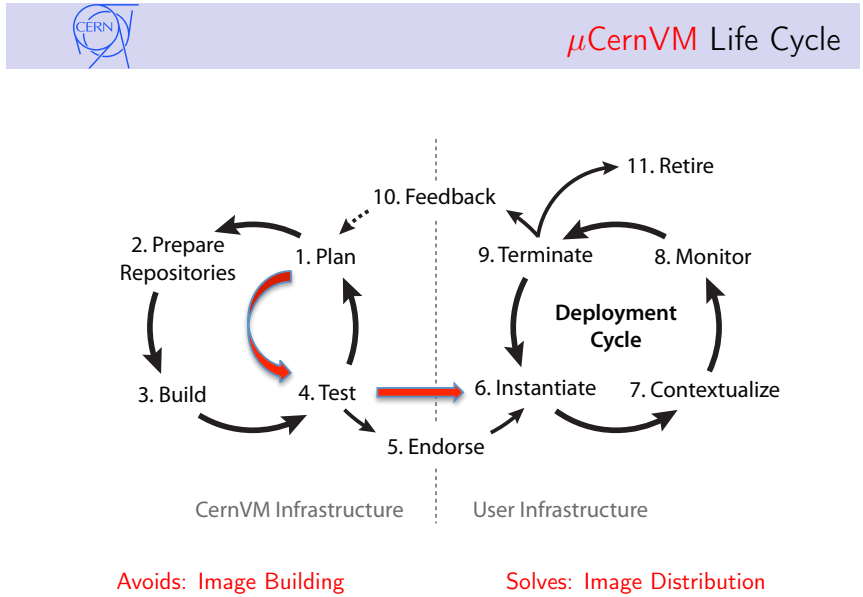


uCernVM

(Jakob Blomer)



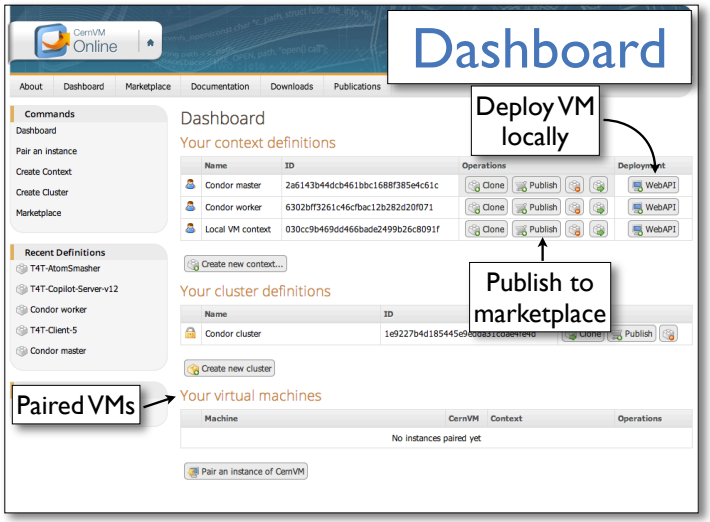
Concept: deploy a ~ 12MB boot loader image with CVMFS client and load the remaining operating system later via CVMFS.



Shall shorten the testing / deployment cycle of newly prepared VMs

Also uses now rpm (was conary)

CernVM ecosystem



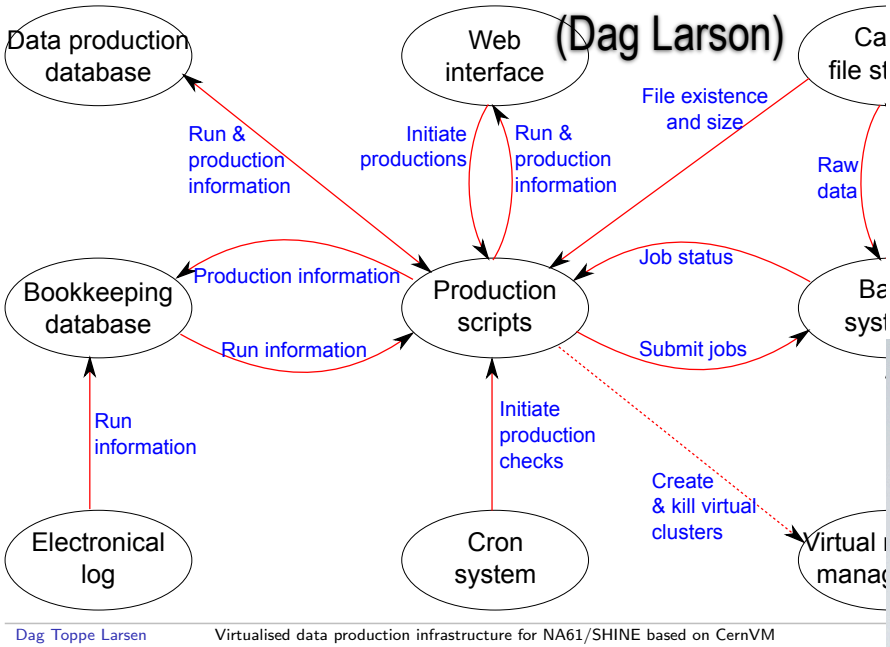
CernVM Online, a place to store and share contexts and deploy local virtual machines

(Georgios Lestaris)

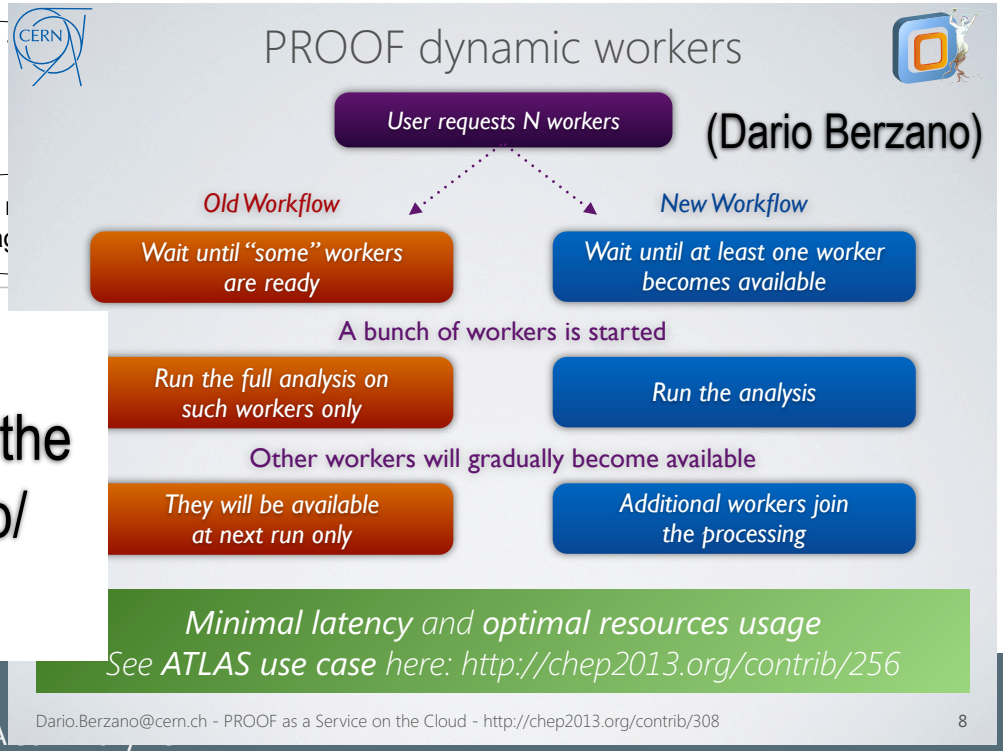
CernVM Cloud, define a virtual cluster and deploy it on infrastructures

Use of CernVM

Overall component interaction overview

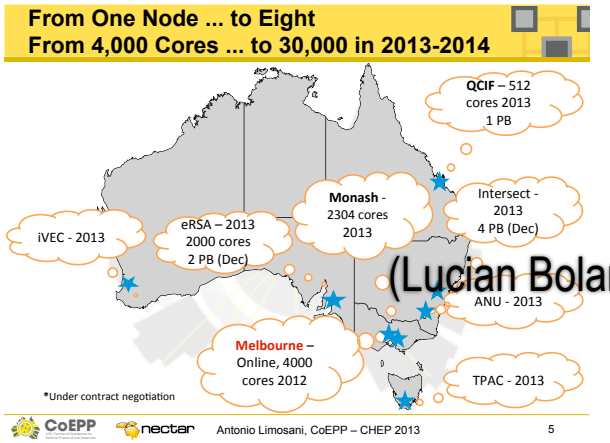


NA61 / shine have created a production system for cloud interaction which ineegrates the CernVM Online/Cloud services.



Proof on GCE, new feature as workers become available will join the processing immediately. Scaling up/ down of HTCondor with "elastiq"

Virtualizing & local site infrastructures



Nectar: Federated openstack cloud infrastructure for Australian Tier2 + Tier3 facilities, ramping up to 30k cores

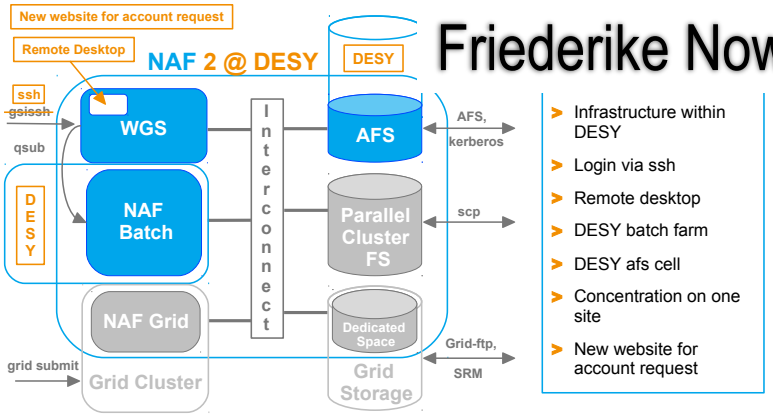
(Lucian Boland)

Our philosophy

KEEP IT SIMPLE	STAY MAINSTREAM	BE USER-ORIENTED
• manageability and flexibility over performance • don't use too many tools • simple images + contextualization (shell-script+puppet)	• use stable and widely-used tools: • **OpenNebula** cloud controller • **GlusterFS** distrib. filesystem • **OpenWRT** for network management	• agile development cycle • provide resources asap • add functionalities when needed

CHEP 2013 - October 14-18, 2013 - Amsterdam S.Vallero 6

NAF 2 Changes (User View)



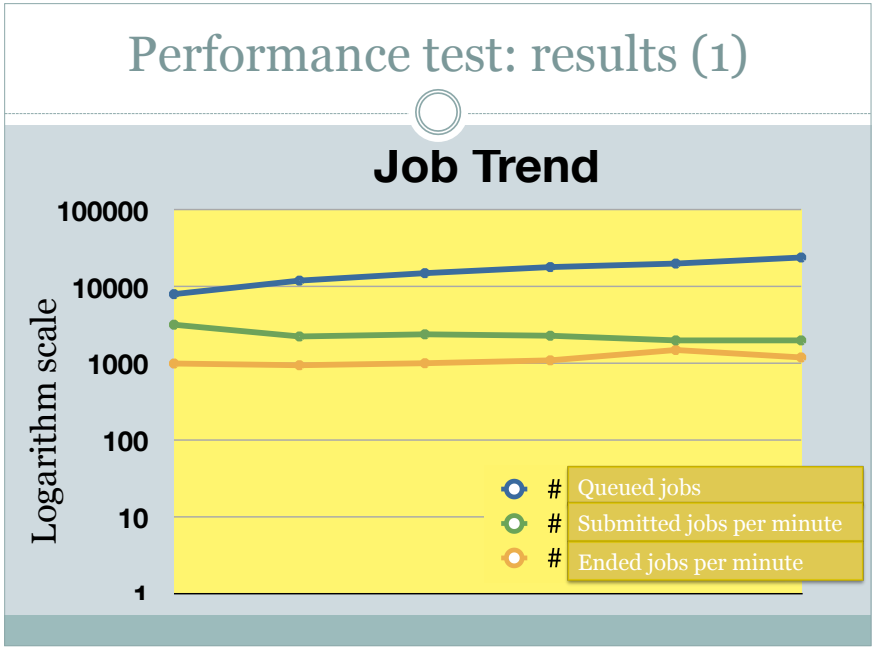
Friederike Nowak

- Infrastructure within DESY
- Login via ssh
- Remote desktop
- DESY batch farm
- DESY afs cell
- Concentration on one site
- New website for account request

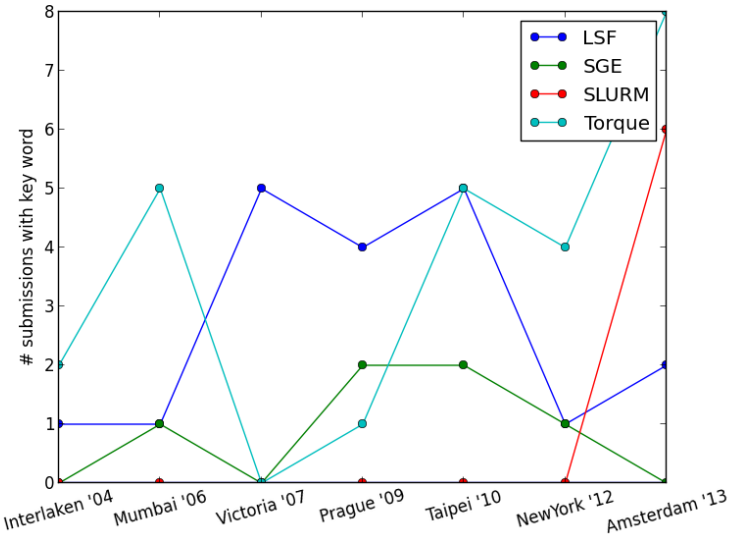
Friederike Nowak | CHEP 2013 | October 14 | Page 16

Private Cloud infrastructure at INFN/ Torino (Sara Vallero), little team relying on standard tools

Batch system



Testing SLURM @ INFN-Bari (Giacinto Donvito), overcome current limitations with queuing jobs and memory footprint of services in torque.



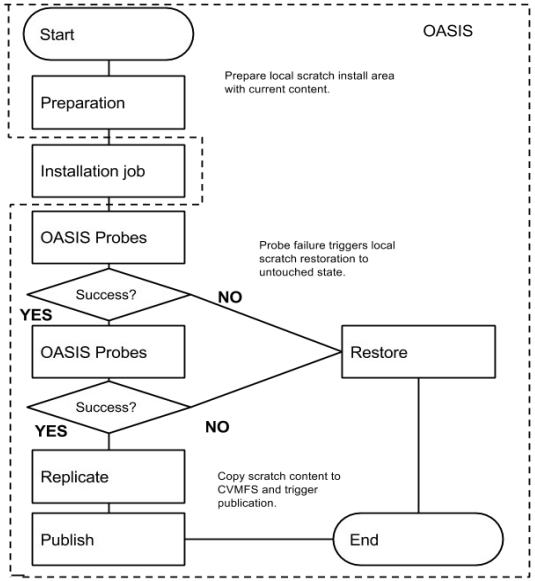


TOOLS

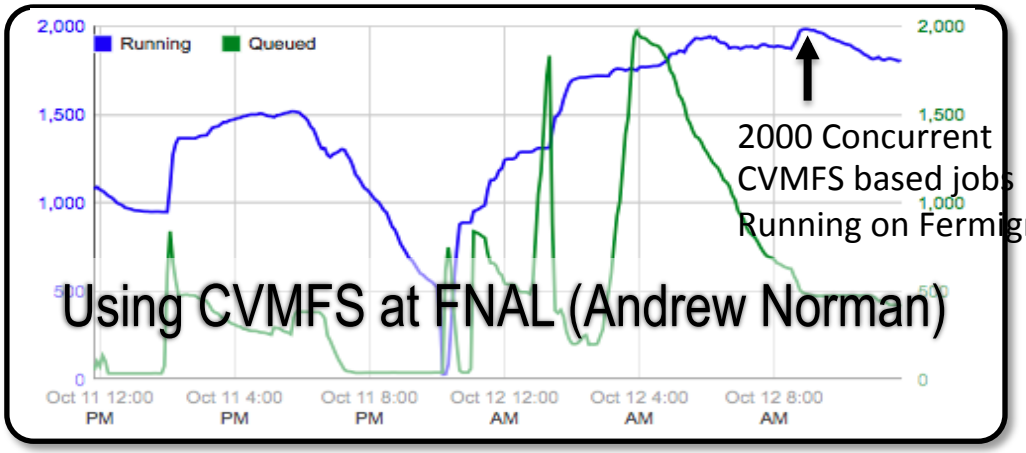
Software Distribution

Oasis: sw distribution for OSG (Jose Caballero Bejar)

Oasis: new service functional flow



Oasis: adding security on top of CVMFS and allow multiple VO hosting



10

Fermilab deployment of CVMFS for both site and OSG

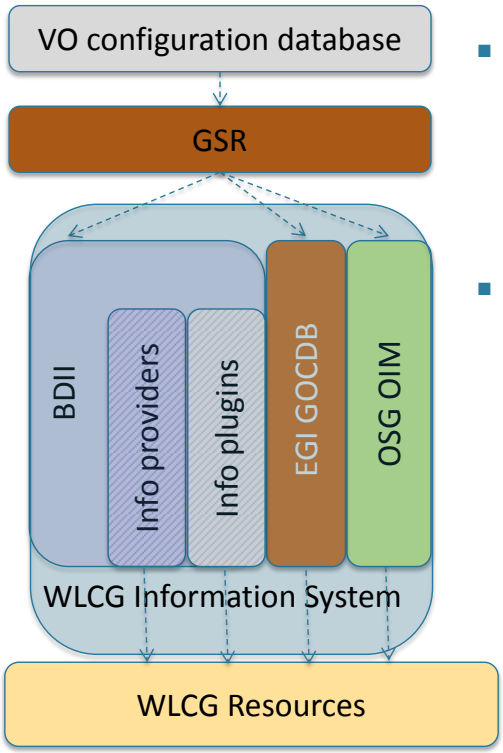
Poster #392 - CernVM-FS - Beyond LHC Computing

Common Tools

WLCG Global Service Registry
(Alessandro Di Girolamo)

Aggregation of multiple "configuration" information sources through a single interface + validation and testing of content provided

Advantages for the LHC VOs

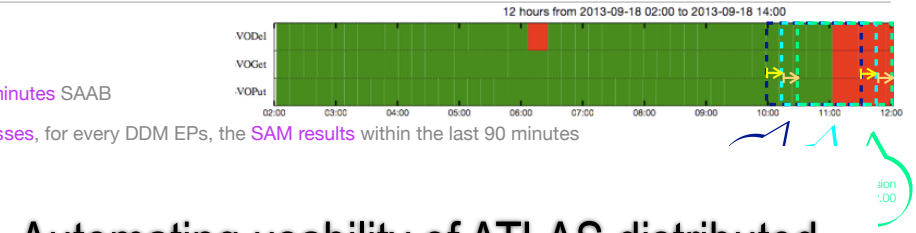


- Unique source of information
  - All VO configuration databases interact with the GSR in the same way
    - Reuse of existing solutions is possible in an easy way
- Full control of the information
  - Enforcing consistent information between registered and actual resources
  - Hiding bugs of the underlying components
    - E.g. Middleware provider bugs in the BDII
  - Ensuring quality by fixing the published information when it is wrong

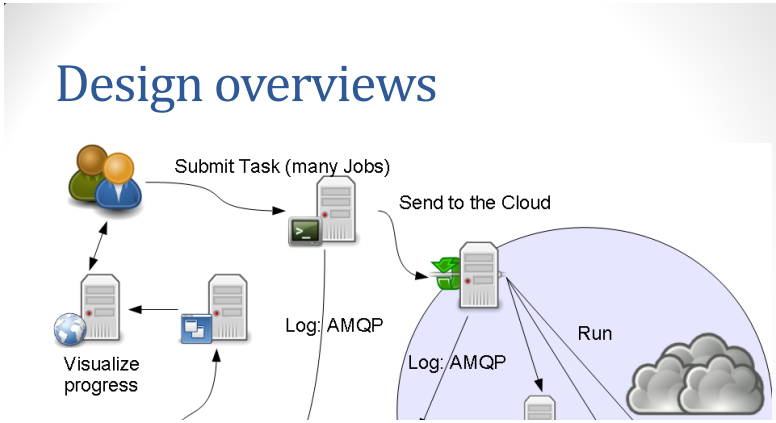
Specific Tools

(… should be common)

… actions with SAM results input



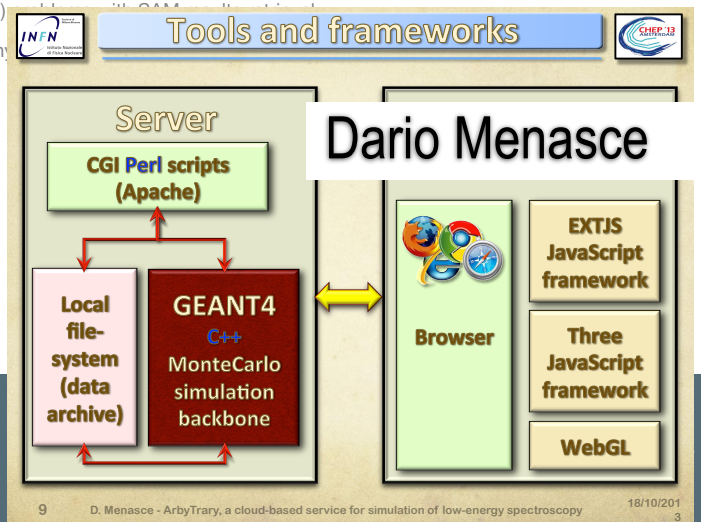
Automating usability of ATLAS distributed computing resources (Salvatore Tupputi)
Extracting information from SAM/Nagios for SEs and taking decision on (un)banning



User centric monitoring in STAR (Jerome Lauret). Multi tier level logging of grid/cloud activities.

CHEP2013, October 14-18 2013, Amsterdam, UCM redesign - Jérôme Lauret for the authors

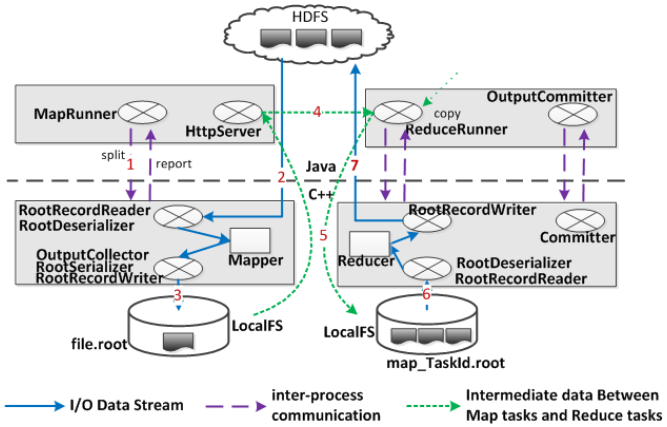
…s mail alerts and notifications:
…elopers in case of (known and tracked)
…s who want to be notified whenever any



Hadoop

HEP MapReduce Procedure

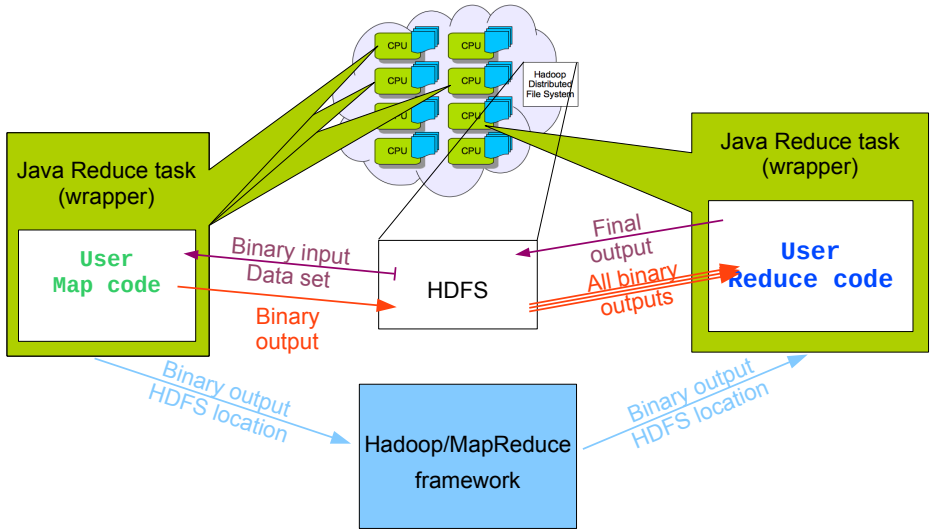
- HEP MapReduce (different from Internet applications):
  - Java side is in charge of job splitting and scheduling
  - C++ side is in charge of I/O and computation



BESIII analysis on Hadoop (Sun Gongxing), wrote ROOT C++ classes to interface to Map/Reduce via libhdfs

Under the hood..

```
# hadoop run RootOnHadoop "user Map code" "user Reduce code" "HDFS input dataset" "HDFS output location"
```

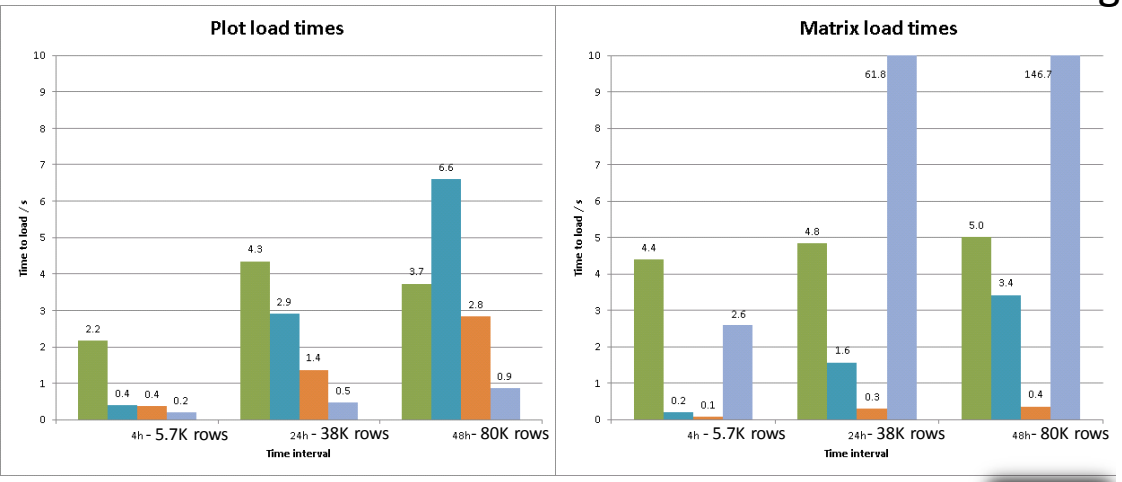


ROOT on Hadoop (Stefano Russo), uses file == chunk on HDFS, wrappers around Map/Reduce -> no ROOT code changes

Oracle / Hadoop / ElasticSearch

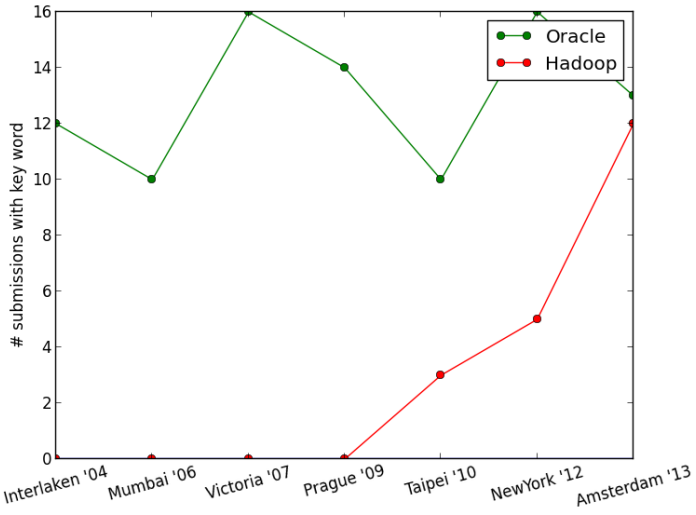
Data Out

WLCG monitoring with NoSQL (Edward Karavakis)



Limitations found for Hadoop/Hbase mainly b/c of "time series" like queries. Evaluated Elasticsearch which outperforms Oracle in several areas.

- ENG is much faster than Oracle for small row counts but won't scale
- EIG is faster than Oracle in all cases but inflexible
- EQG is much faster for few distinct grouping values but won't scale



Two Observations about Abstracts

- Propose also reviewing of abstract texts
  - A good quality abstract will help session conveners in their judgment on acceptance
  - E.g. minimum length, keywords, conveners may contact submitters?
- In case of centralized submissions by experiments ask to check whether a similar topic is submitted twice to different tracks

My (Very Personal) Observations

- Several new developments presented
  - uCernVM, VAC, GSR, ROOT@HADOOP, exploit of opportunistic compute cycles
- Experiments are exploiting new resources for computing successfully
  - At the moment it still takes long time for the setup,
    - maybe we are on a learning curve?
  - Everything that is “non standard” needs a bit of tweaking, either in the technical setup or in the applied workflows.
- We have proven that we can interact with cloud resources, next step is to scale up, but ....
- VOs need a way to “find” cloud resources (BDII?), and ....
- Experiment offline processing teams are growing more and more into the role of “site”-admins
  - E.g. how to monitor these resources efficiently and scaling?
- Developments for/by LHC experiments are/should be re-used by others
  - Good example: CVMFS is *the* tool for software distribution
  - Bad example: between WLCG sites and LHC experiments only little re-usage

Thanks

- To Davide Salomoni for convening Track 3A with me
- To Nurcan, Jeff and Robert for organizing both Tracks 3A + B