Clouds in WLCG

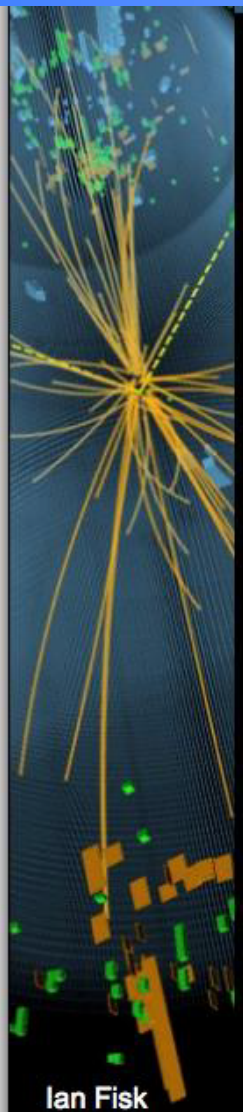
Claudio Grandi (INFN Bologna)

Introduzione

- Fondamentalmente un update della presentazione data al workshop/formazione a Bologna lo scorso febbraio
  - https://agenda.infn.it/getFile.py/access?contribId=3&sessionId=0&resId=0&materialId=slides&confId=5900
- La maggior parte del materiale viene da CHEP13
- Il fuoco per gli esperimenti LHC in questo momento e' sul provisioning delle risorse (WN), come alternativa/complemento a Grid e alle farm locali per il processamento batch e sulla prototipizzazione su farm online e prototipi di siti privati e pubblici
- Il CERN e anche altri siti hanno attivita' anche sul provisioning di istanze singole e servizi

Resource Provisioning

- The resource provisioning layer becomes a potentially common, but complicated interface
  - Goal is to get pilots into a diverse set of computers
    - Dedicated resources, opportunistic computing, local resources, clouds, super computer allocations, etc.



Ian Fisk
FNAL/CD

WLCG Workshop
11-12/11/2013

Pilots

amazon web services

Resource Provisioning

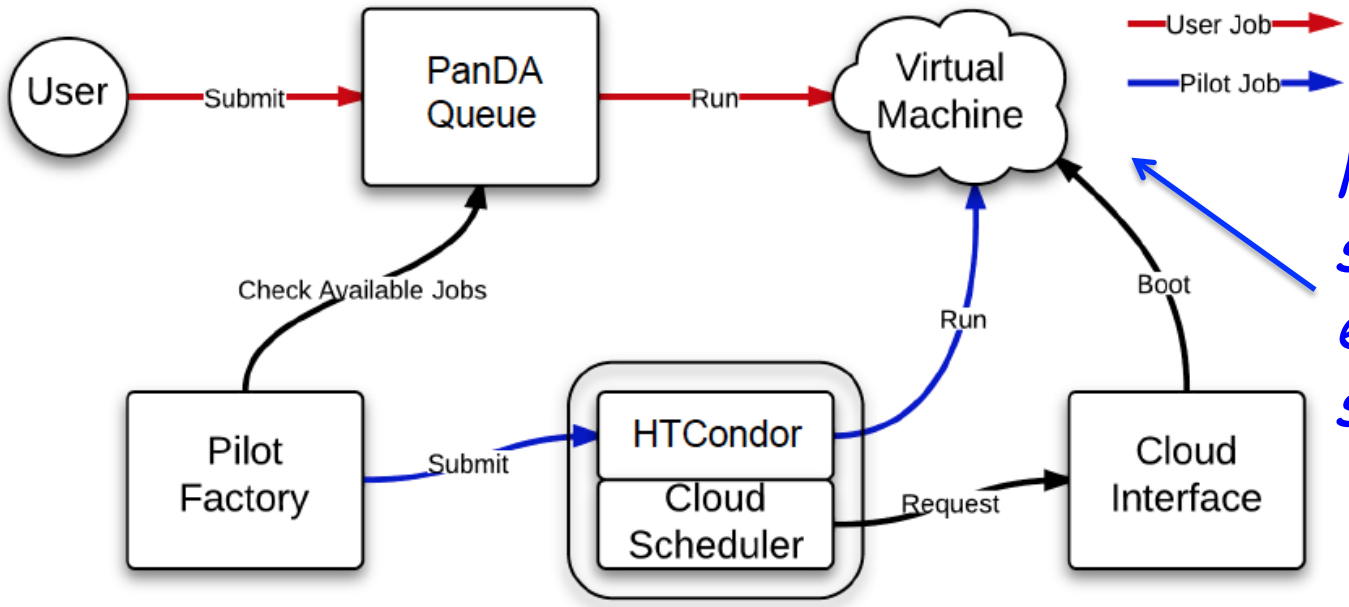
20

Dynamic provisioning

- La creazione dinamica di VM permette al cluster di scalare automaticamente
  - Rendendolo di fatto piu' simile a una infrastruttura Grid
- Problemi da risolvere:
  - Distribuzione delle immagini (per VO? per sito? "golden image"?) e la contestualizzazione
  - Funzionamento in regime saturato
  - …

Integrazione con il framework

Cloud Job Flow



Nei test su AI la VM e' creata staticamente

ATLAS

Sergey Panitkin, CHEP 2013
Amsterdam, NL

6

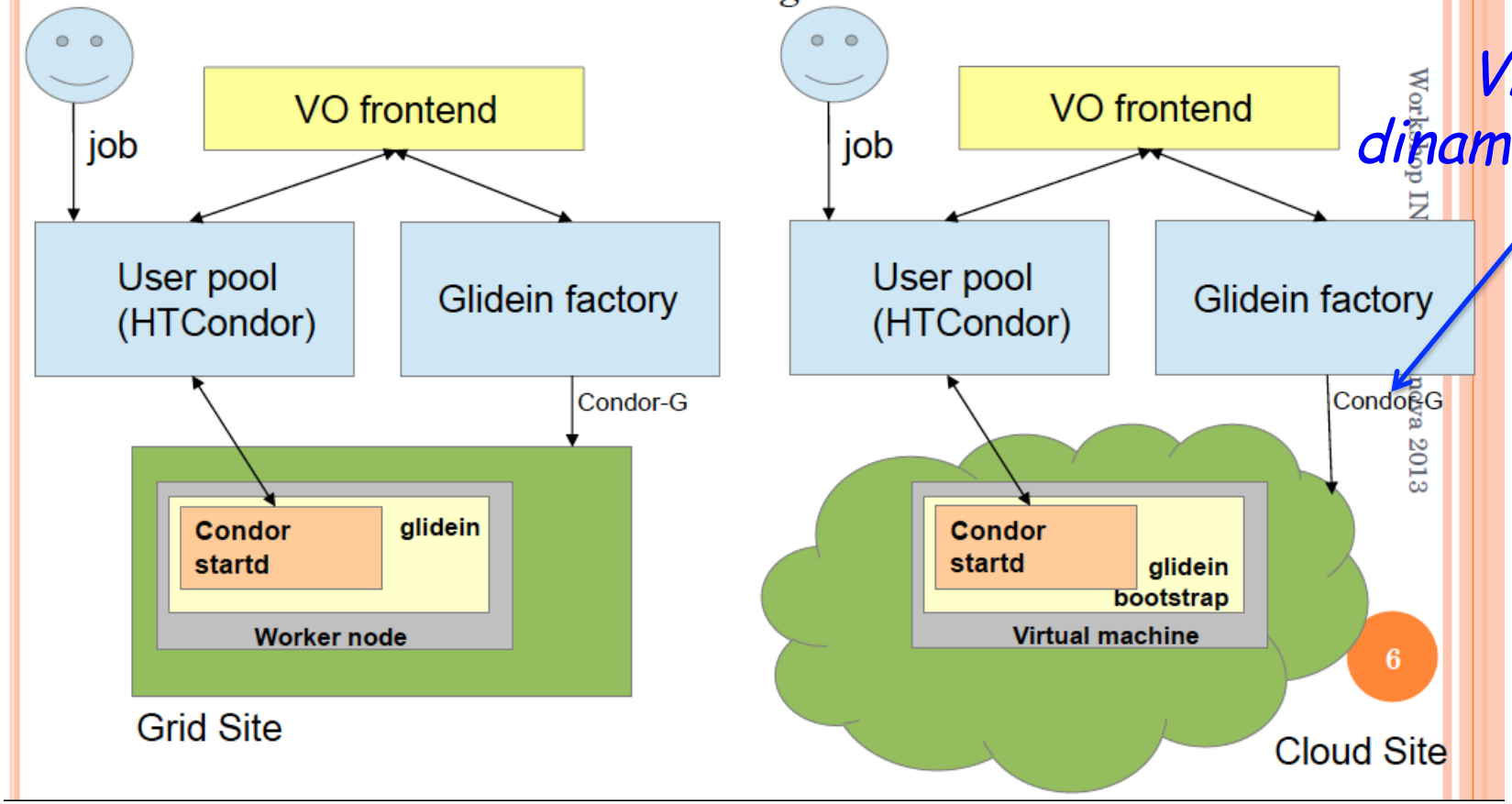
Integrazione con il framework

Dynamic provisioning on Grid & Cloud

- Cloud support introduced with GlideinWMS v3
  - Same architecture: submission of Glideins vs instantation of VMs where glideins run

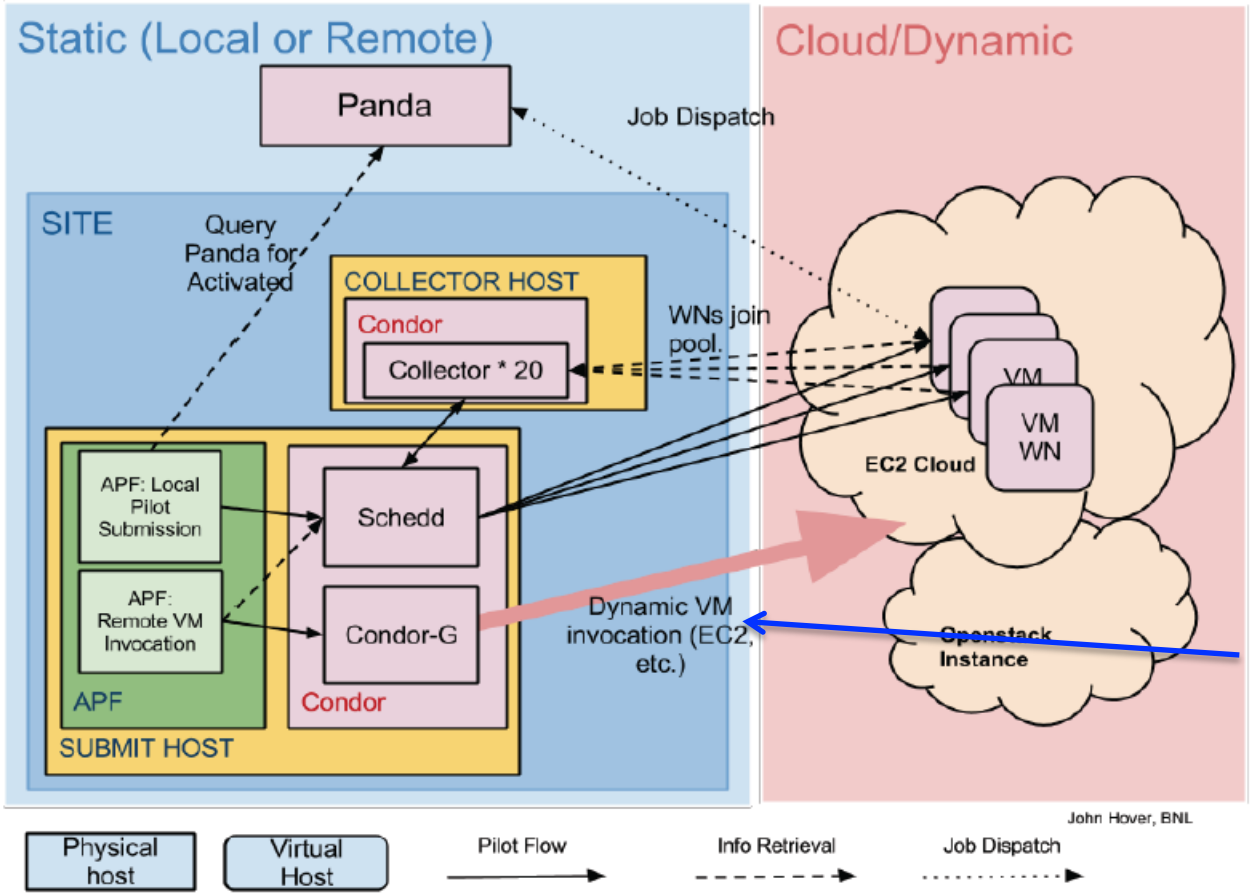
M.Sgaravatto
WS CCR Genova

VM creata dinamicamente



Integrazione con il framework

Hybrid Cloud: BNL OpenStack + EC2



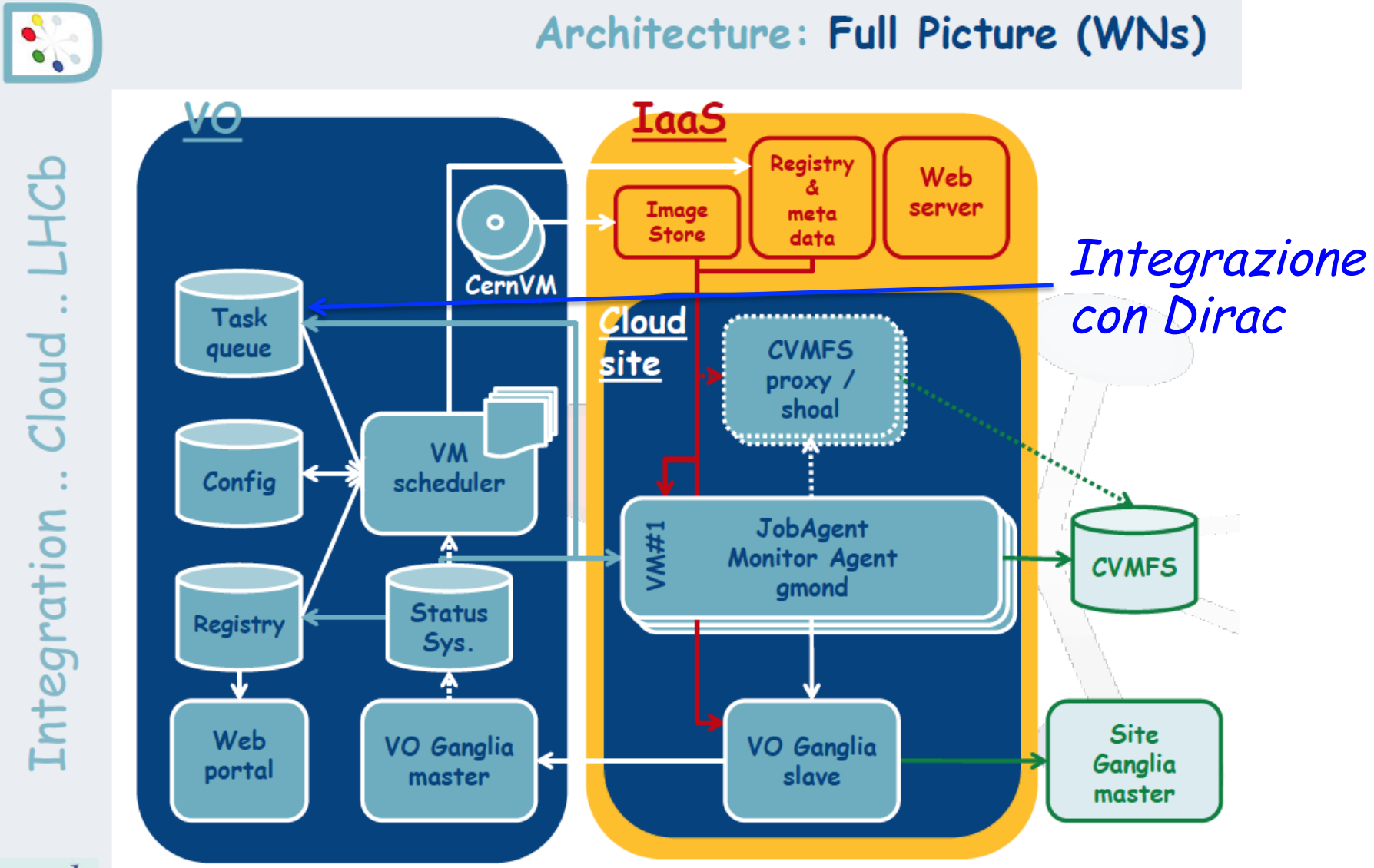
Nei test a BNL la VM e' creata Dinamicamente (evoluzione della APF)

ATLAS

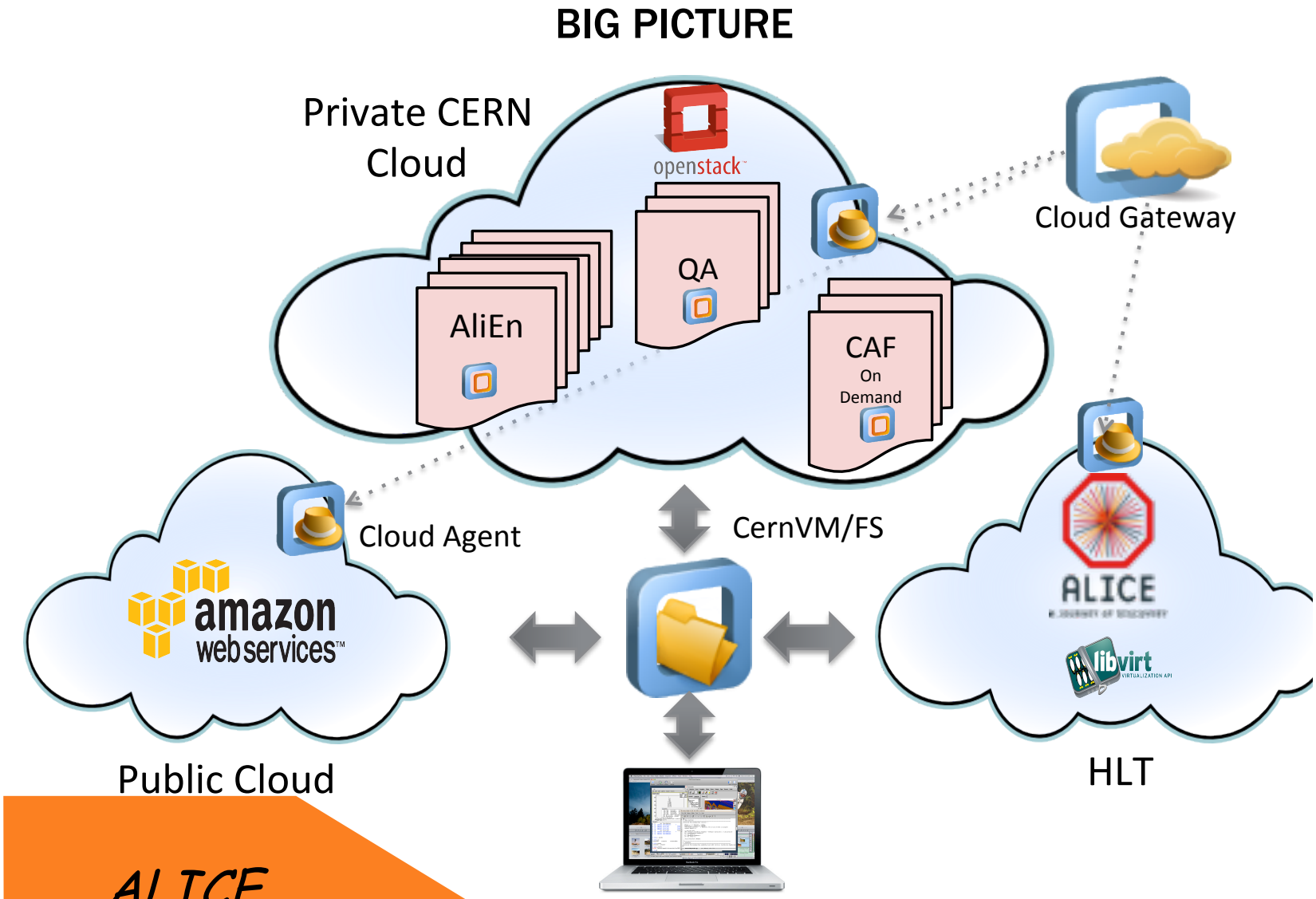
Sergey Panitkin, CHEP 2013
Amsterdam, NL

14

Integrazione con il framework



Integrazione con il framework

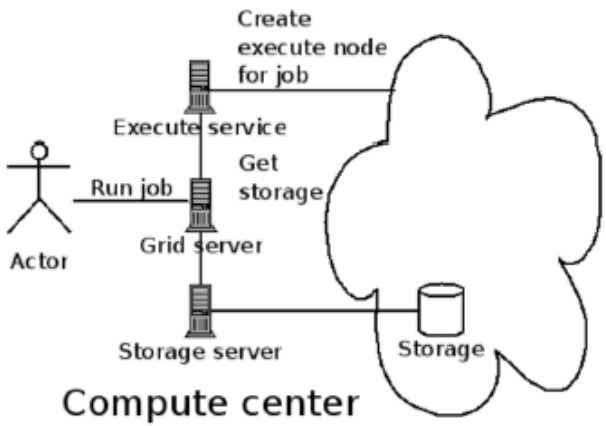


Integrazione con il framework

On-Demand AliEn with OpenStack/EC2

Quick recap

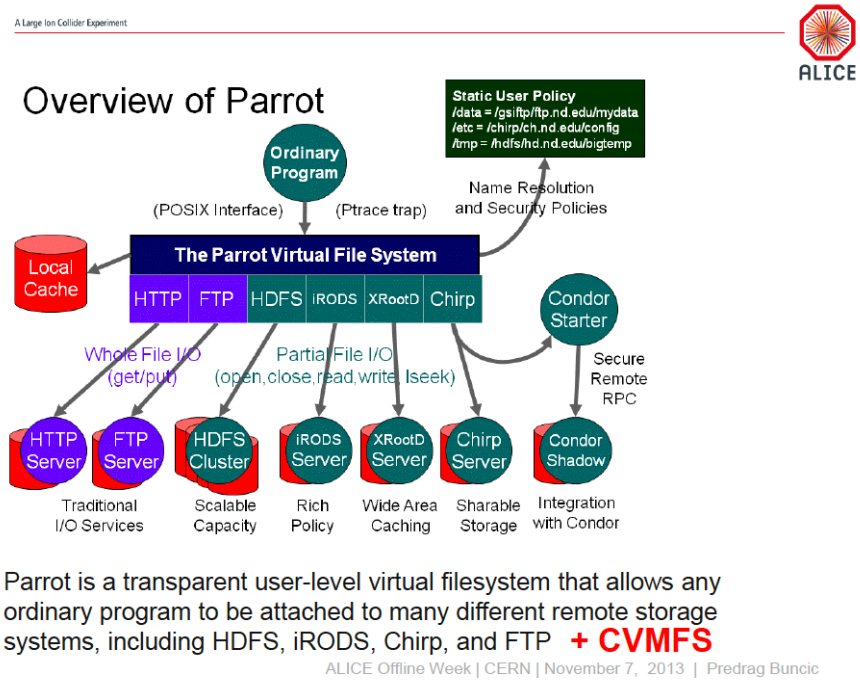
- Use Amazon EC2 API to acquire Grid resources when needed
- Free the resources when done
- OpenStack, CernVM and AliEn



ALICE - Ståle Nestås, Joachim Carlsen

Ambiente

- Per tutti gli esperimenti alcune componenti diventano essenziali per avere immagini simili su siti diversi:
  - Accesso remoto ai dati, ad esempio via xrootd
    - Eliminata la necessita' di complicate configurazioni dipendenti dal sito e di localita' dei dati
    - Il modo per ottimizzare l'accesso locale esiste (tramite la contestualizzazione) ma e' potenzialmente complicato
  - Accesso al software di esperimento tramite CVMFS
    - Anche in user space via Parrot o simili



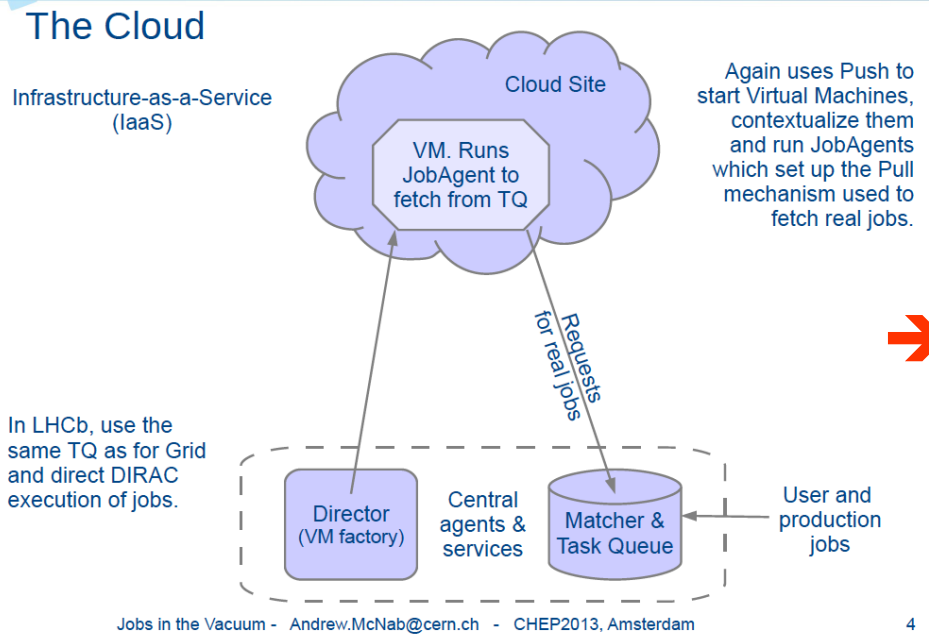
Parrot is a transparent user-level virtual filesystem that allows any ordinary program to be attached to many different remote storage systems, including HDFS, iRODS, Chirp, and FTP **+ CVMFS**

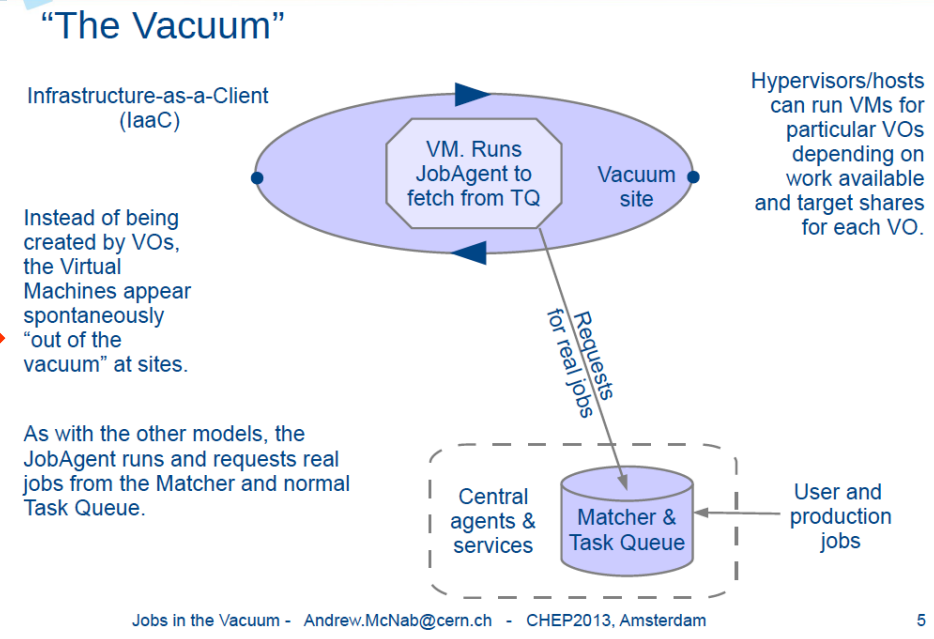
ALICE Offline Week | CERN | November 7, 2013 | Predrag Buncic

Ambiente

- Necessario comunicare al cliente le caratteristiche della macchina allocata (cores, memoria, …)
- Necessario gestire la durata dell'allocazione
  - Che deve essere piu' lunga per compensare le inefficienze di start e stop delle VM
- WLCG *Job/Machine features* working group
  - Uso di files visibili dal pilot job o dalla VM (ad es. sul drive virtuale della VM o dal *magic-IP*) per passare informazioni
  - https://twiki.cern.ch/twiki/bin/view/LCG/MachineJobFeatures

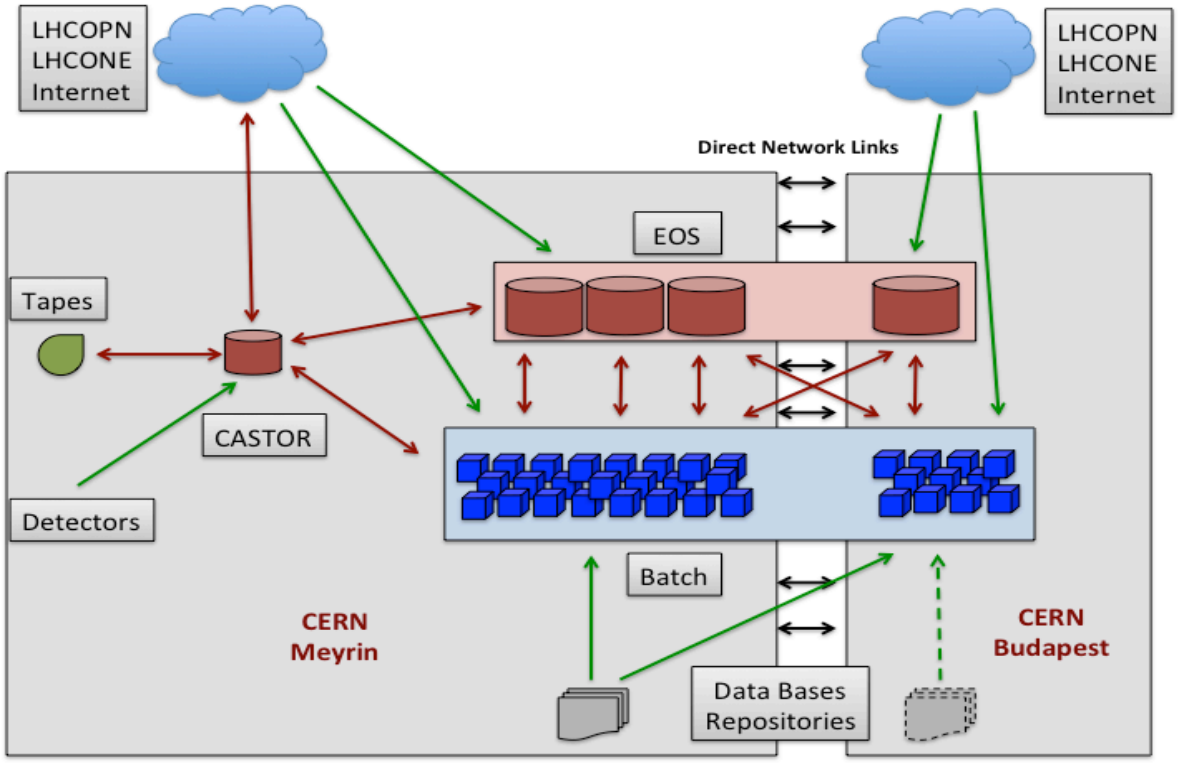
Jobs in the Vacuum (LHCb)





- Eliminata la fase della richiesta di risorse
  - Si assume il pledge dell'esperimento come richeista permanente (*target share*)
- Meccanismi per recuperare le risorse non usate da una VO e ridistribuirle ad altri

Attivita' CERN



- Tier-0 distribuito
  - CERN, Wigner
- Migrazione verso gestione con tools Cloud

⬇

Agile Infrastructure

- Basata su OpenStack (Essex, Folsom, Grizzly)
  - Puppet e Foreman per la gestione
- Previsti 15K HV / 300K VM nel 2015
- Nel 2015 si prevede di fornire le risorse via OpenStack e non piu' via LSF

Testing su CERN AI

Testing summary

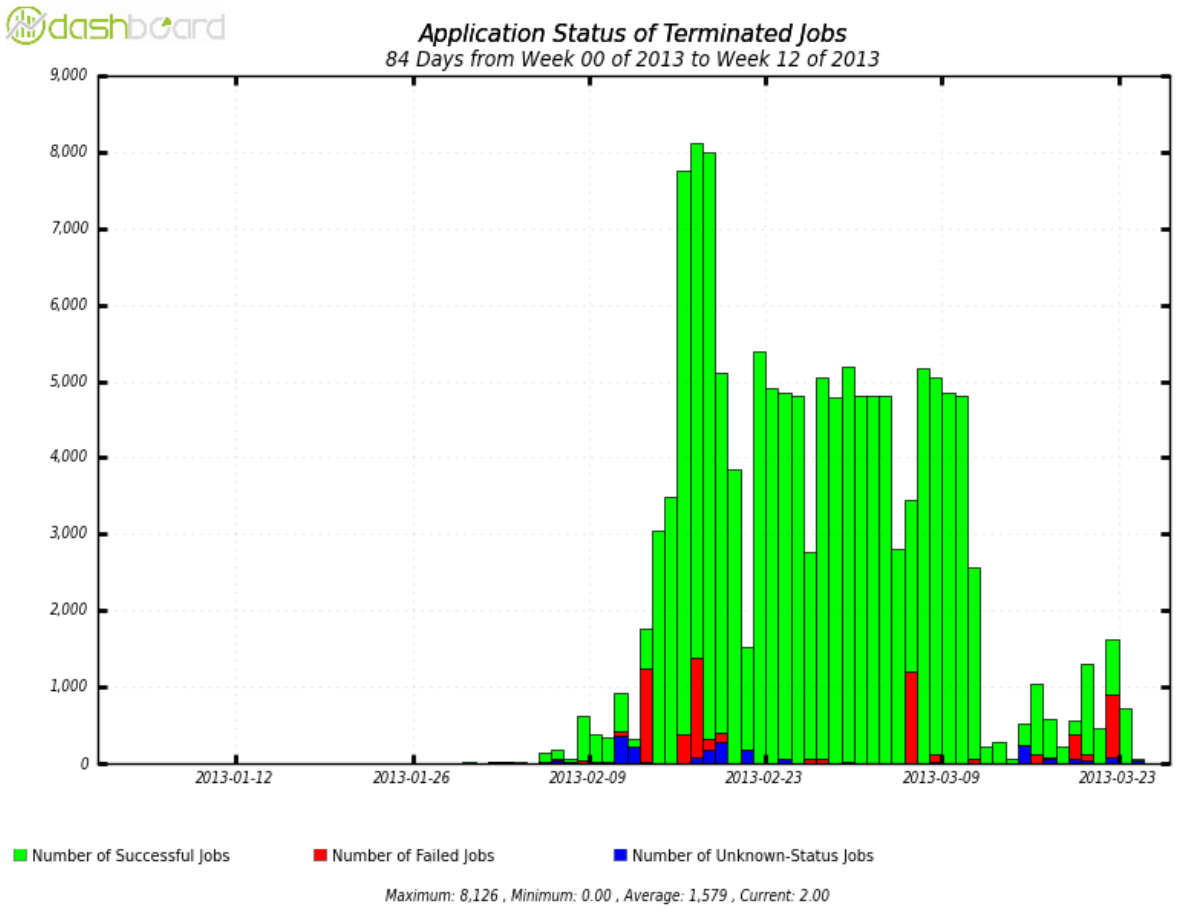
ATLAS	
Scheduler	PanDA
Cluster management	Static
Cluster size	770 cores
Jobs submitted	694,698
Failure rate	9.95%
Job type	Simulation
Typical job duration	31 min.
Duration variance	17.8 min.
Most common error	Failed to read LFC

CMS	
Scheduler	glideinWMS
Cluster management	Dynamic
Cluster size	200 cores
Jobs submitted	337,080
Failure rate	0.31%
Job type	Simulation
Typical job duration	9 min.
Duration variance	4.8 min.
Most common error	App. Error 8020

(file open error)

Testing su CERN AI

Reliability testing



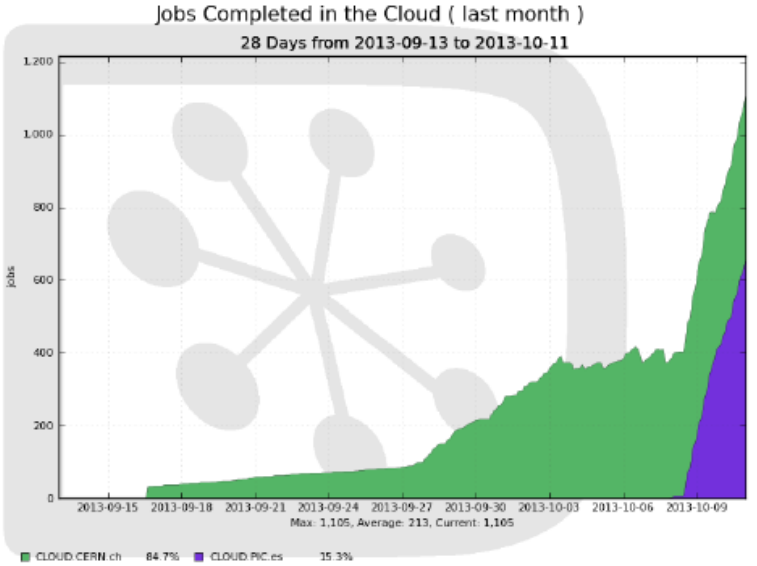
Testing su CERN AI e non solo

Bologna INFN Istituto Nazionale di Fisica Nucleare

Integration .. Cloud .. LHCb

UC-0: Cloud Sites

- No big numbers. In contrast, very solid infrastructure.
- Running on *Production* at:
  - CERN (OpenStack):
    - CLOUD.CERN.ch,
    - CLOUD.CERNMP.ch.
  - PIC (OpenNebula):
    - CLOUD.PIC.es.
  - RAL (StratusLab):
    - VERY SOON !
- Jobs we run:
  - MC,
  - Data processing.



Cloud Site	Count	RAM *	Disk *	VCPUs
CLOUD.CERN.ch	26	4	40	2
CLOUD.CERNMP.ch	4	8	80	4
CLOUD.PIC.es	75	2	10	1
ΣTotal	105	286	2.11K	143

* (GB)

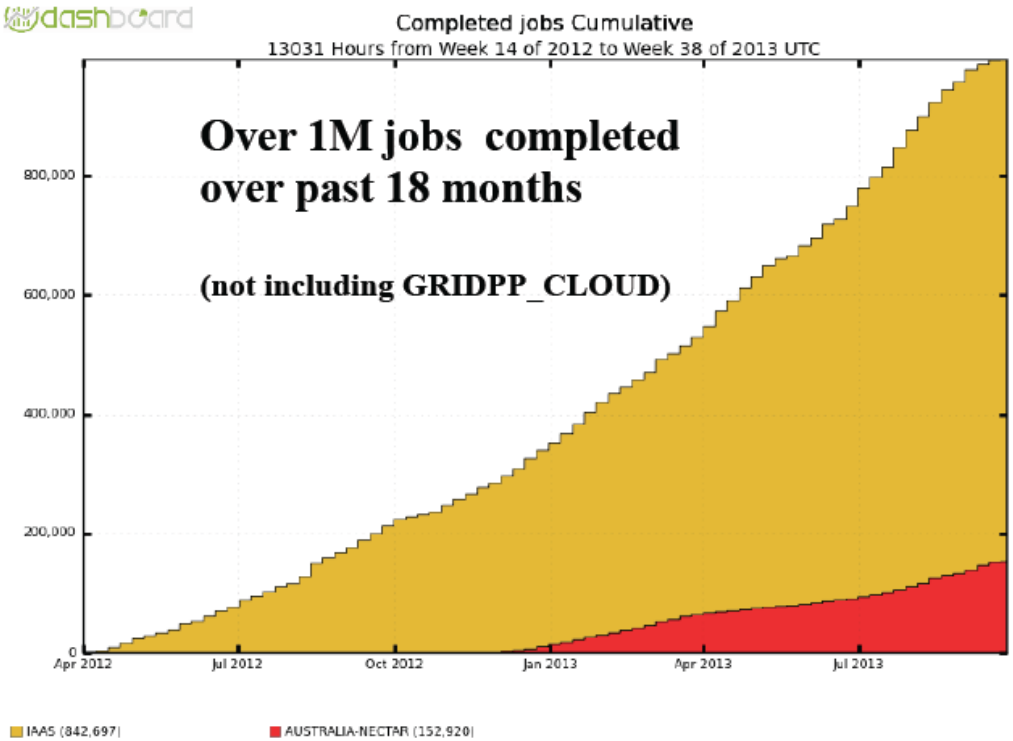
Ancora testing

Production on IaaS Cloud

preliminary tests Oct. 2011

standard operation Apr. 2012

- GRIDPP_CLOUD: added May 2013
- CA-JADE: testing now



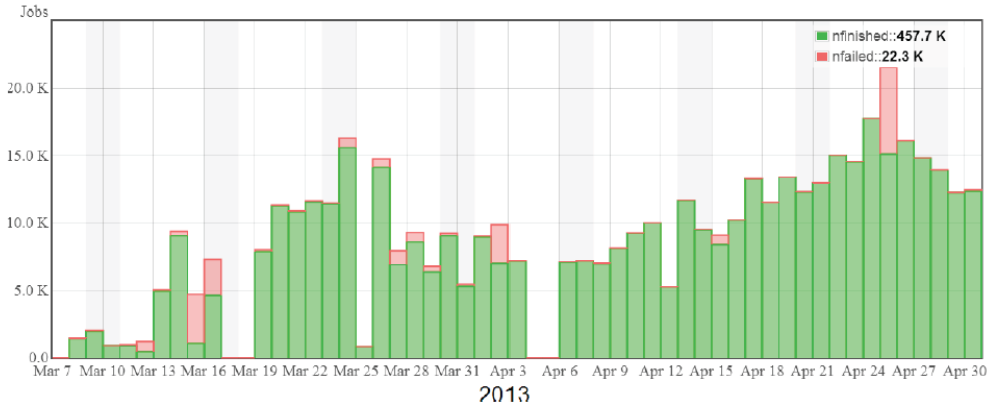
ATLAS

Sergey Panitkin, CHEP 2013
Amsterdam, NL

9

Google Compute Engine

Failed and Finished Jobs on GCE



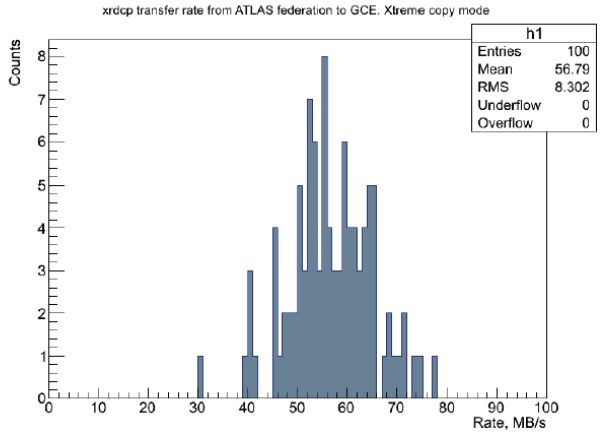
- Most of the job failures occurred during start up and scale up phase – as expected
- Most of the failures were on the ATLAS side – file transfer, LFC problems, HTCondor
- No failures were due to GCE problems

Sergey Panitkin, CHEP 2013
Amsterdam, NL

18

- Test di ATLAS su Google Compute Engine
- 4K cores per due mesi

Data transfers from FAX to GCE



- Data transfer from Federated ATLAS Xroot to GCE in multisource/multi-stream mode
- Xroot cluster on GCE using ephemeral storage with 1.7 TB volumes per node
- Average transfer rate: 57 MB/s (single source xrdcp rate 40 MB/s)
- Note, this is over public networks

Sergey Panitkin, CHEP 2013
Amsterdam, NL

19

- Jobs compute-intensive
- Infrastruttura statica
- Alta efficienza
- Remote access

Farms online

ATLAS

TDAQ HLT Point1 Infrastructure

Compute nodes

#	Type	CPU Cores: non-HT / HT	Memory	Local disk
341	Dell PE 1950	8 / 8 (*)	16 GB	80 GB
320	Dell PE 6100	8 / 16	24 GB	250 GB
832	Dell PE 6100	12 / 24	24 GB	250 GB
1493	Total	15272 / 27816	33 TB	315 TB

(*) No HT support

17.1k used by Sim@P1 already

Network

Type	Current status
P1 ↔ Castor	70 Gbps (20 Gbps reserved for Sim@P1)

IT-SDC 24 September 2013

- Installato OpenStack - risorse rese disponibili quando possibile
- Disaccoppiamento con l'ambiente online

- Risorse ingenti (paragonabili a quelle dei Tier-1)
- Disponibili fuori dalla presa dati

CMS

Imperial College London

HLT in terms of resources

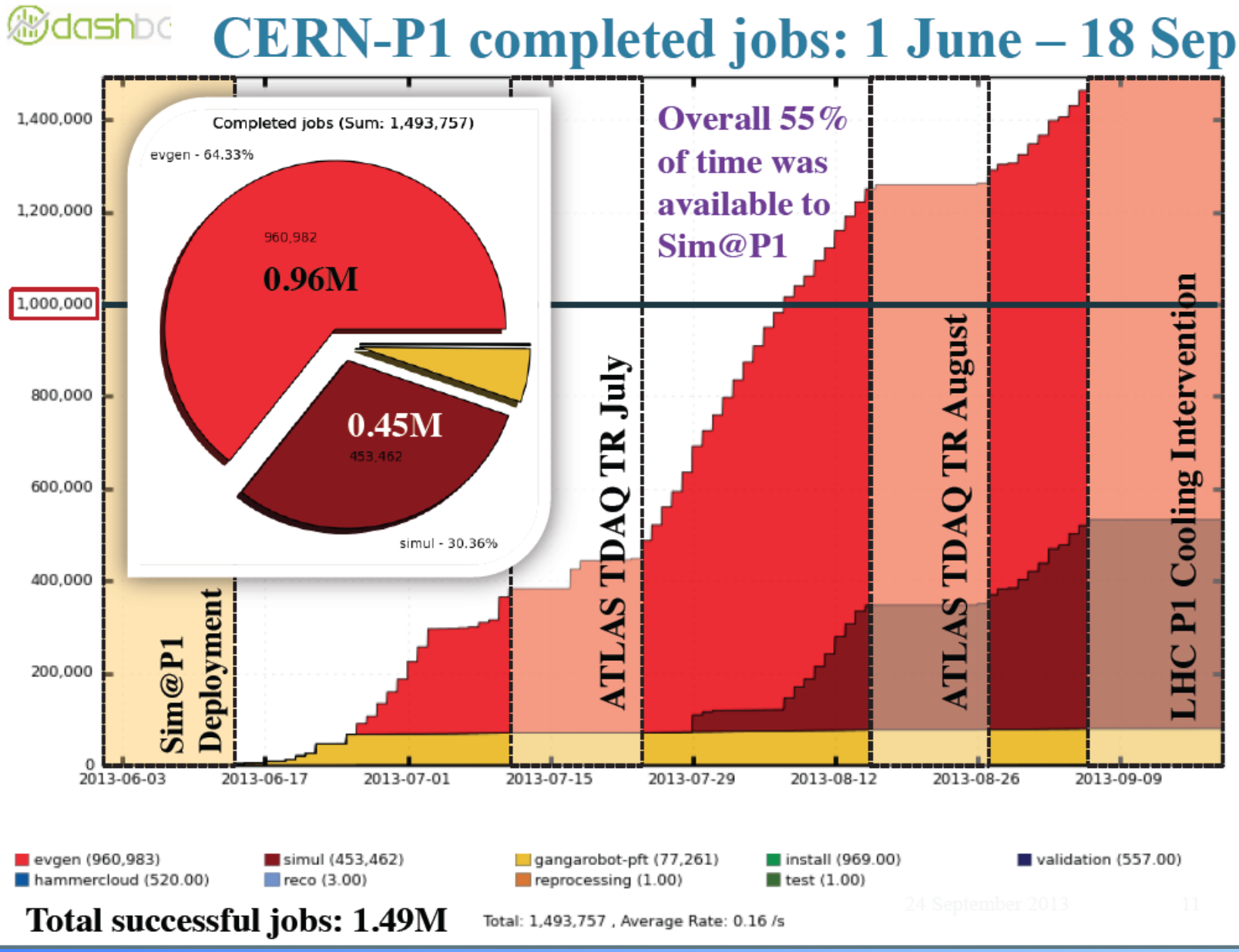
At the start of 2013 the HLT consisted of:

cluster	Nodes	cores (HT on)/ node	cores	Memory (Gbyte/node)	Disk (Gbytes/node)
(1)	720	8	5760	16	72
(2)	288	12 (24)	3456	24	225
(3)	256	16 (32)	4096	32	451
(1)+(2)+(3)	**1264**		**13312**	**26 Tbytes**	**227 Tbytes**

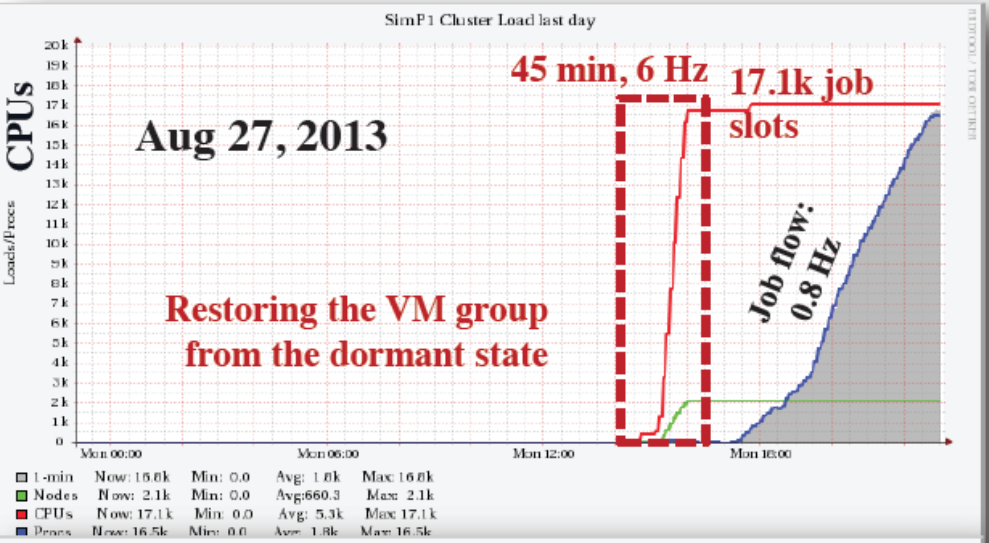
- Three generations of hardware
- Variety of network links to CERN and outside world (1Gb/ and 20Gb/s)
- (Essentially) no storage
- Wholly owned by CMS
- Single use cluster – did one particular job very well indeed, not obviously suited to be a general purpose cluster

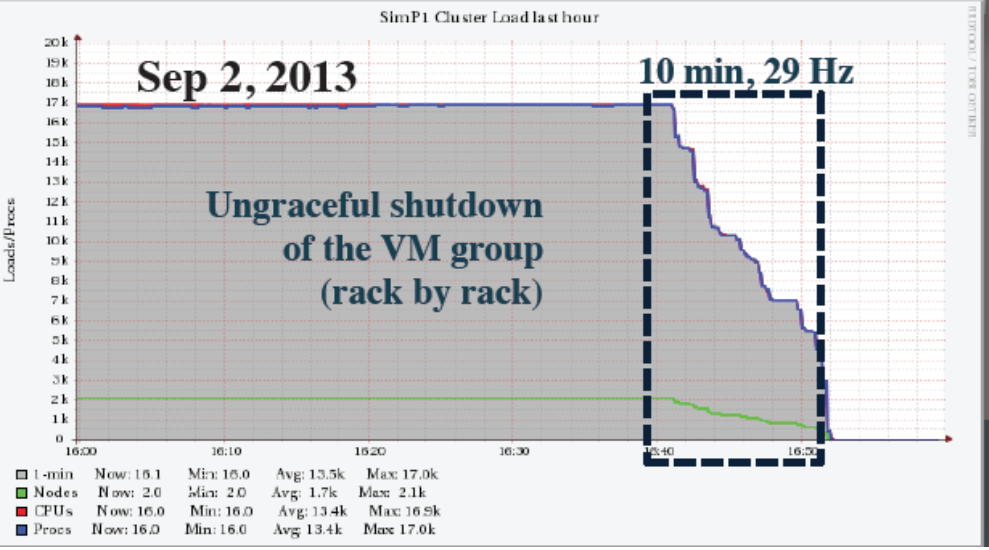
15/10/2013 5

Farms online - ATLAS



Sim@P1: Start & Stop



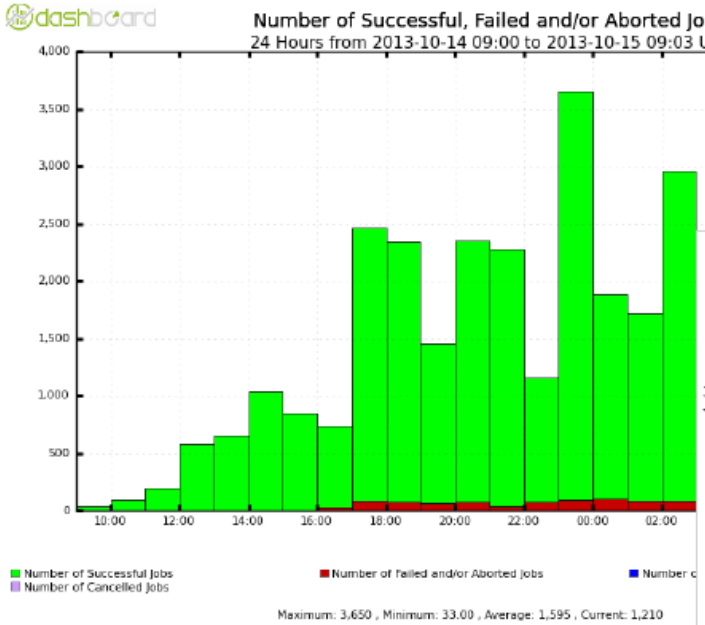


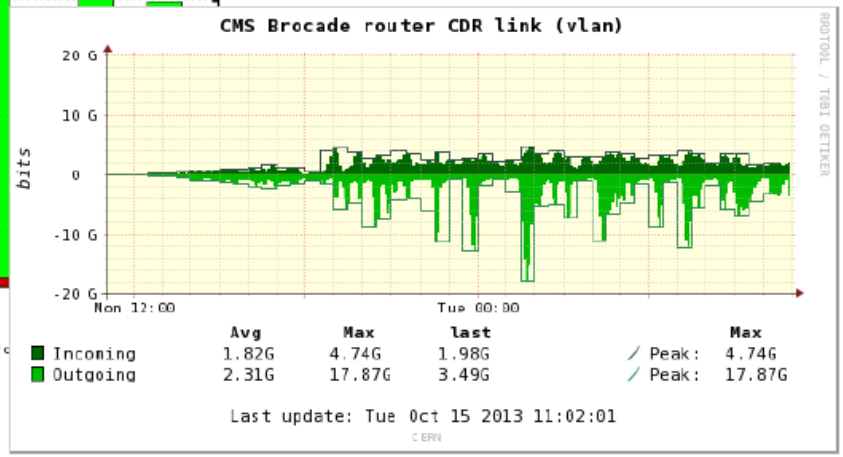
- Restoring Sim@P1:
  - VMs all up and running within 45min (6Hz)
  - MC Production jobs flow 0.8 Hz
    - ✓ now improved to almost 1.5 Hz
- Shutdown:
  - 10min (29Hz) the infrastructure is back to TDAQ

Farms online - CMS

Current Status

Running production workflow as final test before restarting real work on the HLT cloud …





Attivita' future

- Finalizzazione dell'integrazione coi sistemi attuali basati su Grid (che non scompare)
- Federazioni di Cloud – test di nuove infrastrutture
- Utilizzo di risorse sature (shares and all that…)
- Accounting, monitoring
- Comunicazione provider -> client
- *Service discovery* e servizi "location aware"
- Integrazione con risorse opportunistiche e volunteer computing