ATLAS sur le cloud du CC

Workshop LCG-France/France Grilles,
CCin2p3(Lyon), 26-28 novembre 2013

Vamvakopoulos Emmanouil

dapnia
cea
saclay

CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

Outline

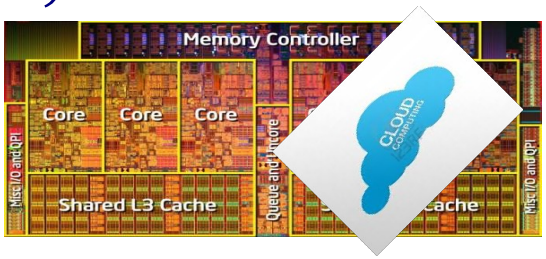
- *Motivation*
- *Current Status : Atlas point of view*
- *Integration of Cloud Resources in the GRID*
- *Preliminary Base Lines*
- *Further plans*

Motivation

- Recent years, Cloud Computing become mature and popular
  - Commercial, public and private Clouds base on different platform (openstack, stratuslab, open-neboula, ...etc)
- Benefits from Cloud infrastructures
  - Industrial Standard of Computer Infrastructure Management (CERN remote-T0)
  - Homogeneous Provisioning of the Resources
    - CPU cycles, Storage, network
  - Decoupling the OS specific binding from applications
  - Reduce (or translate) the operational/management cost
- European effort to develop homogeneous computer infrastructure for e-science (EGI Federation, Helix Neboula, ...etc)
- Business model

Atlas usage cases

- *CPU intensive jobs can easily dispatch to Virtual-Worker Nodes on a IAAS Cloud (Private, Public Commercial)*
  - *Physics Event Generators*
  - *Fast/Full detector simulation (MC)*



- *Ideas for ANALYSIS*
  - *Large scale PROOT cluster O(1000)*
  - *Ephemeral XROOTD storage (cache)*

Google I/O 2013 - Cloud Computing and High-Energy Particle Physics: ATLAS Experiment at CERN & GCE :
Andrew Hanushevsy, Garrick Evans, Sergey Panitkin
http://www.youtube.com/watch?v=LRkLQw5rLy8

WLCG Storage

Attention the Grid/WLCG Storage is still here

- There are "barriers" in order to perform heavy-IO analysis in the "Cloud" (commercial):
  - Nature of the infrastructure (Visualization layer increases the local I/O overheads ~ ?)
  - Extra cost ($) to copy in and out the data on Commercial Cloud environment
  - No dedicated network paths between various Cloud Providers and T1/T2 WLCG/Grid sites (LHONE-LHCOPN).

Cloud of Grids or Grid of Cloud?

The GRID component which cross the Cloud boundary is the WorkerNode

from Bare-metal WN in dedicate sites to Virtual WNs in cloud infrastructure (Commercial, Public, Private Clouds,etc)

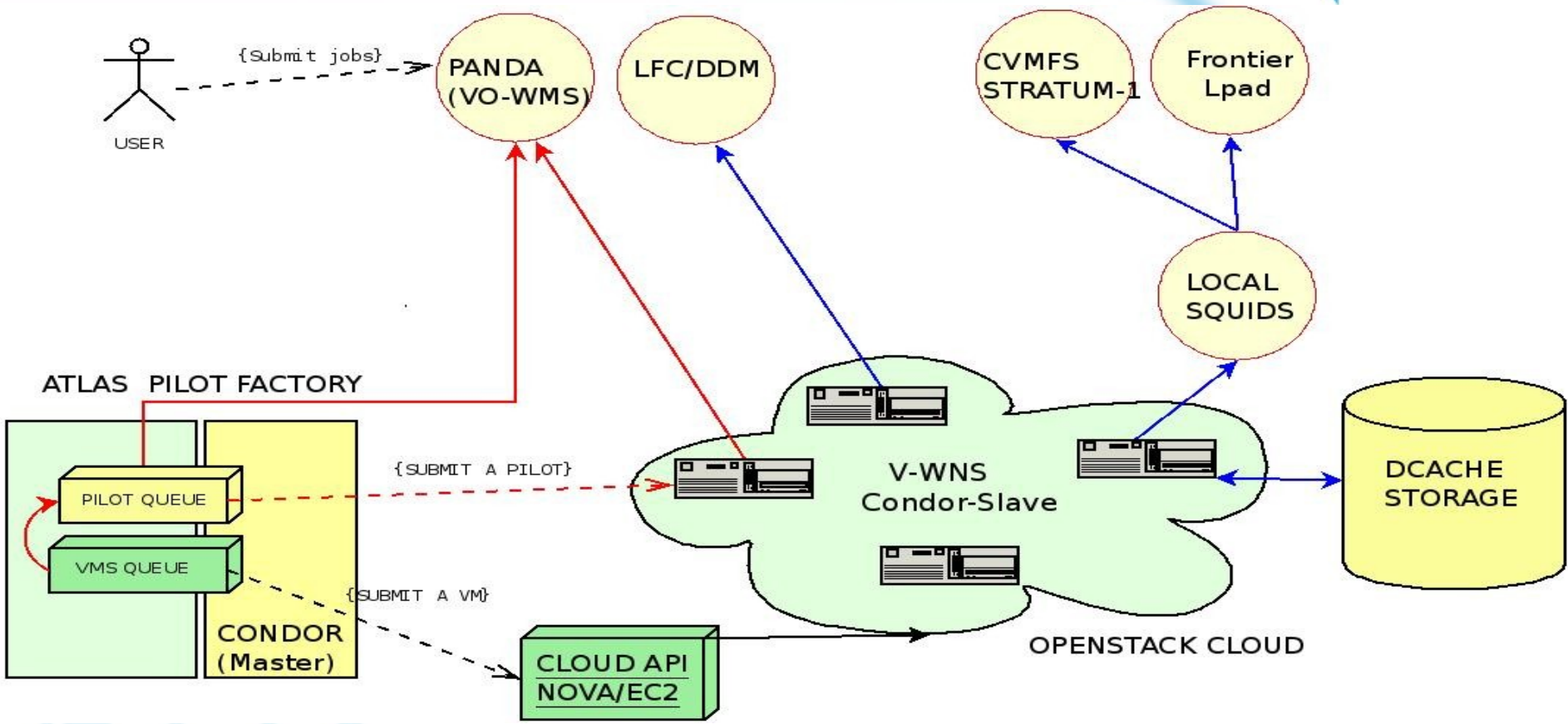
Current Status

- *Atlas HLT farm as a private cloud ~1500 machines ~17K cpu (also for CMS)*
  - *Overlay Opportunistic Clouds in **CMS/ATLAS** at CERN: The CMSoooooCloud in Detail "By: Jose Antonio Coarasa Perez, CERN OpenStack Summit Presentations, Porland 2013, Link*
- *CERN Open-Stack farm (Agile infrastructure)*
- *Queue-boost from commercial clouds (e.g. Amazon, Google, Rackspace, Atos,..etc)*
- *Ordinary WLCG sites with "Cloud" CPU power e.g.*
  - *Australian ATLAS T3 on "Nectar" CLOUD*
  - *IAAS on CA cloud*

IN2P3-CC_OPENSTACK

- Local private cloud infrastructure HTC
- Base on Openstack & qemu-kvm
- 2-3 Project is supported
- ~64 HyperVisors x PE 1950 (E5450@3.00GHz)/16MB + 160GB HD
- Nova api & EC2 api
- 1Gbit/sec nat outbound network connectivity
- 20Gbit/sec per rack uplink connectivity with CC backbone network
- Firewall-ed control connectivity with the vast of majority of the service of CC (two levels)

CONTEXT DIAGRAM



- Job manager (pilot) kept untouched
- Development extensions in APF permits the dynamic provisioning of the resources (Vms)

HTCondor

- Native support of X509 grid proxy delivery mechanism.
- Condor Slaves can communicates with Master scheduler over Firewall/nat (CCB).
- Condor master/ slaves communications are encrypted
- Supports of submission of Virtual machines via EC2 api.
- Decent scalability.

Image for a "Virtual-WN"

- The image should be prepare with all necessity rpms (must avoid update/installation on the FLY!)
- Image should contain some valid default parameters
- We need Uniform Contextualization mechanism (Credentials & tuning) for cloud federations(e.g. puppet)
- We need a standard framework to create images (e.g boxgrinder, oz, image-factory,..,etc).
- CernVM v 3.0 in beta testing
- Monitoring ?

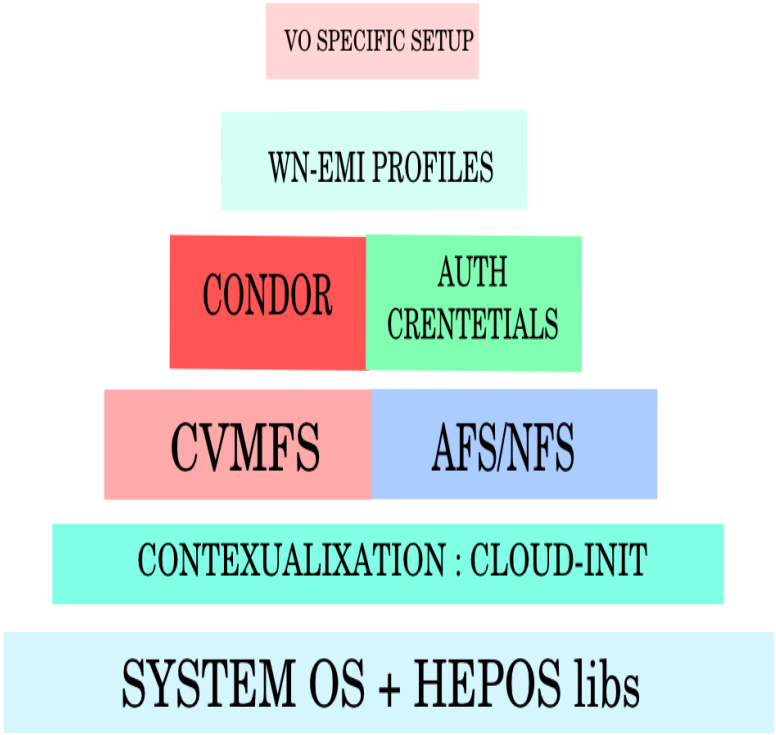


Image Flavors (size) ?

- VO recommendation
  - 2GB RAM per jobs slot (no-swap)
  - 20GB scratch space
  - ? GB for CVMFS cache
- Local openstack
  - 2 GB RAM per 1 VCPU (no-swap)
  - 5GB scratch space per job
  - 15GB (20GB) for CVMFS cache

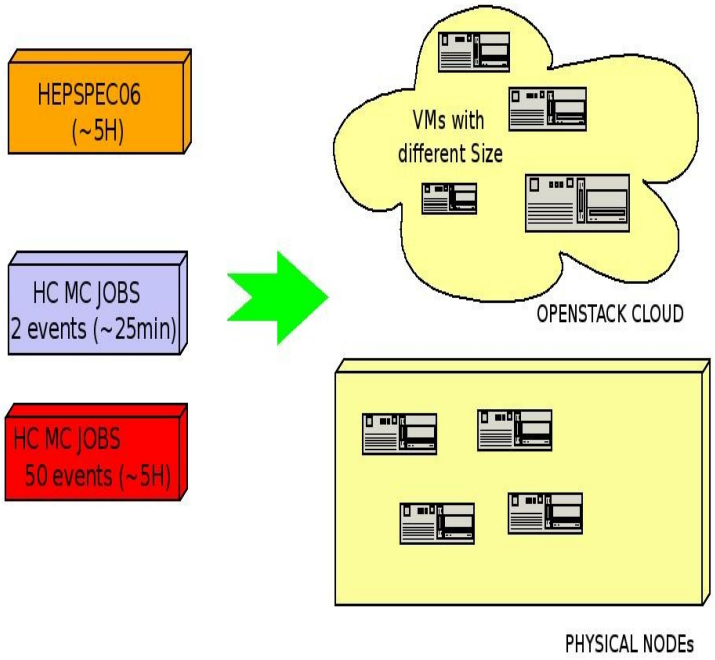
- **We have to take into account the provisioning of extra local disk space for CVMFS cache (~20-30GB per VM)**
- **1-VCPU vs N-VCPU images ?**

Preliminary Base lines

MC PRODUCTION

Test jobs ...

- MC 2 event per jobs and ~20-30min expected wall-clock time
- MC 50 event per jobs ~ wall-clocklock ~5h per job
- Hespec06-64bit (~5h) per job, gcc 4.7 default opts ?
- We run concurrent jobs up to the maximum number of VCPU per instance.



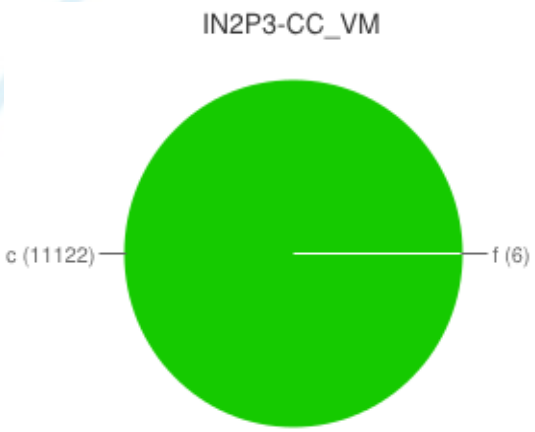
HammerCloud

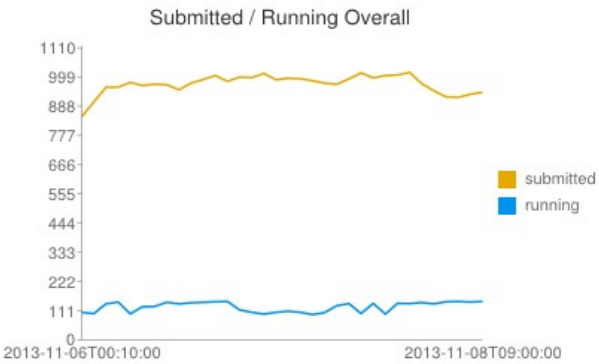
- "... HammerCloud is a Distributed Analysis testing system. It can test yours site(s) and report the results obtained on that test. Used to perform basic site validation, help commission new sites, evaluate SW changes, compare site performances ... ".
- Moreover, HC is used to perform functional test
  - **https://twiki.cern.ch/twiki/bin/view/Main/HammerCloud**
  - **http://hammercloud.cern.ch/hc/**

MC 2 events

- Prod mc12 AtlasG4_trf 17.2.6.2 use_all_spacetokens (HC 498)
- AtlasProduction/17.2.6.2 (?)
- mc12_8TeV.175590.Herwigpp_pMSSM_DStau_MSL_120_M1_000.evgen.EVNT.e1707_tid01212395_00_derHCBMGanga

NO GRID FAILURE!



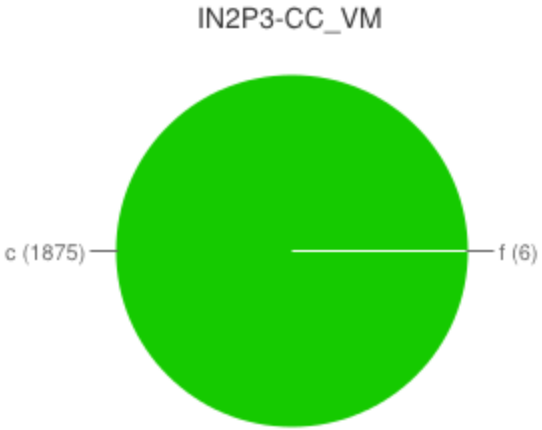


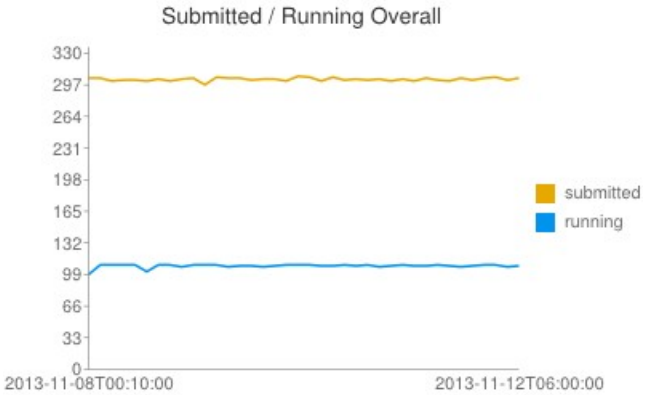
http://hammercloud.cern.ch/hc/app/atlas/test/20027976/

MC 50 events

- CERN-P1 StressTest - mc12 AtlasG4_trf 17.2.2.2 (512)
- AtlasProduction/17.2.2.2
- mc12_8TeV.175590.Herwigpp_pMSSM_DStau_MSL_120_M1_000.evgen.EVNT.e1707_tid01212395_00_derHCBM

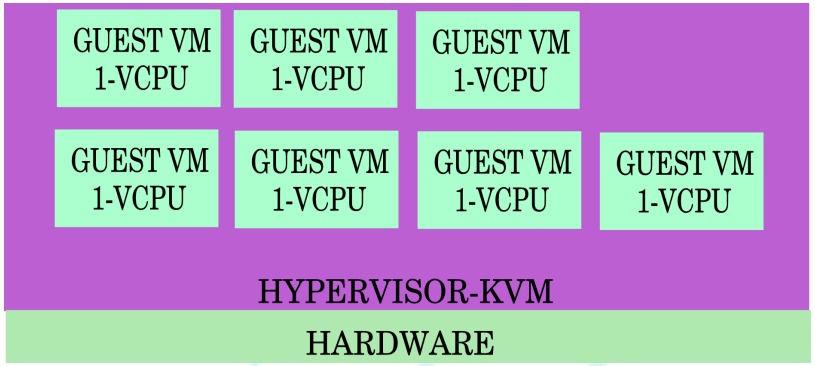
NO GRID FAILURE!



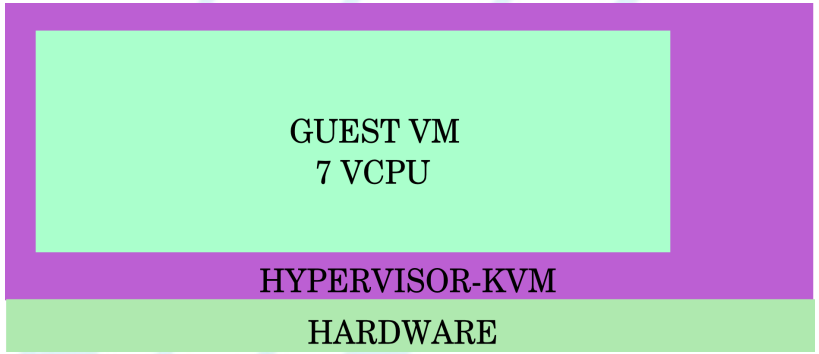


http://hammercloud.cern.ch/hc/app/atlas/test/20028042/
http://hammercloud.cern.ch/hc/app/atlas/test/20028049/

Partitioning



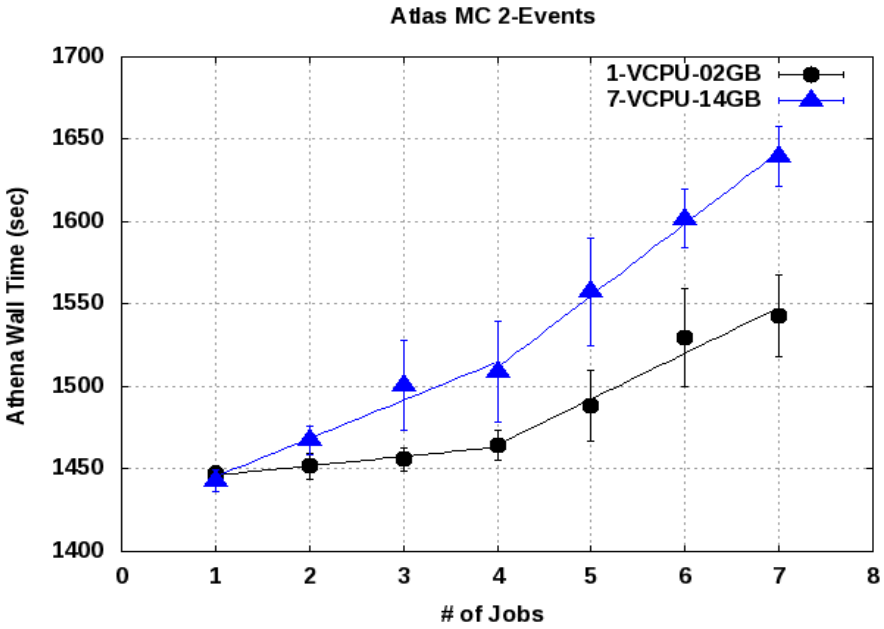
7 X 1 VCPU 2GB RAM +20 GB EHD

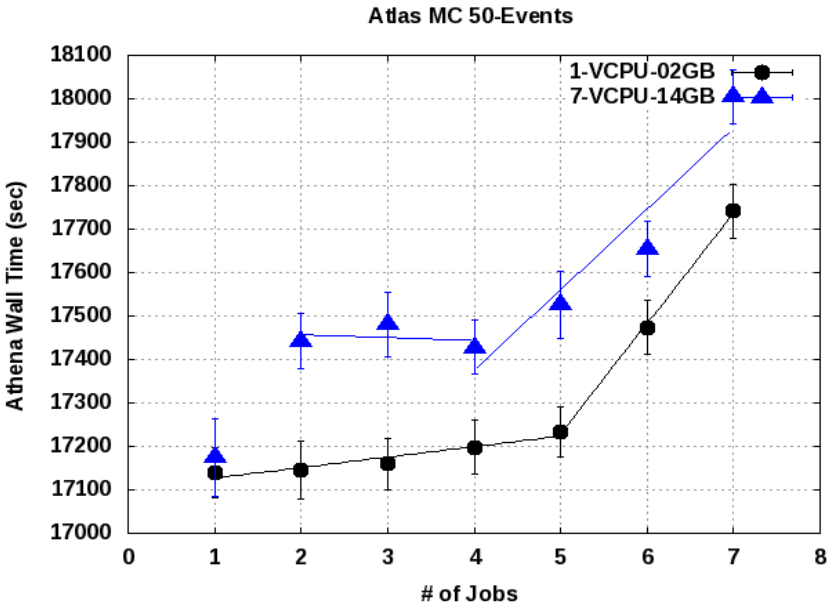


1 X 7 VCPU 14GB RAM +100 GB EHD

- We tried different combination of CPU partitioning and memory and we ran concurrent jobs equal to the number of VCPU for each case.
  - **7 X 1 VCPU 2G +20 GB EHD**
  - **1X 7 VCPU 10G +100 GB EHD**
  - **1X 7 VCPU 12G +100 GB EHD**
  - **1X 7 VCPU 14G +100 GB EHD**
  - **1X 8 VCPU 10G +100GB EHD**
  - **1X 8 VCPU 12G +100 GB EHD**
  - **1X 8 VCPU 14G +100 GB EHD**

MC 2/50 events

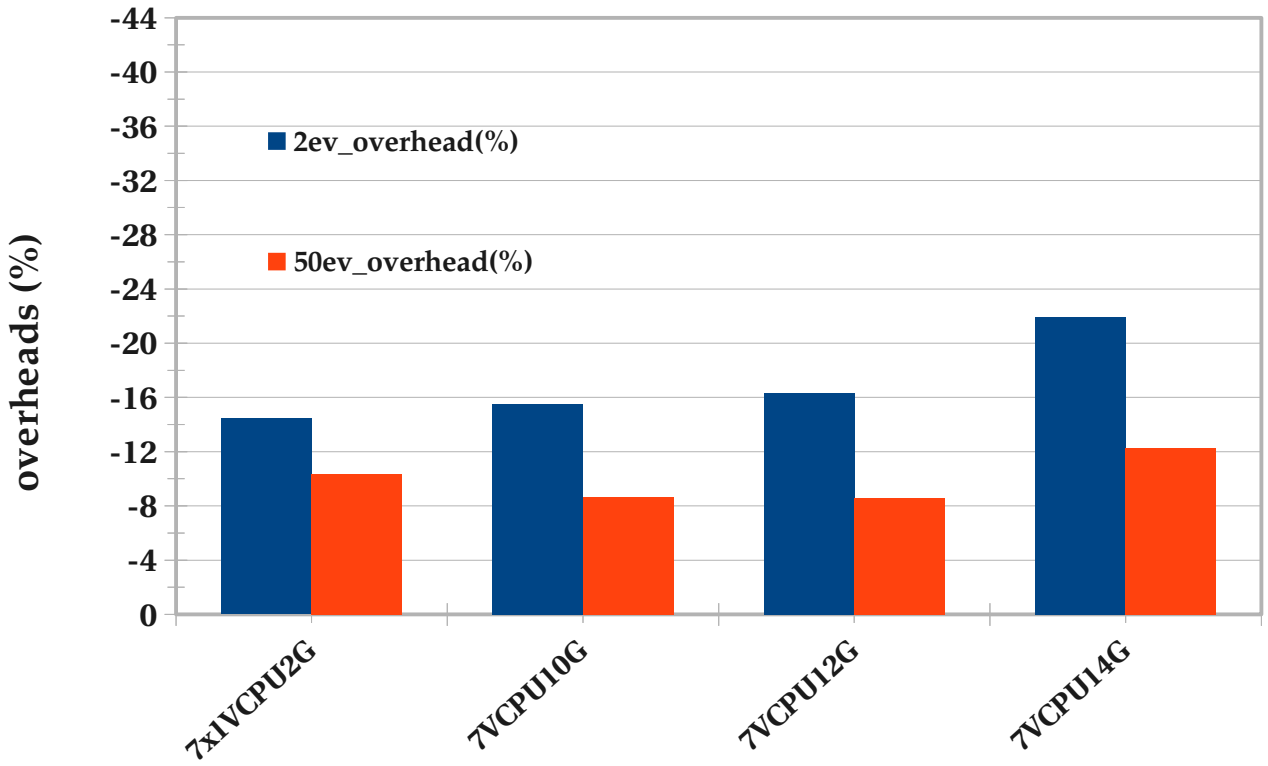




Walltime Overheads: 7-VCPU

DRAFT

$$OVERHEAD = 100\,x\,\frac{PHYSICAL - VIRTUAL}{PHYSICAL}$$

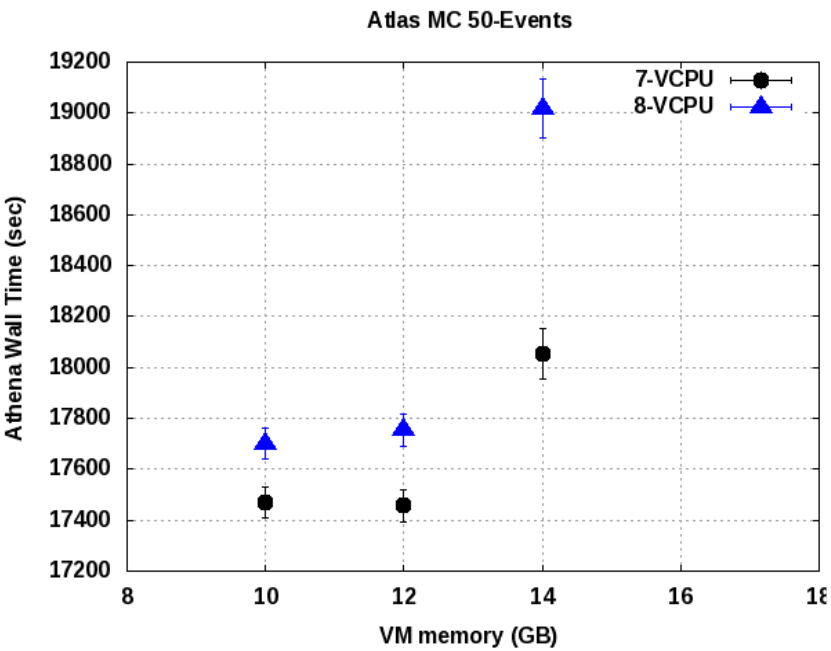


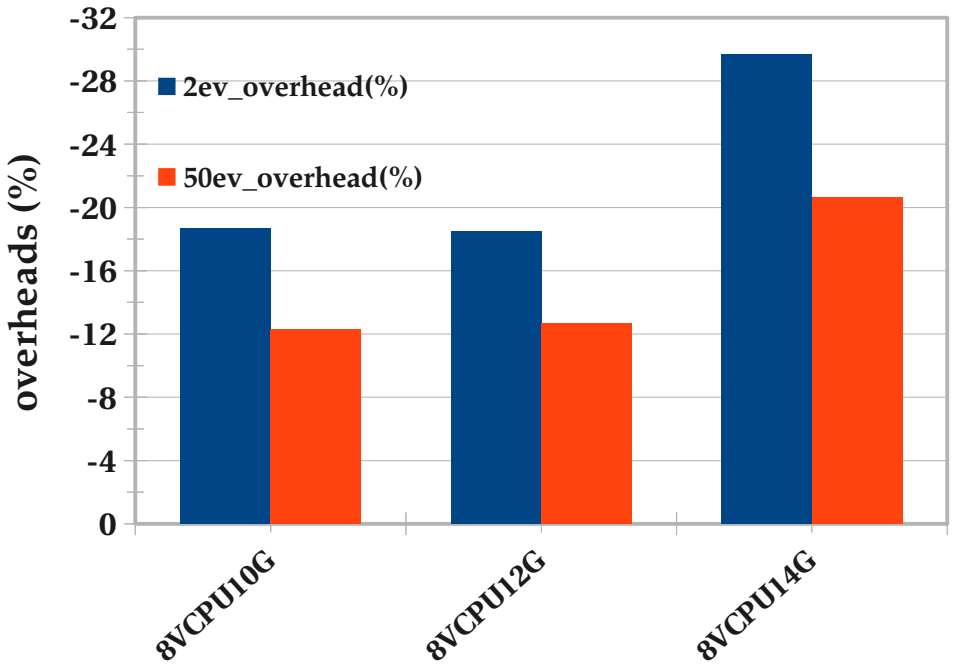
Attention the runs on <u>physical reference machine</u> took place manually due to restrict network connectivity.

VM with 7/8-VCPU

DRAFT

$$OVERHEAD = 100\,x\,\frac{PHYSICAL - VIRTUAL}{PHYSICAL}$$





Attention, the runs on <u>physical reference machine</u> took place manually due to restrict network connectivity.

Preliminary remarks

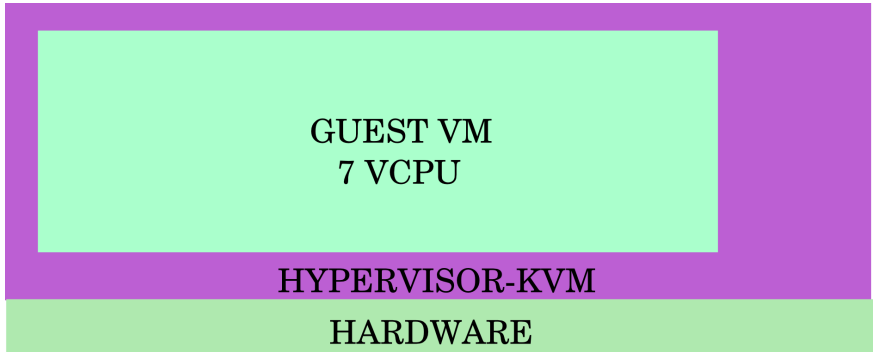
- We have a stable virtual-WN for atlas MC jobs :
  - OPENSTACK INFRASTRUCTURE + IMAGE
- The N x 1-VCPU vs 1 x N-VCPU Vms exhibits difference ~ **2%** for a long MC job for N=7, with total memory allocation of 14GB RAM on the specific H/W.
- We made a first estimation that the walltime overheads due to virtualization for a long MC job about -9% for a hypervisor away from the saturation (7VCPU+10GB Ram over 8CPU +16GB Ram - physical host).
- There are indications that virtualization overheads depend from the amount of the memory that is left for Hypervisor OS.
- The question which arises is : how much memory we should leave for host OS ?

ANALYSIS

ANALYSIS tests

- 23 hypervisors in total
- One vm per hypervisor
- 1Gbit Ethernet per vm
- 1 job per VM
- Maximum 10 jobs running jobs
- Compare with the real farm ?
- Xrdcp-cp vs direct reading

- We can try to test analysis code (ROOT) just to define the limits of I/O : local HD & network.



1 X 7 VCPU 10GB RAM +100 GB EHD

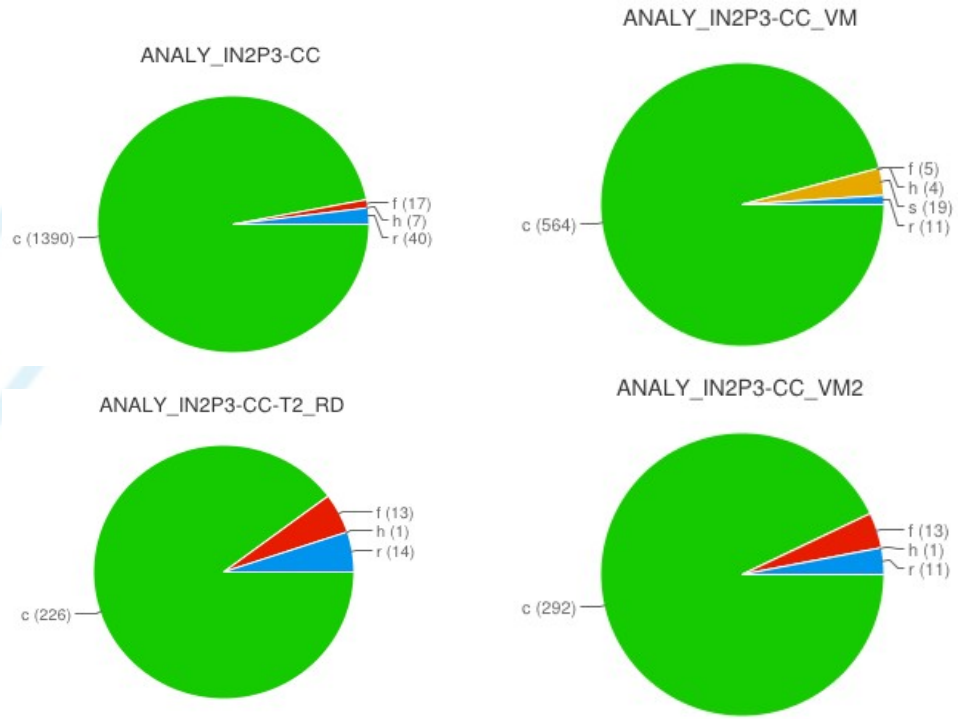
HC ANALYSIS TEST

- ROOT HWWNtupleCode-00-02-07 DATA p1067 17.2.7 Panda SL6 (533)
- user.flegger.*.data12_8TeV*physics_Muons.merge.NTUP_SMWZ*

DRAFT

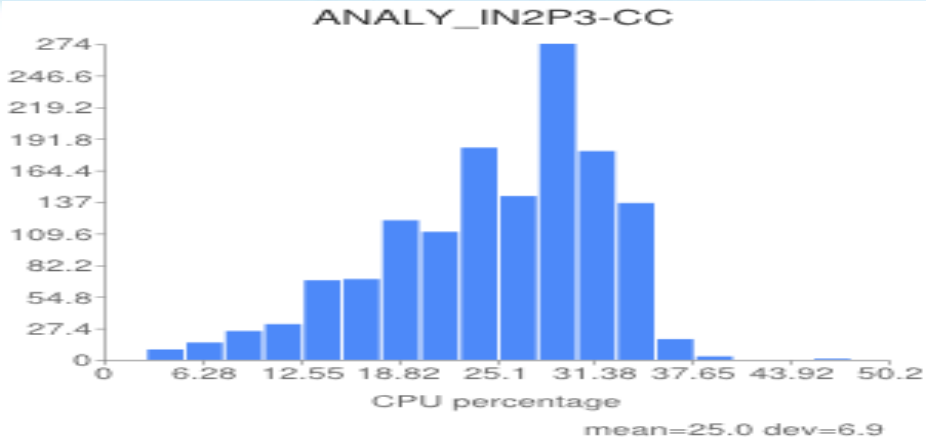
Xrdcp (cp-2-scratch)

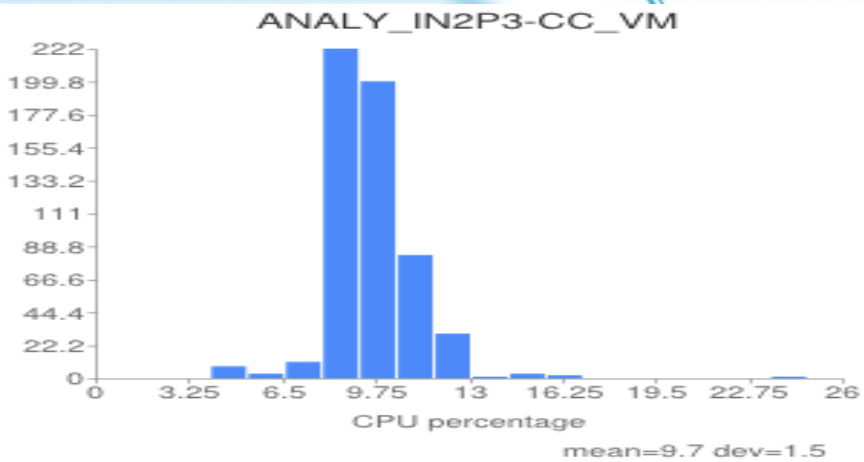
Xrootd direct

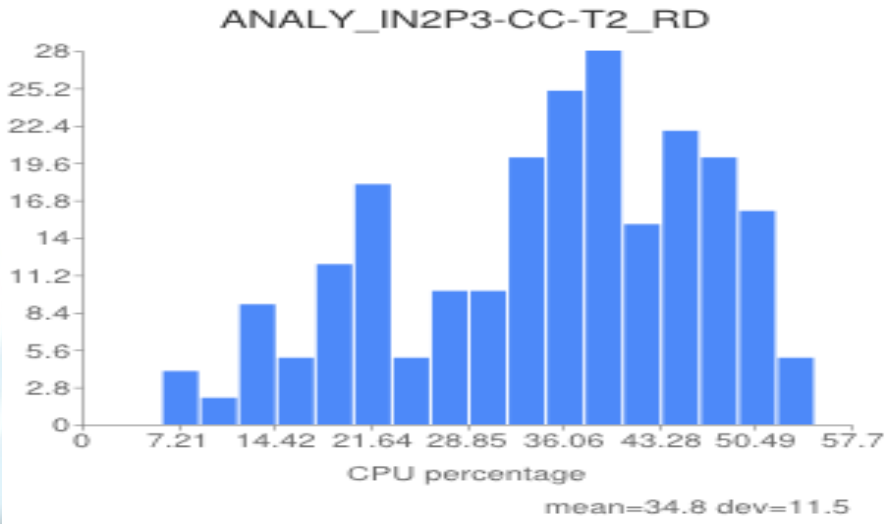


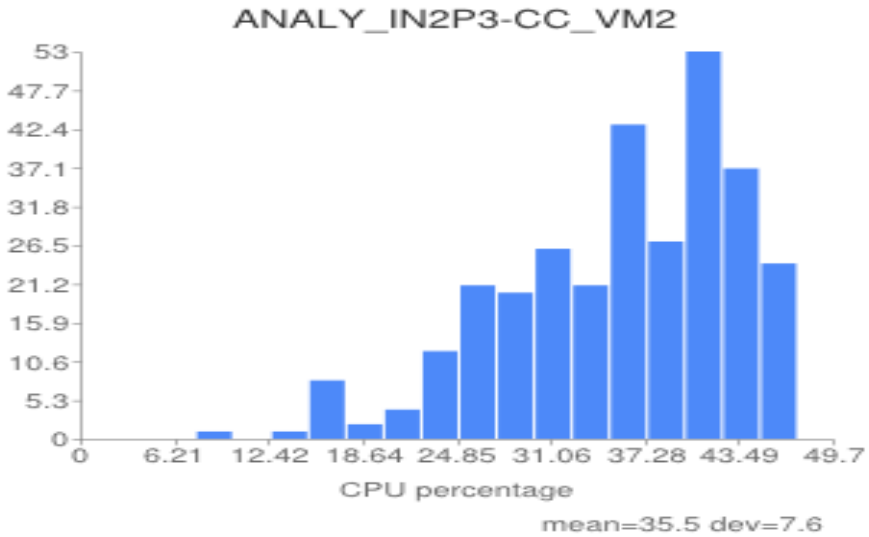
- Real Wns Farm
- 1 vm with 7 VCPU 10GB RAM +100GB SCRATCH
- MAX 10 concurrent jobs for both configuration

CPU Efficiency (%)

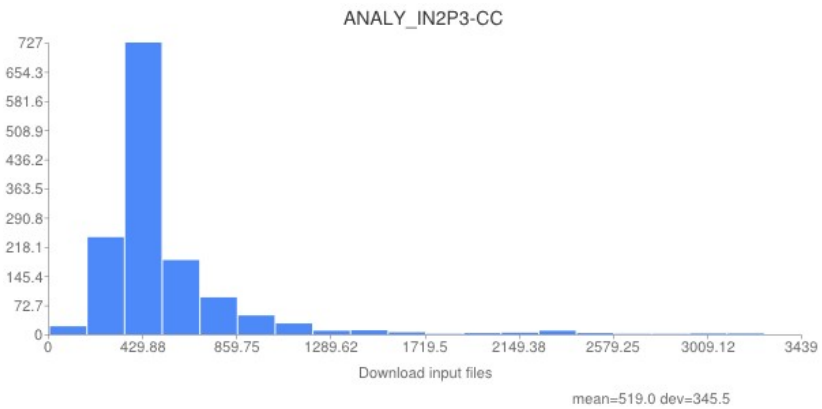


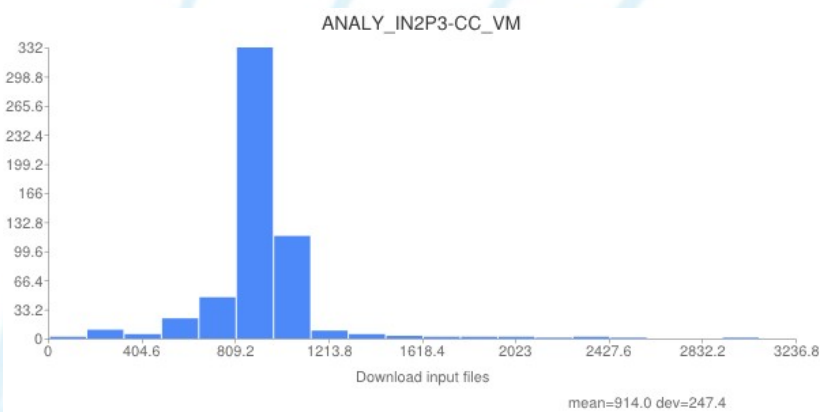






Stage in timing for (xrdcp)

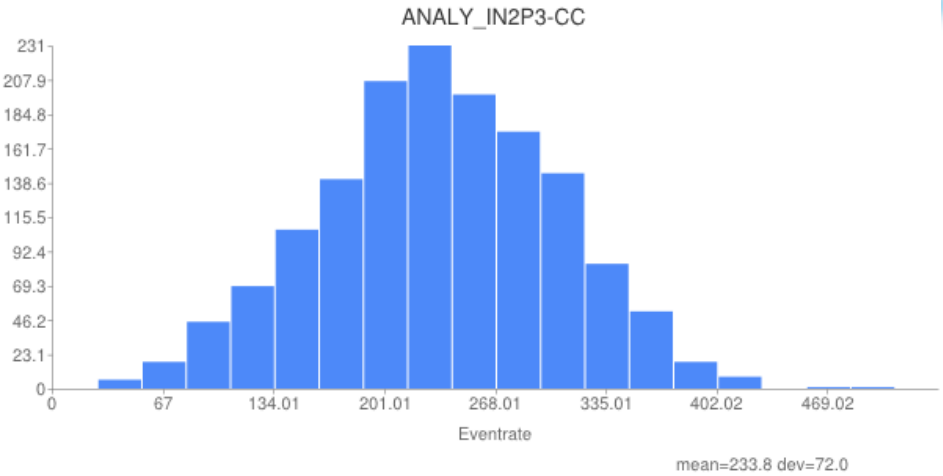


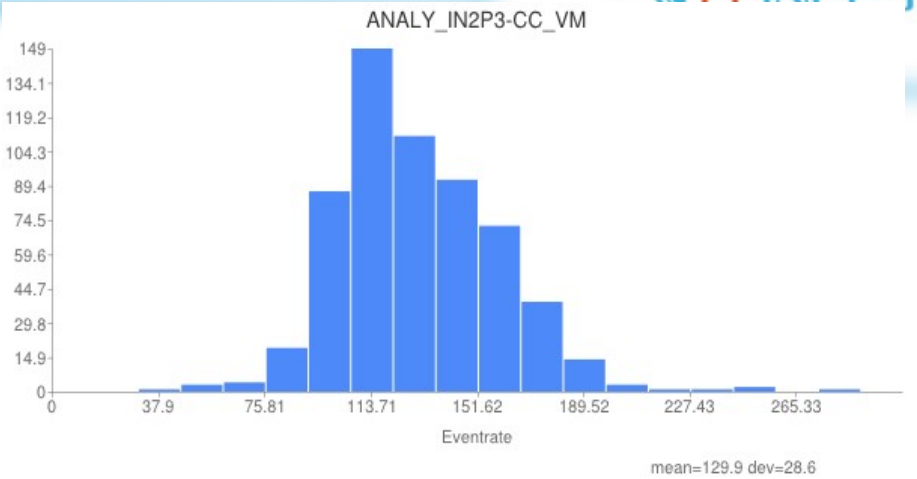


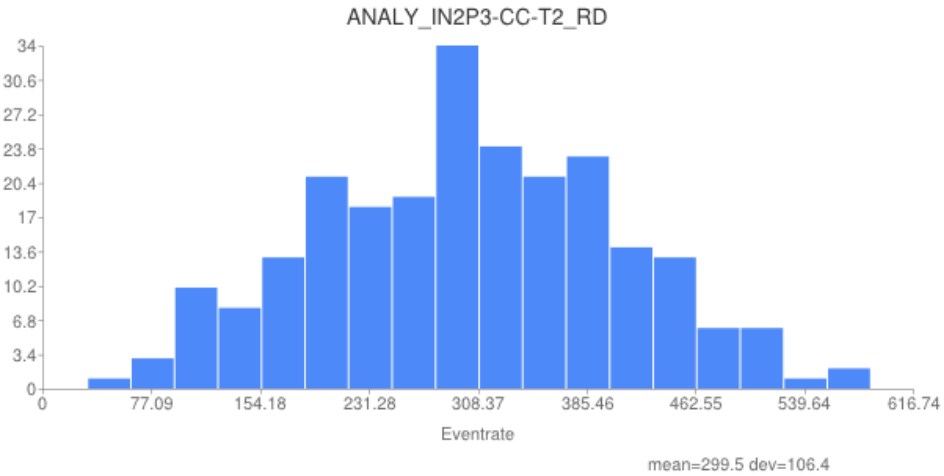
- Xrdcp took ~ 2 times more in vm than in the real farm to download

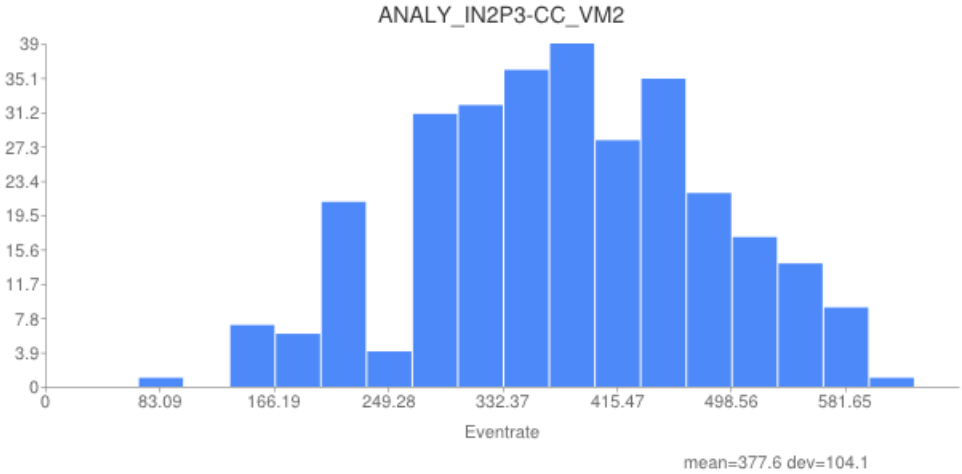
  ~22GB data in 5 files
- Attention! Not same H/Ws

EventRate

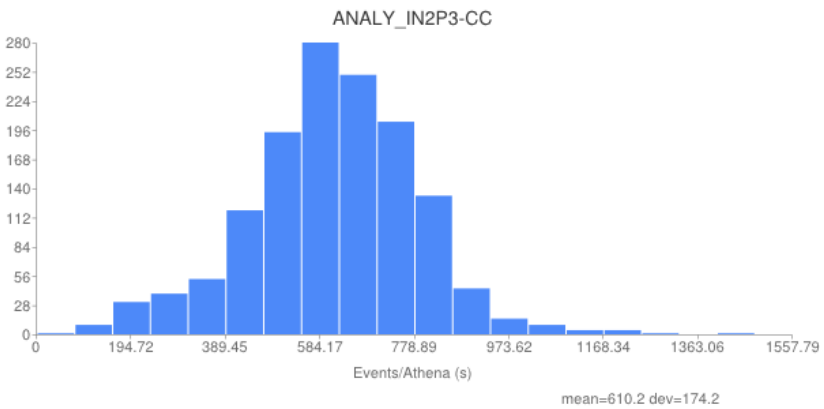

ANALY_IN2P3-CC
231
207.9
184.8
161.7
138.6
115.5
92.4
69.3
46.2
23.1
0
0
67
134.01
201.01
268.01
335.01
402.02
469.02
Eventrate
mean=233.8 dev=72.0

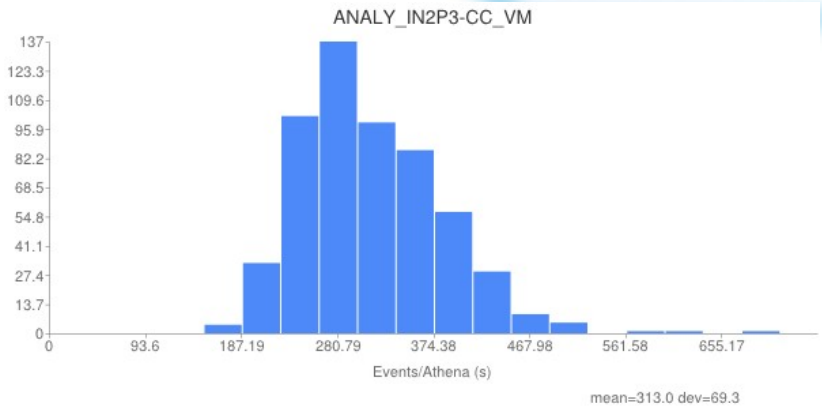

ANALY_IN2P3-CC_VM
149
134.1
119.2
104.3
89.4
74.5
59.6
44.7
29.8
14.9
0
0
37.9
75.81
113.71
151.62
189.52
227.43
265.33
Eventrate
mean=129.9 dev=28.6

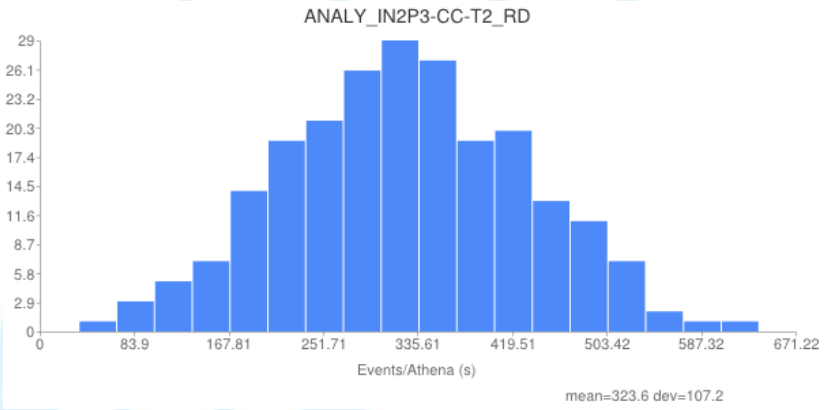

ANALY_IN2P3-CC-T2_RD
34
30.6
27.2
23.8
20.4
17
13.6
10.2
6.8
3.4
0
0
77.09
154.18
231.28
308.37
385.46
462.55
539.64
616.74
Eventrate
mean=299.5 dev=106.4

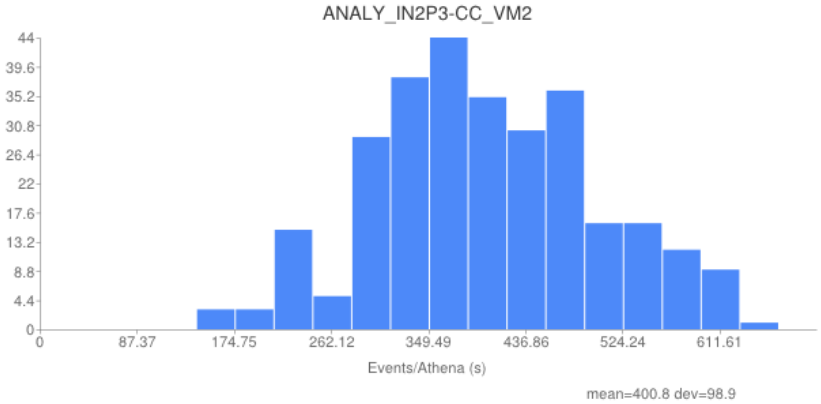

ANALY_IN2P3-CC_VM2
39
35.1
31.2
27.3
23.4
19.5
15.6
11.7
7.8
3.9
0
0
83.09
166.19
249.28
332.37
415.47
498.56
581.65
Eventrate
mean=377.6 dev=104.1

Event/athena_walltime

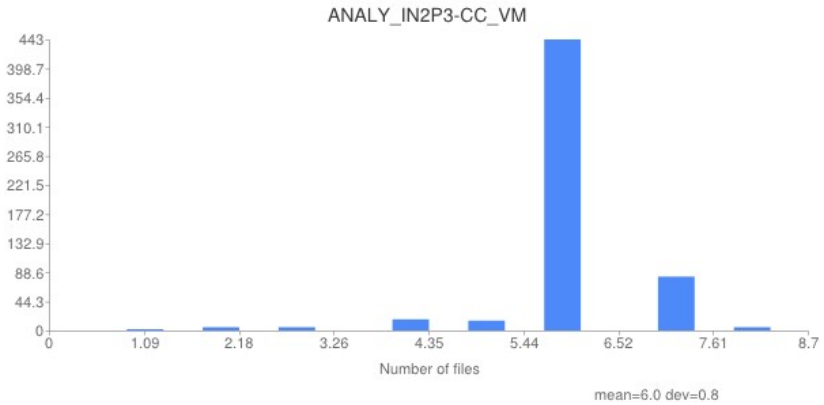


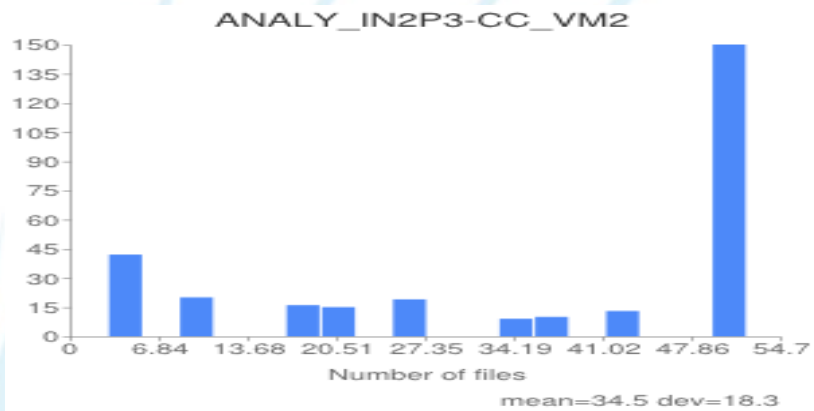






Number of Files (!)





- the number of process files is not the same between the cp to scratch and direct reading
- RD jobs have bigger wall times with respect to cp-scratch jobs

Remarks

- Analysis test was only for demonstration of I/O overhead due to reduce local I/O in VM

- We expect some improving with tuning and upgraded hardware

Last but not least

- *All this work it is a **first effort** to understand better the new arriving environment and available testing tools*
- *address potential direction of further studies and tests*
- *Also we did not make any image tuning …*

Further plans

Further plans I

- We should repeat these tests on modern hardware with more memory and better local I/O subsystem (flexible choice of image size).
- And with standard way of sampling in a “tandem” configurations (VM+HyperVisor vs Physical).
- We should define some Benchmark-ing tactic with respect to interplay of qemu-kvm, linux kernel and underline hardware technology.

Further plans II

- **Standardized the image creation and maintenance**
  - **Updates of the images it is an issue (~GRID STABILITY).**
- **Make next tests with Condor EC2 + APF (or Cloud Sceduler) for the provisioning of v-WNs according with the number of incoming jobs**
- **The configuration with multiple condor collector (scalability).**
- **The configuration with ephemeral squids in side to the CLOUD (shoal project: https://github.com/hep-gc/shoal).**
- **Standardized the usage of the HammerCloud tests based on wide used templates + interact more with HC group.**
- **We want to pass to pre-production state with real MC jobs (HIGH PRIORITY).**
- **Test the submition of VM/jobs to other French Clouds, this it will be a interesting exercise over the WAN.**

Further Reading …

- Overlay Opportunistic Clouds in **CMS/ATLAS** at CERN: The CMSoooooCloud in Detail "By: Jose Antonio Coarasa Perez, CERN OpenStack Summit Presentations, Porland 2013, Link
- **Cloud Computing Patterns** – Fundamentals to Design, Build, and Manage Cloud Applications - Christoph Fehling (Uni Stuttgart), Gridka School 2013, Link
- GridKa School 2013: **OpenStack** tutorial,Link
- EGI- Fedcloud-tf:WorkGroups:Scenario4,Link
- https://twiki.cern.ch/twiki/bin/view/LCG/CloudTesting

Acknowledgments

- CC - Openstack Group
- CC Network Group
- Operation/Support Group
- Sysgrid/sysLinux/Storage

- Atlas France Squad Team
- APF Developers
- AtlasCloud R/D Group
- Atlas HammerCloud Group
- Atlas Frontier Group
- Atlas ADC experts

BACKUP

Motivation: *Helix-Neboula*

- *"… The project aims to pave the way for the development and exploitation of a Cloud Computing Infrastructure, initially based on the needs of European IT-intense scientific research organizations, while also allowing the inclusion of other stakeholders' needs (governments, businesses and citizens …" [I].*
- *"Helix Nebula – the Science Cloud: a public-private partnership building a multidisciplinary cloud platform for data intensive science", Bob Jones (CERN) , GRIDKA School 2013 [II].*

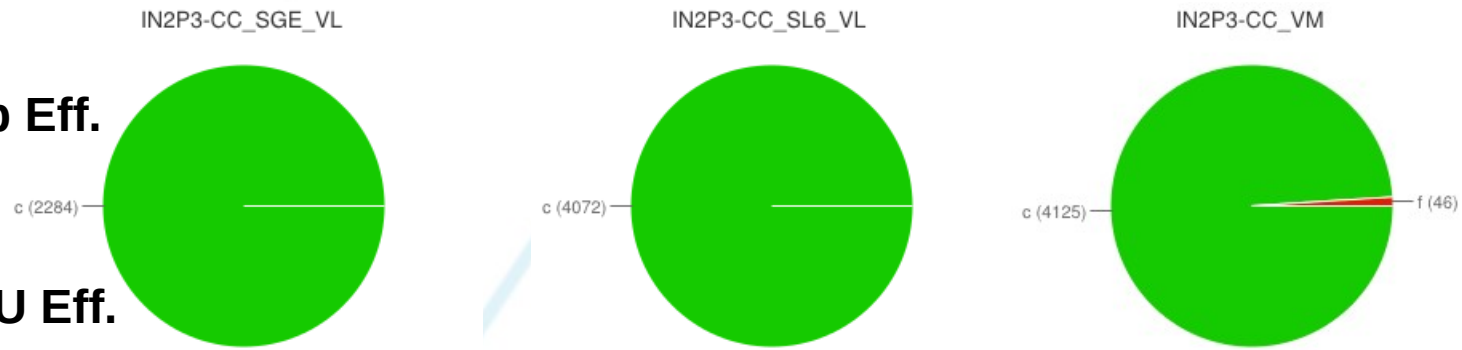
I. ***http://helix-nebula.eu/***

II. ***http://indico.scc.kit.edu/indico/materialDisplay.py?contribId=19&sessionId=1&materialId=slides&confId=26***

HammerCloud I

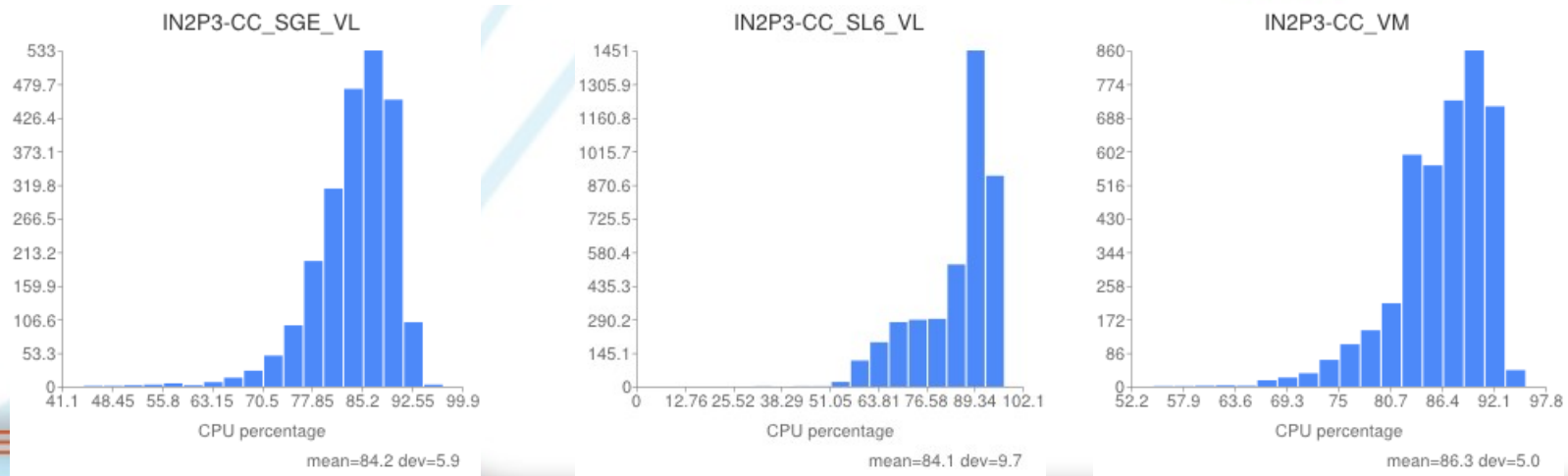
- **Input DS Patterns: mc12_8TeV*evgen.EVNT***
- **Ganga Job Template: ProdTrans/G4_17262.tpl (498)**

Job Eff.



~1%

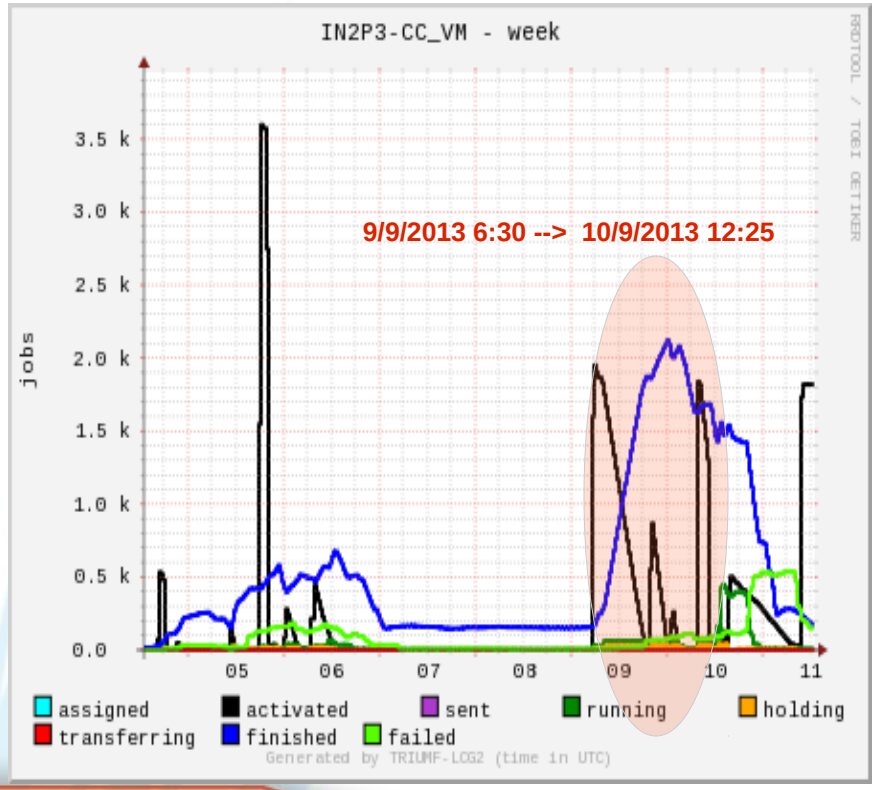
CPU Eff.

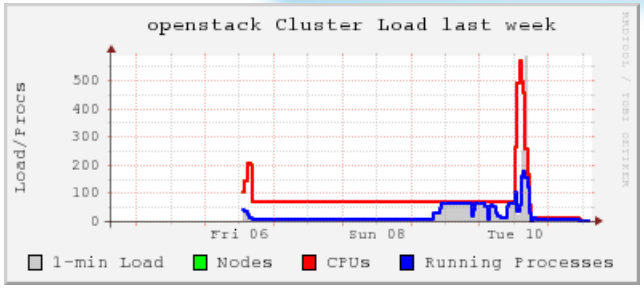


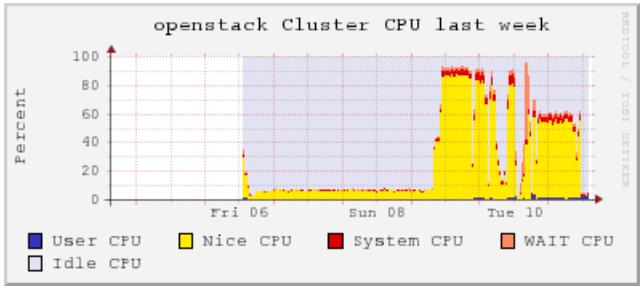
VMs ramp-up

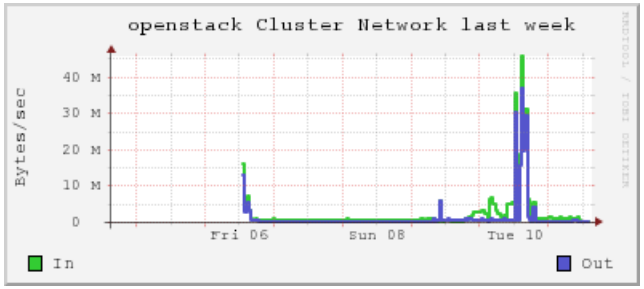
- **Input DS Patterns: mc12_8TeV*evgen.EVNT***
- **Ganga Job Template: ProdTrans/G4_17262.tpl (498)**

64 V-Worker nodes

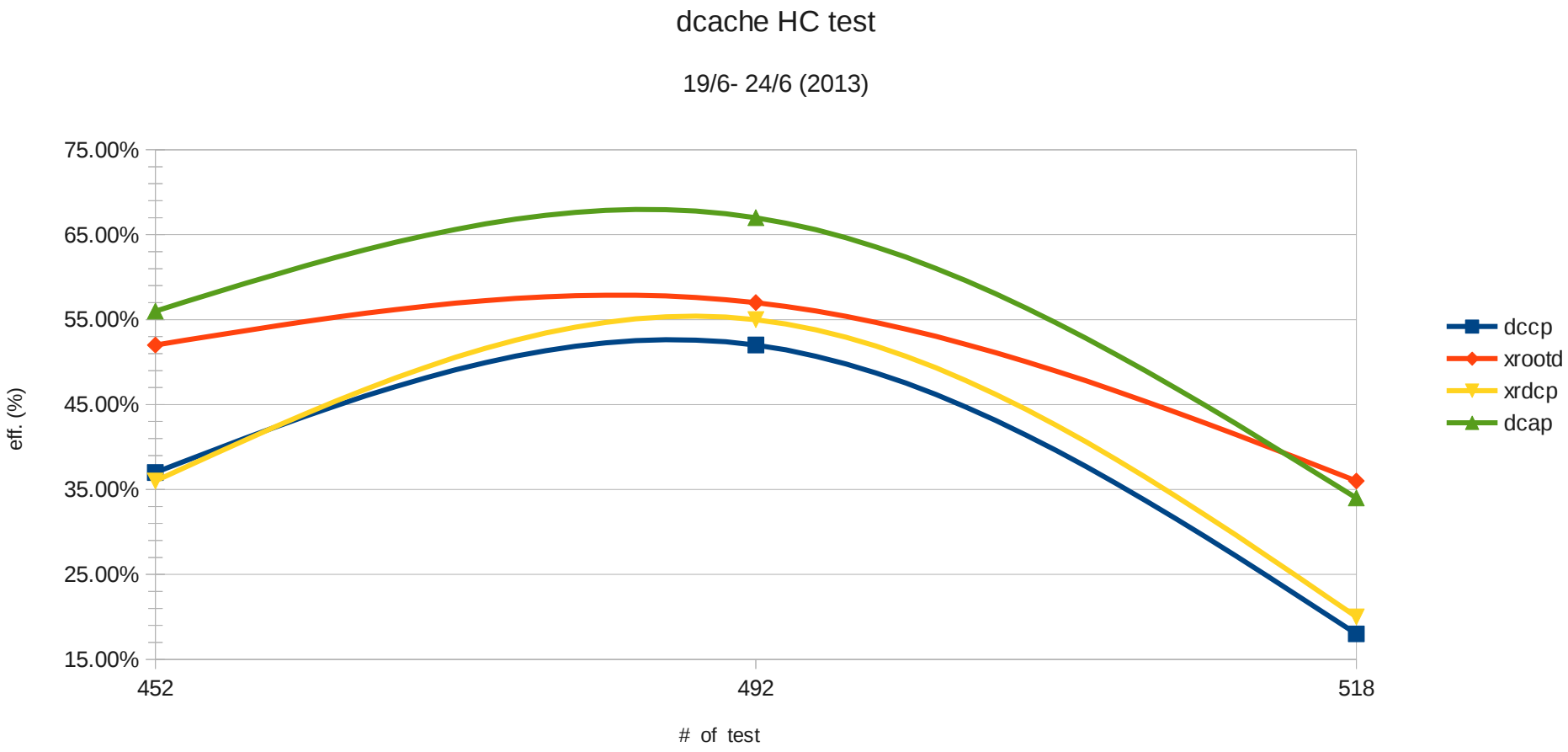








HC tests(ROOT): 9/6 to 23/6



http://hammercloud.cern.ch/hc/app/atlas/test/20022513/ -->(492)
http://hammercloud.cern.ch/hc/app/atlas/test/20022530/ --> (452)
http://hammercloud.cern.ch/hc/app/atlas/test/20022543/ --> 518 (Fax)