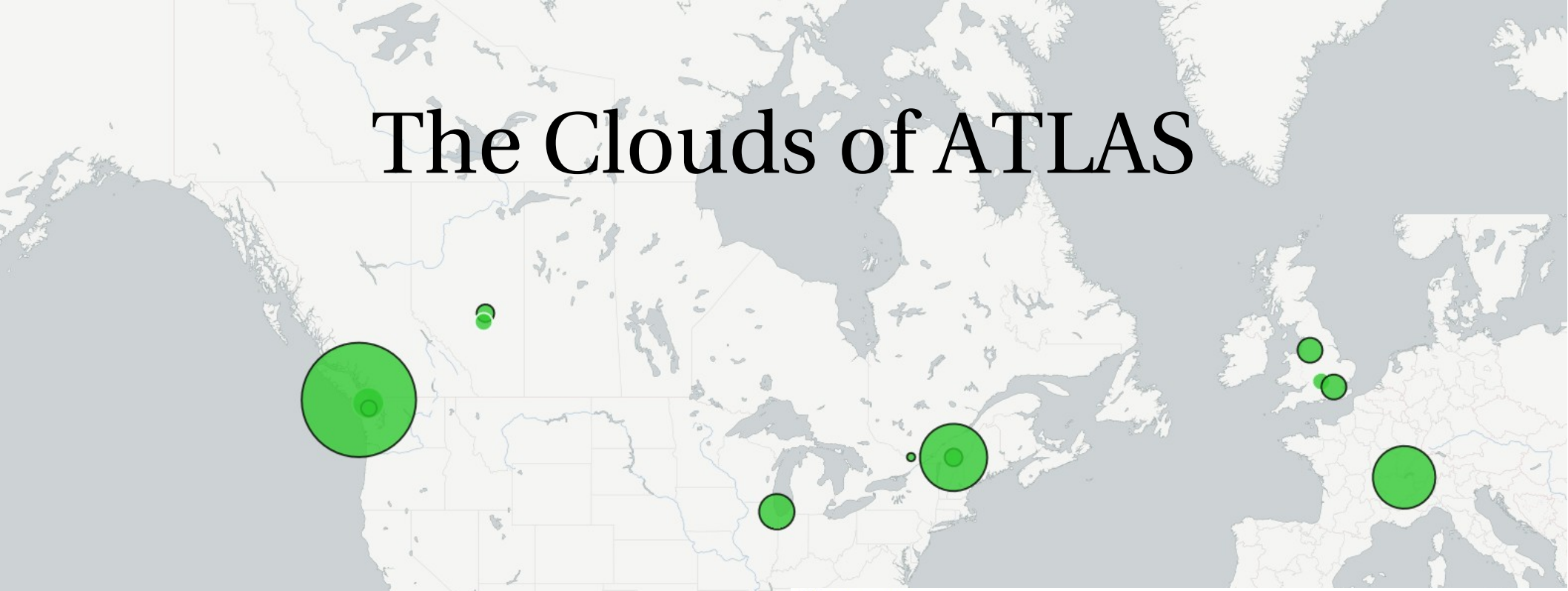
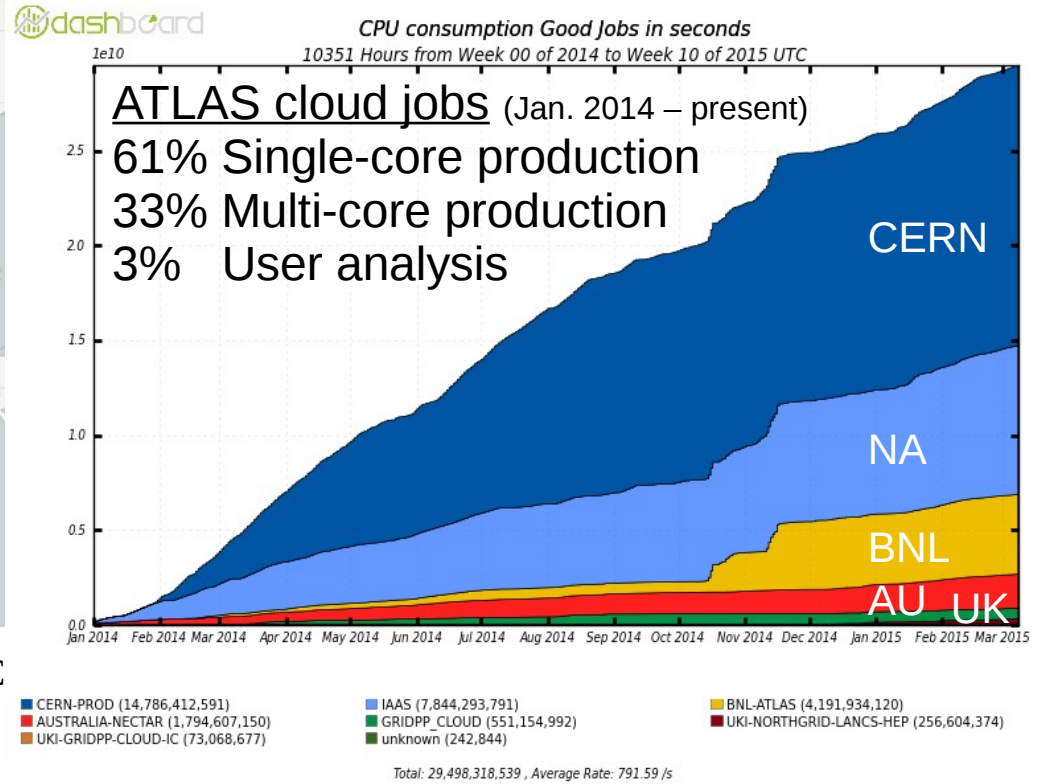
Evolution of Cloud Computing in ATLAS

Ryan Taylor
on behalf of the ATLAS collaboration

21st International Conference on Computing in High Energy and Nuclear Physics **CHEP2015** Okinawa Japan: April 13 – 17, 2015

The Clouds of ATLAS



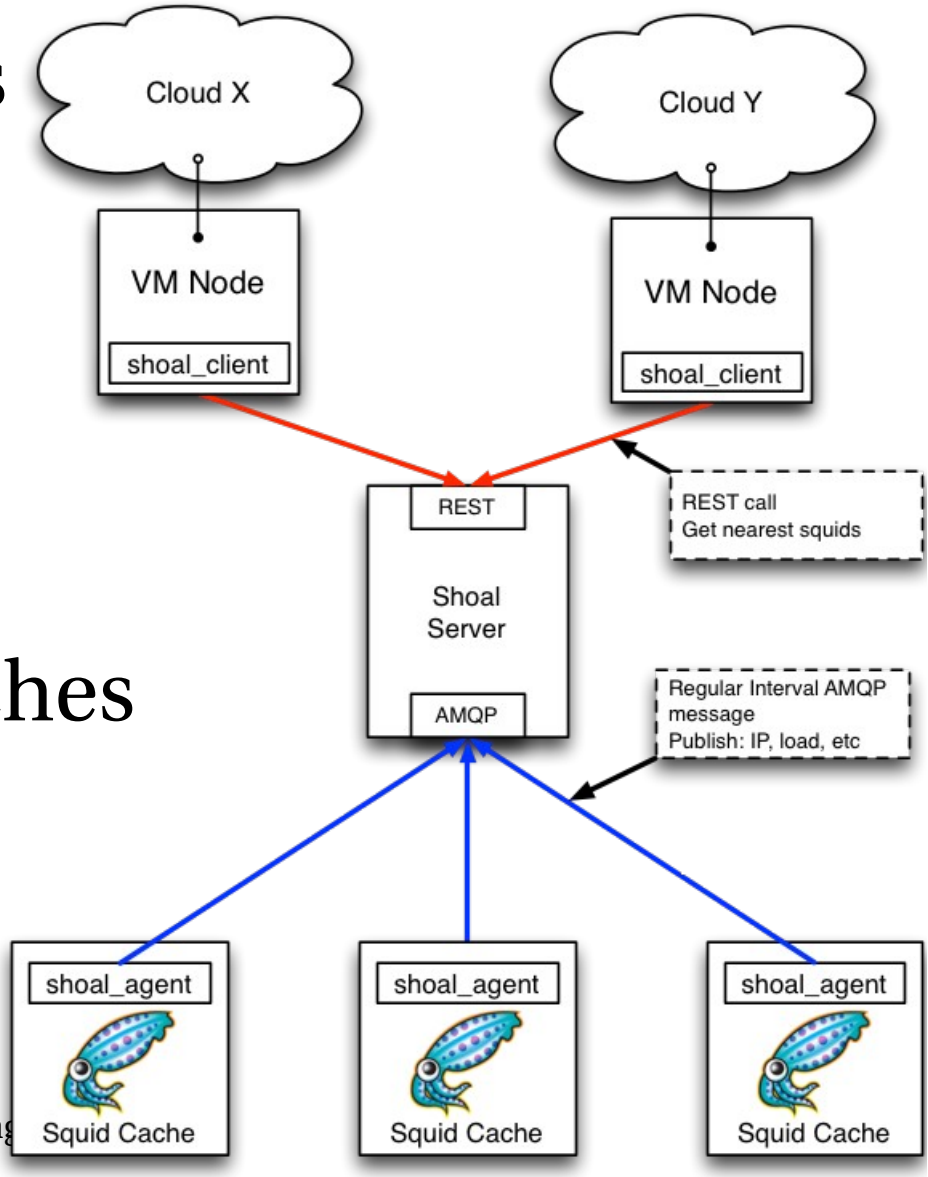


IaaS Resource Management

- Primarily using HTCondor + Cloud Scheduler
  - See talk "HEP cloud production using the CloudScheduler/HTCondor Architecture"
  - In operation for ATLAS for > 3 years
- Dynamic Condor slots to handle arbitrary job requirements
  - e.g. single-core, multi-core, high-mem
- uCernVM image
- Contextualization using cloud-init
- Using *Glint* Image Management System (see poster)

Shoal
Proxy Cache "Federator"

- Shoal Server tracks squids
  - discover new ones
  - verify functionality
  - remove missing/faulty ones
- Workers find best squids
  - based on squid proximity and load
- Build a fabric of proxy caches
  - configurationless topology
  - robust
  - scalable



Shoal | Latest | Services

http://shoal.heprc.uvic.ca

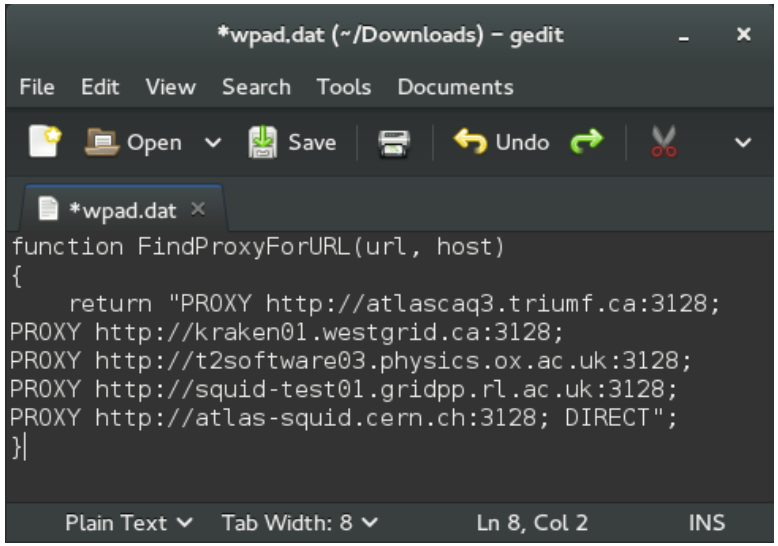
List of Active Squids

5 active in the last 180 seconds

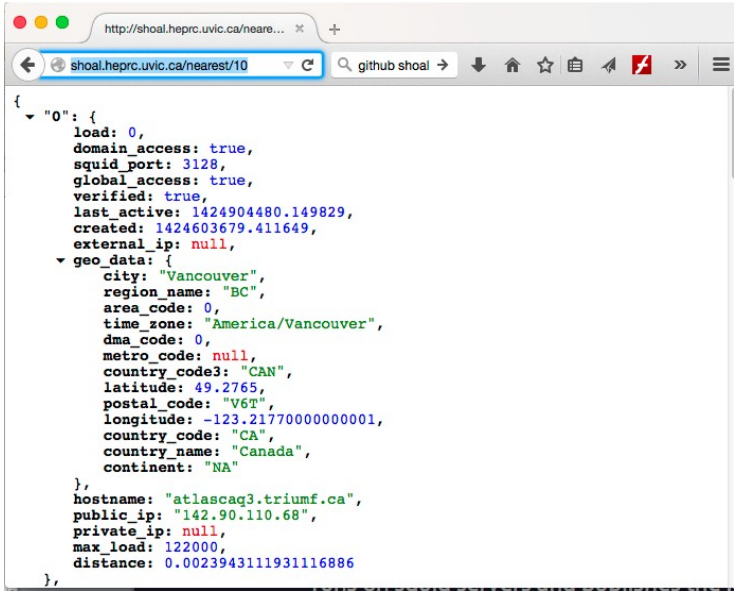
#	Hostname	Public IP	Private IP	Bytes Out	City	Region	Country	Latitude	Longitude	Last Received	Alive	Verified	Access Level
1	squid-test01.gridpp.rl.ac.uk	130.246.183.249		0 kB/s	Appleton		United Kingdom	51.7	-1.35	7s	42h40m43s	✔	Global
2	kraken01.westgrid.ca	206.12.48.249	172.22.2.25	809 kB/s	Vancouver		Canada	49.2836	-123.1041	10s	107h49m9s	✔	Global
3	atlascaq3.triumf.ca	142.90.110.68		0 kB/s	Vancouver		Canada	49.2765	-123.2177	20s	166h20m3s	✔	Global
4	atlas-squid.cern.ch	128.142.200.105		0 kB/s	Geneva		Switzerland	46.1956	6.1481	22s	166h19m59s	✘	Global
5	t2software03.physics.ox.ac.uk	163.1.5.175		35 kB/s	Oxford		United Kingdom	51.75	-1.25	26s	166h18m56s	✔	Global

Shoal-Server v0.7.1

PAC Interface



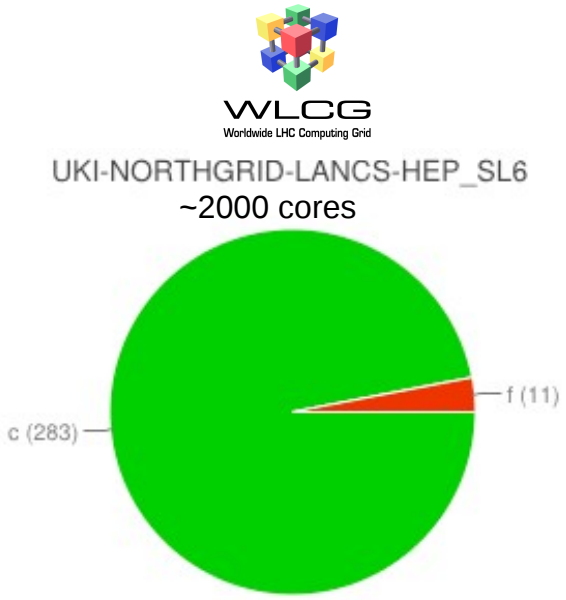
JSON REST Interface

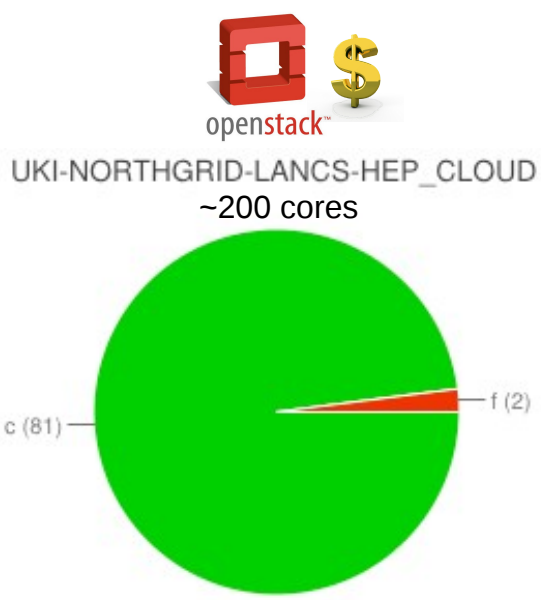


- github.com/hep-gc/shoal
- CHEP 2013 Poster

Grid/Cloud Site Performance Comparison

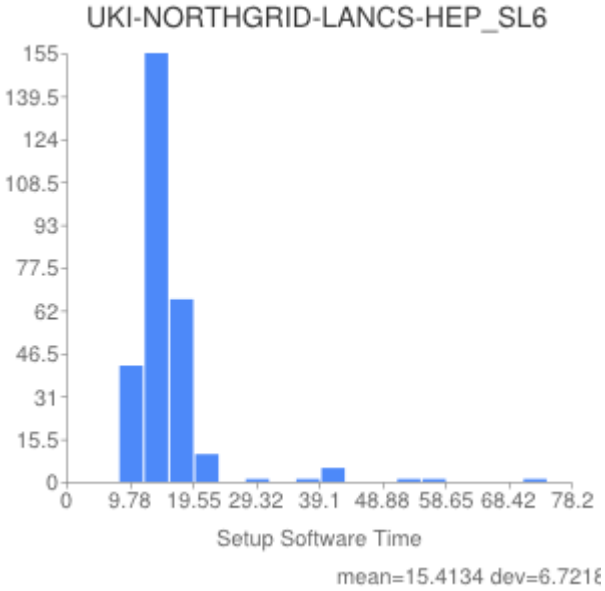
- Used Hammercloud stress tests
- Ran continuous stream of jobs on each site for 24 hours
- Using a single input dataset on the grid storage

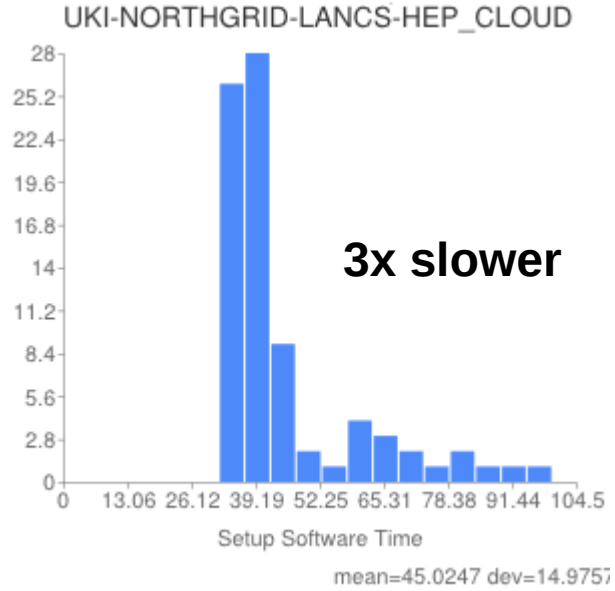




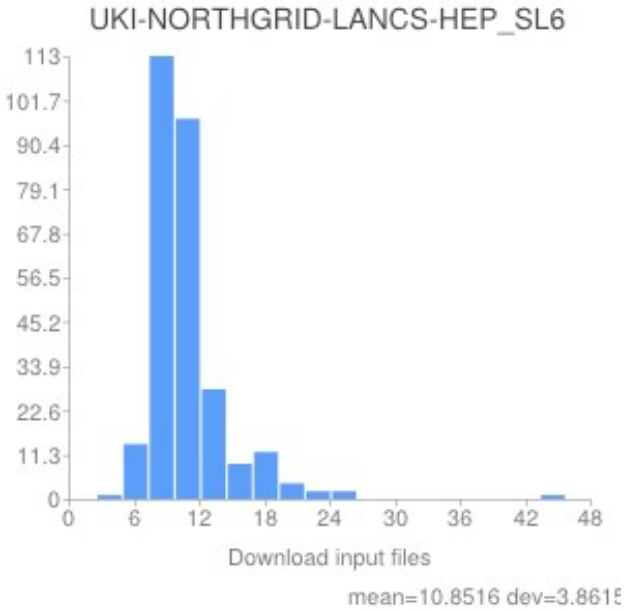
Success rate similar. Grid site processed four times more jobs

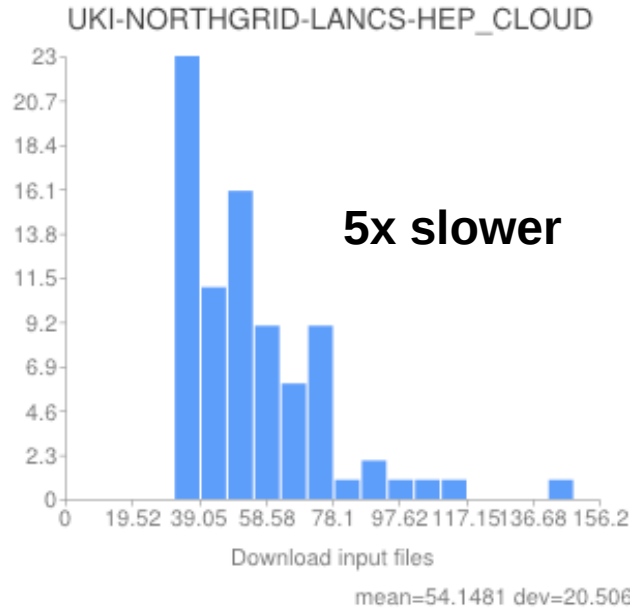
- Software setup time
  - Relies on CVMFS cache and Squid proxy





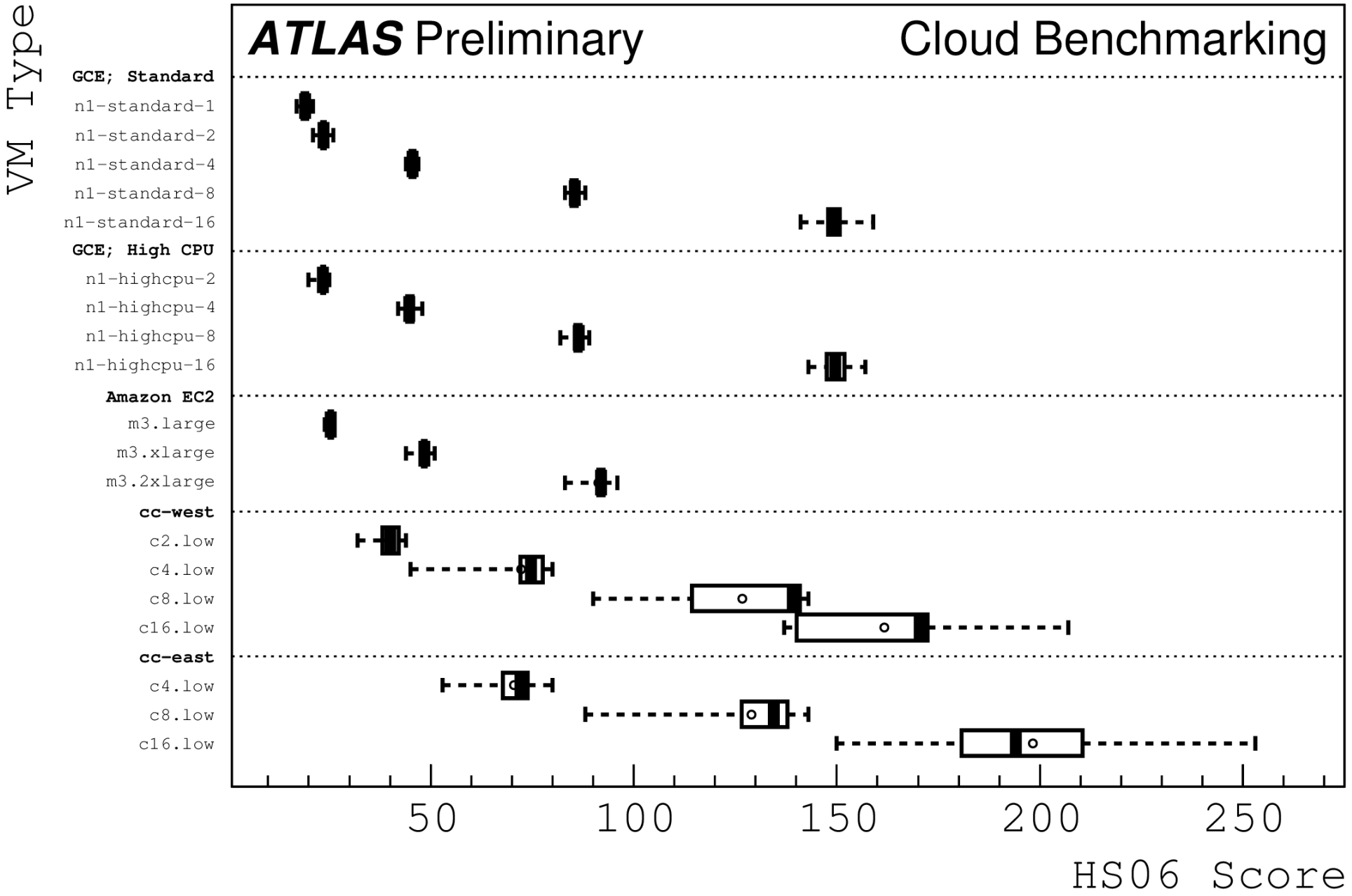
- Data stage-in time
  - Remote vs. local storage access

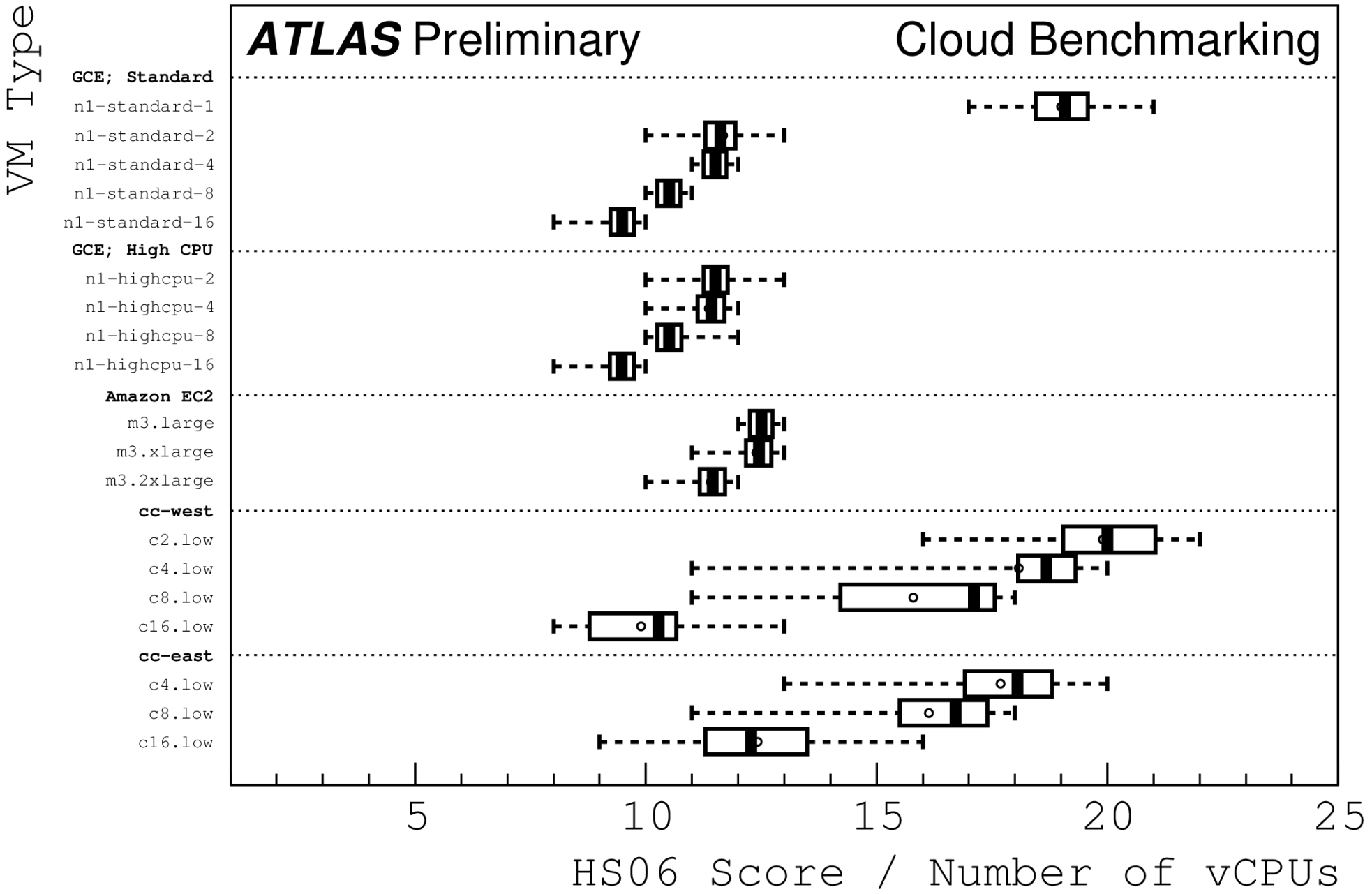




HS06 Benchmarking Study

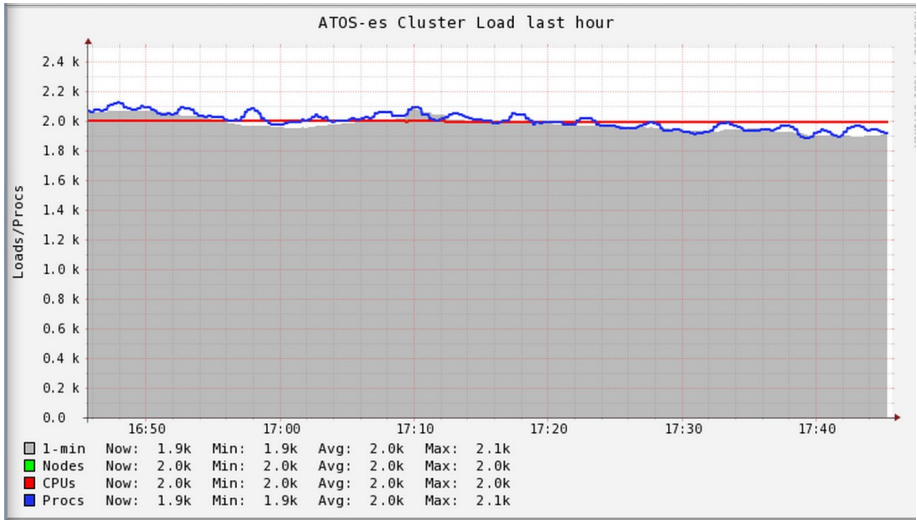
- Commercial clouds provide on-demand scalability
  - e.g. urgent need for beyond pledged resources
- But how cost-effective are they?
- Comparison to institutional clouds


ATLAS Preliminary
Cloud Benchmarking
VM Type
GCE; Standard
n1-standard-1
n1-standard-2
n1-standard-4
n1-standard-8
n1-standard-16
GCE; High CPU
n1-highcpu-2
n1-highcpu-4
n1-highcpu-8
n1-highcpu-16
Amazon EC2
m3.large
m3.xlarge
m3.2xlarge
cc-west
c2.low
c4.low
c8.low
c16.low
cc-east
c4.low
c8.low
c16.low
50
100
150
200
250
HS06 Score


ATLAS Preliminary
Cloud Benchmarking
VM Type
GCE; Standard
n1-standard-1
n1-standard-2
n1-standard-4
n1-standard-8
n1-standard-16
GCE; High CPU
n1-highcpu-2
n1-highcpu-4
n1-highcpu-8
n1-highcpu-16
Amazon EC2
m3.large
m3.xlarge
m3.2xlarge
cc-west
c2.low
c4.low
c8.low
c16.low
cc-east
c4.low
c8.low
c16.low
5
10
15
20
25
HS06 Score / Number of vCPUs

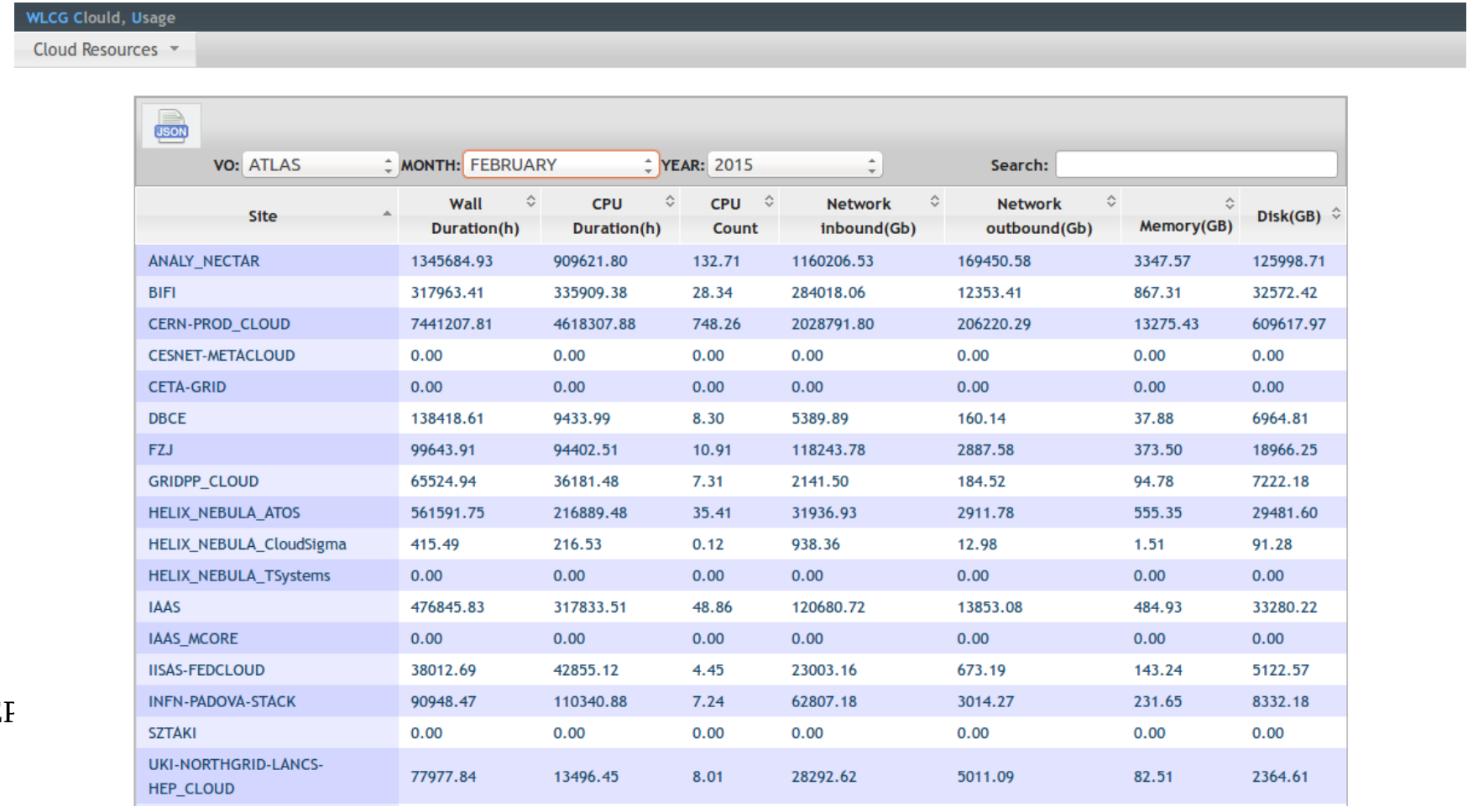
Cloud Monitoring

- VM management becomes the responsibility of the VO
- Basic monitoring is required
  - Detect and restart problematic VMs
  - Identify “dark” resources (deployed but unusable)
  - Can identify inconsistencies in other systems through cross-checks
- Common framework for all VOs
- Implemented with Ganglia
- http://agm.cern.ch/



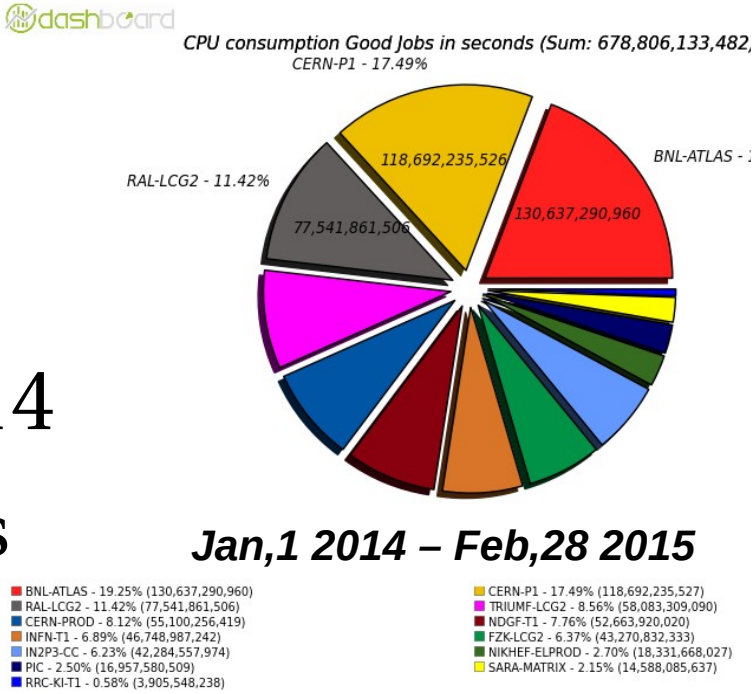
Cloud Accounting

- Provider-side: commercial invoice for resources delivered
- Consumer-side: record resources consumed
- Need to cross-check invoice against recorded usage!



Site	Wall Duration(h)	CPU Duration(h)	CPU Count	Network inbound(Gb)	Network outbound(Gb)	Memory(GB)	Disk(GB)
ANALY_NECTAR	1345684.93	909621.80	132.71	1160206.53	169450.58	3347.57	125998.71
BIFI	317963.41	335909.38	28.34	284018.06	12353.41	867.31	32572.42
CERN-PROD_CLOUD	7441207.81	4618307.88	748.26	2028791.80	206220.29	13275.43	609617.97
CESNET-METACLOUD	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CETA-GRID	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DBCE	138418.61	9433.99	8.30	5389.89	160.14	37.88	6964.81
FZJ	99643.91	94402.51	10.91	118243.78	2887.58	373.50	18966.25
GRIDPP_CLOUD	65524.94	36181.48	7.31	2141.50	184.52	94.78	7222.18
HELIX_NEBULA_ATOS	561591.75	216889.48	35.41	31936.93	2911.78	555.35	29481.60
HELIX_NEBULA_CloudSigma	415.49	216.53	0.12	938.36	12.98	1.51	91.28
HELIX_NEBULA_TSystems	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IAAS	476845.83	317833.51	48.86	120680.72	13853.08	484.93	33280.22
IAAS_MCORE	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IISAS-FEDCLOUD	38012.69	42855.12	4.45	23003.16	673.19	143.24	5122.57
INFN-PADOVA-STACK	90948.47	110340.88	7.24	62807.18	3014.27	231.65	8332.18
SZTAKI	0.00	0.00	0.00	0.00	0.00	0.00	0.00
UKI-NORTHGRID-LANCS-HEP_CLOUD	77977.84	13496.45	8.01	28292.62	5011.09	82.51	2364.61

Sim@P1



- 2nd-most productive site in 2014
  - 64M CPU hours, 1.7B MC events
- Used for LHC stops > 24h
- Fast automated switching via web GUI for shifters
  - TDAQ to Sim@P1: 1h (check Nova DB, start VMs)
  - Sim@P1 to TDAQ: 12m (graceful VM shutdown, update DB)
  - Emergency switch to TDAQ: 100s (immediate termination)