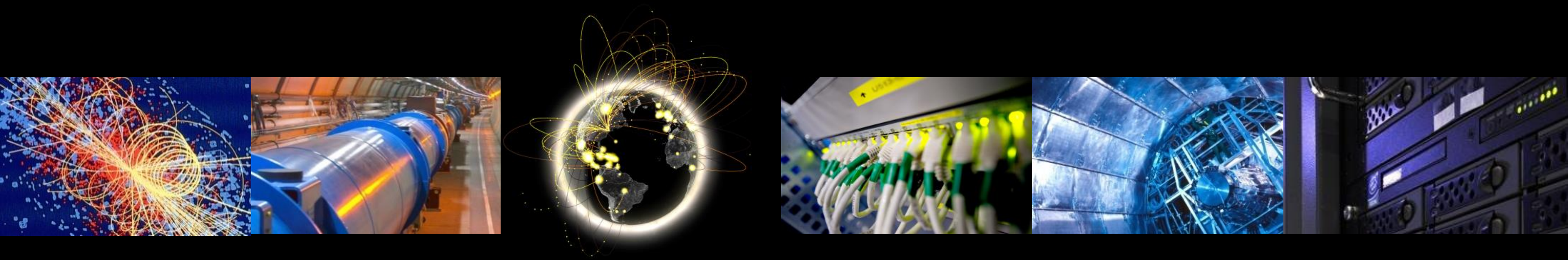
Accounting

John Gordon

WLC Workshop 2016, Lisbon



Outline

- Not a technical discussion of what is happening in the next couple of months.
- What is the long-term vision?

- Cloud and Grid
- Required Future Developments
- Viewing and Downloading

Why Accounting?

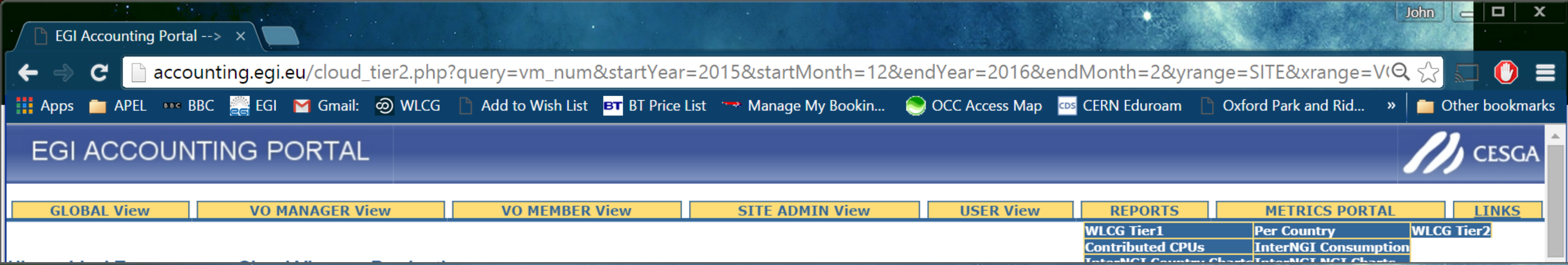
- Accounting should provide an independent, neutral record of resource usage from the point of view of:
  - User
  - VO
  - Site
  - e-infrastructure
  - ‘Management’ (Country, Project, other)
- Does your bank trust you to produce your own bank statements and balances?

Grid and/or Cloud

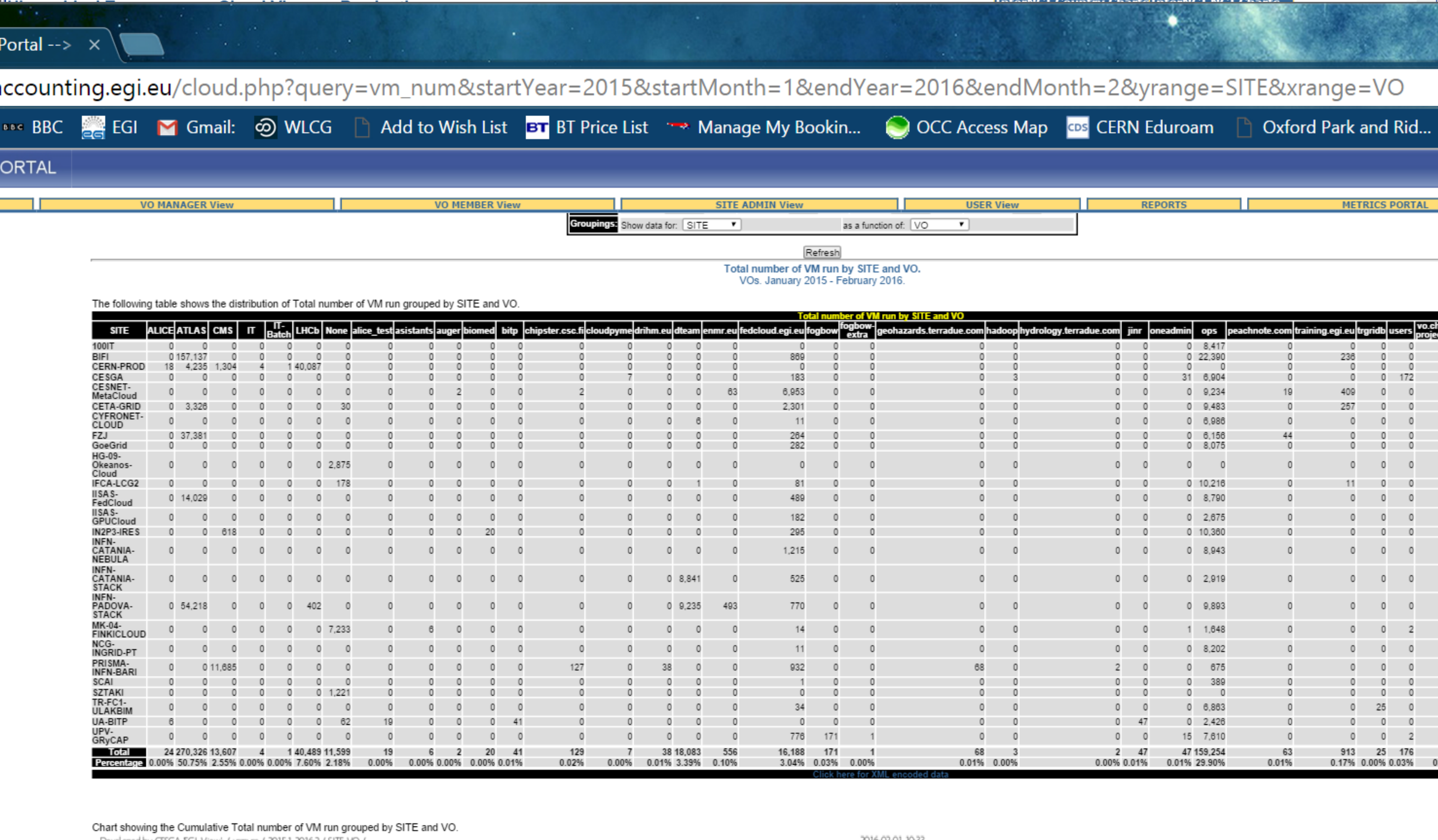
- While cloud accounting exists and is actively being developed, most existing use of cloud by LHC seems to be using VM-based batch workers which use traditional grid accounting.
    - This includes VAC and Condor. Is this going to change?
    - Experiment-based VMs which handle workloads like a pilot job does but run for a long time (a la Dirac) will bypass grid accounting and can/should use the cloud accounting of VMs.
    - If all work ends up running in some cloud then can we move to cloud-only accounting?
    - How to handle the grid+cloud mix?
    - How do we handle the public/private cloud division?
    - Is this future known or does it depend on other discussions at the workshop?

Cloud

- Tier2 view but only 5 T2s are reporting cloud usage to APEL. (no Tier1s). Of these only 1 runs LHC work. There is LHC usage at non T2 sites.
  - I know there are many tests using cloud infrastructures. Can more of them please report accounting of their VMs.
  - It is not necessary to join the EGI FedCloud but if you don't meet their criteria you may not be visible in EGI accounting, only in the WLCG views.
- The infrastructure is in place.
- Working on Monthly reporting (currently whole duration of VM gets accounted once)
- Biggest omission is cputime. I know one pays for wall but the user has a right to know what use they have made of the VM paid for.
- Issue – how to combine with commercial cloud usage.



accounting.egi.eu/cloud_tier2.php?query=vm_num&startYear=2015&startMonth=12&endYear=2016&endMonth=2&yrange=SITE&xrange=V(



ccounting.egi.eu/cloud.php?query=vm_num&startYear=2015&startMonth=1&endYear=2016&endMonth=2&yrange=SITE&xrange=VO

Groupings: Show data for: SITE as a function of: VO

Total number of VM run by SITE and VO.
VOs. January 2015 - February 2016.

The following table shows the distribution of Total number of VM run grouped by SITE and VO.

Total number of VM run by SITE and VO

SITE	ALICE	ATLAS	CMS	IT	IT-Batch	LHCb	None	alice_test	asistants	auger	biomed	bitp	chipster.csc.fi	cloudpyme	drihm.eu	dteam	enmr.eu	fedcloud.egi.eu	fogbow	fogbow-extra	geohazards.terradue.com	hadoop	hydrology.terradue.com	jinr	oneadmin	ops	peachnote.com	training.egi.eu	trgridb	users	vo.ch proje
100IT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8,417	0	0	0	0	
BIFI	0	157,137	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	869	0	0	0	0	0	0	0	22,390	0	236	0	0	
CERN-PROD	18	4,235	1,304	4	1	40,087	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
CESGA	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	183	0	0	0	3	0	0	31	6,904	0	0	0	172	
CESNET-MetaCloud	0	0	0	0	0	0	0	0	0	2	0	0	2	0	0	0	63	6,953	0	0	0	0	0	0	0	9,234	19	409	0	0	
CETA-GRID	0	3,326	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	2,301	0	0	0	0	0	0	0	9,483	0	257	0	0	
CYFRONET-CLOUD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	11	0	0	0	0	0	0	0	6,986	0	0	0	0	
FZJ	0	37,381	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	264	0	0	0	0	0	0	0	6,156	44	0	0	0	
GoeGrid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	282	0	0	0	0	0	0	0	8,075	0	0	0	0	
HG-09-Okeanos-Cloud	0	0	0	0	0	0	2,875	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
IFCA-LCG2	0	0	0	0	0	0	178	0	0	0	0	0	0	0	0	1	0	81	0	0	0	0	0	0	0	10,216	0	11	0	0	
IISAS-FedCloud	0	14,029	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	489	0	0	0	0	0	0	0	8,790	0	0	0	0	
IISAS-GPUCloud	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	182	0	0	0	0	0	0	0	2,675	0	0	0	0	
IN2P3-IRES	0	0	618	0	0	0	0	0	0	0	20	0	0	0	0	0	0	295	0	0	0	0	0	0	0	10,360	0	0	0	0	
INFN-CATANIA-NEBULA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,215	0	0	0	0	0	0	0	8,943	0	0	0	0	
INFN-CATANIA-STACK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8,841	0	525	0	0	0	0	0	0	0	2,919	0	0	0	0	
INFN-PADOVA-STACK	0	54,218	0	0	0	402	0	0	0	0	0	0	0	0	0	9,235	493	770	0	0	0	0	0	0	0	9,893	0	0	0	0	
MK-04-FINKICLOUD	0	0	0	0	0	0	7,233	0	6	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	1	1,648	0	0	0	2	
NCG-INGRID-PT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	8,202	0	0	0	0	
PRISMA-INFN-BARI	0	0	11,685	0	0	0	0	0	0	0	0	0	127	0	38	0	0	932	0	0	68	0	2	0	0	675	0	0	0	0	
SCAI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	389	0	0	0	0	
SZTAKI	0	0	0	0	0	0	1,221	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
TR-FC1-ULAKBIM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	0	0	0	0	0	0	0	6,863	0	0	25	0	
UA-BITP	6	0	0	0	0	0	62	19	0	0	0	41	0	0	0	0	0	0	0	0	0	0	0	47	0	2,426	0	0	0	0	
UPV-GRyCAP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	776	171	1	0	0	0	0	15	7,610	0	0	0	2	
Total	24	270,326	13,607	4	1	40,489	11,599	19	6	2	20	41	129	7	38	18,083	556	16,188	171	1	68	3	2	47	47	159,254	63	913	25	176	
Percentage	0.00%	50.75%	2.55%	0.00%	0.00%	7.60%	2.18%	0.00%	0.00%	0.00%	0.00%	0.01%	0.02%	0.00%	0.01%	3.39%	0.10%	3.04%	0.03%	0.00%	0.01%	0.00%	0.00%	0.01%	0.01%	29.90%	0.01%	0.17%	0.00%	0.03%	0

Click here for XML encoded data

Chart showing the Cumulative Total number of VM run grouped by SITE and VO.

Accessing Accounting Information

- The current portal allows limited data mining via 2-D views of a small number of parameters driven by an interactive web portal.
- The portal will develop a REST interface that will allow a more programmatic download of data into experiment (or other) tools.
- Experiments should be aware and can influence this development.
  - What do you want to download? In what format(s) do you want it?
- Dynamic access to low latency accounting for global allocations and real-time access control.

What Else Can We Account?

- Storage under development.
- Data Usage
- Many other fields which can be recorded but we don’t currently bother (I/o, networking, memory, ???)
- GPU? FPGA?
- Network

Other Issues

- Benchmarking
- Wallclock vs CPUtime
  - APEL currently collects and displays both.
  - A political decision

Benchmarking

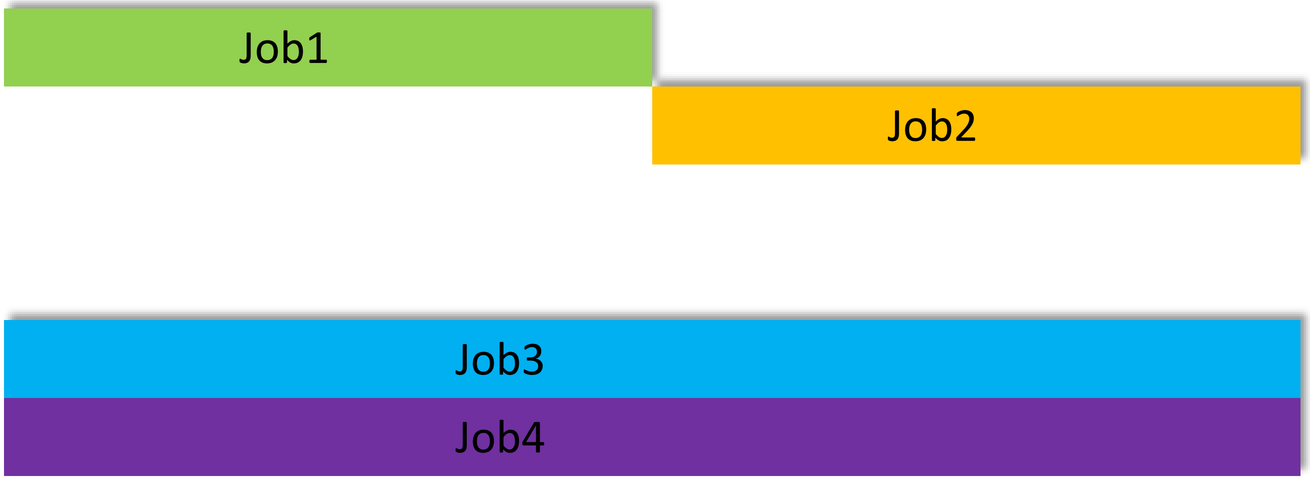
- APEL Repository needs benchmarking information to calculate normalised values for the accounting reports.
  - sites that send job records send us raw cpu, wall and benchmark. We normalize.
  - sites that send summaries normalize at their end and send us both raw and normalized cpu&wall.
  - Non-APEL clients gather data and populate the same schema.
  - In both cases the client obtains benchmark from TL BDII.
- Although APEL was designed to read SubCluster benchmarks these are overriden when a site's batch system scales its reported times. In this case (almost all sites) CPUScalingReference is used to normalise.
  - When a batch system scales cpu the results are exact for each WN. No error introduced by averaging benchmark over the cluster.
  - For systems that don't scale, (GE, LSF) the APEL parser uses the scale factor providedb to normalise
- APEL allows reporting one of a set of benchmarks. (SI2K, HS06)
  - This allows a smooth migration when changing but comparing data cross sites and time requires an agreed conversion.
  - In theory the UR could be extended to allow multiple benchmarks but the algorithms for handing, converting, etc would need to be clear and agreed.
- APEL benchmark retrieval is a simple query. Could be moved to an alternative source within the timescale of a client update at all sites.

Wallclock or CPU?

- APEL currently collects and displays both.
- No technical work to collect
- A political decision on what matters
- Reports would need reworking

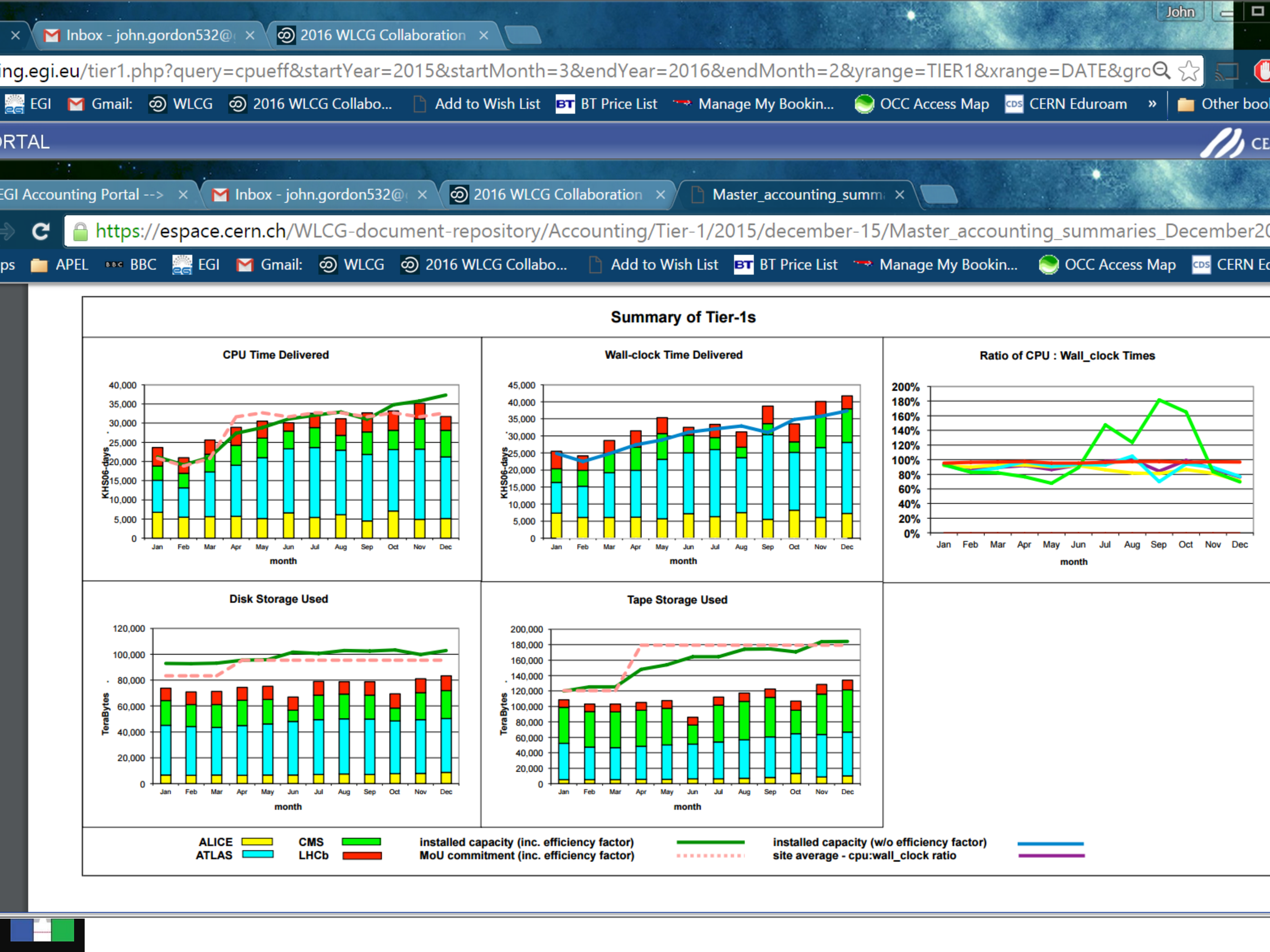
Discuss!

Wallclock and Overcommitment



Cpu=3	Wall=3
Cpu=3	Wall=3
CPU=6	**Wall=6**
Cpu=3	Wall=6
Cpu=3	Wall=6
CPU=6	**WALL=12**

- CPU is reproducible and measured by OS. Wall can change depending on conditions.
- Many reasons for overcommitment, not all planned. I/O, expedited jobs, low pri work.
- Licence to generate wallclock with the uncertainties that introduces.
- Can be managed by (eg) benchmark/jobslot but who can guarantee it will. Major variations will be spotted in efficiency, but will minor?



Summary of Tier-1s

Job Features

- MJF gathers benchmarking information from the resource and gives this to the payload. Is this consistent? Raw power/cpu or normalised by batch system?
- How? Sites supply $JOBFEATURES/hs06_job to each job so the information should be there to work it out from each host's HS06/processor rather than just working with cluster-wide averages.
- Some experiments like ATLAS use benchmarking information to calculate resource utilisation. How? From REBUS? Is it consistent?
- Who/what else uses benchmarking?