# WLCG Node in Pakistan – Challenges & Experiences

*Adeel-ur-Rehman Zafar*

***National Centre for Physics (NCP), Islamabad***

# Scheme of Presentation

- WLCG model and its Requirements
- Proposed Setup for a CMS Tier-2 Site
- Introduction of NCP-LCG2
- Challenges & Experiences (Phase-I)
- Challenges & Experiences (Phase-II)
- Performance Statistics
- Conclusions

# WLCG Figures and Tier-2 Requirements

# WLCG at a Glance

- WLCG/EGEE Grid infrastructure offers resources with a large amount of:
  - computing (around 90K CPUs[1])
  - storage (over 200PB[1])
- Rapidly growing setup
- Provides the natural environment for
  - large-scale physics
  - detector simulations
- Data Analysis requires efficient access to grid

[1] http://goc.grid.sinica.edu.tw/gstat/

# WLCG at a Glance

- At present, the WLCG collaboration is comprised of:
  - Over 270 sites
    - More than 50 countries
      - 5 continents[1]
- Four major experiments to constitute LHC
  - ATLAS
  - ALICE
  - CMS
  - LHCb
    - Each relying on WLCG setup
- Resultant data to be produced/year 10-20PB[2]

1 http://goc.grid.sinica.edu.tw/gstat/

2 F. Donno, M. Litmaath, “Data Management in WLCG and EGEE”, CERN, February – 2008

# Tier-2 Responsibilities

- Major data analysis to be performed by them
- Requirements include
  - Download data from corresponding Tier-1
  - Provision of managed disk storage
  - Provision of access to data stored by other centers of WLCG
  - Dealing with end-user analysis facility
  - Provision of other services e.g.
    - Data simulation
    - Ensuring network bandwidth and services for data exchange with Tier-1s

# CMS Tier-2 site – Proposed Setup

- According to WLCG C-RRB[1], in 2009 with 33 CMS Tier-2s, cumulative requirements are:
  - 112400 HEP-SPEC06 computational power[2] (around 46MSI2K according to WLCG pledged resources doc.)
  - 5700 TB of disk storage[2]
- Yielding average individual requirements as:
  - 3406 HEP-SPEC06 computational power (around 1.4MSI2K)
  - 173 TB of disk storage
- Where HEP-SPEC06 is:
  - A newer HEP-wide benchmark standard for measuring CPU performance
  - Based on C++ benchmarks from CPU2006 standard
- As far as the minimum required bandwidth between Tier-1 & Tier-2s are concerned, it is 1Gbps[3]

[1] Computing Resources Review Board, http://lcg.web.cern.ch/LCG/crrb.htm

[2] http://lcg.web.cern.ch/LCG/Resources/WLCGResources-2008-2013-05FEB2009.pdf

[3] http://cdsweb.cern.ch/record/838359/files/lhcc-2005-023.pdf, 20th June 2005.

# History of NCP-LCG2

# Emergence of NCP-LCG2

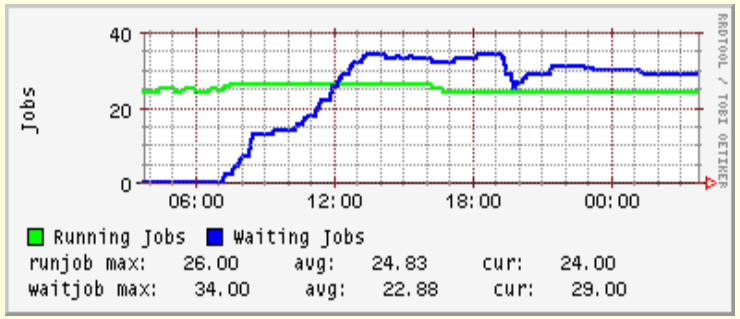
- The effort to bring Pakistan on the WLCG map as a Grid Node was started in October, 2003
- A Grid Technology Workshop was organized by NCP from October 20-22, 2003
- The first ever testbed was deployed during the workshop for tutorial
- LCG1 tag 1.1.1.2 was used for deployment
- This testbed consisted of 09 machines serving various functionalities required by a grid setup

# Emergence of NCP-LCG2

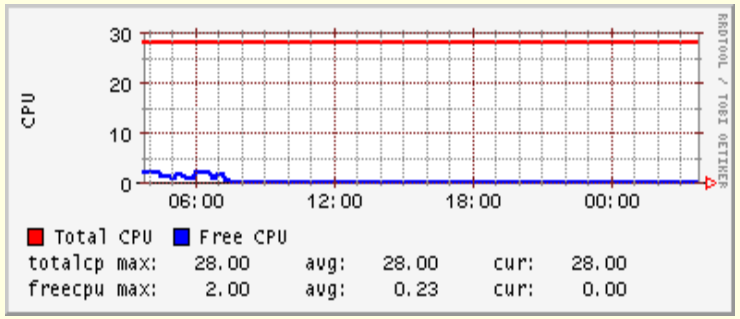
- 30 machines were used during the tutorial for enabling users to communicate with the deployed testbed
- The effort continued and NCP deployed a new LCG version which is LCG2 tag 2.0.0 in June 2004 and then to tag 2.2.0 in August 2004
- Since then, NCP is a Tested & Certified Grid Node in Pakistan
- Testing & Certification has been performed by the Grid Deployment Team (dteam) at CERN
- The node is constantly upgraded to the newer middleware versions, currently we are using
  - GLITE-3_1_0

# Emergence of NCP-LCG2

- NCP-LCG2 Node now appears on the Grid Operations Centre (GOC) website
  - A certified Tier-2 Grid Node, first in South Asia and fifth in Asia




**Running Jobs**




**CPU Usage**

# NCP at the WLCG Map

# *Challenges & Experiences Phase I*

# 2004 – 2006 (1/4)

- Focus was to establish a grid facility in the country to carry out physics data analysis efficiently
- Country-wide noteworthy issue:
  - Ideal level of grid resources not easily available
  - If found, cost was very high
- Overall infrastructure was very weak w.r.t. Tier-2 requirements especially for network
  - Started from 128Kbps
  - No additional SAN
    - Only SE hard-disks -> around 80GB ONLY
  - 512MB – 1GB memory considered sufficient w.r.t. CERN Scientific Linux based on Redhat kernel 7.3
- The setup employed all non-branded machines

# 2004 – 2006 (2/4)

- Due to significantly limited network pipe,
  - Some of the critical network bound services were difficult to deploy locally
  - Utilized those service nodes externally (from CERN) such as
    - RB, top-BDII, ProxyServer
- In the meantime, network bandwidth got increased to 1Mbps
  - Equally shared b/w user & grid data
  - Still not enough to resume the highly network oriented services locally

# 2004 – 2006 (3/4)

- Meanwhile our site unexpectedly began facing abnormal behaviors during job processing
  - Jobs hanging during execution
  - Untraceability of the jobs being run
  - Node-types becoming inaccessible (especially CE)
  - Site failures reported intermittently
- Conducted several hit-n-trials on CE, covering:
  - Average-Load inspection
  - Possibility of unsynchronized time on clients
  - Mis-configuration of important services
- Finally, it got revealed that the culprit was incompatible RAM

# 2004 – 2006 (4/4)

- During the troubleshooting, by chance, we chose to deploy a branded CE machine
  - The irregular messy behavior suddenly got vanished
  - Site monitoring tests passed successfully
  - The situation never got reproduced at CE
  - Site performance was greatly improved
- The conclusion turned out to be hardware incompatibility issues with the OS distribution deployed
  - Particularly if the hardware is not well tested against compatibility with other components and OS release
  - A promising decision has been then taken to shift our Grid Node to a completely branded set of machines

# *Challenges & Experiences Phase II*

# 2007 – Present (1/4)

- Grid Computing is based on Commodity Computing
  - A farm of branded machines is another form of commodity computing
  - The new NCP-LCG2 setup still obeys the model
- Network connectivity reached 2Mbps
  - Initially shared b/w grid and user traffic
  - Still far behind; Tier-2 setup advises in GBs

# 2007 – Present (2/4)

- With the passage of time,
  - Manpower at NCP kept on increasing
  - 2Mbps pipe caused choking the bandwidth
  - Joint nature of network transformed into
    - 75% of total dedicated for grid traffic
    - 25% for the general users
- It was observed that,
  - still we can't deploy back the externally used grid services
- Due to lack of sufficient hardware and network resources, we couldn't provide more than around 60 KSI2K computational power (1 WN corresponds to 1.2 KSI2K)
- NCP shifted to its new campus premises
  - Ultimately, the network connectivity jumped to 10Mbps
  - A milestone achievement for us due to very expensive infrastructure prices
  - 8Mbps out of10Mbps dedicated for grid facility

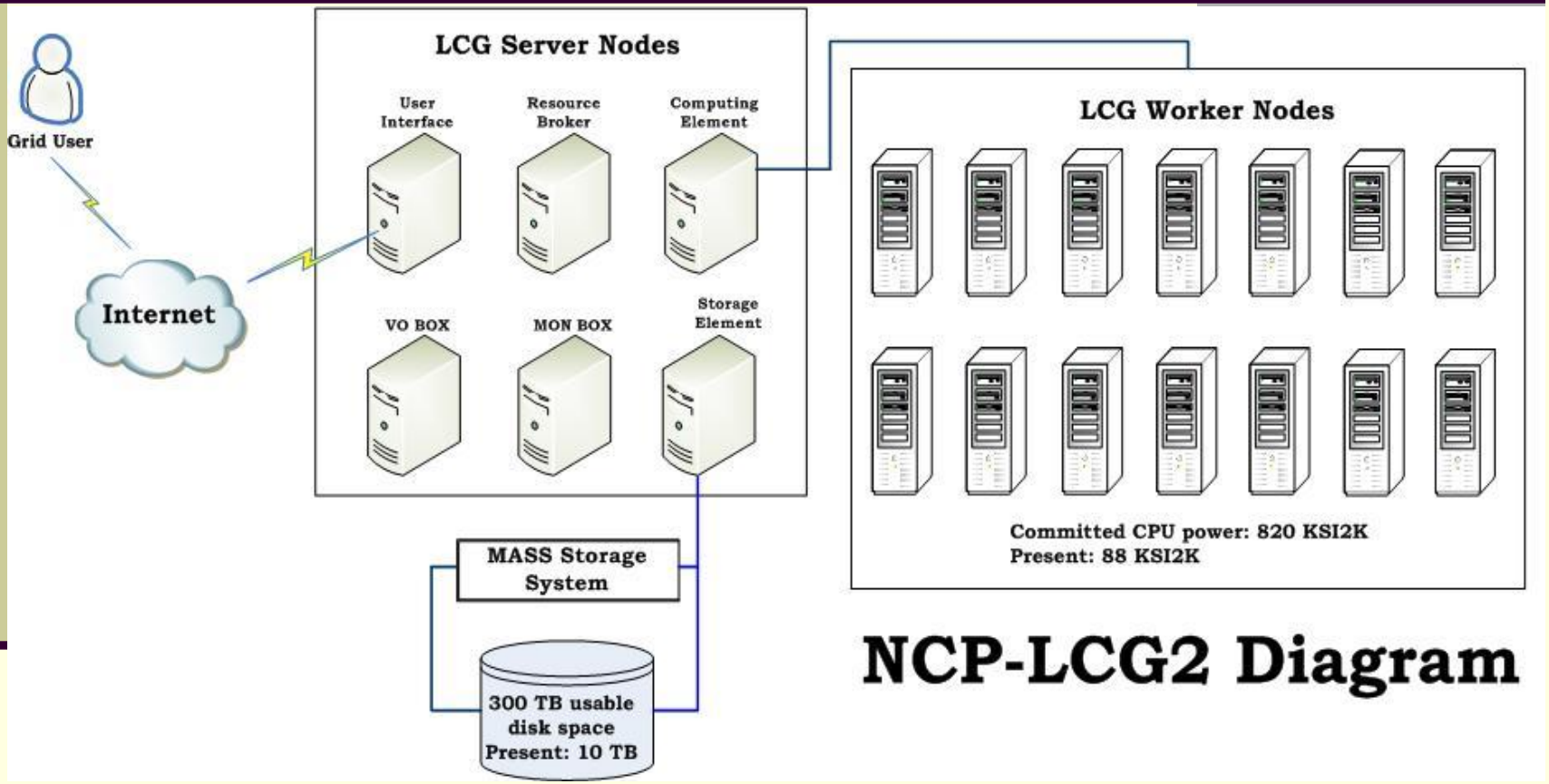
# 2007 – Present (3/4)

- Then it became viable to get independent of CERN in terms of utilizing the network oriented services
  - The situation got much stabled for RB, top-BDII, ProxyServer
- Around the same time, Pakistan hassled dreadfully by the severe power-cuts on daily basis
  - Extensive load-shedding harmfully impacted the battery-backups for power generation – not gaining enough time to get fully charged
- Moreover, due to multiple times abrupt power cuts, certain services got severely damaged and we had to re-install some of the nodes including CE and WNs

# 2007 – Present (4/4)

- NCP-LCG2, sporadically faced network disruptions due to
  - Configuration issues at ISP end
  - Packet losses due to transition b/w designated fibre and copper wire till the destination (applicable when in old premises)
- The day-by-day increase in grid sites within WLCG/EGEE infrastructure
  - Deteriorated our top-BDII efficiency while manipulating LDAP URLs for other top-BDIIs
  - Also turned out to be the root-cause of Replica Management failures i.e. CE-sft-lcg-rm-rep test
  - Attempted to adjust important BDII related time-outs but in vain
  - At this point, Taiwan's top-BDII was exercised leaving ours unused once again which was recovered until the middleware was almost fully upgraded to version glite-3.1.0
- So it was concluded that, with a quite reasonable number of grid sites around the collaboration, the slower network speed could be error-prone. But with 10Mbps along with applying in time update patches, services could be run in a much stable manner.

# WLCG Setup at NCP Network



# NCP-LCG2, Performance Statistics

# NCP-LCG2 Performance

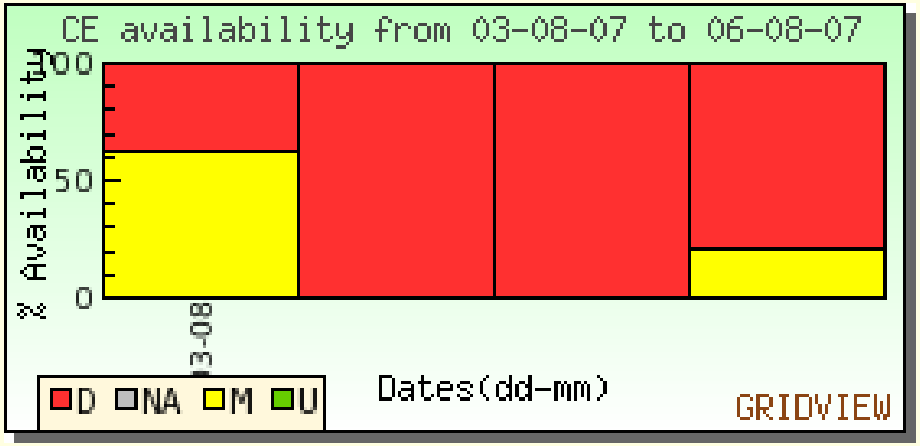
- There exist a number of monitoring tools to watch sites' performances in a desired span of time
  - Tools could be either managed centrally or
  - Help evaluating site statistics locally
  - A few of those tools along with the visibility of NCP-LCG2 within those are presented next covering SAM, APEL and EGEE availability & reliability statistics

# Service Availability Monitoring

- SAM is a successor of
  - TestZoneTests
  - SiteFunctional Tests 1
  - SiteFunctional Tests 2
    - Initial steps towards availability testing of sites
    - Based on bash scripts for data processing
    - Incorporated Perl CGI for results generation
- To cope with the limitations of its predecessors, SAM was devised covering
  - Improved database schema
  - Introduced the concept of sensors as containers of tests based upon grid service types
  - Integration with other grid monitoring tools (e.g., SAM & GridView)
  - Automatic availability metrics calculation w.r.t. various parameters such as VO, date, site etc.

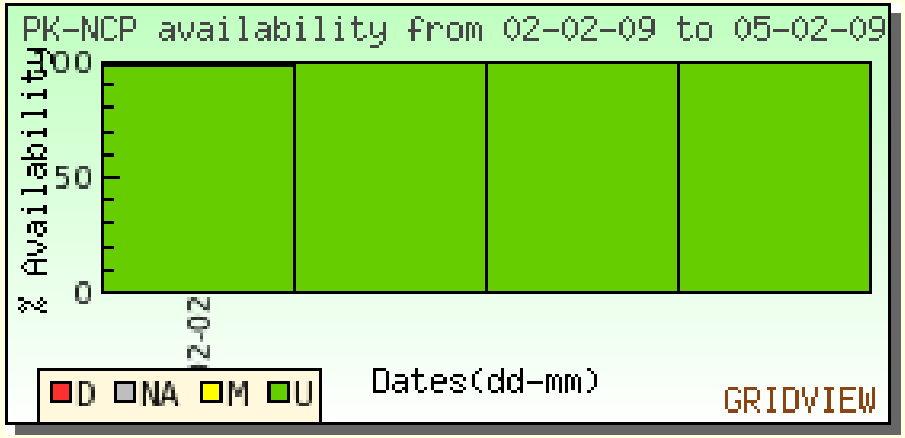
# SAM Graphs for NCP-LCG2

- **Problematic Situation due to Job submission and network limitation problems**
- Test Criticality Defining VO: OPS



# SAM Graphs for NCP-LCG2

- **Stable situation in the current year after the improved infrastructure**
- Test Criticality Defining VO: OPS

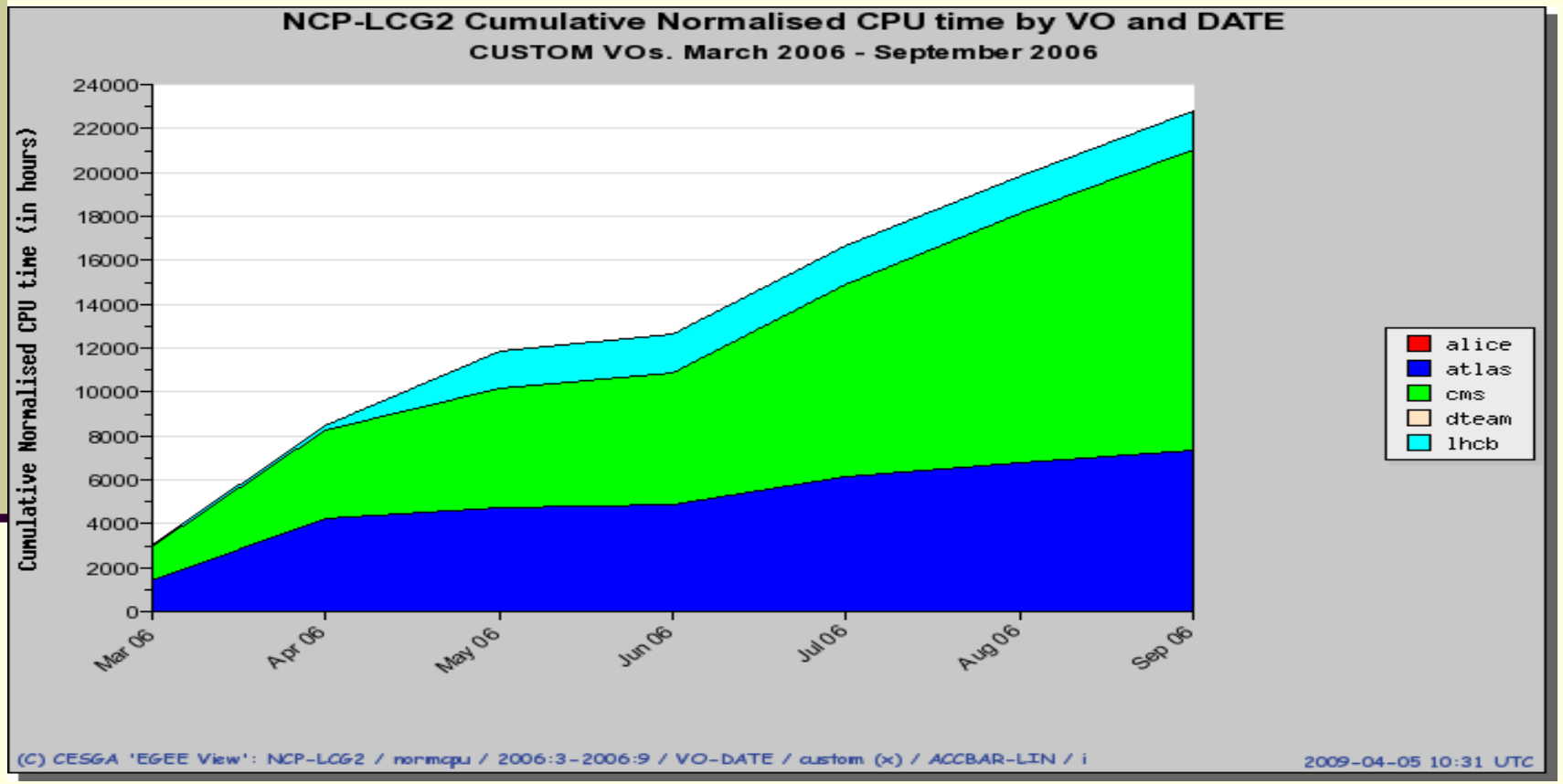


# APEL & RGMA

- APEL stands for **A**ccounting **P**rocessor for **E**vent **L**ogs
- Internally a log-processing application used to interpret gatekeeper and batch server logs
  - To produce CPU job accounting records
- Job based resource usage metrics including
  - CPU time
  - Wall Clock Time
  - Memory
  - Grid User DN etc.
    - Formed into a single usage record in the WLCG accounting schema
- The parsed records are then
  - Published into the R-GMA
  - Archived in a GOCDB (centralized relational database)

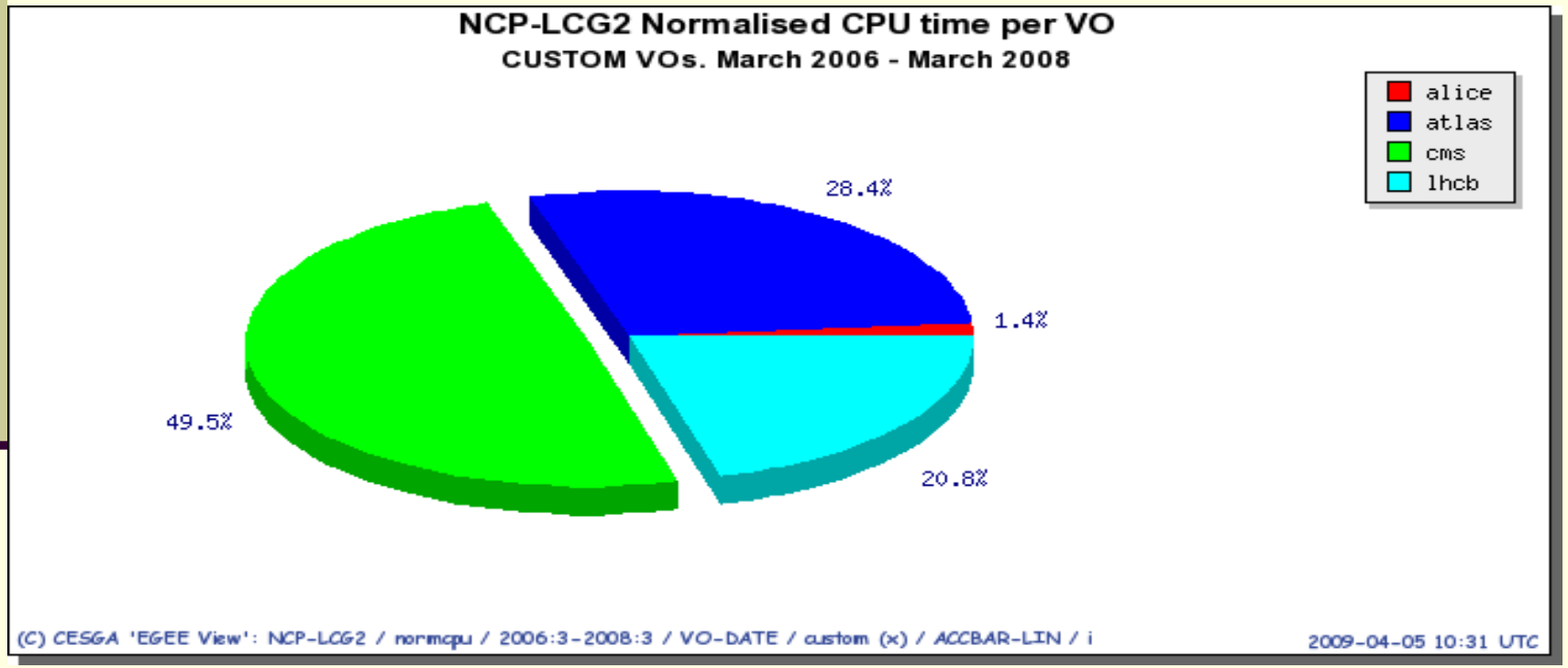
# APEL graphs for NCP-LCG2

- Depicting extensive CMS job processing in 2006



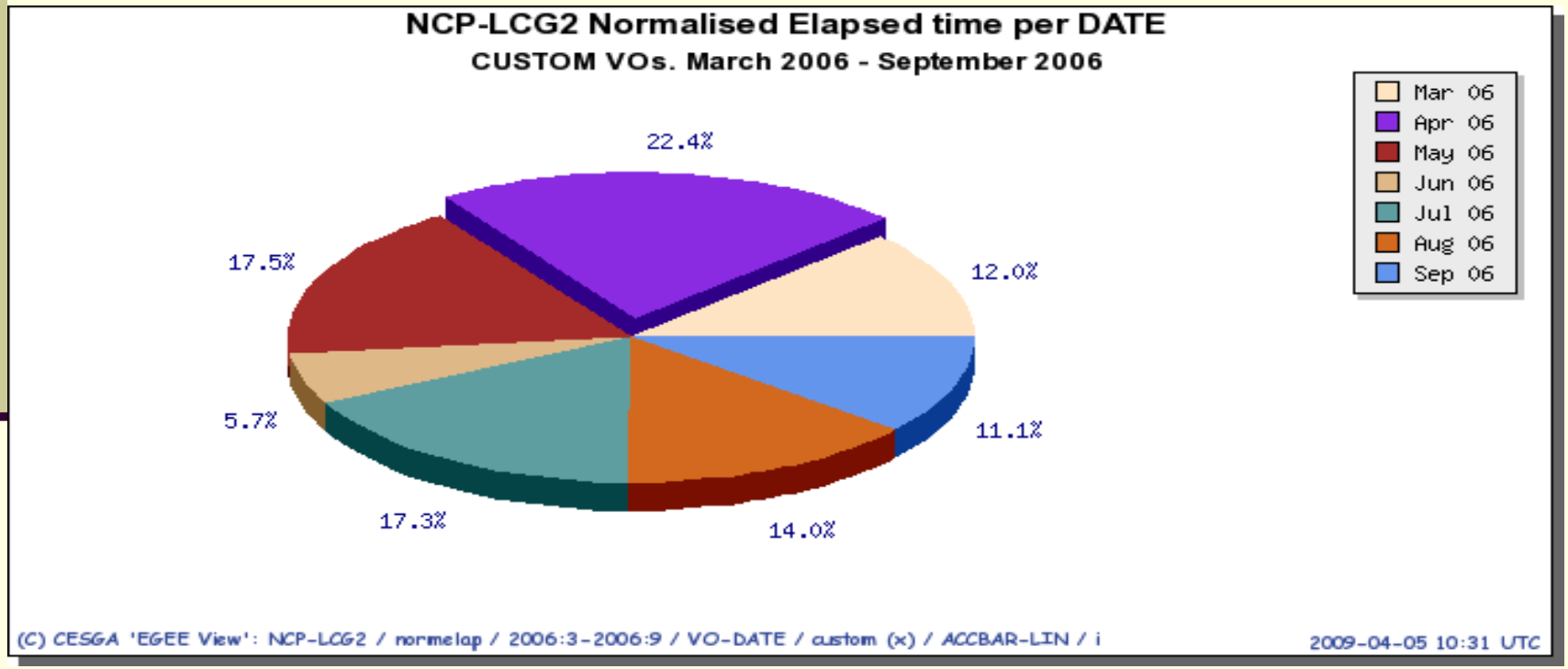
# APEL graphs for NCP-LCG2

- Ratio of CMS jobs allocated CPU time vs. other VOs b/w 2006 - 2008



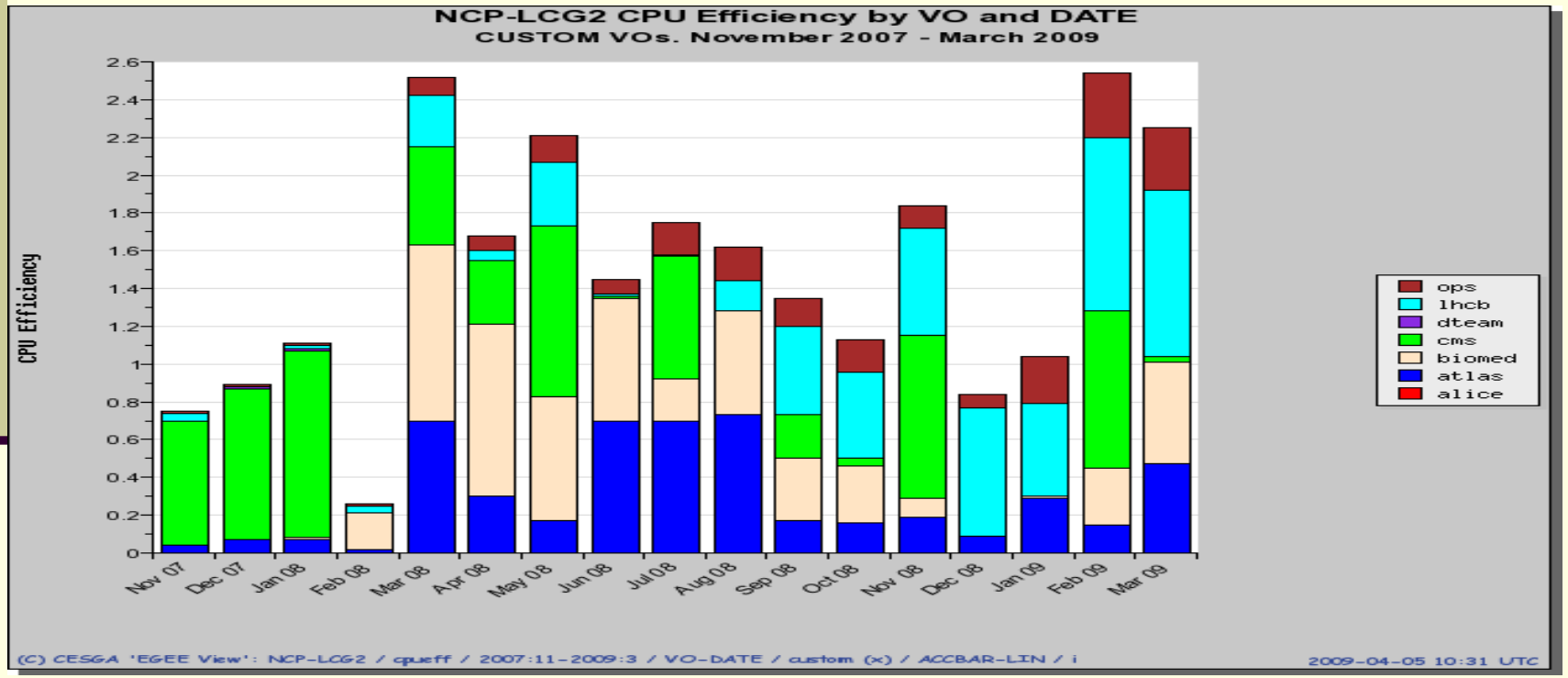
# APEL graphs for NCP-LCG2

- Normalized Elapsed time by VO during the same period covering mid of 2006



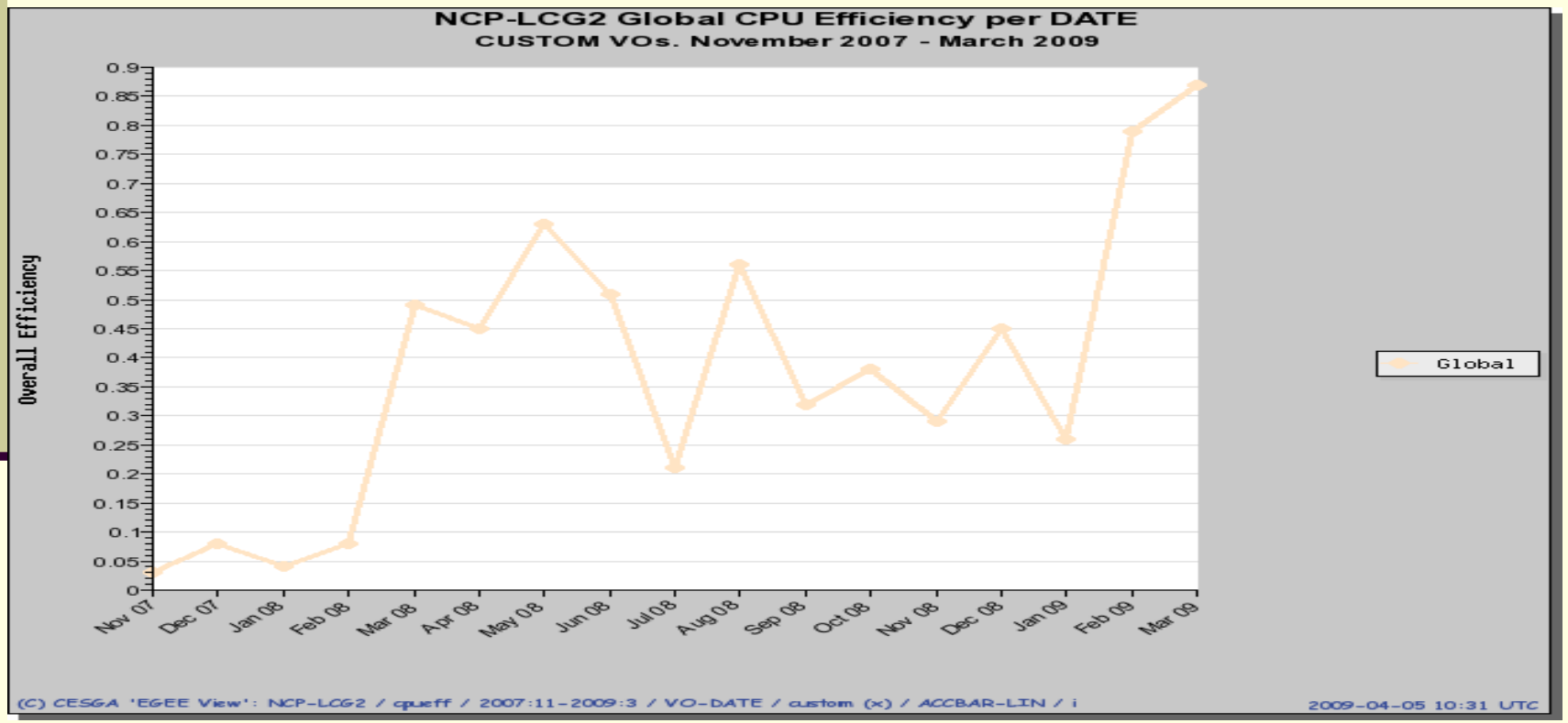
# APEL graphs for NCP-LCG2

- CPU Efficiency from end of 2007 till present by VO



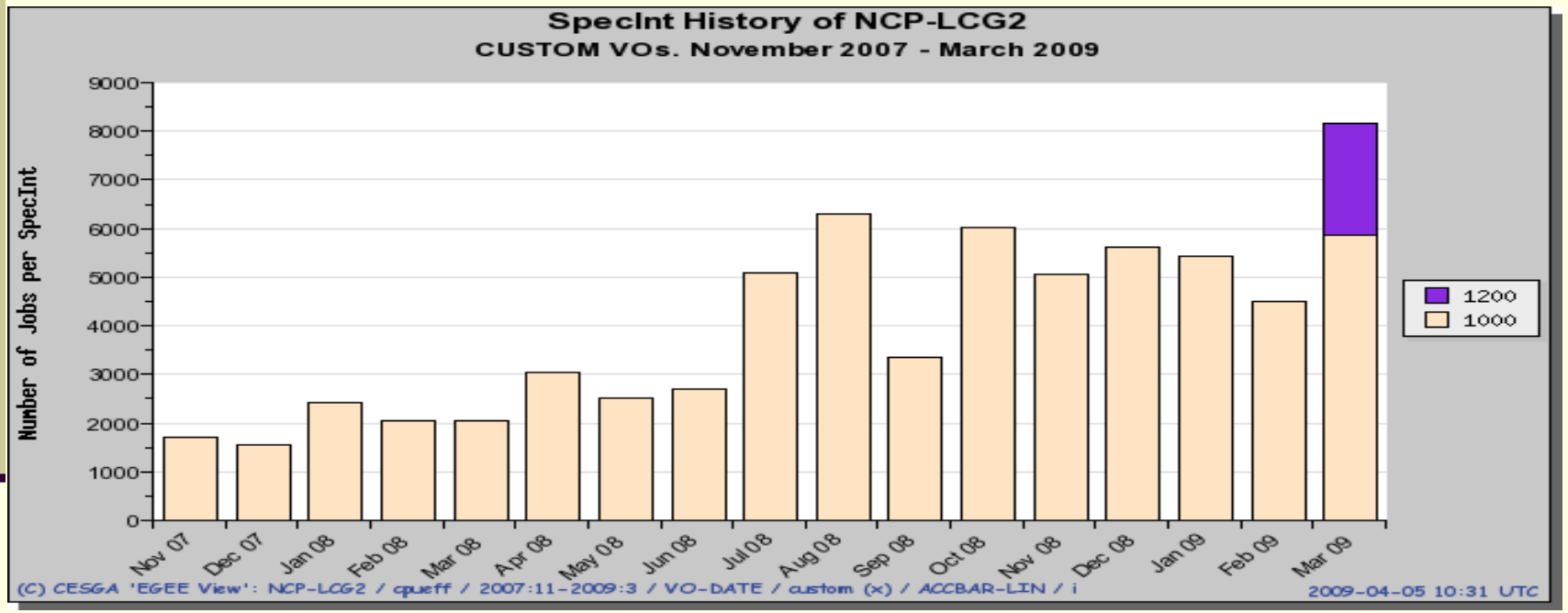
# APEL graphs for NCP-LCG2

- Progress of CPU Efficiency from end of 2007 till present by VO



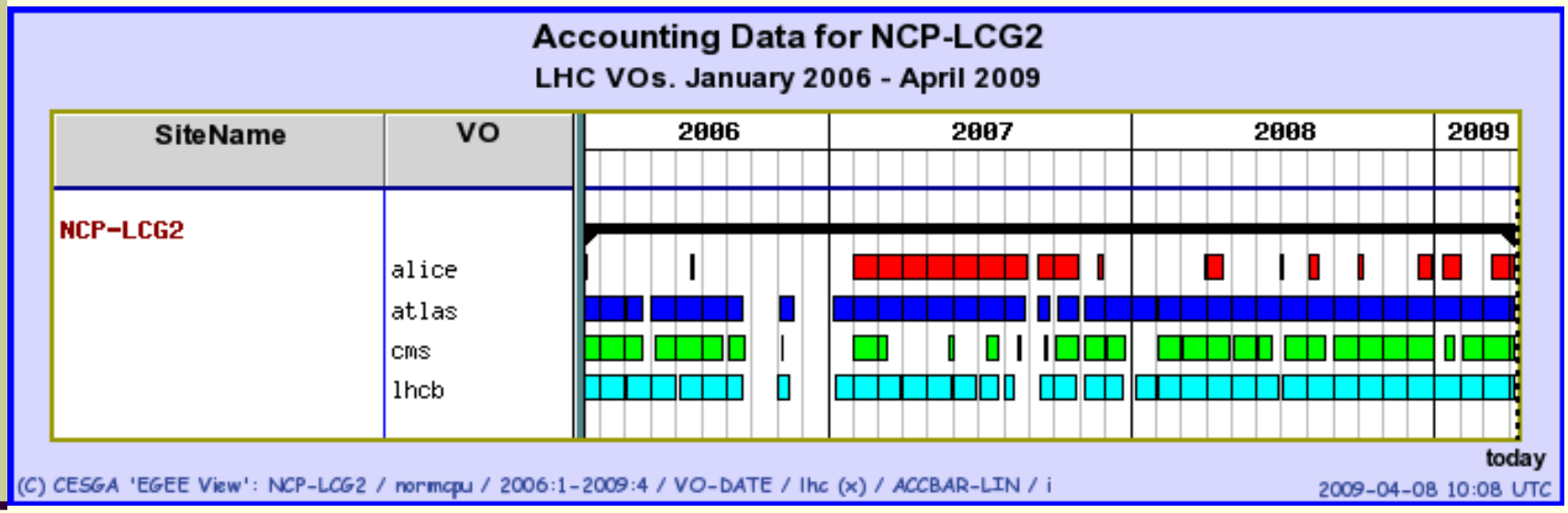
# APEL graphs for NCP-LCG2

- SpecInt History from end of 2007 till present



# APEL graphs for NCP-LCG2

- Accounting data from 2006 - 2009



# EGEE Availability & Reliability Statistics

- EGEE measures grid accounting information in terms of following two important metrics:
  - Site Availability
    - % of successful tests for a site in a given time span
  - Site Reliability
    - Site Availability / Scheduled Availability
- Site Availability above 90% is considered excellent
- Sites are notified about their ratings on monthly basis from their ROC

# EGEE Availability & Reliability Statistics

- Under the tough times, NCP-LCG2 reflected 0% Availability and therefore 0% Reliability
- Recently, under enhanced setup, NCP-LCG2 reached a landmark of over 95% availability as well as reliability
- The average figures are around 70% now-a-days as per published statistics
- Unfortunately, graphs older than 2006 are not available to reflect the true picture of our Phase-I of grid deployment

# Conclusion

- NCP-LCG2 is running by NCP since third quarter of 2004
- Initial grid infrastructure available was fairly insubstantial
- At that time, NCP-LCG2 faced multiple times with chaotic situations, in terms of job submission, network pipe limitation, packet loss etc.
- The effort kept continued in order to perform healthier which lead to gradual expansion of infrastructure
- w.r.t. challenges met and the experience gained in resolving those, it could be inferred that in case of non-branded machines deployed in the setup provided that:
  - hardware of which is not properly tested and
  - for which thorough compatibility testing has not been performed
    - then those may not be suitable for running such a setup

# Conclusion

- In spite of having a weaker setup, once could represent its visibility in WLCG map
- Even with a limited setup, a contribution towards productivity of EGEE/WLCG could be made possible
- With the briskly growing number of WLCG sites, weaker sites do not get a large chunk of real jobs
  - thus not posing much of a performance challenge to WLCG world by their existence

***Thank You!!***

***Questions are Welcome***