



# ASGC Site Report

Felix Lee

HEPiX Spring, 2011



# Outline

- Resource update.
- Storage system re-configuration
  - Castor upgrade
- 10GbE cluster.
- Cluster/Distributed file system survey
- e-Science activities
- Plan

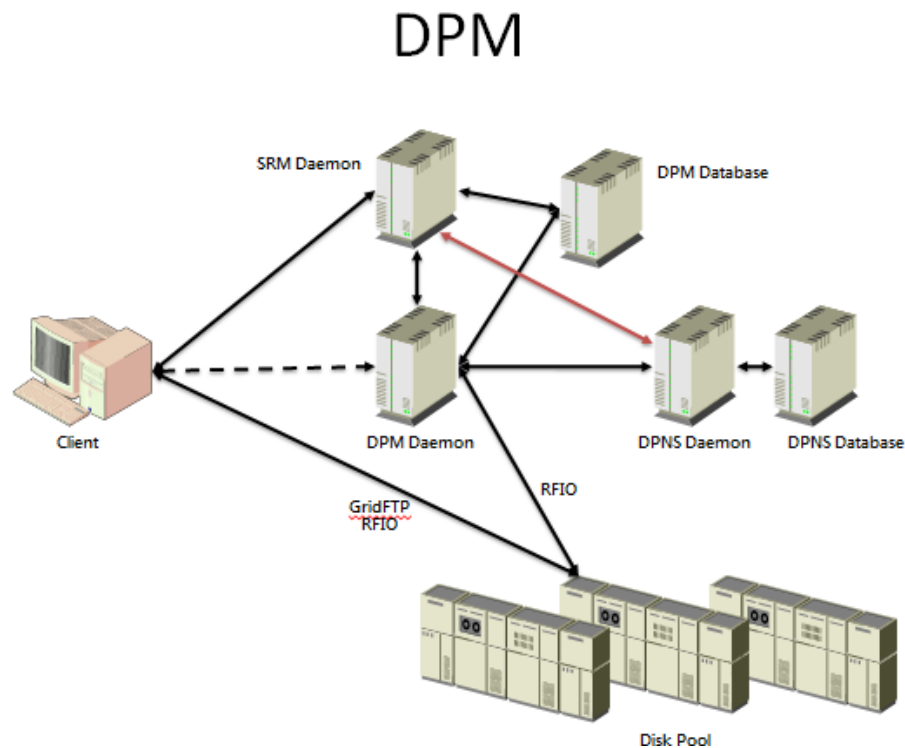
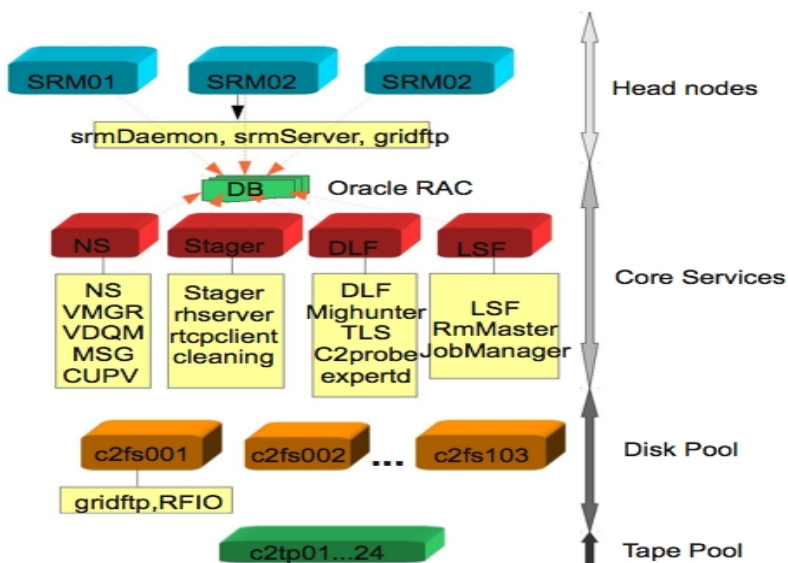


# Current Resource Status

- New 2.3PB disk has been delivered in Jan 2011.
- Purchased 8 \* 10GbE extreme switch to build up 10GbE cluster.
- Added 150 six core nodes to support e-Science.

Resource Groups	CPU Cores	Disk (TB)	Tape (TB)	Inter-Conn	User Groups
WLCG, Life Science, Earth Science, Environmental Changes, Social Simulation & general HPC, E-Science	4,600	4,000	4,000	Ethernet	WLCG, TWGrid
	4,488	700	0	10G Ethe + IB (DDR/QDR)	Earth, Env. Changes, EUAsiaGrid Astronomy
	2,369	470	0	Ethernet	Cloud, Other e-Science

# ASGC WLCG Storage System



1. Discussed and made decision during the CHEP with Atlas
2. CASTOR for RAW Data, DPM for AOD/ESD
3. Optimize DPM for more stable service



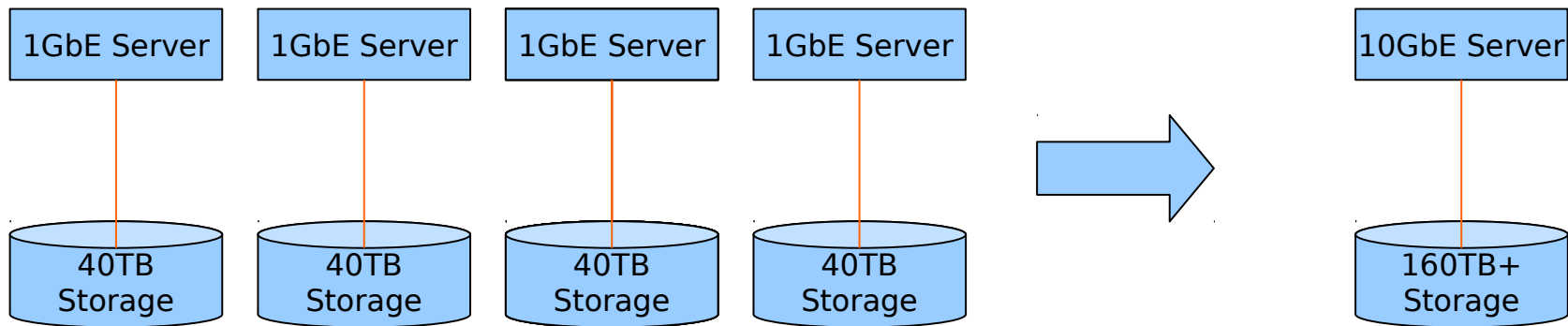
# ReConfiguration of storage

- Merging ATLAS D1T0 storage class to DPM
  - Reduce unnecessary data transmission between ASGC and TW-FTT (which are actually located at the same data centre) – \*.DISK (MCDISK, DATADISK, SCRATCHDISK, HOTDISK)
- Take new delivered 2.3 PB storage as buffer for Castor to DPM data migration.
- DPM Optimization – Configure multiple dpm SRM.
- CASTOR Upgrade to 2.1.10
  - Upgrade to 2.1.10 in April, 2011
  - It's eventually done!



# ReConfiguration of storage

- To support more disk space with single server.
  - 1 disk server handles more disk capacity.
  - DPM has already migrated to this configuration.
  - Castor is on going.

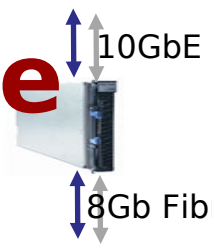


- Why we do this?

- To save rack space, money and so do power consumption.
- In past, we used 7 blade chassis to serve 2PB, now it only takes 1 chassis!



# Disk server Performance

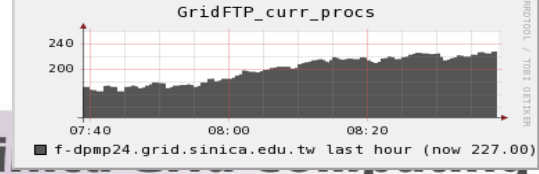
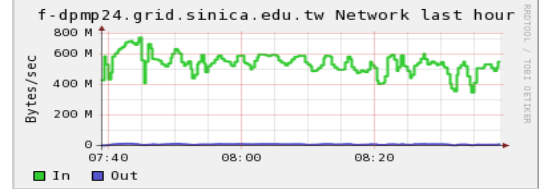
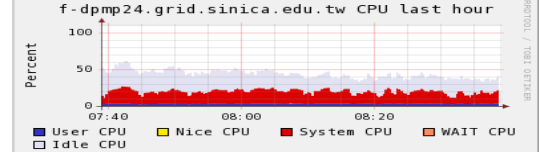
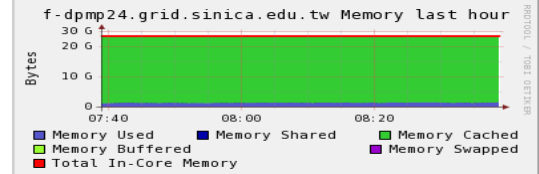
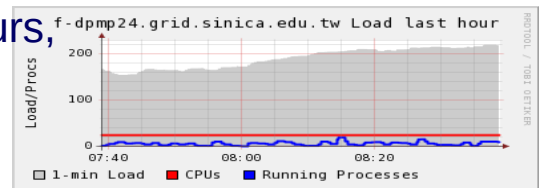
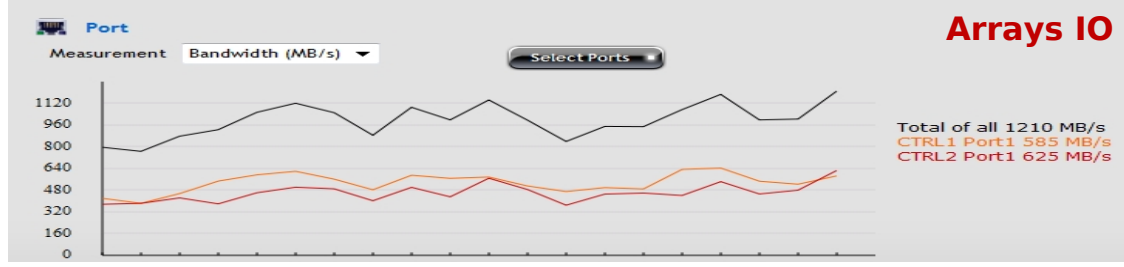


Observation while testing by ATLAS data transfers

- Peak network ~ **950MB/s**
- Peak arrays IO ~ **1210MB/s**
- Completed **4916** file transfers with **6.66TB** data flow in few hours,  
**2** errors (1 SRM\_FAILURE; 1 globus\_xio )
- It can handle more than **200** concurrent GridFTP transfers



Time	IFACE	rxpck/s	txpck/s	rxbyt/s	txbyt/s	rxcmp/s	txcmp/s	rxmst/s
07:45:30 AM	eth2	629640.59	56313.86	952530045.54	3041432.67	0.00	0.00	2.97
07:45:31 AM	eth2	586384.00	46396.00	887168556.00	2512998.00	0.00	0.00	2.00
07:45:32 AM	eth2	572838.00	57892.00	866534085.00	3126657.00	0.00	0.00	3.00
07:45:33 AM	eth2	574588.89	50842.42	869258776.77	2756614.14	0.00	0.00	0.00
07:45:34 AM	eth2	666344.55	87286.14	<b>1007847553.47</b>	4713586.14	0.00	0.00	0.99
07:45:35 AM	eth2	545350.00	90933.00	824455144.00	4942066.00	0.00	0.00	0.00
07:45:36 AM	eth2	399546.00	93752.00	604046130.00	5077206.00	0.00	0.00	1.00





# 10GbE Cluster

- Our first try.
- Use 150 IBM blades
  - Six core Intel L5640 with 24 GB Memory.
- Extreme x650 edge switch
  - Stack 8 \* switch.
- Performance is not quite good against IB, but it's acceptable.
- Cluster is a bit unstable due to IBM blade hardware issue.



# 10GbE Cluster

Nodes	1	2	4	8	12	14	28	42	54
Num_Core	12	24	48	96	144	168	336	504	648
Time	183m10s	86m10s	44m38s	21m37s	12m40s	10m17s	5m8s	3m43s	2m49s
Mutiple	1	2.125	4.103	8.473	14.46	17.811	35.681	50.282	65.029
Ideal	1	2	4	8	12	14	28	42	54



# Cluster/Distributed FS

- Ceph survey.
  - Features
    - MDS HA is available.
    - Data replication is also available.
  - But,
    - It's currently still buggy.
    - The core filesystem “btrfs” with Ceph also has bug.
    - MDS and OSD will get hang by unknown reason.



# Cluster/Distributed FS

- pNFS survey
  - Looking for available open source solution.
    - EXOFS, PVFS2..
  - EXOFS
    - Takes pNFS object layout as MDS
    - Can reach 450MB/s write throughput with 7DS and 7 clients.
      - It only gets 350MB/s with read, but over 1GB/s with re-read.
    - It's a bit unstable with openOSD while single client produces large I/O request.
      - Still investigating openOSD.
  - PVFS2
    - Takes pNFS file layout as MDS.
    - It's on going..



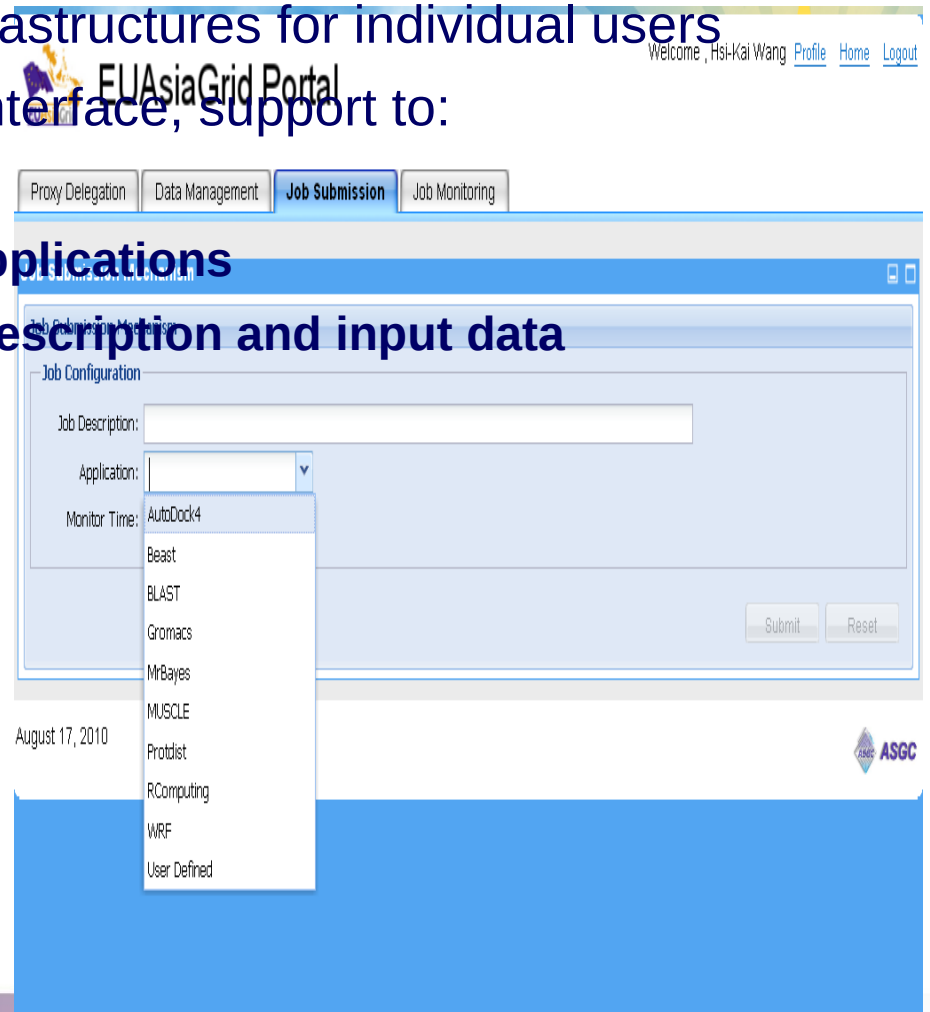
# Cloud/virtualization activities in TW

- Cloud controller(VMM) as CE.
  - Handling virtual machine as pilot worker node.
- Open nebula integration.
  - Developing tm\_gsirfio to allow open nebula to access gLite-dpm by gsirfio.
    - Uses dpm as VM repository.
    - It currently only uses single user proxy to do gsirfio.
- Evaluating Openstack.
- SL6 with KVM hypervisor
  - Performance is not that bad with virtio
    - SLC5 quest with 64bit CPU emulator and 1Gb memory.
    - 879 Mbits/sec with iperf.
    - 130MB/s write, 59.4 MB/s read with qcow2 + virtio (vm image is running on pNFS)
- VM image distributing in Asia-pacific
  - gLite service VM image provision and distributing.
  - Split 1GB image to be 10 \* 100MB files and user uses ftp to download it.
    - The efficiency is quite bad, 1GB takes 3 days to transfer from TW to Singapore.
    - Looking for p2p network, and so do cooperation with VMIC team.



# eScience Portal

- Convenient access to grid infrastructures for individual users
- Provides, through the portal interface, support to:
  - **Submission of jobs**
    - Specific forms for individual **applications**
    - **Helping to prepare the job description and input data**
    - Data management
    - Allow sharing with other users
    - Job Monitoring

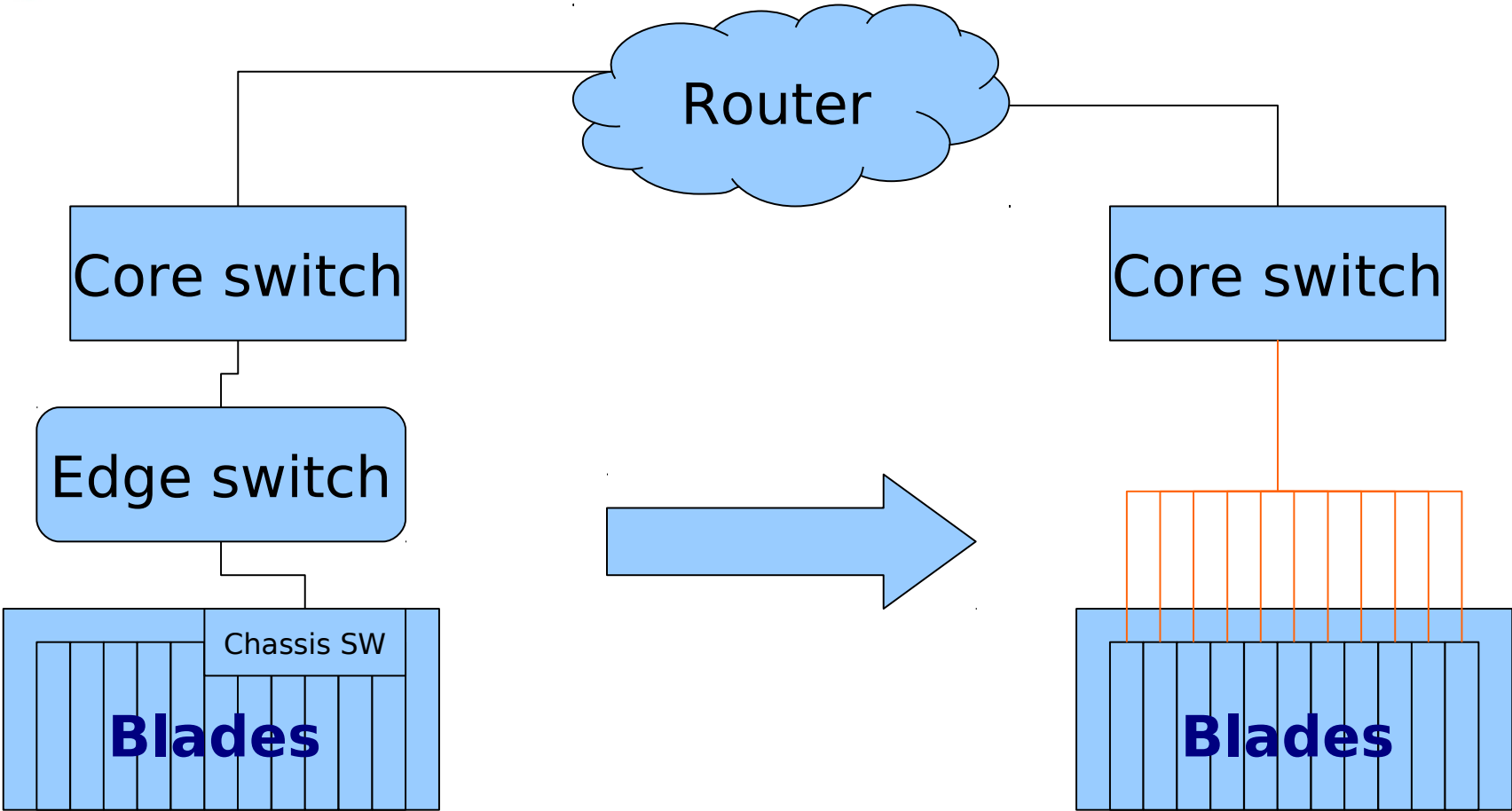




# Future plan

- pNFS implementation.
  - Looking for consultant from Desy dCache and cooperation opportunities with CERN dpm team.
- 10GbE Data Center implementation
  - Procuring 512+ ports 10GbE core switch.
  - Designing two tiers 10GbE LAN.
    - Bypass blade chassis switch by using pass-through module.
    - 10GbE fibre cabling/patch panel are also implementing.

# Future plan





Thank you very much