New infrastructure for the Tokyo ALTAS Tier-2

H. Matsunaga

I. Ueda

(ICEPP, The University of Tokyo)

Introduction

- Tier-2 dedicated to ATLAS
  - the system optimized to the ATLAS computing
- 3-year lease contract for the whole computer system (including system engineers)
  - 2007-2009 for the previous system
  - 2010-2012 for the current system
- Switchover of the whole system every 3 years
  - the electrical and the cooling systems stay the same

Procurements

- Policy is 3-year lease for the whole computer system (including contract work by system engineers)
  - lower price by a big tendering
  - 2007-2009 for the previous system
  - 2010-2012 for the current system
  - Need to replace the whole system at a time
    - The tape library was retained in the last replacement
    - Continue using the UPS and the cooling systems
  - Bidding was conducted last summer

System migration

- During the replacement, reserved resources in another computer room was used temporarily (~1 month)
  - Much less CPU power (with old hardware)
  - Data storage system in the Tier-2 has to migrate twice
    - 200~300 Tbytes of data files
    - Manual alteration of the tables for the DPM database (MySQL)
- All the hardware (inc. racks and cables) was replaced in last November
  - Completed within ~1 week
  - OS installation, configuration, and burn-in tests was partially done in advance outside the University
  - Started production use one month later

The new system

- Same concept as the current system
  - LCG (Tier-2) part and Non-LCG part are logically divided
    - Still good network connection between them
  - Blade server for CPU resource
    - good density and maintainability
  - External RAID Box with Fibre-Channel link for disk storage
    - Good performance (and experience)

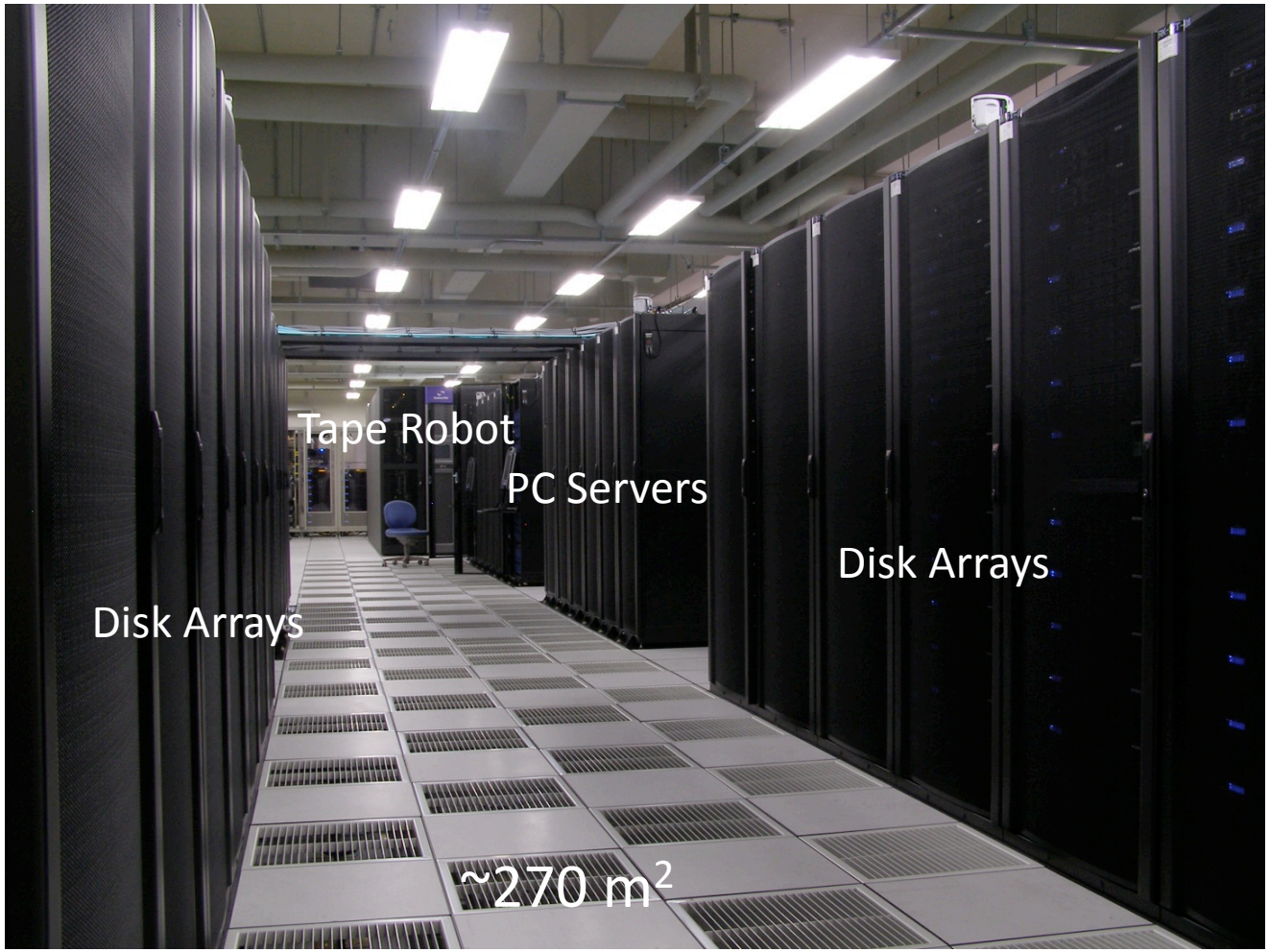
Enhancements

- Technology evolution over the last 3 years
  - More powerful CPU and more cores
  - Faster RAID controller and larger hard disk drive
- 10 Gb Ethernet NIC for worker nodes
  - But oversubscribed into the switch. At most 2-3 Gb per node available when used concurrently.
- Reduce bottlenecks at the interlinks between Ethernet switches
  - More of this later

Component details

- Blade server
  - CPU: WoodCrest (dual core) to Nehalem-EP (quad core)
  - RAM: 8 Gytes to 16 Gbytes
  - HDD: 73 Gbytes (mirrored) to 300 Gbytes (mirrored)
- Disk array
  - 16 slots to 24 slots (faster RAID controller)
  - HDD: 500 Gbytes to 2 Tbytes
  - Fibre Channel: 4 Gbps to 8 Gbps
  - 2 disk arrays/server (5 in the previous system)
- Ethernet switches
  - 3 x RX-16 (16 cards/switch) to 2 x RX-32 (32 cards/switch)
- Tape system
  - LTO3 to LTO4 drives (in part)

Old system



New system

Software for Tier-2

- No major change from the previous system
- DPM as the disk storage system
  - 1 head node (SLC5)
  - 15 file servers (SLC5, XFS)
    - 60 filesystems = 15 x 2 disk arrays x 2 filesystems/array
  - 1.2 Pbytes in one pool
    - 1.0 Pbytes pledge
    - All ATLAS space tokens
    - Can add more disk arrays if needed
      - E.g. From AOD+DESD to ESD model

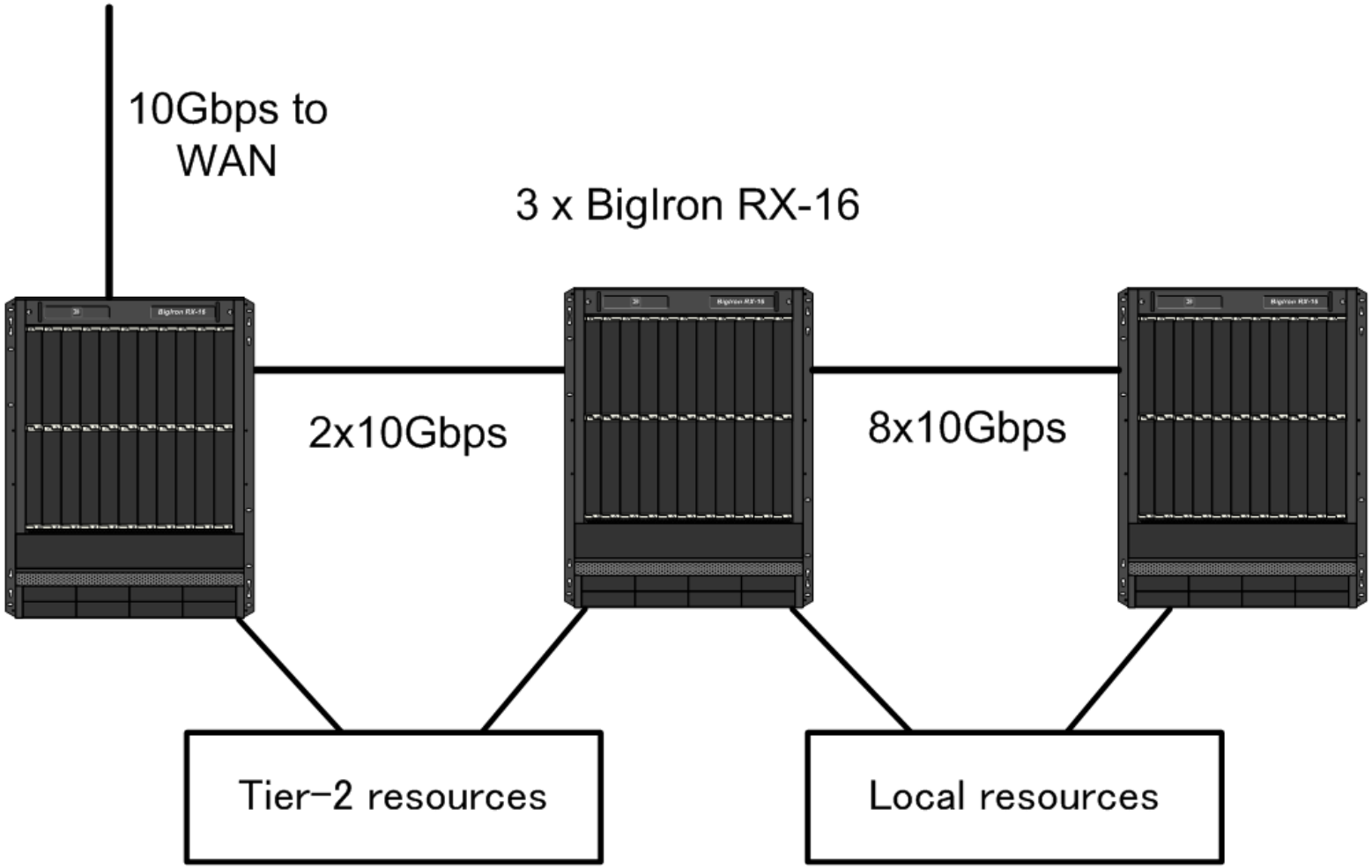
Software for Tier-2 (cont)

- LCG-CE
  - Torque/Maui
  - 896 cores (112 SLC5 nodes)
  - 12k HS06
- NFS server for ATLAS software area
  - Same fileserver and RAID as DPM ones
- Other services
  - Top BDII, WMS/LB, MyProxy
  - (ActiveMQ based) APEL
  - Squid proxy for ATLAS conditions database cache

Network connectivity

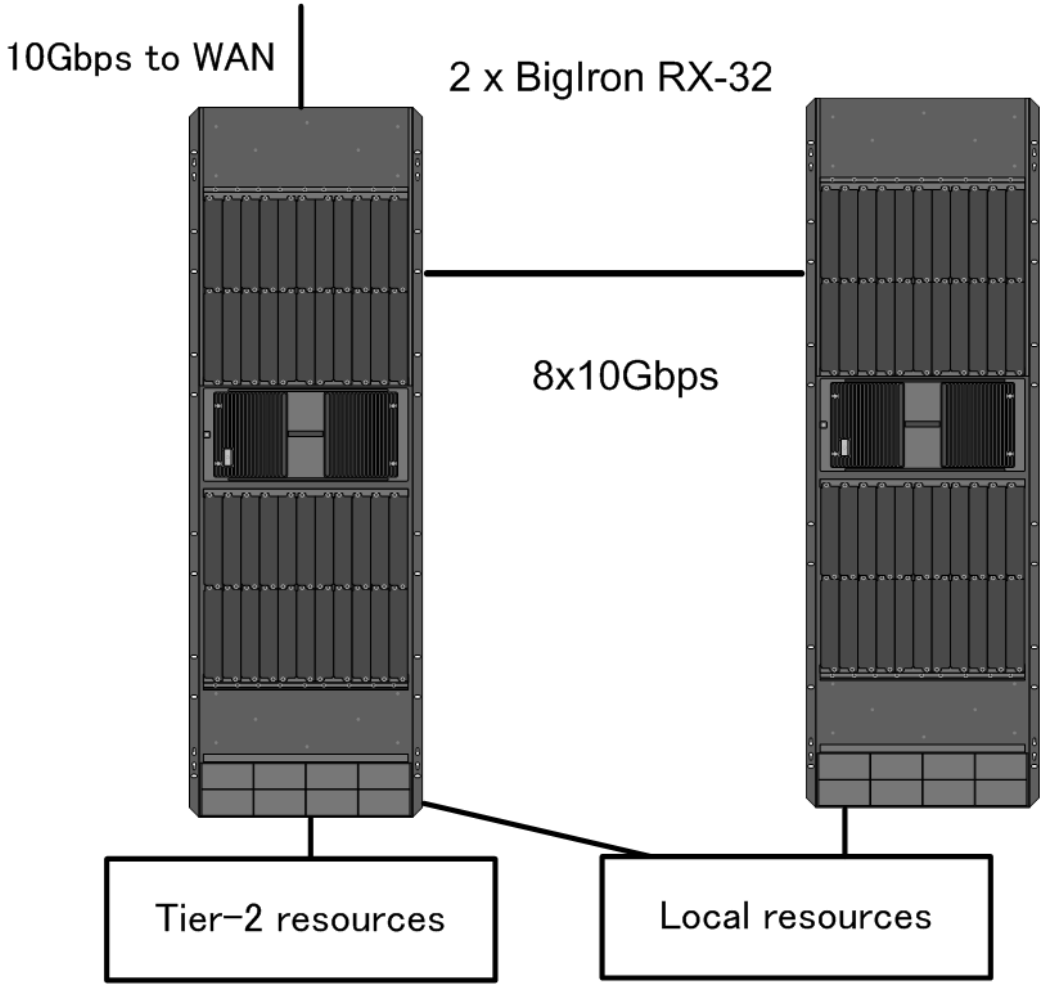
- 10 Gbps WAN (same as before)
- All Tier-2 resources connected to one Ethernet switch
  - No bottleneck due to inter-switch link
  - In the previous system, servers were connected to two switches, and the interlink was 20Gbps.
- Local resources connects to both two switches
  - Local CPUs can read data in Tier-2 disks

Old configuration



8 x 10Gbps is the limit of link aggregation for this switch

New configuration



Local resources

- Blade servers for batch and interactive nodes
  - Same architecture as those in Tier-2
  - 500+ servers with LSF batch system
- NFS (v3) servers
  - 10+ servers (~PB)
- Tape library
  - CASTOR for HSM
  - LTO-3 or 4 drives
- Oracle RAC
- AFS servers for the software repository

Data access and management

- DPM storage in Tier-2
  - Access from WNs within ATLAS (and WLCG) framework
    - (<2Gbps x) 112 WNs vs. (<10Gbps x) 15 fileservers (30 RAIDs)
  - Access also possible from local CPU servers
    - With X.509 credentials
    - Number of jobs should be adjusted if necessary (by user)

Data access and management (cont)

- Local NFS servers
  - Users' data are stored mainly in NFS servers
    - Users like plain filesystem
    - LOCALGROUPDISK is not very popular
  - Official data can also be written with dq2-get etc.
  - Scalability is a biggest issue
    - Each user should throttle jobs if they are I/O intensive
  - Even Tier-2 CPUs can read data
    - But user mapping is different between Tier-2 and local resources

Data access and management (cont)

- CASTOR
  - Cache disk + Tapes (3PB)
  - Still under construction with new hardware
    - New software version: 2.1.9
  - Data archive in the future
    - For both user data and official data
    - Easier for importing from NFS servers
    - Need additional staging disks for importing from DPM?
  - Can be used for user analysis in the long term

Summary

- The whole computer system has been replaced at the end of last year
  - System design is nearly the same as the previous one
- Most of the components are already working well
  - Operations are stable
- CASTOR deployment is a big challenge.
  - For large data archive