Grid Operation at Tokyo Tier-2 Center for Atlas

ICEPP, University of Tokyo
Hiroyuki Matsunaga

ISGC2008, Taipei
April, 2008

TOKYO-LCG2

- ICEPP, University of Tokyo
  - Involved in international collaboration since 1974
- Unique ATLAS Tier-2 site in Japan
  - Main analysis platform for Japanese ATLAS collaborators (including CERN-based persons)
    - No Tier-3 site anticipated in Japan for now
- Site operation supported by Asia-Pacific ROC
- Working in association with French Tier-1 (CC-IN2P3 at Lyon) for ATLAS
  - Data transfer and MC production are performed within the French “cloud”

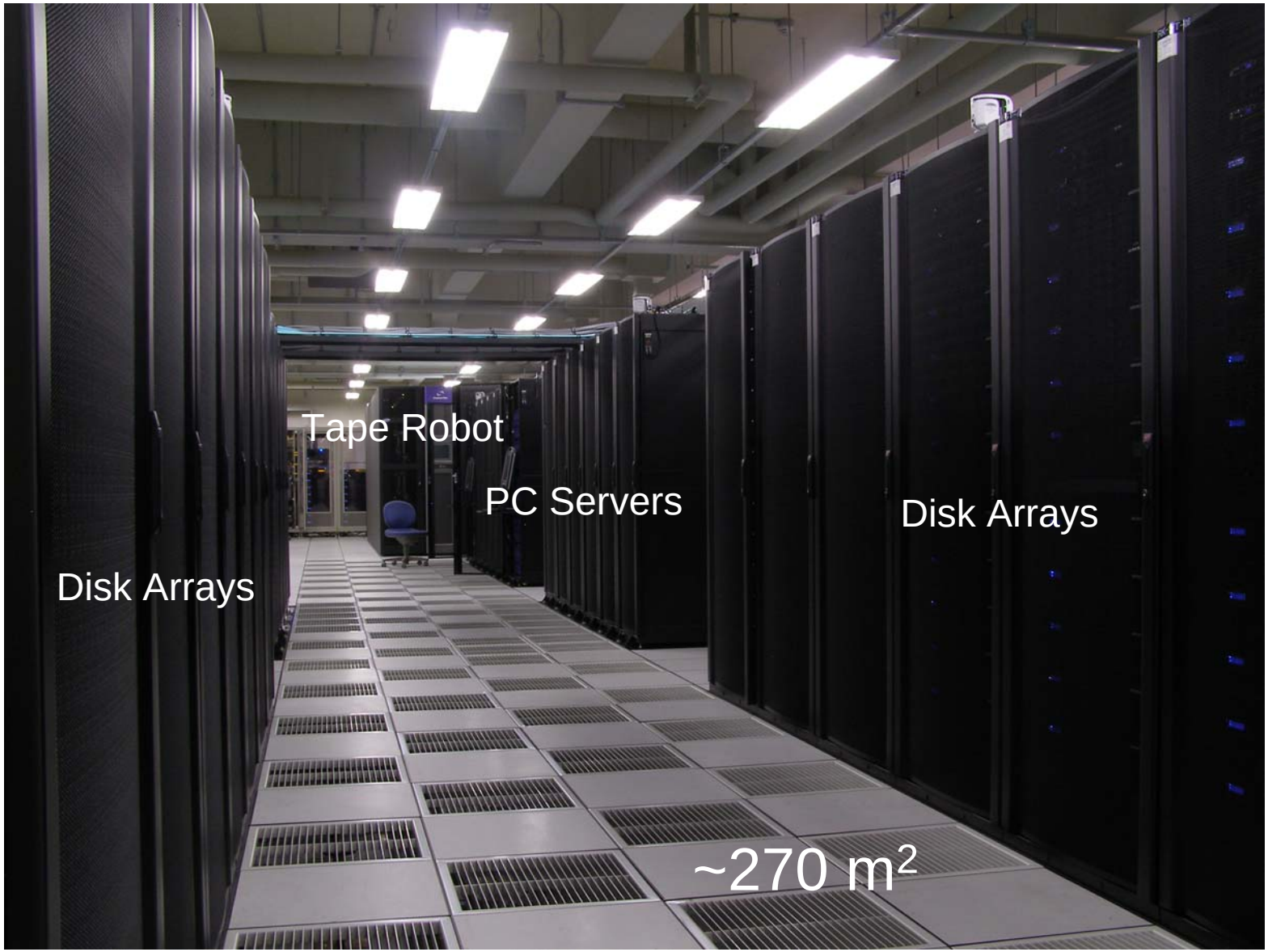
ATLAS-Japan

- 16 institutes: ~70 researchers, ~40 students
  - Endcap Muon Trigger system (TGC)
  - Muon TDC electronics
  - Barrel SCT
  - Online system
  - Central solenoid
  - Trigger simulation
  - Physics analysis
- Expect 30~50 users at Tokyo Regional Center (Grid site + local resource)
  - Few heavy users already
  - Give tutorial on Grid and ATLAS software every year

Computer System

- Current computer system started working early last year
  - Lease for three years (2007 – 2009)
  - Ramped up in production slowly, migrating from the old system with overlapping operations (almost finished)
- Electric power is supplied through UPS to all equipments
  - Though power supply in Tokyo is very stable
- A part of resources is used for only Japanese collaborators (without requiring Grid access)

Computer room



Hardware Components

- Blade servers
  - Dual 3.0GHz Woodcrest (Xeon 5160) CPUs (4cores/node)
  - 8Gbytes RAM
  - 73GB SAS HDDs (mirrored)
    - ~10 Gbytes/core for working space
- Disk arrays
  - ~ 6 Tbytes/box
    - 16 x 500Gbytes SATA HDDs
    - RAID-6 (hardware RAID)
    - 4Gb Fibre-Channel
  - Access through file servers with 10Gb NIC
    - 5 disk arrays/server

Resources for WLCG

From WLCG MoU

	Pledged	Planned to be pledged		
	2007	2008	2009	2010
CPU (kSI2k)	1000	1000	1000	3000
Disk (Tbyes)	200	400	400	600
Nominal WAN (Mbits/sec)	2000	2000	2000	2000

- As of March 2008: 1080 kSI2k, 190 Tbytes, 10000 M (10 G) bits/sec
  - Additional ~200 Tbytes will be deployed soon
- Major upgrade in 2010

Grid services

- gLite 3.0 (SLC3) in production
  - gLite 3.1 (SLC4, 32bit) WNs since last year
  - CE: Torque+Maui
    - LSF is used for non-Grid CPU servers
  - SE/SRM: DPM with MySQL
    - Disk servers: SLC4 x86_64, XFS (6Tbytes/filesystem)
    - Only one pool for ~190Tbytes
    - SRMv2.2: Space token deployed since February
      - For ATLAS FDR-1 and CCRC-1
  - BDII, MON, UI, RB, LFC, MyProxy

Other softwares

- Quattor
  - OS installation (+ update except non-Grid nodes)
- Lemon
  - Fabric monitoring
- MRTG, SmokePing
  - Network monitoring
- OpenManage (Dell), RAIDWatch (Infortrend)
  - Hardware administration
- Home-made scripts
  - Disk usage and file transfer statistics on SE
  - Accounting on CE

Networking

- 10Gbps to the University and the Japanese NREN (SINET) since July 2007
  - Bypassing a firewall in the University
- 10Gbps through CC-IN2P3 (Lyon)
  - SINET + GEANT2 + RENATOR
    - Bottleneck (2.4 Gbps) between SINET and GEANT2 routers at New York was removed in February 2008
- 1Gbps to ASGC
  - Will be upgraded to 2.4Gbps soon
- Started discussion with SINET people for better network performance

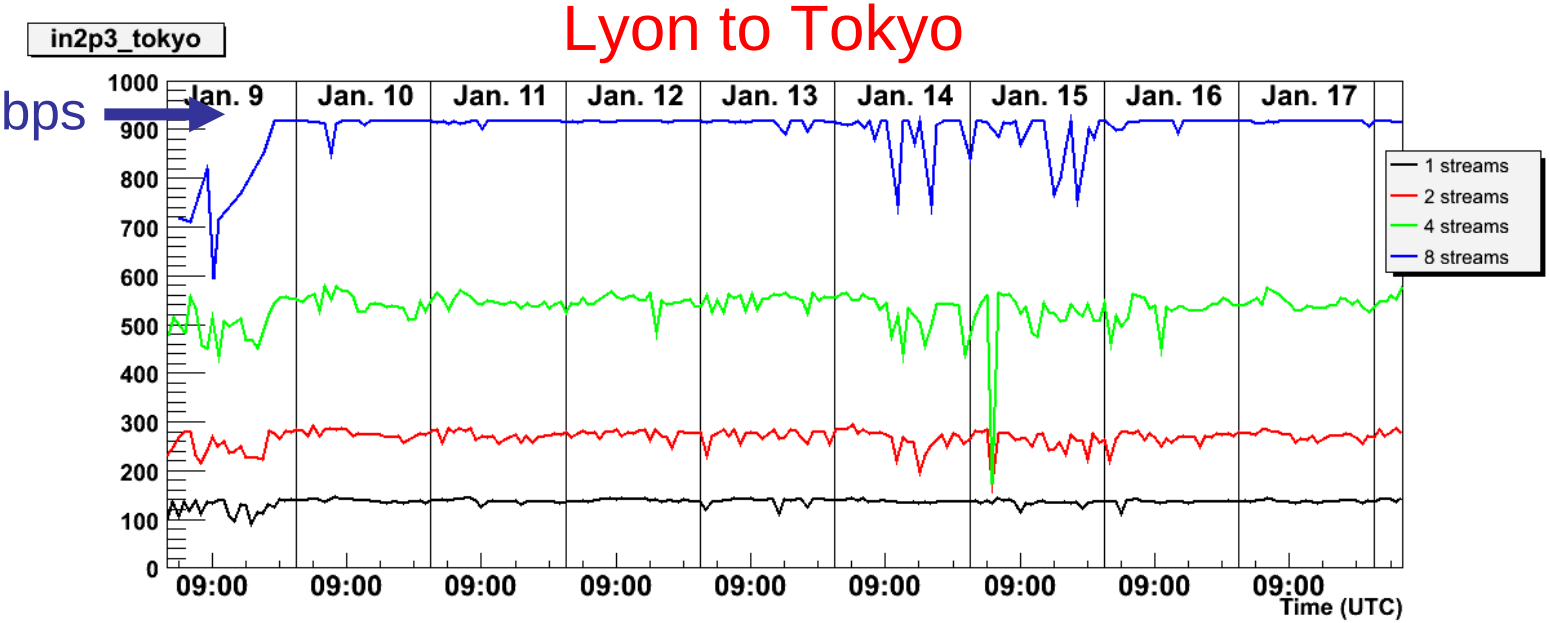
File transfers

- One of main concerns
  - Long path between Lyon and Tokyo (“Long Fat Network”)
    - RTT ~280 msec, ~10 hops
    - most of the path is shared with other traffics
      - Need better TCP implementation for congestion control
      - Should optimize many parameters
        - » Window sizes, number of streams
- Performed many (disk-to-disk) tests with SEs at both ends, in addition to the iperf (memory-to-memory) tests
  - Not try to use extra software or patch if possible

Network tests with iperf

- Use Linux boxes dedicated for iperf test, each with 1Gbps NIC
- Since late last year, throughput has been improved, in particular from Lyon to Tokyo
  - Lyon-to-Tokyo direction is more important for real data transfers
- Saturates 1Gbps NIC with 8 streams, using SL4, for long period
  - ~500 Mbps in the past
  - Worse with SL3
    - ~300 Mbps (due to different TCP implementation)

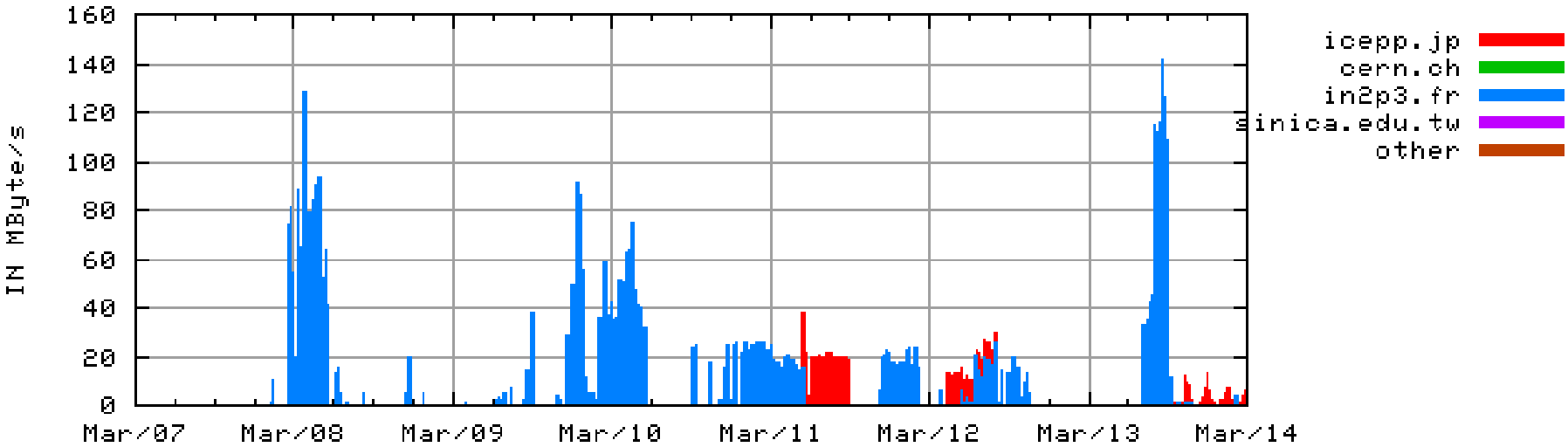
~930Mbps



FTS

- Main tool for bulk data transfers
  - Mostly used in DQ2 (Distributed Data Management system) in ATLAS
- Important to optimize “# of concurrent files” and “# of streams” for better throughput
  - 20 files and 10 streams seem good for now
    - Many streams lead to slow read due to file fragmentation (even with XFS)

Data transfers at M6 run



(Bin width =10min)

- M6 cosmic data transferred via the IN2P3-TOKYO FTS channel
  - dCache (Solaris , ZFS) at Lyon
  - DPM (Linux, XFS) at Tokyo
- Sustained >100 MB/s in case both DQ2 and FTS are healthy
  - DQ2 or FTS sometimes in trouble
  - Non-negligible overhead time with FTS
  - Achieved ~140 MB/s at the peak rate

Site operation

- Resources are (logically) divided into two parts (Grid and non-Grid)
- Administrated by ~5 FTEs
  - Including infrastructure
  - 2 system engineers from company
  - 1~2 FTEs for the Grid operations
- Working on more software deployment
  - LSF, AFS, Oracle, Castor

Site Operation (cont)

- Priority is given to high availability and reliability (for physics outputs)
  - Uses of UPS, RAID, DNS round-robin (still only for UI nodes), Oracle RAC (to be deployed)
  - Extensive burn-in and torture tests before deployment
  - Careful middleware upgrades
    - Choose well-established (and supported) software
    - Validation on testbed beforehand
- Availability of 0.98 and Reliability of 1.00 in January 2008

Plans

- Upgrade gLite middleware to 3.1
- Add 200Tbytes of disk space
- Deployment Oracle RAC
  - ATLAS conditions DB?
  - gLite backend ?
- Set up Castor (tape + disk) for local users
- AFS for home and common software areas
- Try to deploy LSF instead of Maui/Torque (?)

Summary

- Tokyo Tier-2 center is in good shape
  - Resources growing according to the plan
- Gain operational experience
  - Within ATLAS distributed computing model
  - In collaboration with French Tier-1
- International network is improving
  - Data transfer performance (from France to Japan) is getting better

Backup

Traceroute from Tokyo to Lyon

- 1 157.82.112.1 (157.82.112.1) 0.306 ms 0.266 ms 1.032 ms
- 2 tokyo1-dc-RM-GE-1-0-0-119.sinet.ad.jp (150.99.190.101) 1.758 ms 27.654 ms 0.846 ms
- 3 150.99.203.58 (150.99.203.58) 175.752 ms 175.840 ms 175.785 ms
- 4 150.99.188.202 (150.99.188.202) 175.974 ms 175.999 ms 202.050 ms
- 5 so-7-0-0.rt1.ams.nl.geant2.net (62.40.112.133) 259.296 ms 259.250 ms 259.374 ms
- 6 so-4-0-0.rt1.lon.uk.geant2.net (62.40.112.138) 267.461 ms 267.351 ms 267.522 ms
- 7 so-4-0-0.rt1.par.fr.geant2.net (62.40.112.105) 274.785 ms 274.803 ms 274.782 ms
- 8 renater-gw.rt1.par.fr.geant2.net (62.40.124.70) 275.223 ms 274.683 ms 288.517 ms
- 9 in2p3-nri-b.cssi.renater.fr (193.51.186.177) 280.296 ms 280.378 ms 280.331 ms
- 10 Lyon-CORE.in2p3.fr (134.158.224.3) 280.312 ms 280.457 ms 280.351 ms
- 11 cclcgbdiili01.in2p3.fr (134.158.105.155) 280.258 ms 280.272 ms 280.164 ms

New York