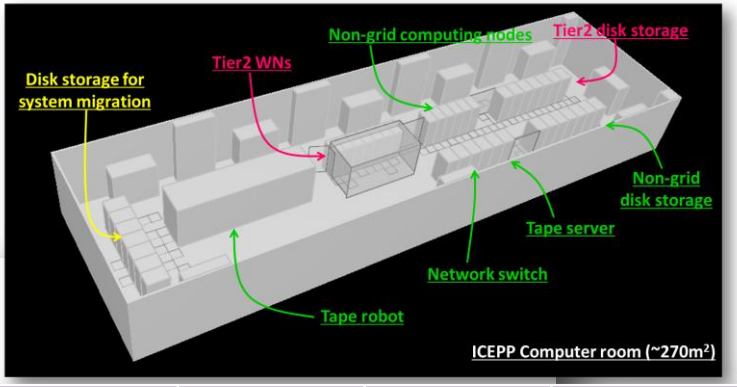
4th system upgrade of Tokyo Tier2 center

Tomoaki Nakamura
KEK-CRC / ICEPP UTokyo

ICEPP regional analysis center



Resource overview

Support only ATLAS VO in WLCG as Tier2.
Provide ATLAS-Japan dedicated resource for analysis.
The first production system for WLCG was deployed in 2007.
Almost of hardware are prepared by three years rental.
System have been upgraded in every three years.
~10,000 CPU cores and 6.7PB disk storage (T2 + local use).

	2013	2014	2015
CPU pledge	16000 [HS06]	20000 [HS06]	24000 [HS06]
CPU deployed	43673.6 [HS06-SL5] (2560core)	46156.8 [HS06-SL6] (2560core)	46156.8 [HS06-SL6] (2560core)
Disk pledge	1600 [TB]	2000 [TB]	2400 [TB]
Disk deployed	2000 [TB]	2000 [TB]	2400[TB]

18.03HS06/core

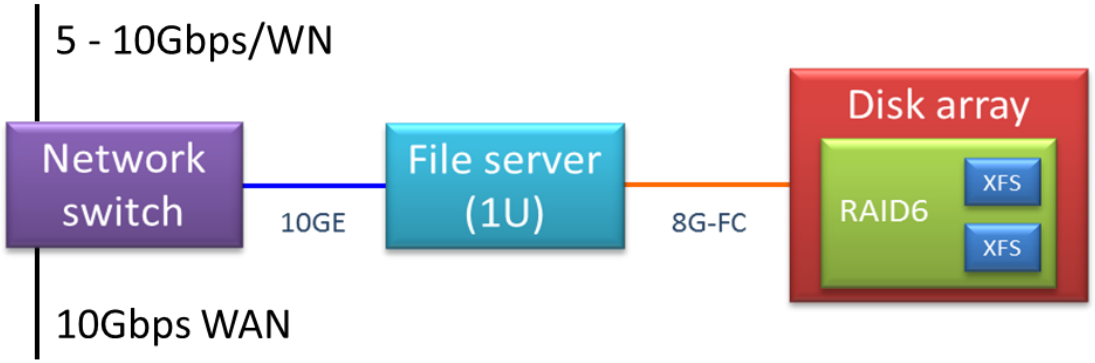
Single VO and Simple and Uniform architecture

Dedicated staff

Tetsuro Mashimo:	fabric operation, procurement
Nagataka Matsui:	fabric operation
Tomoaki Nakamura (KEK-CRC):	Tier2 operation and setup, analysis environment
Hiroshi Sakamoto:	site representative, coordination, ADCoS
System engineer from company (2FTE):	fabric maintenance, system setup

Configuration of the 3rd system

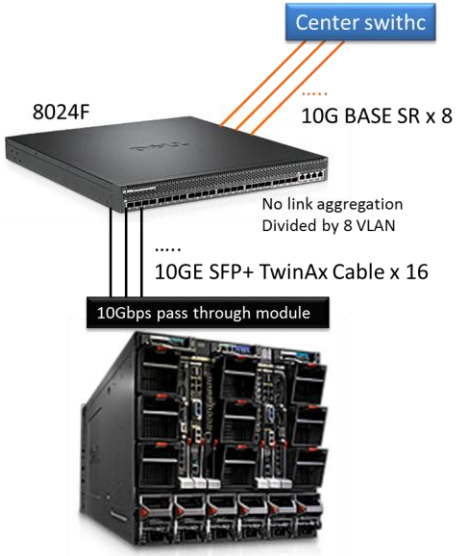
Disk server x48



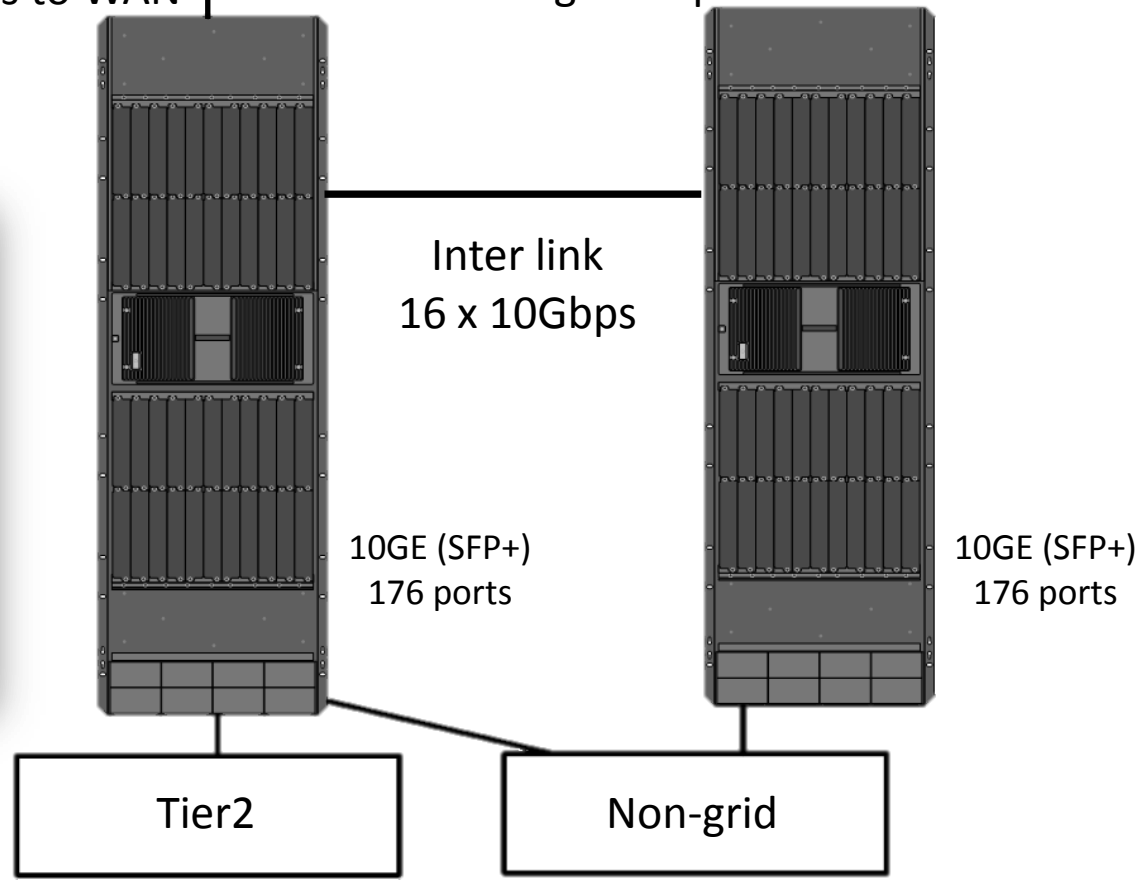
- 66TB x 48 servers
- Total capacity 3.168PB (DPM)
- 10Gbps NIC (for LAN)
- 8G-FC (for disk array)
  500~700MB/sec
  (sequential I/O)

Worker node x160

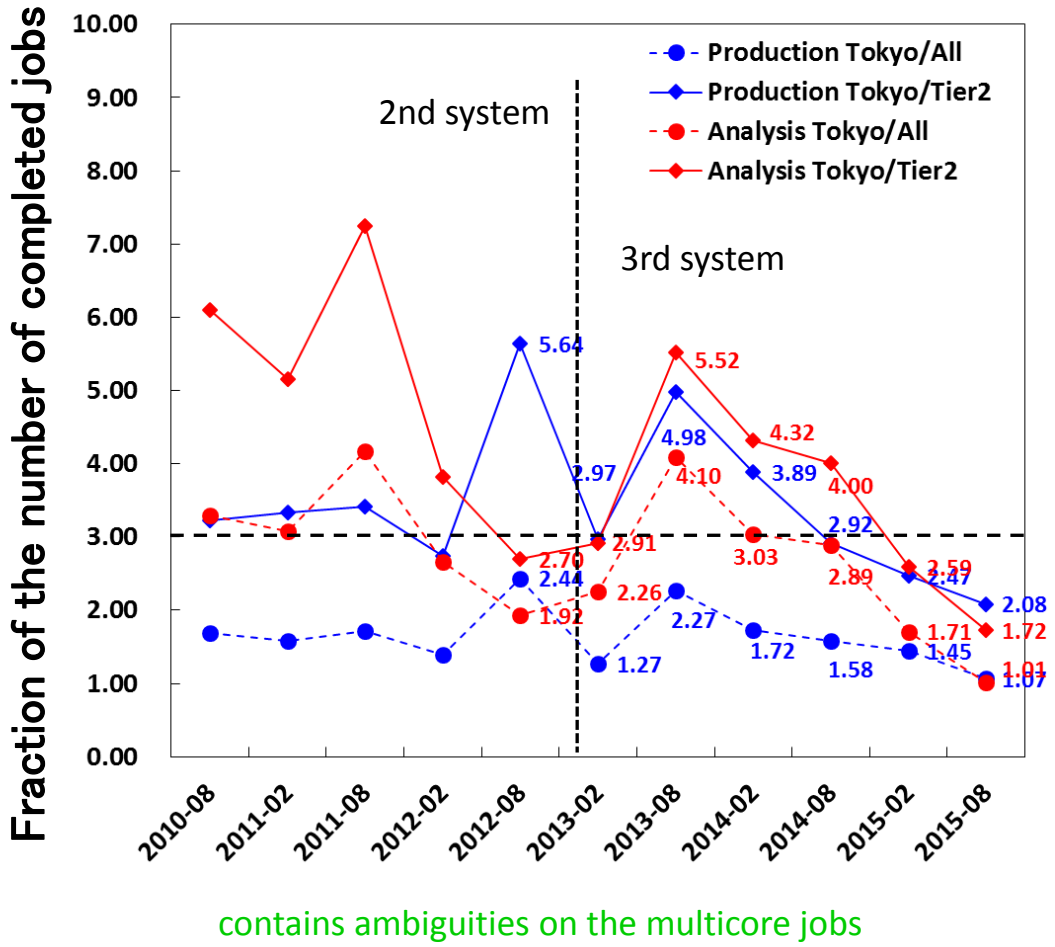
- CPU: 16CPU/node (18.03/core)
- Memory: 2GB/core (80nodes) + 4GB/core (80nodes)
- 10Gbps pass through module (SFP+ TwinAx cable)
- Rack mount type 10GE switch (10G BASE SR SFP+)
- Bandwidth
  80Gbps/16nodes
  minimum 5Gbps
  maximum 10Gbps



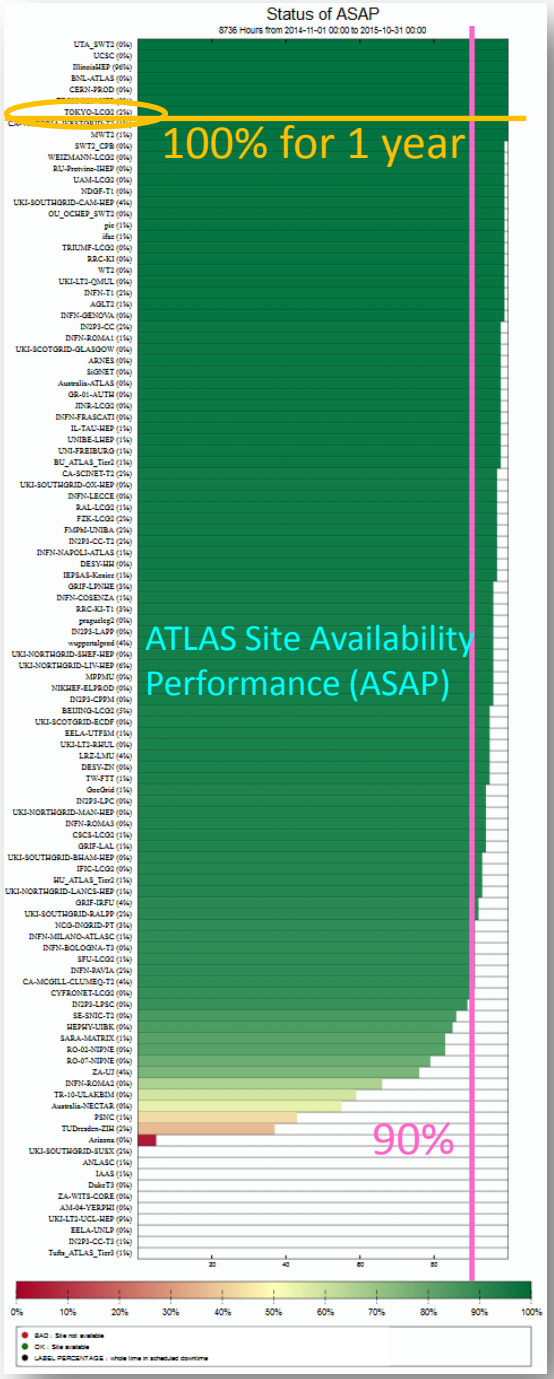
Network configuration



Status in ATLAS



contains ambiguities on the multicore jobs



Multicore queue (8 cores/job)

CE configuration

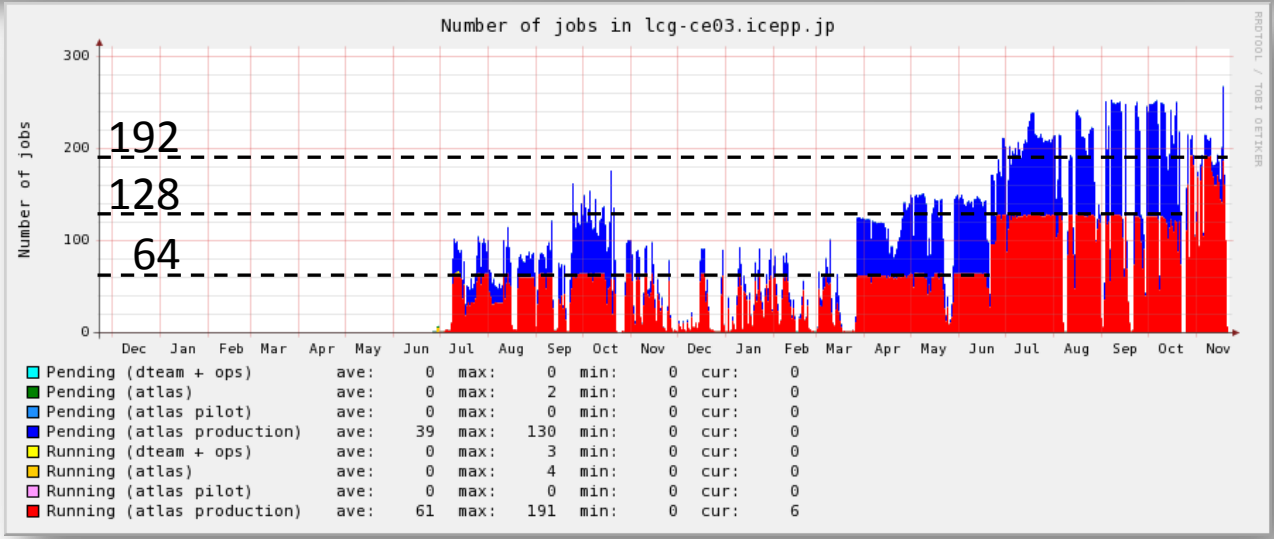
- lcg-ce01.icepp.jp: Dedicated to single core jobs (Analysis and Production jobs)
- lcg-ce02.icepp.jp: Dedicated to single core jobs (Analysis and Production jobs)
- lcg-ce03.icepp.jp: Dedicated to multi core jobs (Production jobs by static allocation)

Squids

- 2 squids for CVMFS (dynamic load balancing and fail-over, active-active)
- 2 squids for Conditional DB (static load balancing and fail-over, active-active)

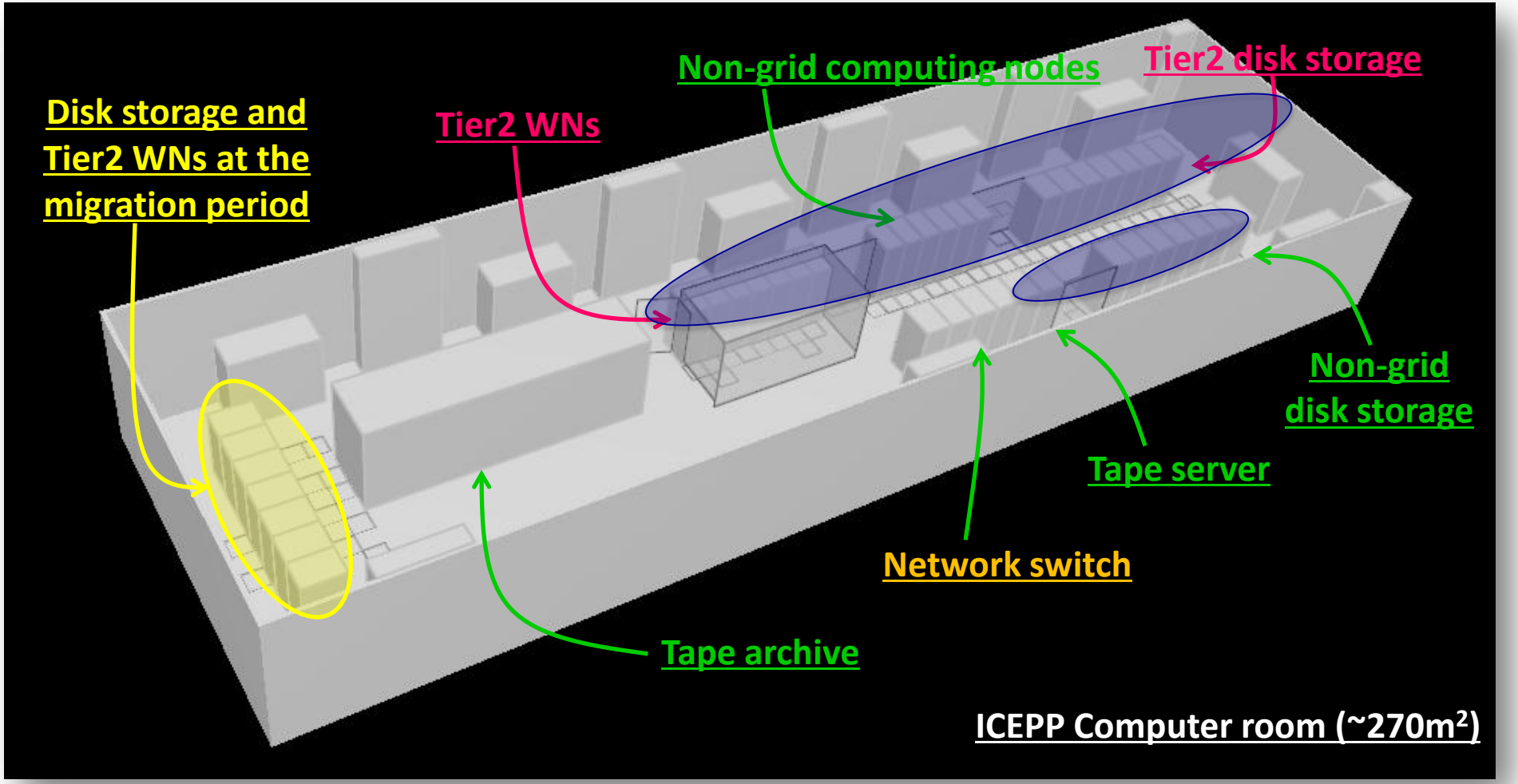
WN allocation for multicore queue

• Jul. 2014: first deployment	(512 cores, 64 job slots, 20%)	Analysis 50%
• Jul. 2015: re-allocation	(1024 cores, 128 job slots 40%)	Analysis 50%
• Oct. 2015: re-allocation	(1536 cores, 192 job slots, 60%)	Analysis 25%



2014 - 2015

System upgrade (Dec. 2015)



System migration (Dec. 2015 - Jan. 2016)

3rd system

Clearance: 2 days

4th system

Construction: 1 week

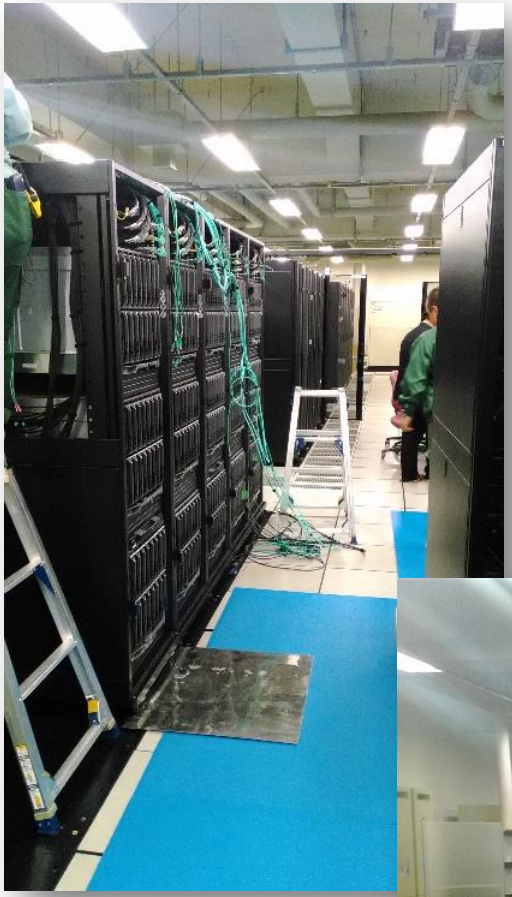
Copy back
several week

Migration period

Running with reduced
number of WNs

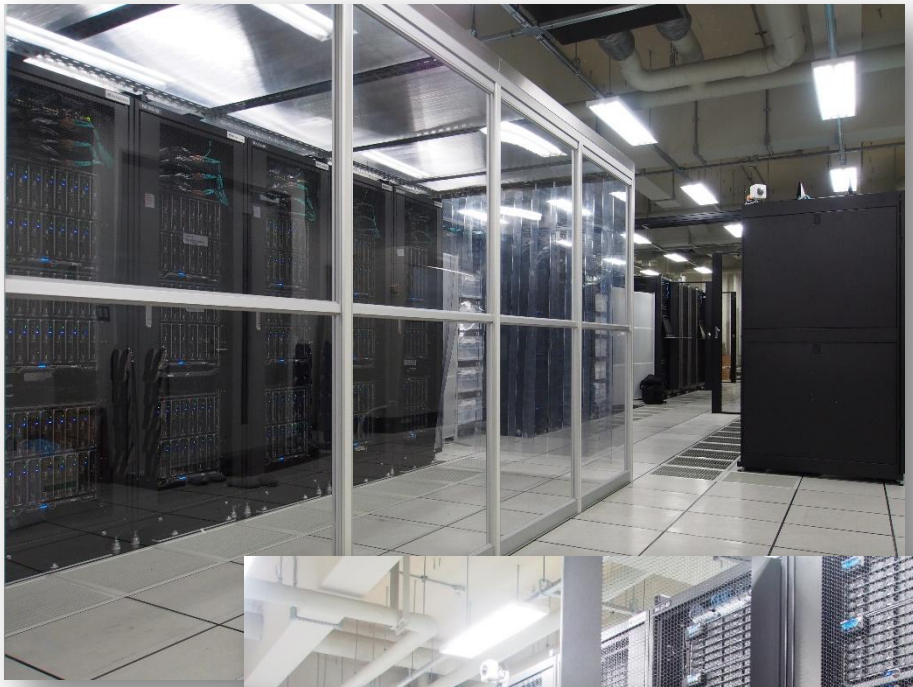
Data copy:
several week

HW clearance (2days in Dec. 2015)

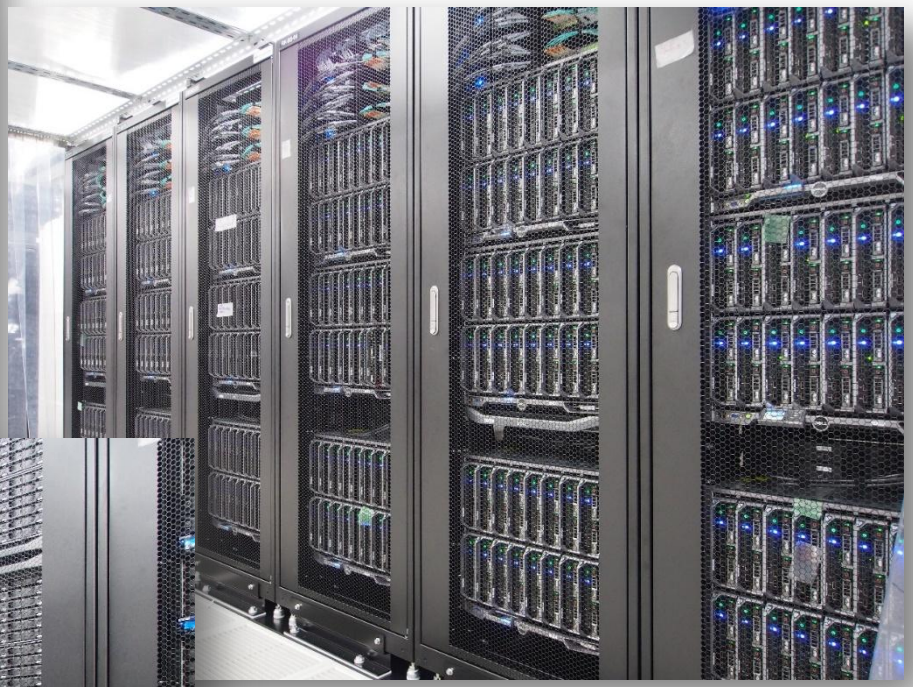


Constructing new HWs (~5 days)

4th system



worker nodes



Disk arrays

4th system

		3rd system (2013-2015)	4th system (2016-2018)
Computing node	Total	Node: **624 nodes, 9984 cores** (including service nodes) CPU: Intel Xeon E5-2680 (Sandy Bridge 2.7GHz, 8cores/CPU)	Node: **416 nodes, 9984 cores** (including service nodes) CPU: Intel Xeon E5-2680 v3 (Haswell 2.5GHz, 12cores/CPU)
	Tier2 pledge 2016 28 kHS06 pledge 2017 32 kHS06	Node: **160 nodes, 2560 cores** Memory: 32GB/node, 64GB/node NIC: 10Gbps/node Network BW: 80Gbps/16 nodes Disk: 600GB SAS x 2	Node: **160 nodes, 3840 cores** Memory: 64GB/node (2.66GB/job slots) NIC: 10Gbps/node Network BW: 80Gbps/16 nodes Disk: 1.2TB SAS x 2
Disk storage	Total	Capacity: **6732TB** (RAID6) Disk Array: 102 (3TB x 24) File Server: 102 nodes (1U) FC: 8Gbps/Disk, 8Gbps/FS	Capacity: **10560TB** (RAID6) + α Disk Array: 80 (6TB x 24) File Server: 80 nodes (1U) FC: 8Gbps/Disk, 8Gbps/FS
	Tier2	DPM: **3.168PB**	DPM: **6.336PB (+1.056PB)**
Network bandwidth	LAN	10GE ports in switch: 352 Switch inter link : 160Gbps	10GE ports in switch: 352 Switch inter link : 160Gbps
	WAN	ICEPP-UTNET: **10Gbps** SINET-USA: 10Gbps x 3 ICEPP-EU: 10Gbps (+10Gbps)	ICEPP-UTNET: **20Gbps (+20Gbps)** SINET-USA: 100Gbps + 10Gbps ICEPP-EU: 20Gbps (+20Gbps)

Grid middle ware

- Simplify for the dedicated services of ATLAS
- CE (3), SE (SRM, WebDAV, Xrootd), Squid (4), APEL, BDII (top, site), Argus, exp-soft
  are migrated from EMI3 to UMD3/SL6
- 3 perfSONAR are kept by the same server
- WMS, LB, MyProxy will be decommissioned (currently running)

Scale-down system (Dec. 2015 to Jan. 2016)

Scale-down system

32 WNs (512 cores)
Full Grid service

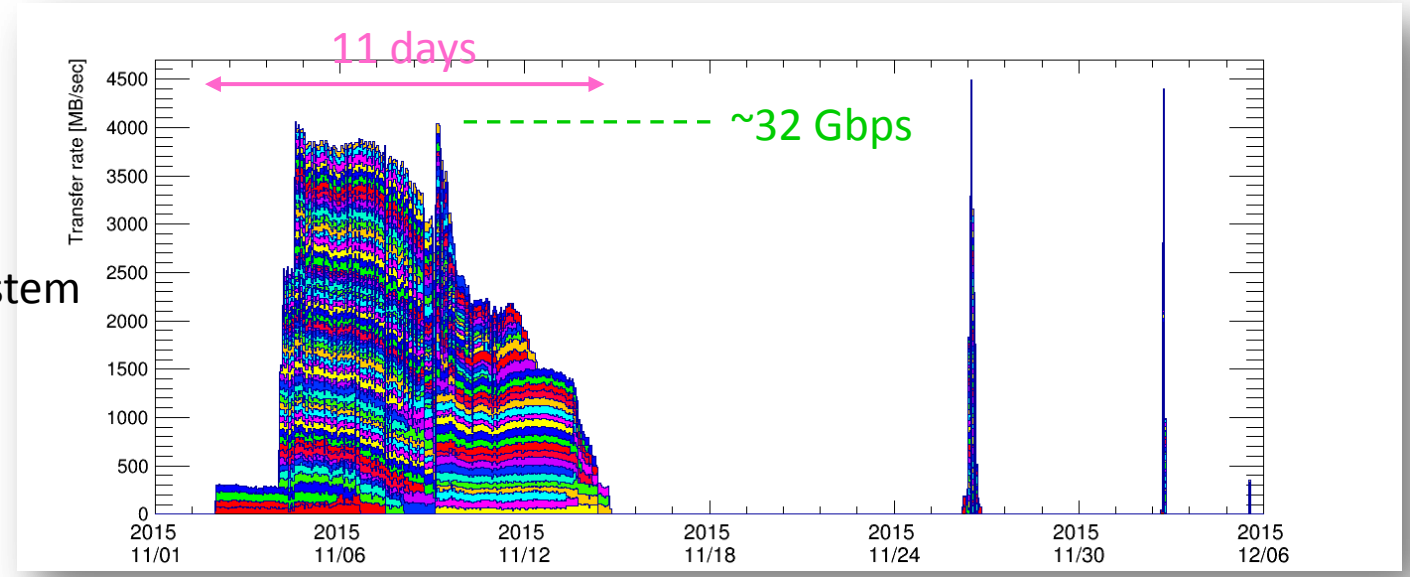
Temporal storage

All of data stored in Tokyo (3.2PB) was accessible from Grid during the migration period.

Data migration

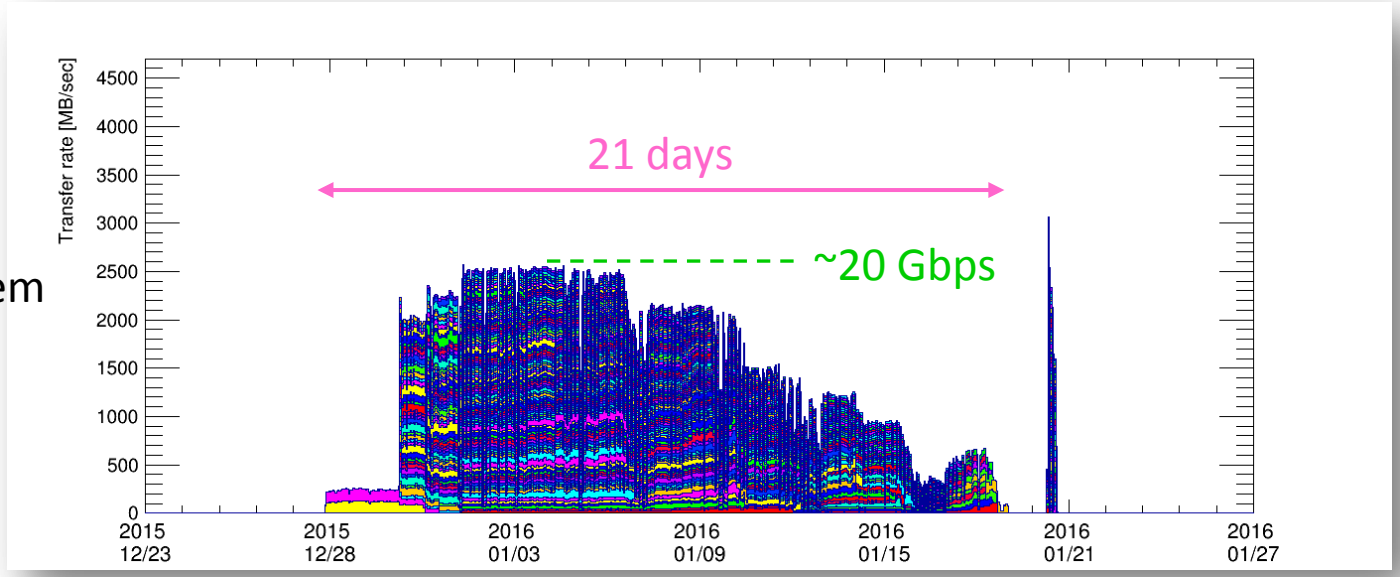
~2.4 PB, 1.5 M files

Copy to scale-down system

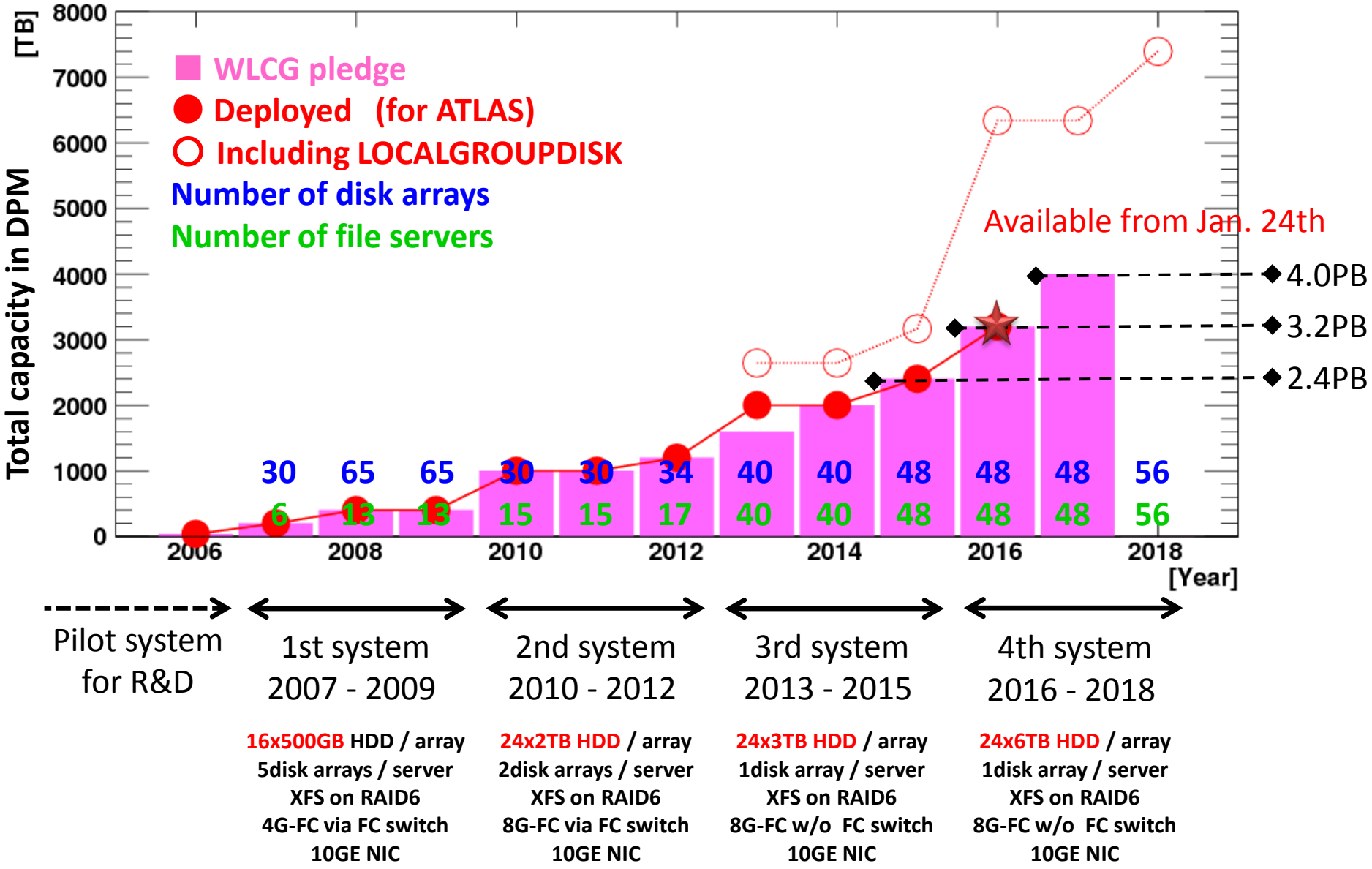


10G x 8 link aggregation (1h average)

Copy back to new system

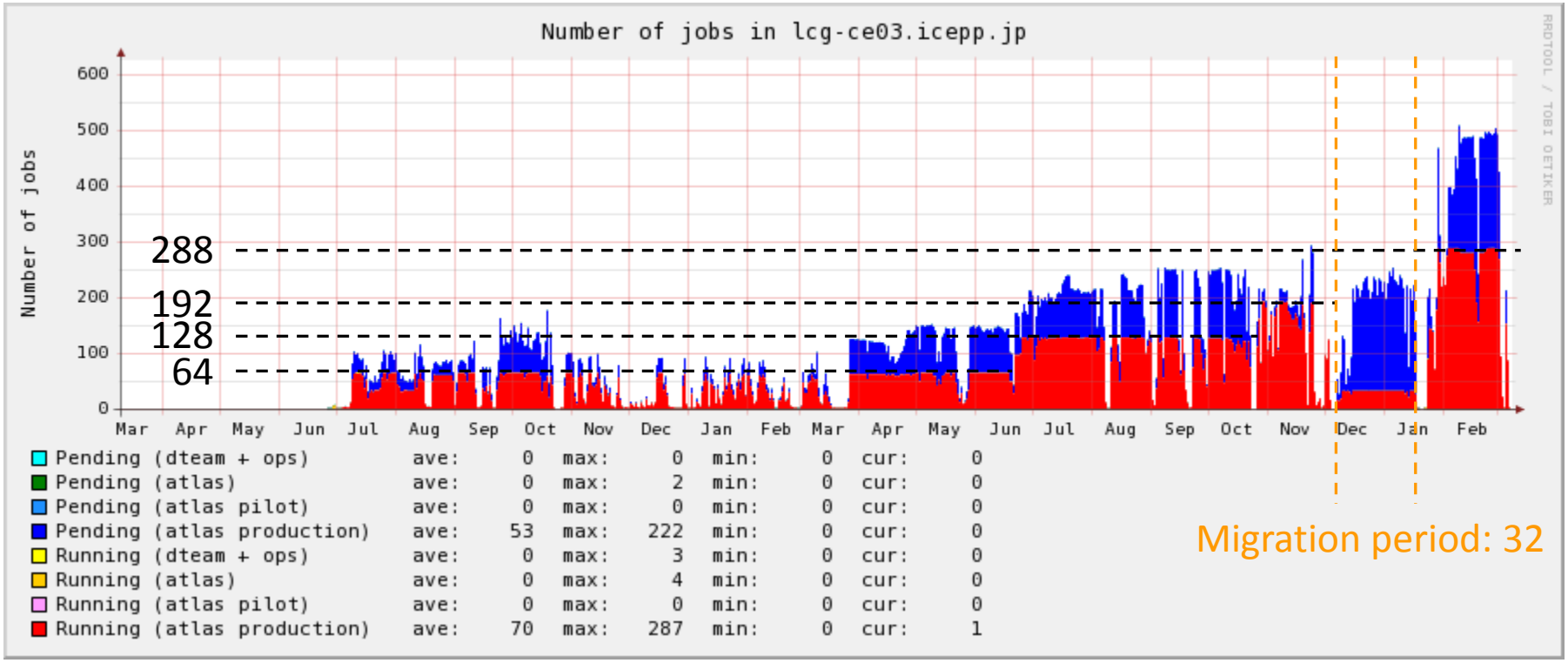


Disk storage for Tier2



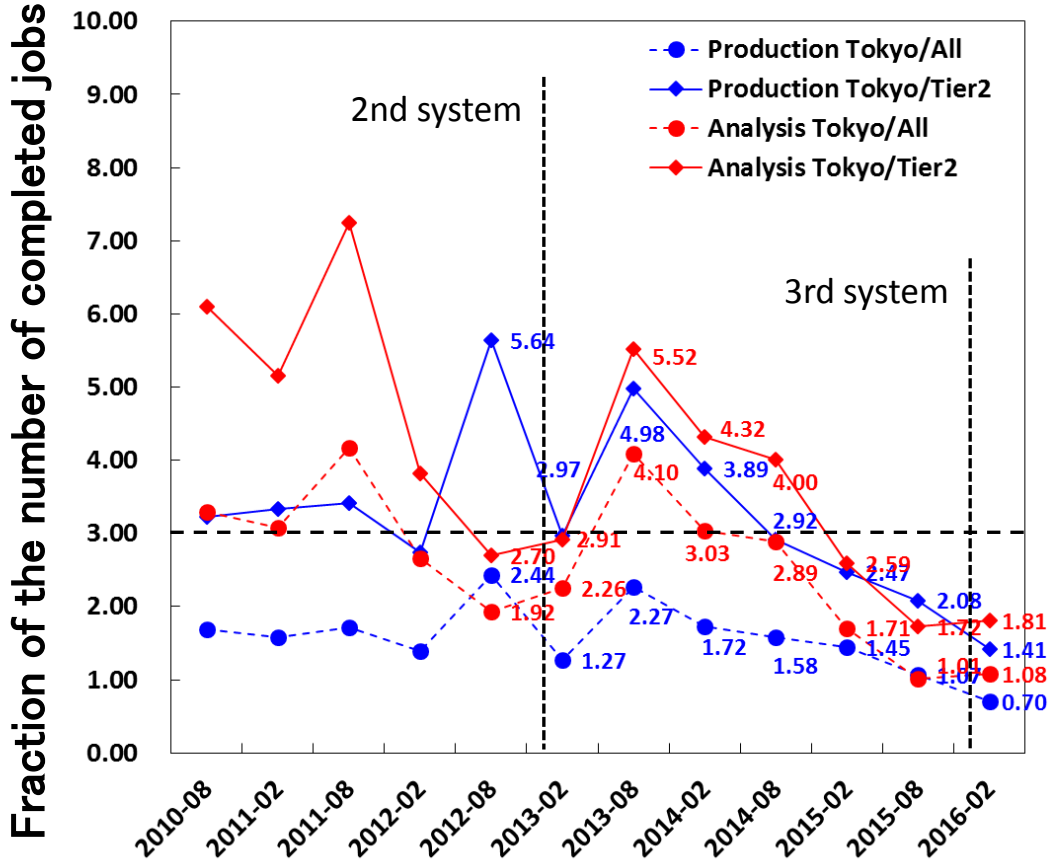
Running CPUs

Multi core jobs (8 cores/job)



288 (8 core job, 2304 cores) slots + 1536 (single job) slots
= 3840 CPU cores in total

Latest month (Feb. 2016)



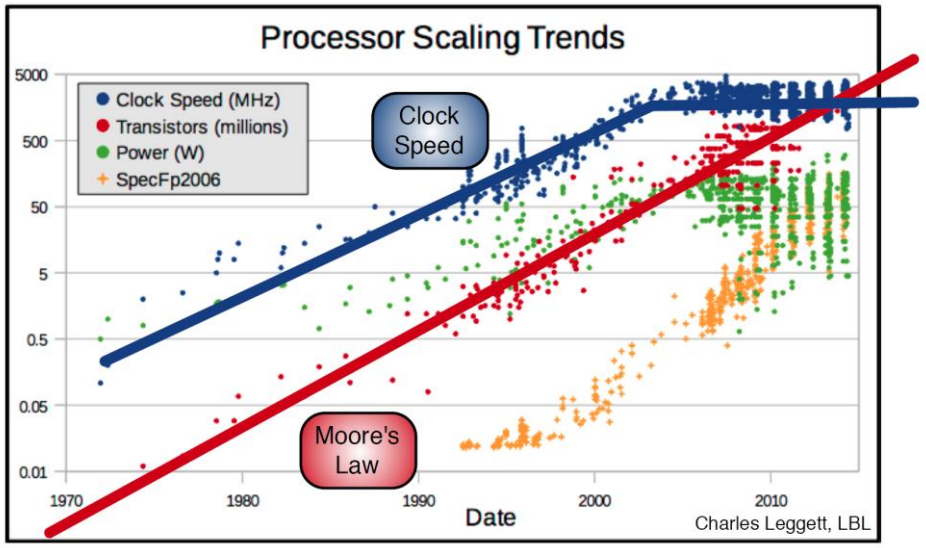
contains ambiguities on the multicore jobs

Latest one month (Feb. 2016)

Production Tokyo/All:	0.84
Production Tokyo/Tier2:	1.82
Production (8cores) Tokyo/All:	1.47
Production (8cores) Tokyo/Tier2:	2.76
Analysis Tokyo/All:	1.73
Analysis Tokyo/Tier2:	2.73

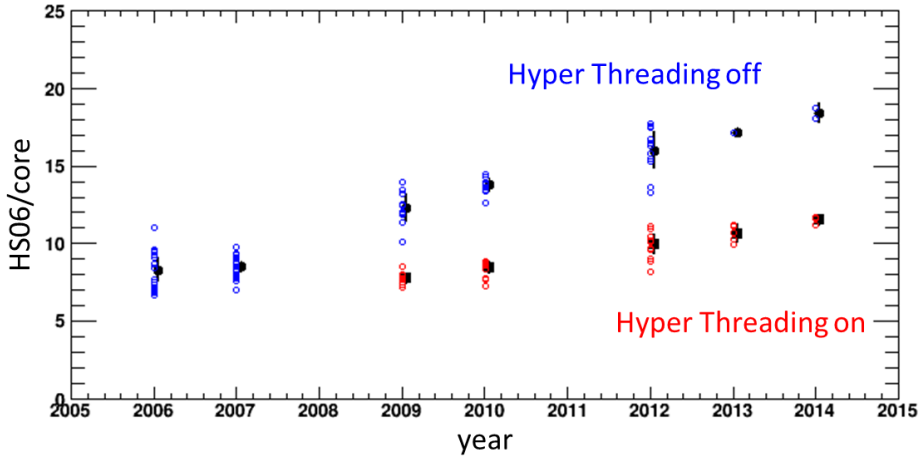
Planning to add 80 WNs to Tier2
(+1920 CPU cores, 5760 CPU cores in total)

CPU performance

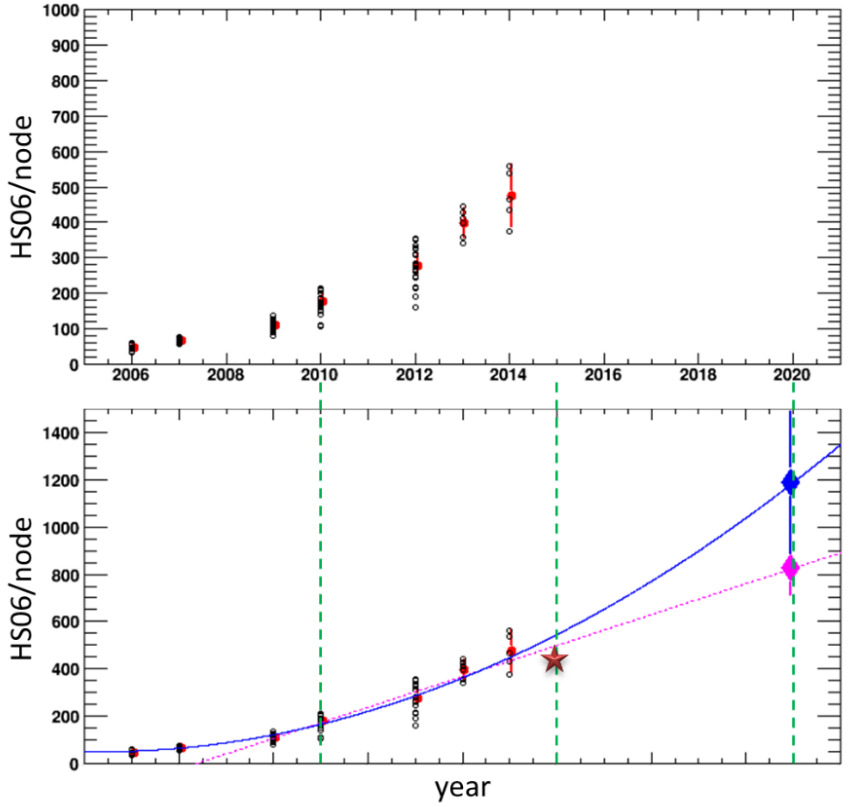


3rd system (8cores/CPU × 2)
E5-2680 (Sandy Bridge) 2.7GHz: 18.03

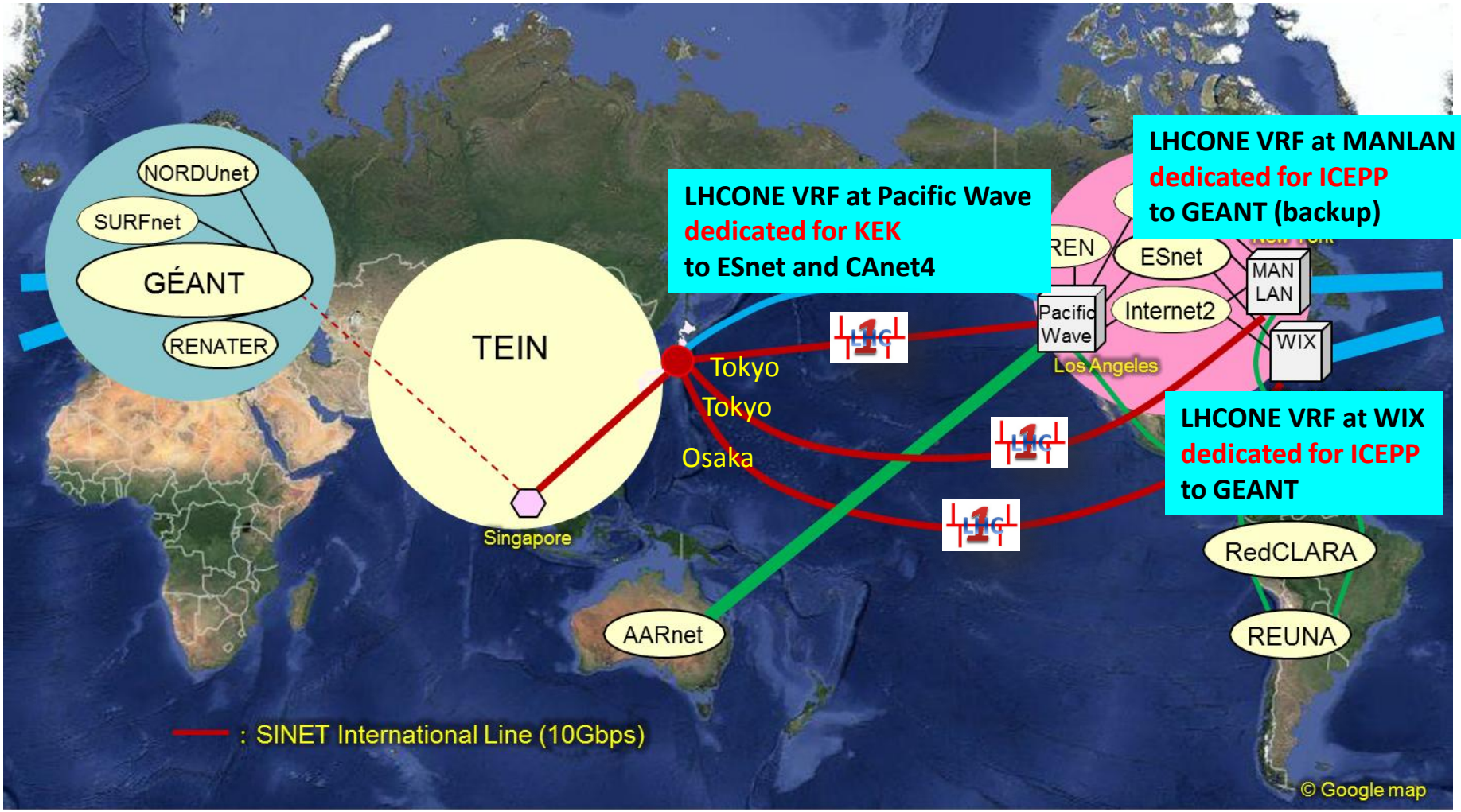
4th system (12cores/CPU × 2)
E5-2680 v3 (Haswell) 2.5GHz: 18.11



2% improve / year



Current LHCONE peering



Y. Kubota (NII)

Data transfer with the other sites

Sustained transfer rate

Incoming data: ~100MB/sec in one day average

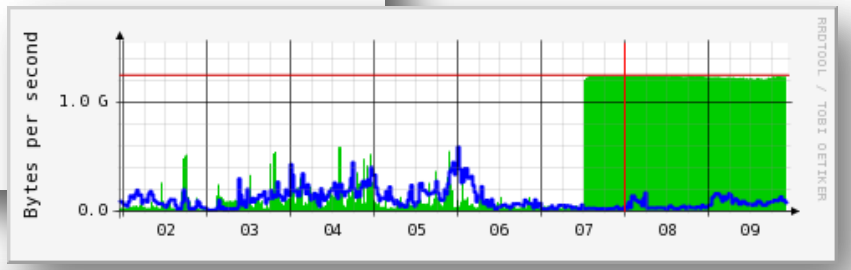
Outgoing data: ~50MB/sec in one day average

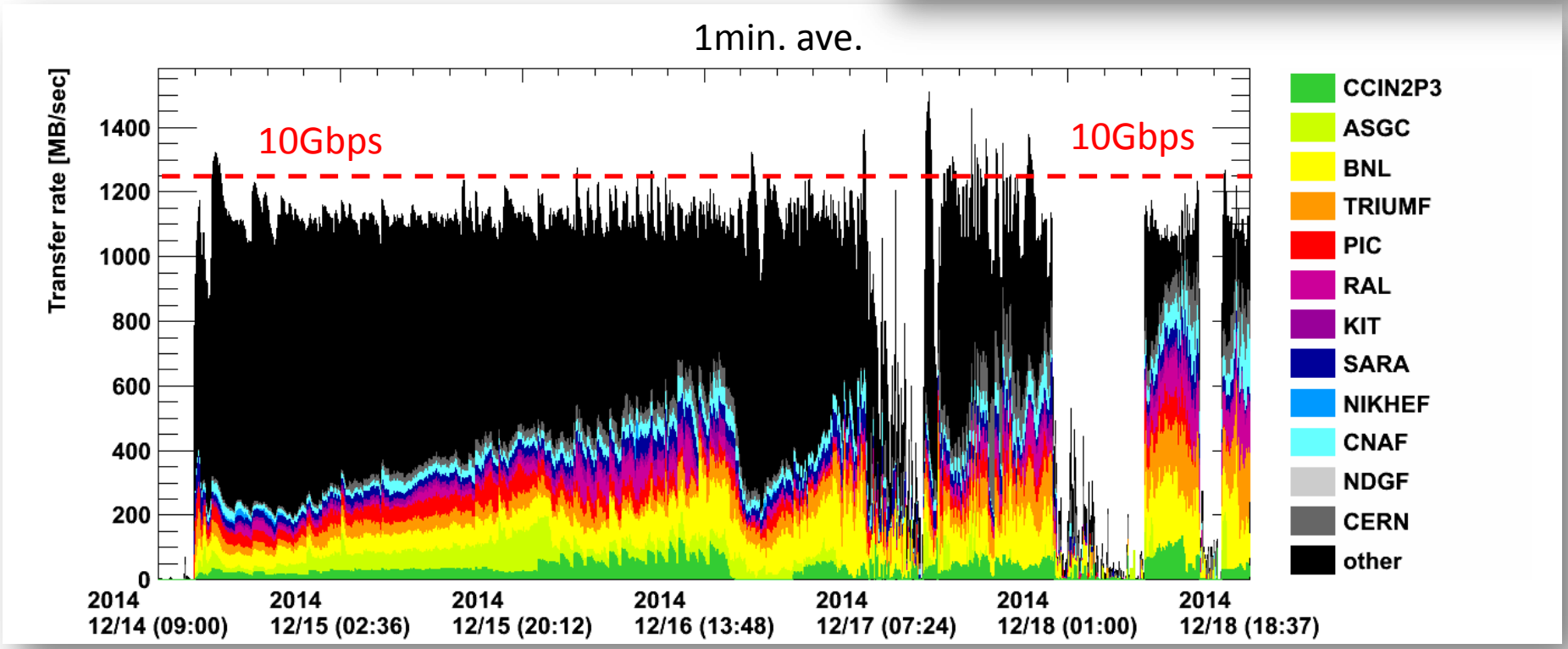
300~400TB of data in Tokyo storage is replaced within one month!

Peak transfer rate

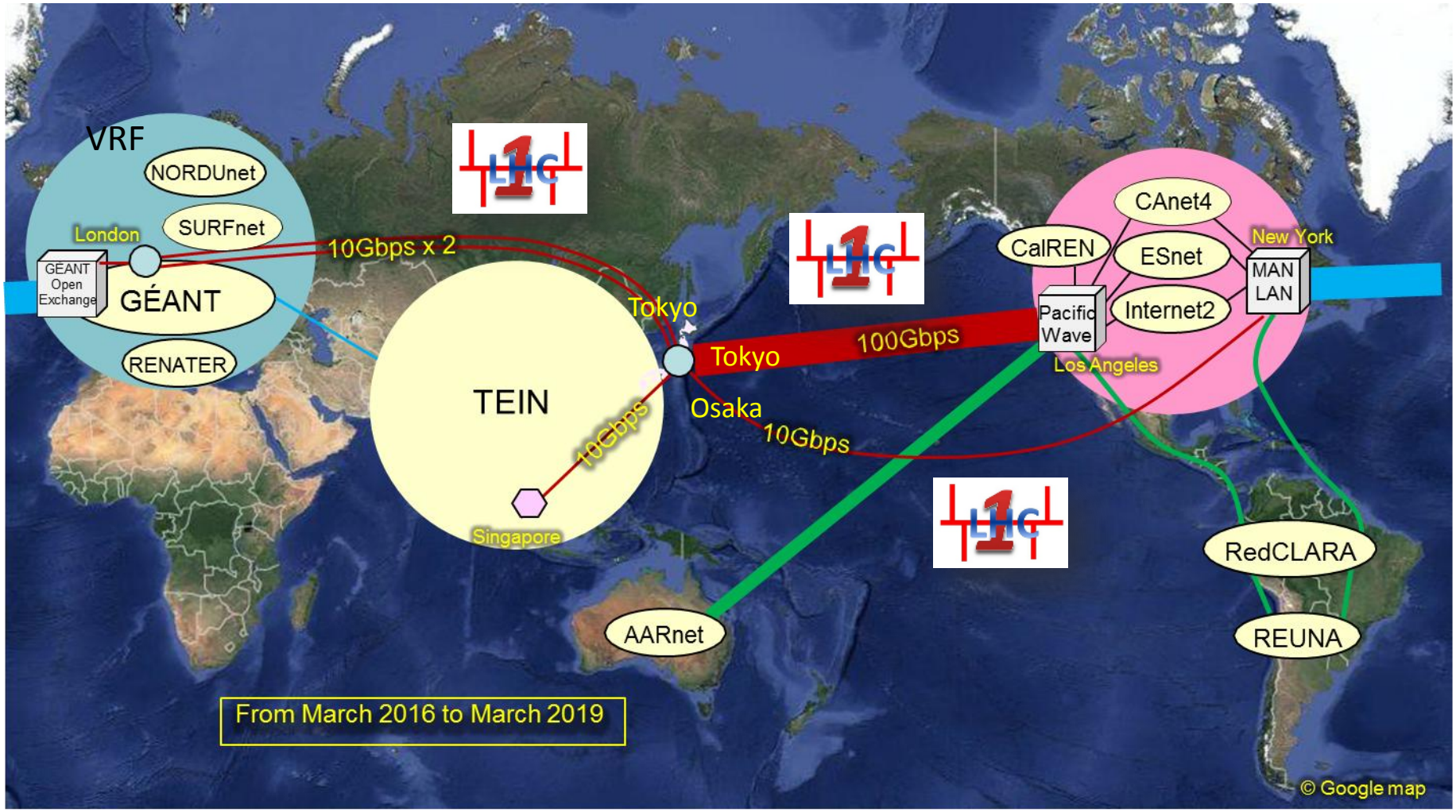
Almost reached to 10Gbps

Need to increase bandwidth and stability!





Upgrade (Apr. 2016) SINET5



Y. Kubota (NII)

Summary

System migration of Tokyo-Tier2 has been completed except for minor performance tuning (all of basic Grid service is already restarted).

Tokyo-Tier2 can provide enough computing resource for ATLAS for the next three years by the stable operation as ever.

The inter national network connectivity for Japan will be quite improved from Apri (Thanks to NII, Japanease NREN). Tokyo-Tier2 will also increase the bandwidth to the WAN.

Concerns for the next system migration after the three years operation:

- **Total data size and the number of files will be increased (8PB for Tier2).**
- **LAN bandwidth and I/O performance will not be enough for migration.**
- **CPU performance (per cost) will not be improved as before.**

Concept needs changing...