STATUS OF KISTI TIERS

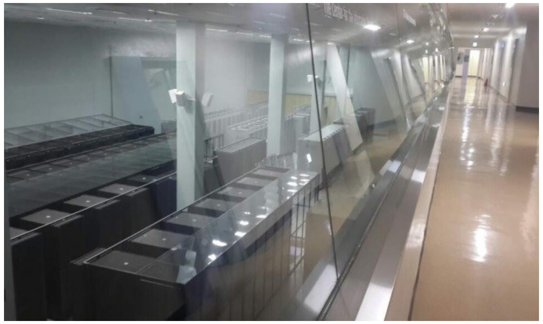
Sang-Un Ahn (sahn@kisti.re.kr)
On behalf of the GSDC

ALICE T1/T2 Workshop
University of Tsukuba, Tsukuba, Japan
5 March 2014

OUTLINE

- KISTI GSDC Overview
- Status of Tier-2
- Status of KIAF
- Status of Tier-1
- Plan & Summary

KISTI GSDC Overview



GSDC Facility

Korea Institute of Science and Technology Information - KISTI

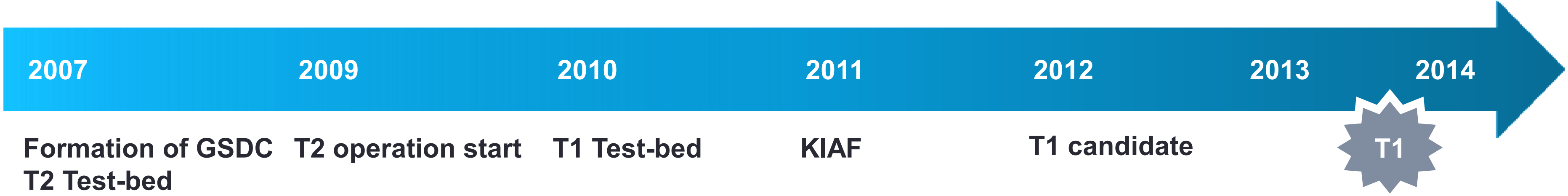
- National research institute for information technology since 1962
  - Around 600 people working for National Information Service (development & analysis), Supercomputing and Networking
- Running High-Performance Computing Facility
  - Total 3,398 Nodes (30,592 CPUs, 300 TFlops at peak), 1,667 TB storage (introduced from 2008)

Global Science experimental Data hub Center - GSDC

- National project to promote research experiment providing computing and storage resources: HEP and other fields of research
- Running Data-Intensive Computing Facility
  - 30 Staffs: system administration, experiment support, external-relations, administration and students
  - Total 484 Nodes (12k cores), 3,5 PB disk and 1 PB tape storage (accum. since 2008)

History of GSDC

- 5 years of the experience running grid computing centre with the collaboration with the ALICE experiment and WLCG



Formation of GSDC
T2 Test-bed

T2 operation start

T1 Test-bed

KIAF

T1 candidate

Status of Tier-2

Representative change (2013. 5): Dr. Jang, Haeng Jin

- Tier-2 operation is now included in the Tier-1 and KIAF project

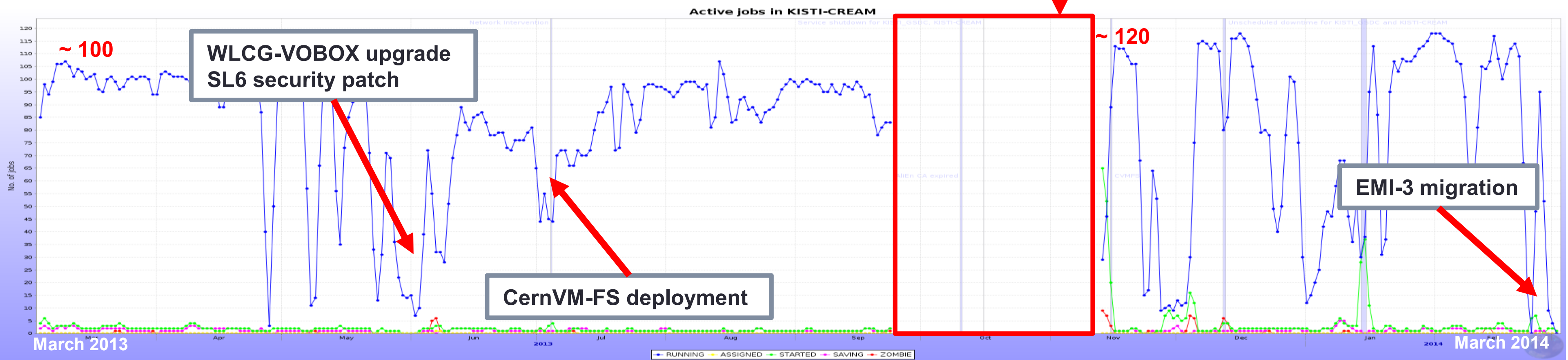
Resources for ALICE

- 120 job slots; 80 TB disk
- Pledges: 600 HS06, 50 TB (since 2011)
- EMI-3 middleware (done on 28th Feb. 2014), WLCG-VOBOX, PBS batch

Operational issues

- Frequent system down due to old system: mostly disk failure on node and storage
  - Main disk failure on 3 servers (end of maintenance contract): permanent job slot reduction (~100)
  - SE failure: old SAN volume configuration needed to be migrated, collapsing XROOTD capacity
- System migration done:
  - Server: HP Blade (2008) → IBM x3550 (2009)
  - Storage: NetApp (50 TB, SAN) → EMC (80 TB, NAS)

Internal network re-structuring
System migration



WLCG-VOBOX upgrade
SL6 security patch

CernVM-FS deployment

EMI-3 migration

Status of KIAF

KISTI Analysis Facility based on PROOF

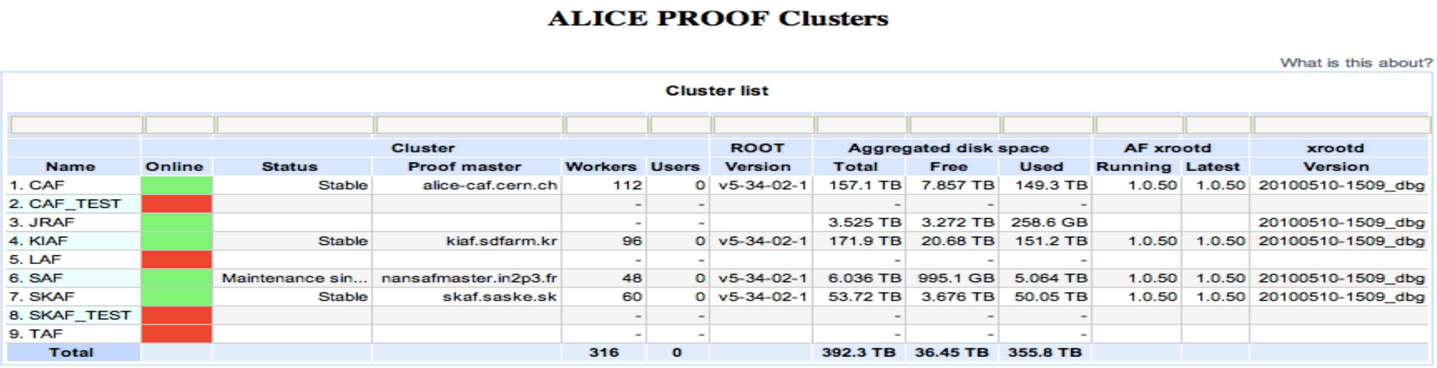
- In operation since 2011, ALICE use only

Resources

- 1 master, 8 worker nodes (12 workers per node = total 96 workers)
- Local disk storage to get better I/O performance: 22 TB disk per node (= total 198 TB)
- Similar size as CERN AF

Operational issue

- Heavy usage by Korean researchers (disk space for data is almost full)
- Synchronizing list of ALICE packages is not working properly
- Waiting for enabling CernVM-FS use in PROOF nodes



ALICE PROOF Clusters

What is this about?

Cluster list

	Cluster					ROOT	Aggregated disk space			AF xrootd		xrootd
Name	Online	Status	Proof master	Workers	Users	Version	Total	Free	Used	Running	Latest	Version
1. CAF	🟩	Stable	alice-caf.cern.ch	112	0	v5-34-02-1	157.1 TB	7.857 TB	149.3 TB	1.0.50	1.0.50	20100510-1509_dbg
2. CAF_TEST	🟥			-	-		-	-	-			
3. JRAF	🟩			-	-		3.525 TB	3.272 TB	258.6 GB			20100510-1509_dbg
4. KIAF	🟩	Stable	kiaf.sdfarm.kr	96	0	v5-34-02-1	171.9 TB	20.68 TB	151.2 TB	1.0.50	1.0.50	20100510-1509_dbg
5. LAF	🟥			-	-		-	-	-			
6. SAF	🟩	Maintenance sin...	nansafmaster.in2p3.fr	48	0	v5-34-02-1	6.036 TB	995.1 GB	5.064 TB	1.0.50	1.0.50	20100510-1509_dbg
7. SKAF	🟩	Stable	skaf.saske.sk	60	0	v5-34-02-1	53.72 TB	3.676 TB	50.05 TB	1.0.50	1.0.50	20100510-1509_dbg
8. SKAF_TEST	🟥			-	-		-	-	-			
9. TAF	🟥			-	-		-	-	-			
Total				**316**	**0**		**392.3 TB**	**36.45 TB**	**355.8 TB**			

Status of Tier-1

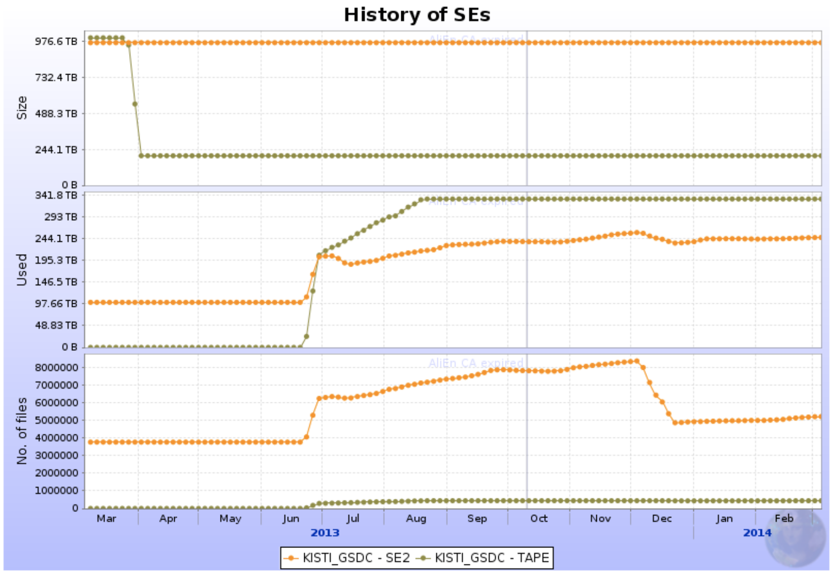
Full Tier-1 site

- Approved at the last WLCG Overview Board (15 Nov 2013)
- Special thanks to ALICE collaboration

Milestones

2010	10	Setup Tier-1 test-bed (100 job slots, 100 TB disk)
2011	11	1 PB disk attached
2012	3	Tier-1 candidate (ALICE Full membership at the same time)
	4	Dedicated 1 Gbps circuit (KISTI-CERN) established
	7	1,500 job slots
	10	Tape system delivered (1 PB)
	12	Tape system installed and operational
2013	1	Integration into WLCG APEL
	2	Integration into SAM tests for both OPS and ALICE
	3	Functional test done and RAW replication (310 TB) started
	4	Stable grid services (job capacity, SE) achieved (for at least 2 months)
	8	RAW replication finished
	9	Plan for 10 Gbps connectivity and on-call submitted to WLCG Management Board
	11	Approved as a full Tier-1 at WLCG Overview Board
	12	Dedicated 2 Gbps circuit (KISTI-CERN) established
2014	1	Integration into LHC OPN

Status of Tier-1

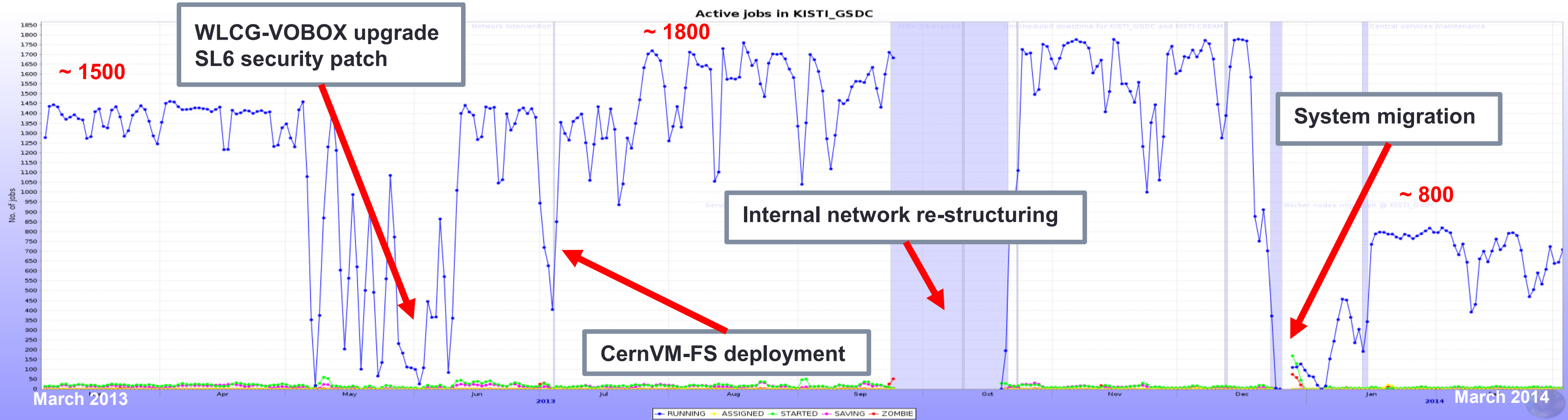


Resources for ALICE

- 832 Job slots (14.5 kHS06), 1 PB disk, 1 PB tape
- Pledges (2013): 25 kHS06, 1 PB disk, 1.5 PB tape
- EMI-2 middleware, WLCG-VOBOX, PBS batch, XRootD

Operational issues

- Long-shutdown (3-week) for internal network re-structuring (see next slide)
- Job slots reduction due to system migration for worker nodes (1,800 → 832)
  - Exchange between T1 nodes (no 10GbE card) and servers having 10GbE card (while serving other experiments)
  - 75 nodes (24 cores/node) providing 1,800 job slots were pulled out and provided to other services
  - Migrated to newly delivered 52 nodes (16 cores/node, HT off) providing 832 job slots



~ 1500

WLCG-VOBOX upgrade
SL6 security patch

~ 1800

CernVM-FS deployment

Internal network re-structuring

System migration

~ 800

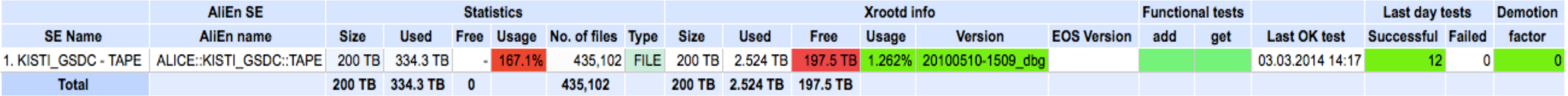
March 2013

March 2014

Status of Tier-1

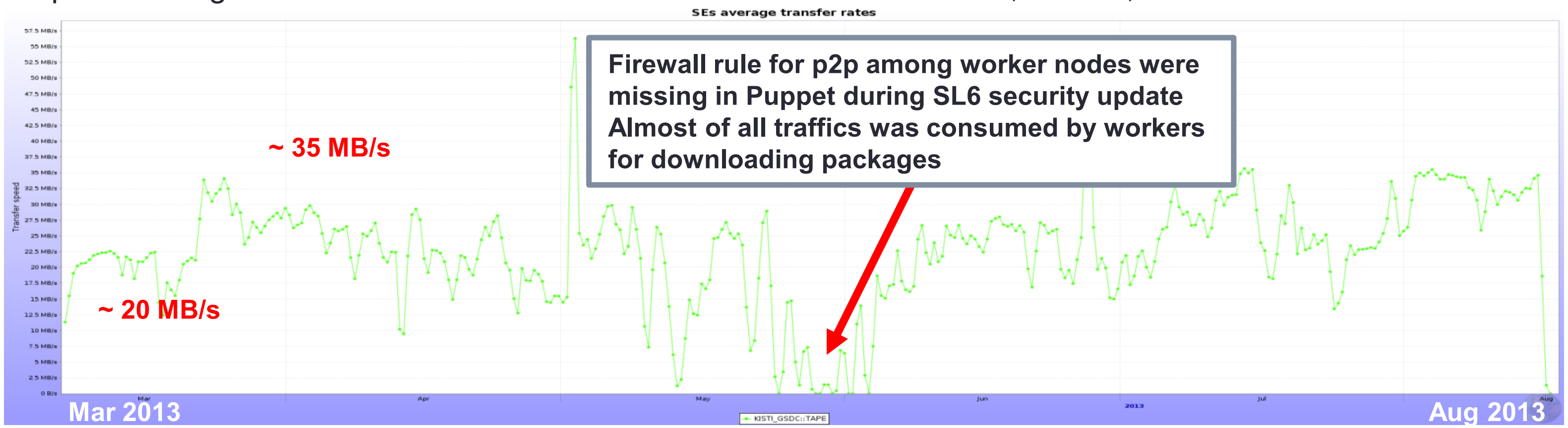
Tape

- 475 TB disk buffer: 200 TB for XRootD, 275 TB for GPFS
  - 10 XRootD servers with 1 redirector
  - 5 GPFS servers
- Home-made script performs data transfer between XRootD and GPFS via FRM
- ITLM (IBM) policy integrated in GPFS used for data migration towards tape
- Currently keeps 1% of data on disk buffer to prepare for server migration (details later)



	AliEn SE	Statistics						Xrootd info						Functional tests			Last day tests		Demotion
SE Name	AliEn name	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage	Version	EOS Version	add	get	Last OK test	Successful	Failed	factor
1. KISTI_GSDC - TAPE	ALICE::KISTI_GSDC::TAPE	200 TB	334.3 TB	-	167.1%	435,102	FILE	200 TB	2.524 TB	197.5 TB	1.262%	20100510-1509_dbg		🟩	🟩	03.03.2014 14:17	12	0	0
Total		200 TB	334.3 TB	0		435,102		200 TB	2.524 TB	197.5 TB									

- Performance issue concerning network during RAW replication
  - 22 MB/s on average; expected 60 MB/s with dedicated 1 Gbps link
  - Problem not understood; cannot investigate further since network configuration changed completely
  - Suspicious things: doubled network traffic due to use of NAS for XRootD, firewall, etc.



~ 35 MB/s

~ 20 MB/s

Firewall rule for p2p among worker nodes were missing in Puppet during SL6 security update
Almost of all traffics was consumed by workers for downloading packages

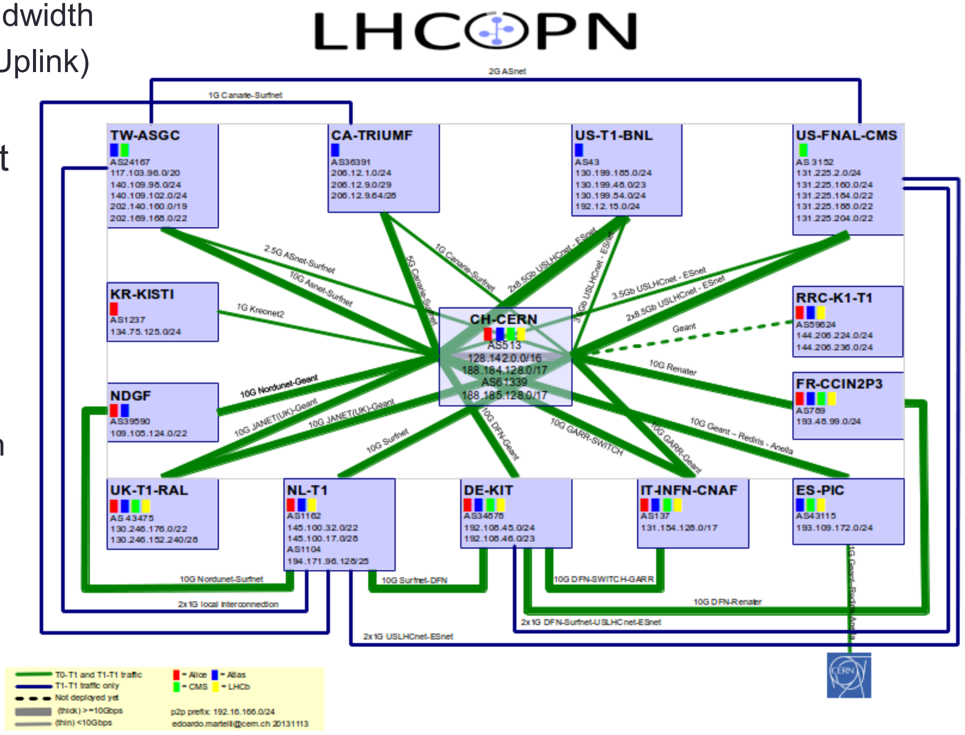
Mar 2013

Aug 2013

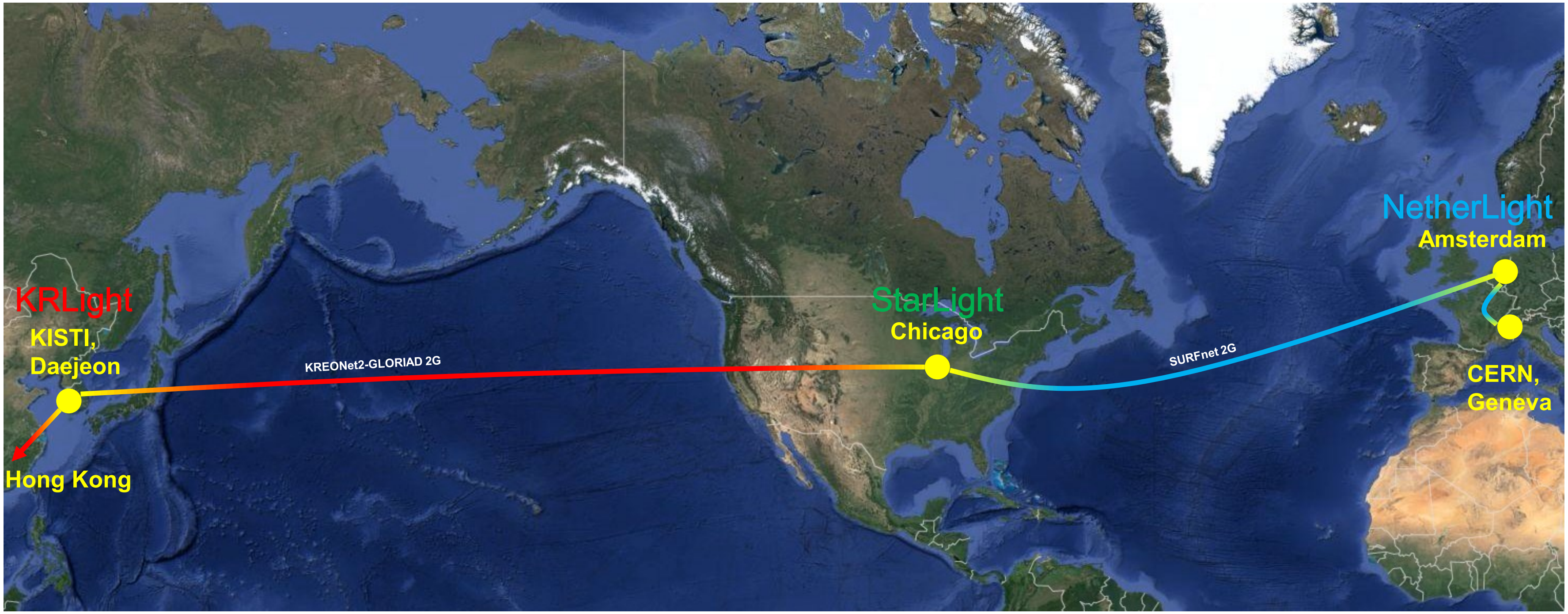
Status of Tier-1

Network

- Full 20 Gbps (Incoming/Outgoing) network configuration for internal network
    - Core switch upgrade: 960 Gbps → 2.5 Tbps bandwidth
    - Rack switches upgrade: 1 G → 10 G (10G * 16 Uplink)
    - Required 10GbE equipped servers
- Dedicated 2 Gbps circuit for KISTI-CERN Net
- Joined LHC-OPN
- perfSONAR deployed
    - 2 nodes at the same switch with tape XRootD
    - Trying to resolve network security rule at KISTI
    - perfSONAR test would help to ensure the best performance of the current network configuration



KISTI-CERN Network (2G)



NREN	KREONet2	SURFnet
Provider	KISTI	SURFnet
Section	Daejeon-Chicago	Chicago-Amsterdam-Geneva
Country	Korea, US	US, the Netherlands, Switzerland

✓ Thanks to Department of KISTI KREONET service

Plan

Pledges

- 25 kHS06
  - New machines were deployed → 52 nodes * 16 cores (HT off) = 832 job slots
  - Additional job slots will come by the end of March: 32 nodes * 32 cores (HT on) = 1024 job slots
  - Total 26 kHS06 (if HT on, 29 kHS06) will be provided
- 1.5 PB tape
  - Procurement for 500 TB will be proceed soon

Tape

- XRootD nodes and GPFS servers will be migrated and extended:
  - 20 XRootD servers with 2 headers for HA
  - 8 GPFS servers to have the best performance (= the same number of tape drives)
- Totally 600 TB will be allocated for disk buffer and all provided via SAN
  - In our experience, use of NAS as storage doubles the network traffic
- Rack switch will be replaced by 10 G

System migration

- Target to service nodes (e.g. CREAM-CE, WLCG-VOBOX, etc.) on old machines
  - End of maintenance contract by this year
  - Multiple (x2) CREAM-CE for HA

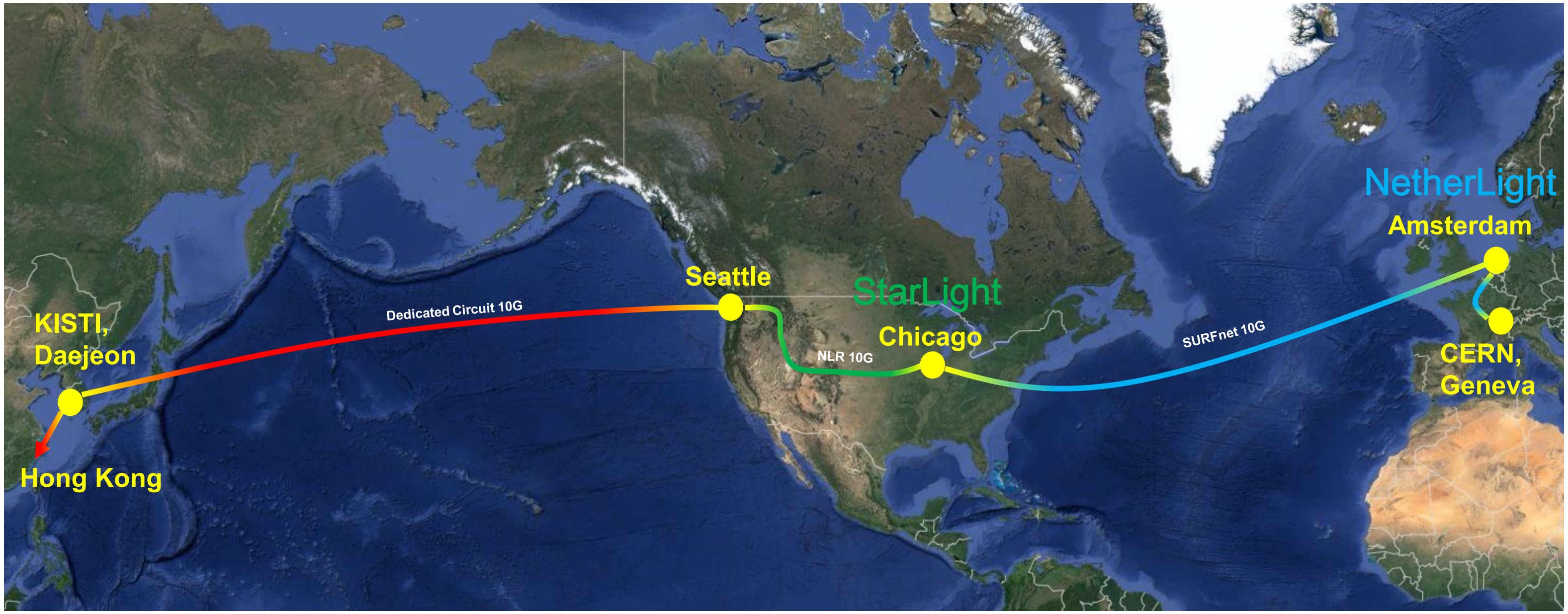
Plan

Middleware

- EMI-3 middleware migration by the end of March
  - Security support for EMI-2 will end by the end of April

Network

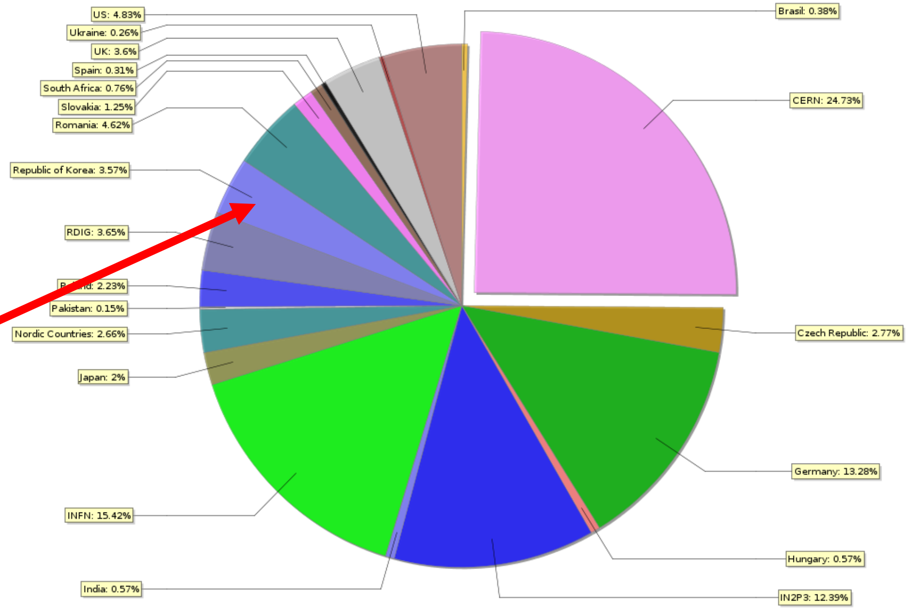
- Open tenders in May for a dedicated circuit between Daejeon to Seattle
  - Up to 10Gbps bandwidth
  - Cooperation with Department of KISTI KREONET service



Summary

- Smooth operations of T1 and T2
- Heavy migration activities in 2013
- Approval as a full Tier-1

Total wall clock hours for ALICE jobs (last year)



KISTI, 3.57 %
(Including Tier-2)

Summary of total resources for ALICE in 2014

Resource	# of Node	Phys. Cores	DISK	TAPE
T1	134	2,212	1,600 TB	1,500 TB
T2	22	88	80 TB	-
KIAF	9	108	198 TB	-
Total	165	2408	1878 TB	1,500 TB

THANK YOU
감사합니다

Questions?