# ALICE Tier-1 Status and Plan

*KoALICE National Workshop 2019 @ High1*
*6 January 2020*

Sang-Un Ahn
On behalf of KISTI-GSDC



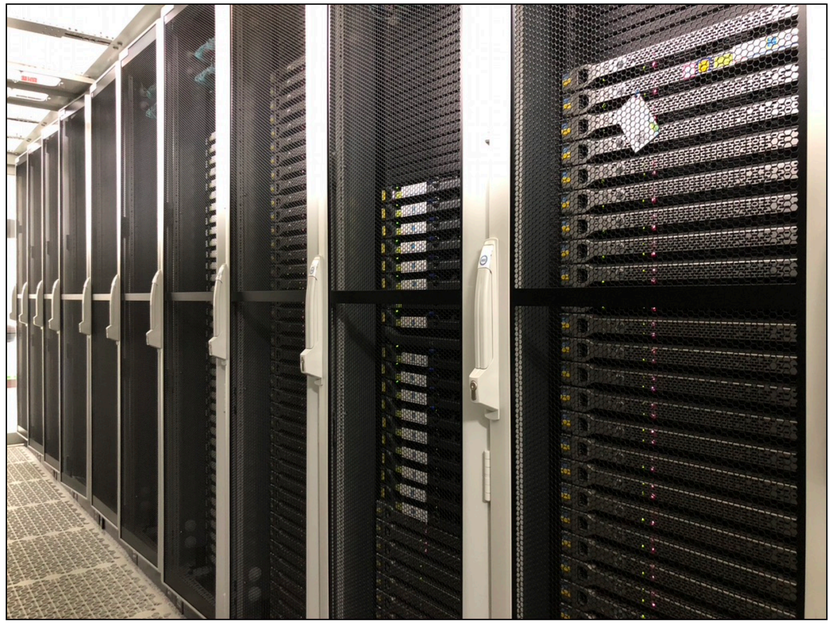

# Contents

- Operations
- Seeking an alternative to tape-based custodial storage - CHEP2019
- International Relations
- Plan & Summary

# Operations


국가슈퍼컴퓨팅센터


D
C



# WLCG Tier-1 System Topology

**Asia Tier Center Forum**

LHCOPN

IPv4 prefix = 134.75.125.0/24

**IPv6 prefix = 2001:320:15:125::/64**

Juniper MX80

Firewall

DELL Z9K

134.75.125.0/24

**80G**

perfSONAR

cream-ce (x3)
**htcondor-ce**
vobox (x2), site-bdii, argus, apel
eos-mgm (disk)
xrootd-redirector (disk)
squid-proxy (x2)

oVirt Cluster

Batch Cluster #1

Batch Cluster #2

Batch Cluster #3

**Batch Cluster #4**

XRootD Cluster (Disk)

EOS Cluster (Disk)

XRootD Cluster (Tape)

Brocade 6510

DELL Z9K

192.168.125.0/24

**80G**

**SAN (2000 TB)**

Dell Compellent

SAN (1500 TB)

EMC VNX

Tape Buffer (600 TB)

Hitachi VSP

Tape Library (3196 TB)

IBM TS3500

GPFS Cluster

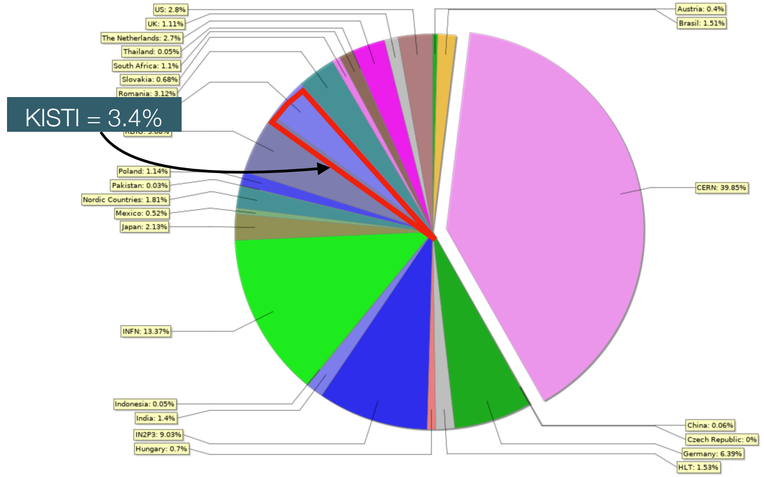
TSM

Public 10G
Private 10G
SAN 8G

# Tier-1 Operations Summary

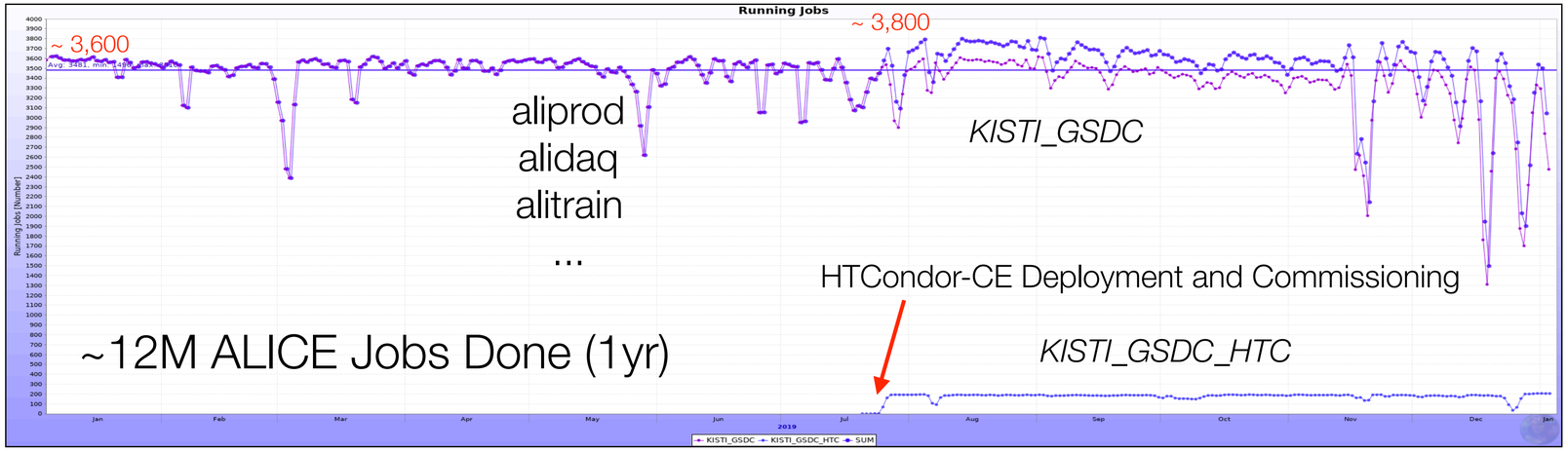
## ~ 3% Contribution to Total(T0+T1+T2+AF) ALICE Computing

Total wall clock hours for ALICE jobs (1yr)

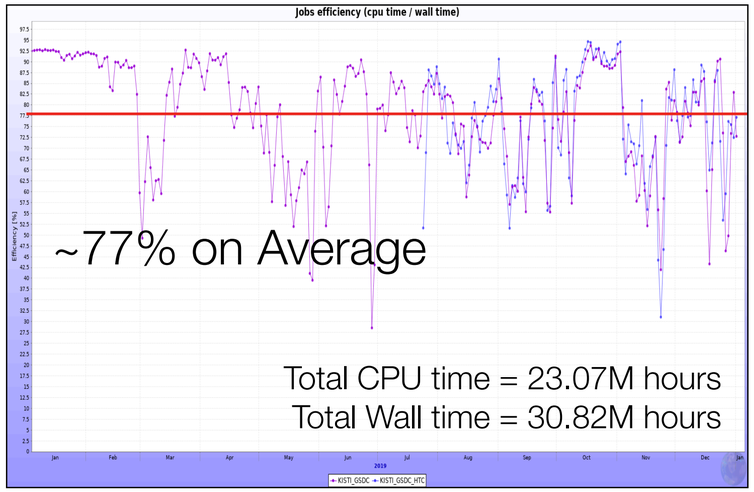



Running Jobs

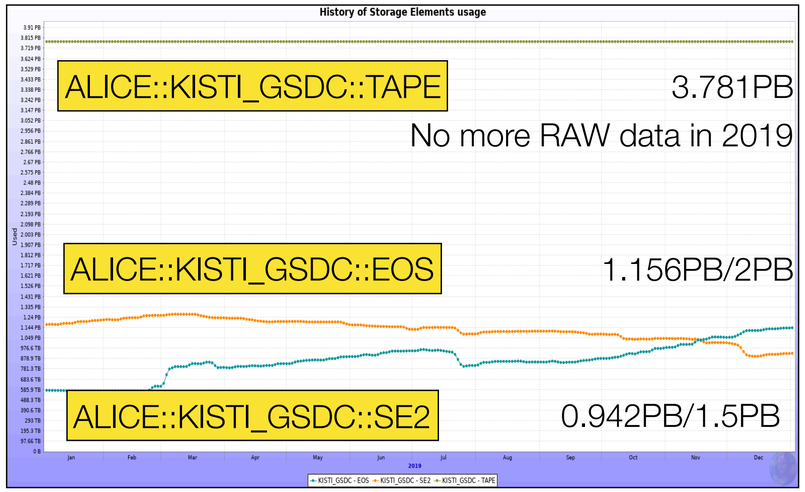
http://alimonitor.cern.ch?3008

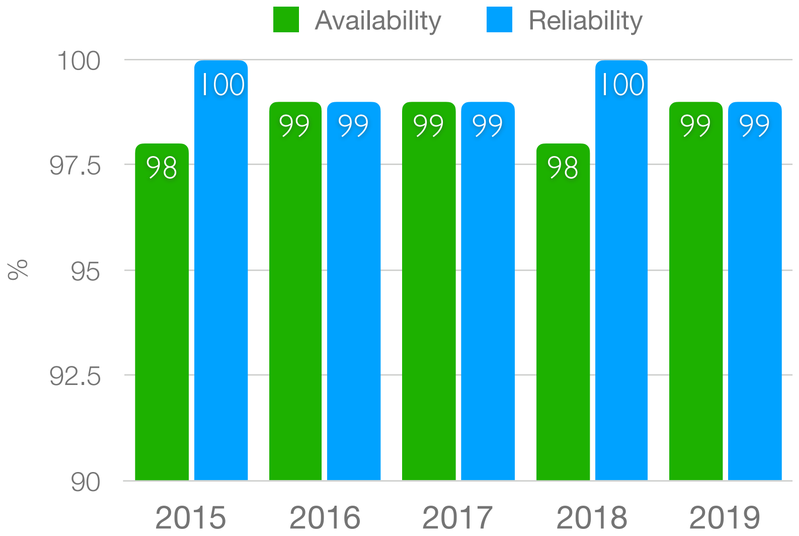



Job Efficiency




SE Usage



# Pledges

## 10% Contribution to ALICE Tier-1 Computing Requirements

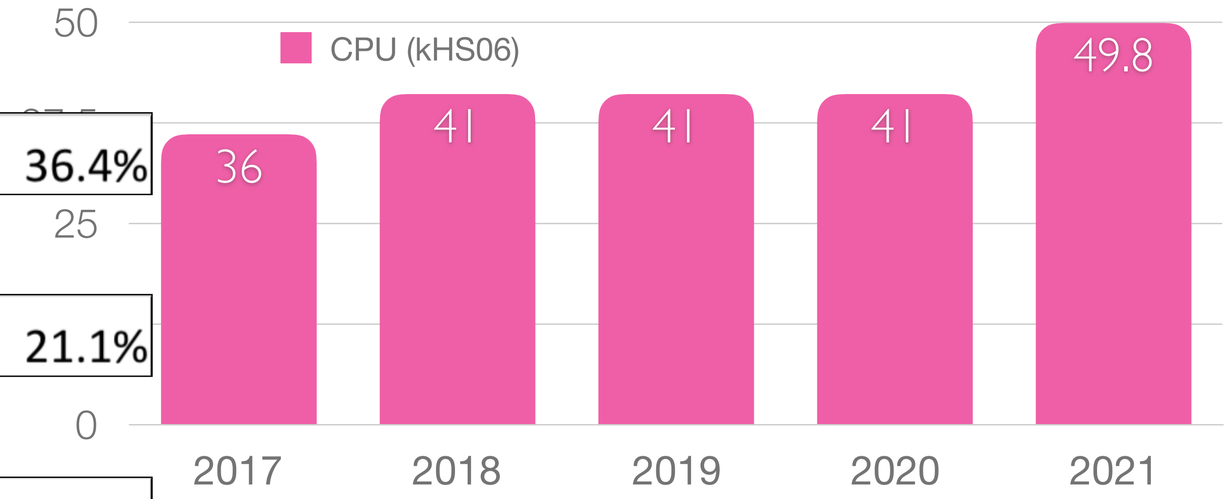
A Large Ion Collider Experiment

### Resources requirements projection

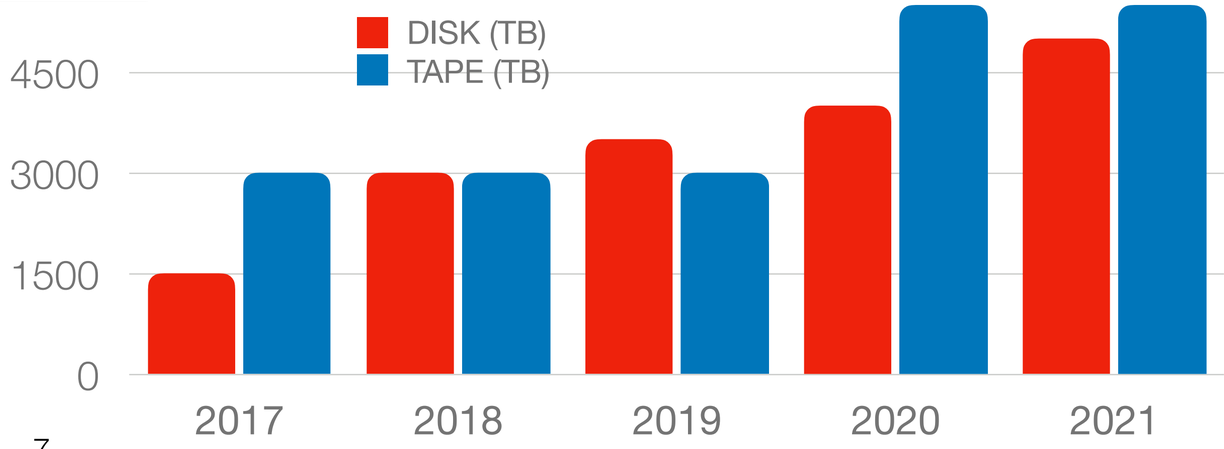
- Projections based on discrete resources simulation, includ
  detector performance and LHC beam schedule s
  growth (without tapes) - compatible with ***flat bud***

| | | 2019 | | | 2020 | | |
|---|---|---|---|---|---|---|---|
| ALICE | | Req. | C-RSG | Pledge | Req. | Pledge | 2020 Pledge/Req. |
| | Tier-0 | 430 | 430 | 350 | 350 | 350 | |
| | Tier-1 | 365 | 365 | 331 | 365 | 353 | |
| | Tier-2 | 376 | 376 | 370 | 376 | 410 | |
| CPU | **Total** | **1171** | **1171** | **1051** | **1091** | **1113** | **2.0%** |
| | Tier-0 | 34.3 | 34.3 | 31.2 | 31.2 | 31.2 | 0.0% |
| | Tier-1 | 37.9 | 37.9 | 35.1 | 44 | 41.8 | -5.0% |
| | Tier-2 | 33.9 | 33.9 | 33.5 | 39 | 41.0 | 5.0% |
| Disk | **Total** | **106.1** | **106.1** | **99.8** | **114.2** | **114.0** | |
| | Tier-0 | 44.2 | 44.2 | 44.2 | 44.2 | 44.2 | |
| | Tier-1 | 37.7 | 37.7 | 41.1 | 37.7 | 44.4 | |
| Tape | **Total** | **81.9** | 81.9 | **85.3** | **81.9** | **88.6** | **8.2%** |

| 2021 | |
|---|---|
| Req. | 2021/2020 Req |
| 498 | 36.4% |
| 515 | 37.0% |
| **1484** | **36.0%** |
| 53.3 | 21.1% |
| 44.8 | 14.9% |
| **143.6** | **25.7%** |
| 55.0 | 45.9% |
| **135.0** | **64.8%** |




| | 2017 | 2018 | 2019 | 2020 | 2021 |
|---|---|---|---|---|---|
| **CPU (cores)** | 3,800 | 3,800 | 3,800 | 3,800 | 4,500 |
| **DISK (TB)** | 1,500 | 1,500 | 3,500 | 4,000 | 5,000 |
| **TAPE (TB)** | 3,000 | 3,000 | 3,000 | 5,500 | 5,500 |



# Seeking an alternative to tape-based custodial storage

Sang Un Ahn[1], Latchezar Betev[2], Eric Bonfillou[2], Heejune Han[1], Jeongheon Kim[1], Seung Hee Lee[1], Bernd Panzer-Steindel[2], Andreas Joachim Peters[2], Heejun Yoon[1]

[1]KISTI, Daejeon, South Korea
[2]CERN, Geneva, Switzerland

*24th International Conference on Computing in High Energy and Nuclear Physics*
*4 - 8 November 2019*

Adelaide Convention Centre
Adelaide, Australia

# ATAS Project

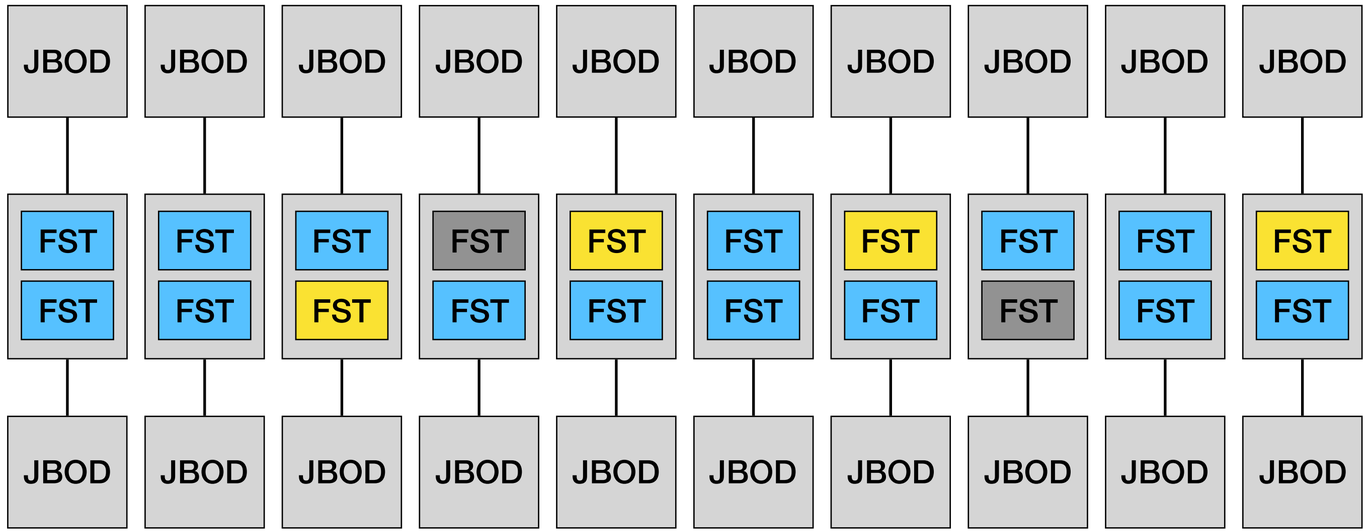
- Motivation:
  - Reduction operational costs of Tape-based custodial storage and risks of tape market, the monopoly of IBM & Japanese cartridge manufacturers (Sony/Fujifilm)
  - In accordance with the recent WLCG R&D activities and CERN EOS storage development - CERN abandoning 2 replica policy; Erasure-coding (software RAID) implementation
  - Hyper-converged infrastructure with cheap commodity hardware - JBOD (Just-Bunch-Of-Disks)
- CERN-KISTI R&D Collaboration:
  - Experts meetings @ KISTI & CERN focusing on design of disk-based custodial storage within a budget constraint ~ 1M USD
    - Market search (JBOD), Limitation study in the combination of SAS HBA and PCIe 3.0, Optimization on between data protection and usable capacity

# Initial System Design

- 10 EOS front-end node, each hosts 2 EOS FSTs, each EOS FST serves 1 JBOD box
  - EOS EC (M, K) = (14, 4) to balance between usable space (77.7% of physical capacity) and data security
  - Data loss probability ~ 0.000000005% (acceptable for ALICE)
- Each front-end node equipped with 2 SAS HBA cards (2 ports for each)
  - 1 HBA = 1 JBOD, SAS multi-path configuration to be tested for HA

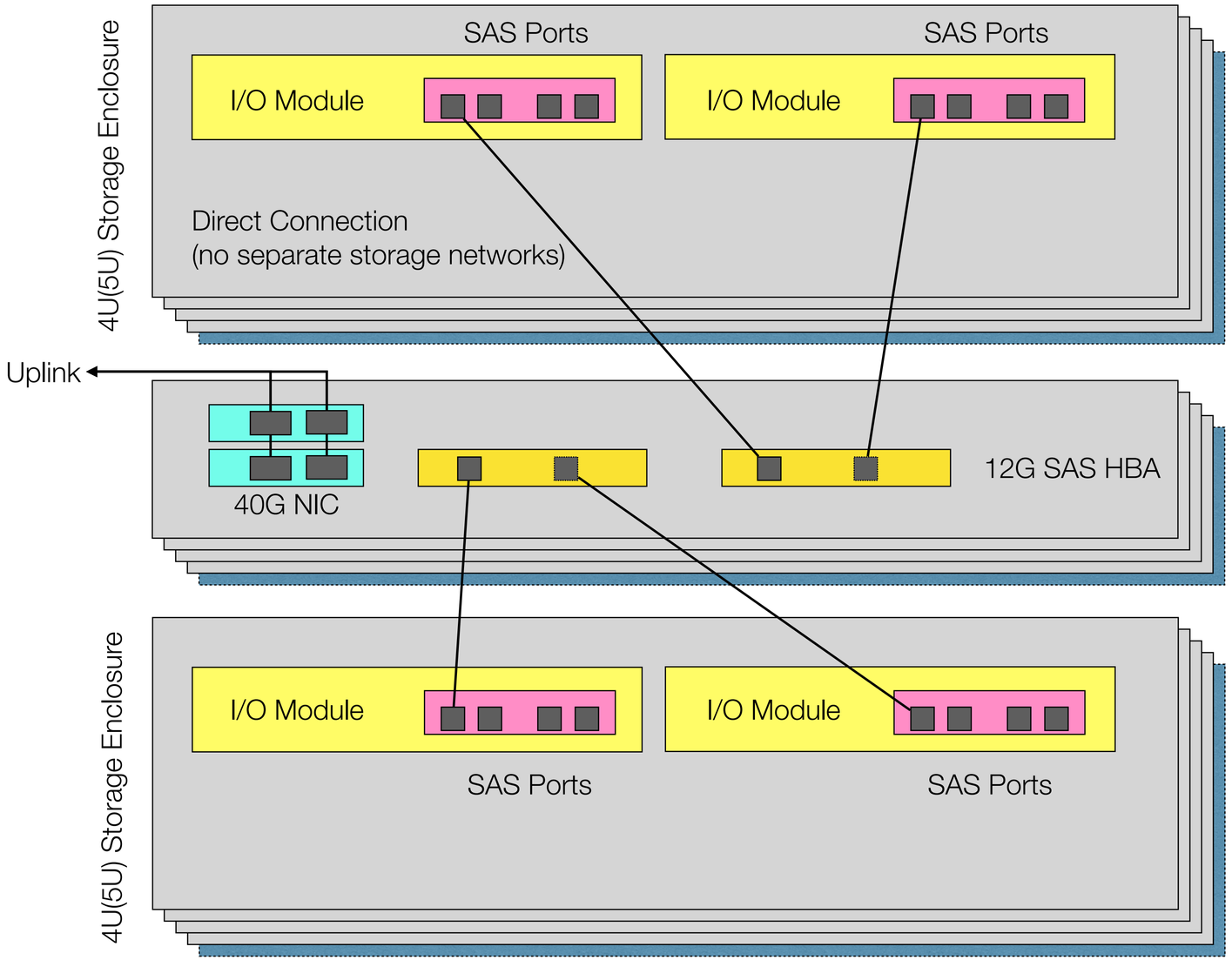
M = data node
K = parity node

EOS RAIN6 (14,4)




- (x2) EOS FSTs based on Docker container
- EOS decides where to store data fragments across FST nodes randomly (no fixed scheme)

# Deployment Setup




- This is a setup similar in all aspects to the CERN EOS current/future deployment


Specifications

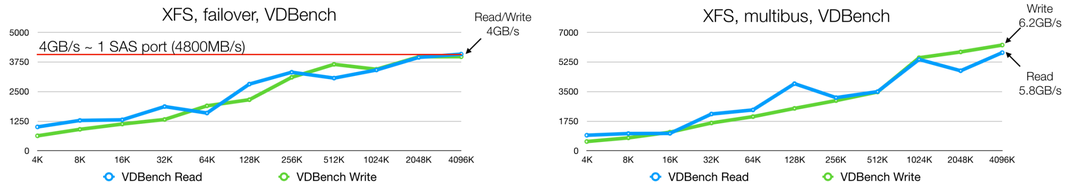
- x10 2U x86 servers
  - x2 40G NICs, x2(x4) 12G SAS HBA cards
- x2 40G network switches
- Even number of JBOD boxes filled up to 18PB

# Performance Test Results

## I/O Test: Multipath Mode

- Multipath mode: failover (active-standby) vs. multibus (active-active)
  - **multibus** mode showed the maximum I/O speed up to 6GB/s for read/write
    - Bottleneck on PCIe 3.0 (6400MB/s)
  - **failover** could not fulfill the available bandwidth, limited under 1 SAS port (48Gb) pipe

I/O Test: VDBench

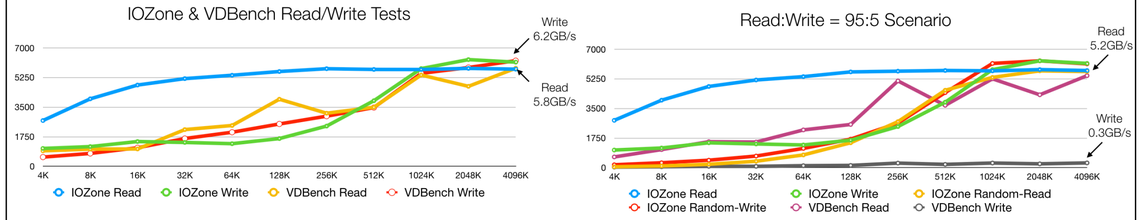



18

## I/O Test: Read/Write

- XFS read/write performance (simultaneous read and/or write from 70 disks)
  - **VDBench** shows full read/write transfer performance @ transfer size >= 2048k (6GB/s)
  - **IOZone** shows full read/write transfer performance @ transfer size ~ 2048k (6GB/s)

Disk: 70EA
Filesize: 2GB

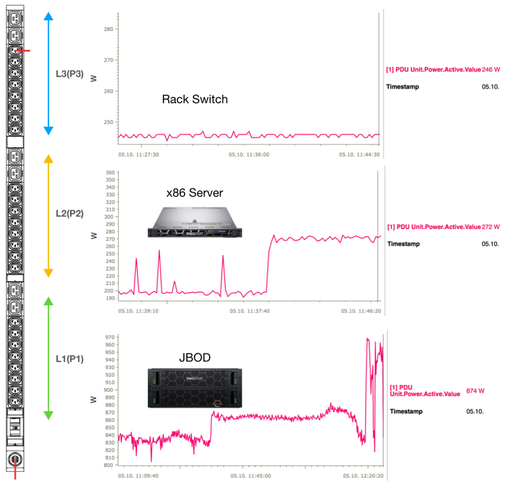



* IOZone tests with different Read/Write ratio Scenario did not much affect on the performance

19

## Power Consumption

- JBOD Test Equipment (70 Disks)
  - JBOD (DELL ME484): idle = 830W; load = 860W (Max 960) **(1.12W/TB)**
  - Server: idle = 200W; load = 270W
  - Switch: idle = 246W; load = 246W
  - **1.75W/TB** including JBOD, Server and Switch
- Disk Storages (Full Load)
  - DellEMC SC7020, 2.5PB - 12,120W **(4.8W/TB)**
  - EMC Isilon, 16 Nodes, 2.95 PB- 13,730W **(4.6W/TB)**
  - EMC VNX, 12 Nodes, 2.36 PB - 5,100W **(2.2W/TB)**
  - HITACHI VSP, 2 PB - 18,300W **(9.15W/TB)**
  - EMC Isilon, 15 Nodes, 1.43 PB - 12,880W **(9W/TB)**
  - EMC CX4-960, 1.5PB - 14,900W **(9.9W/TB)**
- Tape Library (Full Load)
  - **IBM TS3500 5-Frame (3.2PB) - 1,600W (0.5W/TB)**




20

- Confirmed the upper cap of read/write performance ~ 6GB/s (intrinsic limit by PCIe 3.0)
- Power consumption shown ~ 1.75W/TB, not uncomfortably higher than Tape (0.5W/TB)
  - High-end Enterprise Class Storage 5 ~ 9W/TB

# International Relations

# 5th Asia Tier Center Forum & 1st Asia HTCondor workshop

**24-26 October 2019.**

**Jointly organized by TIFR Mumbai and KISTI, South Korea**

**Venue: TIFR, Mumbai India.**

http://indiacms.res.in/atcf5.html

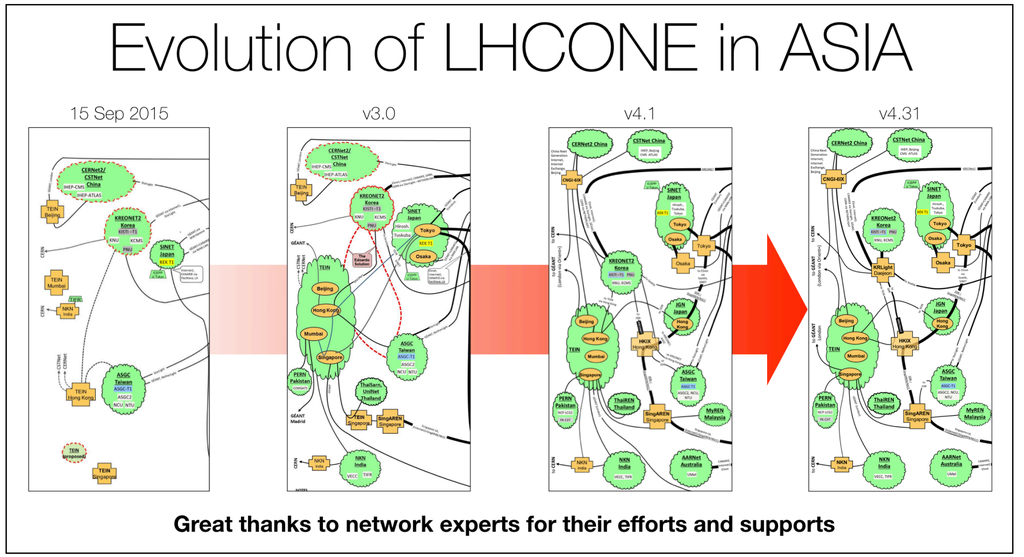
**Registration - https://indico.cern.ch/e/atcf5**

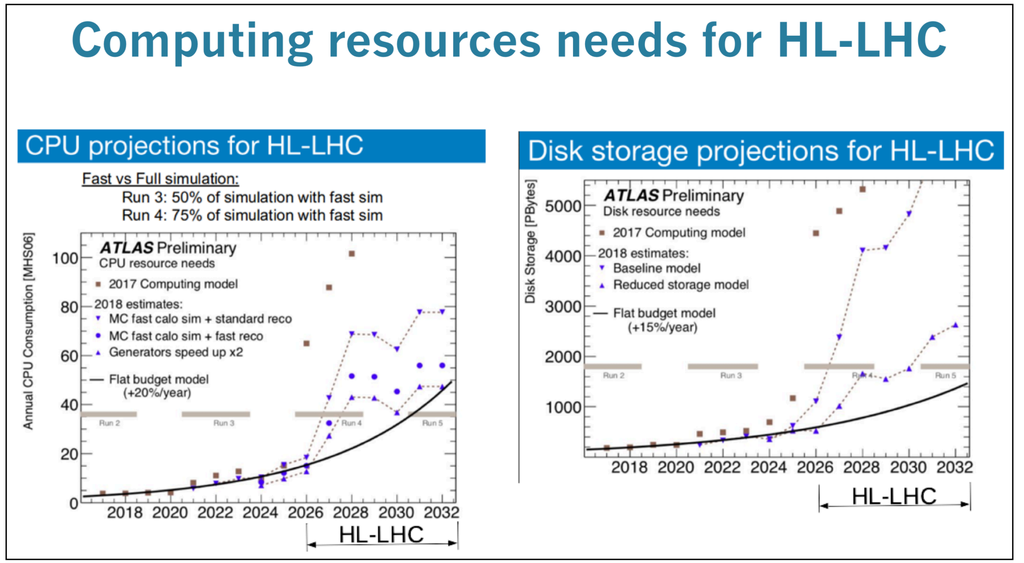
tifr

KiSTi
www.kisti.re.kr
Korea Institute of
Science and Technology Information

# Asia Tier Center Forum

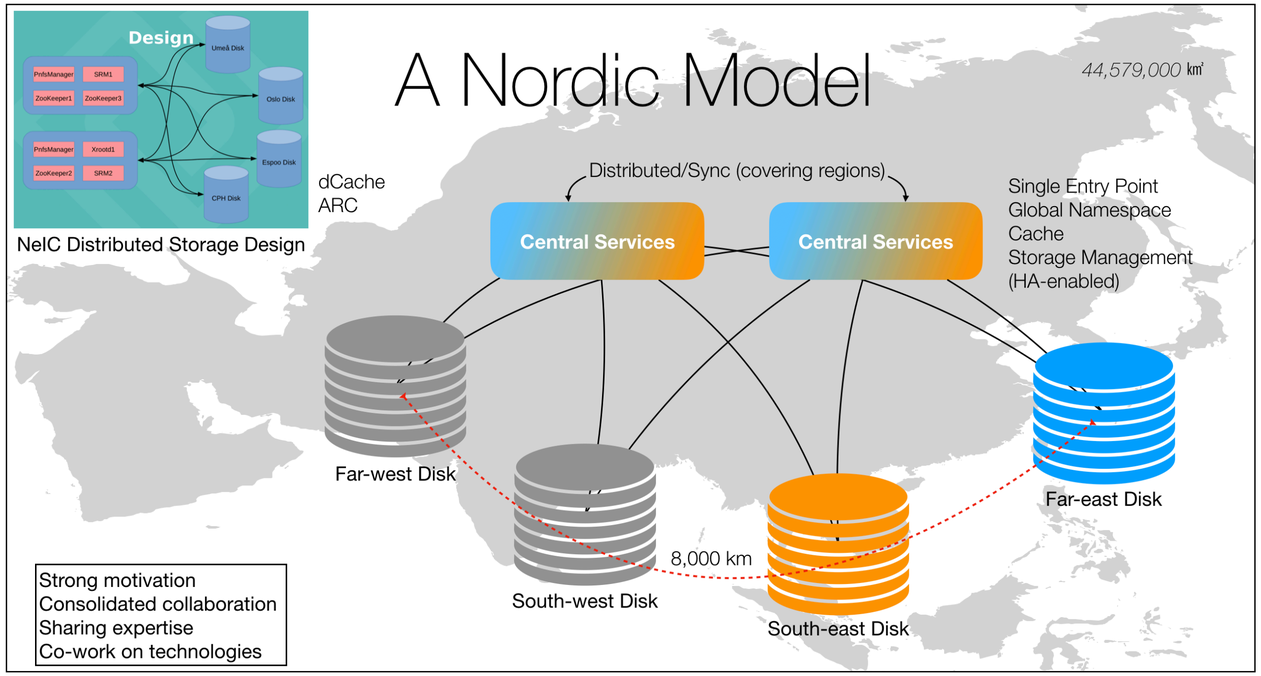
- Started in 2015 led by KISTI, focusing on Asian-wise issues: enhancing network connectivities among regional sites
  - Great success on establishing LHCONE network in the region
  - The fifth event held at TIFR in Mumbai, India - Visit atcforum.org
- Emerging agenda: distributed storage spanning the region
  - WLCG Tier becomes blurred; network-driven disruptive paradigm change - Nucleus-Satellite model, storage consolidation, caching => WLCG DOMA
  - Flat budget scenario, harder to deliver what the LHC experiments require for RUN3, RUN4 and beyond
    - Innovation on the site operations and management are key to reduce the costs and the consolidated efforts are needed

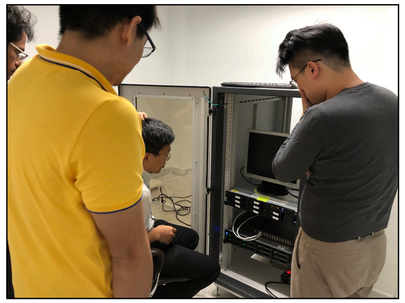


# KISTI-SUT Distributed Storage

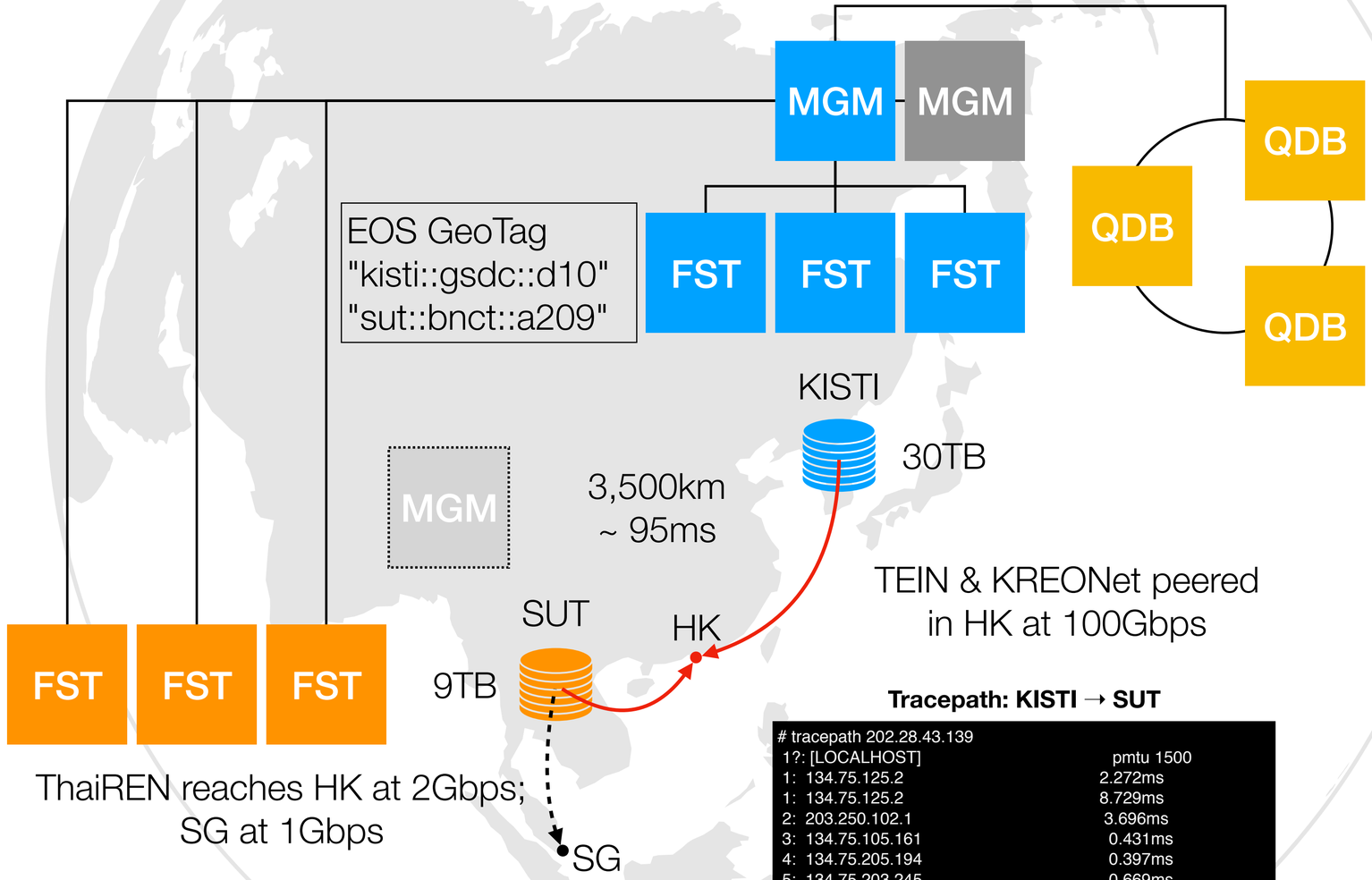
- Motivation:
  - Pursuing the technology evolution in WLCG and answer to the questions e.g. what the benefit of storage consolidation to Asian sites, how we could realise the cost reduction
- The working model: NeIC (NDGF), CloudStor (AARNet)
- Technology: EOS, Docker, Ansible, LHCONE
- Pilot deployment done in August 2019
  - 3-day workshop @ SUT in Nakhon Ratchasima, Thailand
  - Training program in parallel for students: EOS deployment based on Docker container using Ansible playbook



# Topology

- EOS @ KISTI
  - MGM (Master/Slave)
  - QuarkDB cluster (3 nodes)
  - 3 FSTs (30TB HDD NAS)
- EOS @ SUT
  - 3 FSTs (9TB SSD NAS)
- EOS Instance Name = testatcf




```
[root@eos-mgm-01 /]# eos fs ls
```

| host | port | id | path | schedgroup | geotag | boot | configstatus | drain | active | health |
|---|---|---|---|---|---|---|---|---|---|---|
| eos-fst-0001.eoscluster.sdfarm.kr | 1095 | 1 | /data/disk0001 | default.0 | kisti::gsdc::d10 | booted | rw | nodrain | online | N/A |
| eos-fst-0002.eoscluster.sdfarm.kr | 1095 | 2 | /data/disk0002 | default.0 | kisti::gsdc::d10 | booted | rw | nodrain | online | N/A |
| eos-fst-0003.eoscluster.sdfarm.kr | 1095 | 3 | /data/disk0003 | default.0 | kisti::gsdc::d10 | booted | rw | nodrain | online | N/A |
| eos-fst-0004.eoscluster.sdfarm.kr | 1095 | 4 | /data/disk0004 | default.0 | sut::bnct::a209 | booted | rw | nodrain | online | N/A |
| eos-fst-0005.eoscluster.sdfarm.kr | 1095 | 5 | /data/disk0005 | default.0 | sut::bnct::a209 | booted | rw | nodrain | online | N/A |
| eos-fst-0006.eoscluster.sdfarm.kr | 1095 | 6 | /data/disk0006 | default.0 | sut::bnct::a209 | booted | rw | nodrain | online | N/A |

**Tracepath: KISTI → SUT**

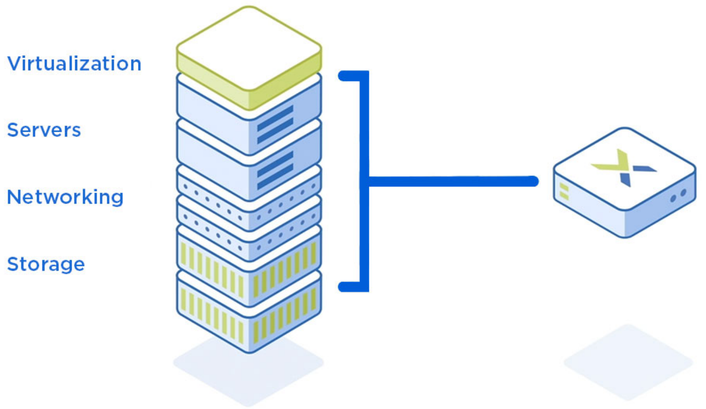
```
# tracepath 202.28.43.139
 1?: [LOCALHOST]                      pmtu 1500
 1:  134.75.125.2                      2.272ms
 1:  134.75.125.2                      8.729ms
 2:  203.250.102.1                     3.696ms
 3:  134.75.105.161                    0.431ms
 4:  134.75.205.194                    0.397ms
 5:  134.75.203.245                    0.669ms
 6:  134.75.203.241                    0.976ms
 7:  134.75.203.18                    39.954ms
 8:  202.179.241.205                  44.706ms
 9:  202.179.241.210                  91.354ms
10:  pyt-to-02-bdr-pyt-link-1.uni.net.th   91.229ms
11:  100.64.253.13                    96.071ms asymm 14
12:  202.28.208.254                   94.953ms asymm 16
13:  202.28.43.139                    95.587ms reached
     Resume: pmtu 1500 hops 13 back 17
```

# Plan & Summary

# Plan

- Operations:
  - VM environment migrating to Hyper-converged Infrastructure (oVirt 4.3)
    - A software-defined infrastructure virtualizing all conventional hardware systems
  - Upgrade to CentOS 7 or 8 (if applicable) for all Grid services and Batch clusters
- ATAS Project:
  - EOS Workshop @ CERN in Feb - Presenting ATAS status and KISTI-SUT DS
  - LBL (ALICE T2) interested in implementing ATAS design, collaboration meeting scheduled in Feb
  - Targeting in production before the start of RUN3 in 2021
- ATCF:
  - ATCF6 venue and schedule (TBD)
  - Re-deployment of distributed storage hardening HA on EOS management, Network tuning, New site - University of Tokyo (ICEPP, ATLAS T2)

Hyper-converged Infrastructure



# Summary

- Flawless KISTI Tier-1 operations for ALICE experiment
- Continuous resource growth to meet the ALICE computing requirement for LHC RUN3, RUN4 and beyond
- Substantial on-going international projects and collaboration pursuing technology evolution in accordance with WLCG R&D activities

# Q & A