ALICE computing and network in Japan

Tatsuya Chujo
University of Tsukuba

AFAD 2016
February 1, 2016
Kyoto University, Uji campus

Outline

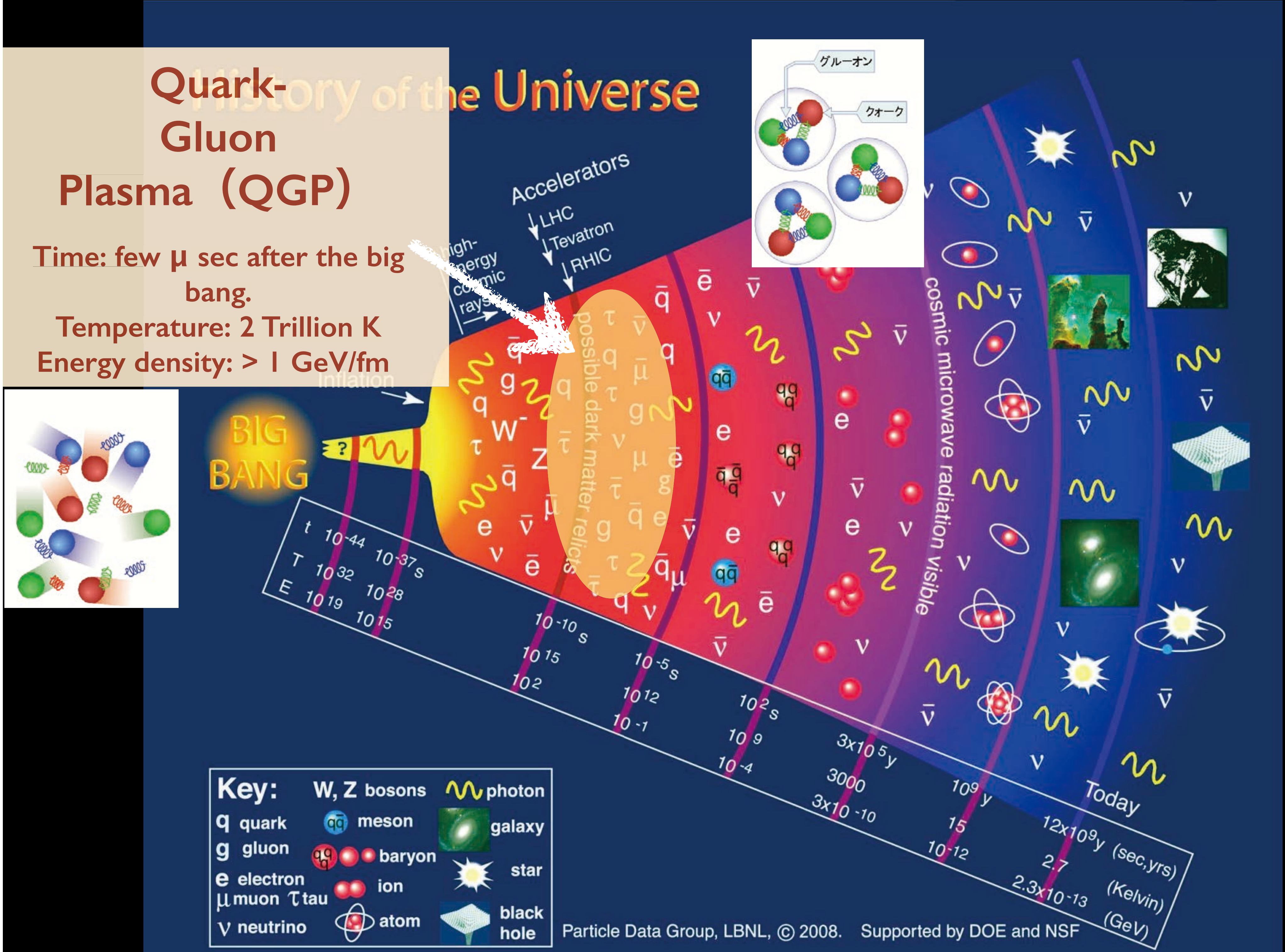
1. Introduction
   - ALICE computing in Run-1 and Run-2
2. ALICE computing in Run-3 and Run-4 (2021-)
3. Current ALICE O^2 project status
4. T2 site(s) in Japan and network
5. Summary

Quark-Gluon Plasma（QGP）

Time: few μ sec after the big bang.

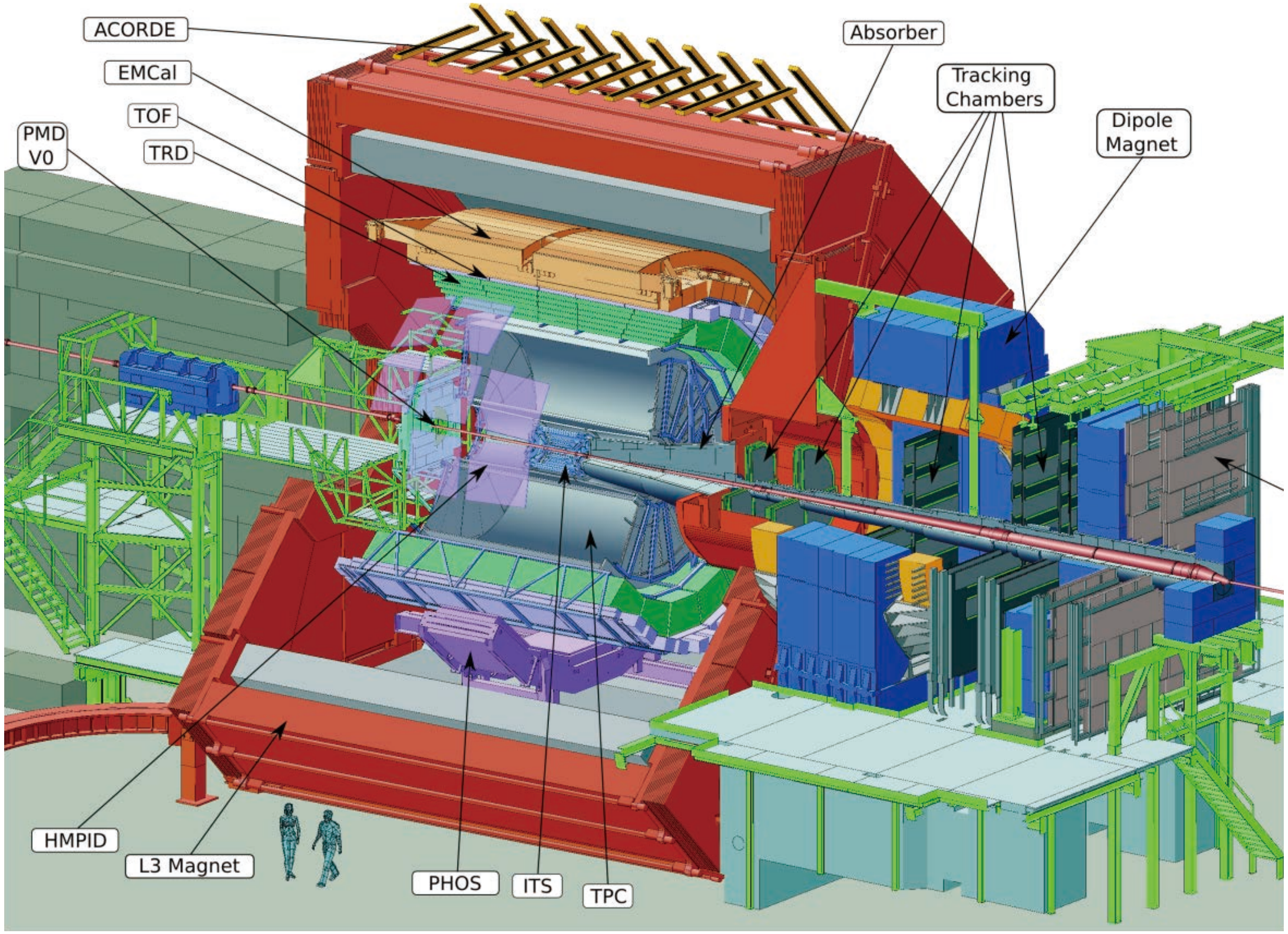
Temperature: 2 Trillion K

Energy density: > 1 GeV/fm

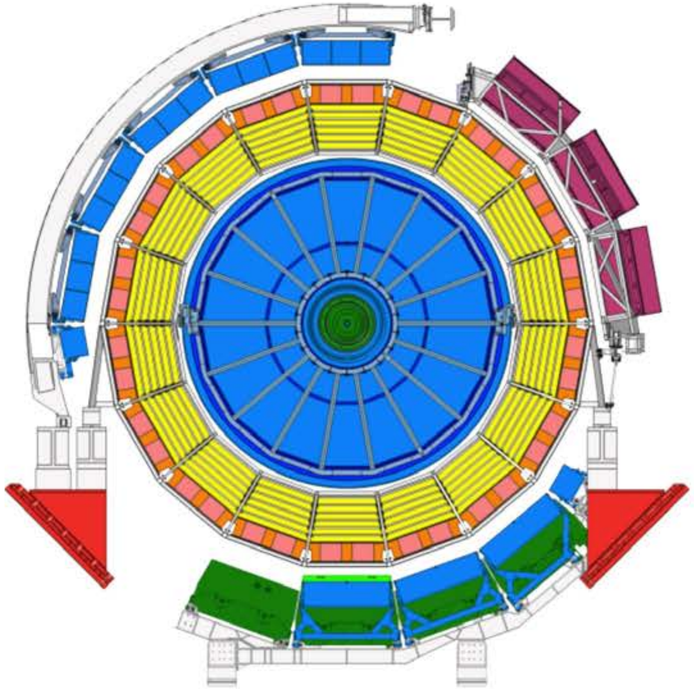


Particle Data Group, LBNL, © 2008. Supported by DOE and NSF

ALICE Experiment



16m x 16m x 26m, 10,000 tons

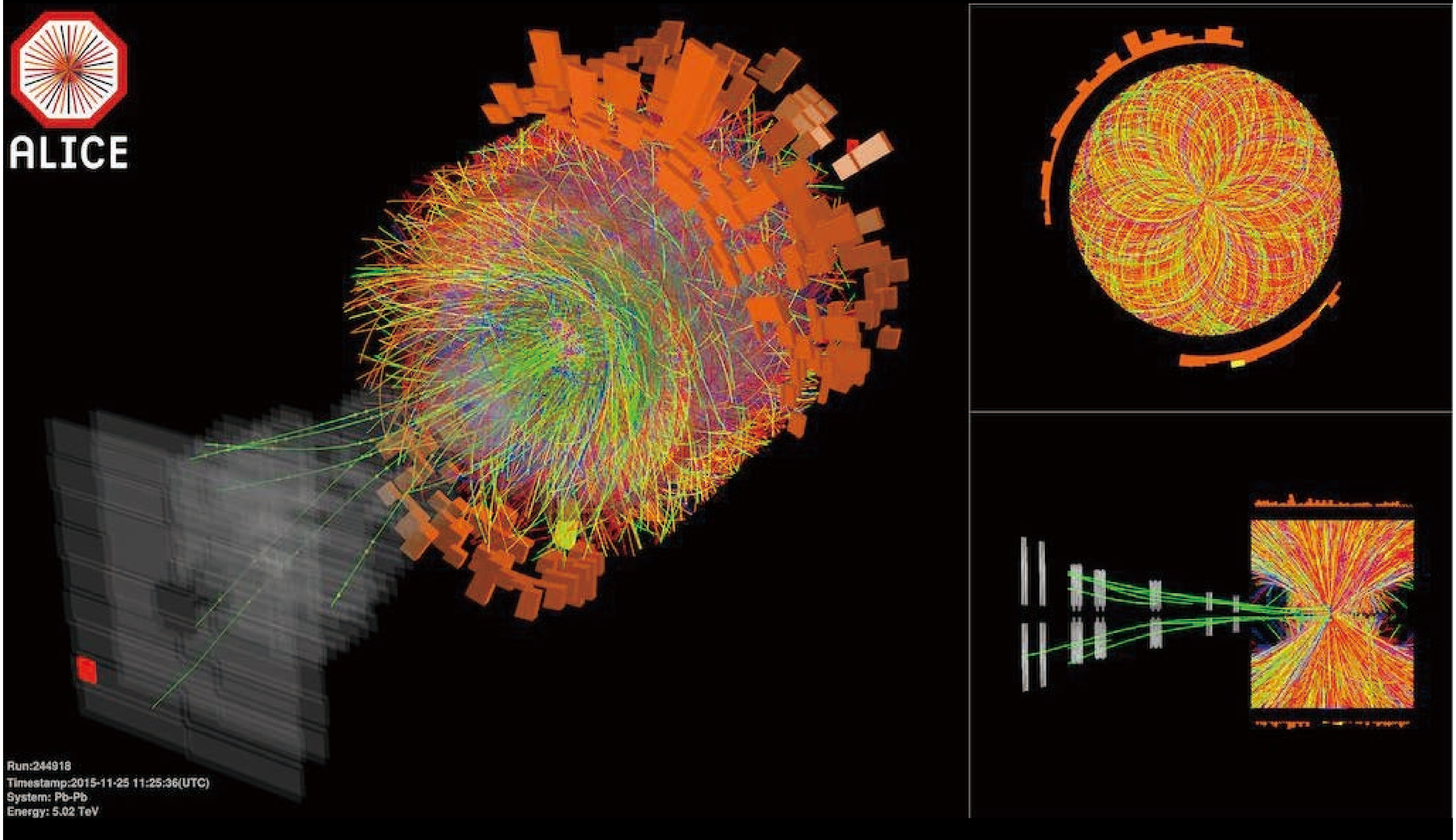


18 different sub-detectors:
tracking, particle identification, energy measurement, event trigger

>1400 scientists from 149 Institutes in 40 Countries

The dedicated experiment in LHC experiments to study Quark Gluon Plasma (QGP) by using heavy ion beams.

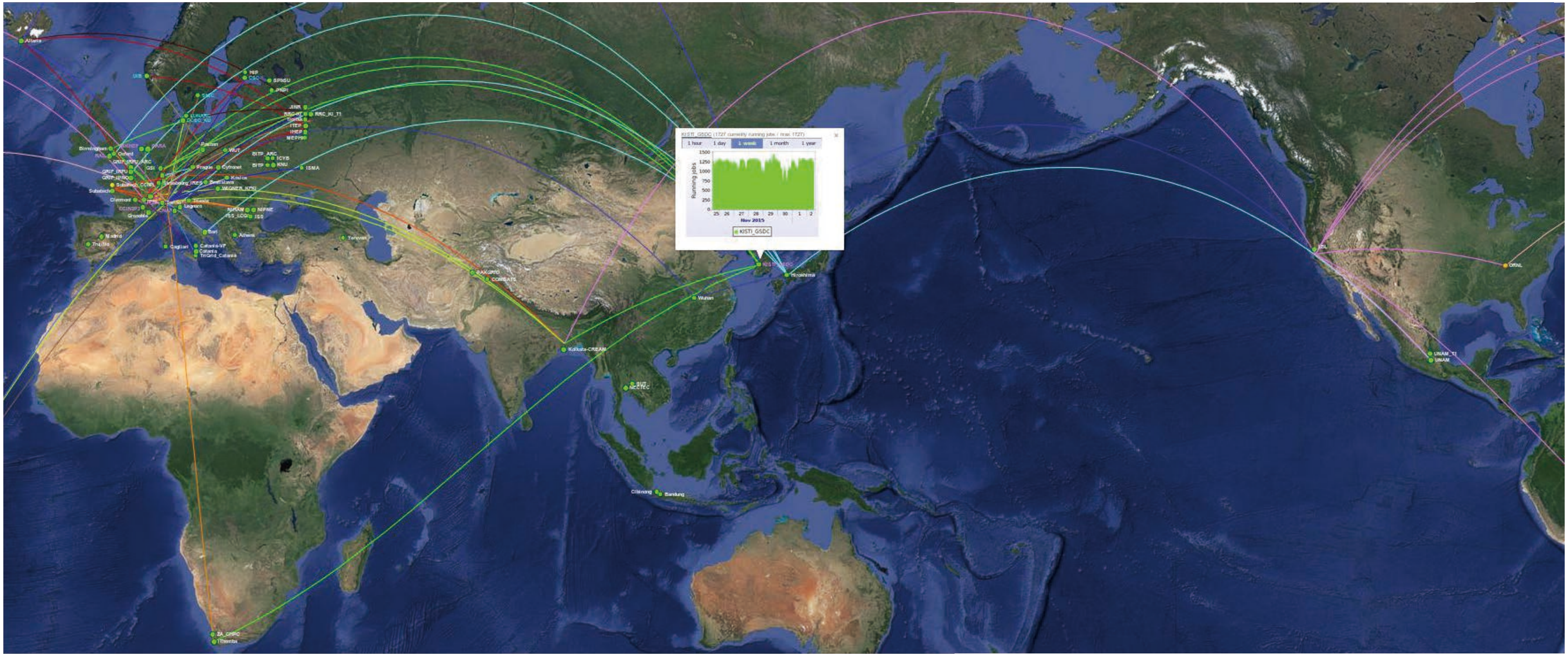
Pb-Pb 5.02 TeV (One PeV collisions , Nov. 2015) !!



Current ALICE computing (some numbers)

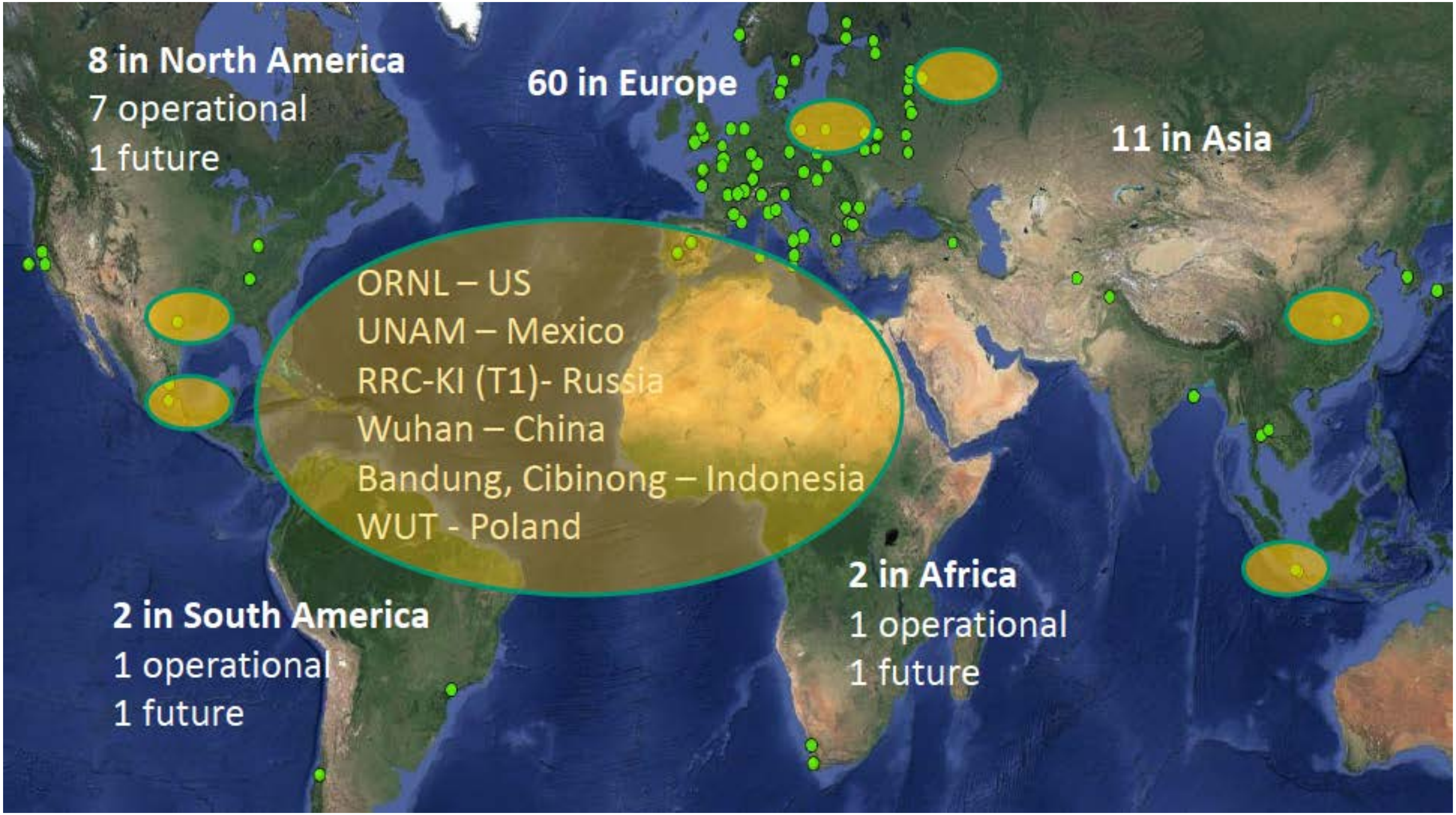
- ▶ 80 computing centers around the world
  - T0, T1, T2
- ▶ Up to 100K concurrently running jobs
- ▶ 600M executed jobs
  - 1600 users
- ▶ 10 Tape + 56 Disk storage elements
  - 25 + 25 PB of data, 1B+ files
- ▶ Up to 40 GB/s read rates (10GB/s avg.)
  - Writing at 1/10th the read rate

Distributed resources on Grid



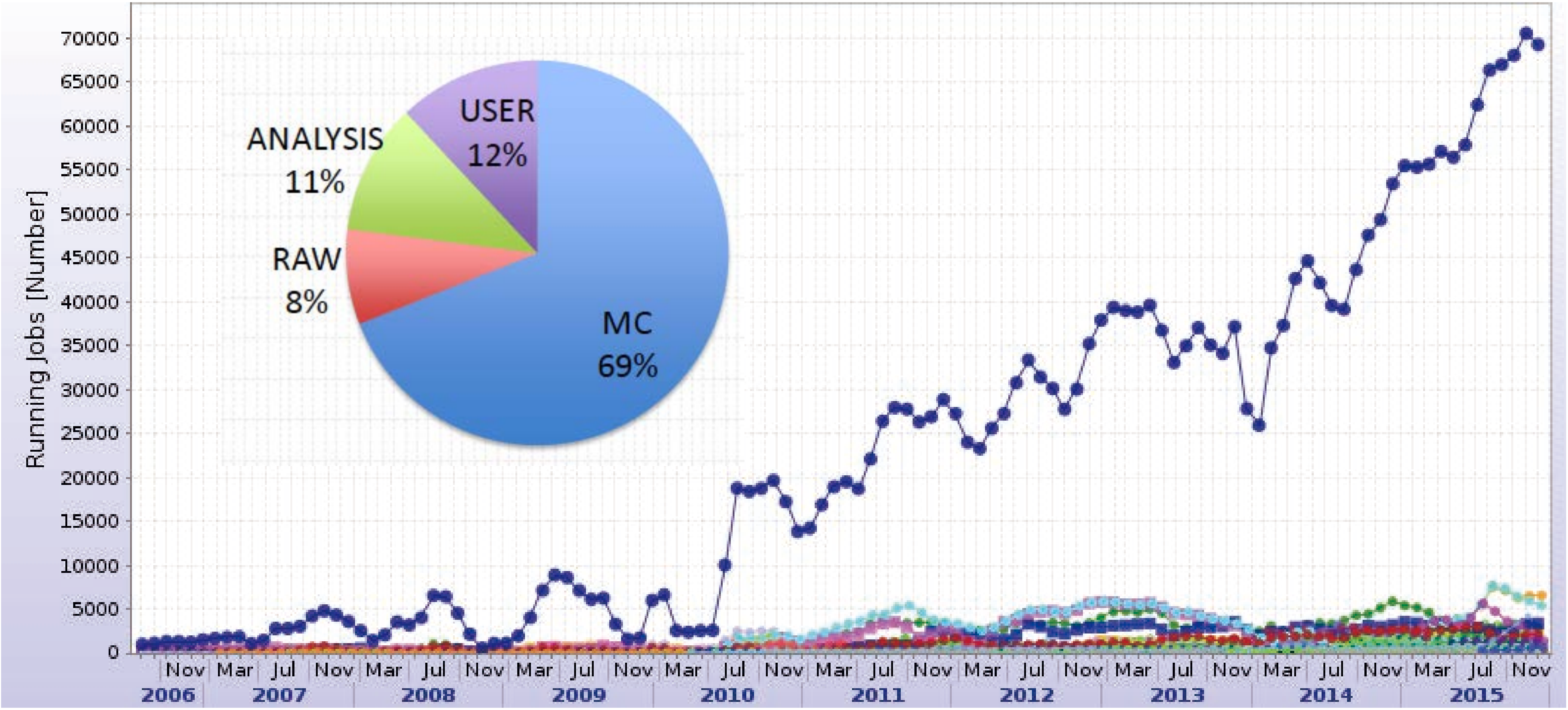
- **Federated computing and storage resources**
  - Users interact with the entire Grid through AliEn
- **Tightly coupled central task queue and file catalogue**
  - Tasks are typically sent to where a copy of the input data is, but one can also read from anywhere in the world

The ALICE grid keeps growing



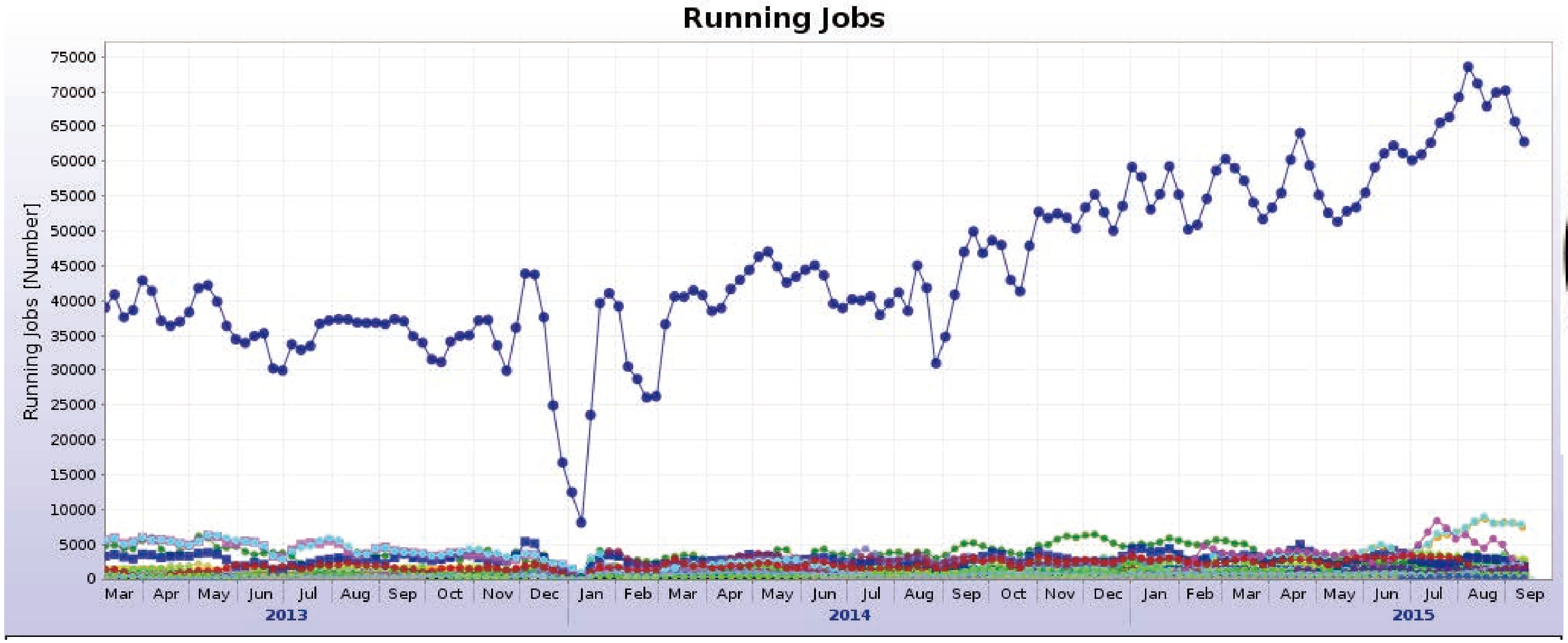
*** 10 Gbs line from KISTI Tier-1 operational**

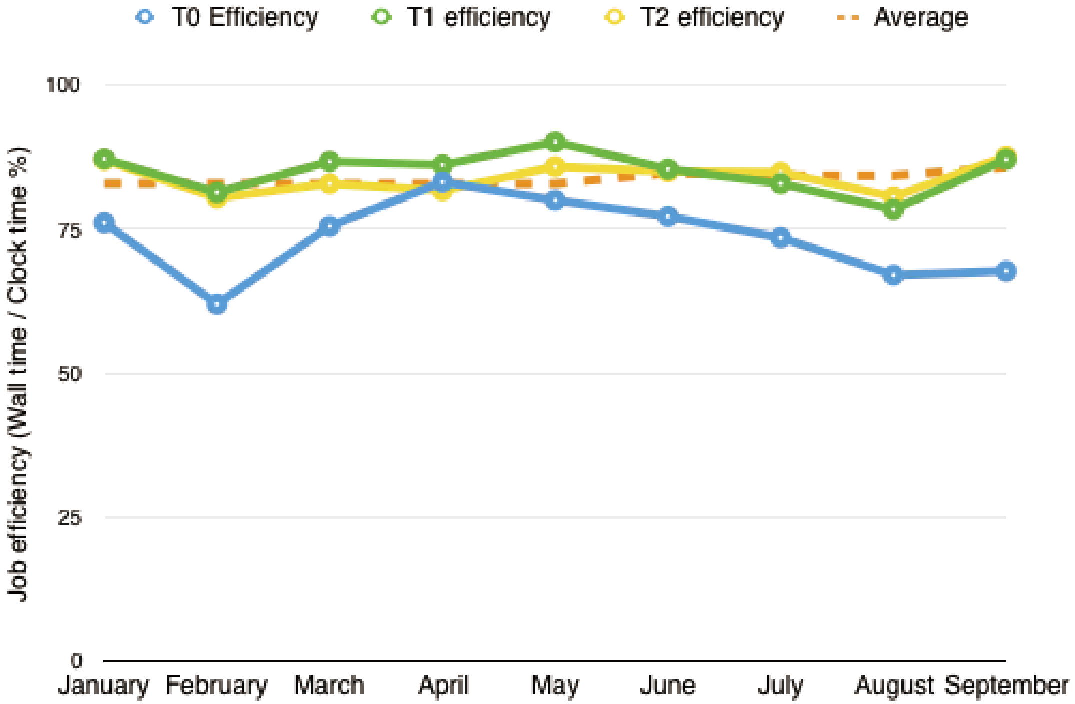
ALICE Grid evolution (10 years history)



Successfully using all available capacity

Grid jobs after the end of Run-1





Maximum:

- **96K parallel jobs**

- Good efficiency in all computing centres (80% on avg.)
- Continue this computing model until the end of Run-2 (LS2: 2019-2020)

ALICE computing in Run-3 and Run-4 (2021-)

LHC schedule

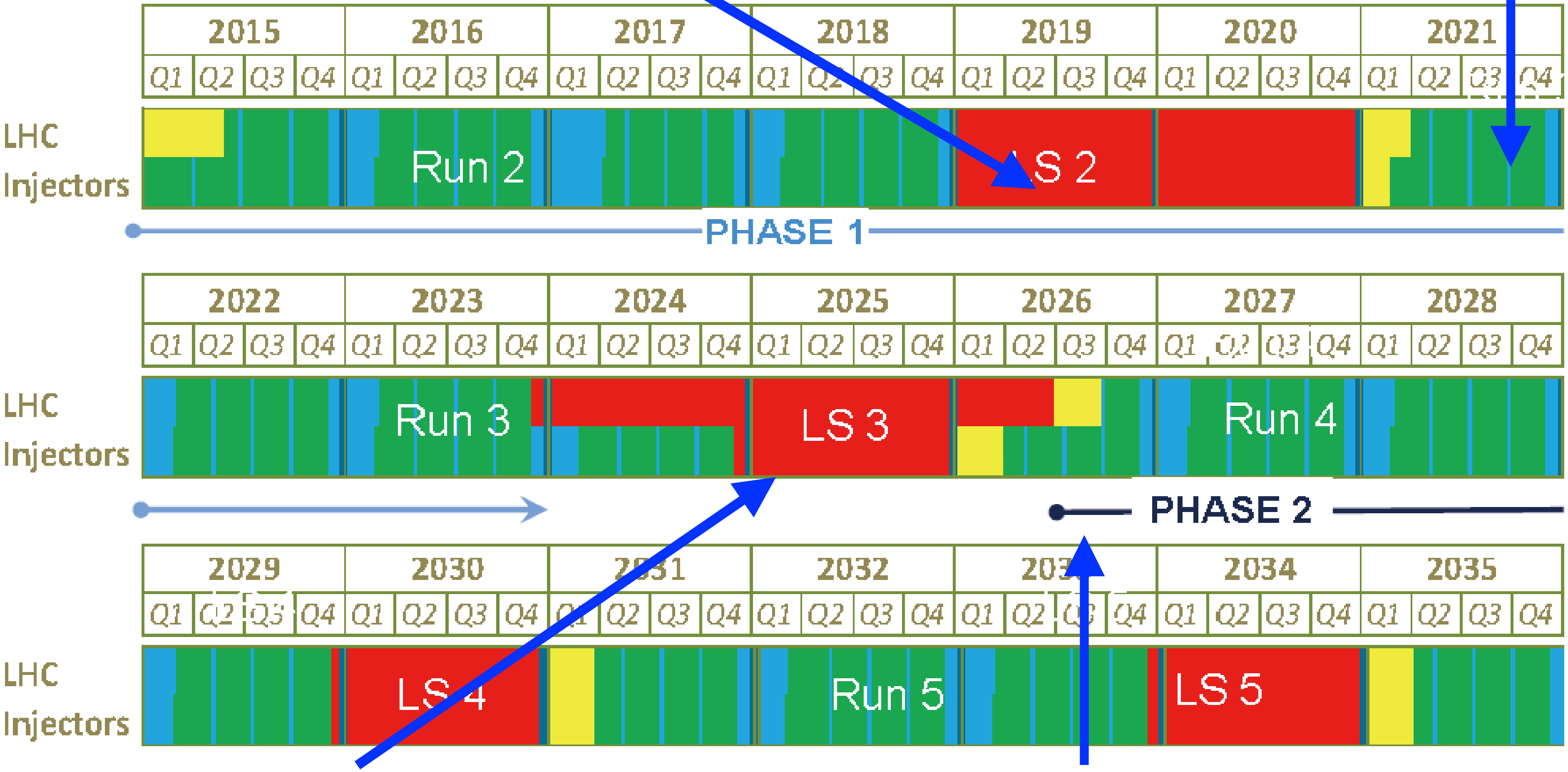
PHASE I Upgrade

ALICE, LHCb major upgrade

ATLAS, CMS ,minor upgrade

Heavy Ion Luminosity

from 10^{27} to 7 x10^{27}



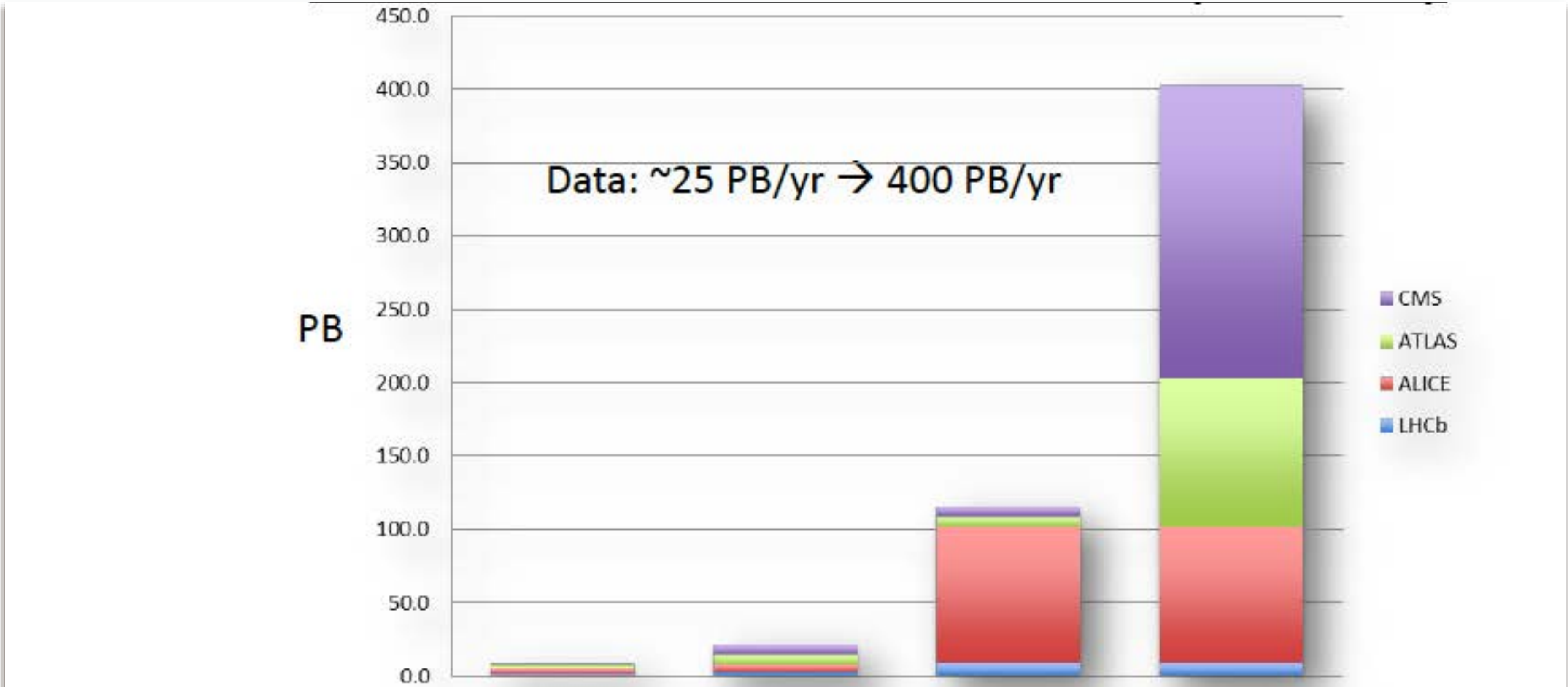
PHASE II Upgrade

ATLAS, CMS major upgrade

HL-LHC, pp luminosity

from 10^{34}(peak) to 5 x10^{34}(levelled)

Resource needed in LH-LHC (WLCG)



Data: ~25 PB/yr → 400 PB/yr

ATLAS/CMS	2010-2012	2015-2018	2020-2022	2025-2035
Integrated luminosity	$25\ fm^{-1}$	$50\ fm^{-1}$	$300\ fm^{-1}$	$3000\ fm^{-1}$
collision energy	7-8TeV	13-14TeV	14TeV	14TeV

ATLAS 400Hz (400MB/s)

ATLAS 1kHz (1-1.5GB/s)

LHCb 20kHz (2GB/s)
ALICE 50kHz (75GB/s)

ATLAS 5-10kHz (10-20GB/s)
CMS 10kHz (40GB/s)

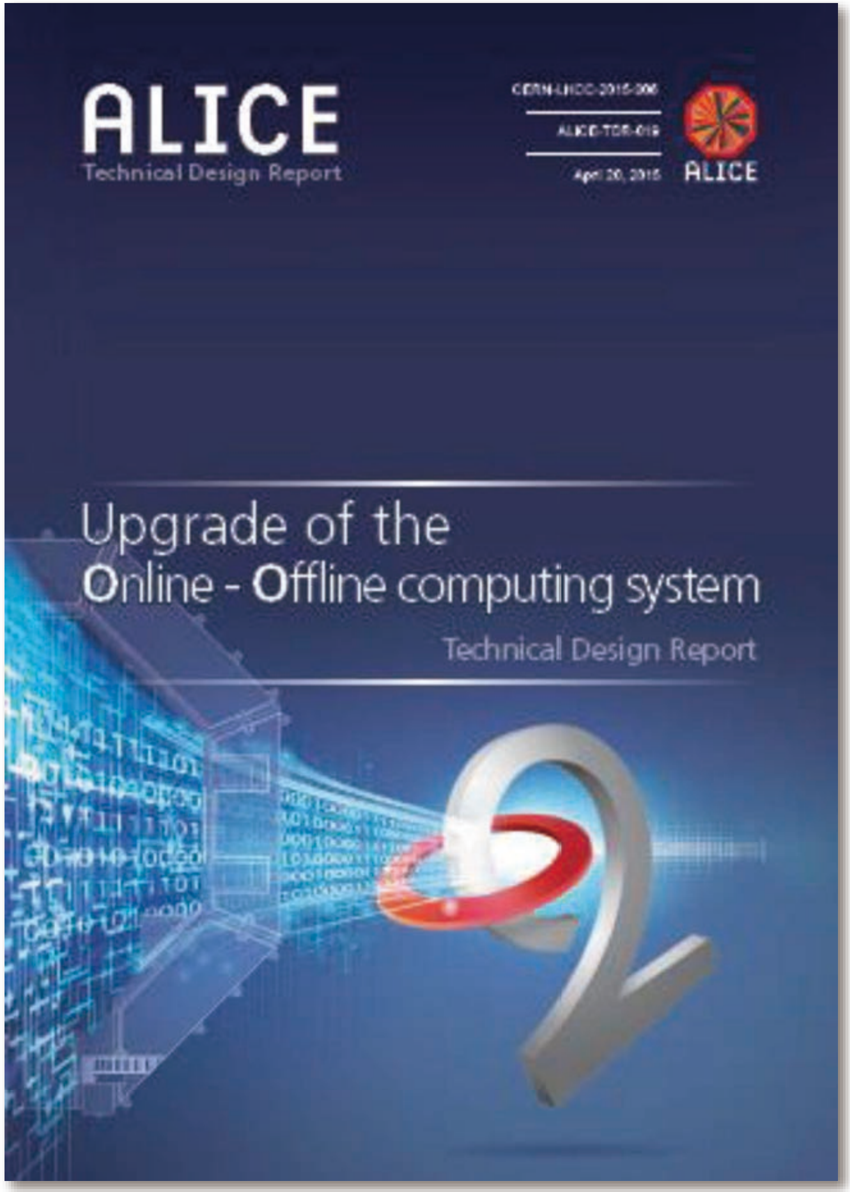
2014/9/20 Tomoaki Nakamura ICEPP, UTokyo/KEK-CRC

The ALICE Online-Offline (O^2) Project

- Handle **>1 T Byte /s** detector input
- Support for continuous readout
- Online reconstruction to reduce data volume
- Common hardware and software system developed by the DAQ, HLT (High Level Trigger), Offline teams

ALICE O^2 project

* O^2 TDR Approved in September 2015 (with some conditions)



ALICE
Technical Design Report

CERN-LHCC-2015-006
ALICE-TDR-019
April 20, 2015

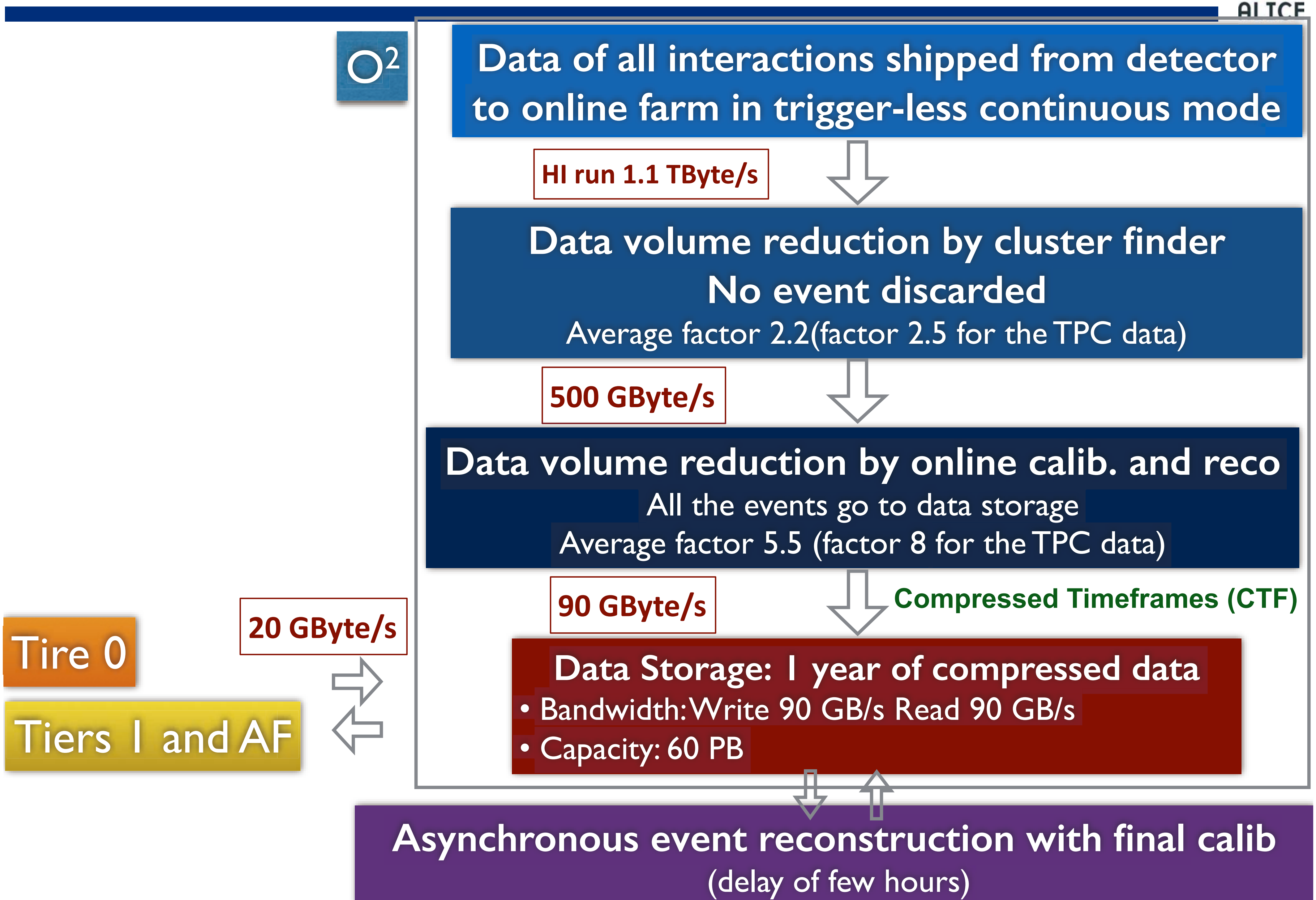
ALICE

Upgrade of the
Online - Offline computing system
Technical Design Report

New O^2 facility

- 463 FPGAs
  - Detector readout and fast cluster finder
- 100K CPU cores
  - To compress 1.1 TB/s data stream 14x
- 5000 GPUs
  - Reconstruction speed-up
  - 3 CPU + 1 GPU == 28 CPU
- 60 PB disk space
  - Buffer space to allow for a more precise calibration
- The current Grid and more in a single computing center
  - Heterogeneous computing capacity
- Identical software should work in both Online and Offline environments

Computing Strategy for Run-3 and 4



O^2

Data of all interactions shipped from detector to online farm in trigger-less continuous mode

HI run 1.1 TByte/s

Data volume reduction by cluster finder

No event discarded

Average factor 2.2(factor 2.5 for the TPC data)

500 GByte/s

Data volume reduction by online calib. and reco

All the events go to data storage

Average factor 5.5 (factor 8 for the TPC data)

90 GByte/s

Compressed Timeframes (CTF)

Data Storage: 1 year of compressed data

- Bandwidth: Write 90 GB/s Read 90 GB/s
- Capacity: 60 PB

20 GByte/s

Tire 0

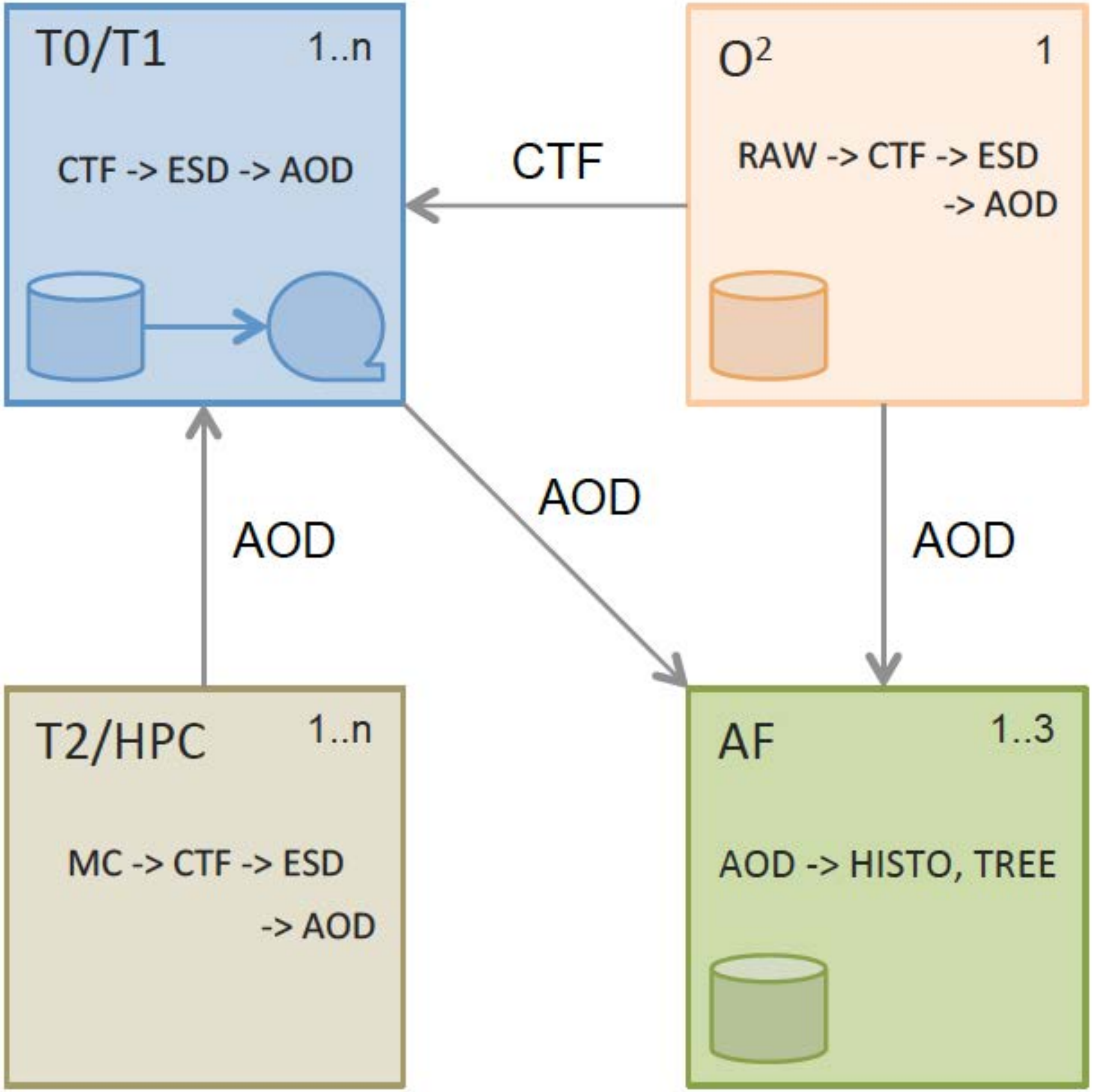
Tiers 1 and AF

Asynchronous event reconstruction with final calib

(delay of few hours)

Roles of Tiers in Run-3

Reconstruction
Calibration
Archiving
Analysis

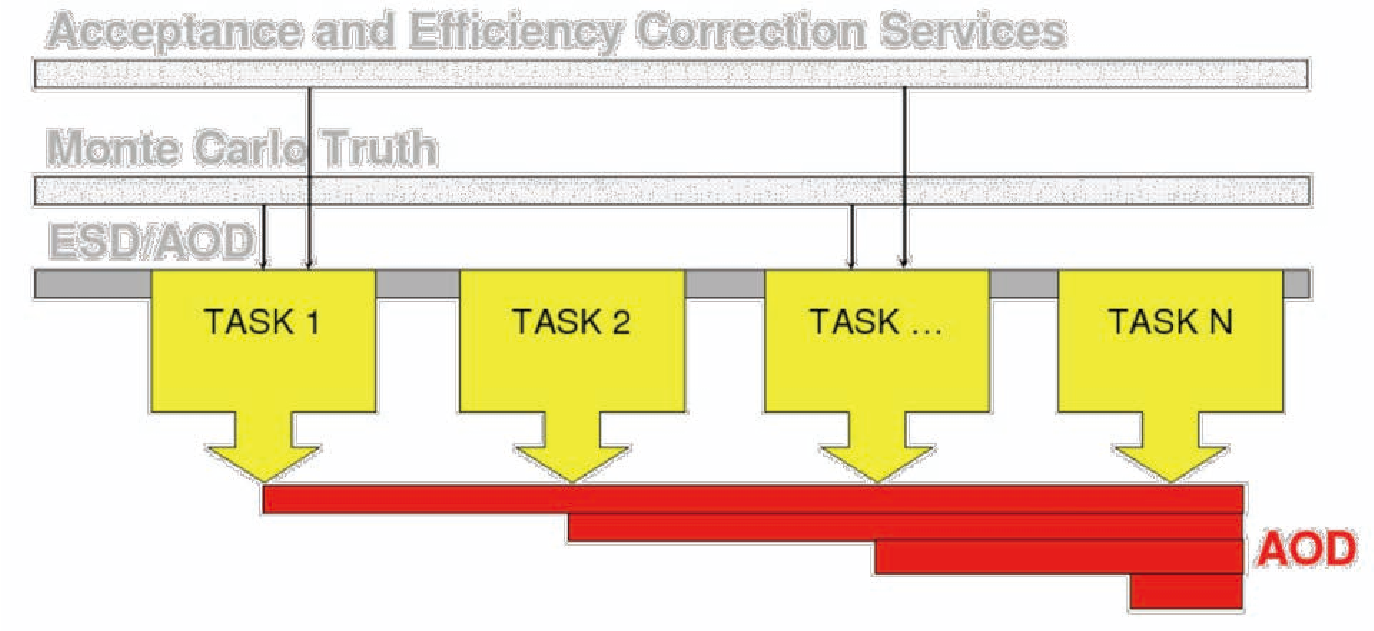


DAQ
Reconstruction
Calibration

Simulation

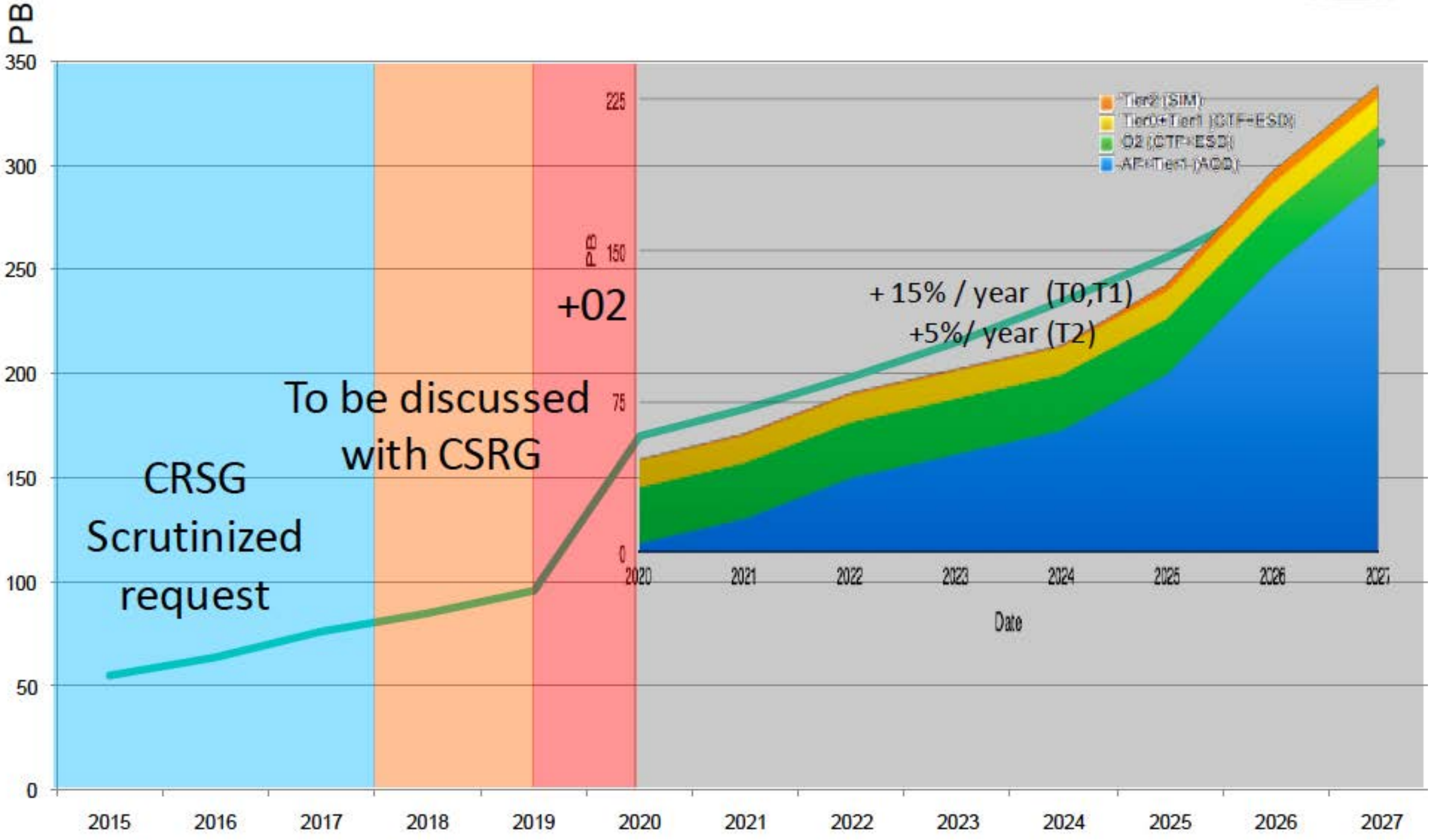
Analysis

Analysis facilities (AF)



- Analysis is still an **I/O bound operation**, even after adopting the analysis trains
- Merging stages could be sped up on well connected, high memory machines
  - Leading to **shorter turn-around time for entire trains**
- Solution is to have **dedicated analysis facility/facilities**
  - Sites optimized for fast processing of large local datasets
  - Run organized analysis on local data, similar as today's Grid
  - Requires 20-30K CPUs and 5-10 PB of very well connected persistent storage space
  - Could be any of the T1s or T2s,
  - But ideally this would be a purpose build facility optimized for such workflow

Storage requirements



CRSG Scrutinized request

To be discussed with CSRG

+02

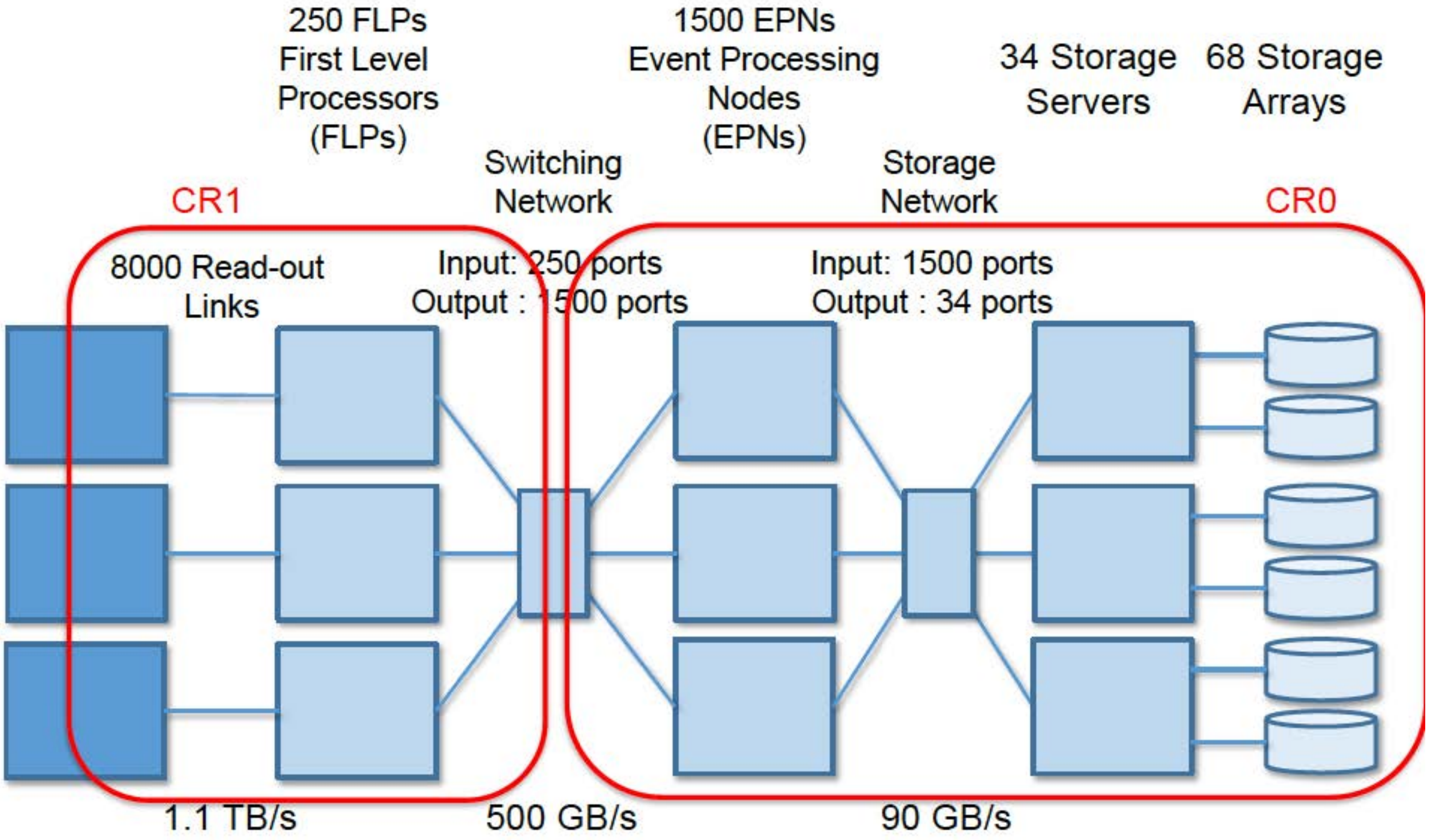
+ 15% / year (T0,T1)
+5%/ year (T2)

* Computing Resources Scrutiny Group (CRSG)

Hardware O^2 facility

Computing rooms

- CR1: reuse existing room (adequate power and cooling for the detector read-out).
- CR0: new room for the computing part of the farm.

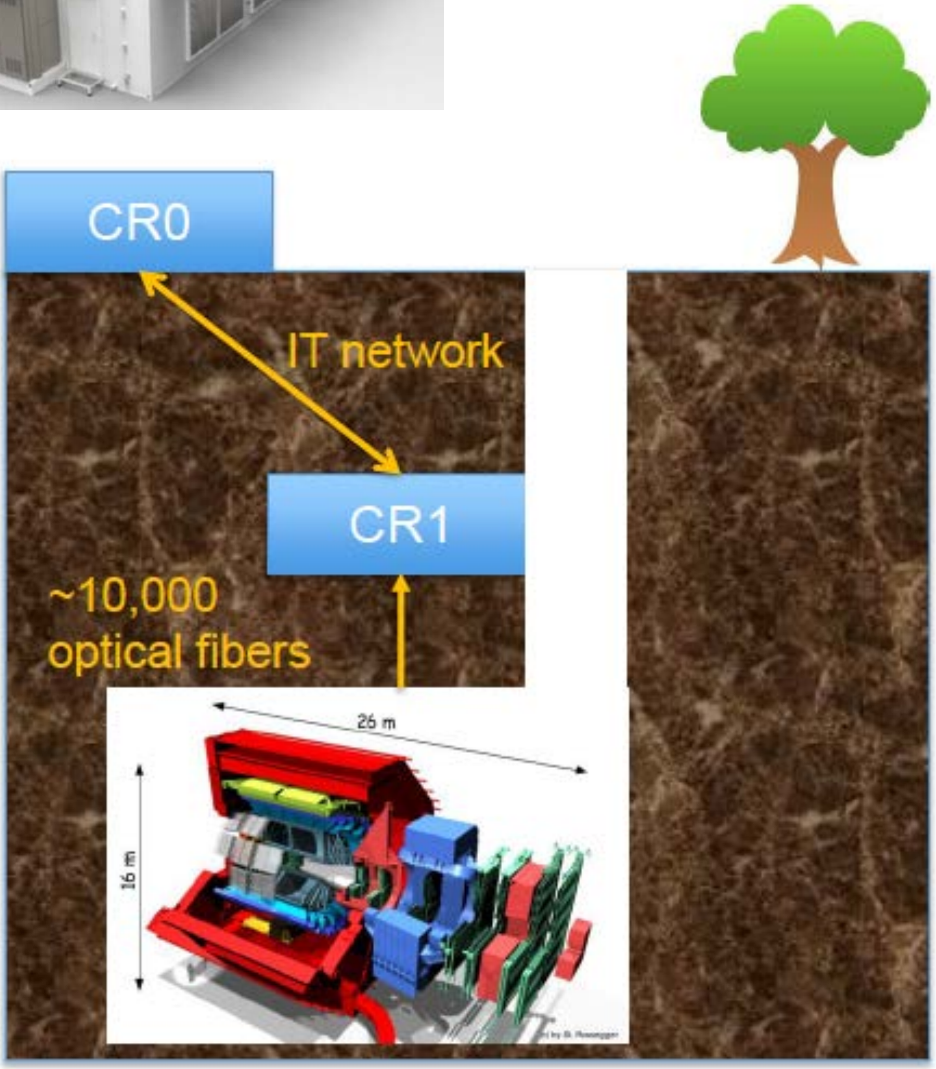


Computing room

CR0: new room and infrastructure needed on the surface

Commercial Container Data Centers

- Used by several Internet giants but at a lower power density



CR1: existing room adequate

T2 site(s) in Japan
and network

ALICE Tier-2 at Hiroshima

- The ALICE T2 site "JP-HIROSHIMA-WLCG" with grid middleware EMI-3 on SL6.5… as stable as possible.
- GRID service; APEL, sBDII, CREAM-CE, XROOTD, DPM-SE, VOBOX… as compact as possible.
- WN resources; 1356 Xeon-cores in total
  Xeon5355(4c@2.6GHz) x 2cpu x 16 boxes
  Xeon5365(4c@3.0GHz) x 2cpu x 20 blades
  Xeon5570(4c@2.9GHz) x 2cpu x 26 blades
  Xeon5670(6c@2.9GHz) x 2cpu x 3 blades
  Xeon5660(6c@2.8GHz) x 2cpu x 42 blades
  E5-2470v2(10c@2.4GHz) x 2cpu x 16 blades
- Storage; 1,056TB disks on 9 servers, but no MS
- Around 3/4 resource deployed to ALICE GRID, and the rest for a local cluster
- Network B/W: 1Gbps on 40Gbps-SINET4 in Japan
- WLCG support by ASGC in Taiwan
- Responsible by Prof. Toru Sugitate
- Operated by TS and K.Tarunaga (M2) under remote technical support by *SOUM* corp., Tokyo.

Tsukuba AlICE T2 status

Members:

- T. Chujo (responsible), S. Kato (technical staff)

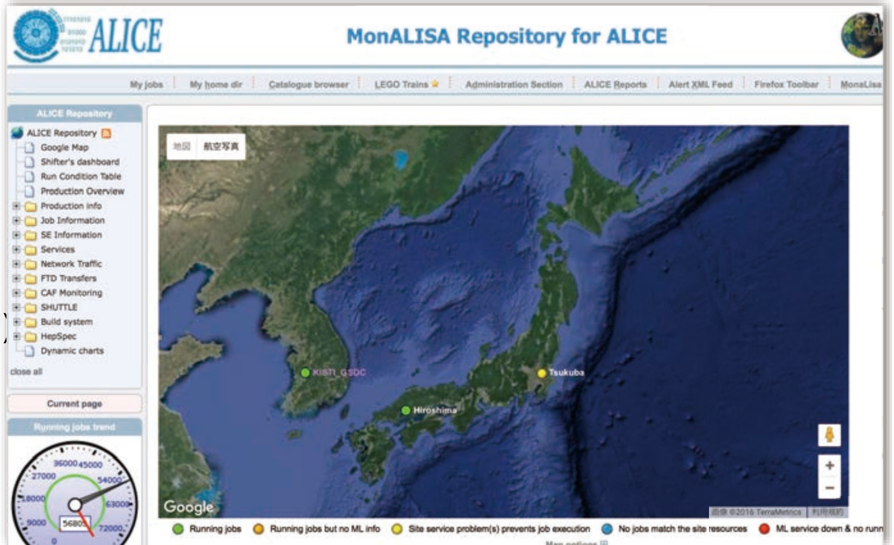
Status:

- Infrastructures, MW (EMI 3.1), and service have been set-up.
- Setting up T2 for the test job submission by ALICE.
- 16 WN's (X5355; 4 cores x 2 cpu,@2.6GHz) in a rack.
- Used IP: HepNet-J.
- Connected to SINET-4/(5) (via HepNet-J).

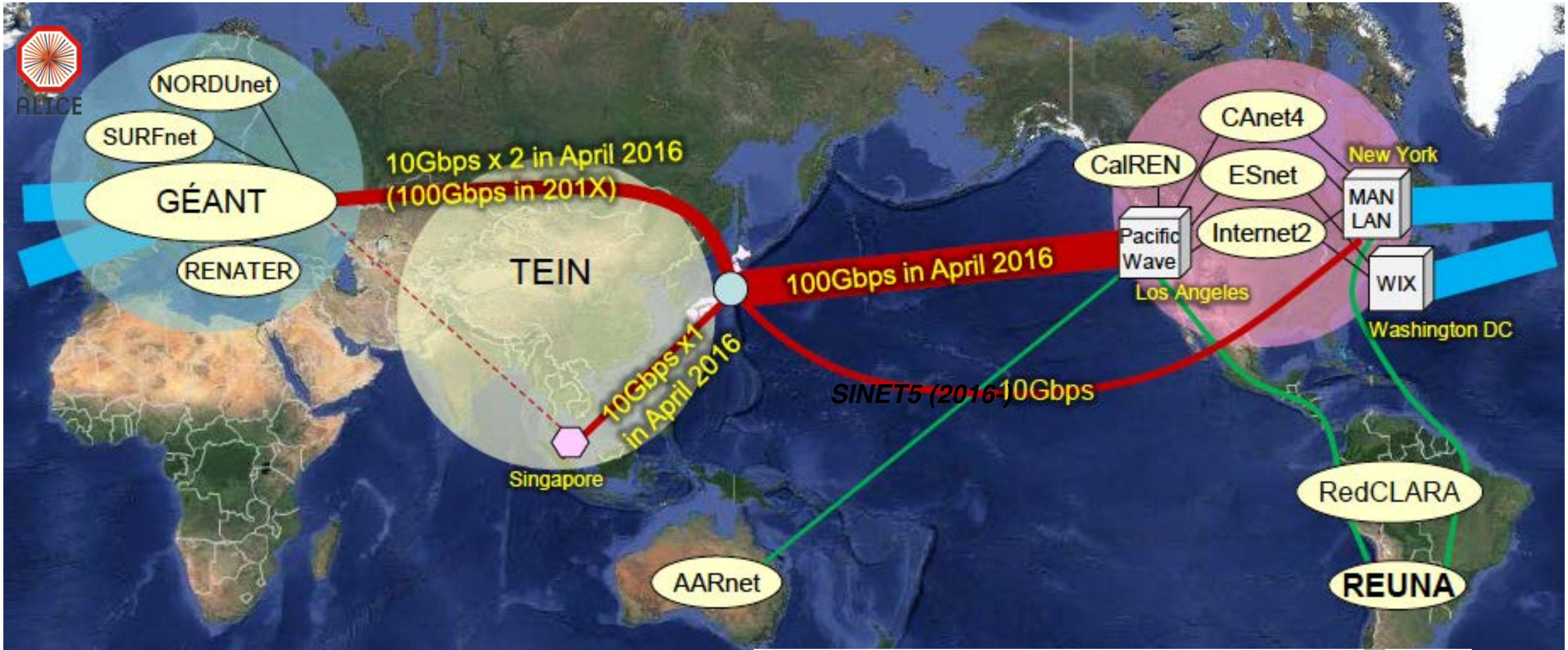
Plan:

- Sign-up WLGC (2016).
- Will use University's IPs for head nodes for the future connection to WLCG (and LHCONE), with the support by U.Tsukuba info. center and KEK.

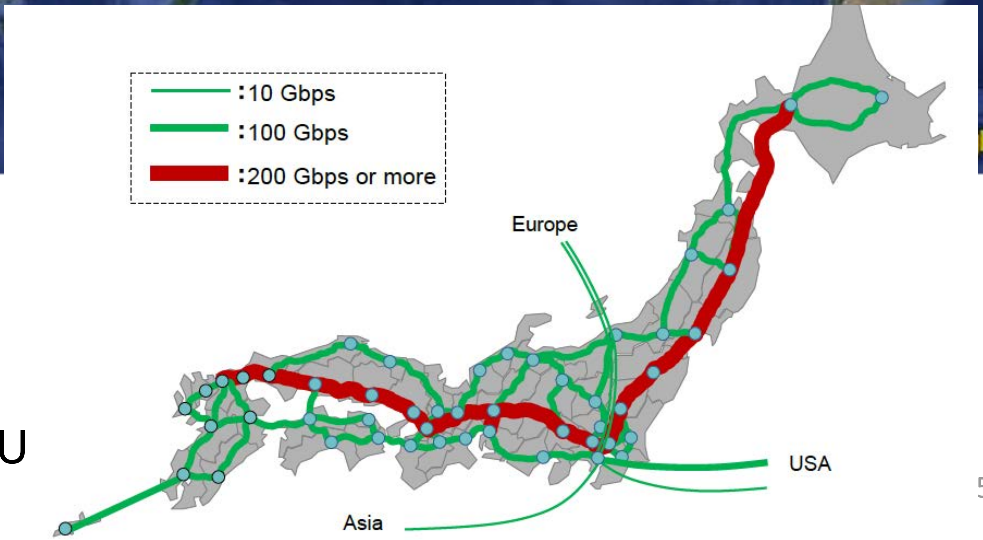
←16 WNs (provided by Hiroshima U)
as a prototype of T2 in Tsukuba
(marked by yellow labels)
2015, May



SINET5 (2016-)



Slide by M. Nakamura (NII) @ AFAD2015



SINET 5 (April 2016-), by NII

Domestic: 200Gbps backborn,

International: 100 Gbps direct link JPN ⇄US & EU

Summary

- ALICE computing usage in Run-1 & 2.
- ALICE Run-3 &4 (2021-)
  - Continuous trigger-less readout at 50 kHz in Pb-Pb collisions.
  - 1 TB/s raw data from detector, need a significant data reduction down to 90 GB/s to storage → O^2 project
  - Computing model (O^2, T0, T1/AF, T2; re-defined the roles)
  - Status of O^2
- Japanese T2 site(s) status & networking (SINET-5).
  - Japanese involvement for O^2.