Introduction to GSDC

(Data Center for Data-intensive Research)

Global **S**cience experimental
Data hub **C**enter

January 29, 2018

Seo-Young Noh

KiSTi
www.kisti.re.kr

Contents

Data & Infra-driven R&D Era

Research Paradigm Shift

Data & Infrastructure are Key in Scientific Discovery

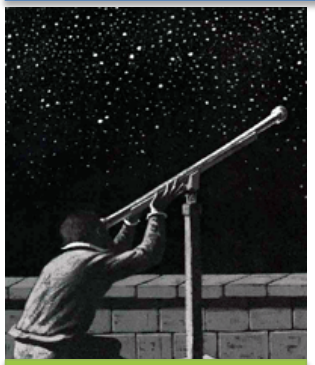
Describing natural phenomena based on Observation

Modeling and Theory

Computing Simulation

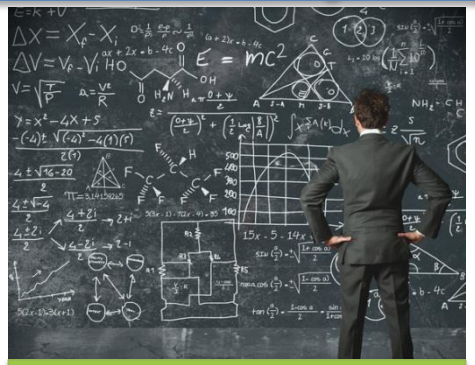
Data Analysis of tremendous data produced from large experimental facilities

Research Paradigm Shift to Data Intensive Scientific Discovery



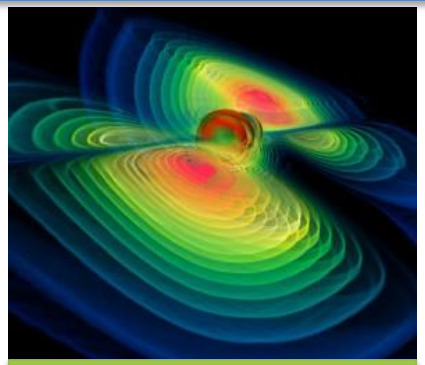
1st Generation: Observation

Galileo's telescope



2nd Generation: Theory

Higgs Theory



3rd Generation: Simulation

Black Hole Simulation

4th Generation: Data

CERN's CMS and ATLAS experiments → Higgs discovery

More chance to do research with advanced equipment, higher chance to get Nobel prize

87% of Nobel prizes have been given to researchers who produced outstanding scientific discoveries using advanced experimental equipment since 1914.

Source: The Fourth Paradigm

Trust in Data … data is leading science

- **CERN** noticed a signal like a new particle in CMS & ATLAS experiments in December 2015.
- The **750 GeV diphoton excess** in particle physics was an anomaly in data collected at the Large Hadron Collider(LHC) in 2015, which could have been an indication of a new particle.
- However, the anomaly was absent in data collected in 2016, suggesting that the diphoton excess was a statistical fluctuation.
- In the interval between the December 2015 and August 2016 results, the anomaly generated considerable interest in the scientific community, including about **500 theoretical studies**.

We are in data-driven science era!!!

Our trust is in data

Lots of theory papers submitted to PRL

PRL 116, 150001 (2016) PHYSICAL REVIEW LETTERS week ending 15 APRIL 2016

Editorial: Theorists React to the CERN 750 GeV Diphoton Data

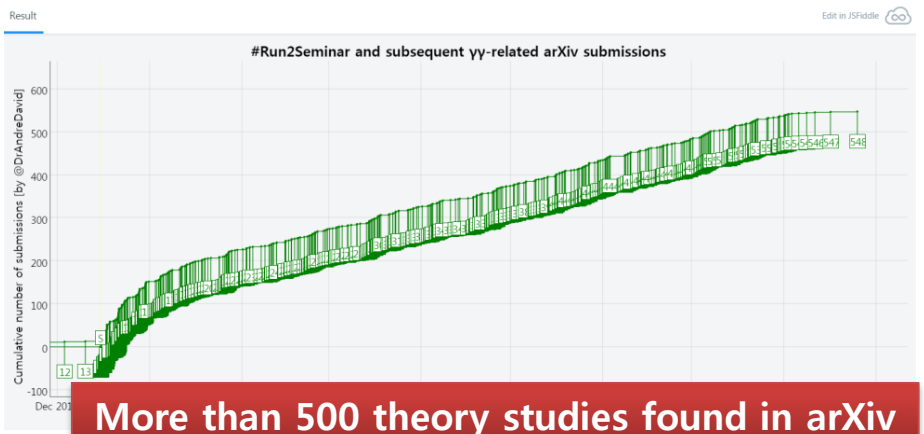
Last December, the ATLAS and CMS Collaborations at the Large Hadron Collider reported preliminary data with a small excess of diphoton events at an invariant mass of about 750 GeV [1,2], which, if verified, would require unexpected new elementary particles. The collaborations have recently reanalyzed their data [3,4], and the signal has become slightly stronger. Though the results are extremely intriguing, more data are required to establish if the excess is real, or a statistical fluctuation.

Over 250 theory papers have appeared following the December announcement, and a number of them were submitted to us. We found it appropriate to publish a small sample of them. To maximize the coherence and fairness of our choices, we obtained informal advice from several experts.

Four such Letters appear in this issue [5–8]. Others may follow, but we think that this set gives readers a sense of the kind of new physics that would be required to explain the data, if confirmed.

Robert Garisto
Editor

Published 12 April 2016
DOI: 10.1103/PhysRevLett.116.150001



More than 500 theory studies found in arXiv

“Open Science”…hot keyword among Policy Makers

- OECD produced [the first Open Science report](), mainly focusing on Open Access, Open Collaboration and Open Data (2015)

- Several expert groups in GSF have been formed to build advisory policy for Open Science: [Research Infrastructure, Data Infrastructure for Open Science]()

Open Data, Open Access and Open Collaboration through Information and Communication Technology

Acknowledgment on the importance of openness

Open science is ***more than open access to publications or data***; it includes many aspects and stages of research processes. […]

… is ***a broader concept*** that includes
- interoperability of scientific infrastructure
- open and shared research methodologies

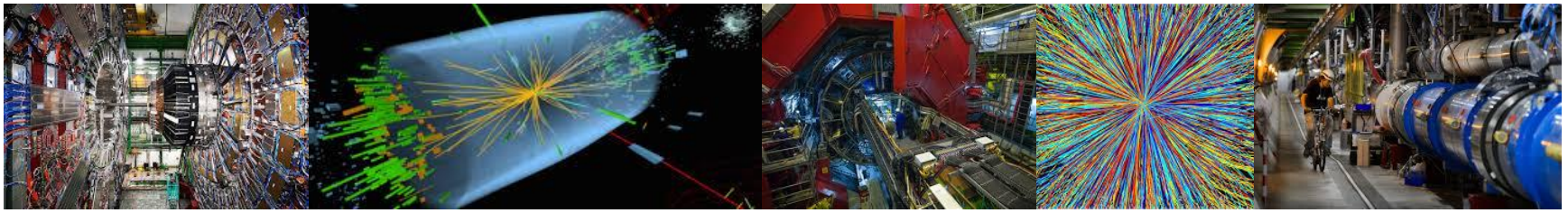
➲ Provides cost-effective access to digital research data ***from public funding***

➲ Enhances utilizations of research data ***to scientific communities as well as societies*** including corporate sectors

Data Infrastructure…that is what we need

Science relies on data, requiring infrastructure for data.

Data is getting more important and growing fast.



Data Infrastructure is the one of key factors for successful science and tackling big problems of humankind.

KISTI has been in preparation for big data research era. Our mission is gradually expanding to national role for data intensive research.

Data Infrastructure: KISTI-GSDC

KISTI…providing powerful ICT infra. service

Unique organization in Korea, our mission is to support various R&D activities through ICT infra. service

Supercomputer…not for specific, but for open to various R&D

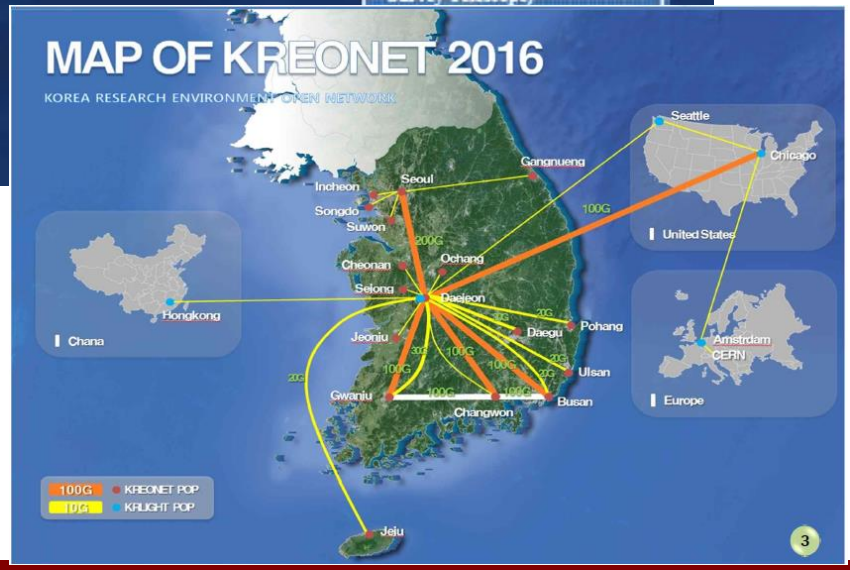
1st

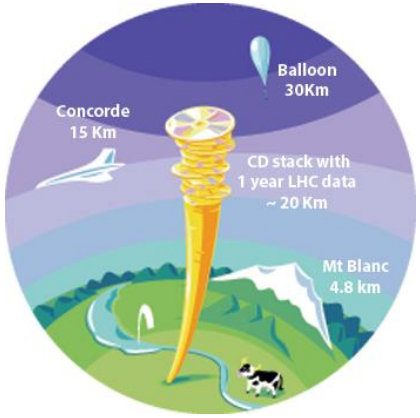
Cray-2S
Nov 1988~Oct 1993
KISTI-1

1988
2GF

Advanced KREONET Center…fast & secure data transmission

Providing domestic researchers with a [constraint free collaborative research environment](#) through KREONET(locally) and GLORIAD(globally)

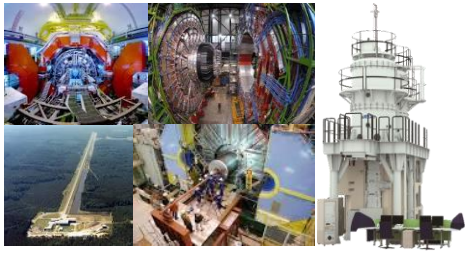




Large-scale Scientific Data:
20Km CD stack with data produced per year in CERN

Global Science experimental Data hub Center

Collaboration with global laboratories



Data from
large and high-valued research equipment

- (Global)
  Asia representative Data Hub
- (Domestic)
  Scientific data management and analysis platform service

History

Collaborations

1. Particle Physics
2. Detector Construction and Exploitation
3. **LHC Computing Grid**
4. CERN's training programs and schools

Korea-CERN Agreement — 2006.10

Korea-CERN (LHC) Protocol — 2007.07

Enhancement of Grid Computing Support for large-scale research facility (Science & Technology Master Plan 577) — 2008.08

Strategy Study on Computing Infrastructure for experimental Data sharing — 2009.12

2010.07 — **Launched Global Science experimental Data hub Center@KISTI**

2014.04 — **WLCG Tier-1 Approved (11th Nation)**

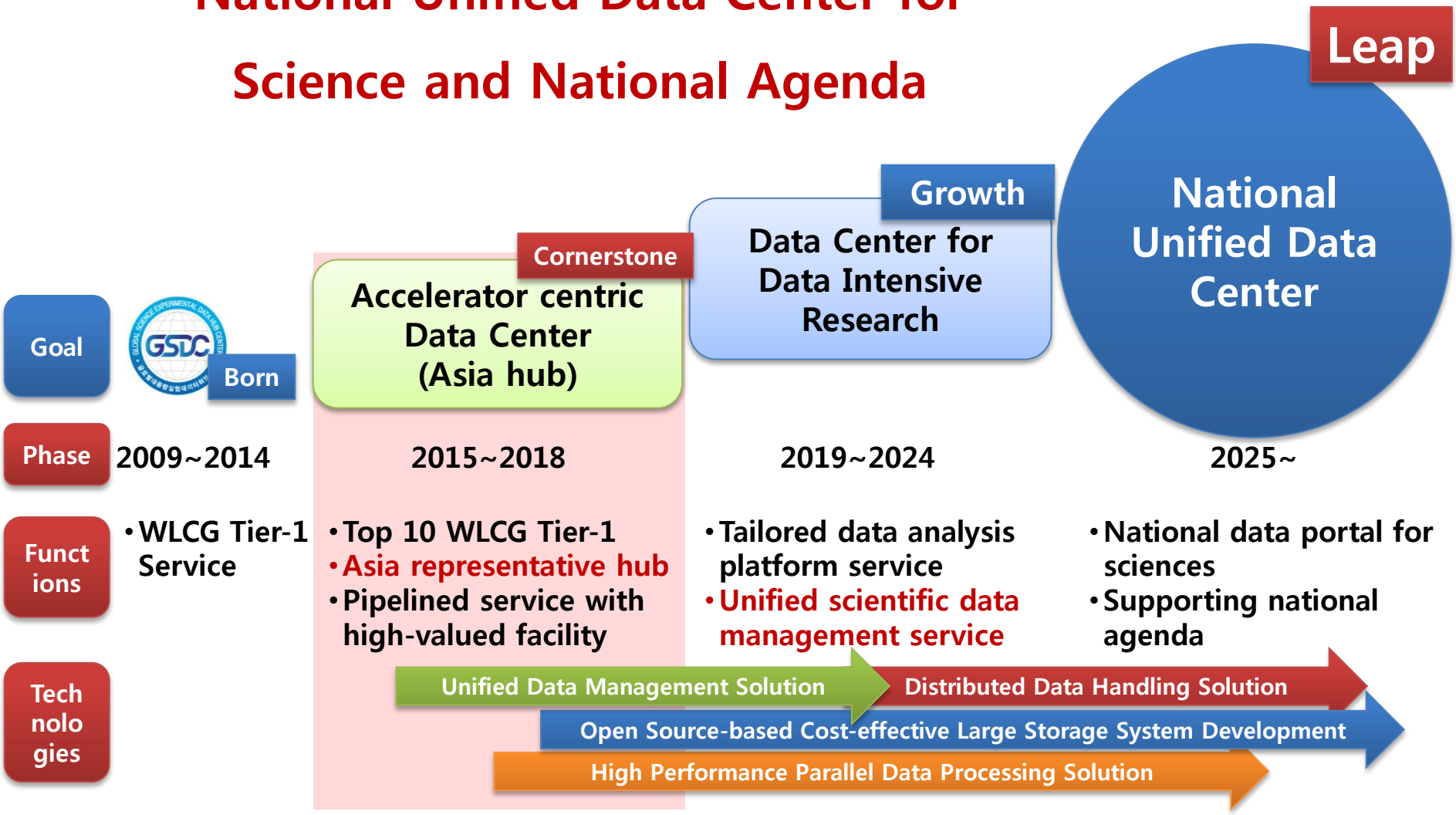
2015.5 — **KISTI-CERN 10Gbps Network Established**

2016 — **Top Quality of Service (~11th ranked)**

National Data Center for R&D

Goal and Roadmap

National Unified Data Center for Science and National Agenda



Strategy

Promotion of Data Intensive Research

GSDC Promoting Science

R&D Partner for World-class Scientific Achievement

Role of GSDC

National Unified Data Center for Science and National Agenda

Service · Development

Technology

- Data Mgmt.
- Efficient Storage
- Infra Unification

Tailored Service for Research

Edu

KiAF (ALICE) | LDG (LIGO) | CMS T3 | RAON

Med/Genome | Climate | TEM | Brain

NA

Global Data Hub

WLCG Tier-1 Asia Hub

Nat'l Agenda

Collaboration

Infrastructure

Open Science Data Platform

Data

Basic Science | Medical · Bio | High-valued | Safety· Infra

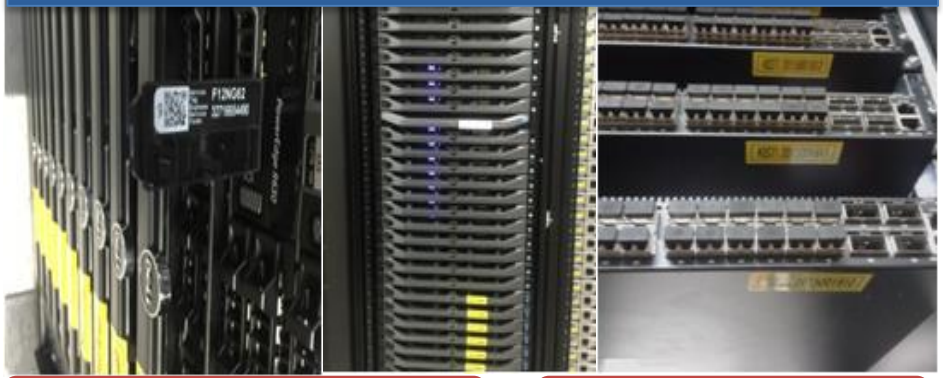
World-class CERN Tier-1 Center

WLCG Tier-1 officially certified in 2014 (Applied in 2012)

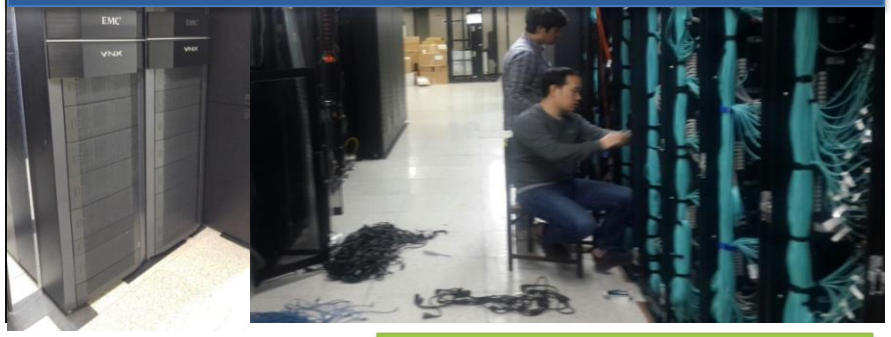
Worldwide LHC Computing Grid

Service-level ranked top 11th among 164 WLCG Tier centers*

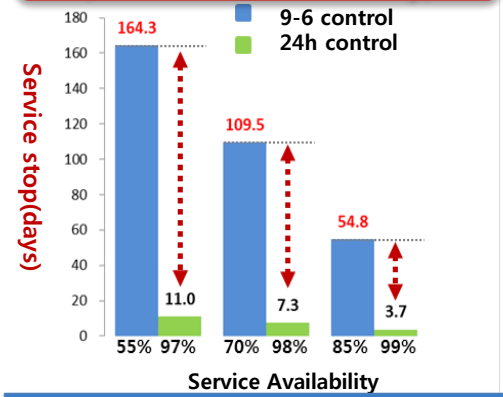
Best equipment procured every year



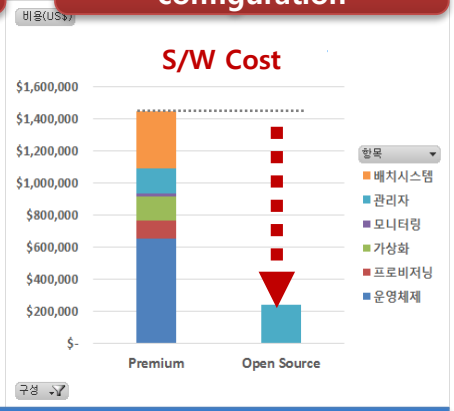
Interlocking with existing systems done by 100% KISTI experts



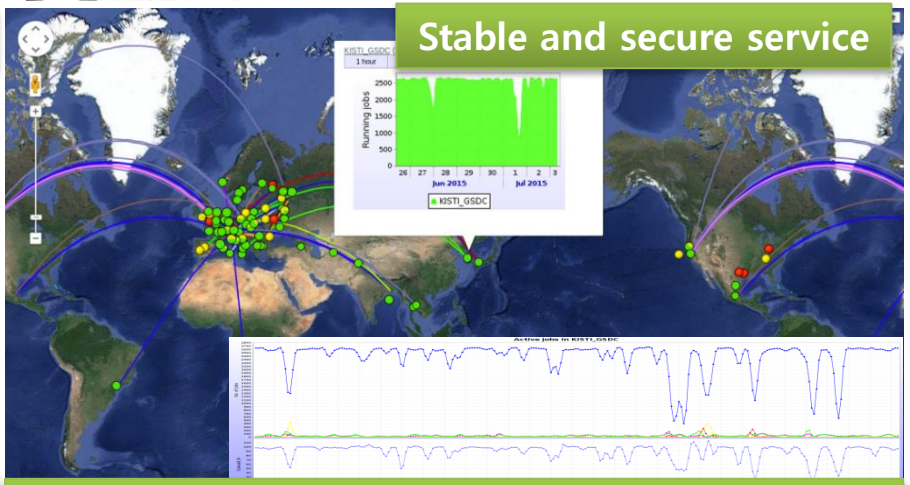
24x365 service quality control (at least 97% availability)



Cost-effective service configuration



Stable and secure service

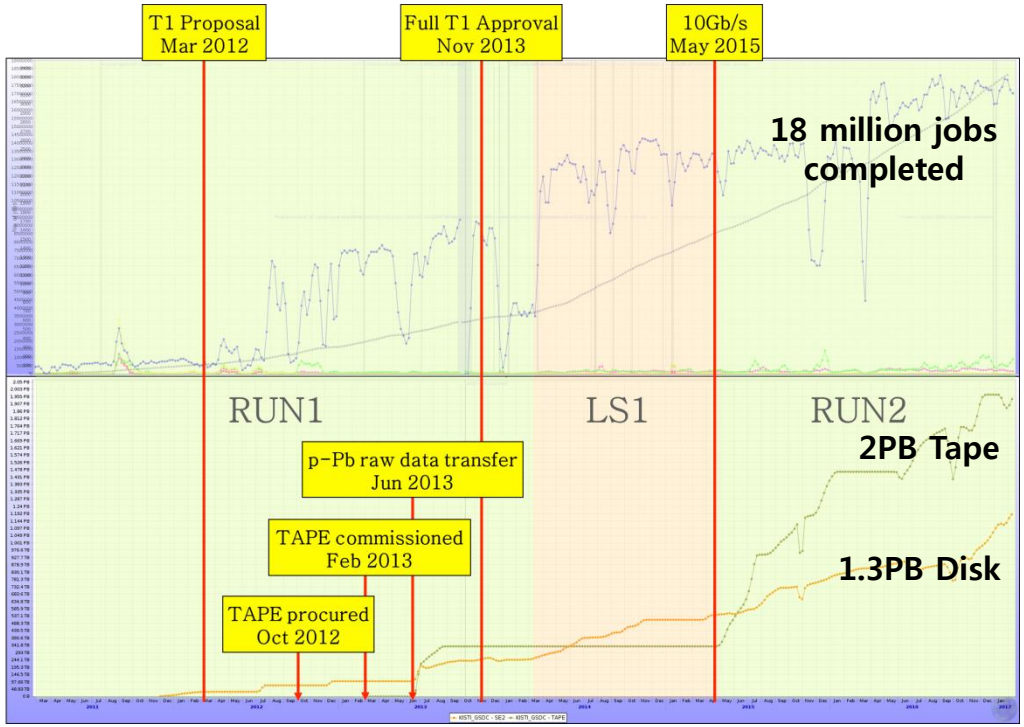


100% open source used, requiring expertise and advanced skills [NOT FREE]

4.5 million data analysis completed per year

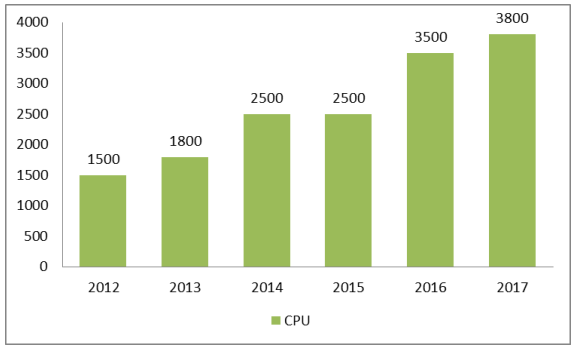
* Service quality measurement based on annual WLCG report

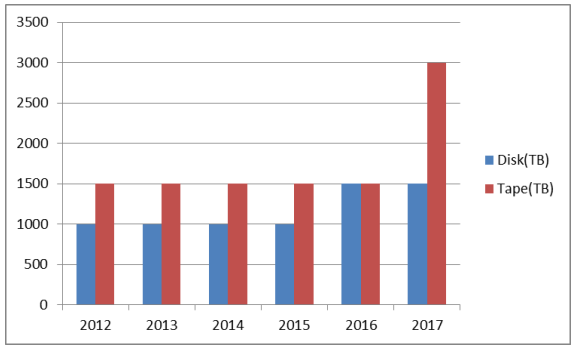
WLCG Tier-1 (ALICE) Resources & Usage

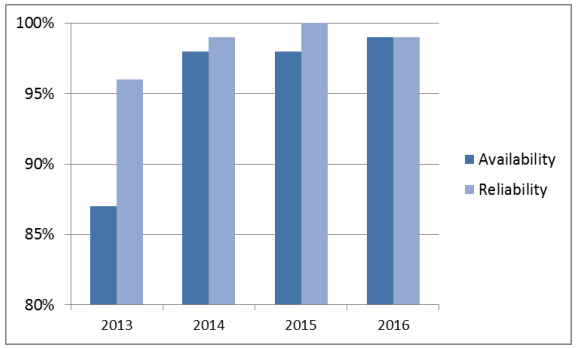


- KISTI WLCG Tier-1 is the latest officially approved Tier-1
- the unique Tier-1 joined after LHC operation (the others joined 5 years before)
- However, it provides very stable service in short time and achieved fast catch-up

More than 1.5PB of raw data transferred from ALICE during RUN2







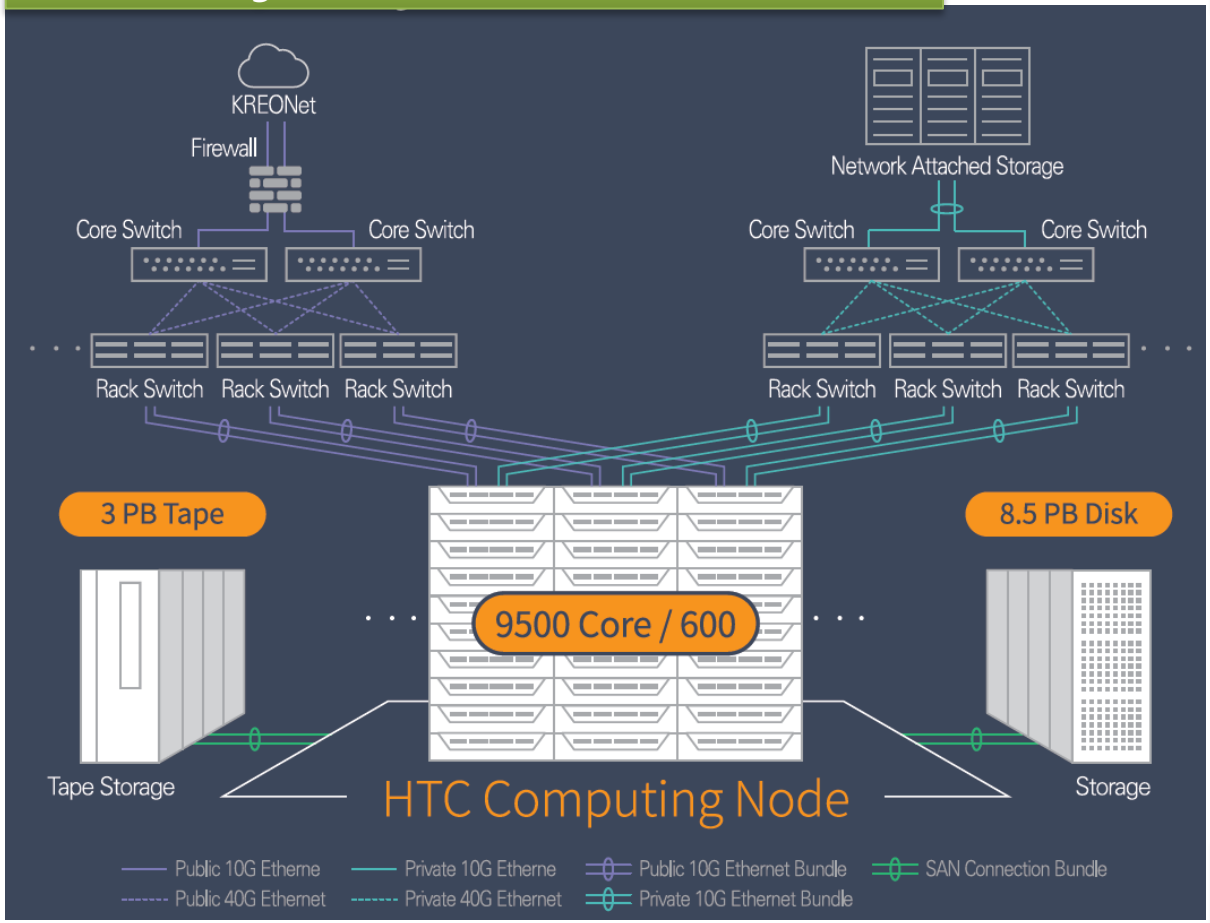
Keeping smooth increment of its capacity in computing and storages as pledged

Keeping top quality of service

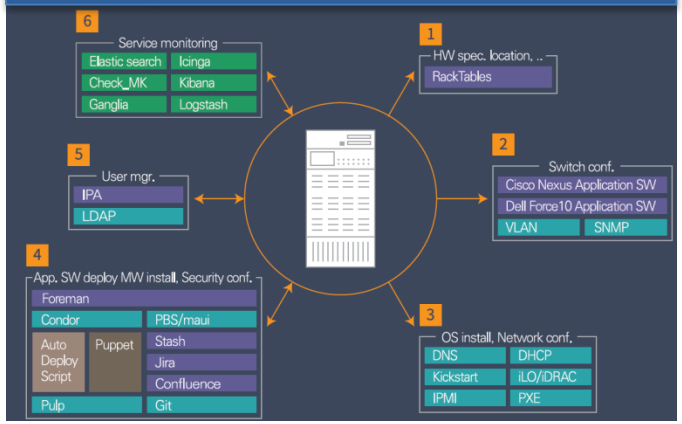
Infrastructure @ KISTI-GSDC … keep growing

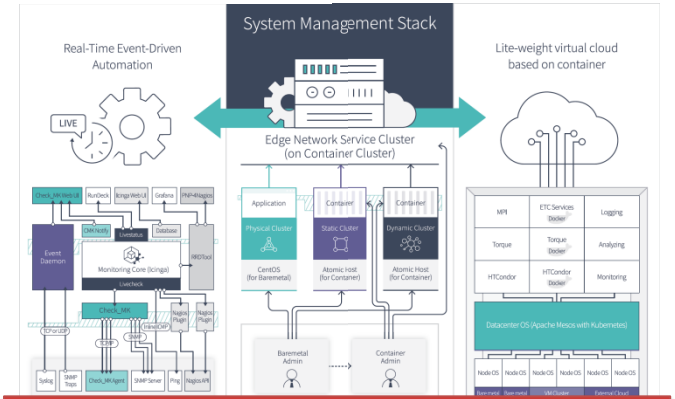
Major vendors' competition place due to every year procurement, requiring big efforts. It is impossible without expertise.

25 storage racks with 5 different vendors



Six steps to enable basic functions





Automatic System Management Stack for Service

Expanding to other Scientific Domains

Experience on WLCG Tier-1 operation and service has given many benefits to expand its service availability to other scientific domains in Korea

and it is still expanding to many other research areas.

Service for additional domestic experiments is under preparation.

Asia Tier Center Forum … central place for Asian community

Steering wheel to solve common issues and troubles faced by Asia Tier centers

7 Tier Centers, ESnet, TEIN, KREONET, CERN


Registration


Net Tuning


Site Reports


Opening


GEANT


GLORIAD-KR


Discussion

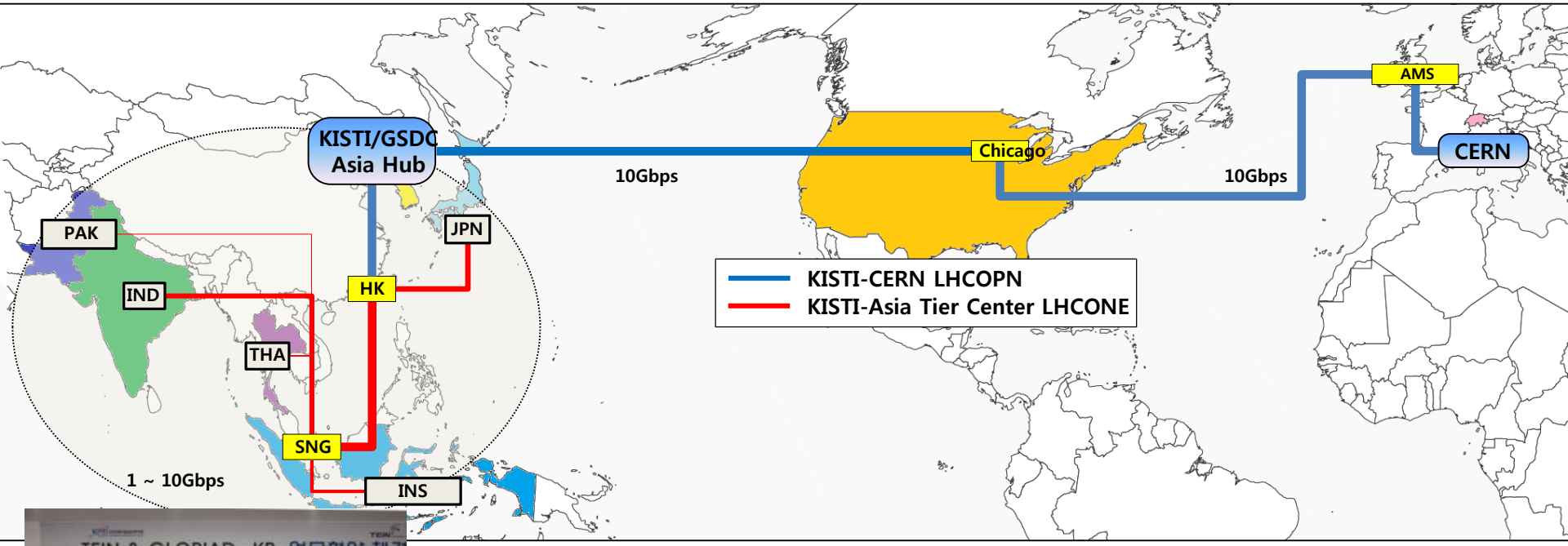
Plan of 2017

- 3rd ATCF in Korea
- Forum Regulation
- Secretariat Office at KISTI

http://www.atcforum.org/

Network Connection

10Gbps direct link between KISTI and CERN through Chicago (sharing 100Gbps GLORIAD as a backup)

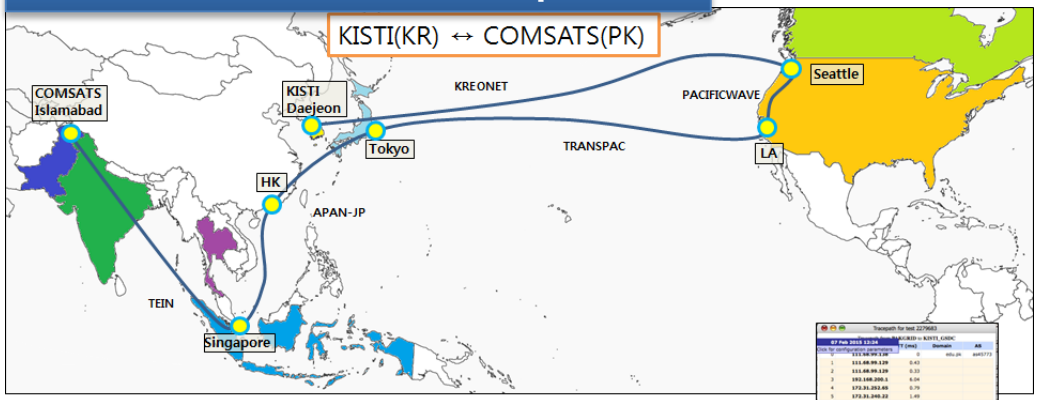


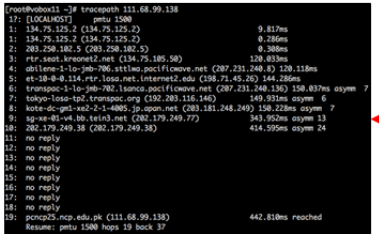
MOU Signup
(June, 2016)

Routing information sharing between TEIN-GLORIAD, improving network connectivity among Asia Tier centers

Network Connectivity Improvement in Asian Tiers

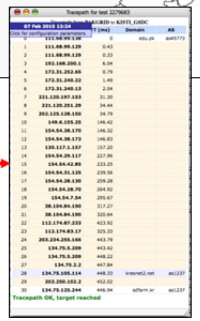
Before TEIN-GLORIAD interoperation



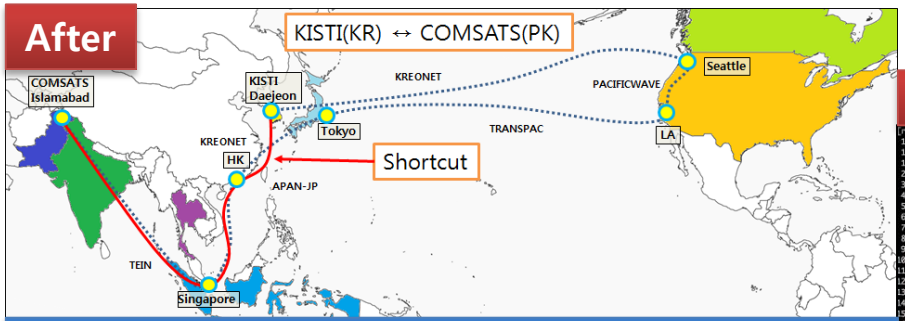


Tracepath

30 hops



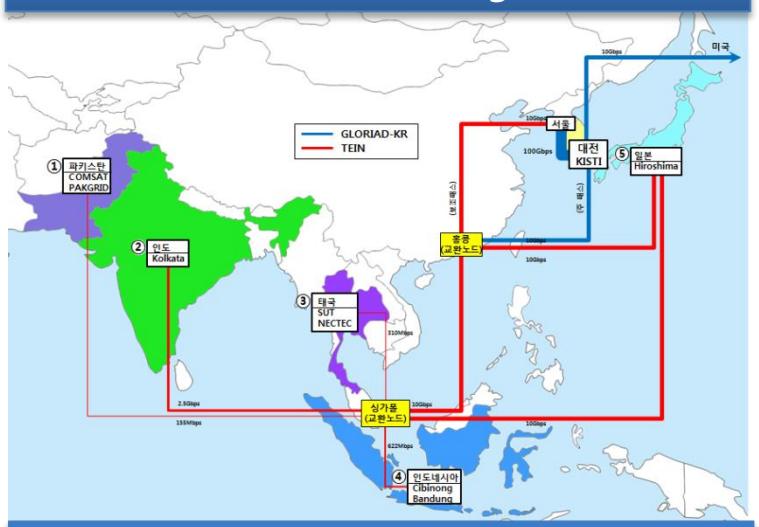
After



Not need to travel Pacific Ocean twice, data is delivered to Asia Tier-2s directly

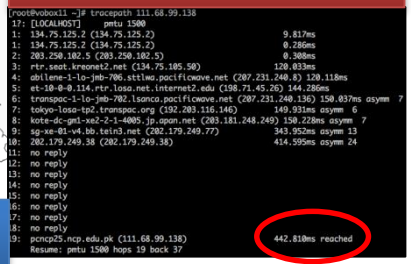
Not an easy task, but done in time with support from Advanced Network Center@KISTI

Interconnection & exchange route info.

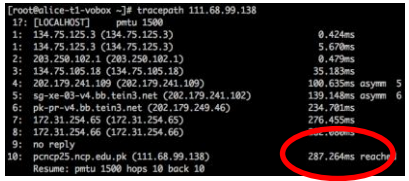


- All Tier-2 centers are connected to TEIN
- GLORIAD line reaches Hong Kong
- Possible to exchange routing information

Connectivity Comparison



442.810ms



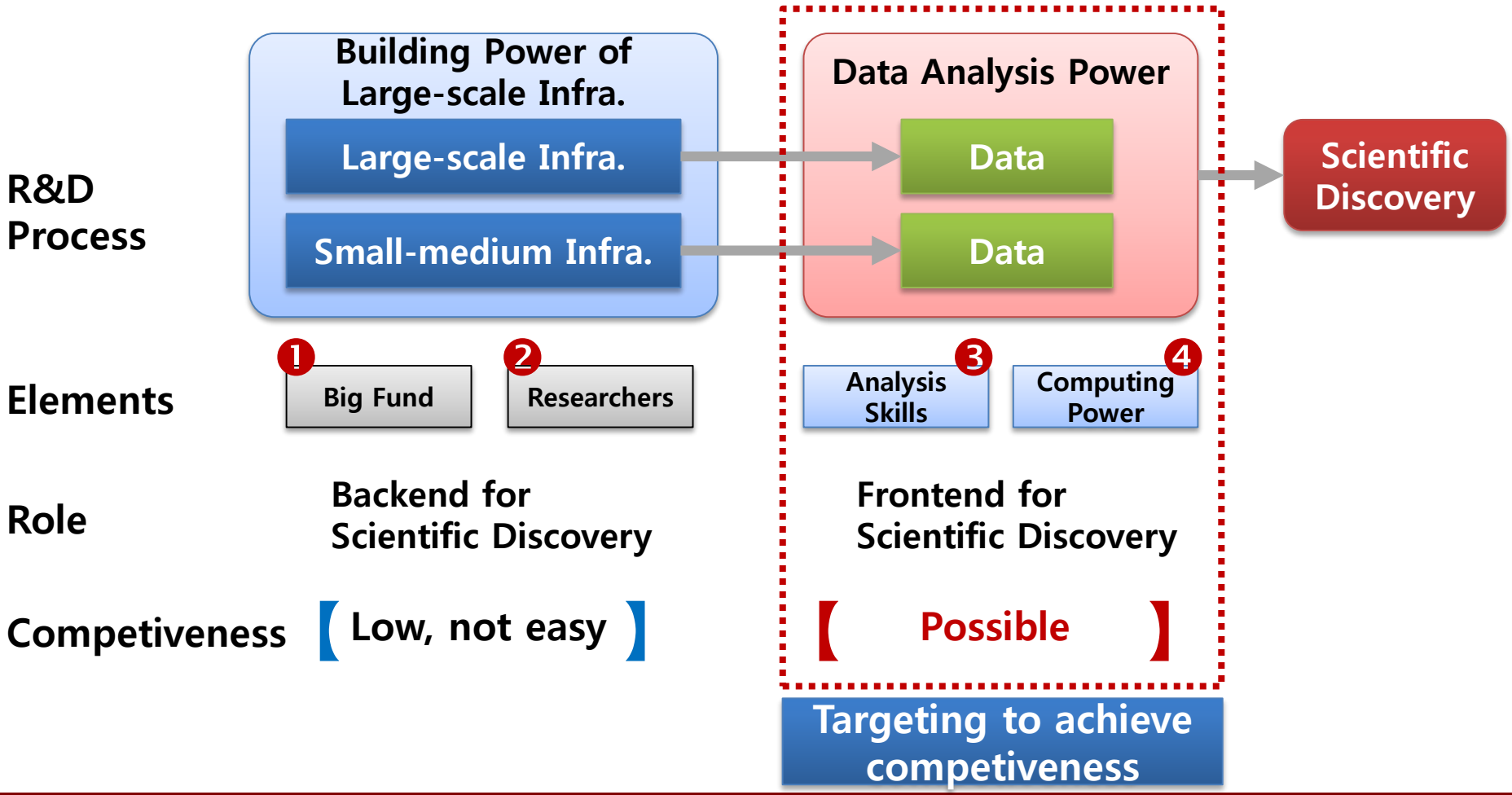
287.264ms

Response time: 155.546ms ↑
10 hops removed

Role Expansion to National Data Center

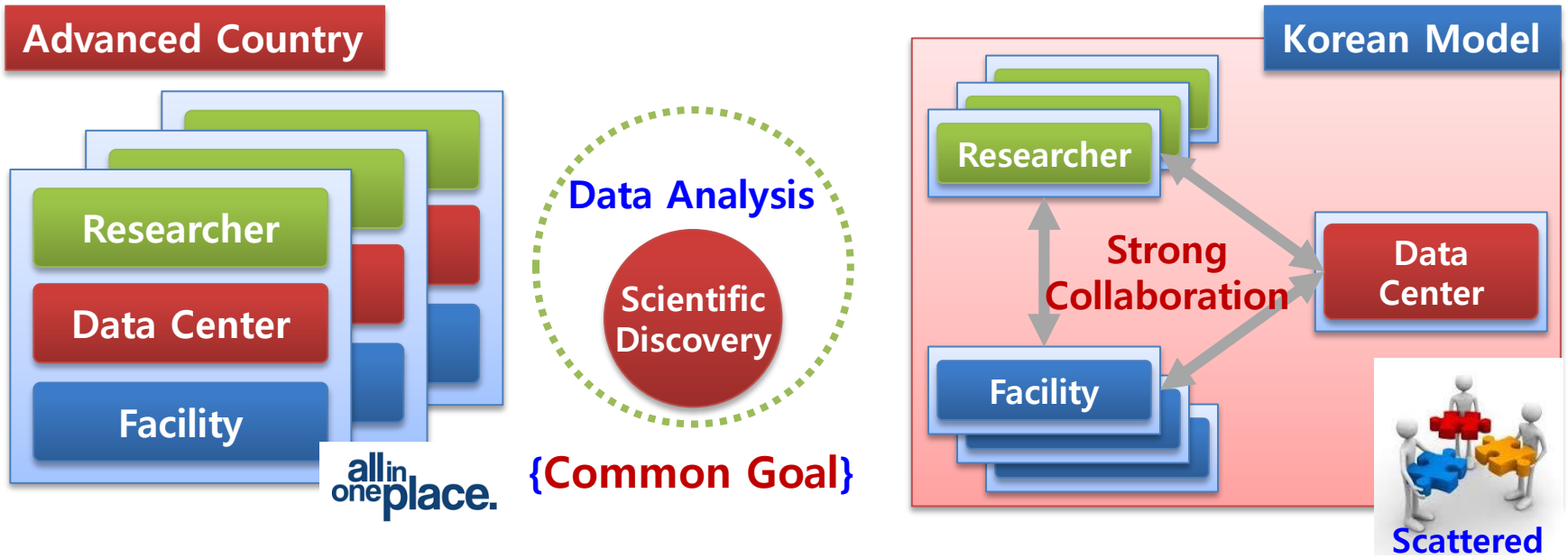
Why we have to focus on data analysis

Data intensive R&D through large-scale infrastructure requires a new strategic model suitable for Korea circumstance



A Korean Model in Data & Infra-driven R&D Era

A model of considering scale is required to keep up competition level with advanced countries of having large facilities and research groups

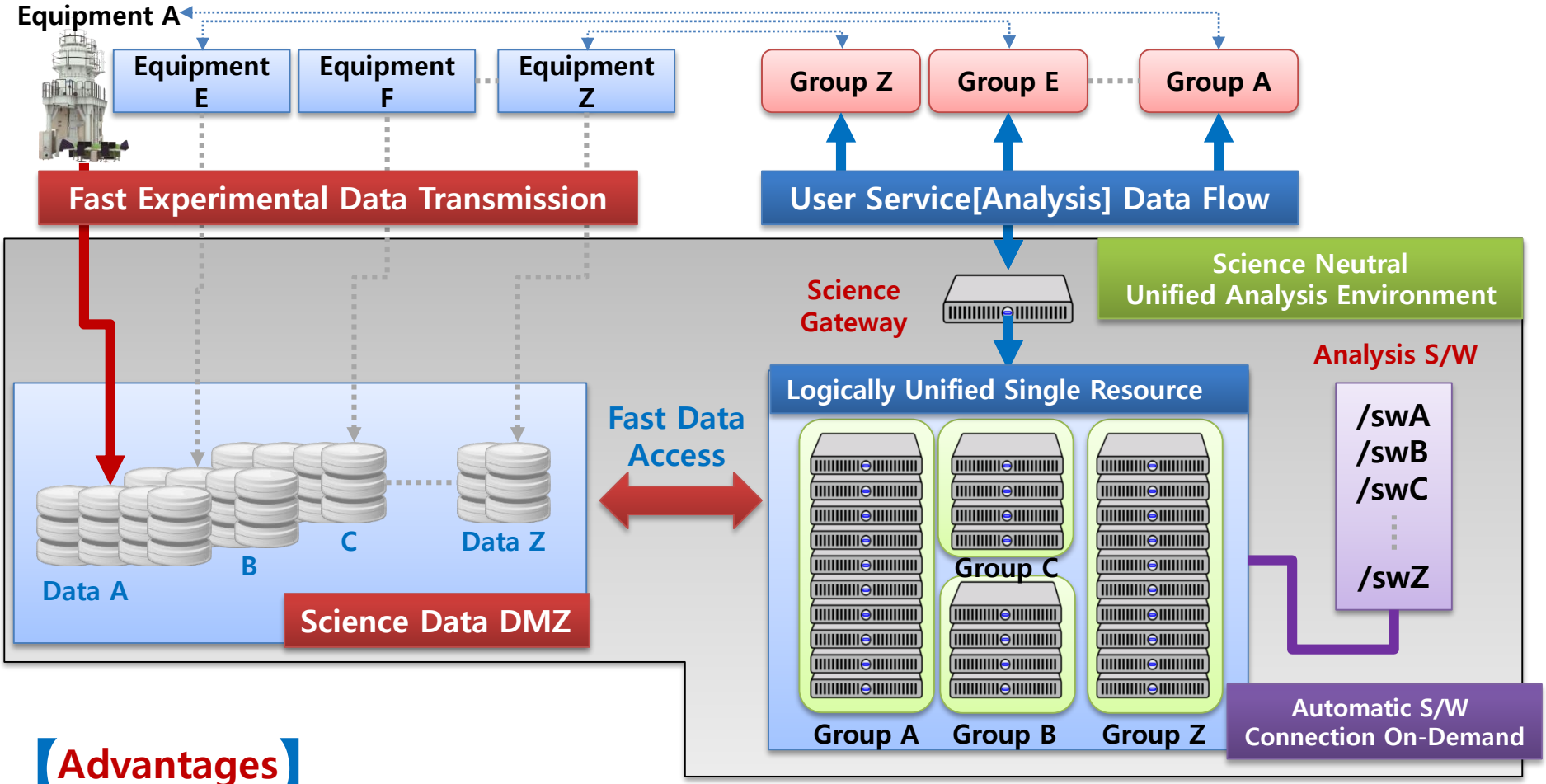


- Large scale research group
- Large scale research facility
- Dedicated data center

vs.

- Small scale research group
- Small or medium scale facility
- Not easy to have a dedicated data center (in size and experts)

Unified Data Analysis Platform @ KISTI-GSDC

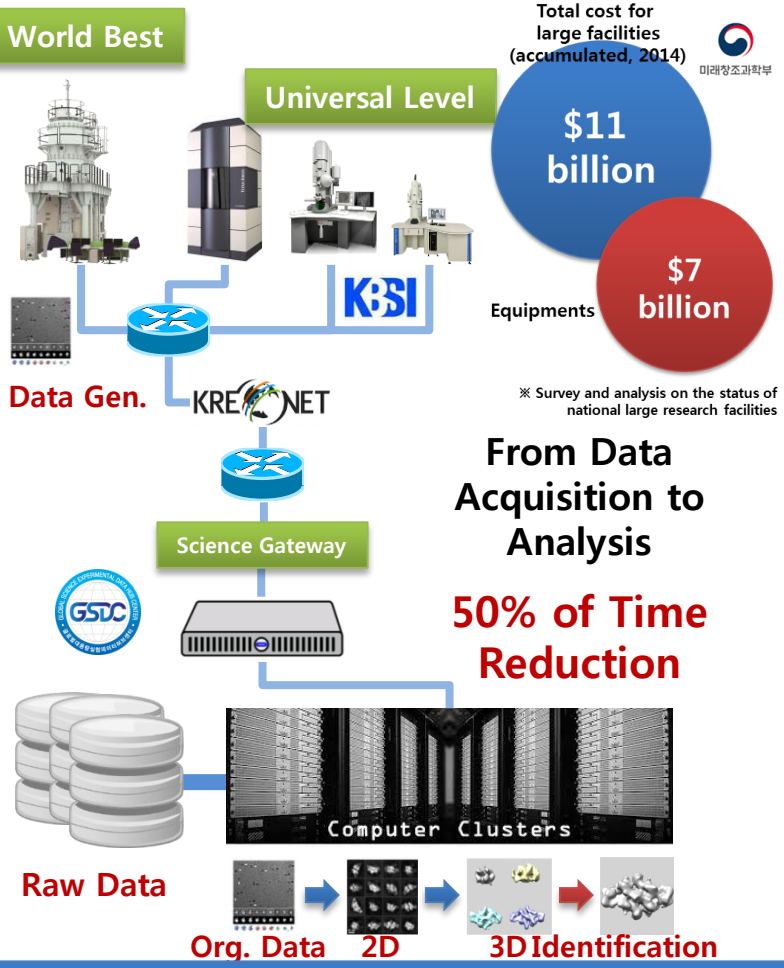


[Advantages]

1. Pluggable Science → Supports in unified way for various groups and equipment
2. Data Infra. Sharing → Reuse and full utilization of infra. saving tax-payer's money
3. Simple R&D Process → Fast results from data acquisition to data analysis

Role Expansion

Transmission electron microscopy Data Sharing·Analysis Farm



Improving R&D efficiency by linking data center and large research facility

Officially joined KEK Belle II Computing Grid

work among Belle II Grid sites.

Sincerely,

Toshiaki Kaneko, Ph.D.
Computing Research Center, Head
High Energy Accelerator Research Organization (KEK)

MOU
KISTI & Belle II

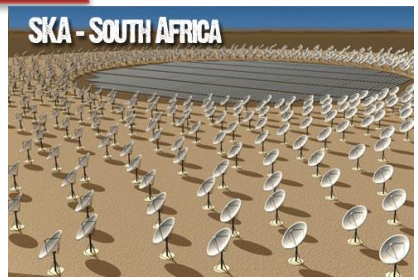
2016.11.16

RAON

New Accelerator in Korea

Utilization of Tier-1 know-how for data management

SKA



Regional data center (under discussion)

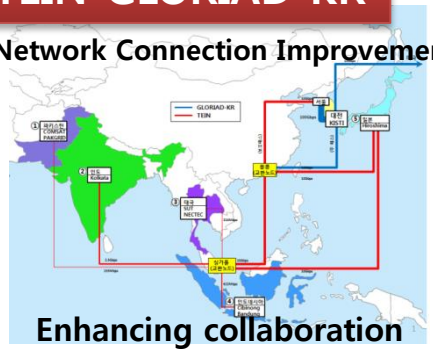
KAGRA

Gravitational Wave Detector in Japan

Officially participation in data management

TEIN-GLORIAD-KR

Network Connection Improvement



Enhancing collaboration in Asian community

Conclusions

Conclusions

Data-driven R&D

- Data and infrastructure are the key in scientific discovery
- CERN's recent 750GeV thing shows that we are in data-driven research era and we trust in data
- Three driving forces - openness of access, collaborations and data

WLCG Tier-1

- helps to have competiveness for data intensive research in Korea through synergy with KISTI-ICT professional institute
- helps to expand WLCG knowledge to various science domains smoothly and to build strategy plan for Open Science

Future

- Tight collaboration with CERN WLCG project, extending to development project in computing general beyond operational service
- KISTI-GSDC, unique infrastructure for data intensive research in Korea, our role is being expanded to national data center for science and national agenda

Thank you.