

# KISTI-SUT Distributed Storage

The 5th ATCF @ Mumbai, India  
24 - 25 October 2019

Sang-Un Ahn, Jeong-Heon Kim, Chinorat Kobdaj

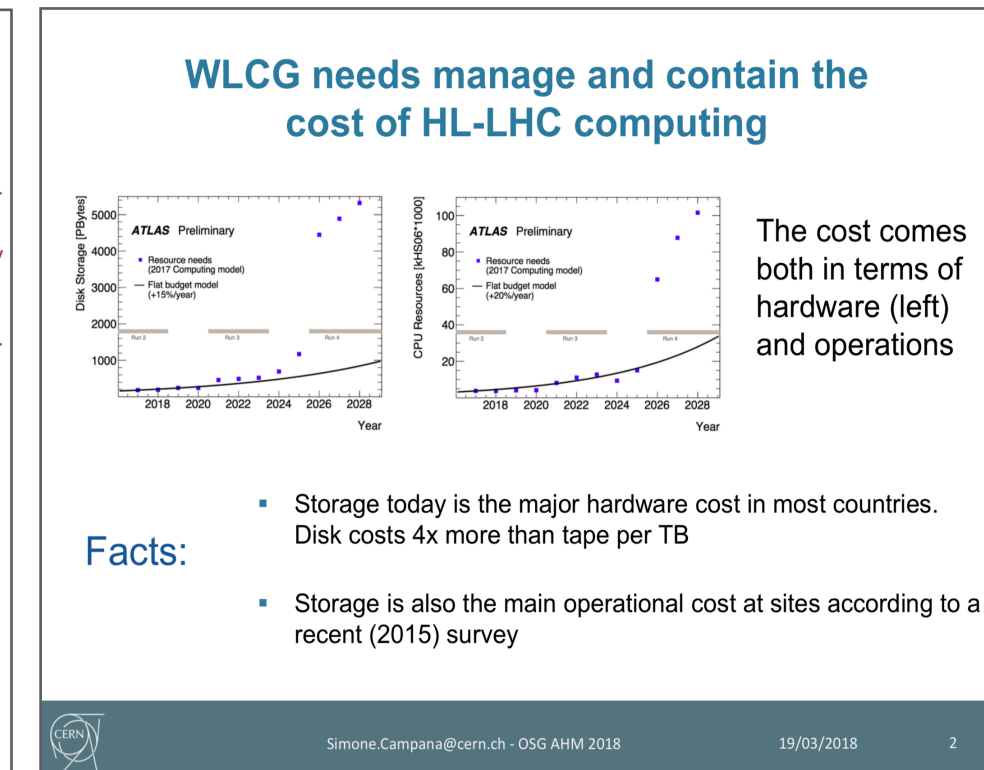
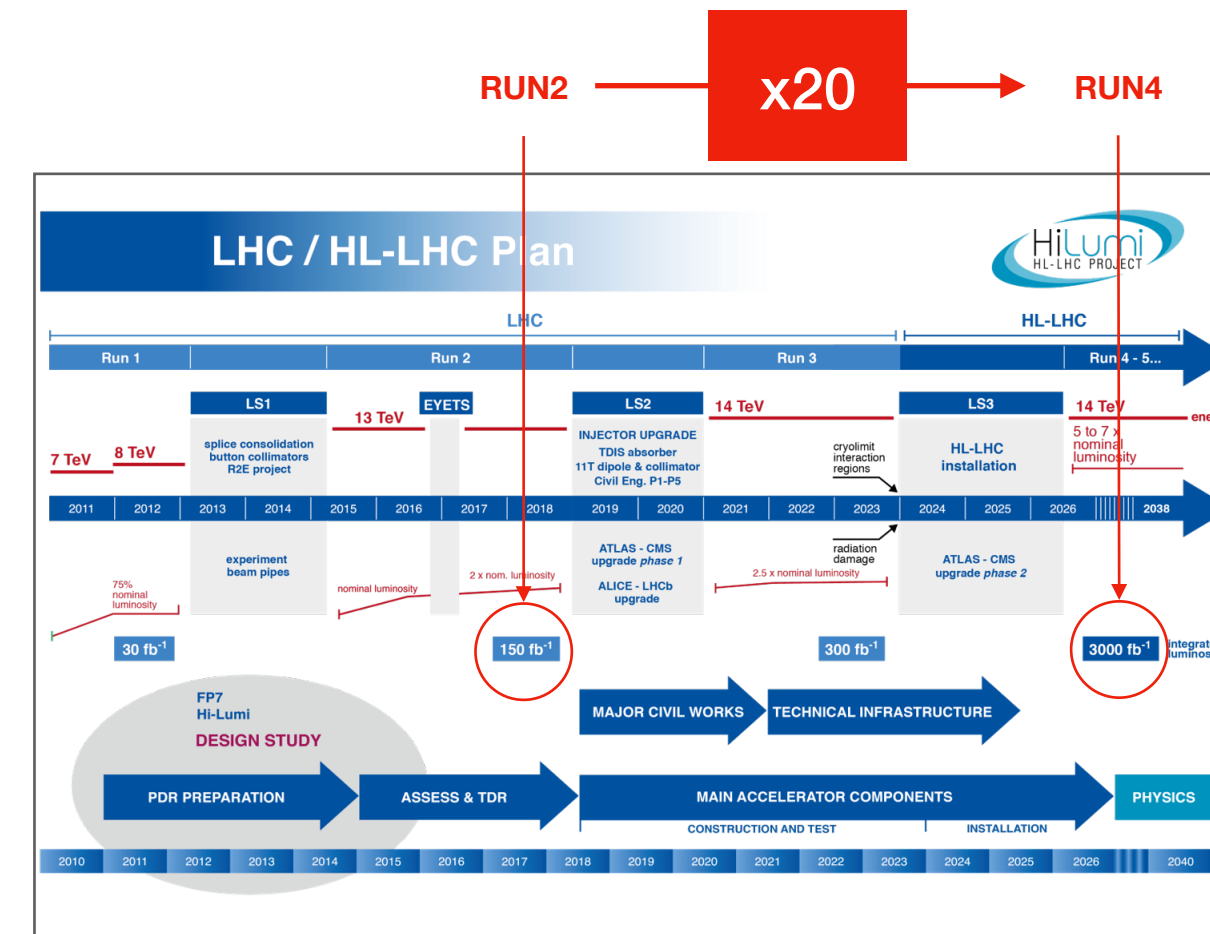


한국과학기술정보연구원  
Korea Institute of Science and Technology Information

Background

# Data Challenges for Upcoming LHC RUNs

- Data challenges foreseen in HL-LHC (RUN4) starting 2026
  - x20 more integrated luminosity compared to RUN2
  - High increase of compute and storage capacities are demanded
  - How do we deliver them in flat-budget scenarios?

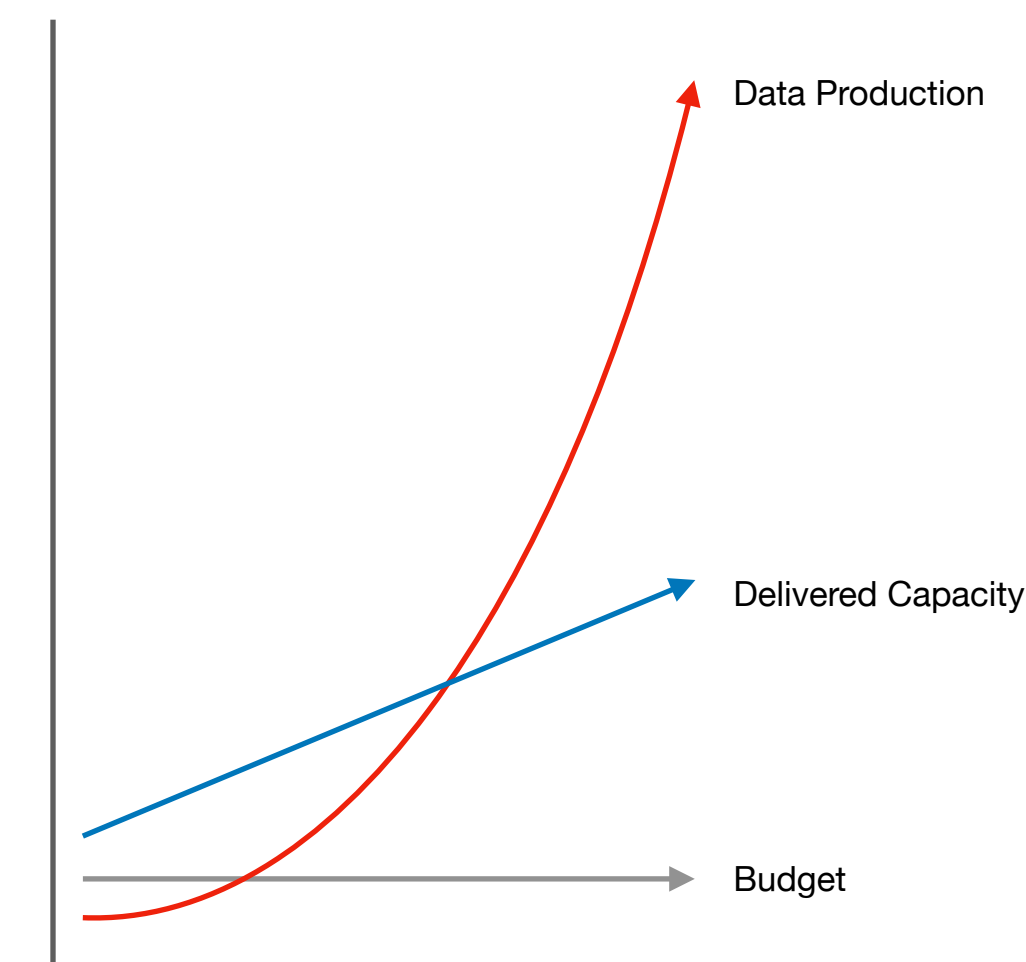
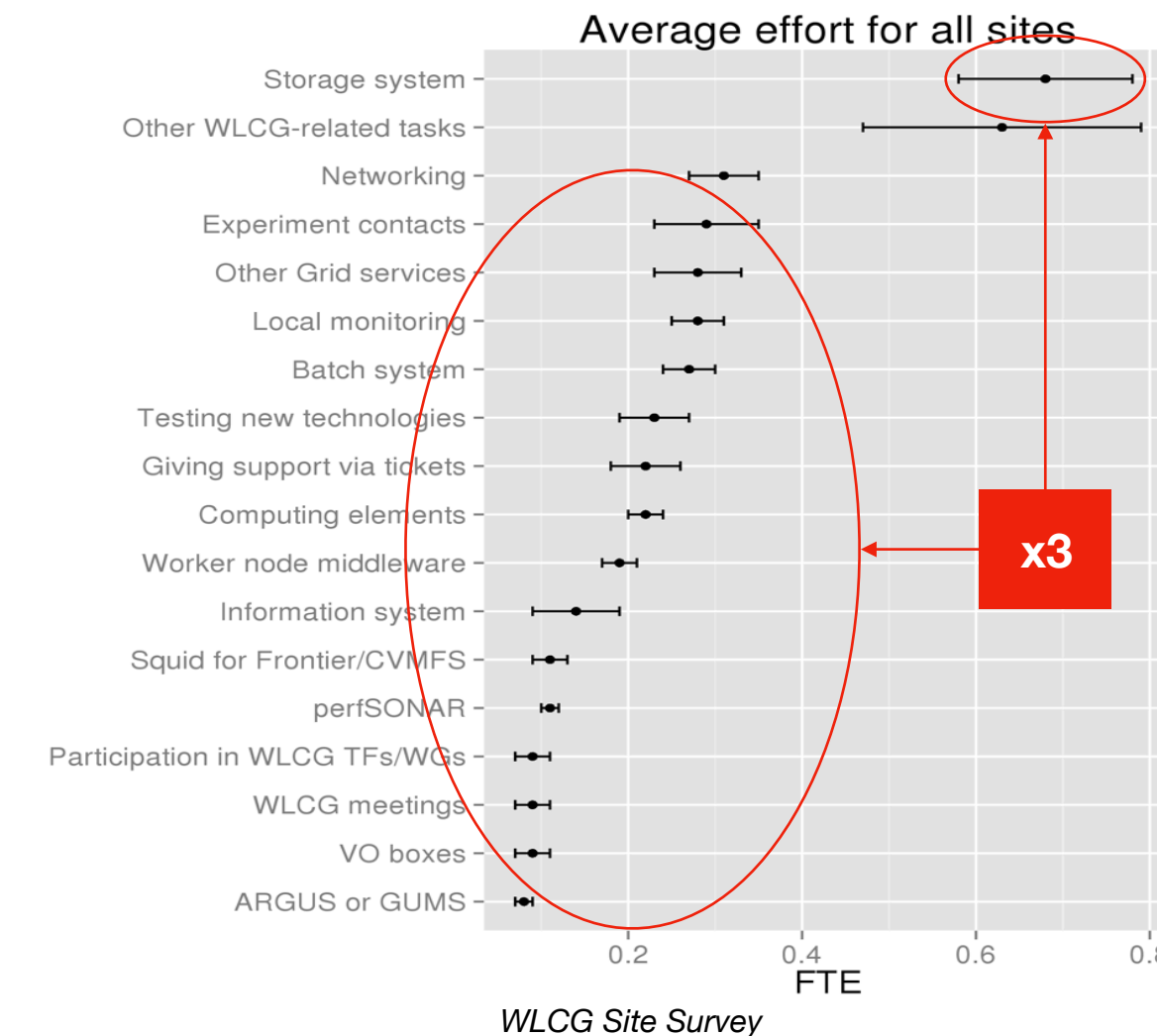


- *“Storage is the main operational cost at sites” (Simone Campana)*

- WLCG 2015 Survey (<https://twiki.cern.ch/twiki/bin/view/LCG/WLCGSiteSurvey>)
- Disk costs 4x more than tape per TB

- Implementing a data lake model based on distributed storages is on-going

- EU-Lake (Europe, WLCG), Open Storage Network (US)
- With the help of advanced high-bandwidth networking available throughout the world



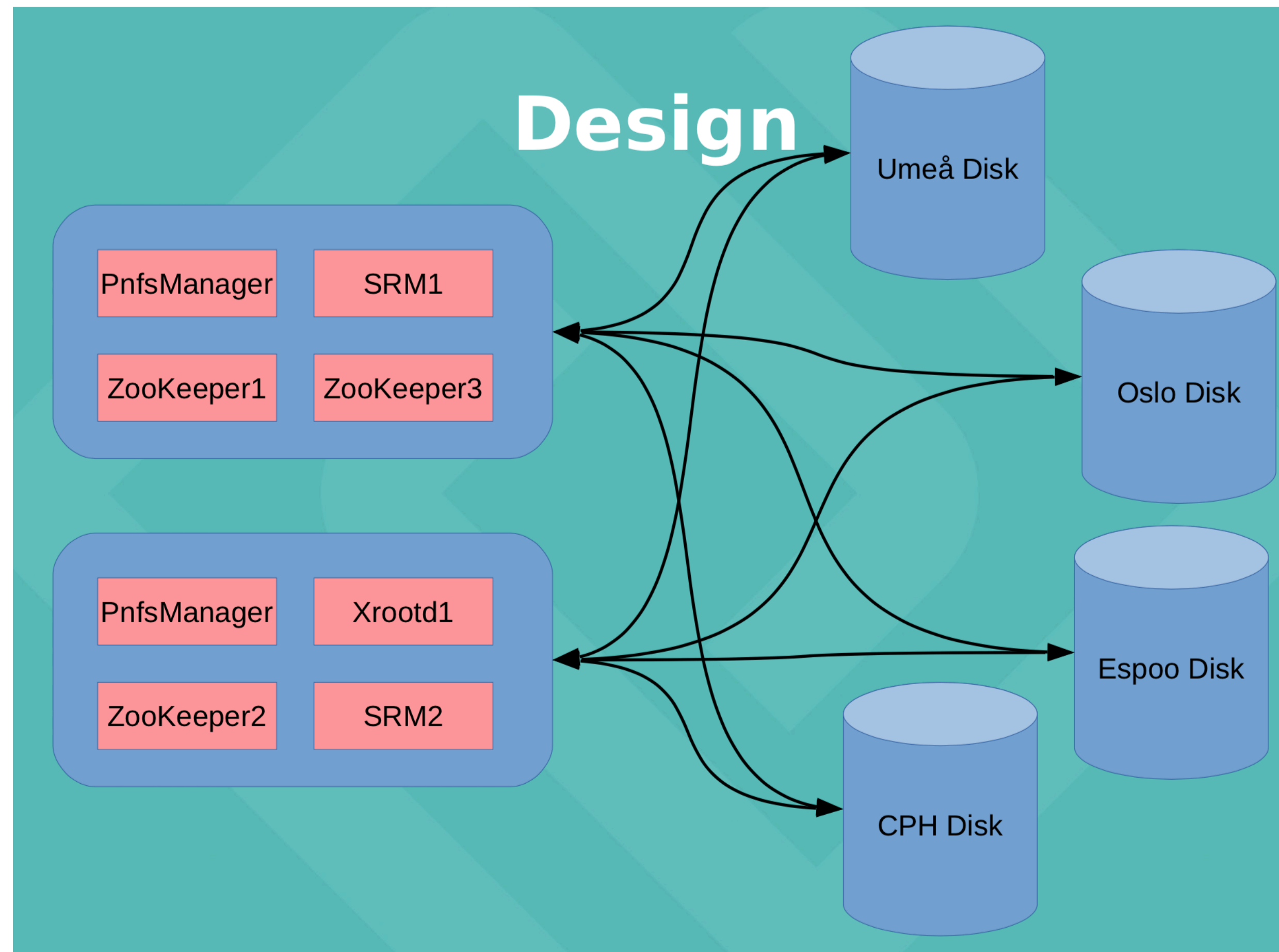
# Distributed Storage in Asia

- A strong collaboration is needed to overcome **Data Challenges foreseen in HL-LHC era**
  - Resource requirements to T1/T2 sites from experiments will increase accordingly
  - Reducing the operational costs is the key; Technology advances? → Consolidated efforts are needed
- **Distributed Storage across Asian sites**
  - *A handful tool to exploit and evaluate the advanced networking in Asia*
  - ATCF4 was a starting point to discuss this

# Discussion

- Improve latencies and bandwidths among distributed sites(storages)
- Prove data transfer capacity between distributed sites upon the current networking configuration
- Consider how reflect different requirements from different VOs, e.g. ATLAS, CMS, ALICE with a single distributed storage
- Consider how reduce operational costs meeting diverse use cases
- Share expertise and technologies
- Propose to setup a distributed storage between KISTI and SUT to address issues above
  - Consolidate distributed storage with EOS and provide a single entry point

# The Nordic Model

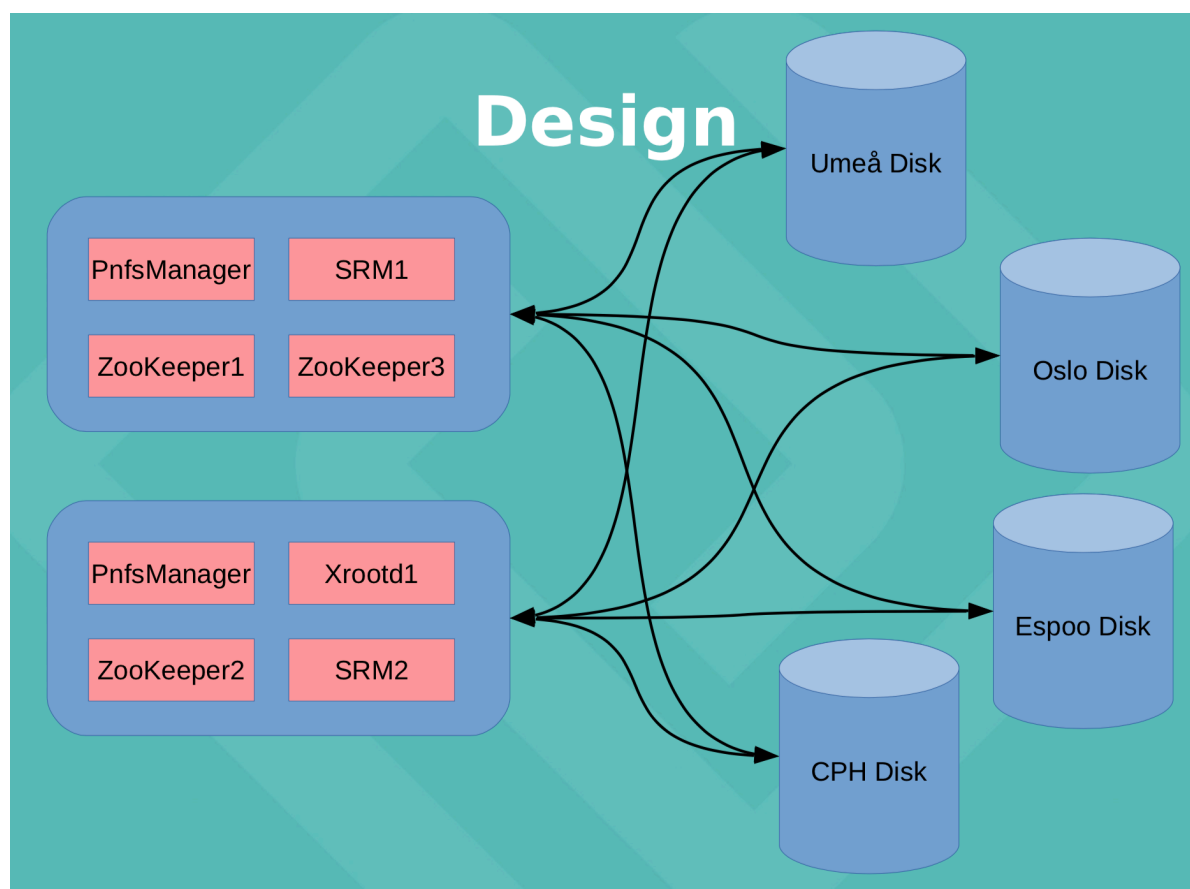


Strong motivation  
Consolidated collaboration  
Sharing expertise  
Co-work on technologies

NeIC Distributed Storage Design

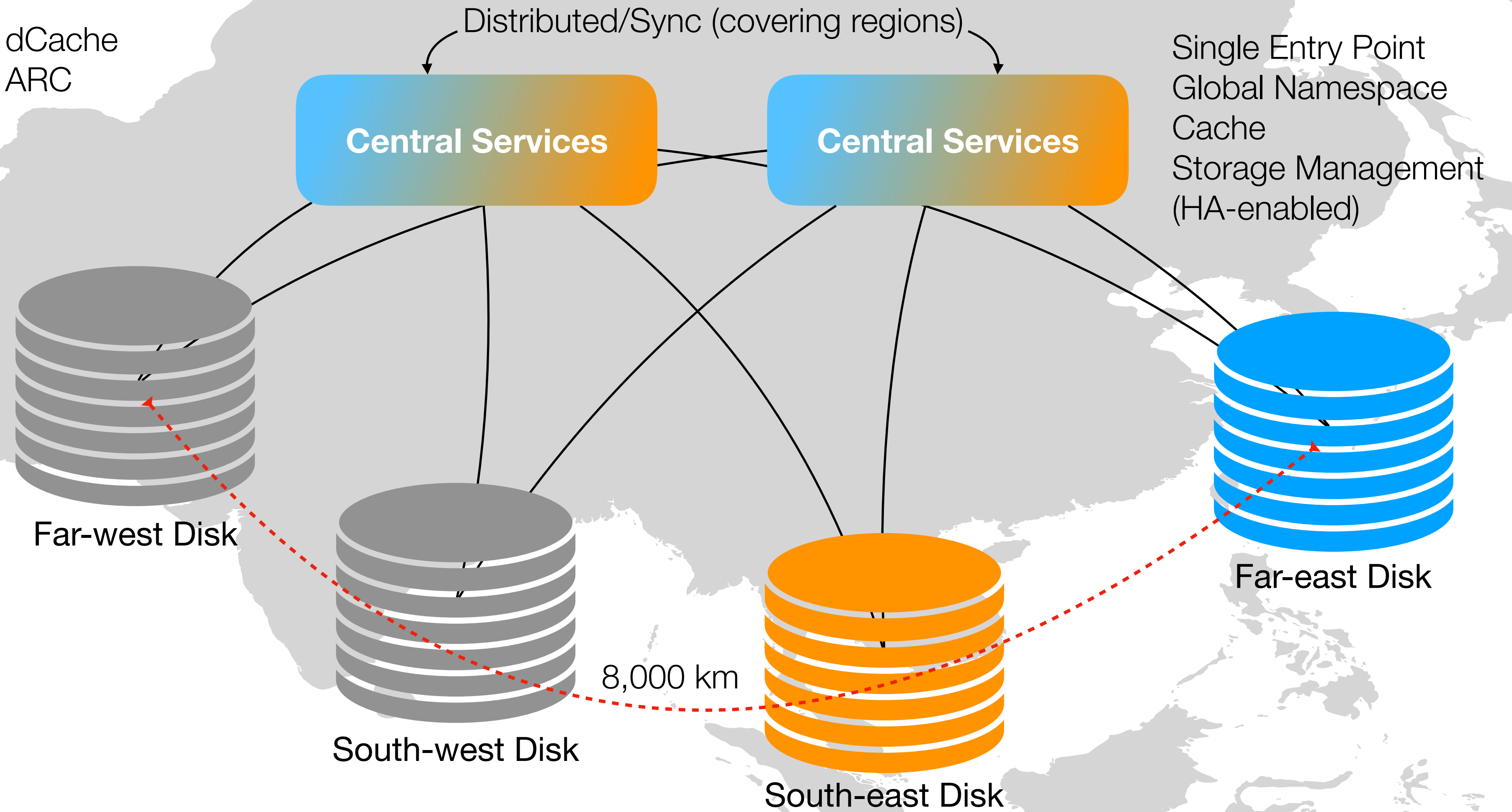
# A Nordic Model

44,579,000 km<sup>2</sup>



NeIC Distributed Storage Design

dCache  
ARC



Strong motivation  
Consolidated collaboration  
Sharing expertise  
Co-work on technologies

# Project Status

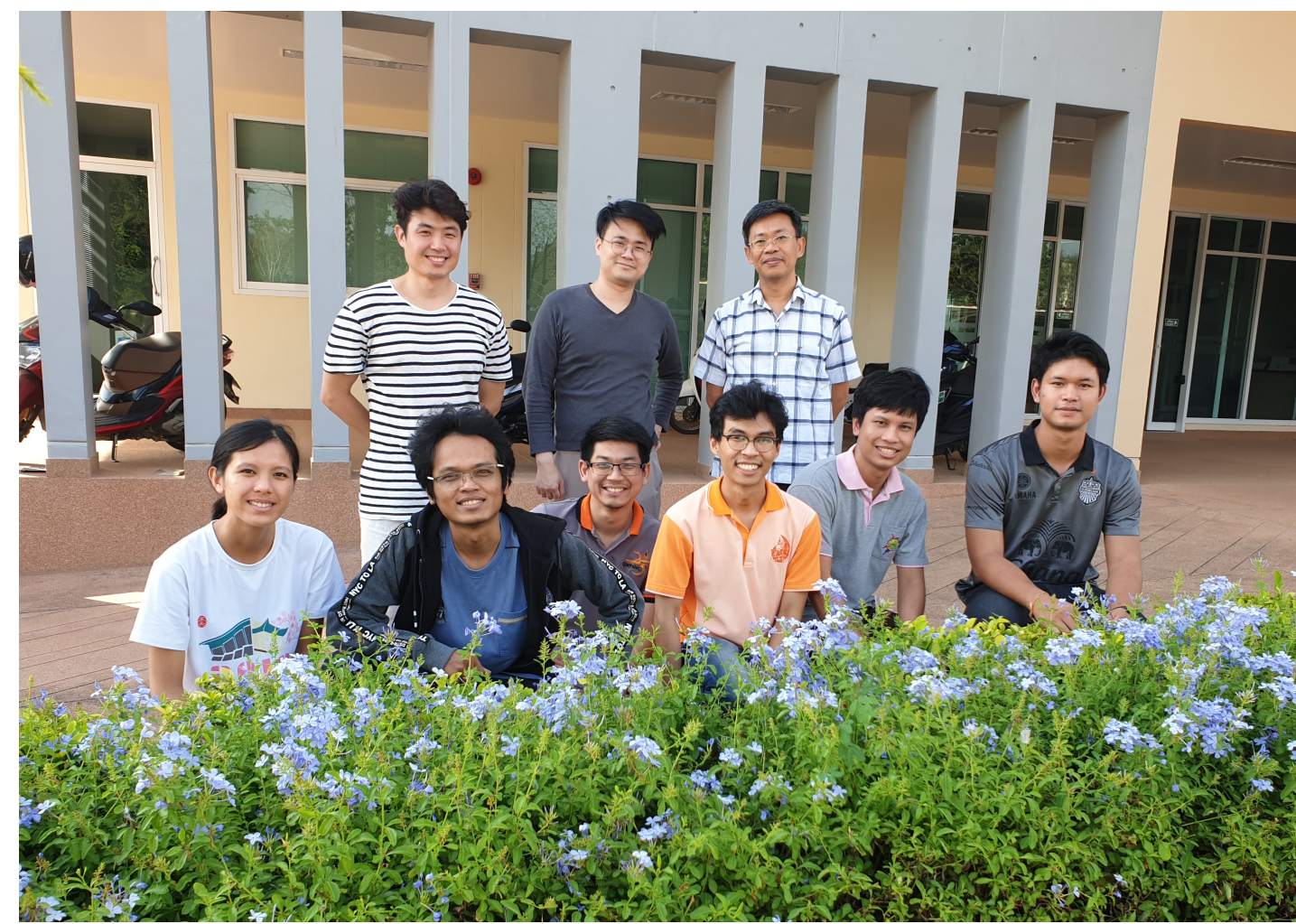
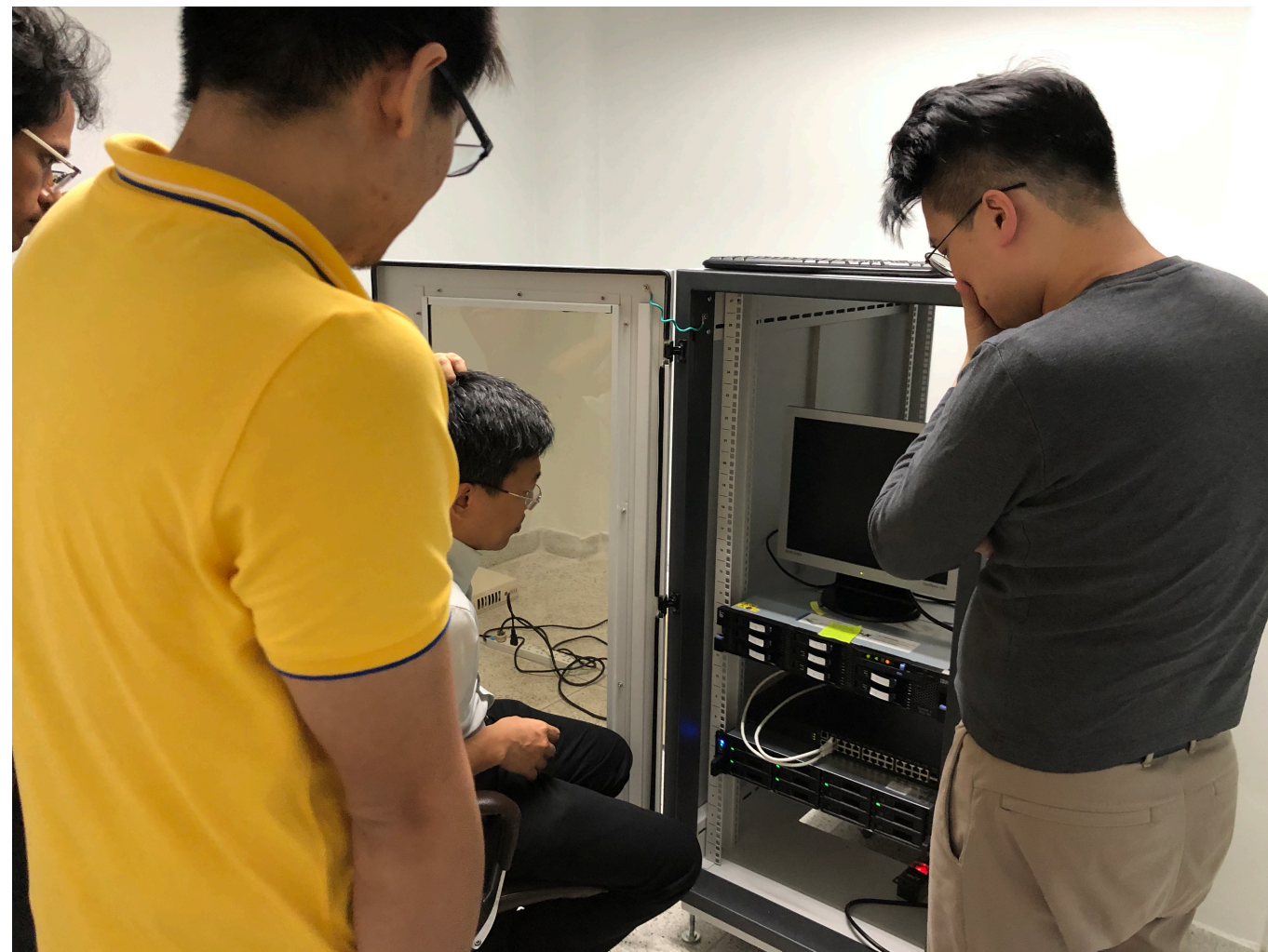


# A bit of History

- Agreement @ ATCF4 in November 2019
  - KISTI and SUT agreed to establish a distributed storage setup in order to demonstrate the feasibility of storage consolidation in the region
  - Experts and students exchange program
- ATCF SC meeting @ Seoul in April 2019
  - Prof. Chinorat Kobdaj proposed a small workshop regarding EOS deployment as well as a training program for students
- EOS Workshop @ SUT in August 2019
  - EOS deployment based on Docker container and using Ansible playbook (written by Dr. Jeong-Heon Kim)
  - 3 Days of Tutorial for SUT students including EOS

# EOS Workshop @ SUT

- EOS Docker Deployment via Ansible Playbook + Tutorials for students
  - Hardware Checking, CentOS 7 Installation, NAS Volume Partitioning
  - EOS Admin Guide, Docker Basics, YAML format for Ansible
- Standalone EOS instance setup on a single server using containers



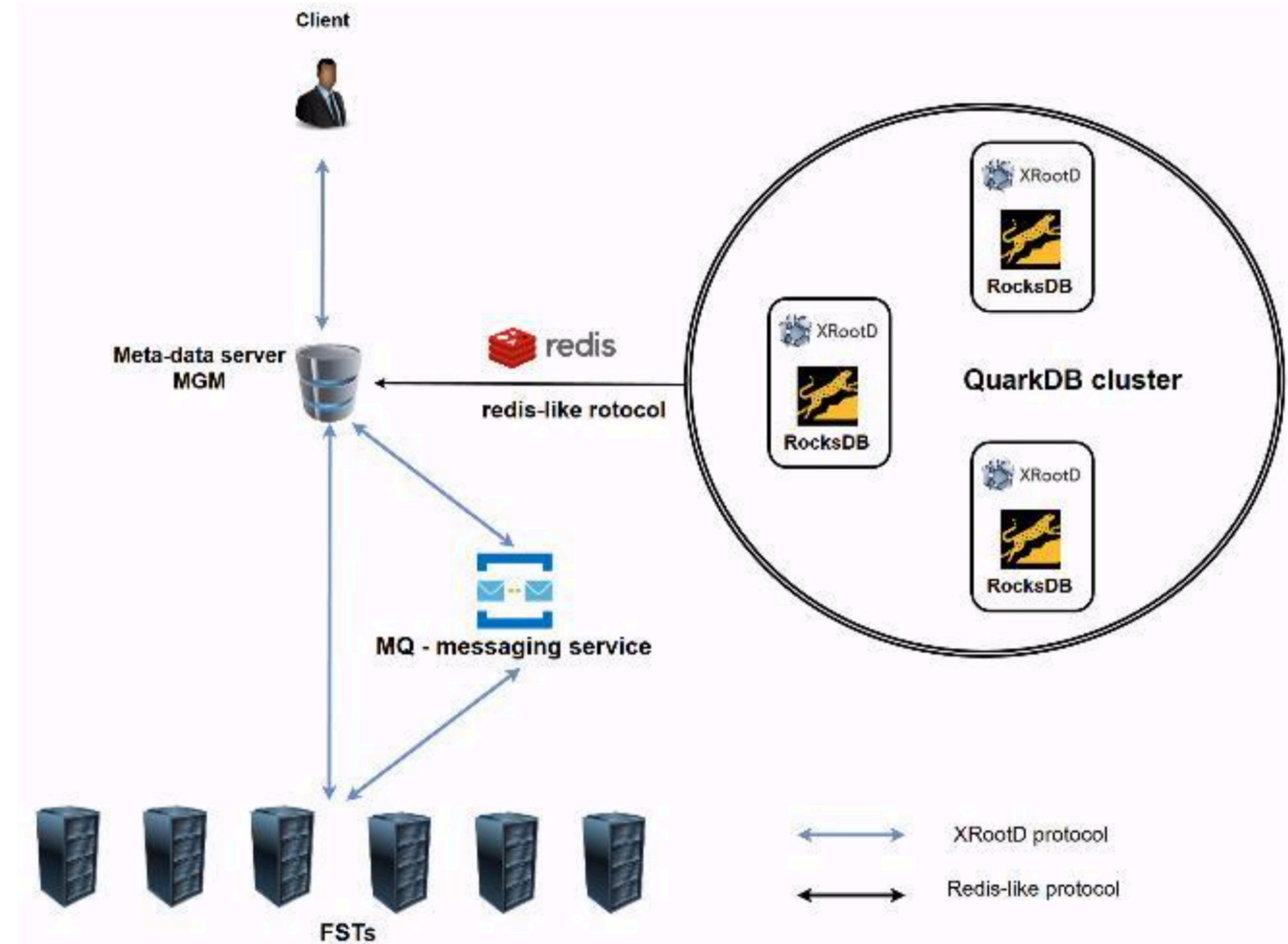
# EOS at a Glance

- Components

- MGM : Management Server
  - ▶ In-memory Namespace will be deprecated soon
- MQ : Message Queue
- FST : File Storage Server

- Important features for a distributed storage

- Location awareness
- GEO Scheduling functionality



# Initial Setup

- Two separate EOS instances @ KISTI and SUT using different GeoTag
  - “kisti::gsdc::d10” for KISTI
  - “sut::bnct::a209” for SUT
- Complete Docker container set for all EOS components
  - 1 MGM, 3 FSTs, 3 QDBs, 1 MQ, 1 KRB
- Deployment via the automation script using Ansible playbook (YAML format)
- EOS Components were deployed and started successfully, local tests were done

# Issues

- Mixed authentication with sss and krb
  - Resolution: enforcing krb for admin user (client)
  - Still this issue persists, need to understand authentication mechanism of EOS
- Federating two separate EOS instances
  - MGM Master/Slave fail-over between the instances
  - In theory, a kind of "Global" MGM should be required, however...

# Case Study

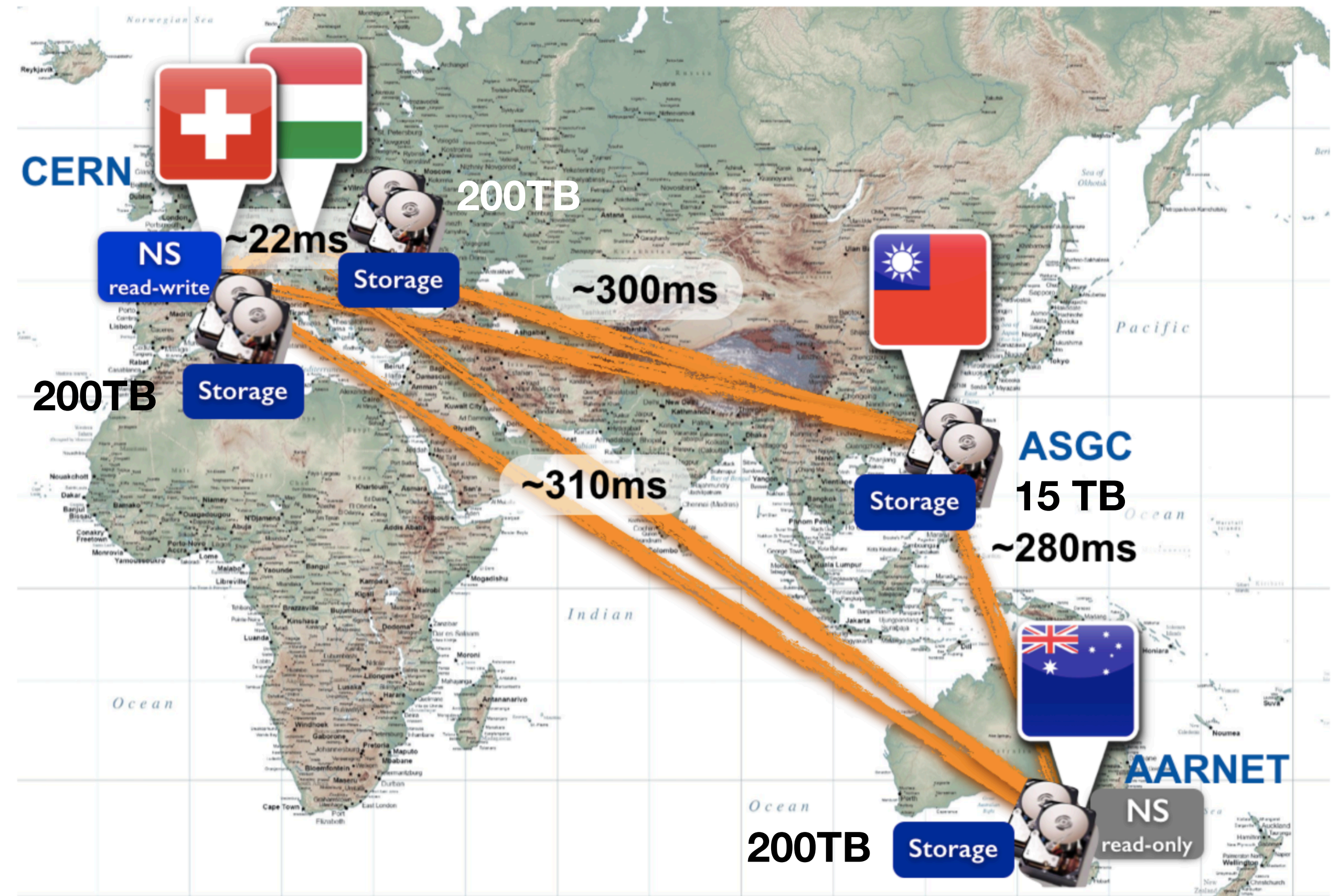
- CERN tested a distributed storage setup using EOS between Meyrin and Wigner
  - "di-EOS - "distributed EOS": Initial experience with split-site persistency in a production service" presented @ CHEP2013
  - 22ms latency, 100Gbit/s between the two sites
- CERN, AARNET(AU), and ASGC(TW) tried to setup and test EOS deployment in wide area network
  - "Global EOS: exploring the 300-ms-latency region" presented @ CHEP2016
  - Latency > 300ms, 16,500km apart

# Global EOS

- Goal
  - "... to test if the EOS software components were able to cope with latencies much higher than 30ms and how the entire software stack was affected by this."
  - "... to explore and discover possible flaws caused by heartbeats retries and default timeouts in such environments."
  - "... to measure how easy it is to deploy this global infrastructure ... and describe how it is possible to improve its performance (hiding network latencies)."

# Global EOS cont'd

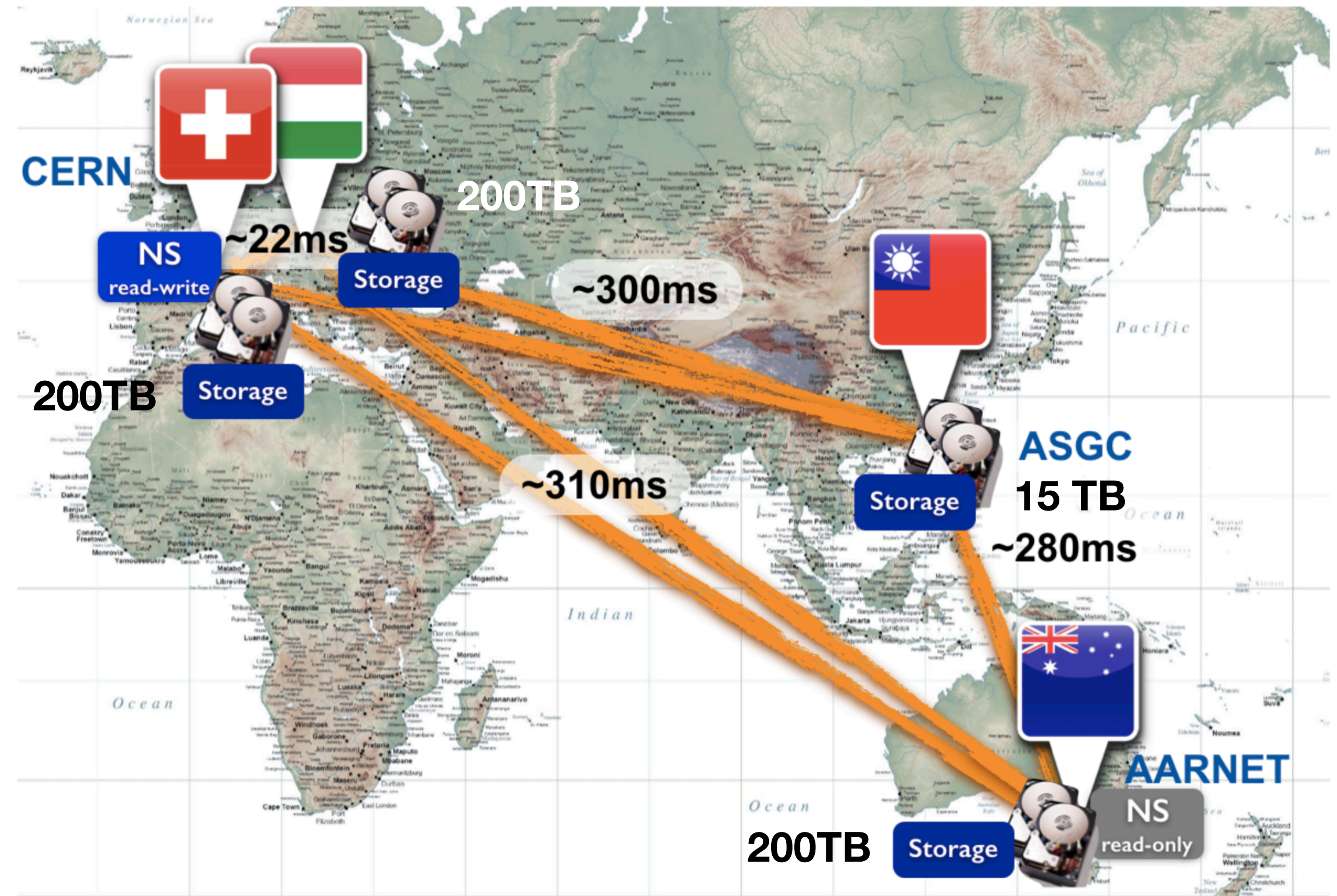
- MGM Master @ CERN; Slave @ Melbourne
  - "EOS keeps constantly in sync the two namespaces located between 290 and 320 milliseconds away"
    - ▶ EOS sync is required for In-memory Namespace; no longer needed for QuarkDB
- Routing Asymmetry
  - "Latencies between storages were computed as averages over time, since the network underneath was not fully dedicated and the routing was changing on a daily or weekly bases"





# Global EOS Conclusion

- Confirmed that,
  - "... the stability and the robustness of EOS in working with such latency, no adaptation of timeouts or other parameter was needed in order to set up the system on this very large geographical scale,"
  - "the system worked immediately out of the box."
- Client behaviour @ Melbourne writes to disk pool @ Melbourne
  - "... contacted the read-write namespace located in Geneva and the data transfers is scheduled to a Melbourne disk."
  - Read is not affected by such a big round trip time
- Average speed of data transfers in MEL-GVA ~ 45MB/s



# Service Status

POWERED BY 

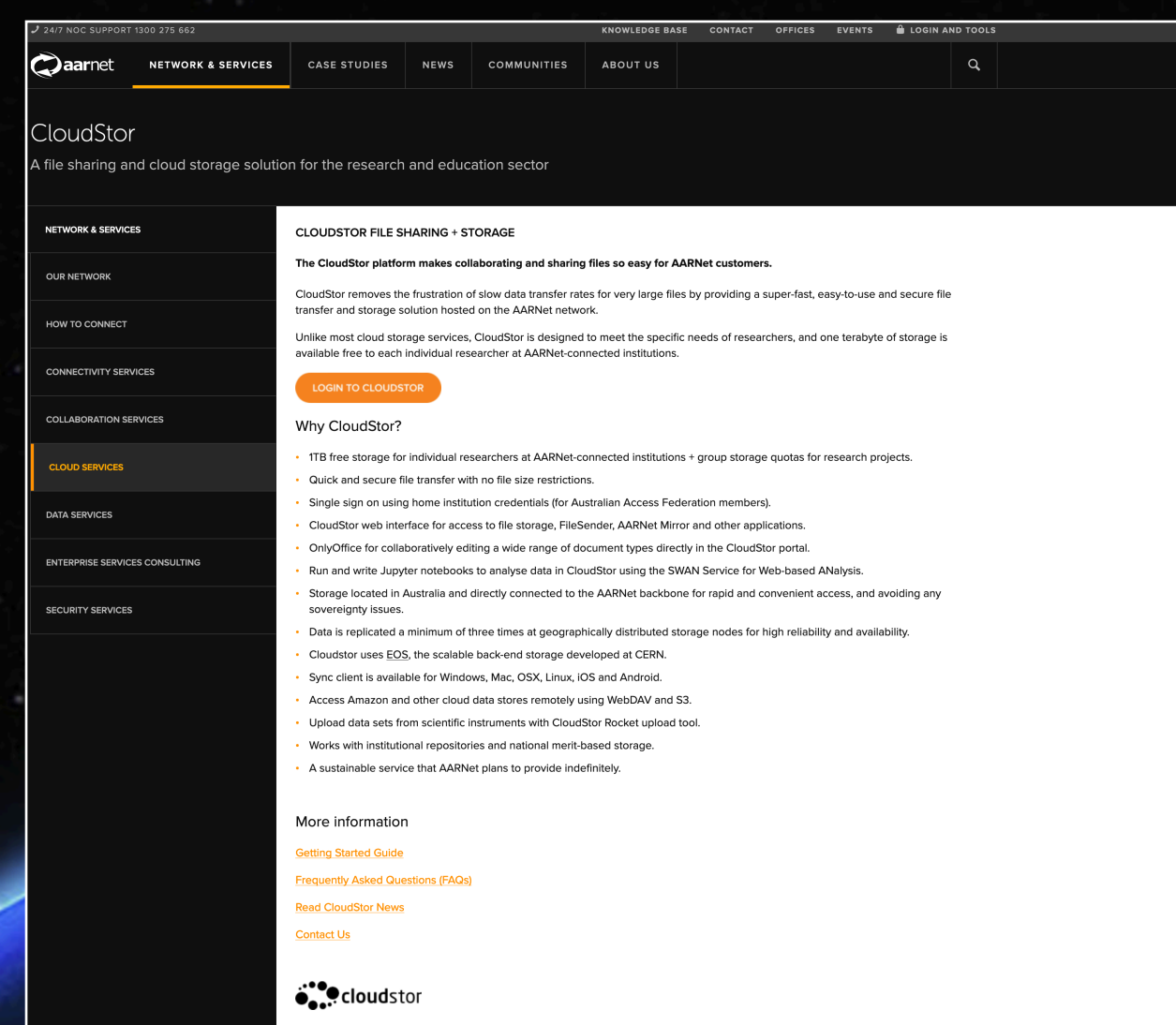
# CloudStor Status

CloudStor services are operational

(~ 1 second ago)

A Good Example of Science Box

- EOS Docker Installation
- CERNBox Deployment
- SWAN (Jupyter-hub)



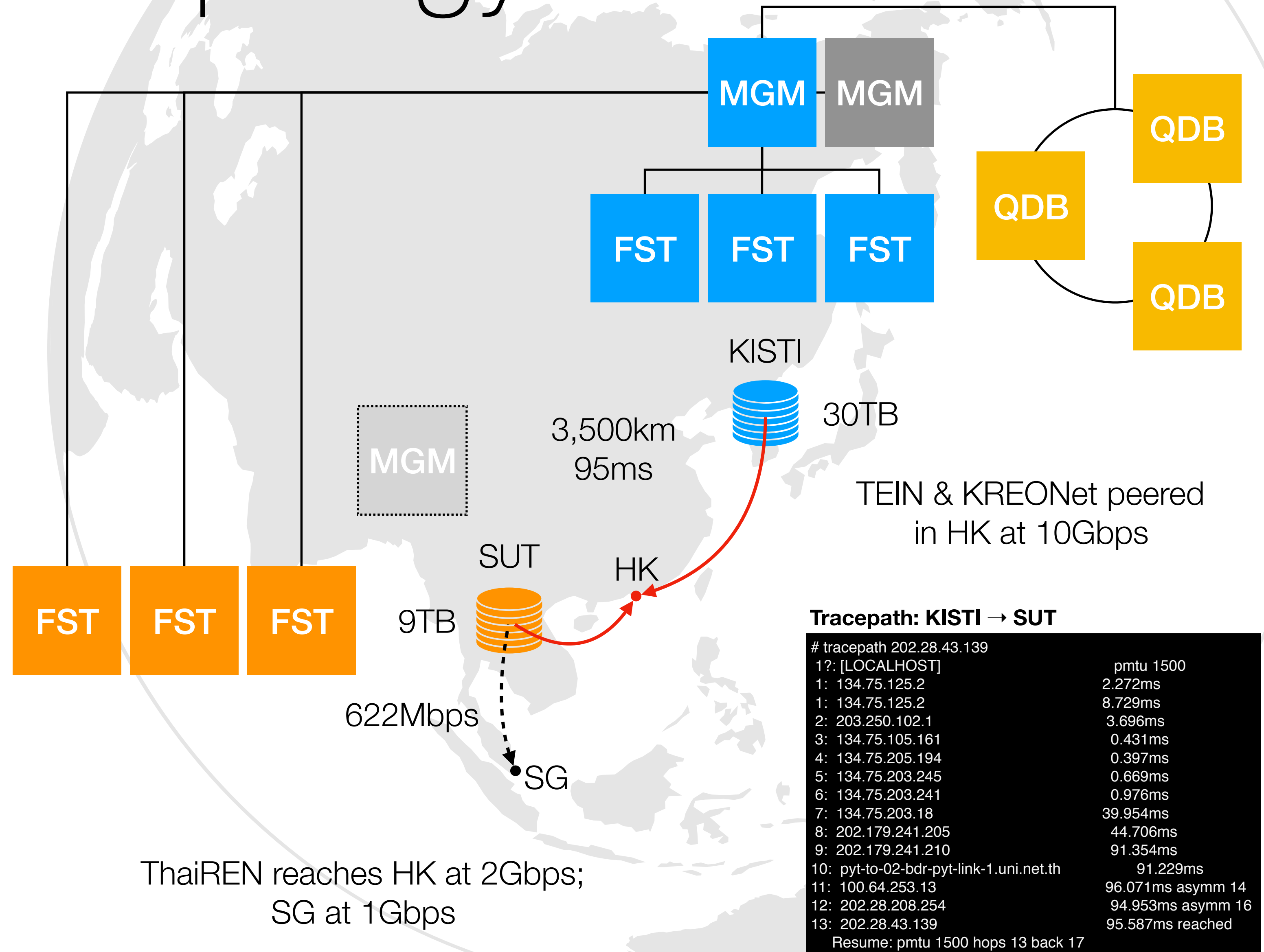
The screenshot shows the CloudStor website interface. At the top, there's a navigation menu with 'NETWORK & SERVICES' highlighted. Below the menu, the 'CloudStor' logo and tagline 'A file sharing and cloud storage solution for the research and education sector' are visible. The main content area is titled 'CLOUDSTOR FILE SHARING + STORAGE' and contains text describing the service's benefits, such as fast data transfer and secure storage. A 'LOGIN TO CLOUDSTOR' button is present. Below this, there's a 'Why CloudStor?' section with a list of features and benefits. At the bottom of the page, there's a 'More information' section with links to 'Getting Started Guide', 'Frequently Asked Questions (FAQ)', 'Read CloudStor News', and 'Contact Us'. The CloudStor logo is also displayed at the bottom left of the page.



- The outcome of CERN-AARNET collaboration concerning EOS deployment in wide-area network (> 300ms latency)
- Cloud storage provided to individual researchers
- Integration with ID-Federation (e.g. EduGAIN)

# 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



# Current Issues

- Operation expired for data transfers (> 10MB files) to FSTs @ SUT (sut::bnct::a209)
  - Small files copy (< 10MB) looks OK
  - SSH Copy (SCP) performs well between the two container hosts: ~17MB/s, which is equivalent to 120Mbps
  - Local data transfer within KISTI performs well: ~ 500MB/s (about 4Gbps)
- Mixed authentication problem still persists

# Next Step

- Further investigation into,
  - Data transfer performance issue
  - Mixed authentication problem
- GSI authentication will be enabled
- Deploy a MGM slave at SUT site + distributed QuarkDB cluster
  - No use case with having off-site QuarkDB cluster setup
  - EOS developers confirmed that replication across QuarkDB should work fine in such high latencies
    - ▶ <https://eos-community.web.cern.ch/t/mgm-sync-and-qdb-replication-in-tens-or-hundreds-milliseconds-of-distance/366/10>

# Conclusion

- The pilot project on KISTI-SUT Distributed Storage based on EOS has started
  - Facilitating the advanced networking environments in Asia
  - Prototyping the storage consolidation for a Data Lake in the Region
  - Provision for LHC Data Challenges beyond RUN3
- Seeking for new candidates to expand the distributed setup