

# A pilot project deploying EOS for the distributed storage between Korea and Thailand

Sang-Un Ahn<sup>1</sup>, Heejune Han<sup>1</sup>, Jeong-Heon Kim<sup>1</sup>, Chinorat Kobdaj<sup>2</sup>, Hee Jun Yoon<sup>1</sup>

<sup>1</sup>KISTI, Daejeon, South Korea

<sup>2</sup>SUT, Nakhon Ratchasima, Thailand

*EOS Workshop*

*2 - 4 February 2020*

IT Amphitheater @ CERN





# 5<sup>th</sup> Asia Tier Center Forum & 1<sup>st</sup> 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>

[https://indico.cern.ch/event/739884/contributions/3632257/attachments/1947840/3231810/atcf5\\_summary\\_sahn.pdf](https://indico.cern.ch/event/739884/contributions/3632257/attachments/1947840/3231810/atcf5_summary_sahn.pdf)

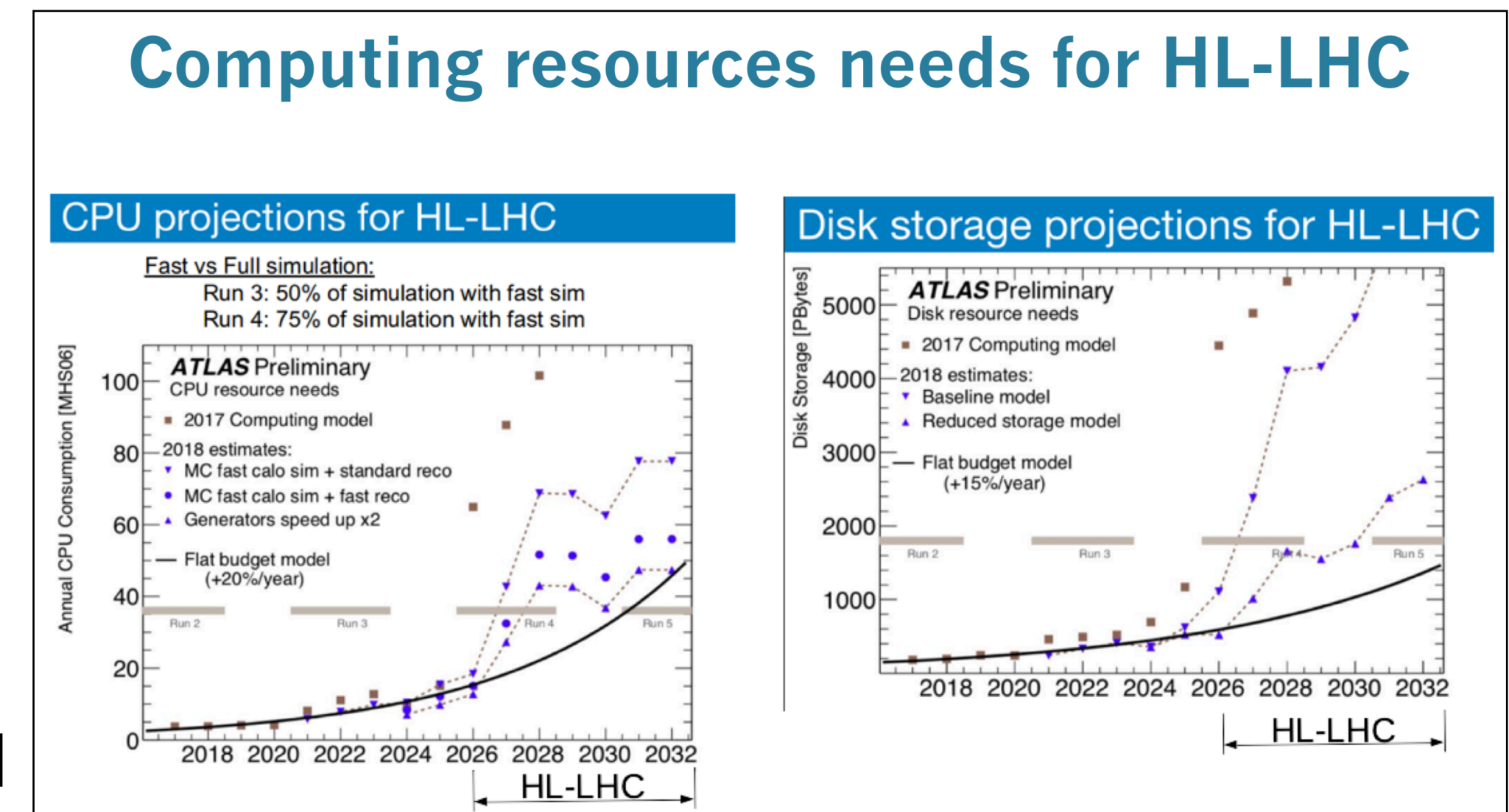
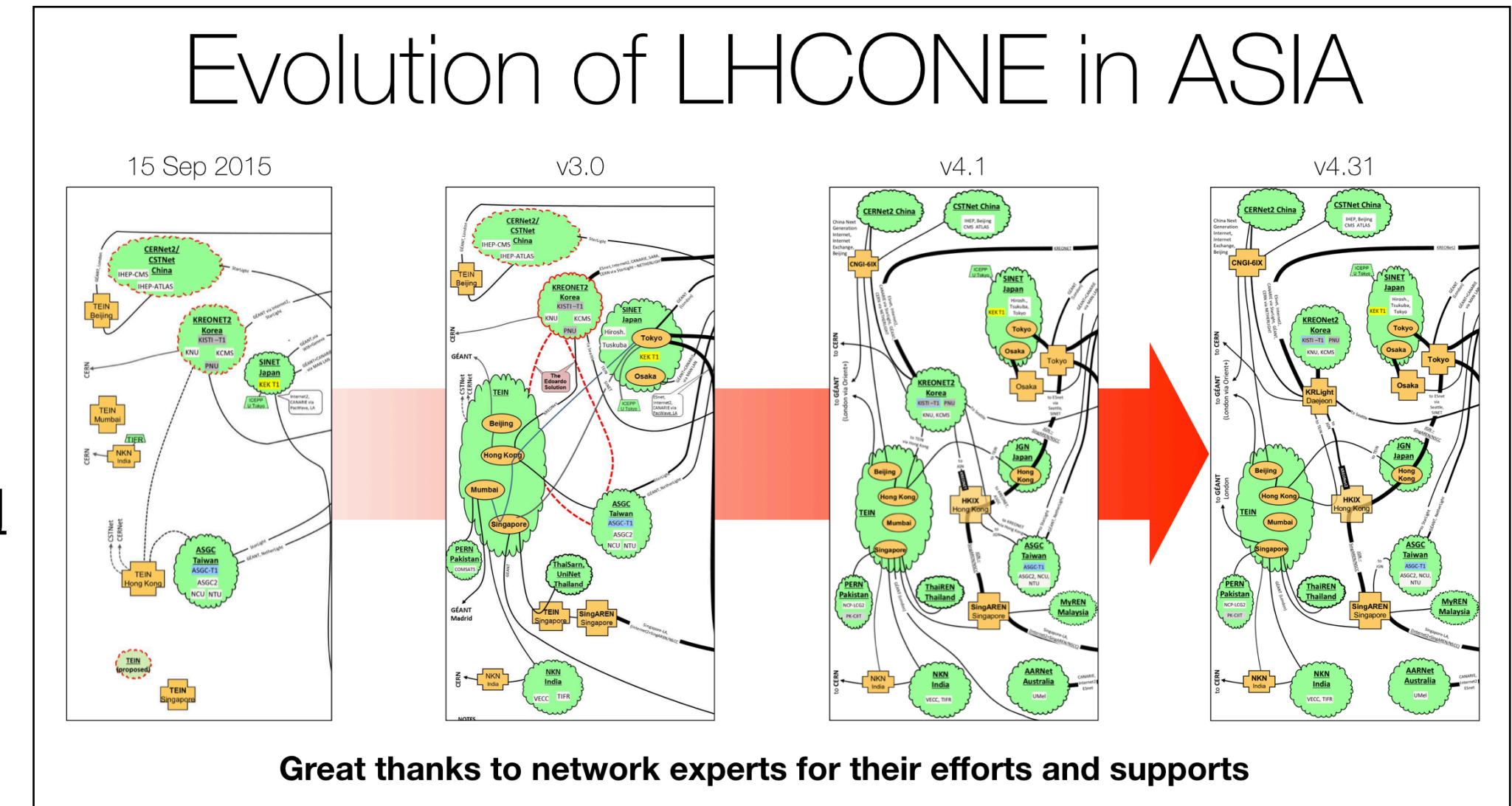


Korea Institute of  
Science and Technology Information



# Asia Tier Center Forum

- Started in 2015, 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](http://atcforum.org)
- Emerging agenda: distributed storage spanning the region
  - Tier can be blurred; network-driven disruptive paradigm change - nucleus-satellite model, caching, storage consolidation → 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





# Distributed Storage in Asia

- A strong collaboration is needed to overcome **Data Challenges** foreseen in HL-LHC
  - 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



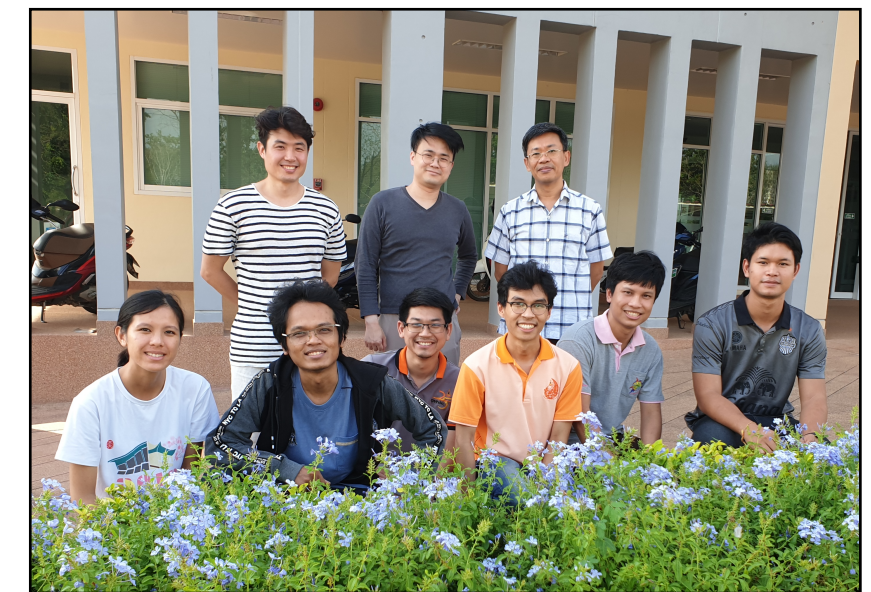
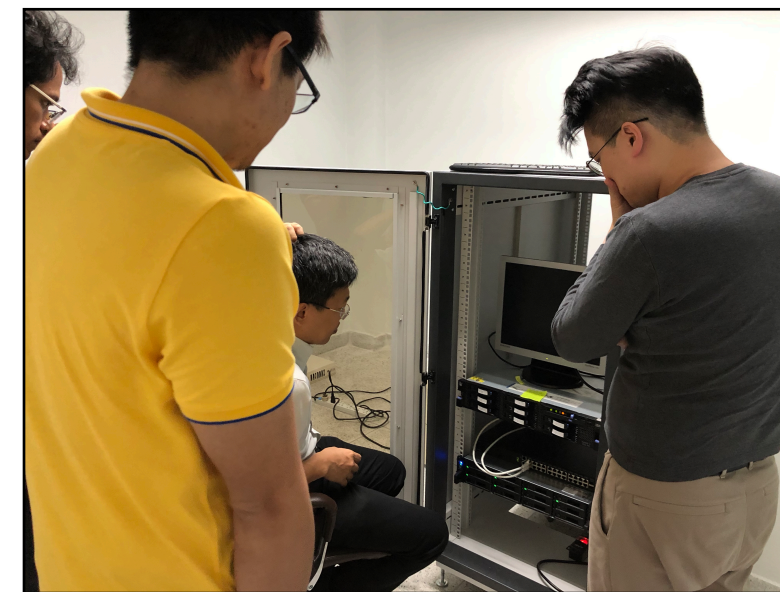
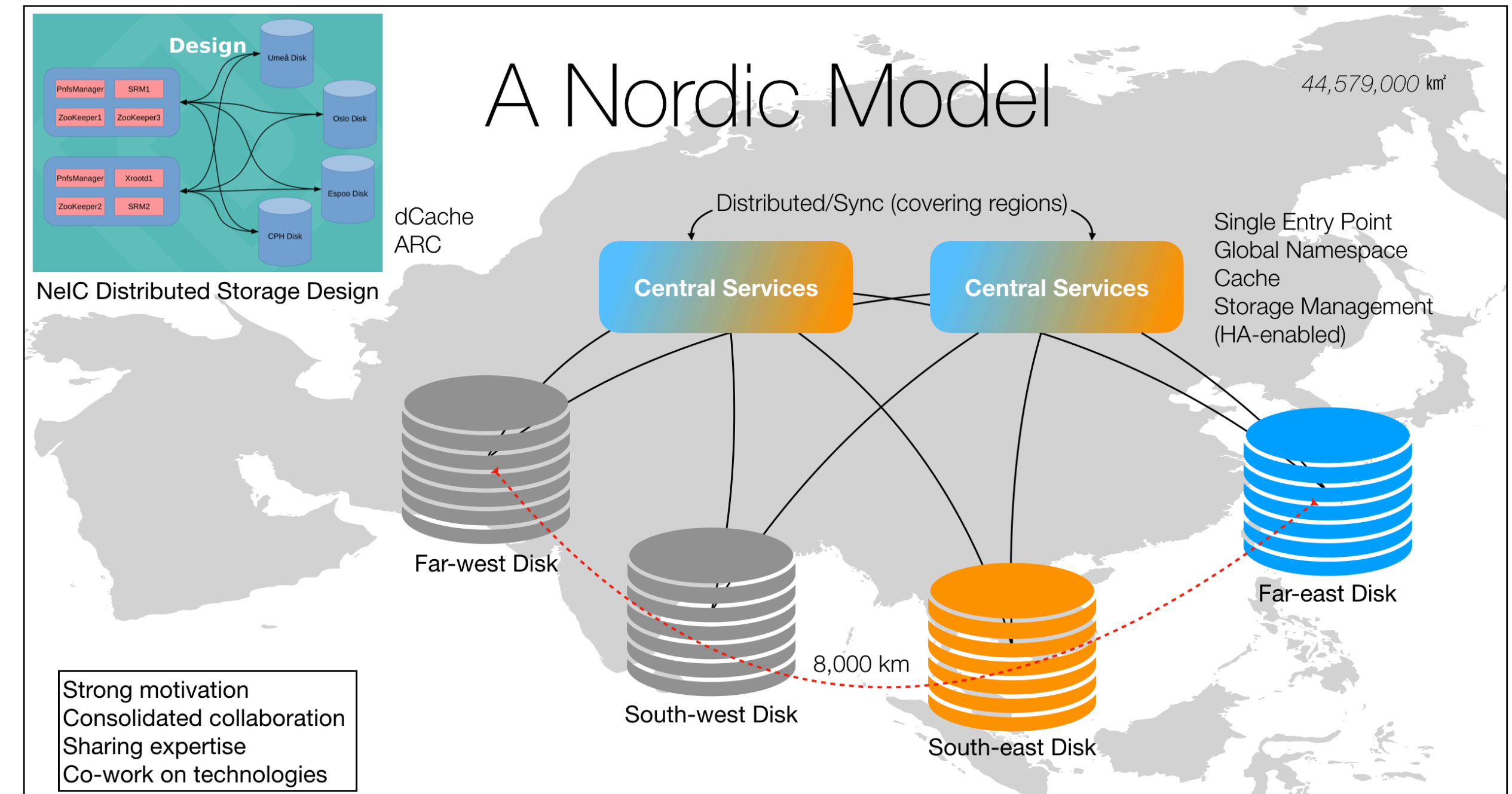
# Discussion

- 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



# 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





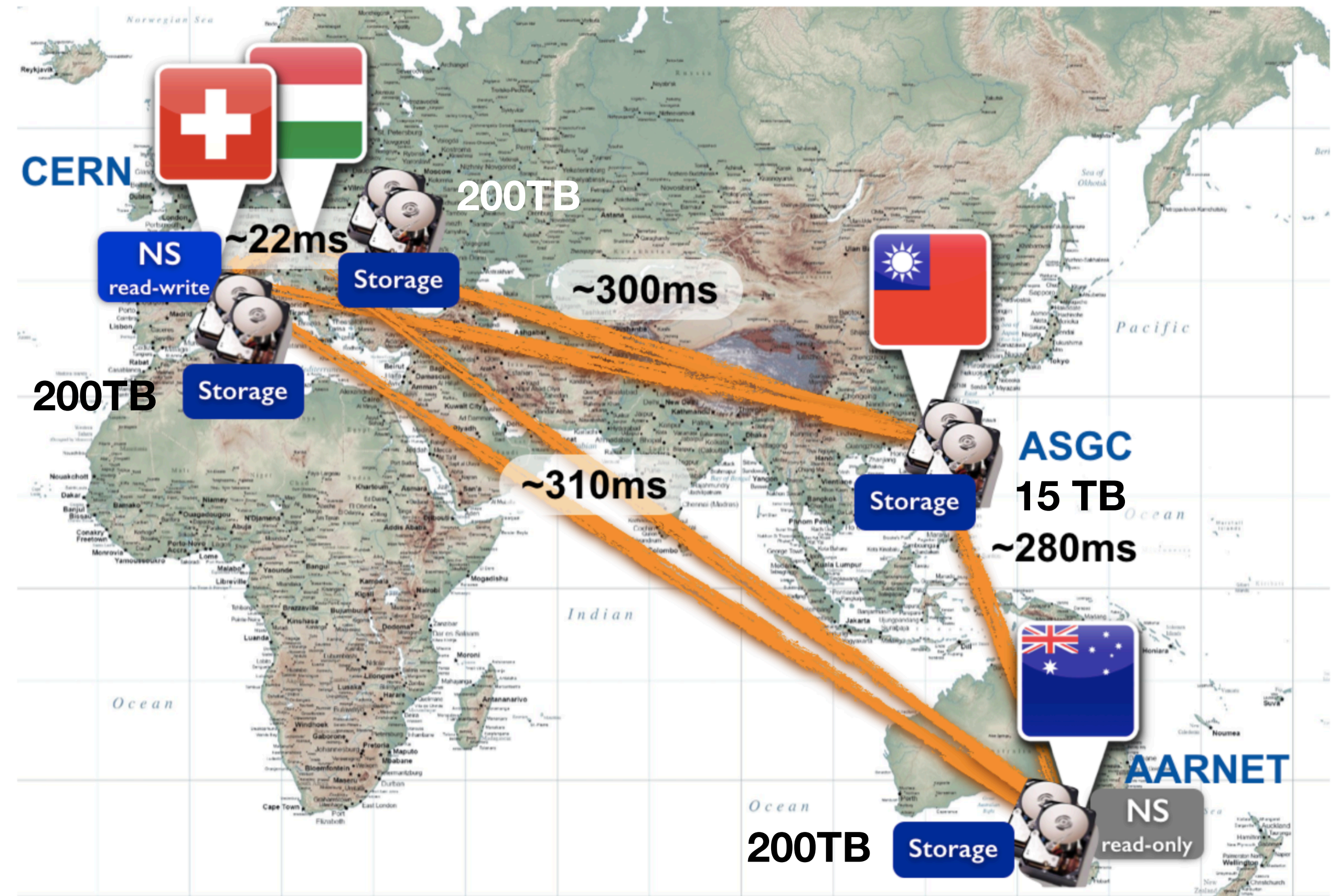
# Case Study

- 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" 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)

24/7 NOC SUPPORT 1305 275 682



NETWORK & SERVICES

CASE STUDIES

NEWS

COMMUNITIES

ABOUT US

CloudStor

A file sharing and cloud storage solution for the research and education sector

NETWORK & SERVICES

OUR NETWORK

HOW TO CONNECT

CONNECTIVITY SERVICES

COLLABORATION SERVICES

CLOUD SERVICES

DATA SERVICES

ENTERPRISE SERVICES CONSULTING

SECURITY SERVICES

CLOUDSTOR FILE SHARING + STORAGE

The CloudStor platform makes collaborating and sharing files so easy for AARNET customers.

CloudStor removes the frustration of slow data transfer rates for very large files by providing a super-fast, easy-to-use and secure file transfer and storage solution hosted on the AARNET network.

Unlike most cloud storage services, CloudStor is designed to meet the specific needs of researchers, and one terabyte of storage is available free to each individual researcher at AARNET connected institutions.

LOGIN TO CLOUDSTOR

Why CloudStor?

- 1TB free storage for individual researchers at AARNET-connected institutions + group storage quotas for research projects.
- Quick and secure file transfer with no file size restrictions.
- Single sign on using home institution credentials (for Australian Access Federation members).
- CloudStor web interface for access to file storage, FileSender, AARNET Mirror and other applications.
- OnlyOffice for collaboratively editing a wide range of document types directly in the CloudStor portal.
- Run and write Jupyter notebooks to analyse data in CloudStor using the SWAN Service for Web-based Analysis.
- Storage located in Australia and directly connected to the AARNET backbone for rapid and convenient access, and avoiding any sovereignty issues.
- Data is replicated a minimum of three times at geographically distributed storage nodes for high reliability and availability.
- CloudStor uses EOS, the scalable back-end storage developed at CERN.
- Sync client is available for Windows, Mac, OSX, Linux, iOS and Android.
- Access Amazon and other cloud data stores remotely using WebDAV and S3.
- Upload data sets from scientific instruments with CloudStor Rocket upload tool.
- Works with institutional repositories and national merit-based storage.
- A sustainable service that AARNET plans to provide indefinitely.

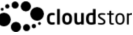
More information

Getting Started Guide

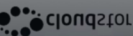
Frequently Asked Questions (FAQs)

Read CloudStor News

Contact Us



cloudstor



cloudstor

CloudStor File Sharing + Storage

CloudStor Mirror


CloudStor Rocket

CloudStor Sync

CloudStor Web Interface

CloudStor WebDAV

CloudStor OnlyOffice



- 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)

© 2019 AARNET

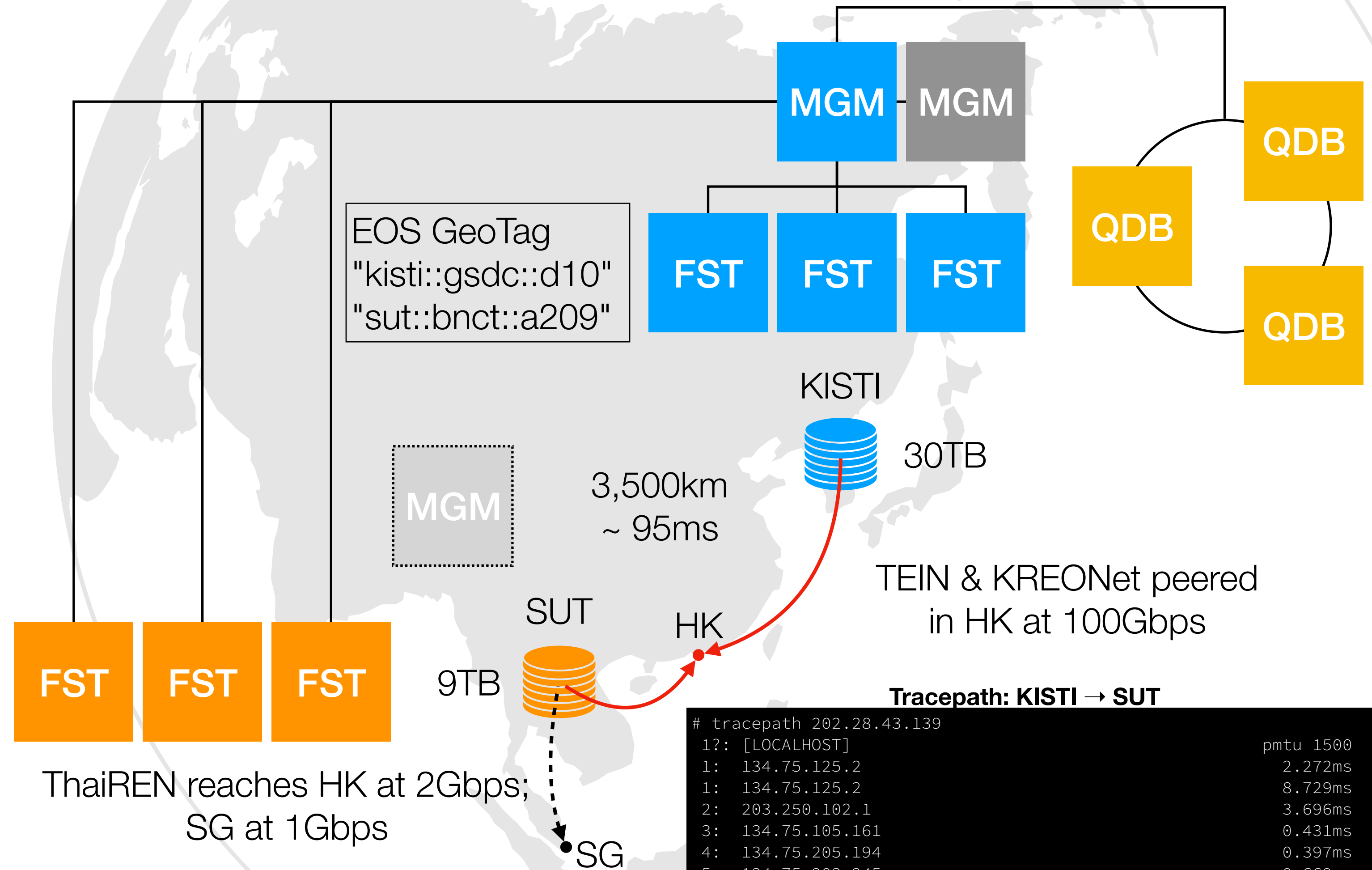
LEGALS

PRIVACY POLICY



# 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



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
```

[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



# 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: need to learn more on EOS



# Next Step

- Further investigation into,
  - Data transfer performance issue
  - Mixed authentication problem
- GSI authentication to be tested
- 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>



# Summary

- The pilot project on KISTI-SUT Distributed Storage based on EOS 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



# Thank you

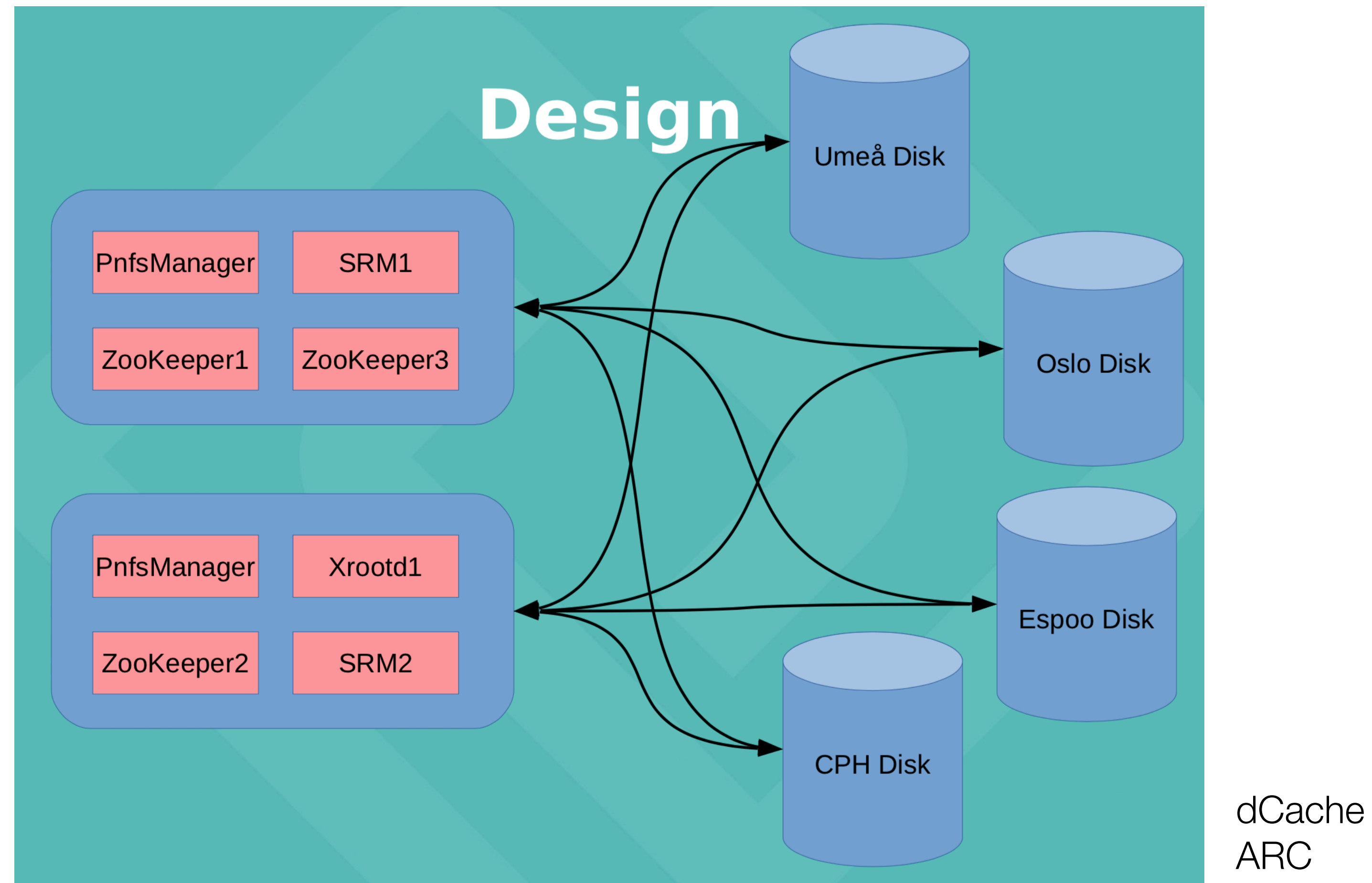
*Questions?*



Back Up



# The Nordic Model



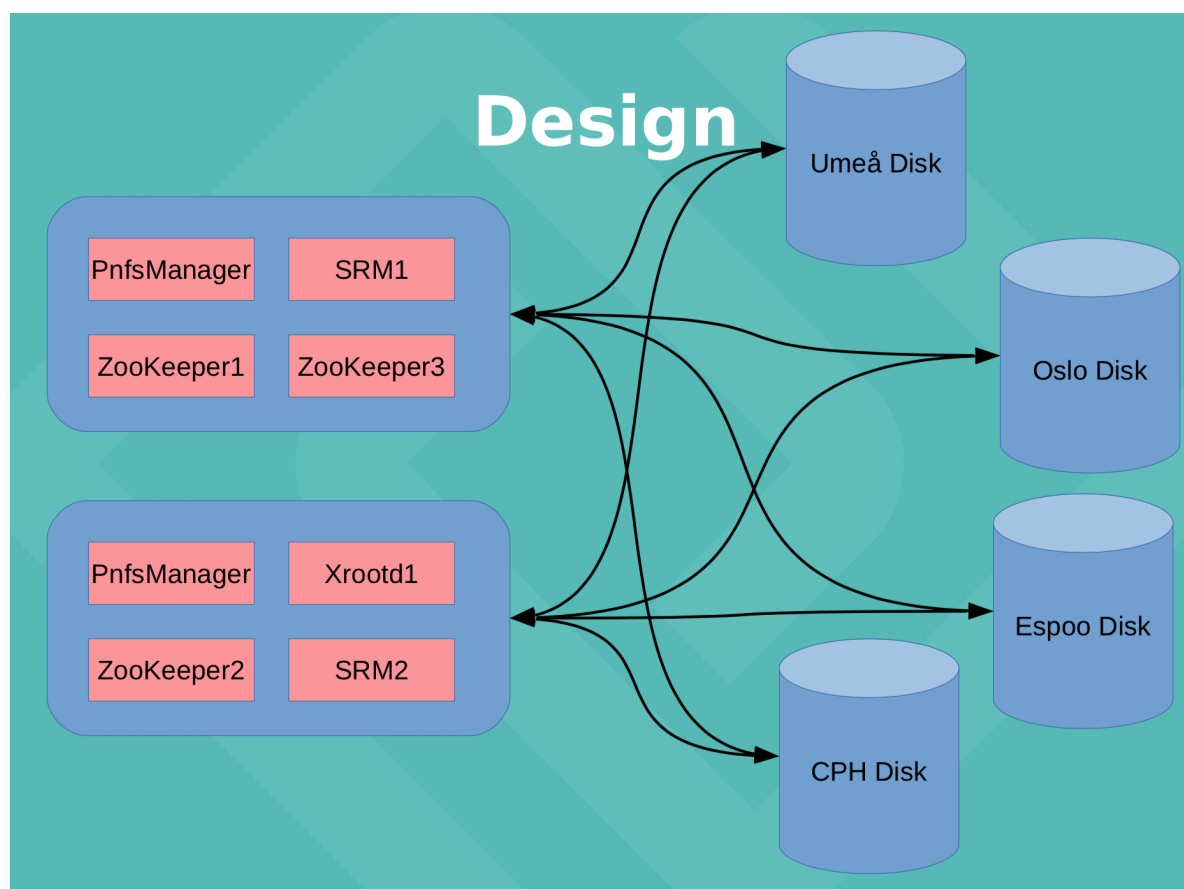
Strong motivation  
Consolidated collaboration  
Sharing expertise  
Co-work on technologies

NeIC Distributed Storage Design



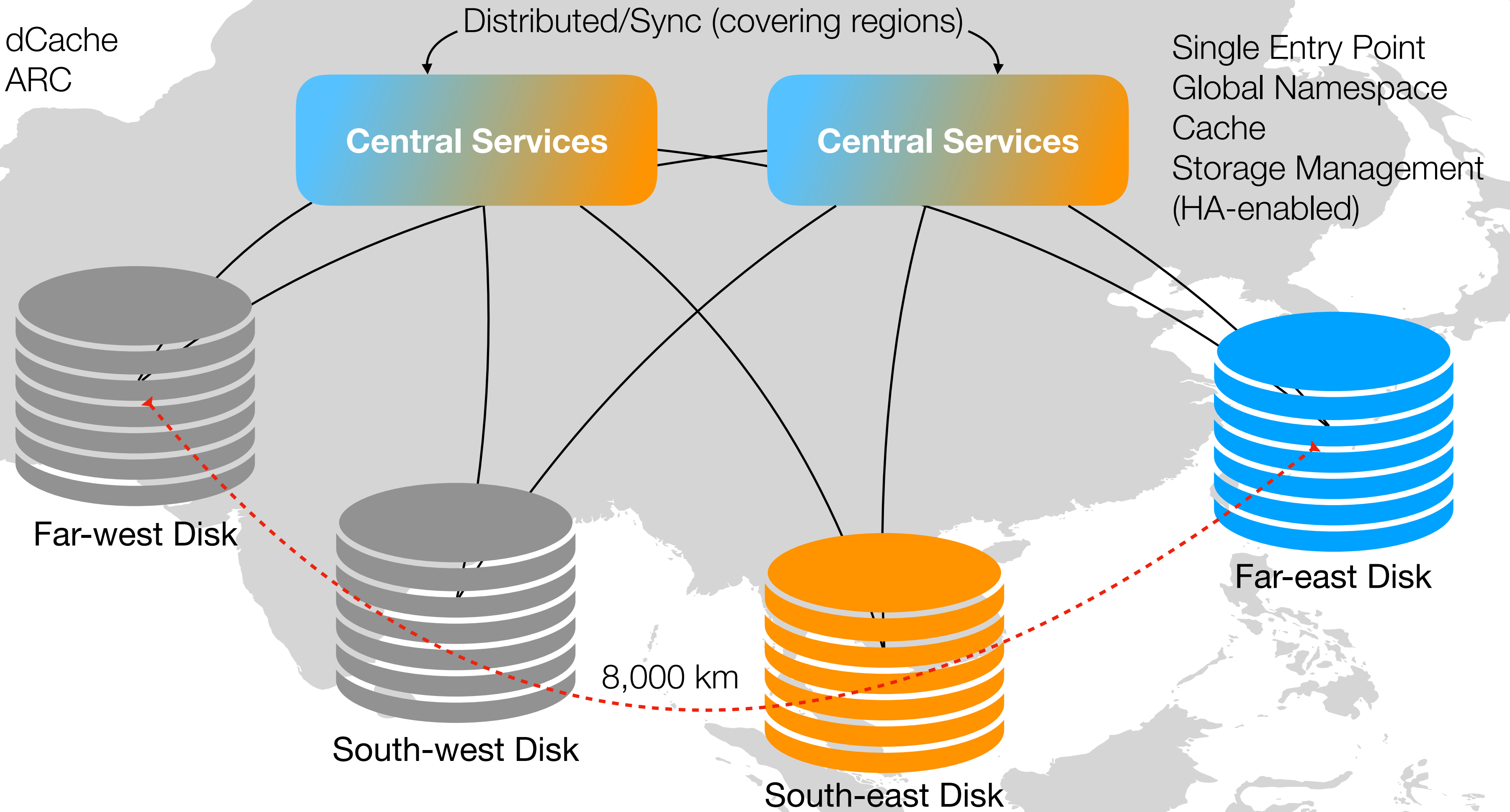
# A Nordic Model

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



NeIC Distributed Storage Design

dCache  
ARC



Strong motivation  
Consolidated collaboration  
Sharing expertise  
Co-work on technologies



# 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...



# 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"

