Inter-Cloud
Global Distributed File System

Ikuo Nakagawa

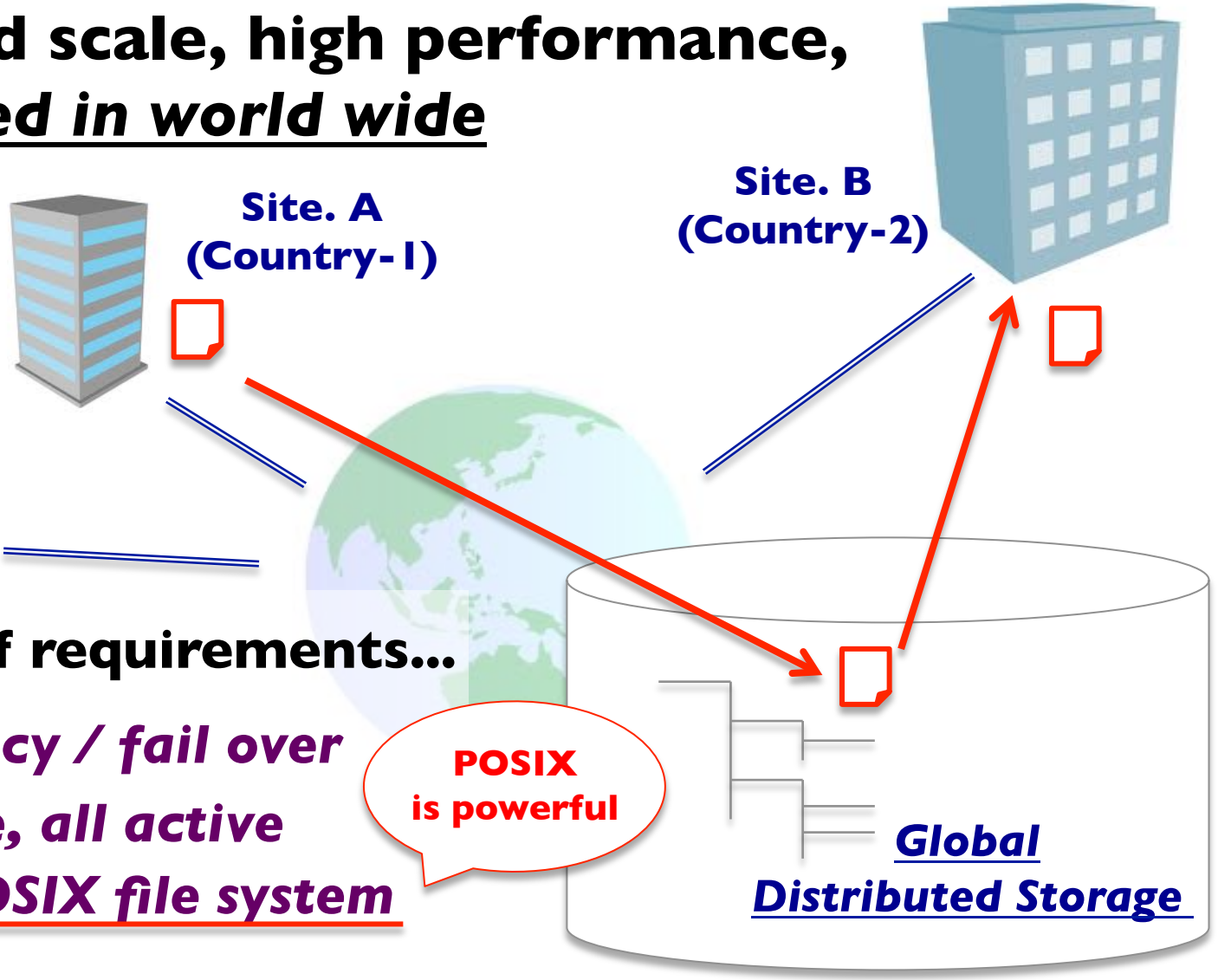
Distcloud Project, Osaka Univ., Intec, Inc.

April, 2015

Sharing data ?

Global Distributed Storage

unlimited scale, high performance,
distributed in world wide



with lots of requirements...

redundancy / fail over

multi-site, all active

native POSIX file system

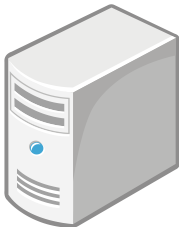
POSIX is powerful

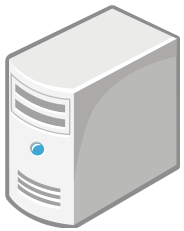
POSIX runs virtual machines

Long Distance Live Migration for Virtual Machines

Lab. / University (ex. Japan)

Workshop (ex. USA)



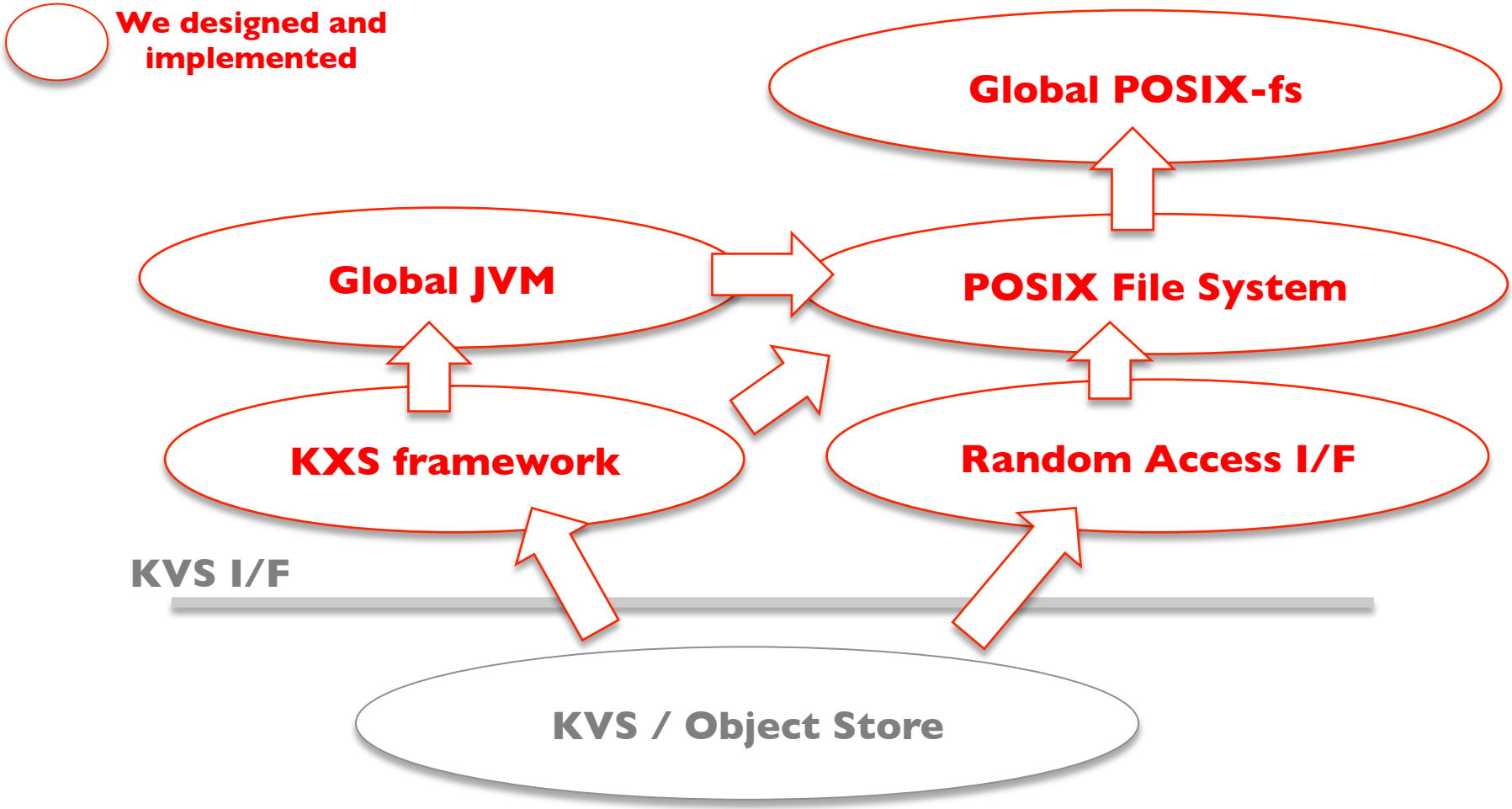


over Global Distributed File System

So, we designed ...

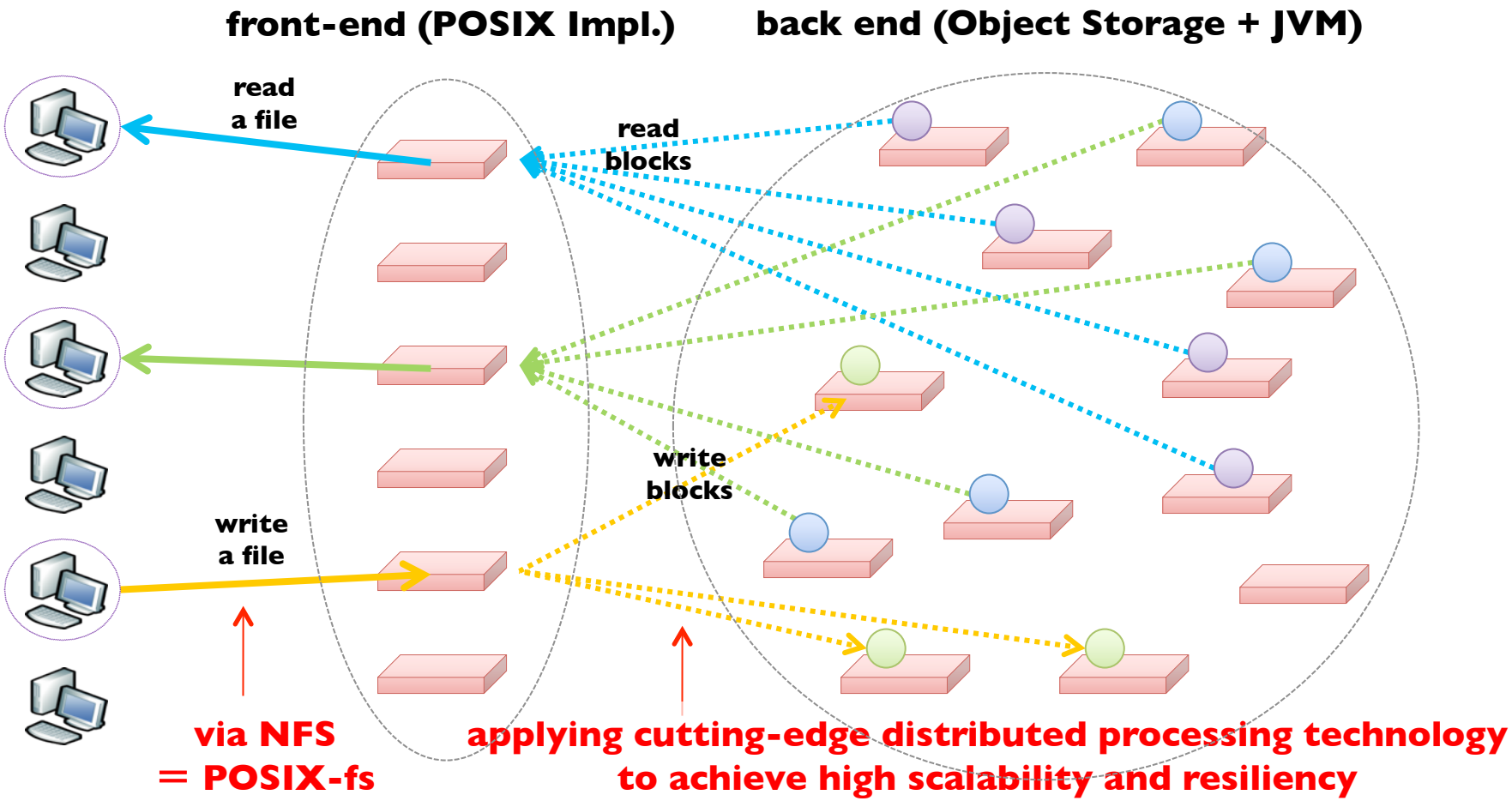
On the top of { KVS / Object Store }

Fully distributed systems on the top of Distributed KVS / Object Store



Distributed POSIX file system

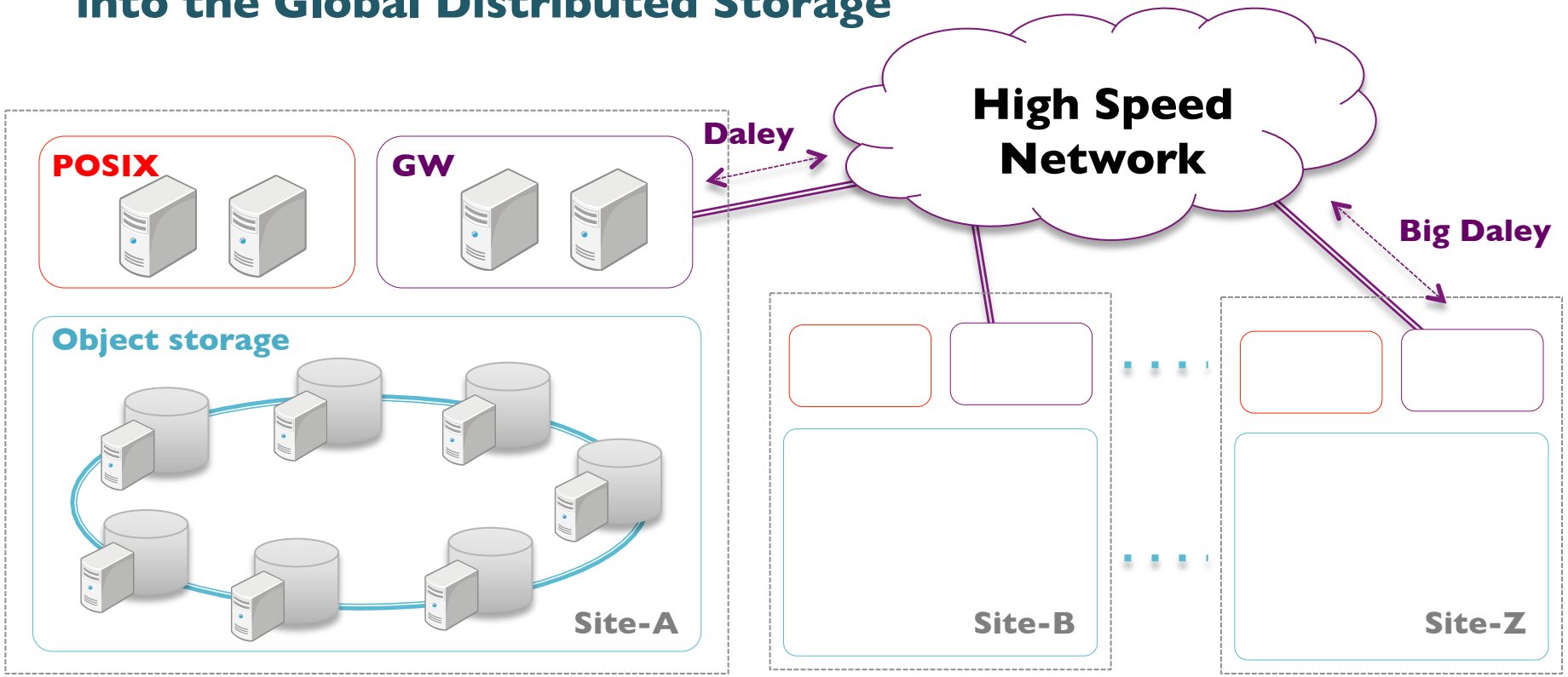
POSIX file system on the top of Scale-out Object Storage



Inter-Cloud & Global distribution

We achieved inter-cloud & global POSIX file system,
Fully distributed, No RDB, No centralized system

Aggregating independent sites via Net,
into the Global Distributed Storage



Global Distributed File System, means

Features as Distribute File System

"POSIX = Native file system"
standard, simple and user-friendly (via NFS)

"Cluster storage"
Scale-out, replication & failover and automatic recovery

"Full Distributed"
No-RDB, no-SPOF, distributing both metadata & userdata

Global Distribution

"Global Distribution"
Multiple site (N > 2) and All Active

Technical Challenges

Light speed ...

Challenges and dilemmas...

Big Latency (because of light speed)

Consistency?
(strict consistency, for POSIX)

Synchronous storage
up to 100Km or 200Km

Storage Live migration
Copy storage image before live migration

Post copy
Copy after live migration

Performance?
(distribution & scale-out)

Amazon S3 / REST based
Object, NOT POSIX

GFS / HDFS
Appendable, NOT POSIX

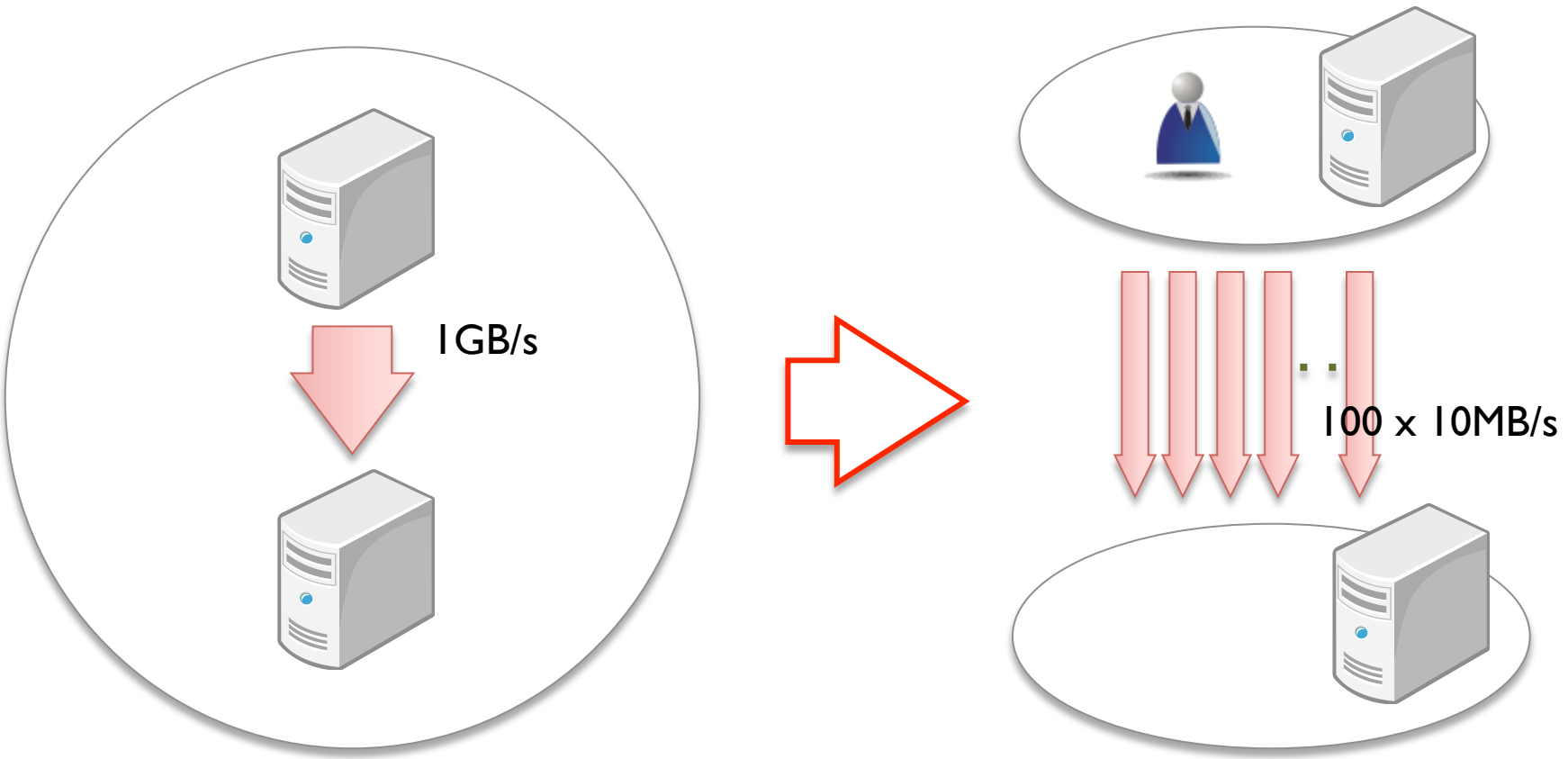
Cloudian, Riak, etc...
S3 compliant, NOT POSIX

We want both Consistency and Performance !

Improve throughput

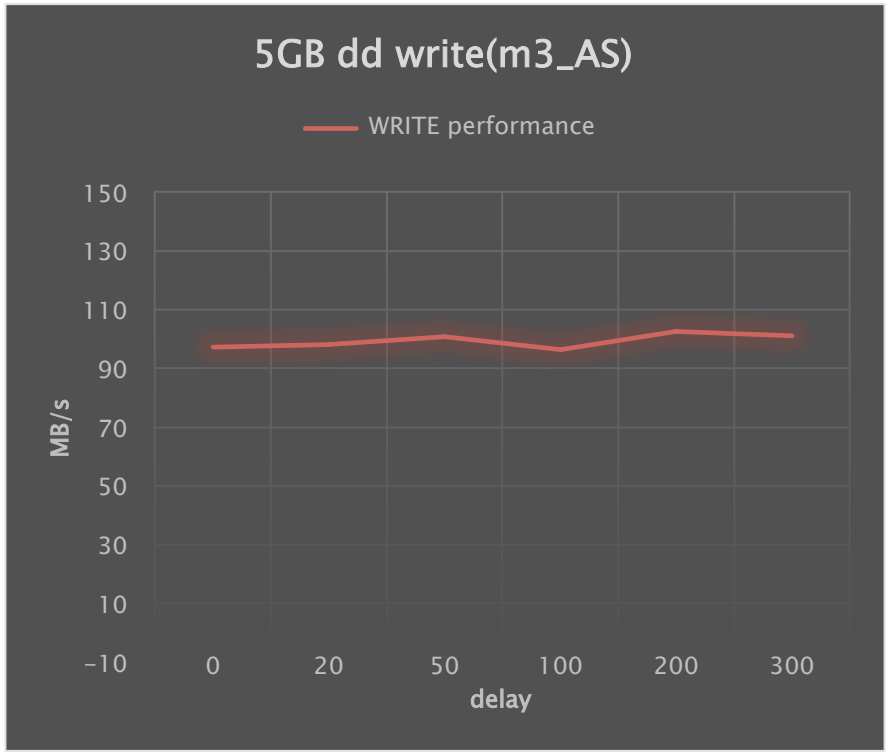
Parallelization improves "throughput"

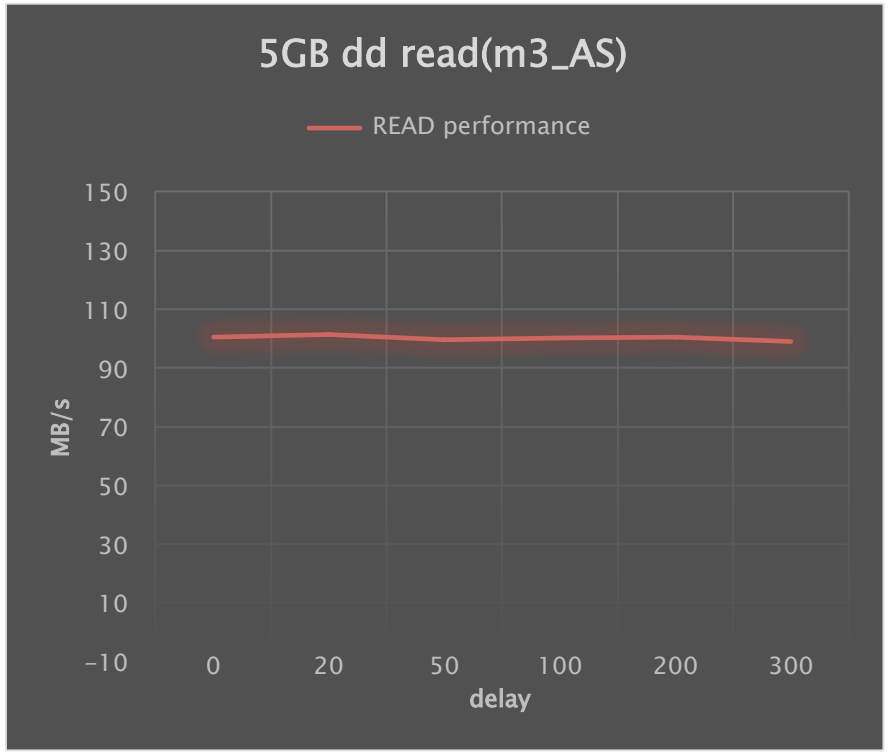
Just increase # of sessions between areas, with buffering for write / pre-fetching for read



Read/Write throughput in the lab.

Very stable performance for both R/W even if there is *300ms RTT !*

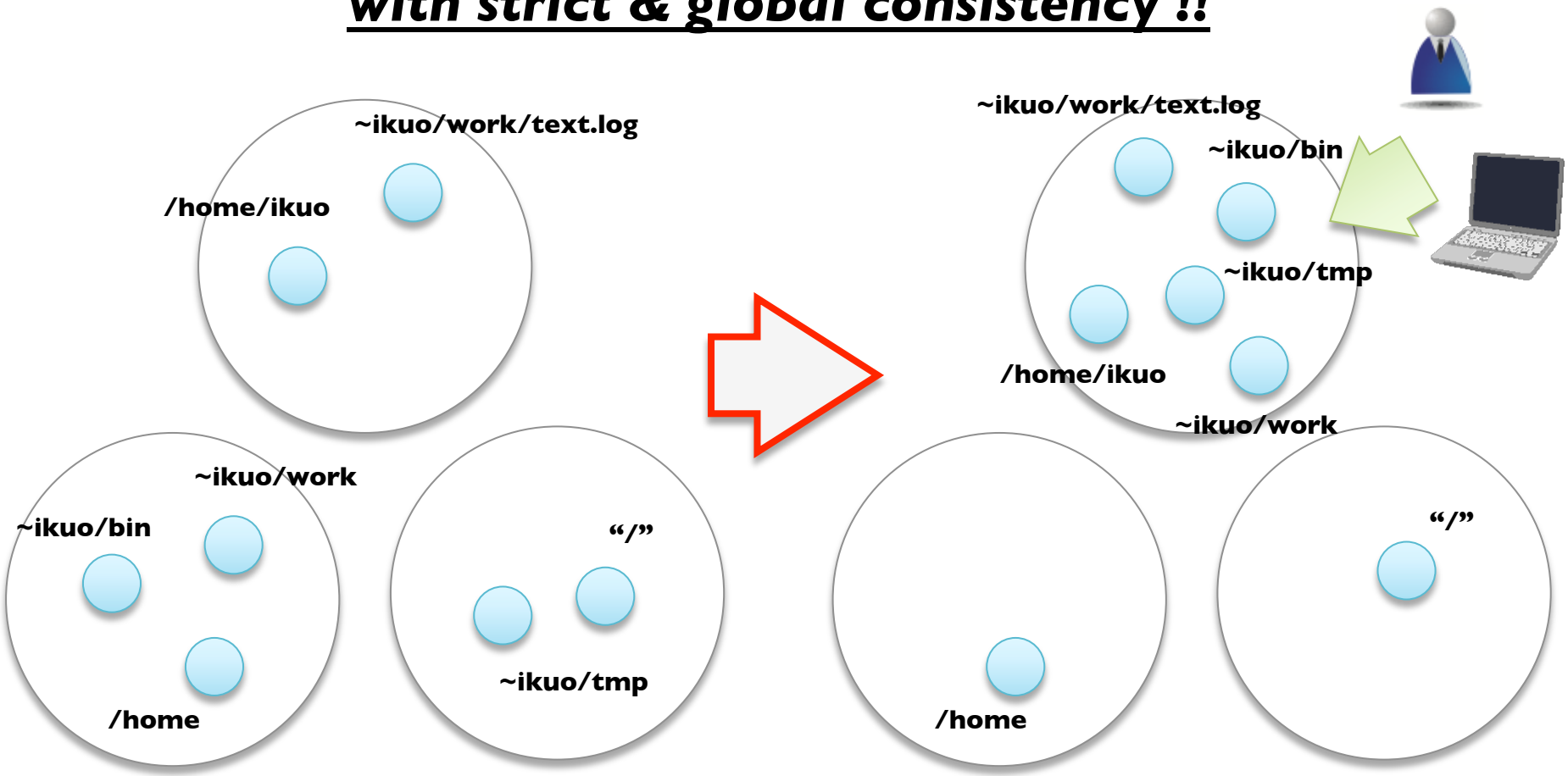




Lower latency

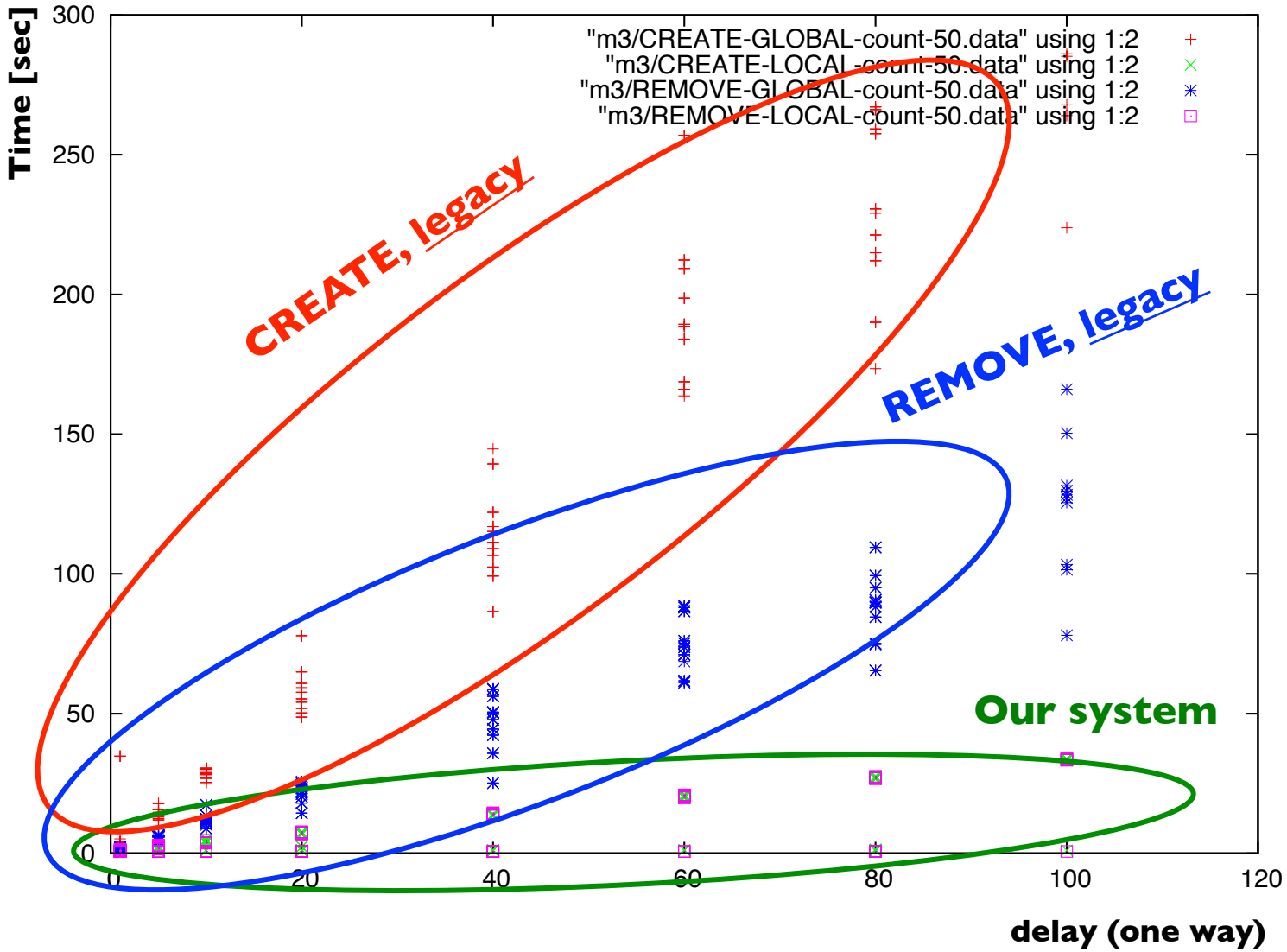
Big challenge against “latency”

Move metadata into the nearest location,
with strict & global consistency !!



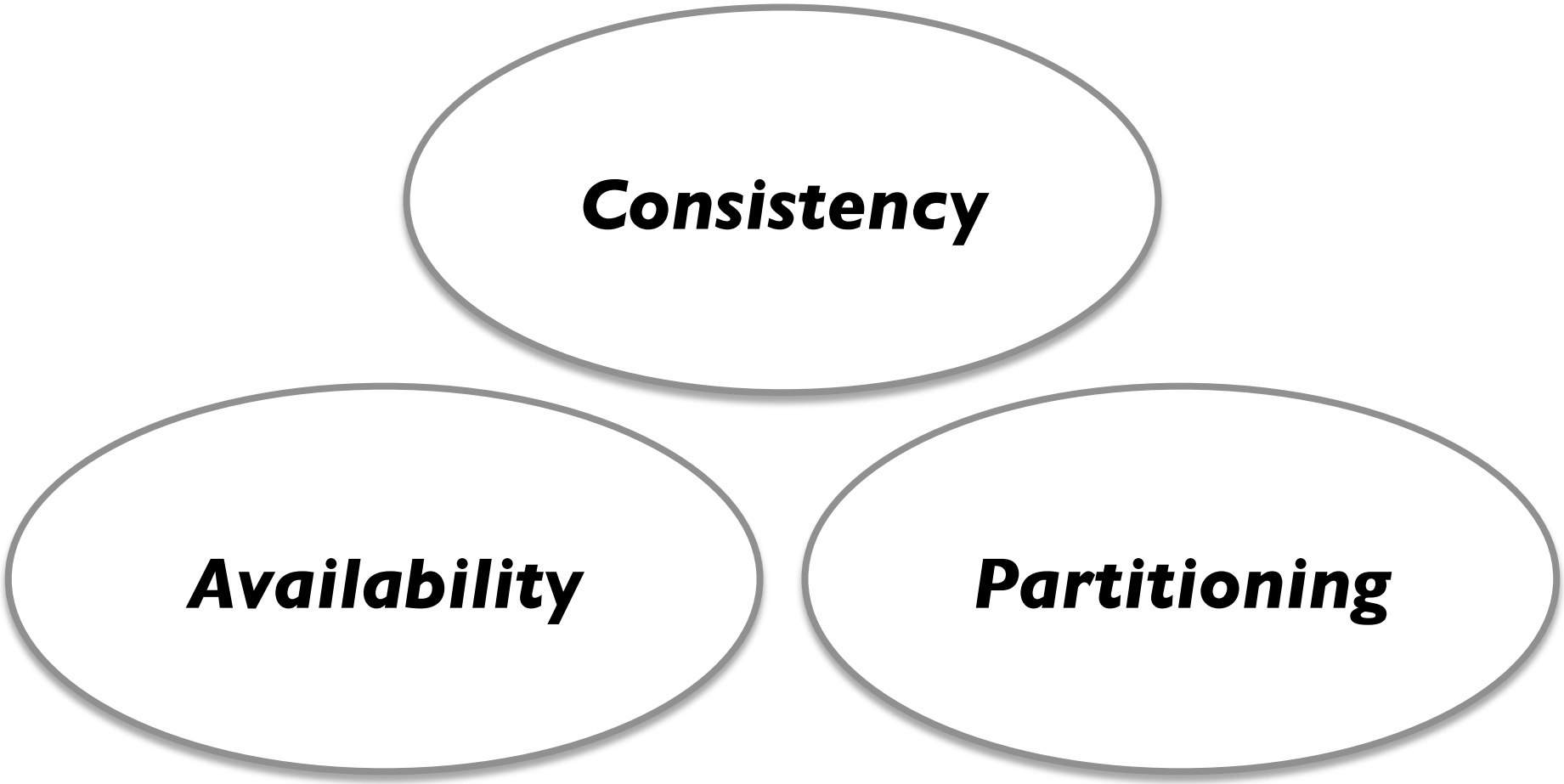
in lock-free and non-blocking model

Latency for metadata operations in the lab.



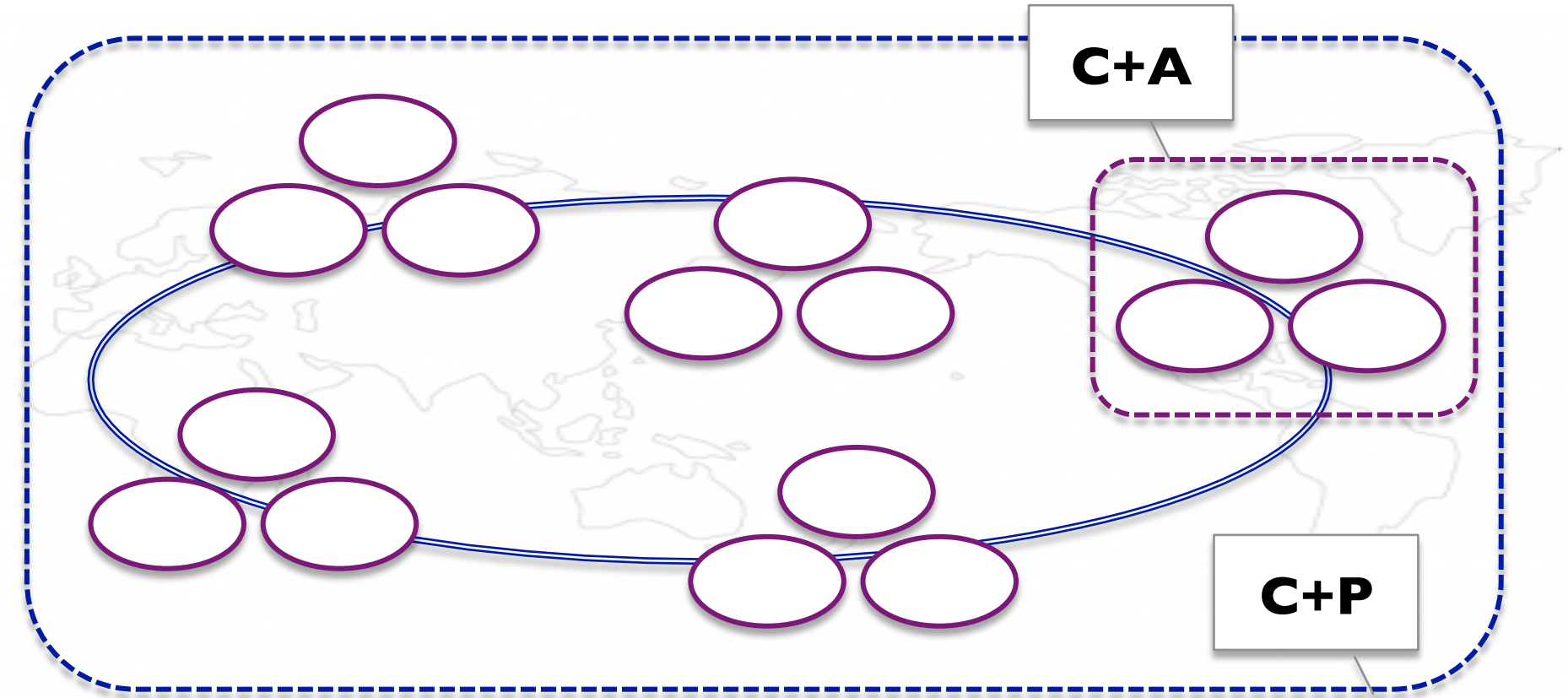
Design concept ?

Design concept – Challenge for CAP theory



Consistency + Availability + Partitioning = ∅

Design concept – Our solution is...

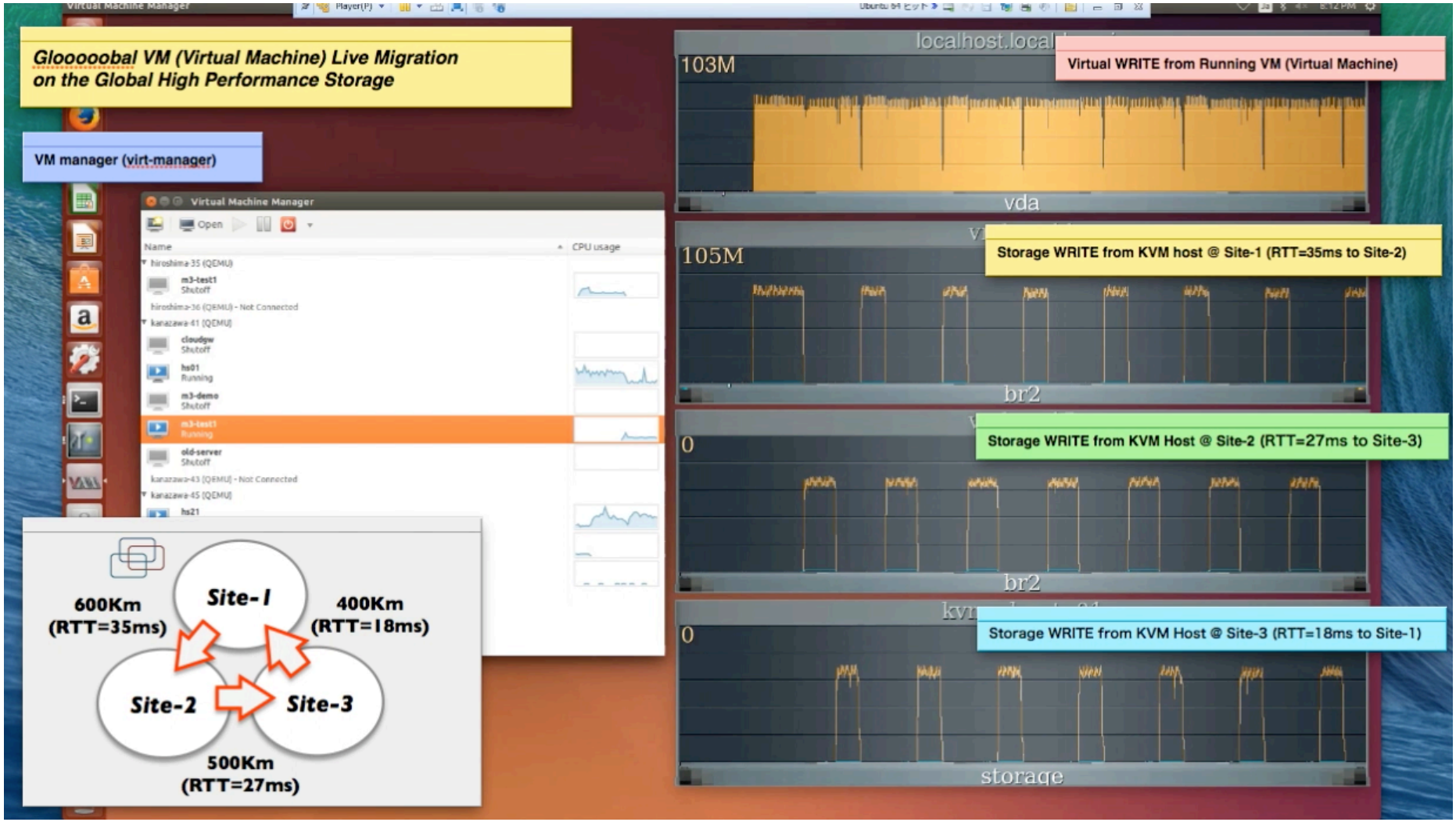


Consistency + (Availability + Partitioning) / 2

intra-continent **inter-continent**

and, experiment

VM Live migration among 3 sites



We are ...

Distcloud Project (Chair: Prof. Shimojo)

experimental consortium by universities, research entities, etc.

Hiroshima Univ.

JAIST Japan Advanced Institute of Science and Technology

Kanazawa Univ.

Kyoto Univ.

UC San Diego

UC San Diego

Osaka Univ.

National Institute of Information and Communication Technology

Intec, Inc.

INTEC IT Holdings Group

Kyushu Univ.

Nara Institute of Science and Technology

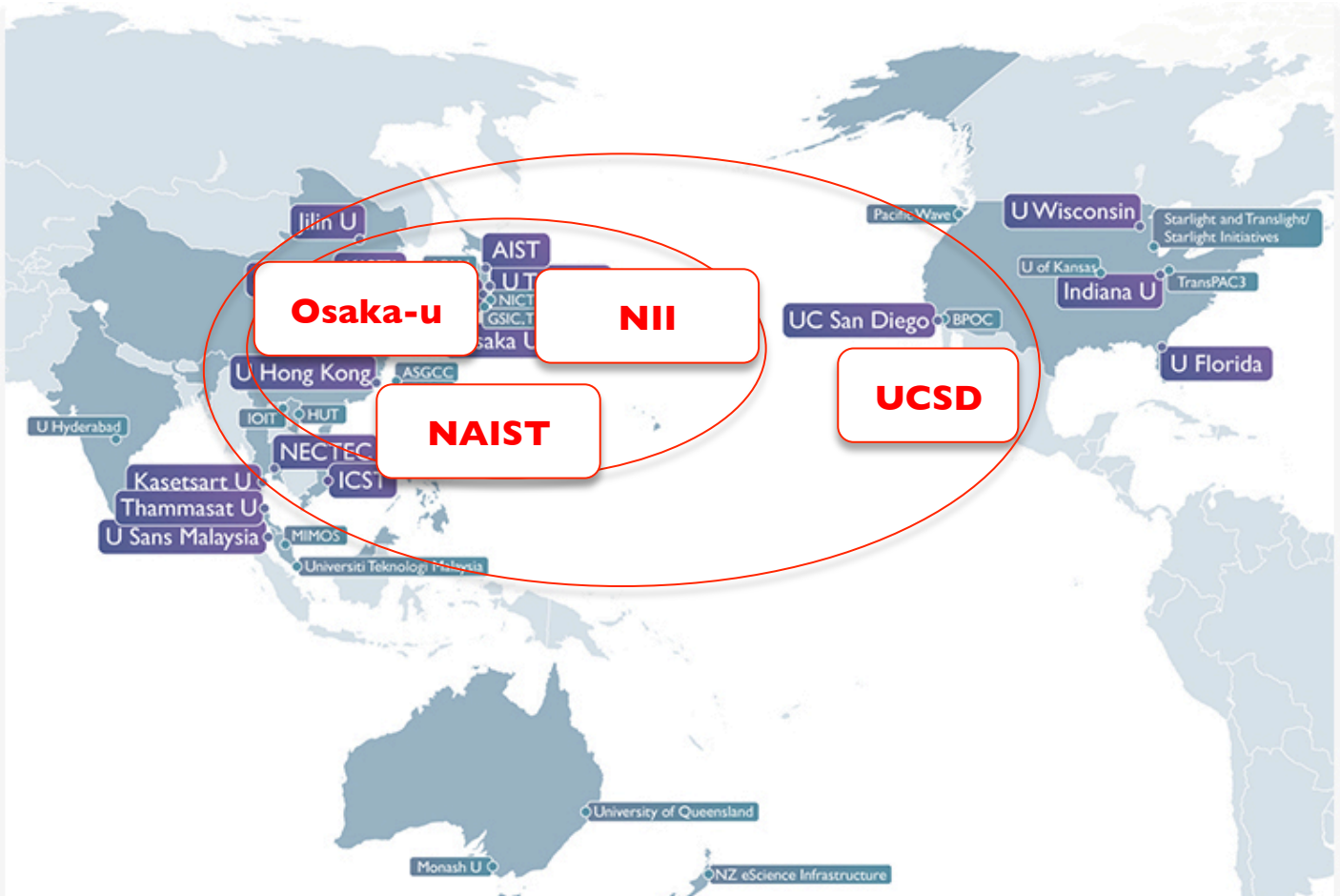
Kyushu Institute of Technology

Kochi Institute of Technology

National Institute of Informatices

Distcloud Project Members

for PRAGMA-ENT...



and, we'd like to propose our file system as an experimental file system for PRAGMA-ENT

Thank you !!

Please feel free to contact to

Ikuo Nakagawa