SDC

STORAGE DEVELOPER CONFERENCE

SNIA ■ SANTA CLARA, 2017

DATEstor: Highly-available Metro Area Distributed Storage Systems

Takaki Nakamura, Ph.D.

Hitachi, Ltd. / SNIA Japan

Agenda

- Introduction
- Overview of DATEstor
- Evaluation on DATEstor
- Conclusions

Agenda

- Introduction
- Overview of DATEstor
- Evaluation on DATEstor
- Conclusions


Great East Japan Earthquake
Mar. 11th, 2011
Magnitude 9.0-9.1
15,000+ deaths
1,000,000+ buildings damaged
Level 7 meltdowns in 3 nuclear reactors
Image provided by Sendai city.

Background

- Information Services at times of disasters
  - Resident registries to identify whether residents are safe or not.
  - Medical histories to sustain their health.

- In the case of the Great East Japan Earthquake and Tsunami of 2011
  - Network connections from/to the disaster area were lost
  - Therefore, data at remote sites was inaccessible from the disaster area.

NTT Onagawa network station damaged by the earthquake (Dec. 2012)

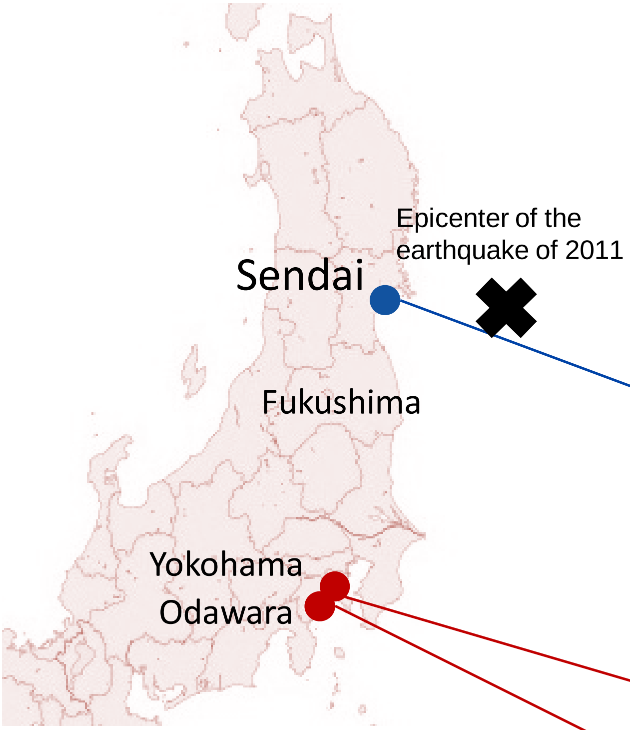
Objective

- To develop a Disaster Resilient Storage System
  - To sustain information service immediately after such a serious disaster

- We had been engaged in this work as Japan National Project from Sep. 2012 to Mar. 2017.

Project Team Members (until Mar. 2017)

Project Leader：
Tohoku univ., RIEC Prof. Hiroaki MURAOKA



Organization	Name
Tohoku univ., RIEC	Takaki NAKAMURA
Hitachi, R&D Group	Shinya MATSUMOTO Hitoshi KAMEI
Hitachi Solutions East Japan	

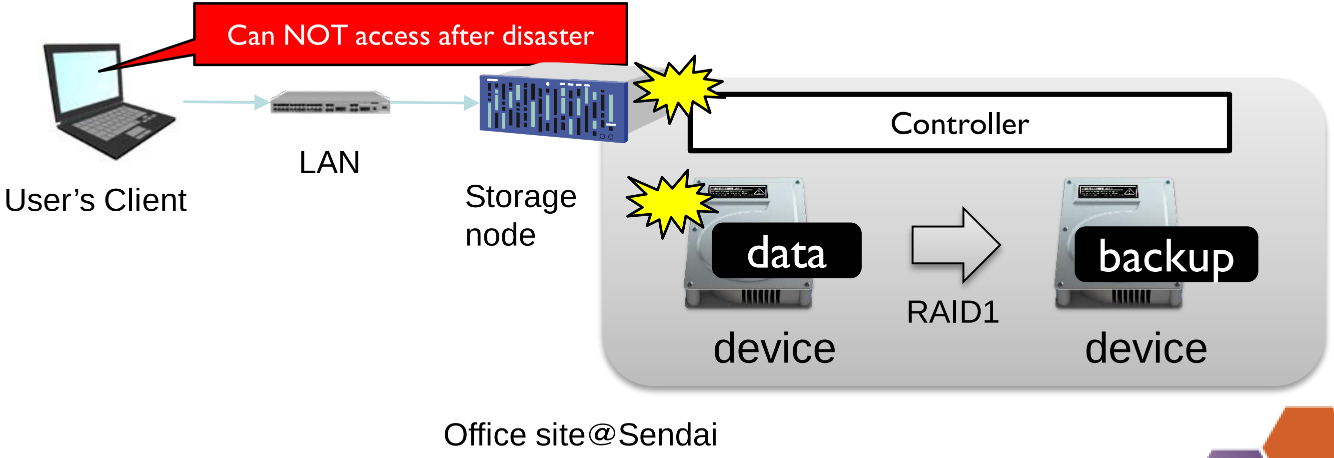
Organization
Hitachi, R&D Group
Hitachi, IT Platform Division Group

Case studies of Great East Japan Earthquake of 2011

- Site/Building (which may have a storage apparatus)
  - Half of total number of buildings were damaged in seriously damaged cities.
    - such as Rikuzen-takata city.
- Network
  - Wide area network (Internet) connections were unavailable for up to 1 month.
  - However, a part of local area network was still available.
- Power
  - Blackouts occurred in many places right after the earthquake.
  - After a few days, power supply recovered in almost all places.

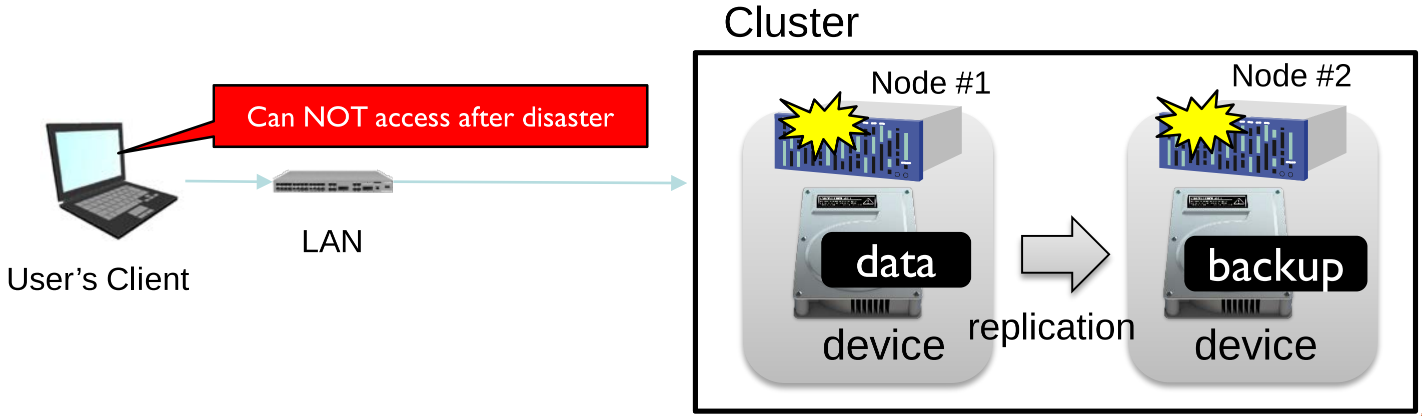
Existing Highly-Available Storage Systems

- (1) RAID / Erasure Code
  - Data has a redundancy among multiple storage devices in a storage node
  - Data is available unless the number of damaged devices is beyond a redundant value.
  - However when the storage node controller is damaged, data becomes unavailable.



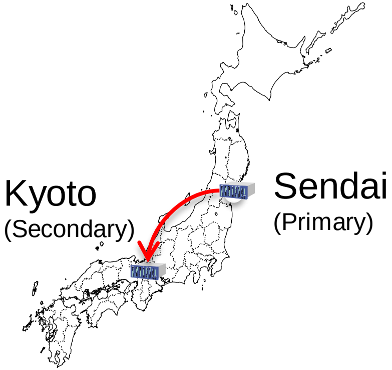
Existing Highly-Available Storage Systems (cont'd)

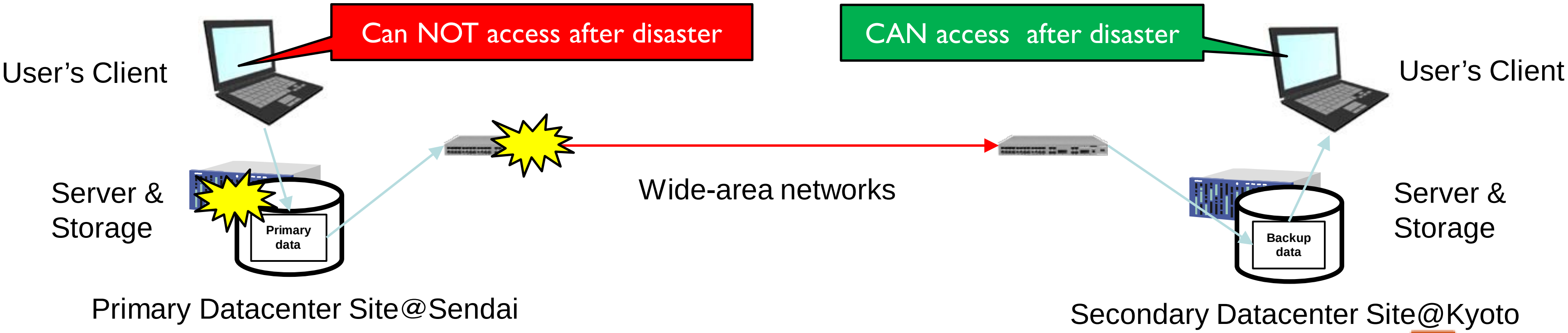
- (2) Local Replication (e.g. rsync, robocopy)
  - Data has a redundancy among multiple storage nodes in a cluster.
  - Data is accessible unless the number of damaged nodes is beyond a redundant value.
  - As the nodes are generally installed in the same or a nearby rack, many nodes may be damaged at the same time by a disaster



Existing Highly-Available Storage Systems (cont'd)

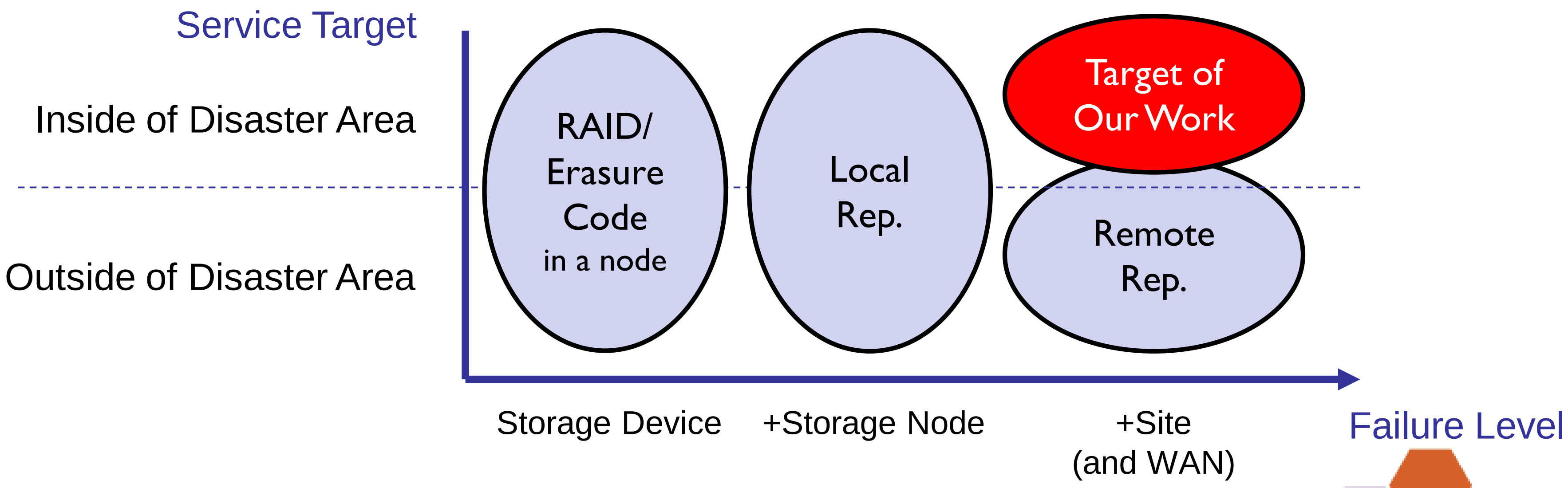
- (3) Remote Replication (e.g. SnapMirror, Cassandra)
  - Data on the primary site is replicated into the storage on the secondary site.
  - The secondary site takes over the services once a disaster occurs.
  - This combination is called "Disaster Recovery" feature.





Target of Our Work

Continue providing information to the inside of disaster area even if both wide area network and storage nodes are damaged.

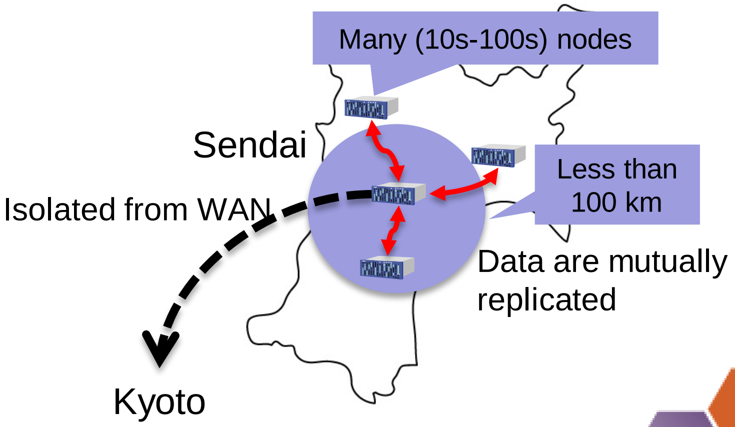


Approach: Metro Area Distributed Storage

- Replicate data at a primary site to nearby sites in addition to a distant site.
  - Accessible by metro/local area network or by physical means even if isolated from WAN.
  - Data is mutually replicated to **<u>many low-end storage nodes</u>** in the metro area
- Two Key Features: Risk-aware Data Replication and Multi Route Restoration

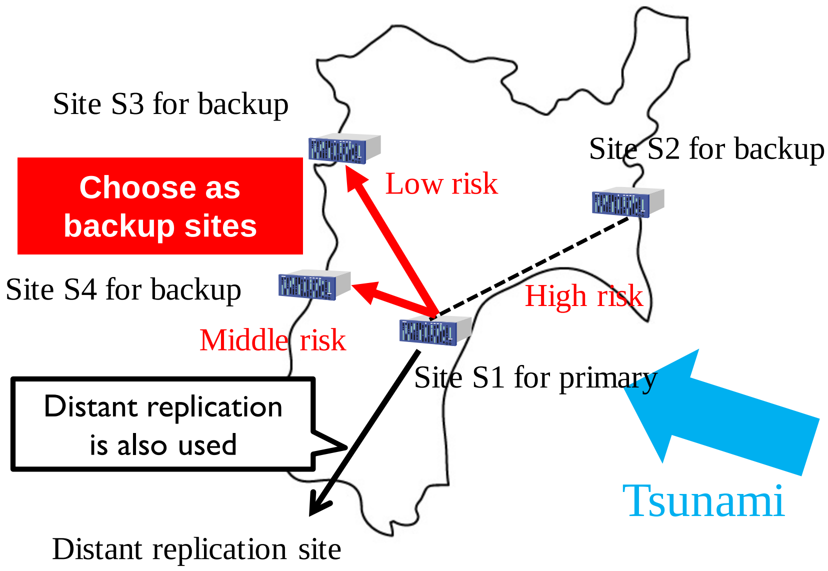
Existing Approach (Distant Replication)

Proposed Approach (Distant + Nearby Replication)



Risk-aware Data Replication (RDR)

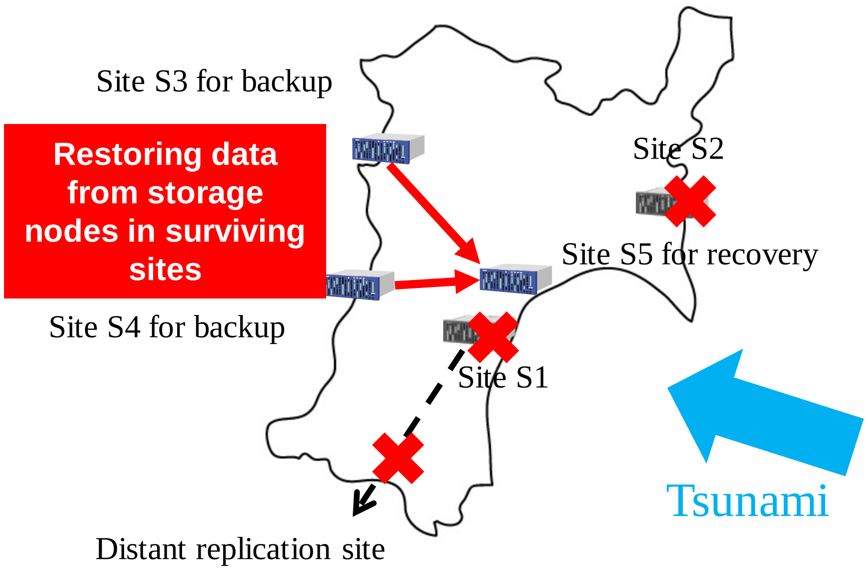
- Replicate data to a safe and a nearby backup site in the metro area
  - Quantify a risk indicator of site-pair based on geographical conditions
  - Choose a backup site with the low risk indicator in the metro area
  - Replicate data to the selected backup site
  - Especially for **<u>many (10s -100s) primary sites</u>**, a method for automatic selection is required



Conceptual diagram of risk-aware data replication feature
(The number of replica: 2 for nearby +1 for distant)

Multi Route Restoration (MRR)

- Data is restored from storage nodes at surviving sites simultaneously in response to a damaged situation.
- Even when the metro/local networks are also disrupted, data is accessible via operations as follows:
  - Transporting the surviving storage node to an area where the service is required.
  - Approaching the surviving storage node



Conceptual diagram of multi-route restoration feature

Use Cases of Proposed Architecture

- Local government services
  - Storage nodes are installed in city offices (including branch offices)
  - Residents information is replicated in a metro area.

- Medical institutions services
  - Storage nodes are installed in hospitals, clinics, and pharmacies.
  - Medical information is replicated in a metro area.

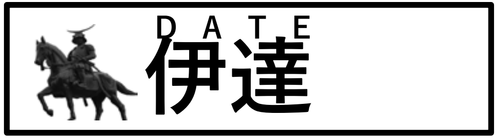
Related Work

- Metro Area Distributed Storage
  - Generally, the relation of the existing remote replication is one-to-one (or a few like three data centers). We approach many-to-many (beyond 100 nodes).
- Risk-aware Data Replication
  - Cassandra has replication policies such as “Rack-aware” and “Datacenter-aware”. Availability zone is also a similar idea. These are kinds of “Risk-aware”. We approach not qualitative (0 or 1) but quantitative policies.
- Multi Route Restoration
  - Basic idea is the same as parallel download technologies such as GridFTP. MRR uses both replicating data and erasure coding data as the data type.
    - Replication for Metadata and small sized file data
    - Erasure code for large sized file data

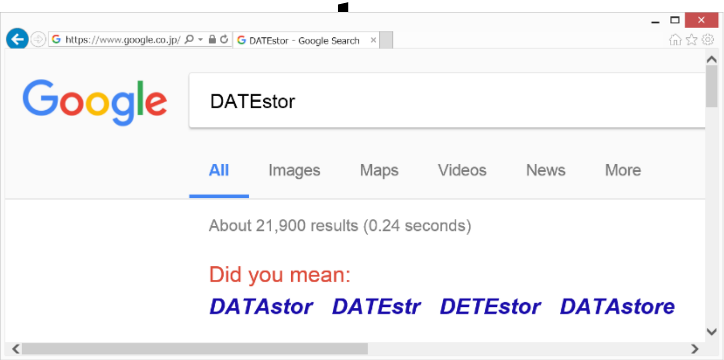
Agenda

- Introduction
- Overview of DATEstor
- Evaluation on DATEstor
- Conclusions

What is DATEstor?



- "DATEstor" is NOT a typo.
- Pronunciation is not [déit] but [date]
- It stands for Disaster-resilient, Autonomous, Tactical, and Economical Storages
- It's also inspired by *DATE Masamune* (伊達政宗)
  - He was a lord of the Tohoku(Sendai) area
  - He lived a long life despite losing his right eye
  - He achieved recovery from Keicho-sanriku earthquake in 1611
- DATE is a prefix used in the sense of "cool"
  - DATEotoko in Japanese means "cool guy" in English
  - Therefore, DATEstor means cool storage!

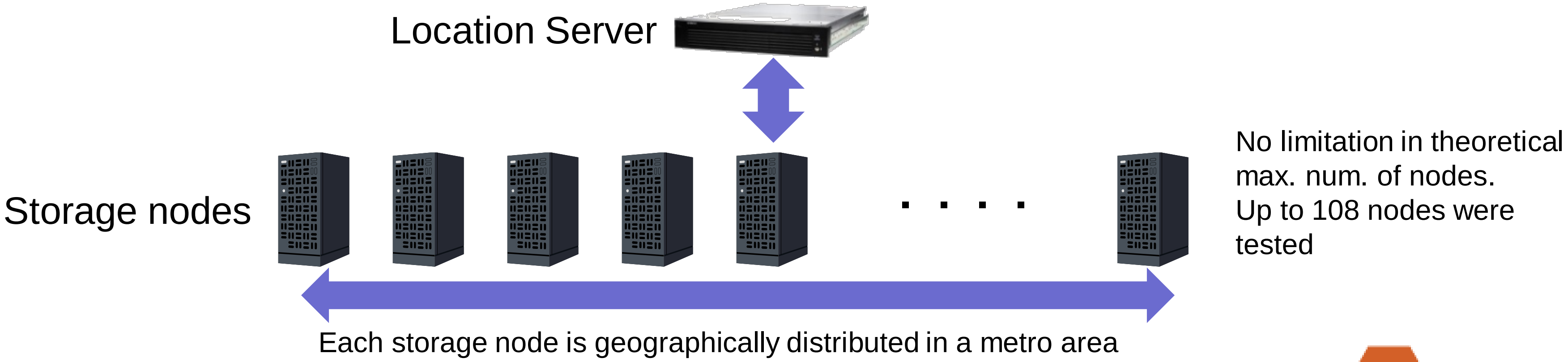


Search result of "DATEstor"

Bronze statue of DATE Masamune

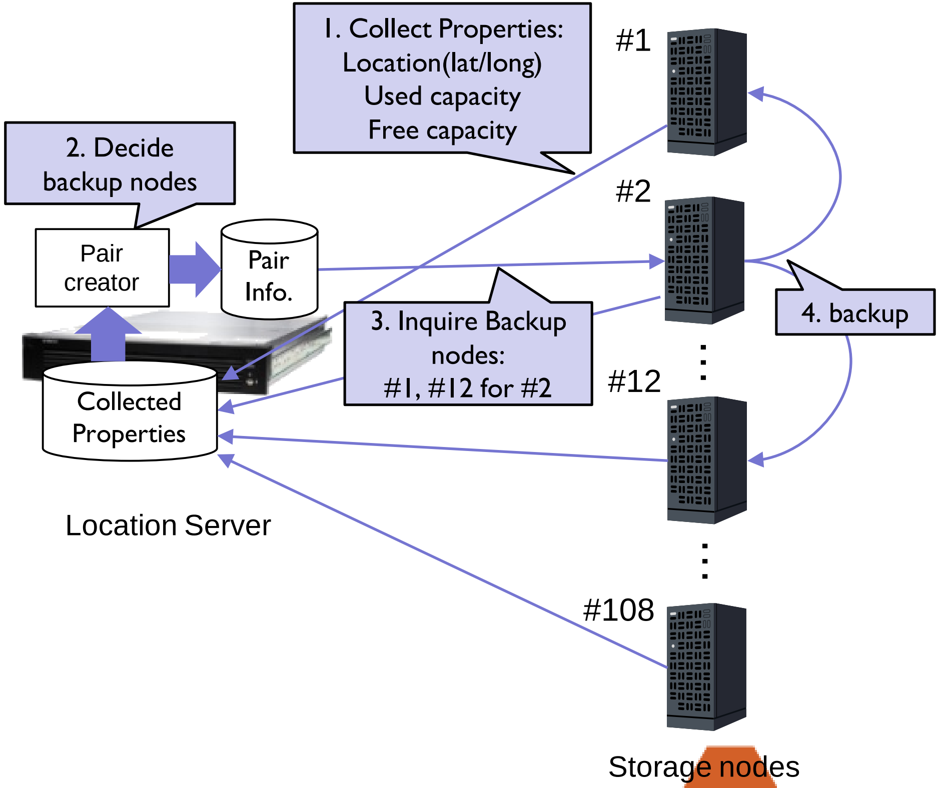
System Architecture of DATEstor

- The system consists of a location server and storage nodes.
- The location server manages properties of the storage nodes:
  - Location, Used capacity, and Free capacity for backup.
- Each storage node stores primary data and backup data of the other nodes.



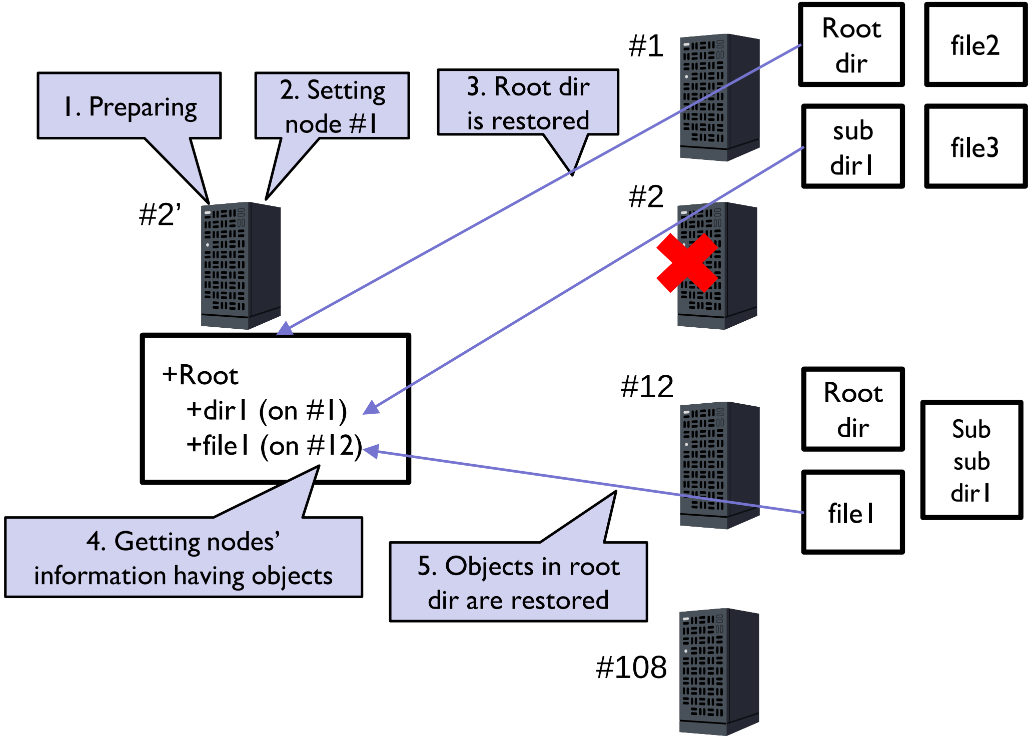
How the replication feature (RDR) works

1. The location server collects the properties of all storage nodes via REST API.
2. The location server decides appropriate(safe) storage nodes to back up data for each storage node using Integer Programming Problem (IPP) Technique.
3. Each storage node inquires the appropriate backup storage nodes to the location server via REST API before starting the first backup
4. Each storage node backs up its data to the appropriate storage nodes.



How the restoration feature (MRR) works

1. Preparing a new node for recovery (or an existing survived node also can be used for recovery)
2. Setting information of one of nodes, which has backup data of root directory, to the new node (#2').
3. Data of the root directory is restored from the node with backup data.
4. The restoring node gets nodes' information, which has backup data of objects stored in the root directory, from the metadata of root directory.
5. Data of the objects stored in the root directory is restored from the indicated nodes.
6. Continuing to dig into the directory.



Prototype System installed at Tohoku univ.

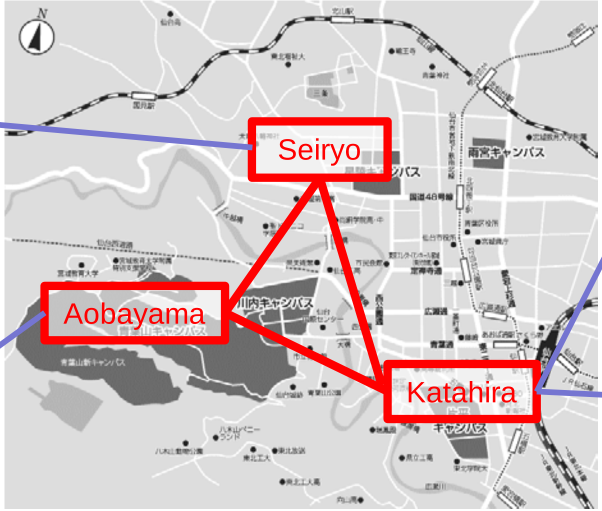
- It consists of 108 storage nodes.
  - The nodes are geographically distributed among 3 campuses and 4 buildings.
  - Each node emulates the node at each of 108 medical institutions around the Sendai area.

Medical dept.

dstor006

Cyber-science center

dstor004



Campuses map of Tohoku univ.

Main building of RIEC

dstor003, dstor005

IT center of RIEC

dstor001, dstor002, dstor007
vdstor001-vdstor101

Specifications Sheet of Prototype System

- The prototype system has been expanded step-by-step for over 4 years.

	gen. 0	1st gen.	2nd gen.
Dates in operation (CY)	2014 1Q	2014 2Q – 2015 1Q	2016 2Q – 2017 1Q
Num. of nodes (bare-metal, virtual)	10 (4, 6)	24 (4, 20)	108 (7, 101)
Num. of virtual sites	10	24	108
Num. of physical sites (buildings)	1 (1)	1 (1)	3 (4)
Average num. of replicas	1	1	1.5
Hint information for determinations of rep. pairs	Distance between sites	Distance between sites	Hazard-map information (J-SHIS)
Implemented features	Risk-aware Data Rep.	Risk-aware Data Rep.	Risk-aware Data Rep. Multi Route Restoration

Agenda

- Introduction
- Overview of DATEstor
- Evaluation on DATEstor
  - Evaluation of Availability
  - Evaluation of Recovery Time
- Conclusions

How to evaluate availability

We use following simulation steps to evaluate availability.

(1) Risk calculator

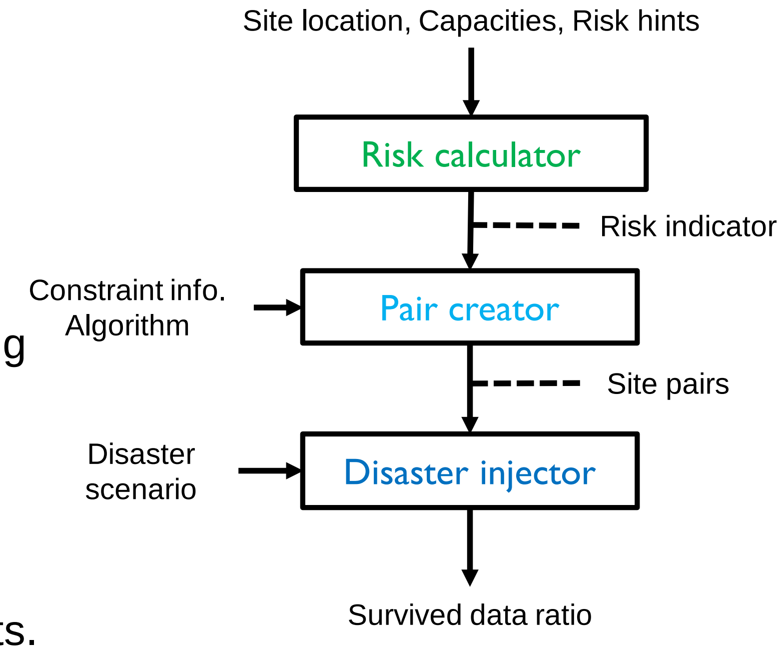
- Calculate the risk indicator based on a risk hint such as hazard map information

(2) Pair creator

- Decide backup sites using Integer Programming Problem (IPP) Techniques.

(3) Disaster injector

- Generate a virtual disaster along with disaster scenarios.
- Survived data ratio is calculated from the results.



Mathematical model of RDR pair creator

- Objective function
  - Sum of weighted risk indicator of each site

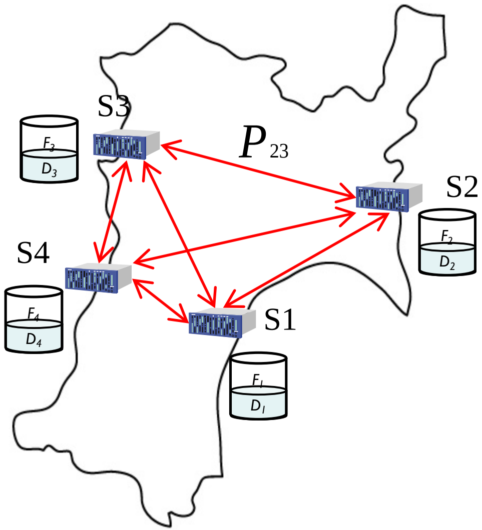
$$\min f(x) = \sum_{i, i \neq j} \sum_{j} \boxed{D_i P_{ij} x_{ij}}$$

Weighted risk indicator of site i

D_i :Amount of data in site i (weight)

P_{ij} :Risk indicator between site i and site j

$$x_{ij} = \begin{cases} 1 & \text{:Replicate data in site i to site j} \\ 0 & \text{:Do NOT replicate data in site i to site j} \end{cases}$$

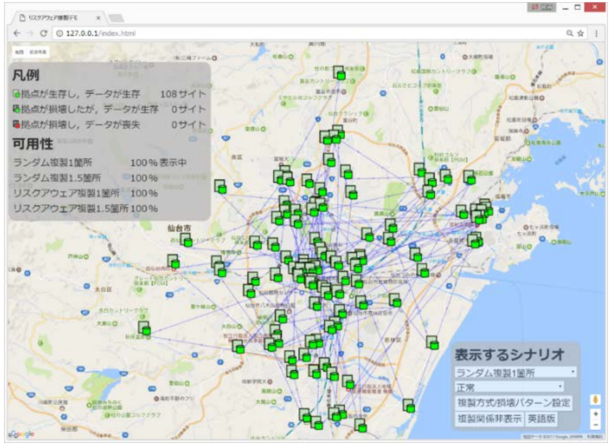


- Constraints
  - The num. of replicas: R_i $\quad \sum_j x_{ij} = R_i, \forall i$
  - Free capacity: F_j $\quad \sum_i D_i x_{ij} \leq F_j, \forall j$

(Note: Above objective function is limited to 1 replica)

Simulation Condition

- Supposed field and sites installing storage nodes
  - A field around Sendai-city, 108 medical institutions in the field
- Supposed disaster scenarios
  - 15 fault zones influencing the field
  - We use one replication pair pattern common to all disaster scenarios.



Supposed field and sites

List of disaster scenarios

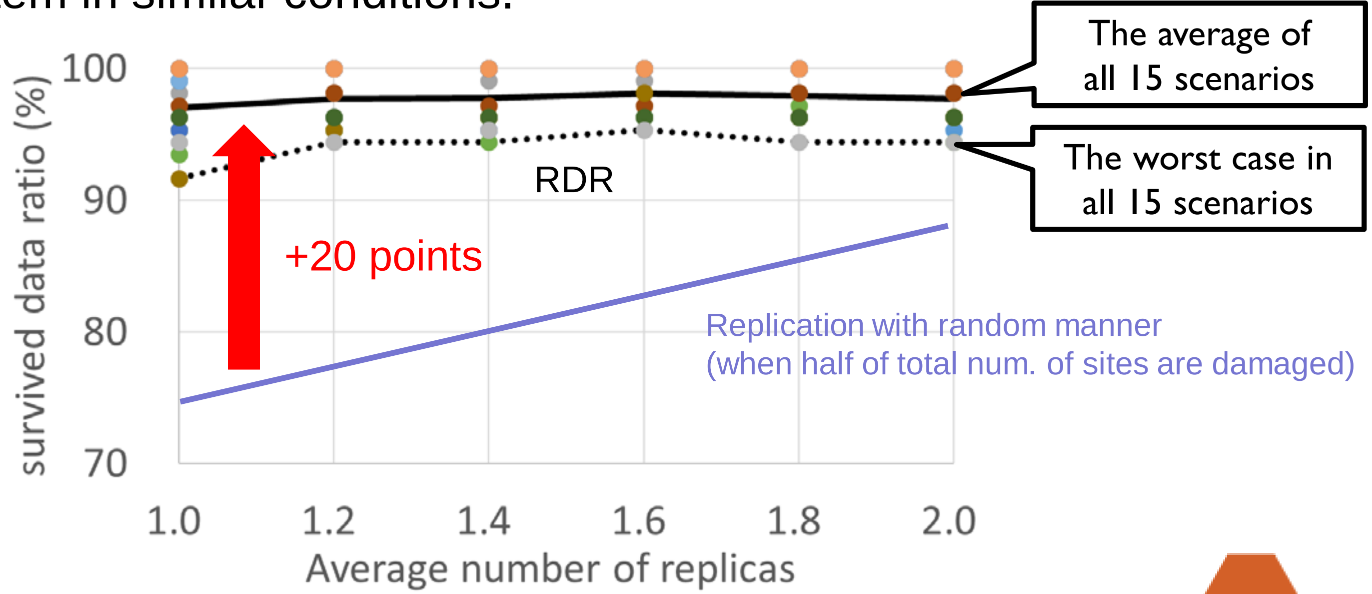
#	code	Name of fault zone
1	F001301	Western edge fault zone of the Kitakami lowland
2	F001502	Southern part of the Yokote basin fault zone
3	F001701	Eastern part of Shinzyo basin fault zone
4	F001801	Northern part of Yamagata basin fault zone
5	F001802	Southern part of Yamagata basin fault zone
6	F002001	Nagamachi-Rifu line fault zone
7	F002101	Western edge fault zone of Fukushima basin
8	F002201	Western edge fault zone of Nagai basin
9	F002301	Futaba fault zone
10	G030025	Asahiyama flexure
11	G030026	Medeshima estimated fault zone
12	G030027	Sakunami_Yashikidaira fault zone
13	G030028	Toogatta fault zone
14	G030029	Obanazawa fault zone
15	ATHOP	Great East Japan Earthquake

Simulation Condition (cont'd)

- Average number of replicas
  - 1.0, 1.2, 1.4, 1.6, 1.8, 2.0
- Assumptions
  - Only one disaster scenario occurs at a time. It damages up to half of nodes.
  - Most frequently damaged patterns occur for each disaster scenario based on hazard map information.
- Evaluation steps
  - If at least one of either the primary data or the backup data survive, the data is scored as survived data.
  - Survived data ratio of all nodes is calculated on each disaster scenario

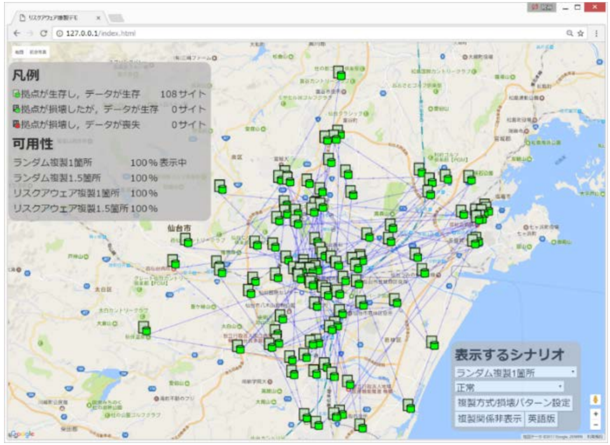
Simulation Results

- RDR achieves more than 90% of availability for all disaster scenarios.
- In addition to the simulation, the improvement of availability was confirmed on the prototype system in similar conditions.



Demo

- Supposed Field and supposed sites installing storage nodes
  - A field around Sendai-city, 108 medical institutions in the field
- Supposed Disaster scenarios
  - 15 fault zones influencing the supposed field



Supposed field and supposed sites

List of disaster scenarios

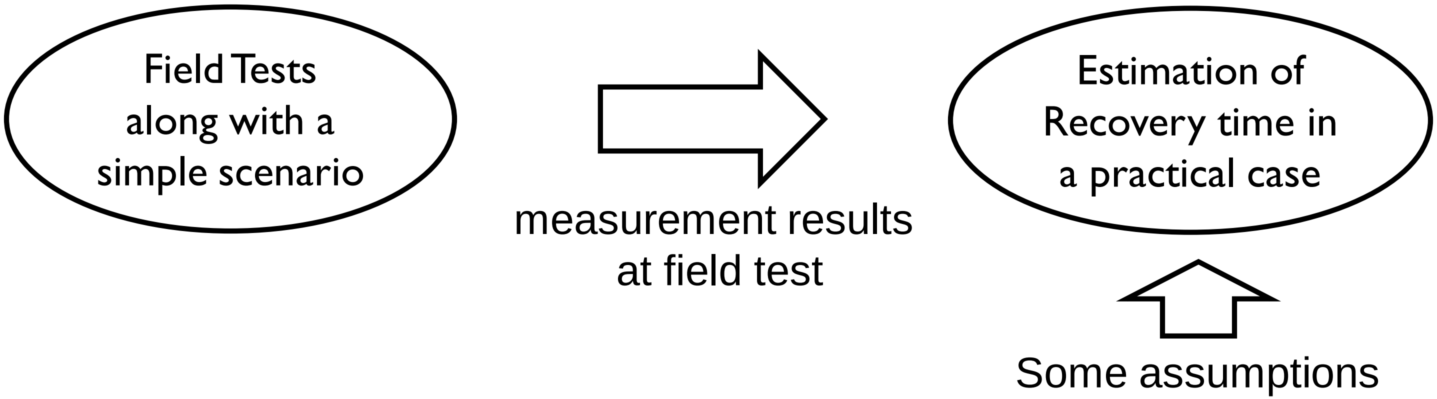
#	code	Name of fault zone
1	F001301	Western edge fault zone of the Kitakami lowland
2	F001502	Southern part of the Yokote basin fault zone
3	F001701	Eastern part of Shinzyo basin fault zone
4	F001801	Northern part of Yamagata basin fault zone
5	F001802	Southern part of Yamagata basin fault zone
6	F002001	Nagamachi-Rifu line fault zone
7	F002101	Western edge fault zone of Fukushima basin
8	F002201	Western edge fault zone of Nagai basin
9	F002301	Futaba fault zone
10	G030025	Asahiyama flexure
11	G030026	Medeshima estimated fault zone
12	G030027	Sakunami_Yashikidaira fault zone
13	G030028	Toogatta fault zone
14	G030029	Obanazawa fault zone
15	ATHOP	Great East Japan Earthquake

Agenda

- Introduction
- Overview of DATEstor
- Evaluation on DATEstor
  - Evaluation of Availability
  - Evaluation of Recovery Time
- Conclusions

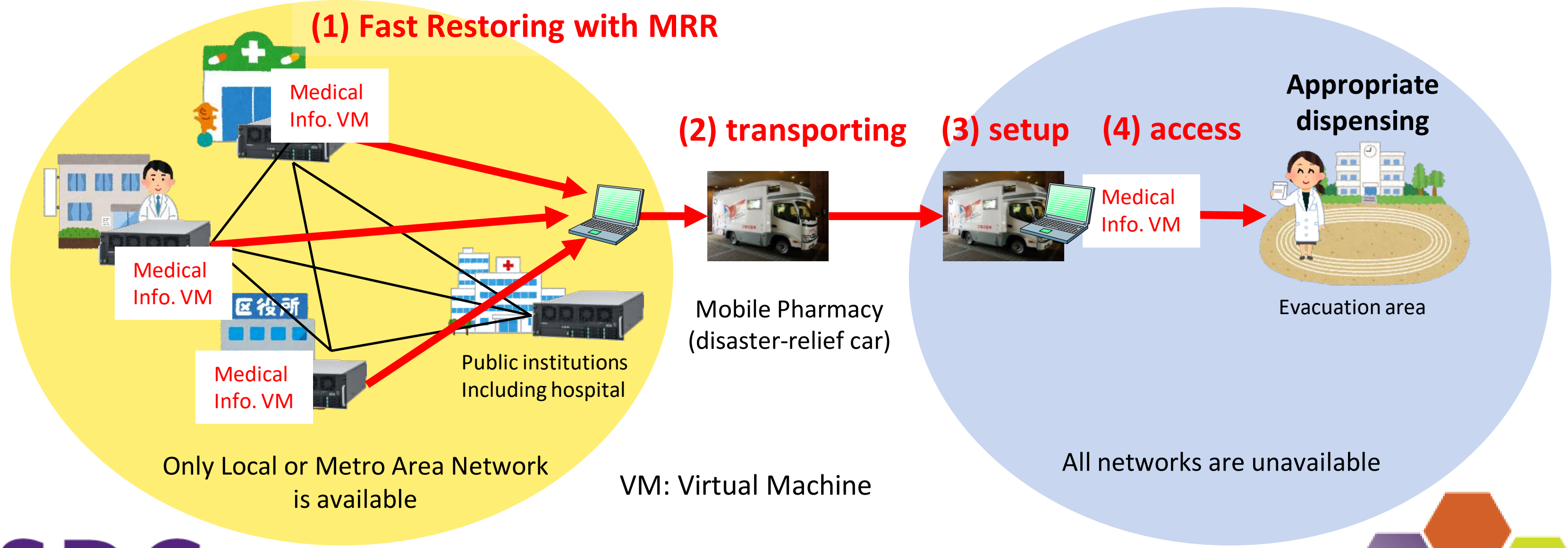
How to estimate Recovery Time

- We estimate recovery time from the measurement results at field tests along with a simple scenario
  - To make it more practical, the scenario is discussed with the board members of the Miyagi pharmacist association



Scenario of field test

- The scenario is to restart dispensing operation in an evacuation area after a disaster
- We kept track of the time from (1) to (4) in the field test.



VM: Virtual Machine

Field test with Pharmacist Association

- Date
  - 23rd Nov. 2016
- Location
  - Katahira campus in Tohoku univ.
- Attendees
  - Project members
  - Members from the Miyagi pharmacist association
  - General participants

Group photo after finishing field test

Pictures of the field test

An engineer restoring VM Images including medical Information to a laptop PC

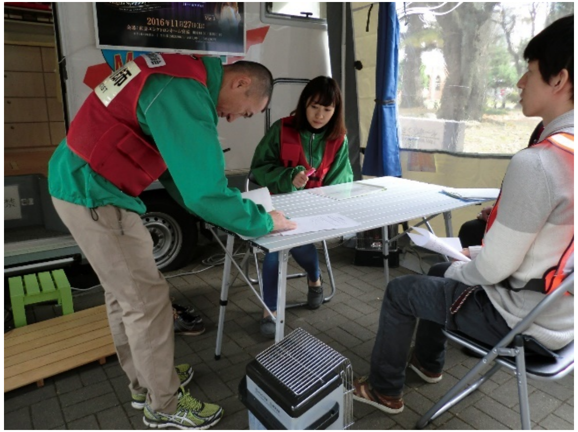
Pharmacists setting up a temporal pharmacy with the Mobile Pharmacy

An engineer restarting the VM on the laptop PC

Pictures of the field test (cont'd)

A health interview by a medical doctor

A pharmacist answering an inquiry of medicine histories



Pharmacists checking an emergency prescription

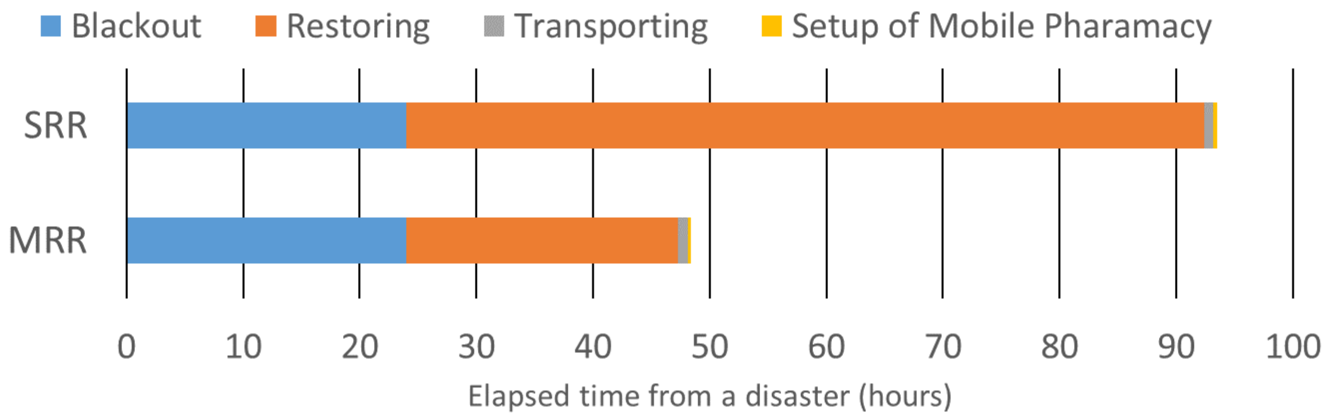
Assumptions in a case

- We suppose some items in a practical case of restoring VM images of all pharmacies in a city.

Items	Assumptions	Remarks
Target Data Size	66 VM disk files (100GB for each)	Corresponding to the number of pharmacies in a city (Ishinomaki city)
Network Delay	100 ms	Because of congestion
Blackout time	24 hours	At many buildings, power is resumed within a day
Transporting speed	3 times slower	Because of traffic jam

Estimated recovery time

- We estimated a recovery time to restart dispensing operation in the case.
- It is confirmed that MRR shortens the time compared with restoring from one node (SRR: Single Route Restoration)
  - From 93 hours to 48 hours in the case.



Agenda

- Introduction
- Overview of DATEstor
- Evaluation on DATEstor
- Conclusions

Conclusions

- The case studies of the Great East Japan Earthquake and Tsunami of 2011 were presented.
  - It damaged not only storage nodes but the network between the disaster area and the outside of the disaster area.
- Highly-available Metro Area Distributed Storage Systems were proposed.
  - Two key features were also proposed:
    Risk-aware Data Replication and Multi Route Restoration
- Effectiveness of the proposed system were confirmed
  - Availabilities are improved to more than 90%.
  - Recovery Time to restart dispensing operation is decreased by almost half at a rough estimation from the field test results.

Further information

- Journal/Transaction Papers
  - Takaki Nakamura, Shinya Matsumoto, and Hiroaki Muraoka, "Discreet Method to Match Safe Site-Pairs in Short Computation Time for Risk-aware Data Replication," IEICE Transactions on Information and Systems, Vol. E98-D, No. 8 (2015), pp. 1493-1502
  - Shinya Matsumoto, Takaki Nakamura, and Hiroaki Muraoka, "Redundancy-based Iterative Method to Select Multiple Safe Replication Sites for Risk-aware Data Replication," IEEJ Transactions on Electrical and Electronic Engineering, Vol. 11, No. 1 (2016), pp. 96-102
  - Takaki Nakamura, Shinya Matsumoto, Masaru Tezuka, Satoru Izumi, and Hiroaki Muraoka, "Comparison of Distance Limiting Methods for Risk-aware Data Replication in Urban and Suburban Area," Journal of Information Processing, Vol. 24, No. 2 (2016), pp. 381-389
- Conference Papers
  - Shinya Matsumoto, Takaki Nakamura, and Hiroaki Muraoka, "Risk-aware Data Replication to Massively Multi-sites against Widespread Disasters," Proc. of the 2nd Asian Conference on Information Systems (ACIS) (2013), 7 pages
  - Shinya Matsumoto, Takaki Nakamura, and Hiroaki Muraoka, "Risk-based Method for Data Redundancy Determination to Improve Replica Capacity Efficiency," Proc. of the 3rd Asian Conference on Information Systems (ACIS) (2014), 8 pages
  - Hitoshi Kamei, Shinya Matsumoto, Takaki Nakamura, and Hiroaki Muraoka, "REC2: Restoration Method Using Combination of Replication and Erasure Coding," Proc. of 5th IIAI International Congress on Advanced Applied Informatics (2016), pp 936-941

Acknowledgments

- This is the collaboration work with:
  - Tohoku University
  - Hitachi
  - Hitachi Solutions East Japan
- This work is also supported as “Research and Development on Highly Functional and Highly Available Information Storage Technology,” sponsored by the Ministry of Education, Culture, Sports, Science and Technology in Japan.

Q & A

Thank You!