クラウドコンピューティング基礎論

第5回

創造情報•小林克志

ikob@acm.org

Outline

1.Administravia

2.Quiz and homework review

3.Distributed Data Store
  - CAP theorem -

4.Global Services

Course Outline

1. Administrivia
2. Cloud computing
3. Service reliability
4. Scale-up / Scale-out
5. Distributed data stores
6. Global services
7. Datacenter networkings (1)
8. Datacenter networkings (2)
9. Network performance
10. User experiences
11. Network latencies
12. Advanced topics

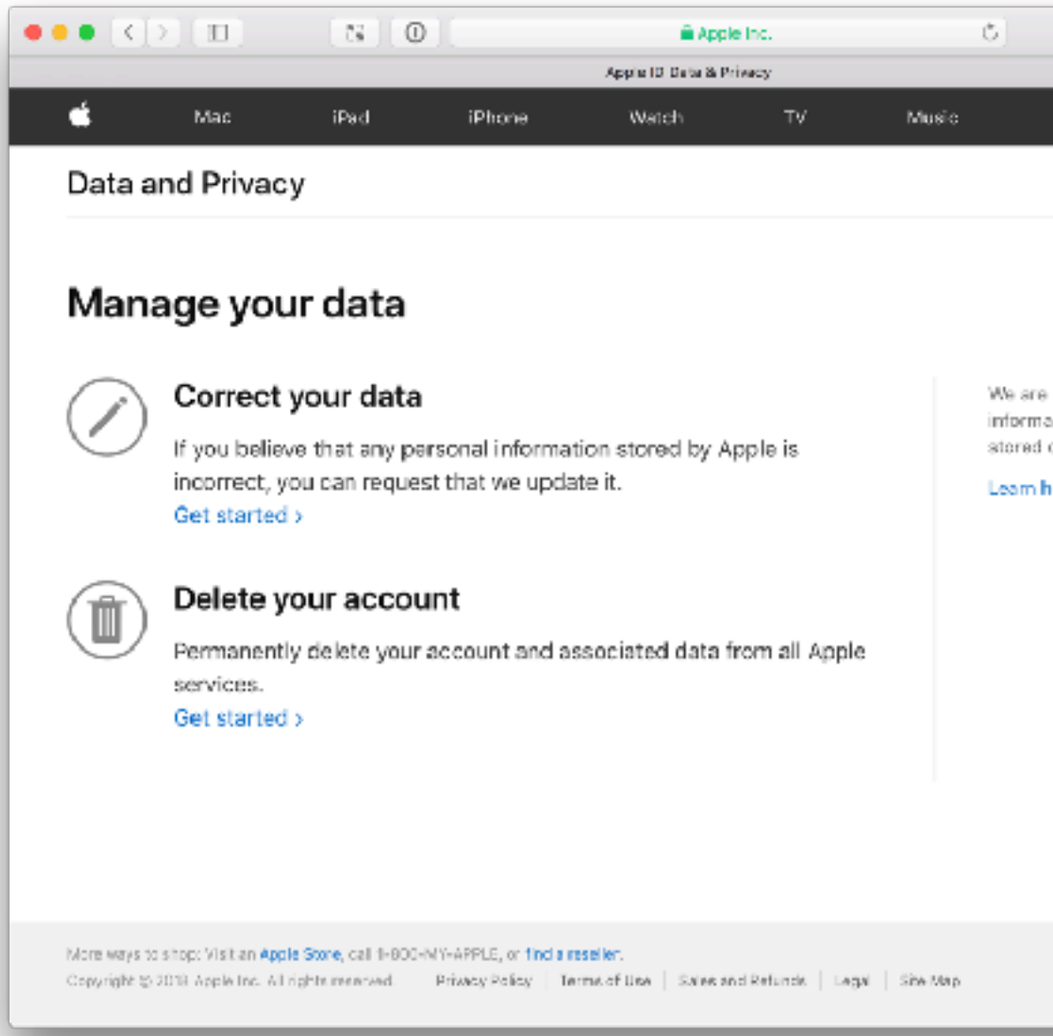
Topic relating to GDPR

“Apple launches new privacy portal, users can download a copy of everything Apple knows about them

Apple has today launched its new Data and Privacy website, allowing Apple users to download everything that Apple personally associates with your account, from Apple ID info, App Store activity, AppleCare history to data stored in iCloud like photos and documents. This is currently only available for European Union accounts, to comply with GDPR, and **will roll out worldwide in the coming months.**

There are also simple shortcuts to updating your info, temporarily deactivating your account and options to permanently delete it. Here’s how to do it …”

9to5mac.com より



Today’s quiz

- Tell an actual example of a systems that are designed scale-up and scale-out approaches other than computer system.

- Ex.)
  Scale-up: To switch larger lecture-room so that more student can be accepted.
  Scale-out: To switch distance learning using video on demand.

- Submit your answers either in Japanese or in English via the course web.

- Bonus points will be given for excellent jokes and stories.

本日のクイズ

- スケールアップ／スケールアウトについて（計算機以外の）具体例を示せ。

- 例）スケールアップ：多くの学生を受け入れるためにより大きな教室に変更する。
  スケールアウト：VoD を利用する遠隔学習に切り替える。

- 秀逸な小咄には加点する

- 講義 Web フォームから記入すること。

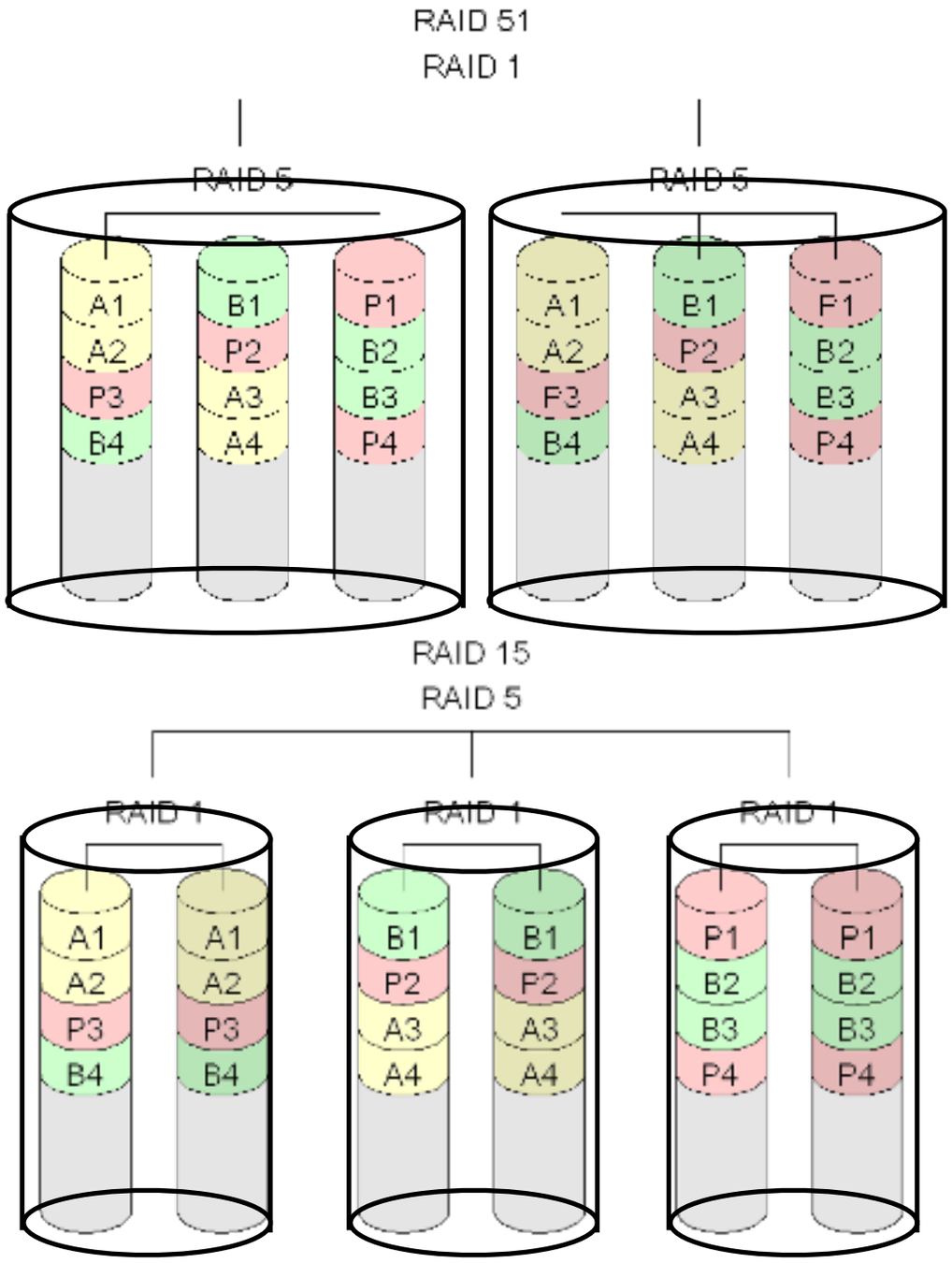
Today’s assignment

- Compare couple of nested RAID configurations
  (1) RAID-51, and (2) RAID-15.

- Discuss which configuration is better, and why ?

    - Consider the operation after HDD crash.

- Submit your answers either in Japanese or in English via the course web.



Wikipedia より

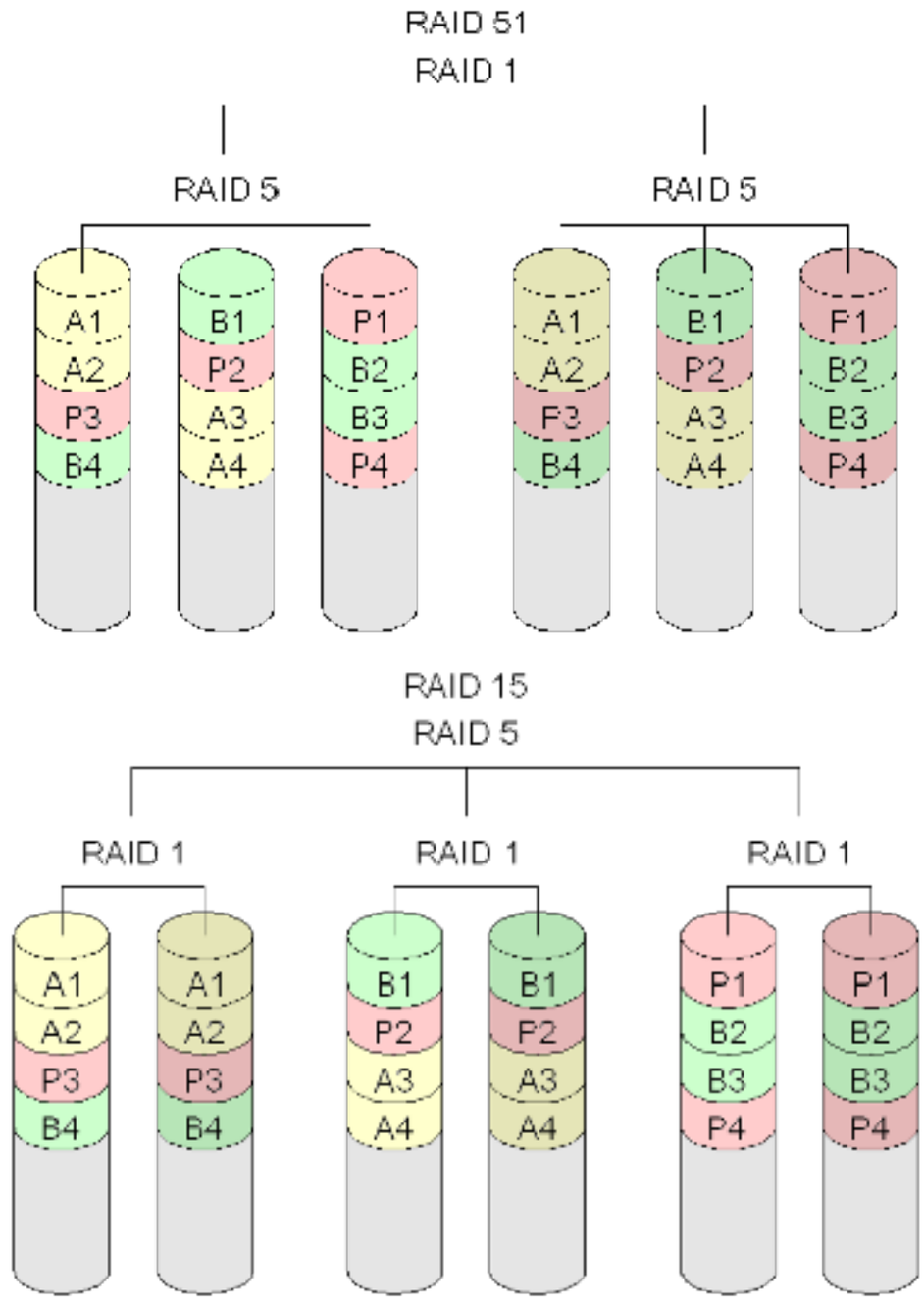
本日の課題

- 2 種類の nested RAID 構成、(1) RAID-51, and (2) RAID-15 を比較せよ。

- いずれの構成が好ましいか議論せよ？

  - ディスク障害後の運用について検討してみる。

- 講義 Web フォームから記入すること。



Wikipedia より

Outline

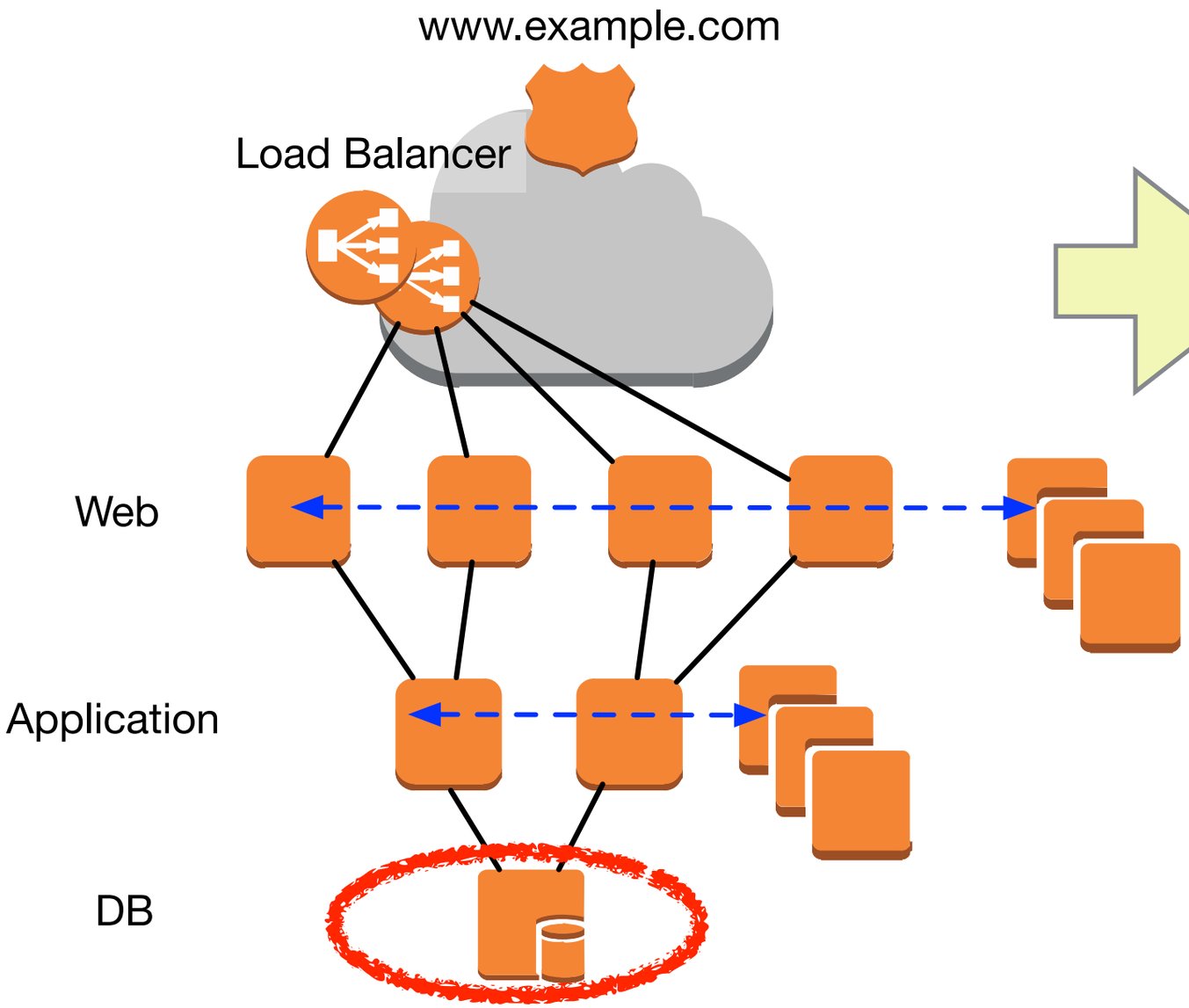
1.Administravia

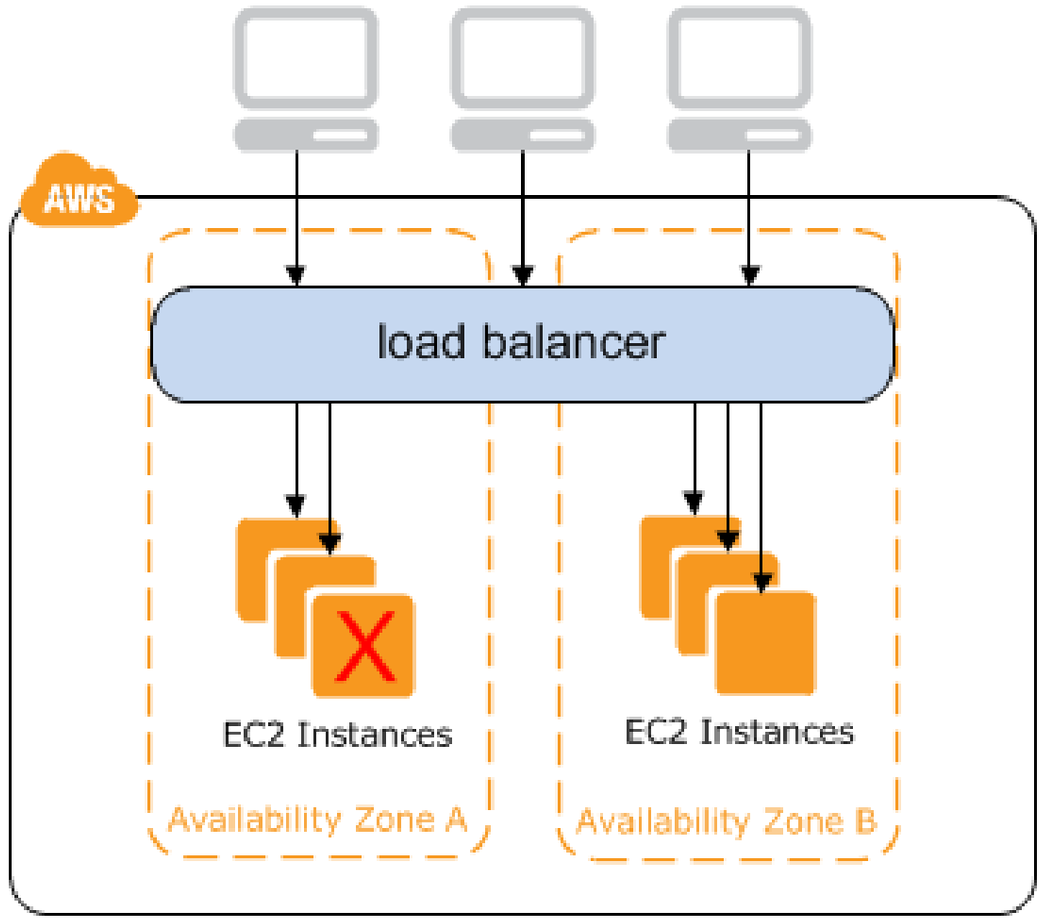
2.Quiz and homework review

3.Distributed Data Store
  - CAP theorem -

4.Global Services

3-tier Web architecture and scalability





“Elastic Load Balancing Developer Guide”, Amazon W

aws.amazon.com

Amazon RDS Multi-AZ Deployments

Menu | amazon webservices | English | My Account | Sign In to the Console

PRODUCTS & SERVICES

Amazon RDS Multi-AZ Deployments

RELATED LINKS

Amazon RDS

Amazon RDS for MySQL

Amazon RDS for Oracle

Amazon RDS for SQL Server

Management Console

Release Notes

Get Started for Free

Create Free Account

Amazon RDS Multi-AZ Deployments

Amazon RDS Multi-AZ deployments provide enhanced availability and durability for Database (DB) Instances, making them a natural fit for production database workloads. When you provision a Multi-AZ DB Instance, Amazon RDS automatically creates a primary DB Instance and synchronously replicates the data to a standby instance in a different Availability Zone (AZ). Each AZ runs on its own physically distinct, independent infrastructure, and is engineered to be highly reliable. In case of an infrastructure failure (for example, instance hardware failure, storage failure, or network disruption), Amazon RDS performs an automatic failover to the standby, so that you can resume database operations as soon as the failover is complete. Since the endpoint for your DB Instance remains the same after a failover, your application can resume database operation without the need for manual administrative intervention.

Get Started with AWS for Free

Create a Free Account

AWS Free Tier includes 750hrs of Micro DB Instance each month for one year, 20GB of Storage, and 20GB for Backups with Amazon Relational Database Service (RDS).

View AWS Free Tier Details »

Benefits

Enhanced Durability

Multi-AZ deployments for the MySQL, Oracle, and PostgreSQL engines utilize synchronous physical replication to keep data on the standby up-to-date with the primary. Multi-AZ deployments for the SQL Server engine use

メニューを表示

aws.amazon.com

Amazon RDS Multi-AZ Deployments

Menu | amazon webservices | English | My Account | Sign In to the Console

PRODUCTS & SERVICES

Amazon RDS Multi-AZ Deployments

RELATED LINKS

- Amazon RDS
- Amazon RDS for MySQL
- Amazon RDS for Oracle
- Amazon RDS for SQL Server
- Management Console
- Release Notes

Get Started for Free

Create Free Account

Benefits

Enhanced Durability

Multi-AZ deployments for the MySQL, Oracle, and PostgreSQL engines utilize synchronous physical replication to keep data on the standby up-to-date with the primary. Multi-AZ deployments for the SQL Server engine use synchronous logical replication to achieve the same result, employing SQL Server-native Mirroring technology. Both approaches safeguard your data in the event of a DB Instance failure or loss of an Availability Zone.

If a storage volume on your primary fails in a Multi-AZ deployment, Amazon RDS automatically initiates a failover to the up-to-date standby. Compare this to a Single-AZ deployment: in case of a Single-AZ database failure, a user-initiated point-in-time-restore operation will be required. This operation can take several hours to complete, and any data updates that occurred after the latest restorable time (typically within the last five minutes) will not be available.

Amazon Aurora employs a highly durable, SSD-backed virtualized storage layer purpose-built for database workloads. Amazon Aurora automatically replicates your volume six ways, across three Availability Zones. Amazon Aurora storage is fault-tolerant, transparently handling the loss of up to two copies of data without affecting database write availability and up to three copies without affecting read availability. Amazon Aurora storage is also self-healing. Data blocks and disks are continuously scanned for errors and replaced automatically.

Increased Availability

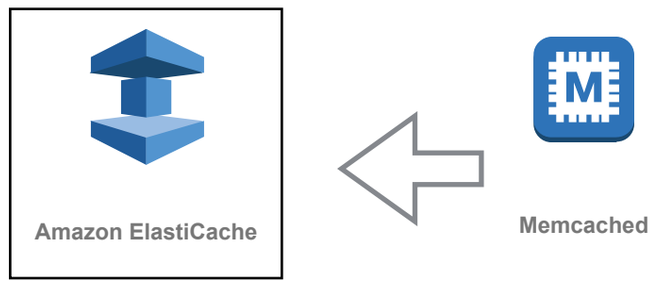
You also benefit from enhanced database availability when running Multi-AZ deployments. If an Availability Zone failure or DB Instance failure occurs, your availability impact is limited to the time automatic failover takes to complete: typically under one minute for Amazon Aurora and one to two minutes for other database engines (see the RDS FAQ for details).

The availability benefits of Multi-AZ deployments also extend to planned maintenance and backups. In the case of system upgrades like OS patching or DB Instance scaling, these operations are applied first on the standby, prior to the automatic failover. As a result, your availability impact is, again, only the time required for automatic failover to complete.

メニューを表示

Cloud Datastore: Why many services ?

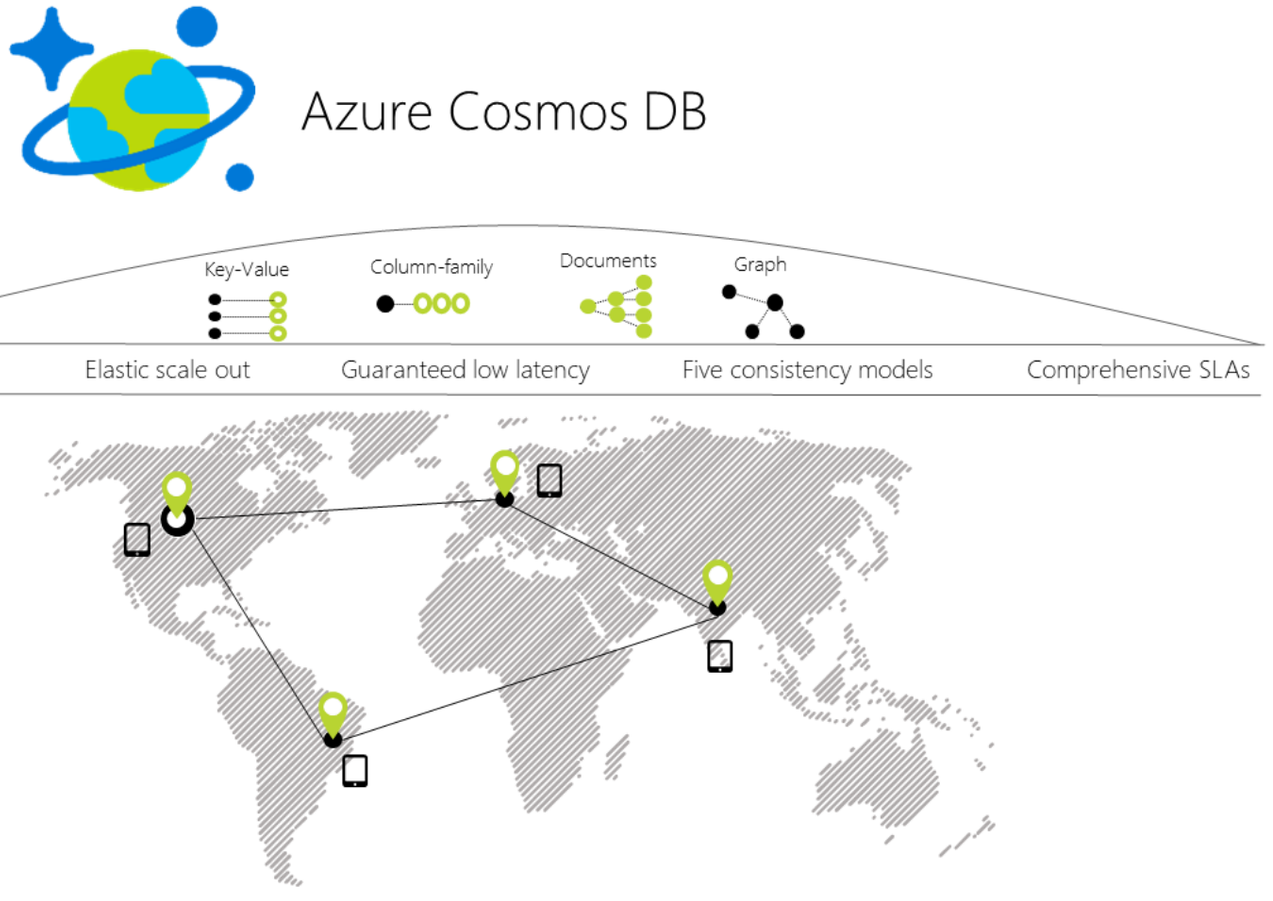
- Advantages against own datastore on customers computing instances.
  - Operation cost: provisioning, patching, backup, recovery, failure detection, and repair.
  - Availability: Multi-AZ deployment
  - Scalability: Read-replica, storage capacity.
- Relational Database (RDB)
  - e.g., Amazon Relational Database Service (RDS) for Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle, Microsoft SQL Server.
- NoSQL
  - Key-Value, Column Oriented/Tabular, Document-oriented…
    - e.g., Google BigTable, Cassandra, Amazon DynammoDB, Redshift
- In memory cache
  - Key-value, faster but volatile.
    - e.g., Memcached, Amazon Elastic Cache



Welcome to Azure Cosmos DB

5/10/2017 • 9 min to read • Edit Online

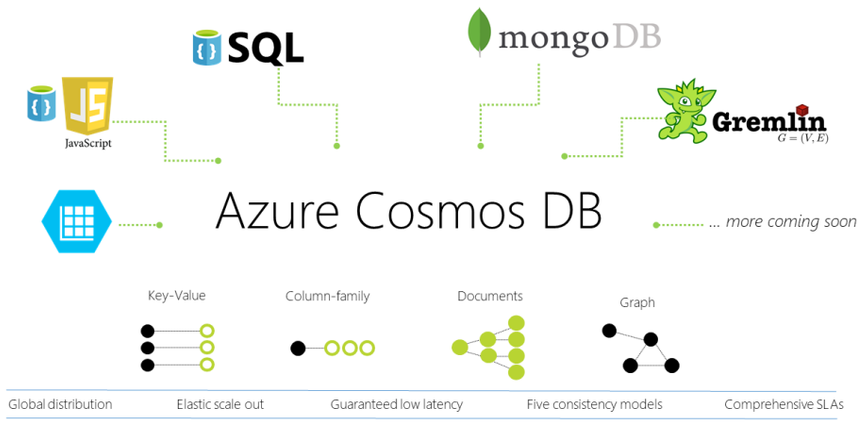
Azure Cosmos DB is Microsoft's globally distributed, multi-model database. With the click of a button, Azure Cosmos DB enables you to elastically and independently scale throughput and storage across any number of Azure's geographic regions. It offers throughput, latency, availability, and consistency guarantees with comprehensive service level agreements (SLAs), something no other database service can offer.



remove an existing region or take a region that was previously associated with their database account offline.

Multi-model, multi-API support

Azure Cosmos DB natively supports multiple data models including documents, key-value, graph, and column-family. The core content-model of Cosmos DB's database engine is based on atom-record-sequence (ARS). Atoms consist of a small set of primitive types like string, bool, and number. Records are structs composed of these types. Sequences are arrays consisting of atoms, records, or sequences.



A globally-distributed, multi-model database service

The database engine can efficiently translate and project different data models onto the ARS-based data model. The core data model of Cosmos DB is natively accessible from dynamically typed programming languages and can be exposed as-is as JSON.

The service also supports popular database APIs for data access and querying. Cosmos DB's database engine currently supports DocumentDB SQL, MongoDB, Azure Tables (preview), and Gremlin (preview). You can continue to build applications using popular OSS APIs and get all the benefits of a battle-tested and fully managed, globally distributed database service.

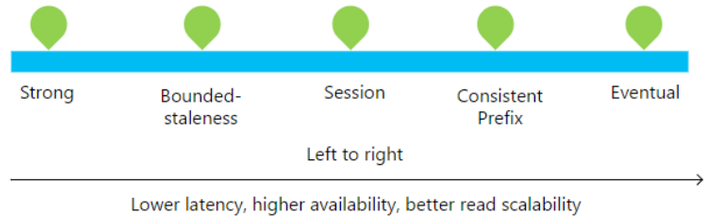
Horizontal scaling of storage and throughput

All the data within a Cosmos DB container (for example, a document collection, table, or graph) is horizontally partitioned and transparently managed by resource partitions. A resource partition is a consistent and highly available container of data partitioned by a customer specified partition-key. It provides a single system image for a set of resources it manages and is a fundamental unit of scalability and distribution. Cosmos DB is designed to let you elastically scale throughput based on the application traffic patterns across different geographical regions to support fluctuating workloads varying both by geography and time. The service manages the partitions transparently without compromising the availability, consistency, latency, or throughput of a Cosmos DB container.

Multiple, well-defined consistency models

Commercial distributed databases fall into two categories: databases that do not offer well-defined, provable consistency choices at all, and databases which offer two extreme programmability choices (strong vs. eventual consistency). The former burdens application developers with minutia of their replication protocols and expects them to make difficult tradeoffs between consistency, availability, latency, and throughput. The latter puts a pressure to choose one of the two extremes. Despite the abundance of research and proposals for more than 50 consistency models, the distributed database community has not been able to commercialize consistency levels beyond strong and eventual consistency.

Cosmos DB allows you to choose between five well-defined consistency models along the consistency spectrum – strong, bounded staleness, session, consistent prefix, and eventual.



The following table illustrates the specific guarantees each consistency level provides.

Consistency Levels and guarantees

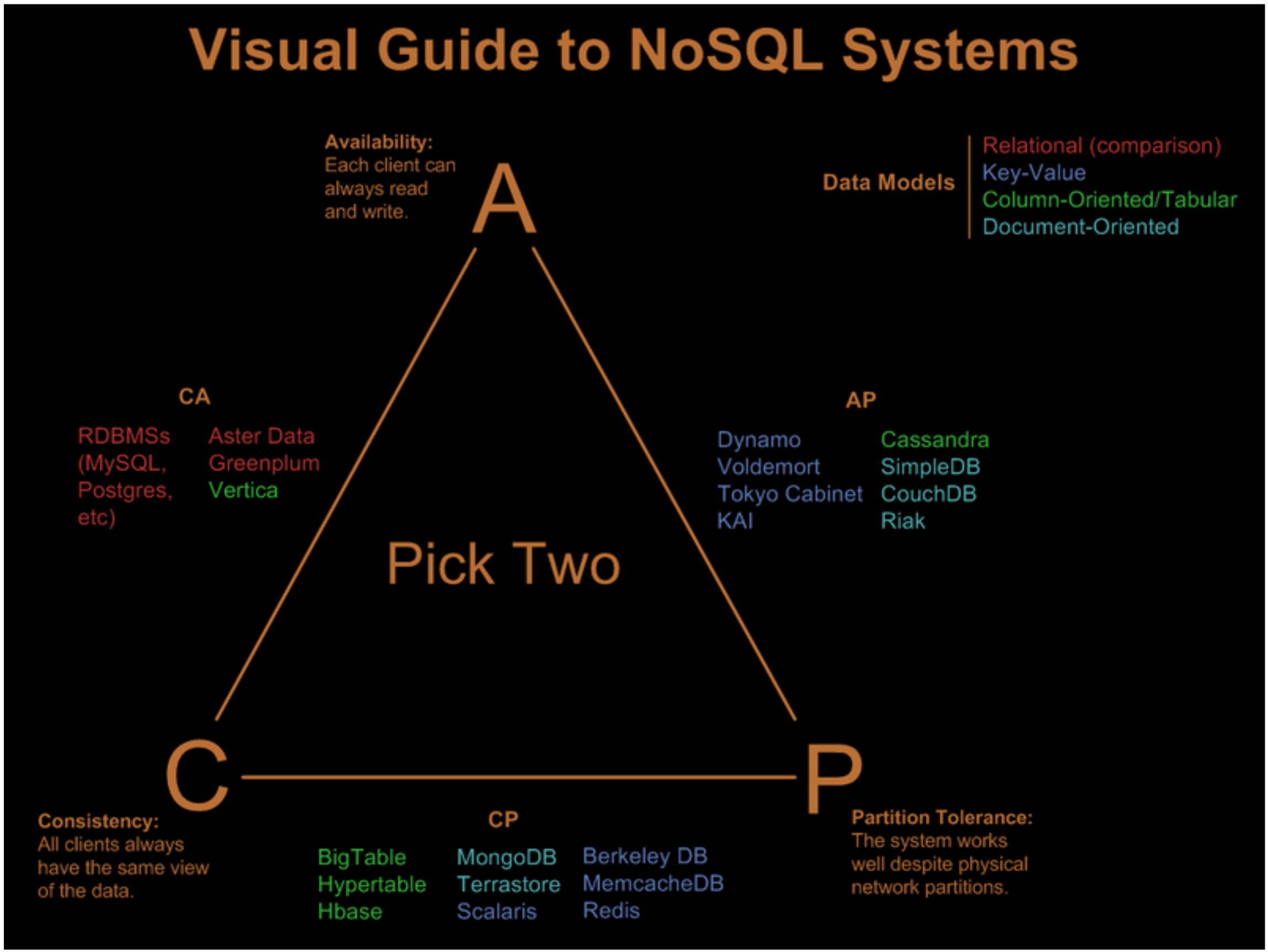
CONSISTENCY LEVEL	GUARANTEES
Strong	Linearizability
Bounded Staleness	Consistent Prefix. Reads lag behind writes by k prefixes or t interval
Session	Consistent Prefix. Monotonic reads, monotonic writes, read-your-writes, write-follows-reads
Consistent Prefix	Updates returned are some prefix of all the updates, with no gaps
Eventual	Out of order reads

You can configure the default consistency level on your Cosmos DB account (and later override the consistency on a specific read request). Internally, the default consistency level applies to data within the partition sets which may be span regions.

Guaranteed service level agreements

Cosmos DB is the first managed database service to offer 99.99% SLA guarantees for availability, throughput, low latency, and consistency.

- Availability: 99.99% uptime availability SLA for each of the data and control plane operations.
- Throughput: 99.99% of requests complete successfully
- Latency: 99.99% of <10 ms latencies at the 99th percentile
- Consistency: 100% of read requests will meet the consistency guarantee for the consistency level requested by you.

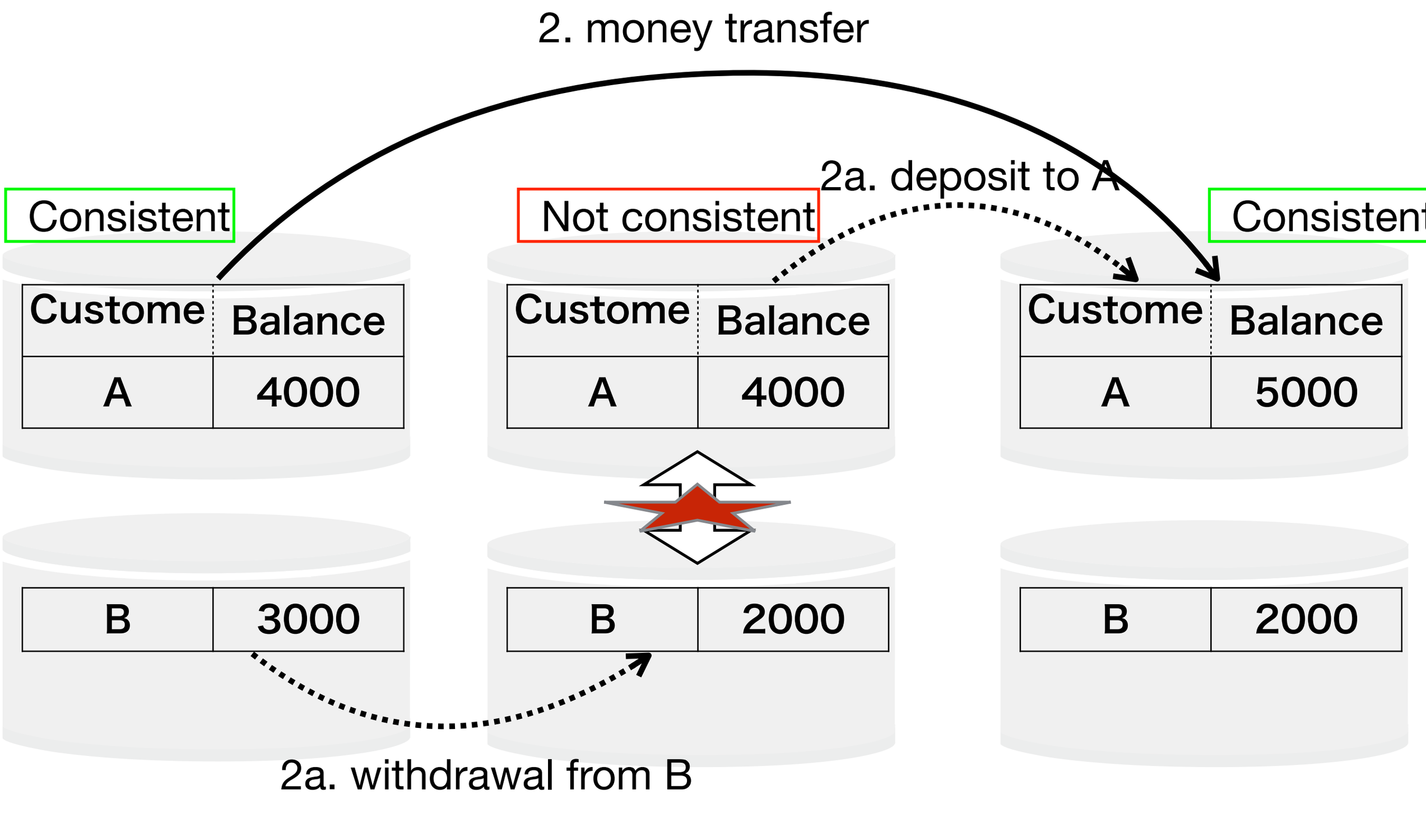


Nathan Hurst, “Visual Guide to NoSQL Systems”, 2010

Ensuring ACID in RDBMS transaction.

- Atomic
- Consistency
- Isolation
- Durability

A transaction on Distributed DB



ACID vs. BASE

- DBMS research is about ACID (mostly)
- But we forfeit "C" and "I" for availability, graceful degradation, and performance

This tradeoff is fundamental.

BASE:

- **B**asically **A**vailable
- **S**oft-state
- **E**ventual consistency

PODC Keynote, July 19, 2000

Eric Brewer, "Towards Robust Distributed Systems",2000

ACID vs. BASE

Inktomi

ACID

- Strong consistency
- Isolation
- Focus on "commit"
- Nested transactions
- Availability?
- Conservative (pessimistic)
- Difficult evolution (e.g. schema)

BASE

- Weak consistency
  - stale data OK
- Availability first
- Best effort
- Approximate answers OK
- Aggressive (optimistic)
- Simpler!
- Faster
- Easier evolution

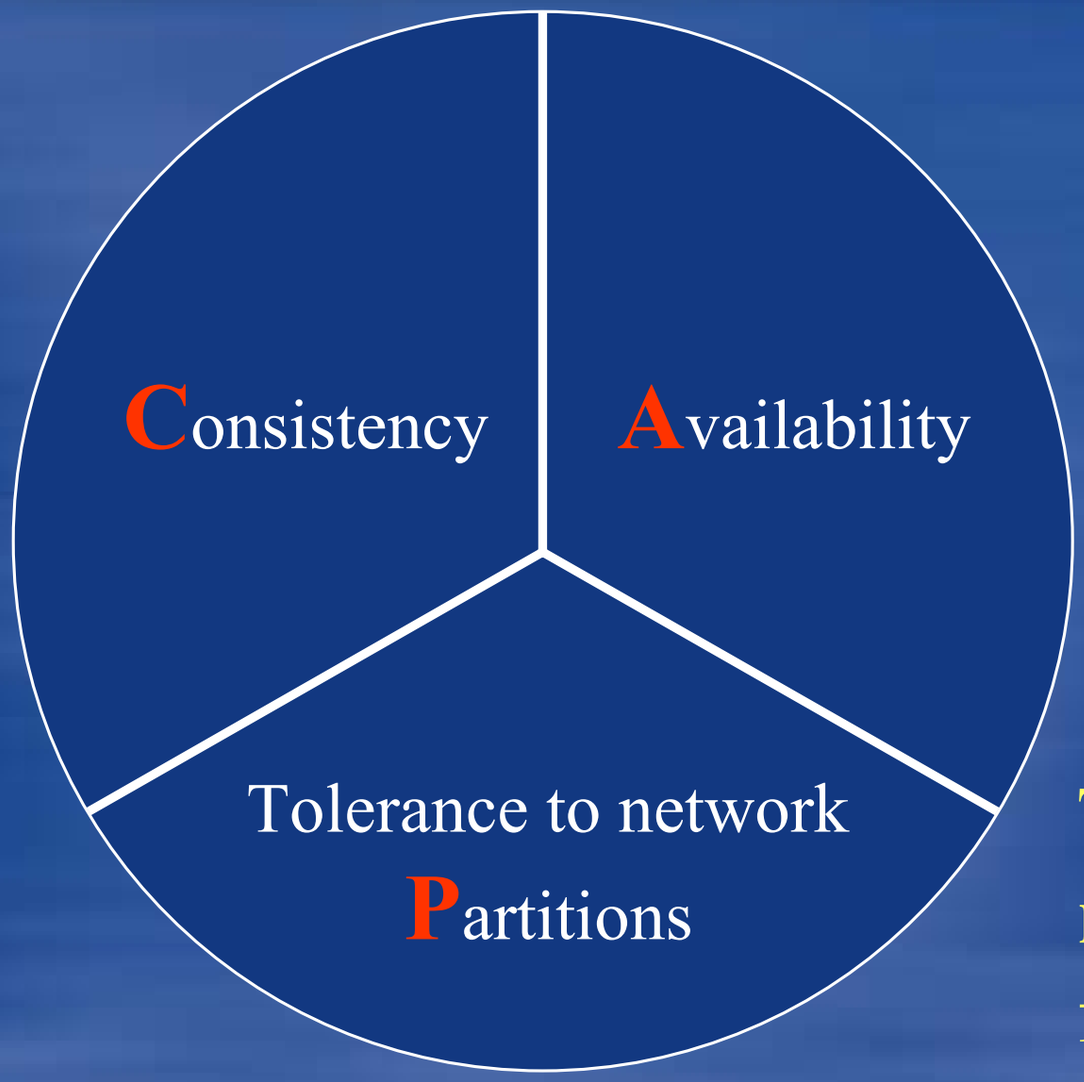
← But I *think* it's a *spectrum* →

PODC Keynote, July 19, 2000

Eric Brewer, "Towards Robust Distributed Systems",2000

The CAP Theorem

Inktomi



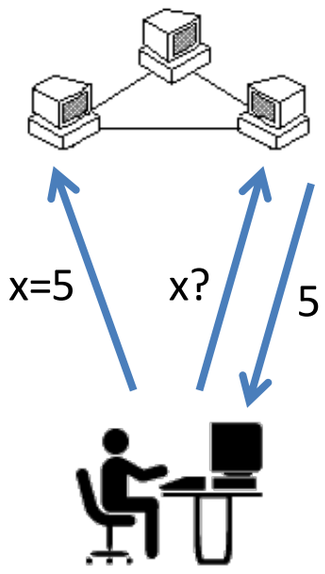
Theorem: You can have **at most two** of these properties for any shared-data system

PODC Keynote, July 19, 2000

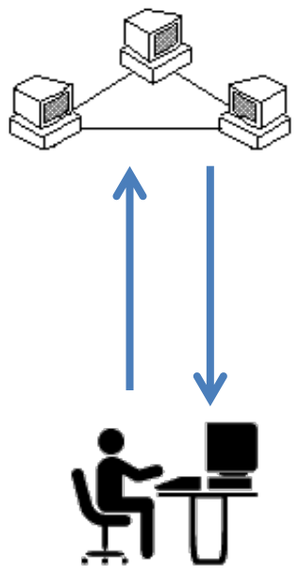
Eric Brewer, "Towards Robust Distributed Systems",2000

CAP Theorem

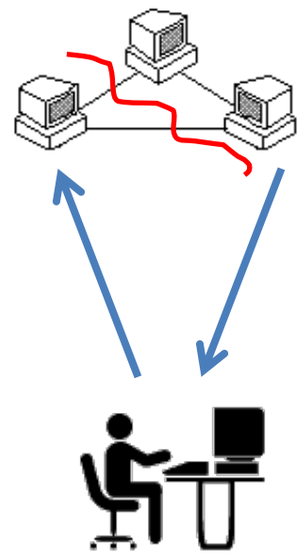
Consistency



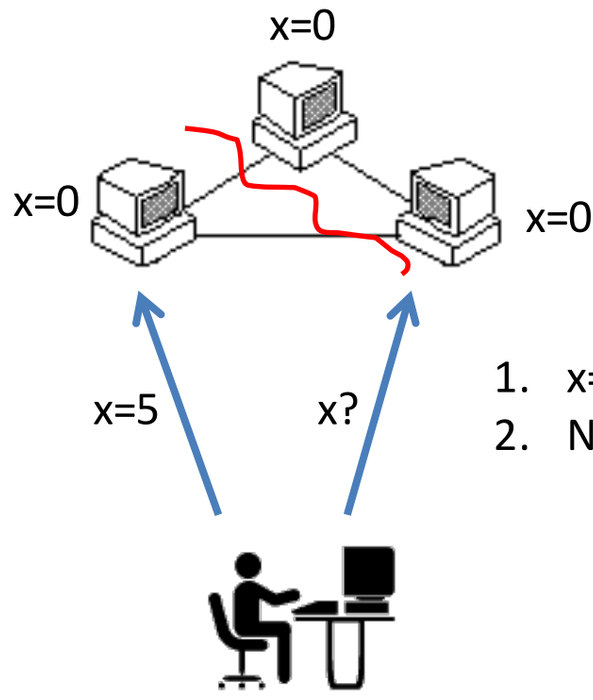
Availability



Partition tolerance



CAP Theorem - Proof



1. x=0 ➔ Not consistent
2. No response ➔ Not available

Cassandra Write Data Flows

Single Region, Multiple Availability Zone

1. Client Writes to any Cassandra Node
2. Coordinator Node replicates to nodes and Zones
3. Nodes return ack to coordinator
4. Coordinator returns ack to client
5. Data written to internal commit log disk

If a node goes offline, hinted handoff completes the write when the node comes back up.

Requests can choose to wait for one node, a quorum, or all nodes to ack the write

SSTable disk writes and compactions occur asynchronously

NETFLIX

Netflix, “Benchmarking Cassandra Scalability on AWS - Over a million writes per second”, Nov. 2011

Today’s Assignment

- Amazon Web Services provides a lot kind of database services.

1. Read documents for “AWS RDS Multi-Availability Zone(Multi-AZ) Deployment”. Tell the Multi-AZ advantages over ordinary RDS. In addition, explain the consideration requirements on Multi-AZ when designing services.
   - https://aws.amazon.com/rds/details/multi-az/
2. Pick one AWS database service other than RDS. Discuss pros and cons of it against RDS Multi-AZ deployment from the viewpoint of Brewer’s CAP theorem.

- Submit your answers in Japanese or in English via the course web.

本日の課題

•Amazon Web Service (AWS)では多数のデータベースサービスを提供している。

1.AWS RDS Multi-Availability Zone(Multi-AZ) Deployment の技術文書を読み、AWS RDS Multi-AZ の通常の AWS RDS に対する優位点、サービス設計に当たって考慮すべき点を述べよ。

•https://aws.amazon.com/rds/details/multi-az/

2.AWS が提供する RDS 以外のデータベースサービスを一つ選択せよ。これと AWS RDS Multi-AZ の優劣を Brewer の CAP 定理の観点から議論せよ。

•講義 Web フォームから記入すること。

Outline

1.Administravia

2.Quiz and homework review

3.Distributed Data Store
  - CAP theorem -

4.Global Services

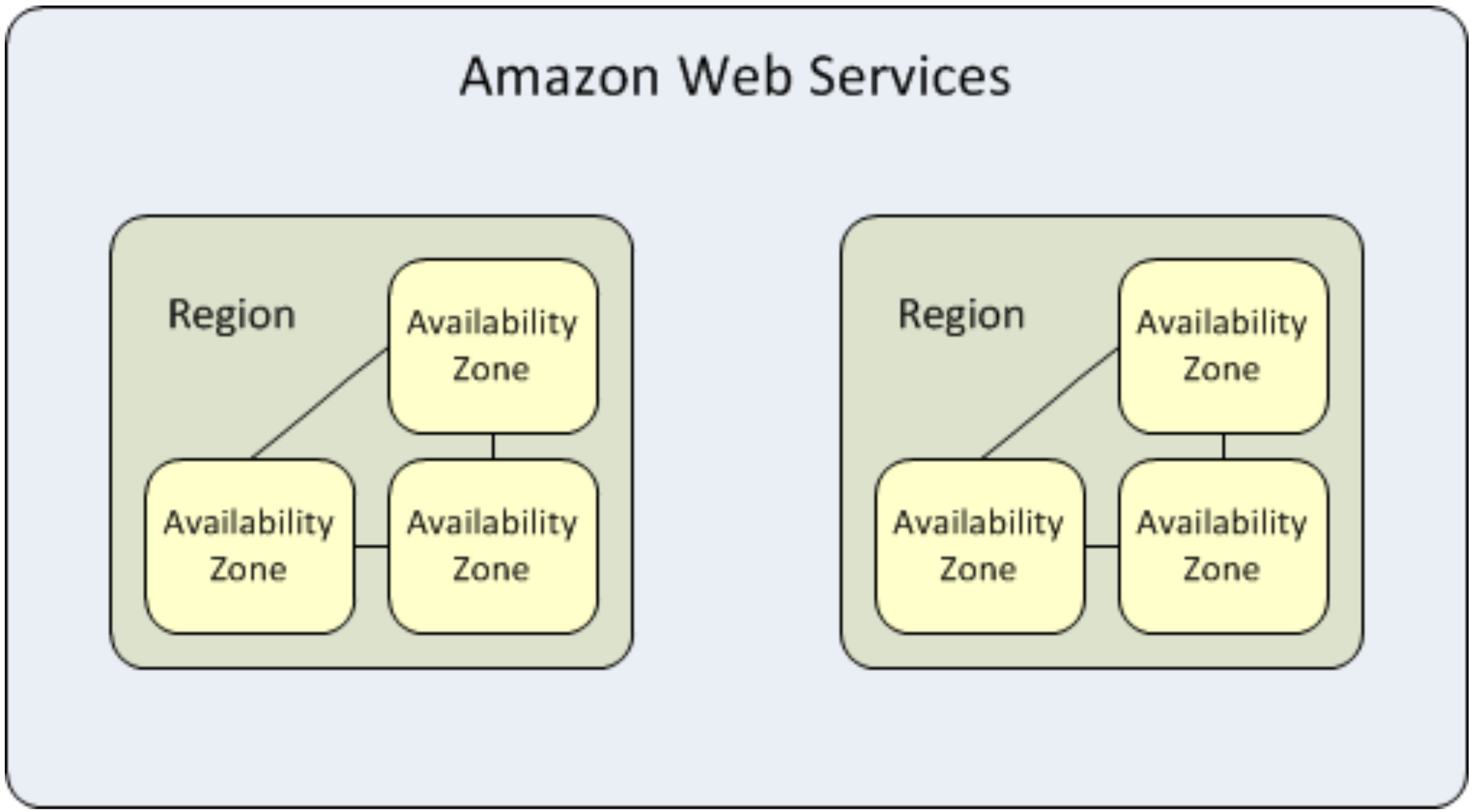
Components of Datacenter (DC)

- Equipments
  - ☑ Server
  - ☑ Network switch, router
  - ☑ Storage
  - ☐ Uninterruptible Power Supply (UPS)
  - ☐ Rack
- Facility
  - ☐ Design: Building vs. Modular
  - ☐ Electric power :
    To connect to high-voltage grid.
    Self power generator.
  - ☐ Environmental :
    Cooling chiller, Air-flow control
  - ☐ Network cabling : Internal, Leased circuit access

http://ishikari.sakura.ad.jp

Courtesy of IIJ

AWS region and zone



Amazon, “Amazon Elastic Compute Cloud: User Guide for Linux”, March, 2015.

超低空の米軍ジェット機

九大工学部に墜落

電算センター全焼

新築中の6階建て

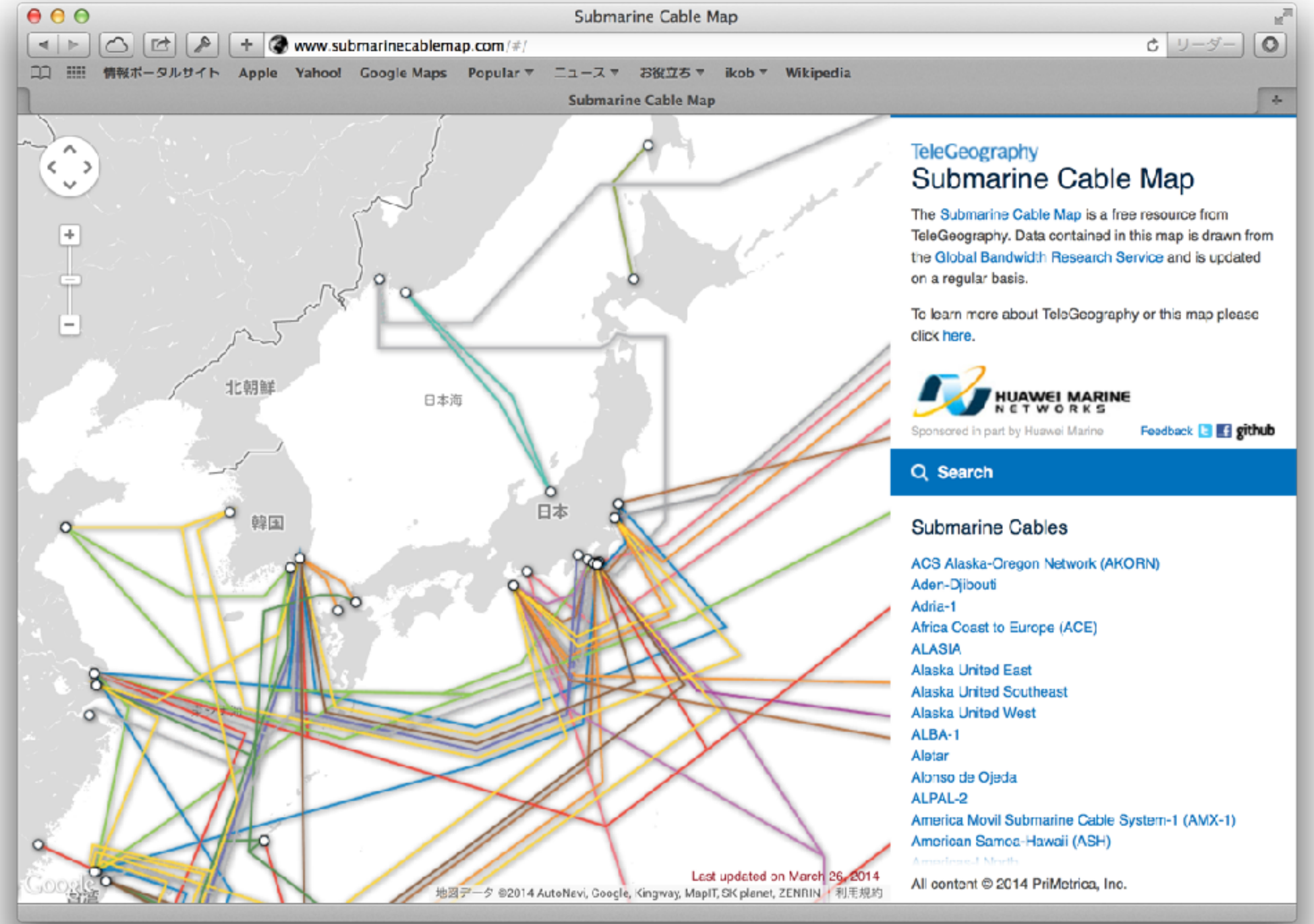
進入路直下の九大

http://www.yomiuri.co.jp/kyushu/50th/shimen/20140827-OYS1T50055.html

Regional : Network connectivity

- Cloud services require reliable network access.
    - Most of cloud services has more than one network connectivities for redundancy.
- Such networks strongly depends on geographical location mainly due to economic factors.
    - Circuit cable systems connect major cities / countries as shorter as possible.
    - Submarine network building cost is $$$. The cost is too much for one company.
    - A limited number of landing points and of paths for submarine cables.

➡Many submarine cables share risks of disaster.

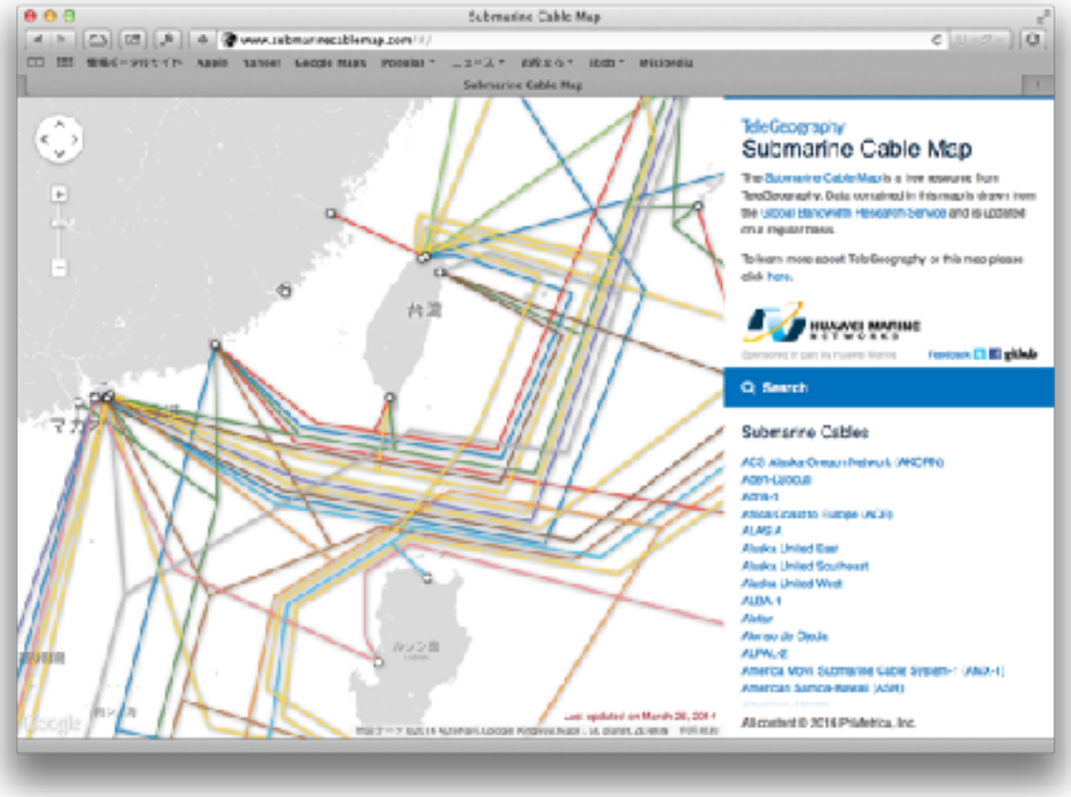

Submarine Cable Map
www.submarinecablemap.com/#/
リーダー
情報ポータルサイト
Apple
Yahoo!
Google Maps
Popular
ニュース
お役立ち
ikob
Wikipedia
Submarine Cable Map
北朝鮮
日本海
韓国
日本
TeleGeography
Submarine Cable Map
The Submarine Cable Map is a free resource from TeleGeography. Data contained in this map is drawn from the Global Bandwidth Research Service and is updated on a regular basis.
To learn more about TeleGeography or this map please click here.
HUAWEI MARINE NETWORKS
Sponsored in part by Huawei Marine
Feedback
github
Search
Submarine Cables
ACS Alaska-Oregon Network (AKORN)
Aden-Djibouti
Adria-1
Africa Coast to Europe (ACE)
ALASIA
Alaska United East
Alaska United Southeast
Alaska United West
ALBA-1
Aletar
Alonso de Ojeda
ALPAL-2
America Movil Submarine Cable System-1 (AMX-1)
American Samoa-Hawaii (ASH)
Last updated on March 26, 2014
地図データ ©2014 AutoNavi, Google, Kingway, MapIT, SK planet, ZENRIN
利用規約
All content © 2014 PriMetrica, Inc.

Submarine cables cut by an earthquake

- A lot of submarine cable systems of South-Asia concentrated on one areas (Luzon strait between Philippine and Taiwan).

- Most systems were damaged at Southern-Taiwan earth quake on Dec. 26 2006.

- Significant impact to Internet connectivity between Japan and Asian countries, especially service for Japanese customers hosted in Hong-Kong.

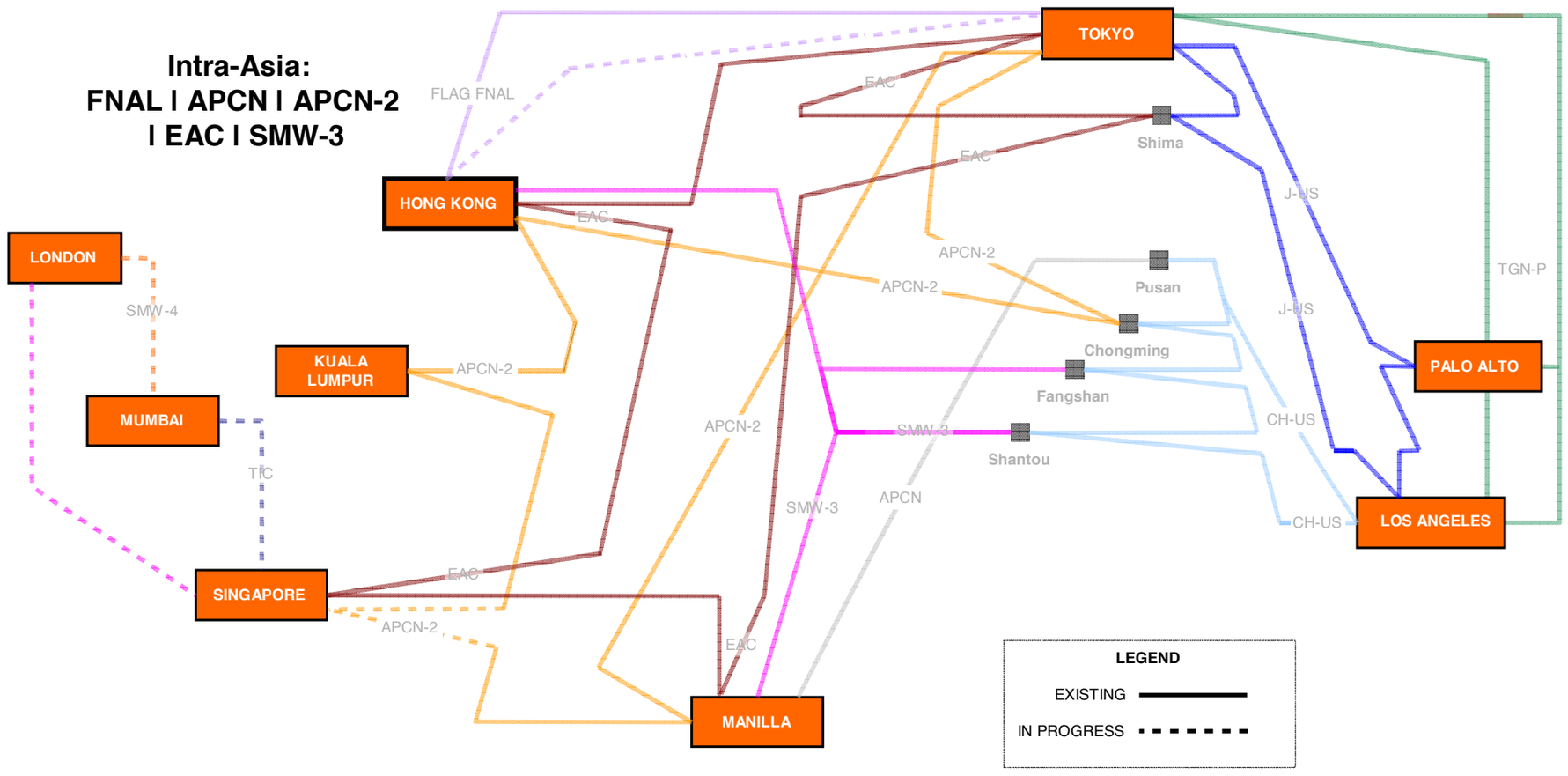


TELEGLOBE

AS6453 Asia Backbone I Physical Routes Diversity

TransPac: C-US I J-US I TGN-P

Intra-Asia:
FNAL I APCN I APCN-2 I EAC I SMW-3



Sylvie LaPerrière ,“Taiwan Earthquake Fiber Cuts: a Service Provider View “, NANOG39, 2007

TELEGLOBE

What makes the Luzon Strait so attractive to cable builders?

Three routes are available to link South East & Northern Asia (Japan-Korea):

1. Luzon Strait between Taiwan & Philippines
   - 320 km width
   - 2600m sill depth in Bashi Channel (north)
2. Route south of the Philippines
   - adds lots of mileage & hence latency
3. Formosa Strait
   - Narrowest part is 130 km width
   - 70 m depth (too close to fishermen)

Given the current market requirements, the Luzon Strait is the best subsea cable route alternative.

Even though many of the cables are **ring-protected**, *both legs* pass through the Bashi Channel (earthquake epicenter) and the cable systems suffered *multiple failures* causing the entire cable system to be **out of service**.

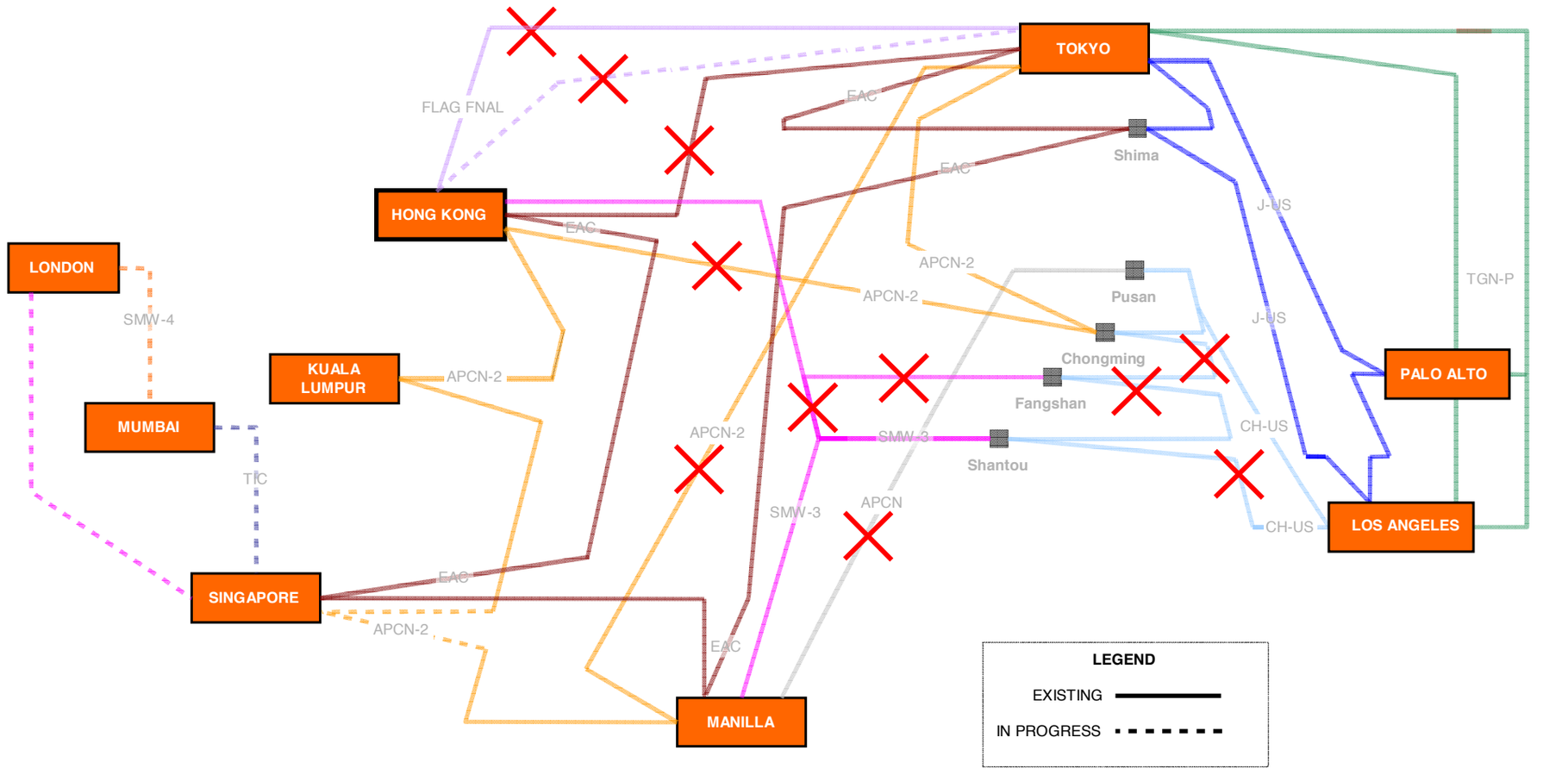
- 9 Cables via Luzon strait: 7 down.

No cables are available to offer restoration. Wait on cable repairs.

Sylvie LaPerrière ,"Taiwan Earthquake Fiber Cuts: a Service Provider View ", NANOG39, 2007

TELEGLOBE

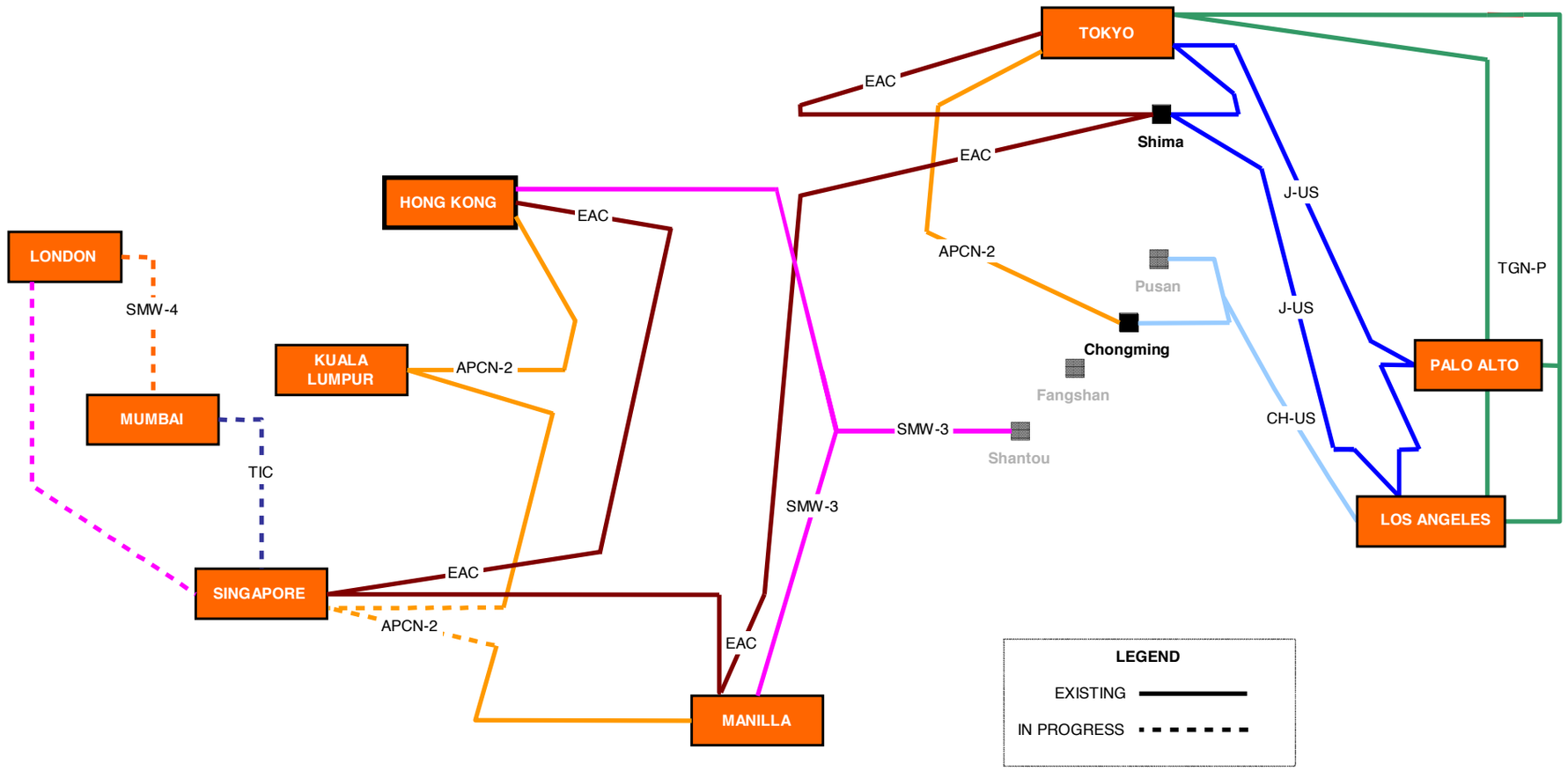
Taiwan Earthquake December 26, 2006 I Cable Faults



Sylvie LaPerrière ,“Taiwan Earthquake Fiber Cuts: a Service Provider View “, NANOG39, 2007

TELEGLOBE

Taiwan Earthquake December 26, 2006 I Remaining Cable Routes



Sylvie LaPerrière ,“Taiwan Earthquake Fiber Cuts: a Service Provider View “, NANOG39, 2007

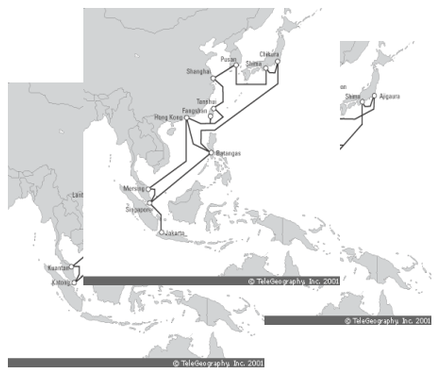
TELEGLOBE

Reparing a subsea cable system I The means…

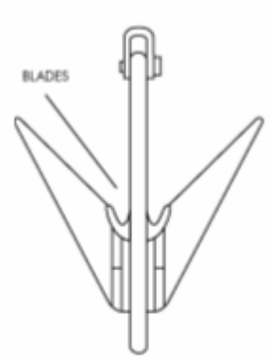
Cable repair ship in the area.

Powerful vessel equipped to maintain station and perform cable repair in rough weather conditions.

All spares, including spare cable, a number of cable bodies and jointing kits.



Not so rough weather



A grapnel fitted with a cutter and a grabbing tool.

45 by 60 cm (18 by 24 in)

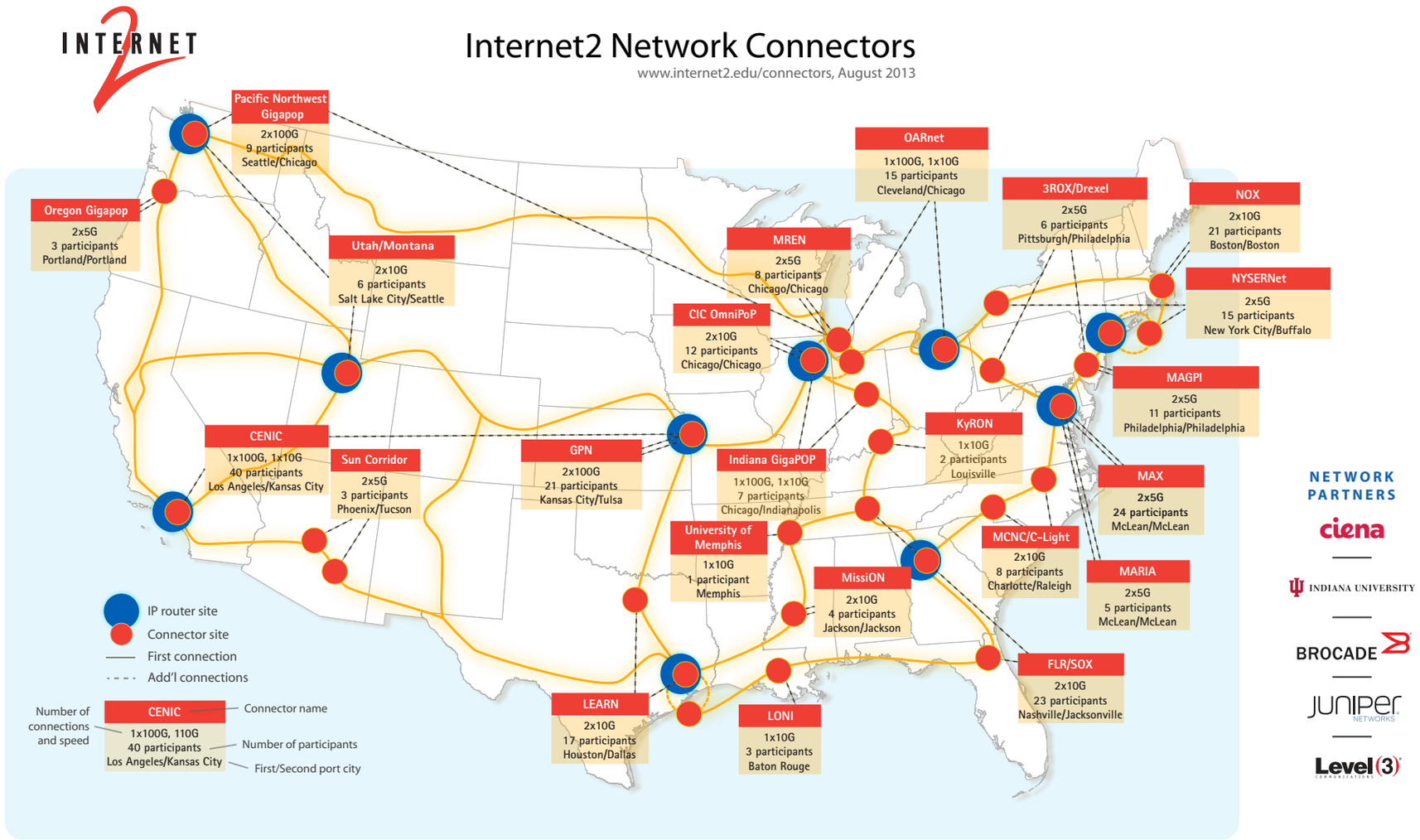
Dropping grapnel + dragging oceanfloor + recover cable = 16 hours

Average repair duration = 7 days

Sylvie LaPerrière ,“Taiwan Earthquake Fiber Cuts: a Service Provider View “, NANOG39, 2007

US R&E Network

- Research and Education network service covering US.
  - Started as “Abilene” network using Qwest circuits.
  - Now as “Internet2 network” using Level3 infrastructure.



Availability Map (05/01-12)

Availability / Disrupt count / Longest down time (sec.)

Regional : impact on applications (To be discussed later class)

- End-to-end network latencies (,or delays) between cloud services and end-systems give significant impacts to user-experiences (UX) and to application performances.
    - Latencies to end applications depend on the location of service.
        - Distance is the biggest factor, such as, 10ms of Tokyo - Osaka, 100ms of Tokyo - LAX, in round-trip-time(RTT).
- Re-locating server is the most effective approach to reduce latencies.
    - Cloud service providers deploy their services to major cities.
    - Contents Distribution Network (CDN) providers deploy their point of presence (PoP) as well.

Components of Datacenter (DC)

- Equipments
  - ☑ Server
  - ☑ Network switch, router
  - ☑ Storage
  - ☐ Uninterruptible Power Supply (UPS)
  - ☐ Rack
- Facility
  - ☐ Design: Building vs. Modular
  - ☐ Electric power :
    To connect to high-voltage grid.
    Self power generator.
  - ☐ Environmental :
    Cooling chiller, Air-flow control
  - ☑ Network cabling : Internal, Leased circuit access

http://ishikari.sakura.ad.jp

Courtesy of IIJ

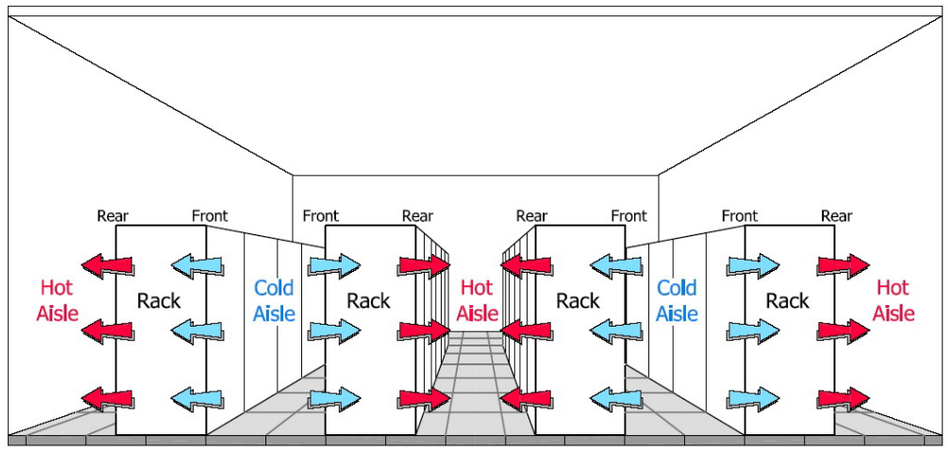
DC Energy efficiency

- One of few differentiation points in DC business.
    - IT equipments, servers and networks are not different from others.
    - Metric: Power Usage Effectiveness (PUE) How much energy is used by non-IT facilities.

      PUE = Total Facility Power / IT Facility Power

      Google : 1.06/1.12@2015, Legacy inefficient DC : 2.0 - 3.0

Figure 1-4. Typical Data Center HVAC Hot Aisle / Cold Aisle Layout



Source: ASHRAE (2004)

- To reduce air conditioning energy/cost is one of the most efficient approach.
    - Alternating cold- and hot-aisle.
    - <u>Open air cooling, Shifting to chiller-less.</u> ⇦ **Prefer cool region**

Report to Congress on Server and Data Center Energy Efficiency Public Law 109-431, Aug. 2007

Today’s quiz

- You want to lure, or invite big datacenter to the city of your birth*.
  Tell the technical advantages of your city for datacenter business.

- Bonus points will be given excellent answers which convince me even with a difficult location.
  (*) If many datacenters are already working in your city, you can adopt “Naha, Okinawa” instead.

- Submit your answers in Japanese or in English via the course web.

本日のクイズ

- 自身の出生地に巨大データセンターを誘致したい*。その都市の技術的優位性を示せ。
- 困難な立地で説得力のある回答には加点する

(*) すでに多くのデータセンターが自身の出生地で稼働している場合は、「那覇、沖縄」に代えても良い。

- 講義 Web フォームから記入すること。

Wikipedia より

Today’s Assignment

- Amazon Web Services provides a lot kind of database services.

1. Read documents for “AWS RDS Multi-Availability Zone(Multi-AZ) Deployment”. Tell the Multi-AZ advantages over ordinary RDS. In addition, explain the consideration requirements on Multi-AZ when designing services.
   - https://aws.amazon.com/rds/details/multi-az/
2. Pick one AWS database service other than RDS. Discuss pros and cons of it against RDS Multi-AZ deployment from the viewpoint of Brewer’s CAP theorem.

- Submit your answers in Japanese or in English via the course web.

本日の課題

•Amazon Web Service (AWS)では多数のデータベースサービスを提供している。

1.AWS RDS Multi-Availability Zone(Multi-AZ) Deployment の技術文書を読み、AWS RDS Multi-AZ の通常の AWS RDS に対する優位点、サービス設計に当たって考慮すべき点を述べよ。

•https://aws.amazon.com/rds/details/multi-az/

2.AWS が提供する RDS 以外のデータベースサービスを一つ選択せよ。これと AWS RDS Multi-AZ の優劣を Brewer の CAP 定理の観点から議論せよ。

•講義 Web フォームから記入すること。

Outline

1.Administravia

2.Quiz and homework review

3.Distributed Data Store
  - CAP theorem -

4.Global Services