CLOUD DATA MANAGEMENT

XUEGANG CHAI
XUEGANG.CHAI@HELSINKI.FI

OUTLINE

- What is the cloud?
- Three characteristics of cloud
- Should transactional data and analytical data both deployed into cloud?

WHAT IS THE CLOUD?

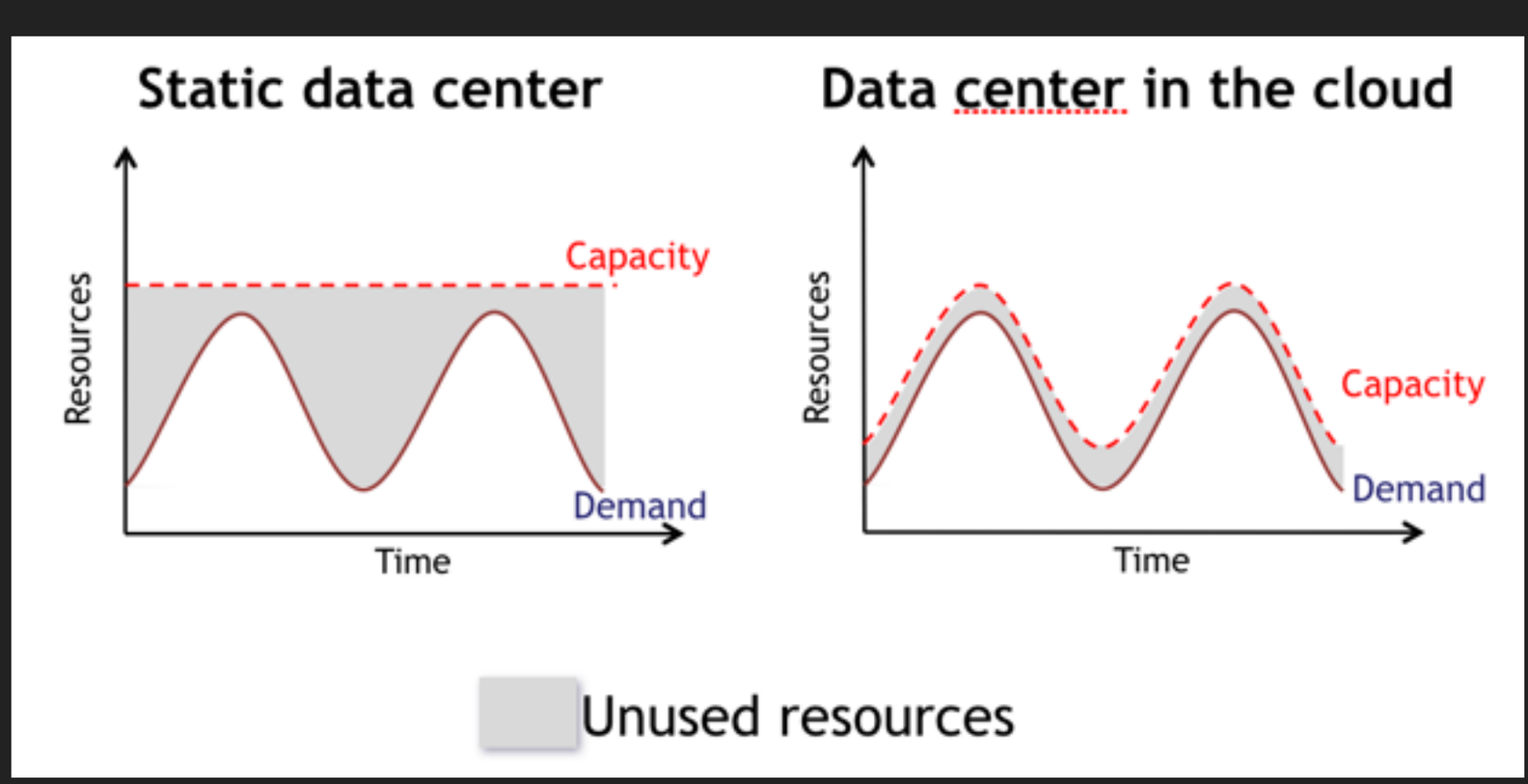
You don’t like making your own cheeseburgers, so you hire a chef to make a cheeseburger for you. You invite all your friends next day for a cheeseburger party. But the chef can’t make 100 cheeseburgers for your 100 friends, he’s too busy! Lots of your friends go hungry! So, the next day you hire another 99 chefs, to make sure everyone can get a cheeseburger. However, that day, half your friends can’t come round, so 50 of the chefs make 50 cheeseburgers for the friends who did come round, and the other 50 chefs stand around doing nothing. But, you still have to pay them! So you’re wasting money.

So, someone comes along and builds a massive kitchen with 2000 chefs, and all the people on your street get rid of their own chefs and kitchen, and instead pay that person to do cheeseburger deliveries for them.

- Reddit

WHAT IS THE CLOUD?

- New trend
- Frees up resources and allows businesses to concentrate on growth
- Cost savings
- Scalability
- Availability



WHAT IS THE CLOUD?

- Cloud Services

THREE CHARACTERISTICS OF CLOUD

1. Computer power is elastic
2. Data is stored at an untrusted host
3. Data is replicated, often across large geographic distances

COMPUTER POWER IS ELASTIC

- Capacity can be scaled up and scaled down quickly by adding or removing server instances (nodes)
- Only if workload is parallelizable
- For example. Amazon EC2
- Shared-nothing architecture are on best to parallelize.

DATA IS STORED AT AN UNTRUSTED HOST

- Risks on storing sensitive data on untrusted host
- Data could be stored in any country, and governed by the law of the country

DATA IS REPLICATED, OFTEN ACROSS LARGE GEOGRAPHIC DISTANCES

- Cloud service providers own data centers all over the world
- CAP theorem: it is impossible for a distributed computer system to simultaneously provide more than two out of three CAP guarantees. (Consistency, Availability, Partition tolerance)

IS CLOUD ONE-FITS-ALL?

- Should transactional data and analytical data deployed into the cloud?

TRANSACTIONAL DATA MANAGEMENT

- Banking, ticket reservation, online e-commerce
- Short and fast inserts and updates initiated by end users, write-intensive, do not typically use a shared-nothing architecture
- Data are fairly small (~1TB)
- Includes sensitive data

TRANSACTIONAL DATA MANAGEMENT

- Transactional data are not appropriate for the cloud
  - Transactional database market is dominated by Oracle, IBM DB2, Microsoft SQL Server and Sybase, none of them can be deployed using a shared-nothing architecture
  - Rely heavily on the ACID guarantees
  - Can not take full advantage of the scalability
  - Risks in storing sensitive data on an untrusted host

ANALYTICAL DATA MANAGEMENT

- Historical data, may come from the various transactional databases
- To help with planning, problem solving, and decision support
- Read-only, good match for the shared-nothing architecture
- Data are typically large (~1PB)

ANALYTICAL DATA MANAGEMENT

- Analytical data is a perfect match for the cloud
  - Read-only, the workloads can easily be parallelized across nodes in a Shared-nothing network
  - Can take the full advantage of scalability
  - Deals of historical data of very large sizes, ACID not needed
  - Sensitive data can be easily anonymized

SUMMARY

- What is the cloud?
- Three characteristics of cloud
- Transactional data management applications are not well suited for cloud deployment
- Analytical data management applications are well suited for cloud deployment

- Source: Daniel J. Abadi: Data Management in the Cloud: Limitations and Opportunities. IEEE Data Eng. Bull. 32(1): 3-12 (2009)

Thank you!