Advanced Topics in DB

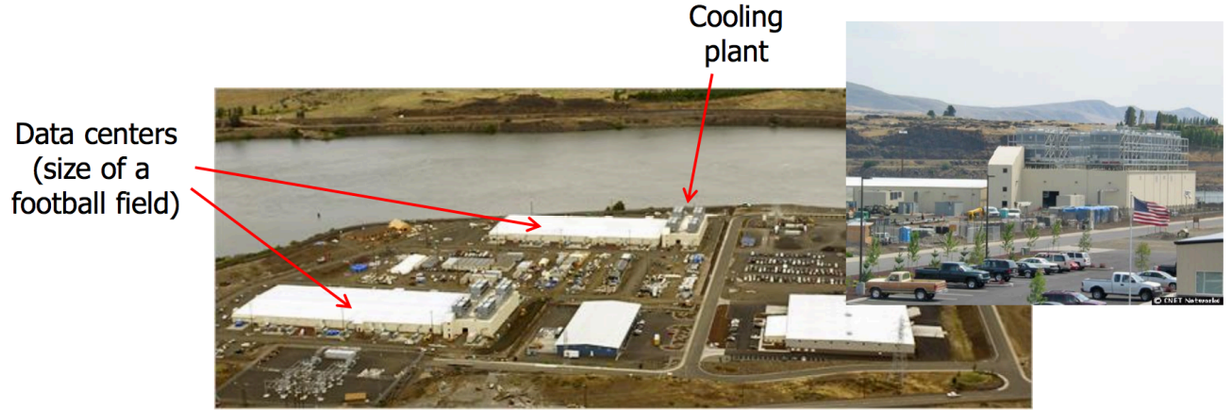
Big Data
Parallel Computing

Recap

- To change big data into value, we need
    - infrastructure that supports
    - computation power
    - analytic tools that works at scale
- Companies own huge amount of data
    - No single machine can store and process this data
    - We need cluster of computers
- Thus, we need new technologies to deal with the huge and the rapid increasing amount of data

Data Center

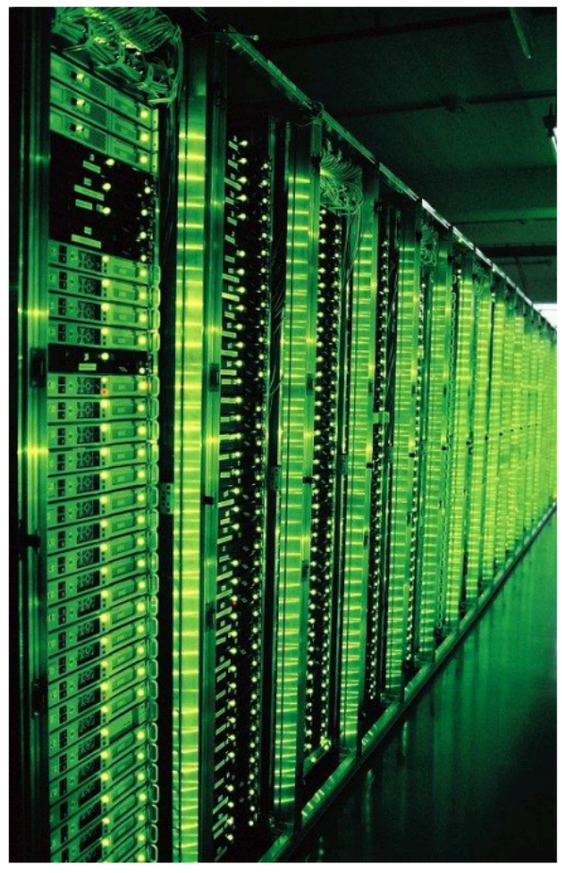
- If the cluster becomes too big to fit in one building
  - build another building
- This will result in the need to have a data center



Google data center in The Dalles, Oregon

What is in the data center?

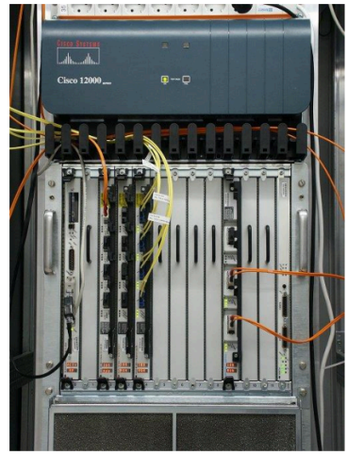
- Hundreds or thousands of racks



Source: 1&1

What is in the data center?

- Massive networking



Source: 1&1

What is in the data center?

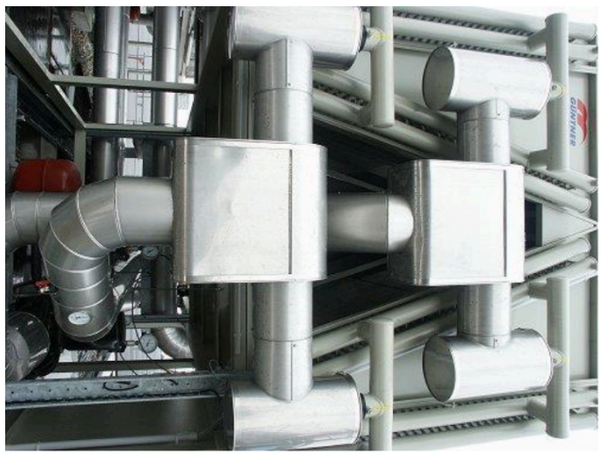
- Emergency power supplies



Source: 1&1

What is in the data center?

- Massive cooling



Source: 1&1

Energy Matters

- Google pays $ 38 M per year, for a data center of 500,000 servers
- build cluster near sources of cheap electricity
- power consumed is turned into heat
  - thus we need cooling, and cooling is a big issue

Global Distribution

- Data centers are often globally distributed
  - need to be physically close to users
  - close to cheaper resources
  - to save for cooling put them in cold areas

Problems with scaling up

- Difficult to dimension
  - load can vary considerably
    - at peak hours load can exceed average load by 2x-10x factor
    - server utilization is 5%-20%
  - waste of resources

Problems with scaling up

- Expensive
  - invest a lot of many on hardware
    - for example, Microsoft invest $ 499 million on one data center
  - need experts
    - setting up cluster is not a trivial task
  - need maintenance

Problems with scaling up

- Difficult to scale
  - scaling up is difficult
    - need to order new machines, install them, integrate them with existing machine
    - major scaling might require redesign; new storage design, new interconnect, new building
  - scaling down is difficult
    - what to do with unused hardware
    - energy is consumed even if the server is not doing work

Summary

- Modern applications require processing huge mount of data
  - measured in petabytes, millions of users
  - need special hardware & algorithms to store, organize, and process the data
- Clusters & data centers may provide the resources we need, but ..
  - they are difficult to dimension, expensive, and difficult to scale

Utility Computing / Cloud Computing

- The cluster/data center challenge is similar to the power plant problem
    - need large up-front investment, expertise to operate, difficulty of scaling up/down
- For the power plant
    - it used to be that every one had their own power resources
    - then it was scaled to use large, centralized power plants with very large capacity
    - usage is metered and customers pay for what they use

Cloud Computing

- The idea of power plant was utilized in the computing power → Cloud Computing

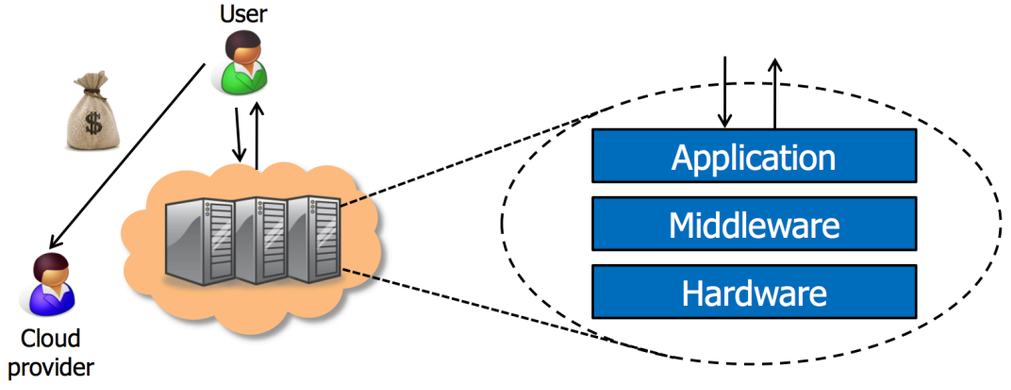
- is a model for enabling convenient, on-demand access to a shared pool of computing resources; networks, servers, storage, applications, and services

- it can be rapidly provisioned (created) and released with minimal management effort

- pay as you go

Everything as a Service

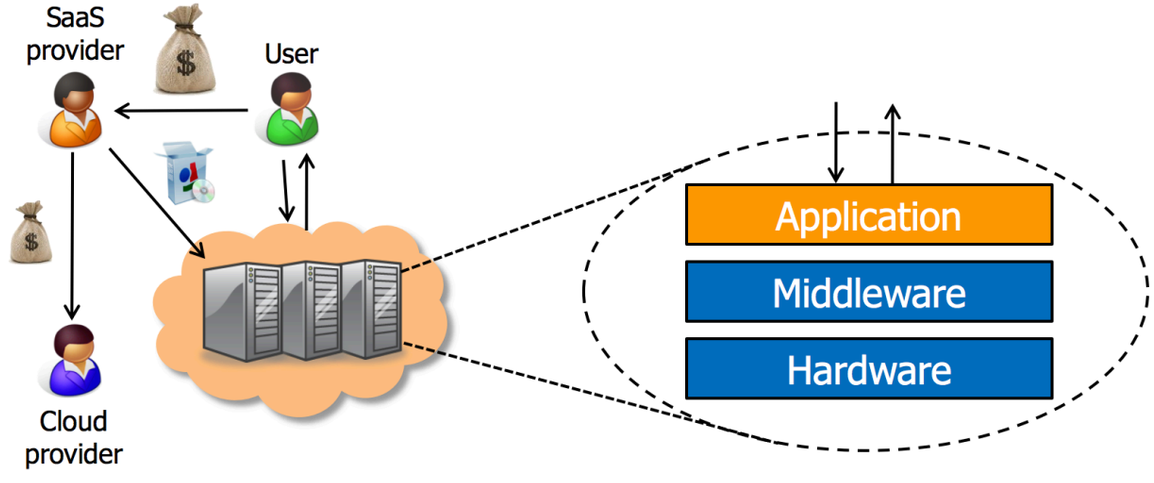
- Software as a Service (SaaS)
  - cloud provides an entire application
  - a third-party provider hosts an application and makes it available for users on the Internet
  - customers pay cloud provider



- Example:
  - **Google Apps**
  - **DropBox**

Everything as a Service

- Platform as a Service (PaaS)
  - cloud provides infrastructure
  - customers pay to PaaS provider, and PaaS provider pay to cloud for the infrastructure



Everything as a Service

- Infrastructure as a Service (IaaS)
  - cloud provides raw computing resources
  - customers pay to IaaS provider, and IaaS provider pay to cloud for the infrastructure
- Amazon EC2

Wrap Up

- Modern applications require processing huge mount of data
  - measured in petabytes, millions of users
  - need special hardware & algorithms to store, organize, and process the data
- Clusters & data centers may provide the resources we need, but ..
  - they are difficult to dimension, expensive, and difficult to scale
- Cloud Computing
  - on-demand access to shared resources on the network, in easy and convenient way

Processing Big Data

- divide & conquer
  - make use of thousands of CPUs
- challenges:
  - **unit of work:** how to split data into partitions
  - **assigning work units to workers:** what if number of units > # of workers
  - **aggregating & combining results:** aggregate results from all workers
  - **coordination & communication between workers:** if a worker needs data from other workers
  - **tracking of workers progress:** how do we know when workers finish
  - **managing work failure:** what will happen if a worker die

Parallel Precessing Problems

- All of the previous challenges are common problems in parallel computing
    - Work load balancing
    - communication between workers
    - access to shared locations, for example accessing the same data
- we need a synchronization mechanism or ..

Source: Ricardo Guimarães Herrmann

Ideas behind the development of Big Data Technologies

- Hide system level details from the developers
  - no need to worry about, reliability, fault tolerance, data partitioning, assigning tasks to workers
- Separating the how from the what
  - developers specify computation power they need
  - execution framework handle that
- MapReduce was the first instantiation of these ideas (Next)