How can we crunch the massive amount of data? => cloud data center
(store and process big datas)

-> 94% global workloads was processed in datacenter (computation)

elastic computing cloud (EC2).

Read papers.
The first pass (5~10min): title, abstract, introduce, conclusion.
category, context (which other paper is it related to), correctness, contribution (是否 signifi), clarity.
The second pass (1 hour): content, no details
× proof and implementation ✓ figures, diagrams. ? convincing -> conclusion
The third pass (several hours and days) technical details.
interested in and have time.

Commodity server --(You 80.)--> rack --> cell (独立供电).

a GPU server : memory
↕ CPU
PCIE bottle neck
GPU ⊃ NVLink RDMA used to connect to remote GPUS bypass CPUS

inter-rack: optical circuits switch. based programmable network

spine
Spine-leaf arch: ~~aggregation~~ switch. + distribution switch (fat tree). 每一层带宽不同，靠近根部带宽增加.

GPUs have dedicated interfaces with rack switches which can be used to connect with remote GPUs bypass CPU

users → network → border router → cluster router → Layer 2 switch

→ rack switch

Inter-DC WAN is very slow.

(普遍存在)
ubiquitous, convenient, on-demand network access.
on-demand computing, delivered over the Internet with pay as you go pricing.
Infrastructure as software.

Lecture 2.

For most of users the best way to deploy large scale internet services is just to deploy on cloud.

VS

Economies of scale. | statistical multiplexing | profitable business.

Cloud pricing

Computing (EC2), Storage (S3, EBS; tiered pricing 阶梯式的). Data transfer (inbound free, outbound ×) data gravity price.

1. Reserved pricing: cost(t) = U + discount × R × t

Guaranteed availability ↓ 一次性费用 ~ 按需价格的常规费用

2. Spot pricing.

user's bid > spot price, win until a new one is posted. ⇒ running users with a lower bid get their instances terminated.

× service guarantee.

e.g. machine learning

on demand service guarantee: if instance is launched, it will not be terminated by platform

It is possible the price of (on demanding) is lower than (spot), because cloud privide want to

kick out all spot users.

* brokerage service. 中间件服务. : make instance acquisition strategies.

private cloud can be cheaper than public cloud e.g. GPU cloud.

public cloud

hybrid cloud.

Cloud computing stack

application.

platform

Infrastructure e.g. AWS

virtualization

Hardware

Storage network

cloud service model

flex ↓ base touse.

Infrastructure as a Service: VMs [computation, storage, network].

platform : SDK. API. automatic scalability; × control of OS etc.

software : office365

Other : function: write apps in the form of cloud functions and define events [trigger only when functions running.]

machine learning: pipeline & tools)

model : model api (chatgpt).

e.g. AWS Lambda.

Issues of cloud.

Availability

data loss

Vendor lock in : lock into the current provider.

⇓

sky

security

privacy.

challenge: storage, scale. faults failures. networking