Cloud Computing Paradigm Shift

Jan Šedivý

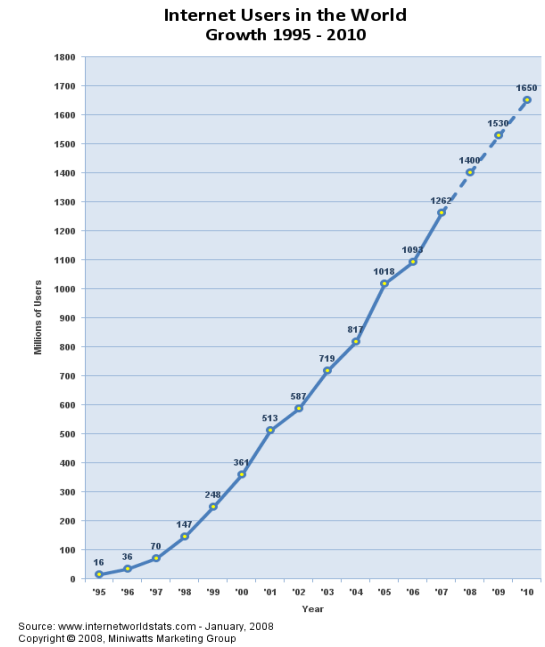
Computing Evolution

2000 Cloud computing

1990 Internet

1980 IBM PC

1970 IBM 370



Source: www.internetworldstats.com - January, 2008

Connect Anywhere, Anytime

Cloud, Web-based applications give users full access to their information **anywhere, anytime**,

- Windows, OS X, Linux
- FireFox, IE, Chrome, Safari …
- mobile phone, tablet computer

The data is stored in the cloud – not on employees computers
Multiple users can access and contribute to projects simultaneously

Employees need only Internet connection.

The name cloud computing was inspired by the cloud symbol that's often used to represent the Internet in flow charts and diagrams.

Central administration

- Applications accessed through browser -
  - no workstation apps to license,
  - no configuration, no personalization
  - no patches,
  - mobile access
- Cloud central administration - fast distribution of new functionality and upgrades
- New apps are complex and expensive
- Employees easier adapt to small improvements than to disruptive changes.
- Traditional software - long learning cycle
- Cost advantage, no upfront payment, pay as you go, Freemium

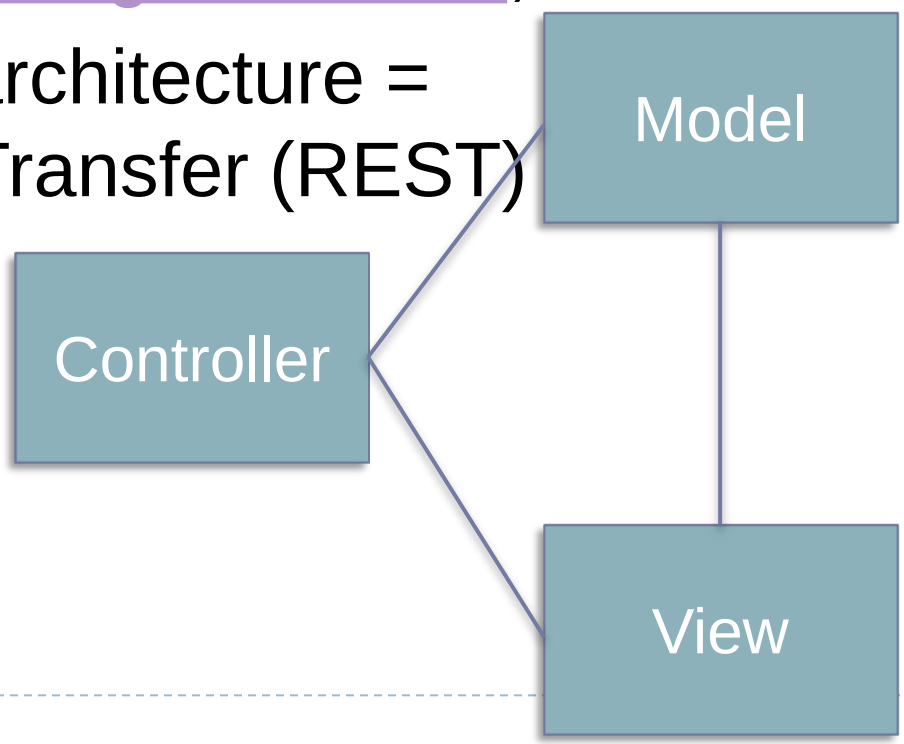
Cloud applications

- Storage – backup
- General number crunching applications
- Web applications - services

  - Autonomic computing
  - Client–server model
  - Grid computing
  - Mainframe computer
  - Utility computing
  - Peer-to-peer
  - Service-oriented computing

WEB applications - Architecture

- multiple *cloud components* communicating with each other
- 3-tier architecture. Architectonical pattern MVC model
- API application programming interfaces,
- web services software architecture = Representational State Transfer (REST)
- View, interaction – browser, native application, different platforms



Cloud architecture

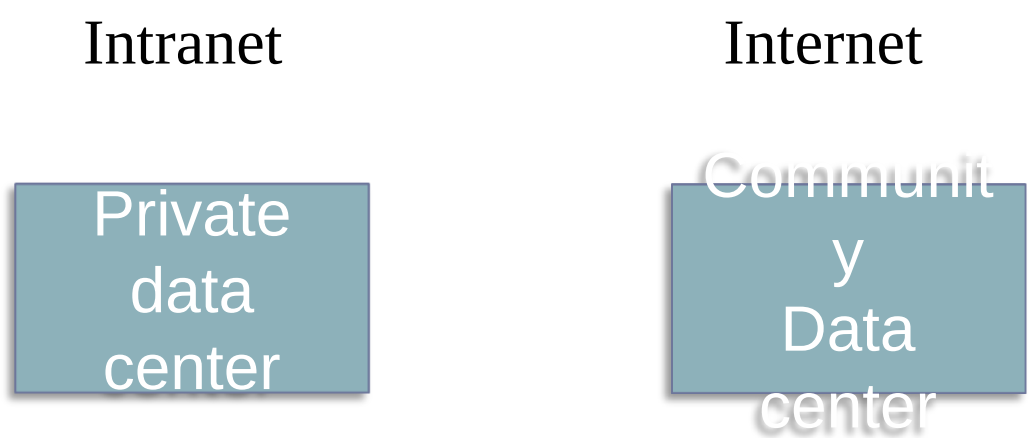
- Client
  - any browser, native app
- Software as a Service – SaaS
  - Google,
  - Microsoft
- Platform as a Service – PaaS
  - Google app engine
  - Microsoft Azure
- Infrastructure as a Service – IaaS
  - Amazon AWS,
  - Rack space

Deployment models

- Public cloud – internet

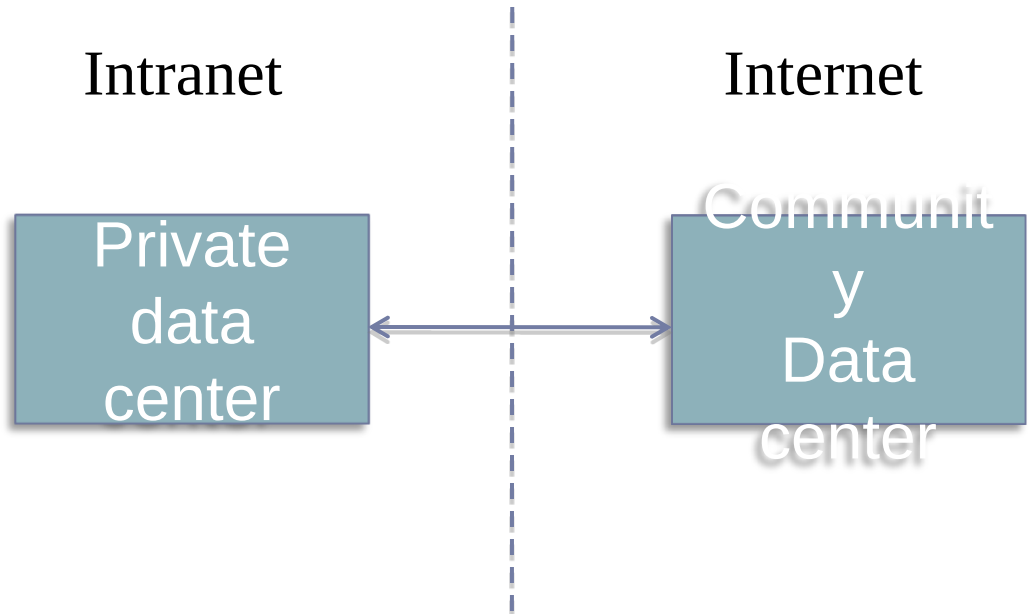
Deployment models

- Public cloud – internet
- Private cloud – intranet



Deployment models

- Public cloud – internet
- Private cloud – intranet
- Hybrid cloud – internet + intranet
- Community cloud – intranet



Higher Reliability and Uptime

- Service Level Agreement (SLA)
- Disaster recovery plan
- Recovery point objective (RPO) design target is zero
- Recovery time objective (RTO) design target is instant fail over

- Disaster recovery plan – 2-4% of IT budget
- data replication - build in redundancy
- fall back to secondary data center
- 99.9% uptime guarantee
- this level of reliability with on-premises or hosted technology is tremendously costly and complex

Example: If a system running perfectly as at 10:00am and the system down at 10:01am. The system later up at 01:00pm. At this point the system RPO is 10:00am. Although, if the RTO stated as 4hours, but since the system was up at 01:00pm (10am to 1pm = 3 hours) the RTO stated as RTO achieved in 3hours and RPO is 10am where the system need to recover 10am data.

What is the size of a data center

- Small server in the cellar – small business data server
- Medium company 50 servers?
- Large company 1000 servers?
- Google, AWS, … 10^6?
- Elasticity – appearance of infinite computational power and memory

Economy of scale

- Cloud data centers are built from computer with minimum number of parts in massive quantities,
- Centers with thousands of servers allow efficient ratio of staff to machines.
- On a per-user basis, these economies of scale allow higher levels of efficiency than can be achieved by SMBs.
- Electricity cost is decreased 5-7 times in a large center (5MW)
- The most commonly used metric to determine the energy efficiency of a data center

$$\mathrm{PUE} = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$$

Public - private center

Advantage	Public cloud	Private cloud
Appearance of infinite computing resources	yes	no
Elimination of up front payment by cloud user	yes	no
Pay as you use	yes	no
Economies of scale due to very large center	yes	Usually not
Higher utilization by multiplexing resources	yes	Depends on size

Security

- Google, Amazon, Sun, IBM and other major cloud players investment in physical and process-based security.
- SAS 70 Type II audit
  - **Logical security**: reasonable assurance of providing access to authorized individuals only
  - **Privacy**: reasonable assurance of data privacy, implement proper policies and procedures
  - **Data center physical security**: good protection of data centers and corporate offices
  - **Incident management and availability**: reasonable assurance data centers and applications are redundant and incidents are properly reported, responded to, and recorded
  - **Change management**: reasonable assurance that development and changes are properly tested
- Browser-based applications do not need to save sensitive data on local devices.
- Lost laptops and memory stick minimize the amount of sensitive data stored on these devices
- Regular staff training and educating in all aspects internet safety and security is must.

Tier	Requirements
1	•Single non-redundant distribution path serving the IT equipment •Non-redundant capacity components •Basic site infrastructure guaranteeing 99.671% availability
2	•Fulfills all Tier 1 requirements •Redundant site infrastructure capacity components guaranteeing 99.741% availability
3	•Fulfills all Tier 1 & Tier 2 requirements •Multiple independent distribution paths serving the IT equipment •All IT equipment must be dual-powered and fully compatible with the topology of a site's architecture •Concurrently maintainable site infrastructure guaranteeing 99.982% availability
4	•Fulfills all Tier 1, Tier 2 and Tier 3 requirements •All cooling equipment is independently dual-powered, including chillers and Heating, Ventilating and Air Conditioning (HVAC) systems •Fault tolerant site infrastructure with electrical power storage and distribution facilities guaranteeing 99.995% availability

Data center requirements

Tier	Requirements
1	•Single non-redundant distribution path serving the IT equipment •Non-redundant capacity components •Basic site infrastructure guaranteeing 99.671% availability
2	•Fulfills all Tier 1 requirements •Redundant site infrastructure capacity components guaranteeing 99.741% availability
3	•Fulfills all Tier 1 & Tier 2 requirements •Multiple independent distribution paths serving the IT equipment •All IT equipment must be dual-powered and fully compatible with the topology of a site's architecture •Concurrently maintainable site infrastructure guaranteeing 99.982% availability
4	•Fulfills all Tier 1, Tier 2 and Tier 3 requirements •All cooling equipment is independently dual-powered, including chillers and Heating, Ventilating and Air Conditioning (HVAC) systems •Fault tolerant site infrastructure with electrical power storage and distribution facilities guaranteeing 99.995% availability

Utilization of servers