Core Technologies for Cloud Computing and Future Internet

Prof. Kai Hwang
University of Southern California

This talk attempts to address a few critical issues on cloud computing technology, environment, and emerging new applications in business, science, and society.

- What are the desired future Internet architecture and protocols to achieve much greater bandwidth, efficient packet routing, and enhanced security than today’s Internet? How to use clouds to boost the global economy?
- What roles clouds can play with a healthy cloud ecosystem to realize the Internet of Things (IoT), to secure pervasive datacenters, and to make the social networks trustworthy?

The Big Switch in Early 21st Century

- Industries race to build massive datacenters (capacity)
- High bandwidth network and pervasive connectivity
  - Terabit network in sight
  - Wideband wireless mobility
- Virtualization help realize the economics of scale

Some IT Trends

- The Data Deluge is a clear trend from Commercial (Amazon, e-commerce), Community (Facebook, Search) to Scientific applications
- Light weight clients from smartphones, tablets to sensors
- Multicore reawakening parallel computing
- Exascale initiatives will continue drive to high end with a simulation orientation
- Clouds with cheaper, greener, easier to use IT for (some) applications
- New jobs associated with new curricula
  - Clouds as a distributed system (classic CS courses)
  - Data Analytics (Important theme at SC11)
  - Network/Web Science

What are the Cloud Computing Services?
(Software, Platform, Infrastructure, Data, Storage, Compute, …) as a Service (SaaS, PaaS, IaaS, Daas, SaaS, CaaS, …)

Public Cloud

Clients

Cloud Manager

Private Cloud

Other Cloud Services

Govt. Cloud Services
Cloud Applications

Science and Technical Applications

Business Applications

Consumer/Social Applications

Network Cloud Services

Today’s Cloud Services Stack

Application Cloud Services

Cloud Services

Platform Cloud Services

Compute & Storage Cloud Services

Co-Location Cloud Services

Network Cloud Services

Cloud Business Potential:
a trillion $ business/year by 2020?

Distributed and Cloud Computing

Kai Hwang, Geoffrey Fox, Jack Dongarra,
published by Morgan Kaufmann, Oct. 2011
(648 pages)
Cloud Computing dealing with too many issues yet to be resolved by ACCTA

Virtualization
QoS
Resource Metering
Scalability
Pricing
Reliability
Billing
Energy Efficiency
Utility & Risk Management
Provisioning on Demand
Service Level Agreements
Security
Legal & Regulatory
Trust
Programming Env. & Application Dev.
Software Eng. Complexity

The Google Gmail Example
• Clouds win by efficient resource use over datacenters

<table>
<thead>
<tr>
<th>Business Type</th>
<th>Number of users</th>
<th># servers</th>
<th>IT Power per user</th>
<th>PUE (Power Usage effectiveness)</th>
<th>Total Power per user</th>
<th>Annual Energy per user</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>50</td>
<td>2</td>
<td>8W</td>
<td>2.5</td>
<td>20W</td>
<td>175 kWh</td>
</tr>
<tr>
<td>Medium</td>
<td>500</td>
<td>2</td>
<td>1.8W</td>
<td>1.8</td>
<td>3.2W</td>
<td>28.4 kWh</td>
</tr>
<tr>
<td>Large</td>
<td>10000</td>
<td>12</td>
<td>0.54W</td>
<td>1.6</td>
<td>0.9W</td>
<td>7.6 kWh</td>
</tr>
<tr>
<td>Gmail (Cloud)</td>
<td>∞</td>
<td>∞</td>
<td>&lt; 0.22W</td>
<td>1.16</td>
<td>&lt; 0.25W</td>
<td>&lt; 2.2 kWh</td>
</tr>
</tbody>
</table>

Challenges/Issues in Cloud Computing

Q: Rate the challenges/issues ascribed to the 'cloud/on-demand model
(1=not significant, 5=very significant)

<table>
<thead>
<tr>
<th>Challenge</th>
<th>% responding 4 or 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security</td>
<td>74.4%</td>
</tr>
<tr>
<td>Performance</td>
<td>63.1%</td>
</tr>
<tr>
<td>Availability</td>
<td>63.1%</td>
</tr>
<tr>
<td>Hard to integrate with in-house IT</td>
<td>61.1%</td>
</tr>
<tr>
<td>Not enough ability to customize</td>
<td>55.8%</td>
</tr>
<tr>
<td>Worried on-demand will cost more</td>
<td>50.4%</td>
</tr>
<tr>
<td>Bringing back in-house may be difficult</td>
<td>50.0%</td>
</tr>
<tr>
<td>Regulatory requirements prohibit cloud</td>
<td>49.2%</td>
</tr>
<tr>
<td>Not enough major suppliers yet</td>
<td>44.3%</td>
</tr>
</tbody>
</table>

Source: IDC Enterprise Panel, August 2018, n=244

What Applications Work better in The Cloud?

• Workflow and Services
• Pleasingly parallel applications of all sorts analyzing roughly independent data or spawning independent simulations including
  – Long tail of science
  – Integration of distributed sensor data
• Science Gateways and portals
• Commercial and Science Data analytics that can use MapReduce (some of such apps) or its iterative variants (most analytic apps)
• Data Analysis requirements not well articulated in many fields – See [http://www.delsall.org for life sciences](http://www.delsall.org)
Example 1: Game Cloud
built at USC
GamePipe Lab,
No.1 Game program
among the to 10
in the USA

Reduction of Gaming Latency
and Improvement of the QoS and QoE (Frame Rate)
on The Game Cloud at USC GamePipe Lab. 2012

Major Technological Challenges in Developing the Future Internet
- Programmable Networking Architectures
- Fusion of The Internet, Mobile and TV Networks
- Named Data Networking beyond the TCP/IP
- Personalized communication protocols
- Intelligent Routing and Content Distribution
- New Ideas for Security and Privacy Protection
- Service migration and disaster recovery
- Massive data protection beyond encryption

Three Approaches To Building The Future Internet Architecture
- OpenFlow for Programmable Virtual Networking
  (Stanford, Princeton, etc., 2008)
- Content-Centric Networking (CCN)
  : Named Data Networking
  (HP Lab, UCLA, etc. 2009)
- Service-Oriented Future Internet Architecture
  (SOFIA) : Chinese Academy of Sciences, Institute of Computing Technology (2011)
Cloud Ecosystem Requirements:

- At the system level, the cloud ecosystem include the cloud platform and infrastructure, resource management policies, etc.
- At the service level, the SLAs, globalized standards, reputation system, billing and accounting system, cloud business models, etc.
- At the user (client) level, Application programming interfaces (APIs), cloud programming environment, Quality of Service control, etc.

Cloud is built on Massive Datacenters

Range in size from “edge” facilities to megascale (100K to 1M servers)

Economies of scale
- Approximate costs for a small size center (1K servers) and a larger, 400K server center.

<table>
<thead>
<tr>
<th>Technology</th>
<th>Cost in small-sized Data Center</th>
<th>Cost in Large Data Center</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network</td>
<td>$95 per Mbps/ Month</td>
<td>$13 per Mbps/ month</td>
<td>7.1</td>
</tr>
<tr>
<td>Storage</td>
<td>$2.20 per GB/ Month</td>
<td>$0.40 per GB/ month</td>
<td>5.7</td>
</tr>
<tr>
<td>Administration</td>
<td>~140 servers/ Administrator</td>
<td>&gt;1000 Servers/ Administrator</td>
<td>7.1</td>
</tr>
</tbody>
</table>
Ecosystem for Market-Oriented Clouds


Autonomic Cloud Management

Develop methodologies and tools to automate the process of cloud management in 4 objectives

1. Manage resources to provisioning of service quality assurance and adaptation
2. Automate the configuration process of VMs and virtual clusters
3. Manage energy consumption under SLA constraints
4. Develop fault prediction models for proactive failure management

MapReduce “File/Data Repository” Parallelism

Instruments
Map = (data parallel) computation reading and writing data
Reduce = Collective/Consolidation phase e.g. forming multiple global sums as in histogram

Example 2: MapReduce Skyline Composition of Web Services in Inter-Cloud Applications

(Courtesy of L. Chen, K. Hwang, and J. Wu, Jan. 2011 [6])
Reduction of Web Service Composition Time from 929 ms to 220 ms using fewer Skyline representatives

Amazon’s Lesson

- Down for 3 days since 4/22/2011
- 1000x of businesses went offline. e.g. Pfizer, Netflix, Quora, Foursquare, Reddit,…
- SLA contract
  - 99.95% availability (< 4.5 hour down)
  - 10% penalty, otherwise

Architecture of The Internet of Things

Cloud Support in Internet of Things and Social Network Applications

1. Smart and pervasive cloud applications for individuals, homes, communities, companies, and governments, etc.
2. Coordinated calendar, itinerary, job management, events, and consumer record management (CRM) services
3. Coordinated word processing, on-line presentations, web-based desktops, sharing on-line documents, datasets, photos, video, and databases, content distribution, etc.
4. Deploy conventional cluster, grid, P2P, social networking applications in the cloud environments, more cost-effectively.
5. Earthbound applications that demand elasticity and parallelism to avoid large data movement and reduce the storage costs
Example 3: Internet of Things: Sensor Grids
A pleasingly parallel example on Clouds

A sensor (“Thing”) is any source or sink of time series
In the thin client era, smart phones, Kindles, tablets, Kinects, web-cams are sensors
Robots, distributed instruments such as environmental measures are sensors
Web pages, Googledocs, Office 365, WebEx are sensors
Ubiquitous Cities/Homes are full of sensors
They have IP address on Internet
Sensors – being intrinsically distributed are Grids
However natural implementation uses clouds to consolidate and control and collaborate with sensors
Sensors are typically “small” and have pleasingly parallel cloud implementations

Research on Cloud Security

Users’ privacy may be leaked while their data are stored in the data center or processed in the cloud. It’s very important to ensure users’ data security during the whole life cycle of the data.

Data Security
The virtualization technology improves the resource utilization in cloud platform, but we need an effective isolation between VMs.

Resource Isolation

Cloud Security Services

Trusted Execution
As cloud users lose control of their resources, it’s necessary to ensure the security of tasks and data processed in the cloud.

Monitoring and Management
As the cloud platform provides services for many users at the same time, it must carry out effective monitoring of user behavior to prevent users from attacking the cloud.

Cloud Services Reliability
In order to provide cloud users uninterrupted services, we must ensure service reliability in cloud computing platforms.

Technology Integration

Cloud Security Management System / Cloud Based Proactive Defense System

Trusted Zones for VM Insulation

Federate identities with public clouds
Identity federation

Data loss prevention
Encryption & key mgmt
Tokenization

Insulate infrastructure from Malware, Trojans and cybercriminals
Anti-malware
Cybercrime intelligence
Strong authentication

Physical Infrastructure
Enable end to end view of security events and compliance across infrastructures
GRC

Insulate information from other tenants
Insulate information from cloud providers’ employees

Access Mgmt
Segregate and control user access

Security Info. & Event Mgmt
(Courtesy of Dr. L. Nick, EMC 2008)
Clouds and Grids/HPC

- Synchronization/communication Performance
- Grids > Clouds > HPC Systems
- Clouds appear to execute effectively Grid workloads but are not easily used for closely coupled HPC applications
- Service Oriented Architectures and workflow appear to work similarly in both grids and clouds
- Assume for immediate future, science supported by a mixture of
  - Clouds – data analysis (and pleasingly parallel)
  - Grids/High Throughput Systems (moving to clouds as convenient)
  - Supercomputers (“MPI Engines”) going to exascale

To Probe Further:

- Geoffrey Fox: Grid of various clouds -- from Raw Data to Wisdom.
- SS = Sensor service, fs = filter services

Conclusions:

- At present, cloud technology is still far from being mature and user-friendly. Cloud industrialization demands first to upgrade of our computer science/engineering education significantly.
- Interesting IoT applications must leverage the clouds to store and process massive amounts of data, that are changing rapidly on a daily basis. With storming and distrustful clouds, the IoT will be a just dream which will never become true.
- Clouds, IoT and social networks are rapidly changing our world, reshaping all human relationships, affecting our daily life, and impacting the global economy and political system. We have to be ready to face the new challenges, like it or not!

References:

3. G. Xie, et al. “Service-Oriented Future Internet Architecture (SOIFA)” IEEE Infocom, Shanghai, China, April 2011