

Cloud Computing and the Future of Internet Services

Wei-Ying Ma

Principal Researcher, Research Area Manager
Microsoft Research Asia

Computing as Utility

Grid Computing

Web Services in the Cloud

What is Cloud Computing?

Software as a Service (SaaS)

雲計算

雲端計算

Platform as a Service (PaaS)

網雲計算

Sky Computing

Comparison

- Grid Computing

- Refer to resource-pooled environments for running compute jobs (like image processing) rather than long running processes (such as a Web site or e-mail server)

- Utility Computing

- Refer to resource-pooled environments for hosting long running processes, and tends to be focused on meeting **service levels** with the optimal amount of resources necessary to do so

- Cloud Computing

- Refer to a variety of **services** available over the Internet that deliver compute functionality on the service provider's **infrastructure**
- Its environment (**infrastructure**) may actually be hosted on either a grid or utility computing environment, but that doesn't matter to a service user
- The **data** in the cloud, as “Intel inside” (or intelligence inside), is often an important part of the services

Cloud Computing – My Simple Definition

Cloud Computing

= **Software** as a Service

+ **Platform** as a Service

+ **Data** as a Service

+ **Infrastructure** as a Service

Two Important Aspects for Cloud Computing

- Operate Internet-based services in the cloud
 - Host and run high-level web apps such Hotmail, Search, Virtual Earth, Office Online, Xbox Live
 - “software as a service”
- Offer an **Internet-based platform to developers** who want to create services but don't have their own cloud to run them on
 - Rent storage, computation, and maintenance (datacenters) from someone else
 - **“platform as a service”** and **“infrastructure as a service”**

Cloud Computing

- **Software as a Service (SaaS)**
 - From user's point of view
 - Apps are located in the cloud (i.e. somewhere in the Web)
 - Software experiences are delivered through the Internet

Cloud Computing

- Platform as a Service (PaaS)
 - From developer's point of view
 - Similar to Windows Platform for ISV developers in the PC ecosystem
 - Computing as utility
 - Cloud computing providers builds and maintains the infrastructure and provides the cloud platform for developers to build applications and end-user services
 - Companies that build end-user services can buy time/resources on this infrastructure

A Hypothetical Scenario

- “The Next Facebook”
 - What platform will be chosen by three Harvard students for their internet service idea?

- They will need to

- Design web pages
- Develop applications (database, business logic)
- Find contents and users

Application

- Buy machines
- Find a place to host machines
- Network

Platform issues

A Service to Solve Platform Issues

- Should be cheap and easy to get up and running
- Should provide all essential building blocks
 - e.g., computation, structured storage, network, load balancing, service management
- Should scale seamlessly
 - e.g., storage capacity, blob size, bandwidth, computation...
- Should be highly available
 - 99.99% availability for all components
- Should provide rich development environment
 - “Visual Studio for Services” that provides end-to-end developer support

Cloud Computing

- **The data in the cloud**
 - The large amount of data available on the web form the “engine” for new Web applications and Internet services
 - The “meshes” created by various services and applications further “interconnect” the data
 - These data form “intelligence inside” for new Web applications and services
- **Data as a Service**
 - Data is a strategic asset for a company

Cloud Computing

- **Infrastructure as a Service**
 - Build datacenters
 - Include power, scale, hardware, networking, storage, distributed systems, etc
 - Distributed data management and geo-distribution for locality, scalability, load balancing, and disaster tolerance
 - Support web-scale “offline” and “real-time” computation
 - Companies that build cloud services can buy time/resources on the infrastructure (like utility)
 - Data-intensive computing becomes a commodity that enables more data-intensive business

Microsoft's Mega-Datacenters



Microsoft's Cloud Platform

Applications & Solutions

This layer does on-demand utility computing back end hooks up with rich Internet applications on the front end, e.g. Office Live CRM, Windows Live services, Xbox Live, MSN, Live Search

Live Platform Services

This layer provides services designed specifically to serve the needs of applications, e.g. identity services, contact lists, social graph, communication services, and advertising platform

Cloud Infrastructure Services

This layer provides

- Virtualization support for all online services to run on
- Application frameworks designed to support a variety of application models
- Automatic deployment, load balancing, performance optimization, and horizontally scalable storage including searchable storage functionality

Global Foundation Services

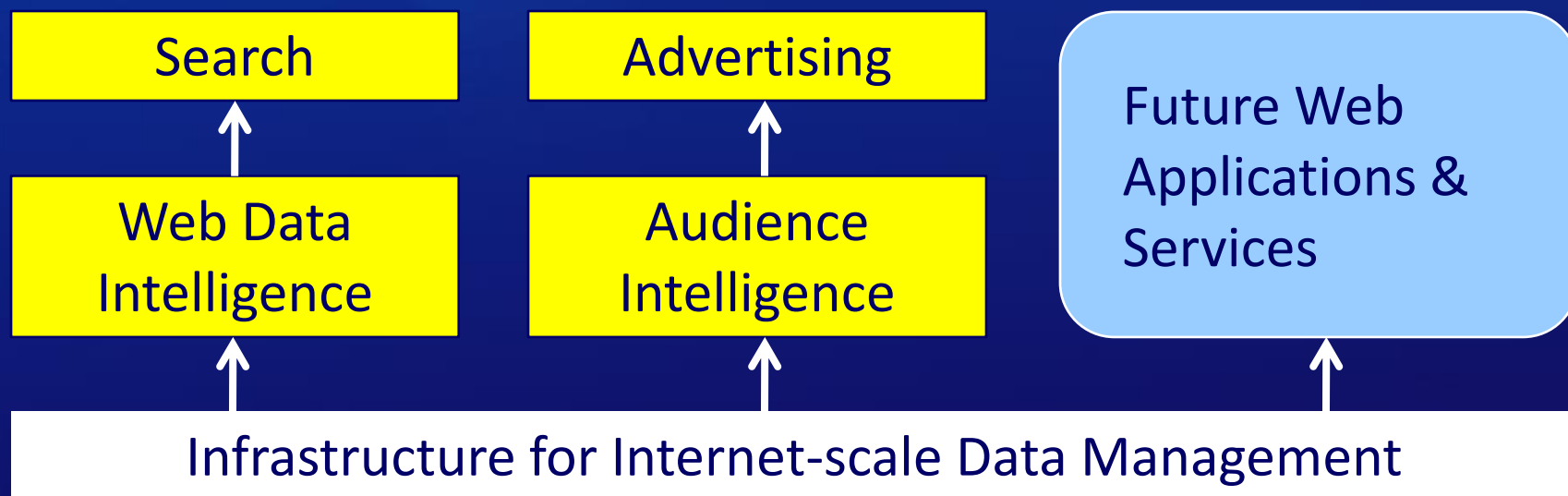
This layer contains physical data centers along with the networks that connect them

Two Major Types of Computing

- Computation-intensive computing
 - Focus on raw computing resource in the cloud
 - Computing as utility and paid for usage
- Data-intensive computing
 - Data as an important part of the services
 - Need to manage and process lots of data
 - E.g. Search, Advertising, Mapping, etc

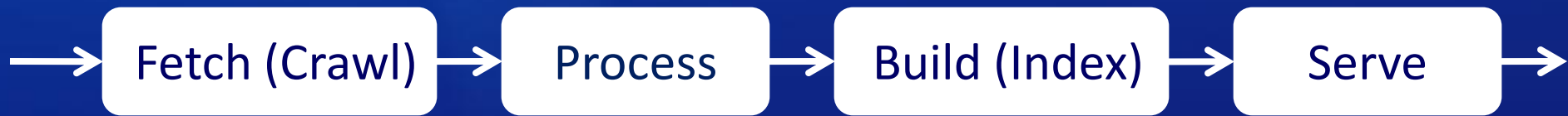
Search and Data-intensive Computing

- Search leads the development of infrastructure for data-intensive computing
 - Deep analysis and integration of web data
 - “Intelligence inside” for future web applications & services



Collect ➡ Store ➡ Aggregate ➡ Analyze ➡ Present ➡ Act

Traditional Search Engine Pipeline



- Guarantee the smooth running of the pipeline
 - Highly optimized for **performance** and **reliability**

Problems and Limitations

- Limited to light-weight, freshness-critical data processing -> Shallow analysis
- Lack of data management capability -> Transient and scattered data
 - Data is second class citizen
 - Difficult to accumulate knowledge (meta-data)
- Tightly coupled system -> Less flexibility & extensibility
 - A small change in one place could create a “ripple effect” in the entire system

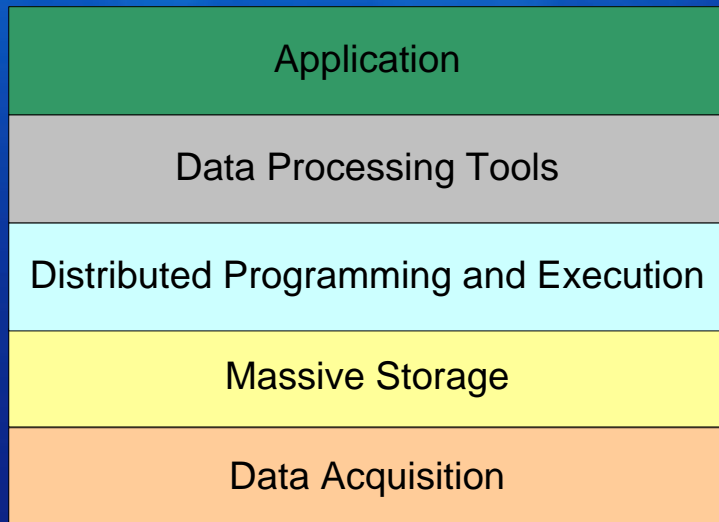
Why Search Infrastructure

- Provide Web data management for search
 - Highly optimized for **flexibility** & **extensibility**
 - Allow expensive, asynchronous data processing
- Support deep analysis and understanding
 - Web documents
 - Queries
 - Users

Experimental Infrastructure

- Search needs “continuous experiments”
 - An infrastructure for “scale” experiments is critical
- Three systems
 - Offline experiments (-> offline computation)
 - Online experiments (-> real-time computation)
 - Production (-> computing as utility)

Infrastructure for Data-intensive Computing



Script Language

Control API

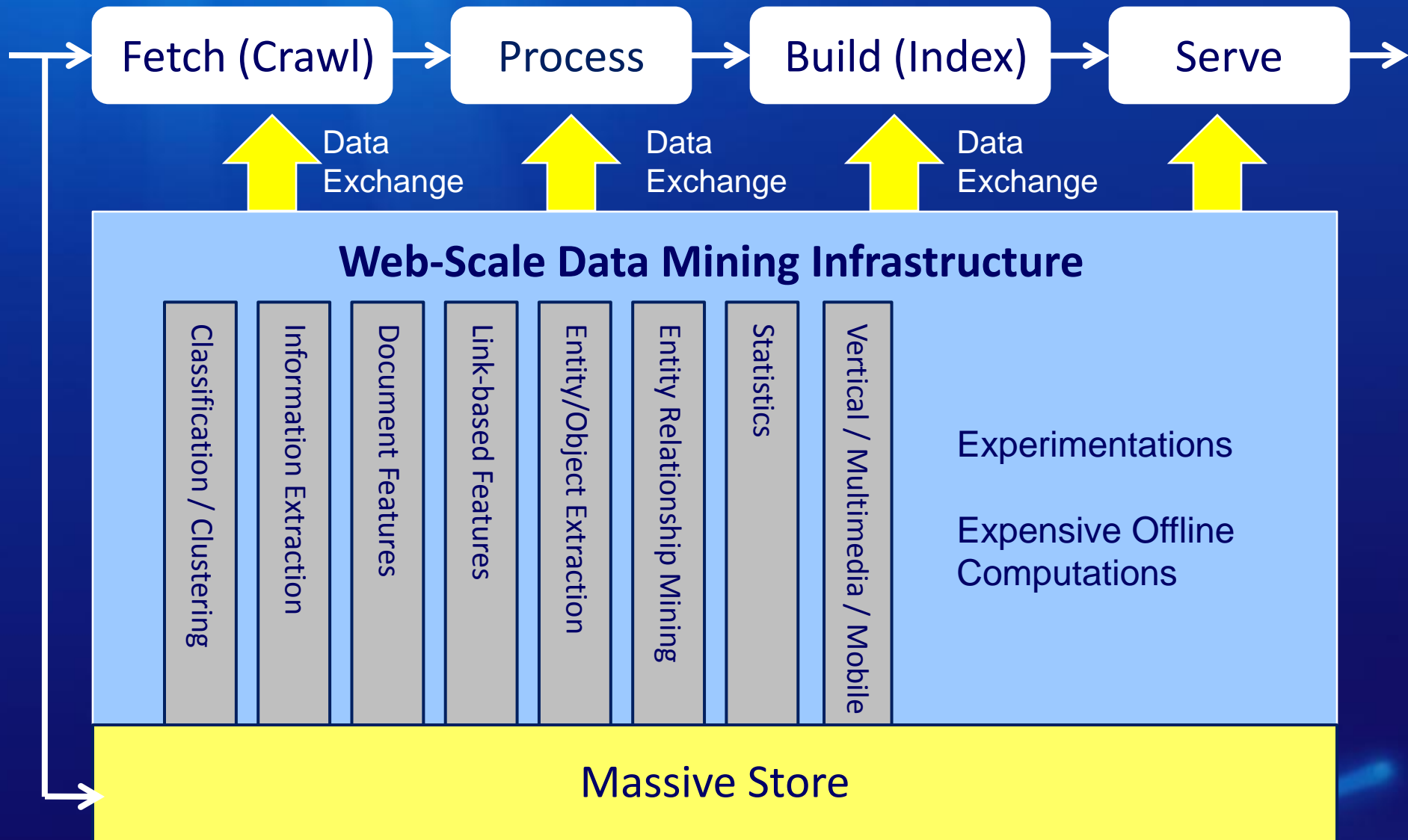
Distributed Programming API

Storage System API

Function **Easy-to-use**



Web-Scale Data Mining for Search





Renlifang (人立方) Search:

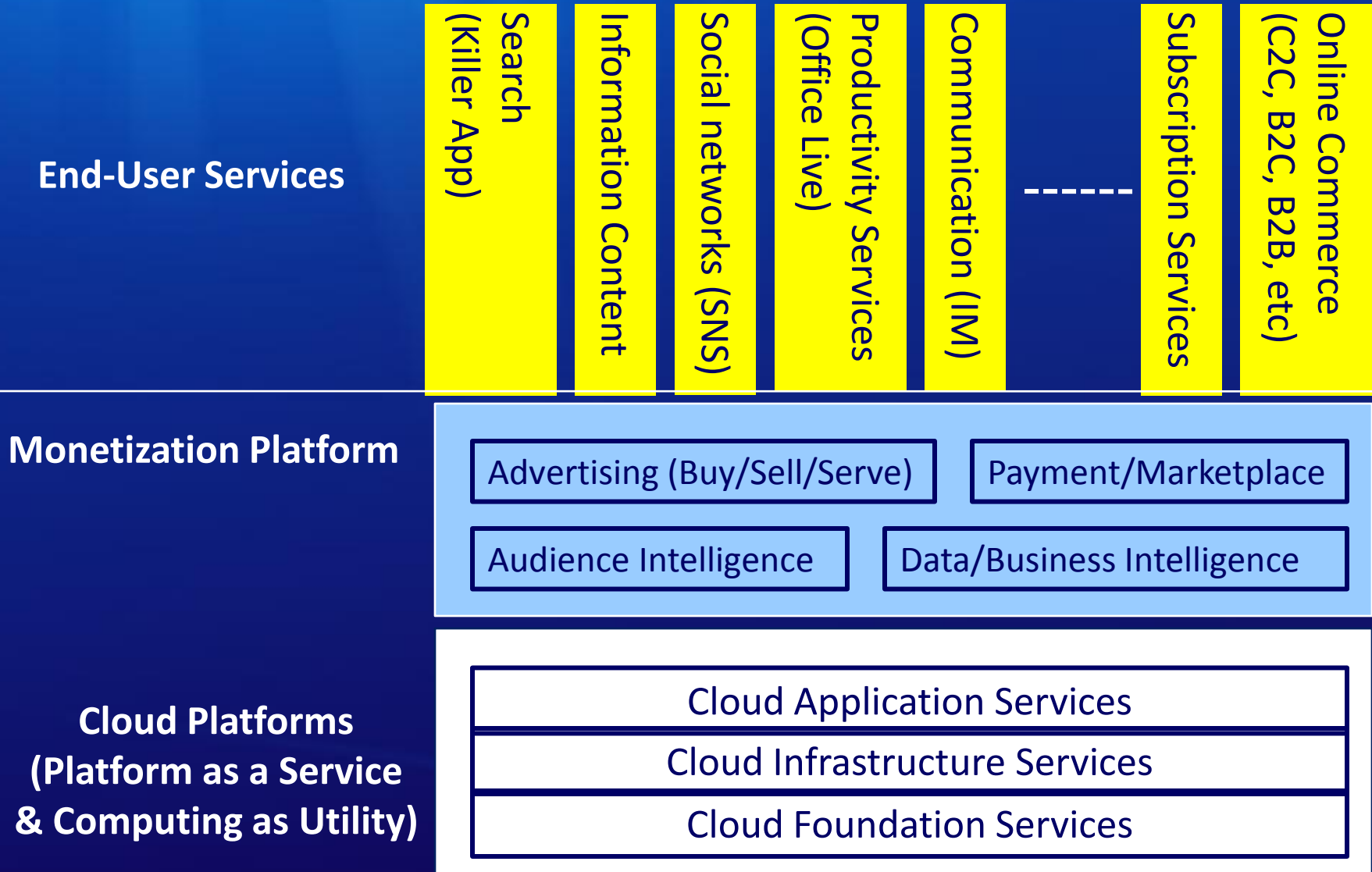
- Renlifang: <http://renlifang.msra.cn>
 - Mining relationships among people, organizations and locations from crawled Web pages
 - Covers head and tail queries
 - On both regular and well-known people, locations, and organizations
- Key technologies
 - Entity Extraction and Summarization
 - Expert Finding (based on Surrounding Text)
 - Entity Relationship Mining and Exploration (based on Web co-occurrence)

Search is a Driving Force for Cloud Computing

- *Build Infrastructure for data-intensive computing*
 - Datacenters and layered infrastructure
- *Web-scale data mining & machine learning*
 - Understand, extract, and expose semantics for future web services and applications
- *Deep integration of Web data*
 - Data as “intelligence inside”
- *Killer App for cloud platform to drive Internet economies on a global scale*

Cloud Computing and Internet Services

Microsoft provides the cloud platform for developers to build Internet services



Datacenter Evolution

- **Generation 1:** Traditional data centers that focus on uptime, reliability, and redundancy
- **Generation 2:** Built with more energy efficiency and sustainability in mind
- **Generation 3:** Built for massive scalability and cost efficiency
- Problems and Challenges
 - Need to build data centers to be just right in size in order to squeeze out the most energy efficiency
 - The worst thing we can do in delivering facilities for the business is not have enough capacity online, thus limiting the growth of our products and services
 - The second worst thing we can do in delivering facilities for the business is to have too much capacity online

Microsoft's Gen 4 Modular Datacenter

- **Generation 4: Modular datacenter and smart growth**
 - Commoditize the build-out of data centers, assembling data center components on-site
 - Use shipping containers filled with servers as their basic building block
 - Modular units of prefabricated mechanical, electrical, and security components
 - A container facility helps ensure that we don't overbuild server capacity, while allowing the company to reduce the time to build a data center
 - Decrease the need for water chillers by using ambient air to cool data centers

Microsoft's Gen 4 Modular Datacenter

- Video:

<http://video.msn.com/video.aspx?vid=b4d189d3-19bd-42b3-85d7-6ca46d97fe40>

Summary

- The real underlying value of cloud computing is that it transparently makes software and data available everywhere
- Computing as utility and data-intensive business
- Cloud computing will have profound impacts on Internet economies
 - “Intelligence inside” for Internet economies (data driven)
 - Speed of light (algorithms driven)
 - Scale to global (infrastructure driven)

Thanks You!

wyma@microsoft.com