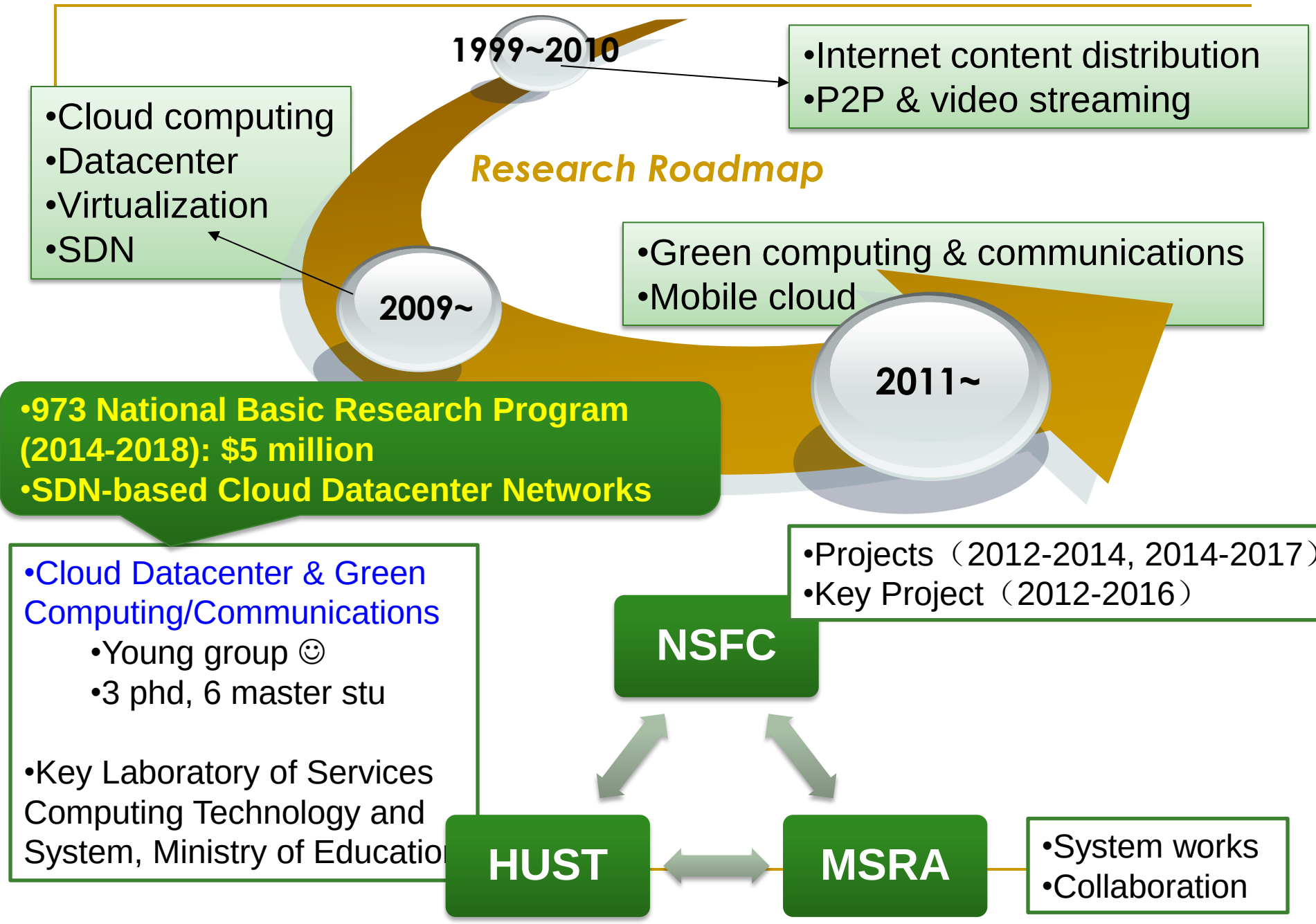
From Online Cloud Services to Underlying Datacenter Optimization

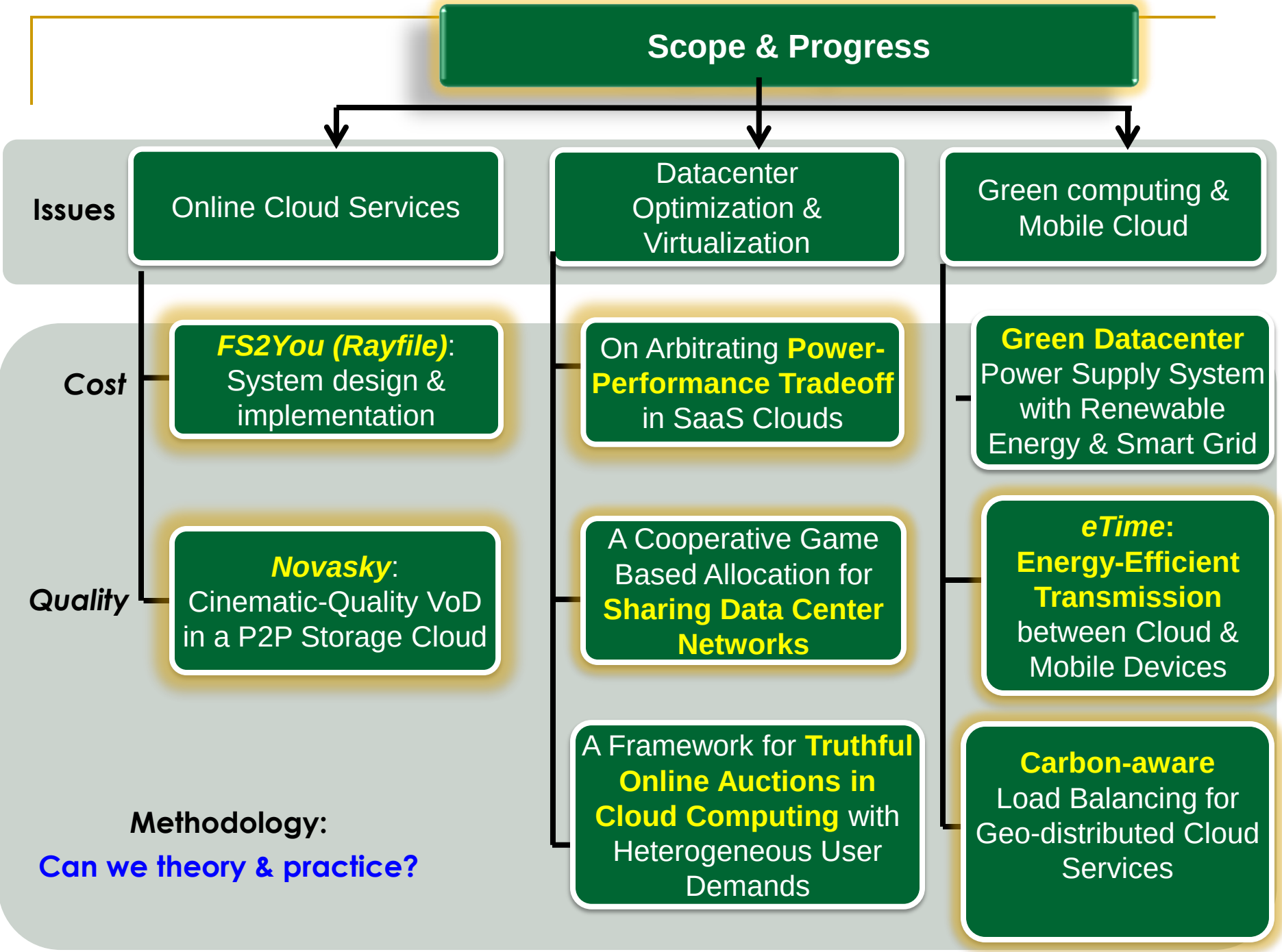
——— *A Top-Down Approach*

Dr. Fangming Liu (fangminghk@gmail.com)
刘方明 : http://grid.hust.edu.cn/fmliu/

Associate Professor
School of Computer Science & Technology
Huazhong University of Science & Technology

Nov. 2013@HKU


1999~2010
•Internet content distribution
•P2P & video streaming
Research Roadmap
•Cloud computing
•Datacenter
•Virtualization
•SDN
2009~
•Green computing & communications
•Mobile cloud
2011~
•973 National Basic Research Program (2014-2018): $5 million
•SDN-based Cloud Datacenter Networks
•Cloud Datacenter & Green Computing/Communications
•Young group ☺
•3 phd, 6 master stu
•Key Laboratory of Services Computing Technology and System, Ministry of Educatio
•Projects（2012-2014, 2014-2017）
•Key Project（2012-2016）
NSFC
HUST
MSRA
•System works
•Collaboration


Scope & Progress
Issues
Online Cloud Services
Datacenter Optimization & Virtualization
Green computing & Mobile Cloud
Cost
Quality
FS2You (Rayfile): System design & implementation
Novasky: Cinematic-Quality VoD in a P2P Storage Cloud
On Arbitrating Power-Performance Tradeoff in SaaS Clouds
A Cooperative Game Based Allocation for Sharing Data Center Networks
A Framework for Truthful Online Auctions in Cloud Computing with Heterogeneous User Demands
Green Datacenter Power Supply System with Renewable Energy & Smart Grid
eTime: Energy-Efficient Transmission between Cloud & Mobile Devices
Carbon-aware Load Balancing for Geo-distributed Cloud Services
Methodology:
Can we theory & practice?

Outline

- Introduction
- Online Cloud Services ➔ case study
  - ***FS2You*: Online Hosting & Content Distribution**
  - *Novasky*: Cinematic-Quality VoD in a P2P Storage Cloud
  - *eTime*: Mobile Cloud
- Underlying Datacenter Optimization
  - Network Virtualization for Multi-tenants DC
  - Green DC Power-Performance Tradeoffs
- Future Plan & Collaboration

Computing, Storage & Distribution as **Utility**

- Based on large-scale datacenters or CDNs with millions of servers

Online social networks

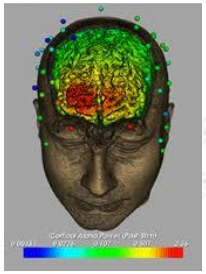
E-commerce

IoT & CPS

Service Cloud

Mobile services

Scientific computing



Online storage & distribution

One critical building block towards full-fledged cloud storage & computing services

Online Hosting and Content Distribution

- A new type of cloud storage & content sharing service

→ online hosting and distribution has become increasingly popular

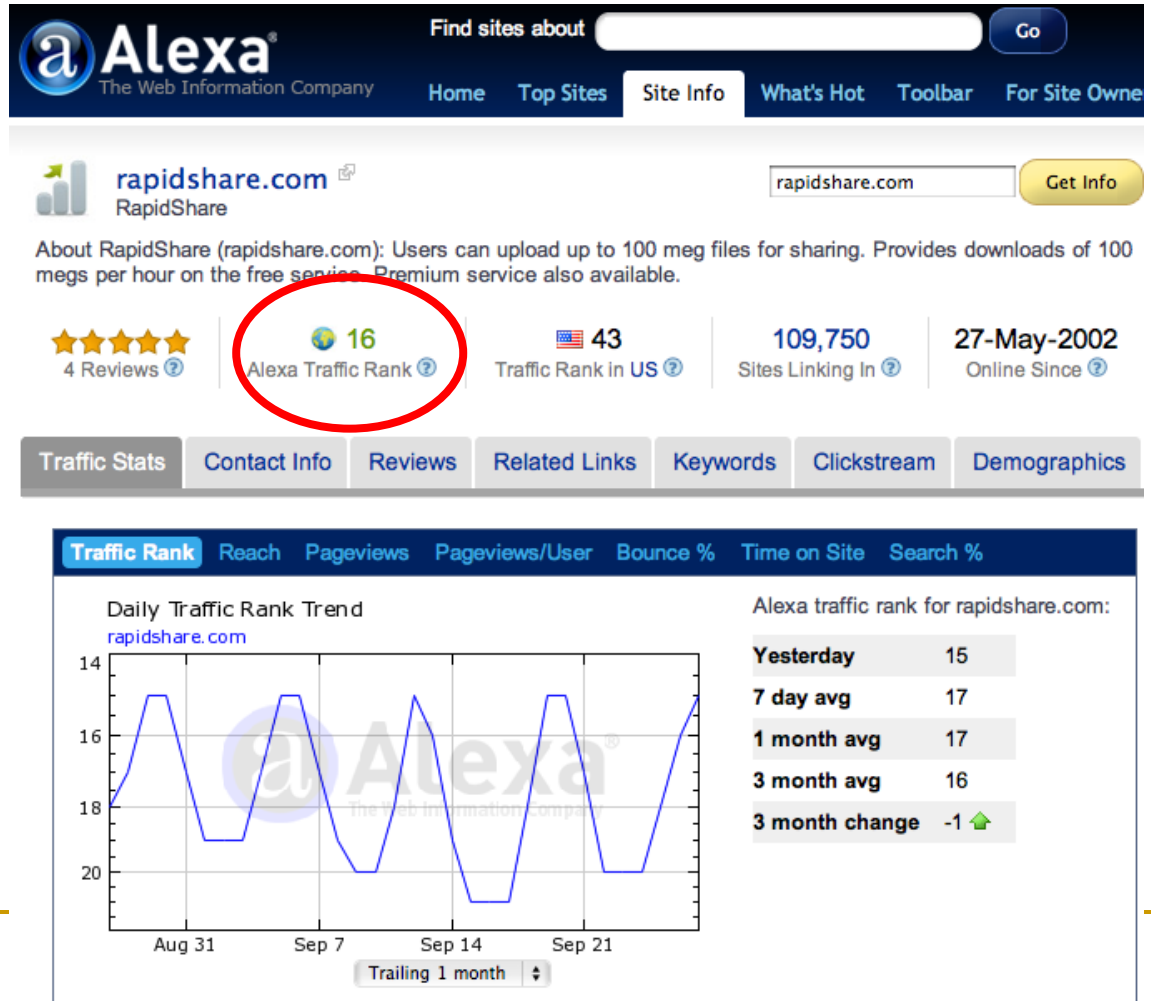
1-click hosting

Easily upload files, of both small & large sizes, onto dedicated servers, to be shared among a potentially large group of interested users

Online Hosting and Content Distribution

- A new type of cloud storage & content sharing service

→ online hosting and distribution has become increasingly popular



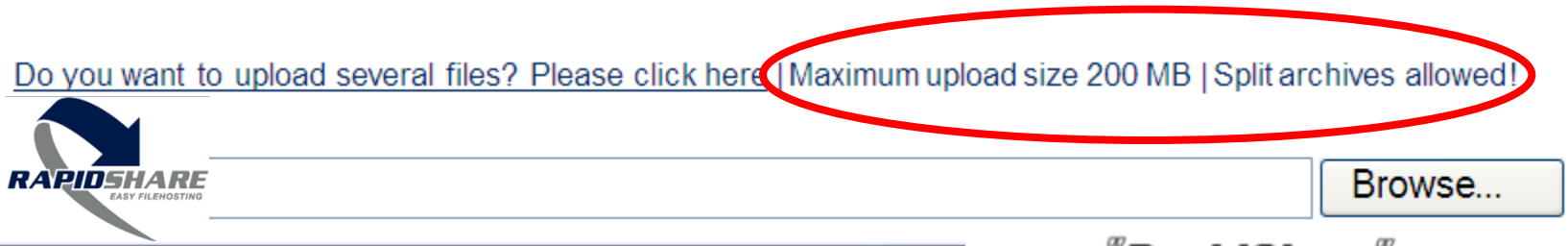
Online Hosting and Content Distribution

- Features compared with conventional P2P file sharing such as *BitTorrent*
  - Better reliability and service guarantee
    - Uploading (storage) and downloading (availability)
  - Ease of use
    - simple URL shared to others, one-click service
    - no software download and configuration

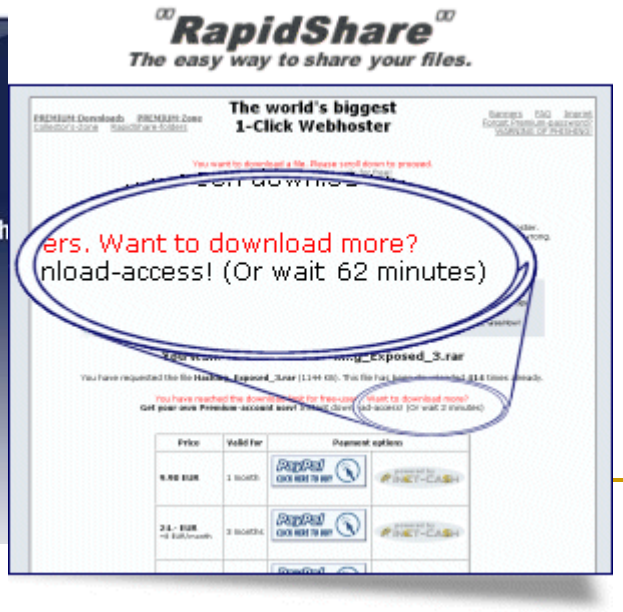
1. Download Link:
http://rapidshare.com/files/230183792/cg.png.html
MD5: 59D49AF36788FC0EE3E018270E6597D3

Limited server storage/bandwidth ⟷ **Extensive files of various popularity & sizes**

- Files hosted in either CDNs or dedicated large datacenters
- *Rapidshare*, >1500 TB of storage in its datacenters
- Skyrocketing server bandwidth costs: yearly 15~20 million USD
  - impose usage restrictions or/and paid service

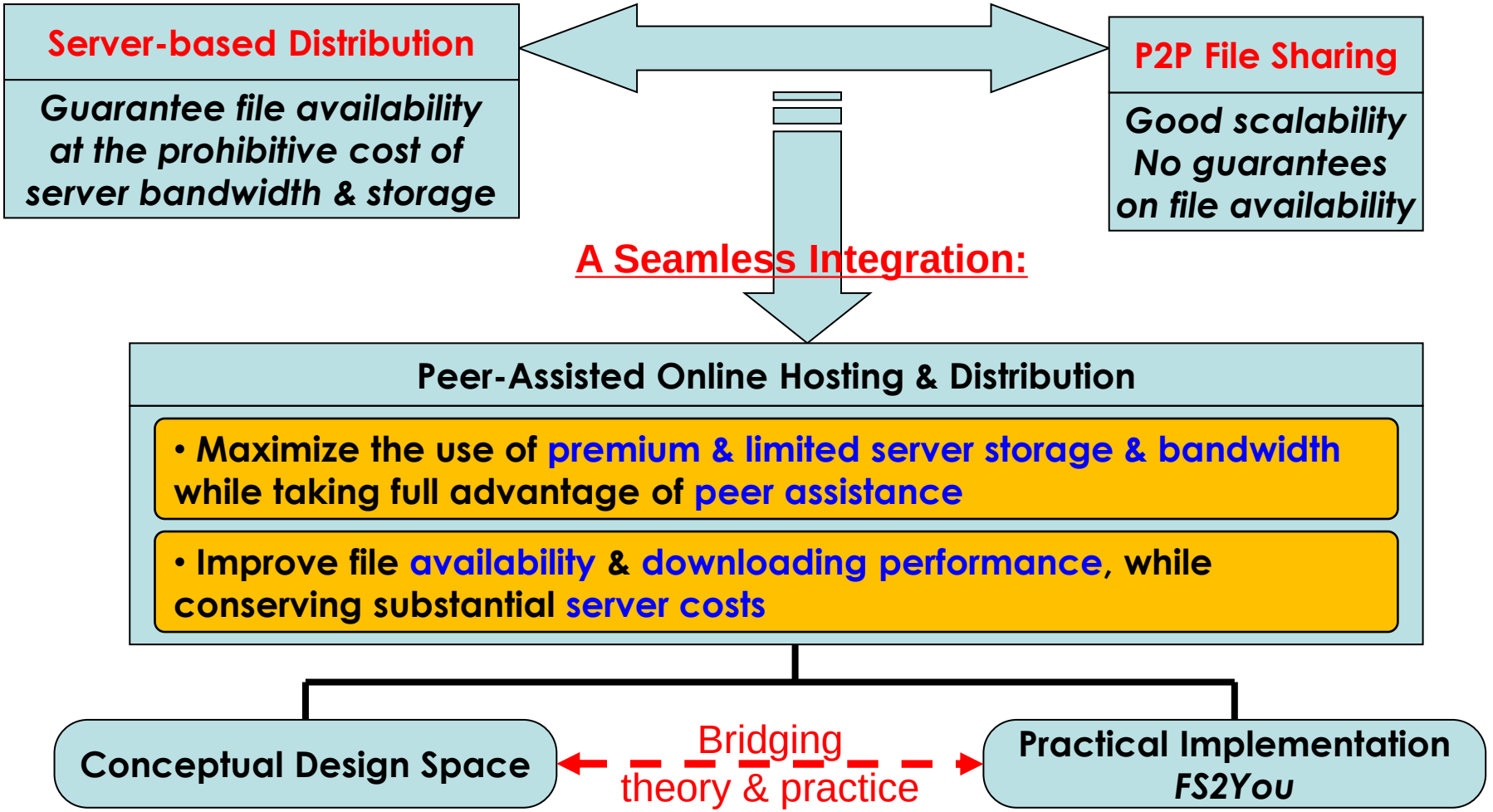


Features	Free	Premium
Max. file size for Uploads	200 MB	2.000 MB [1]
Personal webspace	-	500 GB
Inclusive download traffic	-	150 GB per month
Deletion of files	After 90 days without Download	Never [3]
Instant start of Download	No	Yes
Download speed	Limited [4]	Unlimited
Max. parallel Downloads	1	Unlimited
Support of Download-Accelerator	No	Yes
Resume of broken Downloads	No	Yes



Design Objective & Challenges

- Peer-assistance ➔ natural but non-trivial in design
- Balance two extremes ➔ cost-performance tradeoff
  - Couple peer assistance & server provisioning in a complementary & transparent manner



General Model & Performance Metrics

- Important performance metrics to characterize "good" online hosting and distribution systems from different perspectives

Multiple files: $\mathcal{M} = \{1, 2, \ldots, M\}$

of diverse popularity & sizes: $\lambda_i \;\; f_i$

Limited server storage: $F \leq \sum_{i \in \mathcal{M}} f_i$

Limited server bandwidth: $\sum_{i \in \mathcal{F}} S_i \leq S$

Peer assistance effectiveness: η_i

Peer upload/download capacity: μ_j, c_j

File availability: $\lambda = \sum_{i \in \mathcal{F}} \lambda_i$

attract & serve as many users as possible

$d = \sum_{i \in \mathcal{F}} \overline{x_i} d_i / \sum_{i \in \mathcal{F}} \overline{x_i}$

maintain as high downloading performance as possible

System throughput: $D = \sum_{i \in \mathcal{F}} \overline{x_i} d_i$

Design Space: Semi-Persistent Storage & Replacement

Given a constrained server storage capacity $F \leq \sum_{i \in \mathcal{M}} f_i$
a server storage & replacement strategy determines
which set of files $\mathcal{F} \subseteq \mathcal{M}$ to be stored on the server

Problem abstraction

A classical 0-1 knapsack problem

To attract & serve as many users as possible

Maximize $\lambda = \sum_{i \in \mathcal{M}} \lambda_i w_i$

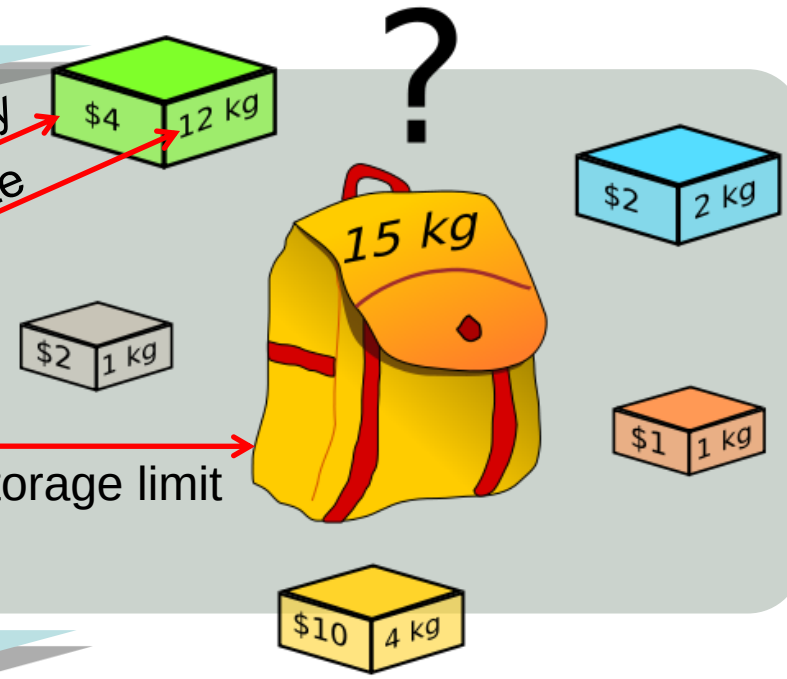
Subject to: $\sum_{i \in \mathcal{M}} f_i w_i \leq F,$

$w_i \in \{0, 1\}, i \in \mathcal{M},$

popularity

size

Server storage limit



- NP-complete ➔ can be solved using a dynamic programming algorithm
- The static nature ➔ not efficient to be used in practical systems
- Not suitable to be used for the eviction or replacement operation
  - dynamic evolution of user interests on currently stored files
  - a continuous flow of newly uploaded files from users

Our Solution: Storage & Replacement

Simplicity & efficiency are more of a concern in practical system implementation & operation, at a cost of acceptable sub-optimal solution

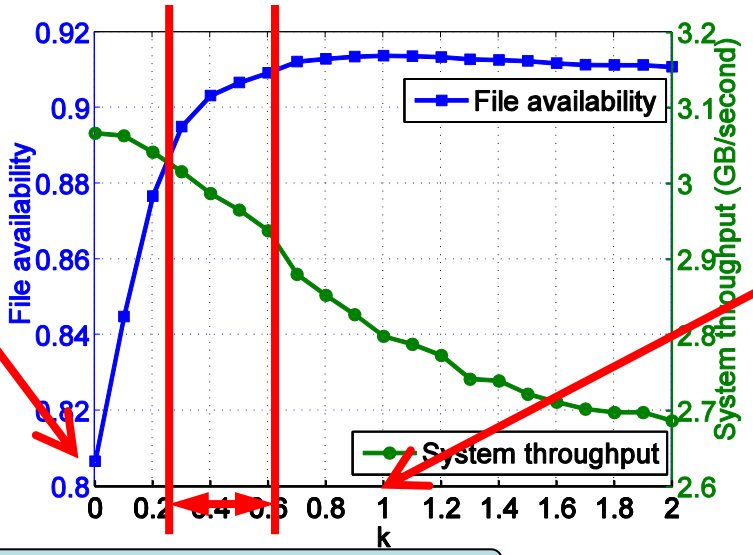
- A simple framework obeying a greedy algorithm
  - each file with a *profit-to-weight index:* $H_i \leftarrow \lambda_i (f_i)^{-k}$
  - files are ranked in descending order by their indices
    - files with relatively high ranks are preferentially stored
    - alternatively, can simply identify those with lower ranks, and perform evictions/replacements whenever necessary
      - *start from lowest ranks until a certain volume of files are evicted*
      - *customize a threshold of Hi below which are the candidates for eviction*

Illustration: Storage & Replacement

- Unify important aspects with tunable design knobs ➔ flexibility

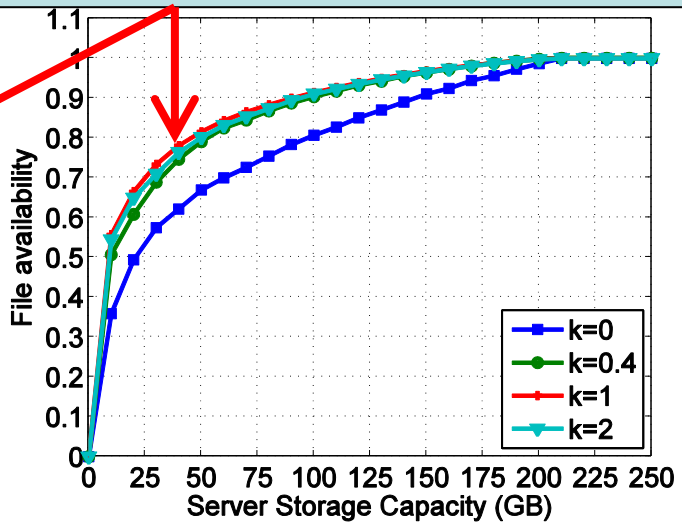
$$H_i \leftarrow \lambda_i (f_i)^{-k}$$

- **K=0** ➔ unpopular-first-eviction strategy ➔ maximize system throughput
- **K=1** ➔ balanced consideration between file popularity & size ➔ maximize system-wide file availability
- **K∈(0,1)** ➔ various degree of throughput & availability



Unpopular-first eviction

With more emphasis on file availability, a real-world system with this customization will be demonstrated later



Opportunity to achieve both high availability & throughput

Design Space: Bandwidth Allocation

Given a specific total amount of server bandwidth S, a server bandwidth allocation strategy determines how to assign the bandwidth to each file $S_i,\ \forall i \in \mathcal{F},\ s.t.\ \sum_{i\in\mathcal{F}} S_i \le S$

What is the optimal server bandwidth allocation across files to achieve the upper bound of system-wide average downloading rate?

A critical factor ➔ server bandwidth provisioning relative to the peer uploading contributions (across files)

Based on Little's Law ➔ system-wide average downloading rate in steady state

To maximize $d = \sum_{i\in\mathcal{F}} \lambda_i f_i \left(\sum_{i\in\mathcal{F}} \frac{\lambda_i f_i}{\mu\eta_i} - R\right)^{-1}$

$$\text{Maximize} \quad R = \sum_{i\in\mathcal{F}} \left(\frac{1}{\mu\eta_i}\right) S_i$$

$$\text{Subject to:} \quad \sum_{i\in\mathcal{F}} w_i S_i \le S,$$

$$S_i \le S_{maxi} = \left(1 - \frac{\mu\eta_i}{c}\right)\lambda_i f_i, i \in \mathcal{F}$$

Problem abstraction

A classical continuous knapsack problem with bounded variables

FS2You: Architecture

Seamless integration of Online storage & peer-assistance

- www.rayfile.com (collaboration with a startup company, venture capital: $30 million)
- One of the most popular online hosting systems in China
- Google.cn entries over 4,000,000 in 2009
- **Measurements on 350 GB trace data** collected from over **3.3 million** users

- **Tracking Server**
  - Channels' (files) Info & MD5
  - Bootstrapping
  - List of peers in channels
- **Hosting Servers**
  - conditions

System

- **FS2You: Peer-Assisted Semi-Persistent Online Hosting at a Large Scale**, **IEEE Transactions on Parallel and Distributed Systems**, vol. 21, no. 10, October, 2010.

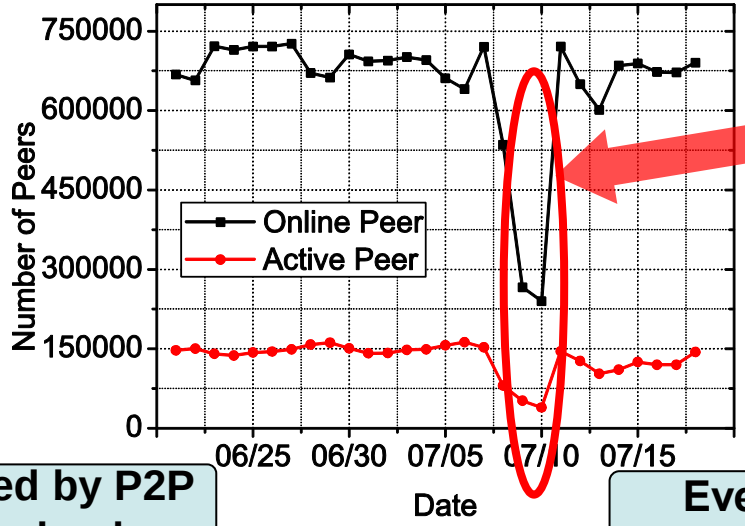
Theory

- **Quota: Rationing Server Resources in Peer-Assisted Online Hosting Systems**, in IEEE **ICNP**, Princeton, New Jersey, October, 2009.

Measure

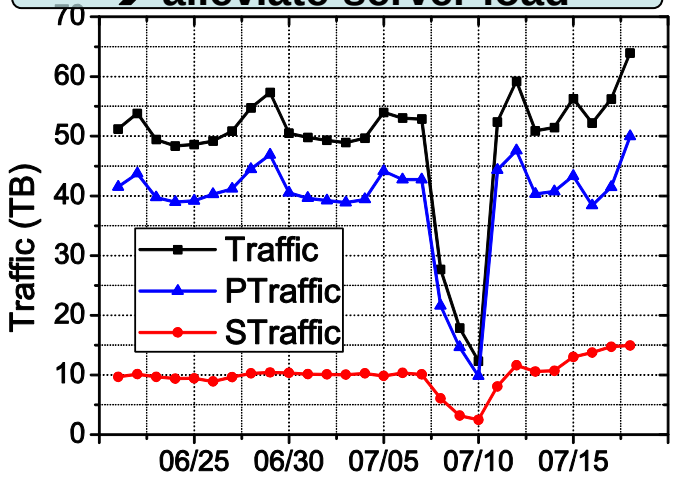
- **FS2You: Peer-Assisted Semi-Persistent Online Storage at a Large Scale**, in IEEE **INFOCOM**, Rio de Janeiro, Brazil, April, 2009.

Overall Scale & Performance

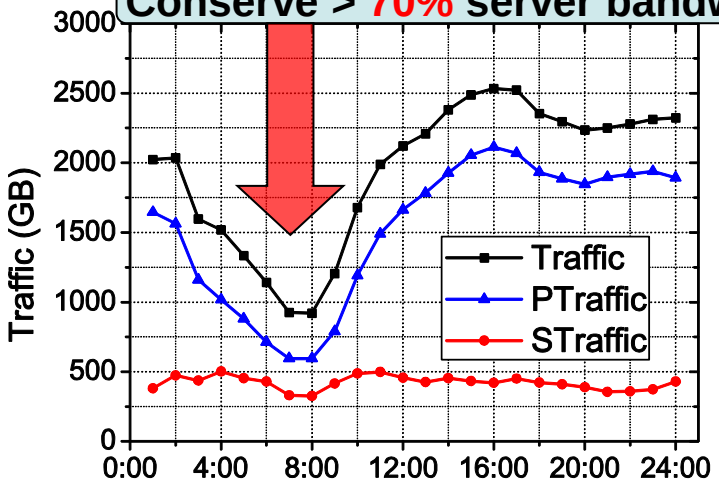


Crash failures of log server

Up to 80% contributed by P2P ➔ alleviate server load



Even during "calm" period, Conserve > 70% server bandwidth



- The designs in FS2You can scale to a large number of peers, and to withstand the test of a tremendous volume of traffic over a long period of time
- The cost of server capacity has been substantially saved by peer assistance

Outline

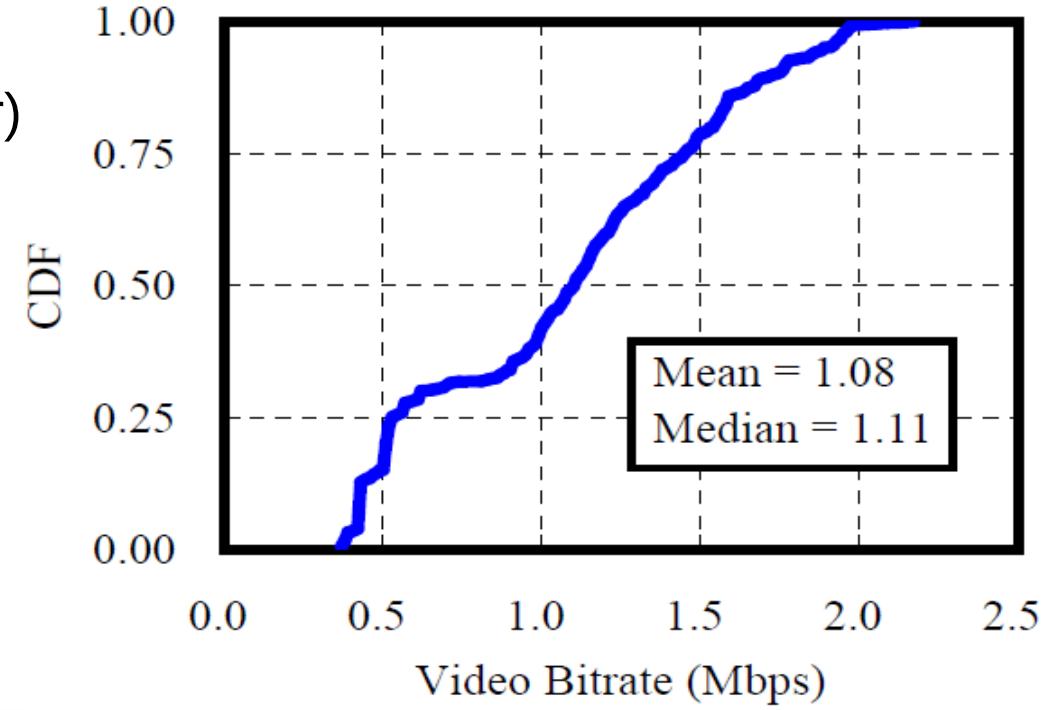
- Introduction
- Online Cloud Services ➔ case study
  - FS2You: Online Hosting & Content Distribution
  - ***Novasky*: Cinematic-Quality VoD in a P2P Storage Cloud**
  - *eTime*: Mobile Cloud
- Underlying Datacenter Optimization
  - Network Virtualization for Multi-tenants DC
  - Green DC Power-Performance Tradeoffs
- Future Plan & Collaboration

Cinematic-Quality VoD in a P2P Storage Cloud

- Hybrid solution of cloud(server) & P2P
  - Peer cache: 1-2 GB
  - Inter-connected with a high-bandwidth network
  - **Deployed in Campus**

>100,000 lines of code in C++

> 1,000 videos with 1– 2 Mbps



"It's Not just Cost, it's the Quality!" [1]

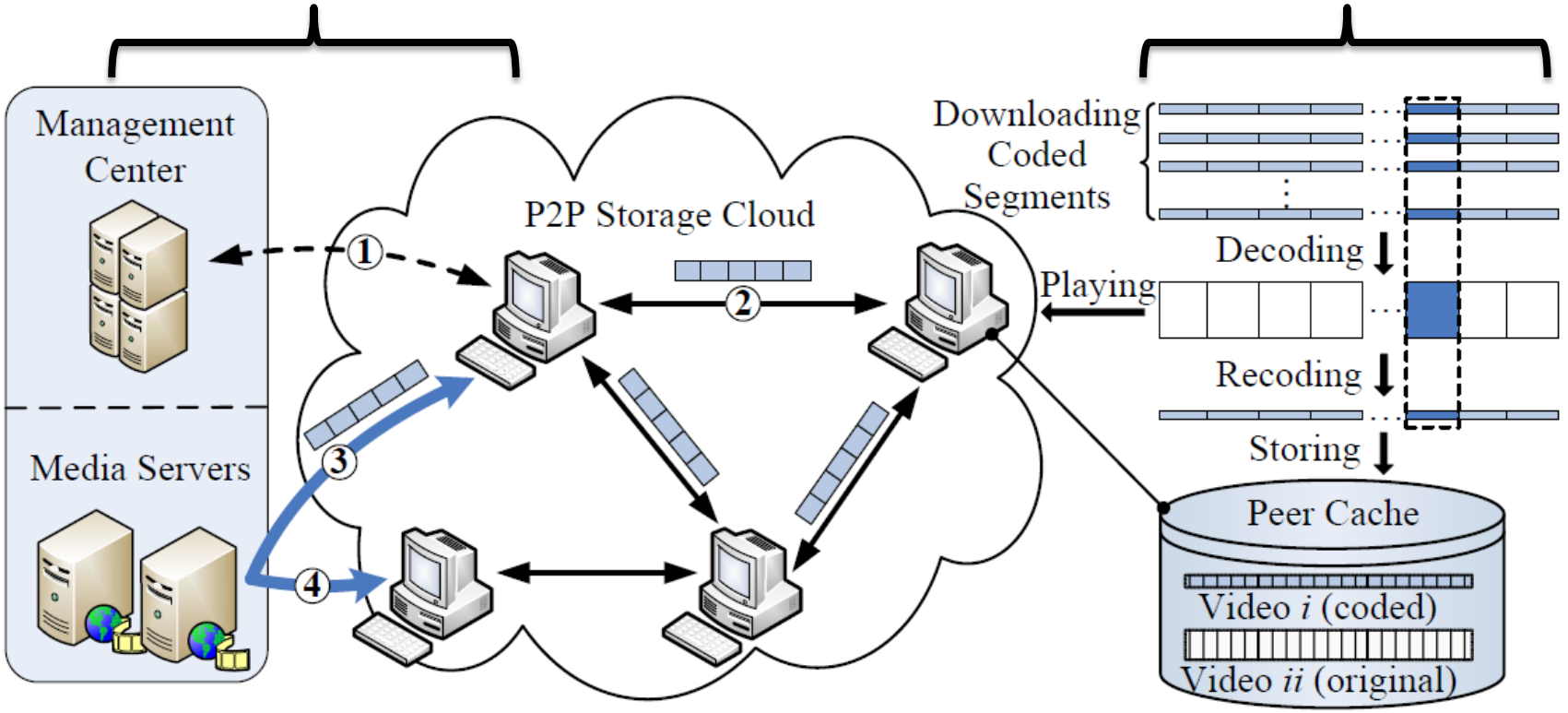
Real System

Theory

- **Cinematic-Quality VoD in a P2P Storage Cloud: Design, Implementation and Measurements**, **IEEE Journal on Selected Areas in Communications (JSAC)**, Special Issue on Emerging Technologies in Communications, 2013
- **Peer-Assisted On-Demand Streaming: Characterizing Demands and Optimizing Supplies**, **IEEE Transactions on Computers**, 2012.

Novasky system architecture with **2 unique designs**

- Servers ➔ Adaptive push-to-peer strategy
- Peers ➔ Coding-aware peer storage & replacement strategy

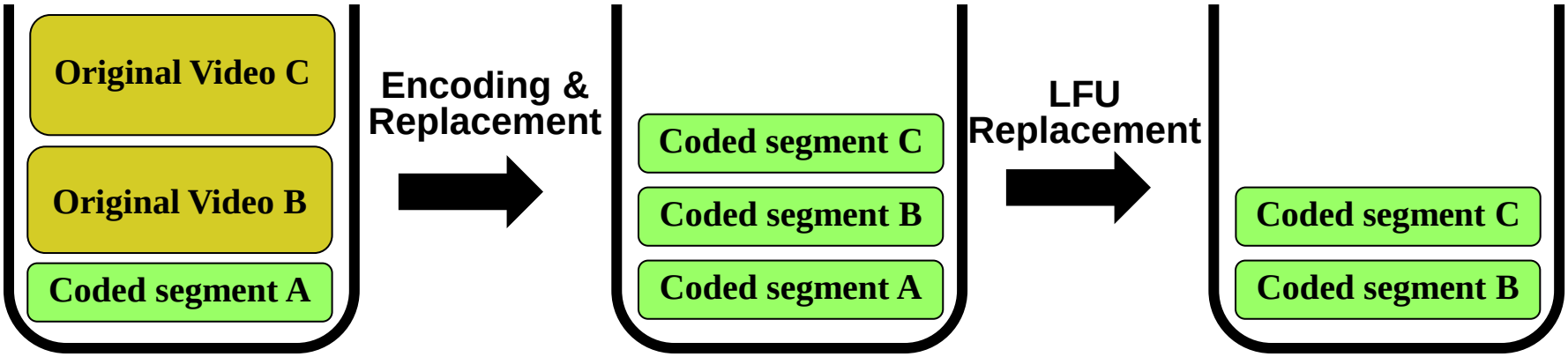


- Transparent "cloud" service & feature-rich user interface

Our Solution: Peer Storage & Replacement using Reed-Solomon Codes

① *When the current cache space is saturated*

② *If cache space is still insufficient to accommodate new incoming video*

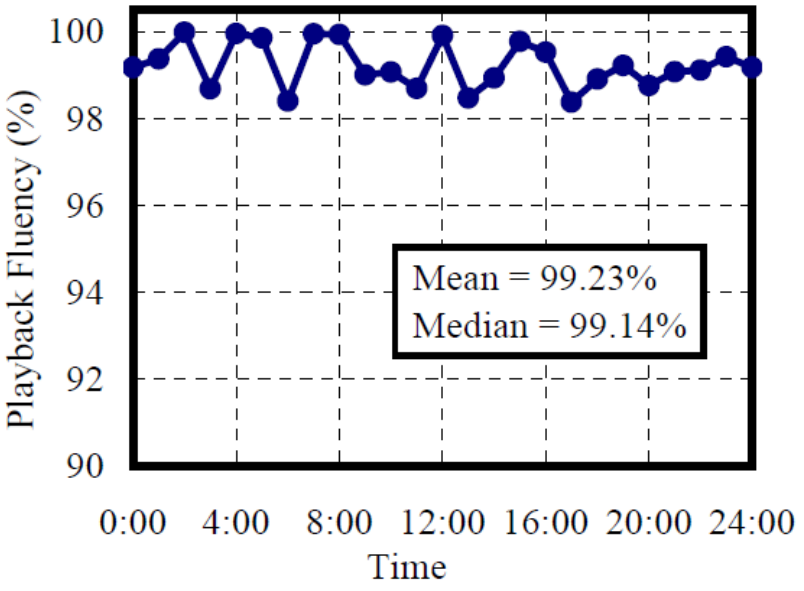


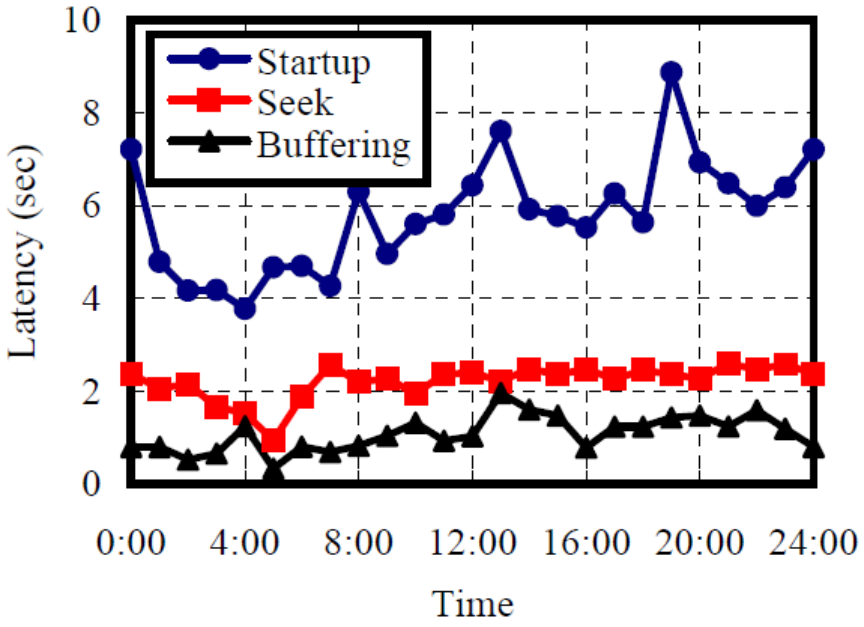

```
sort currently cached original videos in LFU order;
for each original video do
    // code-based replacement
    use Reed-Solomon codes to generate a coded segment
    to be stored, with a randomly selected index from the
    Vandermonde matrix;
    evict the original video from the cache;
    if the cache space is enough for the new incoming data
    then
        store the incoming data; return;
```

```
sort currently cached coded segments in LFU order;
for each coded segment do
    evict the coded segment from the cache;
    if the cache space is enough for the new incoming data
    then
        store the new incoming data; return;
```

On-Demand Streaming Fluency & Latencies

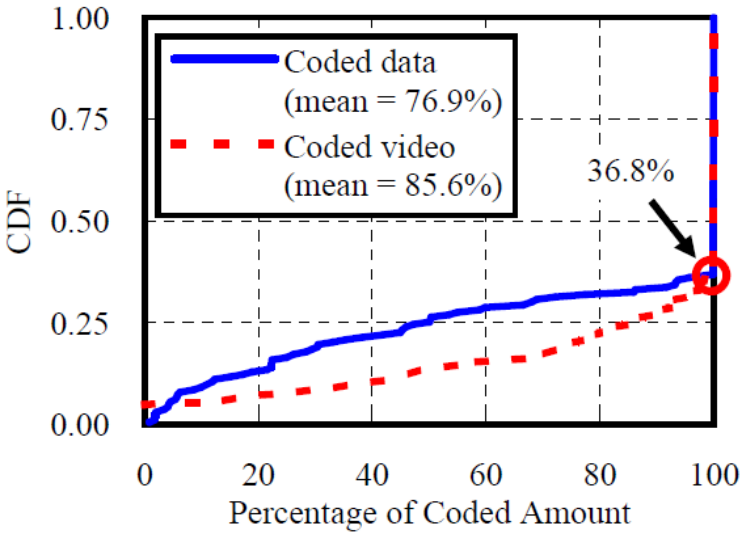
- Measurements over 6 months (2011)
- > 10,000 users to date
- Playback fluency index: fraction of uninterrupted watching time out of the total watching time
- Sustainedly> 98% and the average playback fluency over time is 99.23%





- Seek latencies <= 3 seconds < 10−30 seconds in existing P2P VoD systems
- Fast startups within 4-9 seconds <15−40 seconds in existing systems

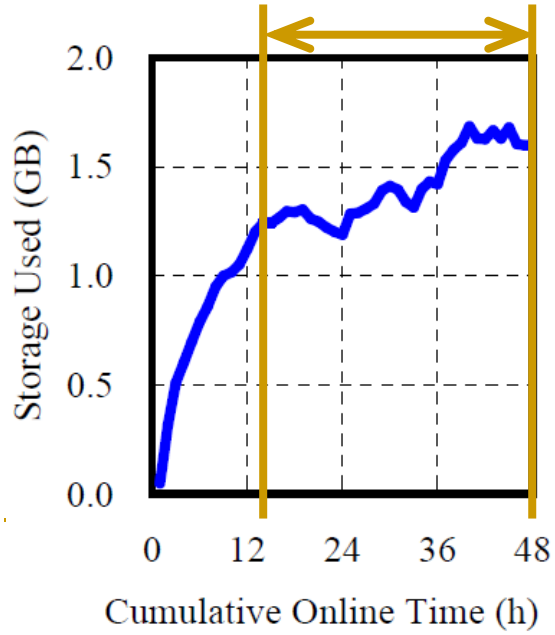
Storage Efficiency with Coding

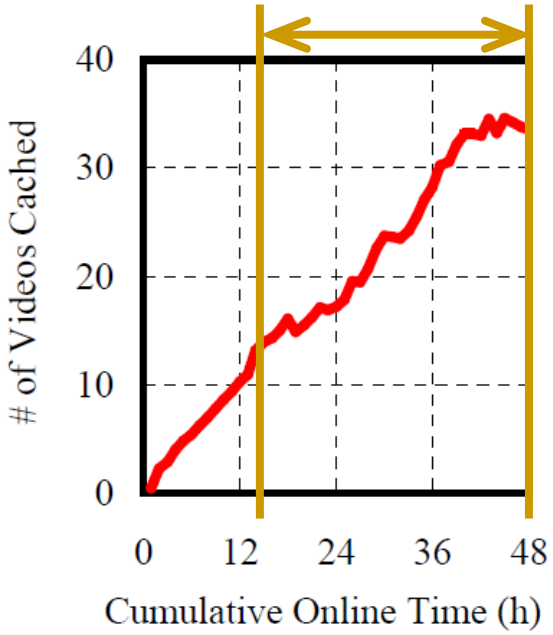


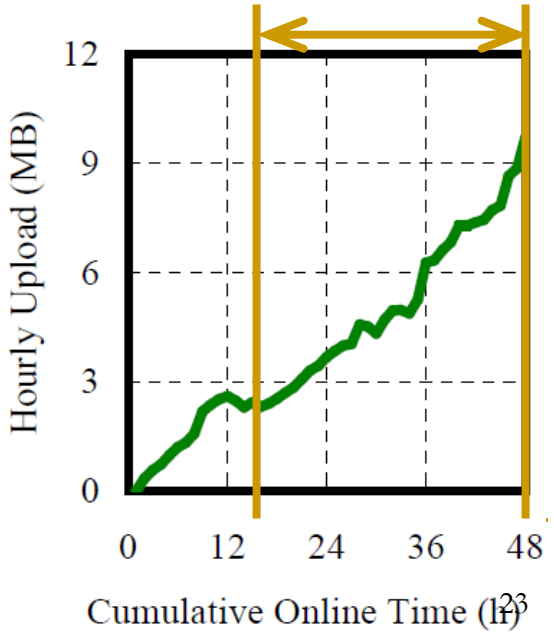
- 63.2% of peers' caches are completely filled with coded data ➔ coding-aware storage & replacement strategy plays an important role
- Improve peer upload contribution and accommodate video enrichment, under a same limit of cache space (2GB)

The increase of storage usage slows down, while no. cached videos keeps increasing rapidly

The increase of peer upload contribution is in proportion to that of no. cached videos, rather than that of storage usage



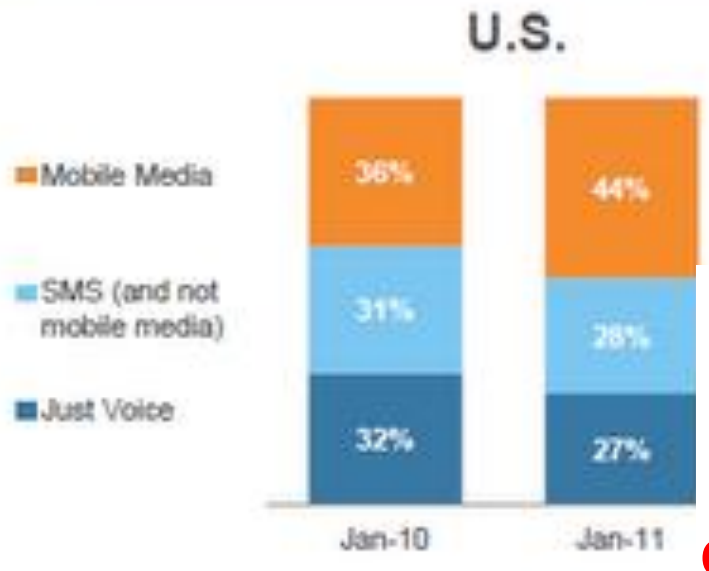


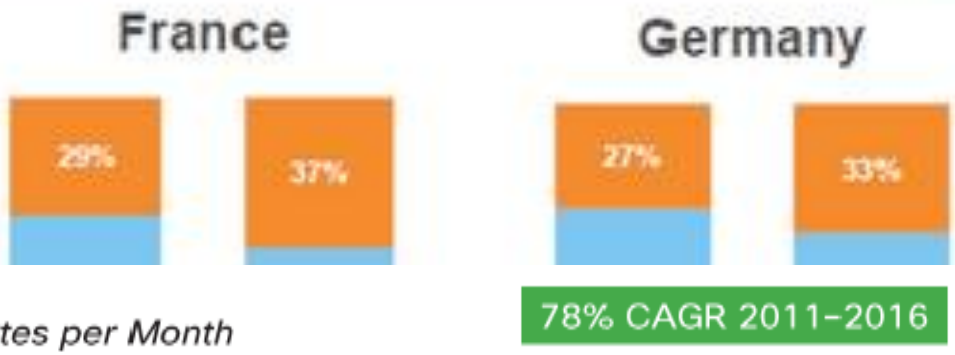


Outline

- Introduction
- Online Cloud Services ➔ case study
  - FS2You: Online Hosting & Content Distribution
  - Novasky: Cinematic-Quality VoD in a P2P Storage Cloud
  - ***eTime*: Mobile Cloud**
- Underlying Datacenter Optimization
  - Network Virtualization for Multi-tenants DC
  - Green DC Power-Performance Tradeoffs
- Future Plan & Collaboration

Mobile Media Users Growing in All Markets





78% CAGR 2011–2016

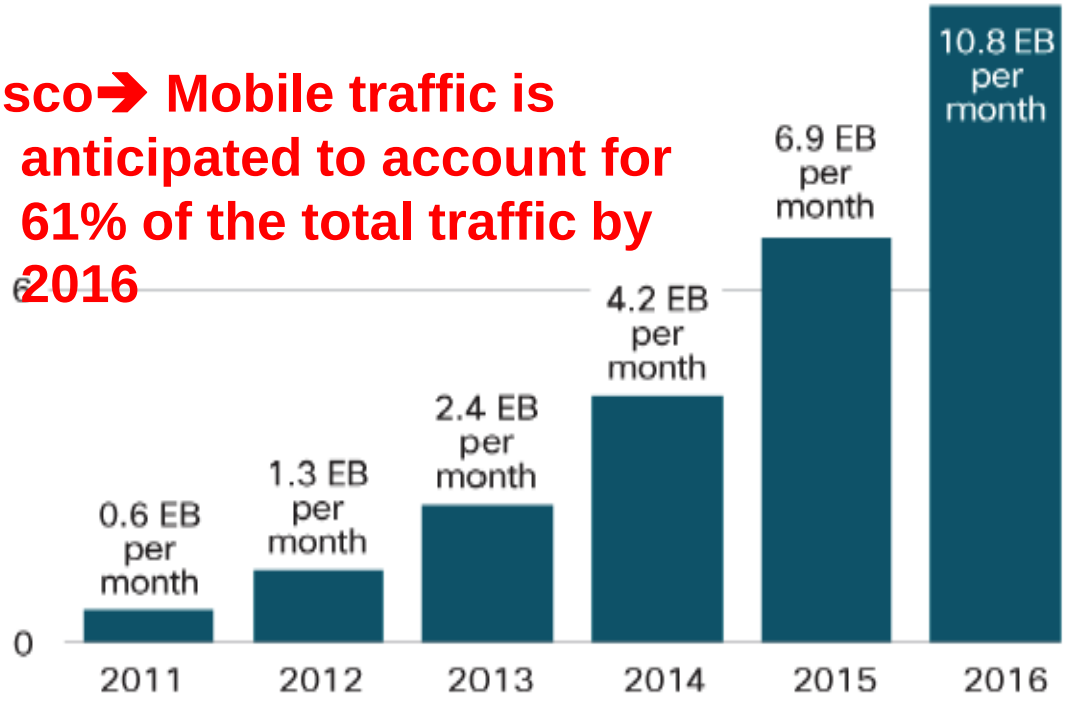
2011

Google has seen

400%

increase in the number of searches on mobile devices

Cisco➔ Mobile traffic is anticipated to account for 61% of the total traffic by 2016

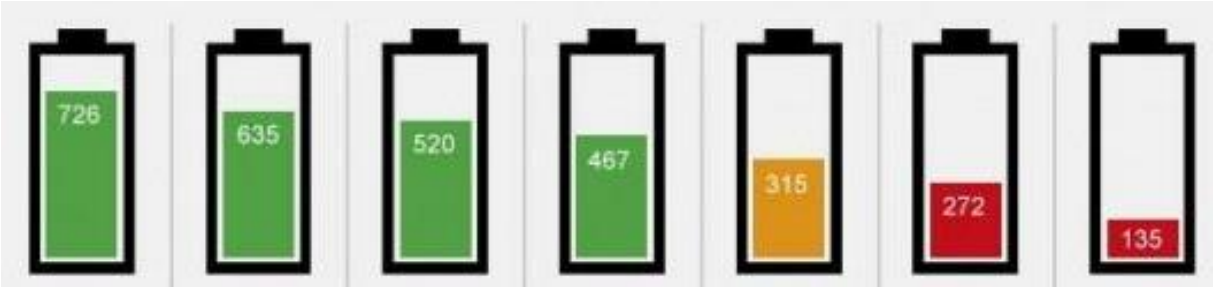


Source: Cisco VNI Mobile, 2012

Green Computing & Communications: Mobile Cloud

- Resource-constrained mobile devices
  - **Battery**, CPU, Storage, Bandwidth

primary bottleneck!

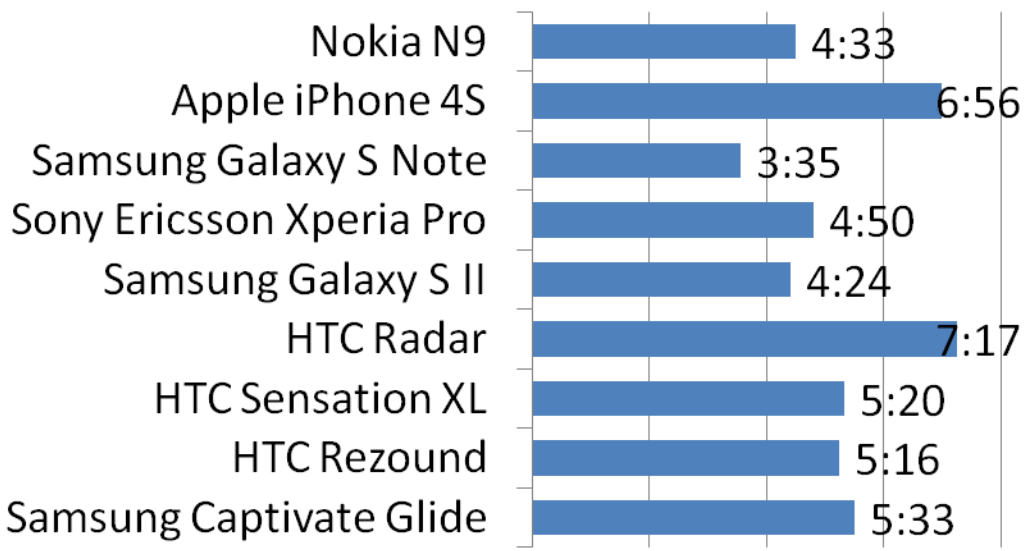


- Wireless network availability & intermittency
  - Barrier, Moving, Weather effect, Radio interference
- Network bandwidth and latency
  - Network coverage, Request congestion

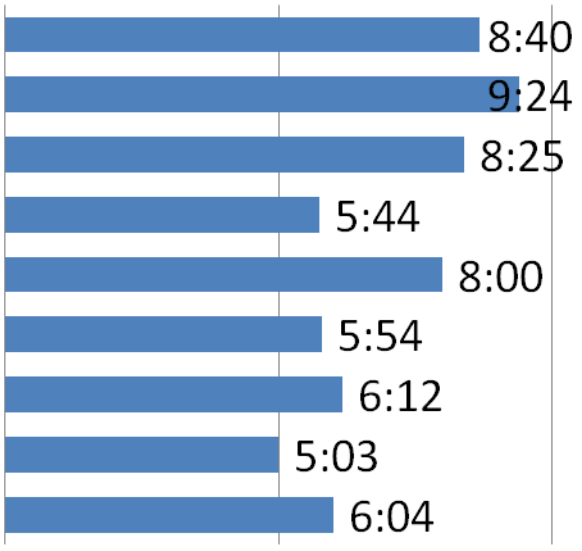
How severe is the energy problem on today's mobile handheld devices?

- Limited battery capacity (online video 1~2 hours…)
- Increasing demand for energy-hungry apps

MAUI[ACM MobiSys 2010]



Web Browsing



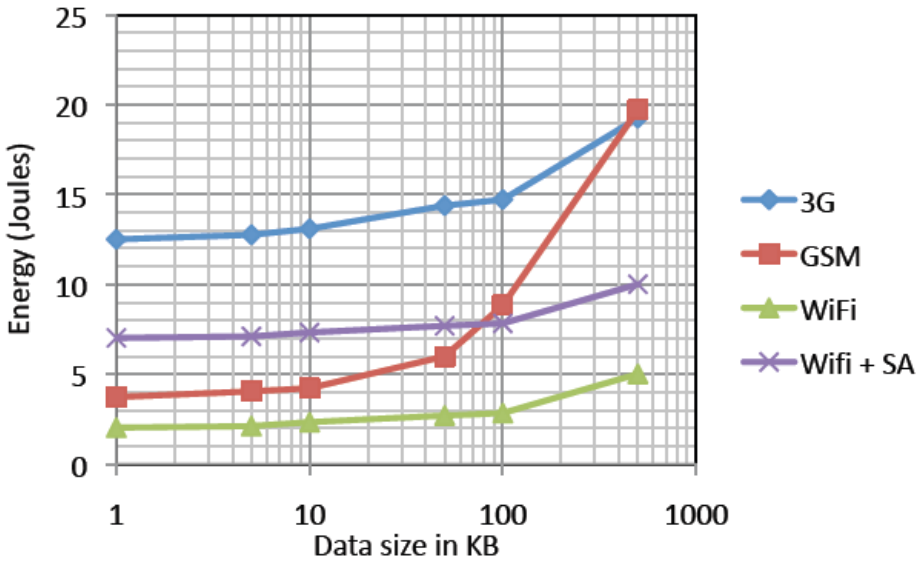
Local Video Playback

From www.GsmArena.com

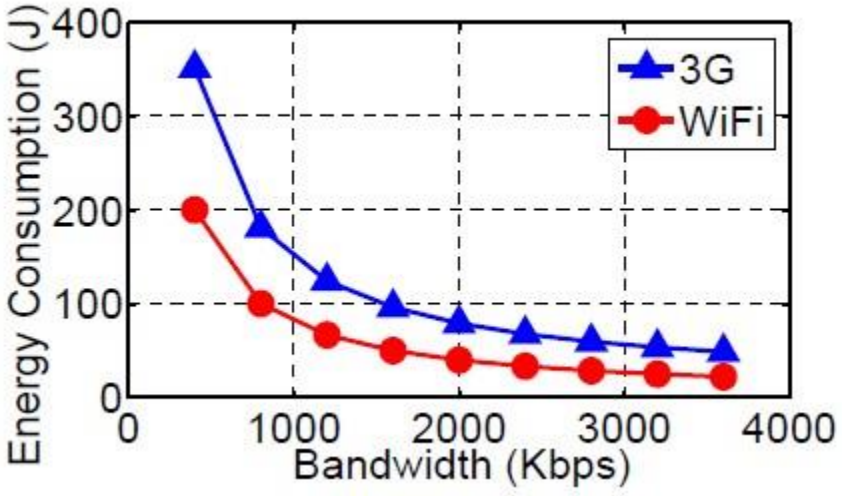
Our Real Measurements:

Energy consumption vs. Network connectivity

Energy consumption
&
Selected Link



Energy consumption
&
Bandwidth



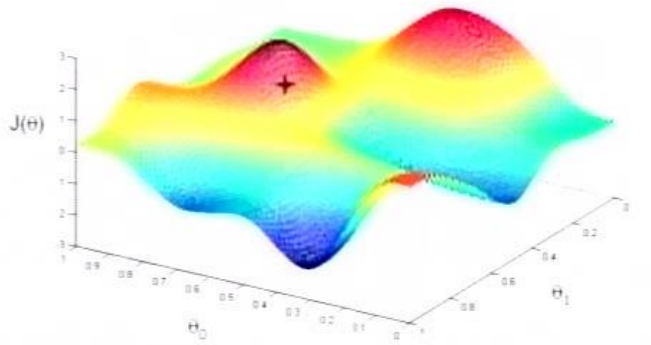
MAUI[ACM MobiSys 2010]

Kelényi[PM2HW2N 2009]

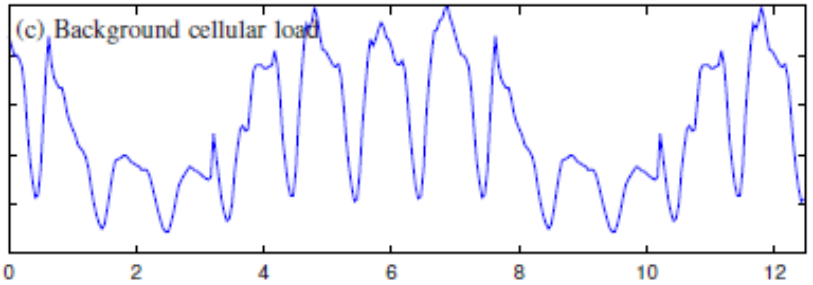
① Try to transmit data in good network connectivity to achieve energy-efficiency.

Network connectivity and the User Context

Signal intensity & Location

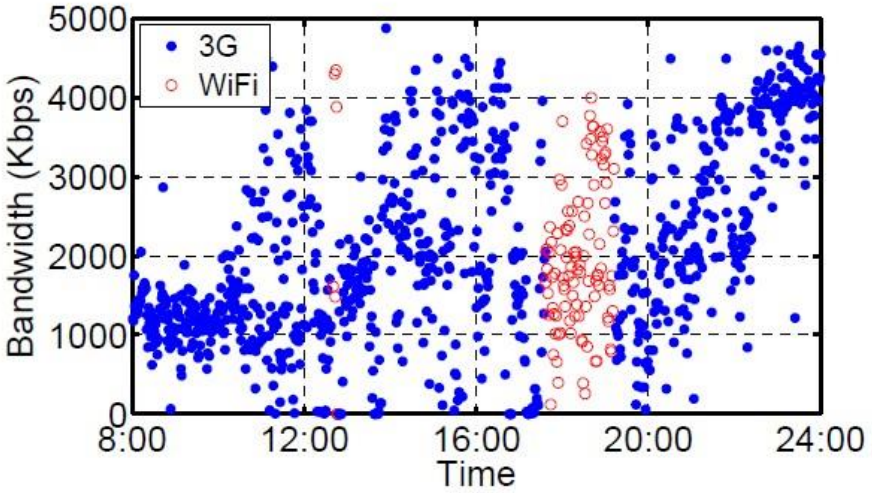


Cellular Load & Time



Time[days]

Bandwidth and Time

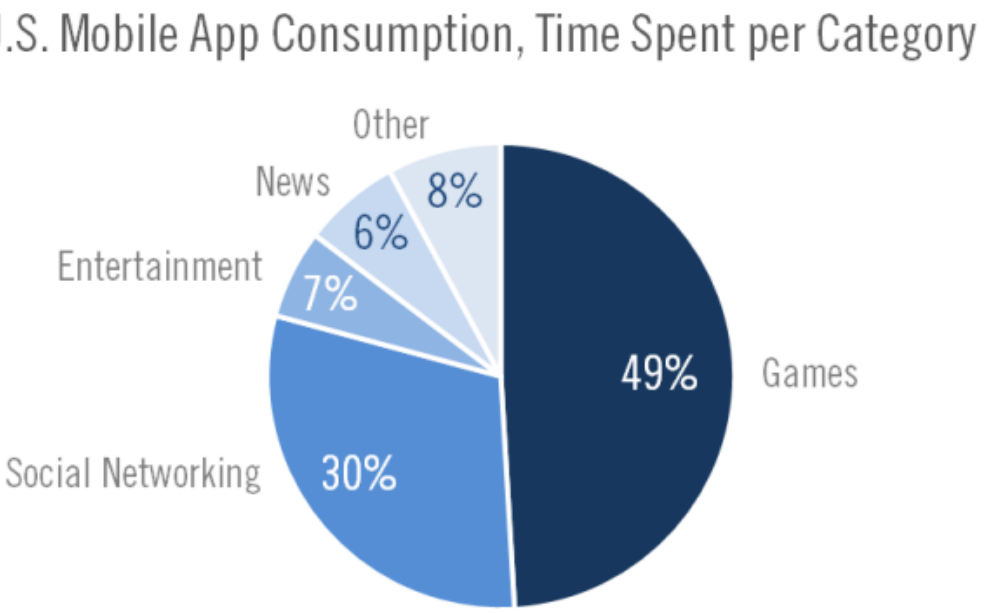


② Choose the good timing to transmit data

The **flexibility** to **schedule** data transmissions in many Apps

Prefetching Friendly
- News Feed, Maps, SNS

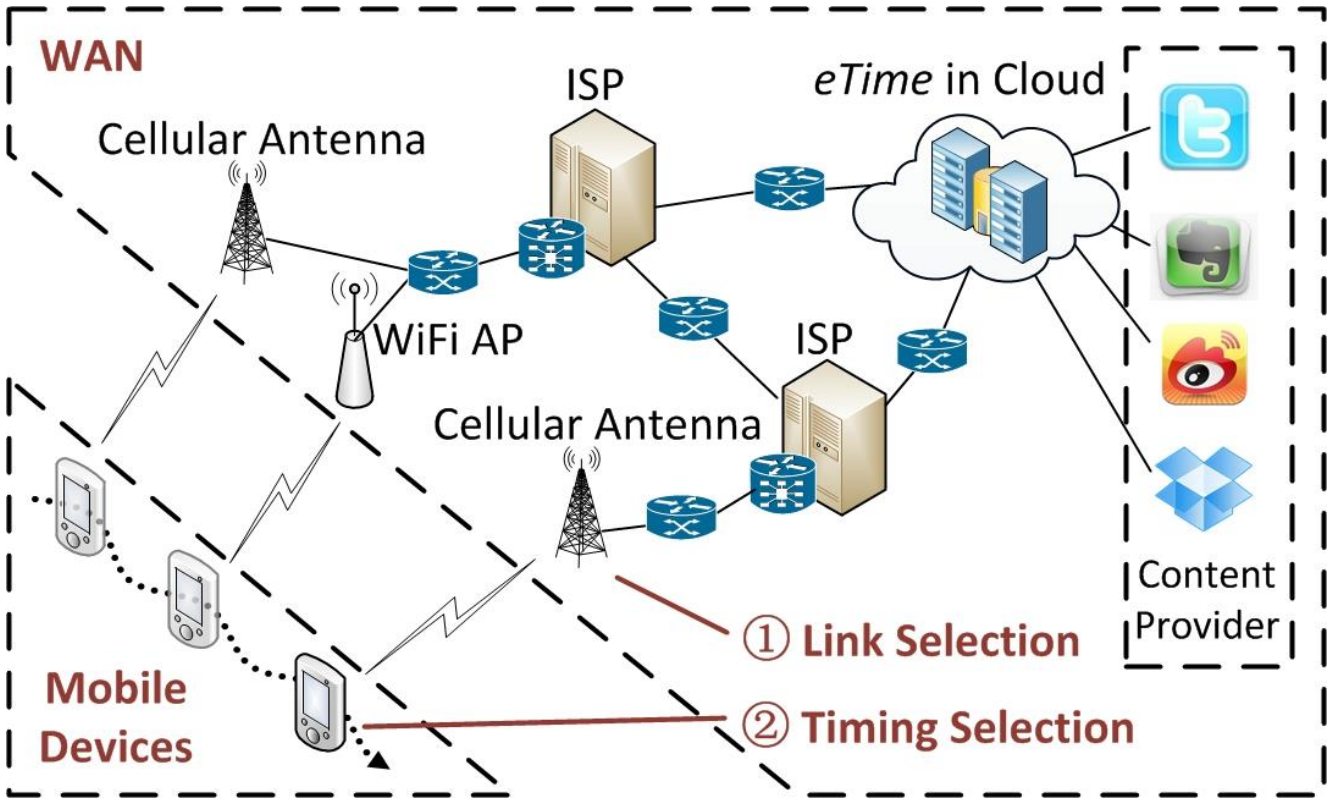
Delay-tolerant
- iCloud, File-sharing
- Software update



③Adaptively seize the good connectivity to prefetch frequently used data while deferring delay-tolerant data.

eTime:

Energy-Efficient Transmission between Cloud & Mobile Devices



1. Employ cloud to manage data in different Apps.

2. Energy optimization under the volatility of wireless network

eTime:
Energy-Efficient Transmission between Cloud & Mobile Devices

- **eTime: Energy-Efficient Transmission between Cloud and Mobile Devices, IEEE INFOCOM (Mini-conference), 2013.**
- **Complete software to be deployed in Android market**

- **Gearing Resource-Poor Mobile Devices with Powerful Clouds: Architectures, Challenges and Applications, IEEE Wireless Communications Magazine, 2013.**
- **Latest comprehensive survey on mobile cloud**

Outline

- Introduction
- Online Cloud Services ➔ case study
  - FS2You: Online Hosting & Content Distribution
  - Novasky: Cinematic-Quality VoD in a P2P Storage Cloud
  - *eTime*: Mobile Cloud
- **Underlying Datacenter Optimization**
  - **Network Virtualization for Multi-tenants DC**
  - Green DC Power-Performance Tradeoffs
- Future Plan & Collaboration

IaaS Clouds Hosting Increasingly More Apps

Datacenters for IaaS cloud services

amazon web services™

Windows Azure

Google Compute Engine

NETFLIX

dolphin

PPTV

Unilever

360 www.360.cn

Pinterest

Flipboard

NEXON

yelp

36% growth

$20 billion

IaaS DCN: Challenges & Opportunities

Today's IaaS cloud

- Shared & Multiplexed across many tenants
- Pay-per-usage charging model via different types of virtual machines (VMs)
  - Only true for: CPU, memory, storage

However

- Intra-DC network resources shared in best effort manner based on traditional protocols, e.g., TCP
- Bandwidth is not fairly shared based on payment

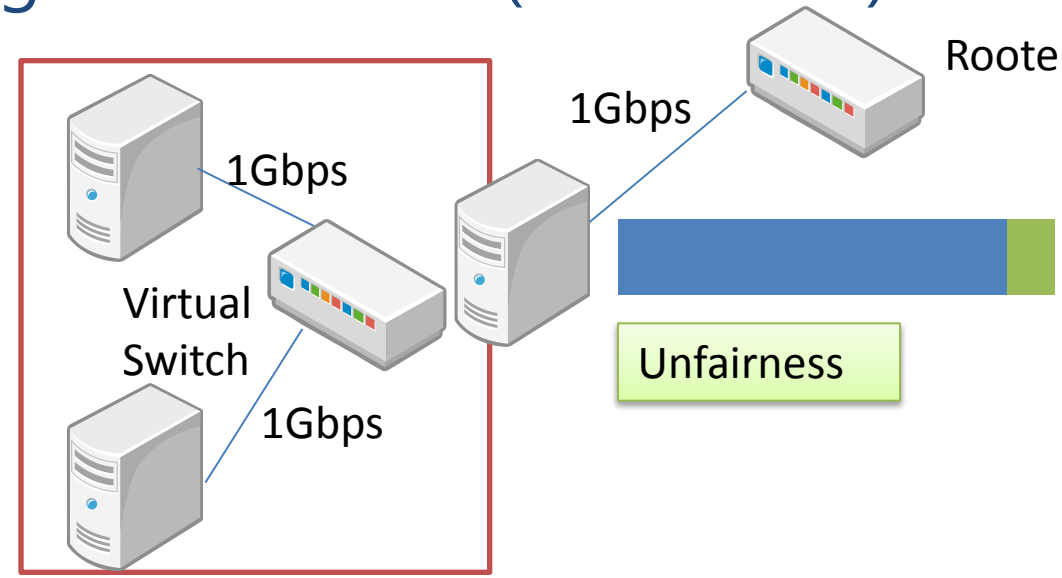
- Unpredictable/varying performance, *e.g., job finish times*
  - ➔ Lack of performance isolation/performance guarantee for VMs
  - ➔ NO charge on quantified intra-DCN bandwidth
  - Remind that Providers do charge you for CPU, Memory, Storage…
  - **Virtualization became mature except for Networking….**

Issue I: Fairness

Example of sharing DC networks (best effort):

VM0: infinite demand
Less connections

VM1: infinite demand
Large amount connections



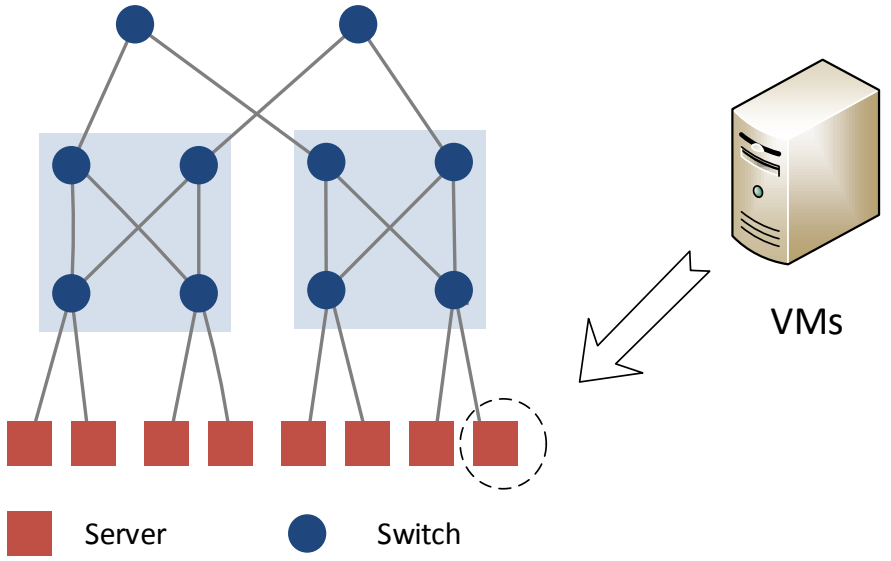
- Relying on TCP's congestion control : flow-level fairness
- The network allocation depends on: 1) VMs running on the same machine, 2) cross-traffic on each link used by the VM

Issue II: Guarantee

An existing approach

Allocate VMs in the topos

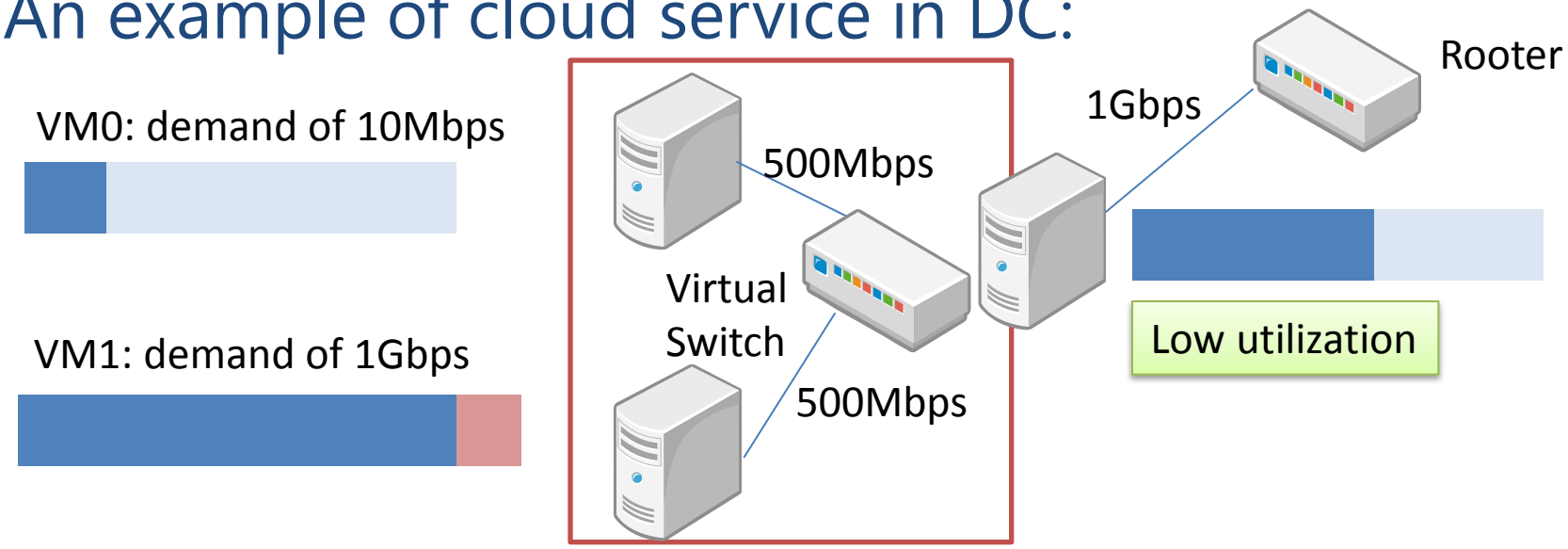
Reserve bandwidth for virtual clusters



Bandwidth guarantee

Issue III: Utilization

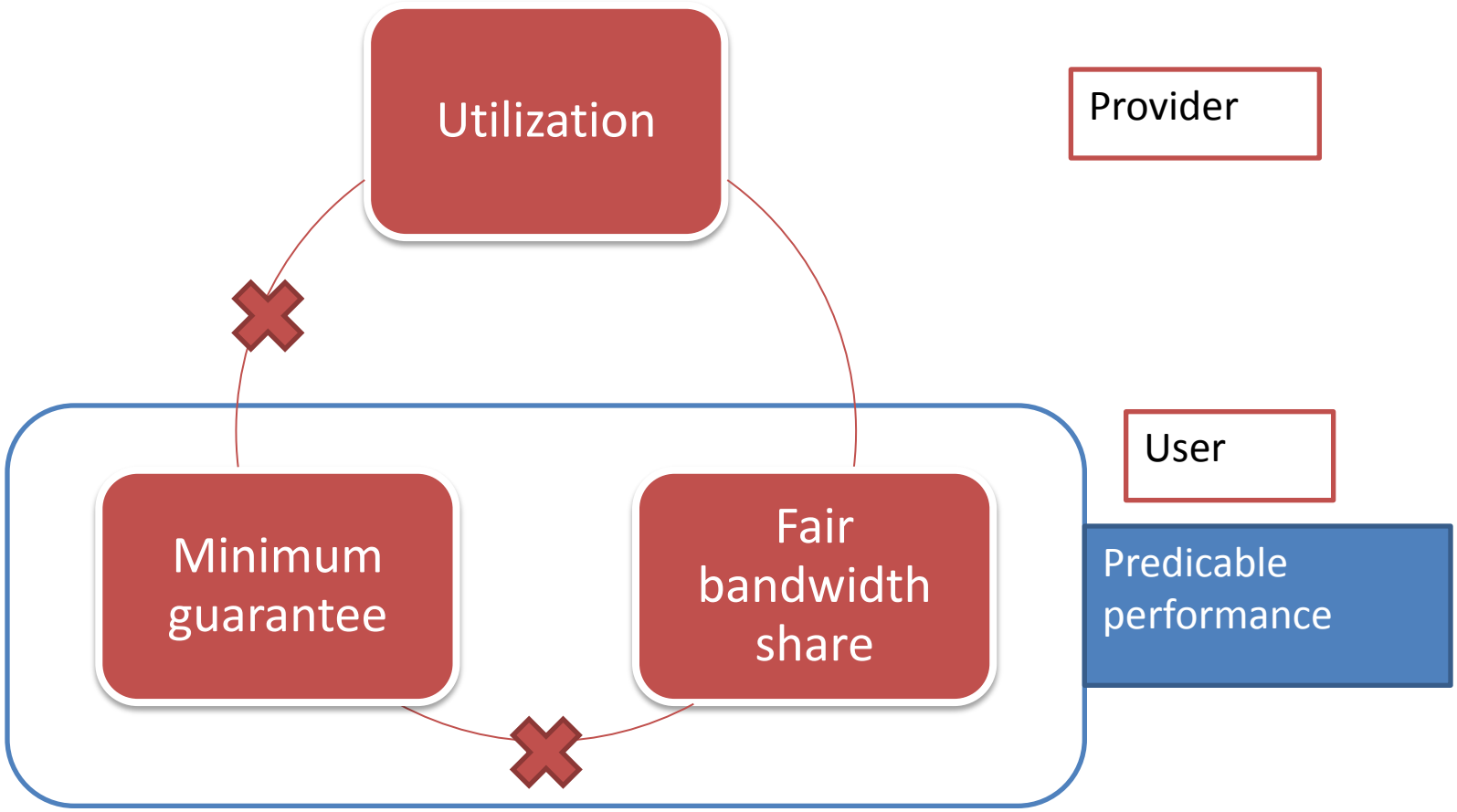
An example of cloud service in DC:



- The networking demands of cloud applications are time-varying
- Low network utilization if statically reserved

A Large Design Space for 3-way Tradeoffs

Providers & Tenants are mutually interested

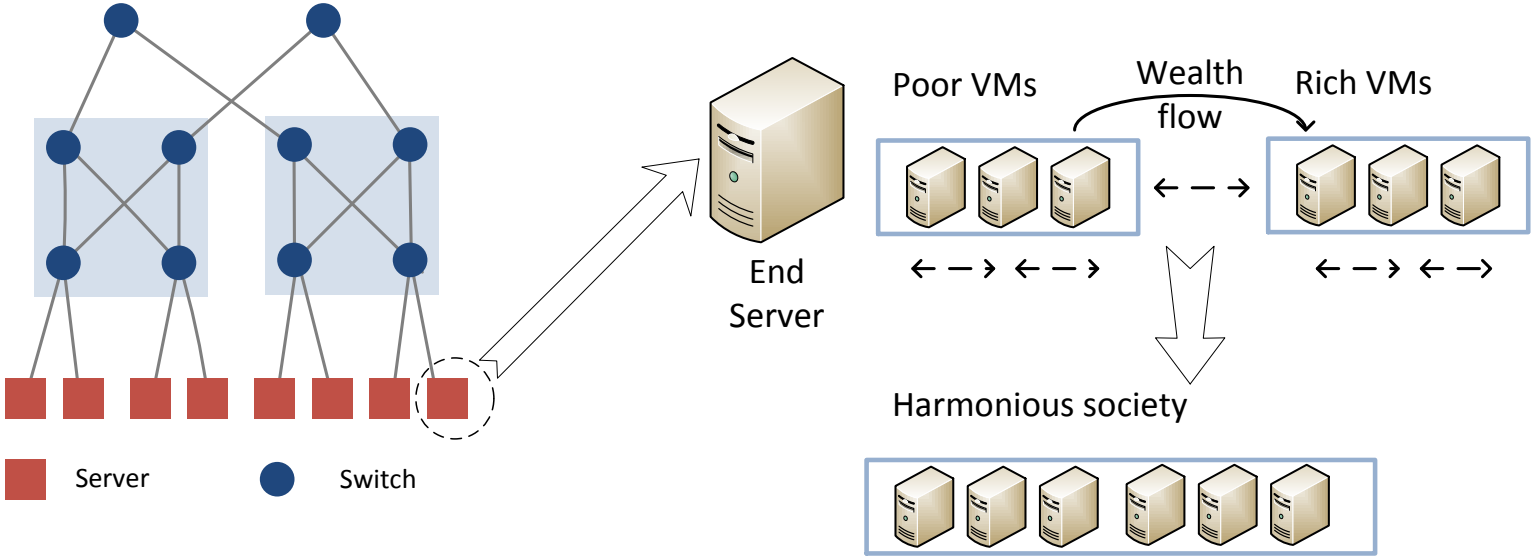


A Cooperative Game Based Allocation for Sharing Data Center Networks

IEEE INFOCOM 2013

1st paper to apply game theory in DCN sharing

- Poor VMs: base bandwidth > bandwidth demand (Bi>Di)
- Rich competitor: base bandwidth ≤ bandwidth demand (Bi≤Ni)

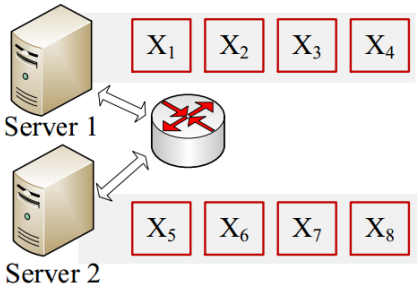


- **Fairness-Minimum Guarantee Tradeoff**
  - 1) **Minimum bandwidth guarantee** for the poor
  - 2) Maintain **proportionality** among the rich

- **System Resource Utilization ➔ Social welfare**

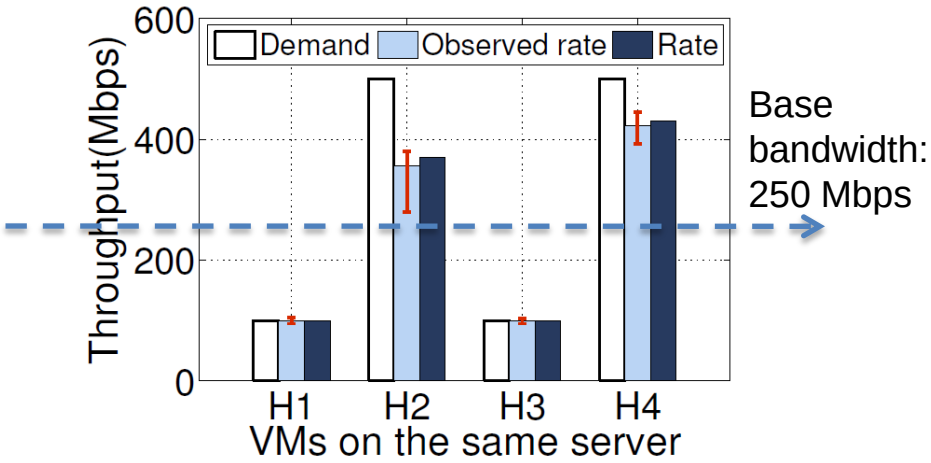
Experimental Results

Consider two servers

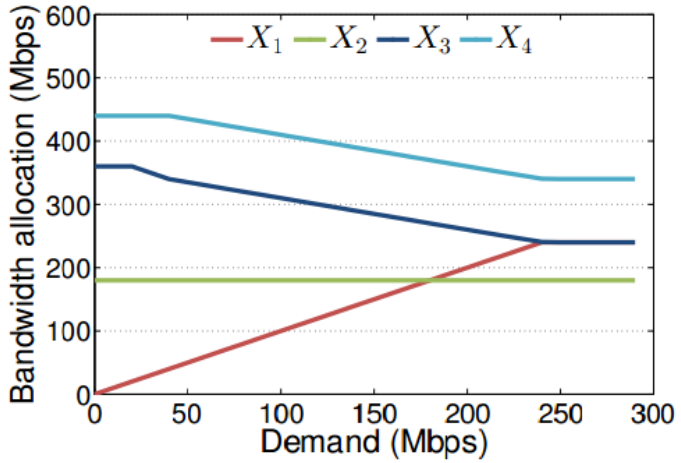


How the VMs compete for the bandwidth

Bandwidth for poor/ rich

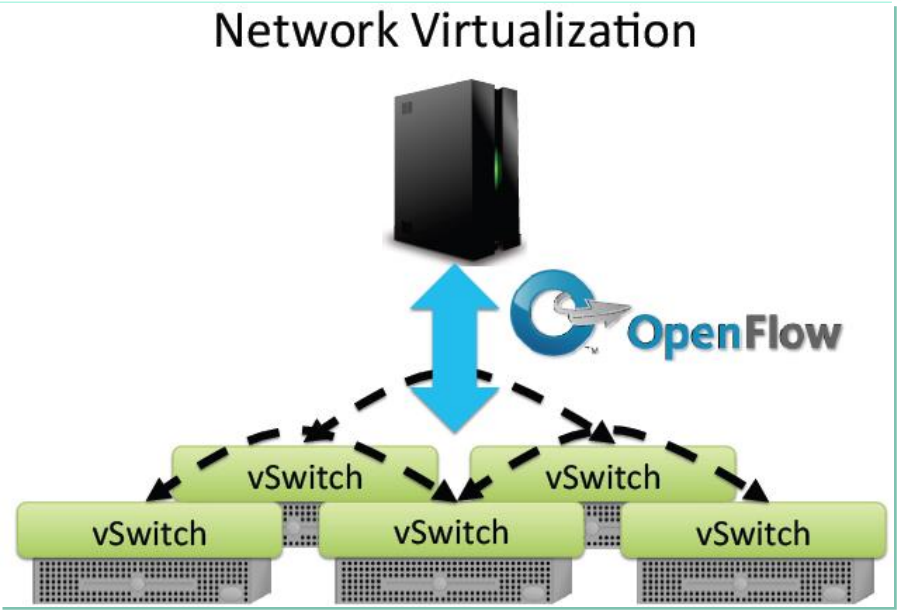


Increasing the demand of one VM



Falloc: Fair Network Bandwidth Allocation in IaaS Datacenters Via a Bargaining Game Approach

IEEE ICNP 2013



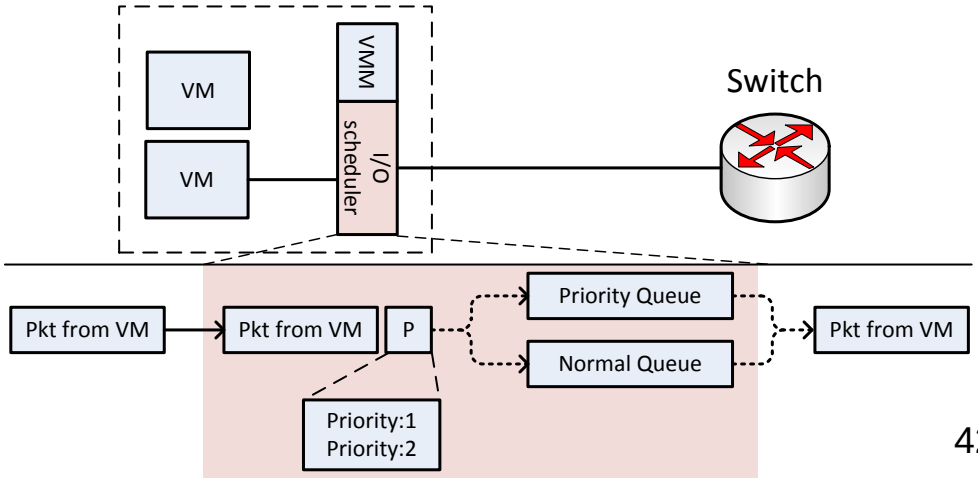
w.r.t Related Works (*e.g.*, Sigcomm'12 on "reasoning" & heuristics)

➔Fill the gap: rigorous theory foundation & understanding

➔1st paper for applying game theory in DC network sharing

Implemented with OpenFlow

- run our proposed bandwidth allocation algorithm in a centralized **controller**
- **Enforce** the allocation result by forwarding packets through specified queues in the switches
- **Mininet Evaluation** ➔ a SDN platform running real network protocols and workloads, ➔ the developed code can be moved to a real OpenFlow network without any change



Outline

- Introduction
- Online Cloud Services ➔ case study
  - FS2You: Online Hosting & Content Distribution
  - Novasky: Cinematic-Quality VoD in a P2P Storage Cloud
  - *eTime*: Mobile Cloud
- Underlying Datacenter Optimization
  - Network Virtualization for Multi-tenants DC
  - **Green DC Power-Performance Tradeoffs**
- Future Plan & Collaboration

More cross-discipline issues …

Energy costs & operational economics

On Arbitrating Power-PerformanceTradeoff in SaaS Clouds
TPDS, INFOCOM 2013

SmartDPSS: Cost-Minimizing Multi-source Power Supply for Datacenters with Arbitrary Demand
ICDCS 2013, ACM e-Energy 2013

Green computing & environmental impact

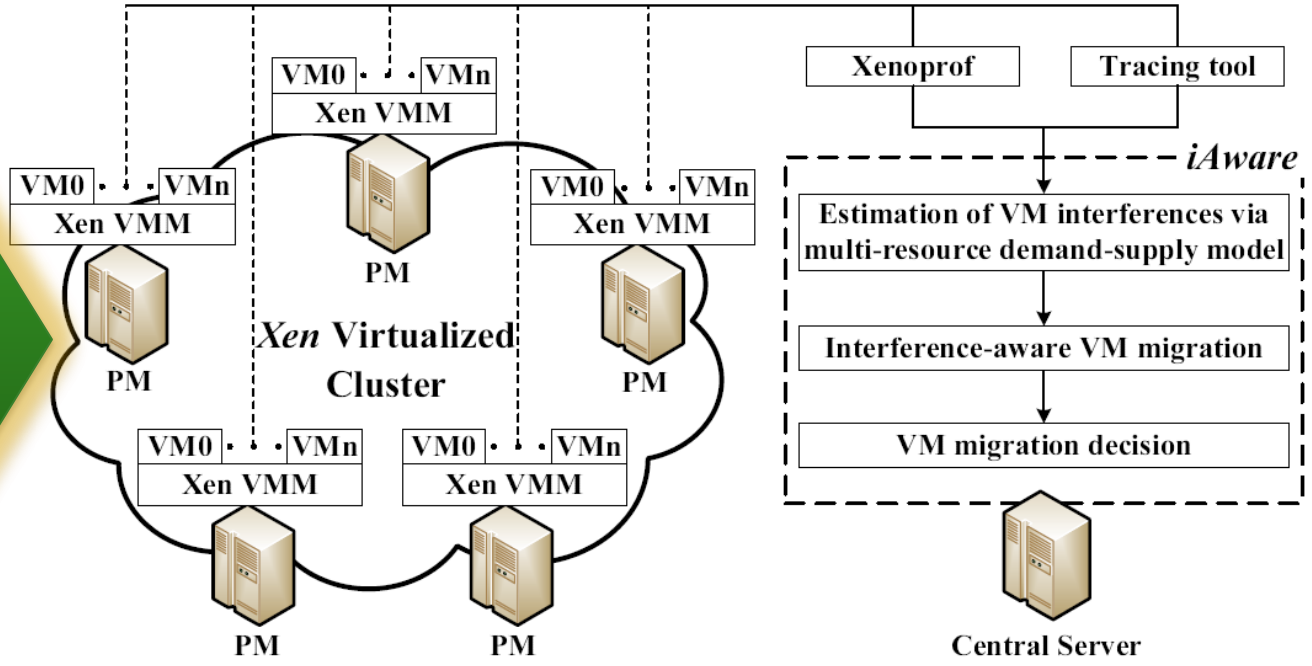
Carbon-aware Load Balancing for Geo-distributed Cloud Services
MASCOTS 2013

Harnessing Renewable Energy in Cloud Datacenters: Opportunities and Challenges
IEEE Network

虚拟机性能干扰问题 ➔ 严重影响云计算SLA

- **We have built *iAware* ➔ realistic *Xen*-virtualized clusters for extensive measurements**
  - **Quantify ➔ What/Where is VM migration interference?**
  - **Quantify ➔ What/Where is VM co-location interference?**
  - **Unveil those unrealistic assumptions on VM-related costs & overheads, that are made in many existing papers…….**

Our *iAware* Virtualized System Implementation



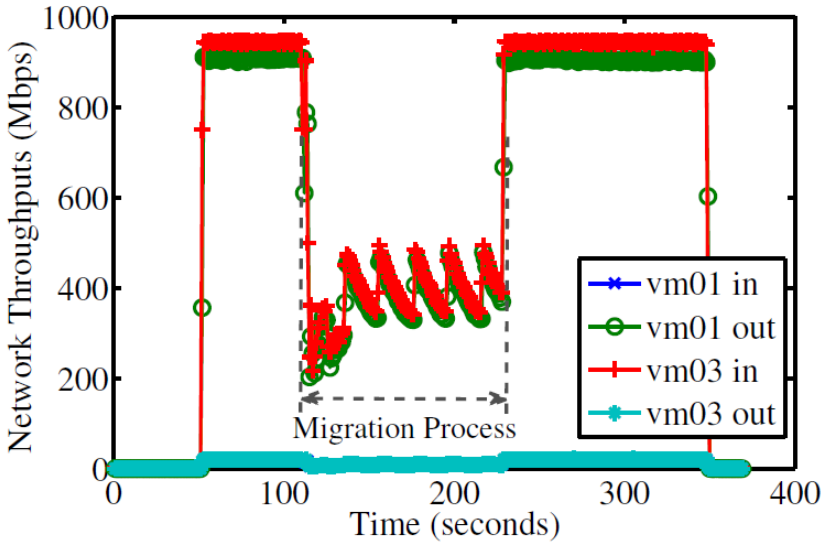
Our Solution

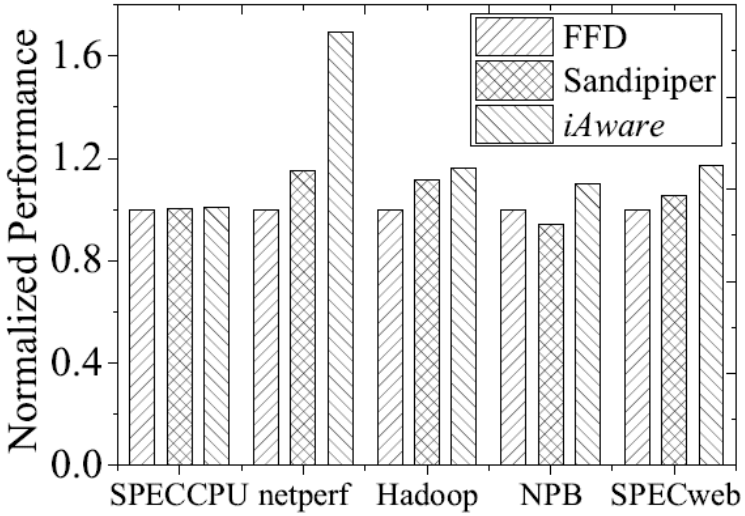
iAware: Interference-Aware VM Live Migration in the Cloud

IEEE Transactions on Computers 2013

Proceedings of the IEEE (Impact Factor: 6.911)

- **Our solution can achieve load balancing and power saving purposes without sacrificing performance (➔mitigating VM performance interferences)**





Outline

- Introduction
- Online Cloud Services ➔ case study
  - FS2You: Online Hosting & Content Distribution
  - Novasky: Cinematic-Quality VoD in a P2P Storage Cloud
  - *eTime*: Mobile Cloud
- Underlying Datacenter Optimization
  - Network Virtualization for Multi-tenants DC
  - Green DC Power-Performance Tradeoffs
- **Future Plan & Collaboration**

GPC 2014 Call For Papers

- **The 9th GPC highlights a new theme on Green, Pervasive & Cloud Computing**
  - **Paper Submission ➔ Jan. 16, 2014 (textended ☺)**
  - **Wuhan, China, May 26-28, 2014**
- **Publication Model**
  - **20 top ranked papers ➔ SCI-indexed journal**
  - **Other selected papers ➔ Springer Lecture Notes in Computer Science**
- **Program Chair**
  - **Fangming Liu (刘方明), Huazhong University of Sci. & Tech.**
- **General Co-Chairs**
  - **Erol Gelenbe, Imperial College, UK, Fellow of IEEE, ACM and IET**
  - **Laurence T. Yang, Huazhong University of Sci. & Tech.**
- **Steering Committee Chair**
  - **Hai Jin, Huazhong University of Sci. & Tech.**
- **TPC Members (Approved)**
  - **Ivan Stojmenovic, Francis Lau… see more: http://grid.hust.edu.cn/gpc2014**

欢迎投稿！ 敬请宣传！

Thank You

Q&A

Fangming Liu (刘方明), Associate Professor
Huazhong University of Sci. & Tech.
http://grid.hust.edu.cn/fmliu/