The Impact of Mobile Multimedia Applications on Data Center Consolidation

Kiryong Ha*, Padmanabhan Pillai[†], Grace Lewis[‡], Soumya Simanta[‡], Sarah Clinch [§], Nigel Davies [§], and Mahadev Satyanarayanan*
*Carnegie Mellon University, [†]Intel Labs, [‡]CMU-SEI, [§] Lancaster University

3/27/2013

Consolidation is the key to Cloud computing

- Cloud computing achieves economies of scale by consolidating resources
 - It lower the marginal cost of operation and management
 - Amazons EC2 spans the entire planet with 7 data centers
- Large separation between a mobile device and its cloud
 - High latency in end-to-end communication
 - Challenges for latency sensitive applications



From http://gcn.com/articles/2012/10/26/agency-data-centers-idc-report.aspx

Emerging Mobile Applications

A new class of applications using video/voice in mobile context

- Apple's Siri, Augmented reality applications
- Many of these are interactive as well as resource-intensive
 - Beyond the processing and storage limits of mobile device
 - Use Cloud to offload execution
- Wearable Devices like Google Glass will push this trend more





From http://www.google.com/glass/start/ and http://www.flickr.com/photos/x1brett/4600461689/

Questions

- 1. Do we really need to offload?
- 2. What's the effect of Cloud location?
- 3. In situation where public Clouds are inadequate, what would be an alternative Cloud architecture?

Applications Studied

Applications

• Resource intensive and latency sensitive applications

Application	Input	Output	Execution Environment
Face Recognition	Image	Name and position of a detected face	OpenCV on Windows
Speech Recognition	Audio	Words in plain text format	Java on Linux
Object Recognition	Image	Names and positions of found objects	C++ on Linux
Augmented Reality	Image	Name (information) of recognized landmark	OpenCV & Intel IPP on Windows
Fluid Simulation	Acc data	Location and pressure of simulated particles	C++ on Linux

1. Do we need to offload?

Extremes of resource demands

- It may appear that today's smartphone are already powerful enough
 - Built-in support for face detection in Android SDK
- But, computation varies dramatically depend on operational conditions
 - Face recognition
 - 1. Increasing number of possible faces,
 - 2. Reducing the constraints on the observation conditions

1. Do we need to offload?

High variability of execution time

1. SPEECH recognition

• Execution time increases with the number of recognized words

	no words	1-5 words	6-22 words
	recognized	recognized	recognized
Measured average time	0.057 s	1.04 s	4.08 s

2. FACE recognition

• Execution time varies depends on the contents of images

Image with single big face	Image with no face	
0.30 s	3.92 s	



* Tested with a netbook as a mobile device

1. Do we need to offload?

Improvement from Cloud offloading

• Though variability still exists, the absolute response times are improved

	No C	Cloud	With Cloud	
	median	99%	median	99%
SPEECH (500 requests)	1.22 s	6.69 s	0.23 s	1.25 s
FACE (300 requests)	0.42 s	4.12 s	0.16 s	1.47 s

* We used Amazon EC2 for SPEECH and private Cloud for FACE.

** We used netbook as a mobile device.

2. Effects of Cloud location

Experiment setup

- Mobile (No offloading) : Local execution within mobile device
- Cloud : Offload to Cloud (Amazon EC2)
 - 4 locations: US-East, US-West, EU, Asia
- 1WiFi (One hop)
- : Offload to nearby compute resource



- Fast mobile/Cloud and a slow Cloudlet configuration
- 1 EC2 Compute Unit \approx 1.0-1.2 GHz 2007 Xeon processor



2. Effects of Cloud location

Network measurement

- Average RTT from 260 global vantage points to an "optimal" Amazon EC2 instance is 73.68 ms [1]
- [CMU Amazon East] has amazingly good connection, but becomes similar with Amazon West case if it is measured off-campus

EC2	Throughp	ut (Mbps)	Latency (RTT, ms)	
	То	From	median	
East	28	34	9.2	
West	12	14	92	
EU	3.6	0.9	99	
Asia	10	0.5	265	

< From mobile to different Cloud via 802.11n >



Impact on response time



Impact on response time



Impact on Energy Usage

Energy consumption per query on mobile device

- Reduce energy consumption by a factor of 3 in 1WiFi
- Offload to far Cloud wastes energy while waiting the response.



* netbook's baseline idle power dissipation is 10W

Effects of Cloud location

- Proximity to the data center is essential
 - Response time & Energy consumption
- 1WiFi Cloud
 - The best attainable proximity
 - The emerging of such applications can be accelerated by deploying infrastructure that assures continuous proximity to the cloud
- Contradictory requirements
 - 1WiFi is valuable for mobile computing
 - However, it needs many data centers at the edges the Internet
 - Lowers the advantages of consolidation

 \rightarrow How can we reconcile this conflict?



3. What's an alternative Cloud architecture?

Hierarchical organization of data center

- Level 1
 - Today's unmodified Cloud
- Level 2
 - Stateless data centers at the edge
 - Appliance-like deployment model
 - 1. Only soft state
 - Cached virtual machine images
 - Cached files from DFS
 - 2. No hard state keeps management overhead low.



* soft state is state which can be regenerated or replaced

3. Alternative Cloud Architecture

Physical realization

- Hardware technology is already out today
- Repurpose as Level 2 data center by removing hard state and adding self-provisioning



Myoonet's Outdoor micro data center [2]



AOL's micro data center [3]



3. Alternative Cloud Architecture

Operating Environment

- Commonalities in the requirement between Levels 1 and 2
 - 1. Strong **isolation** between untrusted user-level computations
 - 2. Mechanisms for authentications, access control and metering
 - 3. Dynamic resource allocation
 - 4. The ability to support a wide range of user-level computations
 - → Virtual machine as Cloud of today like EC2

Industrial efforts for 1WiFi Cloud

Nokia Siemens Networks and IBM

- Announced Liquid Applications at MWC 2013 ۲
- Deploys the cloud into the (cellular) base station ۲
 - *"Improved latency can enable high-value vertical solutions"..*



Discussion & Future work

Rapid Provisioning

Provisioning delay directly impact usability*

Discovery

• How to dynamically find the right Level 2 data center?

Data placement

- Appropriate data placement is important for many "Cloud" applications
- Two extreme ends
 - 1. Application with relatively small data set for its operation
 - 2. A very large data set accessed in an unpredictable manner
- Most applications likely fall between the two ends
 - Map Data: physical location is highly correlated with accessed data map
 - Automatic caching of distributed file system exploiting locality information

* Kiryong Ha et al. "Just-in-Time Provisioning for Cyber Foraging", In MobiSys 2013 (to appear) 3/27/2013

Conclusion

- Do we really need to offload?
 Cloud offload is here to stay
- 2. What's the effect of Cloud location?

Emerging mobile applications endanger cloud consolidation

 In situation where public Clouds are inadequate, what would be an alternative Cloud architecture?
 Hierarchical solution is desirable and feasible

Reference

- 1) A. Li, X. Yang, S. Kandula, and M. Zhang, "Cloudcmp: comparing public cloud providers," in Proceedings of the 10th annual conference on Internet measurement. ACM, 2010, pp. 1–14.
- 2) Myoonet, "Unique Scalable Data Centers," December 2011, <u>http://www.myoonet.com/unique.html</u>.
- 3) R. Miller, "AOL Brings Micro Data Center Indoors, Adds Wheels," <u>http://www.datacenterknowledge.com/archives/2012/08/13/aol-brings-micro-data-center-indoors-adds-wheels</u>, August 2012.
- 4) Kiryong Ha, Padmanabhan Pillai, Wolfgang Richter, Yoshihisa Abe, Mahadev Satyanarayanan, "Just-in-Time Provisioning for Cyber Foraging", In MobiSys 2013 (to appear)
- 5) D. Kucera, "Amazon apologizes for Christmas Eve outage affecting Netflix," December 2012, http://articles.washingtonpost.com/2012-12-31/business/361036071amazon-web-serviceschristmas-eve-outage-largest-online-retailer
- 6) M. Satyanarayanan, "The Evolution of Coda," ACM Transactions on Computer Systems, vol. 20, no. 2, May 2002.
- 7) S. Rhea, P. Eaton, D. Geels, H. Weatherspoon, B. Zhao, and J. Kubiatowicz, "Pond: the oceanstore prototype," in Proceedings of the 2nd USENIX Conference on File and Storage Technologies, 2003, pp. 1–14.