Let Latency Guide You: Black-box characterisation of Cloud Application Performance

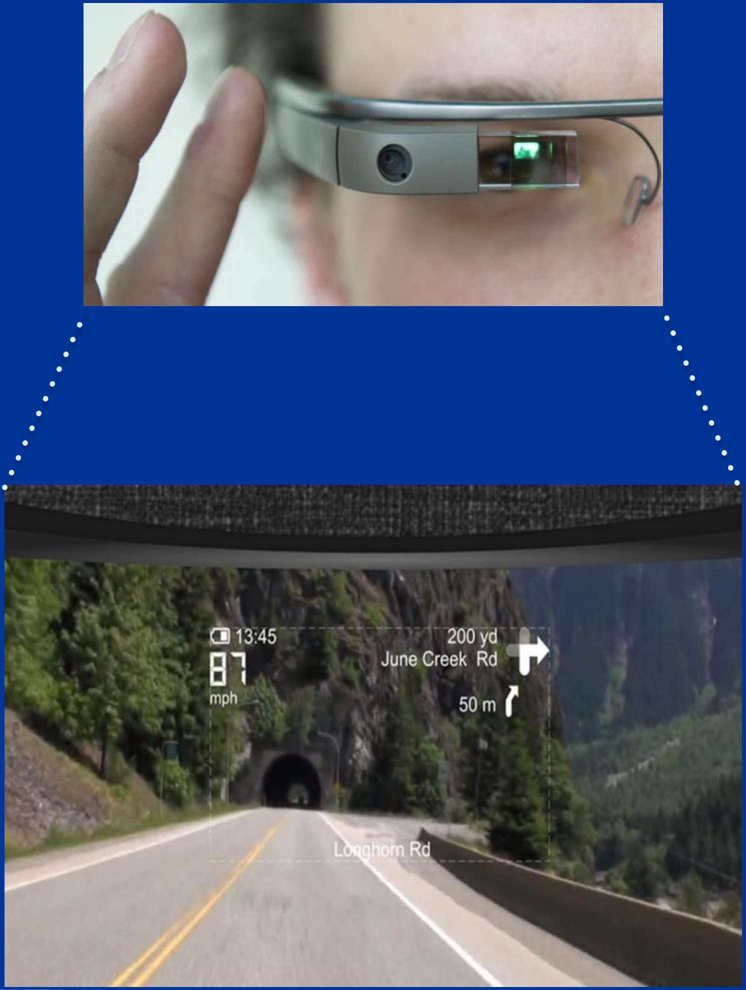
Hamed Saljooghinejad, Felix Cuadrado, Steve Uhlig
Networks Research Group

PhD Student

11th July 2014

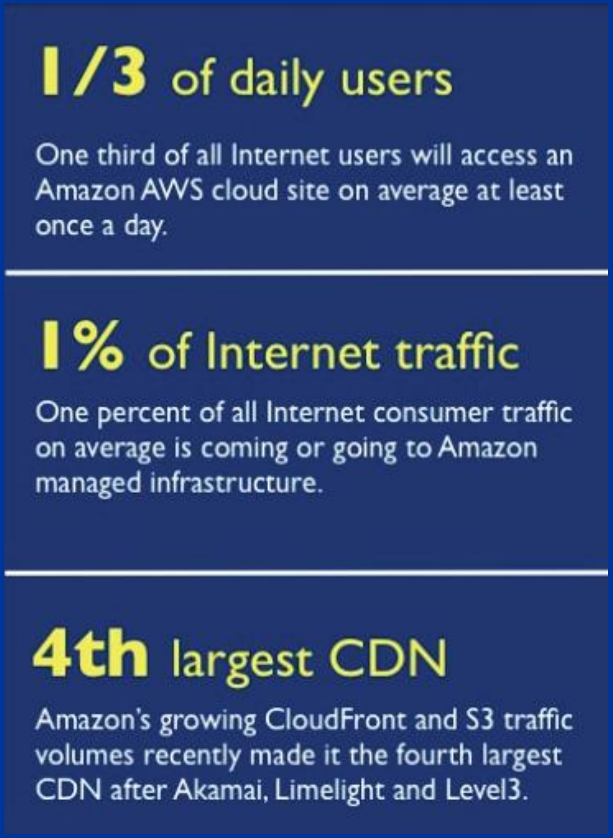
LATENCY MATTERS

Third-Party Apps



PUBLIC CLOUDS ARE USED EVERYWHERE

Amazon Cloud



1/3 of daily users

One third of all Internet users will access an Amazon AWS cloud site on average at least once a day.

1% of Internet traffic

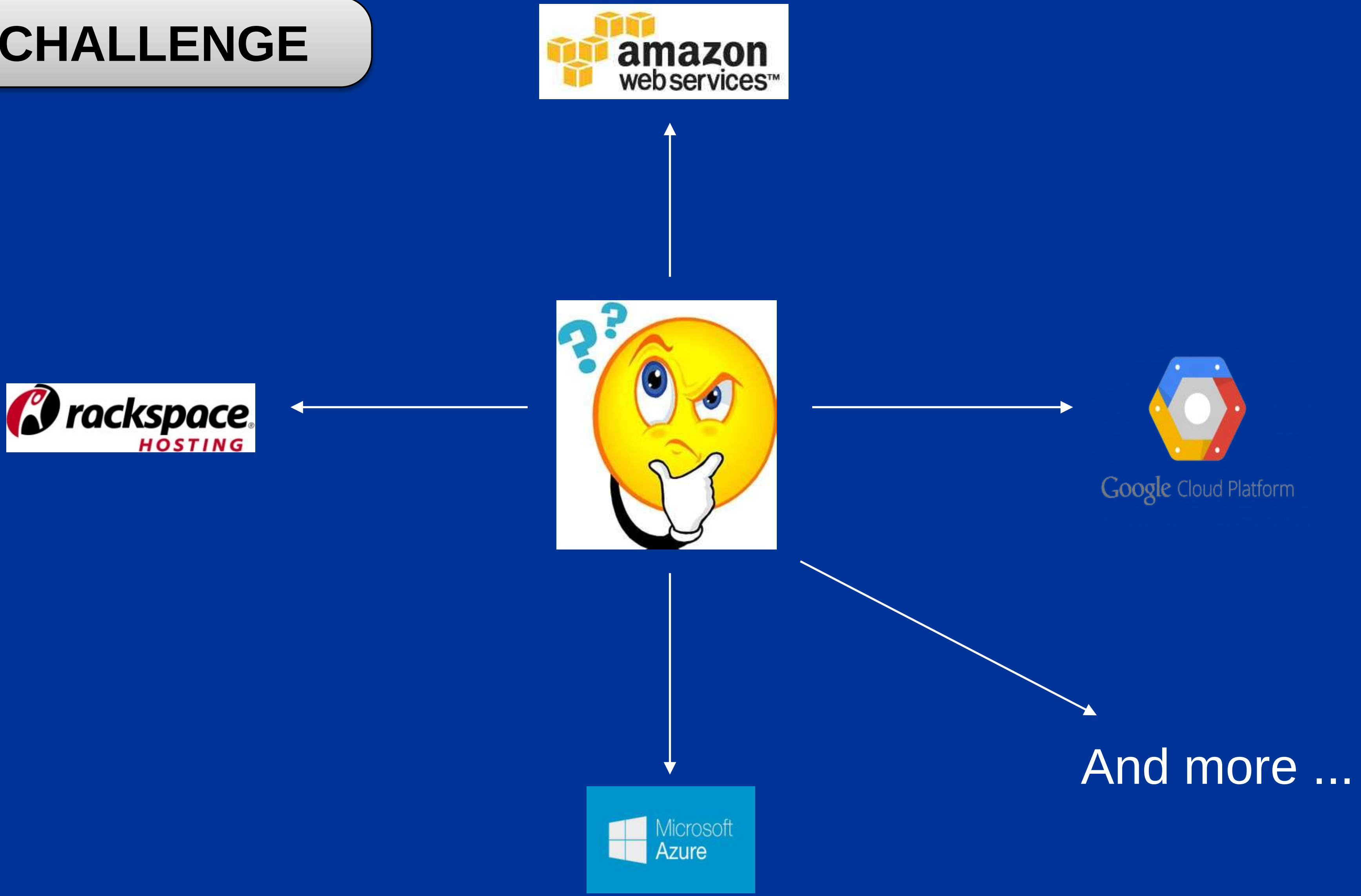
One percent of all Internet consumer traffic on average is coming or going to Amazon managed infrastructure.

4th largest CDN

Amazon's growing CloudFront and S3 traffic volumes recently made it the fourth largest CDN after Akamai, Limelight and Level3.

Craig Labovits, Deep Field, April 2012

CHALLENGE



And more ...

CHALLENGE

Model= t2.micro
vCPU=1
CPU Credits/hour=6
Mem (GiB)=1
Storage (GB)=EBS Only

Model= t2.small
vCPU=1
CPU Credits/hour=12
Mem (GiB)=2
Storage (GB)=EBS Only

Model= t2.medium
vCPU=2
CPU Credits/hour=24
Mem (GiB)=4
Storage (GB)=EBS Only

Model= m2.medium
vCPU=1
Mem (GiB)=3.75
SSD Storage (GB)=1*4

Model= i2.4xlarge
vCPU=16
Mem (GiB)=122
SSD Storage (GB)=4*800

Model= hs1.8xlarge
vCPU=16
Mem (GiB)=177
SSD
Storage (GB)=24*2048800

amazon web services™

Model= r3.large
vCPU=2
Mem (GiB)=15.25
SSD Storage (GB)=1*32

Model= m3.large
vCPU=2
Mem (GiB)=7.5
SSD Storage (GB)=1*32

Model= i2.2xlarge
vCPU=8
Mem (GiB)=61
SSD Storage (GB)=2*800

Model= i2.8xlarge
vCPU=32
Mem (GiB)=244
SSD Storage (GB)=8*800

Model= r3.xlarge
vCPU=4
Mem (GiB)=30.5
SSD Storage (GB)=1*80

Model= m3.xlarge
vCPU=4
Mem (GiB)=15
SSD Storage (GB)=2*40

Google Cloud Platform

Model= c3.8xlarge
vCPU=32
Mem (GiB)=60
SSD Storage (GB)=2*320

Model= m3.2xlarge
vCPU=8
Mem (GiB)=30
SSD Storage (GB)=2*80

rackspace HOSTING

Model= i2.xlarge
vCPU=4
Mem (GiB)=30.5
SSD Storage (GB)=1*800

Model= g2.2xlarge
vCPU=8
Mem (GiB)=15
SSD Storage (GB)=1*160

Microsoft Azure

Model= c3.4xlarge
vCPU=16
Mem (GiB)=30
SSD Storage (GB)=2*160

Model= c3.large
vCPU=2
Mem (GiB)=3.75
SSD Storage (GB)=2*16

Model= r3.8xlarge
vCPU=32
Mem (GiB)=244
SSD Storage (GB)=2*320

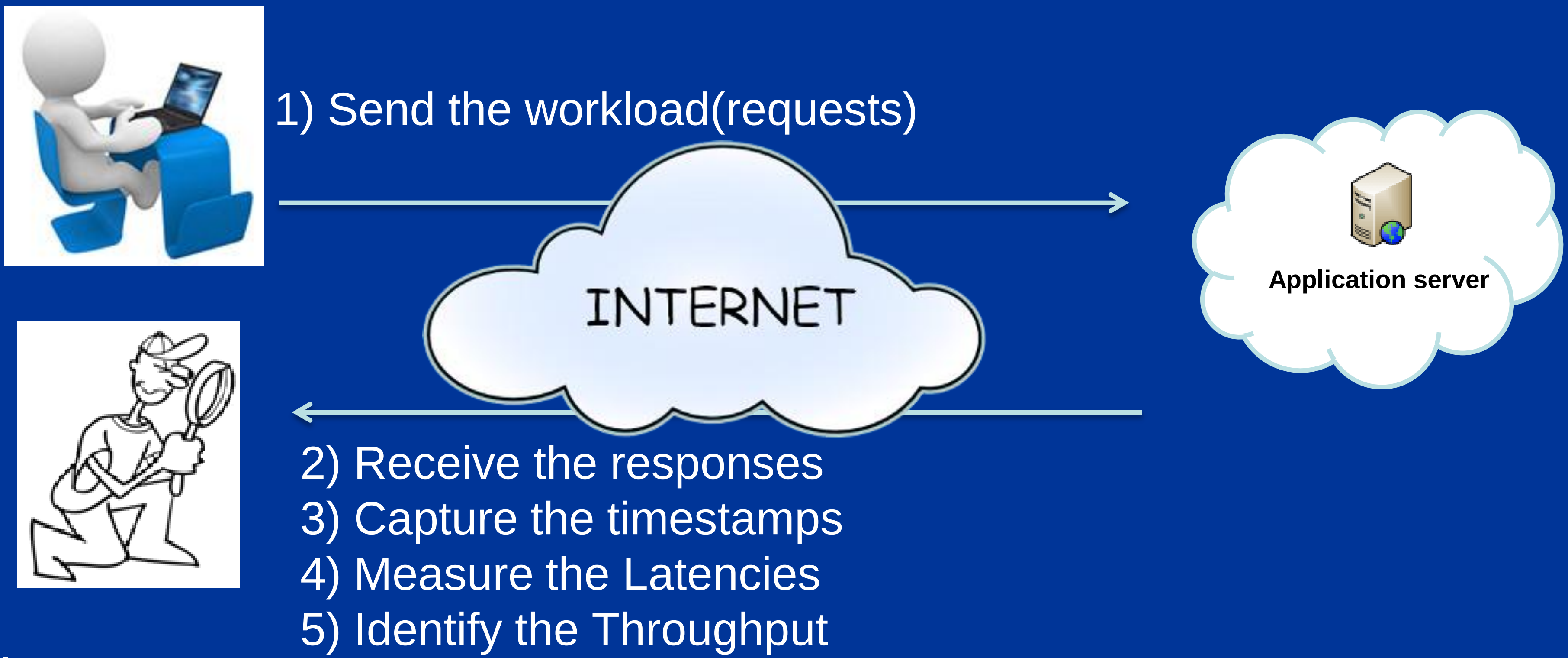
Model= r3.4xlarge
vCPU=16
Mem (GiB)=122
SSD Storage (GB)=1*320

Model= r3.2xlarge
vCPU=8
Mem (GiB)=61
SSD Storage (GB)=1*160

Model= c3.2xlarge
vCPU=8
Mem (GiB)=15
SSD Storage (GB)=2*80

Model= c3.xlarge
vCPU=4
Mem (GiB)=7.5
SSD Storage (GB)=2*40

TOOL



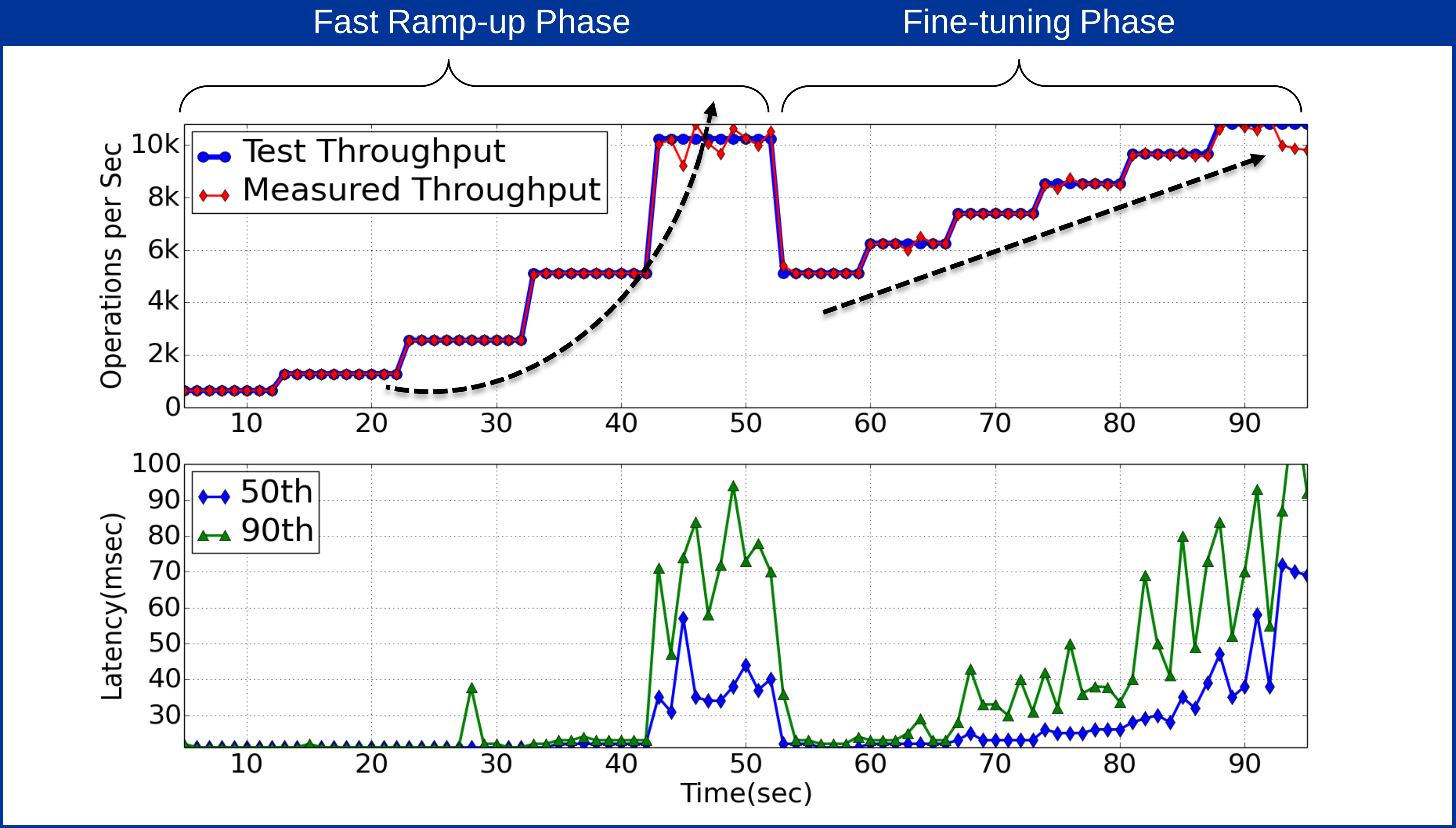
1) Send the workload(requests)

2) Receive the responses
3) Capture the timestamps
4) Measure the Latencies
5) Identify the Throughput

Goals:

- ❑ Sample the application responsiveness vs. various workloads
- ❑ Helps to pick VMs with better performance

METHODOLOGY



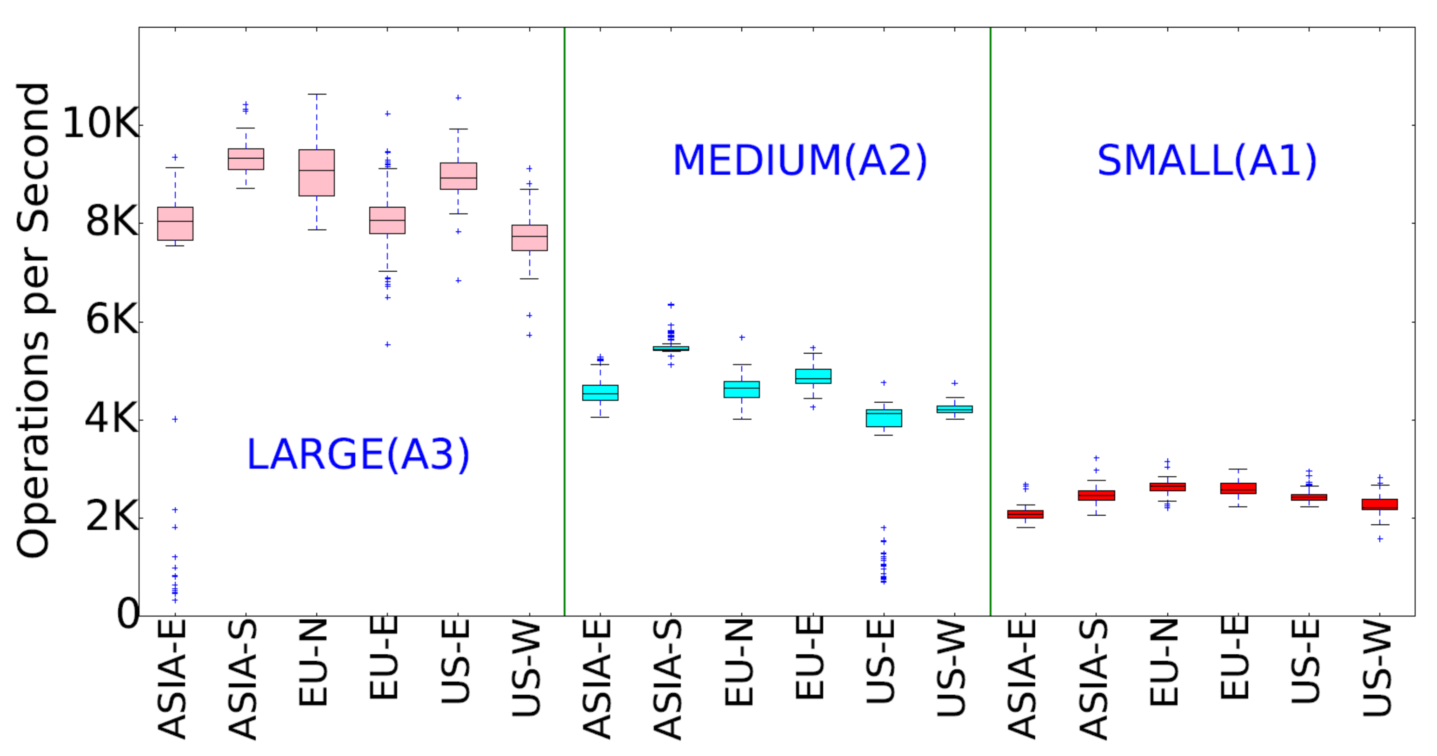
- ❑ RTT is a hint to detect the server side latency status

MEASUREMENTS SETUP

❑ Implementing the methodology as a Plugin for Apache-Jmeter

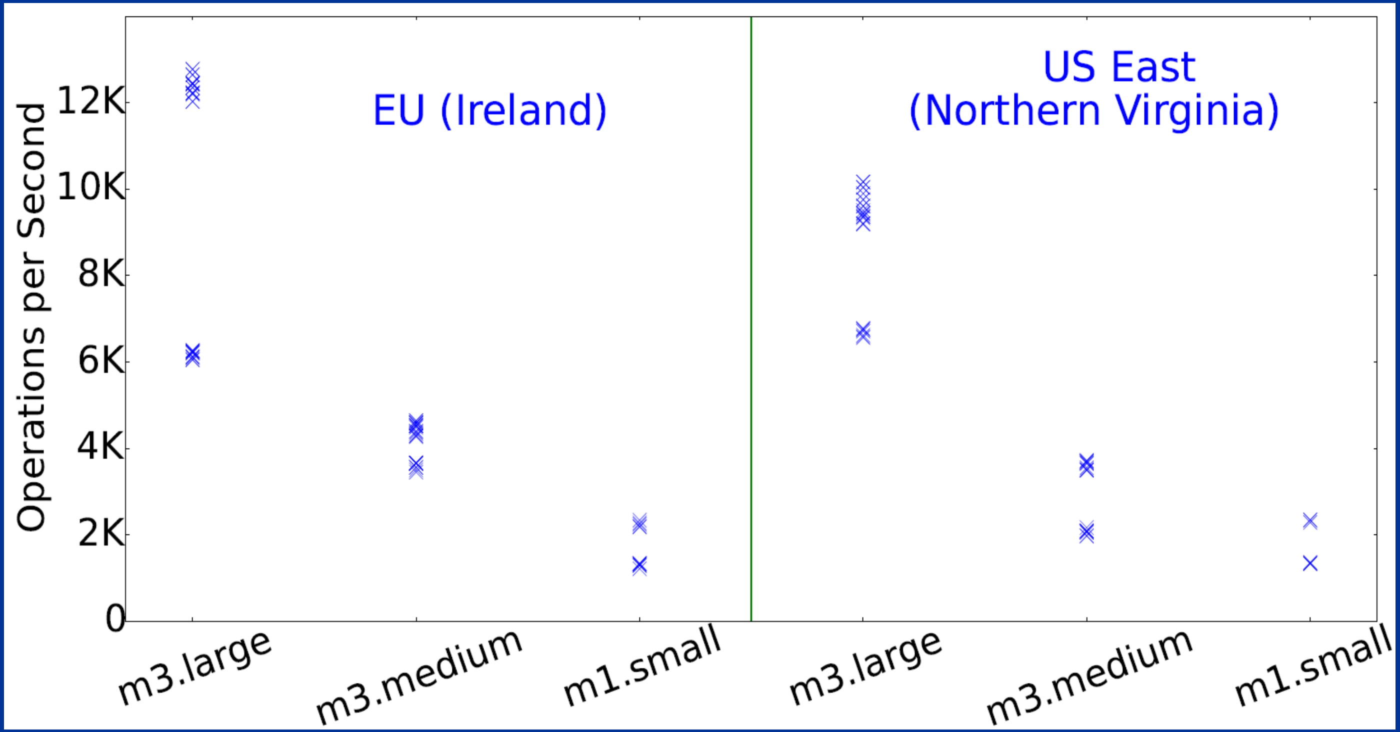
❑ Using a real Cloud application for our benchmarking (Apache Cassandra)

1-BENCHMARKING (MICROSOFT AZURE)



- ❑ 6 Availability Zones (Data Centers)
- ❑ 3 types of Instances
- ❑ Observed more variation in larger instances

1- BENCHMARKING (AMAZON EC2)

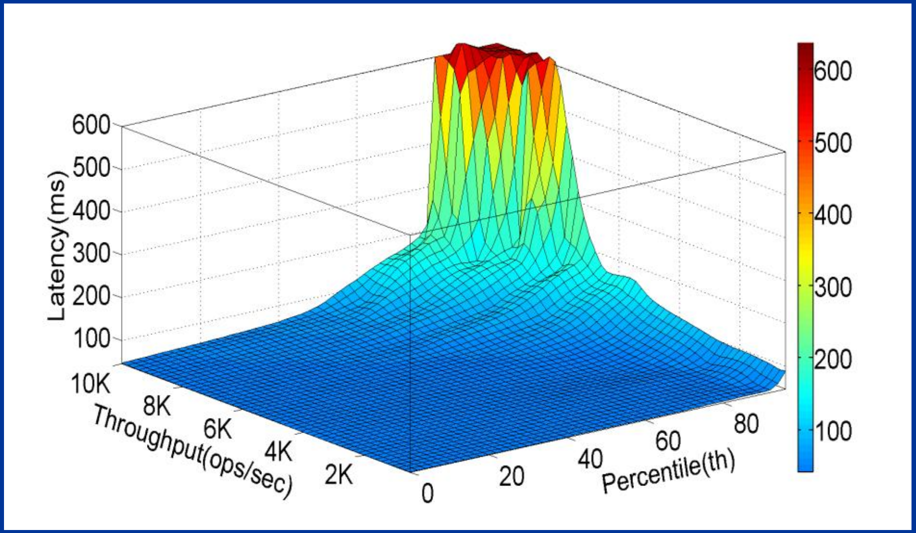


- ❑ Two separate performance bands(Same behavior seen in Google Compute Engine platform)
- ❑ Most likely because of the Hardware heterogeneity[1]

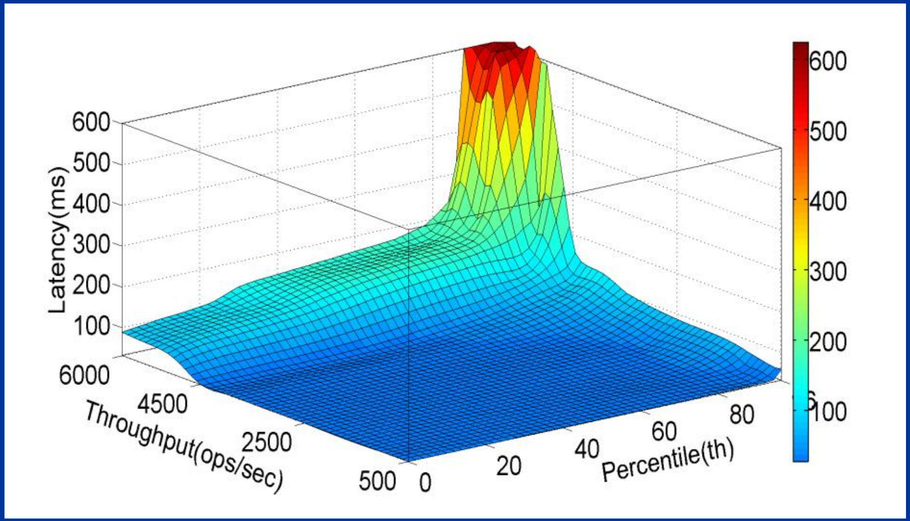
[1] Ou, Zhonghong, et al. "Exploiting hardware heterogeneity within the same instance type of Amazon EC2." *4th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)*. 2012.

2- IDENTIFY LATENCY/THROUGHPUT TRADE-OFF

Large instance(A3)



Medium instance(A1)



❑ Helps application providers in their deployment and provisioning decisions

❑ Future work direction

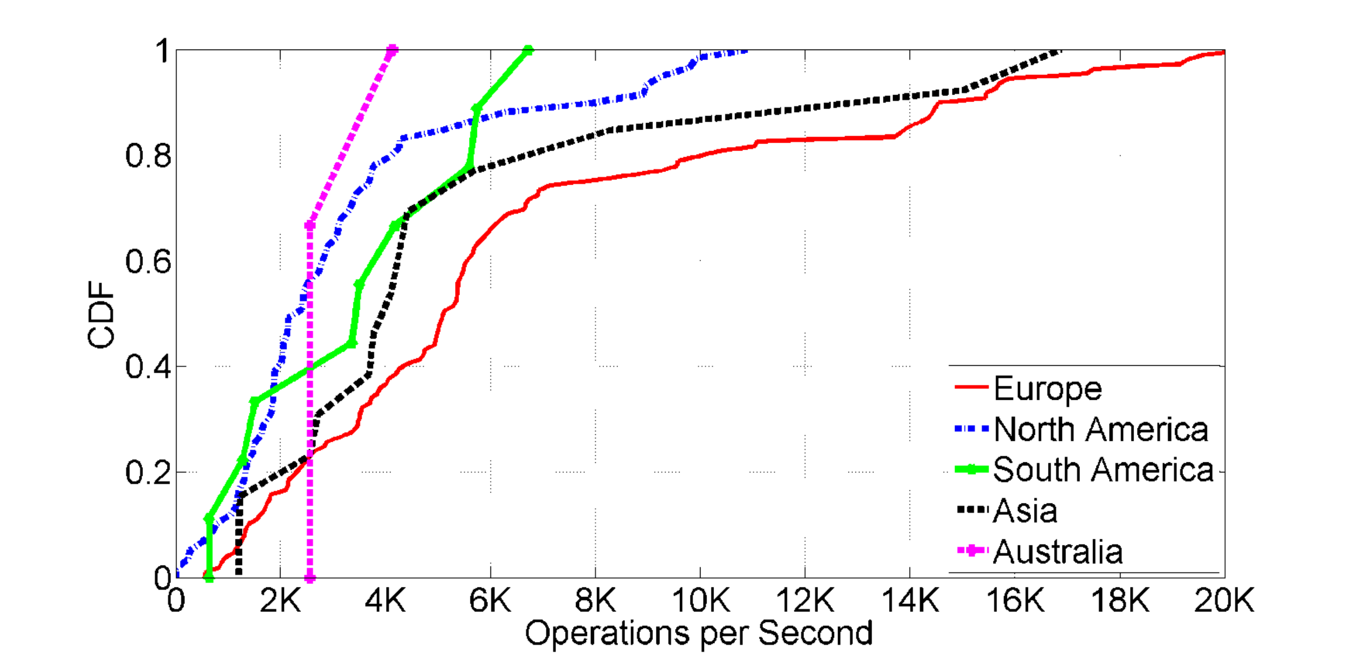
CONCLUSION

- Statistical properties of latency to detect the fine-grained behavior of an application
- A black-box methodology that estimates the workload a Cloud application can sustain
- Benchmarking a cloud application in various cloud platform
- Identify a trade-off between the throughput and latency of application servers, which can help application providers in their deployment and provisioning decisions

Thanks!

BENCHMARKING (PLANETLAB)

- 193 nodes
- 109, 59, 9, 13, 3 nodes in Europe, North and South America, Asia and Australia
- Some of the nodes have performance equivalent to commercial platforms



Queen Mary
University of London