The Performance and Cost Variability of Amazon EC2

Gary A. McGilvary

THE UNIVERSITY of EDINBURGH
informatics

edinburgh
data-intensive
research

OUTLINE

1. Introduction

Motivation
Amazon EC2 and SPRINT

2. Cost Variations

End-User Location
Data Transfer Usage Charges

3. Performance Variations

Cloud Load
Underutilization
Instance Processors

4. Conclusions/Future Work

INTRODUCTION

- Investigate the variability of cost and performance

- Optimal cloud configuration

- Cloud platform selection

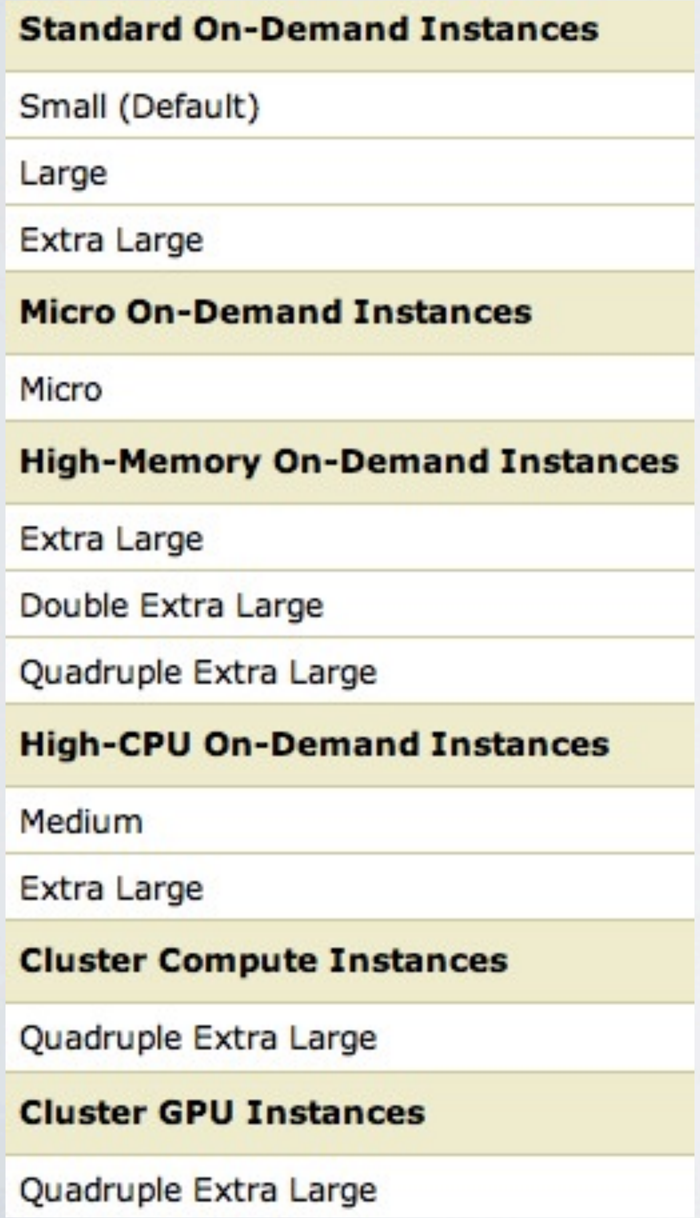
- Cloud vs HPC

- Business impact

AMAZON EC2

- Amazon's Elastic Compute Cloud

Compute



Standard On-Demand Instances
Small (Default)
Large
Extra Large
Micro On-Demand Instances
Micro
High-Memory On-Demand Instances
Extra Large
Double Extra Large
Quadruple Extra Large
High-CPU On-Demand Instances
Medium
Extra Large
Cluster Compute Instances
Quadruple Extra Large
Cluster GPU Instances
Quadruple Extra Large

Size	Memory	Storage	Compute	Cores	I/O	Cost $
Small	1.7 GB	160 GB	1 CU	1	Moderate	0.085
Large	7.5 GB	850 GB	4 CU's	2	High	0.34
XLarge	15 GB	1690 GB	8 CU's	4	High	0.68

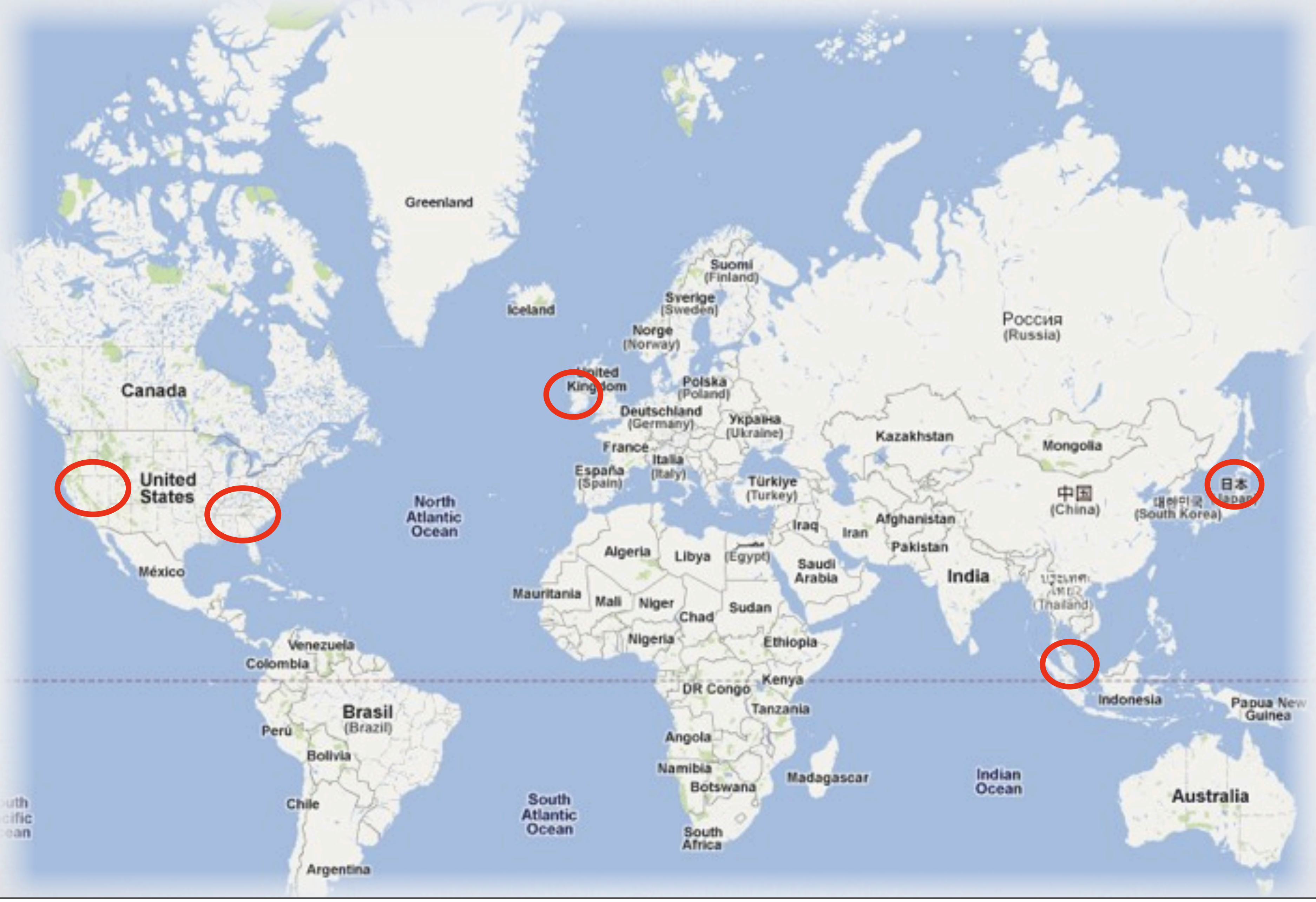
EC2 Compute Unit

- 1 EC2 CU: 1.0 - 1.2 GHz Xeon 2007 processor

AMAZON EC2

- Amazon’s Elastic Compute Cloud

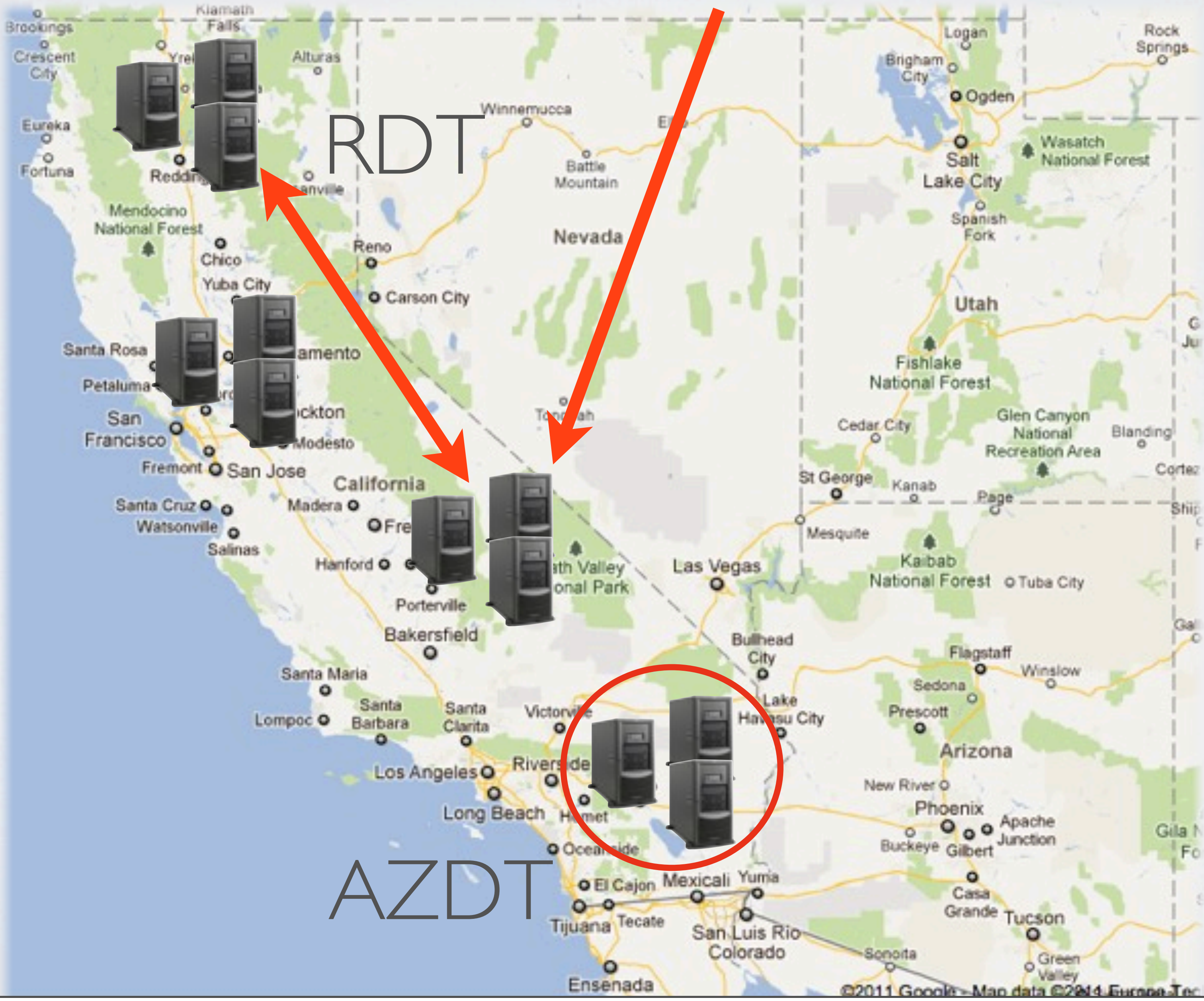
Network - Regions



AMAZON EC2

- Amazon's Elastic Compute Cloud

Network - Availability Zones



IDT

SPRINT

- Simple Parallel R INTerface
  - provides parallel functions of R

HPC	Multi-core desktops	Servers	Shared Memory Machines	Network of Workstations	GPU	Cloud	supercomputers
SPRINT Compatibility	✓	✓	✓	✓		✓	✓

- Functions:
  - *pcor:* parallel correlation (memory/compute-intensive)

    pcor(t (x, y = x)

  - *pmaxT:* parallel permutation test (compute-intensive)

    pmaxT(x, classlabel, B=150000)

OUTLINE

1. Introduction
2. Cost Variations
3. Performance Variations
4. Conclusions/Future Work

Motivation
Amazon EC2 and SPRINT

End-User Location
Data Transfer Usage Charges

COST VARIABILITY

Cost vs User Location:

- Instance Location:
  - US East Region
  - us-east-1b

Greenland
Norwegian Sea
Sverige
Suomi Finland
Norge
Россия Russia
North Sea
United Kingdom
Danmark
Polska
Deutschland
France
España
Portugal
Italia
Kazakhstan
Mongolia
Uzbekistan
Turkmenistan
Türkiye
Ελλάς
Mediterranean Sea
Afghanistan
Pakistan
Iraq
Iran
中国 China
India
Saudi Arabia
Libya
Algeria
Morocco
Mauritania
Mali
Niger
Chad
Sudan
Nigeria
Yemen
Ethiopia
Arabian Sea
North Atlantic Ocean
Labrador Sea
Caribbean Sea
Venezuela
Colombia
Thailand

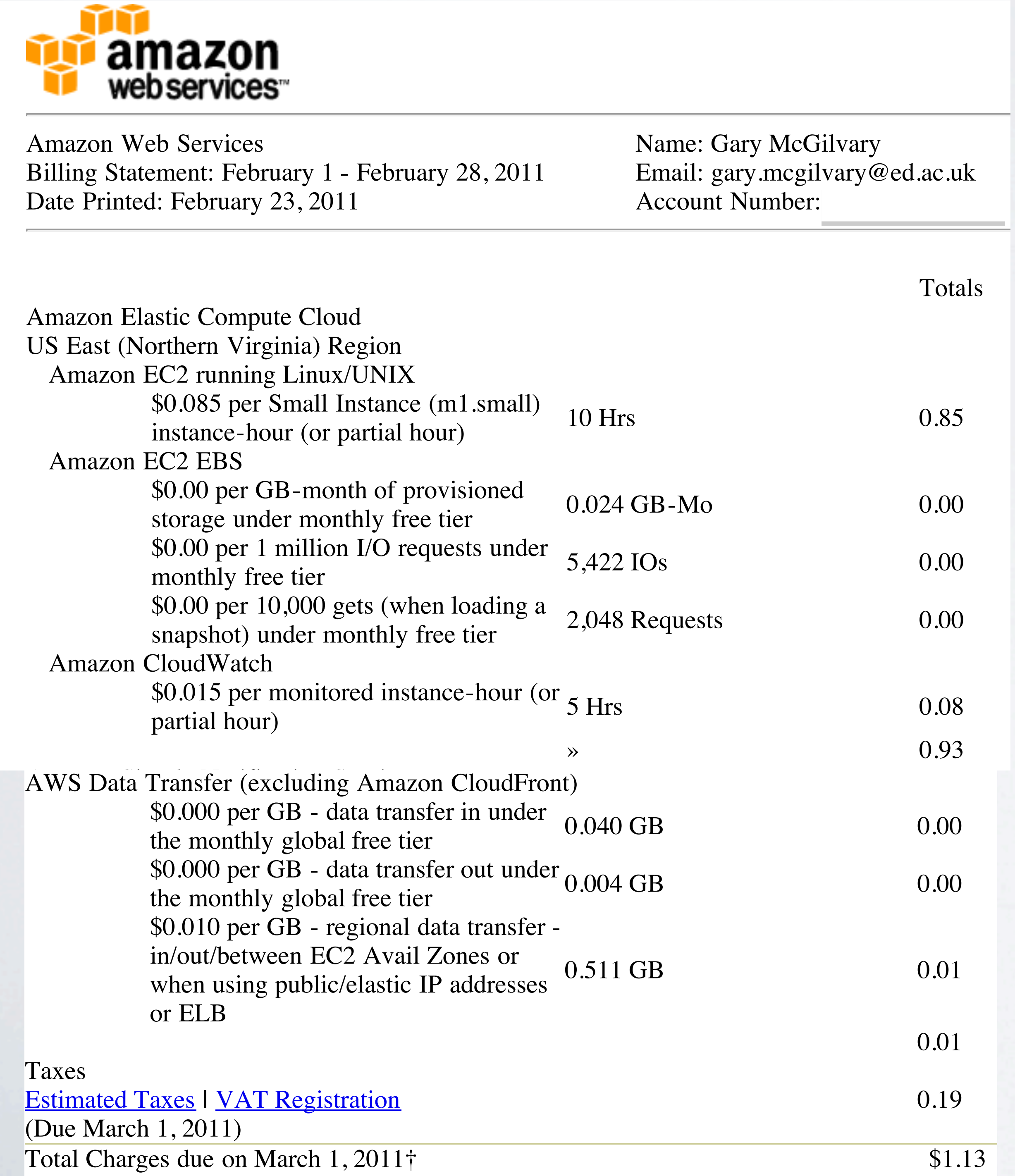
- Submit from: Thailand and UK

Experiment

- Copy of SPRINT pcor data
- SPRINT package installation from EC2 repository
- Execution
- Results: Invoice and Usage report

COST VARIABILITY

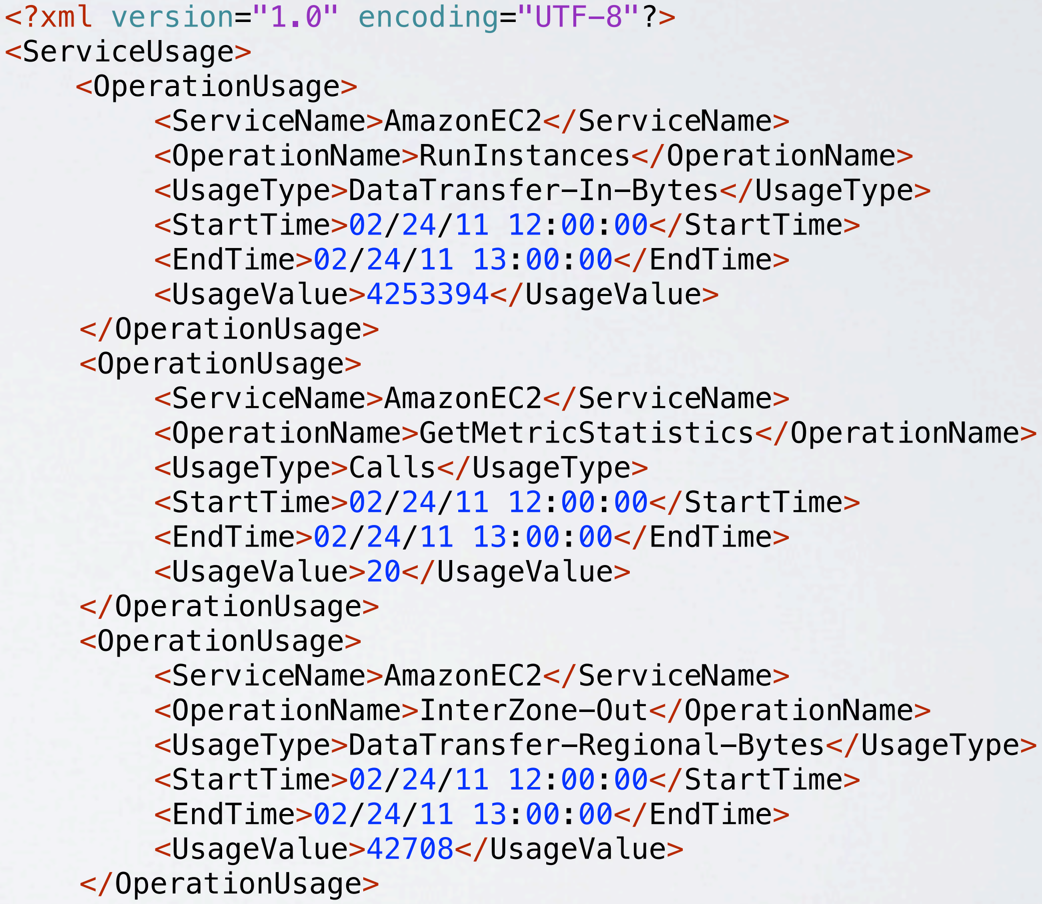
Cost vs User Location:



amazon webservices™

Amazon Web Services	Name: Gary McGilvary
Billing Statement: February 1 - February 28, 2011	Email: gary.mcgilvary@ed.ac.uk
Date Printed: February 23, 2011	Account Number:

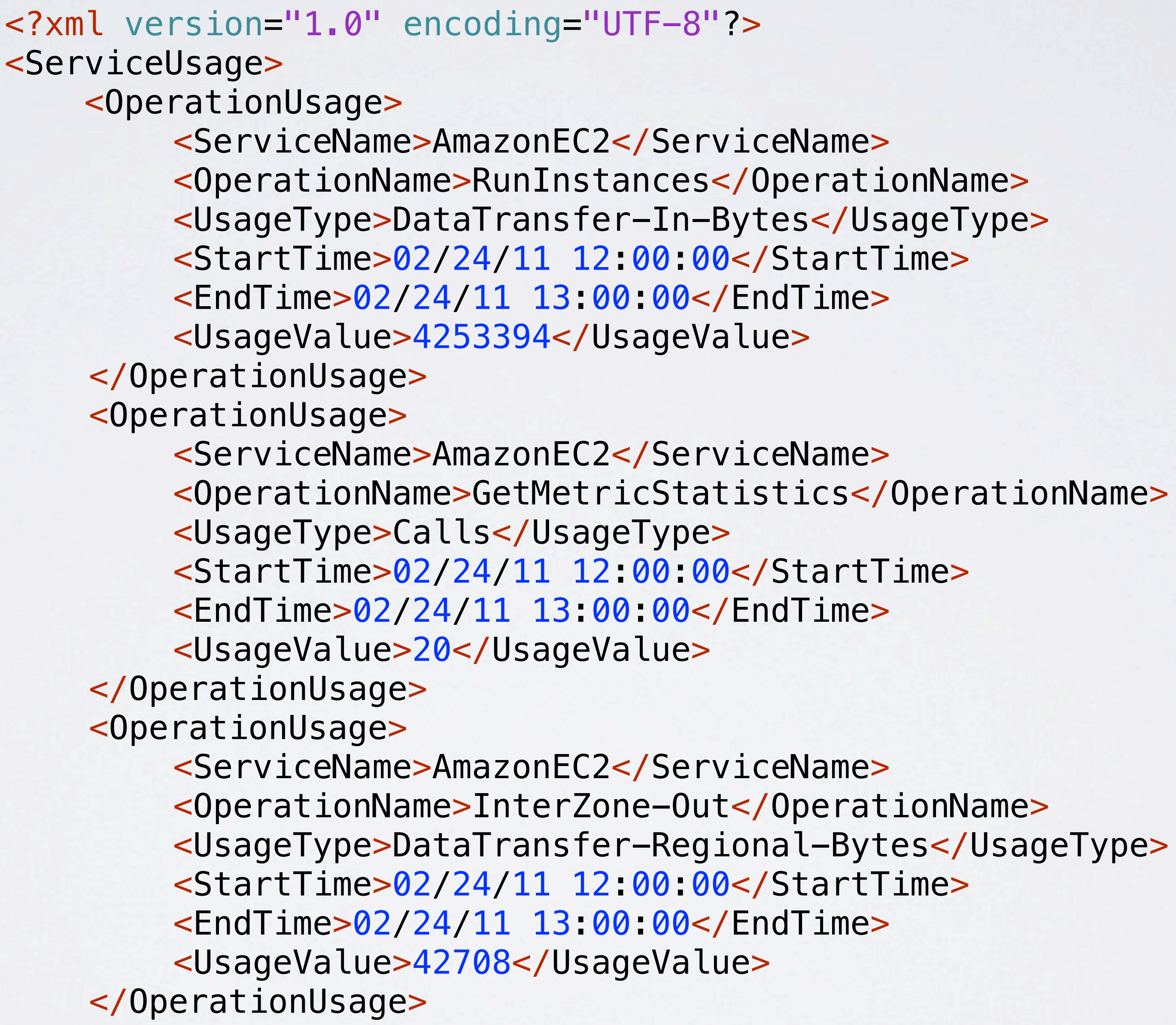
		Totals
Amazon Elastic Compute Cloud US East (Northern Virginia) Region		
Amazon EC2 running Linux/UNIX		
$0.085 per Small Instance (m1.small) instance-hour (or partial hour)	10 Hrs	0.85
Amazon EC2 EBS		
$0.00 per GB-month of provisioned storage under monthly free tier	0.024 GB-Mo	0.00
$0.00 per 1 million I/O requests under monthly free tier	5,422 IOs	0.00
$0.00 per 10,000 gets (when loading a snapshot) under monthly free tier	2,048 Requests	0.00
Amazon CloudWatch		
$0.015 per monitored instance-hour (or partial hour)	5 Hrs	0.08
	»	0.93
AWS Data Transfer (excluding Amazon CloudFront)		
$0.000 per GB - data transfer in under the monthly global free tier	0.040 GB	0.00
$0.000 per GB - data transfer out under the monthly global free tier	0.004 GB	0.00
$0.010 per GB - regional data transfer - in/out/between EC2 Avail Zones or when using public/elastic IP addresses or ELB	0.511 GB	0.01
		0.01
Taxes Estimated Taxes | VAT Registration (Due March 1, 2011)		0.19
Total Charges due on March 1, 2011†		$1.13



```xml
<?xml version="1.0" encoding="UTF-8"?>
<ServiceUsage>
    <OperationUsage>
        <ServiceName>AmazonEC2</ServiceName>
        <OperationName>RunInstances</OperationName>
        <UsageType>DataTransfer-In-Bytes</UsageType>
        <StartTime>02/24/11 12:00:00</StartTime>
        <EndTime>02/24/11 13:00:00</EndTime>
        <UsageValue>4253394</UsageValue>
    </OperationUsage>
    <OperationUsage>
        <ServiceName>AmazonEC2</ServiceName>
        <OperationName>GetMetricStatistics</OperationName>
        <UsageType>Calls</UsageType>
        <StartTime>02/24/11 12:00:00</StartTime>
        <EndTime>02/24/11 13:00:00</EndTime>
        <UsageValue>20</UsageValue>
    </OperationUsage>
    <OperationUsage>
        <ServiceName>AmazonEC2</ServiceName>
        <OperationName>InterZone-Out</OperationName>
        <UsageType>DataTransfer-Regional-Bytes</UsageType>
        <StartTime>02/24/11 12:00:00</StartTime>
        <EndTime>02/24/11 13:00:00</EndTime>
        <UsageValue>42708</UsageValue>
    </OperationUsage>
```

COST VARIABILITY

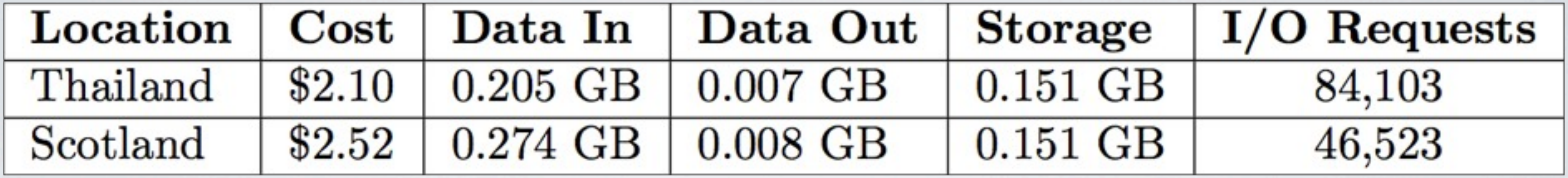
Cost vs User Location:



```xml
<?xml version="1.0" encoding="UTF-8"?>
<ServiceUsage>
    <OperationUsage>
        <ServiceName>AmazonEC2</ServiceName>
        <OperationName>RunInstances</OperationName>
        <UsageType>DataTransfer-In-Bytes</UsageType>
        <StartTime>02/24/11 12:00:00</StartTime>
        <EndTime>02/24/11 13:00:00</EndTime>
        <UsageValue>4253394</UsageValue>
    </OperationUsage>
    <OperationUsage>
        <ServiceName>AmazonEC2</ServiceName>
        <OperationName>GetMetricStatistics</OperationName>
        <UsageType>Calls</UsageType>
        <StartTime>02/24/11 12:00:00</StartTime>
        <EndTime>02/24/11 13:00:00</EndTime>
        <UsageValue>20</UsageValue>
    </OperationUsage>
    <OperationUsage>
        <ServiceName>AmazonEC2</ServiceName>
        <OperationName>InterZone-Out</OperationName>
        <UsageType>DataTransfer-Regional-Bytes</UsageType>
        <StartTime>02/24/11 12:00:00</StartTime>
        <EndTime>02/24/11 13:00:00</EndTime>
        <UsageValue>42708</UsageValue>
    </OperationUsage>
```

COST VARIABILITY

Cost vs User Location:



Location	Cost	Data In	Data Out	Storage	I/O Requests
Thailand	$2.10	0.205 GB	0.007 GB	0.151 GB	84,103
Scotland	$2.52	0.274 GB	0.008 GB	0.151 GB	46,523

- Difference in taxation levels
  - Scale with use and expensive for prolonged time periods

- Difference in resource usages
  - ~~dependent on location~~ or cloud load?

- Consequences of user location:
  - Businesses/Individuals in a tax free zone will benefit

  - reduction in performance?

COST VARIABILITY

Variability of Data Transfer Usage Charges

- Run SPRINT's pcor function multiple times

- Small Ubuntu instance

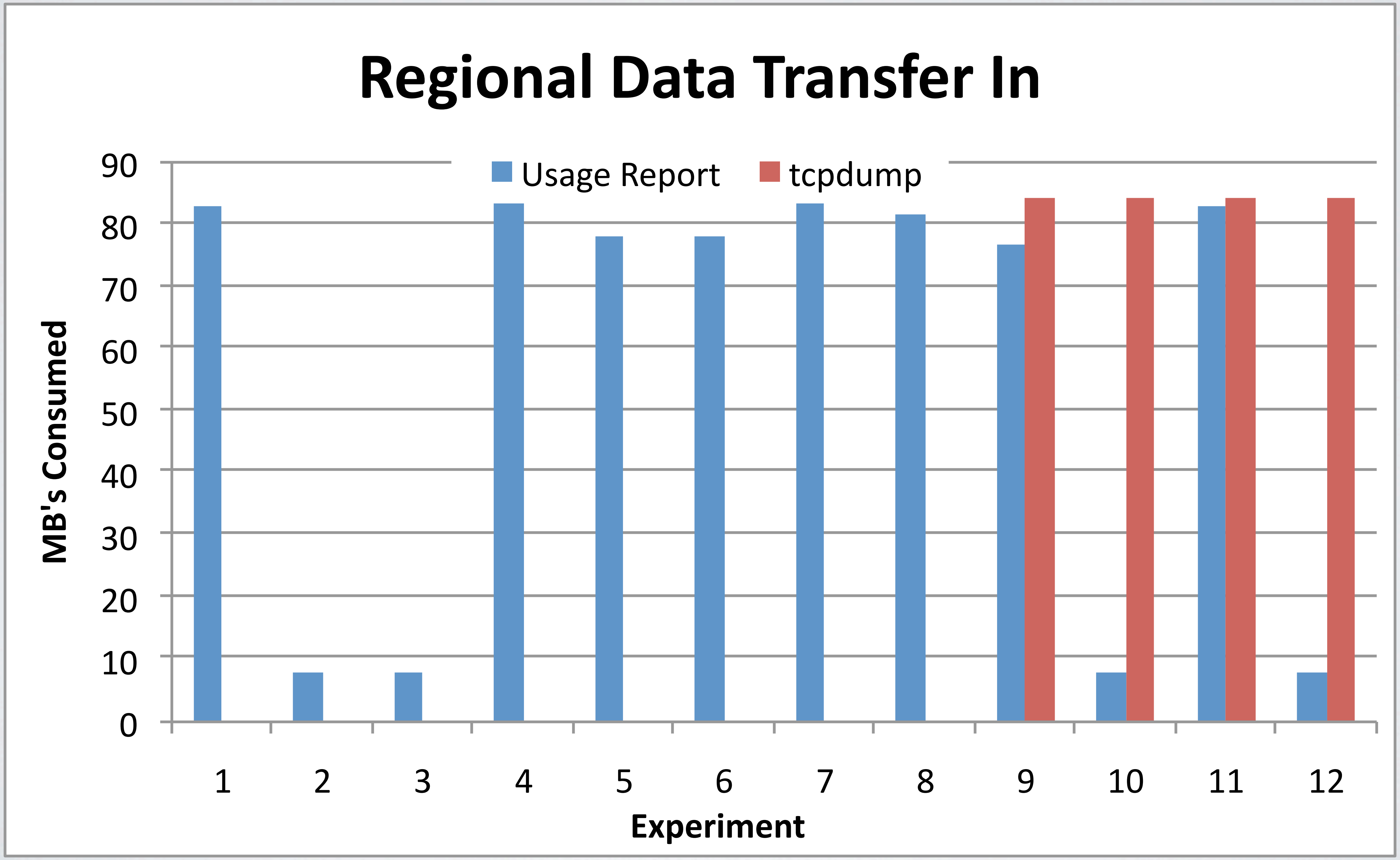
- Regional Data Transfer (RDT)

- Transferred 84.3 MB's from EC2 Ubuntu Repository to the instance

Nevada
Utah
California
Arizona
Reno
Carson City
Sacramento
San Francisco
San Jose
Stockton
Modesto
Fresno
Bakersfield
Las Vegas
Los Angeles
Long Beach
Riverside
Phoenix
Tucson
Salt Lake City
Ogden
Flagstaff
Redding
Santa Rosa
Santa Barbara
Tijuana
Mexicali
Ensenada
©2011 Google - Map data ©2011 Europa Tec

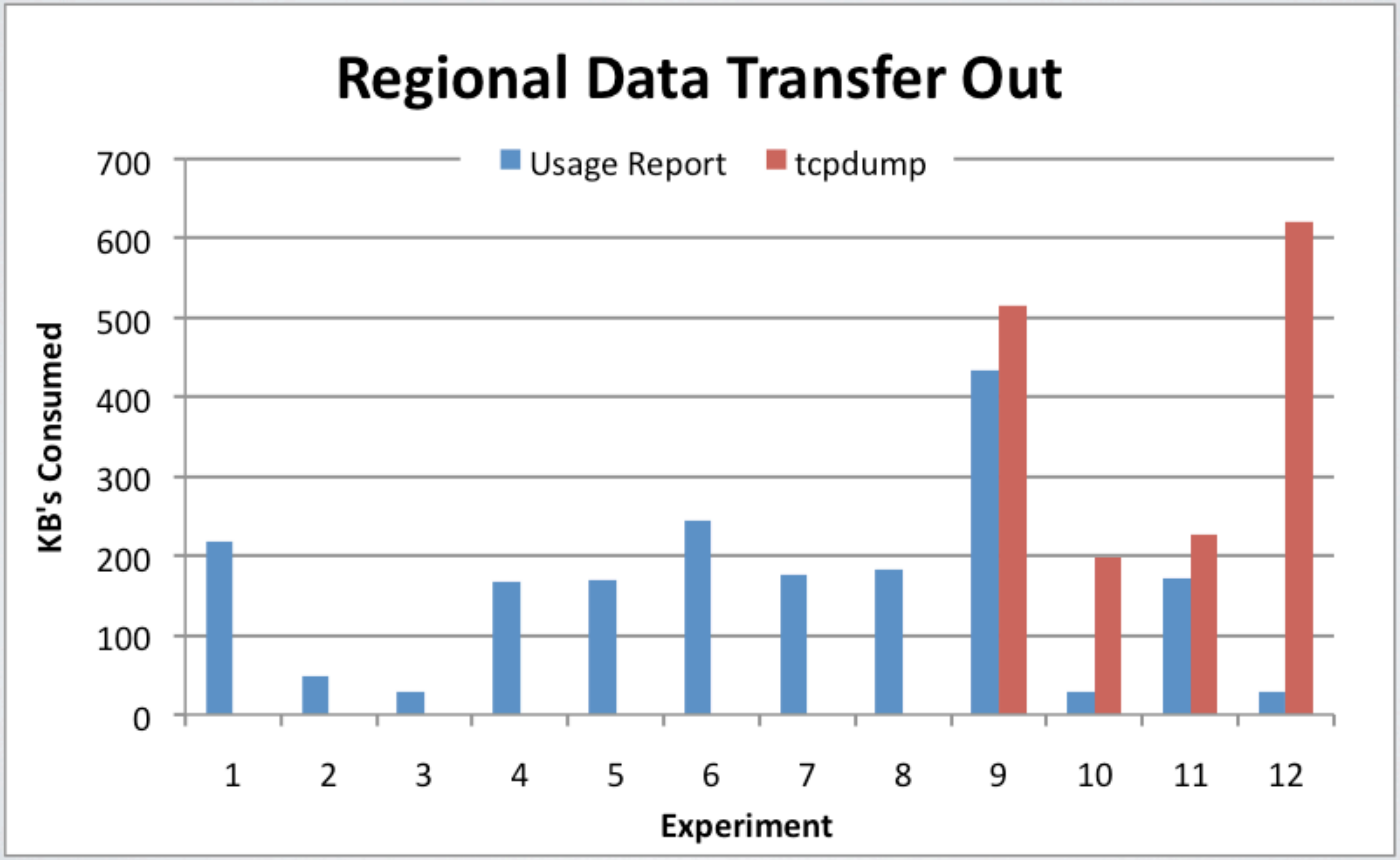
COST VARIABILITY

Variability of Data Transfer Usage Charges



COST VARIABILITY

Variability of Data Transfer Usage Charges



COST VARIABILITY

Variability of Data Transfer Usage Charges

- Consequences of incorrect data usage recording:
  - some free data transfer!
  - substantial savings for prolonged use!
  - GB's can go unrecorded on Azure

OUTLINE

1. Introduction

Motivation
Amazon EC2 and SPRINT

2. Cost Variations

End-User Location
Data Transfer Usage Charges

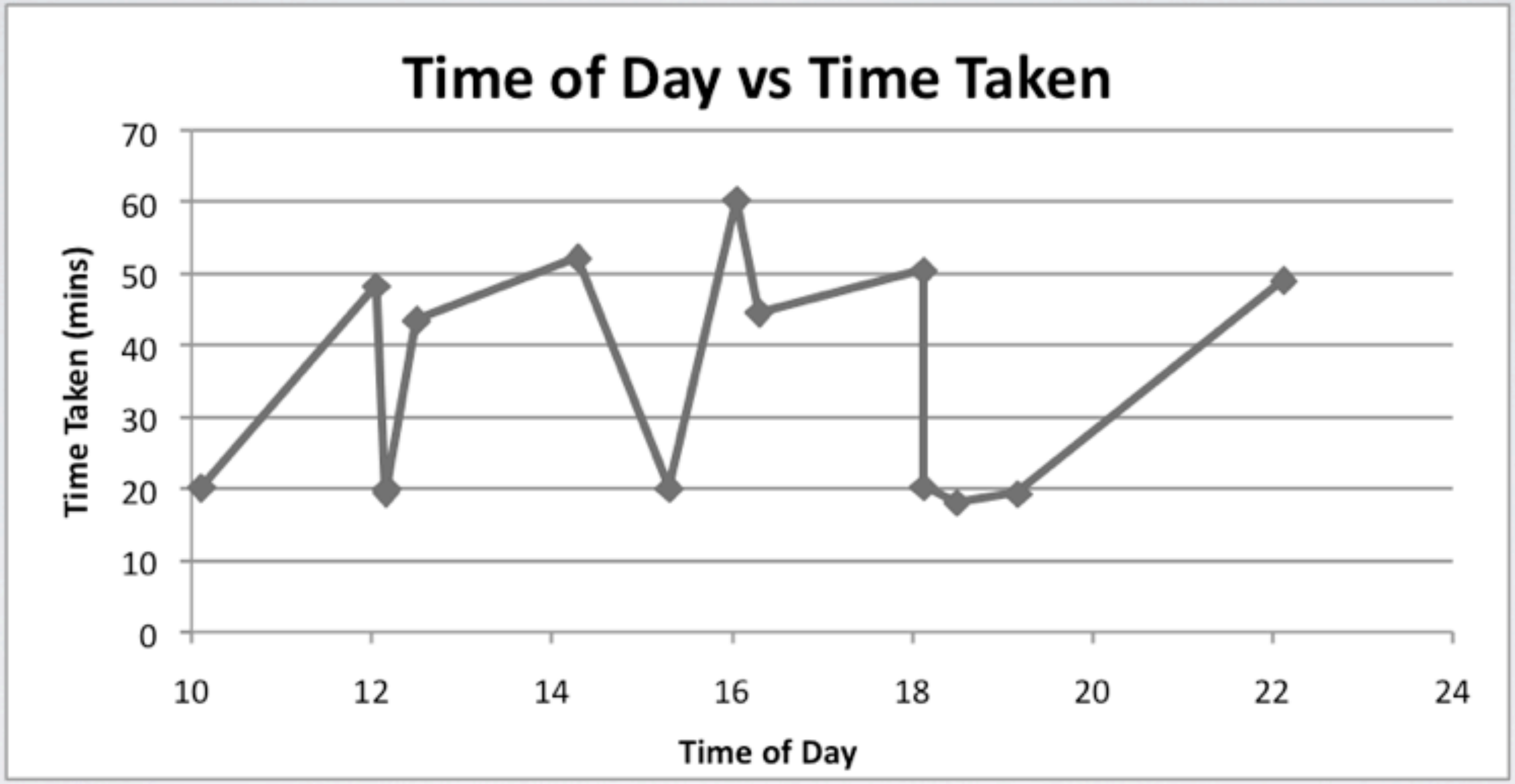
3. Performance Variations

Cloud Load
Underutilization
Instance Processors

4. Conclusions/Future Work

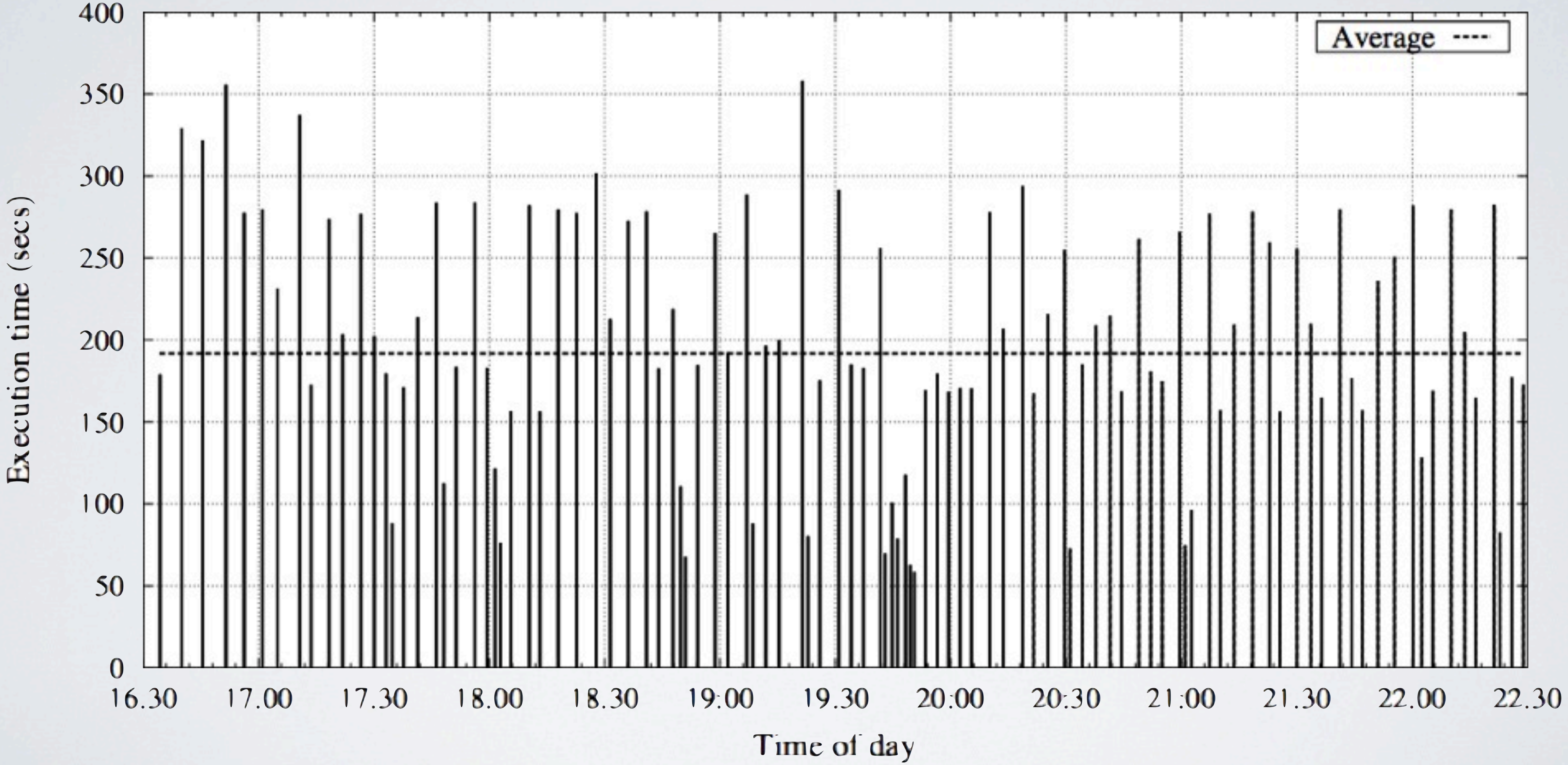
CLOUD LOAD

Cloud Load vs Time of Day?



CLOUD LOAD

Cloud Load vs Time of Day?



Improving High-Performance Computations on Clouds Through Resource Underutilization

UNDERUTILISATION

- Reserving more resources while using a small % of each
  - optimum configuration!

- SPRINT's *pcor* and *pmaxT* functions and EC2 Large instances

- Two Cases: (Large: 2 cores at 2 EC2 CU's)

1. Each SPRINT process per instance (1 core - 50%)
   - e.g *4 processes = 4 instances*

2. Each SPRINT process per instance core
   - e.g *4 processes = 2 instances (4 cores)*

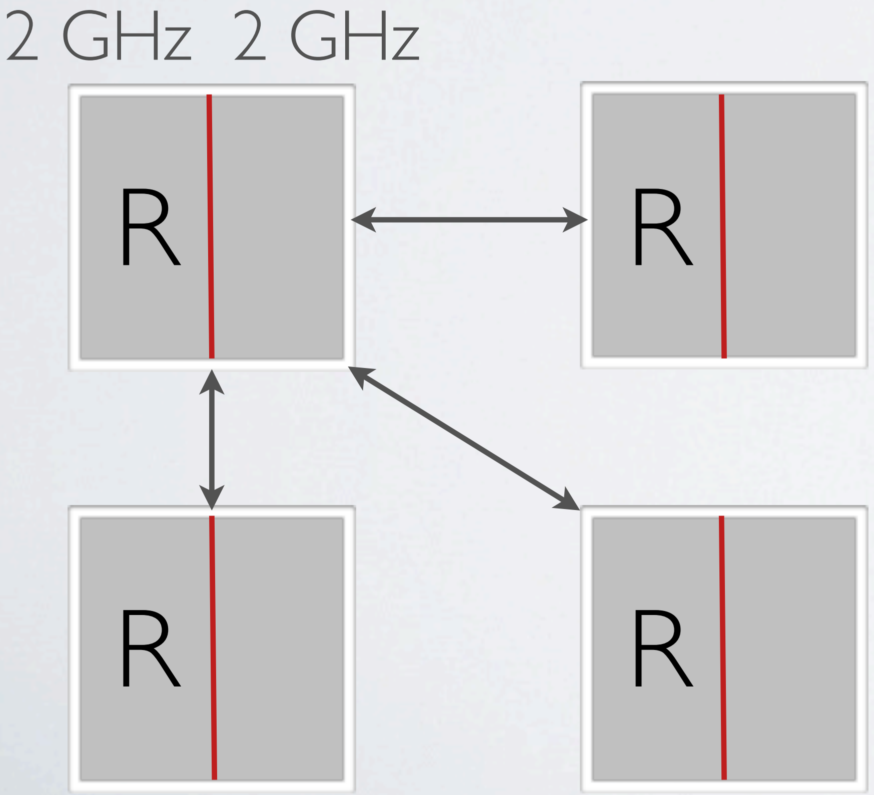
UNDERUTILISATION

- Two Cases: (Large: 2 cores at 2 EC2 CU's)

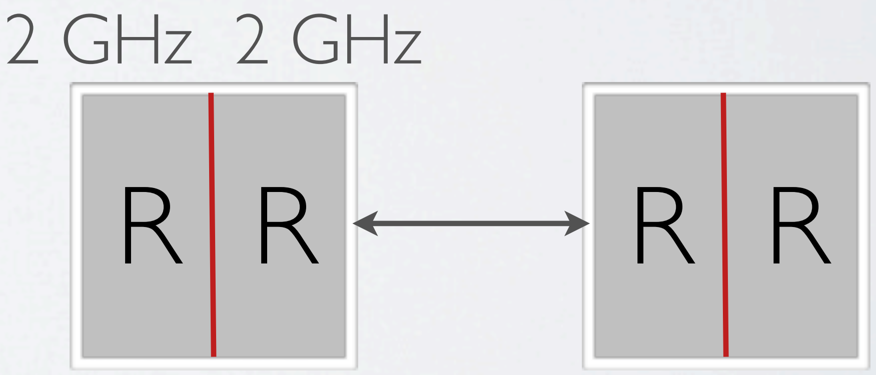
1. Each SPRINT process per instance (1 core - 50%)
   - e.g *4 processes = 4 instances*

2. Each SPRINT process per instance core

- e.g *4 processes = 2 instances (4 cores)*

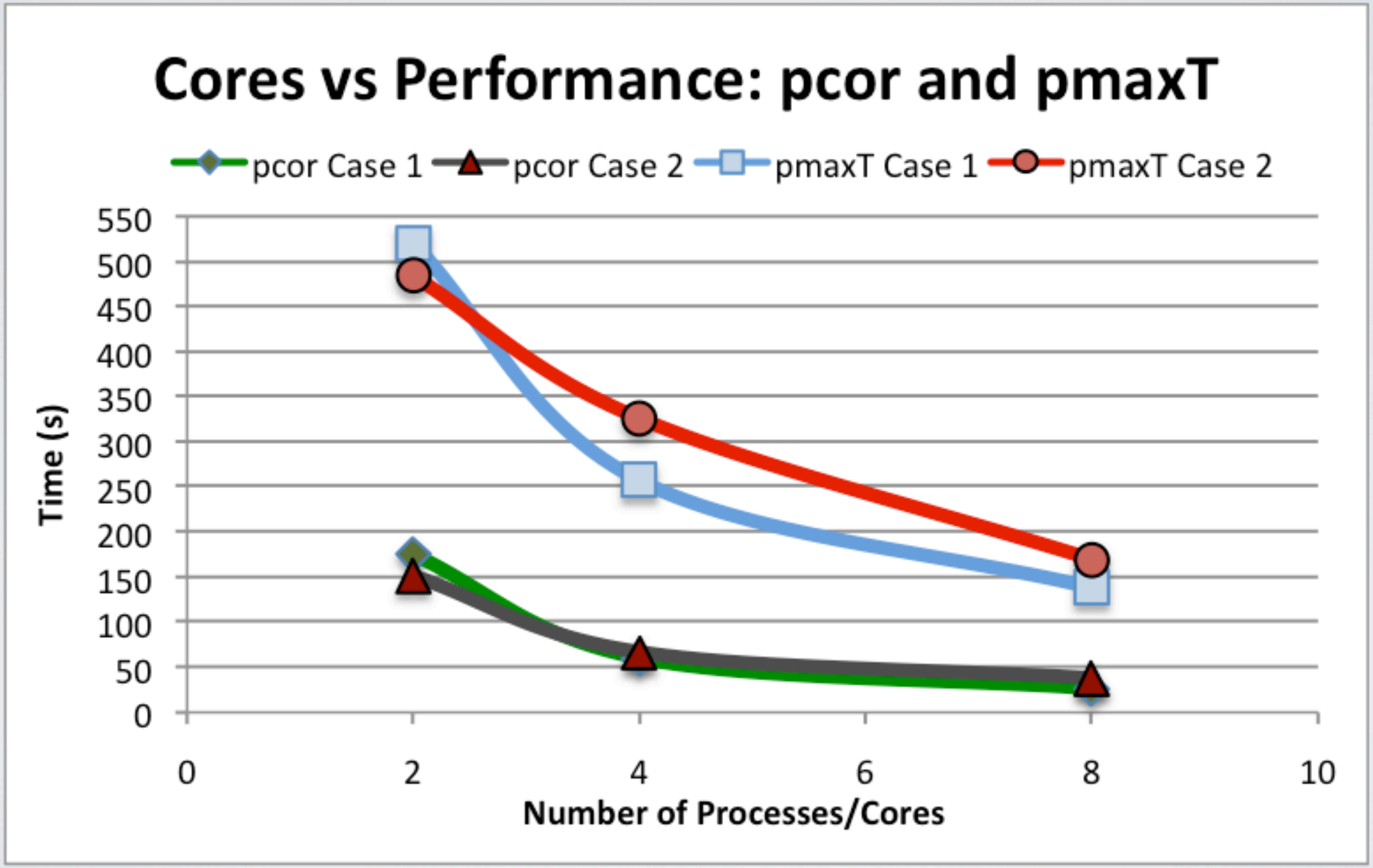


VS



UNDERUTILISATION

Case Study: SPRINT



UNDERUTILISATION

- Consequences of Underutilization:
  - Reserving more resources will increase costs
    - *cost vs performance?*
  - Paper: *Improving High-Performance Computations on Clouds Through Resource Underutilization*
    - "Underutilization improves the expected execution time by two orders of magnitude"
    - "... it is more than 3 times cheaper to use 50% of the resources than 100%"
  - Geared towards finding the optimal cloud configuration
    - Could potentially save businesses/individuals a substantial amount of time and money

How do we determine the utilisation rate, and hence optimal configuration for a job?

INSTANCE PROCESSORS

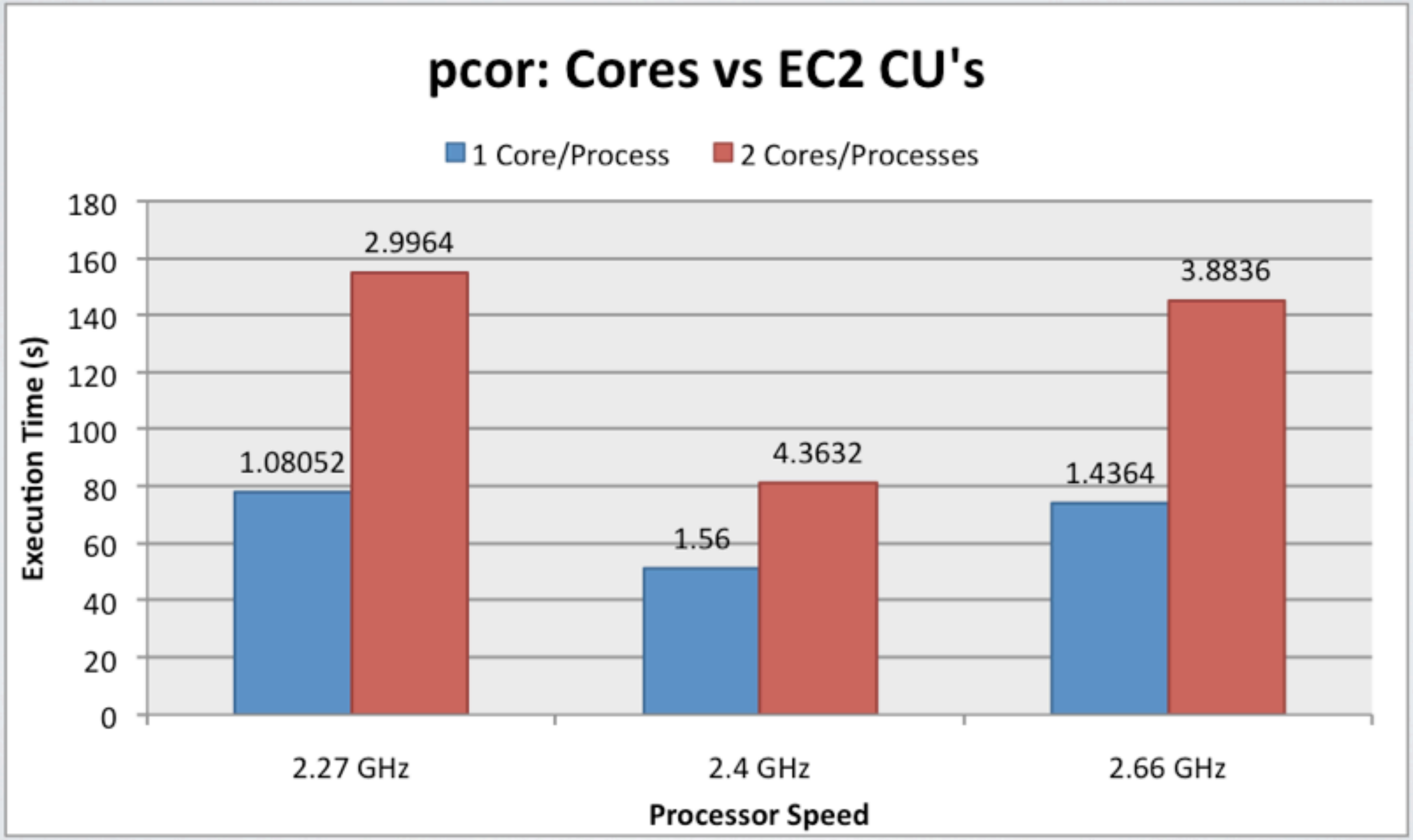
- Instances deployed on varying processors
- <u>Availability Zone</u>: us-east-1d   <u>Instance Type</u>: Large
- Remember: EC2 Compute Units (1.0 - 1.2 GHz, Xeon 2007)

  - *Large instance = 2 EC2 CU's per core (2cores), 4.0 -4.8 GHz*
  - *Large instance has 2 processors*

Processor Type	**Min Usage**	**Max Usage**
Intel Xeon E5507 2.27 GHz (x2 - 4.54 GHz)	88.1%	100%
Intel Xeon E5645 2.4 GHz (x2 - 4.8 GHz)	83.3%	100%
Intel Xeon E5430 2.66 GHz (x2 - 5.32 GHz)	75.1%	90.22%

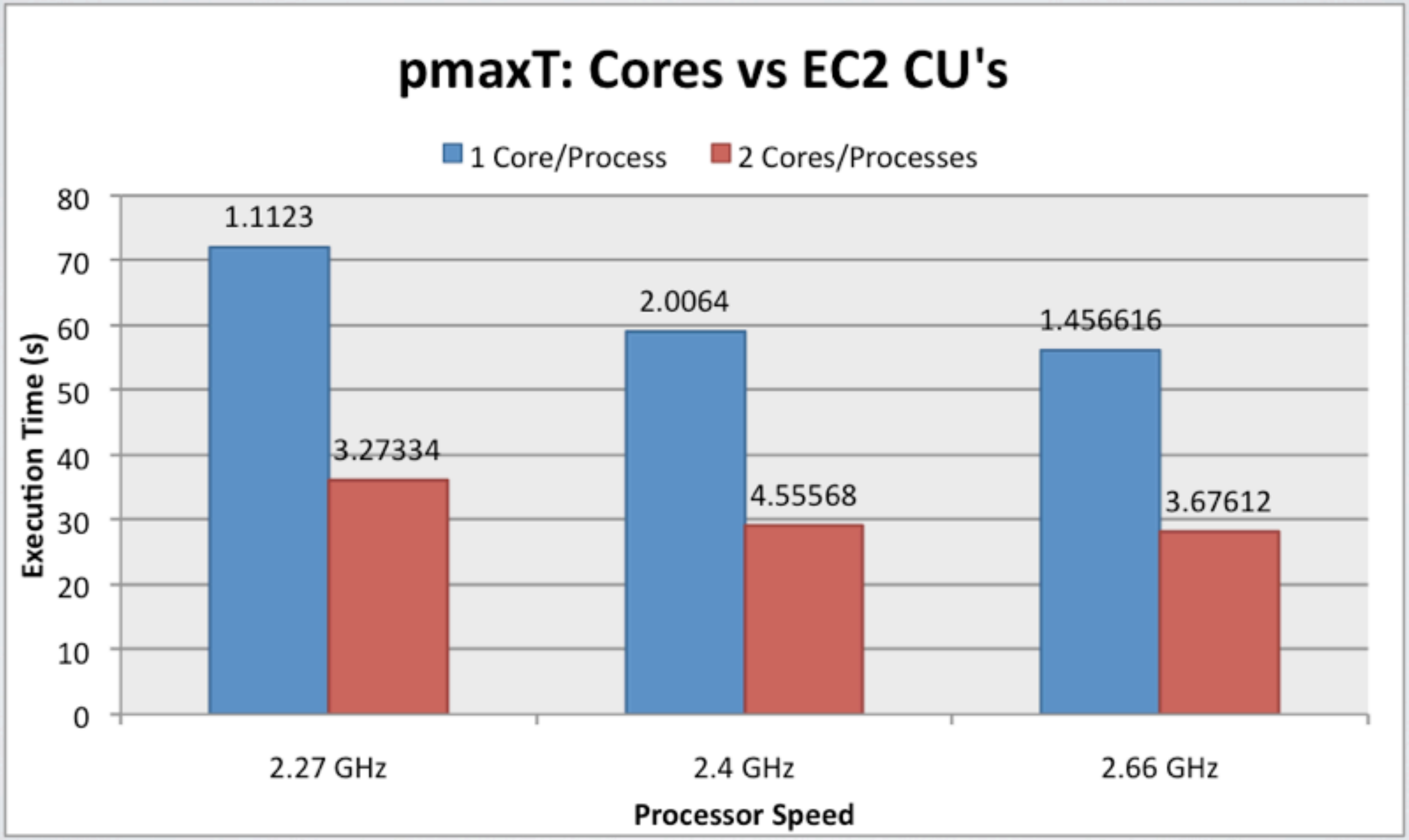
INSTANCE PROCESSORS

1 core: 2-GHz   2 cores: 4-GHz



INSTANCE PROCESSORS

1 core: 2-GHz   2 cores: 4-GHz



CONCLUSIONS

- Dependent on user location, costs may differ as well as resources used

- EC2's data transfer usage mechanism may be incorrect at times
  - cloud load?

- Application performance can vary significantly (execution times)

- Underutilization can increase performance
  - in every case?
  - utilization rate?

- The underlying instance processors affect performance
  - correct EC2 Compute Units specified?

Optimal cloud configuration == increased performance == lowest cost

THANK YOU!

Questions?

gary.mcgilvary@ed.ac.uk