CHALMERS UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

Estimating Cloud Application Performance Based on Micro-Benchmark Profiling

Joel Scheuner, Philipp Leitner

Joel Scheuner

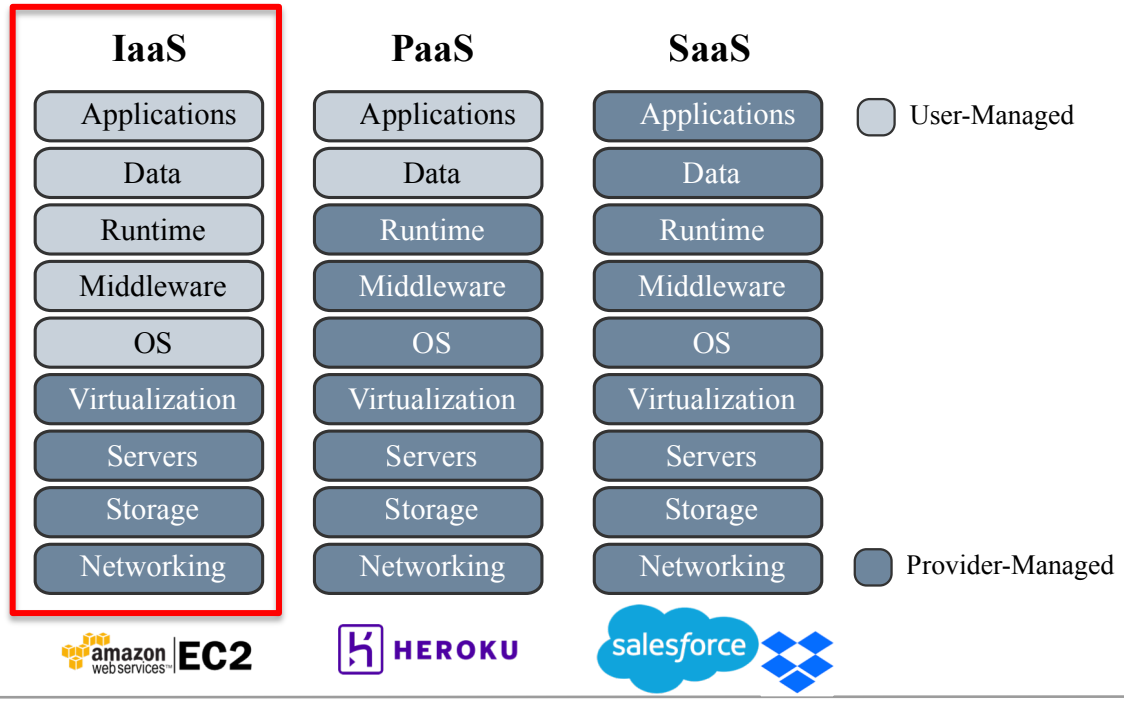
scheuner@chalmers.se

joe4dev

@joe4dev

Supported by WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Context: Public Infrastructure-as-a-Service Clouds



Infrastructure-as-a-Service (IaaS)
Platform-as-a-Service (PaaS)
Software-as-a-Service (SaaS)

Motivation: Capacity Planning in IaaS Clouds

What cloud **provider** should I choose?

Google Compute Engine

amazon web services EC2

Microsoft Azure

Joyent

GOGRID

rackspace

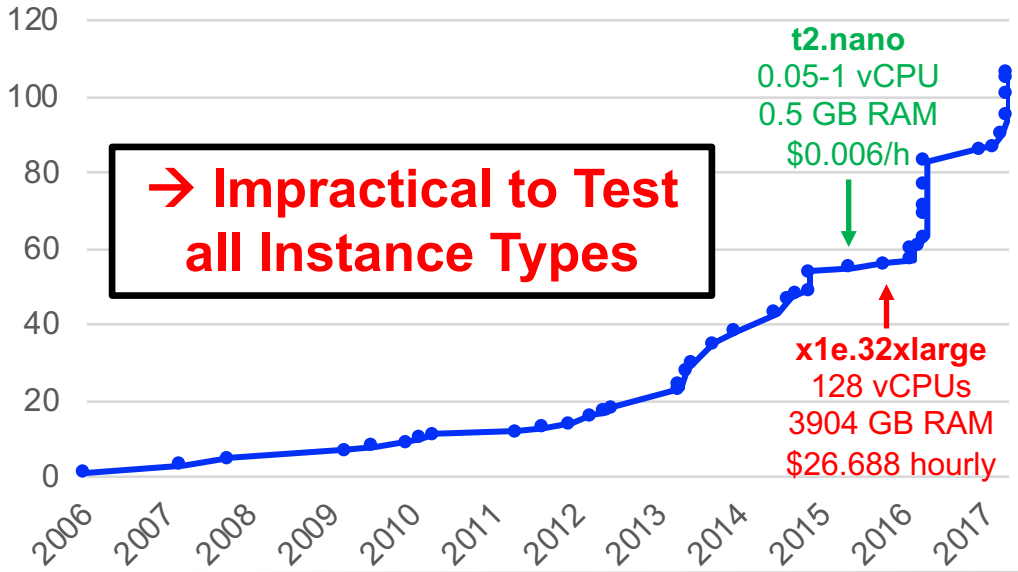
CloudSigma

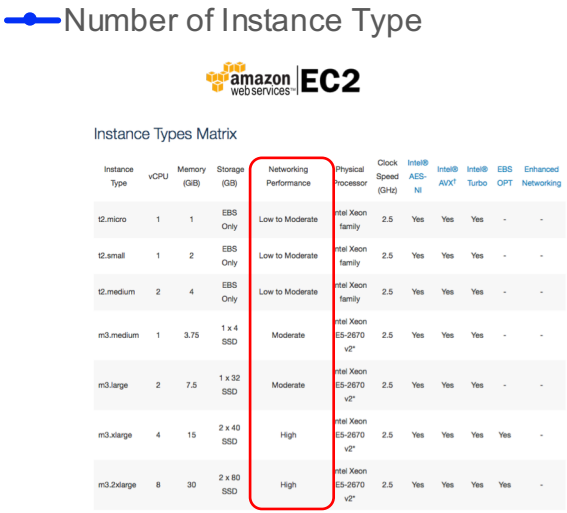
IBM Cloud

https://www.cloudorado.com

Motivation: Capacity Planning in IaaS Clouds

What cloud **service (i.e., instance type)** should I choose?





Topic: Performance Benchmarking in the Cloud

NETFLIX

"The instance type itself is a very major tunable parameter"

@brendangregg re:Invent'17

https://youtu.be/89fYOo1V2pA?t=5m4s

Background

Micro Benchmarks

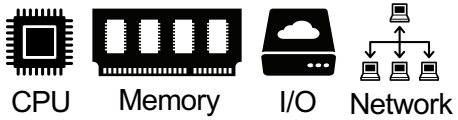
Application Benchmarks

Overall performance
(e.g., response time)

Generic	**Domain**	Specific
Artificial	**Workload**	Real-World
Resource-specific	**Resource Usage**	Resource-heterogeneous

Problem: Isolation, Reproducibility of Execution

Micro Benchmarks



Generic

Artificial

Resource-specific

Application Benchmarks

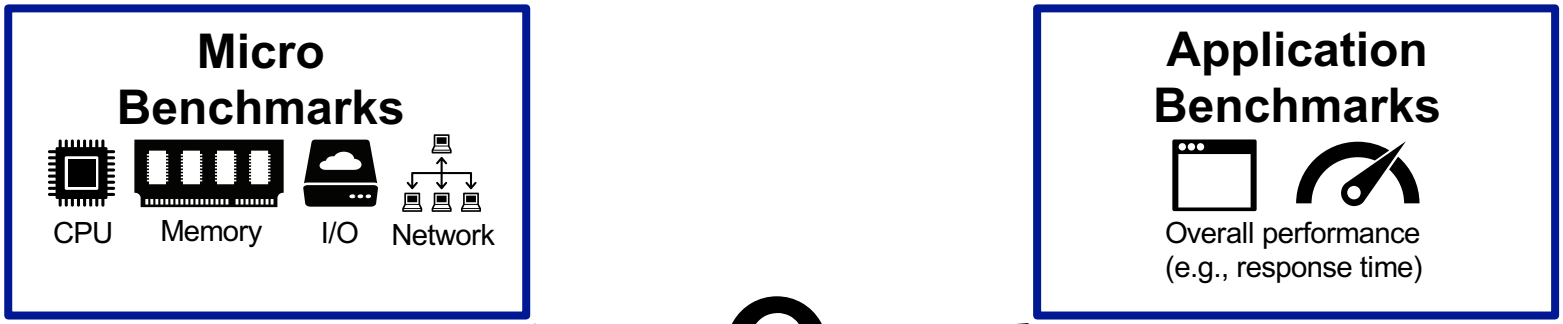
Overall performance
(e.g., response time)

Specific

Real-World

Resource-heterogeneous

Question


Micro Benchmarks
CPU
Memory
I/O
Network
Application Benchmarks
Overall performance
(e.g., response time)
Generic
Artificial
Resource-specific
How relevant?
Specific
Real-World
Resource-heterogeneous

Research Questions

PRE – Performance Variability

Does the performance of equally configured cloud instances vary relevantly?

RQ1 – Estimation Accuracy

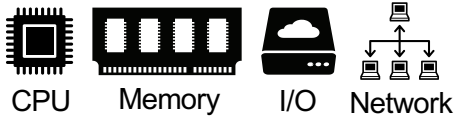
How accurate can a set of micro benchmarks estimate application performance?

RQ2 – Micro Benchmark Selection

Which subset of micro benchmarks estimates application performance most accurately?

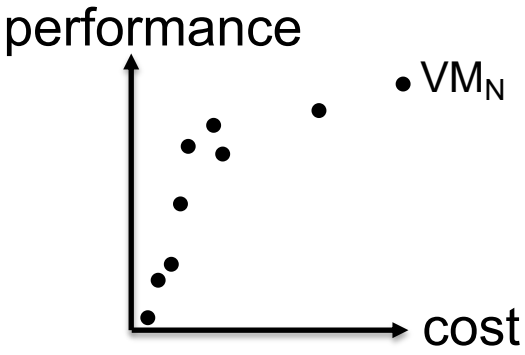
Idea

Micro Benchmarks

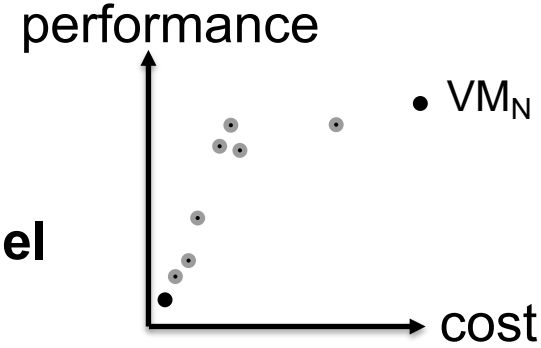


Application Benchmarks

Overall performance
(e.g., response time)

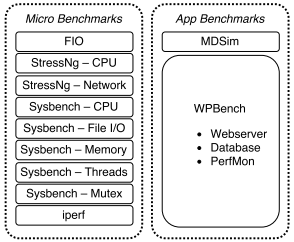


Evaluate a Prediction Model

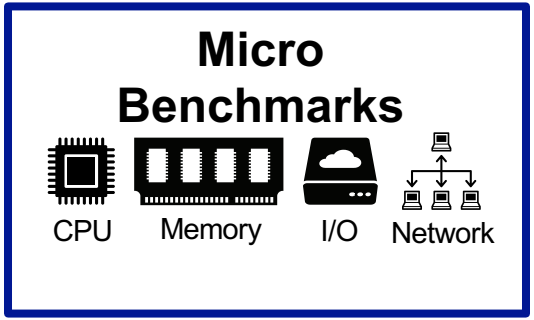


Methodology

Benchmark Design



Micro Benchmarks



Broad resource coverage and specific resource testing

CPU
- sysbench/cpu-single-thread
- sysbench/cpu-multi-thread
- stressng/cpu-callfunc
- stressng/cpu-double
- stressng/cpu-euler
- stressng/cpu-ftt
- stressng/cpu-fibonacci
- stressng/cpu-int64
- stressng/cpu-loop
- stressng/cpu-matrixprod

Memory
- sysbench/memory-4k-block-size
- sysbench/memory-1m-block-size

I/O
- [file I/O] sysbench/fileio-1m-seq-write
- [file I/O] sysbench/fileio-4k-rand-read
- [disk I/O] fio/4k-seq-write
- [disk I/O] fio/8k-rand-read

Network
- iperf/single-thread-bandwidth
- iperf/multi-thread-bandwidth
- stressng/network-epoll
- stressng/network-icmp
- stressng/network-sockfd
- stressng/network-udp

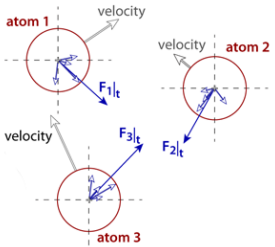
Software (OS)
- sysbench/mutex
- sysbench/thread-lock-1
- sysbench/thread-lock-128

Application Benchmarks

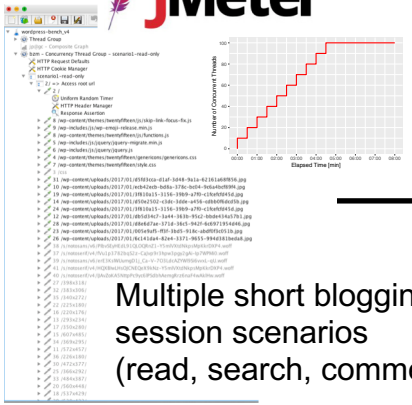
Application Benchmarks

Overall performance
(e.g., response time)

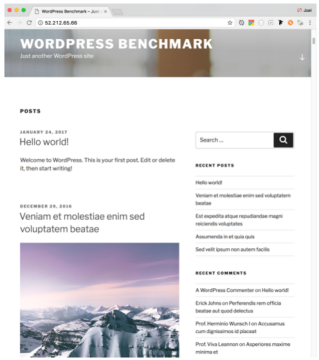
Molecular Dynamics Simulation (MDSim)



APACHE JMeter™ WordPress Benchmark (WPBench)

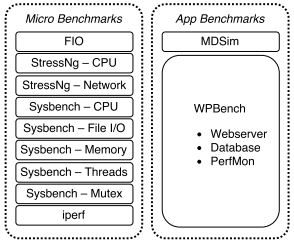


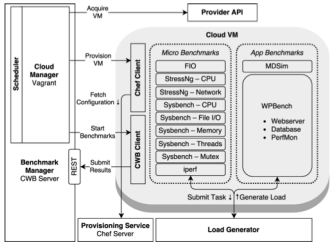
Multiple short blogging session scenarios (read, search, comment)



Methodology

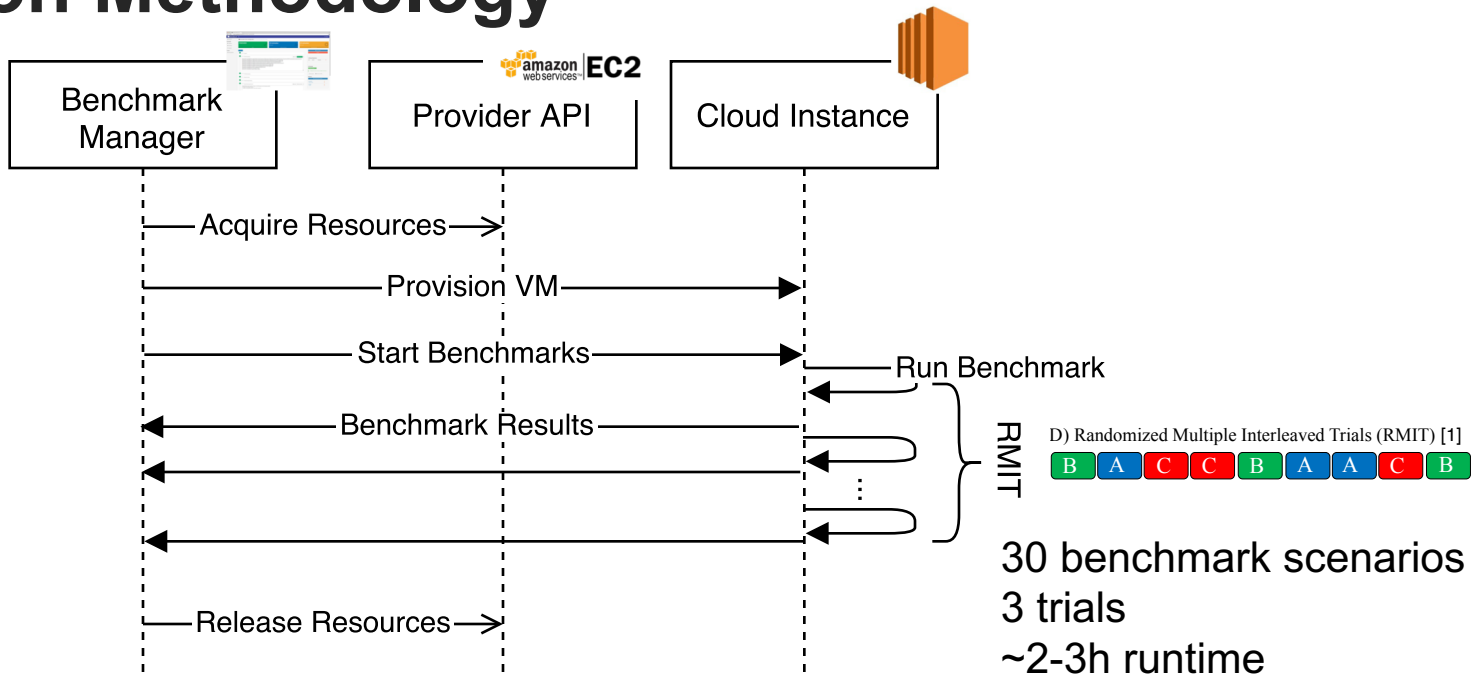
Benchmark Design → Benchmark Execution





A Cloud Benchmark Suite Combining Micro and Applications Benchmarks
QUDOS@ICPE'18, Scheuner and Leitner

Execution Methodology

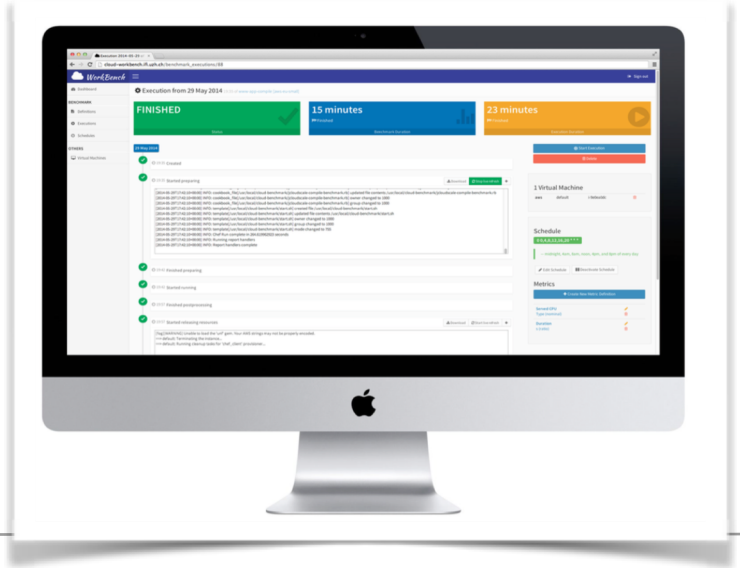


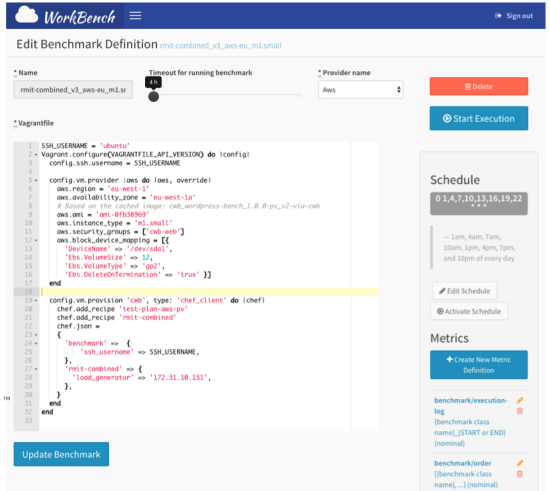
[1] A. Abedi and T. Brecht. Conducting repeatable experiments in highly variable cloud computing environments. ICPE'17

Benchmark Manager

Cloud WorkBench (CWB)

Tool for scheduling cloud experiments





HashiCorp Vagrant

CHEF

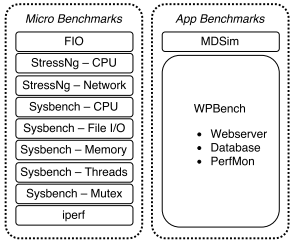
sealuzh/cloud-workbench

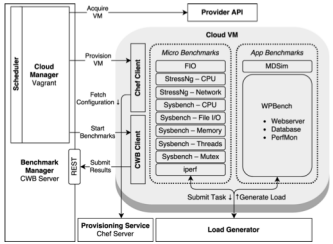
Cloud Work Bench – Infrastructure-as-Code Based Cloud Benchmarking
CloudCom'14, Scheuner, Leitner, Cito, and Gall

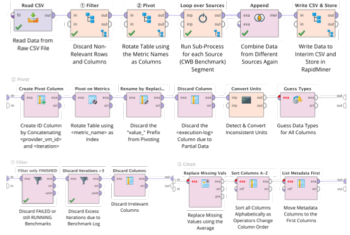
Cloud WorkBench: Benchmarking IaaS Providers based on Infrastructure-as-Code
Demo@WWW'15, Scheuner, Cito, Leitner, and Gall

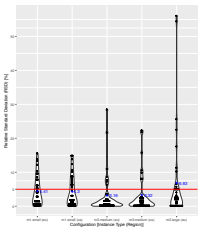
Methodology

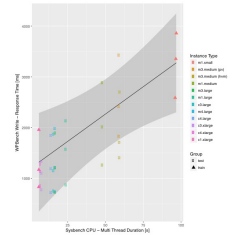
Benchmark Design → Benchmark Execution → Data Pre-Processing → Data Analysis











A Cloud Benchmark Suite Combining Micro and Applications Benchmarks
QUDOS@ICPE'18, Scheuner and Leitner

Estimating Cloud Application Performance Based on Micro Benchmark Profiling
CLOUD'18, Scheuner and Leitner

Performance Data Set

amazon web services EC2

RQ1+2

Instance Type	vCPU	ECU *	RAM [GiB]	Virtualization	Network Performance	
m1.small	1	1	1.7	PV	Low	eu + us
m1.medium	1	2	3.75	PV	Moderate	
m3.medium	1	3	3.75	PV /HVM	Moderate	eu + us
m1.large	2	4	7.5	PV	Moderate	
m3.large	2	6.5	7.5	HVM	Moderate	eu
m4.large	2	6.5	8.0	HVM	Moderate	
c3.large	2	7	3.75	HVM	Moderate	
c4.large	2	8	3.75	HVM	Moderate	
c3.xlarge	4	14	7.5	HVM	Moderate	
c4.xlarge	4	16	7.5	HVM	High	
c1.xlarge	8	20	7	PV	High	

PRE

* ECU := Elastic Compute Unit (i.e., Amazon's metric for CPU performance)

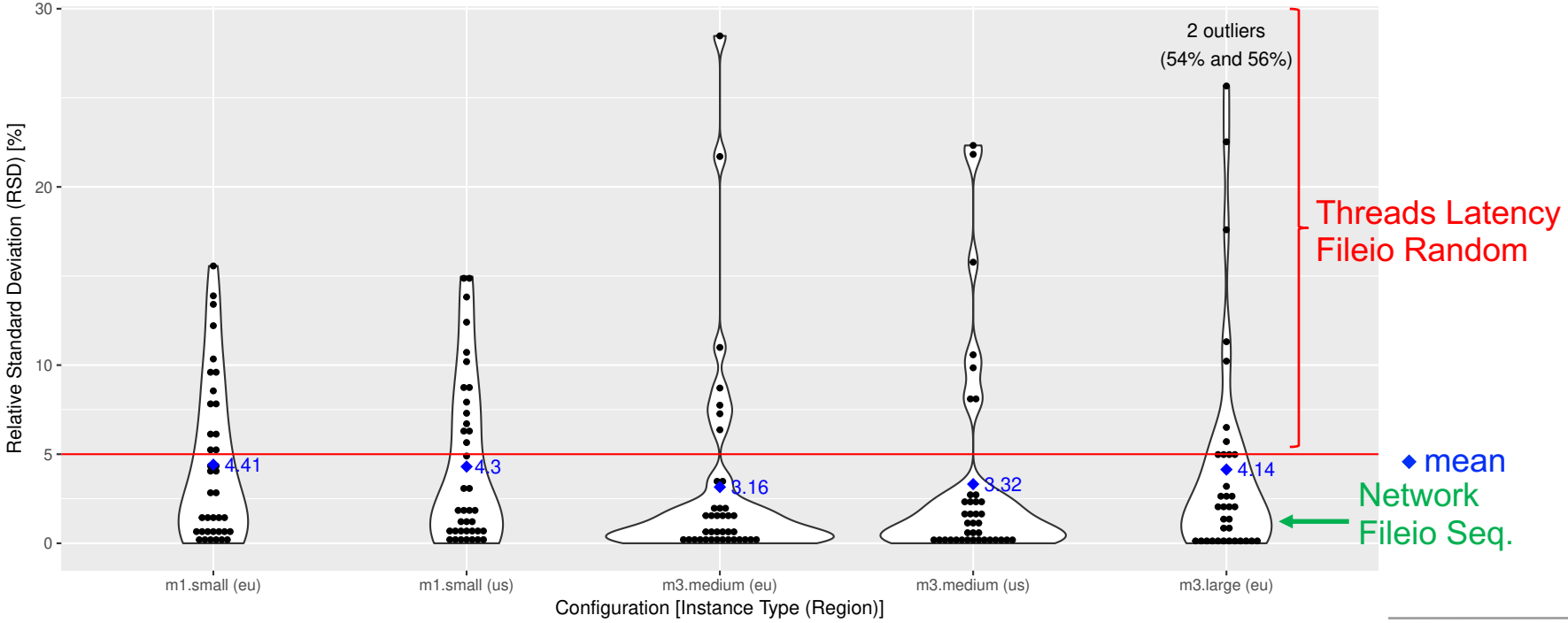
>240 Virtual Machines (VMs) à 3 Iterations → ~750 VM hours

>60'000 Measurements (258 per instance)

PRE – Performance Variability

Results

Does the performance of equally configured cloud instances vary relevantly?

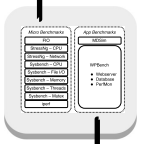


RQ1 – Estimation Accuracy

Approach

How accurate can a set of micro benchmarks estimate application performance?

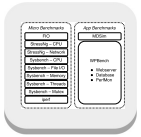
Instance Type$_1$
(m1.small)



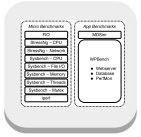
$micro_1$, $micro_2$, …, $micro_N$

app_1, app_2

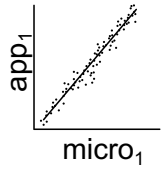
Instance Type$_2$



Instance Type$_{12}$
(c1.xlarge)



Linear Regression Model

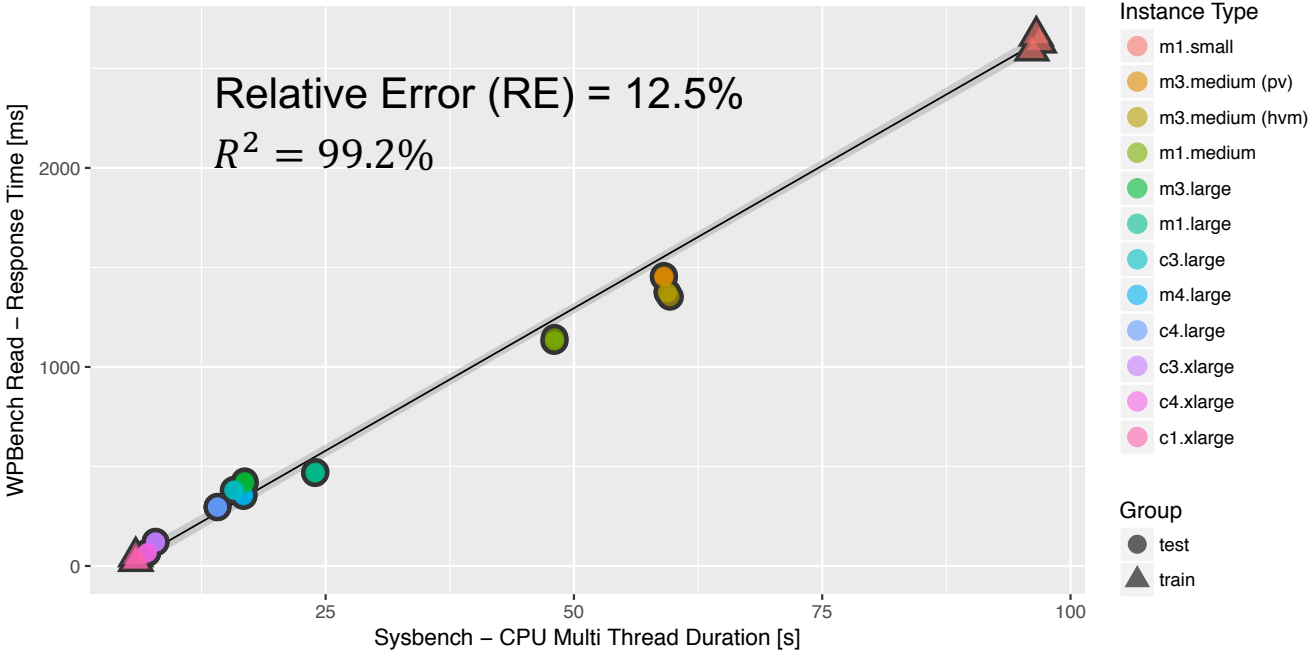


Forward feature selection to optimize relative error

RQ1 – Estimation Accuracy

Results

How accurate can a set of micro benchmarks estimate application performance?



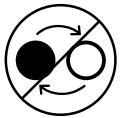
RQ2 – Micro Benchmark Selection — Results

Which subset of micro benchmarks estimates application performance most accurately?

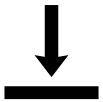
	Relative Error [%]
Micro Benchmark	
Sysbench – CPU Multi Thread	12
Sysbench – CPU Single Thread	454
Baseline	
vCPUs	616
ECU (i.e., Amazon's metric for CPU performance)	359
Cost	663

RQ – Implications

Suitability of **selected** micro benchmarks to estimate application performance



Benchmarks cannot be used interchangeable
→ Configuration is important



Baseline metrics vCPU and ECU are insufficient

Threats to Validity

Construct Validity

Almost 100% of benchmarking reports are wrong because benchmarking is "very very error-prone"[1]
[senior performance architect @Netflix]

→ Guidelines, rationalization, open source

Internal Validity

the extent to which cloud environmental factors, such as multi-tenancy, evolving infrastructure, or dynamic resource limits, affect the performance level of a VM instance

→ Variability PRE, stop interfering process

External Validity (Generalizability)

Other cloud providers?
Larger instance types?
Other application domains?

→ Future work

Reproducibility

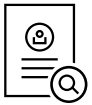
the extent to which the methodology and analysis is repeatable at any time for anyone and thereby leads to the same conclusions

⚠ dynamic cloud environment

→ Fully automated execution, open source

[1] https://www.youtube.com/watch?v=vm1GJMp0QN4&feature=youtu.be&t=18m29s

Related Work



Application Performance Profiling

- System-level resource monitoring [1,2]
- Compiler-level program similarity [3]

[1] Athanasia Evangelinou, Michele Ciavotta, Danilo Ardagna, Aliki Kopaneli, George Kousiouris, and Theodora Varvarigou. **Enterprise applications cloud rightsizing through a joint benchmarking and optimization approach**. *Future Generation Computer Systems*, 2016

[2] Mauro Canuto, Raimon Bosch, Mario Macias, and Jordi Guitart. **A methodology for full-system power modeling in heterogeneous data centers**. In *Proceedings of the 9th International Conference on Utility and Cloud Computing (UCC '16)*, 2016

[3] Kenneth Hoste, Aashish Phansalkar, Lieven Eeckhout, Andy Georges, Lizy K. John, and Koen De Bosschere. **Performance prediction based on inherent program similarity**. In *PACT '06*, 2006

Application Performance Prediction

- Trace and reply with Cloud-Prophet [4,5]
- Bayesian cloud configuration refinement for big data analytics [6]

[4] Ang Li, Xuanran Zong, Ming Zhang, Srikanth Kandula, and Xiaowei Yang. **Cloud-prophet: predicting web application performance in the cloud**. *ACM SIGCOMM Poster*, 2011

[5] Ang Li, Xuanran Zong, Srikanth Kandula, Xiaowei Yang, and Ming Zhang. **Cloud-prophet: Towards application performance prediction in cloud**. In *Proceedings of the ACM SIGCOMM 2011 Conference (SIGCOMM '11)*, 2011

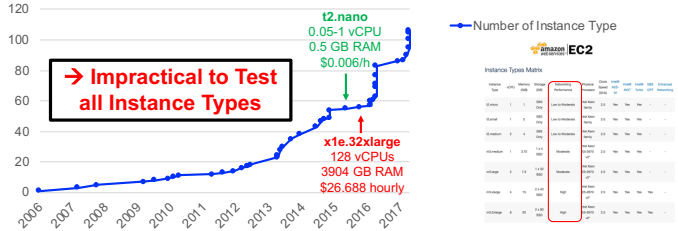
[6] Omid Alipourfard, Hongqiang Harry Liu, Jianshu Chen, Shivaram Venkataraman, Minlan Yu, and Ming Zhang. **Cherrypick: Adaptively unearthing the best cloud configurations for big data analytics**. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, 2017

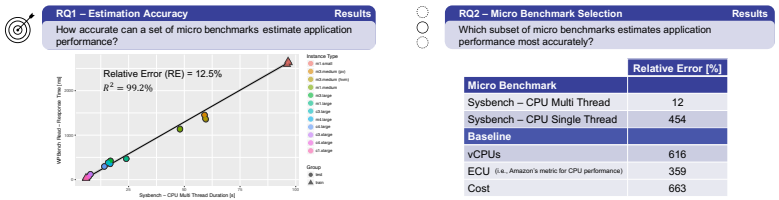
Conclusion

scheuner@chalmers.se
joe4dev

Motivation: Capacity Planning in IaaS Clouds

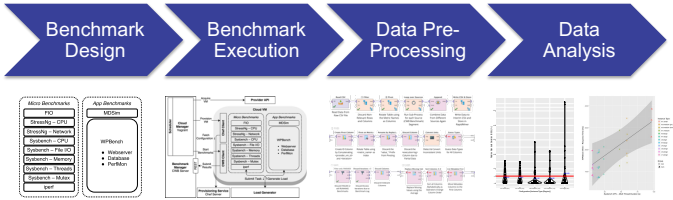
What cloud **service (i.e., instance type)** should I choose?





	Relative Error [%]
Micro Benchmark	
Sysbench – CPU Multi Thread	12
Sysbench – CPU Single Thread	454
Baseline	
vCPUs	616
ECU (i.e., Amazon's metric for CPU performance)	359
Cost	663

Methodology



A Cloud Benchmark Suite Combining Micro and Applications Benchmarks
QUDOS@ICPE'18, Scheuner and Leitner

Estimating Cloud Application Performance Based on Micro Benchmark Profiling
CLOUD'18, Scheuner and Leitner

RQ – Implications

- Suitability of **selected** micro benchmarks to estimate application performance
- Benchmarks cannot be used interchangeable → Configuration is important
- Baseline metrics vCPU and ECU are insufficient