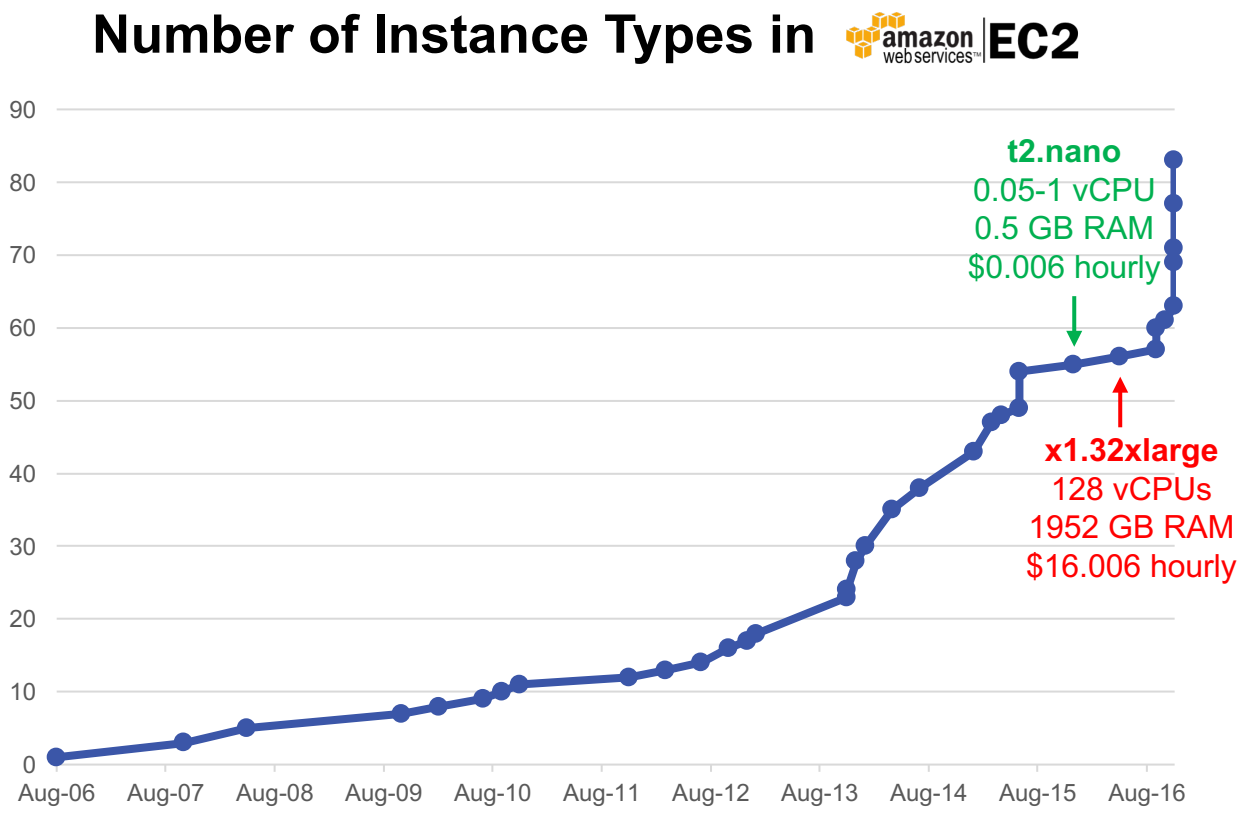
Cloud Benchmarking

Estimating Cloud Application Performance
Based on Micro Benchmark Profiling

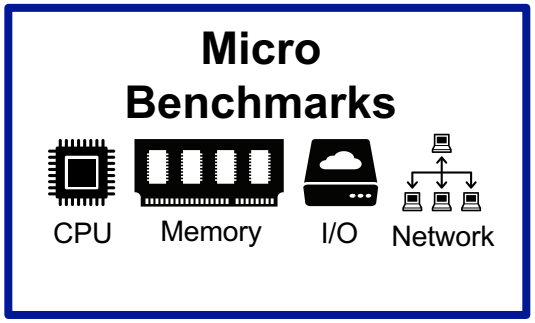
Joel Scheuner

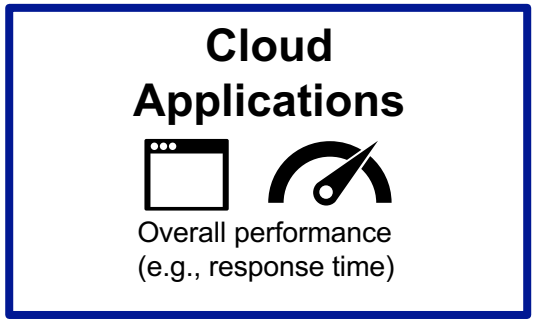
Problem



→ **Unpractical to Test all Instance Types**

Motivation





Generic

Artificial

Resource-specific

How relevant?

Specific

Real-World

Resource-heterogeneous

Research Questions

RQ1 – Performance Variability *within* Instance Types

Does the performance of equally configured cloud instances vary relevantly?

RQ2 – Application Performance Estimation *across* Instance Types

Can a set of micro benchmarks estimate application performance for cloud instances of different configurations?

RQ2.1 – Estimation Accuracy

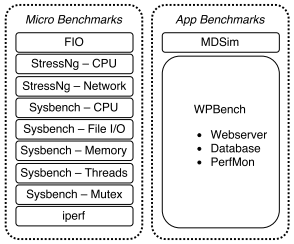
How accurate can a set of micro benchmarks estimate application performance?

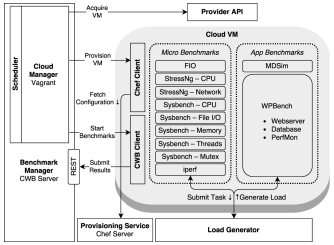
RQ2.2 – Micro Benchmark Selection

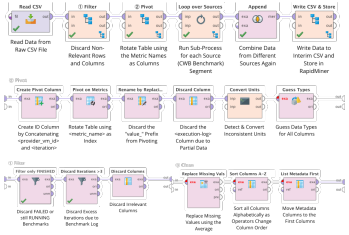
Which subset of micro benchmarks estimates application performance most accurately?

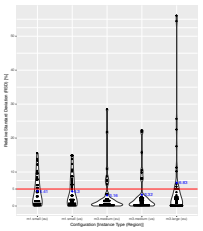
Methodology

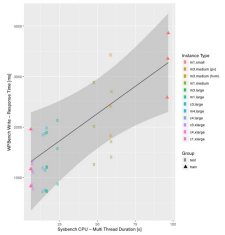
Benchmark Design → Benchmark Execution → Data Pre-Processing → Data Analyses











Performance Data Set

amazon web services EC2

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	

RQ1

RQ2

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

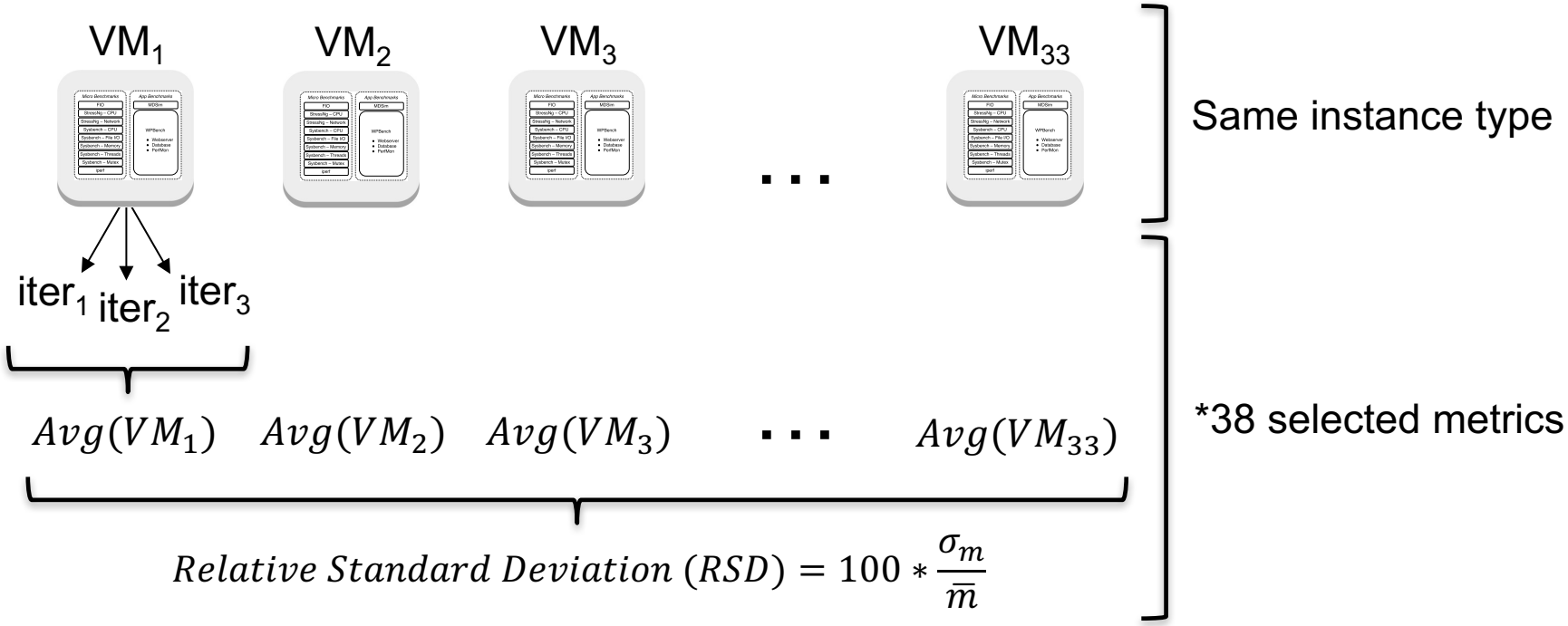
>240 Virtual Machines (VMs) à 3 Iterations → ~750 VM hours

>60'000 Measurements

RQ1 – Approach

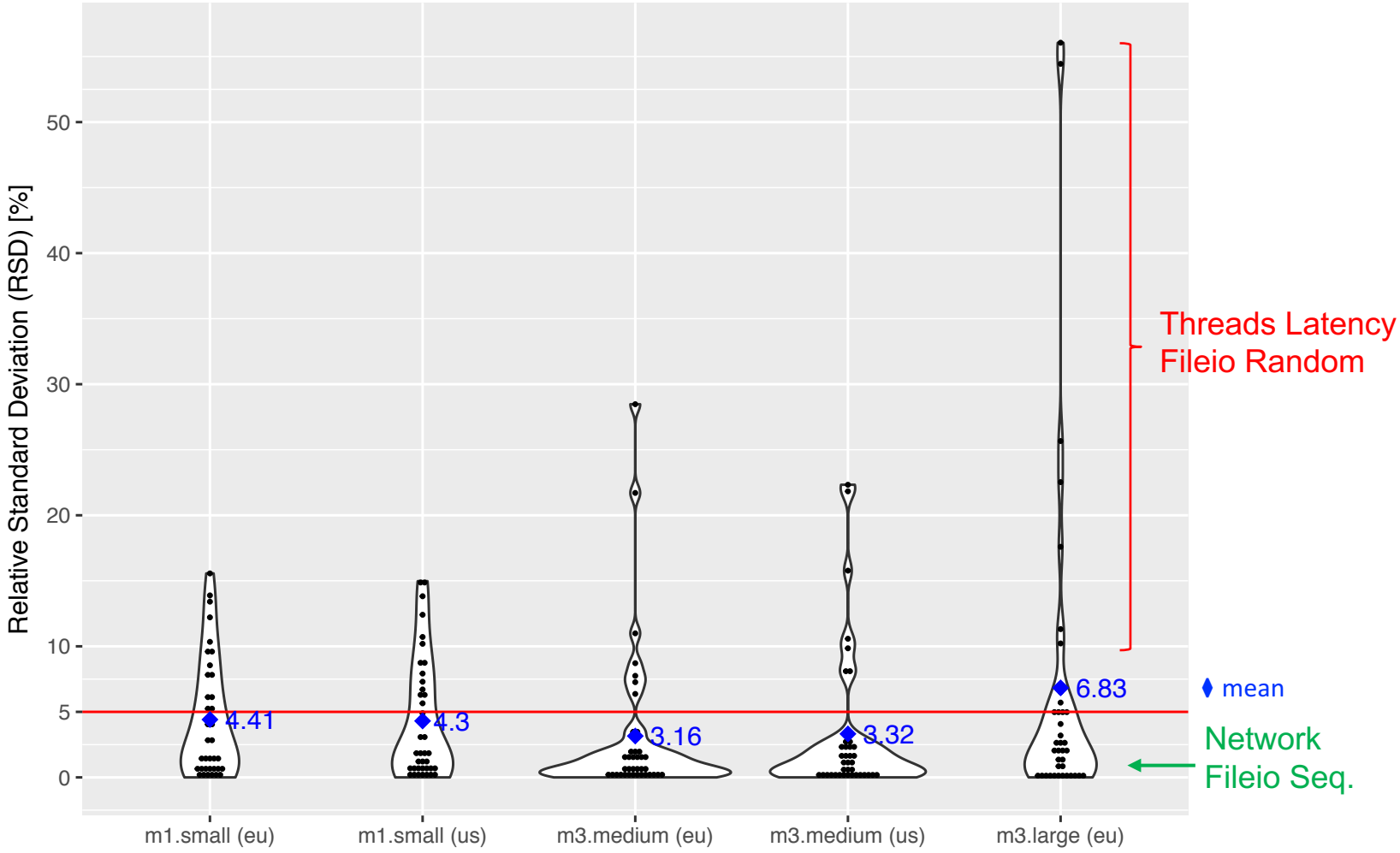
RQ1 – Performance Variability *within* Instance Types

Does the performance of equally configured cloud instances vary relevantly?



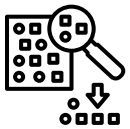
σ_m:= absolute standard deviation
$\bar{m}$:= mean of metric m

RQ1 – Results


Relative Standard Deviation (RSD) [%]
0
5
10
20
30
40
50
4.41
4.3
3.16
3.32
6.83
mean
Threads Latency
Fileio Random
Network
Fileio Seq.
m1.small (eu)
m1.small (us)
m3.medium (eu)
m3.medium (us)
m3.large (eu)
Configuration [Instance Type (Region)]

RQ1 – Implications

Hardware heterogeneity exploiting approaches are not worthwhile anymore [OZL+13, OZN+12, FJV+12]



Smaller sample size required to confidently assess instance type performance

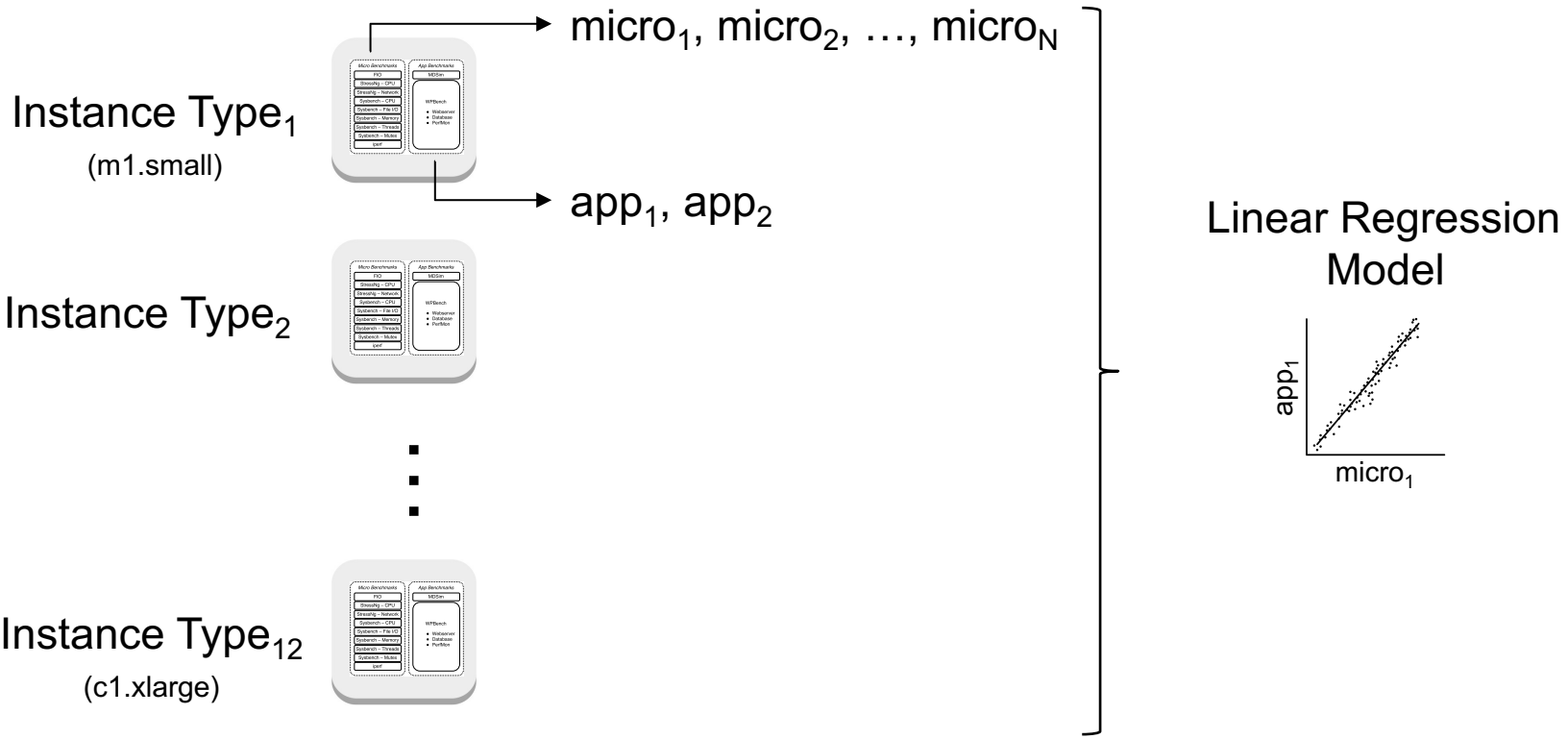
Fair offer

[OZL+13] Z. Ou, H. Zhuang, A. Lukyanenko, J. K. Nurminen, P. Hui, V. Mazalov, and A. Ylä- Jääski. **Is the same instance type created equal? exploiting heterogeneity of public clouds**. *IEEE Transactions on Cloud Computing*, 1(2):201–214, 2013
[OZN+12] Zhonghong Ou, Hao Zhuang, Jukka K. Nurminen, Antti Ylä-Jääski, and Pan Hui. **Exploiting hardware heterogeneity within the same instance type of amazon ec2**. In *Proceedings of the 4th USENIX Conference on Hot Topics in Cloud Computing (HotCloud'12)*, 2012
[FJV+12] Benjamin Farley, Ari Juels, Venkatanathan Varadarajan, Thomas Ristenpart, Kevin D. Bowers, and Michael M. Swift. **More for your money: Exploiting performance heterogeneity in public clouds**. In *Proceedings of the 3rd ACM Symposium on Cloud Computing (SoCC '12)*, pages 20:1–20:14, 2012

RQ2 – Approach

RQ2 – Application Performance Estimation *across* Instance Types

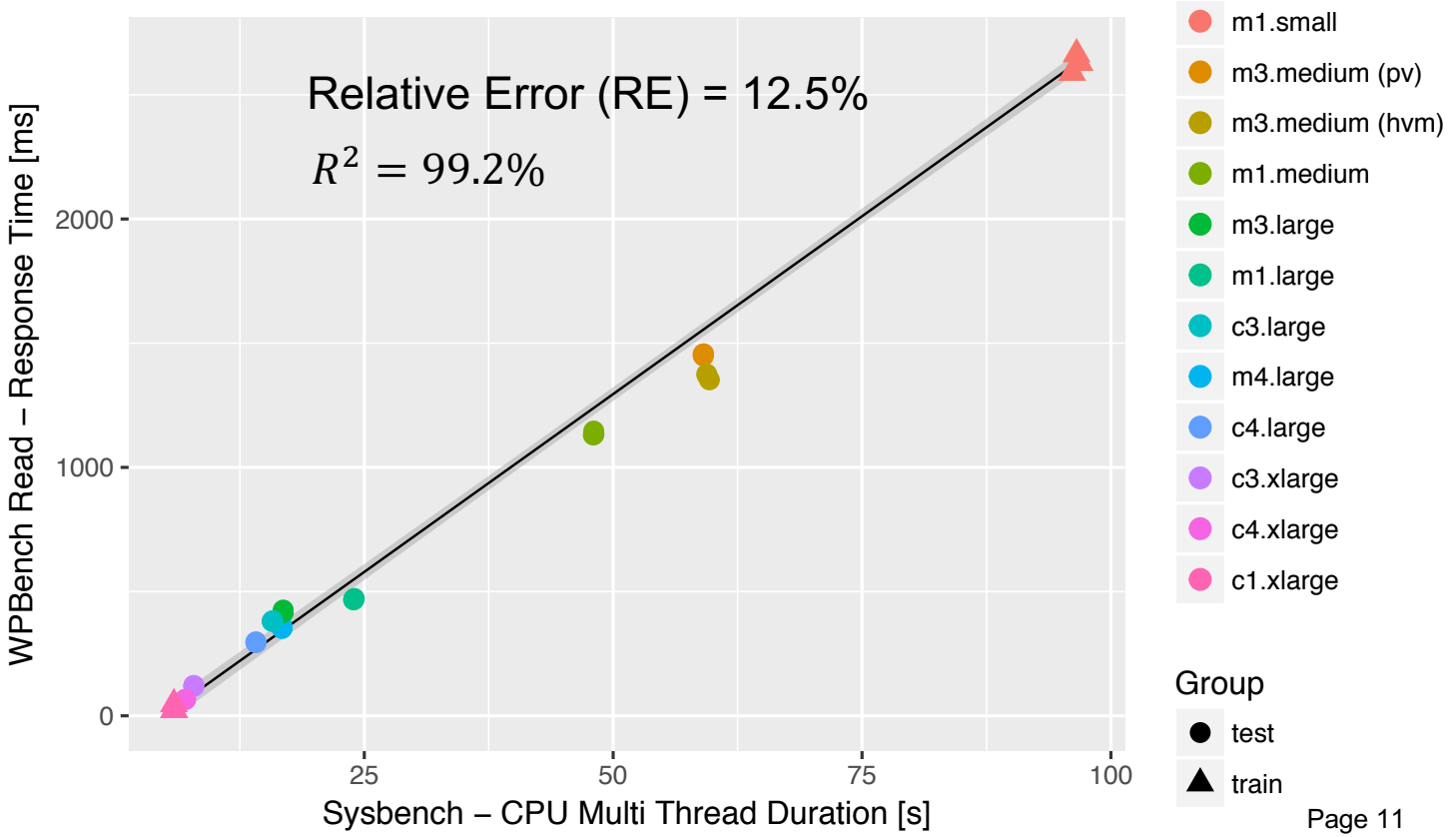
Can a set of micro benchmarks estimate application performance for cloud instances of different configurations?



RQ2.1 – Results

RQ2.1 – Estimation Accuracy

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



RQ2.2 – Results

RQ2.2 – Micro Benchmark Selection

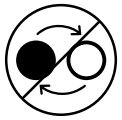
Which subset of micro benchmarks estimates application performance most accurately?

Estimation Results for WPBench Read – Response Time

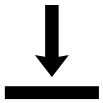
	Relative Error [%]	R^2 [%]
Benchmark		
Sysbench – CPU Multi Thread	12.5	99.2
Sysbench – CPU Single Thread	454.0	85.1
Baseline		
vCPUs	616.0	68.0
ECU	359.0	64.6

RQ2 – 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

Related Work

Application Performance Profiling

- System-level resource monitoring [ECA+16, CBMG16]
- Compiler-level program similarity [HPE+06]

Application Performance Prediction

- Trace and reply with Cloud-Prophet [LZZ+11, LZK+11]
- Bayesian cloud configuration refinement for big data analytics [ALC+17]

[ECA+16] 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

[CBMG16] 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)*, pages 20–29, 2016

[HPE+06] Kenneth Hoste, Aashish Phansalkar, Lieven Eeckhout, Andy Georges, Lizy K. John, and Koen De Bosschere.
Performance prediction based on inherent program similarity.
In *Proceedings of the 15th International Conference on Parallel Architectures and Compilation Techniques (PACT '06)*, pages 114–122, 2006

[LZZ+11] Ang Li, Xuanran Zong, Ming Zhang, Srikanth Kandula, and Xiaowei Yang.
Cloud-prophet: predicting web application performance in the cloud. *ACM SIGCOMM Poster*, 2011

[LZK+11] 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)*, pages 426–427, 2011

[ALC+17] 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

RQ1 – Performance Variability *within* Instance Types

Does the performance of equally configured cloud instances vary relevantly?

Outcome: No. Performance does *not* vary relevantly for most benchmarks in Amazon's EC2 cloud for all intensively tested configurations in two different regions.

RQ2 – Application Performance Estimation *across* Instance Types

Can a set of micro benchmarks estimate application performance for cloud instances of different configurations?

Outcome: Yes. Selective micro benchmarks are able to estimate certain application performance metrics with acceptable accuracy.

RQ2.1 – Estimation Accuracy

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

Outcome: Scientific computing application relative error below 10%
Web serving application relative error between 10% and 20%

RQ2.2 – Micro Benchmark Selection

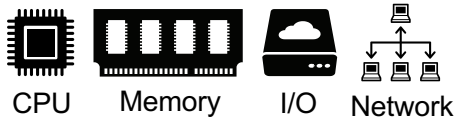
Which subset of micro benchmarks estimates application performance most accurately?

Outcome: Multi Thread CPU benchmark

APPENDIX

Motivation

Micro Benchmarks

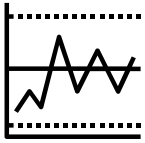


Cloud Applications

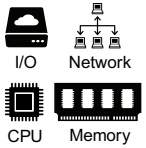
Overall performance
(e.g., response time)

Fast execution			Long-running
Straightforward setup			Complex setup
Clear interpretation			Bottleneck analysis

Related Work (1)



Cloud Performance Variability



Micro Benchmarking

Application Benchmarking

★ Hardware heterogeneity

[OZL+13, OZN+12, FJV+12, DPC10]

Analyst agencies

CLOUD SPECTATOR CloudHarmony

⚠ Reproducibility

[SSS+08, PSF16]

http://cloudsuite.ch/

[OZL+13] Z. Ou, H. Zhuang, A. Lukyanenko, J. K. Nurminen, P. Hui, V. Mazalov, and A. Ylä- Jääski. **Is the same instance type created equal? exploiting heterogeneity of public clouds**. *IEEE Transactions on Cloud Computing*, 1(2):201–214, 2013

[OZN+12] Zhonghong Ou, Hao Zhuang, Jukka K. Nurminen, Antti Ylä-Jääski, and Pan Hui. **Exploiting hardware heterogeneity within the same instance type of amazon ec2**. In *Proceedings of the 4th USENIX Conference on Hot Topics in Cloud Computing (HotCloud'12)*, 2012

[FJV+12] Benjamin Farley, Ari Juels, Venkatanathan Varadarajan, Thomas Ristenpart, Kevin D. Bowers, and Michael M. Swift. **More for your money: Exploiting performance heterogeneity in public clouds**. In *Proceedings of the 3rd ACM Symposium on Cloud Computing (SoCC '12)*, pages 20:1–20:14, 2012

[DPC10] Jiang Dejun, Guillaume Pierre, and Chi-Hung Chi. **EC2 Performance Analysis for Resource Provisioning of Service-Oriented Applications**, pages 197–207. Springer, 2010

[SSS+08] Will Sobel, Shanti Subramanyam, Akara Sucharitakul, Jimmy Nguyen, Hubert Wong, Arthur Klepchukov, Sheetal Patil, Armando Fox, and David Patterson. **Cloudstone: Multi-platform, multi-language benchmark and measurement tools for web 2.0**, 2008

[PSF16] Tapti Palit, Yongming Shen, and Michael Ferdman. **Demystifying cloud benchmarking**. In *2016 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 122–132, 2016

Related Work (2)

Application Performance Profiling

- System-level resource monitoring and micro benchmarks [ECA+16, CBMG16]
- Compiler-level program similarity [HPE+06]

Application Performance Prediction

- Trace and reply with Cloud-Prophet [LZZ+11, LZK+11]
- Bayesian cloud configuration refinement for big data analytics [ALC+17]
- Multi-component queuing models [SS05]

[ECA+16] 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
[CBMG16] 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)*, pages 20–29, 2016
[HPE+06] Kenneth Hoste, Aashish Phansalkar, Lieven Eeckhout, Andy Georges, Lizy K. John, and Koen De Bosschere. **Performance prediction based on inherent program similarity**. In *Proceedings of the 15th International Conference on Parallel Architectures and Compilation Techniques (PACT '06)*, pages 114–122, 2006

[LZZ+11] Ang Li, Xuanran Zong, Ming Zhang, Srikanth Kandula, and Xiaowei Yang. **Cloud-prophet: predicting web application performance in the cloud**. *ACM SIGCOMM Poster*, 2011
[LZK+11] 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)*, pages 426–427, 2011
[ALC+17] 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
[SS05] Christopher Stewart and Kai Shen. **Performance modeling and system management for multi-component online services**. In *Proceedings of the 2nd Conference on Symposium on Networked Systems Design & Implementation - Volume 2*, NSDI'05, pages 71– 84, Berkeley, 2005

Benchmark Design

a) Broad resource coverage

b) Specific resource testing

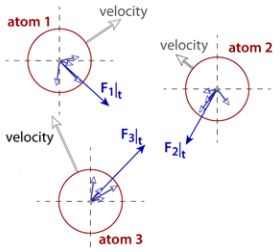
Micro Benchmarks

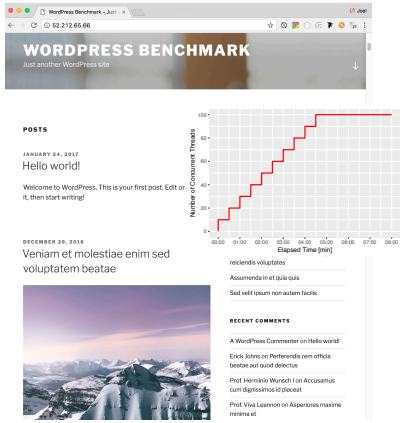
- FIO
- StressNg – CPU
- StressNg – Network
- Sysbench – CPU
- Sysbench – File I/O
- Sysbench – Memory
- Sysbench – Threads
- Sysbench – Mutex
- iperf

App Benchmarks

- MDSim
- WPBench
  - Webserver
  - Database
  - PerfMon

Molecular Dynamics Simulation (MDSim)

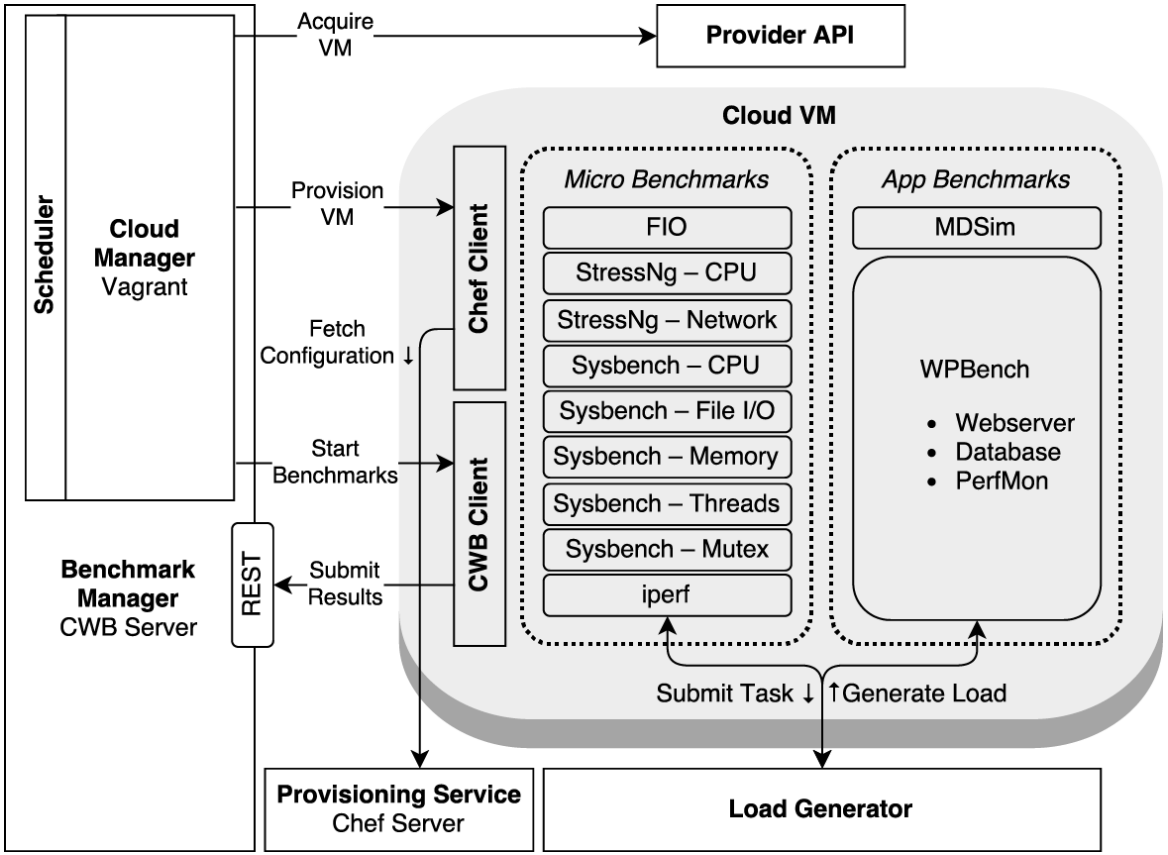




Wordpress Benchmark (WPBench)

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

Benchmark Execution



Benchmark Design → **Benchmark Execution** → Data Pre-Processing → Data Analyses

Benchmark Execution – Data Set

amazon web services EC2

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	

RQ2

RQ1

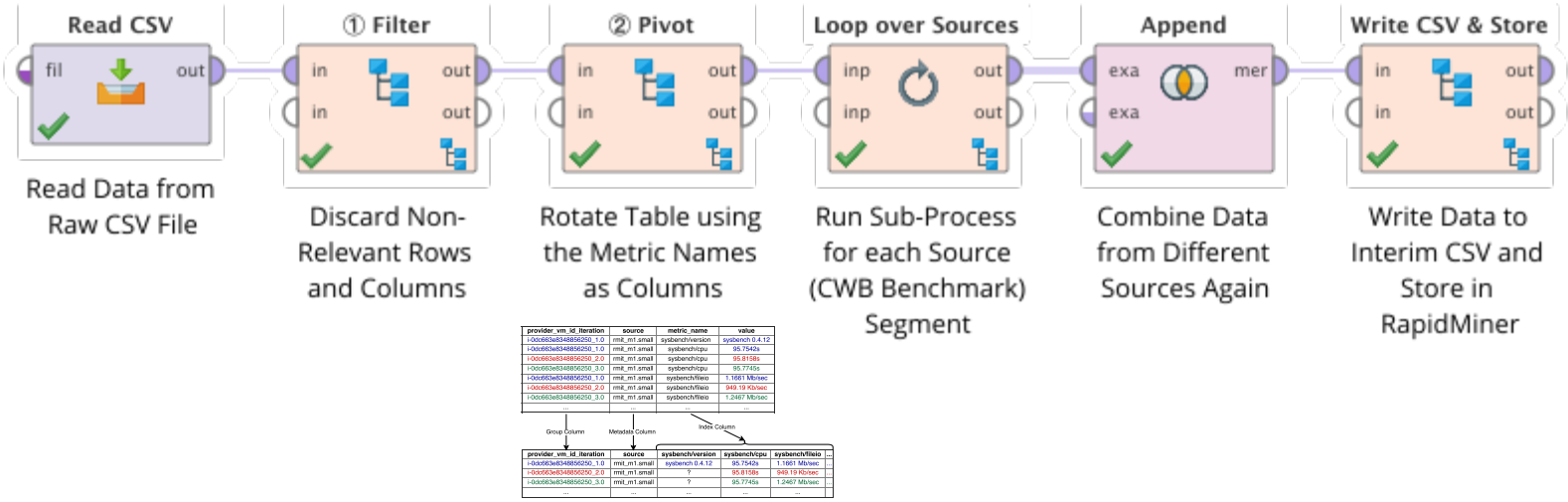
* ECU := Elastic Compute Unit (i.e., Amazon's metric for CPU performance)

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

>60'000 Measurements

Benchmark Design → **Benchmark Execution** → Data Pre-Processing → Data Analyses

Data Pre-Processing

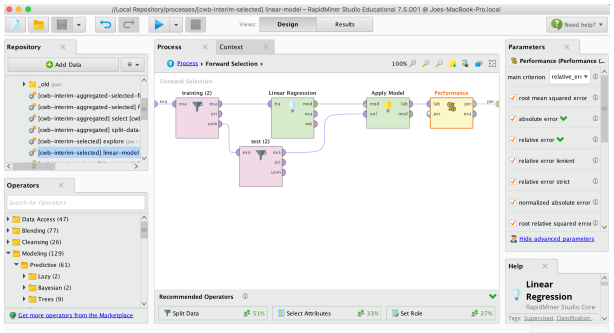


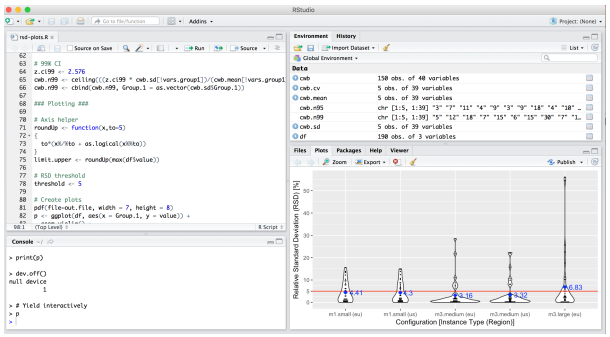
Data Analyses – Implementation

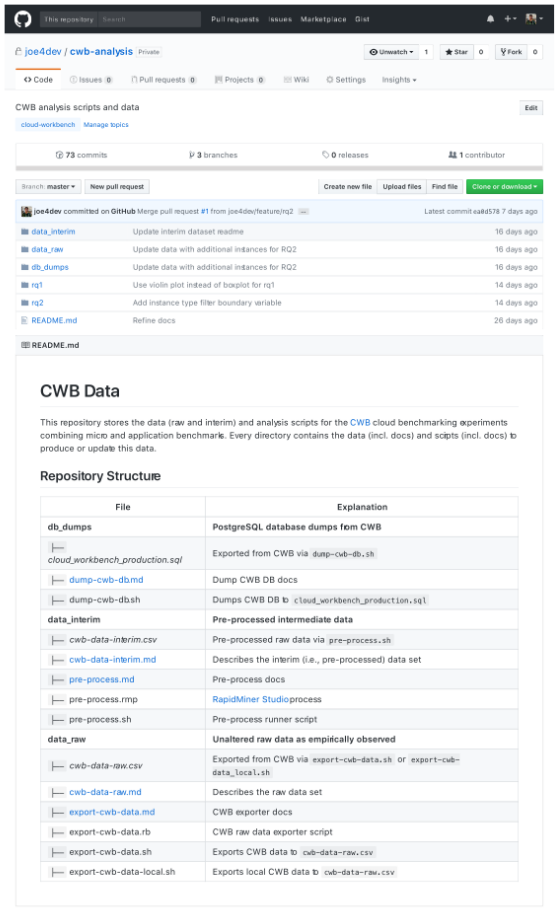
joe4dev/cwb-analysis

rapidminer STUDIO

RStudio®







BASH
THE BOURNE-AGAIN SHELL

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

Data Analyses – Results

Guided by the Research Questions …

RQ1 – Performance Variability *within* Instance Types

Does the performance of equally configured cloud instances vary relevantly?

RQ2 – Application Performance Estimation *across* Instance Types

Can a set of micro benchmarks estimate application performance for cloud instances of different configurations?

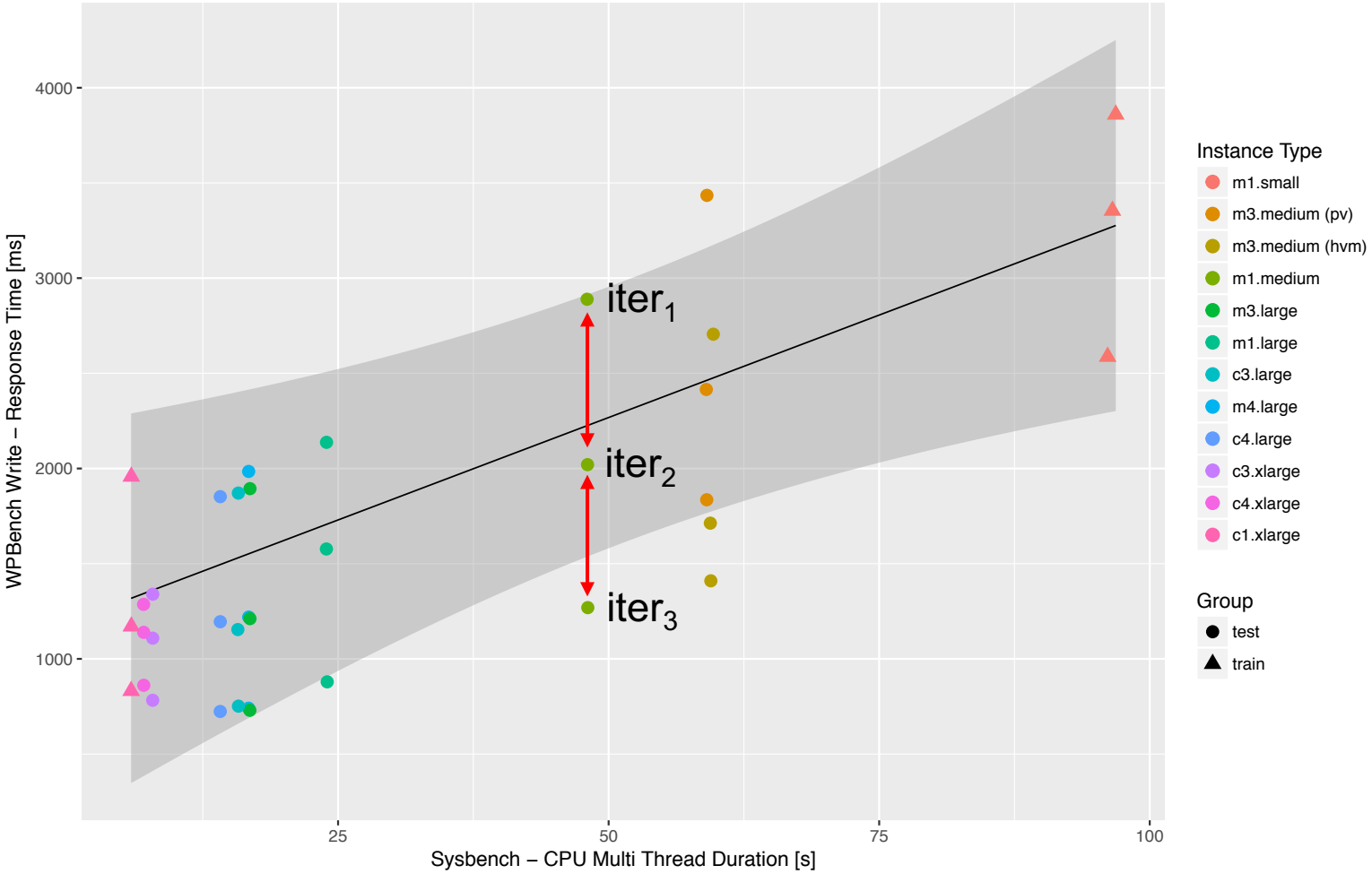
RQ2.1 – Estimation Accuracy

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

RQ2.2 – Micro Benchmark Selection

Which subset of micro benchmarks estimates application performance most accurately?

WPBench Write – Root Cause Analysis



Contributions

1. Extension of Cloud WorkBench (CWB) [SLCG14, SCLG15] with modular plugin system
2. Newly crafted Web serving application benchmark WPBench with three different load scenarios
3. Automated CWB benchmark that combines single-instance and multi-instance micro and application benchmarks
4. Raw and cleaned data set with performance metrics from Amazon EC2
5. Evaluation of an estimation model for application performance based on micro benchmark profiling

[SLCG14] Joel Scheuner, Philipp Leitner, Jürgen Cito, and Harald Gall. **Cloud WorkBench - Infrastructure-as-Code Based Cloud Benchmarking**. In *Proceedings of the 6th IEEE International Conference on Cloud Computing Technology and Science (CloudCom'14)*, 2014

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 RQ1, stop interfering process

External Validity (Generalizability)

Other cloud providers?
Larger instance types?
Other application domains?
→ Future work

Reproducibility

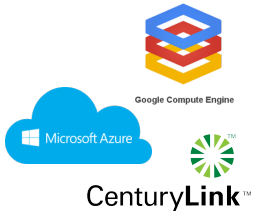
the extent to which the methodology and analysis is repeatable at any time for anyone and thereby leads to the same conclusions
⚠ dynamic cloud environment
→ Fully automated execution, open source

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

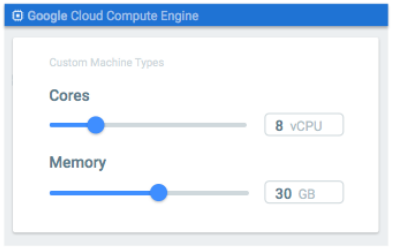
Open Challenges

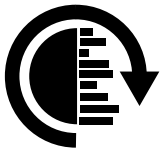
- How to identify a suitable micro benchmark estimator?

Future Work

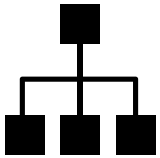


More IaaS providers → Custom instance types





Runtime performance data
Dedicated performance testing → Instance type selection as integral part of (vertical) scaling strategies



Multi-instance application architectures

Conclusion (1)

RQ1 – Performance Variability *within* Instance Types

Does the performance of equally configured cloud instances vary relevantly?

Outcome: No. Performance does *not* vary relevantly for most benchmarks in Amazon's EC2 cloud for all intensively tested configurations in two different regions.

RQ2 – Application Performance Estimation *across* Instance Types

Can a set of micro benchmarks estimate application performance for cloud instances of different configurations?

Outcome: Yes. Selective micro benchmarks are able to estimate certain application performance metrics with acceptable accuracy.

Conclusion (2)

RQ2.1 – Estimation Accuracy

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

Outcome: A scientific computing application achieves relative error rates below 10% and the response time of a Web serving application is estimated with a relative error between 10% and 20%.

RQ2.2 – Micro Benchmark Selection

Which subset of micro benchmarks estimates application performance most accurately?

Outcome: A single CPU benchmark was able to estimate the duration of a scientific computing application and the response time of a Web serving application most accurately.

Summary

Research Questions

RQ1 – Performance Variability *within* Instance Types

Does the performance of equally configured cloud instances vary relevantly?

RQ2 – Application Performance Estimation *across* Instance Types

Can a set of micro benchmarks estimate application performance for cloud instances of different configurations?

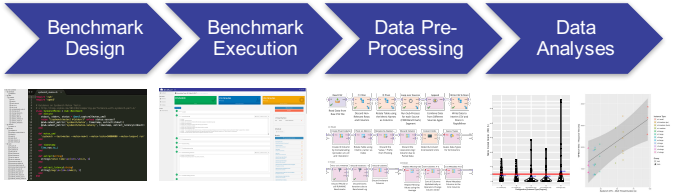
RQ2.1 – Estimation Accuracy

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

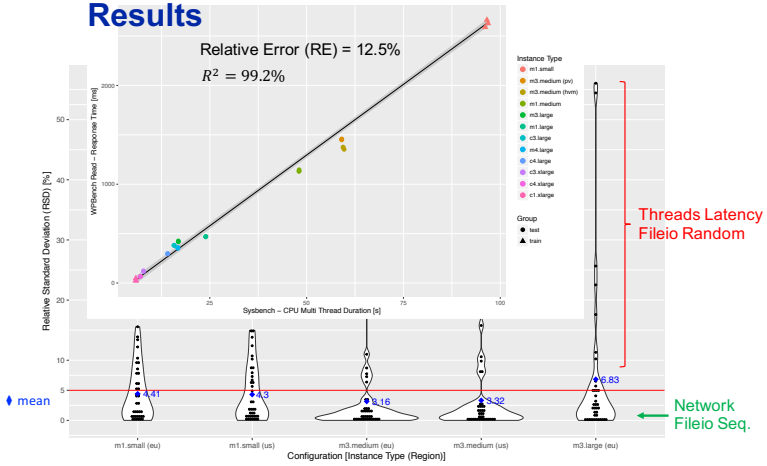
RQ2.2 – Micro Benchmark Selection

Which subset of micro benchmarks estimates application performance most accurately?

Methodology – Overview



Results



Background – Micro Benchmarks

File I/O: 4k random read

I/O

Prepare

```
ubuntu@ip-172-31-4-0:/tmp$ sysbench --test=fileio --file-total-size=4G
 --file-block-size=4K --file-test-mode=rndrd prepare
sysbench 0.4.12:  multi-threaded system evaluation benchmark

128 files, 32768Kb each, 4096Mb total
Creating files for the test
```

Command

```
ubuntu@ip-172-31-4-0:/tmp$ sysbench --test=fileio --file-total-size=4G
 --file-block-size=4K --file-test-mode=rndrd run
sysbench 0.4.12:  multi-threaded system evaluation benchmark

Running the test with following options:
Number of threads: 1

Extra file open flags: 0
128 files, 32Mb each
4Gb total file size
Block size 4Kb
Number of random requests for random IO: 10000
Read/Write ratio for combined random IO test: 1.50
Periodic FSYNC enabled, calling fsync() each 100 requests.
Calling fsync() at the end of test, Enabled.
Using synchronous I/O mode
Doing random read test
Threads started!
Done.

Operations performed:  10000 Read, 0 Write, 0 Other = 10000 Total
Read 39.062Mb  Written 0b  Total transferred 39.062Mb  (3.5793Mb/sec)
  916.31 Requests/sec executed

Test execution summary:
    total time:                          10.9133s
    total number of events:              10000
    total time taken by event execution: 10.9092
    per-request statistics:
         min:                                  0.00ms
         avg:                                  1.09ms
         max:                                596.05ms
         approx.  95 percentile:               3.27ms

Threads fairness:
    events (avg/stddev):           10000.0000/0.00
    execution time (avg/stddev):   10.9092/0.00

ubuntu@ip-172-31-4-0:/tmp$ sysbench --test=fileio --file-total-size=4G
 --file-block-size=4K --file-test-mode=rndrd cleanup
sysbench 0.4.12:  multi-threaded system evaluation benchmark

Removing test files...
```

Result

3.5793 MiB/sec

Cleanup

Bandwidth

Network

Server

```
ubuntu@ip-172-31-4-0:/tmp$ iperf --server --len 128k
------------------------------------------------------------
Server listening on TCP port 5001
TCP window size: 85.3 KByte (default)
------------------------------------------------------------
[  4] local 172.31.4.0 port 5001 connected with 172.31.10.209 port 46432
[ ID] Interval       Transfer     Bandwidth
[  4]  0.0-30.0 sec  3.39 GBytes   971 Mbits/sec
```

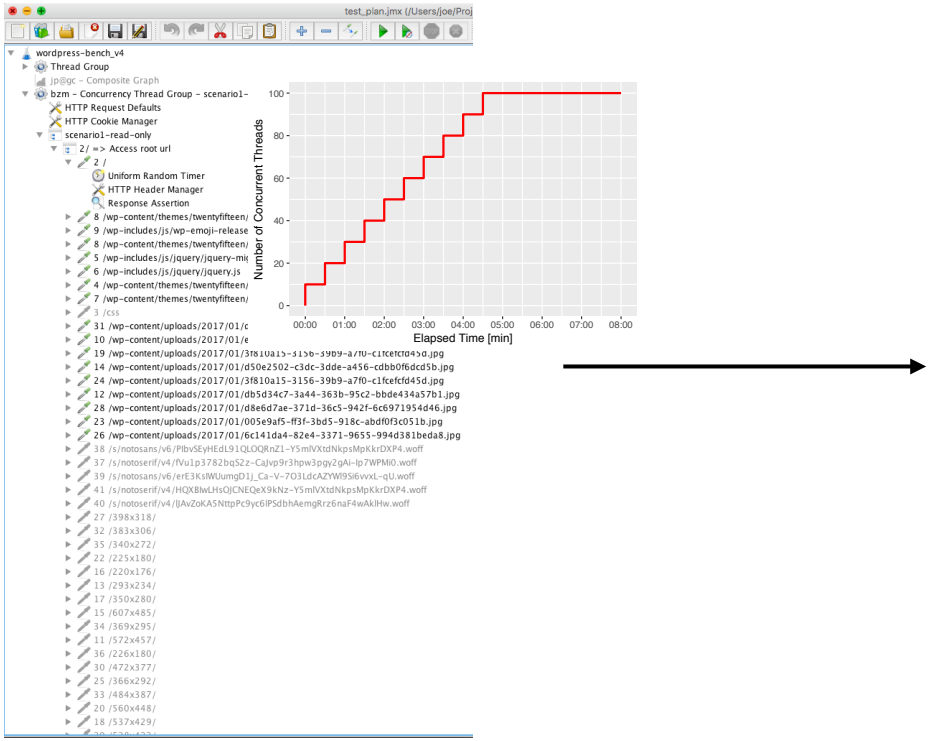
Client

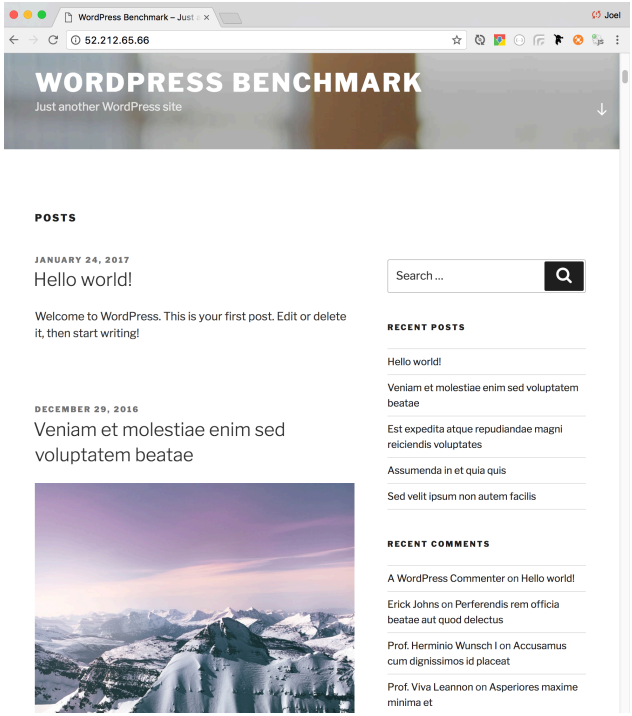
```
ubuntu@ip-172-31-10-209:/tmp$ iperf -c 172.31.4.0 -l 128k -t 30
------------------------------------------------------------
Client connecting to 172.31.4.0, TCP port 5001
TCP window size:  325 KByte (default)
------------------------------------------------------------
[  3] local 172.31.10.209 port 46432 connected with 172.31.4.0 port 5001
[ ID] Interval       Transfer     Bandwidth
[  3]  0.0-30.0 sec  3.39 GBytes  972 Mbits/sec
```

Result

972 Mbits/sec

Background – Application Benchmarks





Test plan (JMeter)

Webapp (Wordpress)

Related Work

Combining Micro and Application Benchmarks

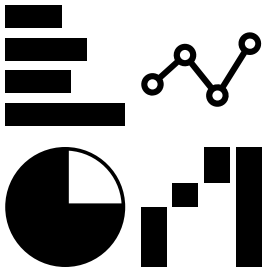
- B. Varghese and O. Akgun and I. Miguel and L. Thai and A. Barker, "**Cloud Benchmarking For Maximising Performance of Scientific Applications**" in IEEE TRANSACTIONS ON CLOUD COMPUTING (2016)
- A. Evangelinou and M. Ciavotta and D. Ardagna and A. Kopaneli and G. Kousiouris and T. Varvarigou, "**Enterprise applications cloud rightsizing through a joint benchmarking and optimization approach**" *Future Generation Computer Systems* - (2016)

Hardware / Performance Heterogeneity

- Farley, Benjamin and Juels, Ari and Varadarajan, Venkatanathan and Ristenpart, Thomas and Bowers, Kevin D. and Swift, Michael M., "**More for Your Money: Exploiting Performance Heterogeneity in Public Clouds**" Proceedings of the Third ACM Symposium on Cloud Computing (SoCC '12)
- Z. Ou and H. Zhuang and A. Lukyanenko and J. K. Nurminen and P. Hui and V. Mazalov and A. Ylä-Jääski, "**Is the Same Instance Type Created Equal? Exploiting Heterogeneity of Public Clouds**" *IEEE Transactions on Cloud Computing (2013)*

Typical Performance Data

- Instance metadata
  - CPU model
  - # CPU cores
- Benchmark metadata
  - Version number (including compiler)
- Benchmark
  - Execution time
  - Latency / response time
  - Throughput / Bandwidth
  - Operations / sec



Execution Methodology

A) Single Trial

B) Multiple Consecutive Trials (MCT)

C) Multiple Interleaved Trials (MIT)

D) Randomized Multiple Interleaved Trials (RMIT)

Ali Abedi and Tim Brecht, **Conducting Repeatable Experiments in Highly Variable Cloud Computing Environments** (ICPE'17)