An Empirical Study on the Discrepancy between Performance Testing Results from Virtual and Physical Environments

Muhammad Moiz Arif

Weiyi Shang

Emad Shihab

Department of Engineering & Computer Science
Concordia University, Montreal, Canada.

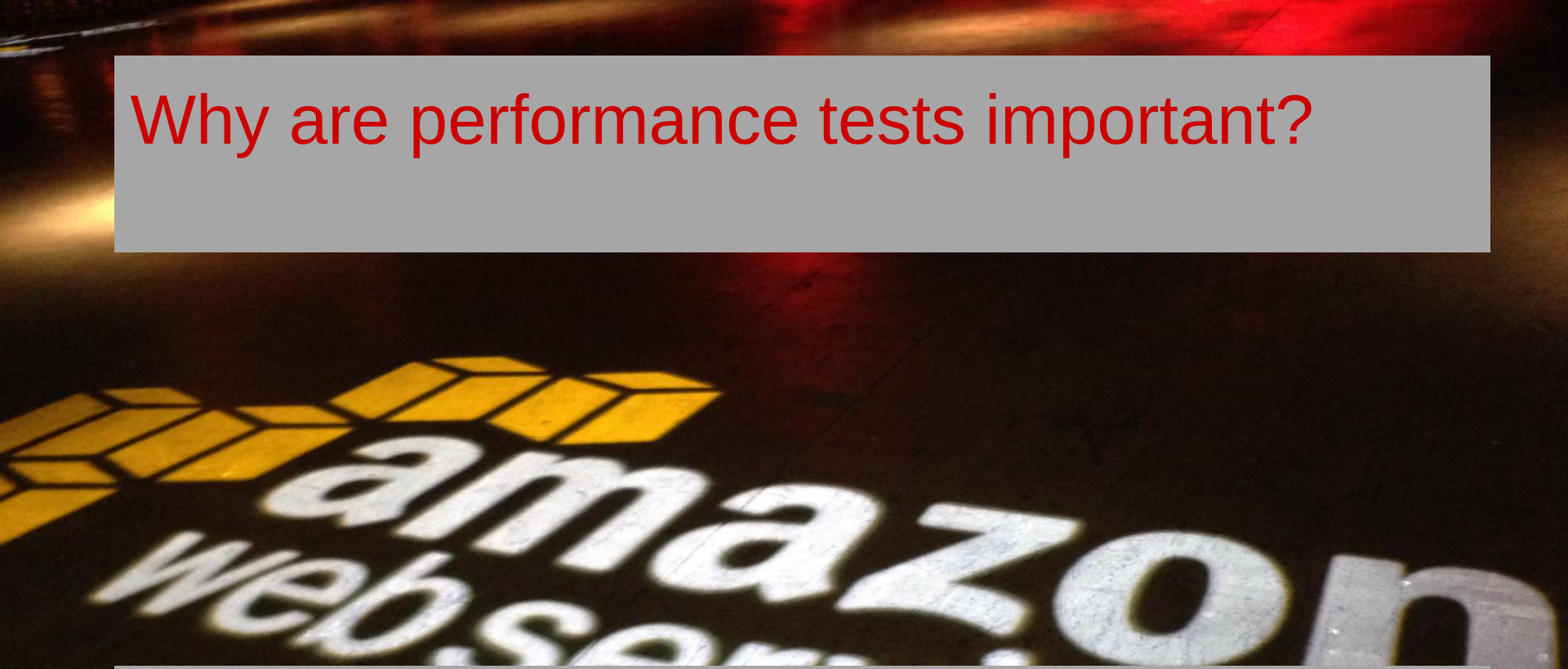
Concordia University
Engineering and Computer Science

What are performance tests?

- Performance assurance activities: ensuring software meets performance requirements.

- Mimicking user behaviour.

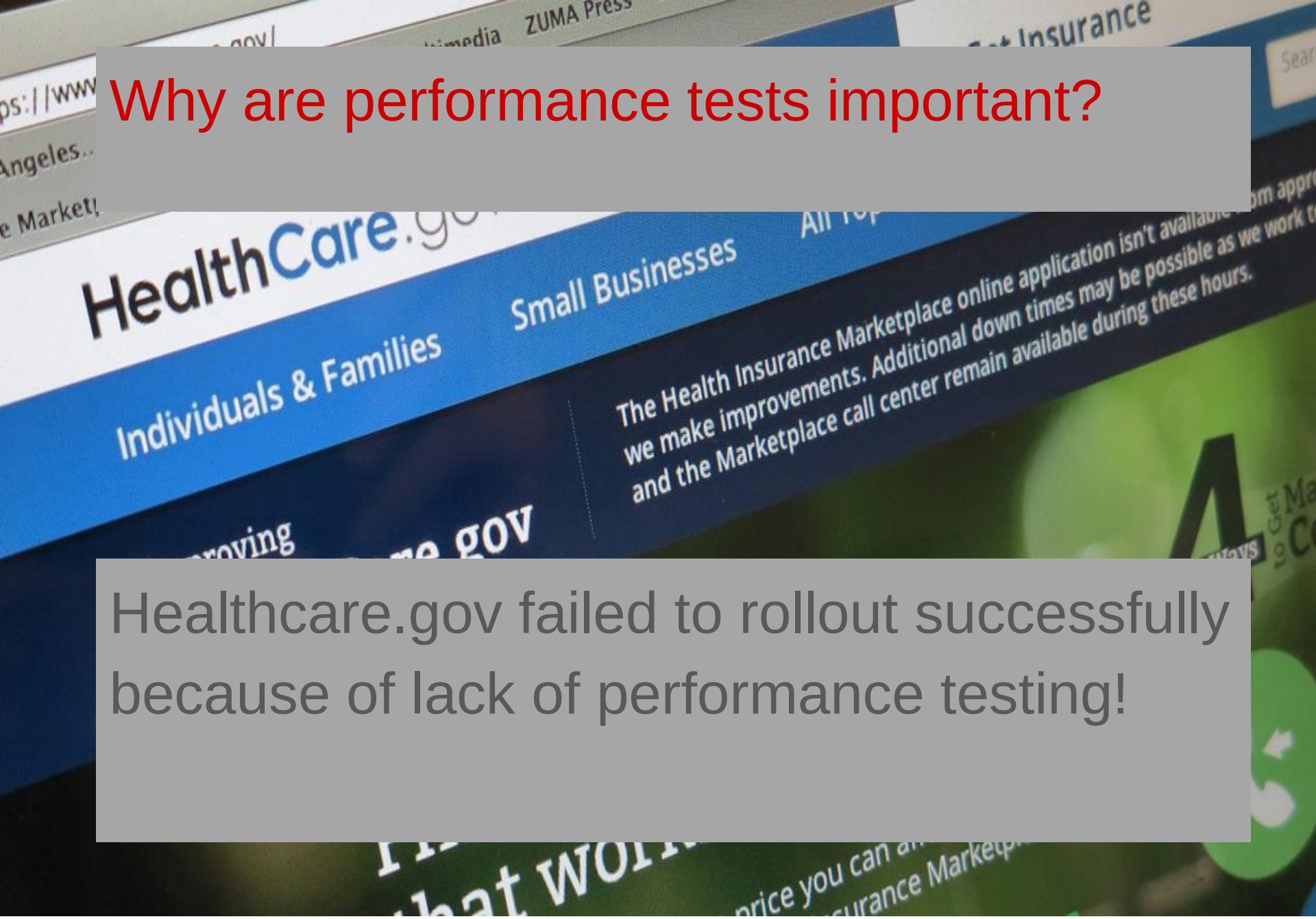
Why are performance tests important?



Amazon estimates that a one-second page-load slowdown can cost up to $1.6 billion!

https://tctechcrunch2011.files.wordpress.com/2014/01/aws_logo.jpg

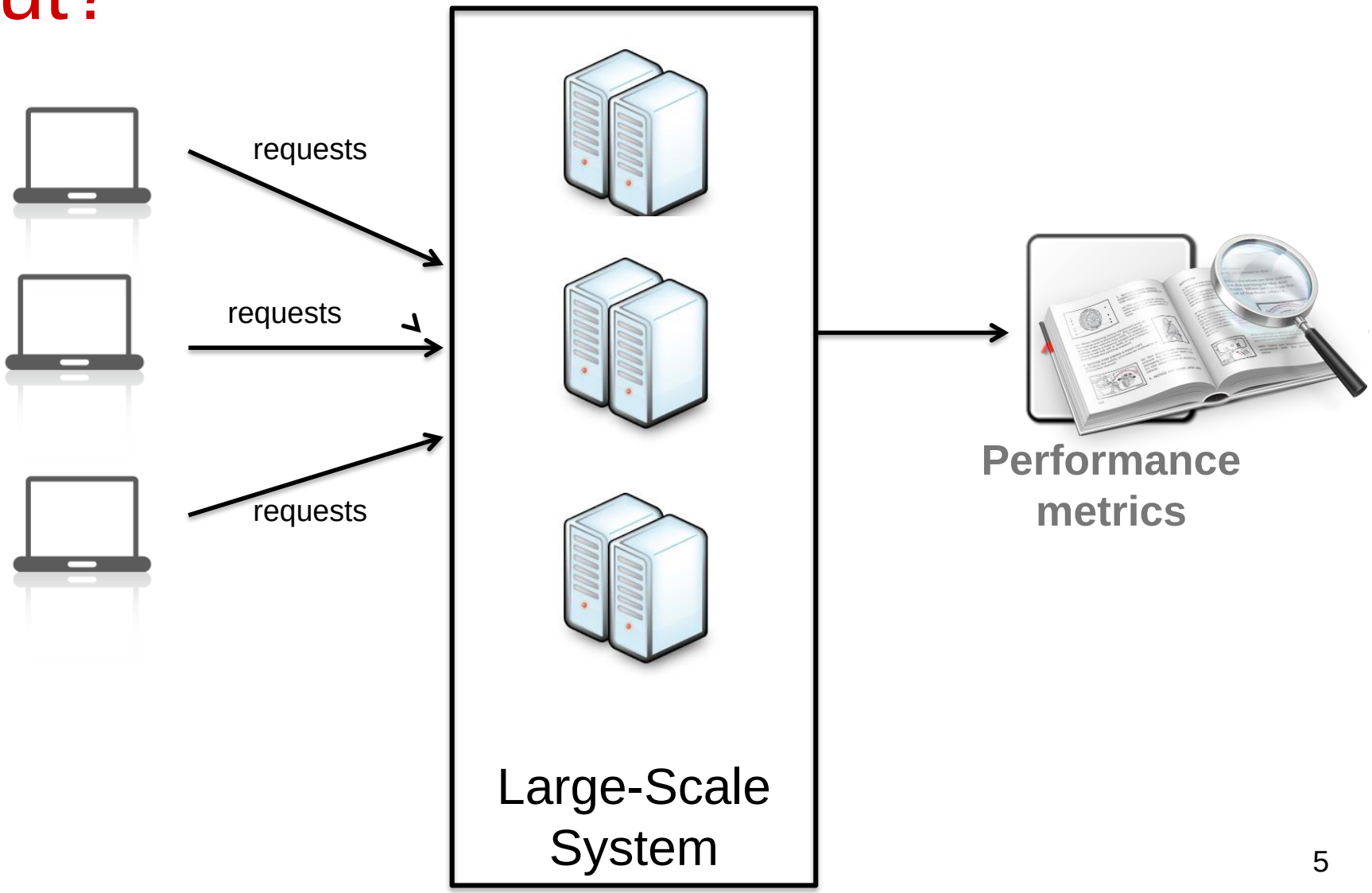
Why are performance tests important?



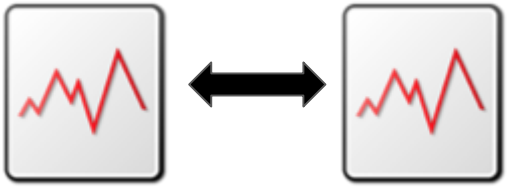
Healthcare.gov failed to rollout successfully because of lack of performance testing!

http://dailysignal.com/wp-content/uploads/140918_HealthCare_Attkisson.jpg

How is a performance test carried out?



Literature review: Types of performance analysis



- Single performanc e metric

[Nguyen, T.H. et al,ICPE,2012]

- Relationship between performance metrics

[Cohen, I.. et al,OSDI,2004]

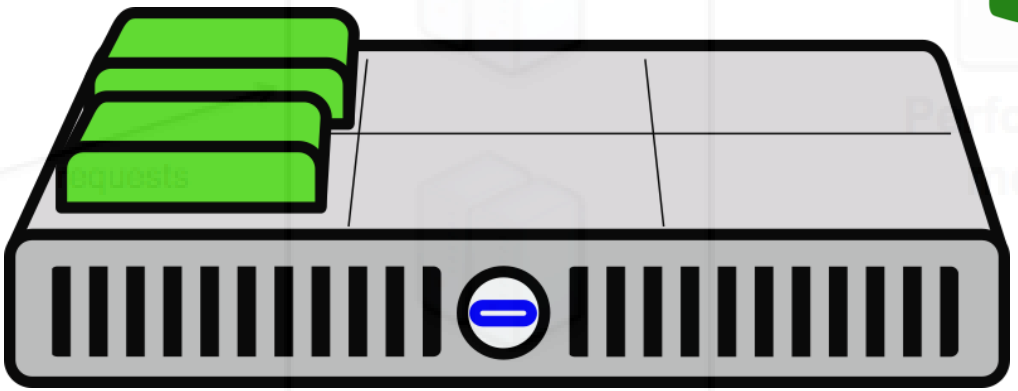
- Statistical modeling on performance metrics

[Shang, W. et al,ICPE,2015]

What if you need to modify your environment?

Image:
https://openclipart.org/detail/191766/question-guy

What if you need to modify your environment?



VIRTUAL MACHINES!

Image:
https://openclipart.org/detail/191766/question-guy
https://openclipart.org/detail/167198/two-virtual-machines

Examples from the industry

Load Testing SugarCRM in a Virtual Machine

Determining the CPU cost of virtualization with VMware ESX

Christopher L Merrill

v1.1

Read and post comments

Summary

The performance of our reference application under load (a default SugarCRM installation) on a virtualized server showed a 14% decrease, measured by total system capacity, compared to the same system running natively on equivalent hardware.

Overview

In a typical virtualization deployment scenario, a virtual machine replaces an existing physical machine. Multiple VMs will usually be deployed on a single host machine, sharing the resources of the host. For instance, VMs with a total of 16 cores might be deployed on a 4 core host machine. As long as utilization remains low, the processor resources can be shared while maintaining application performance goals. However, when the utililization of a single VM becomes very high, performance of one or all the VMs will suffer if the resource sharing continues. With the right hardware and software

https://www.webperformance.com/library/reports/Virtualization2/index.html

Examples from the industry

*"The **performance of our reference application** under load (a default SugarCRM installation) on a **virtualized server** showed a 14% decrease, measured by total system capacity, **compared to the same system running natively on equivalent hardware**."*

https://www.webperformance.com/library/reports/Virtualization2/index.html

Examples from the industry

performance-testing systems on virtual machines that normally run on physical machines

7

1

My employer runs some of our systems on physical machines with attached hard drives. I am charged with performance-testing those systems. For cost reasons, I've been asked to test those systems running on virtual machines (using Xen) attached to a SAN. This is clearly not an apples-to-apples comparison. Some systems use a lot of disk I/O, and so the SAN issue is especially worrisome. Rather than responding with "can't be done" or "not reliable", I want to recommend what *is* possible.

Here are some things that come to mind or that I've found with Google searches:

- Measure SAN speed vs. hard drive and calculate a ratio
- Borrow a physical machine long enough to run a benchmark, do the same with a virtual machine and calculate a ratio
- Even if you can't predict absolute performance on physical machines, you may be able to predict relative performance (i.e. whether the candidate release will be faster or slower than what's currently in production)
- Measure multiple times at different times of day to mitigate resource contention issues, i.e. conflicts with other virtual machines running on the same physical machine or with other clients using the SAN.

Are there other things you can do to mitigate differences between physical machine performance and virtual machine performance in an environment similar to mine? I am particularly interested in actual experiences rather than educated guesses.

performance virtualization xen

aske
view
activ

Ge

Rela
1
3

https://sqa.stackexchange.com/questions/7709/performance-testing-systems-on-virtual-machines-that-normally-run-on-physical-ma

Examples from the industry

performance-testing systems on virtual machines that normally run on physical machines

My employer runs some of our systems on physical machines with attached hard drives. I am charged with performance-testing those systems. For cost reasons, I've been asked to test those systems running on virtual machines (using Xen) attached to a SAN. This is clearly not an apples-to-apples comparison. Some systems use a lot of disk I/O, and so the SAN issue is especially worrisome. Rather than responding with "can't be done" or "not reliable", I want to recommend what

7

Here are some things that come to mind or that I've found with Google searches:

- Borrow a physical machine long enough to run a benchmark, do the same with a virtual
- Even if you can't predict absolute performance on physical machines, you may be able to
- Measure multiple times at different times of day to mitigate resource contention issues, i.e. conflicts with other virtual machines running on the same physical machine or with other clients using the SAN.

Are there other things you can do to mitigate differences between physical machine performance and virtual machine performance in an environment similar to mine? I am particularly interested in actual experiences rather than educated guesses.

performance virtualization xen

*...“For **cost reasons**, I've been asked to test those systems running on **virtual machines** (using Xen) attached to a SAN. **This is clearly not an apples-to-apples comparison**.”...*

https://sqa.stackexchange.com/questions/7709/performance-testing-systems-on-virtual-machines-that-normally-run-on-physical-ma

Research hypothesis

- *For software testing activities there exists a discrepancy between **physical** and **virtual** environments.*

- *We believe that the approaches used so far do not take into account the heterogeneous environments.*

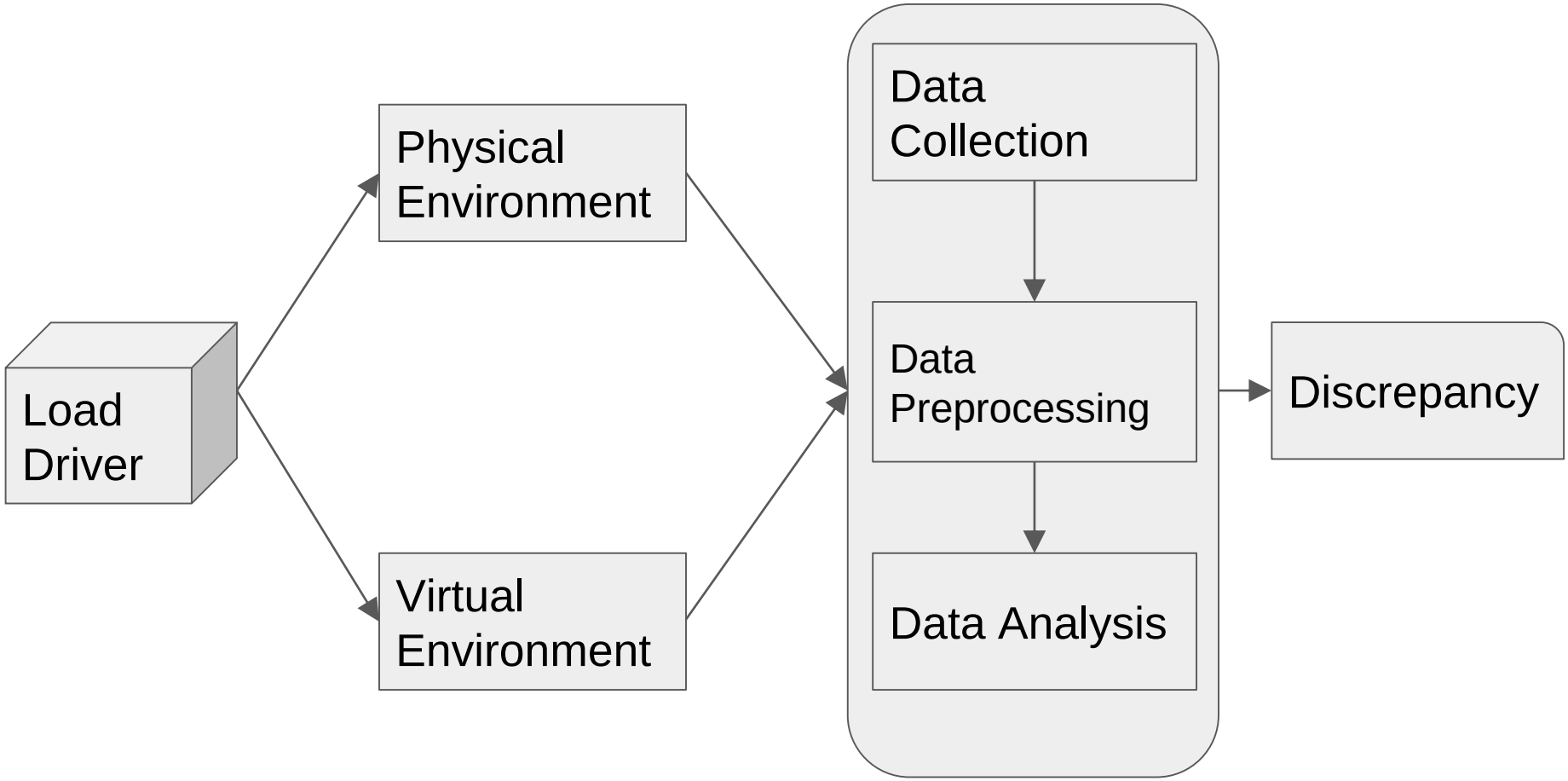
Research Questions

RQ1: Are the trend and distribution of a single performance metric similar across environments?

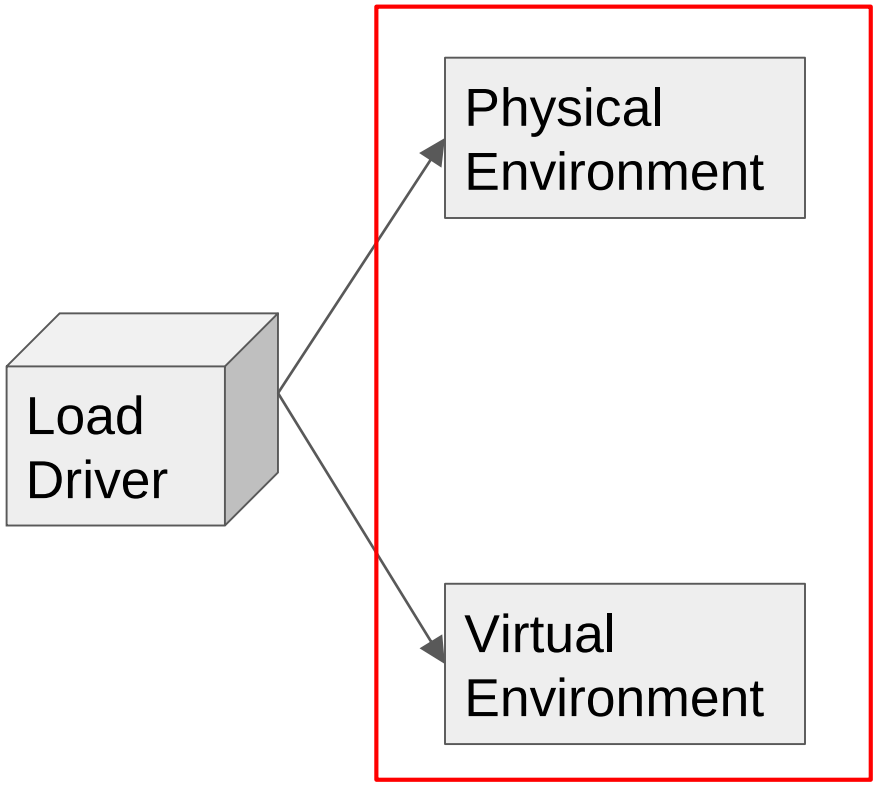
RQ2: To what extent does the relationship between the performance metrics change across environments?

RQ3: Can statistical performance models be applied across virtual and physical environments?

Approach


Load Driver
Physical Environment
Virtual Environment
Data Collection
Data Preprocessing
Data Analysis
Discrepancy

Approach



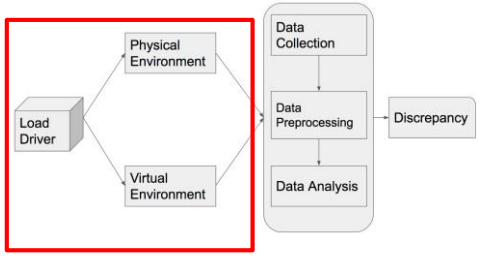
1.Dell DVD Store (DS2)
2. Cloudstore

Approach

- Both the system are used in prior studies.
- DS2 comes with a load driver.
- Cloudstore, we used JMeter to replicate the load.

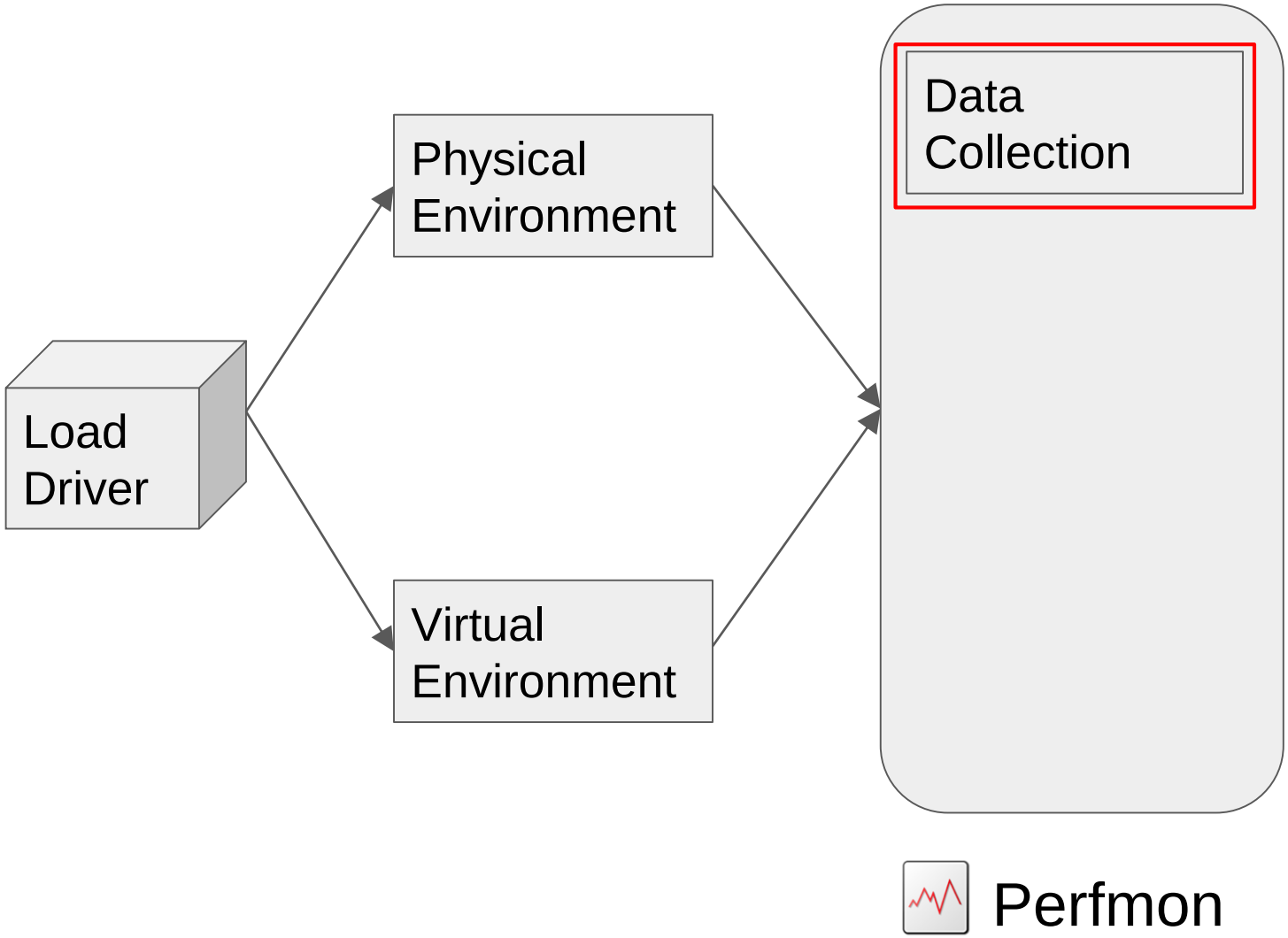
1.Dell DVD Store (DS2)
2. Cloudstore

Approach

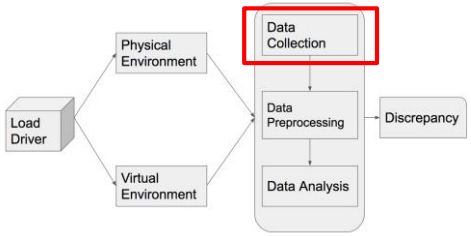


- We set up 3 nodes: Application, database and load driver.
- One virtual environment set up on each node: *single tenancy.*
- We set up identical config. between the two environments. i.e. CPU 2 cores, Memory 3GB.

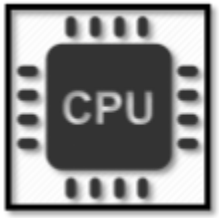
Approach


Data
Collection
Physical
Environment
Load
Driver
Virtual
Environment
Perfmon

Approach



- We used ***PERFMON*** to monitor application and database server.
- We recorded *all* the performance metrics available.

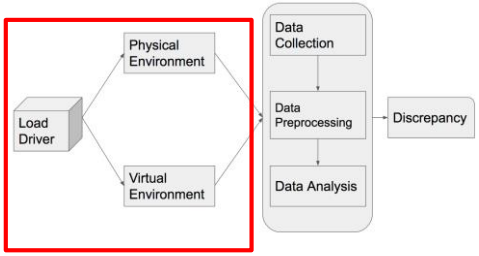


CPU

I/O Ops

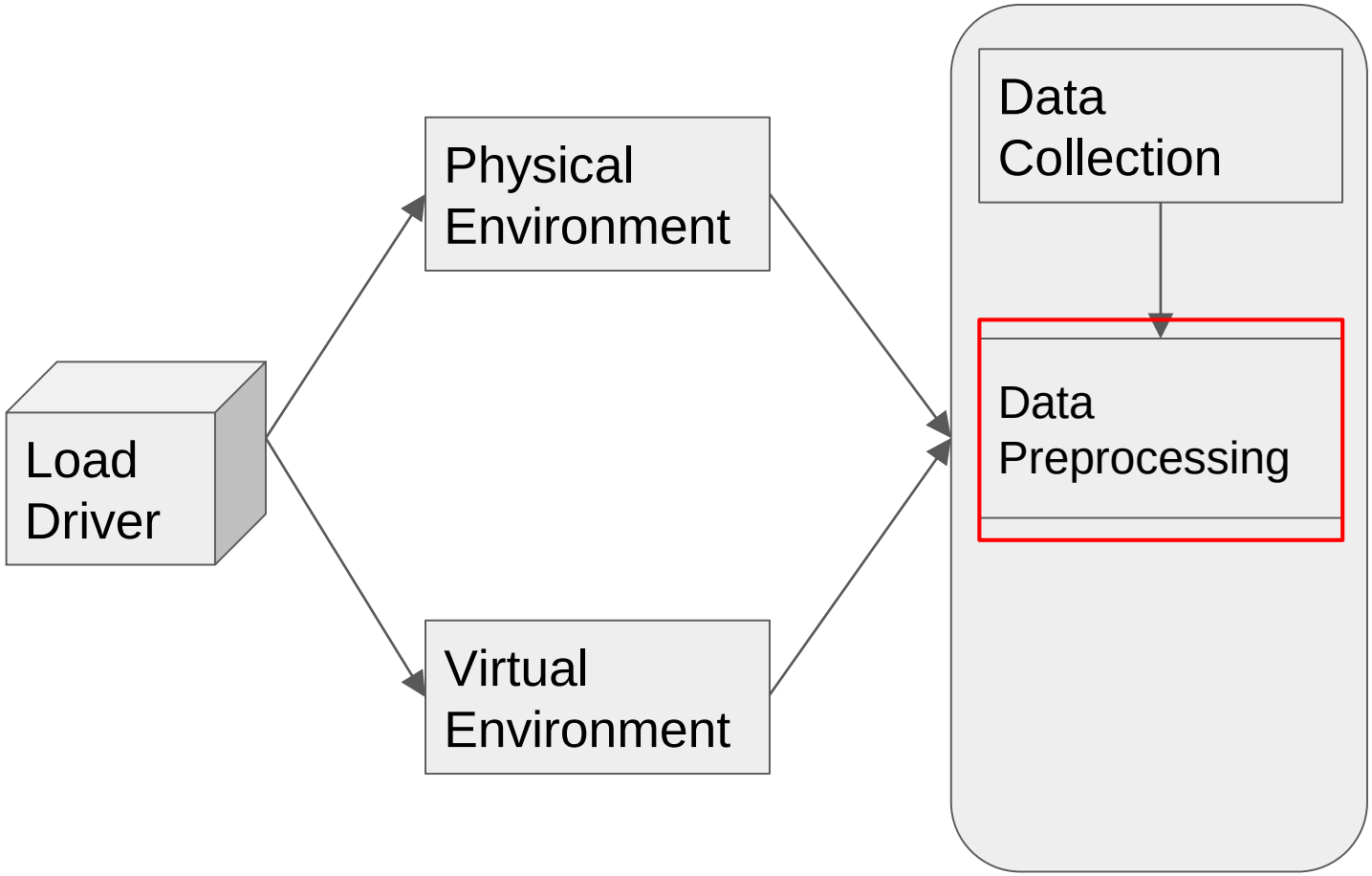
Memory

Approach

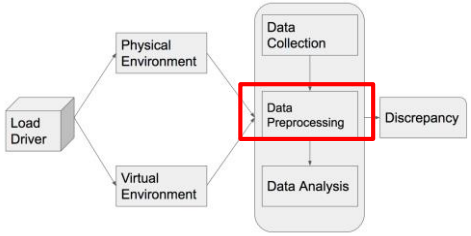


- *Workload* variation introduced by number of threads (threads=users) .
- The variation of *# of threads* was **identical**, **periodic** and **random** across the environments.
- Runtime for the test: 9+ hours.

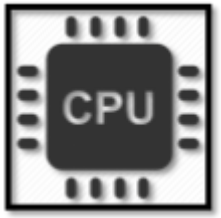
Approach


Load
Driver
Physical
Environment
Virtual
Environment
Data
Collection
Data
Preprocessing

Approach



System Level



CPU

I/O Ops

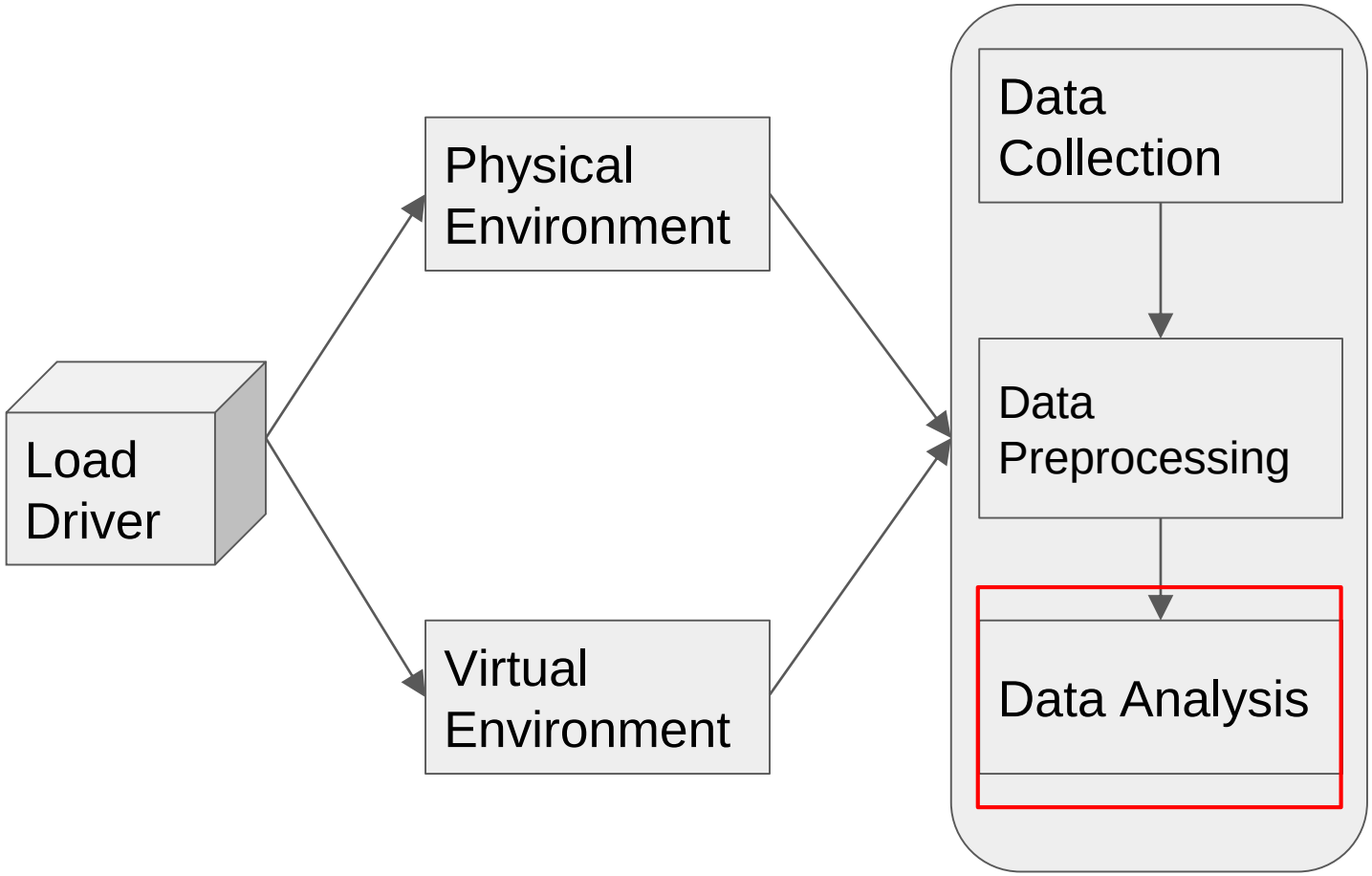
Memory

Application Level

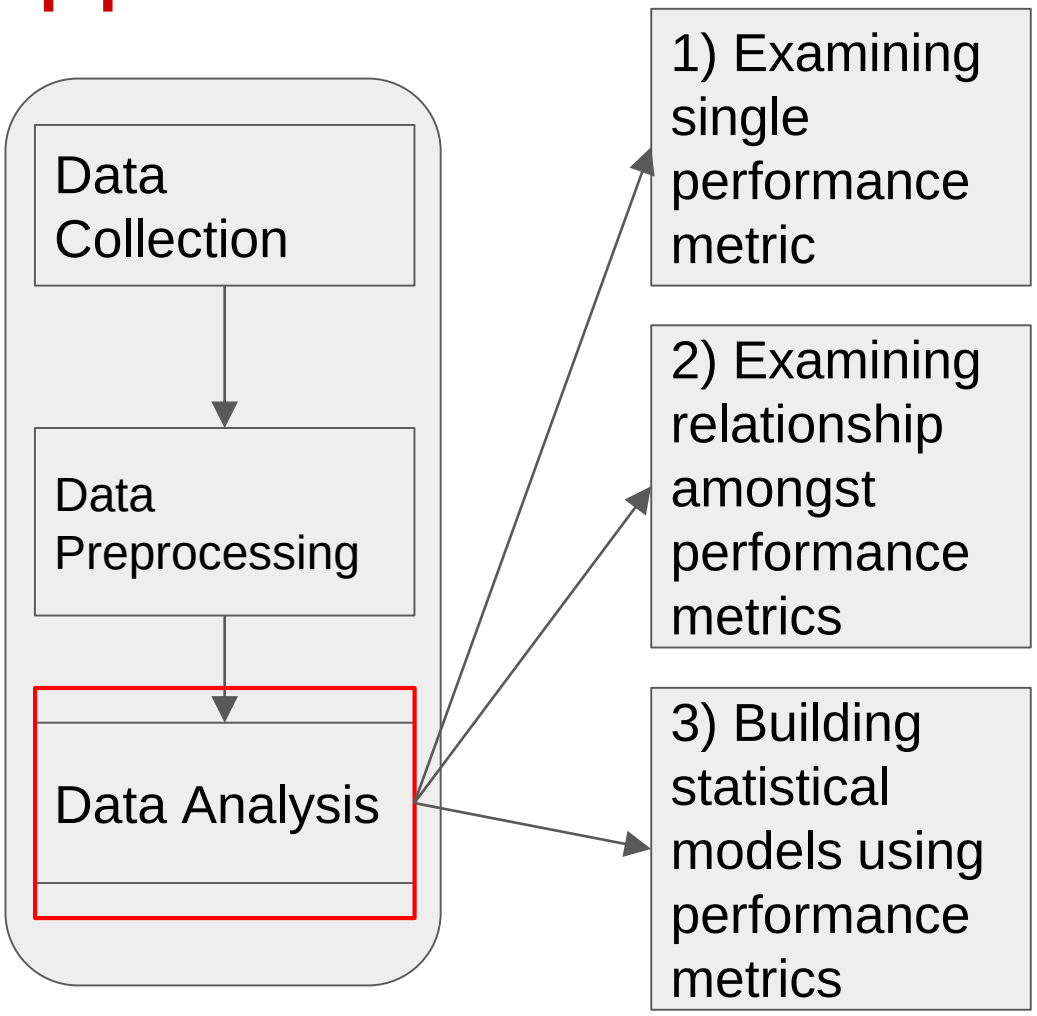
Throughput

Timestamps on logs used to calculate # of request/minute.

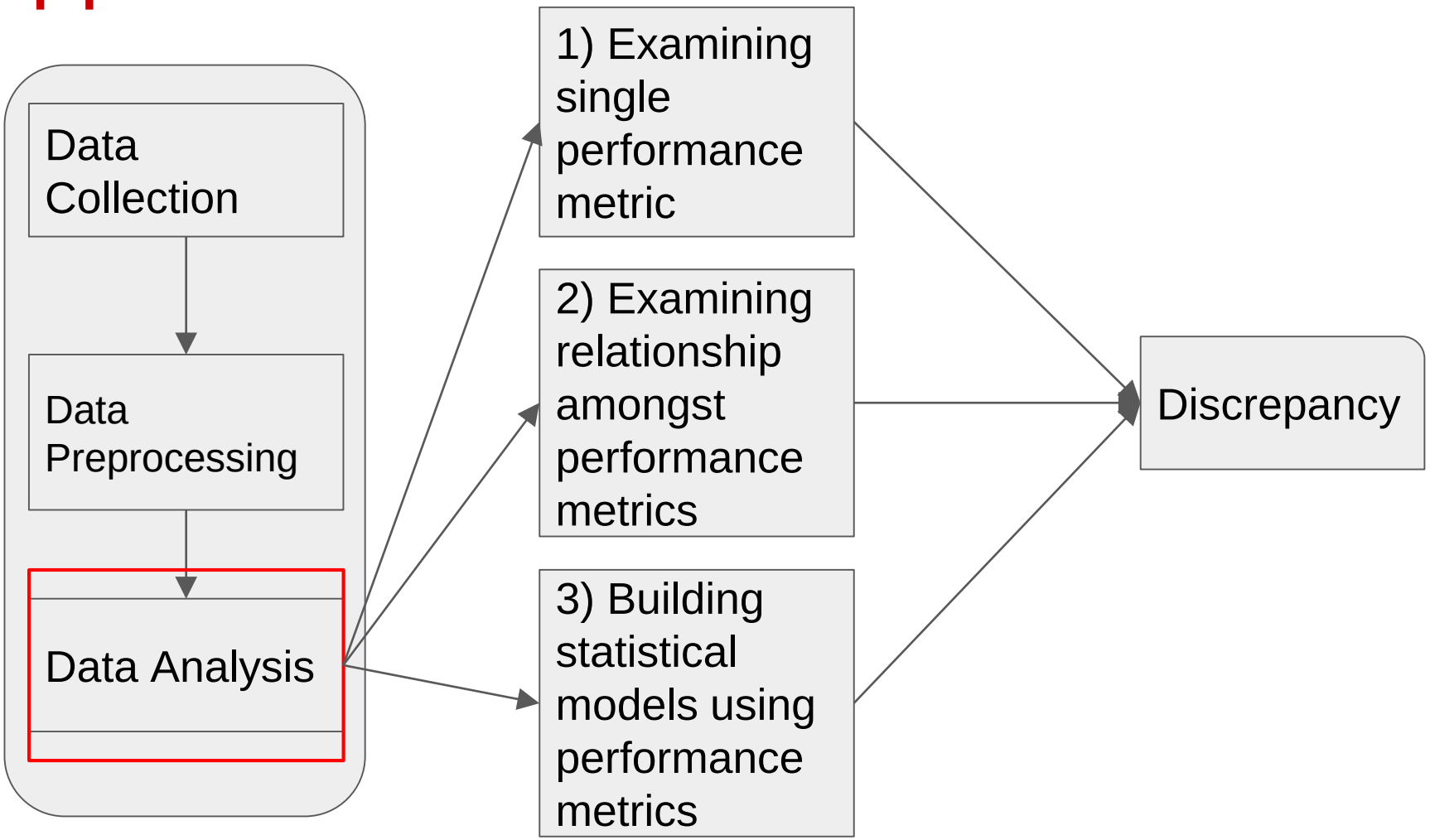
Approach


Load Driver
Physical Environment
Virtual Environment
Data Collection
Data Preprocessing
Data Analysis

Approach


Data Collection
Data Preprocessing
Data Analysis
1) Examining single performance metric
2) Examining relationship amongst performance metrics
3) Building statistical models using performance metrics

Approach


Data Collection
Data Preprocessing
Data Analysis
1) Examining single performance metric
2) Examining relationship amongst performance metrics
3) Building statistical models using performance metrics
Discrepancy

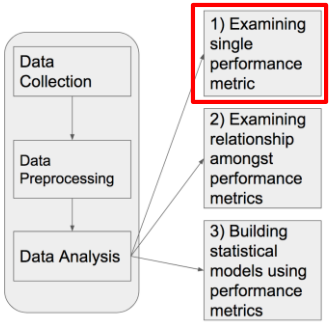
Research Questions

RQ1: Are the trend and distribution of a single performance metric similar across environments?

RQ2: To what extent does the relationship between the performance metrics change across environments?

RQ3: Can statistical performance models be applied across virtual and physical environments?

RQ1: Approach



Shape of the distribution

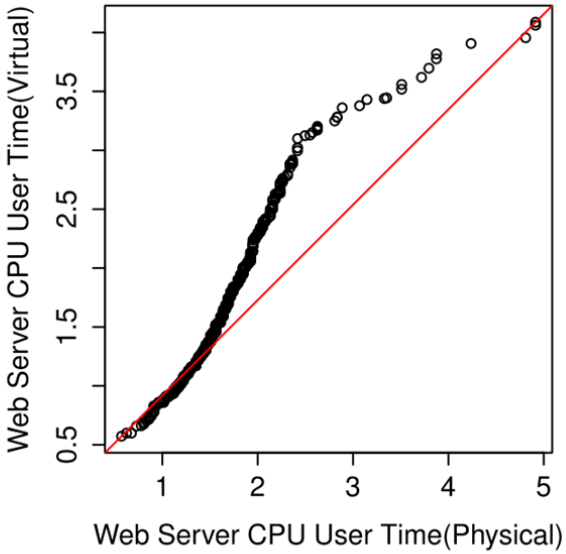
Trend

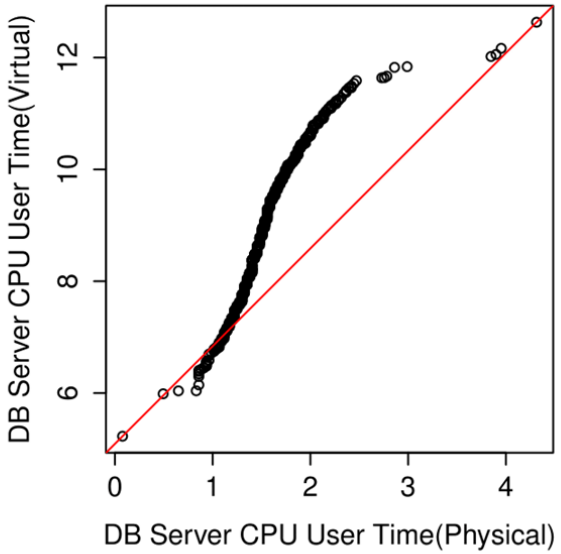
Quantile-Quantile (QQ) Plots

Normalized KS-Test

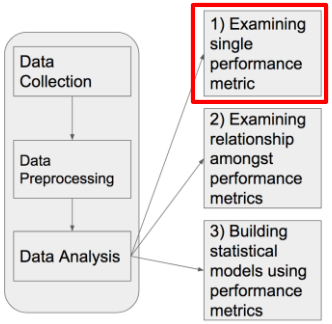
Spearman's rank correlation coefficient

DS2: Most performance metrics do not follow the same shape of the distribution in virtual and physical environments.



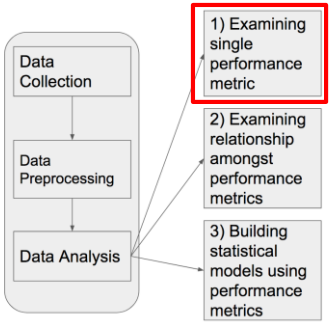


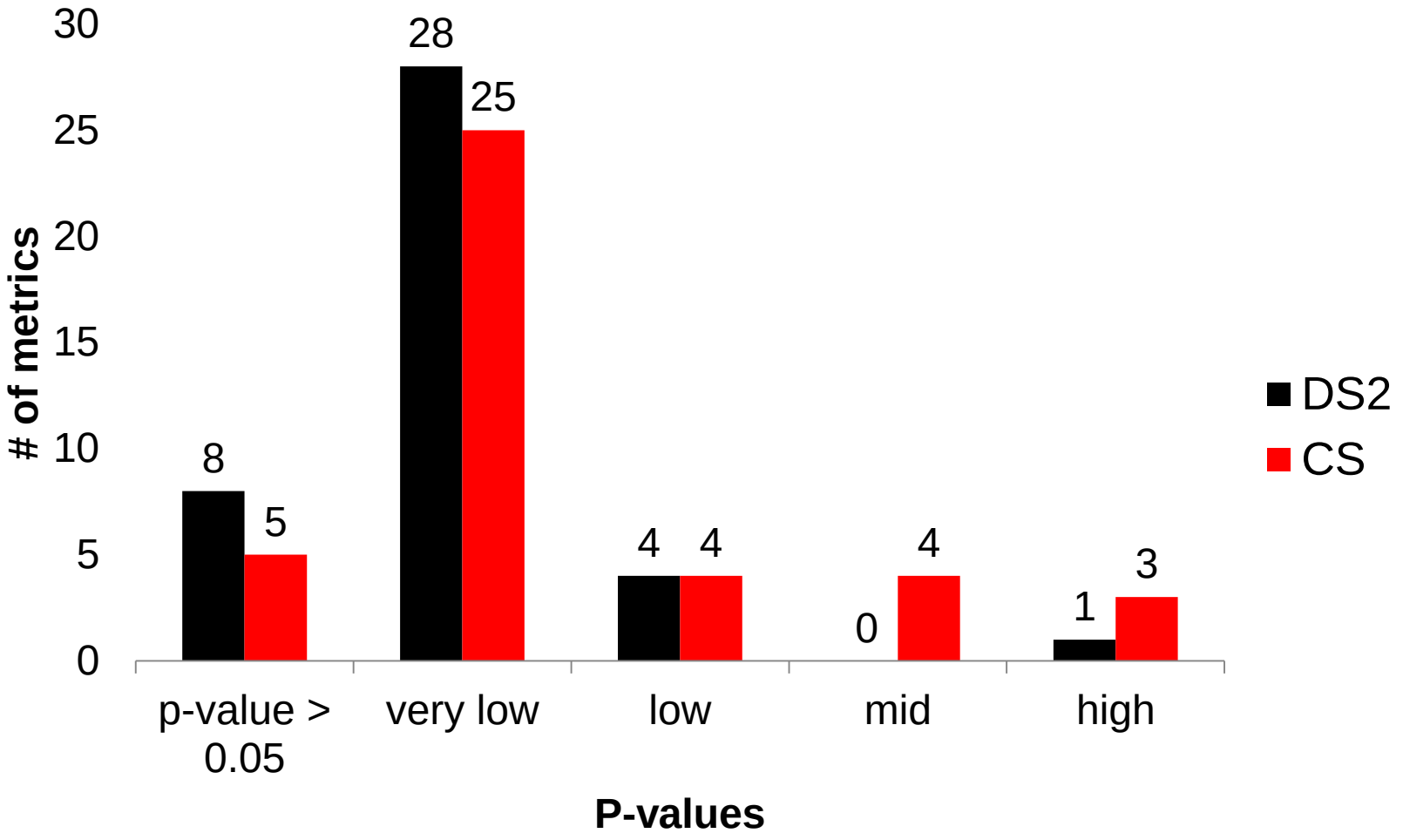
KS-Test: Performance metrics that do not follow the same distribution.



KS-Test (P-value > 0.05)	
DS2	CloudStore
13	12

Spearman: Most performance metrics do not have the same trend in virtual and physical environments.





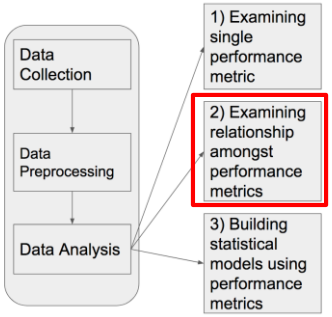
Research Questions

RQ1: Are the trend and distribution of a single performance metric similar across environments?

RQ2: To what extent does the relationship between the performance metrics change across environments?

RQ3: Can statistical performance models be applied across virtual and physical environments?

RQ2: Approach

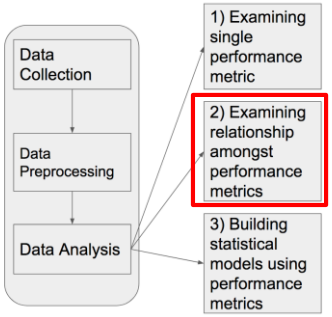


- Spearman's rank correlation coefficient:

  a. We calculate against throughput.

  b. We calculate the absolute difference for each pair of metrics. (represented by heatmaps)

Spearman: The rank of the metrics changes across the environments.

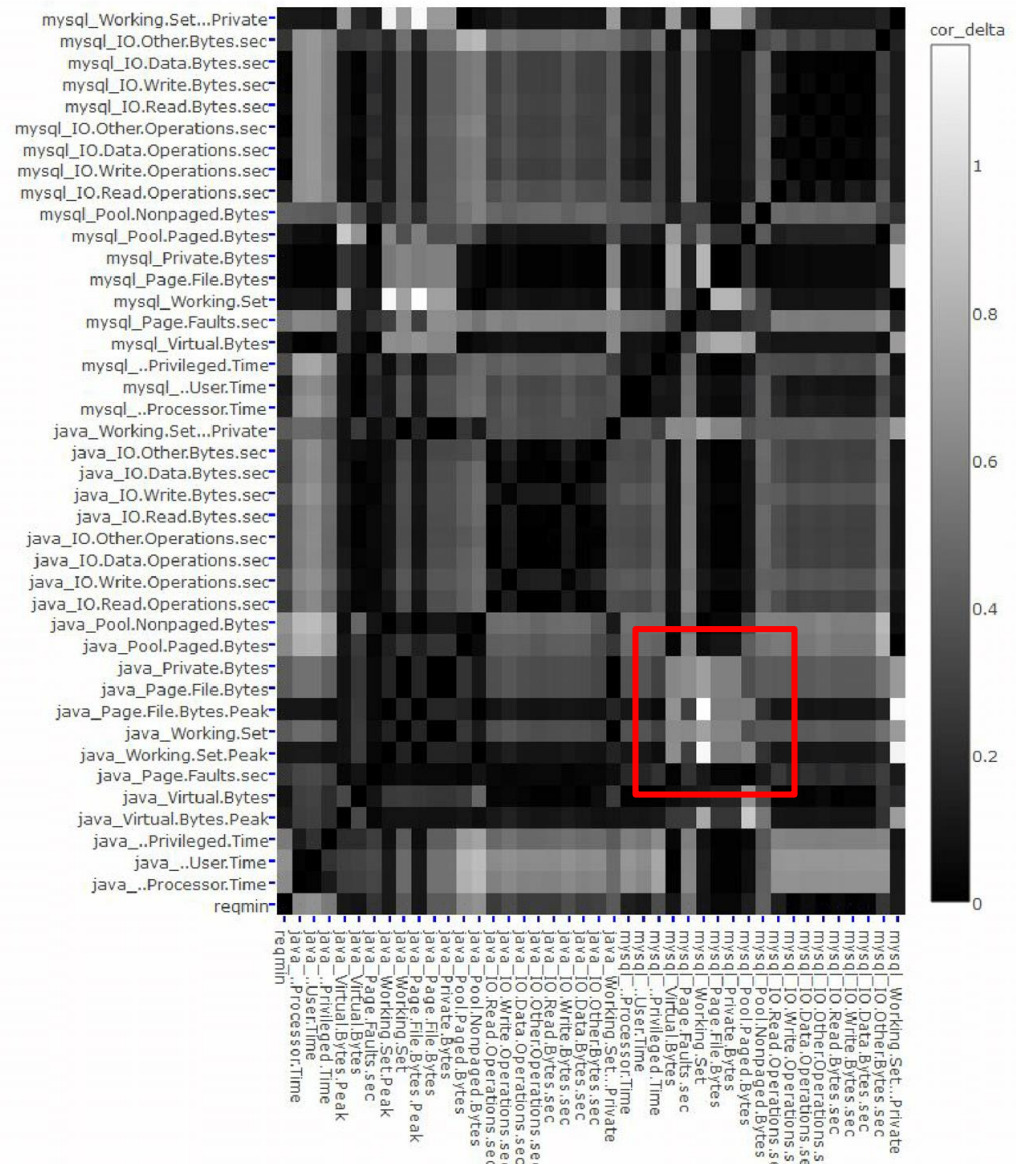


The *rank (vs. throughput)* of the metrics is not the same in the two environments.

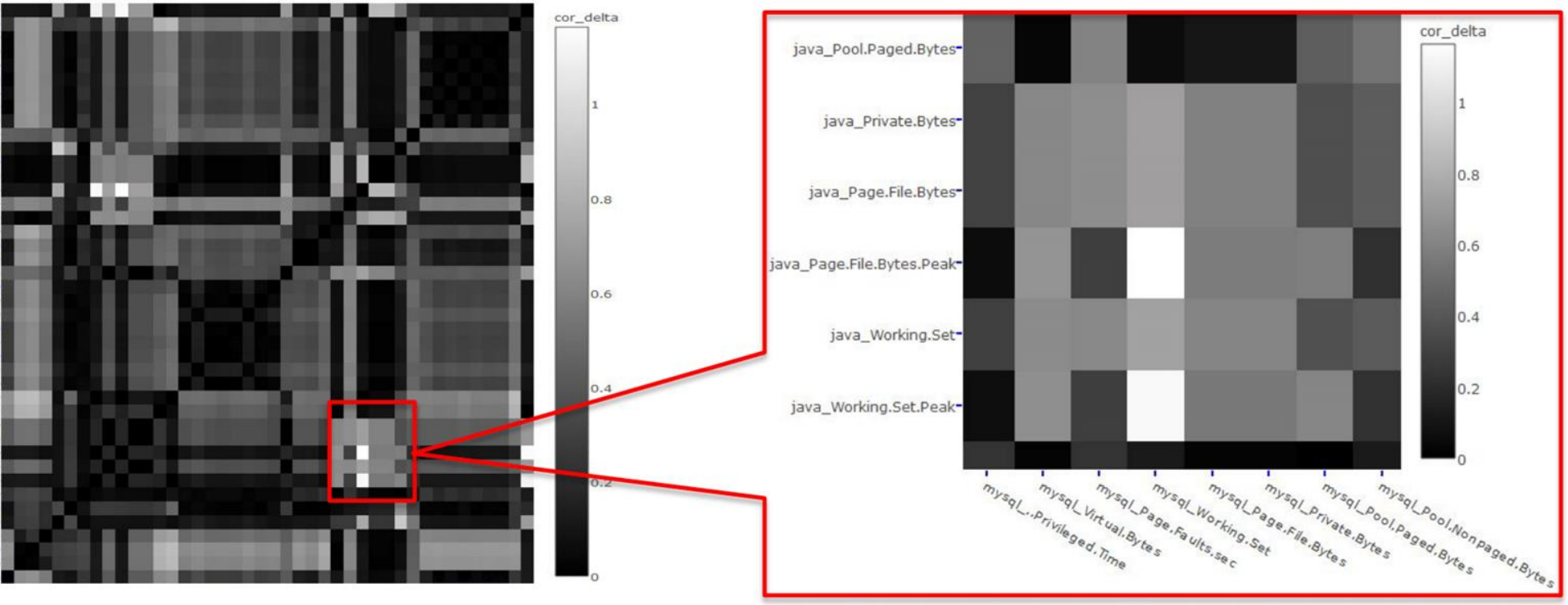
DB I/O Other Bytes/sec #9 → #40

DB I/O Other Bytes/sec #1 → #10

CS: Heatmap


mysql_Working.Set...Private
mysql_IO.Other.Bytes.sec
mysql_IO.Data.Bytes.sec
mysql_IO.Write.Bytes.sec
mysql_IO.Read.Bytes.sec
mysql_IO.Other.Operations.sec
mysql_IO.Data.Operations.sec
mysql_IO.Write.Operations.sec
mysql_IO.Read.Operations.sec
mysql_Pool.Nonpaged.Bytes
mysql_Pool.Paged.Bytes
mysql_Private.Bytes
mysql_Page.File.Bytes
mysql_Working.Set
mysql_Page.Faults.sec
mysql_Virtual.Bytes
mysql_..Privileged.Time
mysql_..User.Time
mysql_..Processor.Time
java_Working.Set...Private
java_IO.Other.Bytes.sec
java_IO.Data.Bytes.sec
java_IO.Write.Bytes.sec
java_IO.Read.Bytes.sec
java_IO.Other.Operations.sec
java_IO.Data.Operations.sec
java_IO.Write.Operations.sec
java_IO.Read.Operations.sec
java_Pool.Nonpaged.Bytes
java_Pool.Paged.Bytes
java_Private.Bytes
java_Page.File.Bytes
java_Page.File.Bytes.Peak
java_Working.Set
java_Working.Set.Peak
java_Page.Faults.sec
java_Virtual.Bytes
java_Virtual.Bytes.Peak
java_..Privileged.Time
java_..User.Time
java_..Processor.Time
reqmin
cor_delta
1
0.8
0.6
0.4
0.2
0

CS: There exist differences in correlation among the performance metrics from virtual and physical environments.



[Kraft, S. et al, SIGSOFT,2011. Menon, A. et al. ACM ICVEE, 2005]

Research Questions

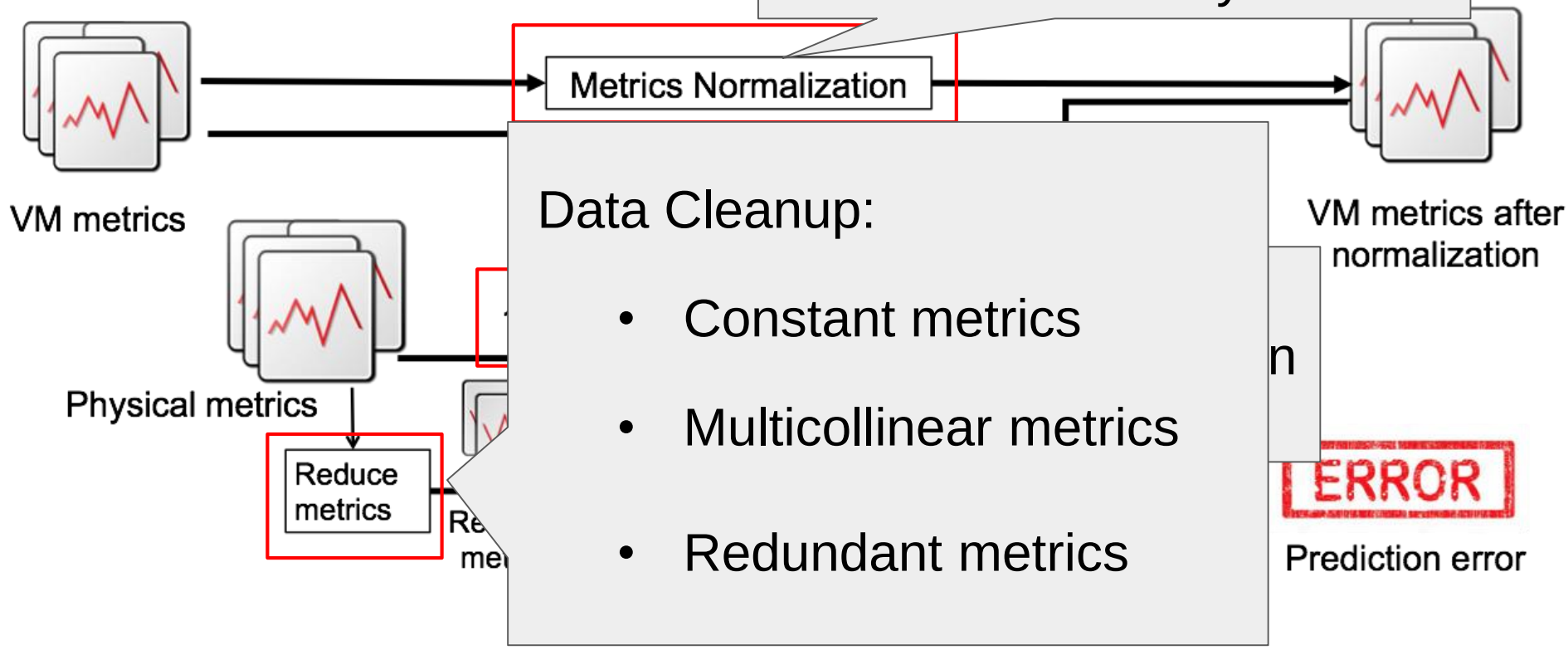
RQ1: Are the trend and distribution of a single performance metric similar across environments?

RQ2: To what extent does the relationship between the performance metrics change across environments?

RQ3: Can statistical performance models be applied across virtual and physical environments?

RQ3: Approach - ng

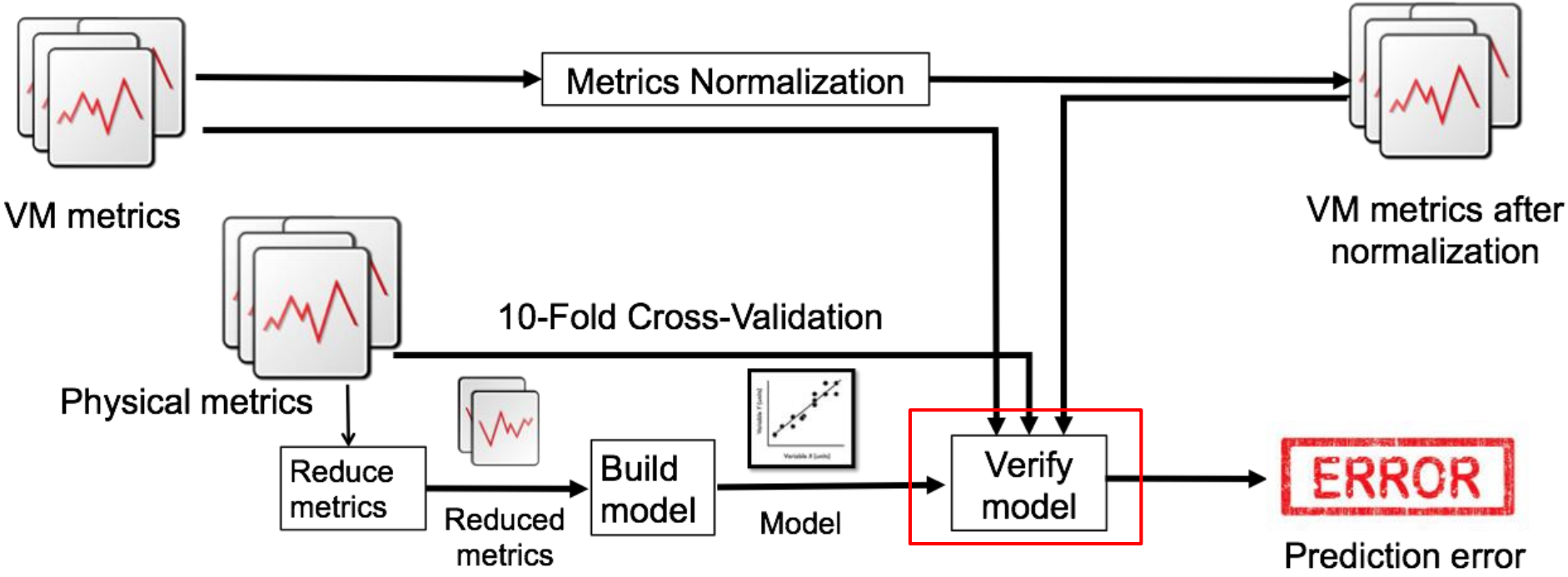
- Normalization by deviance
- Normalization by load



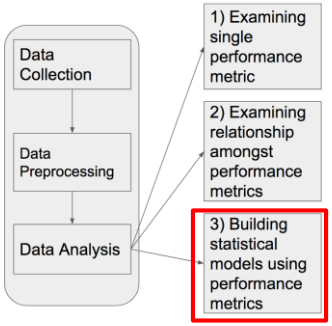
Data Cleanup:

- Constant metrics
- Multicollinear metrics
- Redundant metrics

RQ3: Approach - Statistical modelling



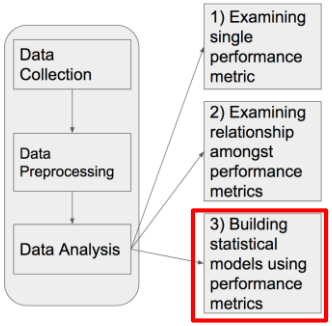
RQ3: Model Verification



Two types of model verification:

1. Internal Validation
2. External Validation after:
   - We normalize by deviance
   - We normalize by load

RQ3: Model Verification

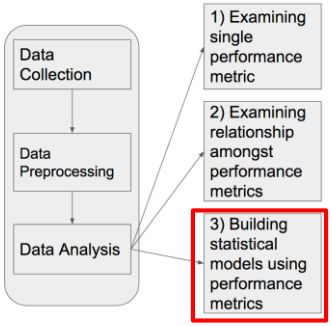


External Validation

- Normalization by deviance

$$M_{normalized} = \frac{M - \tilde{M}}{MAD(M))}$$

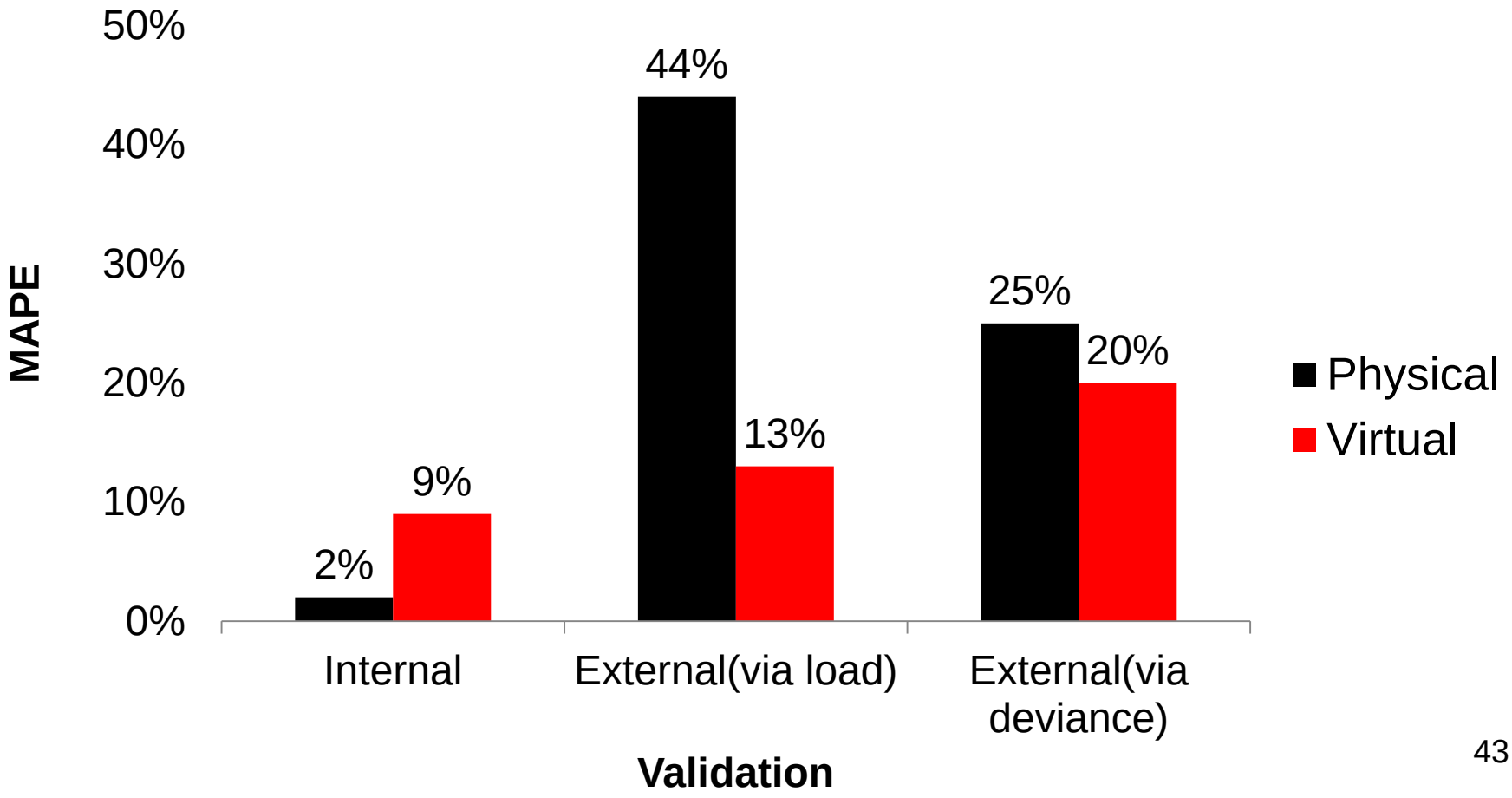
RQ3: Model Verification



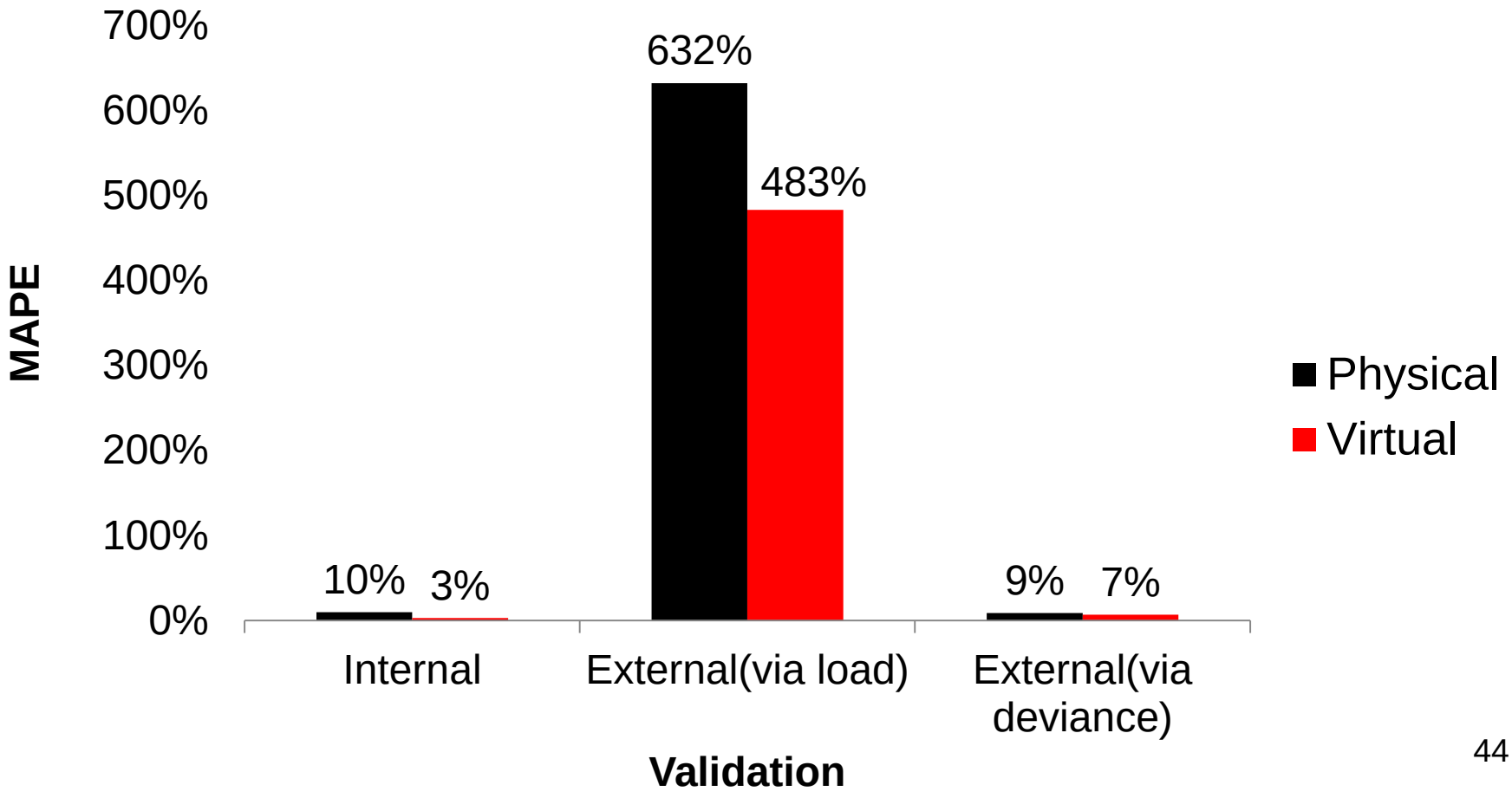
External Validation

- Normalization by deviance
- Followed by a min-max normalization to avoid calculation errors for negative values.

Discrepancy - DS2: We find that the statistical models built by performance testing results in an environment cannot advocate for the other environment due to discrepancies present.



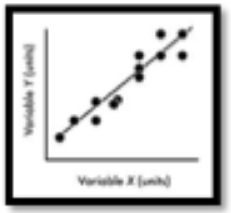
Discrepancy - CS: We find that the statistical models built by performance testing results in an environment cannot advocate for the other environment due to discrepancies present.



Examining the impact from other factors on our results

1. Instability of the virtual environment
2. Impact of the specific virtual machine software
3. Impact of allocated resources

1. Instability of the virtual environment

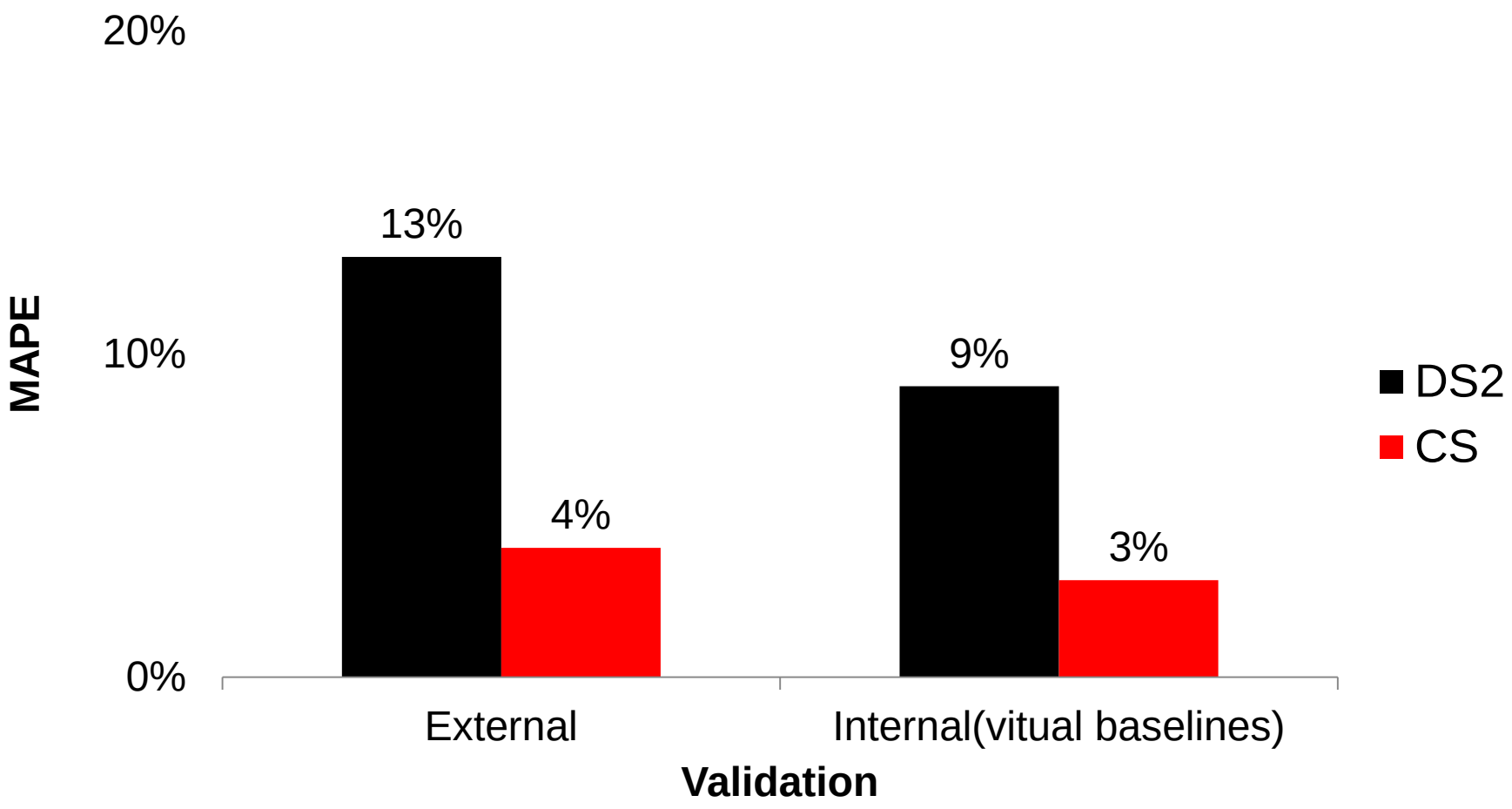


Repeat tests

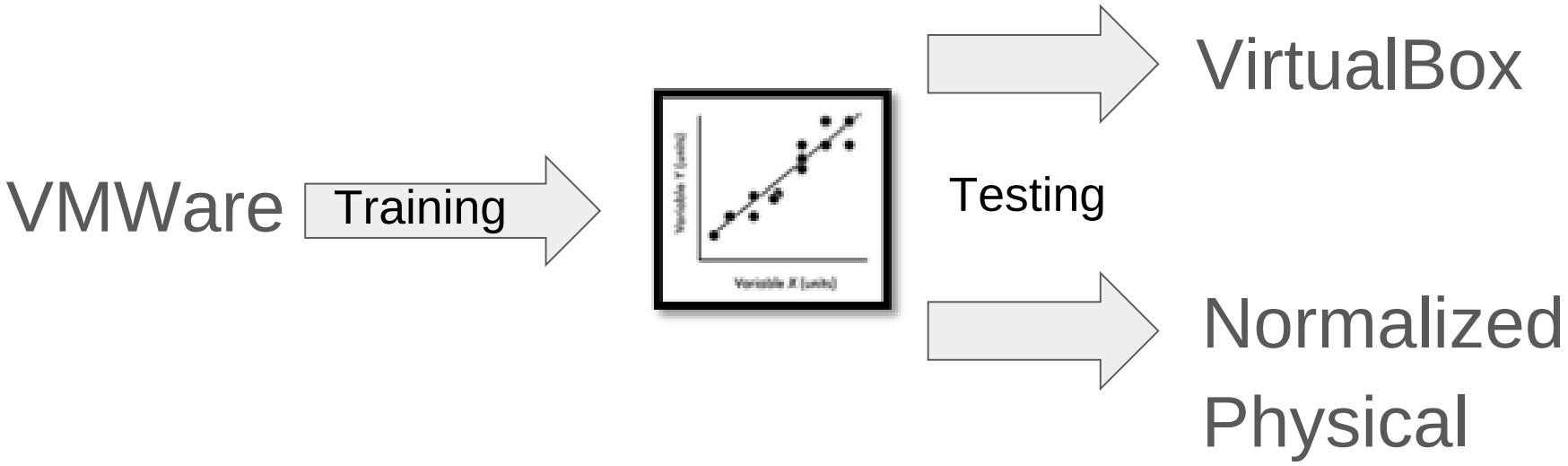
Build models

Compare externally

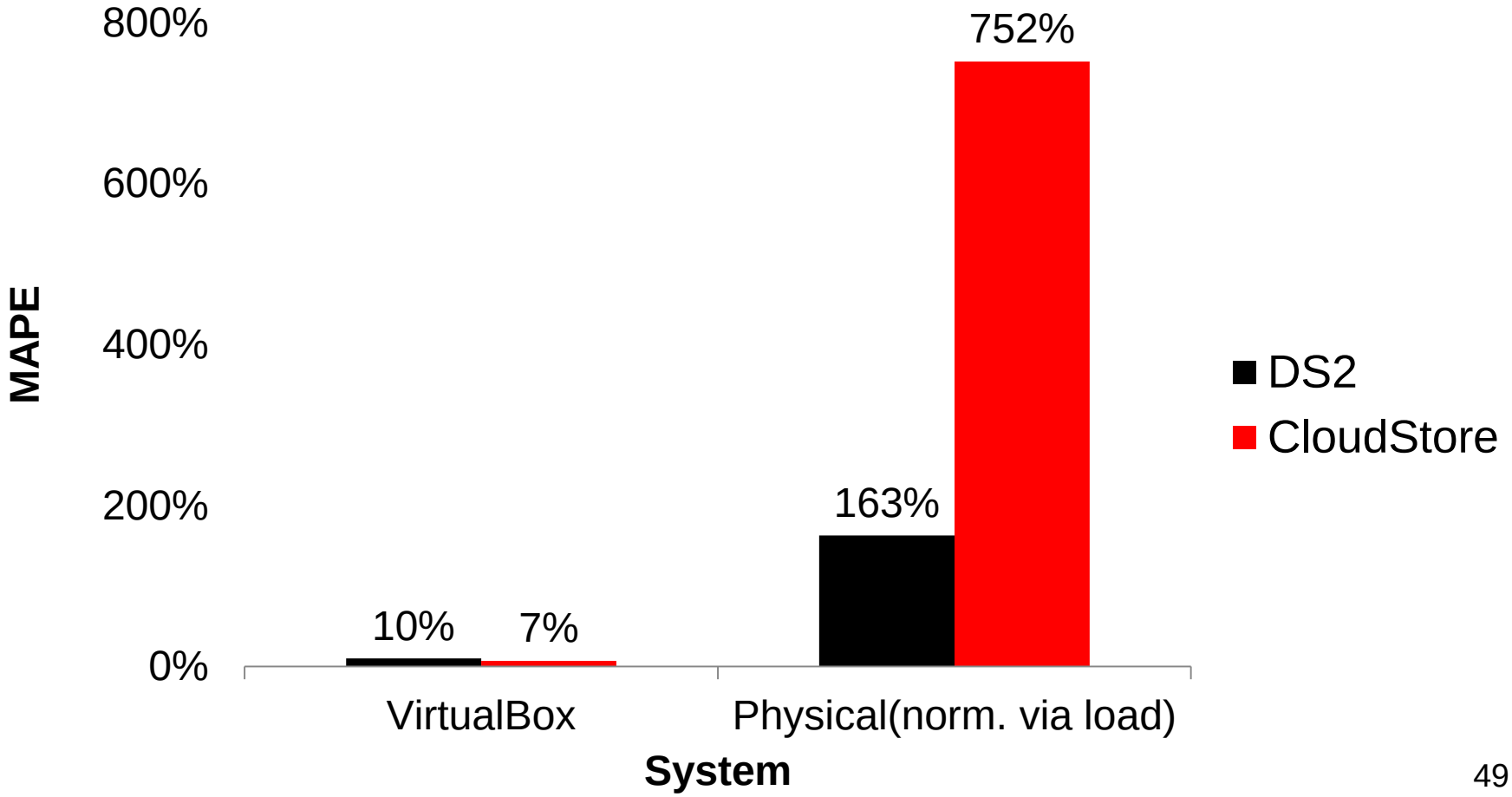
1. Performance testing results from the virtual environments are stable.



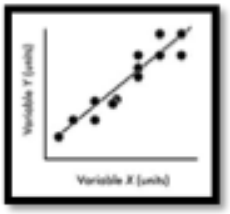
2. Impact of the specific virtual machine software



2. Discrepancy observed during our experiment also exists with the virtual environments that are set up with VMWare.



3. Impact of allocated resources



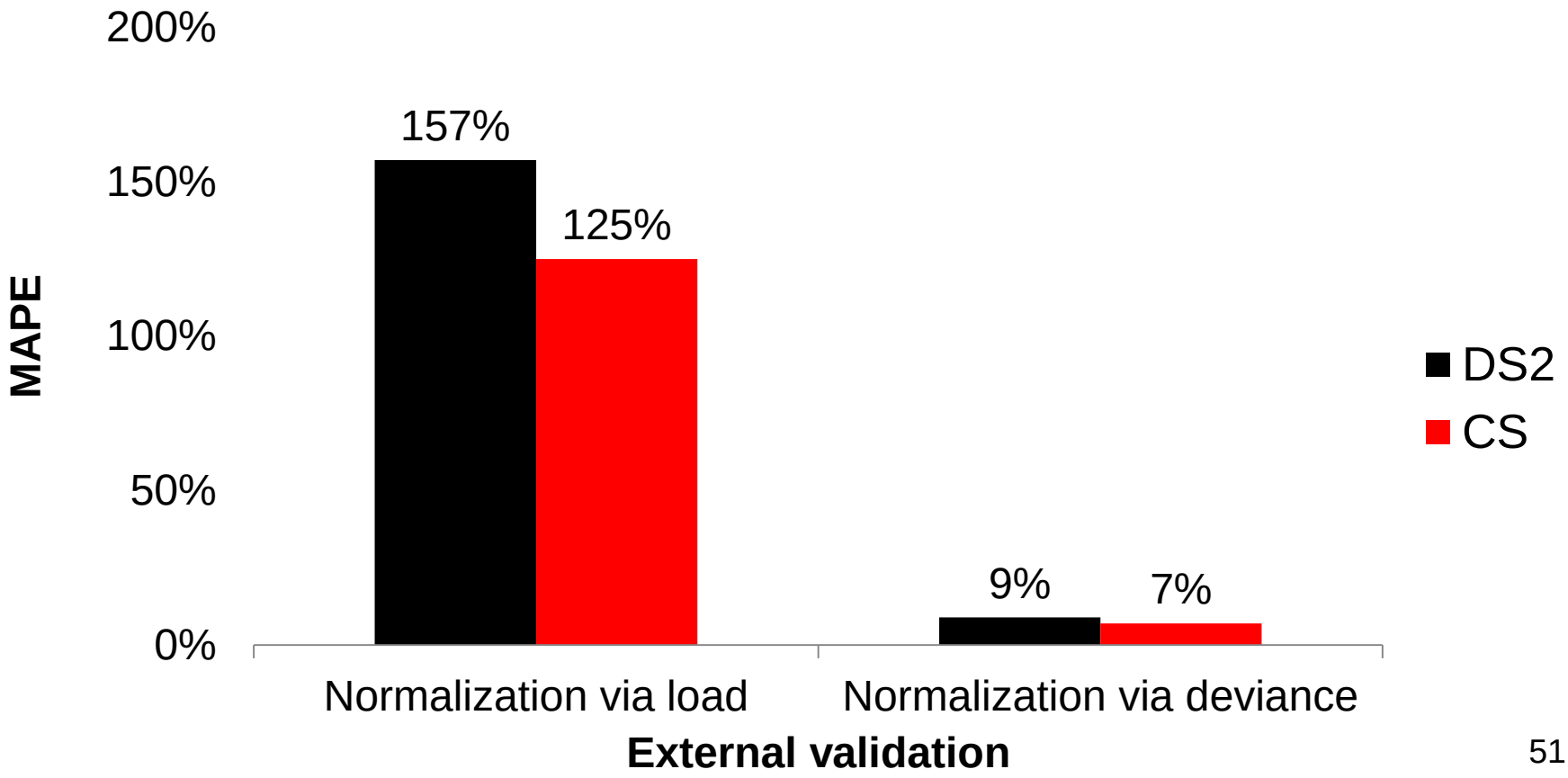
Modify configuration

Build models

Compare externally

We increase the resources to 3 cores and 5GB memory.

3. Our findings still hold when the allocated resources are changed and this change has minimal impact on the results of our case studies.



Contributions

This is the ***first research attempt*** to evaluate the ***discrepancy*** between performance testing results in virtual and physical environments.

We identified the ***performance metrics that contribute the most to the discrepancies***.

We find that **normalizing performance metrics based on deviance** may ***reduce*** the discrepancy

Implications

Practitioners ***cannot assume a straight forward overhead***.

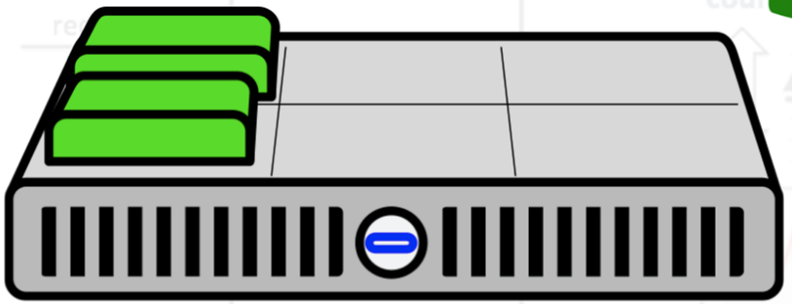
Practitioners should always verify whether the ***inconsistency of correlations between performance metrics*** are due to virtual environments.

Normalization by deviance may minimize such discrepancy.

Summary

How is a performance regression detected?

What if you need to modify your environment?

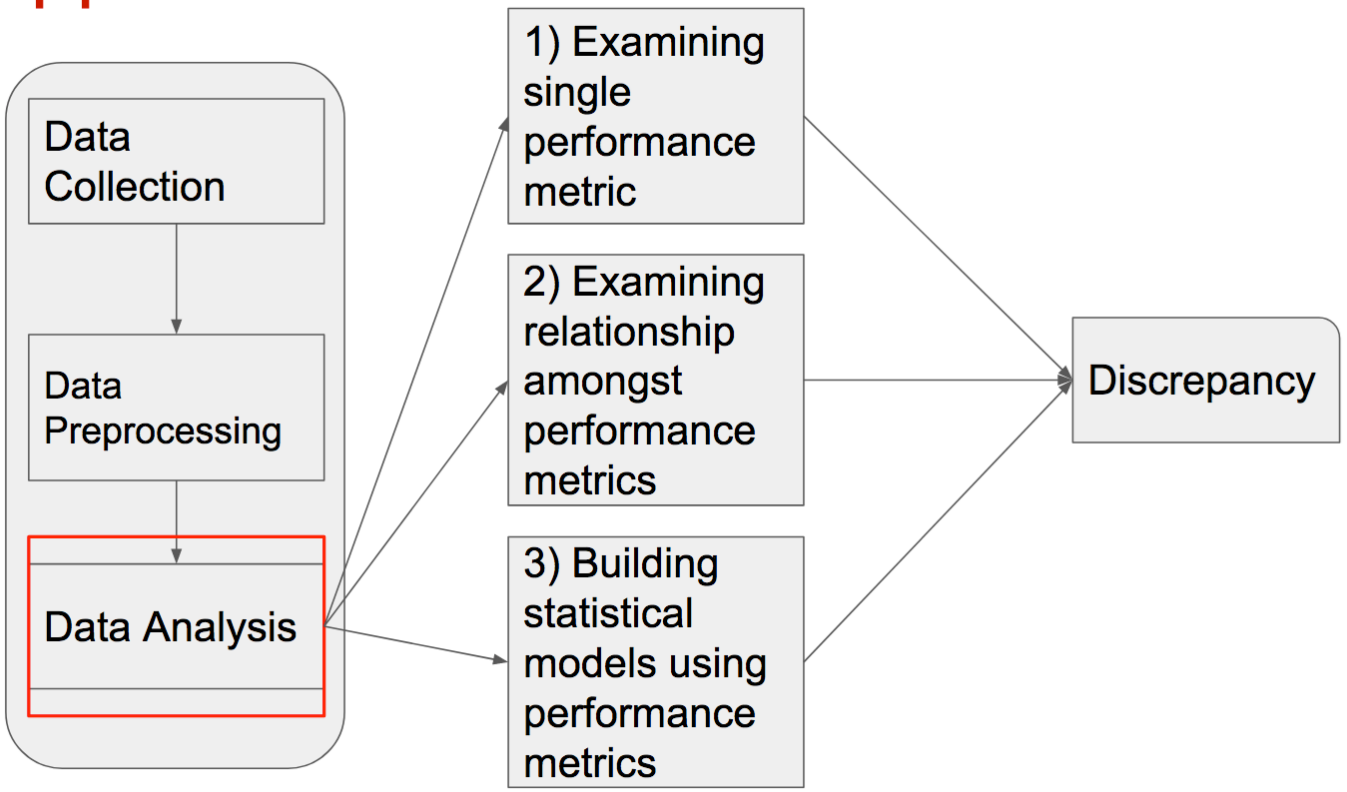


VIRTUAL MACHINES!

Image:
https://openclipart.org/detail/191766/question-guy
https://openclipart.org/detail/167198/two-virtual-machines

Summary

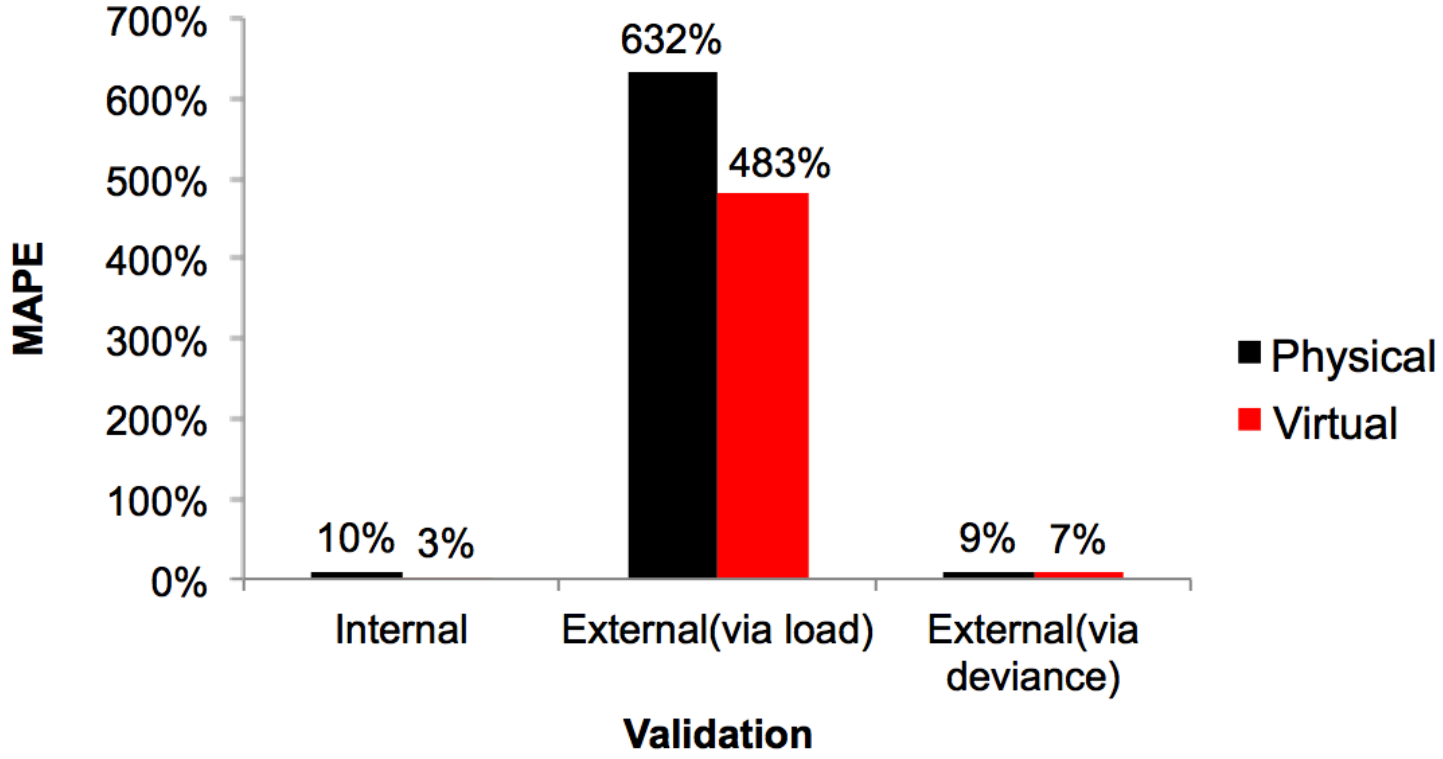
Approach



25

Summary

Discrepancy - CS: We find that the statistical models built by performance testing results in an environment cannot advocate for the other environment due to discrepancies present.



Summary

Contributions

- This is the ***first research attempt*** to evaluate the ***discrepancy*** between performance testing results in virtual and physical environments.
- We identified the ***performance metrics that contribute the most to the discrepancies***.
- We find that **normalizing performance metrics based on deviance** may ***reduce*** the discrepancy

51

Summary

How is a performance regression detected?

What if you need to modify your environment?

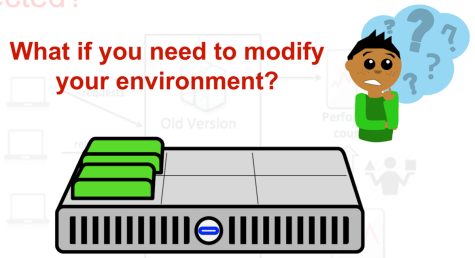
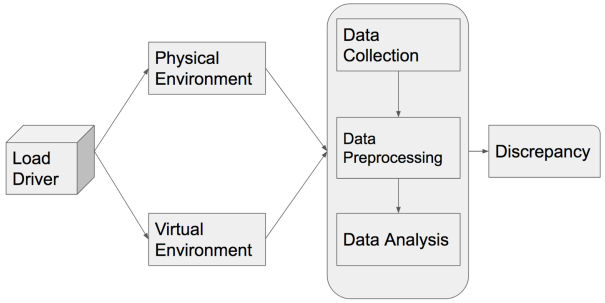


Image:
https://openclipart.org/detail/191766/question-guy
https://openclipart.org/detail/167198/two-virtual-machines

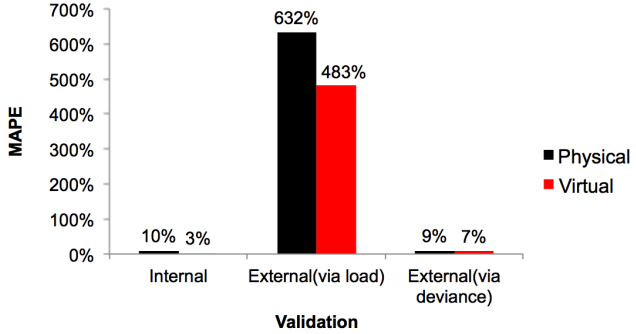
6

Approach



14

Discrepancy - CS: We find that the statistical models built by performance testing results in an environment cannot advocate for the other environment due to discrepancies present.



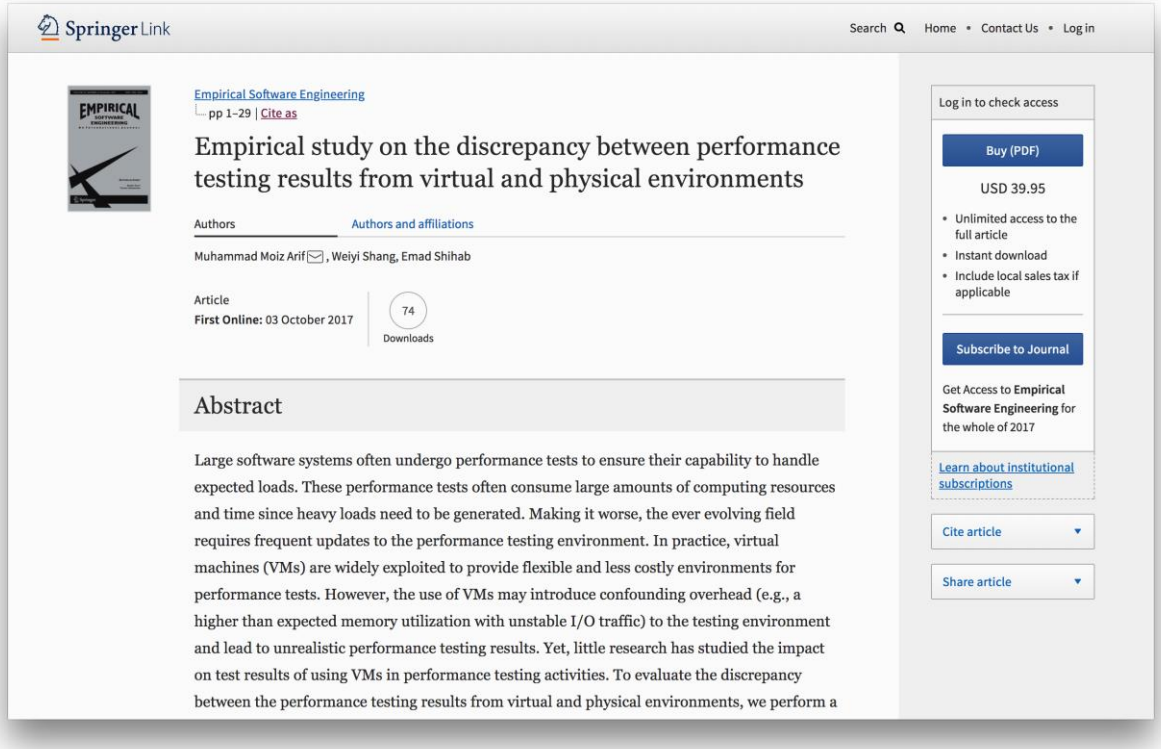
Contributions

- This is the ***first research attempt*** to evaluate the ***discrepancy*** between performance testing results in virtual and physical environments.
- We identified the ***performance metrics that contribute the most to the discrepancies***.
- We find that **normalizing performance metrics based on deviance** may ***reduce*** the discrepancy

51

Publications

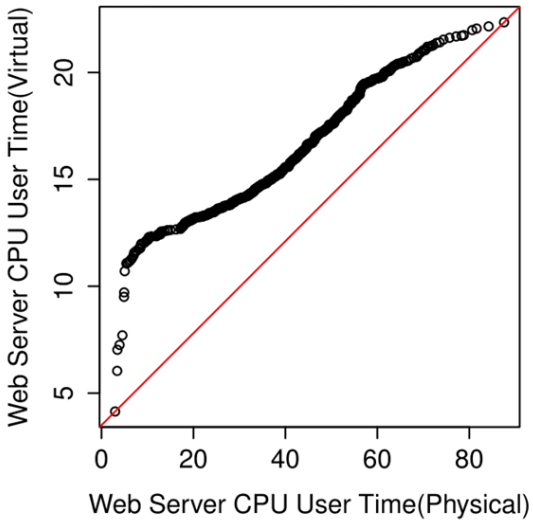
Arif, M.M., Shang, W. & Shihab, E. Empir Software Eng (2017)

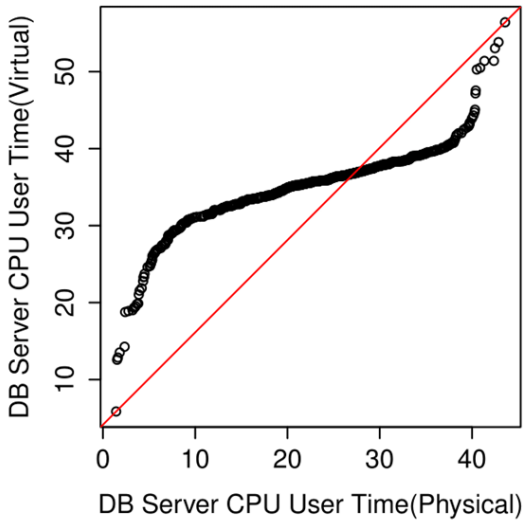


https://users.encs.concordia.ca/~shang/pubs/emse_moiz.pdf

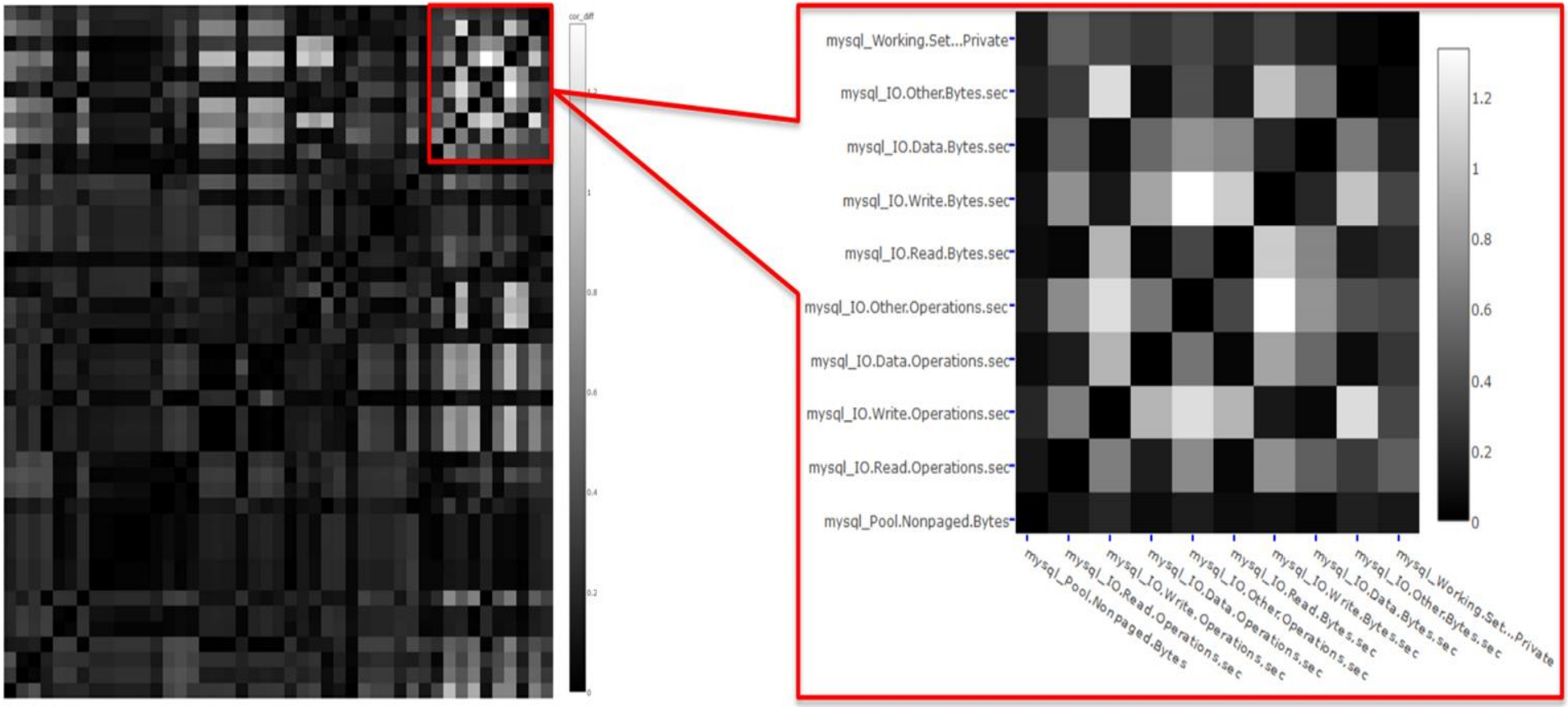
m.moizarif@gmail.com

CS: Most performance metrics do not follow the same shape of the distribution in virtual and physical environments.

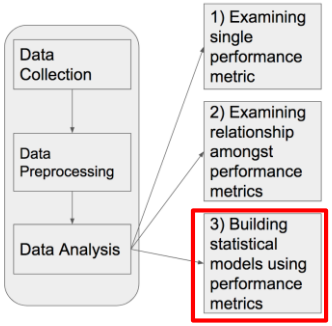




DS2: Heatmap


mysql_Working.Set...Private
mysql_IO.Other.Bytes.sec
mysql_IO.Data.Bytes.sec
mysql_IO.Write.Bytes.sec
mysql_IO.Read.Bytes.sec
mysql_IO.Other.Operations.sec
mysql_IO.Data.Operations.sec
mysql_IO.Write.Operations.sec
mysql_IO.Read.Operations.sec
mysql_Pool.Nonpaged.Bytes
1.2
1
0.8
0.6
0.4
0.2
0

Model Verification

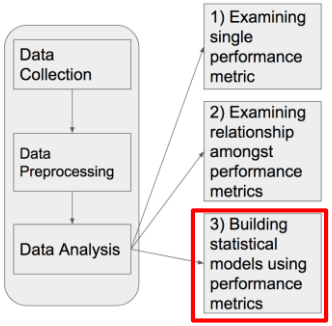


Internal Validation

Ten-fold Cross validation

k=1	Test	Train	Train	Train	Train	Train	Train	Train	Train	Train
k=2	Train	Test	Train	Train	Train	Train	Train	Train	Train	Train
k=3	Train	Train	Test	Train	Train	Train	Train	Train	Train	Train
...	...	...	...	...	...	...	...	...	...	...
k=10	Train	Train	Train	Train	Train	Train	Train	Train	Train	Test

Model Verification



External Validation

Normalization by load

$$throughput_p = \alpha_p \times M_p + \beta_p$$

$$throughput_v = \alpha_v \times M_v + \beta_v$$

$$M_{normalized} = \frac{(\alpha_v \times M_v) + \beta_v - \beta_p}{\alpha_p}$$

Discrepancy: Absolute % Error

Median Absolute Percentage Error (*MAPE*)

Throughput: 100 requests/minute

Predicted: 110 requests/minute

$$MAPE = (\frac{|110 - 100|}{100}) = 0.1$$

Limitations

More case studies

Quality of performance tests and recorded values.

Other types of workload variation.

Hint of overfitting of models.

Choice of intervals (for e.g. 9 hours, 10 seconds)

The complete system: user’s point of view.

Future Work

Reproducing known performance regressions in heterogeneous environments.

Replicating our experiments in cloud environments.

Designing automating techniques.