Taming Performance Variability

Aleksander Maricq* Dmitry Duplyakin* Ivo Jimenez†
Carlos Maltzahn† Ryan Stutsman* Robert Ricci*

* University of Utah
† University of California Santa Cruz

Motivation: Performance Variability

How confident should I be that my results are correct?

How many times do I need to run my experiments?

As a testbed builder, how can I help users figure this out?

11 months
~892,000 data points
835 servers

Memory
Disk
Network

Examine performance variability of testbed hardware

Within servers
Across servers

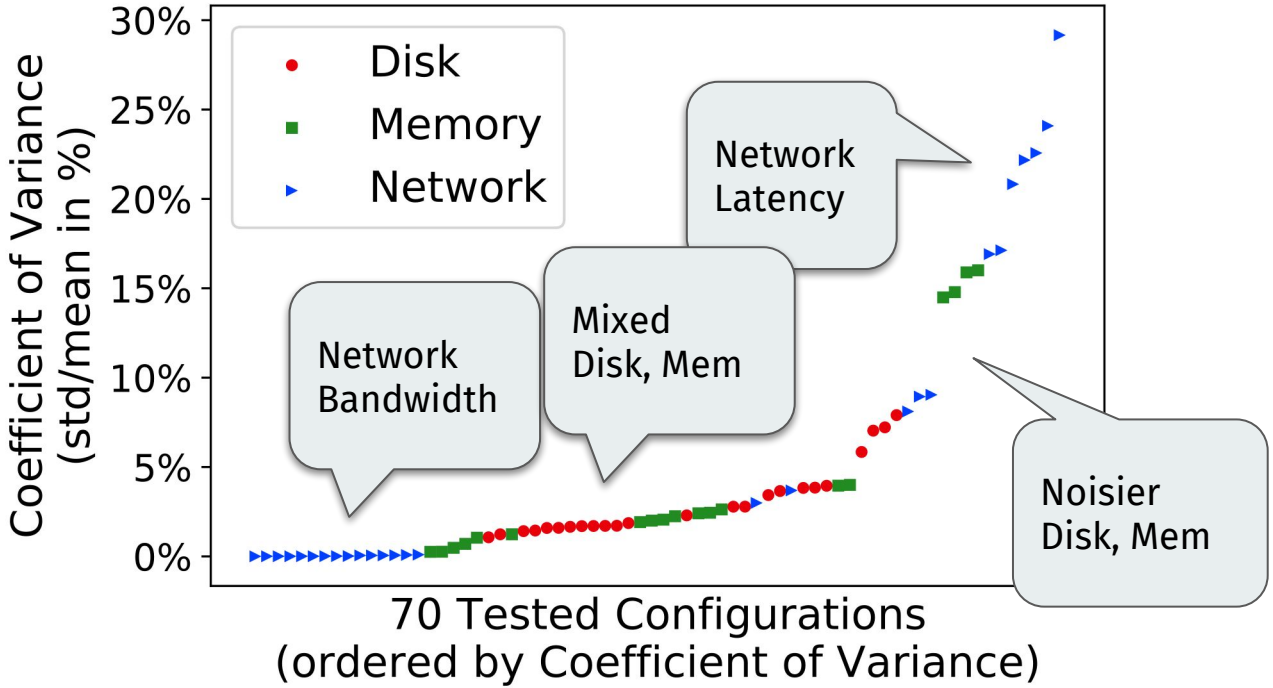
- 1,500 servers at three sites
  - Several distinct ‘types’ of identical servers
- Exclusive, raw access to hardware
  - No interference on servers from simultaneous users
  - Doesn’t add virtualization overhead / variability
- Our experiments were run on servers allocated only to us
- Configuration: Combination of hardware type, workload, parameters

c220g1, single-threaded mem copy, dvfs off

m510, net bw, rack-local

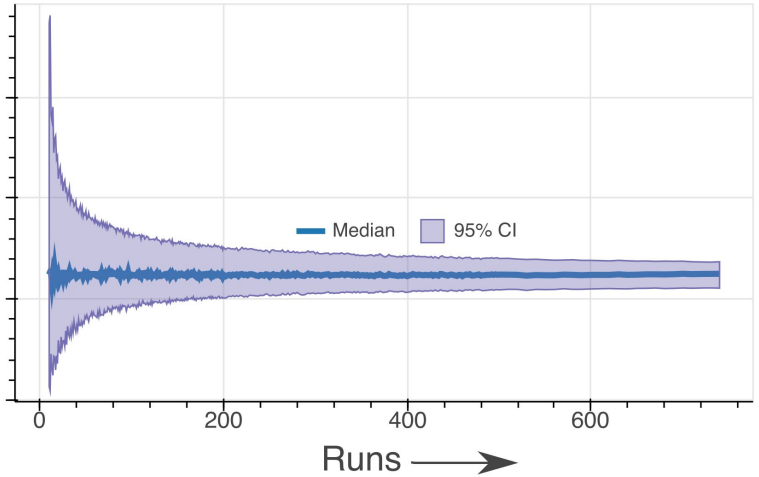
How confident can we be in the correctness of our results?

How much trouble are we in?



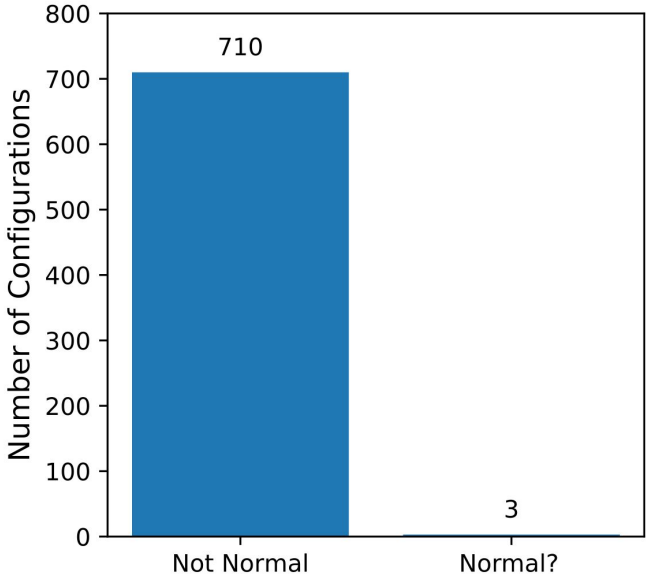
Confidence Intervals

- Range for your mean (different than stdev)
- Represents some % confidence (eg. 95%) the true mean lies between
- More runs -> narrower CI



Testing Normality

- Many statistical models assume normal (gaussian) bell-curve
- Is our data normal? Shapiro-Wilk test (95% confidence)



Use Non-Parametric Statistics to Avoid Assumptions of Normality

How confident can we be in the correctness of our results?

- Some variation is unavoidable
- Results are often non-normal
- More runs → more confidence

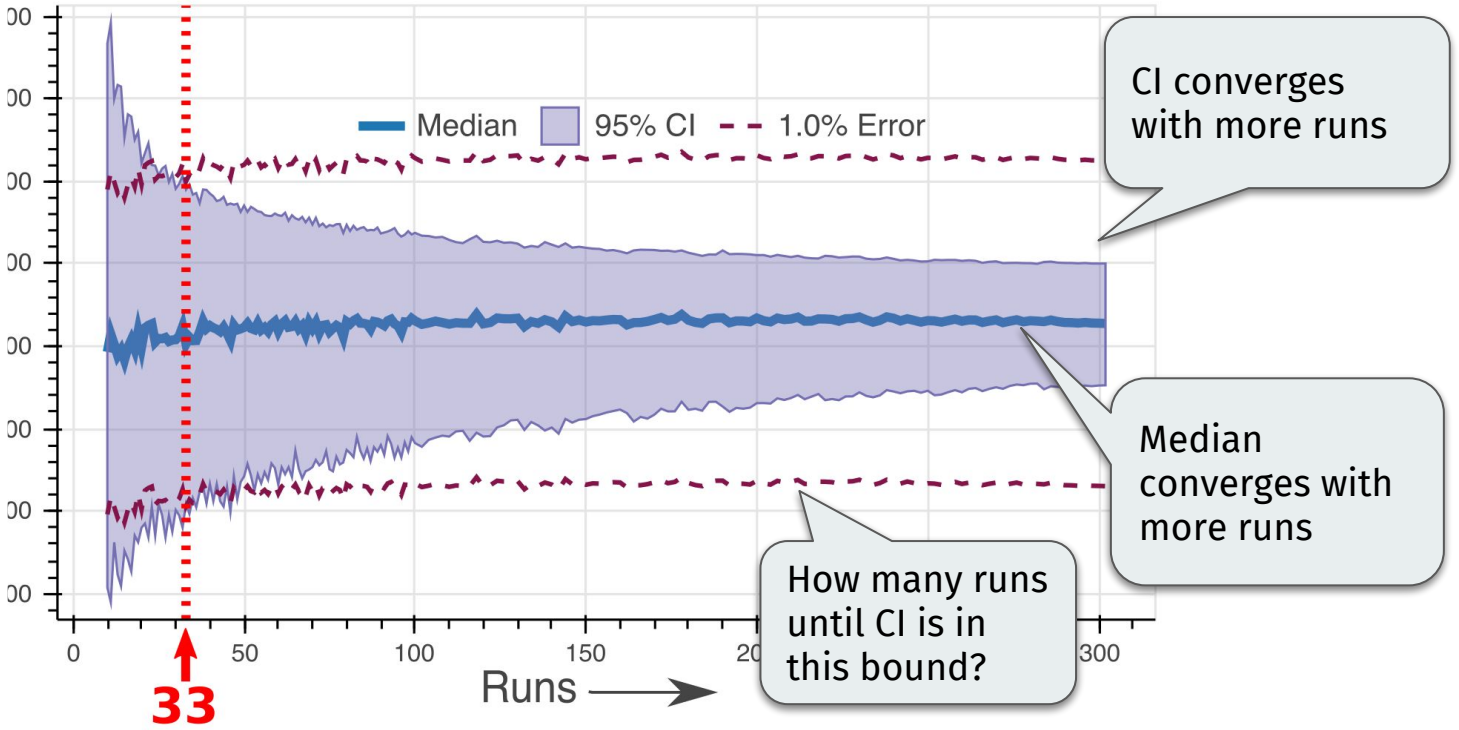
How many times
should we run our experiments?

CONFIRM - CONFIdence-based Repetition Meter

- Uses all our collected data to build *estimates* of how many runs are needed
  - For configurations on a single server or group of servers
- Uses random sub-samples of historical data
  - Takes many sub-samples, computes mean and CI
- Calculating observed empirical CIs still necessary
- Integrated into CloudLab

CONFIRM

From past data, uses random subsets to model median and CI behavior for increasing numbers of runs



CONFIRM Recommendations

	CoV	Recommended Runs
Mem Config A (c8220, ST copy, no dvfs, socket 1)	0.262	10
Disk Config B (c8220, /dev/sda4, seqwrite, iodepth 4096)	1.708	37
Mem Config C (c220g1, ST copy, dvfs, socket 1)	6.139	74
Net Config D (m400, not rack-local, iperf3 (bw), forward)	6.309	10
Net Config E (m510, not rack-local, latency, forward)	8.086	230
Disk Config F (c8220, /dev/sda4, randread, iodepth 4096)	8.122	610

Trend: Higher CoV → More Runs

CoV and recommended runs are not perfectly correlated

Recommended runs rise fast with higher CoV

How many times should we run our experiments?

- Enough for target confidence
- Trend: high CoV → more runs
- Use past data to estimate

Can the facility help?

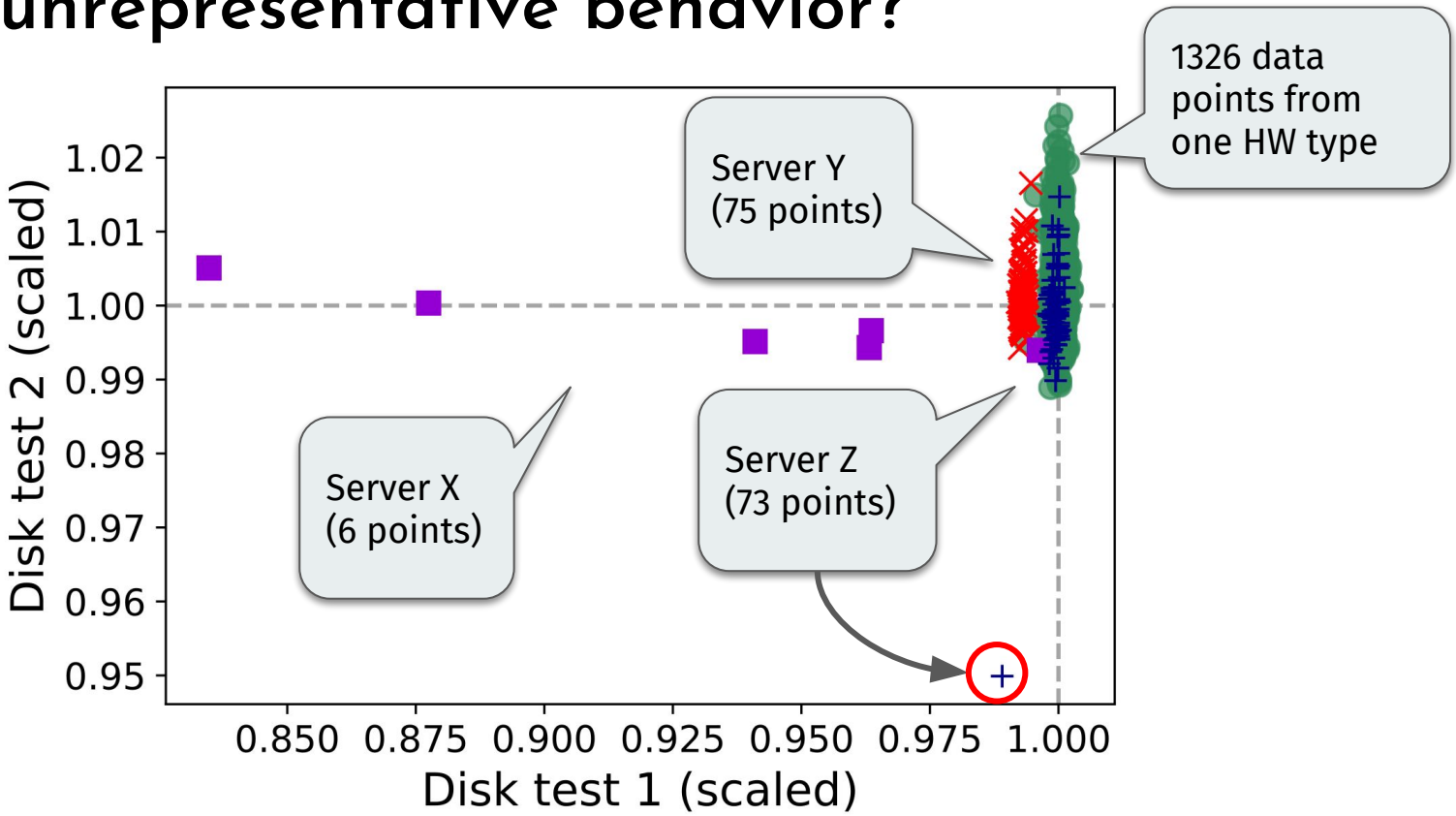
Can The Facility Help?

- Provide indistinguishable resources

Indistinguishable:

Performance results gathered on *any* server should be representative of the *population as a whole*.

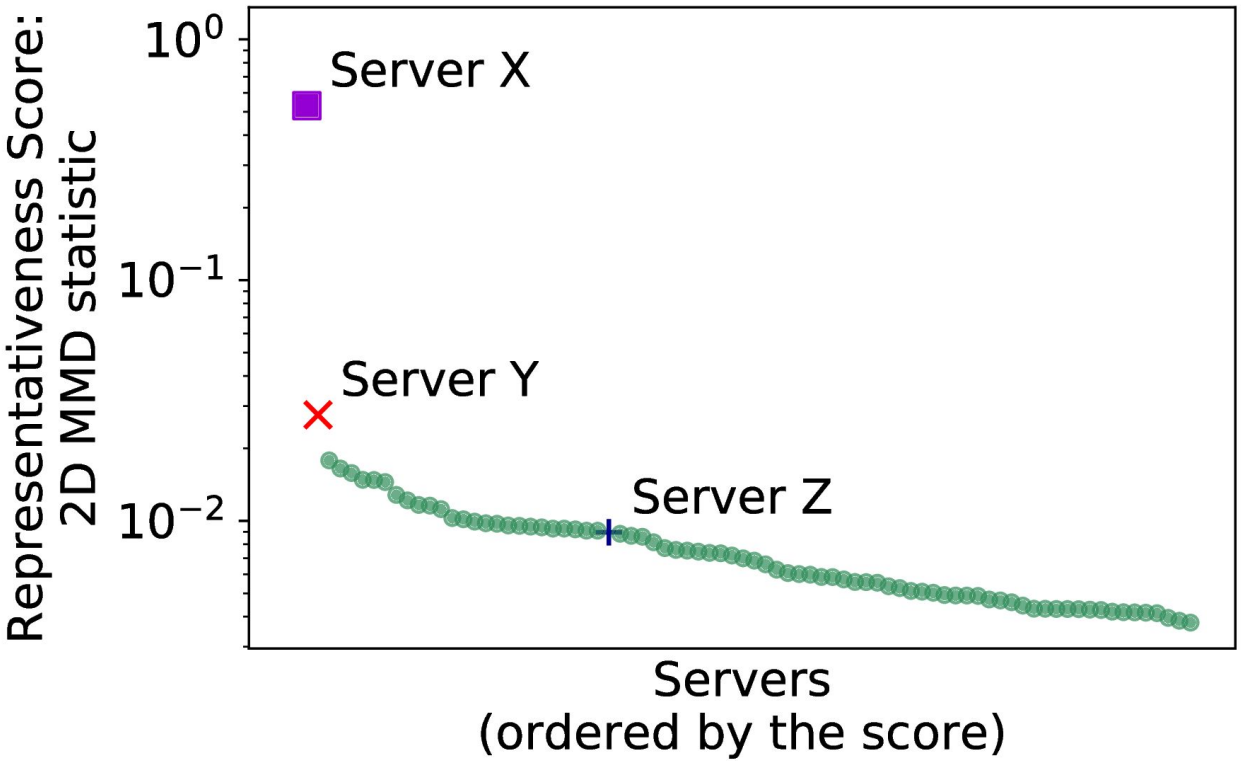
What is unrepresentative behavior?



Detecting Unrepresentative Resources

- Kernel two-sample test based on Maximum Mean Discrepancy (MMD)
  - Provides a measure of similarity between two non-parametric distributions
- We compare:
  - Each server to all others of its type
  - ... using many dimensions: disk, memory, and network
- Remove servers that are statistically dissimilar from the rest

Removing Unrepresentative Servers



Can The Facility Help?

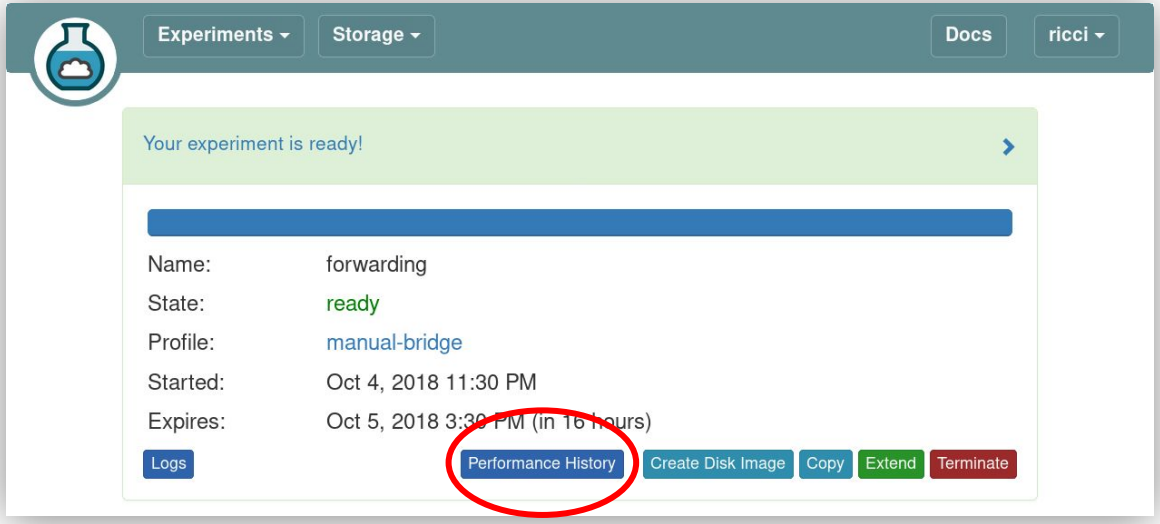
- Fix/remove < 2% of servers

Related Work

- Profiling
  - Cloud-scale (distributed) (Kanev et al., 2015, [1]) (Kozyrakis et al., 2010, [2])
  - Single-node (VM) applications (Yadwakar et al., 2014, [3])
- Quantifying Variability
  - Virtualized clouds (Iosup et al., 2011, [4])
  - Warehouse-scale computers (Dean and Barroso, 2013, [5])
- Other experimentation platforms
  - Baselining performance for Grid'5000 (Nussbaum, 2017, [6])

Summary

- How confident can we be in the correctness of our results?
  - Measure confidence with (non-parametric) CIs to account for unavoidable variability
- How many times should we run our experiments?
  - CONFIRM - Pick a target CI width, estimate the number of runs using past performance data
- Can the facility help?
  - Provide statistically indistinguishable resources
- More results, experiences with pitfalls in the paper



https://confirm.fyi

Poster #7

CloudLab Users BoF: 9:30, Las Palmas II

References

[1]: Kanev et al., Profiling a warehouse-scale computer. ACM SIGARCH News, 2015.

[2]: Kozyrakis et al., Server engineering insights for large-scale online services. IEEE micro, 2010.

[3]: Yadwadkar et al., Predictable and faster jobs using fewer resources. SOCC'14.

[4]: Iosup et al, On the performance variability of production cloud services. CCGrid'11.

[5]: Dean and Barroso. The tail at scale. Communications of the ACM, 2013.

[6]: Nussbaum. Towards trustworthy testbeds thanks to throughout testing, IPDPSW'17.