

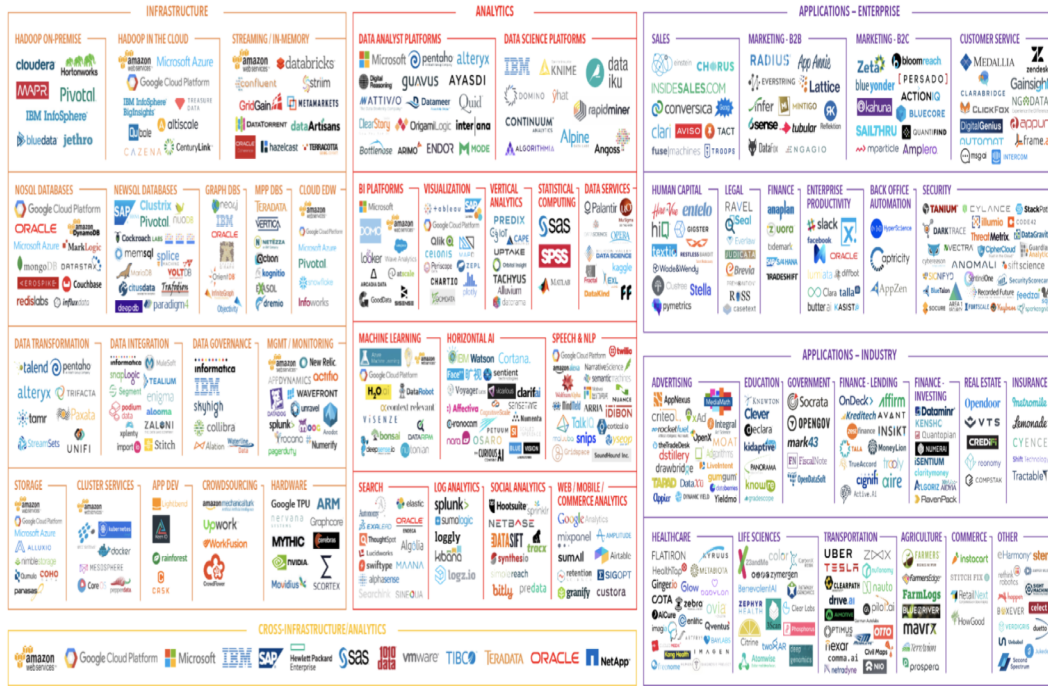
# Is Big Data Performance Reproducible in Modern Cloud Networks?

**Alex Uta**, A.Custura, D. Duplyakin, I. Jimenez, J.S. Rellermeier,  
C. Maltzahn, R. Ricci, A. Iosup

[a.uta@vu.nl](mailto:a.uta@vu.nl)

# Big data infrastructure in the cloud

## BIG DATA LANDSCAPE 2017



- Many big data frameworks, infrastructure
- Designed by both industry and academia
- Highly embedded in the cloud
- Every new system/release, research paper means performance evaluation

Image courtesy of mattturck.com

# How do we assess **performance** (in the cloud)?

Performance evaluation is inherently difficult in systems research.

- Do we understand the underlying conditions?

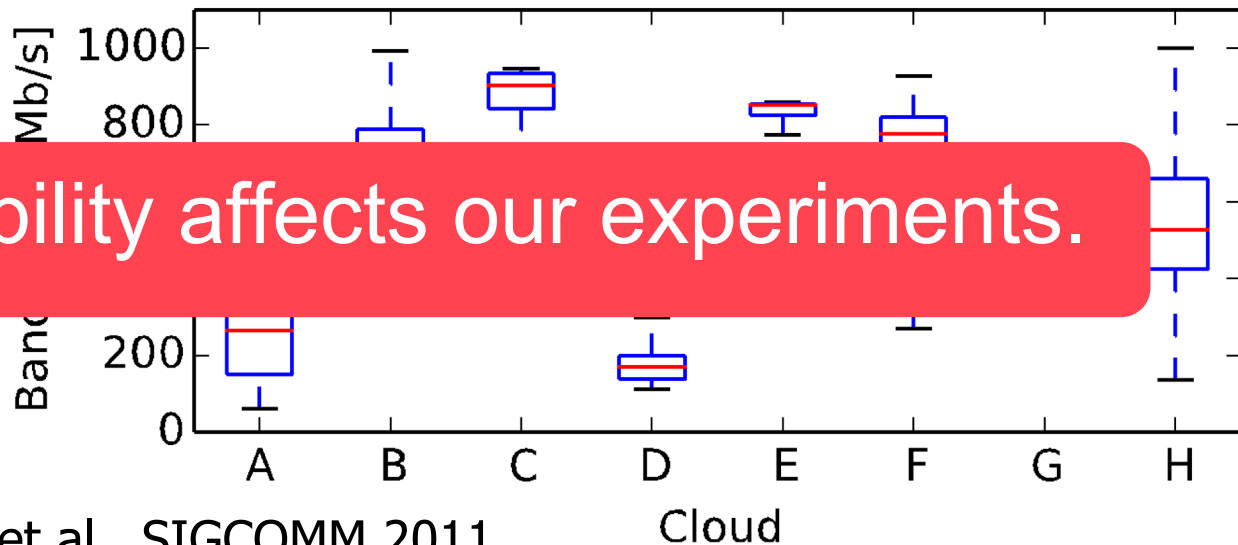
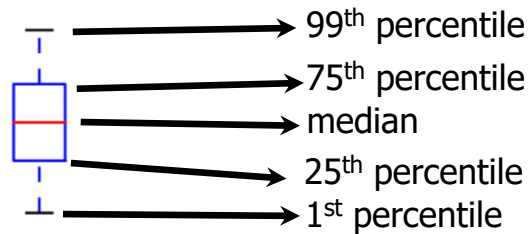
Cloud exacerbates performance variability.

- How many times do we run experiments in the cloud?
- What do we report? (means, medians, distributions)

# Cloud performance is **variable**

- Due to:
  - Co-location
  - Virtualization
  - Network congestion
  - “noisy neighbors”
  - Provider QoS policies

- Focus on networks:



Ballani et al., SIGCOMM 2011

# Agenda & Findings

1. Variability disregarded in cloud performance evaluations
2. Clouds (still) highly variable in network performance
3. Big data performance evaluations impacted by cloud network variability
4. Advice to achieve (more) reproducible performance

# Variability is disconsidered in performance evaluations

- Systematic survey -- 4 top systems conferences
- Time frame: 2010 – 2018
- Articles on: big data, analytics, data science, graph processing, streaming, MapReduce, Hadoop, Spark
- Systems evaluated on clouds



**44 articles > 11,000 citations**

# Variability is disconsidered in performance evaluations

## Questions asked:

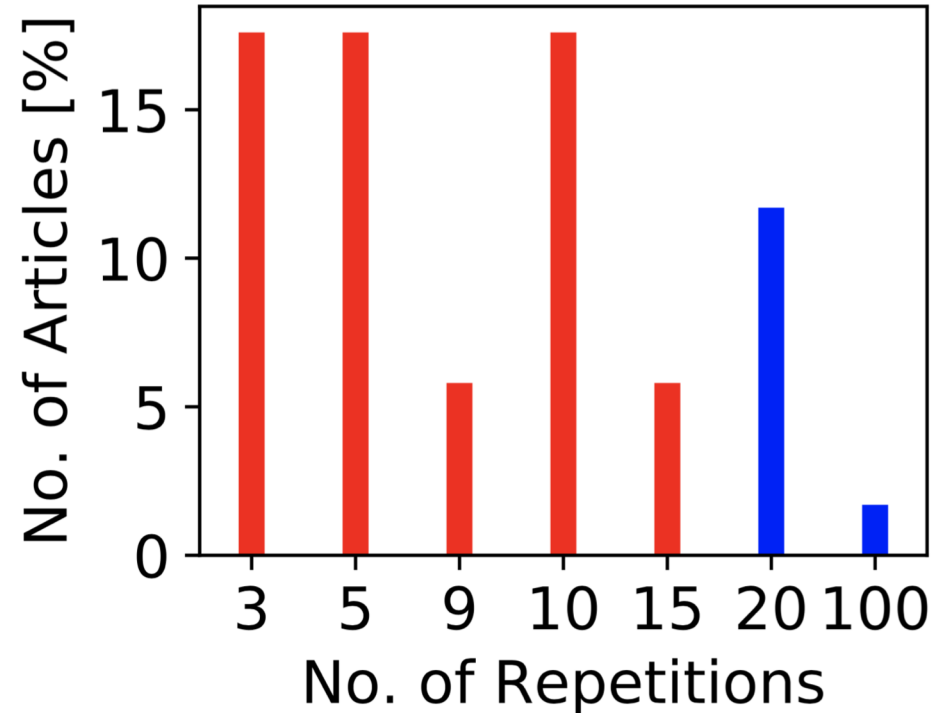
- How many runs/trials?
- Is the article reporting averages/medians over a number of runs/trials?
- Is the article reporting experiment variability? (error-bars, min-max, percentiles)



# Variability is disconsidered in performance evaluations

## Questions asked:

- **How many runs/trials?**
- Is the article reporting averages/medians over a number of runs/trials?
- Is the article reporting experiment variability? (error-bars, min-max, percentiles)

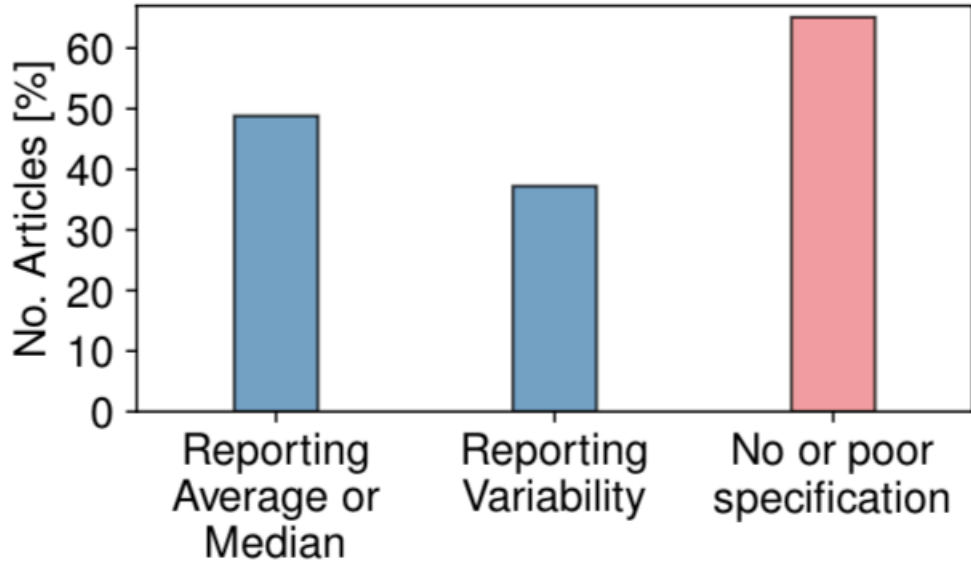




# Variability is disconsidered in performance evaluations

## Questions asked:

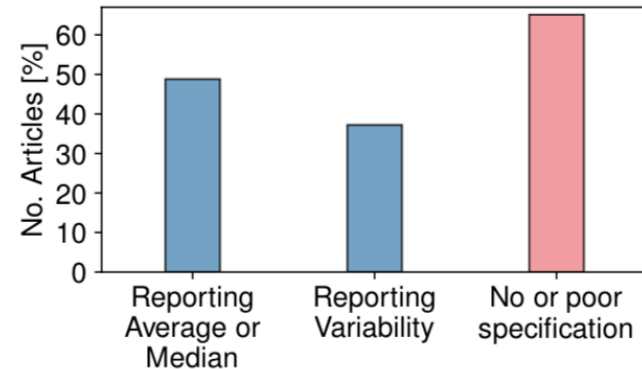
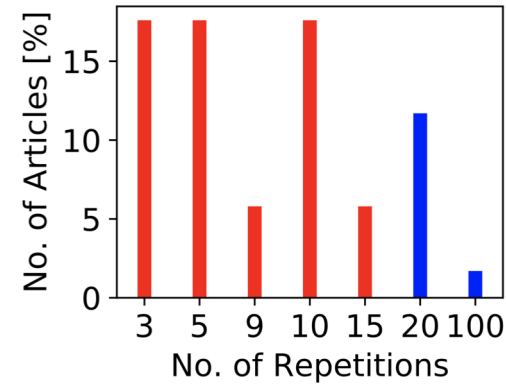
- How many runs/trials?
- **Is the article reporting averages/medians over a number of runs/trials?**
- **Is the article reporting variability? (error-bars, min-max, percentiles)**



# Variability is disconsidered in performance evaluations

## Main findings:



- **Most articles report 3-10 repetitions, few report > 10**
- **> 50% of articles have no or poor experiment specification!**
- **< 50% report only average or median**
- **~ 40% report variability**
- **Cited articles > 11,000 citations**



# Is Big Data Performance Reproducible in Modern Cloud Networks?

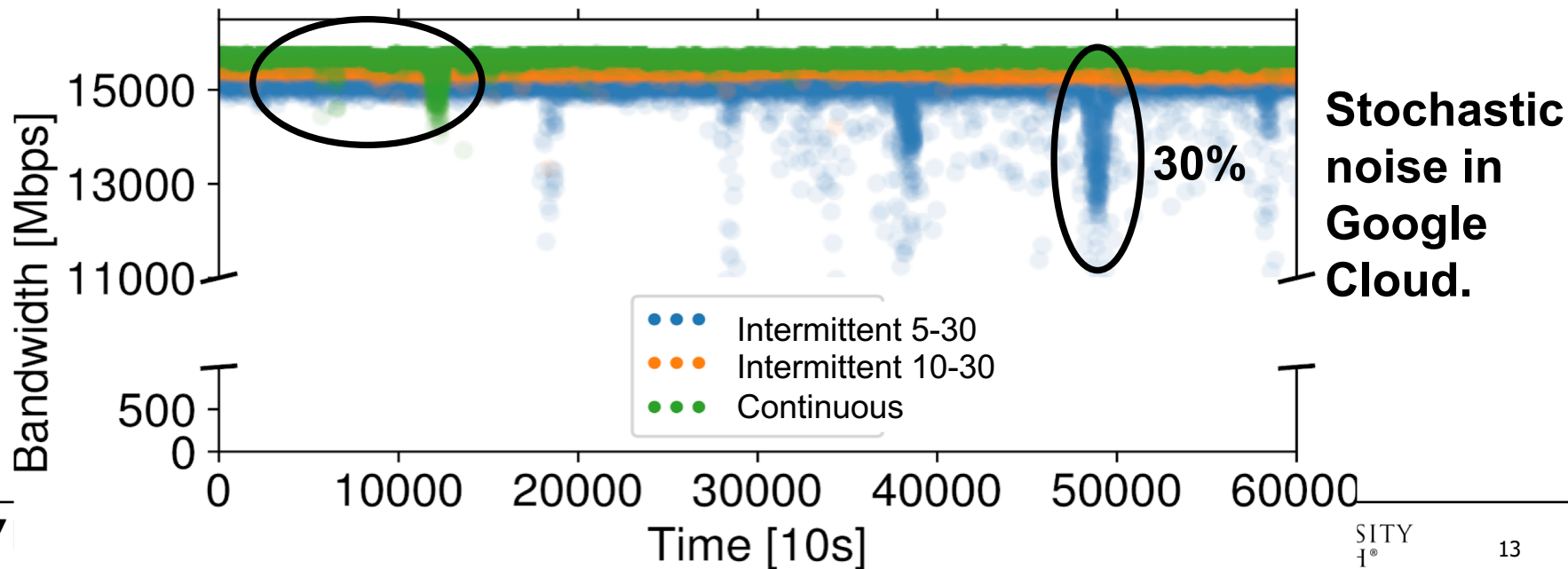
Spoiler alert: not so much...

# Experiment Design (1) -- Measuring the Cloud

- 2 commercial clouds: Amazon EC2, Google
- 1 private, research cloud: HPCCloud – SURF, in NL
- Measured bandwidth and latency for multiple instance types for a **week**
  - Only some instance types, budget limited
- Multiple communication patterns, to mimic real-world applications:
  - Continuous streaming 
  - Intermittent communication: 
    - 5 second stream, 30 second pause
    - 10 second stream, 30 second pause

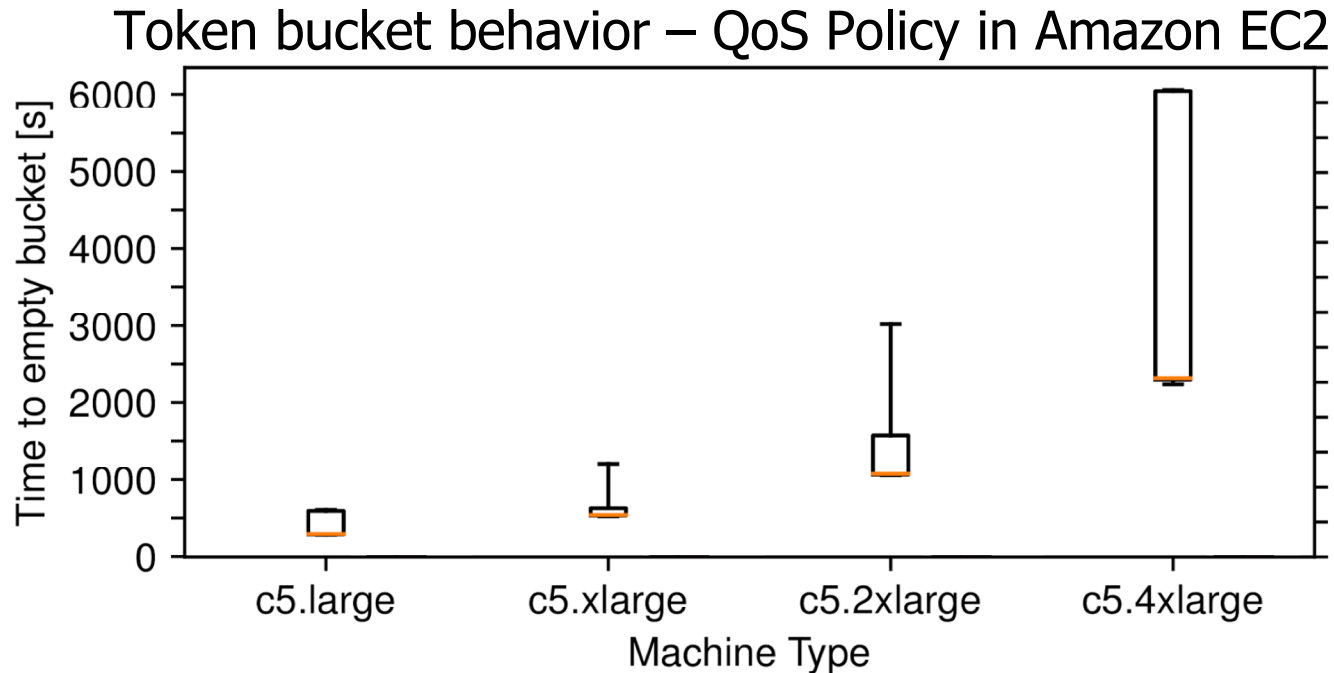
# Performance vs. stochastic noise

- **Problem:** Underlying cloud performance varies between and during experiments.
- **Mitigation:** Establish baselines of platform performance (microbenchmark the platform before running apps) and publish together with application performance.



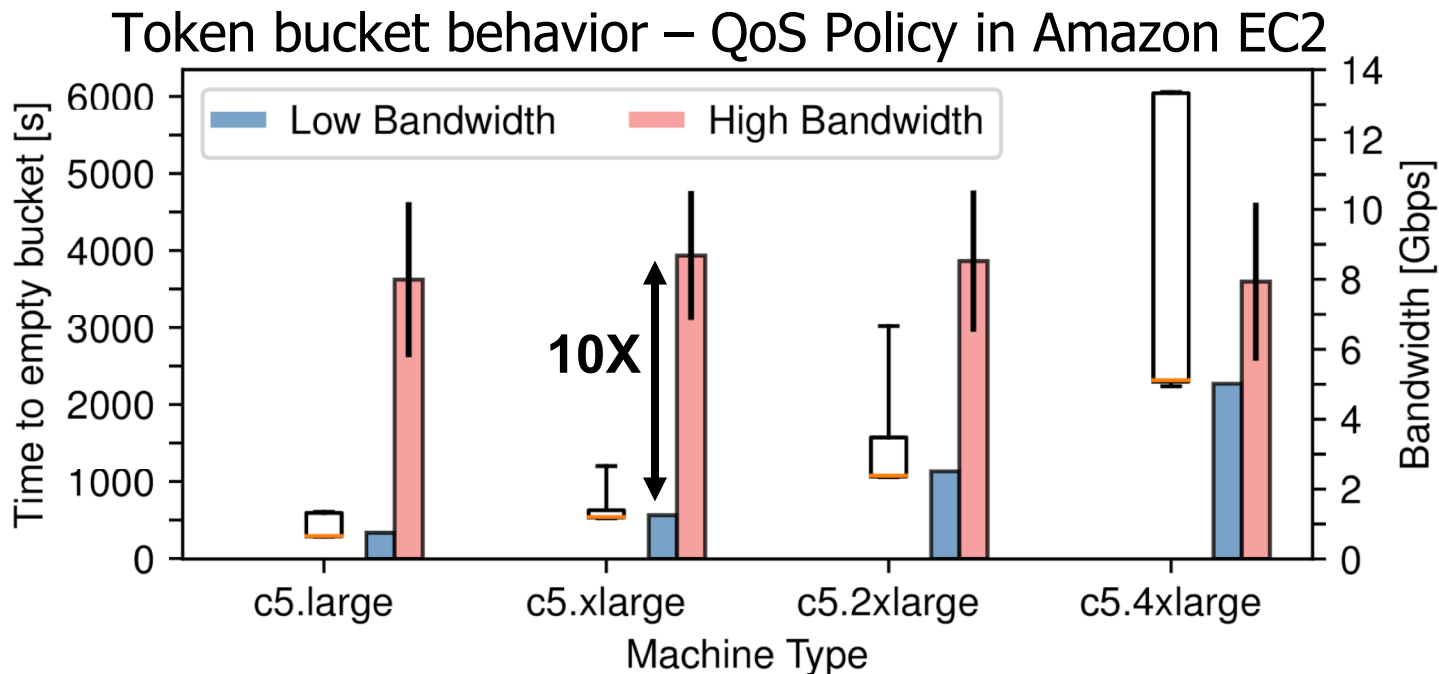
# Performance vs. Provider QoS policies

- **Problem:** Experiment might interact with provider policies.



# Performance vs. Provider QoS policies

- **Problem:** Experiment might interact with provider policies.
- **Mitigation:** Detect interaction with the provider (look for behavior that breaks assumptions, i.e., token buckets) and document these.



# Main Findings – Modern Cloud Networks

- **Large amounts of variability, quantified**
- **Variability depends on access pattern**
- **Different behavior between providers**
- **Two types of variability:**
  - **(Similar to) Stochastic noise**
  - **Given by provider QoS policies**

First to quantify

What does this mean for  
big data experiments?

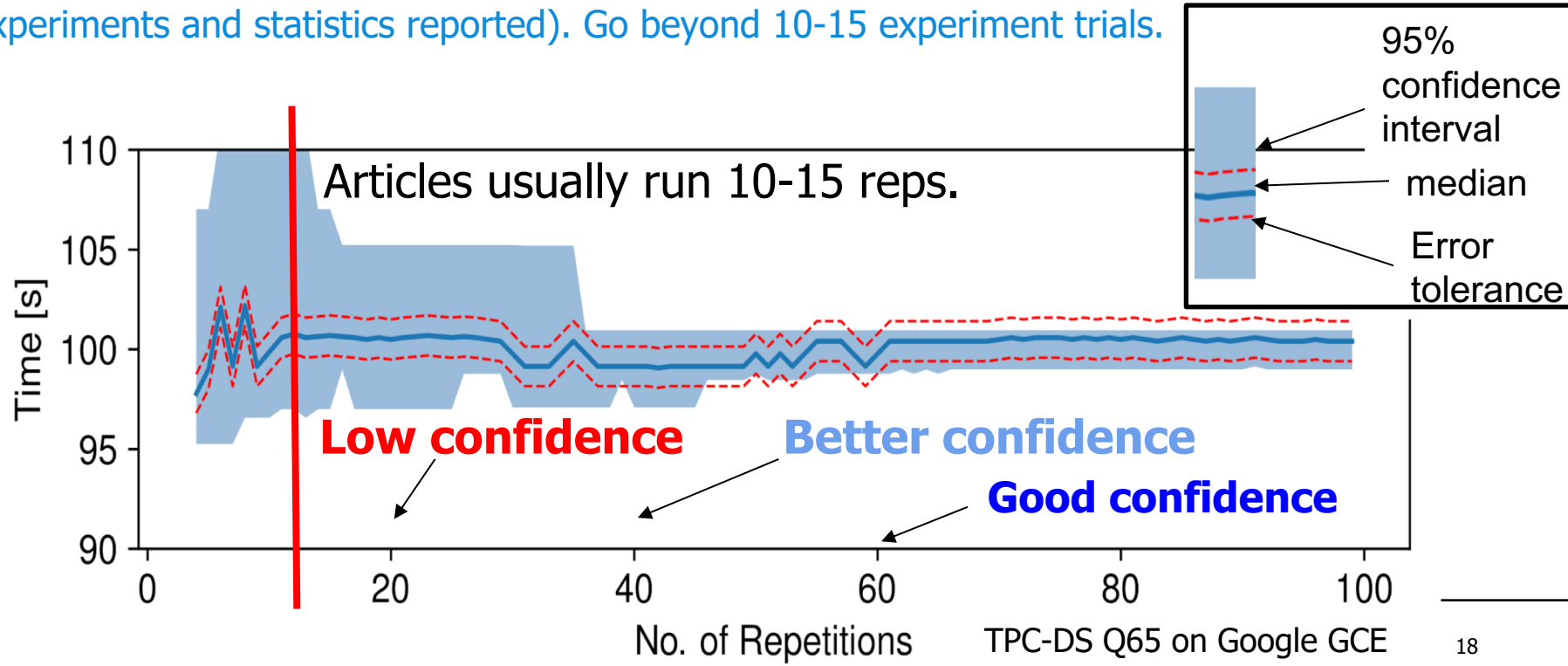


# Experiment Design (2) -- Reproducing App Performance

- Apps = Benchmarking suites: HiBench and TPC-DS
- Platform: Apache Spark
- **Stochastic noise:** ran directly on Google GCE and HPCCloud, NL
- **QoS Policies:** emulate AWS Token Bucket on own infrastructure

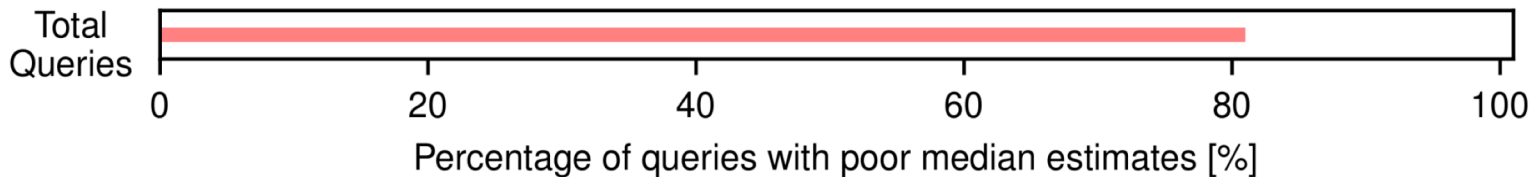
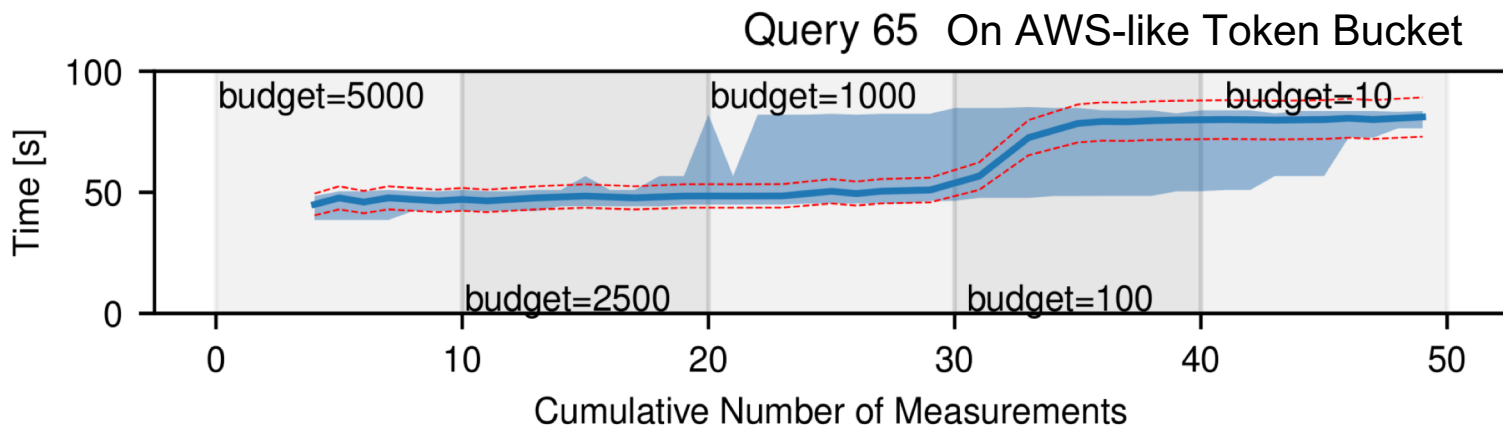
# How to run repeatable experiments?

- **Problem:** Interference (stochastic noise) affects experiments.
- **Mitigation:** Stochastic variability is tamed through robust experimentation (repeated experiments and statistics reported). Go beyond 10-15 experiment trials.



# How to run repeatable experiments?

- **Problem:** Repeated experiments might influence each other.
- **Mitigation:** Randomize experiment order, re-use machines sparingly, try different days, “rest” infrastructure.



# TL;DR: run more experiments

- Network performance variability is a **widespread** phenomenon
- Systems community often **neglects** performance variability in evaluations
- Important **reproducibility problem** in computer systems
- Reliable, repeatable performance == **more runs** than most articles do
- **1<sup>st</sup> to quantify**, but further **community-effort** needed to tackle it

Data released:

DOI [10.5281/zenodo.3576604](https://doi.org/10.5281/zenodo.3576604)

[DOI: 10.5281/zenodo.3576604](https://doi.org/10.5281/zenodo.3576604)