Conducting Repeatable Experiments in Highly Variable Cloud Computing Environments

Ali Abedi and Tim Brecht
Motivation

• Performance is highly variable in cloud environments

[Schad et al. 2010]
Motivation

• Performance is highly variable in cloud environments

Repeating exactly the same experiment/benchmark!
Motivation

• Performance is highly variable in cloud environments

[Schad et al. 2010]  [Iosup et al. 2011]  [Ballani et al. 2013]  [Leitner et al. 2016]
Motivation

- Performance is highly variable in cloud environments

High variability for CPU, Memory, Disk and Network
Problem

• Given high variability (lack of repeatability)
  – How to compare competing alternatives (FAIRLY) (e.g., A, B, and C)?
Existing Approaches

Single Trial

A  B  C
Existing Approaches

Single Trial

A  B  C

Potential issue: Variation in performance
Existing Approaches

Single Trial

A  B  C

Multiple Consecutive Trials (MCT)

A  A  A  B  B  B  C  C  C  C
Existing Approaches

Single Trial

A + CI

Multiple Consecutive Trials (MCT)

A + CI

B + CI

C + CI
Existing Approaches

Single Trial

A   B   C

Multiple Consecutive Trials (MCT)

A   A   A   B   B   B   C   C   C

In 2016 Symposium on Cloud Computing (SoCC):
• Use only these two approaches (18 evaluations)
• Mostly Single Trial
  Public Clouds: Single Trial (7), MCT (4)
Existing Approaches

Single Trial

A B C

Multiple Consecutive Trials (MCT)

A A A B B B C C C

Potential issue: Alternatives are separated by time
Proposed Approaches

Single Trial

A  B  C

Multiple Consecutive Trials (MCT)

A  A  A  B  B  B  C  C  C  C

Multiple Interleaved Trials (MIT)

A  B  C  A  B  C  A  B  C  C
Proposed Approaches

Single Trial

A  B  C

Multiple Consecutive Trials (MCT)

A  A  A  B  B  B  C  C  C  C

Multiple Interleaved Trials (MIT)

A  B  C  A  B  C  A  B  C  C

Potential issue: Periodic changes in cloud environment
Proposed Approaches

Single Trial
A B C

Multiple Consecutive Trials (MCT)
A A A B B B C C C C

Multiple Interleaved Trials (MIT)
A B C A B C A B C

Randomized Multiple Interleaved Trials (RMIT)
B A C C B A A C B
Single Trial

EC2: CPU performance, US, large instance

[Schad et al. 2010]
Methodology: Multiple Trials

- All data used comes from same set of experiments
- Examine different orderings for trials
  - MCT, MIT, RMIT
- Consider two different alternatives

[Schad et al. 2010]
Methodology: Consecutive Trials
Methodology: Consecutive Trials

[A] [B]
Methodology: Consecutive Trials

A

B

A
Methodology: Consecutive Trials
Methodology: Interleaved Trials

[Schad et al. 2010]
Methodology: Interleaved Trials

- Use exactly same data but label as different alternatives
- Consider interleaved ordering for conducting trials

[Schad et al. 2010]
Methodology: Interleaved Trials
Methodology: Interleaved Trials
Methodology: Interleaved Trials
Methodology: Interleaved Trials
Methodology: Randomized, Interleaved
Methodology: Randomized, Interleaved
Methodology: Randomized, Interleaved
Methodology: Randomized, Interleaved
Methodology: Randomized, Interleaved

\[ \text{Methodology: Randomized, Interleaved} \]

\[ \text{Methodology: Randomized, Interleaved} \]
Methodology: Randomized, Interleaved
• 20 trials of each alternative (95% Confidence Intervals)
  – All results **SHOULD BE SIMILAR**
  – If not, approach to running experiments is flawed
Multiple Consecutive Trials (MCT)

A A A B B B B C C C C

EC2: CPU performance, US, small instance
Multiple Consecutive Trials (MCT)

A A A B B B C C C C

EC2: CPU performance, US, small instance
Multiple Consecutive Trials

![Graph showing data over time with two alternatives labeled as (A) and (B).]
Multiple Consecutive Trials

Flawed
Multiple Consecutive Trials

CI’s provide false sense of validity and rigour
Multiple Interleaved Trials

Multiple Interleaved Trials (MIT)

EC2: CPU performance, US, large instance
Multiple Interleaved Trials (MIT)

A  B  C  A  B  C  A  B  C

EC2: CPU performance, US, large instance
Multiple Interleaved Trials

Flawed
Randomized Multiple Interleaved Trials

Randomized Multiple Interleaved Trials (RMIT)

EC2: CPU performance, US, large instance
Randomized Multiple Interleaved Trials (RMIT)

EC2: CPU performance, US, large instance
Randomized Multiple Interleaved Trials

All Are Now Similar!
Randomized Multiple Interleaved Trials

All Are Now Similar!

RMIT approach is sound
Conclusions

• When comparing competing alternatives
• Use Randomized Multiple Interleaved Trials (RMIT)
  – Or risk invalid comparisons and incorrect conclusions

Because:
  – Performance in clouds is highly variable
  – Environment changes
  – Not possible to know about or detect changes!
Conclusions

• When comparing competing alternatives
• Use Randomized Multiple Interleaved Trials (RMIT)
  – Or risk invalid comparisons and incorrect conclusions

Because:
  – Performance in clouds is highly variable
  – Environment changes
  – Not possible to know about or detect changes!

Questions?

Data used is available: URL is in the paper
The End
CDF

Percentage difference between consecutive trials
## Efficacy of MCT

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Instance</th>
<th>Region</th>
<th>RFE</th>
<th>Max Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>Small</td>
<td>US</td>
<td>10.2</td>
<td>37.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EU</td>
<td>9.8</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>US</td>
<td>7.9</td>
<td>30.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EU</td>
<td>14.9</td>
<td>36.5</td>
</tr>
<tr>
<td>Memory</td>
<td>Small</td>
<td>US</td>
<td>9.8</td>
<td>15.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EU</td>
<td>6.7</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>US</td>
<td>10.2</td>
<td>16.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EU</td>
<td>7.0</td>
<td>11.1</td>
</tr>
<tr>
<td>Disk I/O</td>
<td>Small</td>
<td>US</td>
<td>4.0</td>
<td>15.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EU</td>
<td>0.0</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>US</td>
<td>5.9</td>
<td>23.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EU</td>
<td>0.0</td>
<td>NA</td>
</tr>
<tr>
<td>Network</td>
<td>Mixed</td>
<td>US</td>
<td>6.1</td>
<td>19.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EU</td>
<td>5.1</td>
<td>12.5</td>
</tr>
</tbody>
</table>
CPU Performance Large Instance, US

![Graph showing CPU performance over time for even and odd trials.](image-url)
Methodology: Interleaved Trials

• 20 trials of each alternative (95% Confidence Intervals)
  – All results **SHOULD BE SIMILAR**
  – If not, approach to running experiments is flawed
Paying $1,500 to browse Twitter and hang out on Slack

Half-listening to Conference Talks

In Depth

@ThePracticalDev
Proposed Approaches

Single Trial

A  B  C

Multiple Consecutive Trials (MCT)

A  A  A  B  B  B  C  C  C  C

Multiple Interleaved Trials (MIT)

A  B  C  A  B  C  A  B  C  A  B  C

A + CI
Proposed Approaches

Single Trial

\[ A \quad B \quad C \]

Multiple Consecutive Trials (MCT)

\[ A \quad A \quad A \quad B \quad B \quad B \quad C \quad C \quad C \quad C \]

Multiple Interleaved Trials (MIT)

\[ A \quad B \quad C \quad A \quad B \quad C \quad A \quad B \quad C \]

- \[ A + CI \]
- \[ B + CI \]
Proposed Approaches

Single Trial

A  B  C

Multiple Consecutive Trials (MCT)

A  A  A  B  B  B  C  C  C

Multiple Interleaved Trials (MIT)

A  B  C  A  B  C  A  B  C

A + CI  B + CI  C + CI
Proposed Approaches

Single Trial

A  B  C

Multiple Consecutive Trials (MCT)

A  A  A  B  B  B  C  C  C

Multiple Interleaved Trials (MIT)

A  B  C  A  B  C  A  B  C

A + CI  B + CI  C + CI

Potential issue: Periodic changes in cloud environment
Proposed Approaches

Single Trial
A B C

Multiple Consecutive Trials (MCT)
A A A B B B C C C C

Multiple Interleaved Trials (MIT)
A B C A B C A B C

Randomized Multiple Interleaved Trials (RMIT)
B A C C B A A C B

A + CI
Proposed Approaches

Single Trial

A  B  C

Multiple Consecutive Trials (MCT)

A  A  A  B  B  B  C  C  C  C

Multiple Interleaved Trials (MIT)

A  B  C  A  B  C  A  B  C

Randomized Multiple Interleaved Trials (RMIT)

B  A  C  C  B  A  A  C  B

A + CI  B + CI
Proposed Approaches

Single Trial

\[ A \, B \, C \]

Multiple Consecutive Trials (MCT)

\[ A \, A \, A \, B \, B \, B \, C \, C \, C \, C \]

Multiple Interleaved Trials (MIT)

\[ A \, B \, C \, A \, A \, B \, B \, C \, A \, B \, C \]

Randomized Multiple Interleaved Trials (RMIT)

\[ B \, A \, C \, C \, C \, B \, A \, A \, A \, C \, B \]
Multiple Consecutive Trials

Flawed
Multiple Interleaved Trials

Flawed
Multiple Interleaved Trials
Multiple Consecutive Trials