Automated Detection of Performance Regressions Using Regression Models on Clustered Performance Counters

Weiyi Shang, Ahmed E. Hassan
Queen’s University, Kingston, Canada

Mohamed Nasser, Parminder Flora
BlackBerry, Waterloo, Canada
What is a performance regression?

Does the new version have worse performance than the old version?
How to detect performance regression?

Old Version

New version

requests

requests

requests

Performance counters

Performance counters
Large software systems generate large amounts of performance counters
Performance engineers pick counters to compare using T-test

Are these counters significantly different between old and new versions?

Picking the wrong counters will not detect performance regression!
Performance engineers build a model using all counters

Target Counter

Independent Counters

CPU

Memory

I/O read

I/O write
Model-based performance regression detection

Target counter is selected by experience!
Selecting a wrong target counter will fail to detect performance regression.

- Memory has a low correlation with CPU
- I/O write has a high correlation with CPU

Memory-related performance regressions will not be detected by this model!
Our approach

Reduce counters

Cluster counters

Reduced counters

Clusters of counters

For each cluster

Select target counter

Target counter

Build model

Model

Verify model

Prediction error

Old version

New version

Reduce counters

Model prediction error
Step 1: Reduce counters

Removing zero variance counters
Available CPU cores: 4, 4, 4, ...4, 4

Removing redundant counters
I/O write byte/sec = a*I/O write op/sec + b*CPU
Step 2: Cluster counters

Hierarchical clustering

- CPU
- I/O
- Memory

Dendrogram cutting

- CPU
- I/O
- Memory

Flowchart:
- Reduce counters
- Cluster Counters
- Select target counter
- Build model
- Verify model
Step 3: Select target counter

We select the counter that has significant difference between two versions with highest certainty.

We select the counter with smallest p-value.

We select the counter with smallest p-value.
Step 3: Build model

We build a linear regression model.

\[ a \times \text{Memory} + b \times \text{I/O read} + c \times \text{I/O write} \]
Step 4: Verify model

Linear regression models for each cluster

New version counters left in cluster 1

Linear regression model for cluster 1

Calculate prediction error

Reduce counters → Cluster Counters → Select target counter → Build model → Verify model

Mean prediction error

< threshold

> threshold
Case study: subject systems

DELL DVD Store

BlackBerry

Enterprise Application

CPU overhead
Memory overhead
I/O overhead

Removing text index
Removing column index

Real-life performance regression
Applicability

How many target counters does our approach pick?

Accuracy

Can our approach detect performance regressions?

Comparison

Is our approach better than traditional approaches?
Our approach picks a small number of target counters

Our approach picks 2 to 4 target counters.

Picked target counters are different across runs of our approach.

Performance engineers cannot pick target counters based on experience
Applicability

How many target counters does our approach pick?

Our approach picks a small number of target counters.

Accuracy

Can our approach detect performance regressions?

Comparison

Is our approach better than traditional approaches?
Applicability

How many target counters does our approach pick?

Accuracy

Can our approach detect performance regressions?

Comparison

Is our approach better than traditional approaches?

Our approach picks a small number of target counters.
Our approach can detect performance regressions.

Max prediction error
4% to 11% without regression

24% to 1622% with regression

Our approach is not heavily impacted by the choice of threshold value.
Can our approach detect performance regressions?

Our approach can detect performance regressions and is not heavily impacted by threshold value.
**Applicability**

How many target counters does our approach pick?

- Our approach picks a small number of target counters.

**Accuracy**

Can our approach detect performance regressions?

- Our approach can detect performance regressions and is not heavily impacted by threshold value.

**Comparison**

Is our approach better than traditional approaches?
Our approach outperforms picking one target counter to build a model.

Building one model using memory as target counter may fail to detect performance regression.
T-test does not perform well in detecting performance regressions.

There are a large number of counters with significant differences in the T-test results even though no regressions exist.
Can our approach detect performance regressions?

Our approach can detect performance regressions and is not heavily impacted by threshold value.

Is our approach better than traditional approaches?

Our approach outperforms traditional approaches.

Our approach picks a small number of target counters.

How many target counters does our approach pick?
How to detect performance regression?
How to detect performance regression?

- Requests to Old Version
- Requests to New Version
- Performance counters

Diagram illustrating the process of detecting performance regression.
Model-based performance regression detection

Select target counter → Build model → Verify model

Old version

New version

Target counter is selected by experience!
How to detect performance regression?

Model-based performance regression detection

Target counter is selected by experience!
Our approach

Select target counter

Build model

Verify model

Prediction error

For each cluster

Cluster counters

Clusters of counters

Reduce counters

Reduced counters

Old version

New version

Target counter

Model

Model prediction error

Error
How to detect performance regression?

Model-based performance regression detection

Our approach

Target counter is selected by experience!
Applicability

How many target counters does our approach pick?

Our approach picks a small number of target counters.

Accuracy

Can our approach detect performance regressions?

Our approach can detect performance regressions and is not heavily impacted by threshold value.

Comparison

Is our approach better than traditional approaches?

Our approach outperforms traditional approaches.