Ensembles of Models for Automated Diagnosis of System Performance Problems

Steve Zhang (steveyz@cs.stanford.edu)
Armando Fox

In collaboration with:
Ira Cohen, Moises Goldszmidt, Julie Symons of HP Labs

Motivation

- Complexity of deployed systems surpasses ability of humans to diagnose, forecast, and characterize behavior
  - Scale and interconnectivity
  - Levels of abstraction
  - Lack of closed-form mathematical characterization
  - Changes in input, applications, infrastructure
- Need for automated diagnosis tools

Metric Attribution

- Service Level Objectives (SLO)
  - Usually expressed in terms of high-level behaviors, e.g. desired average response time or throughput
- Which low level metrics are most correlated with a particular SLO violation
  - Root cause diagnosis usually requires domain-specific knowledge
  - Often does point to root cause
- Technique can be applied to more than just performance problems

Background and Previous Work

- Modeling system behavior as a pattern classification problem
  - Learn a function F that maps state of low level metrics (M) to SLO state (s^+/s^-)
  - Tree Augmented Naïve Bayes (TAN) models are used to represented F's underlying probability distribution
  - Models evaluated using Balanced Accuracy (BA) =

    $$0.5(\text{Prob}(F(M) = s^- | s^-) + \text{Prob}(F(M) = s^+ | s^+))$$
- First step is feature selection
  - Find subset of metrics that produces F with best BA
  - Usually use heuristic search (greedy search)
  - Avoids curse of dimensionality as well as overfitting

Background (Cont.)

- TAN Models have key property of interpretability: (Metric Attribution)
  - TAN models represent joint probability estimate $P(s^{+/-},M)$
  - For each SLO violation, invert the joint probability estimate of each metric to obtain $P(M_i|s^+)$ and $P(M_i|s^-)$.
  - Metric is attributed with an SLO violation if difference
    $$P(M_i|s^+) - P(M_i|s^-) > 0.$$
  - Output is a list of metrics that are implicated with each SLO violation. (out of the subset of the metrics chosen during the feature selection process)
  - Not straightforward with many popular classifiers like Support Vector Machines or Neural Networks.

Key Results of Previous Work

- Combinations of metrics more predictive of SLO violations than individual metrics.
  - Different combinations under different workload conditions
  - Different combinations under different SLO definitions
- Small numbers of metrics (3-8) usually sufficient to predict SLO violation
  - Selected metrics yield insight into cause of problem
- Relationships among multiple metrics of a model simple enough to be captured by TAN
  - In most cases models yield a BA of 90% - 95%

Why Multiple Models?

- Previous work focused on forensic analysis, after all data available
- Experiments show models built using data collected under one condition do not work well on data collected under different conditions
- By using multiple models, we can incorporate new data as needed by learning new models
- Multiple models would help keep a history of past behavior and eliminate need to re-learn conditions that occurred previously.

Managing an Ensemble of Models

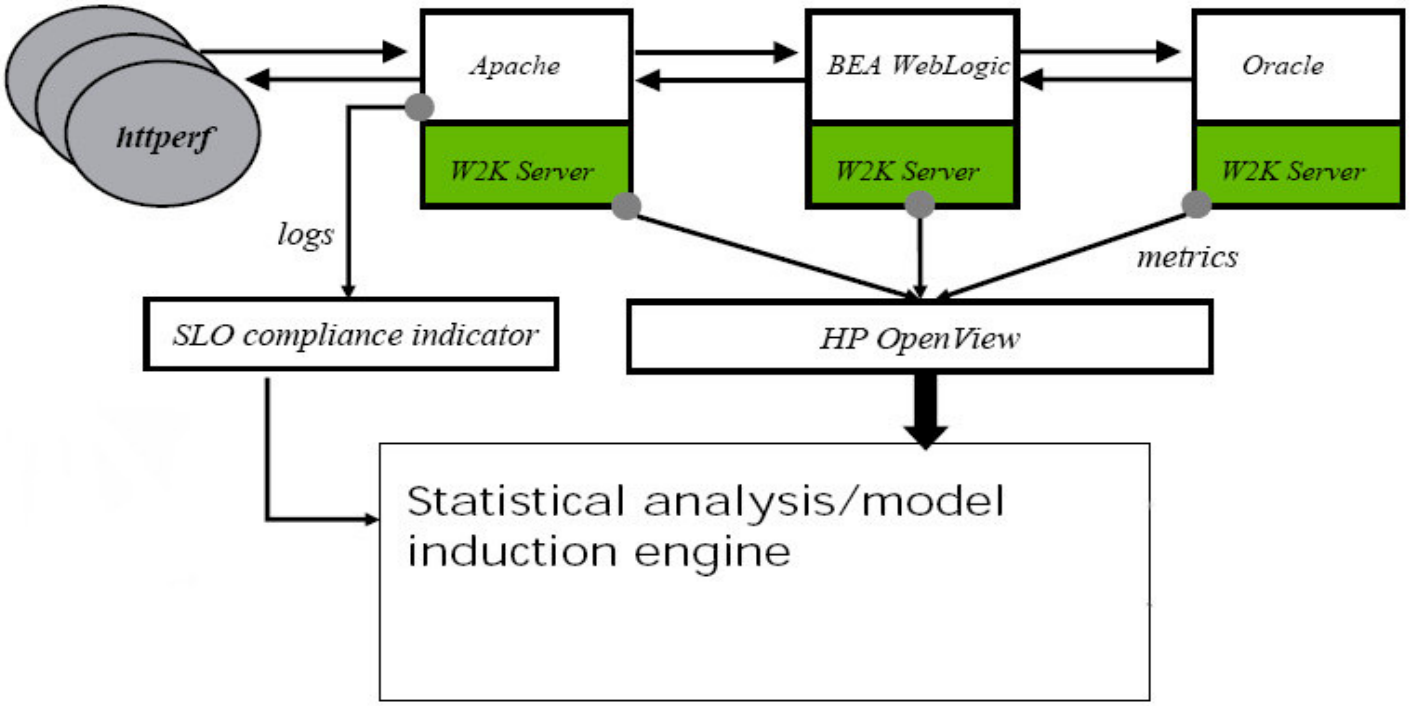
- How much data is needed to build a model?
  - Determined empirically using learning surfaces
- When do we add a new model to the ensemble?
  - When a new model does statistically significantly better over its training window than all current models
- How to fuse information from multiple models?
  - Use metric attribution from model that has best Brier score
    - Brier score is similar to Mean Squared Error (MSE) and offers a finer grain evaluation of a model, important when using small evaluation windows

Algorithm Outline

```
Start with empty TrainingWindow and empty Ensemble
for every new sample do
   Add sample to TrainingWindow
   if TrainingWindow has enough samples then
        Train new Model M on TrainingWindow
        Compute acc(M) using cross validation
        if acc(M) is significantly higher than
           accuracy of all models in Ensemble then
               Add new model to Ensemble
        end if
   end if
   Compute Brier Score over TrainingWindow for all models
   if new sample is SLO violation then
        Do metric attribution using model with best Brier
   Score
   end if
end for
```

Experimental Setup

- Apache Web server + WebLogic App Server + Oracle DB
- Petstore installed on app server



Workloads

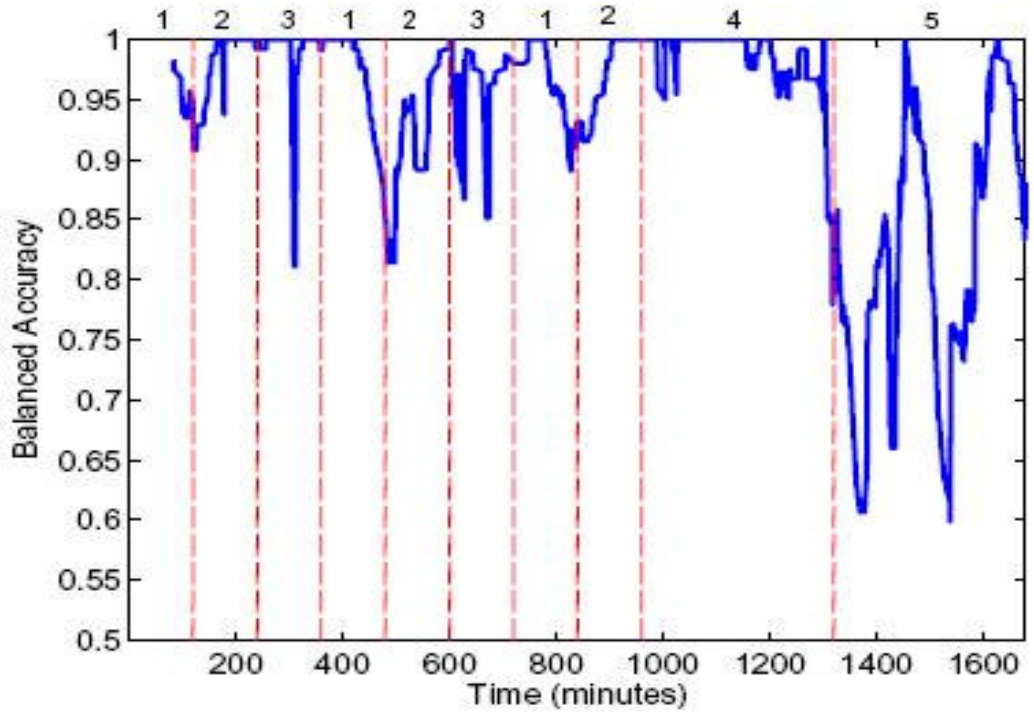
- Collected system metric data under 5 different workload conditions
- RAMP: gradual increase in concurrency and request rate
- BURST: sudden, sustained bursts of increasingly intense workload against backdrop of moderate activity
- BUYSPREE: alternates between normal periods (1 hour) and periods where clients are doing heavy buying (but number of clients remain the same)
- DBCPU & APPCPU: External program takes 30% of CPU for DB or APP server every other hour.
- Used a combination of data from different workloads to evaluate our techniques

Accuracy of the Ensemble

- Ensemble significantly outperforms single model
- Also does slightly better than a workload specific approach
  - Indicates that some workload conditions too complex for single model

	# metrics chosen	avg # attr metrics	BA	FA	Det
Ensemble W80	64	2.3	95.67 ±0.81	4.19± 1.12	95.53± 1.21
Ensemble W120	52	2.5	95.12 ±0.86	4.84± 1.2	95.19± 1.23
Ensemble W240	33	3.7	94.68 ±0.90	5.48± 1.27	94.85± 1.27
Workload specific models	9	2	93.62±0.97	4.51± 1.16	91.75±1.58
Single model	4	2	86.10± 1.38	21.61 ± 2.31	93.81± 1.38

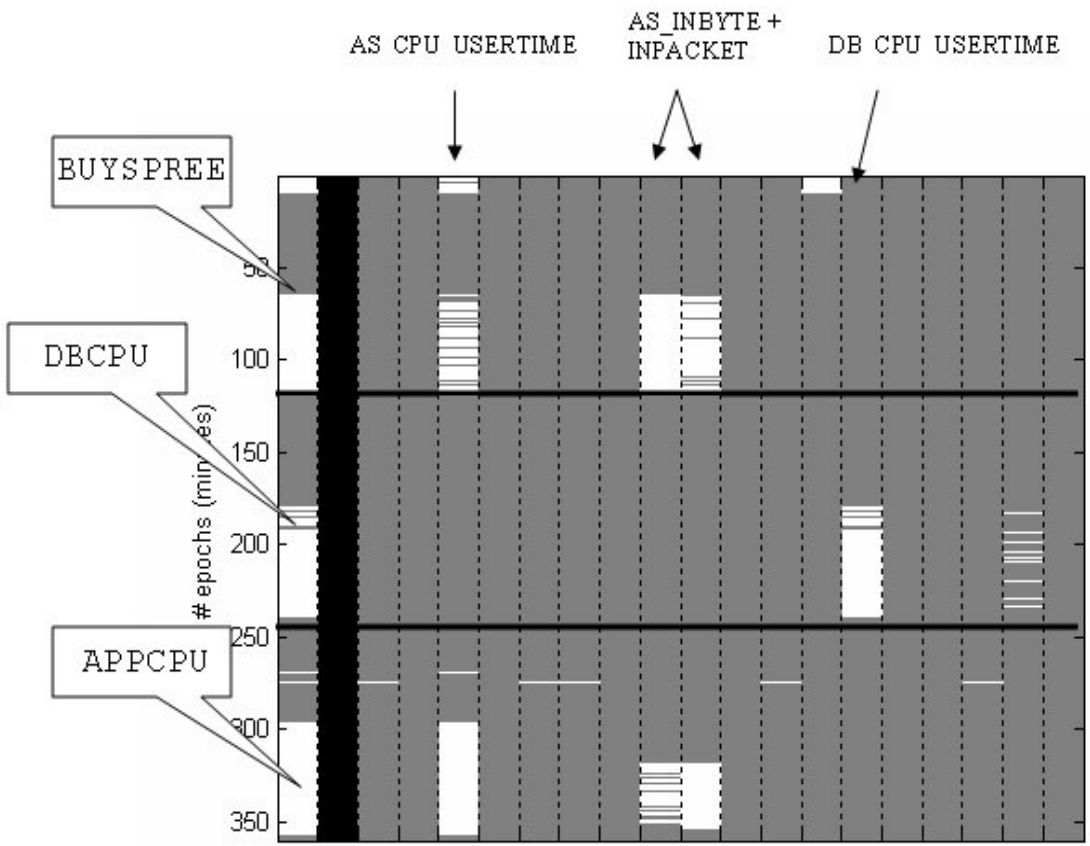
Accuracy During Training



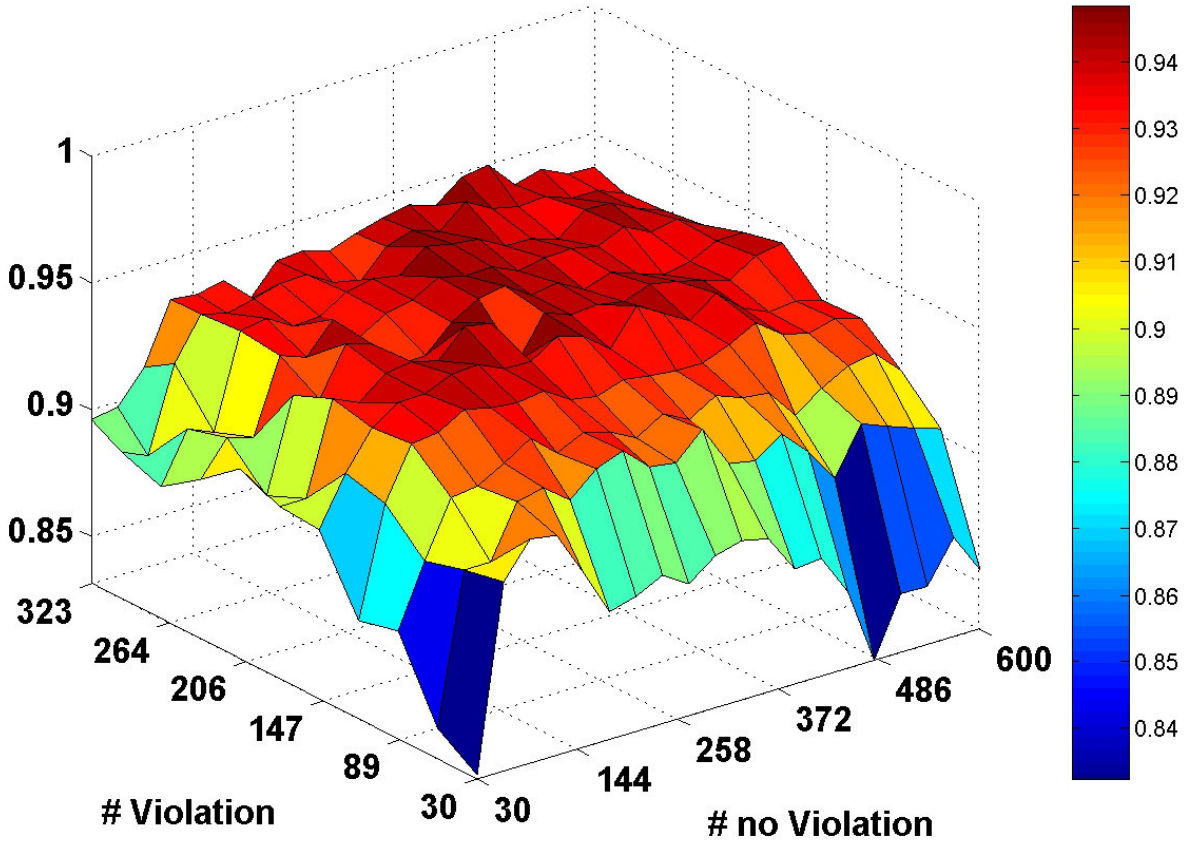
DBCPU(1), APPCPU(2), BUYSPREE(3), RAMP(4), BURST(5)

Metric Attribution

- Each instance of SLO violation may be indicated by several metrics.
- Which metrics indicate violations change as the workload changes
  - First column indicates SLO violation (white) or compliance
  - Other columns indicate if a particular metric indicated violation (white) or not



Learning Surfaces



Learning Surfaces (Cont.)

- Simpler workload conditions require fewer samples of each class
- Accuracy of model depends on ratio of class samples, not just number of samples
  - less sensitive with sufficient samples of each class

Workload	# vio	# no vio	max BA(%)
RAMP	90	80	92.35
BURST	180	60	81.85
BUYSPREE	40	40	95.74
APPCPU	20	30	97.64
DBCPU	20	20	93.90

Summary

- Ensemble of models perform better than single models or even workload specific models
- Our approach allows for rapid adaptation to changing conditions
- No domain specific knowledge is required
- Different workloads seem to be characterized by different metric-attribution “signatures” (future work)