Self-adaptive techniques for the load trend evaluation of internal system resources

Evan Behar
CS 895 – Fall 2010

Material in these slides is adapted from the paper “Self-adaptive techniques for the load trend evaluation of internal system resources”, S. Casolari, M. Colajanni and S. Tosi, ICAS '09. All copyrights belong to the authors or the paper's publishers.
Overview

- Problem Definition
- Previous Work
- Trend Modeling
  - Qualitative Trend Model
  - Quantitative Trend Model
- Experimental Results
- Conclusions
In a self-balancing load regulation system, being able to evaluate the metrics relevant to load is crucial to actually performing regulation.

In many systems, especially Internet-based systems, load is unpredictable and highly variable.

We need to be able to model and view system load in a way that accurately captures the behavior of the system with minimal lag.
Previous Work

- Instantaneous evaluations of the system-state
  - Short-sighted view of load
  - High variability in metrics, does not react well to temporary spikes in demand

- Average system state
  - Reduces variability
  - Introduces unacceptable lag into taking corrective action, significantly reducing the usefulness of the load model
Previous Work

- Stochastic models (e.g., ARIMA, ARFIMA)
  - Precise reproduction of the data set behavior
  - Too slow for real-time computational use
- Qualitative models
  - Based strictly on the qualitative behavior of the load
  - Example: Use of derivative values to define load behavior as increasing, decreasing or oscillating
Trend Modeling

- **Major properties**
  - Linear model simplicity
  - Match the variability pattern of the data set
  - Noise smoothing

- In addition to actual observed values in the system, try to predict the load trend
Qualitative Trend Model

- Evaluate the behavioral trend of the data
- Consider at least $m \geq 2$ selected points consisting of the most recent observed data point, $x_i$, and $m-1$ additional points according to a selection frequency.
- For 2 points, we can define 3 trend patterns between them: increasing, decreasing and stable
For $m > 2$, we can define up to $3^{m-1}$ classes of behavior, but we consider only five: increasing, unstable+, stable, unstable-, and decreasing

- unstable+ and unstable- are both oscillating patterns
- unstable+ ends with increasing behavior, unstable- ends with decreasing behavior

They select these five because their evaluations indicated that using more classes did not improve performance.
Qualitative trend modeling does not give us enough information about the state of the system to make accurate predictions.

- When $m = 2$, transition from qualitative to quantitative by taking the gradient of the segment between the two points.
- When $m > 2$, a linear combination of past values can be used to evaluate the predicted behavior.
Quantitative Trend Model

For each pair of successive points selected, compute the trend coefficient

\[ \alpha_j = \frac{x_{i-j} \frac{n}{m} - x_{i-(j+1)} \frac{n}{m}}{\frac{n}{m}}; \quad 0 \leq j \leq m - 1 \quad i < m \]

Then compute a weighted linear regression of the trend coefficients.

\[ \bar{\alpha}_i = \sum_{j=0}^{m-1} p_j \alpha_j; \quad \sum_{j=0}^{m-1} p_j = 1 \]
In a model based solely on this quantitative trend analysis, the \( i \)-th load state \( S_i \) would be equal to the \( i \)-th weighted linear regression value.

This can be extended to incorporate the actual load observations though:

\[
S_i = \bar{\alpha}_i + \sum_{j=0}^{n-1} q_j x_{i-j}; \quad \sum_{j=0}^{n-1} q_j = 1
\]

*It appears that \( q \) is simply another set of weights to be applied to the load observations to compute a weighted moving average. This is not clearly specified.*
Experimental Setup

- Use of three trend-aware models
  - Qualitative-only
  - Quantitative-only
  - Quantitative plus moving average
  - All trend-aware models use $m = 3$ sample points

- Two traditional models
  - Instantaneous state representation
  - Moving average over most recent 10 points
Experimental Setup

- **Performance factors**
  - 90-percentile of response time for an entire Web request
  - Level of load balancing among the servers (Load Balancing Metric)
    - LBM is normalized from 1 to $N$ where $N$ is the number of servers used for balancing
    - Smaller LBM is better.
Experimental Results

Traditional moving average model
Experimental Results

Qualitative-only Model
Experimental Results

Quantitative-only Model
Experimental Results

Qualitative with load representation
Experimental Results

The qualitative and quantitative metrics perform about at parity, and out-perform the traditional models, but the quantitative with load representation outperforms these models.

<table>
<thead>
<tr>
<th>Support for load-aware decision algorithms</th>
<th>LBM</th>
<th>90-perc. response time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure-aware model</td>
<td>2.15</td>
<td>501ms</td>
</tr>
<tr>
<td>Moving average model</td>
<td>1.85</td>
<td>450ms</td>
</tr>
<tr>
<td>Qualitative trend model</td>
<td>1.40</td>
<td>417ms</td>
</tr>
<tr>
<td>Quantitative trend model</td>
<td>1.41</td>
<td>410ms</td>
</tr>
<tr>
<td>Quantitative trend model with load rep.</td>
<td>1.29</td>
<td>370ms</td>
</tr>
</tbody>
</table>
Conclusions

- The model proposed in the paper improves the 90-percentile response time of the web services and improves the load balancing.

- No significant difference between quantitative-only and qualitative-only.

- Inadequate experimentation—a single one hour run per model is not sufficient to validate the behavior of the model under widely variable load patterns, and no confidence intervals on performance metrics.

- Graphs are unclear—what is the 'lag' parameter these are being graphed against? Is it the time from run-time for the model to adjust? If so, this is just graphed against sampling intervals, and "time" would be a better label for the axis.