3D-04

マクロな動作状況に着目したWebサーバの異常状態の予測と制御

谷村 優介† 笹井 一人‡ 北形 元‡ 木下 哲男‡

†東北大学大学院情報科学研究科 ‡東北大学電気通信研究所

1 はじめに

仮想化技術の成熟に伴い，ネットワークサービスを容易に構築し，需要に応じて計算資源の調整を行うことでサービスをスケールさせるような，柔軟なサービス運用が可能となった．仮想化された計算資源に基づいて構築されたサービスの運用には，サービス品質 (QoS) を維持しつつ，計算資源を最大限に利用した効率的な資源管理を行うことが重要であるが，サービスのスケールに伴ってサーバの動作特性が頻繁に変化するため，QoS の急激な低下を予測し，計算資源の調整を行うタイミングを適切に判断することは困難である．

本稿では，ネットワークサービスの中でも Web サービスに着目し，Web サーバのマクロな動作状況の分析に基づいてサービス異常の予兆を検知する手法を提案する．本手法の有効性を確認するためにシミュレーション実験を行い，本手法に基づいて異常状態の予兆を検知可能であることを示す．さらに，試作システムを用いた実験より，本手法に基づいて実際に異常状態を回避可能であることを示す．

2 関連研究

近年普及している仮想化ソフトは，計算資源の使用率に基づいて計算資源の調整を行うことができる [1]．また，使用率を指標とした手法では，仮想化環境において計算資源を最大限に利活用することが困難であるという考えから，よりサービス運用の実態に即した調整を行うため，QoS を指標とした手法が提案されている [2, 3]．このように計算資源の使用率や QoS を指標とした手法が提案されている一方で，これらの手法ではサーバの動作特性を調査するなどの事前準備が必要となるため，サービスの拡張性が制限されることが指摘されている [4]．サービスのスケールを行うたびにサーバの動作特性を測定することは実運用上困難であり，多様な環境に適用できる，汎用的な手法が必要である．

3 マクロな動作状況に着目したWebサーバの異常状態の予測と制御

本稿では，マクロな動作状況に着目した Web サーバの異常状態の予測と制御手法を提案する．本提案の特徴は次の 2 点である．

- スループットやレスポンスタイムなどのサービスの動作状況，すなわち QoS を巨視的な視点から分析することで，QoS そのものからは予測することが難しい，急激な QoS 低下の予兆を検知する
- サービスの異常状態の予兆検知に際して，サーバの動作特性の測定に基づく予測モデルを用いないため，事前準備が不要であり，サービスの柔軟なスケールを妨げない

本手法では Web サーバのマクロな動作状況の分析として，Web サーバのスループットのゆらぎ成分を用いる．スループットのゆらぎ成分は，スループットの計測値から移動平均との差分として求めることができる．ゆらぎ現象については，相転移現象の臨界点付近でゆらぎの分散が増大することが知られている [5]．また，待ち行列モデルに基づく Web サーバの性能分析の研究より，Web サーバの QoS の低下現象は相転移的に発生することが示されている [6]．そこで本研究では，破局的に発生する Web サービスの QoS の低下現象についても，その異常状態の予兆としてスループットのゆらぎ成分の分散が増大すると予想し，これを確認するためにシミュレーション実験を行った．

4 実験1：ゆらぎに基づく異常状態の予測

Web サーバの異常状態が発生する予兆としてスループットのゆらぎ成分の分散が増大するという仮説を検証するため，Web サーバに負荷をかけながらスループットを計測する実験を行った．本実験では，Web クライアントから 1 秒ごとに HTTP リクエストを発生させ，徐々にリクエスト数を増加させることで Web サーバに負荷を与えた．リクエスト数はポアソン分布に従う乱数に基づいて 1 秒ごとに決定した．Web サーバとして，CPU4 コアの仮想マシン 1 台を使用し，データベースと連動して動的に Web ページを生成する Web アプリケーションを配置した．Web クライアントには十分な性能をもつ仮想マシンを使用し，HTTP リクエストを生成するために Apache Bench [1] を使用した．

4.1 実験結果

Web サーバのスループットの計測結果，スループットのゆらぎの分散をそれぞれ図 1， 2 に示す．図 1 より，リクエスト数 (Concurrency Level) が増大した際に，破局的なスループット低下現象の発生が確認できる．ここでスループット計測値の移動平均 (10 点) をとることで平準化し，各計測値についてそれぞれ平均値との差をとることでゆらぎ成分を算出した．算出したゆらぎ成分の分散 (10 点) を求めたものが図 2 であり，スループットが急激に低下する 12:30 頃に向けて，ゆらぎの分散が増大する現象が確認できる．

実験結果より，破局的に発生する Web サービスの異常状態について，その予兆としてスループットのゆ

Anomaly Prediction Based on the Analysis of Macroscopic Behavior of Web Server Systems
Yusuke TANIMURA†, Kazuto SASAI‡,
Gen KITAGATA‡, and Tetsuo KINOSHITA‡
†Graduate School of Information Sciences, Tohoku University
‡Research Institute of Electrical Communication, Tohoku University

[1] http://httpd.apache.org/docs/2.4/programs/ab.html

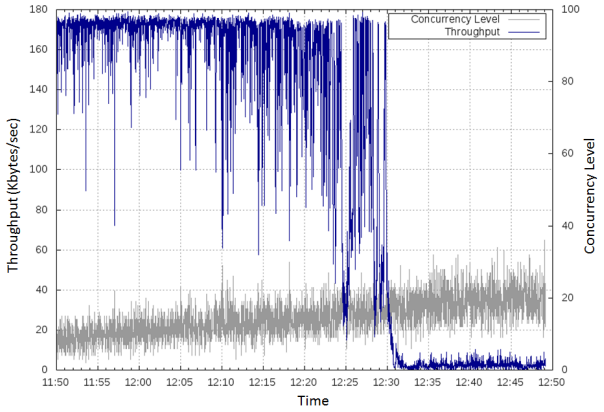

図 1: 負荷の推移とスループット計測値 (実験 1)

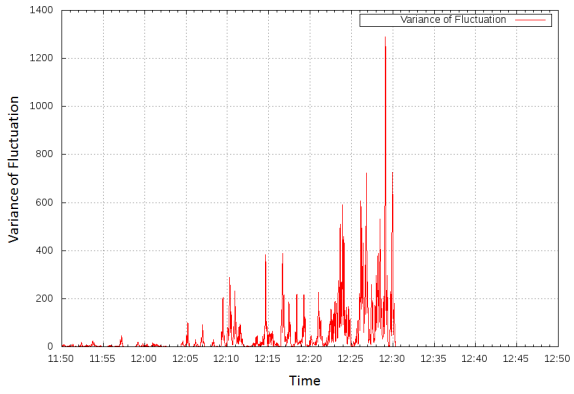

図 2: スループットのゆらぎの分散

らぎ成分の分散が増大することを確認した．これより，Web サーバのマクロな動作状況としてゆらぎの分析を行うことで，事前の動作特性の測定に基づく予測モデルを用いずとも，破局的に発生する Web サーバの異常状態の予兆を検知可能であることを示した．本節の検証に基づき，オンラインでスループットのゆらぎを分析し，ゆらぎの分散の増大を検知した場合にアラートを発報する仕組みを試作した．次節にて試作システムを用いた実験を行い，異常状態を回避可能であるかを検証する．

5 実験 2 : 予兆検知に基づく異常状態の回避

実験条件は前節と同様であり，アラートが発報された際にはロードバランサを介して CPU1 コアの仮想マシンを 1 台ずつ追加することでサービスのスケールアウトを行う．また，本実験では一定時間 (180s) にゆらぎの分散が閾値 (200) を 3 回超えたらアラートを発報するように設定した．

5.1 実験結果

Web サーバのスループットの計測結果を図 3，スループットのゆらぎの分散とアラート発報ログを図 4 に示す．図 3 より，Web サーバの負荷が大きくなった場合でも，計算資源を逐次追加しながらサービスの正常状態を維持したことが確認できる．計算資源の追加は 2 度行われており，追加のタイミングは図 4 に示すアラートの発報ログと対応している．1 度目のアラートを受けて計算資源の追加を行った後も，更なる負荷の増大によって再度アラートが発報されたことが分かる．以上の結果より，ゆらぎの分析による予兆検知に基づいて計算資源を追加することで，実際に Web サーバの異

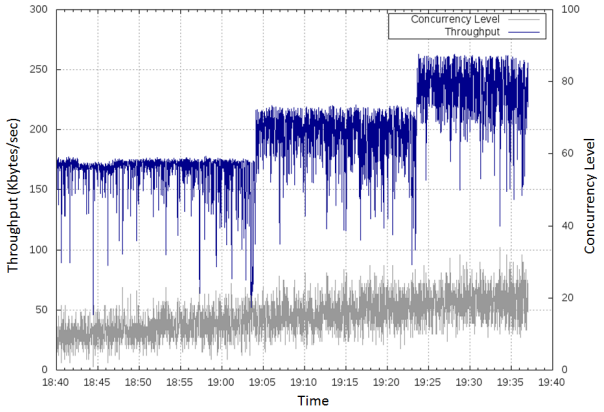

図 3: 負荷の推移とスループット計測値 (実験 2)

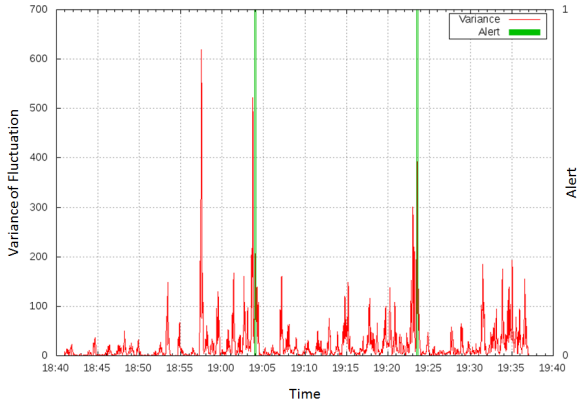

図 4: ゆらぎの分散とアラートの発報ログ

常状態を回避可能であることを確認した．また，サービスのスケールアウトが発生した後も同様に異常状態の予兆検知が可能であることを示した．

6 まとめ

本稿では，Web サーバのマクロな動作状況としてスループットのゆらぎ成分の分析を行うことで，事前にサーバの動作特性の測定を行わずとも，破局的に発生する異常状態の予兆を検知可能であることを検証した．また，試作システムを用いた実験より，本手法に基づいて異常状態を回避可能であることを示した．今後は本手法を応用し，適切なスケールインのタイミングを計る手法について研究を進めたい．

謝辞 本研究の一部は，科研費 No.16K00292 の支援を受けて実施された．

参考文献

[1] A. Gulati, G. Shanmuganathan, A. Holler, and I. Ahmad, “Cloud-Scale Resource Management: Challenges and Techniques,” in Proceedings of HotCloud 2011, pp. 1–6, 2011.

[2] A. Cockcroft, “Utilization is Virtually Useless as a Metric!,” in Proceedings of CMG ’06, 2006.

[3] T. Lorido-Botran, J. Miguel-Alonso, and J. A. Lozano, “A Review of Auto-scaling Techniques for Elastic Applications in Cloud Environments,” Journal of Grid Computing, vol. 12, no. 4, pp. 559–592, 2014.

[4] A. R. Hummaida, N. W. Paton, and R. Sakellariou, “Adaptation in cloud resource configuration: a survey,” Journal of Cloud Computing, vol. 5, no. 1, pp. 1–16, 2016.

[5] 安久 正紘, 寺町 康昌, 山中 一雄, 住谷 正夫, “ゆらぎ現象の工学的応用について,” 応用物理, vol. 61, no. 7, pp. 690–697, 1992.

[6] K. M. Elleithy and A. Komaralingam, “Using a Queuing Model to Analyze the Performance of Web Servers,” in Proceedings of SSGRR2002w, 2002.