O-038

不規則な監視情報を処理するための二段フィルタを用いたリソースキャパシティ予測方式
Resource Capacity Prediction Method with Two-Stage Data Filter for Non-Stationary Workloads

外川 遼介† 大野允裕†
Ryosuke Togawa Mitsuhiro Oono

1. はじめに

ネットワーク経由で IT サービスを提供するクラウドサービスでは、安価なコストや迅速なサービス提供などのメリットがある一方で、可用性や性能というサービスの品質要件(サービスレベル)の管理が重要となる。特に、企業に対して提供されるクラウドサービスでは、その企業からの要求に応じたサービスレベルの目標値を指標とし、それを満たす運用管理が重要となっている。

サービスレベルの目標値を満たすため、サービス提供者には、必要となるリソース処理能力（リソースキャパシティ）の過不足のない適切な管理が求められる。サービス提供者は、サービスの提供前に、想定される需要予測に基づき、必要となるリソースキャパシティを設計する。需要予測には、本番環境よりも小規模なテスト環境を用いて収集した負荷発生時の監視情報を用いる。しかし、本番環境では、テスト時の想定とは異なる利用状況により、リソースキャパシティに過不足が生じる可能性がある。従って、サービス提供者は、本番環境の監視情報をもとにリソースキャパシティを見直す必要がある。

IT サービスの本番環境では、アクセス状況の混在や計測誤差によって、監視情報のばらつきが大きくなる。このように不規則・不安定な監視情報を用いてリソースキャパシティを予測することは、テスト環境の監視情報を用いて予測することよりも困難となる。

そこで我々は、ばらつきが大きく不規則な監視情報を用いる場合に、サービスレベルの目標値を満たすリソースキャパシティの予測性能を向上させる方式を提案する。

2. 課題

ばらつきが大きく不規則な監視情報を用いる場合、高負荷時のリソース使用率を特定し、その推移の推定式に基づいて、目標とするサービスレベル指標に必要なリソースキャパシティを予測する方法が知られている[1]。しかし、臨時処理や障害などによるリソース使用率が瞬間的に突出して高い状況が多く含まれるとリソースキャパシティは過剰に予測される。同様に、縮退運転などによって平常時よりもリソース使用率が低い状況が含まれると、リソースキャパシティは過少に予測される。さらに、リソースキャパシティの予測は、監視情報のばらつきが大きいほど不正確になるため、監視情報のばらつきに応じて安全係数を適切に設計する必要がある

従って、課題は、瞬間的に突出した極端な負荷状況や平常時よりも低い負荷状況を含んだ監視情報であっても、高負荷時の監視情報を適切に特定し、リソースキャパシティを正確に予測することである。

†NEC 情報・ナレッジ研究所 Knowledge Discovery Research Laboratories, NEC Corporations

3. 提案方式

負荷とリソースキャパシティとの真の関係を示す監視情報は、瞬発的な監視情報よりも、発生頻度が高いはずである。そこで、負荷とするサービスレベル指標とリソース使用率との頻度分布を用いて、確率的に頻度の高い監視情報を高負荷時の監視情報として特定し、リソースキャパシティを予測する方式を予測する方式を提案する

3.1 節に示す二段フィルタリングによって高負荷時の監視情報を特定し、3.2 節に示す推定方法によってリソースキャパシティを予測する方式を提案する。

3.1 二段フィルタリング

サービスレベル指標に対するリソース使用率の上位領域にある測定値をその発生頻度に基づいてフィルタリングできるよう、**1)**サービスレベル指標に対するリソース使用率の空間における測定値の頻度分布を生成し、**2)**頻度分布をもとに、頻度の高い領域を選択し（一次フィルタ）、**3)**リソース使用率の上位領域を選択する（二次フィルタ）。

1)サービスレベル指標に対するリソース使用率の空間における頻度分布の生成

サービスレベル指標(X)に対するリソース使用率(Y)の X-Y 空間を正方形の格子状に分割し、各格子での監視情報の頻度を計算することで、頻度分布を生成する。X 軸および Y 軸の分割数は、必要とする分析精度をもとに設定する。また、格子分割による頻度の偏りを排除するため、カーネル密度推定[2]を用いて頻度を計算する。提案方式では、下記に示す二次元カーネル密度推定の式を用いる。

$$\hat{f}(x,y)=\frac{1}{nh_xh_y}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h_x}\right)\sum_{i=1}^{n}K\left(\frac{y-Y_i}{h_y}\right)$$

n はデータポイント数、 h_x と h_y は各変数のバンド幅、K はカーネル関数である。例えば、トランザクション処理件数に対する CPU 使用率の頻度分布において、X 軸の値域は 0～50000、Y 軸の値域は 0～100 と設定する。また、格子の分割数は、CPU 使用率の 1%程度の分析精度を想定し、100 と設定する。

2)サービスレベル指標の階級毎のリソース使用率の高頻度領域の選択

1)で生成した頻度分布からサービスレベル指標(X)の階級毎に、閾値を超える頻度のリソース使用率の領域を選択する。頻度の閾値は、ノイズの発生頻度をもとに設定する。例えば、トランザクション処理件数に対する CPU 使用率の頻度分布の場合、ノイズの発生頻度を 5%と想定

し、頻度の閾値を 95%とする。そして、X 軸の階級毎に 95%以上の頻度がある Y 軸の領域を選択する。

3)サービスレベル指標の階級毎のリソース使用率の上位領域の選択

2)で選択したサービスレベル指標(X)の階級毎のリソース使用率の高頻度領域のうち、リソース使用率(Y)の値が高い領域を上位領域として選択する。そして、上位領域に該当する測定値を高負荷時の監視情報とする。例えば、トランザクション処理件数に対する CPU 使用率の頻度分布において、2)の処理で、X が 0～500 の階級で Y が 0～10 の領域を選択した場合、Y が 10 の領域を選択する。選択した領域の測定値を高負荷時の監視情報とする。

3.2 リソース使用率の推移推定方法

二段フィルタリングによって特定された高負荷時の監視情報をもとに、サービスレベル指標(X)に対するリソース使用率(Y)の推移を示す近似式を求める。また、二段フィルタリングによって特定した高負荷時の監視情報の標準偏差値を求め、安全係数として用いる。そして、サービスレベル指標の目標値と近似式と安全係数から、必要とするリソースキャパシティを予測する。

スケーラビリティがある本番環境では、サービスレベル指標の増加に伴いリソース使用率も線形に増加する傾向にある。従って、リソース使用率の推移の近似式として増加もしくは維持する一次線形式（傾きと切片が 0 以上）を用いる。

一次線形式の推定には、サービスレベル指標に対してリソース使用率が増加もしくは維持する一次線形式を近似できるように、ハフ変換による直線検出法[3]を用いる。なお、安全係数として、高負荷時の監視情報の標準偏差値を近似式の切片に加える。

4. 評価

4.1 実験環境

実験環境として、ユーザ数 1,000 程度を処理可能なトランザクション件数 75,000Tx/sec を想定し、ロードバランサ 1 台、WEBAP サーバ 2 台と DB サーバ 1 台から成るシステムを用いた。また、クライアント端末を用いてシステムにアクセスするユーザ数、およびアプリケーションのトランザクションタイプ（参照系の更新系の比率）が異なるトランザクションを発生させ監視情報を取得した。

4.2 評価方法

クライアント端末を用いるユーザ数を 100 から 500 とし、タイプが異なる 2 種類のトランザクション（高負荷、低負荷）を同じ割合で混在して発生させた際に、収集した監視情報（トランザクション処理件数の集計値とサーバの CPU 使用率）を、評価対象の監視情報として用いた。なお、クライアント端末を用いるユーザ数を 750、875、1,000 とし、評価対象と同様の 2 種類のトランザクションを発生させて、収集した監視情報を用いて算出したサーバの CPU 使用率の 95%信頼区間を、正解とする監視情報とした。評価対象から算出した予測値と正解とする監視情報を比較した。

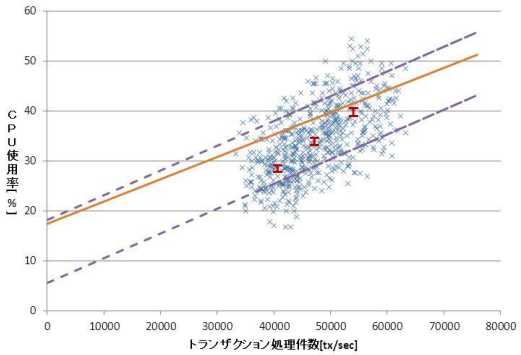

図 1　リソースキャパシティ予測結果

4.3 評価結果

図 1 に、ユーザ数 750、875、1,000 における、WebAP サーバのトランザクション処理件数に対する CPU 使用率の監視情報の分布と統計値、および各方式の予測結果を示す。実線は提案方式の予測結果、上部の破線は最小二乗法の予測値を 1.96 倍の監視情報の標準偏差で補正した予測結果、下部の破線は最小二乗法の予測結果である。エ印は、左側から、それぞれユーザ数 750、875、1,000 における、CPU 使用率の 95%信頼区間である。×印は、CPU 使用率の測定値である。

図 1 より、提案方式による予測結果は、最小二乗法と比較し、傾きが小さくなる。ユーザ数 750、875、1,000 の 95％信頼区間の上限値と提案方式、最小二乗法(95%信頼区間の予測値+1.96×標準偏差)それぞれの予測値を比較すると、どちらの方式も、95％区間より大きな値を予測している。しかし、上限値と予測値の差は、提案方式の方が小さい。特に、ユーザ数 1,000 の場合、従来手法との差は、4.32、提案手法との差は 0.88 となり、提案手法の方がより適切にキャパシティを予測できることを確認した。

5. おわりに

本稿では、本番環境のばらつきが大きい監視情報において、サービスレベルの目標値を満たすリソースキャパシティを予測する方式を提案した。提案方式は、リソース使用率の高い領域にある測定値を頻度に基づいて選択することで高負荷時の監視情報を特定し、その監視情報の標準偏差値を安全係数としてサービスレベル指標の目標値に必要なリソースキャパシティを予測する。

業務アプリケーションの振る舞いを模倣したベンチマーク環境による評価により、提案方式は、高負荷時の監視情報を適切に特定でき、最小二乗法による方式よりも余剰が少ないリソースキャパシティを予測することを確認した。

参考文献

[1]Mylavarapu, S., Sukthankar, and V., Banerjee, P., “An Optimized Capacity Planning Approach for Virtual Infrastructure Exhibiting Stochastic Workload”, Proc. Symposium on Applied Computing, ACM Symposium, (2010).

[2]Izenman, A. J., "Recent Developments in Nonparametric Density Estimation", Journal of the American Statistical Association, Vol. 86, No. 413, (1991)

[3]Dida, R. O., and Hart, P. E., “Use of the Hough transformation to detect lines and curves in pictures”, Magazine Communications of the ACM, Vol. 15, Issue 1, (1972).