クラウドコンピューティングにおける複数区間のワークロード予測に基づく仮想マシン配置手法

合田　和樹　【 分散処理 OS 研究室 】

A Virtual Machine Placement Method Based on Multi-Interval Workload Prediction in Cloud Computing

GODA, Kazuki　【 Distributed System and Operating System Lab. 】

1　はじめに

クラウドサービスにおけるサービスレベルアグリーメント（SLA）遵守には，負荷変動に応じた動的なリソース配置が不可欠である．特にホストのリソース枯渇を防ぐため，過負荷を事前に予測し，VM を移動させるライブマイグレーションが活用されている．関連研究では，統計的手法などを用いた負荷予測に基づく VM 配置手法が提案されてきた [1]．しかし，短期の予測結果に基づく従来手法は，突発的な負荷変動や予測誤差に過敏に反応し，SLA 違反を十分に抑制できない場合や，不要なライブマイグレーションの増加が課題である．そこで本研究では，SLA 違反の低減を目的として，短期的予測と長期的予測を考慮した VM 配置手法を提案する．

2　システムモデル

(1) データセット

配置手法の有効性について評価を行うためにトレンドなどの周期性を持ったデータセットの作成を行う．具体的には，バッチ，Web，データベース，アイドル，予測困難といった異なる周期や特徴を定義することで，それぞれ特徴を持ったデータの作成を行う．

(2) ワークロード

ワークロードが日次や週次といった周期的な需要変動を持つと仮定する．これは，多くの Web サービスや企業システムにおいて一般的に観測される特性であり，本研究では，これらの特性のうち CPU 使用率に着目したワークロードを疑似生成し，評価に用いる．

(3) 予測モデル

本研究では，VM の将来負荷を予測するために，LSTM を用いる．LSTM は長期的な依存関係を考慮可能な再帰型ニューラルネットワーク (RNN) であり，負荷変動のトレンドを捉えることができる．

3　提案手法

本研究では，複数区間のワークロード予測を組み合わせた仮想マシンの配置・再配置手法を提案する．本研究における SLA 違反は，ホストの処理容量を超過した回数として定義する．

関連研究 [2] では，エッジコンピューティング環境を対象に，短期的なワークロード予測に基づく動的なリソース貸借手法が提案されている．本研究ではこのアプローチを参考にし，複数区間のワークロード予測を統合的に活用する MP-SLP (Mixed Prediction Short-term and Long-term Placement) を提案する．本研究の比較手法としては，短期予測区間のみを用いた配置手法を採用する．

3.1　基本的な考え方

提案手法は，複数区間の予測情報を合成し，それに基づいて配置および再配置を行う点にある．

3.2　予測

各 VM は履歴に基づき負荷予測を行い，履歴不足時は直近の値を用いる．履歴が十分な場合は短期・長期予測を行い，予測の重複部分では，各区間の大きい値を選択する．以下に具体例を示す．ここでは，説明を簡潔にするため，短期予測期間を $t_s = 3$，長期予測期間を $t_l = 6$ とし，表 1 に示す．

表 1　短期・長期の仮予測結果

	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+6$
短期予測	10	20	10	–	–	–
長期予測	20	10	30	20	50	55

このとき，各時刻で最大値を取ることで，合成波形は

$$(20,\ \mathbf{20},\ 30,\ 20,\ 50,\ 55) \tag{1}$$

となる．実際に本研究では，以下の二種類の予測を行う．

- 短期予測: 過去 60 区間 (現時刻 t から $t-59$) から，将来 $t+1$ から $t+12$ 区間の予測
- 長期予測: 過去 60 区間 (現時刻 t から $t-199$) から，将来 $t+1$ から $t+60$ 区間の予測.

3.3　配置手法

新規 VM の初期配置では，各ホスト h_i が有する CPU 容量 Cap_i と，短期予測によって推定される将来時刻に

おける負荷 $L_i(k)$ を用いる．時刻 $t+k$ における予測上の空き CPU 容量を，

$$C_i(k) = Cap_i - L_i(k) \tag{2}$$

と定義する．また，時刻が近いほど配置判断に与える影響が大きいと考え，時間減衰を表す重み関数 $w(k)$ を導入する．このとき，ホスト h_i に対する空き容量スコア S_i を次式で定義する．

$$S_i = \sum_{k=1}^{K} w(k)\, C_i(k) \tag{3}$$

新規 VM は，この空き容量スコア S_i が最大となるホストに配置される．

3.4　再配置手法

本手法では，複数区間のワークロード予測を合成した結果に基づき，過負荷状態にあるホストから余裕のあるホストへ VM の再配置を行う．各ホスト h_i について，現在時刻または短期予測により得られるピーク負荷 L_i^{peak} と，予測誤差を考慮した安全係数 α を用い，余剰容量 R_i を次式で定義する．

$$R_i = Cap_i - \alpha \cdot L_i^{\mathrm{peak}} \tag{4}$$

$R_i < 0$ となるホストは，不足量が最大のものから順に処理する．移動対象の VM は，当該ホスト上の VM のうち，短期予測におけるピーク負荷が最大となるものとする．

移動先ホストは，$R_i > 0$ を満たすホストを対象とし，VM を仮割り当てした際の予測負荷が容量制約を満たすかを評価する．各候補ホストは，短期・長期予測結果に基づき，以下のスコアに分類される．

- Score 0：短期予測で違反あり（移動不可）
- Score 1：短期予測では違反なし，長期予測で違反あり
- Score 2：短期・長期予測の双方で違反なし
- Score 3：短期・長期予測で違反はないが，予測が不十分

移動先ホストは Score 2 を最優先で選択し，Score 2 が存在しない場合に限り，Score 1 および Score 3 を代替候補とする．

4　評価

比較手法および MP-SLP を実装し，特徴の異なる複数のワークロード環境下でシミュレーション評価を行った．評価には，5 種類の VM プロファイルが均等に混在する汎用的なワークロードや，Web サーバーに偏りを持たせたワークロード等を用いた．

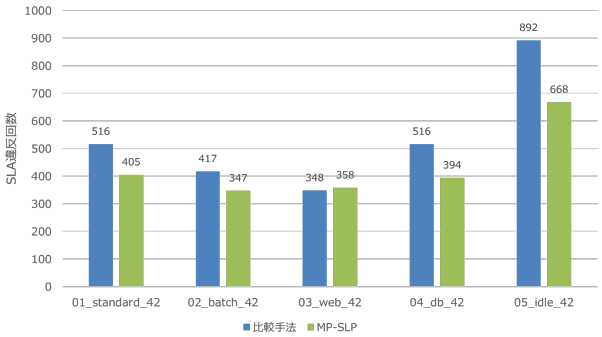

図 1　各ワークロードでの SLA 違反回数

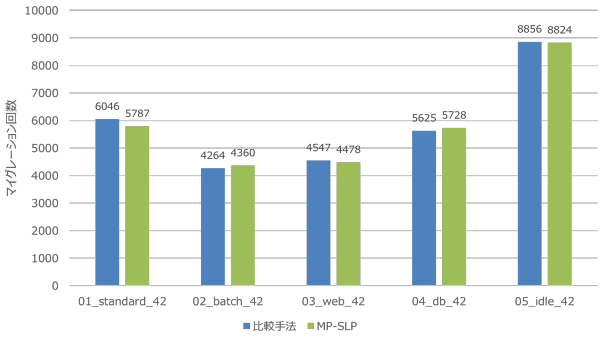

図 2　各ワークロードでのマイグレーション回数

図 1，2 より，比較手法に対し MP-SLP はマイグレーション回数を比較的同程度に抑えながら，SLA 違反回数の低減が確認できた．これは，単一区間の予測のみの比較手法がリスクを見落とすのに対し，MP-SLP は複数予測によって将来のリスクを検知し，予防的な移動を行った結果であると考えられる．一方で，一部ワークロードでは比較手法に比べて改善が見られなかった．これは，誤った予測や，過剰な VM の見積もりが蓄積した結果 SLA 違反回数が増加したのではないかと考える．

5　おわりに

本研究では，複数ワークロードの予測を用いた仮想マシン配置手法について検討し，本手法の有効性を評価した．

参考文献

[1] F. Farahnakian, T. Pahikkala, P. Liljeberg, J. Plosila, and H. Tenhunen, “Utilization prediction aware VM consolidation approach for green cloud computing,” in Proc. IEEE 8th Int. Conf. Cloud Comput. (CLOUD), pp. 381–388, 2015.

[2] 堀口 幹展，高橋 竜一，深澤 良彰，“エッジコンピューティングにおける予測を用いた負荷分散手法”，情報処理学会研究報告，Vol.2023-SE-214, No.29, pp.1-8, 2023.