複数のエッジサーバを用いた物体検知処理の負荷分散手法

1220304　江崎 洲　【 分散処理 OS 研究室 】

1　はじめに

近年, 負荷分散を目的としたエッジコンピューティングが注目されている．エッジコンピューティングでは, エンドユーザの近くに設置したエッジサーバでクライアントやクラウドの処理の一部を行うことで，計算資源への負荷を分散させることやアプリケーションのリアルタイム性を向上させることが可能となる．先行研究 [1] では，クラウドサーバ・エッジサーバ・クライアント複数を対象とした負荷分散手法が提案された．本研究では, 複数のエッジサーバを対象とした負荷分散手法を提案すると共に, 既存手法の欠点であるタスクの配置に偏りが発生する問題を解決する．

2　負荷分散手法

2.1　物体検知アプリケーションの構成

図 1 に物体検知アプリケーションの構成を示す．本研究ではエッジサーバ 3 台とデータセンタ 1 台で Kubernetes クラスタを構成し，クライアントからのアクセスをマスターノードに配置したタスク配置プログラムによってワーカーノードに割り振る．マスターノードにはタスク配置プログラムのみが起動しており，ワーカーノードには人検知アプリケーションと距離推定アプリケーションが起動している．

2.2　提案手法

既存手法では, 前回の実行時間と転送時間を計測し, データ単位あたりの実行時間を求めその値を保存し, タスクの配置は, データ単位あたりの実行時間から次のタスクの実行時間を推定し, 値が最小となるように決定している. しかし, この手法では負荷がかかりレスポンスが遅くなってしまった場合に, データ単位あたりの実行時間の更新が遅くなり, 次の実行時間の推定に古いデータが使われることになり, タスクの偏りが発生する. そこで, この問題を解決するために連続で同じサーバにタスクを割り振らないようにする. 既存手法と同じ手法で

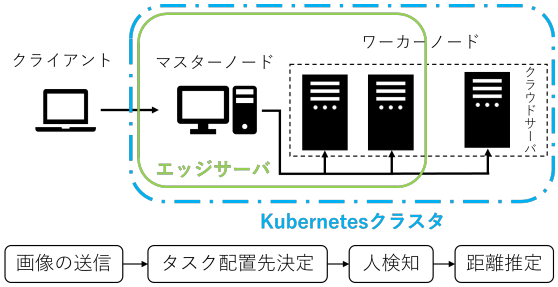

図 1　システム構成

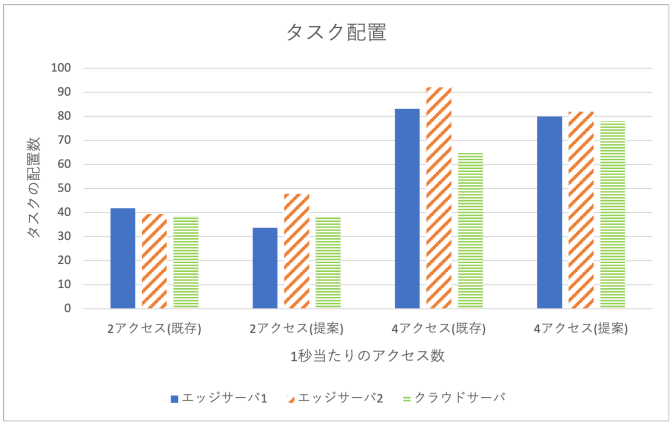

図 2　タスク配置

実行時間を推定し, タスクの配置先を仮決定したあと前回割り振ったサーバと同じであれば次に推定実行時間が小さいサーバに配置を決定する．

3　評価

タスクの偏りを比較するために既存手法と提案手法をマスターノードに配置し，クライアントからの 1 秒あたりのアクセス数を 2,4, として負荷をかけた．タスク配置の結果を図 2 に示す．結果から 4 アクセスの時には提案手法の方がタスクの偏りを抑えることが出来ていることが確認できる.2 アクセスの時は既存手法に比べて提案手法のほうがタスクが偏っている. しかし, 各アクセスに対しての平均実行時間を確認したところ, 提案手法の方が実行時間が短いことが確認できた. これは, サーバ 1 にサーバ 2,3 に比べて性能が低いマシンを使用したため実行時間が遅く, 配置先に選択されにくくなっているからだと推測する．また, 平均実行時間が短い理由として, 既存手法では連続して同じサーバにタスクを配置するため, 計算資源への負荷が集中するが, 提案手法では連続して同じサーバにタスクを配置しないため負荷が集中せず, 実行時間の低下が緩和されるからだと推測する．

4　おわりに

本研究では, 複数のエッジサーバを対象として, 既存手法の問題を連続した配置を行わないことで改善する手法を提案し，有用性を評価した．

参考文献

[1] 永元陽一，横山和俊，”エッジコンピューティング環境での顔認識アプリケーションを対象とした負荷分散機能の実現と評価”，研究報告マルチメディア通信と分散処理 Vol.2021-DPS-187,No12,2021.5