# エッジコンピューティング環境の顔認識アプリケーションを対象とした負荷分散手法

1210349　永元　陽一　【 分散処理 OS 研究室 】

## 1　はじめに

IoT アプリケーションにおいて，スマートフォンや IoT 機器などのモバイルデバイスには，処理能力やネットワーク帯域幅といった制約がある．そのため，負荷分散を目的として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，クライアントやクラウドの処理の一部を，エンドユーザの近くに設置したエッジサーバへ移動させることで，アプリケーションの遅延軽減や計算資源の負荷を軽減することができる．このタスク配置について，顔認識アプリケーションを対象とした手法が提案されている [1]．本研究では，既存手法の欠点である実行時間の推定に時間がかかる問題を解決する手法を提案する．

## 2　提案手法

### 2.1　顔認識アプリケーションの構成

図 1 に顔認識アプリケーションの構成を示す．顔認識アプリケーションは LBP 特徴量による顔認識を，顔検出，前処理，特徴量抽出＆マッチング，の 3 つのタスクに分割している．これらのタスクは，タスク配置プログラムによってクライアント，エッジサーバ，クラウドサーバに負荷分散される．なお，計算機上にはそれぞれのタスクが起動していると仮定しており，処理負荷が大きい特徴量抽出＆マッチングは，クライアント上には配置しない．また，タスク実行システムでは，提案手法に必要な実行時間等を測定するために，タスクの実行時間と以降のタスク完了にかかった時間を測定する．

### 2.2　タスク配置先決定アルゴリズム

本研究では，パラメータ収集の時間を短縮するために前回の実行結果から求めたパラメータを使用する．また，既存手法では計算能力の小さいクライアントで実行時間の推定を行っていたため，提案手法ではエッジサーバで実行時間の推定を行う．パラメータの計算に必要な実行結果は，エッジサーバ上のデータベースに収集し保存する．この際，実行時間と入力データサイズから単位データ量あたりの実行時間を，転送データサイズと転送時間から単位時間あたりのデータ転送速度を求める．タスク配置は，これらのパラメータとデータサイズ，タスクによるデータの縮小率から，各タスクの各計算資源における実行時間と各タスクごとの転送時間を推定し，全ての組み合わせの中から最小となる配置先に行う．これにより，パラメータの収集に必要な時間を短縮しつつサーバの負荷増加やネットワーク帯域の圧迫に対応できると考える．

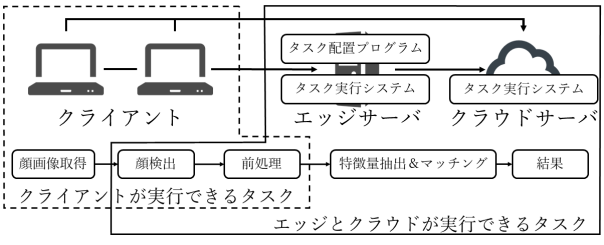


図 1　実験環境とアプリケーションのタスク

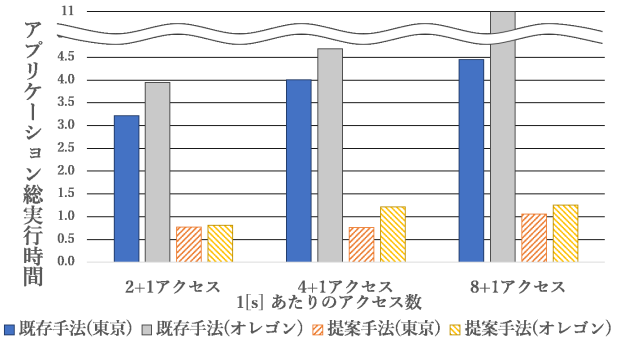


図 2　アプリケーション総実行時間

## 3　評価

実験環境としてエッジサーバをクライアントの近くに設置し，クラウドサーバは比較的近い東京リージョンとオレゴンリージョンの 2 種類を使用した．性能比較を行うために，既存手法のタスク配置プログラムの実行場所を提案手法と同じエッジサーバへ移動し，負荷用クライアントからの 1 秒当たりのアクセス数を 2，4，8 として計算資源に負荷をかけた．この状態で，評価用クライアントからアクセスを行った結果を図 2 で示す．既存手法では，パラメータの収集による計算資源を消費と帯域幅の圧迫により，アクセスが増えるとアプリケーションの総実行時間が大きくなっている．それに比べ提案手法では，アクセス数が増えたとしても総実行時間が大きく増えることなく実行できている．既存手法の総実行時間がここまで大きくなった原因として，アクセス集中による帯域幅の圧迫により，帯域幅の測定を行う iperf でエラーが頻発していることを確認した．

## 4　おわりに

本研究では, エッジコンピューティングにおける負荷分散手法として前回の実行時間をもとに実行時間を推定する手法を提案し有用性を評価した．

## 参考文献

[1] 佐竹颯太，谷遼太郎，重野寛, “エッジコンピューティングにおける顔認識アプリケーションのためのタスク配置システムの提案”，マルチメディア，分散，協調とモバイルシンポジウム 2019 論文集, pp.1190–1195 (2019).