6D-02

リアルタイムコミュニケーション向けエッジコンピューティングシステムにおけるタスク配置とリソース割り当てに関する検討

戴 競択†

TIS†

1 はじめに

近年、遠隔操作や高画質会議などのリアルタイムコミュニケーションの需要が高まっている。特に、エッジコンピューティングを用いて、ストリーミングデータをエッジサーバを通させて、コンピュータービジョンなどの高付加価値化のデータ処理を加えるサービス形態が注目を集めている。一方、エッジサーバのリソースが有限であるため、低遅延大容量という QoS の実現と高効率のサーバ利用を両立させる必要がある。本研究では、ネットワーク状況とサーバのリソース使用率とを考慮し、メディア処理タスクの配置先とサーバリソースの配分の動的決定方法を検討する。

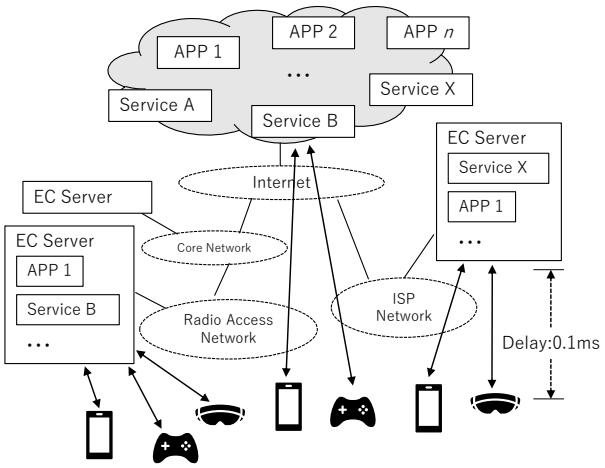

図1 アプリケーション／サービスはEdge Computing Server (EC Server)上でも実装されたため、アクセスが分散する。

2 エッジ側におけるメディア処理

AR/VR(Augmented Reality/Virtual Reality)、クラウドゲーミング、多視点映像配信、1対多／多対多ライブ配信など、サービスプロバイダーはパブリッククラウドでメディア処理サーバを設置し、様々な高付加価値なリアルタイムコミュニケーションサービスを提供する。一方、動画像、音声をはじめとして、多数のユーザ間の大量なデータがクラウドで処理される必要がある。ユーザ数の増加、サービス種類の増加、データ量の増加に応じて、クラウドでの通信負荷及び計算負荷が指数関数的に増加する。これによって、クラウドでの情報処理速度が遅くなり、レスポンスまでの遅延時間が増大する。

一方、高速・大容量化・多接続を実現する 5G ネットワークの商用化と、NFV（Network Functions Virtualization）技術の応用に伴い、MEC（Multi Access Edge Computing）[1]をはじめとして、RAN（Radio Access Network）などのユーザに近い場所に設置する汎用サーバ（エッジコンピューティングサーバ）をサービスプロバイダーに提供するというエッジコンピューティングシステムが注目を集めている。図 1 の示すように、エッジコンピューティングの導入により、汎用サーバで実装されたアプリケーション・サービスは低遅延でユーザとコミュニケーションすることができる。そして、クラウドでの通信負荷及び計算負荷も軽減される。

3 タスク配置とリソース割り当て

一方、エッジサーバのリソース（CPU、メモリ、ネットワーク帯域など）は有限であるため、高効率の利用が重要である。これはメディア処理タスクの配置方法とリソースの割り当て方法に大きく依存する。本報告で、下記 3 つのベーシックな方法を使用し、シミュレーションにおいて力任せ探索でその最適解を見つけ、それぞれの特性を明らかにする。

方法 1：同セッションにある送信者と受信者間の通信遅延を最短にするエッジサーバにメディア処理タスクを配置する。

方法 2：方法1で選出したエッジサーバにタスクを配置するとリソース使用率が 100%以上にな

Resource Allocation and Task Placement Method for Real-Time Communications with Edge Computing
† DAI Jingze, TIS Inc.

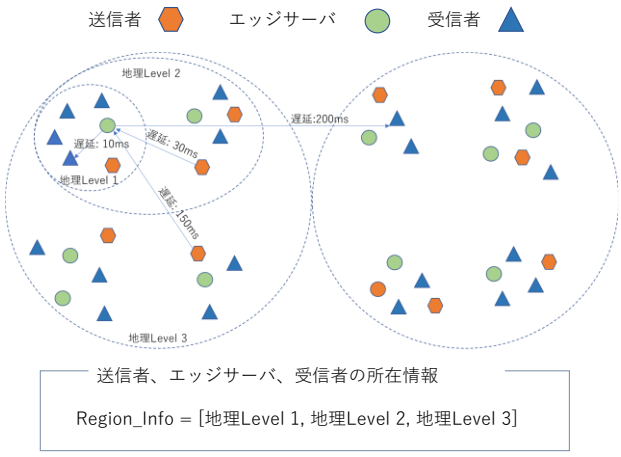

図2　システムモデル。送信者、受信者、エッジサーバの所在情報と両者間の遅延との対応関係の例

る場合、方法 1 を用いて次善を選出する。

方法 3：方法 1 で選出したエッジサーバにタスクのリソース使用率が 100%以上にならないように、リソース使用率の一番低いサーバのリソースを上記サーバに再割り当てする。（必要十分のリソースがあるのは前提である。）

3．1　シミュレーション設定

図 2 に示すように、本シミュレーションでは、送信者がエッジサーバ(1 台)経由で複数の受信者にデータストリームを送信する。送信者と受信者とエッジサーバが「所在情報」を持ち、他者との通信遅延がこの情報に基づて算出する。エッジサーバでの送信帯域幅を有限なリソースとし、扱っている全ストリームの合計送信データレートがこの帯域幅を超えないように、同じ圧縮率で全ストリームデータを圧縮する。表 1 に他の重要なパラメータ設定情報を示す。

3．2　シミュレーション結果と分析

各方法を使用したときの送信者数と受信データレートとの関係、並びに、通信遅延の平均値を図 3 に示す。（本システムでは通信遅延と送信者数は無相関である。）

システムが飽和状態になるまで、方法 2 に比べて方法 1 の平均受信データレートが低下する。その理由として、エッジサーバ及び送受信者の分布に疎密があり、送信開始時間もランダムであるため、エッジサーバが一時的に大量なアクセスを受け付けて、送信帯域幅の制限により送信データを圧縮してしまうことが考えられる。一方、方法 2 が時折次善のサーバを使うので、平均通信遅延が方法 1 のより長い。方法 3 はリソースの再割り当てを行うので、方法 1 と 2 に比べて最短な通信遅延と最大の受信データレートの両立をより長く維持できた。

表1　重要なパラメータ設定情報

送信者数	{50, 70, 90, …, 770 }台
受信者数	送信者数× 5
エッジサーバ数	20台
送信データレート	3Mbps
送信者1台の最大送信先受信者数	5台
エッジサーバの受信帯域幅	無限
すべてのエッジサーバの合計送信帯域幅	2000Mbps
エッジサーバ1台の最大送信帯域幅	方法 1 と 2 ：100Mbps 方法 3 の初期状態：100Mbps
地理Level 1の値域	{0, 1}
地理Level 2の値域	{0, 1}
地理Level 3の値域	{0, 1}

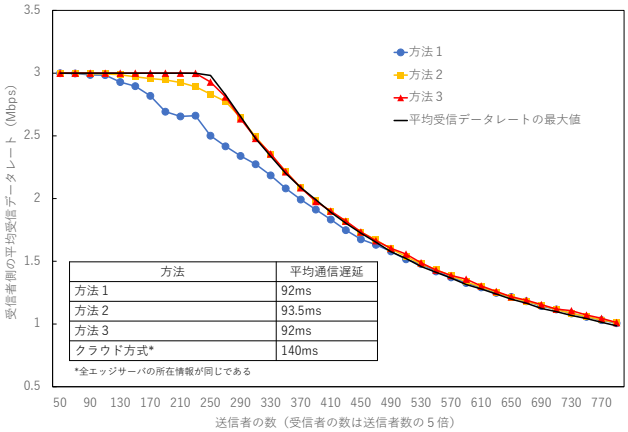

図3　各方法を用いた送信者数と平均受信データレートとの関係、並びに通信遅延の平均値

図 3 に、すべてのエッジサーバが全リソースをシェアリングし必要十分な量のみ利用し、利用完了後返還する、というリソース割り当て方法での結果（最大受信データレート）も示している。現実では、リソースの再割り当てを行う際、種々制限と対価を考慮する必要がある。

4　まとめ

本報告では、メディア処理タスクの配置先とサーバリソースの配分を決定する方法を検討した。大規模運用でリクエストを受信する都度直ちに配分決定を下す、組み合わせ爆発問題を回避した最適化アルゴリズムが必要で、今後の課題である。

参考文献

[1] ETSI - Multi-access Edge Computing - Standards for MEC, https://www.etsi.org/technologies/multi-access-edge-computing