

(続紙 1)

京都大学	博士（情報学）	氏名	張 晨露
論文題目	Service Deployment and Resource Allocation in Distributed Computing Networks (分散コンピューティングネットワークにおけるサービス配備とリソース割り当て)		
(論文内容の要旨) Today's constantly evolving real-world networks are characterized by their dynamic and large-scale nature. Distributed computing networks have been widely adopted in various real-world infrastructures as a flexible and scalable solution in response to the changing demands of real-time users, such as edge computing, cloud computing, and network function virtualization (NFV). NFV is a promising paradigm that virtualizes network functions, such as firewalls or load balancers, by decoupling them from dedicated hardware. NFV allows network functions to be implemented as software instances, referred to as virtual network functions (VNFs). Under NFV architecture, a network service can be provided by a service function chain (SFC), which consists of a chain of VNFs within a specific order, to enable flexible and agile network management. The distributed nature of NFV allows systems to handle dynamic conditions with high-quality performance, providing lower latency, higher scalability, and more flexible resource allocation. To fully leverage the benefits of distributed architectures, efficient service deployment and resource allocation strategies must be considered. This thesis studies three specific problems about service deployment and resource allocation in distributed networks. The first and second problems focus on parallelized service deployment in NFV, considering both computing and transmission resource allocation. The third problem examines the distributed server selection problem under a general distributed network, with particular attention to server resource allocation. This thesis is organized as follows. Chapter 1 introduces the background of service deployment and resource allocation in distributed computing infrastructures. Chapter 2 investigates the related work in literature. Chapter 3 proposes a parallelized service deployment model that handles the imbalance issue with considering that the delay of each VNF in the SFC depends on both arriving traffic and allocated computing resources. The proposed model determines the placement and routing of the given parallelized SFCs and minimizes the deployment cost with satisfying the end-to-end delay constraints. The technique of VNF sharing is applied to improve the efficiency of resource allocation. The cost-saving effect of applying adjustable sojourn time and VNF sharing for a parallelized SFC deployment problem is explicitly analyzed. A mixed integer second-order cone programming (MISOCP) problem is formulated to minimize the deployment cost. A heuristic algorithm is also introduced to solve larger-size problems with less computation time, where the MISOCP approach is intractable to handle. The performance of the proposed model is evaluated by comparing it with three baseline models in different experiments. The numerical results observe that the proposed model outperforms the baseline models in terms of deployment cost. Chapter 4 proposes a robust deployment model for parallelized SFCs against traffic uncertainty. The uncertain traffic is described by an uncertainty set that considers two aspects, including the variation of service traffic arrival rates and the fluctuation of parallel structures. M/M/1 queueing systems are applied to estimate the average sojourn time experienced by each VNF instance. The proposed model is formulated as an MISOCP problem with the objective of minimizing the total deployment cost, which consists of the initial cost to activate VNF instances and the proportional cost to allocate computing capacity on substrate nodes. A			

simulated annealing (SA) algorithm is introduced to obtain the solution with less computation time compared to the MISOCP approach. The numerical results show the advantages of the proposed model with its uncertainty set over the baseline models.

Chapter 5 proposes a robust distributed server selection model for multiple users in real-time applications against delay uncertainty. Both user-server and server-server link delays are considered uncertain due to the varying network conditions, which are defined by two uncertainty sets, respectively. An MISOCP problem is formulated to minimize the weighted sum of data synchronization delay and total interaction delay over the network, which represents the overall delay performance. A heuristic algorithm based on the SA algorithm is introduced to solve larger-size problems with less computation time, where the MISOCP approach becomes intractable to handle. The numerical results show that the proposed model outperforms the baseline models in terms of suppressing the delay.

Finally, Chapter 6 concludes this thesis and discusses the future works to extend this work.

(論文審査の結果の要旨)

本論文は、分散コンピューティングネットワークにおけるサービス配備とリソース割り当てに関する3つの課題に焦点を当てて研究を行っている。第1および第2の課題は、それぞれトラヒックが確定的である場合と不確定的である場合におけるネットワーク仮想化において、並列化されたサービス配備に注目し、コンピューティングと伝送リソースの割り当てを扱う。第3の課題は、一般的な分散ネットワークにおける分散サーバ選択問題を取り上げ、特にサーバのリソース割り当てに重点を置いている。本研究の成果は以下の通りである。

第1に、サービス機能チェーン (SFC: Service Function Chain) 内の仮想ネットワーク機能 (VNF: Virtual Network Function) の遅延がトラヒックの到着率と割り当てられたコンピューティングリソースに依存することに着目し、並列化サービス配備モデルを提案している。このモデルは、並列化されたSFCの配備とルーティングを最適化し、エンドツーエンドの遅延制約を満たしながら配備コストを最小化するものである。さらに、調整可能な滞在時間とVNF共有の適用によるコスト削減効果を分析している。提案された並列化SFC配備問題は、混合整数二次錐計画 (MISOCP: Mixed Integer Second-Order Cone Programming) 問題として定式化されており、大規模な問題を効率的に解くための発見的アルゴリズムも導入している。数値評価の結果、提案モデルが従来モデルよりも配備コスト削減において優れることが示された。

第2に、トラヒックの不確実性に対してロバストな並列化SFC配備モデルを提案している。不確実性はトラヒック到着率の変動および並列構造の変動を考慮した不確実性集合として記述されている。さらに、VNFインスタンスの平均滞在時間を推定するためにM/M/1待ち行列システムが採用されている。このモデルは、VNFインスタンスのアクティブ化に伴う初期コストとノード上のコンピューティングリソースの割り当てに伴う比例コストからなる全体の配備コストを最小化するように設計されており、MISOCP問題として定式化されている。また、大規模な問題に対して短時間で解を得るために焼きなまし法を用いたアルゴリズムを導入している。数値評価では、提案モデルが従来モデルと比較して、不確実性のある条件下で優れた性能を発揮することが確認された。

第3に、リアルタイムアプリケーションの複数ユーザに対して、遅延の不確実性にロバストな分散サーバ選択モデルを提案している。このモデルでは、ユーザとサーバ間およびサーバ間のリンク遅延をそれぞれ二つの不確実性集合として定義している。ネットワーク上のデータ同期遅延とインタラクション遅延の重み付き和を最小化するように、分散サーバ選択問題をMISOCP問題として定式化している。数値評価の結果、提案モデルが従来モデルを遅延抑制の観点で上回ることを明らかにしている。

以上、本論文は、分散コンピューティングネットワークにおけるサービス配備とリソース割り当てに関する最適化モデルを提案し、適応シナリオに対応した解法を提示することで、分散コンピューティングネットワーク技術の発展に寄与するものである。本論文の内容は、学術上、実用上ともに寄与するところが少なくない。よって、本論文は博士(情報学)の学位論文として価値のあるものとして認める。また、令和7年1月21日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。なお、本論文は、京都大学学位規程第14条第2項に該当するものと判断し、公表に際しては、当面の間、当該論文の全文に代えてその内容を要約したものとすることを認める。