ネットワーク機器の消費電力を削減する仮想マシン移送を考慮したネットワークトポロジ

白柳広樹† 山田浩史† 河野健二†

1. はじめに

データセンタにおいて，データセンタを構成するネットワーク機器にかかる消費電力が問題となっている．通常，データセンタではサービスの冗長性を考え Fat Tree のような冗長構成をとり，多くのネットワークスイッチが稼働している．また，現在のネットワークスイッチはトラフィック量に応じた電力効率は得られない．そのため，アイドル状態のようなほとんどネットワーク帯域を使用していない時も，トラフィック量が多い時も消費電力にあまり差がなく一定の電力を消費する．実際，2006 年の U.S. 全体のデータセンタにおけるネットワーク機器の消費電力は年間 1.9 億ドルにのぼることが分かっている[1]．その額は年々増加しており，無視できないものとなっている．

こうした問題に対処するべく，ネットワークトポロジの構成を工夫することでネットワーク機器の消費電力を削減する手法が提案されている．たとえば ElasticTree[2] では，データセンタのトラフィックを監視し，ネットワークのフローをリンクの通信容量を超えないように集約する．集約することで通信を行わなくなる不要なネットワークスイッチの電源を切る．

本研究では，仮想マシン (VM) 移送を考慮してネットワークトポロジを構成することで，ネットワーク機器の消費電力をより削減できることを示す．VM 移送は VM を稼働させたまま，他の物理マシンに移す方法である．既存手法ではトポロジの構成段階で仮想マシン移送を考慮していなかった．本研究では，データセンタ内で一般的に用いられている Fat Tree を題材とし，VM 移送を考慮してネットワークトポロジを構成する．動的に VM を移動させ，ネットワーク機器のより高い消費電力削減を狙う．

† 慶應義塾大学
Department of Information and Computer Science, Keio University

2. 提案手法

2.1 アプローチ

ここで，VM の移送を行う際にはサーバの冗長性を考慮しなければならない．データセンタでは，ラックに障害が発生した場合でも稼働し続けられるようレプリカサーバを別のラックに配置している．そのため，レプリカサーバが稼働しているラックに VM を移送すると，ラックに障害が起きた場合，そのサーバが提供しているサービスが停止してしまう．そのため，この場合にはたとえラック内に VM が 1 台のみ稼働していても，他のラックには VM を移送することができない．結果として，そのラックに接続しているネットワークスイッチの電源を切ることができない．

提案方式では，近年ネットワークスイッチのポート数が増加傾向にあることに着目し，通常の物理マシンが接続されている階層より上の階層の空いているポートに集約用の物理マシンを接続する．この物理マシンを待避用マシンと呼ぶ．待避用マシンを接続することで冗長性を考慮する上で集約できなくなってしまうケースを解消する．そして，VM がなくなった物理マシンおよびネットワークスイッチの電源を切ることで，ネットワークスイッチの冗長性は保ったまま消費電力を削減できる．

待避用マシンが特に効果を発揮する二つのケースについて説明する．一つ目は，レプリカの制約により集約ができない場合である．具体的には，ラック間で移送を行った際に，移送先にレプリカがあったために移送ができなかったようなケースである．この場合も同様に，残ってしまった VM を移送するだけの空きがある待避用マシンが存在すれば，移送することでラック内の VM をなくすことができる．

二つ目は，集約を行ったが，計算資源が足りなかったためにラック内に VM を保持する物理マシンが数台残ってしまったような場合である．ネットワークスイッチの電源を切ることを考えなければ，集約は行われていることになる．しかし，集約しきれなかった VM 分

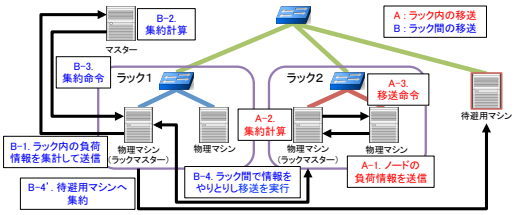

図 1 移送アルゴリズムの全体像

の空きがある待避用マシンが存在すれば，VM をそのまま待避用マシンに移送することでラック内の VM をなくすことができる．

2.2 移送アルゴリズム

本手法の効果を示すために，VM 移送アルゴリズムとして資源使用率に応じて VM を移送するという一般的な方法を採用した．全体の流れを図 1 に示す．A の流れはラック内での集約を表し，B の流れはラック間および待避用マシンへの集約を表す．まず，待避用マシンへの移送によりネットワークスイッチの電源を切ることができるか判断するためにデータセンタ全体を管理するサーバであるマスターを配置する．マスターが集約や待避用マシンへの移送の指示を行っていく．しかし，データセンタのサーバ数は数千〜数万におよぶことが多く，全てのサーバをマスターが管理するとマスターの負担が大きくなってしまう．そこで，各ラック内で 1 台ラック内の管理を行うサーバ（以下ラックマスターと呼ぶ）を配置しマスターの負荷を軽減する．ラックマスターがラック内の移送を管理し，マスターがラック間の移送を管理する．

マスターやラックマスターにおける集約や情報の収集は周期的に行っていく．集約の計算などは計算資源をある程度使用するため頻繁に行うと他の VM が使用できる計算資源が減ってしまう．そこで，データセンタでは CPU 使用率などの資源使用率は周期的に変化する場合がほとんどであるため，ある程度の周期で行っていくようにする．

3. シミュレーション実験

提案手法の省電力効果を検証するために，ElasticTree[2] と同じ実験環境である k = 12 の Fat Tree トポロジを構成し，シミュレーションを行った．k = 12 の時の構成は，総物理マシン数が 432 台，ネットワークスイッチの総数が 180 台である．トポロジの末端のネットワークスイッチとそれにつながっている物理マシン 6 台を 1 ラックとした．実際のデータセンタの 5 日間の資源使用率を測定したデータ[2] をもとにして，資源使用率を作成した．

シミュレーションはレプリカサーバ数，VM 数，資源使用率の割合そして，待避用マシン数を変えて行った．各実験において，待避用マシンの数を 0, 6, 12 台配置した場合のネットワークスイッチの電源を切ることができる時間を測定した．現実の 10 分をシミュレーションでは 1 秒として 1 週間分のシミュレーションを行った．各物理マシンの資源使用率とメモリの情報送信の周期を 10 分，ラックの情報送信，ラック内および全体の集約の計算の周期を 1 時間とした．また，各物理マシンのメモリは 8192 MB とし，各 VM に割り与えられるメモリ量は 1024 MB とした．集約および分散の計算時に用いる資源使用率の閾値は，集約時は 75%，分散時は 85%とした．

本実験では，いずれの場合も消費電力効果を確認でき，最大で約 7 % の消費電力効果を得られることがわかった．詳細のデータはポスターセッションにて発表する予定である．

4. おわりに

本研究では，ネットワークトポロジを考慮した VM の移送によるデータセンタの省電力化手法を提案した．ネットワークスイッチのポートが増加傾向にあることに着目し，本来物理マシンの接続されている階層より上の階層に新たに待避用マシンを接続し，そこへ VM を集約することでネットワークスイッチの電源を切る．

参 考 文 献

1) Richard Brown, Eric Masanet, Bruce Nordman, Bill Tschudi, Arman Shehabi, John Stanley, Jonathan Koomey, Dale Sartor, and Peter Chan. Report to congress on server and data center energy efficiency: Public law 109-431. Technical report, Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US), August 2007.

2) Brandon Heller, Srini Seetharaman, Priya Mahadevan, Yiannis Yiakoumis, Puneet Sharma, Sujata Banerjee, and Nick McKeown. Elastictree: Saving energy in data center networks. In *Proceedings of the 7th USENIX Symposium on Networked Systems Design and Implementation (NSDI '10)*, pp. 249–264, April 2010.