Vol.2011-ARC-195 No.20
Vol.2011-OS-117 No.20
2011/4/14

ネットワークトポロジを考慮した仮想マシンの移送によるデータセンタの省電力化手法

白柳広樹[†1]　山田浩史[†1]
吉田哲也[†1]　河野健二[†1]

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が消費する電力が問題となっている．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ため，ネットワーク帯域の使用量に関わらず一定の電力を消費してしまう．本研究では，VM の移送によりデータセンタ内のネットワークスイッチの消費電力を低下させる手法を提案する．提案手法では，既存のネットワークトポロジに変更を加え，レプリカサーバの配置制約やネットワークの冗長性を考慮しつつ，ネットワークスイッチ配下のマシン上で動作する VM を移送させ，不要なネットワークスイッチを作り出す．提案手法の有効性をシミュレーションで確認したところ，提案手法を用いなかった場合と比べて，最大 7% 程度のネットワークの省電力効果を得ることができた．

Reducing Power Consumption in Data Center Networks by Virtual Machine Migration

Hiroki Shirayanagi,[†1] Hiroshi Yamada,[†1]
Tetsuya Yoshida[†1] and Kenji Kono[†1]

Network switches are a major cause of the large energy consumption in data centers. Data center networks are built from a large amount of network switches whose power consumption is constant regardless of their usage. This paper describes an approach to saving the energy of network switches using a virtual machine (VM) migration technique. We migrate VMs to make unused network switches, taking into account the constraint of replica server positions. To do so, we change the network topology in a way that keeps the existing network switch redundancy. The results of our simulation say that our approach can reduce network energy consumption by up to 7%.

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．

データセンタ内の消費電力を削減する手法として，仮想マシン (VM) の移送がある．多くのデータセンタでは 1 台の物理マシンの上に複数の VM を稼働させており，1 台の物理マシンの使用率を高めている．VM の移送は VM を稼働させたまま別の物理マシンへと移すことができ，これにより，仮想マシンや物理マシンにかかる負荷に応じて，仮想マシンの集約度合を調整することができる．サービスに影響が出ない範囲で VM を集約し，不要な物理マシンを作り，その電源を落とすことで，省電力化を行うことができる．

本研究では，VM の移送によりデータセンタ内のネットワークスイッチの消費電力を低下させることを目的とする．本研究では，Fat Tree を対象に，ネットワークスイッチ配下にある VM を別のネットワークスイッチ配下へと移送することで，より多くの不要なネットワークスイッチを作り出す．これにより，できるだけ多くのネットワークスイッチの電源を切ることを狙う．

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は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．

†1 慶應義塾大学
Keio University

Vol.2011-ARC-195 No.20
Vol.2011-OS-117 No.20
2011/4/14

そこで，提案手法では，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．この物理マシンをアグリゲータと呼ぶ．アグリゲータ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．

ここで，アグリゲータ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アグリゲータが存在すれば，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 分の空きがあるアグリゲータが存在すれば，VM をそのままアグリゲータに移送することでラック内の VM をなくすことができる．

提案手法の有効性をシミュレーションを用いて確認した．Java でシミュレーション用のプログラムを作成し，実際のデータセンタを想定したワークロードを発生させて提案システムを動作させ，ネットワークの省電力効果について検証を行った．レプリカの数やワークロードの比率，VM の数，アグリゲータの数を変更しながら，ネットワークスイッチの電源を落とすことができる時間を測定した．提案システムを用いることで，単純にネットワークトポロジだけ考慮して集約を行っただけのときよりも数%〜 10 %程度のネットワークの省電力効果を得ることができた．

本論文の構成を以下に示す．２章では現在のデータセンタについて説明し，３章では提案システムについて説明し，４章と５章では実験用のシミュレーションプログラムおよび実験内容について説明する．６章では関連研究を紹介する．そして，７章でまとめを述べる．

2. データセンタ

2.1 データセンタの構成

データセンタでは仮想化を行い膨大な計算資源を管理している．仮想化技術とは，1 台の物理マシン上で複数の仮想マシン（VM）を稼働させる技術のことである．物理マシンの計算資源を分割して，VM に割り当てることで，複数の VM を稼働させることができる．

仮想化環境では，稼働中の VM 上で動作する OS やソフトウェアを停止することなく他の仮想マシンへ移し変えることが可能である[2)]．これにより，計算資源をより柔軟に活用することができるようになる．例えば，ワークロードが増加して計算資源が足りなくなった場合，計算資源に余裕のある物理マシンへ VM を移送することで負荷を緩和できる．

また，VM の移送を用いた省電力化も行われている．データセンタでは，計算資源を常に最大限に使用することはほとんどない．これは，データセンタは基本的にワークロードが急増した場合でも耐え得るように設計されるためである．また，データセンタのワークロードは時間帯により増減し，多くなる時もあれば，少なくなる時もある．ワークロードが少ない時間帯には，物理マシンの計算資源に余裕が出てくる．このようなデータセンタの性質をうまく利用して VM を集約する．VM がなくなった物理マシンは電源を切ることができ，その間物理マシンの電力を削減できる．このように，サービスに影響が出ない範囲で VM 集約し，不要な物理マシンを作り，その電源を切ることで省電力化を行うことができる．

また，データセンタは冗長構成をとることで稼働率の向上を図っている．データセンタではまれに，ワークロードが急増するということが起こり得る．もし，急増によりシステムがダウンしてしまうと莫大な被害を被ってしまう．例えば，通販サイトである Amazon では，サービスが 1 時間停止すると約 2000 万円の損害を被ると言われている[3)]．そのため，基本的にデータセンタはワークロードが急増した場合にも耐え得るように設計される．

冗長構成の一つとして，データセンタでは一般的に 図 1 のような Fat Tree トポロジを用いていることがあげられる．Fat Tree トポロジは，通常の Tree 構造のルートの部分を冗長化したものである．通常の Tree 構造はルート付近にトラフィックが集中しやすいという欠点がある．そこで，冗長化によってネットワークの混雑を軽減するようにしたものが Fat Tree トポロジである．ネットワークが冗長化されているのでいずれかのネットワークが使えなくなっても別の道をたどって通信を継続することが可能である．

2.2 ネットワークスイッチ

現在のネットワークスイッチについて説明する．第一に，近年，ネットワークスイッチの性能向上に伴い，スイッチ 1 台当たりのポート数が増加してきている[4)]．そのため，大規模なデータセンタの構築が可能となったり，今まではポートが少なくて構築できなかったようなトポロジを組むことも可能となっている．

第二に，現在のネットワークスイッチではネットワークの使用率に応じた電力効率は得られない．すなわち，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

Vol.2011-ARC-195 No.20
Vol.2011-OS-117 No.20
2011/4/14

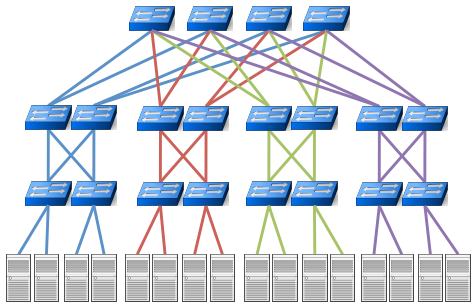


図 1 Fat Tree トポロジ

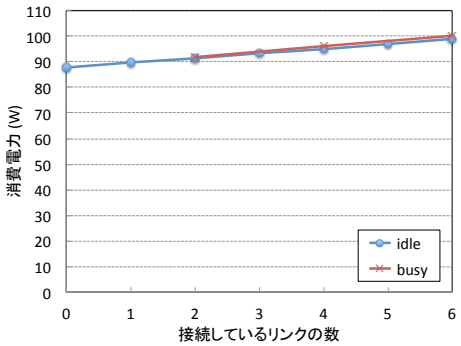


図 2 ネットワークスイッチの消費電力

も，トラフィック量が多い時も消費電力にあまり差がない．

実際に，CISCO のネットワークスイッチを用いて実験を行った．ネットワークスイッチに接続する物理マシンの数を 0 〜 6 台と増やしていき，その時の何もトラフィックを発生させなかった場合と，最大容量と同じだけのトラフィックを発生させた場合の消費電力を測定した．

結果を図 2 に示す．横軸が接続しているリンク数で，縦軸がその時のネットワークスイッチの消費電力である．idle は何もトラフィックを発生させなかった場合，busy は最大容量のトラフィックを発生させた場合である．図 2 を見て分かるように，何もトラフィックが発生していない idle 状態でも最大容量だけトラフィックを発生させた時と消費電力にほとんど変化がなく，数 W 増加しただけであった．また，新たに物理マシンを接続したときのネットワークスイッチの消費電力の増加量は 1 台あたり約 2 W であった，これより，ネットワークスイッチは理想的な直線は描かず，idle 状態でも電力を相当消費してしまうことが分かる．

2.3 データセンタにおける制約

データセンタではネットワーク上にあるサーバと同じ役割を果たすレプリカサーバ複数台配置していることがほとんどである．この冗長構成をとることにより稼働率の向上をはかっている．たとえば、突然サーバに障害が発生した場合でもレプリカサーバが別に稼働していれば，代わりに処理を行うことで継続してサービスを提供できる．

レプリカサーバを配置する場合には，元のサーバとレプリカサーバを別のラックに配置することが望ましい．ラックとは複数のサーバを集めた一つのまとまりであり，障害が発生した場合はラック内全てのサーバに影響する可能性が高いためである．例えば，ラックからつながっているネットワークスイッチに障害が発生してしまった場合は，ラック内の全てのサーバが通信を行うことができなくなる．また，電源はラックごとに管理されていることが多く，電源に障害が発生した場合はラック内の全てのサーバの電源が切れてしまうこともあり得る．もし，元のサーバとレプリカサーバが別のラックで管理されていれば，どちらか一方は障害を回避することができる．そのため，新たにレプリカサーバを立ち上げる際や VM を移送する際にはレプリカサーバの有無を意識し，冗長性を損なわないようにする必要がある．

3. 提　　案

3.1 概　　要

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する．ネットワークスイッチの電源を切ることができるように，ネットワークスイッチ配下にある VM を別のネットワークスイッチ配下へと移送する．しかし，データセンタの冗長性を考えると，集約しきれないケースが出てくる．例えば，移送を行おうとした際に移送先にレプリカサーバが存在すると移送できない．これは，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は別のラックに配置することが望ましいためである．そこで，提案システムでは，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**アグリゲータ**と呼ぶ．アグリゲータ 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り，消費電力を削減する．

3.2 アグリゲータ

アグリゲータ を新たに接続した場合，その分の電力を余分に消費してしまう．そこで，通常時はアグリゲータ の電源を切っておき，もし，アグリゲータ があればラック内の VM をなくすことができるなら，アグリゲータを起動して集約を行う．見た目上はネットワークスイッチ配下で稼働していた物理マシンが上の階層に移動したように見える．すなわち，物理マシンの消費電力は変わらないままネットワークスイッチの消費電力だけをそのまま削減することができる．

ここで，アグリゲータ 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，

Vol.2011-ARC-195 No.20
Vol.2011-OS-117 No.20
2011/4/14

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落とす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 分の空きがあるアグリゲータ が存在すれば，制約を満たしつつ，VM をそのままアグリゲータ に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，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アグリゲータ 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．

このようなケースの時にアグリゲータ がある場合，アグリゲータ へ残った VM を移送してしまうことでレプリカの制約やネットワークの冗長性を満たしつつラック内の VM をなくすことができる．同時にネットワークスイッチの電源を切ることができるので効率よく電力を削減できると考えられる．

4. 移送アルゴリズム

4.1 全 体 像

全体の流れを図 3 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アグリゲータへの集約を表す．まず，アグリゲータ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．マスターが集約やアグリゲータ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．

各物理マシンにおいて，その上で稼働している VM の情報を収集する．今回は VM の移送を指示する指標として，多くの手法[5),6)] で用いられている資源使用率を採用し，物理マシンの負荷状況が把握しやすい CPU 使用率と各 VM に割り当てられるメモリ使用量を用

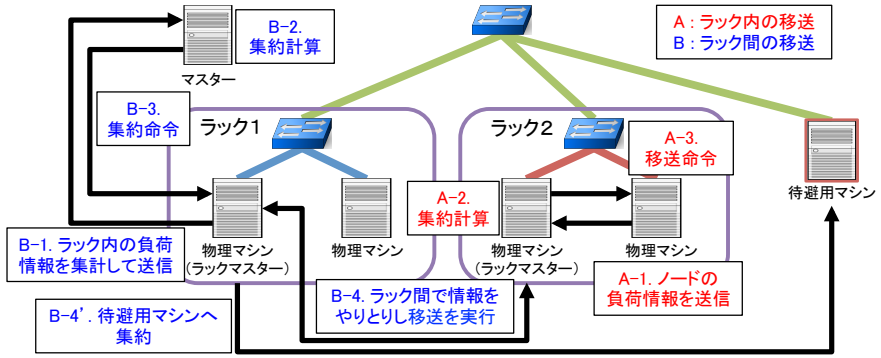


図 3 移送アルゴリズムの全体像

いる．各 VM の資源使用率および物理マシンのメモリ量は周期的にラックマスターへ送信される．

ラックマスターの挙動について説明する．情報を受け取ったラックマスターは決まった周期で集約および，負荷分散の計算を行う．計算によって移送が決定すると，それらの物理マシンに移送の指示を行う．さらに，ラックマスターは周期的にラック内の VM に関する情報をマスターへ送信する．全ての情報を送るとデータ量が膨大になってしまうので，ラック内の情報をまとめてから送る．まとめる情報は，ラック内の物理マシン数，稼働している物理マシン数，物理マシンのメモリ量の合計，使用しているメモリ量の合計および資源使用率の平均である．

マスターは，ラックマスターから受け取った情報をもとにラック間の移送を指示する．稼働物理マシン数，メモリ量，資源使用率から移送元のラックと移送先のラックの候補を選ぶ．候補のラックに移送元と移送先である旨を通知し，その後の集約はそれらのラックに委ねることでマスターの負荷を軽減し，スケーラビリティを高める．これと同時に，アグリゲータへの移送によりラック内の VM をなくすことができるかを判定し，できそうならばアグリゲータを起動し，アグリゲータへの移送を指示する．

移送元と移送先を通知されたラックでは，まず，ラック間でラック内の情報をやりとりする．移送は，通常の場合と同様に行っていく．しかし，ラック間の移送では，ラック内の移送とは異なりレプリカを考慮しなければならない．移送先にレプリカが存在した場合には，その VM は移送しないようにする．

Vol.2011-ARC-195 No.20
Vol.2011-OS-117 No.20
2011/4/14

4.2 アグリゲータへの移送

アグリゲータへ移送を試みる場合に二つの場合ついて説明する．一つ目は，集約の際に移送先のラックの候補が発見できなかった場合である．この時，ほかのラックは既に容量いっぱいになってしまっていると考えられる．そこで，残ってしまっている VM をアグリゲータに移送し，ラック内の VM をなくせれば，冗長性は保ったままネットワークスイッチの電源を切ることができる．

二つ目は，ラック間の集約において，移送に失敗してしまった場合である．すなわち，3.2 章で述べたような計算資源が不足していて全ての VM を移送できなかった場合か，移送先にレプリカがあったために移送できなかった場合である．この場合，ラック内に数台だけ VM が残ってしまっていることが多い．そのため，アグリゲータに移送できる可能性が高く，残ってしまった VM をアグリゲータへ移送してしまうことで冗長性を保ちつつ効率よくネットワークスイッチの電源を切ることができる．

アグリゲータを追加した場合その分の電力が増加してしまうので，通常時は電源を切っておき使用する時のみ入れるようにする．こうするとアグリゲータへの移送後，代わりに移送元の物理マシンには VM がなくなり電源を切ることができるため，全体として物理マシンの消費電力は変わらない．

4.3 ラック内の移送

各物理マシンはその上で稼働している VM の資源使用率を定期的に収集する．各 VM の CPU 使用率は時間で変化していき，メモリ量は開始時に割り当てる．そして，集めた情報を各ラック内にあるラックマスターへと周期的に送信する．

ラックマスターは受け取った情報を用いて周期的に集約を行っていく．まず，総 CPU 使用率が一番低い物理マシン内の VM をなくすことを考える．ラック内の VM をなくすをことを考えた時，資源使用率が小さい方が収容できる物理マシンを見つけやすいためである．移送する VM は，その物理マシン内の VM で一番 CPU 使用率が大きい順に計算していく．また，移送先の物理マシンとして，総 CPU 使用率が一番大きい物理マシンを選ぶ．もし，CPU 使用率が小さい VM から順に見ていく場合，はじめの内は CPU 使用率が小さいのですぐに移送先を決定できるが，CPU 使用率の大きい VM にしわ寄せがくる．例えば，最初の時点では移送できても移送を決定していく内に計算資源の空きが減少していくことで移送できなくなってしまうことが起こりうるため，CPU 使用率の大きい VM から移送先を決定していく．移送可能かどうかの条件として，

- 移送する VM の CPU 使用率と移送先の物理マシンの CPU 使用率の合計が閾値以下
- 移送する VM のメモリ使用量と移送先のメモリ使用量が物理マシンの物理メモリ量以下

の両方を満たした場合に移送可能とする．移送が決定したら，移送元の物理マシンと移送先の物理マシンに移送を指示する．

また，各ラックでは物理マシンの負荷が高くなっていないか周期的にチェックする．負荷分散時の移送先の決定に用いるアルゴリズムは基本的に集約時と同じである．異なる点は，移送する VM として CPU 使用率の低いものから順に見ていく．負荷分散時はラック内の負荷を軽減できれば良く，全ての VM を移送する必要はない．また，負荷が高い状態だとパフォーマンスが低下してしまう恐れがあるのでできるだけ素早く緩和させたい．そのため，移送先を決定しやすい CPU 使用率の低い VM から移送を考えていく．移送を進めていき高負荷な状態が軽減できれば負荷分散処理を終了する．もし，移送先が見つからなかった場合，そのラック内だけでは計算資源が足りていないと考えられる．その場合マスターにその旨を通知し，計算資源の余っているラックへの負荷分散を試みる．具体的なアルゴリズムに関しては，4.4 章で述べる．

4.4 ラック間の移送

ラックマスターは，各物理マシンから送られてきた情報を周期的にまとめて，マスターへと送信する．送信する情報は，物理マシン数，使用している物理マシン数，使用しているメモリ量，総メモリ量，CPU 使用率の平均である．ラック内の全ての情報は送らず，ある程度情報をまとめることでマスターの計算量を減らし，負荷を軽減する．

マスターはラックマスターから送られてきた情報を用いて周期的に集約の計算を行う．まず，ラック内の VM をなくすために，稼働している物理マシン数が一番少ないラックを集約するラックとして選択する．各ラックではすでに集約が行われていることを想定しているため，各物理マシンはどれも同じくらい多く計算資源を使用している考えられる．そのため，ラック内の VM をなくすことを考えた場合、物理マシン数が少ない方が効率よく移送できる可能性が高いと考えられるためである．

集約できるかの条件として物理マシン数とメモリ量を用いる．まず，ラック内に収容するには使用メモリ量分の空きがある必要がある．また，物理マシンの数を用いている理由は，集約するには使用している物理マシン分の空き物理マシンが移送先に必要である可能性が高いからである．各ラックではすでに集約が行われていることを想定しているので，使用中の物理マシンの計算資源が余っていることは期待できなく，移送元の使用している物理マシン分の空き物理マシンが移送先に必要である可能性が高いと考えられるためである．

マスターから移送元，移送先の命令を受けたラックは図 4 のように集約を行っていく．

Vol.2011-ARC-195 No.20
Vol.2011-OS-117 No.20
2011/4/14

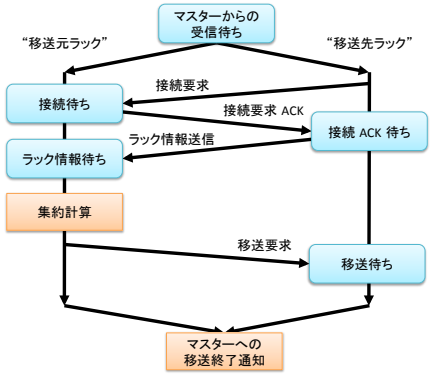


図 4 ラック間の移送における状態遷移図

まず，移送先が移送元に接続を行い，ラック内の物理マシンおよび VM の構成を送る．情報の送信が終了すると，移送元はそれらの情報をもとに集約の計算を行っていく．集約は基本的にラック内の集約で述べたものと同じアルゴリズムである．しかし，ラック間の移送ではレプリカの有無を考慮する必要があるため，移送先に移送する VM のレプリカがないという条件を加える．移送計算が終わると各物理マシンへ移送を指示する．集約が終わると各ラックはマスターへ集約が終わったことを通知して集約を終了する．もし，移送先の物理マシンにレプリカがあり移送できなかった場合，または，計算資源が足りなくて全ての VM を移送できなかった場合は集約が失敗したことをマスターに伝える．なお，ラック間の集約が行われている最中にラック内で移送をしてしまうと，ラック内の移送終了後すぐにまた別のラックへ移送することになり計算資源を無駄に使用してしまうので，この間はラック内の集約および負荷分散処理は行わない．

次に，マスターがラックマスターから高負荷状態の通知を受け取った場合の挙動を説明する．高負荷状態のラックはできる限り早く緩和させたいので，通知を受け取ったマスターはすぐに移送先の検索を始める．集約時と同様に移送中でないラックを選択し，空き物理マシンを昇順に並べ，順に次の条件を満たすかどうかを調べていく．

- 移送先の空き物理マシンが 0 でない
- 総 CPU 使用率が総物理マシン数と CPU 使用率の割合の閾値をかけた値より小さい

この条件を満たすラックに移送元，移送先の通知を行う．移送元，移送先のラックは情報をやりとりし，ラック内の移送で述べた負荷分散のアルゴリズムを用いて負荷分散を行っていく．

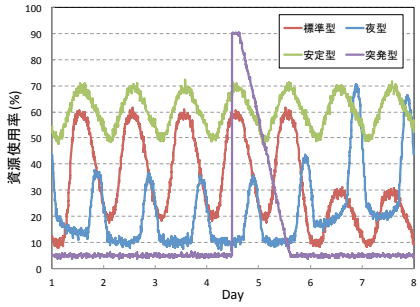


図 5 資源使用率の変化

5. シミュレーション実験

5.1 目　　的

本実験の目的は，提案システムを実際のデータセンタを想定した構成で，シミュレーションを行うことで，ネットワークスイッチの省電力効果について検証することである．

5.2 資源使用率

提案手法の有効性を検証するために，データセンタ内のワークロードを擬似的に作成した資源使用率は次の 4 つである．

- **標準型** - 平日の勤務時間に資源使用率が多くなる
- **安定型** - 資源使用率が平均的に多い
- **夜型**　- 夜に資源使用率が多くなる
- **突発型** - 突発的に資源使用率が増加する

これらの資源使用率を 1 週間分発生させた時の変化を図 5 に示す． 縦軸が資源使用率を表し，横軸は時間を表す．なお，開始は月曜日の 0 時である．

これらの資源使用率を設定した理由を述べる．まず，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[7] をもとにして資源使用率を作成し，これを標準型とした．しかし，データセンタでは様々なサーバが稼働しているため，必ずしも全てのサーバがこの変化するとは言えない．そのため，標準型の資源使用率を補足するようなものが必要と考えられる．そこで，標準型の資源使用率と反対の変動をする夜型と時間に関係なく利用されるような安定型を加えた．さらに，データセンタでは資源使用率が急増することがあるため，

Vol.2011-ARC-195 No.20
Vol.2011-OS-117 No.20
2011/4/14

表 1 レプリカサーバの数を変えた場合のネットワークスイッチの電源を切れる時間

アグリゲータ数 \ Replica	0	5	11	23	35
0	37.6%	35.4%	33.4%	27.0%	16.4%
6	37.9%	36.4%	34.8%	27.8%	17.4%
12	38.5%	37.5%	35.0%	28.8%	18.9%

表 2 資源使用率の比率を変えた場合のネットワークスイッチの電源を切れる時間

アグリゲータ数 \ Workload	100：0：0：0	25：25：25：25	50：25：20：5
0	37.6%	47.5%	35.5%
6	37.8%	48.4%	36.2%
12	38.7%	48.5%	37.5%

突発型も考慮しなければならない．

この 4 つの資源使用率を用いてシミュレーションを行っていく．標準型，安定型，夜型の資源使用率は全て図 5 のように変動する．突発型は，1 週間の内にいつ急増するかは各 VM でランダムに決定される．

5.3 シミュレーション

提案手法の省電力効果を検証するために，ElasticTree[7] 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．

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，アグリゲータ数を変えて行った．各実験において，アグリゲータの数を 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．

5.3.1 レプリカサーバ の数を変化させた場合

まず，レプリカサーバの数が異なる場合について実験を行った．レプリカサーバが何もない場合，レプリカサーバを 5 台持つ VM が 72 種類ある場合，レプリカサーバを 11 台持つ VM が 36 種類ある場合，レプリカサーバを 23 台持つ VM が 18 種類ある場合，レプリカサーバを 35 台持つ VM が 12 種類ある場合と変えて行った．用いた資源使用率は全て標準型，VM は開始時に各物理マシンに 1 台ずつ，合計 432 台を割り当て稼働させた．

結果を表 1 に示す． レプリカサーバがない場合はアグリゲータの数を増やしても 1% 程度しか変わらないが，レプリカサーバの数が増えていくと 2% 程度であるが，ネットワークスイッチの電源を切ることができる時間が増加した．レプリカサーバの数が多いと移送先にレプリカサーバが存在する可能性が高くなり，移送ができないといった状況が起きやすくなる．その場合に数台残ってしまった VM をアグリゲータへと移送できることで性能が向上したと考えられる．また，レプリカサーバの数を 24 台から 35 台へ増やした時に急激に性能が低下した．レプリカサーバの数が非常に多い場合，制約によって移送できないことが頻発すると考えられる．

5.3.2 資源使用率の比率を変化させた場合

次に，資源使用率の比率が異なる場合について実験を行った．今回実験で用いた比率は標準型，安定型，夜型，突発型の順に，100：0：0：0 と 25：25：25：25 と 50：25：20：5 の三種類とした．標準型のみのもの，基本となる標準型の比率を多く，ごくまれに発生する突発型は少なくなる設定するようにしたもの，平均的なもの 3 種類行った．レプリカサーバはないものとし，VM は各物理マシンに 1 台ずつ，合計 432 台稼働させた．

結果を表 2 に示す． 表 2 を見て分かるように，ネットワークスイッチの電源を切ることができる時間は資源使用率の比率に大きく依存する．資源使用率が様々な資源使用率があるものが一番効率が良かった．これは，集約する際に物理マシンの計算資源の残量以内に収まる VM が存在しやすく，集約の効率が非常に良くなるためであると言える．例えば，資源使用率が 60% の VM が 3 台ある場合だと，どの 2 台を合計しても 100% を超えてしまうので 3 台の物理マシンが必要であるが，資源使用率が 60% の VM が 2 台， 30% の VM も 2 台ある場合では，集約を行うことができ， 2 台の物理マシンがあれば良い．そのため，逆に，資源使用率が偏ったものは，資源使用率が少ない時はうまく集約できるが，ある程度多くなると分散してしまう．資源使用率が偏っていると計算資源余っている部分に収容できる VM の候補が少ないため効率が悪くなると考えれる．

5.3.3 VM の数を変化させた場合

次に，VM の数が異なる場合について実験を行った．VM の数を 432, 468, 504, 540 と変えて行った．資源使用率の比率は 50：25：20：5 とし，レプリカサーバの数はそれぞれ

Vol.2011-ARC-195 No.20
Vol.2011-OS-117 No.20
2011/4/14

表 3 VM の数を変えた場合のネットワークスイッチの電源を切れる時間

アグリゲータ数 \ VM	432	468	504	540
0	35.5%	30.2%	25.4%	19.8%
6	36.2%	31.0%	25.8%	20.3%
12	37.5%	31.8%	26.3%	21.5%

1, 2, 0, 0 台あるとした．

結果を表 3 に示す． VM 数が増加するとネットワークスイッチの電源を切ることができる機会が大幅に減る特に VM 数が 540 台の時は資源使用率が一番高くなる値がデータセンタの最大容量付近までいき，電源を切ることができる時間が 20% ととても低い値となった．VM を詰め込んでいるようなデータセンタではネットワークスイッチの省電力は期待できないと言える．

5.4 考　　察

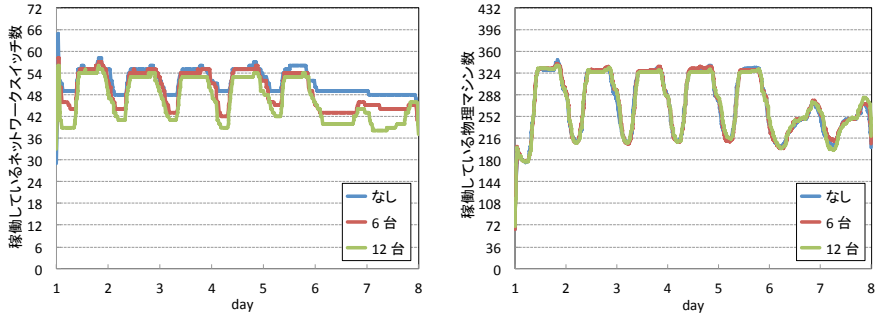
以上の実験では，省電力効果はアグリゲータがない場合に比べて 2% 程度の増加した．これは，ラック内の全ての物理マシンが稼働している状態の時に，その全ての物理マシンをそのままアグリゲータへ移し換えてネットワークスイッチの電源を落とした場合．アグリゲータが特に効果を発揮する状況というのは起こっていないと考えられる．

アグリゲータが効果を発揮するのは 3.2 章で述べたような場合，特にレプリカによる制約で移送できずに残ってしまう状況である．しかし，ラック間で集約する際にレプリカサーバのないラックを探すのだが，ラックの数が多いデータセンタではレプリカサーバのないラックが見つかる可能性が高くなる．だが，レプリカサーバの数を変化させた場合の実験において，レプリカサーバの数を多くしても期待した効果は得られなかった．これは，確かにレプリカサーバの数は多いが，その組の数も多かったために集約できないためである．例えば，本実験のレプリカサーバが 35 台ある VM が 12 組ある場合を考えると．確かにレプリカサーバがあるために移送できないことが多いが，他の 12 組も同様に移送できずに残る可能性が高い．そうなると，ラック内に数台だけ VM が残るというような状況が起きづらく，アグリゲータへ移送できない．提案システムが効果を発揮するのは一部の VM のレプリカサーバが多いような場合と考えられる．

そこで，ある VM が 1 台が 47 台のレプリカサーバを持つという状況でシミュレーションを行った．この VM の資源使用率は標準型である．また，他の VM はレプリカサーバはないものとし，資源使用率の割合は 50 : 25 : 20 : 5, VM の総数は 432 台とした．

表 4 理想時のネットワークスイッチの電源を切れる時間

アグリゲータ数	スイッチの電源を切れる時間
0	27.9%
6	31.4%
12	34.6%



(a) 稼働しているネットワークスイッチの数　　(b) 稼働している物理マシンの数

図 6 理想時の稼働しているネットワークスイッチと物理マシンの数の時間遷移

結果を表 4 に示す． 通常時より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 7% 程度まで増やすことができている．

また，この時の稼働しているラックと物理マシンの数の時間遷移を図 6 に示す． 図 6 を見て分かるように，稼働している物理マシンの数はアグリゲータの数に関わらず終始同じ数となっている．しかし，稼働しているラックの数に大きなずれが生じている．アグリゲータが何もない場合，ラックの数が 48 台を下回ることがない．これは，稼働に必要なラックが 48 台以下になっても，レプリカサーバが 47 台があるためにレプリカの制約で移送ができないからである．一方で，アグリゲータの数を 6 台に増やした場合，稼働しているラックの最小数は，42 台となっている．移送できずに残ってしてしまったレプリカサーバのあるラックをアグリゲータへと移すことができていると言える．すなわち，ネットワークスイッチ 6 台分の電力を削減できている．しかし，アグリゲータの数を 12 台に増やした場合，稼働しているラックの最小数は期待していた 36 台とはなっていない．これは，例えば，VM のレプリカサーバを保持していない状態で容量いっぱいのラックが 6 台あった場合，レプ

Vol.2011-ARC-195 No.20
Vol.2011-OS-117 No.20
2011/4/14

リカの制約があるためにどれだけ集約しても $36+6=42$ 台のラックを必要としてしまうような状況が起こっているためであると考えられる．

6. 関連研究

VM の再配置に関する研究として，pMapper[5] と Entropy[6] などが挙げられる．pMapper は各物理マシンの負荷状況を監視し，電力，移送コスト，パフォーマンスが最小となるように VM の配置を決定する．Entropy も同様に各物理マシンの CPU 使用率，メモリ量を監視する．そして，メモリ割当量と移送時間の関係に着目し，移送時間がが最小となるように VM をの配置を決定する．しかし，これらはネットワークの電力は考慮していない．

ネットワークの消費電力削減に関する研究としては ElasticTree[7] がある．ElasticTree 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ことで消費電力を削減する．本研究と同じようにネットワーク機器の消費電力を削減するが，冗長構成を考慮せずに電源を切るので冗長性が下がってしまう．また，仮想マシンの動的な配置は考慮していない．

7. 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．また，提案手法では，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アグリゲータを接続し，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．普段はアグリゲータの電源は切られており，アグリゲータへ移送することでネットワークスイッチの電源を切ることができる時にのみ起動し，VM の移送を行う．これにより，物理マシンの消費電力はそのままにネットワークスイッチの分の電力を削減することが可能である．

提案システムのシミュレーション用プログラムを Java で実装した．また，提案システムの省電力効果を測定するために，ワークロードを用意し，レプリカの数，ワークロードの比率，VM の数，アグリゲータの数を変更しながら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．提案システムを用いることで，通常通りに集約を行う場合よりも，数〜10%程度ネットワークスイッチの電力を削減することができる．

今回，シミュレーションにより省電力効果の測定を行った．しかし，シミュレーションで全ての状況を再現することは非常に難しい．例えば，ネットワーク帯域などが考えられる．集約のための通信や VM の移送のタイミング，ルーティングなどによりネットワークの負荷は変化し，遅延時間も変わってくる．そのため，実環境での実験も行い，省電力効果について検証する必要がある．

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) Christopher Clark, Keir Fraser, Steven Hand, JacobGorm Hansen, Eric Jul, Christian Limpach, Ian Pratt, and Andrew Warfield. Live migration of virtual machines. In *Proceedings of the 2nd Conference on Symposium on Networked Systems Design and Implementation (NSDI '05)*, pp. 273–286, May 2005.
3) Robert Kembel. *Fibre Channel: A Comprehensive Introduction*. Northwest Learning Assoc, 2000.
4) Dennis Abts, MichaelR. Marty, PhilipM. Wells, Peter Klausler, and Hong Liu. Energy proportional datacenter networks. In *Proceedings of the 37th annual International Symposium on Computer Architecture (ISCA '10)*, pp. 338–347, June 2010.
5) Akshat Verma, Puneet Ahuja, and Anindya Neogi. pMapper: Power and Migration Cost Aware Application Placement in Virtualized Systems. In *Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware (Middleware '08)*, pp. 243–264, December 2008.
6) Fabien Hermenier, Xavier Lorca, Jean-Marc Menaud, Gilles Muller, and Julia Lawall. Entropy: a consolidation manager for clusters. In *Proceedings of the 2009 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE '09)*, pp. 41–50, March 2009.
7) 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.