人工知能学会
第12回知識流通ネットワーク研究会
SIG-KSN-012-06

研究目的に最適なベアメタル・クラウドの提案

○桑田喜隆(*1), 横山重俊(*2), 吉岡信和(*2)
(*1) NTTデータ, (*2) 国立情報学研究所

Proposal of Bare-Metal Cloud System for Researchers

Yoshitaka Kuwata(1), Shigetoshi Yokoyama(2) , Nobukazu Yoshioka (2)
(1) NTT DATA CORPORATION, (2) National Institute of Informatics

概要

クラウドコンピューティング技術を利用した「クラウドサービス」の普及が進み、企業や官公庁、教育分野などで利用が広がりはじめている。仮想化技術を活用しリソースをオーバーコミットすることで、IaaS(Infrastructure as a Service)は安価に提供されるようになった。他方、科学技術計算などを扱う研究分野では、マシン性能を最大限に利用するために物理マシンをそのまま使いたいとの要求が存在する。
筆者らは、仮想化技術を利用せずに計算機リソースの貸し出し管理を行う「ベアメタル・クラウド」を提案する。

Abstract:

Cloud Computing Technologies becomes popular and are widely used in many fields such as private companies, governments, and education. By making use of virtualization technologies, and by over-committing the computing resources to users, IaaS (Infrastructure as a Service) provides virtualized computer resources at very low-cost. On the other hands, in academia where demand of computing intensive tasks is high, researchers need to use as much resources as possible. We propose 'Bare-Metal Cloud' architecture which does not rely on virtualization technologies.

1. はじめに

クラウドコンピューティング技術を利用した様々なサービスは普及期を迎え、情報システムを構築する際の当たり前の技術の一つとして認識されてきている。初期の頃はコンシューマ向けのサービスをインターネット経由で提供するケースが多かったが、近年は企業の情報システムや官公庁の利用など応用される分野が拡大している。

クラウドコンピューティング技術の中でInfrastracture As A Servce (IaaS)と呼ばれるサービスは、コンピューティングのための資源をセルフサービスで払いだすことが可能であり、必要なときに必要なだけのリソースを用意することができるため、オンデマンドでリソースの必要になる教育分野や研究開発分野で特に役に立つと考えられる。参考文献 1)は教育分野のクラウドのアーキテクチャに関して分析をおこなったものであり、参考文献 2)はセルフサービス型のクラウドの有効性の評価を実施したものである。

クラウドコンピューティングの最も重要な技術項目の一つは、仮想化技術である。仮想化技術は計算リソースを最適化し、その利用効率を向上させることで大幅なコストダウンを実現させることが出来る技術である。

たとえば、CPU の仮想化を行うことで、実際に存在する物理マシンの台数より多くの仮想マシンを作成することが可能となる。たとえば、8 コアの CPU を持つ物理マシンを一人で利用する場合には、8 コア全てを占有することになるが、

[1] Yoshitaka Kuwata
NTT データ　基盤システム事業本部
東京都江東区豊洲 3-3-9 豊洲センタービルアネックス
kuwatay@nttdata.co.jp

仮想化技術を利用して仮想マシンとして貸し出す場合、各ユーザに 1 コアを割り当てるとすると 8 ユーザが利用可能となる。更に、オーバーコミットをすることも可能で、一つの CPU を 16 ユーザ以上の利用が可能である。多くのアプリケーションは CPU の利用率が 10%以下と低く、オーバーコミットを行っても影響が少ないことが多い。オーバーコミットによって、CPU の利用効率を向上させることで、安価にサービスを提供することが可能となる。

ストレージの仮想化についても同様に、全ユーザの利用するストレージ領域を共有しユーザの使っていない領域を最適化することで、ストレージの利用効率を向上する。ストレージシステムに、余剰分のリソースを持たないことでコストダウンを可能とすることが可能である。

他方、仮想化技術をベースとしたオーバーコミットやリソースの最適化による弊害も存在する。たとえば、仮想化によって CPU を共有した場合、他のユーザの利用状況によって、使える CPU リソースの最大値が変わり、場合によっては処理に十分なパフォーマンスが発揮できないような状況も考えられる。このため、仮想化技術ではユーザ間の隔離を確実にすることが重要な課題の一つとされており、各種の工夫が行われている。たとえば、CPU リソースについてはユーザの利用可能な絶対値をパーセント単位で設定することで、他のユーザからの影響は最小限になるような制御を行うことが可能となっている。しかし、制御の複雑さゆえに完全に隔離して制御を行うことが難しい。たとえば、CPU の内部に持つキャッシュの影響によって処理に大きな差が出るため、他のユーザの影響を排除することが難しい。

筆者らは、仮想化技術を使わずにクラウドのオンデマンド性を提供することのできる「ベアメタル型」のクラウドを提案する。

2. 研究のための計算機リソース管理の課題と貸し出し方式比較

研究目的に計算機リソースを利用する場合の課題を以下にあげる。

(1) 仮想化によるオーバーヘッド
仮想マシンを利用する場合には、仮想化によってオーバーヘッドが生じる。仮想化技術の実装に依存するが、一般にベアマシンを使う場合に比べて低下する。たとえば、Xen 仮想マシンのオーバーヘッドに関しては参考文献 3)で論じられている。またネットワーク性能や IO 性能も低下することが知られている。参考文献 4)はネットワーク性能を向上させるための取り組みである。

(2) 仮想化によるマシン内の時間の揺らぎ
仮想化を司る Virtual Machine Monitor (VMM)は管理下の仮想マシンに CPU を順番に割付を行う。このため、仮想マシン内で動作する OS のスケジューリング外の管理で動作する仕組みであるため、OS から見た場合には突然時刻が飛んだように見え、OS の時刻には揺らぎが生じる。このため、アルゴリズムの検証などを行うために精密なベンチマークなどをする際に、測定値が正確でなくなる可能性がある。

(3) 他利用者の影響による性能のゆらぎ
前章で議論したとおり、仮想化技術の制約でユーザの隔離が完全でないため、他の利用者によって仮想マシンの性能に影響がでる可能性がある。

(4) リソースの利用効率の向上
本課題は、上記 3 課題とは観点が異なる。リソースを提供する立場からは、研究用に利用するクラウドのリソースは限られるため、利用効率を上げ、できるだけ多くの利用者が利用できること望ましい。
ユーザにオンデマンドでリソースを払い出し返却するための仕組みを実現することで必要なユーザが利用しやすくなり、全体として利用効率が向上する効果が期待できる。また、計画的にリソース配分を決定することが出来る場合には、リソース配分の最適化が可能となると考える。

他方、計算機リソースを貸し出す方式としては、以下の方法が考えられる。

(a) 物理マシンをそのまま貸し出して利用する。
(b) 仮想化技術を使い、複数の仮想マシンを利用するが、物理マシンは占有する。ただし、他のユーザと共有しない。
(c) 仮想化技術を使い、複数の仮想マシンを作成し、他のユーザと物理マシンを共有する。ただし、仮想マシンの数は CPU コア数以下とし、1 個の CPU コアの共有はしないこととする。（オーバーコミットはしない）
(d) 仮想化技術を使い、他のユーザと物理マシンを共有する。さらに利用効率を向上する

ために、リソースのオーバーコミットをする。

表 2-1 に各方式に対する課題の対応状況を示す。

表 2-1 研究のための計算機リソース管理の課題と貸し出し方式の比較

方式	(1) 仮想オーバーヘッド	(2) 時間揺らぎ	(3) 他ユーザの影響	(4) リソース利用効率
(a) 物理マシン	なし	なし	なし	小
(b)仮想化(占有)	小	あり	なし	小
(c) 仮想化(共有)	小	あり	小	中
(d) 仮想化（オーバーコミット）	小	あり	大	大

物理マシンをそのまま利用する方式では、仮想化に起因する性能的な問題はないが、計算機リソースの利用効率が悪くなる。このため計算機リソースを効率よく利用するための工夫が必要となる。そこで、物理マシンを効率よく貸し出し管理可能なベアメタル方式のクラウドを提案する。

3. ベアメタル・クラウドの提案

本稿では、仮想マシンの課題を避け、セルフサービスによるオンデマンドの払い出しなどクラウドならではの使い勝手を実現するためにベアメタル方式のクラウドを提供する。ベアメタル方式のクラウドは仮想化を使わずに物理マシンを貸し出す機能を有するクラウドである。通常の仮想化を利用するクラウド(IaaS)と区別するため、本稿では「ベアメタル・クラウド」と呼ぶこととする。

ベアメタル・クラウドは IaaS 基盤と同様に、以下の機能を有することが必要である。

(1) ユーザ管理機能
(2) マシンの貸し出し管理機能
(3) マシンイメージ管理機能
(4) ネットワーク制御機能

また、上記に加えてベアメタル固有の要件として、以下の機能が必要である。

(5) 物理マシンの管理機能

図 3.1 にベアメタル・クラウドの構成の概念図を示す。

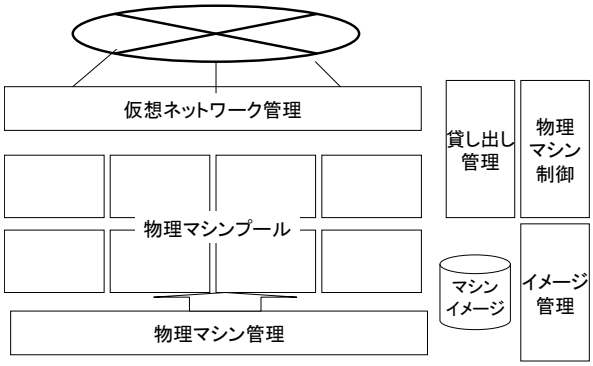

図 3.1 ベアメタル・クラウドの構成概念図

以下のベアメタル・クラウドで提供する必要のある機能について説明をする。

3.1 ユーザ管理機能

クラウドを利用する権限を持つユーザを管理する。利用者の利便性のため、別に認証システムを導入している場合には連携することが望ましい。

3.2 マシンの貸し出し管理および制御機能

ユーザからの要求に応じて物理マシンを払い出し、ユーザから返却されたマシンを管理下に戻す機能。また、借り出した物理マシンを起動したり停止したりする機能が必要である。

3.3 マシンイメージ管理機能

貸し出した物理マシンで利用するマシンイメージを管理する機能。通常はマシンイメージには OS を含む必要最小限の機能が含まれている。

3.4 仮想ネットワーク制御機能

貸し出した物理マシンの仮想ネットワークを設定する機能。通常はネットワーク経由で貸し出したマシンを利用することになるため、仮想ネットワーク制御機能は必須である。クラウドの利用者間のセキュリティを確保するため、閉域ネットワークを構成するようにして、異なるユーザ間のマシンは通信が出来ないような構成とすることが必須である。

3.5 物理マシンの管理機能

物理マシンの電源の管理や、状態などを監視し管理する機能。物理マシンのコンソールを遠隔操作可能なように利用者に提供することで、利便性の向上が期待される。

物理マシンの管理機能については、ハードウエア層で実現する機能である。標準化が進められてはいるが、物理マシンを提供するベンダに固有の機能として実装されている機能も多く、ベンダ独立で細かな機能を提供することは難しい。

ベアメタル・クラウドは直接物理マシンを扱う必要があるため、仮想マシンを扱うクラウドに比べ、起動のシーケンスやネットワークの割り当てシーケンスが異なる。

また、ベアメタル・クラウドでは物理マシンの制御をエンドユーザに渡すことになる。このため、悪意のあるユーザが OS レベルで設定を変更することでセキュリティホールになりえる。たとえば、ネットワーク設定を変更し、クラウド上の全てのネットワークトラヒックを覗き見ることも可能になるため、対策が必要である。

4. ベアメタル・クラウドの試作

筆者らは、Open Source Software(OSS)である dodai-compute [6] を使って NII 内にベアメタル・クラウドを試作した。本章では試作システムの概要について説明する。なお、ここでは試作したクラウドシステムを「物理マシンクラウド」と呼ぶこととする。

4.1 dodai-compute について

dodai-compute は NII の主催する dodai プロジェクトの一つであり、物理マシンをベースとしたクラウド機能を実現するソフトウエアである。OpenStack プロジェクトの一部として開発が進められている。

4.2 物理マシンクラウドの特徴

前章で述べたベアメタル・クラウドを実現するための機能を実現している他に以下の特徴がある。

- 物理マシンとして、主に計算リソースとして利用する「コンピュートノード」と主にデータ格納用に利用する「ストレージノード」の２タイプを用意した。ただし、目的に応じて適宜異なるタイプのノードを追加することも可能である。
- 貸し出したマシンは NII の研究室のネットワークセグメントに L2 でブリッジする仕組みを取っており、研究室のマシンの一部としてネットワーク設定を行える。
- 利用者同士のネットワークセキュリティを確保するため、物理マシンのサービスセグメントは OpenFlow スイッチを使って論理的に隔離する方式を取っている。VLAN を使う方式に比べ、仮に利用者が物理マシンのネットワーク設定を変更しても他のユーザのデータを見ることは出来ない仕組みを実現している。
- 利用者に貸し出されておらず、貸し出し可能な物理マシンを「マシンプール」と呼ばれる単位で管理することで、物理マシンの貸し出し要求に対して直ちに貸し出すことが可能である。また、物理マシン返却時には、ディスクの内容を完全に消去する仕組みを実現しており、次の利用者が読み出せないようにしてセキュリティを確保している。
- 物理マシンはストレージ専用のセグメントを持っており、外部のストレージ装置へのアクセスを高速化することが出来る。
- ユーザの認証には NII の idp(IDentity Provider)を利用しており、他のシステムと共通の学認 ID を使ってログインすることが可能である。学認 [8] は Shibboleth をベースとした認証局で、大学間で認証局同士が連携し各種の情報システムに共通の認証情報を提供する仕組みである。

4.3 物理マシンクラウドの構成

図 4.1 に物理マシンクラウドの構成を示す。

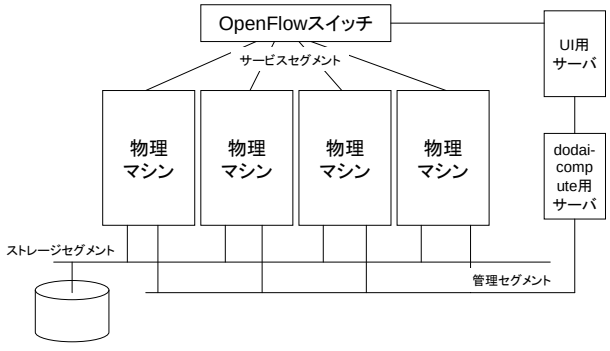

図 4.1 物理マシンクラウドの構成図

物理マシンは 3 種類のネットワークに接続されている。目的に応じて、ネットワークセグメントを使い分けるアーキテクチャとしている。利用者は、UI サーバ上に用意されているユーザインタフェースを経由してマシンの払い出しなどの依頼を行う。UI サーバから、dodai-compute 用サーバに依頼を出し、各所の物理マシンやネットワークを制御する仕組みである。

また、表 4.1 に物理マシンクラウドの主な仕様を示す。

表 4.1 物理マシンクラウドの主な仕様

項目	仕様
貸し出し可能マシン台数	72
利用可能コア数	864
CPU スペック	Xeon 5650 (2.66GHz, 6core)×2
メモリ	96GB memory
ディスク	RAID5 2TB (compute node) RAID5 10TB (storage node)
マシンタイプ数	2(compute,storage)
サービスネットワーク帯域幅	1Gbps
ストレージセグメント帯域幅	10GBbps
管理セグメント帯域幅	10GBbps
利用可能 OS	Ubuntu, CentOS など

4.4 オブジェクト・ストレージ

物理マシンクラウドの特徴の一つとして、利用者向けのオブジェクト・ストレージの提供がある。オブジェクト・ストレージは

実現の方法として、物理マシンクラウド上でストレージタイプのノードを借り出し、その上にオブジェクト・ストレージ・システムとして OpenStack プロジェクトの Swift を導入している。オブジェクト・ストレージは物理マシンクラウド上の物理マシンからアクセス可能なほか、外部のネットワークからも利用することが可能である。また認証の方法としては学認と連携している。

オブジェクト・ストレージはノードの動的な追加が容易であるため、ストレージ容量が不足した場合には物理マシンクラウドから追加でノードを借り出しオブジェクト・ストレージのノードとして追加することで拡張可能である。また、オブジェクト・ストレージの提供が不要になった場合には、オブジェクト・ストレージ・システム自身を削除し、使っていた物理マシンを物理マシンクラウドに返却することも可能である。

4.4 ユーザインタフェース

利用者は Web ページで提供されるユーザインタフェースを使って仮想マシンを操作することが出来る。

(1) クラスタの操作

物理マシンクラウドでは、「クラスタ」と呼ぶ単位で物理マシンを管理する。貸し出した物理マシンはクラスタに所属する。また、クラスタ単位でネットワークセグメントを設定する。

クラスタの管理画面から、物理マシンの払い出しや返却、起動停止などの操作を行うことが可能である。物理マシンの払い出し時に、初期導入する OS の種類を選択することが出来る。

図 4.2 にクラスタ管理画面の例を示す。

図 4.2　クラスタ作成画面の例

(2) クラウド基盤の操作

物理マシンの払い出しが済めば、物理マシンに遠隔ログインしてマシンを操作することが可能である。この段階では OS 以外のソフトウエアは導入されていないため、利用者が必要に応じてソフトウエアを導入することが必要である。

物理マシンクラウドでは、クラスタ上の物理マシンに、Hadoop, OpenStack,, SunGridEngine などのクラウド基盤ソフトウエアを導入する機能を提供している。

図 4.3 にクラウド基盤作成画面の例を示す。

図 4.3　クラウド基盤作成画面の例

5. 類似の研究開発事例

Dodai-compute と類似のミドルウエアとして、OpenStack プロジェクトで開発中の OpenStack General Baremetal Provisioning Framework [5]（以下、OpenStack/GBPF）が存在する。OpenStack/GBPF は dodai-compute と類似の機能を持っているが、実装の方式が異なる。

両者とも開発中のプロジェクトであるため、今後機能的に融合してより洗練されたソフトウエアとなってゆくことが期待される。

また、オープンソースソフトウエアとして提供されるコンピュートリソースの管理ソフトウエアとして Virtual Computing Lab(VCL)[9]があげられる。

6. まとめと今後の課題

本稿では、仮想化技術を使わずに物理マシンを直接貸し出すことの出来るベアメタル・クラウドを提案し、そのメリットおよびデメリットについて論じた。また実装例として dodai-conpute を取り上げて、そのアーキテクチャについて述べた。

本稿では詳細の説明は省略したが、ベアメタル・クラウドで構築したクラスタ上にソフトウエアをデプロイするためのフレームワークとして dodai-deploy [7] がある。dodai-deploy を使うことで、自動的に Hadoop や SunGridEngine などのミドルウエア基盤ソフトウエアを自動的にインストールすることを実現している。

課題としては、仮想化技術とベアメタルを組み合わせることで、専有性とリソースの利用効率の向上の両方を達成することがあげられる。

A. 参考文献

1) アカデミッククラウドアーキテクチャーの検討，横山重俊，桑田喜隆，吉岡信和，情報処理学会論文誌, 第 54 巻 第 2 号（平成 25 年 2 月）
2) オンデマンド・セルフサービス型 プライベートクラウドの企業内での利用モデルの提案と検証，武田健太郎，桑田喜隆，飛内拓弥，岩谷正広，中村竜也，情報処理学会論文誌, 第 54 巻 第 2 号（平成 25 年 2 月）
3) Aravind Menon, Jose Renato Santos, Yoshio Turner, G. (John) Janakiraman, and Willy Zwaenepoel, Diagnosing performance overheads in the xen virtual machine environment. *Proc. of the 1st ACM/USENIX international conference on Virtual execution environments* (VEE '05). ACM, New York, NY, USA, 13-23.
4) Muhammad Bilal Anwer, Ankur Nayak, Nick Feamster, and Ling Liu, Network I/O fairness in virtual machines. *Proc. of the second ACM SIGCOMM workshop on Virtualized infrastructure systems and architectures* (VISA '10). ACM, New York, NY, USA, 73-80.
5) OpenStack Project, General BareMetal Provisioning Framework, http://wiki.openstack.org/GeneralBareMetalProvisioningFramework
6) Dodai Project, dodai-compute, https://github.com/nii-cloud/dodai-compute
7) Dodai Project, dodai-deploy, https://github.com/nii-cloud/dodai-deploy
8) 学術認証フェデレーション, http://www.gakunin.jp/ja/
9) Apache VCL project, Virtual Computing Lab, https://cwiki.apache.org/VCL/apache-vcl.html