仮想ディスクストレージ「UKAI」

仮想環境において、位置管理が困難な仮想ディスクの実データ配置を柔軟に制御できるストレージシステム「UKAI」[*1]の研究をご紹介します。

3.1 仮想化前提基盤

もはや仮想化という言葉から新しい響きを感じることも少なくなってきました。今や仮想化技術はサービスを支えるための要素技術の1つとして重要な地位を占めています。もっとも、仮想化技術自体は新しいものでもなんでもなく、コンピュータが実用化された頃から様々な場面で利用されてきました。今では携帯電話用OSにすら実装されているマルチタスク技術は、最初は高価な物理CPUを仮想的に複数のプログラムで平行利用するために考えられたものでした。また、広くエンタープライズサービスで使われているJavaも、Java仮想マシンを活用する一種の仮想化技術と言えます。それにも関わらず、近年仮想化が大きく注目されたのは、普段利用しているサーバ環境を、許容できる速度で仮想マシンとして実行できる目処が立ったからです。

仮想マシンが実用的に運用できるようになったことで、ネットワーク上で提供されるサービスを仮想機材で構成する場面もでてきました。仮想化技術を利用することで、物理マシンを使っていたときには難しかったサービスの迅速な展開、高負荷時の緊急設備拡張、また利用減に伴う設備の縮退などが比較的簡単に実現できるようになります。もちろん、物理マシンを直接使う場合よりも性能が劣りますが、その差と運用の容易さが釣り合う時代になったと言えるでしょう。仮想マシンが物理マシンの性能に追いつくことはありませんが、基礎性能が向上するに従い、多くのサービスの要求品質を満たす仮想環境がいずれ実現されるであろうことは想像に難くありません。もはやCPUをひとりで占有するアプリケーションがほとんど存在しないのと同様に、一台のハードウェアの上に直接サービスOSが載ることはなくなるでしょう。一部の特殊な用途を除けば仮想マシンがサービス基盤の基本部品になる時代がやってきます。

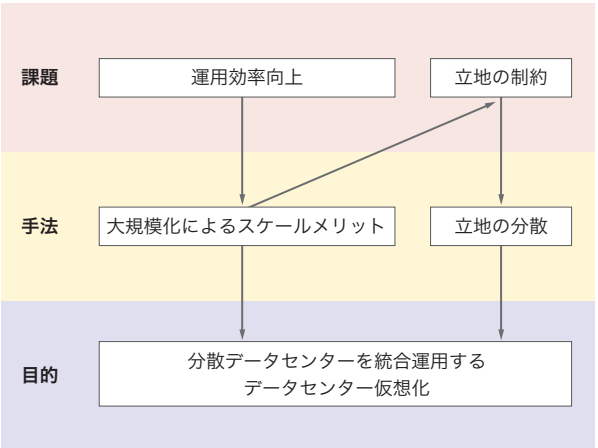

図-1 仮想データセンターの必要性と課題

仮想基盤を効率的に運用するための最初の一歩は仮想マシン専用のデータセンター構築です。IIJの松江データセンターパーク[*2]はそういったデータセンターの1つで、IIJのクラウドサービスIIJ GIOの基盤として利用されています。データセンターの効率をあげる手法の1つに大規模化が挙げられます。しかし、日本のように国土が狭い環境や、大規模化に見合う需要が即座に見込めない場合に大きな投資をして巨大なデータセンターを建設することは困難です。そこで、地理的に独立した中規模のデータセンターを仮想的に大規模データセンターとみなして運用する「仮想データセンター」の考え方がでてきます。分散したデータセンターを横断して基盤サービスを構成する場合、地理条件やネットワーク遅延を考慮して仮想資源を効率的に配置運用しなければなりません（図-1）。今後は仮想資源を柔軟に配置、再配置するための仕組みが重要になってくるのです。

＊1　Keiichi Shima, "UKAI: Centrally Controllable Distributed Local Storage for Virtual Machine Disk Images", In Proceedings IEEE Globecom 2013 Workshop on Cloud Computing Systems, Networks, and Applications（CCSNA）, December 9, 2013.
＊2　IIJの松江データセンターパーク（http://www.iij.ad.jp/DC/dcpark/）。

3.2 仮想マシン用ディスクストレージの必要性

仮想マシンは大きく3つの資源から成り立っています。1つ目は計算機の中核となるCPUとメモリ、2つ目は計算機同士を相互接続するためのネットワーク、最後がシステムやデータを保持するストレージです。これらの資源を柔軟に配置、再配置できる技術が今後の柔軟な仮想基盤の運用に大きく貢献します。

ご存知のとおり、1つ目の資源(CPUとメモリ)の再配置に関しては既に実運用段階です。仮想マシンが前提とする仮想ハードウェア環境はある程度統一することができるので、仮想マシンの仕様が決まっていればそれを実行するハイパーバイザーがどこで運用されていても問題ありません。仮想化技術によっては、動作中の仮想マシンを別のハイパーバイザーに移動させるライブマイグレーションの機能を提供しているものもあります。

2つ目の資源(ネットワーク)の再配置技術については、長らく研究開発が進められてきましたが、Software Defined Networking(SDN)技術がその有力候補として浮上してきました。ネットワークの仮想化と言えば、これまではVLANが標準的に使われてきましたが、データセンターの大規模化に伴いその限界が見えています。SDN技術を使えば仮想マシン単位でネットワークを切り出すことができるようになり、これまでVLANでは実現が困難だった柔軟なネットワーク構成が可能になります。

残る資源はストレージですが、これに関しては今現在よい選択肢が存在しない状況です。仮想マシンに追加する仮想ディスクを提供する技術として、NFSやiSCSIを採用することが多いと思います。ディスクボリューム管理体系の粒度や耐障害性のための二重化などを考慮すると、大規模基盤ではiSCSI製品を選択する場合がほとんどでしょうか。ストレージといってもネットワーク経由で接続しているので、仮想マシンのネットワーク資源が正しく再配置されていれば理屈上はストレージへのアクセスも継続できます。しかしながら、ストレージに保管されたデータは物理的なストレージサーバに固定されたままです。今後、仮想データセンターの考え方に基づく分散運用を想定すると、一度停止した仮想マシンを再起動した際、ハイパーバイザーの資源割り当て状況によっては、これまでとは異なるデータセンターで再起動する可能性があります。ストレージ資源がどこかの機器に固定的に割り当てられていると、遠隔地で起動した仮想マシンの性能に影響を与えてしまいます。CPUやネットワーク資源と同様、ストレージ資源も運用状況に応じて柔軟に再配置できなければなりませんが、NFSやiSCSIではそういった要求に応えることができないのです。

IIJの技術研究所では、分散運用されているデータセンターを仮想データセンターとして統合運用する環境を念頭に置き、仮想マシンで用いられる仮想ディスクの実データ配置を柔軟に制御できるストレージシステム「UKAI」の研究開発を進めています。

3.3 位置配慮型仮想ディスクストレージUKAI

UKAIは以下の3つの目的を実現すべく設計されています。

1. 管理者が運用上の判断で仮想ディスクの実データ配置場所を制御できること
2. ネットワーク遅延が均一でない環境で実用的に動作すること
3. 障害に対する冗長性を持つこと

仮想ディスクの実データは各データセンターに分散配置されるストレージノードに格納します。仮想マシンを作成するとき、その仮想マシンに割り当てる仮想ディスクの実データは可能な限り仮想マシンが動作する場所の近くに配置したいでしょう。各データセンターにストレージノードを配置し、運用的に仮想ディスクの実データ位置を指定することで、仮想マシン本体と仮想ディスクの実データ間の遅延を短縮し、分散環境での性能劣化を防ぐ必要があります。しかし、前節で述べたとおり、仮想マシンが永久に同じハイパーバイザーに配置されるとは限りません。仮想基盤には複数の利用者が乗り入れて共同利用しています。サービスの利用状況によっては、やむなく別のハイパーバイザー、場合によっては別のデータセンターのハイパーバイザーで起動しなければならない場合もでてくるでしょう。このような状況でも、透過的に仮想ディスクへのアクセスを提供できることが大前提です。しかし、別のデータセンターのストレージ資源にアクセスすることは仮想マシンの性能を大きく落とすことになります。UKAIでは、仮想ディスクの管理と実データの管理を分離し、仮想ディスクの情報を固定したまま、実データの移動を実現します。これにより、仮想マシンを稼働させたまま、仮想ディスクを構成する実データの位置を、運用上の判断に応じて自由に変更することができるようになります。

分散環境ではネットワーク遅延への対応も重要な課題です。特に、仮想データセンターのような広域に分散した拠点にまたがってサービスを提供する場合は遅延が与える影響を見積もる必要があります。広域運用では大きく二種類の遅延が考えられます。1つはデータセンター内の通信遅延です。ストレージノードが複数運用されている状況で、かつそれらのノードが同じデータセンターに配置されている場合に相当します。この場合、遅延は小さく（ほぼ1ms以内）、ある程度均一であると想定できます。もう1つはデータセンター間の通信遅延です。こちらはデータセンター同士の位置関係によって変化し、データセンター内と比べて遅延の揺らぎも大きくなると想定できます。UKAIでは、実データの配置場所の選択を管理者に任せ、発生する遅延を予測した運用を実施します。仮想マシンが遠隔で起動してしまった場合でも、その仮想ディスクの実データとの間の遅延が定量的に分かっていれば、性能劣化の度合いが予測できます。それに応じて実データの移動計画を検討することになります。

最後の目的である冗長性については言わずもがなでしょう。ストレージノードが分散運用されている状況で、どこかのノードの障害によってサービス全体を停止するわけにはいきません。ネットワーク障害によってストレージノードへのアクセスが失われる場合も同様です。もちろん、あらゆる障害すべてに対応することはできませんから、ある程度障害を想定して、それに対応できる冗長性を確保することになります。UKAIで想定する障害はストレージノードのディスク及びネットワークハードウェア故障です。データセンター内の機器障害（スイッチ障害など）については、別途データセンター側の機材で冗長化されていることが前提です。UKAIでは仮想ディスクと実データの管理を分離しているため、ある仮想ディスクに対して複数の実データのコピーを持たせることが可能になっています。同一データセンター内に複数のコピーを置けばネットワーク的なミラーディスクとして運用できます。またコピーを他の拠点に持つことで、ストレージのアクセス性能は低下しますがディザスタリカバリの手段として運用できます。

3.4 UKAIの実装

UKAIの設計を検証するためにプロトタイプ実装を公開しています[*3]。実装は以下の点に配慮してあります。

a. 既存のハイパーバイザーとの親和性
b. 分散システムによる一点障害の排除
c. クラウドコントローラとの連携

仮想マシンに仮想ディスクを提供する方式は2つ考えられます。直接的な方法は、ハイパーバイザー本体に新しい仮想ディスク形式を拡張実装することです。直接組み込むことによる性能向上が見込めますが、ハイパーバイザーごとに個別対応する必要があります。もう1つの方法は、仮想ディスクのイメージをファイルとして提供することです。多くのハイパーバイザーは最もシンプルな方式として、1つのファイルを1つの仮想ディスクイメージとみなして仮想マシンに提供する仕組みを持っています。UKAIでは後者の方式を採用し、ハイパーバイザーと独立した仕組みで仮想ディスクストレージを構築します。

分散システムを設計する場合に最も気になるのは、障害発生時の冗長性をどこまで持たせるかです。前節で述べたとおり、UKAIはストレージノードを複数運用することができるため、ストレージノードの障害にはある程度対応可能です。もう1つの問題は仮想ディスクを構成する情報（仮想ディスクのメタデータ）をどのように分散システムとして保存するかということになります。プロトタイプ実装では分散協調支援システムであるApache ZooKeeperを利用しています。

図-2にプロトタイプシステムのモジュール構成図を示します。ハイパーバイザーはローカルシステムにマウントされたファイルシステム上に作られたファイルを仮想ディスク

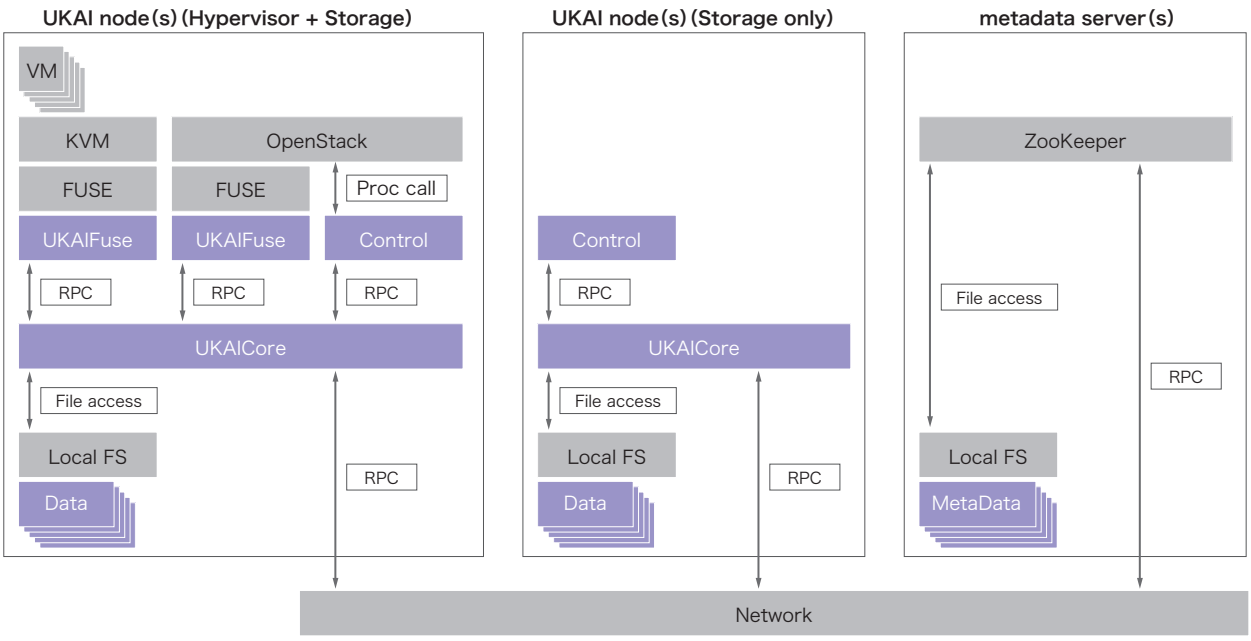

図-2 UKAIプロトタイプシステムのモジュール構成

*3 UKAI：A Location-Aware Distributed Storage Software for Virtual Machine Disk Images（https://github.com/keiichishima/ukai）。

として利用します。ハイパーバイザーから見ると、仮想ディスクイメージは単なるファイルに見えるため、ローカルファイルを仮想ディスクとして利用する場合と同様に運用できます。実際には、マウントポイントの先にはUKAIが存在しており、ファイル入出力はすべてUKAIのサブシステムを経由する形になります。プロトタイプでは、UKAIサブシステムの実装のためにFUSE[*4]とそのPythonバインディングの一種であるfusepy[*5]を使っています。UKAI自体もPythonで実装されています。

仮想ディスクはその名のとおり実態を持たず、実データへのポインターの集合になります。図-3に仮想ディスクと実データ及びストレージノードの関係を示します。図に示したとおり、仮想ディスクは複数のデータブロックに分割され、それぞれのデータブロックが実データへのポインターを持ちます。1つのデータブロックは複数の実データポインターを持つことができ、ストレージノードの障害に対応します。

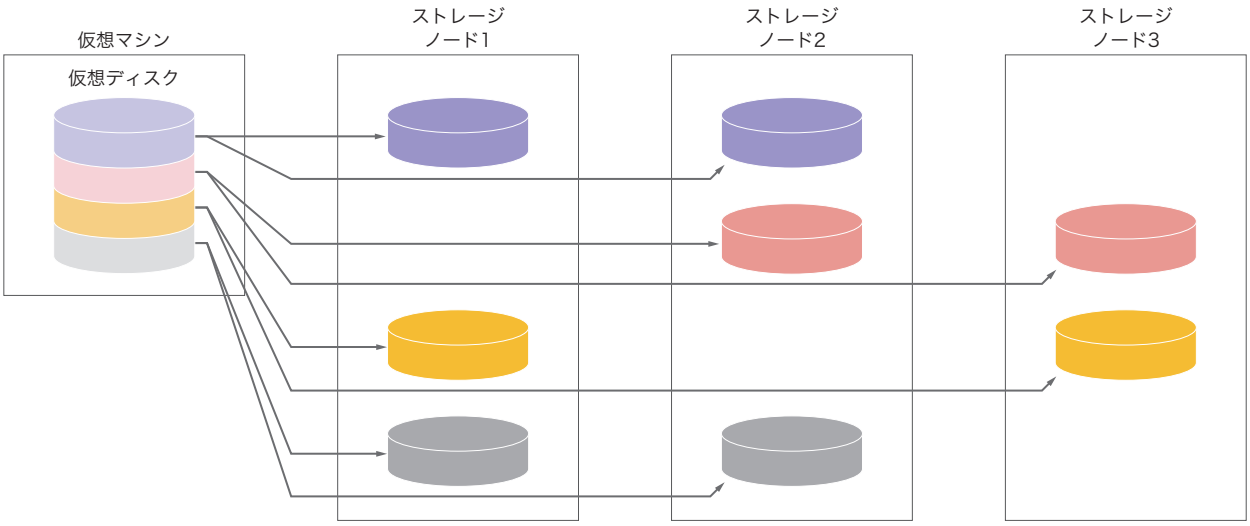

図-3 UKAI仮想ディスクと実データ格納の概念図

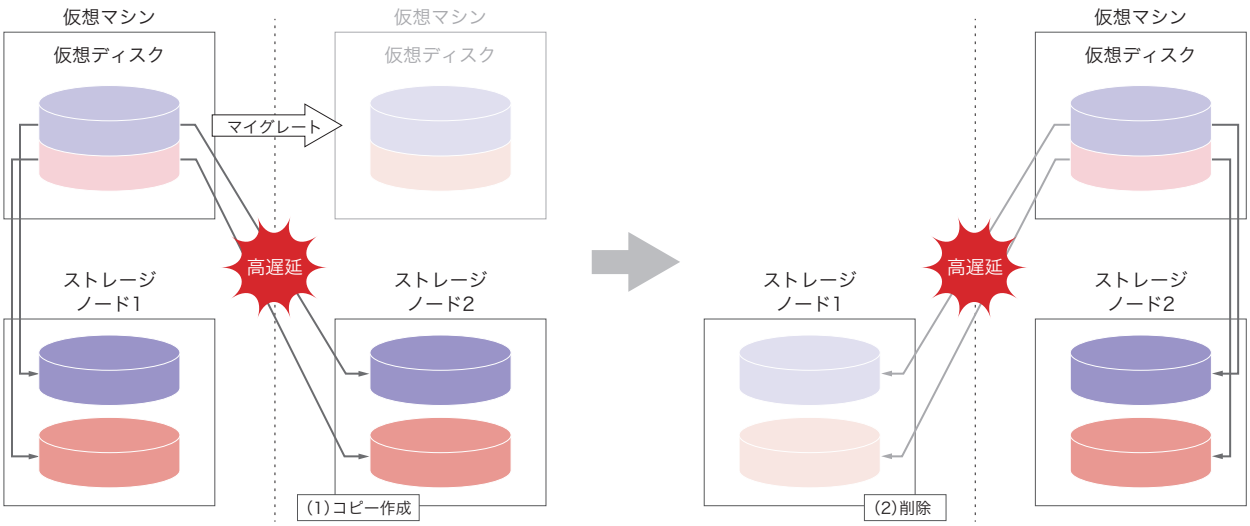

図-4 仮想ディスクの再配置

*4 FUSE(http://fuse.sourceforge.net)。
*5 fusepy(https://github.com/terencehonles/fusepy/)。

仮想ディスクの再配置は図-4に示したように実行されます。仮想マシンが異なるハイパーバイザーへマイグレート、もしくは異なるハイパーバイザーで再起動する場合に、実データを格納していたストレージノードとの間に大きな通信遅延が予測されたとします。この場合、まず移動先のハイパーバイザーに近いストレージノードに実データのコピーを作成し、仮想マシンを移動した後、遅延の大きいストレージノードの実データを削除します。仮想ディスクは単なるポインターの集合なので、稼働中の仮想マシンは実データのコピーが作成されたことや、遠方の実データが削除されたことを意識する必要はありません。再配置が完了するまでの間はディスクアクセス性能が劣化しますが、仮想マシンを停止することなく仮想ディスクの実データの位置を最適化できます。

3.5 まとめ

仮想環境の性能が向上していくにつれ、今後多くの物理マシンが仮想マシンに置き換わっていくと思われます。サーバレンタルのような物理マシン時代から続くサービスはもちろんですが、ソフトウェア的にマシンの追加削除が可能な仮想環境は、SaaSやPaaSなどの構成要素として使われるときにこそ、その真価を発揮するはずです。サービス基盤がクラウド化していく流れの中で、構成資源の柔軟な配置運用技術は不可欠です。今回、位置管理が困難な資源の1つであるストレージの再配置技術の研究をご紹介しました。IIJでは安定したインターネットサービスの基盤となる技術革新を引き続き進めていきます。

島 慶一(しま けいいち)
株式会社IIJ イノベーションインスティテュート 技術研究所 主幹研究員。広域分散コンピューティング環境における仮想計算機のアーキテクチャ及び仮想資源の柔軟な配置技術の研究開発を進めている。