クラスタ構築システム Rocks を用いた仮想クラスタの構築

中田秀基[†] 横井威[†] 江原忠士[†,††]
谷村勇輔[†] 小川宏高[†] 関口智嗣[†]

1. はじめに

データセンタにおける計算機運用の運用率を向上させる方法として，資源の一部を適当なサービスレベルアグリーメントの元に，予約ベースで貸し出すことによる方法が考えられる．この方法は，実際の計算機，ネットワーク，ストレージを用いて実現することもできるが，配線の変更，OS のインストール，アプリケーションのデプロイなど，膨大な作業が必要となる．

これを低コストで実現する方法として，仮想化を用いる方法が考えられる．仮想的な環境に仮想的なクラスタを構築してユーザに提供することによって，管理コストの大幅な低減が実現できる．

われわれは，クラスタの 3 つの側面，すなわち，計算機，ネットワーク，ストレージをそれぞれ仮想化することで，仮想クラスタを実現した [1)]．計算機の仮想化には VMWare[2)] を用い，ネットワークの仮想化には VLAN を，ストレージの仮想化には iSCSI を用いた．

また，仮想クラスタの構築にはクラスタデプロイシステムである NPACI Rocks[3),4)] を用いる．さらに，本システムそのものも Rocks によって簡便に配備することを可能とした．

2. システムの想定

本システムには，クラスタプロバイダ，サービスプロバイダ，ユーザの三者が関与する．クラスタプロバイダは本システムを使用して所有するクラスタを管理し，サービスプロバイダに提供する主体である．サービスプロバイダはクラスタを利用して，ユーザにサービスを提供する．

サービスプロバイダに提供される仮想クラスタは 1 つ以上のゲイトウェイとなるフロントエンドノードと 1 つ以上の計算ノードから構成される．ゲイトウェイと計算ノードはプライベートアドレスのローカルネットワークで接続されている．クラスタプロバイダは，ゲイトウェイのグローバルネットワークへのインターフェイスの IP アドレスをサービスプロバイダに提供する．

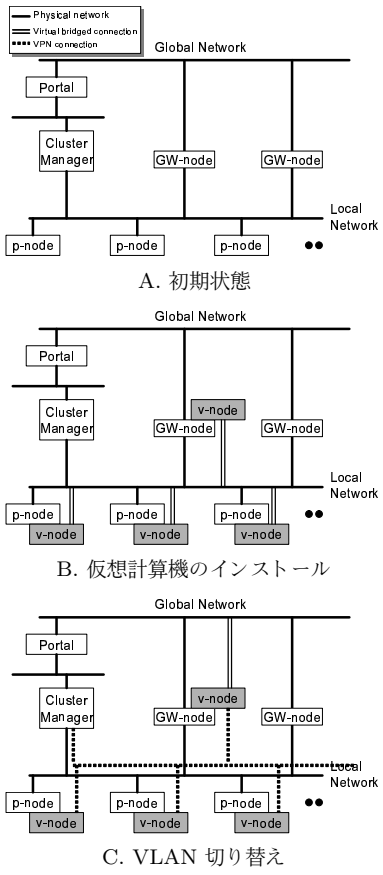
2.1 クラスタ提供シナリオ

- まず，クラスタプロバイダは本システムを利用し，クラスタをインストールする．
- 次にサービスプロバイダがクラスタプロバイダに対して仮想クラスタ構築を依頼する．その際にサービスプロバイダは，使用開始/終了時刻，使用計算機台数，デプロイされるべきアプリケーション，必要メモリなどの情報を提供する．アプリケーションは，なんらかの形でサービスプロバイダが用意する．
- クラスタプロバイダは本システムを用いて仮想クラスタを実クラスタ上に構築し，サービスプロバイダに提供する．
- サービスプロバイダはアプリケーションを利用したサービスをユーザに対して提供する．

3. Rocks によるクラスタのインストール

Rocks は，NPACI(National Partnership for Advanced Computational Infrastructure) の一環として SDSC(San Diego Supercomputer Center) を中心に開発されたクラスタ管理ツールである．クラスタのノード群に対して一括で同じソフトウェアパッケージをインストールすることができる．OS としては，RedHat Enterprise Linux をベースとした CentOS を使用している．Rocks では Roll と呼ばれるメタパッケージによってアプリケーションを管理する．クラスタ管理者は Roll を新たに追加することで，クラスタ

† 産業技術総合研究所
†† 数理技研



A. 初期状態

B. 仮想計算機のインストール

C. VLAN 切り替え

図 1 仮想クラスタのインストール

に新たな機能を追加することができる．

4. システムの設計

4.1 実クラスタの構成

上記の要請を実現するために想定したクラスタの構成を図 1A に示す．クラスタは，クラスタマネージャ，計算ノード (p-node)，ゲイトウェイノード (GW-node) から構成される．クラスタマネージャはクラスタ全体を管理するサーバで，Rocks の Frontend ノードでもある．p-node は，実際にジョブを実行する仮想計算機を動かすノード，GW-node は，仮想クラスタの外部インターフェイスとなる仮想計算機を動かすノードである．p-node は内部のローカルネットワークにのみ接続されているが，GW-node は外部ネットワークへも接続されている．

4.2 実クラスタのインストール

仮想クラスタを提供するために，クラスタプロバイダは実クラスタをインストールしなければならない．この過程は Rocks で行う．まず，クラスタマネージャを Rocks の Frontend としてインストールする．この際の Roll としてクラスタマネージャの他の機能もインストールする．

次に，その他のノードのインストールを行う．p-node と GW-node は基本的に同じアプライアンスとしてインストールするが，GW-node にはその後，グローバルネットワークへのインターフェイスを設定する．

4.3 仮想クラスタのインストール

図 1 に，仮想クラスタのインストールの様子を示す．図 1A は，初期状態である．

クラスタインストール時には，まず p-node および，GW-node 上で仮想ノード (v-node) を起動し，OS およびユーザの指定した Roll を配備する (図 1B)．起動した v-node はそれぞれブリッジネットワーク接続をローカルネットワークに対して持ち，このネットワーク経由でインストールが行われる．

個々のノードのインストールが終了したら，v-node 間の通信路を VLAN に移行する (図 1C)．個々の仮想クラスタごとに異なる VLAN タグを用いることで仮想クラスタ間でのデータ通信を切り分ける．また，クラスタマネージャは，すべての仮想クラスタの VLAN に参加する．これは，仮想ノード上の情報をクラスタマネージャで収集するためである．また，この段階で，GW-node からグローバルネットワーク空間へのブリッジ接続を行う．サービスプロバイダはこの接続を用いてクラスタにアクセスする．

5. 今後の課題

- フロントエンドを含めた仮想クラスタのデプロイ
  現在の実装ではゲイトウェイを提供することはできるが，Roll を利用した一貫性のあるクラスタを運用することはできていない．各仮想クラスタにそれぞれ仮想フロントエンドをインストールすることで，Roll を有効に活用したクラスタ構築を可能にしたい．
- 複数クラスタの統合的運用
  単一のクラスタでは，提供できる資源には限界がある．複数のクラスタ上に，仮想的な単一クラスタを形成し，運用することも検討する必要がある．

参 考 文 献

1) 中田秀基, 横井威, 関口智嗣: Rocks を用いた仮想クラスタ構築システム, 情報処理学会 HPC 研究会 2006-HPC-106 (2006).
2) : VMWare. http://www.vmware.com.
3) Papadopoulos, P. M., Katz, M. J. and Bruno, G.: NPACI Rocks: Tools and Techniques for Easily Deploying Manageable Linux Clusters, *Cluster 2001: IEEE International Conference on Cluster Computing* (2001).
4) : Rocks. http://rocks.npaci.edu/.