仮想クラスタ管理システムの設計と実装

中田 秀基 [1], 横井　威 [1], 江原 忠士 [1, 2]
谷村 勇輔 [1], 小川 宏高 [1], 関口 智嗣 [1]
1.産業技術総合研究所
2.数理技研

Grid Technology Research Center AIST

AIST

背景

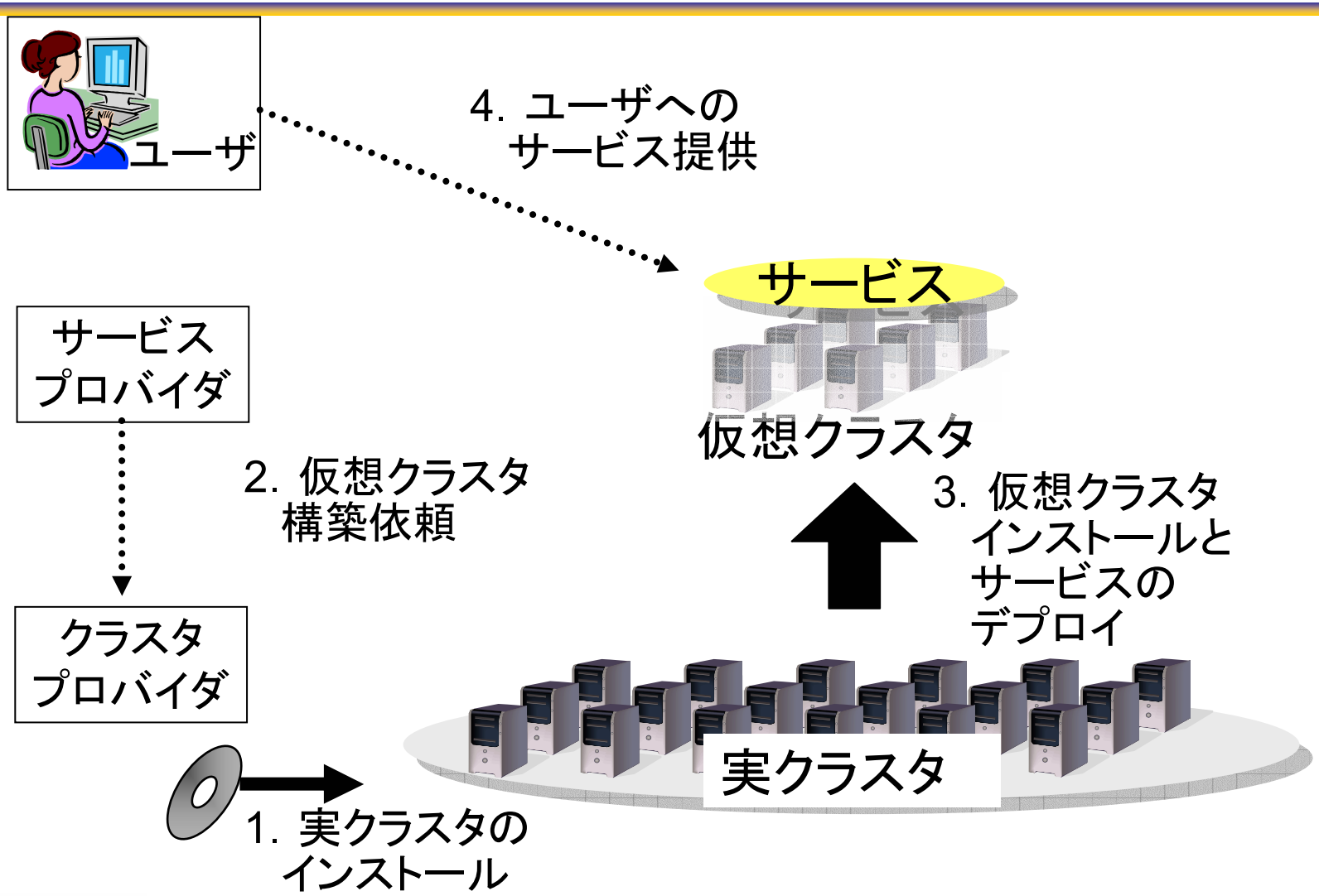
- 仮想化技術の普及
  - 仮想ノードによる管理コストの低減
- 仮想ノード → 仮想クラスタ
  - さらなる管理コストの低減を目指す

- 仮想クラスタ
  - 単なる仮想ノードの集合ではない
    - 管理ソフトウェアなどの設定
    - 名前空間の管理など
  - 計算機だけの仮想化では不十分
    - ストレージ
    - ネットワーク

研究目的

- 仮想クラスタ
  - 事前に予約された特定の期間，ソフトウェアがインストールされた仮想的なクラスタが提供される
  - 提供後はユーザが自由に追加インストール，設定可能
  - 期間としては数日−数ヶ月を想定
- 仮想クラスタ管理システムの提案
  - クラスタインストールツールRocksを用いて，管理用のソフトウェアも含めてインストール
  - 計算機，ストレージ，ネットワークの仮想化
    - 計算機 − VMware Server
    - ストレージ − iSCSI
    - ネットワーク − VLAN

利用シナリオ

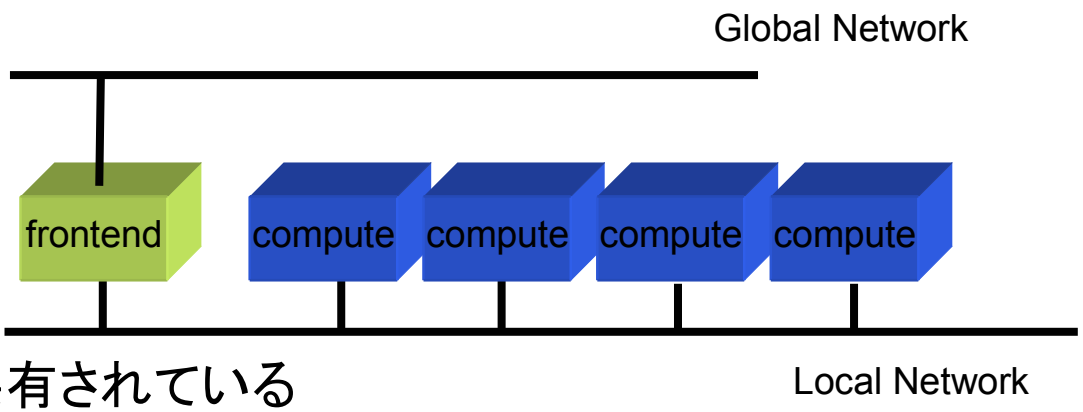

ユーザ
4. ユーザへの
サービス提供
サービス
仮想クラスタ
サービス
プロバイダ
2. 仮想クラスタ
構築依頼
3. 仮想クラスタ
インストールと
サービスの
デプロイ
クラスタ
プロバイダ
実クラスタ
1. 実クラスタの
インストール

利用シナリオ

- データセンターでの利用
  - サービスプロバイダが一定期間だけリソースを利用
- 大学の授業用クラスタ
  - 各授業に専用の仮想クラスタを割り当て
  - アプリケーション，設定を自由に変更可能
    - 失敗したら元に戻せる
  - 毎週定時に起動，終了
- 計算機ファームの拡大
  - 科学技術計算を行う計算機ファームを一時的に拡張
    - グリッド技術を使って透過的に
  - データベース，アプリケーションを自由に配備可能
  - 利用が終わったら解放

仮想クラスタへの要請

- サービスプロバイダから見ると一般的なクラスタと同じ
- ノード構成とネットワーク
  - フロントエンド 1台+ ワーカノード群
  - フロントエンドがデュアルホストのルータ
  - ワーカノード群はLANに接続
    - LANは安全



- 設定
  - 名前空間，ファイル空間が共有されている
  - 運用ソフトウェア
    - モニタリングシステム
    - バッチキューイングシステムなど
- ストレージ
  - 共有ストレージ
  - 個別ノード上のテンポラリストレージ

仮想クラスタ管理システムへの要請

- アプリケーションの自動配備，設定
  - 複数のノードにまたがった複雑な設定の自動化
- ノード構成の自動化
  - ルーティング設定
- 計算機の仮想化
  - 単一の物理ノードで複数の仮想ノードを運用可能
- ストレージの仮想化
  - 柔軟なストレージ管理
    - 物理ディスクにとらわれない容量設定
  - 集中管理による管理コストの削減
- ネットワークの仮想化
  - 一般に仮想計算機はブリッジ接続
    - 実計算機とネットワークを共有
    - クラスタのローカルネットワークには不十分
      - 実計算機のネットワークからの分離が必要

提案システムの概要（1）

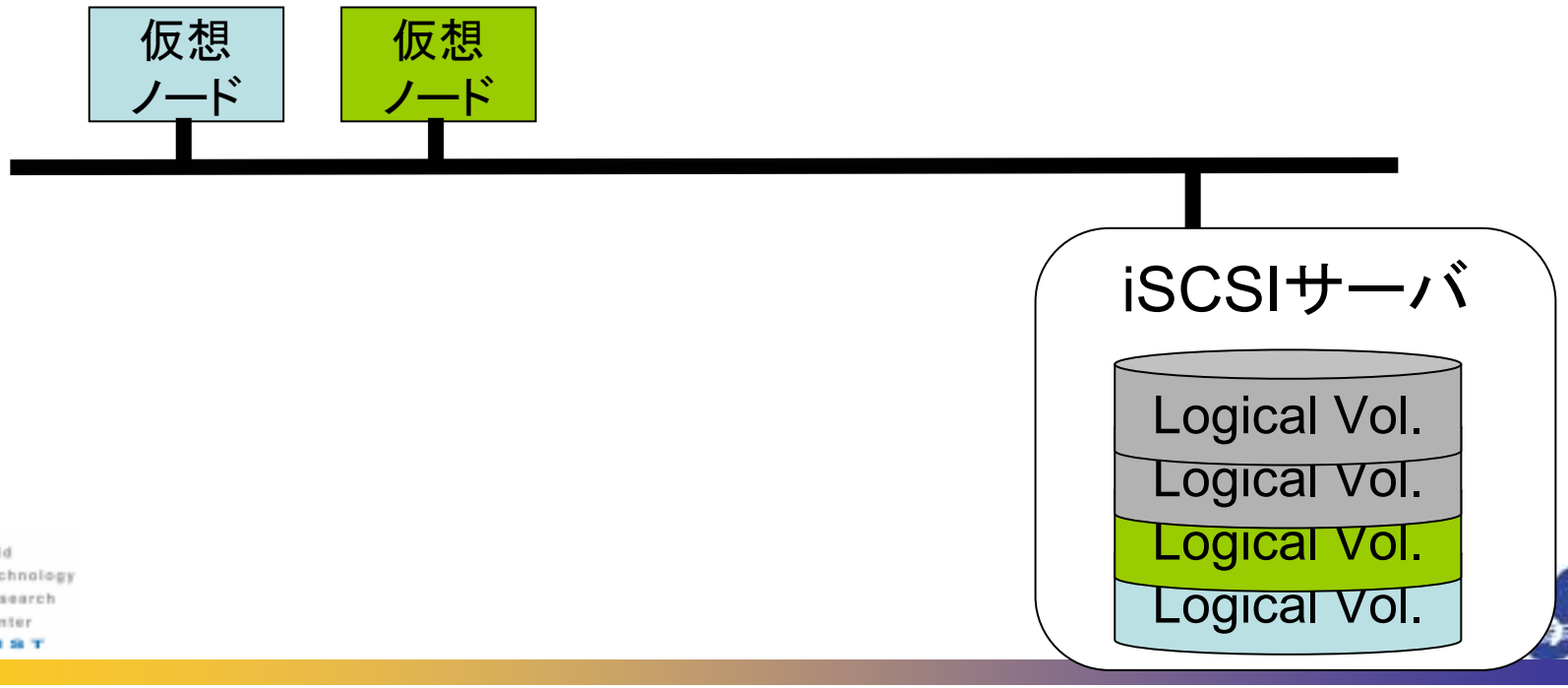
- アプリケーションのインストール，ノード構成の自動化
  - クラスタインストールツール Rocks を利用
    - UCSD でNPACIプロジェクトの一環として開発
    - 世界的に広く活用されている
    - Roll(メタパッケージ) が充実
      - 主要な科学技術ソフトウェアに関しては改めて開発する必要がない.

提案システムの概要 (2)

- 計算機仮想化
  - VMware Server
    - full virtualization を行う仮想計算機
- ストレージ仮想化
  - iSCSI + LVM (Logical Volume Manager)
    - iSCSI でロケーションを分離
    - LVMによる管理の容易化
- ネットワーク仮想化
  - タグVLAN
    - 仮想クラスタのネットワークを相互に分離

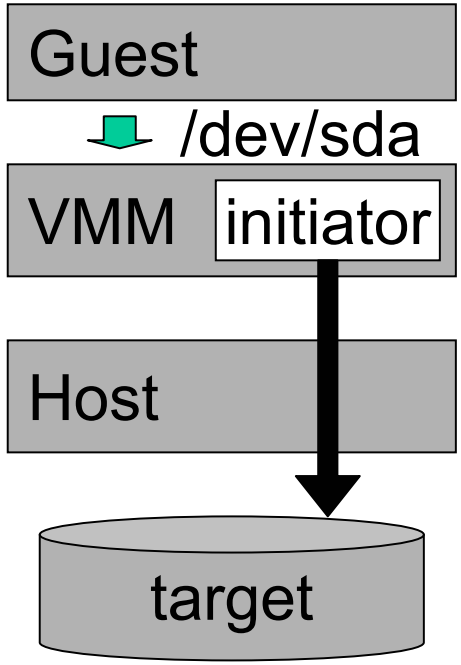
ストレージ仮想化

- ストレージを物理的な実体から切り離すことで，管理コストを低減
  - iSCSIを用いてリモート化，集中管理
  - LVMを用いて物理ディスク構成にとらわれない構成を実現

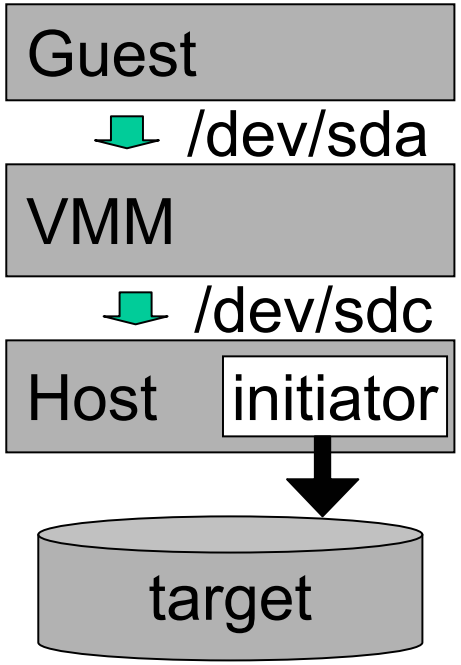


iSCSIと仮想計算機

- VMware Server はiSCSIを直接サポートしていない
  - ホストOSがアタッチしたものをVMに見せて回避



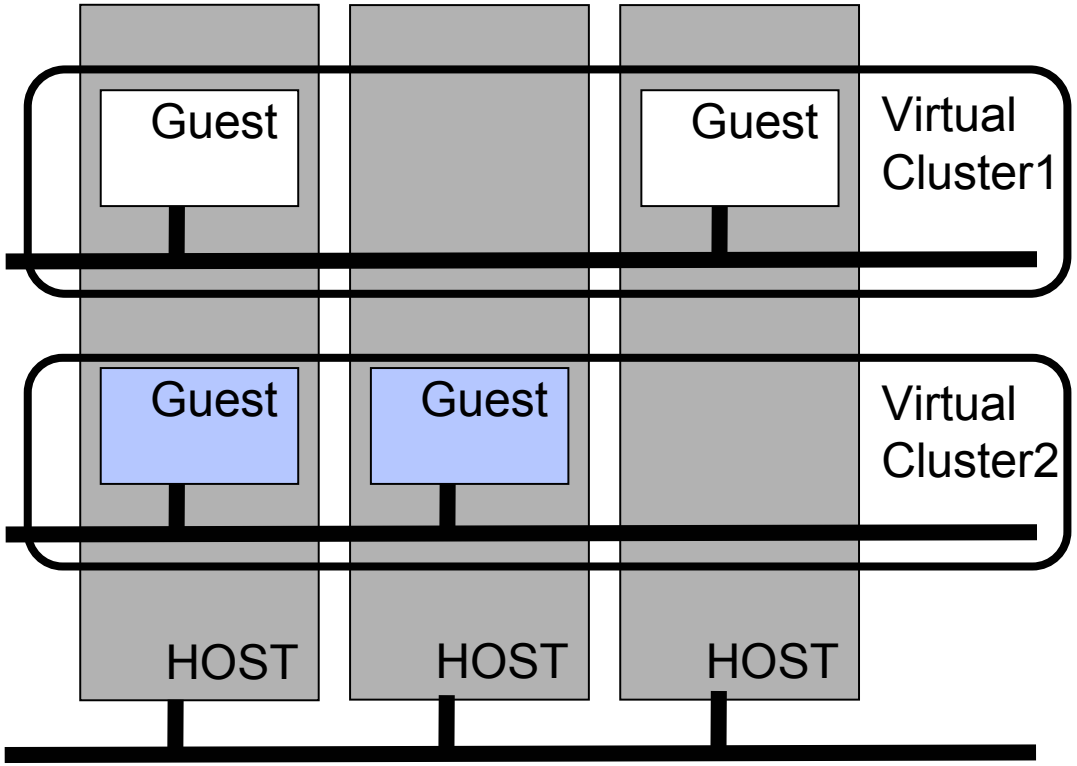
VMMがiSCSIを直接
サポートする場合



VMMがiSCSIを直接
サポートしない場合

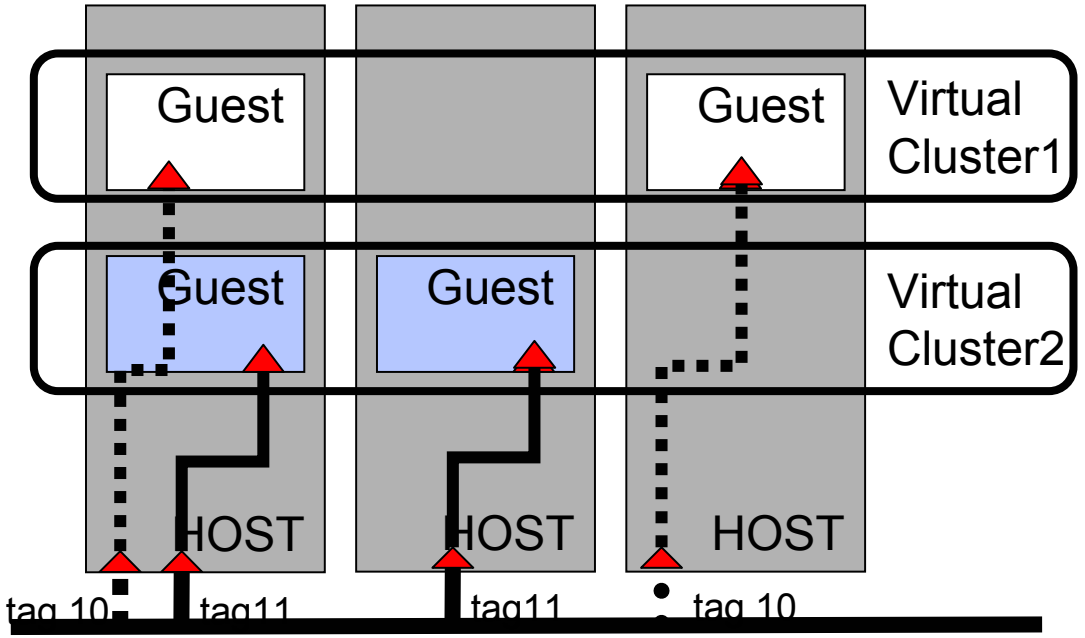
VLANによる仮想クラスタの分離

- 各仮想クラスタが専用の内部ネットワークを擬似的に持つ
- 相互に覗き見ることは不可能



タグVLAN によるネットワークの分離

- ホストノードでタグと仮想クラスタをマッピング
  - ホストノードが，複数のタグつきネットワークインターフェイスを保持
  - 仮想ノードのネットワークインターフェイスへマップ
- 仮想ノードの設定は必要ない
  - 仮想ノードの内部から制約を回避することはできない

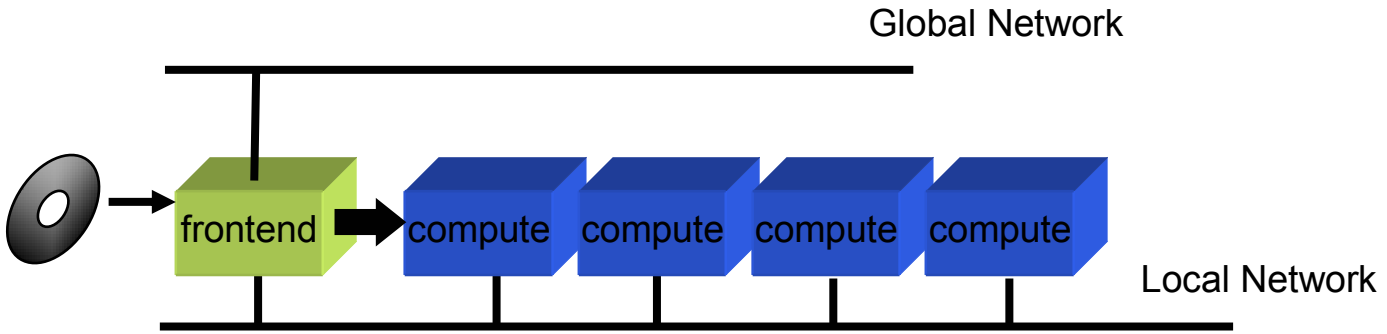


Rocks の概要

- NPACIの一環としてUCSDで実装されたクラスタ管理システム
- クラスタ全体のインストールと，インストール後の管理をサポート
  - 「Roll」という形で比較的祖粒度のアプリケーションパッケージを提供
    - 例： HPC Roll, Grid Roll
  - 「アプライアンス」で，各ノードの役割を規定
    - 例： Compute Node, Database Node
  - Ganglia によるクラスタモニタリングを提供
  - 411によるユーザ名空間管理

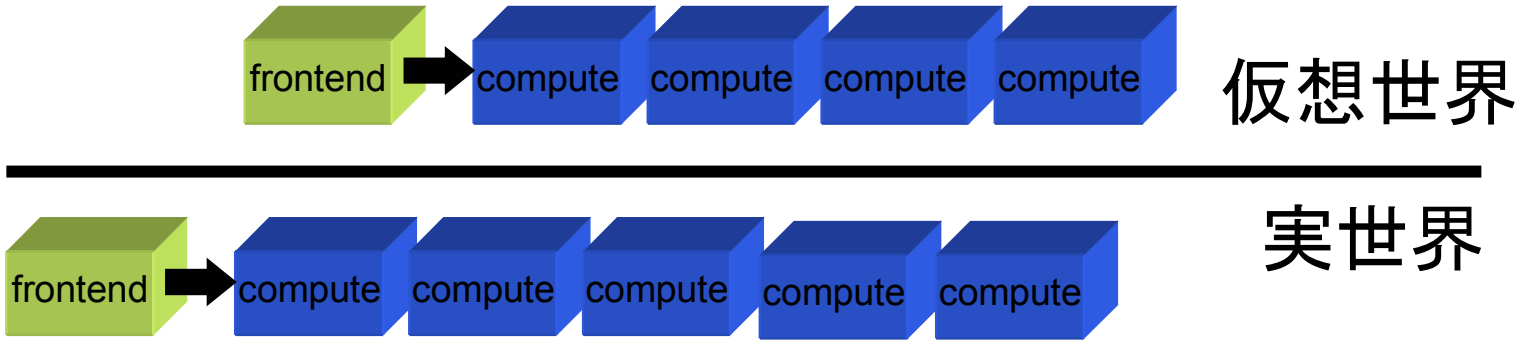
Rocksによるクラスタのインストール

- CD（もしくはネットワーク上のセントラルサーバから）フロントエンドをインストール
- Compute ノードを順番に電源投入
  - 各ノードが自動的にフロントエンドからイメージを取得してインストール
  - 順番に電源を入れることで，ノード名を暗黙裡に指定



仮想クラスタとRocks

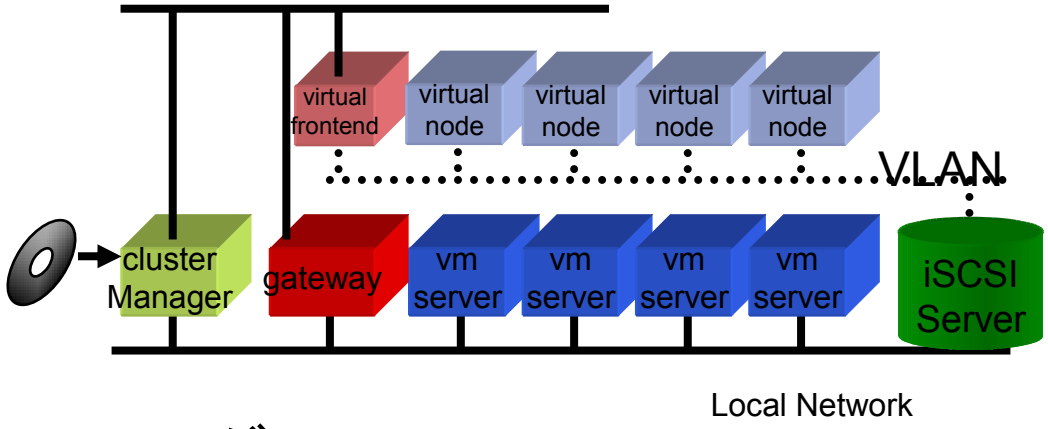
- 仮想クラスタ上に仮想フロントエンドをインストール
  - 仮想フロントエンドから仮想ノード群をインストール



- 仮想クラスタ管理システムを含む実クラスタもRocksを用いてインストール
  - 実クラスタの管理も容易

仮想クラスタの構成

- 4種類のノード
  - クラスタマネージャ
    - クラスタ全体に1機
  - ゲイトウェイノード
    - 仮想フロントエンドをホスティング
    - 外部ネットワークにも足を持つ.
  - VMサーバノード
    - 仮想計算ノードをホスティング
  - ストレージノード
    - iSCSI によるストレージの提供.



動作の概要

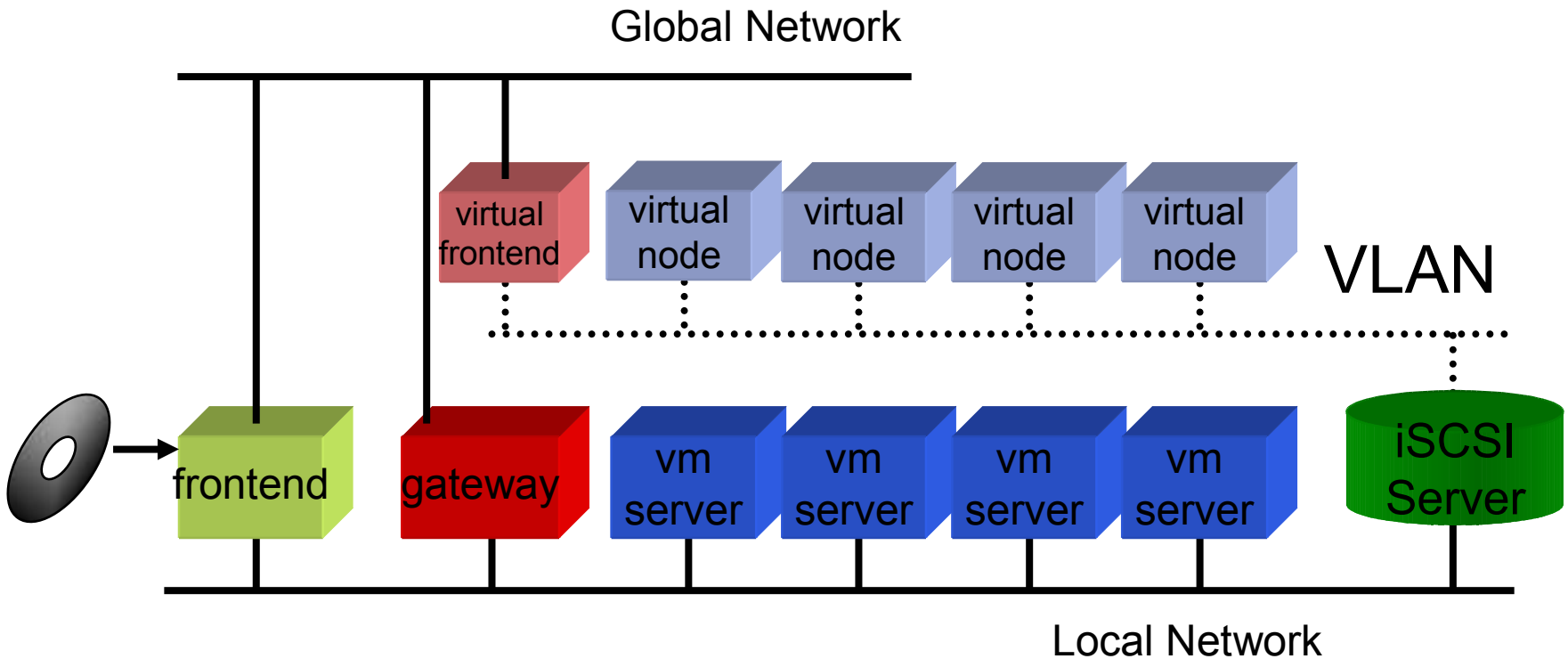
1. サービスプロバイダがWeb インターフェイスを通して、仮想クラスタを予約
   - 開始時刻. 終了時刻, メモリ, ストレージ
   - Roll, Appliance
   - ssh 公開鍵
2. 予約開始時刻
   - 仮想クラスタが起動
     - ストレージとノードが確保される.
   - Rock のクラスタを仮想空間上に自動構築
     - まず仮想フロントエンドを構築
     - 仮想フロントエンドから仮想ノードを構築

動作の概要(2)

3. サービスプロバイダに仮想クラスタを提供
   設定したssh公開鍵が登録され，ログイン可能

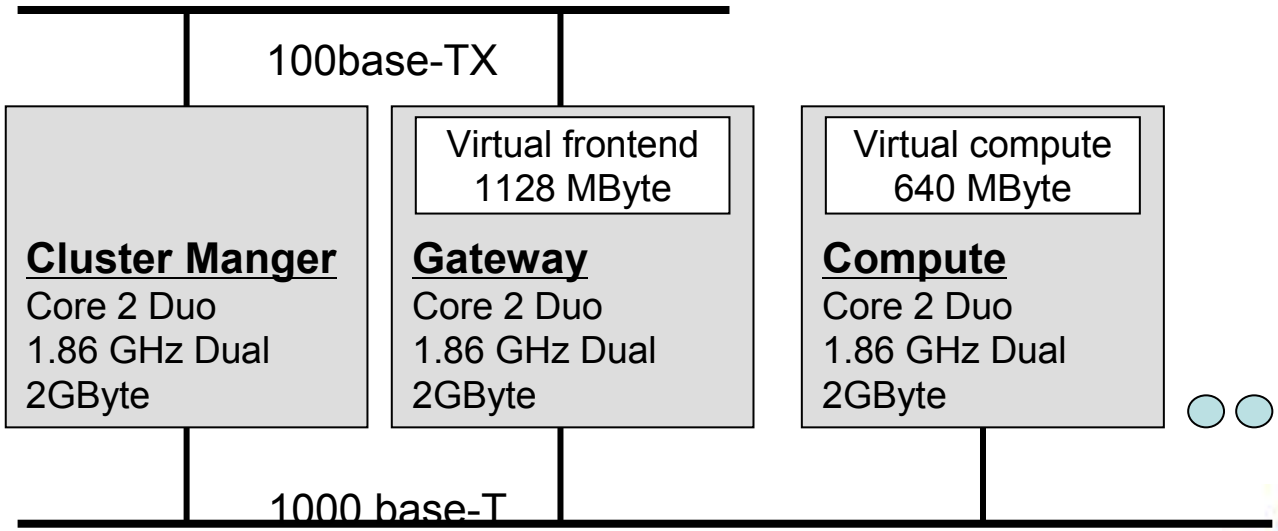
4. 終了時刻
   - ストレージと計算ノードを解放
   - OSとしては,特に終了処理をしない

仮想クラスタインストール



測定

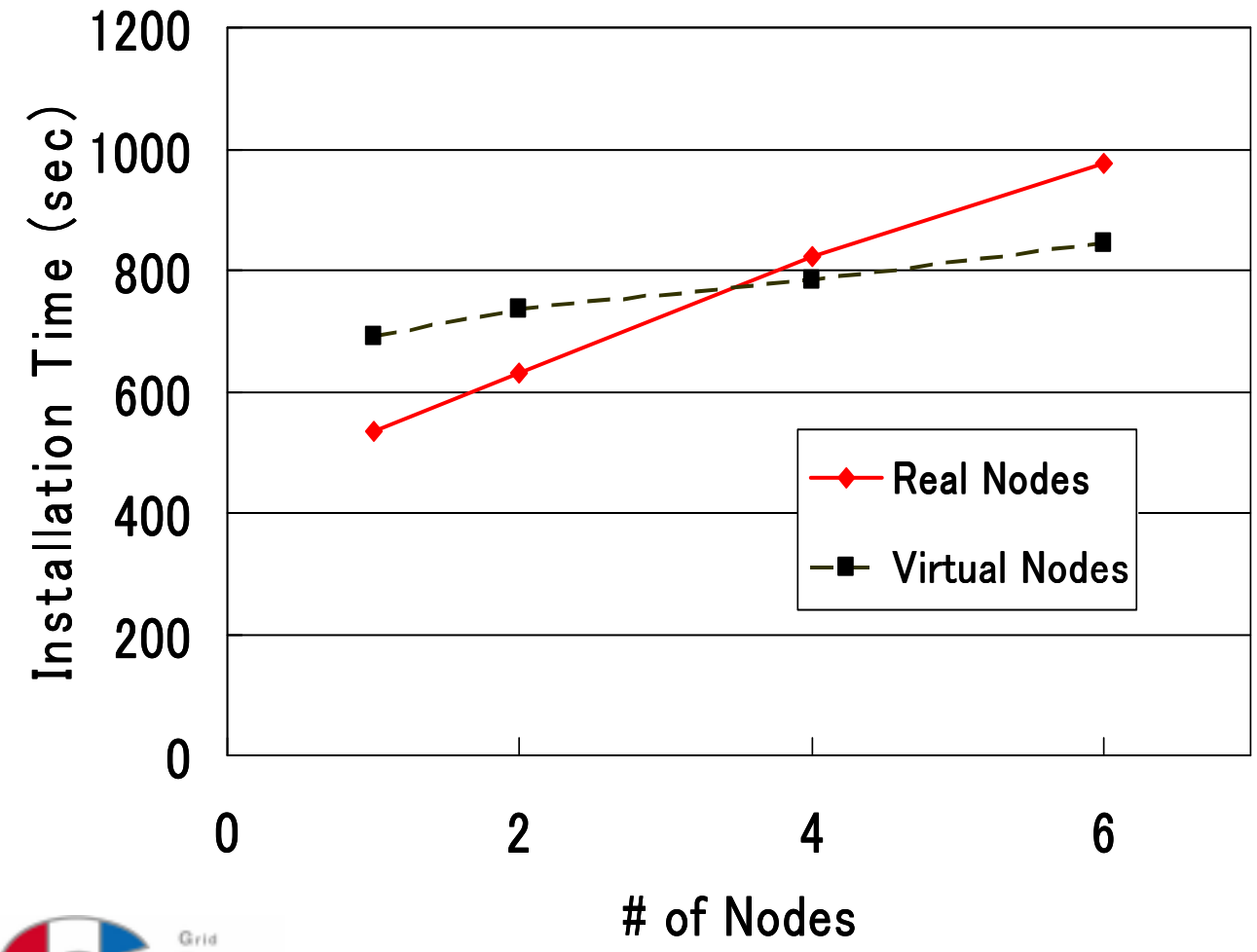
- クラスタのインストール時間を測定
  - 物理クラスタのインストール
  - 仮想クラスタのインストール
  - ノード数を変更して計測



VMware EULA

- 3.3 Restrictions. You may not (i) sell, lease, license, sublicense, distribute or otherwise transfer in whole or in part the Software or the Software License Key to another party; (ii) provide, disclose, divulge or make available to, or permit use of the Software in whole or in part by, any third party (except Designated Administrative Access) without VMware's prior written consent; or (iii) modify or create derivative works based upon the Software. Except to the extent expressly permitted by applicable law, and to the extent that VMware is not permitted by that applicable law to exclude or limit the following rights, you may not decompile, disassemble, reverse engineer, or otherwise attempt to derive source code from the Software, in whole or in part. You may use the Software to conduct internal performance testing and benchmarking studies, the results of which you (and not unauthorized third parties) may publish or publicly disseminate; provided that VMware has reviewed and approved of the methodology, assumptions and other parameters of the study. Please contact VMware to request such review.

測定結果



- 実ノードのインストール時間と，仮想ノードのインストール時間はほぼ同程度
- グラフの傾きが違う理由は現在調査中
  - インストールしているパッケージの集合も異なるので，単純な比較は難しい

関連研究

- ORE Grid [高宮 '06, Nishimura '07]
  - クラスタインストールツール lucieを利用
  - 超高速ノードインストール
- Virtual workspace [Keahey '06]
  - Globus プロジェクトの一環
  - Web Services ベースのインターフェイスで実行環境を構成し，そこにジョブを投入
  - 基本的に，ジョブ単位，ノード単位

関連研究(2)

- OSCAR によるXenクラスタ [Vallee ‘06]
  - OSCAR
    - Rockに相当するクラスタデプロイツール

- Cisco VFrame
  - Infiniband ネットワークとSAN,専用スイッチを用いてストレージとネットワークを仮想化
  - 計算機は仮想化されていない
  - 非常に高価な専用ハードウェアが必須

結論

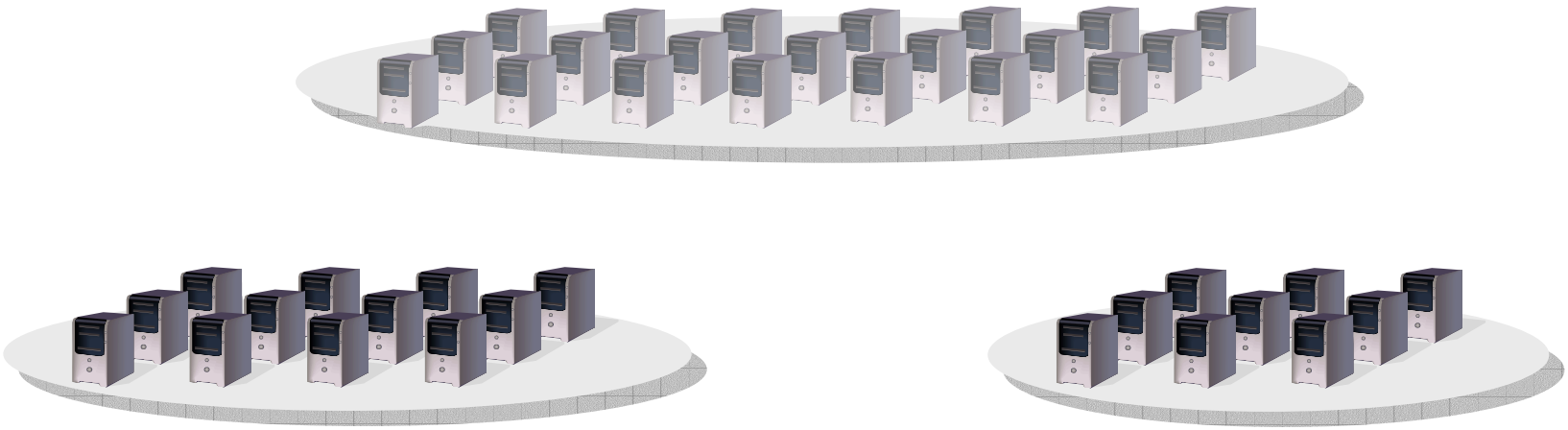
- 仮想クラスタ構築システムを実装
  - Rocksによるクラスタソフトウェアのインストールとコンフィギュレーション
  - 計算機・ストレージ・ネットワークの仮想化
    - VMware Server
    - iSCSI
    - VLAN
- インストール時間を測定
  - 物理クラスタと同程度であることを確認

今後の課題

- インストール時間の詳細な内訳解析
  - インストール時間の短縮
- Xenへの対応
  - CentOS4はXen上のインストールに対応していない
  - RocksがCentOS 5に対応するのを待って対応
- クラスタファイルシステムの提供
  - ストレージへの高速なアクセス
- 他のオペレーティングシステム・ディストリビューションへの対応
  - Windows CCS?

今後の課題(2)

- 複数の物理クラスタにまたがる仮想クラスタ
  - 単一の物理資源では提供できない量の資源をシングルシステムイメージで提供
  - VPNなどの技術を利用



謝辞

- SDSC Rocks team に感謝する
  - Mason Kats
  - Greg Bruno
  - Anoop Rajendra