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[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]

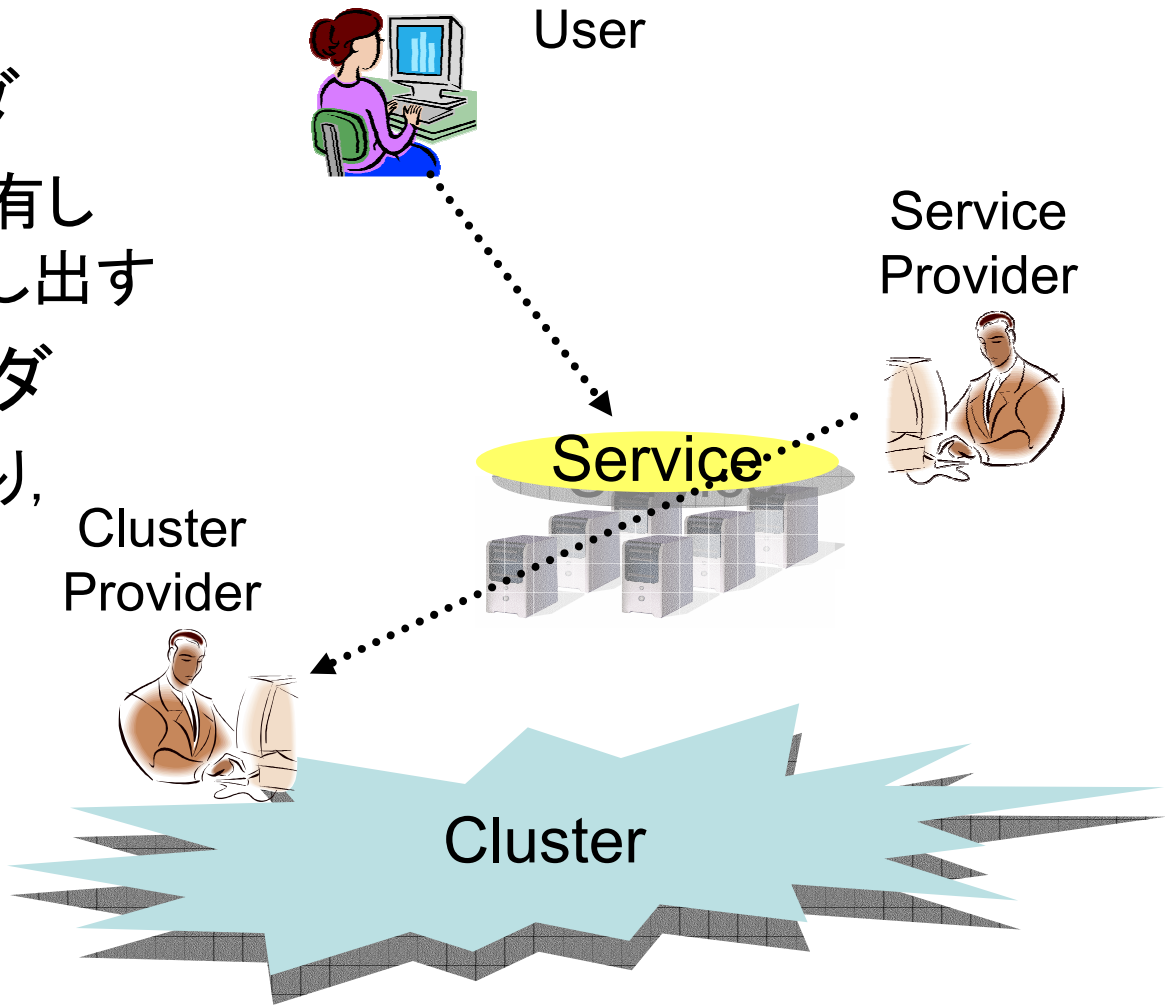
1.産業技術総合研究所
2.数理技研

Grid Technology Research Center AIST

AIST

背景

- 参加者
  - クラスタプロバイダ
    - 物理クラスタを所有し
      仮想クラスタを貸し出す
  - サービスプロバイダ
    - 仮想クラスタを借り，
      サービスを提供
  - ユーザ
    - サービスを利用
- 仮想クラスタの寿命
  - 数日 – 数ヶ月.



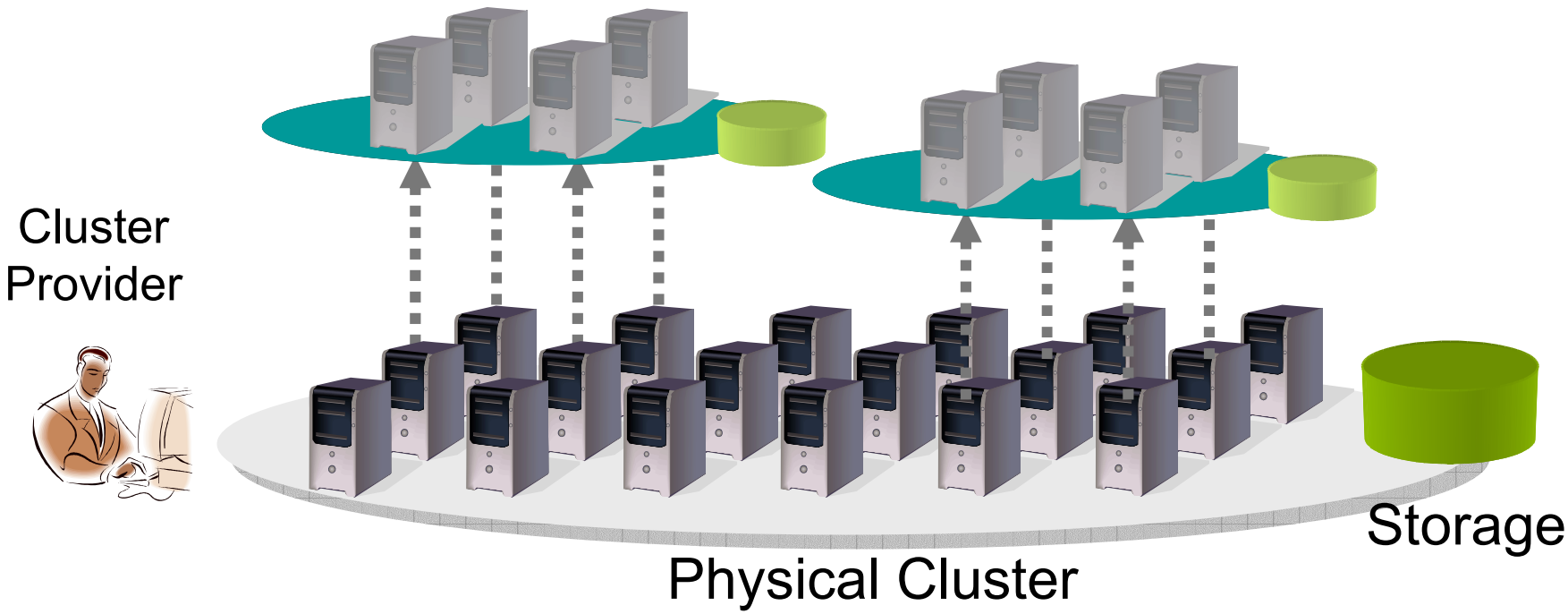
目的

- 「仮想ノード」ではなく「仮想クラスタ」を貸し出す
  - 指定されたノード数，メモリサイズ
  - 共有ディスクスペース
  - 他の仮想クラスタから分離されたネットワーク

クラスタの3つの側面を仮想化

- 計算機仮想化
  - VMWare Server
  - Xen の利用も検討中
- ネットワーク仮想化
  - Tagged VLANによるネットワークの分離
  - ある仮想クラスタから他のクラスタのメッセージを除き見ることはできない.
- ストレージの仮想化
  - iSCSIを用いた，集中管理ストレージの提供

目的

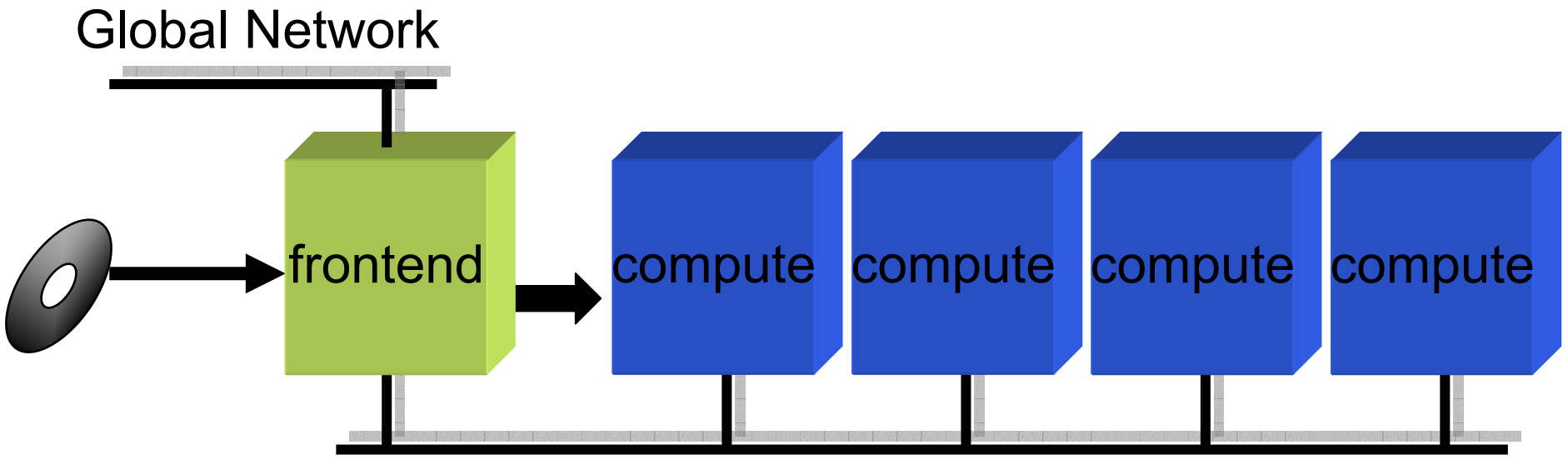


NPACI Rocksによるクラスタのプロビジョニング

- NPACI Rocks
  - UCSDで開発されたクラスタプロビジョニングシステム
    - クラスタをインストールするだけでなく，管理用のツールも提供
      - Ex. 411, Ganglia
    - Roll : “メタパッケージ”
      - Ex. HPC Roll, Grid Roll
    - Appliance : ノードの役割を定義
      - Ex. Compute Node, Database Node
  - MySQL を利用

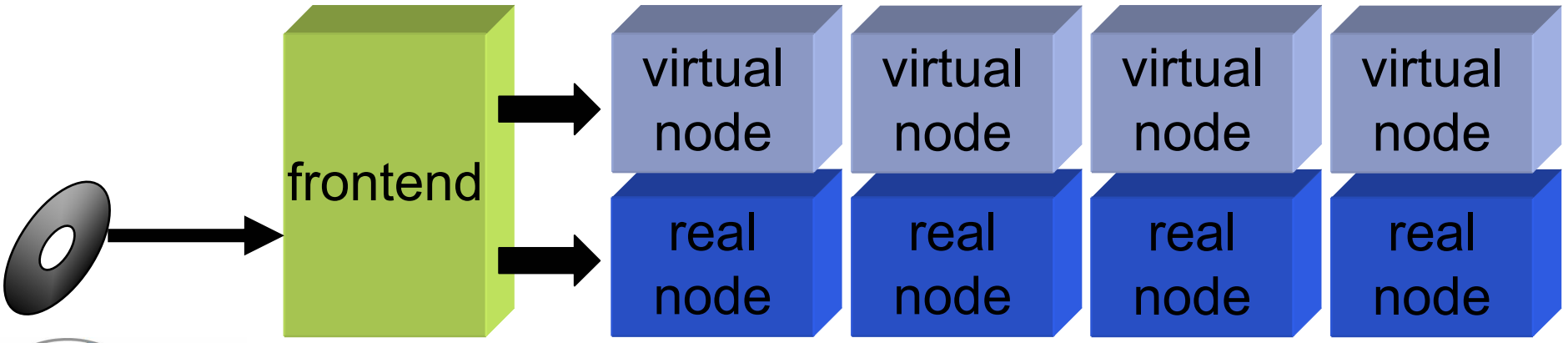
Rocksによるクラスタのインストール

- ‘Frontend’ノードをCDでインストール（もしくはネットワーク経由で‘Central’ノードからインストール）
- 計算ノードをひとつずつ起動
  - PXEでブートして，自動的にインストール

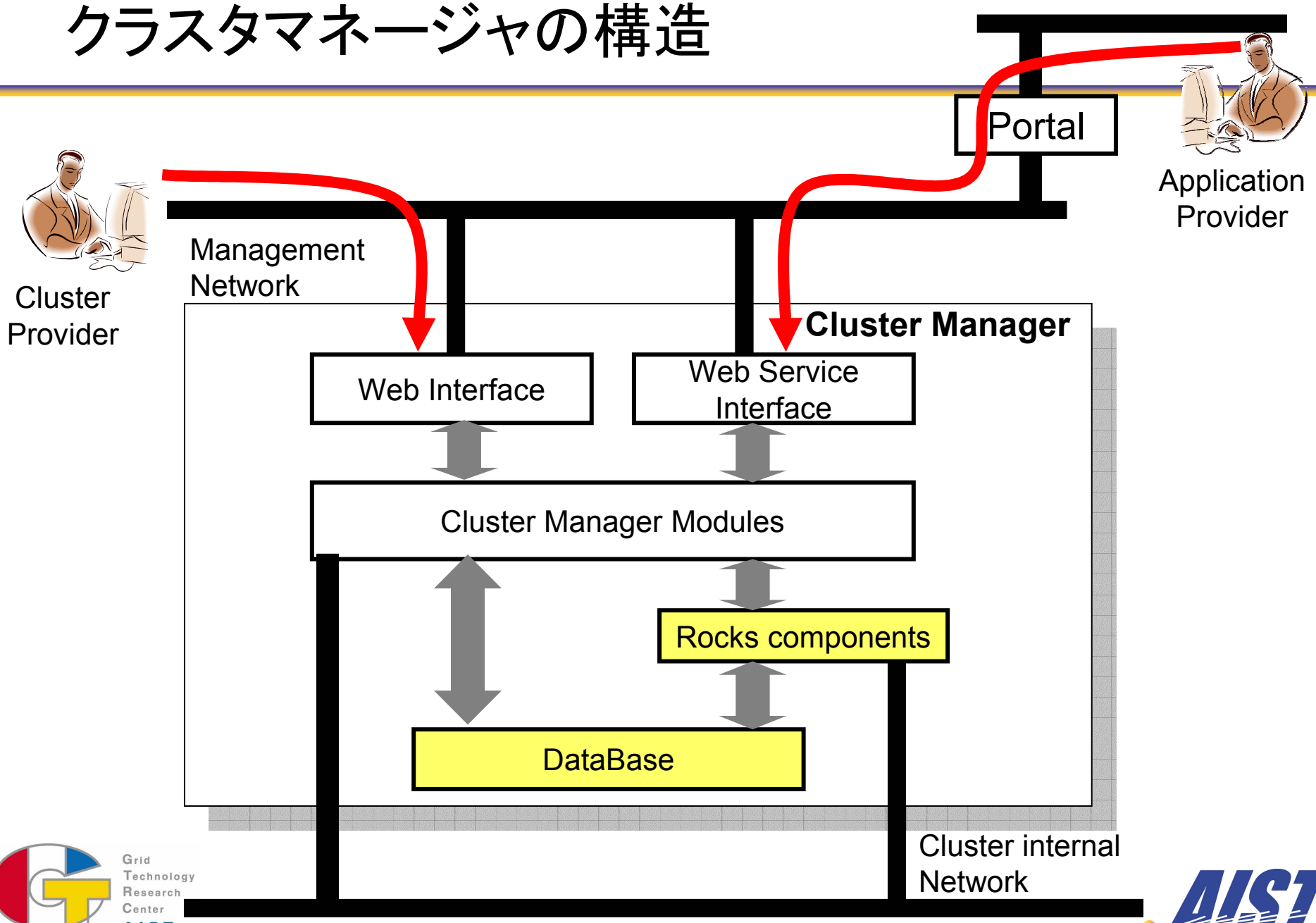


本システムでのRocks の利用

- 物理クラスタと仮想クラスタの双方をRocksでインストール

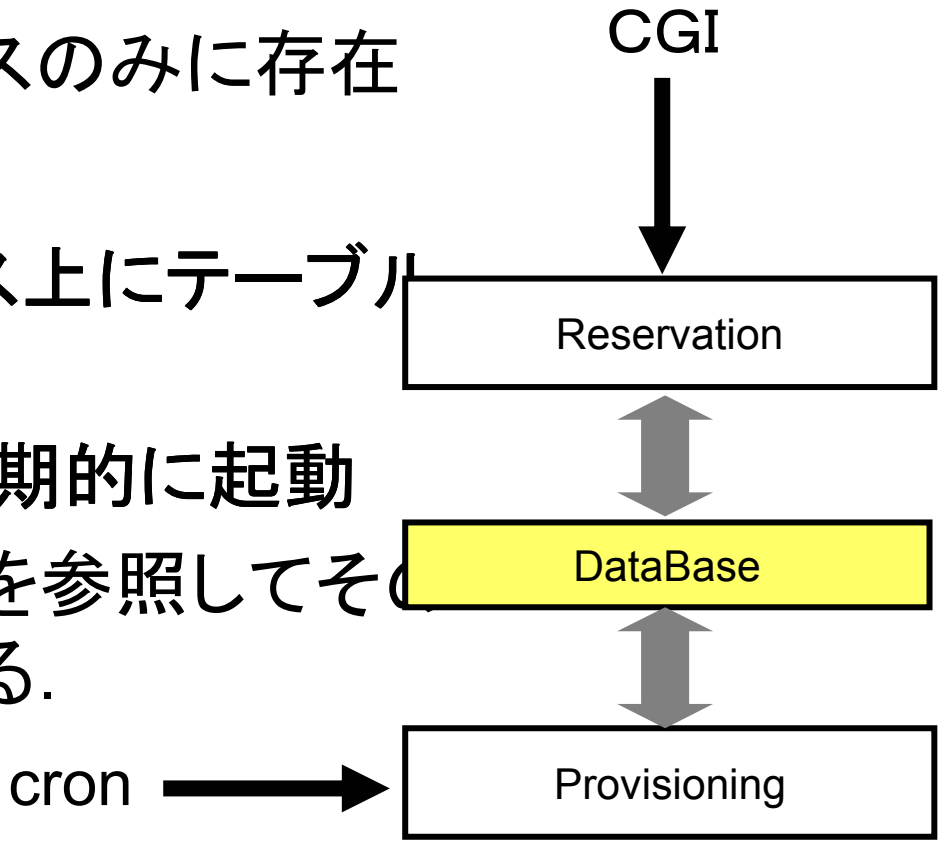


クラスタマネージャの構造



予約とプロビジョニング

- データベースを中心としたアーキテクチャ
  - 「状態」はデータベースのみに存在
  - 耐故障性に貢献
- 予約時にはデータベース上にテーブルを作成するだけ
- ‘cron’がスクリプトを定期的に起動
  - スクリプトがテーブルを参照してそのときするべきことをする.



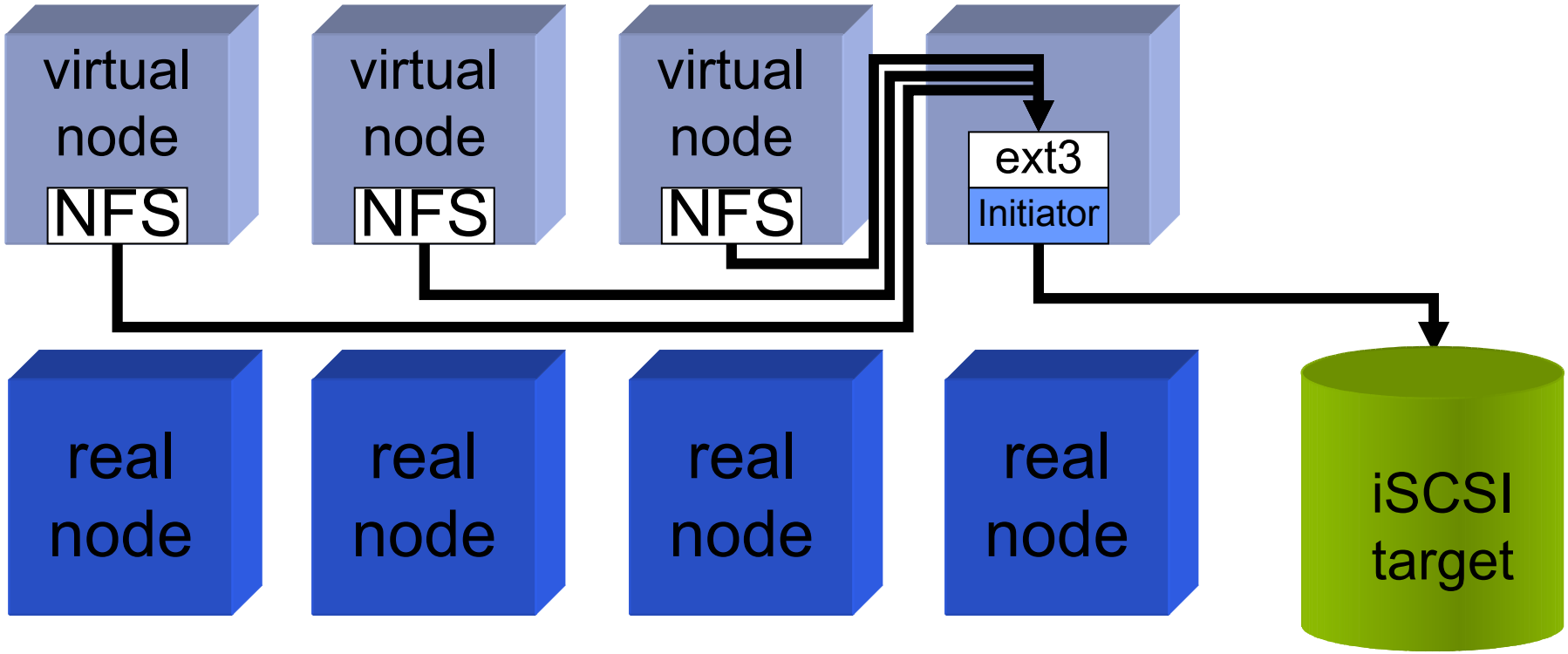
VMWare Server の制御

- ファイルの生成
  - 初期ディスクイメージ
  - VMWareの設定ファイル
- ノード上のVMWare Server を SSHで制御
  - SSH key はRocksが自動的に設定.

iSCSI によるストレージの仮想化

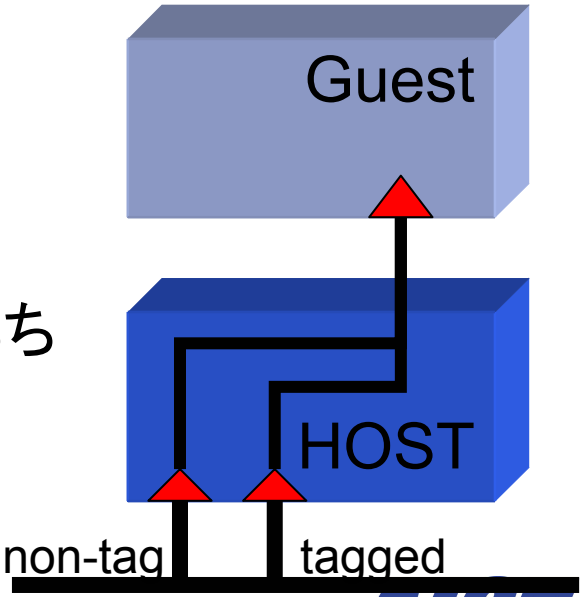
- iSCSI
  - SCSI over IP
  - クライアントからは通常の SCSI デバイスに見える
- 利点
  - ターゲットは論理ボリュームを提供
    - 物理的なディスクの容量と関係なく提供できる.
  - 特別なハードウェアは不必要
- 欠点
  - それほど高速ではない.

仮想ストレージの構造


virtual node
NFS
virtual node
NFS
virtual node
NFS
ext3
Initiator
real node
real node
real node
real node
iSCSI target

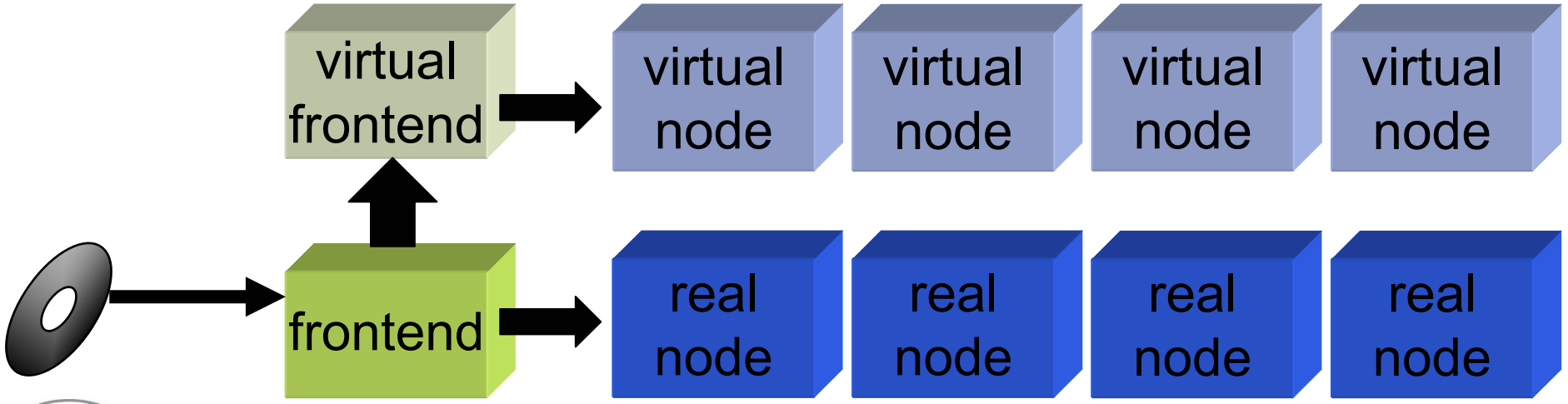
タグ VLANによる分離

- 仮想クラスタのネットワークを他の仮想クラスタから分離
  - 仮想クラスタのノードからは自分のクラスタ内のパケットしか見えない.
- 困難な点
  - ノードインストール時には通常のタグのネットワークが必要.
  - 回答: インストール終了後に動的にタグを切り替える.
- ホストノードででタグの切り替えを実行
  - ホストノードが複数のインターフェイスをもち仮想インターフェイスのマップを変更
  - 仮想ノード内では何もする必要はない.



今後の課題 (1)

- 個々の仮想クラスタに固有の仮想フロントエンドを.
  - Rocks のRollを活用するためには必須.
  - 物理フロントエンドを centralとして仮想計算機上にフロントエンドをインストール



今後の課題 (2)

- 複数の物理クラスタから単一の仮想クラスタを作成
  - VPNを利用してネットワーク接続を確保