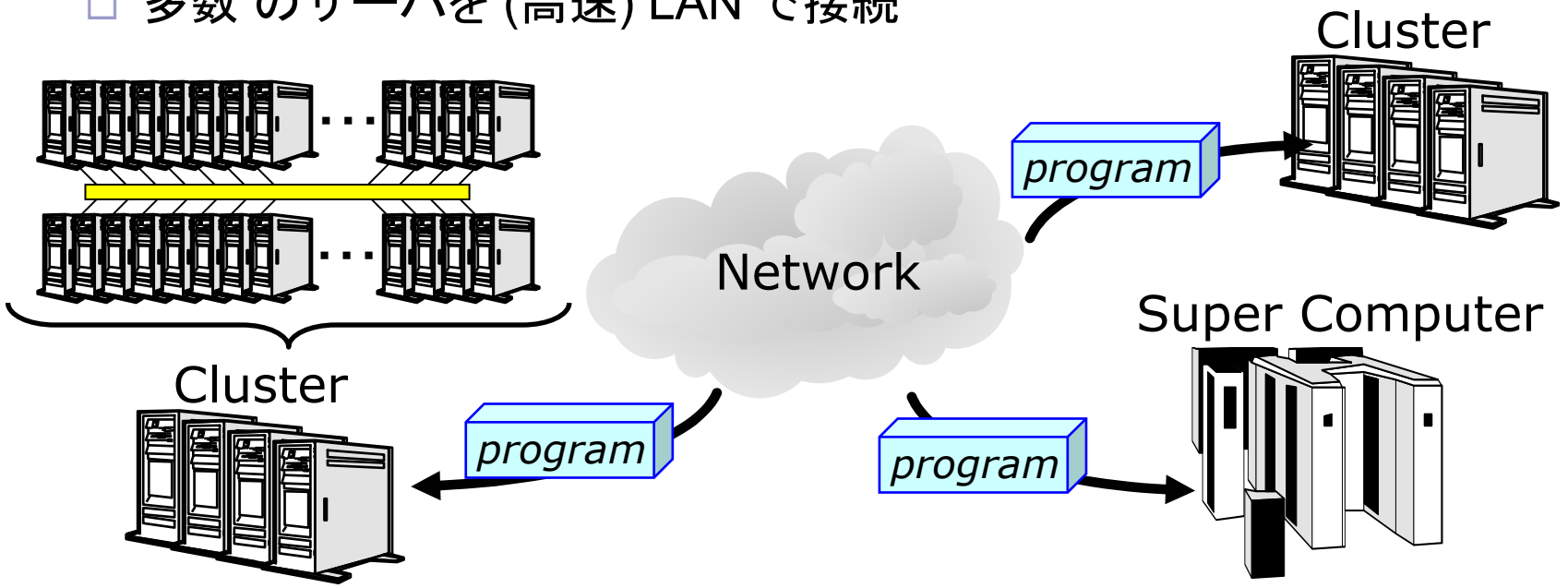
ORE Grid:
仮想計算機を用いた
グリッド実行環境の高速な
配置ツール

高宮 安仁*1, 山形 育平, 青木孝文 (東工大)
中田 秀基(産総研), 松岡 聡(東工大/NII)

＊1 現在 日本電気(株)

背景

- グリッド
  - スパコンやクラスタ等のリソースを広域ネットワークで接続・利用
- クラスタ
  - 多数 のサーバを (高速) LAN で接続



「安い」「汎用」「高速」という価格特性を生かせることから，
クラスタがグリッドリソースの主流となりつつある

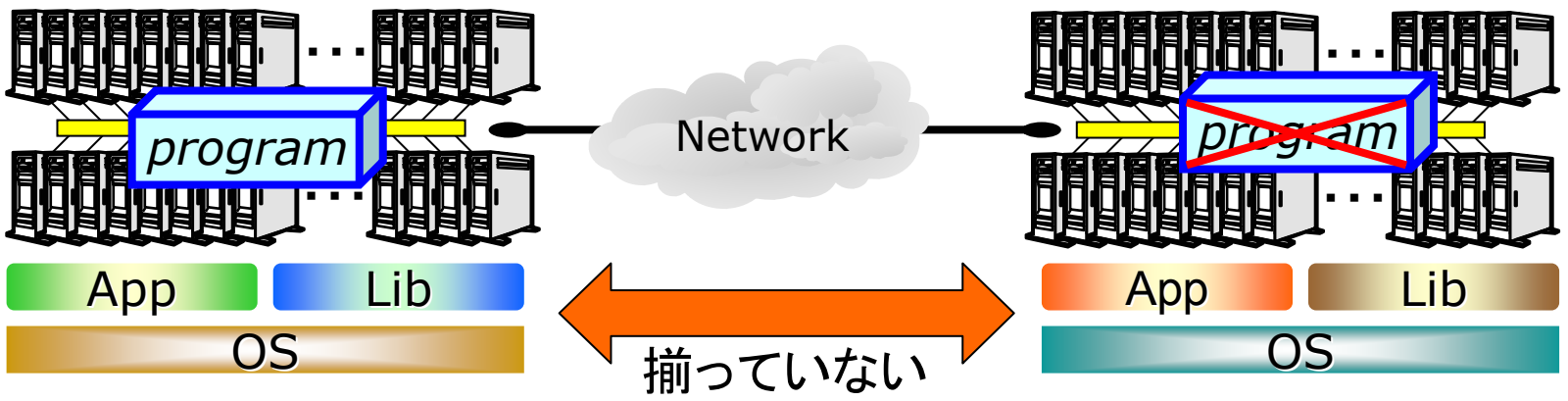
グリッドの現状と問題点

現状:

- ジョブの多様化
- 研究の中心: ミドルウェアやサービス
  - あらかじめ整備された計算リソース (クラスタ) を前提

問題点:

- 相互運用性
- 管理や利用コストの削減



目的と成果

目的:

- グリッド上に多様な実行環境を素早く構築したい
- 管理コストやグリッド利用の障壁を下げたい

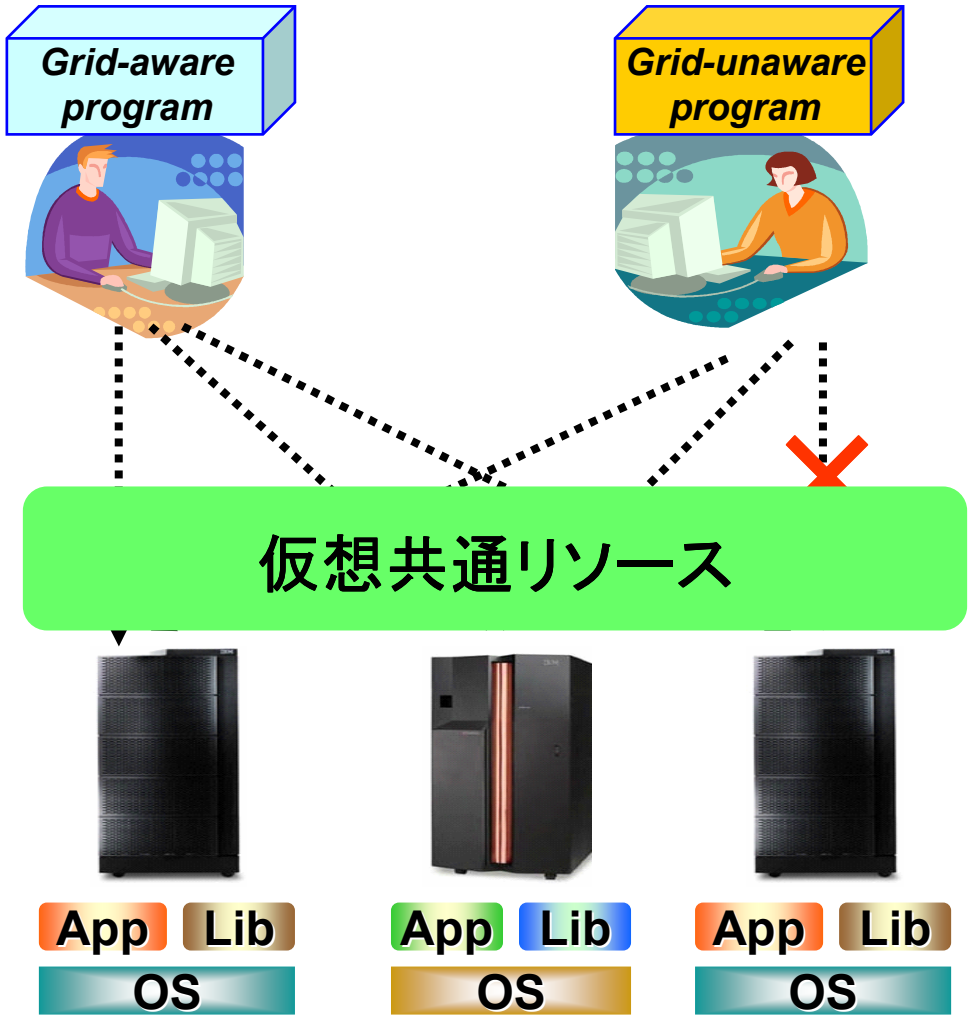
成果:

- 動的なジョブ実行環境構築サービス
  - 低い管理コスト
  - 多様な環境に対応
  - 利用が簡単
  - 高速な環境の構築

背景と問題点:クラスタの相互運用

- ユーザの視点
  - Grid-aware ジョブ:
    どのクラスタでも実行可能
  - レガシーアプリ:
    クラスタの SW 構成を選ぶ
- 管理者の視点
  - 個別対応は大変 OR 不可能

下の構成に依らない
リソースの仮想化・相互運用

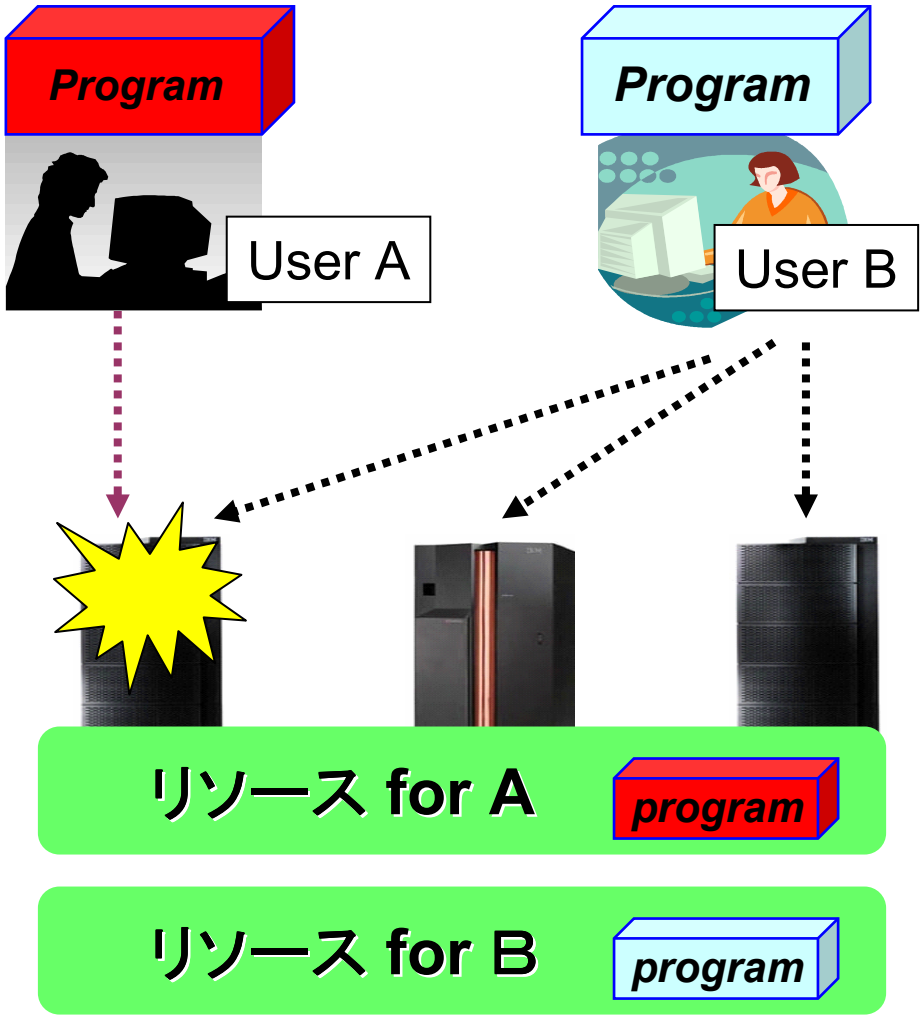


背景と問題点:リソースの安全な共有

- 悪意のある/不安定なジョブ
  - リソースの無駄遣い
  - 攻撃やクラック行為

⇩

他ユーザへの悪影響



管理ドメイン間での
リソース多重化による,
安全な共有機構が必要

解決策: 仮想計算機によるグリッド環境

- VM によるグリッド環境の提案 [Renato et al. ’03]
- 実装
  - The Virtuoso Model [A. I. Sundararaji et al, ’04]
  - VMPlants [I. Krsul et al,’04]
  - Virtual Cluster Workspace [ANL, ’05]
- 安全性 → 解決
- 相互運用性 → 未解決
- とくに，管理者やユーザのコスト，使いやすさが問題

VM グリッドサービス: 関連研究

- 仮想計算機と仮想ネットワークの提供 (Virtuoso)
  - ソフトウェア環境は自分でインストール・設定

  → **環境構築が大変**
- 自動インストール手段の提供 (VMPlants)
  - DAG (有向非巡回グラフ) を用いてインストール手順を手続き的に定義

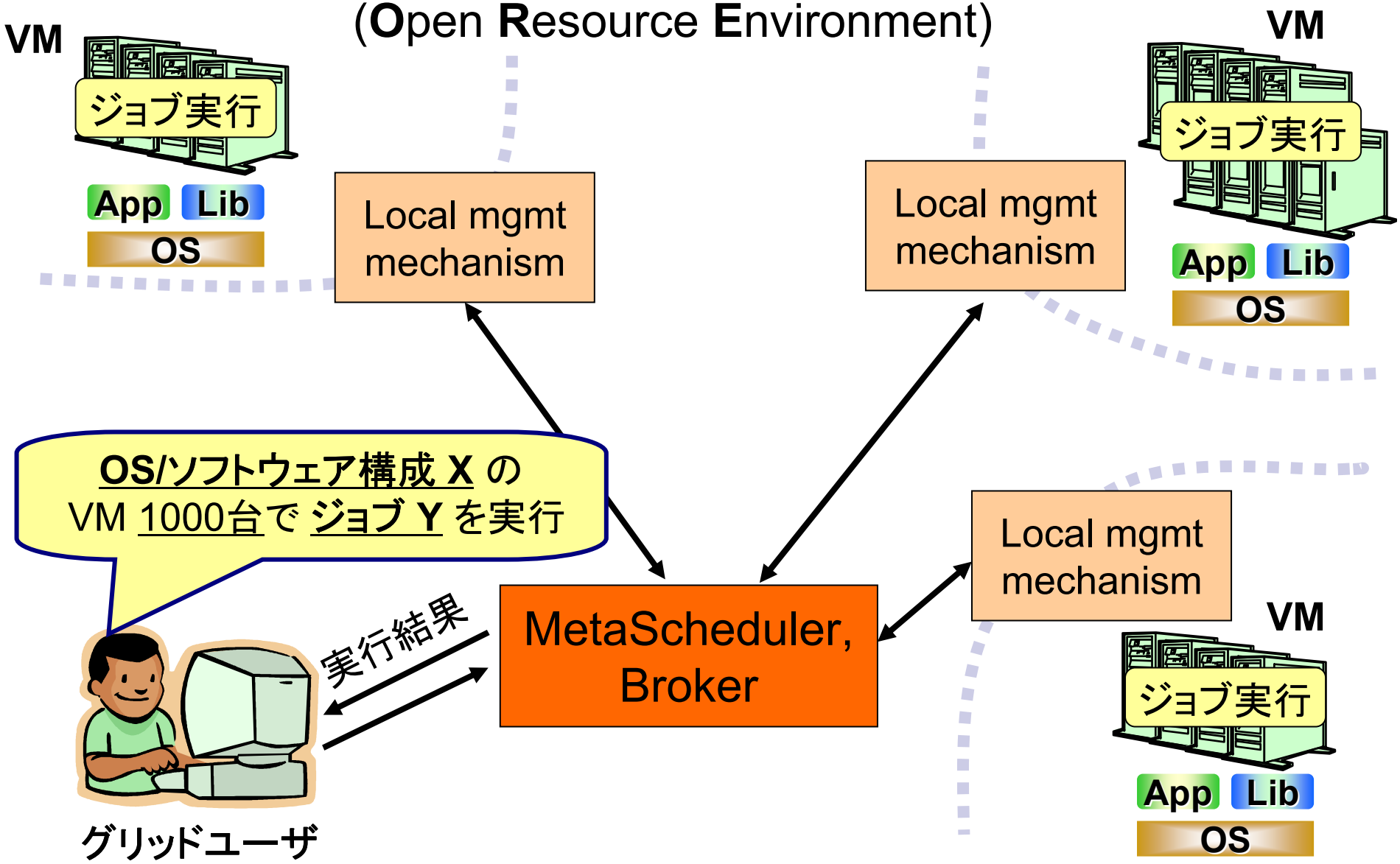
  → **インストール記述が大変**
- ディスクイメージのコピー (Workspace)
  - 管理者が作成したイメージを GridFTP で全ノードへコピー

  → **管理者の手間が大きい**

	Virtuoso	VMPlants	Workspace
容易さ	×	×	×
自動化	×	○	×
多様性	×	×	△

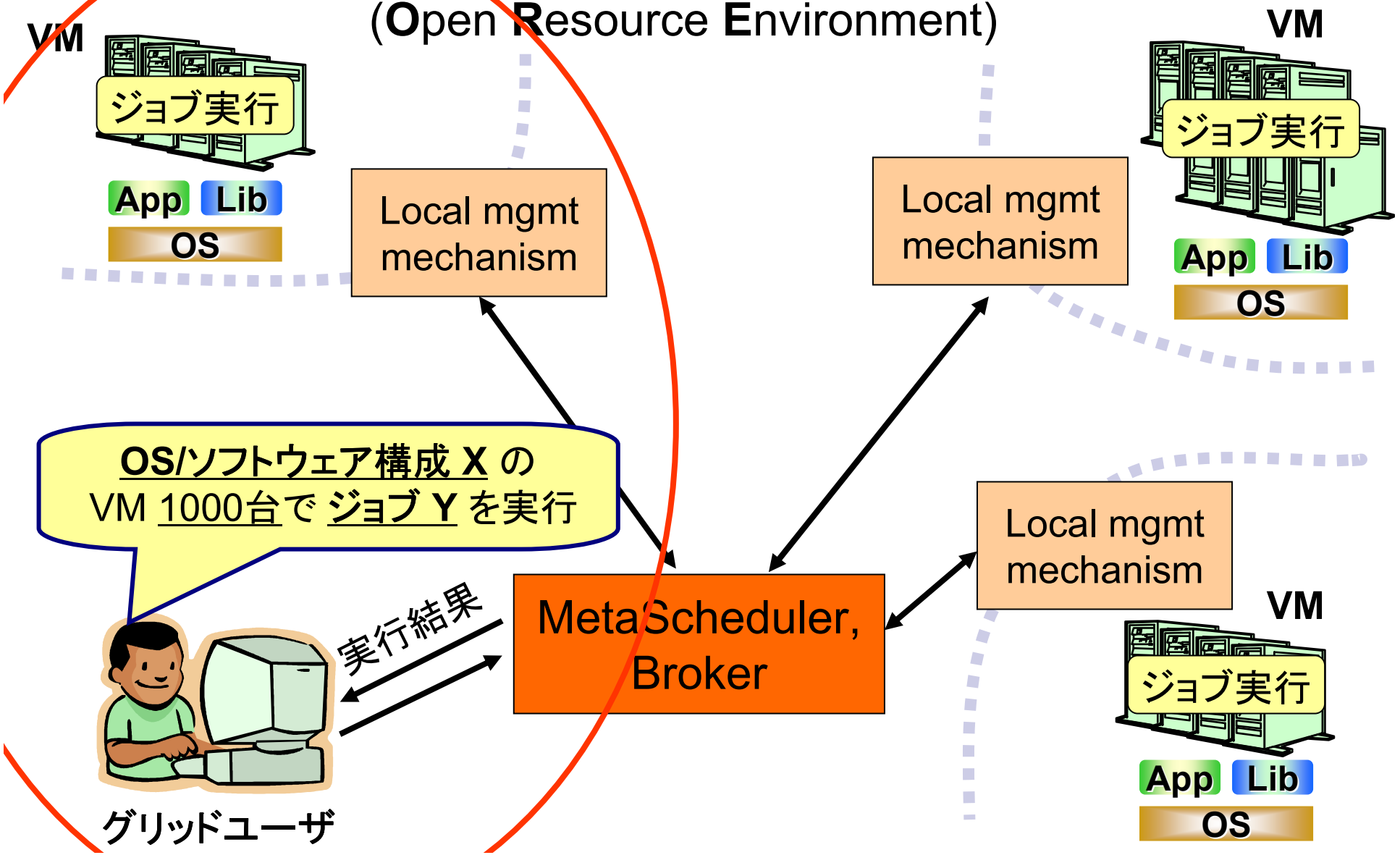
提案システム ORE Grid 全体像

(Open Resource Environment)

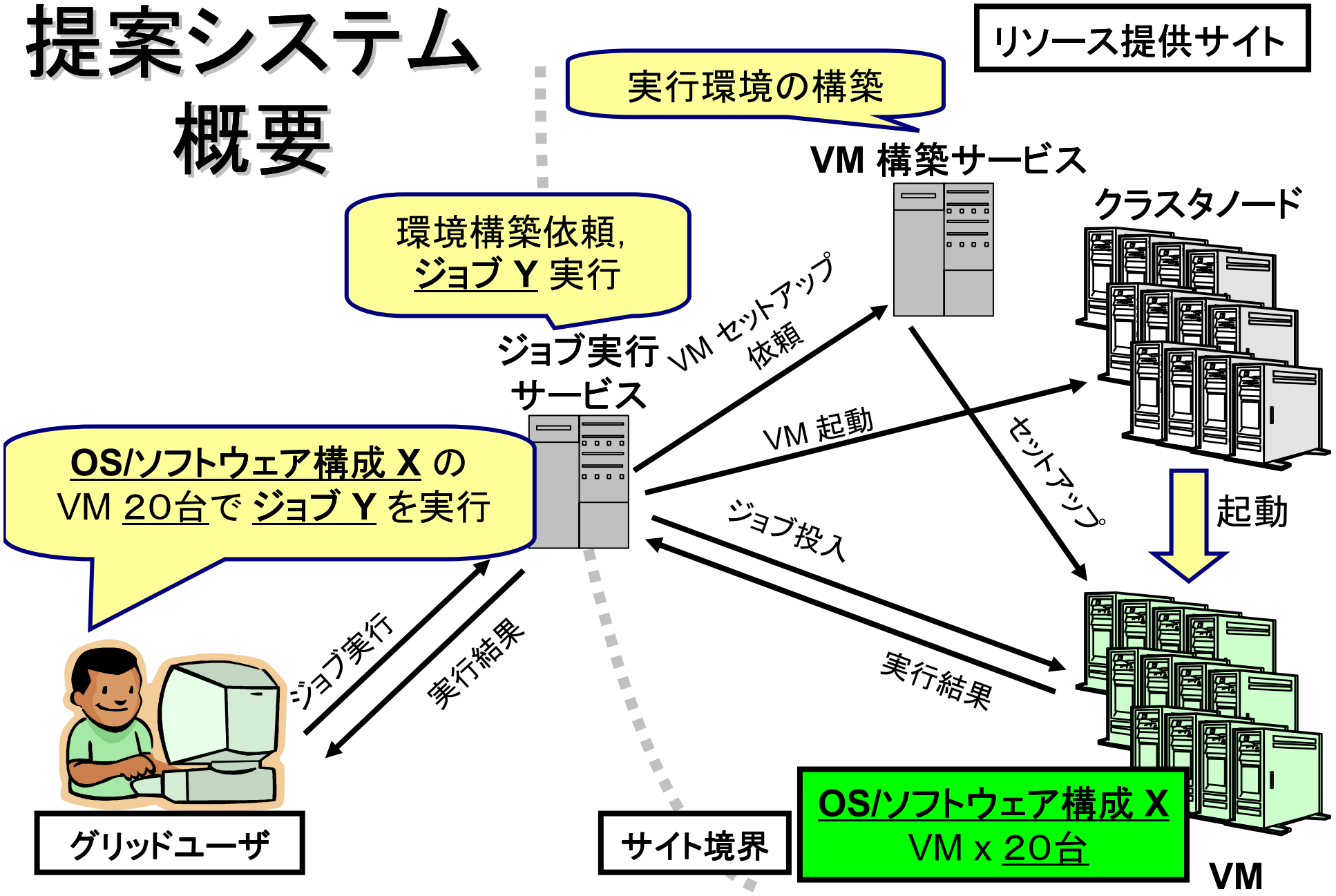


提案システム ORE Grid 全体像

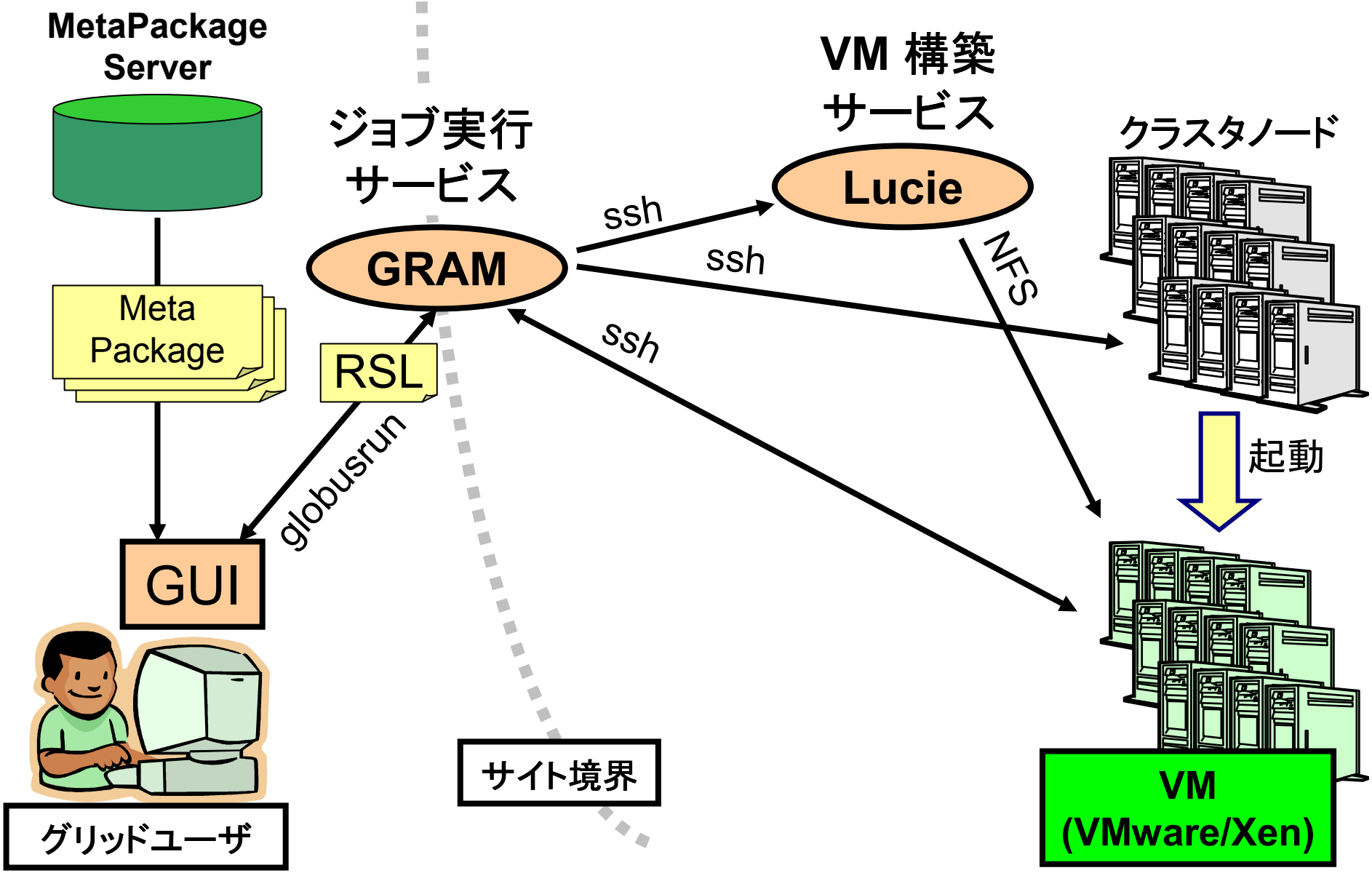
(Open Resource Environment)



提案システム概要


リソース提供サイト
実行環境の構築
VM 構築サービス
クラスタノード
環境構築依頼,
ジョブ Y 実行
ジョブ実行
サービス
VM セットアップ
依頼
VM 起動
セットアップ
起動
OS/ソフトウェア構成 X の
VM 20台で ジョブ Y を実行
ジョブ投入
ジョブ実行
実行結果
実行結果
グリッドユーザ
サイト境界
OS/ソフトウェア構成 X
VM x 20台
VM

提案システム: 実装


リソース提供サイト
MetaPackage Server
ジョブ実行サービス
VM 構築サービス
クラスタノード
Lucie
GRAM
ssh
ssh
ssh
NFS
Meta Package
RSL
globusrun
起動
GUI
サイト境界
グリッドユーザ
VM (VMware/Xen)

自動インストーラツール Lucie

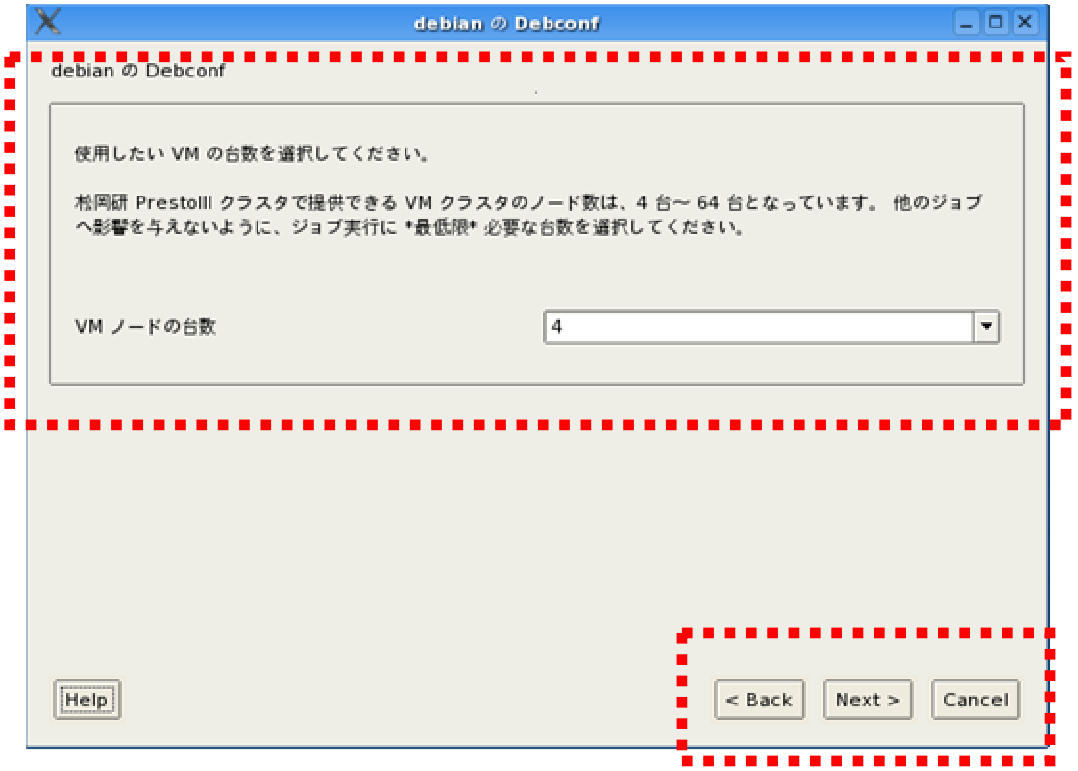
- クラスタ用自動セットアップツール
  - 柔軟な環境構成が可能
  - 完全に自動化されたインストール
  - GUI によるインストール設定

- 特長
  - スクリプトによるインストールのカスタマイズ
  - 対応ディストリビューション: RedHat, Suse, Debian
  - インストール可能な VM: Xen, VMware

メタパッケージ

- ウイザード形式で Lucie の設定を作成
- さまざまな設定テンプレート* をダウンロード可能

説明文，
入力フィールド

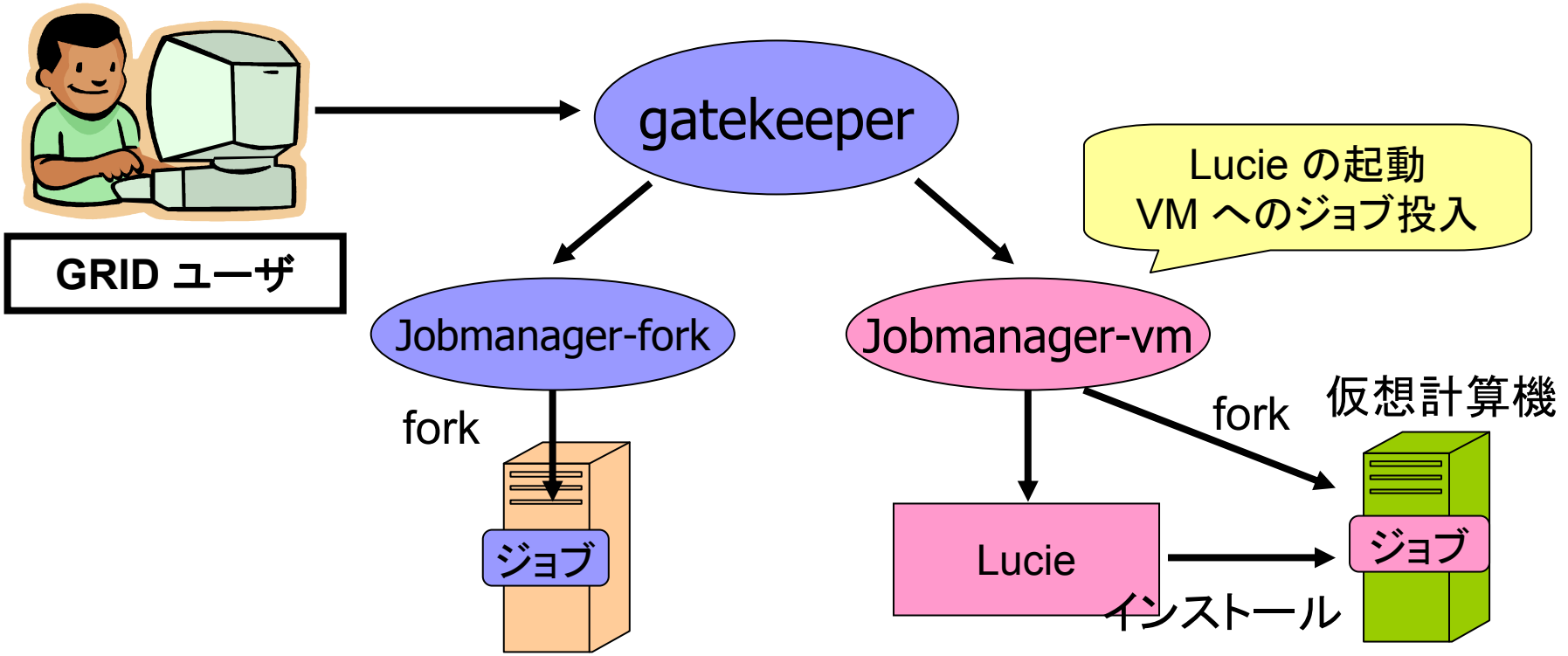


進む＆戻る
ボタン

* http://lucie-dev.titech.hpcc.jp:2500/

実装: GRAM の VM ジョブ対応

- GRAM: GlobusToolkit [I. Foster et al] のジョブ起動機構
  - Gatekeeper デーモンを通じて適切な Jobmanager を起動
- 実装: VM Jobmanager の追加
  - 仮想計算機の起動、Lucie インストーラのセットアップ機能等



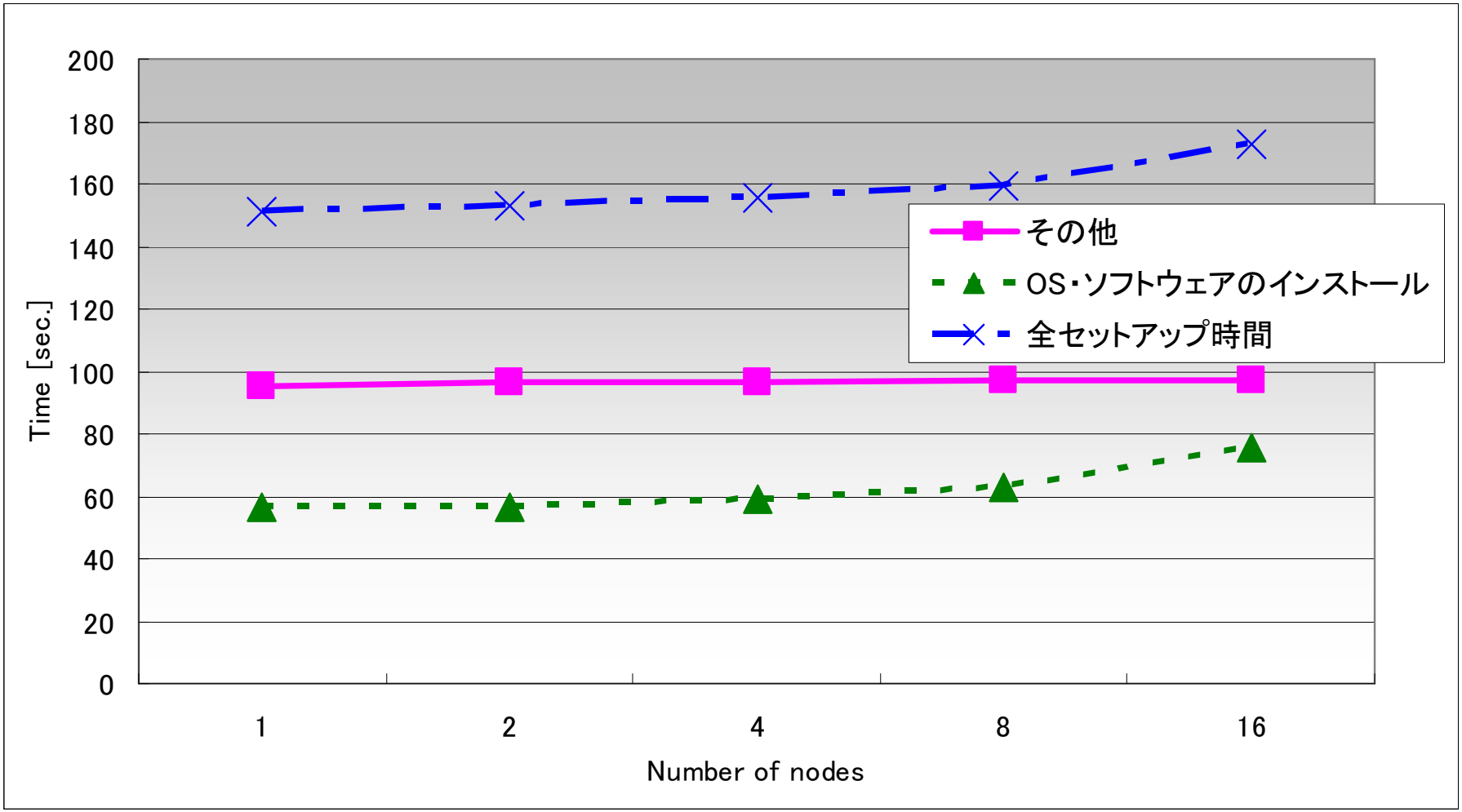
評価:

- ORE Grid 1～16ノード構成の構築時間を計測
- 構築する VM 環境
  - BLAST [S. F. Altschul et al, '90] の実行環境
  - HDD:20GB、Memory:256MB、Kernel: 2.4.28

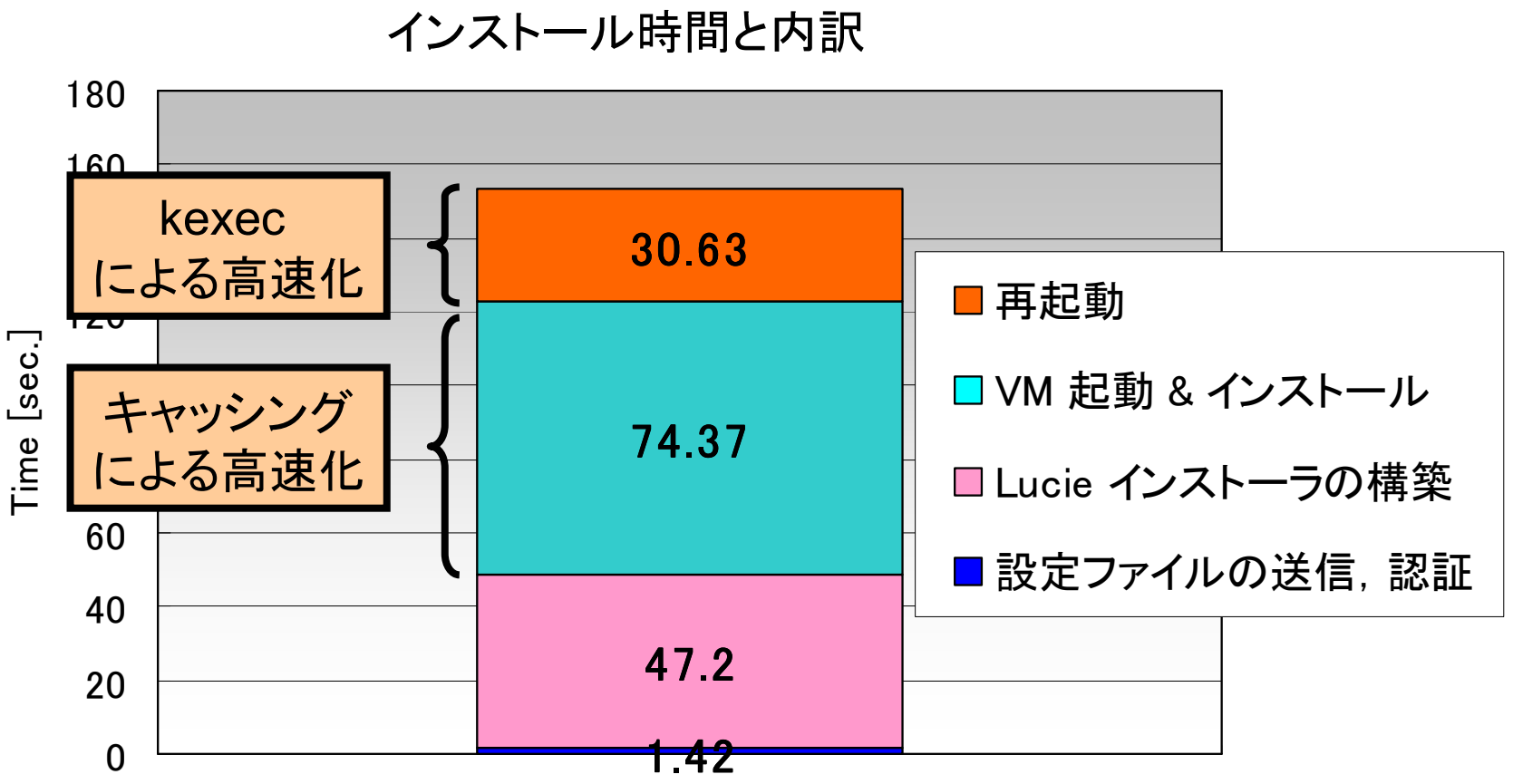
評価環境

CPU	AMD Opteron™ Processor250(2.4GHz)×2
メモリ	PC2700 2GB
OS	Debian GNU/Linux sarge
カーネル	2.4.27
ネットワーク	Gigabit Ethernet
VM	VMware 5.0 Workstation (Linux版)

ORE Grid 環境のセットアップ時間



インストール時間と内訳



- kexec などの再起動高速化によって改善が可能
- VM イメージのキャッシングによって、2回目以降の高速化が可能

高速化の方針

- Kexec [Eric Biederman]
  - リブート処理の高速化
  - ブートローダを経由せず，新カーネルを直接起動
  - ルートファイルシステムの付け替え

- キャッシング [山形ら ’05]
  - 一度セットアップした環境（ディスクイメージ）をキャッシュ
  - インストール性能のモデリング
  - 高速化が望める場合，キャッシュを利用

VM Lucie まとめ

- ジョブ実行環境構築サービスの実現
  - Lucie による VM の自動的な構築
    - GRAM の拡張による VM 構築サービスの追加
    - メタパッケージによる構成のカスタマイズ
- 1ノードあたりの構築時間: 153 秒
  - 構築時間 (153秒) << グリッドジョブ実行時間
  - Kexec やキャッシングによって改善が可能

	Virtuoso	VMPlants	Workspace	ORE Grid
容易さ	×	×	×	○
自動化	×	○	×	○
多様性	×	×	△	○

今後の課題

- 性能向上: ボトルネックの排除
  - 現状: インストーライメージは NFS で取得
    → Rembo などのマルチキャストツールに変更
- メタパッケージライブラリの充実
  - 標準的ツール (NPACI Rocks) 互換化
- スケジューラ
  - 全体のとりまとめ – メタスケジューラ
  - VM を起動するノードの自動選択 – ローカルスケジューラ
- リソース制御
  - サイトポリシに合わせてリソースを制御