’07.2.9

修士論文発表

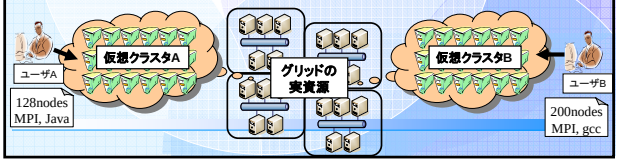
「高速でスケーラブルな仮想クラスタ構築機構」

数理・計算科学専攻
05M37270
西村 豪生

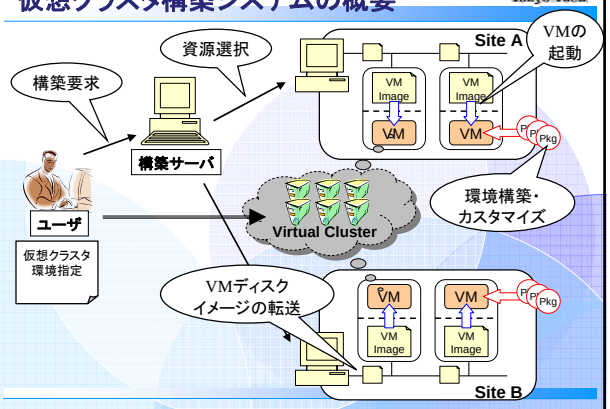
指導教員: 松岡聡 教授

背景：仮想クラスタによる大規模資源共有

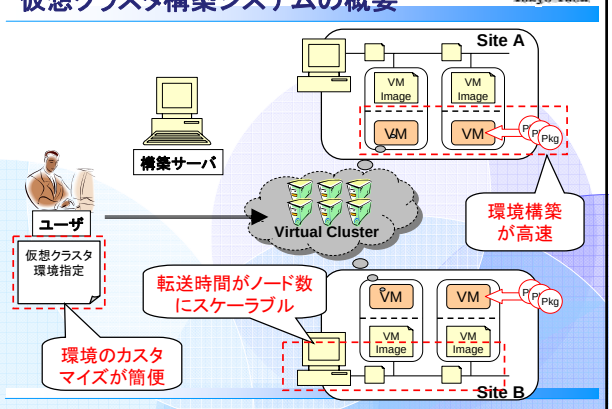
- 仮想クラスタ
  - 仮想計算機(VM)を計算基盤として利用
    - 既存環境に影響を与えずカスタマイズ可能
    - ソフトウェアの不均質性を隠蔽
  - 仮想ネットワークでVM同士を接続
    - ネットワークの非対称性を隠蔽
- ユーザ毎の仮想クラスタを提供することによって大規模資源を効率よく共有



仮想クラスタ構築システムの概要



仮想クラスタ構築システムの概要



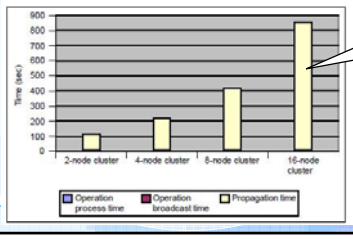
関連研究

関連研究：VMPlants [Krsul *et al.* ’04]

- ユーザはカスタマイズ処理とそれらの依存関係を有効非巡回グラフで記述
- 頻繁に用いられるゴールデンイメージの差分処理のみ行うことで高速化
- ✕ グラフの記述が煩雑であり、ユーザの負担大
- ✕ ユーザが多種多様な場合、ゴールデンイメージを事前に用意するのは困難
- ✕ イメージ転送のスケーラビリティを考慮していない

関連研究：Virtual Cluster Workspace [Foster *et al.* '06]

- Globus プロジェクトによる仮想クラスタ構築機構
- Globus Toolkit と高い相互運用性を持つ
- ✕ 環境のカスタマイズ機構を持たない
- ✕ イメージ転送時間がノード数に線形増加



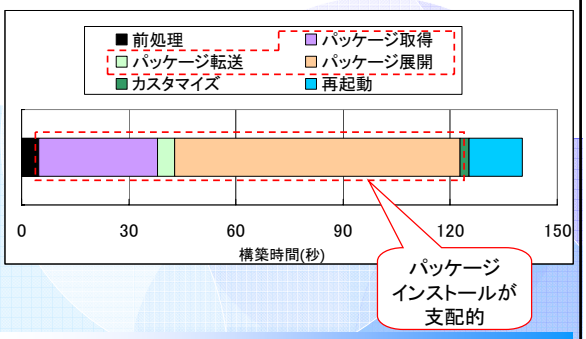
16ノードで850秒

[Foster *et al.* '06]

関連研究：ORE Grid [高宮ら '06]

- ジョブの実行環境を備えた仮想クラスタをクラスタインストーラを用いて動的構築
- 長時間のジョブ実行を想定
- ✕ 単一ノードのインストール時間が大きい
- ✕ NFSを用いているためスケーラビリティが不十分

仮想クラスタ構築時間の内訳



パッケージインストールが支配的

提案

提案手法

- 一部のパッケージを事前に含んでいるディスクイメージ(キャッシュイメージ) からの差分インストール
  - → **支配的なパッケージインストール時間を減少**
- スケーラブルなパイプライン転送を使用
  - → **構築時間がノード数にスケーラブル**
- 既存のクラスタインストーラと連携
  - → **従来のインストーラと同じ要領でカスタマイズ可能**

	提案	VMPlants	Virtual Cluster Workspace	ORE Grid
カスタマイズ性	○	△	×	○
スケーラビリティ	○	×	×	△
構築時間	○	△	×	×

本研究の目的と成果

- 目的
  - ◆提案手法により大規模資源の共有に適した仮想クラスタ構築機構を実現する
    - ●千台規模の仮想クラスタを数十秒で構築することが目標
- 成果
  - ◆過去の履歴からディスクイメージを自動生成する仮想クラスタ高速構築手法を提案
  - ◆プロトタイプ実装を用いて提案手法の有効性を確認

200VM規模の仮想クラスタを30~40秒で構築

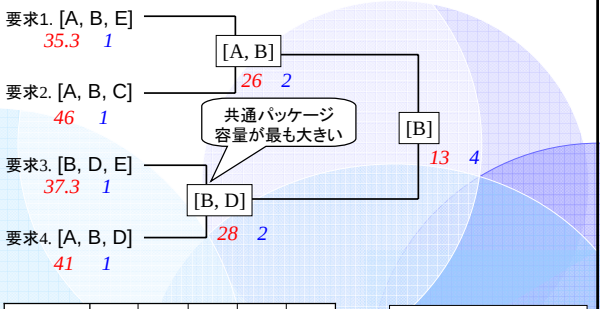
キャッシュイメージの生成アルゴリズム

- ナイーブな方法
  - × 事前に作っておく → どんな要求が来るかわからないので不可能
  - × 要求された全ての組合せを生成 → ディスクがあふれるので不適
- 仮定
  - ◆ユーザのパッケージ要求は偏っており、一定の傾向があると仮定
- 提案
  - ◆過去の履歴から、頻繁に要求されるパッケージの組合せについてのみキャッシュイメージを生成
    - 限られたディスクスペース内で最大の効果を発揮させる
  - ◆過去の要求履歴から頻繁に出現する組合せを抽出
    - 階層的クラスタ解析を応用したアルゴリズム
    - 入力：過去のパッケージ要求集合
      出力：頻出するパッケージの組み合わせの集合

例）{ [MPI make gcc], [condor blast], [MPI gcc], [condor java], … }
→ { [MPI gcc], [condor], … }

提案アルゴリズムによるキャッシュ生成

1. 階層的クラスタ解析によるグルーピング

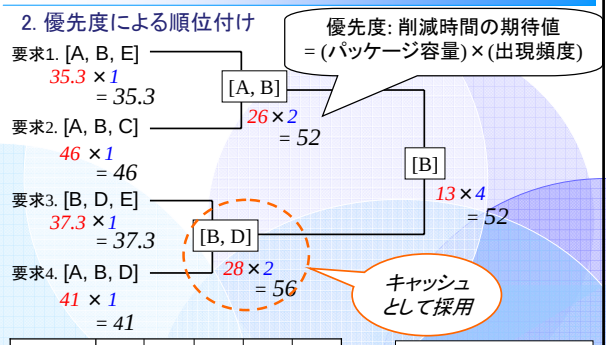


Package	A	B	C	D	E
Size	13	13	20	15	9.3

赤字: パッケージ容量
青字: 出現頻度

提案アルゴリズムによるキャッシュ生成

2. 優先度による順位付け

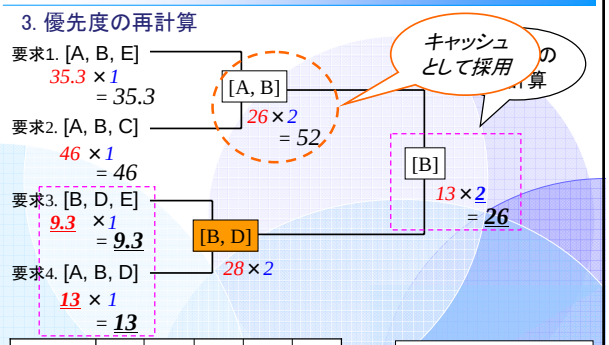


Package	A	B	C	D	E
Size	13	13	20	15	9.3

赤字: パッケージ容量
青字: 出現頻度

提案アルゴリズムによるキャッシュ生成

3. 優先度の再計算



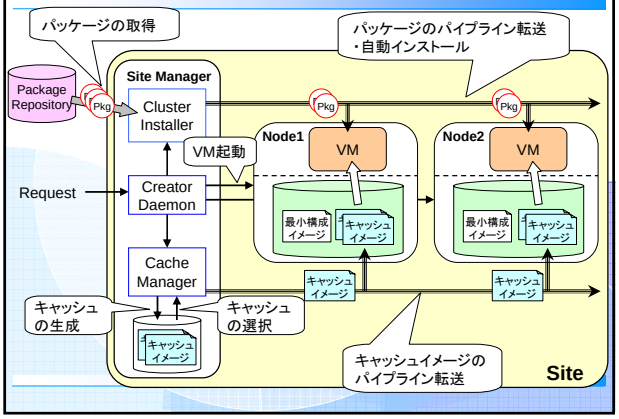
Package	A	B	C	D	E
Size	13	13	20	15	9.3

赤字: パッケージ容量
青字: 出現頻度

キャッシュイメージ選択

- 共通パッケージ容量が最も多いものを最適なキャッシュイメージとして選択
- 最適なキャッシュによって構築時間が減少すると予測されるときのみ使用
  - ◆ノード数、パッケージ容量をパラメータとして過去の履歴から回帰分析
    - ＋ キャッシュイメージの転送時間
    - － キャッシュに含まれるパッケージの取得・転送・インストール時間
  - ◆キャッシュが有効でない場合は計算ノードに配備済みの最小構成イメージを使用

提案システムによるインストール処理の概要



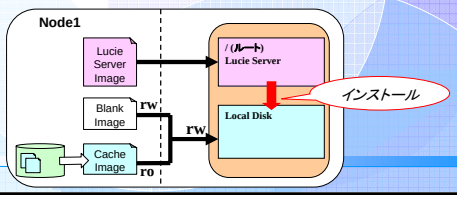
実装

プロトタイプ実装の概要

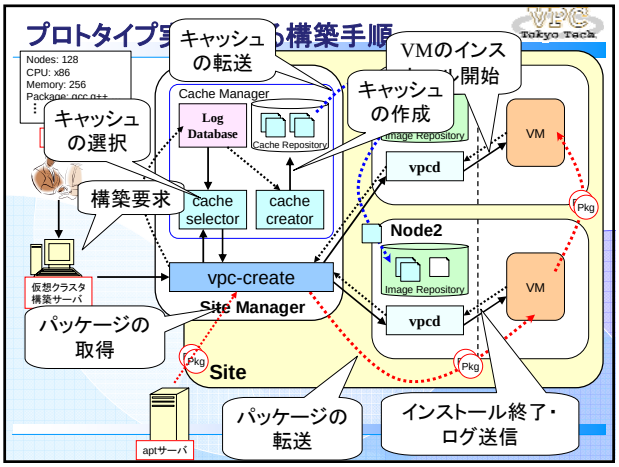
- 仮装計算機モニタ Xen [Barham *et al.* '03] を使用
- クラスタインストーラ Lucie [高宮ら '03] を改変
  - Debian Linux とそのパッケージ管理機構 DPKG/APT を利用
- 二種のパイプライン転送を用途に応じて使用
  - Dolly+ [Manabe '01]
    - 大容量イメージの転送に適している
  - MPICHのブロードキャスト [Thakur et al. '06]
    - 比較的少量のデータ転送に有効

自動カスタマイズ機構の実装

- Lucie のカスタマイズ機構を応用
  - Lucie サーバのディスクイメージを計算ノードに配備
  - ネットワーク通信を介さない
- Stackable Filesystemによるキャッシュイメージコピー時間の削減
  - Read-Onlyのキャッシュイメージに Read-Writeのブランクイメージを重ね合わせて読み書き可能に



プロトタイプ実装による構築手順



評価

評価実験

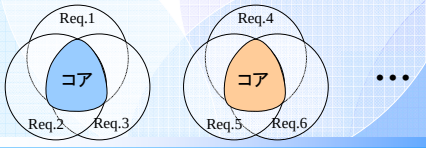
- 評価項目
  - 提案手法による構築時間削減
    - 一定の傾向を持つサンプル要求の構築時間推移を評価
  - 提案システムのスケーラビリティ
    - ノード数の増加による構築時間変化を評価
- 評価環境

	CPU	RAM	HDD	Network	OS
Site Manager	Athlon2000+	1GB	IDE	Gigabit Ethernet	Linux-2.6.12.6
構成1	Opteron242	2GB	IDE		Linux-2.6.16-xen
構成2	Opteron280	4GB	SATA		
構成3	Opteron250	2GB	SCSI		

Software	Xen	Lucie	Dolly+	MPICH	UnionFS
Version	3.0.2-2	0.0.5	0.93-1	1.2.7p1	1.2

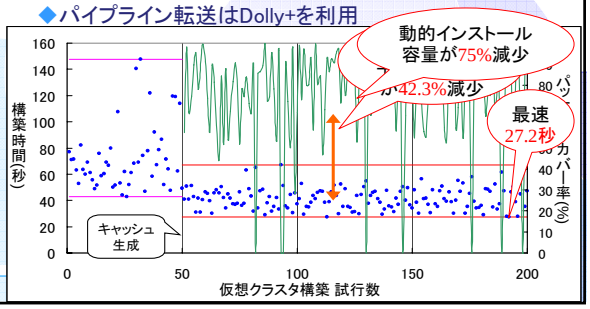
サンプル要求

- ある一定のパッケージ要求傾向を持った要求集合
  - 全て異なった要求であるが、共通部分を持つ
- 基本となるコアパッケージにエキストラパッケージを追加した 720 種
  - コアパッケージ: 16種
  - エキストラパッケージ: 45種



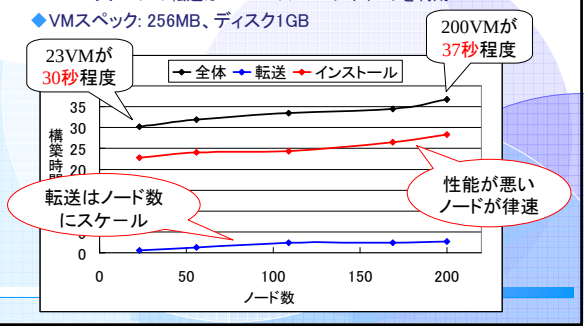
キャッシュ作成前後の構築時間推移

- 50試行毎に5GBのキャッシュイメージを生成
- 50VMの仮想クラスタ構築200試行について測定
  - VMスペック: 256MBメモリ、2GBディスク
  - パイプライン転送はDolly+を利用

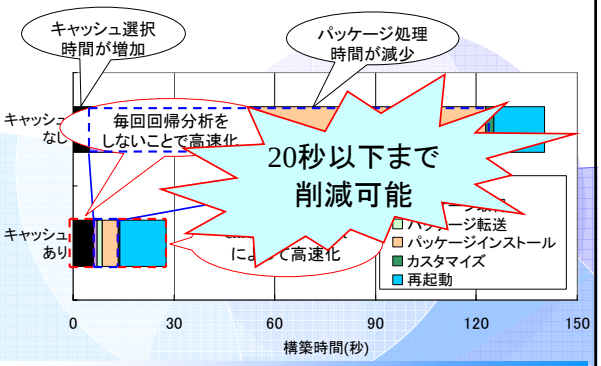


提案システムのスケーラビリティ

- 23～200VMの仮想クラスタ構築時間を測定
  - 5.8MBのパッケージをインストール
    - パッケージの転送はMPICHのブロードキャストを利用
  - VMスペック: 256MB、ディスク1GB



構築時間の内訳



まとめ

- まとめ
  - キャッシュイメージを自動生成する高速でスケーラブルな仮想クラスタ構築手法を提案した
  - プロトタイプ実装を用いて提案手法の有効性を示した
    - 50VMの仮想クラスタが最速30秒以内で構築
    - 1000VMにスケールしても数十秒以内で構築可能
- 今後の課題
  - パッケージインストール部分以外の高速化
  - 計算資源のスケジューリング
    - 各サイトのキャッシュ保持状況を考慮
    - 計算ノードの性能を考慮
  - 精緻な構築時間の予測
    - 計算ノードの性能などのパラメータを追加

対外発表成果

- 国内研究会発表
  - 西村豪生, 中田秀基, 松岡聡.
    仮想計算機と仮想ネットワークを用いた仮想クラスタの構築.
    In **SACSIS 2006 – 先進的計算基盤システムシンポジウム**
    (ポスター発表)
  - 西村豪生, 中田秀基, 松岡聡.
    仮想計算機と仮想ネットワークを用いた仮想クラスタの構築.
    In **2006年並列／分散／協調処理に関する『高知』サマー・ワークショップ (SWoPP高知2006)**
- 査読付き国際会議
  - Hideo Nishimura, Naoya Maruyama, and Satoshi Matsuoka.
    Virtual Clusters on the Fly — Fast, Scalable, and Flexible Installation.
    In *Proceedings of the 7th IEEE/ACM International Symposium on Cluster computing and the Grid 2007 (CCGrid07)*. (to appear)