

資源予約と連携した階層型分散 資源モニタリングシステムの設計

竹房あつ子¹, 中田秀基¹, 柳田誠也^{1,2},
岡崎史裕¹, 工藤知宏¹, 田中良夫¹

¹ 産業技術総合研究所, ² 数理技研

発表概要

- 背景
- 資源予約と連携した階層型分散資源モニタリングシステムの設計
 - 先行研究の情報サービスシステムの概要
 - システムアーキテクチャ
 - 資源予約との連携
 - インタフェースと情報取得プロセス
 - 資源モニタリング情報のデータ表現
 - モニタリング情報の収集方針
- まとめと今後の課題

グリッドとネットワーク

- グリッドとネットワークプロビジョニング技術により、複数管理組織を跨る**高品質仮想計算基盤**の動的構築が可能に
- **GridARS**コアロケーションフレームワーク[SAC SIS2006]
 - 複数資源マネージャと連携
 - 要求資源性能を保証する資源群を事前予約で確保
 - WSRF(Web Services Resource Framework)標準インタフェース
- 高品質仮想計算基盤の構築事例
 - 2006年, 2007年のG-lambda(日本), EnLIGHTened Computing(米国)の共同実験

G-lambda, EnLIGHTenedの共同実験

- GridARS(GL), HARC(EL)によるコアロケーション

- 資源群

- 3ネットワーク
ドメイン: EL, KDDI研, NTT
- 10クラスタ

- アプリケーション

- MPI
- 可視化
- HDビデオ
ストリーム通信



仮想計算基盤提供の課題

- 構成資源が多様かつ分散しており，複数管理ドメインが存在
→ 利用者による予約資源群の状況把握が困難
- GL, EL実験の予約資源モニタ
 - 分散資源の予約完了／利用可能／解放状態の可視化
 - 全ドメインの全資源情報を開示
 - 予約関連情報のみで資源状況なし
 - 利用帯域, 遅延, CPU利用率, ホスト間のコネクティビティなどが必要
- 開示ポリシーを考慮した資源モニタリング情報提供必要



資源予約と連携した階層型分散資源 モニタリングシステム(DMS)の提案

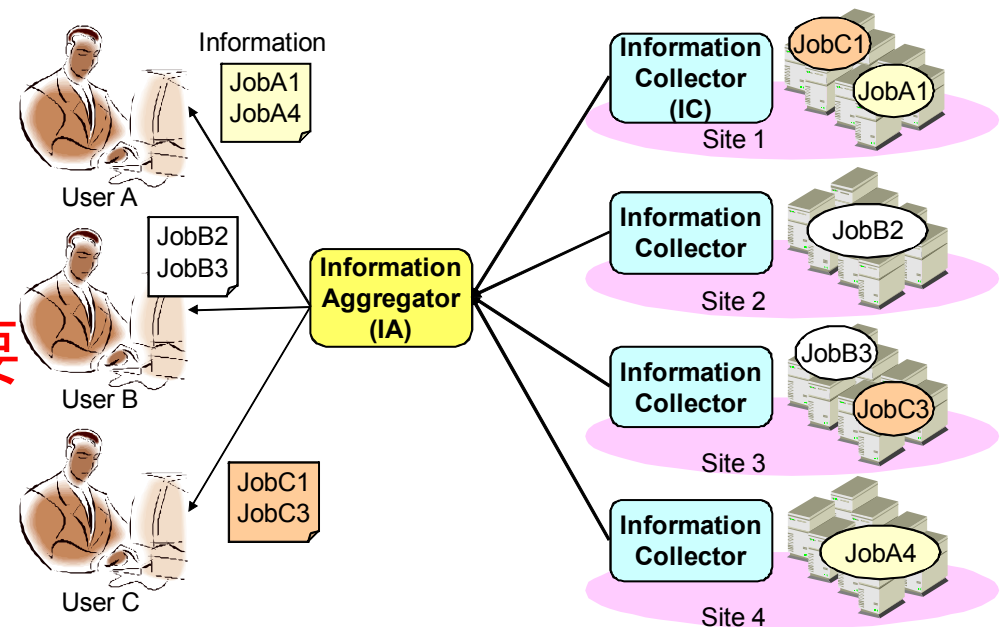
- 先行研究のXACMLを用いた情報サービスシステム [ComSys2007]を改良
- 複雑なドメイン構成を考慮した階層型アーキテクチャ
– ドメインごとに情報開示ポリシーを定義可能
- 資源予約との連携: GridARSフレームワークを利用
- 標準データ表現の拡張(GLUE, v. 2.0) の採用
- 資源モニタリング機構を追加

XACMLを用いた情報サービスシステム

[ComSys2007]

- InformationAggregator(IA)と InformationCollector(IC)の2層構造
 - IA: 分散するICから情報を取得し, ユーザに提供
 - IC: XACMLによる認可に基づき, 情報を提供

- WSRFインタフェース
- 複雑なドメイン構成の実環境への対応と
モニタリング機構が必要



分散資源モニタリングシステム(DMS) の概要

- 実環境では、一組織が複数資源を提供したり (Domain1), 間接的に資源提供したりする(Domain3)

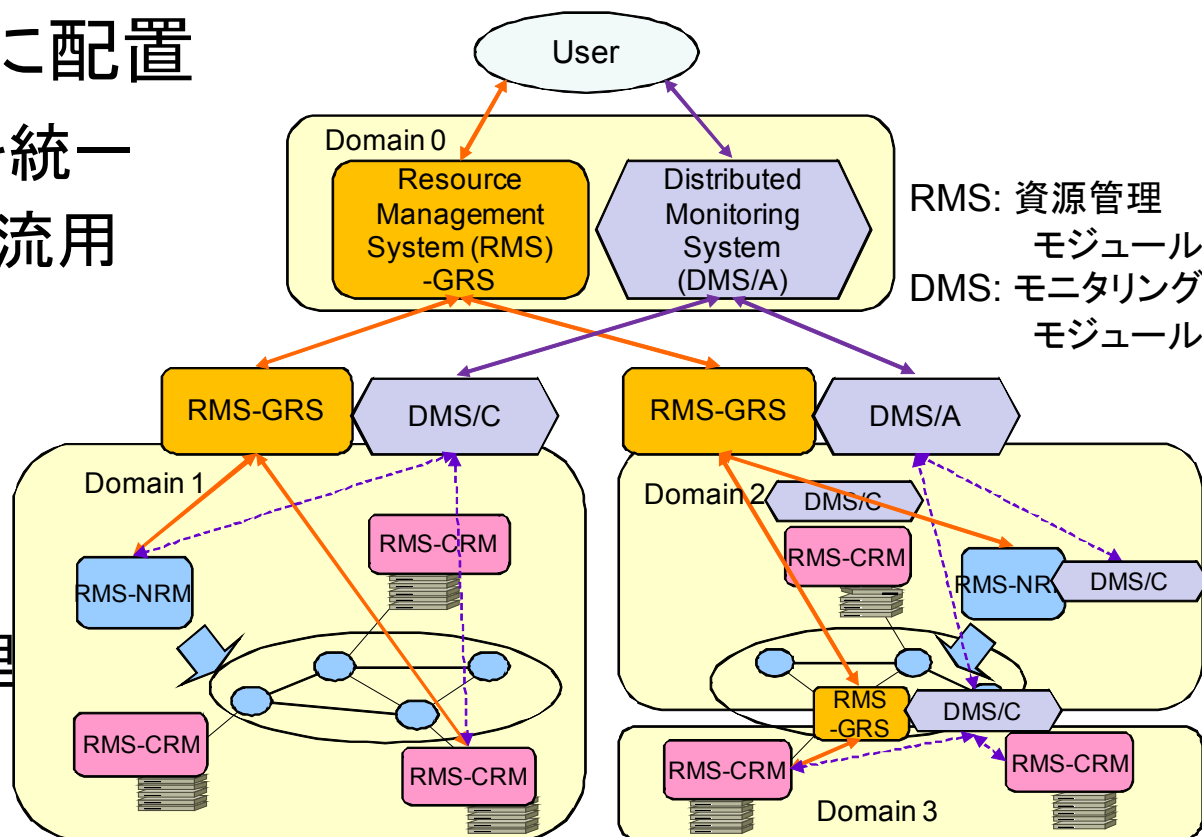
- IA, ICを階層的に配置

- インタフェースを統一
- IA, ICの機能を流用

- ドメインごとに機能を選択

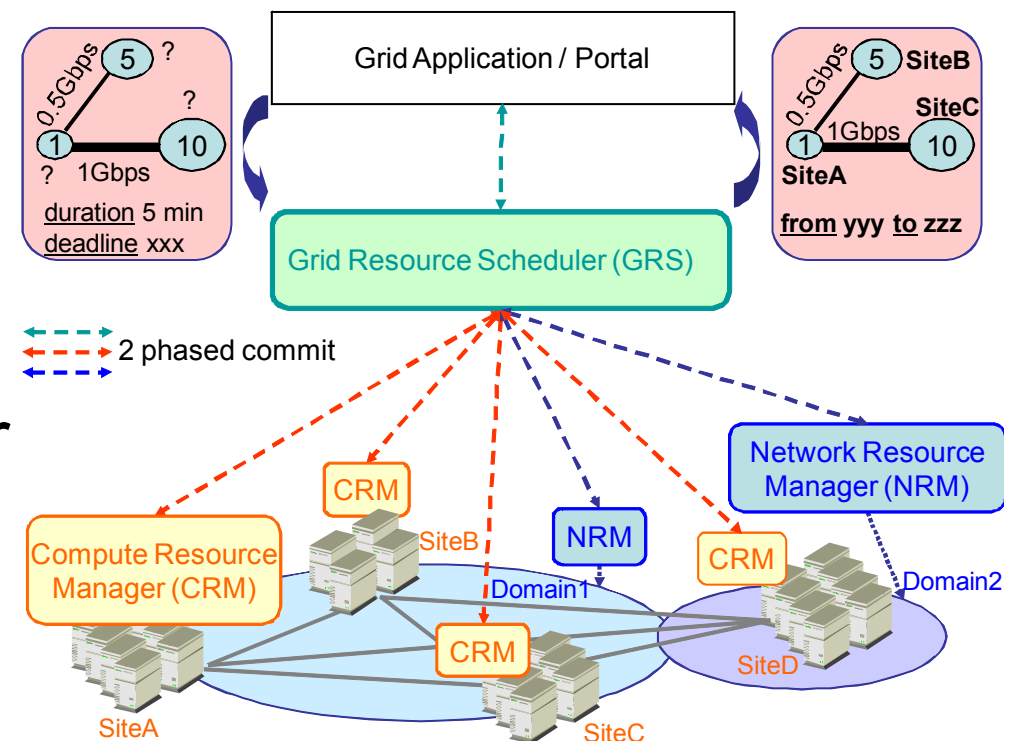
- 資源予約と連携

- GridARSを用いた資源管理システムと連携



GridARSの概要

- 事前予約インターフェースを提供するコアロケーションフレームワーク
- GridARS-WSRFとGridARS-Coschedulerを提供
- GridARS-WSRF
 - WSRFに基づく予約インターフェースを提供
 - 分散トランザクション
- GridARS-Coscheduler
 - ユーザの要求に基づき適切な資源を確保
 - GRSで利用

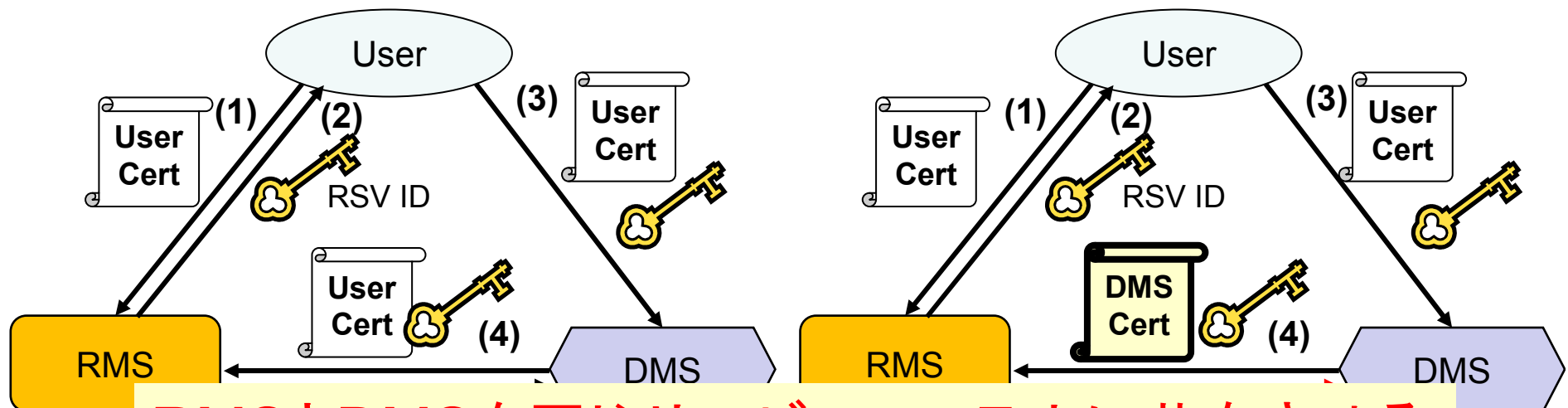


資源予約との連携と認証

- GridARSの資源予約手続きでは, **予約ID**を用いる
→DMSからの資源予約情報取得に**予約ID**を利用
- GSIにおける認証の問題
 - **商用サービスでは権限委譲しない可能性が高い**

権限委譲あり

権限委譲なし



RMSとDMSを同じサービスコンテナに共存させる

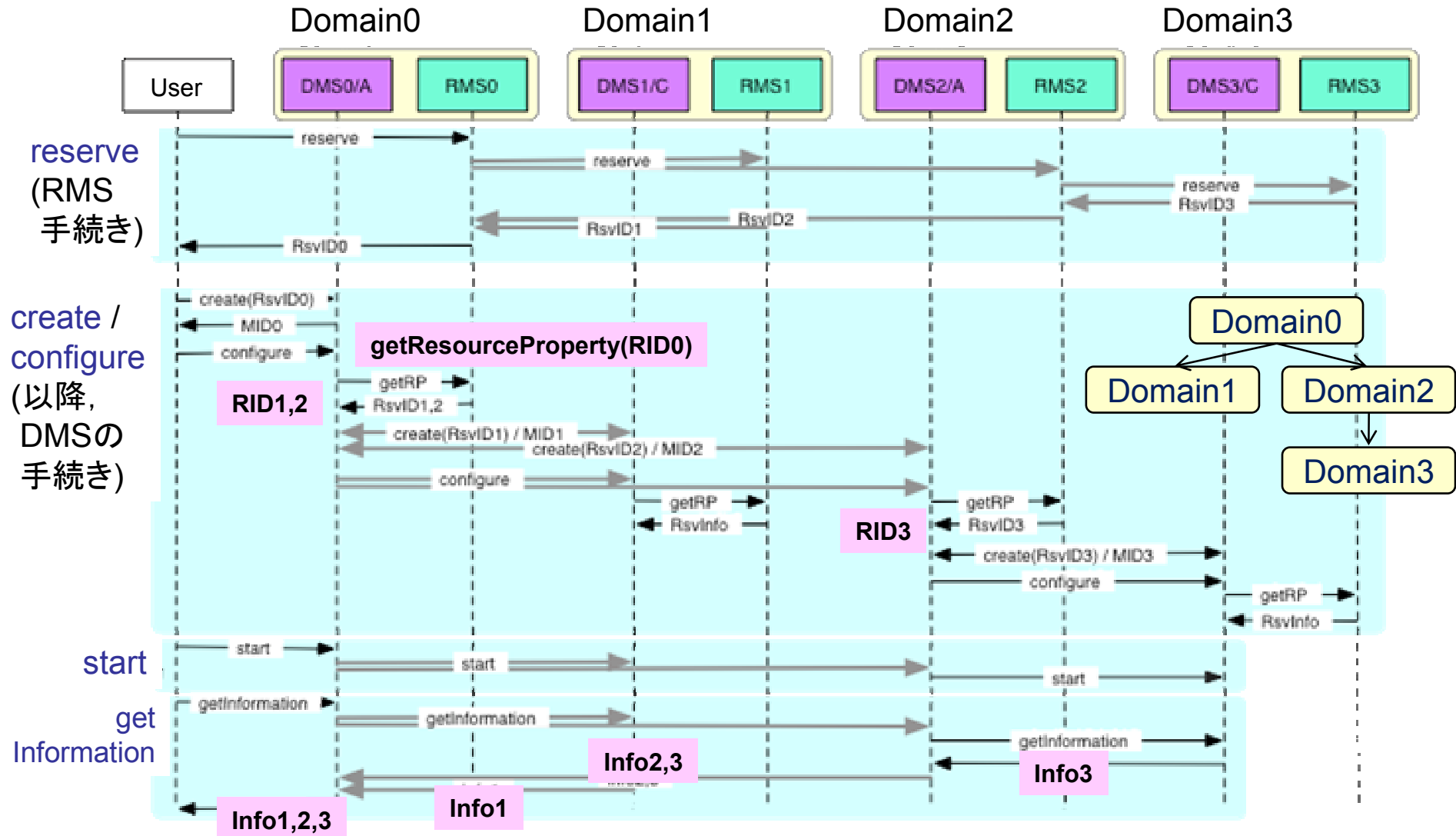
DMSインタフェース

オペレーション名	機能	入力／出力
create	モニタリング手続き開始	予約ID／モニタリングID
configure	収集する情報セットを指定	モニタリングID, 情報セット／ -
start	モニタリング情報収集を開始	モニタリングID (, 時刻)／-
stop	モニタリング情報収集を停止	モニタリングID (, 時刻)／-
getInformation	指定したモニタリング情報の取得	モニタリングID (, 時間帯)／ モニタリング情報

XACMLを用いた認可

- 先行研究のIA, ICのインタフェースを共通化
- モニタリングID: 個々のモニタリング要求に対してDMSが提供
- 情報セット: モニタリング対象情報のリスト
- WS-Notificationに基づくインタフェース (subscribe/unsubscribe)もオプションで提供

情報取得プロセス



資源モニタリング情報のデータ表現

- DMS間で授受される資源表現にGLUE, v. 2.0の拡張を用いる
 - OGFで標準化
 - 計算資源が主対象で, ネットワーク資源はない
 - 時系列情報に関する定義なし
- ネットワーク資源, 時系列情報について拡張
 - ネットワーク資源情報はG-lambdaプロジェクトのGNS-WSI (Grid Network Service - Web Services Interface)に基づく

計算資源例: ComputingActivity

```

<ComputingActivity>
  <ID>test_comp</ID><Name>test application</Name>
  <StartTime>2008-08-01T00:00:00Z</StartTime>
  <EndTime>2008-08-02T00:00:00Z</EndTime>
  <RequestedTotalWallTime>86400</RequestedTotalWallTime>
  <State>ACTIVATED</State>
  <RequestedSlots>2</RequestedSlots>
  <RequestedApplicationEnvironment>gridmpi</RequestedApplicationEnvironment>
  <Monitoring>
    <Timestamp>2008-08-01T01:00:00Z</Timestamp>
    <UsedTotalWallTime>3600</UsedTotalWallTime>
    <UsedTotalCPUTime>1800</UsedTotalCPUTime>
    <UsedMainMemory>1024</UsedMainMemory>
  </Monitoring>
  <Monitoring>...</Monitoring>
</ComputingActivity>
  
```

← 時系列情報

ネットワーク資源例: NetworkActivity

```

<NetworkActivity>
  <ID>test_net</ID>
  <StartTime>2008-08-01T00:00:00Z</StartTime>
  <EndTime>2008-08-02T00:00:00Z</EndTime>
  <RequestedTotalWallTime>86400</RequestedTotalWallTime>
  <State>ACTIVATED</State>
  <Path xmlns="http://www.glambda.net/schemas/gnswsi3/ndl-aist">
    <APoint><Name>AKB</Name></APoint>
    <ZPoint><Name>TKB</Name></ZPoint>
    <Bandwidth>1000</Bandwidth>
  </Path>
  <Monitoring stream="Down">
    <Timestamp>2008-08-01T01:00:00Z</Timestamp>
    <UsedTotalWallTime>3600</UsedTotalWallTime>
    <Transfer>65536</Transfer>
    <CurrentBandwidth>16</CurrentBandwidth><PeakBandwidth>256</PeakBandwidth>
    <Ping><Roundtrip>2.352</Roundtrip><TTL>64</TTL></Ping>
    <Jitter>0.1234</Jitter><Latency>3.456</Latency>
  </Monitoring>
</NetworkActivity>

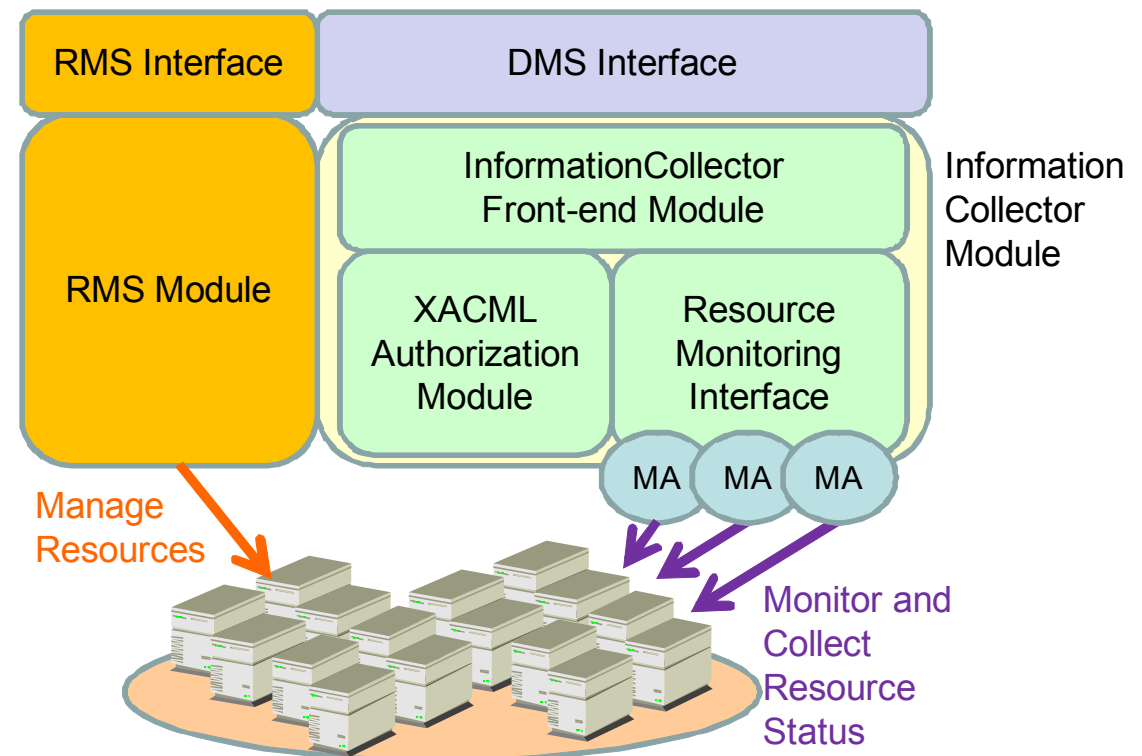
```

← ネットワーク資源定義

← 時系列情報

モニタリング情報の収集

- モニタリング情報収集にICの機能を利用
- モニタリングエージェント(MA)による情報収集
- MAが収集した情報を認可後、ユーザに提供



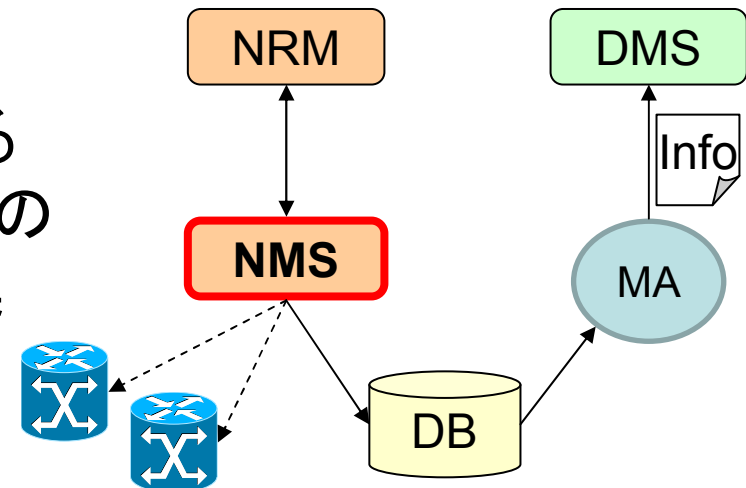
MAでの計算資源の情報収集

- 収集する情報
 - a. 予約自体に関する情報
予約ID, CPU数, メモリ量, 開始・終了時刻, ステータス
 - b. 予約資源利用に関する情報
予約単位に集計された使用CPU時間, メモリ量, ディスク量
 - c. アプリケーションの実行に関する情報
登録されたアプリケーションの実行ログ
- 収集方法
 - a. CRM(計算資源マネージャ)への問い合わせ
 - b. OSのプロセス情報I/Fやプロセスアカウンティングパッケージのログファイルを利用(Linuxではproc, psacct)
 - c. 事前に指定されたファイルを参照

MAでのネットワーク資源の情報収集

- 収集する情報
 - a. 予約自体に関する情報
予約ID, 拠点情報, ピークバンド幅, 開始・終了時刻, ステータス
 - b. 予約資源(パス)利用に関する情報
リンクステータス(Up/Down), パケット統計情報(入出力パケット数(転送/破棄/エラー), 転送量)

- 収集方法
 - a, bとも, NRMに一般に常駐するネットワーク管理システム(NMS)の収集情報を, 内部DB経由で収集
 - NMSはパスのセットアップ, ドメイン内の健全性監視を行う



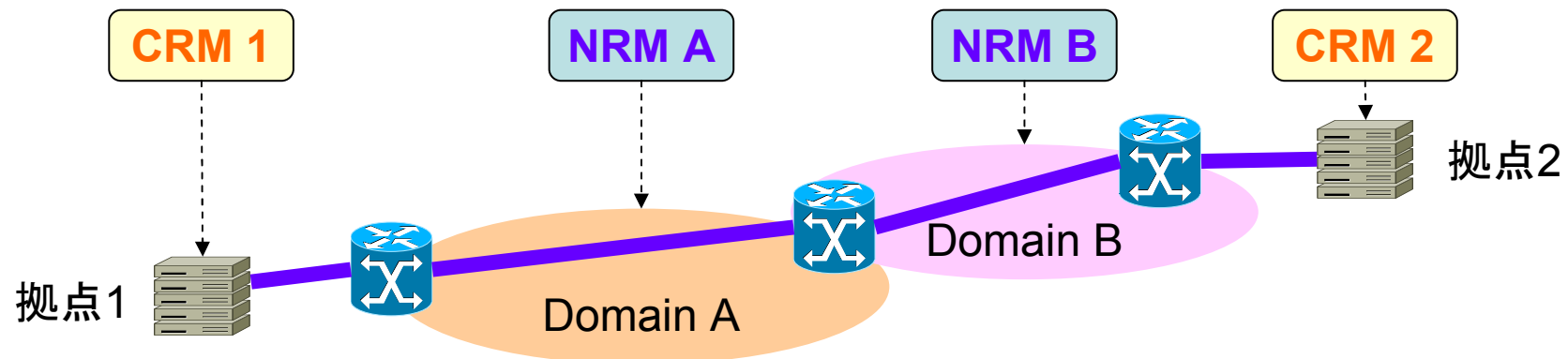
NMSでの収集方法

端点のネットワーク機器のインタフェースに対して行う

- SNMP での問い合わせ
 - SNMP (Simple Network Management Protocol)
 - MIB (Management Information Base)で定義されている標準的な項目をサポートしており, ほぼ全ての機器が対応
 - 仮想インタフェース(e.g., VLAN)への対応は機種依存
- SNMP非対応の場合, CLI (Command Line Interface)を利用

拠点間情報の収集

- 拠点間情報(拠点間スループット, 遅延とその揺らぎ)はネットワークの資源情報に属するが, NRMでは測定できない→CRMで行う
- CRMでのモニタリング
 - 確保した計算資源のCRM間で動的構成情報を共有
 - 他拠点のノードに対してping等で拠点間情報を取得



関連研究 (1/2)

- MonALISA [CHEP2004]
 - GangliaやMRTG等のツールで収集された情報をレポジトリに格納し, クライアントI/Fから提供
 - 情報開示に関する認可は行っていない
- Inca [SC2004]
 - TeraGridにおけるモニタリングシステム
 - エージェントによりユーザ権限で情報を収集
 - 情報は集中管理され, 管理者向けに情報提供

関連研究 (2/2)

- WMSMonitor [Grid2008]
 - EGEEのgLiteを対象としたモニタリングシステム
 - ワークロード, ジョブライフサイクルのモニタリング
 - 管理者, 開発者, VOユーザなど, 複数のユーザカテゴリをサポート
- AMon [Grid2008]
 - D-GridのHEPCGを対象としたモニタリングシステム
 - ステータス, 資源利用状況, アプリケーションの出力のモニタリング
- いずれのシステムもネットワーク資源予約との連携はない

まとめと今後の課題

- 資源予約と連携した階層型分散資源モニタリングシステムDMSの提案
 - 先行研究のXACMLを用いた情報サービスシステム [ComSys2007]を改良
 - 複雑なドメイン構成を考慮した階層型アーキテクチャ
 - ドメインごとに情報開示ポリシーを定義可能
 - 資源予約との連携: GridARSフレームワークを利用
 - 標準データ表現の拡張(GLUE, v. 2.0)を採用
 - 資源モニタリング機構を追加
- 今後は本研究の設計をもとに, 実装を進める

謝辞

本研究の一部は、情報通信研究機構(NICT)の委託研究「ダイナミックネットワーク技術の研究開発」により実施した。