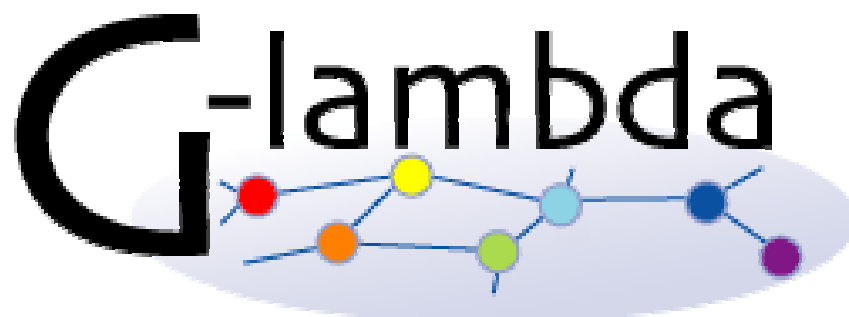


ミドルウェア連携による計算・ネットワーク 資源の日米間グリッドコアロケーション実験

竹房 あつ子¹, 林 通秋², 築島 幸男³, 岡本 修一⁴,
柳田 誠也^{1,5}, 宮本 崇弘², 平野 章³, 鮫島 康則^{4,3},
中田 秀基¹, 谷口 篤^{4,3}, 工藤 知宏¹

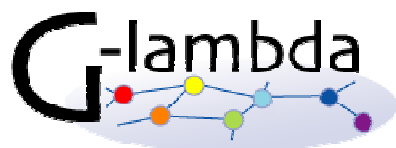
¹産総研グリッド研究センター, ²株式会社KDDI研究所,
³日本電信電話株式会社, ⁴情報通信研究機構, ⁵数理技研



<http://www.g-lambda.net/>

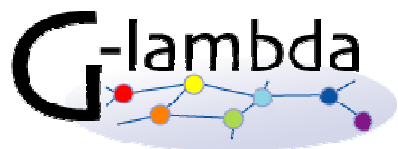
グリッドコアロケーションの課題

- 広域に分散した多様な資源を同時確保し, 1つのグリッド環境としてユーザに提供
 - 資源マネージャは標準的なインタフェースで事前予約による資源提供
- 計算資源では多数キューイングスケジューラが開発
 - プラグインスケジューラ等により事前予約機能を実現
- ネットワーク資源の動的な提供は実験段階
 - 電話や電子メールによる事前の交渉の後, オペレータによる手動設定
 - サービスインタフェースからの自動的な資源提供は実験段階
[Hokke06]
 - **複数管理ドメインに跨る拠点間の帯域確保**は困難
- 異なるプロジェクト間の計算・ネットワーク資源を**横断的かつ自動的に**確保する試みはない



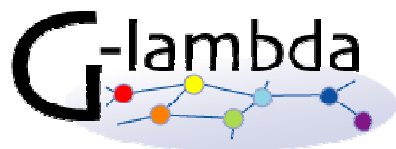
本研究の成果

- 複数ドメインに跨るネットワーク資源を確保することを考慮した**GNS-WSI2**を**G-lambda**プロジェクトで規定
 - ネットワーク資源を事前予約により確保するためのウェブサービスインタフェース**GNS-WSI**を規定[Hokke06]
 - GNS-WSI2 (GNS-WSI ver. 2)では, **WSRF**を採用
 - 分散トランザクションに対応するため, **2相コミット**の予約手続き
 - KDDI研, NTT, 産総研が参照実装
- G-lambdaと**Enlightened Computing**プロジェクト(以降, Enlightened)が共同で, **日米間に跨る計算・ネットワーク資源をミドルウェア連携により事前予約で確保する実証実験** [GLIF2006, SC06]
 - G-lambda, Enlightenedのコアロケーションシステム, 資源マネージャが連携
 - アーキテクチャ, インタフェースの差異を隠蔽する**ラップモジュール**
 - 分散する予約資源の監視のための**予約資源モニタサービス**



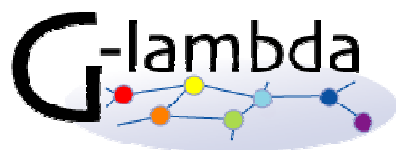
G-lambda , Enlightened日米間資源予約実証実験

- 日米間の**帯域**と日米に分散する**計算機群**の**同時事前予約**
- ネットワーク資源においては,異なる資源マネージャの管理するドメイン間の事前予約での連携は**世界で初めて**
 - 各資源マネージャは異なるI/Fを持ち,それぞれが開発
- 2国間のグリッドコンピューティングリサーチテストベッドのネットワーク・計算資源間の**自動相互運用**



発表の概要

- G-lambdaプロジェクトの概要
- GNS-WSI2 (GNS-WSI ver. 2)の概要
- 日米間資源予約実証実験
 - G-lambda, Enlightenedのミドルウェア構成
 - 複数ドメインでの異種ミドルウェア連携
 - ラッパモジュール開発
 - 予約資源モニタサービス
 - 日米間資源予約実証実験の概要
- まとめ



G-lambdaプロジェクトとGNS-WSI

- 産業技術総合研究所, KDDI研究所, NTT, 情報通信研究機構の共同研究
- 2004年12月より開始
- ネットワークオペレータの資源マネージャがグリッドミドルウェアなどのアプリケーションに提供する標準的なウェブサービスインターフェースGNS-WSI (Grid Network Service - Web Services Interface)の確立が目的



*National Institute of
Advanced Industrial Science
and Technology*

AIST

NICT

National Institute of
Information and
Communications
Technology



JGNII



KDDI
KDDI R&D LABS



<http://www.g-lambda.net/>

G-lambdaプロジェクトメンバ



*National Institute of
Advanced Industrial Science
and Technology*
AIST

- Tomohiro Kudoh
- Hidemoto Nakada
- Atsuko Takefusa
- Yoshio Tanaka
- Fumihiro Okazaki
- Satoshi Sekiguchi
- Hiroshi Takemiya
- Motohiko Matsuda
- Seiya Yanagita
- Katsuhiko Okubo



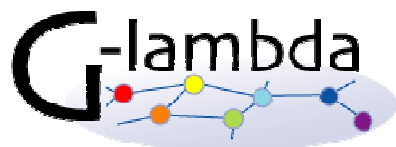
- Shuichi Okamoto
- Tomohiro Otani
- Yasunori Sameshima
- Atsushi Taniguchi



- Masatoshi Suzuki
- Hideaki Tanaka
- Tomohiro Otani
- Munefumi Tsurusawa
- Michiaki Hayashi
- Takahiro Miyamoto

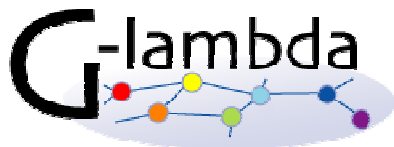
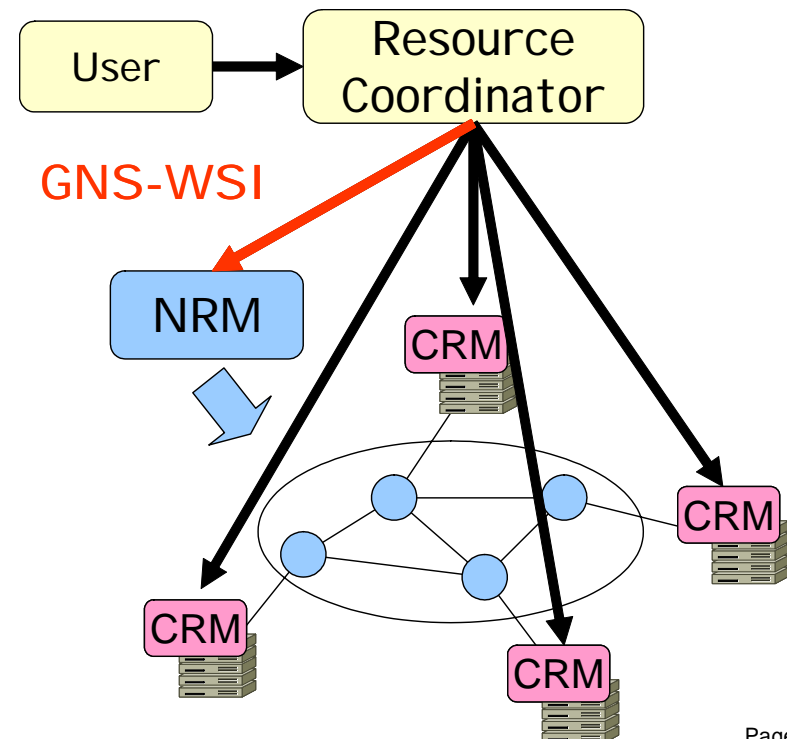


- Akira Hirano
- Yasunori Sameshima
- Wataru Imajuku
- Takuya Ohara
- Yukio Tsukishima
- Atsushi Taniguchi
- Masahiko Jinno
- Yoshihiro Takigawa



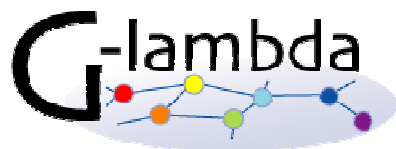
GNS-WSI

- Grid Network Service - Web Services Interface
- グリッドアプリケーションやミドルウェアから、帯域の事前予約を可能にするウェブサービスインタフェース
- ポーリングベース、ノンブロッキングオペレーション
 - 利用可能な資源情報の問い合わせ
 - エンドポイント間のパスの事前予約
 - 予約の修正
(i.e. 予約時刻, 予約時間)
 - 予約資源状況の問い合わせ
 - 予約のキャンセル

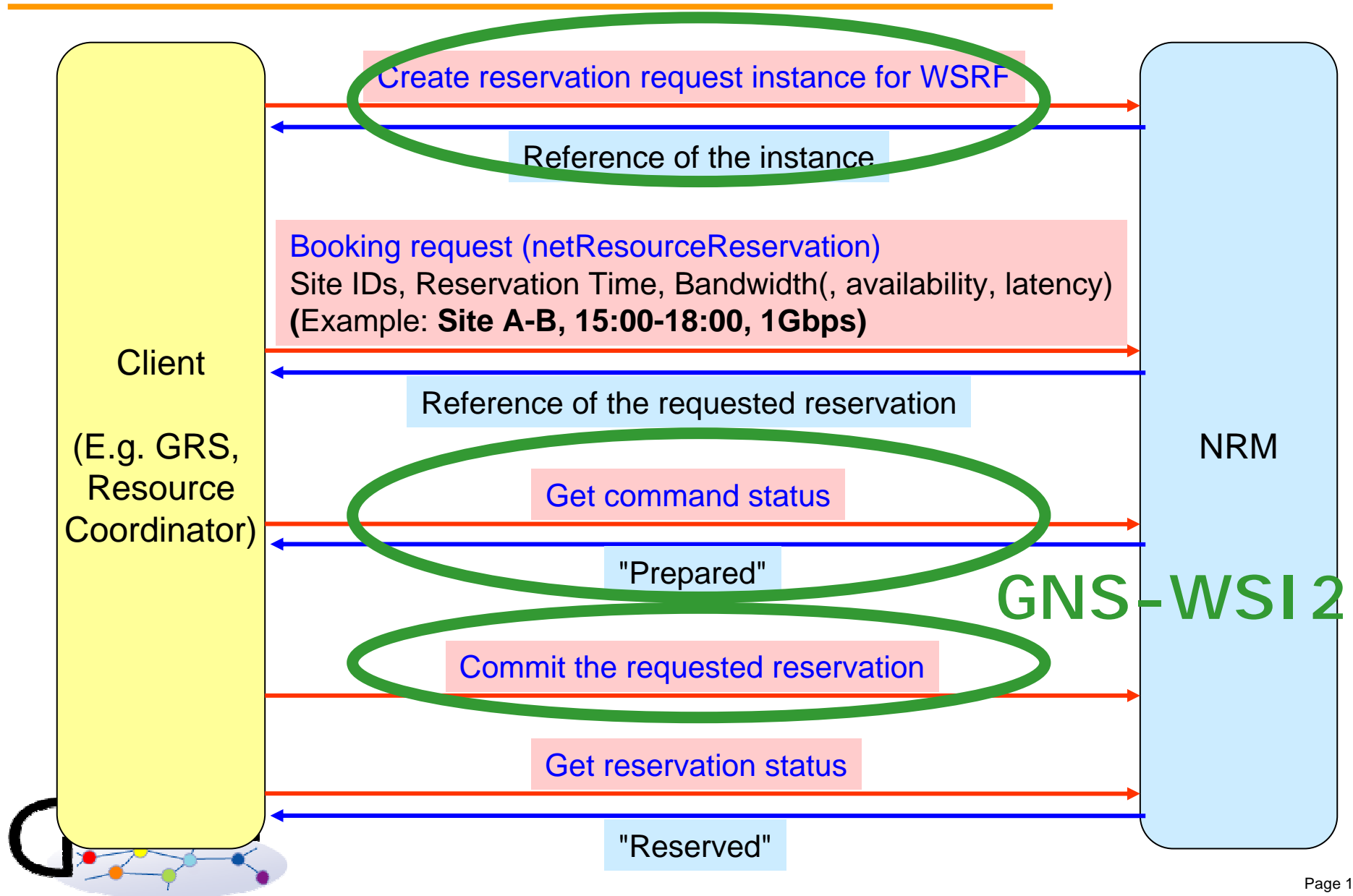


GNS-WSI2 (GNS-WSI ver. 2)

- GNS-WSI ver. 1 (2005年度)
 - 通常のウェブサービスインタフェース
 - 標準的なステートフルインタフェースなし
 - 各予約要求は"パス予約ID"パラメータにより識別
 - 1相コミット
- GNS-WSI2 (GNS-WSI ver. 2) (2006年度)
 - **WSRF** (Web Services Resource Framework)に基づく
 - ステートフルサービスのための標準WSインタフェース
 - エンドポイントリファレンス(EPR)で各予約要求を識別可能
 - **2相コミットプロトコル**をサポート
 - 上位のグリッド資源コアロケータによる分散トランザクション処理が可能
 - **異なるドメインを跨る拠点間の同時予約処理が可能**

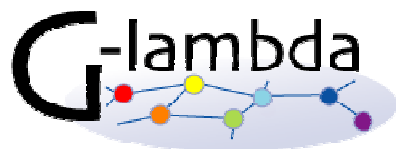


An example XML exchanged through GNS-WSI2



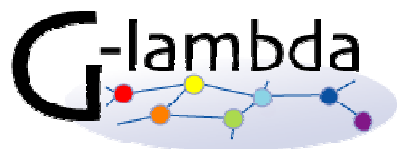
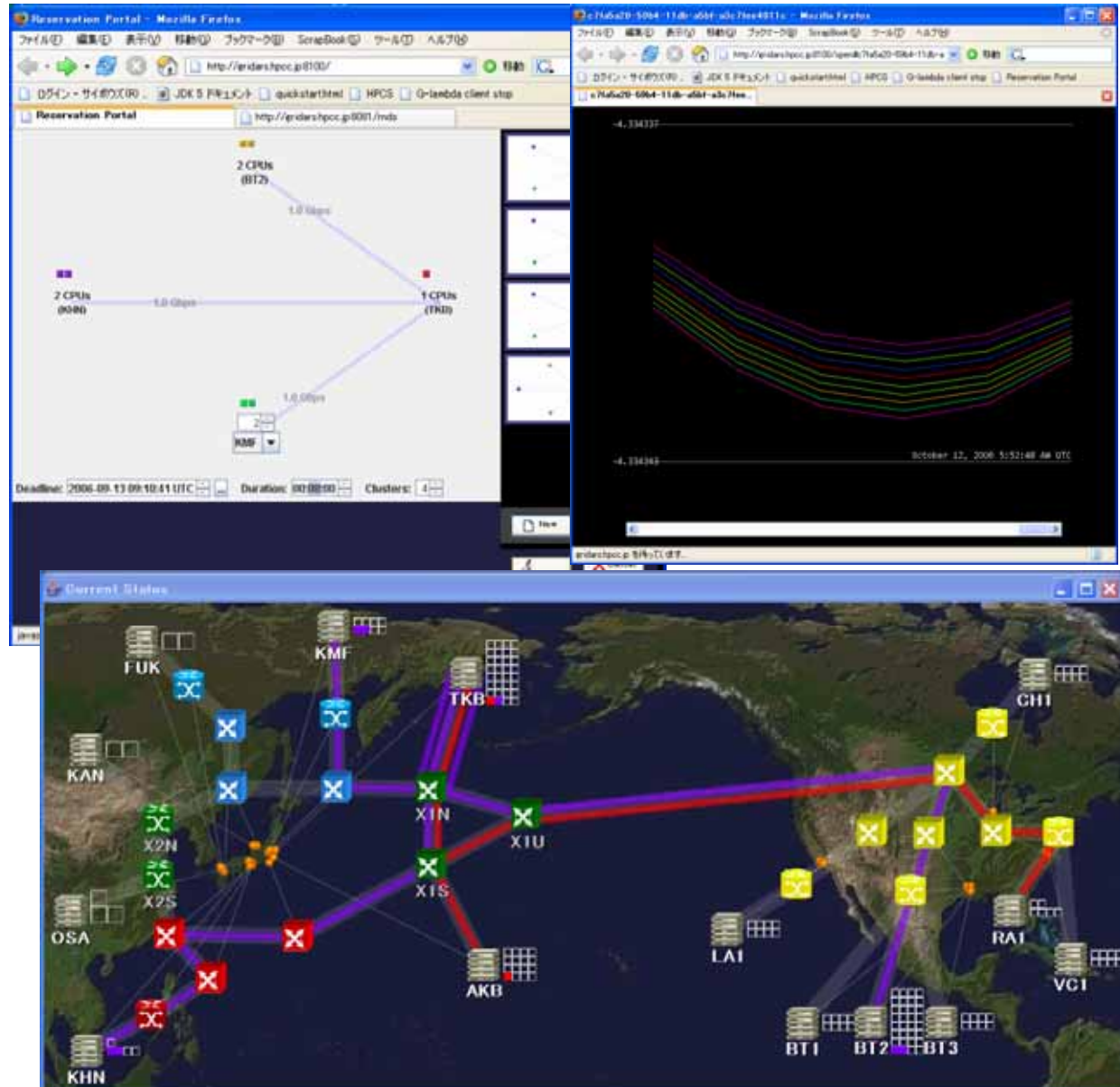
GNS-WSI2の参照実装

- KDDI研, NTT, 産総研が参照実装
- 産総研では, GNS-WSI2のインタフェースモジュール **GridARS-WSRF [HPCS2007]**を開発
 - ネットワーク資源提供者は煩雑なWSRFのコードを開発することなく, GNS-WSI2インタフェースで資源提供が可能
 - Globus Toolkit 4のJava WS Coreを用いて実装
 - 予約手続き情報の永続化
 - GSIによる認証, 認可も可能



日米間資源予約実証実験[GLIF2006, SC06]

- G-lambda (GL)と Enlightened (EL) の共同実験
- 日米間に跨る資源をそれぞれのコアロケーションシステムから事前予約で確保
- 予約資源上でそれぞれアプリケーションプログラムを実行



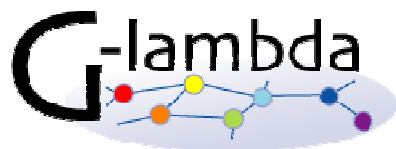
Enlightened Computingプロジェクト



- 2005年に設立され、NSFの資金を得ている
- アプリケーションからの動的な事前予約要求に対応した、光コントロールプレーンとグリッドミドルウェアの統合を目的とした学際的な共同研究
- MCNC, LSU, NCSU, RENCi, Cisco, AT&T, Calientが参加
- ネットワークに接続された地理的に分散して配置された高性能計算装置や科学的な観測装置を、動的、適応的かつ最適に連係させて利用する手法の確立を目指す



NC STATE UNIVERSITY



Enlightened Computingプロジェクトメンバ



- Yufeng Xin
- Steve Thorpe
- Bonnie Hurst
- Joel Dunn
- Gigi Karmous-Edwards
- Mark Johnson
- John Moore
- Carla Hunt
- Lina Battestilli
- Andrew Mabe



- Ed Seidel
- Gabriele Allen
- Seung Jong Park
- Jon Maclaren
- Andrei Hutanu
- Lonnie Leger
- Dan Katz



- Joe Mambretti
- Alan Verlo



- Savera Tanwir
- Harry Perros
- Mladen Vouk



- Olivier Jerphagnon
- John Bowers



- Steven Hunter



- Javad Boroumand
- Russ Gyurek
- Wayne Clark
- Kevin McGrattan
- Peter Tompsu



- Rick Schlichting
- John Strand
- Matti Hiltunen



- Yang Xia
- Xun Su



- Dan Reed
- Alan Blatecky
- Chris Heermann
- Ilia Baldin



複数ドメインに跨る帯域の確保方法

(1) 下位ネットワーク制御層での連携

ユーザは複数ドメインを意識する必要なし

制御層のプロトコルによるトラフィックエンジニアリングが可能

- × 制御層では事前予約をサポートしていない
- × 競合事業者間では困難

(2) ネットワーク資源マネージャ間での連携

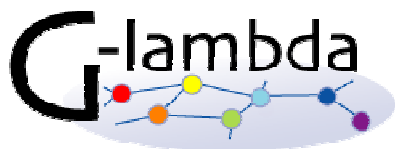
複数ドメインの隠蔽可能

- × 要求をうけたNRMが自ドメインに都合のよい経路選択の可能性

(3) 上位スケジューラ(ユーザ)での連携

- × ユーザに詳細なネットワーク構成に関する知識が必要
- ユーザ側でドメイン間接続の制御が可能
- 下位層での連携不要

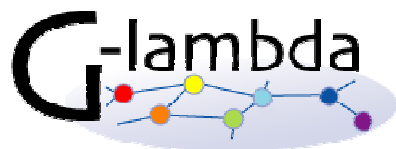
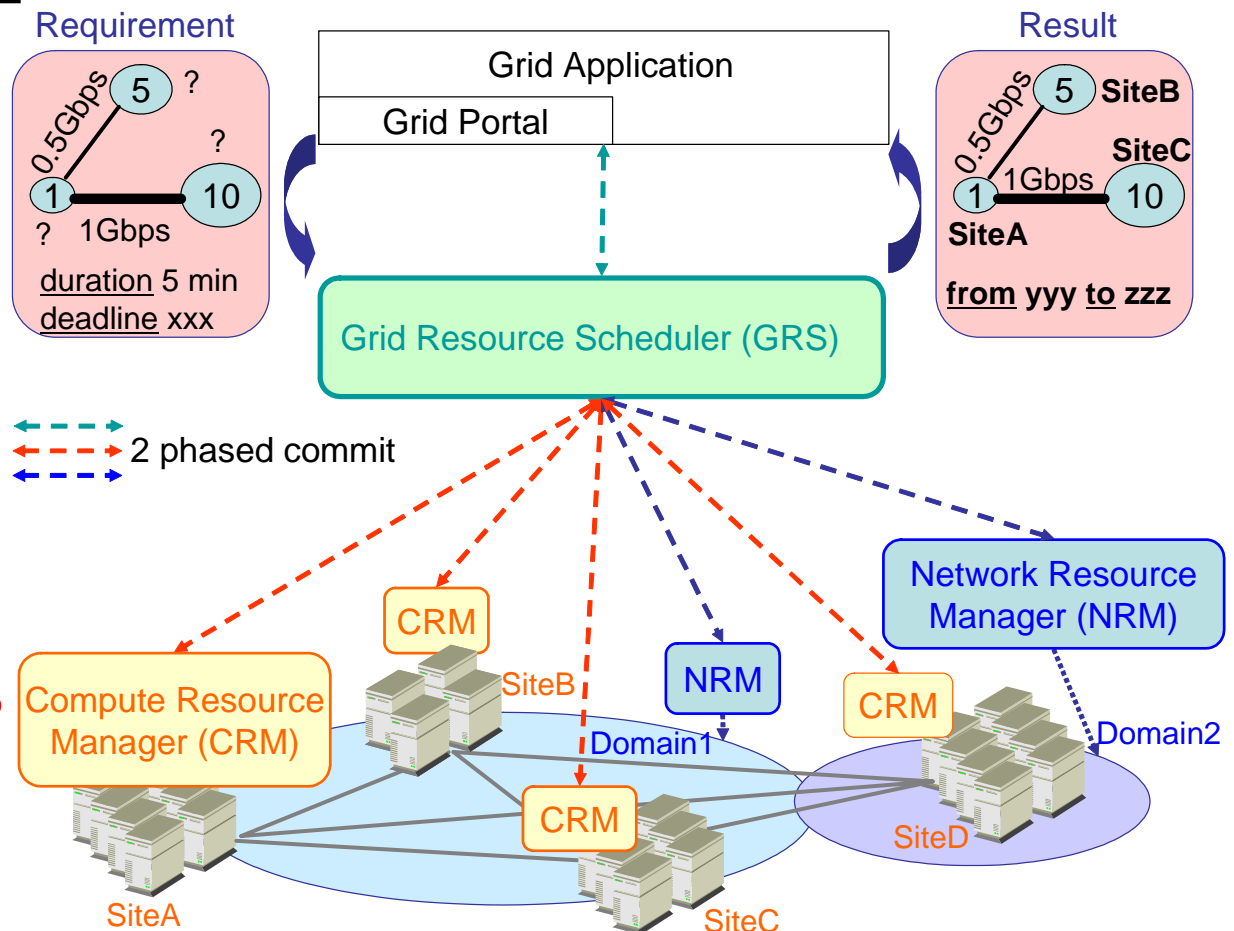
このモデルにより実証実験を実施



G-lambda(GL)ミドルウェア構成 (1/2)

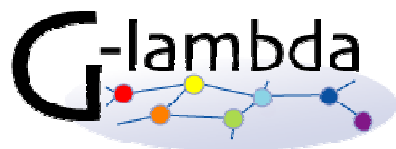
- グローバル資源スケジューラ(GRS), ネットワーク資源マネージャ(NRM), 計算資源マネージャ(CRM)で構成
- 2相コミット, WSRF I/F

- GRS-NRM間:
GNS-WSI2
- GRS-CRM間:
拡張JSDL
- ユーザ-GRS間:
上記を集約したもの
- 階層的な2相コミット
により, GRSも
資源マネージャの
1つになることができる



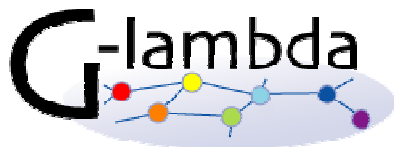
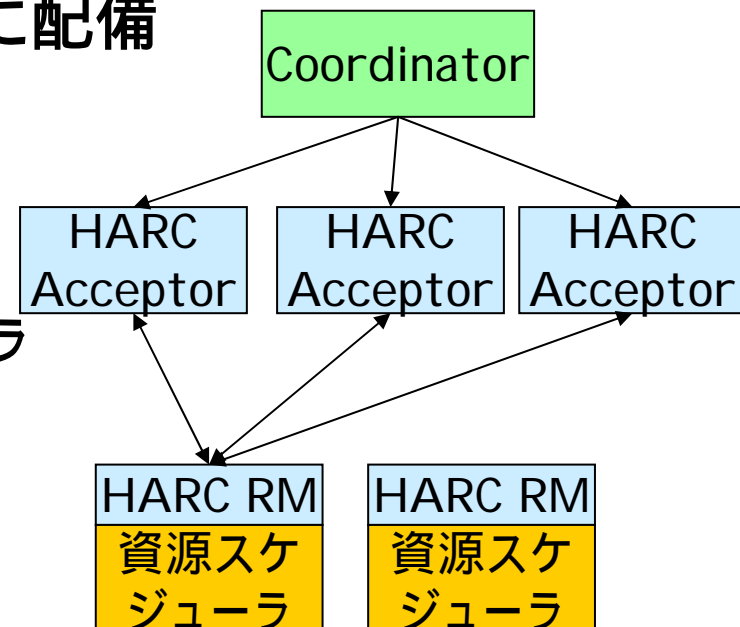
G-lambda(GL)ミドルウェア構成 (2/2)

- GRSはユーザの要求に従い, 複数NRM, CRMと連携して資源を同時確保する
 - クラスタ数, 各クラスタのノード数, 帯域, 予約時間
- 各NRMは管理ドメイン内の拠点間のパスを提供
 - 詳細なパス構成を隠蔽
 - パススケジューリング, 管理
- 各CRMは計算資源を事前予約で提供
 - 既存キューイングスケジューラを利用
- 実験で用いたミドルウェア
 - GRS: **GridARS** [SAC SIS2006, HPCS2007]のGRS
 - NRM: **KDDI研およびNTTがそれぞれ開発したNRM**
 - CRM: **GridARS-WSRF**, **PluS**(産総研), GridEngine
 - PluSは既存キューイングスケジューラに事前予約機能を付加



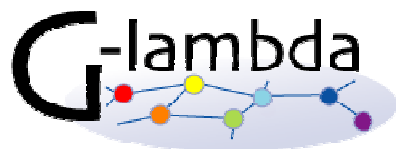
Enlightend(EL)ミドルウェア構成

- **HARC** (Highly-Available Robust Co-scheduler) [LSU, Jon Maclaren] コアロケーションシステムで, 分散する資源を確保
- Acceptorと呼ばれるコアロケータと資源マネージャ間で Paxosコミットプロトコルに基づく手続きを行う
 - 複数のAcceptorを配備し, Acceptor側の耐故障性を高めている
- 各資源スケジューラはHARCの資源マネージャ(HARC RM)の下に配備
- 実験で用いたミドルウェア
 - コアロケーションシステム: HARC
 - NRM: HARC RM, Enlightenedの開発するネットワーク資源スケジューラ
 - CRM: HARC RM, MAUI, TORQUE



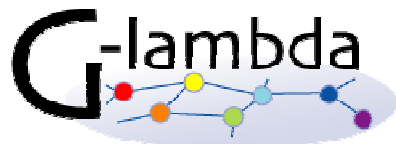
GLミドルウェアとELミドルウェアの主な違い

	G-lambda (GridARS,NRM,CRM)	Enlightened (HARC)
通信方式	WSRF (SOAP)	REST (XML over HTTP)
インタ フェース	資源ごとに異なる スタブライブラリを利用 (WSDLが異なる)	異なる資源の要求も同 じI/F (XMLで記述した資 源要求を授受)
スケジュー リング 機能	GRSはあり (上位スケジューラで行 うことも可)	HARC Acceptorは メッセージの受け渡しの み (上位スケジューラが 行う)



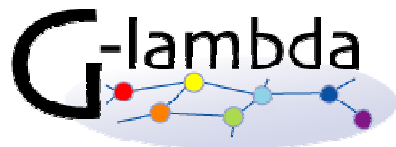
予約資源モニタサービス

- 複数のミドルウェアを介して分散資源の予約と、資源の利用
 - 計算資源はキューイングスケジューラの実績から安定稼動を期待
 - ネットワークの帯域提供は実験段階で不安定
- 分散資源の予約情報を可視化する**予約資源モニタサービス**を開発
 - 制御パケットの監視[iGrid2005]では不可能
 - **複数コアロケータ, NRM, CRMから予約資源情報収集**
- 予約資源モニタサービスの構成
 - ビューア
 - アグリゲータ
 - 予約資源情報の収集して提供
 - OOWeb(Javaの簡易ウェブサーバ)を用いて実装
 - HTTPで通信

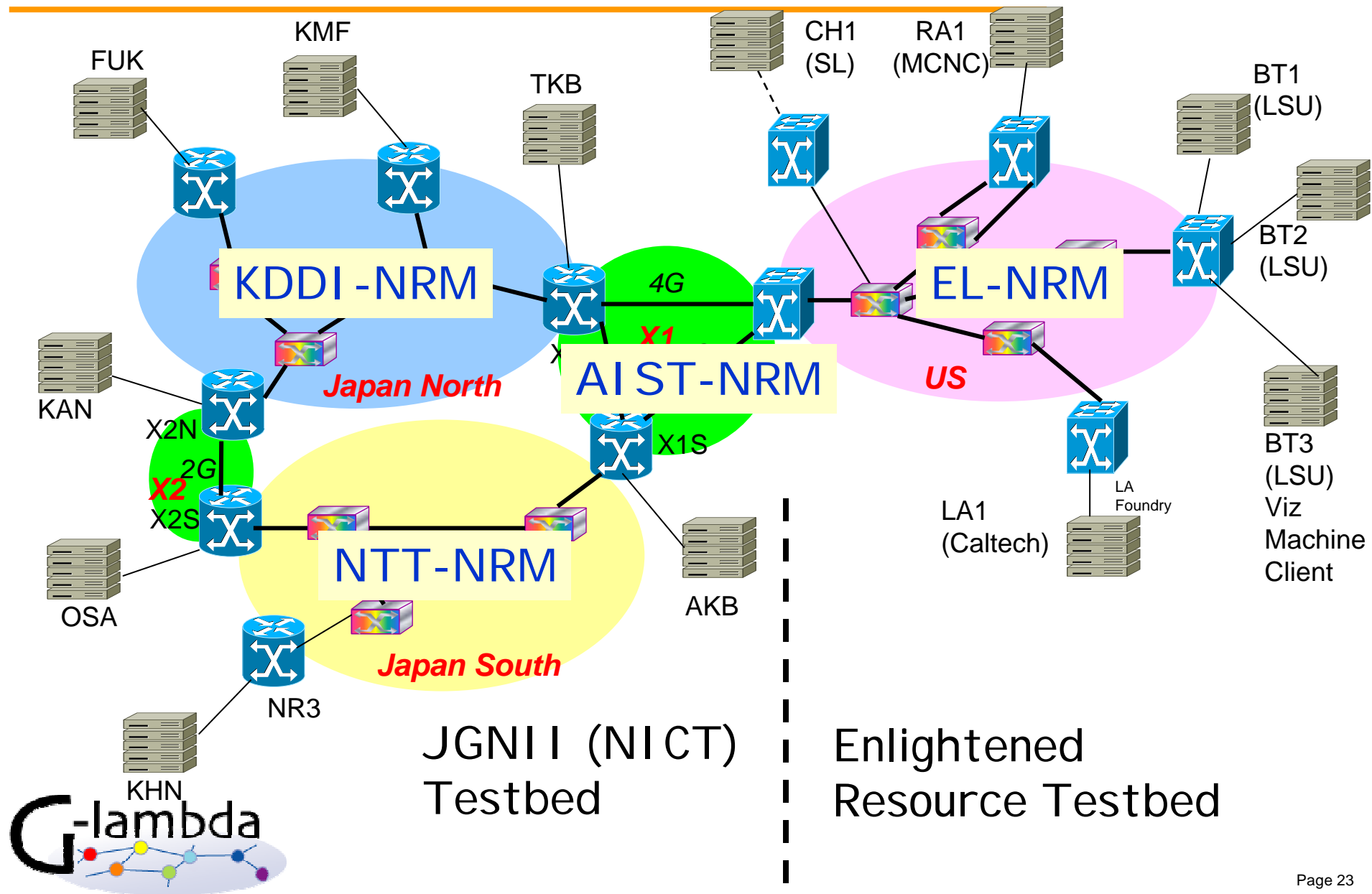


予約資源モニタサービスの収集情報

- コアロケータからユーザの要求と資源予約の**マッピング情報**, 各資源マネージャから**予約資源の状態**(予約済み / 利用可能 / 予約時間終了 / エラー)を収集
- 収集情報(XML)
 - コアロケータ
コアロケータ名(GL/EL), CRMサイト名, NRM名,
他のコアロケータ名, 各RM/コアロケータでの資源予約ID
 - NRM
開始・終了時刻, 端点情報, 帯域, 予約資源状況
 - CRM
開始・終了時刻, 端点情報, CPU数, 予約資源状況

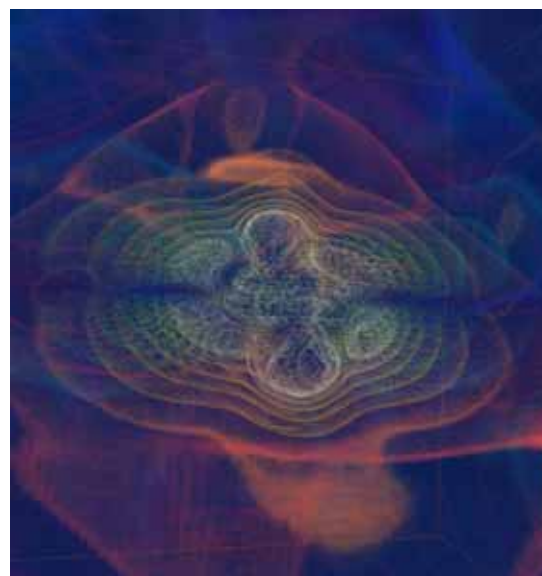


実証実験環境

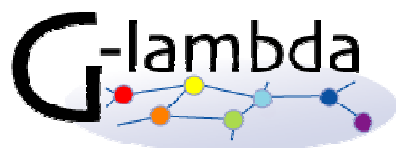


日米間資源予約実証実験の概要

1. G-lambda側から予約・予約状況を表示
2. Enlightened側から予約・予約状況を表示
3. 予約時刻になるとアプリケーションを実行
4. MonALISA(EL)、予約資源モニターサービスビューア(GL)上に
現在確保されている
帯域, 計算機を表示
5. 実行結果が表示
 - Enlightened: ブラックホールの
ビジュアライゼーション
 - G-lambda: 並列科学技術計算
(GridMPIで開発された
量子力学/分子動力学(QM/MD)
連成シミュレーション)

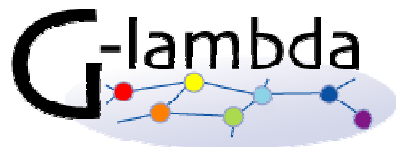


(Enlightened提供)



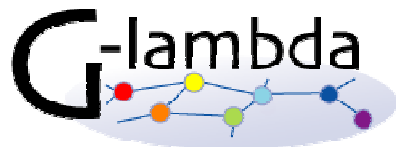
議論

- **標準資源予約インタフェースの必要性**
 - 実証実験ではラッパモジュールの開発により連携
 - 他方のシステムの内部構造を理解しなければいけない
 - **N×Nのラッパ開発は非現実的**
- **より詳細なモニタ情報の提供**
 - 資源予約モニタサービスは直感的でデバッグに有効, ただし, 予約資源が利用可能かどうかの情報のみ
 - 利用可能状態でも, アプリケーションが実行されないこともあった
 - 設定上の問題, 高負荷により実行ジョブが開始しない, 一部サイトでMPIジョブ起動が失敗 → **手作業でのデバッグが必要**
- **帯域提供に関する課題**
 - オンデマンド帯域提供のため, 各計算ノードは制御用・データ通信用の2つのIPアドレスを持つ
 - データ通信用は**予約時しかコネクティビティがない**ため, 並列アプリケーションのコンフィグレーションが複雑に(すべてのパスを予約するわけではない)
 - IPアドレスに対して帯域を提供する → **ユーザ(要求)毎に帯域提供できない**



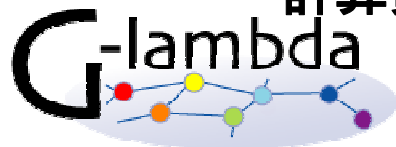
関連研究

- ネットワーク資源の事前予約を考慮したコアロケーションはあまり行われていない
- VIOLA (Phosphorusプロジェクト, EU)
 - 本研究同様, 計算・ネットワーク資源の事前予約ベースの提供
 - WS-Agreement/Negotiation [OGF]に基づくメタスケジューラの開発を進める
 - UNICOREベースグリッドシステム
- グリッドでのネットワーク資源も考慮したインターオペラビリティ実験は本研究のみ
- G-lambda, Enlightened, Phosphorusの連携を計画



まとめ

- 複数ドメインに跨るネットワーク資源を確保することを考慮した**GNS-WSI2**をG-lambdaプロジェクトで規定
 - WSRFを採用
 - 分散トランザクションに対応するため, 2相コミットの予約手続き
 - KDDI研, NTT, 産総研が参照実装
- G-lambdaとEnlightenedが共同で, **日米間に跨る計算・ネットワーク資源をミドルウェア連携により事前予約で確保する実証実験**
 - G-lambda, Enlightenedのコアロケーションシステム, 資源マネージャが連携
 - ラッパモジュールの開発
 - 資源状況の監視のための予約資源モニタサービス
 - 2国間のグリッドコンピューティングリサーチテストベッドのネットワーク・計算資源間の**自動相互運用**を実証



謝辞

- 実証実験にあたり、テストベッド構築、ミドルウェア開発、アプリケーション実行に携わった、Enlightened ComputingプロジェクトおよびG-lambdaプロジェクトの皆様に深く感謝いたします。

