# NAREGI-6 to NAREGI-10

大学・研究機関をつないだ現実の運用を考えた実証評価

Manabu Higashida

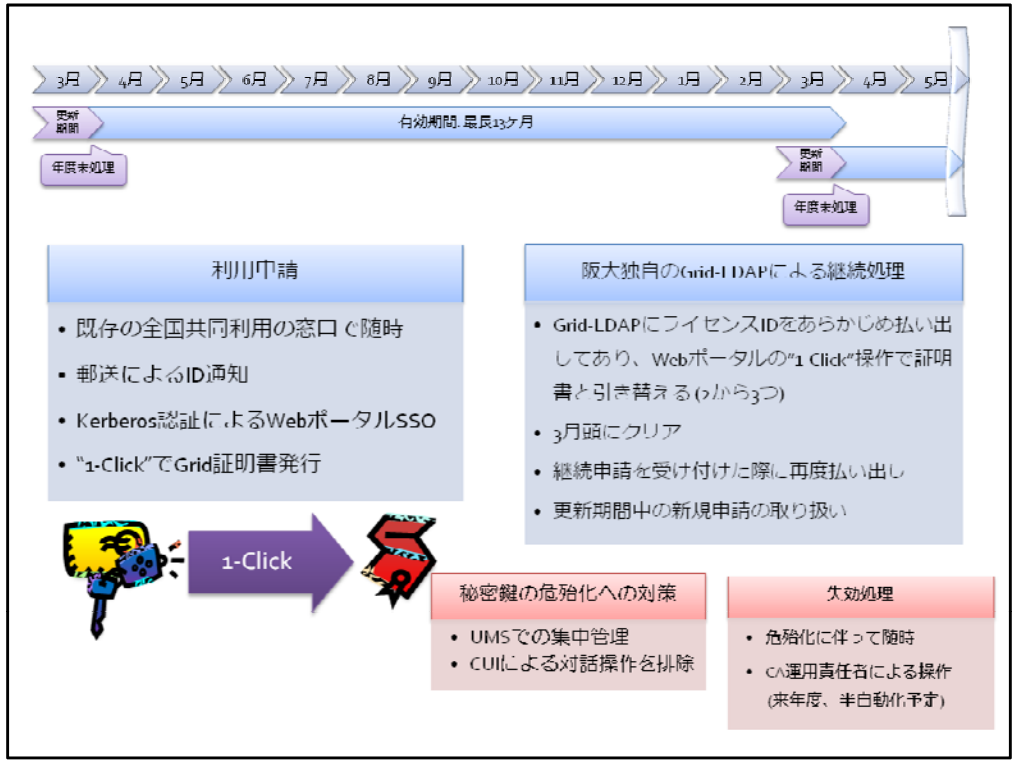
manabu@cmc.osaka-u.ac.jp

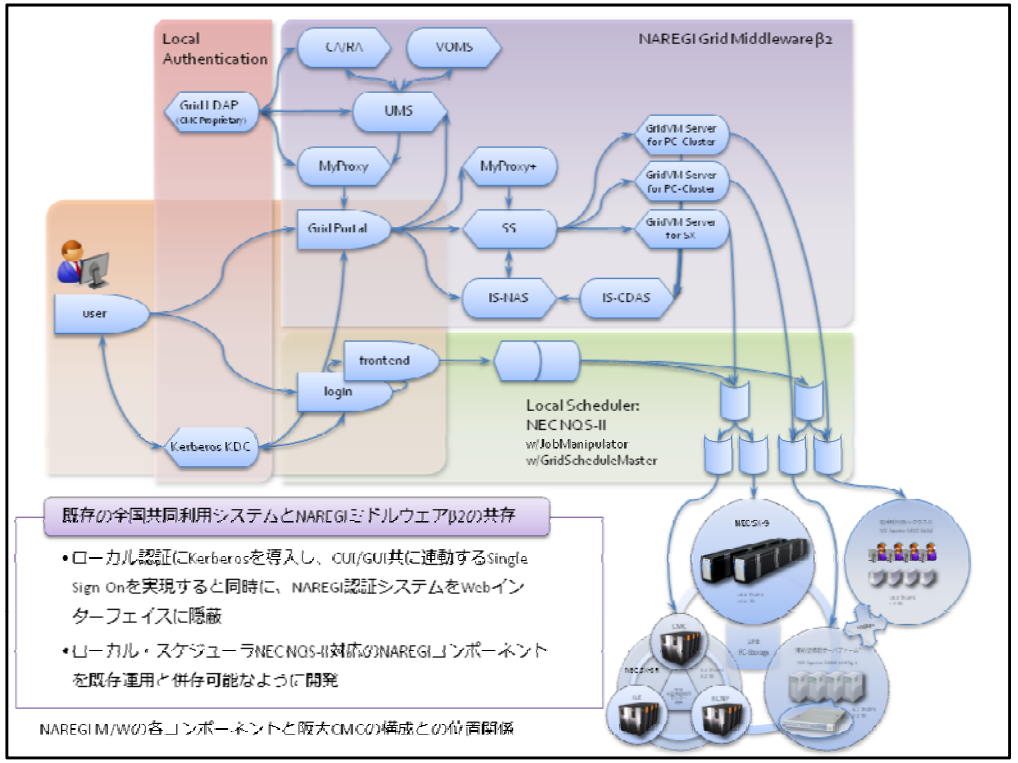
2008/09/26

# 阪大CMCのアプローチ – その1

## 「NAREGI連携なんて無理」と考えていた頃･･･阪大だけでも

- 第1段階
  - すべての登録ユーザにグリッド証明書を
    - “1-click” によるグリッド証明書発行
  - すべての計算機資源をグリッドに提供
    - ローカルスケジューラのパイプキューを閉塞・開放することで提供資源を適宜制御



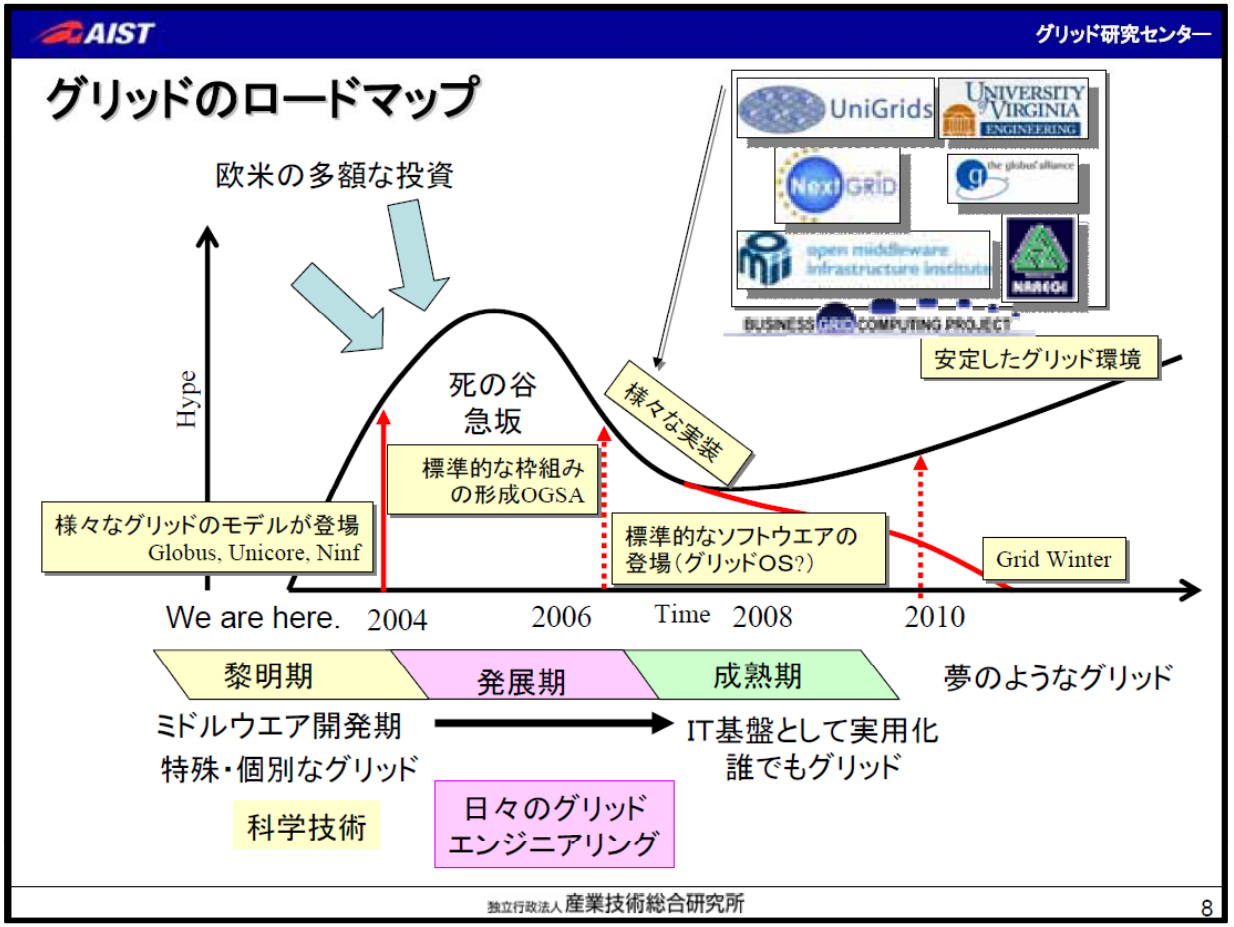



# NAREGI連携 (2007年度)

- 動機
  - 「SINET3のオープニングセレモニーでデモするから」 by 下條 真司 (2007/04くらいに)
- スモールスタート: SC07出展対応
  - 2拠点 (-2007/08)
    - 大阪大学、東京工業大学
  - +2拠点 (2007/08-)
    - 九州大学、国立情報学研究所
- "NAREGI 100T Project" : NAREGIプロジェクト最終年度の課題
  - +1拠点 (2008/01-)
    - 分子科学研究所
  - +1拠点 (2008/03-)
    - 名古屋大学

# NAREGIの「死の谷」

- 数多の国産プロジェクトが越えられなかった壁
  - 金の切れ目が縁の切れ目
- “グリッドOS”と呼べる標準的なソフトウェアは登場せず・・・
  - 米国: Globus Toolkit
  - 欧州: egee gLite
  - 日本: NAREGI
- ソフトウェアなのか?
  - オペレーション
  - コーディネーション




関口 智嗣, “「グリッドの旬」の味わい–Grid World 2006総括–”, Grid World 2006.

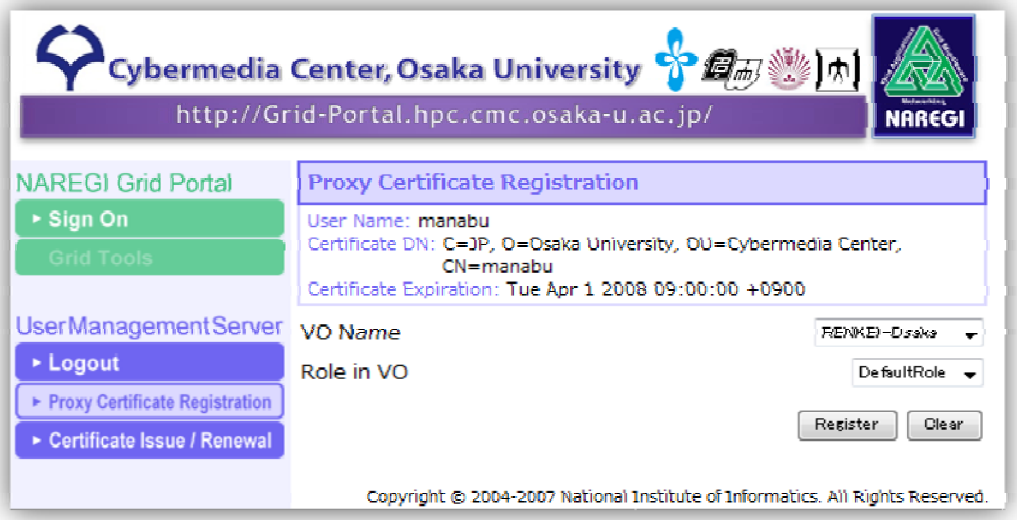
# NAREGI連携 (2008年度)

- 様々な経緯から急展開し、技術的には後退した「ミニマル構成」での7基盤センター連携へ仕切り直し
  - +4-1拠点
    - 北海道大学、東北大学、東京大学、京都大学
    - ~~東京工業大学~~
  - +2拠点
    - 東京工業大学、筑波大学

# NAREGI-10 (仮称)

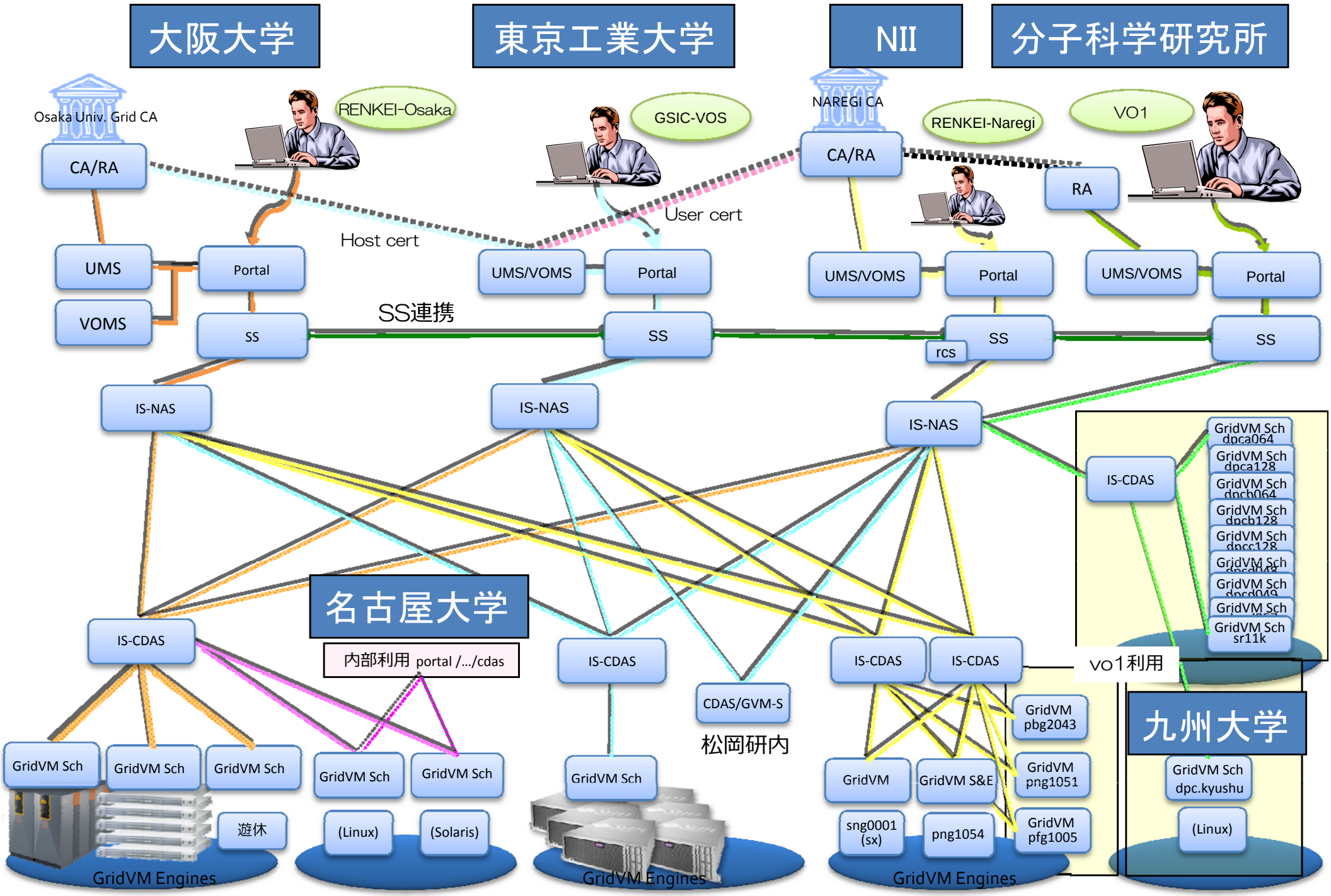
# “NAREGI 100T Project” (仮称) の課題: SC07の反省

- とにかく一度はきちんと「仕様どおり」に動かしてみたい
  - GridVMのことはかなりよくわかった
    - ローカル・スケジューラのことを熟知しているから
  - IS (Information Service) も何故かよくわかった
    - OGSA-DAI≒SQLだから
  - SS (Super Scheduler) のダークサイドぶりが際だつ
    - まるでちゃんと動いている気がしない･･･
      - “Missing”　: 落ちるとジョブが消失、ただしローカル・スケジューラにはジョブが残ったまま･･･
      - “Exception” : 何が原因で落ちたか一切不明･･･
    - このままでは何がどう「仕様どおりでないか」すらフィードバックできない･･･
- 再挑戦するなら「連携実証」の名に相応しい連携をやってみたい
  - ポリコム分配器が許す限り多くの拠点で連携したい
    - 阪大、東工大、九大、国情研、分子研、名大
  - ミドルウェアの連携機能を実証したい
    - IS 連携
    - SS連携 (V1の機能を先出し導入)



# "NAREGI 100T Project"

## Phase-2: 3/27時点のノード構成



# Fact Sheet 1: 管理ノード構成

| | Phase-1 / Phase-2 | SINET3 接続 | グリッド認証局 | ポータル | SS | IS | |
|---|---|---|---|---|---|---|---|
| | | | | | | NAS | CDAS |
| 大阪大学 | 2007年8月 | 10Gbps | ○ | ○ | ○ | ○ | ○ |
| 東京工業大学 | | 4Gbps | ー | ○ | ○ | ○ | ○ |
| 九州大学 | | ー | ー | ー | ー | ー | ー |
| NII/NAREGI | | 1Gbps | ○ | ○ | ◎※1 | ○ | ○ |
| 分子科学研究所 | 2008年1月 | 1Gbps | ー | ○ | ○ | ー | ○ |
| 名古屋大学 | 2008年3月 | 1Gbps | ー | △※2 | △※2 | △※2 | ○ |
| | | | 2ケ所 | 4ケ所 | 4ケ所 | 3ケ所 | 5ケ所 |

※1 NII/NAREGIにSS/RCS (NAREGIv1の予約サービス機能) を設置

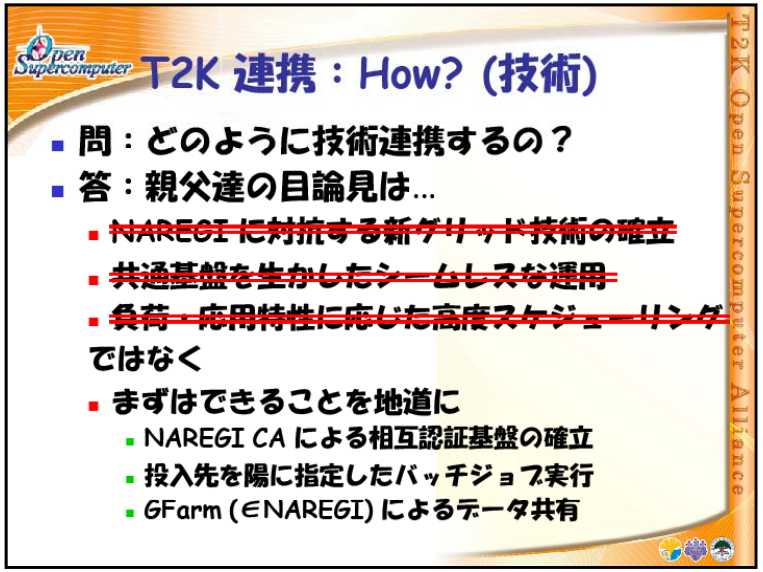
※2 名古屋大学の管理ノードは、学内サービス向けの設定のまま、IS-CDAS以下を連携用に追加設定

# Fact Sheet 2: 計算ノード構成

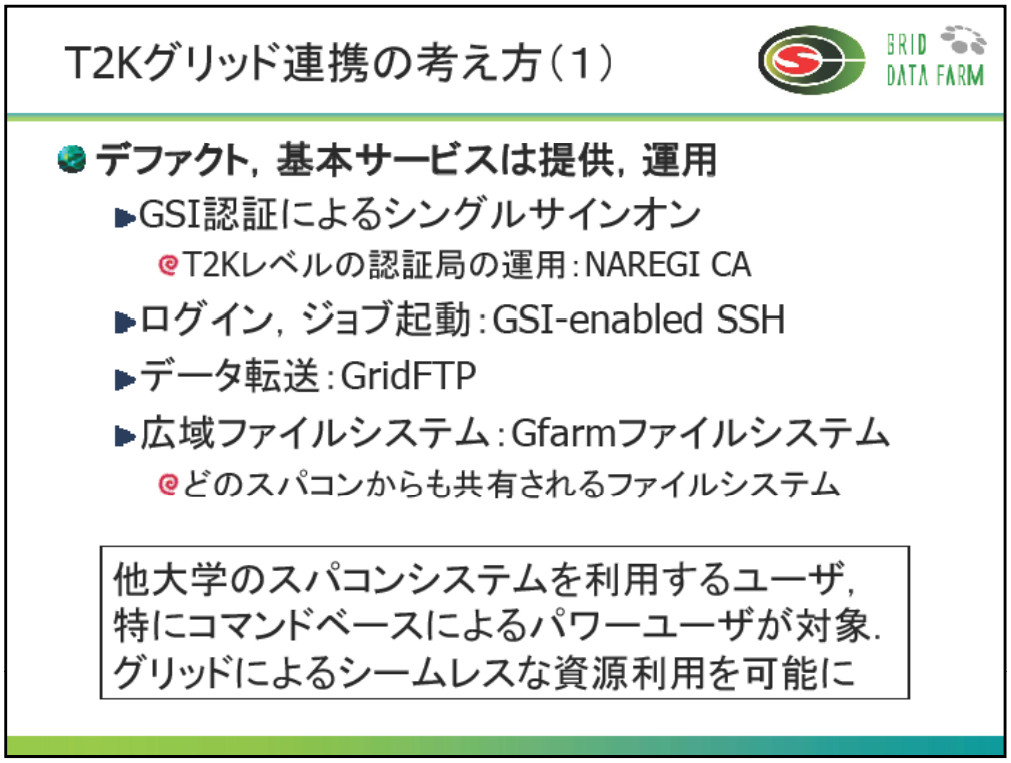
| | | アーキテクチャ | OS | スケジューラ | ノード数 | TFLOPS |
|---|---|---|---|---|---|---|
| 大阪大学 | gridvms1.hpc.cmc | SX-8R | SUPER-UX | NEC NQS-II | 1 | 0.3 |
| | gridvms2.hpc.cmc | x86 | Linux | NEC NQS-II | 8 | 0.4 |
| | gridvms3.hpc.cmc | x86 | Linux | NEC NQS-II | 450 | 16.8 |
| 東京工業大学 | tggn-vms2.grp.gsic | x86+ClearSpeed | Linux | Sun GridEngine | 120 | 18.9 |
| 九州大学 | dpc.kyushu.grid | x86 | Linux | PBS Pro | | 0.1 |
| NII/NAREGI | pbg2043 | SX-8 | SUPER-UX | NEC NQS-II | 2 | 0.2 |
| | pfg1005, png1051, png1053, png3000 | x86 | Linux | PBS Pro | 14 | 0.1 |
| 分子科学研究所 | dpca064.grid, dpca128.grid, dpcb064.grid, dpcb128.grid, dpcc128.grid, dpcd048.grid, dpcd049.grid, dpcd057.grid | x86 | Linux | PBS Pro | 278 | 3.4 |
| | sr11k.grid | POWER5 | AIX | LoadLeveler | 32 | 3.5 |
| 名古屋大学 | naregi4.cc | x86 | Linux | PBS Pro | 6 | 0.2 |
| | ngrd1.cc | SPARC | Solaris | Parallelnavi | 2 | 0.3 |
| | | | | | | 44.1 |

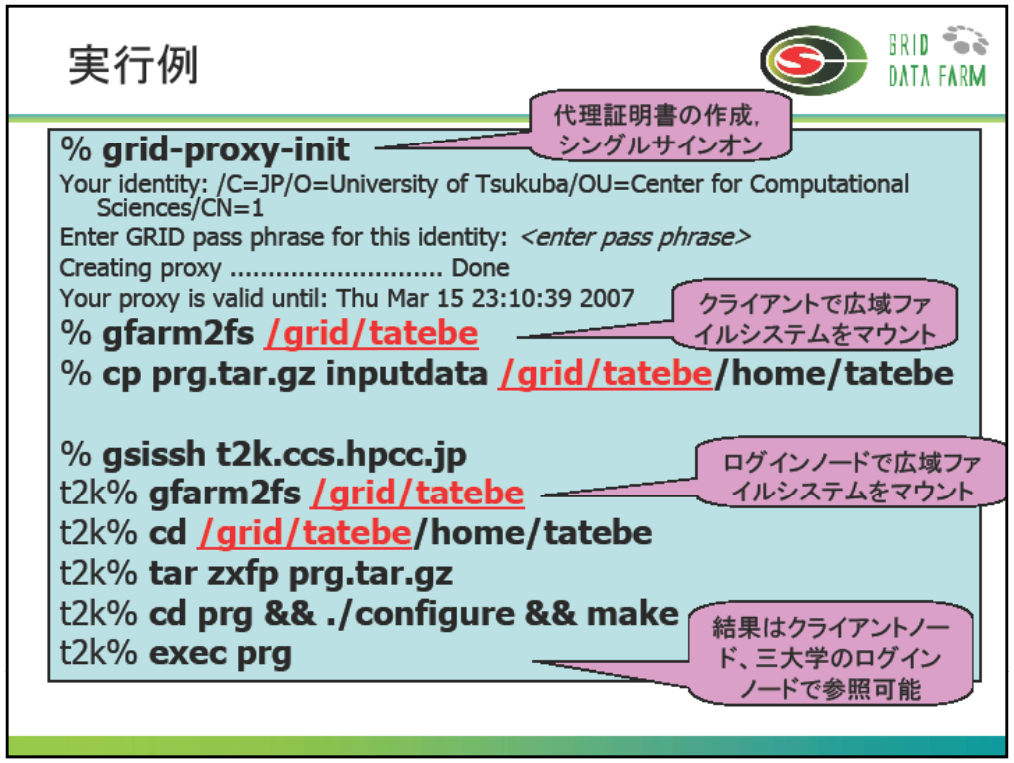
# T2Kのアプローチ

- NAREGI CAによる相互認証基盤の確立
- 投入先を陽に指定したバッチジョブ実行
- Gfarmによるデータ共有




中島 浩, “T2k連携とグリッド運用”,
T2Kシンポジウムつくば 2008.



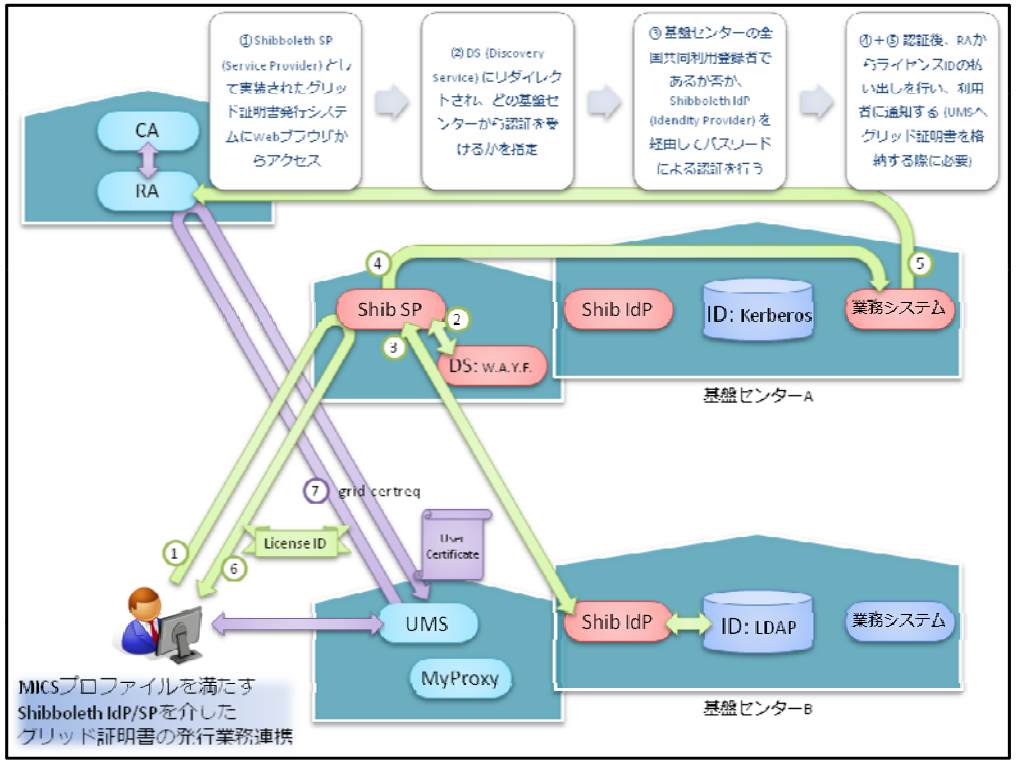



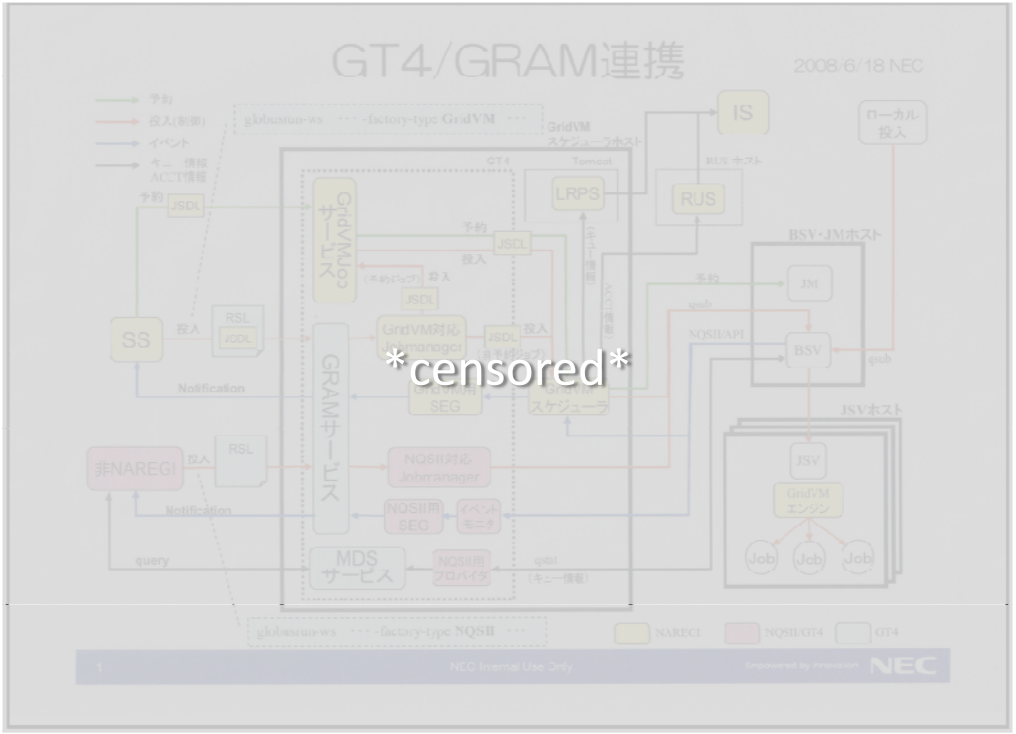

建部 修見, “T2k連携とグリッド運用”, T2Kシンポジウムつくば 2008.

# 阪大CMCのアプローチ - その2

## T2Kグリッド連携の刺激を受けて共存を考え始める

- 第2段階
  - 他の基盤センターの登録ユーザにもグリッド証明書を発行
    - MICSプロファイルを満たすShibboleth SP/IdPによる連携
  - 提供資源を非排他的に共有
    - ローカルスケジューラの予約マップをメタスケジューラに後方からインジェクション

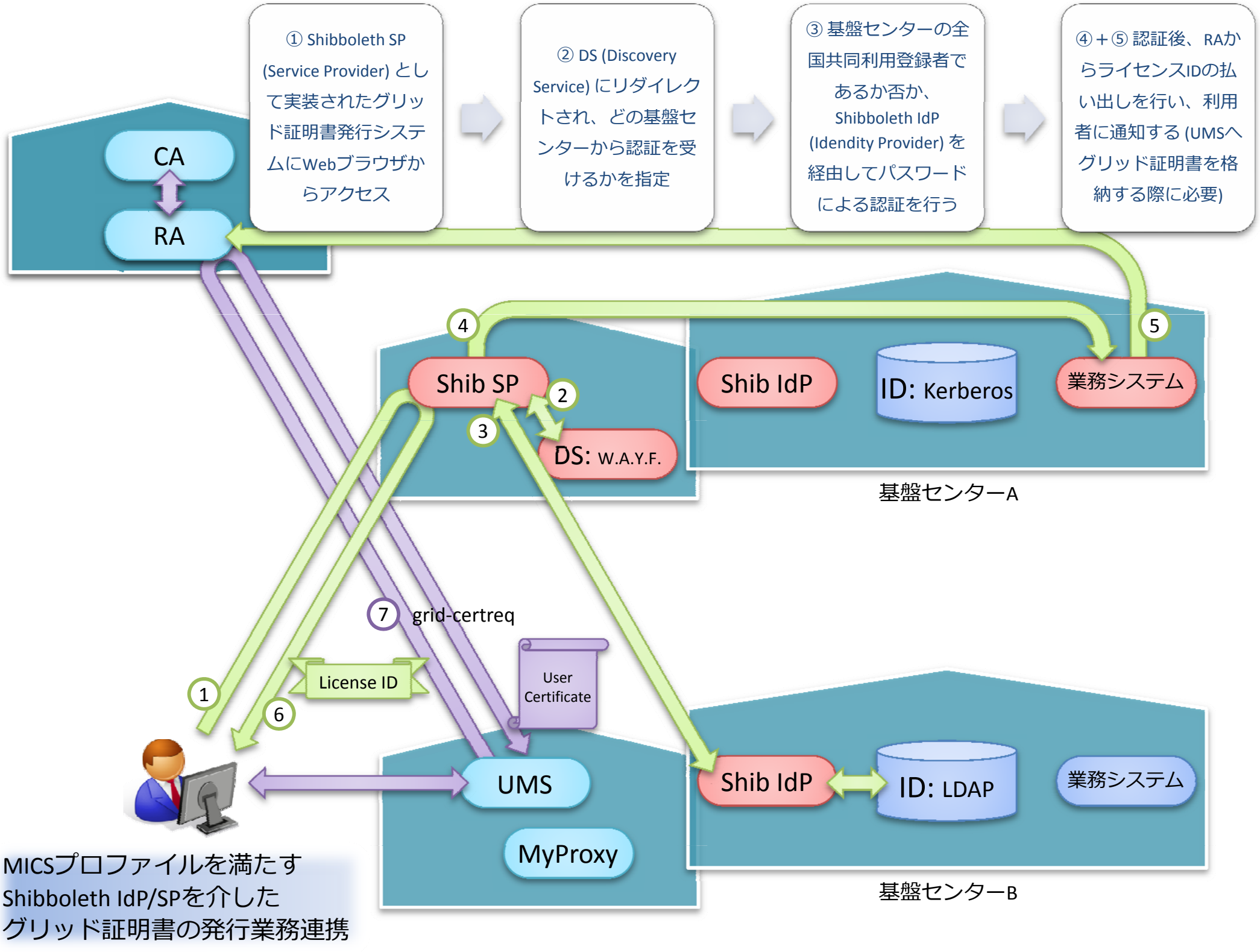


# OpenNAREGI (仮称)

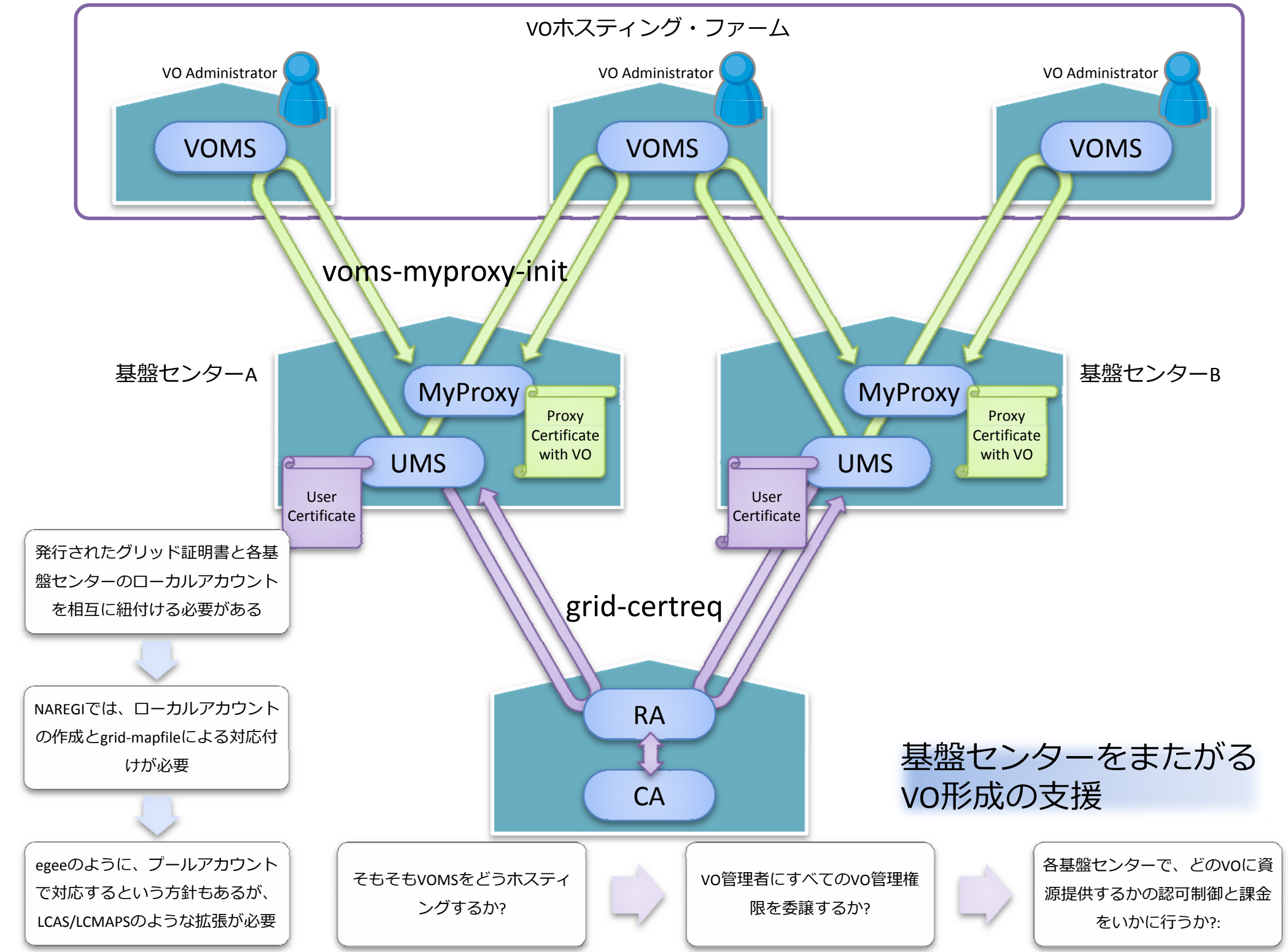
- NAREGIの開発主導権を国情研から基盤センターへ
  - 現場の要望に即した開発項目の洗い直し
    - CUIによる簡素な操作
    - GridVMの簡素化、対応プラットフォームの拡張
    - 運用管理機能の整備
  - ソースコードの開示だけでなく開発プロセスのオープン化
    - Webによるコードの閲覧と静的解析
    - 外部開発者によるコードの貢献
    - 自動ビルト、自動パッケージ化、自動統合テスト

# 基盤センターの
# 共同利用登録と連動した
# グリッド認証局業務の自動化




MICSプロファイルを満たす
Shibboleth IdP/SPを介した
グリッド証明書の発行業務連携

# VO管理者への
# VO管理権限の委譲


VOホスティング・ファーム
VO Administrator
VO Administrator
VO Administrator
VOMS
VOMS
VOMS
voms-myproxy-init
基盤センターA
基盤センターB
MyProxy
MyProxy
Proxy Certificate with VO
Proxy Certificate with VO
UMS
UMS
User Certificate
User Certificate
grid-certreq
RA
CA
発行されたグリッド証明書と各基盤センターのローカルアカウントを相互に紐付ける必要がある
NAREGIでは、ローカルアカウントの作成とgrid-mapfileによる対応付けが必要
egeeのように、プールアカウントで対応するという方針もあるが、LCAS/LCMAPSのような拡張が必要
そもそもVOMSをどうホスティングするか?
VO管理者にすべてのVO管理権限を委譲するか?
各基盤センターで、どのVOに資源提供するかの認可制御と課金をいかに行うか?:
基盤センターをまたがるVO形成の支援

# NAREGI-6

## 認証ポリシーが異なるセンター間の相互連携

- 複数の認証局が発行した証明書を利用できるNAREGI計算機資源環境を構築する
- 実際に運用中の計算機センターの大規模資源に対して、NAREGIミドルウェアからジョブ投入できる環境を構築する

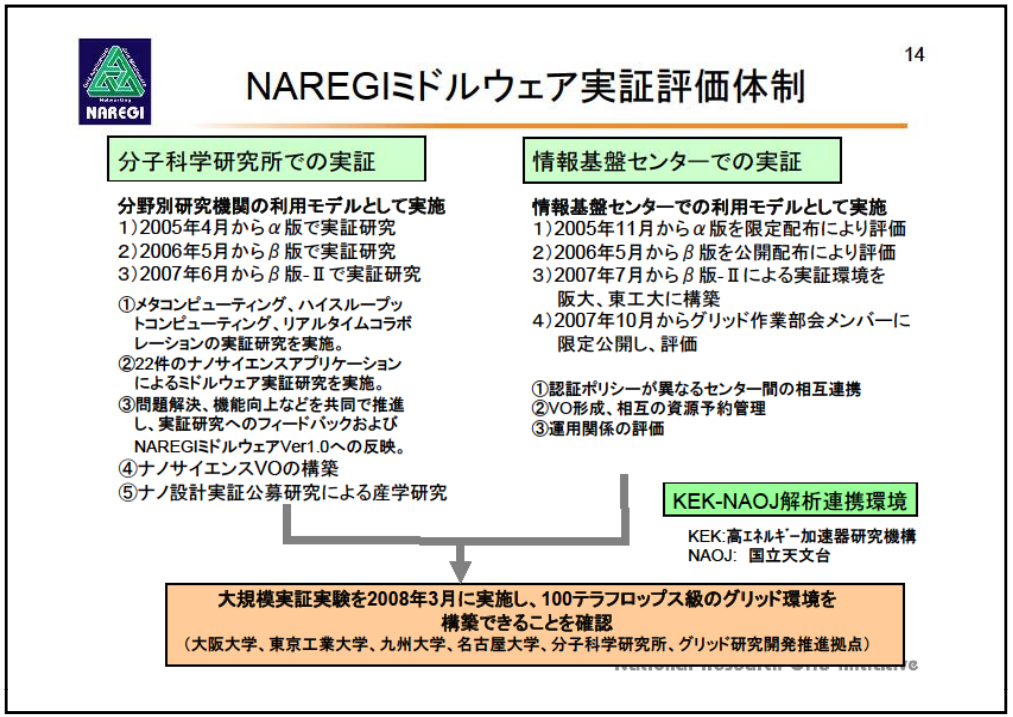
## VO形成、相互の資源予約管理

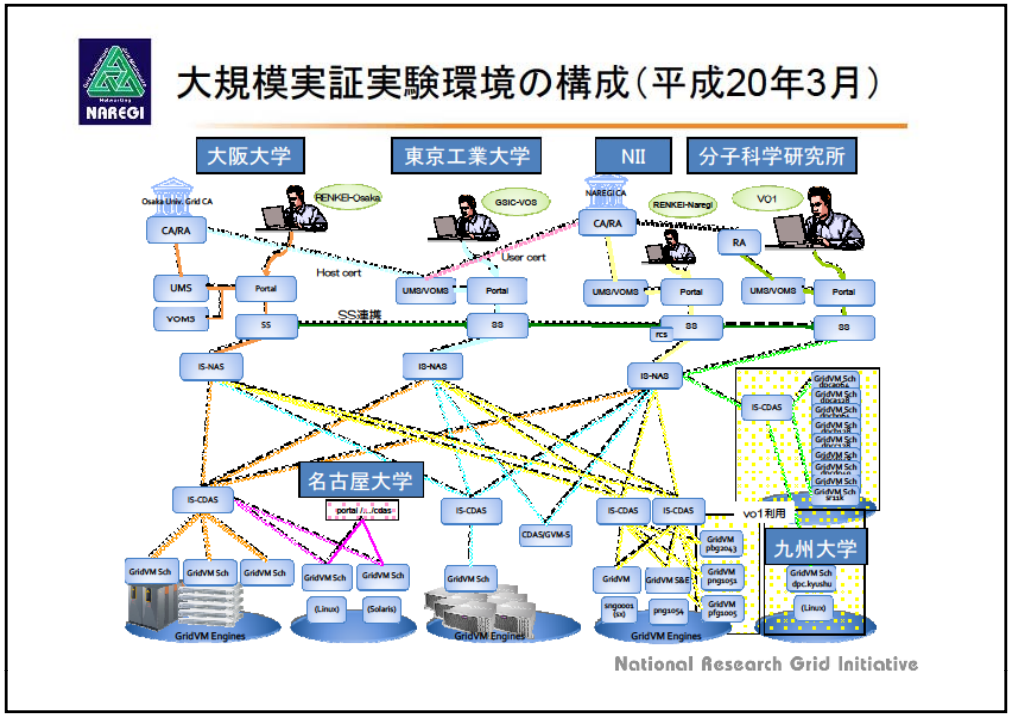
- 各拠点から提供された計算機資源のAUP (利用規定) に対応するVO (仮想組織) を形成する
- 複数のメタスケジューラが他スケジューラの資源予約の状況を反映した資源予約を行ったうえで、実アプリケーションによるジョブ投入ができる環境を構築する

## 運用関係の評価

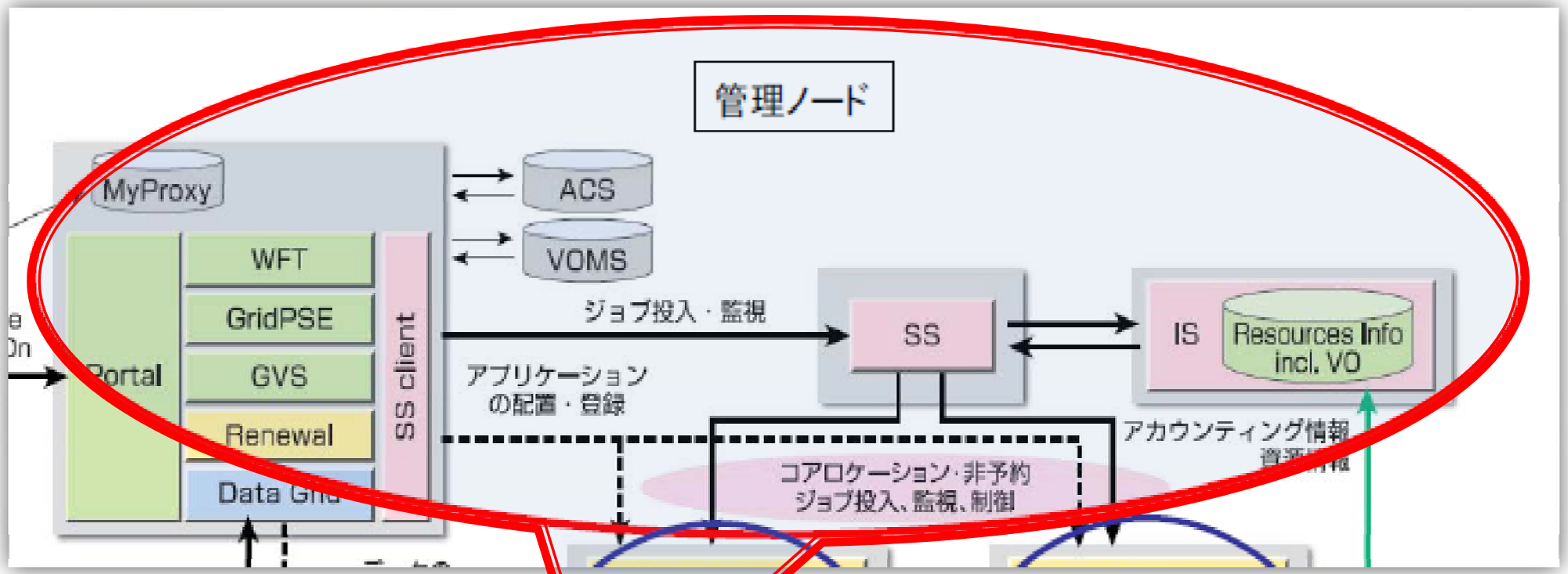
- 実運用環境に展開するに先だって支援体制の実地評価を行う
  - GOC (Grid Operation Center)
  - PERT (Performance Enhancement and Response Team)

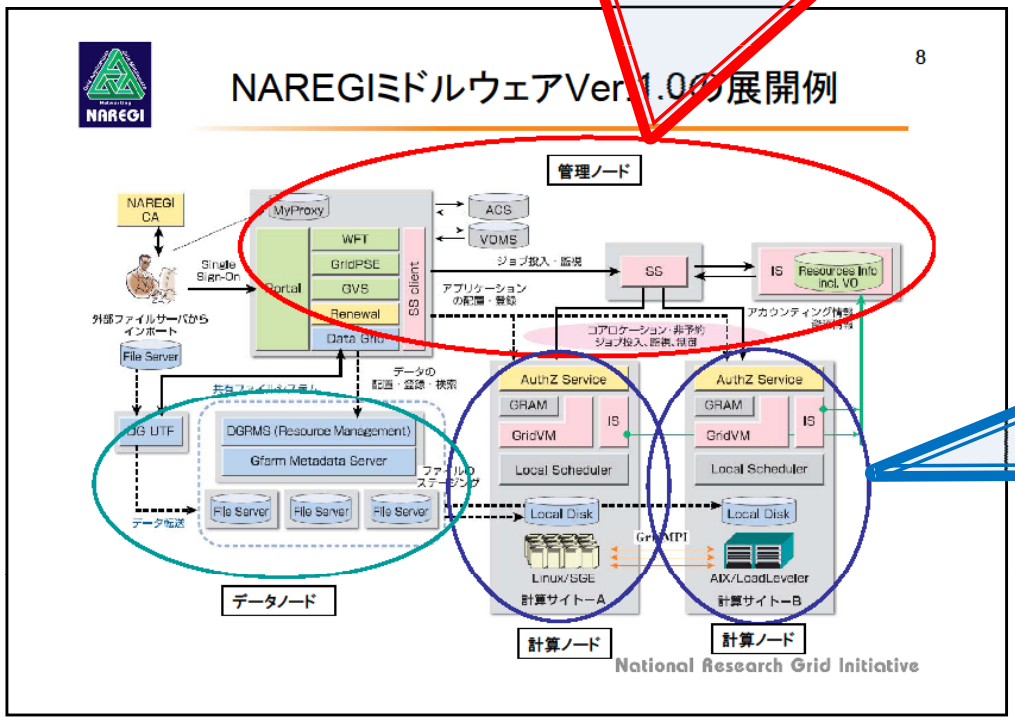
# NAREGIミドルウェアで100TFLOPS級のグリッド環境を構築できるか!?

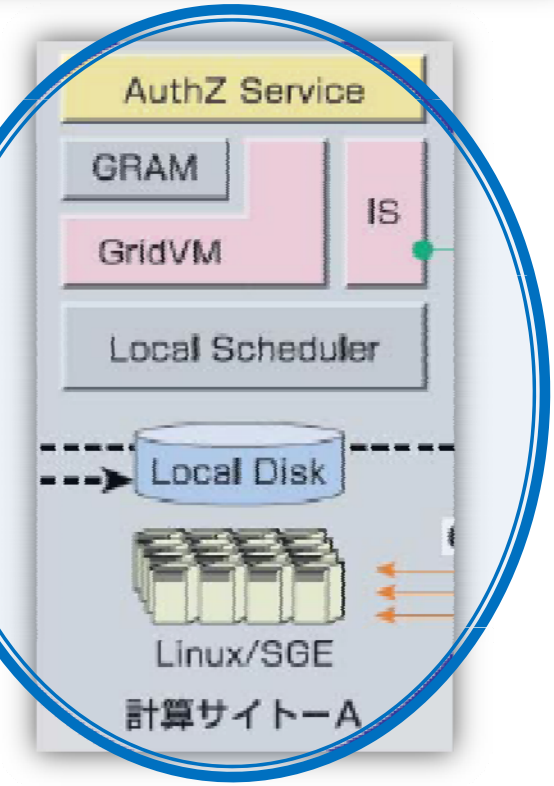



動き出したサイエンスグリッドNAREGI
―研究リソース共有の世界を広げるミドルウェアを公開―
平成20年5月9日付けプレスリリースより抜粋


管理ノード
MyProxy
ACS
VOMS
WFT
GridPSE
GVS
Renewal
Data Grid
Portal
SS client
ジョブ投入・監視
SS
IS
Resources Info
incl. VO
アプリケーション
の配置・登録
アカウンティング情報
コアロケーション・非予約
ジョブ投入、監視、制御


NAREGI
NAREGIミドルウェアVer.1.0の展開例
8
管理ノード
NAREGI CA
Single Sign-On
外部ファイルサーバからインポート
File Server
MyProxy
ACS
VOMS
WFT
GridPSE
GVS
Renewal
Data Grid
Portal
SS client
ジョブ投入・監視
SS
IS
Resources Info incl. VO
アプリケーション の配置・登録
コアロケーション・非予約 ジョブ投入、監視、制御
アカウンティング情報
データの配置・登録・検索
OG UTF
DGRMS (Resource Management)
Gfarm Metadata Server
File Server
File Server
File Server
ファイルのステージング
データ転送
AuthZ Service
GRAM
GridVM
IS
Local Scheduler
Local Disk
GridMPI
Linux/SGE
計算サイトA
AIX/LoadLeveler
計算サイトB
データノード
計算ノード
計算ノード
National Research Grid Initiative


AuthZ Service
GRAM
GridVM
IS
Local Scheduler
Local Disk
Linux/SGE
計算サイトA

## 大阪大学

- 実運用システムと混在し、ローカルスケジューラの仮想キューを組み替えることで柔軟に資源提供

## 東京工業大学

- 運用中のシステムを短期間に組み替えて資源提供

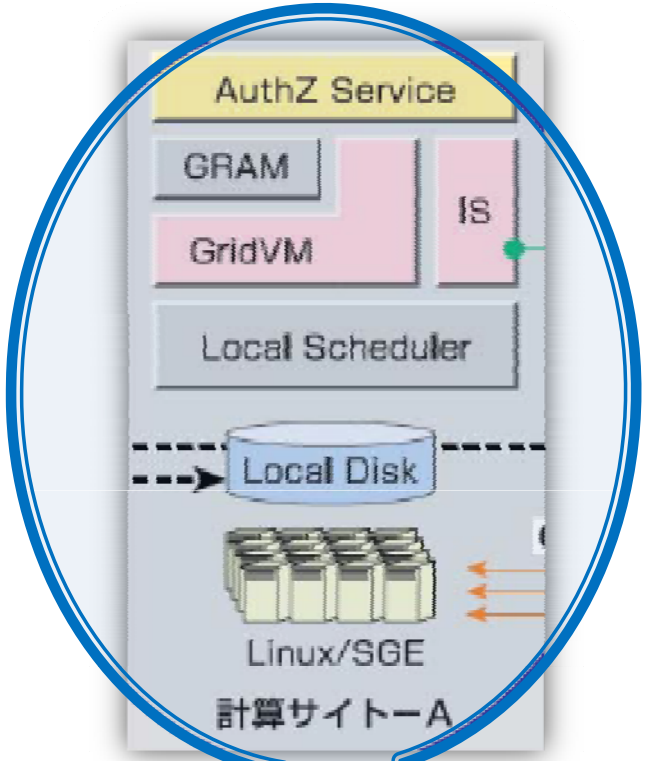
## 名古屋大学

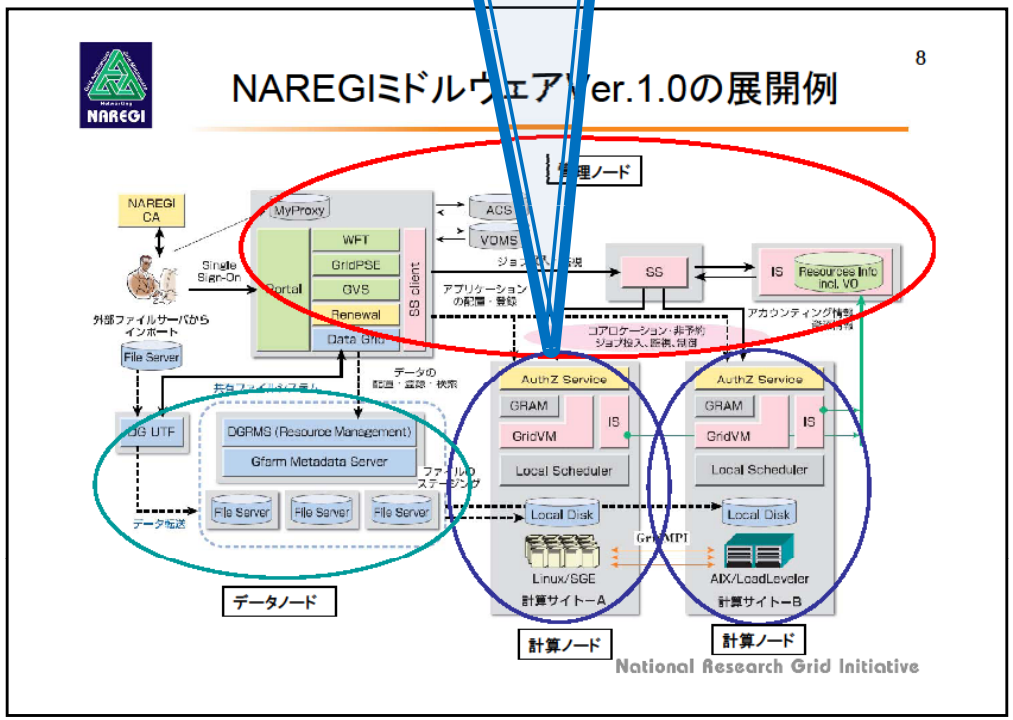
- NAREGIミドルウェア評価用のシステムを学内向け管理システムと共有しつつ連携に資源提供

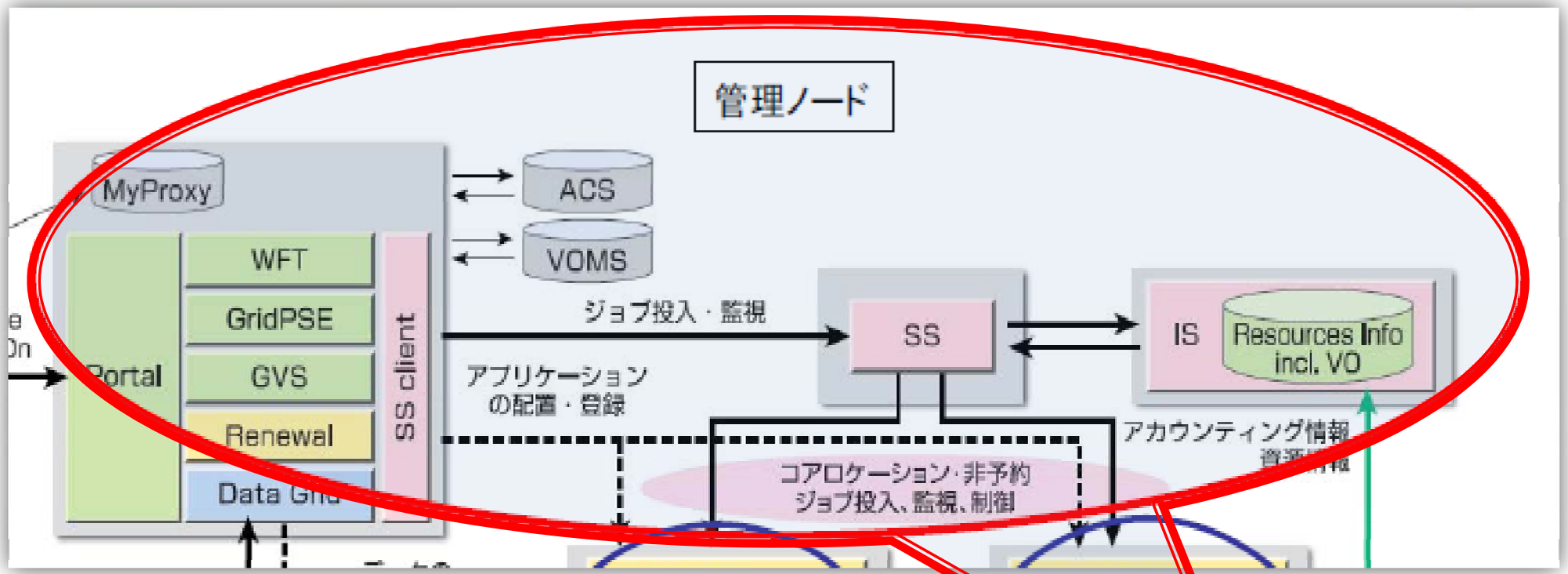
## NII/NAREGI/分子科学研究所/九州大学

- NAREGIミドルウェアで運用中のシステムから可能な限り資源提供
- AUPの差違は適切なVOを形成することによって吸収

各拠点の実情に合わせた
多様な計算機資源の提供を受けた連携



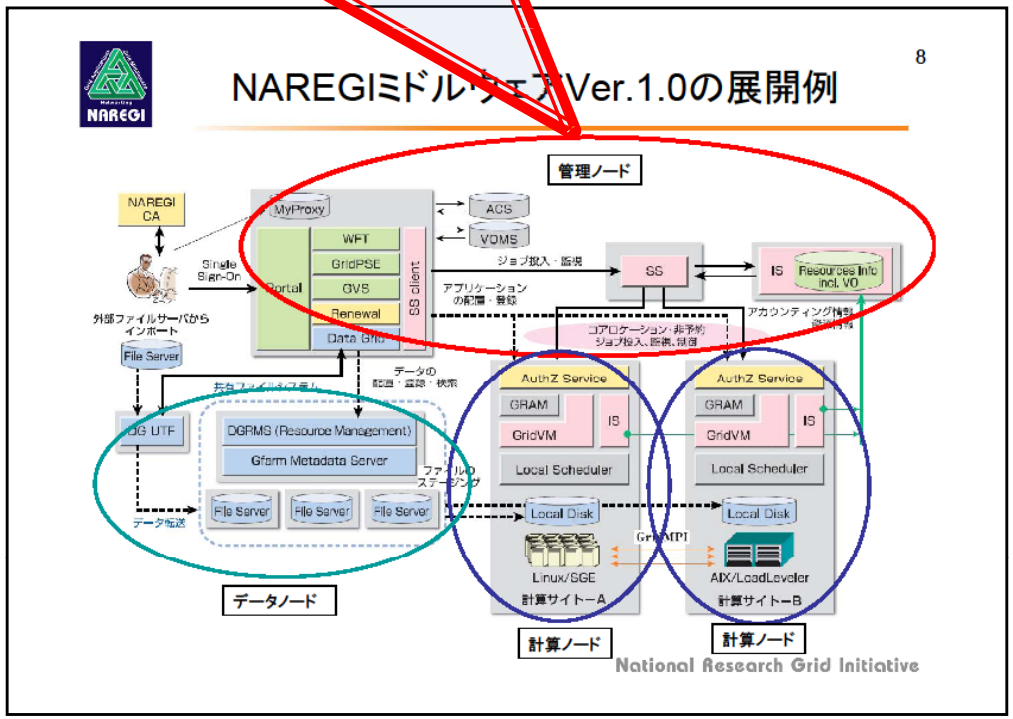



## SS: スーパースケジューラ

- 予約/非予約ジョブの投入・監視
- ジョブ投入したユーザが所属する仮想組織に提供された計算機資源から適切なものを予約

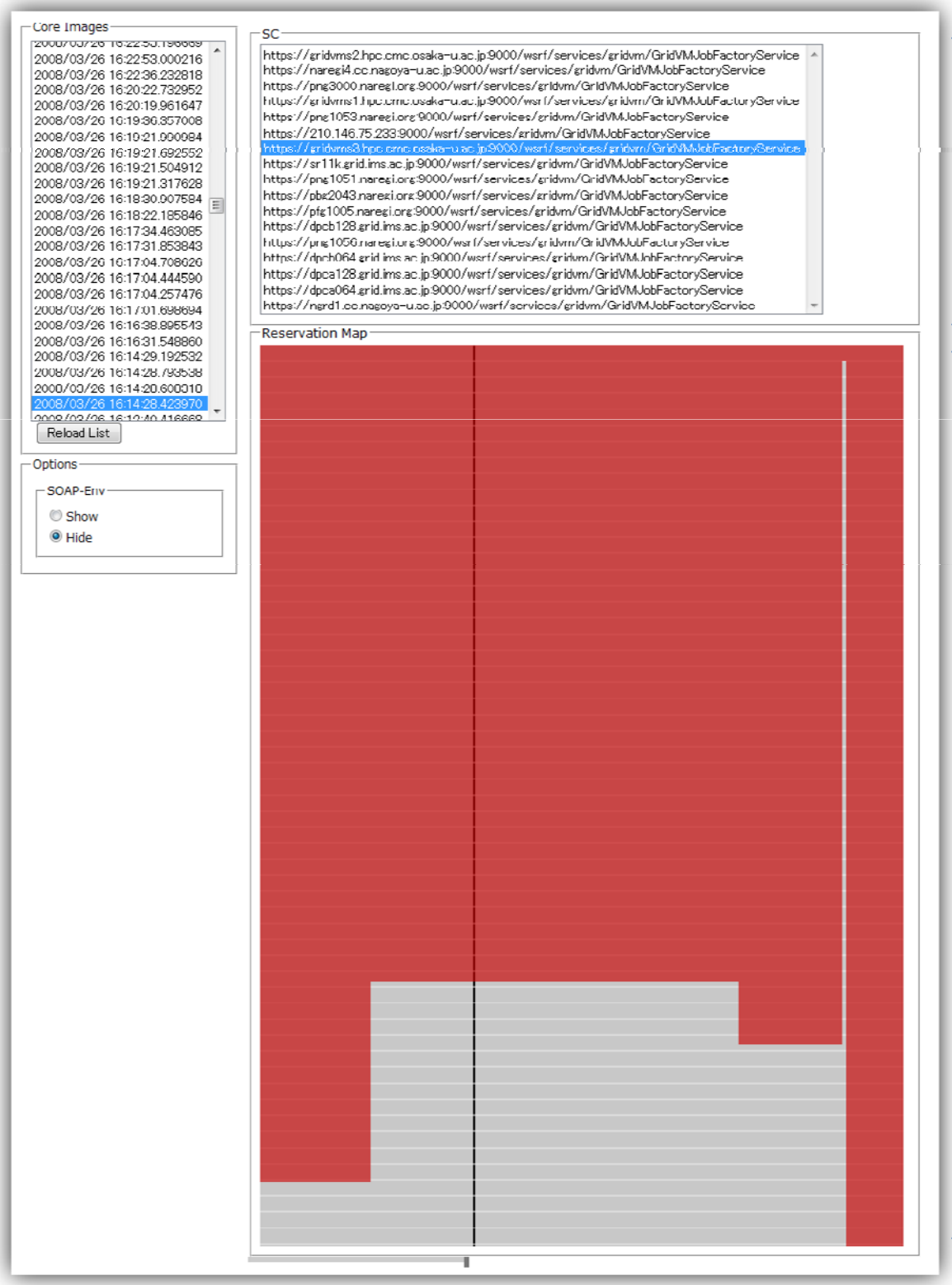
## IS: 情報サービス

- OSやCPU、メモリ量など計算機資源の情報とユーザや仮想組織の情報を集約
- 利用統計情報の蓄積



Webサービスのメッセージ交換による状態推移をすべて記録

Webサービスのメッセージ (SOAP Envelope) を解析するためのオプション




今回の連携に参加したクラスタ群:

3/26 16:14:28 時点で

- 17クラスタ
  - 887ノード

上記のうち、阪大CMCの遊休時利用型クラスタ (449ノード) の予約マップ

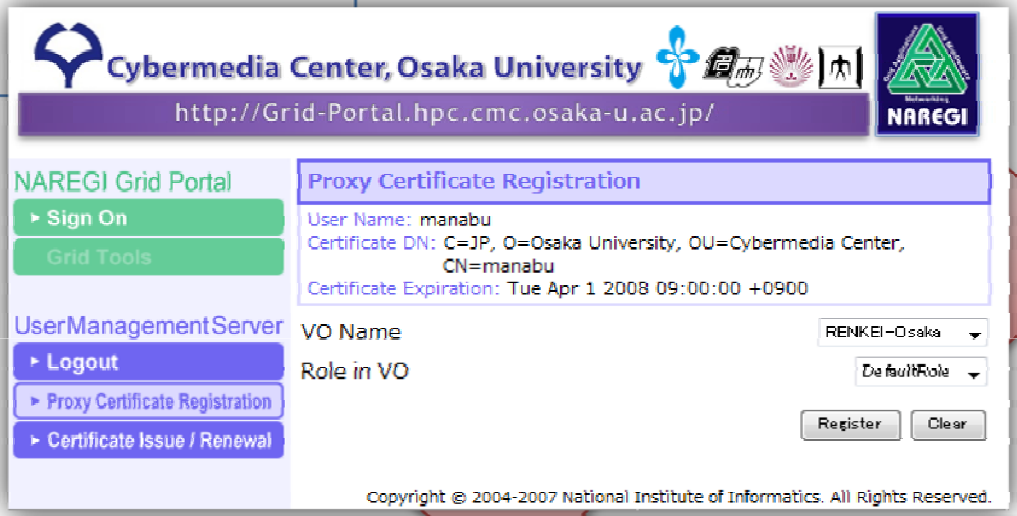
## How SS Works:

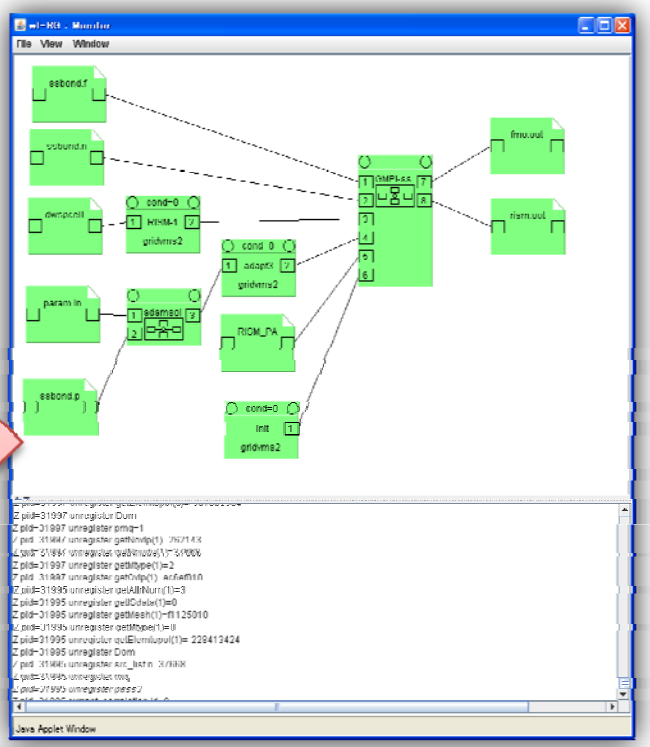
モニタツールによる監視

## 各拠点から様々なジョブを投入

- サイトを跨ったGridMPIジョブ
- ISVアプリケーション：Gaussian
- RISM-FMO連成計算

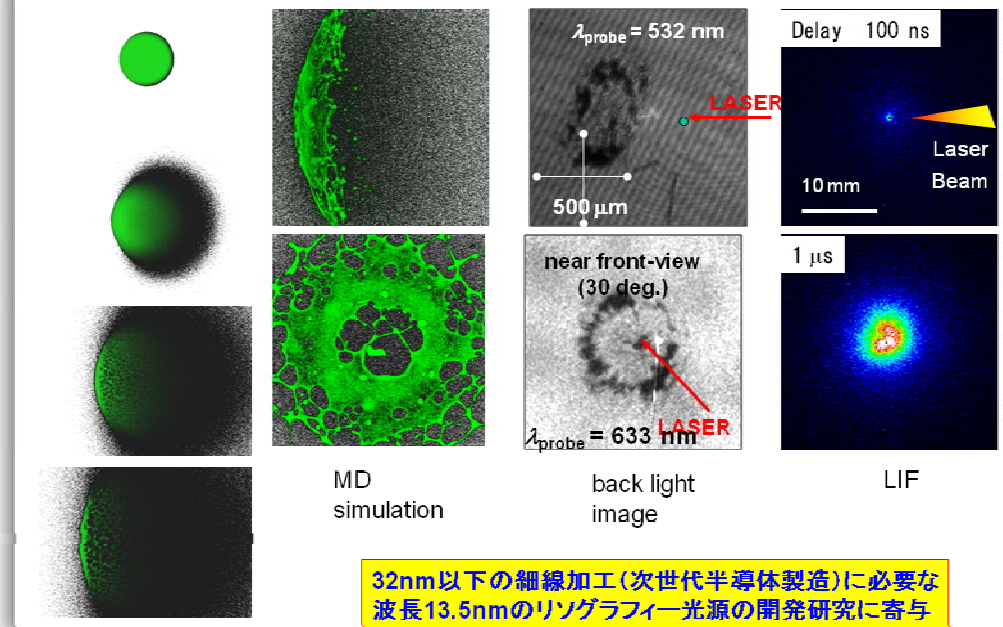
資料協力: 九州大学青柳研究室



GRID-MPI3

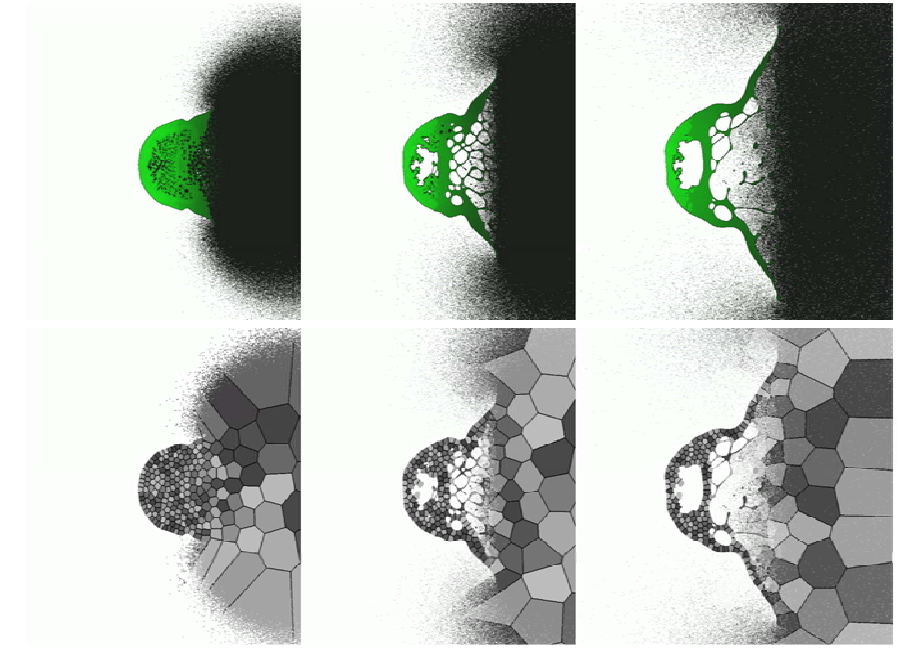
### 大規模GRID-MPI分子動力学シミュレーションにより，レーザー生成高エネルギー密度金属の破壊ダイナミックス実験を再現



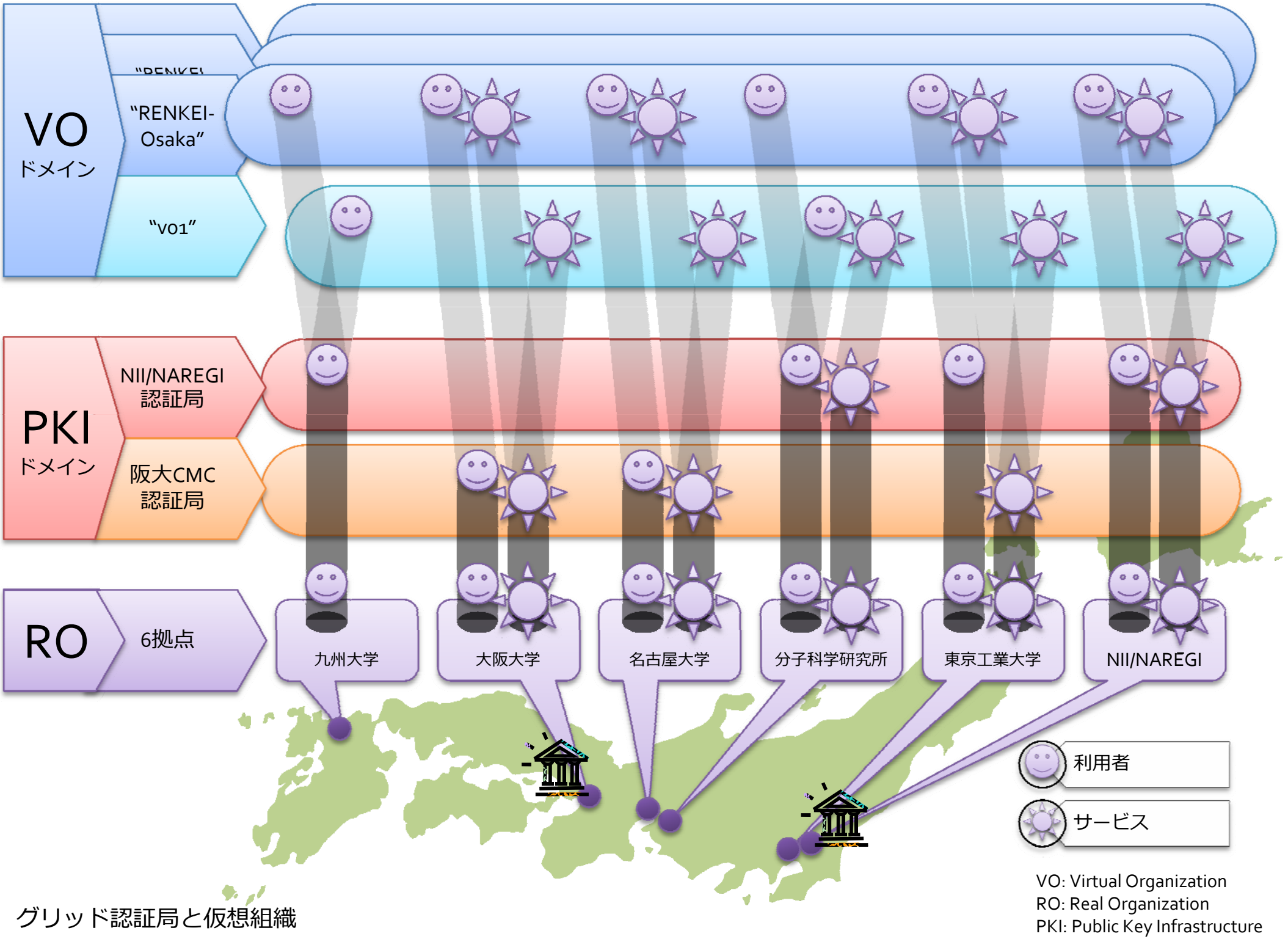

32nm以下の細線加工(次世代半導体製造)に必要な波長13.5nmのリソグラフィー光源の開発研究に寄与

GRID-MPI4

### 粒子のダイナミックスに応じて動的に計算分割領域を最適化し、各CPUの計算機負荷を等分にする



資料協力: 大阪大学レーザーエネルギー学研究センター (平成19年度CSI委託事業報告交流会発表予定資料より抜粋)

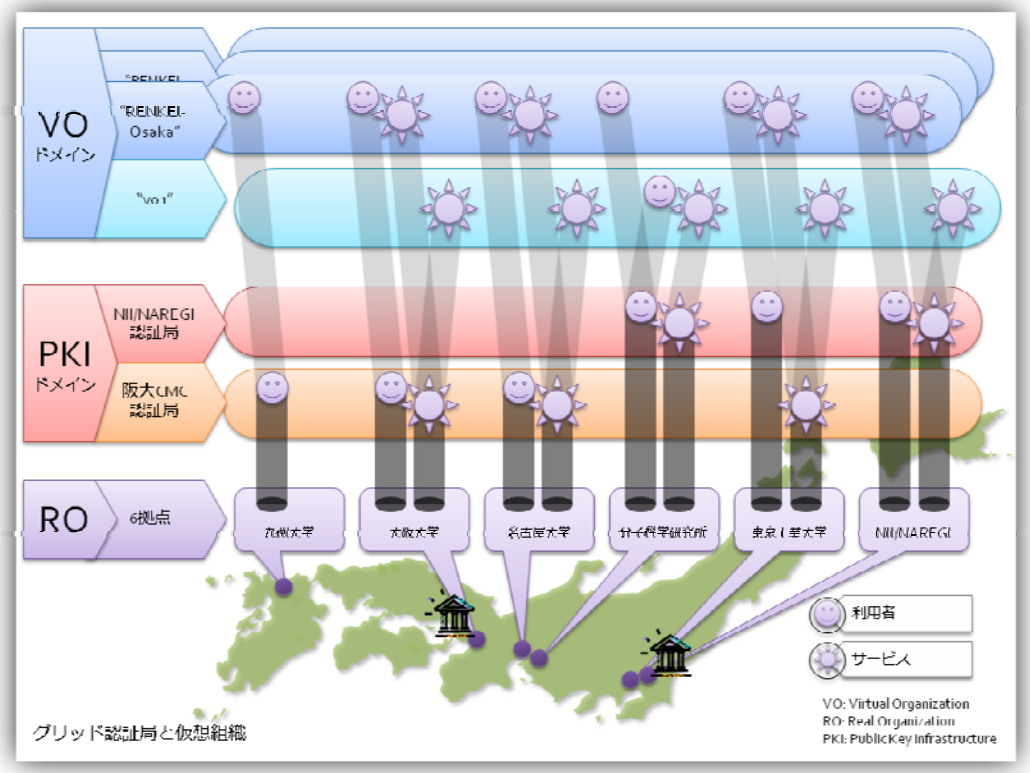



グリッド認証局と仮想組織

# “Registration Agency” 構想に向けて

## 今回のアカウンティング・ポリシー

- ローカル・アカウントを発行し、grid-mapfileで証明書と紐付ける
  - 大阪大学: 通常の全国共同利用アカウント発行にグリッド証明書が付随
  - 東京工業大学: 通常の全国共同利用アカウントを発行し、別途発行されたグリッド証明書を紐付け
  - その他: 一時アカウントを発行



グリッド認証局と仮想組織

## 代理店業務

- NII/NAREGIにて連携アカウントの代理発行業務
- 各拠点のアカウント発行に必要な情報を包括して収集
  - 氏名、職名、所属、研究分野、メールアドレス、電話番号など
- 各拠点に一括して代理申請
- 各拠点にて証明書との紐付けを行う

# 「死の谷」を越えて

- 「京速コン」への集中と最近よく聞く「All Japan」体制
  - 基盤センター群の呉越同舟、護送船団は昔から
  - 京速コンのお零れを巧く拾おう
- NAREGIという かつての泥船
  - その中でもNAREGI CAは悪くないソフト
  - 使ってみると他もそんなに悪くないと思えてくる!?
- NAREGIミドルウェアを使う
  - それよりも
    - 認証局をどう運用するかとか
    - VOをどうホスティングするかとか

    考える方が遙かに重要
  - その後でNAREGIが使えたら使えばいい
    - そんな感覚でいけるんじゃないかという感触を得られた連携実証でした