

# 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

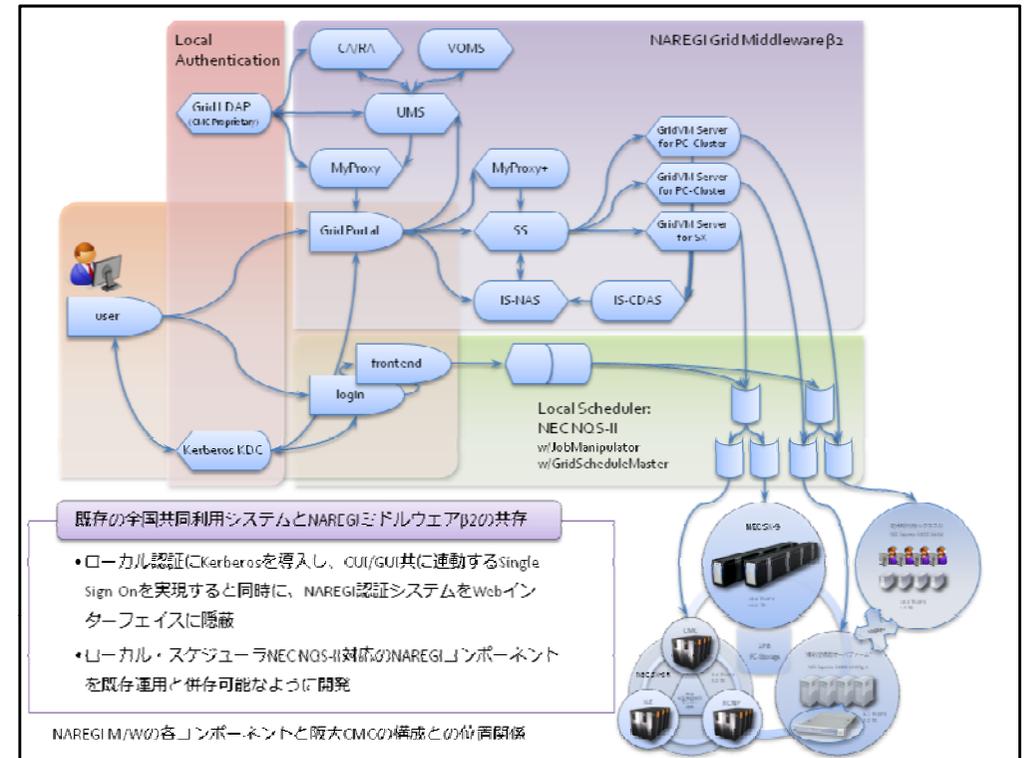
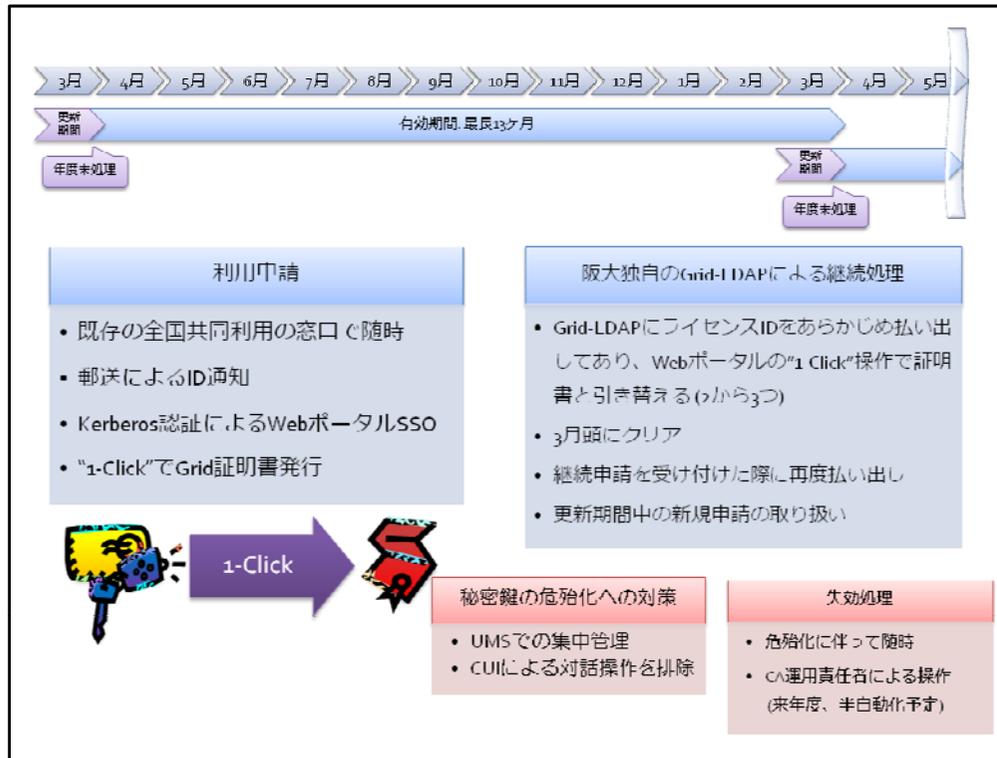
2008/09/26

# 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

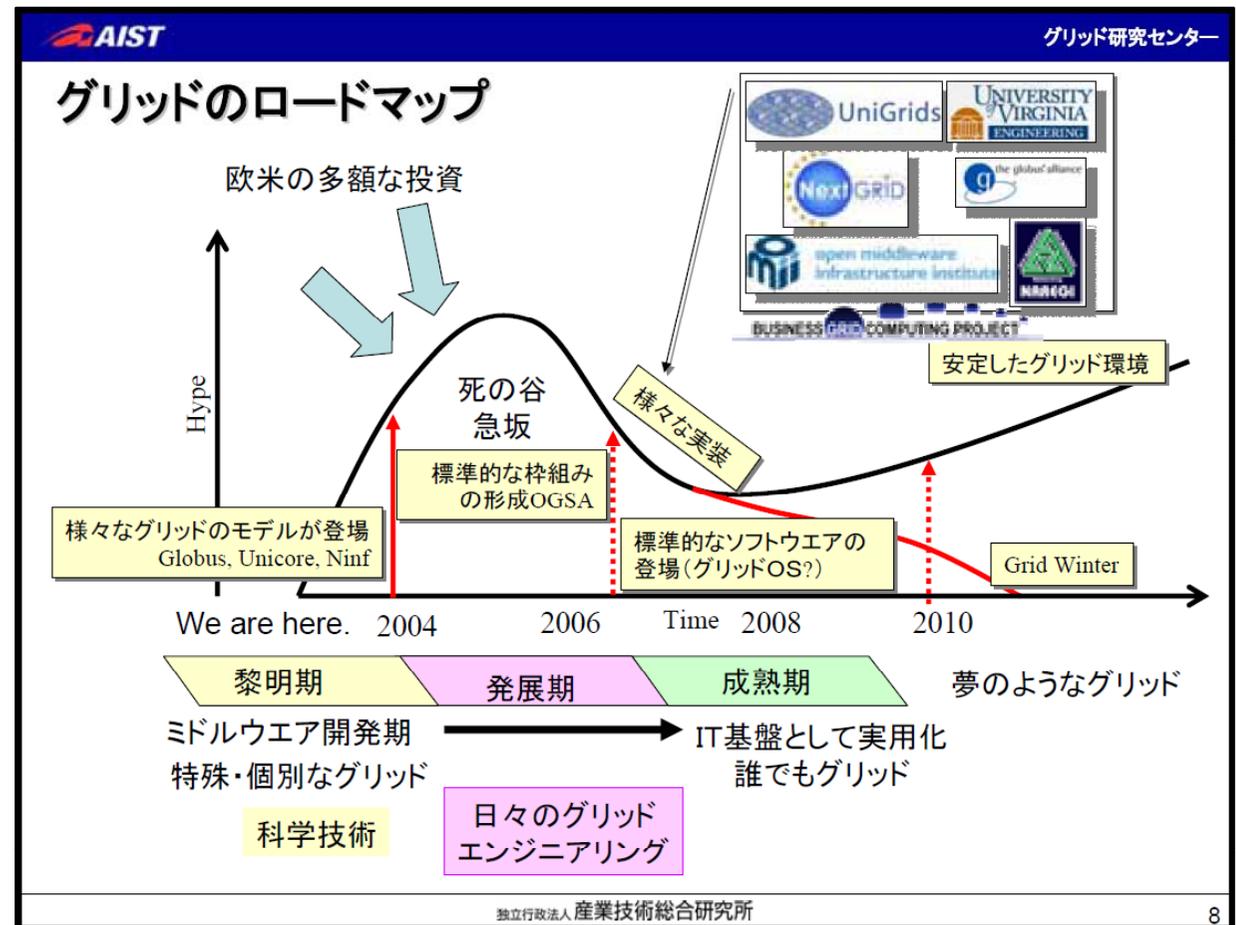


# NAREGI連携 (2007年度)

- 動機
  - 「SINET<sub>3</sub>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



# 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「三  
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: SCo7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)

The screenshot shows the NAREGI Grid Portal interface. At the top, there are logos for Cybermedia Center, Osaka University, and NAREGI. The URL is <http://Grid-Portal.hpc.cmc.osaka-u.ac.jp/>. The main content area is divided into two sections: "NAREGI Grid Portal" and "Proxy Certificate Registration".

**NAREGI Grid Portal**

- ▶ Sign On
- Grid Tools

**User Management Server**

- ▶ Logout
- ▶ Proxy Certificate Registration
- ▶ Certificate Issue / Renewal

**Proxy Certificate Registration**

User Name: manabu  
Certificate DN: C=JP, O=Osaka University, OU=Cybermedia Center, CN=manabu  
Certificate Expiration: Tue Apr 1 2008 09:00:00 +0900

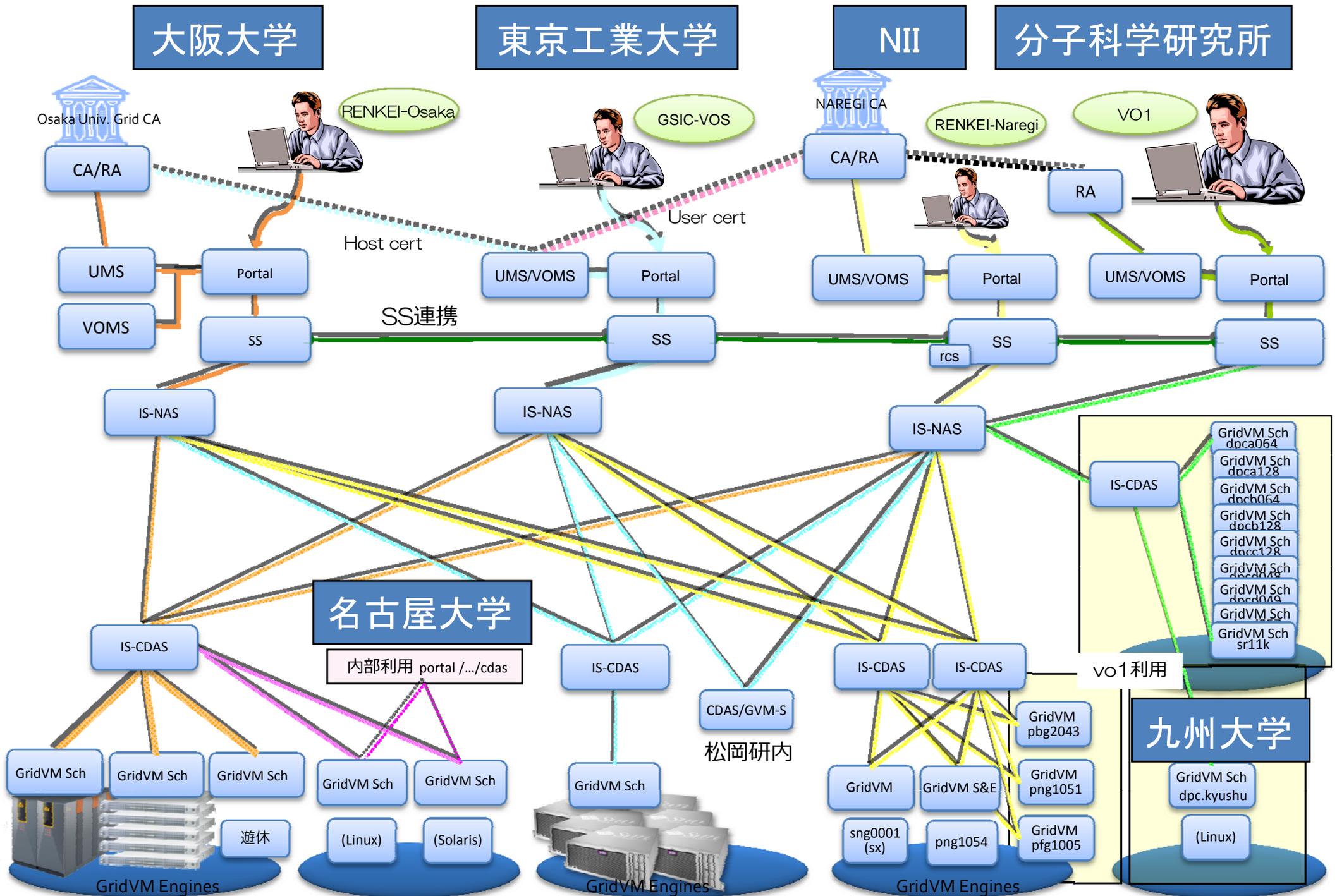
VO Name: FENKE-Osaka  
Role in VO: DefaultRole

Buttons: Register, Clear

Copyright © 2004-2007 National Institute of Informatics. All Rights Reserved.

# "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	—	○	○	○	○
九州大学		—	—	—	—	—	—
NII/NAREGI		1Gbps	○	○	◎※1	○	○
分子科学研究所	2008年1月	1Gbps	—	○	○	—	○
名古屋大学	2008年3月	1Gbps	—	△※2	△※2	△※2	○
			2ヶ所	4ヶ所	4ヶ所	3ヶ所	5ヶ所

※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 NOS-II	1	0.3
	gridvms2.hpc.cmc	x86	Linux	NEC NOS-II	8	0.4
	gridvms3.hpc.cmc	x86	Linux	NEC NOS-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 NOS-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, dpcbo64.grid, dpcb128.grid, dpcc128.grid, dpcdo48.grid, dpcdo49.grid, dpcdo57.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

Open Supercomputer T2K 連携 : How? (技術)

問 : どのように技術連携するの?

答 : 親父達の目論見は...

- ~~NAREGI に対抗する新グリッド技術の確立~~
- ~~共通基盤を生かしたシームレスな運用~~
- ~~負荷 応用特性に応じた高度スケジューリングではなく~~
- **まずはできることを地道に**
  - 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 (∈NAREGI) によるデータ共有

T2K Open Supercomputer Alliance

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

## T2Kグリッド連携の考え方(1)



### ● デファクト, 基本サービスは提供, 運用

- ▶ GSI認証によるシングルサインオン
  - ◎ T2Kレベルの認証局の運用: NAREGI CA
- ▶ ログイン, ジョブ起動: GSI-enabled SSH
- ▶ データ転送: GridFTP
- ▶ 広域ファイルシステム: Gfarmファイルシステム
  - ◎ どのスパコンからも共有されるファイルシステム

他大学のスパコンシステムを利用するユーザ, 特にコマンドベースによるパワーユーザが対象. グリッドによるシームレスな資源利用を可能に

## 実行例



```
% grid-proxy-init
Your identity: /C=JP/O=University of Tsukuba/OU=Center for Computational Sciences/CN=1
Enter GRID pass phrase for this identity: <enter pass phrase>
Creating proxy ..... Done
Your proxy is valid until: Thu Mar 15 23:10:39 2007
% gfarm2fs /grid/tatebe
% cp prg.tar.gz inputdata /grid/tatebe/home/tatebe

% gsissh t2k.ccs.hpcc.jp
t2k% gfarm2fs /grid/tatebe
t2k% cd /grid/tatebe/home/tatebe
t2k% tar xzfp prg.tar.gz
t2k% cd prg && ./configure && make
t2k% exec prg
```

代理証明書の作成, シングルサインオン

クライアントで広域ファイルシステムをマウント

ログインノードで広域ファイルシステムをマウント

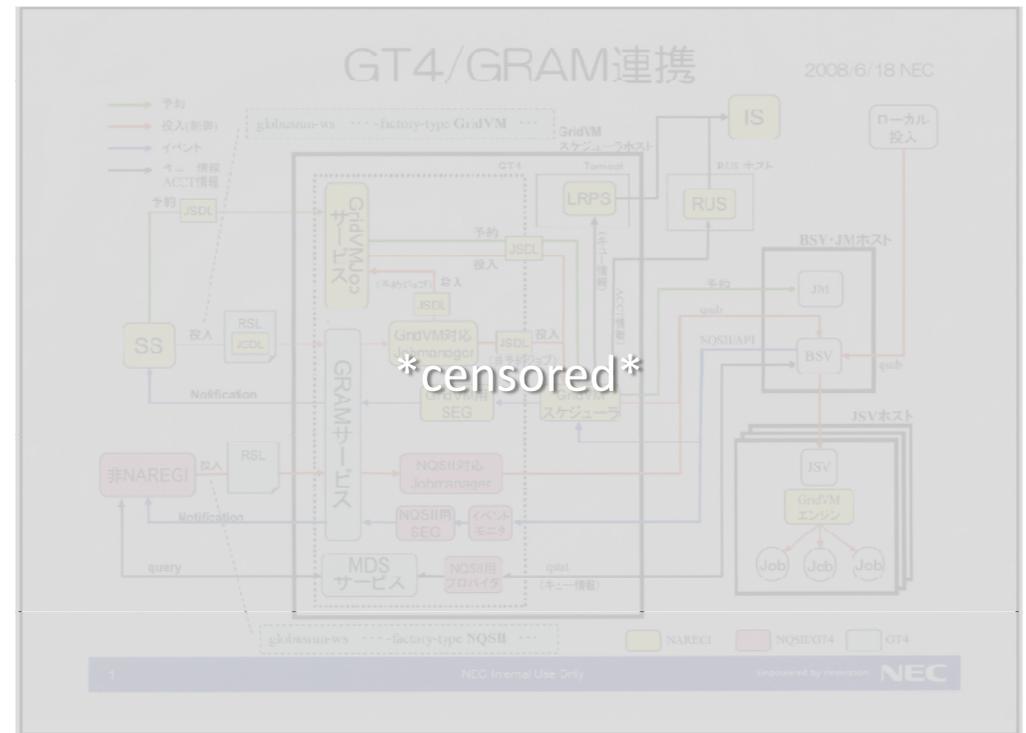
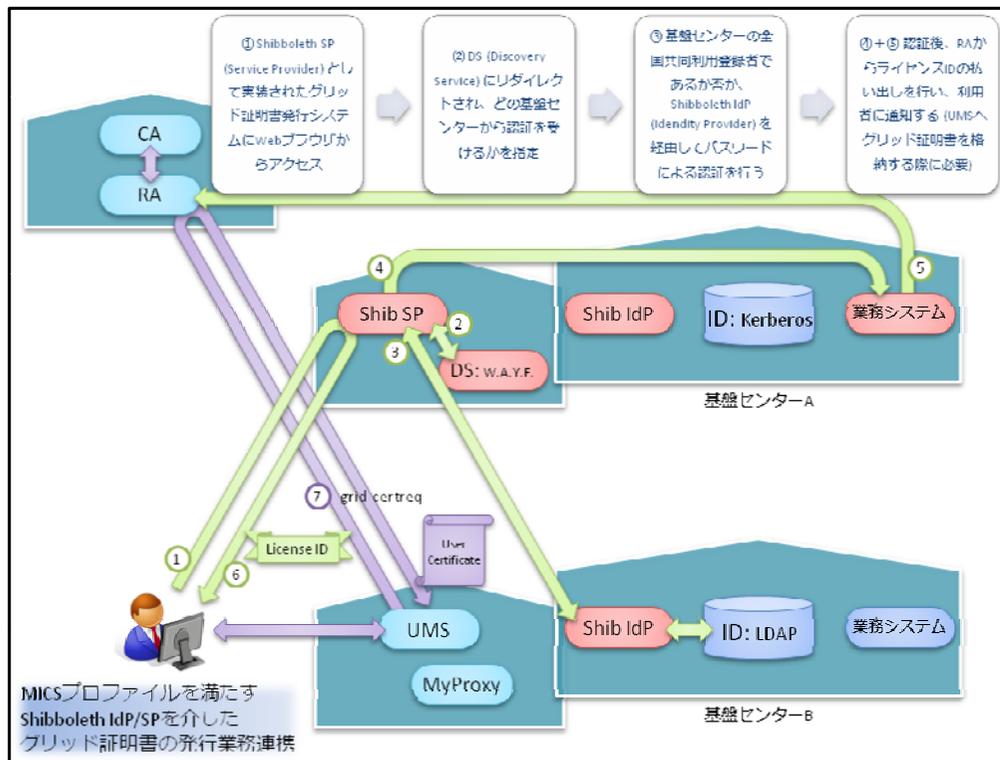
結果はクライアントノード, 三大学のログインノードで参照可能

# 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

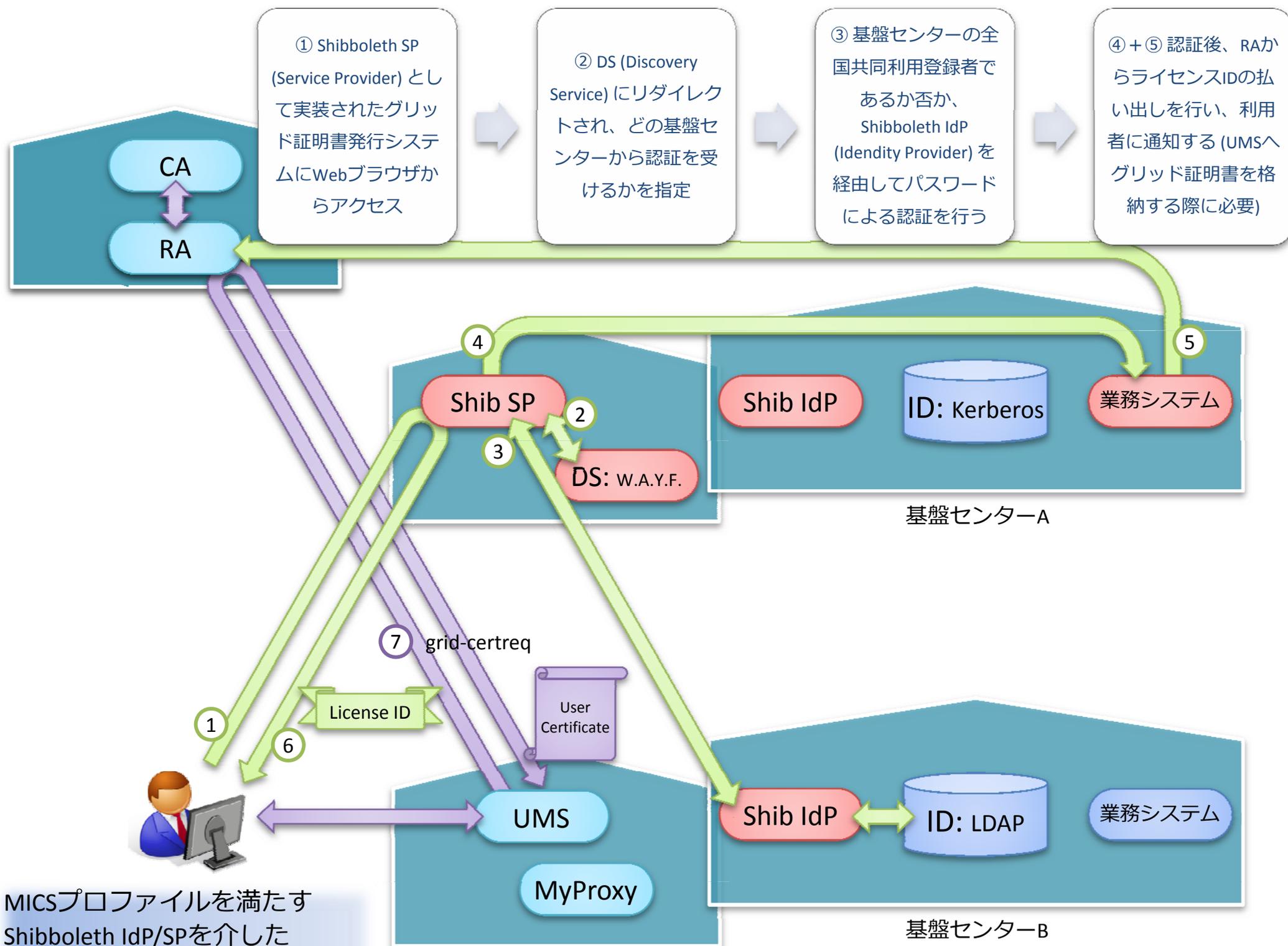
- 他の基盤センターの登録ユーザにもグリッド証明書を発行
  - MICSプロファイルを満たすShibboleth SP/IdPによる連携
- 提供資源を非排他的に共有
  - ローカルスケジューラの予約マップをメタスケジューラに後方からインジェクション



# 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

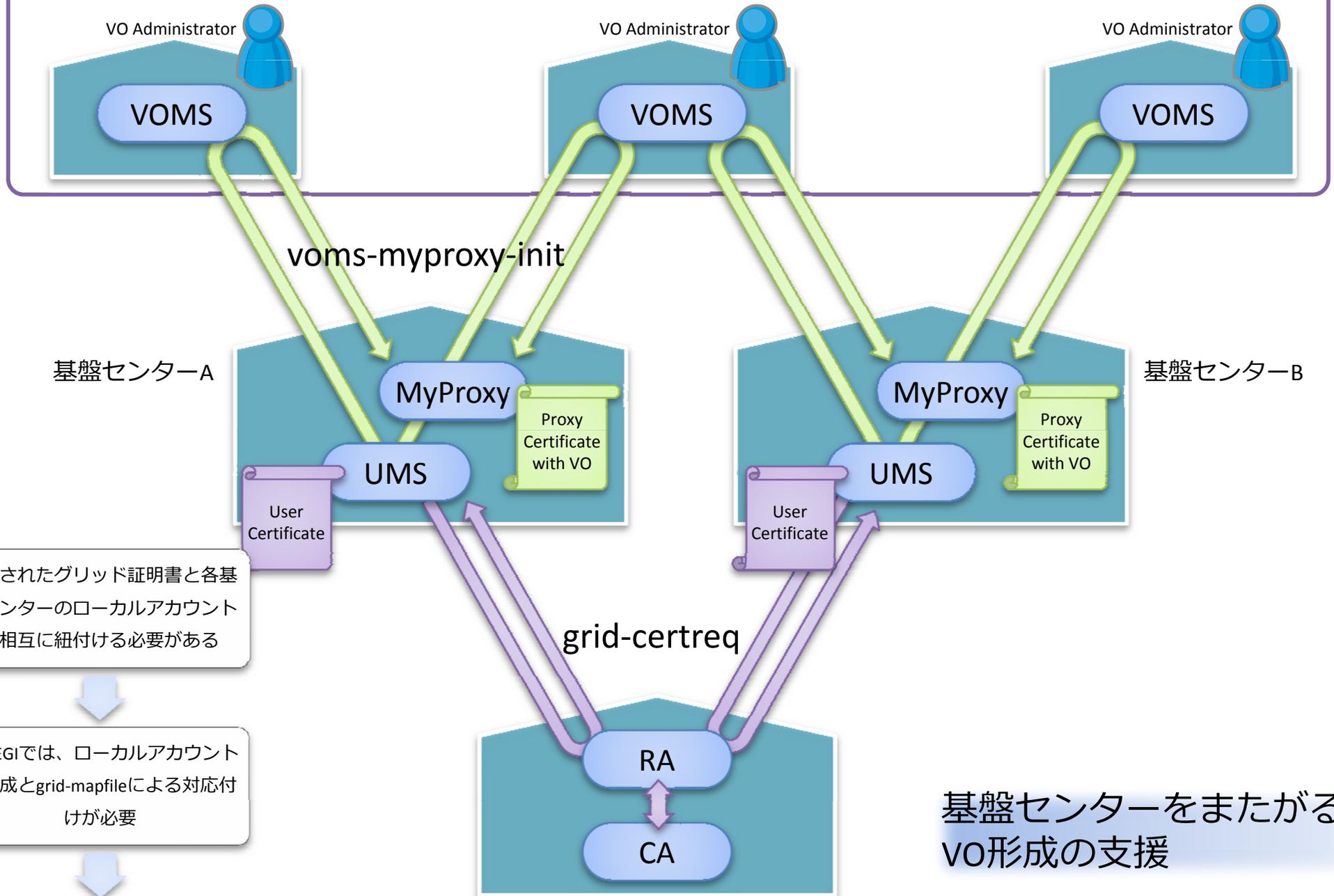
基盤センターの  
共同利用登録と連動した  
グリッド認証局業務の自動化



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



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に資源提供するかを認可制御と課金をいかに行うか?:

**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現地評価を行う
  - 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!?



## NAREGIミドルウェア実証評価体制

### 分子科学研究所での実証

- 分野別研究機関の利用モデルとして実施
- 2005年4月からα版で実証研究
  - 2006年5月からβ版で実証研究
  - 2007年6月からβ版-IIで実証研究
- メタコンピューティング、ハイスループットコンピューティング、リアルタイムコラボレーションの実証研究を実施。
  - 22件のナノサイエンスアプリケーションによるミドルウェア実証研究を実施。
  - 問題解決、機能向上などを共同で推進し、実証研究へのフィードバックおよびNAREGIミドルウェアVer1.0への反映。
  - ナノサイエンスVOの構築
  - ナノ設計実証公募研究による産学研究

### 情報基盤センターでの実証

- 情報基盤センターでの利用モデルとして実施
- 2005年11月からα版を限定配布により評価
  - 2006年5月からβ版を公開配布により評価
  - 2007年7月からβ版-IIによる実証環境を阪大、東工大に構築
  - 2007年10月からグリッド作業部会メンバーに限定公開し、評価
- 認証ポリシーが異なるセンター間の相互連携
  - VO形成、相互の資源予約管理
  - 運用関係の評価

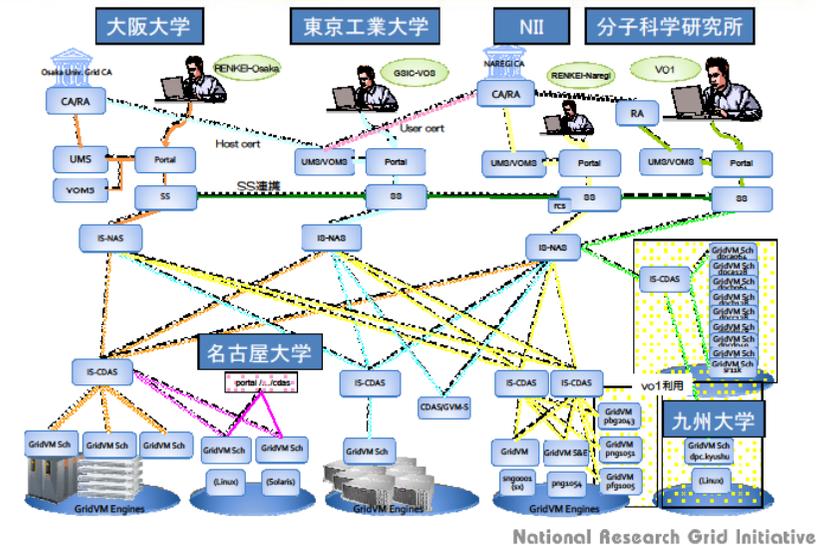
### KEK-NAOJ解析連携環境

KEK: 高エネルギー加速器研究機構  
NAOJ: 国立天文台

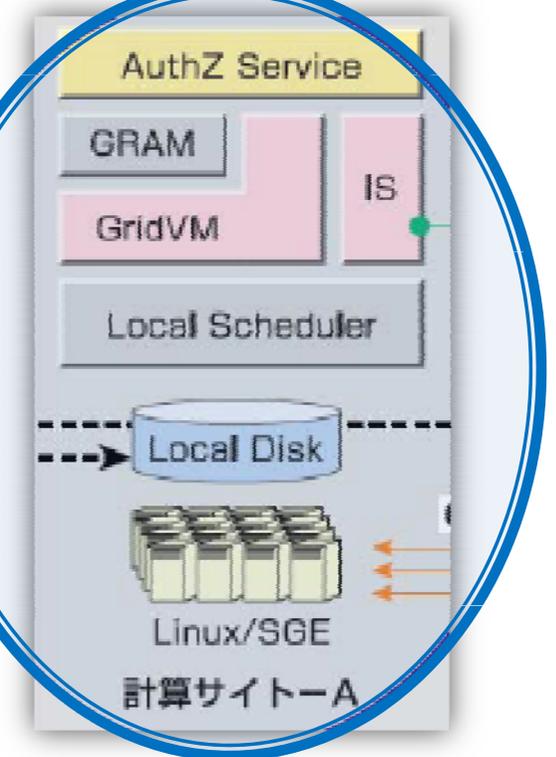
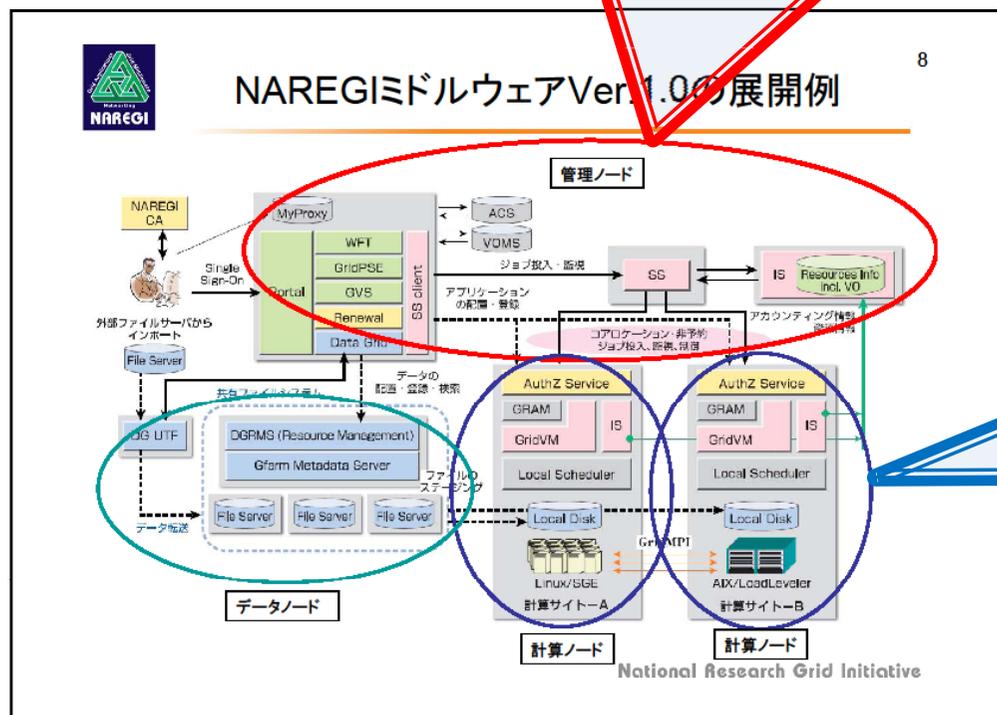
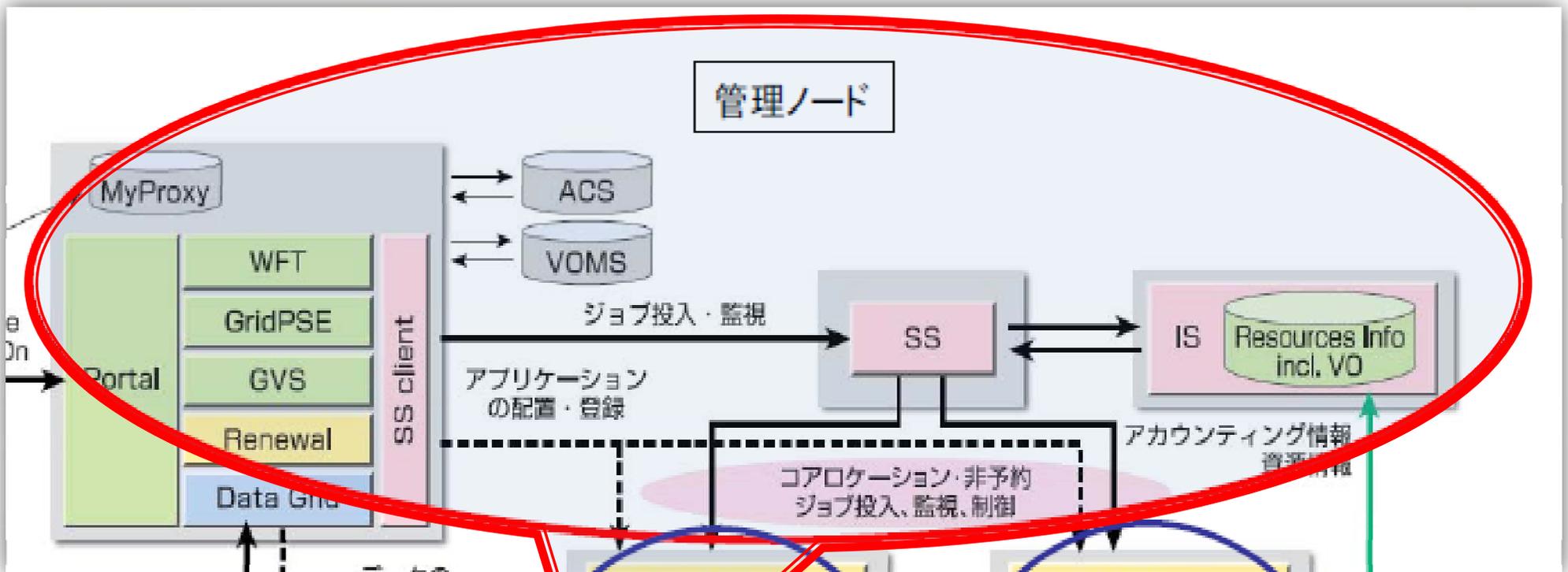
大規模実証実験を2008年3月に実施し、100テラフロップス級のグリッド環境を構築できることを確認  
(大阪大学、東京工業大学、九州大学、名古屋大学、分子科学研究所、グリッド研究開発推進拠点)



## 大規模実証実験環境の構成 (平成20年3月)



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



### 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

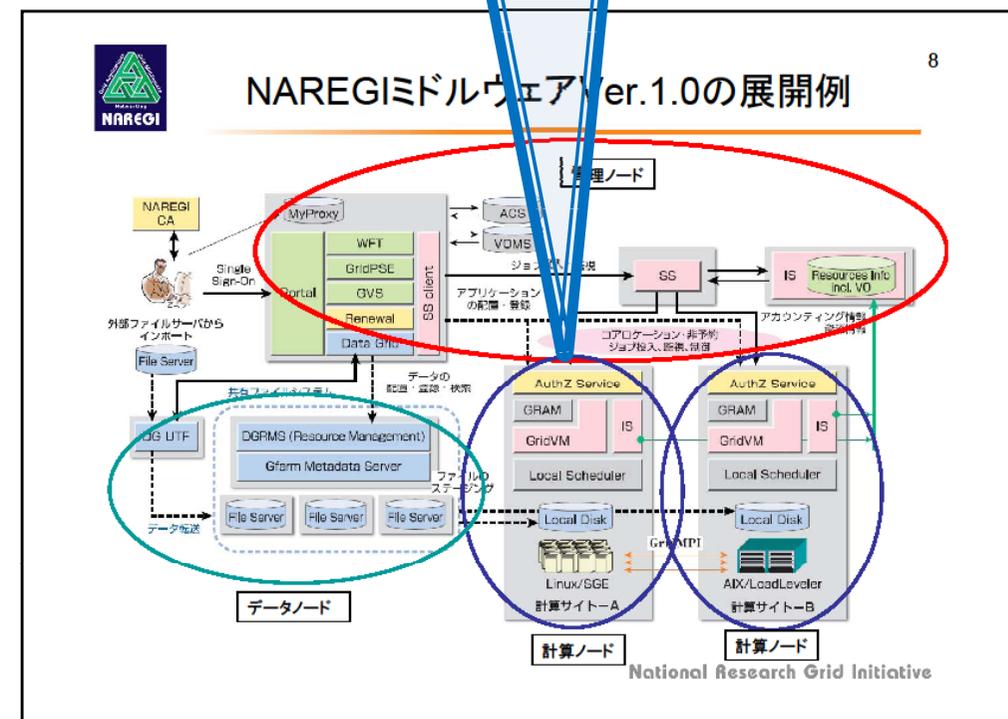
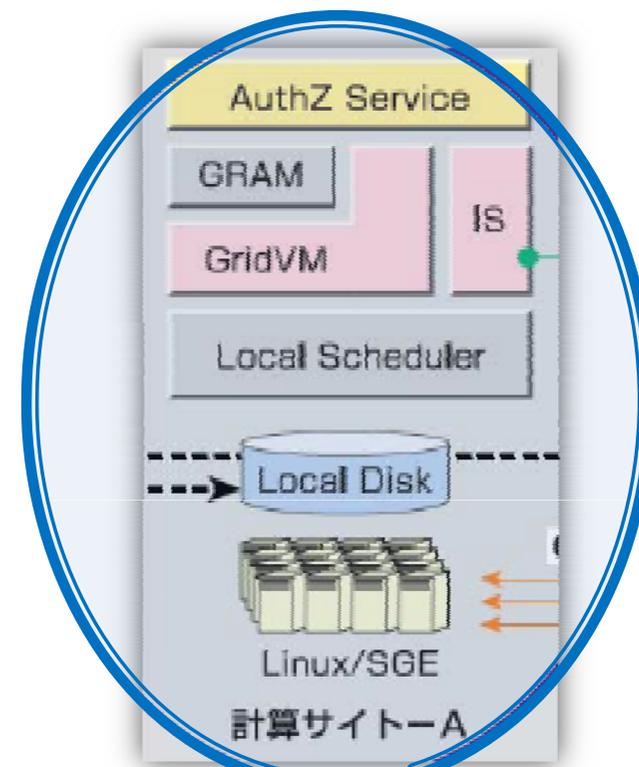
- 運用中のシステムを短期間に組み替えて資源提供

### 名古屋大学

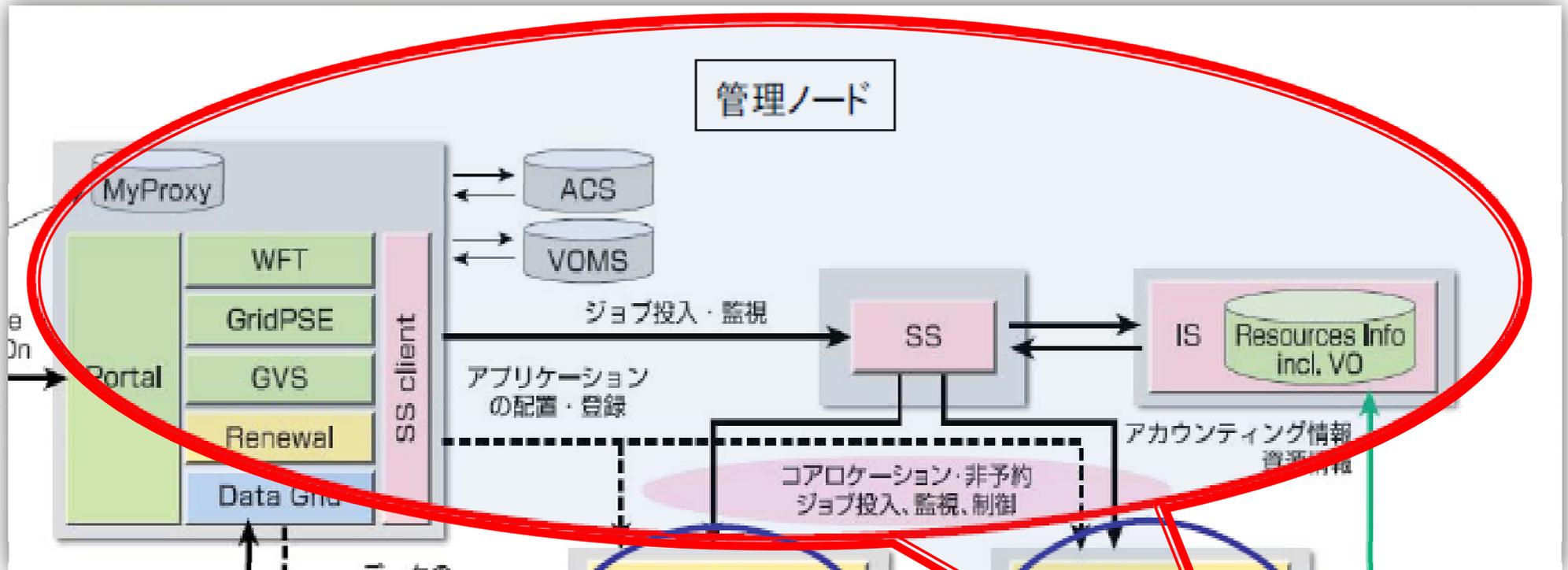
- NAREGIミドルウェア評価用のシステムを学内向け管理システムと共有しつつ連携に資源提供

### 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

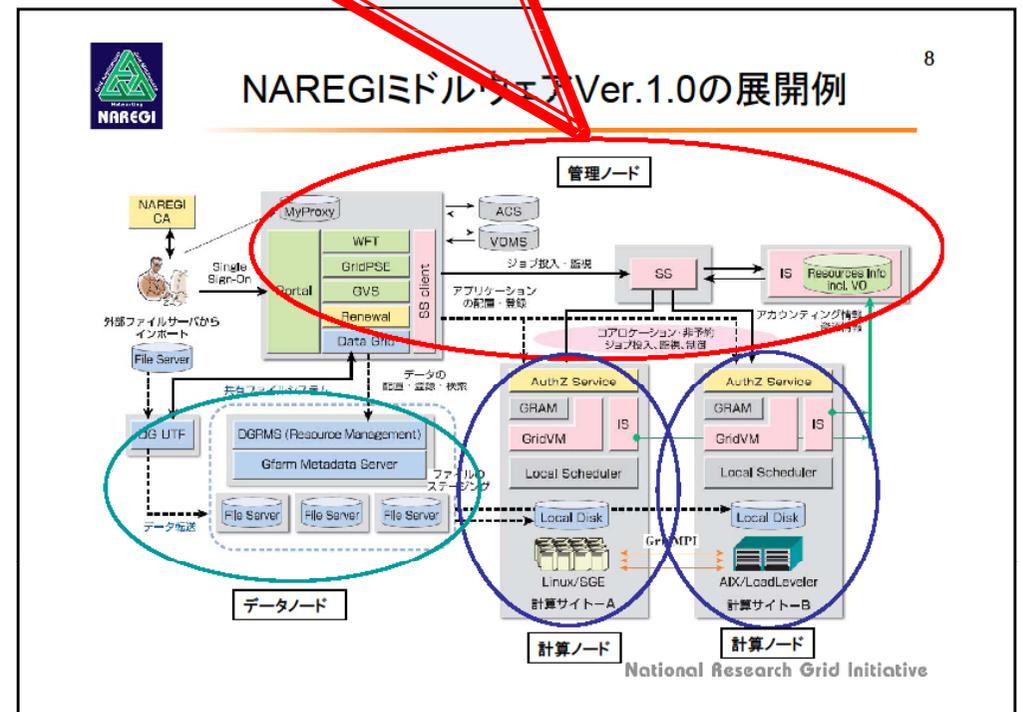


### SS: スーパースケジューラ

- 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

How SS Works:  
モニタツールによる監視

The screenshot displays a monitoring tool interface with three main sections:

- Core Images:** A list of timestamps and IDs, with the entry "2008/03/26 16:14:28.423970" highlighted in blue. A "Reload List" button is located below the list.
- SC:** A list of URLs, with the entry "https://gridvms3.hpc.cmc.osaka-u.ac.jp:9000/wsrf/services/gridvm/GridVMJobFactoryService" highlighted in blue.
- Reservation Map:** A large red rectangular area representing a reservation map. At the bottom, there are several grey rectangular blocks of varying sizes, indicating reservations. A vertical dashed line is present in the center of the map.

Below the SC list, there is an "Options" section with a "SOAP-Env" label and two radio buttons: "Show" (unselected) and "Hide" (selected).

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

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

Cybermedia Center, Osaka University  
<http://Grid-Portal.hpc.cmc.osaka-u.ac.jp/>

NAREGI Grid Portal

- ▶ Sign On
- Grid Tools

UserManagement Server

- ▶ Logout
- ▶ Proxy Certificate Registration
- ▶ Certificate Issue / Renewal

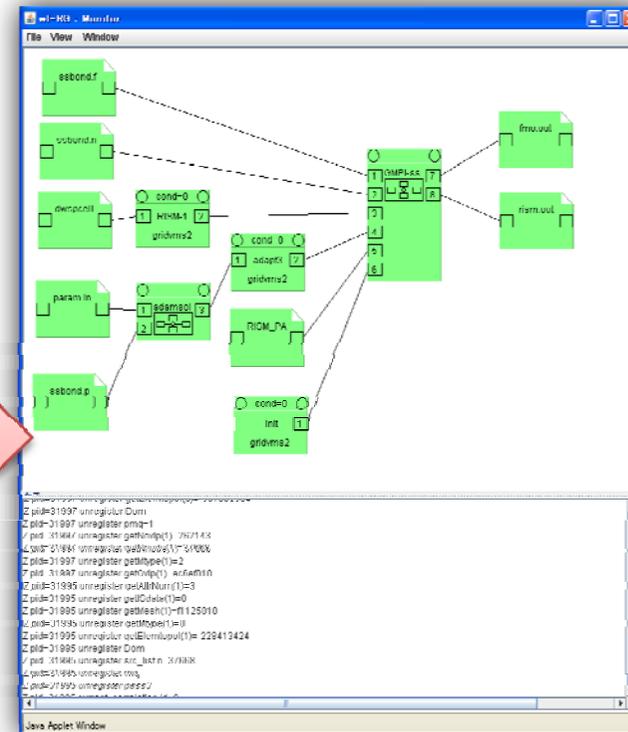
Proxy Certificate Registration

User Name: manabu  
 Certificate DN: C=JP, O=Osaka University, OU=Cybermedia Center, CN=manabu  
 Certificate Expiration: Tue Apr 1 2008 09:00:00 +0900

VO Name: RENKEI-Osaka  
 Role in VO: DefaultRole

Register Clear

Copyright © 2004-2007 National Institute of Informatics. All Rights Reserved.



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クス実験を再現

MD simulation

back light image

LIF

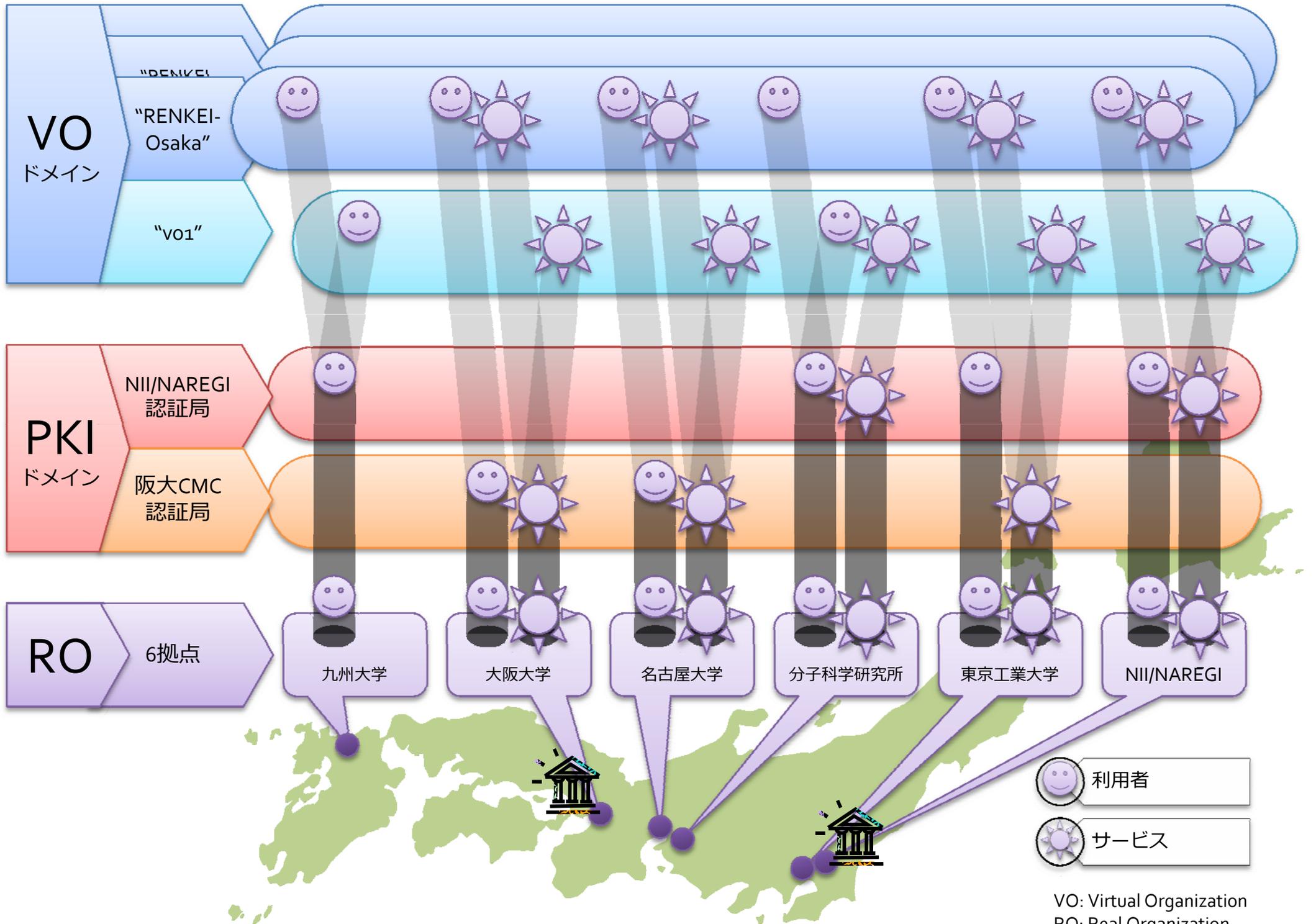
32nm以下の細線加工(次世代半導体製造)に必要な波長13.5nmのリソグラフィ光源の開発研究に寄与

GRID-MPI4 粒子のダイナミ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

MD simulation

back light image

LIF



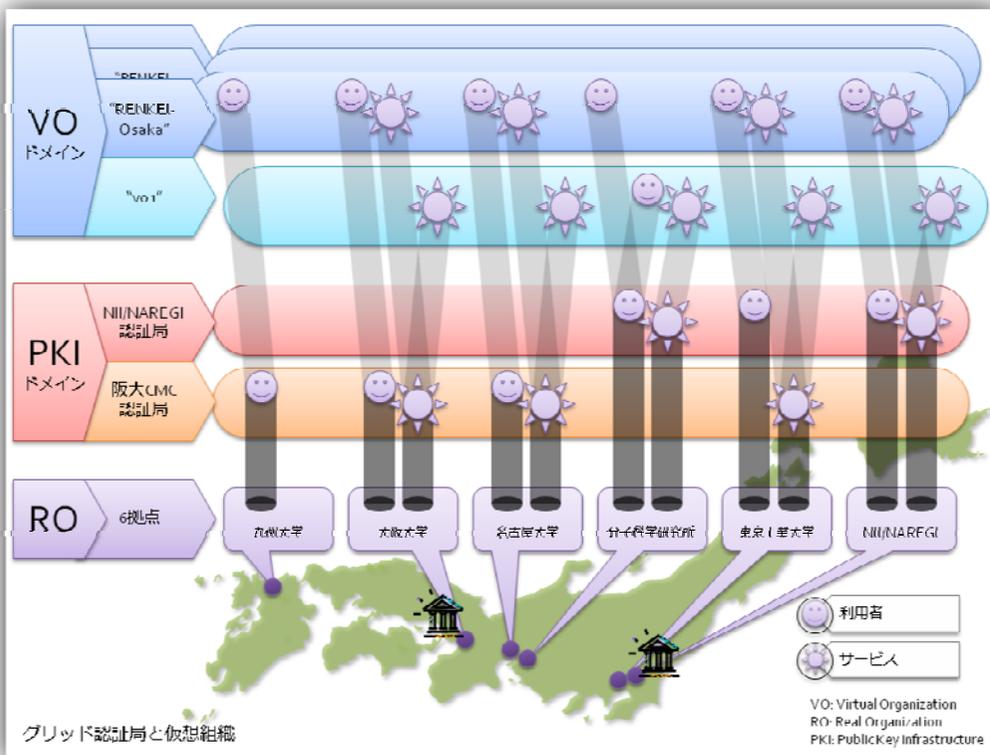
# “Registration Agency” 構想に向けて

## 今回のアカウント発行・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

## 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