

RENKEI
(REsource liNKage for E-scIence)
プロジェクトにおける
PoPの展開配備

東田 学
(大阪大学)

松岡 聡、友石 正彦、實本 英之、滝澤 真一郎
(東京工業大学)

ADVNET2008 ⇒ SACSIS2009

Linuxカーネルの TCPウィンドウサイズ自動最適化と 連動したバルクデータ転送 — Linux TCP Autotuning —

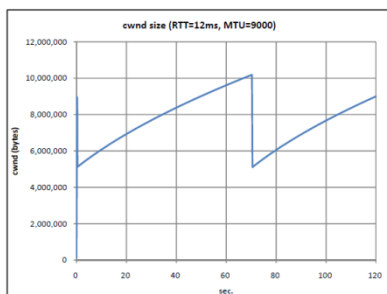
2008/07/15
Manabu Higashida (manabu@cmc.osaka-u.ac.jp)
Shinji Shimono (sshinji@nict.go.jp)

HPCユーザが知っておきたい TCP/IPの話 ～クラスター・グリッド環境の落とし穴～

産業技術総合研究所
情報技術研究部門
高野 了成

2009年5月29日 SACSIS2009チュートリアル

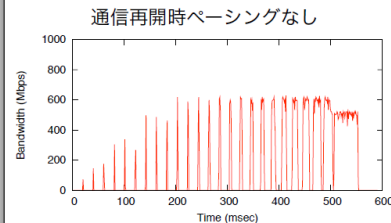
“Web100 Kernel Instruments” を 組み込んで調べてみた



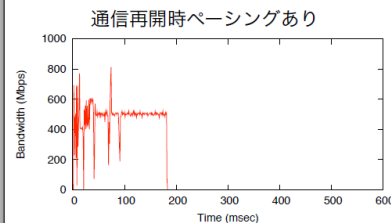
<http://www.web100.org/>

東田 学, 下條 真司, “LinuxカーネルのTCPウィンドウサイズ自動最適化と連動したバルクデータ転送”, ADVNET2008

通信再開時ペーシングの効果

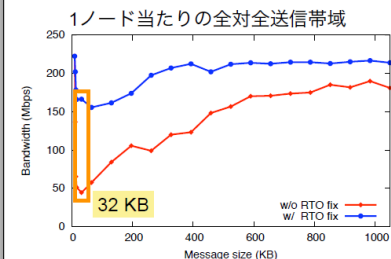


2秒ごとに10MBデータの連続送信の繰り返し



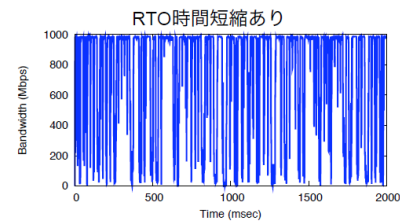
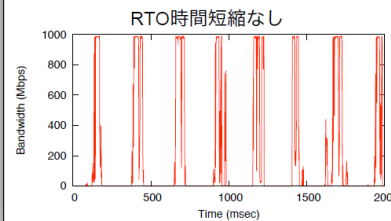
- スロースタートを省略し、アイドル前のcwndから再開
- さらに、バースト送信回避のため、通信再開時ペーシングを使用

RTO時間短縮の効果



$$RTO = SRTT + \max(G, 4 \times RTTVAR)$$

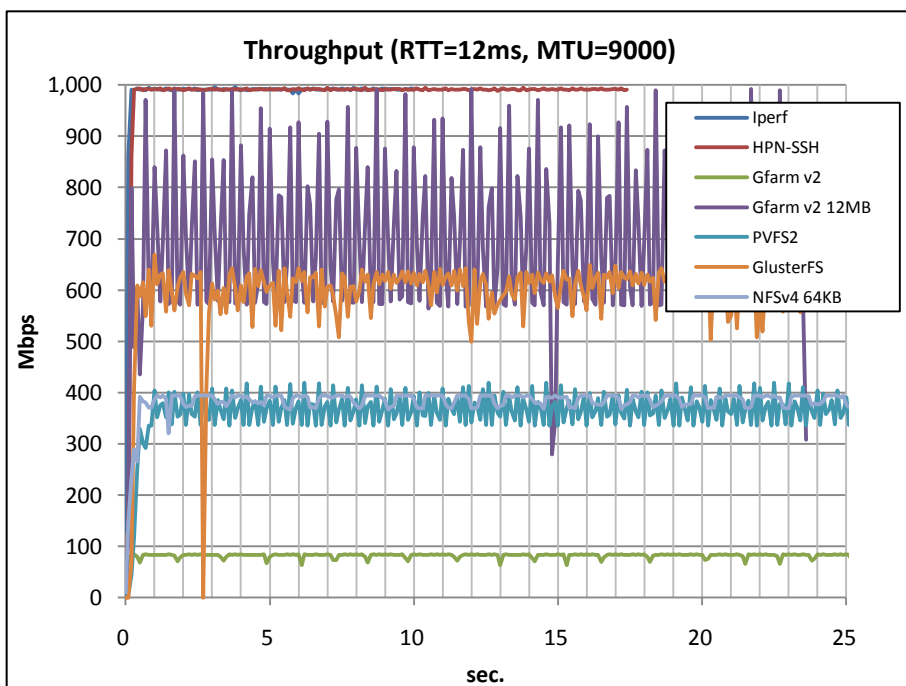
- 変更前: $G = 200$ ミリ秒
- 変更後: $G = 0$ ミリ秒
- (kernel 2.6.23以降) ip route コマンドにより設定変更可能



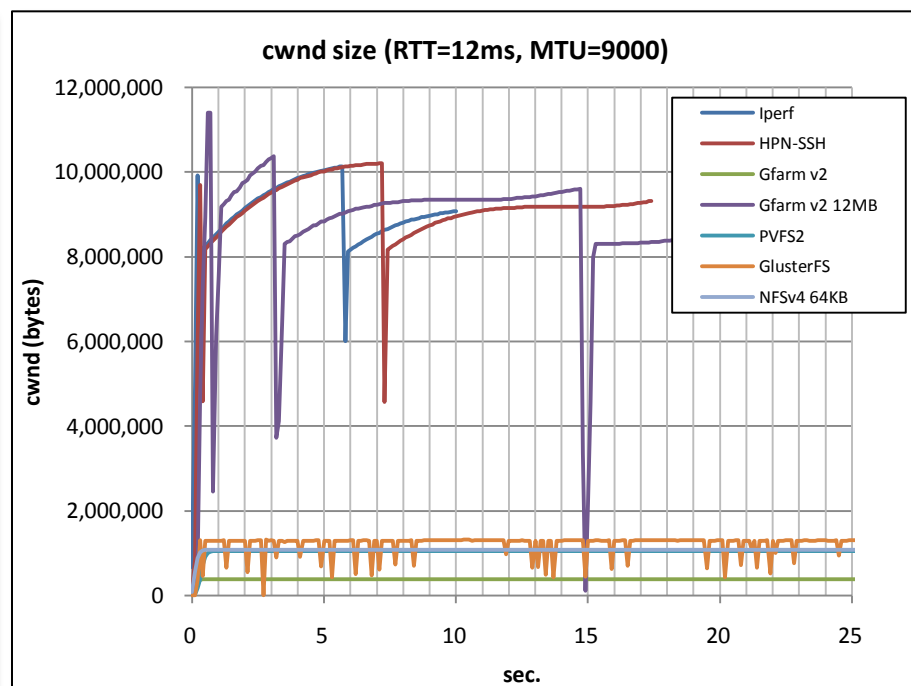
高野 了成, “HPCユーザが知っておきたいTCP/IPの話～クラスター・グリッド環境の落とし穴～”, SACSIS2009チュートリアル

広域分散ファイルシステムの 性能検証例

- ウィンドウサイズの変動に伴うスループットの変化を様々な競合システムと比較検証
 - スループットを追うだけではなく、TCPウィンドウサイズの変動などこれまで実地検証されてこなかったパラメータも観測可能

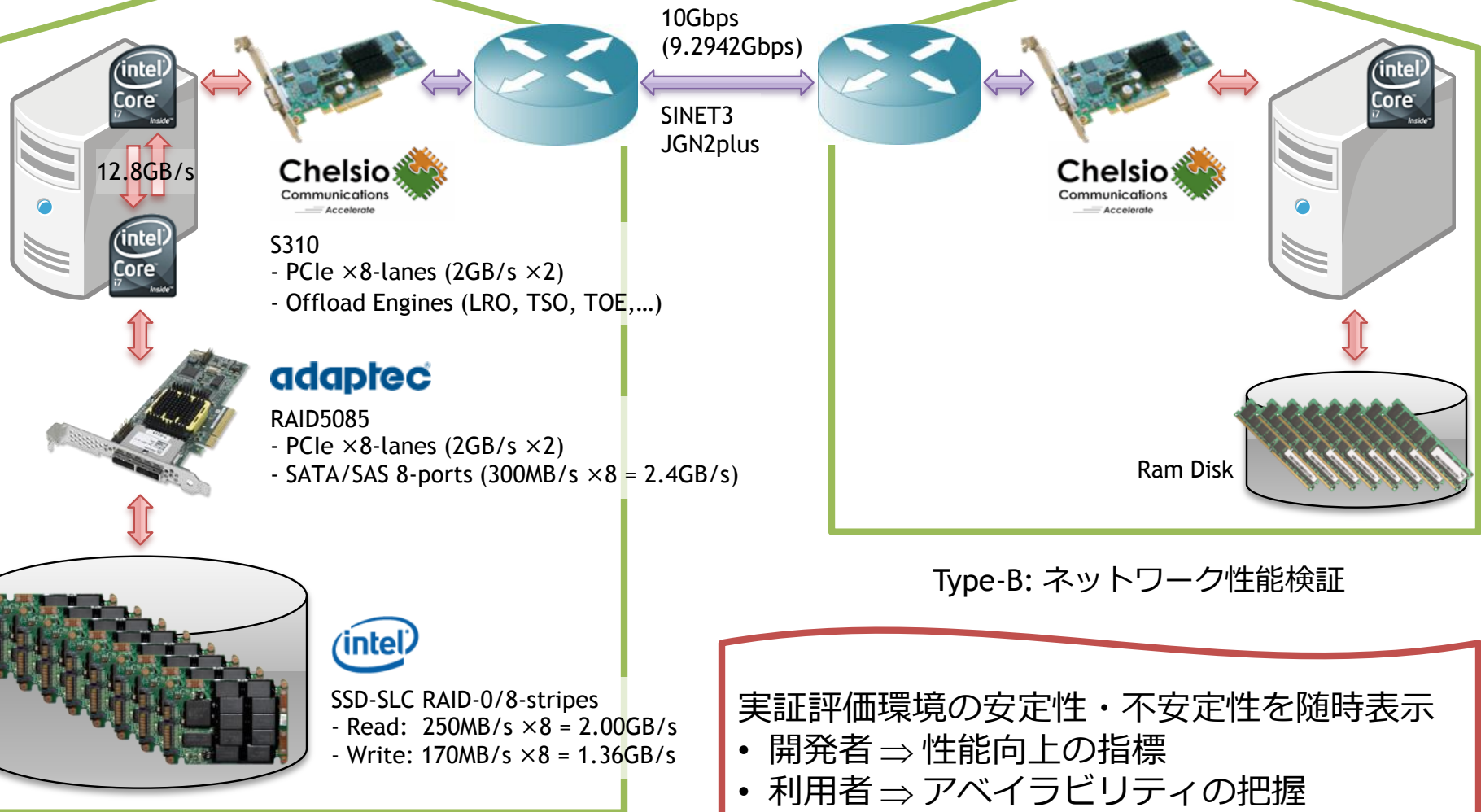


スループット



TCPウィンドウサイズ

1GB/sを維持可能な性能計測原器の配置



Type-A: ディスクI/Oを伴うネットワーク性能検証
(初年度に東工大・阪大に設置予定)

実証評価環境の安定性・不安定性を随時表示

- 開発者 ⇒ 性能向上の指標
- 利用者 ⇒ アベイラビリティの把握

H20成果: RENKEI-PoP (Point of Presence) の仕様策定

目的: 拠点間的高速データ転送をサポートするアプライアンスを定義

- 広域分散ファイルシステムによる利用推進のインセンティブ
 - 高速なローカルI/OとネットワークI/Oの両立した妥協のないサービス拠点
- NAREGIミドルウェアと仮想マシンを組み合わせたアプリケーション・ホスティング
 - グリッド認証基盤とアプリケーション実行基盤を平準化
 - 仮想マシンによる安定版・開発版の同時展開による利用と開発の並行的な促進

マルチプロトコル・低ハードウェアコスト・高性能・グリッドセキュリティ
他拠点連携研究のテストベットとして**安価**に利用可能



RENKEI PoP プロトタイプ

CPU	Core i7 965 Extreme (3.20 GHz)
Memory	12GB (DDR3 PC3-10600 , 2GB*6)
NIC	10GbE (with TCP/IP Offload Engine)
System Disk	500GB HDD
SSD RAID	256GB (RAID 0, Intel SLC 32GB SSD * 8)

iozoneによる入出力テスト
(blocksize = 256KB)

File size	Write	Re-write	Read	Re-read
16GB	996MB/s	827MB/s	908MB/s	773MB/s
32GB	953MB/s	681MB/s	767MB/s	772MB/s

そもそも

“RENKEI”

ってなに!?

RENKEI: REsource liNKage for E-science

- “RENKEI” is directly not post NAREGI-project
 - NAREGIの後継プロジェクトではない
 - NAREGIミドルウェアのオープンコミュニティによる開発プロジェクト “OpenNAREGI” (仮称) は別途企画中...

“RENKEI” is one of derivatives from NAREGI

次世代IT基盤構築のための研究開発:「e-サイエンス実現のためのシステム統合・連携ソフトウェアの研究開発」

研究コミュニティ形成のための 資源連携技術に関する研究の概要 (RENKEI* Project)

2009年6月19日

研究代表者

国立情報学研究所
リサーチグリッド研究開発センター
教授 三浦 謙一

*RENKEI: Resources Linkage for e-Science

背景と目的

■ e-サイエンス

- 計算機やデータ、ユーザ等からなる研究コミュニティを柔軟に形成することにより実現する新たな科学技術研究手法

■ e-サイエンスの実現

- 単一のグリッドによる研究基盤とコミュニティの形成からさらに発展の段階へ
- 運用形態の異なる資源間の連携が新たに必要。
 - 研究室レベルの資源と情報基盤センターレベルの資源の連携
 - 異種グリッドミドルウェア間の連携

目的

研究室の資源(LLS)、情報基盤センターの高性能資源群(NIS)、異種グリッドミドルウェア上で提供される計算、データ、データベース等を柔軟に共有・連携させ、仮想研究コミュニティを形成し、かつ運用するための基盤技術を確立する。

e-サイエンス実現のためのシステム統合・連携ソフトウェアの研究開発

計算アプリケーションユーザ

データベースユーザ

アプリケーション開発者

サブテーマ(1) 計算連携

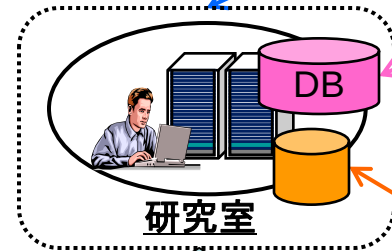
目的: 研究室レベル・情報基盤センターレベル間でのシームレスなジョブ実行技術に関する研究
H20年度計画:
・ワークフローシステムの検討, プロトタイプ設計および実装着手
・異種グリッド環境間の相互運用アーキテクチャの検討, プロトタイプ設計および実装着手
・アプリケーション共有方式検討, プロトタイプ設計

サブテーマ(3) データベース連携

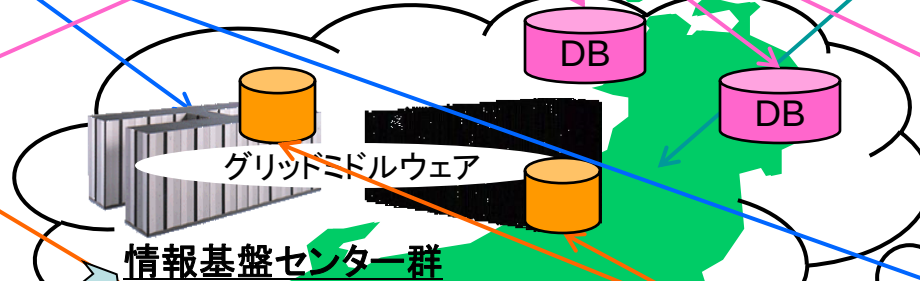
目的: 異種データベースの連携・統合, ユーザ認証情報の管理技術に関する研究
H20年度計画:
・データベース連携・統合ミドルウェアの設計, 実装方法の検討
・ユーザ認証情報管理ツールの設計, プロトタイプの実装

サブテーマ(4) アプリケーションインタフェース

目的: 異種グリッド環境下でのアプリケーション開発者向けインタフェースに関する研究
H20年度計画:
・要素技術およびユースケース調査
・設計着手



研究室



情報基盤センター群

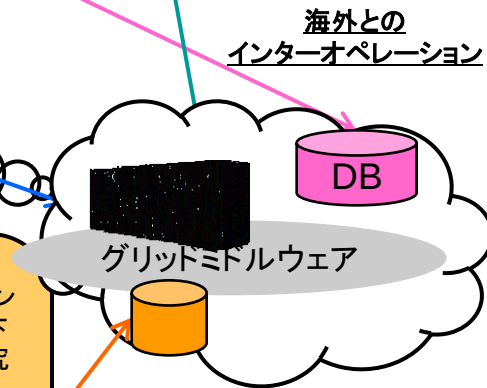
情報基盤センターユーザ

NII CSI委託事業

サブテーマ(5) 実証評価・ユーザ連携
目的: 実証評価基盤の構築, 情報基盤センターおよびエンドユーザと連携した実証評価
H20年度計画:
・事前実証評価環境の構築および計測評価技術の開発
・研究室レベル・情報基盤センターレベル間でのデータ共有技術に関する事前評価

サブテーマ(2) データ共有

目的: 研究室レベル・情報基盤センターレベル間, 異種グリッド環境下でのデータ共有技術に関する研究
H20年度計画:
・研究室レベル・情報基盤センターレベル間の分散ファイルシステムの設計, 基本機能の実装
・異種グリッドミドルウェア環境下のファイルカタログシステムの基本機能の開発



海外との
インターオペレーション

計算/データグリッド
アプリケーションユーザ



超高速

コンピュータ網

形成プロジェクト

の夢の跡

“100テラフロップス級”

- Good Experiences

- 2拠点連携

- 2007/08: 阪大-東工大

- 6拠点連携

- 2007/08: +九大、NII
 - 2007/12: +分子研、名大

- 10拠点連携

- FY2008: 北大、東北大、東大、名大、京大、阪大、九大
+東工大、筑波大

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!?



NAREGIミドルウェア実証評価体制

14

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

- 1) 2005年4月からα版で実証研究
- 2) 2006年5月からβ版で実証研究
- 3) 2007年6月からβ版-IIで実証研究
- ①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。
- ②22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。
- ③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。
- ④ナノサイエンスVOの構築
- ⑤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

- 1) 2005年11月からα版を限定配布により評価
- 2) 2006年5月からβ版を公開配布により評価
- 3) 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
- 4) 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
- ①認証ポリシーが異なるセンター間の相互連携
- ②VO形成、相互の資源予約管理
- ③運用関係の評価

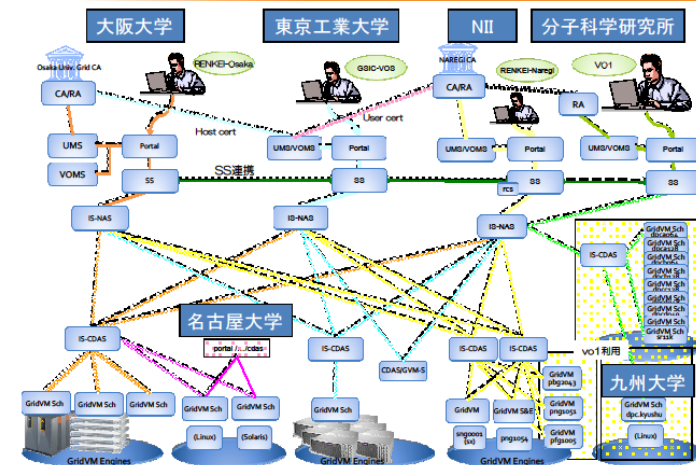
KEK-NAOJ解析連携環境

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)

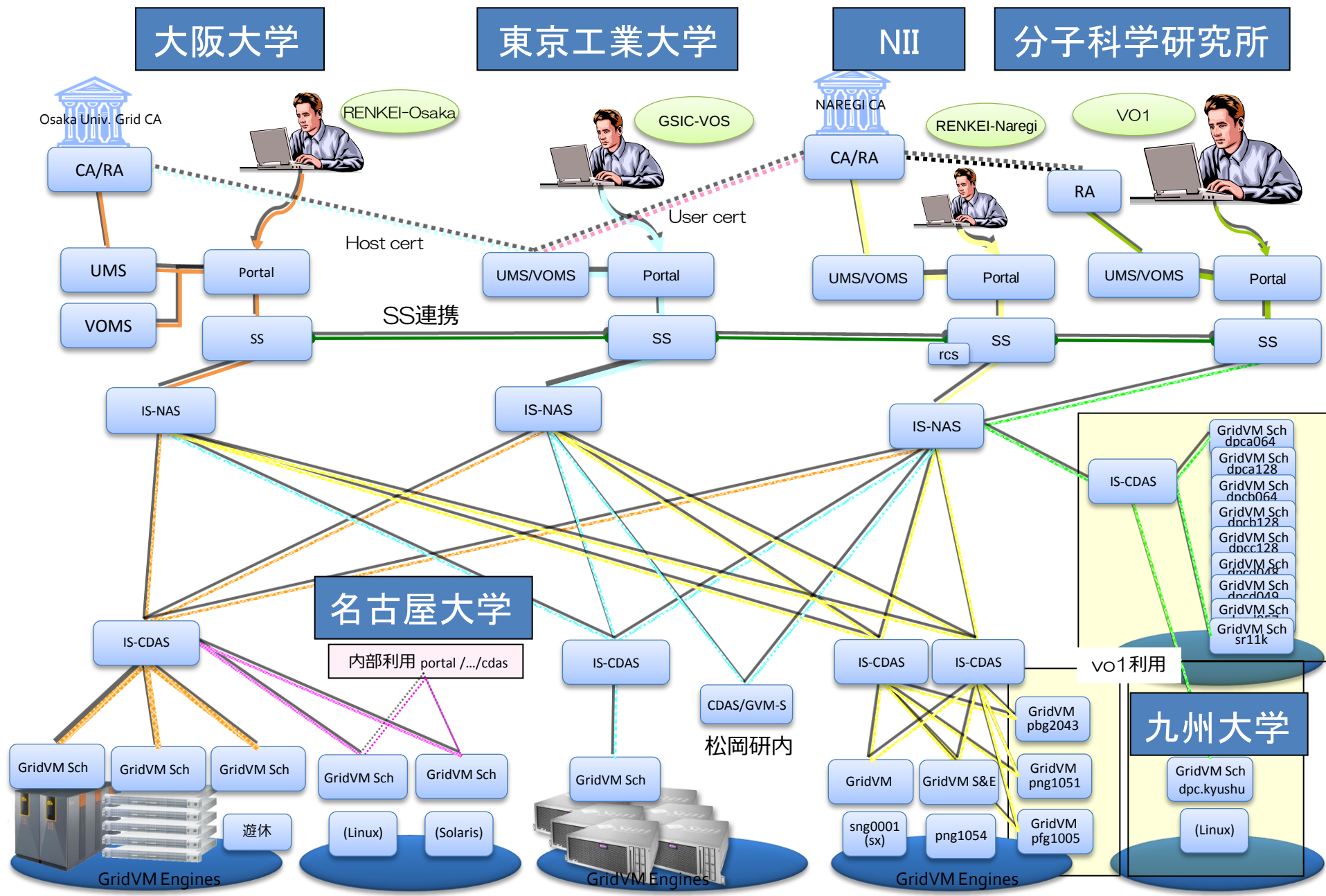


National Research Grid Initiative

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

"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

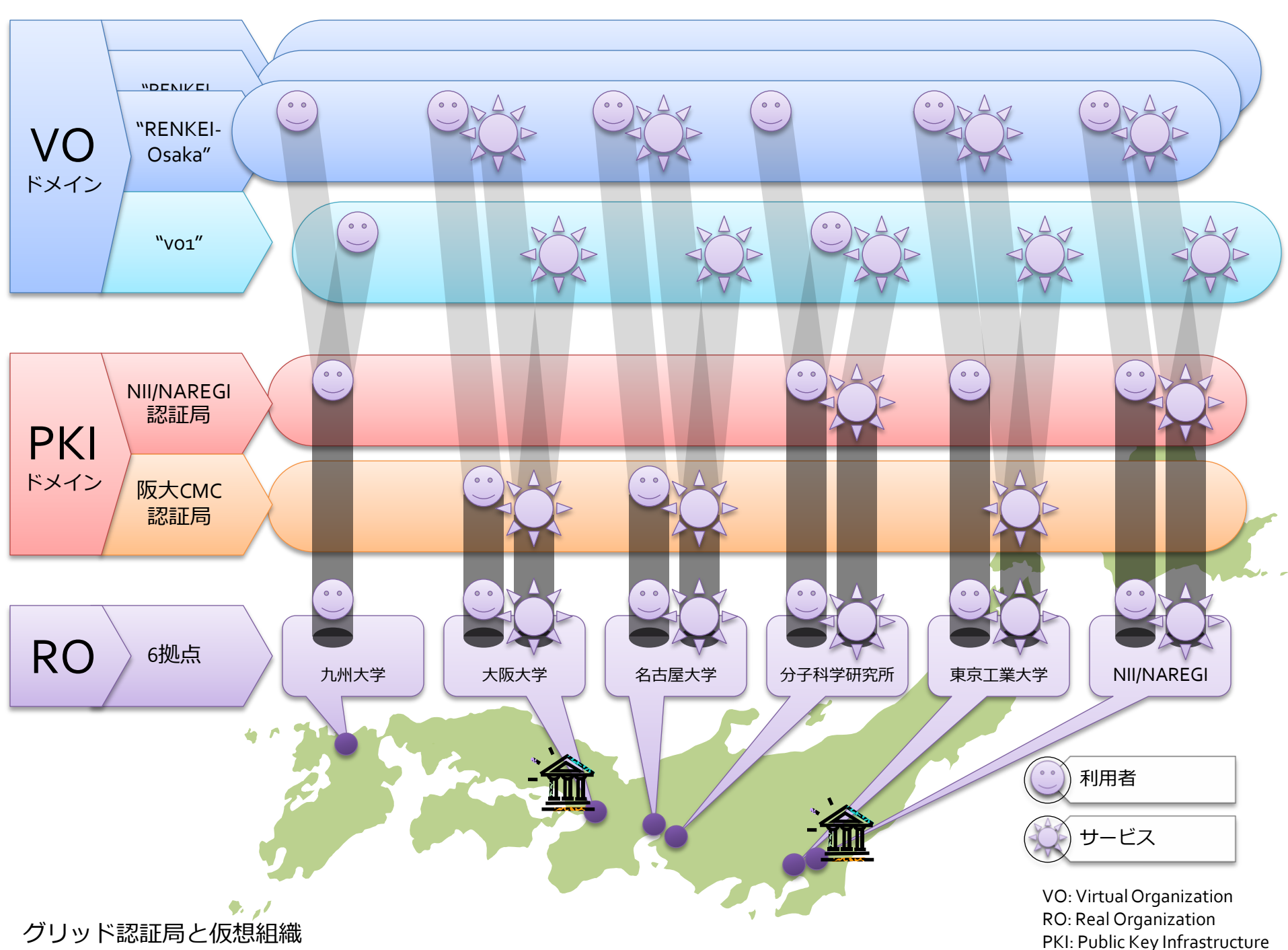
	Phase-1 / Phase-2	SINET3 接続	グリッド 認証局	ポータル	SS	IS	
						NAS	CDAS
大阪大学	2007年8月	10Gbps	○	○	○	○	○
東京工業大学		4Gbps	—	○	○	○	○
九州大学		—	—	—	—	—	—
NII/NAREGI		1Gbps	○	○	◎※ ¹	○	○
分子科学研究所	2008年1月	1Gbps	—	○	○	—	○
名古屋大学	2008年3月	1Gbps	—	△※ ²	△※ ²	△※ ²	○
			2ヶ所	4ヶ所	4ヶ所	3ヶ所	5ヶ所

※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	dpcao64.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



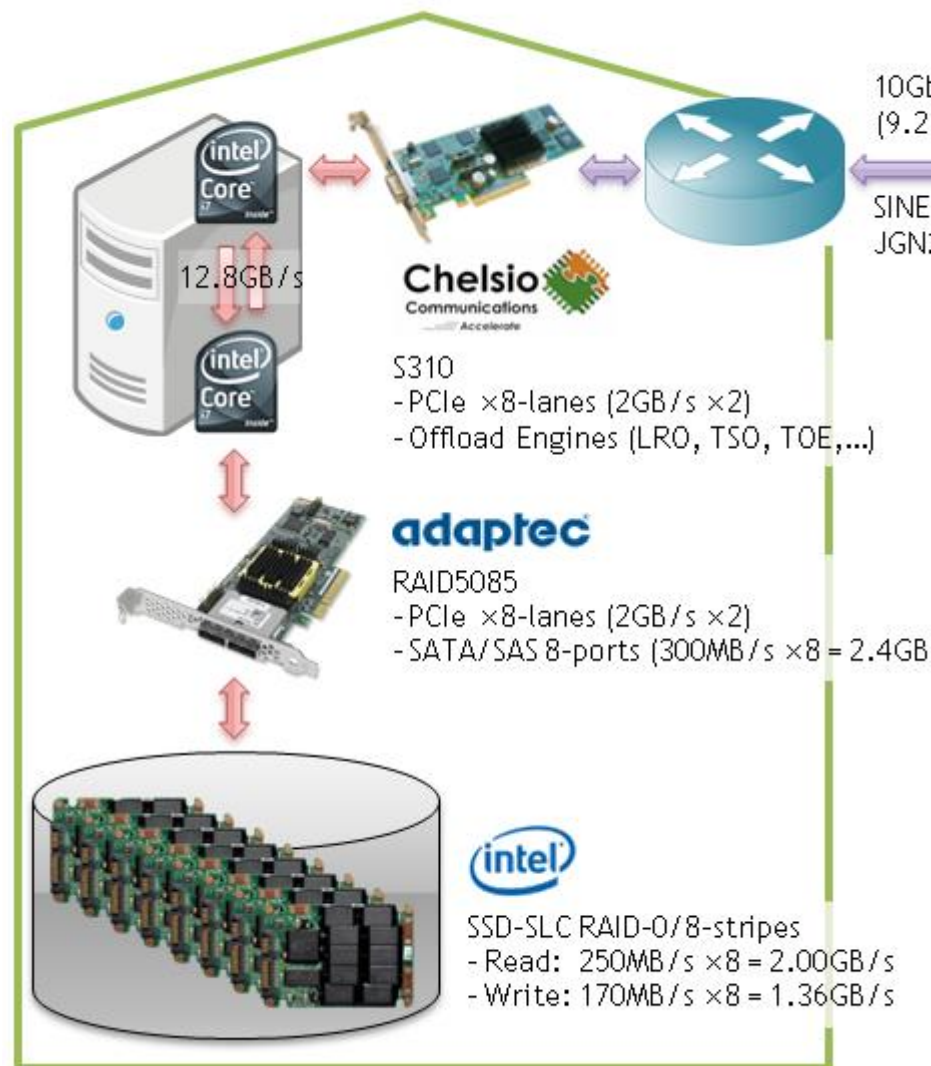
ハリボテでもなんとか動くミドルウェア

- “We reject kings, presidents and voting; we believe in rough consensus and running code.” — David Clark (1992)
 - 意訳) プロジェクトが終わって、やっかいな人たちが去って、まがりなりにも動くコードが残ったってチャンス!?

 OpenNAREGI (仮称) を企画中

経験を詰め込んだ「NAREGI箱」

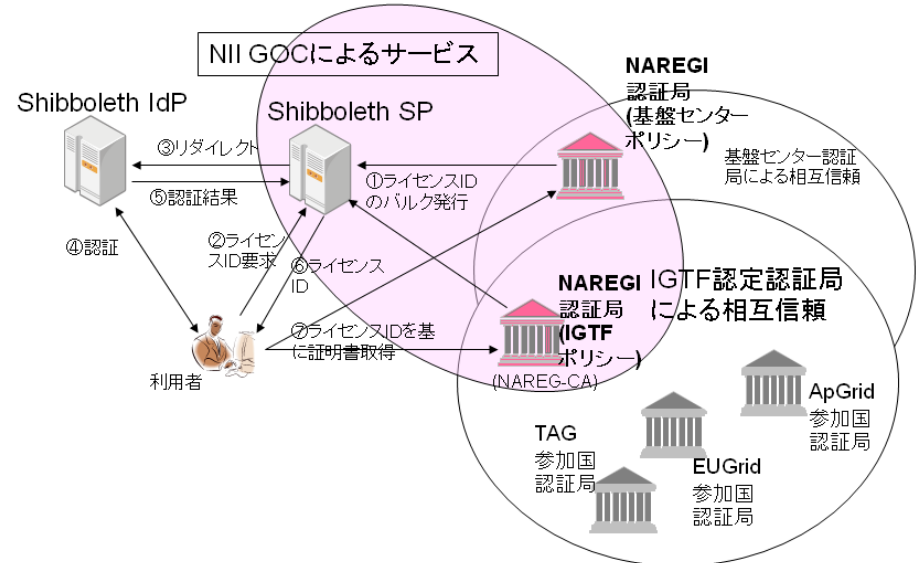
- “Point of Presence”
 - インターネットの黎明期:
 - 回線は来た、あとは箱さえあれば・・・「きっと加藤さんならなんとかしてくれる」作戦再び!?
- そんなの送りつけられて傍迷惑じゃないの!?
 - 認証基盤の普及
 - 自律的なアップデート



Last One Mile: グリッド登録機関の事業化

- UPKIの成果を取り入れたNAREGIミドルウェアの拡張とNII+情報基盤センターによる事業化
 - MICSプロファイルに対応したグリッド認証局の運用
 - 運用局規定 (CP/CPS) の策定
 - NAREGI-CAのShibboleth対応と
既存認証局のアップデート
 - NII/NAREGI認証局
 - 阪大CMCグリッド認証局
 - NAREGIミドルウェアのShibboleth SP対応

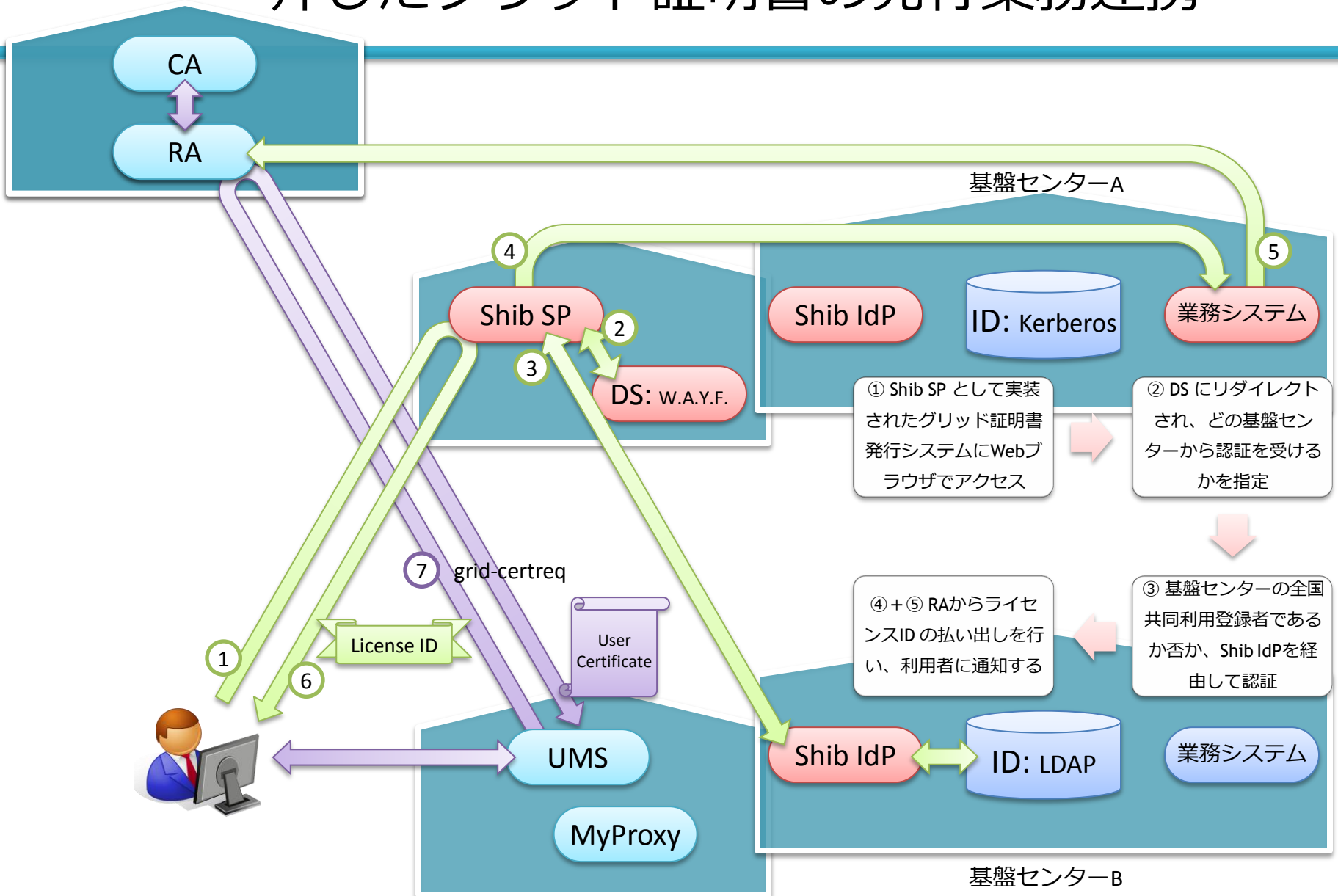
パターン2: NII-GOCでの活用



NII GOCによるポリシー毎の証明書一括発行サービスを実施

既存の情報基盤センターの
共同利用窓口がそのまま
グリッド登録局として機能

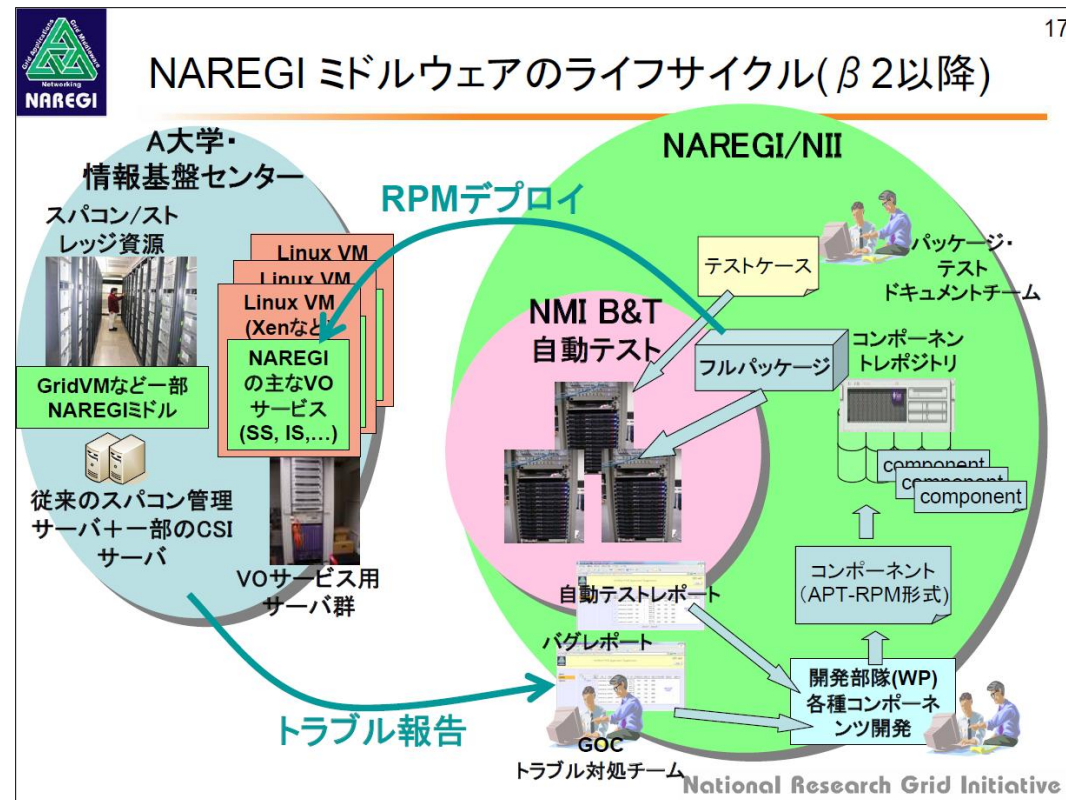
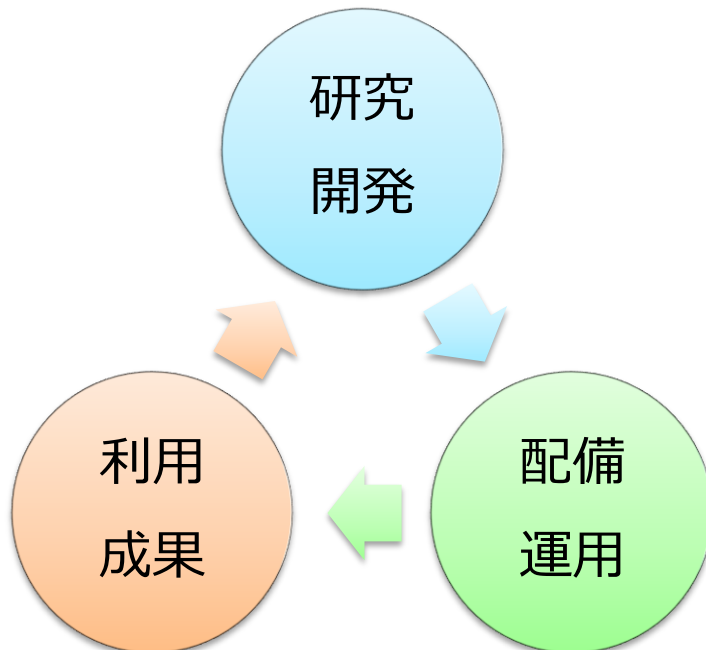
MICSプロファイルを満たすShibboleth IdP/SPを 介したグリッド証明書の発行業務連携



Last One-more Mile: NAREGIミドルウェアの継続的な維持

- オープンソース型の開発サイクルの立ち上げと
統合試験環境との一体化によるライフサイクルの持続

NAREGIミドルウェアの
エコシステムを確立



サブテーマ (5)

実証評価・コミュニティサポート

1. 研究コミュニティ構築のための実証基盤としてのPoP (Point of Presence) の展開

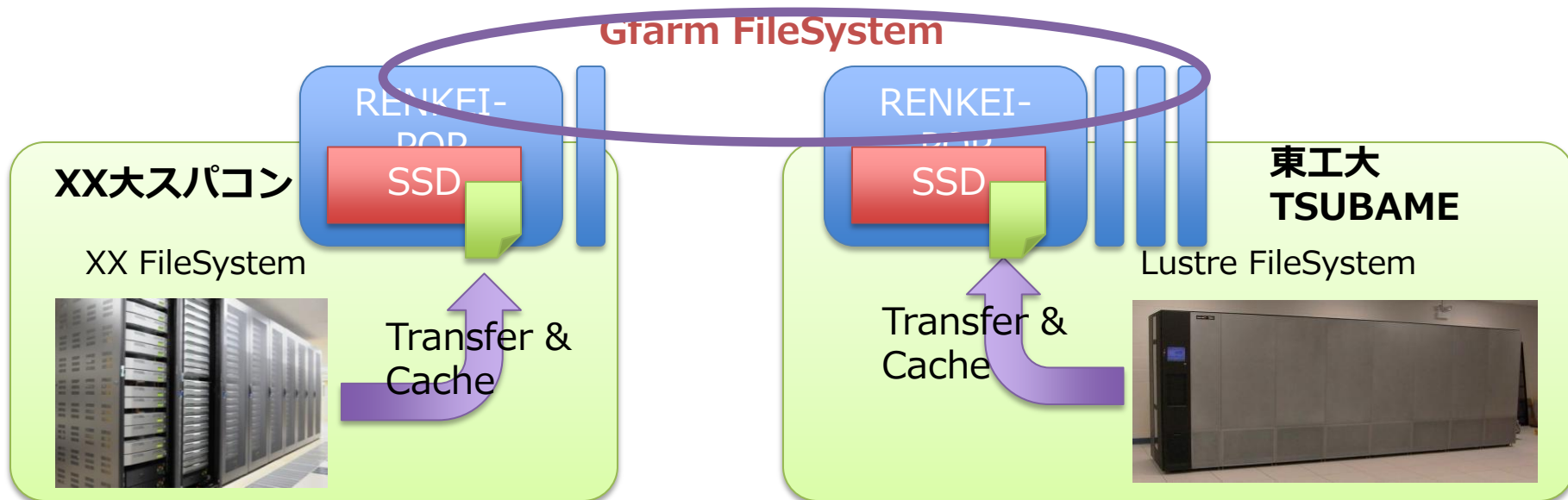
- 仮想マシンによるアプリケーション・ホスティング
 - 各サブテーマの成果物をNLS, NISと連携し広域展開するための橋頭堡
 - 「安定版」と「開発版」など複数の世代を管理
 - » アプリケーション・サービスを維持しながら開発を加速
- PoP間では、1GByte/s以上のデータ転送を維持
 - ローカルI/Oから広域ネットワークまでボトルネックを解消
 - 様々な環境を再現し検証するための広域ネットワーク・シミュレータの役割

2. 研究コミュニティの構築と運営

- 次世代スパコン(NLS)、各研究機関スパコン(NIS)、研究室PC/Cluster (LLS) の連携
 - 各拠点や研究室の持つマンパワー・計算能力・データ・アプリケーションの柔軟な融合を目指す

H21計画: RENKEI-POPによるNISストレージ統合

- RENKEI-POPを介した複数NIS間高速データ共有サービスの実現
 - 高速データ転送と広域分散FSを用いた複数NIS資源の平準化
 - NIS上のデータをRENKEI-POP経由でアクセス
 - RENKEI-POP間は10Gbps閉路網により接続
(ディスク帯域 ≡ ネットワーク帯域)



H21計画: RENKEI-POP上の仮想ホスティング環境

- ネットワーク構成を含めた柔軟なアプリケーション、ミドルウェア実証環境の実現
 - 各種設定のパッケージング、デプロイの自動化

様々なネットワーク構成の実現

- 10Gbps閉路網を直接用いたベスト環境
- 高遅延、低バンド幅環境

実証環境の世代管理

- ユーザが望む環境を実現
- 開発版、安定版の使い分け

仮想マシンを用いたホスティング

Eucalyptus、Rocks Xen Roll
などを評価、検討

ユーザによるネットワーク構成

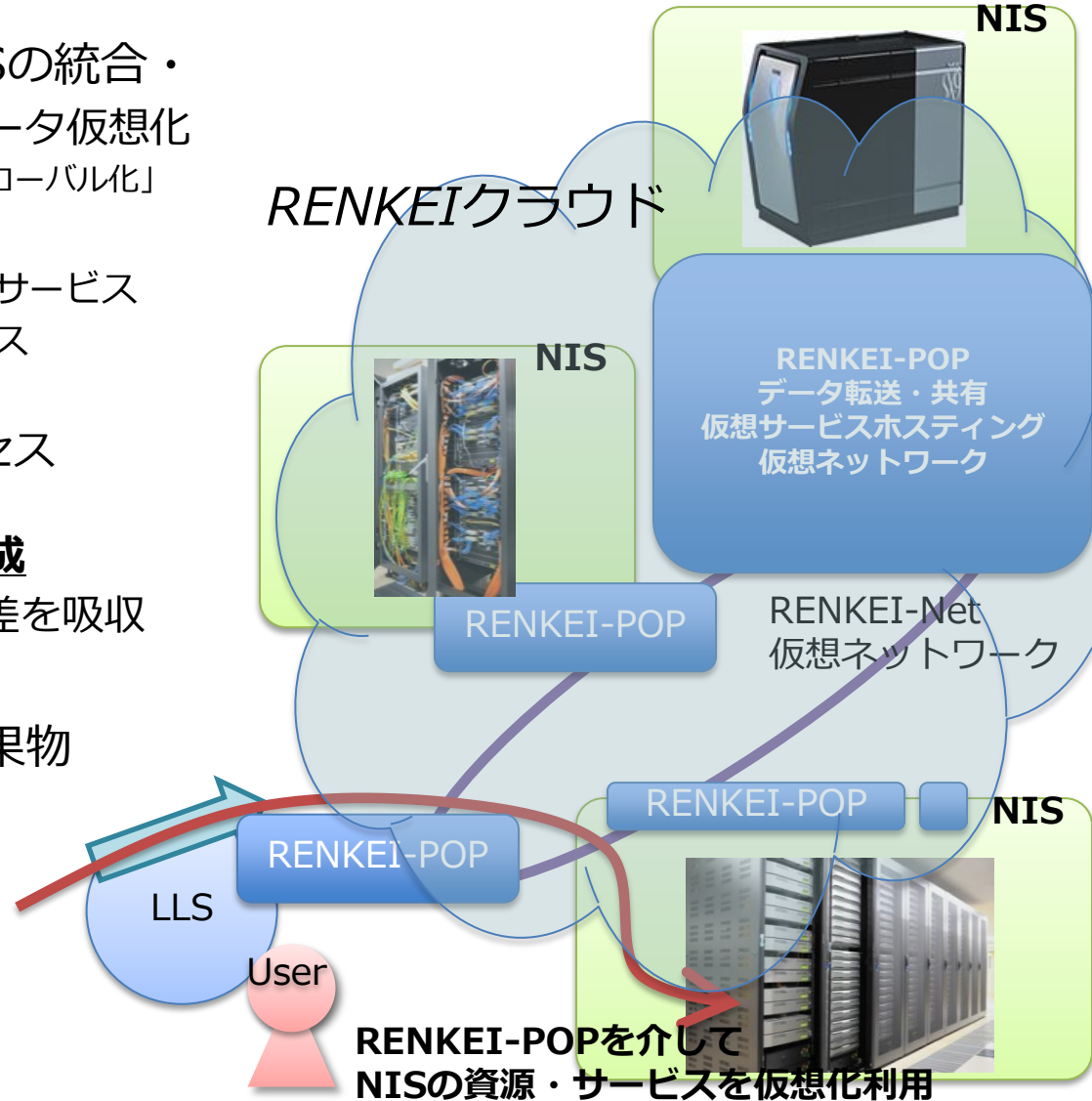
ネットワークエミュレータ、
各種ライブラリを提供し、
ユーザが設定・取捨選択

RENKEI-POPが作る将来の E-Scienceインフラ — RENKEIクラウド

- RENKEI-POPによるNIS、NLSの統合・
 - データ共有、データ転送・データ仮想化
 - NISのローカルストレージの「グローバル化」
 - 仮想サービスホスティング
 - 各種基礎グリッド・クラウドサービス
 - 各種アプリケーションサービス
 - 異なるバージョンの両立
 - LLSからのシームレスなアクセス
 - 仮想ネットワーク

仮想研究コミュニティクラウド形成

- RENKEI-POPによるNIS間の差を吸収
- サブテーマ（1）～（4）の成果物
評価環境として運用・提供し、
有用性を検証



<http://www.e-sciren.org/>

プロジェクトロゴは三浦先生が鋭意作成中・・・