

広域センサネットワークと オーバレイネットワークに関する ワークショップ

言いだしっぺ: 首藤, 松浦, 吉田 (50音順)



発端と意図

- SIProp 勉強会 (主宰 今村さん) Skype チャット
 - Y: RDFのP2P検索の話や、時間重要度コンテングのP2P検索の話、indexの配置の話とテーマは面白いです。
 - S: RDF の P2P 検索、産総研で、OW公開前に OW 使ってやってくれた研究員が いたなあ
 - S: 考えてみれば、RPCみたいな機能とか、けっこう**インテリジェン**トなオーバレイを組むための機能だった。
 - S: オーバレイ上での処理って、なんか足回りを作りたいんですが DHT 使ってデータ再利用とか (put/get)時 RPC とか それこそ MapReduce みたいな処理とか シーズ側からの夢はいろいろあります。
 - Y: こっちは、データフュージョン層の部品提供とか、話してました。でも、その前にやることがいっぱい。 大量に存在するセンサー情報の間引きとか、特徴素解析とか、履歴からの推定とか、です。コンテキストアウェアな話が多いです。
 - S: 何にせよ、ほんとの需要と接触しないと始まらないですねー
 - S: 純粋研究職だったら、Live E! の人にミーティングをお願い、かなあ 別に、今でもやりゃいいんだけど。松浦さんとこに遊びに行きゃいいだけだ。
 - Y: よびだしゃいい。 <あやや</p>
 - Y:でも、首藤さん一人でなく一緒にやりましょうよ
 - S: Live E! のオーバレイへの展開をまじめに考える会



発端と意図

- ・メール
 - To: 松浦さん, 吉田さん

From: 首藤

Subject: Live E! のオーバレイへの展開 (仮) ワークショップ

Seoul A

SOUTH

松浦さん、吉田さん、首藤です。

掲題のような勉強会をやりたいなあ、と考えてます。

僕の動機は、こんなところ↓です。 吉田さんも似た感じだと思います。

- あわよくば OW ベースの仕掛けをほんとに運用する。
- (Live E! から) オーバレイへの要件を深く知る。 オーバレイを発展させるべき方向、新世代ネットワークの要件を考 える材料を手に入れる。

じゃんじゃん議論できるように、人数は多すぎず、 時間的には1日のワークショップかな、と想像してます。

足まわり (オーバレイ) と
 応用 (センサネットワーク) が
 お互いをより深く知って、ますますの発展を (ry



Thanks!

- 奈良先端の皆様
 - ローカルアレンジメント
 - 会場提供,送迎,宿泊・お弁当の手配,...
- 松浦さん
 - 運営のためのウェブシステムや ML の用意
 - スケジュール調整,プログラム編成,...
- 皆樣
 - 御発表の急なお願い
 - 遠方からの御足労
- NICT, 皆様の所属組織
 - 後援, 旅費などの支援

ネタがある限り、 どんどんやりましょう!

ネタを作るべく、どんどんいい仕事しましょう!



PlanetLab の実際

首藤一幸



テストベッドの必然性

- なぜオーバレイ?
 - スケーラビリティ
 - 頑健性
 - 管理主体の分散
 - ファイル共有ソフト, グループウェア, …
 - アンダーレイの効率的な利用
 - オーバレイルーティング, ALM, ...
- それなりの規模で実験しないと!
 - できれば実ネットワーク上で。
- 要 大規模な実験を行う手段
 - シミュレータ, エミュレータ
 - 大規模テストベッド → 本題へ



テストベッド

- 大きなクラスタ
 - 北陸先端 StarBED
 - PC 680 台
 - 2002年設置の 512台は 1 GHz Pentium III, RAM 512 MB
 - Internet につながっている。持ち込み PC もつなげる。
 - NICT北陸RCの活動
 - 無線エミュレータ QOMET, ネットワーク環境(?)エミュレータ Rune, Internet全ASシミュレーション, ...

NORTH

Seoul A

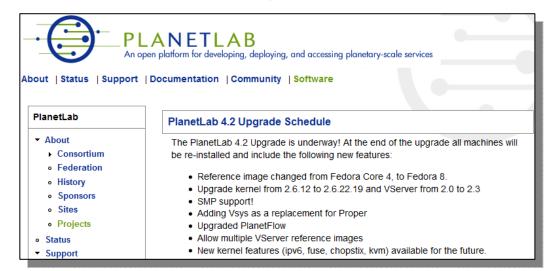
KOREA

- 関係者経由でお願いすると、使わせてもらえる。
- 東工大 TSUBAME
 - スパコン。
 - 2.4 GHz Opteron (dual core) 8プロセッサ × 639 台 + ...
 - 産業利用を募集してる。文科省予算 → 企業側の支払いはなし。P2Pアプリの実験もお誘い頂いた。
- Emulab
 - ネットワーク実験とその管理を行うためのソフトウェア。Internet や無線ネットワークのエミュレーションが可能。
- 広域ネットワーク
 - JGN2plus とか
 - ◆ 全国規模の高速ネットワーク。アクセスポイントまでの線は参加組織側が準備。
 - 今年度、PlanetLab のノードを配置しようとしている。
- 広域に分散したマシン群
 - PlanetLab



PlanetLab

- 地球規模テストベッド
- 800台以上が登録されている。
 - 実際に使えるのは 400台強? ← 2007年 1月の首藤実績。
- 研究プロジェクトごとに Linux の仮想マシンが払い出される。
 - 仮想マシン: sliver, その集合: slice
 - VM と言っても、Linux-VServer による chroot 的な仕掛けなので、軽量。
- 貢献すれば使える。
 - 大学・非営利の研究所なら、PC 2台を供出すれば、使える。
 - Industrial が使うためには \$25,000 / 年~。その場合 slice は 2つまで。



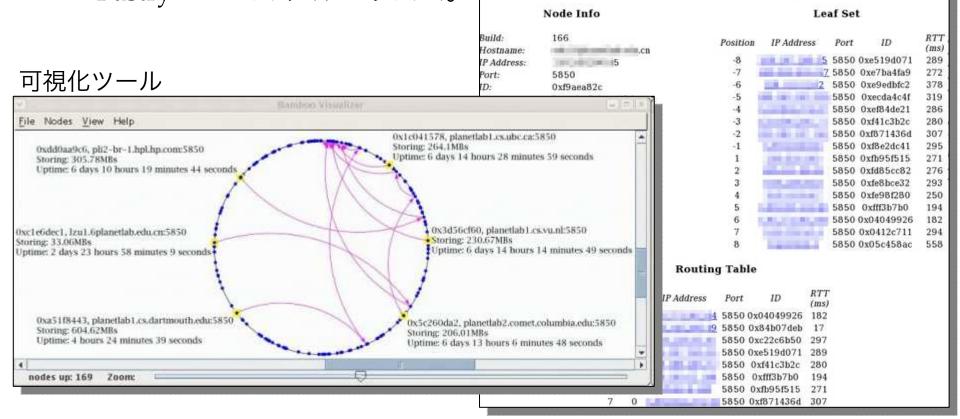


AGUGASAR

Seoul A

OpenDHT

- 誰でもアクセスできる DHT サービス。
 - XML-RPC, Sun (ONC) RPC で put, get できる。
- Bamboo DHT を使って運用されている。
 - Pastry ベースのアルゴリズム。



SHOPPER

Pyongyang

Bamboo Node Status Page

Seoul A

Coral CDN

- PlanetLab 上に deploy された CDN。
- 使い方: URL の末尾に .nyud.net を付ける。

"You make it fun; we'll make it run"



The Coral Content Distribution Network

Home .. Project home page

LAGUSASAR

Seoul A

Overview .. Brief overview and news

Usage .. Coral Wiki and FAQ

Lists .. Mailing Lists

Pubs .. Publications and people

Download • Plugins and source code

Illuminati •• Network measurement project

OASIS -- OASIS anycast service

2004年 8月 30日の記事

Our Goal

INTERNET

Watch

記事検索

検索

最新ニュース

【 2008/07/01 】

「GyaO」刷新、トッ ブページIこMS 「Silverlight」採用 [19:19] P2P技術で"25ゃんねる効果"を軽減できる無料CDNが正式公開

分散P2P技術を使ってWebサーバーへの極端な負荷を軽減できる無料のコンテンツ配信ネットワーク(CDN)をニューヨーク大学が開発し、30日に正式公開した。無料で利用でき、日本語サイトでも問題なく使用可能だ。

このサービスは「Coral:The New York University Distribution Network」と名付けられている。ニューヨーク大学コンピュータサイエンス学部がP2P技術に関する研究プロジェクトとして発足したもので、全米科学財団より支援を受けて2004年3月から公開テストが行なわれてきた。

nd that the website is temporarily off-line because
Does your network have a really low-bandwidth connection
from slow downloads? Have you ever run a website, only t
s, overloading your server and possibly causing high
e problems!

nt



Overlay Weaver

- 誰でもアクセスできる DHT サービス。
 - として PlanetLab 上で運用。OpenDHT のマネっこ。
 - XML-RPCで put, get できる。ライブラリ、ツールも提供。
- 最大 35カ国 418台で稼動。

Overlay	Weaver Node S	Status				
Node Inform	nation					
URL:	http://pl	. jp:3998/				
Node ID:	7a17c1434ef5858f0fbfe052a08b318bbb385433					
Lookup algorithm	n: Kademlia					
Lookup style:	Iterative	I				
# of stored keys:	0					
Routing Tal	ble					
http://pl	duke.edu:3998/	7abe663cbc720375b37dd7dcb71e7153aa85da0e				
152						
http://pl ibbt.be:3998/		7bc89d2027953a7bb7c7872290afe025abfce687				
http://pl		7ba15fb06db8821e233f94ab7497658e359ce865				
http://pl	.mit.edu:3998/	7b2da219f7d1c520d0fec850d2b66318ad46ccd4				
http://pl	📰 🖪: washington.edu:3998/	7b638aa42cf93c1f81c8caecdc5e4cf2e4cdabc0				
153						
http://p	ac.jp:3998/	793bfb7e552e859a63a2eff10d956ebbe678e226				



Overlay Weaver: 学んだこと

- 実環境での動作から学ぶことは多い。
 - Chord の欠陥 (?) に気づいた。
 - predecessor の更新がない。
- PlanetLab はアプリにかなり厳しい環境。 それだけに、育つ。育たざるを得ない。
 - 非常に負荷が高い。
 - load average 数十は当たり前。
 - そのせいもあり、遅延がとても大きい。数秒も普通。
 - 到達性が低い。しかも、非対称。
 - 信頼性が低い。
 - UDP データグラムの分割 → 破損が起きた??? → アプリの改善
 - Bamboo / OpenDHT の開発者 Sean Rhea は、このあたりを扱った論文を何本か出してる。
 - Churn, 非対称な到達性, DHTへのput/getの時間短縮, ...

なぜ PlanetLab?

- 「Internet 上に分散」している
 「大規模」なテストベッドであり、
 各「実験プロジェクトがよく分離」されている。
 - Internet 上に分散
 - 実ネットワークをまたいだ実験が可能。
 - 「それ本当に動くの?」というツッコミにある程度応えられる。
 - とはいえ、大学・北米に偏っている、という批判もある。
 - 大規模
 - 実機の台数としては望みうる最大規模。
 - 分離
 - プロジェクト (slice) ごとに仮想マシンの占有。
 - ●他の利用者にあまり気兼ねなく使える。
 - とはいえ、AUPはあるし、リソースを食い過ぎると迷惑。
 - 「PlanetLab のノードから DoS 攻撃が!」



利用までの手順

- NORTH
 A Pyongyang
 KOREA
 Seoul A
 SOUTH
 KOREA
- 所属組織が PlanetLab に参加
- ユーザとしての利用準備
- slice の作成・管理
- slice へのノードの追加・削除
- 実験



所属組織が PlanetLab に参加

paperwork

- Princeton U. とやりとり。
- 学科長級 (?) のサインが要る。

• 機材を用意

- PC 2台以上と PCU
 - PC: 性能に条件あり。 2006年末: DELL Dimension C521, 14万円
 - PCU: Power Control Unit, 遠隔電源制御装置 APC AP7900, 約 5万 5千円

• 機材を設定

- ノード情報が書き込まれた USB メモリ or CD-R を準備する。
- PC を起動する。
- ベースとなる Linux がネット越しに ガーッと流し込まれる。
- 以降のマシン管理は Princeton U. 側の仕事。

			etLab p Agreemen	t	
Type of Membership	(select one –	fees and benefits f	or each level ava	ailable	at www.planet-lab.org):
Charter Member Full Member Associate Member Sponsor Member Managing Party Academic Member	(\$300,000/yrUS) (\$ 75,000/yrUS) (\$ 25,000/yrUS) (\$ 10,000/yrUS) (\$0) (\$0)				
Member Information	n:				
Institution Name: Institution Address:		University Totsukamachi,	Shinjuku-ku	1, 70	kyo, 168-805Q JAPAN
Primary Administrative Contact Signature	直流	一辛		Date:	November 20, 2006
Name: Address:	Kazuyuki	Shudo		Title:	
114414551	100				hand the street will be a
Phone Number:		M. 18			AND A PROPERTY OF
E-mail:		Land March			_
Primary Technical Contact Signature		800			Secretary to a second
Name:		The same			
Address:	777	The second second		. 10	Committee Commit
Phone Number:	10000	367		100	The second second
E-mail:	Section 1			-	
Authorized Official Signature (Can bind the Institution contractually)		员 泽一		Date:	November 20, 2006
Name:	YOICHI	MURAUKA		/	
Address:	3-4-1 OKI	IBO SHINJUKU	TOKYO JAP	AN	
Phone Number:	100	40-1-00			A 60 L. Anna L. (2009)
E-Mail:		the state of the	Strange		
By the signature of the Membership Agreement		ficial above, the Me	mber Institution a	grees t	o the terms and conditions of this
Approved:					
Princeton University	Date	University of Cali	fornia, Date	Univ	ersity of Washington Date

PlanetLab Membership Application

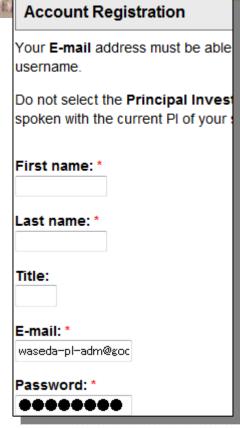
Terms and Conditions of Membership Review Carefully

Membership to the PlanetLab Consortium ("PlanetLab") shall be granted by PlanetLab in its sole and absolute discretion. Notice of Membership shall be sent to the above-named applicant ("Member") within thirty (30) days of submission of this Membership Agreement to the address above. Such notice shall be sent to both the Primary Technical Contact and the Primary Administrative Contact.

Upon receiving such notice (the "Effective Date"), Member shall abide by and be subject to the terms and conditions as set forth in this Membership Agreement and as set forth in the PlanetLab Consortium Governance Plan (the "Governance Plan"), the Acceptable Use Policies (the "AUP"), and PlanetLab Design Notes (PDN), (collectively "PlanetLab Policies"), all of which are available at www.planet-lab.org (the "PlanetLab Website"), and are incorporated herein by reference. Member hereby represents that it has reviewed the PlanetLab Policies on the PlanetLab Website as of the date hereof.

ユーザとしての利用準備

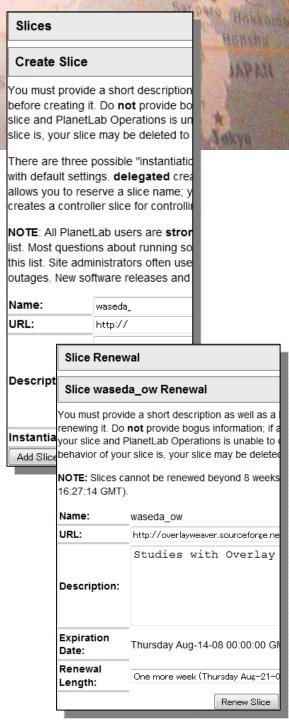
- PlanetLab の AUP を読んで、理解する。
 - AUP: Acceptable Use Policy
- ウェブサイトにて、アカウントを作成
- PI が承認。
 - PI: Principal Investigator (研究推進者) PlanetLab の利用者は 3種類: PI, tech contact, user
- ノードへのログインに使う SSH の鍵を 準備
 - SSH の鍵ペアを公開鍵を生成し、ウェブから登録する。





slice の作成・維持

- slice の作成
 - ウェブから申請
 - slice の名前 waseda_○○
 - PI が承認
 - Academic の場合、slice は 10 まで。
- slice の維持
 - 2ヶ月 (未満) ごとに renewal



slice へのノードの追加・削除

- 追加して15分かそこら待つと、 仮想マシンが起動され、 SSHログインできるようになる。
- 手で数百台を追加するの!?
 - 可能。でも、エラく大変。これまで何十時間費やしたことか。
 - 頻繁に、落ちたり復活したりする。→ 要 メンテナンス
- ノード, slice, ユーザ等の 管理作業はすべて、 ウェブ、RPC (XML-RPC/SOAP) の双方で行える。
 - XML-RPC / SOAP の API: PlanetLab Central API (PLCAPI)
 - Python 等のライブラリがある。

Slice waseda_ow - Nodes
Select a site to add nodes from.
Waseda University
planetlab1.muraoka.info.waseda.ac.jp
planetlab2.muraoka.info.waseda.ac.jp
Add Nodes
Nodes currently associated with slice
Check boxes of nodes to remove:
planet2.scs.cs.nyu.edu
pl1.cs.utk.edu
pl2.cs.utk.edu
planetlab2.cs.wisc.edu
$\begin{tabular}{l} \hline \blacksquare \ planet2.pittsburgh.intel-research.net \\ \hline \end{tabular}$
planetlab1.postel.org
planetlab03.cs.washington.edu
planetlab1.cs.unibo.it
planetlab3.cs.duke.edu
planetlab1.csail.mit.edu
planetlab3.csail.mit.edu
planetlab2.xeno.cl.cam.ac.uk

vn1 cs wustliedu

planetlab2.netlab.uky.edu

slice へのノードの追加・削除

- スマートな方法
 - CoMonから稼動ノードの一覧を得る。
 - % wget -O nodes.txt
 "http://comon.cs.princeton.edu/status/tabulator.cgi
 ?table=table_nodeviewshort&format=nameonly&select='
 resptime>0'"
 - ファイル nodes.txt として、ノード一覧が得られる。
 - PLCAPI を叩いて、一気に slice に追加する。
 - 手元のマシンに PLCAPI (RPC するライブラリ) をインストールしておく。
 - % plcsh -h https://www.planet-lab.org/PLCAPI/ -u shudo@ほ げほげ

Password: <パスワードを入力>

shudo@ほげほげ connected using password authentication Type "system.listMethods()" or "help(method)" for more information.

[shudo@ほげ]>>> node_list = [line.strip() for line in open("nodes.txt")]

[shudo@ほげ]>>> ÁddSliceToNodes("waseda_ow", node_list) [shudo@ほげ]>>> exit



実験

- あとは ただ、Linux マシンとして使うだけ。
 - SSH ログインして使う。

```
The authenticity of host 'planetlab–01.naist.jp (163.221.11.71)' can't be established.
RSA key fingerprint is 39:
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'planetlab–01.naist.jp,

' (RSA) to the list of known hosts.
Last login: Mon Jun 23 10:09:45 2008 from
[waseda_ow@planetlab-01 ~]$ [
```

MORTH W Pyengyang

Seoul A

- Fedora 8 ベース。Linux 2.6.22.19。
- yum や rpm コマンドで RPM パッケージを追加・削除できる。
- 数百台を使うのは、それなりに大変。
 - ソフトのインストール, 起動 / 停止, 監視, 制御



ツール

- ソフトウェア管理
 - Stork, CoDeploy, ...設定ファイルを書いてほげほげ。使ったことありません!
- 複数ノードの一括操作
 - parallel SSH
 - pssh, pscp, ...
 - 対象ホスト名/IPアドレス一覧をファイルとして与える。

Seoul A

SOUTH

- GXP (Grid and Cluster Shell, by 東大田浦さん) なんかいいんじゃないでしょうか。
- ... 僕も昔、汎用のツール (シェルスクリプト集) を書きました。
- PlanetLab 専用、というツールはあまりない。
- 結局、皆、自分でスクリプトを書いてなんとかしている。
 - 敷居はなかなか高い。そりゃ、何十台、何百台を使おうというんですから。

ツール: SSH の準備

- ノードへの操作はすべて SSH で行うので...
 - ログインごとにパスワードを打たずに済むように
 - % eval `ssh-agent` % ssh-add ~/.ssh/id_...(秘密鍵のファイル) Enter passphrase for /home/...:
 - ログインごとにホストキーを確認されないように
 - 一度手でログインして「yes」を打っておく。
 or ssh -o StrictHostKeyChecking=no



OWの場合: ソフトウェア管理

• ソフトウェアをインストールするスクリプト ow-setup.sh

```
#!/bin/sh
USER=waseda_ow スライス名
HOST_FILE=$* ホスト名一覧の入ったファイル: ホストファイル
OPT="-0 StrictHostKeyChecking=no -1 $USER -h $HOST_FILE -t 1200"

# copy .bashrc
pscp $OPT .bashrc /home/$USER/
# install JDK
pscp $OPT jdk-*.rpm /home/$USER/
pssh $OPT "sudo rpm -Uvh jdk-*.rpm && rm jdk-* && sudo ln -sf
/usr/java/jdk1.5.0_15 /usr/local/java"
# install Overlay Weaver
pscp $OPT overlayweaver.tar.bz2 /home/$USER
pssh $OPT "rm -rf overlayweaver && tar xjf overlayweaver.tar.bz2"
```

- 出力を元に失敗したノードの一覧を作成し、再度スクリプトを実行。
 - 実行時の出力例:

% ow-setup.sh nodes.txt
Success on planetlab3.williams.edu:22
Error on onelab03.inria.fr:22
Error on pl-1.hip.fi:22

• • •

- grep とか Emacs のキーボードマクロとかを駆使する。
- 全ノードで成功するまで繰り返す。か、どこかであきらめるかする。

OWの場合: ソフトウェア管理

fgrep /usr/java > log-\${i} &

done

ファイルサイズを見ると、インストールされているか否かが判る。

ssh \$OPT \${USER}@\${i} ls -d /usr/java 2>&1 |

- % ls -l log-* -rw-r-r- 1 shudo users 10 2008-07-02 03:43 log-ホスト名
- 1s の結果を編集することで、 インストールし直しのためのホストファイルを得られる。

OWの場合: ソフトウェアの起動・停止

- OW を起動するスクリプト ow-invoke.sh
 - #!/bin/sh
 USER=waseda_ow
 HOST_FILE=\$*
 OPT="-0 ...

```
pssh $OPT 'owdhtshell -a Chord -x -n -p
10101 -acl acl.txt -no-upnp > /dev/null
2>&1 &'
```

- 停止するスクリプト ow-kill.sh
 - #!/bin/sh
 ...
 pssh \$OPT "killall java"



OW の場合: 監視

- プロセスの存在を確認するスクリプト test-process.sh
 - #!/bin/sh

ファイルサイズを見ると、プロセスが存在するか否かが判る。

```
% ls -l log-*
-rw-r-r- 1 shudo users 225 2008-07-02 03:43 log-ホスト名
```

1s の結果を編集することで、 起動し直しのためのホストファイルを得られる。



ソフトウェア・プロセス管理: まとめ

- シェルスクリプト, SSH, parallel SSH を駆使して
 - ソフトウェア管理 ow-setup.sh test-installation.sh
 - プロセス管理 ow-invoke.sh ow-kill.sh test-process.sh
- スクリプトには、毎度、適切なホスト一覧 (ファイル) を与える。

ソフトウェアや実験に固有の処理

- OW の場合、起動後に、 オーバレイの構築が必要。
 - 全ノードを join させる必要がある。
 - 手で制御? 台数が多いので、現実的ではない。% telnet planetlab-01.naist.jp 10101

OWのシェルに接続

Ready.
init pub1-s.ane.cmc.osaka-u.ac.jp
contact: 13X.X.XX.XXX:3997

- OW 自体の機能で、各ノードを制御。
 - ホストファイルを読んで、各ノードに TCP 接続。接続先は OW のシェル。
 - 与えられたシナリオに従って、ホスト群を制御。
 - % owemu -c host_file scenario_file
 - 実は、分散環境エミュレータの機能拡張。
 - (仮想) ノードを起動する代わりに、動作しているノードに TCP 接続する。

まとめ

- PlanetLab は生 Linux。
 - 使うのはそれなりに大変。生ゆえ、自由度が高い。
- 実 Internet での実験からは、学ぶことが多い。
 - アプリが強くなる。というか、ならざるを得ない。
- 実は、実験プロジェクト間の分離がよくなされていることが活用を促進してるのではないか。
 - slice の中で何をやろうと、他への影響は小さい。
 - 管理ソフトが大事な役割を果たしている。
 - cf. そこらの、生の PC クラスタ, StarBED, ...





JGN2plus の姿



サイバー関西プロジェクト 研究会 (7/2-3) での 小林和真 (kazu-k) さんの御講演





Langhou

NORTH * Pyongyang KOREA Seoul SOUTH

KOREA

Howshis 松岭山

Sagnara Bakkalas

JGN Platform Services

- Provisioning of QoS
 - DCN (Dynamic Circuit Network) and extended GMPLS based.
 - Collaborate with Internet2, APAN, possibly GEANT
 - Web service based
- Measurement
 - PerfSONAR based.
 - Collaborate with Internet2, APAN, possibly GEANT
 - Web service based
- Overlay Services
 - Planetlab/CORE (by Prof. Nakao@Hakusan) and their output
 - PIAX (by Prof. Teranishi)
 - Other



研究グループ?

NORTH w Pyengyang KOREA Seoul SOUTH

Houses MAGAI

NICT

Network Virtualization

Lab

CORE

Sagnero Boxkolos

Clusters of Testbeds in JGN2plus Report of MSF Workshop on Network Research Textinos http://guiz.co.umars.edu/jesthed_workshop Network Computing Starbed Application Largescale Network emulation environment Tiled Display wall (Optiportal) Coming Soon E-science Network Measurement Optical Networking Control and ès application Sensor network Live El

PIAX

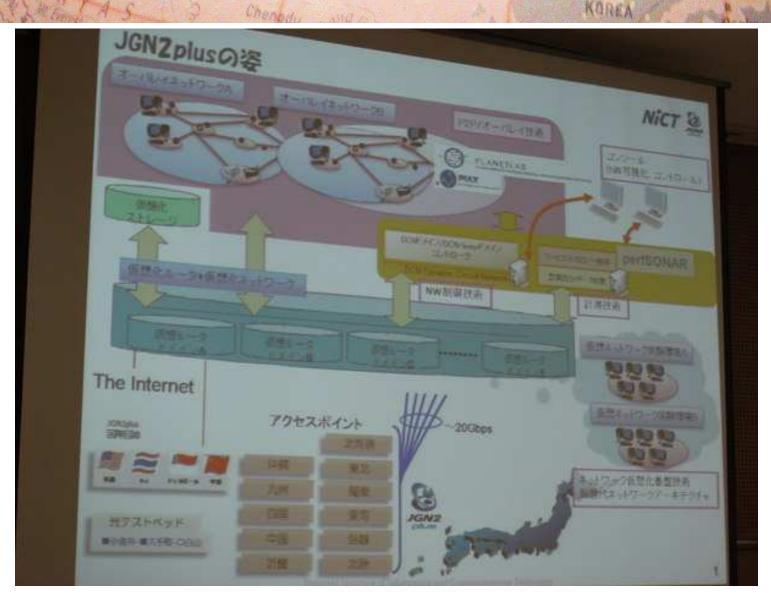
Otemachi (shimojo and esaki)

Sensor network

Live E!

PIAX

JGN2plus の姿 下條先生版???



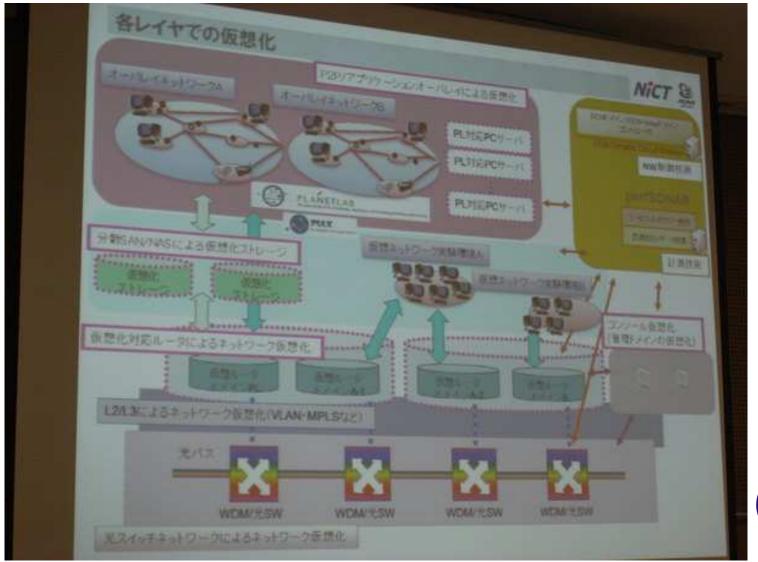


Sagnery Boxkolos

Howshis

MAGAI

JGN2plus の姿 kazu-k さん版





Sagnara Bakkala

MAGAI

議論?

- センサネットワークの実験を助ける
 - エミュレータ, シミュレータ
 - _ テストベッド
 - のあり方。
 - どんな設備 / ソフトがあるといいのか。Emulab ? StarBED 上でやられてるようなやつ?
- JGN2plus や StarBED 等への提言。
- DNS 的 階層型 分散 DB とオーバレイの比較