



Live E! シンポジウム 2008,
東京大学, 2008年9月19日(金)

オーバレイ・ センサネットワーク

～ オーバレイネットワークへの期待

首藤 一幸

情報通信研究機構 (NICT) 客員研究員

内容

- オーバレイネットワーク
 - 非構造化 / 構造化オーバレイ
- 非構造化オーバレイの基本と実例
 - Peer-to-peer ライブ配信
- 構造化オーバレイの基本と実例
 - Overlay Weaver とその PlanetLab 上の運用
- オーバレイへの期待
 - ストレージ, データ処理, ...



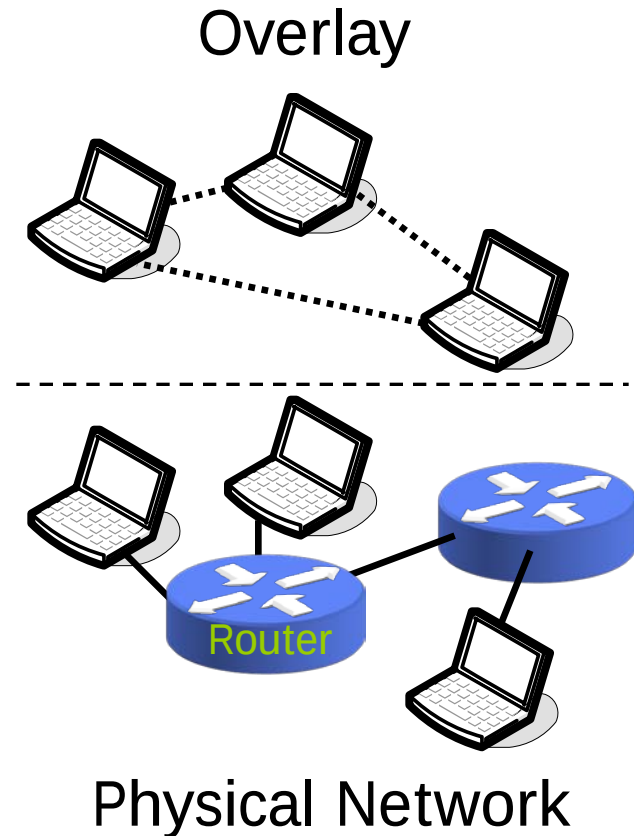
オーバレイネットワーク

非構造化オーバレイ
と
構造化オーバレイ

オーバレイネットワーク (overlay network)

- 下位ネットワークとは独立したトポロジを持つ上位ネットワーク

- UUCP, NetNewsサーバのトポロジ
- 電話網上のインターネット
- インターネット上のVPN, CDN
- MBone, 6bone, ...
- GENI
 - 米国 NSF のテストベッド構築プロジェクト
 - 研究者は、自身の“net”を構築
 - PlanetLab からの強い影響
- ここでは特に、大規模な自律分散系
 - それって (pure) P2P ?
 - 狙い: 多ノード / 耐故障 / 低管理コスト
 - 自律的にオーバレイを構成



Unstructured と Structured

- 非構造化 (unstructured) オーバレイ

- 例： **Gnutella** ネットワーク, Winny ネットワーク
- 誰を隣接ノードとするか、**トポロジに制約がない**。
- 存在するオブジェクトは、**発見できる可能性がある**。
- 一般に、効率は良くないが、柔軟な検索が可能。

- 構造化 (structured) オーバレイ

- 例： **DHT** (分散ハッシュ表) のネットワーク
- 誰を隣接ノードとするか、**トポロジに制約がある**。
- 存在するオブジェクトは、**(たいてい) 発見できる**。
- 一般に、効率は良いが、柔軟な検索が苦手。



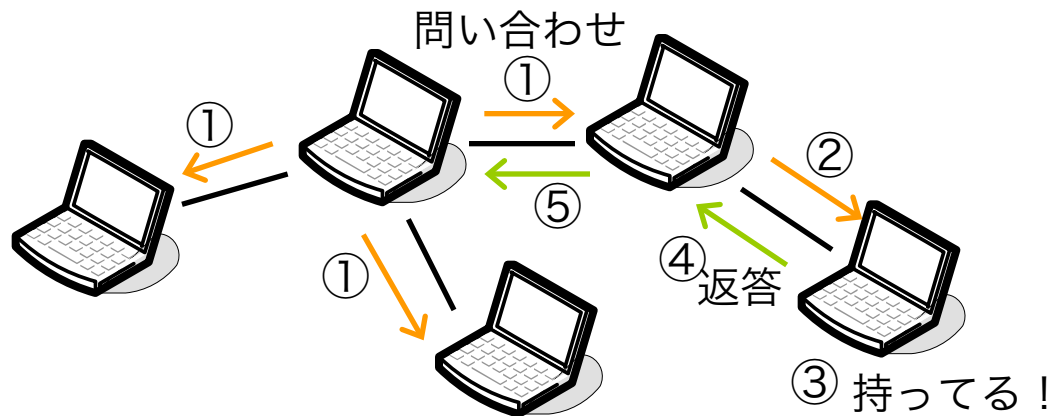
非構造化オーバーレイの 基本と実例

- 基礎: Gnutella プロトコル
- peer-to-peer ライブ配信

Unstructuredオーバレイの例:

Gnutella プロトコル (0.4)

- 各ノードは、一定数の隣接ノード (neighbor) を持つ。
 - どのノードを隣接ノードとするか、および、隣接ノードの数はアプリ依存。例えば 4。
- 検索時、各ノードは問い合わせを全隣接ノードに転送する → **flooding**
 - TTL (time-to-live, 最大ホップ数) は 7。
 - 返答は、転送経路に沿って返される。



- Gnutella プロトコル 0.6 は、super-peers の概念を採り入れた。

UG Live: Peer-to-Peer ライブ配信

- 映像・音声の**ライブ**配信が可能。

ここが技術的に面白い。メッシュ型 (⇔ツリー型)
cf. YouTube, ○○動画

- 配信元の**ネットワーク帯域幅**をほとんど**食わない**。

⇒ 極めて**低い配信コスト**

⇒ 誰でも発信

cf. サーバ-クライアント, CDN

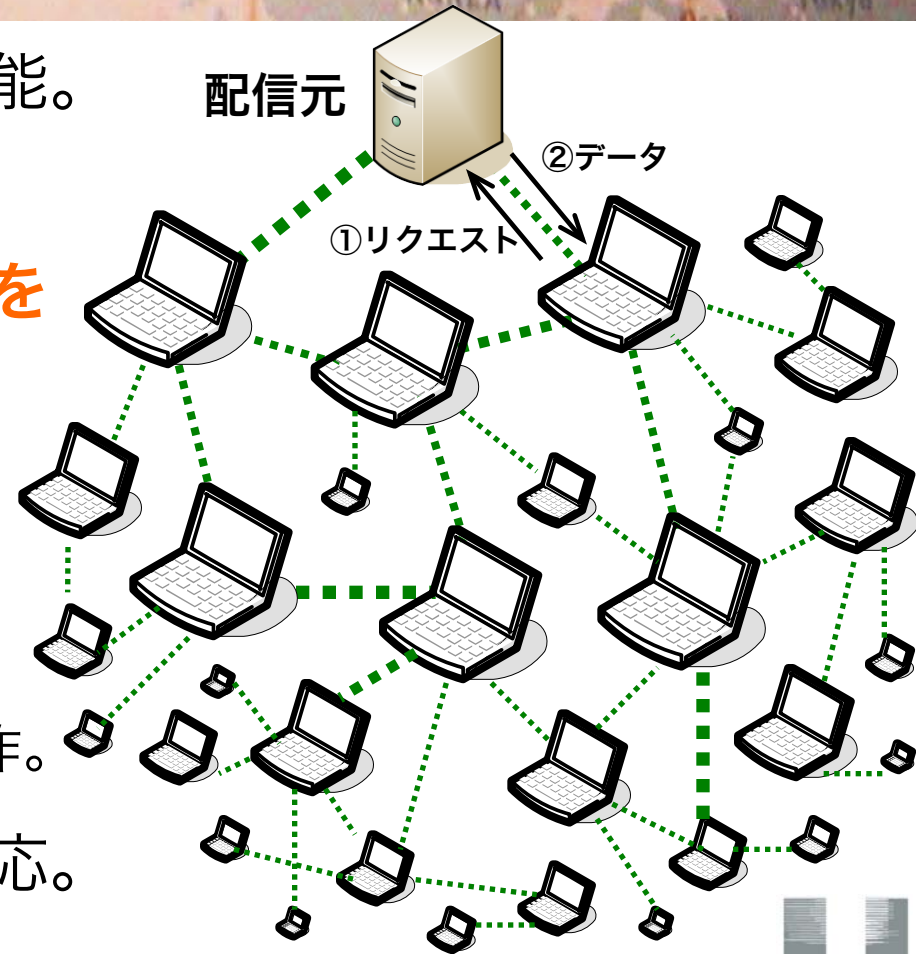
- 1万ノード**での動作実績。

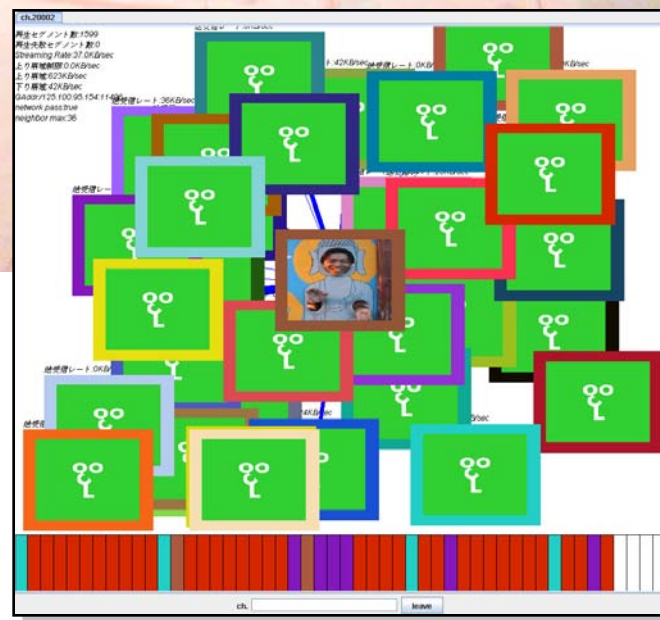
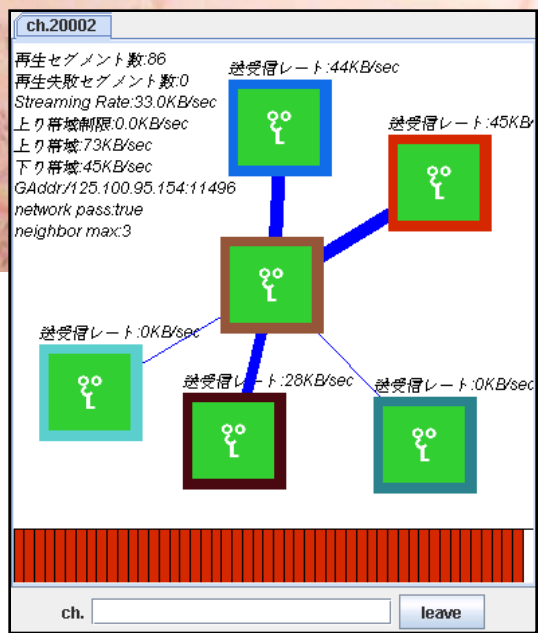
(PC クラスタ上)

実地 & PlanetLab上でも大規模動作。

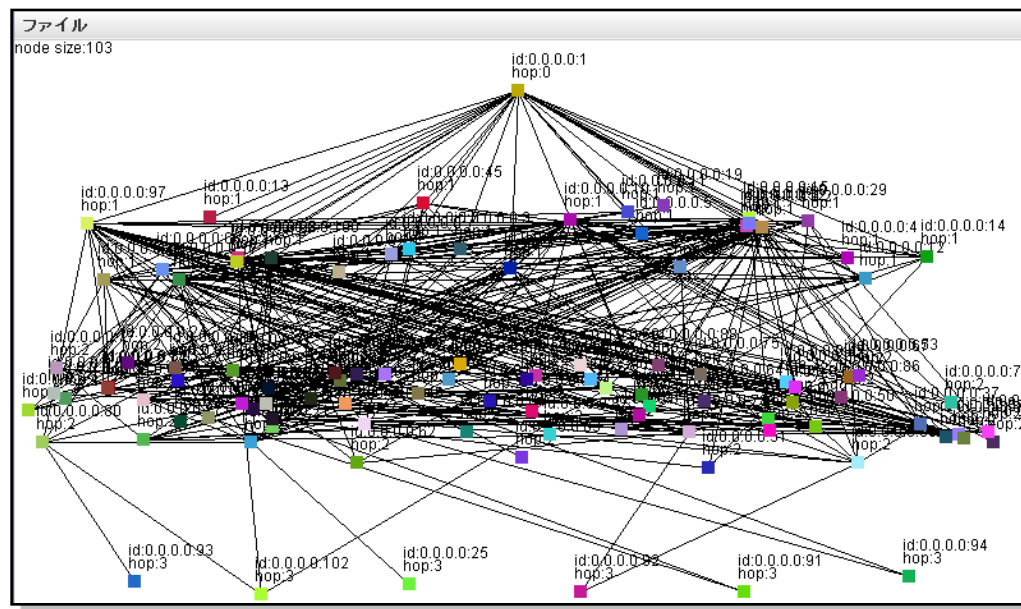
- 様々なネットワーク環境に適応。

xDSL~光ファイバ, ファイアウォール/NAT





各ノードで、周辺状況を可視化

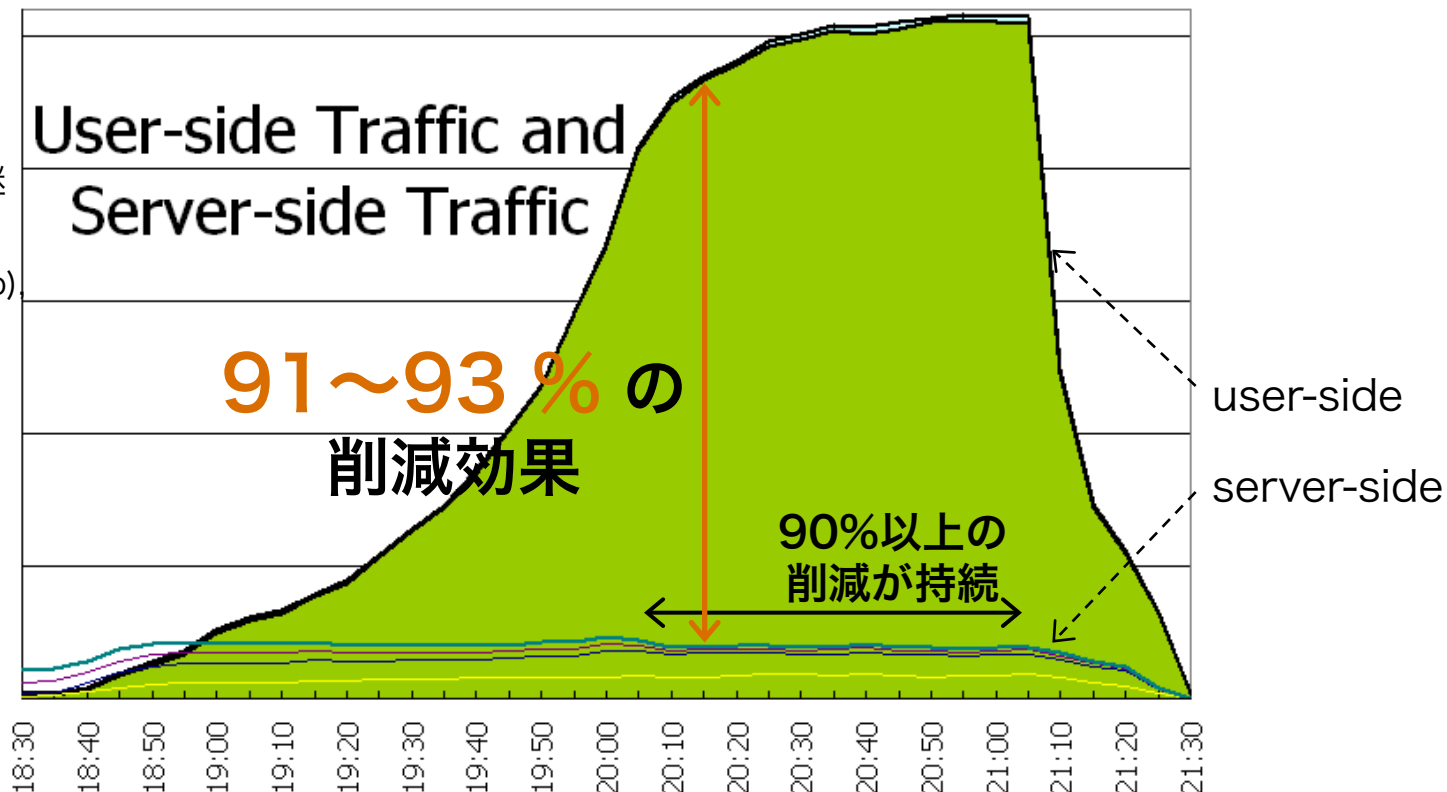


全ノードの接続状況を可視化

UG Live: トラフィック削減効果

- 視聴側トラフィックに対する
配信側トラフィックが、わずか数%
→ 90% 以上の削減効果

2007年 11月 26日
accessのtalk about生中継
配信: Jストリーム,
Castella (www.castella.jp)
ウタゴエ



事例: UG Live



OceanGridについて

OceanGridとはウタゴエ株式会社が開発したグリッド型ライブストリーミング配信システムです。グリッド型ライブストリーミングとは、同じライブ中継を視聴する視聴者間で直接ライブストリーミングのデータをリレー送信することで大規模なライブ中継を実現するためのシステムです。

映像・音声のエンコードおよび再生にはMicrosoft社のWindows Mediaプラットフォームを利用し、OceanGridは視聴者間のデータのリレー送信を実現しています。

本コンテンツを閲覧するためには、Windows Media PlayerとOcean Gridクライアントソフトウェアが必要です。



OceanGridクライアントソフトウェアのダウンロードは[こちら](#)
インストール手順については[こちら](#)
視聴に関するよくあるお問い合わせについては[こちら](#)

鹿島アントラーズ戦 音声 ライブ配信

協力：Jストリーム様

- ・2回 / 月 程度
- ・2007年 4～12月
- ・2008年 4～12月

事例: UG Live



ステージLIVE映像

技術協力: [基幹理工学部 後藤滋樹研究室](#)



300 Kbps
(IPv6配信は[こちら](#)をクリック)



1 Mbps
(IPv6配信は[こちら](#)をクリック)



P2Pによる超高画質配信
2 Mbps

==== P2P配信について ====
[専用ソフトダウンロード](#)
[ソフトのインストール方法](#)
開発元: [ウタゴエ株式会社](#)

早稲田祭 2007 のステージ中継 (2007/11/3~4)

協力: 早稲田大学 後藤研究室 様

非常に高いビットレート 2 Mbps

事例: UG Live



「きずな (WINDS)」 / H-IIAロケット14号機
打ち上げライブ中継 P2P実証実験サイト

映像提供: 宇宙航空研究開発機構 (JAXA)

超高速インターネット衛星「きずな (WINDS)」
H-IIAロケット14号機打ち上げの様を種子島のスタジオからライブ配信中!



24インチLCDでの
視聴に耐える品質



人工衛星「きずな (WINDS)」
/ H-IIAロケット14号機 打ち上げ (2008/2/23)

配信: Jストリーム様, ウタゴエ

事例: UG Live

Peer to Peer
P2P 「きぼう」打ち上げ第2便:1J(STS-124)
ミッションライブ中継 P2P実証実験サイト
P2Pネットワーク実験体

映像提供: 宇宙航空研究開発機構 (JAXA)/NASA

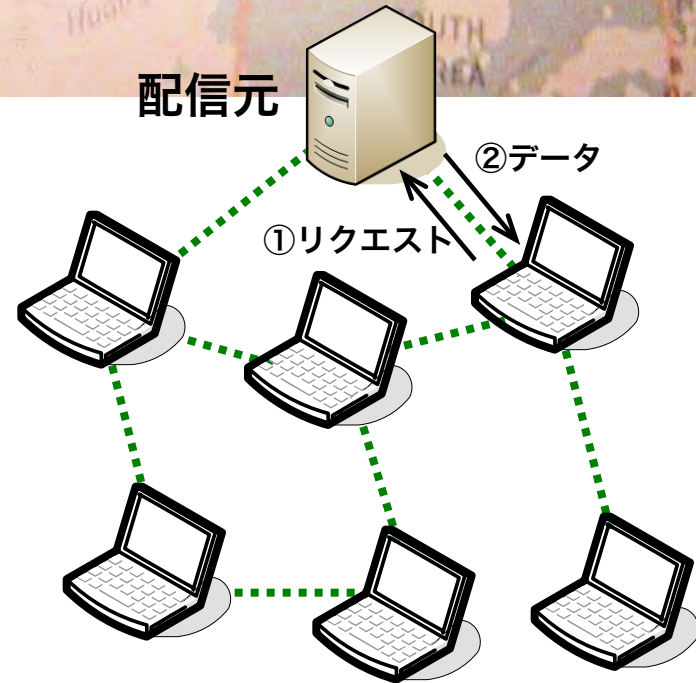
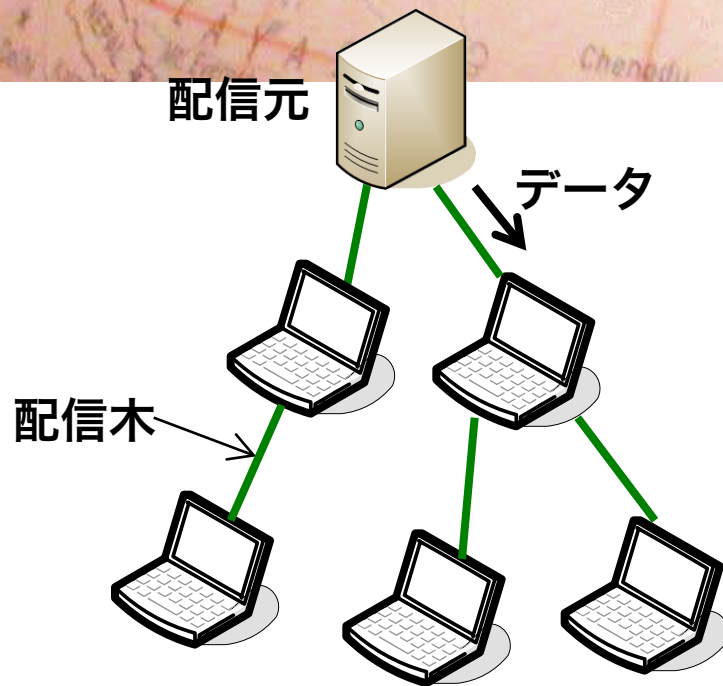
「きぼう」日本実験棟 1J(STS-124)ミッション ライブ中継
国際宇宙ステーションへの組み立て第2便 ～ きぼうの、その先へ ～



スペースシャトル打ち上げ,
「きぼう」日本実験棟 組み立て (2008/6/1-15)

配信: Jストリーム様, ウタゴエ

アプリケーション層マルチキャスト (ALM) の手法: ツリー型 vs. メッシュ型



・ ツリー型

- 明示的に作ったトポロジがデータ流路。
- 根から葉に向かってpush。
- △ ノード故障時は、要 迅速な復旧。
- ○ 末端まで確実に届く。配信遅延 小。
- 広い上り帯域幅を使い尽くし得る。

・ メッシュ型

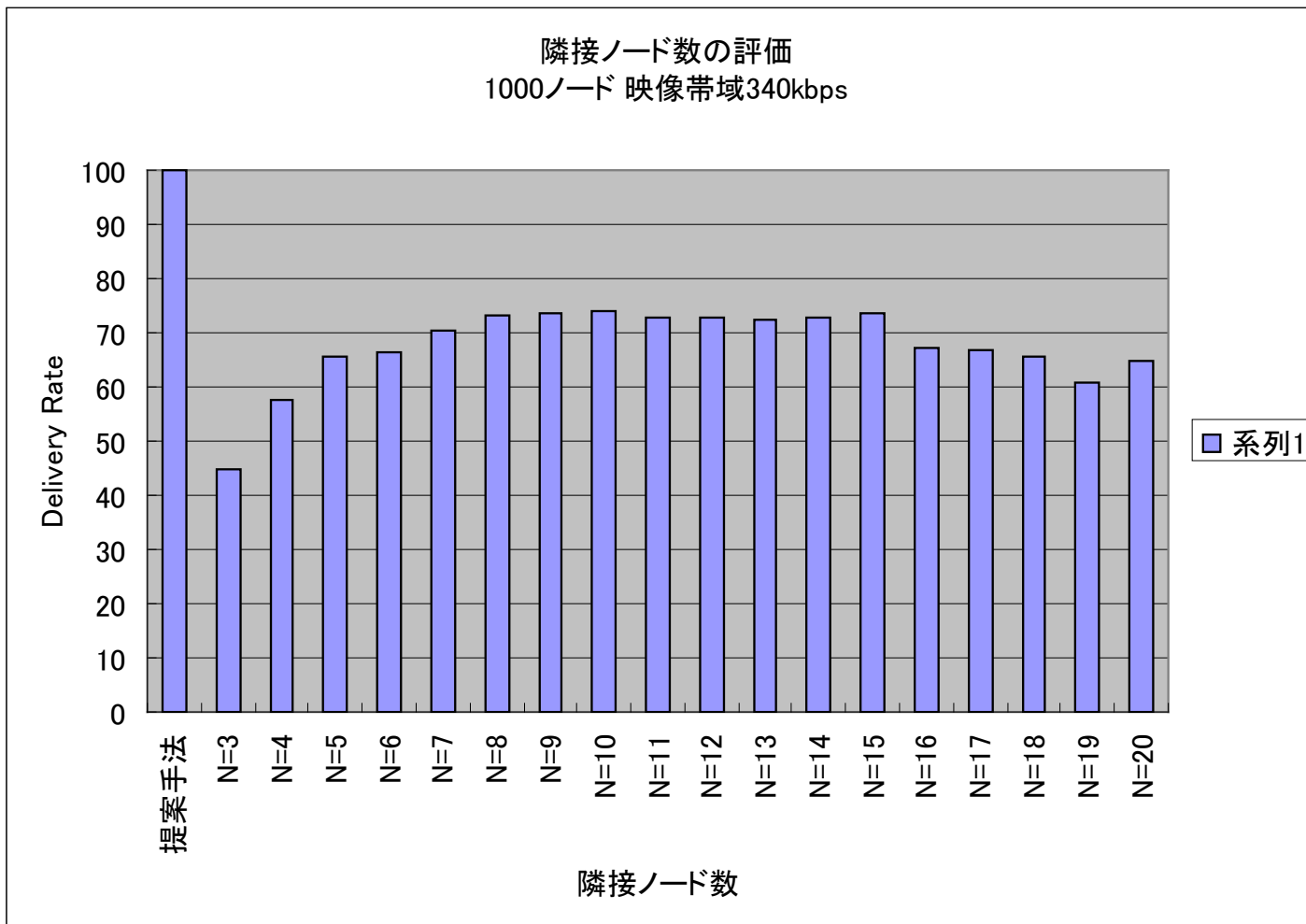
- 隣接ノードと緩やかな関係を保つ。
- 隣接ノードからpull。
- ○ 生来、ノード故障に強い。
- △ 末端まで届く保証なし。要 補償策。
- 狭い上り帯域幅を使い尽くし得る。

技術の詳細

- メッシュ型の最初の提案は、香港の大学 & カナダの大学から。
 - X. Zhang et al.: CoolStreaming/DONet: A Data-Driven Overlay Network for Efficient Live Media Streaming, Proc. INFOCOM 2005.
- ウタゴエでの研究・開発:
現実のインターネットでうまく動かす
 - 隣接ノード数の調整アルゴリズム
 - 上り/下り帯域幅の差への配慮 (秘密の技術アリ)
 - 配信元ノードからデータがスムーズに出て行く工夫
 - バックアップの仕掛け
 - NAT越え, ウェブとの連携, ...

研究 (例): 隣接ノード数の調整

- 帯域幅等に応じて、動的に調整。



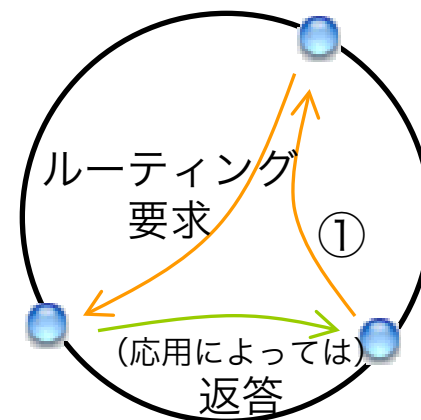
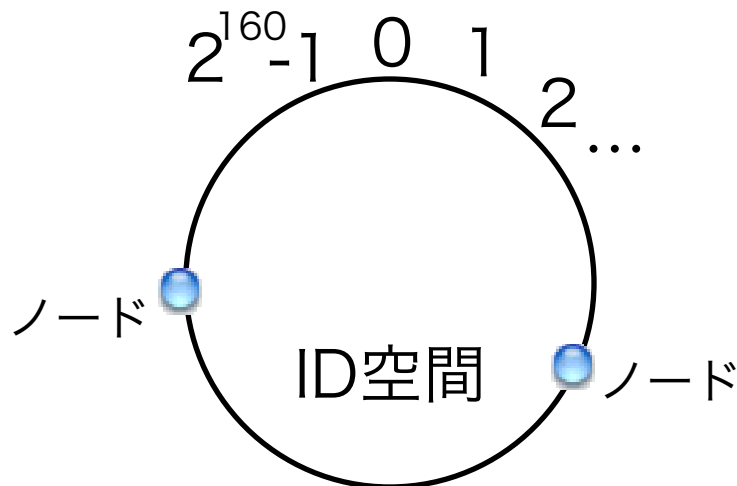


構造化オーバレイの 基本と実例

Overlay Weaver と
その PlanetLab 上の運用

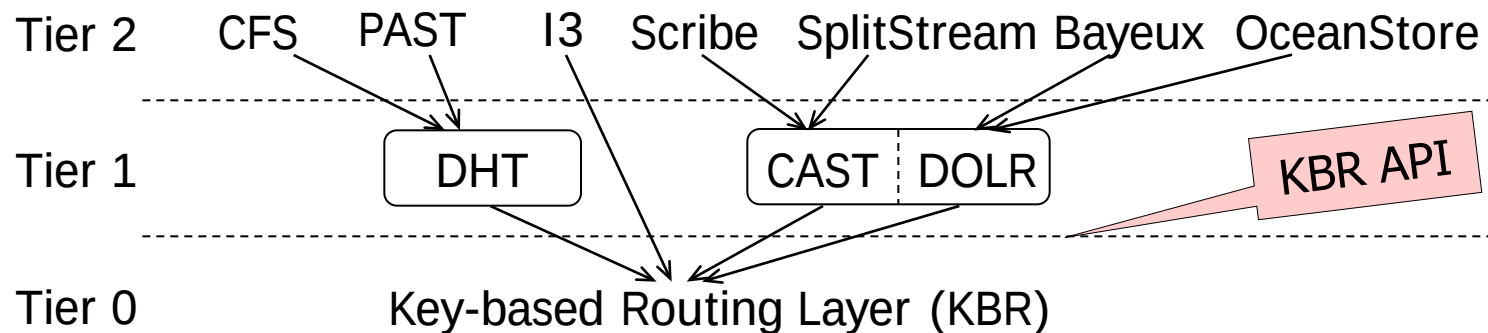
構造化オーバレイの基本

- ノード（計算機）とオブジェクトの両方にIDが振られる。
 - IDはたいてい整数値。160ビットだったり 128ビットだったり。
 - オブジェクト：任意の文字列だったり、ファイルだったり、プロセスだったり。
- ノードは、ID空間中のある範囲を受け持つ。
 - だいたい、ノードのIDと数値的に近い範囲を受け持つ。
- IDを宛先としてルーティングが行われ、その行き着く先は、そのIDを受け持つ担当ノードとなる。



構造化オーバーレイの基本

- 肝は **routing / forwarding**
- 資源にかかる負担が、ノード数 n として $O(\log n)$ 。
 - forwarding 時のメッセージ数など。
 - cf. 非構造化オーバーレイ上の flooding
- 様々なアルゴリズムが提案されてきた。
 - CAN, Chord, Tapestry, Pastry, Kademlia, Koorde, Broose, Accordion, ORDI, DKS, D2B, Symphony, Viceroy, ...
- 構造化オーバーレイ上に、いろいろなサービスが載る。
 - 分散ハッシュ表 (DHT), マルチキャスト, メッセージ配送, ...



Cited from [Dabek03]

構造化オーバーレイの応用例

- BitTorrent クライアント
 - eDonkey2k (hybrid P2Pプロトコル) から派生した Kad Network。eMule 等が実装。
 - BitTorrent、Azureus のトラッカーなし動作
 - Azureus: 100万以上のノード
- DNS
 - “A Comparative Study of the DNS Design with DHT-Based Alternatives”, Proc. INFOCOM 2006.
- その他、もろもろの名前解決や位置解決
 - ホスト名 → IPアドレス, 名前 → 電話番号, 曲名 → 楽曲ファイルやそのURL などなど。

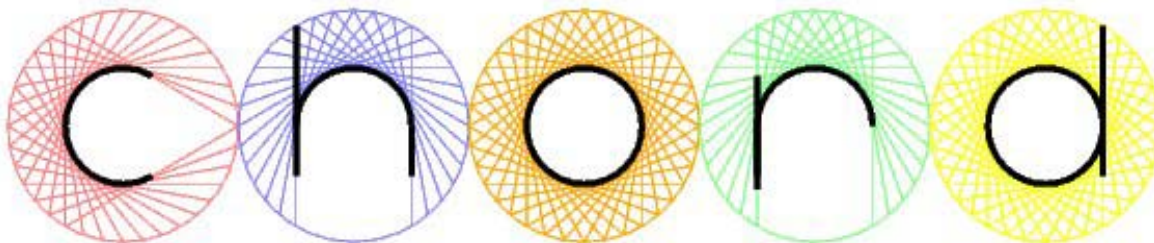
オーバレイ織り器

Overlay
Weaver

- (構造化) オーバレイ関係の名前には、糸を織るというアナロジが見られる。
 - Chord -- 弦
 - Tapestry -- 壁掛けの織物
 - Pastry -- 練り粉, パイの皮 ??
- Overlay Weaver
 - (structured) オーバレイの織り器
 - Weave -- 織る, 編む
 - Weaver -- 織り手, 織工, 編む人



Weaver

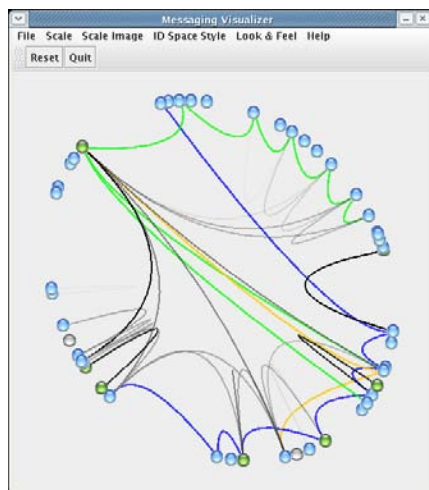


Tapestryの口ゴ

Overlay Weaver

● 構造化オーバーレイライブラリ

- 記述言語は Java。
- 約 3 万ステップ。
 - 4 万 8 千行。
- Apache License 2.0。
 - いろいろな目的に使いやすい。



可視化ツール

● 特徴

- 分散ハッシュ表 (DHT) だけじゃない。アプリケーション層マルチキャストも。
- ルーティングアルゴリズムを差し替え可能：
Chord, Kademlia, Koorde, Pastry, Tapestry
- コーディングなしで運用や実験が可能。
 - サンプルツール (DHTシェル等) を使った運用。
 - エミュレータを使った実験: メッセージ数やホップ数の計測。PC 1 台で 15 万ノード動作。
- 動作状態を可視化して楽しめる。デモできる。
- XML-RPC 経由で DHT を使える。
 - Bamboo, OpenDHT と同じプロトコル → 同じクライアントを使える。
- 地球規模テストベッド PlanetLab 上で運用している。
 - 34 ヶ国 490 台

オープンソースソフトウェア, 研究プラットフォームとして

Overlay
Weaver

- <http://overlayweaver.sf.net/> (SourceForge)

- 2006年 1月 17日、リリース。
- Apache License 2.0

- 状況 (2008/9/10 1時半)

- ダウンロード 10,739件。
- メーリングリスト登録数
 - 英語 76名, 日本語 90名
- Mixi コミュニティのメンバー数 156名

- 研究プラットフォームとして

- ブラジル, ロシア, 英国, ハンガリー, ...
- 第三者が、ルーティング方式を追加実装:
Symphony, EpiChord, ...
- 第三者による利用 / 応用多数:
RDFのDB(産総研), XMLデータ検索(筑波大), 複製数の影響
(U Federal do Rio Grande do Sul), 複数サーバの横断検索
(お茶大), ウェブアクセス動向集計(東工大), 複製配置手法
(早大), コミュニケーション向けユーザ管理サーバ(日立),
多次元範囲検索システム(NEC), ...

Overlay
Weaver

[English |
Japanese]

概要

オーバレイ構築ツールキット

Overlay Weaver

Overlay Weaver はオーバレイ構築ツールキットです。アプリケーション開発に加えて、オーバレイのアルゴリズム設計もサポートします。

アプリケーション開発者に対しては、分散ハッシュ表 (DHT) やマルチキャストといった高レベルサービスに対する共通 API を提供します。

ウェブサイト

スクリーンショット

Messaging Visualizer

円 直線 曲線 渦巻 格子

50 のノードと、それらの間の通信が可視化されています。マルチキャストのための配送木も色付きの線として描かれています。ここで全ノードと Messaging Visualizer は、エミュレータの上で計算機 1 台上で動作しています。Messaging Visualizer は実ネットワーク上でも動作します。

Messaging Visualizer と 300 の (仮想) ノード

円 格子

首藤自身の研究プラットフォームとして

- churn 耐性の評価と向上手法
 - 首藤, “下位アルゴリズム中立なDHT実装への耐churn手法の実装”, 情処ACS論文誌, 2008年3月.
- collective routing
 - 首藤, 中尾, “構造化オーバーレイでの一括ルーティング”, OS研究会 (SWoPP), 2008年 8月.

PlanetLab 上での運用

- 実機上、できればインターネット上で、実際に大規模な試験・運用を。
- PlanetLab: 地球規模テストベッド
 - PC 2台を供出すれば、使える。
 - プロジェクトごとにVM (Linux-VServer) を用意できる。
 - ただし、企業の fee は \$25,000 / 年～！
 - 客員研究員として、大学を PlanetLab に参加させた。
 - 最大で、**34ヶ国 490台**で動作。

Overlay Weaver Node Status

Node Information

URL: <http://planetlab.jp:3998/>
Node ID: 9e0074f964d11454587484e4bb5ca356658da1c8
Lookup algorithm: Chord
Lookup style: Iterative
of stored items: 0

Routing Table

Predecessor

<http://planetlab.fr:3998/> 902cd17e540547a50b5c12c680a4b2b985947972

Successor List

<http://planetlab.cn:3998/> 9e2a85116bbbe229e69399dd8f4470ad8a24259a
<http://planetlab.cn:3998/> a4b5eda65b7726e858d8fbd0fb364fbd84231a61
<http://planetlab.cn:3998/> a6c1c54334ef2a7d2c5e303094393e978692b30a
<http://planetlab.jp:3998/> ab028d1caad8482a8777bf93216515f4e02e19a4
<http://planetlab.edu:3998/> b010cfa1ec63dffb04123eac81fee105f033c4ec
<http://planetlab.com:3998/> b41d267be23ef5ecb5b1ef659f554a925d0bef37
<http://planetlab.jp:3998/> b48851667b142d4253aeba48381e5402e0c8ad5c
<http://planetlab.cn:3998/> b591843d3cc42317a8e065e80a8089c90320cd27

Finger Table

1 <http://planetlab.cn:3998/>
151 <http://planetlab.cn:3998/>
156 <http://planetlab.edu:3998/>
158 <http://planetlab.net:3998/>
159 <http://planetlab.fr:3998/>
160 <http://planetlab.cn:3998/>

PlanetLab上で運用しているノードの
ウェブインタフェース

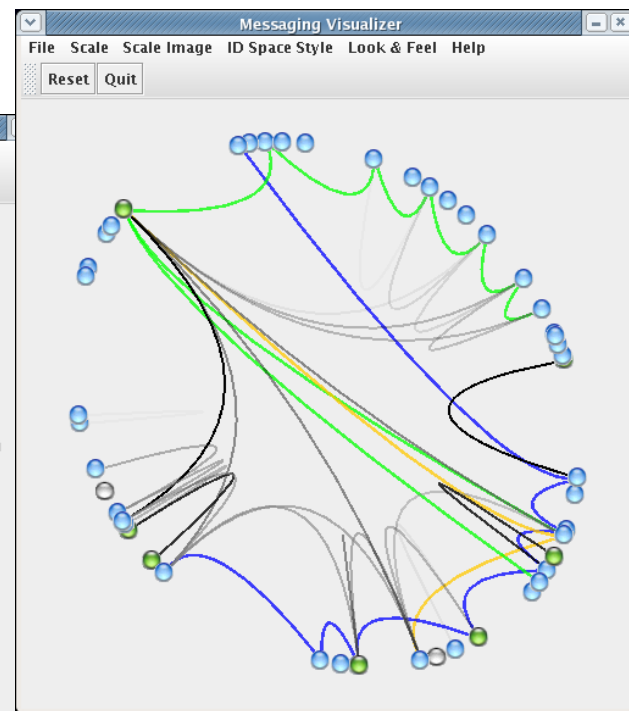
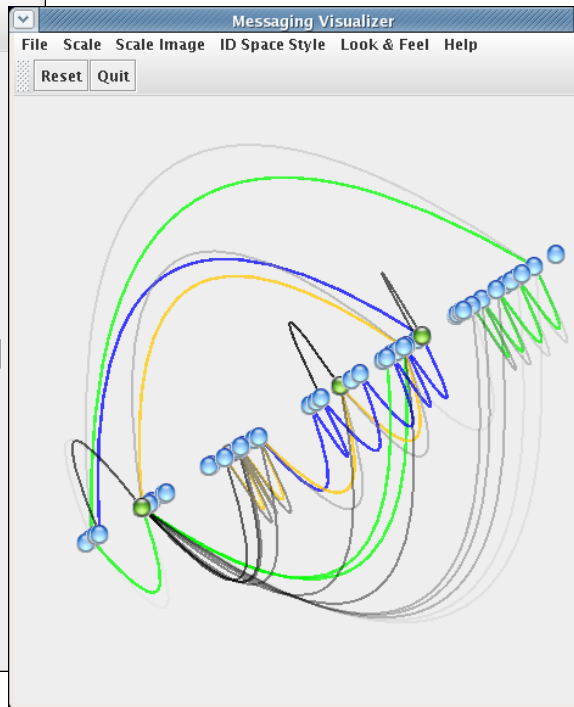
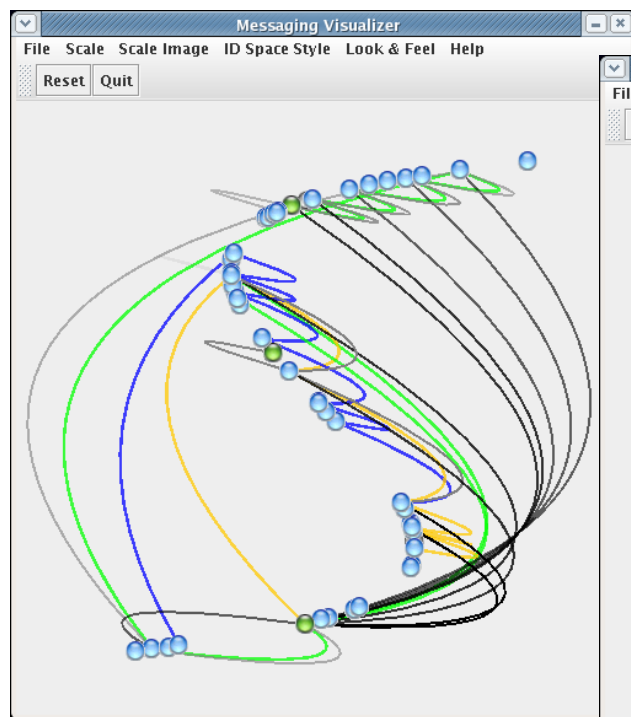


デモ

- 可視化ツール: Overlay Visualizer
- PlanetLab 上の DHT に対する操作

Overlay Visualizer

- ノード, メッセージ, マルチキャストの配送木を、実行中に可視化。
- 実ネットワーク / エミュレータどちらでも動作。



ノードのウェブインタフェース

- ノードについての諸元。

Overlay Weaver Node Status

Node Information

URL:	http://planetlab0.otemachi.wide.ad.jp:3998/
Node ID:	7a17c1434ef5858f0fbfe052a08b318bbb385433
Lookup algorithm:	Chord
Lookup style:	Iterative
# of stored keys:	0

Routing Table

Predecessor

• Chord の経路表。

<http://pub2-s.ane.cmc.osaka-u.ac.jp:3998/> 793bfb7e552e859a63a2eff10d956ebbe678e226

Successor List

http://planetlab1.cs.duke.edu:3998/	7abe663cbc720375b37dd7dcb71e7153aa85da0e
http://planetlab04.cs.washington.edu:3998/	7b638aa42cf93c1f81c8caecdc5e4cf2e4cdabc0
http://planck227.test.ibbt.be:3998/	7bc89d2027953a7bb7c7872290afe025abfce687
http://planet1.l3s.uni-hannover.de:3998/	7f9b10393062377effc2785cf705ab4aad953bcb
http://planet6.berkeley.intel-research.net:3998/	8376275b1547a93b518075bb3af9f53501408455
http://planetlab1.cesnet.cz:3998/	83aa8c5cdd6dd5f9ff441ed6e0e72be2b2c8910e
http://planet1.prakinf.tu-ilmenau.de:3998/	85ad64a1d1ce6cb55b3b1919b01f4f508dca9536
http://planetlab-1.unk.edu:3998/	868c3bbad9adca55d22bdc878c926087f4c7a035

Finger Table

1	http://planetlab1.cs.duke.edu:3998/	7abe663cbc720375b37dd7dcb71e7153aa85da0e
153	http://planetlab-01.kyushu.jgn2.jp:3998/	903f7c72ac22c1bb71ab57a0230c167517d39e96
158	http://planetlab4.cs.uiuc.edu:3998/	a1f4bfcdcb8d56a21ca007de183e3a24d503e3a1
159	http://planetlab02.sys.Virginia.EDU:3998/	bda4cc7207314bf6ea874a04a37d8653954da9ac
160	http://pli2-pa-1.hpl.hp.com:3998/	fd04d9785c525f614ae3384d5f181815d93b827d

ウェブインタフェース

- 分散ハッシュ表 (DHT) への put, get, remove

Put, Get and Remove Operations

operation	key	value	TTL (sec)	secret	
get	foo				submit
put	foo		600	(option)	submit
remove	foo		(option)		submit

ウェブインタフェース

- DHT からの get の結果。

Results

Get results:

key: foo

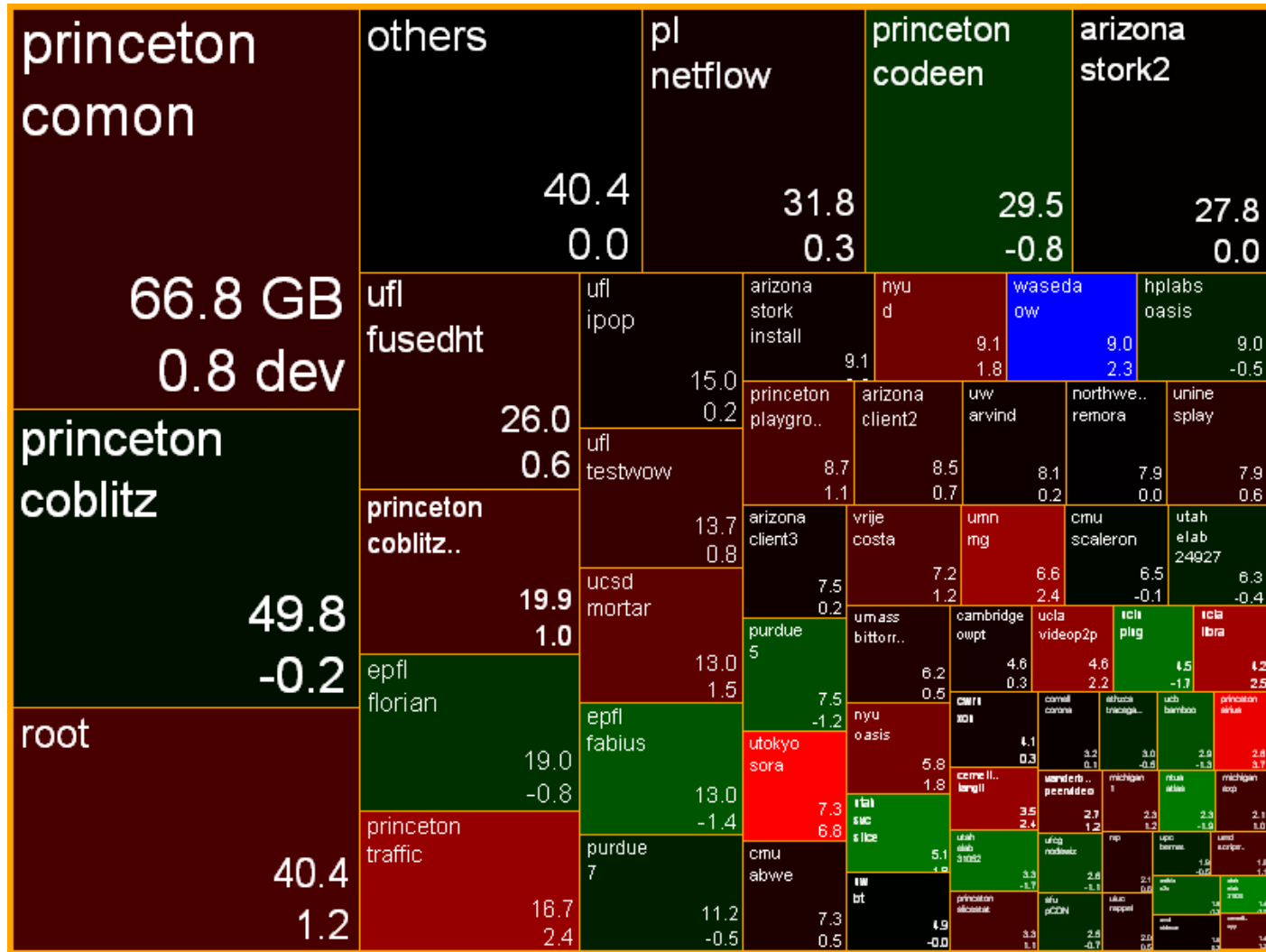
value: bar

Route

Hop	Node	ID	time
0	http://planetlab0.otemachi.wide.ad.jp:3998/	7a17c1434ef5858f0fbfe052a08b318bbb385433	0
1	http://planetlab02.sys.Virginia.EDU:3998/	bda4cc7207314bf6ea874a04a37d8653954da9ac	416
2	http://planetlab2.engr.uconn.edu:3998/	de9aab0095f26bdf3125a0843dd8e0c83676ae38	1245
3	http://planet2.cs.ucsb.edu:3998/	f131d7ec5a7b8b2ce6631506151dcf9af4d8ef16	2074
4	http://planetlab-1.cs.uh.edu:3998/	00e64d0fcab4640596d7d63e73aa0842ce6a49e1	2231
5	http://PLANETLAB-2.CMCL.CS.CMU.EDU:3998/	080d283be8537508023ad9e070b4778bc1cc51b3	2558
6	http://planetslug3.cse.ucsc.edu:3998/	0b25d99513184b13cacc4ca18f43d3079a8f002c	3273
7	http://planetlab0.dojima.wide.ad.jp:3998/	0c66909552301a7d33275cb1716cf418e87b18d3	3715

PlanetLab の資源利用状況

- CoVisualize





オーバレイネットワークへの 期待

センサネットワークの基盤に

- できないか？
- したい！
- ...という人達が盛んに研究を推進中。
- オーバレイの活かし方
 - ストレージ
 - 分散ハッシュ表 (DHT) にセンサデータを格納
 - 範囲検索もしたい → 奈良先端, NEC, ...の研究, SkipGraph, ...
 - ノードが担当地域を動的に調整するオーバレイ by 松浦先生 (奈良先端)
 - データ処理, 計算
 - データ融合 (fusion) - 今後のPIAX (阪大)
 - 分散処理の途中データを DHT に格納して、再利用 (東大)
 - オーバレイをプログラミング
 - E.g. Overlay GHC by 斉藤先生 (慶應)

コミュニティ, 情報交換

- 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

– http://live-e.naist.jp/sensor_overlay2008/

Sensor & Overlay Networks

■ 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

あらゆる場所にセンサが埋め込まれ、その情報をネットワークを通して共有し、誰もが自由にアクセスできる環境が構築可能になりつつある。Live E! プロジェクトでは気象センサを日本およびアジア中心に設置し、その情報を共有している。広範囲なセンサネットワークが構築されるにつれ、オーバレイネットワーク技術を中心とした分散環境の構築が求められている。

第二回を近日開催予定