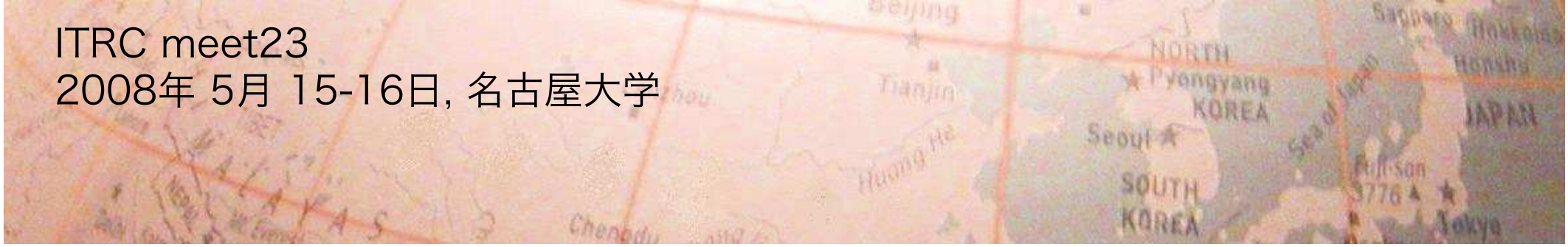


ITRC meet23
2008年 5月 15-16日, 名古屋大学

Overlay Weaver と その PlanetLab 上での運用

首藤一幸
ウタゴエ / NICT

UTA GOE
UNIVERSAL GRID

内容

- Peer-to-Peer コンテンツ配信技術 (ウタゴエ社)
  - ライブストリーミング
  - IPTV

- オーバレイ構築ツールキット
  - Overlay Weaver
    - 産総研グリッド研究センター由来。
      2006年度からは個人で継続。

UTA GOE
UNIVERSAL GRID

Peer-to-Peer ライブ配信

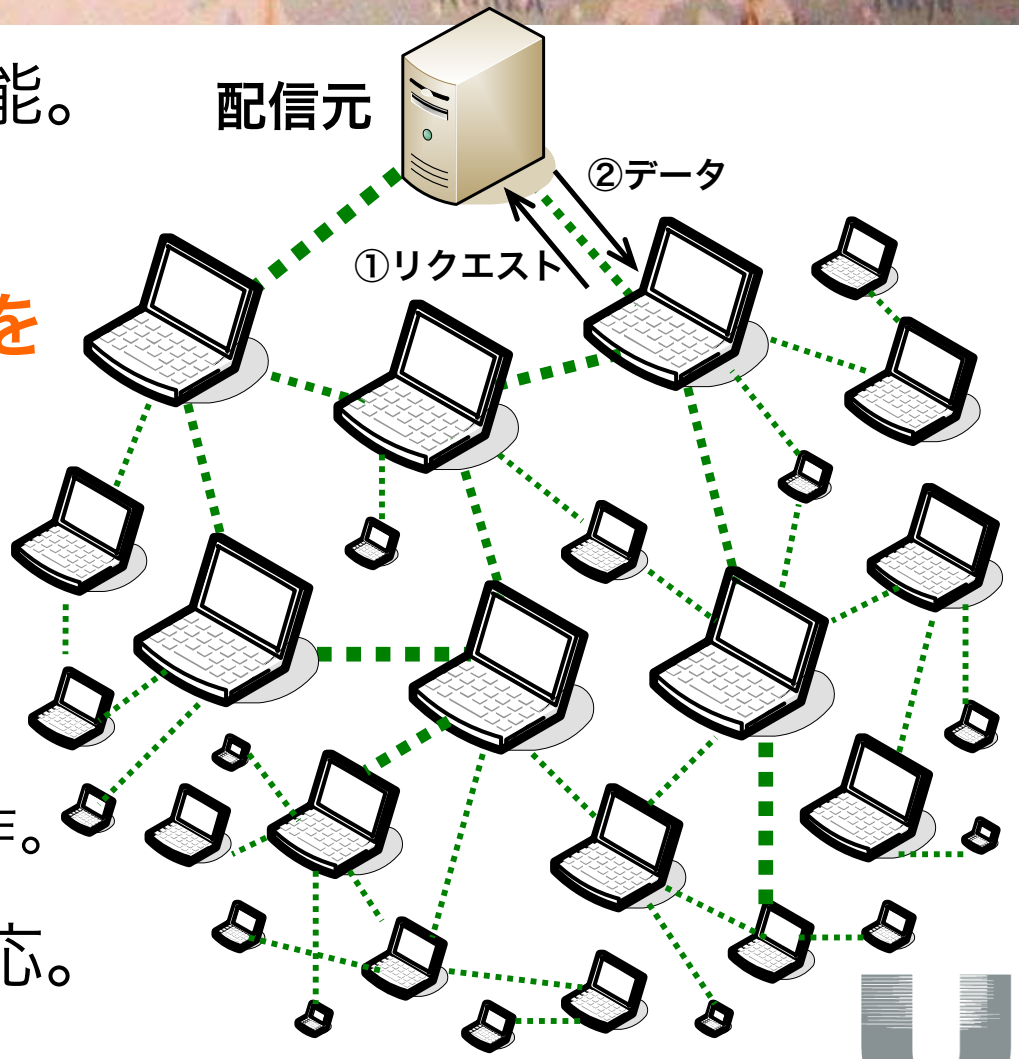
実ネットワークに適応するオーバレイマルチキャスト 放送基盤

- 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



池長慶彦 (早稲田大学 村岡研究室)
首藤一幸 (ウタゴエ (株))

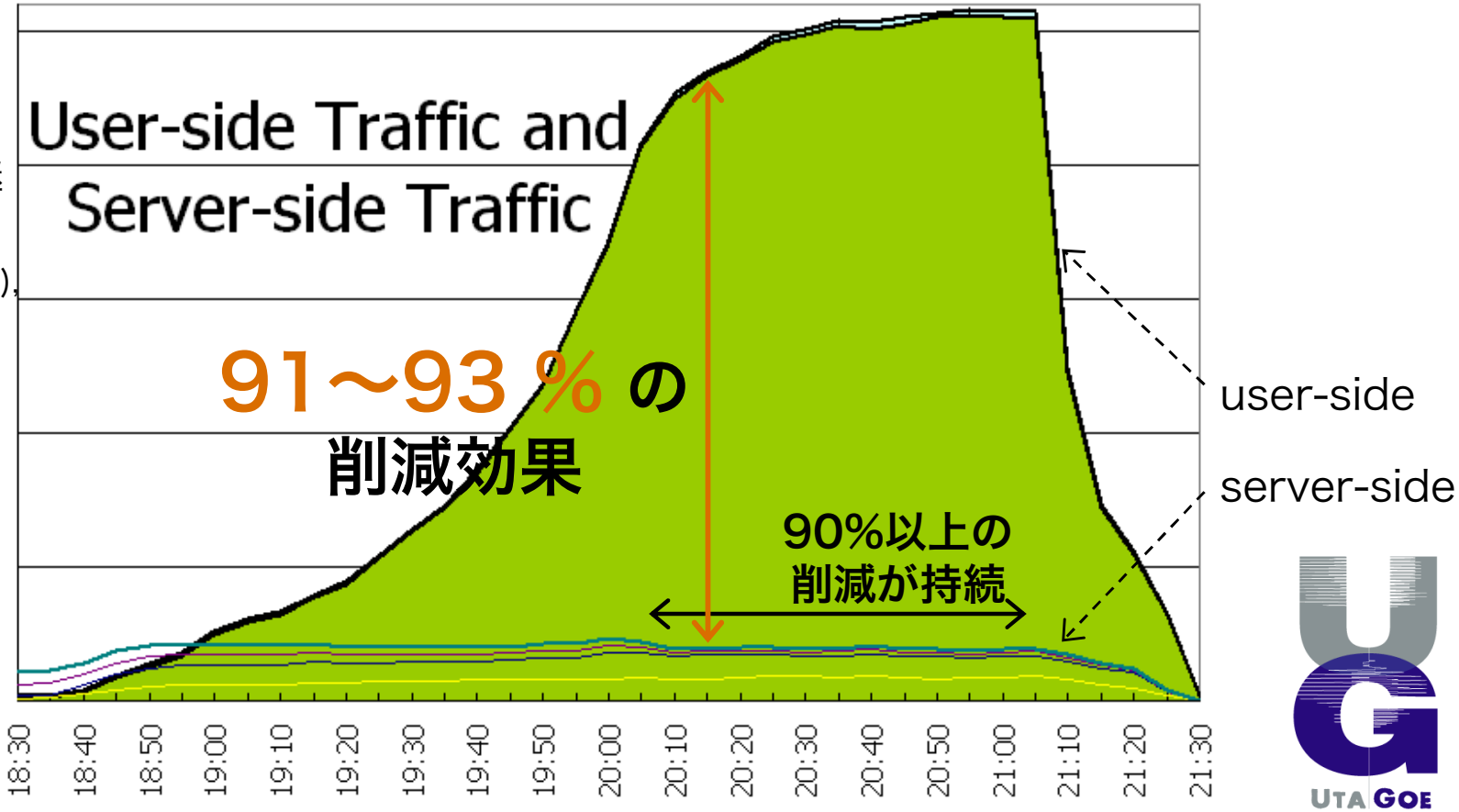
IPA未踏ソフトウェア 2006年度上期 並木PM

UTA GOE UNIVERSAL GRID

トラフィック削減効果

- トラフィック削減効果 90〜95 %
  - ウタゴエ社 UG Live 使用

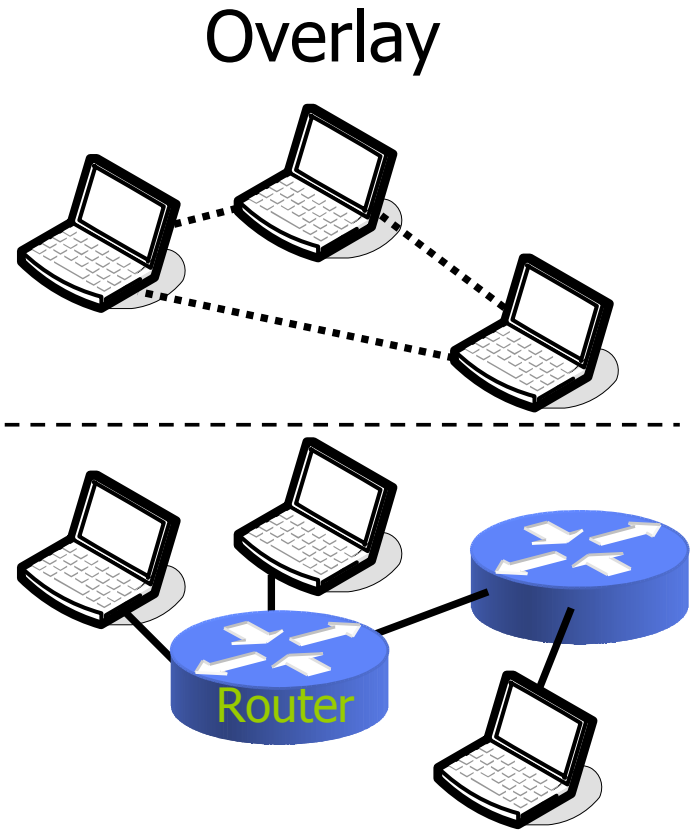
2007年 11月 26日
accessのtalk about生中継
配信: Jストリーム,
Castella (www.castella.jp),
ウタゴエ



UTA GOE
UNIVERSAL GRID

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 ?
      - 多ノード/低管理コスト
        → 自律的にオーバレイを構成



UTA GOE
UNIVERSAL GRID

Unstructured と Structured

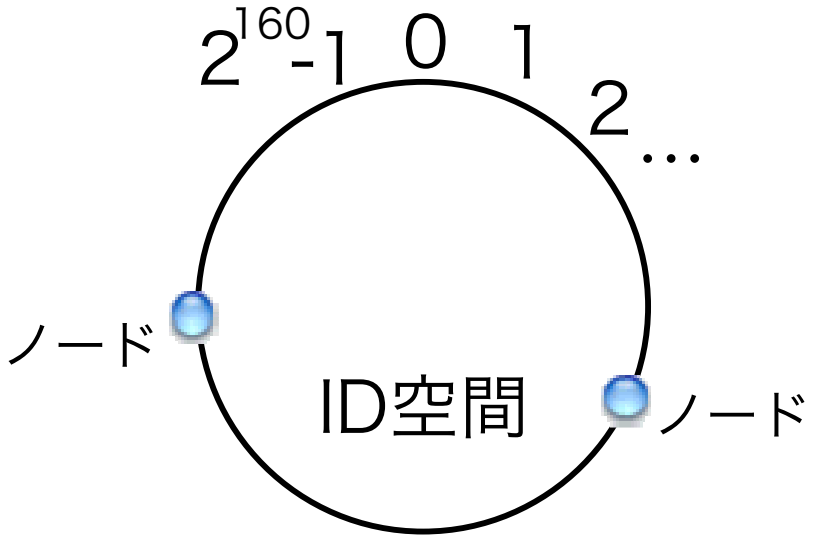
- 非構造化 (unstructured) オーバレイ
  - 例： Gnutella ネットワーク, Winnyネットワーク
  - 誰を隣接ノードとするか、トポロジに制約がない。
  - 存在するオブジェクトは、発見できる可能性がある。
  - 一般に、効率は良くないが、柔軟な検索が可能。

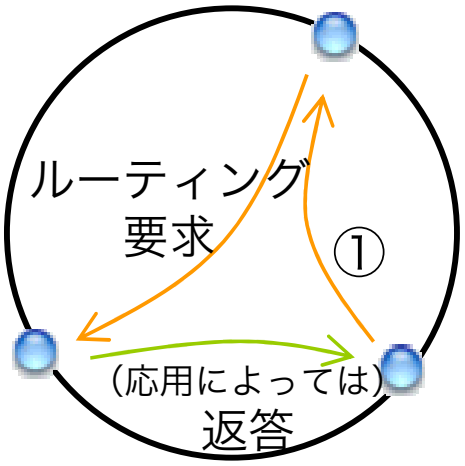
- 構造化 (structured) オーバレイ
  - 例： DHT（分散ハッシュ表）のネットワーク
  - 誰を隣接ノードとするか、トポロジに制約がある。
  - 存在するオブジェクトは、（たいてい）発見できる。
  - 一般に、効率は良いが、柔軟な検索が苦手。

UTA GOE
UNIVERSAL GRID

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。

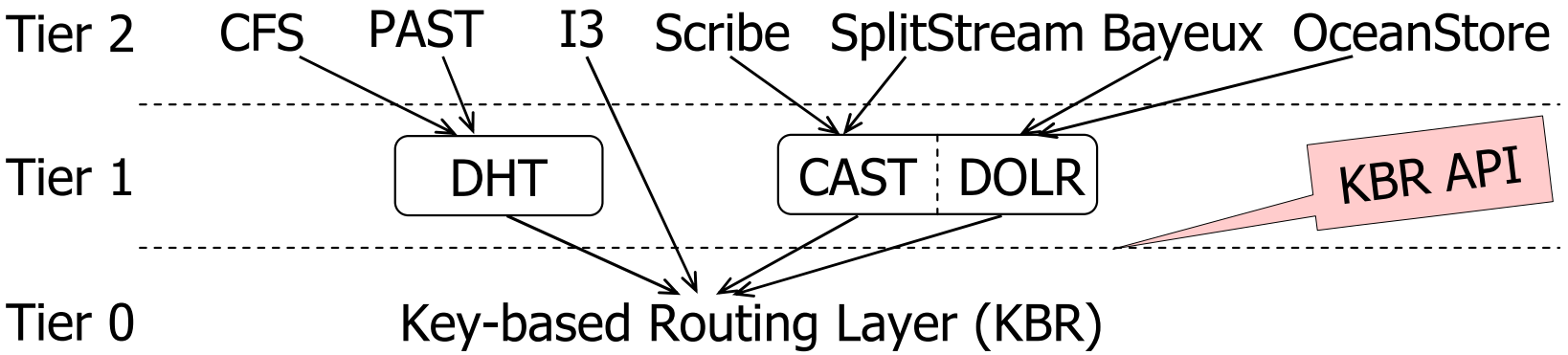




構造化オーバレイの基本

- 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. unstructured オーバレイ上の 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。

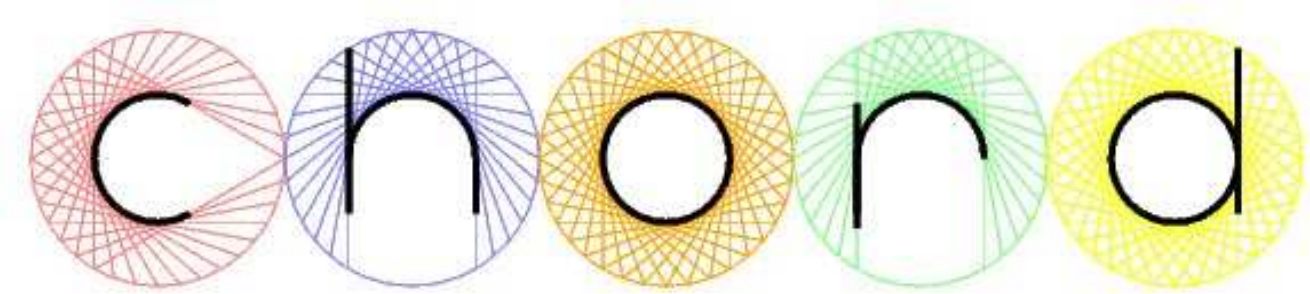
UTA GOE UNIVERSAL GRID

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

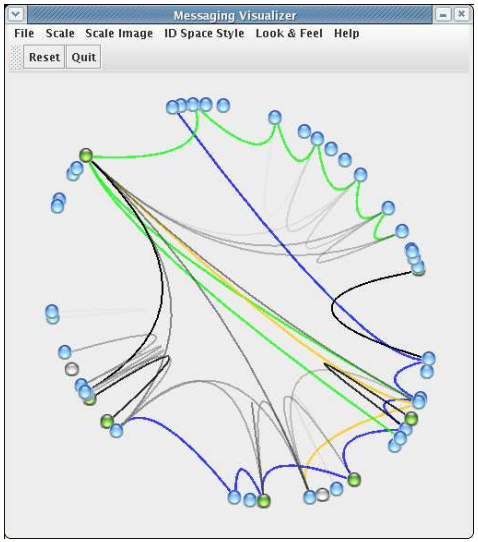


Tapestryのロゴ

Overlay Weaver

- 構造化オーバレイ・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



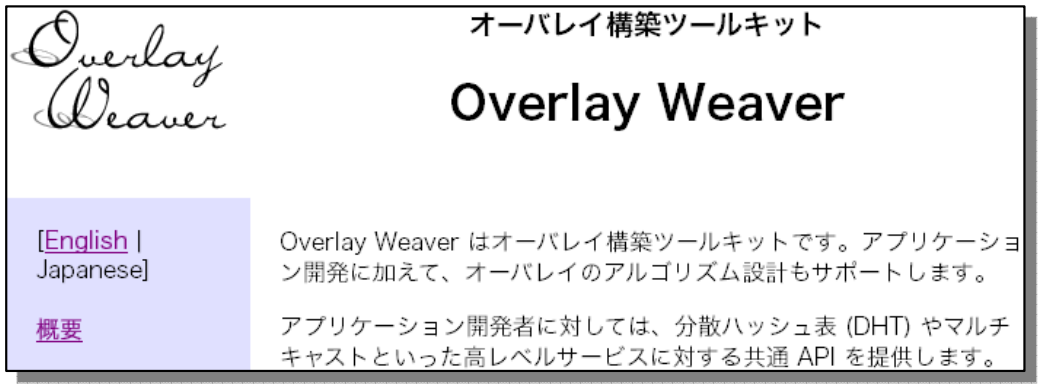
- 特徴
  - 分散ハッシュ表 (DHT) だけじゃない。アプリケーション層マルチキャストも。
  - ルーティングアルゴリズムを差し替え可能：
    Chord, Kademlia, Koorde, Pastry, Tapestry
  - コーディングなしで運用や実験が可能。
    - サンプルツール (DHTシェル等) を使った運用。
    - エミュレータを使った実験: メッセージ数やホップ数の計測。PC 1台で 70,000ノード動作。
  - 動作状態を可視化して楽しめる。デモできる。
  - XML-RPC 経由で DHT を使える。
    - Bamboo, OpenDHT 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。
    - 35ヶ国 418台

UTA GOE UNIVERSAL GRID

オープンソースソフトウェア, Overlay Weaver 研究プラットフォームとして

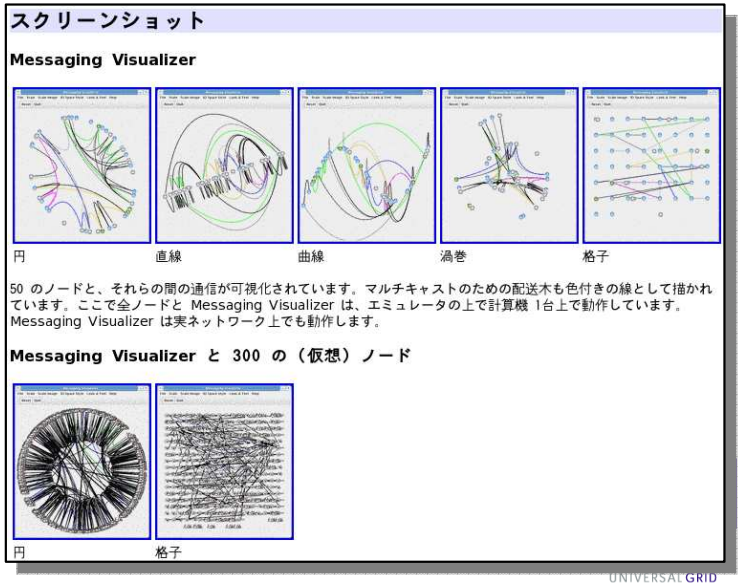
- http://overlayweaver.sf.net/ (SourceForge)
  - 2006年 1月 17日、リリース。
  - Apache License 2.0

- 状況 (2008/5/17 23時)
  - ダウンロード 9,500件。
  - メーリングリスト登録数
    - 英語 71名, 日本語 90名
  - Mixi コミュニティのメンバ数 152名



ウェブサイト

- 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), …



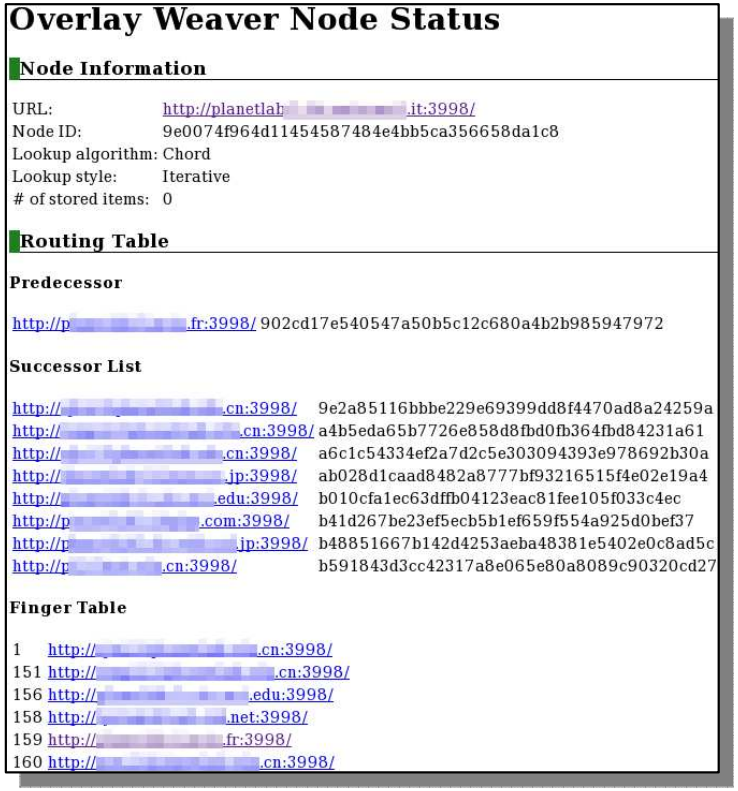
研究プラットフォームとして

- 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, “構造化オーバレイでの一括ルーティング”, OS研究会 (SWoPP), 2008年 8月.

UTA GOE
UNIVERSAL GRID

PlanetLab 上での運用

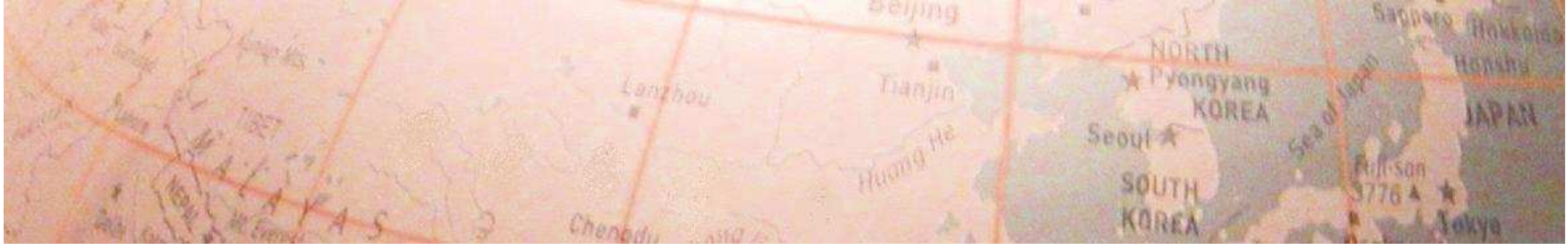
- 実機上、できればインターネット上で、実際に大規模な試験・運用を。
- PlanetLab: 地球規模テストベッド
  - PC 2台を供出すれば、使える。
    - プロジェクトごとにVM (Linux-VServer) を用意できる。
  - ただし、企業の fee は $25,000 / 年～！
  - 客員研究員として、大学を PlanetLab に参加させた。
  - 最大で、35ヶ国 418台で動作。



PlanetLab上で運用しているノードのウェブインタフェース

PLANETLAB
An open platform for developing, deploying, and accessing planetary-scale services

UTA GOE UNIVERSAL GRID



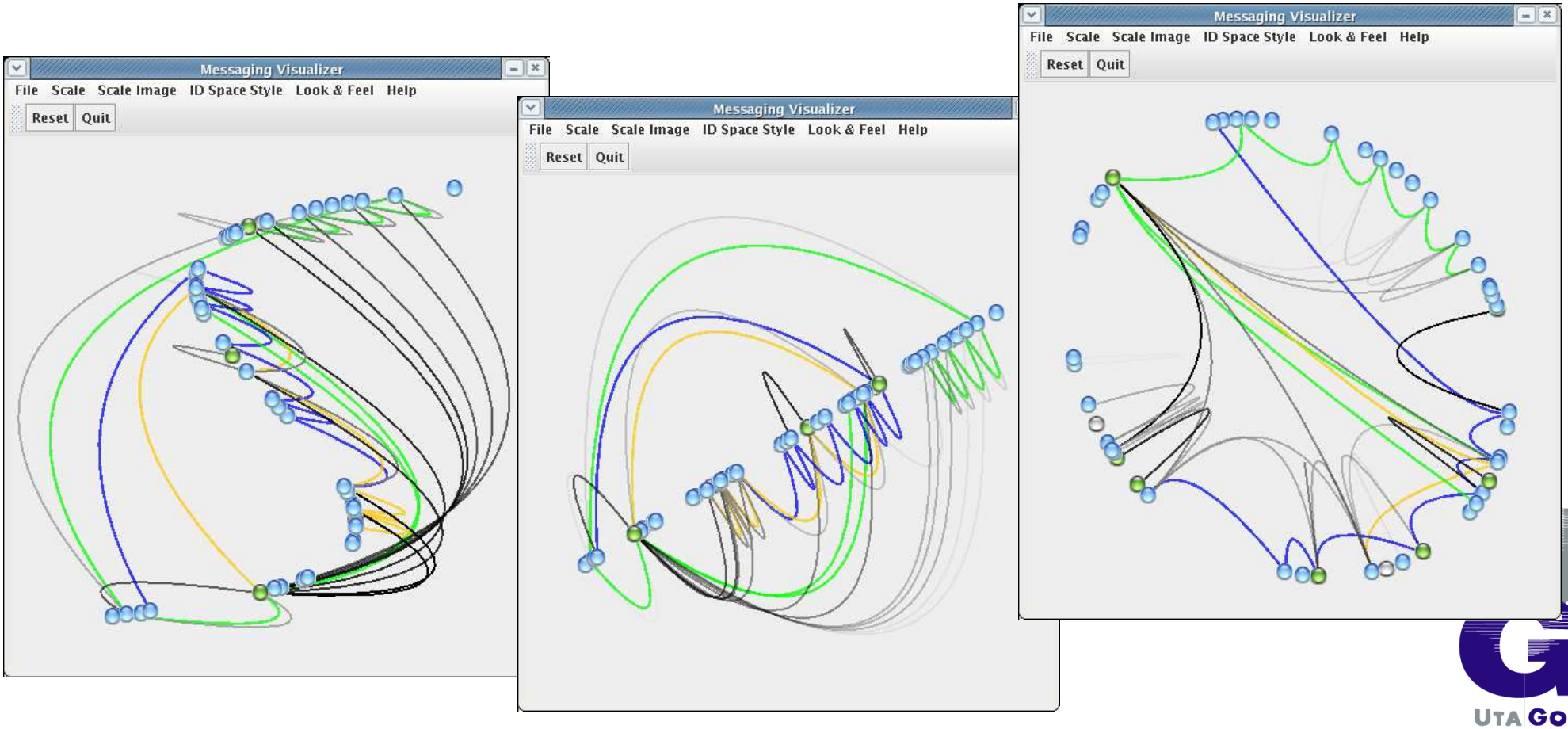
デモ

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

UTA GOE
UNIVERSAL GRID

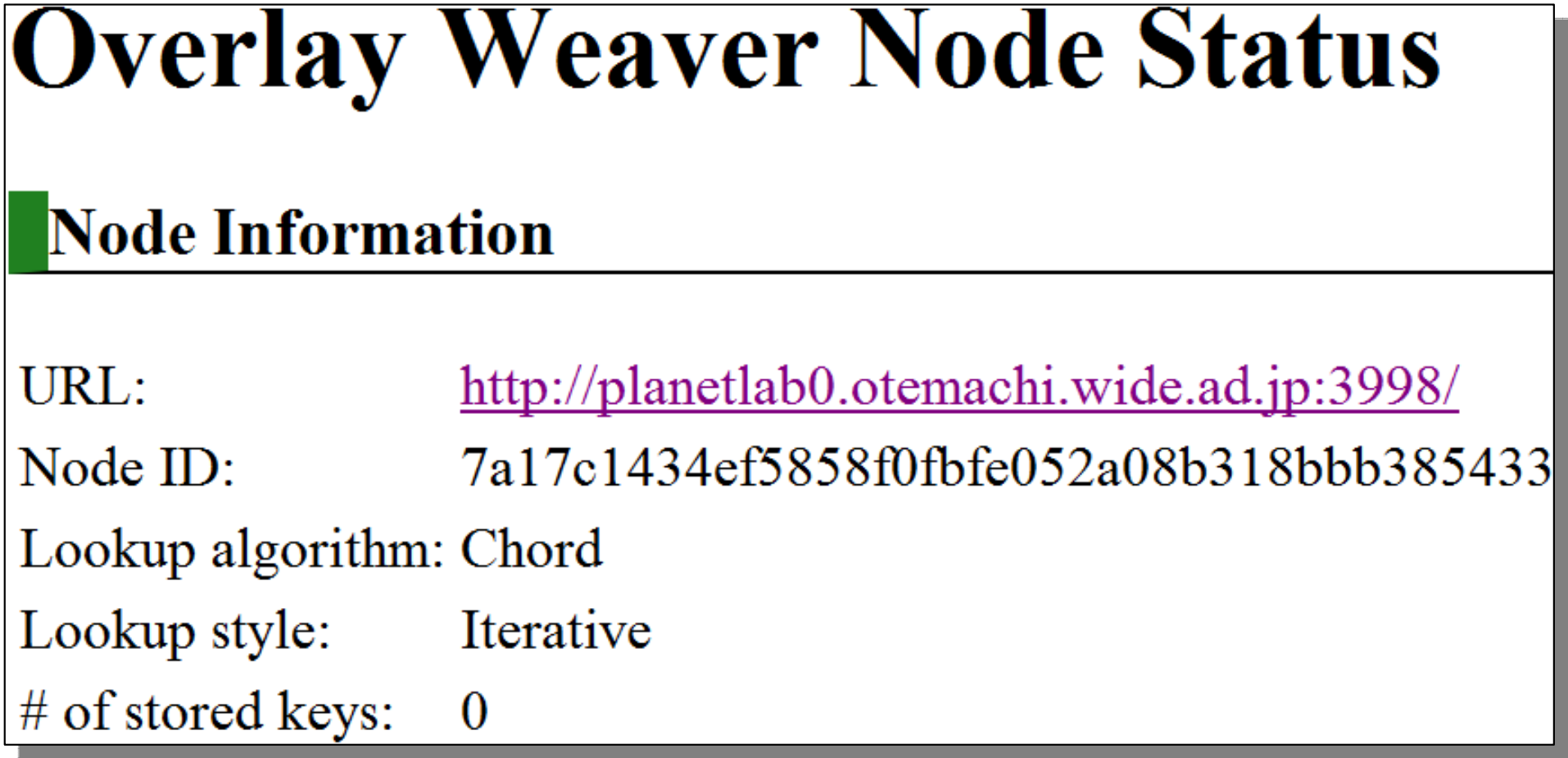
Overlay Visualizer

- 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

- Chord の経路表。

Predecessor

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

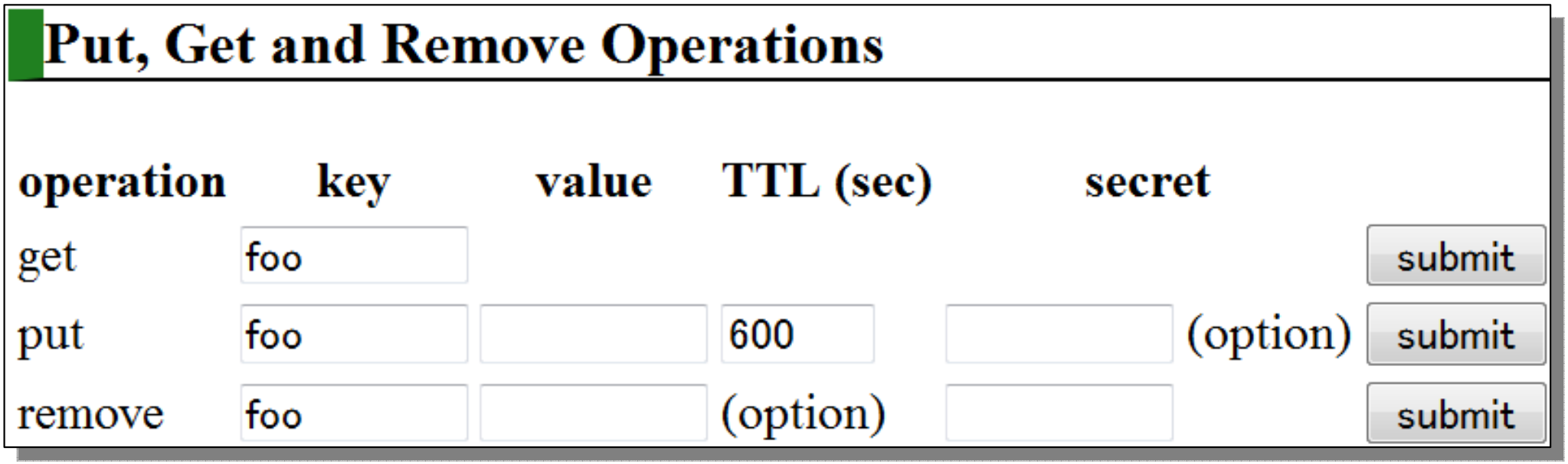
Finger Table

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

UTA GOE UNIVERSAL GRID

ウェブインタフェース

- 分散ハッシュ表 (DHT) への put, get, remove



UTA GOE UNIVERSAL GRID

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