



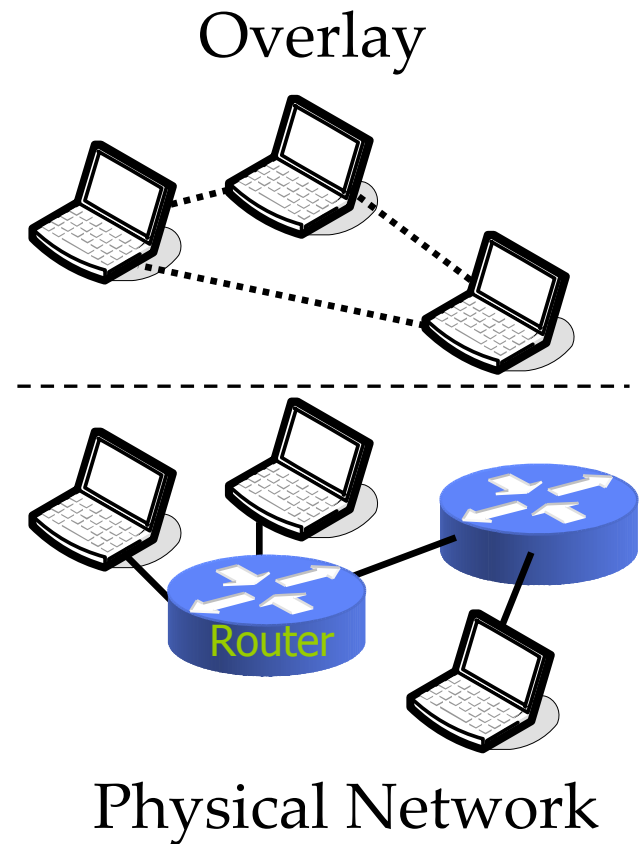
Structured and Unstructured Overlay Networks

Kazuyuki Shudo / 首藤 一幸, Ph.D.

Tokyo Institute of Technology (Tokyo Tech)

Overlay

- An network constructed over another network
 - e.g. Internet over a telephone network
 - e.g. Networks of file sharing programs
 - FastTrack, eDonkey2K, Gnutella, ...
 - with more than 1,000,000 nodes.
- Application-level network
 - constructed in a **autonomic and decentralized** way to keep **performance and fault-tolerance** with 1,000s and 1,000,000s nodes.
- Its topology is independent from the underlying network (i.e. Internet).
 - Then, called **Overlay Network** or **Network Overlay**.






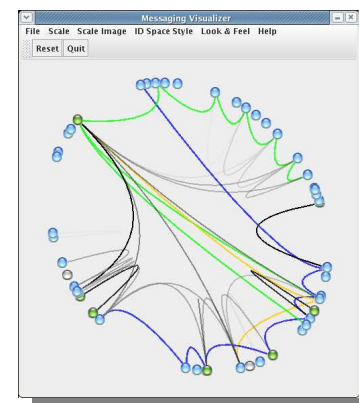
Unstructured and Structured

- Unstructured overlay
 - e.g. **Gnutella** network, Winny network
 - **No (few) constraint** imposed on which node to be a neighbor (topology).
 - An existing object on it may be found.
 - Generally, less-efficient but supports flexible (e.g. range) search.
- Structured overlay
 - e.g. A network for **distributed hash table (DHT)**
 - **An algorithm-based constraints** imposed on which node to be a neighbor.
 - An existing object on it is (almost) certainly found.
 - Generally, efficient but it has been said to be weak in flexible search.

Overlay-related activities

I will demonstrate today

- **Peer-to-peer live streaming**
 - Application-level/layer multicast (ALM), Overlay multicast (OM)
 - Started at Utagoe in 2006
 - **Unstructured** overlay network
 - **Overlay Weaver**: an overlay construction toolkit
 - Started at AIST in 2005
 - **Structured** overlay network
- 





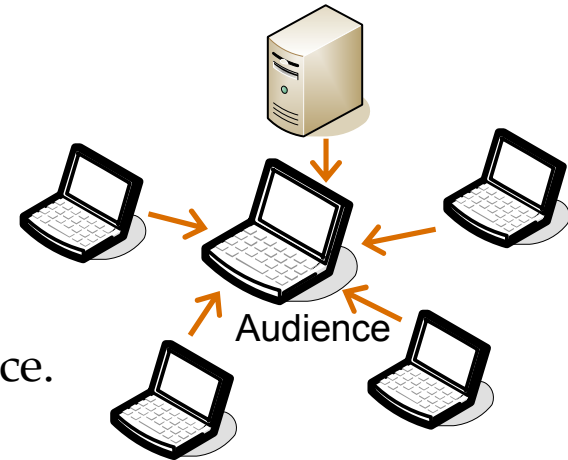
Peer-to-peer live streaming

An example of **unstructured** overlay

P2P Content Delivery

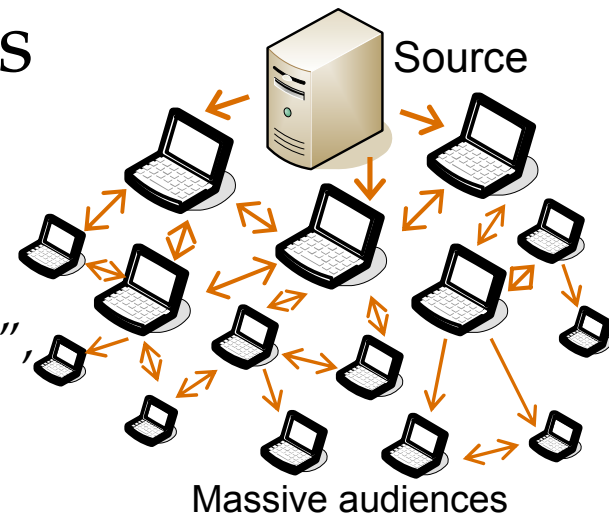
- Gathering technologies

- Download
- On-demand streaming
 - Playing while downloading.
- Gather parts of contents replicated in advance.
- “Swarming”



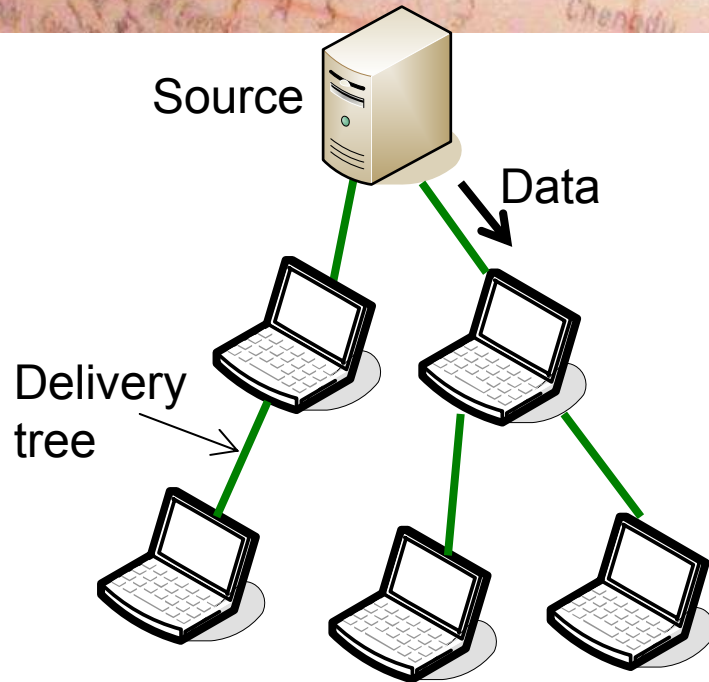
- Disseminating technologies

- (Live) streaming
- Deliver content to massive number of audiences in a short time.
- “Application-{level, layer} Multicast (ALM)”, “Overlay Multicast”, “Endsystem multicast (ESM)”



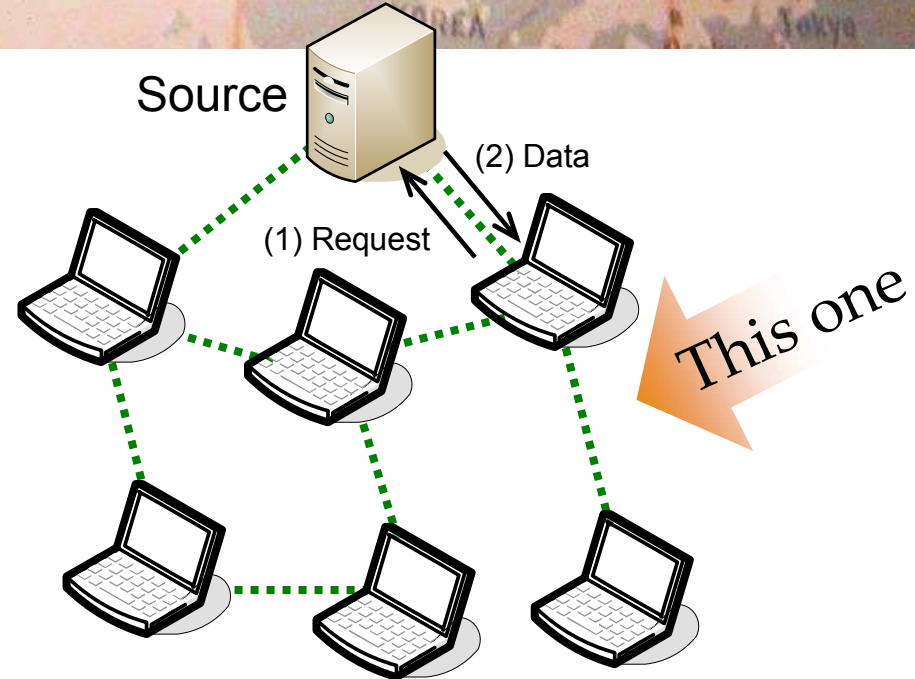
Dissemination: ALM, OM, ESM

Tree-based vs. Mesh-based



- Tree-based

- Data flow along the delivery tree constructed explicitly.
- **Push** from the root toward leaves.
- **△** Requires quick repair in node failure
- **○** Data reach all ends, with low latency.
- Exploits broadband links.



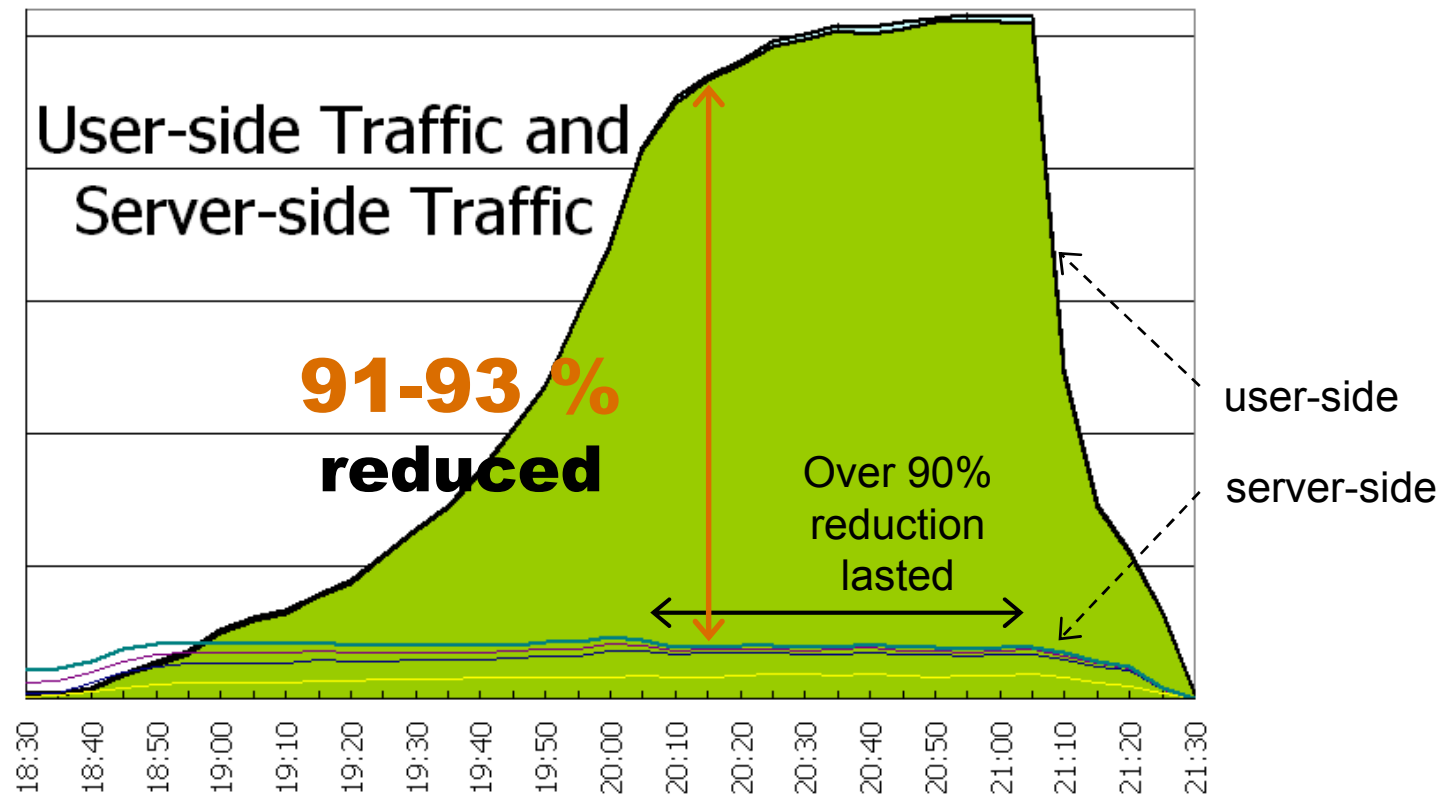
- Mesh-based

- Keeps loose relationships with neighbors.
- **Pull** data from neighbors.
- **○** Robust to node failure by nature
- **△** Delivery to ends not guaranteed. To be compensated.
- Exploits narrowband links.


(Live) streaming

- From 90 to 95 % of traffic reduced around the source
 - With Utagoe's UG Live software


On Nov 26, 2007,
“access talk-about live”
by J-Stream,
Castella (www.castella.jp),
and Utagoe



Application



ウタゴエ株式会社




日経CNBC 視聴ページ

日経CNBCをパソコンで視聴できます

IP LABO
同時再送信実験

日経CNBC 視聴ページ



CMの後は...
チャート分析

芝 700 +27 (2691万株) 7位: ユニー 958
REUTERS ロンドン FTSE100 5650.2 +155
NIKKEI CNBC

43000 Version: 1.3.6 (P)
ch.og20000

再生セグメント数: 993
再生失敗セグメント数: 0
Streaming Rate: 599.8 kbps
上クスループット: 0.0 kbps
下クスループット: 720.0 kbps
総送信量: 9637 KB
総受信量: 48327 KB
GAddr: 219.179.114.147: 3000
network pass: true
lower neighbor size: 2
upper neighbor size: 285
316.0 kbps



サブプライム問題

★★★ただ今放送中の番組
REUTERS ナスダック総合指数 2319.42 -7.33
NIKKEI CNBC

Nikkei CNBC broadcasting (from March 24 to 28, 2008)
delivered by TV Tokyo Broadband Inc. and Utagoe Inc.

Simultaneous with TV

Application



CD quality music

Radio broadcasting (from June 2 to 30, 2008)

delivered by NIPPON BROADCASTING SYSTEM, Impress Imageworks, J-Stream and Utagoe



Overlay Weaver:

An Overlay Construction Toolkit

An example of **structured** overlay

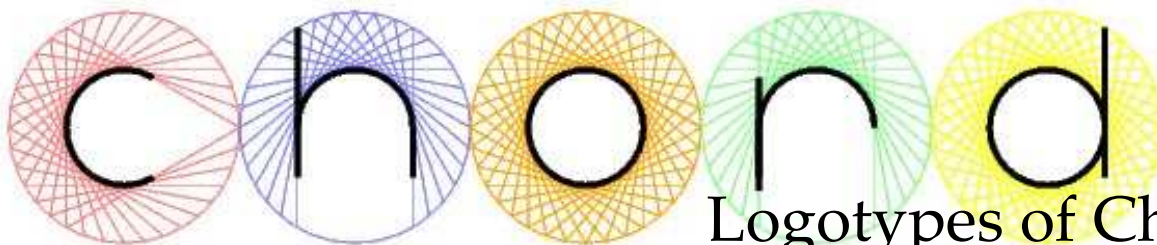
Overlay Weaver

Overlay Weaver

- We see analogies between structured overlays and weaving.
 - Chord, Tapestry, ...
- Overlay Weaver
 - A weaving device of (structured) overlays



Weaver



Logotypes of Chord and Tapestry



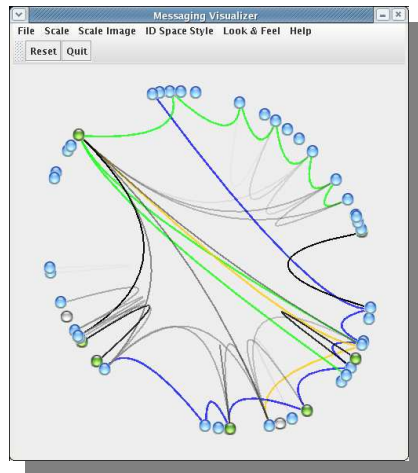
Overlay Weaver

- DHT Library

- in Java
- about 23,000 steps
- licensed under Apache License 2.0
 - ready to be applied to various purpose.

- Properties

- supports application-layer/level multicast (ALM), not only DHT.
- supports multiple routing algorithms:
 - Chord, Kademlia, Koorde, Pastry, Tapestry, ...
- We can conduct experiments without writing code.
 - Operation with sample tools such as DHT shell and Mcast shells.
 - Measurement of # of messages, # of hops and ... with Emulator.
- 150,000 nodes on a single PC.
- A DHT is accessible via XML-RPC-based protocol.
 - the same protocol as Bamboo and OpenDHT.



Overlay
Visualizer

Overlay Weaver as an Open Source Software

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

- released on 17th Jan, 2006.
- Apache License 2.0

- Statistics (as of 10:10, 5th Dec, 2008)

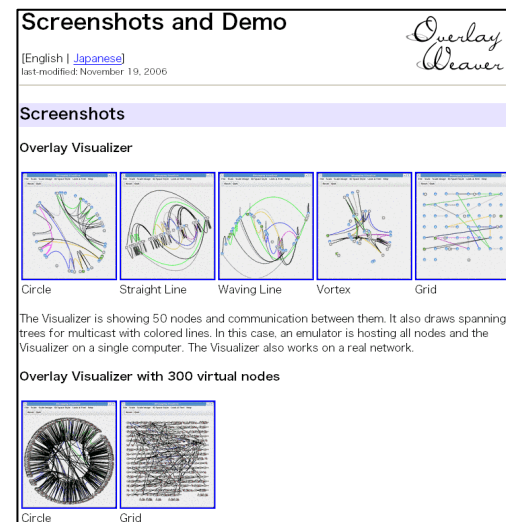
- 11,420 downloads
- # of subscribers of mail lists
 - English: 80, Japanese 86
- # of members of the mixi community: 152

- As a research platform

- Used in **Brazil, Russia, UK, Hungary, ...**
- **Third parties implemented routing algos:**
Symphony, EpiChord, ...
- **Many use cases:**
RDF DB (AIST), XML search (Tsukuba U.), # of replicas (U Federal do Rio Grande do Sul), Search over multiple servers (Ochanomizu U.), Web access investigation (TITECH), Replica placement (Waseda U.), User mgmt server (Hitachi), Multi-attribute range search (NEC), ...



Web site






As a research platform

- Evaluation and improvement of **churn tolerance**
 - Shudo, “Churn Tolerance Improvement Techniques in an Algorithm-neutral DHT”, IPSJ ACS journal, March 2008.
- **Collective forwarding**
 - Shudo, “Collective Forwarding on Structured Overlays” (written up)

Operation on PlanetLab

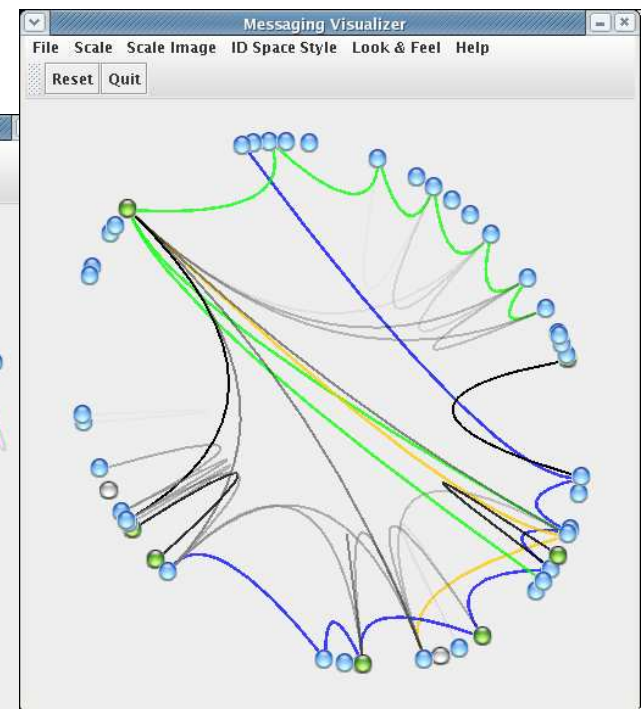
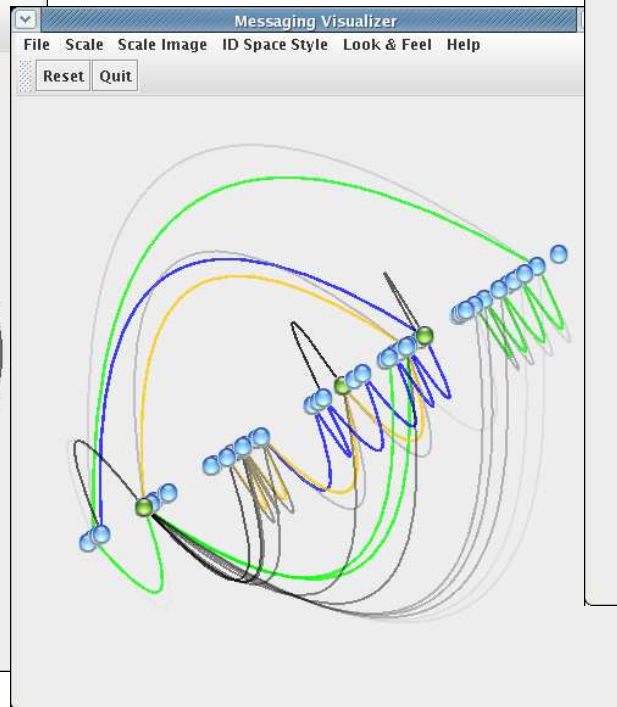
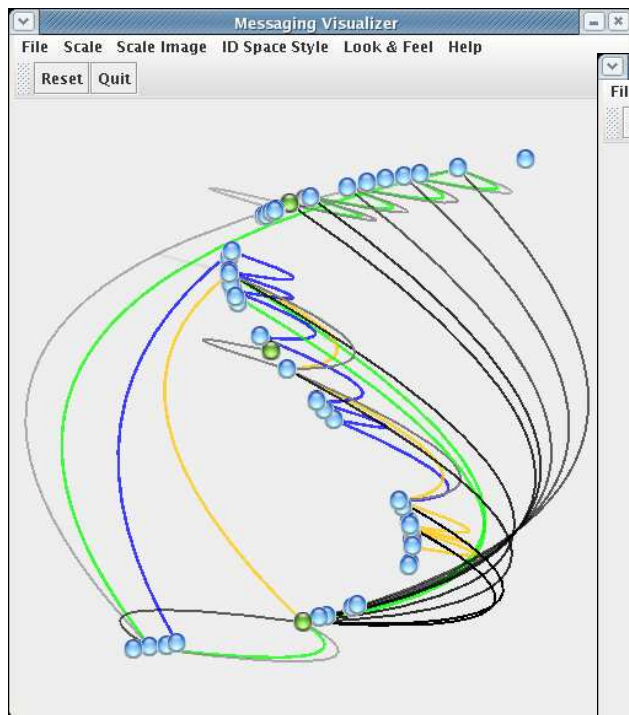
- Real experiments are important
 - As a proof of ...
 - Just for fun ☺
- 
- PlanetLab: a globe-scale testbed
 - Currently consists of 931 nodes at 452 sites
 - An user operates 2 PCs and can use the testbed
 - 561 Overlay Weaver nodes in 35 countries (at most).

Overlay Weaver Node Status	
Node Information	
URL:	http://planetlab.it: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/	

Web interface of a node

Overlay Visualizer

- Shows nodes, message deliveries and delivery tree for multicast on the fly.
- Works both on a real network and a distributed environment emulator.



Future work

- Unstructured overlays
 - Are technologies to select “near” peers consistent with ISPs and carriers’ interests?
 - e.g. P4P, IETF ALTO WG
 - Utilizes routing information like AS numbers
 - Install-less peer-to-peer systems
- Structured overlays
 - Cloud backends with structured overlays
 - e.g. Microsoft’s Azure Services Platform
 - In-memory cache, pub-sub support, distributed computation
 - ID/Locator separation with structured overlays
 - e.g. i3, HIP with DHT, PIAX group’s activity



Spare slides

Web interface of a node

- A node specification

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

• A routing table for Chord

<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

Web interface

- Put, get and remove an item to a distributed hash table (DHT)

Put, Get and Remove Operations

operation	key	value	TTL (sec)	secret	
get	<input type="text" value="foo"/>				<input type="button" value="submit"/>
put	<input type="text" value="foo"/>	<input type="text"/>	<input type="text" value="600"/>	<input type="text"/> (option)	<input type="button" value="submit"/>
remove	<input type="text" value="foo"/>	<input type="text"/>	(option)	<input type="text"/>	<input type="button" value="submit"/>

Web interface

- A result of a get request to a DHT

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

Resources utilization of PlanetLab

- CoVisualize

