PLANETLAB

And Network Virtualization

Timothy Roscoe
Intel Research at Berkeley

Tuesday, May 10, 2005

Desculpa

- Perdão a todos, mas esse é o único slide em português.

Overview

- PlanetLab history
- Concepts and Architecture
- What is it used for?
- Unblocking Network Architecture research
  - Overlay networks
  - Network Virtualization
- Conclusion

PlanetLab is…

- Large collection of machines spread around the world for distributed systems research
- Focus/catalyst for systems and networking community
- Intel project ⇒ consortium of companies and universities

The value proposition

- “Collectively owned”
- Institutions join, provide nodes
  - IA32 architecture servers
  - Hosted outside the firewall
  - Provide power, cooling, & bandwidth
- In exchange, researchers get to use a small “slice” of many machines worldwide.

Origins: wide-area distributed systems research c.2002

- Researchers had no way to try out real systems
  - Architectures, simulations, emulation on large clusters, calling 17 friends before the next deadline…
- but *not* the surprises and frustrations of experience at scale to drive innovation
- How can research systems be validated?

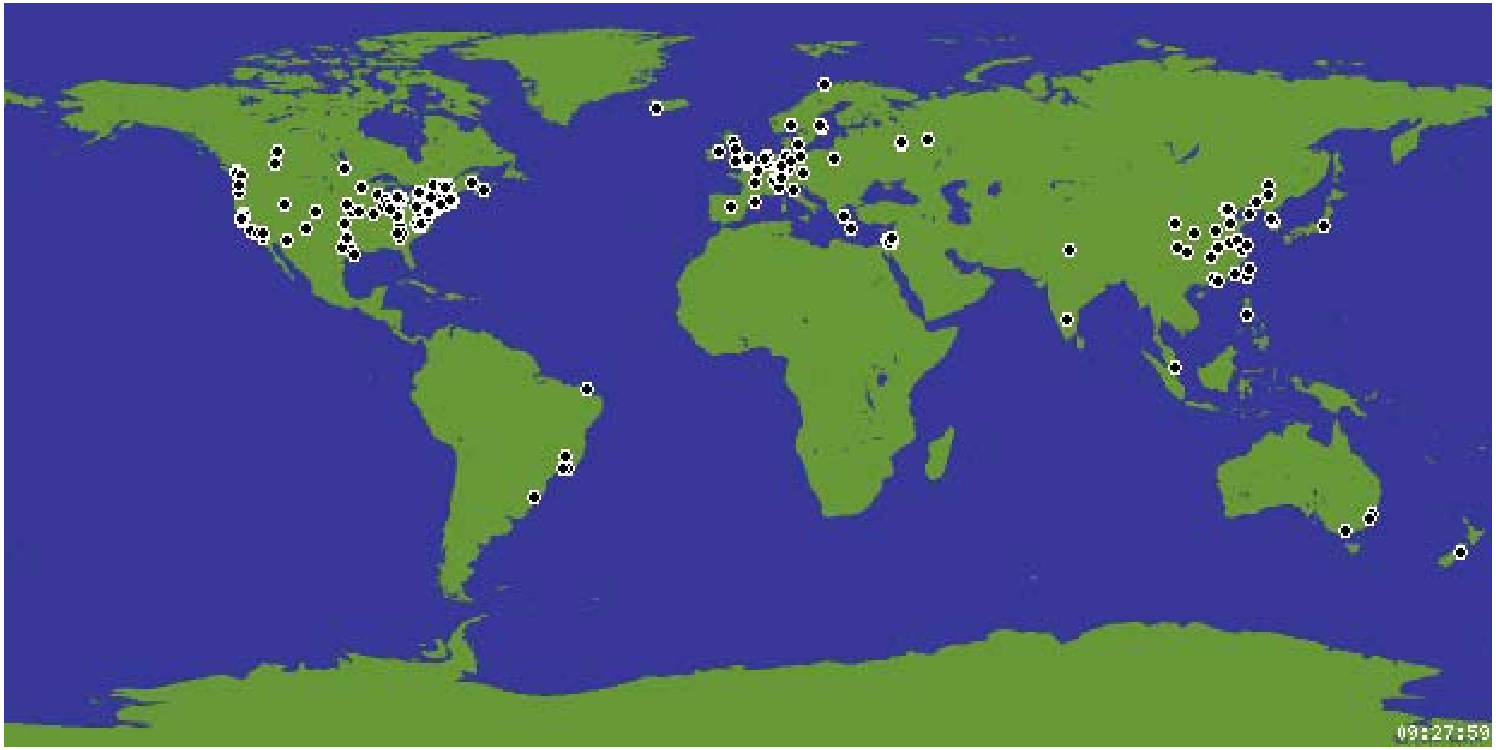
Origins: large-scale networking research c.2002

- Strong feeling the Internet had ossified
  - Intellectually, infrastructure, etc.
  - NRC "looking over fence at networks"
- New ideas abandoned as undeployable
  - Overlays as a way out of the impasse
  - Next internet emerges as overlay (again)
- How can researchers deploy overlays?

Early timeline

- David Culler and Larry Peterson discuss initial idea early 2002
- "Underground" meeting at IRB March 2002
- Position paper (Anderson, Culler, Peterson, Roscoe) June 2002.
- Intel seeds project, core team, 100 nodes
- First node up July 2002
- By SOSP (deadline March 2003) ~25% of accepted papers refer to PlanetLab
- Large presence at SIGCOMM
- 11 out of 27 papers in NSDI 2004

PlanetLab today



About 560 nodes, 269 sites, 30 countries, 5 continents
Universities, Labs, POPs, CoLos, DSL lines
Huge presence in research conferences
Several thousand researchers, students, faculty

PLANETLAB

Intel Research Berkeley

The PlanetLab Consortium

- Modelled on the W3C
- Run by Universities
  - U. Washington, U.C. Berkeley, U. Cambridge, Princeton U.
  - Based in Princeton, NJ, USA.
- Funded by Industry and Govts.
  - NSF, EU, Cernet, etc.
  - Intel, HP, Google, AT&T, FranceTelecom, etc.

The PlanetLab Consortium

- Node resources provided by member institutions
- Small "support" team NOC in Princeton
  - Additional NOCs planned in Europe (Paris), China (Tsinghua)
- Steering Committee
  - University representatives
  - Top-level industrial sponsors

PlanetLab Architecture

Timothy Roscoe, 10 May 2005

Short-term requirements

(March 2002)

To support current research work in distributed & P2P systems and networking:

- Shared by many simultaneous users
- Isolation and protection
- Use familiar API (Linux)
- Networking flexibility
- Manageable
- Must be fully operational in 3 months!

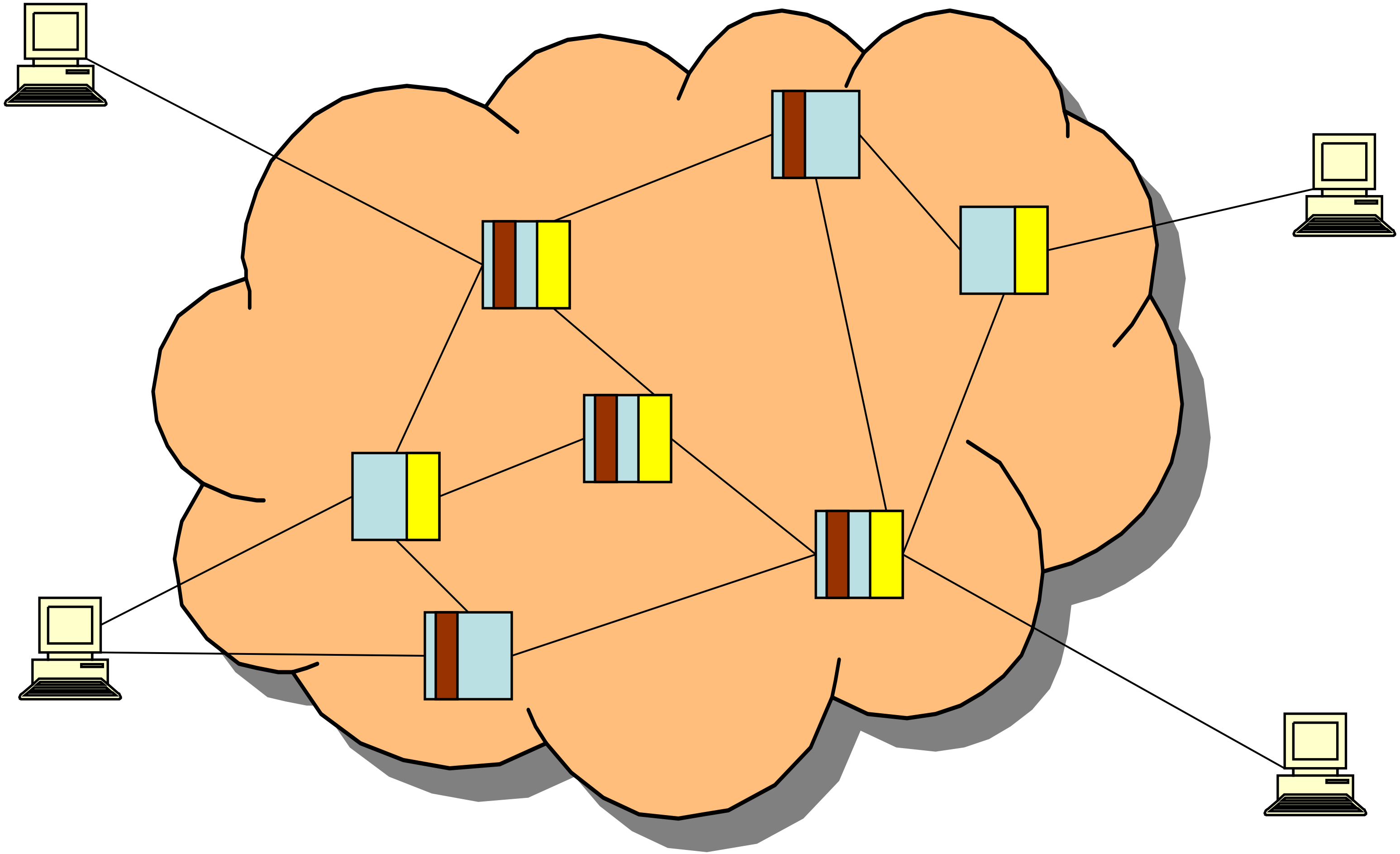
Long-term requirements

- To change the world by incubating the next Internet:
  - Must evolve over time
  - Community replaces all functionality (including O/S)
  - Allow parallel approaches to coexist
  - Produce a viable replacement for the existing network and services

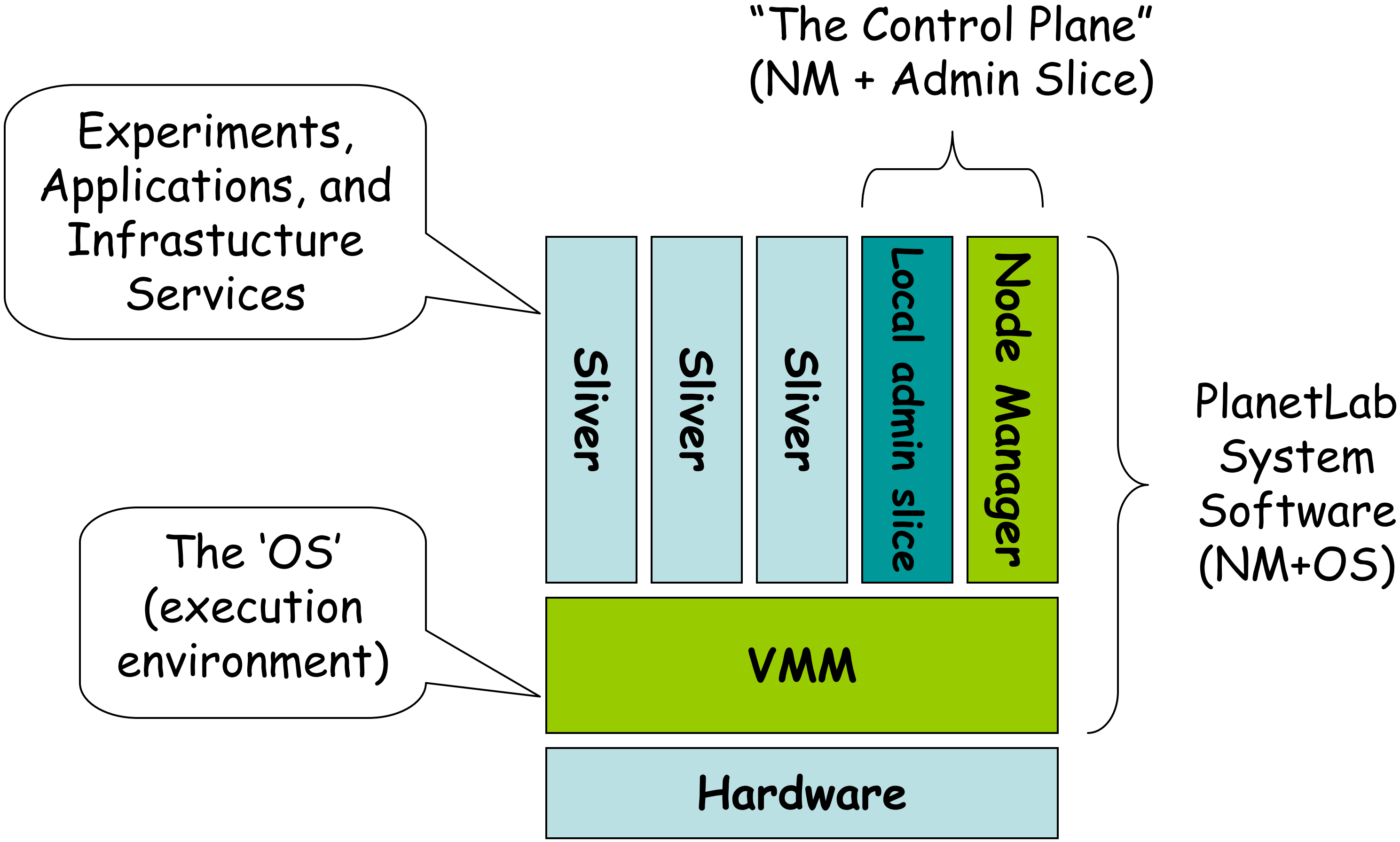
Distributed Virtualization

- *Slices*
  - Basic unit of isolation and sharing
  - Distributed set of virtual machines (slivers)
  - Services & applications run "in" slices
- *Nodes*
  - Physical machines, grouped into *Sites.*
  - One node hosts many slivers
- Infrastructure Services
  - Provide functionality to developers or other services rather than users

Slices



Node architecture


"The Control Plane"
(NM + Admin Slice)
Experiments, Applications, and Infrastucture Services
The 'OS' (execution environment)
Sliver
Sliver
Sliver
Local admin slice
Node Manager
VMM
Hardware
PlanetLab System Software (NM+OS)

Uses and Lessons

What is PlanetLab good for?

- *Planetary-Scale* applications:
  - **Low latency** to widely spread users
  - **Span boundaries:** jurisdictional and administrative
  - **Simultaneous viewpoints:** on the network or sensors
  - **Hardware deployment** is undesirable

What do people use it for?
(a few we know about)

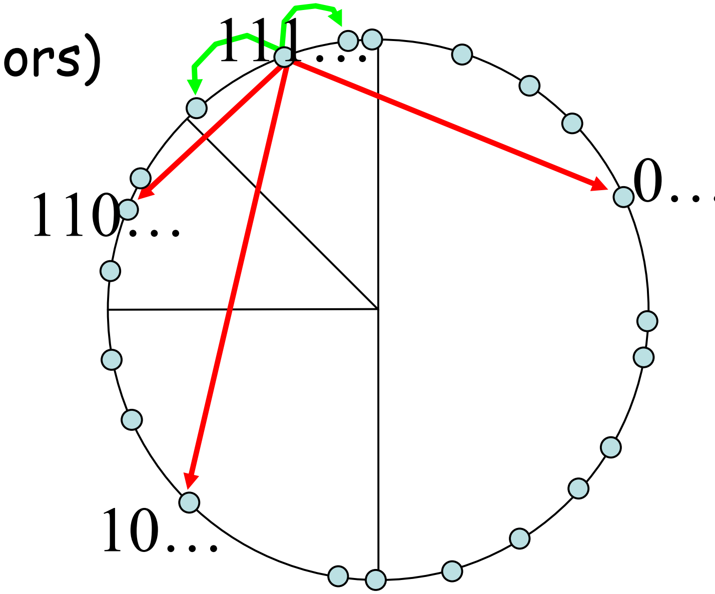
- Overlay Networks
  - RON, Pluto, Violin, etc.
- Network measurement
  - Scriptroute, *Probe, I3, Gnutella mapping.
- Application-level multicast
  - ESM, Scribe, TACT, etc.
- Wide-area P2P distributed storage
  - Oceanstore, SFS, SFS-RO, CFS, Palimpsest, IBP
- Resource allocation
  - SHARP, SHARE, Slices, XenoCorp, Automated contracts
- Distributed query processing
  - PIER, SDIMS, Sophia, etc.
- Content Dist. Networks
  - CoDeeN, Coral, Beehive
- Management and Monitoring
  - Ganglia, InfoSpect, Sword, BGP Sensors, etc.
- Distributed Hash Tables
  - Chord, Tapestry, Pastry, Bamboo, Kademlia, etc.
- Virtualization and Isolation
  - Xen, Denali, VServers, SILK, Mgmt VMs, etc.
- Router Design implications
  - NetBind, Scout, NewArch, Icarus, etc.
- Testbed Federation
  - NetBed, RON, XenoServers
- Etc., etc., etc.

Example: CoDeeN
(Princeton)

- Content Distribution Network
  - ~330 (open) caching proxy servers
  - Open to all users (see URL)
- Highly available (after lots of work!)
- Spawned many subprojects / services:
  - CoBlitz, scalable distribution of large files.
  - CoDeploy, efficient synchronization for slices.
  - CoDNS, fast and reliable name lookup.
  - CoMon, node monitoring for PlanetLab
  - CoTest, login debugging tool for nodes
  - PlanetSeer, distributed network anomaly tracing
- Illustrates how deployment of a real service spurs research
  - http://codeen.cs.princeton.edu/

Example: Bamboo
(Intel Research / U.C. Berkeley)

- P2P Distributed Hash Table (DHT)
  - Like Pastry, Chord, CAN, Tapestry, etc.
- Each DHT node has
  - An identifier in $[0,2^{160})$
  - Leaf set (Predecessors, Successors)
  - Routing table
    - Nodes w/similar prefixes
    - Choose node for each prefix by *proximity* (in network latency)
- Each node responsible for keys closest to its ID



Bamboo, contd.

- Arose from frustration!
  - PlanetLab deployment showed up many problems with existing DHTs
- Pastry topology, new maintenance
  - Highly robust under churn
- Used by new research projects
  - OpenDHT (opendht.org: Intel & UCB)
  - PIER (P2P relational queries: Intel / UCB)
  - Xenosearch (multidim. search: U. Cambridge)
- http://www.bamboo-dht.org/

Lessons from PlanetLab

- Nothing works as expected at scale!
  - Many unintended and unexpected consequences of algorithmic choices
  - Simulation results do not carry over well
    - Simulate, deploy, measure, edit cycle
- Evaluating competing approaches "in the wild" refines techniques
- The ability to try things out "for real" seems to stimulate ideas

Lessons from PlanetLab (2)

- "Unusual" traffic triggers intrusion detection systems
  - Network is often brittle & paranoid!
- UDP w/CC replaces TCP
  - Overlays and P2P applications have many options for next-hop
- Reliable distributed systems from unreliable components *are* possible!
  - OpenDHT: 99.99% availability

What about networking?

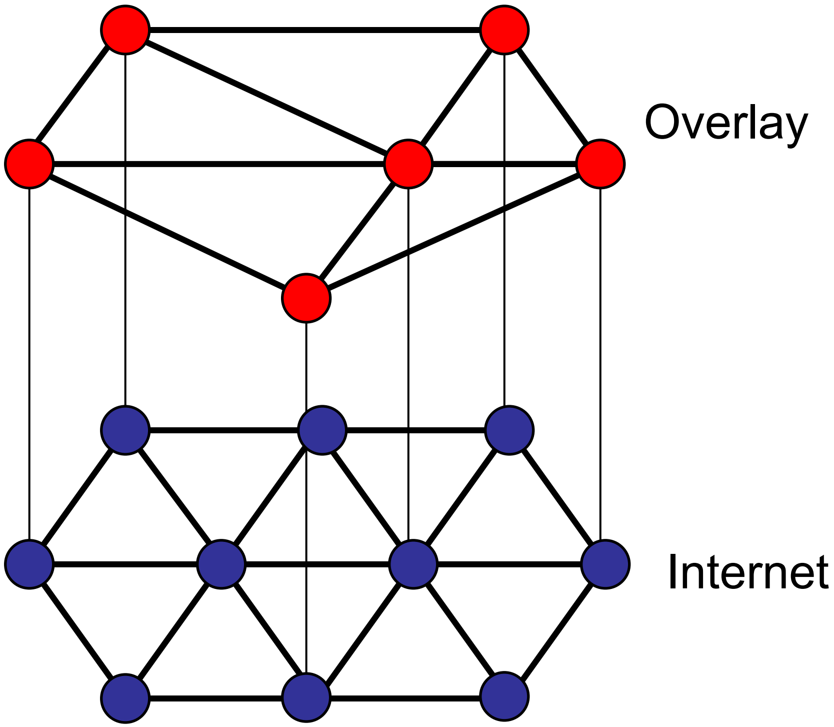
Research into *Internet Architectures*

- Network Management & Provisioning
- What replaces BGP?
- What replaces IP?
- Alternatives to packet switching

- How can University research in this area be validated?
- How can such research achieve impact on industry?

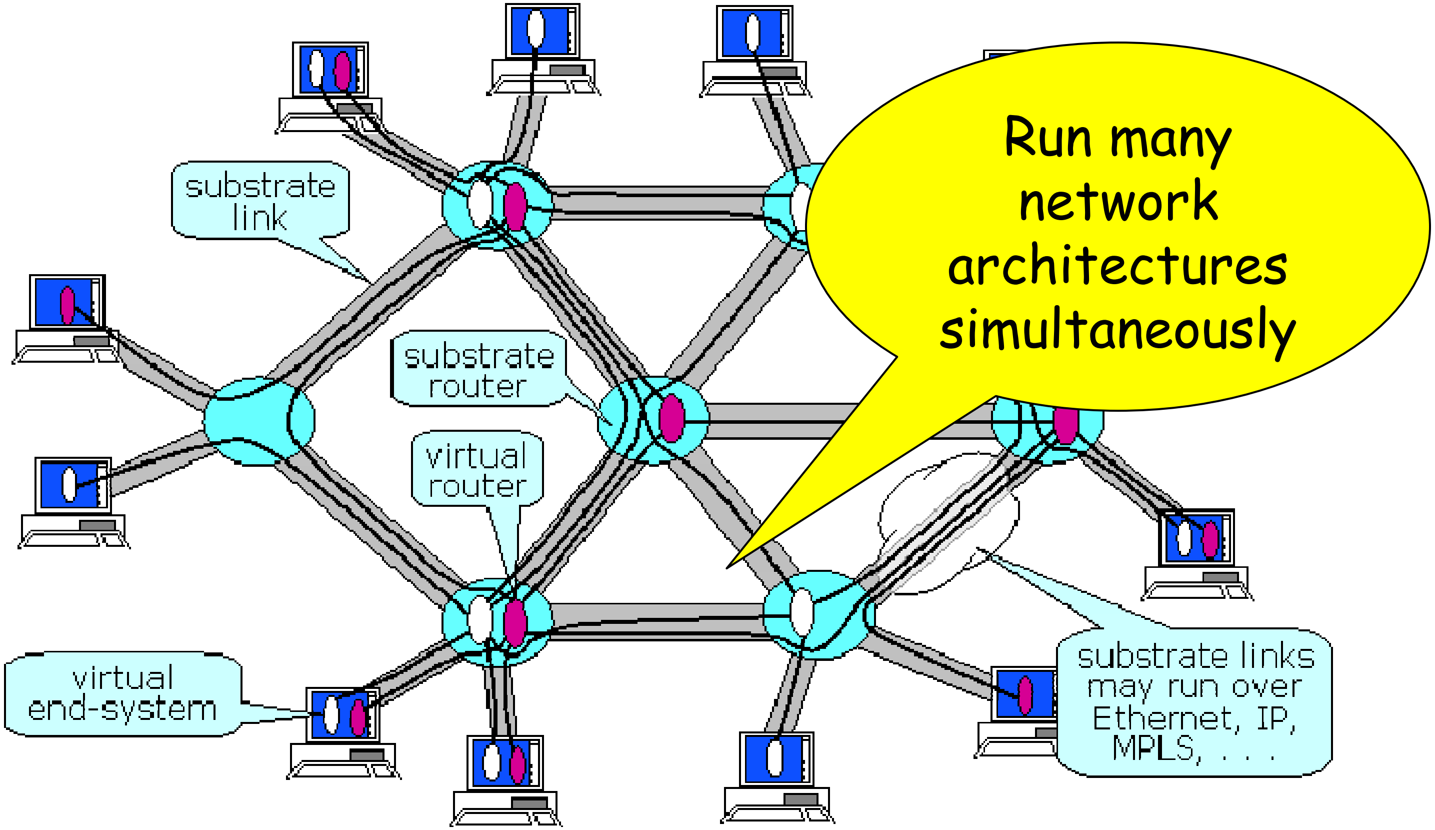
Evolving the Internet: Overlay Networks

- Add new layer to the network architecture.
- Purpose-built virtual networks using existing Internet



Network virtualization

http://www.arl.wustl.edu/netv/



Run many network architectures simultaneously

Enabling architecture research: requirements

- Localizable *computational* resources on demand (control & management)
- Provisioned *links* between computational nodes (emulate physical links)
- Way to *share* resources (parallel research efforts)
- Way for users to *access* overlays over the existing network

Role of PlanetLab

- PlanetLab can provide the computational resources required
- Slices provide the means of sharing them
- Access mechanism: later in talk…
- **Missing piece**: network provisioning

Network provisioning for virtualized networks

- Recent NSF report recommending funding of provisioned links for virtualized network architecture
  - Exploit existing PlanetLab deployment
  - Use IP VPNs, MPLS, LambdaRail, etc.
  - New form for US network testbeds
- Similar efforts discussed in Europe
- Perhaps the only significant extra cost of this line of research

Accessing an IPvN overlay

- “Towards a deployable IP anycast service”: Hitesh Ballani, Paul Francis, WORLDS 2004
- “Towards an Evolvable Internet Architecture”: Sylvia Ratnasamy, Scott Shenker, Steve McCanne, SIGCOMM 2005

Summer Intern Project: Deploy IP Anycast for Overlay Redirection!

- Use a /22 prefix and AS number for anycast
  - Deploy "anycast gateways" at IRB, Cornell
  - Advertise prefix via BGP (help from ISPs)
- Deploy an experimental "IPvN" architecture
  - Actually, a "Default Off" (explicit authorizatoin) network.
- Long term: offer the gateways as generic service for overlay providers.
  - Acts as extension to PlanetLab facilities

Conclusion

- PlanetLab provides an unprecedented platform for experimenting with large distributed systems “in the wild”
- Network virtualization is a technique to unblock radical innovation in network architecture
- Only requires incremental investment over PlanetLab (pipes)
- Enables small groups to perform large-scale architecture research

The Debate: Purism vs. Pluralism

- Purism: virtualization is a way to determine which features the "next" network architecture will have.
- Pluralism: virtualization *is* the next network architecture.

Obrigado!

Acknowledgements: I've talked about many people's work: Tom Anderson, Hitesh Ballami, Mic Bowman, David Culler, Brent Chun, Paul Francis, Brad Karp, Steve McCanne, Vivek Pai, Ruoming Pang, Youngsoo Park, Guru Parulkar, Larry Peterson, Sylvia Ratnasamy, Sean Rhea, Scott Shenker, Jon Turner,

and many others…

PlanetLab is not the Grid, #1

- The Grid aims at location-transparency for large computations
  - “I don’t care where this protein-folding job runs as long as it’s done by Monday”

- PlanetLab is all about small, long-running services in specific locations
  - “I need to run a new secure file cache for the next 6 months in Seoul, Sydney, Tromsø, and Vancouver”

PlanetLab is not the Grid, #2

- The Grid is about *standardizing* one particular paradigm for large-scale utility computing.
- PlanetLab provides a low-level platform over which *many* distributed computing paradigms can be tried.
  - You could build the Grid over PlanetLab's abstractions if you really wanted

PlanetLab is not the Grid, #3

- The Grid starts from scratch in putting together an execution environment for remote computation (e.g. OGSA)
- PlanetLab starts with a well-known, simple interface and encourages the community to evolve multiple, competing execution environments

PlanetLab is not the Grid, #4

- The analogy with networking…

OGSA ⇔ ISO OSI

PlanetLab ⇔ TCP/IP

Enough said.