IP over P2P: Enabling Self-configuring Virtual IP Networks for Grid Computing

Arijit Ganguly, Abhishek Agrawal,
P. Oscar Boykin, Renato Figueiredo

University of Florida

IPDPS 2006

What is the talk about?

- Convergence of Grid and P2P technologies[1]
- Context of network virtualization

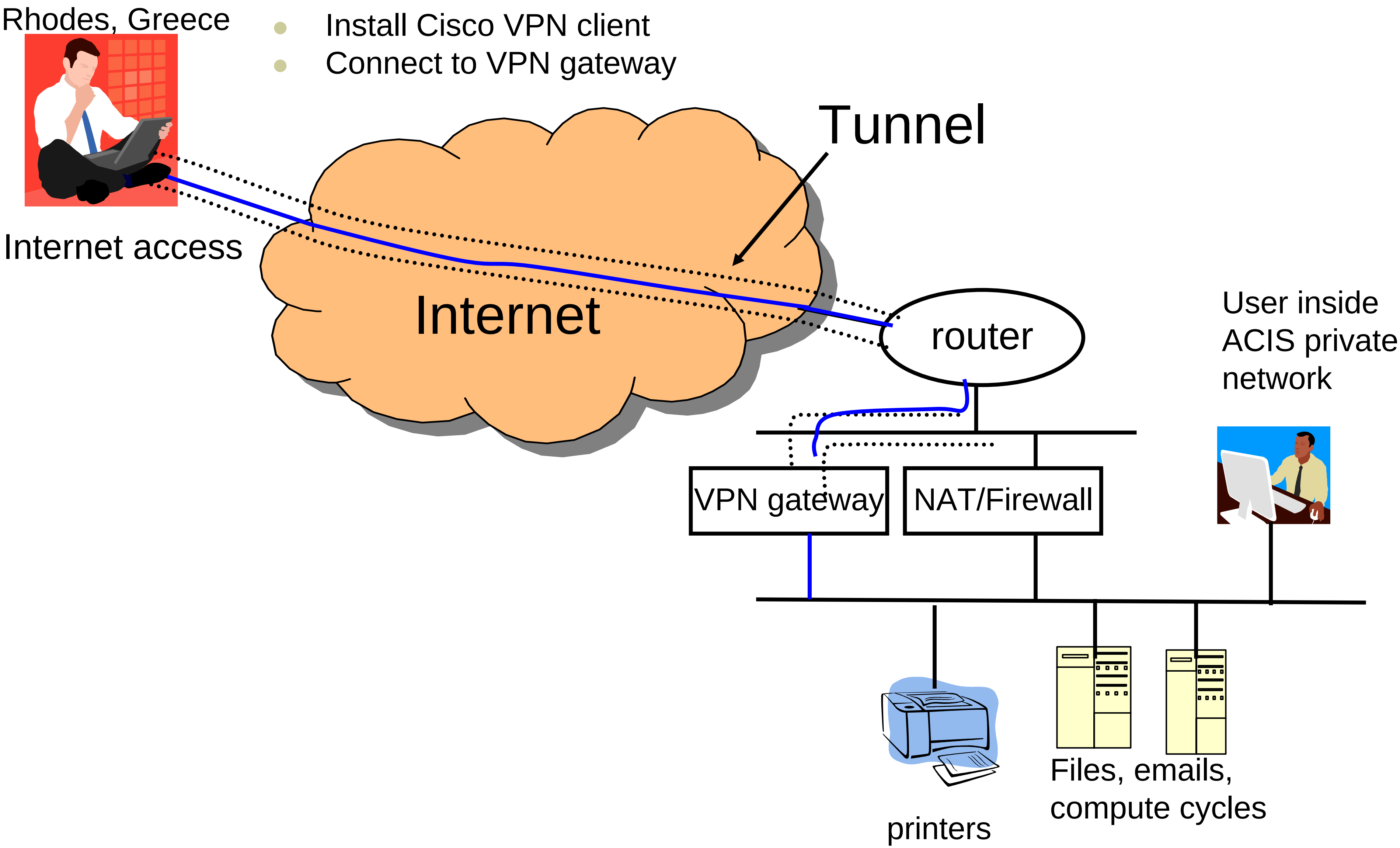
1 *On death, taxes, and the convergence of peer-to-peer and Grid Computing*. Foster et al. *IPTPS 2003*

Outline

- Virtual networking and Grid Computing
- Related work
- Our approach – IP over P2P
- Experimental evaluation
- Conclusion and Future work

Background - Virtual Private Networks

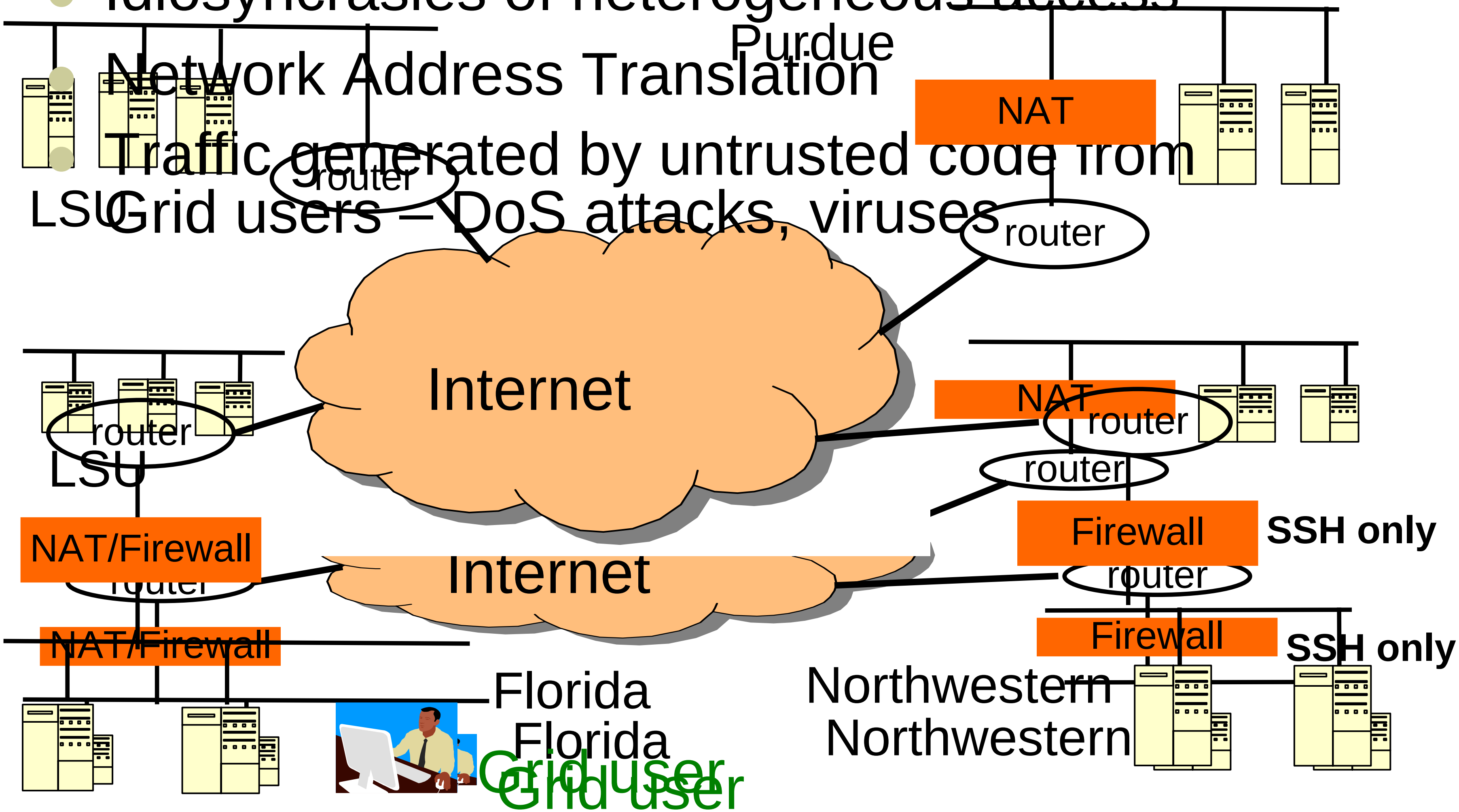
- Install Cisco VPN client
- Connect to VPN gateway



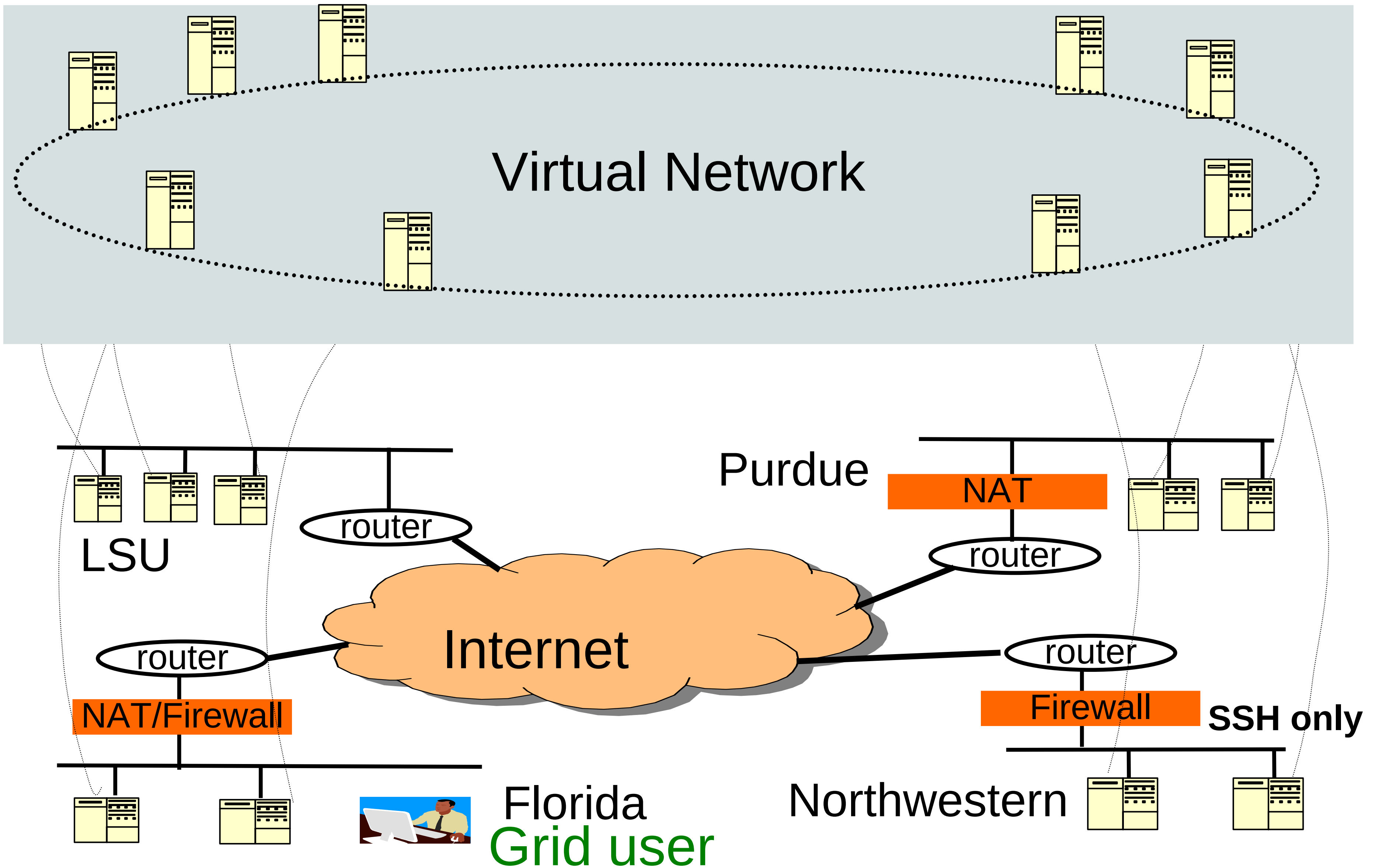
Grid scenario

Issues:

- Idiosyncrasies of heterogeneous access
- Network Address Translation
- Traffic generated by untrusted code from Grid users – DoS attacks, viruses



Virtual network of Grid resources



Virtual networking for Grids

- VNET (Northwestern University)
  - Bridge a remote Virtual Machine (VM) to a client network
- VIOLIN (Purdue University)
  - Virtualized network components
  - Isolated from real physical network
- ViNe (University of Florida)
  - Virtual IP network of Grid resources
  - To be presented on Friday (Session 32)

Common technology: Overlay tunneling

What differentiates us: P2P routing

Motivations for P2P

- Scalability and Self-configurability
  - Manual effort required to add a new node constant
  - Independent of size of the network
- Resiliency
  - Robust P2P routing
- Accessibility
  - Ability to traverse NAT
  - Hole punching[1]

1 RFC 3489 - STUN - Simple traversal of User Datagram Protocol through Network Address Translators

Our approach – IP- over-P2P (IPOP)

- Isolation
  - Virtual addres
    address spac
- Self-configura
  - Automatic set
    - Decentralize
      - No global
      - No centra
    - VM mobility
- Decentralized
  - No changes t
  - No globally de

```
#affiliation
condor_wow
#transport
udp
#port
15000
#number of remote TAs
2
#list of TAs
brunet.udp://planetlab-01.bu.edu:15000
brunet.udp://planetlab1.cs.purdue.edu:15000
#virtual interface
tap0
#virtual IP address of tap0
172.16.1.5
#MAC address of tap0
CB:DF:E7:20:60:35
```

IPOP - Architecture Overview

- IP tunneling over P2P overlay networks
  - UDP, TCP

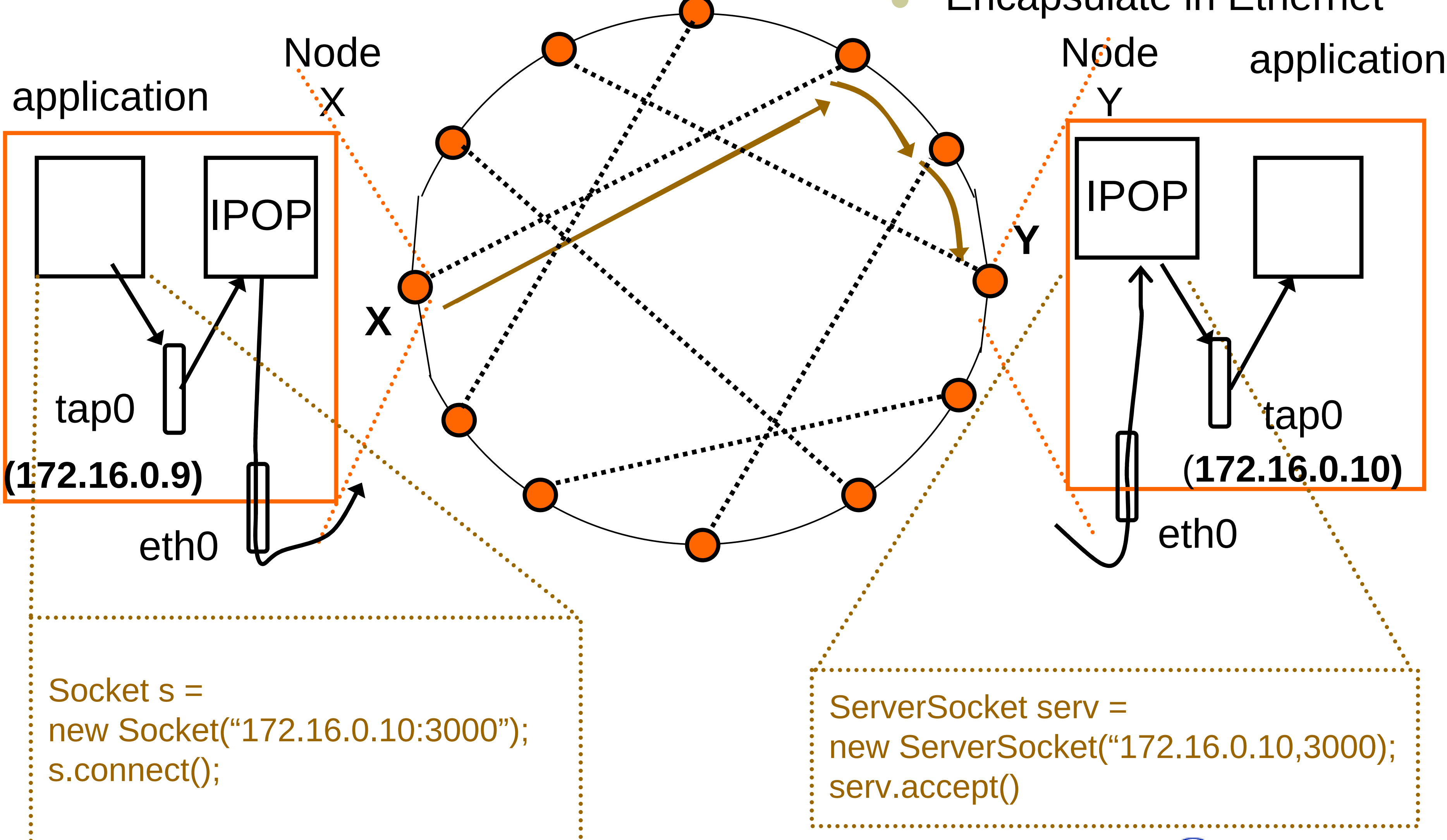
- Virtual IP packet capture and injection through *tap* interface

- Builds upon Brunet P2P library

IPOP – Packet capture and routing

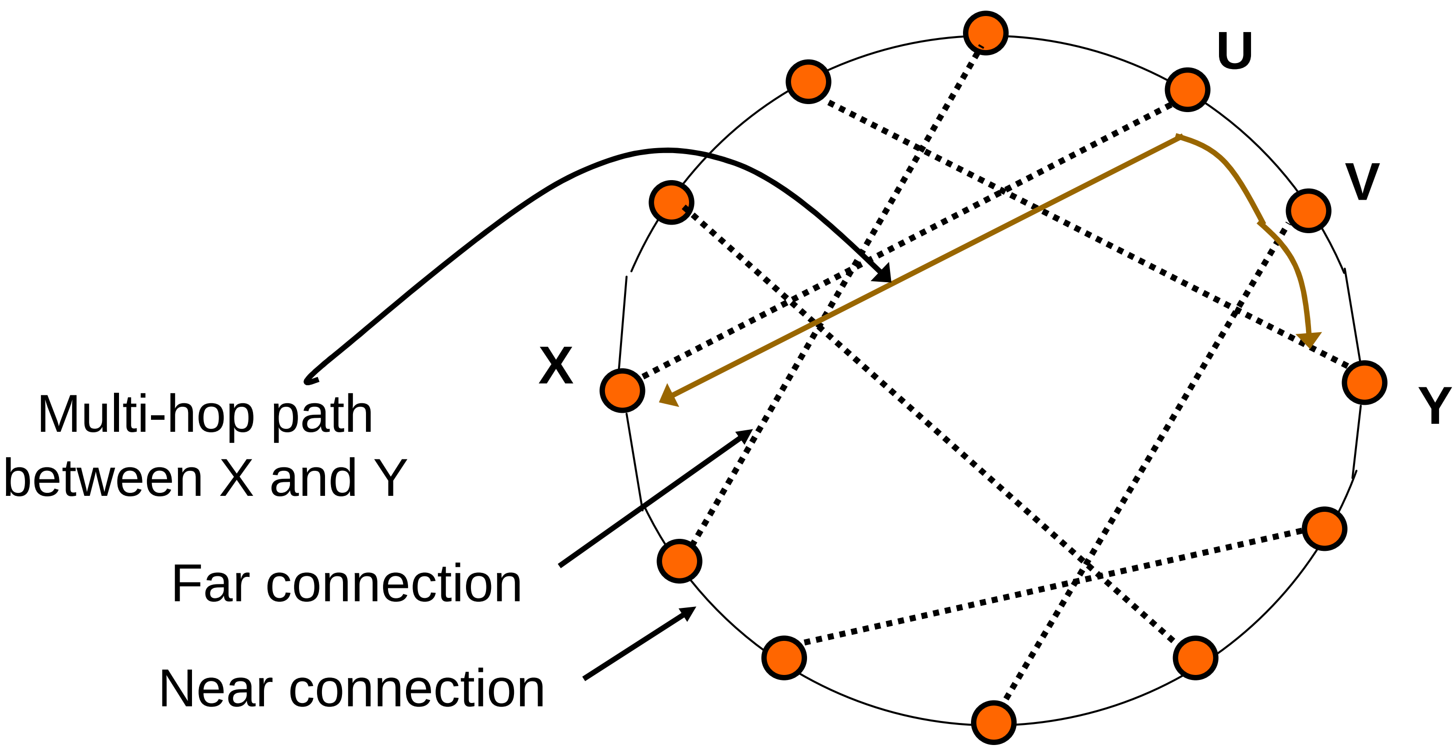
- Extract IP from Ethernet
- Encapsulate IP inside P2P

- Extract IP from P2P
- Encapsulate in Ethernet



Brunet P2P architecture

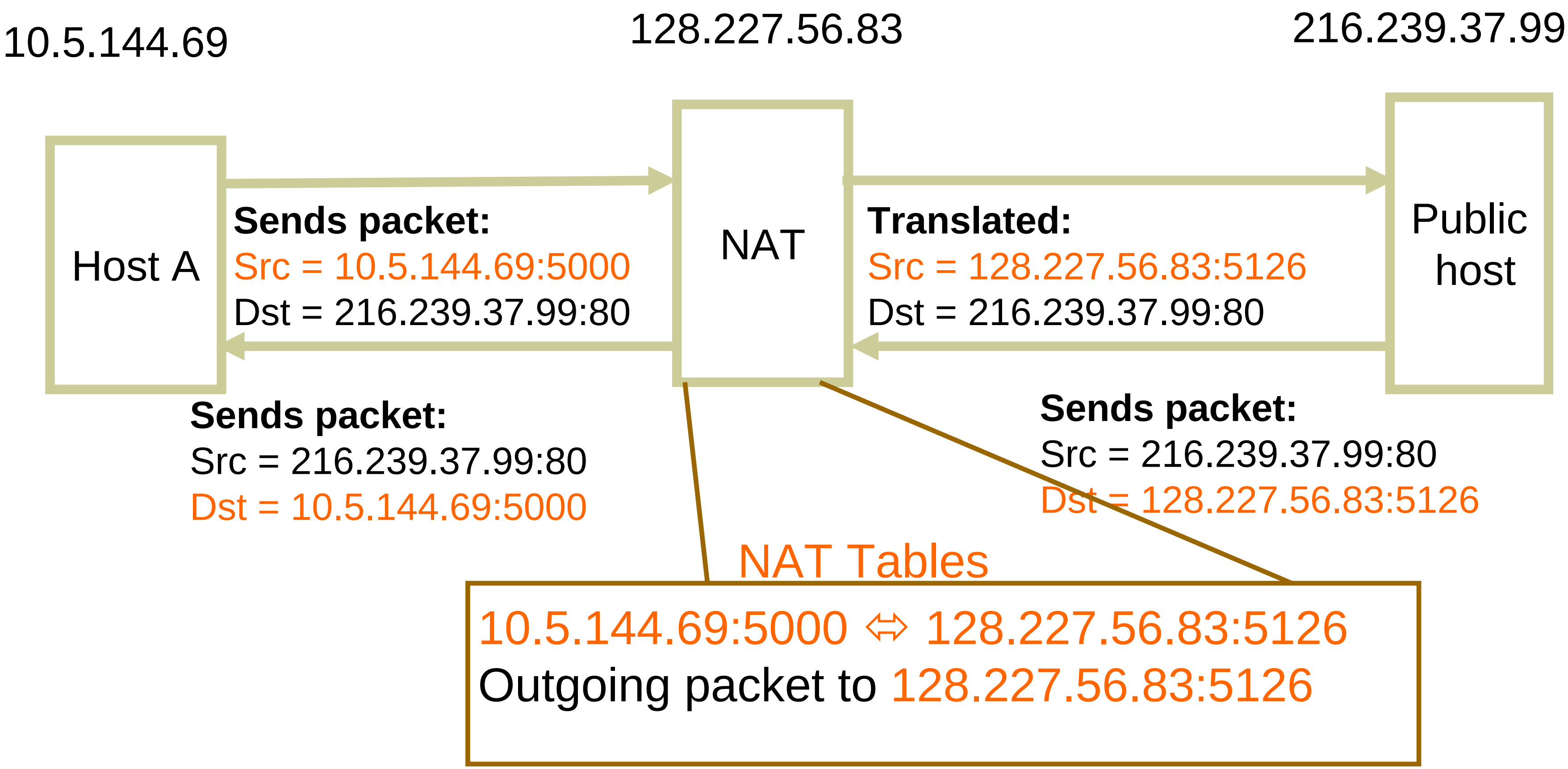
- Ring-structured overlay network topology
  - Nodes ordered on 160-bit addresses
- Overlay link:
  - Near: neighbor connections
  - Far: connections across ring



Brunet P2P architecture (2)

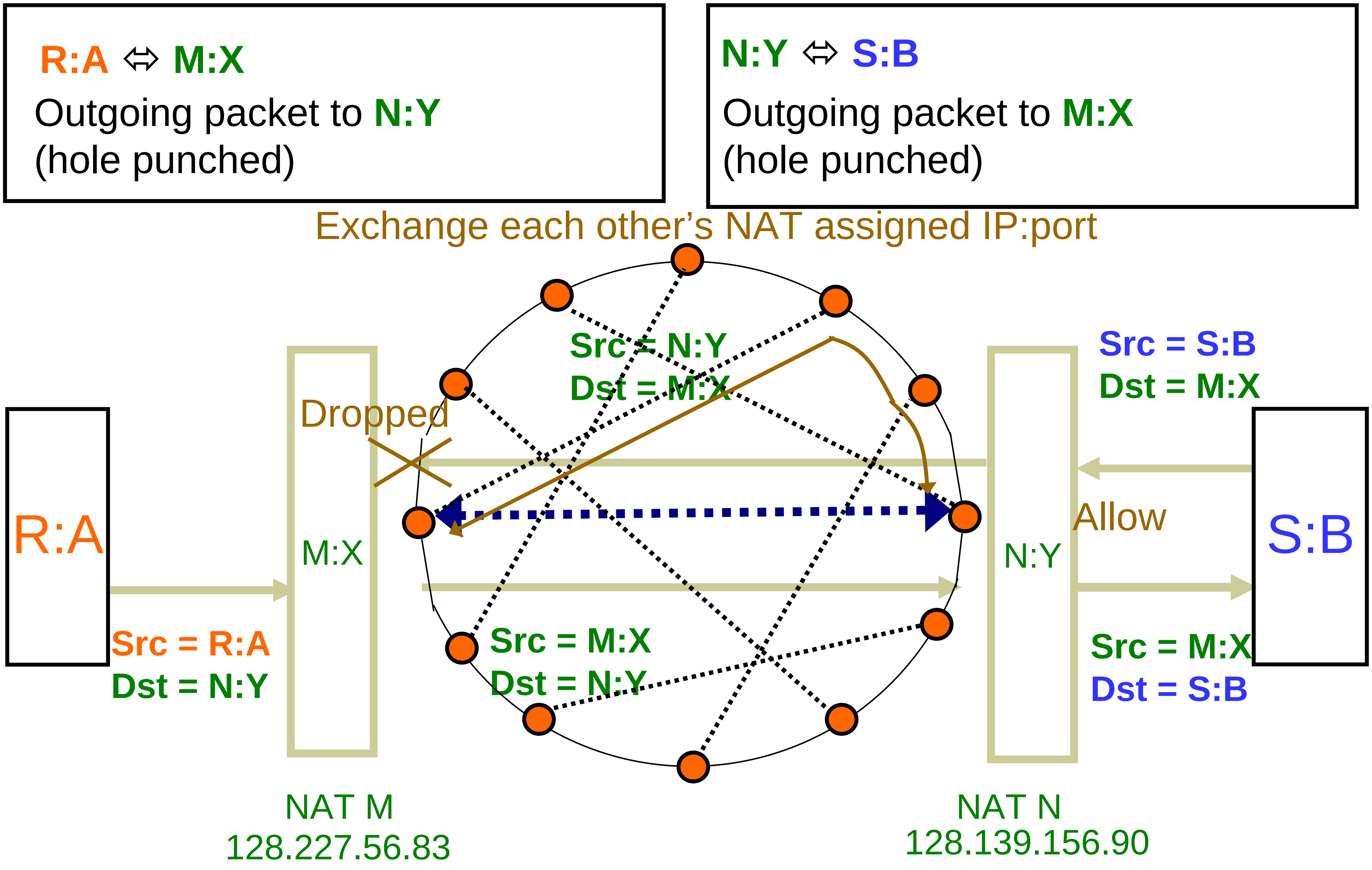
- Routing
  - Constant number of connections
    - $O(\log^2(n))$ overlay hops
  - $O(\log(n))$ connections
    - $O(\log(n))$ overlay hops
  - n connections
    - 1-hop
- C# library, supports:
  - Connection setup and maintenance
  - NAT traversal

Network Address Translation (NAT)



Applications on NATed hosts can learn their NAT assigned IP:port

NAT traversal – Behind NATs



Experiments

- Latency overhead and throughput of single overlay link
  - LAN and WAN
- MPI application over IPOP
  - Light Scattering Spectroscopy (LSS)
- Multi-hop routing experiments
  - More than 100 node network on PlanetLab

Latency (single IPOP link)

- Two IPOP nodes separated by single overlay hop
  - ACIS – ACIS for LAN
  - ACIS – VIMS for WAN
- Ping times between two nodes
- 6ms-11ms overhead per packet for ICMP ping
- Relative overhead is smaller in Wide-Area

ACIS: Florida   VIMS: Virginia

Latency overhead - analysis

- Reasons for high LAN overhead:
  - Double traversal of kernel stack
  - C# runtime
  - User-level overlay – context switches
  - Other user-level overlays (VNET, Violin) report few-ms latency overheads

Throughput (single IPOP link)

- Two IPOP nodes separated by single overlay hop
  - ACIS – ACIS for LAN
  - ACIS – VIMS for WAN

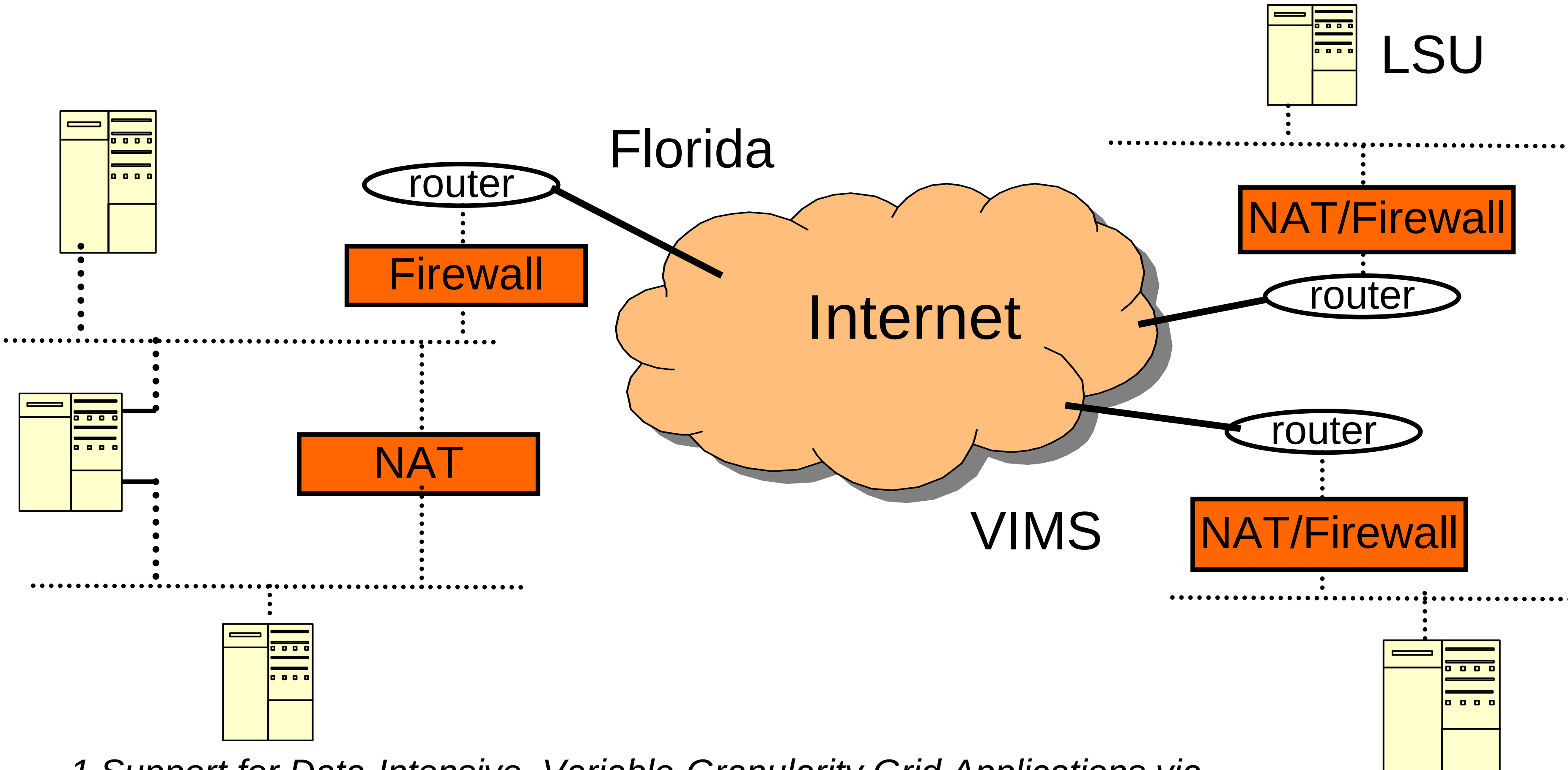
- "ttcp"
  - file transfer sizes (13.09 MB, 92.97 MB)

- 1.9MB/s LAN bandwidth (20% of physical 9.4 MB/s)
- 1.2MB/s WAN bandwidth (80% of physical 1.5 MB/s)

ACIS: Florida VIMS: Virginia

Real Application – Parallel LSS

- MPI + NFS + SSH1[1]



1 Support for Data-Intensive, Variable-Granularity Grid Applications via Distributed File System Virtualization - A Case Study of Light Scattering Spectroscopy. Figueiredo et al. CLADE 2004

Real Application – Parallel LSS

- With IPOP, could run "parallel LSS" unmodified
  - No changes to NAT/Firewall rules
- Achieve parallel speedup

PlanetLab experiments

- Demonstrate ease of adding a new node and achieving IP routability in WAN environment
- 118 node TCP-based overlay on PlanetLab
- Connect two IPOP nodes in ACIS lab to PlanetLab network
- Measure ping times between nodes
  - Average: 1617 ms; Std Dev: 2098 ms

Planetlab experiments (analysis)

- Issues:
  - High-load (>10) on nodes in routing path
  - Geographically unaware p2p routing
    - Packets between machines in Florida routed through machines in California
- Improvements:
  - Direct overlay link setup between communicating nodes
    - No concerns of load and inefficient p2p routing

Conclusion

- Our contribution:
  - Novel virtual IP network based on P2P overlay
    - Scalable and Self-configurable
    - Resilient
    - NAT traversal
  - Experiments showed feasibility of using P2P approach for virtual networking

Future work

- Overhead of TCP or UDP
  - Raw sockets or Ethernet-based overlay edges
- Kernel level extensions
  - Tap module with encapsulation and bridging
  - Reduce context switches

Related Work

- Virtual Networking
  - VIOLIN
  - VNET
  - ViNe (Session 32)
- Internet Indirection Infrastructure (i3)
  - Support for mobility, multicast, anycast
  - Decouples packet sending from receiving
  - Based on Chord p2p protocol
- IPv6 tunneling
  - IPv6 over UDP (Teredo protocol)
  - IPv6 over P2P (P6P)

Acknowledgments

- In-VIGO team at UFL
- National Science Foundation
  - Middleware Initiative (http://www.nsf-middleware.org)
  - Research Resources Program
  - nCn center
- Resources
  - Peter Dinda (Northwestern University)
  - SURA/SCOOP
- IBM Shared University Research

Questions?

Thank You