Towards a Uniform Self-Configuring Virtual Private Network for Workstations and Clusters in Grid Computing

David Wolinsky, Yonggang Liu, and Renato Figueiredo
ACIS P2P Group
University of Florida

So What's the Big Deal

- Support connectivity for Grids across the Internet in constrained locations
- Clear VN to overlay interface
- Auto-configuring virtual networking end points

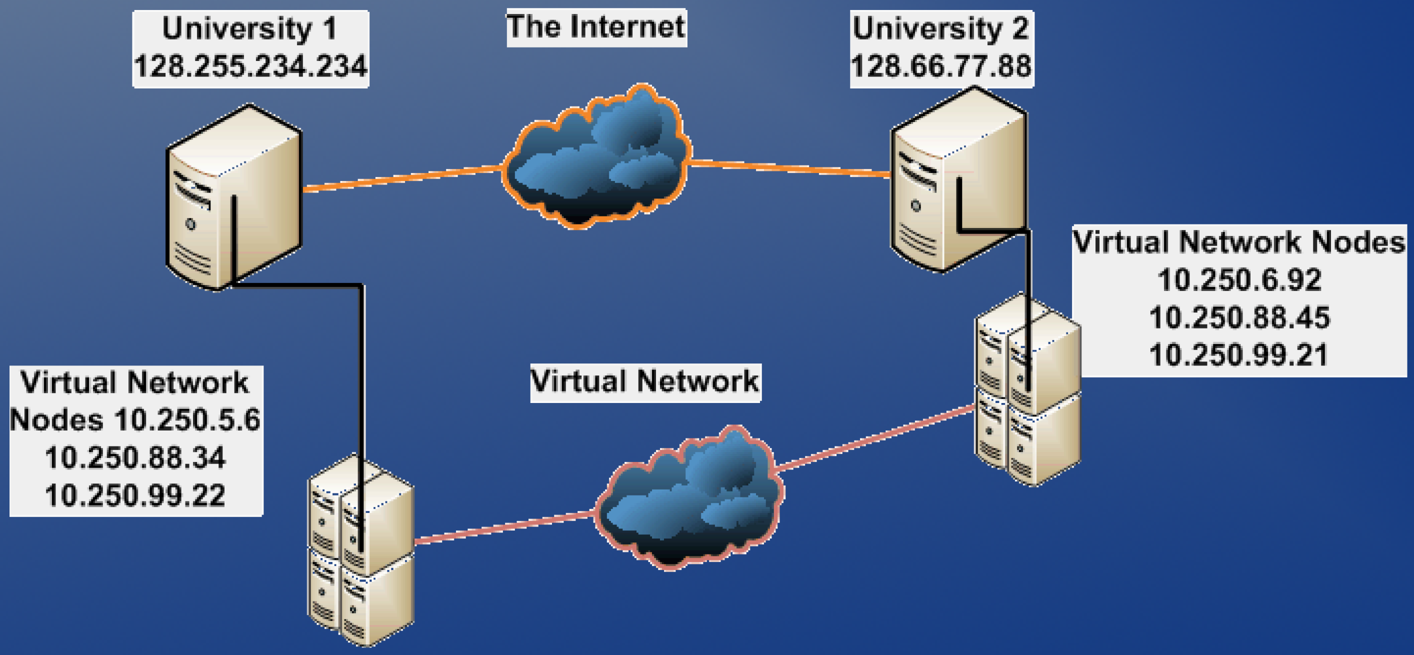
- Focus: What technologies and standards exist that we can reuse to simplify VN development and deployment?

Autonomic Virtual Networking

- What is Virtual Networking
- Virtual Networking in an Overlay
- Virtual Networking Interfaces and Routers
- Address Allocation and Resolution
- Quantitave Evaluation
- Conclusions / Future work

Virtual Networking in Action

- Unified layer 3 (IP) network for all machines
- Cross-site communication without a middleware broker



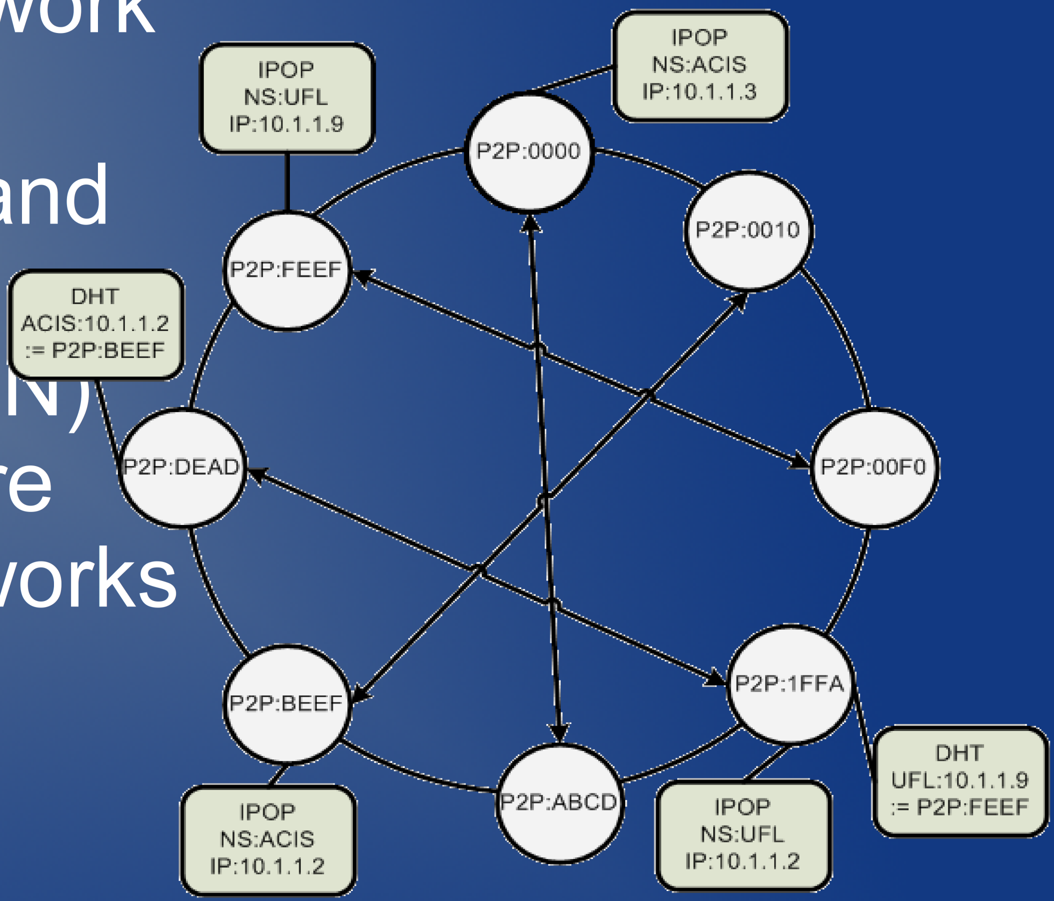
UF UNIVERSITY of FLORIDA ACIS

Recent Virtual Networking Overlays

- ViNe – statically configured IP routes
- VNET – static overlay routes with broadcast ARP
- N2N – dynamic P2P overlay with broadcast ARP
- IPsecVPNs (Cisco, Open, L2TP) – central servers no direct communication
- Hamachi, Wippien, Other upcoming “P2P” VPNs – centralized authenticator, bootstrap, tunnel, decentralized if direct connection capable
- IPOP – next slide, dynamic p2p overlay with distributed address discovery

Virtual Networking in the Overlay

- Structured P2P Network Overlay
- Provides tunneling and direct shortcuts
- NAT Traversal (STUN)
- Distributed data store
- Multiple Virtual Networks Per Overlay



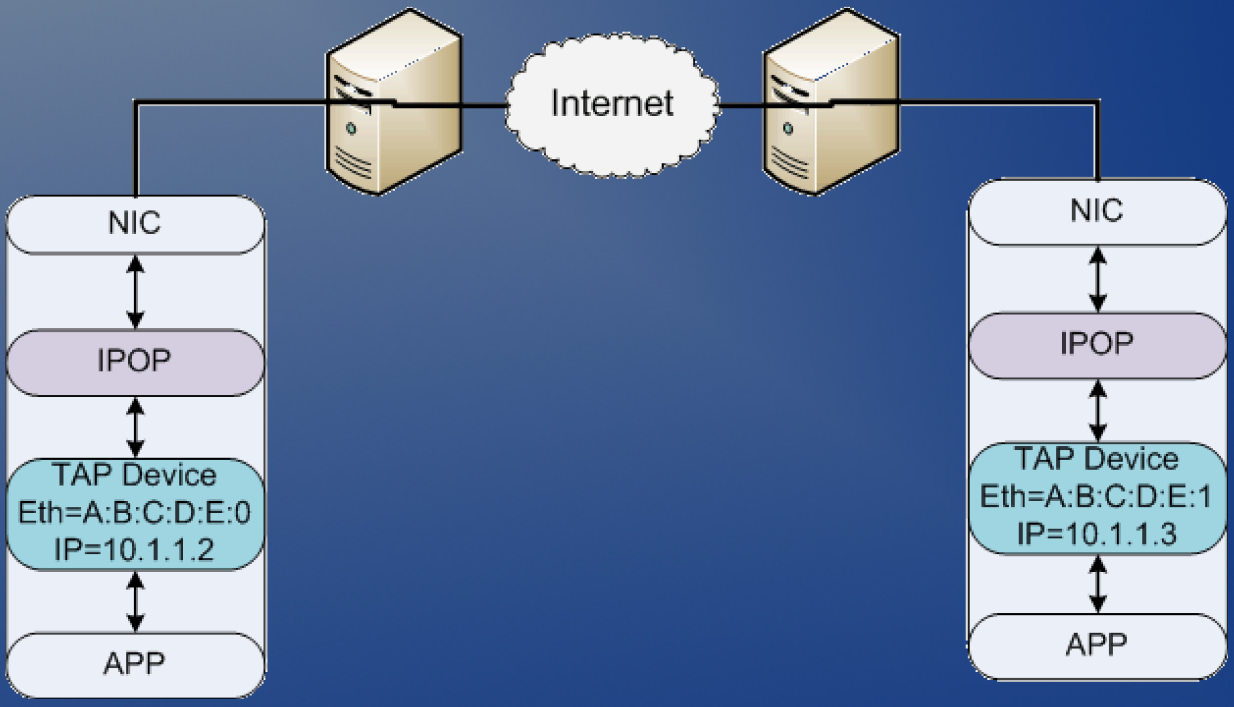
UF UNIVERSITY of FLORIDA ACIS

Recent Virtual Networking Models

- ViNe – Site based router given a unique static address space
- VNET, N2N – Layer 2 VN device
- IPsecVPNs, “P2P” VPNs – Layer 3 VN device
- IPOP – next few slides, a Layer 2/3 VN router and device

VN Interfaces

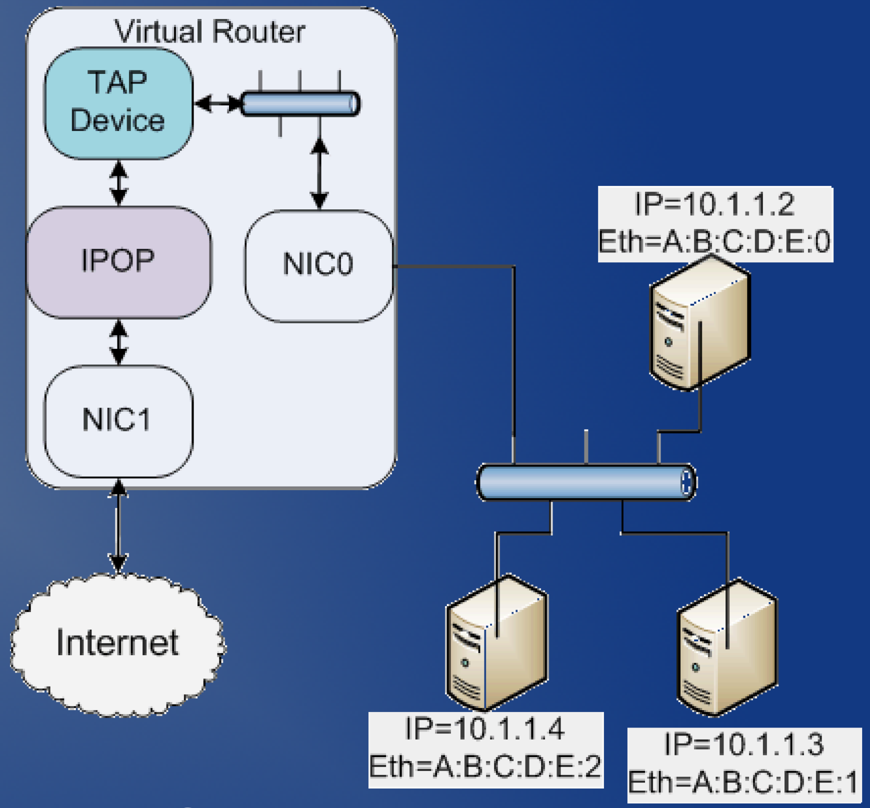
- Each machine has VN Interface running locally
- Machine has VN and “Internet” connectivity



UF UNIVERSITY of FLORIDA ACIS

VN Routers

- Single VN instance (Router) for entire cluster
- Machines have VN connectivity
- May have “Internet” connectivity if there is an “Internet” router



- Limited to no resource configuration
- Reduced virtualization overhead between nodes on same physical network

VN Routers - Downside

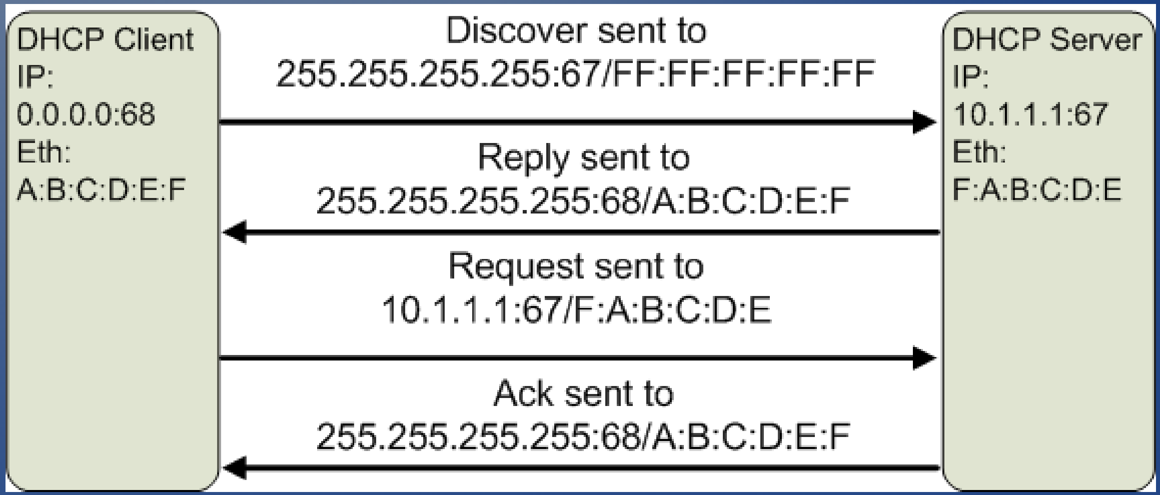
- Mixed address space on the same L2 Network is frowned upon by Sys admins
- Can get around via Virtual LAN, if available
- DHCP for different L3 networks on the same L2 can confuse machines, needs a method to differentiate
  - DHCP packet redirection to a non-standard port
  - DHCP server listening on a non-standard port
  - Pre-registering MAC addresses

Address Allocation and Resolution

- Novelty: machines in multiple domains to think they are all part of the same Layer 2 Virtual Network without supporting Layer 2
- Dynamic Address Allocation - DHCP
  - Dht.Create(DesiredIP, MyP2PAddress)
  - If returns true obtained IP
  - If returns false retry with different IP
- Resolution - ARP
  - Dht.Get(RemoteIP)
  - Returns P2P Address

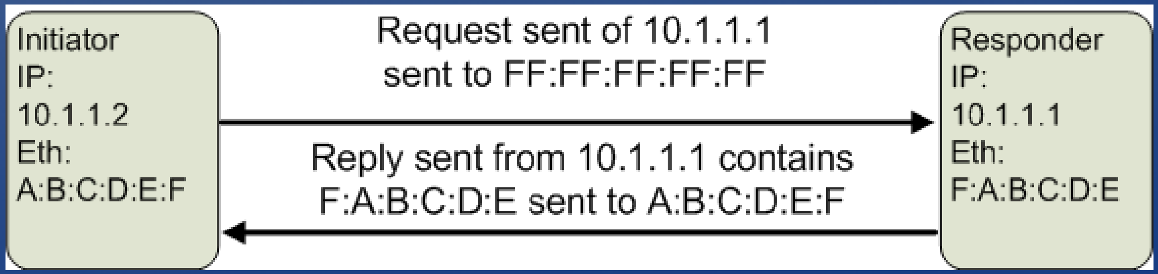
DHCP

- Provides address allocation and DNS settings
- VN Router keeps a history of allocations and ignores packets destined for them sent within the physical network



ARP

- Used to resolve mapping of Ethernet to IP address
- VN Router ignores these if for physical network
- Otherwise acts as a bridge replying with a non-standard Ethernet address and all packets are sent to it
- Operating System (Kernels) keep a table that maps these addresses
- In Layer 3 VNs, like ViNe, each Router is given its own subnet and thus each machine in the Physical network only needs to know the Routers address, but you lose the ability to have a single subnet for the entire network

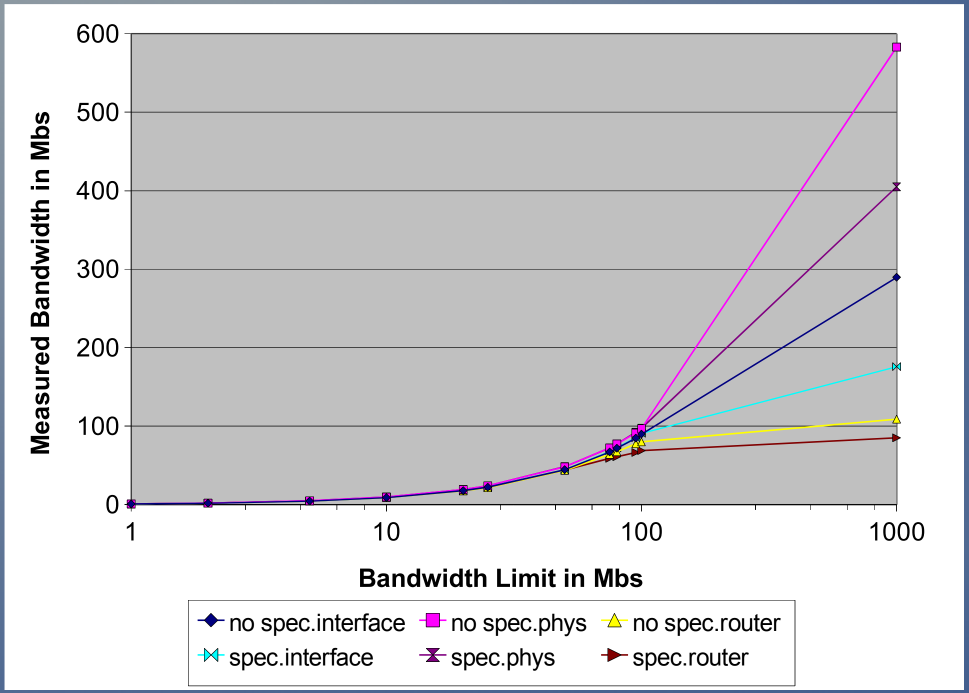


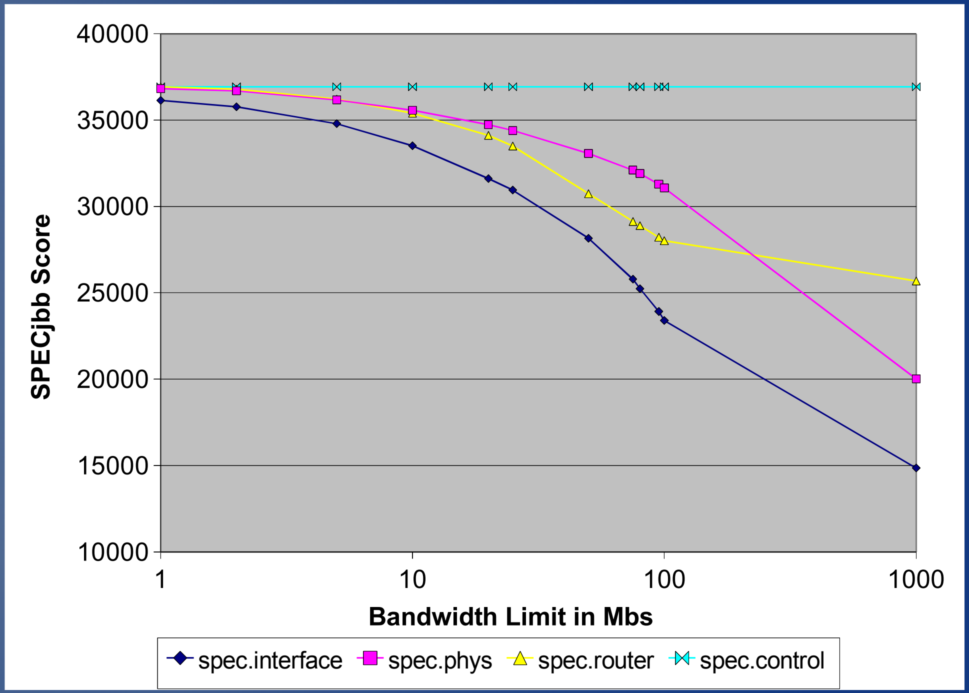
UF UNIVERSITY of FLORIDA ACIS

Evaluation Introduction

- Results from follow up paper accepted at SC '09
- Bandwidth – netperf tcp stream
- Latency – netperf tcp transactions
- CPU - SPECjbb
- Set of 7, quad core machines on the same Gigabit switch, running 4 VMs each
- 1 machine acts as a server, the other 6 act as clients
- Servers share a single link to enforce bandwidth, bandwidth set using tc (traffic control)
- Router mode uses a 5th VM
- Interface mode has an interface in each VM

Bandwidth vs CPU



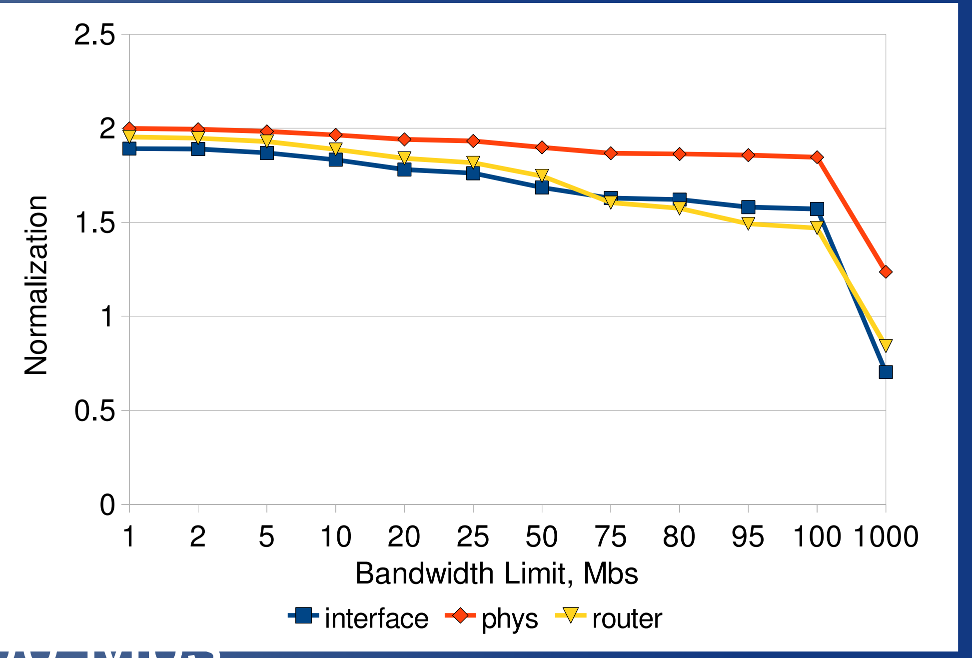


- Router better initially
- Around 80-100 it appears router / interface switch positions

UF UNIVERSITY of FLORIDA ACIS

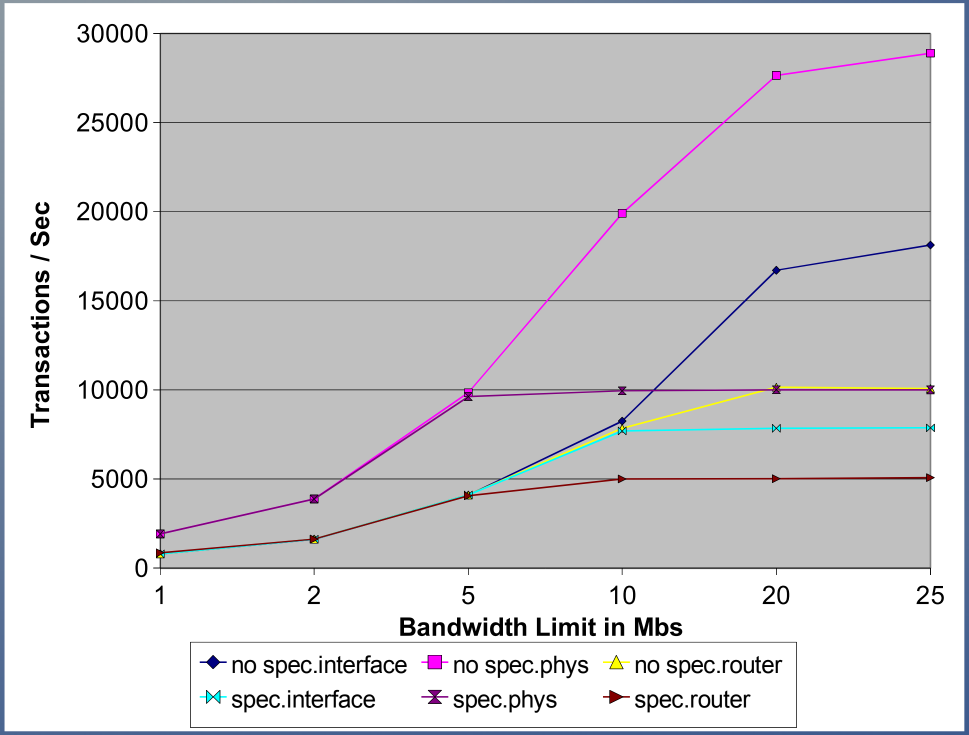
Bandwidth Normalization

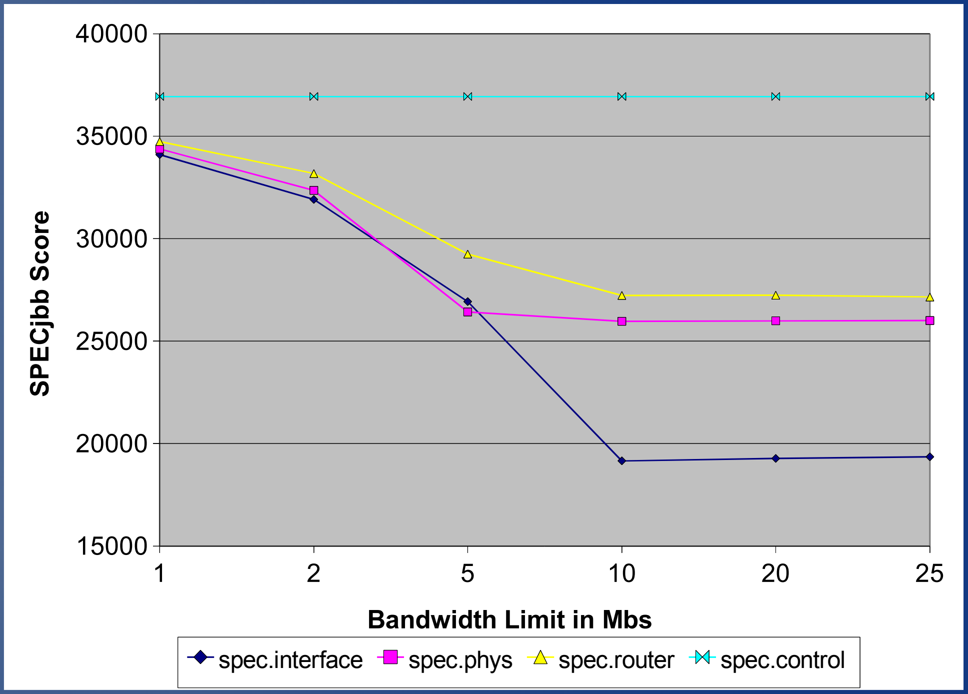
- Meas bw w/spec / Max b
- + Meas SPEC / Max SPE
- Router better < 50 Mbs, interface better > 50 Mbs
- Router has better 1000 Mbs normalization but has significant cap on maximum bandwidth < 10



UF UNIVERSITY of FLORIDA ACIS

Latency vs CPU



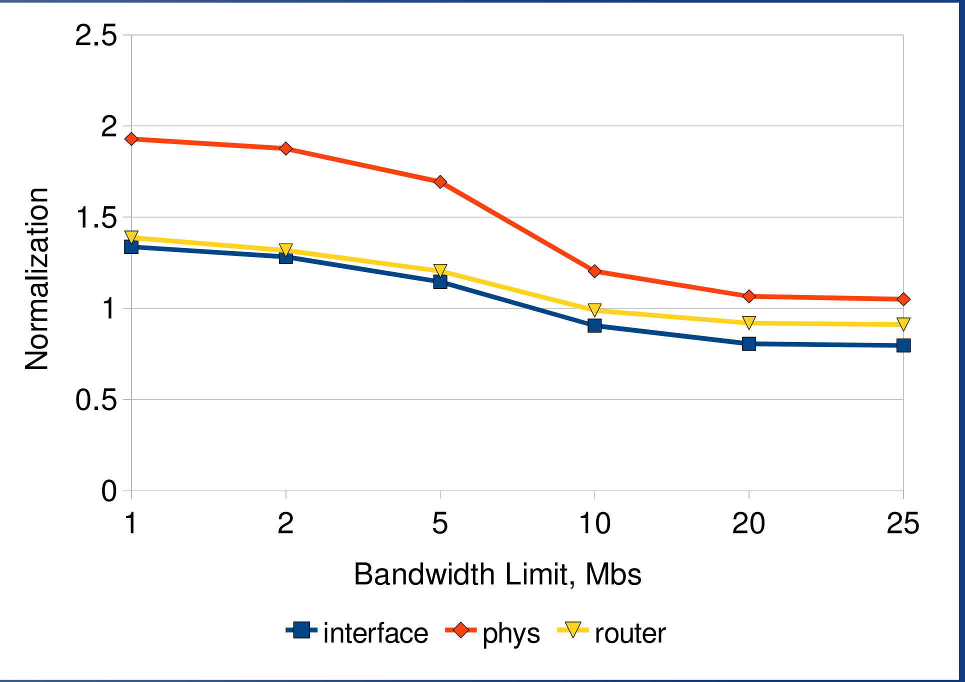


- All have a maximum trans/sec (steady state not presented)
- Interface can perform more trans / sec (with loss of CPU performance) than Router

UF UNIVERSITY of FLORIDA ACIS

Latency Normalization

- Meas ts w/ spec / Max ts Meas SPEC / Max SPEC
- Router has better overall than Interface
- Router suffers transaction hit in favor of more CPU cycles (previous slides)



Evaluation Conclusions

- For networks < 100 Mbs, no clear quantitative choice
- If network performance is key, use interface
- If CPU performance is key, use router

VTDC Results

- Multisystem setup
- Extremely confusing, sorry!
- Meant to help exploit cases that would not be obvious by having similarly configured machines
- Consists of 4 machines dating from 2003 to 2008 all with 1Gbit network cards

Latency Results from VTDC Paper

- Router incurs extra latency (as expected), but not as much as expected

Interface Latency (Ping - Msec)

client / server	m1		m2		m3		m4		average	
	h2h	v2v	h2h	v2v	h2h	v2v	h2h	v2v	h2h	v2v
m1	N/A	N/A	0.48	0.94	0.5	0.7	0.26	0.51	0.42	0.72
m2	0.32	0.82	N/A	N/A	0.38	0.73	0.14	0.57	0.28	0.71
m3	0.49	0.78	0.51	0.96	N/A	N/A	0.12	0.48	0.38	0.71
m4	0.27	0.53	0.3	0.7	0.22	0.55	N/A	N/A	0.26	0.58
average	0.36	0.71	0.43	0.87	0.37	0.65	0.18	0.52	0.33	0.69

Router Latency (Ping - Msec)

client / server	m3 / m1			m2 / m4		
	meas	exp	phys	meas	exp	phys
m3 / m1	N/A	N/A	N/A	0.94	1.17	0.91
m2 / m4	1.13	1.51	1.26	N/A	N/A	N/A

BW Results from VTDC Paper

- Router doesn't inhibit performance and can actually improve performance on slow machines

Interface BW (Iperf - Mbps)

client / server	m1		m2		m3		m4		average	
	h2h	v2v	h2h	v2v	h2h	v2v	h2h	v2v	h2h	v2v
m1	N/A	N/A	700	109	940	192	940	183	860	161.33
m2	576	111	N/A	N/A	622	83	566	105	588	99.67
m3	928	162	701	87	N/A	N/A	939	140	856	129.67
m4	942	202	701	98.3	942	136	N/A	N/A	861.67	145.43
average	815.33	158.33	700.67	98.1	834.67	137	815	142.67	791.42	134.03

Router BW (Iperf - Mbps)

client / server	m3 / m1			m2 / m4		
	meas	exp	phys	meas	exp	phys
m3 / m1	N/A	N/A	N/A	219	183	701
m2 / m4	219	202	576	N/A	N/A	N/A

Wrap Up

- This work focuses on making VN Routers autonomic
- Autonomic features include:
  - Address allocation
  - Address resolution
- Other interesting possibilities
  - Support for more VN Routers to help with scaling
  - The effects of implementing a Layer 2 VN Router

How IPOP Works

- Start with a list of sites that may be online
- Connects with at least one that helps it find other peers online
- Locates its place in the P2P overlay and forms connections with neighbors
- Dht operations will now work and the next DHCP will place an Namespace+IP Address: P2P Address into the Dht
- Another node will want to talk to Namespace + IP Address, will look it up in the Dht, and then send packets to the remote P2P Address via the overlay

Forming Direct Communication

- If sufficient packets are transferred, the two machines will attempt to form direct connections
- Machines exchange all methods of addressing each other, both learned and statically configured
- Machines simulataneously attempt to connect, tricking (some) NATs into opening bidirectional pathways
- If connection is not possible, messages can still be sent via the overlay

A Simple, Static DNS

- All addresses in VN subnet are statically mapped
- Subnet 10.0.0.0/8For example:
  - IP Address 10.5.32.155
  - Hostname C005032155
- Hostnames not of the form CXYZ are quickly rejected to allow other DNS to respond
- In Linux, it must be the first DNS listed
- Works seamlessly in Windows

UF UNIVERSITY of FLORIDA ACIS