NetLord: A Scalable Multi-Tenant Network Architecture for Virtualized Datacenters

Jayaram Mudigonda, Praveen Yalagandula, Jeff Mogul, Bryan Stiekes, Yanick Pouffary

Present by Xiang Wang

Venus Team, NSLab

RIIT, Tsinghua Univ.

Oct. 26, 2011 @ NSLab Seminar

The Goal

- Build the right network for a cloud datacenter?

The Goal

- Build the right network for a cloud datacenter?

Cloud Datacenter

- Provides Infrastructure as a Service
  - Shared across multiple tenants
  - Pay-as-you-go model
- Virtualized
  - Tenants run Virtual Machines (VMs)
  - Time-multiplex
- Examples
  - Amazon EC2

The Right Network

- Virtualization + Multi-tenancy
  - A *Virtual Network* to each tenant
  - No restrictions on addressing or protocols
- Scale
  - Tenants, Servers: 10s of 1000s, VMs: 100s of 1000s
  - Adequate bandwidth
- Inexpensive
  - CAPEX: Cheap COTS components
  - OPEX: Ease of management

The Challenge

- Basic COTS switching gear →
  - Limited functionality and resources:
    - Not enough Forwarding Information Base (FIB) space
- Multi-tenancy →
  - Not full address-space virtualization
    - Only MAC/IP address-space sharing
- Configuration →
  - Careful manual configuration

The Challenge

No switch support and conserve switch resources

to simultaneously achieve:

Scale

Multi-tenancy

Ease of configuration

State of the Art – Scale

- Most prior work is limited by one or more of:
  - New protocols
  - Modified control and/or data planes
  - Preferred topologies
  - Resources (such as table space) on switches

State of the Art – Multi-tenancy

- Traditional VLANs
  - Single tenant
  - Careful configuration
  - Cannot scale beyond 4K
- Mostly on segregation not virtualization

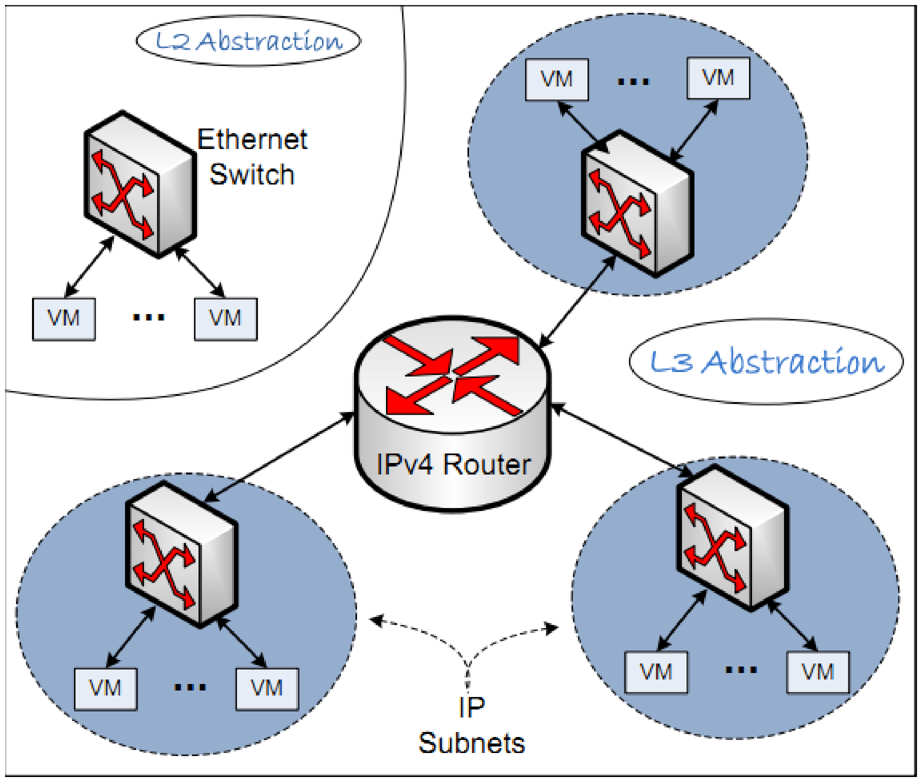
NetLord

- An encapsulation scheme
- A complementary switch configuration

→

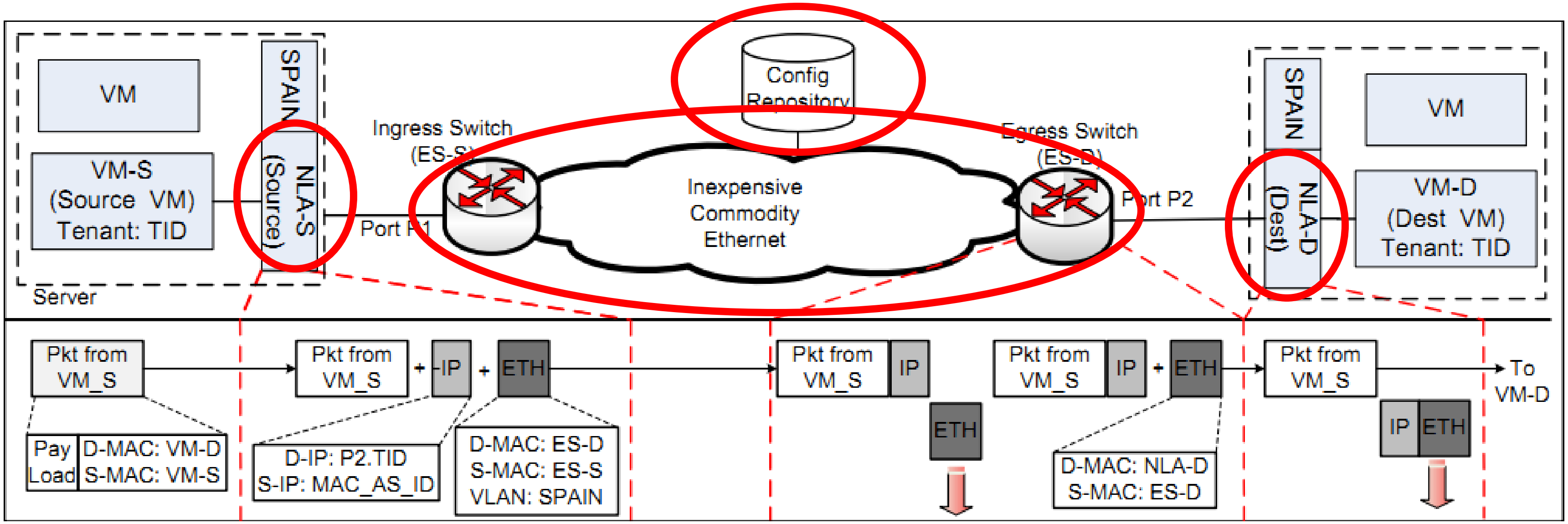
- Scalable multi-tenancy
- Ease of configuration
- Significant reduction in FIB requirements
- High bisection BW

A Tenant's View of NetLord



- One or more private MAC address space
- Full L2 L3 address-space virtualization
- Multiple tenants can use the same address

NetLord Components



- Fabric switches
- Configuration repository
- NetLord agents (NLA)

NetLord Encapsulation

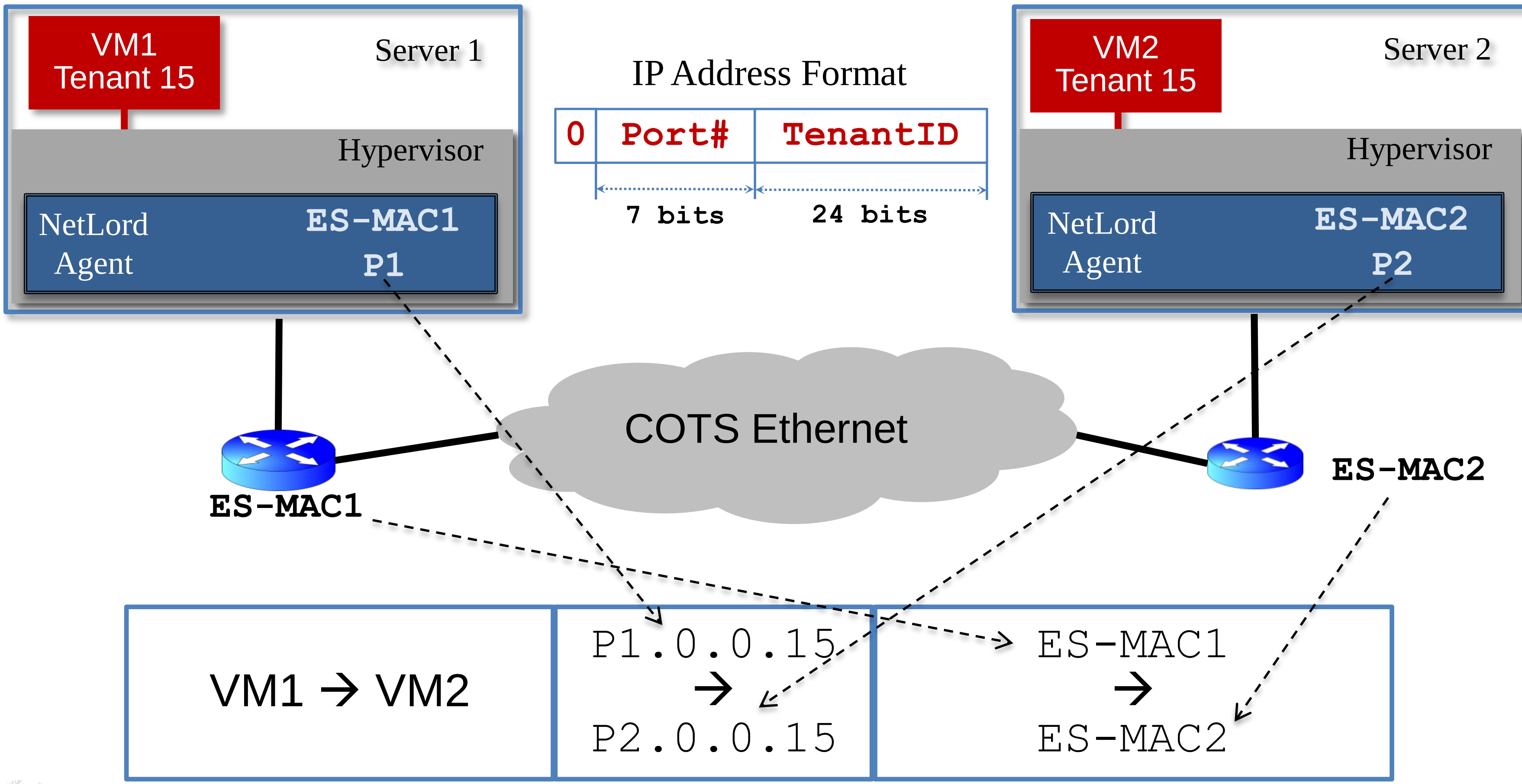
- Why encapsulate?
  - Unmodified VM packets onto the network
  - Excessive FIB pressure, FIB miss
  - MAC/IP address-spaces conflict

- Alternative: Rewrite headers
  - Rewrite with server MAC somewhat reduced FIB
  - Cannot identify the right VM on dst Server

NetLord Encapsulation

- Two headers: a MAC and an IP
- Reduced FIB pressure
  - Outer Src MAC = MAC of the Src edge switch
  - Outer Dst MAC = MAC of the Dst edge switch
- Correct delivery
  - Right edge switch: The outer MAC header
  - Right server: Right port # in the outer dest IP addr
  - Right VM: Tenant-ID frm outer dest IP + Inner dest MAC
- Clean abstraction
  - No assumptions about VM protocols and/or addressing

NetLord Encapsulation



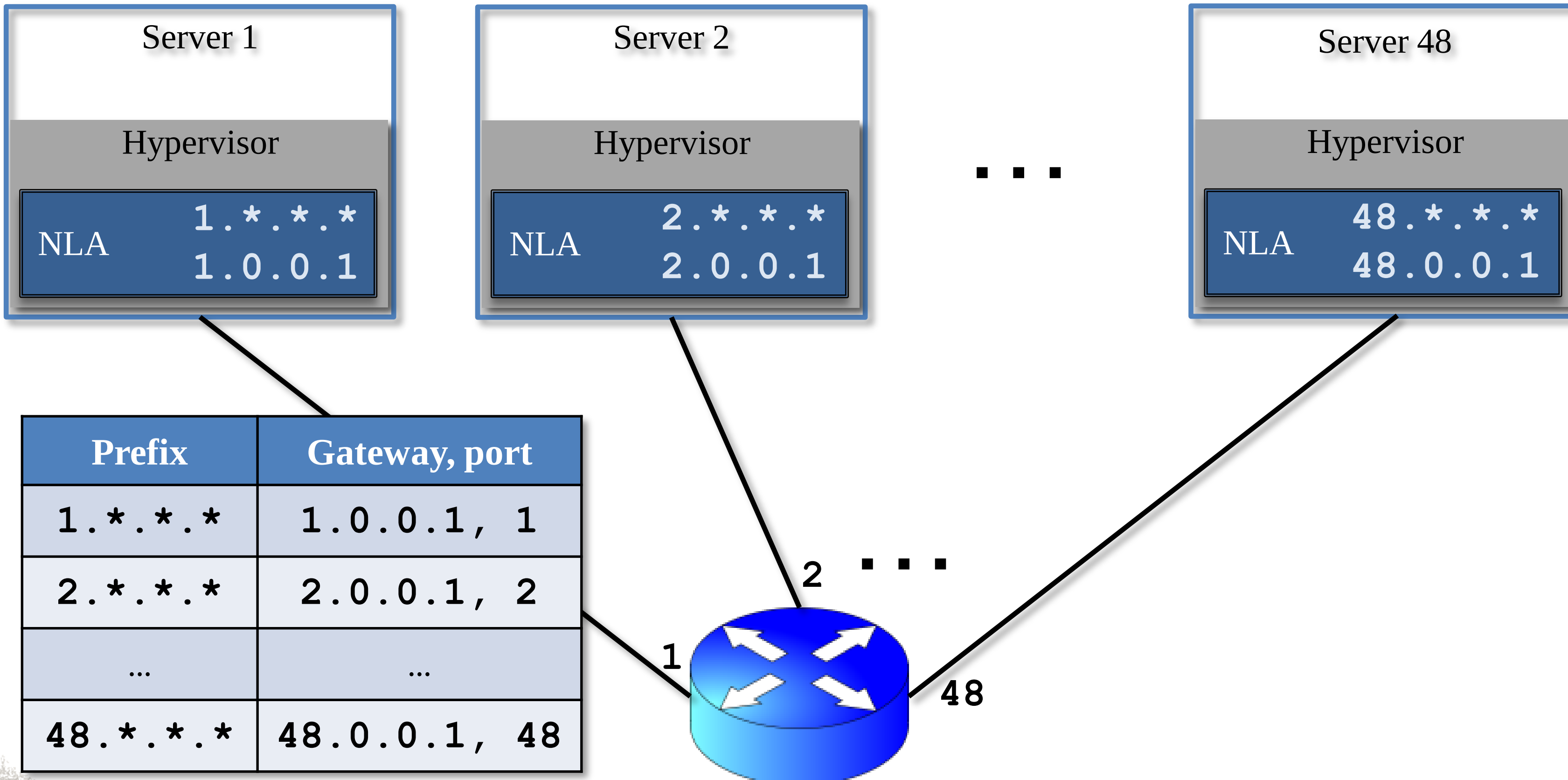
Switch Configuration

- Outer MAC hdr takes pkt to egress edge switch

- A switch on MAC Pkt addressed to itself
  - Strips MAC hdr and forwards based on IP hdr inside
  - Standard behavior

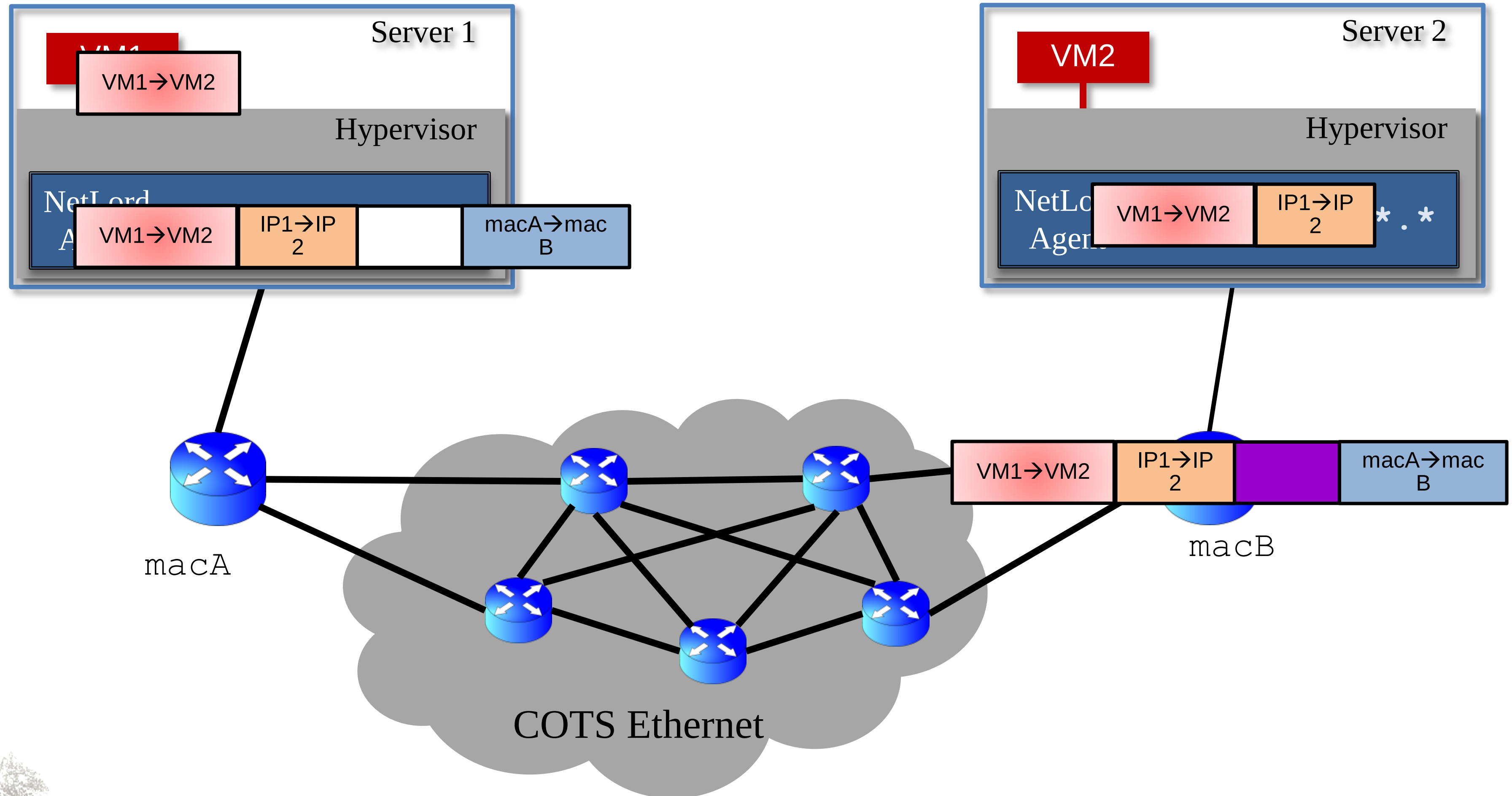
- Correct forwarding
  - Configure the L3 forwarding tables right
  - Make sure to match the server configs

Switch Configuration



Prefix	Gateway, port
1.*.*.*	1.0.0.1, 1
2.*.*.*	2.0.0.1, 2
…	…
48.*.*.*	48.0.0.1, 48

Putting It All Together



Evaluation

- Overhead of NLA
  - "ping" for latency
  - 1 / 2 - way Netperf for throughput
- Scalability of NetLord
  - Multi-tenant parallel shuffle workload

Evaluation - Overhead of NLA

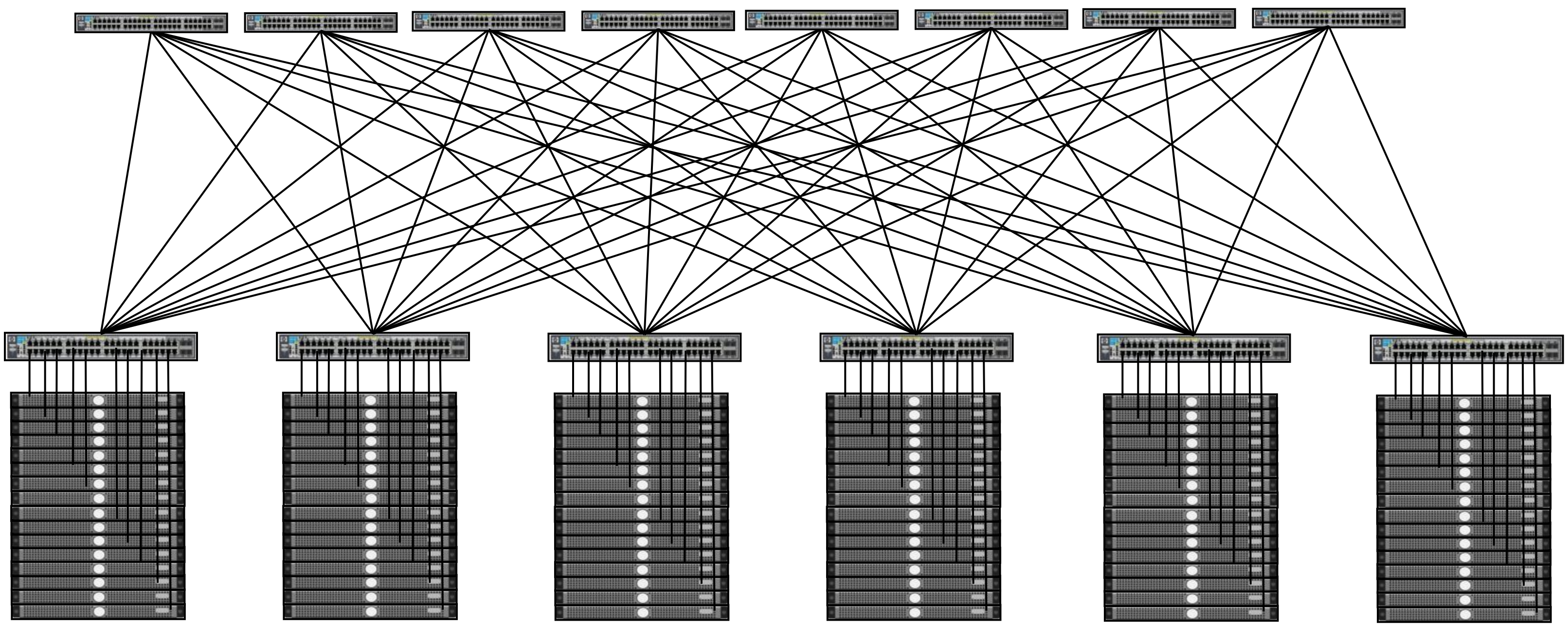
- Overhead of NLA

Case	Metric	PLAIN	SPAIN	NetLord
Ping	avg	97	99	98
(in μs)	min/max	90/113	95/128	93/116
NetPerf	avg	987.57	987.46	984.75
1-way	min	987.45	987.38	984.67
(in Mbps)	max	987.67	987.55	984.81
NetPerf	avg	1835.26	1838.51	1813.52
2-way	min	1821.34	1826.49	1800.23
(in Mbps)	max	1858.86	1865.43	1835.21

- encaping overheads are ignorable

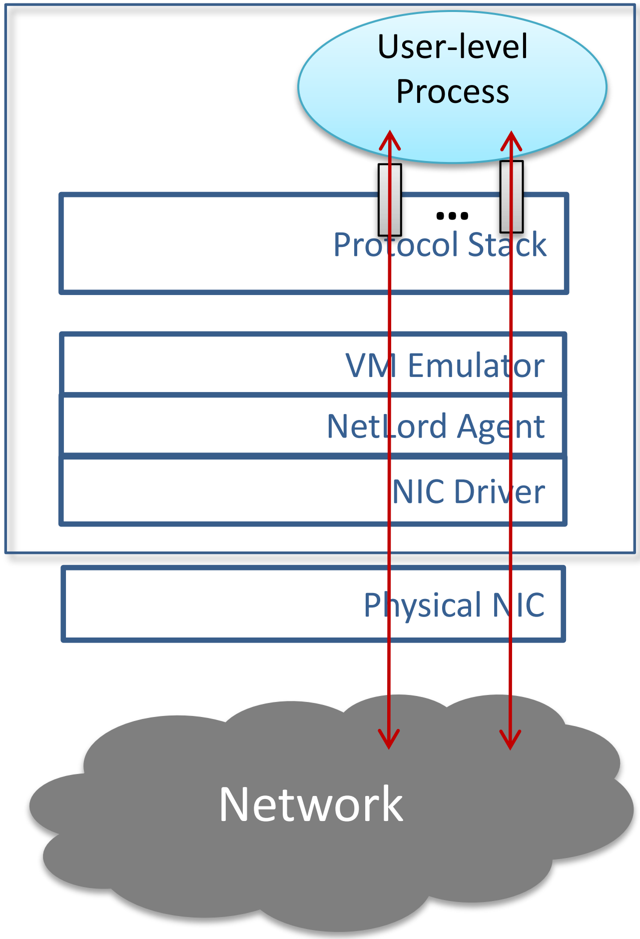
Evaluation - Scalability of NetLord

- 74 Servers in a 2-level fat-tree topology



Evaluation - Scalability of NetLord

- NLA Kernel module
- VM Emulator
  - A thin module above NLA
  - TCP flow -> emulated VM
  - Exports a virtual device
  - Re-writes MAC addresses
- Up to 3K VMs / Server
  - 74 VMs / Tenant
  - 200K VMs in all

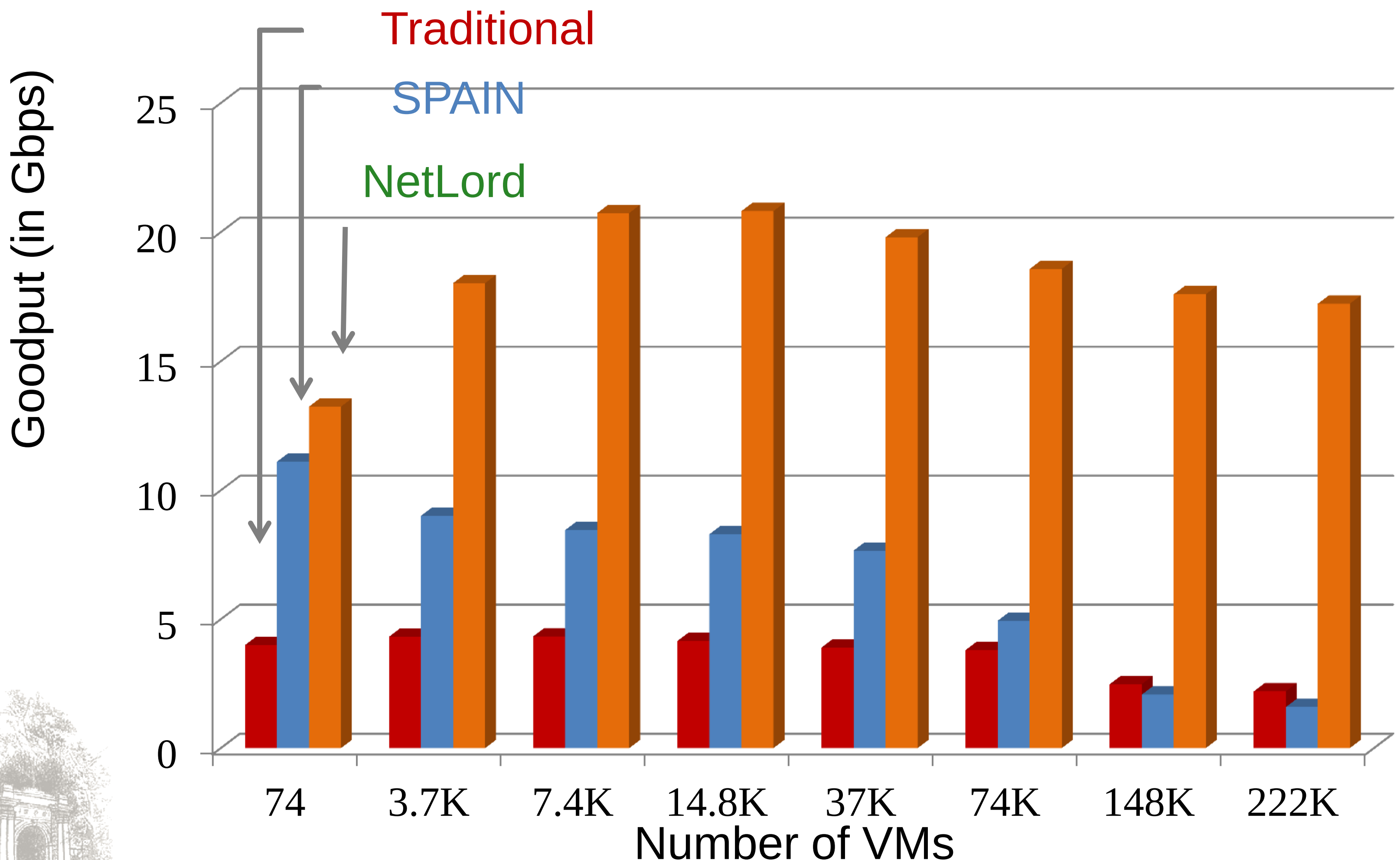


Evaluation - Scalability of NetLord

- Parallel shuffles
  - Emulating shuffle-phase of Map-Reduce jobs
  - Each shuffle: 74 mappers & 74 reducers
  - Each mapper transfers 10MB data to all reducers

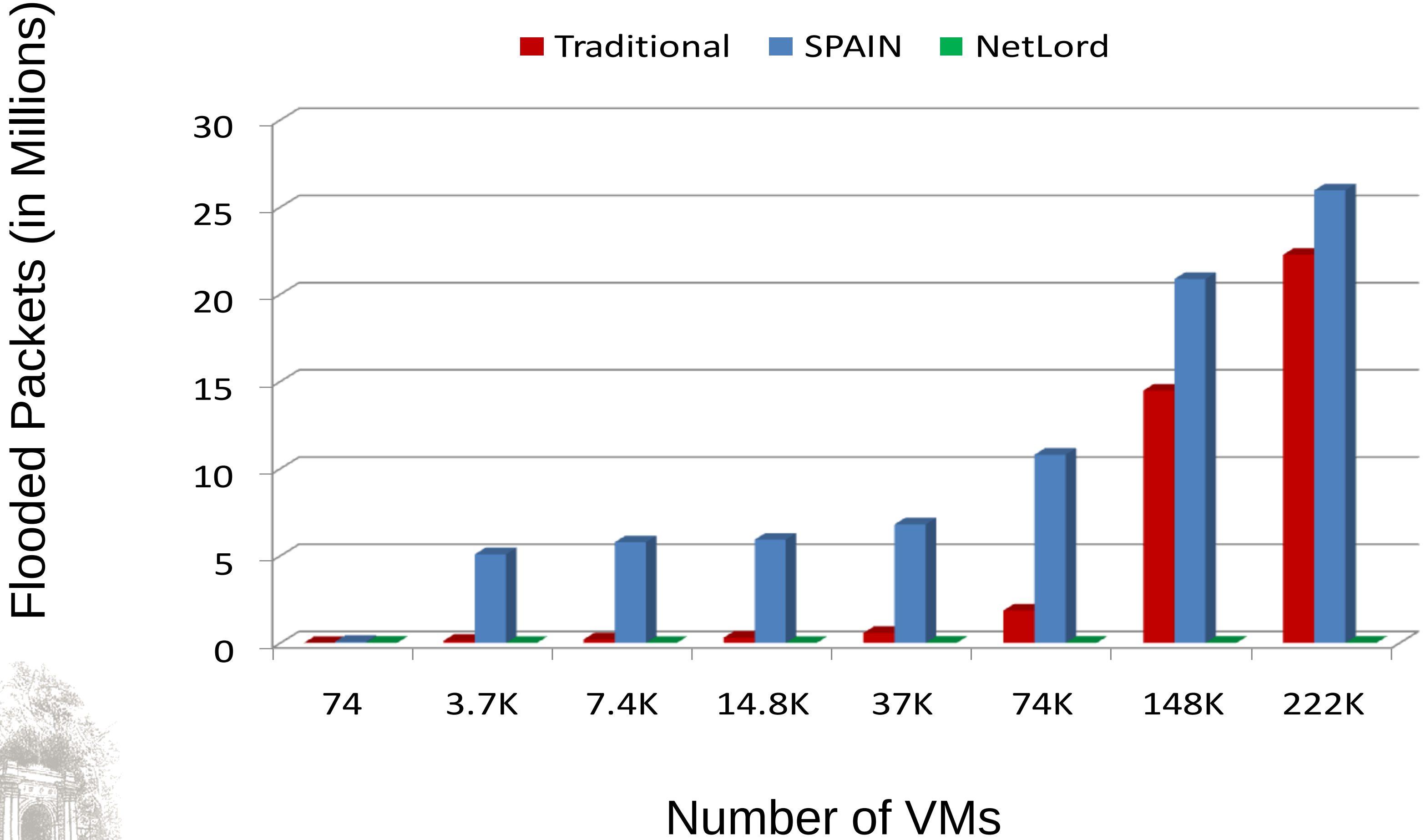
Evaluation - Scalability of NetLord

- Goodput



Evaluation - Scalability of NetLord

- Floods



Summary

- NetLord combines simple existing primitives in a novel fashion to achieve several out-sized benefits of practical importance:
  - Scale
  - Multi-tenancy
  - Ease-of-use
  - Bisection BW

Acknowledgements

- Almost the whole content comes from authors' slides presented at SIGCOMM 2011 and also their paper

- This slides is only for seminar use in NSLab

- For more information, please refer to the following links:
  - http://conferences.sigcomm.org/sigcomm/2011/papers/sigcomm/p62.pdf
  - http://conferences.sigcomm.org/sigcomm/2011/slides/s62.pptx

Discussion