



# Network Virtualization

Future Internet Communications Technologies

Prof. Dr. Panagiotis Papadimitriou



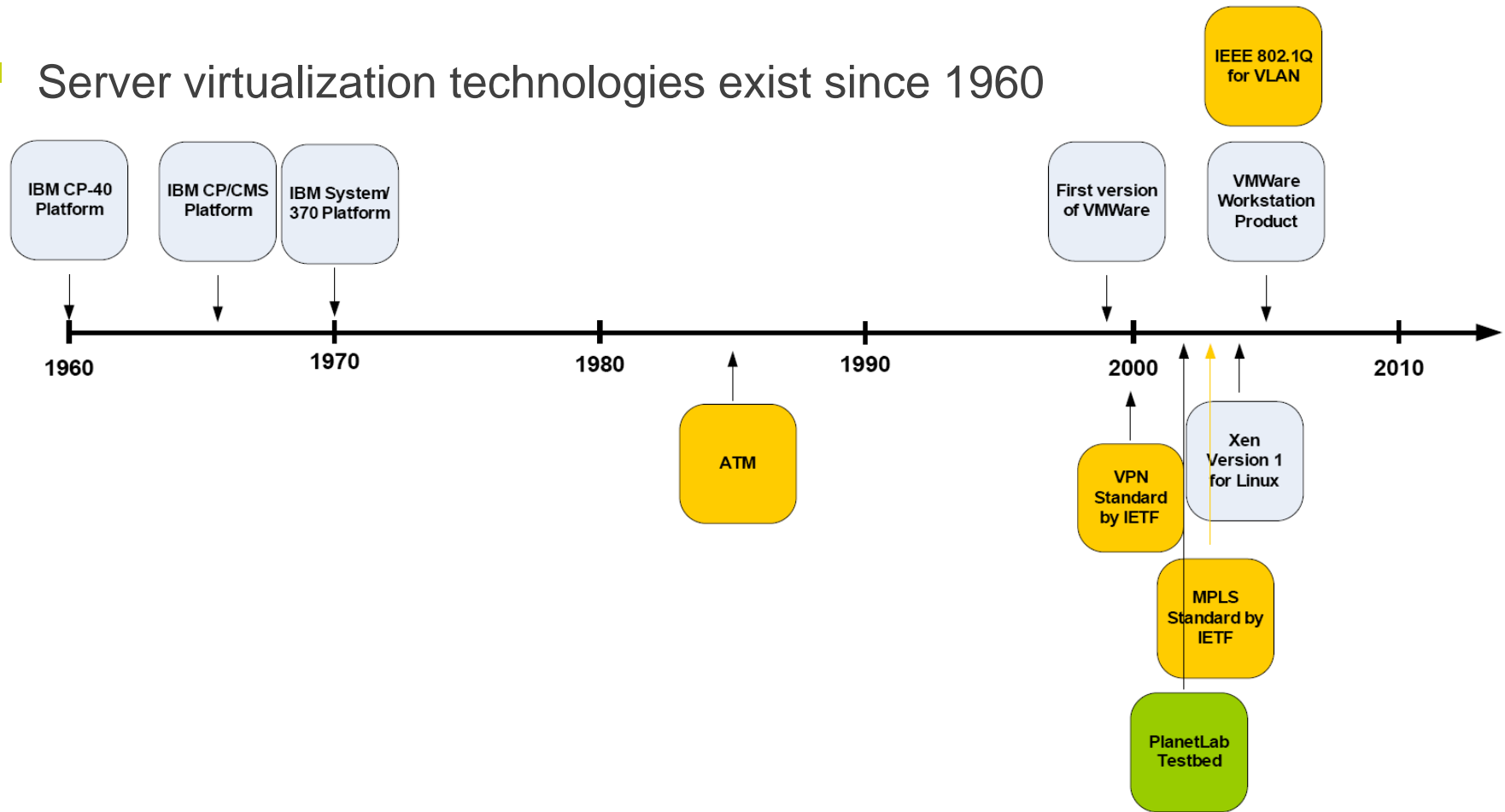
- Introduction to Network Virtualization
- Network Virtualization Technologies
- Server Virtualization
- Link Virtualization
- Network Interface Virtualization
- Router Virtualization
- Virtual Network Embedding
- Virtual Network Embedding Across Multiple Substrate Networks
- Virtual Link Setup



# Introduction to Network Virtualization

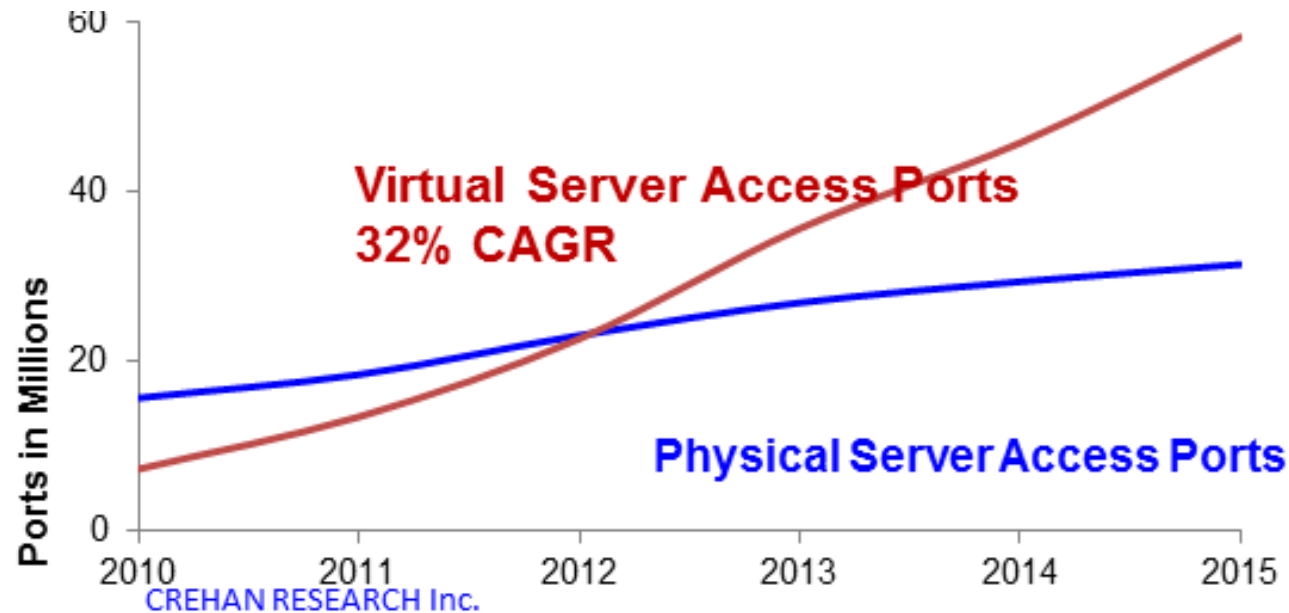


- Virtualization is not a new concept (e.g., virtual memory)
- Server virtualization technologies exist since 1960



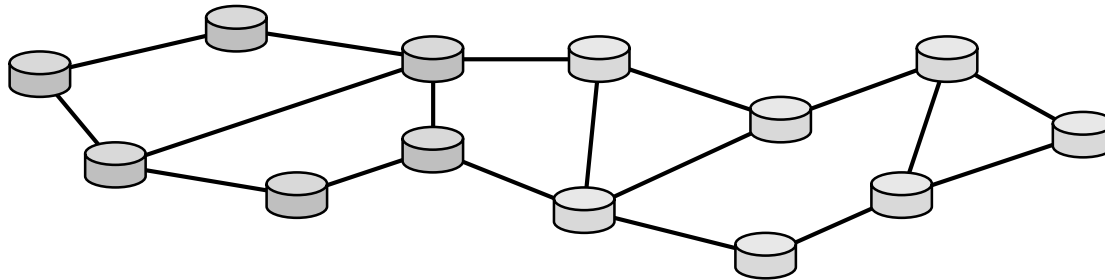
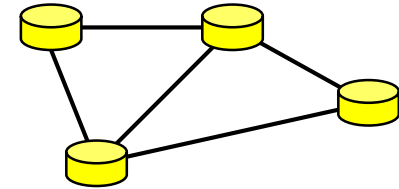
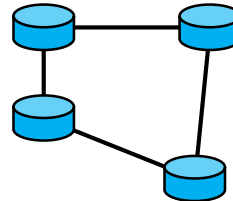
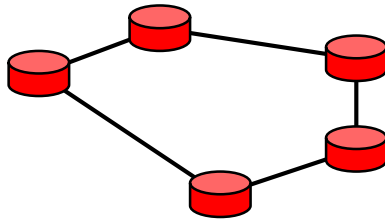


- The number of virtual access ports increases by 32%





- Multiple virtual networks can coexist on top of a shared physical (substrate) network.
- A virtual network is composed of virtual nodes and links.





- Resource sharing (e.g., CPU, bandwidth)
- Independent network management
- Abstraction
- Flexible network management
  - Elasticity
  - Fault tolerance
  - Easier maintenance of physical equipment
- Excellent platform for experimentation
  - Isolating experimental from production traffic
- Innovation in architecture design
  - Concurrent deployment of multiple network architectures



- Virtual networks are conceptually similar to overlays:
  - Virtual networks provide a topology composed of virtual nodes and links on top of one or multiple physical networks
- Overlays:
  - introduce mechanisms above the transport layer to control the routing across the overlay nodes (e.g., multicast)
  - are typically deployed at the network edge
  - do not lead to any fundamental changes in the physical infrastructure
- Virtual networks can provide higher control on routing and packet delivery by utilizing virtual routers with customized protocol stacks and enhanced programmability

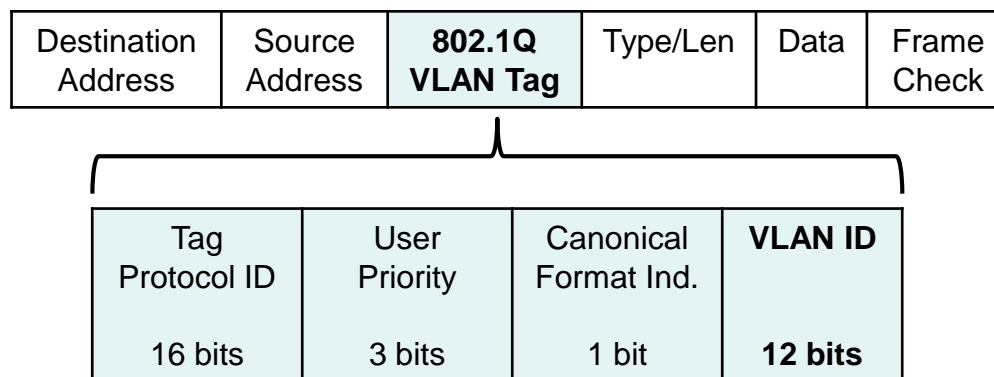


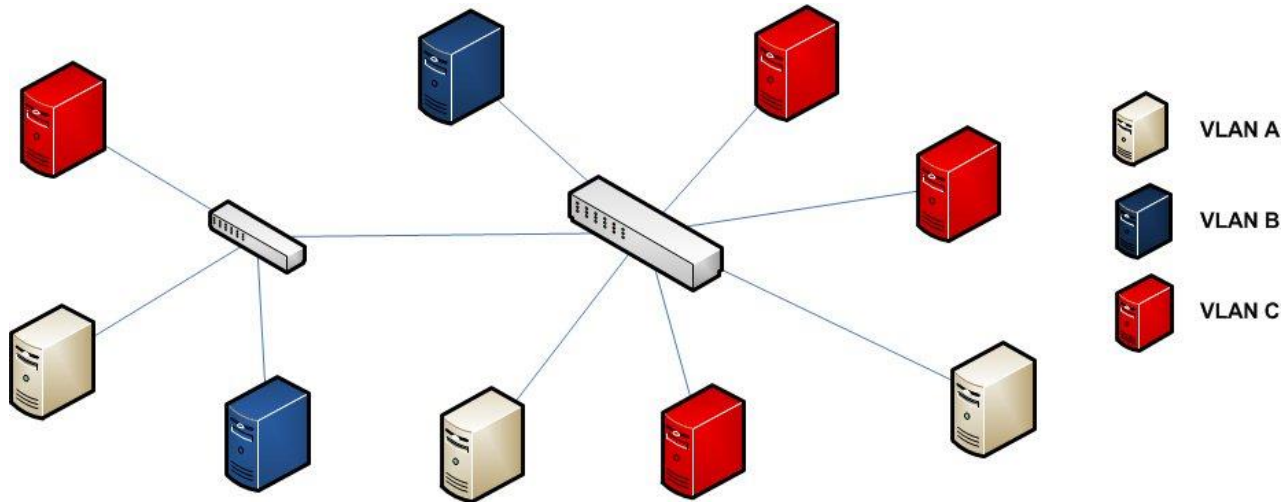


# Network Virtualization Technologies

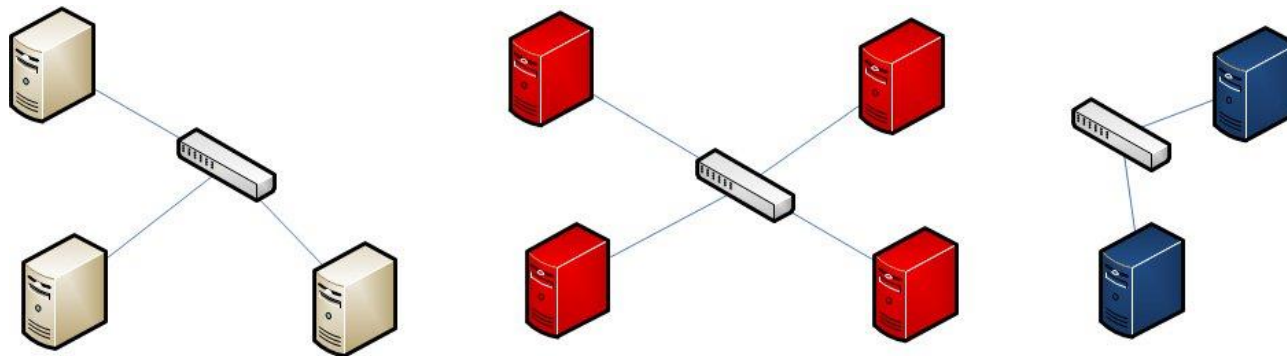


- VLANs isolate groups of hosts into separate broadcast domains using the IEEE 802.1Q standard
  - ✓ Widely available (supported by many switches)
  - ✓ Up to 4096 VLANs
  - ✗ Limited capability for WANs





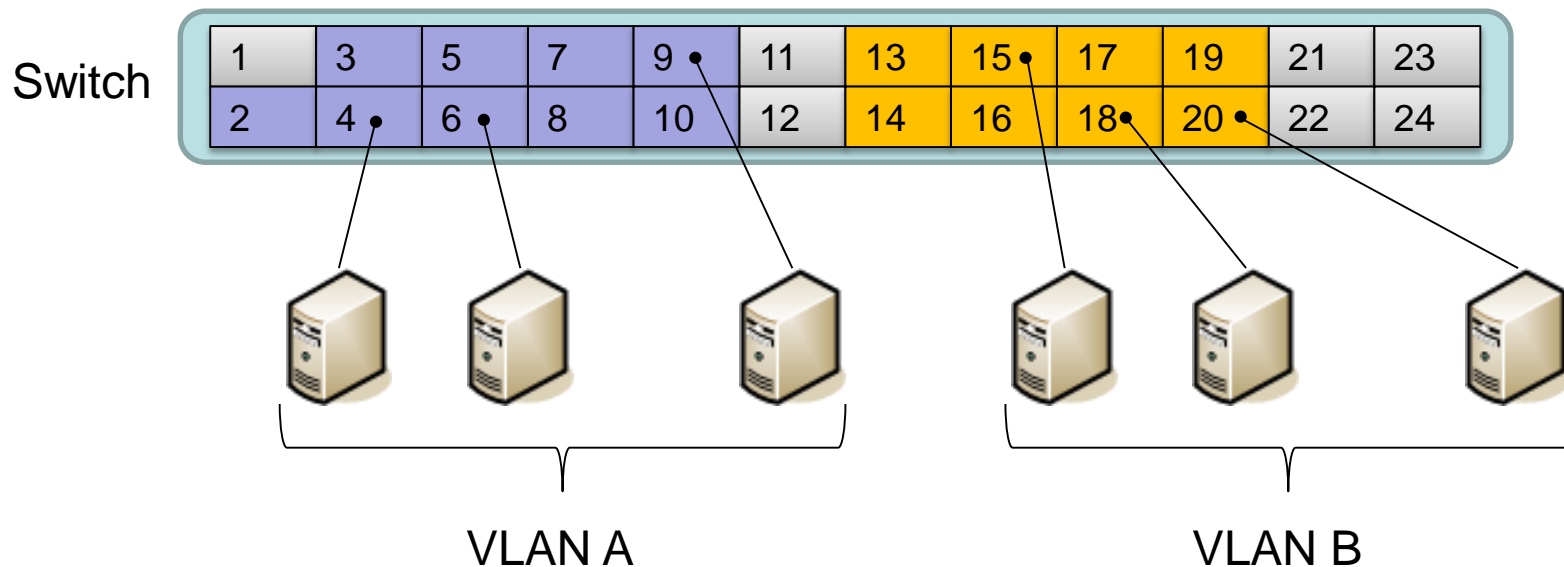
Physical View



Logical View

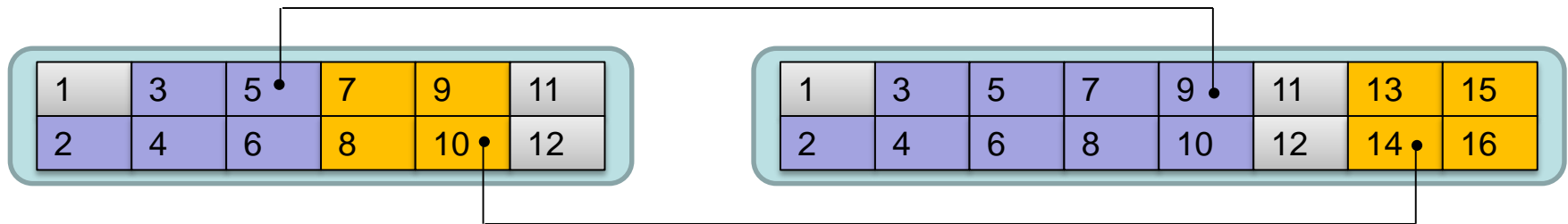


- Switch ports are assigned with VLAN IDs
  - The switch software maintains a table of port-to-VLAN mappings



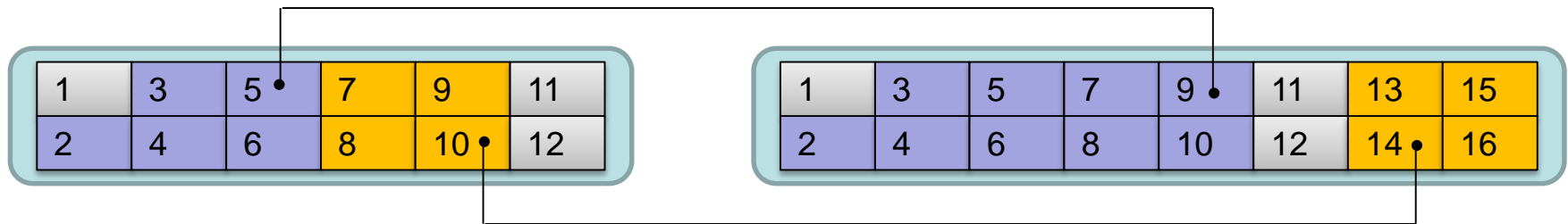


- Separate connection per VLAN
  - Wastes switch ports
  - Static configuration

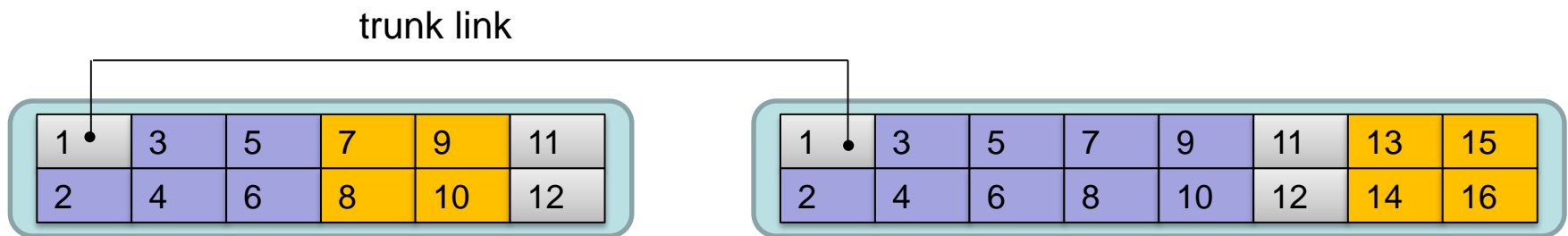




- Separate connection per VLAN
  - Wastes switch ports
  - Static configuration



- VLAN trunking
  - All frames (irrespective of VLAN ID) are forwarded over the trunk link





- VPNs comprise a service that provides:
  - remote access to a private network via tunneling
  - secure access via authentication
  - traffic isolation
- VPNs do not offer any benefits or flexibility in terms of management (e.g., independent resource management) as virtual networks do
- VPN setup usually incurs a long delay

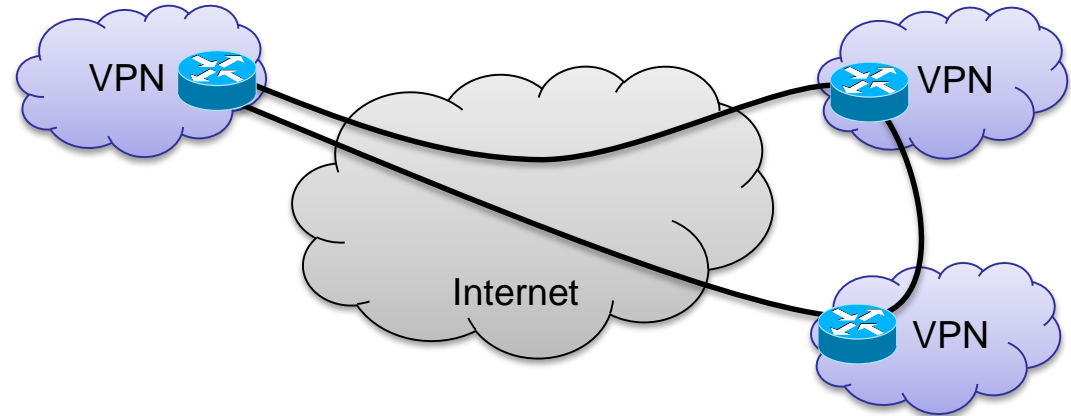


- A tunnel is a method of sending data by encapsulating the data and its protocol information within a different transmission unit
- VPNs encompass various tunneling technologies:
  - Layer 2:
    - Point-to-point:
      - MPLS, L2TP, PPTP, PPP
    - Point-to-multipoint:
      - Virtual Private LANs
  - Layer 3:
    - IP-in-IP
    - Generic Routing Encapsulation (GRE)
    - IPsec





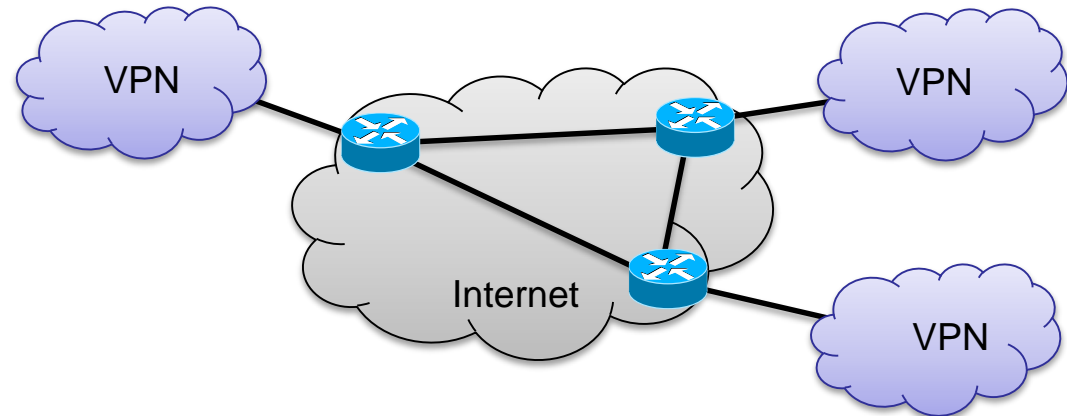
- Customer model



- Provider-provisioned VPN (PPVPN), e.g., BGP/MPLS VPN

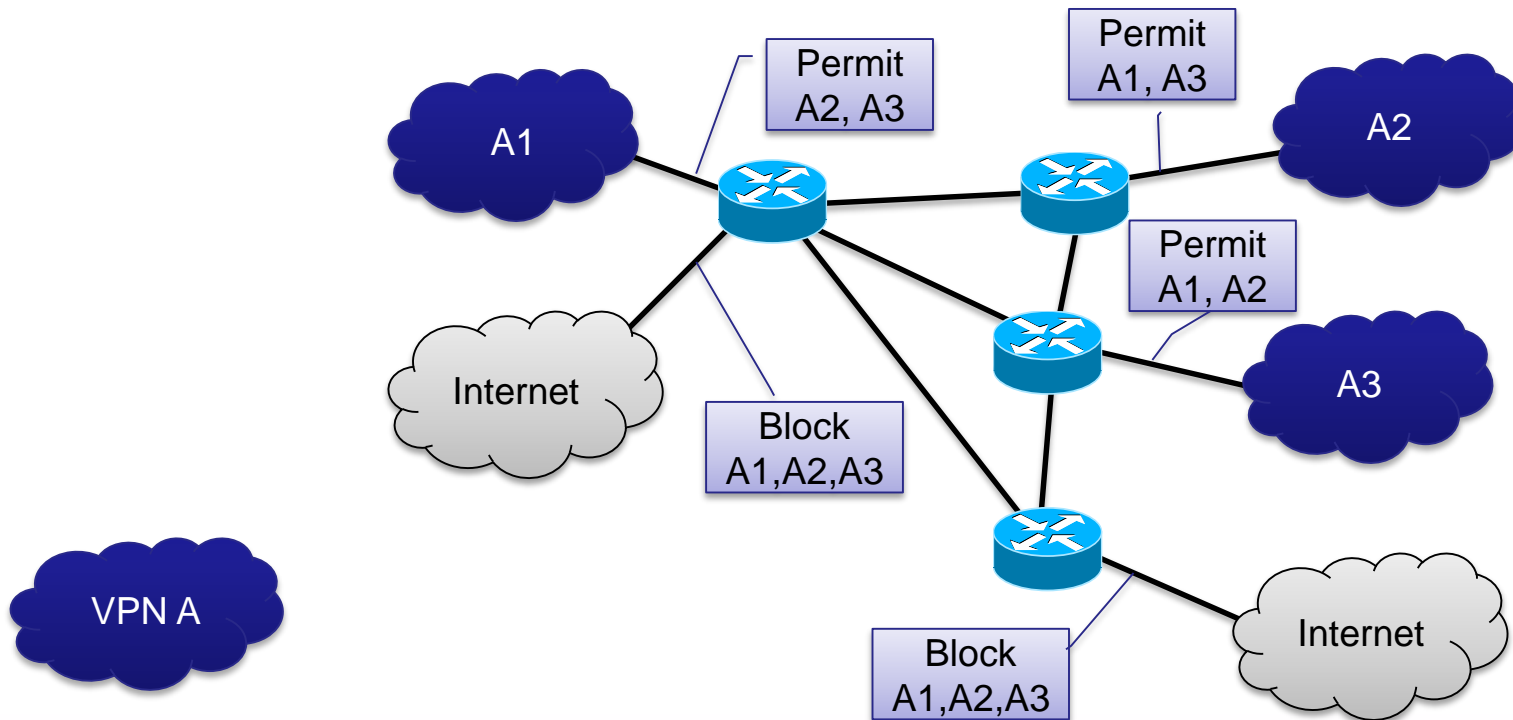
- Logical routers with own:

- Routing protocol
    - Forwarding table
    - Route filtering



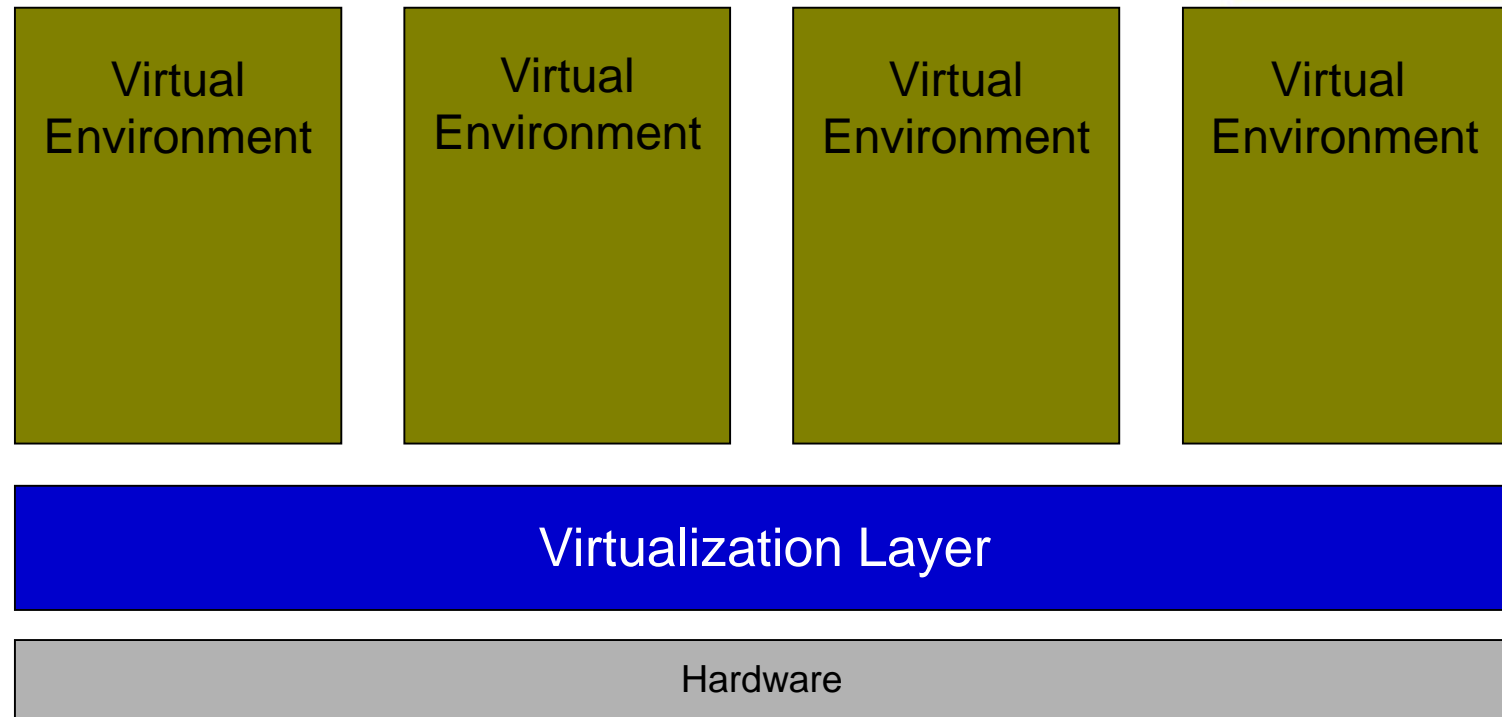


- Route filtering
  - Controls route propagation so that:
    - networks within a VPN receive route advertisements for other networks in the same VPN
    - networks not in same VPN do not receive these advertisements

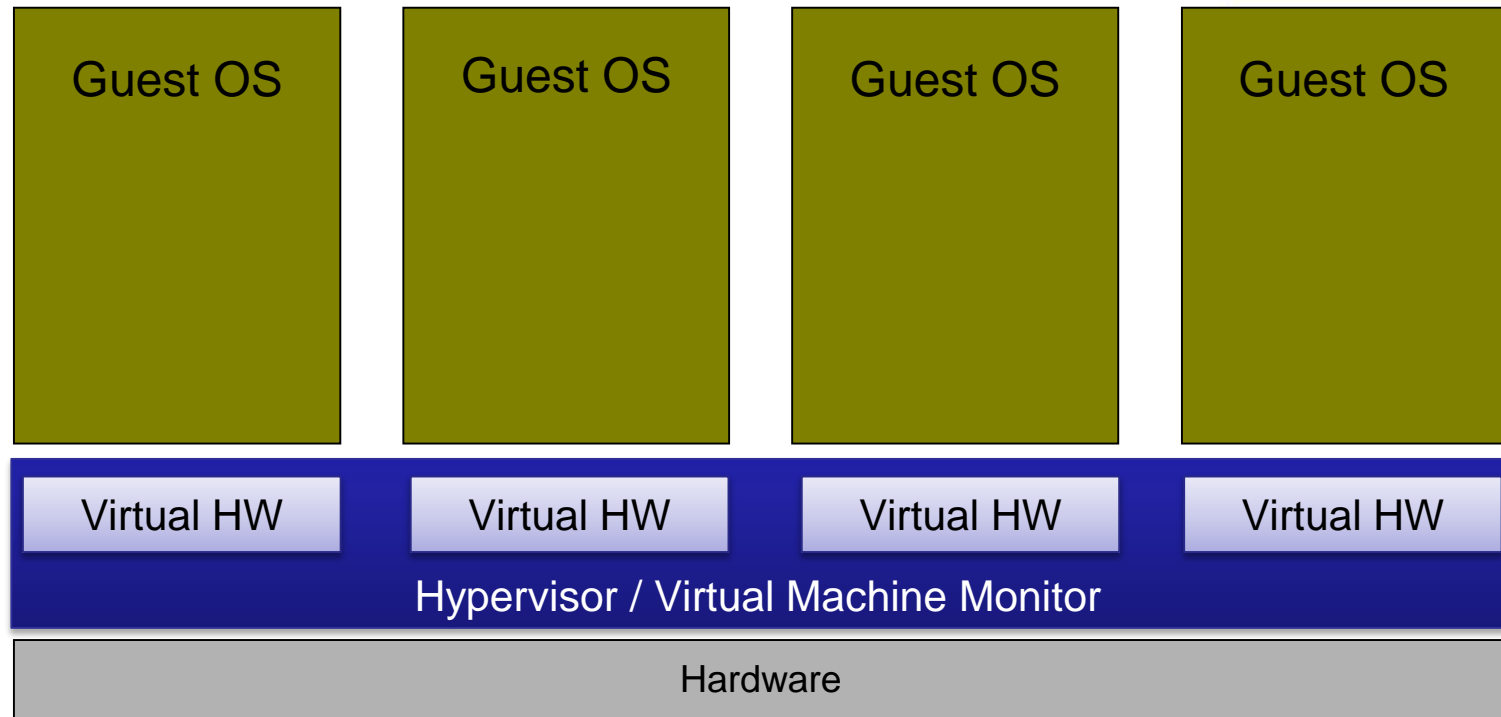




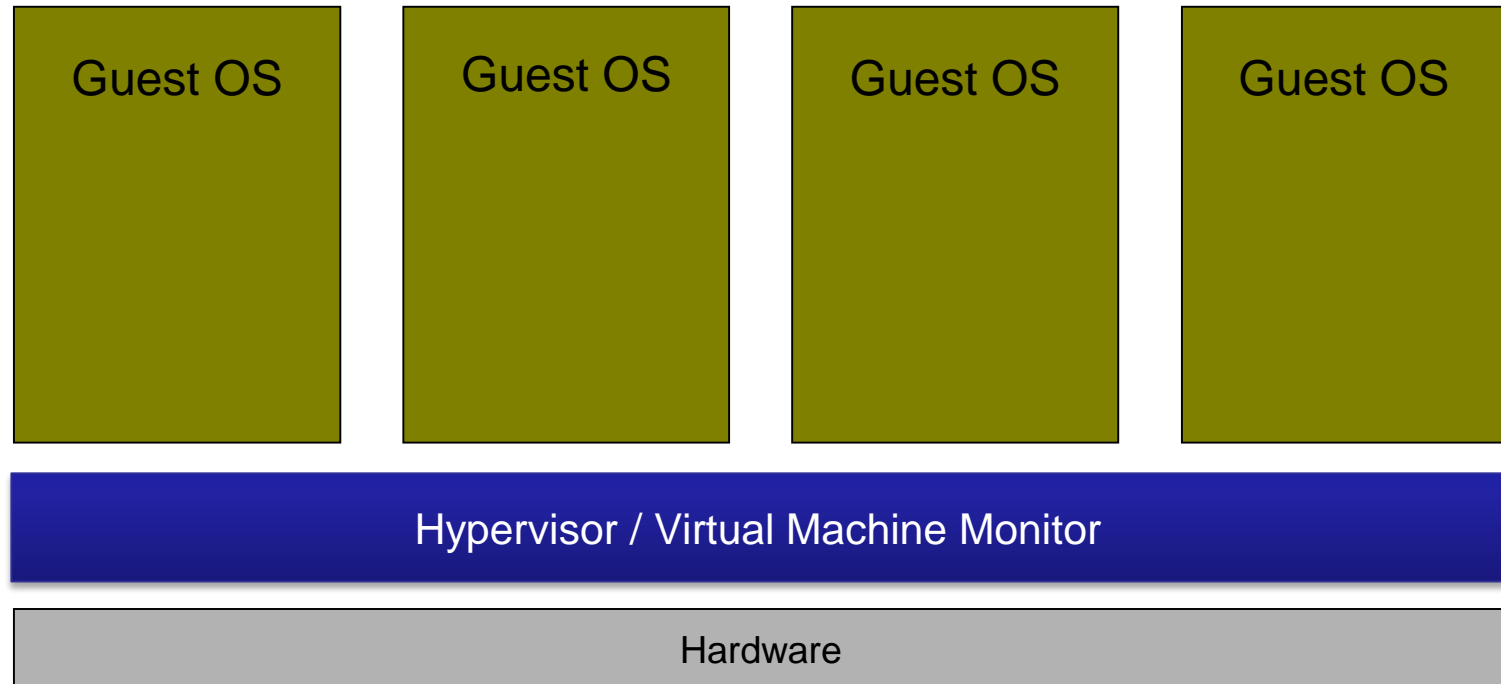
# Server Virtualization



- Server virtualization technologies:
  - aim to provide efficient resource (e.g., CPU, memory) sharing and isolation
  - allow multiple operating systems (OS) to run concurrently on a single host



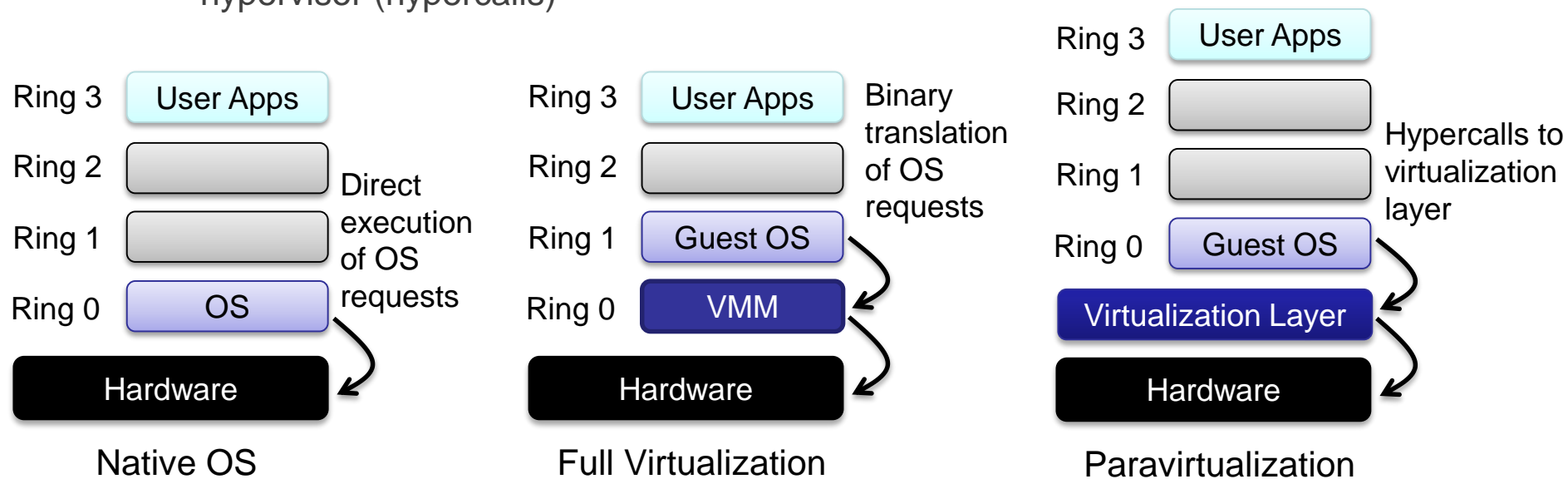
- Full Virtualization (e.g., KVM):
  - provides fully emulated virtual machines (BIOS, devices, memory management) for running multiple OSes
  - allows guest domains to directly execute privileged operations
  - accommodates unmodified guest OSes

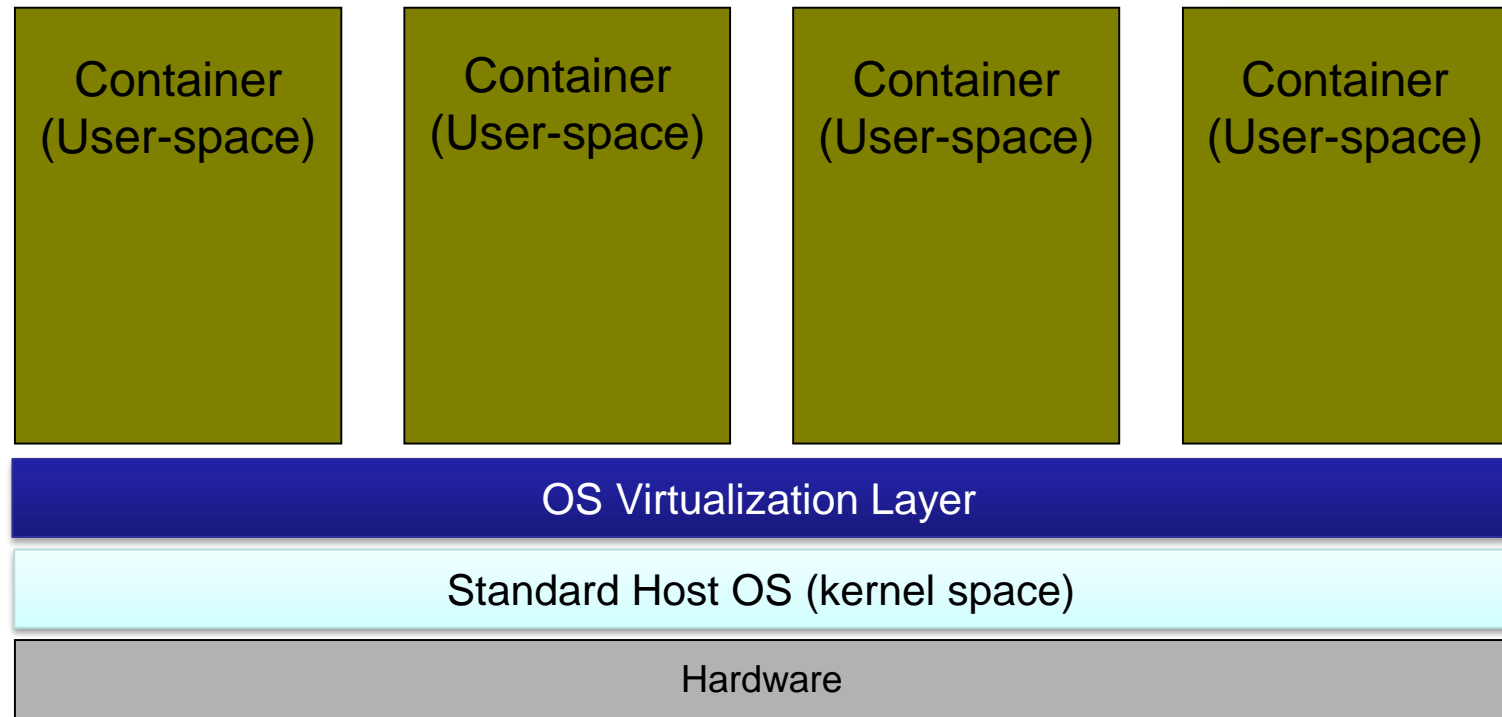


- Paravirtualization (e.g., Xen):
  - provides a hypervisor that runs on top of the hardware
  - does not allow guest domains to directly execute privileged operations (permission is required via hypercalls)
  - requires modified guest OSes



- Protection rings represent hierarchical levels of privilege within the architecture of a computer system
- In full virtualization, VMM provides emulation to handle and modify privileged operations made by unmodified guest OS kernels (as running at Ring 0).
- In paravirtualization, the guest OS kernel is modified to run on the hypervisor
  - Privileged operations that will only run at Ring 0 are replaced with calls to the hypervisor (hypercalls)



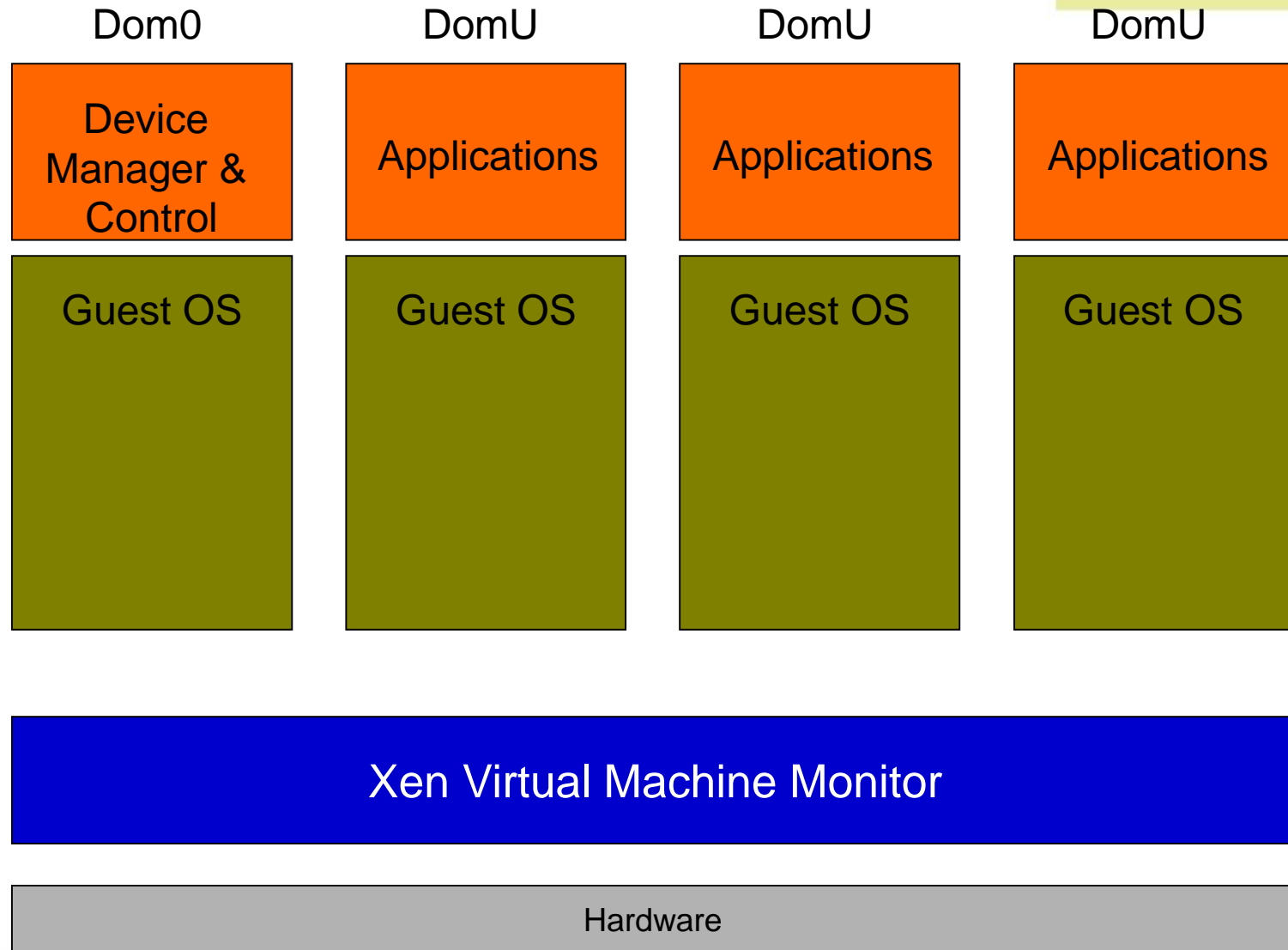


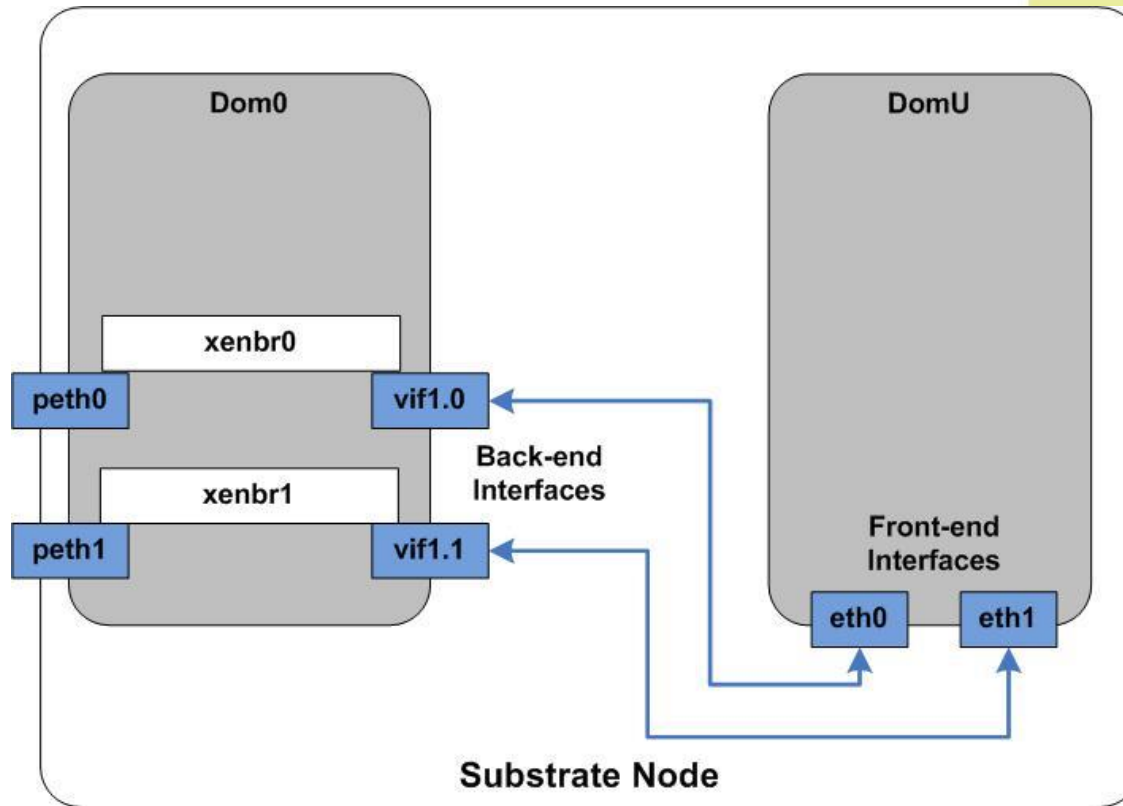
- OS-based virtualization (e.g., OpenVZ, VServer):
  - creates multiple partitions of operating system resources (containers)
  - provides a single kernel instance for all containers





- Full Virtualization
  - Maximum isolation
  - Performance penalty
  
- Paravirtualization (e.g., Xen, Denali)
  - High flexibility (e.g., different guest OSes)
  - Adequate isolation
  - Sub-optimal performance due to overhead
  
- OS-based virtualization
  - High performance due to small overhead
  - Limited flexibility (single instance of kernel)
  - Lower level of isolation





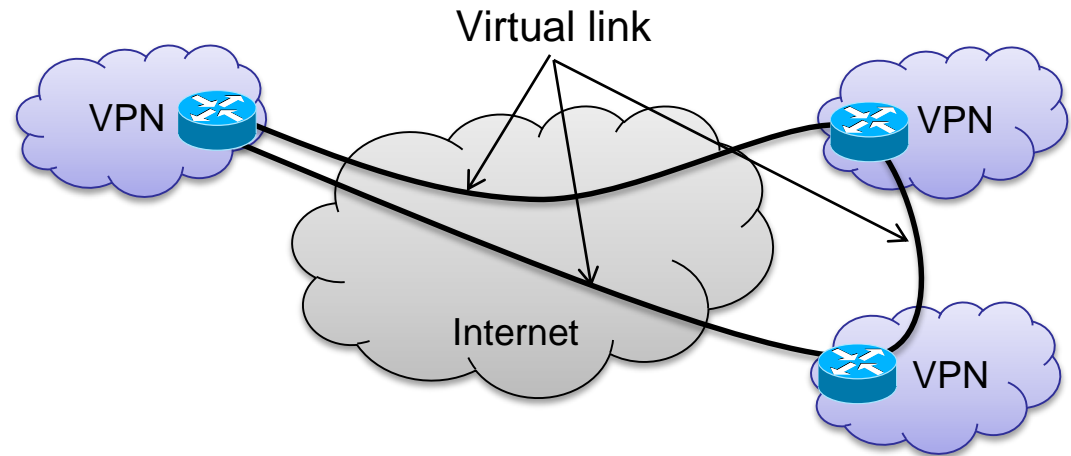
- A NIC driver is split into a front-end and back-end driver
- Xen creates I/O channels to connect front-end with back-end interfaces
- A Linux bridge typically connects a virtual back-end with a physical interface



# Link Virtualization

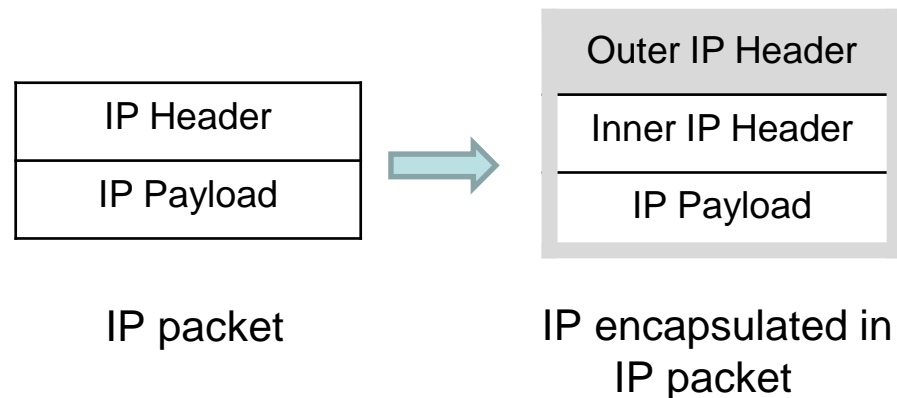


- Virtual links (a.k.a. tunnels) give the appearance of a single link (“pseudo-wire”), although they might span multiple hops
- Various technologies for link virtualization:
  - IP-in-IP
  - Multi-Protocol Label Switching (MPLS)
  - Generic Routing Encapsulation (GRE)
  - Ethernet over IP (EtherIP)

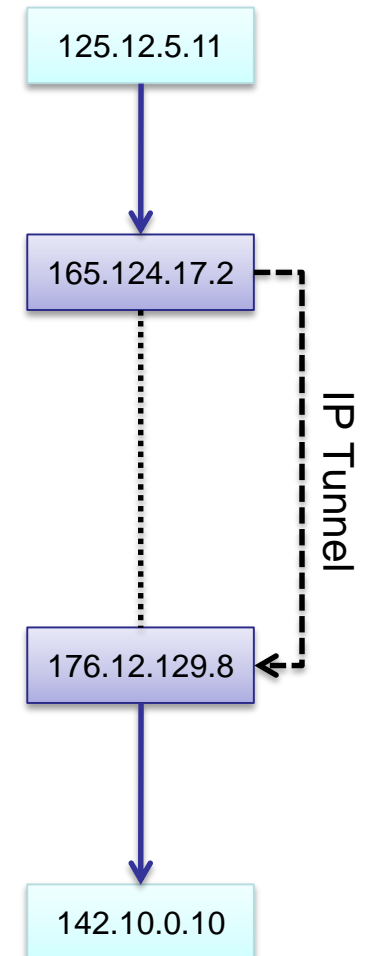
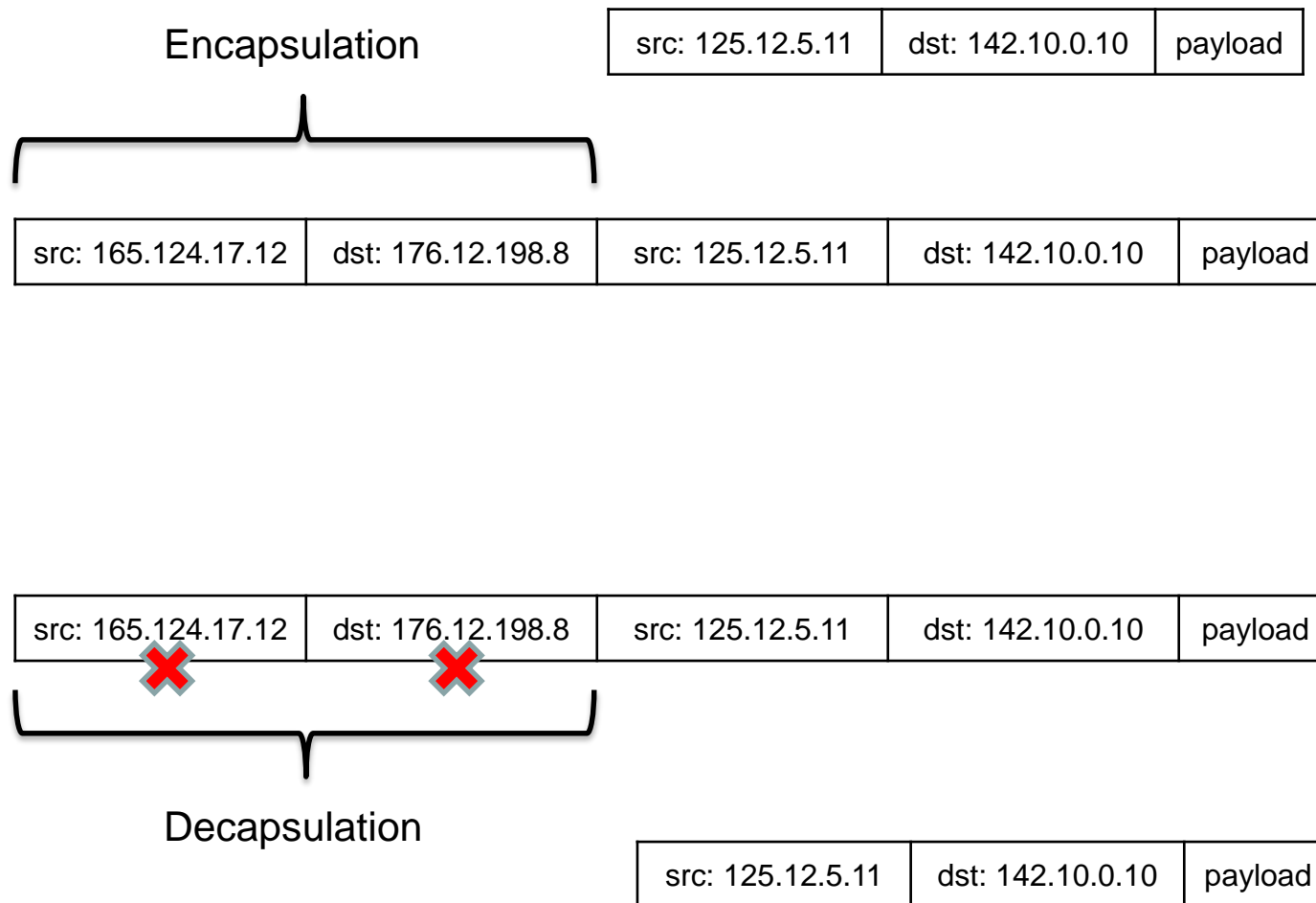




- IP-in-IP encapsulates IP packets into other IP packets (v4 or v6)
  - ✓ Widely available
  - ✗ Encapsulation overhead (20 bytes for IPv4)



# IP-in-IP Example

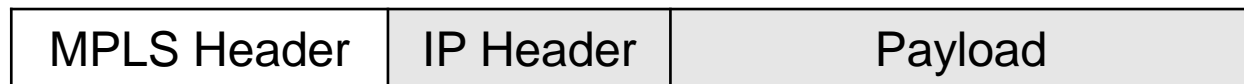




- MPLS encapsulates IP packets into frames (Ethernet, PPP)
  - ✓ Reliable link virtualization
  - ✓ Fast lookup based on a fixed-length label
  - ✓ Deployed by ISPs



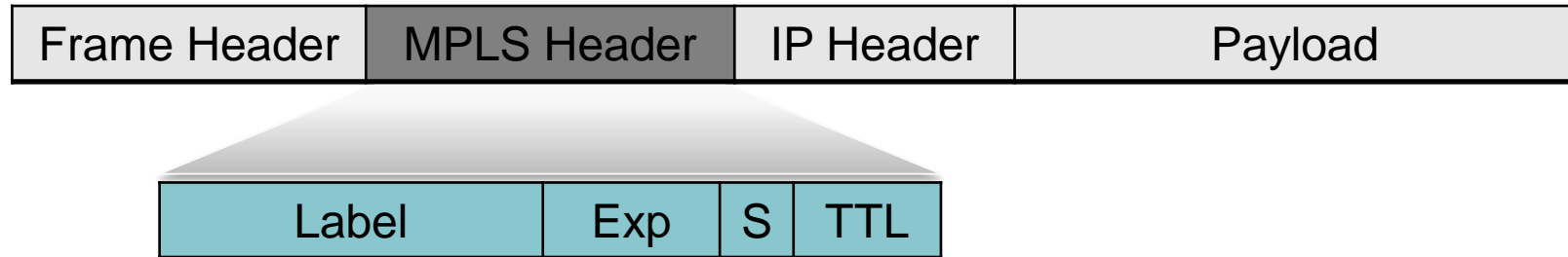
IP packet encapsulation



MPLS packet encapsulation







- MPLS header fields:
  - 20-bit Label value
  - 3-bit Traffic Class field for QoS
  - 1-bit field for stacked MPLS headers
  - 8-bit TTL (Time-to-Live)



- An MPLS-capable router (a.k.a. label switch router – LSR) maintains MPLS forwarding tables:

in label	out label	prefix	output interface
8	13	15.1.3	1
12	4	15.4.6	2

- An LSR forwards packets as follows:
  - examines the packet label value
  - performs a label lookup to determine the “out label” and the output interface
  - swaps the packet label value with the “out label” value
  - forwards the packet to the output interface



in label	out label	prefix	output interface
–	1	15.1.3	1
–	5	15.4.6	2

- An edge MPLS-capable router (a.k.a. label edge router – LER) forwards packets as follows:
  - examines the packet IP header
  - performs an IP lookup to determine the “out label” and the output interface
  - sets the “out label” value to the packet label value
  - forwards the packet to the output interface

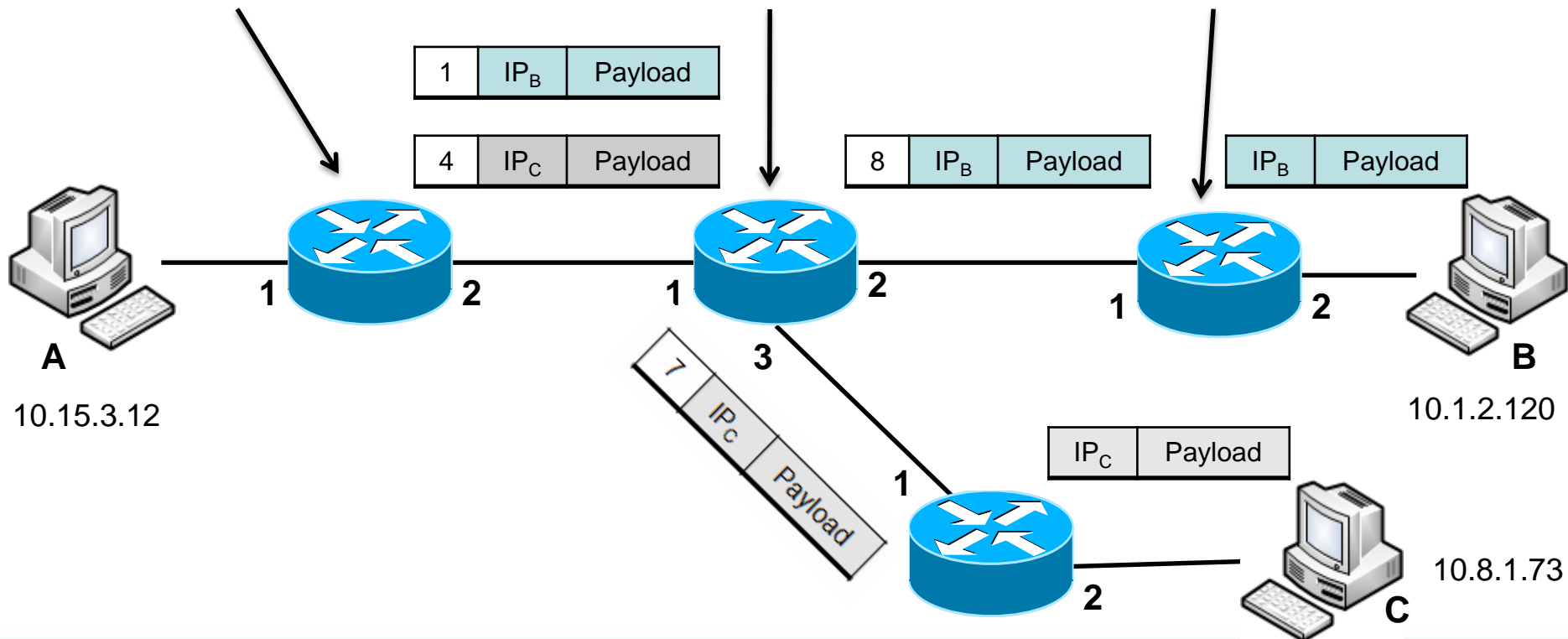
# MPLS Example



in label	out label	prefix	output interface
–	1	10.1.2	2
–	4	10.8.1	2

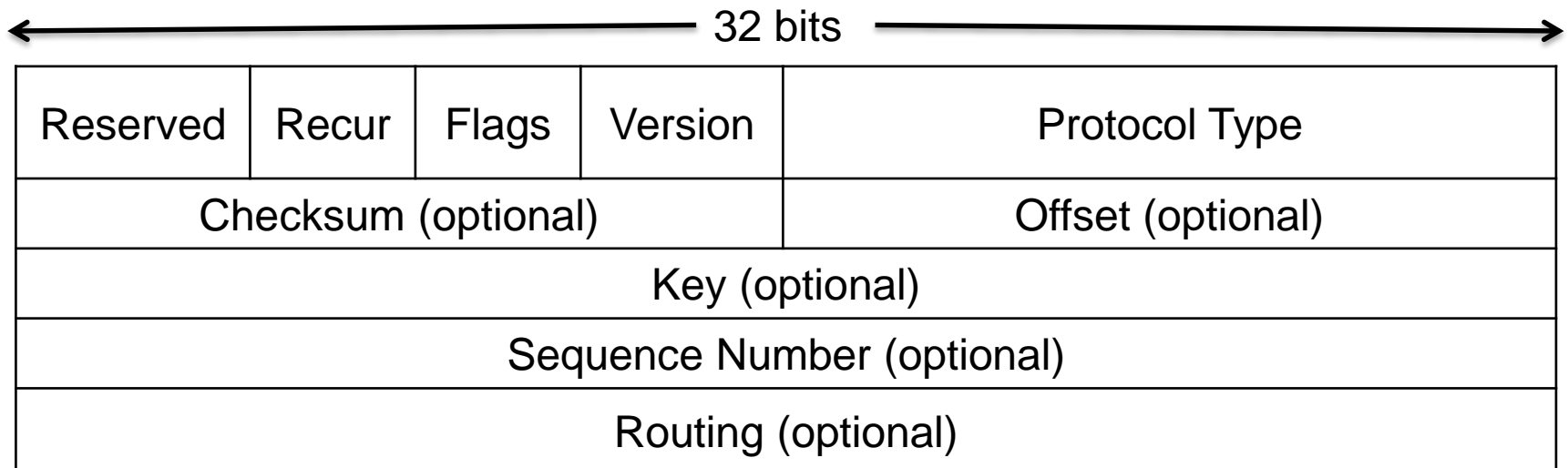
in label	out label	prefix	output interface
1	8	10.1.2	2
4	7	10.8.1	3

in label	out label	prefix	output interface
8	–	10.1.2	2



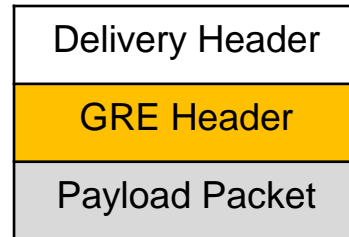


- Generic Routing Encapsulation (GRE)
  - ✓ General packaging protocol (can package any protocol's packets)
  - ✓ 4-byte (optional) key for de-multiplexing
  - ✓ Small encapsulation overhead (4 bytes without optional header fields)

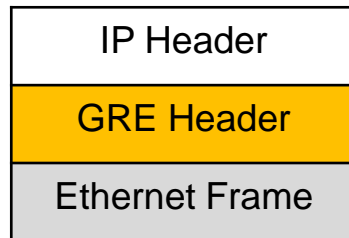




- A GRE encapsulation packet has the form:

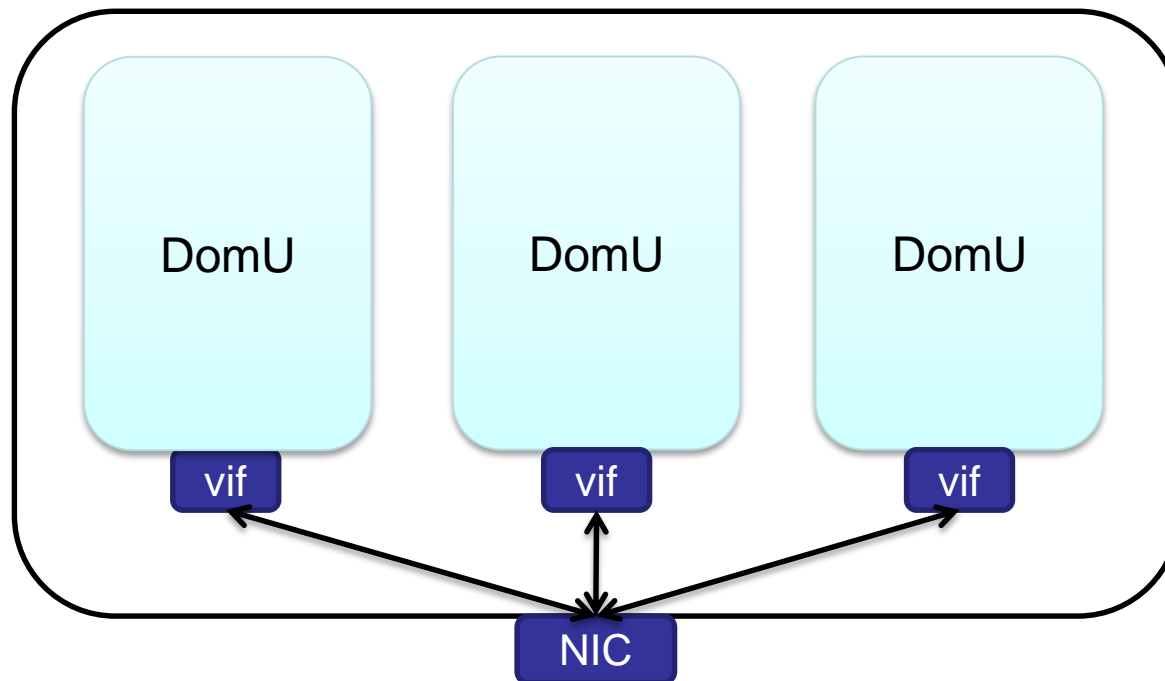


- Ethernet over GRE (EGRE):



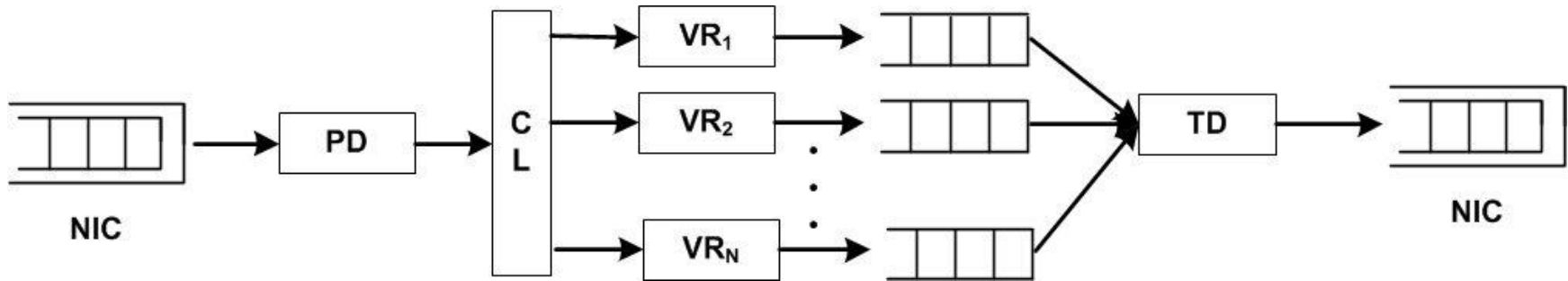


# Network Interface Virtualization



- Multiple virtual nodes/routers might have to share a physical network interface
  - Packets should be classified among the virtual nodes/routers

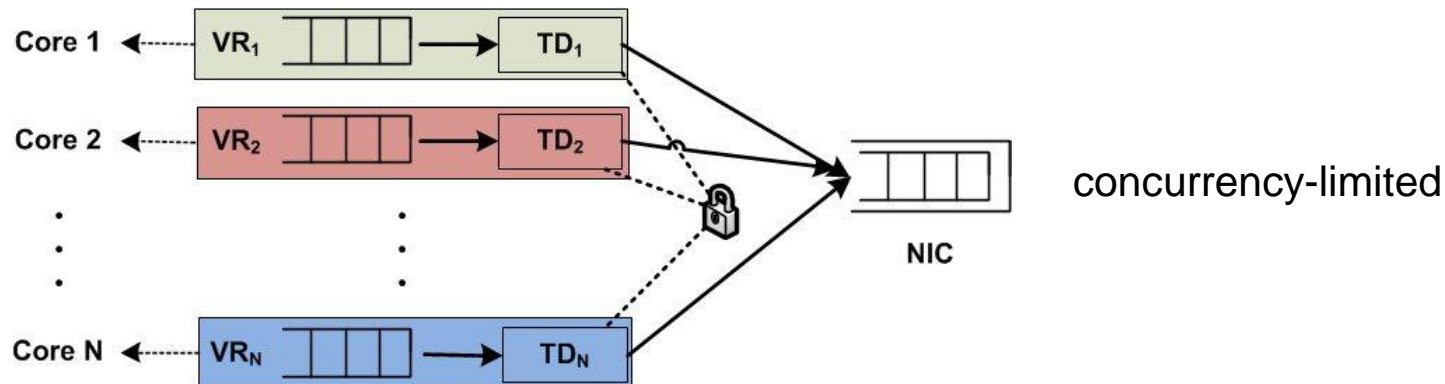
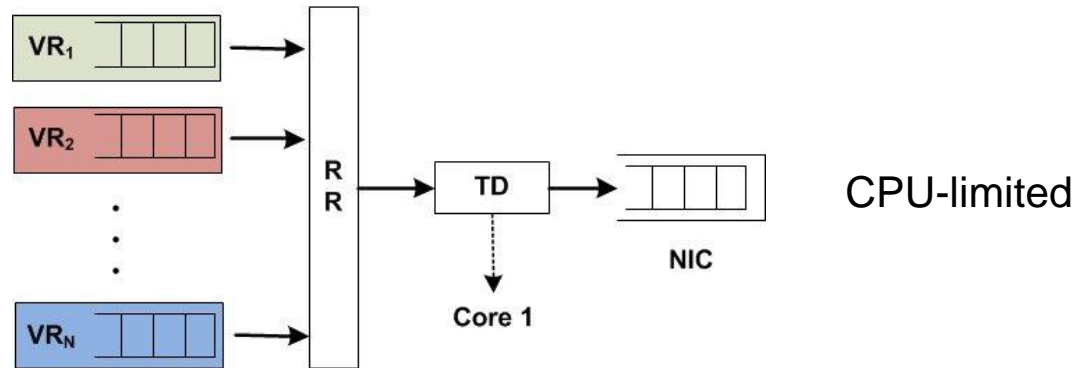


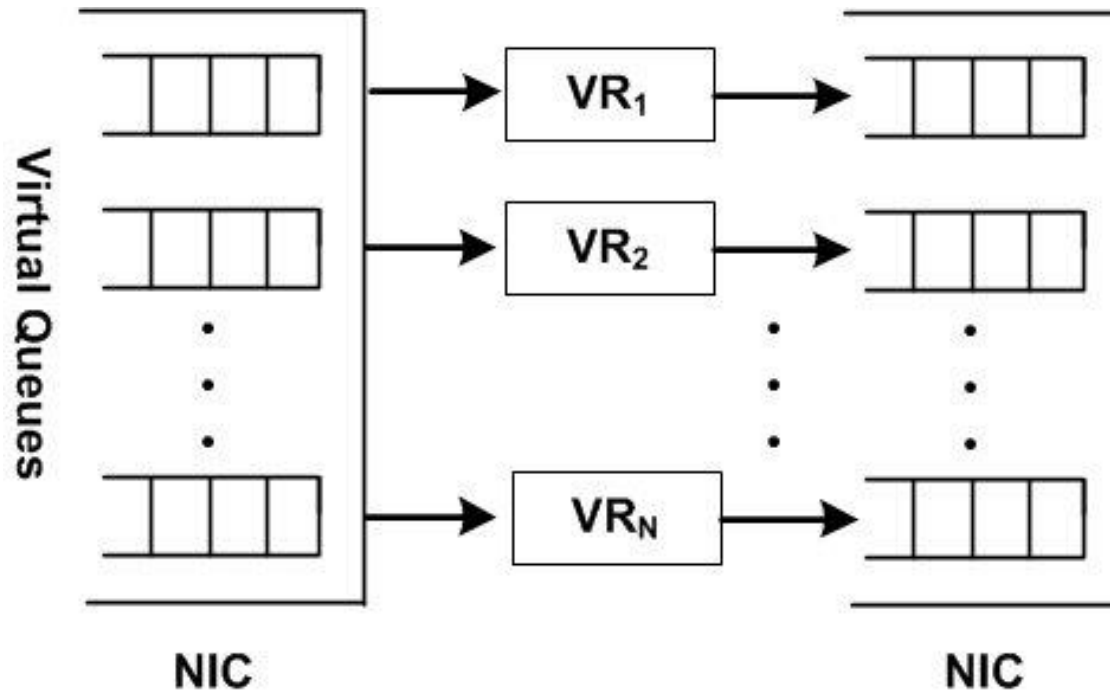


- ✓ Widely available (e.g., Click Modular Router, OpenvSwitch)
- ✓ Easy to enforce policies (e.g., weighted fair queuing)
- ✗ Resource utilization with multi-core CPUs



## ■ Use Case: Output processing



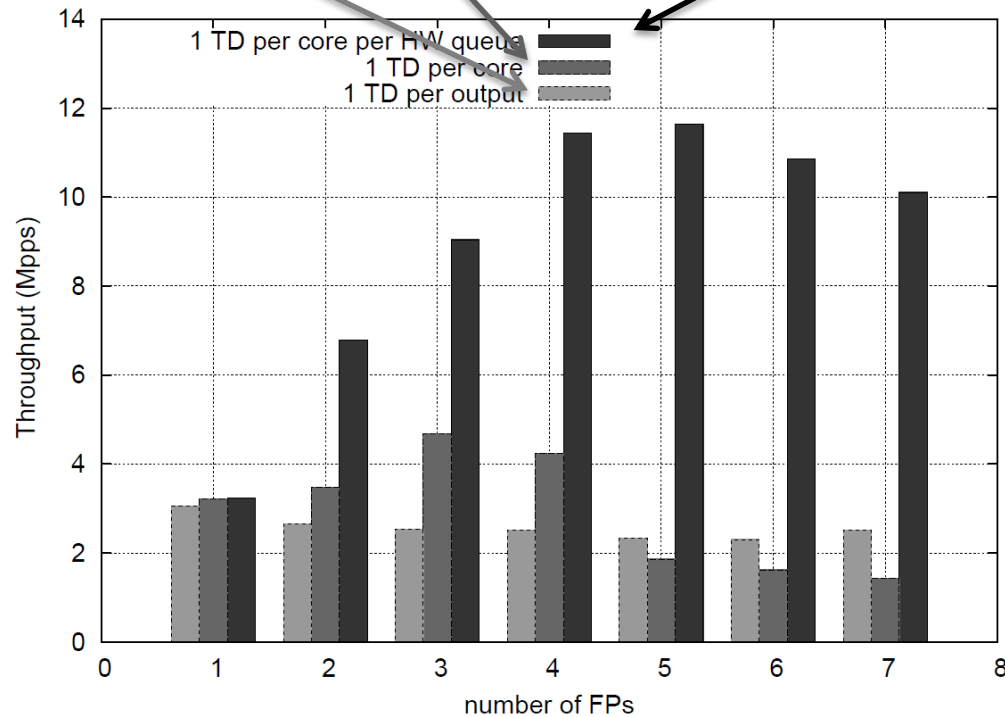
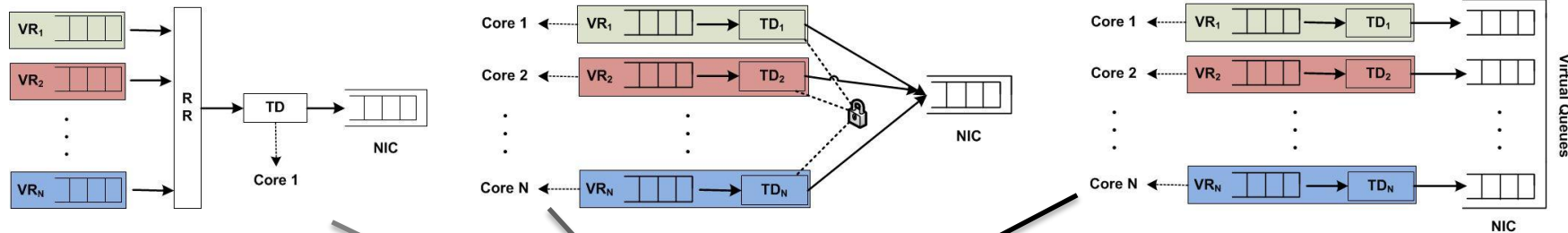


- ✓ Packet classification in hardware, off-loading the CPU
- ✗ Currently only round-robin scheduling is available



- Intel's VMDq:
  - Classification in multiple queues based on the destination MAC address
  - Possible packet reordering
  
- Microsoft's Receiver-side Scaling (RSS):
  - Classification in multiple queues using a hash function
  - Flow-based classification

# Performance with Output Processing

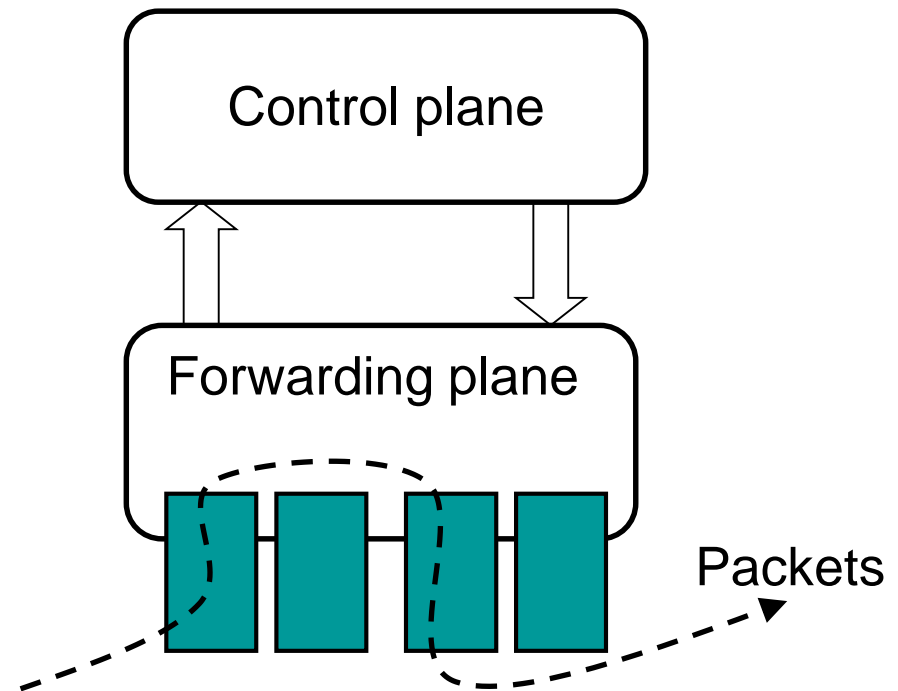




# Router Virtualization

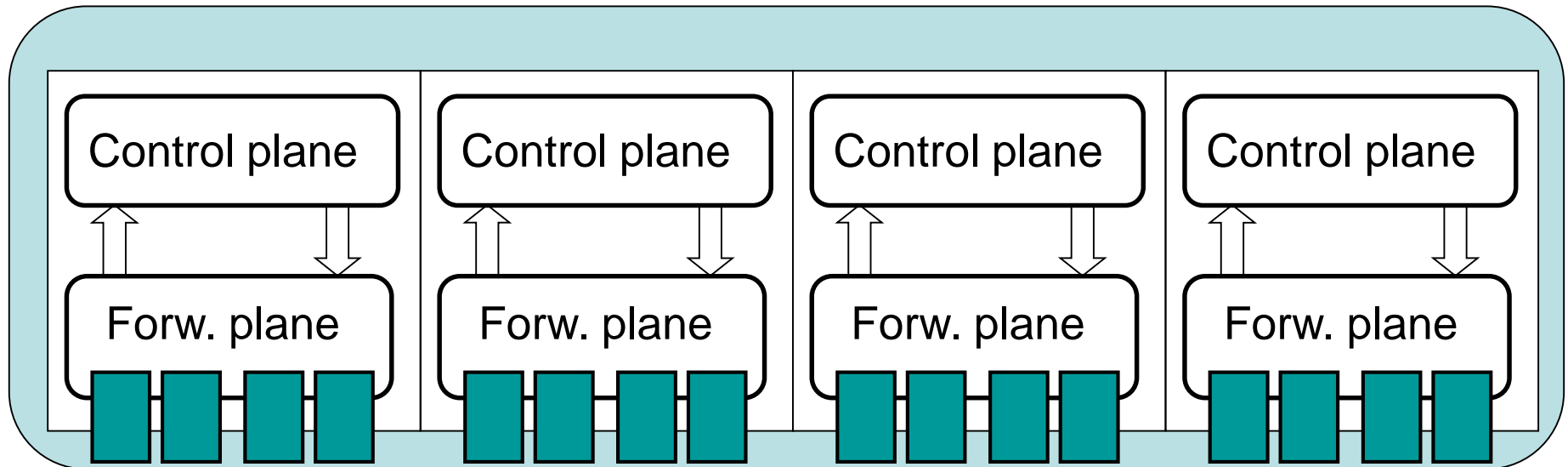


- Programmability:
  - Fully programmable forwarding plane (e.g., Click Modular Router)
  - Extensible control plane (e.g., XORP)
- High Performance:
  - Forwarding small packets at several Mpps
  - Limited by main memory accesses
  - Plenty of spare CPU cycles: much potential for virtualization





- One box fulfills the role of multiple routers
  - Dedicated forwarding and control planes for each virtual router
  - Server virtualization technologies can be used to virtualise and isolate software routers



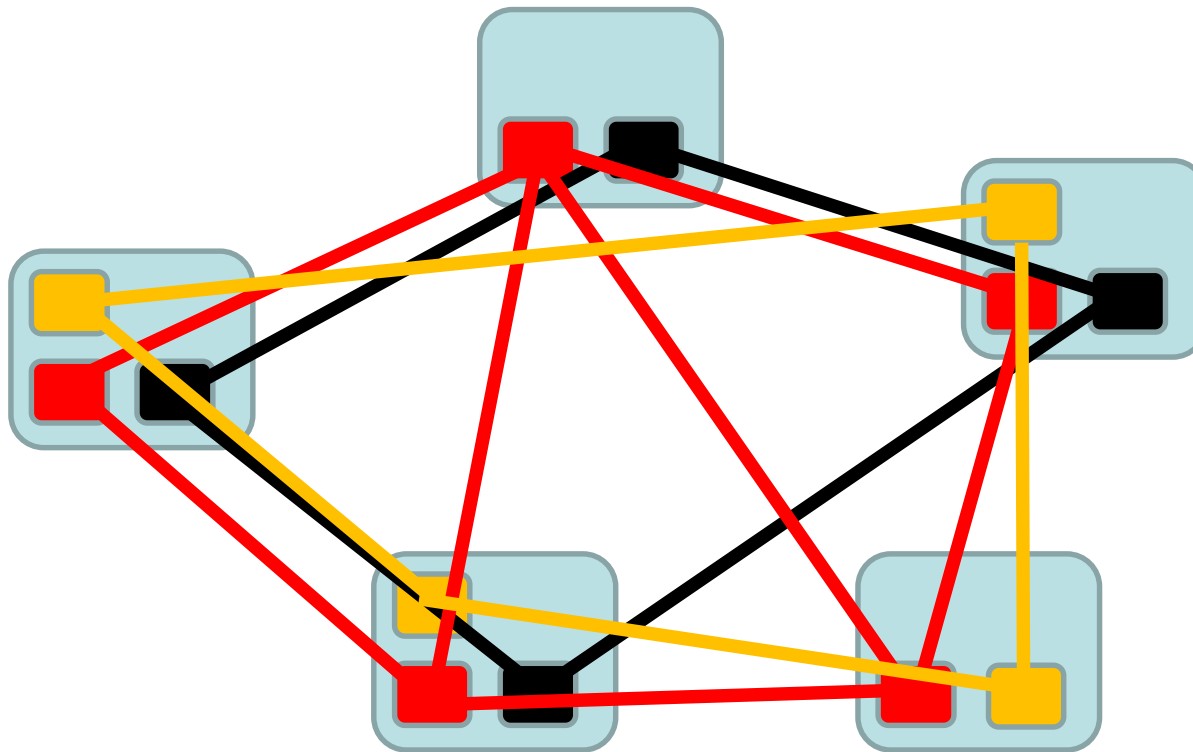




- One box fulfils the role of multiple routers
  - Independent and flexible management, leasing the Internet
  - Resource sharing (many available CPU cycles)
  - Lower hardware and support cost
    - e.g., Small businesses within the same building sharing the same router, each managing its own VR
- Excellent platform for experimentation
  - Rolling out new and unstable solutions without risk

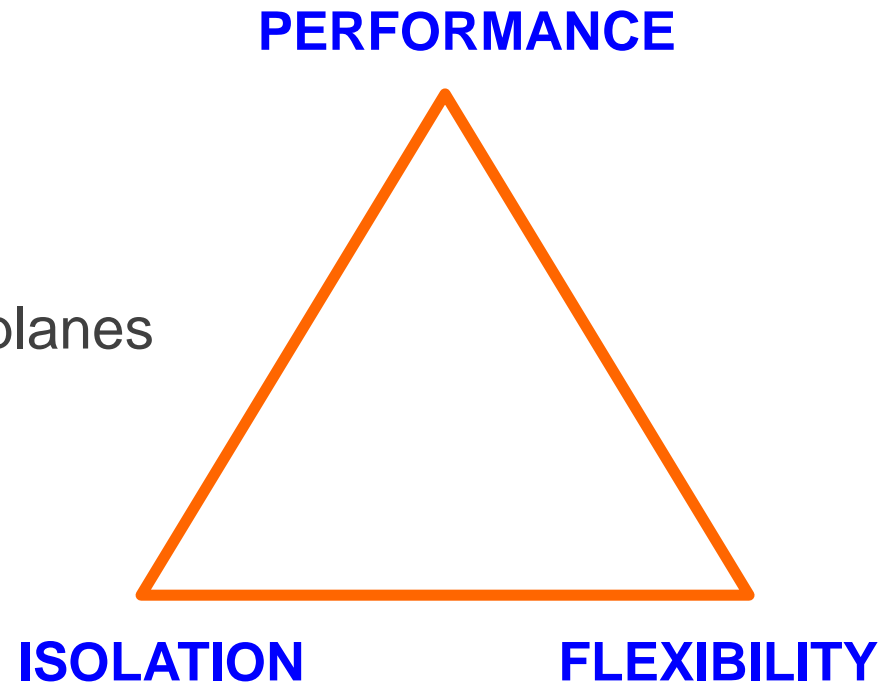


- Innovation in network architecture design
  - Each virtual network has dedicated virtual routers with customized forwarding planes and routing protocols



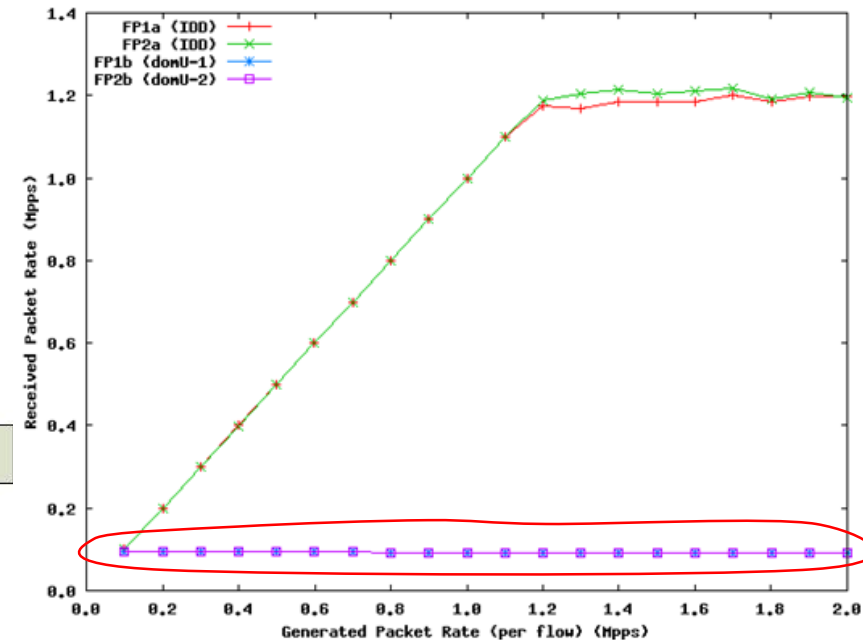
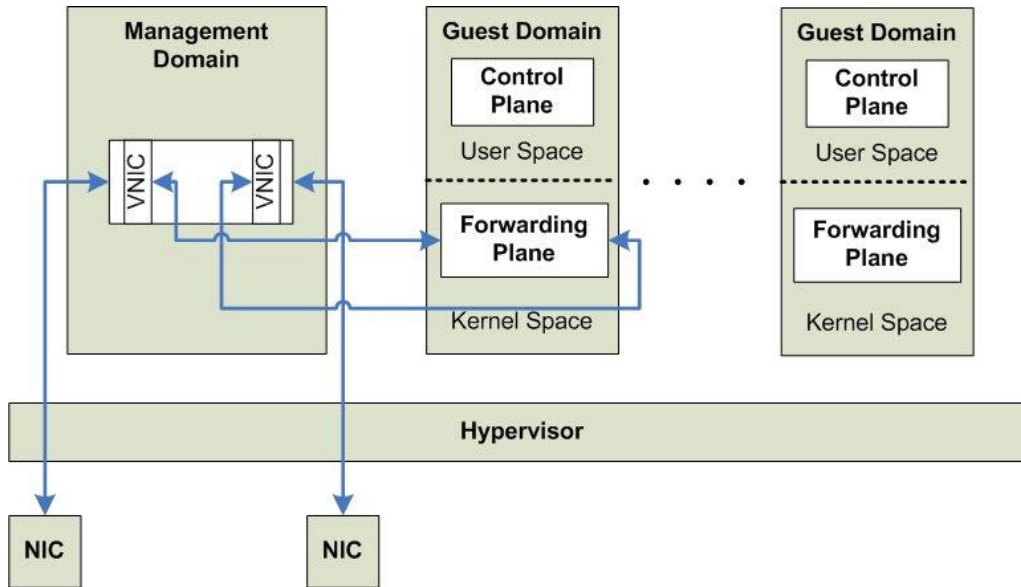


- Isolating the virtual routers from each other
- Minimizing the overhead of virtualization
- Highly configurable forwarding planes

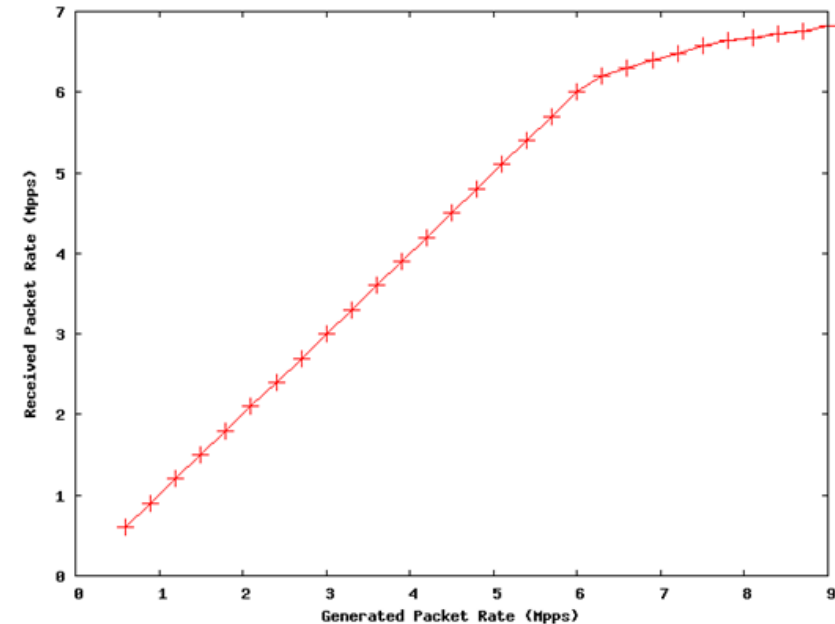
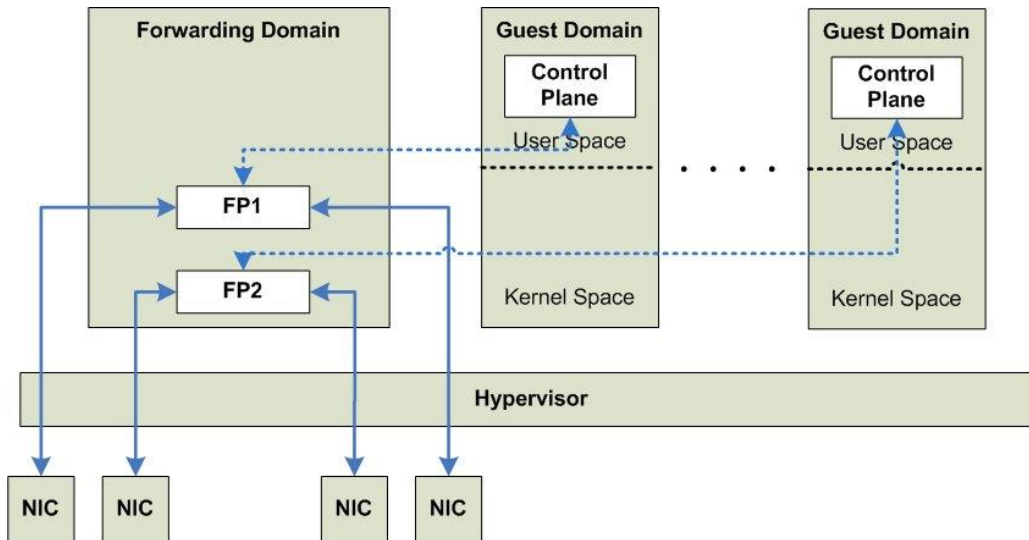




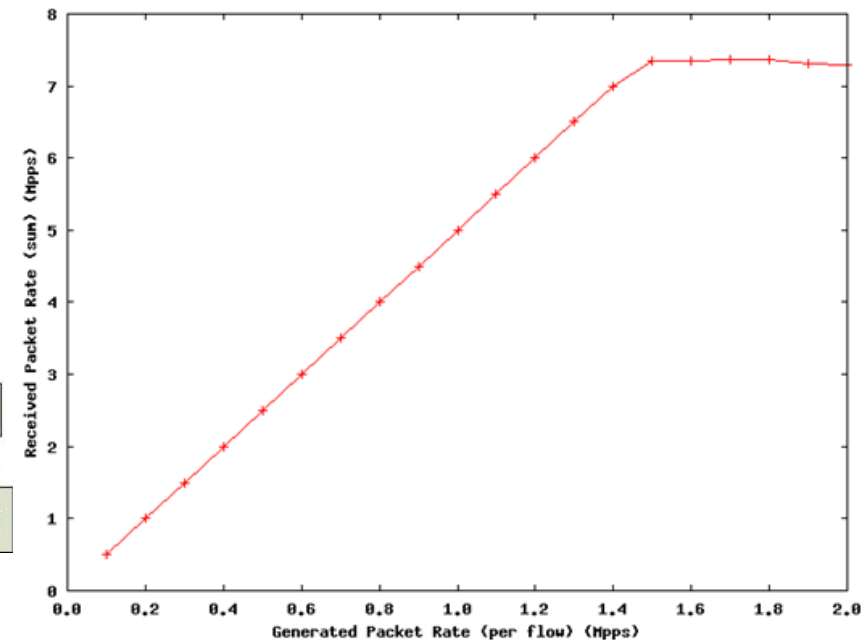
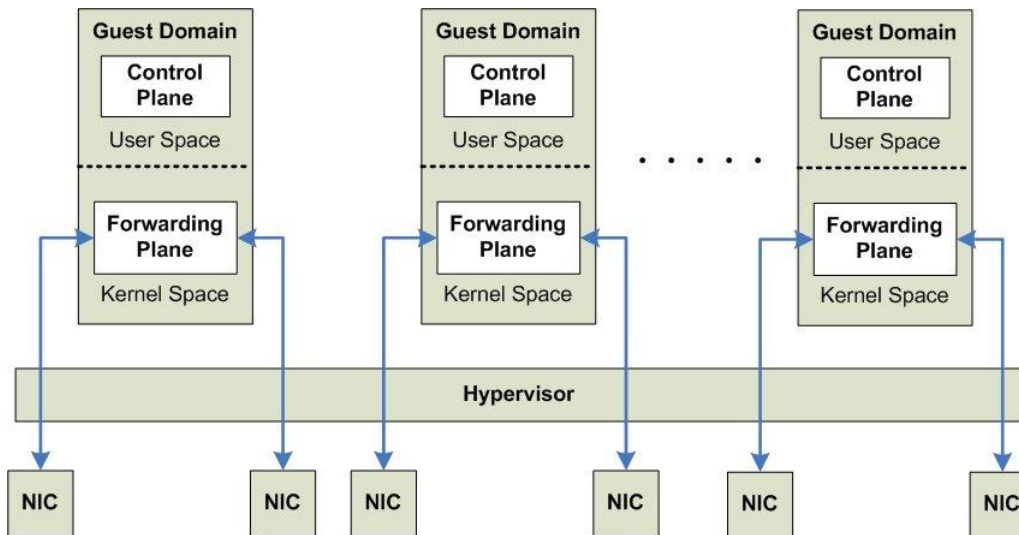
- Hosting forwarding planes
  - Determine the most suitable virtual forwarding configuration for high performance, flexibility and isolation
- Virtual forwarding scenarios:
  - Forwarding using I/O channels
  - Common forwarding plane
  - Forwarding with direct-mapped network interfaces



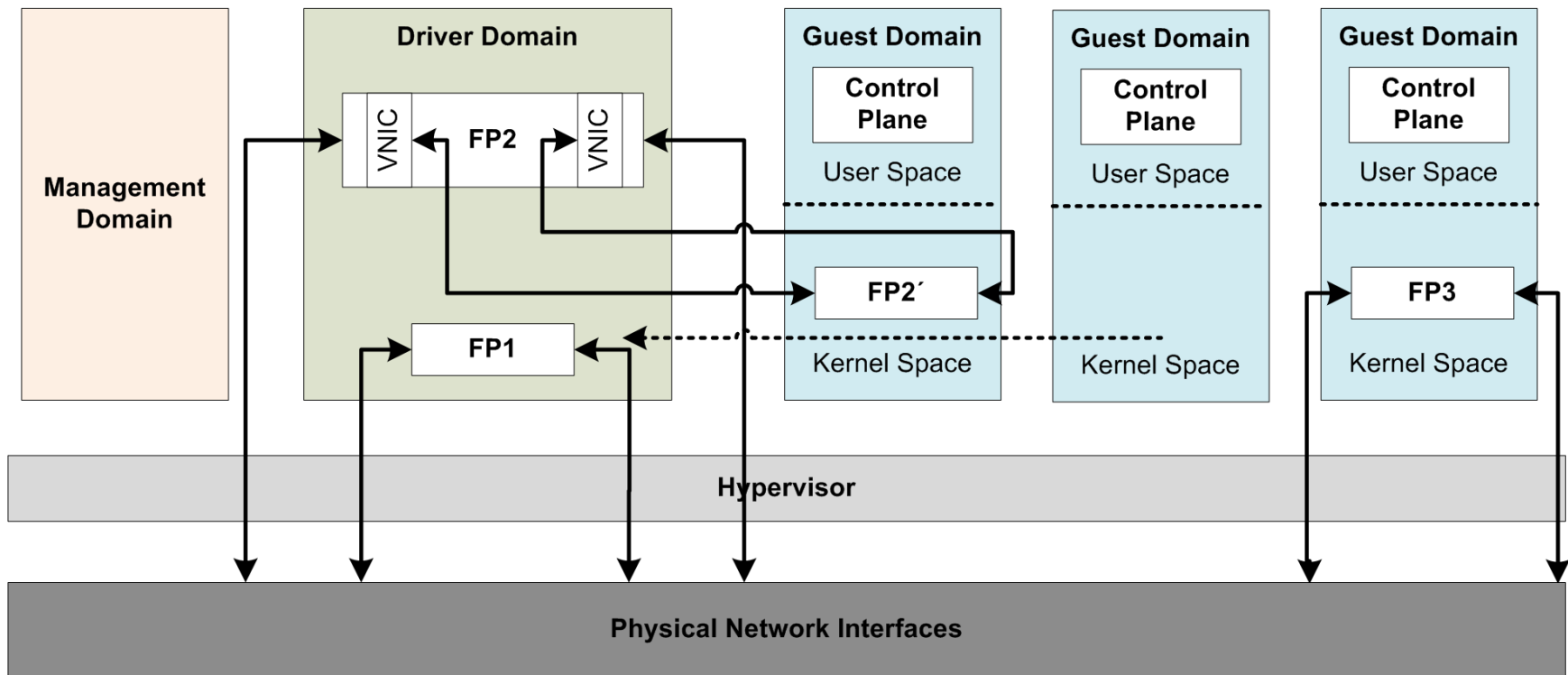
- ✓ Isolation
- ✗ Low performance (due to costly hypervisor domain switching)



- ✓ High performance
- ✗ Isolation (it should be enforced with proper core allocation and scheduling)



- ✓ High performance
- ✓ Isolation
- ✗ Requires NIC multi-queuing to scale



- Forwarding within a common domain (FP1)
- Forwarding with I/O Channels (FP2 ↔ FP2')
- Forwarding in Guest Domains with direct NIC mapping (FP3)

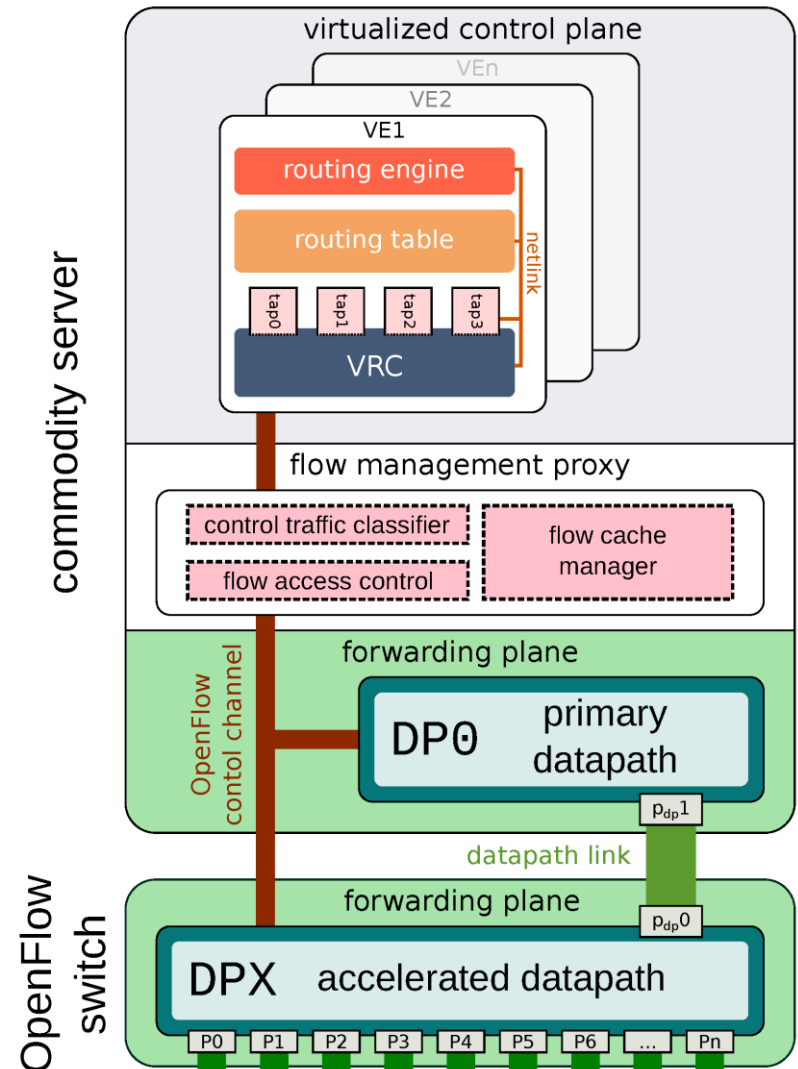




# Accelerated Software Virtual Routers

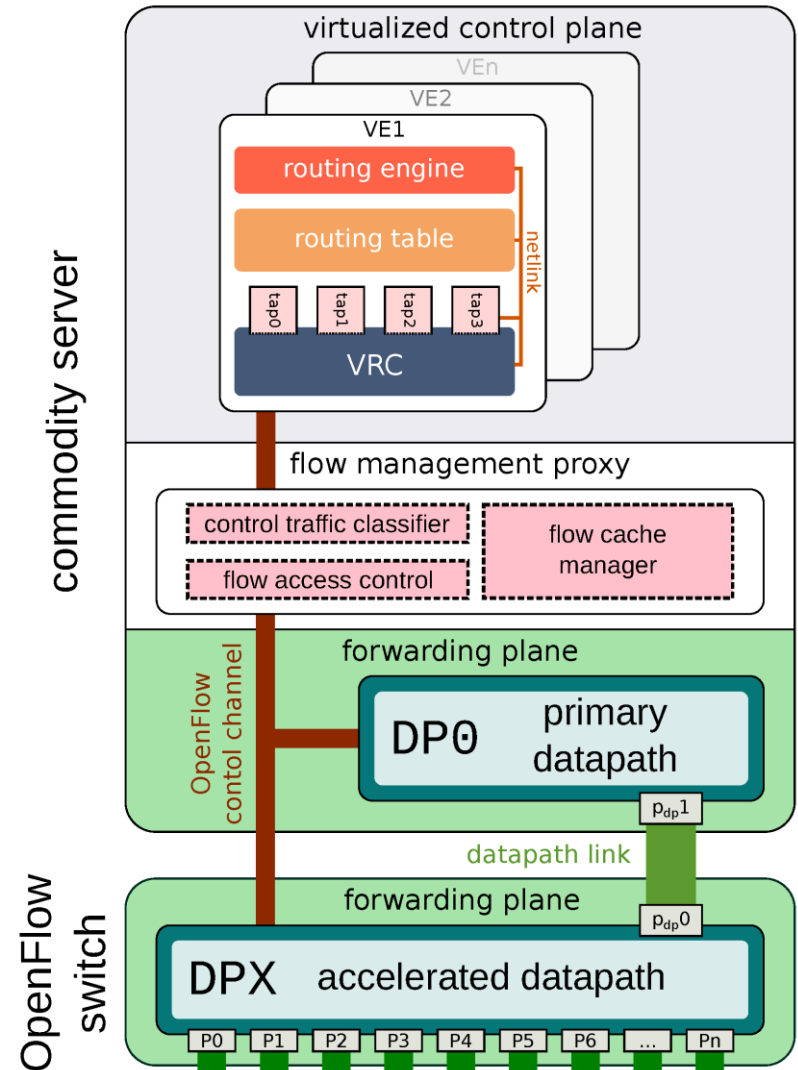


- Features:
  - Control-plane extensibility
  - Forwarding-plane programmability
  - High performance
  - High port density
- Main components:
  - Forwarding plane composed of primary and accelerated datapath
  - Virtualized control plane
  - Flow management proxy:
    - Transparent layer between the forwarding and control plane



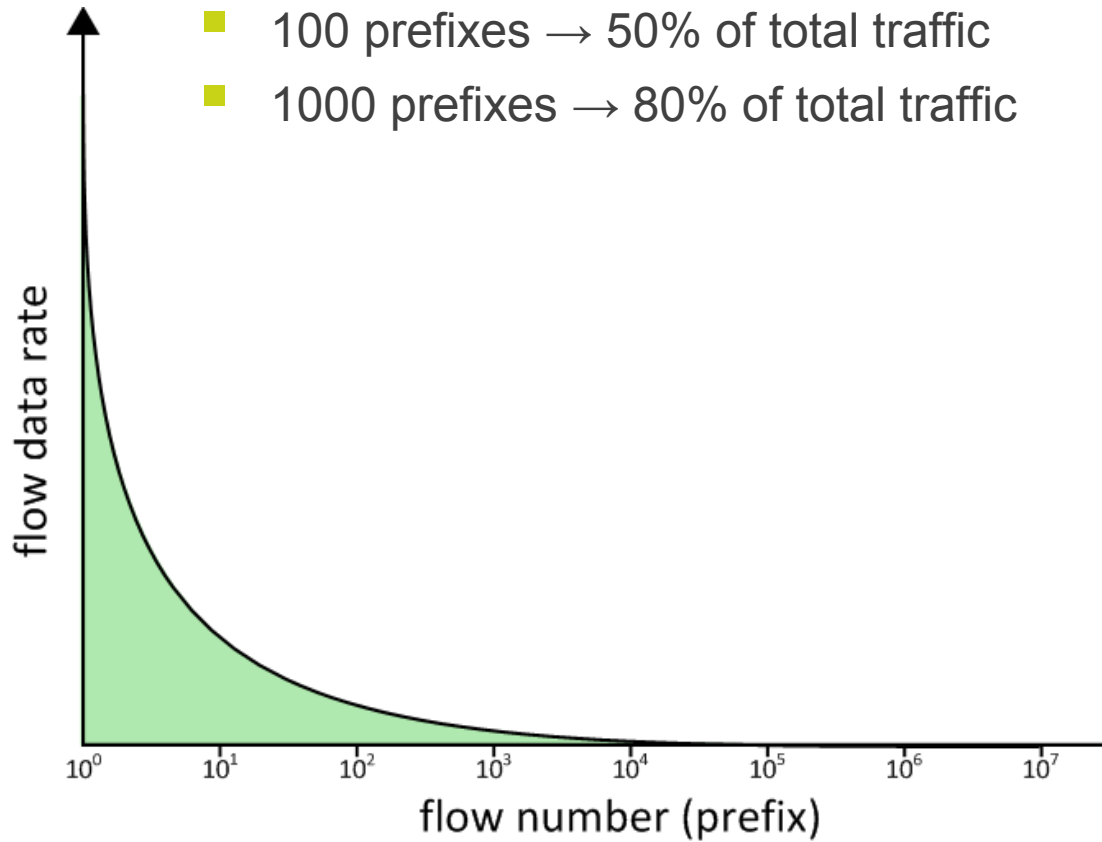


- Main limitation:
  - OpenFlow switch flow table has very small size (few Mbytes)
    - can store only a few thousands of flow entries
- This limitation seems to make such a platform infeasible:
  - A full BGP routing table includes nearly 400K entries



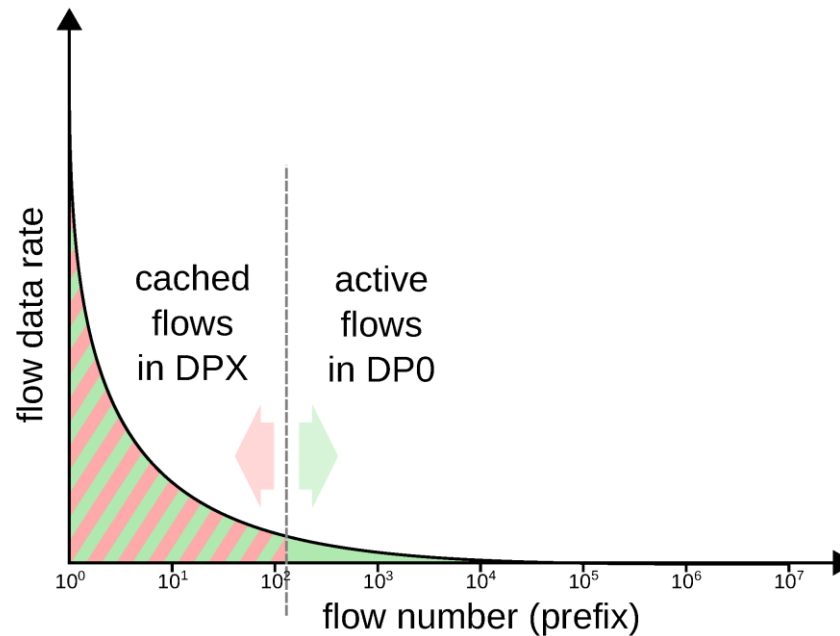


- Flow distribution in the Internet:
  - A small subset of flows carries most of Internet traffic
    - Traffic statistics from an access router at a large European ISP
      - 100 prefixes → 50% of total traffic
      - 1000 prefixes → 80% of total traffic



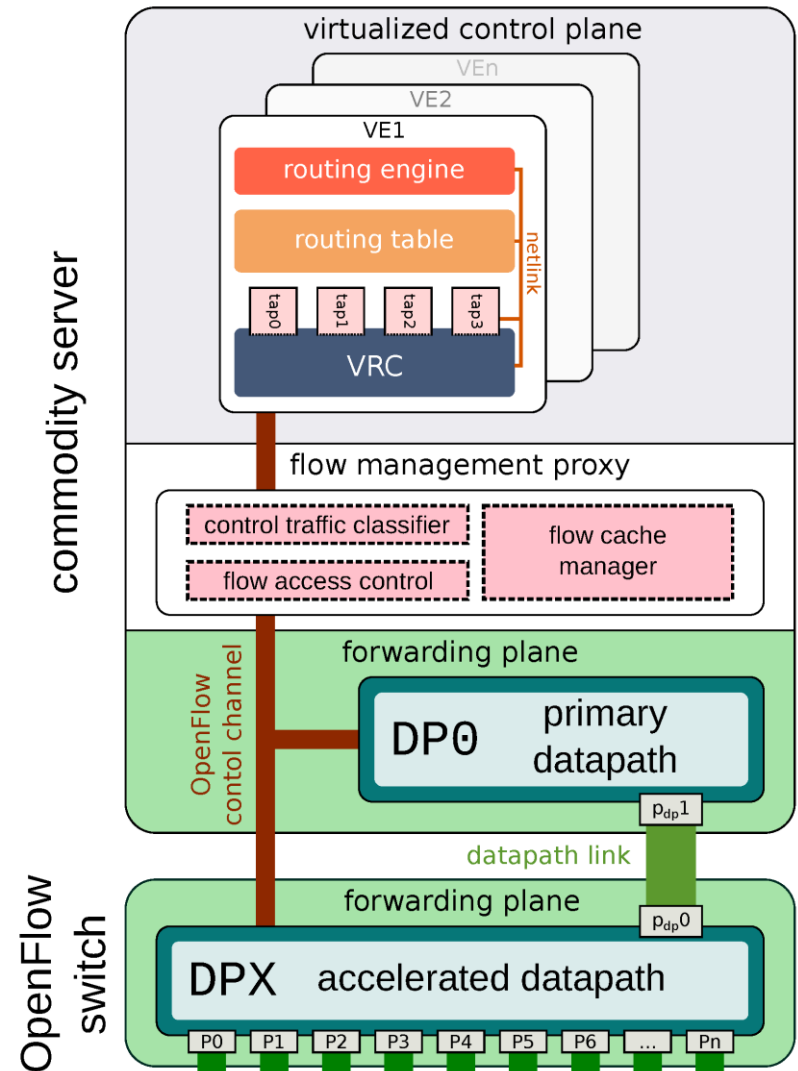


- Leverage on Internet flow distribution:
  - Dual-datapath approach:
    - Primary datapath (DP0) in a commodity server with forwarding entries for all flows
    - Accelerated datapath (DPX) in OpenFlow switch with forwarding entries for the subset of large-volume flows



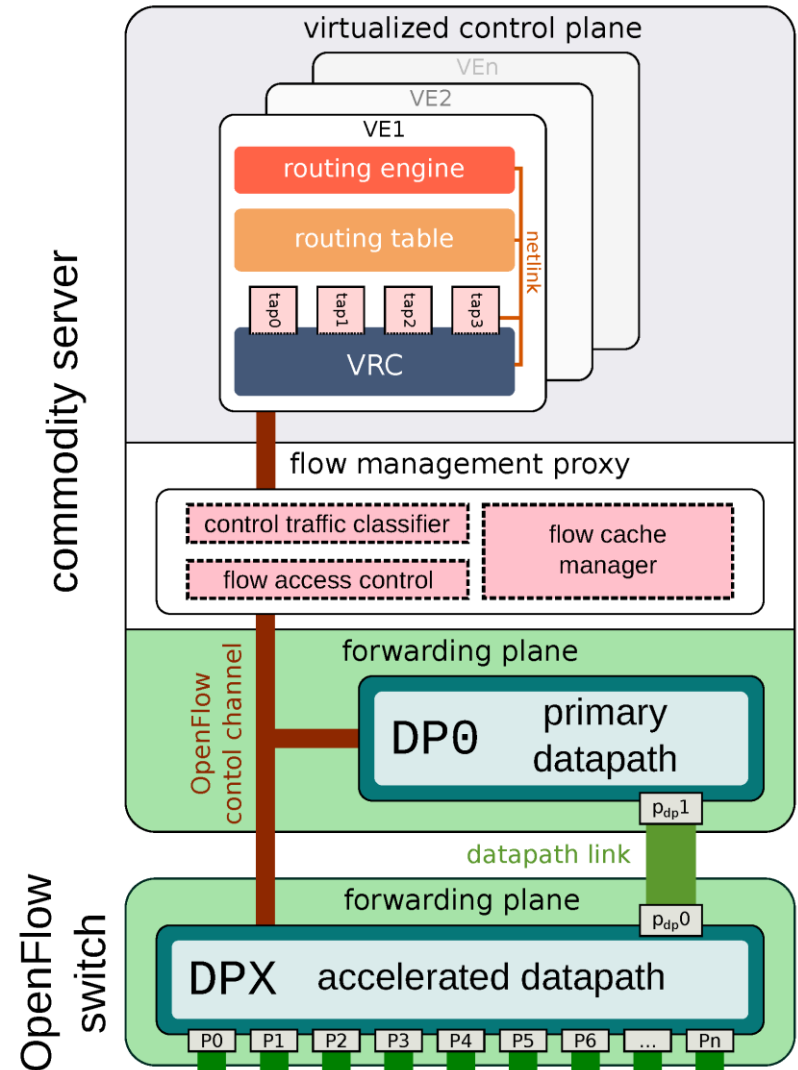


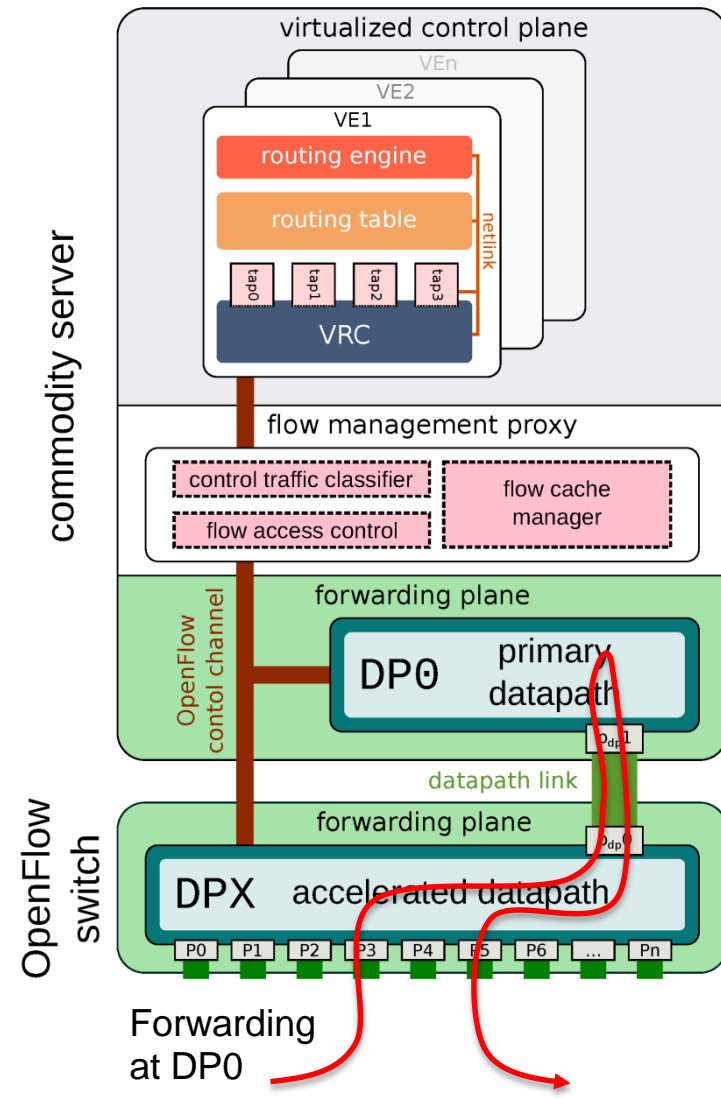
- Virtual interfaces mapped to the physical ports:
  - Mapping registered with FMP
- Virtual Router Controller (VRC):
  - Forwards control messages to virtual interfaces
  - Generates flow insertion/ deletion commands
    - Listens on Netlink socket for routing table or ARP updates





- Control message classification among virtualized control planes
- Access control on flow insertion/deletion commands generated by a VRC
- Flow caching in the DPX
  - Least recently used (LRU)
  - Least frequently used (LFU)

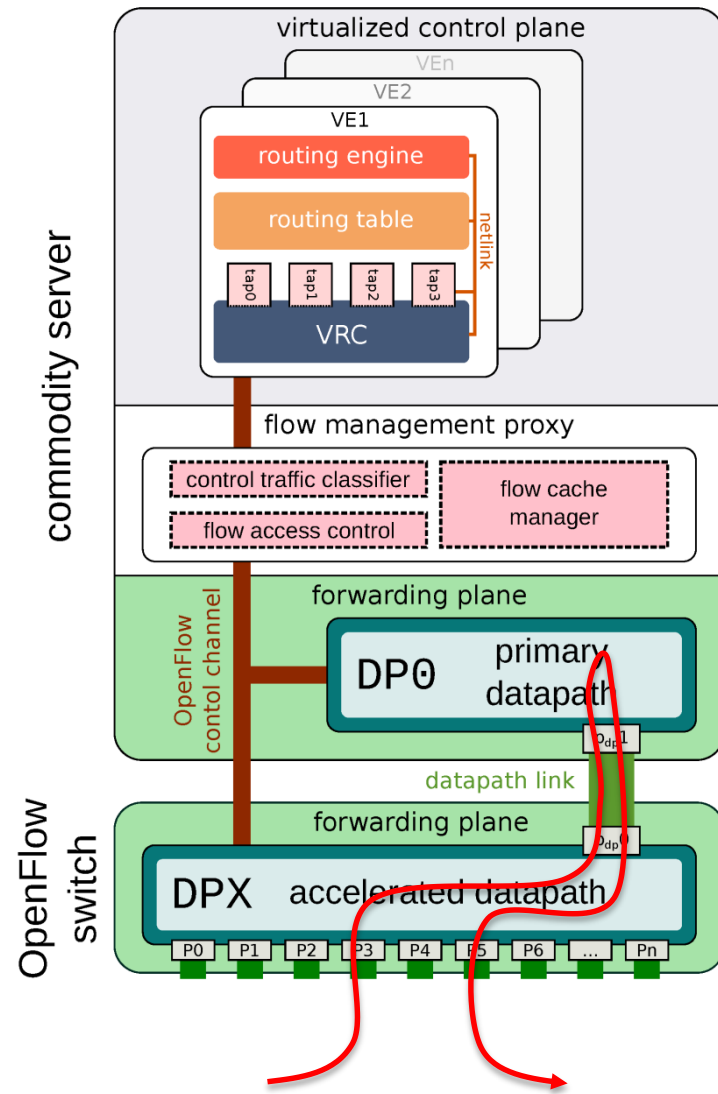




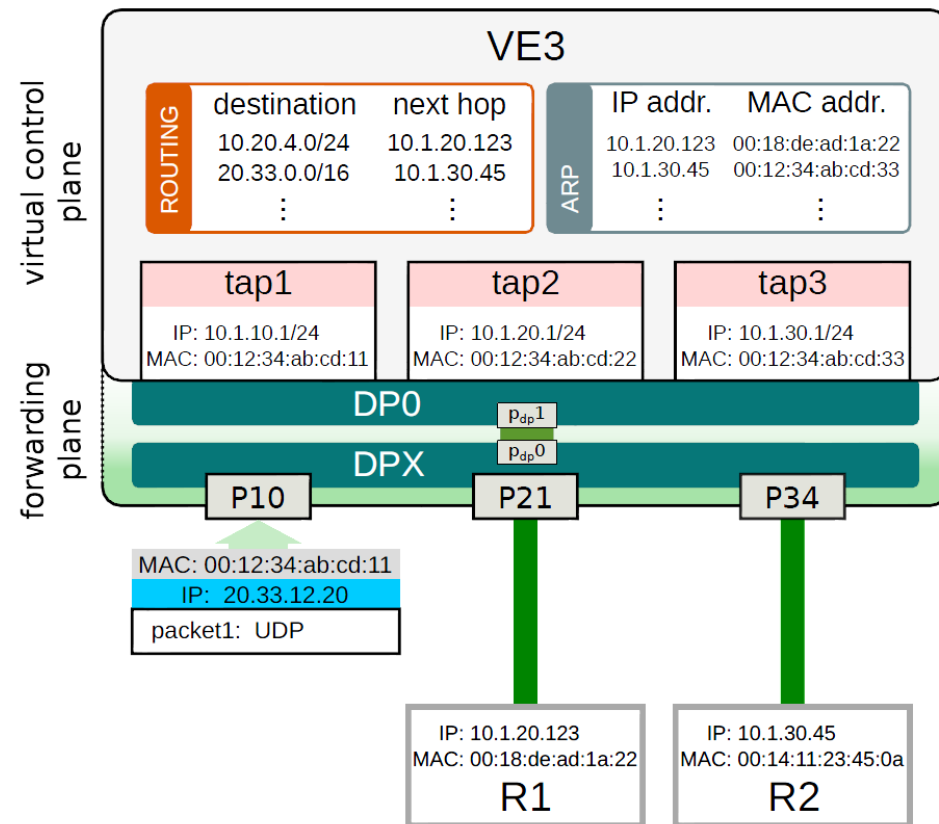




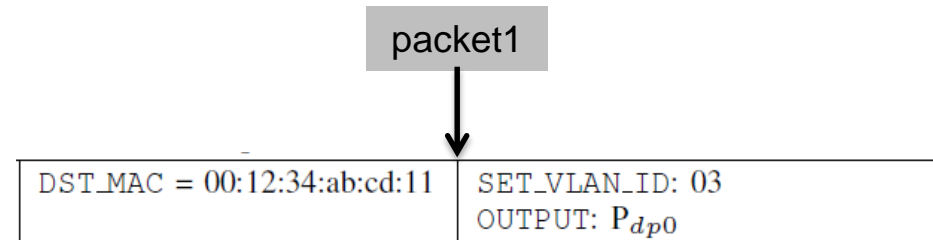
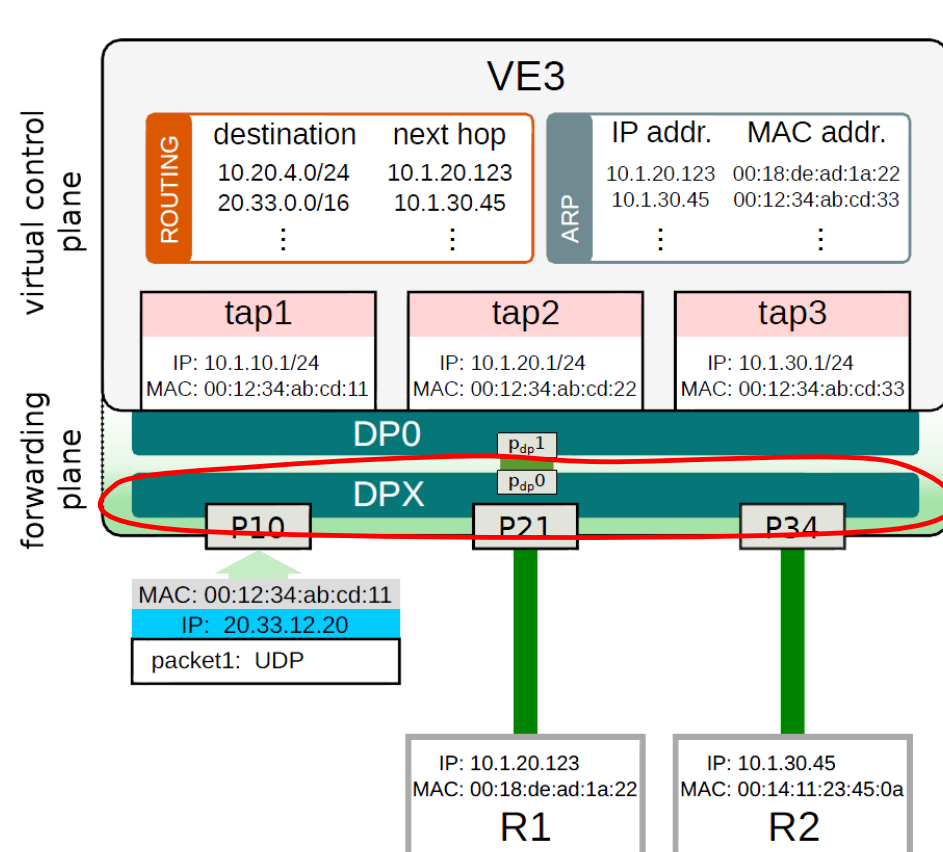
- Input redirection flow entries
- Output flow entries
- Cached routing flow entries
- Local control flow entries:
  - Routing updates
  - ARP resolution



# Exemplary Packet Forwarding at DP0

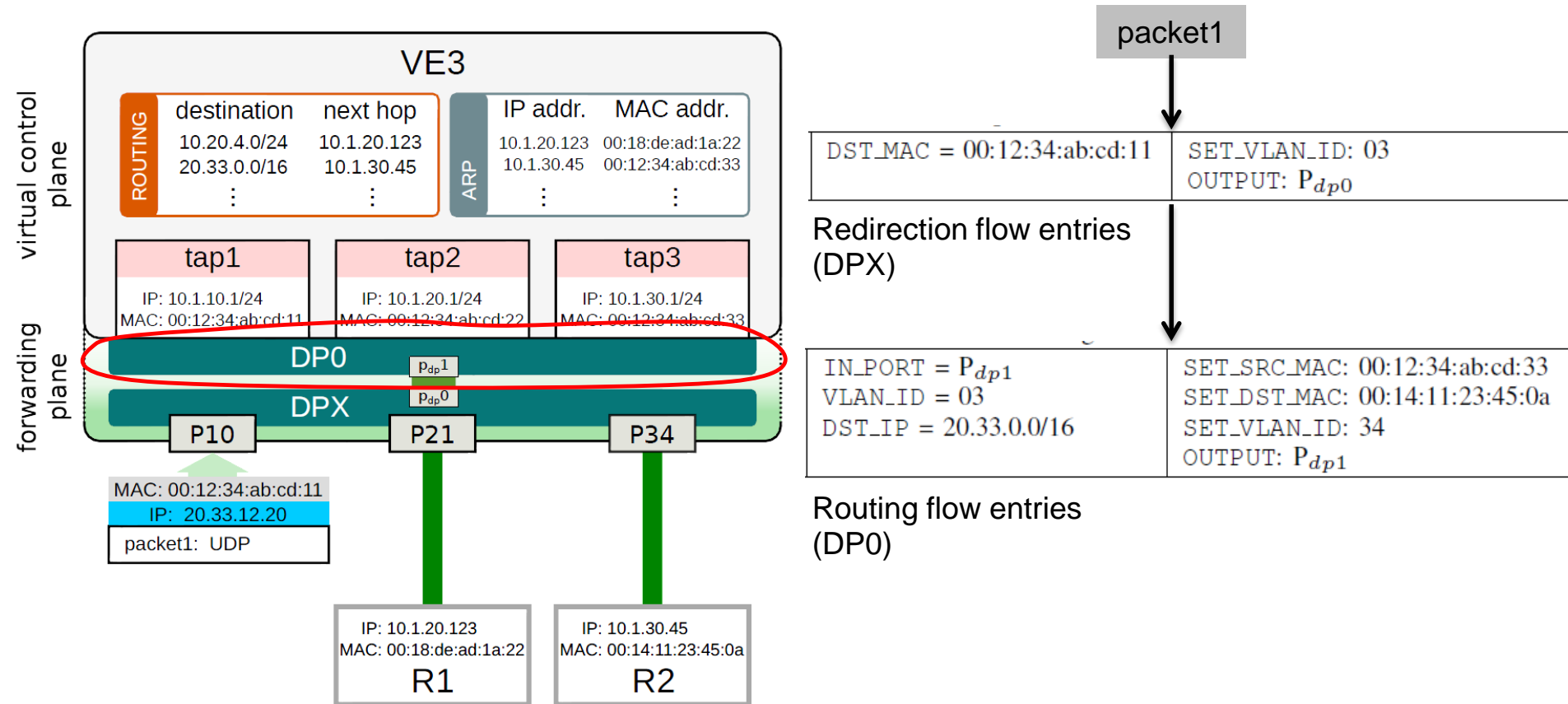


# Exemplary Packet Forwarding at DP0

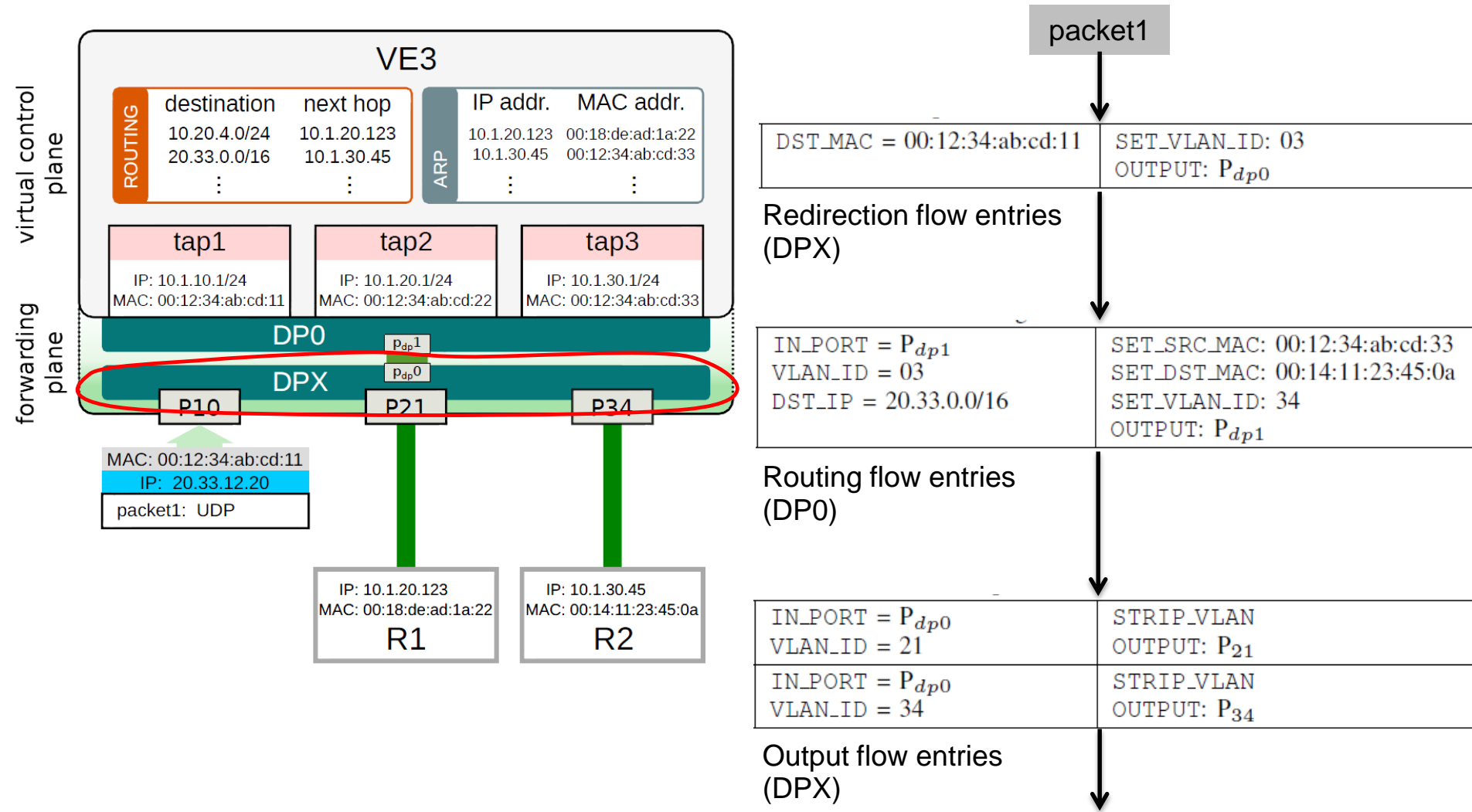


Redirection flow entries  
(DPX)

# Exemplary Packet Forwarding at DP0



# Exemplary Packet Forwarding at DP0





- One input redirection entry per virtual port ( $\sum_{i=1}^N M_i$  entries)
- One local control entry per virtual port ( $\sum_{i=1}^N M_i$  entries)
- One output entry per physical port ( $P$  entries)

$$\left. \begin{array}{l} \text{One input redirection entry per virtual port } (\sum_{i=1}^N M_i \text{ entries}) \\ \text{One local control entry per virtual port } (\sum_{i=1}^N M_i \text{ entries}) \\ \text{One output entry per physical port } (P \text{ entries}) \end{array} \right\} 2 \sum_{i=1}^N M_i + P$$

↓

Hundreds of entries

$N$ : number of virtual routers

$M_i$ : number of virtual ports for  $i^{\text{th}}$  virtual router

$P$ : number of physical ports



- DP0 (Pronto Switch)
  - Forwarding at line rate
  - Latency: 4  $\mu$ s
  - Flow insertion rate: ~1000 flows/sec
  
- DPX (Server)
  - Forwarding rate: ~1Mpps per core
  - Latency: 15  $\mu$ s
  - Flow insertion rate: ~6000 flows/sec

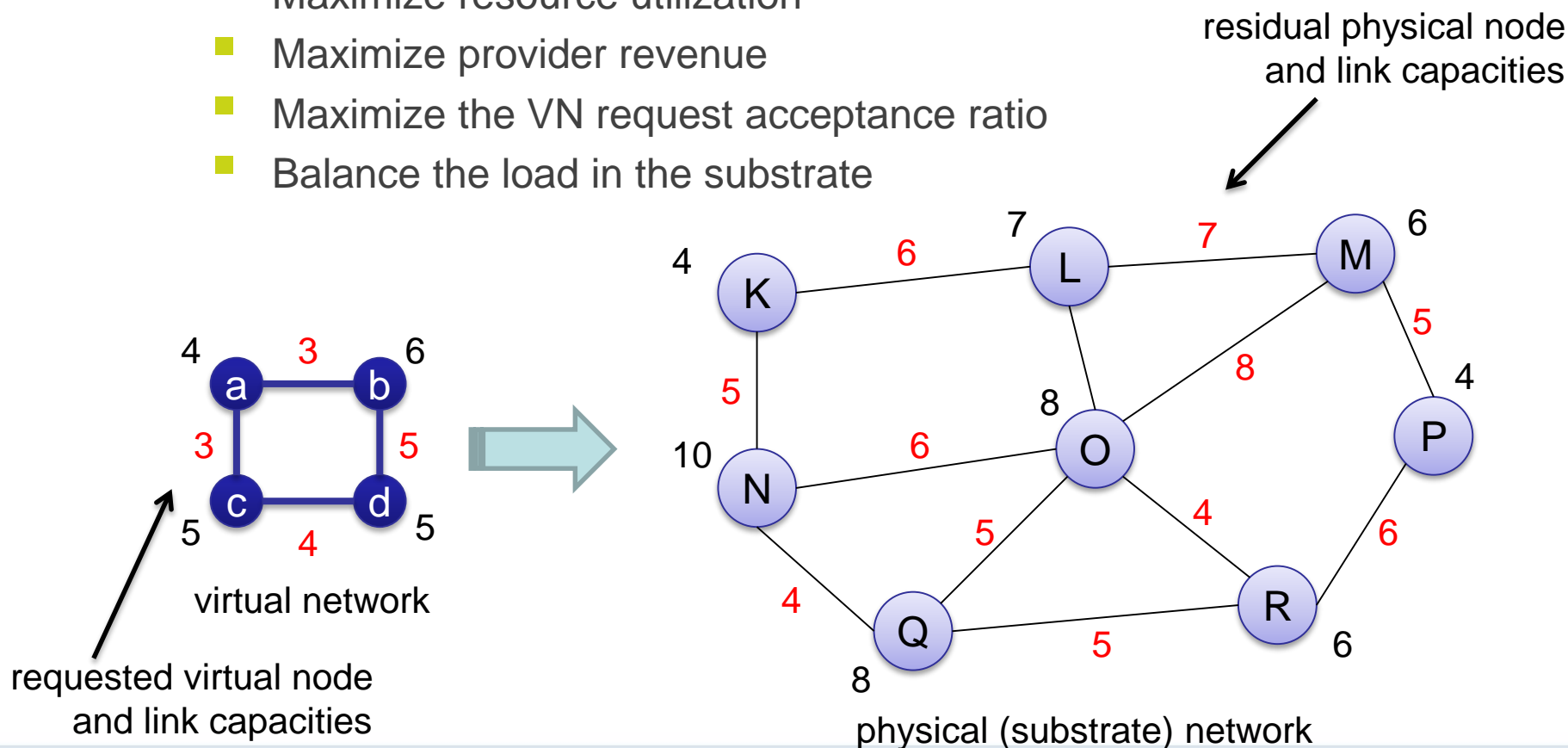


# Virtual Network Embedding



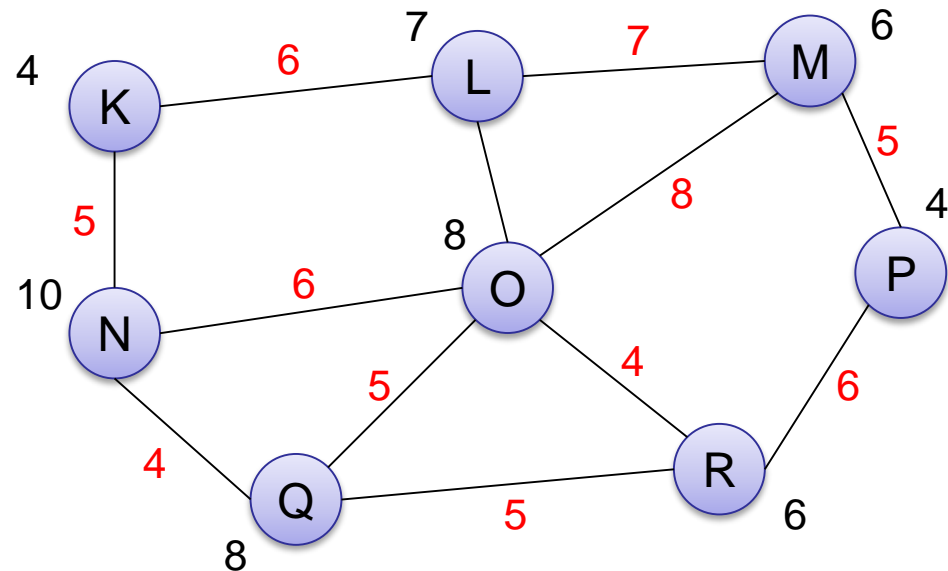
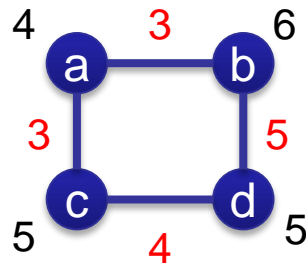


- Virtual network embedding:
  - Compute a mapping of a virtual network onto a physical network given an objective, e.g.,:
    - Maximize resource utilization
    - Maximize provider revenue
    - Maximize the VN request acceptance ratio
    - Balance the load in the substrate



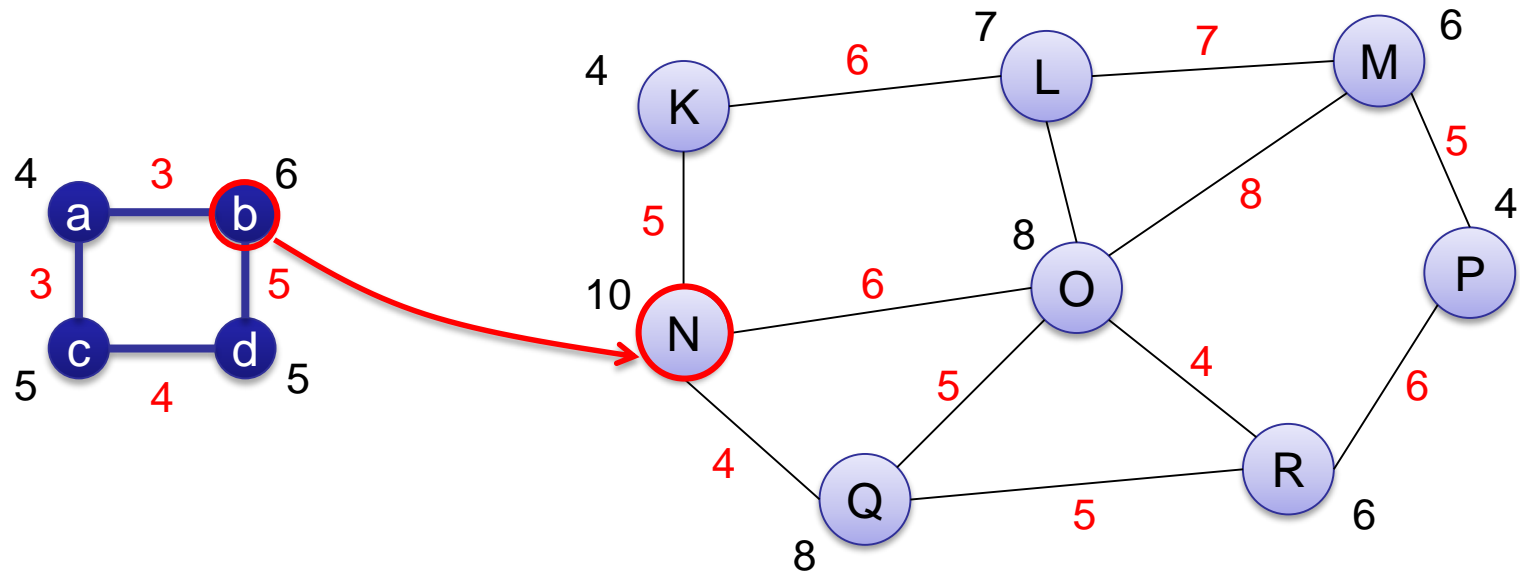


- VN embedding algorithm with simultaneous virtual node and link assignment
  - Objective: Load balancing across the substrate



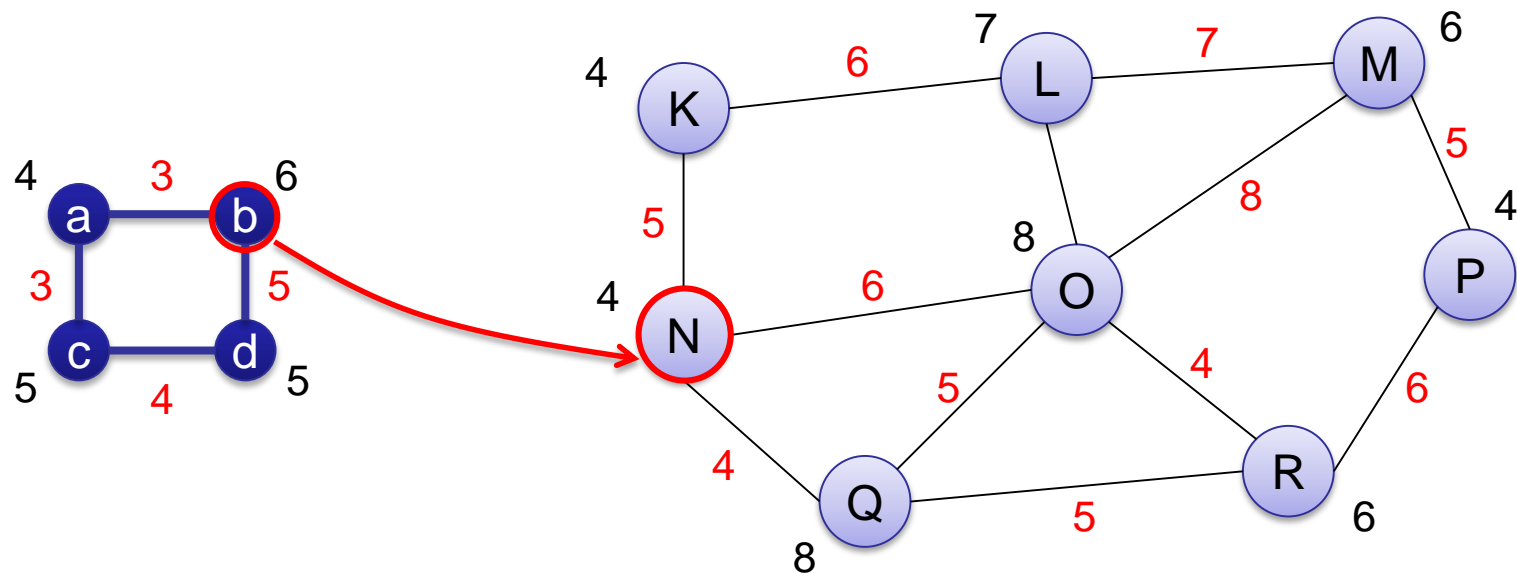


- VN embedding algorithm with simultaneous virtual node and link assignment
  - Objective: Load balancing across the substrate



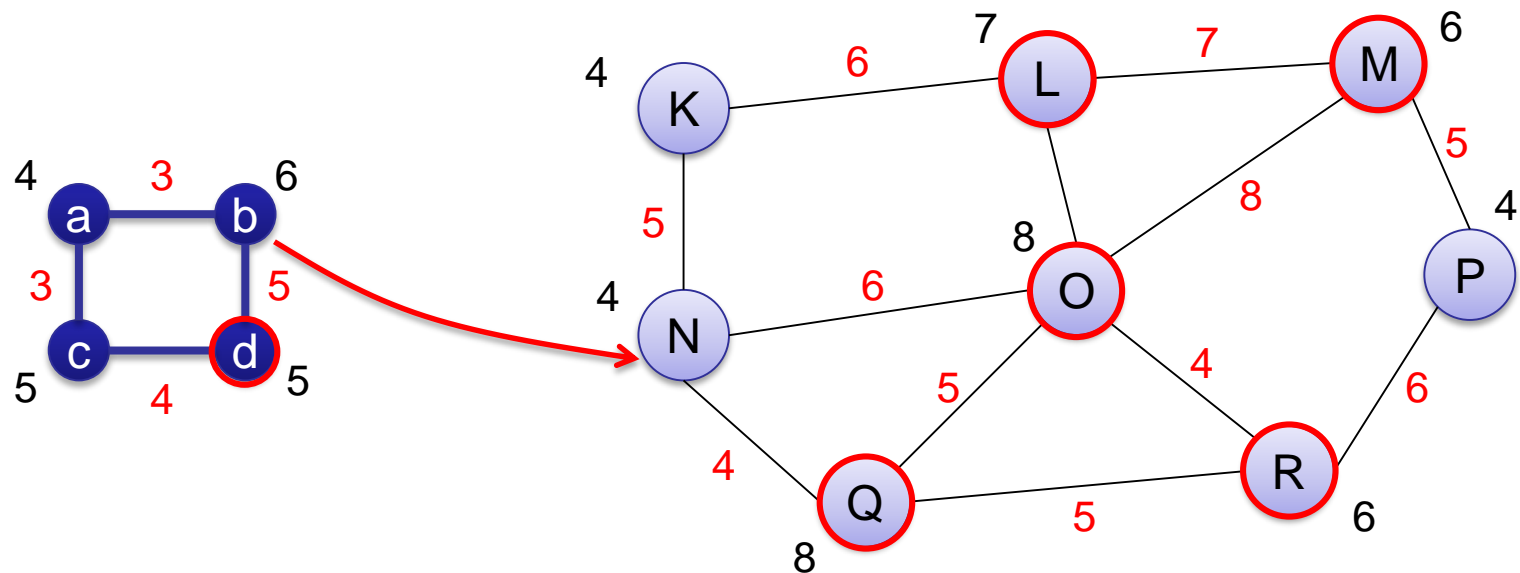


- VN embedding algorithm with simultaneous virtual node and link assignment
  - Objective: Load balancing across the substrate



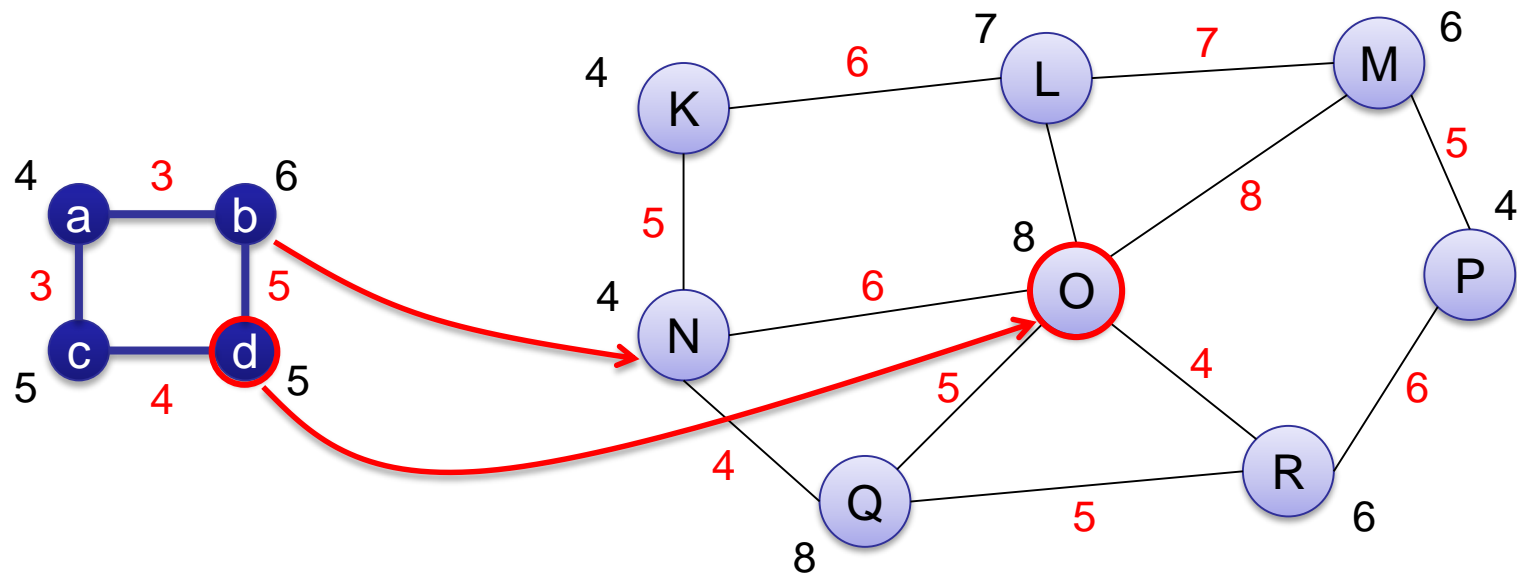


- VN embedding algorithm with simultaneous virtual node and link assignment
  - Objective: Load balancing across the substrate



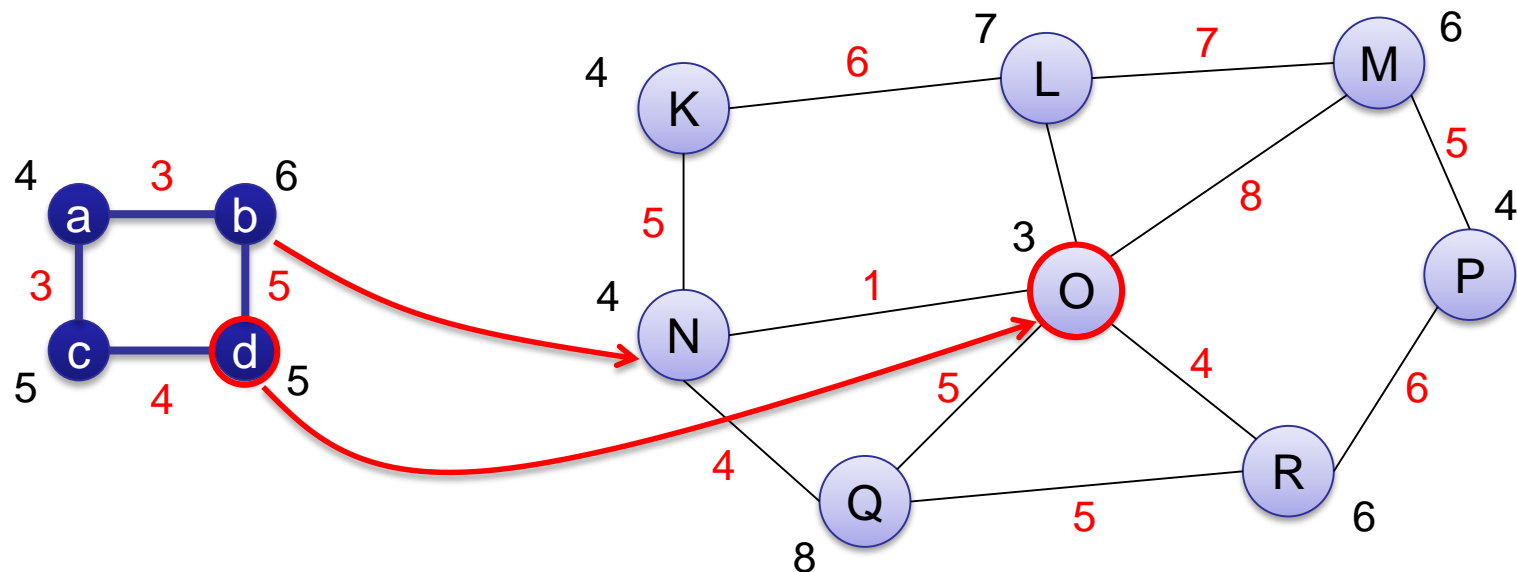


- VN embedding algorithm with simultaneous virtual node and link assignment
  - Objective: Load balancing across the substrate



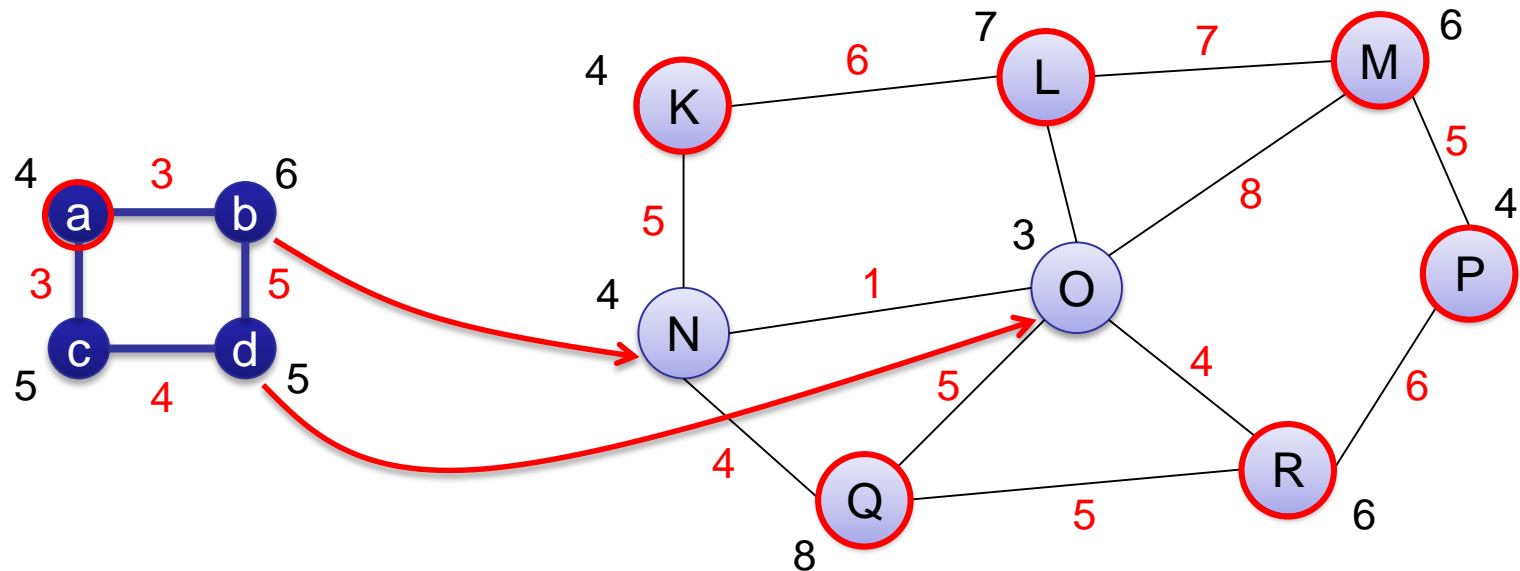


- VN embedding algorithm with simultaneous virtual node and link assignment
  - Objective: Load balancing across the substrate





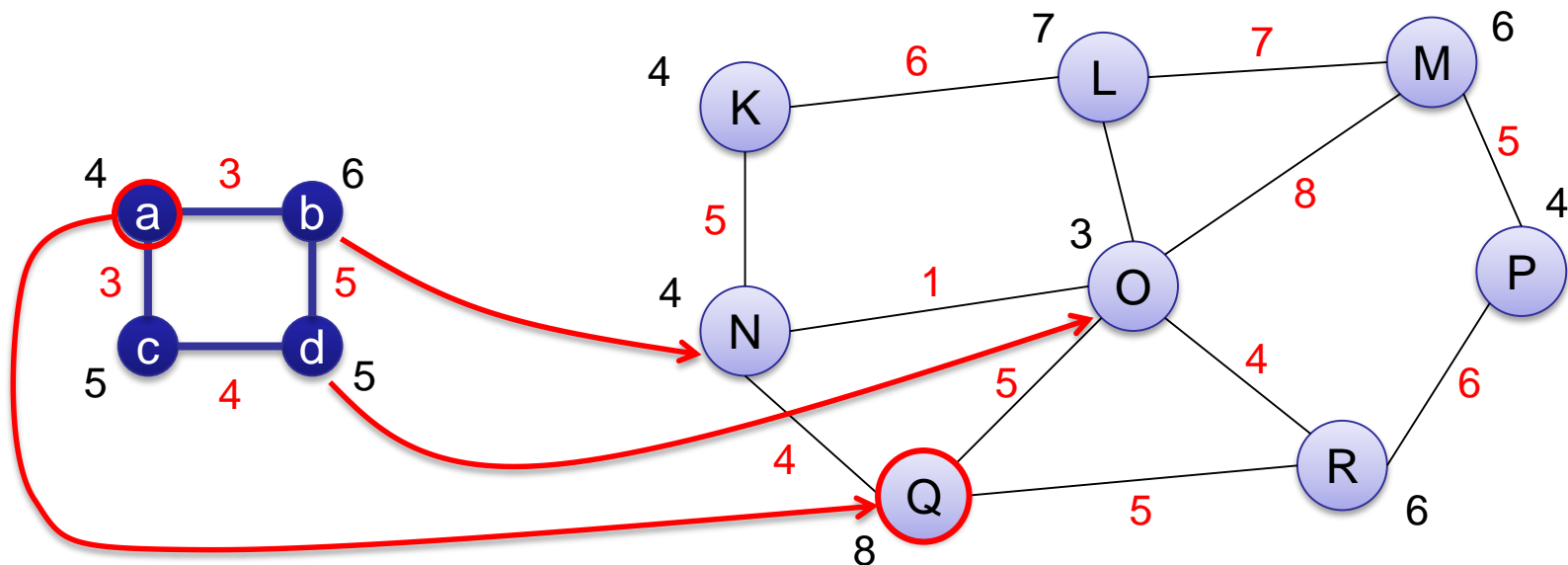
- VN embedding algorithm with simultaneous virtual node and link assignment
  - Objective: Load balancing across the substrate





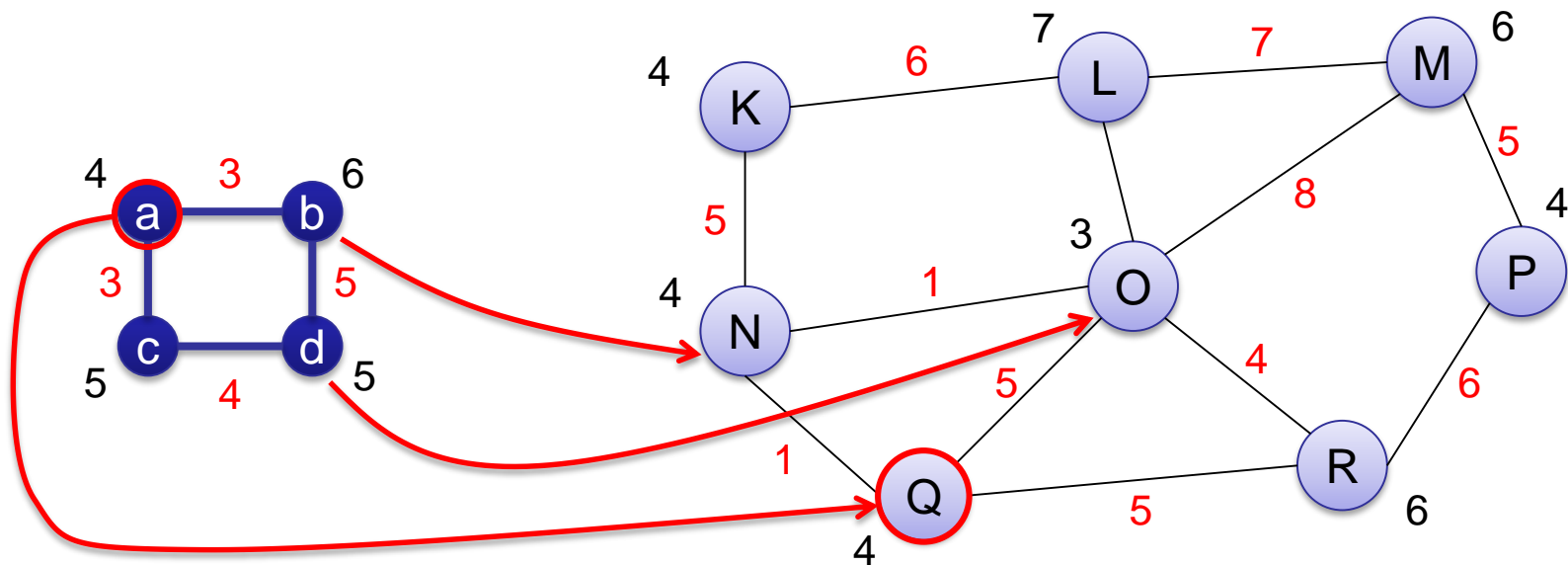


- VN embedding algorithm with simultaneous virtual node and link assignment
  - Objective: Load balancing across the substrate



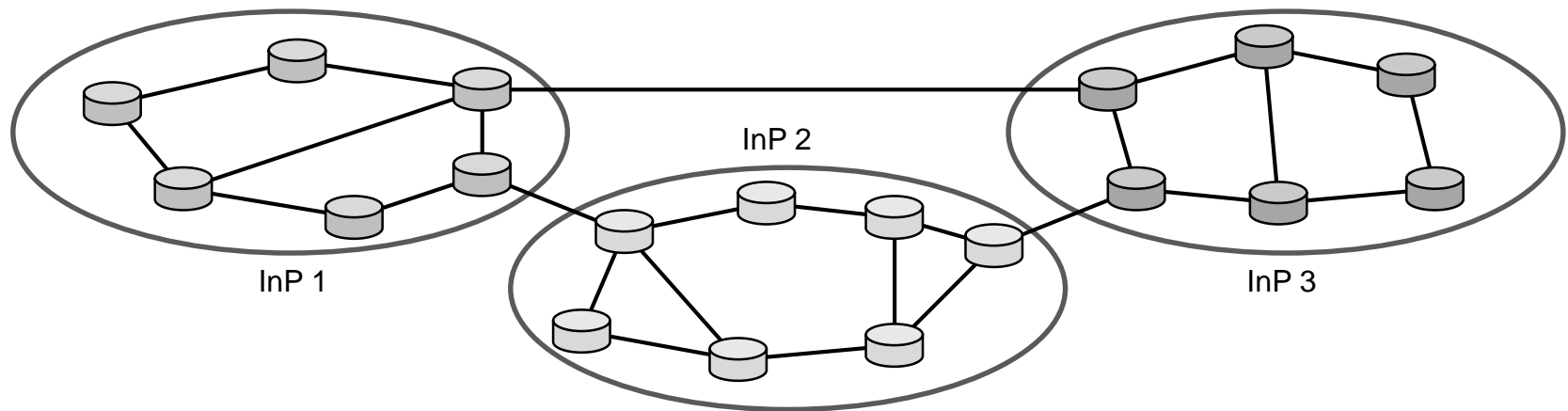


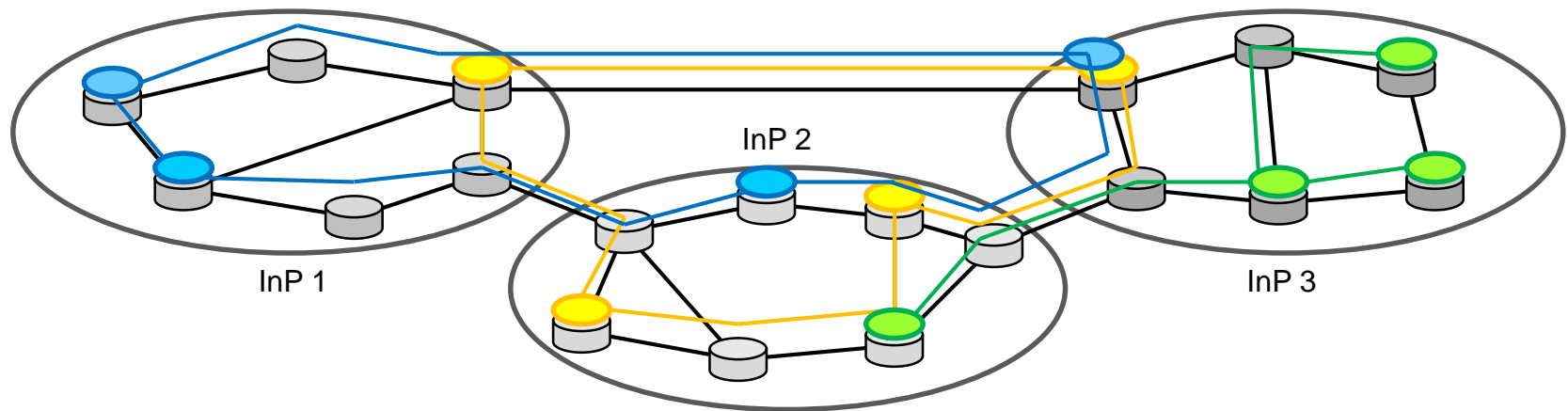
- VN embedding algorithm with simultaneous virtual node and link assignment
  - Objective: Load balancing across the substrate





# Virtual Network Embedding Across Multiple Substrate Networks



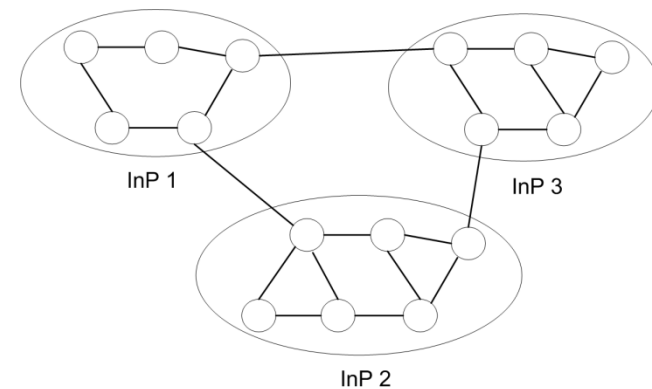




- Service Provider (SP)
  - Deploys services on VNs

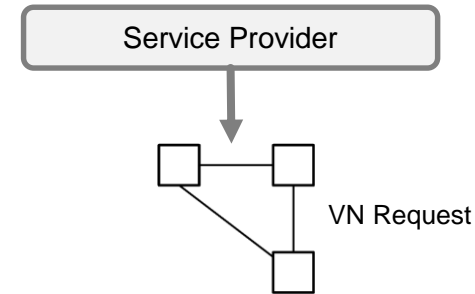
Service Provider

- Infrastructure Provider (InP)
  - Owns and manages the physical infrastructure
  - Leases resources for VNs

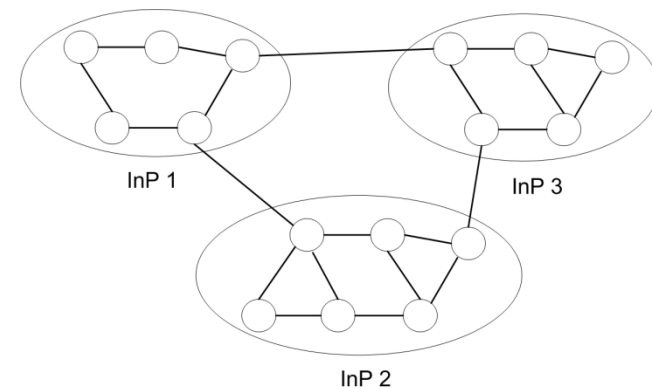




- Service Provider (SP)
  - Deploys services on VNs

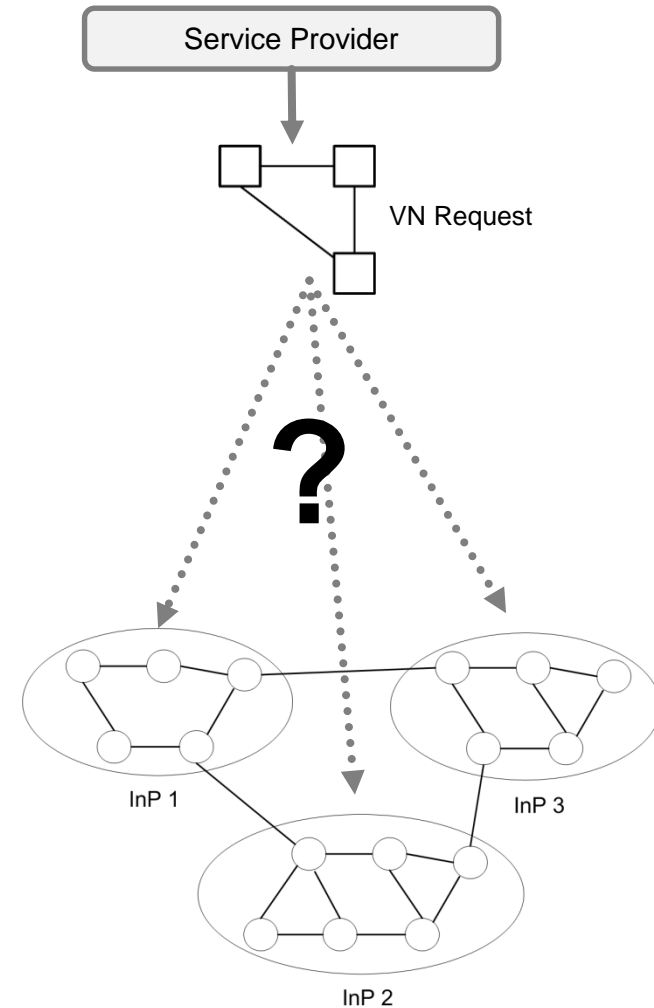


- Infrastructure Provider (InP)
  - Owns and manages the physical infrastructure
  - Leases resources for VNs





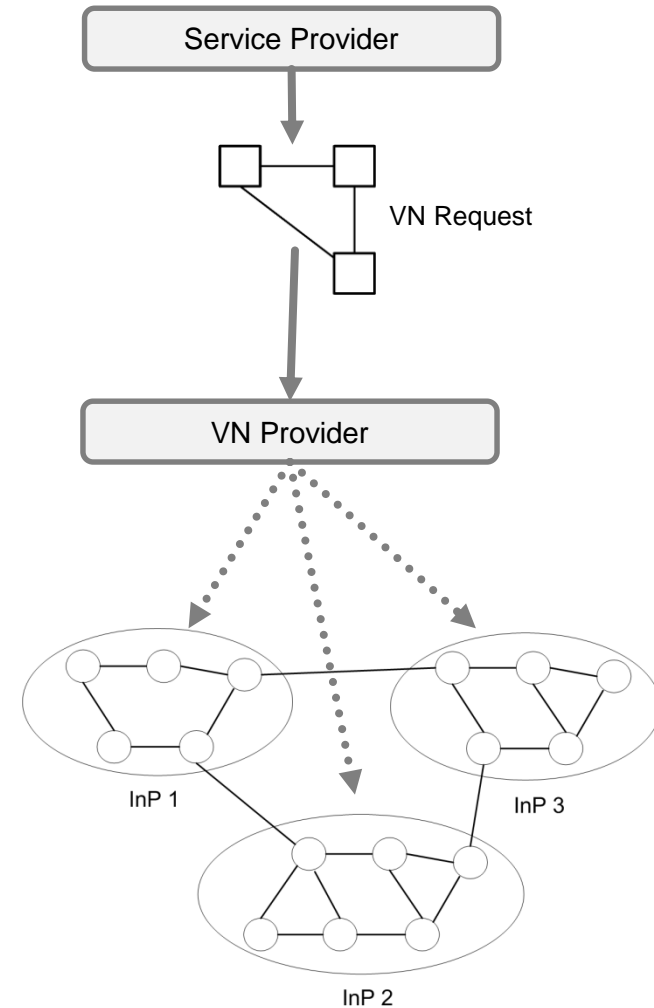
- Service Provider (SP)
  - Deploys services on VNs
- Infrastructure Provider (InP)
  - Owns and manages the physical infrastructure
  - Leases resources for VNs





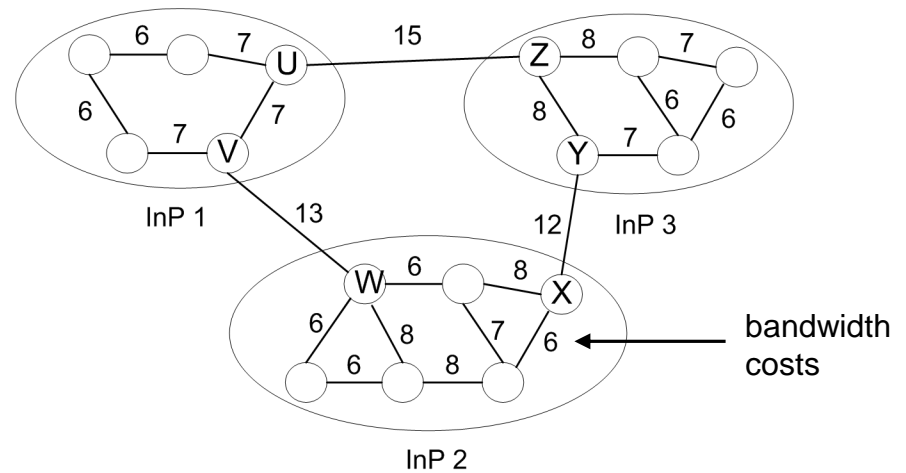
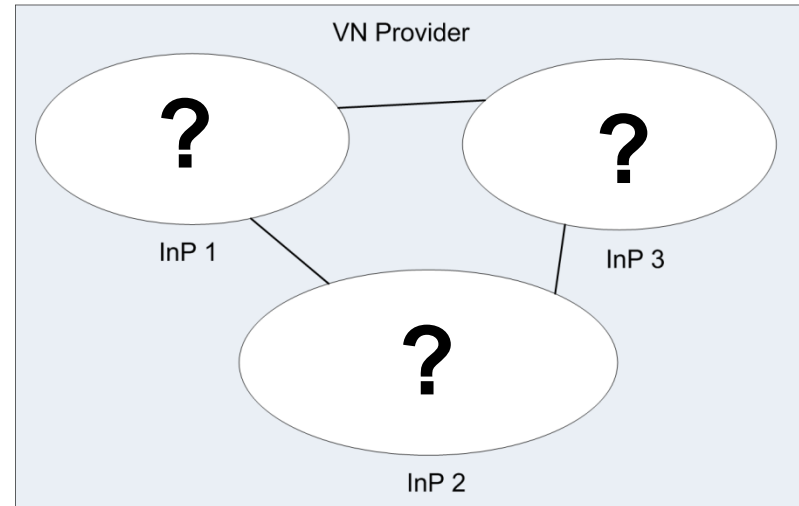


- Service Provider (SP)
  - Deploys services on VNs
- Virtual Network Provider (VNP)
  - Assembles resources from one or multiple InPs into a VN
- Infrastructure Provider (InP)
  - Owns and manages the physical infrastructure
  - Leases resources for VNs



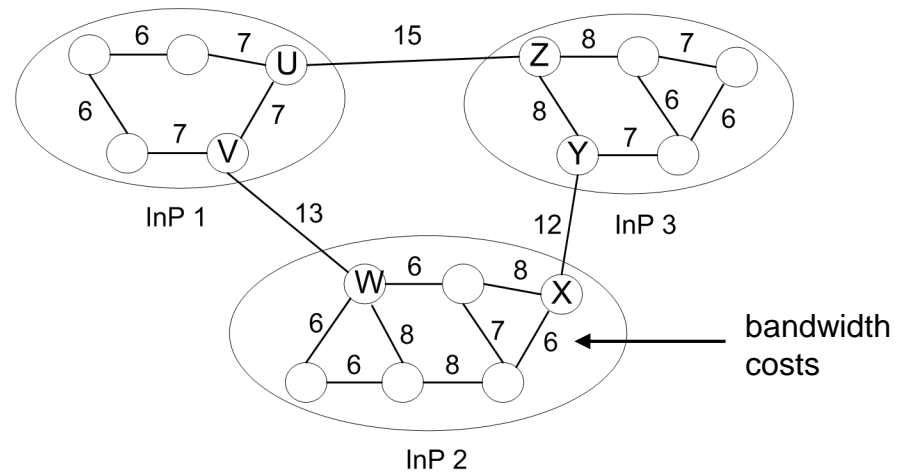
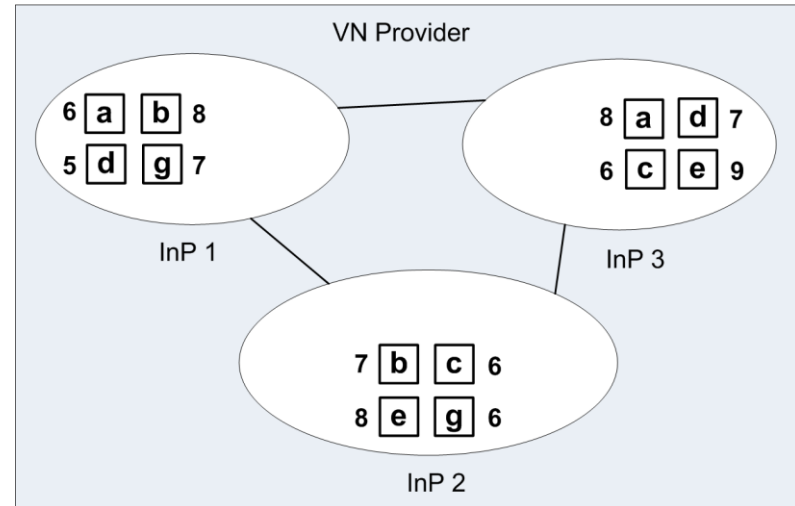


- VN Provider's visibility on substrate network topology and resources is limited to:



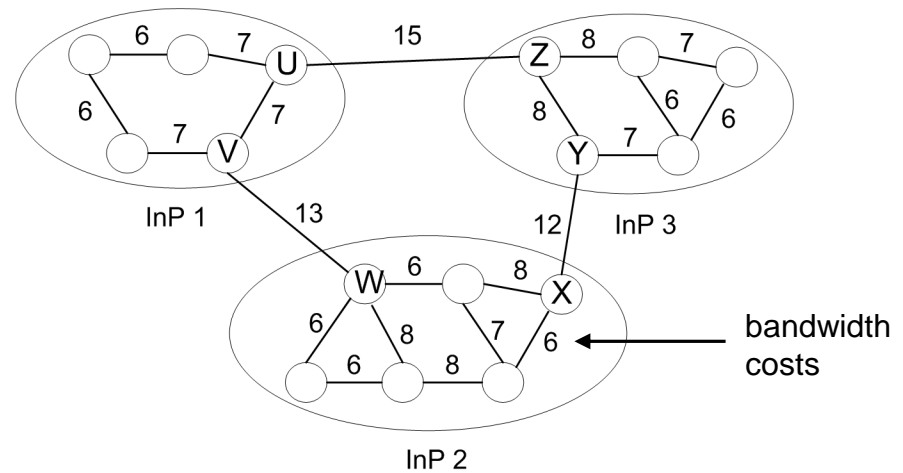
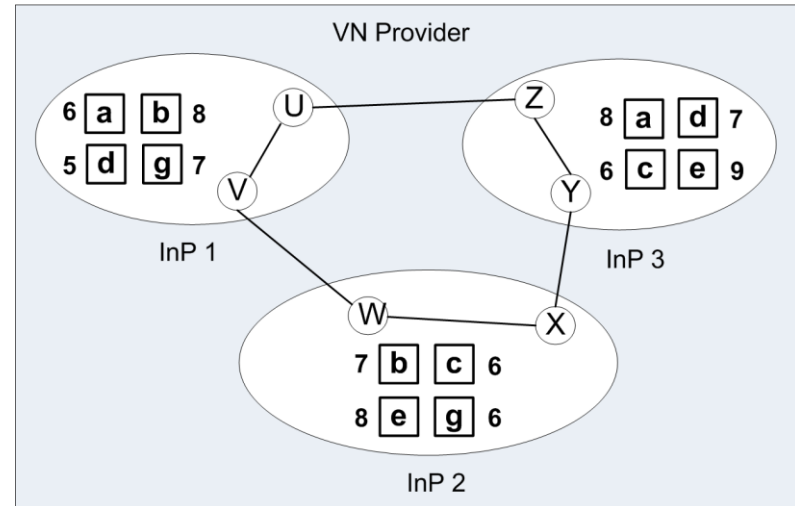


- VN Provider's visibility on substrate network topology and resources is limited to:
  - Offered virtual node types (similar to Amazon EC2)



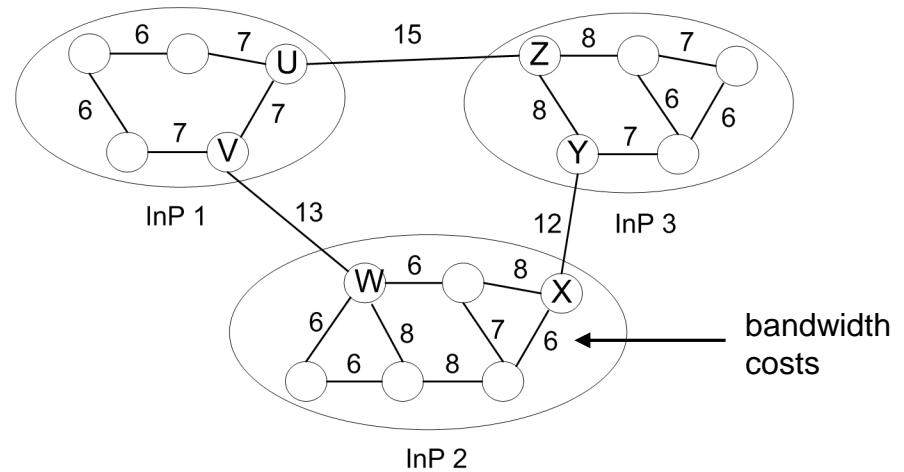
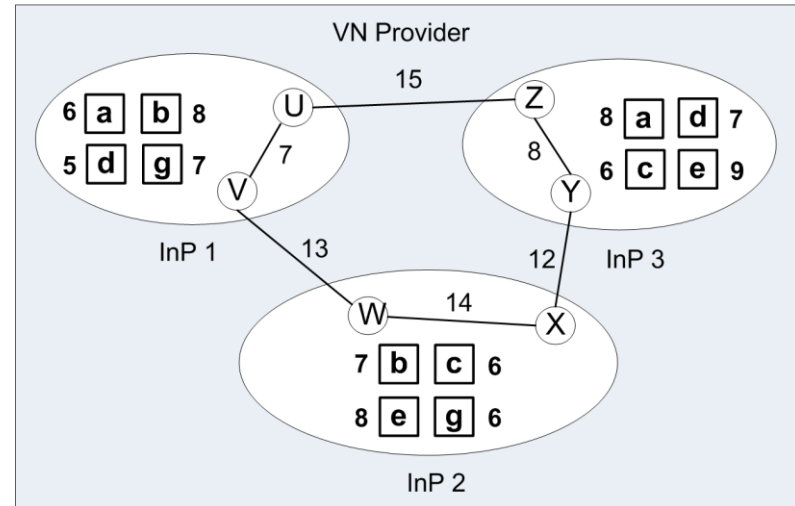


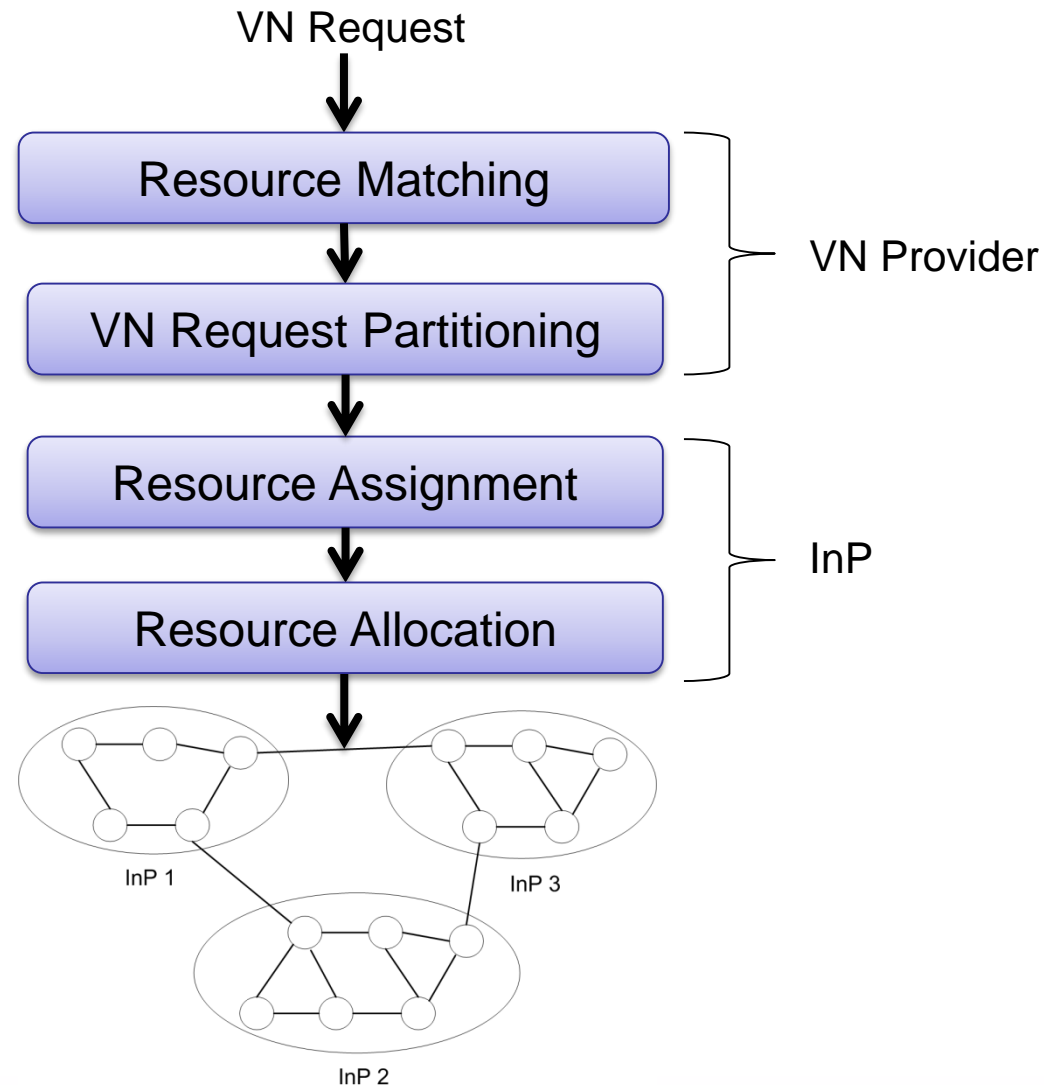
- VN Provider's visibility on substrate network topology and resources is limited to:
  - Offered virtual node types (similar to Amazon EC2)
  - Location of peering nodes

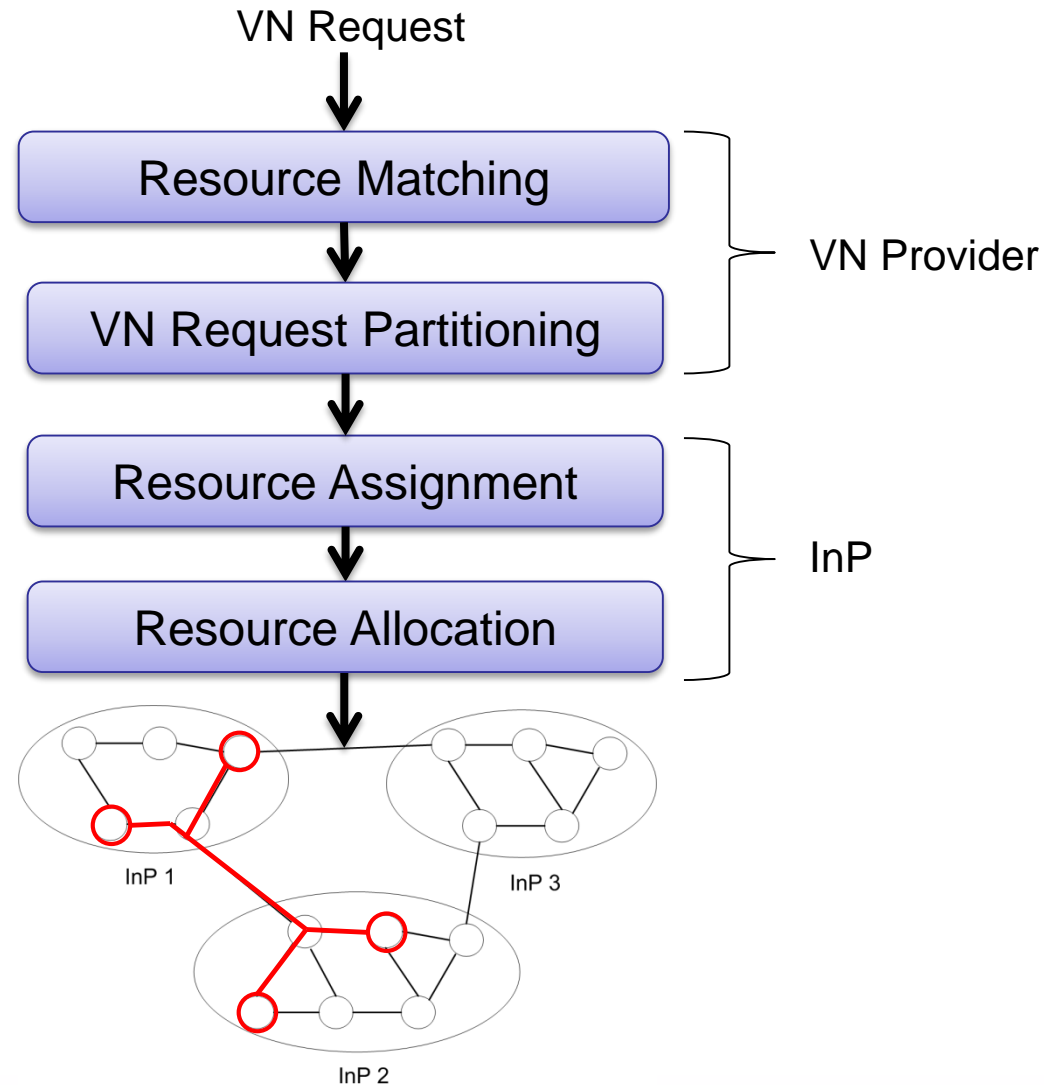




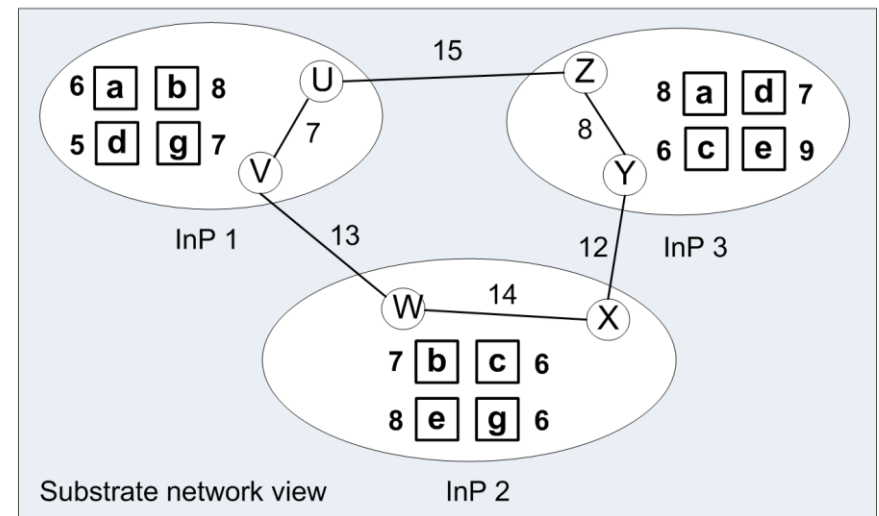
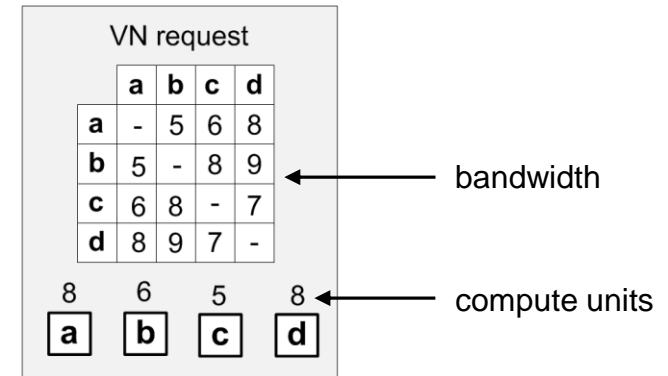
- VN Provider's visibility on substrate network topology and resources is limited to:
  - Offered virtual node types (similar to Amazon EC2)
  - Location of peering nodes
  - Bandwidth cost over links between peering nodes







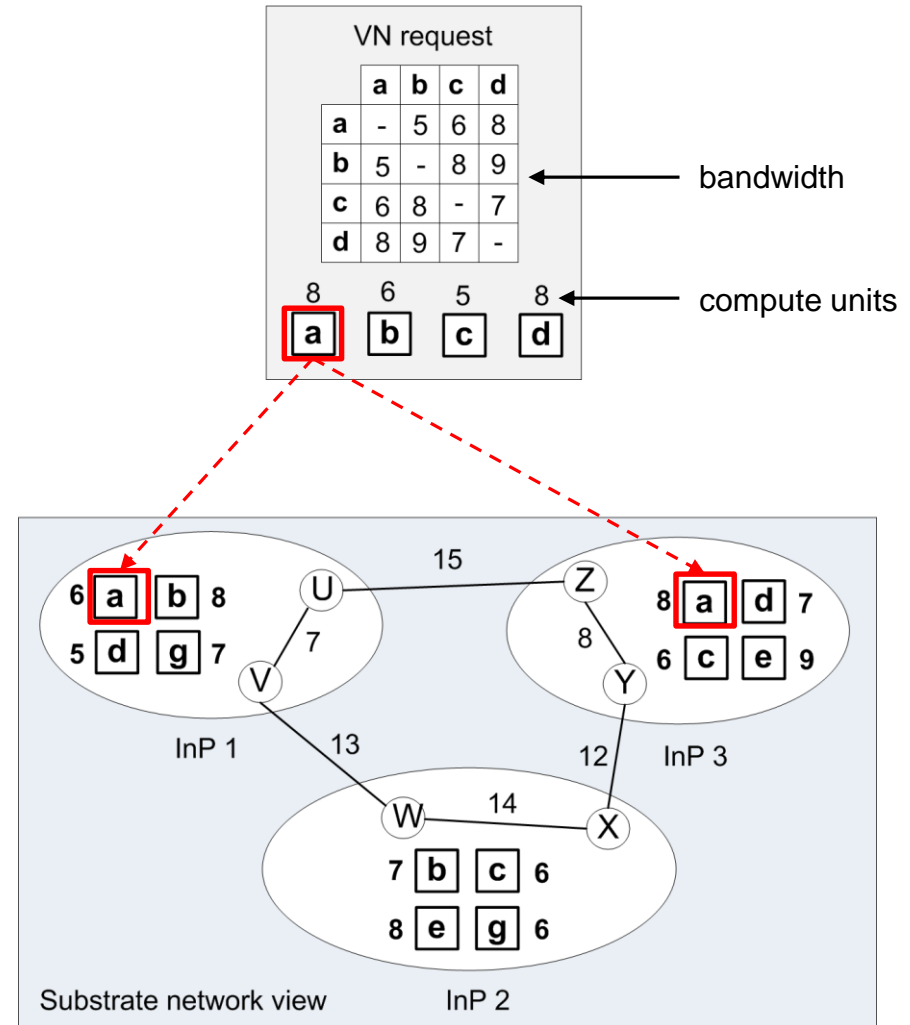
- VN Provider matches requested to advertised resources
  - Candidates for each requested resource are identified





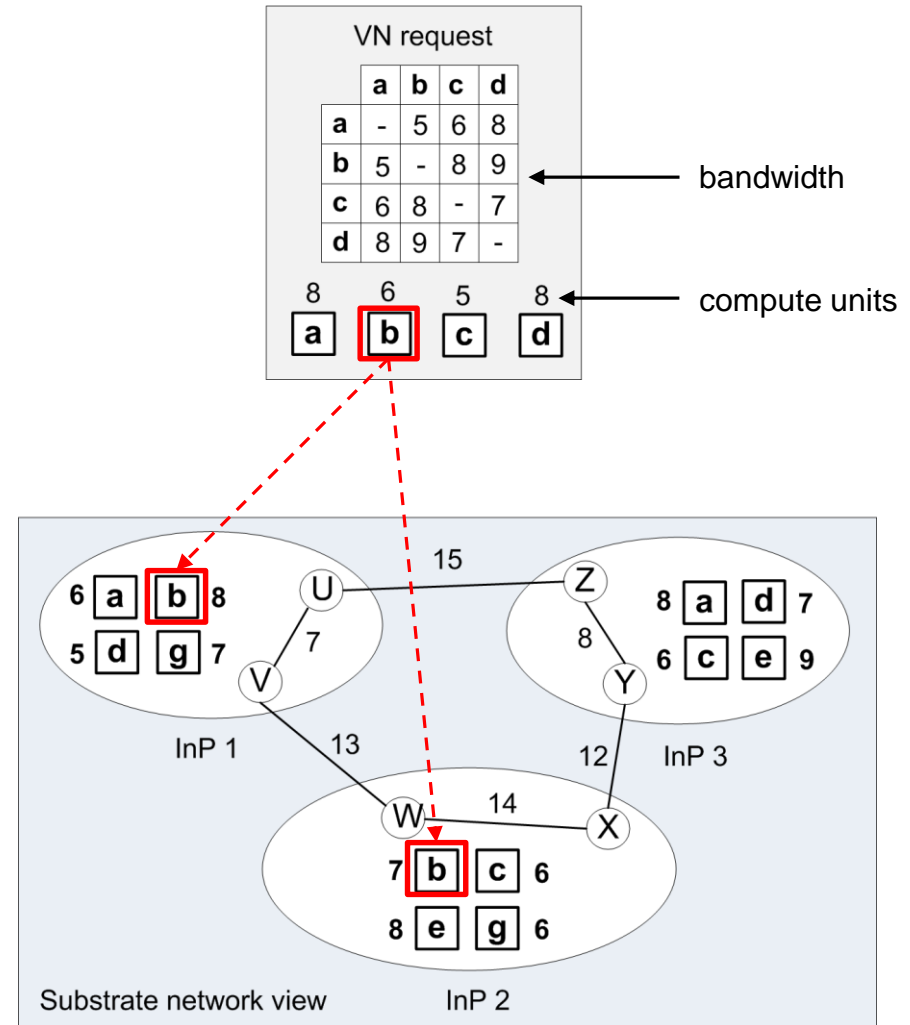


- VN Provider matches requested to advertised resources
  - Candidates for each requested resource are identified
    - $a: \{InP1, InP3\}$



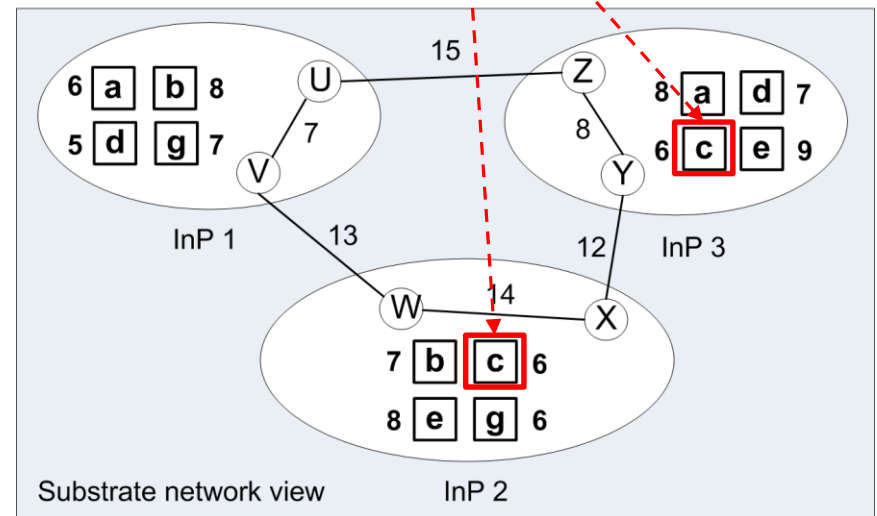
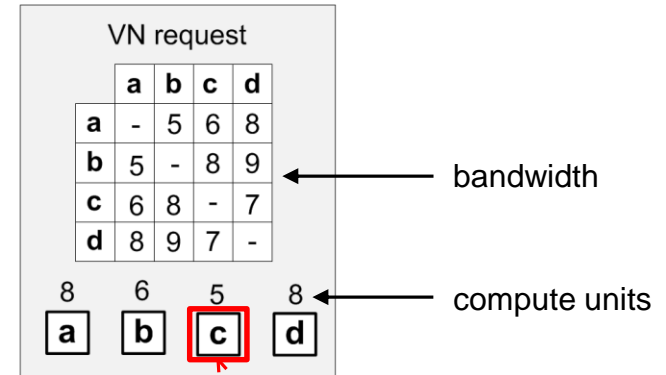


- VN Provider matches requested to advertised resources
  - Candidates for each requested resource are identified
    - $a: \{InP1, InP3\}$
    - $b: \{InP1, InP2\}$



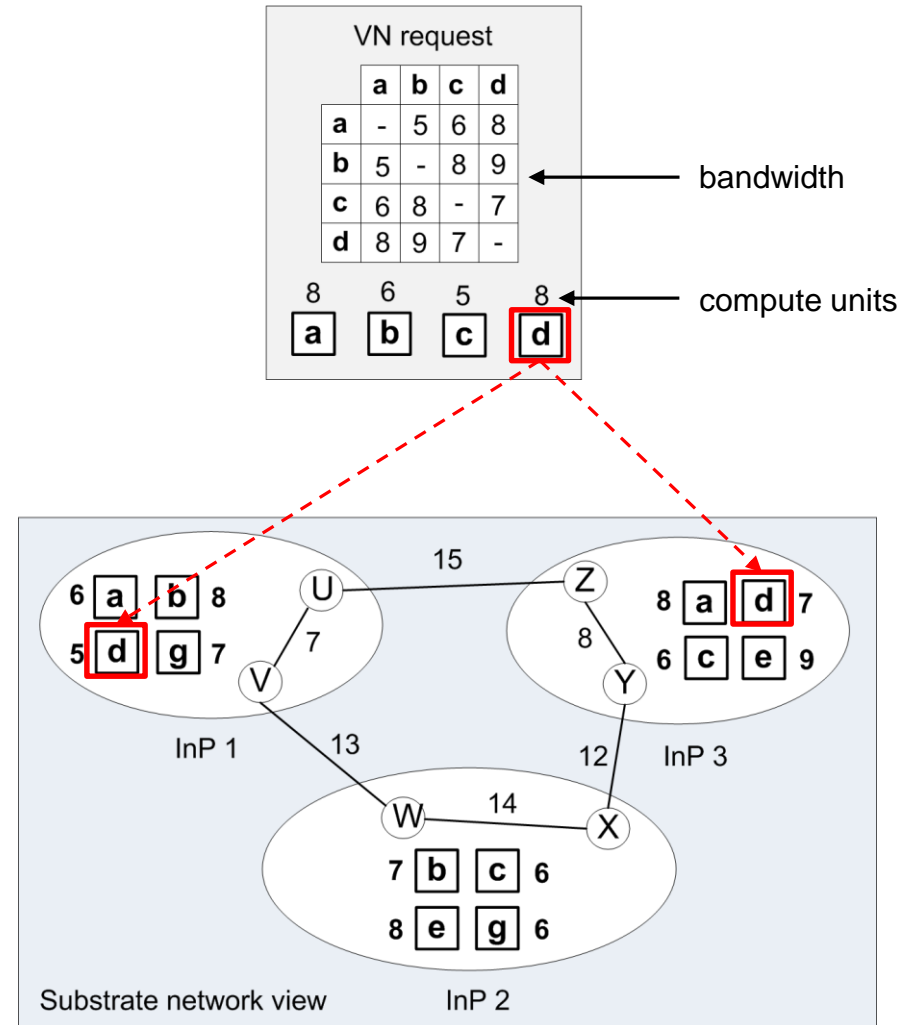


- VN Provider matches requested to advertised resources
  - Candidates for each requested resource are identified
    - a: {InP1, InP3}
    - b: {InP1, InP2}
    - c: {InP2, InP3}



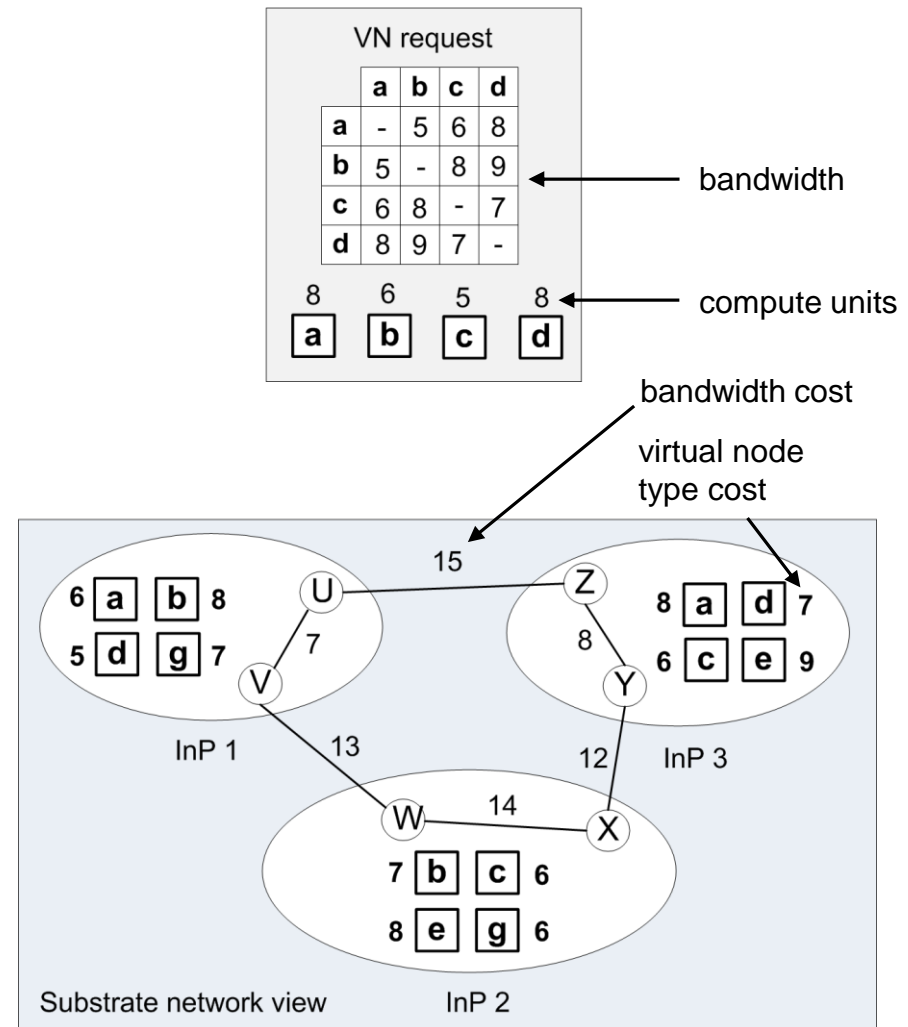


- VN Provider matches requested to advertised resources
  - Candidates for each requested resource are identified
    - a: {InP1, InP3}
    - b: {InP1, InP2}
    - c: {InP2, InP3}
    - d: {InP1, InP3}



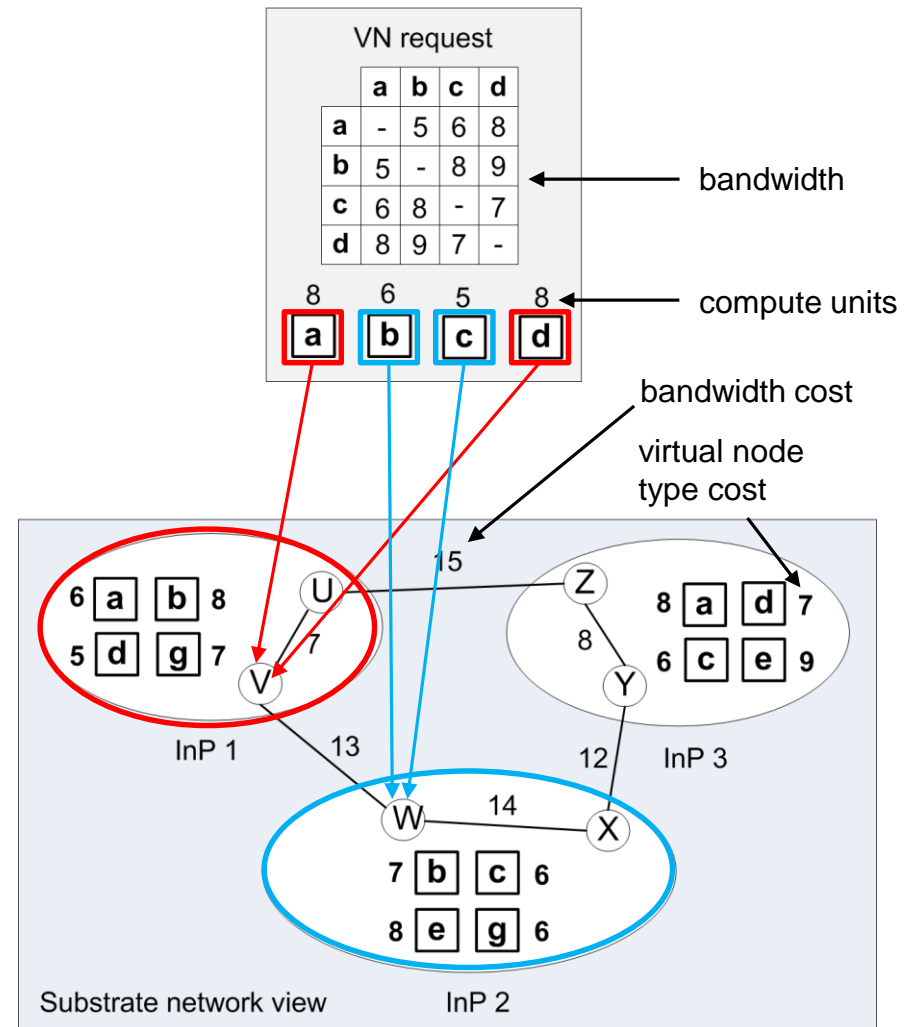


- Binary integer program:
  - Objective:
    - Minimize the cost for the Service Provider



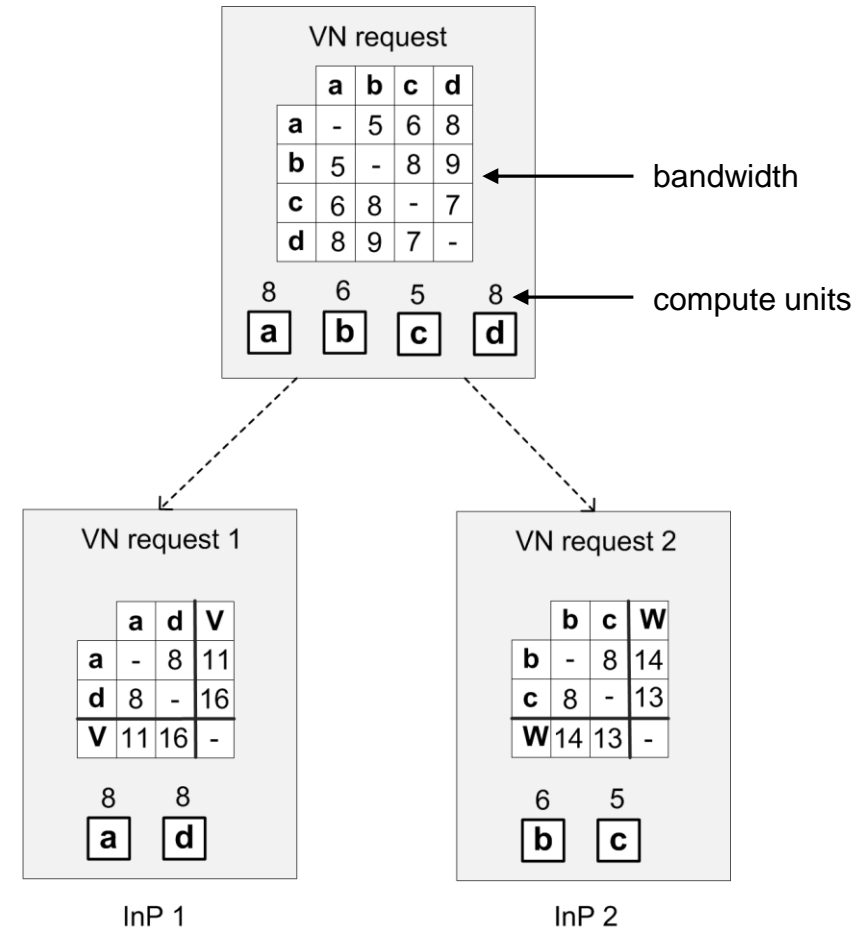


- Binary integer program:
  - Objective:
    - Minimize the cost for the Service Provider
  - Solution:
    - Virtual node to peering node assignment



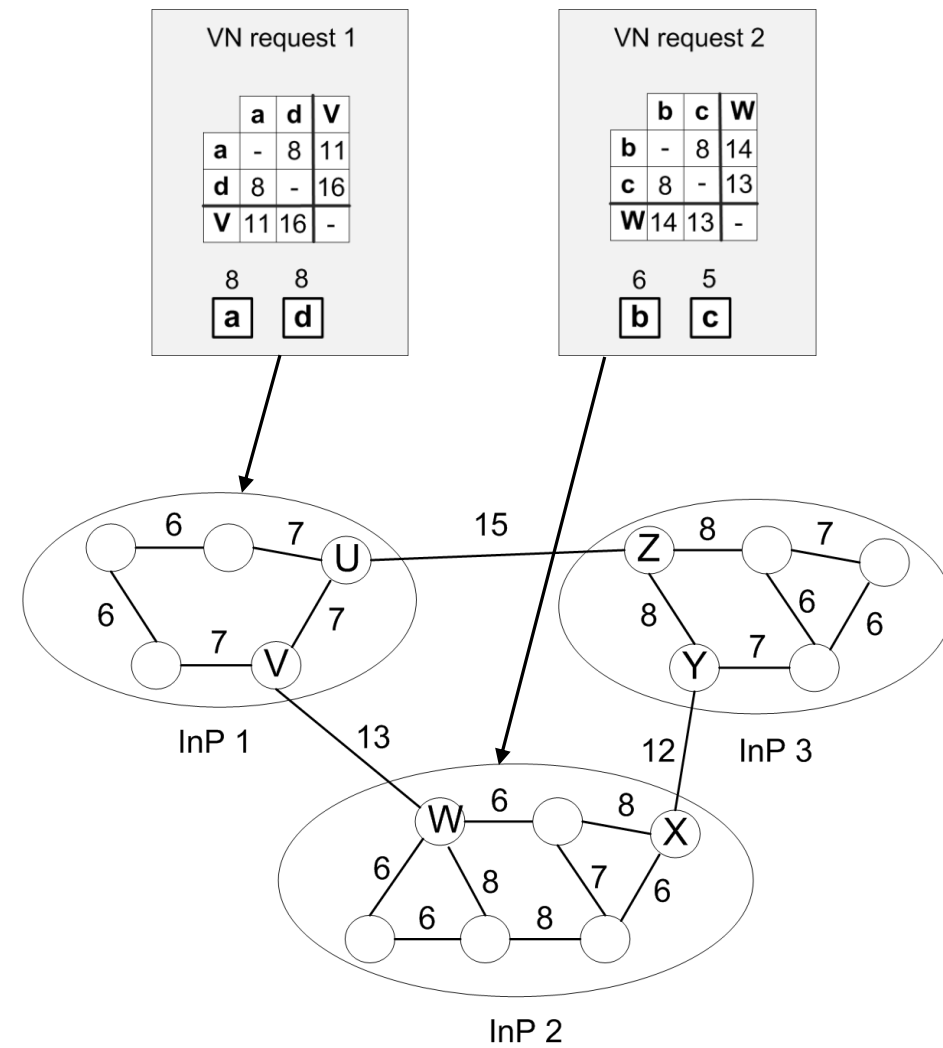


- Binary integer program:
  - Objective:
    - Minimize the cost for the Service Provider
  - Solution:
    - Virtual node to peering node assignment
- VN segment description:
  - Virtual node specifications
  - Bandwidth demands between virtual nodes and peering nodes





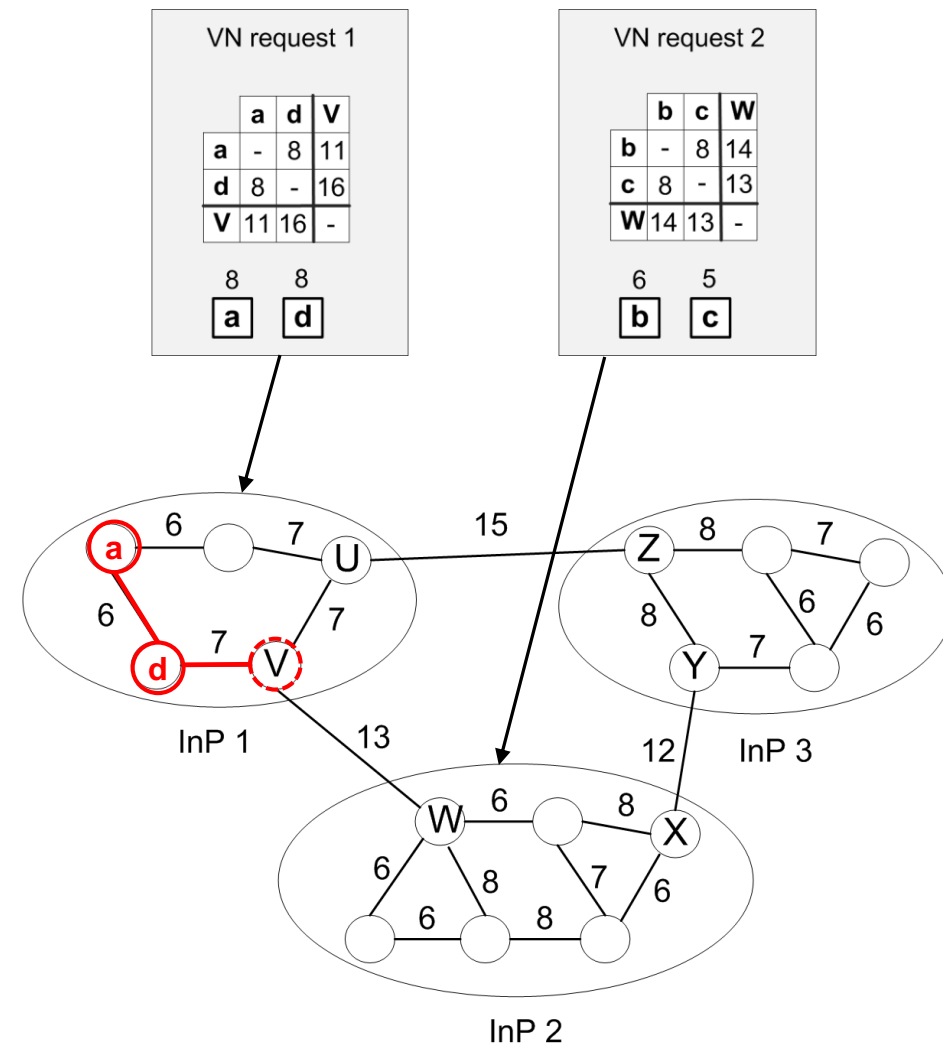
- Mixed integer multi-commodity flow problem:
  - Objective:
    - Minimize the embedding cost





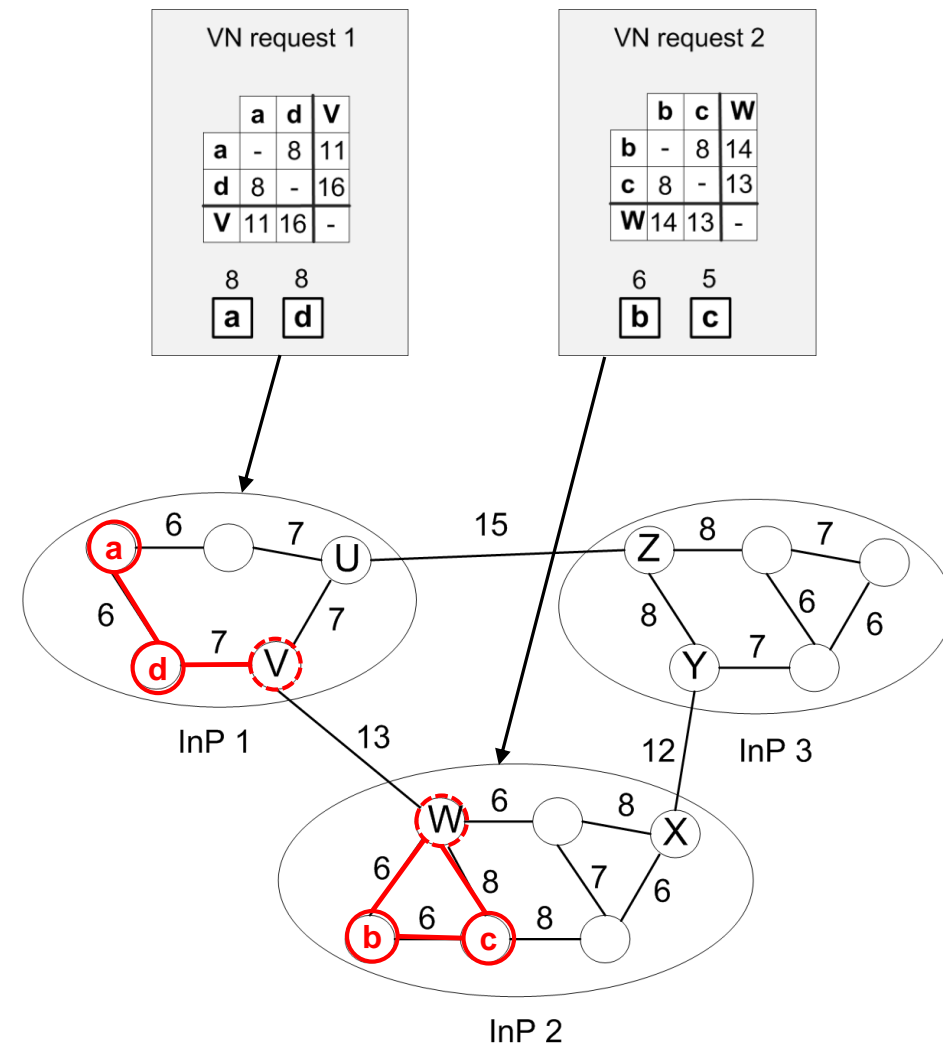


- Mixed integer multi-commodity flow problem:
  - Objective:
    - Minimize the embedding cost
  - Solution:
    - VN segment mapping onto the substrate network
      - Compliance with virtual node to peering node bindings



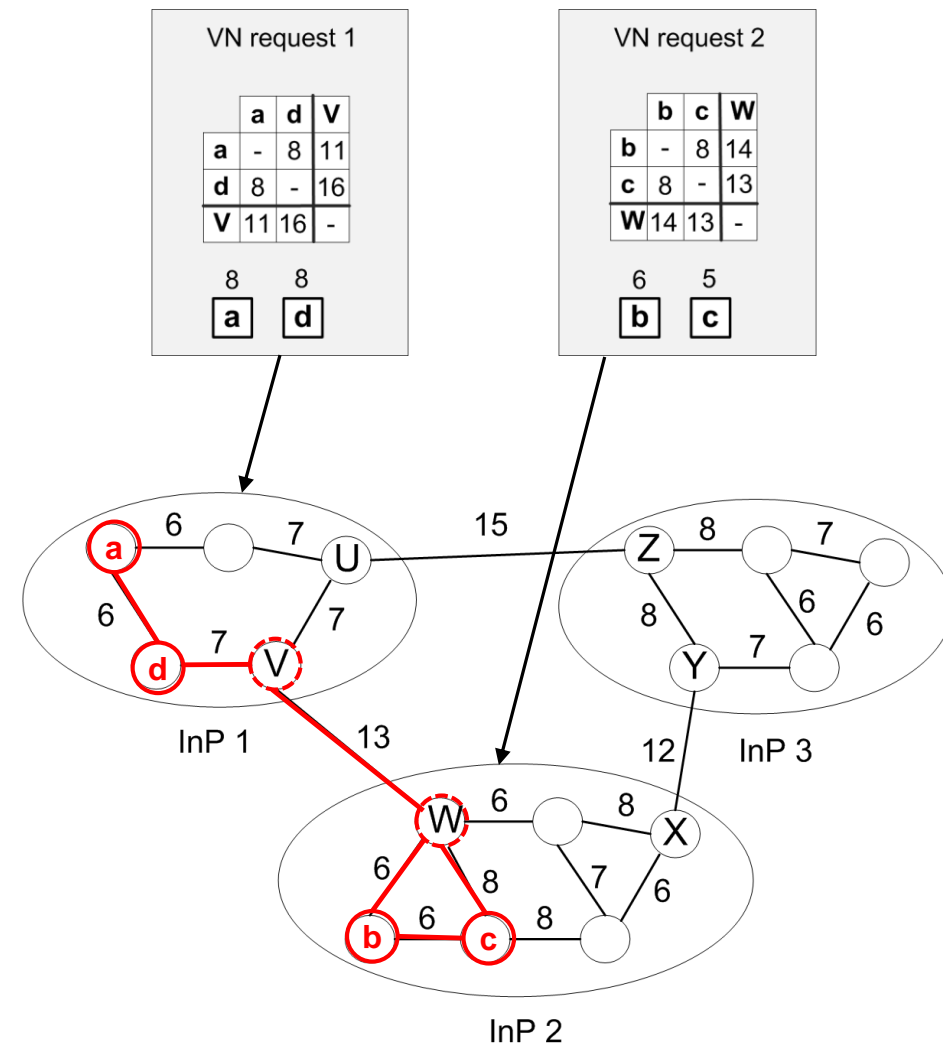


- Mixed integer multi-commodity flow problem:
  - Objective:
    - Minimize the embedding cost
  - Solution:
    - VN segment mapping onto the substrate network
      - Compliance with virtual node to peering node bindings



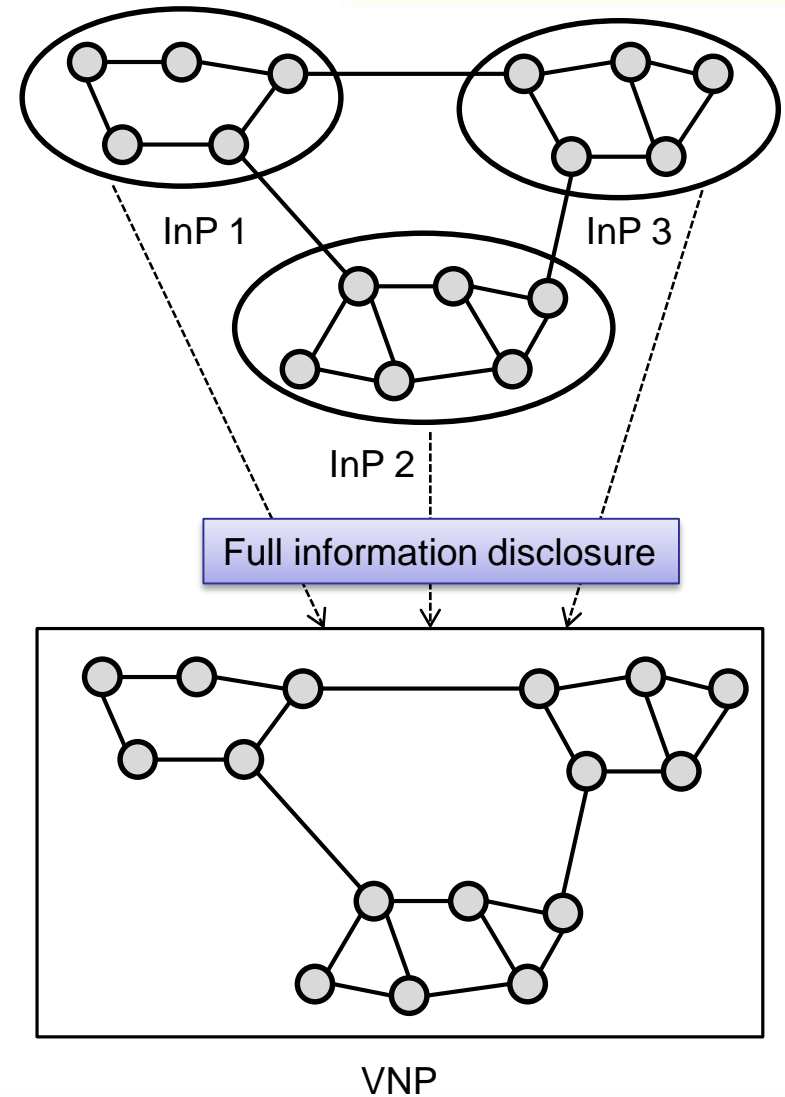
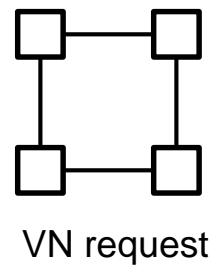


- Mixed integer multi-commodity flow problem:
  - Objective:
    - Minimize the embedding cost
  - Solution:
    - VN segment mapping onto the substrate network
      - Compliance with virtual node to peering node bindings



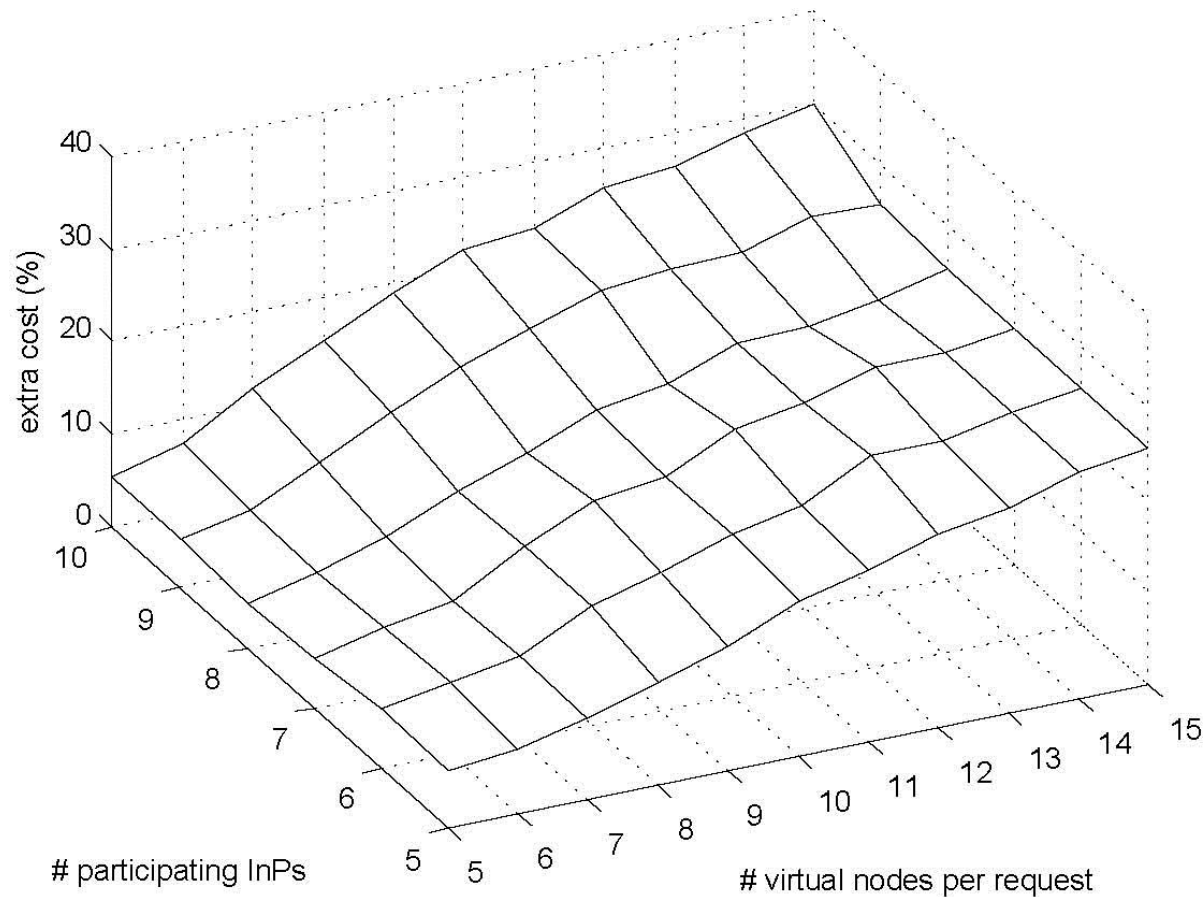


Limited information disclosure (LID)  
vs.  
Full information disclosure (FID)



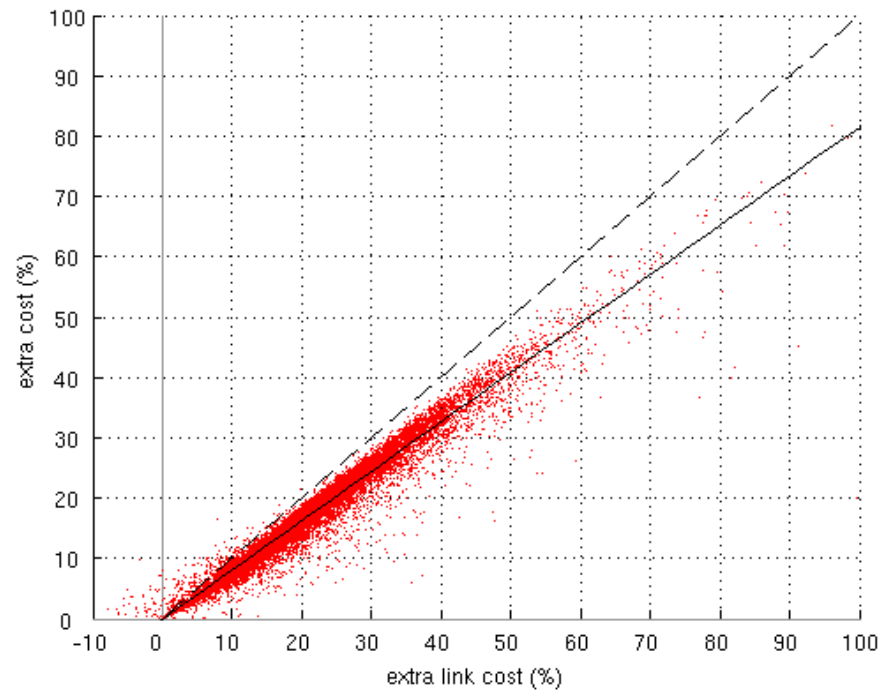
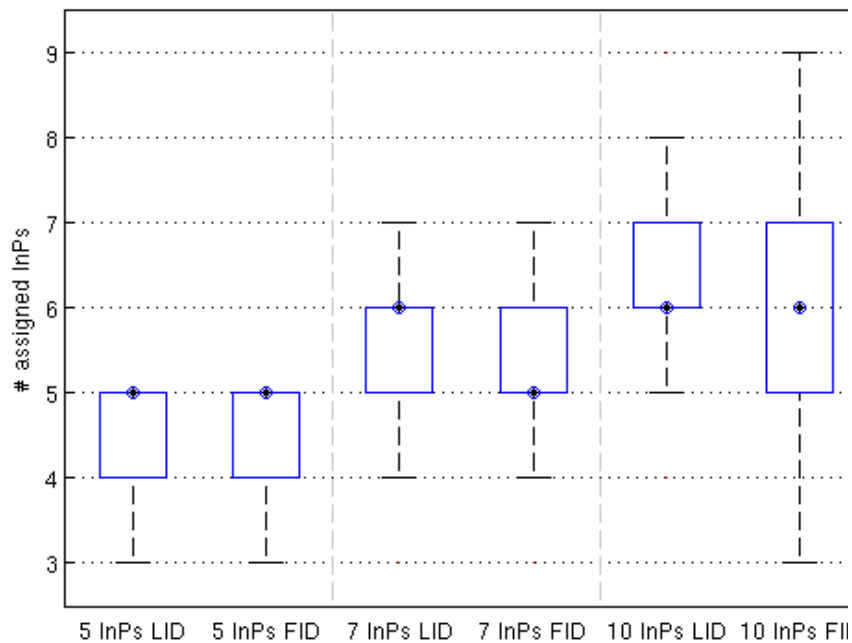


- LID incurs 5-30% extra cost



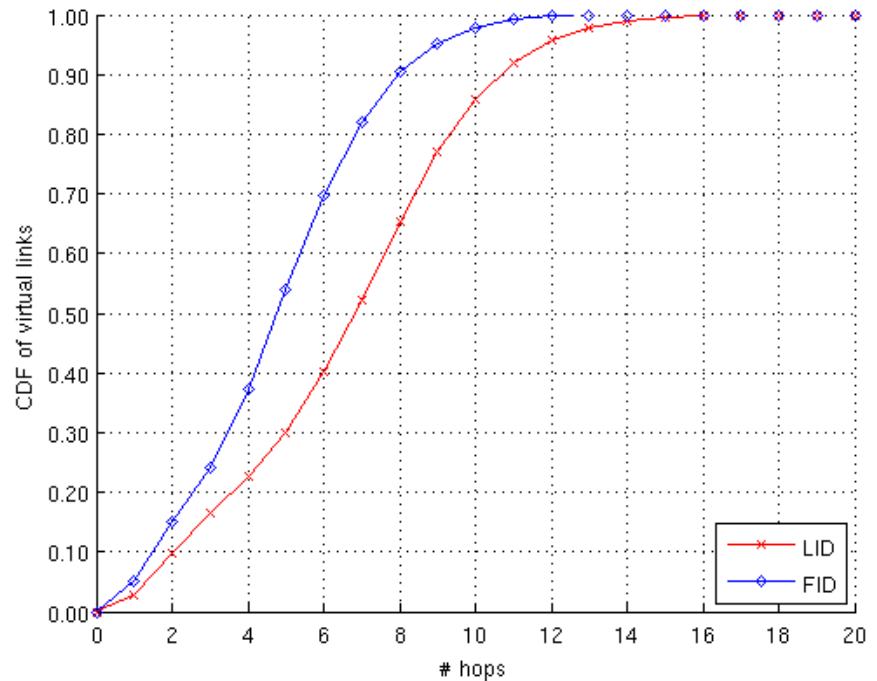
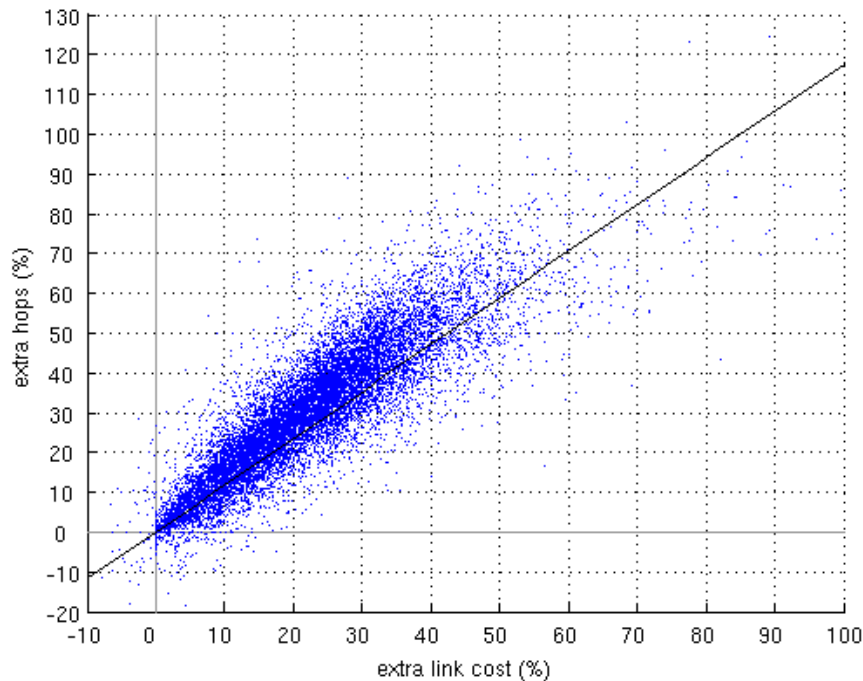


- Same number of assigned InPs under LID and FID
- Extra cost is correlated with extra link cost





- Extra link cost is due to increased hop count





# Virtual Link Setup





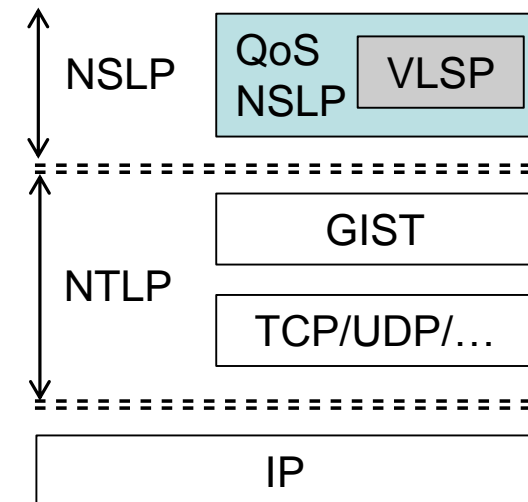
- Objective:
  - Interoperable solution for virtual link setup with QoS guarantees across InPs
  
- Approach:
  - Couple virtual link setup signaling with QoS reservation signaling for efficiency
    - Rely on existing QoS resource reservation protocol (IETF NSIS)
    - Add new object to NSIS QoS NSLP to carry the required information for virtual link setup
  
- Requirements:
  - NSIS support in routers
  - IP-based substrate
  - New QoS NSLP object support (only) in virtual link end-points



- NSIS QoS NSLP extension with new Virtual Link Setup Protocol (VLSP) object:
  - Virtual link setup at the end-points via VLSP
  - Resource reservation and QoS via NSLP object at the intermediate nodes

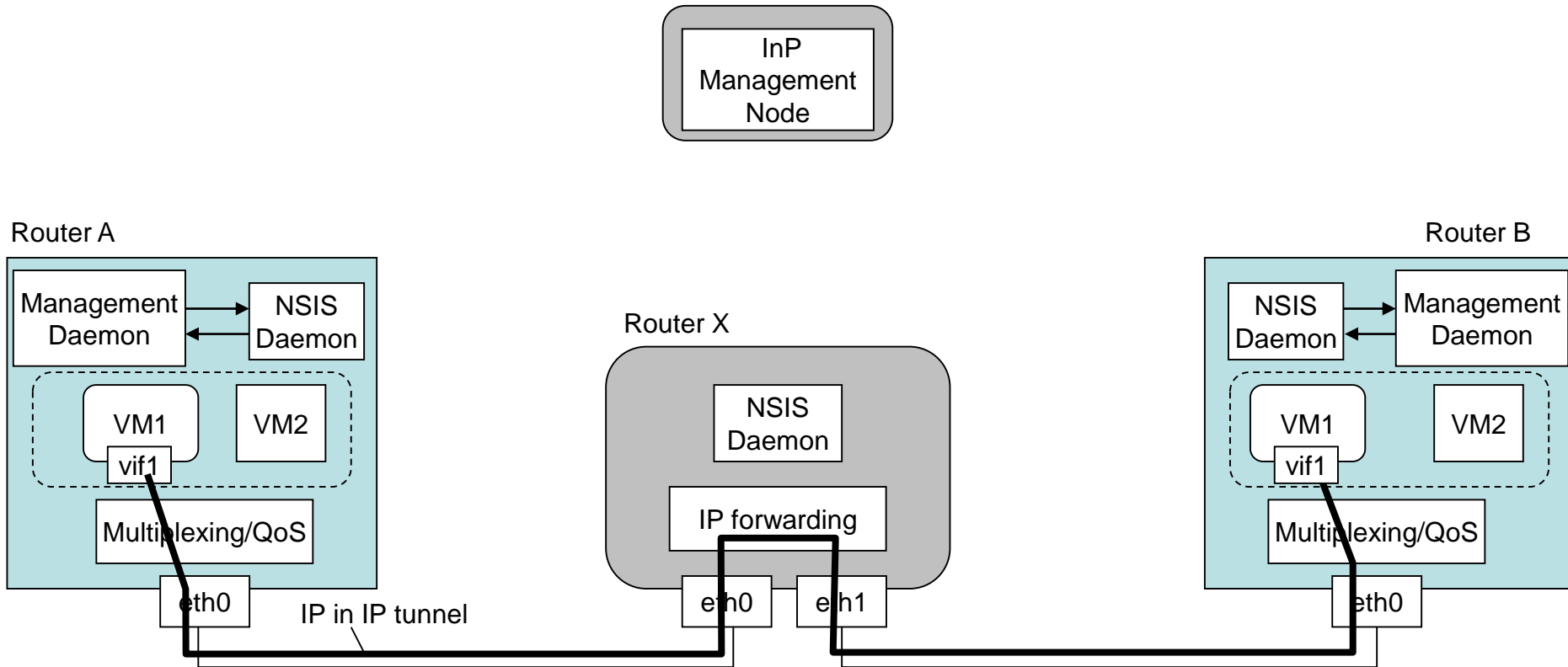
0	64	127
Virtual Network ID		
Source Virtual Node ID		
Destination Virtual Node ID		
Source Virtual Interface ID		Destination Virtual Interface ID
Virtual Link ID (optional)		Virtual Link Type (optional)

VLSP object

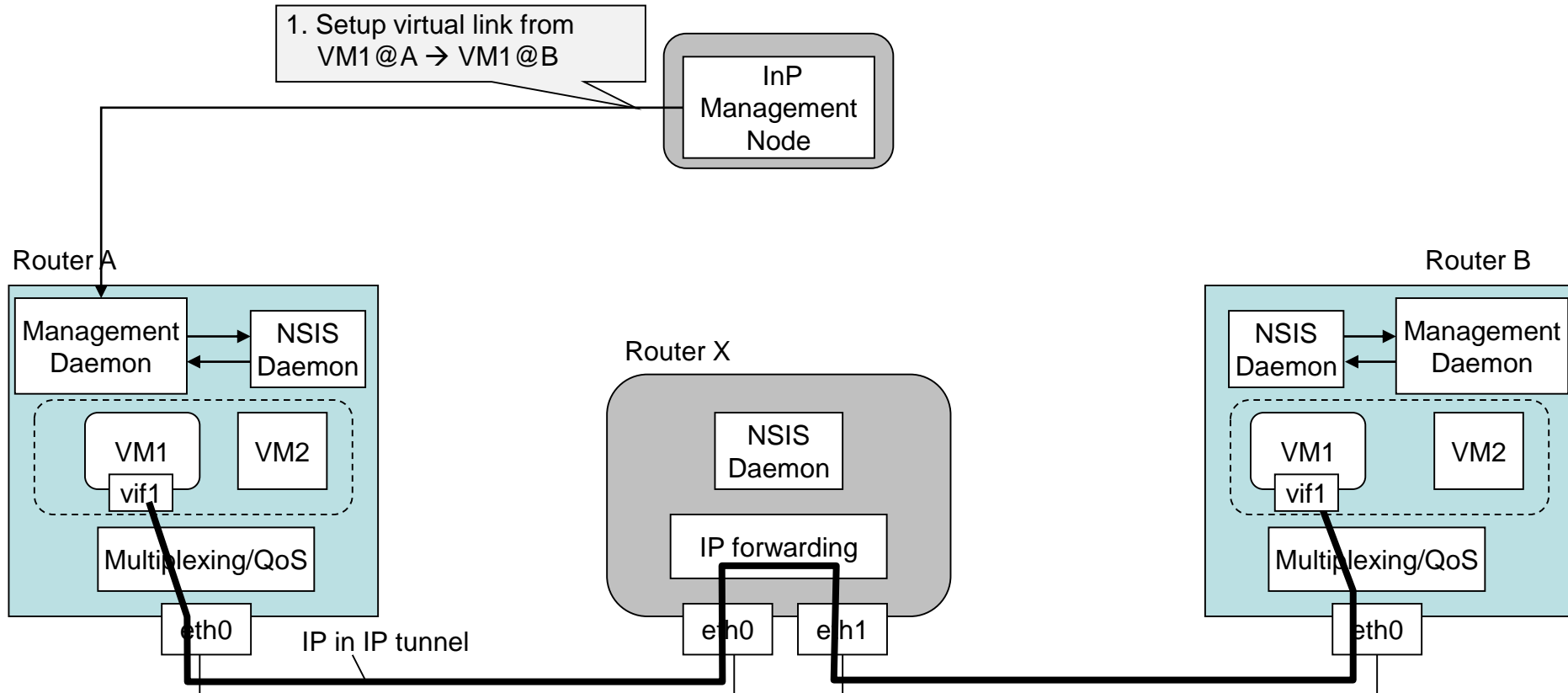


NSIS QoS NSLP/VLSP

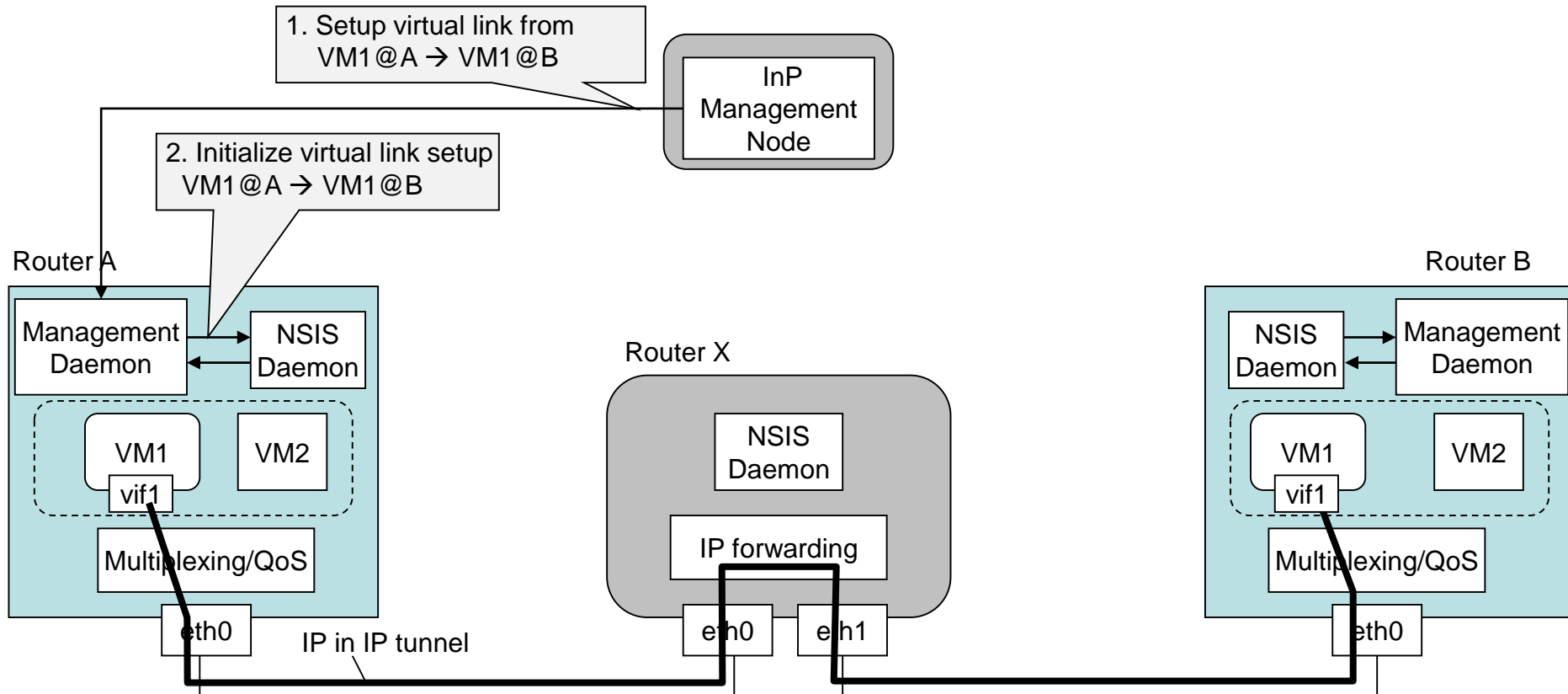
# Virtual Link Setup Workflow



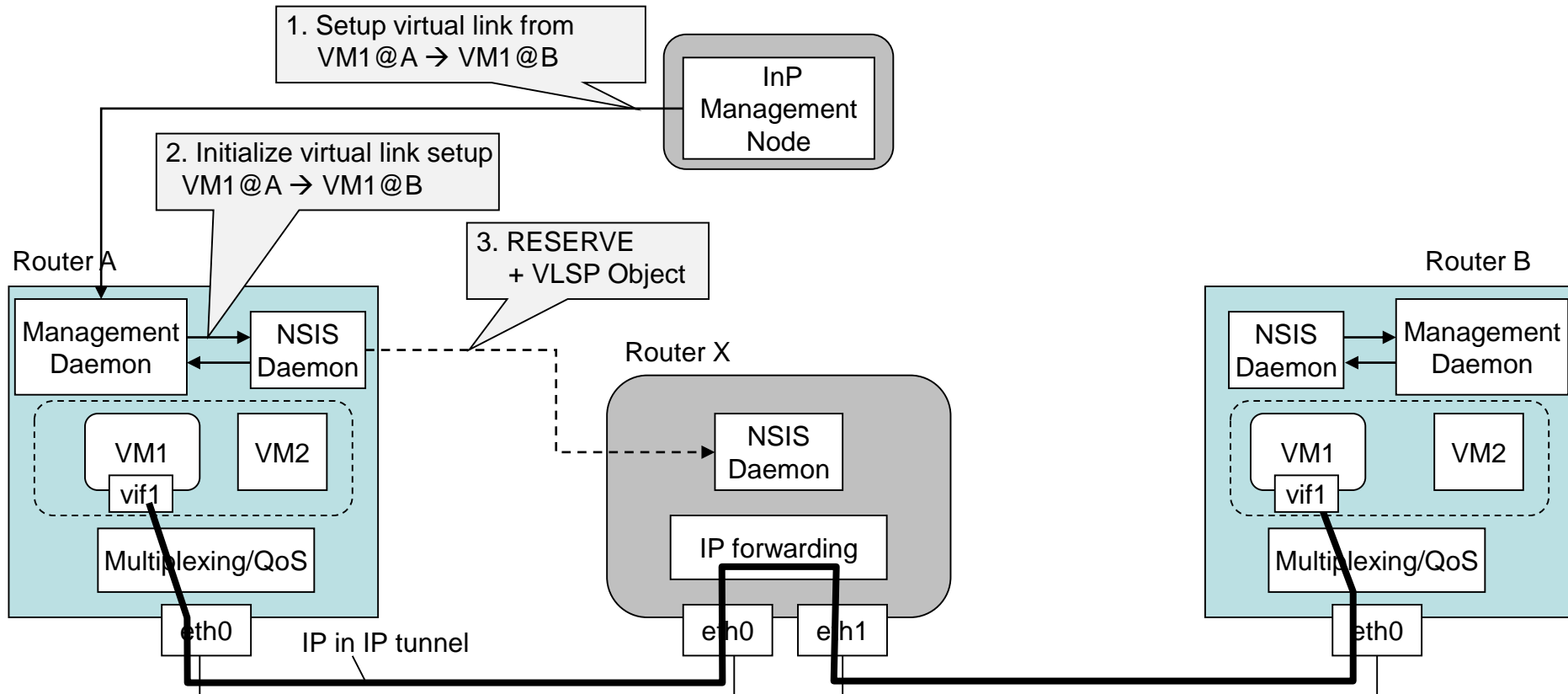
# Virtual Link Setup Workflow



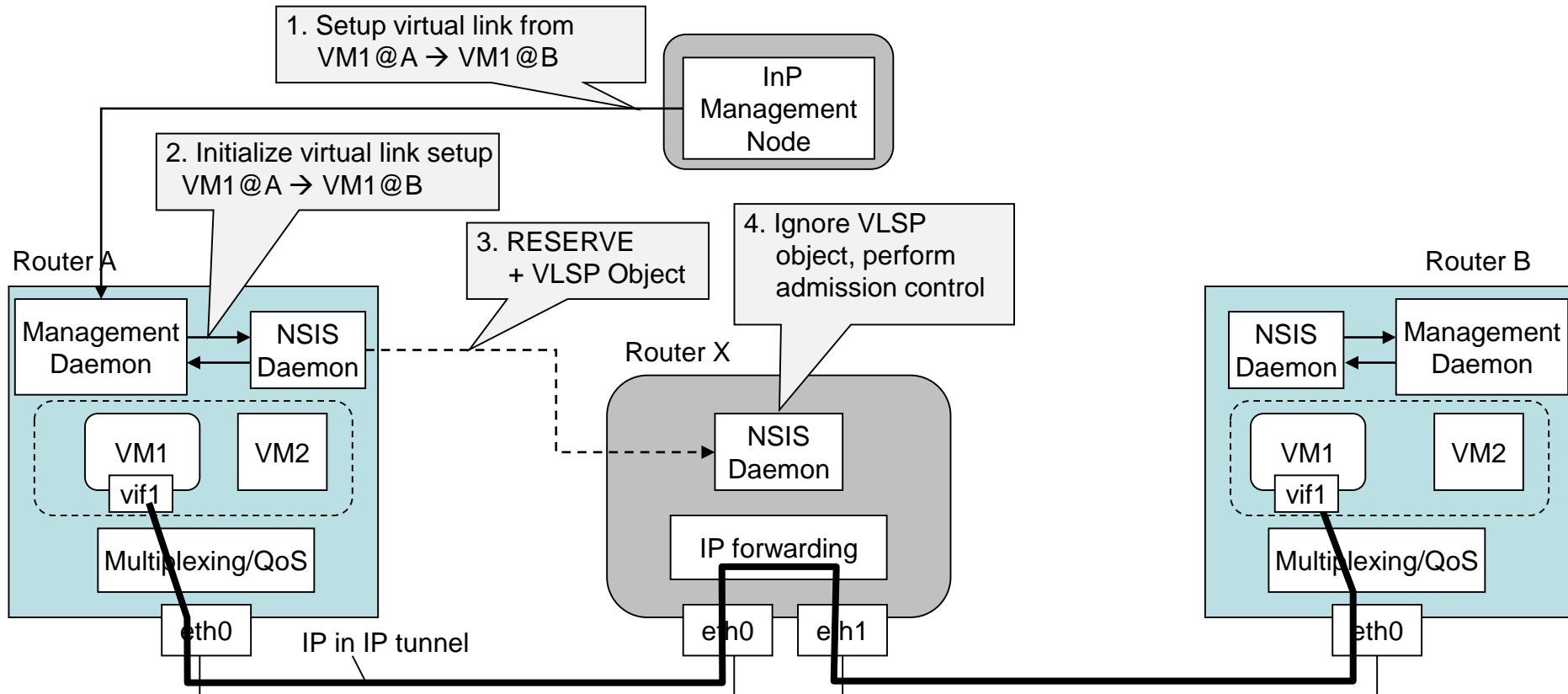
# Virtual Link Setup Workflow



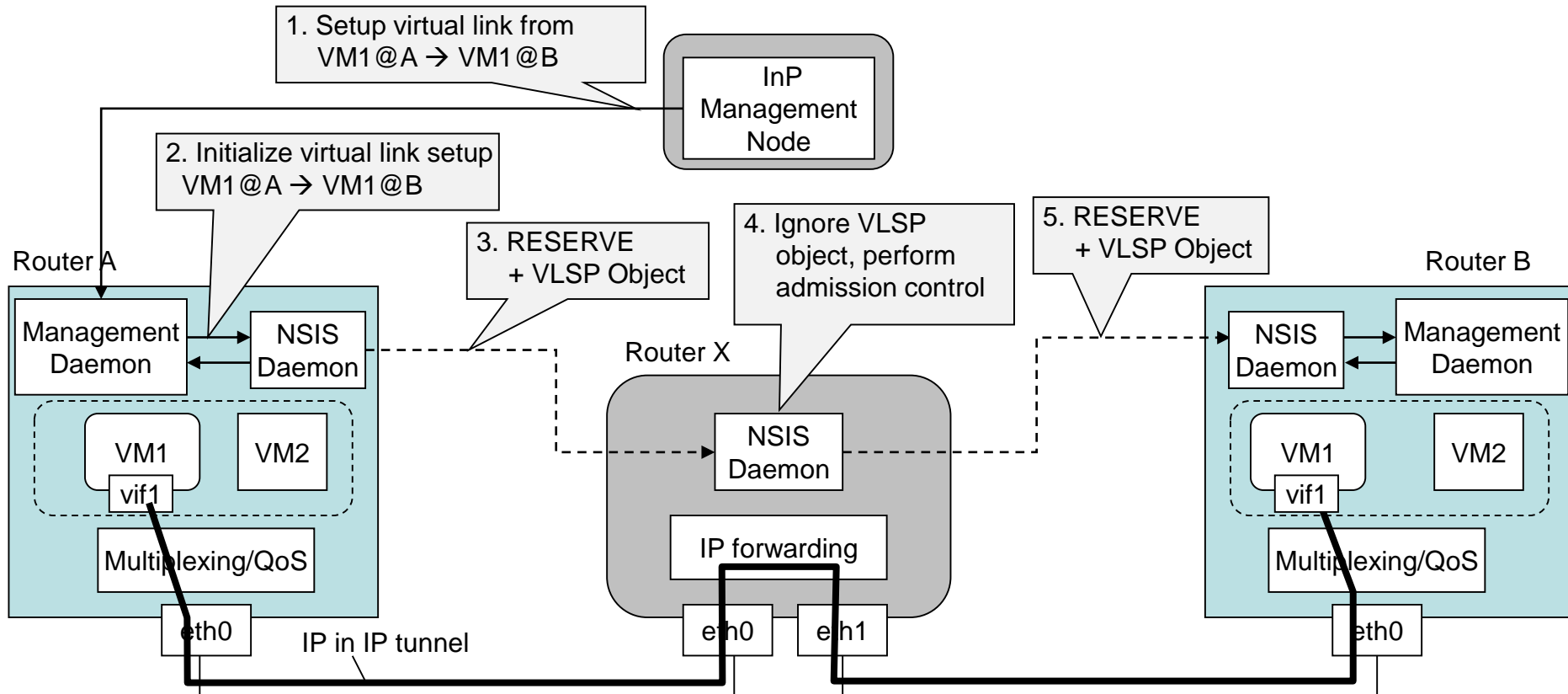
# Virtual Link Setup Workflow



# Virtual Link Setup Workflow

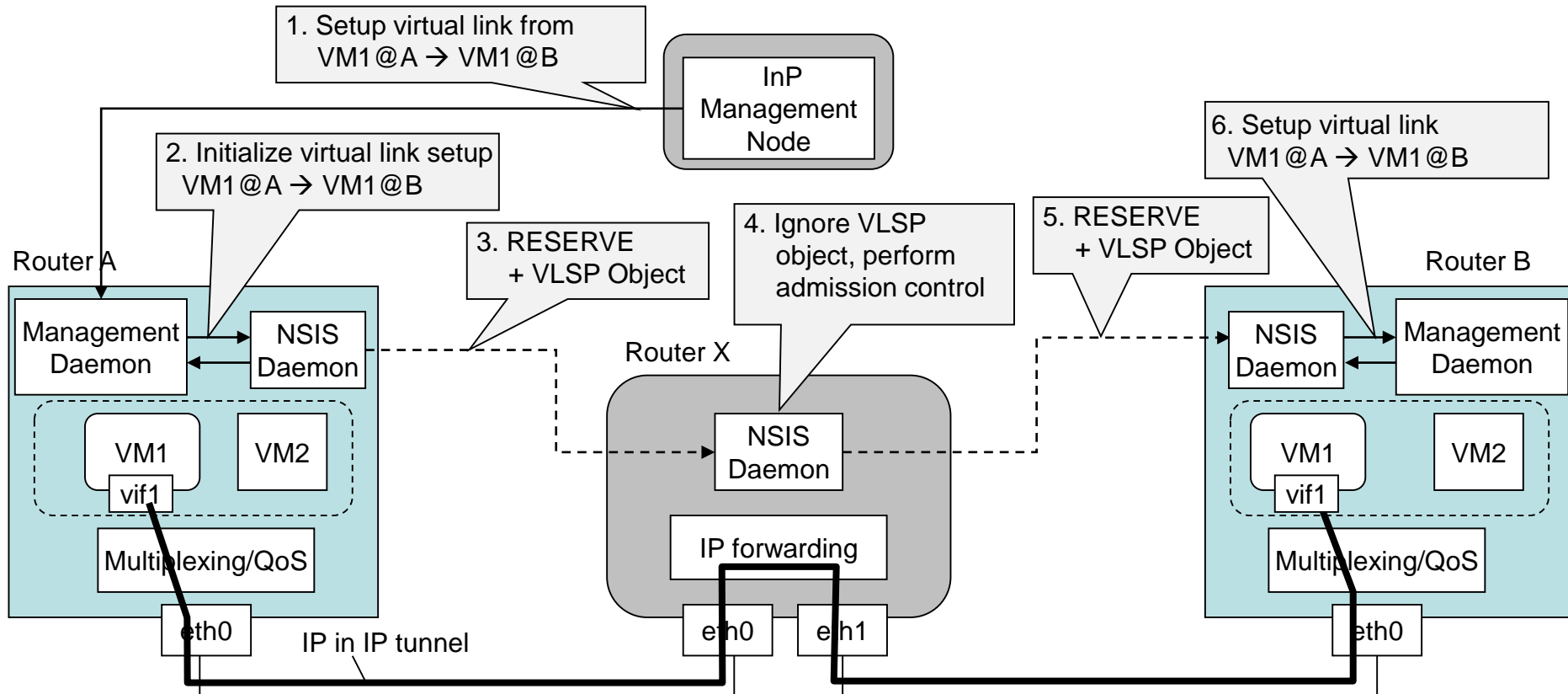


# Virtual Link Setup Workflow

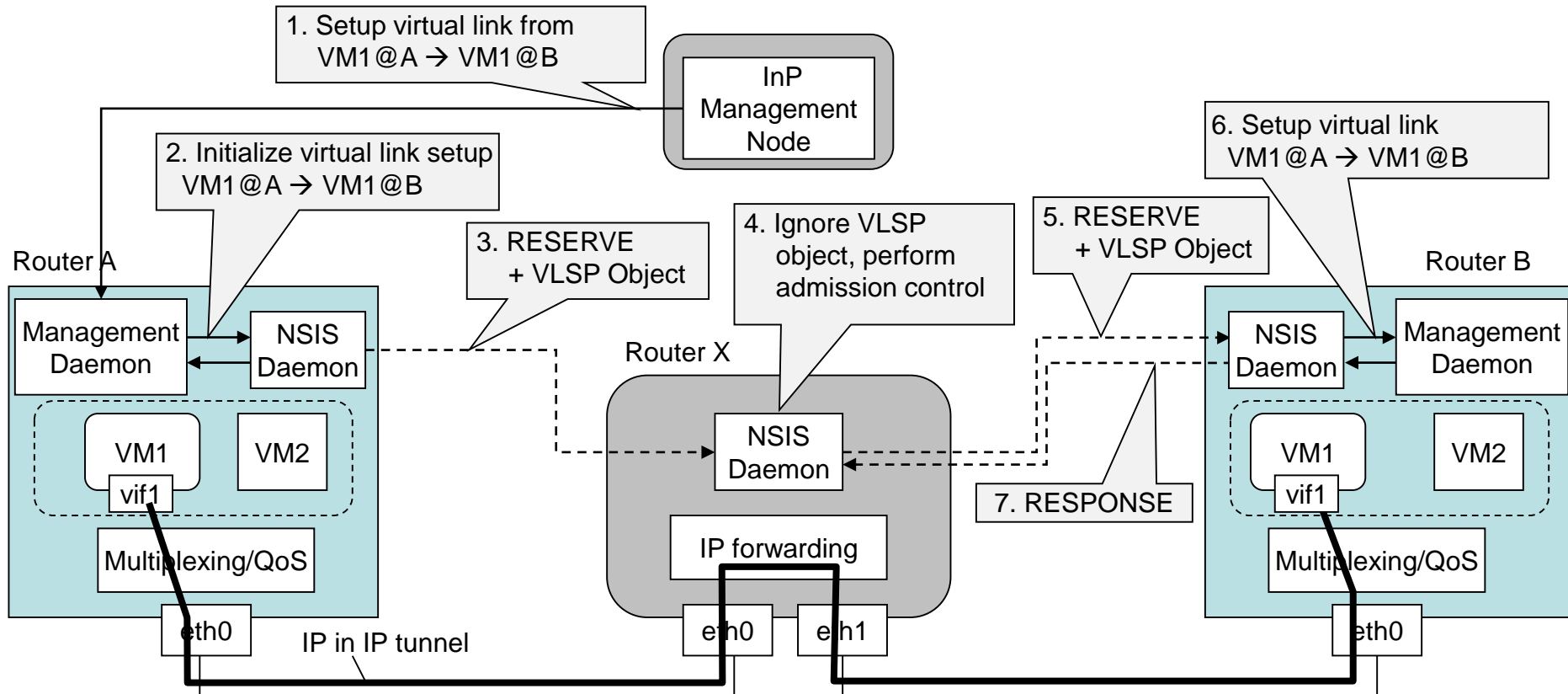




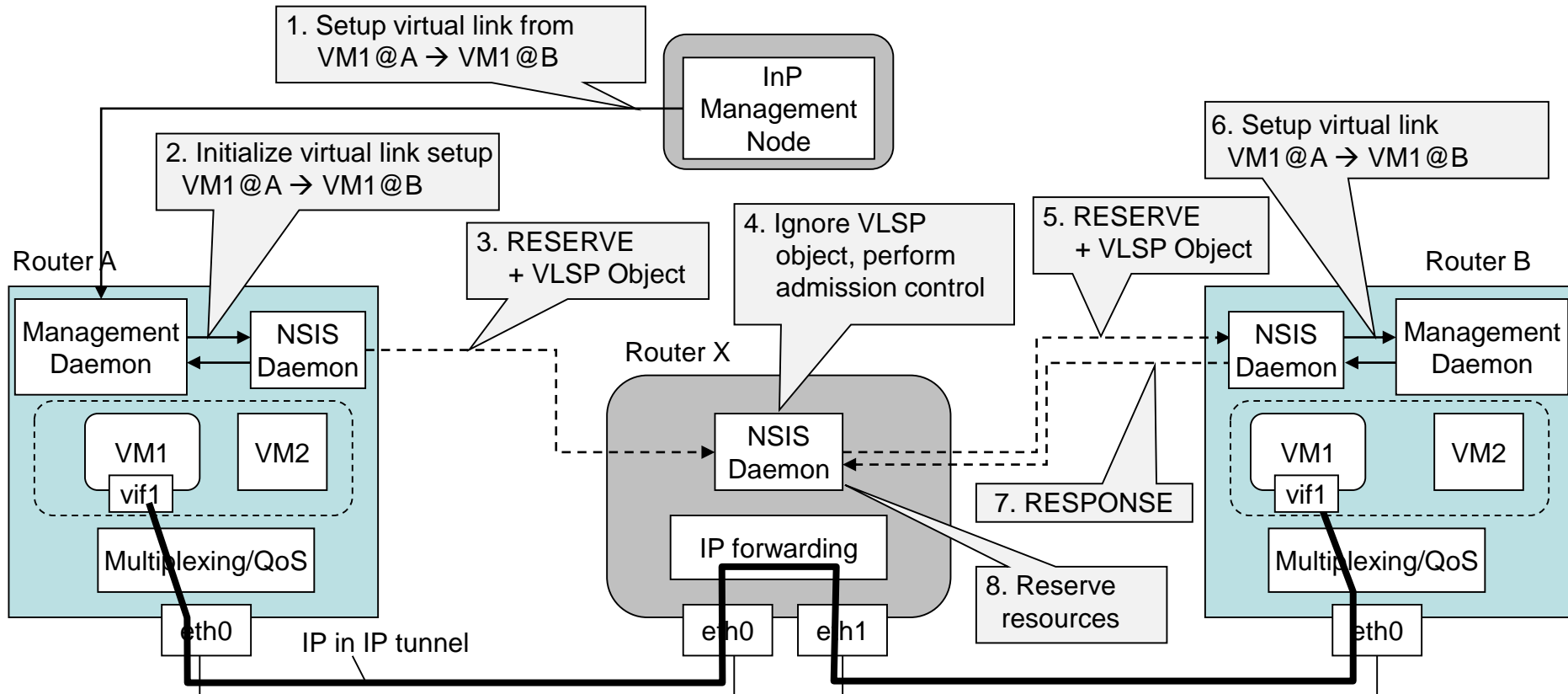
# Virtual Link Setup Workflow



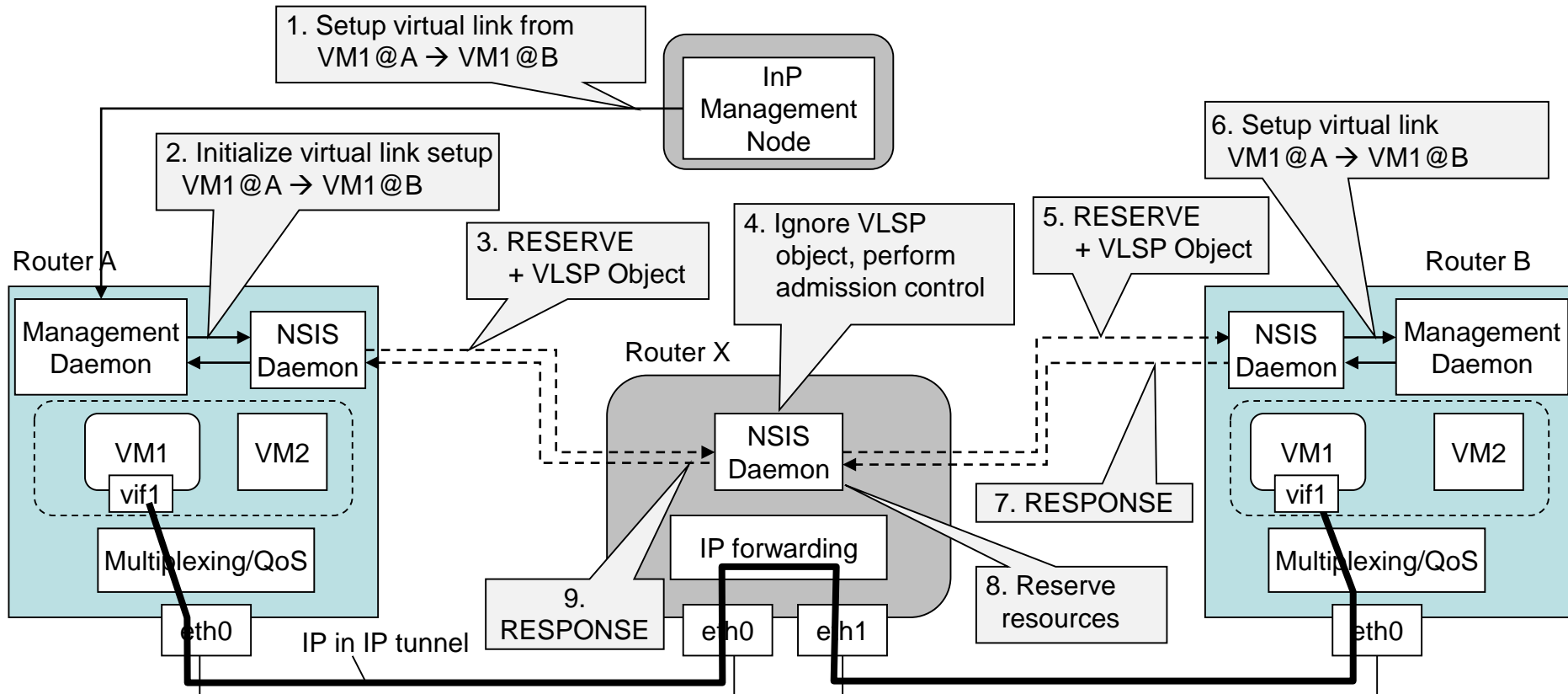
# Virtual Link Setup Workflow



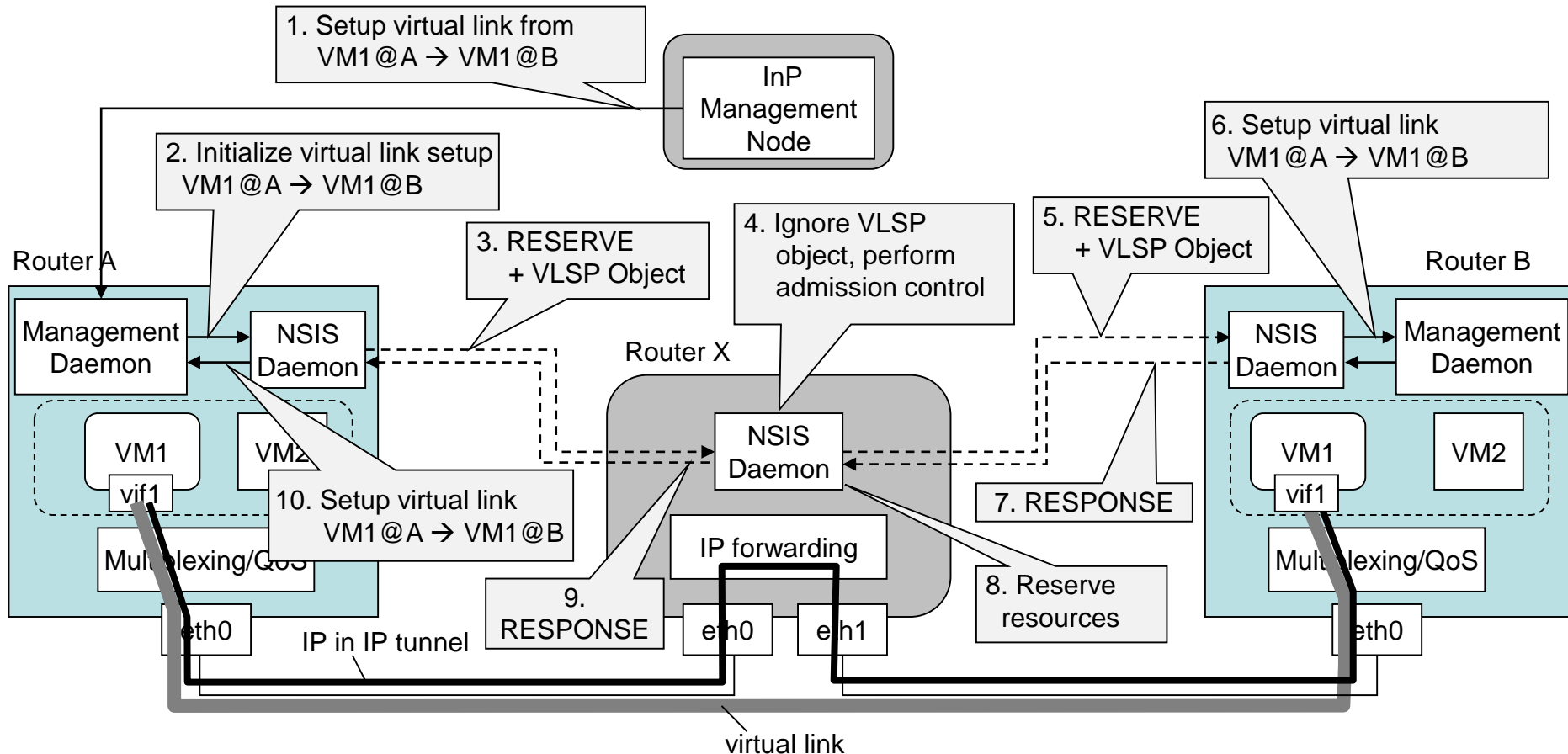
# Virtual Link Setup Workflow



# Virtual Link Setup Workflow



# Virtual Link Setup Workflow





- P. Barham, et al., **Xen and the Art of Network Virtualization**, ACM SOSP 2003
- D. Farinacci, et al., **Generic Routing Encapsulation (GRE)**, RFC 2784, IETF, 2000
- N. Egi, et al., **Towards High Performance Virtual Routers on Commodity Hardware**, ACM CoNEXT 2008
- Z. Bozakov and P. Papadimitriou, **OpenVRoute: An Open Architecture for High-Performance Programmable Virtual Routers**, IEEE HPSR 2013
- G. Schaffrath, et al., **Network Virtualization Architecture: Proposal and Initial Prototype**, ACM SIGCOMM VISA 2009
- D. Dietrich, et al., **Multi-Domain Virtual Network Embedding with Limited Information Disclosure**, IFIP Networking 2013
- P. Papadimitriou, et al., **Towards Large-Scale Network Virtualization**, IFIP WWIC 2012