NVO3: Network Virtualization

Thomas Narten
narten@us.ibm.com

IETF 82 – Taipei
November, 2011

Level Set: Purpose of Today's Session

- Begin process to initiate new work area for IETF
- Focus on the problems that are motivating this work area
- Outline a general framework (i.e., overlays) for solution direction
- Show that industry support for direction already exists
- What we do not want to do:
  - Discuss process, as in whether this is in scope for L2VPN or better done in another WG.
  - Spend months (longer?) surveying the entire solution space before selecting a solution direction

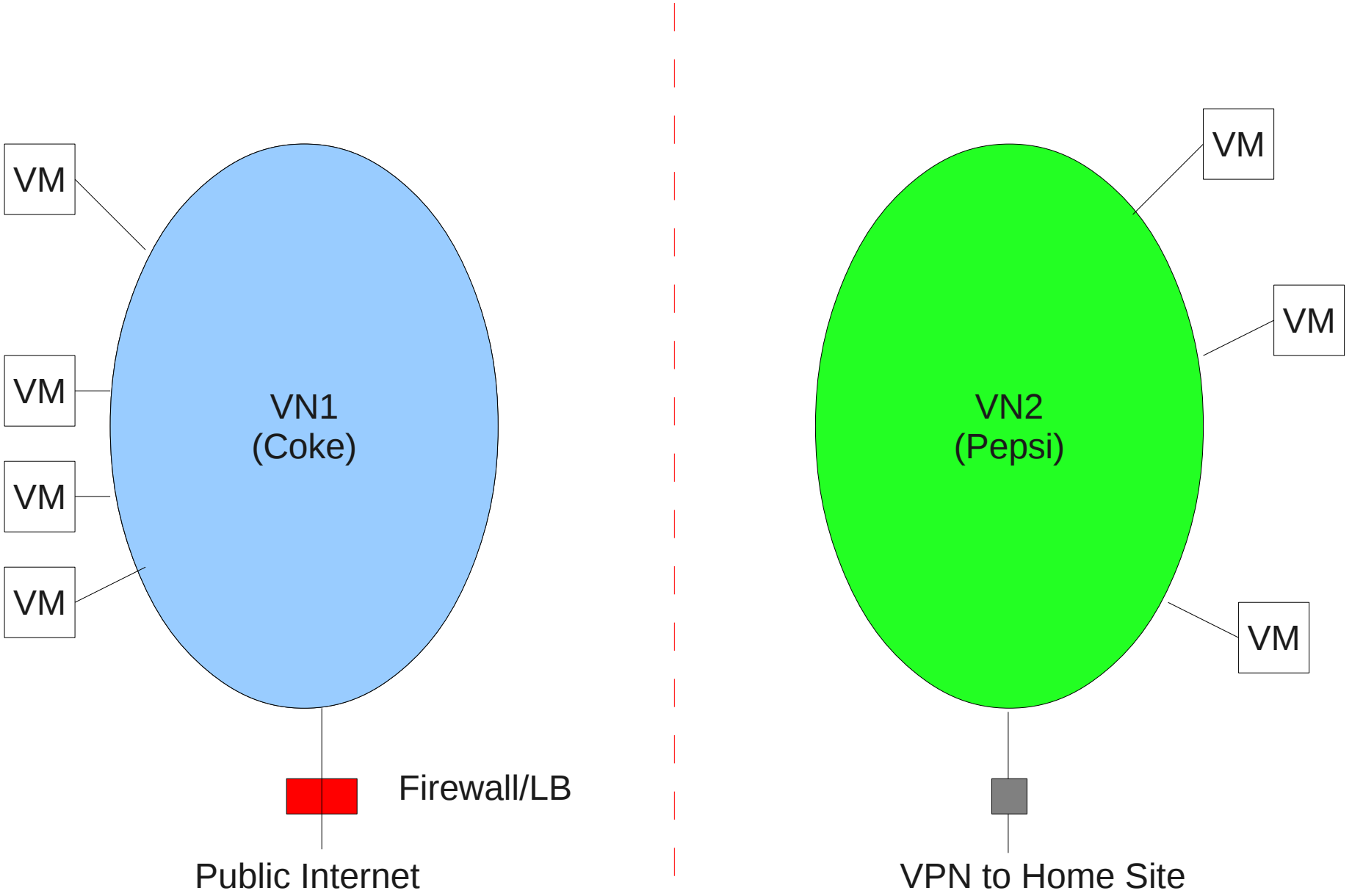
High-Level Motivation

- Imagine a data center
  - Could be cloud provider, hosting center, enterprise
  - Supports multiple tenants (e.g., Pepsi and Coke).
- Tenant wants (and operator wants to sell) ability to:
  - Create a Virtual Network instance
  - Create set of VMs that logically attach to the Virtual Network
  - Network as a Service
- The Virtual Network (with associated VMs) provides a distributed service
  - E.g., web hosting, email service, etc.
  - Or uses VPN to extend back into enterprise network

VN Requirements (Tenant Perspective)

- VMs think they are connected to a "real" network
  - Send/receive Ethernet frames
- Each VN instance uses its own address space
  - Tenant uses whatever addresses it wants (e.g., private addresses)
- VNs are fully isolated from each other (security)
  - One tenant's traffic can't be seen by another tenant
  - Packets stay local to a VN
  - Traffic enters/exits a VN only through controlled entry point
    - Could be a connection to Public Internet
    - Could be a VPN connection back to tenant's home site
    - Could have firewall, ACLs, etc.

Logical View (Tenant)



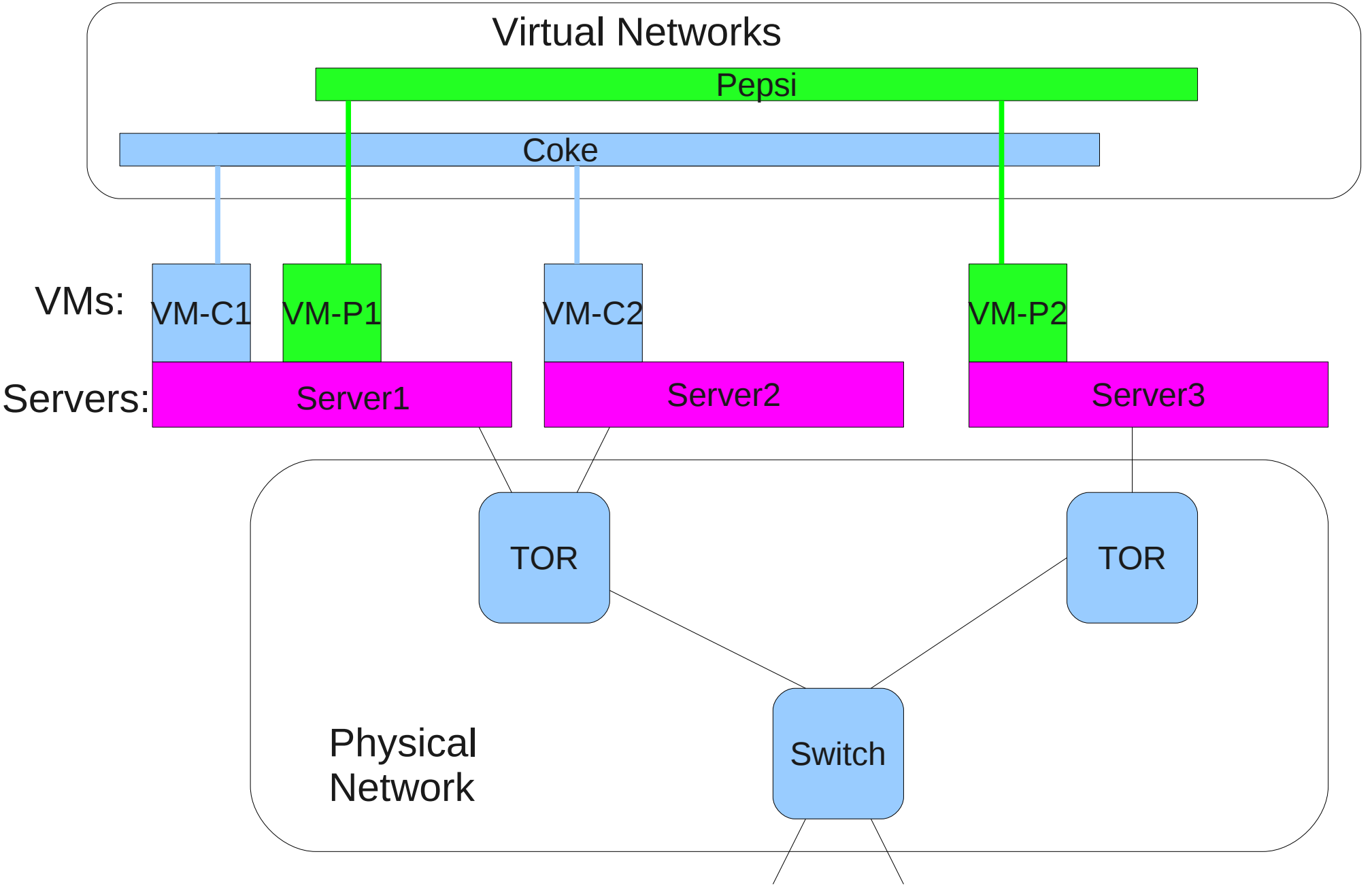
VN Requirements (DC Perspective)

- Want ability to place VMs anywhere in the data center
  - Without being constrained by physical network attributes or concerns (e.g., IP subnet boundaries)
  - Both initial placement and for VM migration
- Reality: L2 VLANs & broadcast domains no longer sufficiently scalable
    - TRILL, SPB, etc. working on this, but no magic bullet
  - Problem today for larger data centers (ARMD work)
  - Will only get worse in future as DCs grow
- Note: Above two are in conflict with each other
  - Can't move a VM (today) to a "different" IP subnet

Requirements (DC Perspective) – Cont

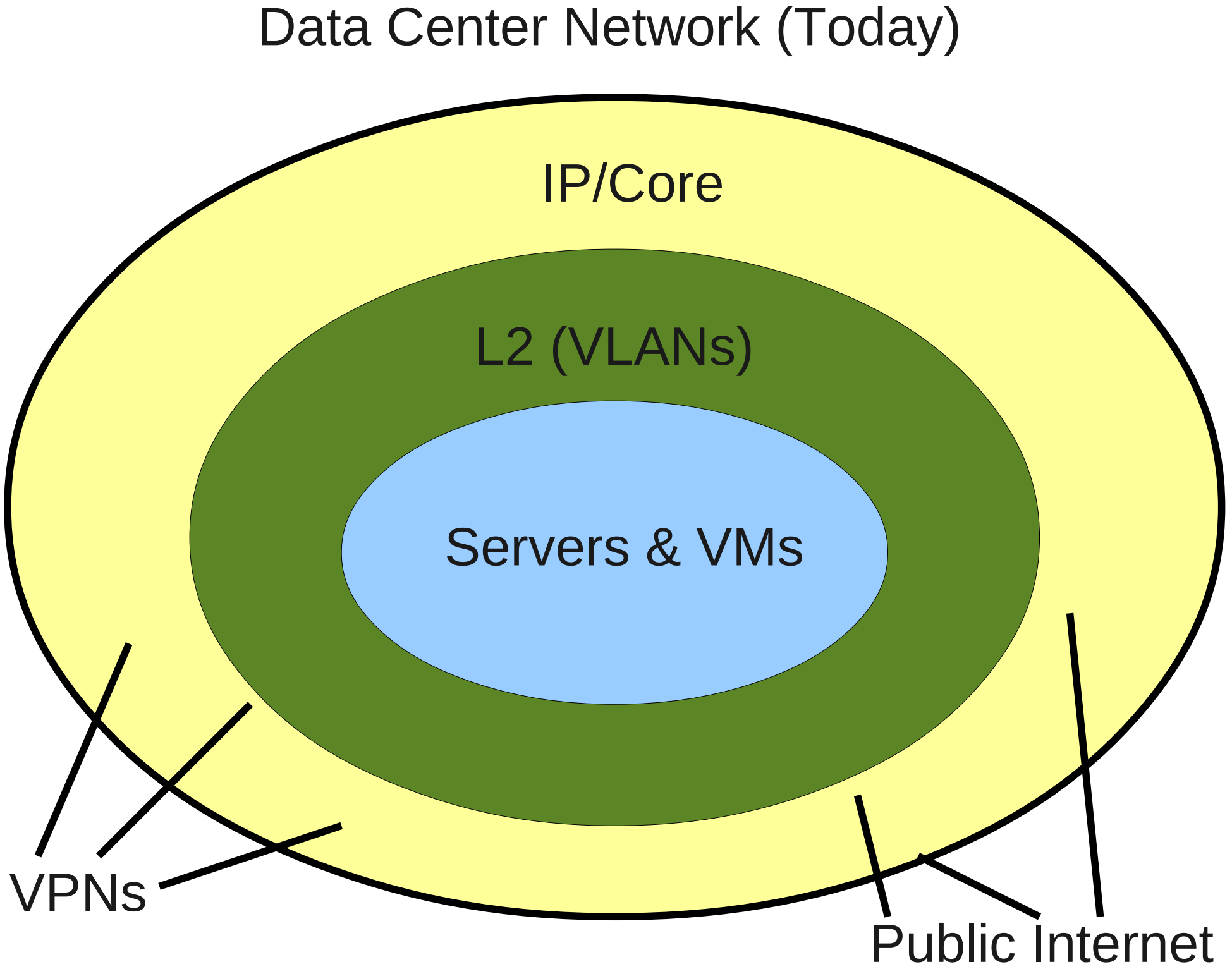
- Want to separate the logical network attributes associated with VM from the physical instantiation
  - e.g, VLAN info, QoS, L2 protocols, IP Subnets, etc.
  - Observation: reconfiguring the network elements when placing VMs is complex, error prone
- Want to abstract away the key network properties
  - Server virtualization allows VMs to abstract away physical properties for memory, processor, I/O, etc.
  - Network properties include VLANs, IP Subnetting, etc.
- Solution needs to scale to cover entire data center (and beyond)
  - Millions of VMs (and beyond)

Physical & Logical View



Summary of Requirements

- Multi-tenant support
- Support VM placement anywhere in data center
  - Both initial placement & migration
- On-demand elastic provisioning of resources
  - Grow/shrink dynamically as workload changes
  - Allow for “stretching” of virtual network
- Small forwarding tables in switches
  - Return to model where switches only know MAC addresses of physical switches
- Decouple logical/physical network configuration
- Scale to millions of VMs (and beyond)


Data Center Network (Today)
IP/Core
L2 (VLANs)
Servers & VMs
VPNs
Public Internet

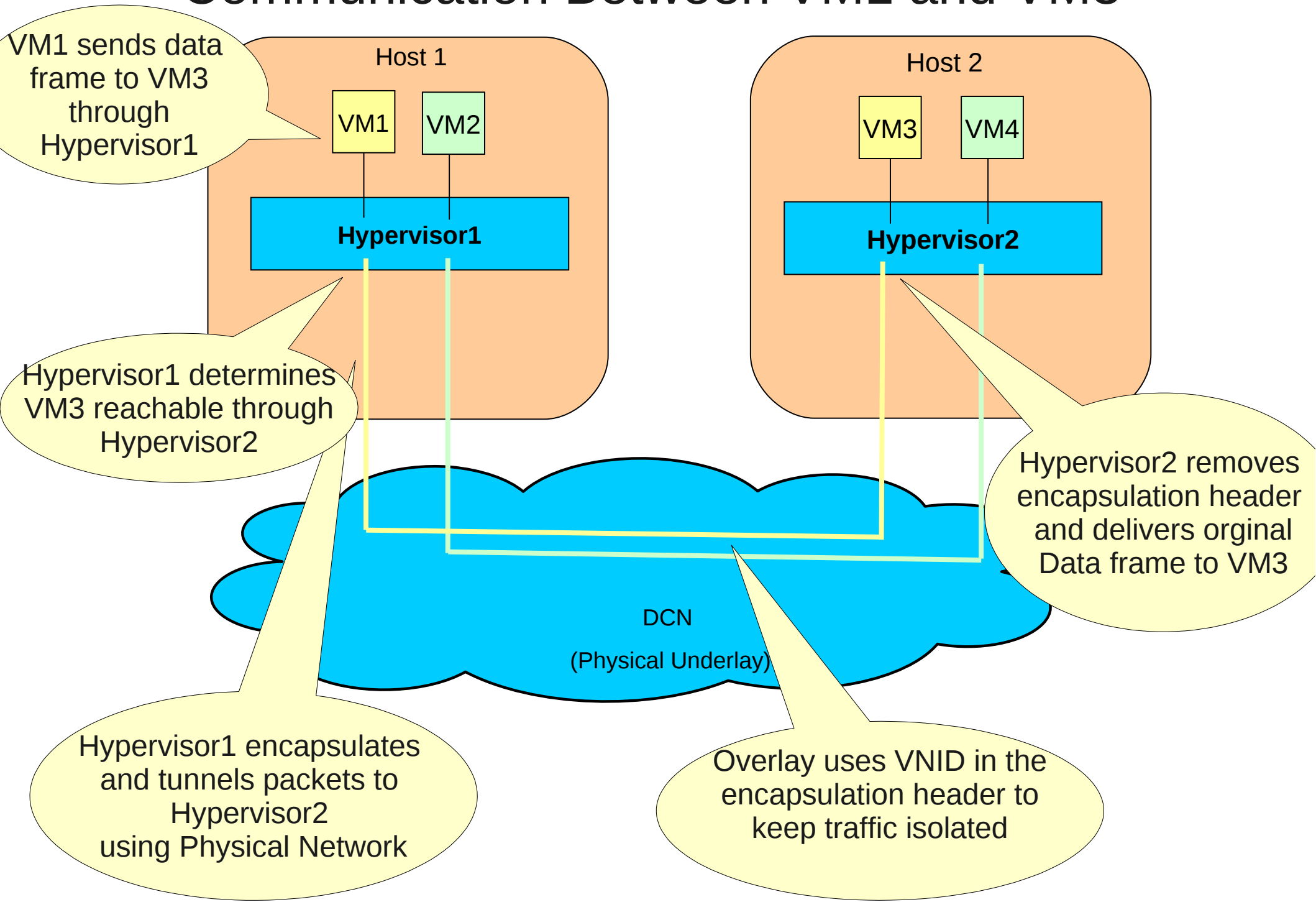
Focus of NVO3

- NVO3 starting point is green part of previous slide
  - Not motivated by VPNs coming into the DC
    - but will (of course) connect to such VPNs
  - Motivated by need for better multi tenancy across entire DCN
- Need a better alternative to (current) L2 VLANs
- Spans entire DCN and even into remote DCNs
- Support highly dynamic changes to VN span as VMs are moved around
  - E.g., responding to highly-dynamic workloads in real time

Overlay Approach

- Layer a virtual network over the infrastructure network
- Use an overlay or “shim” header for encapsulation
  - Overlay header carries a Virtual Network Identifier (VNID)
    - VNID identifies a specific VN instance
    - Analogous to VLAN ID, VPLS Instance, etc.
    - Needs to be “large enough” (e.g., 24 bits)
  - Also encapsulates original packet from VM as data
- Tunnel packet from source to destination
  - Encap/decap done by edge switch or hypervisor
  - VM itself unaware tunneling is taking place

Communication Between VM1 and VM3



IETF Work Area

- Although an overlay/encapsulation header is needed,
  - Exact header details not important (but must meet requirements)
  - Multiple encapsulations not necessarily a problem
  - Existing, already defined encapsulations may suffice
  - Attempt to pick The One encapsulation likely unproductive
- Control plane is where things get interesting
  - This is where IETF can provide value

Control Plane Tasks

- Need mechanism to populate mapping tables used when encapsulating
  - Need to know where to tunnel packet to
- Need mechanism for delivering multi destination frames within a VN instance
  - For implementing tenant broadcast or multicast
- Need registration mechanism for endpoint to inform switch:
  - When it is attaching to a particular VN instance
  - When it is detaching from the network (and VN instance)
- Registration mechanism must include updating of stale information in switches

Address Mapping: Learning Approach

- Reuse control plane from IEEE 802.1 bridging approach
- Build mapping tables by examining inner & outer source addresses of received packets
- Packets to unknown unicast destinations flooded within VN
  - IP multicast group address (from DCN) associated with VN instance
  - Packets sent to DCN multicast group delivered to all endpoints on VN
  - Tenant broadcast/multicast handled in same way (sent to DCN multicast address)
- Simple, well understood, but also inherits known limitations

Address Mapping: Directory Based Service

- Use “centralized” directory service to store address mappings
  - Edge devices query directory service to obtain mappings
- Need to update directory service when:
  - Instantiating a VM
  - Migrating a VM (replace old binding with new)
- Need way to invalidate old cached information in edge switches when directory is updated
- “Centralized” is misnomer – need replication/backup
- Need to develop requirements, select an approach
  - Engineering work, not rocket science

The Big Picture

- A number of things needed to realize overall solution
  - NVO3 is NOT proposing to do them all
  - Some aspects do not even have a standards component and are (necessarily) proprietary
- Orchestration system handles VM placement
  - When instantiating a VM on a specific server,need mechanism for registering attachment of VM to a particular VN instance
  - When moving a VM,need mechanism to deregister attachment of VM from network at VM's previous point of attachment
- Orchestration piece is not IETF work!!!
  - But some pieces used by orchestration system are needed

Industry Status

- Already two existing proposals for implementing overlays:
  - VXLAN (draft-mahalingam-dutt-dcops-vxlan)
  - NVGRE (draft-sridharan-virtualization-nvgre)
- Significant vendor backing behind the efforts
  - If IETF does not engage, work will happen outside of IETF
- Short window of opportunity for IETF to become home for this effort
  - IETF is the obvious place for pursuing this work
  - IETF either engages or becomes irrelevant on this topic

Related Work

- TRILL is an L2 technology
  - Complementary to IP based overlay approach
  - Demand exists for an IP-based approach
- IEEE Shortest-Path Bridging (SPB)
  - Complementary and L2 based
- ARMD not chartered to do protocol work

L2VPN

- L2VPN comes from a strong service provider perspective
- NVO3 driven by DCN operator perspective
  - This difference is fundamental
- L2VPN is fundamentally about using SPs to stitch together L2 networks across a WAN
  - Nothing wrong with that
  - But, NVO3 is about multi-tenancy within the DCN, independent of an external VPN provider

L2VPN (continued)

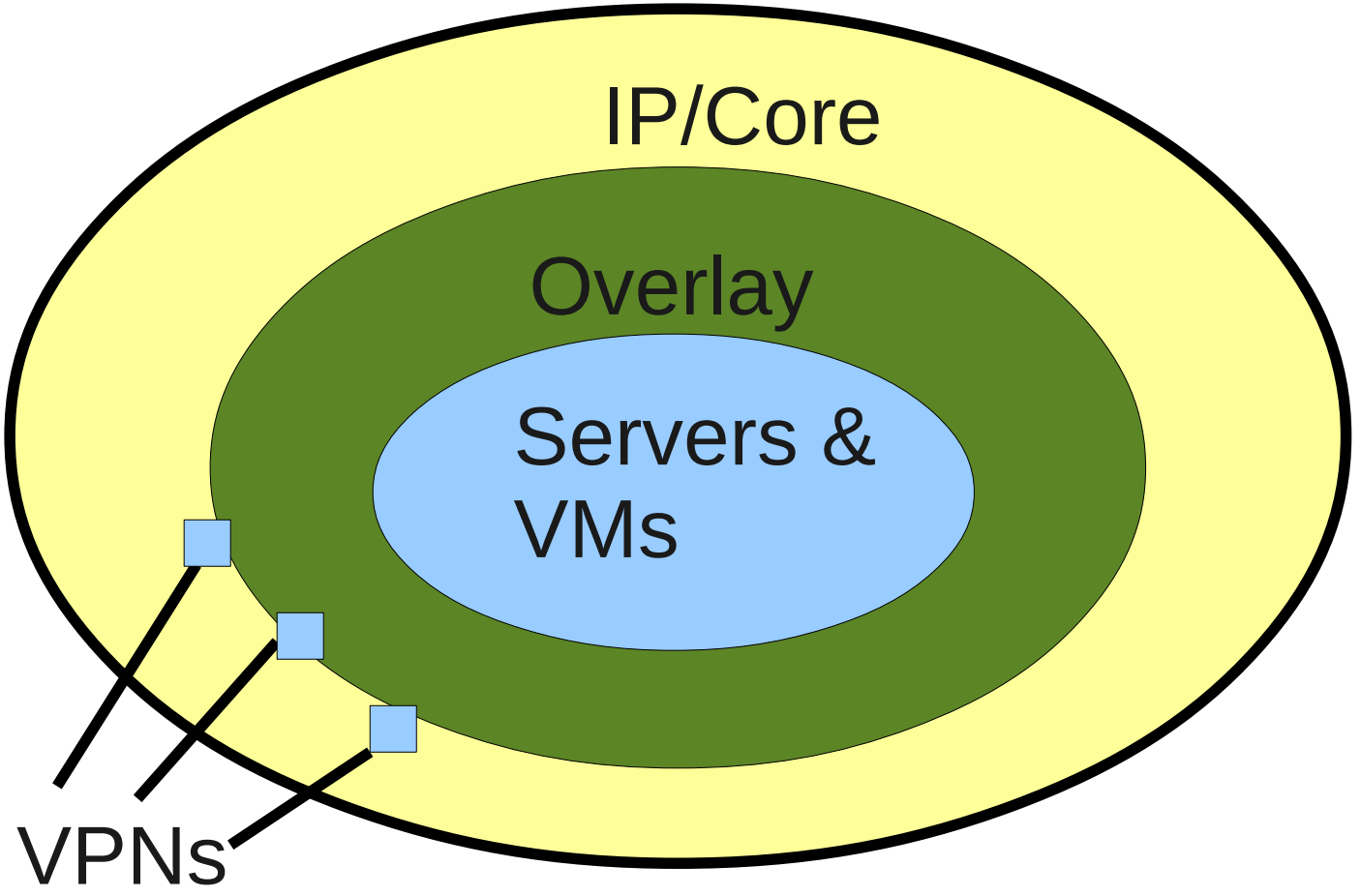
- L2VPN approach (e.g., EVPN) is about:
  - Improved scaling of L2 within DCN (e.g., SPB) and across WAN
  - Pushing L2VPN “edge” deeper into DCN
- This is one approach, but not the only approach
- NVO3 is about providing for multi-tenancy at a higher level (e.g., IP) independent of the underlying L2 technology
  - NVO3 interfaces with L2VPN protocols at boundary between the two

L2VPN vs. L3VPN

- The idea that NVO3 is strictly carrying L2 frames over L3 is overly simplistic
- Mantra: Carry L3 when we can, L2 when we must
- 90% of the control plane issues are layer agnostic
  - Could be used by (say) TRILL to provide directory-assisted mappings
- Would be beneficial if a single control plane framework/architecture could be reused in different contexts

NVO3 and VPNs

- Virtual Networks will need to connect to VPNs, but that is secondary.
- Straightforward to connect VN to a VPN



Summary Points

- NVO3 driven from DC by intra-DC problems
- Needs to span across DCs, but that is secondary
- The cost/benefit of overlays is extremely compelling to DCN operators (works with existing equipment)
  - In fully virtualized systems, can be implemented entirely in hypervisor software
  - In traditional DCNs, edge switches need enablement
  - TRILL, SPB, etc. have a dfferent deployment path
- Major vendors are already committed to moving in the overlay direction
- We have a short window in which IETF either engages, or becomes irrelevant on this topic

Acknowledgments

This effort stems from the work of many others...

See list of authors in problem statement and in the NVGRE and VXLAN documents.

Backup

Background & Definitions

- Term “switch” and “hypervisor” used interchangably when talking about encap/decap
  - Both will be tunnel endpoints, when one term is used, assume other is implied as appropriate
  - Switches will implement functionality in order to support service to non-virtualized servers
- Term VM used throughout, assume non-virtualized server is also intended as appropriate
- DC – Data Center
- DCN – Data Center Network
- VN – Virtual Network (as presented here)