MPLS-Based Hierarchical SDN for Hyper-Scale DC/Cloud
draft-fang-mpls-hsdn-for-hsdc-00

Luyuan Fang lufang@microsoft.com
Vijay Gill vgill@microsoft.com
Fabio Chiussi fabiochiussi@gmail.com

IETF 91, MPLS WG

November 14, 2014

Underlay Network Scalability Challenges

- Scale at low-cost, use commodity HW
  - Use small FIBs/LFIBs in all network nodes, avoid FIB explosion
- Achieve high resource utilization
  - Efficiently support ECMP and any-to-any, server-to-server TE
- Scale at low operational and computational complexity
  - Locally minimize complexity and network state, with no information loss
- Scale while achieving improved cloud elasticity and service velocity
  - Overcome today's challenges of NFV (e.g. SLB) scalability and VM/NFV mobility

MPLS-Based HSDN Design Requirements

- MUST support millions to tens of millions of underlay network endpoints in the DC/DCI.
- MUST use very small LFIB sizes (e.g., 16K or 32K LFIB entries) in all network nodes.
- MUST support both ECMP and any-to-any, end-to-end, server-to server TE traffic.
- MUST support ECMP traffic load balancing using a single forwarding entry in the LFIBs per ECMP group.
- MUST require IP lookup only at the network edges (e.g., server in DC or edge server in core).
- MUST support encapsulation of overlay network traffic, and support any network virtualization overlay technology.
- MUST support control plane using both SDN controller approach, and the traditional distributed control plane approach using any label distribution protocols.

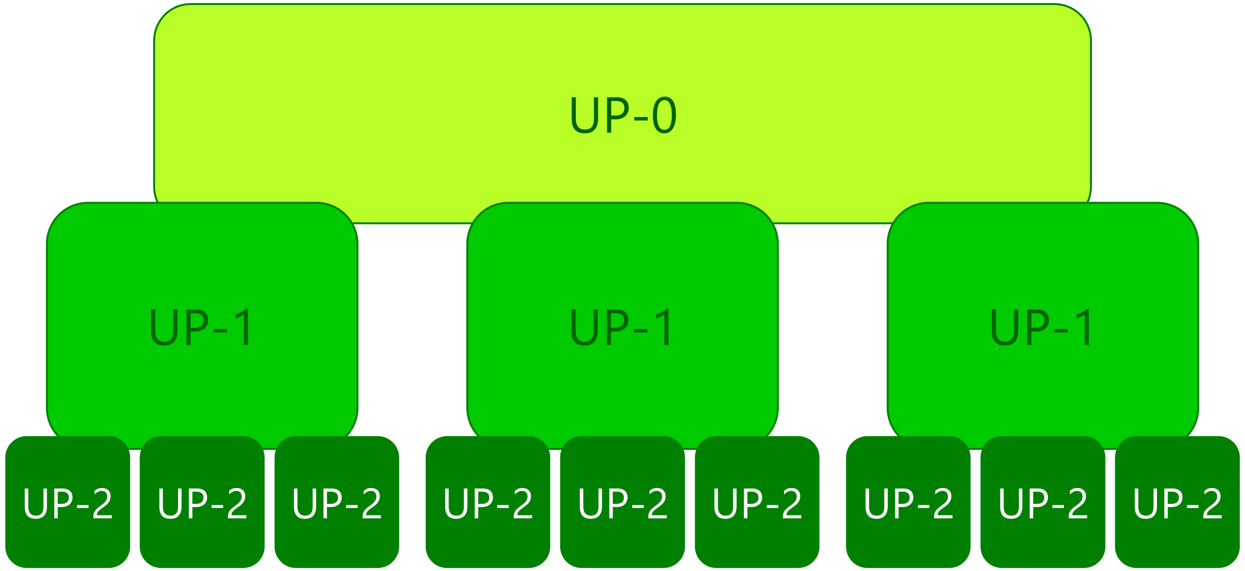
Choice of Technologies: MPLS forwarding + SDN control

- MPLS
  - Unify forwarding (DC and core), no IP lookup other than at the edge/server
  - Flexibility of the label stack, naturally suitable for hierarchical decomposition
  - Ease of redirection, can be leveraged to increase elasticity
  - Security advantages
- SDN
  - Allow decoupling of control plane and data plane, make HW fungible
  - Take ownership of control plane development (short release cycle for bug-fixes and new features)
  - Reduce number of protocols
  - Make global optimization possible

HSDN – One Fundamental Abstraction for Both Forwarding and Control

Forwarding Plane

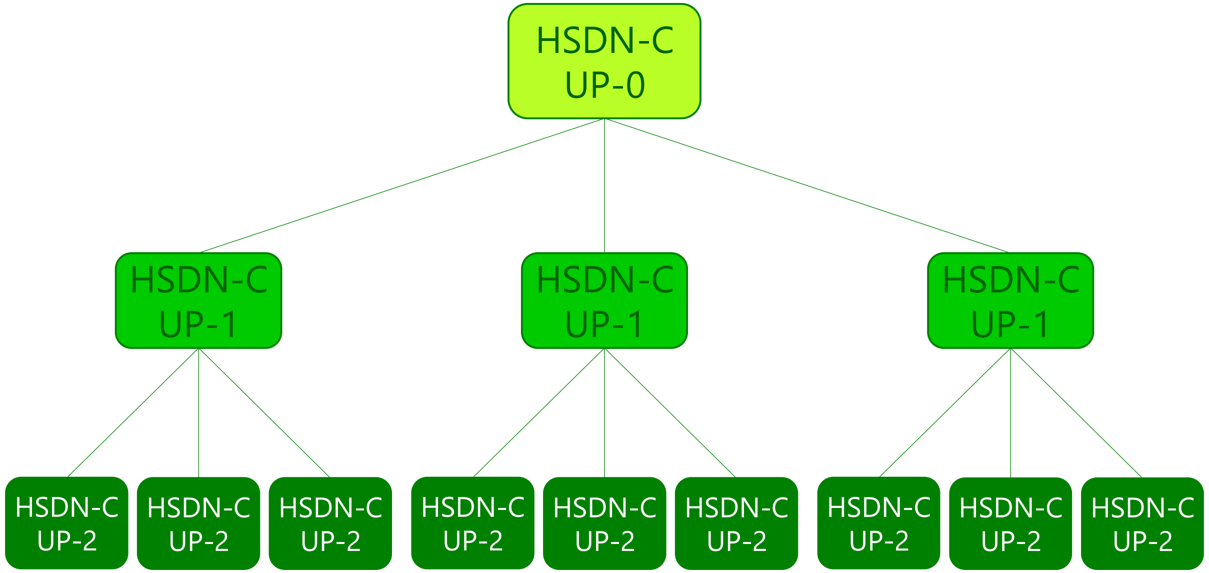
HIERARCHICAL UNDERLAY PARTITION (UP)



- **Keep number of destinations in all domains small**
- **Locally, hide destination explosion using hierarchical partitioning**

Control Plane

HIERARCHICAL CONTROL



- **Keep number of paths per domain manageable**
- **Keep computational complexity per domain small**

One Consistent Abstraction Paradigm

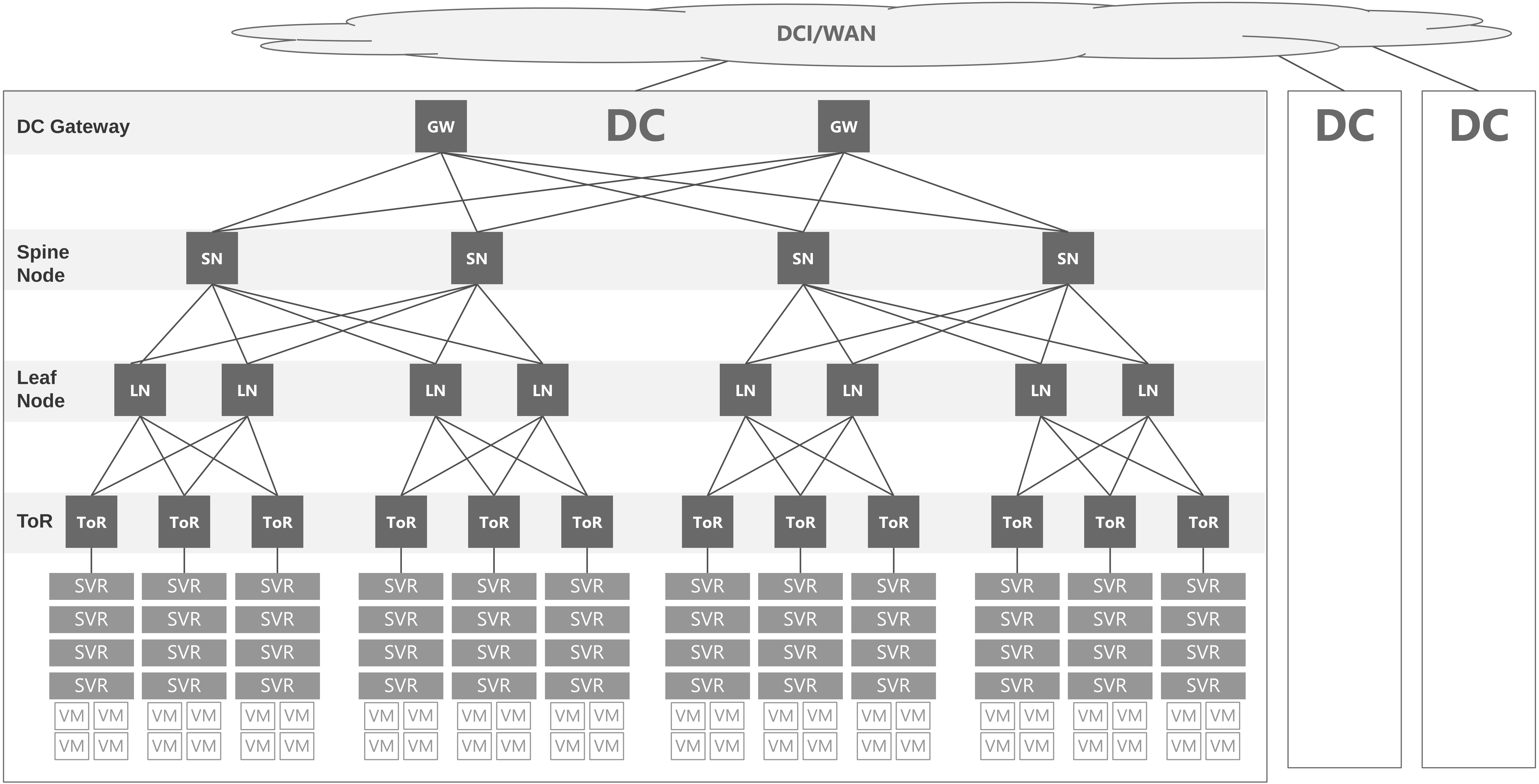
- **Divide and conquer**
- **Keep all domains balanced and small**
- **Locally minimize network state**

→ “Infinite” Horizontal Scaling

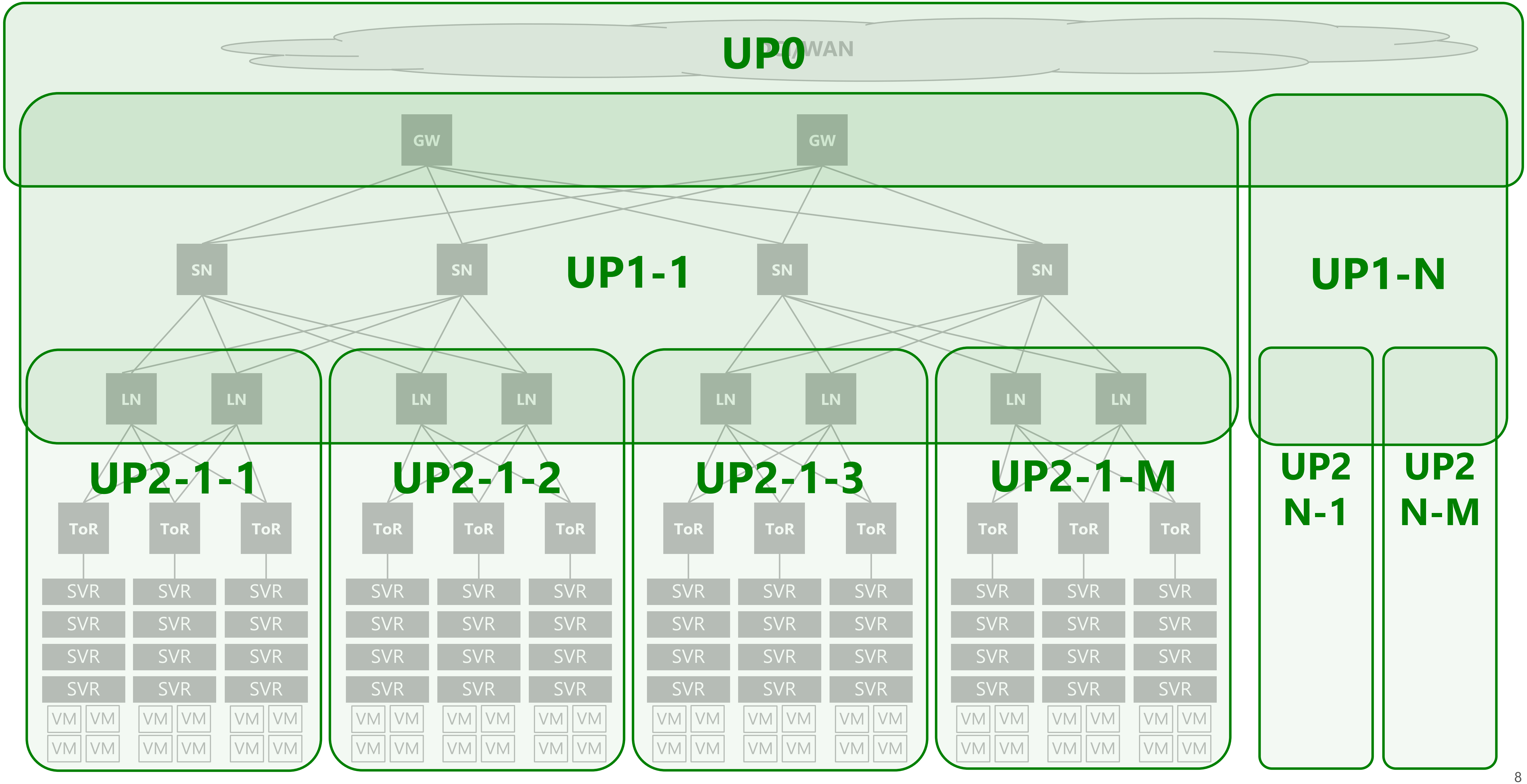
HSDN Forwarding Plane

- Divide the DC and DCI/WAN in a hierarchically-partitioned structure
- Assign groups of Underlay Partition Border Nodes (UPBNs) in charge of forwarding within each partition
- Construct HSDN MPLS label stacks to identify the end points according to the HSDN structure
- Forward using the HSDN MPLS labels

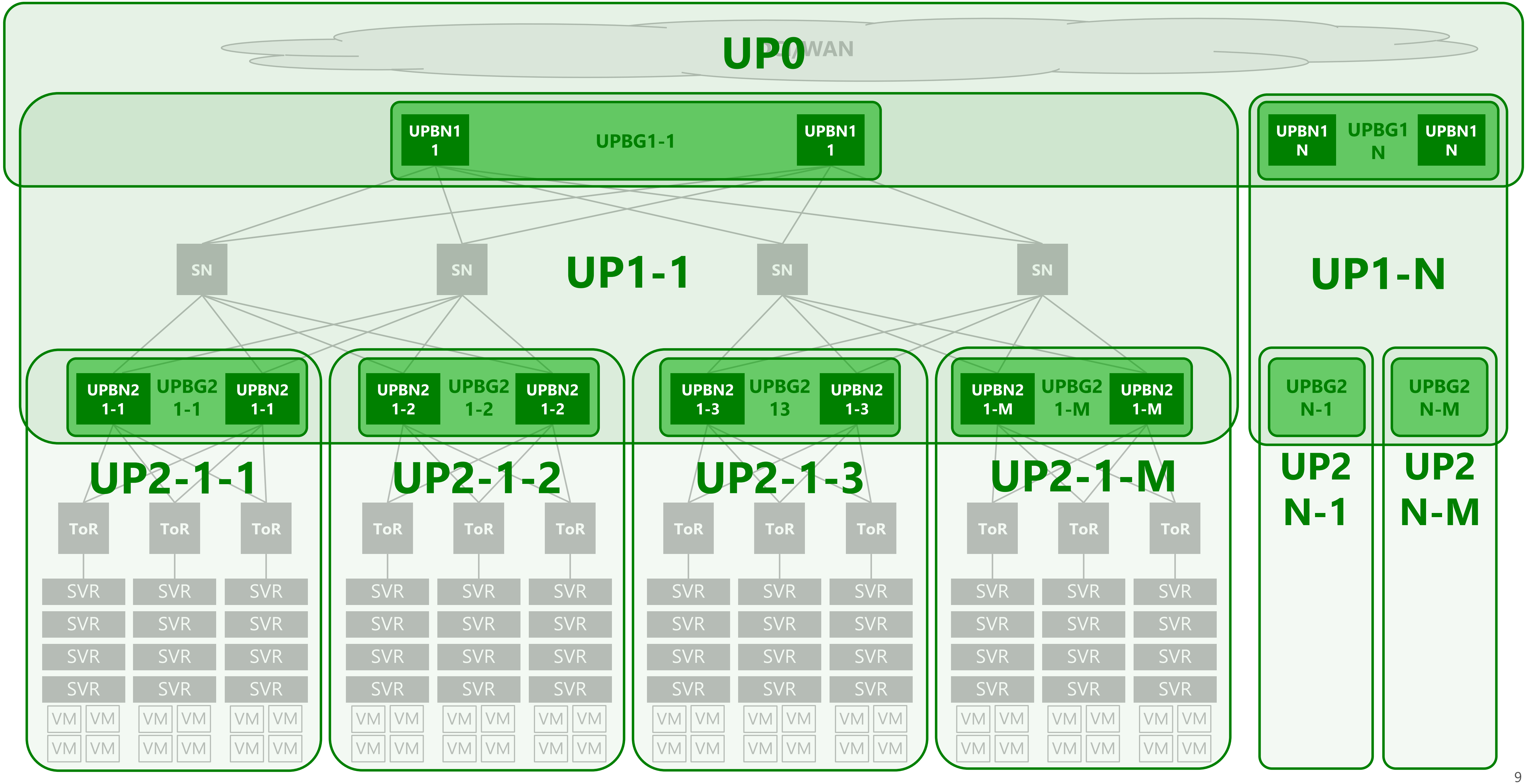
Typical Clos-Based DC Topology, Spine and Leaf Architecture



HSDN: Hierarchical Underlay Partitioning

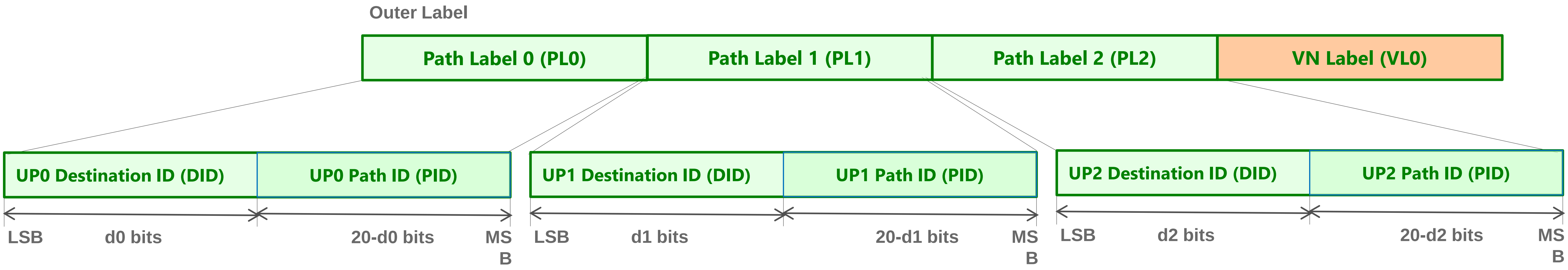


HSDN: Assign UPBNs and UPBGs



HSDN Label Stack

- Stack of path labels, plus one VN label
- One path label per level of underlay partition



HSDN Forwarding: Life of a Packet

HSDN Label Stack
3 Path Labels

PL0	PL1	PL2	VL

PL0

Identifies destination UPBN1 or UPBG1

PL1

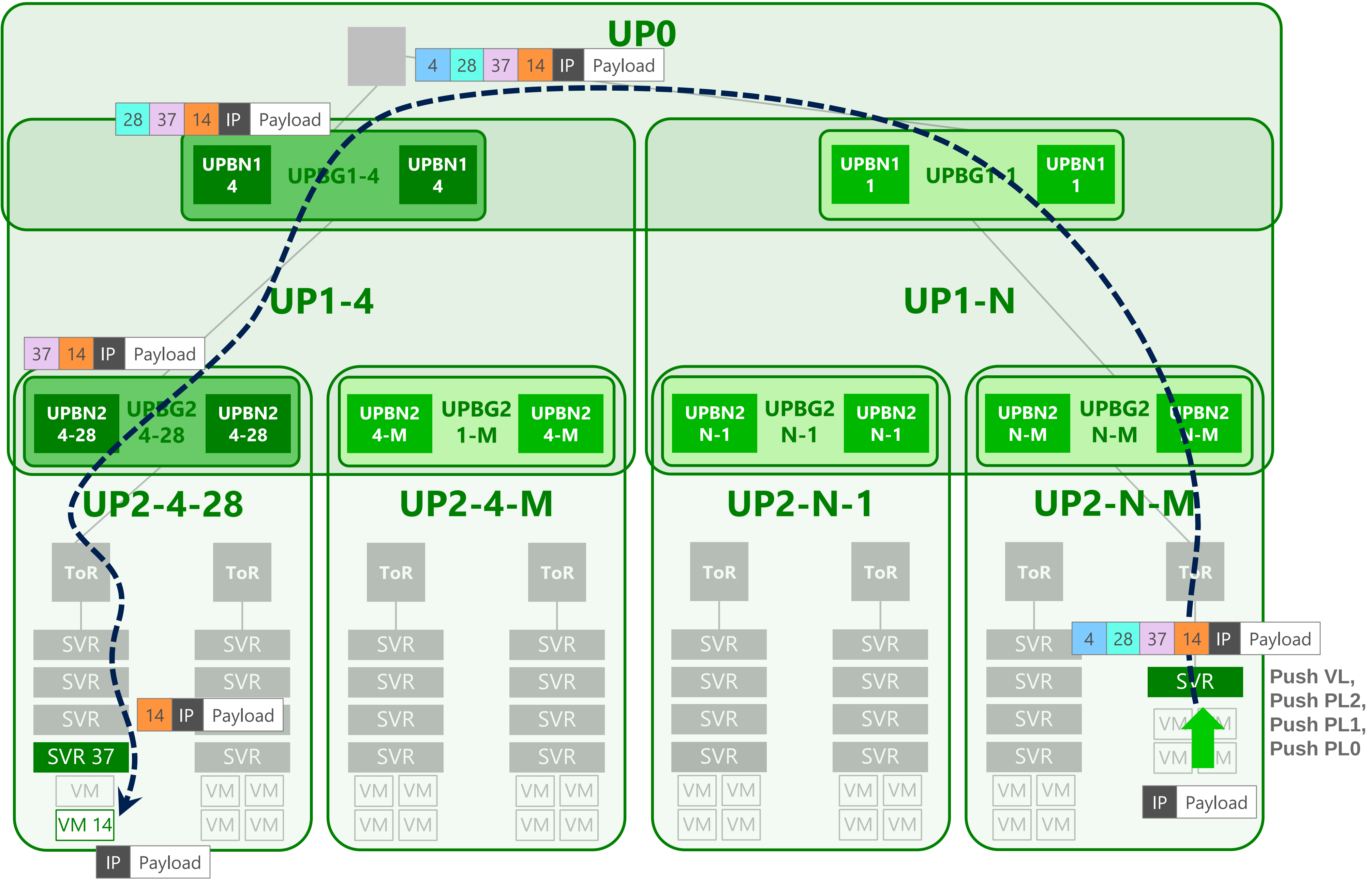
Identifies destination UPBN2 or UPBG2 within UP1

PL2

Identifies destination server within UP2

VL

Identifies VN



HSDN Control Plane

- HSDN Controller (HSDN-C) is horizontally scalable
  - Implemented as a set of local partition controllers HSDN-C-UP, following the HSDN hierarchy
  - Each HSDN-C-UP operates largely independently
  - Locally-reduced computational complexity for many functions, including TE
- Network state also distributed according to the HSDN hierarchy
  - Forwarding state is still in the network nodes, and is locally minimized
- HSDN supports both controller-centric SDN approach and traditional distributed routing/label distribution protocol approach
  - Useful during migration from legacy to full SDN (e.g., use BGP-LU for label distribution, RFC 3107)

HSDN Scaling Examples

HSDN scales to tens of millions of underlay network endpoints with small LFIBs

- Assumptions
  - N hyper-scale DCs interconnected through DCI/WAN
  - DC fabrics are S-stage, asymmetrical, fat-Clos-based
- Support any-to-any, server-to-server
  - non-TE traffic with ECMP load balancing
  - TE traffic
- Max LFIB size (the largest LFIB size among all Tiers of switches) is as follows:

Number of Server endpoints	Max LFIB size ECMP only (No TE)	Max LFIB size ECMP and TE Concurrently
3 M	~ 1K	< 14K
10 M	< 2K	< 24K
40 M	< 3K	< 36K

Next Steps

- Collect feedbacks from WG
- Update the draft based on comments and new developments
- Ask for WG adoption