MPLS-Based Hierarchical SDN for Hyper-Scale DC/Cloud
draft-fang-mpls-hsdn-for-hsdc-02

Luyuan Fang	lufang@microsoft.com
Vijay Gill	vgill@microsoft.com
Deppak Bansal	dbansal@microsoft.com
Fabio Chiussi	fchiussi@cisco.com
Chandra Ramachandran	csekar@juniper.net
Shahram Davari	davari@broadcom.com
Linda Dunbar	linda.dunbar@huawei.com
Barak Gafni	gbarak@mellanox.com
Andrew Qu	andrew.qu@mediatek.com
Jeff Tantsura	jeff.tantsura@ericsson.com
Yakov Rekhter	
Ebben Aries	exa@fb.com
Daniel Voyer	daniel.voyer@bell.ca
Wen Wang	wen.wang@centurylink.com
Himanshu Shah	hshah@ciena.com
Ramki Krishnan	

Changes from 00 version

- Added thirteen new co-authors on the draft
- Lots of feedback and discussion
- Streamlined the description of the architecture and basic forwarding
- Expanded and clarified label stack semantics for both ECMP and TE
- Added section on TE
- Added reference to new companion draft on “hybrid approach” for HSDN control plane using BGP-LU to distribute labels
  - draft-fang-idr-bgplu-for-hsdn-00

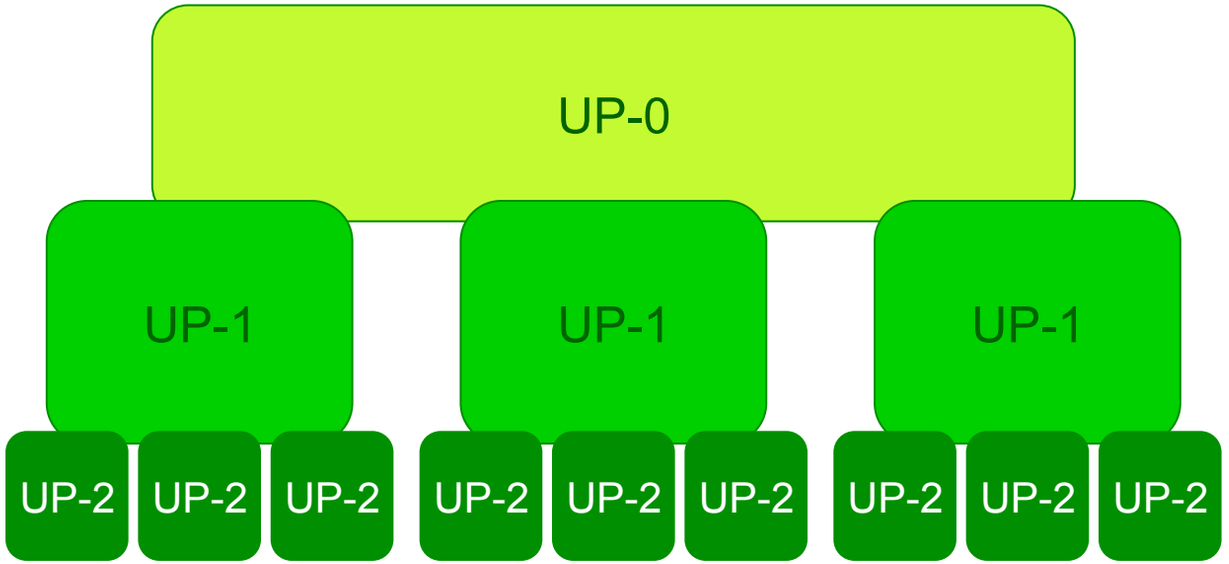
Refresh: MPLS-Based HSDN Design Requirements

- 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.

HSDN – One Fundamental Abstraction for Both Forwarding and Control

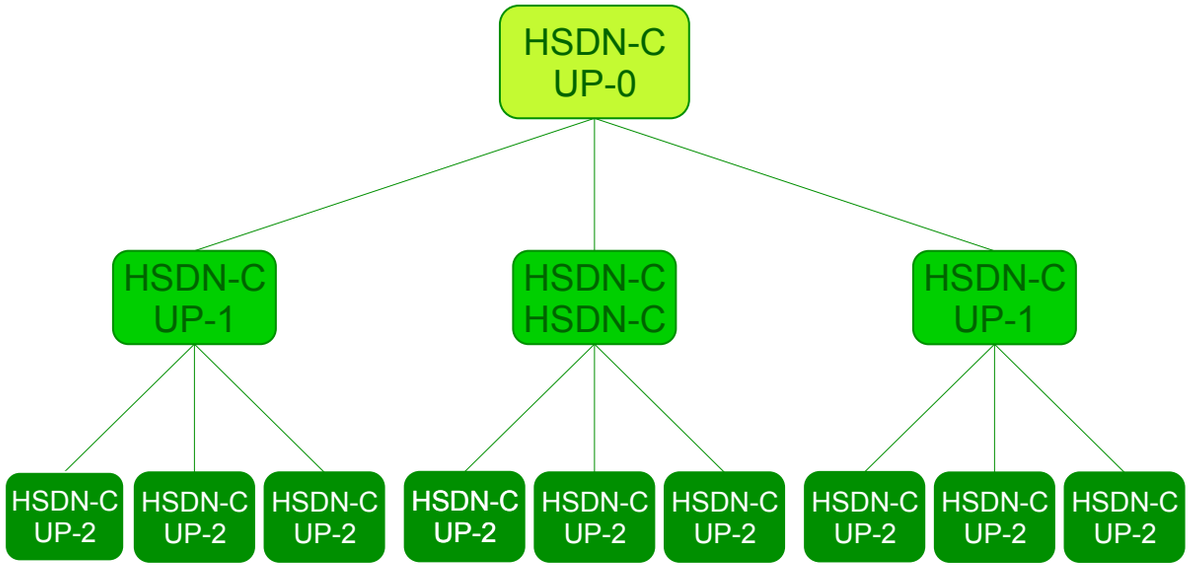
Forwarding Plane

HIERARCHICAL UNDERLAY PARTITION (UP)



Control Plane

HIERARCHICAL CONTROL



One Consistent Abstraction Paradigm

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

→ “Infinite” Horizontal Scaling

HSDN Achieves Hyper Scale

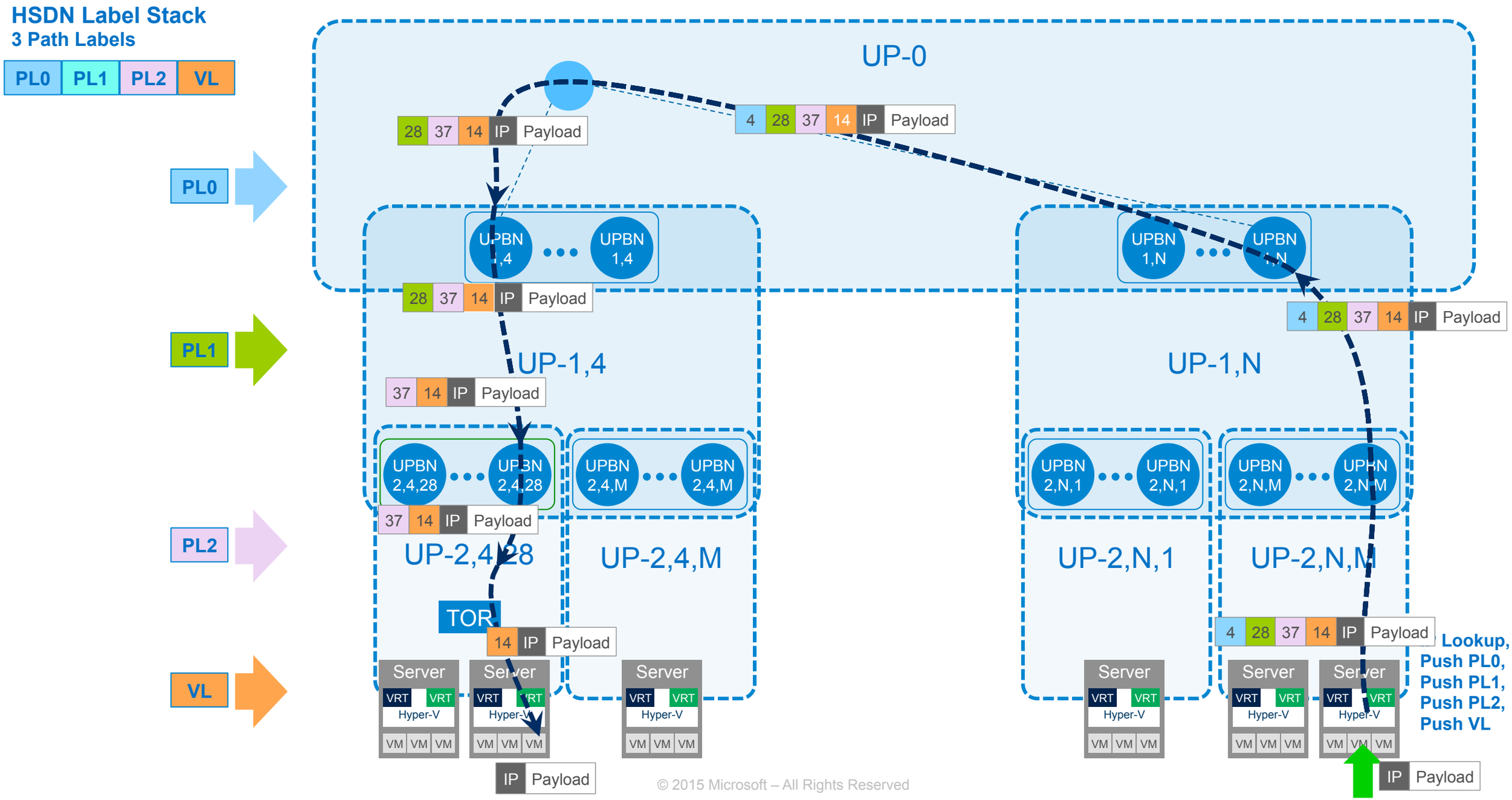
HSDN achieves massive scale using surprisingly small forwarding tables while supporting both ECMP load balancing and any-to-any end-to-end TE

The two fundamental properties of HSDN

- **All** paths in the network are **pre-established** in the forwarding tables
- Labels identify **entire paths** (or group of paths) rather than simply destinations

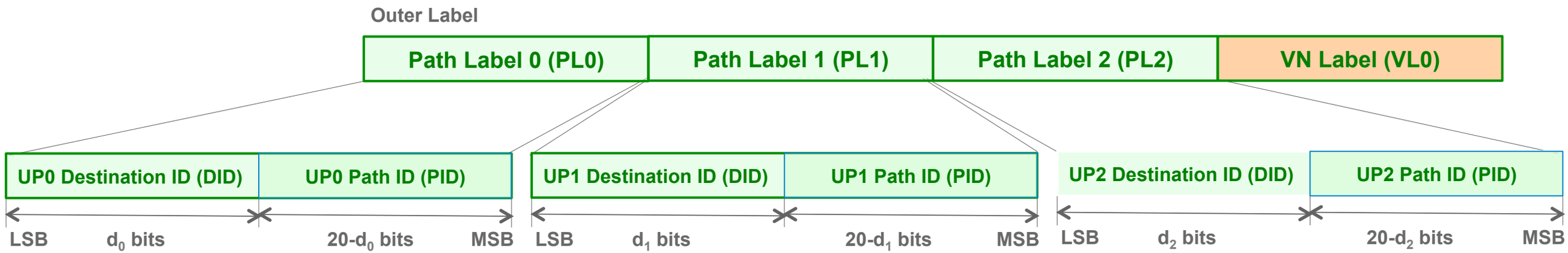
The Life of a Packet

TO SERVER 37 in UP2,4,28



HSDN Label Stack

- Stack of path labels, plus one VN label
- Each path label in the stack associated with a level in the hierarchy
- Labels identify entire paths, rather than just destinations within the partition
- Label stack imposed at entry points
- Labels assigned according to "physical" location of endpoint in the HSDN structure
  - PL associated with a partition globally unique within that partition
  - The LFIBs become rather "static"
  - Single LFIB entry per ECMP group. All paths in each ECMP group use the same outgoing labels.



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 “hybrid-approach” using distributed routing/label distribution protocol to distribute labels, in conjunction with controller
  - Useful during technology migration, to handle non-SDN-capable legacy nodes
  - Example based on BGP-LU in draft-fang-idr-bgplu-for-hsdn-00

HSDN Scaling Examples

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

- UPBNs need to have entries in their LFIBs only to reach destinations in the two partitions to which they belong to
- For route optimization, one hair-pin label value identifies traffic that needs to be kept within the partition once the corresponding UPBN is reached
- 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 more feedback from WG
- Ask for WG adoption