MPLS-Based Hierarchical SDN for Hyper-Scale DC/Cloud
draft-fang-mpls-hsdn-for-hsdc-04

Luyuan Fang, Microsoft
Deepak Bansal, Microsoft
Fabio Chiussi
Chandra Ramachandran, Juniper
Ebben Aries, Facebook
Shahram Davari, Broadcom
Barak Gafni, Mellanox
Daniel Voyer, Bell Canada
Nabil Bitar, Verizon

IETF 93, MPLS WG, July 23, 2015

Changes since 02 version

- Shortened author list on the front page, moved many co-authors to contributor list, and added one co-author and one contributor.
- Changed intended status to Informational
- Updated based on comments from co-authors and non-coauthors
- Added reference to paper "Hierarchical SDN for the Hyper-Scale, Hyper-Elastic Data Center and Cloud," *Proc. ACM SIGCOMM Symposium on SDN Research*, June 2015, http://dl.acm.org/citation.cfm?id=2775009
  - Contains the LFIB computation details with ECMP and TE, scalability analysis, and performance data

Changes since 02 version

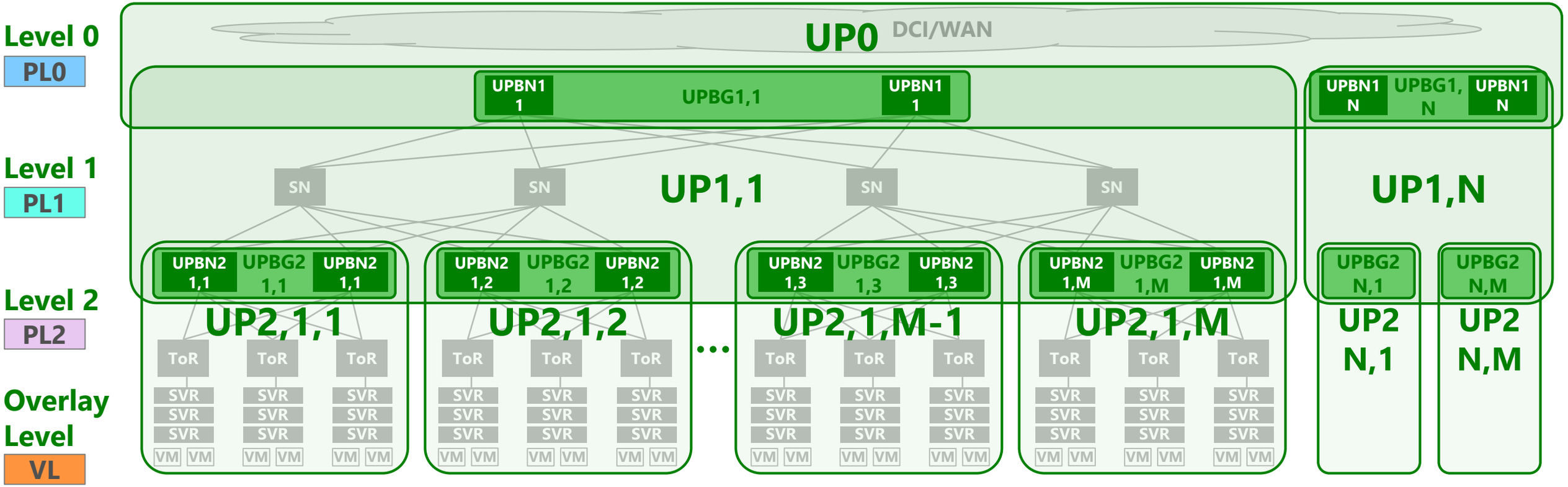
- Spelled out “non-end-to-end” HSDN (i.e., label imposition in the network nodes rather than the servers), in addition to end-to-end.
- Re-written explanation of “turn around” entry for route optimization
- Clarified label scope and uniqueness requirements (UP-wide unique)
- Expanded description of control plane using SDN controller

Benefit of Hierarchical SDN (HSDN)

- Partitioning is crucial for scaling to 10's of millions endpoints
- HSDN is the architecture for partitioning the DC and DCI
  - The principle applies to any forwarding: MPLS, SR, IPv4, and IPv6, L2 or even L1
  - The control plane can be implemented with full SDN approach or using BGP-LU for label distribution (draft-fang-idr-bgplu-for-hsdn-01)
- Two game-changing properties of HSDN
  - All paths in the network can be pre-established in the LFIBs (with small LFIBs)
  - Labels can identify paths, not just destinations

All Paths are set: support End-to-End Any-to-Any TE and ECMP concurrently

HSDN: Hierarchical Underlay Partitioning

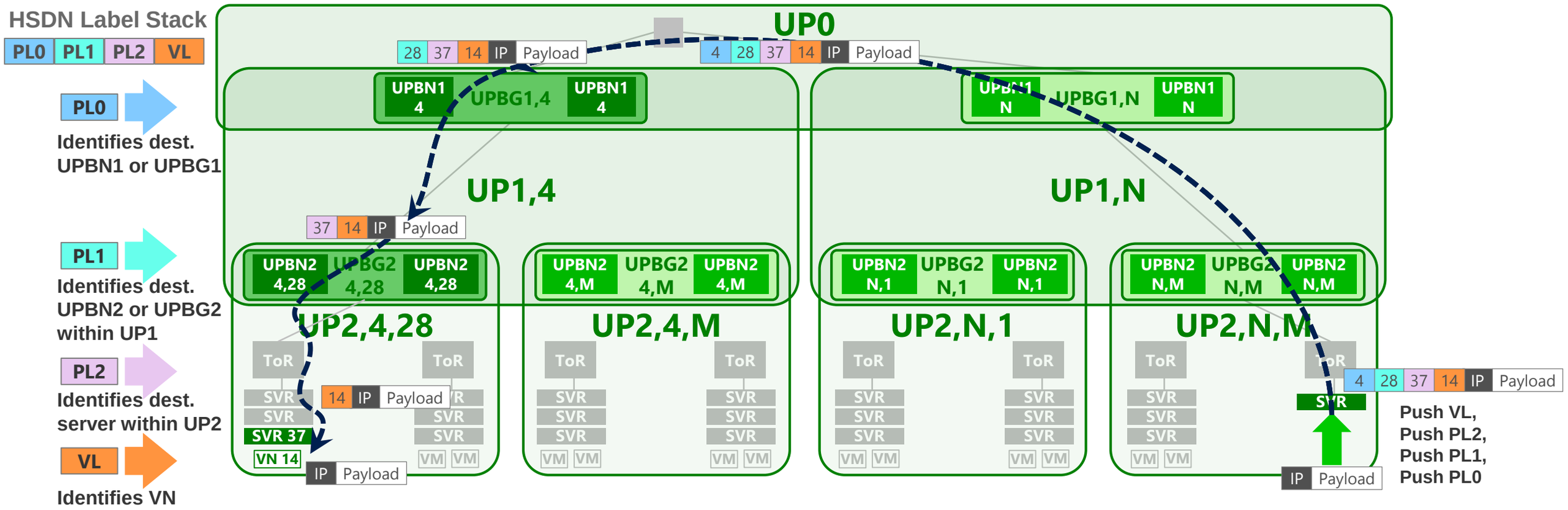


- One path label per level of underlay partition, plus one VN label
- Labels are “static,” globally unique within each partition

Example:

- UP0 = DCI; UP1s = DCs; UP2s = Clusters ➔ With 3 levels, easily scale to 10’s of millions of endpoints

HSDN Forwarding: The Life of a Packet



- Route optimization
  - Forward a packet from any source to any destination using the same (or less) number of hops as in a flat architecture and without introducing any additional latency
  - "Turn Around" entry to optimize label usage

Next Steps

- Request for WG adoption