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

One Consistent Abstraction Paradigm

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

→ “Infinite” Horizontal Scaling
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

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

TO SERVER 37 in UP2,4,28

HSDN Label Stack
3 Path Labels

- PL0
- PL1
- PL2
- VL

UP-0

UP-1,4

UP-2,4,28

UP-2,4,M

UP-1,N

UP-2,N,1

UP-2,N,M

Server

Hyper-V

VM

Payload

IP

UPBN

2,4

UPBN

2,4,M

UPBN

2,4,28

Lookup, Push PL0, Push PL1, Push PL2, Push VL

PL2

VL

PL0

© 2015 Microsoft – All Rights Reserved
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:

<table>
<thead>
<tr>
<th>Number of Server endpoints</th>
<th>Max LFIB size ECMP only (No TE)</th>
<th>Max LFIB size ECMP and TE Concurrently</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 M</td>
<td>~ 1K</td>
<td>&lt; 14K</td>
</tr>
<tr>
<td>10 M</td>
<td>&lt; 2K</td>
<td>&lt; 24K</td>
</tr>
<tr>
<td>40 M</td>
<td>&lt; 3K</td>
<td>&lt; 36K</td>
</tr>
</tbody>
</table>
Next Steps

• Collect more feedback from WG
• Ask for WG adoption