NVA Mapping Distribution Mechanism

draft-dunbar-nvo3-nva-mapping-distribution-02

Linda Dunbar

Donald Eastlake

Tom Herbert

Status

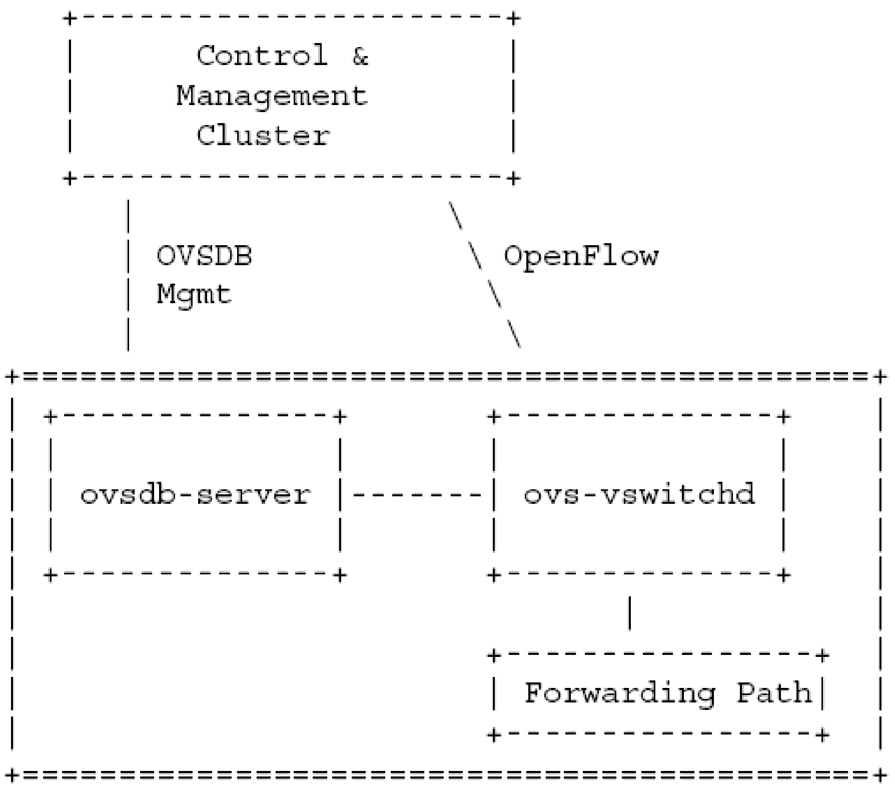
- Reviewed by two NVO3 Interim meetings
- Received a lot of comments with regard to how NVE expressing interested VNs.
- A new subTLV (Enabled-VN TLV) under the IS-IS Router Capability TLV [RFC4971] is specified here for NVE to indicate all its interested VNs in the IS-IS LSP message
- Comparing with OVSDB (Open vSwitch DB Management) mechanism

OVSDB Briefing

Independent Submission by VMware (RFC7047)

The OVSDB management interface is used to perform management and configuration operations on the OVS instance. Compared to OpenFlow, OVSDB management operations occur on a relatively long timescale. Examples of operations that are supported by OVSDB include: Creation, conf, delete, ..

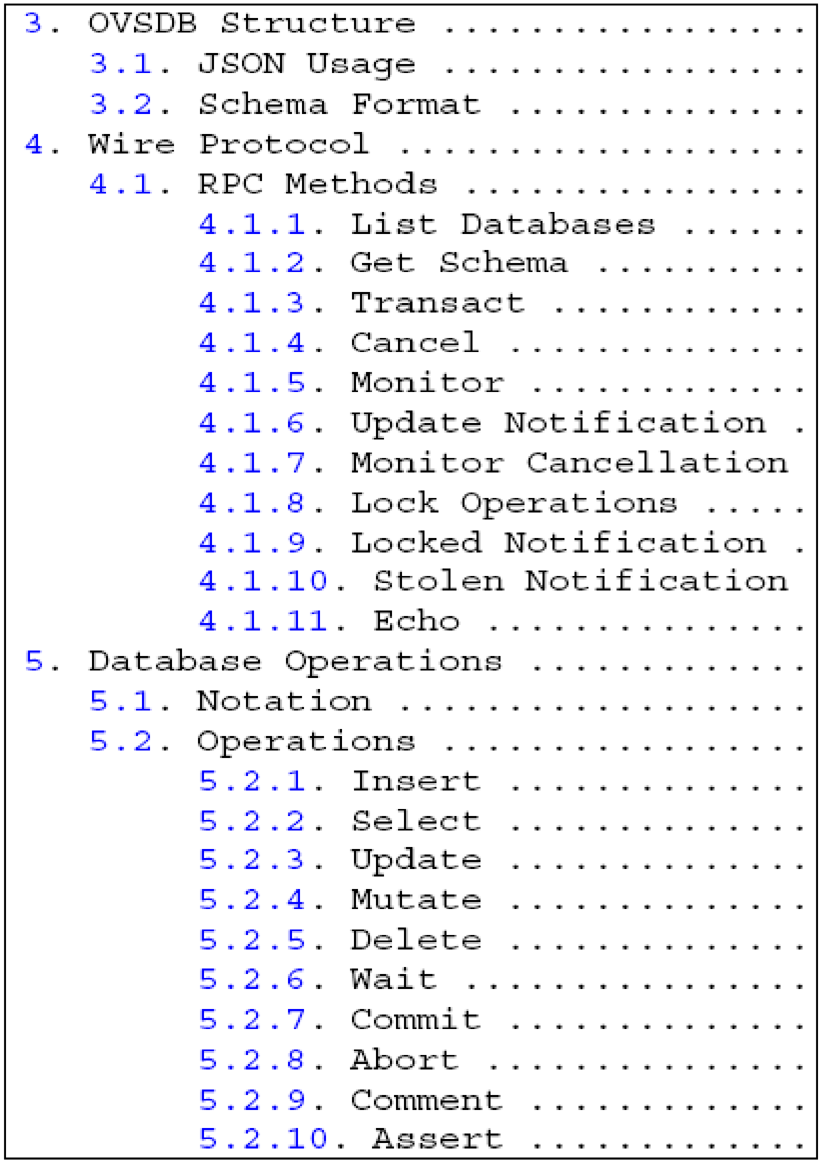
- OVSDB to Virtual Switch is like network element manager (EMS) to switches, i.e. responsible for configuring every aspect of Virtual Switches, including IP addresses for ports, path/link cost, Timer for Spanning Tree, Hello Timer, enabling Multicast snooping, OpenFlow tables, etc.
  - After the vSwitch is configured properly, the controller can use OpenFlow to dynamically send down flow entries. Even though OVSDB can setup the L2/L3 routes (https://github.com/openvswitch/ovs/blob/master/vtep/vtep.xml), dynamic forwarding tables are set up by OpenFlow (for vSwitches)



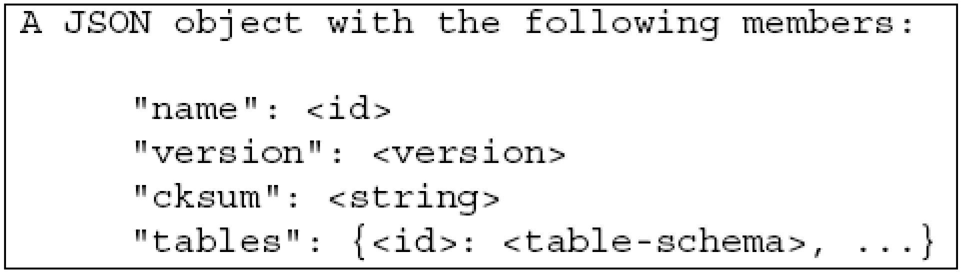
```
   +----------------------+
   |      Control &       |
   |     Management       |
   |      Cluster         |
   +----------------------+
      |                \
      | OVSDB           \ OpenFlow
      | Mgmt             \
      |                   \
+============================================+
| +--------------+       +--------------+    |
| |              |       |              |    |
| | ovsdb-server |-------| ovs-vswitchd |    |
| |              |       |              |    |
| +--------------+       +--------------+    |
|                               |            |
|                        +----------------+  |
|                        | Forwarding Path|  |
|                        +----------------+  |
+============================================+
```

Figure 1: Open vSwitch Interfaces

OVSDB: Push Model

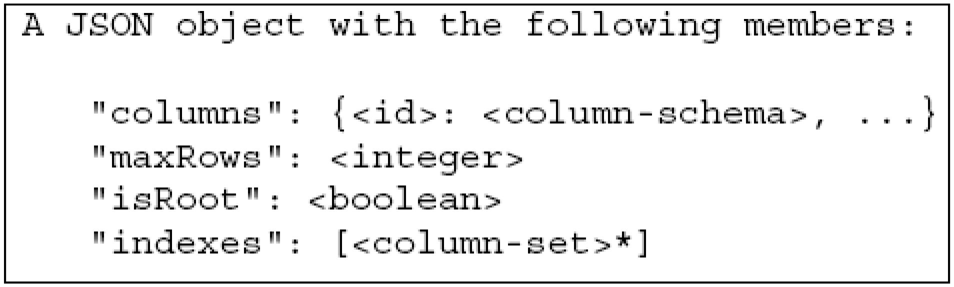


```
3. OVSDB Structure ..................
   3.1. JSON Usage ..................
   3.2. Schema Format ..............
4. Wire Protocol ....................
   4.1. RPC Methods ................
        4.1.1. List Databases ......
        4.1.2. Get Schema ..........
        4.1.3. Transact ............
        4.1.4. Cancel ..............
        4.1.5. Monitor .............
        4.1.6. Update Notification .
        4.1.7. Monitor Cancellation
        4.1.8. Lock Operations .....
        4.1.9. Locked Notification .
        4.1.10. Stolen Notification
        4.1.11. Echo ...............
5. Database Operations ..............
   5.1. Notation ...................
   5.2. Operations .................
        5.2.1. Insert ..............
        5.2.2. Select ..............
        5.2.3. Update ..............
        5.2.4. Mutate ..............
        5.2.5. Delete ..............
        5.2.6. Wait ................
        5.2.7. Commit ..............
        5.2.8. Abort ...............
        5.2.9. Comment .............
        5.2.10. Assert .............
```



```
A JSON object with the following members:

     "name": <id>
     "version": <version>
     "cksum": <string>
     "tables": {<id>: <table-schema>, ...}
```



```
A JSON object with the following members:

   "columns": {<id>: <column-schema>, ...}
   "maxRows": <integer>
   "isRoot": <boolean>
   "indexes": [<column-set>*]
```

NVA-NVE Mapping distribution: Push Model

- **Incremental Push Service Update**
  - *Achieved by Link State Update to distribute the incremental updates.*
- **Requesting Push Service:**
  - Push NVAs use VN scoped reliable *Link State* flooding to announce their availability to push mapping information.
  - NVEs use VN scoped reliable Link State flooding to announce all the Virtual Networks in which they are participating
  - Whenever, there are changes in the mapping entries, NVA uses CSNP messages to only send the changed portion of the entries.
- **Policies:** When ingress edge can’t find entries for the incoming data frame:
  - simply drop the data frame,
  - flood it to all other edges that are in the same VN, or
  - start the “pull” process to get information from Pull NVA

Pull Query Format

- PULL NVA announce its supported VNs
- Pull Requests for the interested VNs or TSs are sent to one specific NVA instance that has the needed information
  - Triggered by:
    - An NVE receives an ingress data frame with a destination whose egress NVE is unknown, or
    - An NVE receives an ingress ARP/ND request for a target whose link address (MAC) or egress edge NVE is unknown.
- Pull Response with instruction on how long entries can be kept by NVE, actions to take if no match is found

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  Ver  | Type  | Flags | Count |      Err      |    SubErr     |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                        Sequence Number                        |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     | QUERY 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...
     | QUERY 2
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...
     | ...
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...
     | QUERY K
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...
```

Figure 4. Pull Query TLV

Open Discussion on Using OVSDB

- How to use OVSDB to distribute incremental changes of inner-outer mappings to all edge nodes?
- OVSDB missing areas:
  - Edge nodes request for Push?
  - Edge nodes express the participated VNs?
  - NVA express the supported VNs ranges/list/?
  - Edge nodes feedback newly discovered attached TSs to NVA
  - Edge nodes exchange mapping among themselves.

Next Step

- NVO3 need at least one way to distribute Mapping;
- Suggest adopt the current draft to NVO3 WG
- Need new proposal for using OVSDB with the open issues addressed.
- NVO3 shouldn’t wait
  - Charter state Control Plane completed by Oct.

BACKGROUND INFORMAITON

bitMap to express interested VNs subTLV

```
   +-+-+-+-+-+-+-+-+
   |INT-VN-TYPE-1  |                  (1 byte)
   +-+-+-+-+-+-+-+-+
   |   Length      |                  (1 byte)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |       |  Start VN ID          |  (4 bytes)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  VNID bit-map....
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Figure 2. Enabled-VN TLV using bit map
```

Range to express interested VNs

```
   +-+-+-+-+-+-+-+-+
   | INT-VN-TYPE-2 |                  (1 byte)
   +-+-+-+-+-+-+-+-+
   |   Length      |                  (1 byte)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |       |  Start VN ID          |  (4 bytes)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |       |  End VN ID            |  (4 byptes)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3. Enabled-VN TLV using Range

List to express interested VNs

```
   +-+-+-+-+-+-+-+-+
   | INT-VN-TYPE-3 |                  (1 byte)
   +-+-+-+-+-+-+-+-+
   |   Length      |                  (1 byte)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |       |        VN ID          |  (4 bytes)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |       |        VN ID          |  (4 bytes)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |       |        VN ID          |  (4 bytes)
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |     .   .   .
   +-+-+-+-+-+-+-+-+-+-+-+-+-+
```

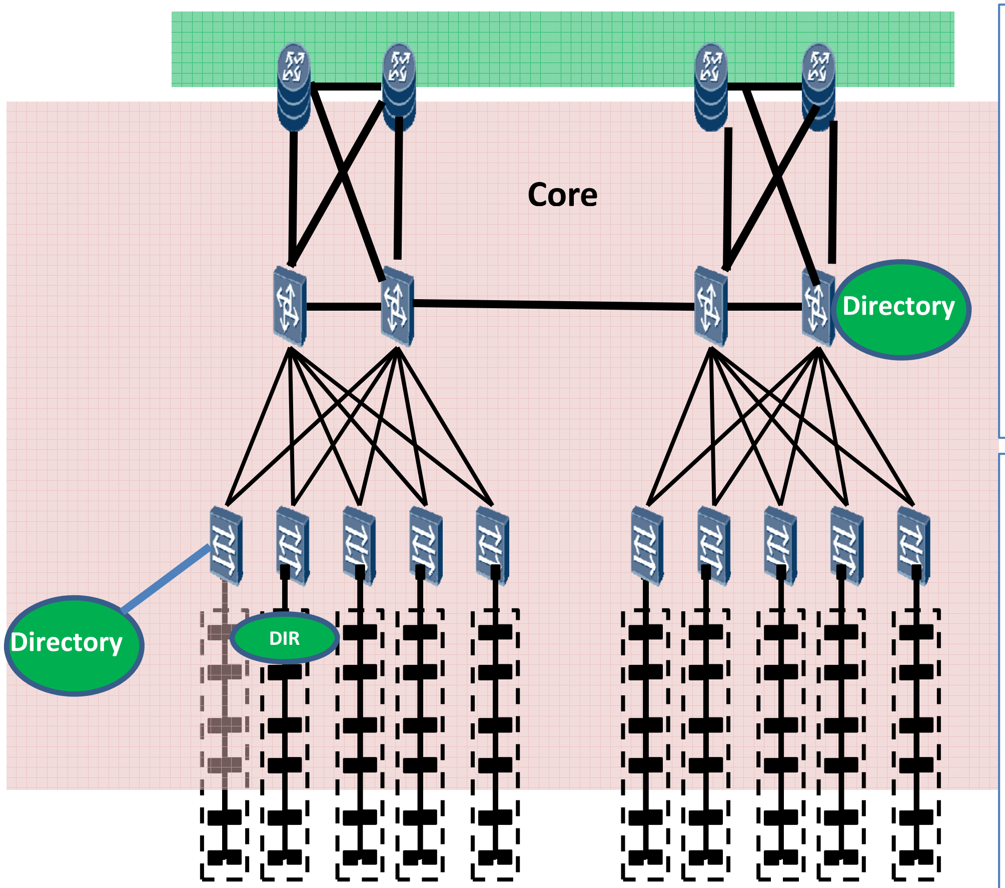
Figure 4. Enabled-VN TLV using list

Incremental Push service

- A new TLV is needed for to carry CSNP timeout value and a flag for NVA to indicate it has completed all updates.

```
            +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+--+
            | Type                            |   (2 bytes)
            +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+--+
            | Length                          |   (2 bytes)
            +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+--+
            |R| Priority    |                     (1 byte)
            +-+-+-+-+-+-+-+-+
            | CSNP Timeout  |                     (1 byte)
            +-+-+-+-+-+-+-+-+
            | Flags         |                     (1 byte)
            +---------------+
            | Reserved for expansion              (variable)
            +-+-+-+-...
   Figure 3. CSNP Complete TLV
Flags: A byte of flags defined as follows:
                  0   1   2   3   4   5   6   7
                +---+---+---+---+---+---+---+---+
                | UN|CP |       RESV            |
                +---+---+---+---+---+---+---+---+
```

Various ways of NVAs connected to NVEs



Locations:

- Embedded in routers/switches in the core, or as standalone servers attached to them.
- Standalone servers or VMs connected to Edges via the client side port

Contents:

- Centralized NVA
- Distributed NVA:
  - Each NVA has mapping for a subset of VNs
  - multiple NVAs have mapping entries for a VN

Reachable Interface Addresses (IA) TLV

- To advertise a set of addresses within a VN being attached to (or reachable by) a specific NVE
- These addresses can be in different address families. For example, it can be used to declare that a particular interface with specified IPv4, IPv6, and 48-bit MAC addresses in some particular VN is reachable from a particular NVE.

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Type = TBD                    |  (2 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Length                        |  (2 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Addr Sets End                 |  (2 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| NVE Address subTLV   ...         (variable)
+-+-+-+-+-+-+-+-+-+-+-+-
| Flags         |                  (1 byte)
+-+-+-+-+-+-+-+-+
| Confidence    |                  (1 byte)
+-+-+-+-+-+-+-+-+-+-
| Template ...                     (variable)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...-+
| Address Set 1    (size determined by Template)    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...-+
| Address Set 2    (size determined by Template)    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...-+
|   ...
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...-+
| Address Set N    (size determined by Template)    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...-+
| optional sub-sub-TLVs ...
+-+-+-+-+-+-+-+-+-+-+-+-+...
```

Query Record

QUERY: Each QUERY Record within a Pull Directory Query message is formatted as follows:

```
    0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
  +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
  |        SIZE           |    RESV   |   QTYPE   |
  +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
    If QTYPE = 1
  +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
  |                      AFN                      |
  +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
  |  Query address ...
  +--+--+--+--+--+--+--+--+--+--+--+...
    If QTYPE = 2, 3, 4, or 5
  +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
  |  Query frame ...
  +--+--+--+--+--+--+--+--+--+--+--+...
```

QTYPE	Description
-----	-----------
0	reserved
1	address query
2	ARP query frame
3	ND query frame
4	RARP query frame
5	Unknown unicast MAC query frame
6-14	assignable by IETF Review
15	reserved

AFN: Address Family Number of the query address.

PULL Responses

- **When the mapping entry is available in the NVA**
  - **Valid Response**
- **When the mapping is not available:**
  - “drop” or “native-forward” (i.e. flooding)
- **cache timer**

```
0                   1                   2                   3
       0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |  Ver  | Type  | Flags | Count |      Err      |    SubErr     |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |                        Sequence Number
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      | RESPONSE 1
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...
      | RESPONSE 2
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...
      | ...
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...
      | RESPONSE K
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...
```

Slide 17

L1 What if removing the sequence number?
L73504, 1/28/2015

Pull Response

RESPONSE: Each RESPONSE record within a Pull NVA Response message is formatted as follows:

```
           0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
         +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
         |         SIZE          |OV|  RESV  |   Index   |
         +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
         |                   Lifetime                    |
         +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
         |                Response Data ...
         +--+--+--+--+--+--+--+--+--+--+--+--...
```

Push-Pull Hybrid Model

- Push model are used for some VNs, and pull model are used for other VNs.
  - It can be operator’s decision (i.e. by configuration) on which VNs’ mapping entries are pushed down from NVA (e.g. frequently used) and which VNs’ mapping entries are pulled (e.g. rarely used).
  - Useful for Gateway nodes where great number of VNs are enabled.
- Or, a portion of hosts in a VN is pushed, other portion has to be pulled.