Benchmarking Workgroup
Internet-Draft
Intended status: Informational
Expires: November 19, 2018

S. Wu
Juniper Networks

May 18, 2018

Network Service Layer Abstract Model
draft-xwu-bmwg-nslam-01

Abstract

While the networking technologies have evolved over the years, the layered approach has been dominant in many network solutions. Each layer may have multiple interchangeable, competing alternatives that deliver a similar set of functionality. In order to provide an objective benchmarking data among various implementations, the need arises for a common abstract model for each network service layer, with a set of required and optional specifications in respective layers. Many overlay and or underlay solutions can be described using these models.

Table of Contents

1. Introduction

This document provides a reference model for common network service framework. The main purpose is to abstract service model for each network layer with a small set of key specifications. This is essential to characterize the capability and capacity of a production network, a target network design. A complete service model mainly includes

Infrastructure - devices, links, and other equipment.

Services - network applications provisioned. It is often defined as device configuration and or resource allocation.

Capacity - A set of objects dynamically created for both control and forwarding planes, such as routes, traffic, subscribers and etc. In some cases, the amount and types of traffic may impact control plane objects, such as multicast or ethernet networks.

Performance Metrics - infrastructure resource utilization.

1.1. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

1.2. Purpose of The Document

Many efforts to YANG model and OpenConfig collaboration are well under way. This document specifies a higher layer abstraction that reuses a small subset of YANG keywords for service description purpose. It SHALL NOT be used for production provisioning purpose. Instead, it can be adopted for design spec, capacity planning, product benchmarking and test setup.

The specification described in this document SHALL be used for outline service requirements from customer perspective, instead of network implementation mechanism from operators perspective.

1.3. Conventions Used in This Document

Descriptive terms can quickly become overloaded. For consistency, the following definitions are used.

- Node - The name for an attrubute.
- Brackets "[" and "]" enclose list keys.
- Abbreviations before data node names: "rw" means configuration data (read-write), and "ro" means state data (read-only).
- Parentheses enclose choice and case nodes, and case nodes are also marked with a colon (":").

2. Network Service Framework

The network service layer abstract model is illustrated by Figure 1. This shows a stack of components to enable end-to-end services. Not all components are required for a given service. A common use case is to pick one component as target service with a detailed profile, with the remaining components as supporting technologies using default profiles.

Network Service Layer Abstract Model

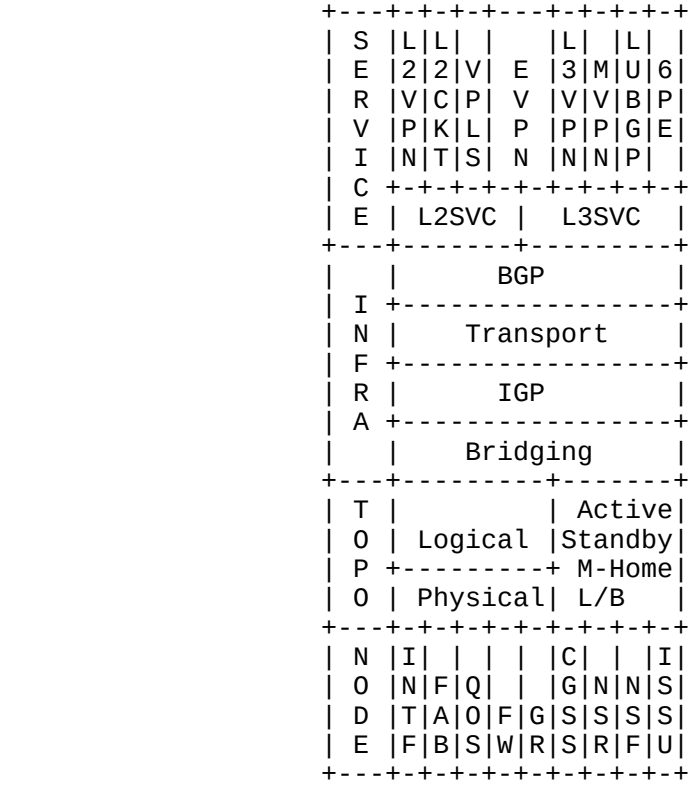

Figure 1

2.1. Node

A network node or a network device processes and forwards traffic based on predefined or dynamically learned policies. This section covers standalone features like the following:

- INTF - Network interfaces that provides internal or external connectivity to adjacent devices. Only physical properties of the interfaces are of concern in this level. The interfaces can be physical or logical, wired or wireless.

- FAB - Fabric capacity. It provides redundancy and cross connect within the same network device among various linecards or interfaces

- QOS - Quality of Services. Traffic queuing, buffering, and congestion management technologies are used in this level

- FW - Firewall filters or access control list. This commonly refers to stateless packet inspection and filtering. Stateful firewall is out of scope of this document. Number of filters daisy chained on a given protocol family, number of terms within same filter, and depth of packet inspection can all affect speed and latency of traffic forwarding. It also provides necessary security protection of node, where protocol traffic may be affected.

- GR - Graceful Restart per protocol. It needs to cooperate with adjacent node

- CGSS - Controller Graceful / Stateful Switchover. A network device often has two redundant controllers to minimize the disruption in event of catastrophic failure on the primary controller. This is accomplished via real time state synchronization from the primary to the backup controller. It, however, should be used along with either NSR or NSF to achieve optimal redundancy.

- NSR - Non-Stop Routing - hitless failover of route processor. It maintains an active copy of route information base (RIB) as well as state for protocol exchange so that the protocol adjacency is not reset.

- NSF - Non-Stop Forwarding for layer 2 traffic, including layer 2 protocols such as spanning tree state

- ISSU - In Service Software Upgrade - Sub-second traffic loss in many modern routing platforms. The demand for this feature continues to grow from the field. Some study shows that the downtime due to software upgrade is greater than that caused by unplanned outages.

2.2. Topology

Placement of network devices and corresponding links plays an important role for optimal traffic forwarding. There are two types of topologies:

- Physical Topology - Actual physical connectivity via fiber, coax, cat5 or even wireless. That could be a ring, bus, star or matrix topology. Even though all can be modeled using point-to-point connections.

- Logical Topology - With aggregated ethernet, extended dot1q tunneling, or VxLAN, a logical or virtual topology can be easily created spanning across geography boundaries. Recent development of multi-chassis, virtual-chassis, node-slicing technologies, and multiple logical units within a single physical node have enabled logical topology deployment more flexible and agile.

With various topology, the following functionalities need to be taken into consideration for feature design and validation.

- Active-Standby - 1:1 or 1:n support. The liveness detection is essential to trigger failover.

- M-Home - Multi-homing support. A customer edge (CE) device can be homed to 2 or more Provider Edge (PE) devices at the same time. This is a common redundancy design in layer 2 service offering

- L/B - Load Balancing - When multiple diverse paths exist for a given destination, it is important to achieve load balancing based on multiple criteria, such as per packet, or per prefix. Sometimes, cascading effect can make issues more complex and harder to resolve

The topology, regardless of physical or virtual, can be better depicted using a collection of nodes and point-to-point links. Some broadcast network, or ring topology, can also be abstracted using same collection of point-to-point links. For example, in a wireless LAN network, each client is a node with wireless LAN NIC as its physical interface. The access point is the node, at which all WLAN cients terminate with airwaves. The Service Set IDentifier (SSID) on this access point can be considered as part of broadcast domain with many pseudo-ports taking airwave terminations from clients.

The default link id can use srcnode-dstnode-linkseq to uniquely identify a link in this topology. If this is a link connecting two ports on the same node, it can use link-id of srcnode-srcnode-

linkseq-portseq. Additional attributes of the node can be added with proper placement for auto topology diagram.

Network Topology Definition

```
node-id-1 {
    maker: maker_name,
    model: model_name,
    controller: controller-type,
    mgmt_ip: mgmt_ip_address,
    links: {
        link-id-1 {
            name: link_name,
            connector: 'sfpp',
            attr: ['10G', 'Ethernet'],
            node_dst: destination node-id,
            link_seq: sequence number for links between the node pair
            ...
        }
    }
    ...
}
node-id-2 {
    ...
}
```

Figure 2

2.3. Infrastructure

Network infrastructure here refers to a list of protocols and policies for a data center network, an enterprise network, or a core backbone in a service provider network.

- Bridging - Spanning Tree Protocol (STP) and its various flavors, 802.1q tunneling, Q-in-Q, VRRP and etc
- IGP - Interior Gateway Protocol - some common choices are OSPF, IS-IS, RIP, RIPng. For multicast, choices are PIM and its various flavors including MSDP, Bootstrap, DVMRP
- Transport - Tunnel technologies including
  - MPLS - Multi-Protocol Label Switching - most commonly used in service provider network
    - LDP - Label distribution protocol - including mLDP and LDP Tunneling through RSVP LSPs

        + RSVP - Resource Reservation Protocol - including P2MP and its various features like Fast ReRoute - FRR.

      * IPSec - IPSec Tunnel with AH or ESP

      * GRE - Generic Routing Encapsulation (GRE) tunnels provides a flexible direct adjacency between two remote routers

      * VxLAN - In data center interconnect (DCI) solutions, VxLAN encapsulation provides data plane for layer 2 frames

   o BGP - Define families and their sub-SAFI deployed, as well as route reflector topology.

2.4. Services

Previous sections mostly outline an operator's implementation of the network, while customers may not necessarily care about these. This section defines service profiles from customer's view.

   o Layer 2 Services

      * Layer 2 VPN - RFC 6624

      * Martini Layer 2 Circuit - RFC 4906

      * Virtual Private LAN Services - RFC 4761

      * Ethernet VPN - RFC 7432

   o Layer 3 Services

      * Type 5 Route for EVPN - draft-ietf-bess-evpn-prefix-advertisement-05

      * Layer 3 VPN - RFC 4364

      * Labled Unicast BGP - RFC 3107

      * Draft Rosen MVPN - RFC 6037

      * NG MVPN - RFC 6513

      * 6PE - RFC 4798

In next section, an abstract model is proposed to identify key metrics for both layer 2 and layer 3 model

3. Service Models

A service model is a high level abstraction of network deployment from bottom up. It defines a set of common key characteristics of customer traffic profile in both control and forwarding planes. The network itself should be considered as a blackbox and deliver the services regardless of types of network equipment vendor or network technologies.

The abstraction removes some details like specific IP address assignment, and favors address range and its distribution. The goal is to describe aggregated network behavior instead of granular network element configuration. It is up to implementation to map aggregated metrics to actual configuration for the network devices, protocol emulator and traffic genrator.

A single network may be comprised of multiple instances of service models defined below.

3.1. Layer 2 Ethernet Service Model

The metrics outlined below are for layer 2 network services typically within a data center, data center interconnect, metro ethernet, or layer 2 domain over WAN or even inter-carrier.

- service-type: identityref, ELAN, ELINE, ETREE
- sites-per-instance: uint32, an average number of sites a layer 2 instance may span across
- global-mac-count: uint32, Global MAC from all attachment circuits, local and remote. This is probably the most important metric that determines the capacity requirements in layer 2 for both control and forwarding planes
- interface-mac-max: uint32, maximum number of locally learned MAC addresses per logical interface, aka attachment circuit
- single-home-segments: uint32, number of single homed ethernet segments per service instance
- multi-home-segments: uint32, number of multi homed ethernet segments per layer 2 service instance
- service-instance-count: uint32, total number of layer 2 service instances. Typically, one customer is
- traffic-type: list, {known-unicast: %, multicast, %, broadcast: %, unknown-unicast: %}
- traffic-frame-size: list, predefined mixture of traffic frame size distribution
- traffic-load: speed of traffic being sent towards the network. This can be defined as frame per second (fps), or actual speed in bps. This is particular important whenever some component along

forwarding path is implemented in software, the throughput might be affected significantly at high speed
- traffic-flow: A distribution of flows. This may affect efficient use of load-balancing techniques and resource consumption. More details discussed in later section of this document.
- layer3-gateway-count: uint32, number of layer 2 service instances that also provide layer 3 gateway service
- arpnp-table-size: uint32. This is only relevant with presence of layer 3 gateway

Integrated routing and bridging (IRB) and EVPN Type 5 route have blurred boundaries between layer 2 and layer 3 services.

3.2. Layer 3 Service Model

This section outlines traffic type, layer 3 protocol families, layer 3 prefixes distribution, layer 3 traffic flow and packet size distributions.

- proto-family: protocol family are defined with three sub-attributes. The list may grow as the complexity
  - proto - list: inet, inet6, iso
  - type - list: unicast, mcast, segment, labeled
  - vpn - list, true, false
- prefix-count, uint32, total unique prefixes
- prefix-distrib, list of prefix length size and percentage. This could be a distribution pattern, such as uniform, random. Or simply top representation of prefix lengths
- bgp-path-count, uint32, total BGP paths
- bgp-path-distrib, top representation of number of paths per prefix
- traffic-frame-size, similar to traffic-frame-size in layer 2 model. The focus is on the MTU size on each protocol interfaces and the impact of fragmentation
- traffic-flow, similar to traffic-flow in layer 2 model, it focuses on a set of labels, source and destination addresses as well as ports
- traffic-load, similar to traffic-load in layer 2 model
- ifl-count, uint32,
- vpn-count, uint32,

3.3. Infrastructure Service Model

- bgp-peer-ext-count, uint32, number of eBGP peers
- bgp-peer-int-count, uint32, number of iBGP peers
- bgp-path-mtu, list, true or false. Larger path mtu helps convergence

- o bgp-hold-time-distrib, list of top hold-time values and their respective percentage out of all peers.
- o bgp-as-path-distrib, list of top as-path lengths and their respective percentage of all BGP paths
- o bgp-community-distrib, list of top community size and their respective percentage out of all BGP paths
- o mpls-sig, list, MPLS signaling protocol, rsvp or ldp
- o rsvp-lsp-count-ingress, uint32, total ingress lsp count
- o rsvp-lsp-count-transit, uint32, total transit lsp count
- o rsvp-lsp-count-egress, uint32, total egress lsp count
- o ldp-fec-count, uint32, total forwarding equivalence class
- o rsvp-lsp-protection, list, link-node, link, frr
- o ospf-interface-type, list, point-to-point, broadcast, non-broadcast multi-access
- o ospf-lsa-distrib, list. OSPF Link Statement Advertisement distribution is comprised of those for core router in backbone area, and internal router in non-Backbone areas. A common modeling can include number of LSAs per OSPF LSA type
- o ospf-route-count, list, total OSPF routes in both backbone and non-backbone areas
- o isis-lsp-distrib, list, similar to ospf-lsa-distrib
- o isis-route-count, list, total IS-IS routes in both level-1 and level-2 areas

TODO: bridging, OAM, EOAM, BFD and etc

3.4. Node Level Features

TODO: node level feature set

3.5. Common Service Specification

For most network services, regardless of layer 2 or layer 3, protocol families, the following needs to be considered when measuring network capacity and baseline.

- o rib-learning-time, uint32 in seconds. This indicates show quickly the route processor learns routing objects either locally and remotely
- o fib-learning-time. In large routing system, forwarding engine residing on separate hardware from controller, takes additional time to install all forwarding entries learned by controller.
- o convergence-time, this is could be as a result of many events, such as uplink failure, ae member link failure, fast reroute, local repair, upstream node failure and etc
- o multihome-failover-time, this refers to traffic convergence in a topology where a customer edge (CE) device is connected to two or more provider edge (PE) devices.

- o issu-dark-window-size. Unlike NSR, the goal of ISSU is not zero packet loss. Instead, there will be a few seconds, or in some cases, sub-second dark window where it sees both total packet loss for both transit and or host bound protocol traffic.
- o cpu-util, total CPU utilization of the controllers in stead state
- o cpu-util-peak, Peak CPU utilization on the controller in event of failure, and convergence
- o mem-util, total memory utilization of the controllers in steady state
- o mem-util-peak, total memory utilization on the controller in event of failure and convergence
- o processes, list of top processes running on the controllers with their CPU and memory utilization.
- o lc-cpu-util, top CPU utilization on the line cards
- o lc-cpu-util-peak, maximum peak CPU utilization among all line cards in event of failure and convergence
- o lc-mem-util, top memory utilization on the line cards
- o lc-mem-util-peak, maximum peak memory utilization among all line cards in event of failure and convergence
- o throughout, in both pps and bps. This is measured with zero packet loss. For virtualized environment, throughput is sometimes measured with a small loss tolerance given the nature of shared resource
- o traffic performance, in both pps and bps. It is measured the rate of traffic received by pumping oversubscribed traffic at ingress
- o latency in us. this is more important within a local data center environment rather than DCI over wide area network. Use of extensive firewall filter or access control lists may affect latency
- o Out of Order Packet - This can happen in intra-node or over ECMP where different paths have large latency/delay variations.

The list of metrics can be used for network monitoring during network resiliency test. This is to understand how quickly a network service can restore during various events and failures

3.6. Common Network Events

A list of events is defined to characterize network resiliency. These attributes require that the provider networks have diverse paths and node redundancy built-in. They directly affect service level agreement and network availability.

3.6.1. Event Attributes

Each network or system event may each be defined with the following aspects

- o event-iteration, uint16, event to be repeated
- o event-interval, unit16, seconds in between consecutive events
- o event-dist, list, random, equal, or other type of event scheduling
- o event-timeout, uint16, seconds when a single event is expected to complete
- o event-convergence, uint16, seconds before the network can be recovered

3.6.2. Hardware Related

Some hardware failures can not easily replicated, or even simulated in a lab environment, like the memory errors. A system or network should be equipped to monitor, detect and contain the impact to avoid global catastrohpic failure that may proagate beyond a single node or the regional network.

- o hw-mod-yank: hardware module removal and insertion.
- o hw-interface: Transceiver on/off or any other simulated link failures.
- o hw-storage: storage failure, ether local or network attached.
- o hw-power: unplanned power failure.
- o hw-controller: Controller failure
- o hw-memory: memory errors

3.6.3. Software Component

- o sw-daemon-watchdog-loss: Induced CPU hog that trigger watchdog failure
- o sw-daemon-restart-graceful: Graceful software daemon restart.
- o sw-daemon-restart-kill: The process is killed and the daemon was forced to restart
- o sw-daemon-panic: Sometimes a panic can introduced to trigger a coredump of software daemon along with restart.
- o sw-os-panic: network operating system may panic under various situations. Many network products with a console access support OS panic with a special sequence keystroke. Sometimes it may also generate a coredump for further debug
- o sw-upgrade: In many provider networks, the most downtime actually come from scheduled maintenance, especially software or firmware upgrade to provide better feature set or as a result of security patch. It is important to understand the downtime requirement for a routine software upgrade on a given network device or the devices in the network. This often presents a challenge to the access network.

3.6.4. Protocol Events

- protocol-keepalive-loss: Loss of keepalive, such as hellos for routing-protocols like OSPF and BGP
- oam-keepalive-loss: there are many OAM protocols, such as EOAM, MPLS OAM, LACP, one of their main purposes is to detect reachability. This is different from routing protocols keepalive
- protocol-adjacency-reset: clear all protocol neighbors and any routes or link states learned from the neighbor
- protocol-db-purge: remove all database objects learned from a particular neighbor, or a group of neighbors, or all neigbors. The database maybe original set of state learned from neighbors, or the consolidated database.

3.6.5. Redundancy Failover

The network protocols and design have a lot of redundancy built-in. It is important to benchmark their effectiveness.

- ha-lag-links: measure packet loss in milliseconds when member link(s) of a link aggregation group is torn down. In case of protected interface, traffic should failover seamlessly to the backup interface in event of primary link failure
- ha-controller: In a system with redundant controller, measure the network recovery time when the primary controller fails. If the advanced non-stop routing/forwarding is enabled, the network should only experience zero or sub-second traffic loss.
- ha-multihome: in addition to device level redundancy, many protocols support network layer redundancy though multihoming such as EVPN.
- ha-mpls-frr: MPLS RSVP Fast ReRoute provides core network redundancy
- ha-uplink: the core network is typically designed to provide path diversity for edge devices, at either layer 2 and layer 3 connectivity. The resiliency of network is measured by how fast the system detects the failure and reroute the traffic

4. Use of Network Service Layer Abstract Model

The primary goal is to characterize and document a complex network using a simplified service model. While eliminating many details such as address assignment, actual route or mac entries, it retains a set of key network information, including services, scale, and performance profiles. This can be used to validate how well each underlying solution performs when delivering same set of services.

The model can also be used to build a virtualized topology with both static and dynamic scale closely resemble to a real network. This

eases network design and benchmarking, and helps capacity planning by studying the impact with changes to a specific dimension.

5. Acknowledgements

The authors appreciate and acknowledge comments from Al Morton and others based on initial discussions.

6. IANA Considerations

This memo includes no request to IANA.

All drafts are required to have an IANA considerations section (see Guidelines for Writing an IANA Considerations Section in RFCs [RFC5226] for a guide). If the draft does not require IANA to do anything, the section contains an explicit statement that this is the case (as above). If there are no requirements for IANA, the section will be removed during conversion into an RFC by the RFC Editor.

7. Security Considerations

All drafts are required to have a security considerations section. See RFC 3552 [RFC3552] for a guide.

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC3552] Rescorla, E. and B. Korver, "Guidelines for Writing RFC Text on Security Considerations", BCP 72, RFC 3552, DOI 10.17487/RFC3552, July 2003, <https://www.rfc-editor.org/info/rfc3552>.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226, DOI 10.17487/RFC5226, May 2008, <https://www.rfc-editor.org/info/rfc5226>.

8.2. Informative References

[L2SM] B. Wu et al, "A Yang Data Model for L2VPN Service Delivery", 2017, <https://www.ietf.org/id/draft-ietf-l2sm-l2vpn-service-model-02.txt>.

[RFC8049] Litkowski, S., Tomotaki, L., and K. Ogaki, "YANG Data Model for L3VPN Service Delivery", RFC 8049, DOI 10.17487/RFC8049, February 2017, <https://www.rfc-editor.org/info/rfc8049>.

Author’s Address

Sean Wu
Juniper Networks
2251 Corporate Park Dr.
Suite #200
Herndon, VA 20171
US

Phone: +1 571 203 1898
Email: xwu@juniper.net