Benchmarking Methodology for Computing-Aware Traffic Steering (CATS)

draft-yl-bmwg-cats-00

K. Yao, P. Liu, China Mobile

Motivations

- **Computing-aware traffic steering(CATS) is a traffic engineering approach considering both computing and network metrics, in order to select appropriate service instances.**
- *https://datatracker.ietf.org/wg/cats/about/*

- **Major Objectives:**
  1. Assess what CATS metrics are appropriate for instance selection
  2. Assess the benefits and trade-offs of multiple deployment solutions
  3. Analyze CATS effectiveness from different perspectives, e.g., system utilization, latency, session continuity, protocol overhead, etc.

Aspects that Impact CATS Solution Choices:

- Multiple levels of CATS metrics that can be chosen for path and instance selection:
  - Three metric levels defined in *daft-ietf-cats-metric-definition*
  - *L0: Raw metrics*
  - *L1: Metrics normalized within categories*
  - *L2: Fully normalized metrics*
  - metric levels differ in encoding complexity, extensibility, stability, and accuracy.

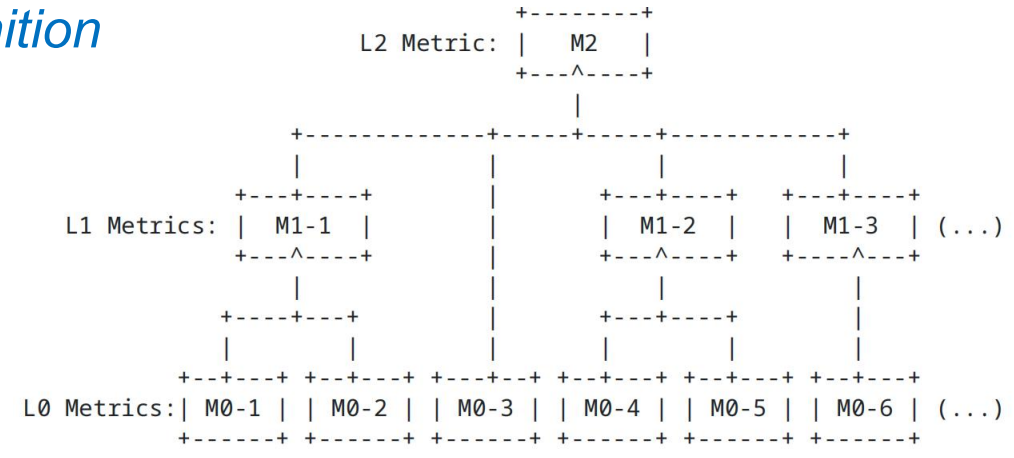


Figure 1: Logic of CATS Metrics in levels

- Placement of aggregation and normalization functions for CATS metrics
  - CATS service contact instance, CATS-forwarder, C-SMA.
- Methods for metrics exposure or distribution:
  - Distributed approach choose distributed control plane protocols, e.g., BGP extensions.
  - Centralized approach choose bidirectional interfaces and protocols, e.g., restful APIs.

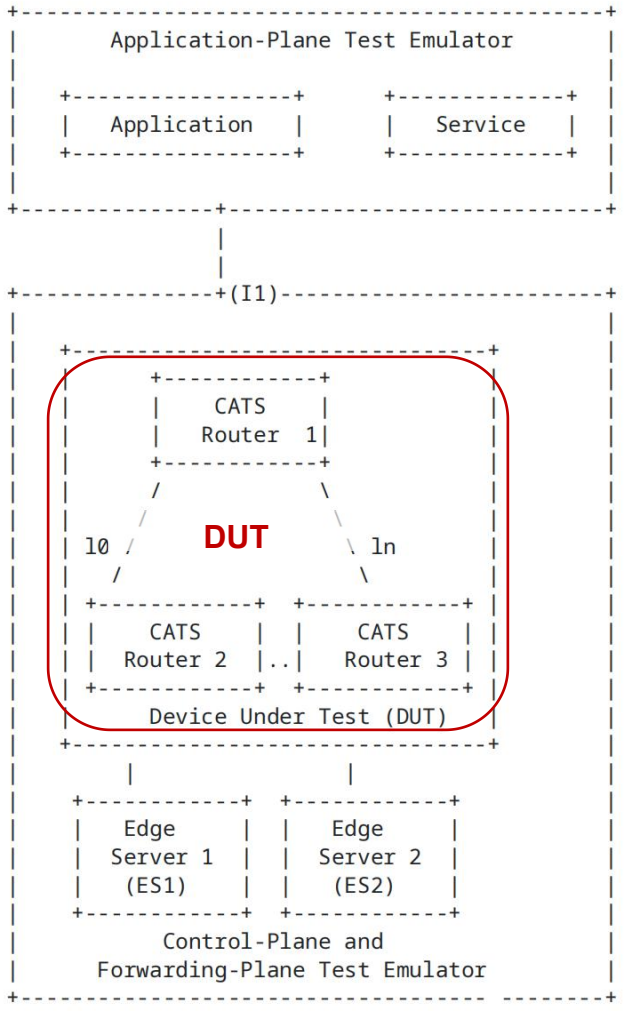
Structure of the I-D:

- **Section 3: Test Methodology**
  - Introduce two deployment approaches on CATS framework
  - Centralized Approach: SDN controller serves as the path selector(C-PS)
  - Distributed Approach: CATS Ingress node serves as the C-PS
- **Section 4: Reporting Format**
- **Section 5: Benchmarking Tests**
  - Test the functionalities of CATS metrics collection and exposure
  - Test session continuity(instance affinity)
  - Test CATS scheduling latency
  - Test the system utilization(CPU, GPU..)

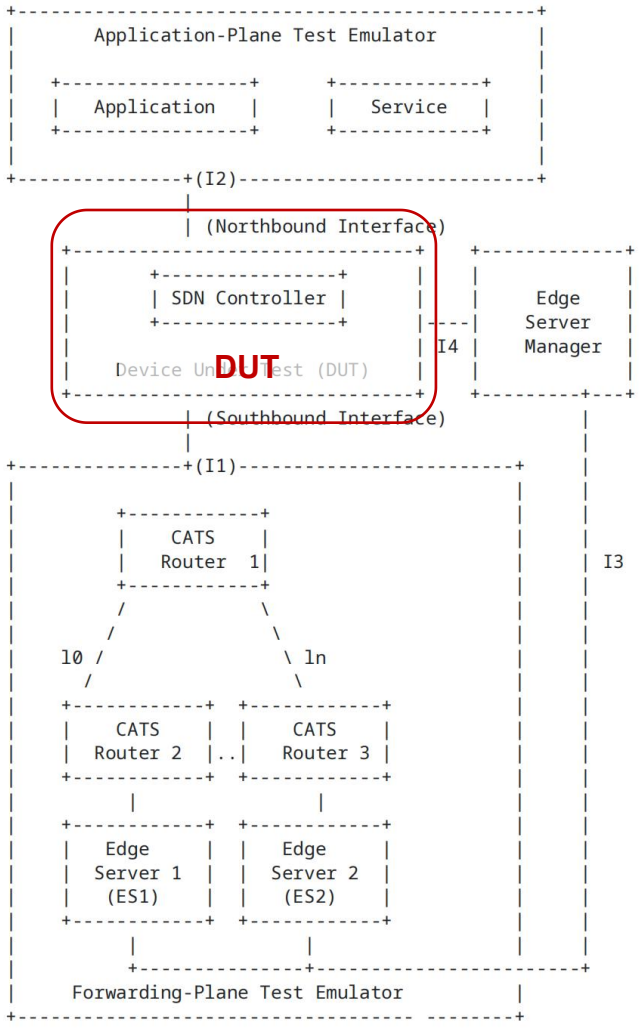
3. Test Methodology
  3.1. Test Setup
    3.1.1. Test Setup - Centralized Approach
    3.1.2. Test Setup - Distributed Approach
  3.2. Control Plane and Forwarding Plane Support
  3.3. Topology
  3.4. Device Configuration
4. Reporting Format
5. Benchmarking Tests
  5.1. CATS Metrics Collection and Distribution
  5.2. Session continuity
  5.3. Latency
  5.4. Sytem Utilization
6. Security Considerations
7. IANA Considerations

Test Setup for Both of the Solutions:

- **Application Plane:**
  - Emulate services that require CATS for service scheduling
  - CDN, cloud rendering...
- **Control plane & Forwarding plane:**
  - Distributed approach - DUT is the CATS router(ingress and egress)
  - Centralized approach - DUT is the SDN controller



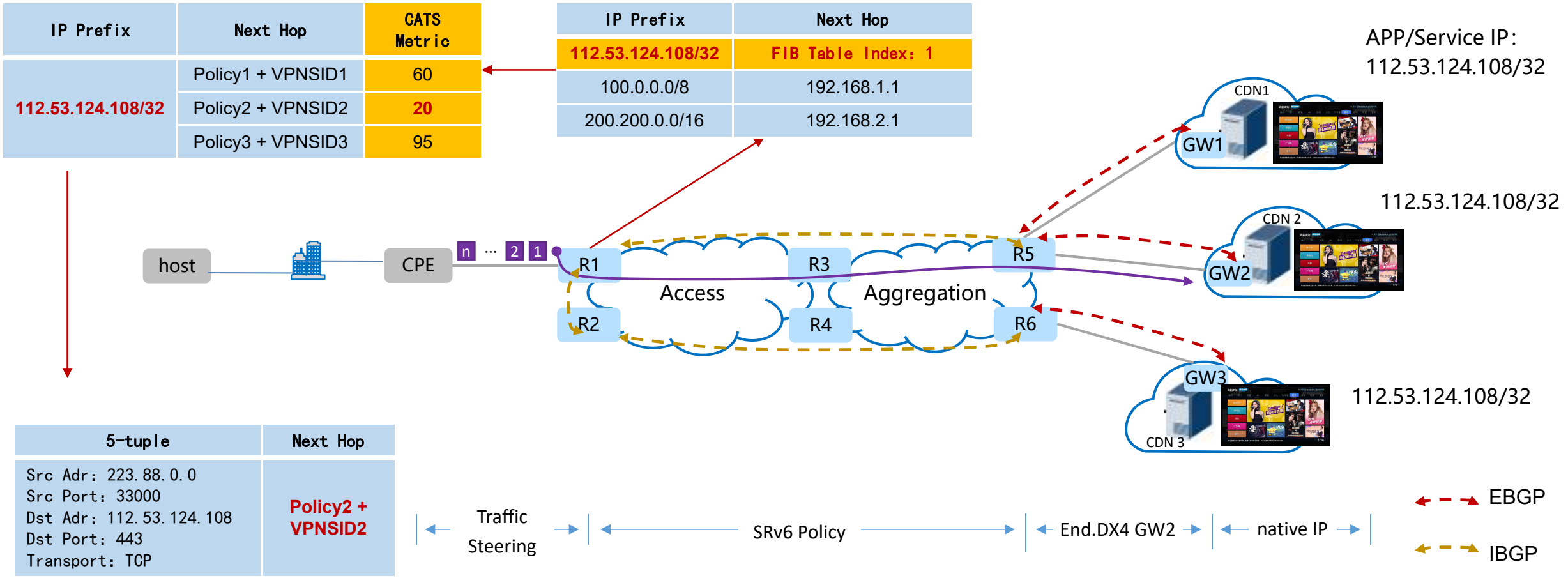
Distributed solution setup



Centralized solution setup

Distributed Approach as An Example:

- CATS metrics are distributed by **extending BGP**
- CATS ingress node serves as C-PS based on CATS metrics and select corresponding **SRv6 policy** to steer traffic



Next Step

- This is the first time to propose this document in bmwg & ippm
- Authors think it within the scope of bmwg, since
  - clear description on the class of network function, system, or service being addressed(CATS system and service)
  - clearly identify the metrics for traffic steering(CATS metrics)
  - methodologies and benchmarking tests(distributed and centralized solutions)
  - tests are done within controlled environments
- Solicit for comments and welcome for contributions
- Revise the I-D and ask for presentation/adoption in IETF 124