

Framework and Applicability of Computation-aware Traffic Steering in Optical Transport Networks

draft-zhao-cats-otn-applicability-00

Yang Zhao (zhaoyangyjy@chinamobile.com)

LiuYan Han (hanliuyan@chinamobile.com)

Xiao Li (lixiao33@huawei.com)

Haomian Zhang (zhenghaomian@huawei.com)

Daniel King (daniel@olddog.co.uk)

Motivation

- Highlight how Computation-aware Traffic Steering (CATS) would be implemented using Optical Transport Network (OTN) technology.
 - Extends CATS thinking into optical transport domains
- Multi-site compute services are now common.
- Best effort dispatch to the “nearest” site is often not enough.
 - Brings deterministic transport into the discussion
- Pure packet-based steering may not meet strict latency, jitter, and determinism targets.

Use Cases

- Some workloads need both:
 - awareness of compute state
 - awareness of network state
- Use Cases include:
 - AI large-model training
 - Some AI inference jobs
 - High performance computing workloads
 - Tele-health and other BW performance-sensitive services
 - Object-based Media (see BBC example)
 - <https://datatracker.ietf.org/doc/draft-rrk-object-based-media-usecase/>

What our I-D Proposes

- Take the agreed concept and functional components of the CATS WG framework, and then:
 - Use OTN to complement packet-based CATS
 - Combine compute metrics with optical-layer metrics
 - Establish end-to-end “hard-isolation” for demanding service flows.
- Service flows can be mapped into optical containers such as:
 - ODUk, fgOTN
- Path selection can consider:
 - deterministic path latency
 - wavelength continuity constraints
 - optical link performance
 - compute resource state

CATS & OTN Framework and Workflow

- Key building blocks
 - Service Sites / Service Instances / Service Contact Instances
 - C-SMA for service and compute metrics
 - C-NMA for optical network metrics
 - C-PS for service/path selection
 - C-TC for traffic classification
 - CATS-aware OTN edge nodes
- How would the service be instantiated?
 - Service identified by CS-ID
 - Metrics distributed to decision points
 - Best instance and path selected
 - Traffic classified and mapped into optical transport
 - Traffic delivered to the selected service instance

What would we like from CATS WG?

- Need deeper thinking on
 - Deployment decision making based on metrics
 - Compute side: service status, GPU or compute load, memory availability
 - Optical side: latency, wavelength or timeslot availability, optical link quality
 - Operational questions
 - Centralised vs distributed deployment
 - Update frequency and dampening
 - OAM and observability
 - Interaction with ACTN, PCE, and optical controllers
 - Security and confidentiality of metrics and topology data

Next Steps

- We believe OTN is a useful applicability case for CATS for deterministic, high-bandwidth services, and we welcome feedback and contributions.
- Request WG members to review and comment.
 - Is OTN a useful CATS applicability case?
 - Is the problem statement clear?
 - Are the functional roles and workflow sensible?
 - Are the right compute and optical metrics identified?
 - What are the main gaps around scaling, OAM, security, and controller interaction?