Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 7, 2020

J. Dang
B. Liu
Huawei Technologies
G. Yang
China Telecom
K. Lee
LG U+

November 4, 2019

Instant Congestion Assessment Network (iCAN) for Traffic Engineering draft-liu-ican-01

Abstract

This draft proposes a new technology named iCAN (instant Congestion Assessment Network), which represents a set of mechanisms running directly on network nodes. These mechanisms allow the nodes adjusting the flows' paths based on real-time measurement of the candidate paths. The measurement is to reflect the congestion situation of each path, so that the nodes could decide which flows need to be switched from a path to another.

This is something that current TE technologies can hardly achieve. In current TE, the paths are usually planned in a centralized controller, which is far from the data plane, thus neither be able to assess the real-time congestion situation of each path, nor able to assure the data plane always go as expected (especially in SRv6 scenarios). In a result, traditional TE is not able to adjust the flow paths in real-time to fit for the change of traffic instantly.

iCAN can work with traditional TE perfectly: the controller plans multi-path transmission in relatively long period (e.g. minutes), and iCAN does the flow path optimization in a much shorter interval (e.g. milliseconds).

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on May 7, 2020.

Table of Contents

1. Introduction

Traditional IP routing is shortest path based on static metrics, which can fulfil basic requirement of connectivity. MPLS-TE brings the capability of utilizing non-shortest paths, thus traffic dispatch is doable; however, MPLS-TE in only a complementary mechanism because of the scalability issue. Segment routing provides even more flexibility that paths could be easily programmed; and along with the controller, it could be scaled.

However, the above mentioned mechanisms all run in the control plane, which implies that they are not able to sense the data plane situation in real-time, thus they are mostly for relative static planning/controlling (minuets, hours or even day-level) of network traffic and not able to adapt to the microscopic traffic change in real-time (e.g. mili-second level). So, in real bearer networks (metro, backbones etc.), it is always underload so that the redundant resources could tolerant the traffic burst, results in a significant waste of network resources.

This draft proposes the iCAN (Instant Congestion Assessment Network) architecture to achieve autonomous adapt to traffic changes in real-time in terms of switching flows between multiple forwarding paths. iCAN includes following mechanisms:

- A mechanism between ingress and egress nodes to assess the path congestion situation in RTT level speed, to recognize which paths are underload and which are heavy loaded.

- Recognizing big flows and small flows in the device, in real time

- Ingress node dispatches flows to multiple paths, to make load balance, or to guarantee SLA for specific flows

Use cases, scenarios and implementation candidates of iCAN are also discussed in this draft.

2. Problem Statement

2.1. Background Problems

2.1.1. Latency issue

New services like AR/VR would require strictly low lantency which would be a big challenge to current network. Other than fixed lantency caused by the forwarding devices' internal processing and transmission time on wire, the prominent factor of lantency is the queuing time caused by congestion. Thus, to controll the latency of a certain path, is mostly to controll the congestion.

2.1.2. Microburst issue

The network users/services are so comprehensive that the traffic model is always uncertain, which results in high bandwidth peak-to-average ratio. In other words, real traffic could often changes dramatically in second or even millisecond level. Thus, even if the bandwidth of paths seem all good in a network management system, there might be congestion happening in real forwarding plane that just could not detected by the management system.

2.2. Gap Analysis

2.2.1. Load balancing

In real networks, the traffic is usually un-balanced. Some links might be idle while some might be heavily congested. The partial congestion in the network has affected the bandwidth planning of the network. Ideally, the bandwidth planning of the network generally depends on the average bandwidth of the link, however, in reality the bandwidth planning more tends to peak bandwidth of the link in order to guarantee the business experience. Especailly for low-latancy servie, since it is very sensitive to congestion, the result is that operators would have to increase investment for network expansion.

Although there are mechanisms against load balancing issue, the real result is usually not as expected.

- Device-level Load Balance (e.g. ECMP)

  1) Not recognize flows' amount. ECMP is mostly deployed in a per-flow manner. Since the devices cannot recognize each flow's amount, they just fork the flows based on the numbers of flows, not the exact amount of flows. Thus, the result of ECMP could be unbalanced.

2) Not consider congestion status of E2E paths. ECMP only cares about the next hop, thus, if one remote link that is on the path is already highly congested, the device would still fork flows to the path due to the ECMP local decision.

- Network-level load balance (e.g. UCMP)

1) Lack of data plane mechanisms to ensure the real sharing ratio between multiple paths. Again, since there is no mechanims to recognize flows' amount, the controller just could not make sure whether the traffic is forked exactly as it expected.

2) Slow reaction: The global path optimization architecture of SDN can sense traffic changes by measuring protocol, but the overall feedback speed is relatively slow.

2.2.2. SLA assurance

- QoS mechanism

When congestion occurs, QoS scheduling will be trigger. The QoS mechanism in a network element selects a portion of the packets into the buffer. The purpose of delaying the retransmission is to increase the network capacity. Even if low-latency services are placed in high-priority queues to avoid additional delays, high-priority queues are still in buffers. Once there are more high-priority services, the packets will still enter the buffer for a period of time.

2.2.3. High availability

Current networks highly relies BFD for high availability. The problems of BFD are:

1) Complex configurations

2) Can only detect path on/off, not able to detect path quality deterioration

3) Cannot distinguish individual paths in multi paths

3. iCAN Architecture and Key Technical Requirements

3.1. Architecture

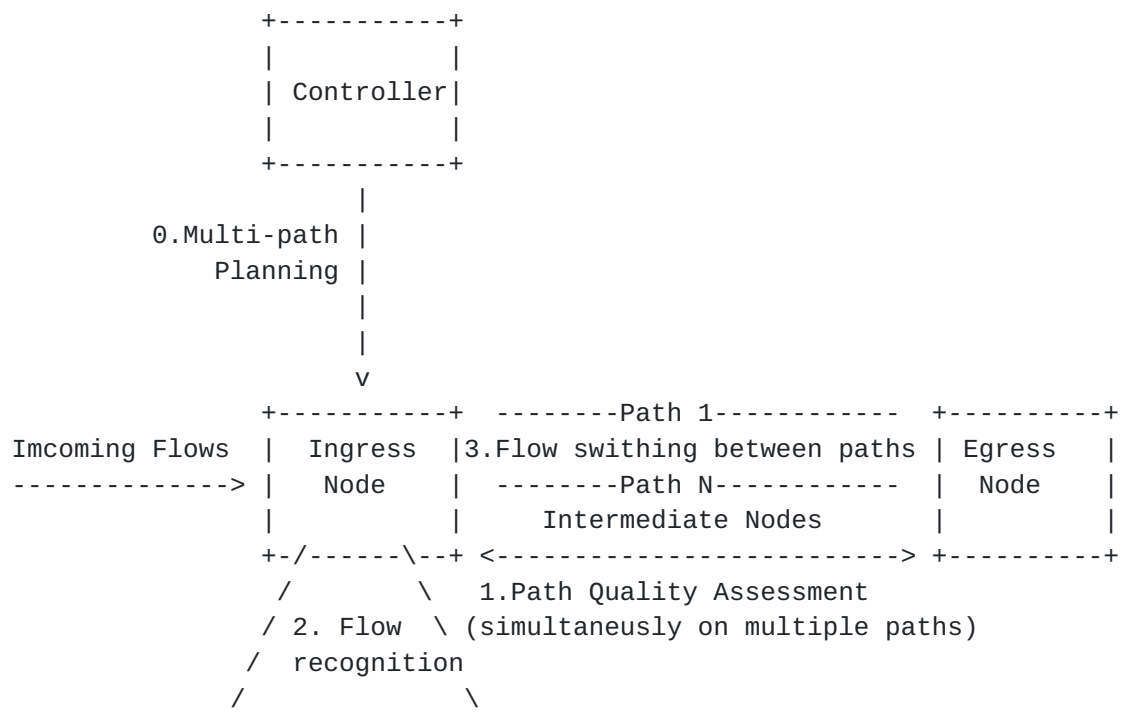

Figure-1: iCAN architecture

As above figure shows, there are 3 entities:

1. Controller

   - Responsible for planning multiple paths for a set of flows that could be aggregated to a pair of Ingress/Egress routers.

   - After delivering the planned paths to the ingress router, the controller would need nothing to do.

2. Ingress node:

   - Serves as a local "controller" for the iCAN system.

   - Responsible for triggering the path congestion assessment, which is coordinated with the egress router through a measurement protocol.

   - After getting the assessment results, the ingress router would calculate which flows need to be switched to a different path, in order to make the paths load balanced or to assure the transport quality of a certain of important flows.

- In order to do the path switching calculation, the ingress router needs to recognize the TopN flow passing by it, since switching the big flows would make the most effort.

3. Egress node:

- Only needs to coordinate with the ingress router to do the path assessment.

4. Intermediate nodes (optional)

- If the intermediate nodes are allowed to participate in iCAN, they can provide useful information (e.g. link utilization) for better measurement of path quality. (TBD)

3.2. Key technical requirements

3.2.1. Path quality assessment

- Req-1: the assessment MUST reflex the congestion status of the paths.

- Req-2: the assessment SHOULD be done within a RTT timeslot. Since iCAN is to adapt the traffic change in real-time, the assessment needs to be done very fast.

- Req-3: the assessment MUST be done for multiple paths between the same ingress/egress routes simultaneously.

Candidate solutions:

- For Req-1:

  In the draft [I-D.dang-ippm-congestion], a new metric is proposed to indicate the congestion degree of a certain path. A specific value could be calculated according to a certain method (described below), so that the congestion situation of different paths could be quantified and compared.

- For Req-2 & 3:

  In the draft [I-D.dang-ippm-multiple-path-measurement], a specific method is proposed to calculate the congestion degree as decribed above. The proposed method supports measuring multiple paths simultaneously, which needs a special mechanism to allign measurements of different paths to the same time slots.

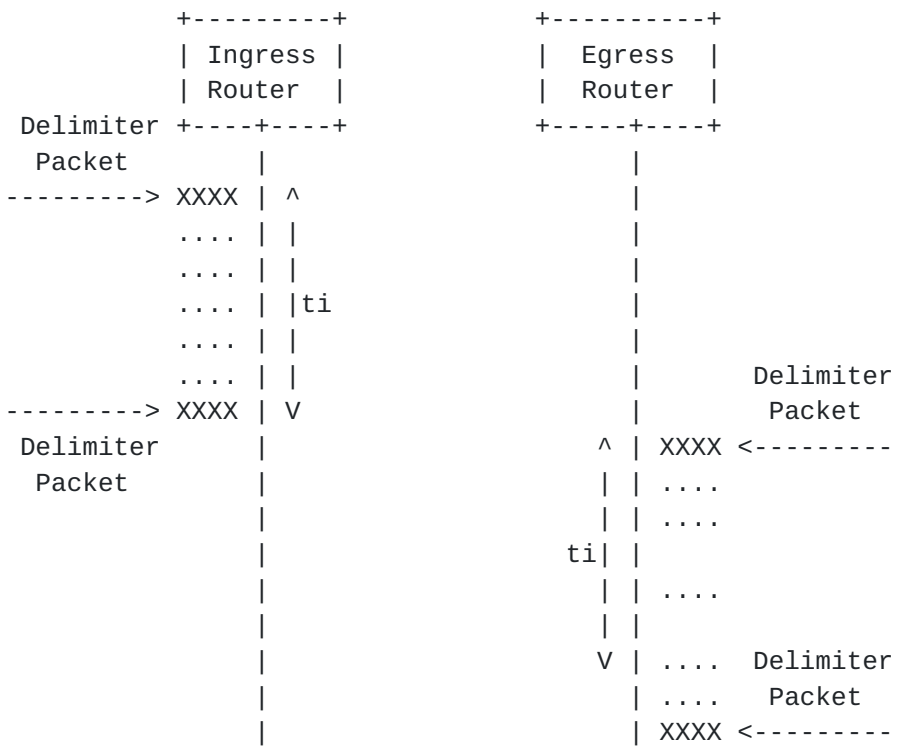

Figure-2 Path congestion assesment between ingress/egress routes

As the figure shows above, the ingress router inserts the delimiter packet in a fixed interval "ti" (e.g. 3.3ms); at the same time, the ingress router would count how many packets are sent during the interval. When the egress routers recieves the same delimiter packet, it also starts count the packet number during the same interval "ti". Because of the path congestion, the gap between the receiving packets might probably differs from the gap between sending packets, the difference is just the key infomation for estimating the congestion degree of the path. The egress router would ignore the gap difference, and just return the packet number to the ingress router when the "ti" time is up.

The ingress/egress router could also count the packets number between two delimiter packes, so that the counts could be compared for detecting packet loss. If packet loss happens, it should be specially taken into consideration by the path congestion degree calculation.

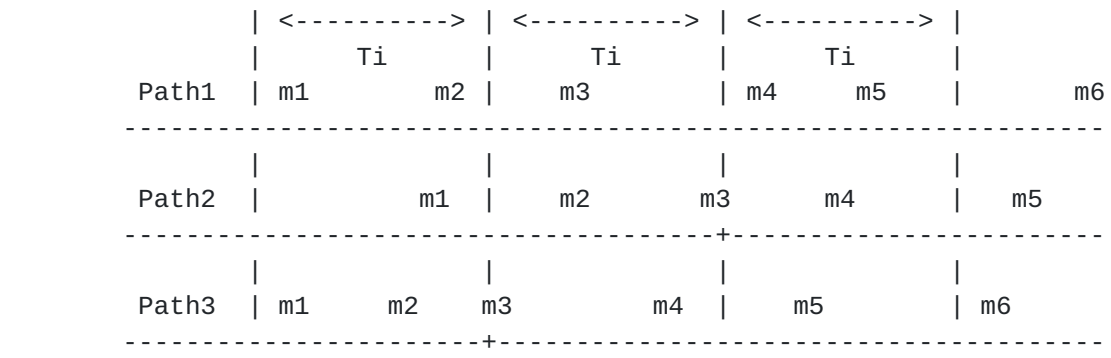

Figure-3 Multiple path measurement allignment

As the figure shows above, m1-mN indicates every measurent made on each path, as every "ti" interval, the egress would return the packet count to the ingress router for path congestion degree measurement (as described above). Due to path congestion, the packets sent by the egress router would arrive at the ingress router at different intervals other than "ti" time. Thus, for path congestion degree comparison between multipla paths, it needs a longer "Ti" interval (e.g. 10ms) to make sure that there would be at least one measurement completed most of the time.

3.2.2. Recognition and statistic of flows in devices

- Req-1: the device SHOULD be able to recognize TopN big flows within a timeslot.

- Req-2: the device MAY need to statistic all flows' amount within a timeslot.

3.2.3. Flow switching between paths

- Req-1: the device SHOULD be able to recognize flow let. The flow switching is done from the next flow let.

- Req-2: the device MAY need to actively generate gap to artificially create flow let. If the flow needs to be switched immediately, then the device would need to make the gap, to avoid out-of-order packets arriving to the destination through multiple paths.

- Req-3: the device SHOULD avoid oscillation of frequently switching flows from one to another.

- Req-4: multiple ingress devices SHOULD be able to coordinate so that they won't switch flows to the shared path at the same time, to avoid potential congestion in the shared path.

4. Use Cases and Scenarios

4.1. Network load balancing

Background problem: traffic is not balanced in current metro network.

While some links are heavily loaded, others might be still lightly loaded: unbalance could lows down the service quality (e.g. SLA could not be guaranteed in the heavily loaded links/paths); unbalance could lows down the network utilization ratio (normally with 30%, e.g. a 100G physical capacity network can only bear at most 30G traffic, a huge waste of network infrastructure).

iCAN could be used for load balance among the multiple paths between a pair of ingress/egress nodes. Once the network is balanced, the real throughput of the network could be elevated significantly. A couple of targeted scenarios are described below.

4.1.1. Multiple-access in backbone networks

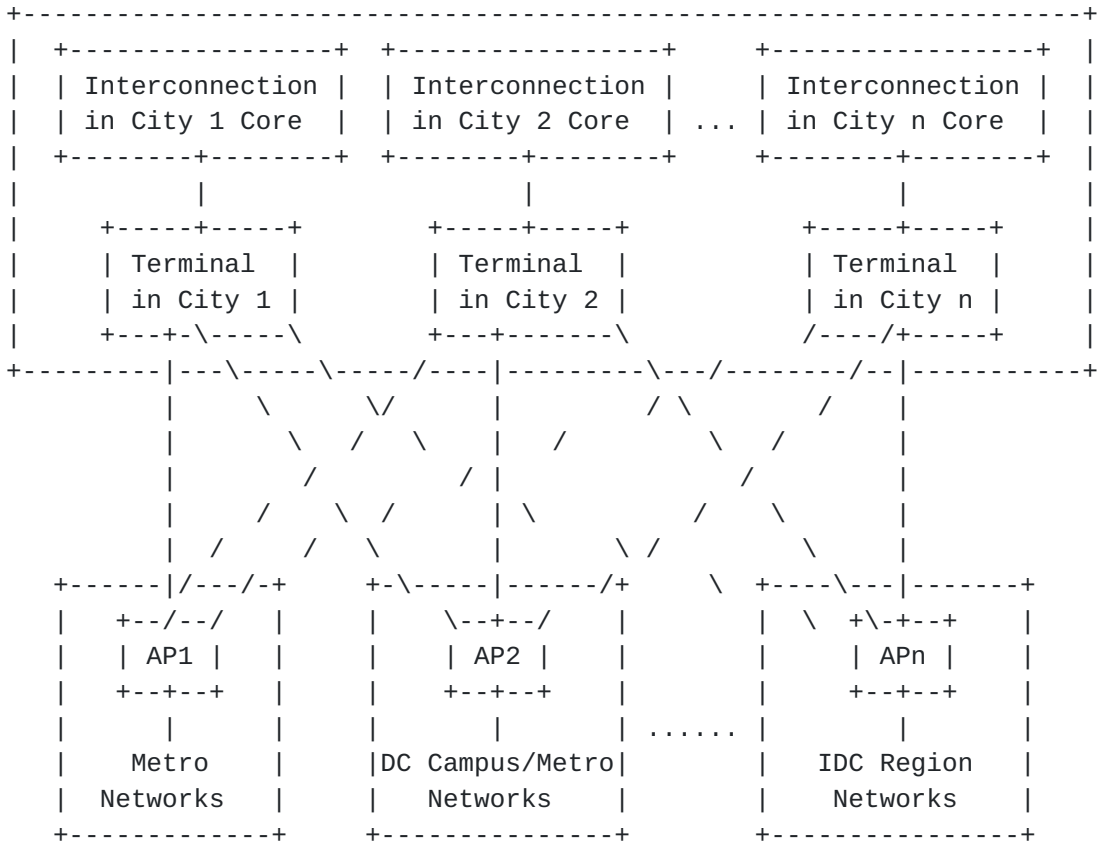

Figure-4: Multiple-access in backbone networks

As Figure-4 shows above, at the edge of the backbone, there are APs (Acess Points) that attached with different networks. The APs connect to the TPs (Terminal Points), via which the traffic from the APs could be exchanged to other networks (e.g. other ISPs). In order to get a high quality connection, each AP could attach to multiple TPs, thus, the IP tunnels are constructed in a mesh manner.

Between APs and TPs, BGP is running for traffic routing. The BGP policies are always static, since it is very error-prone for manully adjusting BGP policies. Thus, the traffic running among the paths between APs and TPs are always un-balanced. With iCAN runs as on top of the mesh tunnels, the un-balanced issue could be solved in principle.

4.1.2. Multiple paths in metro network

Similar as the iCAN architecture showed as Figure-1, in metro networks, there might be multiple paths between ingress/egress router pairs. Thus, iCAN could be used to increase the throughput without hardware expansion.

4.2. SLA assurance

Since iCAN could switch flow in real-time, it can guarantee a set of important flows. Once the path which carries the important flows is to be congested, the other flows could be switched to alternative paths, and the important flows would stablely running in the original path.

(More content TBD)

4.3. Fine-Granularity reliability

Traditional reliability protocols such as BFD, can only assess the link on or off. With the path congestion assessment ability, iCAN could also asses the quality degradation.

(More content TBD)

5. Implementation with Underlay Technologies

5.1. iCAN with SRv6

- SR Multiple Explicit Paths

  For example, there are 3 paths between the ingress and egress nodes, and the multi-path is defined as a SR-List containing LSP1/2/3.

The probe message detects the congestion status of the three SR-list paths. The edge device adjusts the load balancing between the three paths according to the congestion status of the three SR-lists, and switch the flows from the path with a high congestion to the path with a low congestion.

- SR Multiple Explicit+Loose Paths

  In loose path scenario, there needs to be an additional approach to probe the specific paths of a SR tunnel. After that, operations on the probed paths are the same as explicit path scenario.

5.2. iCAN with VxLAN

TBD.

5.3. iCAN with MPLS/MPLS-TE

TBD.

6. Standardization Requirements

1. Multi-path Planning (North Interface between Controller and devices)

2. Path Congestion Assesment (Horizontal Interface between devices), mostly regarding to Req-1&2&3 described in Section 3.2.1 .

3. Flow Switching Negotiation (Horizontal Interface between devices), mostly regarding to Req-3&4 described in Section 3.2.3 .

(More content TBD.)

7. Security Considerations

TBD.

8. IANA Considerations

TBD.

9. Acknowledgements

Very valuable comments were from Shunsuke Homma, Mikael Abrahamsson and Bruno Decraene.

A commercial router hardware based prototype had been implemented to prove the machinisms discussed in the document are workable.

Conventions: [RFC2119][RFC2629]

10. References

10.1. 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>.

[RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, DOI 10.17487/RFC2629, June 1999, <https://www.rfc-editor.org/info/rfc2629>.

10.2. Informative References

[I-D.dang-ippm-congestion] Dang, J. and J. Wang, "A One-Path Congestion Metric for IPPM", draft-dang-ippm-congestion-01 (work in progress), March 2019.

[I-D.dang-ippm-multiple-path-measurement] Dang, J. and J. Wang, "A Multi-Path Concurrent Measurement Protocol for IPPM", draft-dang-ippm-multiple-path-measurement-01 (work in progress), March 2019.

Authors' Addresses

Joanna Dang
Huawei Technologies
Q27, Huawei Campus
No.156 Beiqing Road
Beijing 100095
P.R. China

Email: dangjuanna@huawei.com

Bing Liu
Huawei Technologies
Q27, Huawei Campus
No.156 Beiqing Road
Hai-Dian District, Beijing 100095
P.R. China

Email: leo.liubing@huawei.com

Guangming Yang
China Telecom
109 Zhongshan W Ave
Guangzhou 510630
P.R. China

Email: yangguangm@chinatelecom.cn

Kyungtae Lee
LG U+
71, Magokjungang 8-ro, Gangseo-gu
Seoul
Republic of Korea

Email: coolee@lguplus.co.kr