In-situ Flow Information Telemetry for EVPN

Tianran Zhou
Standard Expert, Huawei
zhoutianran@Huawei.com

Take a long time to find why packet loss or delay?
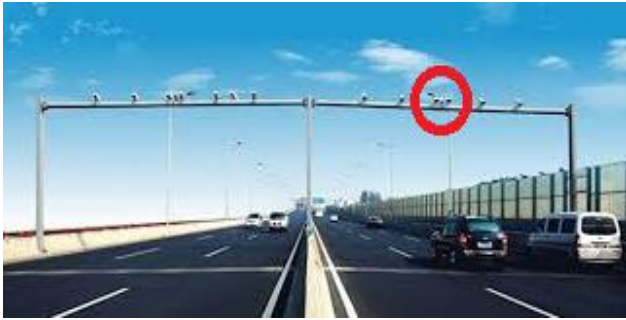

Active OAM

In-situ Flow Information Telemetry

Advantages:
- Fine grained **Flow** SLA monitoring
- Measure the **Real** user traffic
- **Per packet** monitoring
- Abundant data plane information to enable big data **AI**

- The sequence of devices the packet visited along its path.
- The set of rules it matched at every device along the way.
- The time it spent buffered in every switch, to the nanosecond.
- The flows that the packet shared each queue with.
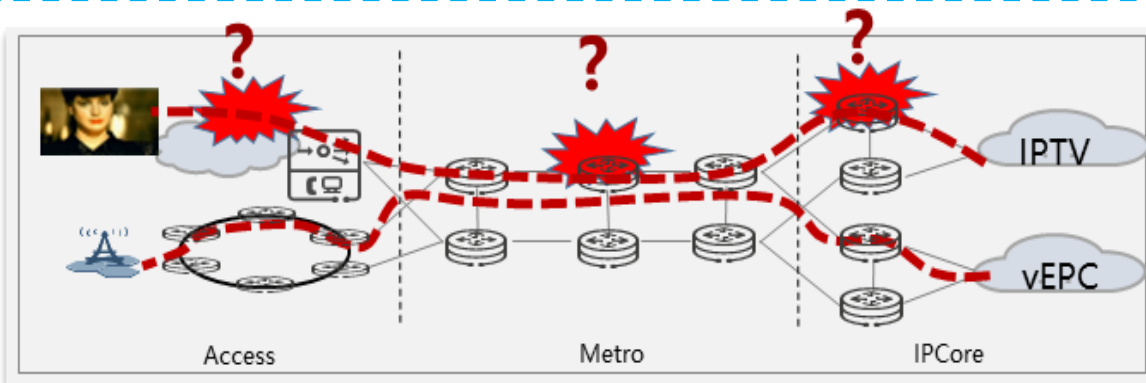
Silent Packet Loss: random packet loss or only drop small packets. Both BFD and TWAMP cannot detect while the service is damaged.

SLA Report: High value customer do not trust the SLA measurement by active probes. Credible SLA report can be provided by IFIT.

Flow Monitoring: Hop by hop SLA measurement per user, per service, or per application.

HUAWEI

Use Case 1: OTT Service is Hard to Operate

Game quality is bad: hard to locate, time consuming, hard to optimize

Device CRC alarm; SLA monitoring is normal (15min sampling rate) Ignored

Game experience degraded suddenly

Expended the link capacity for more traffic.

iSStar/Active probe

Identified an abnormal 100G interface with CRC error. The interface is not well contacted.

June 28 August 29 September 7

2 Months without any action

1 Week Identify the issue hop by hop

Analyze
- wireless
- IPRAN
- Transport
- Metro

Identify issues
- Identify path
- Analyze log
- Field inspect
- …
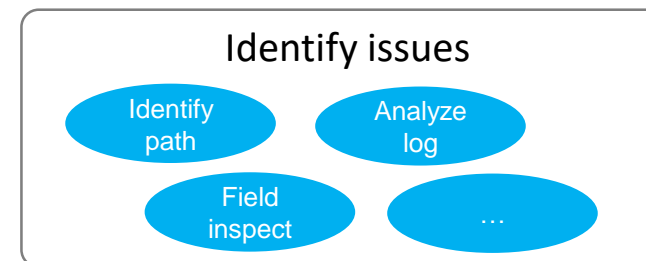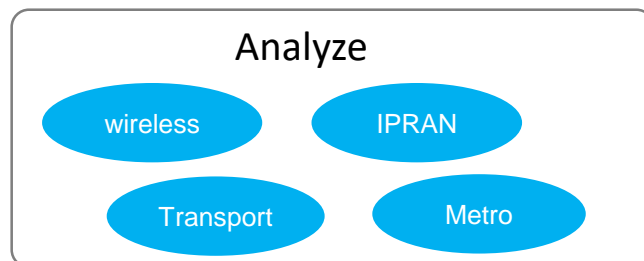
End to end QoS optimization

- Cross various department including access, transport, core network, Internet, etc.

- A lot of man-month cost.

HUAWEI

Use Case 2: Hard to Detect and Locate Silent Forwarding Issue

Failure Classification

- Hardware issue
- Config issue
- Forwarding issue
- Protocol issue

Huawei report
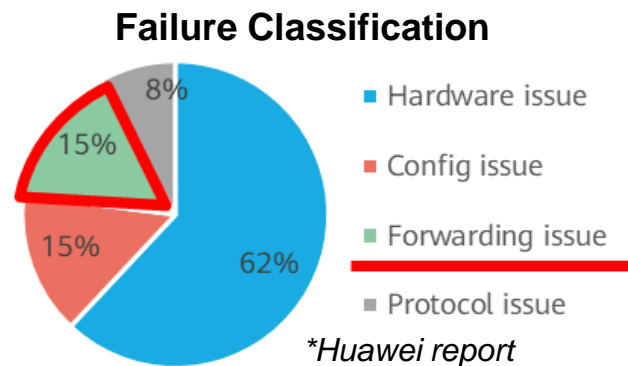
Around **15%** forwarding issue

Will cost **80%** operation time

Silent Forwarding Issue: there is no system alarm when the service experience degrade.

An abnormal chip impacted **6000** customers, the impact lasted for **7** hours.

Traffic

| Get the first complaint. | A lot of complaints from customers on IPTV and Internet. | Operator and vendor work together. | Back to normal. | Failures happen again. |

14:39   16:30   17:00   17:26   19:26   22:00
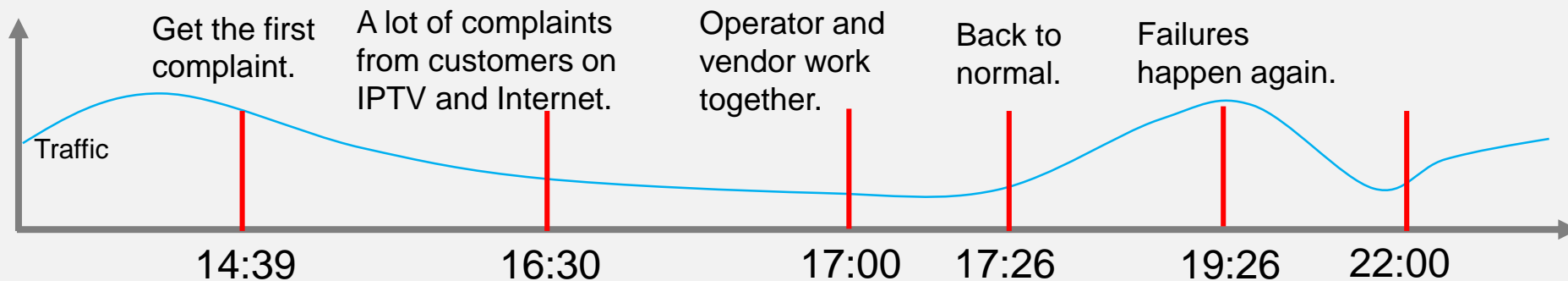
The first defense line:
Both device and OSS get no alarm

The second defense line:
No abnormal by Manual inspection

HUAWEI

Network Visualization with a Holistic Network Telemetry Framework

RFC9232: Network Telemetry Framework

Query		Subscription		Performance
Poll based		Push based		
Reactive	→	Streaming		Efficiency
Unstructured		Structured		
Fixed/Limited Data		Model-driven		Automation

Network Operation & Management Applications

Streaming Analytics ⟷ Data Lake

Control Plane Telemetry ⟷ Management Plane Telemetry

Data Plane Telemetry

IFIT

External Data and Event Telemetry
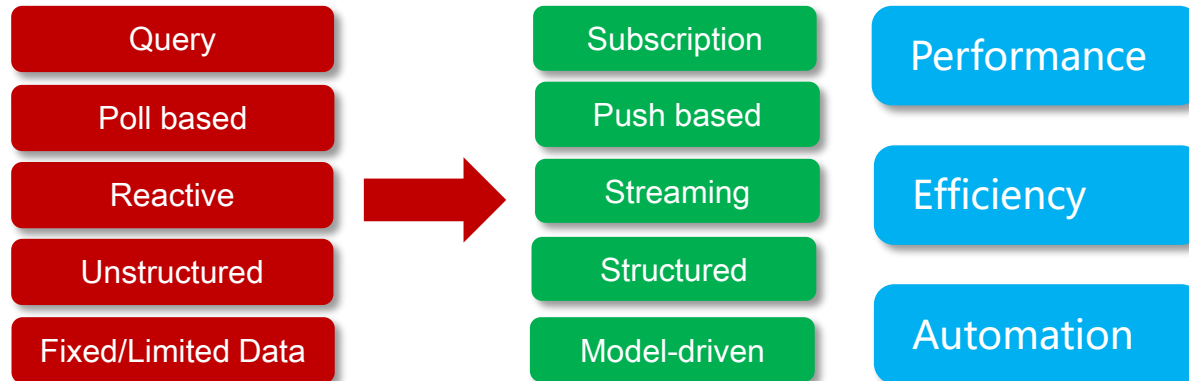
Network Visibility Presents Multiple Viewpoints

- Device viewpoint: takes the network infrastructure as the monitoring object. E.g., network topology, device status and statistics
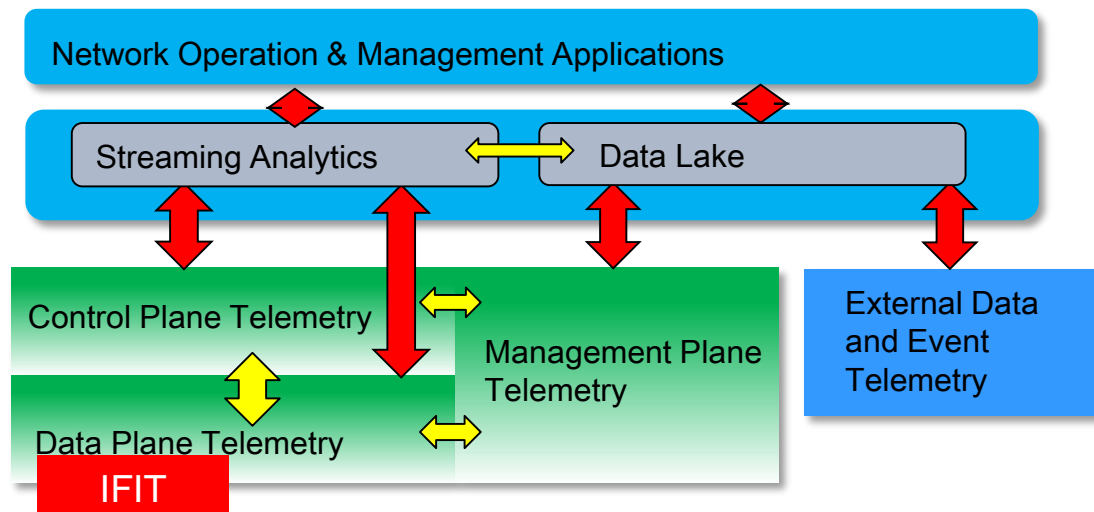- Traffic viewpoint: takes the flows or packets as the monitoring object. E.g., traffic quality, flow path
- Switch viewpoint and/or correlate service experience with network

Elastic Network Telemetry

- Routine network monitoring covers the entire network with low data sampling rate.
- When issues arise or trends emerge, the telemetry data source can be refocused and the data rate can be boosted.

Efficient Data Fusion From Multiple Sources

- Reduce the overall quantity of data
- Improve the accuracy of analysis

HUAWEI

IFIT-based Reactive Telemetry Framework

On demand underlying technique selection

Application and Management System

Intelligent Flow, Packet&Data Selection

Controller

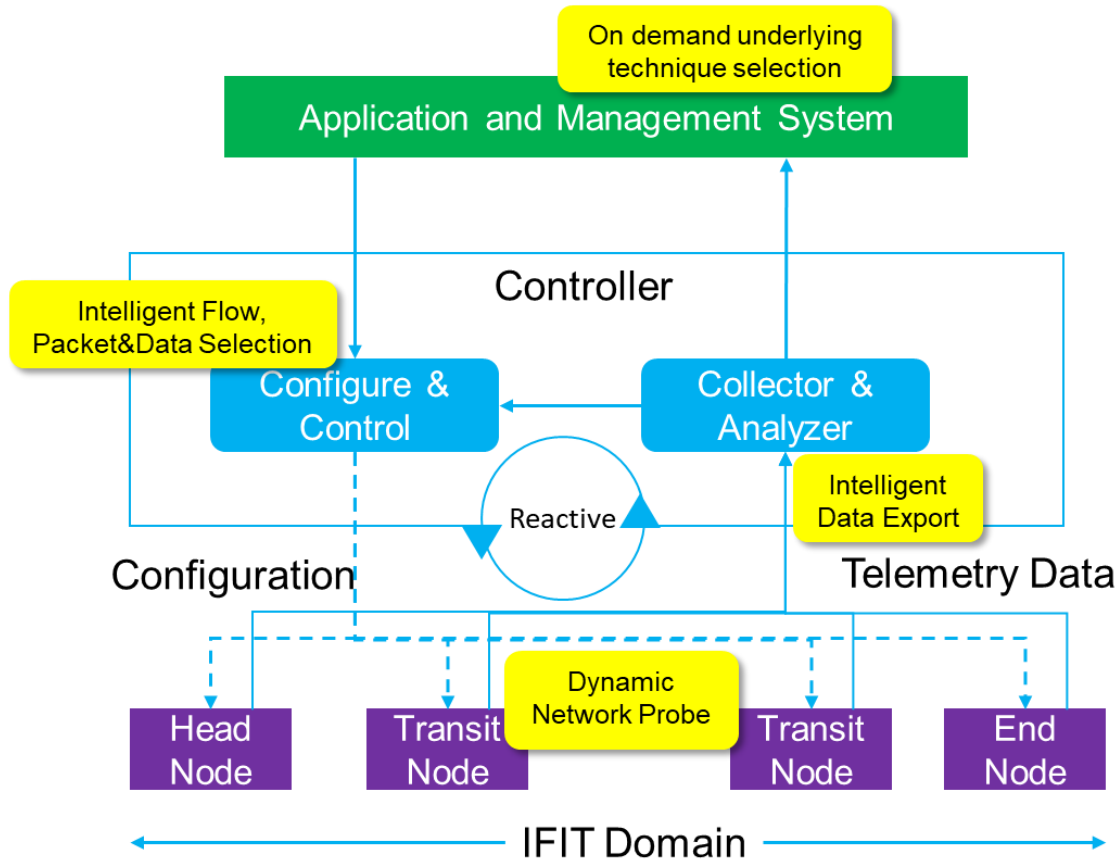Configure & Control

Collector & Analyzer

Reactive

Intelligent Data Export

Configuration

Telemetry Data

Dynamic Network Probe

Head Node

Transit Node

Transit Node

End Node

IFIT Domain

On-demand Underlying Technique Selection:

- Postcard vs Passport
- Hop by hop vs Edge to edge

Intelligent Flow, Packet, and Data Selection

- Select some specific service flows, packets or data according to service or operation and maintenance requirements
- Variable sampling frequency

Intelligent Data Export

- perform de-redundancy and compression processing of the exported data

Dynamic Network Probe

- enables probes for customized data collection in different network planes
- can be loaded into the data plane through incremental programming or configuration.
- In-situ Flow Information Telemetry

IETF:

- RFC9341:Alternate-Marking Method
- RFC9342: Clustered Alternate-Marking Method
- RFC9343: IPv6 Application of the Alternate Marking Method

ETSI:

- GR ENI 012: Reactive In-situ Flow Information Telemetry

HUAWEI

IFIT based Solution: High Precision Service Level Measurement

	Packet Loss	Delay(us)	Jitter(us)
Flow 1	🔴 1%	🔴 50,000	🔴 3,000
Flow 2	🟢 0%	🔴 30,000	🟢 100
Flow 3	🟢 0%	🟢 1000	🟢 10

iMaster NCE-IP

- **Location of the packet loss:**
  Analysis of postcard based statistic from every node.
- **Per-hop delay and jitter:**
  Analysis of link and node delay based on per-hop timestamp.
- **Route playback:**
  Display the real path of each traffic flow.

Ingress → Transit → Egress → Internet / Cloud

Ingress:
- Customize the measurement.
- Attach the IFIT header to flow objects.

Transit/Egress:
- Automatic statistic: packet count, delay.
- Periodically report the result based on streaming telemetry.

High Precision
- ➢ Meets the strict packet loss detection requirements (10^{-6}) of Cloud VR services.
- ➢ per-packet loss detection.

Abundant Scenarios and Metrics
- ➢ Metrics: Delay, Loss, Through put, Path
- ➢ Scenario: EVPN/L3VPN/L2VPN/SRv6/SR-MPLS/MPLS

Easy to Deploy
- ➢ Only configuration at the ingress, no need for transit and egress; On demand E2E/HbH.
- ➢ Bypass the legacy nodes for best compatibility.

HUAWEI

Packet Loss Measurement

iMaster NCE-IP

- **Packet loss[i] = Tx[i]–Rx[i]**
  - a). Packet loss in Period [i] = 5-5 = 0
  - b). Packet loss in Period [i+1] = 5-4 = 1
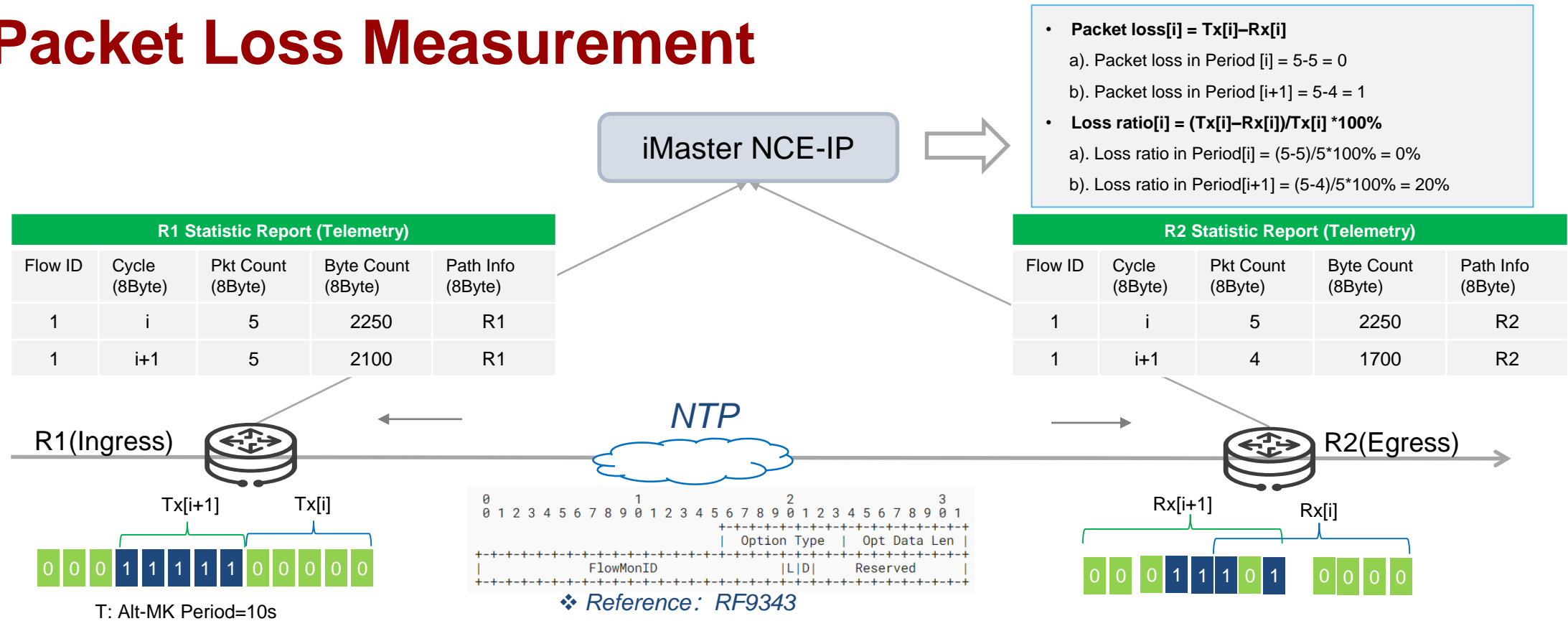- **Loss ratio[i] = (Tx[i]–Rx[i])/Tx[i] *100%**
  - a). Loss ratio in Period[i] = (5-5)/5*100% = 0%
  - b). Loss ratio in Period[i+1] = (5-4)/5*100% = 20%

R1 Statistic Report (Telemetry)

Flow ID	Cycle (8Byte)	Pkt Count (8Byte)	Byte Count (8Byte)	Path Info (8Byte)
1	i	5	2250	R1
1	i+1	5	2100	R1

R2 Statistic Report (Telemetry)

Flow ID	Cycle (8Byte)	Pkt Count (8Byte)	Byte Count (8Byte)	Path Info (8Byte)
1	i	5	2250	R2
1	i+1	4	1700	R2

R1(Ingress)

NTP

R2(Egress)

Tx[i+1]   Tx[i]

`0 0 0 1 1 1 1 1 0 0 0 0 0`

T: Alt-MK Period=10s

```
 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
                                +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                                | Option Type   | Opt Data Len  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            FlowMonID            |L|D|      Reserved           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

❖ *Reference：RF9343*
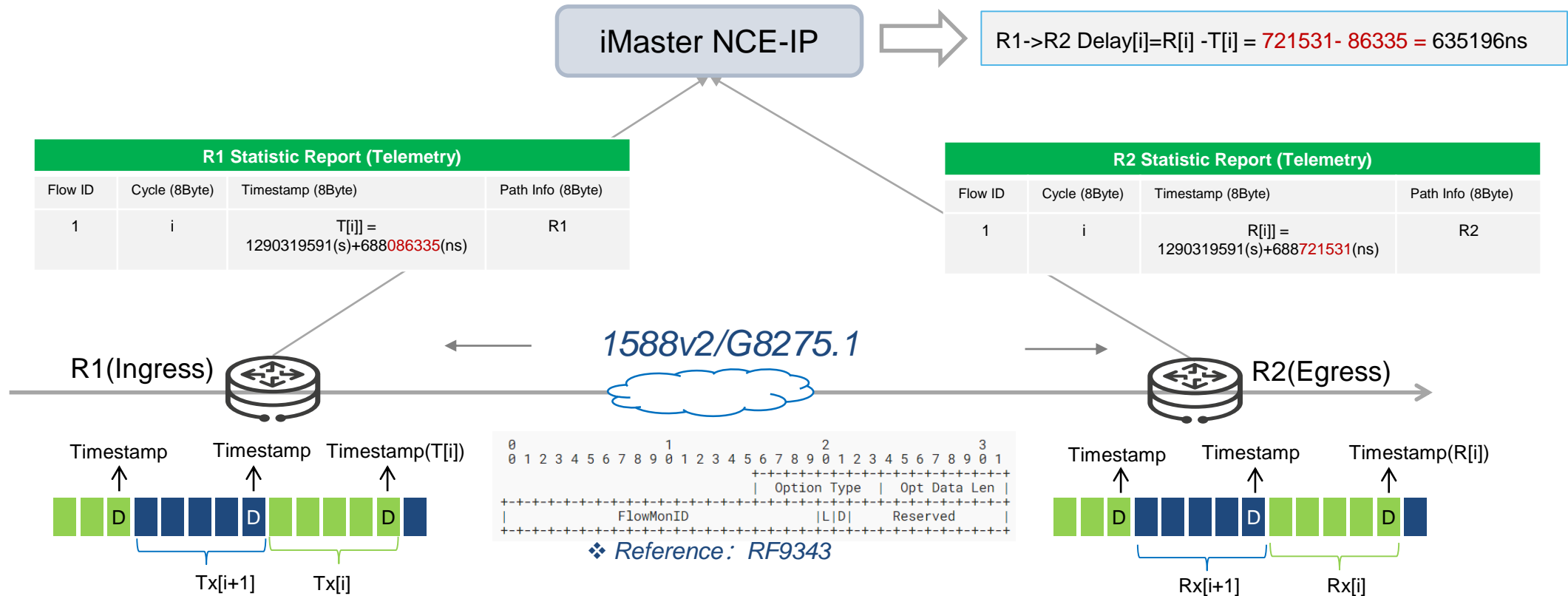
Rx[i+1]   Rx[i]

`0 0 0 1 1 1 0 1 0 0 0 0`

- **Alternate Marking:** Mark the L bit with 0 or 1 alternately by period.
- **Statistic:** Count the packet number and bytes received within each period.

  For example: packet number Tx[i]=5; Tx[i+1]=5

  period = 10s
- **Report Frequency:** once per period.

- **Statistic at Transit and Egress node:** increase 2/3 period to mitigate the miss order; count the packet number and bytes with color 0 or 1 received within each period.

  For example: count 0 marking packets in the first period till 10+2/3*10=16s;

  count 1 marking packets in the second period till 26s;

  Rx[i]=5; Rx[i+1]=4

❖ *Reference：Alternate Marking Method (RF9341)*

Delay Measurement

iMaster NCE-IP ➡ R1->R2 Delay[i]=R[i] -T[i] = 721531- 86335 = 635196ns

R1 Statistic Report (Telemetry)

Flow ID	Cycle (8Byte)	Timestamp (8Byte)	Path Info (8Byte)
1	i	T[i]] = 1290319591(s)+688086335(ns)	R1

R2 Statistic Report (Telemetry)

Flow ID	Cycle (8Byte)	Timestamp (8Byte)	Path Info (8Byte)
1	i	R[i]] = 1290319591(s)+688721531(ns)	R2

1588v2/G8275.1

R1(Ingress)   R2(Egress)

Timestamp Timestamp Timestamp(T[i])

```
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
                                +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                                |  Option Type  |  Opt Data Len |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            FlowMonID               |L|D|      Reserved        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

❖ *Reference：RF9343*

Tx[i+1]   Tx[i]

Timestamp Timestamp Timestamp(R[i])
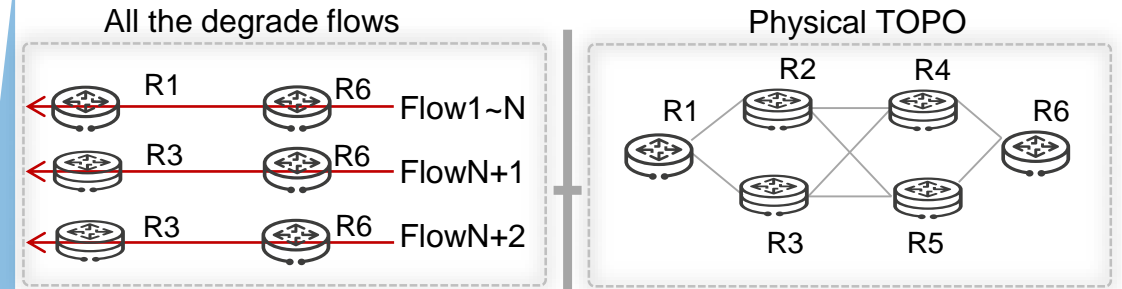
Rx[i+1]   Rx[i]

- **Marking:** Select one packet in each period, and mark D bit with 1.
- **Timestamp:** take the timestamp when the marking packet is received, and report to the controller with the period ID.

- **Statistic at Transit and Egress node:** take the timestamp when the marking packet is received, and report to the controller with the period ID.
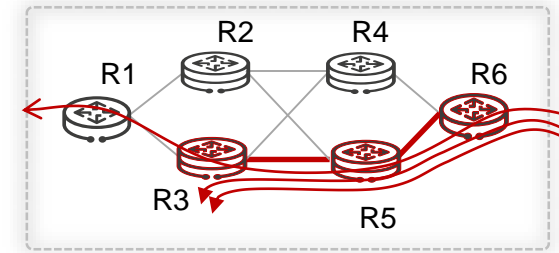
❖ *Reference：Alternate Marking Method (RFC9341)*
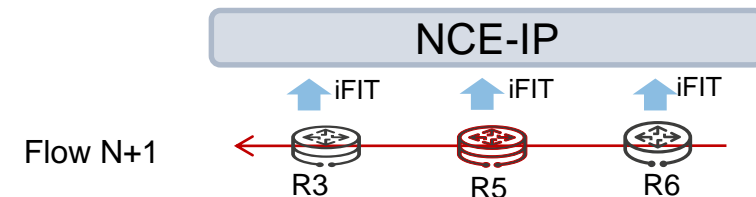
HUAWEI

Cluster Analysis for Degrade Services

Native AI
Cluster Analysis

NCE-IP

R2

R4

R1

Flow1~N

R3

Flow N+1

R5

R6

Flow N+2
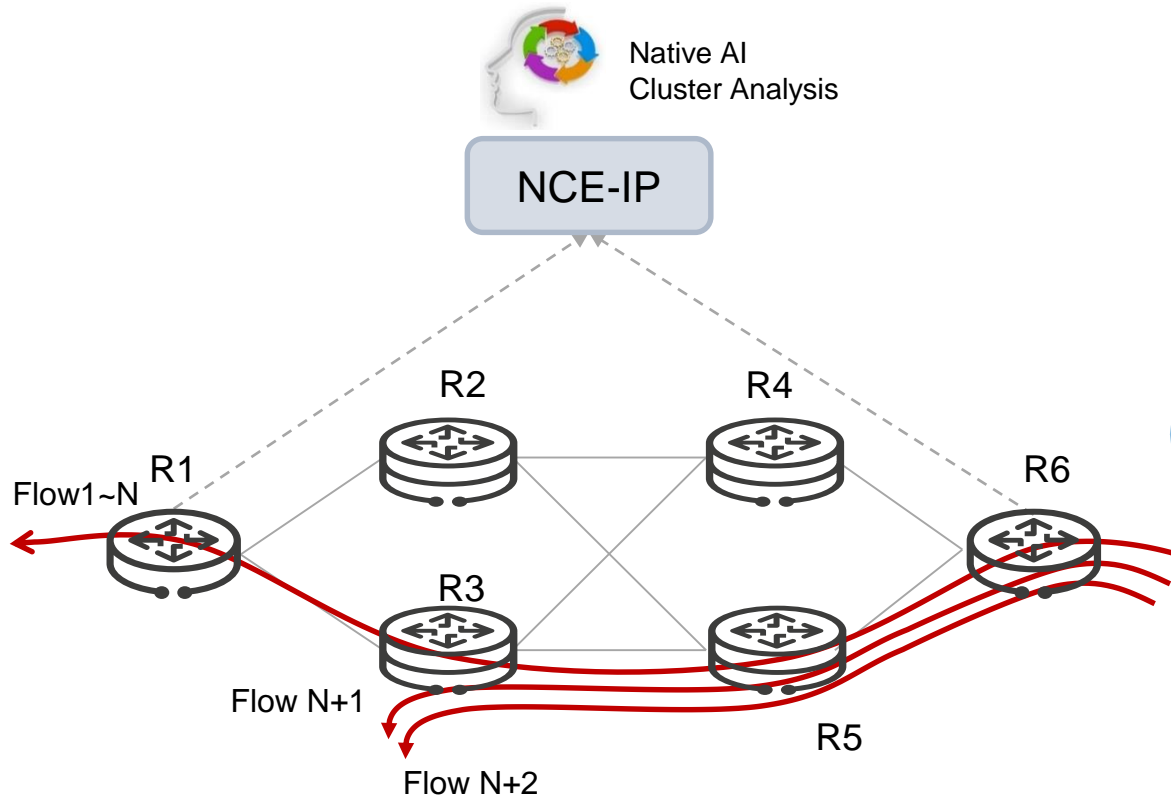
- **Use the cluster analysis to reduce the suspicious range.**
- **Do hop by hop measurement only for limited number of flows.**

All the degrade flows

R1   R6   Flow1~N

R3   R6   FlowN+1

R3   R6   FlowN+2

Physical TOPO

R2   R4

R1   R6

R3   R5

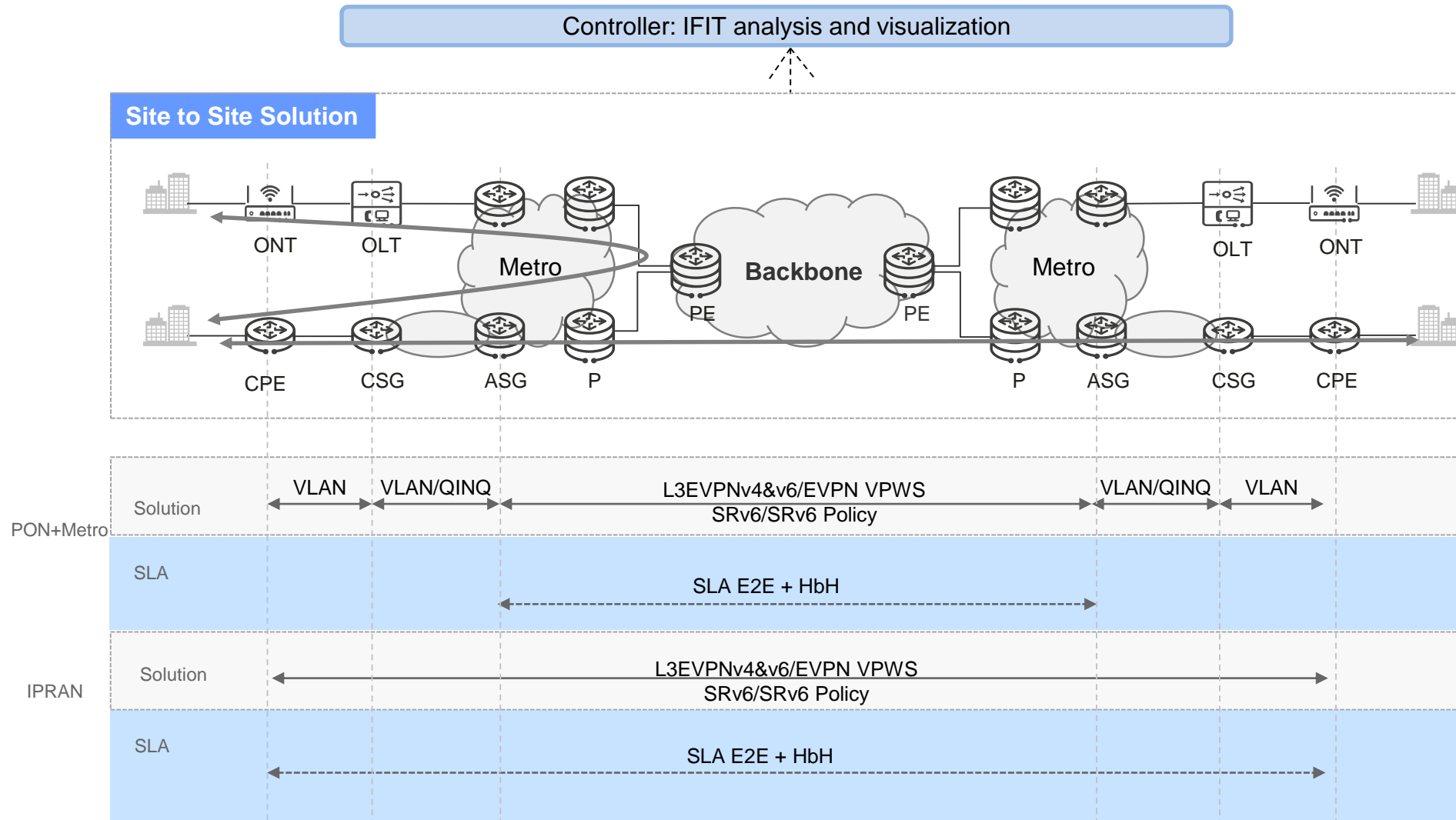Cluster all the degrade service for shared paths: R3~R6
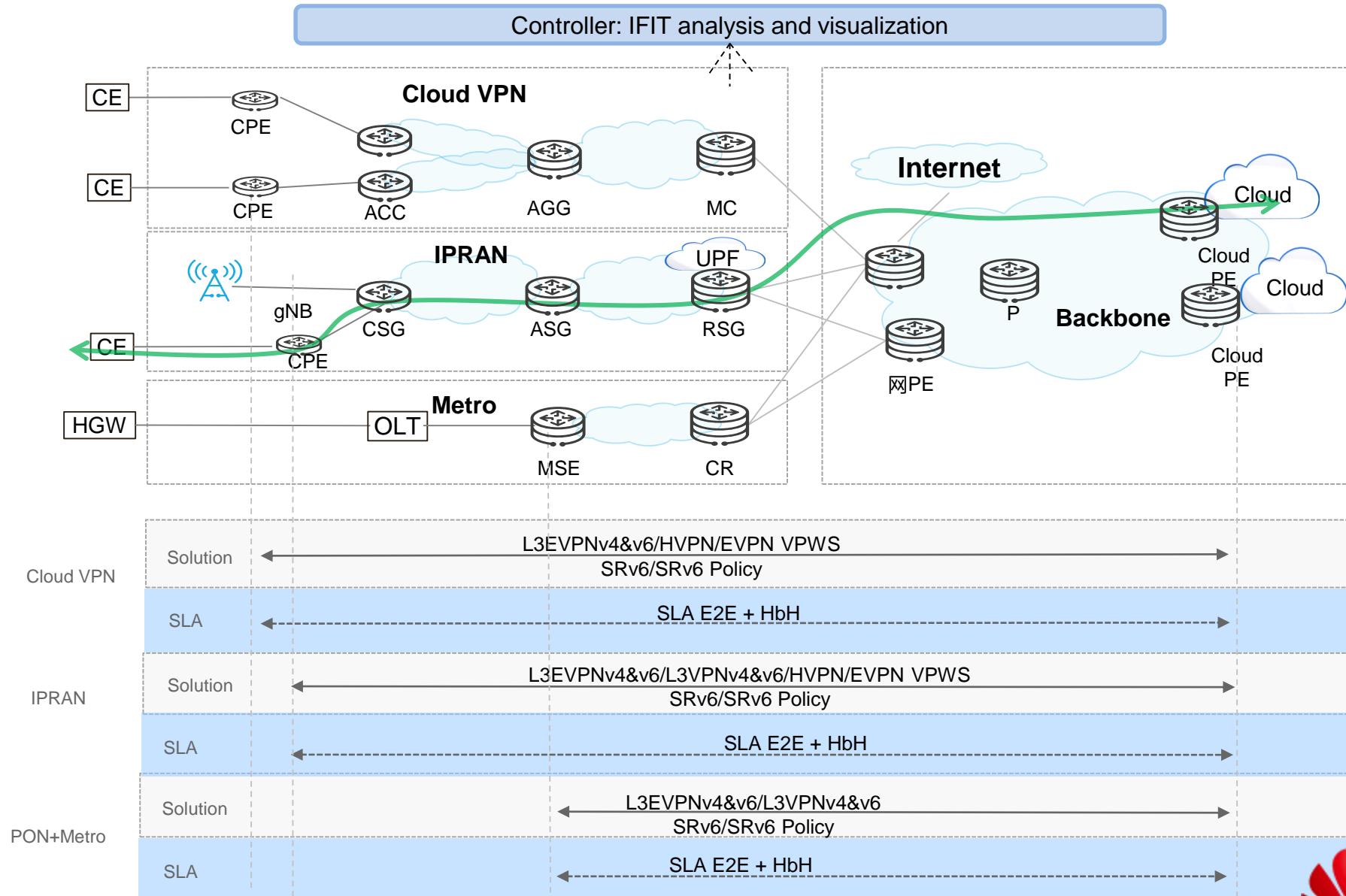
R2   R4

R1   R6

R3   R5

- Select one typical degrade flow from the cluster for hop by hop measurement. E.g., N+1.
- Locate the failure node. E.g., R5.
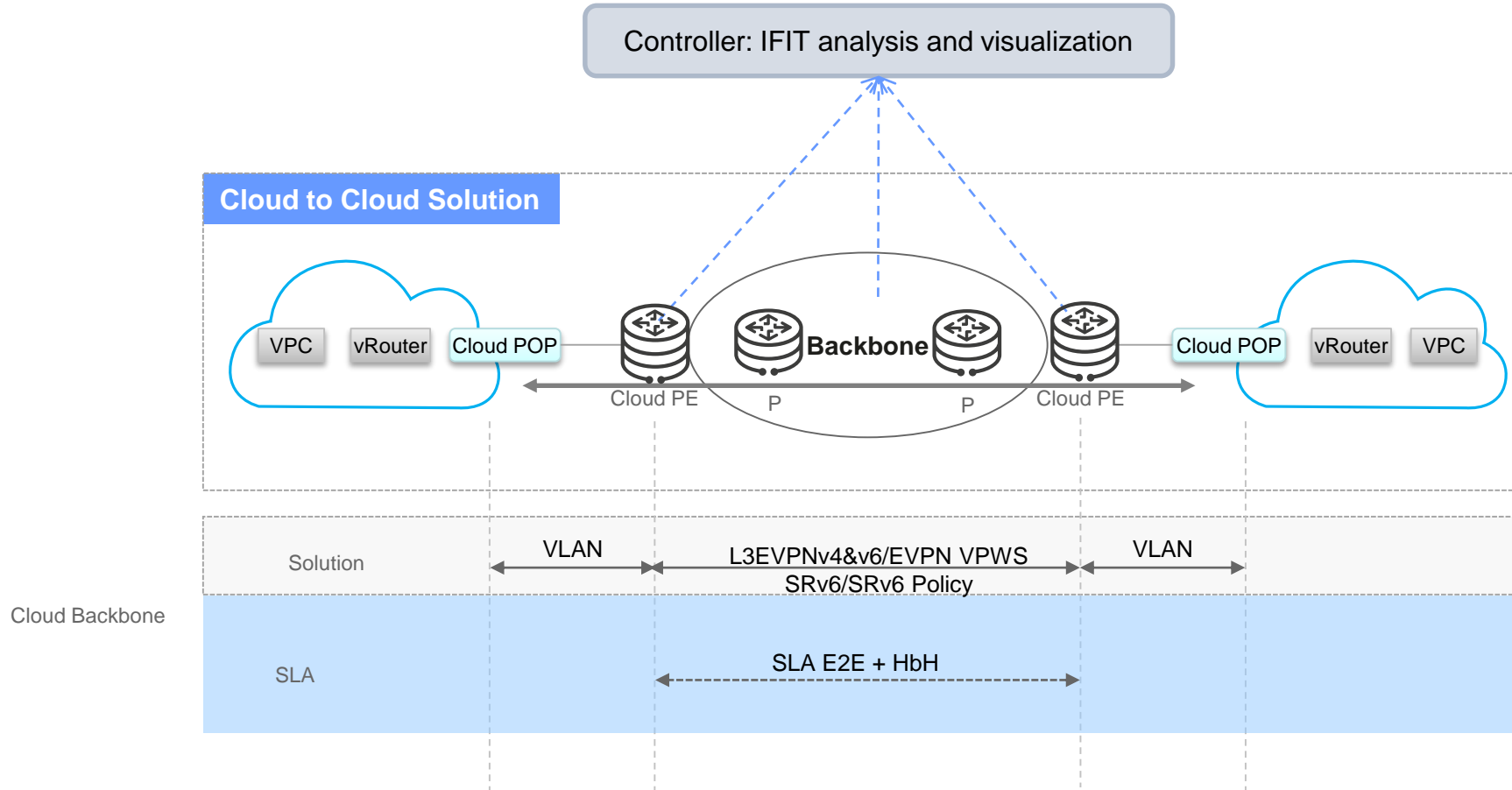
NCE-IP

iFIT   iFIT   iFIT

Flow N+1

R3   R5   R6

HUAWEI

Any Site to Any Site

Controller: IFIT analysis and visualization

Site to Site Solution

ONT OLT Metro Backbone Metro OLT ONT

PE PE

CPE CSG ASG P P ASG CSG CPE

PON+Metro	Solution	VLAN VLAN/QINQ L3EVPNv4&v6/EVPN VPWS SRv6/SRv6 Policy VLAN/QINQ VLAN	
	SLA	SLA E2E + HbH	
IPRAN	Solution	L3EVPNv4&v6/EVPN VPWS SRv6/SRv6 Policy	
	SLA	SLA E2E + HbH	

HUAWEI

Site to Any Cloud

Controller: IFIT analysis and visualization

	Solution	L3EVPNv4&v6/HVPN/EVPN VPWS
Cloud VPN		SRv6/SRv6 Policy
	SLA	SLA E2E + HbH
IPRAN	Solution	L3EVPNv4&v6/L3VPNv4&v6/HVPN/EVPN VPWS
		SRv6/SRv6 Policy
	SLA	SLA E2E + HbH
PON+Metro	Solution	L3EVPNv4&v6/L3VPNv4&v6
		SRv6/SRv6 Policy
	SLA	SLA E2E + HbH

Any Cloud to Any Cloud

Global Deployment

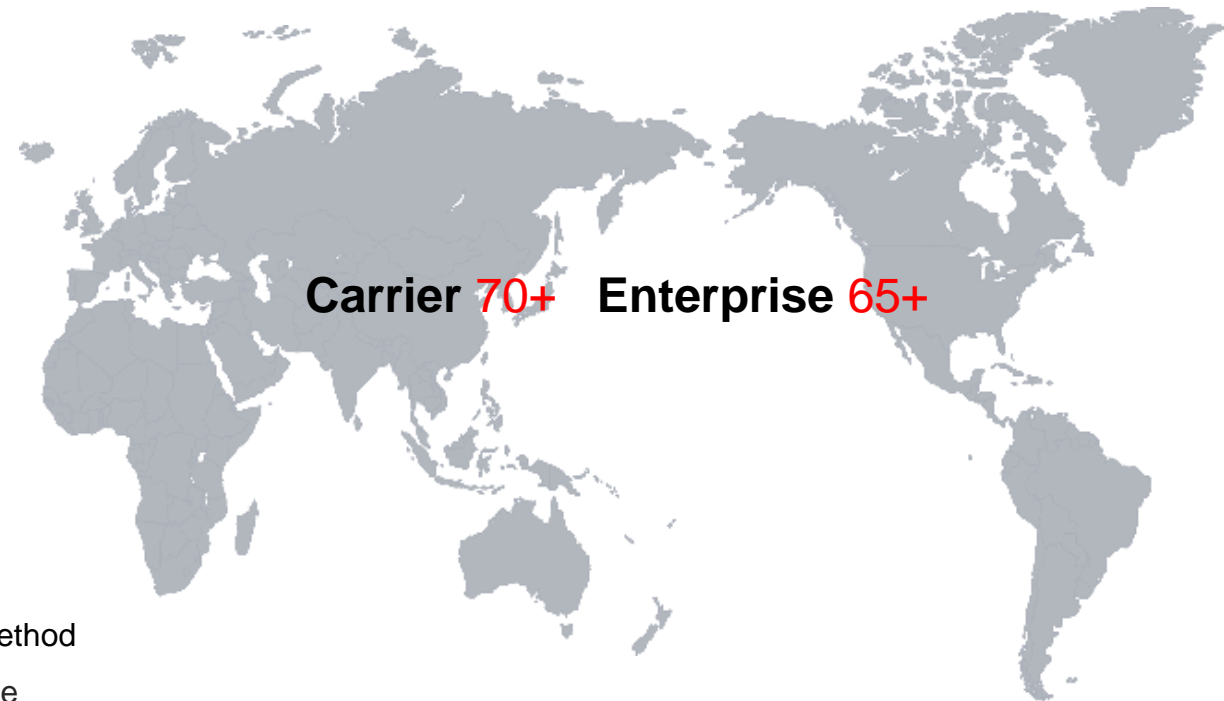2019 Tokyo Interop
Best of Show Award

IFIT: Intelligent Flow
Information Telemetry
Published On SIGCOMM 2019

RFC9232: Network Telemetry Framework
RFC9341: Alternate Marking Method
RFC9342: Clustered Alternate Marking Method
RFC9343: IPv6 Application of the Alternate
Marking Method

Global IFIT Cases

Carrier 70+   Enterprise 65+

HUAWEI