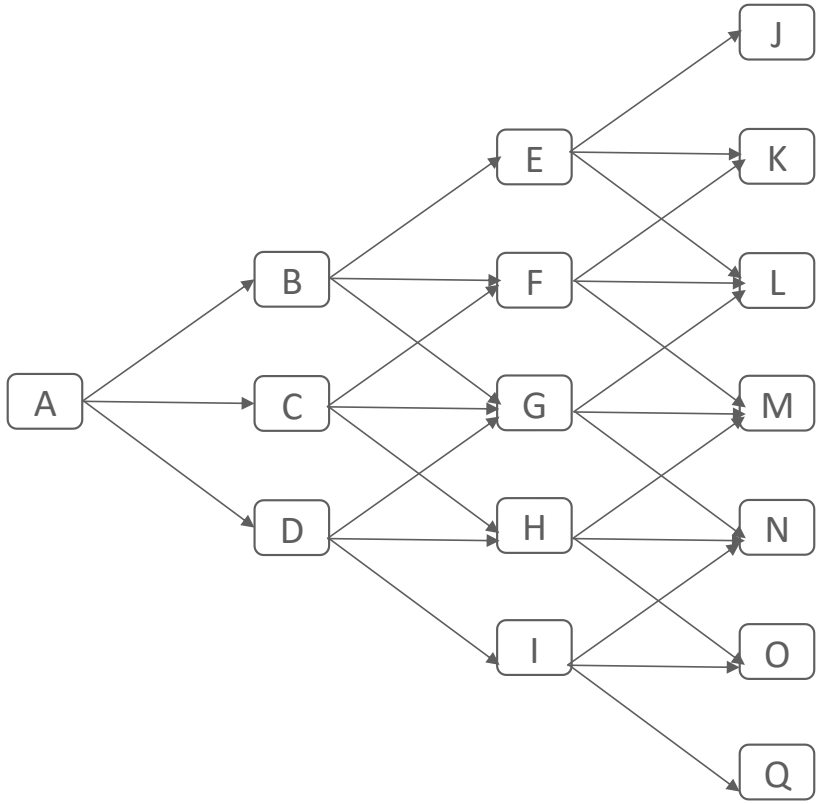
Path Tracing

NANOG 85 – Montreal

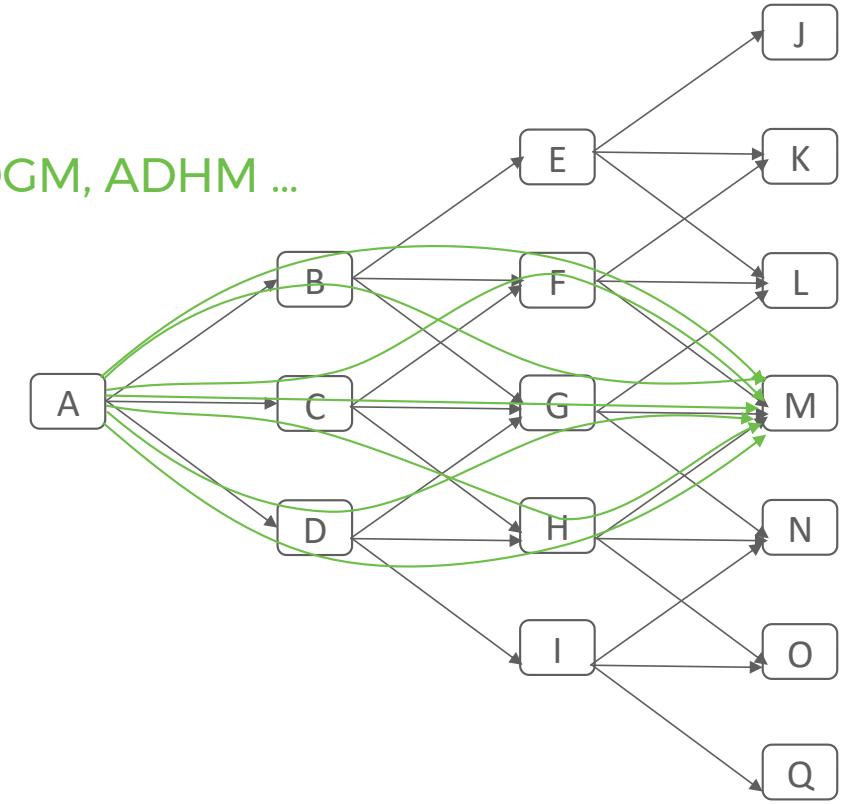
The exact path from A to M is not known

- ECMP is a key in IP networks.



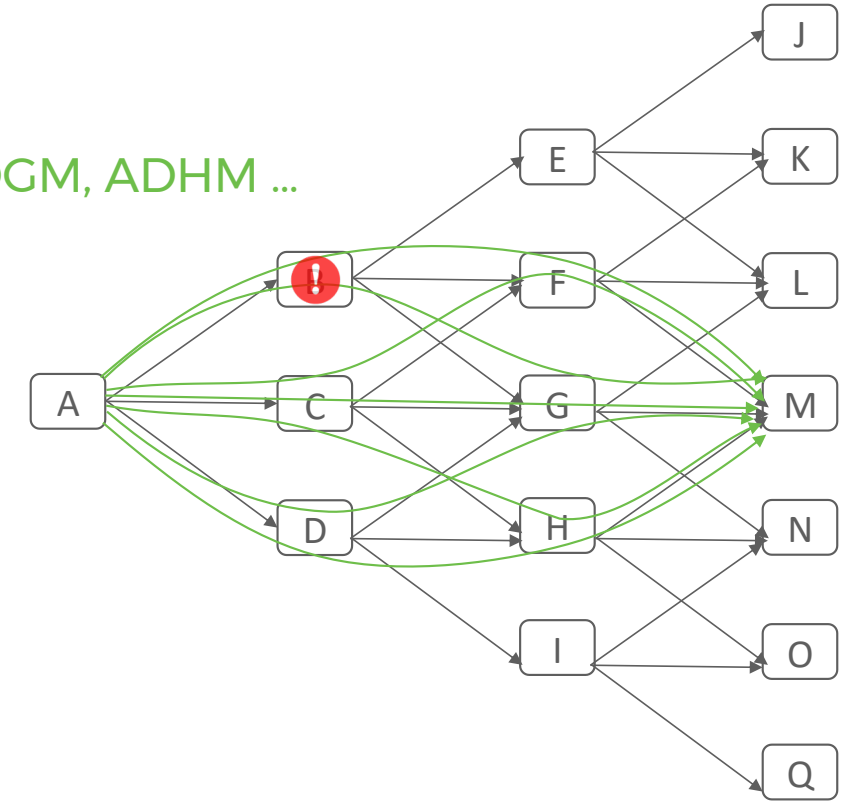
The exact path from A to M is not known

- ECMP is a key in IP networks.
- 7 possible “valid” ECMP path
  - ABFM, ABGM, ACFM, ACGM, ACHM, ADGM, ADHM …



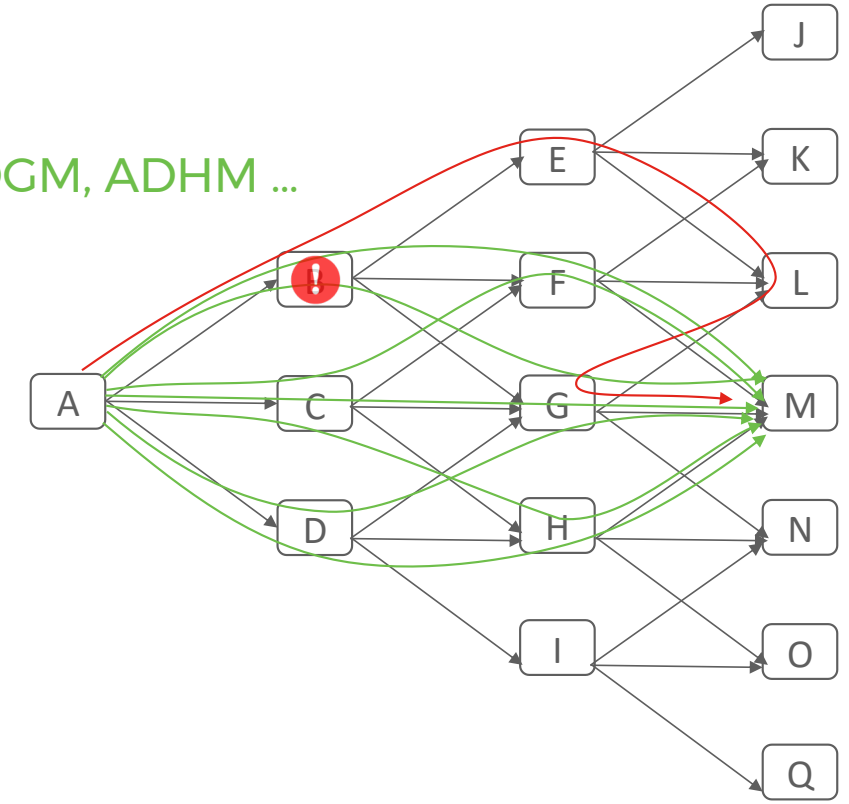
The exact path from A to M is not known

- ECMP is a key in IP networks.
- 7 possible “valid” ECMP path
  - ABFM, ABGM, ACFM, ACGM, ACHM, ADGM, ADHM …
- The path may be invalid
  - Routing or FIB corruption @ B



The exact path from A to M is not known

- ECMP is a key in IP networks.
- 7 possible “valid” ECMP path
  - ABFM, ABGM, ACFM, ACGM, ACHM, ADGM, ADHM …
- The path may be invalid
  - Routing or FIB corruption @ B

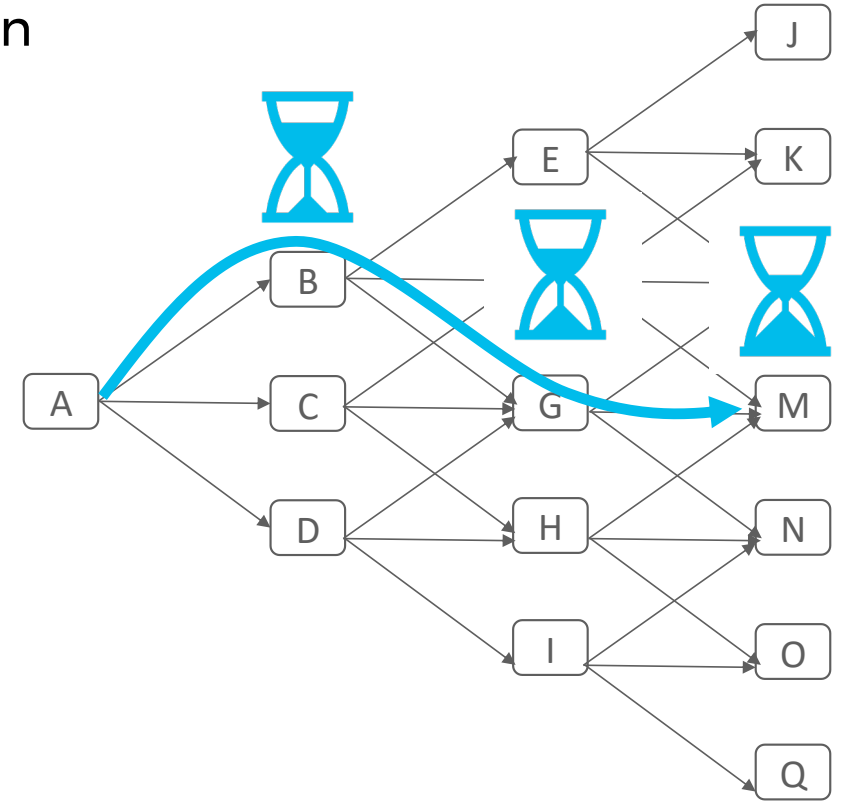


Path Tracing

- Deterministic Per-Packet Tracing
- Implemented at linerate in the base HW pipeline
  - No punting to CPU, no offload to co-processors
- Ultra-MTU-efficient: only 3 byte per hop!
  - 12-bit Interface, 8-bit Timestamp, 4-bit Load
- IPv6 with native SRv6 support
  - MPLS design also available
- Seamless Deployment
  - Interwork with legacy nodes

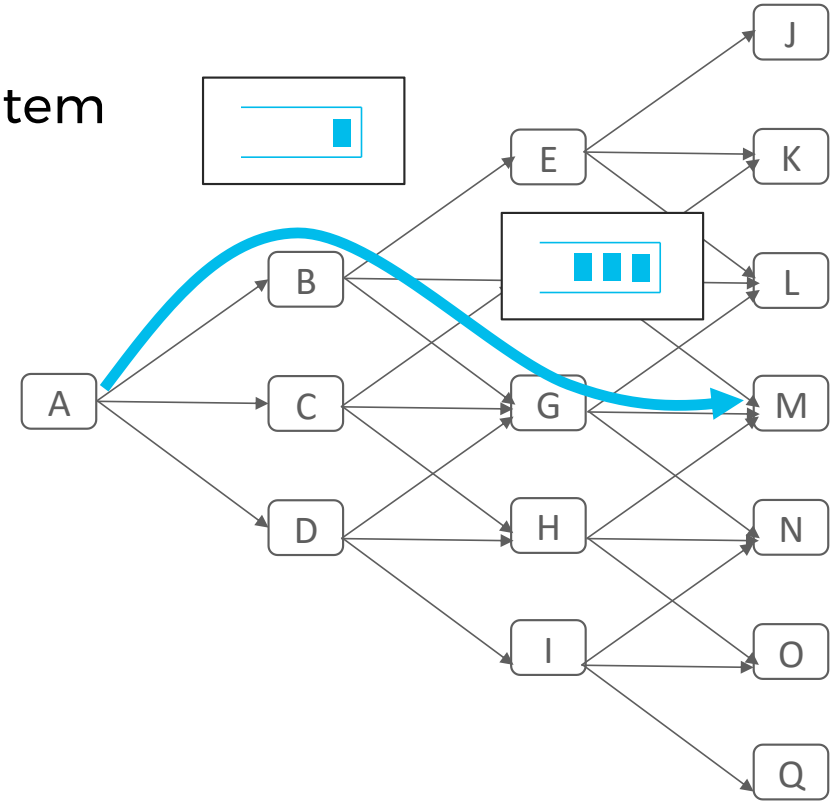
How is time spent from A to M?

- Let's assume we solve the first question
  - ABGM
- How is time spent from A to M?

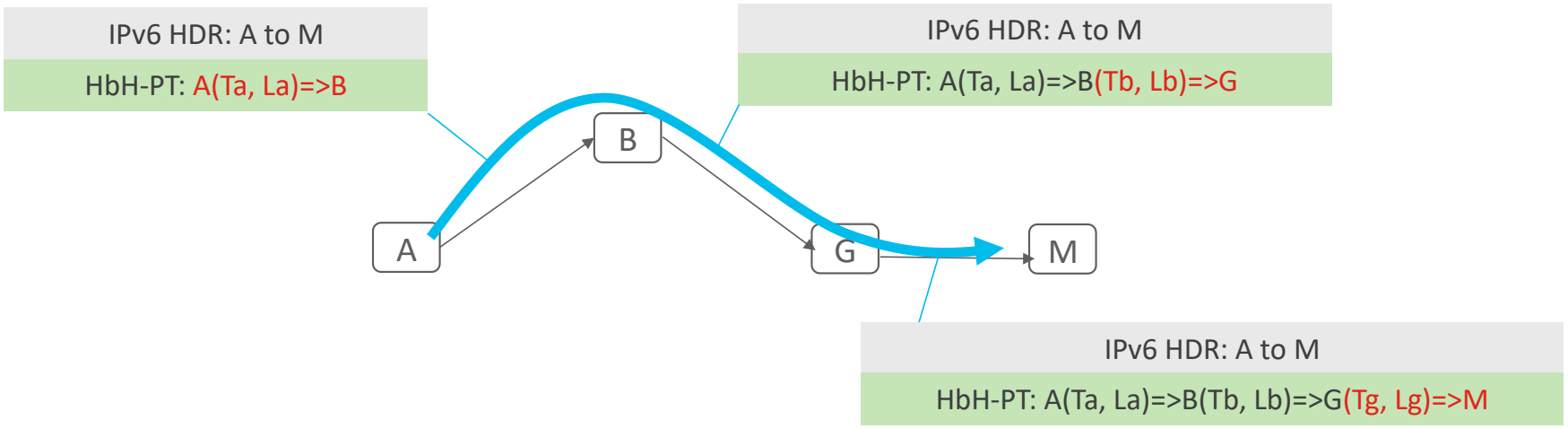


What load at each hop

- Ingress and Egress Queues
- Individual and Aggregate Memory System



Hop-by-Hop Collection



- This is a simplified illustration to introduce the concept

Design intuition

NANOG™

Dataplane Encapsulation

- Minimize NPU parsing
- Minimize # of Read/Write
- Minimize depth of Read/Write
- Maximize Read/Write at fixed positions
- Avoid Header Insert/Resize
- Minimize MTU

Per-Hop Collected Data

- Highly compressed
- Midpoint Compressed Data (MCD)
  - Only 3 Bytes per Hop!
  - 12-bit Interface ID, 4-bit Interface Load, 8-bit Timestamp

Minimize HW complexity by leveraging SDN analytics

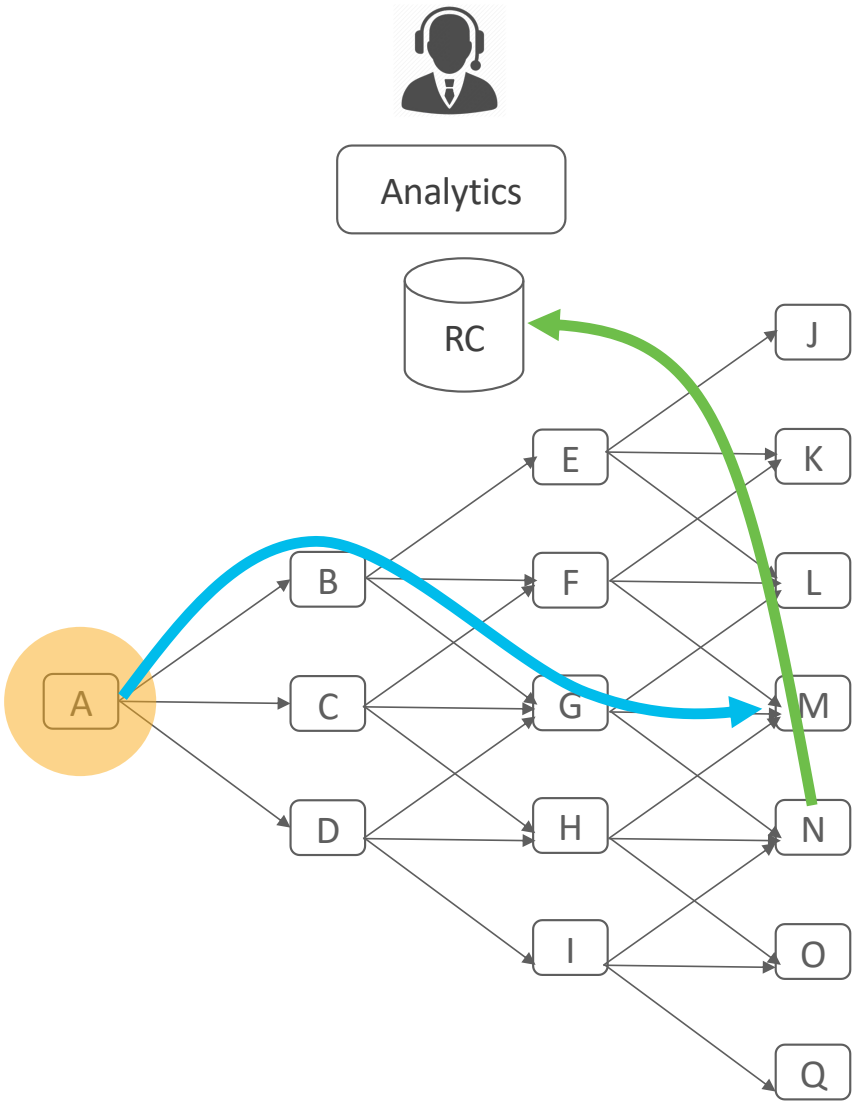
- Analytics
  - translates the list of collected IDs into a path
  - deduces the timing and load history at each hop
  - Highlights hotspots
- GUI visualization
- Feedback loop to applications
  - Trigger a change of path (SR, MTCP)
  - Trigger a change of rate

Roles and Data Model

NANOG™

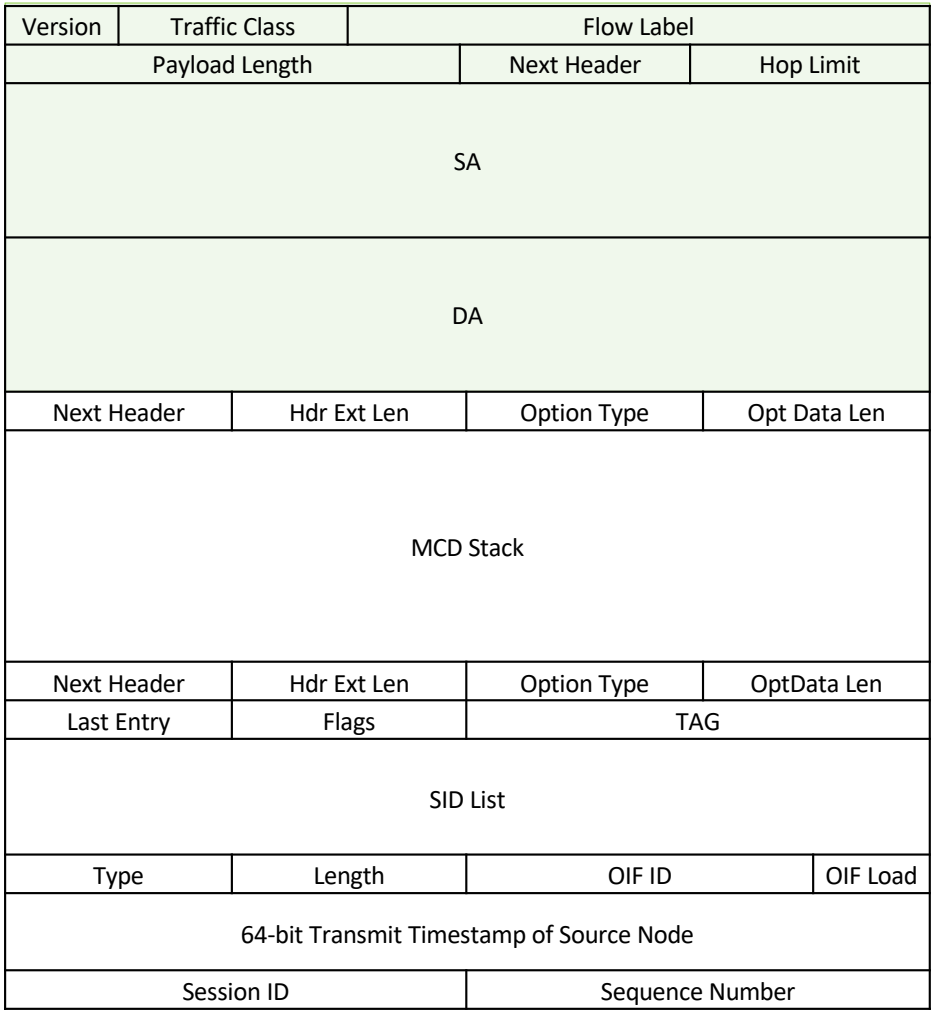
Source

- Originates the probe
- Collects transmission data



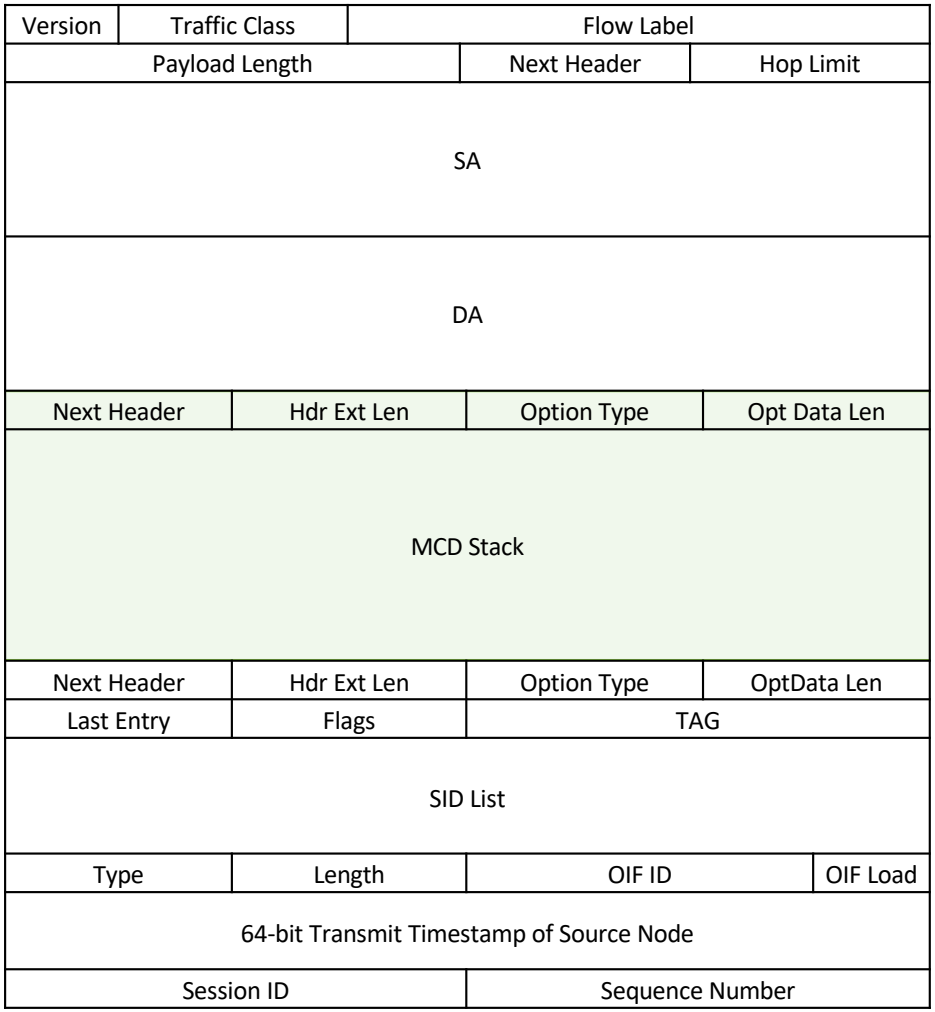
Packet generated @ SRC

- IPv6 Header
  - SA, DA, DSCP, FL, …



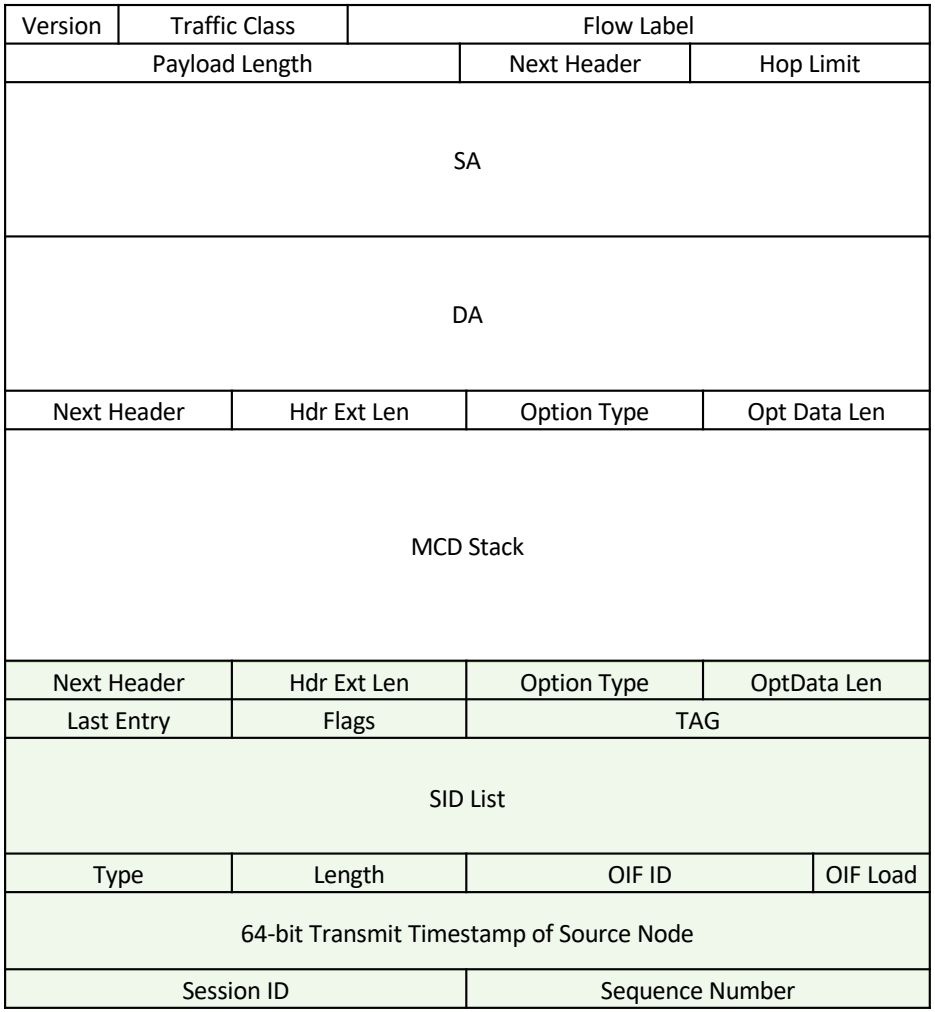
Packet generated @ SRC

- IPv6 Header
  - SA, DA, DSCP, FL, …
- HbH header
  - PT Option Type
  - MCD Stack



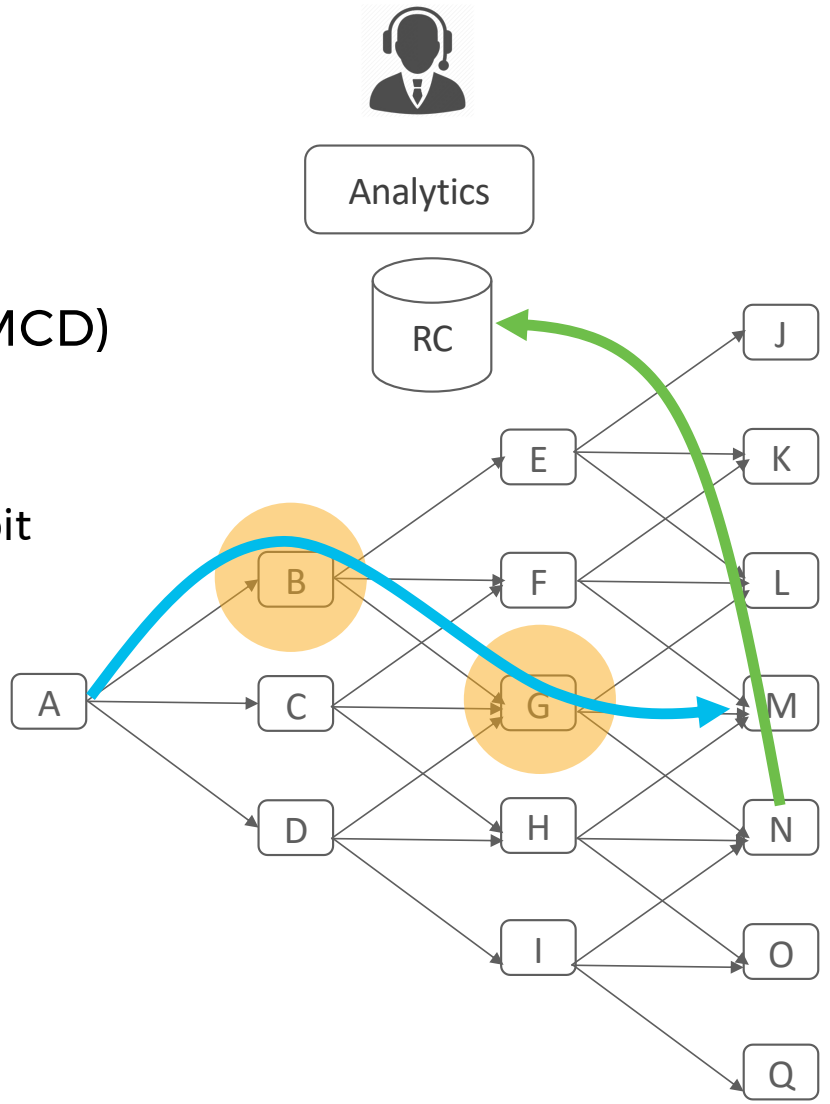
Packet generated @ SRC

- IPv6 Header
  - SA, DA, DSCP, FL, …
- HbH header
  - PT Option Type
  - MCD Stack
- SRH
  - SID List
  - SRH PT-TLV (TS, OIF ID, OIF Load)

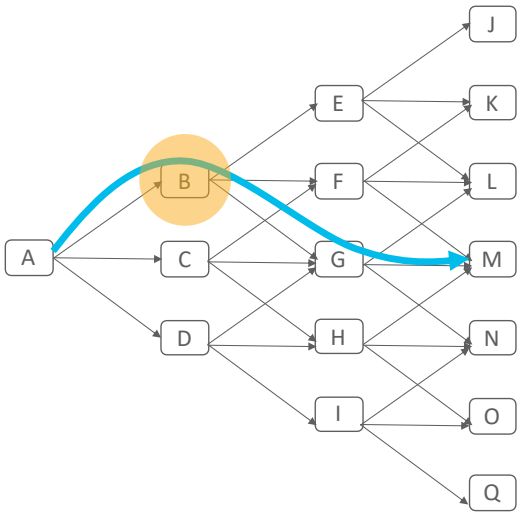


Midpoint

- Collects Midpoint Compressed Data (MCD)
- MCD
  - Only 3 byte per hop!
  - 12-bit Interface ID, 4-bit Interface Load, 8-bit Timestamp
- Shift & Stamp behavior
  - linerate



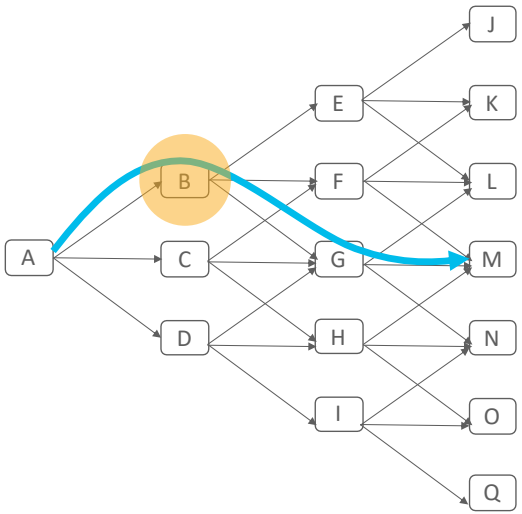
Shift & Stamp @ B



Version	Traffic Class	Flow Label	
Payload Length		Next Header	Hop Limit
SA			
DA			
Next Header	Hdr Ext Len	Option Type	Opt Data Len
Next Header	Hdr Ext Len	Option Type	OptData Len
Last Entry	Flags	TAG	
SID List			
Type	Length	OIF ID	OIF Load
64-bit Transmit Timestamp of Source Node			
Session ID		Sequence Number	

Shift & Stamp @ B

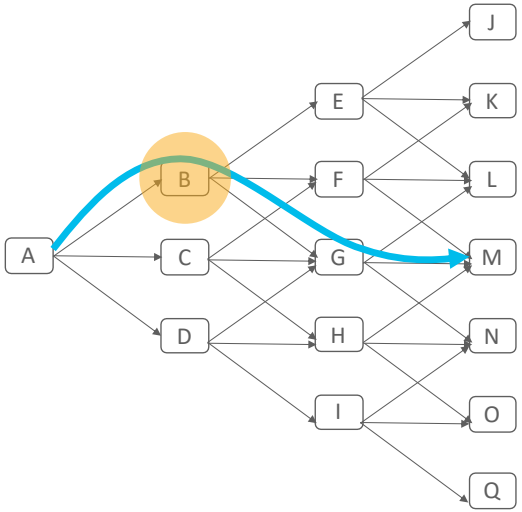
- Shift MCD Stack 3Bytes to the right



Version	Traffic Class	Flow Label	
Payload Length		Next Header	Hop Limit
SA			
DA			
Next Header	Hdr Ext Len	Option Type	Opt Data Len
Next Header	Hdr Ext Len	Option Type	OptData Len
Last Entry	Flags	TAG	
SID List			
Type	Length	OIF ID	OIF Load
64-bit Transmit Timestamp of Source Node			
Session ID		Sequence Number	

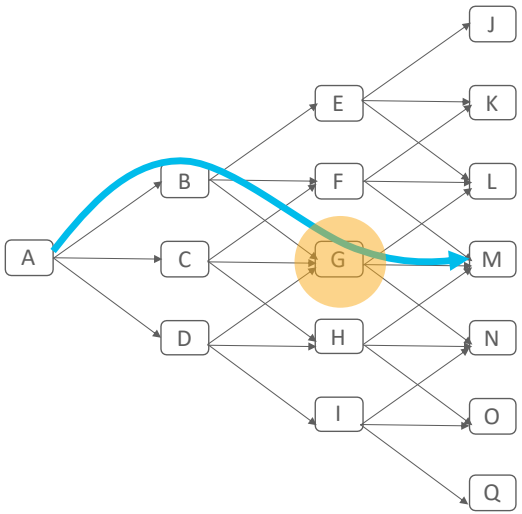
Shift & Stamp @ B

- Shift MCD Stack 3Bytes to the right
- Stamp MCD in the first 3Bytes of the MCD Stack



Version	Traffic Class	Flow Label	
Payload Length		Next Header	Hop Limit
SA			
DA			
Next Header	Hdr Ext Len	Option Type	Opt Data Len
MCD (B)			
Next Header	Hdr Ext Len	Option Type	OptData Len
Last Entry	Flags	TAG	
SID List			
Type	Length	OIF ID	OIF Load
64-bit Transmit Timestamp of Source Node			
Session ID		Sequence Number	

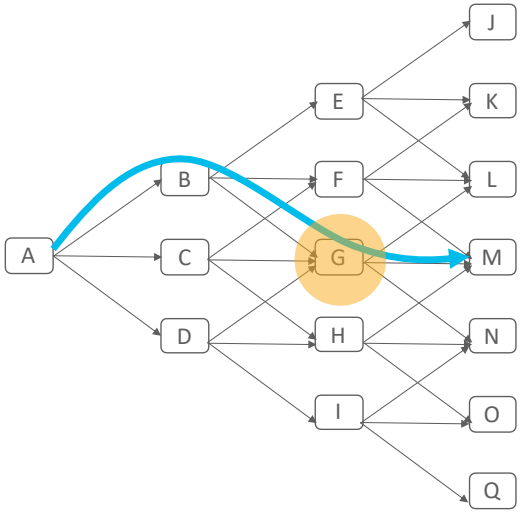
Shift & Stamp @ G



Version	Traffic Class	Flow Label	
Payload Length		Next Header	Hop Limit
SA			
DA			
Next Header	Hdr Ext Len	Option Type	Opt Data Len
MCD (B)			
Next Header	Hdr Ext Len	Option Type	OptData Len
Last Entry	Flags	TAG	
SID List			
Type	Length	OIF ID	OIF Load
64-bit Transmit Timestamp of Source Node			
Session ID		Sequence Number	

Shift & Stamp @ G

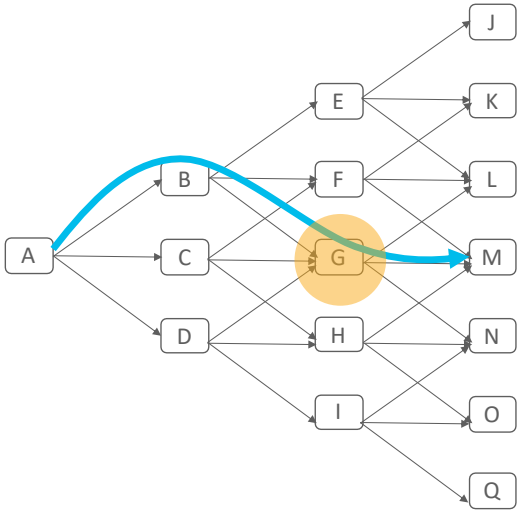
- Shift MCD Stack 3Bytes to the right



Version	Traffic Class	Flow Label	
Payload Length		Next Header	Hop Limit
SA			
DA			
Next Header	Hdr Ext Len	Option Type	Opt Data Len
			MCD (B)
MCD (B)			
Next Header	Hdr Ext Len	Option Type	OptData Len
Last Entry	Flags	TAG	
SID List			
Type	Length	OIF ID	OIF Load
64-bit Transmit Timestamp of Source Node			
Session ID		Sequence Number	

Shift & Stamp @ G

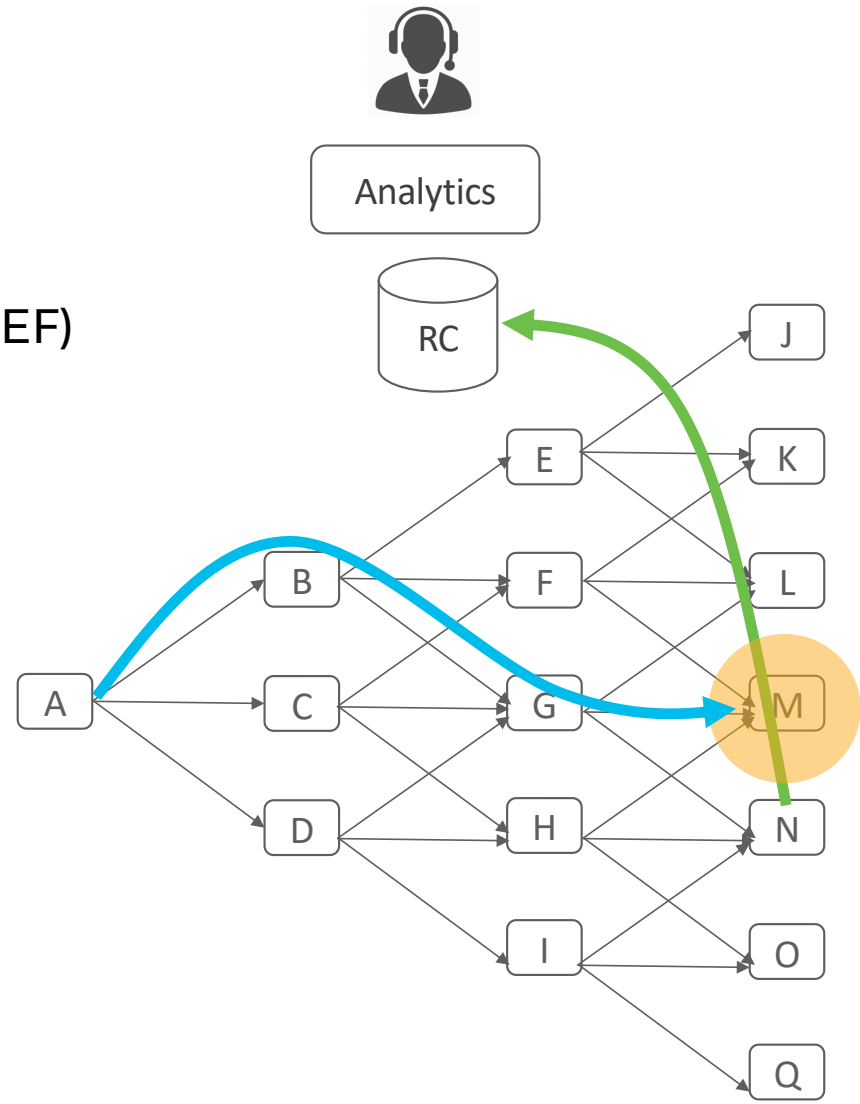
- Shift MCD Stack 3Bytes to the right
- Stamp MCD in the first 3Bytes of the MCD Stack



Version	Traffic Class	Flow Label	
Payload Length		Next Header	Hop Limit
SA			
DA			
Next Header	Hdr Ext Len	Option Type	Opt Data Len
MCD (G)			MCD (B)
MCD (B)			
Next Header	Hdr Ext Len	Option Type	OptData Len
Last Entry	Flags	TAG	
SID List			
Type	Length	OIF ID	OIF Load
64-bit Transmit Timestamp of Source Node			
Session ID		Sequence Number	

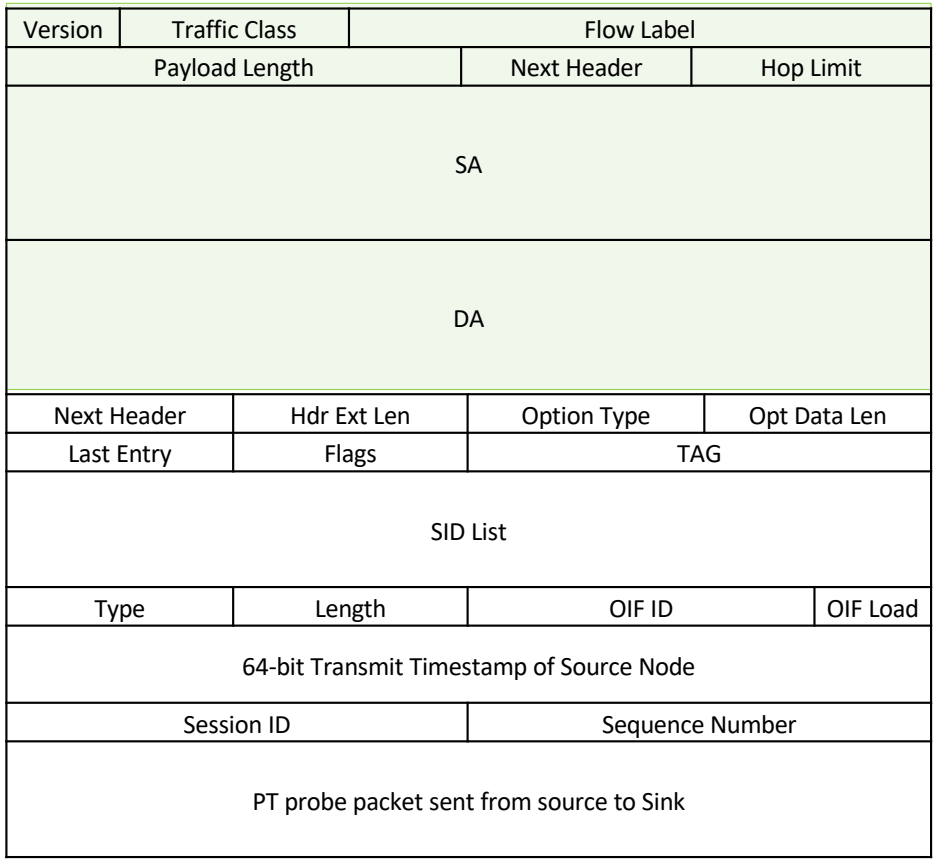
Sink

- Timestamp, Encapsulate and Forward (TEF)
  - Linerate
- Records reception data



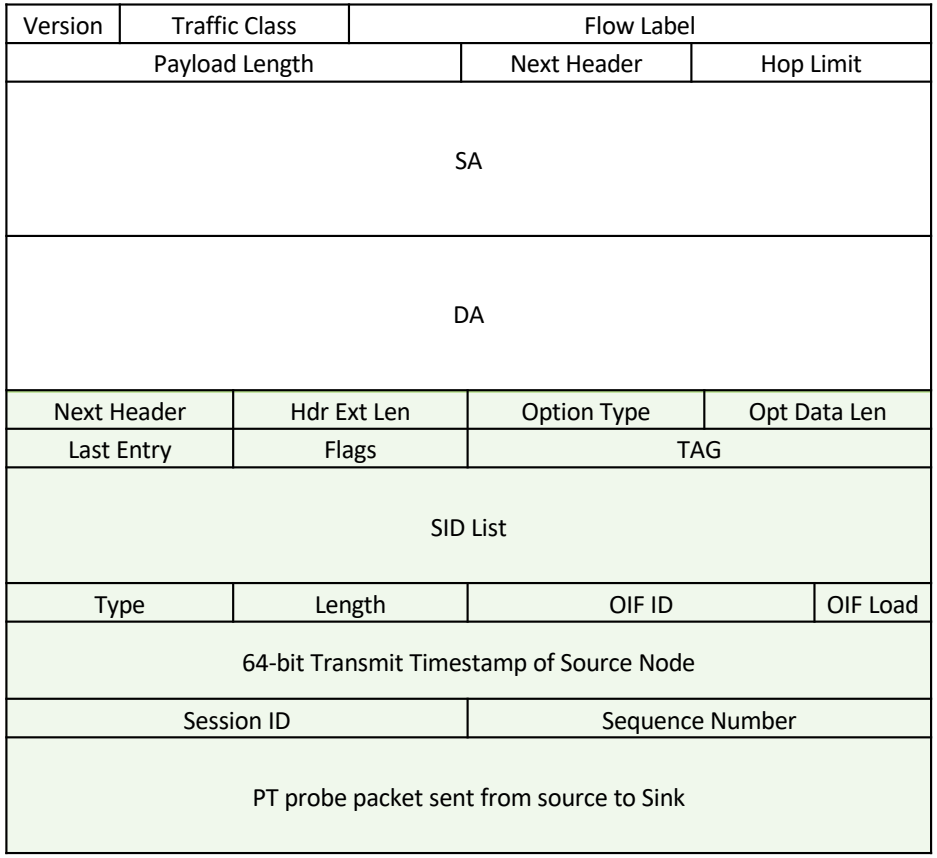
Packet forwarded by Sink

- IPv6 Header
  - DA = Regional Collector(RC)



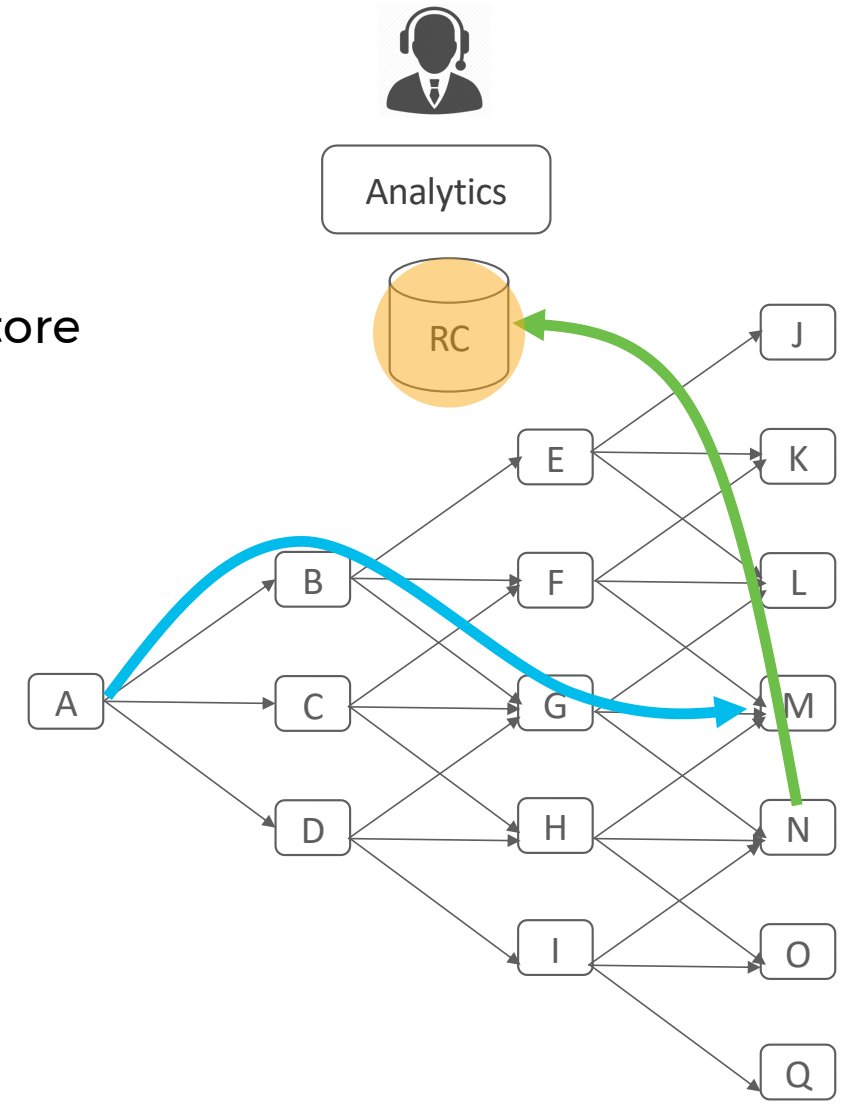
Packet forwarded by Sink

- IPv6 Header
  - DA = Regional Collector(RC)
- SRH
  - SRH PT-TLV (TS, IIF ID, IIF Load)



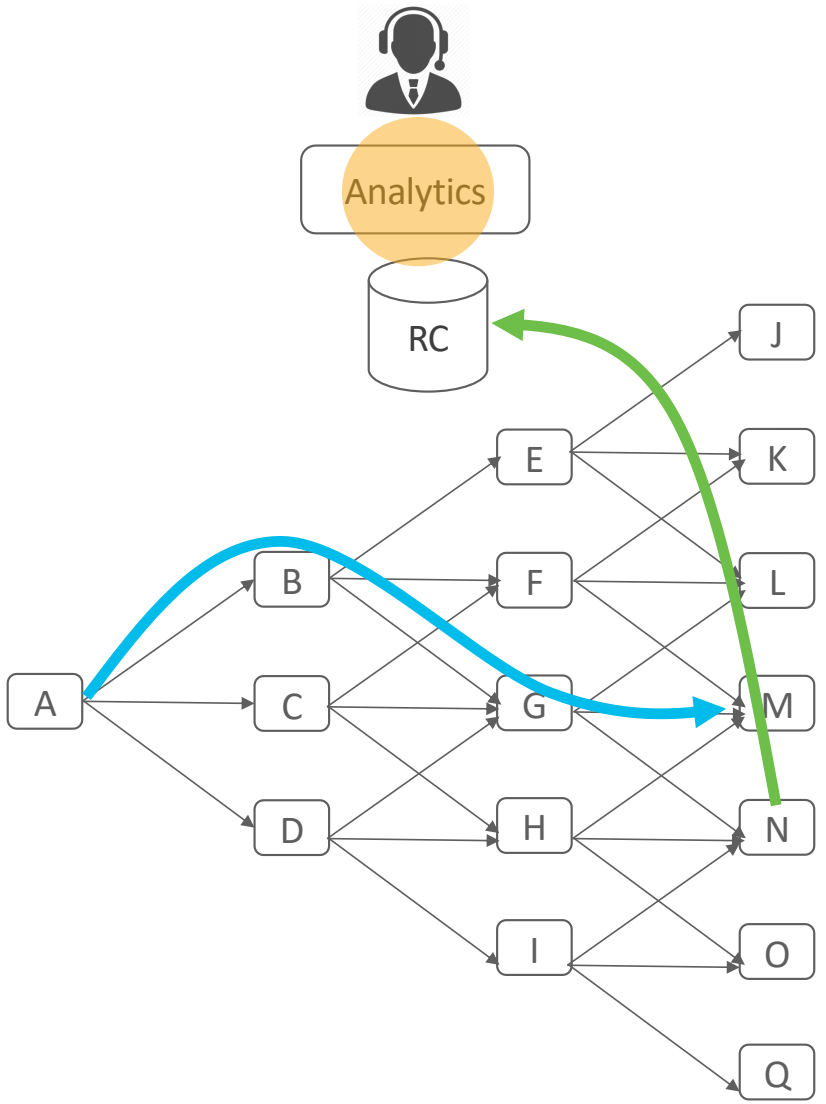
Regional Collector (RC)

- Ingest probes data in Time-Serie Data store



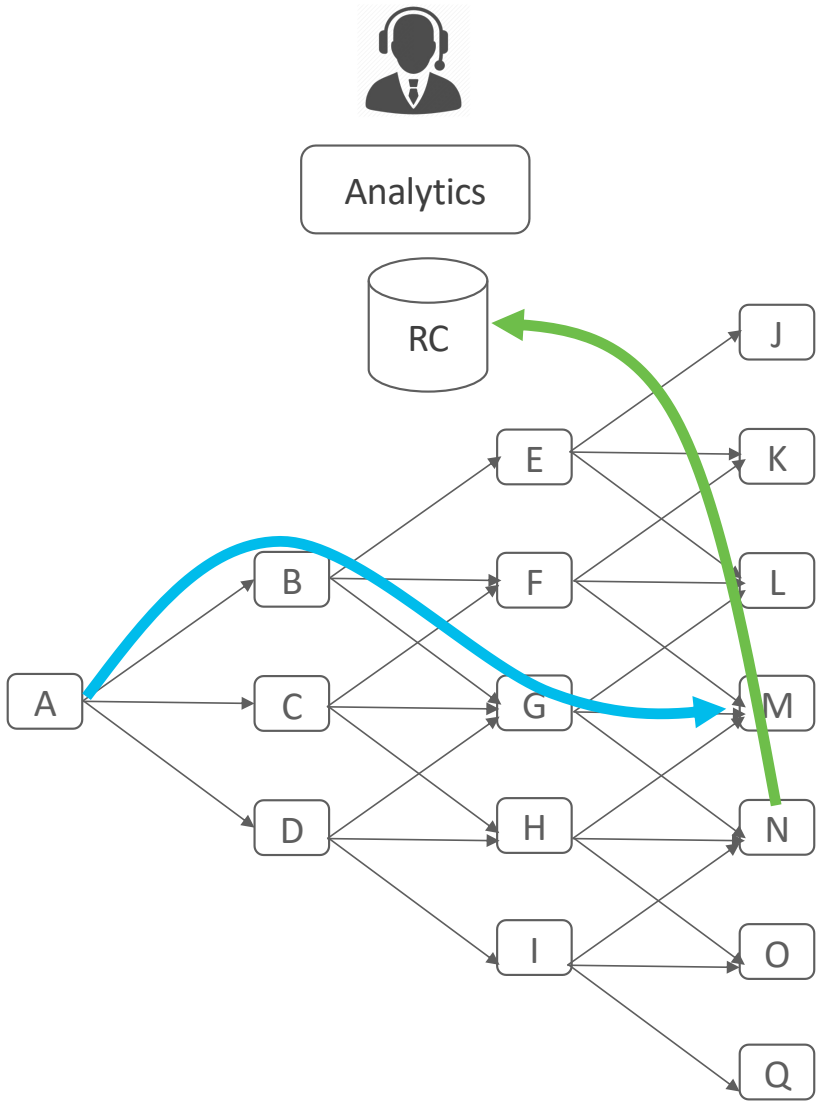
Analytics

- Visualization
- Troubleshooting
- Alerts



Collected Data

- Source
  - 12-bit Outgoing Interface ID
  - 4-bit Outgoing Interface Load
  - 64-bit PTP Tx Timestamp
- Midpoint
  - 12-bit Outgoing Interface ID
  - 4-bit Outgoing Interface Load
  - 8-bit Truncated PTP Tx Timestamp
- Sink
  - 12-bit Incoming Interface ID
  - 4-bit Incoming Interface Load
  - 64-bit PTP Rx Timestamp



Ecosystem & Standardization

NANOG™

Ecosystem

- Rich Eco-System
  - Broadcom, Cisco, Marvell, +others
- Strong Operator Interest
- Rich Open-Source
  - Linux, FD.io VPP, P4, WIRESHARK, TCPDUMP

Standardization

- Submitted to IETF in March 2022
  - https://datatracker.ietf.org/doc/html/draft-filsfils-spring-path-tracing

Conclusion

- Path Tracing
  - Operators can deterministically detect ECMP paths
  - Implemented at linerate in the base HW pipeline
  - Ultra-MTU-Efficiency

- Ecosystem
  - Rich Eco-System (Broadcom, Cisco, Marvell, +others)
  - Strong Operator Interest
  - Rich Open-Source
  - Being standardized at IETF

Thank you

NANOG 85

NANOG™