University of Amsterdam
Informatics Institute

Responsible Internet

DESIRE6G

Lightweight INT on the Tofino Programmable Switch

Angelos Dimoglis, Leandro C. De Almeida, Konstantinos Papadopoulos, Chrysa Papagianni, Panagiotis Papadimitriou, Paola Grosso

November 18, 2024

Contents

1. Background
   1. In-band Network Telemetry
   2. Per-flow Aggregation
2. Implementation on Tofino
3. Performance Evaluation
4. Future work

Network telemetry

Refers to the method of collecting information about the network state.

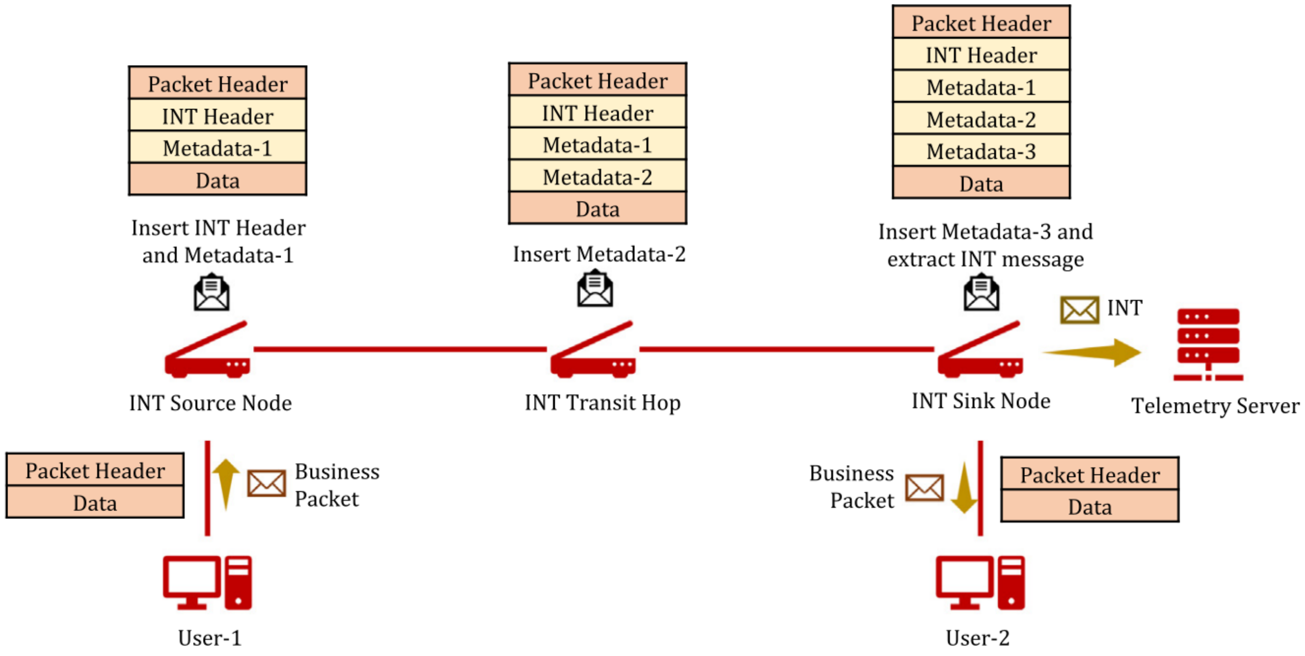
It is a two-step process:

1. Collection of data (e.g. Buffer queue size, delay, etc.) from individual networking devices.
2. Processing of the collected information to take network management decisions to improve:
   - Performance
   - Security
   - Efficiency [1]

   …

In-band Network Telemetry (INT)

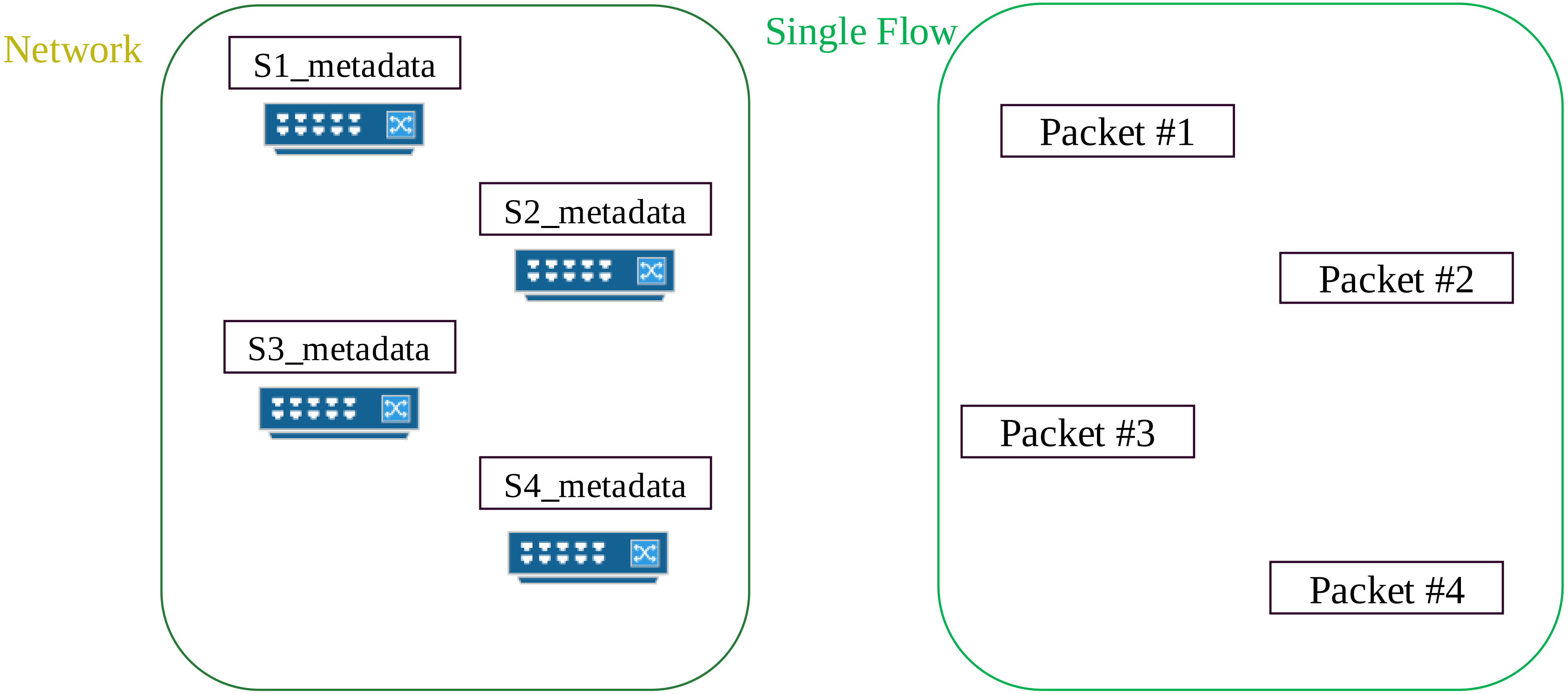
- Combining packet forwarding and network measurements
- Implemented entirely on the data plane
- Improved accuracy, and performance



Typical scheme of In-band Network Telemetry [2]

Lightweight INT: Per-Flow Aggregation (PFA)

Main Idea: The telemetry values are being spread across the packets of a single flow (e.g. TCP).



Proposed methods based on PFA [3]

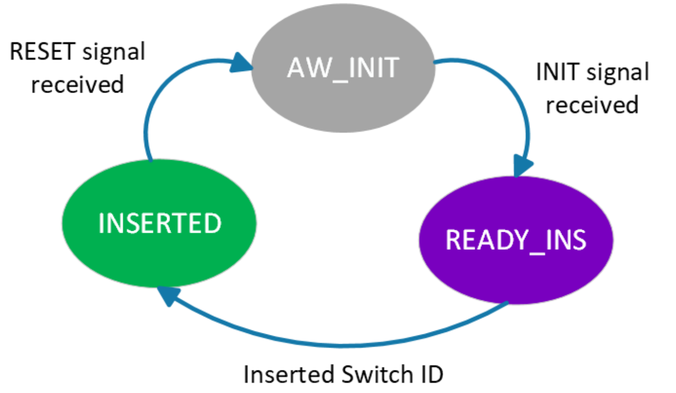
Deterministic Approach (DLINT)

- All the switches are inserting metadata sequentially.
- Requires coordination among the switches
  - Maintaining a per-flow telemetry state on the switch

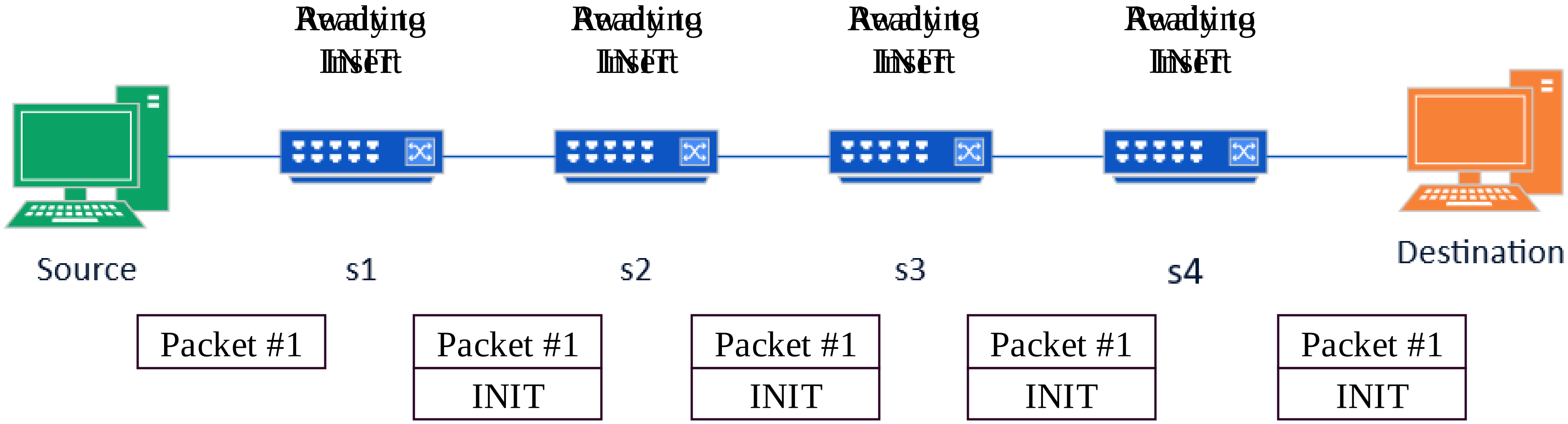
Probabilistic Approach (PLINT)

- The switches are inserting metadata based on a probability.
- No coordination is needed.
  - Stateless

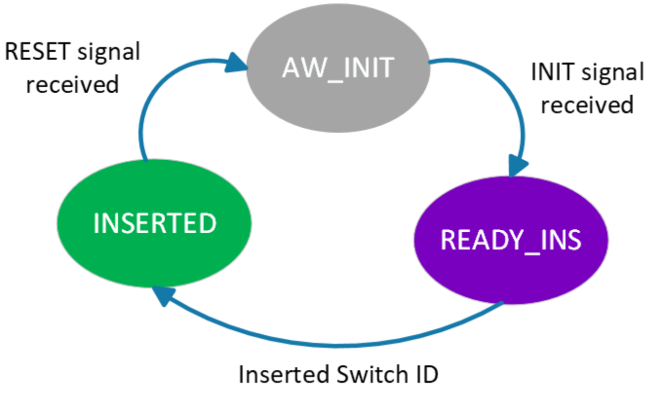
Deterministic PFA-INT Example: Path Tracing



AW_INIT: Awaiting INIT signal
READY_INS: Ready to insert Switch ID
INSERTED: Inserted Switch ID



Deterministic PFA-INT Example: Path Tracing



RESET signal received

AW_INIT

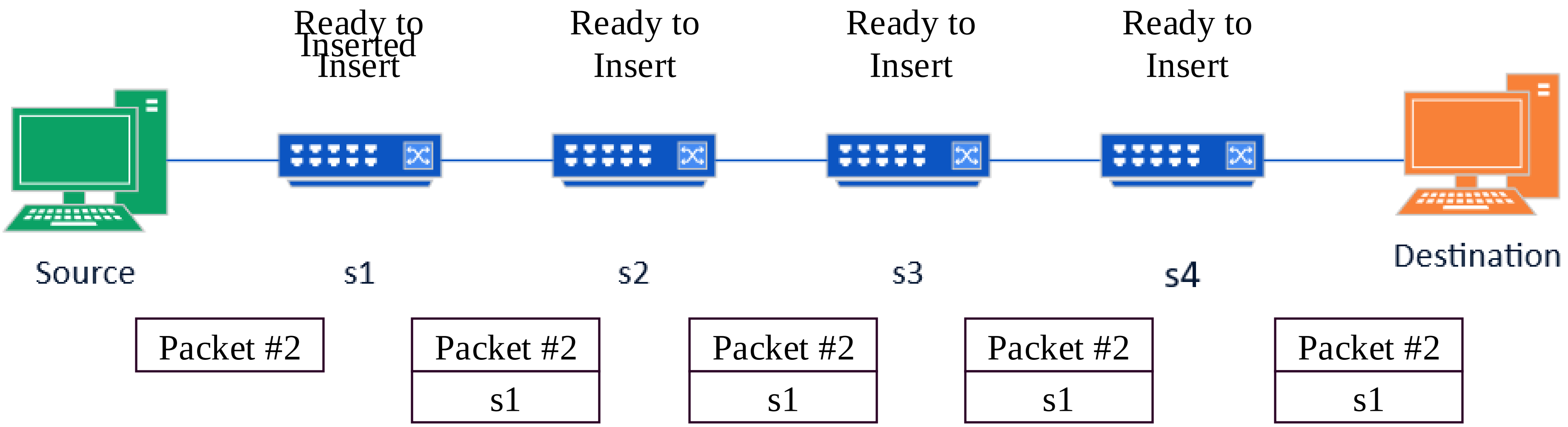
INIT signal received

INSERTED

READY_INS

Inserted Switch ID

AW_INIT: Awaiting INIT signal
READY_INS: Ready to insert Switch ID
INSERTED: Inserted Switch ID



Ready to Insert
Inserted

Ready to Insert

Ready to Insert

Ready to Insert

Source

s1

s2

s3

s4

Destination

Packet #2

Packet #2
s1

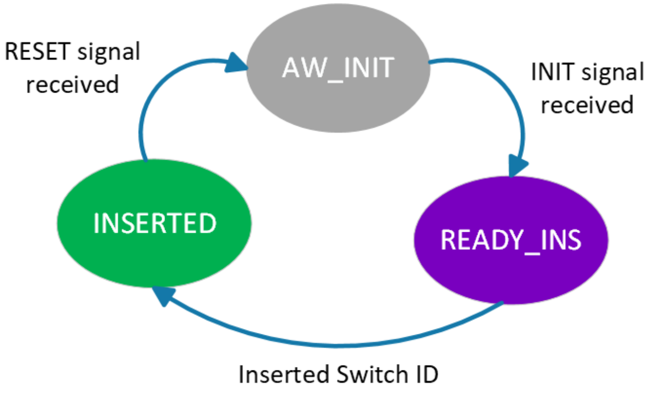
Packet #2
s1

Packet #2
s1

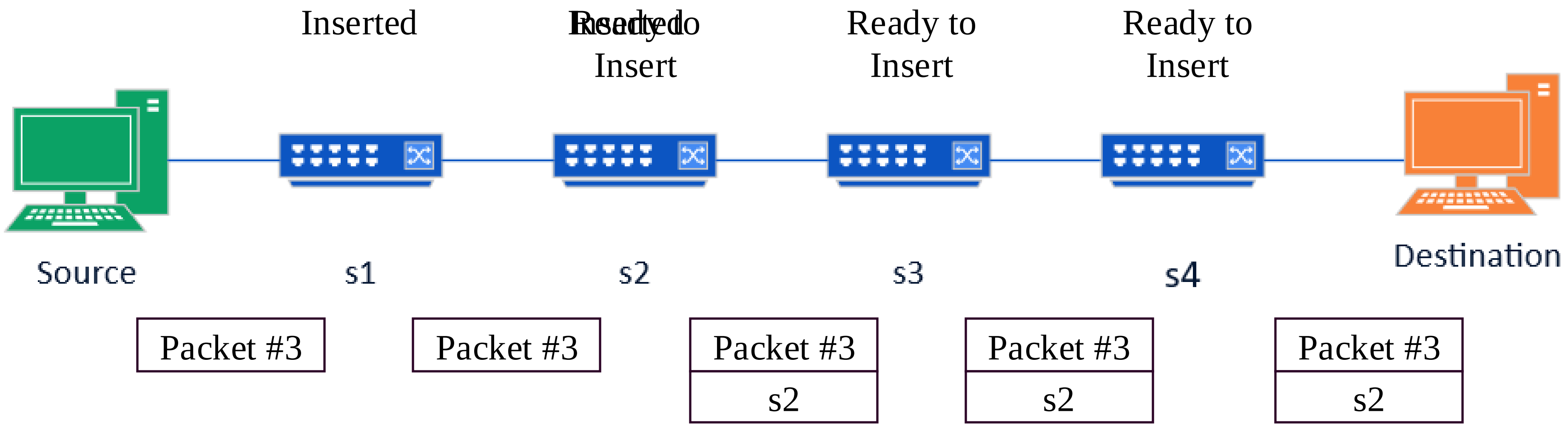
Packet #2
s1

UNIVERSITY OF AMSTERDAM
Informatics Institute

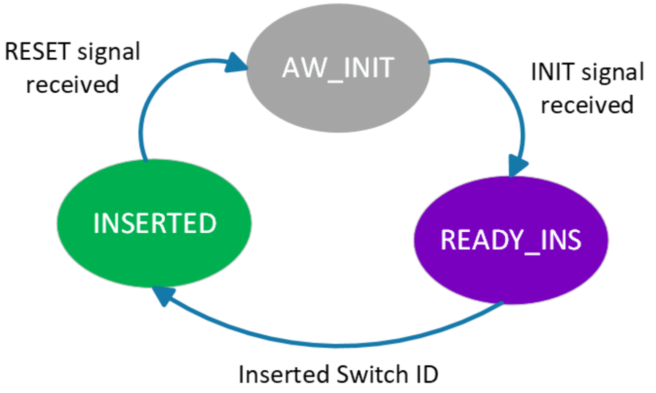
Deterministic PFA-INT Example: Path Tracing



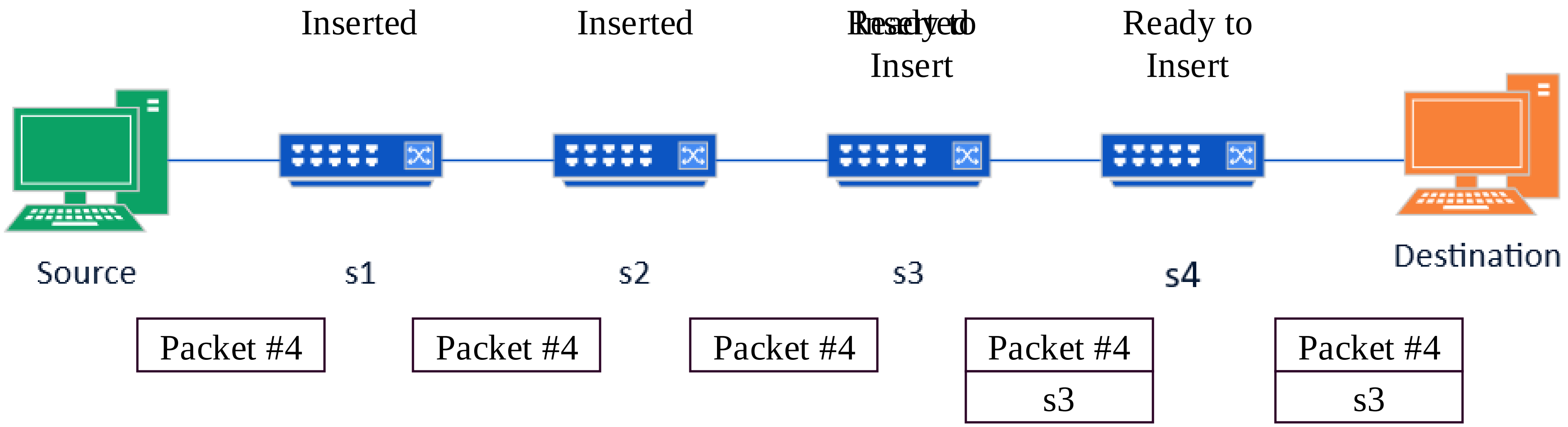
AW_INIT: Awaiting INIT signal
READY_INS: Ready to insert Switch ID
INSERTED: Inserted Switch ID



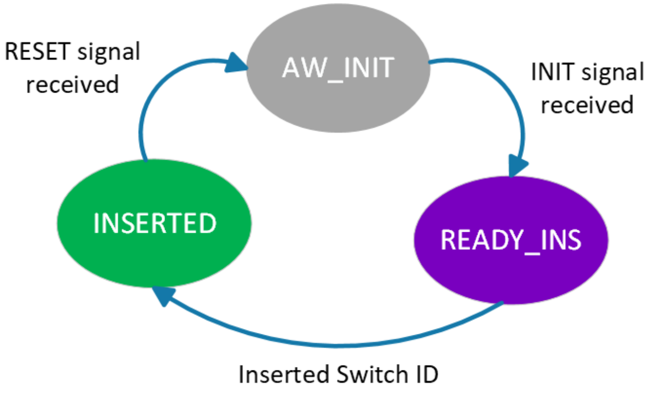
Deterministic PFA-INT Example: Path Tracing



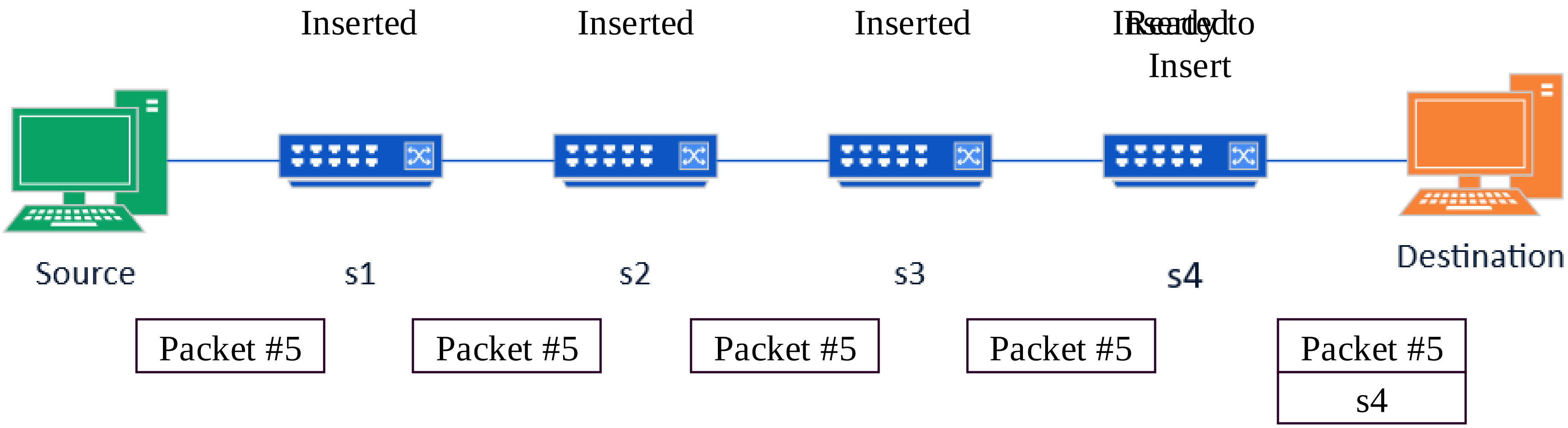
AW_INIT: Awaiting INIT signal
READY_INS: Ready to insert Switch ID
INSERTED: Inserted Switch ID



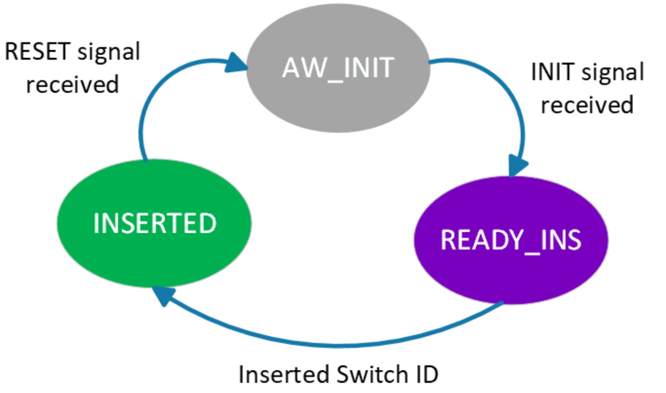
Deterministic PFA-INT Example: Path Tracing



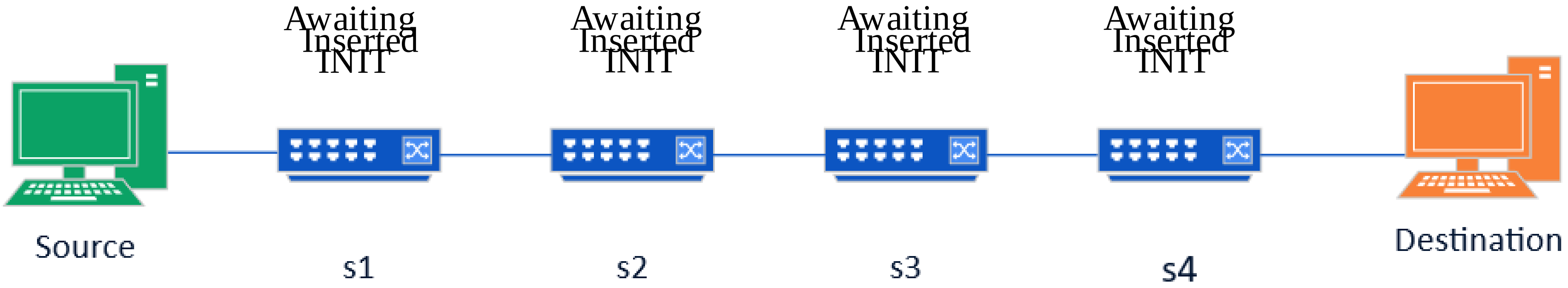
AW_INIT: Awaiting INIT signal
READY_INS: Ready to insert Switch ID
INSERTED: Inserted Switch ID



Deterministic PFA-INT Example: Path Tracing



AW_INIT: Awaiting INIT signal
READY_INS: Ready to insert Switch ID
INSERTED: Inserted Switch ID



Packet (ACK)
RESET

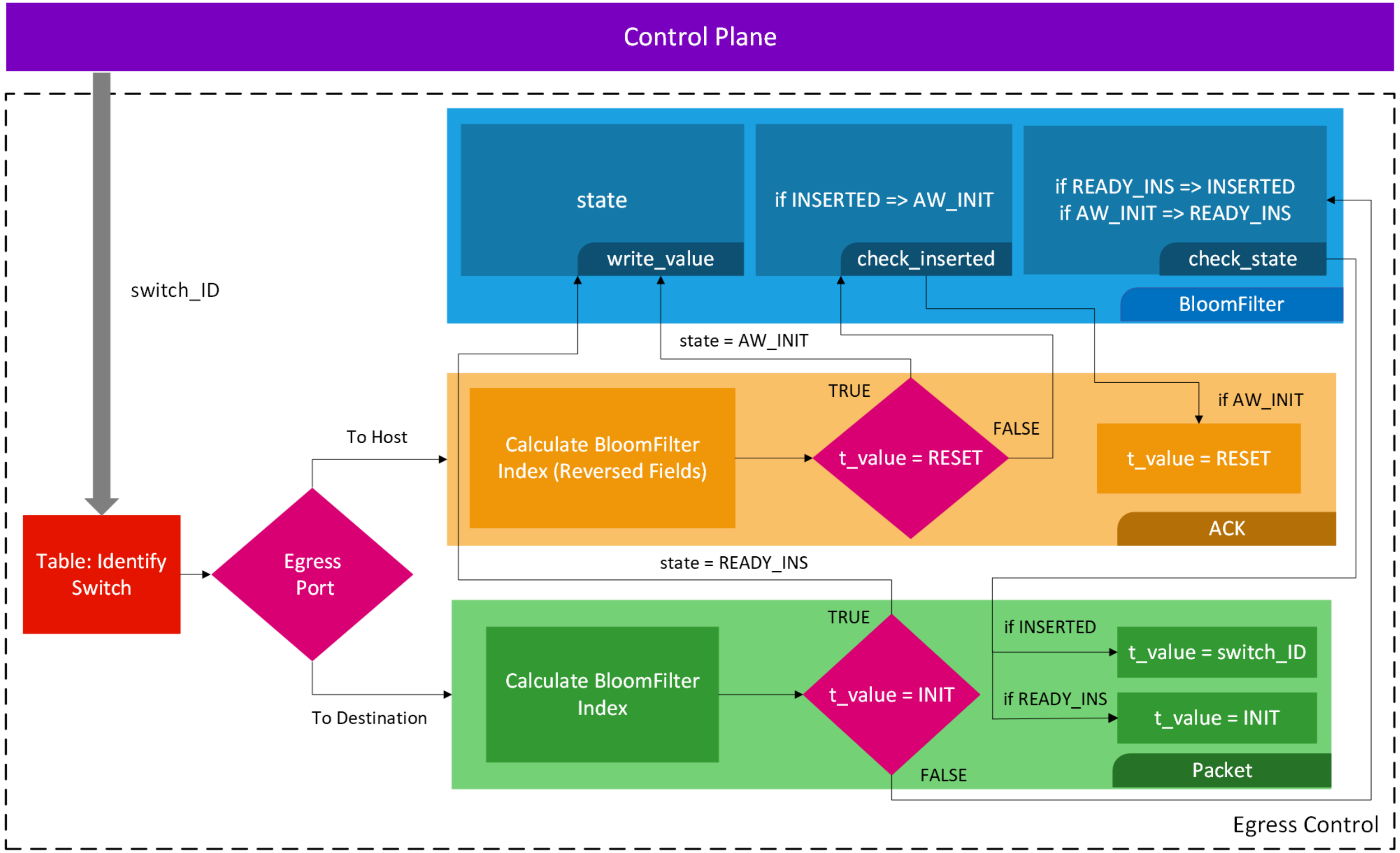
Packet (ACK)
RESET

Packet (ACK)
RESET

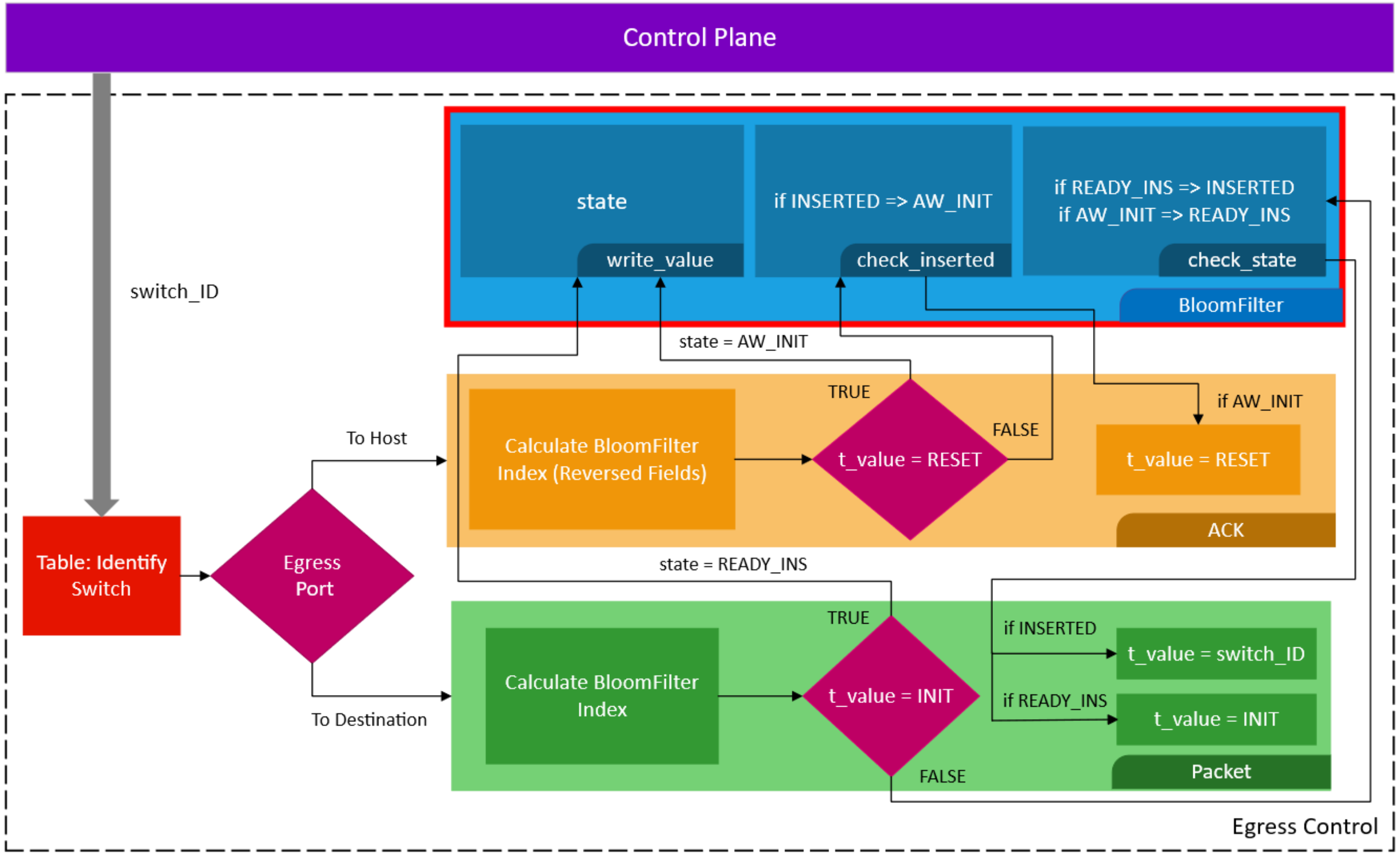
Packet (ACK)
RESET

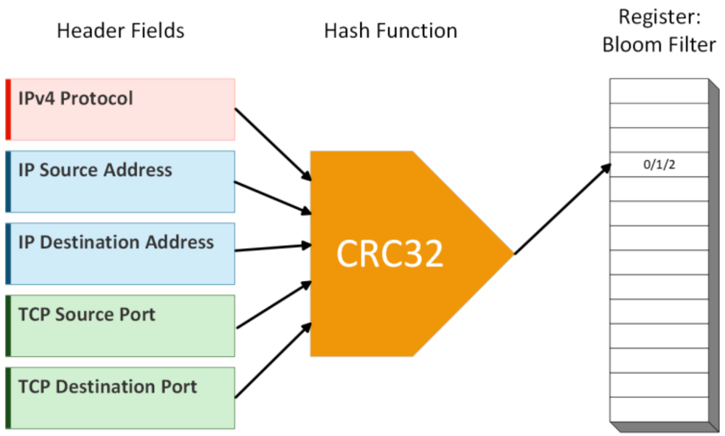
Packet (ACK)

DLINT Implementation on Tofino



DLINT Implementation: Register





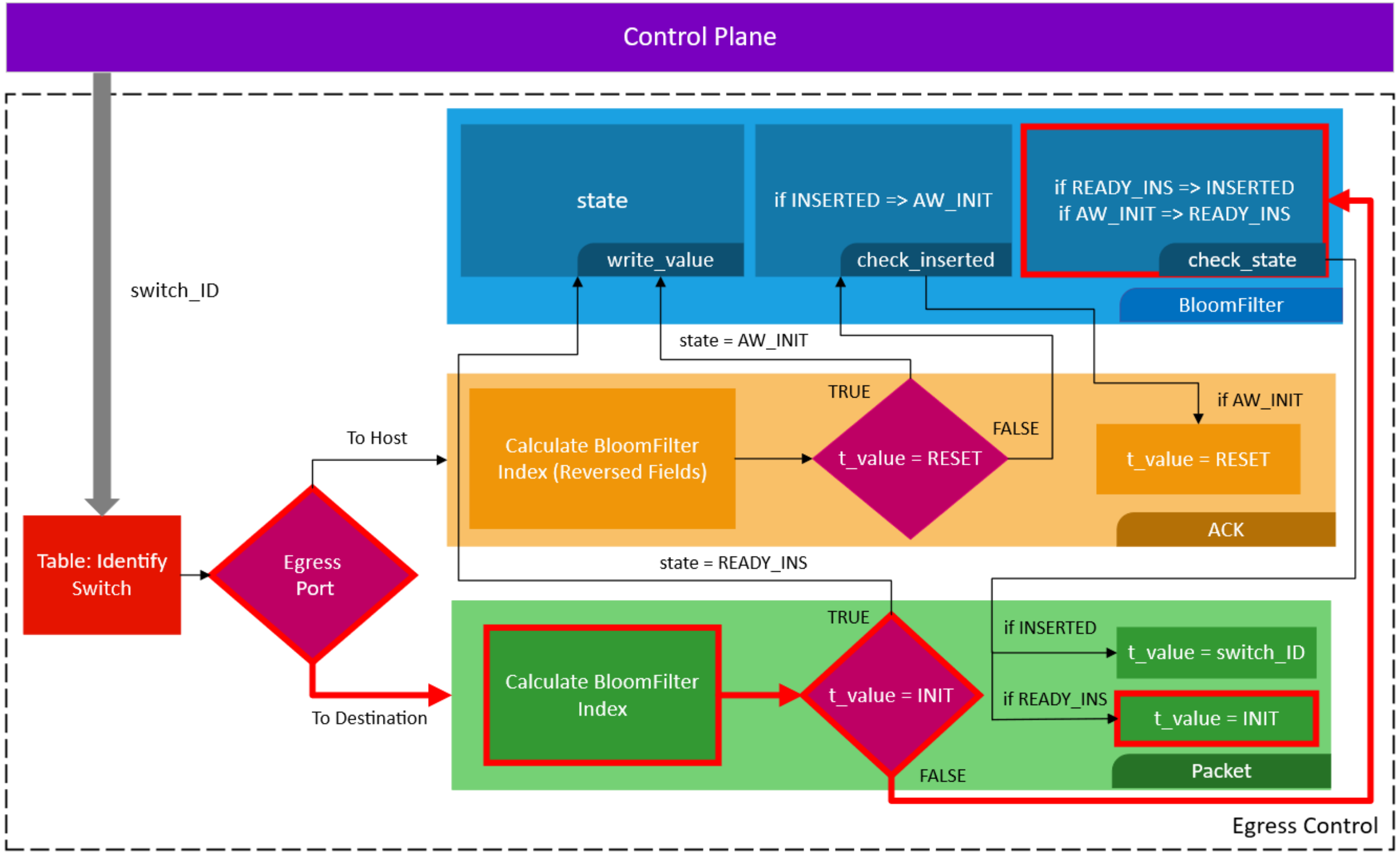
Restrictions of Register Action:

- Limited amount of resources
- Only 1 call per packet

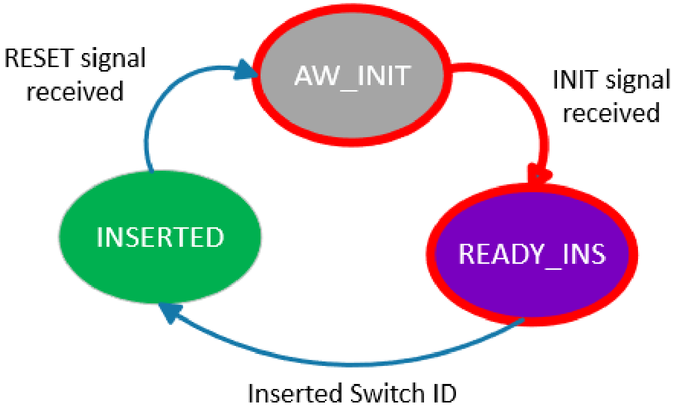
Solution:

- Different Register Action per case
- Minimzing the instructions by checking conditions in advance

DLINT Implementation: 1st transition



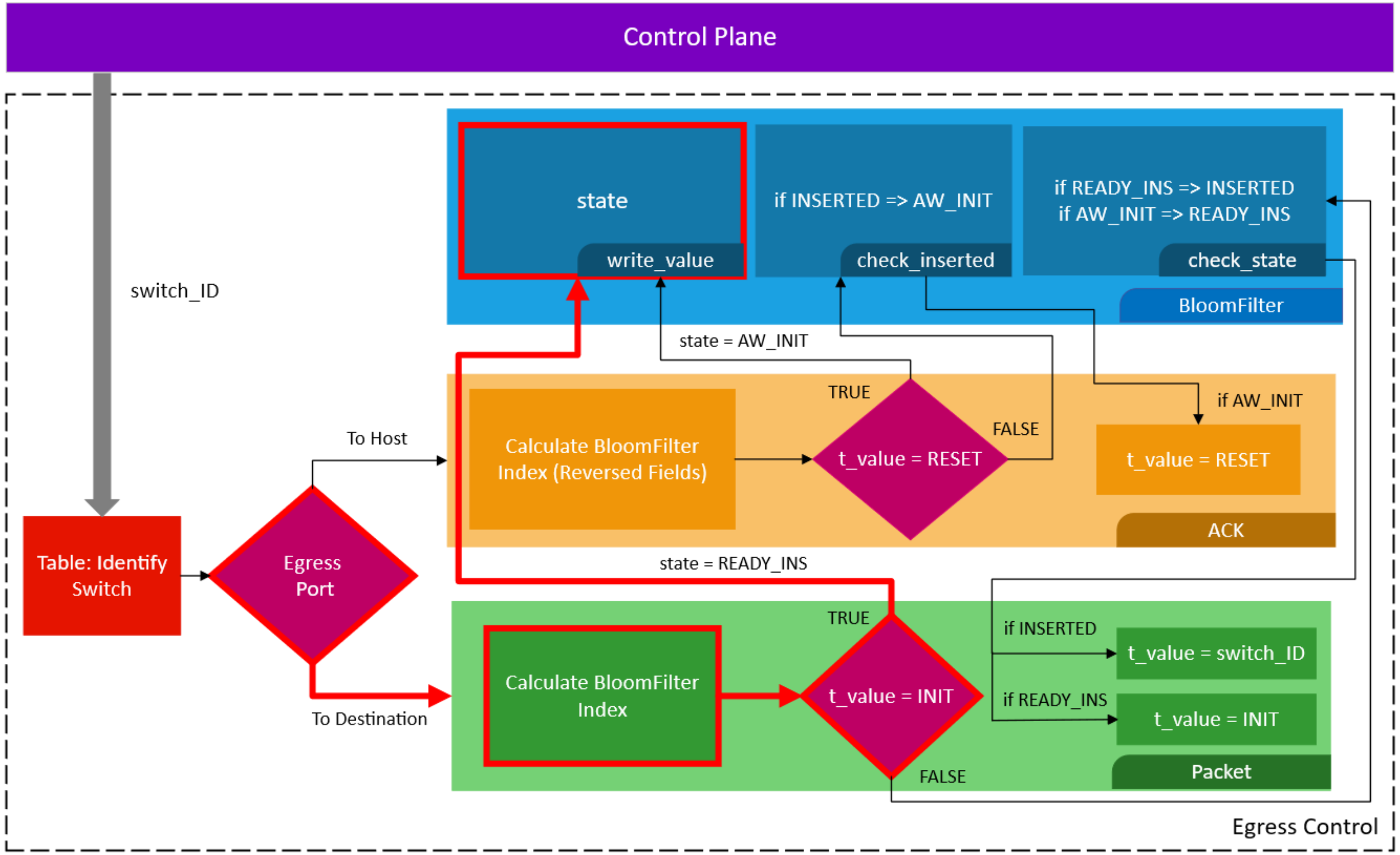
(for the first P4 Switch)



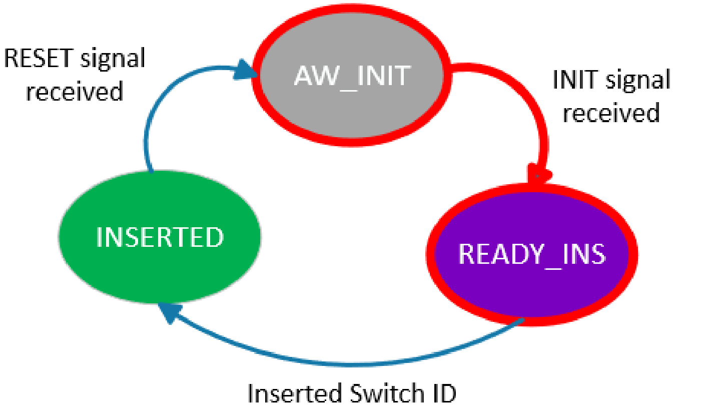
AW_INIT: Awaiting INIT signal
READY_INS: Ready to insert Switch ID
INSERTED: Inserted Switch ID

Note: Only the 1st switch will embed the INIT signal

DLINT Implementation: 1st transition

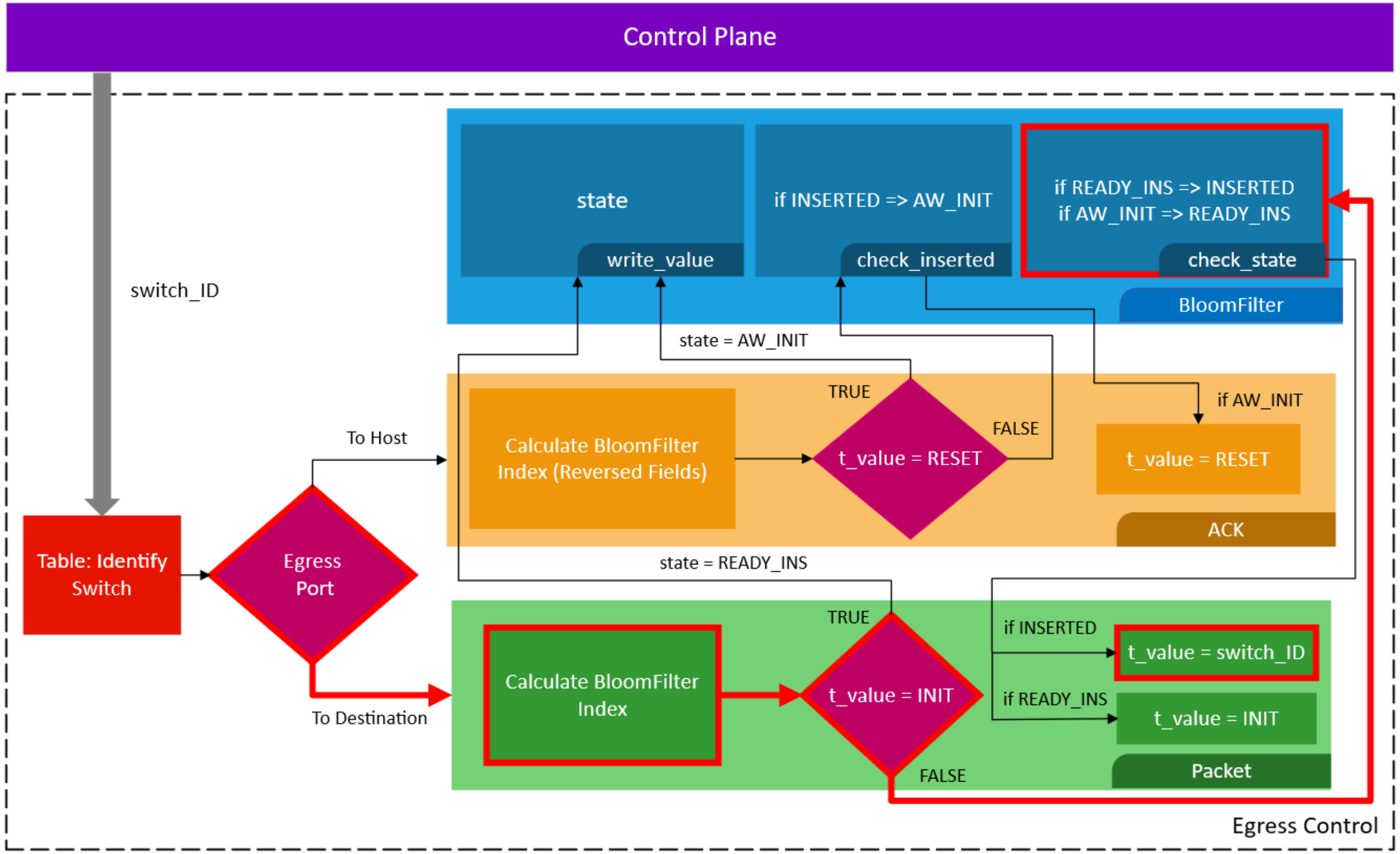


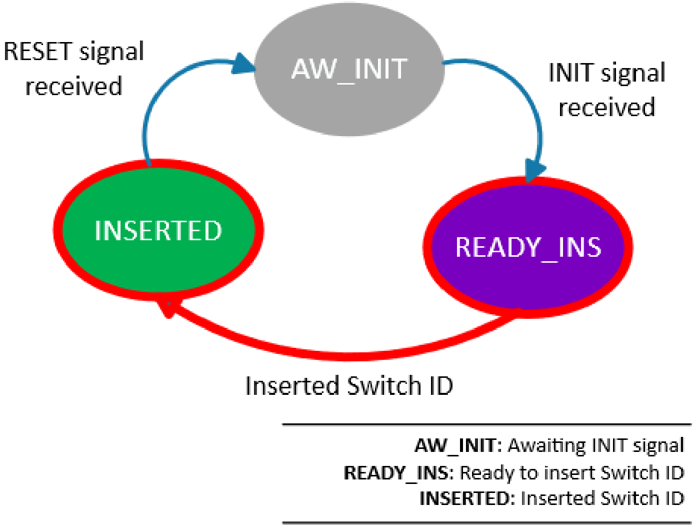
(for the remaining P4 Switches in the path)



AW_INIT: Awaiting INIT signal
READY_INS: Ready to insert Switch ID
INSERTED: Inserted Switch ID

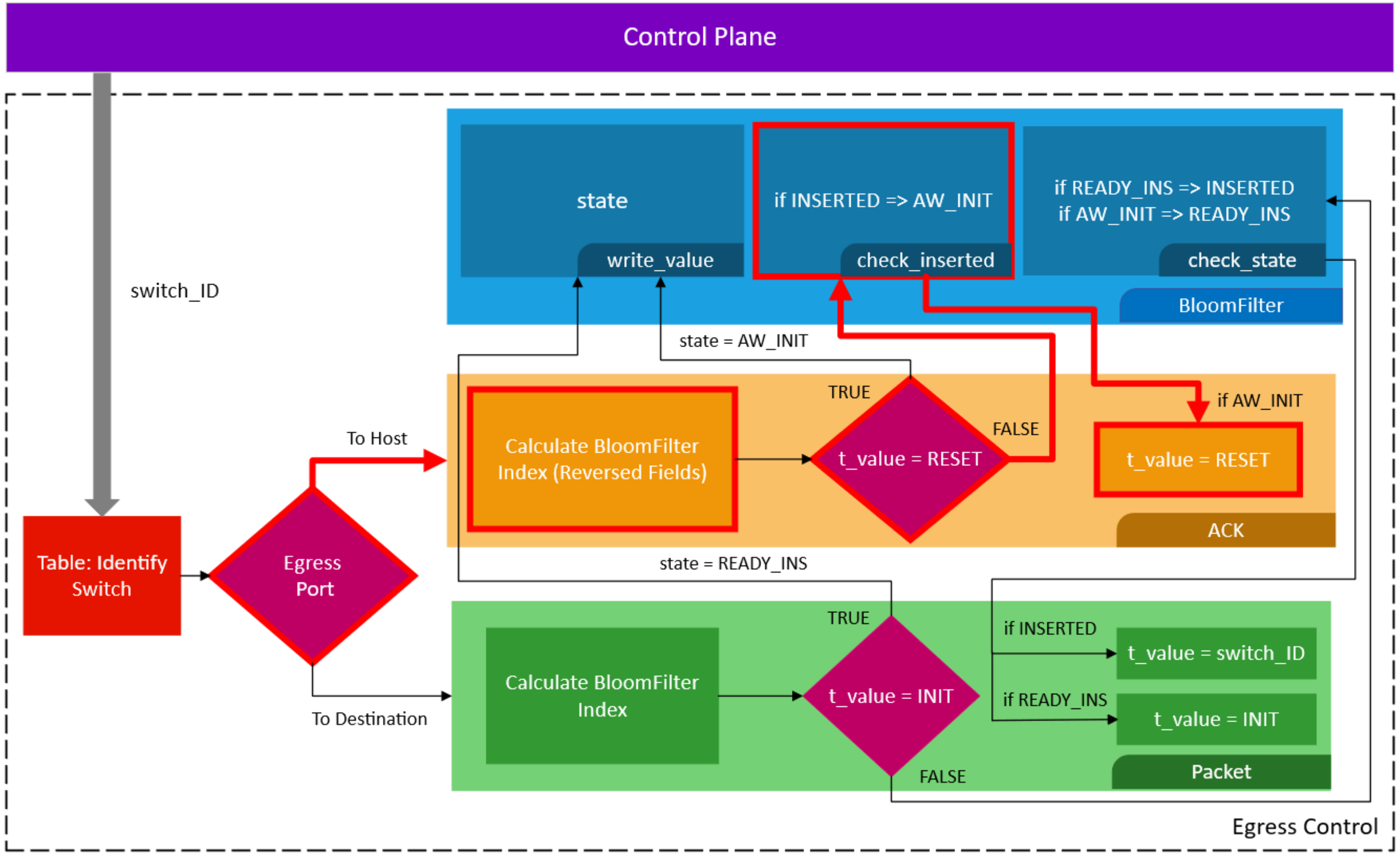
DLINT Implementation: 2nd transition



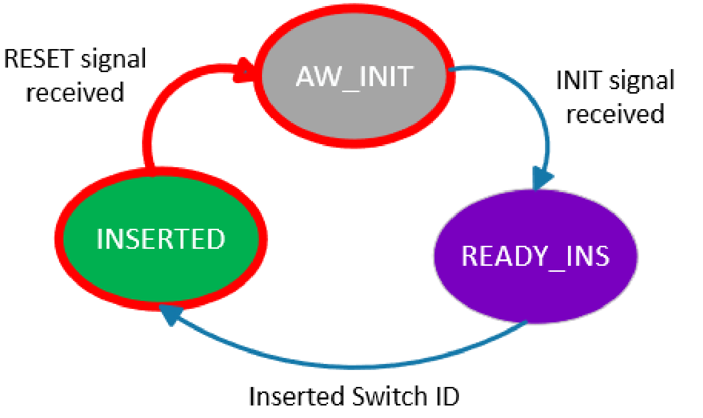

RESET signal received
AW_INIT
INIT signal received
INSERTED
READY_INS
Inserted Switch ID
AW_INIT: Awaiting INIT signal
READY_INS: Ready to insert Switch ID
INSERTED: Inserted Switch ID

Note: Using the same Register Action for 2 different transitions
=> less resource usage!

DLINT Implementation: 3rd transition



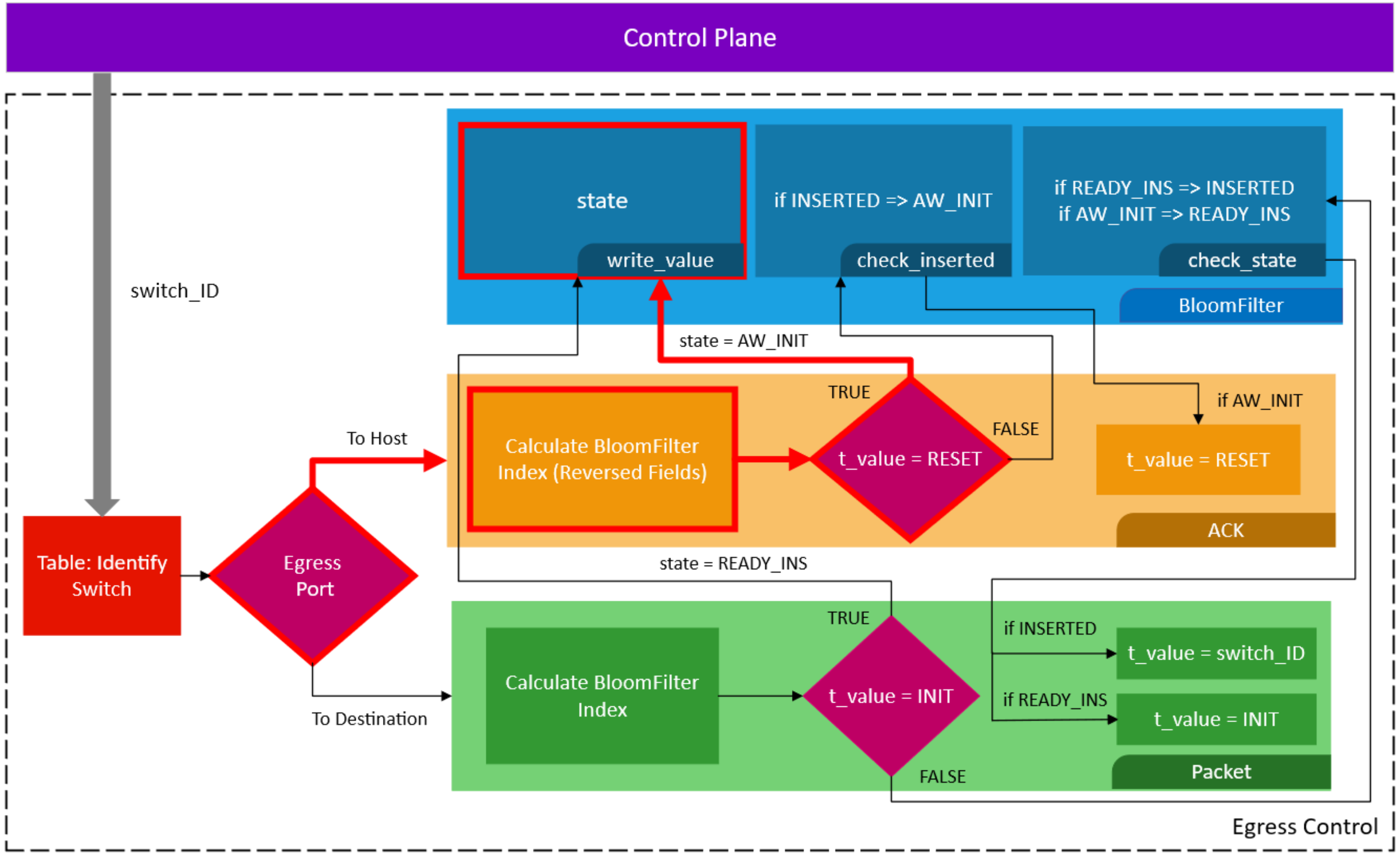
(for the first P4 Switch in the opposite direction)



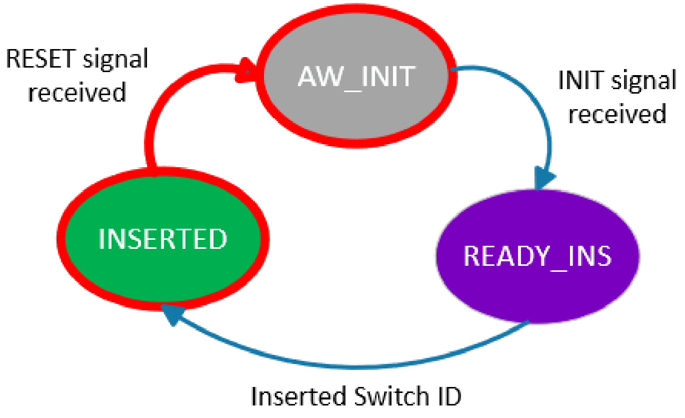
AW_INIT: Awaiting INIT signal
READY_INS: Ready to insert Switch ID
INSERTED: Inserted Switch ID

Note: Only the last switch will embed the RESET signal

DLINT Implementation: 3rd transition

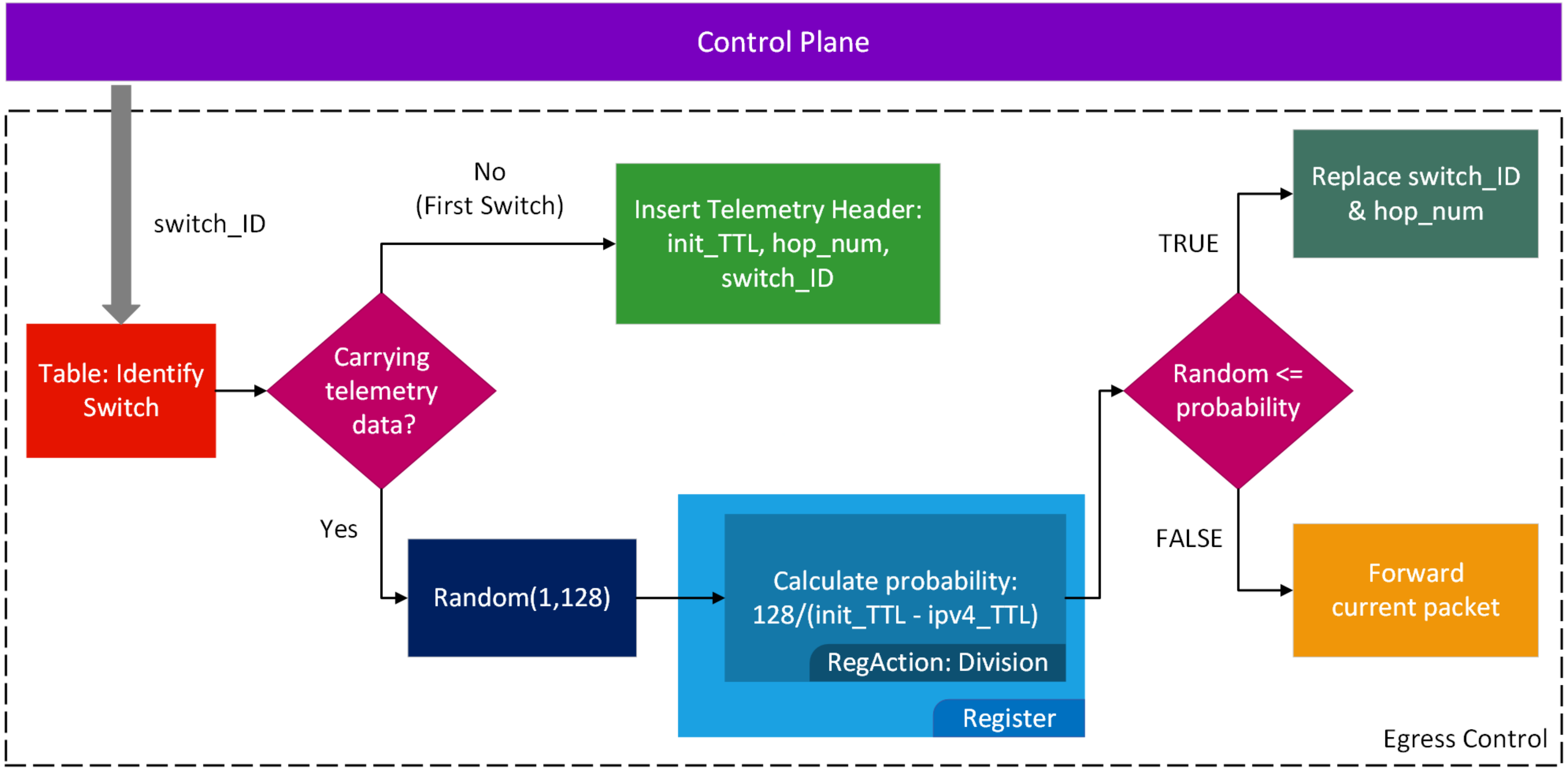


(for the remaining P4 Switches)



AW_INIT: Awaiting INIT signal
READY_INS: Ready to insert Switch ID
INSERTED: Inserted Switch ID

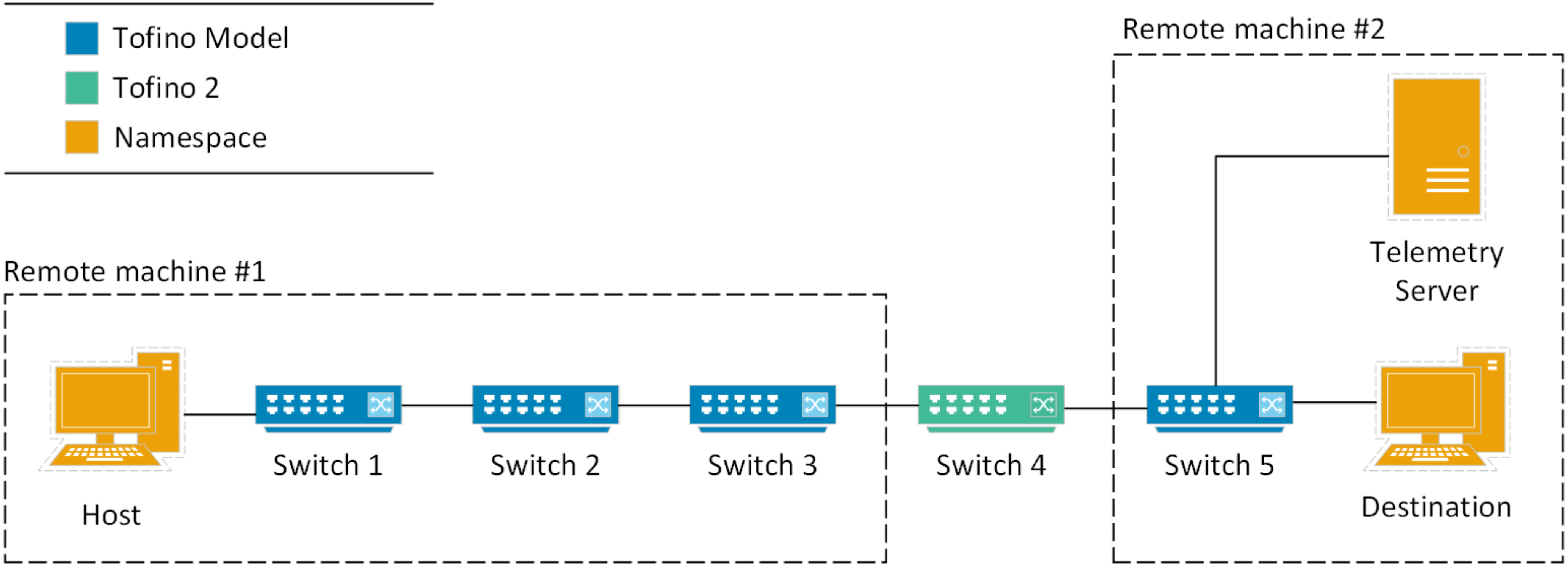
PLINT Implementation on Tofino



Challenge: Performing division on the Tofino Programmable Switch

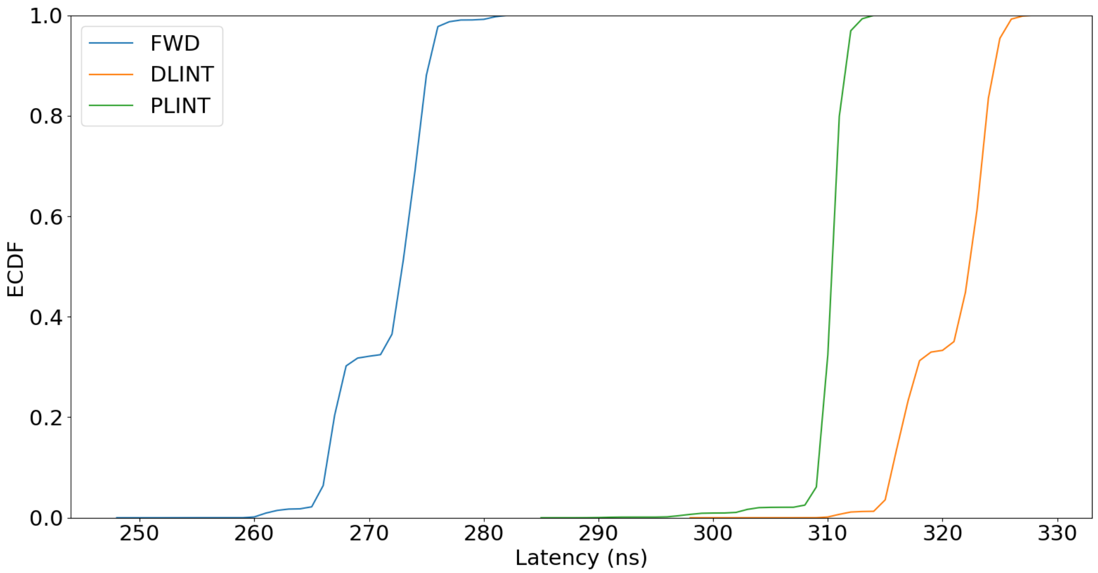
Solution: Division is supported on Tofino 2, by the SALU of the Register

Performance Evaluation



Evaluation Environment

Evaluation Results: Processing Delay



ECDF: 3 Traffic flows: 5, 10, and 20 Mbps

FWD: Simple IPv4 Forwarding

Processing Delay Increase compared to FWD

PLINT: +12%

DLINT: +17%

Evaluation Results: Resources Utilization

Resource	DLINT	PLINT	FWD
Stages	7	7	2
SRAM	0.4%	0.3%	0.1%
TCAM	0.4%	0.4%	0.4%

Comparing stages and RAM per method

Power type	DLINT	PLINT	FWD
Weight	171.5	150.8	36.2
Worst-case Power (W)	1.25	1.12	0.35

Comparing Power Consumption per method

Noticeable difference in SRAM due to the register used in DLINT

Weight: Unit-less metric representing relative resource usage in each block of the pipeline

DLINT: +12% more power-consuming in worst-case scenario

Future work

- Results based on the usage of multiple hash functions for indexing the register (for mitigating hash conflicts)
    - Measuring performance in retrieving all the metadata from the switches
- Performance of machine learning tasks when the values are collected in-band
- Performance impact of encrypting the collected metadata on the data plane
- Deployment of both approaches on a multi-domain P4-programmable network (e.g. 2STIC, FABRIC)

References

[1] M. Yu, “Network telemetry: Towards A Top-Down Approach,” *ACM SIGCOMM Computer Communication Review*, vol. 49, no. 1, 2019.

[2] Tan, Lizhuang, et al. "In-band network telemetry: A survey." *Computer Networks* 186 (2021): 107763.

[3] Papadopoulos, Konstantinos, Panagiotis Papadimitriou, and Chrysa Papagianni. "Deterministic and Probabilistic P4-Enabled Lightweight In-Band Network Telemetry." *IEEE Transactions on Network and Service Management* (2023).

[4] Angelos Dimoglis, Leandro C. de Almeida, Konstantinos Papadopoulos, Chrysa Papagianni, Panagiotis Papadimitriou, and Paola Grosso. “Lightweight INT on the Tofino Programmable Switch”.

Thank you