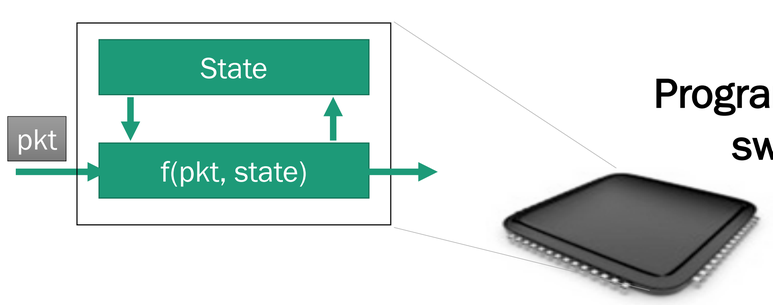
Relaxing constraints in stateful network data plane design

Carmelo Cascone^, Roberto Bifulco*, Salvatore Pontarelli+, Antonio Capone^
^ Politecnico di Milano (Italy), *NEC Laboratories Europe (Germany), +Univ. Roma Tor Vergata (Italy)

1. INTRODUCTION

Program and run at line rate algorithms that read and modify data plane's state

- E.g.: Stateful firewall, dynamic NAT, flowlet load balancing, AQM, measurement, etc.

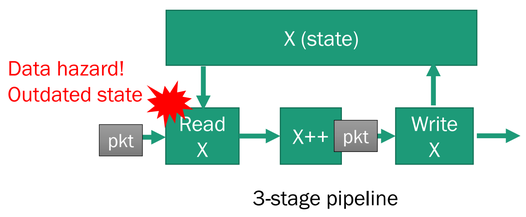


Programmable *stateful* switching chip

2. PROBLEM STATEMENT

- Pipelining is the way to scale for high throughput (Tb/s)
- When pipelining, accessing state at different stages of the pipeline can cause data hazards

Example:
A packet counter. For each packet increase the value of X.



3-stage pipeline

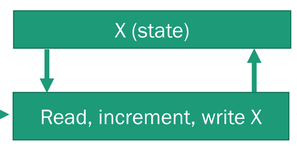
State of the art

- Reconfigurable Match Table (RMT) [SIGCOMM 2013]
  - Programmable parser, actions, table size, stateless
  - 640 Gb/s non-blocking line rate, 6.5 Tb/s Tofino Chip
- Domino/Banzai [SIGCOMM 2016]
  - Extends RMT with support for stateful actions named "Atoms"
  - Atoms can be pipelined to implement complex algorithms
  - Strict constraints on the atom execution time
  - State read-modify-write must be executed in 1 clock tick

Pros

If an algorithm can be mapped to an atom pipeline, then it is guaranteed to run at line rate, with any traffic workload.

Banzai atomic model



Cons

Memory must operate at pipeline frequency. Limited size, can't support slower but larger memories (e.g. DRAM)

Support limited actions with simple circuitry that can meet pipeline timing. E.g. Can't do SQRT in 1 clock cycle at 1Ghz

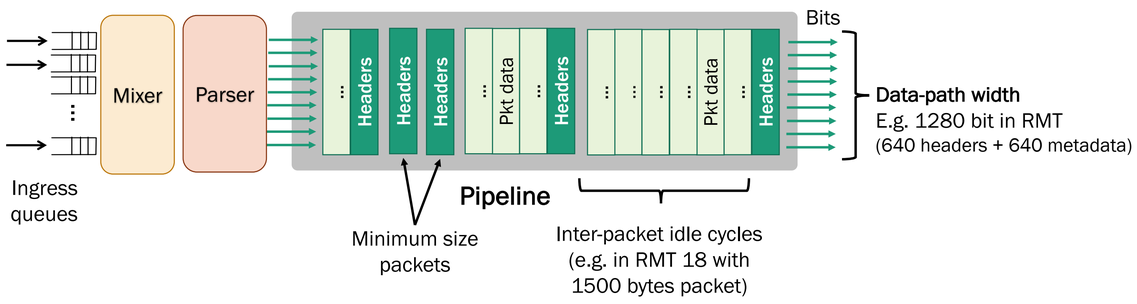
Research question:

- If we allow for stateful processing blocks that span many clock cycles:
  - What is the risk of data hazards with realistic traffic conditions?
  - What is the throughput when using locking to access the memory?
  - How much silicon is needed to implement such a locking scheme?

3. OBSERVATIONS

1. Pipeline's header processing rate depends on packet size

- Packets are read from input ports in chunks, e.g. 80 bytes in RMT
  - 80 bytes * 1 Ghz (chip clock freq.) = 640 Gb/s line rate
- Larger packets cause inter-packet idle cycles
  → Minimize risk of data hazards



2. Distinguish between per-flow and global state

- *Global*: shared among all packets
- *Per-flow*: shared by packets of the same flow
  → Different flows can be processed in parallel
- E.g. a stateful firewall needs only per-flow state, a DNAT both

Partly funded by EU H2020 projects:
"BEBA" grant agreement 644122
"VirtuWind" grant agreement 671648

4. MOTIVATING EXPERIMENTS

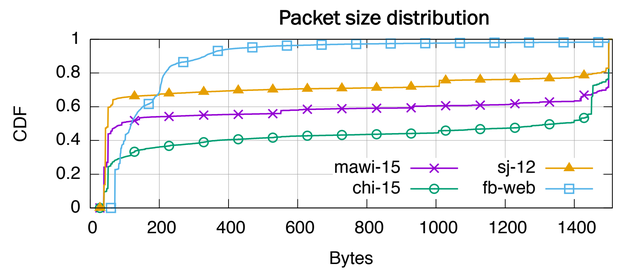
Evaluate the risk of data hazards using simulations with real traffic traces

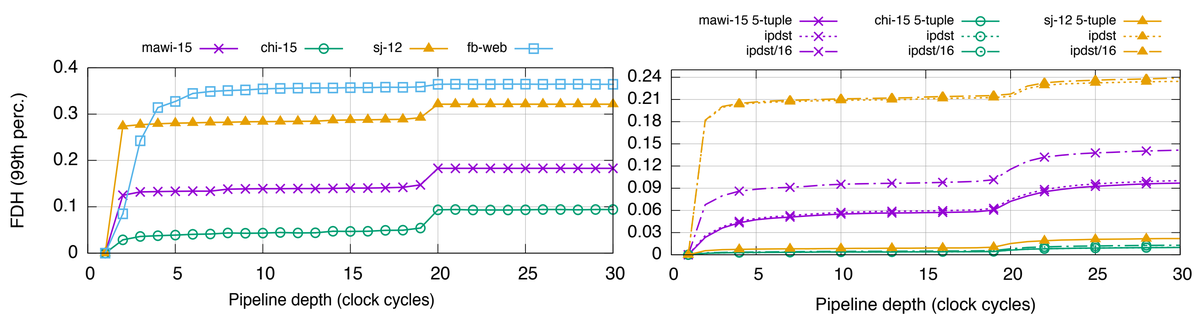
Trace	Provider	Description	Date	Num pkts	Num flows per 1m pkts		
					5-tuple	ipdst	ipdst/16
chi-15	CAIDA	10Gb/s backbone link in Chicago. Usual conditions.	Feb 19, 2015	3.5b	100.6k	57.7k	4.6k
sj-12	CAIDA	10Gb/s backbone link in San Jose. Unusually high number of 5-tuples.	Nov 15, 2012	3.6b	249k	17k	2k
mawi-15	MAWI	1Gb/s backbone link in Japan. High volume of anomalous traffic.	Jul 21, 2015	135m	40.8k	17.3k	1.7k
fb-web	Facebook	Packet samples from 10 most active ToR switches in a web cluster	2015	447m	n/a	n/a	n/a

Simulation results:

Fraction of data hazards (FDH) per batch of 100k packets

- Pipeline's read chunk = 80 bytes
- All packets back-to-back
  i.e. 100% pipeline utilization

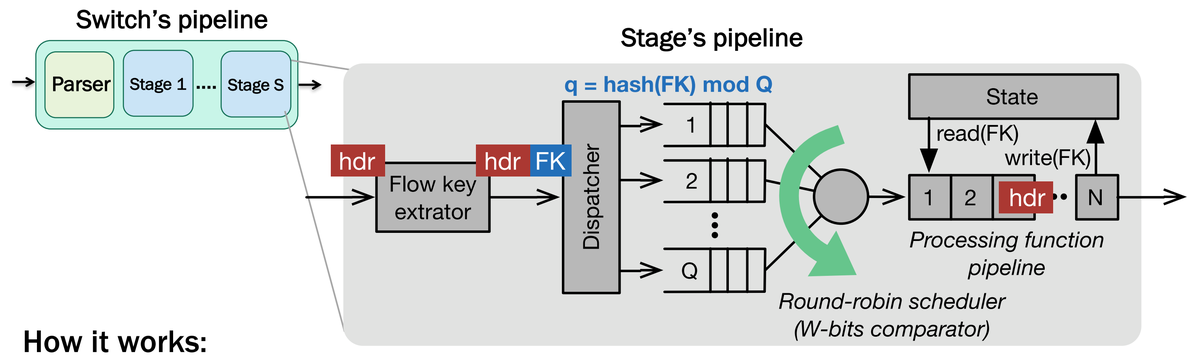




Global state *Per-flow state*

5. APPROACH: MEMORY LOCKING

If two packets of the same flow arrive back-to-back, processing is paused for the second packet until the first one has left the stage pipeline.

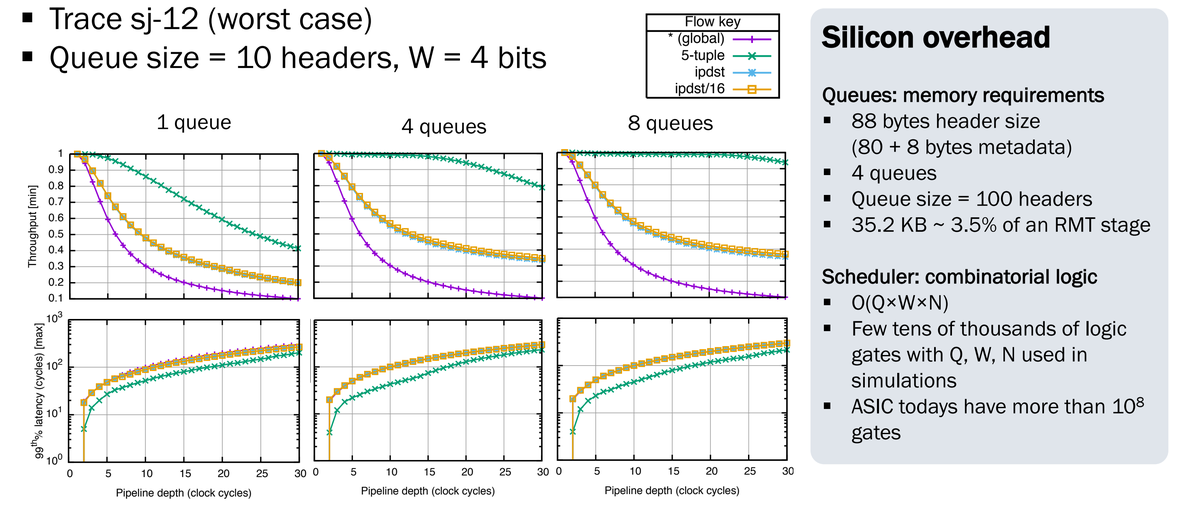


How it works:

1. Each packet is associated with a flow key (FK)
2. A dispatcher enqueues packets by hashing on the FK
3. A round-robin scheduler decides if a queue can be served by looking at the head-of-line's FK and comparing it to what is currently in the stage's pipeline.
4. Comparison is performed by reducing the space of the FK to few bits (W)

Simulation results:

- Trace sj-12 (worst case)
- Queue size = 10 headers, W = 4 bits



Silicon overhead

Queues: memory requirements

- 88 bytes header size (80 + 8 bytes metadata)
- 4 queues
- Queue size = 100 headers
- 35.2 KB ~ 3.5% of an RMT stage

Scheduler: combinatorial logic

- O(Q×W×N)
- Few tens of thousands of logic gates with Q, W, N used in simulations
- ASIC todays have more than 10^8 gates

Clock cycle budget (and latency) for all traces:

Maximum number of clock cycles (limited to 30) per processing function, to sustain a given throughput.
W = 4 bits. Latency values are given for 1 Ghz clock frequency, i.e. 1 clock cycle = 1 ns.

Thrpt	Q_{len}	Q	chi-15 5-tuple	chi-15 ipdst	chi-15 ipdst/16	chi-15 global	sj-12 5-tuple	sj-12 ipst	sj-12 ipdst/16	sj-12 global	mawi-15 5-tuple	mawi-15 ipdst	mawi-15 ipdst/16	mawi-15 global	fb-web global
100%	10	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		4	1	1	1		1	1	1		1	1	1		
		8	1	1	1		1	1	1		1	1	1		
		16	1	1	1		1	1	1		1	1	1		
	100	1	20 (174ns)	20 (190ns)	21 (230ns)	8 (282ns)	4 (49ns)	1	1	1	2 (18ns)	2 (20ns)	2 (35ns)	1	2 (10ns)
		4	30 (175ns)	30 (192ns)	30 (320ns)		8 (152ns)	1	1		2 (12ns)	2 (14ns)	2 (25ns)		
		8	30 (137ns)	30 (144ns)	30 (259ns)		8 (133ns)	1	1		2 (12ns)	2 (14ns)	2 (25ns)		
		16	30 (122ns)	30 (126ns)	30 (221ns)		8 (123ns)	1	1		2 (11ns)	2 (14ns)	2 (24ns)		
99.9%	10	1	8 (16ns)	8 (16ns)	8 (18ns)	4 (18ns)	2 (5ns)	1	1	1	1	1	1	1	2 (10ns)
		4	14 (33ns)	14 (31ns)	14 (38ns)		2 (4ns)	1	1		1	1	1		
		8	16 (39ns)	15 (30ns)	16 (44ns)		2 (4ns)	1	1		1	1	1		
		16	17 (37ns)	18 (43ns)	16 (42ns)		2 (5ns)	1	1		1	1	1		
	100	1	27 (568ns)	27 (618ns)	26 (605ns)	8 (282ns)	6 (143ns)	2 (86ns)	2 (84ns)	1	3 (42ns)	3 (48ns)	3 (100ns)	2 (60ns)	2 (10ns)
		4	30 (175ns)	30 (192ns)	30 (320ns)		15 (526ns)	2 (79ns)	2 (77ns)		4 (41ns)	4 (52ns)	4 (135ns)		
		8	30 (137ns)	30 (144ns)	30 (259ns)		22 (731ns)	2 (79ns)	2 (78ns)		4 (38ns)	4 (50ns)	4 (131ns)		
		16	30 (122ns)	30 (126ns)	30 (221ns)		25 (741ns)	2 (79ns)	2 (72ns)		4 (37ns)	4 (49ns)	4 (129ns)		
99%	10	1	21 (80ns)	21 (80ns)	21 (89ns)	7 (60ns)	3 (14ns)	1	1	1	2 (13ns)	2 (14ns)	1	1	2 (10ns)
		4	30 (142ns)	30 (148ns)	30 (184ns)		10 (45ns)	1	1		5 (34ns)	4 (31ns)	2 (18ns)		
		8	30 (129ns)	30 (138ns)	30 (184ns)		11 (47ns)	1	1		6 (42ns)	4 (31ns)	2 (18ns)		
		16	30 (116ns)	30 (122ns)	30 (180ns)		12 (47ns)	1	1		7 (52ns)	4 (31ns)	2 (18ns)		
	100	1	30 (950ns)	30 (922ns)	30 (1.1us)	9 (842ns)	8 (268ns)	2 (86ns)	2 (84ns)	2 (171ns)	9 (422ns)	8 (380ns)	5 (316ns)	4 (379ns)	2 (10ns)
		4	30 (175ns)	30 (192ns)	30 (320ns)		22 (1.1us)	3 (285ns)	3 (283ns)		24 (1.7us)	17 (1.3us)	7 (572ns)		
		8	30 (137ns)	30 (144ns)	30 (259ns)		30 (1.9us)	3 (290ns)	3 (292ns)		30 (2.1us)	23 (2.2us)	8 (753ns)		
		16	30 (122ns)	30 (126ns)	30 (221ns)		30 (940ns)	3 (296ns)	3 (293ns)		30 (1.3us)	25 (2.5us)	8 (759ns)		