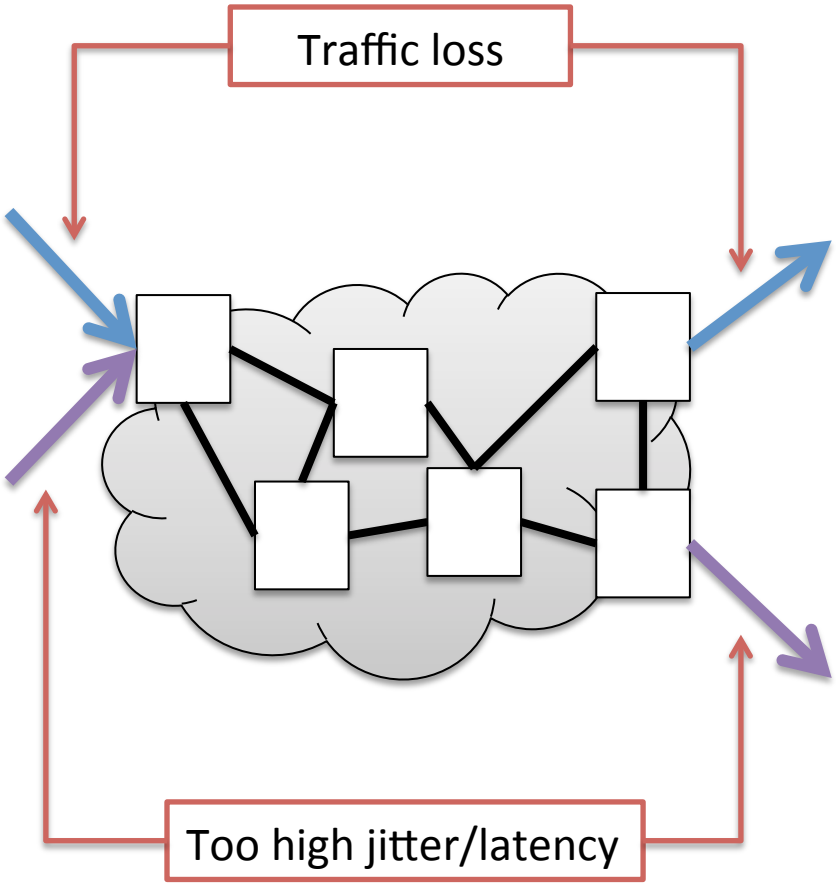
Strategies of packet buffering inside Routers

Rafal Jan Szarecki #JNCIE136
Solution Architect, Juniper Networks

Why should I care ?

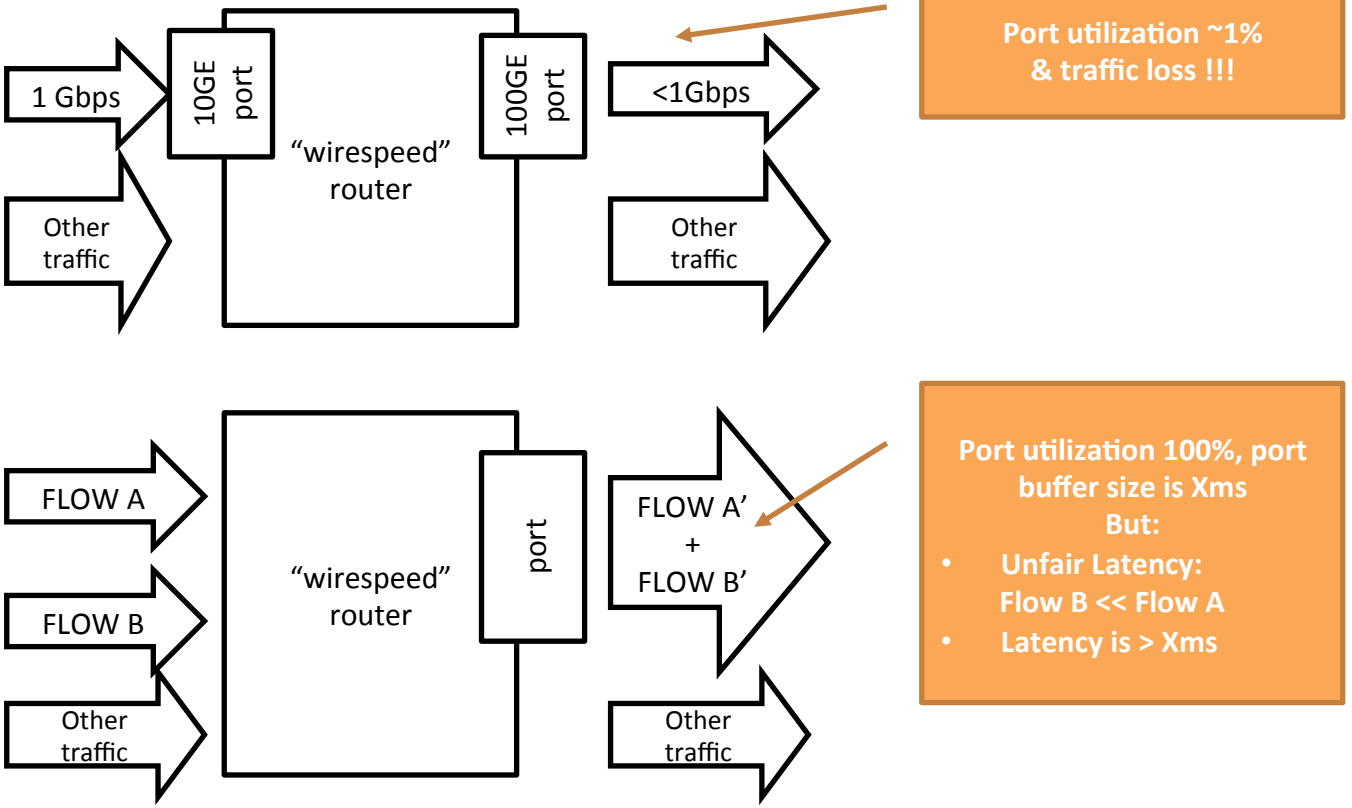
- Under load, Your router's discard behavior according to queuing strategy selected by vendor.
  - Could be quite unintuitive !
  - Better to know, how to live/deal with this artificial "intelligence". And turn it for your benefit.
  - Do not troubleshoot if there is nothing unexpected/misbehaving.

How it manifest



- Something is going on
  - SLA monitoring system rise alarm
  - Customer calls and complains
- Which node in network cause it?
  - You may need go node-by node.
  - Other expert/analitic systems my help.
  - Out of scope
- When guilty node is nail down …

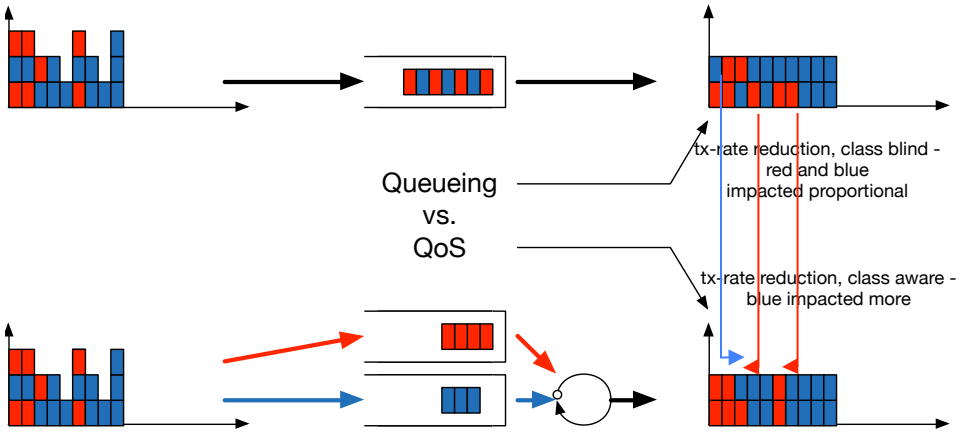
Unintuitive behavior



Defect (bug) or expected behavior ?

Queuing != QoS

- Queuing goal – avoid traffic/ packet drops during temporal congestion

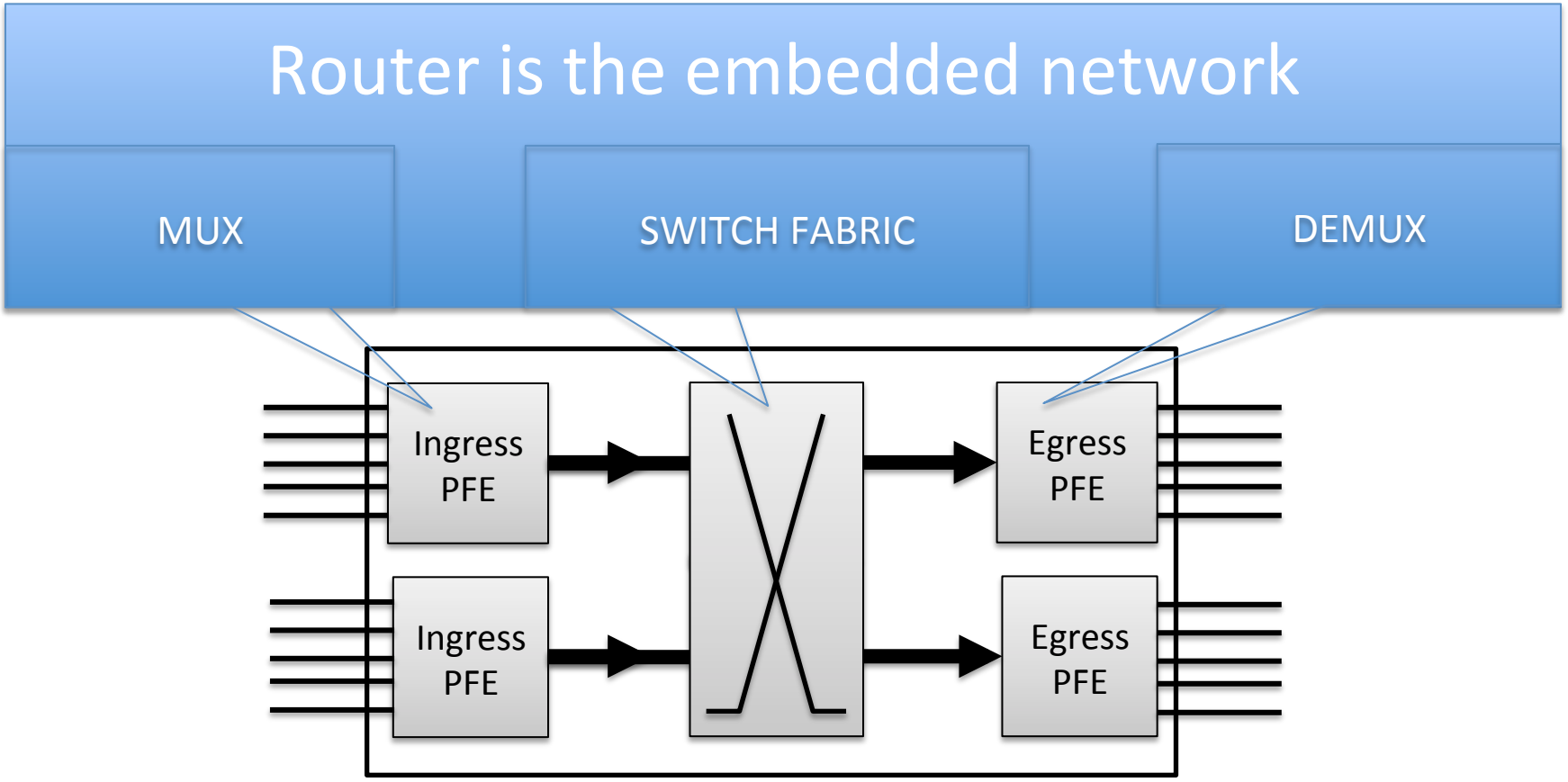


- QoS goal - provide differential treatment and separation among traffic of different classes.
  - Avoid traffic/packet drops during temporal congestion in some classes at expenses of losses in other.
  - Re-order packets in a way to deliver data of some classes as fast as possible.

This talk is not on QoS.

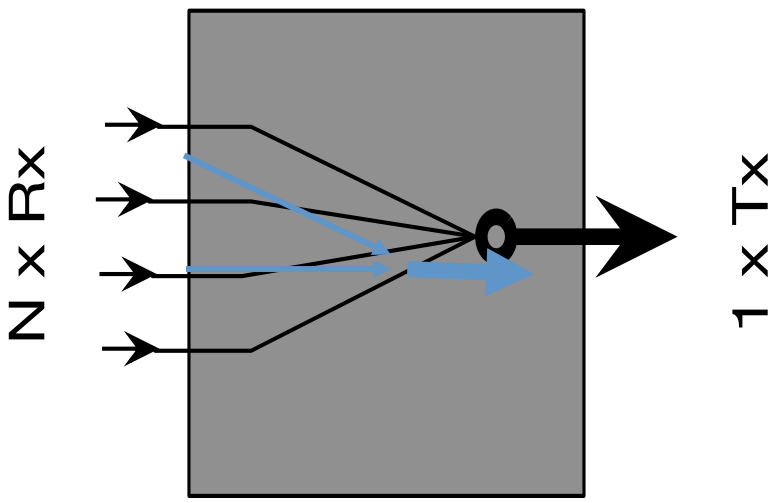
We will look at best-effort only

Router anatomy



PFE – a CPU, NPU or ASIC that process packet

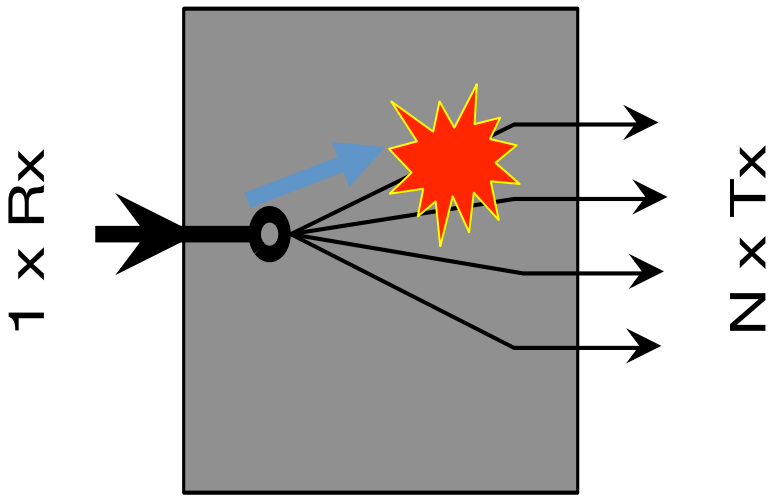
The MUX



N x Rx rate <= 1 x Tx rate

- Multiple low-speed In (Rx) port and
- Single high-speed Out (Tx) interfaces
- Many to One
- No congestion risk – no need for buffer

The de-mux

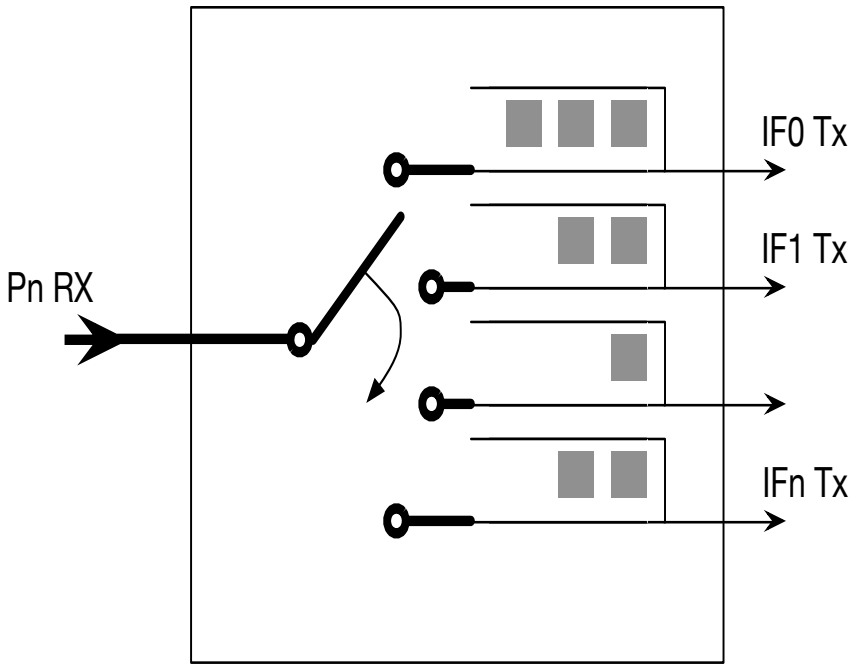


Rx rate == N x Tx rate

- Simple model
- High-speed In (Rx) port and
- Multiple lower-speed Out (Tx) interfaces
- One to many

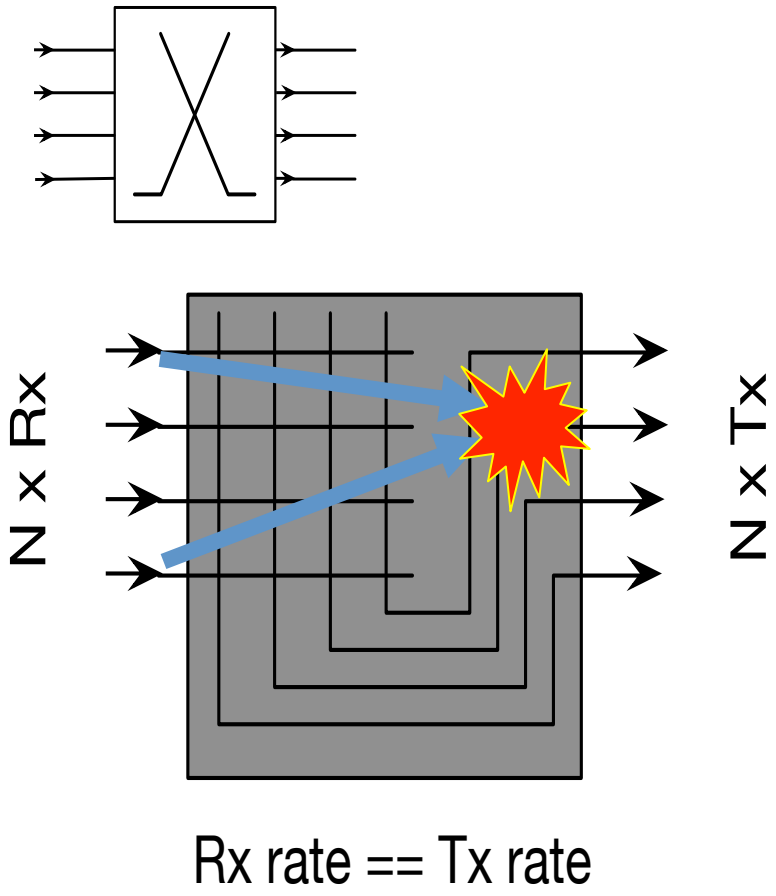
- 1 Rx to 1 Tx @ same time ➔ Congestion.

Queuing architectures for de-mux



- Simple Output Queuing – OQ
- Usually implemented in single (shared) memory

The (asynchronous) switch fabric

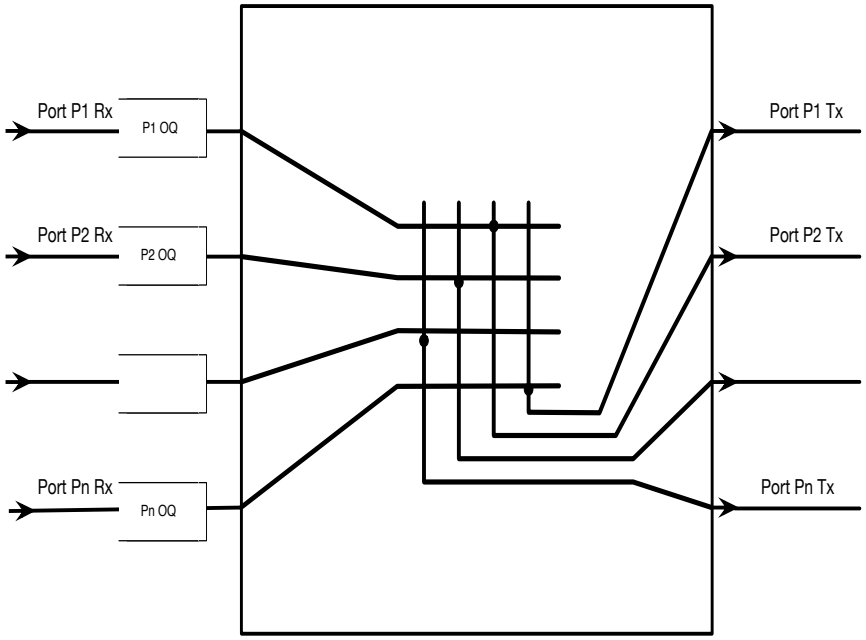


- N x In (Rx) and N x Out (Tx) port of switch of same speed
- Any to Any
- Each ingress port is independent
  - Traffic/datagram may appear at any time
  - Not aware about egress port state
- N Rx to 1 Tx @ same time ➔ Congestion.

Queuing architectures for switch fabric

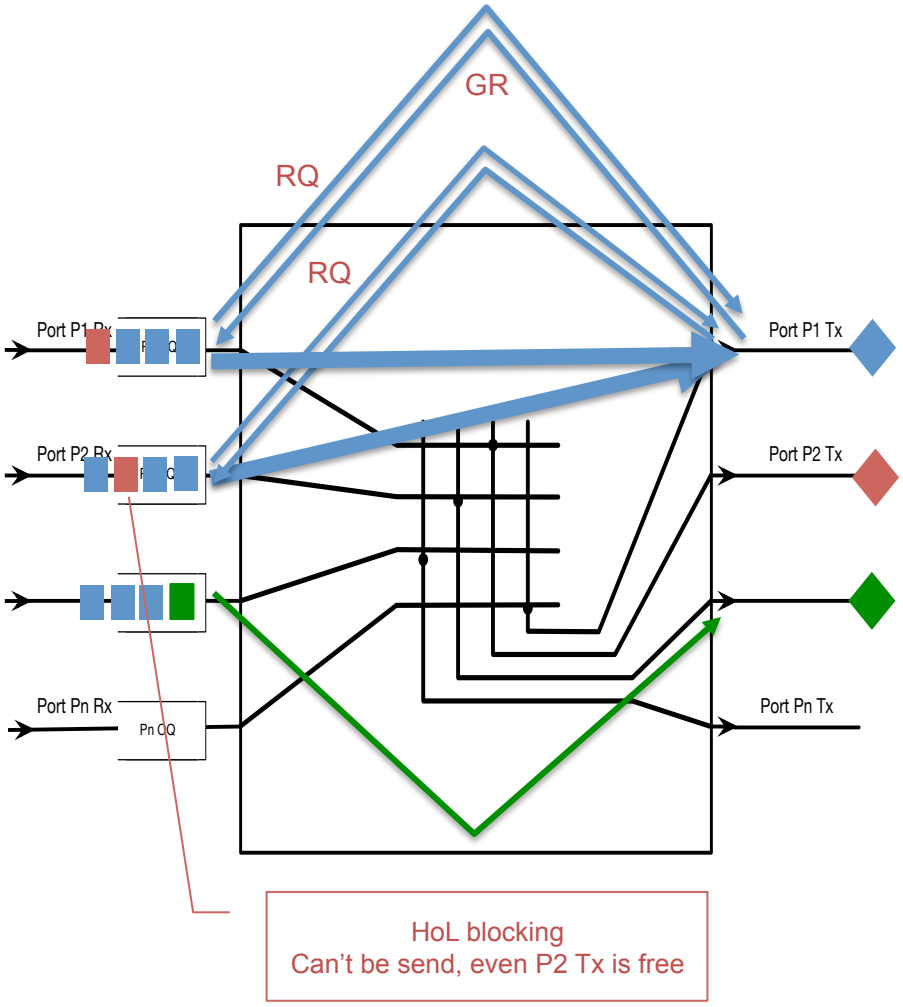
- Output Queuing – OQ – not used due to technological limitations.
- Input Queuing – IQ
- Virtual Output Queuing - VOQ

IQ



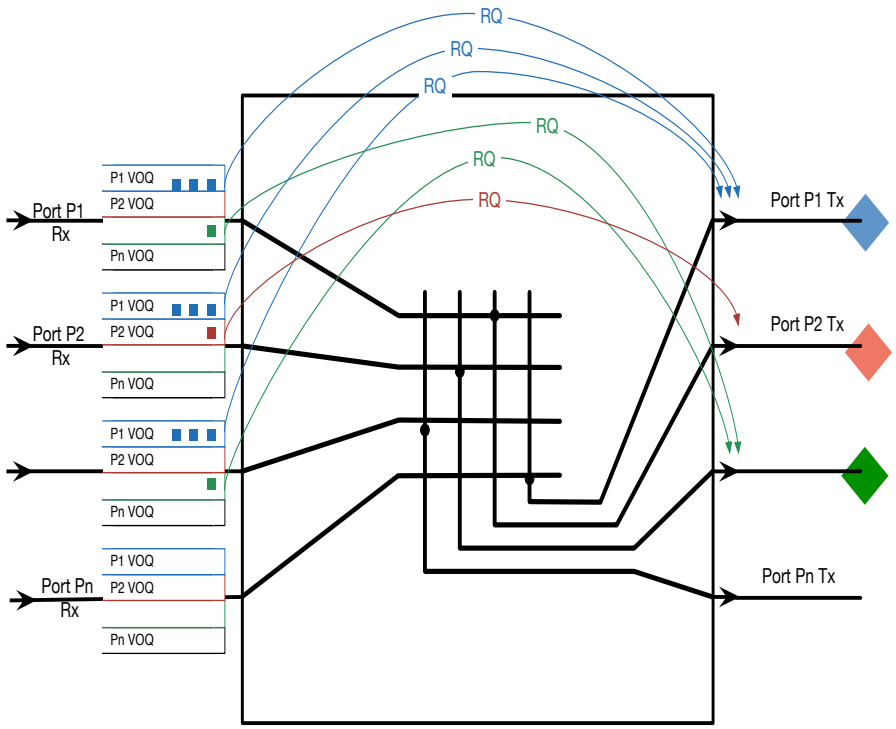
- <100% efficiency
- Queue fan-out need to be over 2 x desired port traffic to get 99%+ efficiency

IQ Flow-control



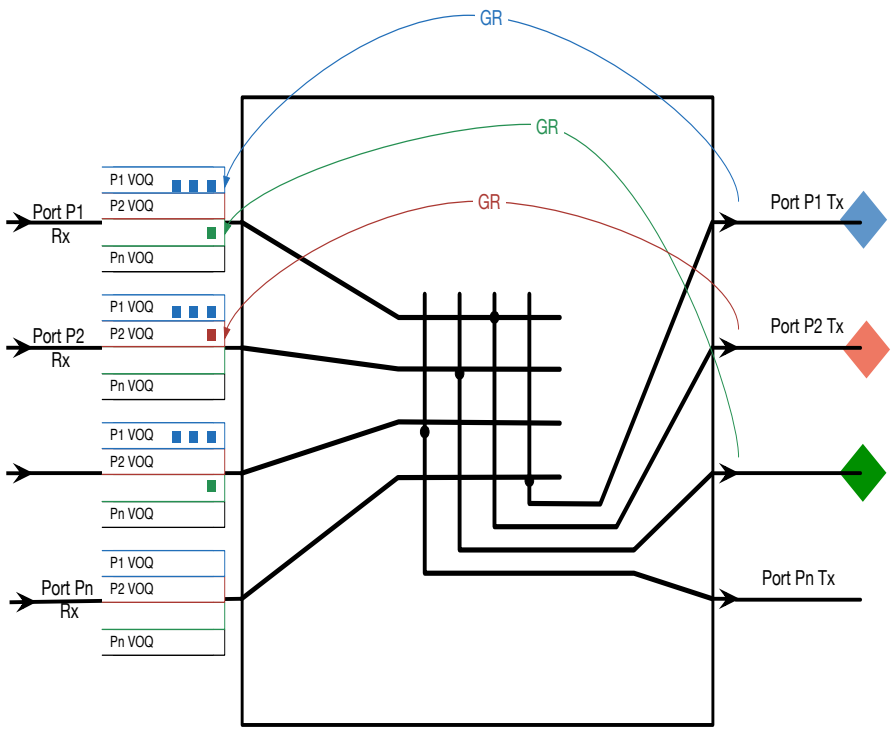
- Asynchronous – each egress is independent
- Ingress PFE sends Request
  - Each has data size
  - Only for packet at head of queue
- egress PFE answer w/ Grant – when egress Fabric port is free
- Egress schedules grants
  - Prevent starvation
  - E.g. RoundRobin or fair-share

VOQ



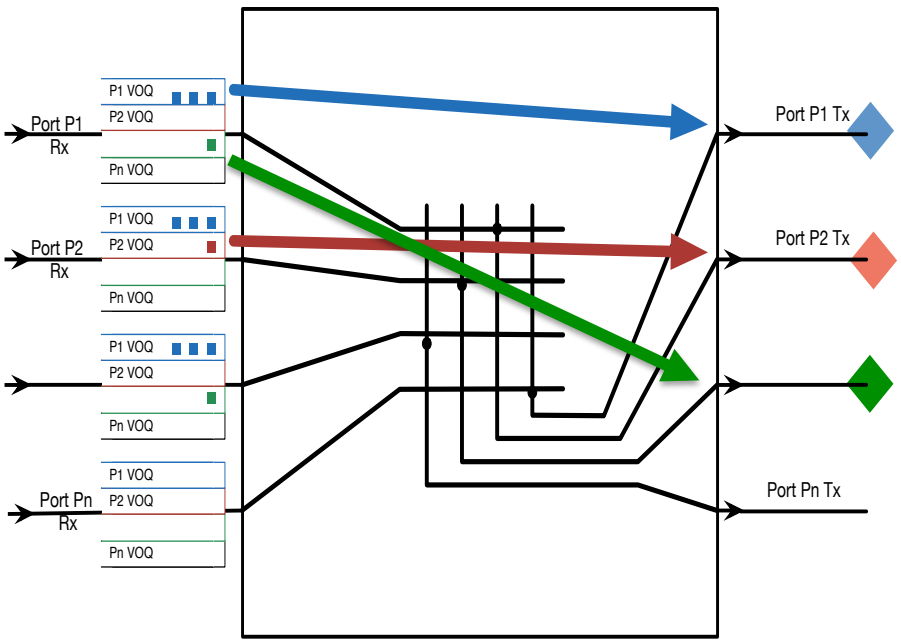
- Variant Input Queuing – "dedicated Input queue for each output port"
- No need for over-speed
- Flow-Control and scheduling
  - Extension to IQ
  - Ingress PFE can send requests to multiple egresses simultaneously

VOQ



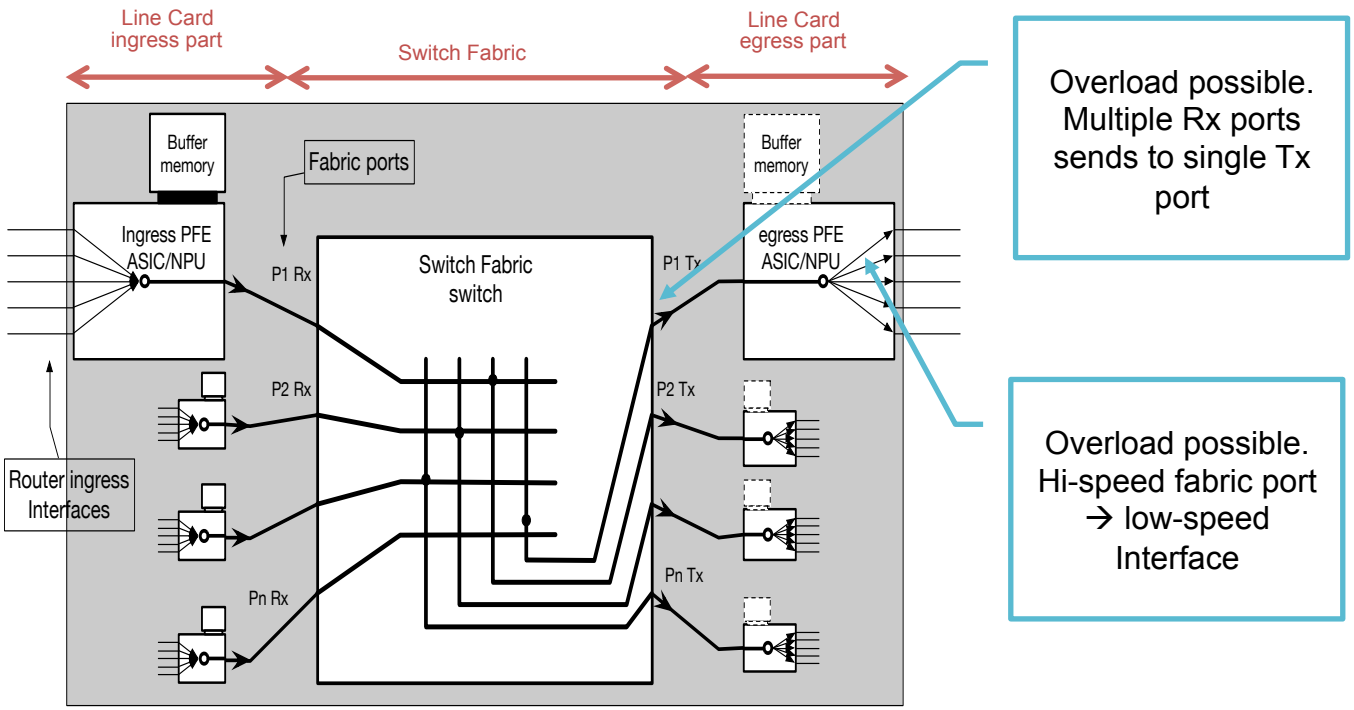
- Variant Input Queuing – "dedicated Input queue for each output port"
- No need for over-speed
- Flow-Control and scheduling
  - Extension to IQ
  - Ingress PFE can send grants to multiple egresses simultaneously

VOQ



- Variant Input Queuing – "dedicated Input queue for each output port"
- No need for over-speed
- Flow-Control and scheduling
  - Extension to IQ
  - Ingress PFE can send grants to multiple egresses simultaneously
- No HoL blocking

The router



- Multistage
  - Ingress mux to fabric
    - No congestion
  - Fabric switch
  - Egress demux to ports
    - hiSpeed (fab) to low speed port
- Congestion points
  - Fabric-out (many→ one)
  - Egress mux (fast → slow)
  - Need queuing

Two approaches

Buffer twice – CIOQ systems

Combined **I**nput **O**utput **Q**ueuing

- Buffer before fabric; de-queue when fabric egress port is available (empty) – fabric **VOQ** (or IQ)
- Buffer before egress interface; de-queue when interface is available (empty) – **OQ**

- Simpler to Implement
- Higher scalability [O(n)]
- Requires more memory
  - Space (size)
  - 2 x bandwidth
- Bigger system residency time and Jitter

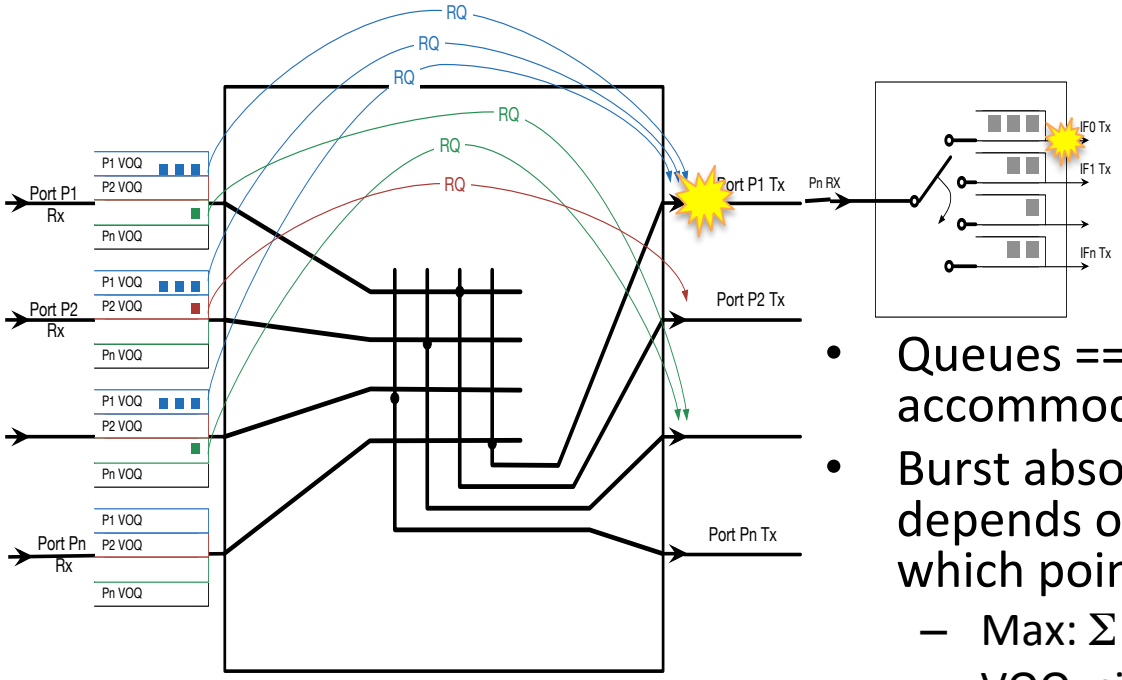
Buffer Once - VOQ systems

Virtual **O**utput **Q**ueuing

- Buffer before fabric;
  De-queue when all way down to egress interface is available (empty) – end-to-end system **VOQ**

- Requires a lot of queues - complex queue management @ scale [$O(n^2)$]
- Requires less memory
  - Space (size)
  - 1 x bandwidth
- Lower residency time (latency inside router)
- Lower power requirements

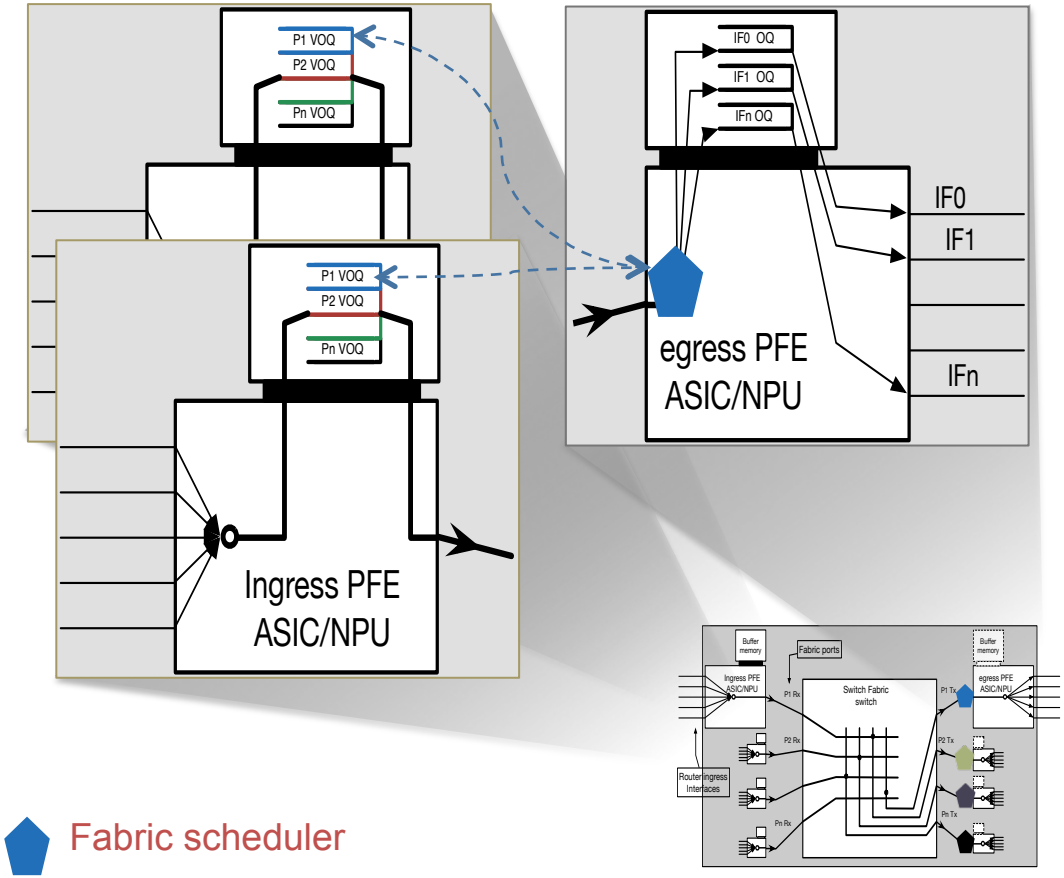
Buffer twice – latency



- Queues == Buffer == accommodate burst and **delay it**
- Burst absorption capability depends on type of burst - at which point congestion appears
  - Max: Σ (VOQ size, OQ size)
  - VOQ_size or OQ _size if congestion in one point only.
- Max latency: Σ (VOQ size, OQ size)
- Wait, there will be example

Potential congestion point

Buffer twice – bandwidth Fabric Scheduler and flow-control

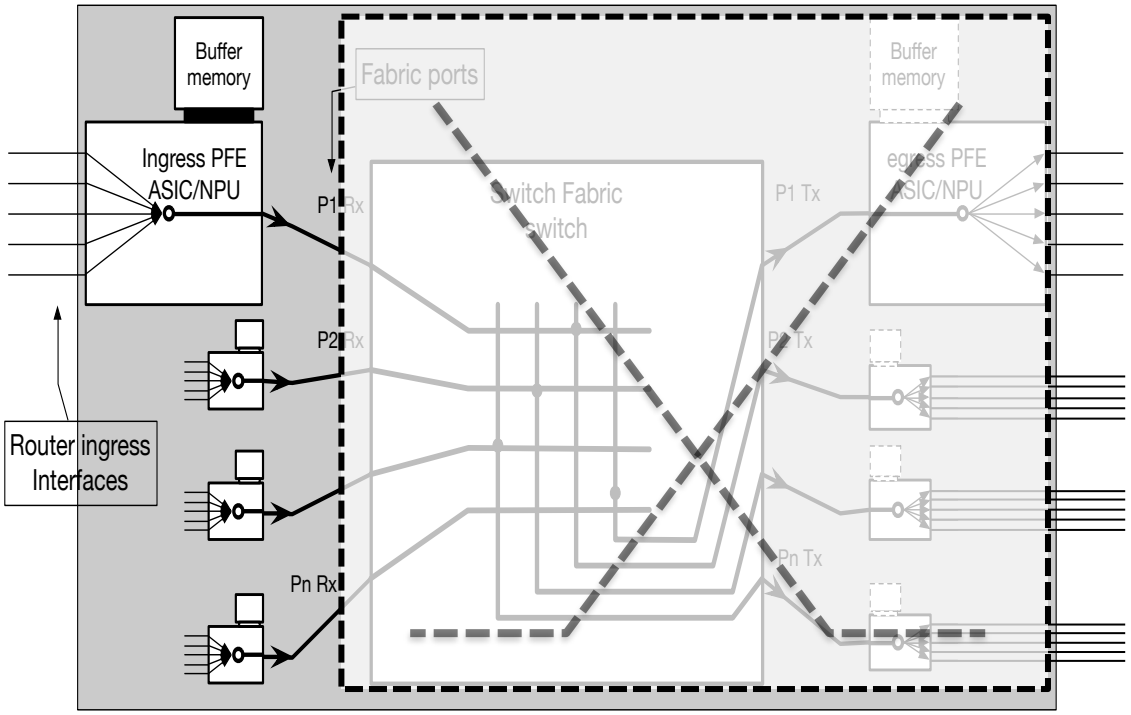


- Which ingress PFE get GRANT next. (e.g. fair-share) from given egress PFE
- Monitor Fabric Egress and stop giving GR from queue on Fabric ingress that suppose to egress fabric via congested port
- Packet received from Fabric are
  - **stored** in egress port output queue.
  - **Or dropped** if queue is full.

Fabric scheduler

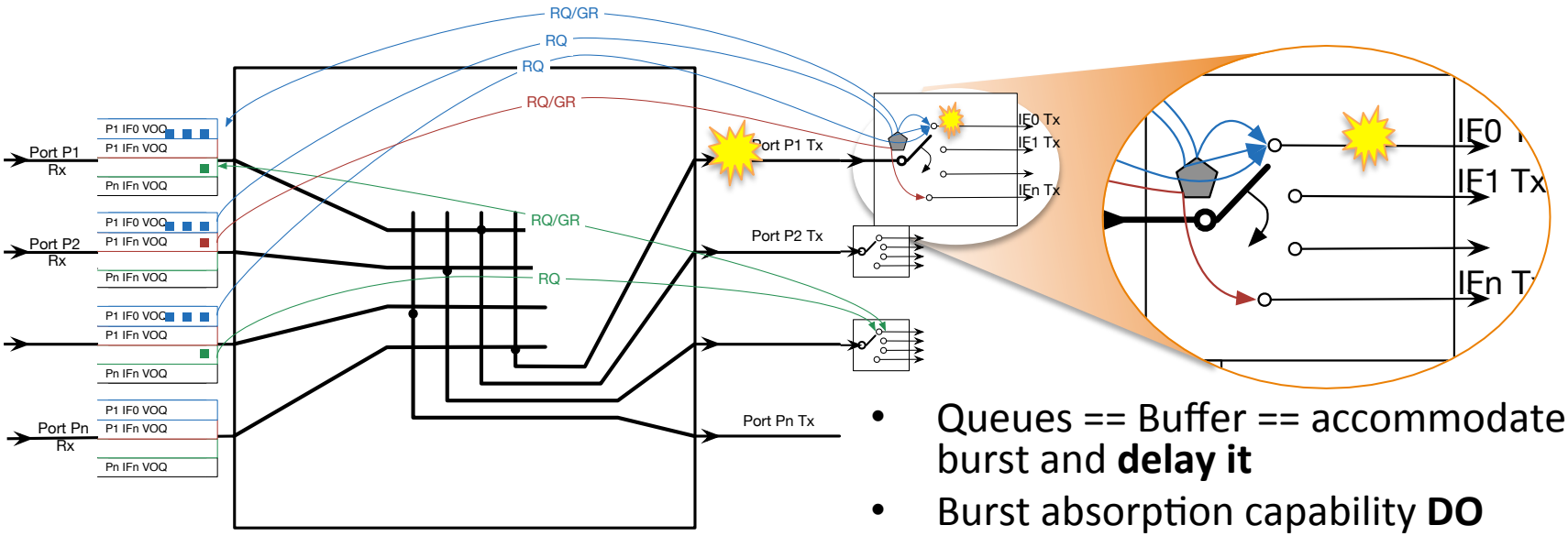
- Fabric queuing and flow-control independent form egress port queuing and scheduling.
- Router behavior – residency time (latency), jitter, drop rate depends on both.

End-to-End VOQ system



- For queuing purpose, switch fabric and all de-mux – seen as single switch
  - N inputs (Rx) → M output (Tx)
  - N x Rx speed == M x Tx speed
  - M >> N;

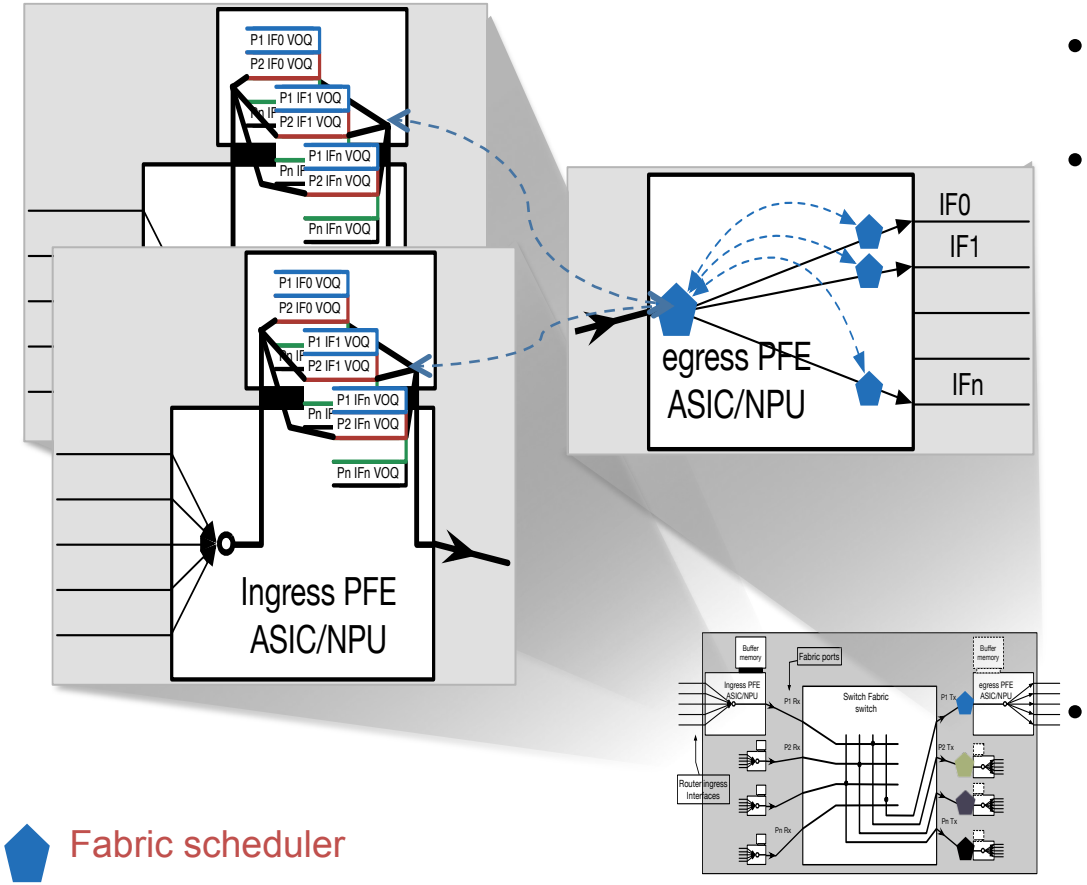
Buffer once - latency



- Queues == Buffer == accommodate burst and **delay it**
- Burst absorption capability **DO NOT** depends on type of burst (regardless at which point congestion appears)
  - Max: VOQ size
  - Min: VOQ size
- Max latency: Σ (VOQ size)
- Wait, there will be example

Potential congestion point

Buffer once – bandwidth Fabric Scheduler and flow-control



- Fabric Queuing - VOQ (per egress interface)
- Fabric scheduler and Flow-Control
  - Which ingress [PFE, VOQ] get GRANT next. (e.g. fair-share) from given egress PFE
  - Monitor egress interface and stop giving GR from queue on Fabric ingress if egress interface is not free
- Packet received from Fabric is immediately send out by egress interface

- Fabric queuing and flow-control depends on egress interface only.

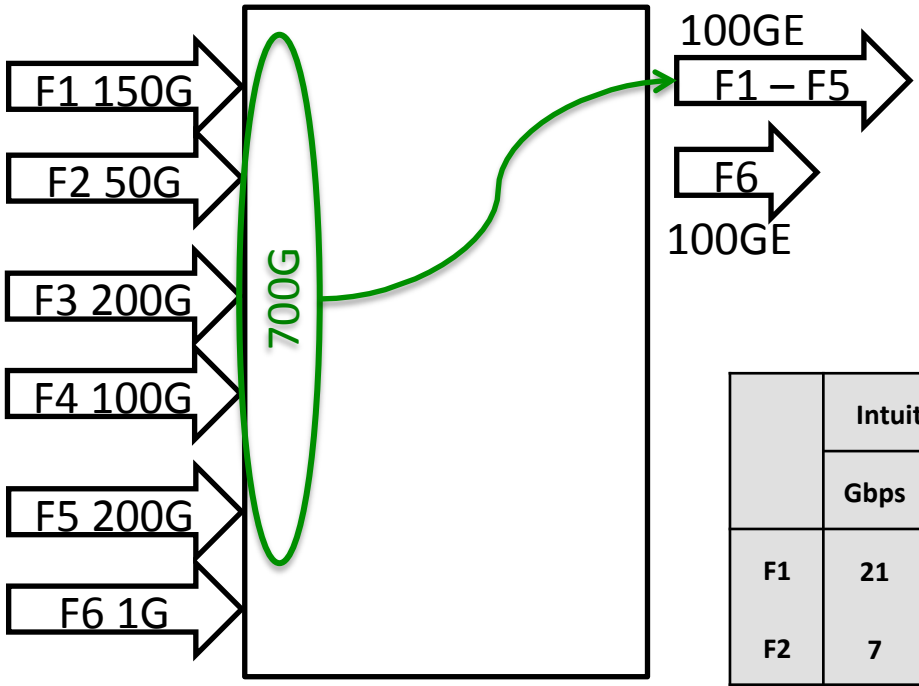
System characteristic vs. Queuing architecture

	CIOQ	CIOQ (w/ fabric VOQ)	E2E VOQ
Low residency time	✖	✖	✔
High load	✖	✔	✔
Low power footprint	✖	✖	✔
High number of interfaces (each with independent queuing. E.g. BNG, BE)	✔	✔	✖
multi-chassis systems	✔	✔	✔
Examples*	C7500, Early C7600	Juniper MX, Cisco ASR9k, CRS-X ALU 7750/7950*	Juniper PTX Cisco NCS600

* Please contact me if you want to update, correct add more – rafal@juniper.net

UNINTUITIVE BEHAVIOR - BANDWIDTH

Intuitive drop behavior

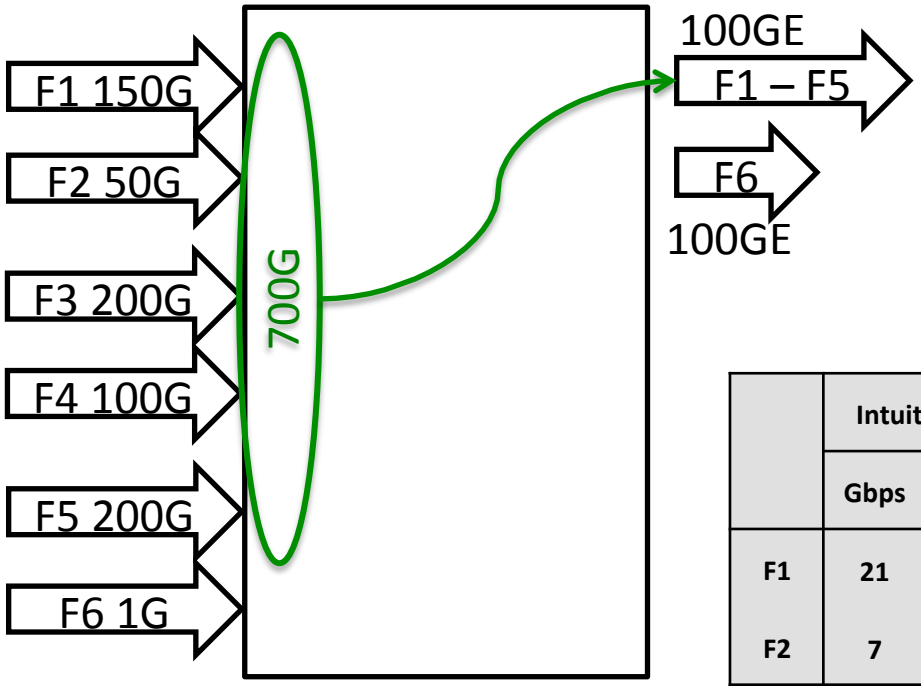


600/700 ➔ 86% loss expected

- All traffic is BE

	Intuitive (loss)?		Observed 2 (router X)		Observed 2 (router Y)	
	Gbps	Loss %	Gbps	Loss %	Gbps	Loss %
F1	21	86%				
F2	7	86%				
F3	29	86%				
F4	14	86%				
F5	29	86%				
F6	1	0%				

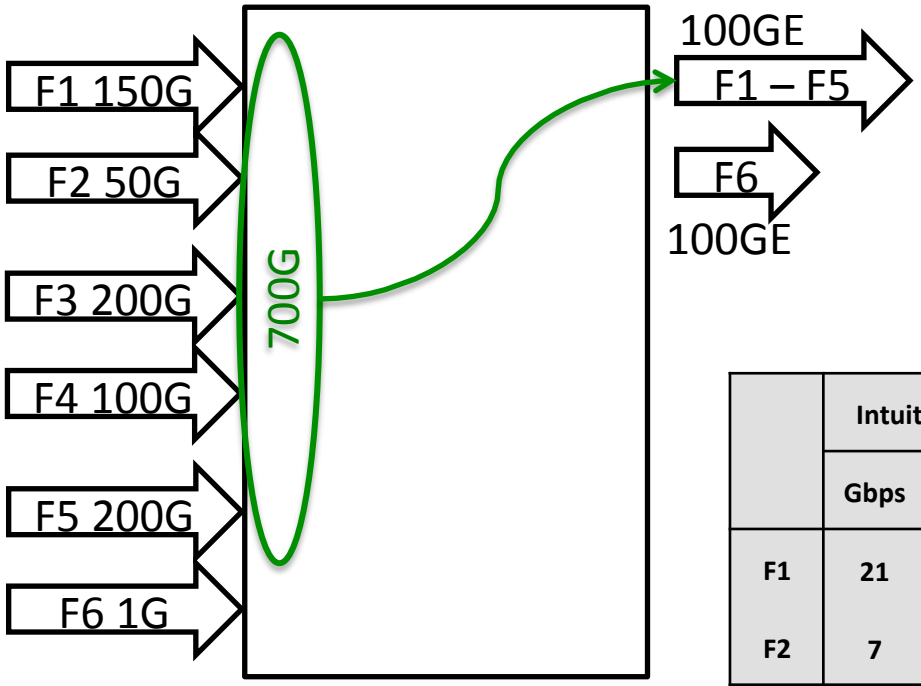
Un-intuitive drop behavior of router "X"



- All traffic is BE
- Why losses in F6?
- Why unequal losses in F1-F6

	Intuitive (loss)?		Observed 2 (router X)			
	Gbps	Loss %	Gbps	Loss %	Gbps	Loss %
F1	21	86%	25	82%		
F2	7	86%	8	82%		
F3	29	86%	22	88%		
F4	14	86%	11	88%		
F5	29	86%	33	83%		
F6	1	0%	0	52%		

Un-intuitive drop behavior of router "Y"

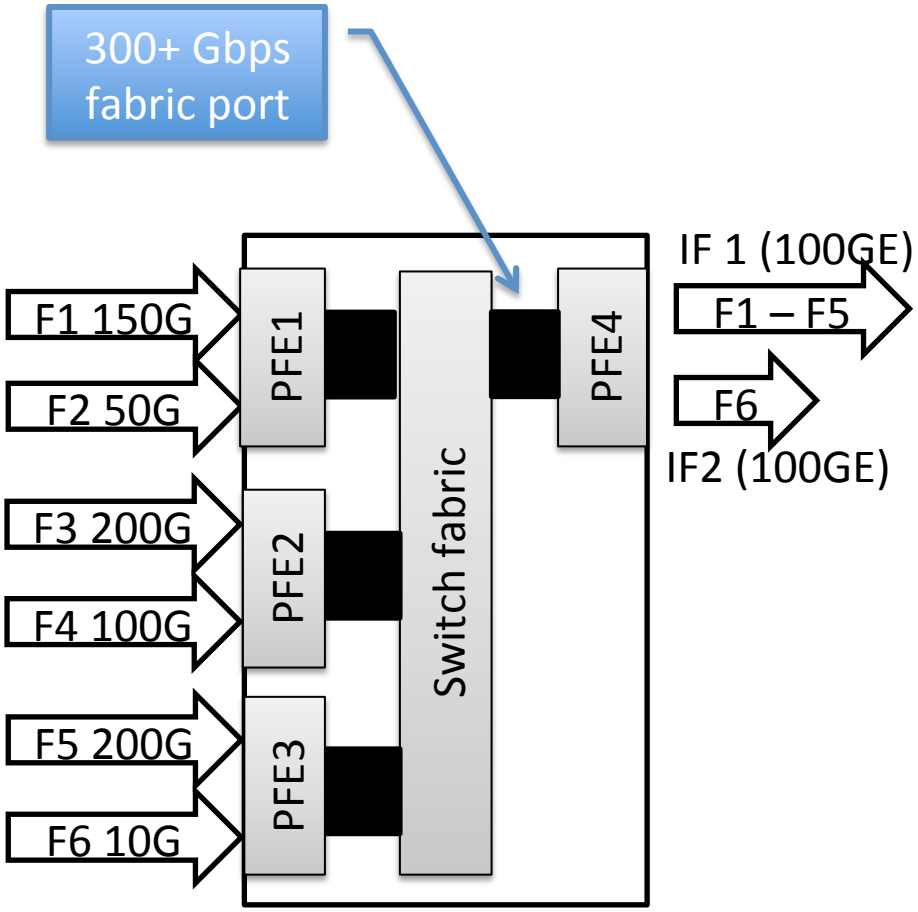


- All traffic is BE
- Why losses in F6?
- Why unequal losses in F1-F6
- No losses in F6 on router Y

	Intuitive (loss)?		Observed 2 (router X)		Observed 2 (router Y)	
	Gbps	Loss %	Gbps	Loss %	Gbps	Loss %
F1	21	86%	25	82%	26	83%
F2	7	86%	8	82%	9	83%
F3	29	86%	22	88%	22	89%
F4	14	86%	11	88%	11	89%
F5	29	86%	33	83%	33	84%
F6	1	0%	0	52%	1	0%

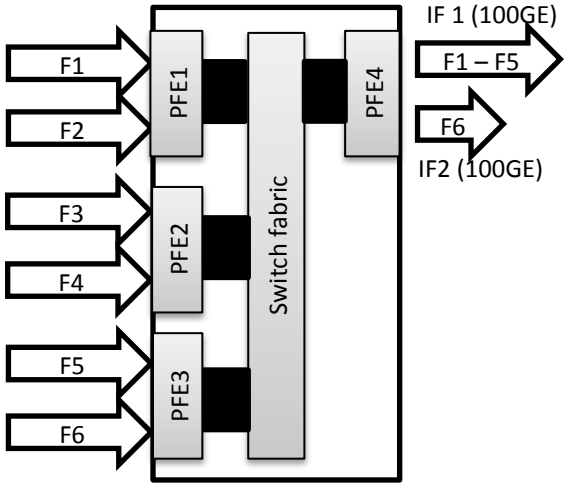
Know your hardware

- Router X -> CIOQ
- Router Y -> VOQ
- Fabric port -> 300Gbps
- Fair-share fabric scheduler



Router X – CIOQ - behavior

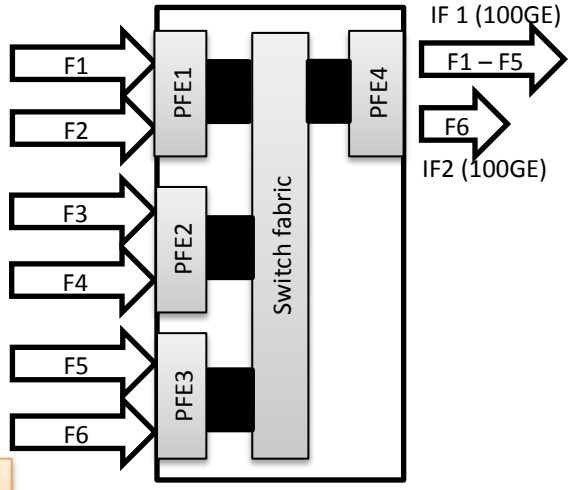
- Flow F1&F2 shares same buffer (queue) on PFE1. Same for F3&F4 @ PFE2 and for F5&F6 @ PFE3
- There is 3 ingress PFE that want to talk to PFE4
  - Fabric Scheduling gives 100Gbps to each ingress PFE



	offered	From fabric @ PFE4		On egress interface	
	Gbps	Gbps	Fabric Loss % (F_loss)	Gbps	Cumulative Loss %
F1	150	75 } 100G	50%		
F2	50	25	50%		
F3	200	67 } 100G	67%		
F4	100	33	67%		
F5	200	95.2 } 100G	52%		
F6	10	4.8	52%		

Router X – CIOQ - behavior

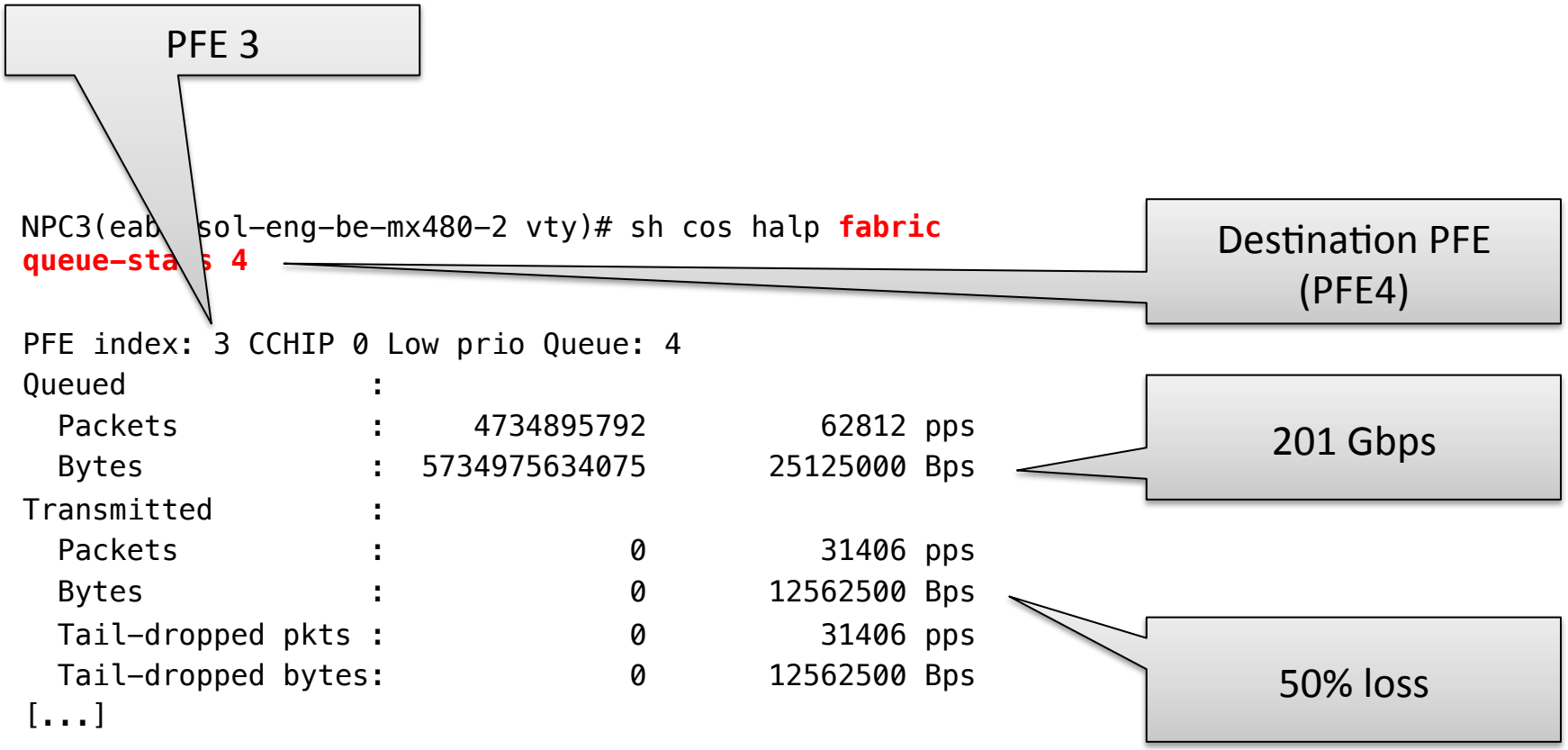
- Flows F1 – F5 (295Gbps) are queued in OQ of IF1, and Tx @ 100Gbps. (**65% loss – Egress interface loss; E-loss**)
- Flows F6 (10Gbps) are queued in OQ of IF2, and tx @ 100Gbps. (0% loss)



Cumulative loss for F1-F5: F_loss + (1-F_loss)*E_loss

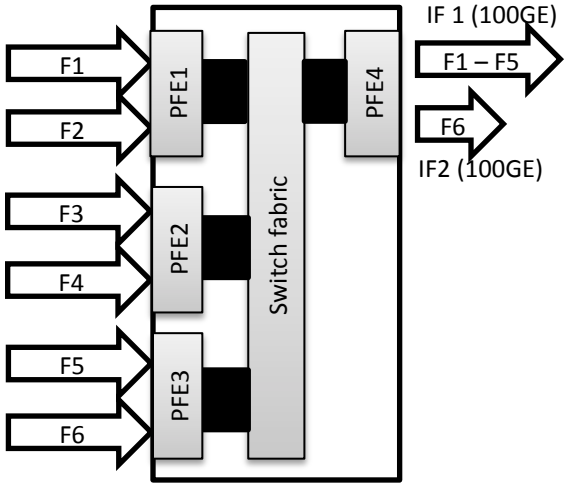
	offered	From fabric @ PFE4		On egress interface	
	Gbps	Gbps	Fabric Loss % (F_loss)	Gbps	Cumulative Loss %
F1	150	75 } 100G	50%	25	82%
F2	50	25	50%	8	82%
F3	200	67 } 100G	67%	22	88%
F4	100	33	67%	11	88%
F5	200	95.2 } 100G	52%	33	83%
F6	10	4.8	52%	4.8	52%

CLI example


PFE 3
NPC3(eab sol-eng-be-mx480-2 vty)# sh cos halp fabric
queue-sta s 4
PFE index: 3 CCHIP 0 Low prio Queue: 4
Queued :
Packets : 4734895792 62812 pps
Bytes : 5734975634075 25125000 Bps
Transmitted :
Packets : 0 31406 pps
Bytes : 0 12562500 Bps
Tail-dropped pkts : 0 31406 pps
Tail-dropped bytes: 0 12562500 Bps
[...]
Destination PFE (PFE4)
201 Gbps
50% loss

Router Y – VOQ - behavior

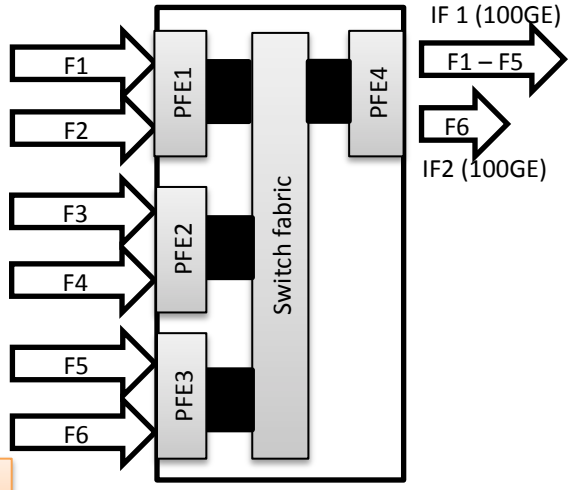
- Flow F1&F2 shares same buffer (queue) on PFE1. Same for F3&F4 @ PFE2 and for F5 @ PFE3.
- **Flow F6 has separate buffer on PFE 3.**
- There is 3 ingress PFE that want to talk to egress interface IF1 (100GE)
  - Fabric Scheduling gives **33Gbps to each ingress PFE for Egress Interface IF1 VOQ**



	offered	From fabric @ PFE4		On egress interface	
	Gbps	Gbps	Fabric Loss % (F_loss)	Gbps	Cumulative Loss %
F1	150	25 } 33G	83%		
F2	50	8	83%		
F3	200	22 } 33G	89%		
F4	100	11	89%		
F5	200	33 } 33G	83%		
F6	10	10 } 100G	0%		

Router Y – VOQ - behavior

- There is 1 ingress PFE that want to talk to egress interface IF2 (100GE)
  - Fabric Scheduling gives 100Gbps to only one ingress PFE (PFE3) for Egress Interface IF2 VOQ
- F6 do not consume it's share in full. Only 10Gbps.



Cumulative loss for F1-F5: F_loss

	offered	From fabric @ PFE4		On egress interface	
	Gbps	Gbps	Fabric Loss % (F_loss)	Gbps	Cumulative Loss %
F1	150	25 } 33G	83%	25	83%
F2	50	8 }	83%	8	83%
F3	200	22 } 33G	89%	22	89%
F4	100	11 }	89%	11	89%
F5	200	33 } 33G	83%	33	83%
F6	10	10 } 100G	0%	10	0%

CLI example

Ingress PFE 3
(200Gbps toward egress interface)

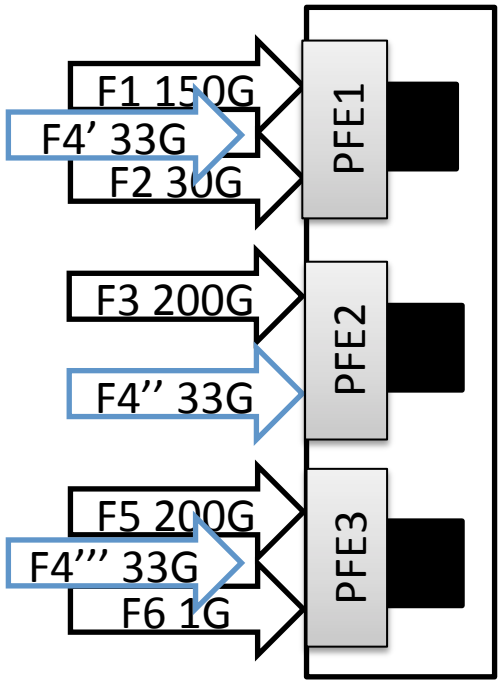
VOQ of egress IF
(100GE)

```
SNGFPC1(Thorax-re0 vty)# debug cos halp qlen tq 3 voq 2048
<snip>

 VOQ|   AQID|   qlen|  qlenold|    tabw|   ntabw|
maxrate(Mbps)| DBB Time(us)|
=============================================================
==================
   560|  29771|   2752|     2304|      64|      14|
1613|       111813|
=============================================================
==================

Number of Samples Collected = 1
 Parm |    Min|   Avg|     Max|
==============================
 qlen |   2752|   2752|   2752|
 tabw |     64|     64|     64|
ntabw |     14|     14|     14|
qdrain|  33000|  33000|  33000|
FreePg| 259387| 259387| 259387|
  UM  |      0|      0|      0|
```

33Gbps drain-rate

Accept your router personality

- Behavior is clear now ☺ - caused by Fabric scheduler
  - Non of vendor (AFAIK) allows for Fabric scheduler tuning.
  - Have to live with it as it is.

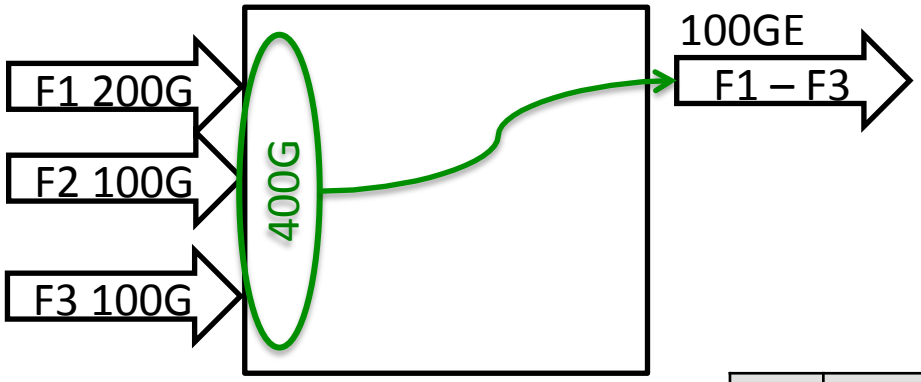


Mitigation: Load PFE fairly

- you get fair results among all flows
- Other goodies behind – blast radius

ROUTER BEHAVIOR CASE STUDY - LATENCY

Intuitive latency

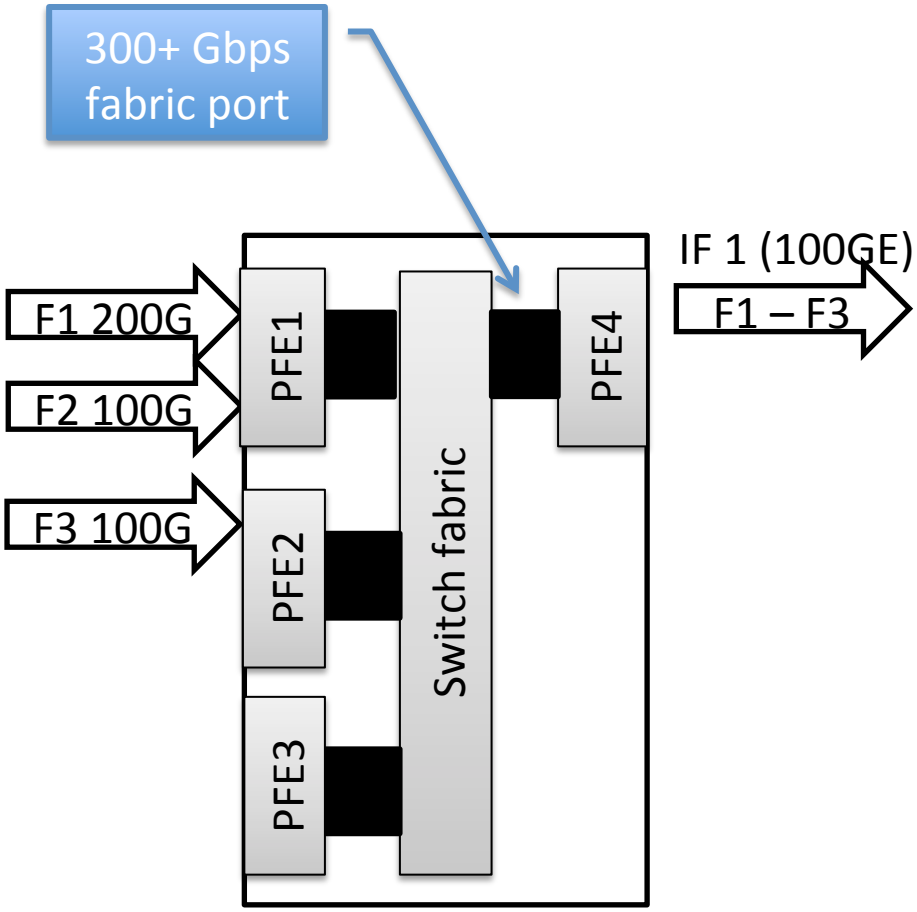


400Gbps --> 100Gbps
Buffer 100% -> 100ms latency
And 75% losses

	Expected (loss)?	Observed 1 (router X)	Observed 2 (router Y)
	ms	ms	ms
F1	100	200	100
F2	100	200	100
F3	100	100	100

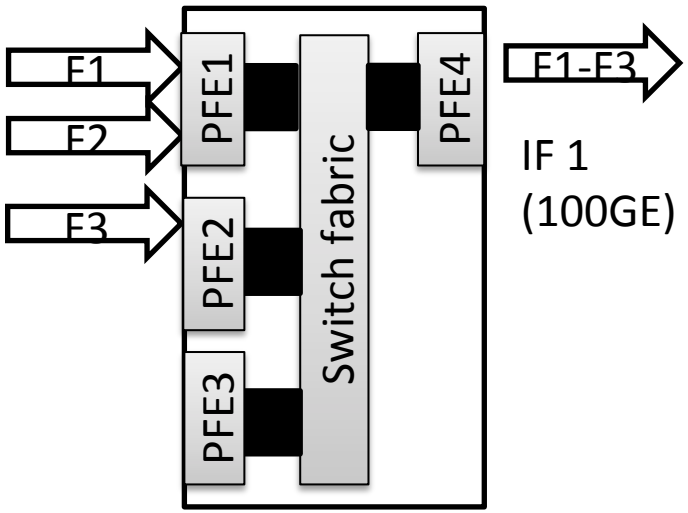
Know your hardware

- Router X -> CIOQ
  - 100ms VoQ before fabric
  - 100ms OQ
- Router Y -> VOQ
  - 100ms VoQ before fabric
- Fabric port -> 300Gbps



Router X - CIOQ

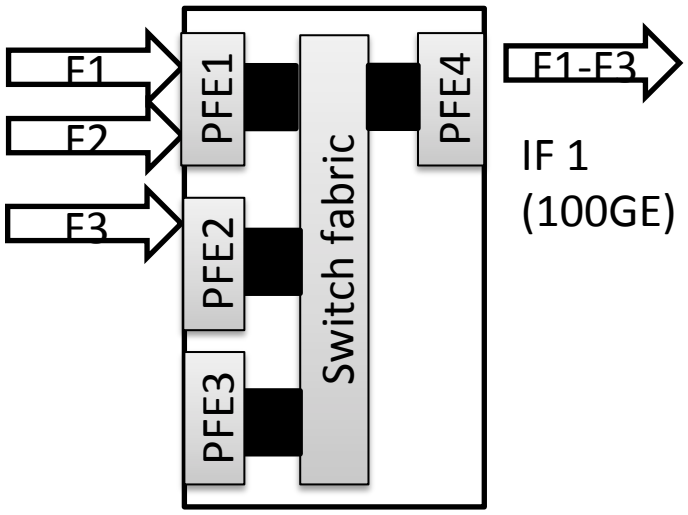
- Flow F1&F2 shares same VoQ buffer (queue) on PFE1. Flow F3 is alone on PFE2
- There is 2 ingress PFE that want to talk to PFE4
  - Fabric Scheduling guarantee 150Gbps to each ingress PFE
- Flows F1-F2 (300Gbps) are queued in VOQ of PFE1, and Tx @ 200Gbps (150Gbps + leftover).
  Latency is 100ms



	offered	Fabric component	Egress Interface	Total
	Gbps	ms		
F1	200	100 (33% loss)		
F2	100	100 (33% loss)		
F3	100			

Router X - CIOQ

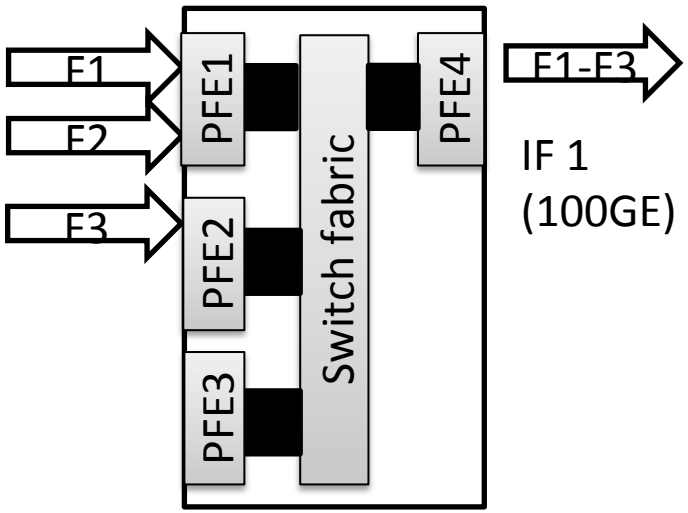
- Flows F3 (100Gbps) are queued in VOQ of PFE1, and Tx @ 100Gbps → no buffering
- Flows F1-F3 (300Gbps) are queued in OQ of IF1, and Tx @ 100Gbps.
  Egress interface latency is 100ms



	offered	Fabric component	Egress Interface	Total
	Gbps	ms	Ms	ms
F1	200	100 (33% loss)	100 (66% loss)	200 (77% loss)
F2	100	100 (33% loss)	100 (66% loss)	200 (77% loss)
F3	100	0 (0% loss)	100 (66% loss)	100 (66%)

Router Y - VOQ

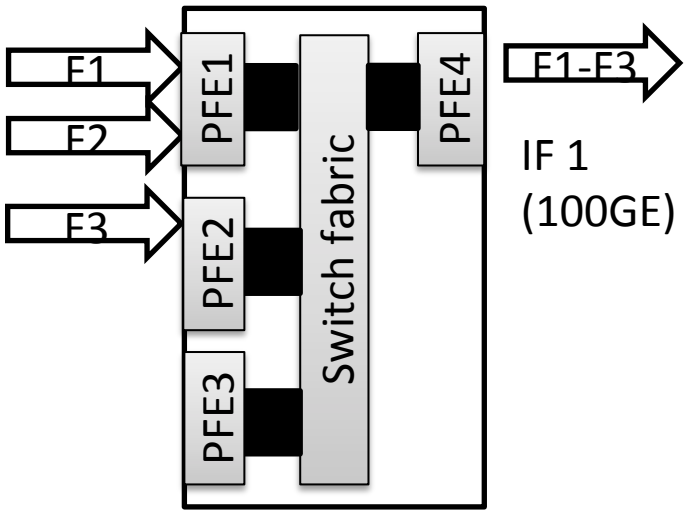
- Flow F1&F2 shares same VoQ buffer (queue) on PFE1. Flow F3 is alone on PFE2
- There is 2 ingress PFE that want to talk to PFE4
  - Fabric Scheduling guarantee 50Gbps to each ingress PFE



	offered	Fabric component	Egress Interface	Total
	Gbps	ms	ms	sm
F1	200	100		
F2	100	100		
F3	100	100		

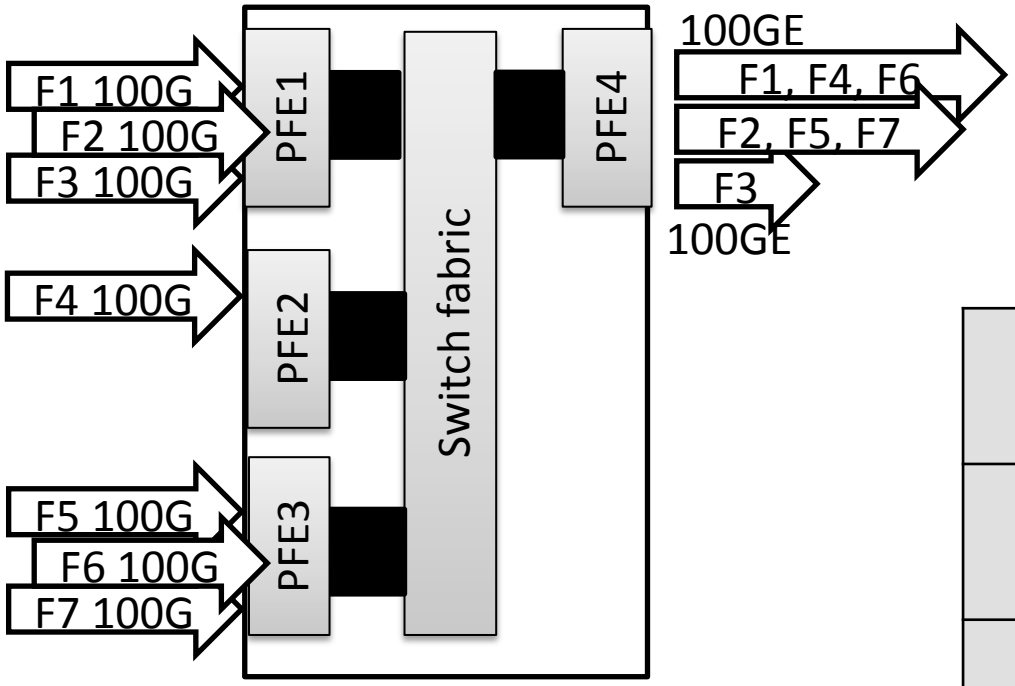
Router Y - VOQ

- Flows F1-F2 (300Gbps) are queued in VOQ of IF1, and Tx @ 50Gbps.
  Latency is 100ms
- Flows F3 (100Gbps) are queued in VOQ of IF1, and Tx @ 50Gbps
  Latency is 100ms



	offered	Fabric component	Egress Interface	Total
	Gbps	ms	ms	sm
F1	200	100 (83% loss)	0	100 (83% loss)
F2	100	100 (83% loss)	0	100 (83% loss)
F3	100	100 (50% loss)	0	100 (50% loss)

Other surprising behavior – router W



- All traffic is BE

	Observed on egress	
	Gbps	Loss %
F1	33	67%
F2	50	50%
F3	100	0%
F4	33	67%
F5	25	75%
F6	33	67%
F7	25	75%

Homework

- What Queuing architecture router W is?
- Explain behavior.

- Answers: rafal@juniper.net
  - Deadline – 11pm today.

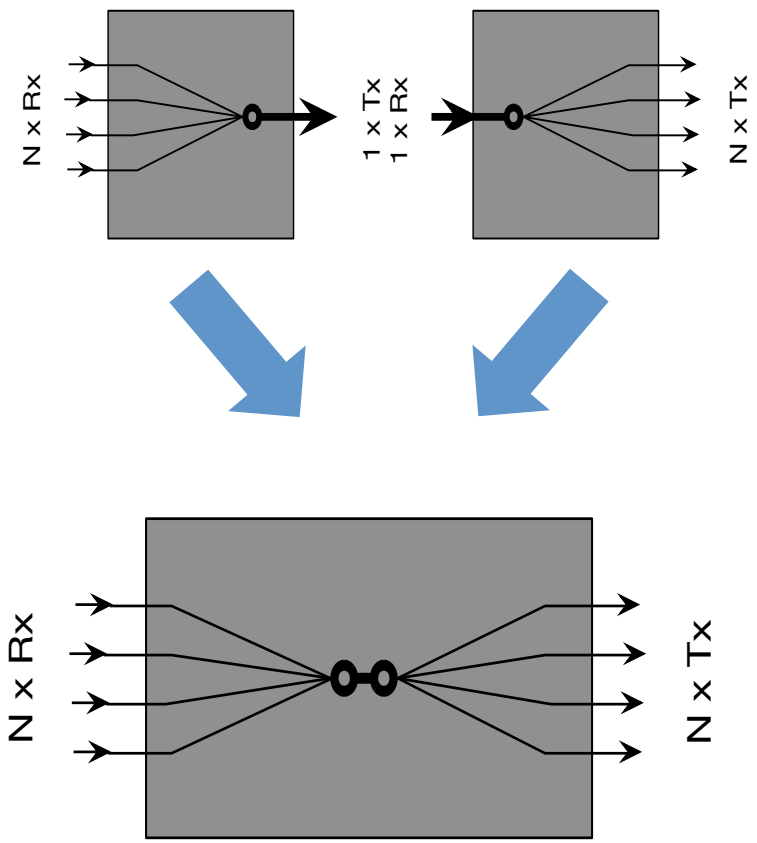
Summary

- Know your router anatomy
- System queuing architecture impacts power consumption and system scaling capabilities.
- System queuing architecture impact residency-time
- System queuing architecture may be a reasoned for non-intuitive traffic loss pattern.
  - (Re)Assign ports to roles smart way.
  - Trying to solve of non-existing problem cost time and headache of writing a incident report – avoid it.

Tank you!

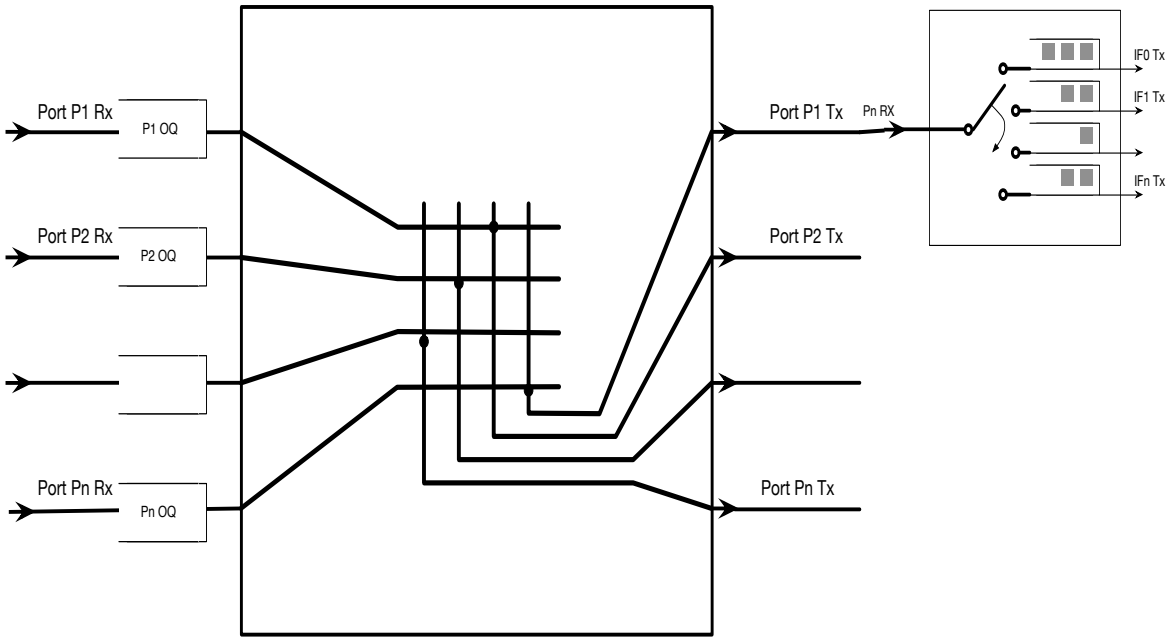
BACKUP SLIDES.

From building blocks to centralized router



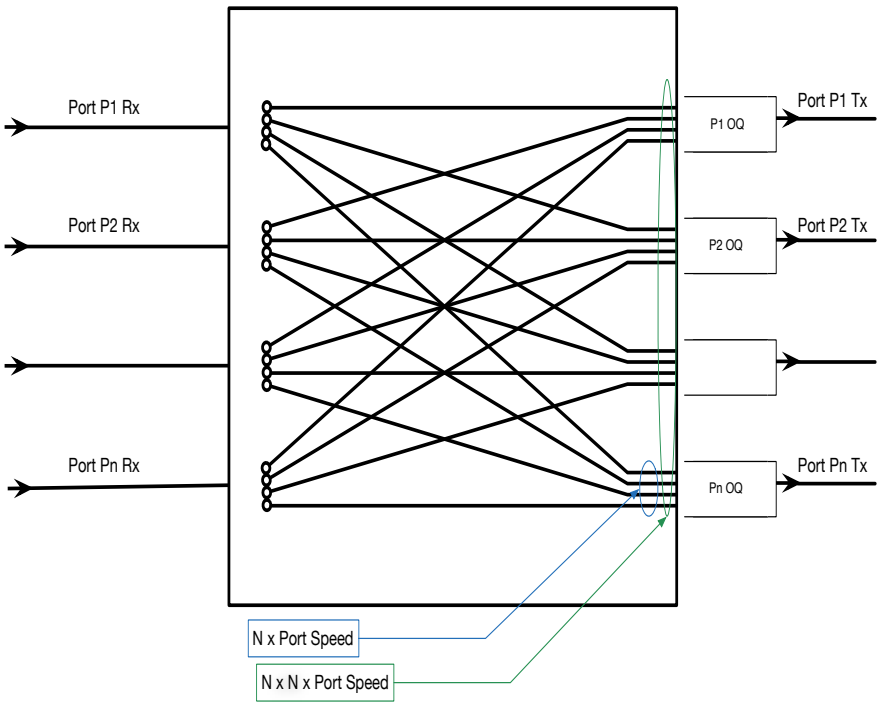
- Single switching element
- No Switch Fabric
- N x N Interfaces
- Interfaces may have different speeds
- Memory used to build egress interface queues
- Different memory options
  - On-chip shared by mux and de-mux
    - Very fast (SRAM) – ~10Tbps+
    - Small and costly (10's MB)
  - Off-chip shared by mux and de-mux
    - Deep queues/buffers (GB)
    - Slower (DRAM) – ~1Tbps
  - Off-chip memory limits PFE performance.

Combined Input Output Queuing



IQ requairments for loss-less:
Fabric Port speed >> Σ (ASIC egress interfaces)

OQ

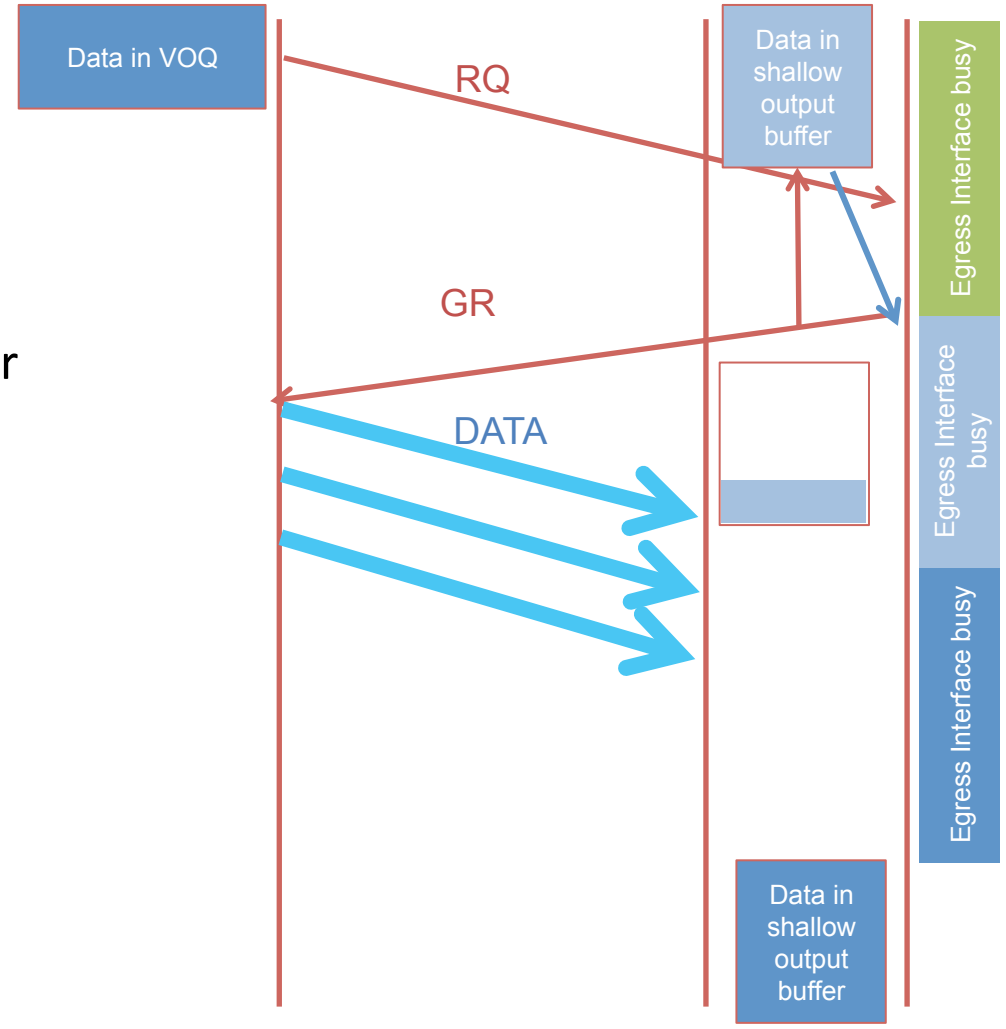


- Fabric switch is not a really switch here.
- Need very high speed (N x) fan-in to buffer.
  - If switch port is 600Gbps, and we have 100 ports →
  - 60T of raw bandwidth into buffers !
- 100% efficiency

- Good on paper only
  - extremely expensive
  - Bound by technology

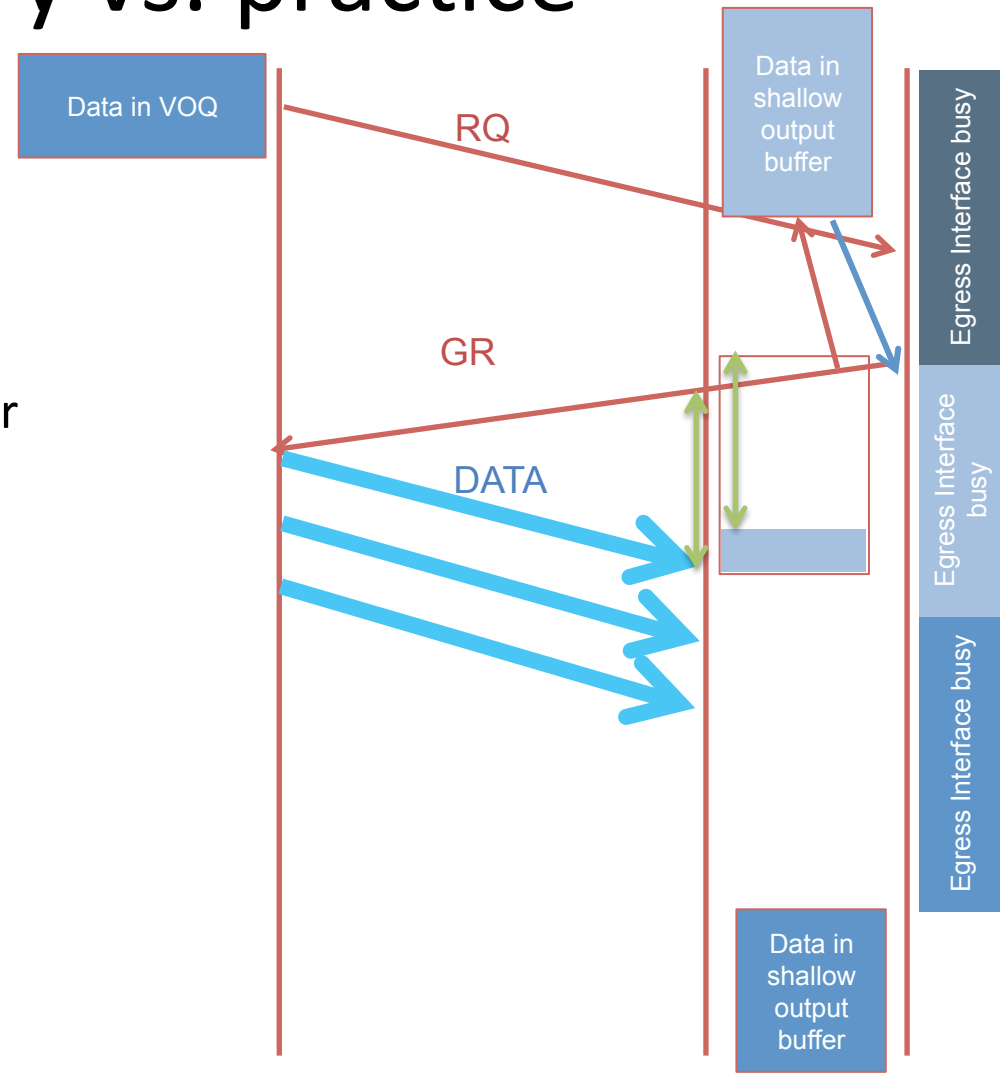
Buffer once – per egress interface VOQ Theory vs. practice

- In stat-mux system it is unpredictable when egress interface becomes free.
- Grant signaling delay affect deficiency
- Need for shallow buffer after fabric (egress PFE) to compensate delay.
  - Need just ~10 usec.
- Similar to CIOQ but:
  - Fabric flow-control ensure that OQ never overflow.
  - Try to keep IQ always full, never empty (if data are in VOQ)
  - Egress interface free/busy indirectly controls fabric flow-control



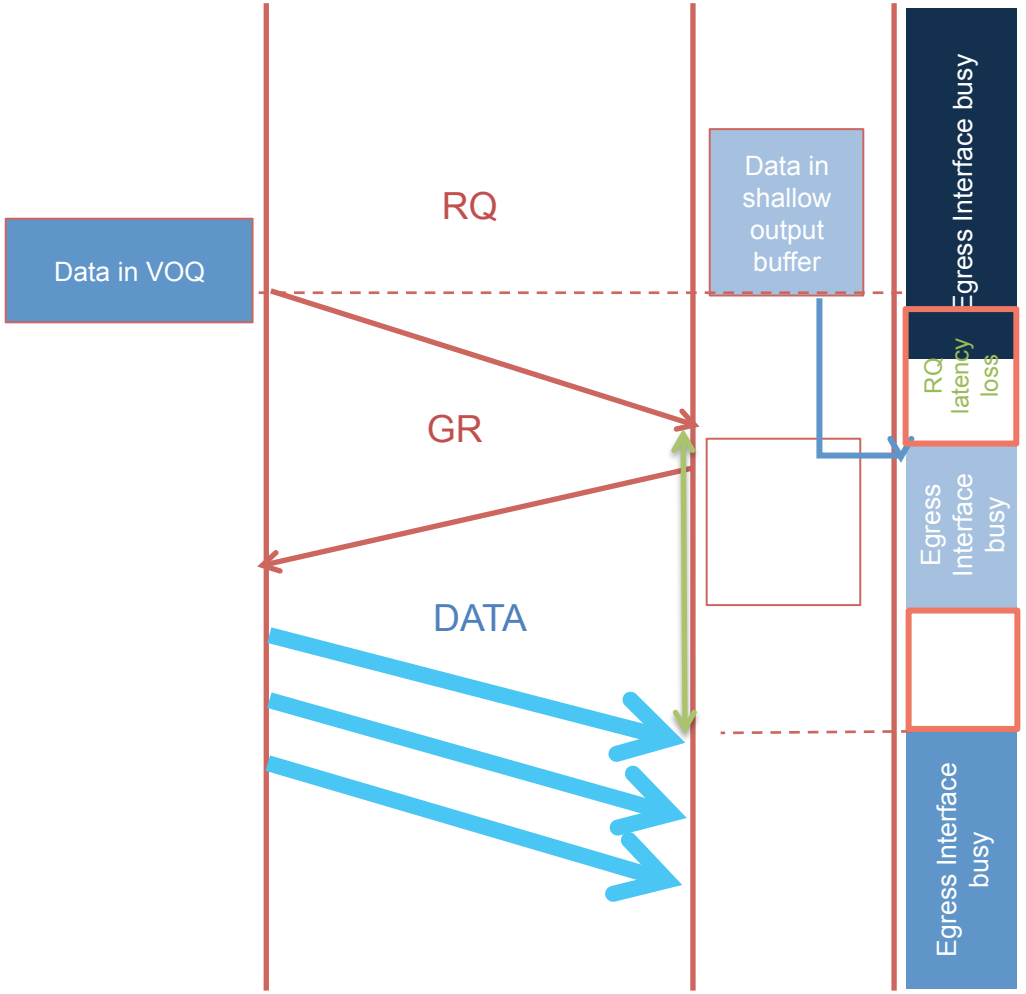
Buffer once – per egress interface VOQ Theory vs. practice

- In stat-mux system it is unpredictable when egress interface becomes free.
- Grant signaling delay affect deficiency
- Need for shallow buffer after fabric (egress PFE) to compensate delay.
  - Need just ~10 usec.
- Similar to CIOQ but:
  - Fabric flow-control ensure that OQ never overflow.
  - Try to keep IQ always full, never empty (if data are in VOQ)
  - Egress interface free/busy indirectly controls fabric flow-control



Too shallow shallow output buffer

- Latency
  - 3 x fabric one-way latency is worst case (RQ-GR-DATA)
  - RQ scheduler on egress
  - GR scheduler on ingress
- RQ latency
  - Can't be compensated
  - Statistically minor problem – asynchronous. RQ could be send while egress interface handles other data
- RQ/GR scheduler - Can't be compensated
- If shallow buffer < 2x latency – inefficient egress IF utilization

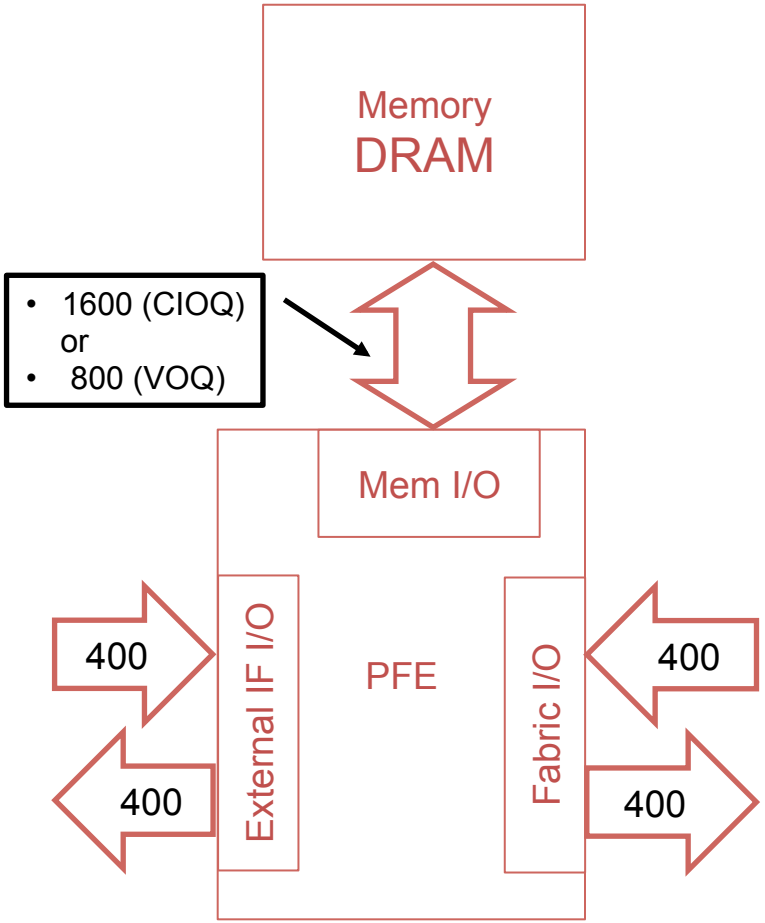


Impact on PFE design

PFE complexity

- PFE for per-egress IF VOQ system != PFE for CIOQ system
- For system of 4k IF and 8 QoS Classes
  - **VOQ** PFE need support 32.000 queues
  - **CIOQ** PFE need 1.040 queues
- If PFE support 400k queues
  - **VOQ** system can support 50k IF
  - **CIOQ** system can support 1.000.000's IF (from queue scaling perspective only. Other limits apply)
- CIOQ PFE has typically less queues but much more of other functionalities.

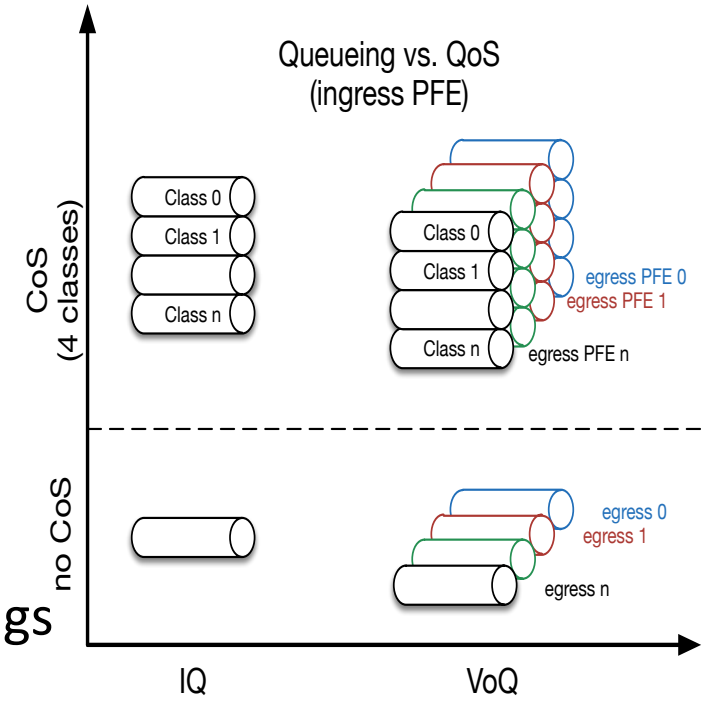
PFE performance



- To handle 4 x 100GE interfaces, PFE need:
  - **CIOQ:**
    - **more then 3.6Tbps** of PFE I/O.
    - All packet goes to and from memory twice – **4 memory accesses**.
  - **VOQ**
    - **more then 2.4Tbps** (25% less)of PFE.
    - All packet goes to and from memory once – **2 memory accesses**.
  - Memory I/O BW need to be oversized - even better reduction (~30%)
  - Each memory access causes latency
- Less I/O == saved gates
  - Have more VOQ (bigger system), OR
  - lower cost and power, OR
  - higher performance, OR

CoS vs. Queuing

- Queuing as discussed so fare
  - manage congestion/overload
  - Assumes all traffic is same class
- CoS
  - Traffic has different classes
  - Each class need other treatment
- CoS + Queueing = QoS
  - Scheduler(s) take into account 2 things
    - Class of traffic
    - Source instance (e.g. ingress PFE)
  - What was 1 Queue becomes set of parallel queues.
  - Classifier needed -> put traffic to this or other queue (out of scope)

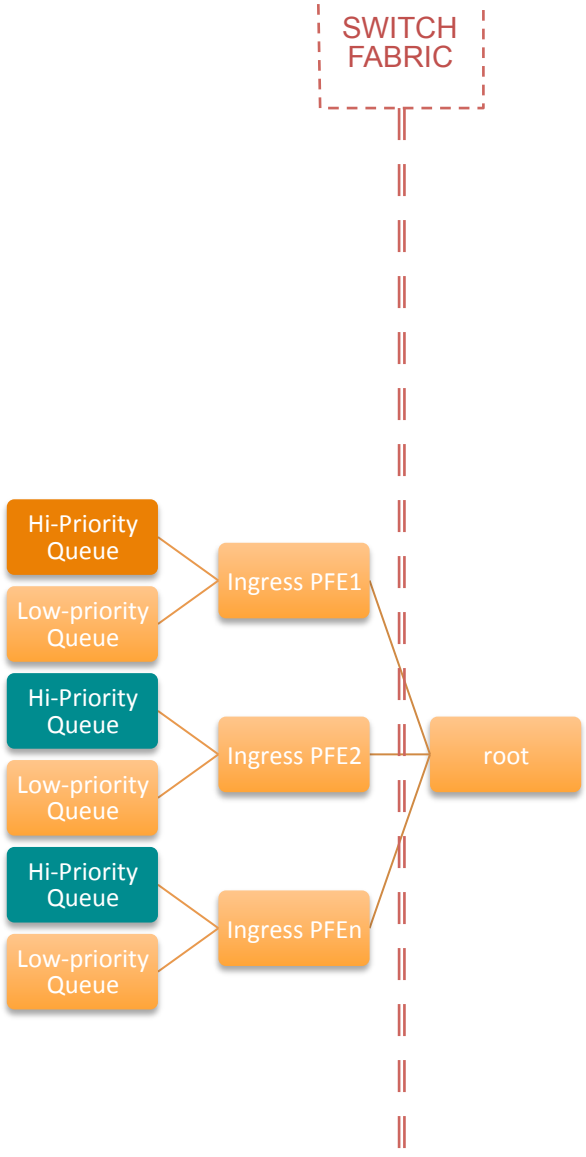


PFE complexity

- PFE for per-egress IF VOQ system != PFE for CIOQ system
- For system of 4k IF and 8 QoS Classes
  - VOQ PFE need support 32.000 queues
  - CIOQ PFE need 1.040 queues
- If PFE support 400k queues
  - VOQ system can support 50k IF
  - CIOQ system can support 1.000.000's IF (from queue scaling perspective only. Other limits apply)
- CIOQ PFE has typically less queues but much more of other functionalities.

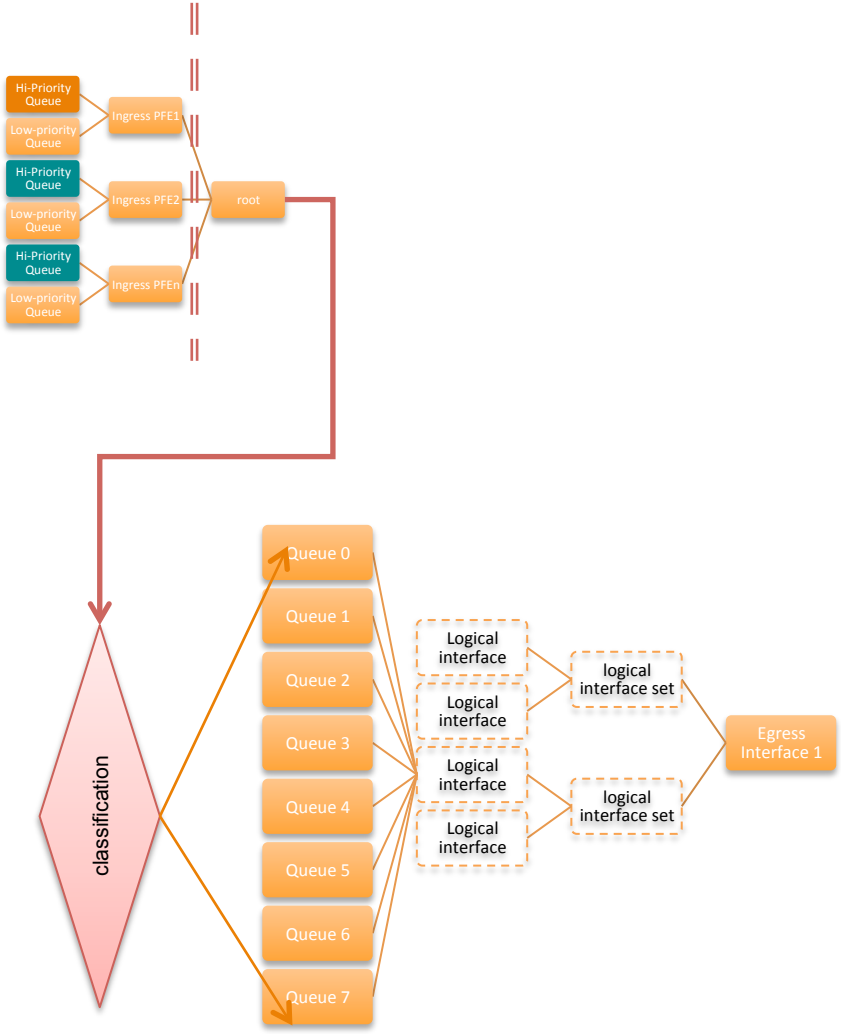
CIOQ- Fabric

- CIOQ
- Fabric VOQ
  - Flow-Controll
    - per Fabric Egress
    - By Request – Grant protocols.
    - Do not depends on Egress interfaces and OQ state.
  - 2 Classes Hi-/Low- priority
    - classification
  - Scheduling
    - ingress PFE fairness 1st
    - CIR-bound Priority scheduling.
  - Delay Bandwidth Buffer:
- Scheduler logic seats on egress PFE; Memory is on ingress side of fabric.



CIOQ – Egress Interface

- CIOQ
- Egress (logical) Interface Queues
  - 8 Queues per (logical) Interface
  - 10M queues per system
  - Scheduling
    - 5 priorities level, 4 RED profiles per queue. Configurable.
    - 2 or 4 scheduling hierarchy level w/ priority propagation
  - Delay Bandwidth Buffer: _____
- Scheduler logic and Memory seats on egress PFE



VOQ – integrated scheduler

- Flow-Control
  - per egress (logical) interface
  - By Request – Grant protocols.
  - Depends on Egress interfaces OQs (Output Queues) state.
- Queues
  - 8 per egress (logical) interface
  - 390.000 per system (HW limit)
- Scheduling
  - 4 priorities level, 4 RED profiles per queue. Configurable.
  - 2 or 3 scheduling hierarchy level w/ priority propagation
- Delay Bandwidth Buffer:
  - Shared memory
  - 100ms upper bound
  - 40ms worst case.
- Scheduler logic seats on egress PFE; Memory is on ingress side of fabric