Distributed Buffer Challenges

François Labonté flabonte@arista.com
Hugh Holbrook holbrook@arista.com

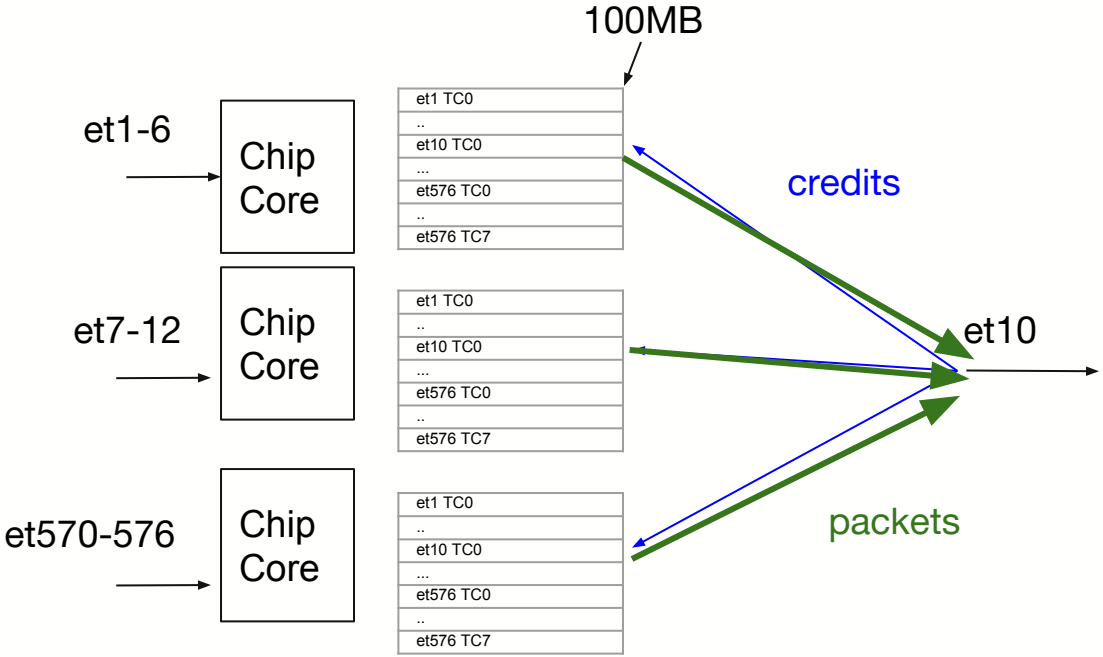
ARISTA

Arista Builds Products based on Merchant Silicon

Arista Products uses different chips that offer different feature/price performance points

- Packet processor table size
- Features (VXLAN, MPLS, etc)
- Buffering is one of those diffentiator
  - Most chips have on-chip buffers on the order of 10s of MB
  - One chip family has external DRAM buffers in the order of GB

Arista 7500 and 7280 distributed buffer system

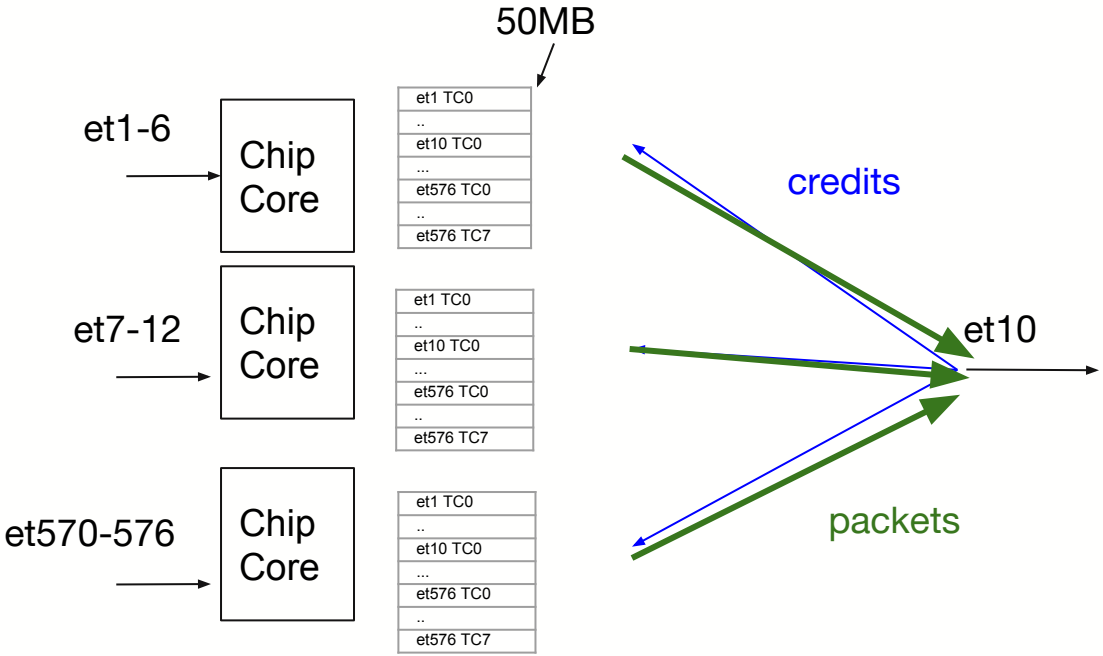


VOQs are on each ingress core for all ports and traffic classes in device

Packets are buffered on ingress and scheduled by the egress port

A device can have 2 to 192 cores

Configuration is not global



We can set tail drop policies/WRED/ECN per ingress core but no guarantees about total buffering for a given port and traffic class or a given port

Multiple Traffic Classes and egress shaping and scheduling make things even worse

Monitoring is also hard in a distributed system

We can get maximum queue sizes on each ingress core (watermark over time period)

Buffering burst can be very quick and fleeting

But we cannot get precisely the highest sum of the queues at any point in time

So we end up with measurements of the sum of the queues was at least as big as the biggest queue on any core and up to the max sum of the

Software bugs / Features / Defaults

We strive for software free of bugs, that said we do sometimes have bugs and some do affect buffer behavior.

We are working on more features to improve buffer configuration and monitoring. I wish they had been in when we first shipped… But hopefully we only get better

Defaults thresholds - about that 500MB… I feel stuck with archaic decisions that were arbitrary. I wished I had the insight of real use case measurements

Knobs/Monitoring that would help

- We currently only allow limiting tail drop per queue
- We could allow limiting the total amount of buffers to reproduce more realistically a device with smaller buffers when multiple queues are competing for resources
- We have a feature that will measure the latency of packets inside a queue. In a system with a single traffic class active per port this would give a great idea of the amount of buffering

We are excited to participate in meaningful experiments to improve the state of the art in understanding the impact of buffers on application performance.