Carousel: Scalable Traffic Shaping at End Hosts

Ahmed Saeed, Nandita Dukkipati, Vytautas Valancius, Vinh The Lam, Carlo Contavalli, Amin Vahdat
Motivation

Figures taken from: Achieving high utilization with software-driven WAN, Hong et al. ACN SIGCOMM 2
Typical rate limiter architecture

Socket Buffers in Host OS or Guest VM

Multi-queue Token Buckets rate limit flow aggregates

Classifier

Scheduler

To NIC
Commonly used Shapers

**HTB**

- Packets traverse the HTB tree. Internal Nodes handle traffic accounting, and leaf nodes shaping traffic.
- Source flows
- Token Bucket

**FQ/Pacing**

- Flow state lookup
- Hashable of red-black trees on flow ids
- Flow pacing
- Flow delays (red-black tree on timestamps)

**Drawbacks**

- Good rate conformance
- High CPU Overhead (Contention)
- High Memory Overhead (Buffering & Data structures)
Tenets

1. Single Queue Shaping

2. Deferred Completions (Backpressure)

3. Silos of one shaper per-core
Architecture

Socket Buffers in Host OS or Guest VM

Timestamp every packet based on policy and rate

Time-based queue ordered based on timestamp

To NIC

Report completion to source
Timestamps

- Determine packets release time
- Each policy sequentially applies a timestamp according to its rate
- Finally largest timestamp is chosen => smallest rate (no policy violation)
Single Time-indexed queue

- **Timing Wheel**: Array of Lists of time stamped packets
- Array represents buckets from *now* until *horizon* => Circular time representation
- Wheel *spins* & dequeues all packets in *now-bucket* with a period of \( g_{min} \)
Deferred Completions

- Provide backpressure by bounding # of enqueued packets per flow
- Completion signal is sent at dequeuing time instead of enqueuing time
Micro Benchmarks: Setup

- Experiments on egress traffic shaping
- Two servers sharing same top of rack switch
- Carousel implemented in software NIC
- HTB / FQ-pacing for comparison
Rate Conformance: Single connection

Figure 12: Comparison between Carousel, HTB, and FQ in their rate conformance for a single flow.
Rate Conformance: Multiple connections

Figure 13: Comparison between Carousel, HTB, and HTB/FQ showing the effect of the number of flows load on their rate conformance to a target rate of 5 Gbps.
Memory Efficiency: Changing Target Rate

Figure 14: A comparison of Deferred Completion (DC) and delay-based congestion control (TCP Vegas) backpressure, for fifteen flows when varying the aggregate rate limits.
Memory Efficiency: Vary # of Flows

Figure 15: A comparison of Deferred Completion (DC) and delay-based congestion control (TCP Vegas) backpressure, while varying the number of flows with an aggregate rate limit of 5Gbps.
Impact at Receiver (100:1 incast)

Figure 19: Comparing receiver side rate limiting with target rate.
Production Experience

• 25 servers in 5 geographically diverse locations (US & EU)

• Serving video traffic (YouTube)

• During peak hours
Production Experience: CPU

Figure 21: Comparison between FQ and Carousel for two machines serving large volumes of video traffic exhibiting similar CPU loads showing that Carousel can push more traffic for the same CPU utilization.
Production Experience: Software NIC

Figure 22: Software NIC efficiency comparison with pacing enforced in kernel vs pacing enforced in software NIC. Bandwidth here is normalized to a unit of software NIC self-reported utilization.
Paper:

- Well written
- Helpful diagrams
- Lots of evaluation (maybe too much?)