SENIC: Scalable NIC for End-Host Rate Limiting

Sivasankar Radhakrishnan

Yilong Geng, Vimalkumar Jeyakumar,
Abdul Kabbani, George Porter, Amin Vahdat

UCSD CSE
Computer Science and Engineering

STANFORD
Computer Science

Google

USENIX NSDI 2014
4 April 2014
Consolidation of Servers

Network resource management and allocation is crucial
Network Resource Allocation

- Performance isolation: Oktopus, Seawall, EyeQ
- Congestion control: QCN, RCP, D3, DCTCP, HULL

Rely on programmable rate limiters

Scalable rate limiting is required
Thousands of rate limiters per server
## Rate Limiter Options

<table>
<thead>
<tr>
<th>Feature</th>
<th>Software</th>
<th>Hardware</th>
<th>SENIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scales to many classes</td>
<td>✔️</td>
<td>✗</td>
<td>✔️</td>
</tr>
<tr>
<td>Works at high link speeds</td>
<td>✗</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Low CPU overhead</td>
<td>✗</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Accurate and precise</td>
<td>✗</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Supports hypervisor bypass</td>
<td>✗</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

Reorganize responsibilities of the NIC and operating system
Current NIC Design

1. OS notifies NIC about packet

2. Schedule DMA transfer from host memory to NIC Tx buffers

3. Rate limit NIC Tx ring buffers

4. Transmit packet

Qdisc queues

Typically 8-32 rings

Packet Scheduler
Current NIC Design

Qdisc queues

TX buffers

Typically 8-32 rings

Packet Scheduler

Host (DRAM)

NIC (SRAM)
Host DRAM is cheap and abundant

Current NIC Design

Host (DRAM)

NIC (SRAM)

Qdisc queues

Packet Scheduler

TX buffers

Typically 8-32 rings

Wire

Host DRAM is cheap and abundant
Current NIC Design

Qdisc queues

TX buffers

Typically 8-32 rings

Host (DRAM)

NIC (SRAM)

Packet Scheduler

Wire

Hardware is good at per-packet scheduling
Typically 8-32 rings

Qdisc queues

Host (DRAM)

NIC (SRAM)

Expensive and limited

Can we get rid of these large buffers?
SENIC Design

1. OS notifies NIC about packet

2. Rate limit per-class queues stored in host RAM

3. DMA packet from host memory to NIC

4. Transmit packet

FIFO queues (or ring buffers)
1. OS notifies NIC about packet

2. Rate limit per-class queues stored in host RAM

3. DMA packet from host memory to NIC

FIFO queues (or ring buffers)

Many Tx queues

Packet Scheduler

Late binding of packet transfers to NIC
SENIC Design

FIFO queues (or ring buffers)

\[ \text{Many Tx queues} \]

Scalability

Host RAM

NIC

Precision and Low CPU overhead

CPU handles control plane operations

( Configuring queues, rate limits, packet classification )
SENIC Prototypes

- **NetFPGA 10G hardware prototype**
  - Demonstrates feasibility
  - Implements simple token bucket scheduler
  - Late binding of DMA transfers from host memory

- **Software prototype**
  - Dedicated CPU core for network scheduling
  - Works with any existing NIC
NetFPGA 10G Microbenchmarks

Synthesized at 100MHz with 1000 rate limiters
Is it Accurate?

- Synthesized at 100MHz with 1000 rate limiters
- Inter-packet delay for a traffic class

Average: within 0.038% of ideal pacer delay
Standard deviation: 1.7% of inter-packet delay

Pkt 1 → Pkt 2 → Pkt 3
1500B packets
Is it Fast?

- Scheduling decision latency:
  - 5 SRAM lookups (50 ns)

- 1500B packet at 40Gb/s: 300ns budget

- Smaller packets: schedule a burst at a time
Macrobenchmark: Tenant Isolation

X 8 machines

X 10 tenants

Memcached One-to-All

6Gb/s

UDP All-to-All

3Gb/s

SENIC
Macrobenchmark: Tenant Isolation

 Metrics:
  1. Memcached tail latency
  2. UDP throughput

 Compare SENIC to:
  1. Hierarchical Token Buckets (HTB)
  2. Parallel Token Buckets (PTB)

 Varying memcached tenant load
Memcached Tail Latency

(Lower is better)
UDP Tenant Throughput
(Closer to 3Gb/s configured limit is better)
UDP Tenant Throughput
(Closer to 3Gb/s configured limit is better)

SENIC accurately enforces rate limits and delivers high throughput

Memcached Load (% of configured 6Gb/s limit)
SENIC Supports Other NIC Features

1. TCP Segmentation Offload
2. Hypervisor Bypass + Untrusted Guest VMs
3. Constant-Time Hierarchical Scheduler
TCP Segmentation Offload

Host Memory

64KB TCP Segment

NIC

Wire

Header cached on NIC
DMA header and payload for each MTU sized packet
2X DMA transfers?
No Problem!
SENIC – TSO

- 40Gb/s, 1500B MTU: 6.5M DMA transfers per second
- Measurement from a Mellanox Connect-X3 NIC:
  - 13 – 14M DMA transfers per second supported
Summary

- Delivers vision of scalable rate limiting
- Accurate and precise
- Easily implementable in hardware and software

Code @ http://sivasankar.me/senic/