Performance Challenges
In
Software Networking

Stephen Hemminger

@networkplumber
Who am I?

- Principal Architect
  Brocade vRouter (Vyatta)

- Fellow
  Linux Foundation

- Sequent
  Unix SMP networking

- DPDK
  - #3 contributor

- Linux
  - 10+ year contributor
  - Maintainer
    - Bridge
    - iproute
Agenda

- Myths
- Requirements
- Benchmarks
- Reality
Myths

- Software networking can never do:
  - 1Gbit
    - 2008 – Linux, FreeBSD, ...
  - 10Gbit
    - 2013 – DPDK, Netmap, ...
  - 40Gbit
    - 2015 – DPDK, ...
  - 100Gbit
    - 2016?
Hardware vs Software

- Clock rate
- TCAM size
- TCAM miss
- Bus transactions

- Clock rate
- Cache size
- Cache misses per packet
- PCI bus operations
Optimization cycle

Measure

Optimize ← Analyze
SDN Measurement

- Forwarding
- RFC2544
- Scaling
- Imix, BGP, Firewall, ...
- Application
- BGP convergence
- Availability

SDN Workload
Performance
Test Environment
Benchmark vs Reality

- Benchmark
  - random flows
  - 10 or less rules
  - 128GB memory
  - 32 or more CPU

- Reality
  - Bursty flows
  - 1000’s of rules
  - 2GB VM
  - 2-4 CPU
System effects

- Data/Control resource sharing
  - CPU cache
  - Background noise
- Power consumption
- Memory footprint
- Virtualization overhead
- Platform differences
### Basics

<table>
<thead>
<tr>
<th></th>
<th>Sandy Bridge</th>
<th>Haswell</th>
<th>Skylake</th>
</tr>
</thead>
<tbody>
<tr>
<td>(bytes/cycle)</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>L1 Peak Bandwidth</td>
<td>2x16</td>
<td>2x32 load 1x32 store</td>
<td>2x32 load 1x32 store</td>
</tr>
<tr>
<td>L2 data access (cycles)</td>
<td>12</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>L2 peak Bandwidth</td>
<td>1x32</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>Shared L3 access (cycles)</td>
<td>26-31</td>
<td>34</td>
<td>44</td>
</tr>
<tr>
<td>L3 peak bandwidth</td>
<td>32</td>
<td>-</td>
<td>32</td>
</tr>
<tr>
<td>Data hit in L2 cache</td>
<td>43 – clean hit 60 – modified</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

memory is ~70+ ns away (i.e. 2.0 GHz = 140+ cycles)

Time Budget

- 10Gbit 64 byte packet
  - 67.2ns = 201 cycles @ 3Ghz
- Cache
  - L3 = 8 ns
  - L2 = 4.3
- Atomic operations
  - Lock = 8.25 ns
  - Lock/Unlock = 16.1

Network stack challenges at increasing speeds – LCA 2015
Jesper Dangaard Brouer
Magic Exlir?
Fast vs Slow

- New software
  - Lockless
  - Single function
  - Tight layering
  - Cache aware

- Legacy software
  - Interrupts
  - Shared resources
  - System calls
  - VM exit
Performance Tradeoffs

- Bulk operations
- Lock-less Algorithms
- Tight integration
- Polling
- Caching

→ Latency
→ Update speed
→ Consistency
→ Inflexible
→ CPU utilization
→ Power management
→ Memory utilization
→ Update overhead
Intel Haswell CPU pipeline
Cache flow

Worst case 7+ cache miss per packet!
Cache Ping/Pong

- Cache line shared between cores
  - Statistics
  - Session state
NFV bucket brigade
Packet batching
New developments

- DPDK
  - Multi-architecture
  - NIC support
  - Packet pipeline
  - ACL
  - LPM
  - ...

- Linux
  - Batched Tx
  - Lockless queue disciplines
  - Memory allocator performance
Conclusions

- Software networking is function of:
  - Algorithmics
  - Low level CPU utilization
  - Cache behavior
Questions?
Thank you

Stephen Hemminger
stephen@networkplumber.org
@networkplumber
Next Generation Software Networking

- Openvswitch + DPDK
- Brocade – vRouter
- 6Wind
- FD.io – VPP
- Juniper - Opencontrail
- Huawei - Fusionsphere
Performance Challenges In Software Networking

Stephen Hemminger

@networkplumber
Who am I?

- Principal Architect
  Brocade vRouter (Vyatta)

- Fellow
  Linux Foundation

- Sequent
  Unix SMP networking

- DPDK
  - #3 contributor

- Linux
  - 10+ year contributor
  - Maintainer
    - Bridge
    - iproute
Agenda

- Myths
- Requirements
- Benchmarks
- Reality
Myths

- Software networking can never do:
  - 1Gbit
    - 2008 – Linux, FreeBSD, ...
  - 10Gbit
    - 2013 – DPDK, Netmap, ...
  - 40Gbit
    - 2015 – DPDK, ...
  - 100Gbit
    - 2016?
## Hardware vs Software

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock rate</td>
<td>Clock rate</td>
</tr>
<tr>
<td>TCAM size</td>
<td>Cache size</td>
</tr>
<tr>
<td>TCAM miss</td>
<td>Cache misses per packet</td>
</tr>
<tr>
<td>Bus transactions</td>
<td>PCI bus operations</td>
</tr>
</tbody>
</table>
Optimization cycle

Measure

Optimize ← Analyze
SDN Measurement

Forwarding
RFC2544

Scaling
Imix, BGP, Firewall, ...

Application
BGP convergence
Availability

SDN Workload
Performance
Test Environment
Benchmark vs Reality

- Benchmark
  - random flows
  - 10 or less rules
  - 128GB memory
  - 32 or more CPU

- Reality
  - Bursty flows
  - 1000’s of rules
  - 2GB VM
  - 2-4 CPU
System effects

- Data/Control resource sharing
  - CPU cache
  - Background noise
- Power consumption
- Memory footprint
- Virtualization overhead
- Platform differences
## Basics

<table>
<thead>
<tr>
<th></th>
<th>Sandy Bridge Ivy Bridge</th>
<th>Haswell</th>
<th>Skylake</th>
</tr>
</thead>
<tbody>
<tr>
<td>(bytes/cycle)</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>L1 Peak Bandwidth</td>
<td>2x16</td>
<td>2x32 load 1x32 store</td>
<td>2x32 load 1x32 store</td>
</tr>
<tr>
<td>L2 data access (cycles)</td>
<td>12</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>L2 peak Bandwidth</td>
<td>1x32</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>Shared L3 access (cycles)</td>
<td>26-31</td>
<td>34</td>
<td>44</td>
</tr>
<tr>
<td>L3 peak bandwidth</td>
<td>32</td>
<td>-</td>
<td>32</td>
</tr>
<tr>
<td>Data hit in L2 cache</td>
<td>43 – clean hit 60 – modified</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

memory is ~70+ ns away (i.e. 2.0 GHz = 140+ cycles)

Time Budget

- 10Gbit 64 byte packet
  - 67.2ns = 201 cycles @ 3Ghz
- Cache
  - L3 = 8 ns
  - L2 = 4.3
- Atomic operations
  - Lock = 8.25 ns
  - Lock/Unlock = 16.1

Network stack challenges at increasing speeds – LCA 2015
Jesper Dangaard Brouer
Magic Exlir?
Fast vs Slow

- New software
  - Lockless
  - Single function
  - Tight layering
  - Cache aware

- Legacy software
  - Interrupts
  - Shared resources
  - System calls
  - VM exit
Performance Tradeoffs

- Bulk operations
- Lock-less Algorithms
- Tight integration
- Polling
- Caching

➔ Latency
➔ Update speed
   Consistency
➔ Inflexible
➔ CPU utilization
   Power management
➔ Memory utilization
   Update overhead
Intel Haswell CPU pipeline
Cache flow

Worst case 7+ cache miss per packet!
Cache Ping/Pong

- Cache line shared between cores
  - Statistics
  - Session state
NFV bucket brigade
Packet batching
New developments

- DPDK
  - Multi-architecture
  - NIC support
  - Packet pipeline
  - ACL
  - LPM
  - ...

- Linux
  - Batched Tx
  - Lockless queue disciplines
  - Memory allocator performance
Conclusions

- Software networking is a function of:
  - Algorithms
  - Low level CPU utilization
  - Cache behavior
Questions?
Thank you

Stephen Hemminger
stephen@networkplumber.org
@networkplumber
Next Generation Software Networking

- Openvswitch + DPDK
- Brocade – vRouter
- 6Wind
- FD.io – VPP
- Juniper - Opencontrail
- Huawei – Fusionsphere