Suricata

Extreme Performance Tuning

With Incredible Courage
By

• Michal Purzynski (@MichalPurzynski )
  – Threat Management, Mozilla
• Peter Manev (@pevma)
  – Suricata Core Team
  – Lead QA and training instructor
  – Stamus Networks
  – Mobster evangelist
For...

The brother and sister mobsters
Life of a packet...
...is full of dark alleys where it can get lost...

If span port Or packet broker

In the switch RX and TX buffers

In a card's FIFO

DMA Hardware IRQ sent Or not

In host's receive buffers

Sofnet NAPI poll

In AFP ring

Suri does not copy anything
All your packets belong to Suricata..... if they make it
The Grand Plan

Use a local CPU cache L3 as a data bus :-(
CPU socket

NIC location and NUMA relation
Suricata with AF_Packet on NUMA

CPU socket
NIC location and NUMA relation (HT)

CPU 0
- Core 0: Housekeeping
- Core 1: Data
- Core 2: Workers

CPU 1
- Core 14: Data
- Core 15: Workers

Thread allocation:
- p1p1 -> core 0-13 on NUMA - 0
- p3p1 -> core 14-27 on NUMA - 1
- p1p1 -> core 28-41 on NUMA - 0
- p3p1 -> core 42-55 on NUMA - 1

Suricon
CPU socket

NIC location and NUMA relation (HT)

p1p1 -> core 0-13 on NUMA - 0
p3p1 -> core 14-27 on NUMA - 1
p1p1 -> core 28-41 on NUMA - 0
p3p1 -> core 42-55 on NUMA - 1

Kernel – 3.x / 4.x with HT
Server

- One card per NUMA slot.
  - Or latency will kill you (™)
- Haswell, for best results
  - Independent P-states and other goodness
- C-states enabled. Turbo mode rocks IDS.
  - Linux will overwrite anyway
- L2 prefetchers disabled
  - Your card will send packets to L3 for you (DCA)
  - L2 prefetching impacts L3 and trashes it
- Snoop early, snoop frequently and never at home
NIC

Keep # of RSS queues == 1

Order in stream engine

avoid asymmetric hashing impact -> packet reordering
NIC indicators

- **rx_missed_errors**
  - packets overwritten in card’s FIFO

- **rx_no_dma_resources / rx_no_buffer_count**
  - means we are dropping packets because we don't have any free descriptors and cannot alloc new.
  - when using RSS=1 - “rx_no_dma_resources” does not increase counter

Use `ethtool -S <interface>` to discover magic
Use `ethtool -g <interface>` make ring smaller (yes)
The Cache consideration

- CPU load cache misses and latency
  - Execution WAITS till the data is fetched from RAM or other cache
  - MESIF (cache and memory coherence protocol) plays for Suri
- Happens when
  - Other software trashes caches (Redis/ES/DB...)
  - Threads are “jumping” between CPU sockets
  - NIC's ring descriptors are too big
  - HW prefetchers are enabled “wild wild west style”
The NUMA consideration

- MESIF plays for Intel
- Remote data access has a huge delay
- Remote node cache L3 access latency +- cache L3 miss latency
- Latency, not a cache misses are a problem for NUMA

Local L3 -+ 20ns - Remote L3 - >80ns
Local RAM - 96ns, remote 140ns
Cache thrashing effect in (bad)action

“A cache miss is a failed attempt to read or write a piece of data in the cache, which results in a main memory access with much longer latency.”
Cache thrashing effect in (bad)action

**CPU can not access RAM directly**

Performance counter stats for 'CPU(s) 0-39':

- 25,845,771,051 LLC-loads [50.00%]
- 7,214,035,111 LLC-load-misses [50.00%] # 27.91% of all LL-cache hits
- 3,768,432,384 LLC-stores [50.00%]
- 12,306,901,314 LLC-prefetches [50.00%]

60.001104971 seconds time elapsed

*If system is under pressure and packets are not in CPU L3 local cache = load_misses*
Which usually results in... 
(watch out for)

- capture.kernel_drops
- tcp.reassembly_gap
- Flow timeouts reached
- A small number of CPUs being pegged to 100%
- Drops in place no one looks for except us :-) 
  - ....and you now :)
So we looked at our set up (20Gbps)

- Suricata 3.2dev (using AF-PACKETv2/3)
- Kernel 4.4.0-38-generic #57~14.04.1-Ubuntu
- 21332 rules from ET Pro
- 128GB RAM, 8 DIMMS, 4 per socket.
- 2x Intel(R) Xeon(R) CPU E5-2697 v3 @ 2.60GHz - 28 cores total, HT enabled and used for 56 hardware threads.
- 2x dual port X520 AKA 82599 And X710 - one port on each card used. Cards installed into separate NUMA nodes.
Dug in research...

- Kernel code, ixgbe, i40e (and long nights)
- Suricata CPU affinity
- Intel e1000-devel mailing lists
- af-packet v2 and af-packet v3 for Suricata
- X520 and x710 NIC testing
- Suricata's new bypass feature
- 1 RSS queues considerations
After some time digging in...
We introduced the CPU affinity move....

- Step one - isolate cores and pin IRQs
- Step two - configure Suricata cpu affinity
- Step three - local bypass
Step one - isolate cores

- Cores are for you, not for a scheduler
- Steal them, leave one per node for housekeeping
- Single work type per dedicated core = no ticking
- Less userspace->kernel transitions, less TLB and cache trashing

```
nsm16 ~ cat /proc/cmdline
BOOT_IMAGE=/vmlinuz-4.4.0-45-generic.efi.signed root=UUID=dedcba7d-1909-4797-bd57-663a423a6a2f ro
processor.max_cstate=3 intel_idle.max_cstate=3 selinux=0 apparmor=0 mce=ignore_ce nohz_full=1-13,15-55 isolcpus=1-13,15-55 rcu_nocb=1-13,15-55
```
Step one - set all IRQs away from Suricata thread workers

- Have an IRQ-->core per node
- If not enough, use RPS but never split processing
Step two – Suricata cpu affinity

set-cpu-affinity: yes
# Tune cpu affinity of suricata threads. Each family of threads can be bound
# on specific CPUs.
cpu-affinity:
  - management-cpu-set:
    cpu: [0,28,14,42]  # include only these cpus in affinity settings
    mode: "balanced"
    prio:
      default: "low"
  - detect-cpu-set:
    # NUMA and Hyper-threading example on kenel 4.x
    # NUMA order -> 0/1/0/1
    # (2x14 cpus =56 total with HT)
    # 2 x NICs. 1 - plp1 and 1 - p3p1
    # plp1 -> 3-13 on NUMA-0 / p3p1 ->17-27 on NUMA-1 /
    # plp1 -> 31-41 on NUMA-0 / p3p1 -> 45-55 on NUMA-1
    cpu: ["3-13","17-27","31-41","45-55"]
    mode: "exclusive"  # run detect threads in these cpus
    # Use explicitely 3 threads and don't compute number by using
    # detect-thread-ratio variable:
    # threads: 3
    prio:
      default: "high"
Step two – Suricata cpu affinity

- interface: p1p1
  threads: 11
  cluster-id: 99
  use-mmap: yes
  tpacket-v3: yes
  ring-size: 400000
  block-size: 393216
  #buffer-size: 1048576
  ##buffer-size: 262144
  cluster-type: cluster_flow

- interface: p3p1
  threads: 11
  cluster-id: 98
  use-mmap: yes
  tpacket-v3: yes
  ring-size: 400000
  block-size: 393216
  #buffer-size: 1048576
  ##buffer-size: 262144
Step three – local bypass

• Local bypass
  – If the corresponding flow is local bypassed then it simply skips all streaming, detection and output and the packet goes directly out in IDS mode and to verdict in IPS mode.
  – In suricata.yaml
    • Set by “stream.bypass: yes”
    • Adjusted by “stream.reassembly.depth”
Performance Before

```
Terminal — ssh nsm16.private.scl3.mozilla.com — zsh

```

```
Tasks: 48, 108 thr; 77 running
Load average: 1.87 1.79
Uptime: 5 days, 15:24:47

```

<table>
<thead>
<tr>
<th>PID</th>
<th>USER</th>
<th>PRI</th>
<th>NI</th>
<th>VIRT</th>
<th>RES</th>
<th>SHR</th>
<th>S</th>
<th>CPU%</th>
<th>MEM%</th>
<th>TIME+</th>
<th>Command</th>
</tr>
</thead>
<tbody>
<tr>
<td>12747</td>
<td>root</td>
<td>20</td>
<td>0</td>
<td>6389M</td>
<td>5471M</td>
<td>1628S</td>
<td>S</td>
<td>3367</td>
<td>4.2</td>
<td>2:01.29</td>
<td>/etc/nsm/suricata-noHT.conf --pidfix</td>
</tr>
<tr>
<td>3785</td>
<td></td>
<td>25</td>
<td>5</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>S</td>
<td>450.22.5</td>
<td>7h25:38</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
<tr>
<td>3856</td>
<td></td>
<td>23</td>
<td>3</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>R</td>
<td>50.8 22.5</td>
<td>41:40:67</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
<tr>
<td>3840</td>
<td></td>
<td>23</td>
<td>3</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>R</td>
<td>24.7 22.5</td>
<td>18:15:21</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
<tr>
<td>3852</td>
<td></td>
<td>23</td>
<td>3</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>R</td>
<td>23.4 22.5</td>
<td>20:02:11</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
<tr>
<td>3849</td>
<td></td>
<td>23</td>
<td>3</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>R</td>
<td>23.4 22.5</td>
<td>18:54:54</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
<tr>
<td>3846</td>
<td></td>
<td>23</td>
<td>3</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>S</td>
<td>23.4 22.5</td>
<td>18:34:35</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
<tr>
<td>3858</td>
<td></td>
<td>23</td>
<td>3</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>R</td>
<td>22.0 22.5</td>
<td>18:22:58</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
<tr>
<td>3859</td>
<td></td>
<td>23</td>
<td>3</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>R</td>
<td>21.4 22.5</td>
<td>25:39:71</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
<tr>
<td>3843</td>
<td></td>
<td>23</td>
<td>3</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>R</td>
<td>20.7 22.5</td>
<td>18:38:01</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
<tr>
<td>3855</td>
<td></td>
<td>23</td>
<td>3</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>R</td>
<td>20.0 22.5</td>
<td>17:45:92</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
<tr>
<td>3847</td>
<td></td>
<td>23</td>
<td>3</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>R</td>
<td>20.0 22.5</td>
<td>18:28:53</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
<tr>
<td>3854</td>
<td></td>
<td>23</td>
<td>3</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>R</td>
<td>19.4 22.5</td>
<td>18:22:27</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
<tr>
<td>3851</td>
<td></td>
<td>23</td>
<td>3</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>S</td>
<td>18.7 22.5</td>
<td>18:01:08</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
<tr>
<td>3845</td>
<td></td>
<td>23</td>
<td>3</td>
<td>30.1G</td>
<td>28.3G</td>
<td>13.3G</td>
<td>R</td>
<td>18.7 22.5</td>
<td>18:17:37</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT.conf --pidfix</td>
<td></td>
</tr>
</tbody>
</table>

```
Performance after

```
Terminal — ssh nsm16.private.scl3.mozilla.com — zsh

```

```
Tasks: 44, 52 thr; 8 running
Load average: 2.19 2.03
Uptime: 1 day, 14:05:00

```

<table>
<thead>
<tr>
<th>PID</th>
<th>USER</th>
<th>PRI</th>
<th>NI</th>
<th>VIRT</th>
<th>RES</th>
<th>SHR</th>
<th>S</th>
<th>CPU%</th>
<th>MEM%</th>
<th>TIME+</th>
<th>Command</th>
</tr>
</thead>
<tbody>
<tr>
<td>21574</td>
<td>root</td>
<td>25</td>
<td>5</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>R</td>
<td>434</td>
<td>23.2</td>
<td>5h34:01</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
</tr>
<tr>
<td>21659</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>R</td>
<td>32.6</td>
<td>23.2</td>
<td>2h26:33</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
<tr>
<td>21643</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>S</td>
<td>30.5</td>
<td>23.2</td>
<td>2h24:35</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
<tr>
<td>21646</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>R</td>
<td>29.2</td>
<td>23.2</td>
<td>3h19:11</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
<tr>
<td>21645</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>R</td>
<td>25.8</td>
<td>23.2</td>
<td>2h22:11</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
<tr>
<td>21658</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>R</td>
<td>23.1</td>
<td>23.2</td>
<td>2h18:05</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
<tr>
<td>21649</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>S</td>
<td>21.0</td>
<td>23.2</td>
<td>2h23:10</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
<tr>
<td>21654</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>S</td>
<td>20.4</td>
<td>23.2</td>
<td>2h16:50</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
<tr>
<td>21662</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>S</td>
<td>19.7</td>
<td>23.2</td>
<td>2h24:21</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
<tr>
<td>21657</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>S</td>
<td>19.0</td>
<td>23.2</td>
<td>2h18:27</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
<tr>
<td>21652</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>S</td>
<td>19.0</td>
<td>23.2</td>
<td>2h25:52</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
<tr>
<td>21641</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>S</td>
<td>17.6</td>
<td>23.2</td>
<td>2h23:41</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
<tr>
<td>21655</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>S</td>
<td>17.0</td>
<td>23.2</td>
<td>2h23:14</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
<tr>
<td>21651</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>S</td>
<td>17.0</td>
<td>23.2</td>
<td>2h22:44</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
<tr>
<td>21642</td>
<td>23</td>
<td>3</td>
<td>30.3G</td>
<td>29.2G</td>
<td>13.3G</td>
<td>S</td>
<td>17.0</td>
<td>23.2</td>
<td>2h20:45</td>
<td>/opt/suricata/bin/suricata -c /etc/nsm/suricata-noHT-clusterondie</td>
<td></td>
</tr>
</tbody>
</table>
```

F1 Help  F2 Setup  F3 Search  F4 Filter  F5 Tree  F6 Sort By  F7 Nice  F8 Nice + F9 Kill  F10 Quit
Cache misses Before and After

```bash
perf stat -e LLC-loads,LLC-load-misses,LLC-stores,LLC-prefetches -C 1

Performance counter stats for 'CPU(s) 0':

<table>
<thead>
<tr>
<th>Counter</th>
<th>Value</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLC-loads</td>
<td>1939117135</td>
<td>(66.67%)</td>
</tr>
<tr>
<td>LLC-load-misses</td>
<td>289305806</td>
<td># 14.92% of all LL-cache hits (66.67%)</td>
</tr>
<tr>
<td>LLC-stores</td>
<td>356341541</td>
<td>(66.67%)</td>
</tr>
<tr>
<td>LLC-prefetches</td>
<td>&lt;not supported&gt;</td>
<td>(66.67%)</td>
</tr>
</tbody>
</table>

114.342786481 seconds time elapsed

Core handling kernel. Ring descriptor size 512 buffers. Each buffer is 2048 bytes in size.
```

```bash
perf stat -e LLC-loads,LLC-load-misses,LLC-stores,LLC-prefetches -C 1 sleep 60

Performance counter stats for 'CPU(s) 1':

<table>
<thead>
<tr>
<th>Counter</th>
<th>Value</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLC-loads</td>
<td>659135009</td>
<td>(66.67%)</td>
</tr>
<tr>
<td>LLC-load-misses</td>
<td>1372230</td>
<td># 0.21% of all LL-cache hits (66.67%)</td>
</tr>
<tr>
<td>LLC-stores</td>
<td>124004599</td>
<td>(66.67%)</td>
</tr>
<tr>
<td>LLC-prefetches</td>
<td>&lt;not supported&gt;</td>
<td>(66.67%)</td>
</tr>
</tbody>
</table>

60.001419120 seconds time elapsed
Suricata packet drops after 0.00137% On 20Gbps
Mobster myths

“AFP is slow, you must use <a new kernel bypass>” - is it? :-)

“Linux cannot deal with interrupts (use BSD)” - we have NAPI from > 16 years

“Use RSS/RPS to load balance” - and get packets reordering and missing events

“Disable HT it hurts performance” - actually, the opposite

“Make <buffer> huge” - and it won’t fit into L3
Lessons Learned

➢ Make no assumptions - verify everything

➢ Understand the packet travel critical path

➢ Understand what your counters _really_ mean

➢ NUMA is awesome - know how to use it

➢ We run out of traffic....

➢ Everything can be undone with a bunch of badly written rules
Things to watch out for

- `rx_missed_errors` (NIC/ethtool)
- `rx_dma_failed /rx_no_buffer_count` (NIC/ethtool)
- Switchport
- Cache misses
- Correct traffic mirroring (esp vlan tags)

- Suricata
  - Memcaps
  - Reassembly gaps
  - Flow emergency
  - Decoder invalid
  - Make sure MTU is same across
Detailed info to come up

• Research paper
  – Containing all the details of this research
  – Commands and scripts
  – Trouble shooting advice and guidance
  – Perf Indicators

• Potential inclusion in Suricata's docs/repo (PR)

• Some other interesting experiments....
Thanks to

- Mozilla (time, traffic, hardware)
- Intel - Alexander Duyck
- Eric Leblond (@Regit – Suri cpu affinity)
- Daniel Borkmann (netsniff-ng, AFP)
- And Dave Miller for AFP :-)
- SuriCon 2016 !!