Is Your Elephant a Gazelle?

How to accelerate IPsec elephant flows

Fan Zhang
Agenda

• What is elephant flow and where is the bottleneck to secure them with open-source IPsec solutions?
• FD.io VPP and VPP IPsec
• VPP Synchronous Crypto Infra Introduction
• VPP Asynchronous Crypto Infra Introduction
• Scale Single IPsec Flow Even More
• Summary
What is Elephant Flow (EP)?

- Extremely large continuous flow in Internet
- 4.7% packets in total, takes 41.3% bandwidth

How Userspace data plane handles IPsec:
- Isolated and limited per core processing resource, including stack and crypto.
- Flow-to-core affinity.

This makes IPsec EP handling difficult
Pain Points of Processing IPsec EP

- Crypto processing requires large amount cycles, while EP is mostly large packets
- Flow-to-core affinity always make one core extremely busy, while other cores relaxing.
- A perfect core extremely powerful to handle the flow also means wasting the cycles most of the time.
- Load-balancing single flow to multiple cores will cause race condition when anti-replay is enabled.
- We propose our answer to resolve the problems with FD.io VPP IPsec
<table>
<thead>
<tr>
<th>Out of the box</th>
<th>Development Toolkit for ...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cloud Infra, Discrete Appliance, Virtual Network Functions ...</td>
</tr>
<tr>
<td>Software</td>
<td>Framework, API &amp; Libraries</td>
</tr>
<tr>
<td>Architecture</td>
<td>Packet Processing Pipeline,</td>
</tr>
<tr>
<td>Interfaces</td>
<td>configuration driven,</td>
</tr>
<tr>
<td>Networking</td>
<td>composable &amp; extensible</td>
</tr>
<tr>
<td>Extensible</td>
<td>Wide Device Support</td>
</tr>
<tr>
<td>Integrations</td>
<td>Select Native Drivers &amp;</td>
</tr>
<tr>
<td></td>
<td>Libraries &amp; Sample Code</td>
</tr>
<tr>
<td></td>
<td>Wide Protocol Support</td>
</tr>
<tr>
<td></td>
<td>SDK Model</td>
</tr>
<tr>
<td></td>
<td>Realized through</td>
</tr>
<tr>
<td></td>
<td>Plugin Model</td>
</tr>
<tr>
<td></td>
<td>Support for</td>
</tr>
<tr>
<td></td>
<td>openstack.</td>
</tr>
<tr>
<td></td>
<td>Kubernetes</td>
</tr>
<tr>
<td></td>
<td>Discrete Appliances</td>
</tr>
</tbody>
</table>

*Other names and brands may be claimed as the property of others.*
FD.io VPP IPsec

- Open-Source Production-ready IPsec Implementation.
- Capable of single server 1Tb IPsec processing.
- Supports AH, ESP (tunnel and transport), ESPoUDP, ESPoGRE.
- Supports major crypto algorithms (AES-CBC, HMAC-SHA*, AES-GCM).
- Supports multiple crypto engine plugins.
- Supports both CPU based crypto (VAES) and Lookaside HW accelerations (QAT)
- Efficient and Cloud-friendly
FD.io VPP Native Crypto Infra (before VPP 20.05)

- A generic infrastructure to provide symmetric crypto service within VPP
- Provides generic API and multiple crypto plugin engines supporting:
  - Key management (add, delete, and update)
  - Crypto operation (cipher, hash, AEAD)
- Advantages
  - Performance
  - Availability
  - Flexible
- Disadvantages
  - No HW offload support
  - Single IPsec Flow crypto Scaling not possible
Scale VPP IPsec Single Flow Throughput With Crypto Offload

- IPsec = packet processing (pps sensitive) + crypto (bps sensitive)
- Offload crypto workload to
  - Dedicated HW (e.g. QAT)
  - Dedicated CPU core(s)
  helps gaining more cycles to packet I/O and stack processing.
- To support both, we need a generic asynchronous crypto infrastructure.
VPP Async Crypto Infra

- Released in VPP 20.05
- Share the same key management as synchronous crypto infra.
- Provides Generic Enqueue and Dequeue Handler.
- User graph node enqueues the packets to the target engine.
- A dedicated dispatch graph node will handle dequeue.
Adding QAT Hardware acceleration with DPDK Cryptodev

• New DPDK Cryptodev RAW API
  • A more compact data structure.
  • Raw buffer pointer and physical address as input.
  • More sophisticated enqueue/dequeue control method.
  • Customizable status field set callback function design.
• ~15% performance improvement.
• New DPDK Cryptodev Raw API will be released in DPDK 20.11
• The change has already been merged in VPP 20.09 as a DPDK 20.08 patch.
Elephant flow without QAT? SW-scheduler Crypto Engine

- A pure SW crypto engine that utilizes dedicated CPU cores to process crypto workload.
- Crypto worker threads actively scan the frame queue, mark unprocessed frame as “WIP”, and processed frame as “Complete”.
- Dispatch dequeue first N “Complete” frames.
... Also cloud friendly!

- Crypto Dispatch Node running in polling mode can achieve best possible performance, but it is unfriendly to cloud-native use case.
- That’s why we made it supporting interrupt mode.
- Active polling within an interrupt handling
- Precise signaling when a crypto frame is enqueued/processed.
Can We Scale Single IPsec Flow Even More?

- With Async Crypto we achieved single IPsec flow processing capability of up to 40Gbps.
- Even with crypto offloaded, there are still heavy I/O and stack processing left.
- Intel® DLB or DPDK SW eventdev offers the way to distribute the packets to multiple CPU cores. The packet ordering is maintained from RX to TX.
- With the help of DLB or SW eventdev, we may load-balance most single flow IPsec workload to more cores.
Can We Scale Single IPsec Flow Even More? (cont.)

- Only non-distributable workload (SQN update and check) is handled by a single core.
- Load-balancing more workload to other CPU cores helps regaining more cycles to receive packets.
  - IPsec Stack Processing
  - Crypto
  - Tx (post IPsec)
- Development ongoing, estimate to finish and upstream EOY 2021.
- Our goal is to achieve 100Gbps single IPsec flow processing capability.
Summary

• VPP Synchronous Crypto Infra provides amazing performance to process IPsec workload, but fails to scale with bigger flow.

• We provided asynchronous crypto infra to make SW and HW offloading possible to scale IPsec single flow throughput. The infra supports interrupt mode to make it cloud-native friendly.

• We also provided Cryptodev and SW scheduler async crypto engines.

• Both async crypto engines helped to achieve 40Gbps IPsec elephant flow processing.

• To scale the single IPsec flow even further, we process offload both crypto and most IPsec stack to other cores with Intel® DLB or DPDK Eventdev.
Thank you very much!

Q&A

For questions please contact: roy.fan.zhang@intel.com