Hardware offload BOF

Roopa Prabhu, Shrijeet Mukherjee

A fantasy agenda

Capabilities

a.Explicit list

Or Query serially, punt to higher level (explicit hierarchy)

Or Model each device uniquely with capability (no hierarchy)

b. Need to understand this for Switch Asic versus VEPA/EVB/SRIOV nic etc

Flow offload :

c. Manage as discrete devices or generic pipeline

How is interop measured aka how to avoid anarchy :)

d. Flow API scheme [John Fastabend]

d’.Model using P4 [Mihai]

f. TC scheme [Jiri Pirko]

f’. EZChip [Gilaad]

A fantasy agenda

Routing tables, FDB, MDB, ACL

g. Capacity indication, aka properties of tables [Roopa]

h. Fine grain capability (e.g. is it sufficient to ask if multicast is supported)

j. Table characteristics LPM versus Logical Hash based LPM's and practical implications

Device model : (Not mutually exclusive in anyway)

k. Maintaining operational consistency is KEY

Make switch look like NIC or vice versa e.g. is learning a basic capability ?

l. Model using OVS (inherently host based)

m. Model using rocker [Scott]

n. Switch Abstraction Interface [Sanjay]

n'. Intel [Uri]

n". Qualcomm [Olivari]

A fantasy agenda

Features :

- l3 offloads [Hannes]
- acl offloads [Pablo]

o. Load Balancing

p. Bonding ++ (MLAG and friends)

q. Stateful packet processing

Etherpad output

- **Hardware offload BoF**
- Netdev01 - Sun, Feb 15, 2015 1:00pm

- Major points:
- Preserve the Linux Networking Model
- Goal is to exchange ideas

- Capability Determination
- Patrick: Can the default route be removed? So that individual routes can be removed. Use route usage statistics to remove least used routes.
- David: Existing tools must continue to work.
- Signal an error if out of space, or
- Provide capacity initially and restrict usage to that limit
- Cares:
- Must be able to support the decisions we make
- Partick: A flag in the route add ??? [ed. help]
- Ben: Need a fairly good switch model to emulate the hardware devices.
- ???: The least recently used method proposed by Patrick may lead to periods of major route updates when network changes occur. With multiple routing suites each one needs to be given the capabilities (capacities) of the underlying switch
- Shrijeet: General agreement: minimal policy in the kernel, most in userspace

- Switch ASIC vs. SRIOV (NIC) Switching models
- Shrijeet: Do these two need to be different?
- Andy: Come to my talk tomorrow.
- Thomas: Bring all the DSA drivers onboard
- Gilaad: Need to define common features, plus need a way to include additional (unique) capabilities
- David: Look to John Fastabend's work. "I'm ok having three abstractions"
- John: Use flow API to query device and provide capabilities
- Gilaad: Offload isn't limited to HW. The offload can be done by software too.
- David: E.g. Hypervisor offload from VMs.
- Ben: We need to define the building blocks. E.g. Hash tables, TCAMs. Hardware implementation should be mimicked as closely as possible.
- John: Can table attributes be used?
- David: There is a learning curve necessary to determine the best way to proceed. Everyone is an expert in his/her own isolated area.

- Discrete devices vs. generic pipeline
- Alexi: NDAs!!!

- Flow API Scheme
- John: Expose headers the HW supports, actions (set field, pop tag, etc.), tables and their attributes, table operations (add, remove), pushed to userspace via Netlink,
- Ben: Who is the master of this information? E.g. "tc". Should it be subservient to this information or should it be the master.
- Patrick:
- Jamal: Offload to the HW until there is no more room?
- Eric B: The hardware is many, many, orders of magnitude faster than software in many cases. Thus fallback to CPU is impractical.
- Patrick: That's OK. Hardware is just speeding things up.
- Thomas: Customers do not expect SW processing rates.
- David: We need a bit "If it doesn't fit in the HW let me know" (to userspace).to address concerns from Lisa
- Patrick: There are some devices which are not under NDA and have (crappy) drivers
- Neil: Sometimes in HW the same resources are used for multiple functions. This makes capability determination very difficult.
- ??? remembers that it is mandatory to maintain the kernel states like conntrack for example
- Patrick agrees and explains that for conntrack, it should not be a problem, only timers need to be updated and there is already APIs to do that.

- There will be another session later today at 4:30pm.

- **Hardware offload BoF**
- Netdev01 - Sun, Feb 15, 2015 4:30pm

- Presentation - Mihai: P4: Specifying Data Planes
- Refer to slides
- Jamal: Q: What's the problem with loops? A: Loops do not have predictable latency. Q: Is everything a table (data types)? A: There are structures
- ???: Q: Can byte processing across the payload be done? A: Some facilities can be done, like checksum, but not crypto. Q: How would you handle TLVs without having loops? A: Parser can have loops.
- Q: What is the dataplane? A: Generated from the P4 program.
- Shrijeet: Q: How does stateful processing work (e.g. ECMP)? A: Counters can be used, but are fairly limited. You can also send to the control plane.

- Flow API Discussion - Jiri
- Jiri: concerned about new proposed flow API bypassing the linux kernel
- Thomas: There is value in moving complexity from the kernel to userspace. And we need to find a middle ground.
- daveM: kinda disappointed by the frankenstein ovs has become. proper protocol should have been added to the kernel
- Thomas: what is your answer to P4
- shm: P4 allows rebuilding your NIC. And allows a P4 software implementation.
- Fastabend: What if people want to support non-standard protocols ?
- DaveM: There are many concerns over what if we cannot support some thing that hardware supports. Lets cross that bridge when we get there. make proper interfaces.
- Jamal: I don't like the bypassing of the Linux kernel. My tools iproute2, tc should work. Extend tools if they don't support it today.
- DM: I have always seen it as a framework
- Thomas: Another example nftables. The kernel takes bytecode or JIT'ed code as input.
- Should this be given directly to hw driver ?.
- shm: You should. The switch driver should be able to support this by interpreting the byte code.
- Patrick: We are not sending bytecodes to the kernel, we are sending netlink messages. It would not be hard to support nftables offload.
- DaveM: Another idea was that Pablo wanted to make a separate nftables HW chain for such offloads.
- DaveM: I will not allow arbitrary exposure of paths to flow offload to hardware I promise.

- Gilaad - NPU Offload Discussion
- Does not favor a userspace API to poke the HW. Used example of SNORT to explain the split between HW and SW offloading. Another example of SSH decryption which can be done in SW or specialized hardware. Offload should encompass more than basic L2/L3 functions which are currently implemented in existing ASICs. Make the offload approach more turing complete.
- DaveM: The crypto example is a good one.

- Matty - What is the best way to allow kernel switch drivers?
- First step: Ability to represent ports, send/receive, link control/status, counters
- Next step: Resource management. Some resources are shared among different functions. Can't simply advertise table sizes.
- Next Step: Start porting resource management into the kernel.
- Next Step: More complex data structures: e.g. ALPM and ISSU (stateful restart)
- DaveM: Get the netdev code committed right away (first step). The resource management should be done in a private playground, until the kernel constructs are defined.
- Matty: Can a driver with limited functionality and some other new APIs for ???
- DaveM: Don't bring into the tree something which defines new APIs that haven't been defined yet.
- Aviad: Offered to educate the community about switch ASIC capabilities

- Roopa - SwitchDev
- Seamless offloads - fib and fdb offloads using NETIF_F_HW_SWITCH_OFFLOAD
- Flags for identifying "software only" or "hardware only" operations.
- Duplicate packet handling (for packets already forwarded in hardware)
- Update kernel counters with hardware counters - Some devices like a bridge have both HW and SW forwarded frames which need to get combined.
- LAG offloads: Can be implemented today using existing notifiers.

- ScottF - Rocker
- An emulated device to fill a gap of a switchdev driver model. Offloads forwarding data plane. Runs on QEMU. No physical devices exist with cumbersome NDA, non-upstreamable code. Currently supports basic L2 functions. In progress is L3 functionality. Others are working on nftables and flow API offloading. Device is not in Linux kernel. Use rocker as an example for real hardware devices. Goal is to solidify the API.
- DaveM: "If you are the first vendor to upstream a switch device you will be so cool"

- Sanjay - OCP SAI
- Refer to presentation slide
- A common switch HW device API.
- Jamal: Is this a wrapper around SDK APIs?
- DaveM: Good news is that this unifies the SDK interface. Bad news is that it encourages perpetuation of proprietary SDK development.
- Aviad: SAI is first time that switch vendors and customers have come together to support a common interface.
- Shrijeet: netdev is the common NIC API.
- Aviad: Need to work within the eco-system to develop this API
- DaveM: Main use of SAI is to discover the scope of the problem

- DaveM - No protocol is an island. It interacts with other components (tc, MTU settings, etc.)

- Mathieu - IPQ806x Hardware acceleration
- Refer to presentation slides.
- John Fastabend: How do you decide which things are offloaded and which are not? Can functions be split?
- Sol: Everything is offloaded until it can't be any longer and then it is no longer.
- JohnF: Do we need userspace to tell when to program the entries in the HW.
- Ben: Detectors weigh in and unanimous decision must be made to accelerate.
- Patrick: Doesn't matter if the policy decision is in the kernel or not.
- Sol: Touchpoints - places in the Linux kernel where information is needed or needs to be provided. These touchpoints should be as light as possible. How do we get them in the kernel?
- Jamal: Can ECM module be generalized?
- Ben: Yes.
- Sol: Some piece of logic must decide if something can be offloaded or not. All code is open sourced, but not upstreamed.
- Patrick: Very close to what he would have done except for the post routing hook and connection tracking hook.
- Shrijeet: HW stats also need to be handled.
- Ben: We also have a switch device, but it looses conntrack stats.

- Pablo - Netfilter Interface for ACL HW Offload
- Refer to presentation slides
- Patrick: Why not use a common flow abstraction layer for the hardware?
- Pablo: Conversion is done in nftables and so it is common
- Roopa: Will a seamless offload work without the special chain?
- Pablo: Explicit semantics are needed to define what goes into HW and what does not.
- DaveM: This is another approach for the policy selection.
- JohnF: Verification code should not be replicated across drivers. Should be common.
- Shrijeet: Seamless iptables sync is being done today.
- Patrick: Semantic equivalence between HW and SW is impossible.
- JohnF: We don't necessarily want the same rules in HW as SW.
- Roopa: This (nftables) can't be the only model
- (Back and forth between Shrijeet and Patrick about ability to generally accelerate rules. Jamal and Thomas join in.)
- JohnF: Can there be a chain per table?
- Patrick: Yes, that should be supported.
- Thomas: It would be great to have this in TC.
- Patrick: That is something I've been thinking about. It will have to be done.
- Jamal: tc has been offloaded to hardware, a long time ago.
- Simon: How does it associate an object to a device?
- Pablo: The base chain is always associated to a device.
- Alexi: [ed: help, got distracted]

- Hannes -

Progress made ..

- Agreement on 2 phase transaction model
- Agreement on hierarchical ops and attribute management
- Mellanox takes pole position

Immediate things worth looking at

- Merge embedded nic modeling with switchdev ops
- Stats, stats and stats
- Ethtool type ops
- Device manager ops

- Resource management

- Datapath modeling is a different BOF

Updates

- Mellanox – Jiri
- NFT – Pablo
- JohnF – TC/classifier offload
- Features – Roopa
- Wireless – anyone ?
- Who else wants to provide some update ?