DPDK SRIOV and control over Embedded Switch

Alex Rosenbaum

DPDK Summit Userspace - Dublin- 2017
Agenda

- Background and Needs for SR-IOV
- New Use Cases for SR-IOV and embedded switch
- Discussion
Background
Para-virt networking
Para-virt networking

DPDK user space networking for high performance (burning CPU)
Para-virt networking

DPDK user space networking for high performance (burning CPU)

SR-IOV with its own management (separate embedded switch)
Para-virt networking

DPDK user space networking for high performance (burning CPU)

SR-IOV with its own management (separate embedded switch)

switchdev and OVS integrated in Kernel (controlled via tc commands)
New Use Cases
Use Case 1: Hypervisor offload

- OVS DPDK with direct data path to VM’s
  - switchdev SR-IOV offloads already implemented in Kernel OVS
  - Use DPDK ‘slow’ path for exception flows or unsupported HW features
- allow DPDK control and data path of embedded switch
  - Representor ports are exposed over the PF
  - Data Path RX & TX queues per representor
    - Send/receive packet to/from VF is done through it’s representor
  - ACL, steering, routing
  - encap/decap
  - flow counters
  - IPSec
- Co-exists with para-virt solutions
Use Case 2: Secured Hypervisor

- **SmartNIC**
  - NIC + CPU/RAM
  - Linux
  - Run applications

- **OVS DPDK to manage VF’s**
  - HW offloads depending on capabilities
  - High perf SW fast path for advanced features

- **3rd party applications**

- **Full SDK**

**Diagram:**
- **Host**
  - Guest PV
    - VF driver
  - Guest PV
    - VF driver
  - Guest VF
    - VF driver

- **Smart NIC**
  - OVS DPDK
  - neutron
    - PMD
    - PMD
    - PMD
  - Linux
    - TC Flow Offloads
  - Embedded Switch
    - uplink
    - Uplink-Rep
    - PF-Rep
    - VF0-Rep

**Note:**
- netdev uplink
Use Case 3: Bare Metal Provisioning

- Manage networking for a Bare Metal Server
  - ACL, FW
  - Transparent Tunnels: VLAN, VXLAN, GRE, IPSec
  - Monitoring: flow counters
VF representors are:
  - a PMD of its own? per port? Holding a switchdev index?
  - Ports in a new rte_switchdev?

Embedded switch control plane
  - TC already controls the eSwitch in upstream kernel – application can use this directly
  - Dedicated DPDK API’s – should match the kernel’s interface/parameter list

Name mapping between VF and it’s VF representors
Questions?

Alex Rosenbaum
alexr@mellanox.com
More about...

- BlueField SoC
- eSwitch model
- Detailed NIC offload capabilities