Linaro

VPP on OpenDataPlane enabled SmartNICs

Openstack Summit 2017 - May 10th

François-Frédéric Ozog

LNG
Networking

Linaro Overview

- Linaro leads software collaboration in the ARM ecosystem
- Instead of duplicating effort, competitors share development costs of core software to accelerate innovation and time to market
- Linaro is member funded and delivers output to members, into open source projects, and into the community
- Founded in 2010 with 6 members, now 35, with 140 staff and ~300 OSS engineers distributed globally

ARM® HISILICON QUALCOMM INNOVATION CENTER, INC. QuIC

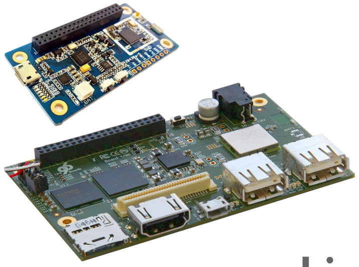
Google socionext for better quality of experience SPREADTRUM®

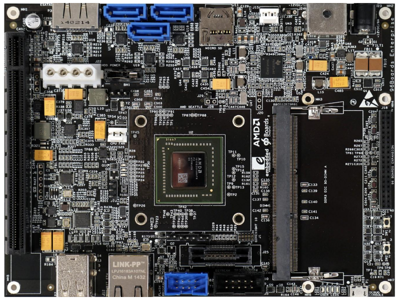
ST life.augmented TEXAS INSTRUMENTS ZTE Tomorrow never waits

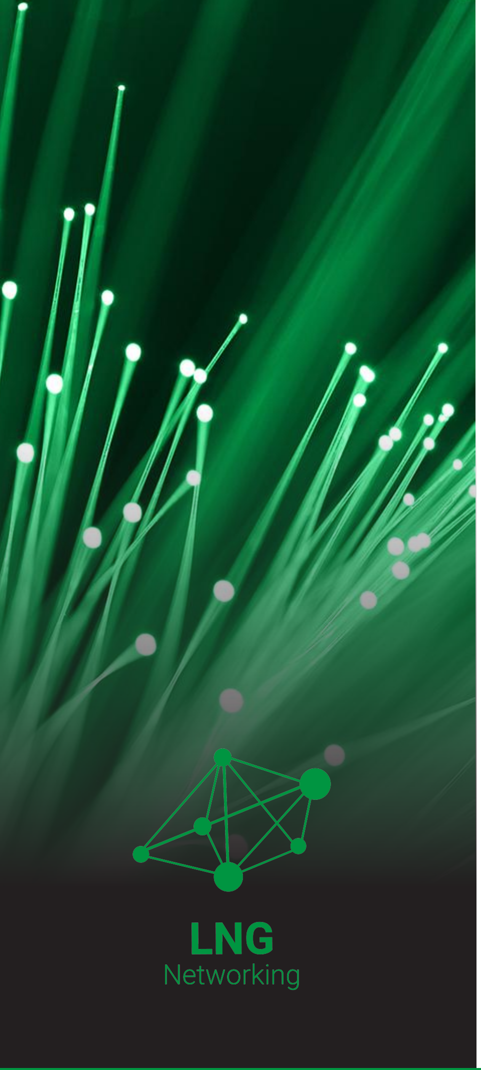
Core and Club Members

AARCH 32 AARCH 64

TrustZone Security Foundation by ARM







Agenda

What OpenDataPlane brings to VPP

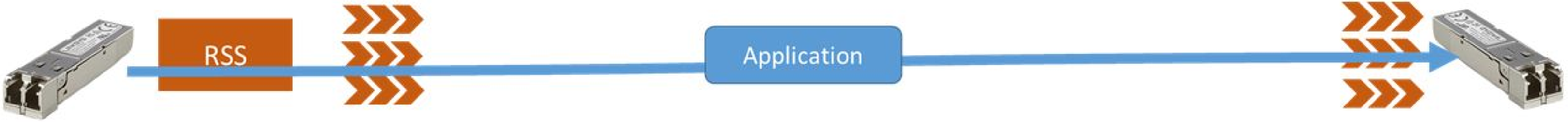
direct-virtio

odp-net

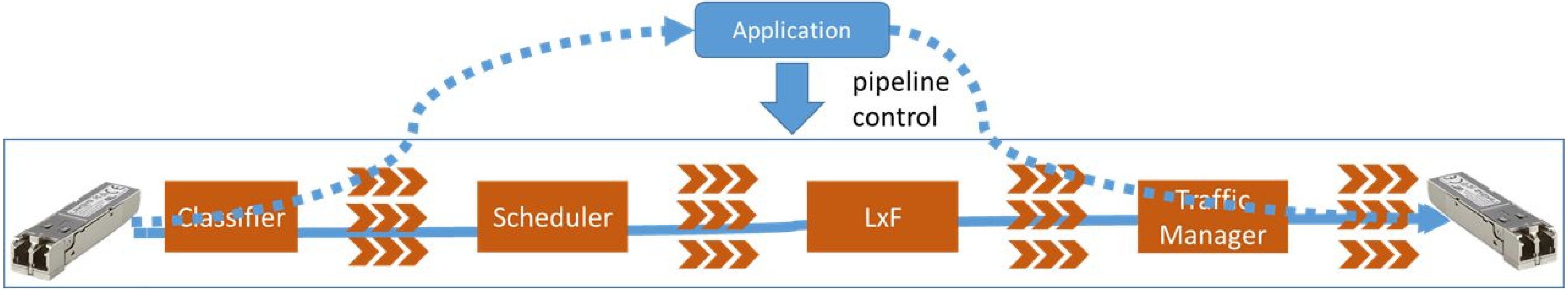
odpCL

Dataplane software - the two families

Software Implemented: DPDK, Netmap, PacketDirect...



Software Defined: OpenDataPlane



ODP software development

- Event based programming framework
  - Key performance element for our members that have large independent development teams
  - -> building block dev team / product teams
- Yet, can be doing poll mode run to completion
- IPsec fat pipe: packet ordering can be very important:
  - Hardware and software frameworks
- Two companion projects
  - OpenFastPath for socket + TCP/IP layer for traffic termination
  - VPP for switching and routing
  - Offloading of TCP from VM to infrastructure will allow transparent DDoS attack mitigation
- Low speed or low trafficing interfaces can share polling cores

ODP : hardware first

- No/almost no software metadata
    - rte_mbuf is a two cacheline object
    - odp_packet_t is an abstract 64 bit quantity
- Hardware driven packet placement (50Gbps)
    - 100Gbps Netcope and Chelsio PMDs need to copy from native buffers to rte_mbufs
- Silicon specific queue policies, RSS doesn’t help for single tunnel IPsec
    - Hardware maintained packet ordering
- When there is no acceleration
    - All or part of the pipeline is in software and then ODP ~ DPDK

ODP multiple implementations

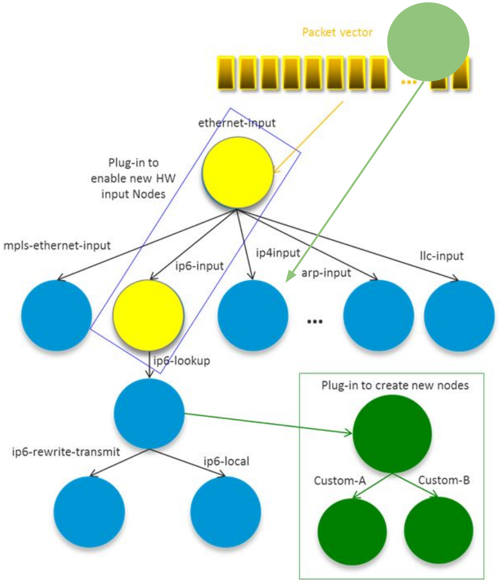
- No execution context constraints or assumptions
  - Do not assume any OS aspects - could be Windows
  - Can be pthreads, processes or any other construct (fibers...)
  - Has been implemented in GPU type processors: Kalray - 256 cores SmartNIC

- Linaro maintained Reference implementation
  - “linux-generic” and validation suite
- Silicon vendor maintained
  - Broadcom, Cavium, Kalray, Marvell, NXP, TI
- “odp-cloud” for hardware agnostic applications

- Unified git repo

How VPP and ODP fit together?

- ODP can be either event driven or poll mode
- VPP has its own software packet metadata
  - No need to maintain a copy in packet-input node -> save L3 cache space @100Gbps...
  - VPP on SoC specific ODP implementation: software metadata can be empty
  - VPP on ODP-CLOUD: software metadata is one cacheline
- Graph shortcuts



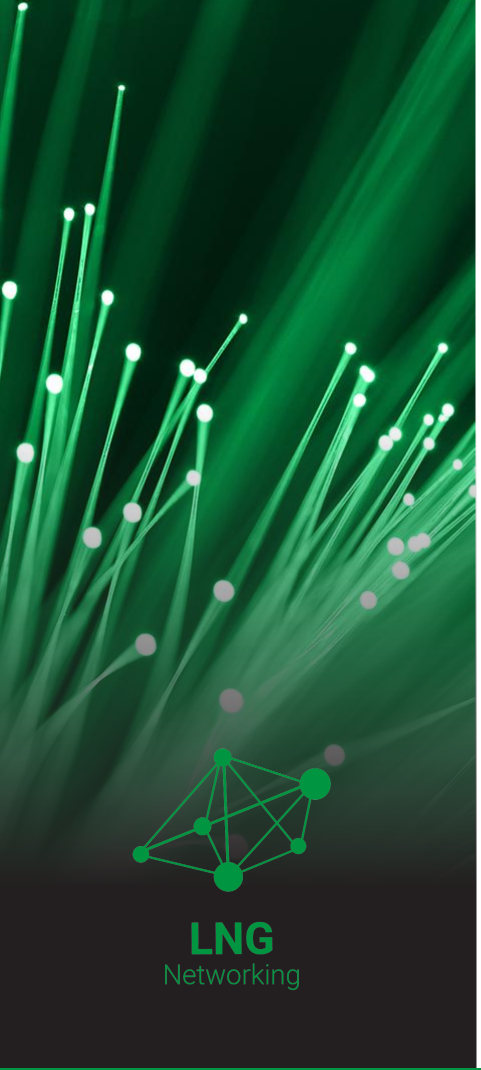
odp-input to directly feed ip4-input with decrypted IPsec packets

Deploying VPP with ODP

- Will be part of Linaro Enterprise Platform 1Q’18
- Available on both ARM and Intel environments
- Cloud/NFV infrastructure
  - Can leveraged pre-installed silicon specific ODP or packaged ODP-CLOUD
- In a VM/Container/VNF
  - Implement a PE router for instance
- In a SmartNIC
  - Honeycomb agent can be on the host (need additional development) or on the SmartNIC
  - Host can be Intel, ARM or IBM Power

VPP on SmartNIC

- Probably a cluster with host (east-west traffic)
- Dedicating CPU and memory to activities
  - The most scarce resource is memory, adding cores to the job may not be enough
  - Stateful firewall need a large memory and its function can impede VM activities
  - VPNaaS: IPsec offload is extremely efficient
  - DVR: some operators have deployed BGP very low and a 600K routing table occupies a few GB
- PLUGABBLE intelligence for more agility
  - Development cycle for server if much larger than SmartNICs
  - Changing SmartNICs sourcing may be more frequent
- Hypervisor domain “sanctuarization”



Agenda

What OpenDataPlane brings to VPP

direct-virtio

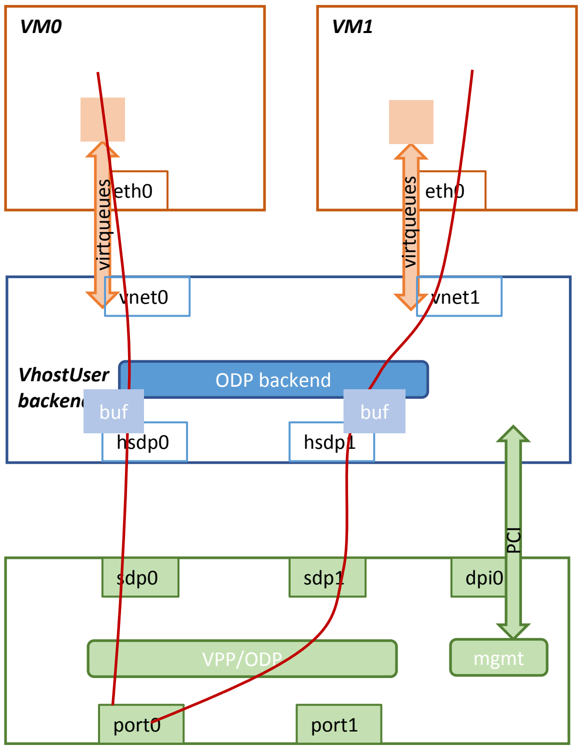
odp-net

odpCL

Direct-virtio highlights

- Boot
  - SmartNIC boots VPP
  - vhostuser backend initializes SmartNIC communications
- Stage 1
  - vhostuser backend negotiate VNICs operations (buffers, capabilities...) and informs SmartNIC about VNIC existence
  - VM can be live migrated
- Stage 2
  - Vhostuser passes buffer information to SmartNIC
  - SmartNIC can position packets directly into VM memory space
  - Can be downgraded to stage 1 for live migration

Stage 1: host relays traffic



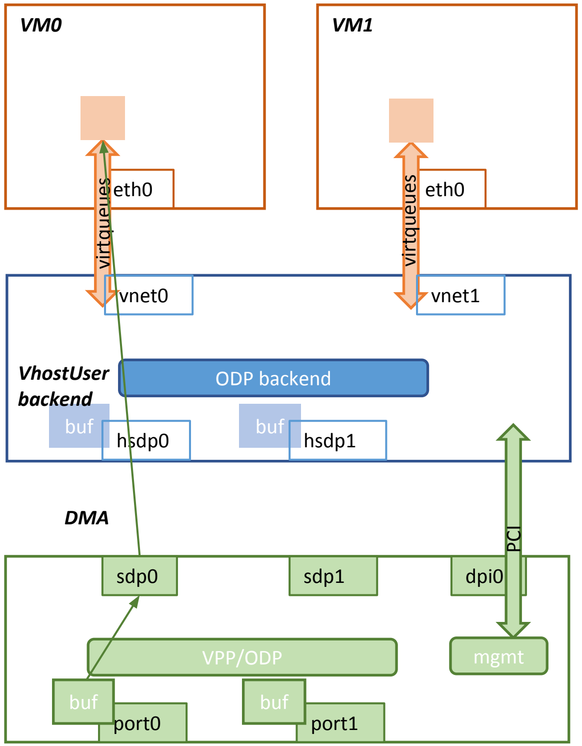
VMs: just use virtio-net, have no clue what’s behind

Host: pass relevant information from VM VNICS to SmartNIC (so that there is one “sdp/hsdp” pair per vnic) and relays to ports in a “macvtap” way

Packets are received in blue buffers (host memory) and copied to orange ones (VM memory).

SmartNIC:
does its work and forwards relevant/modified traffic from network ports to data plane “virtual ports” implemented on PCIe root complex
Configured by Honyecomb residing in the SmartNIC

Stage 2: SmartNIC position packets directly in VM



VMs: just use virtio-net, have no clue what’s behind

Host: pass orange buffers to SmartNIC through their host memory mapping.
Packets are received directly in orange buffers
May forward virtio “kick” from SmartNIC to VM if required

SmartNIC:
Associated orange buffers to each “virtual port” on PCIe side
Call

Direct-virtio buffer management

- Problem with “available” and “used” chains
  - Not hardware friendly, best if same cache hierarchy
- Build on ETSI NFV IFA002 Extensible Paravirtualized Device:
  - Virtio + PCI Virtual Function + plugin

- ODP(/DPDK?) virtio-net driver probes for direct-virtio control PCI VF
  - SmartNIC creates one PCI VF per VM: 256 VM max, thousands of interfaces
  - Each device controls all virtio interfaces and queues of the VM (same approach as Mellanox or Chelsio)
- Stage 2 buffer management
  - Buffers and allocated and controlled through the direct-virtio VF
  - 100Gbps capable buffers: packets are placed by SmartNIC to coalesce many packets per DMA transaction

LNG
Networking

Agenda

What OpenDataPlane brings to VPP

direct-virtio

odp-net

odpCL

odp-net highlights

- ABSTRACT PCI device exposed by an ODP enabled SmartNIC
    - Single device driver
    - Visible in a bare metal environment or instantiated on a virtio bus
    - Very rich control of offloads
- Can be thought of ODP API RPC over PCIe
    - Remember, ODP API is software defined
- Standardized packet format (direct-virtio) optimized for 100Gbps line rate

Agenda

What OpenDataPlane brings to VPP

direct-virtio

odp-net

odpCL

odpCL highlights

- OpenCL allows rich multicore applications collaborate with highly parallel tasks executed on a separate execution environment
- Writing applications that can partly run on a host with tens of cores and a SmartNIC with tens of cores can become very complex
- odpCL would try to leverage tool chain constructs to build, debug and maintain those clustered applications, leveraging odp-net as a communication bearer

Linaro

Thank You

For further information: www.linaro.org

LNG
Networking