Developing OpenStack as a Framework for NFV

Alan Kavanagh (Ericsson)
Adrian Hoban (Intel)
Steven Gordon (Red Hat)

*OpenStack is a registered trademark of the OpenStack Foundation*
AN NFV Transformation has started

Yesterday
- Optimized VNF configuration Build
  - APP
  - H/W

Today
- APP
- OpenStack
- Virtualization Layer
- H/W

NFV | Openstack | ODL Transformation

- OSS/BSS
- PaaS
- Cloud IaaS Mgmt
- Virtualisation Layer
- Physical Infrastructure

Yesterday's services: Firewall, BRAS
Today's services: Optimized VNF configuration Build, APP, OpenStack, Virtualization Layer, H/W
Provision & Configuration of VNF’s

VNF

APP
APP
APP
Network
Compute
Storage

OVF

Service Catalog

VNF
NSD

Attributes for NFV config and expose via API

PaaS
Policy/Security/Governance

Openstack NBI API’s

Neutron
Nova
Cinder

Glance
Swift

VNF Additional Requirements

VLAN Trunk
Security
PCI Device Capability
CPU Pinning
Huge Page Tables
NUMA Topology
Local storage
Network Anti-Affinity
Firmware validation
QoS
CPU Pinning
Huge Page Tables
NUMA Topology
Local storage
Network Anti-Affinity
Firmware validation
QoS
Scheduling is ... placing an application in the Cloud based on the constraints to handle my app to meet my SLA.

Information Exposed to make “Intelligent fine grain placement decisions”

- Onboard Storage
- PCI_Device_Type
- Network Based Anti-Affinity
- High Performance vSwitch
- Link_Type
- CPU Pinning
- NUMA Awareness
- Feature .....
Extensions for NFV Data Planes

- DPDK
- PCIe SR-IOV
- NUMA
- Huge Pages
- Accelerated vSwitches

*OpenStack is a registered trademark of the OpenStack Foundation*
Data Plane Development Kit (www.dpdk.org)

Optimized software libraries and drivers for accelerating packet processing

1: Intel internal estimate
2: Intel Internal measurement of packet processing performance using Intel Xeon processors. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to intel.com/performance
Single Root I/O Virtualisation

PCI-SIG Single Root I/O Virtualization (SR-IOV)

- Allows partitioning of a single Ethernet Server Adapter port into multiple Virtual Functions.
- Virtual Functions may be allocated to VMs each with their own bandwidth allocation.
- They offer a high performance, low latency path for data packets to get into the VM.

Physical Functions (PF):

- This is a full PCIe function that includes the SR-IOV Extended Capability (used to configure and manage the SR-IOV functionality).

Virtual Function (VF):

- This is a light weight PCIe function that contains the resources necessary for data movement (but minimizes the set of configuration resources).
OpenStack PCIe SR-IOV Support

- Juno extended with support for NIC based PCIe SR-IOV allocation

Allocate VF's

Configure whitelist in nova.conf per node

Configure SR-IOV mech driver with VF vendor and product ID's

Configure Neutron to use SR-IOV mechanism driver

Create Neutron port with vif_type=direct

Nova boot VM with allocated Port ID

SR-IOV facilitates highest performance I/O connectivity to the VM from the NIC
Filter Extensions: CPU Features & PCIe Devices

**compute_capabilities_filter**

- Nova libvirt driver extensions to expose all CPU instruction set based features to scheduler
- Select hosts with required features based on `extra_spec` settings

**pci_passthrough_filter**

- Select hosts with requested PCIe device

---

**Premium Flavors**

<table>
<thead>
<tr>
<th>Cloud Customers</th>
<th>Cloud Providers</th>
<th>Telco Use Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhanced capabilities</td>
<td>Revenue opportunities</td>
<td>Enhanced performance</td>
</tr>
</tbody>
</table>

---

1. Intel® Advanced Encryption Standard New Instructions (Intel® AES-NI)
Non Uniform Memory Architecture (NUMA)

- Performance and latency characteristics differ depending on the core a process is executing on and where the memory a process is accessing is located.

Optimising placement for memory proximity enables greater performance & efficiency
Filter Extensions: NUMA

- Helps to co-locate CPU core allocations to a single socket (when possible)
- Resource tracks core/socket consumption and filters to available subset of suitable platforms.

Co-location helps with cache efficiency for faster inter-process data sharing & communication
Filter Extensions: NUMA

- Helps to co-locate CPU core allocations to a single socket (when possible)
- Resource tracks core/socket consumption and filters to available subset of suitable platforms.

Enables the OSes to allocate local memory for greater performance & efficiency
Huge Page Tables: Patches† for Juno

Translation Lookaside Buffer (TLB)
- Memory component that accelerates address translation.
- Caches a subset of address translations from the page table.

Huge page table sizes (e.g. 1 GB)
- TLB caches a greater range of memory translations
- Helps reduces TLB misses.

Memroy Address Translation Request

TLB
- Check TLB Cache
  - Fetch Page Table from memory
    - If translation not in cache fetch page table from memory

Small Page Entries (4 KB)

Small page table entries (4KB) can result in a greater number of TLB misses

† Note: Huge Page patches available on 01.org
Huge Page Tables: Patches† for Juno

Translation Lookaside Buffer (TLB)
- Memory component that accelerates address translation.
- Caches a subset of address translations from the page table.

Huge page table sizes (e.g. 1 GB)
- TLB caches a greater range of memory translations
- Helps reduces TLB misses.

Memory Address Translation Request

TLB
- Check TLB Cache

Fetch Page Table from memory
- If translation not in cache fetch page table from memory

Huge Page Entries (1GB)
- More memory space covered by TLB cache

Configuration of Huge Page (1GB) based memory drives performance & efficiency gains

† Note: Huge Page patches available on 01.org
DPDK Based vSwitches: Patches† for Juno

Open vSwitch

DPDK based vSwitches

OpenStack needs to be able to configure user space vhost methods & patch ports to enable these vSwitches

Nova API

Libvirt Driver

VIF Binding & Huge Page Patches

Neutron API

Neutron-ML2

Open vSwitch Mechanism Driver

External Controller

DPDK based vSwitch / Ericsson vSwitch / Other

Open vSwitch L2 Agent (Patched)

Configuration of DPDK based vSwitches drives performance & efficiency gains

† Note: DPDK accelerated vSwitches patches available on 01.org
Where are we going with this...?
Where are we going with this...?

- NFV requires:
  - Open, standard, APIs to provision and manage VNFs.
  - Performance, determinism, and reliability features not yet present in OpenStack.

- NFV shares:
  - A common desire for simplicity, agility and scale of implementation.
Succeeding requires

• Bringing together:
  • Communications Service Providers
  • Network Equipment Providers
  • OpenStack Vendors
  • OpenStack Developers
  • Industry standardization efforts (ETSI NFV, OPNFV)

• Striking balance between functionality exposure and simplicity.
• Improved functional testing to offset integration testing challenges (e.g. NUMA, SR-IOV).
Succeeding requires

• Working group formed at Atlanta summit.
• Meets weekly on IRC:
  • [http://openstack.org/wiki/Meetings/NFV](http://openstack.org/wiki/Meetings/NFV)

• Current mission:
  • Identify use cases
  • Define and prioritize requirements
  • Create design blueprints
  • Submit and review patches with the help of the relevant OpenStack projects and subteams
Juno status

• Improved SR-IOV support
• Multiple vNICs on the same network
• Evacuate instance to a scheduled host
• Libvirt driver guest vCPU topology
• Libvirt driver NUMA topology awareness (partial implementation)
Tentative Kilo goals

- VLAN trunking into a virtual machine
- Permit unaddressed interfaces and/or interfaces without security groups
- Continuation of NUMA-awareness work:
  - Memory (incl. large pages)
  - I/O device locality
- vCPU pinning
- Userspace vhost support
- Unified, pluggable, scheduler
- Configurable MTU
- Port mirroring
Get involved...

• Collaborate on use cases, requirements, and solutions for Kilo and beyond:
• Thursday, November 6 @ 09:00 – 10:30
• Hyatt Hotel - Batignolles
Summary

- Network Transformation is happening right now
- Incremental requirements for OpenStack APIs
  - Additional attributes for service exposure
  - Policy & SLA based provisioning for the application
  - Fine grained placement
  - Policy Control and Enforcement
  - Unified Scheduling across compute, network & storage resources
  - Enhanced Platform Awareness for Performance & Determinism
  - Additional tuning knobs to accurately instantiate VNFs
- Several communities need to collaborate openly
Backup
A “Mission Critical Application” is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL’S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS’ FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked “reserved” or “undefined”. Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families: Go to: Learn About Intel® Processor Numbers

Intel, the Intel logo, Itanium, Intel Atom, Intel Xeon Phi, Intel AppUp, and Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.

Copyright © 2014 Intel Corporation. All rights reserved