SR-IOV connectivity in an SDN overlay

Industry Situation/Customer Problem

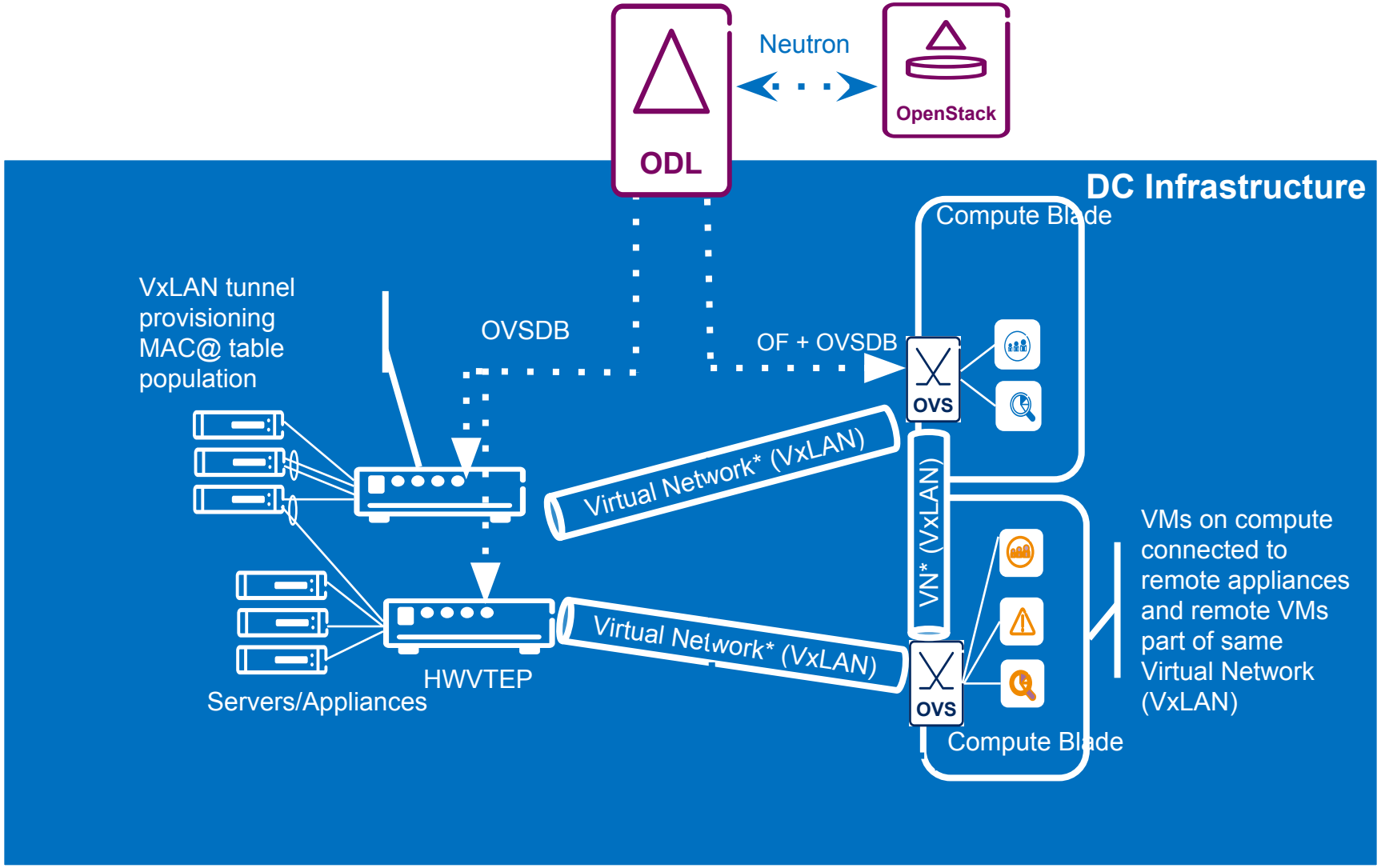
- SR-IOV and PCI-PT technology is the preferred choice for high performance NFV workloads
- Database and other high performance applications are being run on non-virtualized bare metal servers in the data center
- At the same time, a VxLAN based overlay controlled by a central SDN controller is the proven and well-adopted cloud networking deployment model
- There is a need to provide seamless connectivity between bare metal workloads, legacy VLAN domains, SR-IOV enabled workloads and virtualized workloads in the cloud environment

SR-IOV connectivity in an SDN overlay

Technical Opportunity

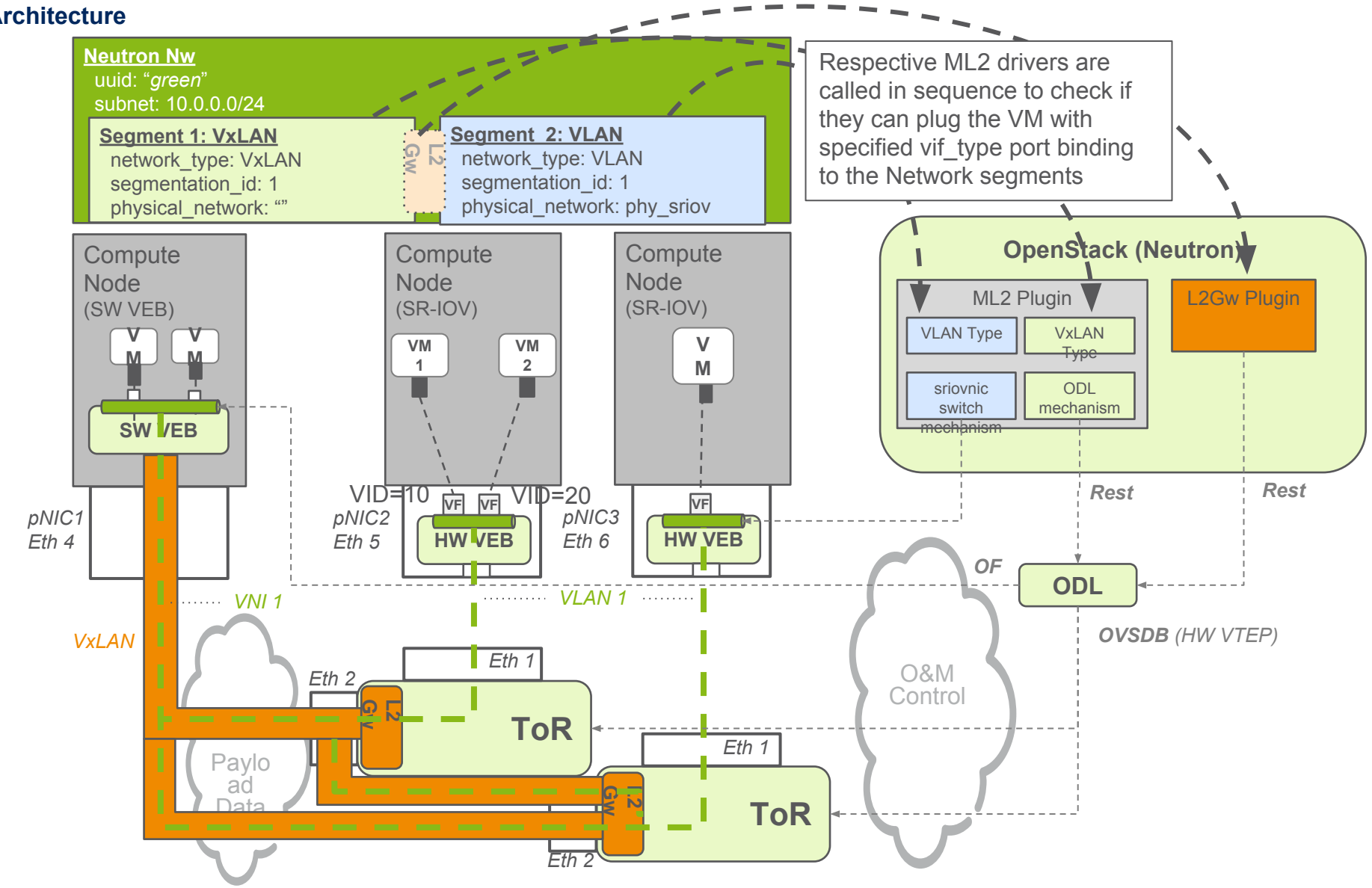
- These issues can be solved by deploying VLAN gateway functionality on TOR VTEP devices connected to such domains
- Ericsson has lead the effort in enabling configuration of HWVTEPs and seamless integration with features in the Netvirt project
- Orchestration in an Openstack environment via Neutron multi-segment networks and the L2GW service plugin
- Southbound configuration implemented using a hwvtep plugin which implements the OVSDB protocol and hardware_vtep schema

Architecture



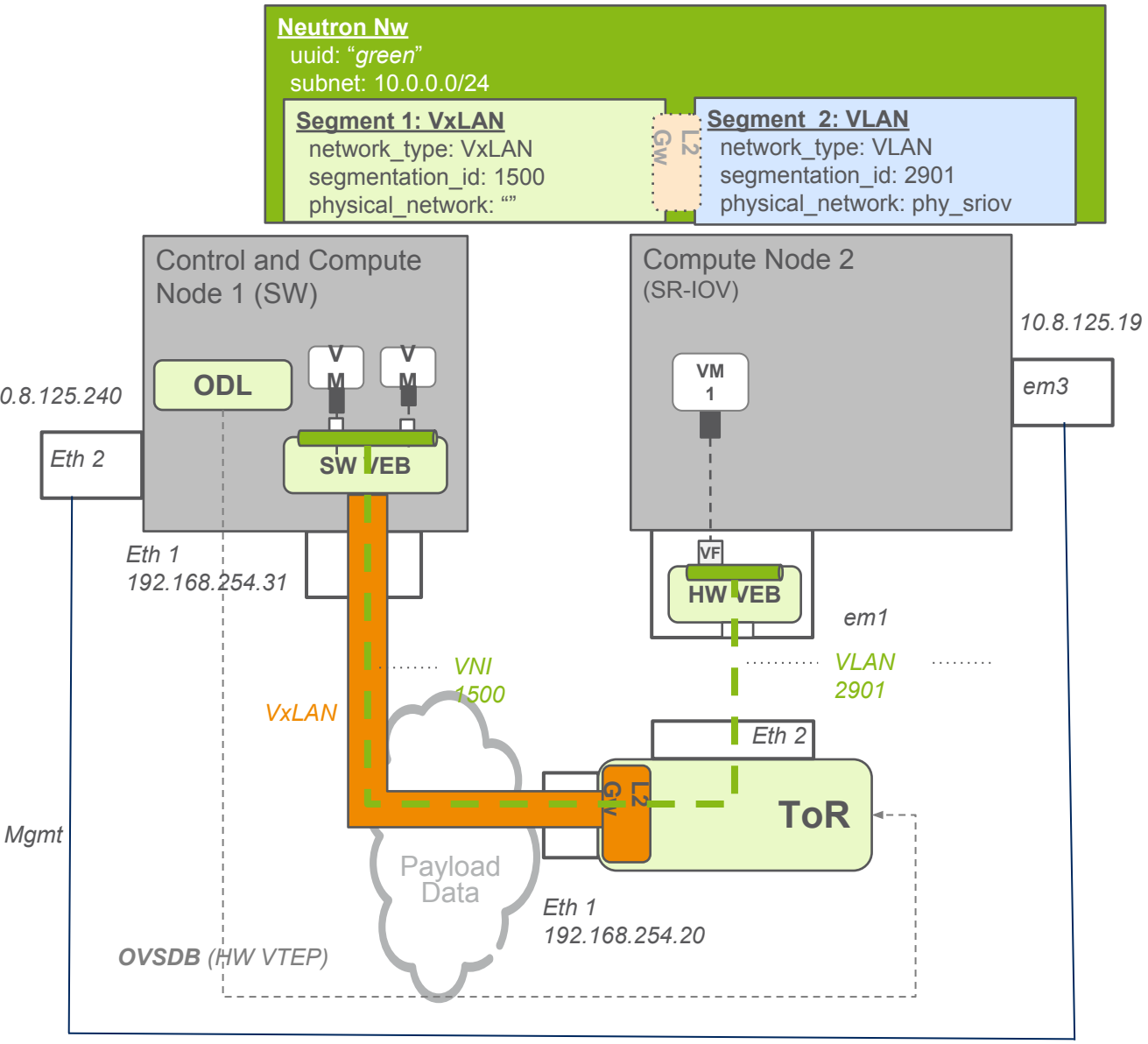
SR-IOV connectivity in an SDN overlay

Architecture



SR-IOV lab setup

Architecture



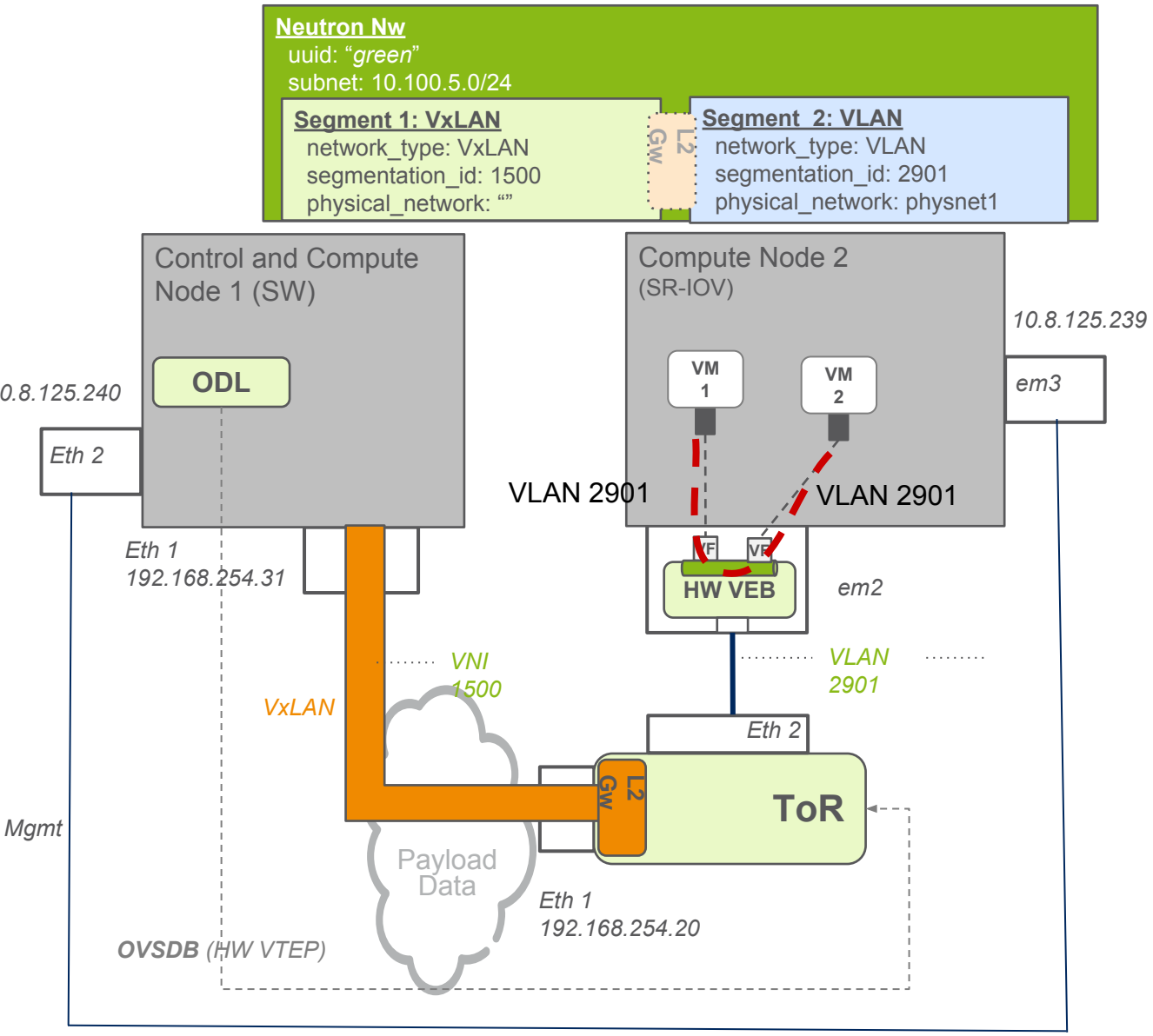
Server 1
10.8.125.18
CentOS7
Openstack Control/Compute
ODL
HWVTEP Emulator VM

VM1
10.100.5.3

Server 2
10.8.125.19
CentOS7
Compute node
SR-IOV NIC em1
TAG = 2901

Use Case 1 - Two VMs on one SRIOV Compute Node

Architecture



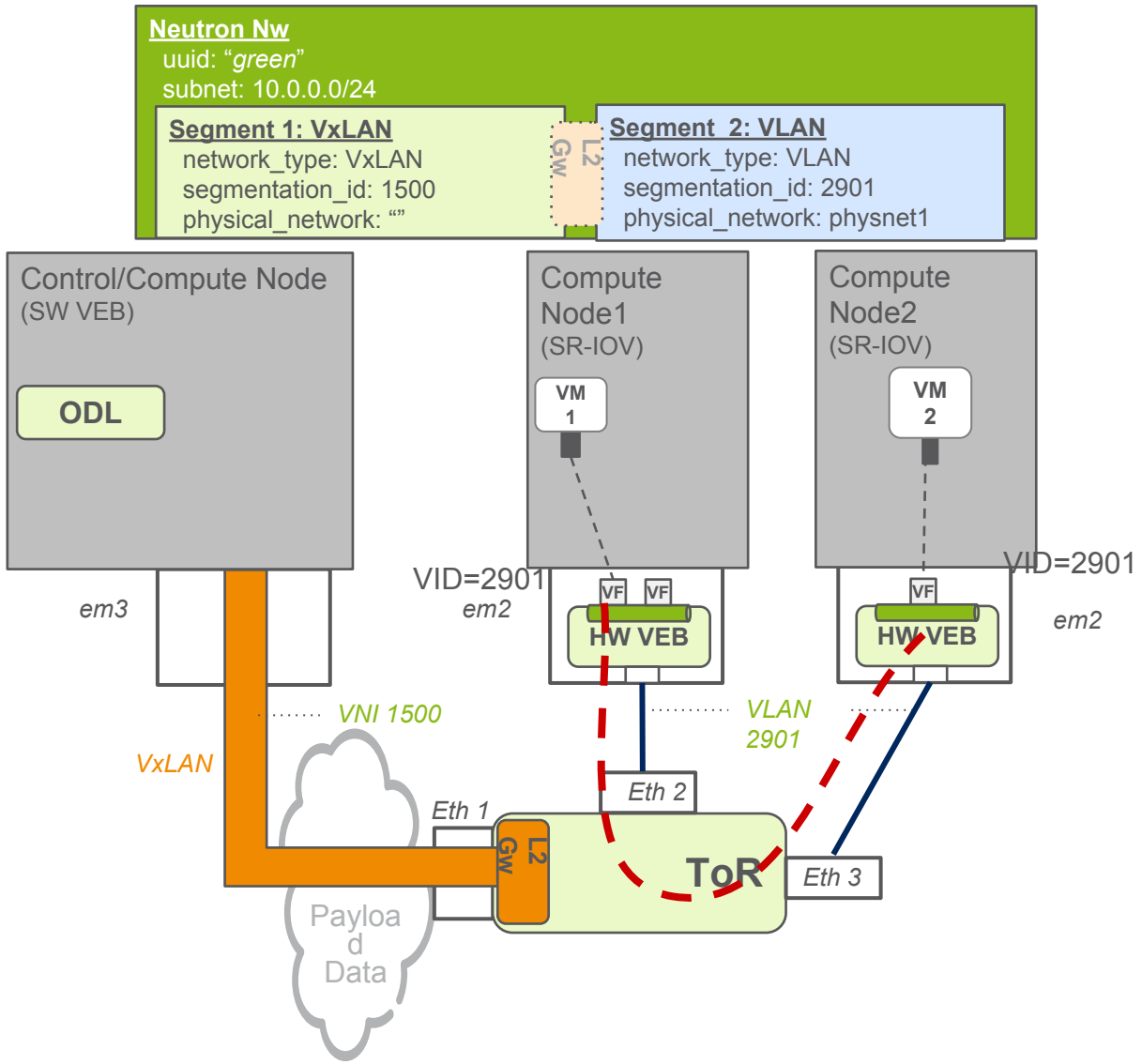
Server 1
10.8.125.18
CentOS7
Openstack Control/Compute
ODL

HWVTEP Emulator VM
eth2=>em3

Server 2
10.8.125.19
CentOS7
SR-IOV NIC em1
TAG = 2901

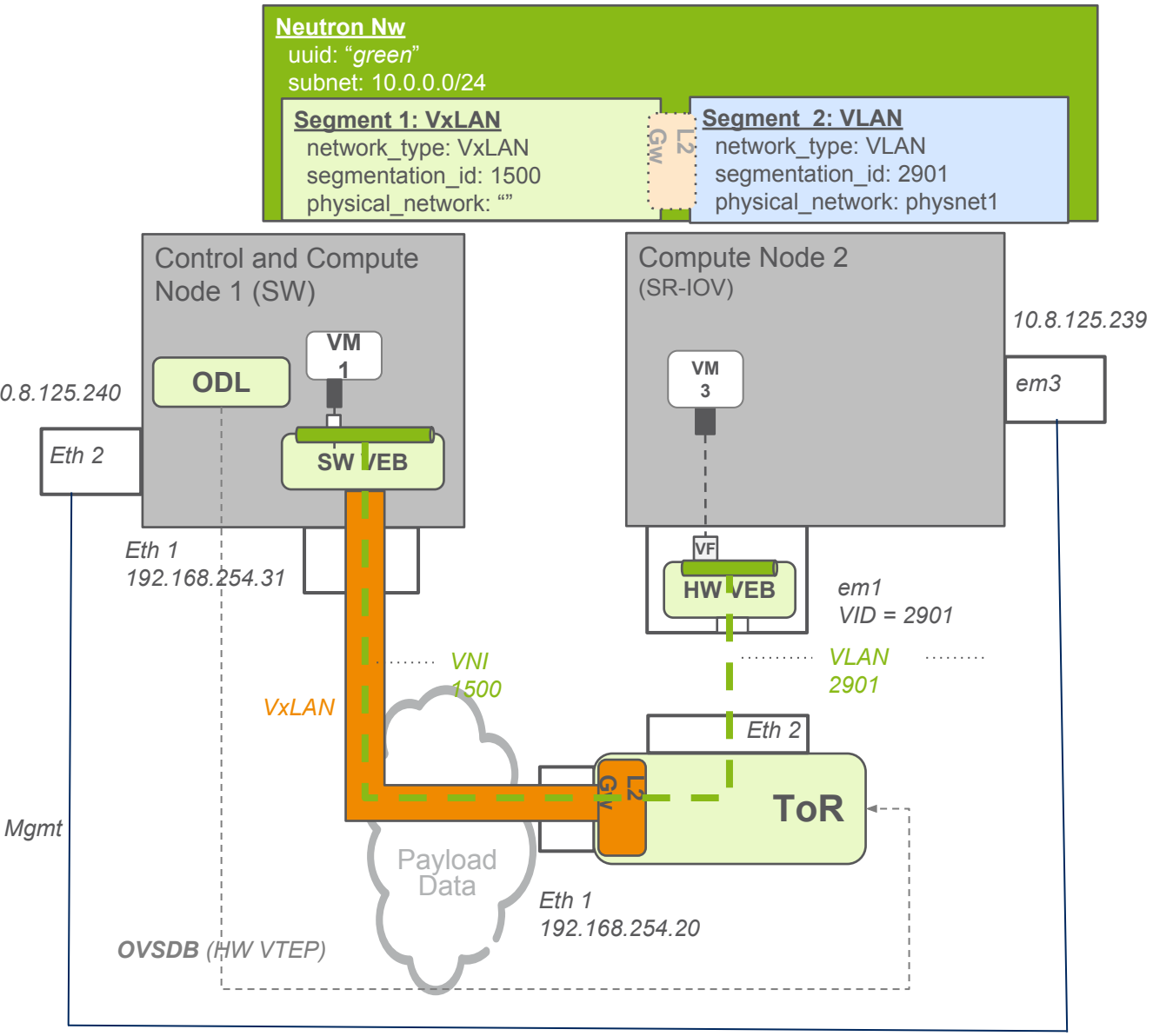
Use Case 2 - VM on separate compute nodes

Architecture



Use Case 3 - VM on control (vxlan) ping VM on SRIOV

Architecture

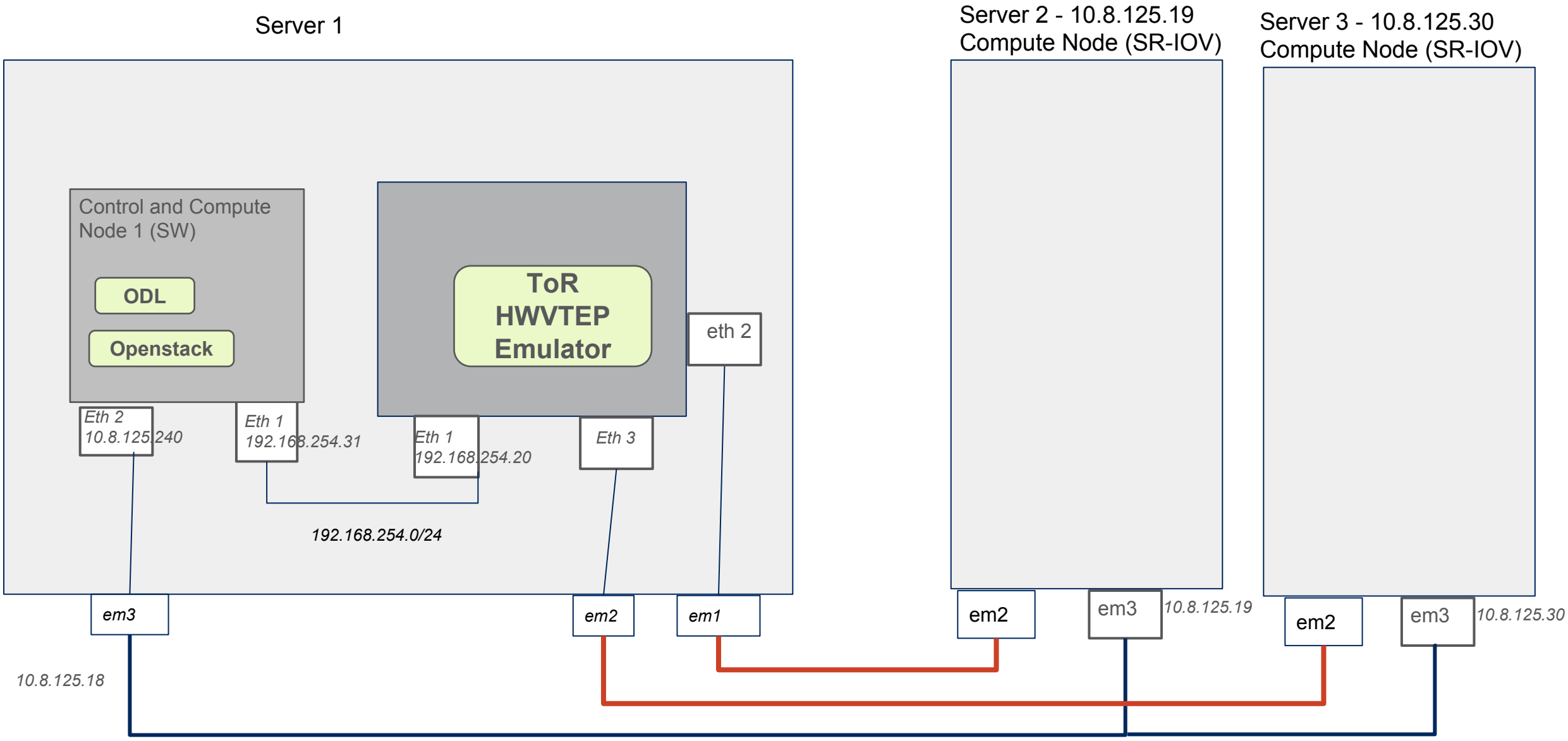


Server 1
10.8.125.18
CentOS7
Openstack Control/Compute
ODL

HWVTEP Emulator VM
eth2=>em3

Server 2
10.8.125.15
CentOS7
SR-IOV NIC em1
TAG = 2900

SR-IOV lab setup - Physical



Bugs/Issues

- Netvirt - When ODL DHCP is disabled, add qdhcp-namespace to L2 unknown DMACS BC group so qdhcp-namespace will see DHCP pkt
  - need to confirm this is what is really needed
- os_ssh.sh - can't ssh to host…. need to debug
  - I've been using console to debug…
- ARP issue from VM on control to VM on SRIOV compute
  - ARP request from SRIOV VM received on control node
  - No ARP response - need to debug
- What other multi-provider network bugs may be lurking?

What is SR-IOV?

- SR-IOV: Single Root I/O Virtualization and Sharing
- SR-IOV spec defines extensions to the PCI Express Spec to allow multiple OS’s or VM’s running simultaneously within a single computer to share PCI Express hardware resources (for example, a single Ethernet Port)
- Physical Functions (PFs) : Full PCI e functions that include SR-IOV Extended capability. Used to configure and manage SR-IOV functionality. The physical Ethernet controller that supports SR-IOV
- Virtual Functions (VFs): Lightweight PCI e functions that contain resources necessary for data movement but minimized config functions. Virtual PCIe device created from a PF

Why use SR-IOV?

- Direct communication between VM and device, bypassing hypervisor and virtual switch layer
- Near native I/O performance for each VM on a physical server
- Standard way of sharing capacity of I/O device
- Data protection between VMs on same physical server
  - Independent memory space, interrupts, and DMA streams for each VM

SR-IOV Networking in OpenStack

- Virtual bridge no longer required
- Each SR-IOV port associated to virtual function (VF)
- SR-IOV ports provided by either:
  - hw-based Virtual Ethernet Bridging (HW VEB)
  - Extended to upstream physical switch (IEEE 802.1br)
- SR-IOV ports connected either:
  - directly to its VF
  - with a macvtap device that resides on the host, which is then connected to the corresponding VF
- https://wiki.openstack.org/wiki/SR-IOV-Passthrough-For-Networking
- https://wiki.openstack.org/wiki/SR-IOV-Passthrough-For-Networking-Mitaka-Ethernet

Limitations

- When using QoS, max_burst_kbps is not supported.
- Max_kbps is rounded to Mbps
- No support for security groups, so firewall driver must be disabled
- Not integrated into horizon, must use CLI or API to configure SR-IOV
- Live migration not supported with SR-IOV ports

Open Questions

- How does ODL know connectivity between servers and TOR ports?
  - Assumption:
    - currently uses REST API to statically configure
      - How does this REST API work?
    - future: need some type of discovery mechanism.
      - possibility: LLDP-based discovery.
- How do you configure VLAN-based SR-IOV virtual functions to support multiple VMs on same server?
- Can more than one VM on a given server be on the same network?

Acknowledgements

Thanks Daya for Architecture slides!

References

- http://www.intel.com/content/www/us/en/pci-express/pci-sig-single-root-io-virtualization-support-in-virtualization-technology-for-connectivity-paper.html
- http://www.intel.com/content/www/us/en/pci-express/pci-sig-sr-iov-primer-sr-iov-technology-paper.html
- http://www.netdevconf.org/1.1/proceedings/slides/duyck-sr-iov-openstack.pdf

SR-IOV lab setup - Physical