Optimizing and extending overlay networking for containers

October 28, 2015

Ton Ngo, Mohammad Banikazemi,
Baohua Yang, Simeon Monov

IBM Cloud

Outline

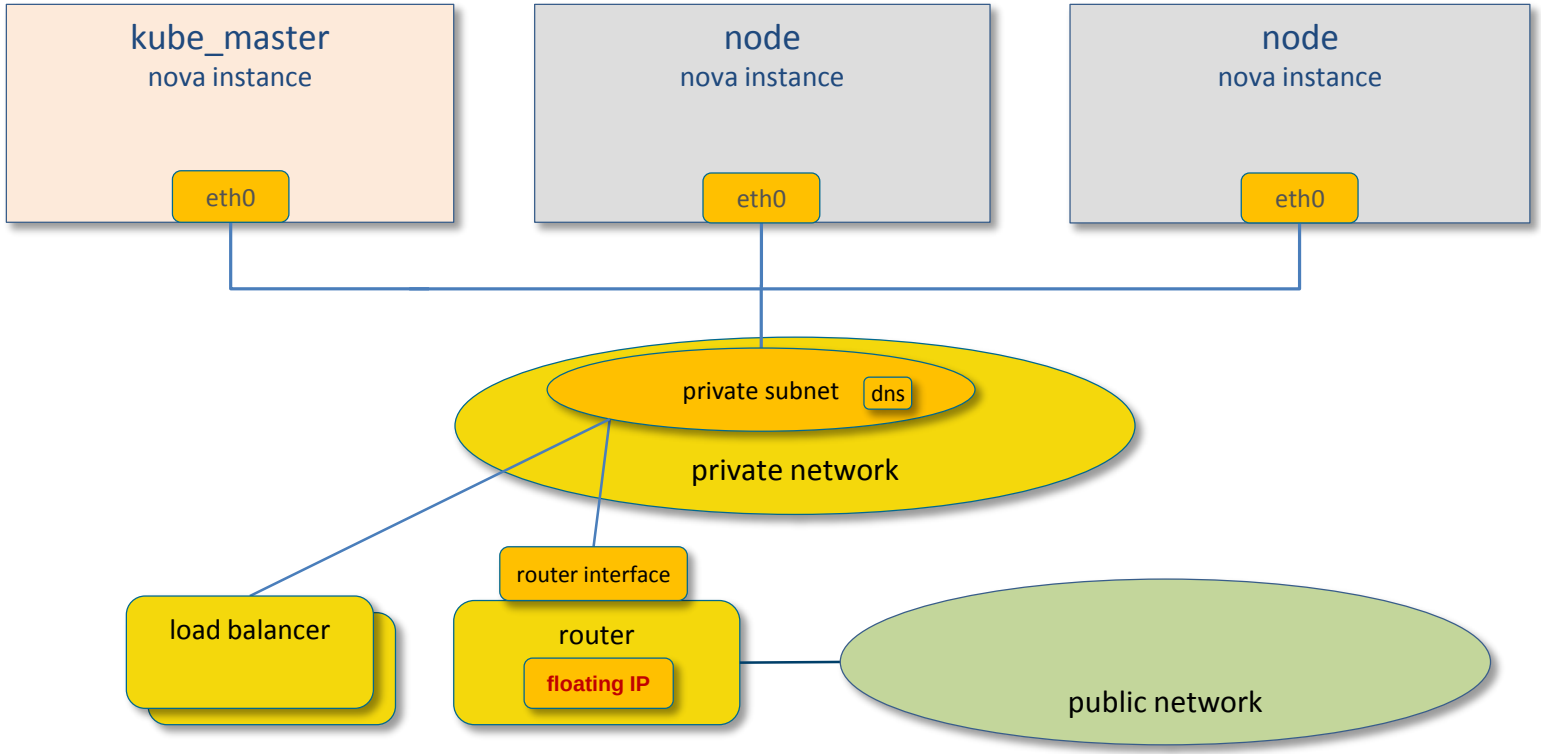
- Container Networking
- Performance Evaluation and Observations
- Opportunities and Future Directions
- Q&A

Container Networking

Background

- Container: abstraction for process
  - Manage IP instead of port
  - Overlay network
- OpenStack context
  - Magnum: container as a service
  - Neutron networking
- Goals
  - Investigate a specific scenario: Kubernetes
  - Performance Evaluation
  - Identify areas for improvement

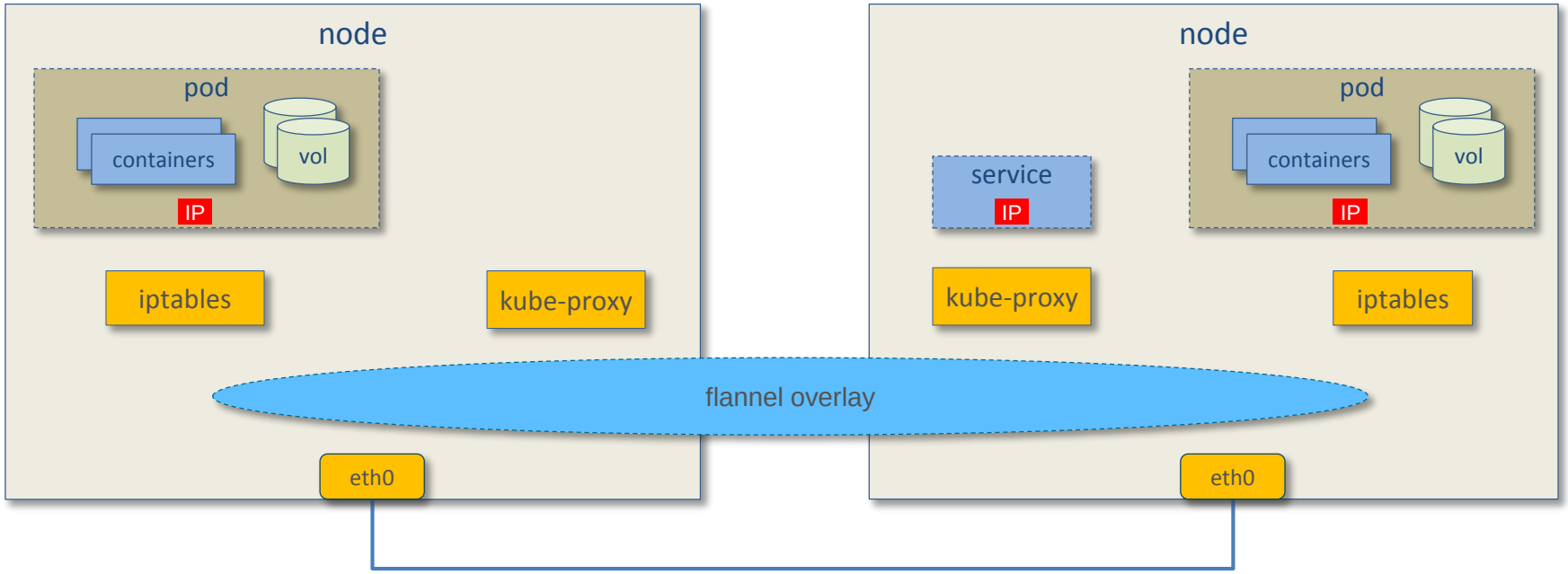
Kubernetes Cluster in OpenStack



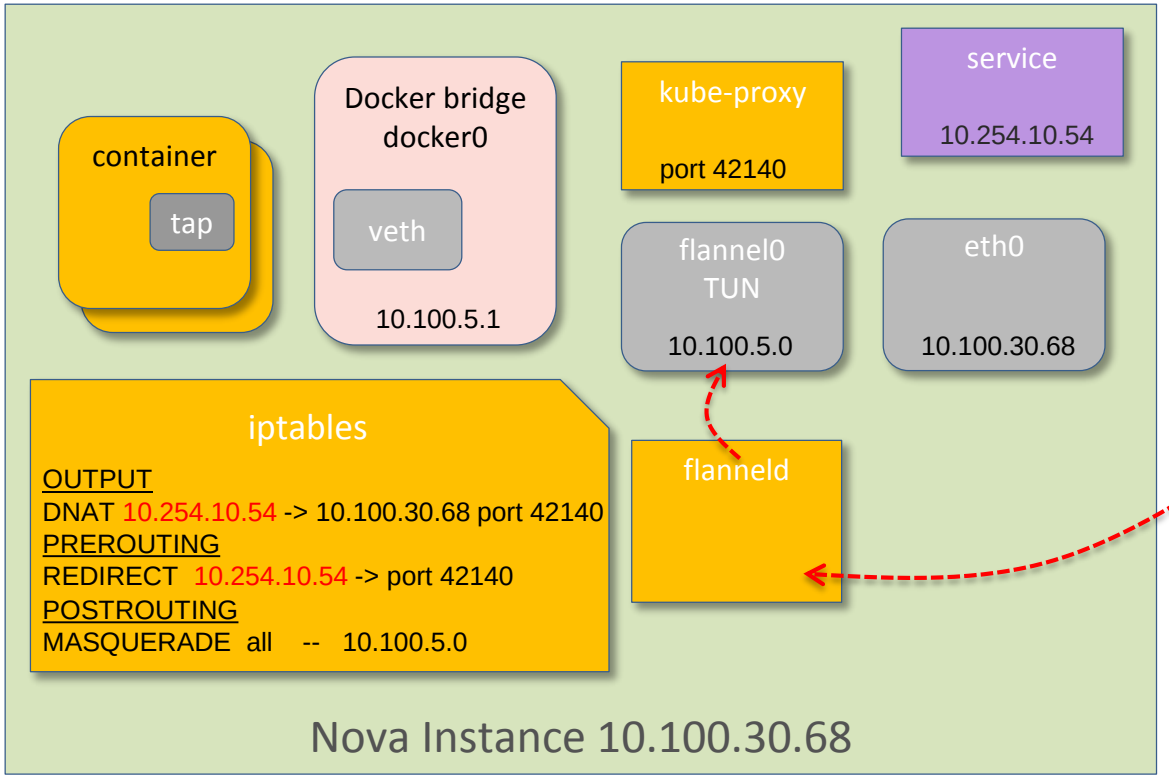
Kubernetes Concepts

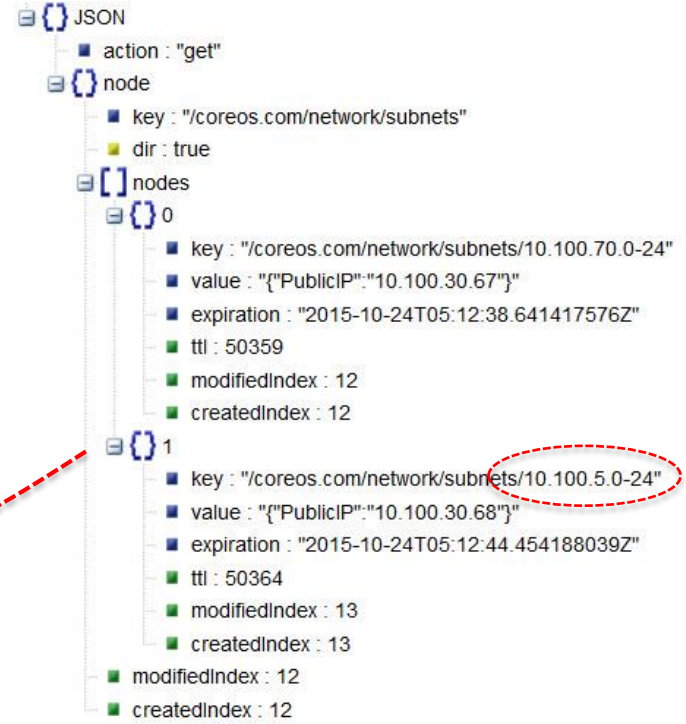
- **Abstractions**
  - pods:
    - group of containers on same host
    - IP per pod
  - service:
    - proxy, load balancing
    - IP per service
  - replication controller: maintain exact number of pods
- **Networking support**
  - kube-proxy: a Kubernetes component
  - flannel: an overlay network (other options available)
  - iptables rules: kernel support

Kubernetes Cluster in Operation



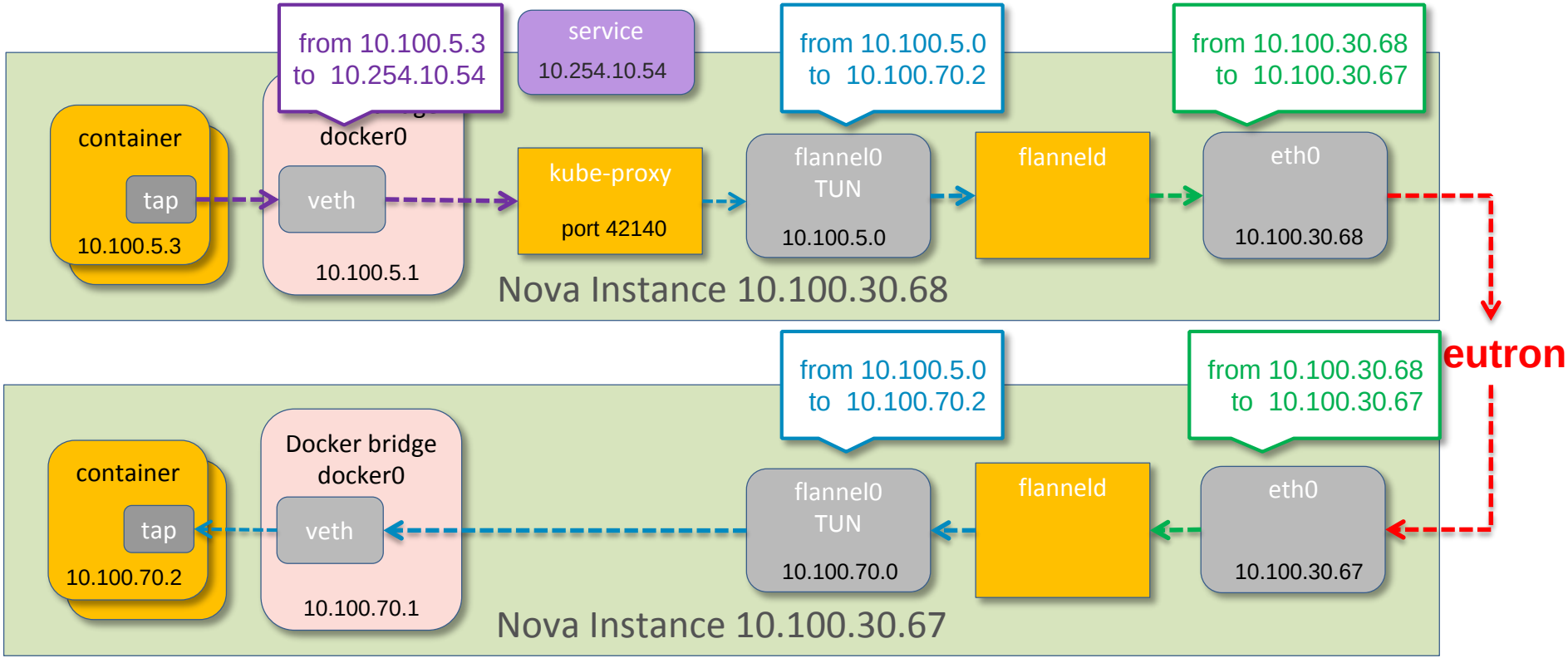
Networking Setup



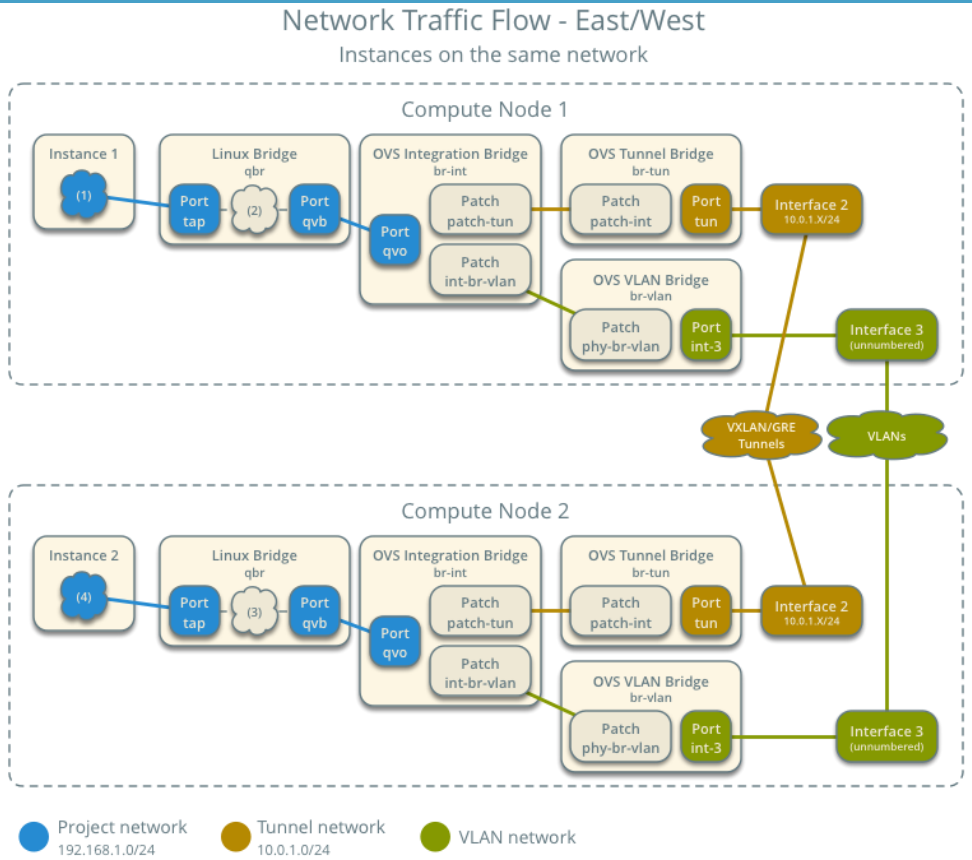


etcd

Container to Container Communication

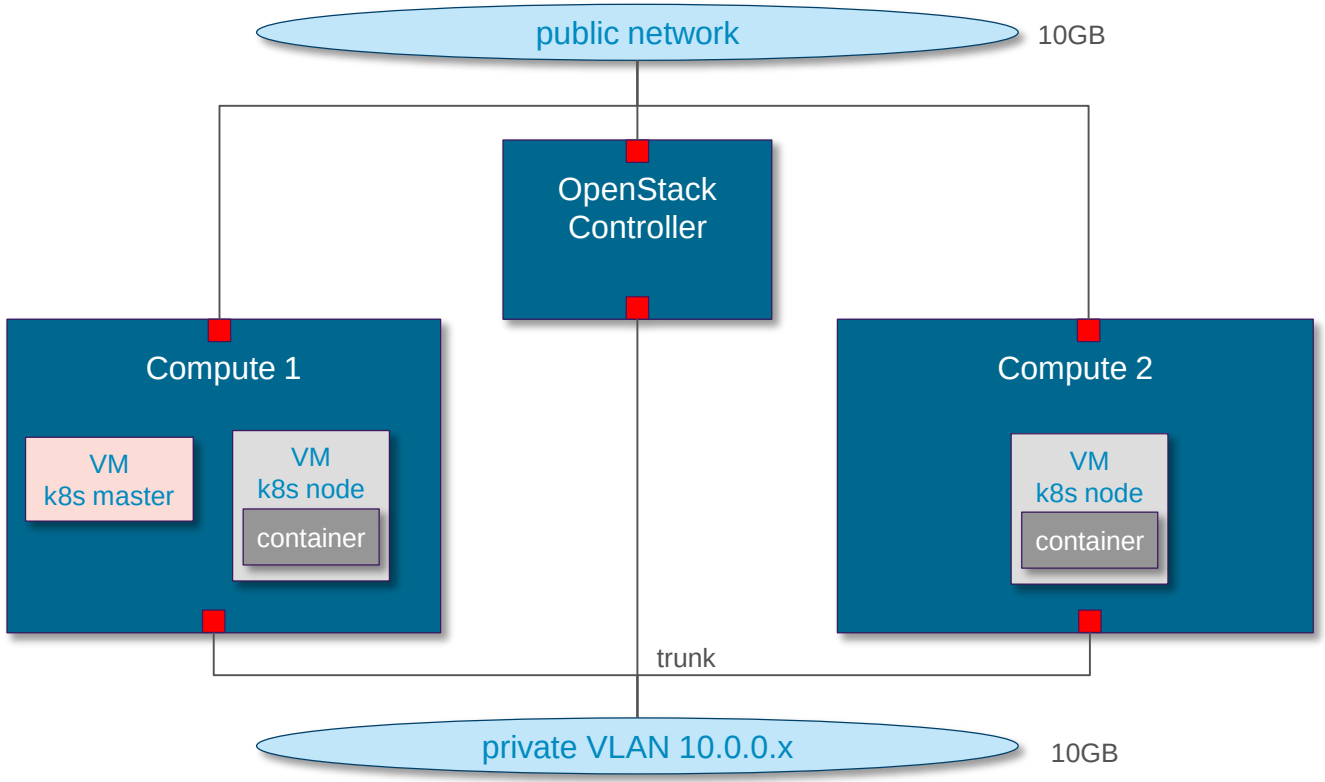


Neutron Path (from the Neutron Document)



Performance & Observation

Test Environment



X3650M4
2x E5-2680v2, 10 cores each
256gb memory
4x 900gb SAS drives RAID 10

Scenarios Considered

- Traffic Path
  - Server to Server
  - VM to VM
  - Container to Container
    - Same pod
    - Different pods on same host
    - Different pods on different hosts

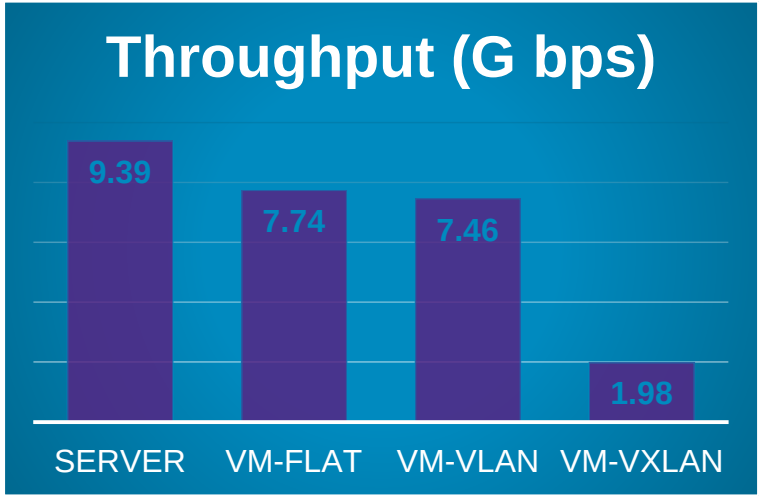
- Neutron Implementation
  - Flat
  - VLAN
  - Overlay

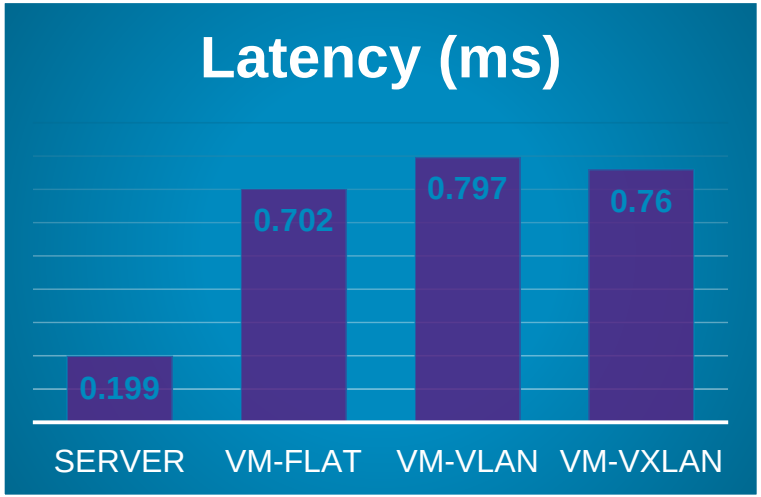
Throughput!

Latency!

1. VM Utilizes Single Overlay

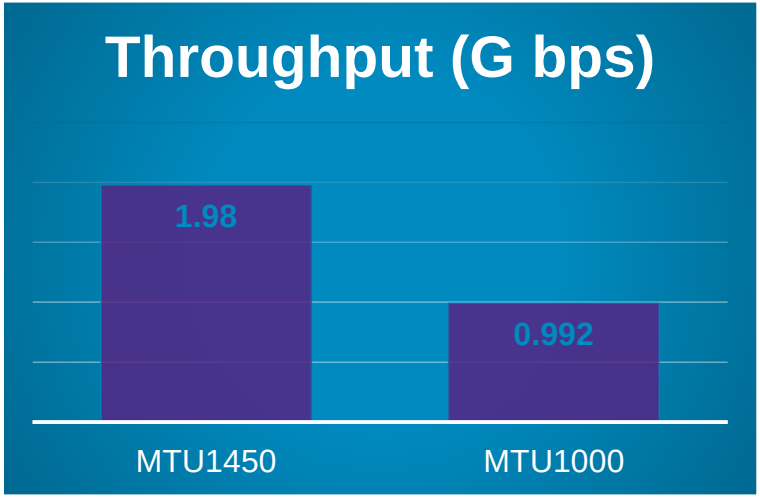
- Server, VM, and Overlay
  - From server to vm-flat: bw to 82%, latency to 350%
  - From vm-flat to vm-vlan: bw to 96%, latency to 114%
  - From vm-flat to vm-overlay: bw to 26%, latency to 108%





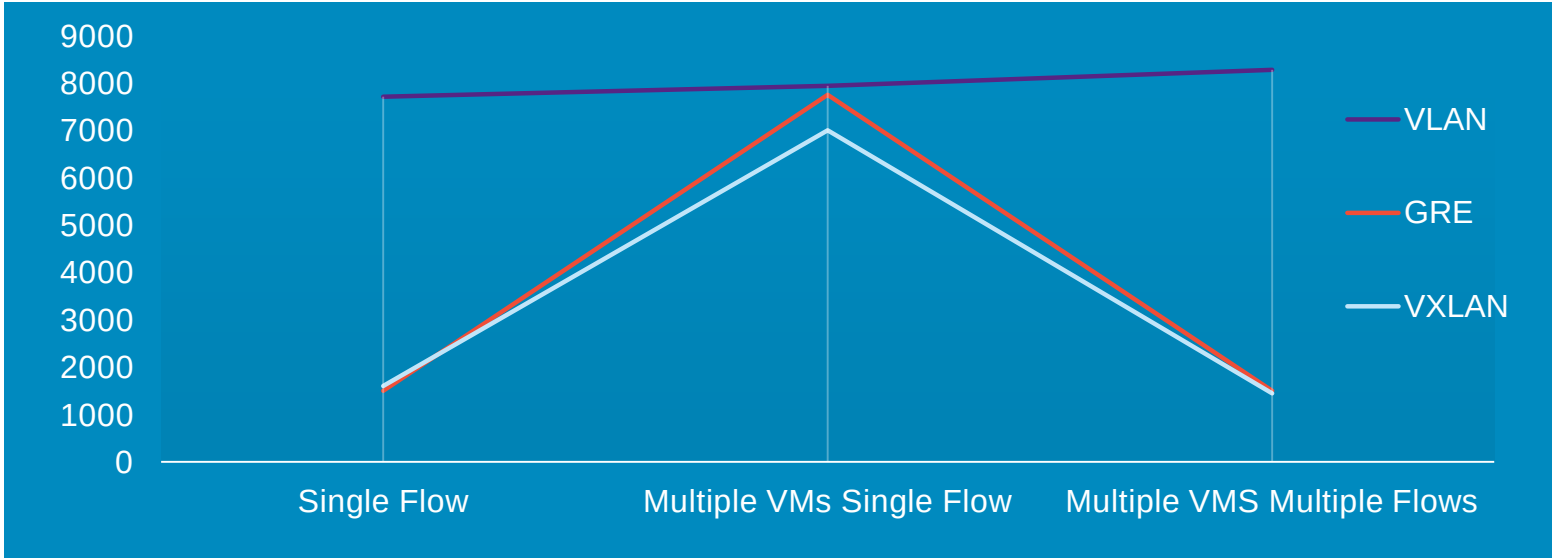
2. Single Overlay Bottleneck is Packet Processing

- VM to VM over VXLAN data
  - Changing MTU from 1450 to 1000, the bw will decrease to 50%.
  - *Hardware offloading could be adopted in physical host.*



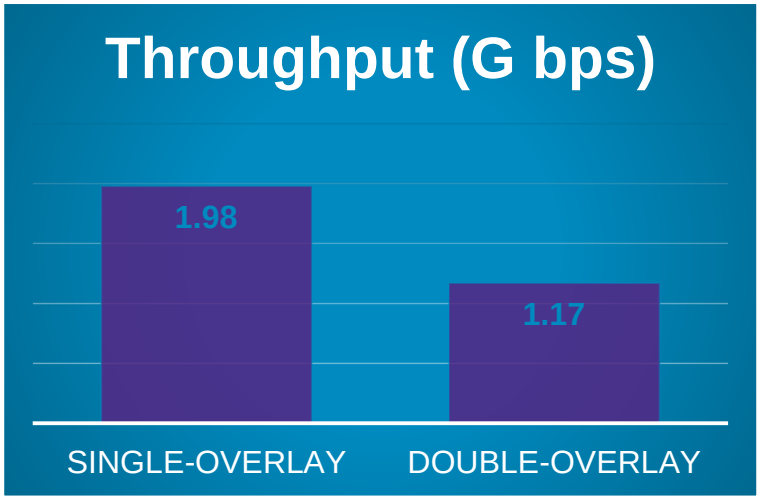
3. Real Cloud Scenarios Have Concurrent Flows

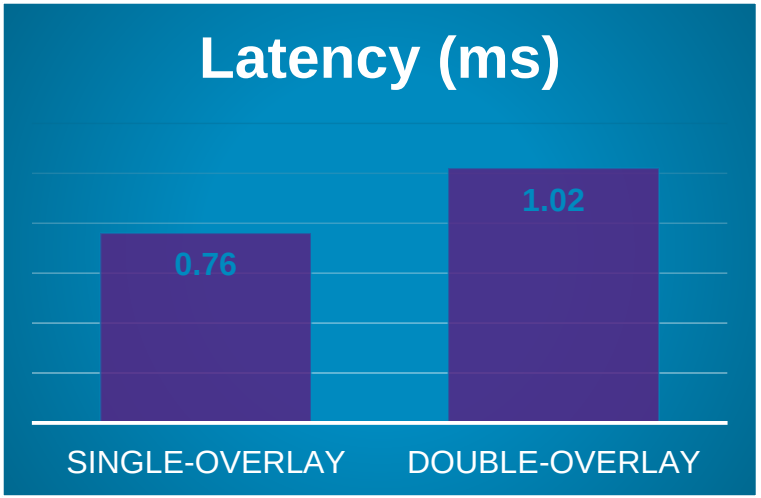
- VXLAN traffic with multiple streams (from OpenStack Summit 2015 Vancouver)
- *Do our applications actually need high throughput single-flow?*



4. Container Utilizes Double Overlay

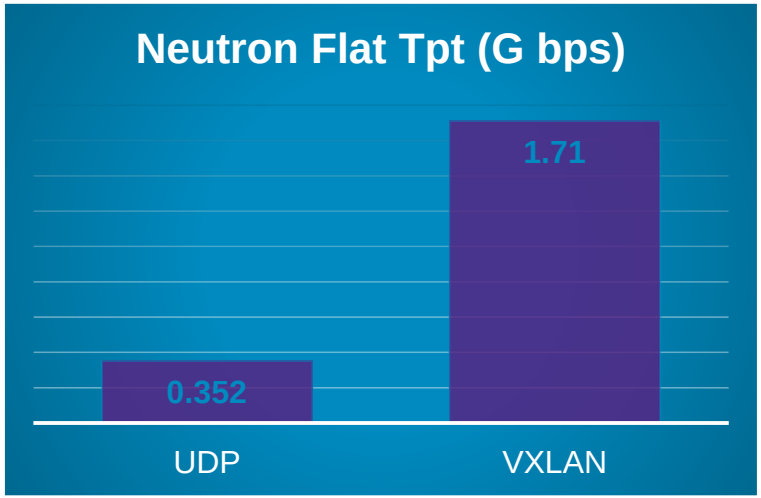
- Compare double overlay (Pod-Pod on Flannel) with single overlay (VM-to-VM on VXLAN)
  - Throughput drops 41% (compare with 74% drop)
  - Latency increases 34% (compare with 250% increasing)

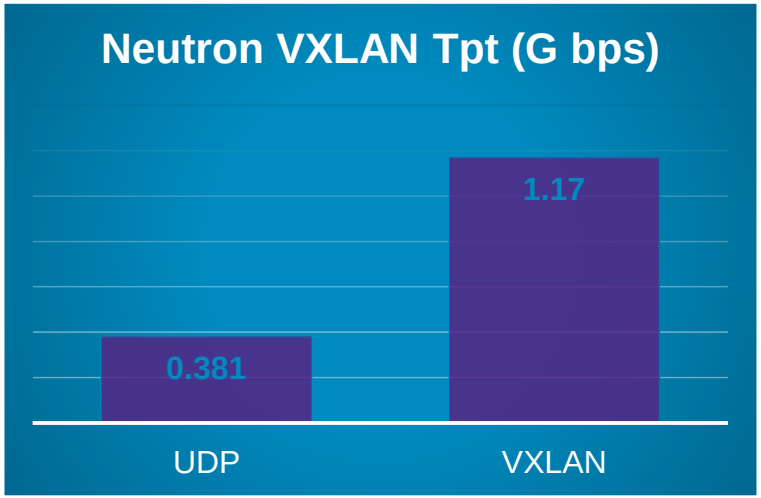




5. Networking Backend Does Matter

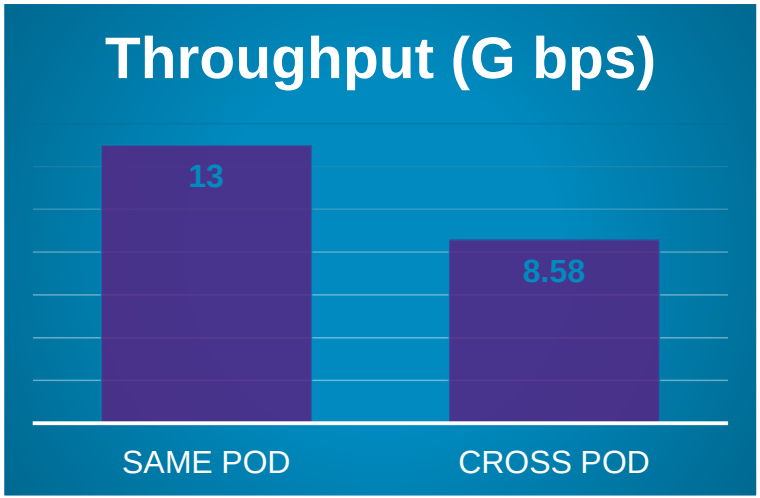
- Compare Flannel UDP and VXLAN backend
  - VXLAN obtain 3~5 x throughput!
  - IPtables, kube-proxy decrease 10% throughput, while increase 5% latency

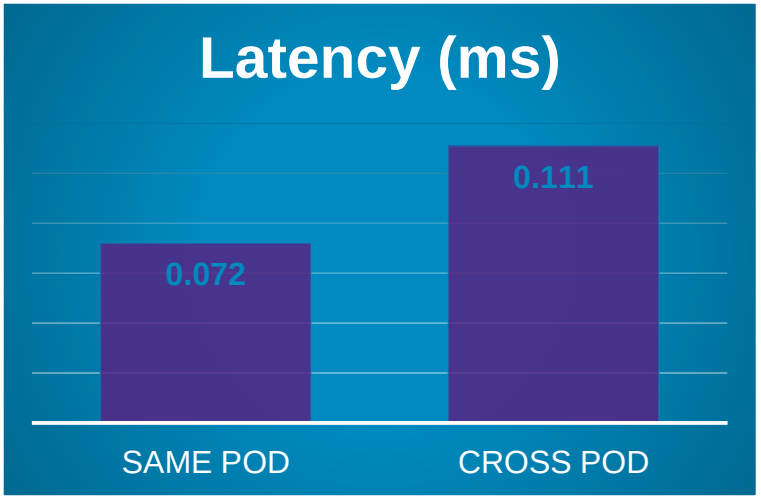




6. Linux Bridge Affects Container Performance

- Same-host pods connected by LB
  - 37% bw drop
  - 54% latency increase





Opportunities and Future Direction

Ongoing Efforts

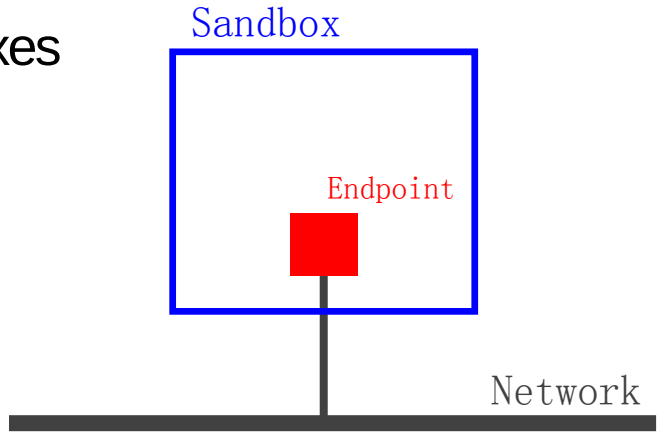
- Magnum: Container Networking Model
- Docker Networking: libnetwork
- Connecting Docker to Openstack: Project Kuryr
- Neutron: VLAN-aware VMs

Docker Networking: libnetwork

- Separated networking module from the core Docker engine
- New pluggable architecture
- Implements the Container Network Model (CNM)
- Available in Docker 1.9.0

Docker Networking: libnetwork

- Main Concepts:
  - Sandbox: Contains the configuration of a container's network stack
  - Network: Collection of Endpoints that can communicate with Each other
  - Endpoint: Connects networks to sandboxes
- Similar to Neutron Core concepts



Pluggable: libnetwork Drivers

- Supported drivers:
  - Null: no network
  - Bridge: traditional Docker networking (new implementation)
  - Overlay: multi-host networking
  - Remote: Connecting to Docker Network plugins
    - Uses JSON-RPC
    - Can be used to utilize Neutron

Project Kuryr

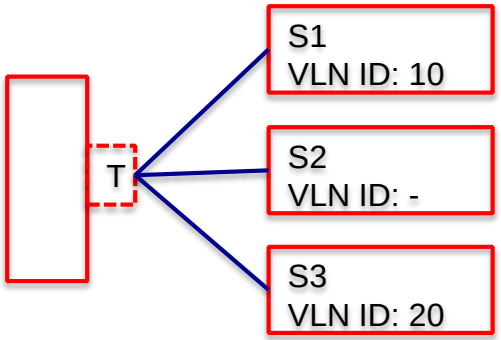
- Docker for Openstack Neutron
- Kuryr is a Docker network plugin that uses Neutron
  - Provides networking services to Docker containers
  - Will provide containerised images for the common Neutron plugins
- Part of OpenStack Ecosystem

Project Kuryr

- Docker Networks: Neutron networks
- Docker Endpoints: Neutron ports
  - Neutron subnets gets created from a predefined Neutron subnetpool
  - Docker IPAM driver for Kuryr in the works
- Docker Join/Leave: plug/unplug

Neutron: VLAN-Aware VMs

- Provides an efficient way for interconnecting containers deployed within VMs
- Avoids using overlays on top of overlays
- Define new types of Neutron Ports
  - Trunk ports
  - Parent/Children relationship
- Initial patches under review
- Building momentum with increased interest



T: Trunk Port
S: Subport

Thank You

IBM Open by design™

“IBM Client Day” on Wednesday October 28th

11:15 am	**OpenStack for Beginners (Putting OpenStack to work for your Business)** *Jesse Proudman*
12:05 pm	**The open cloud: A platform of possibilities - use cases from IBM and Blue Box** *Hernan Alvarez • Sunil Bhargava • Jojari Cannon Edwards • Azmir Mohamed*
2:00 pm	**It’s getting HOT in here: Turning up the HEAT with IBM MobileFirst for iOS apps** *Michael Elder • Tyson Lawrie • Tim Pouyer*
2:50 pm	**Hot topics with IBM: Federation and Containers** *Phil Estes • Andrew Hately • Steve Martinelli • Brad Topol*
3:40pm	**Case Study: Key Information Systems going above and beyond cloud demands with OpenStack and IBM Power Systems** *Clayton Weise (Key Information Systems) • Ann Funai*
4:40 pm	**Enterprise Journey to OpenStack Adoption: Real World Stories** *Markus Winter (SAP SE) • Atsushi Koga (NTT Data) • Nate Ziemann*
5:30 pm	**Gazing into the Crystal Ball, market insights and futures with IBM** *Angel Diaz • Moe Abdula • Jesse Proudman • Monty Taylor*

Visit the IBM Booth in the Marketplace

Tuesday, October 27th

1:00 – 2:00 **Book signing "Identity, Authentication & Access Management in OpenStack"**
Steve Martinelli • Henry Nash • Brad Topol

3:30 – 4:40 **Turning up the HEAT with IBM MobileFirst for iOS Apps**
Tyson Lawrie • Tim Pouyer

Wednesday, October 28th

1:00 – 2:00 **IBM Design Thinking: Making OpenStack Work For You**
Greg Hintermeister • Carly Stevens • Robin Cannon • Karl Vochatzer

3:30 – 4:40 **Object Storage Service on IBM Bluemix**
Dharmesh Bhakta • Riz Amanuddin

Thursday, October 29th

10:00 – 11:00 **Regulated workloads in the cloud: IBM Watson Health Cloud for Life Sciences Compliance**
Greg Bowman • Shawn Mullen

1:00 – 2:00 **Orchestration: Making Magic Happen with OpenStack and IBM Urban Code**
Michael Elder • John Page • Saurabh Agarwal

All sessions in room Heian (New Takanawa)

Scenarios considered

IBM Cloud

Main implementation

message path

		Flat	VLAN	VXLAN / GRE
Server to server		9.39	9.39	3.5
VM to VM		7.74	7.46	1.98
Container to container	same pod	13.0	12.3	13.2
	different pods, same host	8.58	7.33	8.40
	different pods, different hosts	1.71	1.67	1.17