Networking for VMs and Containers under Linux

Guillaume Urvoy-Keller

January 22, 2021

References

- Laurent Bernaille blog: `https://blog.revolve.team/author/lbernail/`
- Docker Networking Cookbook, PacktPub, Jon Langemak
- L3 + VXLAN Made Practical presentation (Openstack summit 2014) by Nolan Leake and Chet Burgess
- [ICIP19] Bacou, Mathieu, et al. "Nested Virtualization Without the Nest." Proceedings of the 48th International Conference on Parallel Processing. 2019.
- [HotCloud19] Lei, Jiaxin, et al. "Tackling parallelization challenges of kernel network stack for container overlay networks." 11th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 19). 2019.

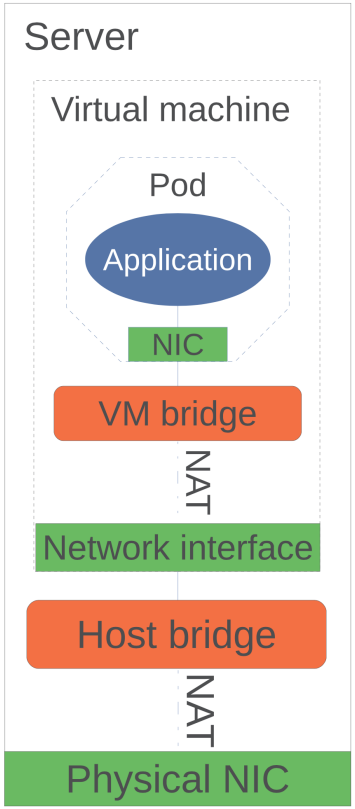
Outline

Typical deployments

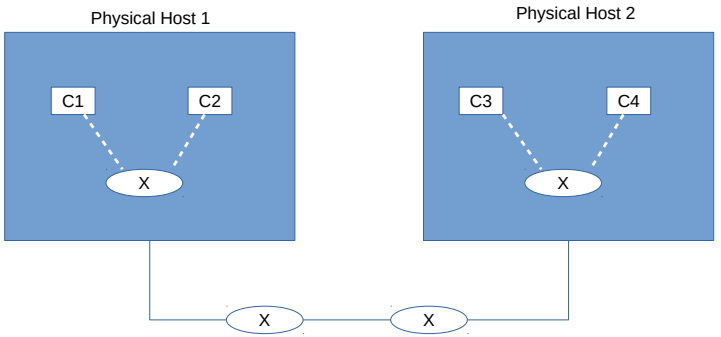
- Physical servers with VMs
- Physical servers with containers.
  Google claims that “everything at Google runs in containers” [1]
- Physical servers with VMs with containers or groups of containers (called PODS in Kubernetes)
  - Nested virtualization
  - Native containers ok for private setup
  - Nested approach enforces isolation among tenants

[1] Containers at Google. https://cloud.google.com/containers/.

Typical deployments with network circuitry



(a) Nested networking.



Containers in Physical servers

Basic network elements

- Physical NIC (pNIC)
- Virtual NIC (vNIC)
- Virtual Wires to interconnect vNIC to pNIC or NICs to switches
- Virtual Switches
- Virtual Routers
- Tunnels to interconnect PMs, VMs or containers
- Network Address Translation: for VMs/containers to borrow the host address

Basic network elements: Linux toolbox

- (Almost) Everything is an "interface" :
  - vNIC or pNIC
  - bridge
  - router: a bridge with an IP: uses the IP routing functions of the kernel
  - veth : a virtual wire with two ends, typically veth0@veth1 and veth1@veth0
  - Tunnel interfaces
- Important: when created, an interface receives an identifier (number) that stays the same during lifetime of interface
- Iptables to do NAT
- Network namespaces to have dedicated NICs, bridges, iptables rules, sockets

The "ip" command in Linux

Swiss knife of Linux for manipulating interfaces[2]

- ip link ... $\Rightarrow$ manipulates interfaces / bridges
- ip add ... $\Rightarrow$ assigns/removes IP addresses
- ip route ... $\Rightarrow$ modifies routing tables ; e.g. *ip route show*

```
user@net2:~$ sudo apt-get install iproute2 # what you need to manipulate network
    settings
user@net2:~$ sysctl -w net.ipv4.ip_forward=1 # transforms your machine into a router
```

[2] Beware of ifconfig (for instance, it does not see all the addresses of an interface if there are multiple addresses).

Creating a Linux Bridge

```
user@net1:~$ sudo ip link add host_bridge1 type bridge
user@net1:~$ ip link show host_bridge1
5: host_bridge1: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN mode
        DEFAULT group default
    link/ether f6:f1:57:72:28:a7 brd ff:ff:ff:ff:ff:ff
user@net1:~$ sudo ip address add 172.16.10.1/26 dev host_bridge1 # assigns an IP
        address to the interface to make it layer 3 aware (enables to use routing facility of
        kernel)
user@net1:~$ sudo ip link set dev eth1 master host_bridge1 # associate an interface to a
        bridge
user@net1:~$ sudo ip link set dev eth1 nomaster # de-associate
```

Creating a Veth pair

Let us create a second bridge (the first one was host_bridge)

```
user@net1:~$ sudo ip link add edge_bridge1 type bridge
user@net1:~$ sudo ip link add host_veth1 type veth peer name edge_veth1 # create a
        VETH pair specifying the ends name
user@net1:~$ ip link show
[...]
13: edge_veth1@host_veth1: <BROADCAST,MULTICAST,M-DOWN> mtu 1500 qdisc
        noop state DOWN mode DEFAULT group default qlen 1000
    link/ether 0a:27:83:6e:9a:c3 brd ff:ff:ff:ff:ff:ff
14: host_veth1@edge_veth1: <BROADCAST,MULTICAST,M-DOWN> mtu 1500 qdisc
        noop state DOWN mode DEFAULT group default qlen 1000
            link/ether c2:35:9c:f9:49:3e brd ff:ff:ff:ff:ff:ff
```

The interface number 13 and 14 remain the same, wherever the interface is, e.g. if you move it to another namespace

Side note...

Put all this up as this is not the default:

```
user@net1:~$ sudo ip link set host_bridge1 up
user@net1:~$ sudo ip link set edge_bridge1 up
user@net1:~$ sudo ip link set host_veth1 up
user@net1:~$ sudo ip link set edge_veth1 up
```

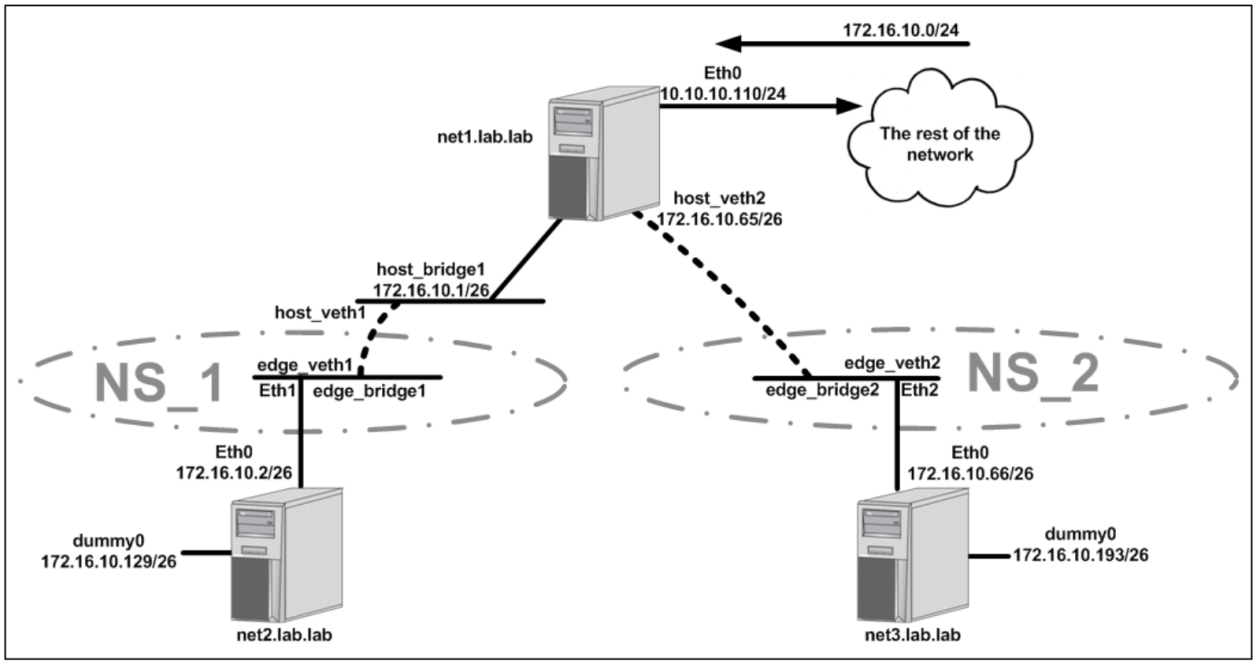
How to distinguish between a bridge or a simple interface or a veth: use **ip -d link** + name of interface:

```
root@ubuntu-xenial:/sys/class/net/enp0s3# ip -d link show dev docker0
6: docker0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP mode DEFAULT group
        default
    link/ether 02:42:86:07:6e:98 brd ff:ff:ff:ff:ff:ff promiscuity 0
    bridge forward_delay 1500 hello_time 200 max_age 2000 ageing_time 30000 stp_state 0 priority 32768 vlan_filtering
            0 vlan_protocol 802.1Q addrgenmode eui64
root@ubuntu-xenial:/sys/class/net/enp0s3# ip -d link show dev enp0s3
2: enp0s3: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP mode DEFAULT group
        default qlen 1000
    link/ether 02:d2:3e:0e:ff:c0 brd ff:ff:ff:ff:ff:ff promiscuity 0 addrgenmode eui64
root@ubuntu-xenial:/sys/class/net/enp0s3# ip -d link show dev veth84e2b4a
17: veth84e2b4a@if16: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue master docker0 state UP
        mode DEFAULT group default
    link/ether 72:14:0f:4d:d1:28 brd ff:ff:ff:ff:ff:ff link-netnsid 0 promiscuity 1
    veth # this is a veth connected to docker0
    bridge_slave state forwarding priority 32 cost 2 hairpin off guard off root_block off fastleave off learning on flood on
            addrgenmode eui64
```

Network Namespaces

- Network namespaces allow you to create isolated views of the network.
- Allows to mimic Virtual Routing and Forwarding (VRF) instances available in most modern networking hardware (e.g. Cisco Switches).

Scenario to implement (Docker Networking Cookbook)



Network Namespaces

```
user@net1:~$ sudo ip netns add ns_1
user@net1:~$ sudo ip netns add ns_2
user@net1:~$ ip netns list
ns_2
ns_1
```

Create the bridges inside the namespaces

```
user@net1:~$ sudo ip netns exec ns_1 ip link add edge_bridge1 type bridge
user@net1:~$ sudo ip netns exec ns_2 ip link add edge_bridge2 type bridge
```

Network Namespaces

Do an **ip link show** inside a given ns namespace

```
user@net1:~$ sudo ip netns exec ns_1 ip link show
1: lo: <LOOPBACK> mtu 65536 qdisc noop state DOWN mode DEFAULT group
default
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
2: edge_bridge1: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN mode
       DEFAULT group default
    link/ether 26:43:4e:a6:30:91 brd ff:ff:ff:ff:ff:ff
```

We next move the interfaces eth1 and eth2 within the namespaces + one side of the VETH pairs

```
user@net1:~$ sudo ip link set dev eth1 netns ns_1
user@net1:~$ sudo ip link set dev edge_veth1 netns ns_1
user@net1:~$ sudo ip link set dev eth2 netns ns_2
user@net1:~$ sudo ip link set dev edge_veth2 netns ns_2
```

For sake of completness

We have done the hard work. For sake of completness, we need to plug the VETH inside NS to the switchs and put everything up:

```
user@net1:~$ sudo ip netns exec ns_1 ip link set dev edge_veth1 master edge_bridge1
user@net1:~$ sudo ip netns exec ns_1 ip link set dev eth1 master edge_bridge1
user@net1:~$ sudo ip netns exec ns_2 ip link set dev edge_veth2 master edge_bridge2
user@net1:~$ sudo ip netns exec ns_2 ip link set dev eth2 master edge_bridge2
```

```
user@net1:~$ sudo ip netns exec ns_1 ip link set edge_bridge1 up
user@net1:~$ sudo ip netns exec ns_1 ip link set edge_veth1 up
user@net1:~$ sudo ip netns exec ns_1 ip link set eth1 up
user@net1:~$ sudo ip netns exec ns_2 ip link set edge_bridge2 up
user@net1:~$ sudo ip netns exec ns_2 ip link set edge_veth2 up
user@net1:~$ sudo ip netns exec ns_2 ip link set eth2 up
```

Docker (host-level) Networking

Docker networks

You have a set of predefined networks:

```
root@ubuntu−xenial: docker network ls
NETWORK ID NAME DRIVER SCOPE
bfb14981a5df bridge bridge local
b7c327787044 host host local
492f4a9fe233 none null local
```

- bridge mode is default (see next slide):
- Host mode is when you connect container directly to the host $\Rightarrow$ leads to port contention, e.g., you cannot run multiple replicas of a web server!
- None is.... none

Docker bridge mode: bridge0 in the Linux main network namespace

```
root@ubuntu−xenial:/sys/class/net/enp0s3# docker network inspect bridge
    {
        "Name": "bridge",
        "Id": "bfb14981a5df08fe27881557d87bc0be18185cd0ba64f7552e196fd1bbad1260",
        "Created": "2017−10−20T14:49:36.899406866Z",
        "Driver": "bridge",
        "EnableIPv6": false,
        "IPAM": {
              "Config": [
                {
                    "Subnet": "172.17.0.0/16",
                    "Gateway": "172.17.0.1"
                }
            ]
        },
        [....]
                },
        "Options": {
            "com.docker.network.bridge.default_bridge": "true",
            "com.docker.network.bridge.enable_icc": "true",
            "com.docker.network.bridge.enable_ip_masquerade": "true",
            "com.docker.network.bridge.host_binding_ipv4": "0.0.0.0",
            "com.docker.network.bridge.name": "docker0",
            "com.docker.network.driver.mtu": "1500"
        },
        "Labels": {}
    }
```

Custom networks

```
root@ubuntu-xenial:~# docker network create mynetwork
0b396f0fc9264ad7153943f293b863e028c858c4d051744b93128bc21b291336
```

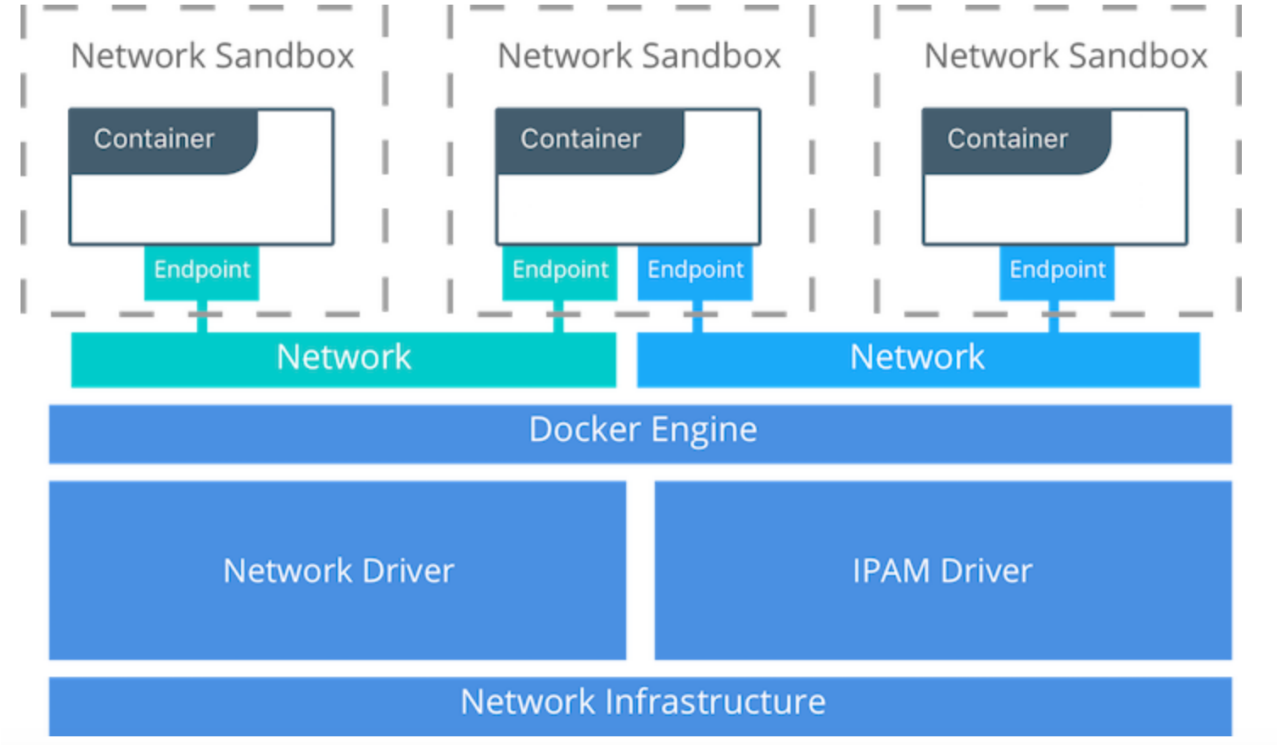
However, the scope is still local (host machine) – see last column.

The real meat will be the overlay.

```
root@ubuntu-xenial:~# docker network ls
NETWORK ID NAME DRIVER SCOPE
bfb14981a5df bridge bridge local
b7c327787044 host host local
0b396f0fc926 mynetwork bridge local
492f4a9fe233 none null local
```

Docker Networking Model

The Container Networking Model

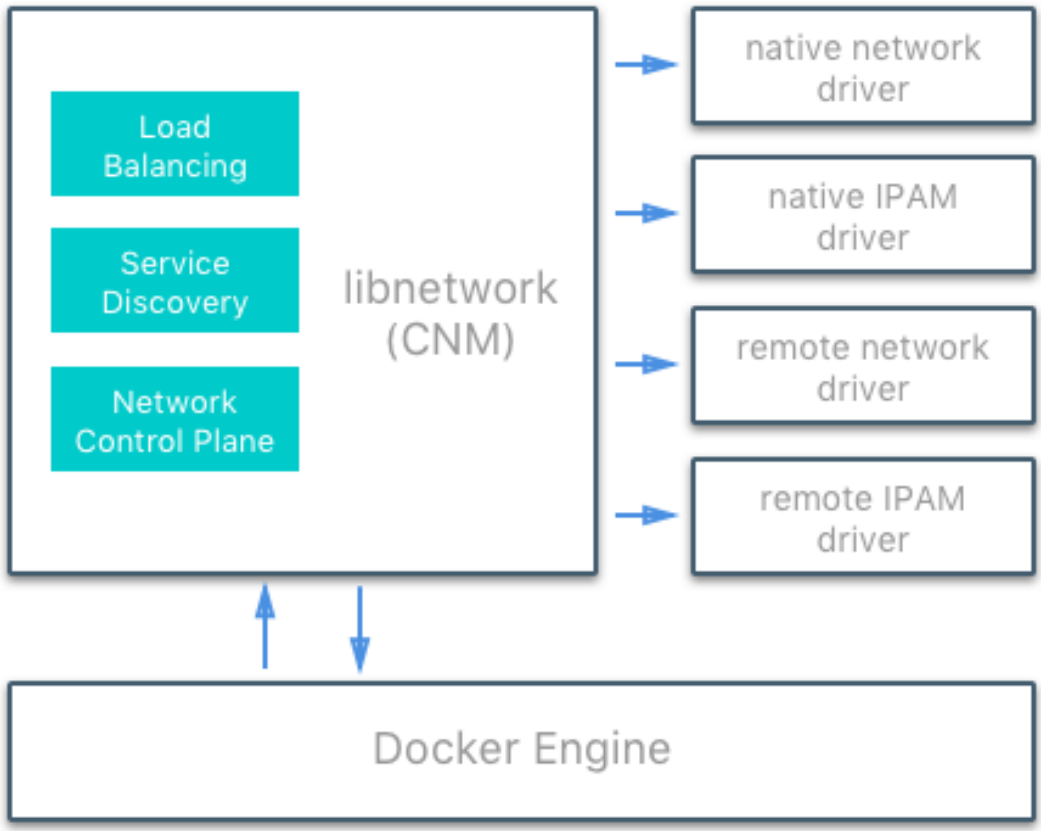


source: https://success.docker.com/Architecture/Docker_Reference_Architecture%3A_Designing_Scalable%2C_Portable_Docker_Container_Networks

The Container Networking Model

- "Sandbox — A Sandbox contains the configuration of a container's network stack. This includes management of the container's interfaces, routing table, and DNS settings. An implementation of a Sandbox could be a Linux Network Namespace, a FreeBSD Jail, or other similar concept."
- Endpoint: enable connection to the outside world, from a simple bridge to a complex overlay network
- Network driver: possibility to use Docker solution (swarm) or third party
- IPAM : IP address management - DHCP and the like

An open Network driver Model



source: https://success.docker.com/Architecture/Docker_Reference_Architecture%3A_Designing_Scalable%2C_Portable_Docker_Container_Networks

Docker Native Network Drivers

Driver	Description
Host	With the `host` driver, a container uses the networking stack of the host. There is no namespace separation, and all interfaces on the host can be used directly by the container
Bridge	The `bridge` driver creates a Linux bridge on the host that is managed by Docker. By default containers on a bridge can communicate with each other. External access to containers can also be configured through the `bridge` driver
Overlay	The `overlay` driver creates an overlay network that supports multi-host networks out of the box. It uses a combination of local Linux bridges and VXLAN to overlay container-to-container communications over physical network infrastructure
MACVLAN	The `macvlan` driver uses the MACVLAN bridge mode to establish a connection between container interfaces and a parent host interface (or sub-interfaces). It can be used to provide IP addresses to containers that are routable on the physical network. Additionally VLANs can be trunked to the `macvlan` driver to enforce Layer 2 container segmentation
None	The `none` driver gives a container its own networking stack and network namespace but does not configure interfaces inside the container. Without additional configuration, the container is completely isolated from the host networking stack

source: `https://success.docker.com/Architecture/Docker_Reference_Architecture%3A_Designing_Scalable%2C_Portable_Docker_Container_Networks`

Remote Network driver

Driver	Description
contiv	An open source network plugin led by Cisco Systems to provide infrastructure and security policies for multi-tenant microservices deployments. Contiv also provides integration for non-container workloads and with physical networks, such as ACI. Contiv implements remote network and IPAM drivers.
weave	A network plugin that creates a virtual network that connects Docker containers across multiple hosts or clouds. Weave provides automatic discovery of applications, can operate on partially connected networks, does not require an external cluster store, and is operations friendly.
calico	An open source solution for virtual networking in cloud datacenters. It targets datacenters where most of the workloads (VMs, containers, or bare metal servers) only require IP connectivity. Calico provides this connectivity using standard IP routing. Isolation between workloads — whether according to tenant ownership or any finer grained policy — is achieved via iptables programming on the servers hosting the source and destination workloads.
kuryr	A network plugin developed as part of the OpenStack Kuryr project. It implements the Docker networking (libnetwork) remote driver API by utilizing Neutron, the OpenStack networking service. Kuryr includes an IPAM driver as well.

source: `https://success.docker.com/Architecture/Docker_Reference_Architecture%3A_Designing_Scalable%2C_Portable_Docker_Container_Networks`

Docker Swarm 101

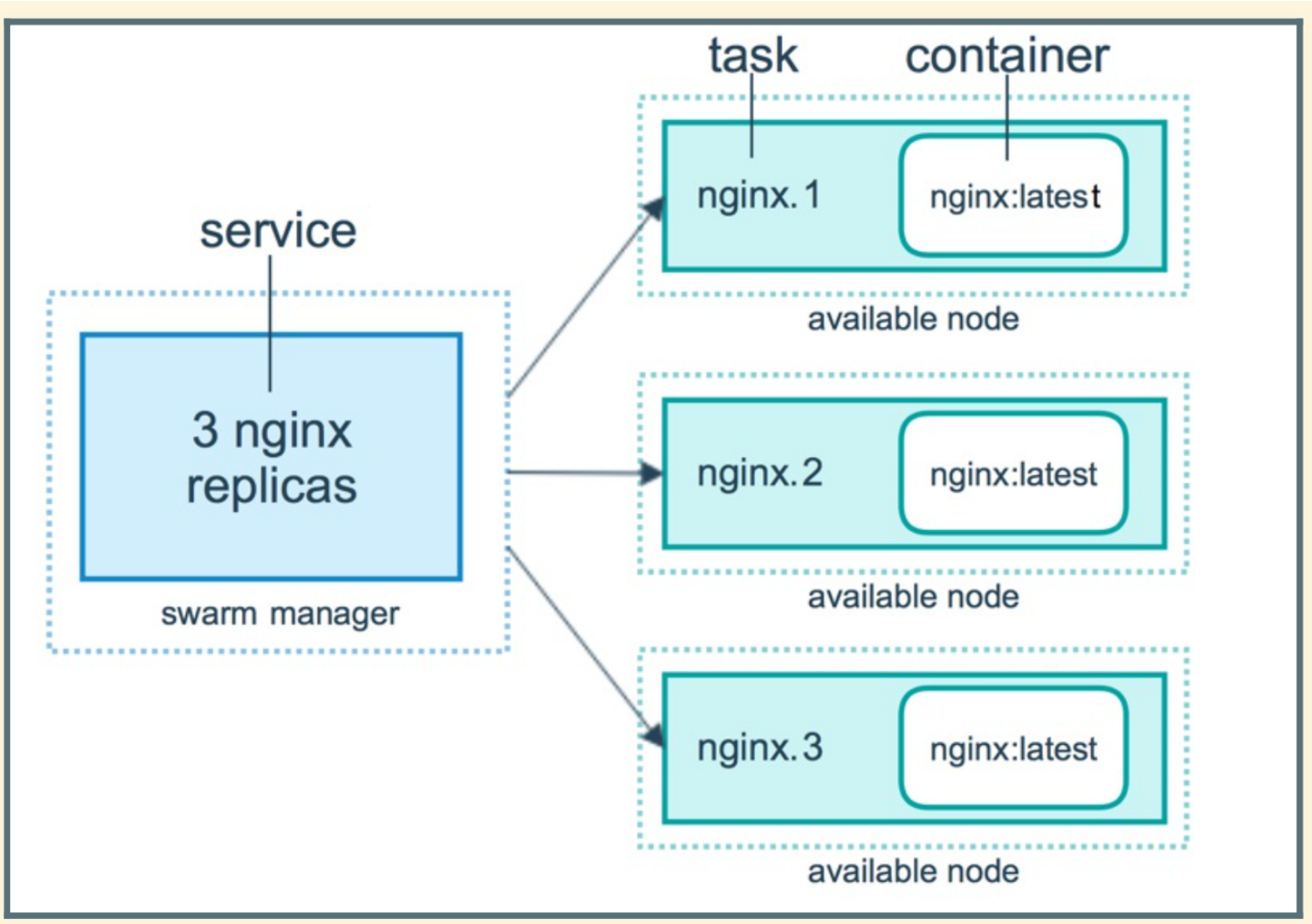
Docker swarm

Several Docker Hosts $\Rightarrow$ Use them in Cluster

Docker Engine 1.12:

- Natively supports swarm
- Clusters organized into workers, managers and leaders
- Dispatching of services : tasks to be executed by servers

Swarm tasks dispatching



Swarm operations

```
$ docker swarm init --advertise-addr <MANAGER-IP>

# Swarm initialized: current node (8jud...) is now a manager. To add a worker to this swarm
      , run the following command:

docker swarm join --token SWMTKN-1-59fl4ak4nqjmao1ofttrc4eprhrola2l87... \
      172.31.4.182:2377
```

Check state:

```
$ docker info Swarm: active
NodeID: 8jud7o8dax3zxbags3f8yox4b
Is Manager: true
ClusterID: 2vcw2oa9rjps3a24m91xhvv0c
```

You have created a first node in the swarm (your host)

```
$ docker node ls
ID HOSTNAME STATUS AVAILABILITY MANAGER STATUS
8jud...ox4b * ip-172-31-4-182 Ready Active Leader
```

Docker swarm

Docker has generated tokens to join the swarm:

```
$ docker swarm join−token worker
$ docker swarm join−token manager
```

You can then join by issuing on the second host:

```
$ docker swarm join −−token TOKEN−WORKER... 172.31.4.182:2377
```

If this works, you should have

```
$ docker node ls
ID HOSTNAME STATUS AVAILABILITY MANAGER STATUS
8jud...ox4b ∗ ip−172−31−4−182 Ready Active Leader
ehb0...4fvx ip−172−31−4−180 Ready Active
```

Docker swarm

You can now execute a service :

```
root@ubuntu−xenial: docker service create −−replicas 1 −−name helloworld alpine ping
    docker.com
```

and observe the services in general or a specific service

```
root@ubuntu−xenial: docker service create −−replicas 1 −−name helloworld alpine ping docker.com
2klpz2bef3ez7w498hw17bwbw

root@ubuntu−xenial: docker service ls
ID NAME MODE REPLICAS IMAGE PORTS
2klpz2bef3ez helloworld replicated 1/1 alpine:latest
root@ubuntu−xenial: docker service ps helloworld
ID NAME IMAGE NODE DESIRED STATE CURRENT STATE ERROR PORTS
5uwod1wobk0m helloworld.1 alpine:latest ubuntu−xenial Running Running 35 seconds ago
```

Docker Network Overlay

Docker Overlay

- Enables multi-host networking
  - A host here is a physical or virtual machine that features the docker daemon
  - Docker hosts be created independently or from a central place using docker-machine
- Docker overlay driver enables to create a VLAN
  (= private IP network e.g. 172.19.1.0/24)
  for groups of distributed (over the Docker hosts) containers

Docker Network Overlay

Create an overlay

```
$ docker network create --driver overlay my-network
```

Inspect network

```
$ docker network inspect my-network
[
    {
        "Name": "my-network",
        "Id": "fsf1dmx3i9q75an49z36jycxd",
        "Created": "0001-01-01T00:00:00Z",
        "Scope": "swarm",
        "Driver": "overlay",
        "EnableIPv6": false,
        "IPAM": {
            "Driver": "default",
            "Options": null,
            "Config": []
        },
        "Internal": false,
        "Attachable": false,
        "Ingress": false,
        "Containers": null,
        "Options": {
            "com.docker.network.driver.overlay.vxlanid_list": "4097"
        },
        "Labels": null
    }
]
```

Docker Network Overlay

What is important in previous listing:

- The driver : overlay!
- The scope : swarm $\Rightarrow$ network extends to a swarm, not local to host
- Attached containers are listed in the docker inspect

You can now attach a service (set of containers) to the overlay

```
$ docker service create replicas 3 −−name my−web −−network my−network nginx
```

Docker Ovelay Network: What is behind the hood?

IETF RFC 7348

Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks

- Tunnelling of Ethernet frames within UDP packets

VXLAN

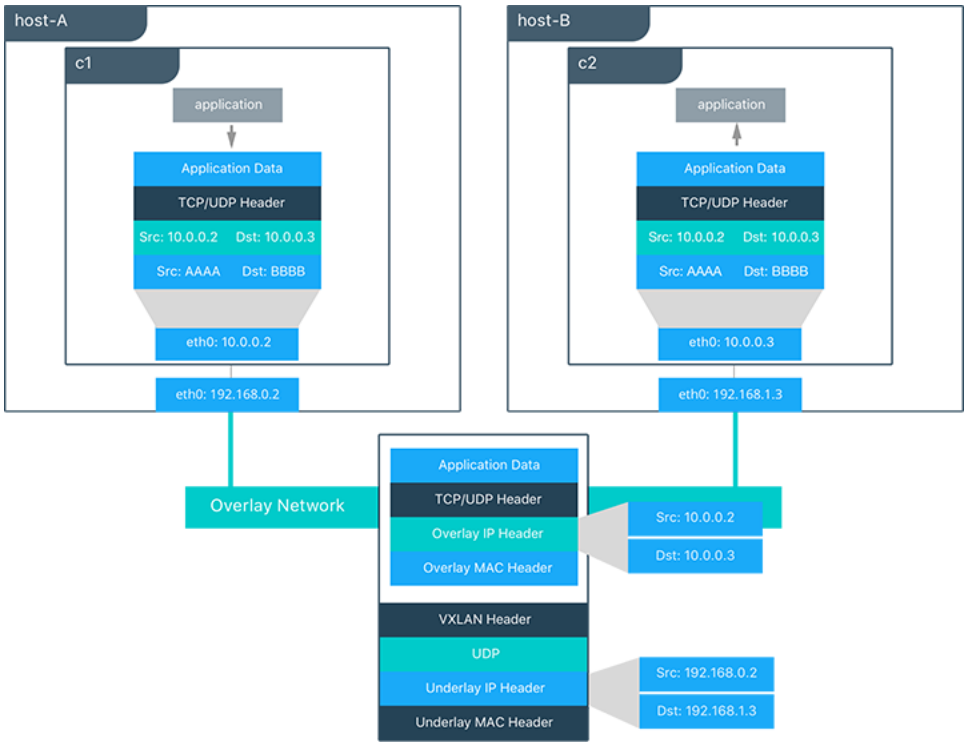


Figure: source: docker web site

"c1 does a DNS lookup for c2. Since both containers are on the same overlay network the Docker Engine local DNS server resolves c2 to its overlay IP address 10.0.0.3." $\Rightarrow$ need to discuss additional services like DNS, load balancing

Docker overlay - what is added in a given host (and in each container)

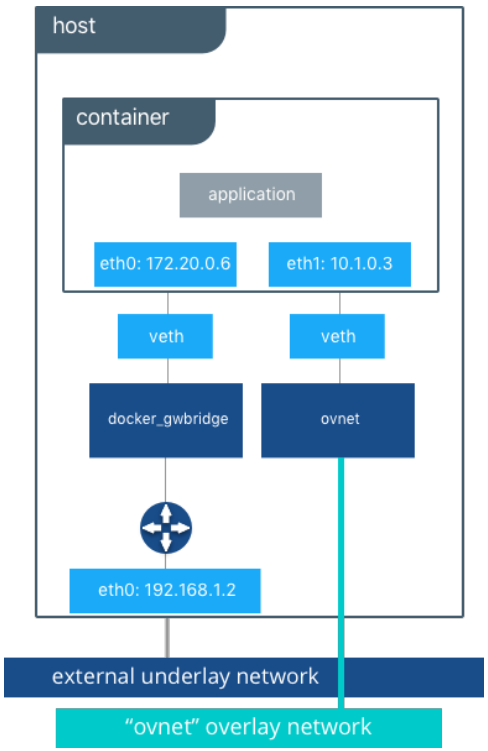


Figure: source: docker web site

Docker overlay - what is added in a given host (and in each container)

- Docker_gwbridge: the egress bridge for traffic that goes outside (underlay means 'not any VXLAN interface')
- ovnet: one bridge per overlay.
  - Created on each docker host with containers in this overlay, ie. no a priori creation on all swarm nodes (just where needed)
    - Called the egress bridge
    - One per host
    - Constitutes the so-called VXLAN Tunnel End Point (VETP).
    - VETPs communicate with each other to maintain the overlay
  - Uses a Linux VXLAN port to attach to the outside.

VXLAN

- VNI–VXLAN Network Identifier
  - 24 bit number (16M+ unique identifiers)
  - Part of the VXLAN Header
  - Similar to VLAN ID
  - Limits broadcast domain
- VTEP–VXLAN Tunnel End Point
  - Originator and/or terminator of VXLAN tunnel for a specific VNI
  - Outer DIP/Outer SIP

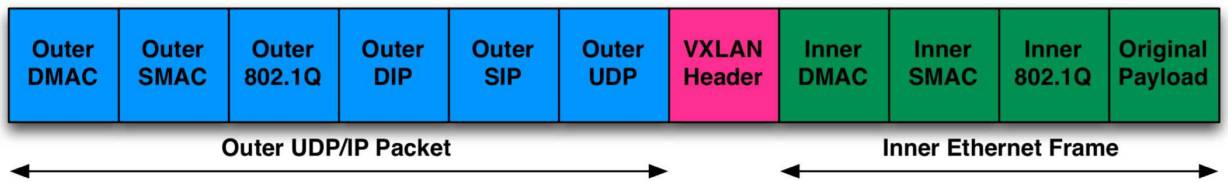


Figure: source: Nolan Leake (cumulus) and Chet Burgess (Metacloud)

VXLAN

Sending a packet

- ARP table is checked for IP/MAC/Interface mapping
- L2 FDB is checked to determine IP of destination on VTEP for destination on MAC on source VTEP
- Source VETP then encapsulates frame with correct destination VETP and destination decapsulates....

```
root@mcpl.dev4.mc:~$ ip addr show dev vxlan12345
140: vxlan12345: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue state UNKNOWN
 group default
    link/ether 4e:3c:82:2e:09:0e brd ff:ff:ff:ff:ff:ff
    inet 10.9.8.1/24 scope global vxlan12345
       valid_lft forever preferred_lft forever
root@mcpl.dev4.mc:~$ arp -an | grep 10.9.8.2
? (10.9.8.2) at 7a:79:96:de:2c:a2 [ether] on vxlan12345
root@mcpl.dev4.mc:~$ bridge fdb show dev vxlan12345
7a:79:96:de:2c:a2 dst 172.16.3.102 self
4e:3c:82:2e:09:0e dst 127.0.0.1 self
```

Figure: source: Nolan Leake (cumulus) and Chet Burgess (Metacloud)

VXLAN

- Need a mechanism to determine the VETP (tunnel end point = physical machin) for a container MAC @
- Option in VXLAN
  - IP Multicast group per overlay (per VXLAN)
  - Unicast also possible

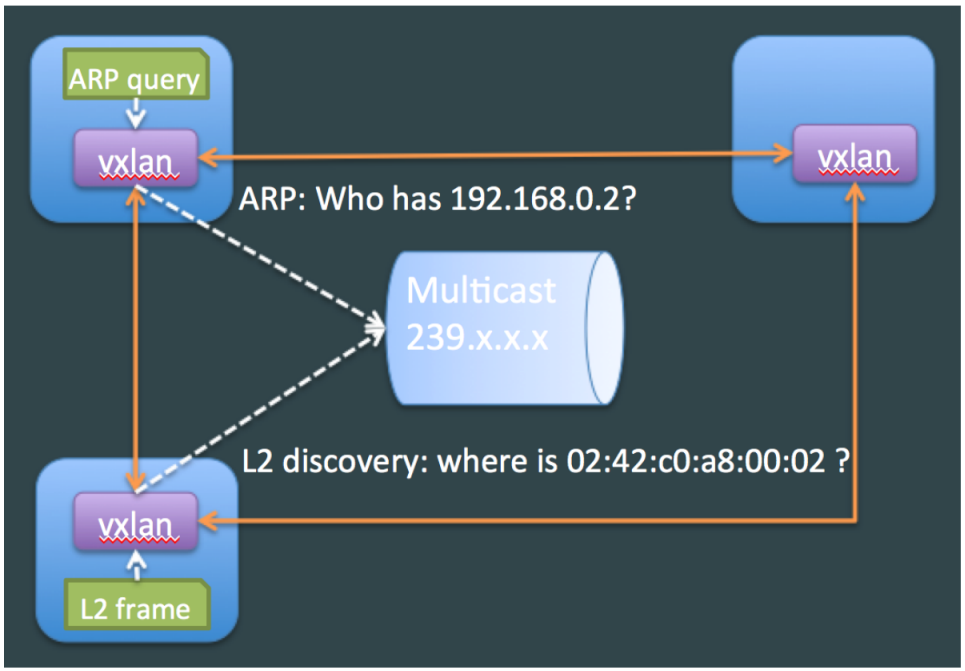


Figure: source: Nolan Leake (cumulus) and Chet Burgess (Metacloud)

VXLAN support in Linux

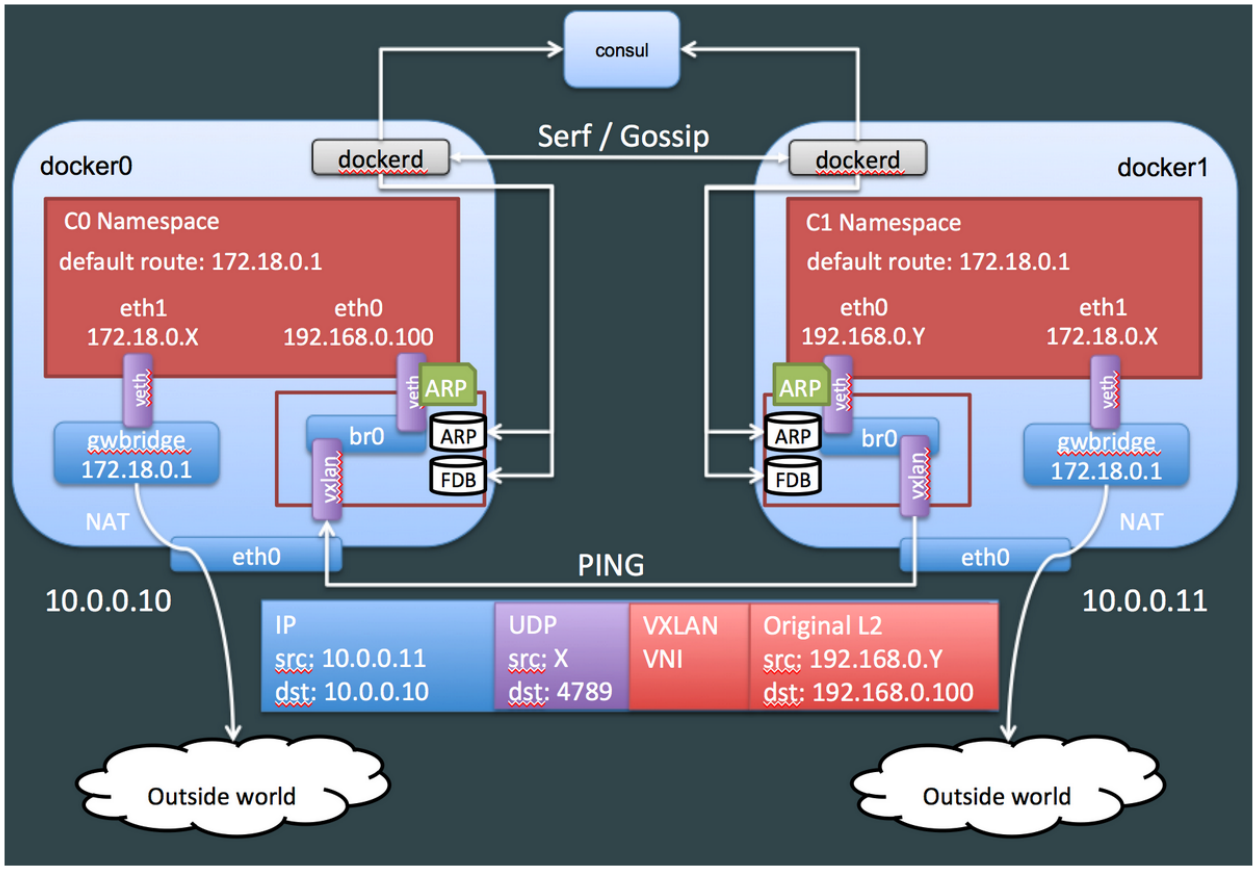
- Well supported in most modern Linux Distros – Linux Kernel 3.10+
  - Linux uses UDP port 8472 instead of IANA issued 4789 – iproute2 3.7+
  - Configured using ip link command

```
root@mcpl.dev4.mc:~$ ip link add vxlanl2345 type vxlan id 12345 remote 172.16.3.100
root@mcpl.dev4.mc:~$ ip link set dev vxlanl2345 up
root@mcpl.dev4.mc:~$ ip -d link show dev vxlanl2345
140: vxlanl2345: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue state UNKNOWN
 mode DEFAULT group default
    link/ether 4e:3c:82:2e:09:0e brd ff:ff:ff:ff:ff:ff promiscuity 0
    vxlan id 12345 remote 172.16.3.100 port 32768 61000 ageing 300
```

Figure: source: Nolan Leake (cumulus) and Chet Burgess (Metacloud)

VLXAN - The Docker way

- Key advantage: Docker knows where MAC addresses appear in the overlay when it creates the containers
- Propagate to each VETP the MAC/VETP mapping: no need for arp, hence no mutlicast or unicast of arp!



Docker Network Control Plane Model

Relies on a gossip protocol (SWIM) to propagate network state information and topology across Docker container clusters. A complex task at large scale when one must reach a consensus. Part of Swarm.

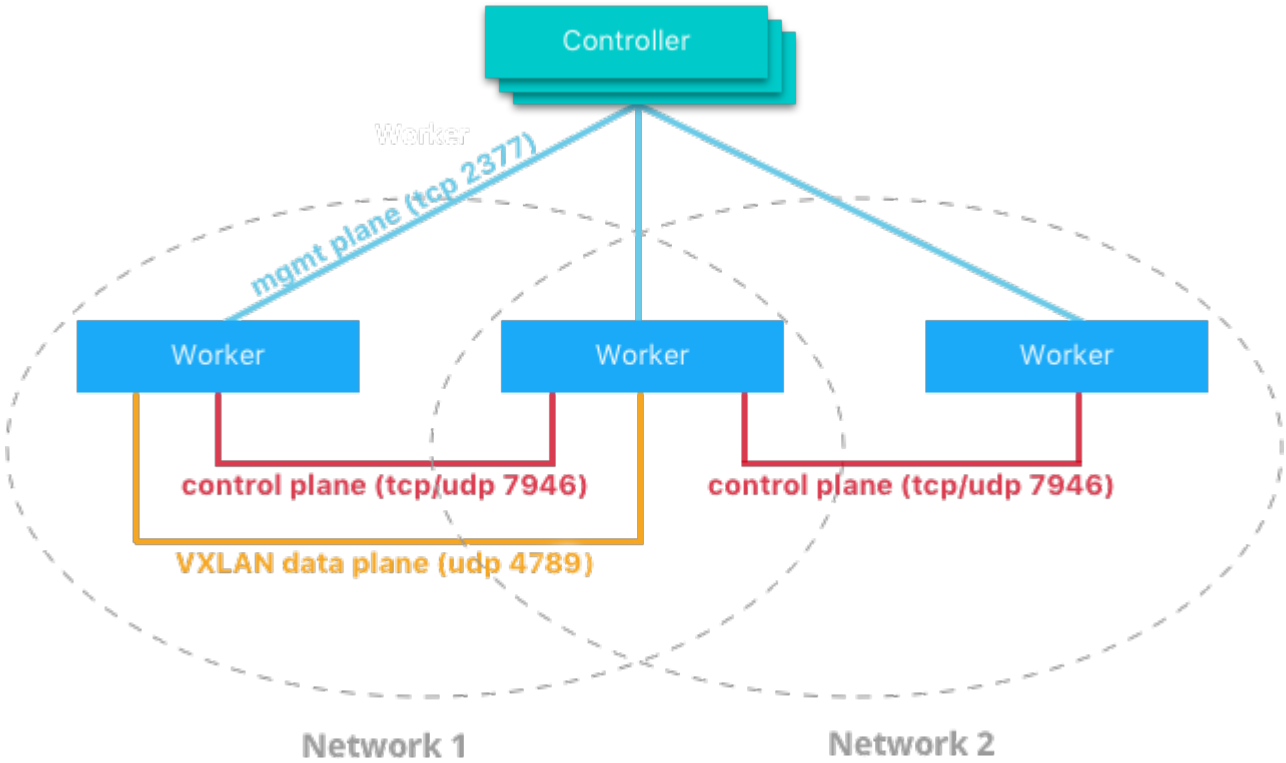


Figure: source: docker web site

Docker networking: What is still missing?

- Port publishing at swarm level and load balancing
- DNS support (omitted)

Exposing ports

- Two modes:
  - Host mode: only exposes the IP of hosts with a replica
  - Ingress mode (default mode) : any host even without replica will answer.

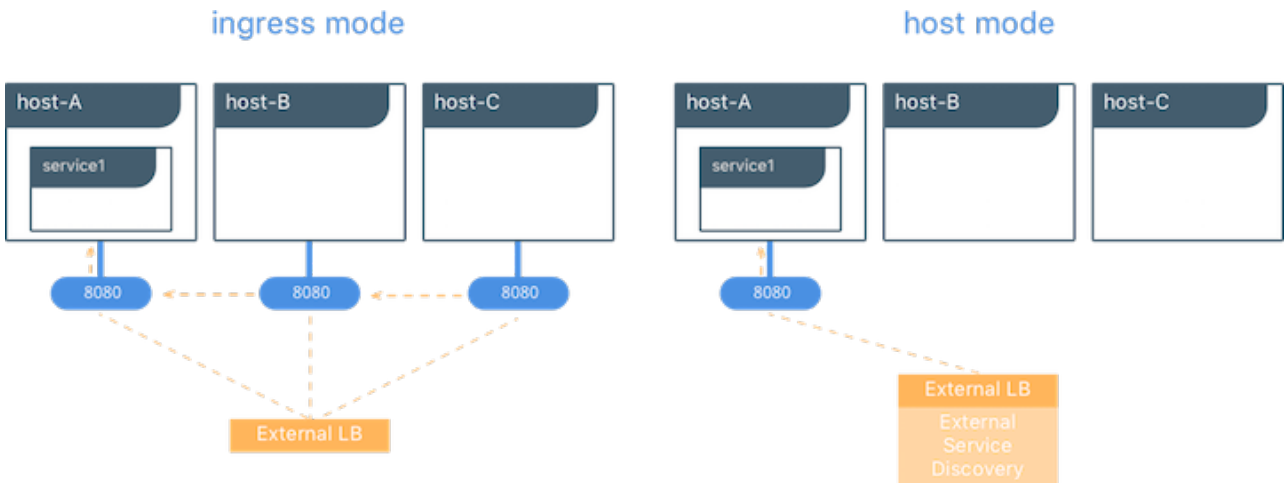


Figure: source: docker web site

Exposing port: Ingress mode

- Ingress mode relies on Swarm Routing mesh protocol (control plane TCP port on docker network control plane slide)
- In general, usage of an external (to docker) load balancer like HAproxy

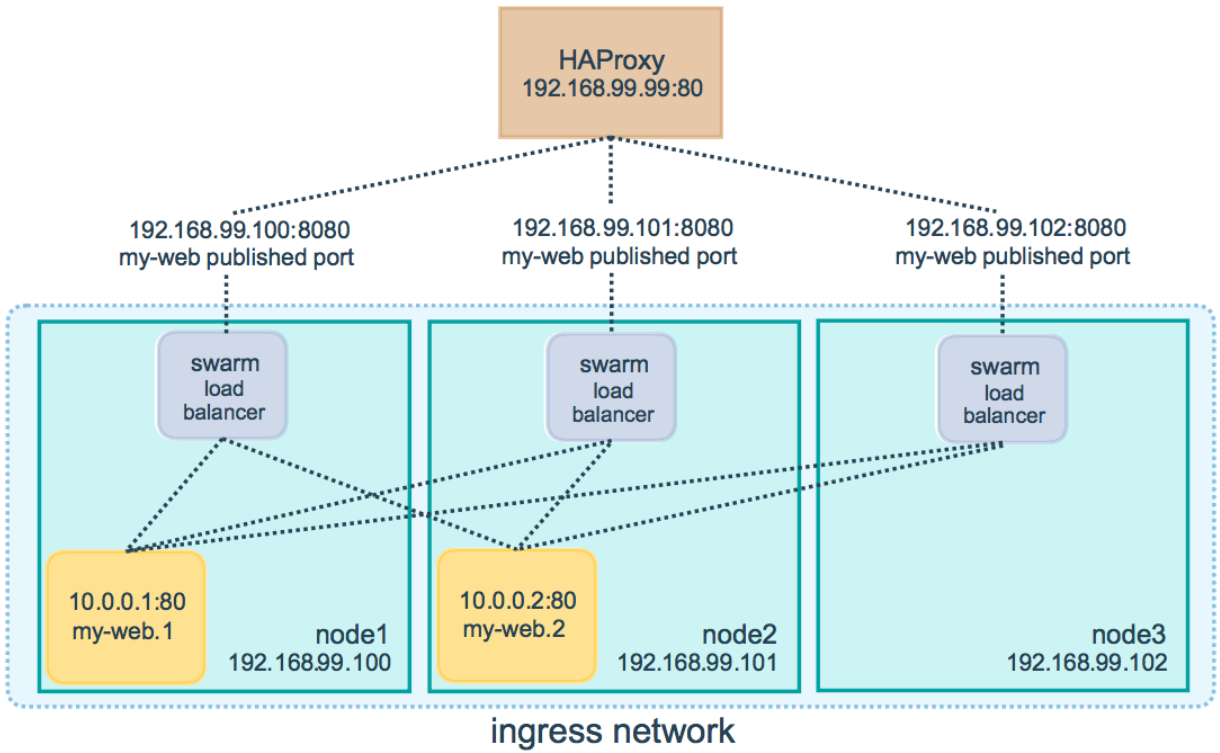


Figure: source: docker web site

Exposing port: Ingress mode

Relies on ipvs (load balacing at layer 4) and iptables (firewall/NAT) of Linux kernel

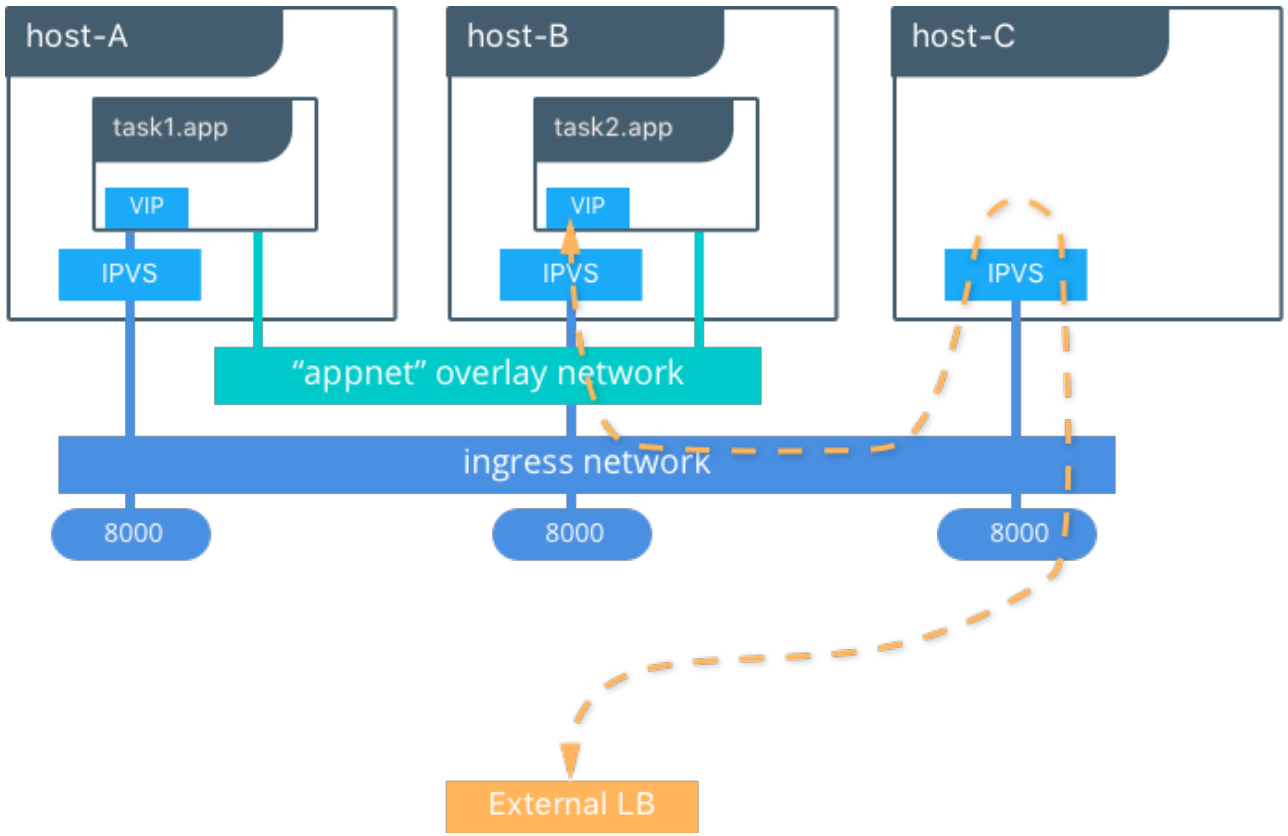


Figure: source: docker web site

Docker Network Overlay

Network Overlay: a peek on performance

- [HotCloud19] Lei, Jiaxin, et al. "Tackling parallelization challenges of kernel network stack for container overlay networks." 11th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 19). 2019.
  - Highlight the inefficiency at high speed of the kernel
- [ICIP19] Bacou, Mathieu, et al. "Nested Virtualization Without the Nest." Proceedings of the 48th International Conference on Parallel Processing. 2019.
  - Proposal of a solution to boost performance of nested virtualization

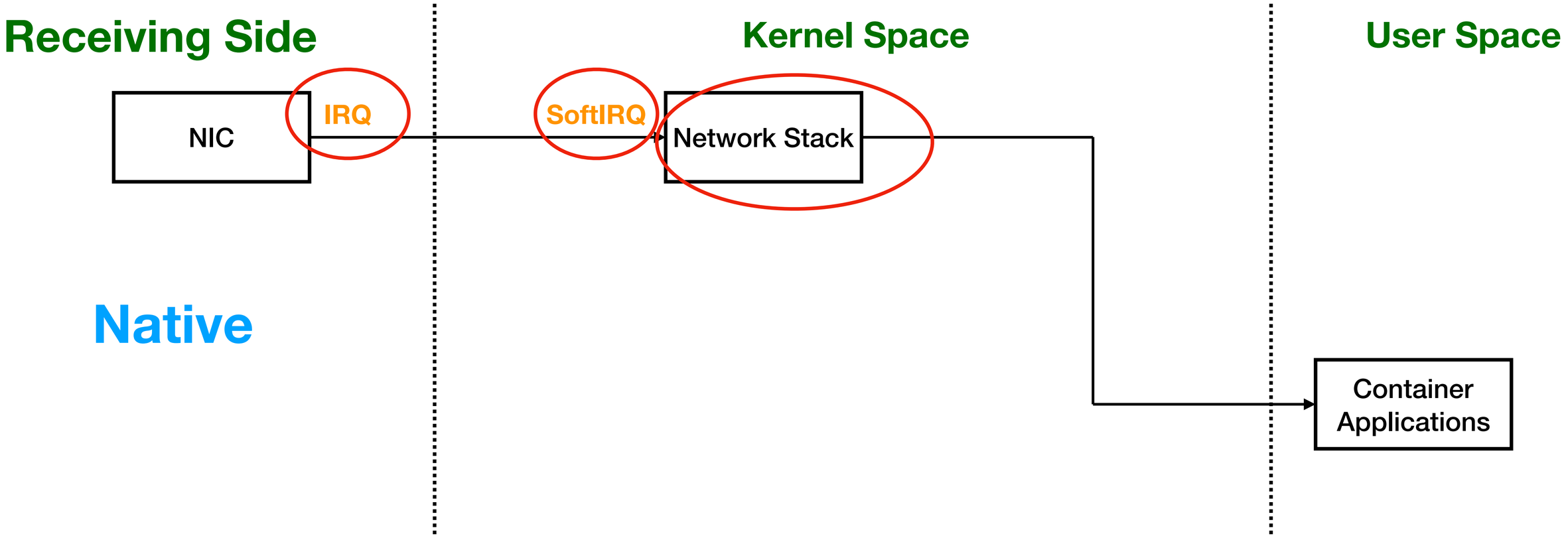
Tackling Parallelization Challenges of Kernel Network Stack for Container Overlay Networks

Jiaxin Lei*, Kun Suo+, Hui Lu*, Jia Rao+

* SUNY at Binghamton
+ University of Texas at Arlington

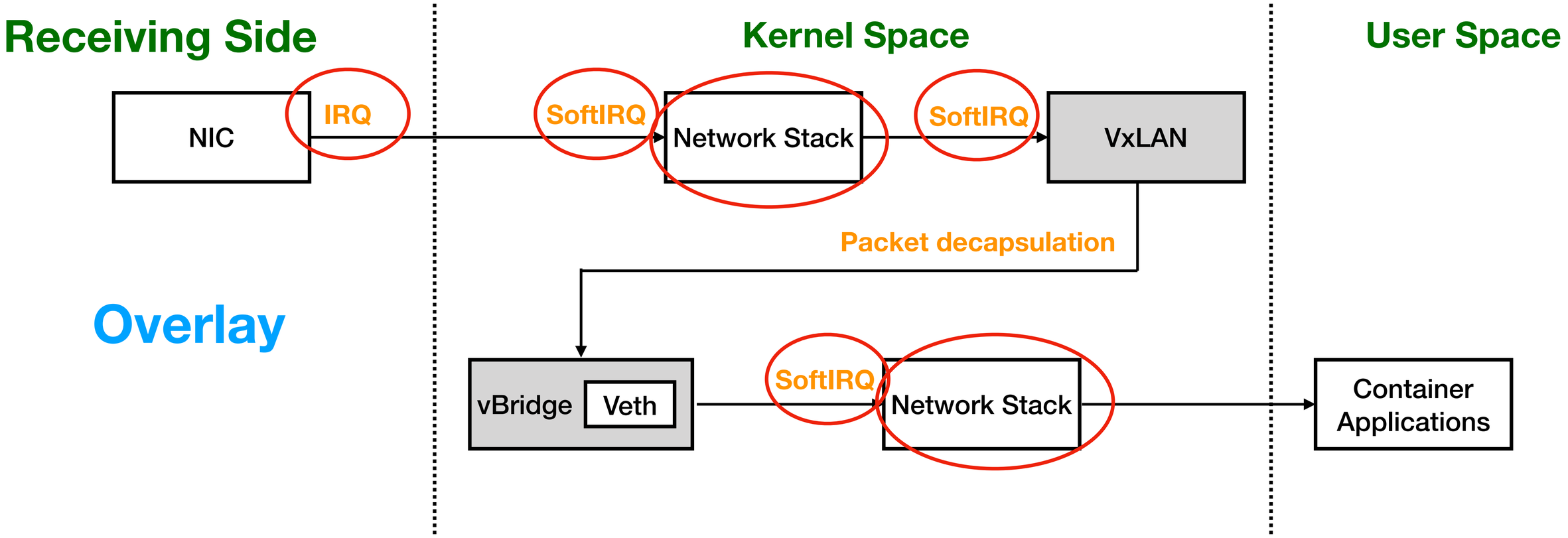
Network Packet Processing Path

- **Prolonged** network packet processing path
- **Additional** virtual devices overhead



Network Packet Processing Path

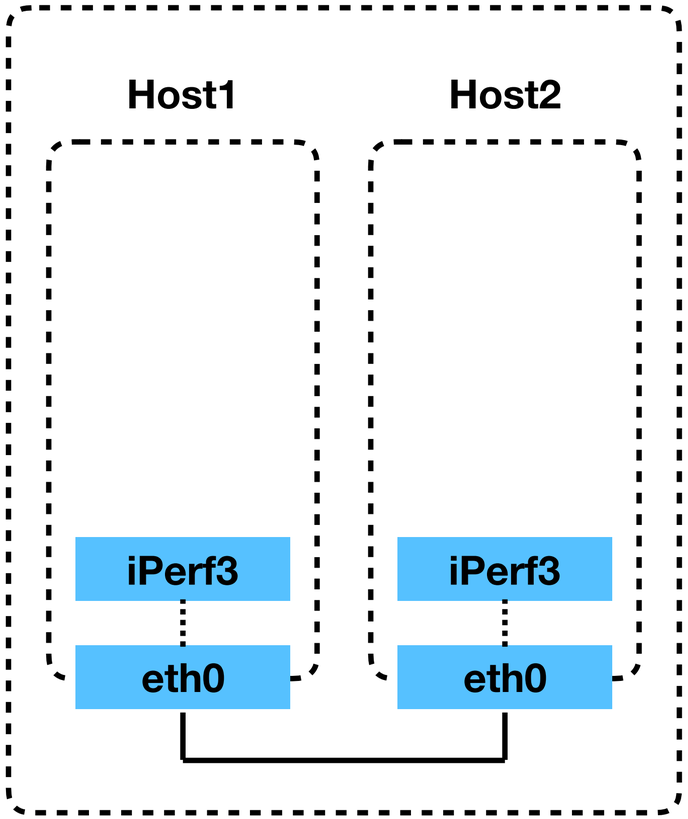
- **Prolonged** network packet processing path
- **Additional** virtual devices overhead



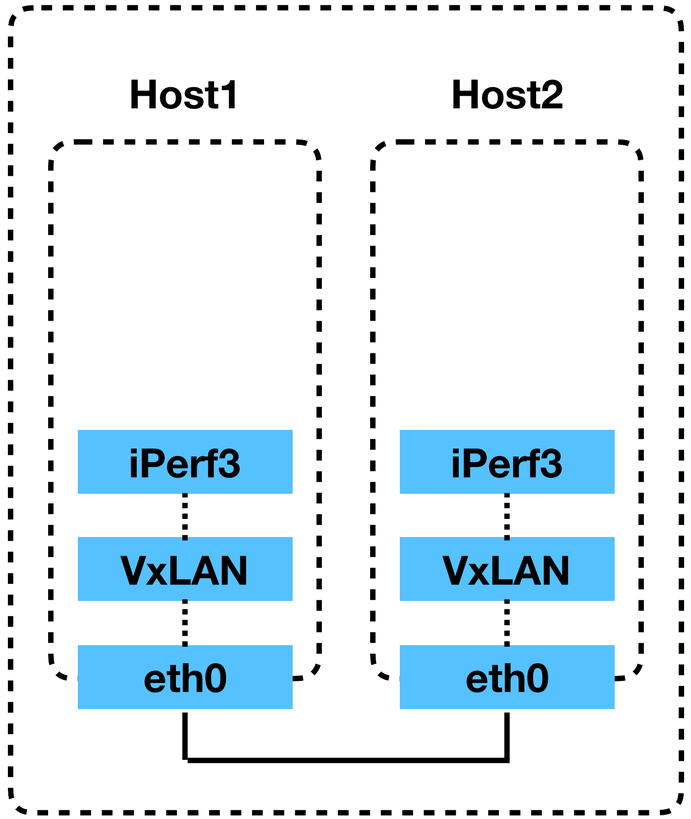
Experimental Settings

Hardware			Software	
CPU	Memory	NIC	Throughput	CPU Usage
2.2 GHz	64GB	40Gbps	iPerf3	mpstat
10 cores		Multi-queue		

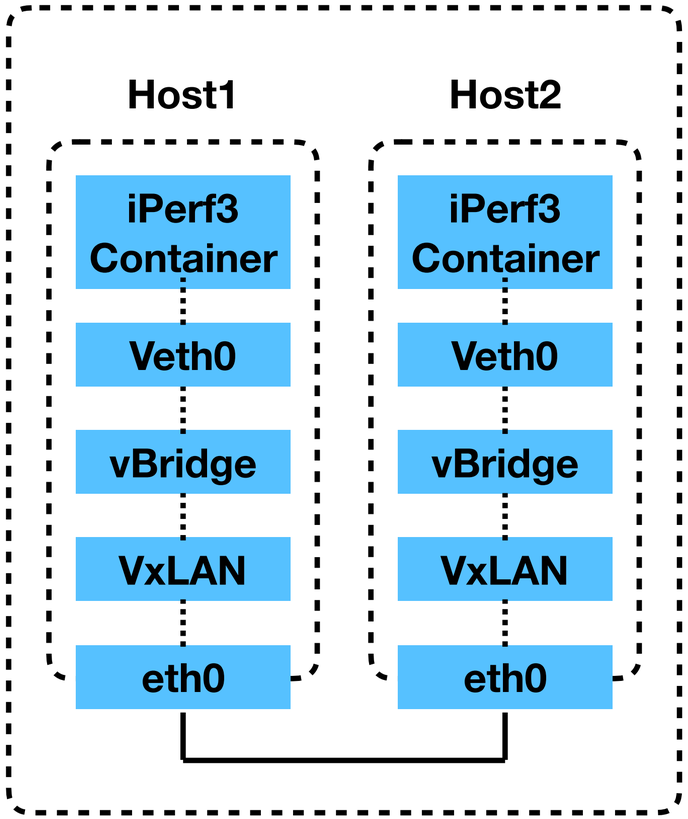
S1 - Native



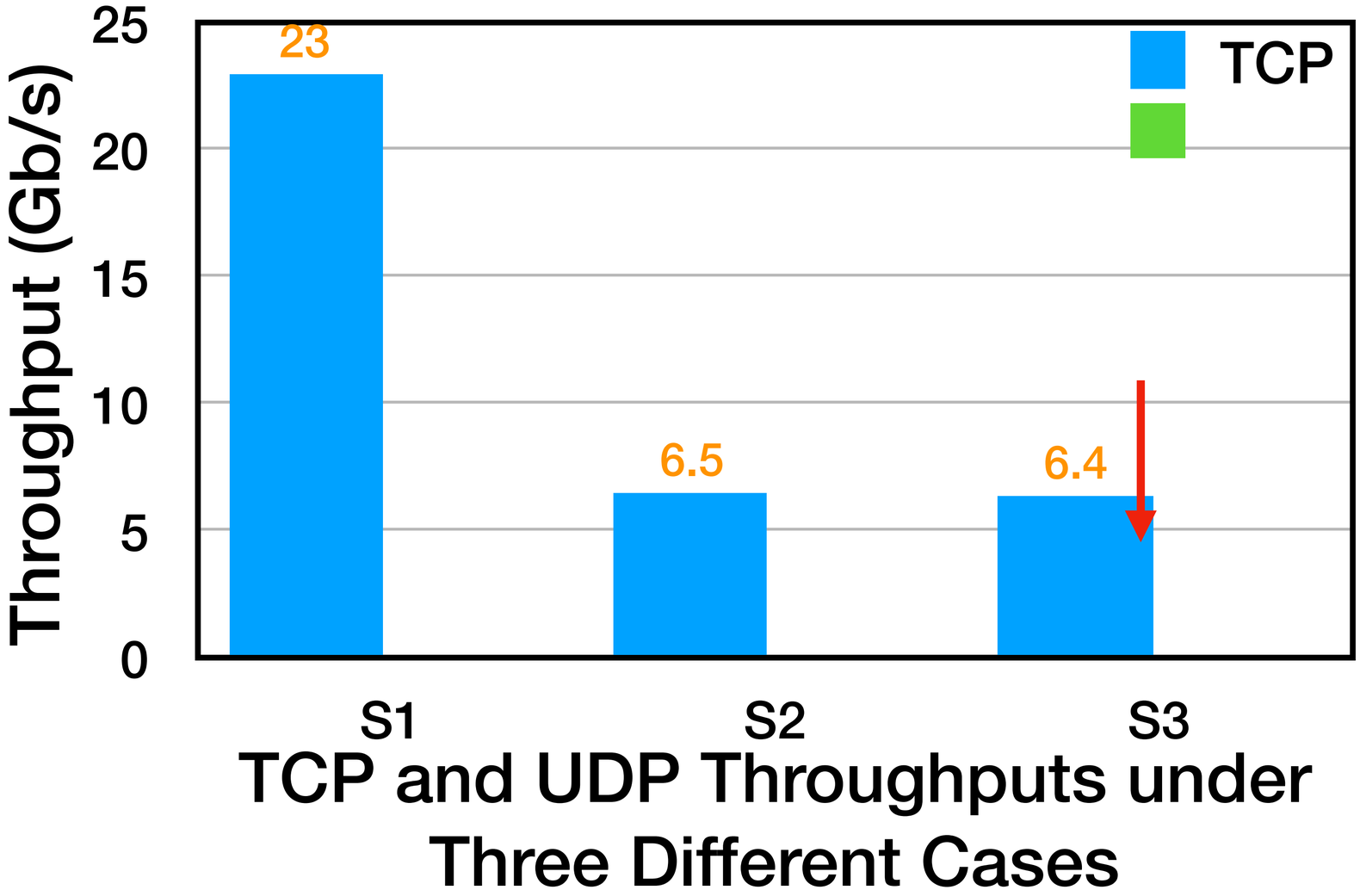
S2 - Linux Overlay



S3 - Docker Overlay

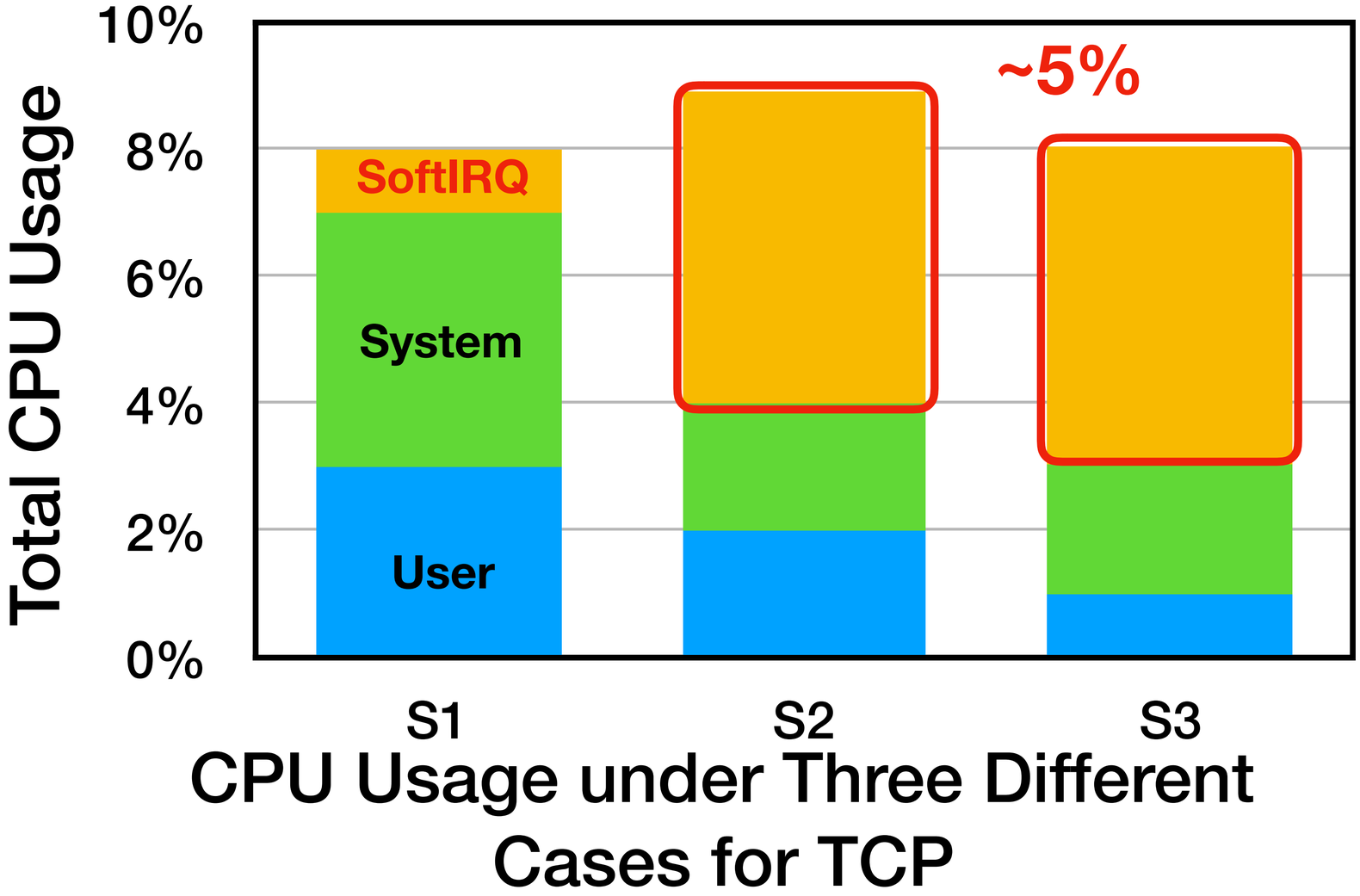


Single Flow Performance



- **TCP Throughput** of Docker Overlay case drops **72%** compared with native case.

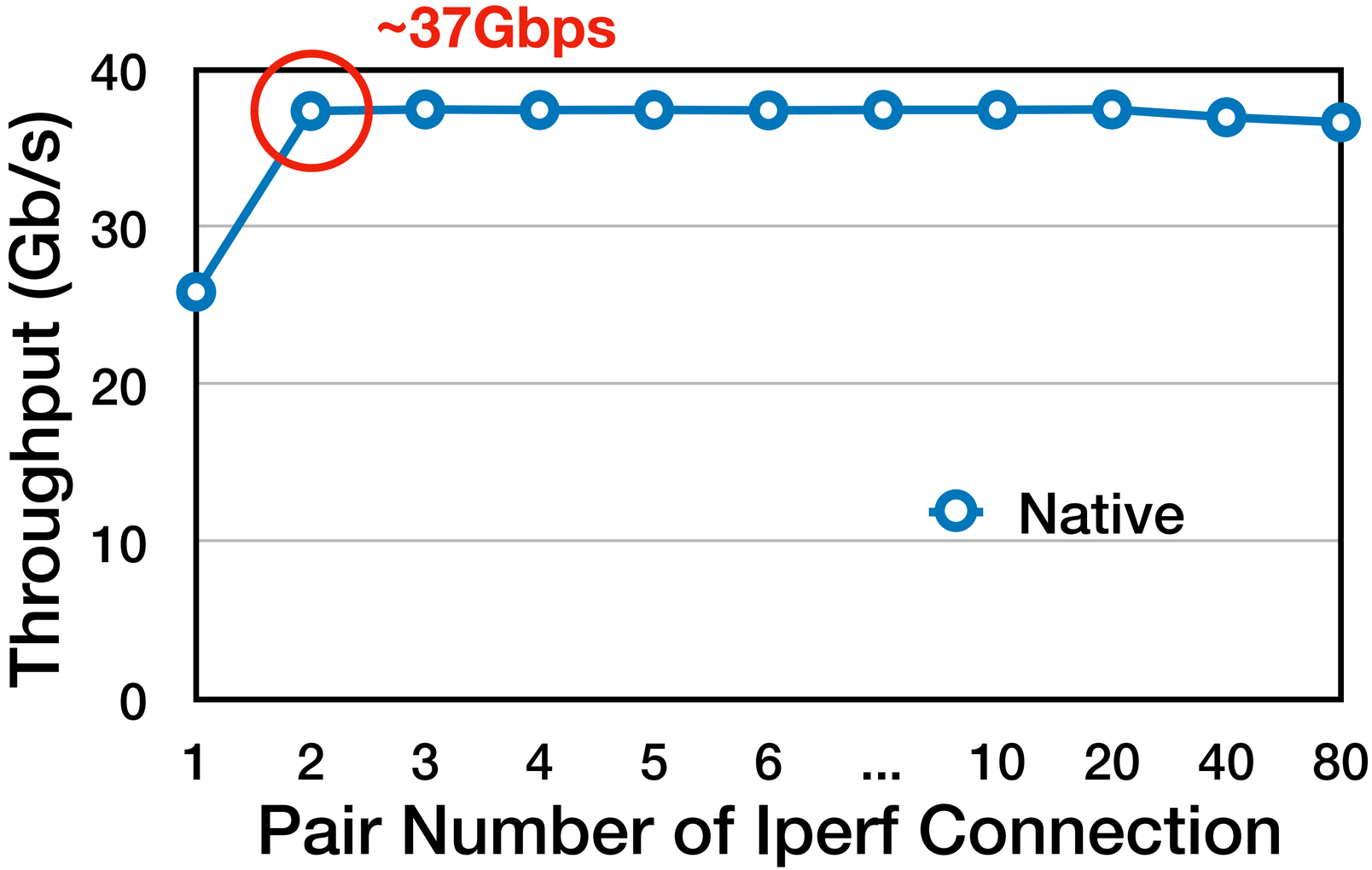
Single Flow Performance



- Packet processing overhead **fully saturates** one cpu core in two overlay cases.
- Current solutions **can't scale** single flow performance.

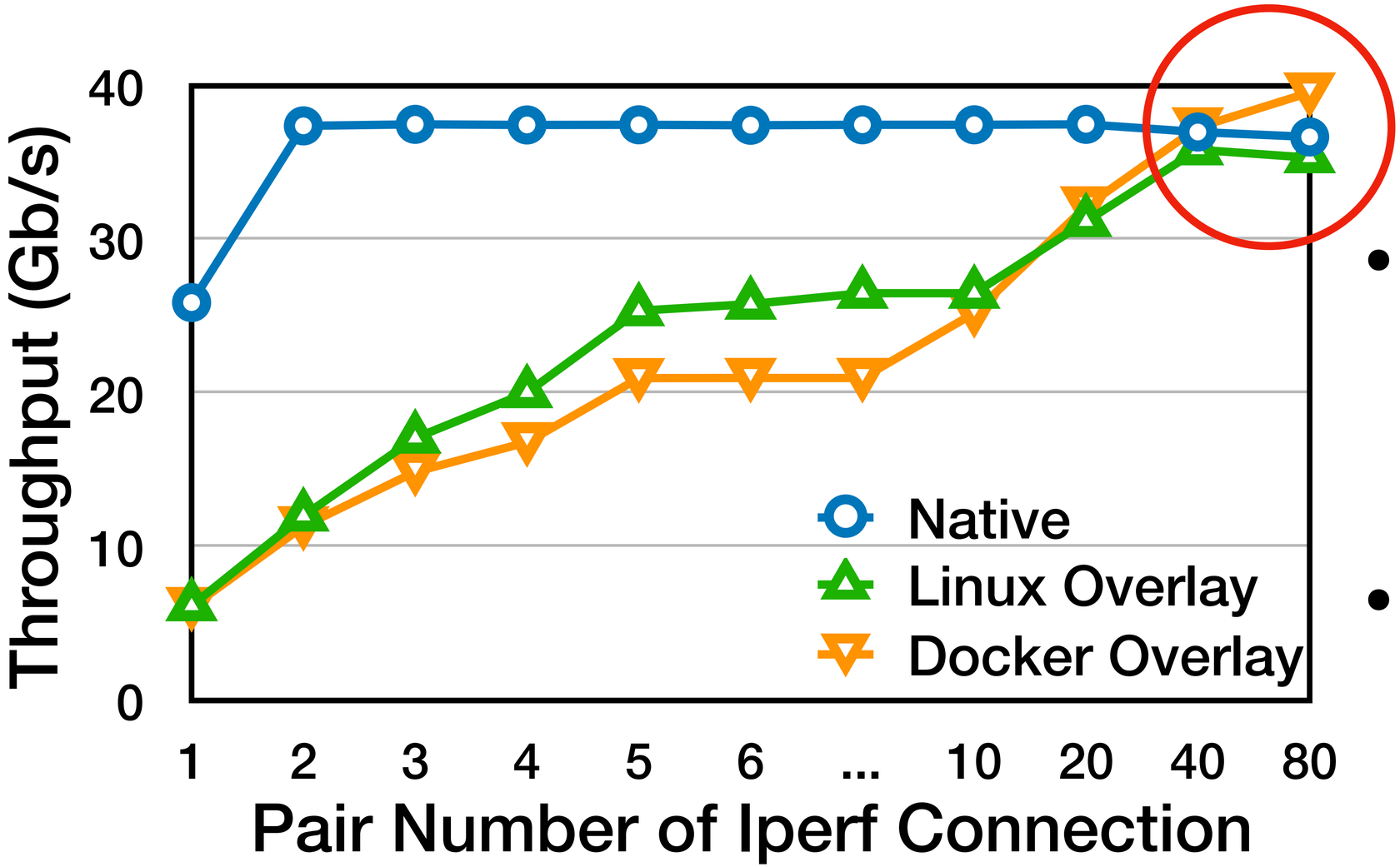
* 5% indicates one cpu core is fully saturated.

Multiple Flows Performance



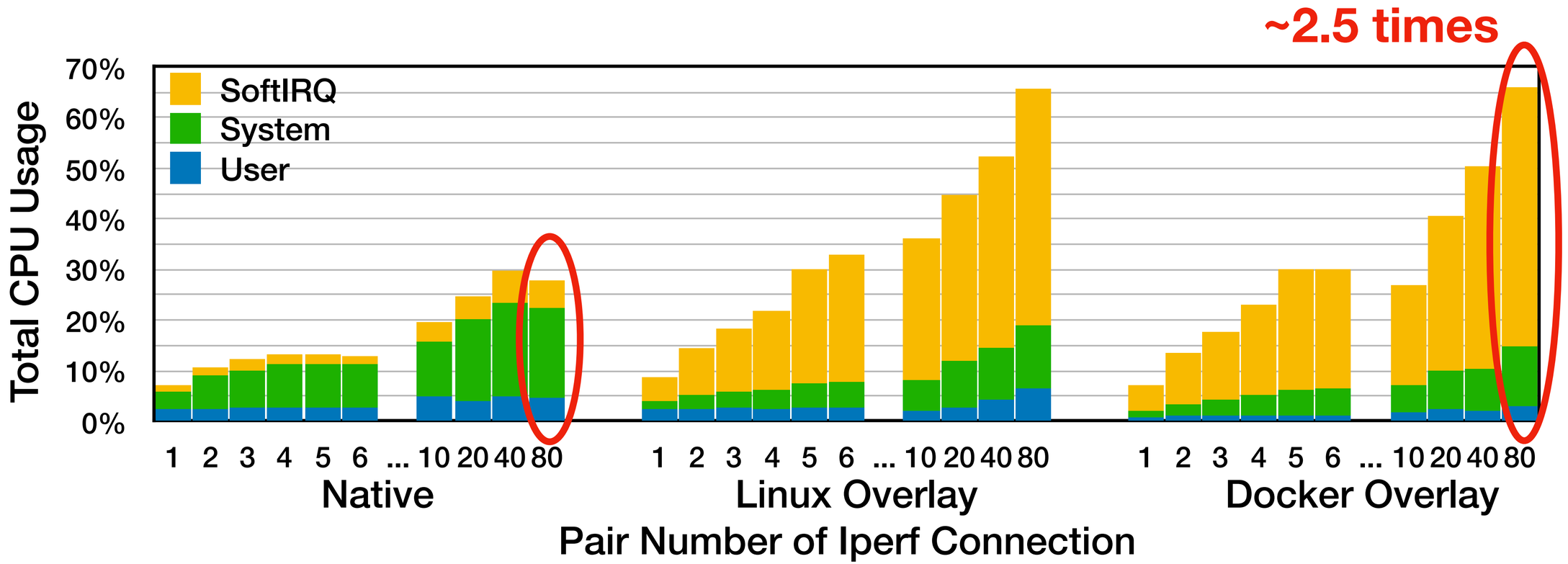
- Native case quickly reaches ~**37 Gbps** under TCP with only **2 pairs**.

Multiple Flows Performance



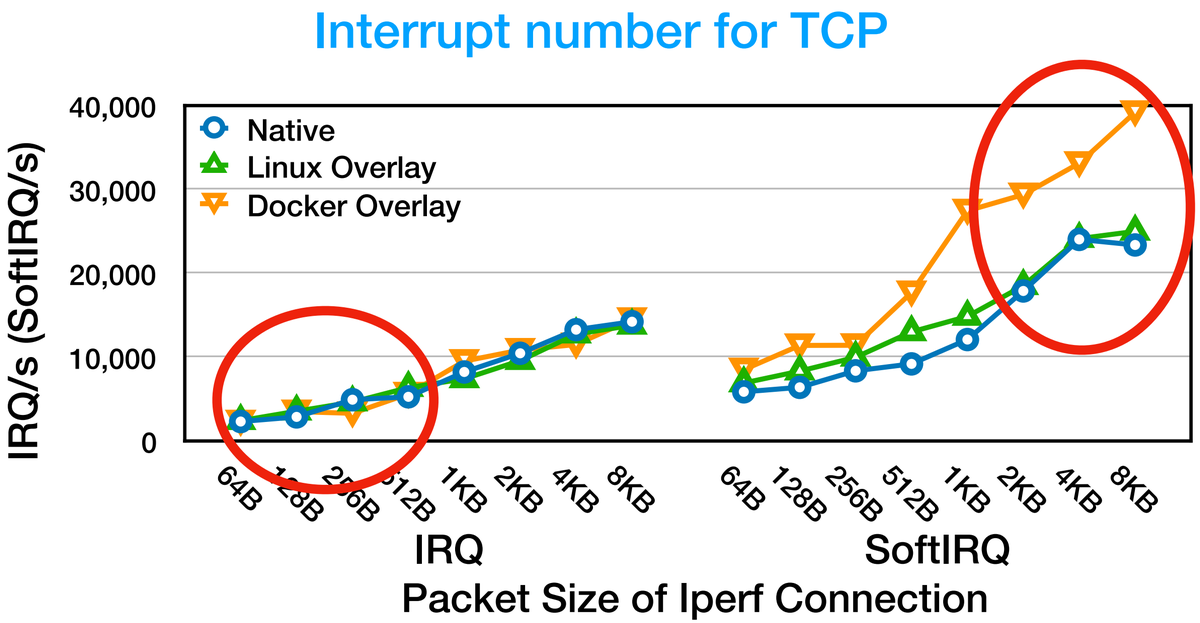
- Native case quickly reaches ~**37 Gbps** under TCP with only **2 pairs**.
- In two overlay cases, TCP throughput **grows slowly.**

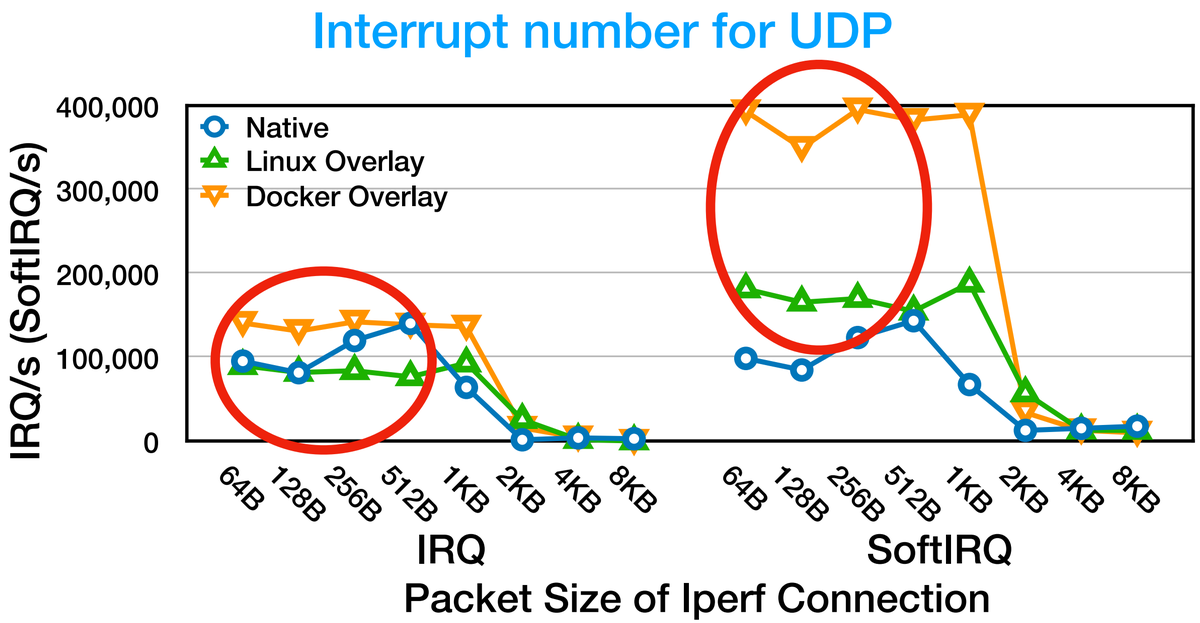
Multiple Flows Performance



- Under the same throughput (e.g., 40 Gbps), overlay networks consume **much more** CPU resources (e.g., around **2.5 times**) than the native case.

Interrupt Number with Varying Packet Sizes



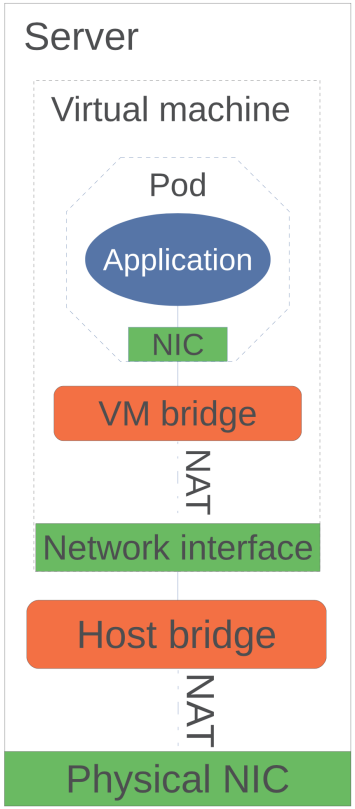


- IRQ number increases dramatically in the Docker overlay UDP case — **10x** of that in the TCP case.
- **3x** softIRQ numbers are observed in Docker Overlay case compared with the IRQ numbers.

Insights and Conclusions

- Kernel does not provide **per-packet level parallelization**.
- Kernel does not efficiently handle **various packet processing tasks**.
- Bottlenecks become more severe for **small packets**.

[ICIP19]: BrFusion



(a) Nested networking.

- Give each pod its own exclusive vNIC
- Created by hypervisor upon instruction of container orchestrator
- Enable to bypass: intermediate NAT+bridge+vNIC

[ICIP19]: BrFusion

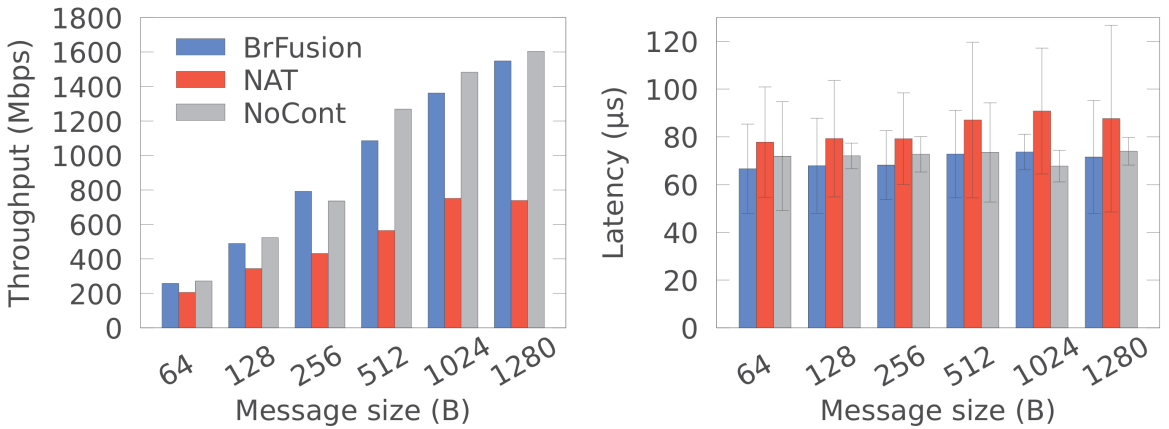


Figure 4: *BrFusion* performance gain using micro-benchmark. Bars are standard deviation of latency during the run.

- Higher throughput as compared to NAT
- Smaller latency as the chain of treatment is reduced