SDC
STORAGE DEVELOPER CONFERENCE
SNIA ■ SANTA CLARA, 2017

Maximizing Network Throughput for Container Based Storage

David Borman
Quantum

Agenda

- Assumptions
- Background Information
- Methods for External Access
  - Descriptions, Pros and Cons
- Summary

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Assumptions

- Services only run one instance
  - No load balancing to multiple back-ends
- External access to the service
  - Clients are not running on cluster nodes
- Throughput is the primary goal.

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Some of the Technology

- Docker
- Kubernetes (k8s)
- Ceph (and Rook)
- iSCSI

Some of the Pieces

- Docker containers
- Kubernetes pods
- Kubernetes deployments

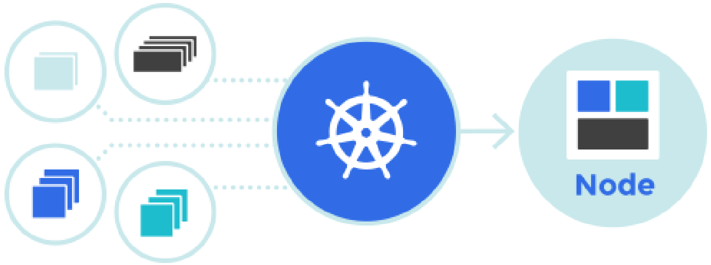
2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Docker

- "Docker is a software technology providing containers … [using] resource isolation features of the Linux kernel such as cgroups and kernel namespaces … [providing] resource limiting, including the CPU, memory, block I/O, and network." - https://en.wikipedia.org/wiki/Docker_(software)

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Kubernetes



- “Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications. It groups containers that make up an application into logical units for easy management and discovery.” - https://kubernetes.io

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Ceph

- “Ceph is a unified, distributed storage system designed for excellent performance, reliability and scalability.” - http://ceph.com
- “[It] implements object storage on a single distributed computer cluster, and provides interfaces for object-, block- and file-level storage” - https://en.wikipedia.org/wiki/Ceph_(software)

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Rook

- “Rook orchestrates battle-tested open-source storage technologies including Ceph … Rook is designed to run as a native Kubernetes service”

  \- https://rook.io

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

iSCSI Target implementations

- User Space
  - iscsi_tgt from the Intel SPDK
    - Using librbd and librados
- Kernel
  - Linux LIO
    - With kernel RBD module

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

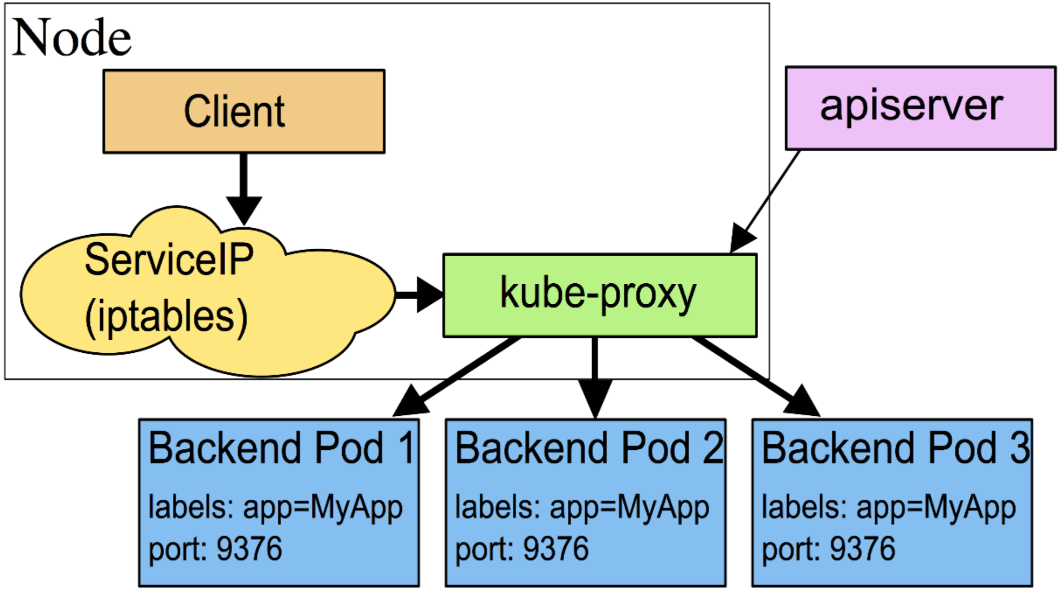
Kubernetes Service

- “A Kubernetes Service is an abstraction which defines a logical set of Pods and a policy by which to access them ... For non-native applications, [it] offers a virtual-IP-based bridge to Services which redirects to the backend Pods”

  - https://kubernetes.io/docs/concepts/services-networking/service/

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

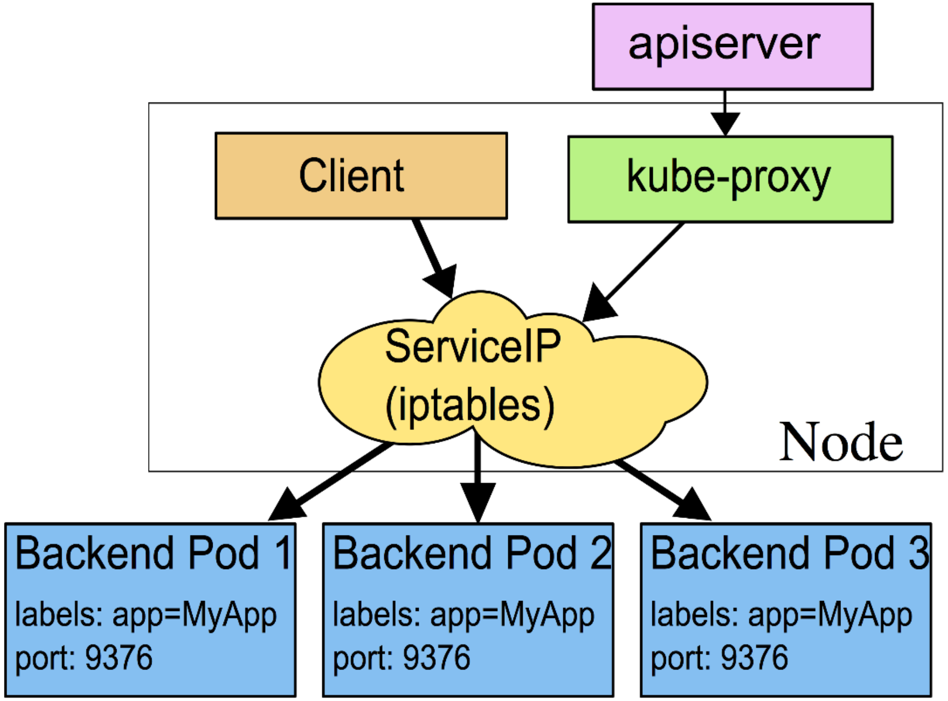
K8S Service: Proxy-mode: iptables



- https://kubernetes.io/docs/concepts/services-networking/service/

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

K8S Service: Proxy mode: User Space



- https://kubernetes.io/docs/concepts/services-networking/service/

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

iSCSI

- “… [an IP based] storage networking standard for linking data storage facilities. It provides block-level access to storage devices by carrying SCSI commands over a TCP/IP network.” - https://en.wikipedia.org/wiki/ISCSI

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Kubernetes pods with IP addresses

NAMESPACE	NAME	READY	STATUS	RESTARTS	AGE	IP	NODE
default	rook-operator-4134348477-hdjxh	1/1	Running	0	36m	10.2.79.2	172.17.4.202
kube-system	calico-node-5z9bm	2/2	Running	0	41m	172.17.4.201	172.17.4.201
kube-system	calico-node-gts4h	2/2	Running	0	39m	172.17.4.203	172.17.4.203
kube-system	calico-node-thwmw	2/2	Running	0	41m	172.17.4.101	172.17.4.101
kube-system	calico-node-v6vwv	2/2	Running	0	41m	172.17.4.202	172.17.4.202
kube-system	calico-policy-controller-7v74k	1/1	Running	0	41m	172.17.4.201	172.17.4.201
kube-system	heapster-v1.2.0-3863399399-32f04	2/2	Running	0	38m	10.2.69.4	172.17.4.201
kube-system	kube-apiserver-172.17.4.101	1/1	Running	0	41m	172.17.4.101	172.17.4.101
kube-system	kube-controller-manager-172.17.4.101	1/1	Running	0	40m	172.17.4.101	172.17.4.101
kube-system	kube-dns-1358247298-wl0mt	4/4	Running	0	41m	10.2.69.2	172.17.4.201
kube-system	kube-dns-autoscaler-2586315044-c8c48	1/1	Running	0	41m	10.2.69.3	172.17.4.201
kube-system	kube-proxy-172.17.4.101	1/1	Running	0	40m	172.17.4.101	172.17.4.101
kube-system	kube-proxy-172.17.4.201	1/1	Running	0	40m	172.17.4.201	172.17.4.201
kube-system	kube-proxy-172.17.4.202	1/1	Running	0	40m	172.17.4.202	172.17.4.202
kube-system	kube-proxy-172.17.4.203	1/1	Running	0	39m	172.17.4.203	172.17.4.203
kube-system	kube-scheduler-172.17.4.101	1/1	Running	0	40m	172.17.4.101	172.17.4.101
kube-system	kubernetes-dashboard-3619675109-q1kz0	1/1	Running	0	41m	10.2.79.5	172.17.4.202
rook	mon0	1/1	Running	0	36m	10.2.87.2	172.17.4.203
rook	mon1	1/1	Running	0	36m	10.2.87.3	172.17.4.203
rook	mon2	1/1	Running	0	36m	10.2.79.3	172.17.4.202
rook	osd-hwdr2	1/1	Running	0	35m	10.2.69.5	172.17.4.201
rook	osd-j6qj4	1/1	Running	0	35m	10.2.87.4	172.17.4.203
rook	osd-m71z8	1/1	Running	0	35m	10.2.15.2	172.17.4.101
rook	osd-td2lc	1/1	Running	0	35m	10.2.79.4	172.17.4.202
rook	rook-api-3722659863-53d33	1/1	Running	0	35m	10.2.87.5	172.17.4.203

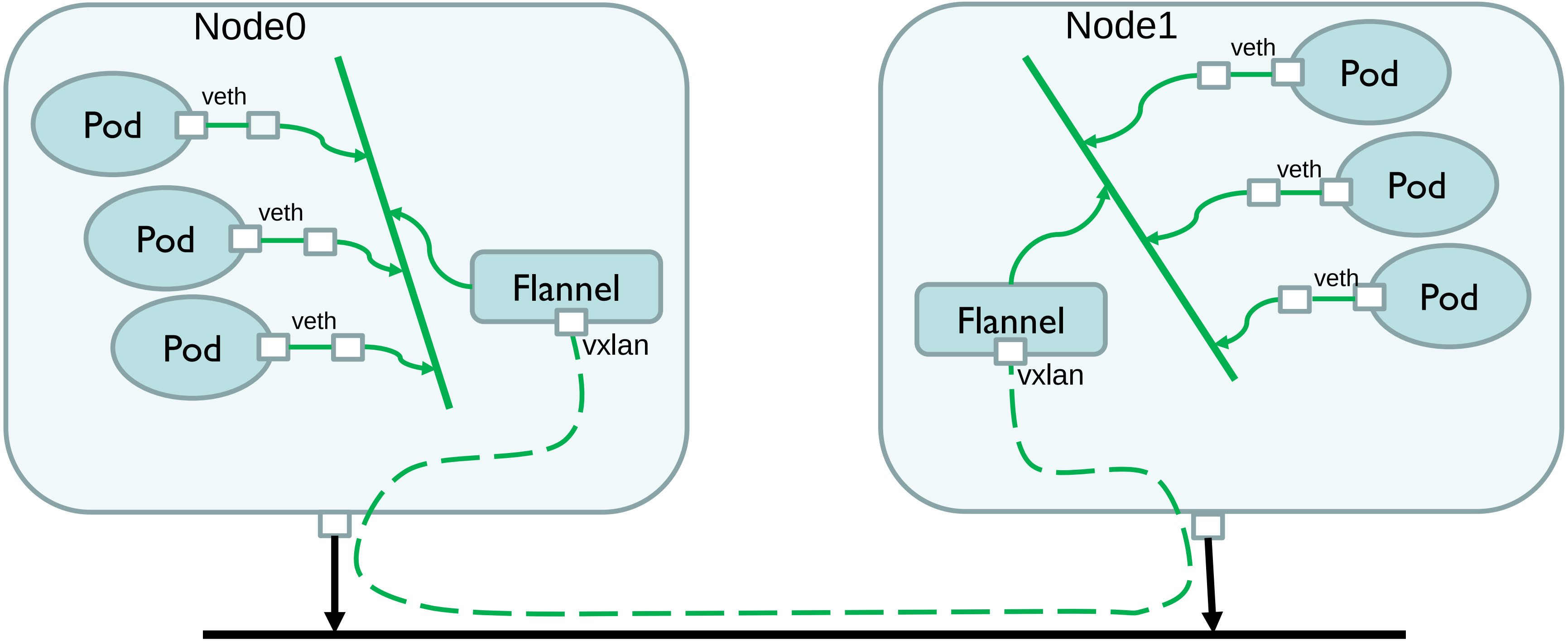
2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Kubernetes pods with IP addresses

NAMESPACE	NAME	IP	NODE
kube-system	calico-node-gts4h	172.17.4.203	172.17.4.203
kube-system	kube-proxy-172.17.4.203	172.17.4.203	172.17.4.203
rook	mon0	10.2.87.2	172.17.4.203
rook	mon1	10.2.87.3	172.17.4.203
rook	osd-j6qj4	10.2.87.4	172.17.4.203
rook	rook-api-3722659863-53d33	10.2.87.5	172.17.4.203

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Flannel and Calico



2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Methods for External Access to Pods

- Kubernetes Service
  - Using Type=NodePort
- Host Networking
  - Kubernetes Pod: hostNetwork: true
- Kernel Bypass
  - Virtual NIC via Intel SPDK

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Kubernetes Service - NodePort

- Kube-proxy runs on each node
- Packets are routed internally to the correct node where the pod is running

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Kubernetes Service - NodePort- Pros

- Pods can be reached through any node
- Pods can restart, on the same or a different node, and still be reachable at the same IP/TCP address

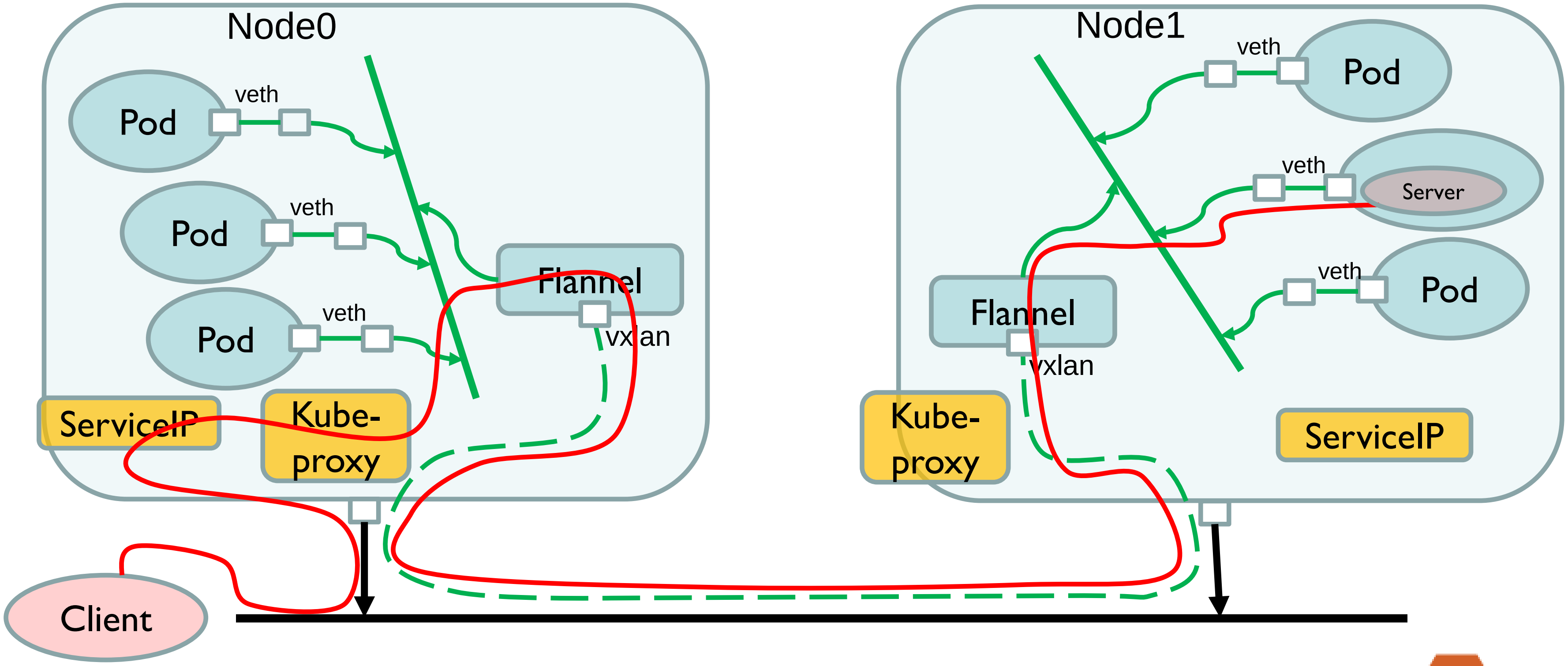
2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Kubernetes Service - NodePort- Cons

- Traffic is NATed (source IP, dest TCP port)
- Traffic sent to wrong node is forwarded
  - And encapsulated over vxlan
- service.spec.externalTrafficPolicy=Local
  - Avoids source IP NAT, but…
  - Traffic sent to the wrong node is dropped

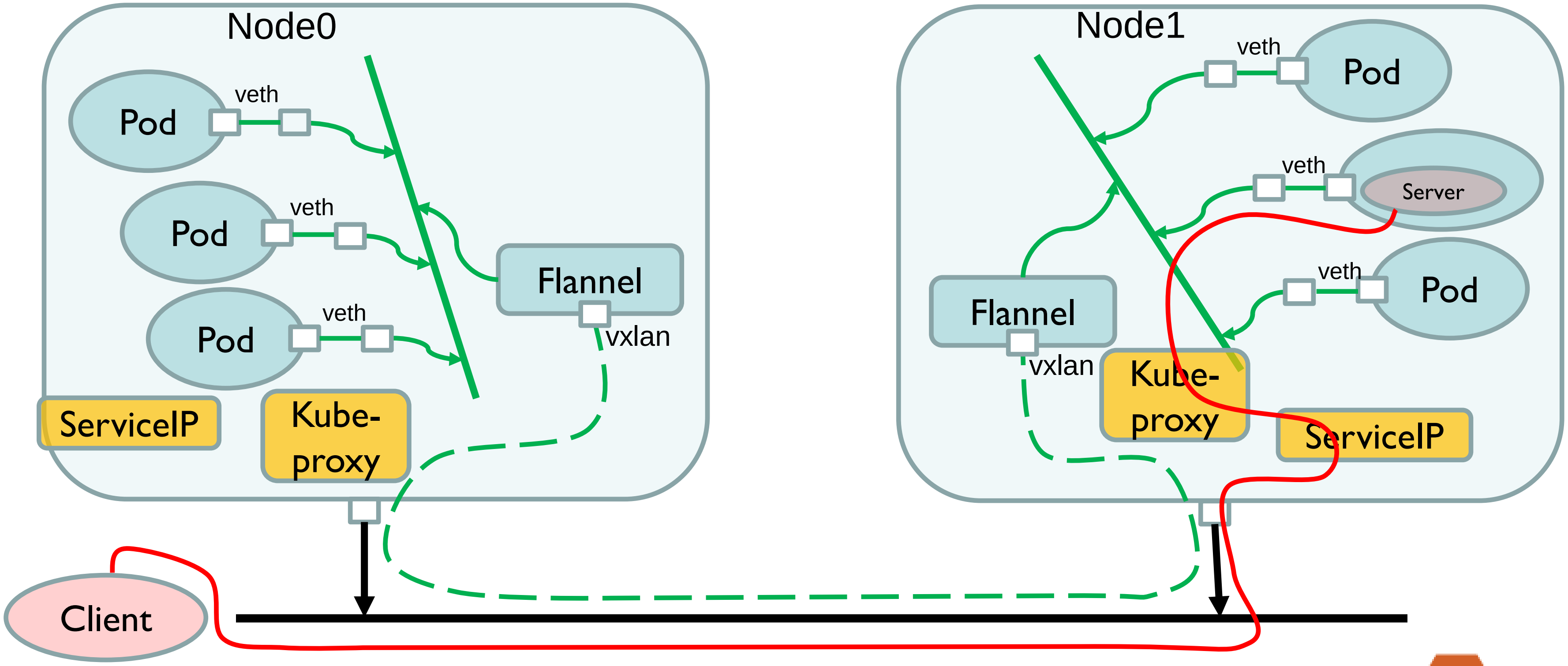
2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Client Connects to Wrong Node



2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Client Connects to Correct Node



2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Host Networking - Pros

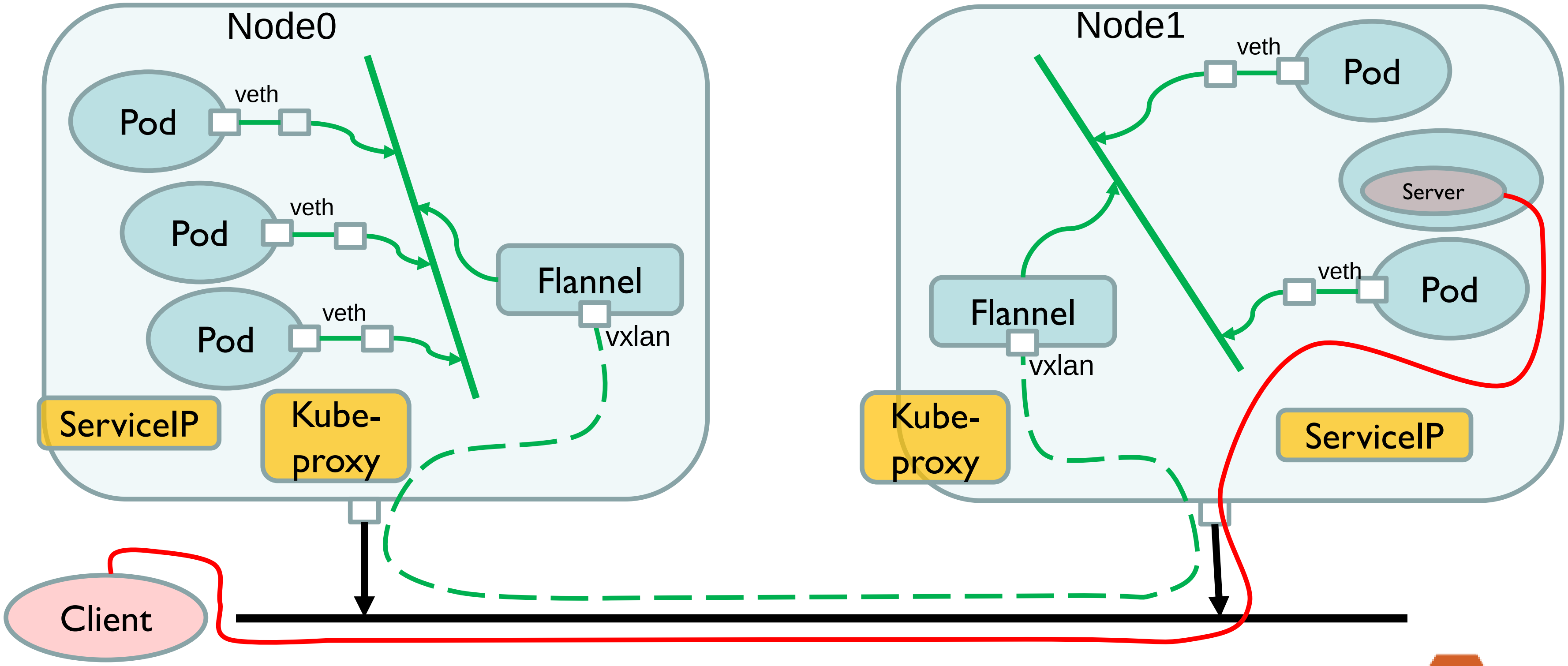
- Exposes the host's interfaces inside the pod
- No NAT
- Traffic goes directly to the correct node

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Host Networking - Cons

- The IP address changes when a pod restarts on a new node

Client Connects with Host Networking



2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Adding keepalived to Host Networking

- K8S contrib project: kube-keepalived-vip
- Manages external Virtual IP (VIP) addresses
- Uses K8S Daemonsets to run on every node
- If a node crashes, VIP is moved to a new node

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Adding keepalived - Pros

- If a node crashes, VIP is moved to a new node
  - Connectivity is not tied to any individual node
- External clients only need to know the VIP

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Adding keepalived- Cons

- Pod must be constrained to node with the VIP
- VIPs are managed outside of K8S networking

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Kernel Bypass

- A virtual NIC is created at the hardware level, and the application manages it directly.

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Kernel Bypass - Pros

- Application has direct access to a virtual NIC
- Higher performance
- IP address moves with App

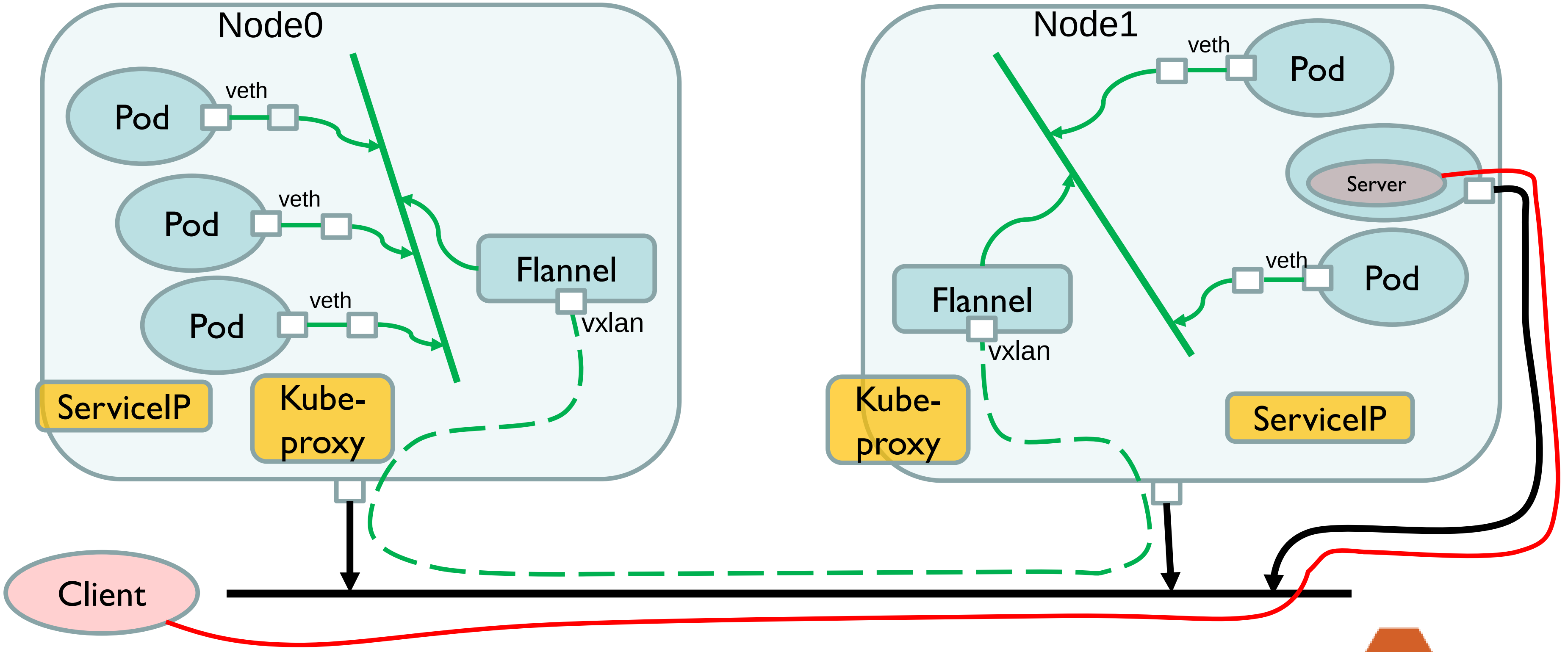
2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Kernel Bypass - Cons

- Application must manage IP address
- Application needs a TCP/IP implementation
- Requires NIC support

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Client Connects via Kernel Bypass



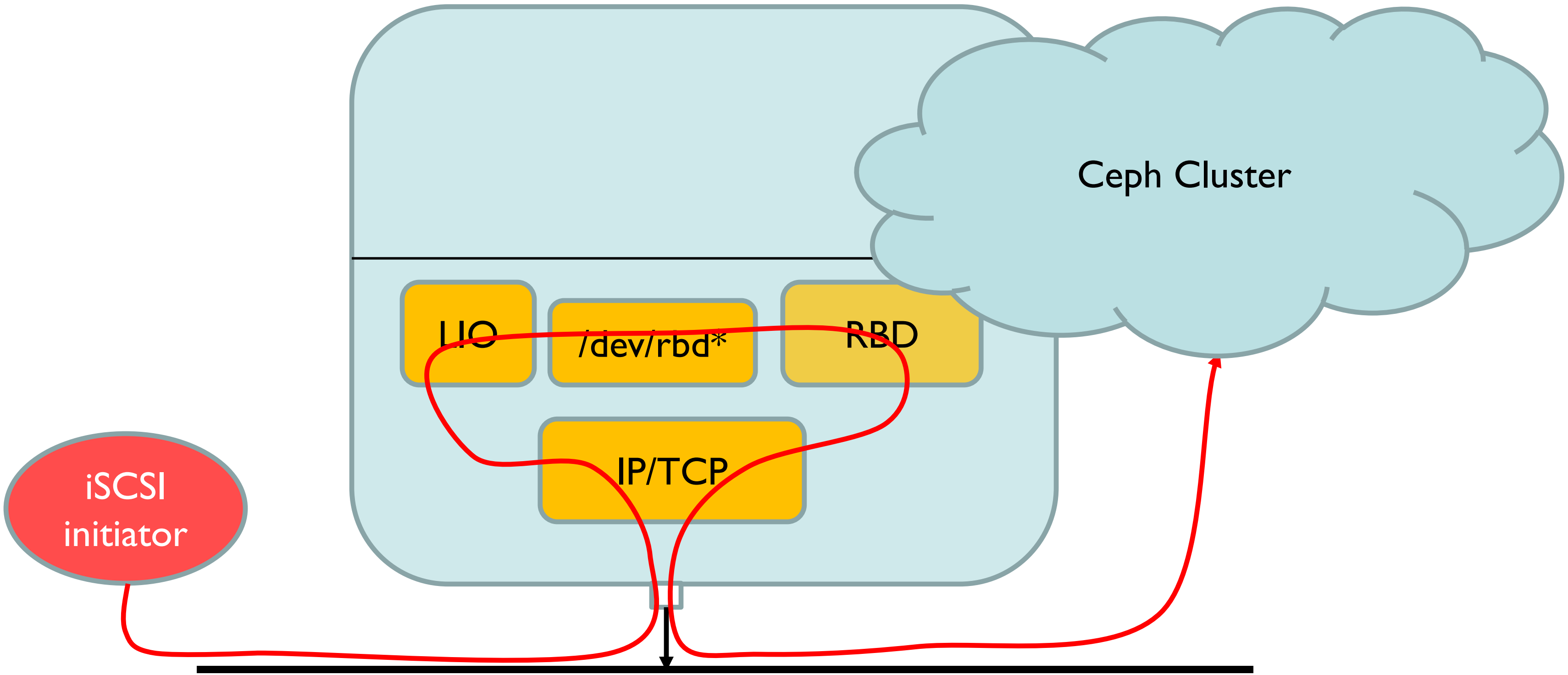
2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

iSCSI Target Backed by Ceph RBD Image

- Export access to Ceph RBD images
  - To clients that are not part of the Ceph cluster
- iSCSI Initiator (client side) is widely implemented
- Options for kernel or user level implementations

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Data path: Linux LIO iSCSI target



2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Kernel Level iSCSI Target - Pros

- Data path is only in the kernel
- No polling needed – interrupt driven
- Uses host networking (or VIPs)
- K8S pod is just for configuration
- No user level memory issues

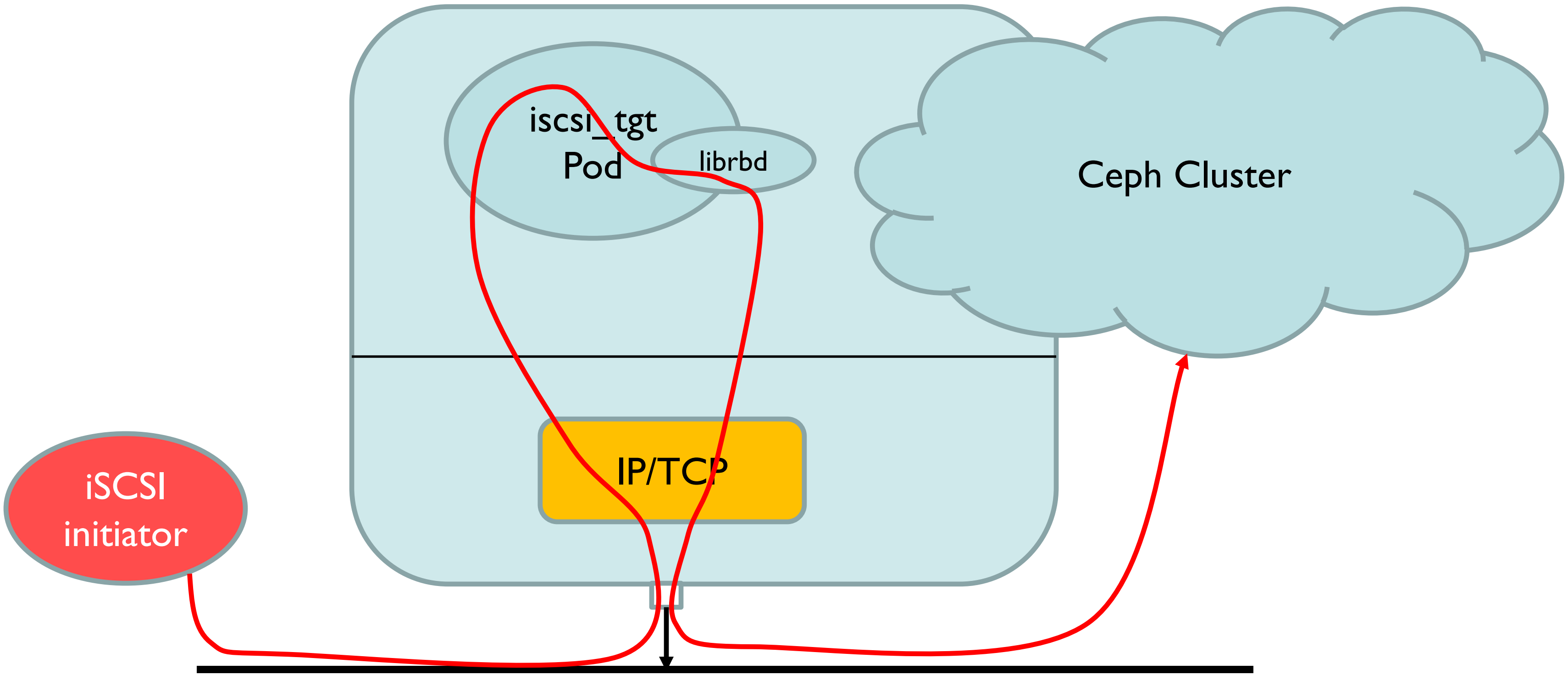
2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Kernel Level iSCSI Target - Cons

- Dependent on kernel modules being loaded
- Multiple targets on the same node could be more difficult to configure (from multiple pods)

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Data path: Intel SPDK iSCSI target



2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

User Level iSCSI Target - Pros

- Not dependent on kernel modules
- Multiple pods can be running on the same node
- Easy to update the container

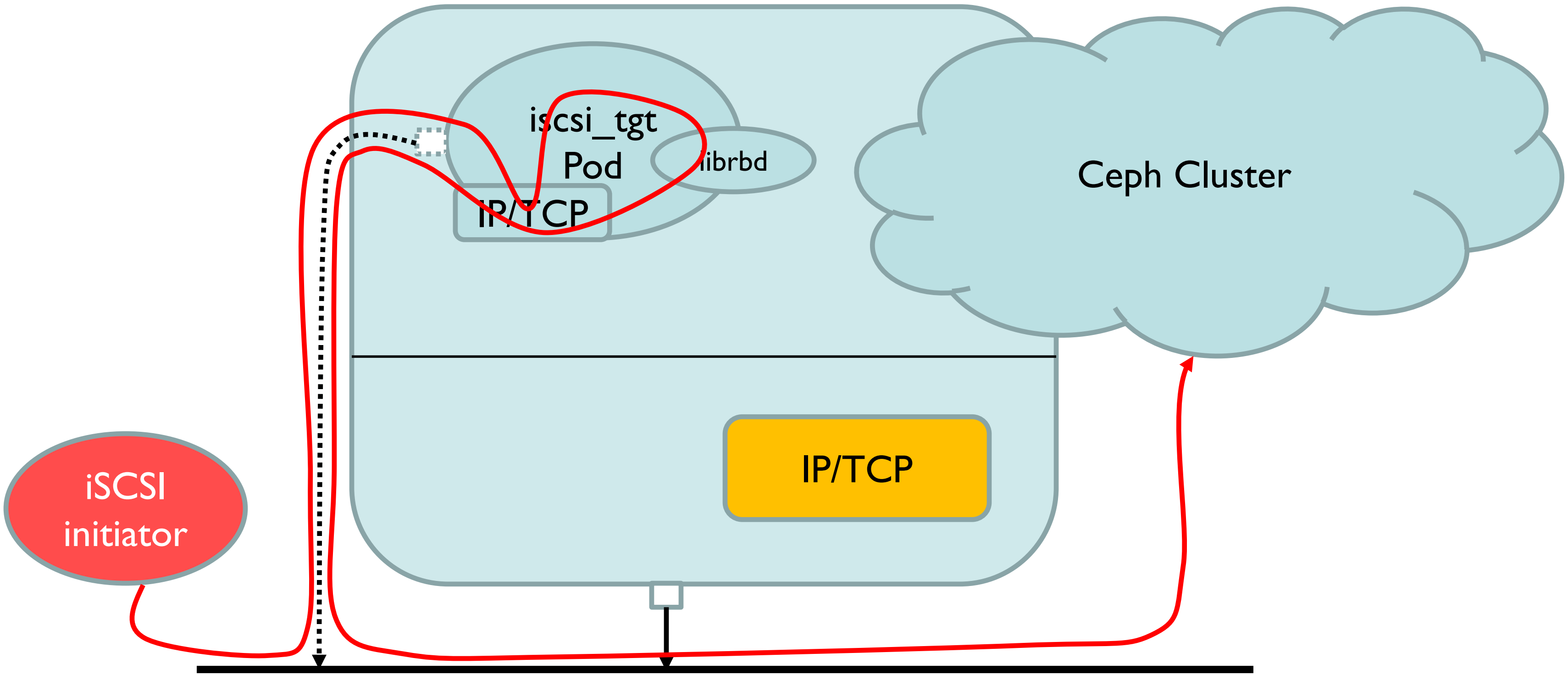
2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

User Level iSCSI Target - Cons

- CPU/memory Usage
  - Linux hugepages
  - Polling for data

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Data path: Intel SPDK with Virtual NIC



2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Virtual NIC - Pros

- Data goes directly to the application
- No transitions to kernel space
- Application has its own IP address

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Virtual NIC - Cons

- Application needs IP/TCP library
- Application must manage its own IP address
- CPU/memory usage

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Summary

- We looked at three options for external access:
  - K8S Service, Host Networking, Kernel Bypass
- The best way to reduce latency is to eliminate packet forwarding, encapsulation and NAT
- The semantics of a particular service will often drive the choice

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Q & A

- Questions?

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Resources

- https://rook.io
- http://ceph.com
- https://kubernetes.io
- https://www.docker.com
- http://www.spdk.io
- http://linux-iscsi.org

2017 Storage Developer Conference. © Quantum Corporation. All Rights Reserved.

Resources (continued)

- https://github.com/kubernetes/contrib/tree/master/keepalived-vip