A Study on Portable Load Balancer for Container Clusters

by

Kimitoshi Takahashi

Dissertation

submitted to the Department of Informatics

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

The Graduate University for Advanced Studies (SOKENDAI)

March 2019

Committee

Kento Aida(Chair)	National Institute of Informatics / Sokendai
Atsuko Takefusa	National Institute of Informatics / Sokendai
Michihiro Koibuchi	National Institute of Informatics / Sokendai
Takashi Kurimoto	National Institute of Informatics / Sokendai
Atsuko Takefusa	National Institute of Informatics / Sokendai
Shigetoshi Yokoyama	National Institute of Informatics / Gunma University

Acknowledgments

[Filled in later]

Abstract

Today a vast majority of the people in the world use PCs or smartphones to communicate with friends, check up the news, watch the videos, play games, etc., through the Internet. These services are called web services because they utilize web technology through HTTP(S) protocols. Web services are generally provided by a cluster of web server programs, database server programs, and load balancers. Web service providers deploy these programs on a cluster of physical servers in an on-premise data center or on a cluster of VMs in cloud infrastructures.

Recently Linux container technology and clusters of the containers have come to draw attention because they are expected to make web services consisting of multiple web servers and a load balancer portable, and thus realize easy migration of web services across the different cloud providers and on-premise data centers. Service migrations prevent a service to be locked-in a single cloud provider or a single location and enable users to meet their business needs, e.g., preparing for a natural disaster, lower the cost of infrastructure and comply the regulations.

In order for a web service to be deployed easily in different base infrastructures, container management systems are often used. However existing container management systems lack the generic capability to route the traffic from the internet into web service container clusters. For example, Kubernetes, which is one of the most popular container management systems, is heavily dependent on cloud load balancers. If users use unsupported base infrastructures, it becomes users responsibility to route the traffic into their cluster while keeping the redundancy and scalability. This means that users are happy only in the major cloud providers including GCP, AWS, and Azure; thus they could easily be locked-in those infrastructures.

In this dissertation, the author proposes a load balancer architecture that is usable

in any of the base infrastructure, including cloud providers and on-premise data centers, in order to free users from lock-ins. The proposed load balancer architecture utilizes software load balancers with container technology to make the load balancers runnable in any base infrastructure. It also utilizes ECMP technology to make multiple load balancers active, and thereby to provide redundancy and scalability.

The author implemented a containerized software load balancer that is run by Kubernetes as a part of container cluster, using Linux kernel's IPVS. In order to discuss the feasibility of the proposed load balancer, performance measurements are conducted in 1 Gbps network environment. It was shown that the proposed load balancers are runnable in an on-premise data center, GCP and AWS. It can be said that the proposed load balancers are portable. The throughput levels of a load balancer are dependent on settings for multi-core packet processing. It was shown to be better to use as many CPU cores as possible for packet processing. The throughput levels are also very dependent on the overlay network backend mode and overhead of the container network, i.e., veth+bridge. The host-gw mode where no tunneling is used resulted in the best performance level, and the vxlan mode resulted in the second best. Although the overheads of the container network are invisible in 1 Gbps network environment, they are visible in 10 Gbps network environment. In the experiment in 1 Gbps network environment, the ipvs-nat load balancer in the container had the same performance level as load balancing function of iptables DNAT in the node net namespace. Furthermore, the performance level of ipvs-tun load balancer in a container with the L3DSR setup was about 1.5 times larger than that of iptables DNAT. Therefore in 1 Gbps network environment, the proposed load balancer is portable while it has the 1.5 times better performance level or the same performance level depending on the mode of operation.

Also implemented is the ECMP setups where multiple of the load balancer containers are deployed, each advertising the route to the service VIP. The ECMP technique makes the load balancers redundant and scalable since all the load balancer containers act as active. The whole system is resilient to a single failure of load balancer container. Also by utilizing multiple of load balancers simultaneously, the throughput of the total system is increased significantly. These characteristics are evaluated by checking the routing table of the upstream router and throughput measurement. The author verified that ECMP routing table was properly created in the experimental system.

The update of the ECMP routing table was correct and quick enough, i.e., within 10 seconds, throughout 20 hours experiment. The maximum performance levels of the cluster of load balancers scaled linearly as the number of the load balancer pods was increased up to four of them.

The author also extended the throughput measurement into 10 Gbps network environment. It was revealed that ipvs-nat and ipvs-tun load balancers in containers had lower performance levels compared with the iptables DNAT. By setting up the load balancing table in node net namespaces, the performance levels of ipvs-nat and ipvs-tun became closer to that of the iptables DNAT, which is suggesting that the overhead of the container network is no longer invisible in 10 Gbps network environment. The author is currently implementing and evaluating a novel software load balancer using XDP technology to provide a better alternative to ipvs as a portable load balancer.

The outcome of this study will benefit users who want to deploy their web services on any cloud provider where no scalable load balancer is provided, to achieve high scalability. Moreover, the result of this study will potentially benefit users who want to use a group of different cloud providers and on-premise data centers across the globe seamlessly. In other words, users will become being able to deploy a complex web service on aggregated computing resources on the earth, as if they were starting a single process on a single computer.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

1.1.1 Web application cluster

Today, a great number of people in the world can not spend a day without using smartphones or personal computers(PCs) to retrieve information from the Internet for work or for daily life. For example, people use these devices to look up web pages, emails, social media and sometimes to play games. These services are often called web applications or web services, where information is delivered using Hyper Text Transfer Protocols(HTTP) or Hypertext Transfer Protocol Secure (HTTPS) from servers at the other end of the Internet. Web applications are provided by various organizations, including commercial companies, government, non-profitable organizations, etc.

For example, Google provides a variety of web services including, Gmail, Search engine, Google Suits, etc. Facebook provides social media service, Amazon provides shopping sites. Governments provides information regarding the service they provide

to their citizens. Schools often provide a syllabus to their students, which is important for campus life. The author calls those organizations that provide web applications, web application providers hereafter.

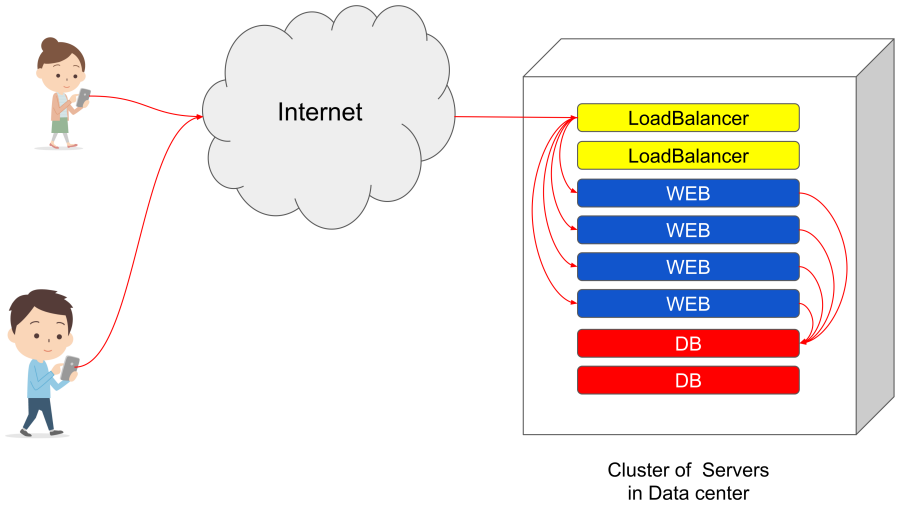


Figure 1.1: An example of web application cluster.
The load balancers distribute requests from clients to multiple web servers. The web servers form responses using data retrieved from the database servers and send it back to the clients. Sometimes the web servers also store and update important data into the database servers.

A client program on PCs or smartphone sends out requests to servers and the servers respond with data that is requested, using HTTP or HTTPS. Servers for web applications are usually computers located in a data center. In the data center multiple servers cooperate to fulfill the need of the clients. A group of these servers is often called a web application cluster or a web cluster. Figure 1.1 shows schematic diagram of an example of a web application cluster.

In this example, there are two load balancers, four web servers and two database(DB) servers that work together to respond to the requests from clients. The load balancers distribute the requests from clients to multiple web servers. Then the web servers form responses using data retrieved from the database servers and send it back to the clients. Sometimes the web servers also store and update important data into the database servers.

1.1.2 Migration of web application cluster

As web applications become an essential part of daily life, an outage of the web application service is getting to be a bigger problem. If something happens to a web application cluster in a data center, people will not be able to access the necessary information.

For example, if web pages run by local government stops, people will not be able to access the information regarding public service. If a shopping site run by a company stops, customers can no longer buy products and the revenue of the company will be negatively affected. Outages of web applications by giant companies can have an even bigger impact. An outage of Gmail or Google search engine will probably stop most of the business activities around the world. Service down of Amazon.com affect buyers and many businesses that sell products on its platform.

In order to prevent such outages, preparing another web application cluster in a different location in the case for disasters is very important. For that purpose, it is desirable if a web application cluster can be easily migrated to a different data center. Migration of a web application cluster becomes more realistic with the use of Linux container technology, which is explained later.

Migration capability of a web application cluster also has other benefits. If an e-commerce service is successful in Japan, the company that runs the service might want to start the same service in other country, for examlpe, in Europe. In this situation, the company probably wants to migrate its web applictation cluster to somewhere in Europe, because, for Europian customers, responses from a web site in Europe is quicker than that from a web site in Japan.

Being able to migrate a web application cluster is very important for a variety of purposes, including disaster recovery, cost performance optimizations, meeting legal compliance and shortening the geographical distance to customers. These are the main concerns for web application providers in e-commerce, gaming, Financial technology(Fintech) and Internet of Things(IoT) field. Therefore it is important for a web application provider to be able to easily deploy and migrate their web applications among different infrastructure around the world. The purpose of this research is to propose infrastructures for web application providers, where they can easily deploy their services across the world, regardless of cloud providers or data centers they use.

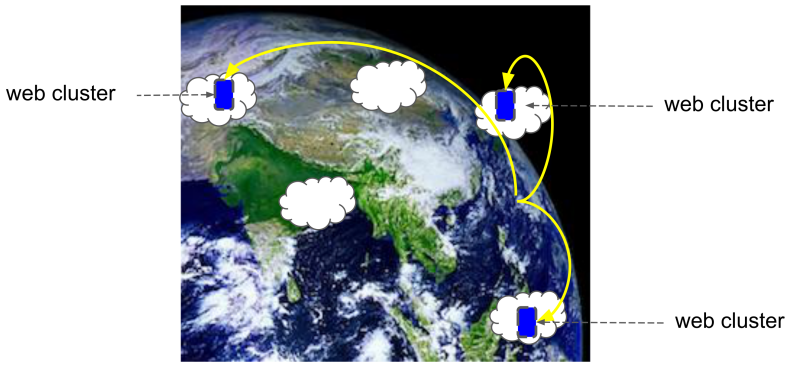


Figure 1.2: Migration of web application cluster to different locations.
It is desirable to be able to migrate a web cluster from one place to another with the easiness of on push button.

1.1.3 Ideal infrastructure for migration of web application

In order to realize an easy migration of web applications, an ideal infrastructure probably have the following features; 1) have universal middleware to manage web clusters, 2) store data in globally consistent data storage, 3) route global traffic based on proximity to the client. Figure1.3 shows an exemplified global container infrastructure having these features. Container orchestrators launch and manage container clusters. Important data are stored in globally consistent data storage, which is simillar to the Google spanner[8, 7] or CockroachDB[32]. Traffic is routed to the closest data center using anycast[28].

Each of these features is important and research efforts are on-going in many institutions. In this study, the author focuses on the research regarding container orchestrator as a universal middleware. By realizing global container infrastructure web application providers will be able to deploy their web applications whenever and wherever they want. Also, they will be able to move their web applications quickly depending on a variety of circumstances, including disaster recovery, cost performance optimization and compliance to government regulations due to trade wars, etc.

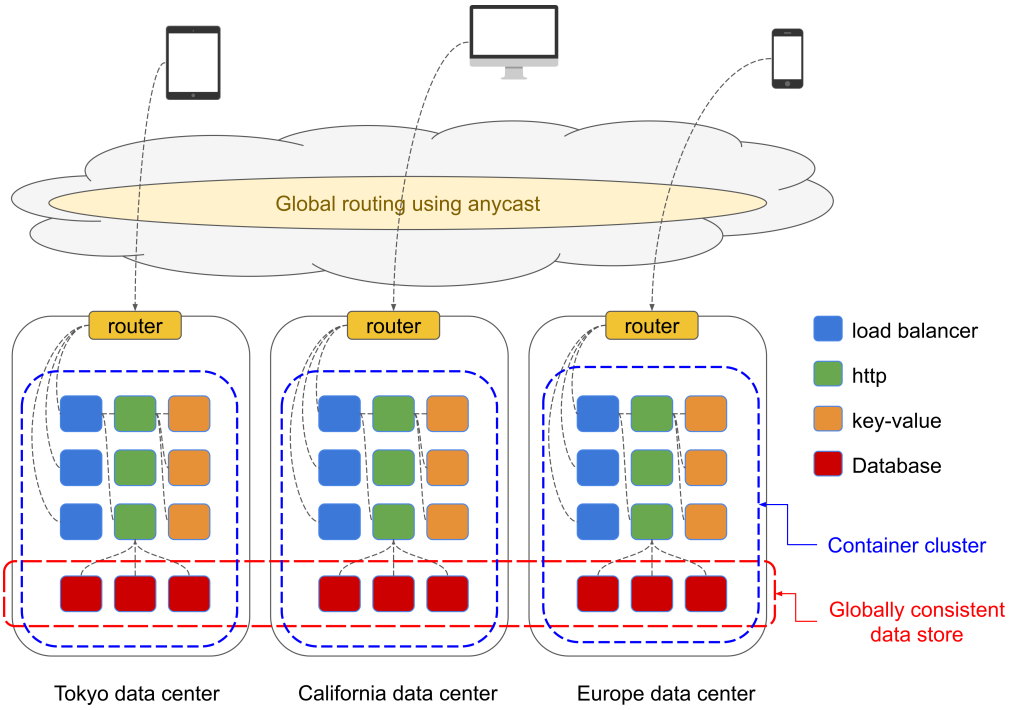


Figure 1.3: An ideal global container infrastructure.
Multiple web application clusters, each of which consisting of a cluster of containers, are deployed in three different data centers as an example. In each of the data center, container orchestrator manages the container cluster. Important data are stored in a globally consistent data store. Access from the client is routed to the closest data center using anycast.

1.2 Inftastructure for web applications

1.2.1 On-premise data center

Historically, most of the web application providers purchased servers and installed them in server housing facilities called data centers. In this type of infrastructure, web application providers typically need to sign a contract with data center company for server housing rack spaces, buy servers and install them in their rented racks by themselves. They also install OS and software stacks needed to run their web applications in the servers. Since web application providers place servers in their facilities(either owned or rented), and they are responsible for managing the servers, this type of infrastructure is often called on-premise infrastructure in contrast to Cloud

Computing infrastructure.

Preparing data centers, installing the servers and configuring software stacks for web application services often require a considerable amount of time, money and effort. If web application providers want to expand their services to different countries or if they want to prepare for natural disasters by preparing an additional web application cluster in a different data center, they probably need about the same amount of time, money and effort required to build their original infrastructures. Therefore migration of web application in this type of infrastructures has always been a daunting task.

1.2.2 Cloud computing

The emergence of Cloud Computing made many things easier for web application providers than before. Cloud computing utilizes a virtual machine(VM) technology, e.g. KVM, Xen, and VMware. Cloud computing service providers offer VMs to web application providers with pay-per-use billings.

Figure 1.4 compares different type of usages of a single physical server and Figure 1.4 (b) shows an example architecture of VM technology. VMs share a single physical server. A full OS including Linux kernel is running on top of the virtual machine represented by the hypervisor. Each VM behaves almost as same as a single physical server. Since VMs are fractions of a single physical server, server resources are utilized with finer granularities. Web application providers can start their services with a cluster of VMs, which is smaller than a cluster of physical servers, and hence resulting in lower cost.

Cloud providers generally prepare physical servers and software stacks for VMs before renting it to users, and they also provide an easy to use web user interfaces. As a result, users only need to click a few buttons on web browsers and wait for a few minutes before obtaining up-and-running VMs. This simplicity will bring agility to web application providers when they launch their services. And since computing resources are offered with per-second pay-per-use billings, web application providers can quickly reduce the cost by stopping excessive VMs, when the demand for computing power is scarce. This was impossible when web application providers purchased physical servers and used them as bare metal servers. In short, cloud computing brought users agility, flexibility, and cost-effectiveness.

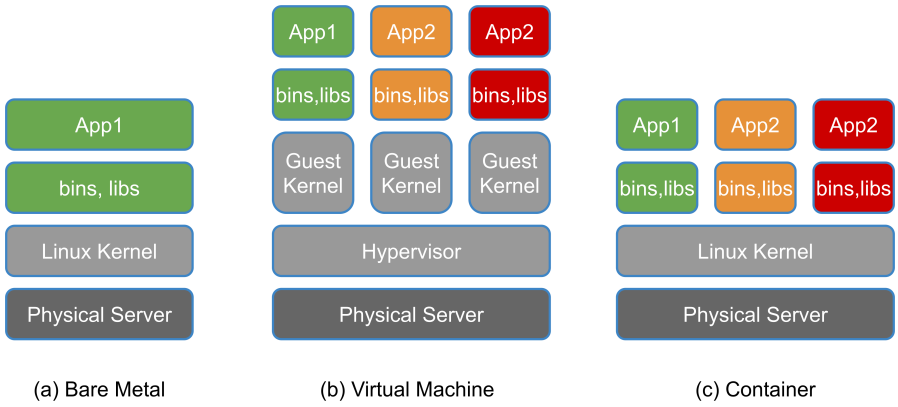


Figure 1.4: The difference in physical server usage between (a) Bare Metal servers, (b) Virtual Machine and (c) Container technology.
(a) Bare Metal servers is a word to describe conventional physical servers in contrast to Virtual Machines. On top of a Bare Metal server, an operating system and application programs are running. (b) Virtual Machine technology utilizes physical server hardware and a hypervisor. The hypervisor provides generic representations of server hardware, which are called virtual machines. A full operating system and applications are running on each of the virtual machines. (c) Container technology separates applications by containing them to their respective namespaces. Applications can not see each other's file systems, networks, users and process IDs unless they belong to the same namespace. Since container technology merely relies on Linux kernel's namespace function and optionally cgroup, a containerized process does not have any additional overhead compared with a process running on a conventional physical server and operating system. Container technology can be also utilized on top of virtual machines.

1.2.3 Container technology

More recently, Linux containers[26] have come to draw a significant amount of attention. Figure 1.4 (c) shows an example architecture of container technology. Linux containers are merely the processes with separate execution environments that are created using the Linux kernel's namespace feature. The namespace feature can isolate visibility of resources on a single Linux server.

Every process in a container is assigned to a certain namespace, and if two processes belong to different namespaces, they can not see each other's resources. Linux kernel implements filesystem namespace, PID namespace, network namespace,

user namespace, IPC namespace, and hostname namespace. For example, every filesystem namespace can have its own root filesystem, and every network namespace can have its own network devices and IP addresses. Therefore, it is possible to configure processes as if they were running in different Linux systems by assigning them to different namespaces, although they share kernel and hardware. While a VM needs to run a full OS on top of a hypervisor and hence imposes extra overhead, a process in Linux container is merely a process with a dedicated namespace and hence imposes no extra overhead.

The Linux container can run on any Linux systems including physical servers and VMs. Due to the widespread usage of Linux systems, the Linux container can run in most of the cloud infrastructures and on-premise data centers, which is beneficial for migrations.

Several management tools are available for Linux containers, including LXC[29], systemd-nspawn[12] and Docker[27]. These tools assign an appropriate namespace to a process upon the launch of itself and make it look like running in its own virtual Linux system. For example, container tools restore a file system from an archive file every time a container is launched. Container tools also set up separate network interfaces with separate IP addresses in the container's namespace.

The fact that each container has its own file system that is restored from a single archive file brings a significant benefit, i.e. a program binary and shared libraries are always exactly the same regardless of the base infrastructure. Therefore a process in a container is guaranteed to behave exactly the same manner, even if totally different data centers or cloud providers are used. This was not easy when there was no container technology. Because there are many flavors of Linux distributions, and even if the same distribution is used, there was always a chance that a slight difference in a program binary version or library versions could have broken the expected behavior.

In addition to that, containers can have own version of libraries in their respective filesystems, in other words, libraries in any container can be independently updated without influencing other containers. In conventional technologies, processes on a single server are dependent on common shared libraries, and hence updates of the library sometimes have caused unexpected side effects. Container tools alleviate these problems by packing necessary libraries into archives.

Thanks to these benefits container technologies are very attractive for web applica-

tions and their migrations. Considerable efforts in utilizing container technologies for web applications are ongoing. And to simplify the deployment of a complex web application that consists of interdependent container clusters, several container orchestrators(container management systems) have been in development.

1.2.4 Container Orchestrator

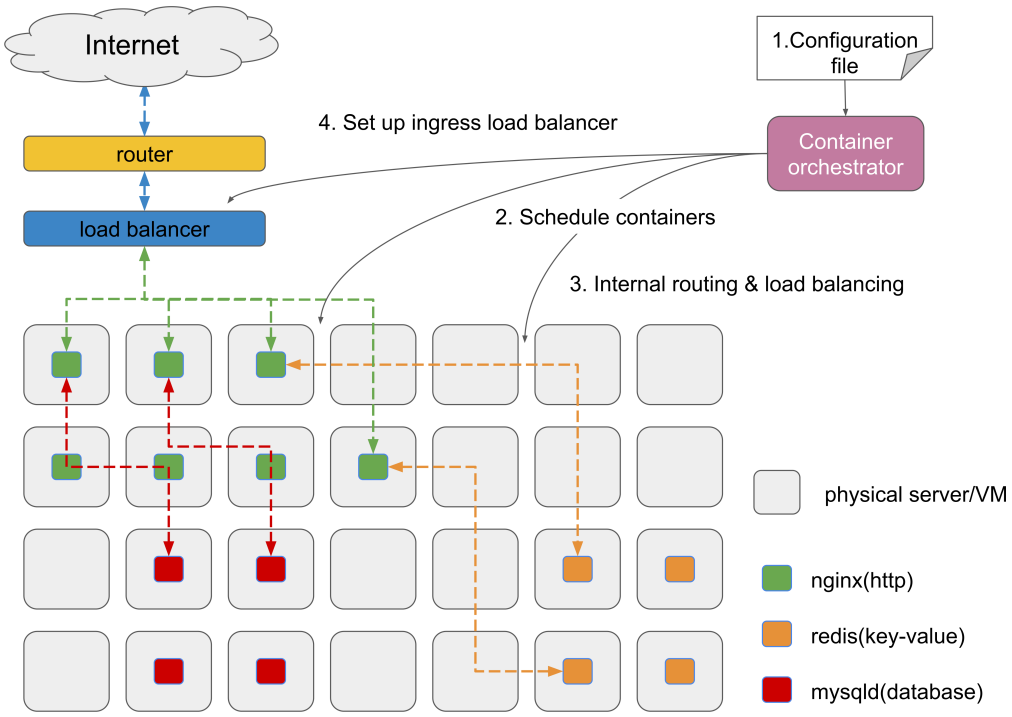


Figure 1.5: A web application cluster and container orchestrator.
A web application cluster consists of nginx(http), redis(key-value store) and mysqld(database) is depicted in this figure. Each of the component consists of a container cluster. Container orchestrator recieve configuration file(1), schdule containers(2), set up ingress routing using load balancer(3) and set up internal routing(4).

A container orchestrator (also called container cluster management system) is a tool to simplify the management of a cluster of containers that are launched on multiple servers. Figure 1.5 shows the important features for the container orchestrator for web applications; 1) **Configuration file:** Orchestrator should manage container

clusters based on a configuration file. The configuration file must be able to describe container cluster internals and relationships between interdependent container clusters. 2) **Scheduling:** Depending on the configuration file, the orchestrator must be able to pick servers and launch containers on them. Orchestrator must also maintain the state described in configuration files, for example, the orchestrator may need to maintain the number of running containers. 3) **Ingress routing:** The orchestrator must be able to set up routes for incoming traffic from the internet to multiple containers in a redundant and scalable manner. 4) **Internal routing:** If the web application consists of multiple interdependent container clusters, the orchestrator must be able to set up routes between them in a redundant and scalable manner.

In a configuration file a user can describe how a container cluster should be configured, and also can describe relationships between different clusters. As a result, a user can launch a web application that consists of interdependent container clusters just by supplying the configuration file to the orchestrator. For example, a web cluster in Figure 1.5 consists of three different functionalities, namely http server, key-value store, and database. Each of those consists of a container cluster. In the configuration file, the relationships between http server cluster, key-value store cluster, and database cluster are specified. The configuration file also contains how each cluster should be configured, including the number of the containers and resources assigned to them. Users only need to feed the configuration file to the orchestrator to launch the web application.

Thanks to these features, an orchestrator can be viewed as if it is an Operating System for a server farm in a data center, which not only schedules and launches containers on the server farm but also routes the traffic to the appropriate containers. By using orchestrators, a user can start a complex web application that consists of multiple interdependent container clusters, on multiple servers in a data center, as easily as starting a single process on a single computer. As a result, a user can also easily migrate their web applications at his or her convenience. And migrated web applications are guaranteed to behave exactly the same manner, not only because the same program binary and libraries are used in container, but also because container orchestrators hide difference among the base infrastructures.

Several container orchestrators are available, including Kubernetes, Docker swarm and Mesos/Marathon. Each of the container orchestrators varies in target applications,

and thus has the strength and weaknesses.

Kubernetes Kubernetes[5] is an open source container orchestrator, originally developed at Google based on their experience of production container orchestrator, Borg[39]. Since Google runs many of large scale web applications, Kubernetes are considered to be best suited to run web applications.

Docker swarm Docker Swarm is a container orchestrator built in Docker daemon itself. Users can execute regular Docker commands, which are then executed by a swarm manager. The swarm manager is responsible for controlling the deployment and the life cycle of containers.

Mesos/Marathon Mesos[19] is a common resource sharing layer for different type of applications like Hadoop, MPI jobs, and Spark in a Data Center. By using Mesos user does not neet to have dedicated pysical server cluster for each applications. Marathon is a framework which uses Mesos in order to orchestrate Docker containers. Because of the broader scope of applications, an out of box Mesos might not be particularly suited for web applications.

	Kuberenets	Docker Swarm	Mesos Marathon
Config file	Yaml	Yaml	Json
Scheduling	Yes	Yes	Yes
Ingress routing	Static Cloud load balancer**	Static	Static
Internal routing	iptables DNAT	ipvs	haproxy

Table 1.1: Container orchestrator comparison.
Important aspects of features as web application infrastructures are compared.
**Support for Cloud load balancer is only avalable in limited infrastructures including GCP, AWS, Azure and Openstack.

Table1.1 compares these orchestrators based on necessary features as an infrastructure for web applications. Although all of these orchestrators mostly satisfy the requirements, they rely on static routing for ingress traffic. Only Kubernetes has the functionality to manage cloud load balancer so that ingress traffic from the

Internet is routed to containers in a redundant and scalable manner, nevertheless, this functionality is applicable only for a few cloud environments.

To the best knowledge of the author, none of the existing container orchestrators has full support for the redundant and scalable ingress routing feature. The author believes this is an open and important topic for research and development, and therefore intends to pursue it.

1.3 Portable software load balancer

1.3.1 Load balancer for container clusters

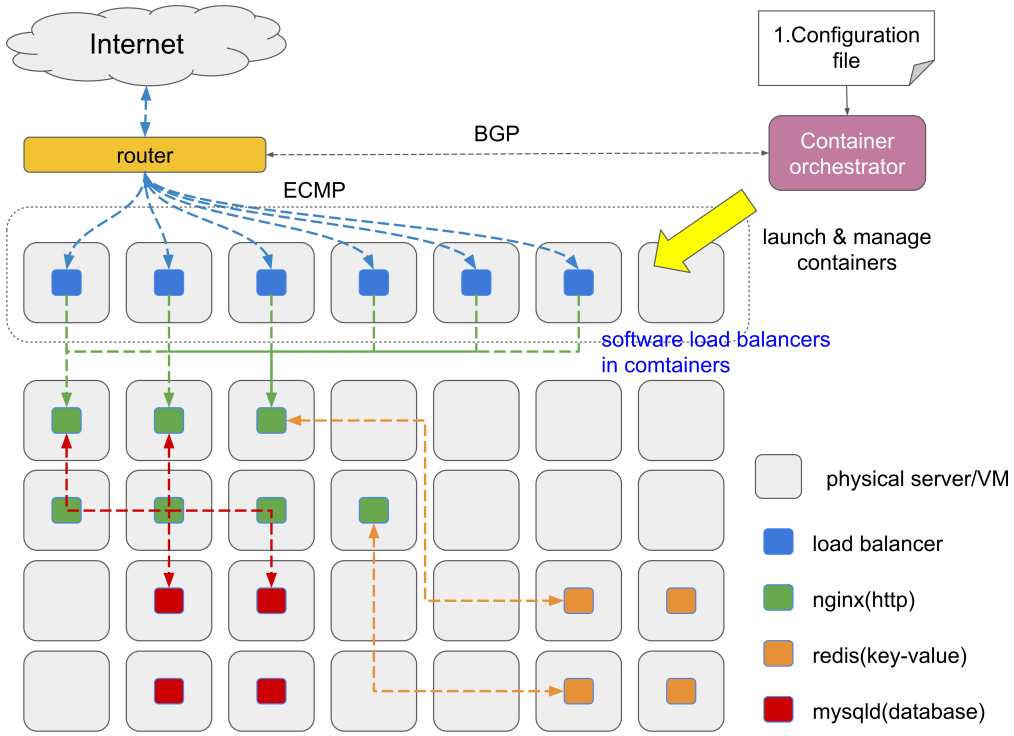


Figure 1.6: Load balancer for container clusters.
In order distribute the traffic, container orchestrator launches a cluster of software load balancer containers. The container orchestrator also communicates with the upstream router through BGP protocol and the router set up an ECMP routing rule in the routing table.

The purpose of this research is to investigate a generic way to route the traffic into container clusters in a redundant and scalable manner and thereby to facilitate web application migrations. In order to bring this into reality, the author proposes a cluster of containerized software load balancers.

Figure 1.6 shows schematic diagram of an example architecture for such load balancers. A web application that consists of nginx, redis, and mysqld, each being a cluster of containers, is running in the server farm. There is also a cluster of software load balancer containers, running in the same server farm. All of the containers are deployed and managed by the container orchestrator. The orchestrator also communicates with the upstream router using Border Gateway Protocol(BGP), so that the ingress traffic is forwarded according to Equal Cost Multi Path(ECMP) routing table to the load balancer containers in a redundant and scalable manner.

Container orchestrators are good at managing a cluster of containers, and they can scale containers, i.e. change the number of containers depending on the needs. Therefore it seems to be a good idea to deploy load balancers as a cluster of containers. In addition to that, by utilizing ECMP routing, redundancy and scalability are accomplished at the same time.

The author investigates a cluster of containerized software load balancers for Kubernetes as a test case since Kubernetes seems most appropriate for web application clusters at the moment. Nevertheless, the author expects general findings of this investigation can be easily applied to the other container orchestrators as well.

1.3.2 Problems of Kubernetes

As is mentioned in the previous section, none of the existing container orchestrators provide full support for automatic set up of ingress traffic routing. In the case of Kubernetes, the problem is its partial support for external load balancers. Here the author elaborates on the situation.

Load balancers are often used to distribute high volume traffic from the Internet to thousands of web servers. They are implemented as dedicated hardware or as software on commodity hardware. For on-premise data centers, there are a variety of proprietary hardware load balancers. Major cloud providers have developed software load balancers[13, 31] dedicated for their infrastructures. Although software load

balancers for cloud infrastructure have APIs, through which Kubernetes can control the behavior, most of the proprietary hardware load balancers do not have such APIs.

Thanks to the expressive syntax of the configuration file, Kubernetes allows users to easily launch complex web applications that consist of multiple interdependent container clusters as if they were launching a single application program. It also allows users to modify the state of their container clusters, just by modifying the configuration file. Kubernetes always try to make the status of containers to match its desired state, which is written in the configuration file. As a result, Kubernetes is currently considered to be most preferable orchestrators for web applications.

As for the routing of the ingress traffic, the problem becomes evident. In environments where there are supported load balancers, namely cloud environments including Google Cloud Platform (GCP), Amazon Web Applications (AWS), or Openstack, Kubernetes can request the infrastructures to automatically set up the load balancers upon the launch of a web application. The cloud load balancers will distribute ingress traffic to every node(physical servers or VMs) that might host containers. Once the traffic reaches the nodes, Kubernetes nicely route them to containers using iptables Destination Network Address Translation(DNAT) based internal load balancer.

However, this scheme does not work in many environments, since there are many load balancers that are not supported by Kubernetes. In such cases Kubernetes expects users to manually set up a route for the ingress traffic, which generally lacks redundancy and scalabilities. Kubernetes fails to provide uniformity that is essential for easy deployment and migration of complex web applications.

Other container orchestrators, e.g. Docker swarm or Mesos/Marathon, do not even have partial support for load balancers and expect users to manually set up the route for ingress traffic. Therefore this is a generic problem that current container cluster orchestrators possess.

1.3.3 Proposed solution

In order to alleviate this problem, the author proposes a portable and scalable software load balancer that can be used in any environment where there is no load balancer supported by container orchestrators, as is shown in the Figure 1.6. By using a proposed load balancer, users no longer need to manually adjust their services to the base

infrastructures.

As a proof of concept the author implements the proposed software load balancer that works well with Kubernetes using following technologies; 1) To make the load balancer runnable in any environment, Linux kernel's Internet Protocol Virtual Server (ipvs)[41] is containerized using Docker[27]. 2) To make the load balancer redundant and scalable, the author makes it capable of updating the routing table of upstream router with Equal Cost Multi-Path(ECMP) routes[16] using Border Gateway Protocol(BGP), which is a standard routing protocol. 3) The author also extends the research into implementing the novel load balancer using eXpress Data Plane(XDP) technology[4] to enhance the performance level to meet the need for 10Gbps network speed.

1.3.4 Contribution

Contributions of this paper can be summarized as follows: 1) The author addresses the problem of ingress traffic routing that is generic to container orchestrators and proposes software load balancer suitable for container environment. This is one of the most important problems for container orchestrators because without solving this problem migration of a web application will never be easy. 2) The author builds a proof of concept load balancers using OSS, which means that anyone can test drive the proposed load balancers and use them in production for free. 3) The author provides quantitative performance analysis to see the feasibility of proposed load balancer architecture in 1Gbps and 10Gbps network. 4) The author clarifies the remaining problems for future improvement and explores other technology to be used in faster networks.

The outcome of this study will benefit users who want to deploy their web applications on any cloud provider where no scalable load balancer is provided, to achieve high scalability. Moreover, the result of our study will potentially benefit users who want to use a group of different cloud providers and on-premise data centers across the globe seamlessly. In other words, users will become being able to deploy a complex web application on aggregated computing resources on the earth, as if they were starting a single process on a single computer.

1.4 Outline

The rest of the paper is organized as follows. Chapter 6 and Chapter 2 provides related works and the background information necessary to understand following chapters. Chapter 3 provides the problems of existing load balancers and proposes suitable architectures. Chapter **??** presents implementation of the proposed load balancer architecture in detail. Chapter **??** discusses portability and performance levels of the proposed load balancer in 1 Gbps network environment. Chapter **??** discusses the redundancy and scalability of the proposed load balancers. Chapter 5 present the performance levels of the proposed load balancer in 10 Gbps network environment and discuss the method to improve the performance of a software load balancer. Chapter **??** discusses the limitation and the future work of this study, which is followed by a conclusion of this work in Chapter 7.1.

2

Background

This chaper provides background information that are important in this research. First two of the most popular overlay networks used in Kubernetes are explaind in detail. Then the author explain how to utilize multicore CPUs for packet proccessing in Linux.

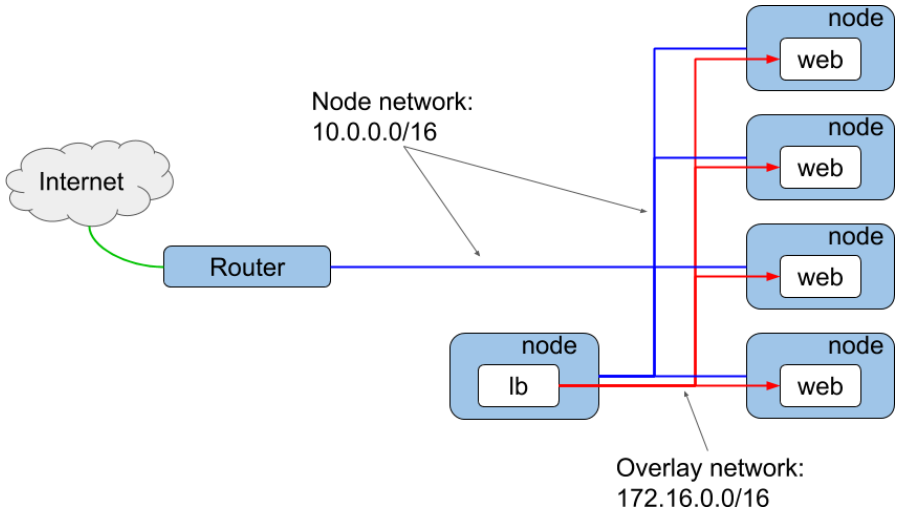


Figure 2.1: The network architecture of an exemplified container cluster system. A load balancer(lb) pod(the white box with "lb") and web pods are running on nodes(the blue boxes). The traffic from the internet are forwarded to the lb pod by the upstream router using the node network, and the distributed to web pods using the overlay network.

2.1 Container Network

2.1.1 bridge+veth

2.1.2 macvlan, ipvlan

2.1.3 hostnetwork

2.2 Overlay Network

2.2.1 General concept

In order to discuss load balancer for container cluster, the knowledge of the overlay network is essential. We briefly explain an abstract concept of overlay network in this subsection.

Fig. 2.1 shows schematic diagram of network architecture of a container cluster system. Suppose we have a physical network(node network) with IP address range of 10.0.0.0/16 and an overlay network with IP address range of 172.16.0.0/16. The node

network is the network for nodes to communicate with each other. The overlay network is the network setups for containers to communicate with each other. An overlay network typically consists of appropriate routing tables on nodes, and optionally of tunneling setup using ipip or vxlan. The upstream router usually belongs to the node network. When a container in the Fig. 2.1 communicates with any of the nodes, it can use its IP address in 172.16.0.0/16 IP range as a source IP, since every node has proper routing table for the overlay network. When a container communicates with the upstream router that does not have routing information regarding the overlay network, the source IP address must be translated by Source Network Address Translation(SNAT) rules on the node the container resides.

The SNAT caused a problem when we tried to co-host multiple load balancer containers for different services on a single node, and let them connect the upstream router directly. This was due to the fact that the BGP agent used in our experiment only used the source IP address of the connection to distinguish the BGP peer. The agent behaved as though different BGP connections from different containers belonged to a single BGP session because the source IP addresses were identical due to the SNAT.

There many overlay network implementations. The author investigated two of the popular ones to see how it works.

2.2.2 Flannel

We used flannel to build the Kubernetes cluster used in our experiment. Flannel has three types of backend, *i.e.*, operating modes, named host-gw, vxlan, and udp[9].

In the host-gw mode, the flanneld installed on a node simply configures the routing table based on the IP address assignment information of the overlay network, which is stored in the etcd. When a *pod* on a node sends out an IP packet to *pods* on the different node, the former node consults the routing table and learn that the IP packet should be sent out to the latter. Then, the former node forms Ethernet frames containing the destination MAC address of the latter node without changing the IP header, and send them out.

In the case of the vxlan mode, flanneld creates the Linux kernel's vxlan device, flannel.1. Flanneld will also configures the routing table appropriately based on the information stored in the etcd. When *pods* on different nodes need to communicate,

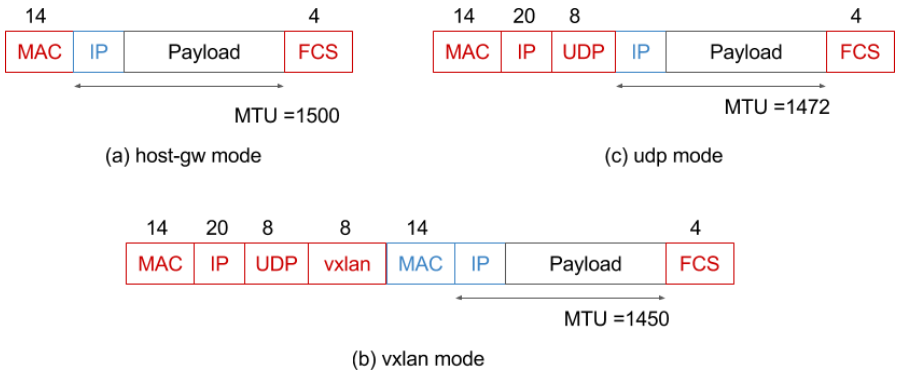


Figure 2.2: Frame diagram

the packet is routed to flannel.1. The vxlan functionality of the Linux kernel identify the MAC address of flannel.1 device on the destination node, then form an Ethernet frame toward the MAC address. The vxlan then encapsulates the Ethernet frame in a UDP/IP packet with a vxlan header, after which the IP packet is eventually sent out.

In the case of udp mode, flanneld creates the tun device, flannel0, and configures the routing table. The flannel0 device is connected to the flanneld daemon itself. An IP packet routed to flannel0 is encapsulated by flanneld, and eventually sent out to the appropriate node. The encapsulation is done for IP packets.

Figure 2.2 shows the schematic diagrams of frame formats for three backends modes of the flannel overlay network. The MTU sizes in the backends, assuming the MTU size without encapsulation is 1500 bytes, are also presented. Since packets are not encapsulated in the host-gw mode, the MTU size remains 1500 bytes. An additional 50 bytes of header is used in the vxlan mode, thereby resulting in an MTU size of 1450 bytes. In the case of the udp mode, only 28 bytes of header are used for encapsulation, which results in an MTU size of 1472 bytes.

Performance of the load balancers can be influenced by the overhead of encapsulation. Thus, the host-gw mode, where there is no overhead due to encapsulation, results in the best performance levels as is shown in Chapter **??**. However, the host-gw mode has a significant drawback that prohibit it to work correctly in cloud platforms. Since the host-gw mode simply sends out a packet without encapsulation, if there is a cloud gateway between nodes, the gateway cannot identify the proper destination, thus drop the packet.

We conducted an investigation to determine which of the flannel backend mode

mode	On-premise	GCP	AWS
host-gw	OK	NG	NG
vxlan	OK	OK	OK
udp	OK	OK	OK

Table 2.1: Viable flannel backend modes. In cloud environment tunnelig using vxlan or udp is needed.

would be usable on AWS, GCP, and on-premise data centers. The results are summarized in Table 2.1. In the case of GCP, an IP address of /32 is assigned to every VM host and every communication between VMs goes through GCP's gateway. As for AWS, the VMs within the same subnet communicate directly, while the VMs in different subnets communicate via the AWS's gateway. Since the gateways do not have knowledge of the flannel overlay network, they drop the packets; thereby, they prohibit the use of the flannel host-gw mode in those cloud providers.

In our experiment, we compared the performance of load balancers when different flannel backend modes were used.

2.2.3 Calico

[Filled in later]

2.3 Multicore Packet Proccessing

Recently, the performance of CPUs are improved significantly due to the development of multi-core CPUs. One of the top of the line server processors from Intel now includes up to 28 cores in a single CPU. In order to enjoy the benefits of multi-core CPUs in communication performance, it is necessary to distribute the handling of interrupts from the NIC and the IP protocol processing to the available physical cores.

2.3.1 rss

Receive Side Scaling (RSS)[38] is a technology to distribute handling of the interrupt from NIC queues to multiple CPU cores. Subsequently, Receive Packet Steering

```
81: eth0-tx-0
82: eth0-rx-1
83: eth0-rx-2
84: eth0-rx-3
85: eth0-rx-4
# obtained from /proc/interrupts
```

Figure 2.3: RX/TX queues of the hardware

(RPS)[38] distributes the IP protocol processing to multiple CPU cores by issuing inter core software interrupts.

Since load balancer performance levels could be affected by these technologies, we conducted an experiment to determine how load balancer performance level change depending on the RSS and RPS settings. The following shows how RSS and RPS are enabled and disabled in our experiment. The NIC used in our experiment is Broadcom BCM5720, which has four rx-queues and one tx-queue. Figure 2.3 shows the interrupt request (IRQ) number assignments to those NIC queues.

When packets arrive, they are distributed to these rx-queues depending on the flow each packet belongs to. Each receive queue has a separate IRQ associated with it. The NIC triggers this to notify a CPU when new packets arrive on the given queue. Then, the notified CPU handles the interrupt, and performs the protocol processing. According to the [38], the CPU cores allowed to be notified is controlled by setting a hexadecimal value corresponding to the bit maps indicating the allowed CPU cores in "/proc/irq/$irq_number /smp_affinity". For example, in order to route the interrupt for eth0-rx-1 to CPU0, we should set "/proc/irq/82/smp_affinity" to binary number 0001, which is 1 in hexadecimal value. Further, in order to route the interrupt for eth0-rx-2 to CPU1, we should set "/proc/irq/83/smp_affinity" to binary number 0010, which is 2 in hexadecimal value.

We refer the setting to distribute interrupts from four rx-queues to CPU0, CPU1, CPU2 and CPU3 as RSS = on. It is configured as the following setting:

RSS=on

```
echo 1 > /proc/irq/82/smp_affinity
echo 2 > /proc/irq/83/smp_affinity
echo 4 > /proc/irq/84/smp_affinity
echo 8 > /proc/irq/85/smp_affinity
```

On the other hand, RSS = off means that an interrupt from any rx-queue is routed to CPU0. It is configured as the following setting:

RSS=off

```
echo 1 > /proc/irq/82/smp_affinity
echo 1 > /proc/irq/83/smp_affinity
echo 1 > /proc/irq/84/smp_affinity
echo 1 > /proc/irq/85/smp_affinity
```

rps

The RPS distributes IP protocol processing by placing the packet on the desired CPU's backlog queue and wakes up the CPU using inter-processor interrupts. We have used the following settings to enable the RPS:

RPS=on

```
echo fefe > /sys/class/net/eth0/queues/rx-0/RPS_cpus
echo fefe > /sys/class/net/eth0/queues/rx-1/RPS_cpus
echo fefe > /sys/class/net/eth0/queues/rx-2/RPS_cpus
echo fefe > /sys/class/net/eth0/queues/rx-3/RPS_cpus
```

Since the hexadecimal value "fefe" represented as "1111 1110 1111 1110" in binary, this setting will allow distributing protocol processing to all of the CPUs, except for CPU0 and CPU8. In this paper, we will refer this setting as RPS = on. On the other hand, RPS = off means that no CPU is allowed for RPS. Here, the IP protocol processing is performed on the CPUs the initial hardware interrupt is received. It is configured as the following settings:

```
RPS=off
echo 0 > /sys/class/net/eth0/queues/rx-0/RPS_cpus
echo 0 > /sys/class/net/eth0/queues/rx-1/RPS_cpus
echo 0 > /sys/class/net/eth0/queues/rx-2/RPS_cpus
echo 0 > /sys/class/net/eth0/queues/rx-3/RPS_cpus
```

The RPS is especially effective when the NIC does not have multiple receive queues or when the number of queues is much smaller than the number of CPU cores. That was the case of our experiment, where we had a NIC with only four rx-queues, while there was a CPU with eight physical cores.

2.4 Linux vritual server

In order to demonstrate a software load balancer that is runnable in any environment, the ipvs is containerized. In addition to that the proposed load balancer uses two other components, keepalived, and a controller. These components are placed in a single Docker container image. The ipvs is a Layer-4 load balancer capability, which is included in the Linux kernel 2.6.0 released in 2003 or later, to distribute incoming Transmission Control Protocol(TCP) traffic to *real servers*[1] [41]. For example, ipvs distributes incoming Hypertext Transfer Protocol(HTTP) traffic destined for a single destination IP address, to multiple HTTP servers(e.g. Apache HTTP or nginx) running on multiple nodes in order to improve the performance of web services. Keepalived is a management program that performs health checking for *real servers* and manages ipvs balancing rules in the kernel accordingly. It is often used together with ipvs to facilitate ease of use. The controller is a daemon that periodically monitors the *pod* information on the master, and it performs various actions when such information changes. Kubernetes provides ingress controller framework as the Go Language(Golang) package to implement the controllers. We implement a controller program that feeds *pod* state changes to keepalived using this framework.

[1]The term, *real servers* refers to worker servers that will respond to incoming traffic, in the original literature[41]. We will also use this term in the similar way.

2.5 Summary

3 Architecture and Implementation

This chapter provides discussion of load balancer suitable for container clusters. First the author discusses problems of conventions architecture in Section 3.1.1. Then the author proposes the best one in Section **??**. After that the author discusses inplementation of a portable and redundant load balancer in Section **??**.

3.1 Architecture

3.1.1 Problem of Conventional Architecture

The problem of Kubernetes is its partial support for the ingress traffic routing. Figure 3.1a shows an exemplified Kubernetes cluster. A Kubernetes cluster typically consists of a master and nodes. They can be physical servers or VMs. On the master, daemons that control the Kubernetes cluster are typically deployed. These daemons include, apiserver, scheduler, controller-manager and etcd. On the nodes, the kubelet daemon will run *pods*, depending the PodSpec information obtained from the apiserver

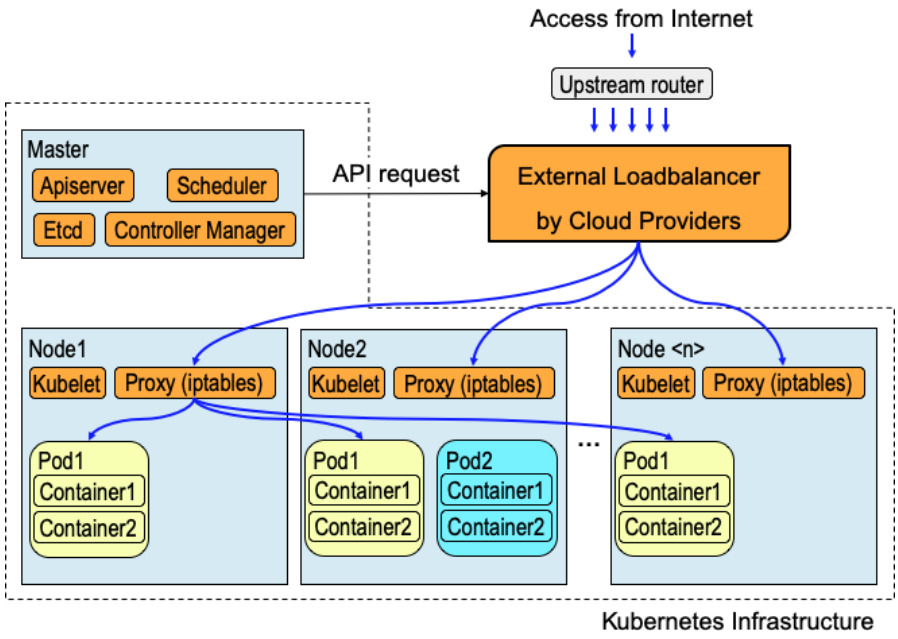
on the master. A *pod* is a group of containers that share the same network namespace and cgroup, and is the basic execution unit in a Kubernetes cluster.

When a service is created, the master schedules where to run *pods* and kubelets on the nodes launch them accordingly. At the same time, the master sends out requests to cloud provider's API endpoints, asking them to set up external cloud load balancers that distribute ingress traffic to every node in the Kubernetes cluster. The proxy daemon on the nodes also setup iptables DNAT[24] rules. The Ingress traffic will then be evenly distributed by the cloud load balancer to nodes, after which it will be distributed again by the DNAT rules on the nodes to the designated *pods*. The returning packets follows the exact same route as the incoming ones.

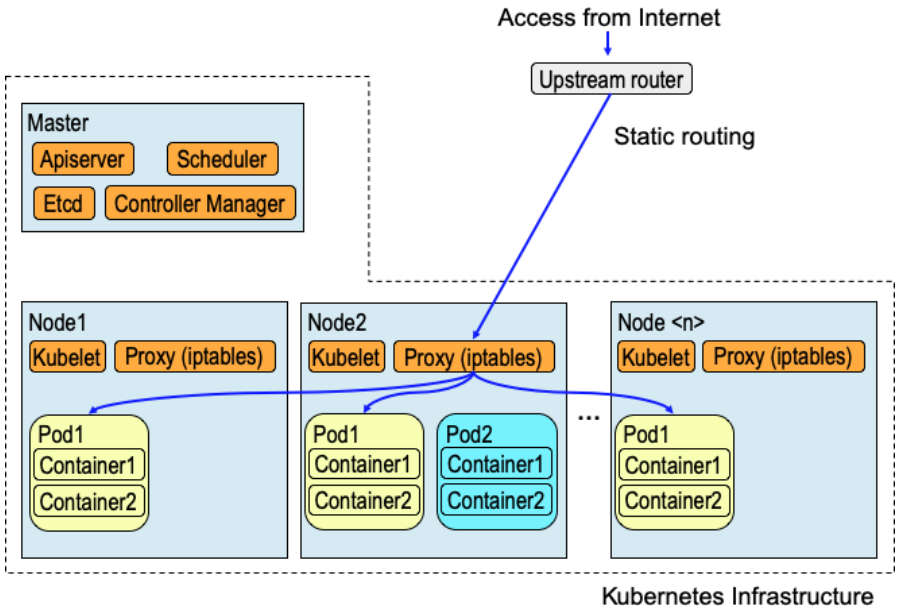
This architecture has the followings problems: 1) There must exist cloud load balancers whose APIs are supported by the Kubernetes daemons. There are numerous load balancers which is not supported by the Kubernetes. These include the bare metal load balancers for on-premise data centers. 2) Distributing the traffic twice, first on the external load balancers and second on each node, complicates the administration of packet routing. Imagine a situation in which the DNAT table on one of the nodes malfunctions. In such a case, only occasional timeouts would be observed, and hence it would be very difficult to find out which node is malfunctioning.

Regarding the first problem, if there is no load balancer that is not supoorted by Kubernetes, users must set up the routing manually depending on the infrastructure. The traffic would be routed to a node then distributed by the DNAT rules on the node to the designated *pods*. However, this approach significantly degrades the portability of container clusters.

In short, 1) Kubernetes can be used only in limited environments where the external load balancers are supported, and 2) the routes incoming traffic follow are very complex. In order to address these problems, the author proposes a containerized software load balancer that is deployable in any environment even if there are no external load balancers.



(a) Kubernetes in cloud infrastructures



(b) Kubernetes in on-premise data centers

Figure 3.1: Conventional architecture of a Kubernetes cluster
In supported infrastructures, e.g., major cloud providers, Kubernetes automatically set up route to the service IPs, when web service providers launch container clusters. In unsupported infrastructures, e.g., on-premise data centers, they have to manually set up the route.

3.1.2 Load balancer in container

The author proposes a load balancer architecture, where a cluster of load balancers are deployed as containers. Figure 3.2 shows the proposed load balancer architecture for Kubernetes, which has the following characteristics; 1) Each load balancer itself is run as a *pod* by Kubernetes. 2) Load balancing tables are dynamically updated based on information about running *pods*. 3) There exist multiple load balancers for redundancy and scalability. 4) The routing table in the upstream router are updated dynamically using standard network protocol.

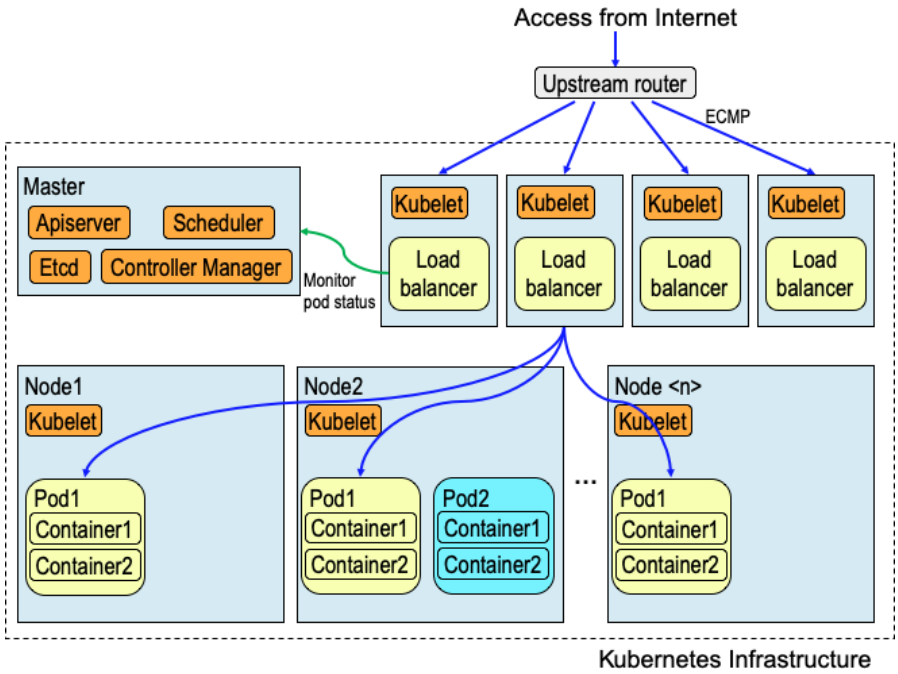


Figure 3.2: Kubernetes cluster with proposed load balancer.

Containerized load balancer can run in any environment including on-premise data centers if only they use Linux systems. Furthermore, load balancers can share the server pool with web containers. Users can utilize existing servers rather than buying dedicated hardware.

In the case of conventional architecture, cloud load balancer distributed ingress traffic to every node and then internal load balancers based on iptables DNAT forwarded the packet again to running pods. However, in the case of the proposed load balancers, the ingress traffic is directly routed to the running *pods*. As a result, the route

becomes simpler and hence finding malfunctions becomes easier than the conventional architecture.

Because a cluster of load balancer containers is controlled by a container orchestrator, Kubernetes, the load balancer becomes redundant and scalable. Kubernetes always tries to maintain the number of load balancer containers at the number specified by the administrator. If a single container fails, Kubernetes schedule and launch another one on a different node, which provides the resilience to failures. Furthermore, when there is an increase in the traffic, it can also scale the size of the cluster depending on the demand.

The routes to the load balancers are automatically updated through the standard protocol, BGP. Therefore users do not need to manually add the route every time new load balancer container is launched, as is the case in the conventional architecture.

3.1.3 Redundancy with ECMP

Fig. 3.3 shows a schematic diagram to explain redundancy architecture with ECMP for the proposed load balancer. The ECMP is a functionality a router often supports, where the router has multiple next hops with equal cost(priority) to a destination, and generally distribute the traffic depending on the hash of the flow five tuples(source IP, destination IP, source port, destination port, protocol). The multiple next hops and their cost are often populated using the BGP protocol. The notable benefit of the ECMP setup is the fact that it is scalable. All the load balancers that claims as the next hop is active, i.e., all of them are utilized to increase the performance level. Since the traffic from the internet is distributed by the upstream router, the overall throughput is determined by the router after all. However, in practice, there are a lot of cases where this architecture is beneficial. For example, if a software load balancer is capable of handling 1 Gbps equivalent of traffic and the upstream router is capable of handling 10 Gbps, it still is worthwhile launching 10 of the software load balancer containers to fill up maximum throughput of the upstream router.

We place a node with the knowledge of the overlay network as a route reflector, to deal with the complexity due to the overlay network. Since the upstream router normally has no knowledge of overlay network used in a Kubernet clusters, and hence IP addresses that containers use, a container must use SNAT to communicate with the

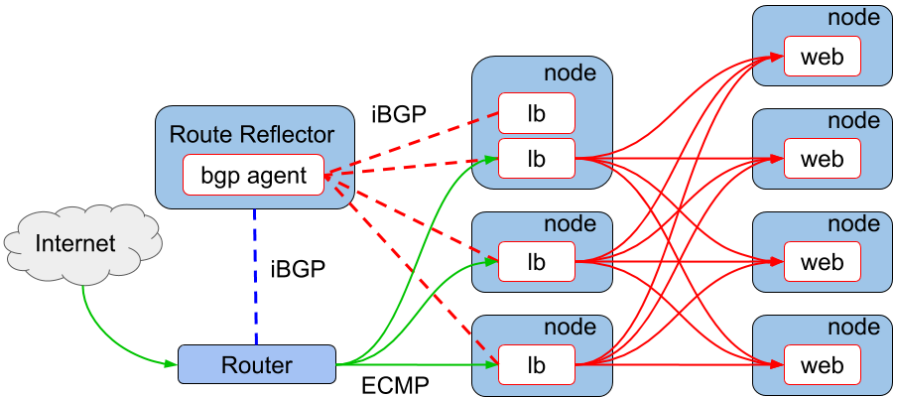


Figure 3.3: The proposed architecture of load balancer redundancy with ECMP. The traffic from the internet is distributed by the upstream router to multiple of lb pods using hash-based ECMP and then distributed by the lb pods to web pods using Linux kernel's ipvs. The ECMP routing table on the upstream router is populated using iBGP.

router. The SNAT caused a problem when we tried to co-host multiple load balancer containers for different services on a single node, and let them connect the upstream router directly. This was due to the fact that the BGP agent used in our experiment only used the source IP address of the connection to distinguish the BGP peer. The agent behaved as though different BGP connections from different containers belonged to a single BGP session because the source IP addresses were identical due to the SNAT.

A route reflector is a network component for BGP to reduce the number of peerings by aggregating the routing information[37]. In our proposed architecture we use it as a delegater for load balancer containers towards the upstream router.

By using the route reflector, we can have the following benefits. 1) Each node can accommodate multiple load balancer containers. This was not possible when we tried to directly connect load balancers and the router through SNAT. 2) The router does not need to allow peering connections from random IP addresses that may be used by load balancer containers. Now, the router only need to have the reflector information as the BGP peer definition.

Since we use standard Linux boxes for route reflectors, we can configure them as we like; a) We can make them belong to overlay network so that multiple BGP sessions from a single node can be established. b) We can use a BGP agent that supports dynamic neighbor (or dynamic peer), where one only needs to define the IP range as a

peer group and does away with specifying every possible IP that load balancers may use.

The upstream router does not need to accept BGP sessions from containers with random IP addresses, but only from the route reflector with well known fixed IP address. This may be preferable in terms of security especially when a different organization administers the upstream router. Although not shown in the Fig. 3.3, we could also place another route reflector for redundancy purpose.

This chapter presents implementation of the proof of the concept system for the proposed load balancer architecture in detail. First overall architecure is explained in Section 3.2.1. Then ipvs containerization is explained in detail in Section 3.2.2. Finally implementation of BGP software container is explained in Section 3.2.3.

3.2 Implementation

3.2.1 Experimental system architecture

Fig. 3.4 shows the schematic diagram of proof of concept load balancer cluster system. Each load balancer pod consists of an exabgp container and an ipvs container. The ipvs container is responsible for distributing the traffic toward the IP address that a service uses, to web server(nginx) pods. The ipvs container monitors the availability of web server pods by consulting apiserver on the master node and manages the load balancing rule appropriately. The exabgp container is responsible for advertising the route toward the IP address that a service uses, to the route reflector. The route reflector aggregates the routing information advertised by load balancer pods and advertise them to the upstream router. The upstream router updates its routing table according to the advertisement.

All the nodes and route reflector are configured using Debian 9.5 with self compiled linux-4.16.12 kernel. The author also used conventional Linux box as the upstream router for testing purpose, using the same OS as the nodes and route reflector. The version of Linux kernel needed to be 4.12 or later to support hash based ECMP routing table. The author also needed to enable kernel config option CONFIG_IP_ROUTE_MULTIPATH[21] when compiling, and set the kernel parameter fib_multipath_hash_policy=1 at run time. Although in the actual production environ-

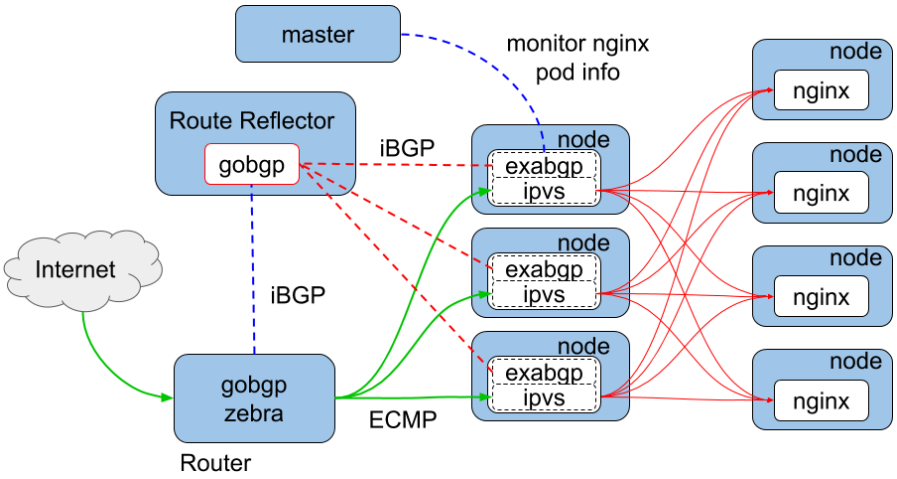


Figure 3.4: Proof of concept system
An experimental container cluster with proposed redundant software balancers. The master and nodes are configured as Kubernetes's master and nodes on top of conventional Linux boxes, respectively. The route reflector and the upstream router are also conventional Linux boxes.

ment, proprietary hardware router with the highest throughput is usually deployed, we could still test some of the advanced functions by using a Linux box as the router.

Exabgp is used in the load balancer pods and gobgp is used in the route reflector and the upstream router. Exabgp can be configured as a static route advertiser in a much simpler way than gobgp or other software, which is preferable for load balancers. Gobgp supports add-path[40] feature needed for multi-path advertisement, which is required for the route reflector. Gobgp also supports Forwarding Information Base(FIB) manipulation[15]feature through zebra[30] for updating routing table in the upstream router. The add-path and FIB manipulation are not supported in exbgp. The configurations for the router is summarised in B.3.

The route reflector also uses a Linux box with gobgp and overlay network setup. The requirements for the BGP agent on the route reflector are dynamic-neighbours and add-paths features. The configurations for the route reflector is summarised in B.2.

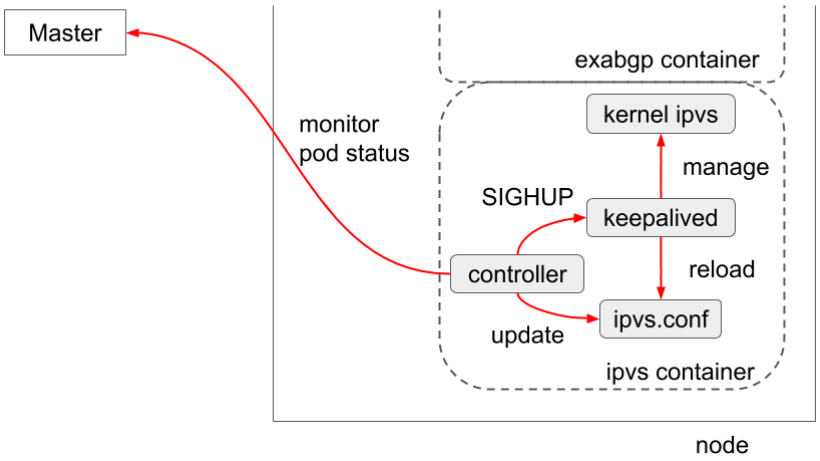


Figure 3.5: Implementation

3.2.2 Ipvs container

The proposed load balancer needs to dynamically reconfigure the IPVS balancing rules whenever *pods* are created/deleted. Figure 3.5 is a schematic diagram to show the dynamic reconfiguration of the IPVS rules. Two daemon programs, controller and keepalived, run in the container inside the LB2 pod are illustrated. The keepalived manages Linux kernel's IPVS rules depending on the ipvs.conf configuration file. It is also capable of health-checking the life of *real server*, which is represented as a combination of the IP addresses and port numbers of the target *pods*. If the health check to a *real server* fails, keepalived will remove that *real server* from the IPVS rules.

The controller monitors information concerning the running *pods* of a service in the Kubernetes cluster by consulting the apiserver running on the master. Whenever *pods* are created or deleted, the controller will automatically regenerate an appropriate ipvs.conf and issue SIGHUP to keepalived. Then, keepalived will reload the ipvs.conf and modify the kernel's IPVS rules accordingly. The actual controller[23] is implemented using the Kubernetes ingress controller[2] framework. By importing existing Golang package, "k8s.io/ingress /core/pkg/ingress", we could simplify the implementation, e.g. 120 lines of code.

Configurations for capabilities were needed in the implementation: adding the CAP_SYS_MODULE capability to the container to allow the kernel to load required

```
virtual_server fwmark 1 {
  delay_loop 5
  lb_algo lc
  lb_kind NAT
  protocol TCP
  real_server 172.16.21.2 80 {
    uthreshold 20000
    TCP_CHECK {
      connect_timeout 5
      connect_port 80
    }
  }
  real_server 172.16.80.2 80 {
    uthreshold 20000
    TCP_CHECK {
      connect_timeout 5
      connect_port 80
    }
  }
}
```

Figure 3.6: An example of ipvs.conf

kernel modules inside a container, and adding CAP_NET_ADMIN capability to the container to allow keepalived to manipulate the kernel's IPVS rules. For the former case, we also needed to mount the "/lib/module" of the node's file system on the container's file system.

Figure 3.6 and Figure 3.7 show an example of an ipvs.conf file generated by the controller and the corresponding IPVS load balancing rules, respectively. Here, we can see that the packet with `fwmark=1`[3] is distributed to `172.16.21.2:80` and `172.16.80.2:80` using the masquerade mode(Masq) and the least connection(lc)[41] balancing algorithm.

3.2.3 BGP software container

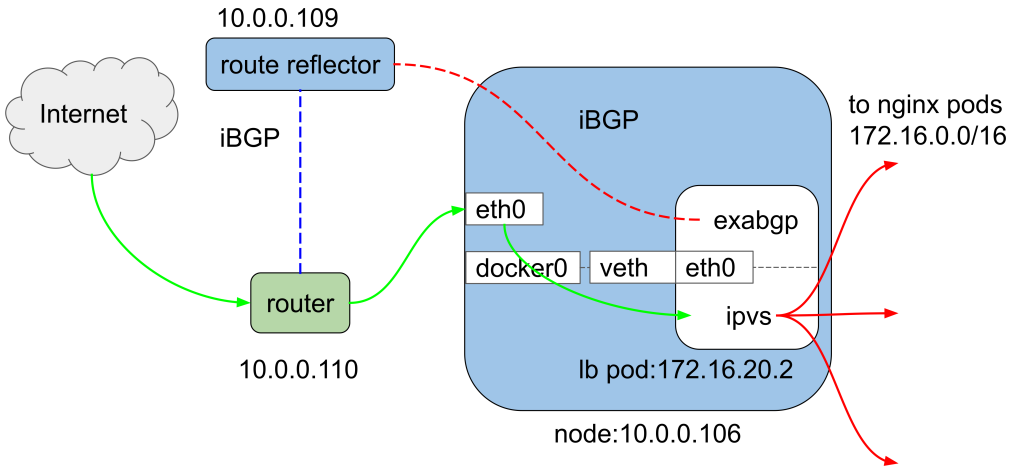
In order to implement the ECMP redundancy, we also containerized exabgp using Docker. Fig.3.8 (a) shows a schematic diagram of the network path realized by the

```
# kubectl exec -it IPVS-controller-4117154712-kv633 --
IPVSadm -L
IP Virtual Server version 1.2.1 (size=4096)
Prot LocalAddress:Port Scheduler Flags
  -> RemoteAddress:Port Forward Weight ActiveConn InActConn
FWM  1 lc
  -> 172.16.21.2:80      Masq    1      0          0
  -> 172.16.80.2:80      Masq    1      0          0
```

Figure 3.7: Example of IPVS balancing rules

exabgp container. We used exabgp as the BGP advertiser as mentioned earlier. The traffic from the Internet is forwarded by ECMP routing table on the router to the node, then routed to ipvs container.

Fig.3.8 (b) summarises some key settings required for the exabgp container. In BGP announcements the node IP address, 10.0.0.106 is used as the next-hop for the IP range 10.1.1.0/24. Then on the node, in order to route the packets toward 10.1.1.0/24 to the ipvs container, a routing rule to the dev docker0 is created in the node net namespace. A routing rule to accept the packets toward those IPs as local is also required in the container net namespace. A configuration of exabgp is shown in B.1.



(a)

[BGP announcement]
route 10.1.1.0/24 next-hop 10.0.0.106
[Routing in node net namespace]
ip netns exec node ip route replace 10.1.1.0/24 dev docker0
[Accept as local]
ip route add local 10.1.1.0/24 dev eth0

(b)

Figure 3.8: (a) Network path by the exabgp container. (b) Required settings in the exabgp container.

3.3 Summary

4 Performance Evaluation

This chapter discusses portability and performance level of a single ipvs load balancer in 1 Gbps environments. First the author investigated general characteristics of a single load balancer using physical servers in on-premise data center and compared performance level with existing iptables DNAT and nginx as a load balancer. Then the author also carried out the performance measurement in GCP and AWS to show that the containerized ipvs load balancer is runnable and has the same characteristics in the cloud environment. The following sections explain these in further detail.

4.1 Throughput measurement for ipvs-nat Load balancer

4.1.1 Benchmark method

A set of throughput measurement was carried out using an HTTP benchmark program, wrk[17]. Figure 4.1(a) illustrates a schematic diagram of the experimental setup.

Multiple *pods* are deployed on multiple nodes in the Kubernetes cluster. In each *pod*, an nginx web server pod that returns the IP address of the *pod* are running. The author set up the ipvs, iptables DNAT, and nginx load balancers on one of the nodes. All the nodes and the benchmark client are connected to a 1Gbps network switch as in Figure 4.1(b).

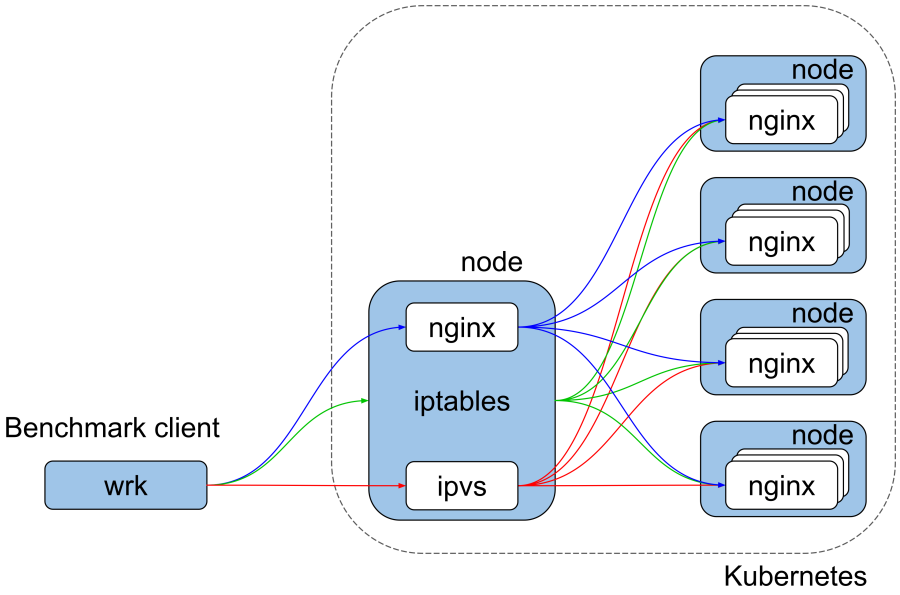
The throughput, Request/sec, is measured cluster as follows: The HTTP GET requests are sent out by the wrk on the client machine toward the nodes, using destination IP addresses and port numbers that are chosen based on the type of the load balancer on which the measurement is performed. The load balancer on the node then distributes the requests to the *pods*. Each *pod* returns HTTP responses to the load balancer, after which the load balancer returns them to the client. Based on the number of responses received by wrk on the client, load balancer performance, in terms of Request/sec can be obtained.

Table 4.1 shows an example of the command-line for wrk and the corresponding output. The command-line in Table 4.1 will generate 40 wrk program threads and allow those threads to send out a total of 800 concurrent HTTP requests over the period of 30 seconds. The output example shows the information including per thread statistics, error counts, Request/sec and Transfer/sec.

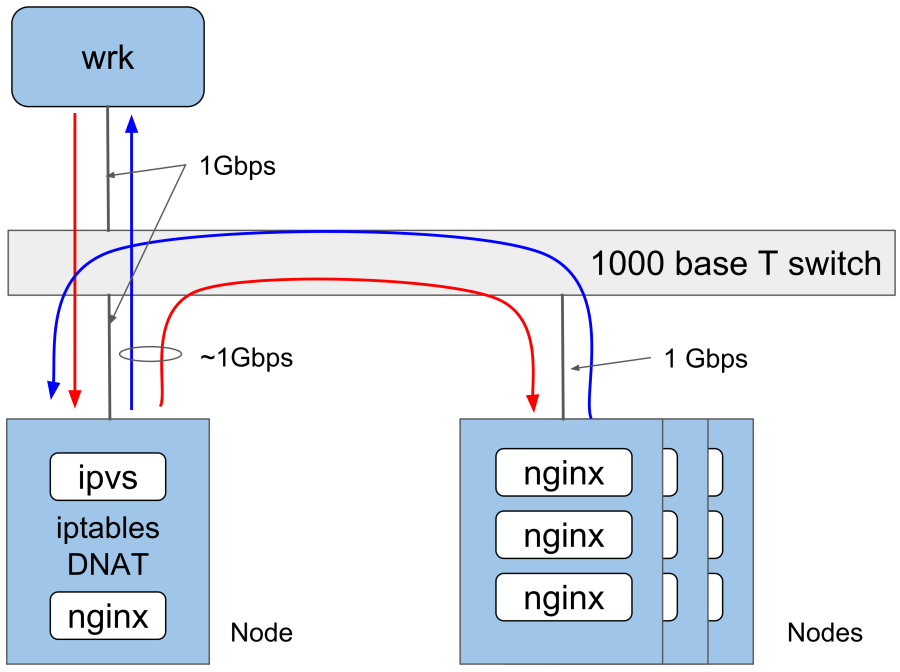
```
[Command line]
 wrk -c800 -t40 -d30s http://172.16.72.2:8888/
-c: concurrency, -t: # of thread, -d: duration

[Output example]
 Running 30s test @ http://10.254.0.10:81/
  40 threads and 800 connections
  Thread Stats   Avg      Stdev     Max   +/- Stdev
    Latency    15.82ms   41.45ms   1.90s    91.90\%
    Req/Sec     4.14k   342.26     6.45k    69.24\%
   4958000 requests in 30.10s, 1.14GB read
  Socket errors: connect 0, read 0, write 0, timeout 1
Requests/sec: 164717.63
Transfer/sec:     38.86MB
```

Table 4.1

Table 4.2 shows hardware and software configuration used in the experiments. All



(a) Logical configuration.



(b) Physical configuration.

Figure 4.1: Benchmark setup.

[Hardware Specification] CPU: Xeon E5-2450 2.10GHz (with 8 core, Hyper Threading) Memory: 32GB NIC: Broadcom BCM5720 Giga bit (Node x 6, LB x 1, Client x 1)
[Node Software] OS: Debian 8.7, linux-3.16.0-4-amd64 Kubernetes v1.10.6 flannel v0.7.0 etcd version: 3.0.15
[Container Software] Keepalived: v1.3.2 (12/03,2016) nginx : 1.11.1(load balancer), 1.13.0(web server)

Table 4.2

of the nginx web server pods are configured to return the IP address of the *pod*, in order to make them return a small HTTP content. This makes a relatively severe condition for load balancers. The size of the character string making up an IP address is limited to 15 bytes. If the author had chosen the HTTP response size so that most of the IP packet resulted in maximum transmission unit(MTU), the performance would have been dominantly limited by the Ethernet bandwidth.

For this experiment a total of eight servers are used; six servers for nodes, one for the load balancer and one for the benchmark client, with all having the same hardware specifications. The software versions used for Kubernetes, web server and load balancer *pods* are also summarized in the Table 4.2. The hardware we used had eight physical CPU cores and a 1Gbps NIC with 4 rx-queues.

4.1.2 Results

Effect of multicore proccessing

Figure 4.2 shows a result of throughput experiment with different multicore proccessing settings. The following three RSS and RPS settings were compared:

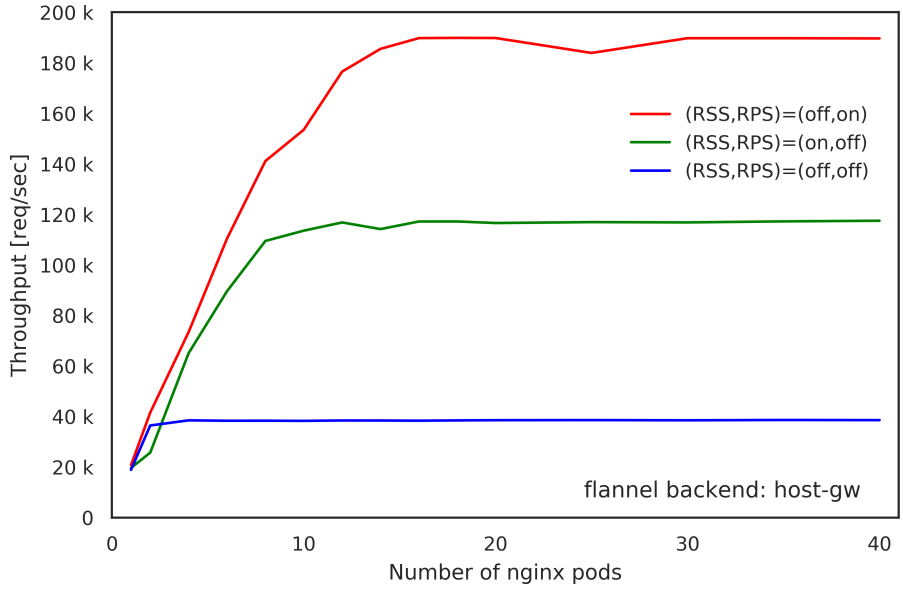


Figure 4.2: Effect of multicore proccessing on ipvs throughput.

```
(RSS, RPS) = (off, off)
           = (on , off)
           = (off, on )
```

The case with "(RSS, RPS) = (off, off)" means that multicore packet processing is completely disabled, i.e., all the incoming packets are processed by a single core. The "(RSS, RPS) = (on, off)" means that the interrupt handling and the following IP protocol processing are performed on four of the CPU cores by assigning four rx-queues to those cores. In this case four of the eight CPU cores are utilized. The "(RSS, RPS) = (off, on)" means that a single core handles all of the interrupts from the NIC then the following IP processings are performed on the other cores. In this case, all of the eight CPU cores are utilized.

We can see a general trend in which the throughput linearly increases as the number of nginx *pod*s increases and then it eventually saturates. The saturated throughput levels indicate the maximum performance level of the ipvs load balancer. The maximum performance levels depend on the (RSS, RPS) settings. From the results in this figure, it can be seen that if we turn off multicore packet processing, *i.e.*, when "(RSS, RPS) = (off, off)", performance degrades significantly.

If we compare the results for the cases when "(RSS, RPS) = (on, off)" and "(RSS,

RPS) = (off, on)", the latter is better than the former. It is clear that the case that utilizes all of the CPU cores better performs than the case with only four CPU cores utilized.

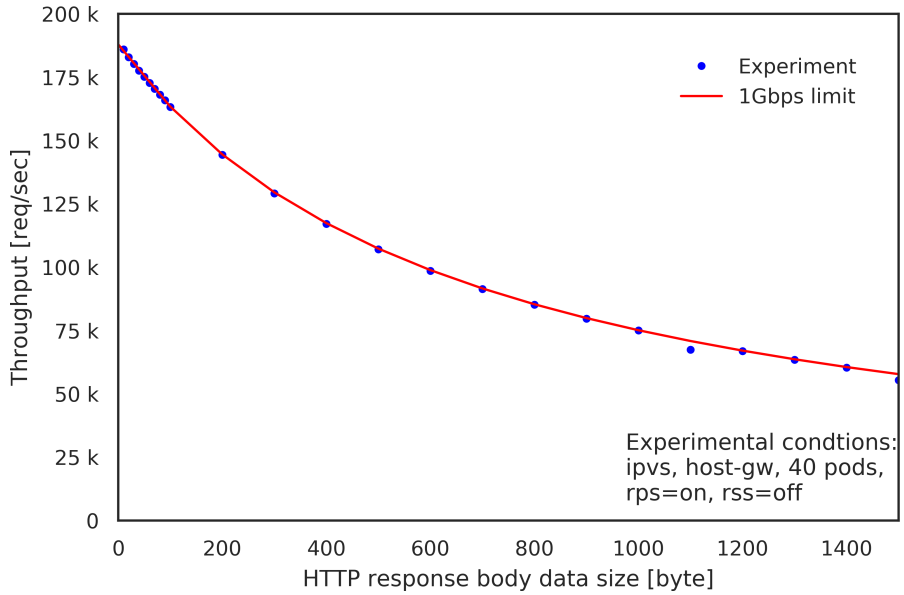


Figure 4.3: Performance limit due to 1Gbps bandwidth

At first, it was not clear what caused the performance limit for the case when "(RSS, RPS) = (off, on)", the author thought it was due to the insufficient CPU performance. However, that was not the case in the conditions of the experiment; it turned out to be due to the 1Gbps bandwidth. A packet level analysis using tcpdump[22] revealed that 665.36 bytes of extra HTTP headers, TCP/IP headers and ethernet frame headers are needed for each request in the case of the wrk benchmark program(Appendix C). This results in the upper limit of 184,267 [req/sec] when the date size of HTTP response body is 13 byte, which agrees well with the performance limit for the case when "(RSS, RPS) = (off, on)" in Figure 4.2. Figure 4.3 shows the theoretical upper limit of the performance level for 1Gbps ethernet together with actual benchmark results for the range of larger data sizes, and they agree very well. Therefore it can be said that when "RPS = on", ipvs performance is limited by 1Gbps bandwidth. The author regarded that "(RSS, RPS) = (off, on)" is the best setting in our experimental conditions, and used this setting throughout this thesis unless explicitly stated otherwise.

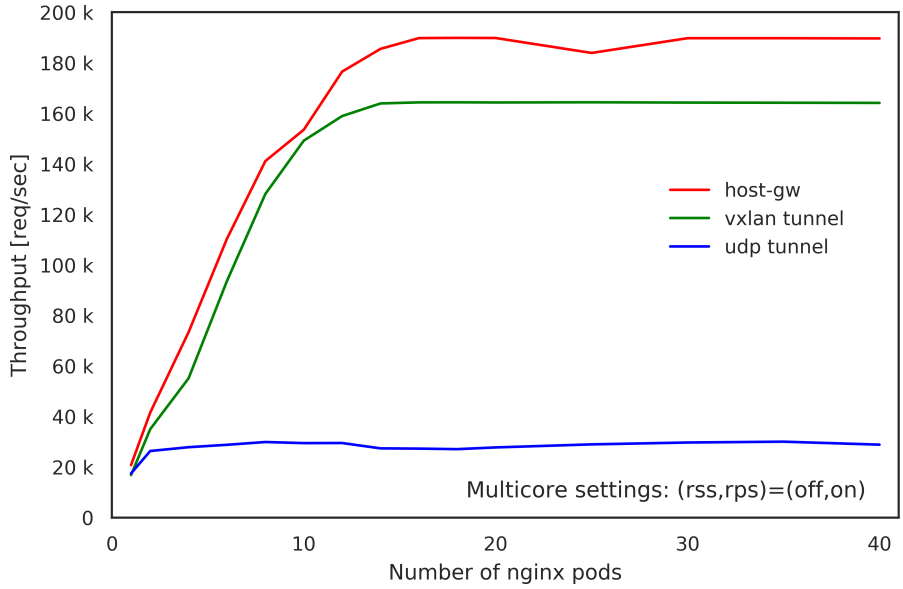


Figure 4.4: Effect of flannel backend modes on ipvs throughput.

Effect of overlay network

Figure 4.4 shows the ipvs throughput results for different overlay network settings. As for the overlay network, the author used the flannel and measured the performance levels for flannel's three backend modes, host-gw, vxlan and udp. Except for the udp backend mode case, we can see the trend in which the throughput linearly increases as the number of nginx *pod* increases and then it eventually saturates. The saturated throughput levels indicate the maximum performance levels of the ipvs load balancer. If we compare the performance levels among the flannel backend modes types, the host-gw mode where no encapsulation is conducted shows the highest performance level, followed by the vxlan mode where the Linux kernel encapsulate the Ethernet frame. The udp mode where flanneld itself encapsulate the IP packet shows significantly lower performances levels. The author considers the host-gw mode is the best, the vxlan tunnel the second best and the udp tunnel mode unusable. As is shown here, overlay network settings greately affect the performance level. The author used host-gw mode for most of the experiments conducted in on-premise data centers and vxlan mode for the experiments conducted in cloud environments.

Comparison of different load balancer

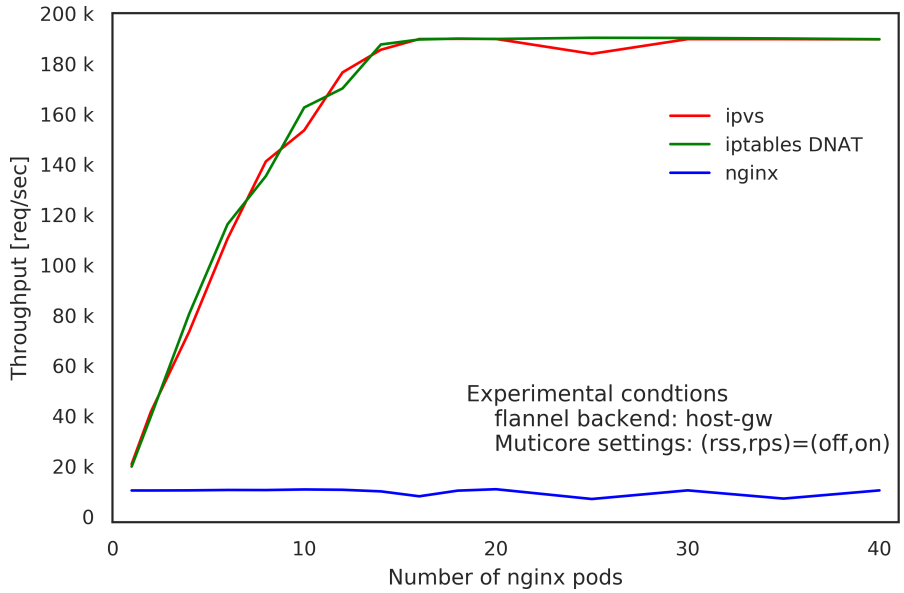


Figure 4.5: Throughput comparison between ipvs, iptables DNAT and nginx.

Figure 4.5 compares the performance measurement results for different load balancer ipvs, iptables DNAT, and nginx. The proposed ipvs load balancer exhibits almost equivalent performance levels as the iptables DNAT based load balancer. The nginx based load balancer shows no performance improvement even though the number of the nginx web server *pods* is increased. It is understandable because the performance of the single nginx as a load balancer is expected to be similar to the performance as a web server.

Figure 4.6 compares Cumulative Distribution Function(CDF) of the load balancer latency at the two constant loads, 160K[req/sec] and 180K[req/sec] for ipvs and iptables DNAT. We can see that the latencies are a little bit smaller for ipvs. For example, the median values at 160K[req/sec] load for ipvs and iptables DNAT are, 1.14 msec and 1.24 msec, respectively. Also, at 160K[req/sec], they are 1.39 msec and 1.45 msec, respectively. These may not be considered a significant difference; however, we can at least say that our proposed load balancer is as good as iptables DNAT. So, to conclude this section, the containerized ipvs load balancer showed equivalent performance levels with the iptables DNAT load-balancing function that is used in Kubernetes cluster.

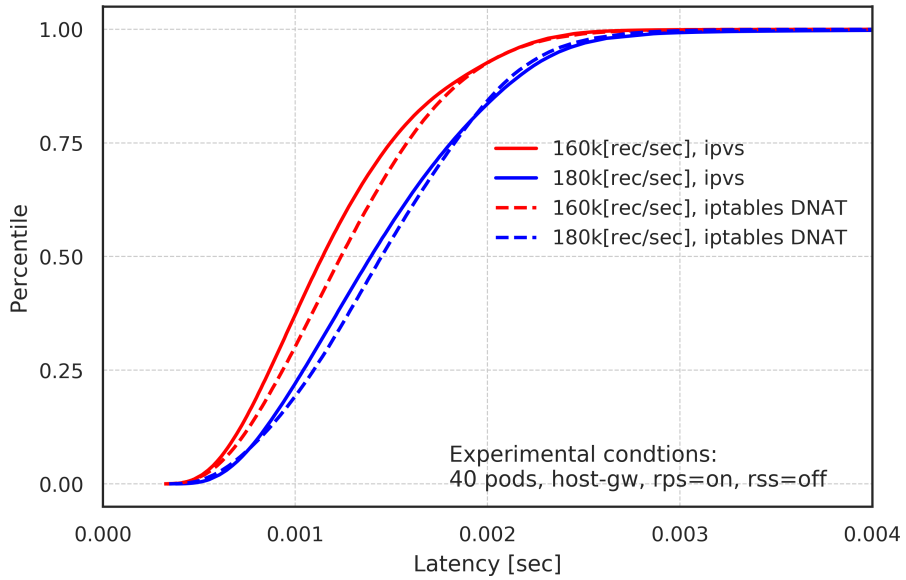


Figure 4.6: Latency cumulative distribution function.

L3DSR using ipvs tun

The performance levels of ipvs and iptables DNAT have been limited by 1 Gbps bandwidth. This can be alleviated in the case of ipvs by using so-called Layer 3 Direct Server Return(l3dsr) setup. Figure 4.7 shows the schematic diagram illustrating packet flow for the HTTP request packet(the red arrows) and response packet(the blue arrow).

The ipvs has the mode called ipvs-tun. When the ipvs-tun send out the packets to real servers, it encapsulates the original packet in ipip tunneling packet that is destined to real servers. The real server receives the packet on a tunl0 device and decapsulates the ipip packet, revealing the original packet. Since the source IP address of the original packet is maintained, the returning packets are sent directly toward the benchmark client. In this scheme, the returning packets do not consume the bandwidth nor the CPU power of the load balancer node.

The iptables DNAT does not have the functions that enable L3DSR settings. Therefore this one of the benefits of the ipvs load balancer.

The author carried out throughput measurement using the physical setup shown in Figure 4.7. Figure 4.8 shows the throughput of the ipvs-tun, conventional ipvs (after here the author call it ipvs-nat) and iptables DNAT. As can be seen in the figure, while the performance levels for ipvs-nat and iptables DNAT exactly match, the performance

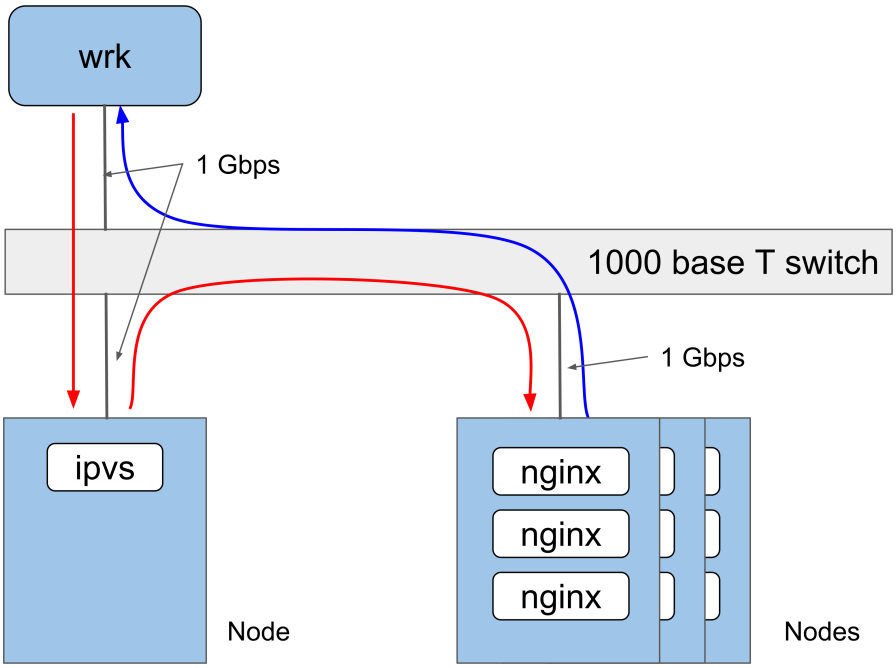


Figure 4.7: Physical configuration for L3DSR experiment.

levels for ipvs-tun is greatly improved, e.g., 1.5 times larger saturated throughput than for ipvs-nat and iptables DNAT cases.

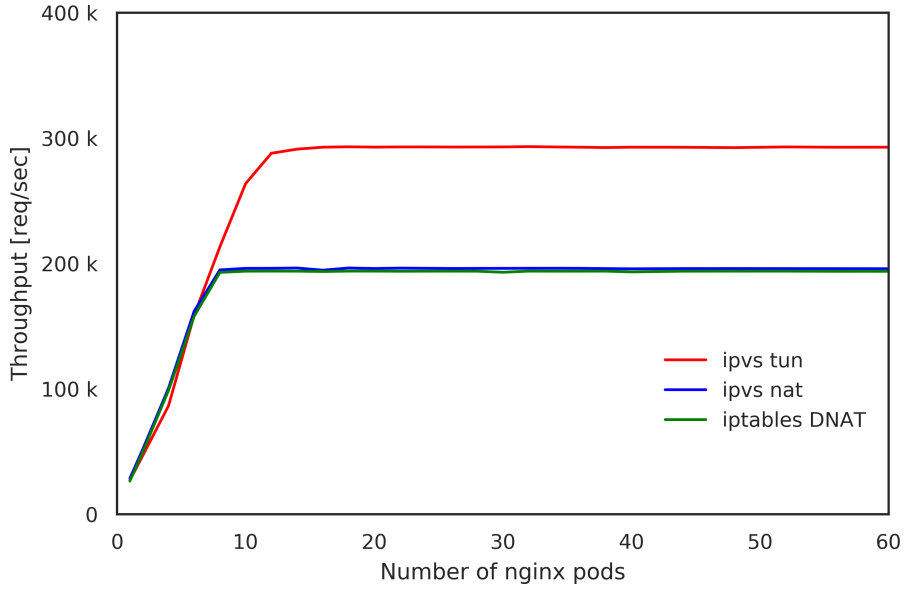


Figure 4.8: Throughput of ipvs l3dsr @1Gbps.

4.2 ECMP redundancy

This section discusses the redundancy and scalability of the proposed load balancers. The ECMP technique is expected to make the load balancers redundant and scalable since all the load balancer containers act as active. The whole system is resilient to a single failure of load balancer container. Also since multiple of load balancers can be utilized simultaneously, it is expected that the throughput of the total system is increased significantly. In order to evaluate these characteristics of the ECMP technique, the author examined if the ECMP routing table is updated correctly when multiple of the load balancer *pods* are started. After that, in order to explore the scalability, the author also measured the throughput of the cluster of load balancers. Finally, the author examined how quick those ECMP routing table updates are. The following sections explain the evaluation in detail.

4.2.1 Evaluation method

Figure 4.9 shows the schematic diagram of the experimental setup. And Table 4.3 summarizes hardware and software specifications for the experiments. Multiple *pods* are deployed on multiple nodes in the Kubernetes cluster. In each *pod*, an nginx web

server pod that returns the IP address of the *pod* are running. There are multiple nodes for load balancers and on each of the nodes, single load balancer *pod* is deployed. Each load balancer *pod* consists of both an ipvs container and an exabgp container. The routing table of the benchmark client is updated by BGP protocol through a route reflector.

Using these hardware and software setups the following four types of evaluations have been carried out; 1) Evaluation of ECMP functionality. The author examined if ECMP routing table is correctly updated. 2) Evaluation of the scalability. The author evaluated how throughput is improved by running multiple ipvs pods simultaneously. 3) Evaluation of ECMP response. The author evaluated the delay between the time ipvs pods are started or stopped until the time ECMP routing table reflected the change.

The throughputs are measured using wrk in the same manner as in Chapter **??**. Notable differences from the previous throughput experiment in Figure 4.1 are; There four nodes for load balancers instead of one. Also, the 10 Gbps NIC is used for the benchmark client since for scalability experiment, there are multiple of ipvs container load balancers that can fill up 1 Gbps bandwidth. Some of the software has also been updated to the most recent versions at the time of the experiment.

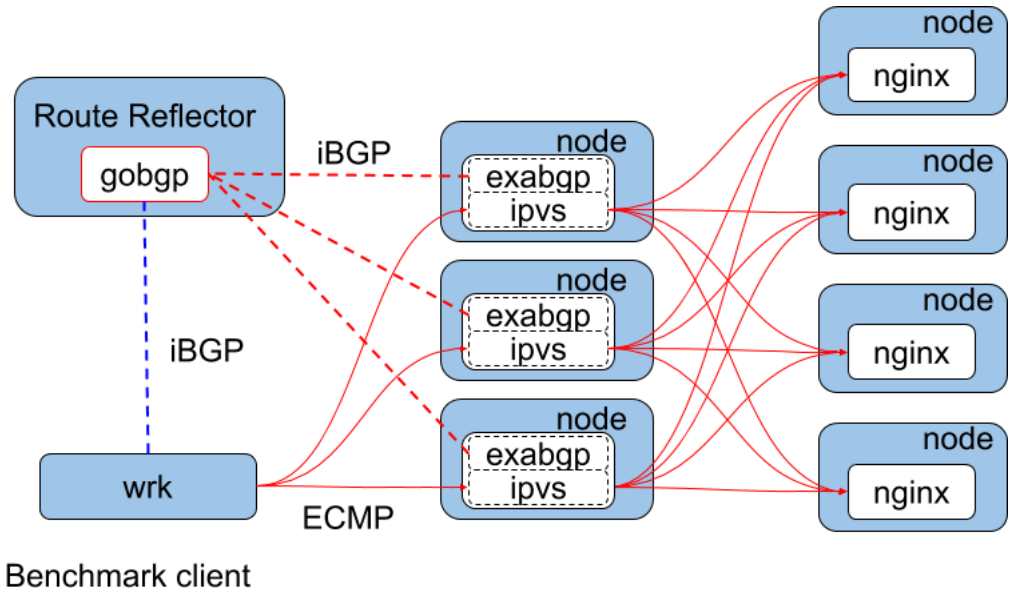


Figure 4.9: Experimental setups.

[Hardware Specification] CPU: Xeon E5-2450 2.10GHz x 8 (with Hyper Threading) Memory: 32GB NIC: Broadcom BCM5720 Giga bit (Node x 6, Load Balancer x 4) CPU: Xeon E5-2450 2.10GHz x 8 (with Hyper Threading) Memory: 32GB NIC: Intel X550 (Client x 1)
[Node Software] OS: Debian 9.5, linux-4.16.8 Kubernetes v1.5.2 flannel v0.7.0 etcd version: 3.0.15
[Container Software] Keepalived: v1.3.2 (12/03,2016) nginx : 1.15.4(web server)

Table 4.3: Hardware and software specifications.

4.2.2 Reults

ECMP functionality

10.1.1.0/24 via 10.0.0.106 dev eth0 proto zebra metric 20

(a) With single load balancer *pod*.

10.1.1.0/24 proto zebra metric 20
nexthop via 10.0.0.105 dev eth0 weight 1
nexthop via 10.0.0.106 dev eth0 weight 1
nexthop via 10.0.0.107 dev eth0 weight 1

(b) With three load balancer *pod*s.

10.1.1.0/24 pro to zebra metric 20
nexthop via 10.0.0.107 dev eth0 weight 1
nexthop via 10.0.0.105 dev eth0 weight 1
nexthop via 10.0.0.106 dev eth0 weight 1
10.1.2.0/24 proto zebra metric 20
nexthop via 10.0.0.107 dev eth0 weight 1
nexthop via 10.0.0.106 dev eth0 weight 1

(c) For a service with three load balancer *pod*s and a service with two load balancer *pod*s.

Table 4.4: ECMP routing tables.

First, the author examined ECMP functionality by monitoring the routing table on the benchmark client. Table 4.4 (a) shows the routing table entry on the router when a single load balancer pod existed. From this entry, we can tell that packets toward 10.1.1.0/24 are forwarded to 10.0.0.106 where the load balancer pod is running. There is also a keyword, zebra, which indicates that zebra controls this routing rule.

When the number of the load balancer pods was increased to three, the routing table becomes to have entries in Table 4.4 (b). There are three next hops towards 10.1.1.0/24 each of which being the node where the load balancer pods are running. The weights of the three next-hops are equal, i.e., 1. The update of the routing entry was almost instant as the author increased the number of the load balancers.

Table 4.4 (c) shows the case where the author additionally started new service with two load balancer pods with service addresses in 10.1.2.0/24 range. It was possible to accommodate two different services with different IP addresses, one with three

load balancers and the other with two load balancers on a group of nodes, 10.0.0.105, 10.0.0.106 and 10.0.0.107. The update of the routing entry was almost instant as the author started the load balancers for the second service.

Scalability

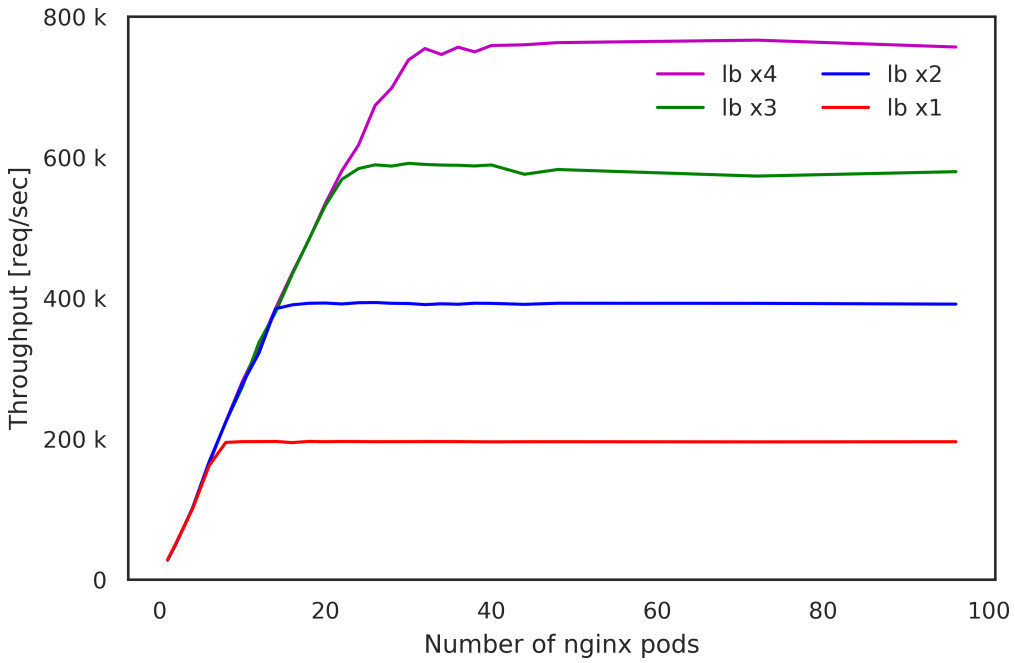


Figure 4.10: Throughput of ECMP redundant load balancer.
The throughputs are measured for a single load balancer(lb x1), two(lb x2), three(lb x3) and four(lb x4) load balancers.

The throughput measurement was also carried out to show that ECMP technique increases the throughput as the number of the load balancers is increased. Figure 4.10 shows the results of the measurements. There are four solid lines in the figure, each corresponding the throughput result when there are one through four of the proposed load balancers.

As can be seen in the figure, as we increased the number of the pod the throughput increased linearly to a certain level after which it saturated. The saturated levels, i.e. performance levels, depend on the number of the ipvs load balancer pods (lb x 1 being the case with one ipvs pods, and lb x2 being two of them and as such). The performance levels increase linearly as we increase the number of the load balancers.

The performance level did not scale further when the number of load balancers was increased more than four. This was because the performance of the benchmark client was hitting the ceiling, i.e., the CPU usage was 100% when the total throughput was around 780k [req/sec]. The author expects that replacing the benchmark client with more powerful machines will further improve the performance level this system.

Response

Figure 4.11 shows the histogram of the ECMP update delay. The author measured the delays until the number of running ipvs pods is reflected into the routing table on the benchmark client. The number of the ipvs pods is changed randomly every 60 seconds for 20 hours. As we can see from the figure, most of the delays are within 6 seconds, and the largest delay was 10 seconds. We can say that ECMP routing update in our proposed architecture is quick enough.

Figure 4.12 shows the throughput measurement results when the number of the load balancers was periodically changed. The red line in the figure shows the number of the ipvs load balancer pods, which was changed randomly in every 60 seconds. The blue line corresponds to the resulting throughput. As we can see from the figure, the blue line nicely follows the shape of the red line. This indicates that new load balancers are immediately utilized after they are created. It also indicates that after removing some load balancers, the traffic to them is immediately directed to the existing load balancers.

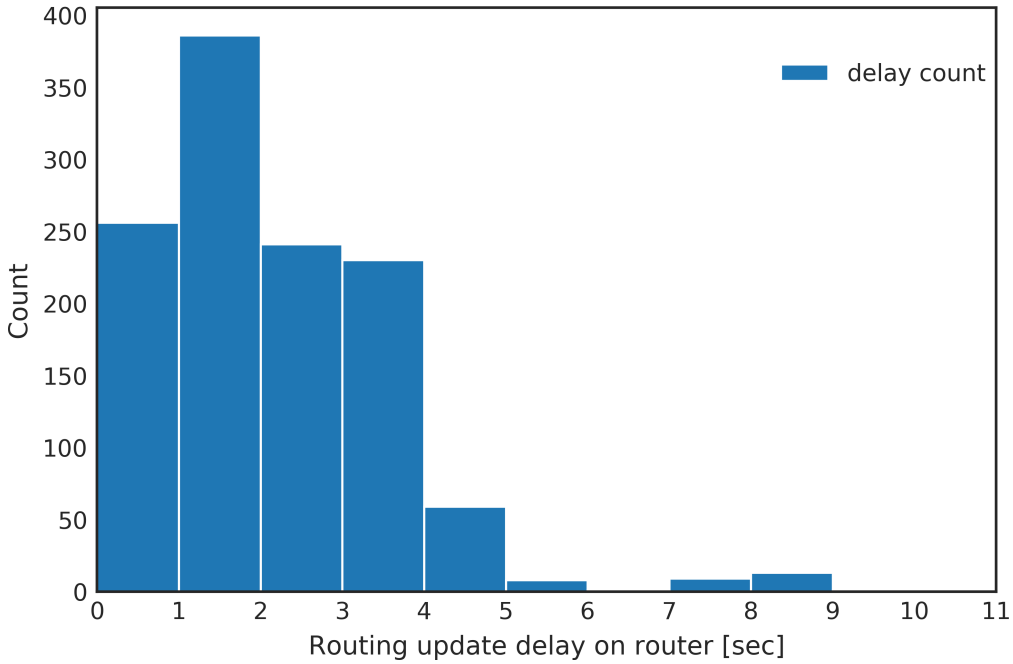


Figure 4.11: A histogram of the ECMP update delay.
This shows the delays until the number of running ipvs pods is reflected into the routing table on the benchmark client, when the number of the ipvs pods is changed randomly every 60 seconds for 20 hours.

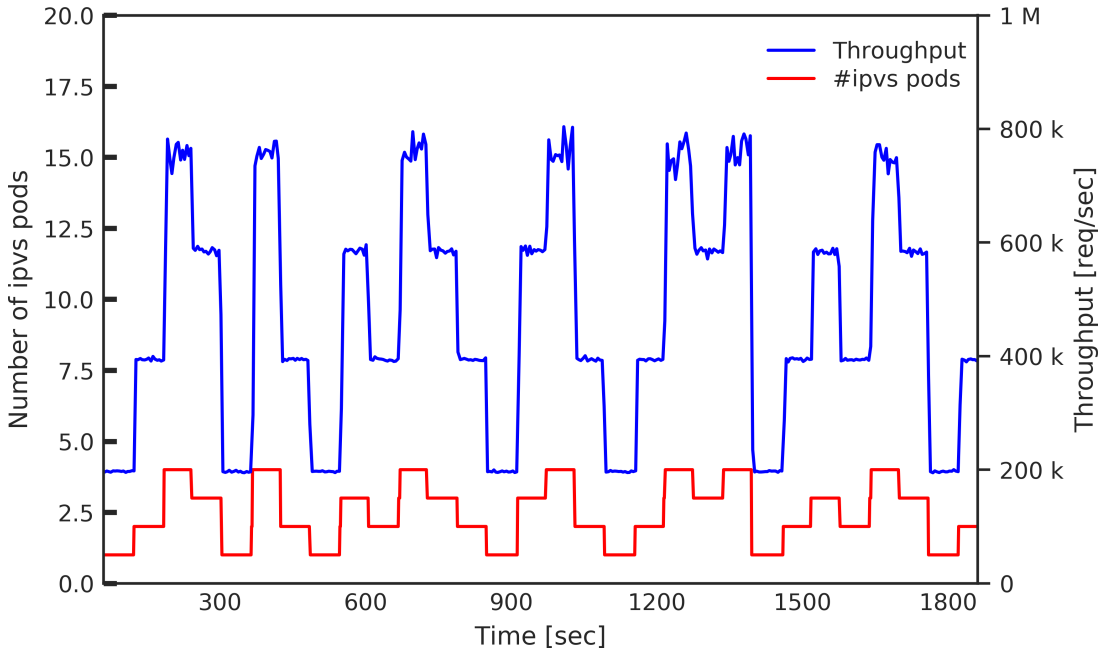


Figure 4.12: Throughput responsiveness.
This shows the throughput responsiveness when the number of the load balancers was changed randomly in every 60 seconds.

4.3 Cloud experiment

4.3.1 Method

4.3.2 Results

So far, it has been shown that the proposed ipvs-nat load balancer in a container has equivalent throughput, and the proposed ipvs-tun load balancer in a container has even better throughputs. In this section, the author shows that the proposed load balancer is portable by showing that it can be run in cloud environments, and also shows that it has the same behavior as in on-premise data centers.

Figure 4.13 and Figure 4.14 show the load balancer performance levels that are measured in GCP and AWS, respectively. For both environments, the author measured throughput with several conditions of CPU counts, since the machine specifications can be easily changed in the cases of cloud environments. Both results show similar characteristics as the experiment in an on-premise data center in Figure 4.2, where throughput increases linearly to a certain saturation level that is determined by utilized CPU core count. In other words, it indicates that the proposed load balancer can be run in cloud environments and also functions properly.

It seems that CPU counts determine the load balancer's throughput saturation levels. The actual throughput numbers are smaller than those of the load balancers in on-premise data centers. This may be because the physical servers in on-premise data center outperform the VMs in a cloud environment, or because network bandwidth is smaller and is limited based on the type of instances. A detailed analysis is further required in the future to clarify which factor limits the throughput in the case of the cloud environment. Nonetheless, we can say that the proposed ipvs load balancers can be run in both GCP and AWS, and the behavior is the same with the load balancers in on-premise data centers.

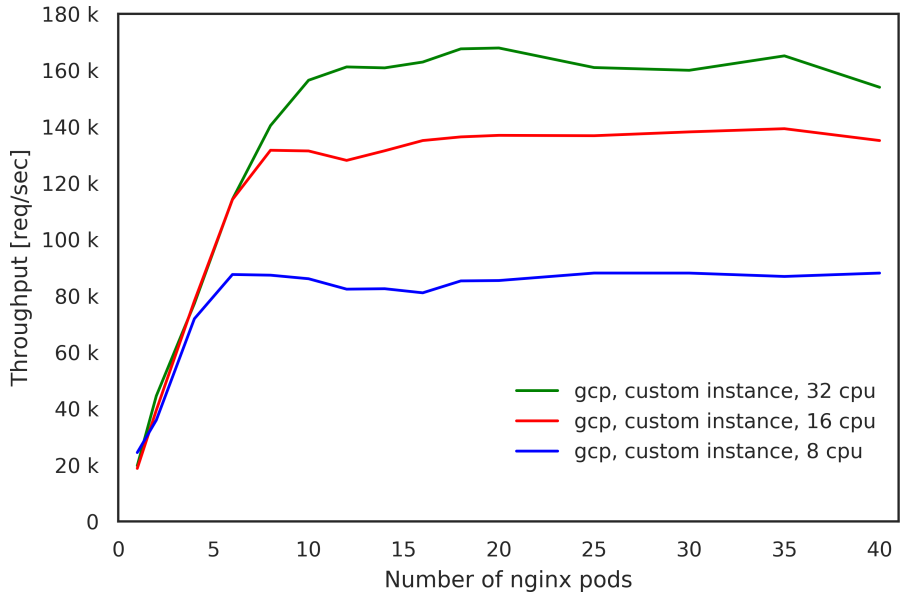


Figure 4.13: GCP

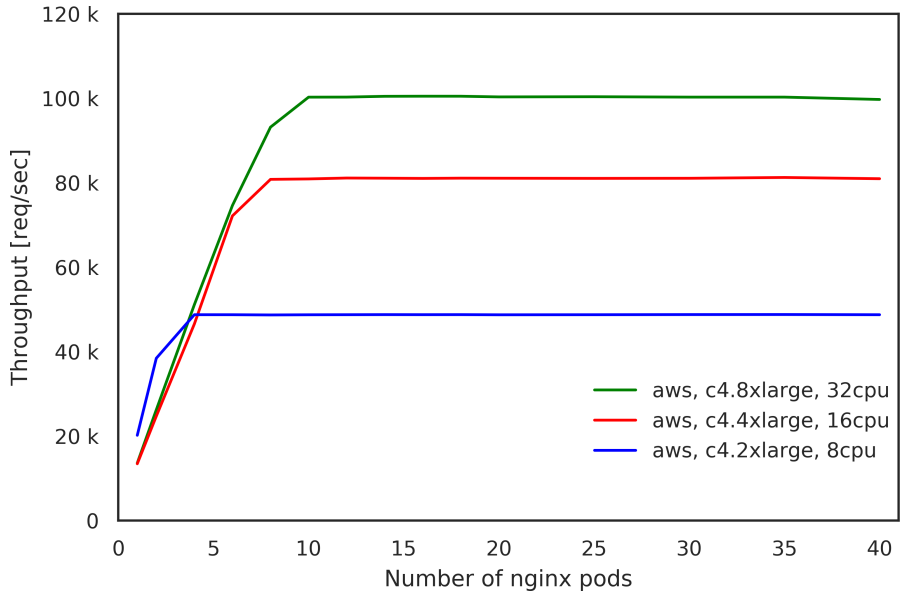


Figure 4.14: AWS with Node x 6, Client x 1, Load balancer x 1. Custom instance.

4.4 Summary

4.4.1 singl lb

In this chapter portability and performance level of proposed load balancer in 1 Gbps network environments has been discussed. The throughput levels of a load balancer are dependent on settings for multicore packet processing. It is clear that the case that utilizes all of the CPU cores better performs than the case with only four CPU cores utilized. It is better to use as many CPU cores as possible for packet processing. The throughput levels are also very dependent on the back end mode of the flannel overlay network. The host-gw mode where no tunneling is used resulted in the best performance level.

The performance levels of ipvs-nat, iptables DNAT and nginx have also been compared. The proposed ipvs-nat load balancer in the container had the same performance level as load balancing function of iptables DNAT. Furthermore, in the case of L3DSR setup, the performance level of ipvs-tun load balancer has about 1.5 times larger than that of ipvs-nat and iptables DNAT.

It is also shown that the proposed load balancer can be run in GCP and AWS. The behavior of the proposed load balancer in those cloud environments is the same as that in the on-premise data center. The author concludes that the proposed load balancer is portable and outperforms the existing iptables DNAT load balancers in 1 Gbps network environments.

4.4.2 ECMP

In this chapter, the redundancy and scalability of the proposed load balancers have been discussed. The author verified that ECMP routing table was properly created in the experimental system. The update of the ECMP routing table was quick enough, i.e., within 10 seconds, throughout 20 hours experiment and the routing table was always correct. The scalability of the load balancer was also examined and it has been found that maximum performance levels scaled linearly as the number of the load balancer pods was increased to four. The maximum throughput level obtained through the experiment was 780k [req/sec], which is limited due to the maximum CPU performance of the benchmark client rather than the performance of the load balancer cluster.

5 Further Improvement

Up until this chapter most of the experiments are done in 1Gbps network environments. The proposed load balancers have shown decent performance levels in 1Gbps environment. However, it is essential to investigate the feasibility of the proposed load balancers in 10Gbps network environments. In this chapter, the author carries out throughput measurements of ipvs-nat, ipvs-tun, and iptables DNAT in 10Gbps environment. Then the author improves the performance levels of ipvs-nat and ipvs-tun by setting up these load balancers in the node net namespace. Also presented is the novel software load balancer using eXpress Data Plane(XDP) technology, as an alternative to ipvs software load balancers.

5.1 Throuput of ipvs-nat, ipvs-tun and iptables DNAT

Figure 5.1 shows the packet flow for ipvs-nat and iptables DNAT, and Figure 5.2 shows that for ipvs-tun. The 10Gbps NICs are used for benchmark client and the node for the load balancer. In the case of the ipvs-nat and iptables DNAT, the response packets from

nginx pods are returned to the load balancer, and then load balancer returns it to the client. In contrast, in the case of ipvs-tun, the response packets from nginx pods are directly returned to the client. Since the load balancer does not have to process the response packet, a better performance level is expected for ipvs-tun.

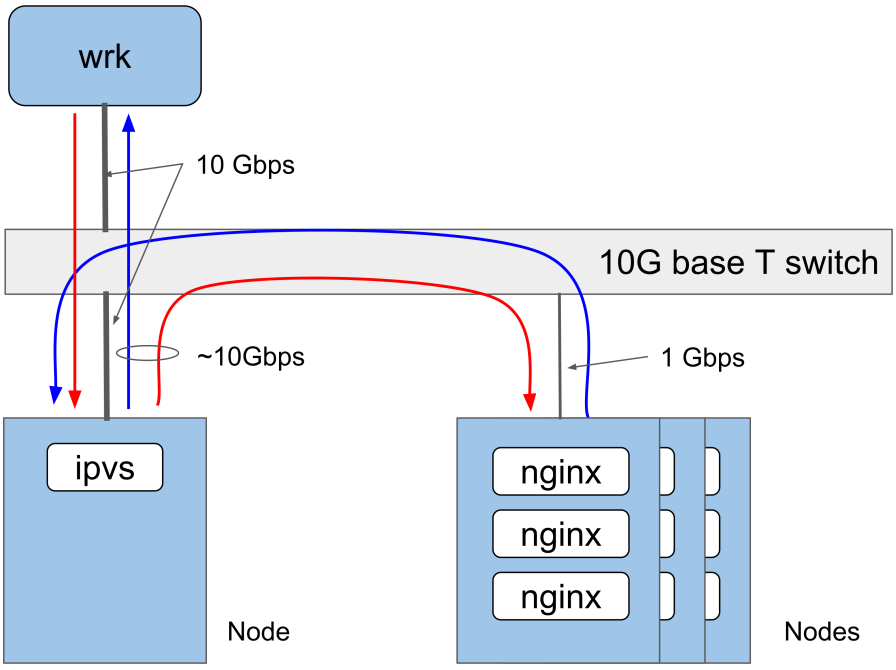


Figure 5.1: Packet flow of ipvs-nat and iptables DNAT.

Figure 5.3 shows the throughput of ipvs-tun, ipvs-nat and iptables DNAT in 10Gbps environment. We can see the general characteristics of a load balancer where the throughput increases linearly to a certain level as the number of nginx container increases, and then eventually saturates. These saturation levels are the performance limits of each of the load balancers, which is determined by packet forwarding efficiency or the bandwidth of the network. The performance limit of the iptables DNAT is close to 780k [req/sec], where the CPU usage of the benchmark client becomes 100%.

Table 5.1 summarizes the throughput of ipvs-tun, ipvs-nat and iptables DNAT at 40 nginx pods in 10 Gbps and 1 Gbps networks. By using 10 Gbps network, the performance levels for all of these load balancer are improved. However, the magnitudes of the improvements are different among the types of the load balancers. While the throughput of the ipvs-nat is 334833 [req/sec], that of the iptables DNAT is 777640

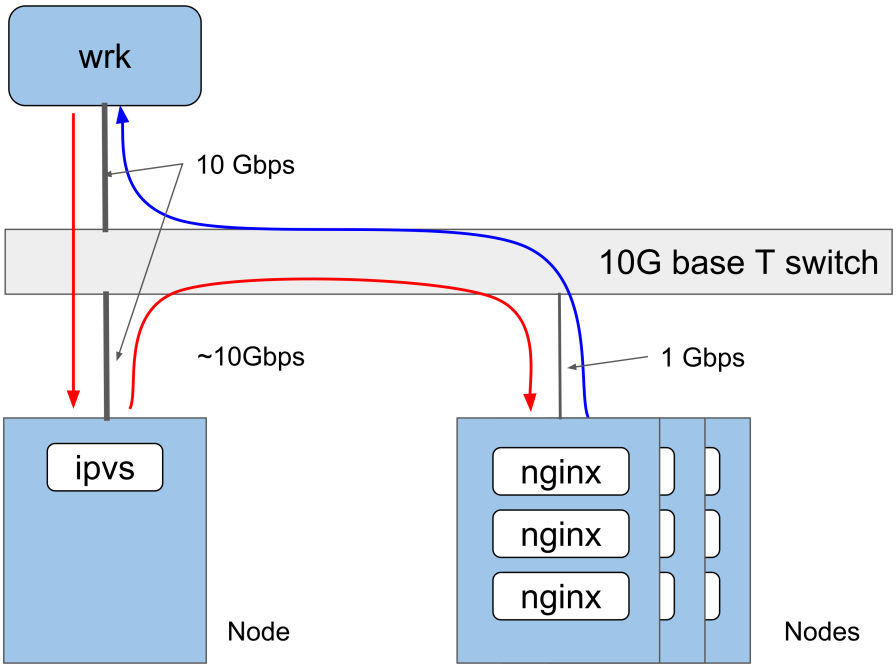


Figure 5.2: Packet flow of ipvs-tun.

[req/sec]. This suggests that the packet forwarding of the iptables DNAT is more efficient than that of ipvs-nat. Although the throughput of the ipvs-tun, 730975 [req/sec] is better than ipvs-nat because of the L3DSR settings, it still falls short of that of iptables DNAT. It seems that containerized ipvs load balancers are inherently less efficient than the iptables DNAT, which could be attributed to either overhead of container network(veth+bridge)[42, 35] or kernel code for ipvs itself. In order to investigate this issue, the author conducted a throughput measurement for ipvs-nat and ipvs-tun that are set up in node net namespaces in the next section.

	Throughput [req/sec]	
Type of load balancer	1Gbps	10Gbps
iptables DNAT	193198	777640
ipvs-nat	195666	334833
ipvs-tun	292660	730975

Table 5.1: Performance levels in 1Gbps and in 10Gbps.
Throughput results of the load balancers at 40 nginx pods from the data for the Figures 4.8 and 5.3 are shown.

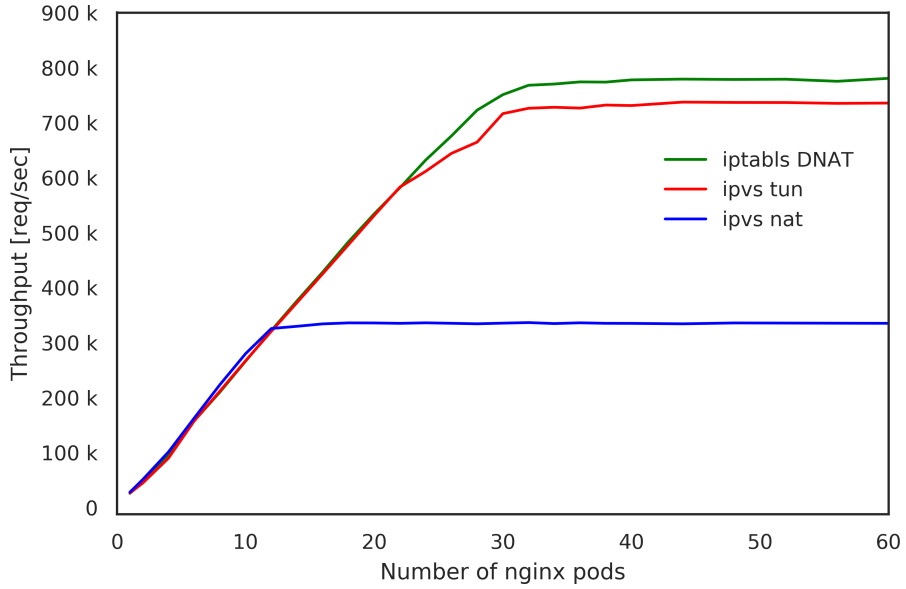


Figure 5.3: Throughput of load balancers in 10 Gbps.

5.2 Throuput of ipvs-nat, ipvs-tun and iptables DNAT

In order to improve the throughput of the ipvs load balancers by removing the overhead of container network, the ipvs load balancers were set up in node net namespaces. Appendix XXX shows inside of ipvs container script that launches the keepalived in node net namespaces. By doing so, the load balancing tables are created in the node net namespace.

Figure 5.4 shows the throughput of ipvs-nat and ipvs-tun in the node namespace together with the throughput of the iptables DNAT. The throughput of the ipvs-tun is almost identical to that of iptables DNAT, which is limited by CPU power of the benchmark client. Although the throughput of the ipvs-nat is smaller than that of the iptables DNAT, it is clearly improved from the result in Figure 5.3.

Table 5.2 compares the throughput of ipvs load balancers in the pod namespace and in the node namespace at 40 nginx pods. The thoughput data are taken from the results in the Figures 5.3 and 5.4. We can see that maximum throughputs can be improved in the case of load balancers in node namespace both for the ipvs-nat and the ipvs-tun.

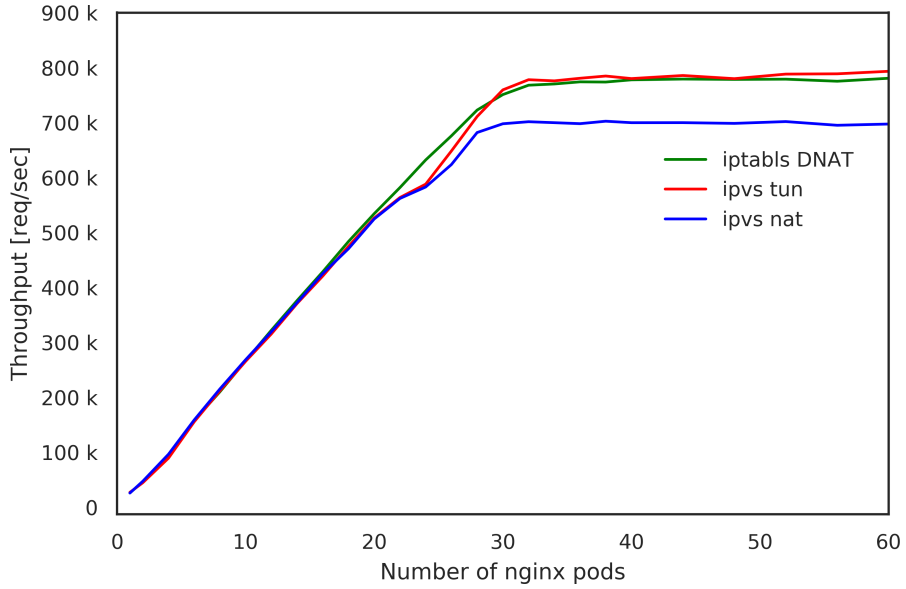


Figure 5.4: Throughput of load balancers in node name space.

	Throughput [req/sec]	
Type of load balancer	pod name space	node name space
iptables DNAT	NA	777640
ipvs-nat	334833	699635
ipvs-tun	730975	779932

Table 5.2: Performance levels in pod namespace and in node namespace. Throughput results of the load balancers at 40 nginx pods from the data for the Figures 5.3 and 5.4 are shown.

5.3 XDP load balancer

[Filled in later]

5.4 Summary

In this chapter, the author carried out throughput measurements of ipvs-nat, ipvs-tun, and iptables DNAT in 10Gbps environment. From the results the general characteristics of a load balancer are observesd. The throughput increases linearly to a certain level as the number of nginx container increases, and then eventually saturates. The performance levels for of the load balancers are improved by using 10Gbps network.

However, the throughputs of ipvs-nat and ipvs-tun are smaller than that of iptables DNAT.

Then the author improved the performance levels by setting up ipvs-nat and ipvs-tun load balancers in the node net namespace to remove overhead of the container network. The throughput of the ipvs-tun became almost identical to that of iptables DNAT, and the throughput of the ipvs-nat also improved to the level close to that of the iptables DNAT.

[Filled in later]

6 Related Work

This Chapter provides related works and the background information of this study.

6.1 Related Work

This section highlights related work, especially that dealing with container cluster migration, software load balancer containerization, load balancer tools within the context of the container technology and scalable load balancer in the cloud providers.

Container cluster migration: Kubernetes developers are trying to add federation[1] capability for handling situations where multiple Kubernetes clusters[1] are deployed on multiple cloud providers or on-premise data centers, and are managed via the Kubernetes federation API server (federation-apiserver). However, how each Kubernetes cluster is run on different types of cloud providers and/or on-premise data

[1]The *Kubernetes cluster* refers to a server cluster controlled by the Kubernetes container management system, in this paper.

centers, especially when the load balancers of such environments are not supported by Kubernetes, seems beyond the scope of that project. The main scope of this paper is to make Kubernetes usable in environments without supported load balancers by providing a containerized software load balancer.

Software load balancer containerization: As far as load balancer containerization is concerned, the following related work has been identified: Nginx-ingress[33, 20] utilizes the ingress[2] capability of Kubernetes, to implement a containerized Nginx proxy as a load balancer. Nginx itself is famous as a high-performance web server program that also has the functionality of a Layer-7 load balancer. Nginx is capable of handling Transport Layer Security(TLS) encryption, as well as Uniform Resource Identifier(URI) based switching. However, the flip side of Nginx is that it is much slower than Layer-4 switching. We compared the performance between Nginx as a load balancer and our proposed load balancer in this paper. Meanwhile, the kube-keepalived-vip[34] project is trying to use Linux kernel's ipvs[41] load balancer capabilities by containerizing the keepalived[6]. The kernel ipvs function is set up in the host OS's net namespaces and is shared among multiple web services, as if it is part of the Kubernetes cluster infrastructure. Our approach differs in that the ipvs rules are set up in container's net namespaces and function as a part of the web service container cluster itself. The load balancers are configurable one by one, and are movable with the cluster once the migration is needed. The kube-keepalived-vip's approach lacks flexibility and portability whereas ours provide them. The swarm mode of the Docker[14, 11] also uses ipvs for internal load balancing, but it is also considered as part of Docker swarm infrastructure, and thus lacks the portability that our proposal aims to provide.

Load balancer tools in the container context: There are several other projects where efforts have been made to utilize ipvs in the context of container environment. For example, GORB[36] and clusterf[25] are daemons that setup ipvs rules in the kernel inside the Docker container. They utilize running container information stored in key-value storages like Core OS etcd[10] and HashiCorp's Consul[18]. Although these were usable to implement a containerized load balancer in our proposal, we did not use them, since Kubernetes ingress framework already provided the methods to

retrieve running container information through standard API.

Cloud load balancers: As far as the cloud load balancers are concerned, two articles have been identified. Google's Maglev[13] is a software load balancer used in Google Cloud Platform(GCP). Maglev uses modern technologies including per flow ECMP and kernel bypass for user space packet processing. Maglev serves as the GCP's load balancer that is used by the Kubernetes. Maglev is not a product that users can use outside of GCP nor is an open source software, while the users need open source software load balancer that is runnable even in on-premise data centers. Microsoft's Ananta[31] is another software load balancer implementation using ECMP and windows network stack. Ananta can be solely used in Microsoft's Azure cloud infrastructure[31]. The proposed load balancer by the author is different in that it is aimed to be used in every cloud provider and on-premise data centers.

7 Conclusion and future work

7.1 Conclusions

In this dissertation, the author proposed a portable load balancer with ECMP redundancy for the Kubernetes cluster systems that is aimed at facilitating migration of container clusters for web services. The proposed load balancer architecture utilizes software load balancers with container technology to make the load balancers runnable in any base infrastructure. It also utilizes ECMP technology to make multiple load balancers active, and thereby to provide redundancy and scalability.

The author implemented a containerized software load balancer that is run by Kubernetes as a part of container cluster, using Linux kernel's IPVS. In order to discuss the feasibility of the proposed load balancer, performance measurements are conducted in 1 Gbps network environment. It was shown that the proposed load balancers are runnable in an on-premise data center, GCP and AWS. Therefore the proposed load balancers can be said to be portable. The throughput levels of a load balancer are dependent on settings for multi-core packet processing. It was shown that better to use

as many CPU cores as possible for packet processing. The throughput levels are also very dependent on the overlay network backend mode. The host-gw mode where no tunneling is used resulted in the best performance level, and the vxlan mode resulted in the second best. In the experiment in 1 Gbps network environment, the ipvs-nat load balancer in the container had the same performance level as load balancing function of iptables DNAT on the node. Furthermore, the performance level of ipvs-tun load balancer in a container with the L3DSR setup was about 1.5 times larger than that of iptables DNAT. Therefore in 1 Gbps network environment, the proposed load balancer is portable while it has the 1.5 times better performance level or the same performance level depending on the mode of operation.

Also implemented is the ECMP setups where multiple of the load balancer containers are deployed, each advertising the route to the service VIP. The ECMP technique makes the load balancers redundant and scalable since all the load balancer containers act as active. The whole system is resilient to a single failure of load balancer container. Also by utilizing multiple of load balancers simultaneously, the throughput of the total system is increased significantly. These characteristics are evaluated by checking the routing table of the upstream router and by throughput measurement. The author verified that ECMP routing table was properly created in the experimental system. The update of the ECMP routing table was correct and quick enough, i.e., within 10 seconds, throughout 20 hours experiment. The maximum performance levels of the cluster of load balancers scaled linearly as the number of the load balancer pods was increased up to four of them. The maximum throughput level obtained through the experiment was 780k [req/sec], which is limited due to the maximum CPU performance of the benchmark client rather than the performance of the load balancer cluster.

The author also extended the throughput measurement into 10 Gbps network environment. It was revealed that ipvs-nat and ipvs-tun load balancers in containers had lower performance levels compared with the iptables DNAT. This has been suspected to be due to the overhead of the container network, i.e., veth+bridge. By setting up the load balancing table in node net namespaces, the performance levels of ipvs-nat and ipvs-tun became closer to that of the iptables DNAT. Although the overheads of the container network were invisible in 1 Gbps network environment, they were no longer invisible in 10 Gbps network environment. The author is currently implementing and evaluating a novel software load balancer using XDP technology to

provide a better alternative to ipvs as a portable load balancer.

The outcome of this study will benefit users who want to deploy their web services on any cloud provider where no scalable load balancer is provided, to achieve high scalability. Moreover, the result of this study will potentially benefit users who want to use a group of different cloud providers and on-premise data centers across the globe seamlessly. In other words, users will become being able to deploy a complex web service on aggregated computing resources on the earth, as if they were starting a single process on a single computer.

[Filled in later]

Although major cloud providers do not currently provide BGP peering service for their users, the authors expect our proposed load balancer will be able to run, once this approach is proven to be beneficial and they start BGP peering services. Therefore we focus our discussions on verifying that our proposed load balancer architecture is feasible, at least in on-premise data centers. For the cloud environment without BGP peering service, single instance of ipvs load balancer can still be run with redundancy. The liveness of the load balancer is constantly checked by one of the Kubernetes agents, and if anything that stop the load balancer happens, Kubernetes will restart the load balancer container. The routing table of the cloud provider can be updated by newly started ipvs container immediately.

The authors limit the focus of this study on providing a portable load balancer for Kubernetes to prove the concept of proposed architecture. However, the same concept can be easily applied to other container management systems, which should be discussed in future work.

Bibliography

[1] The Kubernetes Authors. *Federation*. 2017. URL: https://kubernetes.io/docs/concepts/cluster-administration/federation/.

[2] The Kubernetes Authors. *Ingress Resources | Kubernetes*. 2017. URL: https://kubernetes.io/docs/concepts/services-networking/ingress/.

[3] Bert Hubert et al. *Linux Advanced Routing & Traffic Control HOWTO*. 2002. URL: http://www.tldp.org/HOWTO/Adv-Routing-HOWTO/index.html (visited on 07/14/2017).

[4] Gilberto Bertin. "XDP in practice: integrating XDP into our DDoS mitigation pipeline". In: *Technical Conference on Linux Networking, Netdev*. Vol. 2. 2017.

[5] Brendan Burns et al. "Borg, omega, and kubernetes". In: (2016).

[6] Alexandre Cassen. *Keepalived for Linux*. URL: http://www.keepalived.org/.

[7] Brian F. Cooper. "Spanner: Google's Globally-distributed Database". In: *Proceedings of the 6th International Systems and Storage Conference*. SYSTOR '13. Haifa, Israel: ACM, 2013, 9:1–9:1. ISBN: 978-1-4503-2116-7. DOI: 10.1145/2485732.2485756. URL: http://doi.acm.org/10.1145/2485732.2485756.

[8] James C. Corbett et al. "Spanner: Google&Rsquo;s Globally Distributed Database". In: *ACM Trans. Comput. Syst.* 31.3 (Aug. 2013), 8:1–8:22. ISSN: 0734-2071. DOI: 10.1145/2491245. URL: http://doi.acm.org/10.1145/2491245.

[9] Inc CoreOS. *Backend*. URL: https://github.com/coreos/flannel/blob/master/Documentation/backends.md (visited on 07/14/2017).

[10] Inc CoreOS. *etcd | etcd Cluster by CoreOS*. URL: https://coreos.com/etcd (visited on 07/14/2017).

[11] Docker Inc. *Use swarm mode routing mesh | Docker Documentation*. 2017. URL: https://docs.docker.com/engine/swarm/ingress/ (visited on 07/14/2017).

[12] Jake Edge. *Creating containers with systemd-nspawn*. 2013. URL: https://lwn.net/Articles/572957/.

[13] Daniel E Eisenbud et al. "Maglev: A Fast and Reliable Software Network Load Balancer." In: *NSDI*. 2016, pp. 523–535.

[14] Docker Core Engineering. *Docker 1.12: Now with Built-in Orchestration! - Docker Blog*. 2016. URL: https://blog.docker.com/2016/06/docker-1-12-built-in-orchestration/.

[15] Exa-Networks. *Exa-Networks/exabgp*. July 2018. URL: https://github.com/Exa-Networks/exabgp.

[16] Mohammad Al-Fares, Alexander Loukissas, and Amin Vahdat. "A scalable, commodity data center network architecture". In: *ACM SIGCOMM Computer Communication Review*. Vol. 38. 4. ACM. 2008, pp. 63–74.

[17] Will Glozer. *wrk - a HTTP benchmarking tool*. 2012. URL: https://github.com/wg/wrk.

[18] HashiCorp. *Consul by HashiCorp*. URL: https://www.consul.io/ (visited on 07/14/2017).

[19] Benjamin Hindman et al. "Mesos: A platform for fine-grained resource sharing in the data center." In: *NSDI*. Vol. 11. 2011. 2011, pp. 22–22.

[20] NGINX Inc. *NGINX Ingress Controller*. 2017. URL: https://github.com/nginxinc/kubernetes-ingress.

[21] *ip-sysctl.txt*. URL: https://www.kernel.org/doc/Documentation/networking/ip-sysctl.txt.

[22] Van Jacobson, Craig Leres, and S McCanne. "The tcpdump manual page". In: *Lawrence Berkeley Laboratory, Berkeley, CA* 143 (1989).

[23] ktaka-ccmp. *ktaka-ccmp/ipvs-ingress: Initial Release*. July 2017. DOI: 10.5281/zenodo.826894. URL: https://doi.org/10.5281/zenodo.826894.

[24] Martin A. Brown. *Guide to IP Layer Network Administration with Linux*. 2007. URL: http://linux-ip.net/html/index.html (visited on 07/14/2017).

[25] Tero Marttila. “Design and Implementation of the clusterf Load Balancer for Docker Clusters”. en. Master’s Thesis, Aalto University. 2016-10-27, pp. 97+7. URL: http://urn.fi/URN:NBN:fi:aalto-201611025433.

[26] Paul B Menage. “Adding generic process containers to the linux kernel”. In: *Proceedings of the Linux Symposium*. Vol. 2. Citeseer. 2007, pp. 45–57.

[27] Dirk Merkel. “Docker: lightweight linux containers for consistent development and deployment”. In: *Linux Journal* 2014.239 (2014), p. 2.

[28] Walter Milliken, Trevor Mendez, and Dr. Craig Partridge. *Host Anycasting Service*. RFC 1546. Nov. 1993. DOI: 10.17487/RFC1546. URL: https://rfc-editor.org/rfc/rfc1546.txt.

[29] Vivian Noronha et al. “Performance Evaluation of Container Based Virtualization on Embedded Microprocessors”. In: *2018 30th International Teletraffic Congress (ITC 30)*. Vol. 1. IEEE. 2018, pp. 79–84.

[30] Osrg. *osrg/gobgp*. URL: https://github.com/osrg/gobgp/blob/master/docs/sources/zebra.md.

[31] Parveen Patel et al. “Ananta: Cloud scale load balancing”. In: *ACM SIGCOMM Computer Communication Review* 43.4 (2013), pp. 207–218.

[32] Andrew Pavlo and Matthew Aslett. “What’s really new with NewSQL?” In: *ACM Sigmod Record* 45.2 (2016), pp. 45–55.

[33] Michael Pleshakov. *NGINX and NGINX Plus Ingress Controllers for Kubernetes Load Balancing*. Dec. 2016. URL: https://www.nginx.com/blog/nginx-plus-ingress-controller-kubernetes-load-balancing/.

[34] Bowei Du Prashanth B Mike Danese. *kube-keepalived-vip*. 2016. URL: https://github.com/kubernetes/contrib/tree/master/keepalived-vip.

[35] Cristian Ruiz, Emmanuel Jeanvoine, and Lucas Nussbaum. “Performance evaluation of containers for HPC”. In: *European Conference on Parallel Processing*. Springer. 2015, pp. 813–824.

[36] Andrey Sibiryov. *GORB Go Routing and Balancing*. 2015. URL: https://github.com/kobolog/gorb.

[37] E. Chen T. Bates and R. Chandra. *BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)*. RFC 4456. RFC Editor, Apr. 2006, pp. 1–12. URL: https://www.rfc-editor.org/rfc/rfc4456.txt.

[38] Tom Herbert and Willem de Bruijn. *Scaling in the Linux Networking Stack*. URL: https://www.kernel.org/doc/Documentation/networking/scaling.txt (visited on 07/14/2017).

[39] Abhishek Verma et al. "Large-scale cluster management at Google with Borg". In: *Proceedings of the Tenth European Conference on Computer Systems - EuroSys '15* (2015), pp. 1–17. DOI: 10.1145/2741948.2741964. URL: http://dl.acm.org/citation.cfm?doid=2741948.2741964.

[40] Daniel Walton et al. *Advertisement of multiple paths in BGP*. RFC 7911. RFC Editor, July 2016, pp. 1–8. URL: https://www.rfc-editor.org/rfc/rfc7911.txt.

[41] Wensong Zhang. "Linux virtual server for scalable network services". In: *Ottawa Linux Symposium* (2000).

[42] Yang Zhao et al. "Performance of Container Networking Technologies". In: *Proceedings of the Workshop on Hot Topics in Container Networking and Networked Systems*. ACM. 2017.

A
ingress controller

```
package main

import (
        "log"
        "net/http"
        "os"
        "syscall"
        "os/exec"
        "strings"
        "text/template"
        "github.com/spf13/pflag"
        api "k8s.io/client-go/pkg/api/v1"
        nginxconfig "k8s.io/ingress/controllers/nginx/pkg/config"
        "k8s.io/ingress/core/pkg/ingress"
        "k8s.io/ingress/core/pkg/ingress/controller"
        "k8s.io/ingress/core/pkg/ingress/defaults"
)
```

```
var cmd = exec.Command("keepalived", "-nCDlf", "/etc/keepalived/ipvs.
    ↪ conf")

func main() {
        ipvs := newIPVSController()
        ic := controller.NewIngressController(ipvs)
        cmd.Stdout = os.Stdout
        cmd.Stderr = os.Stderr
        cmd.Start()
        defer func() {
                log.Printf("Shutting down ingress controller...")
                ic.Stop()
        }()
        ic.Start()
}

func newIPVSController() ingress.Controller {
        return &IPVSController{}
}

type IPVSController struct{}

func (ipvs IPVSController) SetConfig(cfgMap *api.ConfigMap) {
        log.Printf("Config map %+v", cfgMap)
}

func (ipvs IPVSController) Reload(data []byte) ([]byte, bool, error)
    ↪ {
        cmd.Process.Signal(syscall.SIGHUP)
        out, err := exec.Command("echo", string(data)).CombinedOutput
            ↪ ()
        if err != nil {
                return out, false, err
        }
        log.Printf("Issue kill to keepalived. Reloaded new config %s
            ↪ ", out)
        return out, true, err
}
```

```
func (ipvs IPVSController) OnUpdate(updatePayload ingress.
    ↪ Configuration) ([]byte, error) {
        log.Printf("Received OnUpdate notification")
        for _, b := range updatePayload.Backends {
                type ep struct{
                        Address,Port string
                }
                eps := []ep{}
                for _, e := range b.Endpoints {
                        eps = append(eps, ep{Address: e.Address, Port
                            ↪ : e.Port})
                }

                for _, a := range eps {
                log.Printf("Endpoint %v:%v added to %v:%v.", a.
                    ↪ Address, a.Port, b.Name, b.Port)
                }

                if b.Name == "upstream-default-backend" {
                        continue
                }
                cnf := []string{"/etc/keepalived/ipvs.d/" , b.Name ,
                    ↪ ".conf"}
                w, err := os.Create(strings.Join(cnf, ""))
                if err != nil {
                        return []byte("Ooops"), err
                }
                tpl := template.Must(template.ParseFiles("ipvs.conf.
                    ↪ tmpl"))
                tpl.Execute(w, eps)
                w.Close()
        }

        return []byte("hello"), nil
}

func (ipvs IPVSController) BackendDefaults() defaults.Backend {
        // Just adopt nginx's default backend config
        return nginxconfig.NewDefault().Backend
```

```
}

func (ipvs IPVSController) Name() string {
        return "IPVS Controller"
}

func (ipvs IPVSController) Check(_ *http.Request) error {
        return nil
}

func (ipvs IPVSController) Info() *ingress.BackendInfo {
        return &ingress.BackendInfo{
                Name:       "dummy",
                Release:    "0.0.0",
                Build:      "git-00000000",
                Repository: "git://foo.bar.com",
        }
}

func (ipvs IPVSController) OverrideFlags(*pflag.FlagSet) {
}

func (ipvs IPVSController) SetListers(lister ingress.StoreLister) {
}

func (ipvs IPVSController) DefaultIngressClass() string {
        return "ipvs"
}
```

B
ECMP settings

B.1 Exabgp configuration on the load balancer container.

exabgp.conf:

```
neighbor 10.0.0.109 {
 description "peer1";
 router-id 172.16.20.2;
 local-address 172.16.20.2;
 local-as 65021;
 peer-as 65021;
 hold-time 1800;
   static {
     route 10.1.1.0/24 next-hop 10.0.0.106;
   }
```

```
}
```

B.2 Gobgpd configuration on the route reflector.

gobgp.conf:

```
global:
  config:
    as: 65021
    router-id: 10.0.0.109
    local-address-list:
    - 0.0.0.0 # ipv4 only
  use-multiple-paths:
    config:
      enabled: true

peer-groups:
  - config:
      peer-group-name: k8s
      peer-as: 65021
    afi-safis:
      - config:
          afi-safi-name: ipv4-unicast

dynamic-neighbors:
  - config:
      prefix: 172.16.0.0/16
      peer-group: k8s

neighbors:
  - config:
      neighbor-address: 10.0.0.110
      peer-as: 65021
```

```
    route-reflector:
      config:
        route-reflector-client: true
        route-reflector-cluster-id: 10.0.0.109
    add-paths:
      config:
        send-max: 255
        receive: true
```

B.3 Gobgpd and zebra configurations on the router.

gobgp.conf:

```
global:
  config:
    as: 65021
    router-id: 10.0.0.110
    local-address-list:
    - 0.0.0.0

  use-multiple-paths:
    config:
      enabled: true

neighbors:
    - config:
        neighbor-address: 10.0.0.109
        peer-as: 65021
      add-paths:
        config:
          receive: true
```

```
zebra:
  config:
    enabled: true
    url: unix:/run/quagga/zserv.api
    version: 3
    redistribute-route-type-list:
      - static
```

zebra.conf:

```
hostname Router
log file /var/log/zebra.log
```

C

Analysis of the performance limit

The maximum throghput in this series of experiment is roughly, 190k[req/sec] for both ipvs an the iptables DNAT. At first, it was not clear what caused this limit. The author analyzed the kind of packets that flows during the experiment using tcpdump[22] as follows; 1) A wrk worker opens multiple connections and sends out http request to the web servers. The number of connections is determined by the command-line option, eg. 800/40 = 20 connection in the case of command-line in Table 4.1. The worker sends out 100 requests to the web server within each connection, and closes it either if all of the responses are recieved or time out occurs. 2) As in seen in Listing C.1, tcp options were mss(4 byte), sack(2 byte), ts(10 byte), nop(1 byte) and wscale(3 byte), for SYN packets. For other packets, tcp options were, nop(1 byte), nop(1 byte) and ts(10 byte). 3) The author classified the types of packes and counted the number of each type in a single connection, which is 100 http requests. Table C.1,C.2,C.3 summarize the data size of 100 request, including TCP headr, IP header, Ether header and overheads. From this analysis, it was found that per each HTTP request and response, request data with the size of 227.68[byte] and response data with the data(http content)+437.68[byte]

were being sent.

Since the node for load balancer recives and transmits both request and response packets using single network interface, each 1Gbps half duplex of full duplex must accomodate request and response data size. Therefore the theoretical maximum throughput can be expressed as;
throughput[req/sec] = band width[byte/sec]/(request + response)
= 1e9/8/(data+665.36)

Figure 4.3 shows plot of theoretical maximum throughput 1Gbps ethernet together with actual benchmark results. Since experimnetal results agrees well with theory, the author concludes that when “RPS = on”, ipvs performance limit is due to the 1Gbps bandwidth.

```
1  curl -s  http://172.16.72.2:8888/1000
2  tcpdmup(response):
3
4  03:09:27.968942 IP 172.16.72.2.8888 > 192.168.0.112.60142:
5  Flags [S.], seq 2317920646, ack 648140715, win 28960, options [mss
       1460,sackOK,TS val 2274012282 ecr 2324675546,nop,wscale 8],
       length 0
6  03:09:27.969685 IP 172.16.72.2.8888 > 192.168.0.112.60142:
7  Flags [.], ack 85, win 114, options [nop,nop,TS val 2274012282 ecr
       2324675546], length 0
8  03:09:27.969945 IP 172.16.72.2.8888 > 192.168.0.112.60142:
9  Flags [P.], seq 1:255, ack 85, win 114, options [nop,nop,TS val
       2274012282 ecr 2324675546], length 254
10 03:09:27.969948 IP 172.16.72.2.8888 > 192.168.0.112.60142:
11 Flags [P.], seq 255:1255, ack 85, win 114, options [nop,nop,TS val
       2274012282 ecr 2324675546], length 1000
12 03:09:27.970846 IP 172.16.72.2.8888 > 192.168.0.112.60142:
13 Flags [F.], seq 1255, ack 86, win 114, options [nop,nop,TS val
       2274012282 ecr 2324675547], length 0
```

Listing C.1: An example of the tcpdump output

Type of Packet	Payload [byte]	Header [byte]	Count	Total [byte]
SYN	0	98	1	98
ACK	0	90	102	9,180
Push(GET)	44	90	100	13,400
FIN+ACK	0	90	1	90
Total				22,768

Table C.1: Request data size for 100 HTTP requests in wrk measurement.

Type of Packet	Payload [byte]	Header [byte]	Count	Total [byte]
SYN+ACK	0	98	1	98
ACK	0	90	2	180
Push(GET)	254	90	100	34,400
Push(DATA)	data	90	100	100x(data+90)
FIN+ACK	0	90	1	90
Total				100x(data+90)+34,768

Table C.2: Response data size for 100 HTTP requests in wrk measurement.

Type of field	SYN	ACK, SYN+ACK, FIN+ACK, PUSH
preamble	8	8
ether header	14	14
ip header	20	20
tcp header	20 + 20(tcp options)	20 + 12(tcp options)
fcs	4	4
inter frame gap	12	12
Total [byte]	98	90

Table C.3: Header sizes of TCP/IP packet in Ethernet frame.

D VRRP

D.1 VRRP

Fig. D.1 shows an alternative redundancy setup using the VRRP protocol that was first considered by the authors, but did not turn out to be preferable. In the case of VRRP, the load balancer container needs to run in the node net namespace for the following two reasons. 1) When fail over occurs, the new master sends gratuitous Address Resolution Packets(ARP) packets to update the ARP cache of the upstream router and Forwarding Data Base(FDB) of layer 2 swicthes during the transition. Such gratuitous ARP packets should consist of the virtual IP address shared by the load balancers and the MAC address of the node where the new master load balancer is running. Programs that send out gratuitous ARP with node MAC address should be in the node net namespace. 2) Furthermore, the active load balancer sends out periodic advertisement using UDP multicast packet to inform existence of itself. The load balancer in backup state stays calm unless the VRRP advertisement stops for a specified duration of time. The UDP multicast is often unsupported in overlay network

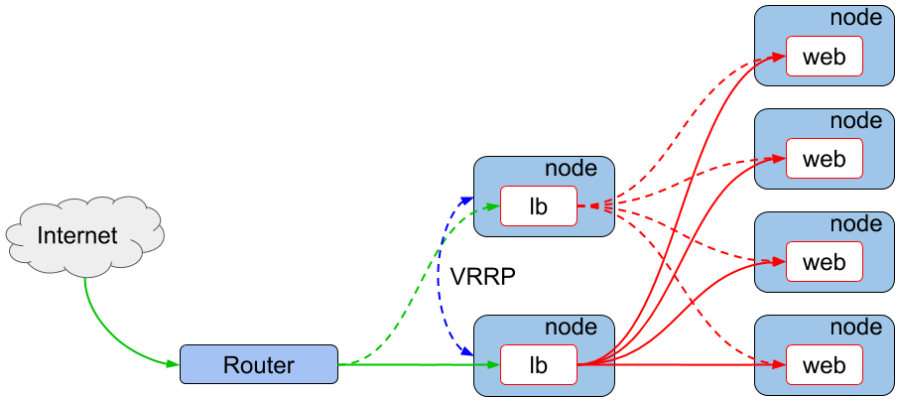


Figure D.1: An alternative redundant load balancer architecture using VRRP. The traffic from the internet is forwarded by the upstream router to a active lb node and then distributed by the lb pods to web pods using Linux kernel's ipvs. The active lb pod is selected using VRRP protocol.

used by container cluster environment, and hence the load balancer needs to be able to use the node net namespace. Running containers in the node net namespace loses the whole point of containerization, i.e., they share the node network without separation. This requires the users' additional efforts to avoid conflict in VRRP configuration for multiple services.

VRRP programs also support unicast advertisement by specifying IP addresses of peer load balancers before it starts. However, container cluster management system randomly assign IP addresses of containers when it launches them, and it is impossible to know peer IPs in advance. Therefore the unicast mode is not feasible in container cluster environment.

The other drawback compared with the ECMP case is that the redundancy of VRRP is provided in Active-Backup manner. This means that a single software load balancer limits the overall performance of the entire container cluster. Therefore we believe the ECMP redundancy is better than VRRP in our use cases.