Solution Brief

Cloud Gateway

Alibaba Increases Intelligence, Performance and Visibility of Cloud Gateway to Manage Peak Traffic Times

Alibaba's Sailfish cloud gateway project leverages Intel® Tofino™ programmable Ethernet switch ASICs and Intel® Xeon® Scalable processors for gateways that can keep up with growth of public cloud services and on-line shopper demand

How do you manage public cloud network traffic during a shopping festival that can draw millions of shopper visits creating terabits per second of data traffic to an e-commerce site? Alibaba faced this problem and used Intel® Xeon® Scalable processors along with Intel® Tofino™ programmable Ethernet switch ASICs to withstand the demands of daily and peak traffic times. Leveraging the benefits of Intel Tofino, Alibaba brought intelligence, performance, visibility and control to their cloud gateways.

INTELLIGENCE

PERFORMANCE

VISIBILITY & CONTROL

Table of Contents

Alibaba Cloud Network Architecture 1

Virtual Private Cloud is the Basis of Cloud Services 2

Cloud Gateway Enables Comunication....2

Sailfish is Designed for Scalable Performance 3

Intel Tofino Brings Packet Throughput Scalability 4

Gateway Performance 4

Conclusion.....5

Learn More.....5

End Notes.....5

Alibaba Cloud Network Architecture

The overall technical architecture of the Alibaba LuoShen Cloud Network Platform is a very typical Software-Defined Networking (SDN) and Network Functions Virtualization (NFV) architecture (see Figure 1). The foundation is the physical network infrastructure, which provides the most basic global connectivity capabilities of the network. The next layer up from the hardware is the forwarding plane. The LuoShen platform uses various types of resources as the forwarding data plane including CPU, FPGA, programmable switch ASICs, etc. On top of the forwarding data plane, one NFV platform called the CyberStar platform is built to manage different underlying forwarding resources and provide unified abstraction capabilities, such as elastic resource expansion and contraction capabilities, resource heterogeneous shielding capabilities, etc. This platform aims to improve the research and development efficiency of upper-layer service network elements. The top layer is the forwarding logic of service network elements themselves.

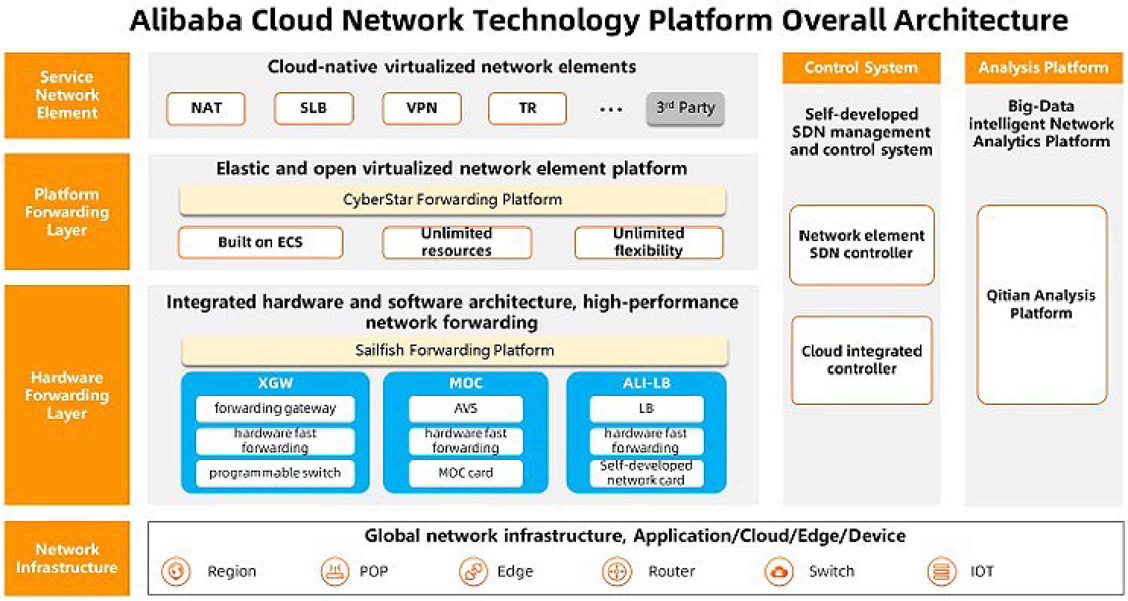
Alibaba Cloud Network Technology Platform Overall Architecture
Service Network Element
Cloud-native virtualized network elements
NAT
SLB
VPN
TR
...
3rd Party
Platform Forwarding Layer
Elastic and open virtualized network element platform
CyberStar Forwarding Platform
Built on ECS
Unlimited resources
Unlimited flexibility
Hardware Forwarding Layer
Integrated hardware and software architecture, high-performance network forwarding
Sailfish Forwarding Platform
XGW
forwarding gateway
hardware fast forwarding
programmable switch
MOC
AVS
hardware fast forwarding
MOC card
ALI-LB
LB
hardware fast forwarding
Self-developed network card
Control System
Self-developed SDN management and control system
Network element SDN controller
Cloud integrated controller
Analysis Platform
Big-Data intelligent Network Analytics Platform
Qitian Analysis Platform
Network Infrastructure
Global network infrastructure, Application/Cloud/Edge/Device
Region
POP
Edge
Router
Switch
IOT

Figure 1. Alibaba LuoShen Cloud Network Platform architecture.

Virtual Private Cloud is the Basis of Cloud Services

The foundation of the economic and service agility of a cloud service rests on the ability to provide dedicated compute and storage resources on a shared infrastructure using virtual private clouds (VPCs). These VPCs can be small or large, can operate across multiple physical servers and provide resource isolation using overlay networking protocols such as VXLAN. Multiple virtual machines (VMs) within the same VPC can communicate with each other while VMs belonging to different VPCs are isolated and transparent to each other.

Cloud Gateway Enables Communication

Sometimes, a VM in one VPC needs to communicate with another VM in a different VPC within the same region or in a remote location. In other cases, a VM in one VPC wants to access the public Internet or the resources inside the enterprise's in-house data centers (IDCs). The cloud gateway addresses these inter-VM and cross-region communication requirements and constructs tunnels if the communication needs to cross a VPC boundary.

Facilitating all of these connections is the cloud gateway (see Figure 2), which is designed to make all of these connections across potentially millions of VPCs. The performance of the cloud gateway impacts the scalability and throughput of the cloud service provider.

The main function of a cloud gateway is providing packet forwarding for fast and reliable connectivity to globally distributed cloud resources in a multi-tenant environment. Gateways are typically deployed in clusters to allow for scalability. But the growth of public cloud data center traffic is so rapid that cluster-based scalability is expensive and a challenge to manage. Something needed to be done to make each of the servers in the cluster more performant while also maintaining the large forwarding tables that must be supported for a multi-tenant gateway.

This challenge was significant for the Alibaba Group, which is one of the world's largest public cloud service providers. The company's Alibaba Cloud offers a complete suite of cloud services, including elastic computing, database, storage, network virtualization, large-scale computing, security, management and applications, big data analytics and machine learning platform services.

In 2019, Alibaba put its entire e-commerce business onto its public cloud[1] which contributed to an increase of dozens of terabits per second of traffic for its cloud gateway. This was the impetus for developing a new scale out multi-tenant cloud gateway design called Sailfish. Working with Intel technology, specifically Intel Tofino programmable Ethernet switch ASICs and Intel Xeon Scalable processors, the company developed Sailfish the first P4-programmable, intelligent gateway. This solution provides the capacity to scale gracefully in response to explosive traffic growth coming from its growing e-commerce business while preventing the gateway from becoming the system choke point.

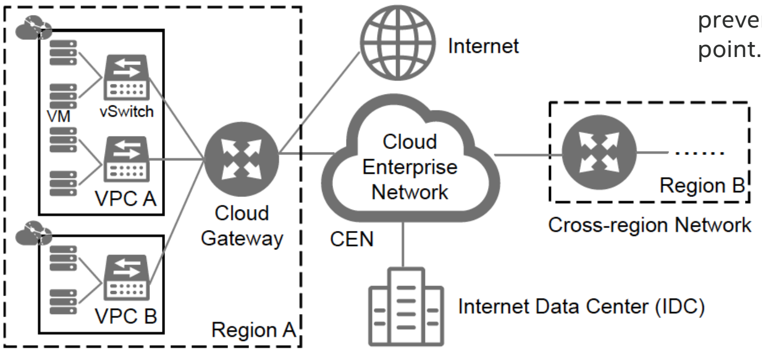

Figure 2. Cloud gateways act as the communications hub for public cloud data centers facilitating communications between virtual private clouds, the internet and local or remote virtual resources.

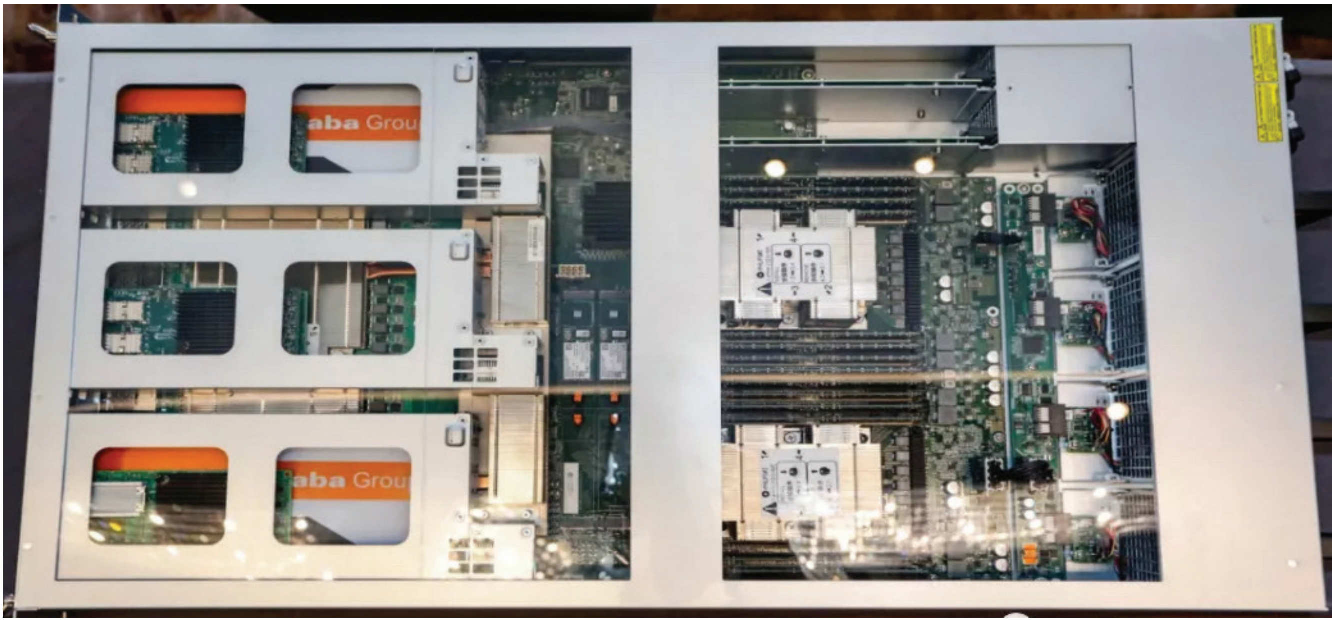

Figure 3. Top view of Alibaba Sailfish cloud gateway.

Sailfish is Designed for Scalable Performance

In designing the Sailfish gateway (see Figure 3), Alibaba identified four highlights that would distinguish the functionality and performance of the gateway:

Highlight 1: *Leading technical architecture:* Sailfish is the industry's first P4-programmable cloud gateway solution in which a single cluster carries dozens of Tbps traffic[2]. It is based on Intel Tofino programmable Ethernet switch ASICs. The ASICs allow the gateway to use an integrated software and hardware design, and uses the programmable ASICs to improve data plane performance for fast packet forwarding performance and programmability to add new functionality to meet rapid business iteration challenges.

Highlight 2: *Leading performance indicators:* A single Sailfish gateway supports bandwidth up to 3.2Tbps, which can be horizontally expanded by adding another gateway to provide throughput of up to 6.4Tbps. The switch reduces packet latency significantly, while also increasing packets-per-second processing capacity significantly.

Highlight 3: *Leading in large-scale applications:* The Sailfish gateway has been deployed in Alibaba Cloud's data centers all over the world for more than two years, achieving large-scale deployment, supporting more than 1 million users for services including VPCs, elastic IP addresses, shared bandwidth, high-speed channels or cloud enterprise network services.

Highlight 4: *Leading cost advantage:* The use of Sailfish in the Alibaba network has reduced both CapEx and OpEx by releasing significant number of servers for revenue-generating workloads needed, which drives reductions in development and operating costs.

As seen in Figure 4, the Sailfish cloud gateway has evolved from separate appliances providing discrete connectivity services, such as connectivity between VPCs, between VPCs and the Internet, between VPCs and internet data centers (IDCs) and more. In this way, the development, testing and online deployment of new services would not impact other, already deployed, services.

Alibaba then integrated these separate features into an eXtendable GateWay (XGW) based on Intel Xeon Scalable processors that can be scaled up by using higher performance or higher core count CPUs.

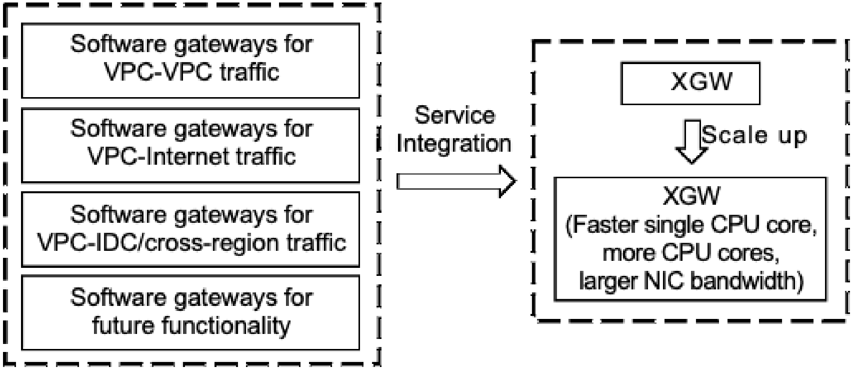
Software gateways for VPC-VPC traffic
Software gateways for VPC-Internet traffic
Software gateways for VPC-IDC/cross-region traffic
Software gateways for future functionality
Service Integration
XGW
Scale up
XGW (Faster single CPU core, more CPU cores, larger NIC bandwidth)

Figure 4. The Sailfish design evolved from multiple discrete service gateways into a multi-service eXtendable GateWay (XGW).

Intel Technology Powers Performance

Sailfish utilizes both the Intel® Tofino™ programmable Ethernet switch ASIC and the 2nd Generation Intel® Xeon® Scalable processor for performance.

Intel Tofino delivers up to 6.4 Tb/s throughput with support for up to 100GbE port speed for hyperscale, cloud, and service provider environments. Intel Tofino is P4-programmable using its Protocol-Independent Switch Architecture (PISA). This architecture allows for precise control over packets, customized support of network protocols, and the ability to run networking applications, such as routing, firewall, load balancing and more, right in the fabric.

Intel Xeon Scalable processors benefit from decades of innovation for the most common workload requirements, supported by close partnerships and deep integration with the world's software leaders and solution providers. Intel Xeon Scalable processors offer a balanced architecture, optimized for many workload types and performance levels, all with the consistent, open, Intel architecture that delivers built-in AI acceleration and advanced security capabilities.

Intel Tofino Brings Packet Throughput Scalability

In addition to the forwarding performance, the cloud gateway needs to support a large number of stateless and stateful forwarding tables for diverse cloud services. Due to the multi-tenancy in the public cloud, the forwarding table entry size is large because it contains both the VPC identifier and the destination address. Also, the huge number of VPCs and VMs in the public cloud impacts the table entry number making it huge. In the Alibaba Cloud, a single cloud region can host millions of VPCs and millions of VMs.

To meet these needs and be scalable, the Sailfish gateway combines a hardware gateway design using the Intel Tofino programmable Ethernet switch ASIC and a software gateway based on the Intel Xeon Scalable processor. Sailfish designers used these design elements to maximize the performance scalability:

Hardware and software co-design: Sailfish consists of both an Intel Tofino-based hardware gateway and Intel architecture-based software gateway. The hardware gateway leverages the line rate packet processing performance of the Intel Tofino. That is paired with the software gateway's huge memory space and full programmability. The Sailfish design places the hardware gateway first in the data path - in front of the software gateway. The hardware gateway stores only a few key routing tables that are frequently used by the majority of traffic. The software gateway holds the remaining volatile tables that are hit by a small portion of traffic and the huge stateful tables that cannot be compressed into the hardware gateway.

Table splitting among hardware gateway clusters: Sailfish is able to split tables horizontally among hardware gateway clusters to reduce the number of table entries stored in each hardware gateway. The horizontal splitting results in each hardware gateway storing only a portion of entries from all of the forwarding tables. Using this design, multiple hardware gateway clusters can cover all the entries in a region, while a single cluster is only responsible for entries of some tenants. Within a cluster, multiple hardware gateways maintain the same table entries, share the traffic load and backup each other.

Single-node table compression: The Sailfish design utilizes table compression on each hardware gateway in order to fit more table entries into a single node. These optimizations include pipeline folding, table splitting among pipelines, table mapping across pipelines, memory resource pooling, TCAM conservation and table entry compression. These memory optimizations were possible due to the programmability of Intel Tofino which enabled adapting the device to Alibaba's exact use case. The single-node table compression increases the number of entries carried in one cluster, thus reducing the number of necessary clusters which reduces CapEx and OpEx.

Gateway Performance

Alibaba's tests of the gateway yielded significant benefits in three areas:

Intelligence: Several configurations were used to attain the best observed performance. In one configuration, traffic load was distributed between data pipelines to halve the number of table entries in each pipeline. The traffic load was split horizontally and could use either hashing or service characteristics with historical data mining to divide the traffic. Traffic was also shared between the hardware and software gateways, which resulted in the software gateway carrying a very small portion of the cloud traffic in order to leverage the Intel Tofino performance. The traffic sharing left a few gigabits per second of traffic to be handled by the software gateway which could easily be handled by the CPU core.

Performance[3]: When measured against a comparable software gateway, Sailfish reduced latency by 95% to 2µs, improved throughput by more than 20 times to 3.2Tbps and with a 71 times increase in packets per second to 1.8Gpps using packet length of < 256 bytes.

Visibility & Control[4]: Packet drops rates were measured at the very low rate of from 10^{-11} to 10^{-10} - six orders of magnitude lower than that of software gateway alone. This demonstrates the stable performance of Sailfish in a production environment and is a result of the large packet-throughput safety margin provided by the Intel Tofino's high-capacity data plane.

Conclusion

The growth of traffic served by the Alibaba Cloud public cloud service and the ever present possibility of a data burst has made it necessary to rethink the design of the cloud gateway. The innovative Sailfish is a hardware gateway based on Intel Tofino programmable Ethernet switch ASICs to maximize throughput, that is complemented by a software gateway based on Intel Xeon Scalable processors for huge memory availability. Together, these gateways provide scalable throughput performance.

Sailfish has been adopted and utilized in the Alibaba Cloud where the company says the performance of the gateway has been worth the effort to develop it. But could Sailfish be ready for use by other cloud companies? Alibaba has stated[5] that other large public cloud services can benefit from the performance and scalability of the hardware and software gateway design of Sailfish. Small or medium-sized public cloud services, however, may be able to utilize only a software-based gateway for now, but that won't always be the case as the volume of data continues to grow.

INTELLIGENCE	PERFORMANCE	VISIBILITY & CONTROL
Horizontal traffic split between hardware gateway plus software gateway combination frees up CPU for other more meaningful tasks.	>20x throughput increase, up to 95% latency reduction vs. software gateway only[6]	6 orders of magnitude packet drop rate reduction vs. software gateway only[7]

Learn More

Alibaba Cloud Services

Intel Intelligent Fabric Processors

Intel Tofino Programmable Ethernet Switch ASIC

Intel Xeon Scalable Processors

End Notes

1. Alibaba Group's comprehensive cloud practice and thinking (06/2020) https://developer.aliyun.com/article/765369
2. "Sailfish: Accelerating Cloud-Scale Multi-Tenant Multi-Service Gateways with Programmable Switches," SIGCOMM '21, https://dl.acm.org/doi/abs/10.1145/3452296.3472889
3. Performance claims made by Alibaba based on its internal testing in August 2021 (for further information, please see https://dl.acm.org/doi/abs/10.1145/3452296.3472889.) For more complete information about Intel performance and benchmark results, please visit www.intel.com/benchmarks.
4. Ibid
5. See end note 2.
6. See end note 3.
7. See end note 3.

Notices & Disclaimers
Intel technologies may require enabled hardware, software or service activation.
No product or component can be absolutely secure.
Your costs and results may vary.
Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.