WHITEPAPER

AMAZON ELB:

Your Master Key to a Secure, Cost-Efficient and Scalable Cloud

www.cloudcheckr.com

Cloud Checkr

TABLE OF CONTENTS

Overview 3

What Is ELB? 3

- How ELB Works 4
- Classic Load Balancer 5
- Application Load Balancer 5
- Network Load Balancer 5

How ELB Helps Maintain Optimal Instance Usage 8

- Performance 8
- Availability 8
- Security 8
- Hybrid Clouds 8
- Containers 9
- Scaling 9

ELB as Part of a Reserved Instance Strategy 10

Why You Need to Monitor ELB 11

- Time to Start Tracking 12

> "Load balancing, auto scaling, and cloud monitoring work together to help you to build highly scalable and highly available applications. Amazon CloudWatch monitors your Amazon EC2 capacity, Auto Scaling dynamically scales it based on demand, and Elastic Load Balancing distributes load across multiple instances in one or more Availability Zones."
>
> **– JEFF BARR** | Chief Evangelist For AWS

OVERVIEW

AWS Elastic Load Balancing (ELB) is Amazon's proprietary load balancing service that harnesses the power of distributed cloud-based workloads.

It boasts a rich set of features and configurations, with three classes of load balancer to support a wide variety of routing requirements.

But it also plays a pivotal role in maintaining performance, availability, security and a cost-efficient cloud environment.

In this paper, we take a deep dive into the workings of ELB, how it can help optimize your cloud resource utilization and reduce costs as part of a Reserved Instance strategy. Then finally we look at why monitoring your ELB deployments is so important.

We've got a lot to cover. So let's get to it.

WHAT IS ELB?

ELB distributes traffic across a group of target EC2 instances, ensuring your application isn't dependent on a single server. ELB provides a single point of access to your application, sharing the workload between resources. This helps your systems to serve client requests more quickly.

You can also use ELB to improve fault tolerance by hosting your application on instances in more than one Availability Zone.

How ELB Works

Your load-balanced application environment consists of your ELB and a group of registered instances in one or more enabled Availability Zones. You can manually add or remove instances as the load on your application increases or decreases. Alternatively, your instances can be members of an Auto Scaling group attached to your load balancer.

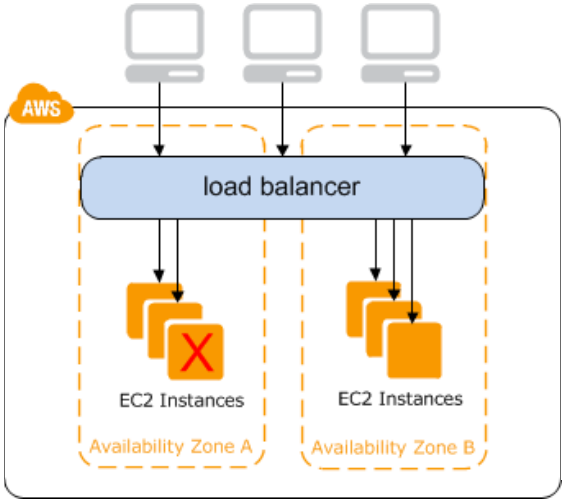

FIGURE 1:

An ELB enabled in two Availability Zones.

(Source: AWS)

ELB accepts incoming traffic using one or more listeners, which continually check for client connection requests and forwards them to your registered targets.

The service is automatically configured to perform health checks. These monitor the availability of your instances so that requests are only sent to healthy targets, hiding network and instance failures from the end user.

By default, ELB distributes load evenly across your enabled Availability Zones. However, it also comes with the option of cross-zone load balancing, which distributes traffic evenly between instances, regardless of Availability Zone. With either configuration, you should assign resources equally across Availability Zones for better fault tolerance.

IPv4 and IPv6 Support

ELB supports IPv4 addresses universally whereas IPv6 support is generally restricted to EC2-Classic environments only. However, AWS now offers IPv6 across all its ELB services in US East (N. Virginia) and EU (Ireland), with more Regions expected to follow suit in the near future.

The following are Amazon's three classes of ELB service:

Classic Load Balancer

Amazon's basic-level load balancer and the cloud vendor's original load balancing product. It was renamed Classic Load Balancer (CLB) when the company launched Application Load Balancer in August 2016.

CLB routes traffic between end users and your backend servers based on IP address and TCP port, operating at Layer 4 of the OSI model.

With widespread use amongst established AWS customers and legacy support for EC2-Classic environments, CLB is still ostensibly Amazon's standard load balancing offering.

Application Load Balancer

Application Load Balancer (ALB) is a more sophisticated Layer 7 technology that's capable of more intelligent load distribution.

With ALB, you can create rules based on application-level content rather than simply IP address and port number. This gives you the scope to create different target clusters for different types of traffic, where each cluster can have its own Auto Scaling group tailored more precisely to the volume and nature of requests.

ALB includes support for both path-based and host-based routing. So you could, for example, process requests to URLs such as example.com/mobile or mobile.example.com using different resources from those that handle the rest of your website.

Network Load Balancer

Network Load Balancer (NLB) is a Layer 4 TCP component providing an extreme level of load balancing performance for large-scale latency-sensitive applications. What's more, it is designed to automatically handle sudden or unexpected surges in traffic without a pre-warming period.

You can also map static IP and Elastic IP addresses to your NLB, eliminating potential issues when using user-friendly aliases to your load balancer DNS name.

And, unlike the other ELB options, NLB is also network transparent, which means your servers are able to see the user's actual IP address and port without the need for complex technical workarounds.

Feature	Application Load Balancer	Network Load Balancer	Classic Load Balancer
Protocols	HTTP, HTTPS	TCP	TCP, SSL, HTTP, HTTPS
Platforms	VPC	VPC	EC2-Classic, VPC
Health checks	✔	✔	✔
CloudWatch metrics	✔	✔	✔
Logging	✔	✔	✔
Connection draining (deregistration delay)	✔	✔	✔
Load Balancing to multiple ports on the same instance	✔	✔	
IP addresses as targets	✔	✔	
Path-Based Routing	✔		
Host-Based Routing	✔		
Configurable idle connection timeout	✔		✔
Cross-zone load balancing	✔		✔
SSL offloading	✔		✔
Sticky sessions	✔		✔

Feature	Application Load Balancer	Network Load Balancer	Classic Load Balancer
Back-end server encryption	✔		✔
Static IP		✔	
Elastic IP address		✔	
Preserve Source IP address		✔	

ELB Pricing

CLB is charged at an hourly rate PLUS a cost per GB of data processed by your load balancer.

Rates vary from Region to Region with the lowest prices currently in US East (N. Virginia), US East (Ohio), US West (Oregon) and Asia Pacific (Seoul) at $0.025 per CLB-hour + $0.008 per GB of data.

The most expensive Region is South America (São Paulo) at $0.034 per CLB-hour + $0.011 per GB of data.

ALB and NLB use a more complicated pricing structure. It comprises a fixed hourly charge PLUS a variable hourly charge based on a complex traffic processing metric called a Load Balancer Capacity Unit (LCU). This is calculated from a combination of dimensions, such as the number of new connections per second and bandwidth used in Mbps.

Prices are also lowest in US East (N. Virginia), US East (Ohio), US West (Oregon) and Asia Pacific (Seoul) and most expensive in South America (São Paulo).

HOW ELB HELPS MAINTAIN OPTIMAL INSTANCE USAGE

Performance

ELB addresses the problem of delivering application performance at scale. Whereas single-machine deployments are limited by the maximum instance size available, there's virtually no limit to how many instances you can host in an ELB target group.

CLB and ALB also support a configurable idle connection timeout feature, which terminates a connection when no data is sent between client and server over your specified time threshold. As a result, you can free up target instances based more closely on the time it takes your servers to fulfill requests, helping you to improve application performance.

Availability

ELB's health checks ensure your workload is spread across only healthy targets. What's more, you can configure an attached Auto Scaling group to replace unhealthy instances in response to failed ELB health checks, ensuring you maintain the optimal size of your target cluster.

Security

ELB supports a variety of features and settings for securing your application stack. For example, you can configure CLB to use Amazon's default security policy or choose from a set of other predefined policies. These support SSL 3.0, TLS 1.0, TLS 1.1 and TLS 1.2 protocols, a wide range of ciphers and a server order preference option, which allows your load balancer to determine which cipher to use in SSL connection negotiations.

Alternatively, you can define your own custom security policy with the ciphers and protocols you need. This can help you meet compliance standards such as PCI and HIPAA.

CLB and ALB also offer centralized management of SSL certificates. This simplifies security management by allowing you to upload your certificates directly to your load balancer. It also offloads decryption from your instances, reducing the CPU workload on your application servers.

Hybrid Clouds

ELB can distribute traffic across AWS and on-premise infrastructure using the same load balancer. You can also assign different target groups to your different environments and use content-based routing to send traffic to either your public cloud or on-premise data center depending on the type of request.

This can help you make optimum use of your hybrid cloud by assigning workloads to the resources that suit them best.

Containers

ELB is also integrated into Amazon's EC2 Container Service (ECS).

The service leverages the content-based routing capabilities of Application Load Balancer to map different container tasks to different ports.

As a result, ELB can play a role in helping you to reduce your infrastructure footprint. This is because multiple container deployments can be hosted on a single instance, making them a lightweight and cost-efficient alternative to virtual machines.

Learn More about Containers

To find out more about containers and how they can help reduce your cloud costs, check out our recent post:

How Containers Can Optimize VM Workload Density and Reduce Your Cloud Costs

"Microservices and container-based applications have changed what customers require from AWS load balancing. Monitoring takes on greater importance in microservices workloads, for example, so IT teams can respond quickly when an app component crashes."

– SITE EDITOR | SearchAWS at TechTarget

Scaling

When you attach an Auto Scaling group to ELB, it will automatically spin up new instances as load increases and terminate them when demand goes down.

This can help you maintain cost-efficient cloud infrastructure—through fine-grained horizontal scaling of resources and a healthy balance of compute performance and instance utilization. ELB automatically responds to the changing number of instances, rerouting requests accordingly.

ELB AS PART OF A RESERVED INSTANCE STRATEGY

You can also leverage ELB and Auto Scaling to take advantage of discounted alternatives to standard on-demand pricing, such as Reserved Instances, helping you to lower your monthly cloud bills even further.

Here's how:

When you set up an Auto Scaling group, you can specify a scaling policy to dynamically adjust your desired capacity as your workload increases or decreases.

At the same time, you can set a minimum number of instances below which the size of your group should never fall. Similarly, you can designate a maximum size to prevent scaling above a specified number of instances.

The following diagram shows an Auto Scaling group with a minimum size of two, maximum size of five and current desired capacity of three instances.

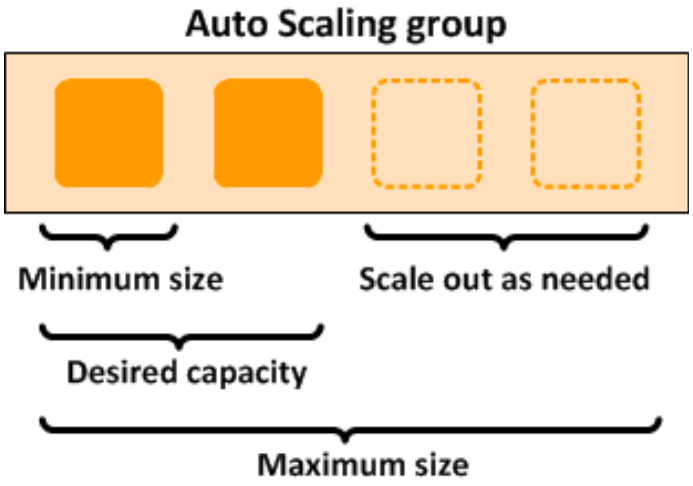

FIGURE 2:

A simple Auto Scaling Group (Source: AWS)

In this example, your Auto Scaling group will always run at least two instances at any one time. So it makes sense to purchase two Reserved Instances that match the specifications of your target instances. This is because Reserved Instances are perfectly suited to steady, predictable workloads, which consume your billing credits all or most of the time, maximizing their savings potential.

What Are Reserved Instances?

AWS Reserved Instances are a financial commitment to reserved discounted capacity, offering potential savings of up to 75% compared with on-demand pricing. They entitle you to a credit against any running instance that matches the instance type you specify at the time of purchase.

AWS also offers a more flexible offering, the Convertible Reserved Instance, which works on the same principle, but comes with the option to change the associated instance family, operating system (OS) or tenancy.

Reserved Instances are available over a fixed term of either one or three years. The level of savings you can expect to achieve depend on a variety of factors, such as the instance type you choose, the length of term and the amount you pay upfront.

However, in order to achieve the maximum possible discount, you should utilize your Reserved Instances as much as possible—as any unused credit goes to waste.

If you also use Amazon's Scheduled Scaling feature, where you scale your fleet of instances up and down at fixed times, you should consider using Scheduled Reserved Instances to meet the cost of running additional capacity each time your target pool increases. This setup is ideal for applications with regular workload patterns, such as peak activity during working hours and quiet periods during evenings and weekends.

Finally, for less predictable workloads, you could also consider developing a scaling strategy that combines on-demand instances with Spot Instances, which allow you bid on spare EC2 capacity. In this case, you'd set a bid price lower than the on-demand price and configure a scaling policy that would always attempt to scale up using Spot Instances first.

WHY YOU NEED TO MONITOR ELB

ELB is a critical component of your cloud infrastructure, providing the gateway between end users and your backend applications. That's why it's so important to monitor your load balancers to ensure your infrastructure delivers the performance you expect.

Visibility into your cloud is key. So the best place to start is to know what you have—what your load balancers are being used for, what target instances they're using and whether these are healthy and distributed evenly across Availability Zones.

You should also keep a close check on the utilization of your ELBs. To keep costs down and reduce your attack surface, you consider reducing the minimum size of underutilized Auto Scaling groups and remove unused ELBs and their target instances altogether.

On the other hand, if requests are regularly being queued or your load balancer often becomes overloaded and cannot fulfil requests then you'll need to increase the maximum size of your Auto Scaling group or use larger instances.

Similarly, high timeout metrics could suggest your instances are being overutilized, requiring changes to your Auto Scaling configuration. Alternatively, they could point to an issue with the ELB itself, which may not be allowing instances sufficient time to process requests.

And, finally, you should track whether requests are generally going up or down, as this can help you reconfigure ELBs before utilization problems arise.

Time to Start Tracking

The sooner you start tracking your ELBs, the sooner you can eliminate configuration issues that reduce application performance and increase your cloud running costs.

However, the problem is that setting up your own ELB monitoring system can be a complex and time-consuming process. So a better option is to use a solution that can already do the job for you.

Not only can third-party monitoring tools do just that, but they can also provide a range of other value-added features for managing your cloud, such as automated resource resizing, cost allocation reports and Reserved Instance recommendations.

And, most importantly, you can have all this up and running in no time at all.

Need CloudCheckr for your organization? Learn more at **www.cloudcheckr.com**.

Cloud Checkr

342 N GOODMAN ST,
ROCHESTER, NY 14607

1-833-CLDCHCK

www.cloudcheckr.com