Server Load Balancer

プロダクト紹介

プロダクト紹介

プロダクトの概要

Alibaba Cloud Server Load Balancerは、複数のバックエンドサーバーに転送ルールとスケジューリングアルゴリズムに基づいてトラフィックを配信するトラフィック分散制御サービスです。

Server Load Balancerサービスは、仮想IPアドレスを設定することによって、同じ地域にあるECSサーバーリソースを高性能で可用性の高いアプリケーションサービスプールに仮想化します。クライアント要求は、定義されたリスニングルールに従ってクラウドサーバープールに配信されます。

Server Load Balancerサービスは、クラウドサーバープール内のECSサーバーの正常性状態をチェックし、異常状態のECSサーバーを自動的に分離して、SPOF（Single Point of Failure）問題を解決し、アプリケーションの全体的なサービス機能を向上させます。標準ロードバランシング機能に加えて、TCPおよびHTTPリスニングはDDoS攻撃を防御することができ、アプリケーションサーバーの保護機能が強化されます。

サーバーロードバランサのコンポーネント

Server Load Balancerサービスは、Server Load Balancerインスタンス、リスナー、およびバックエンドサーバーの3つの部分で構成されています。

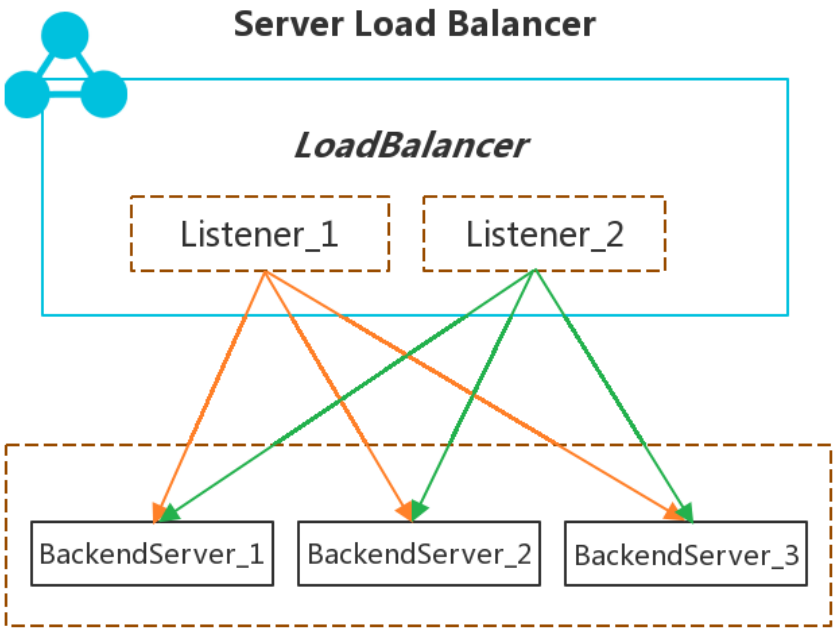
ロードバランサ：Server Load Balancerサービスを使用する場合は、Server Load Balancerサービスを購入してServer Load Balancerインスタンスを作成する必要があります。複数のリスナーとバックエンドサーバーをServer Load Balancerインスタンスに追加できます。

リスナー ：Server Load Balancerサービスを使用する前に、少なくとも1つのリスナーをServer Load Balancerインスタンスに追加する必要があります。このリスナーは、クライアント要求がバックエンドサーバーに転送される方法を定義します。

バックエンドサーバー：フロントエンド要求を受け取るECSインスタンスのグループ。 ECSサーバをサーバプールに個別に追加するか、VServerグループを介してバックエンドサーバを一括して管理することができます。

次の図に示すように、Server Load Balancerインスタンスがクライアント要求を受信すると、リスナーは、

構成されたリスニングルールに従って、要求を対応するバックエンドECSインスタンスに転送します。



利点

高可用性

単一障害点（SPOF）なしで完全冗長モードで動作するように設計されたServer Load Balancerは、DNSと併用するとローカルおよびクロス地域の災害許容度をサポートし、最大99.95%のサービス可用性を提供します。

拡張性

Server Load Balancerは、トラフィックの変動時に外部サービスを中断することなく、アプリケーションの負荷に基づいて柔軟にサービスを拡張できます。

低価格

Server Load Balancerは、従来のハードウェア負荷分散システムに比べて60%もコスト効率が良いです。 O&Mコストを発生させることなくプライベートネットワークインスタンスに無料でアクセスできるため、高価な負荷分散装置を購入する必要が完全になくなります。

セキュリティ

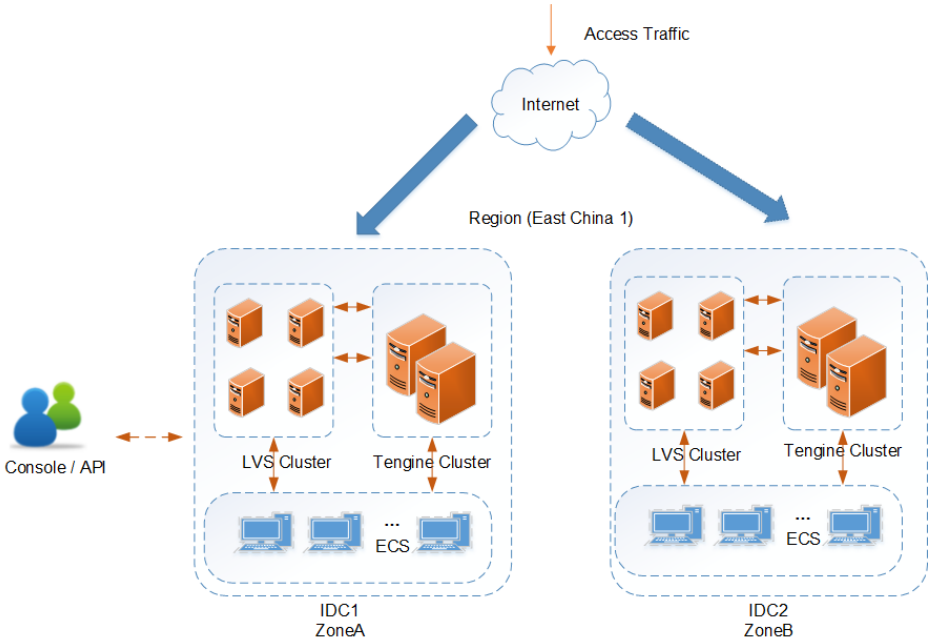
Alibaba Cloud Securityと組み合わせると、Server Load Balancerは、HTTPフラッド攻撃やSYNフラッド攻撃など、最大5GbpsのDDoS攻撃を防御できます。

アーキテクチャ

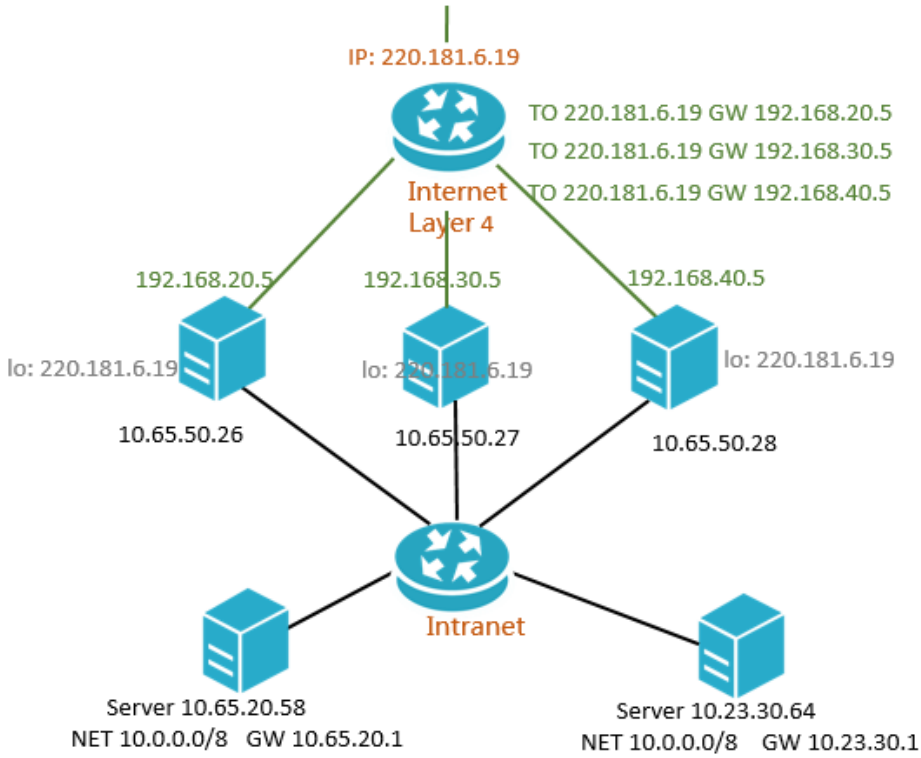
Alibaba Cloud は、レイヤ 4 (TCP プロトコルと UDP プロトコル) とレイヤ7 (HTTP プロトコルと HTTPS プロトコル) のロードバランシングサービスを提供します。クラスタに展開すると、Server Load Balancer はセッションを同期させて、単一障害点 (SPOF) から ECS インスタンスを保護できます。これにより、冗長性が向上し、サービスの安定性が保証されます。

レイヤー 4 Server Load Balancer は、オープンソースソフトウェアの LVS (Linux Virtual Server)と keepalived を基盤にしています。

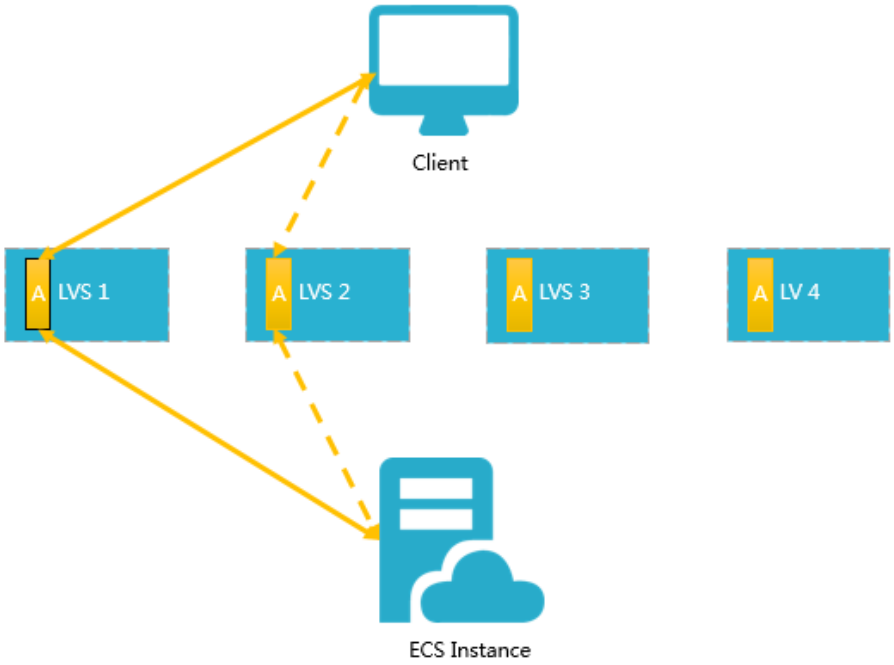
レイヤー 7 Server Load Balancer は、Taobao が開始した Web サーバープロジェクトである Tengine を利用します。Nginx を基盤とする Tengine には、トラフィックの多い Web サイトのニーズに合わせた幅広い高度な機能と特性が追加されています。



以下の図に示しているように、個々のリージョンにあるレイヤー４ロードバランサーは実際には複数のLVSから構成されるクラスターの中で実行されます。 クラスターデプロイモデルにより、不測の事態においても負荷分散サービスによって可用性、安定性や拡張性を保証いたします。



さらに、LVSクラスターの中のLVSはマルチキャストパケットを送信して他のLVSとセッションを同期しています。次の例ではセッションAはLVS1にセットされています。セッションAは3パケット送信された後に他のLVSと同期されます。正常時には、セッションリクエストは実線で示されているようにLVS1に送信されます。もし異常時またはメンテナンス時には点線で示されているようにセッションリクエストは他の正常に起動しているLVSに送信されます。このように業務アプリケーションに影響を与えることなくホットアップグレード、機器故障時のメンテナンスやクラスターメンテナンスを行うことができます。



注意: コネクションが確立していないとき (three-way handshakeが完了していないとき) または、コネクションは確立しているがセッション同期が機能していない場合にはホットアップグレードはコネクションが中断されないことを保証しません。その場合はユーザー自身がコネクションを再接続する必要があります。

機能

Alibaba Cloud Server Load Balancerには、以下の機能があります。

サポートされるプロトコル
現在、Server Load Balancer はレイヤー 4 プロトコル (TCP および UDP) とレイヤー 7 プロトコル (HTTP および HTTPS) の両方で使用できます。

ヘルスチェック
バックエンド ECS インスタンスでのヘルスチェックにより、Server Load Balancer は異常な ECS インスタンスを自動的にブロックし、再び正常に機能するようになったら復元できます。

セッション維持
Server Load Balancer のセッション維持機能は、セッションのライフサイクルの間、クライアントの要求を同じバックエンド ECS インスタンスに転送します。

スケジューリングアルゴリズム

Server Load Balancerは、次のスケジューリングアルゴリズムをサポートしています。

ラウンドロビン

リクエストはバックエンドのECSサーバーに順次配信されます。

加重ラウンドロビン (WRR)

バックエンドサーバーごとに重みを設定できます。より高い重みを持つサーバは、より低い重みを持つサーバよりも多くのリクエストを受け取ります。

重み付け最小接続 (WLC)

バックエンドECSサーバーに設定された重みに加えて、サーバーへの接続数も考慮されます。より高い重み付け値を持つサーバーは、一度に大きな割合のライブ接続を受け取ります。重みが同じである場合、システムは確立された接続の数が最も少ないサーバーにネッ

トワーク接続を指示します。

ドメイン名/URL ベースの転送
レイヤー 7 (HTTP および HTTPS) プロトコルの場合、Server Load Balancer はユーザーのドメイン名または URL に基づいて異なる仮想サーバーグループにトラフィックを転送します。

マルチゾーン
指定されたゾーンの Server Load Balancer インスタンスの作成をサポートします。マルチゾーンリージョンにはマスターゾーンとスレーブゾーンをデプロイできます。マスターゾーンが障害のときは、スレーブゾーンが障害のあるゾーンからサービスを自動的に引き継ぎます。

アクセス制御
ホワイトリスト機能をサポートします。ホワイトリストは、Server Load Balancer モニタリングにアクセスできる IP アドレスを決定するために使用されるアクセス制御方法です。ユーザーアプリケーションが特定の IP アドレスだけを許可する場合に適用されます。

セキュリティ
アプリケーションファイアウォールと CC 保護をサポートします。クラスターの WAF モジュールを使用することで、CNAME を変更することなく WAF 保護を有効にできます。Anti-DDos サービスと組み合わせて使用することで、システムは 5 Gbps 以下の DDoS 攻撃を防ぐことができます。

証明書管理
これは HTTPS プロトコルでの統合証明書管理サービスであり、バックエンド ECS インスタンスに証明書をアップロードする必要がなくなります。復号化は Server Load Balancer で実行されるので、バックエンド ECS インスタンスの CPU オーバーヘッドが減ります。

帯域幅制御
リスナーの結果に基づいて、各サービスに帯域幅ピークを設定できます。

サポートされているインスタンス/ネットワークのタイプ
Server Load Balancer はインターネットネットワークまたはイントラネットネットワークをサポートします。

モニタリング
Server Load Balancer の実行ステータスがリアルタイムでわかるように、豊富なモニタリングデータが提供されます。

管理方法
Server Load Balancer コンソール、API、SDK など、複数の管理方法が提供されます。

適用シナリオ

Server Load Balancer は、次のシナリオに適用されます。

シナリオ 1: トラフィック量の多いアプリケーション用のトラフィックの負荷分散

アプリケーションのトラフィックが高い場合は、Server Load Balancer を使用してトラフィックを複数の ECS インスタンスに分散できます。さらに、セッション維持機能を使用し、クライアントからのセッション要求を同じバックエンド ECS インスタンスに転送して、アクセス効率を高めることができます。

シナリオ 2: アプリケーションのためのサービス機能の拡張

ビジネスニーズに応じて、いつでもバックエンド ECS インスタンスを追加または削除してサービス機能を拡張できます。これは Web および App アプリケーションに適用されます。

シナリオ 3: SPOF (単一障害点) の排除

Server Load Balancer サービスは、ヘルスチェック機能により、異常な ECS インスタンスを自動的にブロックし、正常な ECS インスタンスに要求を分散して単一障害点を排除します。

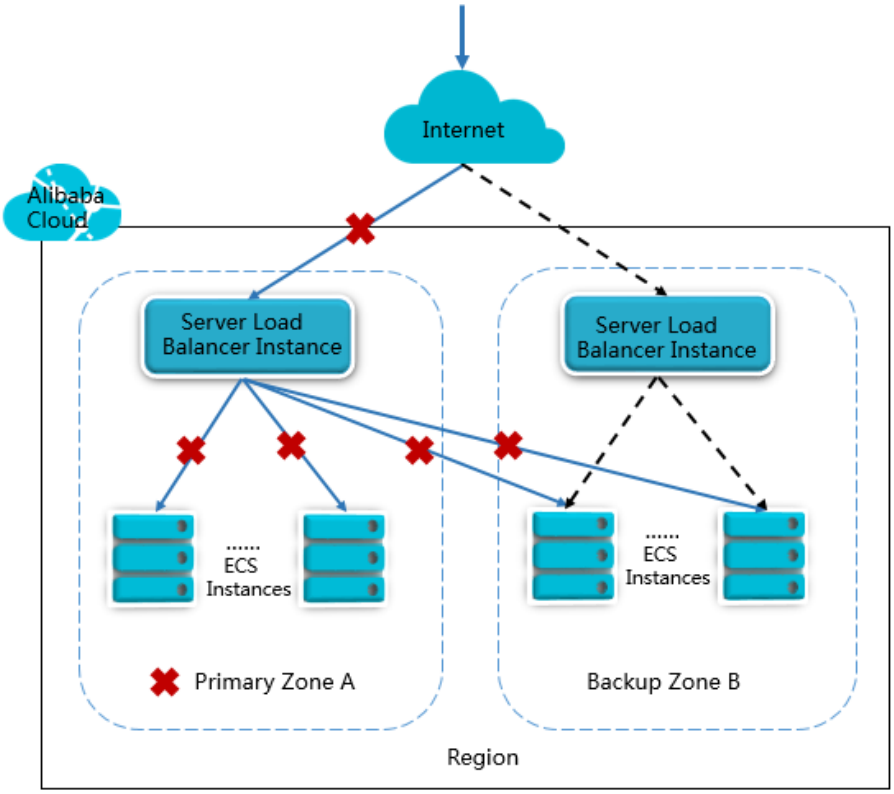
シナリオ 4: 1 つのリージョンでの耐障害性 (マルチリージョンでの耐障害性)

より信頼性が高いサービスを提供するため、Alibaba Cloud Server Load Balancer により既にほとんどのリージョンにマルチゾーンがデプロイされています。

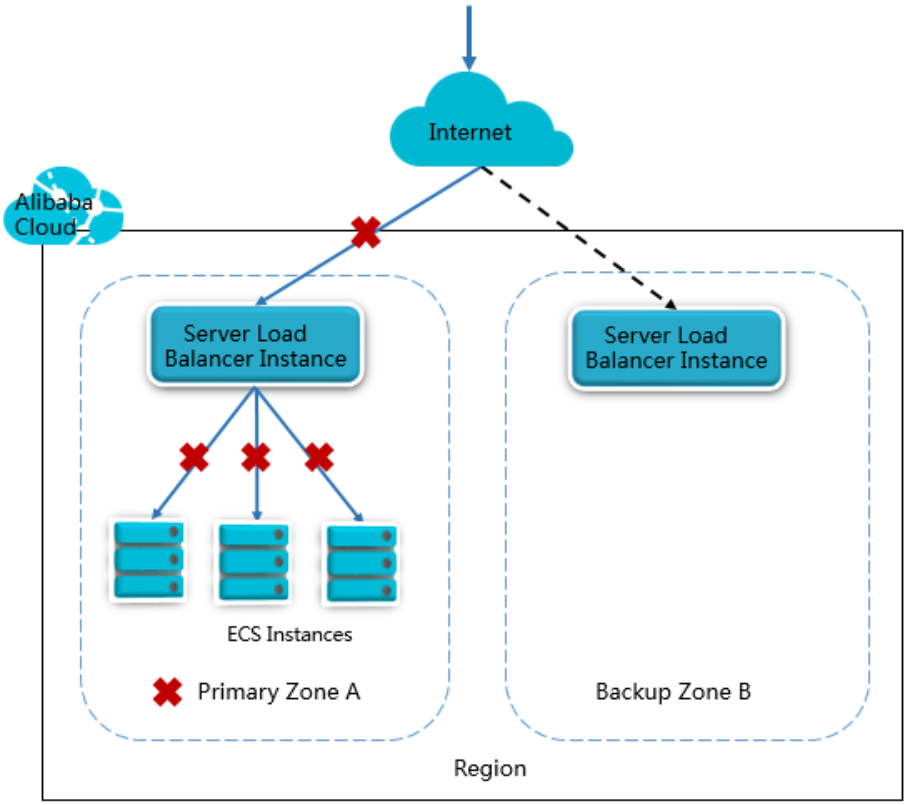
プライマリゾーンが使用できないとき、Server Load Balancer サービスは迅速にバックアップゾーンに切り替えて、30 秒以内にサービス機能を復元します。プライマリゾーンが使用可能になると、Server Load Balancer サービスはプライマリゾーンに自動的に戻ります。

耐障害性のため、同じリージョンの複数のゾーンで Server Load Balancer インスタンスを作成することをお勧めします。さらに、ECS インスタンスを合理的にデプロイして、高可用性または低レイテンシのロードバランシングサービスを利用できます。各ゾーンに少なくとも 1 つの ECS インスタンスを追加することがベストプラクティスです。

次の図に示すように、プライマリゾーン A が正常に動作している場合、青の線のようにプライマリゾーン A の ECS インスタンスにトラフィックが分散されます。プライマリゾーン A が使用できなくなると、黒の点線のようにバックアップゾーンの ECS インスタンスにトラフィックが分散されます。これにより、1 つのゾーンの障害によるビジネスの中断が回避されると共に、さまざまなゾーンのプロダクト間のレイテンシが減ります。

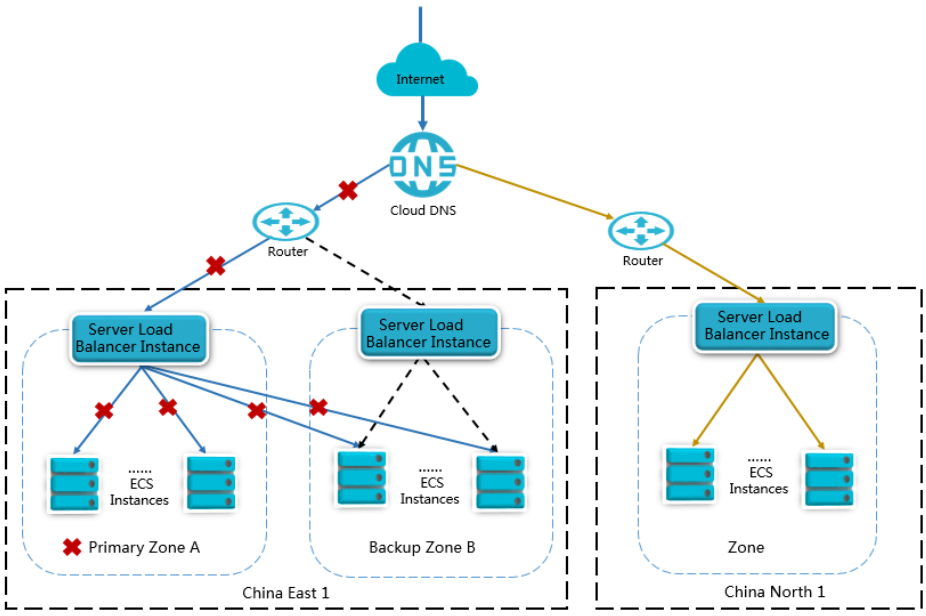


すべての ECS インスタンスをプライマリゾーンに追加し、バックアップゾーンに ECS インスタンスがない場合 (下図を参照)、高可用性を犠牲にして低レイテンシが達成されます。この状況では、プライマリゾーンが使用できなくなると、Server Load Balancer サービスはバックアップゾーンに自動的に戻ります。ただし、分散化された要求を処理する ECS インスタンスがありません。



シナリオ 5: クロスリージョンでの耐障害性

DNS との組み合わせにより、さまざまなリージョンで Server Load Balancer サービスをデプロイし、DNS を使用してドメインを Server Load Balancer インスタンスの IP アドレスに解決することで、クロスリージョンの耐障害性を達成できます。リージョンが使用できなくなったときは、ドメイン名解決を中止するだけで済み、ユーザーアクセスへの影響はありません。



Server Load Balancer プロダクトの制限

制限	一般的なユーザーに対する制限の説明	例外適用方法(例外的な上限)
ServerLoadBalancerインスタンスの作成に関するECS制限	ユーザーは少なくとも1つのECSインスタンスを保持している必要があり、ServerLoadBalancerインスタンスを作成する場合、ECSインスタンスを保持しているリージョンのみを選択できます。	例外なし
使用可能なServerLoadBalancerインスタンスのタイプ	インターネット/イントラネット	例外なし
ServerLoadBalancerインスタンスの課金方法	従量/帯域課金	例外なし
単一リスナーのピークパブリック帯域幅の範囲(従量課金)	1～5,000MB/無制限、デフォルト:1MB	例外なし
従量課金インスタンスのユーザーのデフォルトクォータ	60	チケット
ServerLoadBalancerインスタンスのエイリアスに対する制限	長さは1～80文字(文字、数字、”-”、”/”、”.”、”_”を含む)	例外なし
ServerLoadBalancerインスタンスリスナーの数	50	例外なし
レイヤ7リスナーのドメイン/URL転送追加可能なルールの数	20	例外あり
ServerLoadBalancerモニタリングで使用できるプロトコルのタイプ	HTTP/HTTPS/TCP/UDP	例外なし
ServerLoadBalancerモニタリングのフロントエンド/バックエンドポートの範囲	1～65535	例外なし
ServerLoadBalancerモニタリングの転送ルール	wrr/wlc、デフォルト:wrr	例外なし
HTTPプロトコル-セッション維持-クッキーの処理方法	insert/server、デフォルト:insert	例外なし
HTTP-セッション維持-クッキーのタイムアウト時間	1～86,400、デフォルト:3,600	例外なし

HTTP-セッション維持-クッキー名	長さは1～200文字。クッキーはRFC2965に準拠している必要があります。つまり、ASCII英文字と数字のみが含まれ、コンマ、セミコロン、スペースを含むことはできず、”$”文字を先頭にすることもできません。	例外なし
HTTP-ヘルスチェック-ポート	入力範囲:1～65,535。デフォルトのポートはバックエンドサーバーポートです。	例外なし
HTTP-ヘルスチェック-ドメイン	長さは1～80文字。文字、数字、’-‘および’.’のみが使用できます。	例外なし
HTTP-ヘルスチェック-URI	長さは1～80文字。文字、数字、”-”、”/”、”.”、”%”、”?”、”#”、および”&”のみが使用できます。	例外なし
HTTP-ヘルスチェック-タイムアウト時間	入力範囲:1～300。デフォルト値:5	例外なし
HTTP-ヘルスチェック-チェック間隔	入力範囲:1～50。デフォルト値:2	例外なし
HTTP-ヘルスチェック-正常状態しきい値	入力範囲:2～10。デフォルト値:3	例外なし
HTTP-ヘルスチェック-異常状態しきい値	入力範囲:2～10。デフォルト値:3	例外なし
TCP-セッション維持-タイムアウト時間	0～3,600、デフォルト:1,000	例外なし
TCP-ヘルスチェック-ポート	入力範囲:1～65,535。デフォルトのポートはバックエンドサーバーポートです。	例外なし
TCP-ヘルスチェック-タイムアウト時間	入力範囲:1～300。デフォルト値:5	例外なし
TCP-ヘルスチェック-チェック間隔	入力範囲:1～50。デフォルト値:2	例外なし
TCP-ヘルスチェック-正常状態しきい値	入力範囲:2～10。デフォルト値:3	例外なし
TCP-ヘルスチェック-異常状態しきい値	入力範囲:2～10。デフォルト値:3	例外なし
UDP-ヘルスチェック-ポート	入力範囲:1～65,535。デフォルトのポートはバックエンドサーバーポートです。	例外なし
UDP-ヘルスチェック-タイムアウト時間	入力範囲:1～300。デフォルト値:5	例外なし
UDP-ヘルスチェック-チェック間隔	入力範囲:1～50。デフォルト値:2	例外なし

UDP-ヘルスチェック-正常状態しきい値	入力範囲:2～10。デフォルト値:3	例外なし
UDP-ヘルスチェック-異常状態しきい値	入力範囲:2～10。デフォルト値:3	例外なし
バッチで追加/削除できるバックエンドECSインスタンスの数	20	例外なし
単一キーのAPIアクセス頻度制限	5,000回/日	現在、自動プロセスは使用できません。サポートが必要な場合はカスタマーサービスにお問い合わせください。
単一ユーザーによってアップロードされる証明書の最大数	100	現在、自動プロセスは使用できません。サポートが必要な場合はカスタマーサービスにお問い合わせください。

プロダクト用語

用語	正式名	説明
Server Load Balancer	サーバロードバランサー	Alibaba Cloud Server Load Balancerとはトラフィックの負荷分散制御機能で、設定されたスケジューリングアルゴリズムとリスニングルールに基づき複数のECS間における受信トラフィックの負荷分散を行います。
Server Load Balancer Instance	サーバロードバランサインスタンス	Server Load BalancerインスタンスとはServer Load Balancerのサービスとして起動しているインスタンスのことです。 Server Load Balancerを使用するには初めにServer Load Balancerを作成する必要があります。インスタンスIDとはServer Load Balancerインスタンスの一意の識別子です。
Server Load Balancer IP	サーバロードバランサIP	IPアドレスはServer Load Balancerインスタンス作成後に割り当てられます。インスタンスタイプによりIPアドレスはパブリックIPかプライベートIPかのどちらかになります。外部サービスを提供するためにパブリックIPの名前解決を使用するこ

		とができます。
Listener	リスナー	リスナーとは受信リクエストがどのように配信するかを定義付けするものです。ユーザは少なくとも一つのリスナーをServer Load Balancerインスタンスに追加しなければなりません。
BackendServer	バックエンドサーバ	バックエンドサーバとは配信されたリクエストを処理するECSインスタンスです。
VServer Group	VServerグループ	配信されたリクエストを受信するECSインスタンスグループです。他のリスナーは他のVServerグループを利用でき、listener dimensionの中でリクエスト配信を保持できるようになります。
Multiple Zones	マルチゾーン	Server Load Balancerはすでに災害対策用に複数のゾーンにデプロイされています。デフォルトではプライマリゾーンで負荷分散サービスを行います。しかしプライマリーゾーンが使用不可となったときインスタンスはサービスを継続するために自動的にバックアップゾーンに切り替わります。さらにプライマリーゾーンが復旧した後、元のプライマリーゾーンに切り替わります。このことにより当該地域での可用性が大いに高まります。