Auto-Scaling을 통한 서버 자동 확장

개요

시스템을 구축할 때 최대 동시 접속자나 Peak 때의 자원 사용량 등 시스템의 성능을 고려한 설계를 해야 합니다. 하지만 서비스에 충분한 시스템 용량을 예측하기 어렵거나 비용 확보 등의 문제로 사용량 급증을 대비한 용량 산정을 설계에 미리 반영하기 어려운 경우가 있습니다.

이 경우 SDS Cloud의 VM Auto-Scaling 기능을 이용하여 예측하지 못한 부하나 주기적인 요청 증가에 대비하면서도 비용 효율적인 인프라 서비스 환경을 운영 할 수 있습니다. 자체 모니터링 기능과 연계, 수요에 탄력적으로 반응하여 서버를 증설하고 상단 Load Balancing을 통해 서비스 중단 없는 확장과 원활한 부하분산 효과를 제공합니다

아키텍처 다이어그램

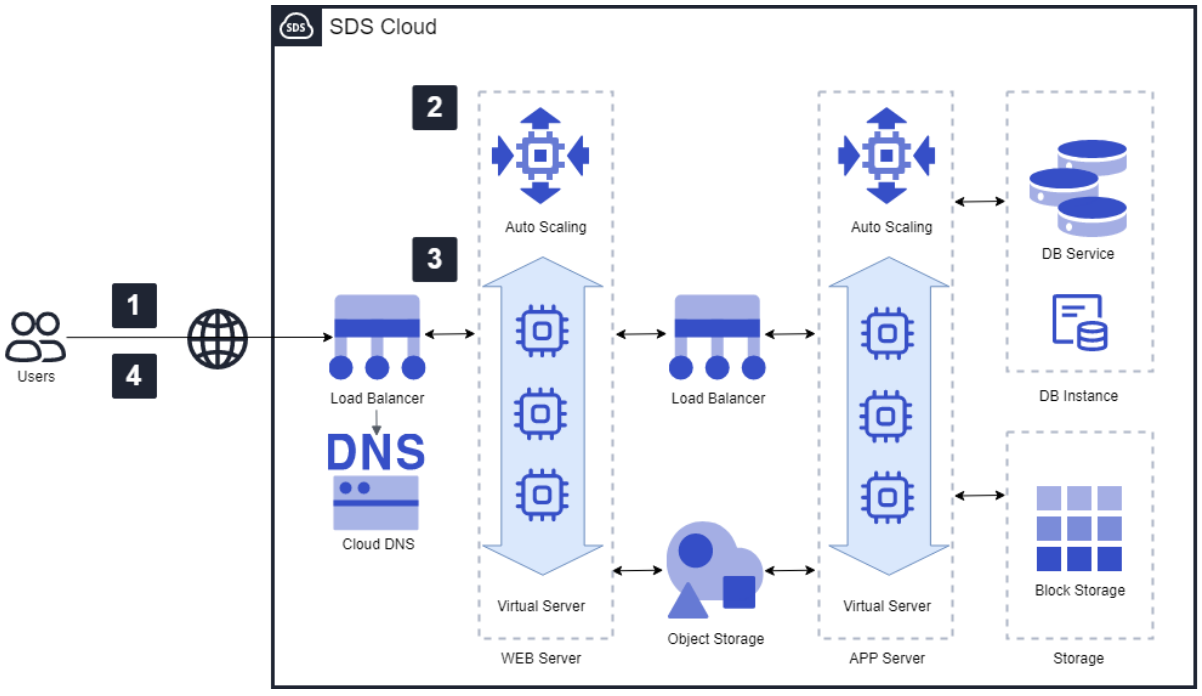

Figure 1. Auto-Scaling을 통한 서버 자동 확장

1. 부하가 많지 않은 상황에서 Auto-Scaling Group은 초기 설정 자원만큼만 가동되어 운

영된다.

2. 서비스 요청량이 증가하여 시스템의 사용률이 운영자가 사전 설정한 임계값(CPU, MEM, Network, Disk I/O)에 도달하면, Auto-Scaling Instance는 사전 정의된 단위만큼 추가 **Virtual Server**를 프로비저닝한다.
3. **Load Balancer** 에 새로 기동된 **Virtual Server** Instance의 IP가 추가되어 부하 분산에 포함된다.
4. 사용자는 시스템의 Scale In/Out과 상관없이 연속적인 서비스 사용이 가능하다.

사용 사례

A. 온라인 쇼핑몰 구축

연말이나 명절 등 특정 시점에 대규모의 이벤트/트래픽이 발생할 수 있는 시스템에서 실시간 사용량 모니터링 및 증설로 수요에 탄력적으로 대응할 수 있습니다.

B. 미디어 서비스 front-end 구축

소셜 미디어 서비스는 출퇴근 시간 트래픽 급증이나 이벤트로 인한 불규칙적인 트래픽 폭증으로 트래픽 증감에 민감합니다. 이러한 경우 적절한 Instance Type과 Auto-Scaling을 활용한 Scale-In/Out 정책 설정으로 안정적인 서비스와 성능을 제공할 수 있습니다.

선결 사항

Auto-Scale Instance 생성 시 사용할 **Load Balancer**가 미리 생성되어 있어야 합니다.

제약 사항

VM Auto-Scaling 상품을 통해 자동 확장되는 대상은 **Virtual Server** 자원으로 한정되며 Compute 상품 군내의 다른 상품(ex. **Bare Metal Server**)에는 적용되지 않습니다. Container 상품의 경우 자체 Auto-Scaling 을 설정하여 사용할 수 있습니다.

고려 사항

A. Instance Type 선택

일반 Virtual Server로 생성한 상품을 VM Auto-Scaling 상품으로 전환할 수는 없으므로 auto-scale 요건이 있는지 시스템 설계 시 미리 고려하여야 합니다.

기본 설정한 Instance type으로 증설되므로 최적의 자원 사용과 과금 플랜을 위해 부하 상황을 모니터링하여 임계값과 증설 단위를 적절히 조절할 필요가 있습니다.

B. 서버/어플리케이션 환경

Instance 증설 시 추가된 IP 주소는 Load Balancer에 추가되어 서비스되므로 Auto-Scaling Group이나 앞 단에서 사용할 어플리케이션이 대표 VIP(Virtual IP)를 통한 통신 방식을 제공하는지 사전 확인이 필요합니다.

Auto-Scaling 임계값 자원 사용률 모니터링을 위해 서버 프로비저닝 될 때 모니터링 에이전트가 설치됩니다. 성능 수집을 위해 해당 모니터링 프로세스는 가동된 상태로 유지되어야 합니다.

VM Auto-Scaling 으로 생성된 Virtual Server는 File Storage 가 지원되지 않아 Auto-Scaling Group 서버와 파일 공유 요건이 있는 경우는 Object Storage 를 통한 파일 공유 방식을 권장합니다.

C. 과금 요소

상품 자체의 과금은 없으나 VM Auto-Scaling을 통해서 생성되는 자원(Virtual Server, Block Storage 등) 에 대해서는 해당 자원 별 기 정의된 요금으로 과금 됩니다.

사용하고자 하는 IP 주소를 사전에 예약해 두고 scale-out 될 때 사용할 수 있지만 예약 IP 주소를 사용하는 경우 과금 될 수 있습니다.

D. Compliance

Auto-Scaling Group 내의 Virtual Server에 상용 어플리케이션을 사용하는 경우 서버 대수와 그에 따른 코어의 유동적인 변경으로 라이선스 이슈가 발생할 수 있습니다. 사전에 어플리케이션의 라이선스 확장과 관련한 사항을 확인할 필요가 있습니다.

관련 상품

- Virtual Server
- DB Service
- Load Balancer

- DNS
- Object Storage
- Block Storage