

Web Servislerinin Amazon EC2 Ölçeklendirilmesi

Z. İlhan¹ and T. Akgün²

¹Okan Üniversitesi, Türkiye, zanailhan@gmail.com

²Okan Üniversitesi, İstanbul/Türkiye, tevfik.akgun@okan.edu.tr

Autoscaling Web Services on Amazon EC2

Abstract—Fault tolerance, high availability & scalability are essential prerequisites for any enterprise application deployment. One of the major concerns of enterprise application architects is avoiding single points of failure. There is a high cost associated with achieving high availability & scalability. We will look at an economical approach towards automatically scaling Web service applications while maintaining the availability & scalability guarantees at an optimum economical cost. This approach, involving the Amazon EC2 cloud computing infrastructure, makes it unnecessary to invest in safety-net capacity & unnecessary redundancy.

I. GİRİŞ

AMAZON EC2; web teknolojileri ile ölçeklendirilebilir bir işlem kapasitesi ve altyapısı sağlayabilen web tabanlı bir servistir. Amazon EC2 bir sunucunun kullanıma hazır olması (bootup time) için gerekli süreci dakikalara indirerek başlangıç ve kapanış sürecinde kaybedilen zamanı minimize eder. Bu sistemin kullanımı sırasında yapılan ücretlendirme sadece kullanım süreci içerisinde kullanılan veri alan / transferine (Computing Capacity) göre düzenlenmiştir. Bu ücretlendirme modeli; bir Cloud Computing çözümünde kullanılacak ağ altyapı aygıtları (Güvenlik duvarları, Router vb.), ilgili aygıtların bakımı, barındırma masrafları ve bunların işletim zorluklarını daha az maliyetlendirmeyi hedefler.

II. EC2 ÖZELLİKLERİ

Amazon.com; web sitesi yöneticilerinin son zamanlarda dile getirdiği “hata toleransı minimize edilmiş dirençli çözümler” geliştirebilmek için tasarlanmış iki yeni özelliği ile Cloud Computing çözümlerinde büyük kolaylık sağlayacağını düşünmektedir.

III. ERİŞİLEBİLİR BÖLGELER

- Cloud Computing çözümlerinin olmazsa olmaz önkoşullarından biriside oluşabilecek sistem hatalarının ölçeklendirilmesi ve bu ölçekler dahilinde sistemin ayakta kalma süreci ve performansının maksimumdaki devamlılığını sürdürebilmesidir. EC2 temelde bu prensiple tasarlanmıştır.

- Cluster ve Cluster desteği; sunucu üzerinde çalışan uygulama bağımsız olup sistem stabilitesini etkileyen en önemli etmenlerdendir. Ayrıca cluster’ı oluşturan sunucuların hepsi tek noktadan izlenebilmeli ve gerektiğinde güncellenebilmelidir. EC2 sahip olduğu web arayüzü ve entegrasyon amaçlı geliştirilen API (Application Programming Interface) ile kullanıcının kendi yönetim sistemlerine de entegre edilebilmektedir.

IV. ELASTİK İP ADRESLERİ

- Elastik IP adresleri Cloud Computing için tasarlanmış static ip adresleridir. Elastik ip adresleri müşteri / kullanıcı hesabına adanmış olup kullanıcının bu ip adresi üzerinden dinamik olarak bu adreste bulunacak sunucuyu değiştirebilme olanağı sağlar. Bu sayede bir elastik ip’nin arkasında bir sunucu cluster’ı veya farklı mantıkta yerleştirilmiş bir dizi sunucu bulunabilir ve böylece herhangi bir sunucuda oluşabilecek sorun durumunda ilgili sunucunun yerini bir başka sunucu alabilmesi amacıyla tasarlanmışlardır.

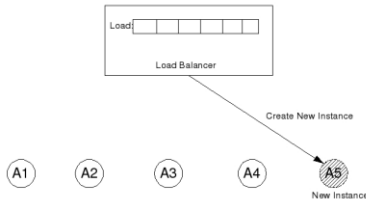
A. Apache Axis2

Axis2 web servisleri için tasarlanmış bir ara katman yazılım platformudur. Axis2 sunucular arasında clustering’i sağlayan middleware’dır. Axis2 aracılığı ile bir düğümde (cluster da bulunan herhangi bir sunucuda) bulunan durum bilgisi diğer düğümlerle gerçek zamanlı olarak paylaşılır ya da düğümlerin takip edildiği bir başka noktada kopyalanır. Aslında temelde clustering uygulaması bir başka grup iletişim katmanı’ü (Group Communication Framework (GCF)) olan Apache Tribes üzerinden sağlanmaktadır, fakat EC2 üzerinde cluster düğümleri arasında multicast membership fonksiyonuna izin verilmediğinden dolayı çalışmamaktadır.

B. Apache Synapse Dinamik Yük Dengeleme (Dynamic load balancing)

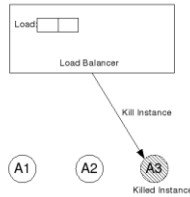
Apache Synapse yüksek performanslı basit bir ESB (Enterprise Service Bus) dir. Apache Synapse basitçe configure edilebilir, geçitleme (gatewaying) ve yük dengeleme (load balancing) entegrasyonunu kolaylaştırabilir bir yapıda tasarlanmıştır. Tipik bir yük dengeleme sisteminde yük dengeleyici sistemin giriş noktası olarak tasarlanır. İşlevsel sunucular cluster’ı yük dengeleyici gelen istekleri sahip olduğu sunucu düğümüne dağıtır. Geleneksel yük dengeleyiciler statik bir konfigürasyona sahiptirler; yani yük dağıtımını yapacakları sunucuların adreslerini sistem çalışmaya başladığında bir veri tabanı veya text dosyadan okurlar ve çalışma zamanı boyunca bu konfigürasyona bağlı

olarak çalışırlar. Fakat dinamik yük dengeleme sistemlerinde ise cluster sistemine ve load balancing sistemine katılan üyeler ve düğümler çalışma zamanında sisteme dahil olabilir veya sistemden ayrılabilir.



Şekil 1: Sunucular cluster'ında yük artışına bağlı olarak cluster'a yeni sunucu düğümü eklenmesi.

Yük artışı durumunda yük dengeleyicisine gelen mesaj uzunluğunun artışına bağlı olarak, Apache Synapse ile yürütülen yük dağıtım sistemi, çalışan EC2 sunucuları sayısını artırarak sistemin kendisine gelen istekleri aynı performans düzeyinde karşılamalarını sağlar.



Şekil 2: Sunucular Cluster'ında yük azalmasına bağlı olarak cluster'a yeni sunucu düğümü eksilmesi.

Yük azalması durumunda, yük dengeleyicisine gelen mesaj sayısının azalmasına bağlı olarak, yük dengeleme sistemi sadece ilgili yükü karşılayacak kadar EC2 sunucusunun çalışmasını sağlayarak sistem performansını korur. Bu durumda sistemin kendisine gelen istekleri mevcut EC2 sunucularıyla karşılaması sağlanır.

C. Dağıtık Sistem Hatalarını Yenmek

Dağıtık sistemlerde oluşan sistem hatalarını yenmede genel yaklaşım; sistemde bulunan kritik sistem bileşenlerini birer cluster'a dönüştürmek yani her kritik bileşen için bu bileşenlerden oluşan bir alt cluster tanımlamaktır. EC2 Deployment modeli bu yaklaşım üzerine kuruludur. Bu modele göre birden fazla EC2 üzerinde çalışan sunucular birbirlerinden bağımsız birer donanım üzerinde çalışan birer sunucu gibi davranırlar.

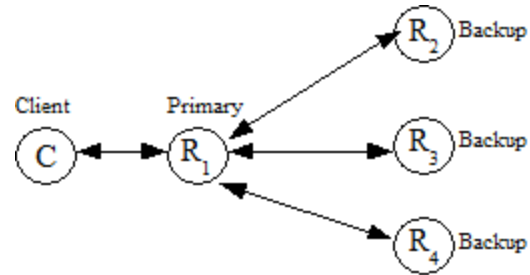
Veri Kopyalanması:

Veri replikasyonu dağıtık sistemlerde veri güvenliğini sağlamanın ve dolayısıyla sistem stabilitesinin sağlanmasıyla doğru orantılı bir parametredir. Pasif ve Aktif olmak üzere iki türde yapılabilmektedir.

• Pasif Veri Kopyalanması:

Birincil Yedekleme (Primary Backup) olarak bilinen bu kopyalama yönteminde, verilere ulaşmak isteyen bütün istemciler sadece birincil düğümlerle iletişime geçer. Birincil düğüm üzerinde saklanan bütün verilerin kopyası ise yedek düğümler üzerine kopyalanarak saklanır. Eğer birincil düğüm bir nedenden dolayı cevap veremez duruma geçerse, yedek düğümlerden herhangi biri sistem tarafından birincil düğüm

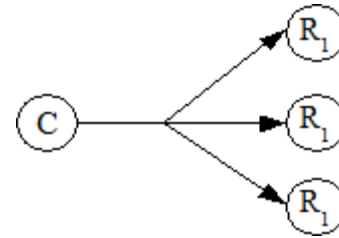
olarak atanır. Geride kalan yedek düğümleri ise yeni atanan birincil düğüme bağlanır.



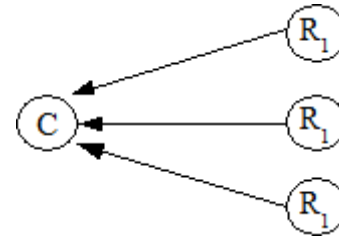
Şekil 3: Sunucular cluster'ında yük artışına bağlı olarak cluster'a yeni sunucu düğümü eklenmesi.

• Aktif Veri Kopyalanması:

İstemciden gelen istemler kopya düğümlerinin oluşturduğu düğümler grubuna gönderilir ve istemcilerin istemlerine cevabı sadece replica yöneticisi gönderir.



Şekil 1: Replikalara multicast request gönderimi



Şekil 2: Replikaların istemci istemini cevaplaması

KAYNAKLAR

- [1] Amazon Elastic Compute Cloud (Amazon EC2) <http://aws.amazon.com/ec2/>
<http://aws.amazon.com/elasticloadbalancing/>.
- [2] Apache Synapse Enterprise Service Bus (ESB) <http://synapse.apache.org/>.
- [3] Apache Axis2/Java <http://ws.apache.org/axis2/>.
- [4] Apache Tribes <http://tomcat.apache.org/tomcat-6.0-doc/tribes/introduction.html>.