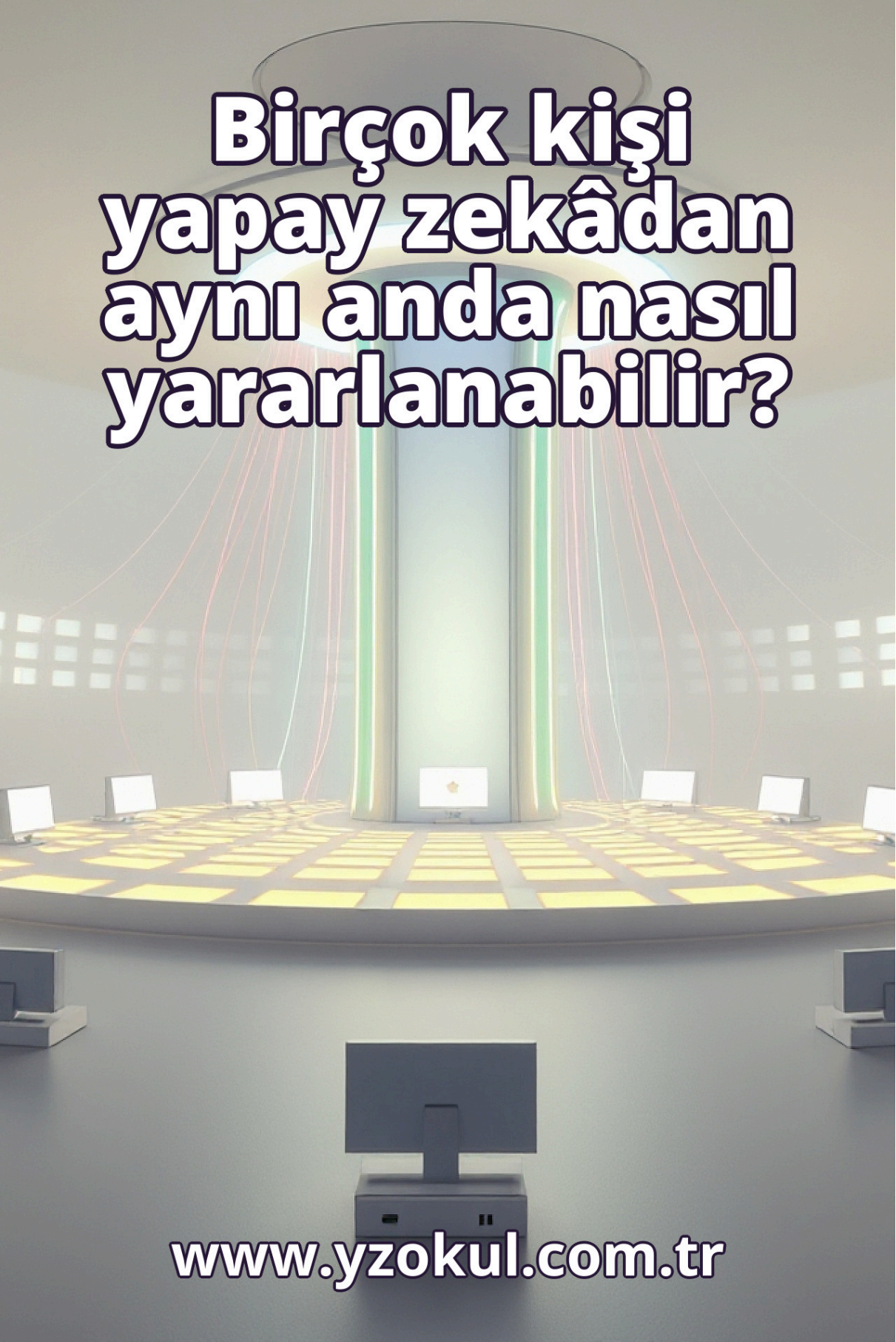

Birçok kişi yapay zekâdan aynı anda nasıl yararlanabilir?
www.yzokul.com.tr

Birçok kişi yapay zekâdan aynı anda nasıl yararlanabilir?



Birçok kişinin yapay zekâdan aynı anda nasıl yararlanabileceğini bilmek istiyorsun. Kısa yanıt, hizmet sağlayıcıların işi birçok makineye dağıttığı ve isteklerin birlikte işlenebildiğidir. Kişi başına bir özel bilgisayar yerine, paylaşılan bir sistem herkese hizmet verir. Bu kılavuz, bu sistemin her parçasını adım adım açıklar.



Bir kişiden gelen tek bir istek tek bir sunucu tarafından işlenebilir. Sunucu, ağ üzerinden diğer bilgisayarlara hizmet veren bir bilgisayardır. Bir yapay zekâ hizmetine soru yazdığında, isteğin bir sunucuya gider. O sunucu işi yapar ve bir yanıt gönderir. Bu, herhangi bir paylaşım başlamadan önceki basit başlangıç tablosudur.



Birçok kişi aynı anda istek gönderebilir. Her istek bir yanıt gerektirir ve hizmet, herkese ayrı bir özel makine vermeden bunların hepsini işlemelidir. Bu, kurulum için çok maliyetli ve çok yavaş olurdu. Bunun yerine hizmet, birçok kişiye hizmet verebilen paylaşılan makineler kullanır. Zorluk, bu paylaşımı iyi düzenlemektir.



Hizmet sağlayıcılar istekleri birden çok işlemciye veya sunucuya dağıtır. İşlemci, bir bilgisayarın talimatları yerine getiren parçasıdır. Her makine, tek bir makinenin her şeyi yapması yerine toplam işin bir bölümünü üstlenir. Bu dağıtım, hiçbir bilgisayarın tüm yük altında ezilmemesi anlamına gelir. İş, birçok isteğin birlikte işlenebilmesi için bölünür.



Bazı hesaplamalar bir araya getirilebilir. Gruplama, donanımın aynı anda daha fazla yararlı iş yapmasını ve kapasitesini daha verimli kullanmasını sağlar. Örneğin, birçok küçük istek tek bir daha büyük işte toplanabilir. Makine daha sonra bu grubu birçok ayrı geçiş yerine tek geçişte işler. Bu, zaman kazandırır ve mevcut donanımdan daha iyi yararlanılmasını sağlar.



Gelen istekler, hiçbir makinenin aşırı yüklenmemesi için paylaştırılır. İşin bu şekilde dağıtılmasına yük dengeleme denir. Yük dengeleyici, her isteği boş kapasitesi olan bir makineye yönlendiren bir sistemdir. Bir sunucu meşgulse, istek başka bir sunucuya gider. Yük dengeleme, tüm sistemin birçok kullanıcı için sorunsuz çalışmasını sağlar.



Bir okulda birçok öğrencinin aynı anda aynı yapay zekâ hizmetini kullandığını düşün. İstekleri farklı makinelere dağıtılır, böylece hizmet herkes için çalışmaya devam eder. Bir öğrenci matematikte yardım isterken bir başkası özet ister. Yük dengeleyici her isteği onu işleyebilecek bir sunucuya gönderir. Okulun her öğrenci için ayrı bir özel makineye ihtiyacı yoktur.



Hizmeti daha fazla kişi kullanmaya başladığında yük artar. Sağlayıcı, ek yükü taşımaya yardımcı olmak için daha fazla kaynak ekleyerek yanıt verebilir. İşi paylaşmak için ek sunucular veya işlemciler devreye alınabilir. Bu ölçek büyütme, hizmetin öncekinden daha fazla kişiye hizmet verebileceği anlamına gelir. Kaynak eklemek, artan talebi karşılamanın bir yoludur.



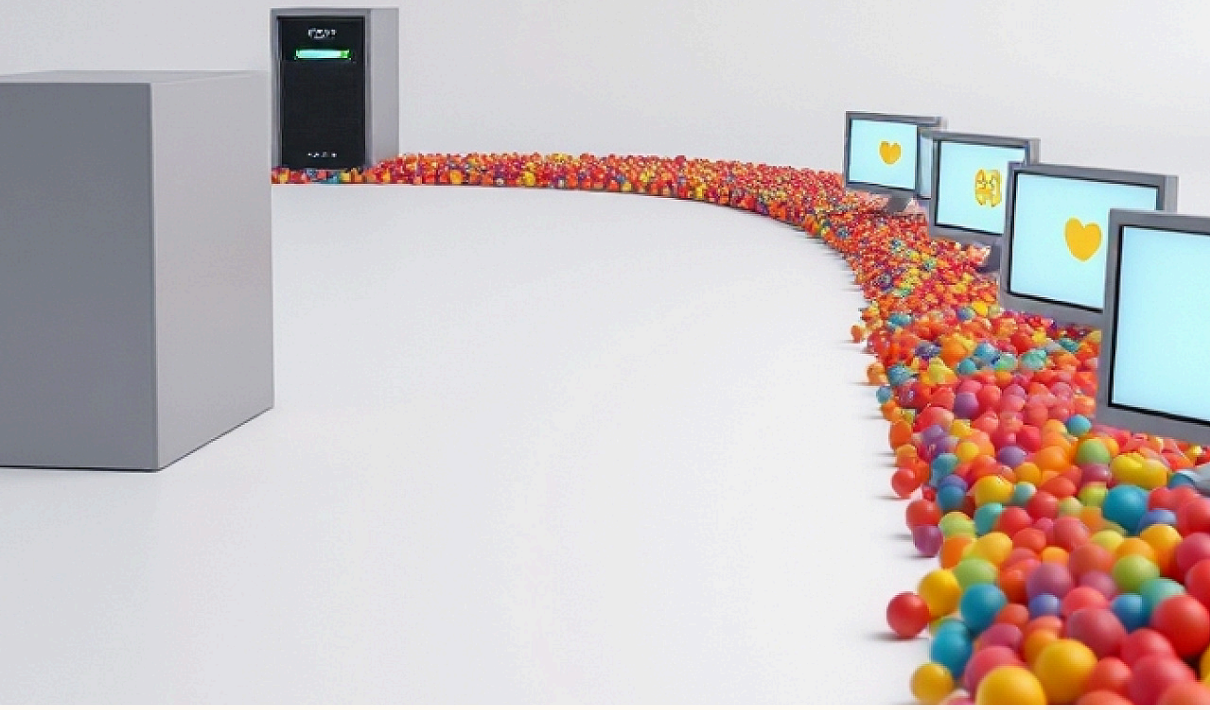
Yük arttığında istekler bir kuyruğa konabilir. Kuyruk, isteklerin sıralarını beklediği bir hattır. Kuyruktaki istekler kaybolmak yerine bekler. Sistem, makineler boşaldıkça bunları sırayla işler. Kuyruk, kimsenin isteğini düşürmeden geçici bir yoğunluğu yönetmeye yardımcı olur. Kuyrukta beklemek normaldir ve hizmetin adil kalmasını sağlar.



Sağlayıcılar ek makineler veya kapasite devreye alabilir. Bu ek yardım, hizmetin öncekinden daha fazla isteği işlemesini sağlar. Örneğin, yoğun dönemlerde daha fazla sunucu açılabilir. Ek makineler paylaşılan havuza katılır ve işin bir bölümünü üstlenir. Kaynak eklemek, daha yüksek talebe verilen planlı bir yanıttır, bir başarısızlık işareti değildir.



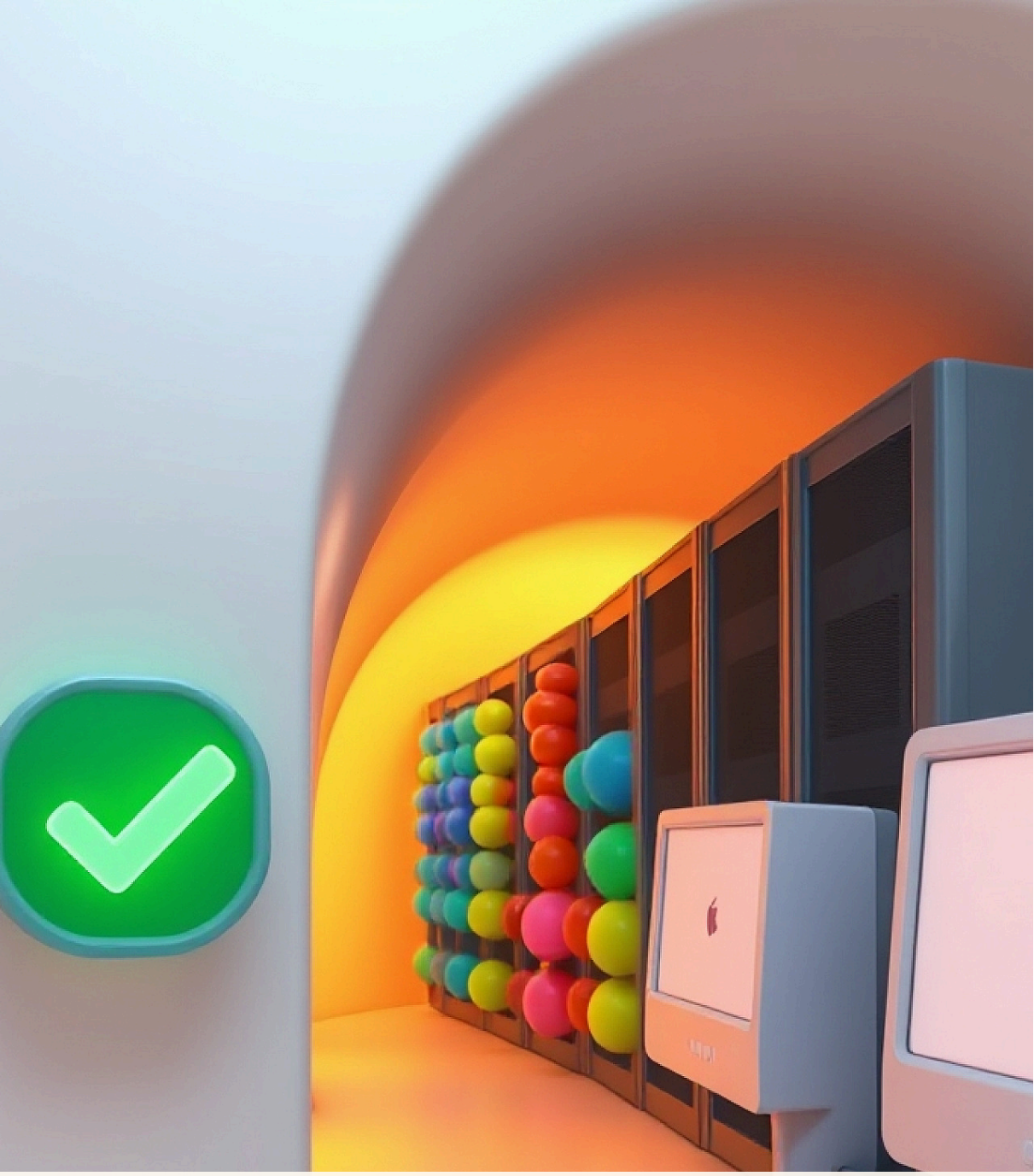
Her kurulumun bir azami kapasitesi vardır. Kapasite, bir sistemin aynı anda işleyebileceği en fazla iştir. Paylaşım ve kuyruklarla bile ne kadar iş yapılabileceğinin bir sınırı vardır. Makinelerin sonlu bir hızı ve belleği vardır. Bu sınırı bilmek sağlayıcıların plan yapmasına yardımcı olur ve gecikmelerin neden olabileceğini anlamana yardımcı olur.



Talep kapasiteyi aştığında gecikmeler olabilir. İsteklerin tamamlanması daha uzun sürebilir çünkü makinelerin hızlıca işleyebileceğinden daha fazla iş vardır. Kuyruk uzar ve bekleme süreleri artar. Bu bir arıza değildir; çok fazla talebin doğal sonucudur. Çok yüksek talep, iyi tasarlanmış sistemlerde bile gecikmelere yol açabilir.



Yaygın bir hata, tek bir büyük makinenin herkese hizmet verdiğini düşünmektir. Gerçekte iş, birçok kişiye hizmet vermek için birçok makineye dağıtılır. Tek bir makine tek başına çok yavaş olur ve kolayca aşırı yüklenebilir. Birçok kişinin aynı anda yararlanabilmesini sağlayan şey, sunucular arasında paylaşımdır. Gücün, birlikte çalışan birçok makineden geldiğini unutma.



Bir başka hata, kuyrukta beklemenin hizmetin bozuk olduğu anlamına geldiğini düşünmektir. Kuyruk, hemen işlenebilecekten daha fazla isteği yönetmenin normal bir yoludur. İstekleri güvende tutar ve sırayla işler. Kuyruklar, başarısız bir hizmetin değil, meşgul bir hizmetin işaretidir. Sistemin baskı altında kararlı kalmasına yardımcı olurlar.



Kapasitenin sınırsız olarak sonsuza kadar büyüyebileceğine inanmak da yanlıştır. Sağlayıcılar kaynak ekleyebilir, ancak yine de bir azami değer vardır. Fiziksel makineler, enerji ve maliyet sınırlar koyar. Talep bu azami değeri aştığında gecikmeler kaçınılmazdır. Bu sınırı anlamak, meşgul hizmetler için gerçekçi beklentiler belirlemene yardımcı olur.



Özetle, istekler sunuculara dağıtılır, verimlilik için gruplanır, dengelenir, kuyruğa alınır ve gerektiğinde ek kaynaklar verilir. Yük dengeleme her isteği boş kapasitesi olan bir makineye yönlendirir. Kuyruklar, makineler hazır olana kadar istekleri tutar. Talep arttıkça ek kaynaklar eklenir. Kapasitenin sınırları olduğunu, bu nedenle çok yüksek talebin gecikmelere yol açabileceğini unutma.