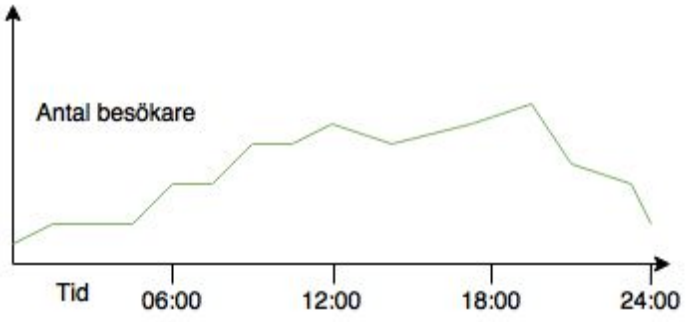
Automatisk skalning av webbtjänster genom att använda ett adaptivt distribuerat system

Tuan Nguyen & David Jaenson

Företag måste ofta göra en avvägning mellan snabba laddningstider av webbtjänster och mängden pengar företaget är villigt att betala för resurser i form av datorkraft. I detta examensarbete undersöker vi metoder för att minska resursanvändning medan vi fortfarande bibehåller, i största möjliga mån, snabba laddningstider.

Hemsidors laddningstider är viktiga för att göra besökare nöjda och inte skapa frustration. Om en hemsida eller tjänst på internet tar för lång tid att ladda finns en stor risk att användaren lämnar sidan. Det är därför viktigt för företag att hemsidor och tjänster tilldelas tillräckligt med resurser så att dessa kan ladda snabbt och felfritt. Att tilldela mer resurser till en tjänst eller hemsida kostar dock pengar. Företag måste därför göra en avvägning; att antingen tilldela mycket resurser, få korta laddningstider och nöjda besökare, men med höga kostnader. Alternativt att tilldela lite resurser, få långa laddningstider och missnöjda besökare men låga kostnader.

Då antalet användare av en viss tjänst bestämmer hur mycket resurser som behöver avsättas för att hantera tjänsten och antalet besökare till en viss tjänst varierar över tid, så använder ofta företag något som kallas *statisk allokering* av resurser till sina tjänster. Den typ av statisk allokering som ofta används är att antalet resurser allokeras så att tjänsten eller hemsidan kan hantera det maximala antalet förväntade besökare. Detta innebär att tjänsten har en konstant mängd resurser. Som visas i Figur 1 så varierar i regel antalet besökare med tiden. Detta leder till att en del av resurserna är oanvända en stor del av tiden. Det innebär i sin tur att både energi och pengar spenderas på resurser som inte används, vilket är problematiskt.



Figur 1. Antalet besökare till hemsidor varierar över tid. Bilden ovan visar ett typiskt användarmönster för antalet besökare över 24 timmar för en hemsida. Som bilden visar så tenderar antalet besökare vara färre på natten. Företag varierar ofta inte mängden resurser i form av datorkraft över tid. Det leder till oanvända resurser när antalet besökare är lågt.

I det här examensarbetet har vi undersökt hur man kan effektivisera resursanvändningen genom att använda något som kallas *dynamisk allokering* av resurser. Det innebär att vi varierar resurserna med tiden baserat på hur mycket resurser som faktiskt krävs av en tjänst vid en viss tidpunkt. Mer specifikt så har vi undersökt hur man med tiden kan variera resurserna för en internettjänst, samtidigt som vi vill garantera en maximal laddningstid av samma tjänst. Vi bygger i det här examensarbetet ett komplett ramverk som automatiskt sköter anpassningen av resursallokeringen.

Vi bygger även i det här arbetet flera olika schemaläggningsalgoritmer som bestämmer hur mycket resurser som ska allokeras baserat på de nuvarande laddningstiderna i systemet. Ett exempel på en sådan schemaläggningsalgoritm är att alltid allokera nya resurser till systemet om ett anrop till systemet har tagit mer än 5 sekunder. Om något sådant anrop inte existerar tar vi i stället bort resurser från systemet. Vi testar vårt ramverk med hjälp av schemaläggningsalgoritmerna samt utvärderar och diskuterar resultaten av våra tester.

Vi visar att vi kan minska resursanvändningen samtidigt som vi fortfarande kan upprätthålla en maxgräns för laddningstider för en webbtjänst för majoriteten av alla anrop. Detta genom att använda oss av vårt egenbyggda ramverk och dynamisk allokering. För en av schemaläggningsalgoritmerna lyckas vi minska resursanvändningen med 30 procent samtidigt som vi lyckas upprätthålla maximala laddningstider på 5 sekunder i 97.6 procent av fallen. Detta kan potentiellt motsvara en kostnadssänkning med upp till 30 procent samtidigt som antalet överskridna laddningstider endast är 2.4 procent.