LUNDS UNIVERSITET

Nya metoder för ett pålitligare moln

Tommi Berner

Institutionen för Reglerteknik

Populärvetenskaplig sammanfattning av doktorsavhandlingen *Modeling and Control for Improved Predictability of Cloud Applications*, juni 2022. Avhandlingen kan laddas ner från `www.control.lth.se/publications`.

När du öppnar din webbläsare och klickar in dig på din favoritbutiks hemsida, skickas en förfrågan från din dator över internet till *molnet* för att hämta innehållet. Molnet utgörs, i alla fall i det här sammanhanget, av stora datacenter med beräkningskapacitet som motsvarar hundratusentals vanliga datorer. Den delen av det stora datacentret som har ansvar för att svara på just din förfrågan kallas en *server*. I den bästa av världar nås din förfrågan av en ledig server direkt, som blixtsnabbt svarar med allt innehåll som behövs för att beskriva din favoritbutik. Tyvärr är så inte alltid fallet, vilket den som till exempel försökt köpa biljetter online till en populär konsert eller fotbollsmatch kan intyga.

Även om beräkningskapaciteten i molnet kan tyckas vara oändlig, är den faktiskt trots allt begränsad. Dessutom måste din favoritbutik, precis som alla andra aktörer, betala för att reservera beräkningskraft för sina servrar i molnet. Samtidigt vill såklart du som använder hemsidan inte behöva vänta särskilt länge på att sidorna ska ladda, då de flesta av oss ju är ganska otåliga. De tider när många andra också vill besöka din favoritbutik på nätet, blir det en utmaning för butiken att både hålla laddningstiderna nere samtidigt som kostnaderna för servrarna i molnet inte får rusa iväg. Situationen är lik de köer som många av oss upplever varje dag vid klockan 18 i mataffärer. De digitala förfrågningar som köar i servrar i molnet, kan väldigt väl liknas vid köerna fulla av stressade människor som väntar på att få betala vid kassan. Lösningen på långa köer i mataffärerna brukar bestå av att öppna fler kassor, och motsvarigheten i molnet är att betala för fler servrar. Det här kan man såklart göra, men det är inte helt lätt eftersom det kan ta ett tag innan de nya servrarna är redo. Dessutom är inte alla servrar lika snabba på att hantera förfrågningar, precis som alla medarbetare inte jobbar lika effektivt i kassan i mataffären.

I den här avhandlingen diskuteras och presenteras nya metoder för att hantera precis det här problemet som uppstår i molnet, då många användare vill komma åt samma hemsida. Utöver hemsidor finns det en hel del andra saker i vår vardag som körs i molnet, till exempel gömd funktionalitet i allt från bilar till smarta klockor. De här olika programmen som körs i molnet kan benämnas som *applikationer*. Avhandlingens mål är att bidra till att alla applikationer som

körs i molnet, kan vara responsiva och pålitliga även under tider med hög belastning. Metodiken som används bygger på koncept från två olika ämnesområden, reglerteknik och köteori.

Reglerteknik är ett ämne där tillämpad matematik utnyttjas för att styra system. Viktiga exempel där reglertekniken gjort stor skillnad är bland annat autopiloter i flygplan, farthållare i bilar och regulatorer i kemiska processer i fabriker. Intressant nog kan den matematiska modellering som tidigare gjorts för att styra vätskenivåer i stora tankar, i princip rakt av kopieras för att istället hålla köer i servrar korta. Det koncept som i den här avhandlingen utnyttjas för att direkt kunna påverka servrars snabbhet bygger på att många applikationer i molnet svarar bättre än nödvändigt på de flesta förfrågningar. Genom att dela upp applikationer i en absolut nödvändig del som alltid används, och i en valfri del som används om det finns tid, är det möjligt att med reglertekniska metoder styra kölängder och väntetider i servrarna. Ett exempel på en applikation som skulle kunna delas upp är just en butiks hemsida. När du klickar på en vara som du är intresserad av att köpa, skulle i så fall den absolut nödvändiga delen vara en beskrivning av varan, inklusive pris och bilder, medan den valfria delen skulle kunna utgöras av rekommendationer av liknande varor. I den här avhandlingen framförs nya och förbättrade modeller och metoder för att låta applikationen själv besluta när ett fullständigt svar kan ges, vilket leder till en automatisk styrning av köerna hos servrarna.

Det andra ämnesområdet som den här avhandlingen hämtat inspiration från är köteorin, ett ämne som faktiskt handlar om precis det man skulle kunna tro. Köteorin används för att matematiskt kunna analysera köer i servrar, telefonsupport och faktiskt även de som uppkommer i mataffärer vid kassorna. Ett problem som uppkommer både i mataffärer och datacenter är konsten att hitta den kö som blir klar snabbast. En strategi som tyvärr inte funkar om man är ensam i affären är att ställa sig i flera köer samtidigt. I ett datacenter kan man däremot utan problem använda sig av den här strategin, som går ut på att inkommande förfrågningar kopieras, eller *klonas*, för att sedan skickas till många servrar samtidigt. I den här avhandlingen framförs nya koncept som förenklar den matematiska modelleringen som beskriver hur kloning påverkar väntetiden för förfrågningar. Utifrån den grunden kan sedan styrlogik utvecklas för att aktivt besluta vid vilka tillfällen som kloningen kan göra nytta för att minska väntetiderna, och därmed bidra till ett pålitligare moln.