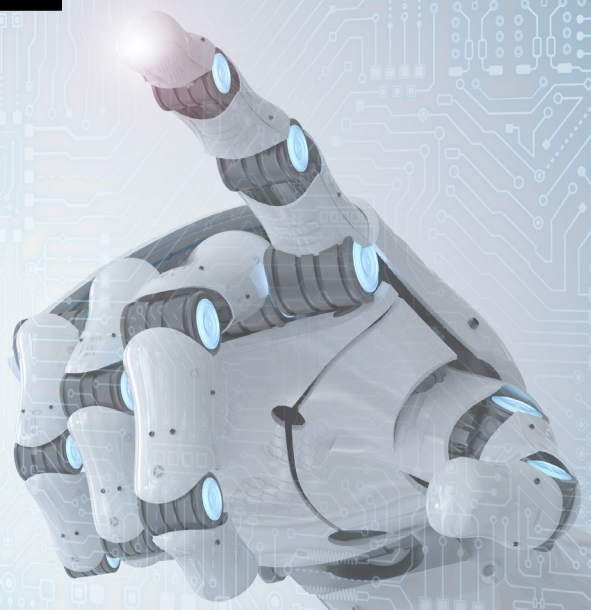
KUNDCASE

Inovia gjorde sin AI-utveckling 10 gånger snabbare med Power-plattformen

MÖT INOVIA

Inovia är ett marknadsledande bolag som är specialiserade på Big Data och AI där lösningarna utvecklas av ett eget team av utvecklare, data scientists och PhD:s i machine learning. Inovia utvecklar egna modeller och algoritmer för att hela tiden ligga i absolut framkant och kunna erbjuda unika lösningar.



UTMANINGAR, UTVECKLINGSTID OCH HÖGA KRAV

PROBLEMET

Inovia använde tidigare en x86-plattform med tillhörande GPU:er. Utmaningen var att träningen tog väldigt lång tid, typiskt runt fyra veckor för att få ut resultat av en körning. Att testa och justera ett antal variabler, ändra en parameter i taget eller kombinationer av dem när en körning tar fyra veckor skulle kräva flera års utvecklings-tid för att få ut en färdig version för lansering.

Inovia har idag många lösningar inom ett flertal områden där en är att översätta **tal till text i realtid på svenska.**

Lösningen som Inovia har utvecklat levererar en sökbar text med kategorisering, identifiering av person, känslor och mycket mer. Samtal kan summeras och man kan bland annat automatiskt kryptera känslig information eller identifiera specifika samtalsämnen. **Man kan lätt förstå att banker och andra aktörer som har krav på sig att dokumentera tycker att den här typen av lösningar är intressanta.**

conoa

RÄTT PUSSELBITAR PÅ RÄTT PLATS

För att bygga banbrytande AI krävs många pusselbitar där personal och kunnande är helt avgörande för att kunna vara ledande. Till det finns ett oerhört stort krav på träningsdata, i både mängd och kvalitet. Den sista pusselbiten handlar om att optimera träningstid, vilket kräver en infrastruktur som klarar mycket träning på så kort tid som möjligt.

För detta finns ett antal grundläggande ramverk och bibliotek där de flesta genomför träningen på GPU:er (Graphical Processing Unit) från Nvidia. CPU (Central Processing Unit) används mestadels till att förbereda data innan den skickas till GPU. Det kan exempelvis handla om frekvensjustering av ljudfiler innan de skickas över till GPU.

conoa

LÖSNINGEN

Inovia behövde hitta en lösning som kunde slutföra träning på mycket kortare tid, för att snabbare uppnå den kvalitetsnivå man satt som mål.

Efter att ha testat ett antal olika lösningar, fick man upp ögonen för IBM:s Power-plattform. Från början var Inovia tveksamma till IBM, eftersom de inte ville låsa in sig i en teknologi. Men efter att ha satt sig in i vad IBM hävdar att Power klarar, vilket är betydligt mer än vad någon annan kunde erbjuda, valde man att kontakta Conoa för att genomföra en ordentlig utvärdering av Power-plattformen.

Conoa lånade ut en POWER8-server så att Inovia fick möjlighet att testa och jämföra mot andra lösningar baserat på sina egna algoritmer och data set. Resultatet var en dramatisk tidsreduktion, en körning som tidigare tagit tio veckor tog nu en vecka.

Beslutet att välja Power var enkelt och när IBM lanserade den ännu kraftfullare POWER9 valde Inovia att uppgradera. Resultatet av investeringen blev att ledtiderna kortades ner, inlärningen blev snabbare och fler språkmodeller kunder valideras. Viktigast av allt är den kraftiga tidsvinsten vilket innebär att man kan fortsätta ligga steget före sina konkurrenter.

– Men det handlar inte bara om att välja rätt hårdvara, säger Marcus Ekendahl på Inovia, utan vi behöver även en leverantör som har kunskap att hjälpa till och ge support, och det är en såpass nischad produkt att det är ett begränsat utbud för detta, både på nätet och framförallt i Sverige.

– Conoa är tekniskt kunniga och alltid hjälpsamma, fortsätter Marcus, vi får alltid den support vi behöver och de är väldigt professionella.

conoa

HUR FUNGERAR POWER PLATTFORMEN?

AI-lösningar, modeller och algoritmer för machine learning och deep learning bygger idag på matematiska modeller där man gör flera beräkningar parallellt med varandra.

En vanlig CPU kan göra ett mindre antal samtidiga beräkningar, begränsningen sätts av hur många cores och trådar som finns på CPU:n. Som ett exempel kan en Intel CPU med 24 cores köra maximalt två trådar per core, vilket innebär att den maximalt kan gör upp till 48 (24x2) beräkningar samtidigt.

En GPU som Nvidia V100 har 5120 CUDA cores och 640 Tensor Cores (cores som är unika för Tensorflow). Med andra ord kan ett enda GPU-kort göra 100 gånger fler samtidiga beräkningar än en vanlig CPU. Det ska dock påpekas att det blir en jämförelse mellan äpplen och päron i den bemärkelsen att en CPU kan göra väldigt många saker som en GPU inte klarar. För just den här typen av beräkningar är dock en GPU väldigt mycket mer effektiv och blir därför det logiska valet.

> ***Conoa är tekniskt kunniga och alltid hjälpsamma, vi får alltid den support vi behöver och de är väldigt professionella.***
>
> *Marcus Ekendahl, Inovia*

conoa

VIKTIGT MED FÖRARBETE

Innan man kan starta sin träning krävs det ofta en del förarbete med den data man ska använda. Data ska också importeras till servern, resultat ska sparas och sedan exporteras över ett nätverk till en annan server eller lagringslösning. Allt detta kräver en CPU och således måste GPU och CPU kopplas ihop och kunna kommunicera med varandra.

På en x86 server sker all kommunikation mellan CPU och GPU över PCIe-protokoll. En utmaning med PCIe är den stora mängd instruktioner som krävs för att skicka ett paket med data, typiskt 20 000 instruktioner per paket vilket medför latens och bandbredd som används av protokolldata snarare än den data man faktiskt vill skicka.

Power-plattformen har en unik koppling från CPU till GPU, och använder inte PCIe för detta. Istället använder man NVIDIA:s NVLink, som sitter direkt på chip-nivå på både CPU och GPU. Man får en direktkoppling med väsentligt lägre latens och mycket högre tillgänglig bandbredd. Resultatet är att CPU och GPU slipper vänta på varandra vilket kan ge dramatiska effekter.

Power kan även koppla ihop olika funktioner och system över OpenCAPI och PCIe-Gen4 vilket ingen annan plattform har i dagsläget. Effekten är densamma, lägre latens och mer tillgänglig bandbredd mellan till exempel server och nätverk eller direktkopplad flashlagring.

SÅ ANVÄNDER INOVIA POWER-PLATTFORMEN

Inovia använder en setup där allting är Linux-baserat, stora delar av ramverket är open source, och mycket av det är från början utvecklat för x86-arkitekturen, vilket gör att koden behöver kompileras om för Power-processorer. Ofta finns färdiga paketeringar att tillgå och mycket delas genom OpenPOWER vilket är en stiftelse och en community för utveckling av plattformen, likt open source fast för hårdvara.

> ***POWER9 har en ohygglig prestanda, den är verkligen byggd för uppgiften.***
>
> *Magnus Halvarsson, Inovia*

När Inovia uppgraderade till POWER9 så byttes även operativsystem från Ubuntu till Red Hat som har en något äldre kodbas. Delar av programvaran fick därför skapas på nytt för att få ut max av utvecklingsmiljön. Detta är också Inovias strategi – att vara pionjärer och leda utvecklingen.

– Jag fick bygga en egen kompilator för att kunna köra på den nya plattformen, berättar Magnus Halvarsson på Inovia. Och vi bygger saker i Tensor Flow som jag inte hittat att någon annan gjort ännu, så mig veterligen är jag först i världen med vissa grejer där.

Eftersom POWER9 är så pass ny har Inovia byggt mycket själva men det börjar dyka upp allt fler bibliotek med kod för den som vill använda sig av det.

– POWER9 har en ohygglig prestanda, fortsätter Magnus, och den är verkligen byggd för uppgiften. GPU:erna sitter ihop bra med datorn i övrigt. Det gäller att få dem som skickar in jobb att skriva den bästa machine learning-koden för att inte applikationen ska behöva flytta data mellan GPU och RAM i onödan. POWER9 har kommit för att stanna, säger han avslutningsvis.

conoa

ANDRA LÖSNINGAR

Måste man använda GPU:er? Kan man inte bygga en CPU som klarar lika många parallella processer? Jo, men då blir det i princip en GPU. Själva poängen med NVIDIA:s GPU är att den klarar många enkla men parallella beräkningar, istället för en CPU som idag har ett maxantal av 192 parallella processer (men med större beräkningskraft lagd på var och en av dem).

Ett annat alternativ är att göra som Intel, att de lägger CPU och GPU i samma chip. På det viset slipper man ifrån problematiken för kommunikation mellan CPU och GPU. Snabbare blir det dock inte då NVIDIA:s GPU:er är snabbare och detsamma gäller för Powers CPU:er. Flexibiliteten minskar också, GPU-utvecklingen går just nu snabbare än CPU med ungefär ett år mellan nya modeller. Med andra ord vill man kanske uppgradera GPU oftare än CPU, vilket inte går med ett integrerat chip.

Det finns ytterligare en aspekt att ta i beaktande. Med en serverpark av GPU-maskiner för AI blir det stora datamängder som ska flyttas mellan olika fysiska servrar och med en x86 plattform är man hänvisad till PCI Gen3-kommunikation över standard IP eller snabbare infiniband. Med POWER kan man använda alternativa tekniker, exempelvis PCI Gen4 eller OpenCAPI, för att även här eliminera PCI protokollet för extern kommunikation. Det finns redan idag till exempel Infiniband-kort för OpenCAPI.

EN NY STANDARD

Använda fler maskiner då? Det är ett alternativ men kommer med sina nackdelar, framförallt på en x86-baserad miljö där varje CPU och GPU är sin egen miljö med dedikerat minne, och ser de andra som externa enheter.

Exempelvis har en NVIDIA V100 GPU 32GB minne vilket innebär att den kan jobba med data set som är max 32GB. Har man större mängd data måste man dela upp det i flera jobb. På en Power-baserad server med NVLink kan GPU och CPU prata direkt med varandra, som om de vore en enhet. Det innebär till exempel att man kan använda serverns minne för de data set som GPU:n jobbar med. Helt plötsligt kan man använda 1TB istället för 32GB vilket ger möjlighet att köra större jobb mycket mer effektivt. Dessutom kan man använda OpenCAPI för att klustra ihop upp till 256 GPU:er i en distribuerad miljö.

Så även om det finns andra lösningar, så är det numera standard bland utvecklare av AI och machine learning att bygga applikationer på GPU, FPGA eller advanced I/O för att åstadkomma workload acceleration.

NÄSTA STEG FÖR INOVIA

– Vi kommer att erbjuda AI-as-a-Service, berättar Marcus Ekendahl igen. Baserat på denna teknologi kommer vi kunna paketera det till andra företag som inte har samma kompetens att bygga modeller och algoritmer. Vi kommer också att expandera till fler länder och språk, tanken är att bygga ett OpenPOWER-kluster som kan ha kapacitet för detta.

– Därtill har vi andra projekt och produkter som vi utvecklar parallellt med denna som kommer kunna dra nytta av denna teknologi med allt vad den innebär av ökad prestanda, tillägger Marcus.

Deltag i Conoas seminarium och få djupare insikt i hur ert AI-företag kan dra nytta av Power.

ANMÄL DIG HÄR

conoa

OM CONOA

Conoa är ett företag med lång erfarenhet av att bygga och optimera traditionell infrastruktur i datacenter med Linux och andra open source-lösningar. Idag fokuserar vi på att hjälpa företag att transformera sin it-infrastruktur till moderna hybrida molnlösningar och container-teknologier.

Vi implementerar innovativa lösningar som stödjer CI/CD, DevOps, automatiserade flöden och configuration management.

Vi gör det genom att erbjuda konsulttjänster av högsta kvalitet, utbildning och ett djupt kunnande om open source-produkter.

Besök conoa.se och lär dig mer om hur vi kan hjälpa ditt företag förbättra sin infrastruktur.

Ring våra experter idag på 08-32 77 00 eller maila info@conoa.se

Conoa AB | Tegnérgatan 35 | 111 61 Stockholm