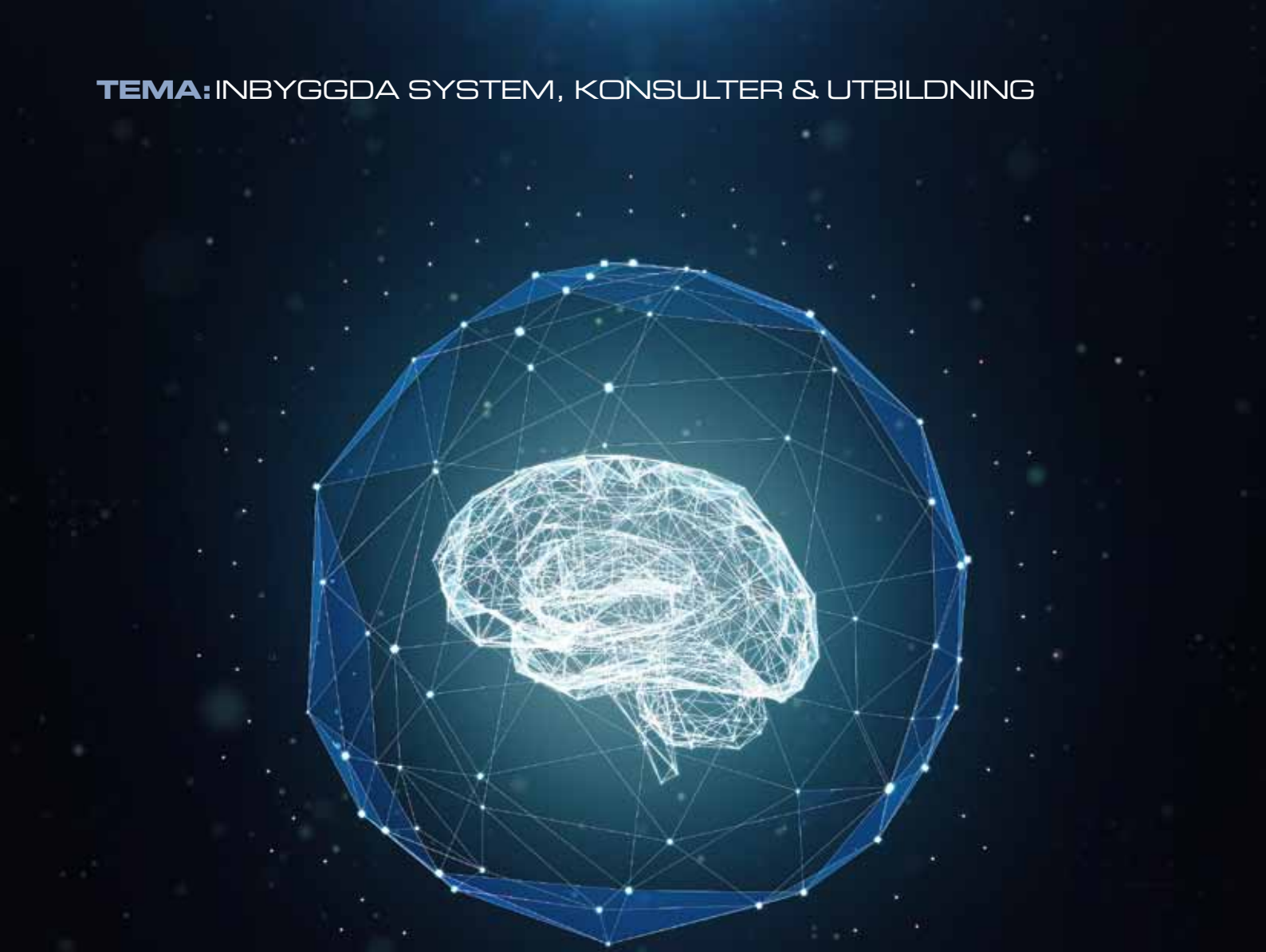

Effektiva inferenser i moln, nät och noder

EXPERTARTIKEL

Av Daniel Eaton, Xilinx

Dan Eaton arbetar med strategisk marknadsföring och är globalt ansvarig för att få fler kunder att använda FPGA:er för att accelerera sina uppgifter. Han är också ansvarig för partnerskapet med Amazon Web Services liksom Xilinx växande ekosystem med mjukvaruföretag som gör acceleratorer. Innan han började på Xilinx år 2017 hade han grundat två företag inom maskininlärning, signalbehandling och dataanalys.

AI-tillämpningar inom deep learning ärnyckeln till en ny era med högre produktivitet där den mänskliga kreativiteten kompletterats och förstärkts av maskiner. Det behövs många terabyte data för träningen liksom miljarder och åter miljarder matematiska operationer. Träningen kan göras offline i en process som tar många dagar. När väl det tränade nätverket ska implementeras är ramarna betydligt snävare.

Serverkorten i datacentren kan uppgraderas för att klara AI-beräkningar men det fysiska utrymmet är begränsat samtidigt som man måste ta hänsyn till energiförbrukningen. Dessutom förväntar sig kunderna snabba svar vilket kräver korta fördröjningar.

I förarstödssystem eller självkörande fordon där säkerhet och liv står på spel är korta svarstider och ett förutsägbart realtidsbeteende kritiskt. Samtidigt är den fysiska storleken och effektförbrukningen än mer begränsade än i datacenter plus att man måste ta hänsyn till vikten och värmeutvecklingen. Vid ett event som Tesla nyligen ordnade kallat Autonomy Day förklarades varför företaget valt att konstruera ett eget chip. Primärt handlar det om en kombination av låg effekt, under 100W, och kortare svars-

tider jämfört med grafikprocessorer.

I takt med att AI börjar användas i allt fler tillämpningar för att ge snabba svar på komplicerade frågor, kommer prestandakraven som ställs på de neurala nätverken att bli allt hårdare.

Oberoende om vi pratar om AI i molnet eller för inbyggnadstillämpningar i fordon måste inferensmotorn som kör tillämpningarna ha kort svarstid, låg effektförbrukning och ha ett litet fotavtryck.

ATT PÅ ETT BRA SÄTT förbereda ett tränat neuralt nätverk på att göra inferenser i den verkliga världen kräver inte bara beskärning och optimering utan också ett väl övervägt val av beräkningsplattform för att vara säker på att önskvärd prestanda (typiskt svarstiden) hamnar inom uppställda ramar. Det gäller effektförbrukning, storlek och termiskt fotavtryck. I takt med att de kommersiella installationerna av AI blir allt fler och slutanvändarnas krav intensifieras lanserar processortillverkarna allt mer sofistikerade arkitekturer för att möta dessa krav.

En del av de kretsar som siktar på tillämpningar som självkörande fordon har en hybridarkitektur med CPU:er och applikationsprocessorer med ett stort antal GPU:er för matematiska operationer. Trots alla resurser som finns på kretsarna är dessa arkitekturer låsta och kräver att användarna arbetar med fasta minnesbredder och antal bitar för data. Normalt är åttabitars heltal det minsta som finns även om algoritmer för deep learning kan fungera tillfredställande med data som har mycket lägre upplösning, i vissa fall ner till två eller en enda bit. De oflexibla CPU- och GPU-arkitekturerna har svårt att klara kraven från de neurala nätverken. Mer flexibla arkitekturer som kan anpassa upplösning och antal kärnor krävs för optimal beräkningsprestanda och effektförbrukning.

BESKÄRNING OCH OPTIMERING av det upplärda neurala nätverket och effektiv implementation i målprocessorn – det är svårt nog i sig. Men dessutom kommer ständigt nya och effektivare neurala nätverk – i högre takt än hårdvaran som den körs på. Ett projekt som vid starten väljer den senaste hårdvaran kommer garanterat att kännas gammalt när det är dags för den kommersiella fasen.

För att möta dessa utmaningar vad gäller prestanda, effektförbrukning och flexibilitet kan utvecklarna dra nytta av flexibiliteten i FPGA:er när de bygger sina AI-acceleratorer. FPGA:er kan konfigureras med många hundra eller till och med tusentals parallella beräkningsenheter vars upplösning är ned till en bit. Minnesgränssnitten kan skräddarsys för att eliminera flaskhalsar. Dessutom går det att programmera om FPGA:er vilket ger utvecklarna en extra möjlighet att uppdatera sina neurala nätverksstrukturer mellan olika generationer av kisel och därmed hålla jämna steg med utvecklingen.

Efter köpet av AI-specialisten DeePhi Tech år 2017 har Xilinx fått större muskler att utveckla verktyg för att beskära och optimera neurala nätverk men också IP för att implementera dessa i FPGA:er. Beskärningen förenklar det neurala nätverket genom att ta bort icke-påverkande vikter som är nära noll och organisera om nätverket där det är möjligt för att minimera antalet beräkningsoperationer och energin som går åt för att göra dem. DeePhi Techs metod för att beskära neurala nätverk är optimerad för FPGA:er och kan ta bort upp till 90 procent av vikterna samtidigt som resultatet från en bildigenkänningsuppgift är acceptabelt. Prestanda är upp till tio gånger snabbare och dessutom ökar energieffektiviteten.

SJÄLVKÖRANDE FORDON är ett lättbegripligt exempel på behovet av korta beräkningstider och minimal storlek, vikt och effekt. Objekt som upptäcks av radarn eller kameran, exempelvis andra fordon, cyklister eller fotgängare, måste identifieras inom bråkdelar av en sekund. Det är välkänt att människor kan reagera på synintryck inom en fjärdedels sekund så system för självkörning behöver vara minst lika snabba, helst lite bättre. För att kunna matcha en människa måste systemet kunna bestämma sig för att nödbromsa på 1,5 sekunder från det att något upptäcks till ett beslut är fattat och systemet bromsar.

Xilinx annonserade nyligen ett samarbete med Mercedes Benz om att använda avancerade FPGA:er och deep learning för analys av data från kamera, radar och lidar för att övervaka föraren, styra bilen och undvika kollisioner. Experter från de båda företagen implementerar AI-algoritmer på en mycket adaptiv fordonsplattform och ska optimera beräkningstekniken för deep learning till Mercedes neurala nätverk. Tekniken ger mycket korta svarstider samtidigt som den är energieffektiv vilket gör att systemet kan arbeta tillförlitligt inom de givna termiska ramarna för fordonsmiljön.

Även i datacenter används FPGA:er för att köra acceleratorer för deep-learning. De är betydligt bättre räknat i prestanda-per-watt än typiska GPU-lösningar. Operatören SK Telekom har på ett lyckat sätt förbättrat sin röstaktiverade assistent kallad Nugu med hjälp av Kintex Ultrascale som AI-accelerator i sina datacenter. Det här är den första implementationen av AI i den koreanska telekomindustrin och har förbättrat SK Telekoms automatiska röstigenkänningstillämpning med så mycket som 500 procent jämfört med konventionella grafikprocessorer. Dessutom har den totala kostnaden sjunkit genom att företaget kunnat addera AI-acceleratorerna till de existerande servrarna som saknar grafikprocessorer.

ETT ANNAT EXEMPEL är AI-baserade hemlarm, som en del av en molnlösning utvecklad tillsammans med Tend Insights, där snabb inferens ger smartare övervakning och innovativa tjänster. Kameror som placerats ut i hemmet har en grundläggande förmåga att identifiera bildrutor som innehåller händelser som kan vara intressanta. Dessa laddas upp till FPGA-bestyckade acceleratorer i molnet som nås via en uppsättning API:er (som finns I Xilinx maskininlärningssvit, ML Suite). Acceleratorerna kan larma genom att se skillnad mellan familjemedlemmar, husdjur, främlingar och främmande djur.

Om ägaren ger sitt samtycke kan bilderna också användas för att identifiera familjemedlemmar som har problem, exempelvis en äldre person som ramlat och inte kan ta sig upp själv. Systemet kan då meddela andra familjemedlemmar eller hemtjänsten.

Det finns många andra scenarios där AI används för att göra komplex mönsterigenkänning och bildanalys med korta svarstider. Det gäller exempelvis genanalys för att snabba upp diagnosticering av sjukdomar. Här använder man FPGA:er för att accelerera AI-inferenserna vilket redan kortat tiden från 24 timmar till 30 minuter för att sekvensera patientens gener och identifiera avvikelser. Arbete pågår med att korta tiden ytterligare.

Sen finns det atomforskning. Experiment med kärnfysik genererar extremt högupplösta bilder som lätt kan ha över hundra miljoner bildpunkter och måste analyseras på 25 ms. Det är en utmaning som kräver snabba neurala nätverk men vanlig processorbaserad inferens har inte en chans. De är en storleksordning för långsamma. Det är FPGA:er som hjälper vetenskapsmännen att få de svar de behöver.

Slutsats

AI har relativt nyligen blivit en användbar teknik men används trots det redan i tjänster som människor interagerar med regelbundet. Möjligheten att sänka driftskostnaderna, korta väntetiderna för kunderna och att hitta nya möjligheter att tjäna pengar är spännande för kommersiella företag vilket ökar kravet på att förbättra prestanda genom att korta svarstiderna, effektförbrukningen och kostnaden.

Det finns två aspekter när man använder AI. Det första är att träna ett neuralt nätverk av en typ som bäst passar uppgiften. Den andra är att beskära och optimera det tränade nätverket så att det kan implementeras som en inferensmotor på en lämplig processor.

Flexibiliteten och prestanda i FPGA-arkitekturer kombinerat med effektiva verktyg för att implementera och optimera, som kompilatorn i ML-sviten och DeePhi:s optimerare, med rekonfigurerbarheten som ger möjlighet att implementera det senaste neurala nätveksarkitekturen utan att behöva vänta på en ny krets, är de tre grundläggande ingredienserna för att realisera accelererad AI-inferens i molnet, i nätverket och i ändnoderna. ■