En jämförelse av AI acceleratorer

Oscar Björkgren

1 Inledning

1.1 AI och hårdvara

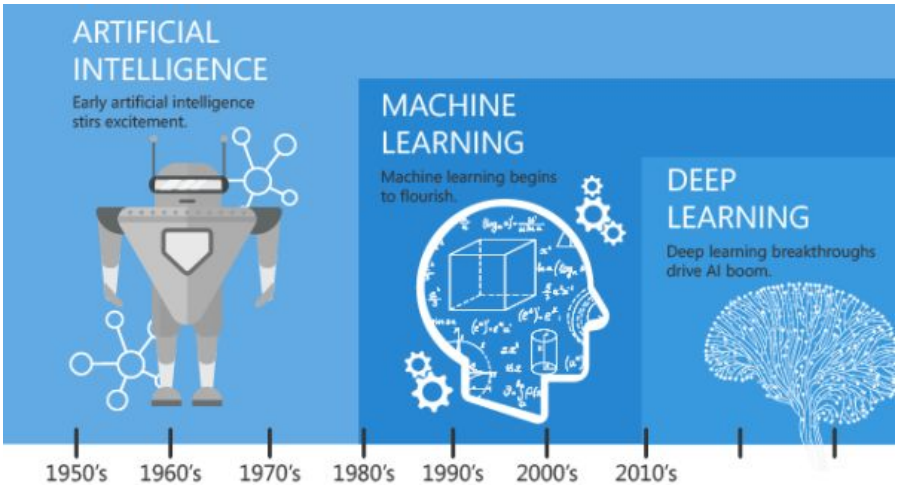
Hårdvara tillverkas med tanke på tillämpningsområde och enligt hur flexibelt den måste kunna fungera. Att komma fram till en optimal hårdvarulösning kräver kunskap både inom tillämpningsområdet och inom hårdvaruindustrin, men i dagens läge tillverkas det mycket hårdvara för testning och experimentering så att man inte behöver lösa hela problemet på en gång, utan steg för steg förbättra implementationen genom att t.ex. bygga eller byta komponenter på en mikrokrets. För att hitta en lämplig lösning för ett system som använder sig av AI spelar hårdvaran en stor roll, speciellt om prestanda är prioritet. I allmänna fall kan man experimentera med AI i generell hårdvara, men ifall systemet ska kunna ställa strikta krav på ett program, måste mjukvaran kunna samspela med hårdvaran.
Eftersom AI har funnits till som koncept sedan flera årtionden har det spridits brett utöver olika branscher och områden. Det som ändå till synes har framskridit mest inom teknologi är prestandan och storleken på hårdvara. Prestanda möjliggör snabbare problemlösning medan storlek kan påverka tillämpningens fysiska aspekt. Numera är det vanligt att AI-problem styrs från användaren, eller klienten, till en server som sköter beräkningen. Google har ett klart övergrepp då det är fråga om moln AI och erbjuder flera tjänster som använder sig av det, men erbjuder även verktyg för användare som själv vill bygga en tjänst med AI. Hårdvaran Google använder för sitt moln AI är deras egenutvecklade och designad för att lösa problem inom vissa ramar. Övriga teknologi jättar har också gjort en storsatsning på AI-specifik hårdvara. Speciellt inom telefon industrin har användningen av AI-chip marknadsförts och väckt mycket diskussion. Den hårdvara som produceras för telefon industrin bör ändå inte jämföras med moln AI-hårdvaran eftersom dessa två system i många fall arbetar tillsammans med varann. Det som eftersträvas med dessa AI-chip är i slutändan att minska beroendet av molnet och utföra beräkningarna så långt som möjligt självständigt. Denna konsumentinriktade

användning av AI har blivit självklar och är en del av vår vardag även om många av de nya teknikerna inte nått allmän kännedom.

1.2 Begrepp

Begreppet AI är ofta en generalisering av olika tekniker som med hjälp av vissa karakteristiska drag kan kopplas till ett och samma koncept. Denna avhandling kommer att så långt som möjligt skilja åt undergrupperna av AI, men omöjligen utesluta en viss grad av generalisering.

Maskininlärning är en undergrupp av AI där en algoritm tränas för att känna igen och behandla en viss typ av data. En vanlig tillämpning av maskininlärning är att träna algoritmen för att känna igen specifika former i bilder, och sedan tillämpa algoritmen för slutlednignar. Neuronnät är en form av maskininlärning var träningen sker genom att sprida algoritmen i neuroner. Varje enskild neuron utför en specifik beräkning som genom träning justeras för att nå det önskade resultatet.



För tillfället är dessa undergrupper av AI i stort fokus eftersom det sker mycket utveckling inom dem. Flera AI acceleratorer är specialiserade i att behandla dem på unikt sätt.

1.3 Accelerering av AI

Det centrala inom AI teknologin är algoritmerna och matematiken som bearbetar problemen. Algoritmerna består ofta av unika lösningar för varje problem, men matematiken som används är i hög grad av samma typ. I grunden är AI något som

används för att behandla information för att skapa en tolkning av det. Denna information kan bestå t.ex. av text, ljud eller bilder, men i grunden är de alla någon form av signaler eller data.
Kalkyl av vektorer och matriser används ofta inom matematiken för att representera och behandla information. Därför är också det beräkningsmässigt största problemet inom AI matematiken inom vektor- och matrisberäkningar. Genom effektivare lösning av dessa problem blir AI system generellt snabbare. AI acceleratorer använder sig därför oftast av processorarkitekturer som är specialiserade för dessa beräkningar.

2 AI-acceleratorer

AI har upprepade gånger varit bland de populäraste och mest omtalade forskningsämnen inom IT. Det som oftast varit i huvudrollen inom ämnet är algoritmerna och tillämpningen för systemet, vilket också är det mest lockande och intressanta för offentligheten. Nu har även hårdvaran kommit mera i fokus eftersom nya uppfinningar som t.ex. självkörande bilar och flygplan blivit rätt så vanliga. Utvecklingen har lett till att dagens mest omtalade form av AI består av olika former av maskininlärning. Tidigare former av maskininlärning har begränsats till en stor del av hårdvaran som funnits tillgänglig. Idag är algoritmerna inom maskininlärning rätt så igenkända, vilket gör att hårdvaran kommer att beaktas mera för att nå bättre prestanda inom systemet.
Universitetet MIT har nyligen erbjudit en kurs som behandlar hårdvaruarkitektur ämnat för AI. Lektorn för kursen, Vivienne Sze beskriver lärande ändamålet för kursen på följande vis ““How can you write algorithms that map well onto hardware so they can run faster? And how can you design hardware to better support the algorithm?”, vilket betyder att vi även efter en hårdvaru tillämpning för AI måste mjukvaran beaktas igen för att utveckla väl fungerande system.
I detta kapitel beskrivs tekniska delen hos hårdvara vars huvuduppgift är att beräkna problem inom maskininlärning, och generellt inom AI. Eftersom AI acceleratorerna har växt upp inom industrin, kommer även de mest insiktsfulla lösningarna att presenteras och jämföras. De centrala frågorna i undersökningen är hurudan den

tekniska lösningen är, vilket tillämpningsområde som lösningen riktar sig mot och hur tillgänglig lösningen är. Med dessa riktlinjer kommer jämförelsen att ge en täckande bild över vad som finns tillgängligt inom hårdvara för AI och hur det kan utnyttjas. Undersökningen fungerar också som grund för kapitel 3 var realtidskrav för AI acceleratorer tas i beaktande.

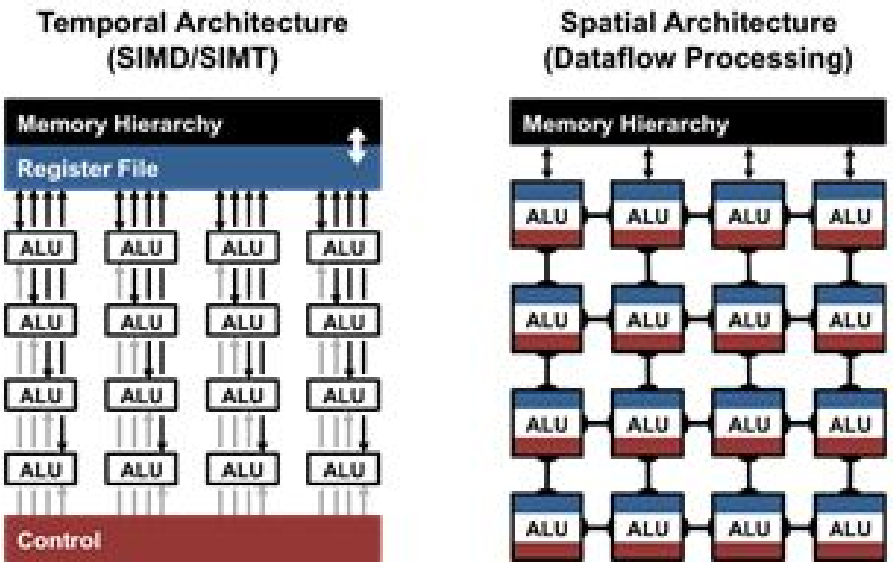
2.1 Teknik och tillämpning

För att förstå sig på egenskaperna av hårdvaran inom industrin presenteras vissa arkitekturer och principer som blivit populära. Hårdvaran kan variera mycket bland tillverkarna och accelereringen kan ske på olika nivåer vilket gör att de gemensamma dragen är få. Tillverkarna har trots allt som mål att utveckla den mest unika lösningen inom marknaden. Det som ändå binder samman olika lösningarna är väldigt fundamentalt för effektiv processering och hantering av data. Att förstå själva processeringen av data och vilka krav det ger upphov till är speciellt viktigt.

2.1.1 Paralell processering

Begreppet “multiply-accumulate operation” eller MAC, används inom digital signalbehandling för att beskriva en operation var man manipulerar en ackumulator med en beräkning, som inom AI ofta innebär matris- och vektorkalkyl. Dessa operationerna kan utföras och parallelliseras med hjälp av hårdvara som stöder MAC.
Parallellisering är en effektiv lösning i många fall där man utnyttjar möjligheten att dela beräkningar i flera delar och låta dem utföras separat. Man kan använda parallell processering inom olika områden och nivåer då man tillämpar det till problemlösning eftersom det omfattar allt mellan hårdvara och mjukvara. Det som används inom AI och huvudsakligen i maskininlärning är parallel processering på hårdvarunivå, där varje enhet måste utöver själva beräkningen kunna både läsa från- och skriva till minne. Enligt Sze et. al. kan man dela på processorarkitekturer som används till parallel processering inom maskininlärning i två grupper, temporär-

och rumslig arkitektur. Temporär arkitekturen framkommer inom vanligare sorters processorer så som bland CPU, och GPU enheter. Rumsliga arkitekturen framkommer i sin tur i mer specialiserad hårdvara så som i FPGA och ASIC kretsar. Dessa två arkitekturer kan användas för att accelerera maskininlärning på var sitt sätt med uppskattningsvis samma resultat.
Inom maskininlärning fås största nyttan av parallell processering i inlärningsskedet, då algoritmen justeras med hjälp av stora mängder träningsdata. Efter att algoritmen är färdigt tränad krävs det i vanliga fall inte så stor processeringskapacitet för att genomföra en slutledning. Eftersom tränings skedet är ofta tidskrävande och har stor inverkan till algoritmens funktion och prestanda är det en stor fördel ifall träningen är så högt paralelliserad som möligt.



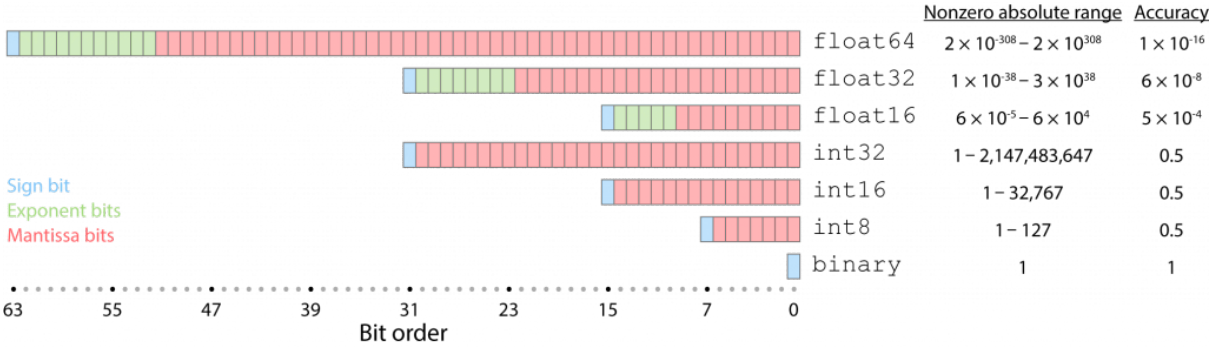
2.1.2 Minneshantering

Minneshantering är ett av flaskhalsarna i hårdvaran när man hanterar mycket data. Parallella processer kan öka problemet ytterligare ifall processerna använder aktivt minnesresurser. För varje MAC operation krävs typiskt tre läsningar från minne, och en skrivning till minne. Kommunikationen blir mycket tät då mängden parallella processer ökar och därför måste minneshanteringen vara effektiv. Detta ställer också ökade krav på bandbredden av minnesbussarna. I vanliga fall räcker inte

processorernas interna minne inte till och istället måste lagring och läsning skötas med hjälp av RAM. Eftersom läsning och skrivning till RAM minnet är kostsamt med tanke på både tid och energi måste det bildas minneshierarkier för att minimera dess användning. Minneshierarkier är olika minnes komponenter inom systemet var de är ordnade enligt responstid. Datat som används inom maskininlärning och speciellt neurala nätverk är ofta upprepande till en viss mån, eftersom alla noder i nätverket använder sig av vissa specifika värden i sina MAC operationer. I detta fall lagras det frekvent använda datat högt upp på minneshirerakin där responstiden är som snabbast.
Acceleratorer använder olika dataflödes processer för att utnyttja minneshierarkin på bästa möjliga sätt. Dataflödet bestämmer vilket minne ska lagra vilket data. Dataflödes processer kan tillämpas till olika behov, men i fallet där datat till en viss del är förutsägbar och upprepande så kan processen optimeras till en hög grad genom att minimera interaktionen med RAM minnet.

2.1.3 Ordlängd och aritmetik

Vanligaste ordlängden i processorer är 64- och 32-bitar med flyttalsaritmetik. Detta är en flexibel lösning till tillämpningar med varierande krav. Inom signalprocessering används både heltals och flyttalsaritmetik för att representera och beräkna signaler, som också blivit vanligt inom AI teknologin.
Ordlängden och aritmetiken påverkar det dynamiska området och precisionen i processorn vilket har direkt inverkan på prestandan. Det dynamiska området innebär skillnaden mellan det största och minsta talet som kan representeras med ett ord, medan precisionen innebär mängden av tal som ryms inom det dynamiska området. För tillämpningar med höga precisionskrav används därför flyttalsaritmetik eftersom ett flyttal kan representera decimaler med en upplösning som beror på ordlängden. Flyttalsoperationer är viktiga inom AI och maskininlärning eftersom det medför en högre precision i träningen av modeller. Heltals operationer är i sin tur snabbare och föredras därför inom slutledning i AI.



Flyttals- och heltalsaritmetik har båda nackdelar som kan uppstå i representation av reella tal. I processorer som använder flyttalsaritmetik är problemet oftast att tal inte ryms inom det dynamiska området dvs. talet är antingen för litet eller för stort med tanke på ordlängden. I processorer som använder heltalsaritmetik är det oftast problem med precisionen eftersom tal med stor noggrannhet måste bli avrundade.

Operation	Energy (pJ)	Area (µm^2)
`int8` addition	0.03	36
`int16` addition	0.05	67
`int32` addition	0.1	137
`float16` addition	0.4	1,360
`float32` addition	0.9	4,184
`int8` multiplication	0.2	282
`int32` multiplication	3.1	3,495
`float16` multiplication	1.1	1,640
`float32` multiplication	3.7	7,700

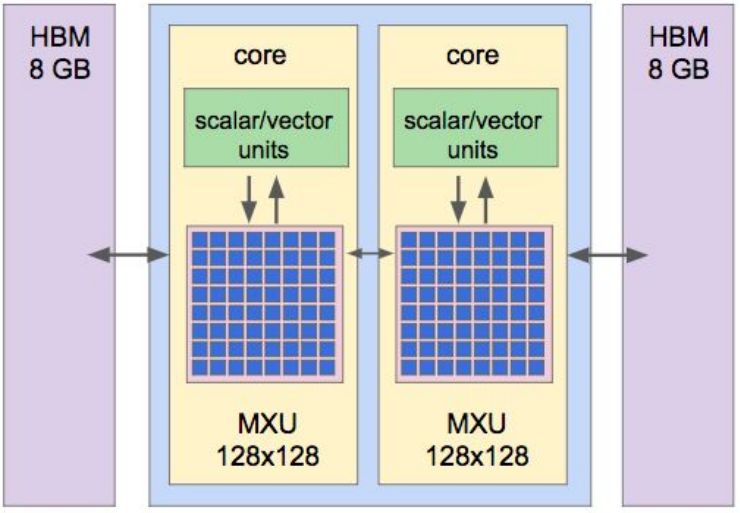
2.2 Google - Tensor Processing Unit

Googles tjänster har använt AI i stor utsträckning och mycket utveckling skedde inom teknologin innan företaget publicerade deras Tensor Processing Unit, eller TPU. Denna AI accelerator är avvikande med tanke på att hårdvaran är Googles egenutvecklade även om de inte är aktiva inom marknaden. Andra generationens TPU publicerades maj 2017 som bl.a. korrigerade bandbegränsnings problemet i minnet i första generationens acceleratorn. Google har varit begränsad i sina tekniska beskrivningar om deras acceleratorer, men i juni 2017 presenterades en analys över första generationens TPU.

Googles första TPU accelerator var specialiserad för att utföra slutledningar med hjälp av färdigt tränade modeller inom neurala nätverk och maskininlärning. Andra generationens TPU har utvidgats till att även träna modellerna som tidigare inom Google utfördes med hjälp av mer konventionella GPU baserade lösningar.

Googles acceleratorer är unika inom branschen eftersom de är baserade på ASIC kretsar, dvs. en “Application Specific Integrated Circuit”. Denna typens kretsar är tillverkade för att lösa problem inom en specifik domän, som i detta fall är matrismultiplikationer. Precisionen av beräkningarna är av lägre format för att öka prestandan. Första generationens TPU använder sig av 8-bitars heltals format, vilket har visat sig vara tillräckligt för utförande av slutledningar. Formatet har utöver precisionen en stor inverkan på elkonsumtionen och fysiska storleken av hårdvaran, vilket första generationens TPU har som fördel. Nackdelen för 8-bitars heltals formatet är att den inte har förmåga att konkurrera inom träning av maskininlärnigsmodeller. Precision är en avgörande faktor när det gäller träningen, vilket ofta sköts av processorer som använder sig av flyttalsformat.

Andra generationens TPU använder 32-bitars flyttals format som möjliggör både träning- och slutleding av neurala nätverk. Acceleratorn består av fyra processorer med en gemensam databehandlingskapacitet av 180 teraflops. I varje processor finns två kärnor och två 8GB minnen med 600GB/s bandbredd.



Acceleratorerna är en del av en datacenter för Googles AI resurser. De används i konfigurationer av 64, där varje TPU styrs av en CPU som representerar en huvudnod. Google har som syfte med dessa system att uppehålla tjänster och erbjuda resurser till utvecklare och företag.

2.3 Facebook - Big Basin

Facebook har aktivt forskat inom AI en längre tid, och 2017 offentliggjordes Big Basin som är efterträdare till Big Sur, deras första AI accelerator. Facebook är en

nykomling i AI accelerator marknaden och har utvecklat sina lösningar i hög grad baserat på Intels och Nvidias hårdvara. Facebook har stora ambitioner för sina AI lösningar och en omfattande mängd data tillgänglig vilket gör att företaget kommer med stor sannolikhet att fortsätta växa inom denna teknologi. Eftersom Facebooks verksamhetsområde är inom sociala medier är det osannolikt att deras hårdvara överlåts för offentligt bruk inom en molntjänst. Försäljning av hårdvaran är också osannolik eftersom den är utvecklad med hjälp av Open Compute Project organisationen som planerar och distribuerar hårdvarulösningar för datacenter produkter. Big Basins tekniska beskrivning var därför väldigt omfattande och tillgänglig.

Big Basin är specialiserad i att träna maskininlärningsmodeller för bild, video och text igenkänning. För att kunna flexibelt anpassa sig till olika tillämpningar bygger Big Basin på en modulär design. Processeringen sker med hjälp av CPU och GPU moduler, som ska kunna uppgraderas med tanke på kommande teknologier. Eftersom maskininlärning och server arkitektur har liknande krav på hårdvaran har Big Basin haft stor influens av Facebooks server infrastruktur. Detta möjliggör enkel integrering till back-end funktionalitet och nya server teknologier.

Big Basin är delad in i tre sektioner varav alla har modulära egenskaper. Sektionerna är till för att minska beroendet bland modulerna vilket leder till att CPU och GPU kretsarna är frånskilda och oberoende. Arkitekturen följer “just a bunch of disks” eller JBOD, var varje enhet är självständig. Denna arkitektur används oftast för hårddiskar men i Big Basin gäller arkitekturen för processor enheter. Big Basin innehåller typiskt 8 GPU som styrs av en extern server huvudnod. Huvud noden är varierande och kan anpassas enligt behov som t.ex. energikonsumtion eller överföringshastighet. Vanligtvis används en Intel Xeon processor i huvud noden.

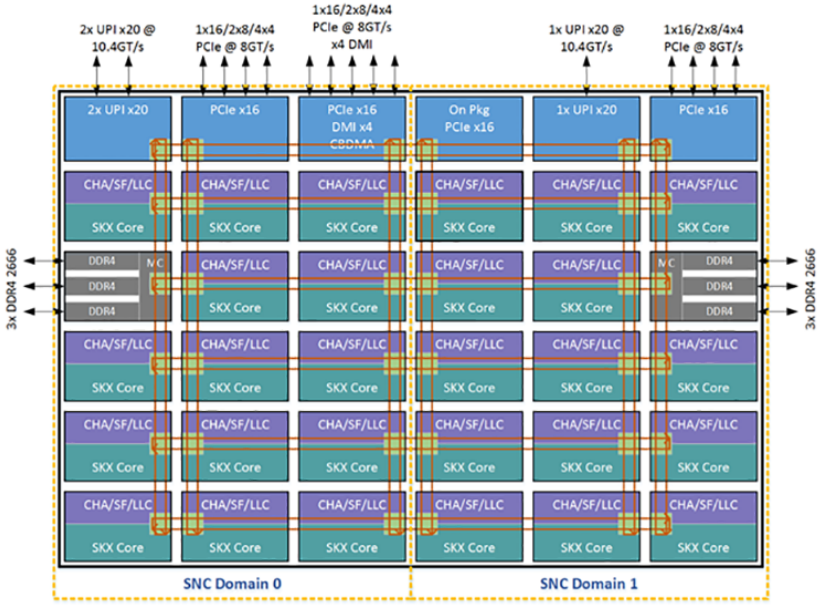
Som accelerator hårdvara använder Big Basin Nvidias nyaste AI accelerator Volta Tesla V100. Processorerna kopplas ihop med Nvidias hög bandbredsbuss NVLink som möjliggör 900Gb/s överföringshastighet.

Service	Resource	Training Frequency	Training Duration
News Feed	Dual-Socket CPUs	Daily	Many Hours
Facer	GPUs + Single-Socket CPUs	Every N Photos	Few Seconds
Lumos	GPUs	Multi-Monthly	Many Hours
Search	Vertical Dependent	Hourly	Few Hours
Language Translation	GPUs	Weekly	Days
Sigma	Dual-Socket CPUs	Sub-Daily	Few Hours
Speech Recognition	GPUs	Weekly	Many Hours

2.4 Intel

En stor del av processorer för konsumenter och företag tillverkas av Intel. Det stora utbudet och långa erfarenheten placerar Intel väldigt centralt inom hårdvaru branschen som har gjort att de också har en märkvärdig roll inom AI teknologin. Utöver egenutvecklad hårdvara erbjuder de produkter som tillverkats av företag som absorberats av Intel, och även produkter som de varit med och utveckla åt andra företag.

Intel Xeon familjens processorer används utöver server lösningar också inom AI. Enligt Intel är det den mångsidigaste produkten som de erbjuder för AI tillämpningar. Intel Xeon Scalable är den senaste produkten inom familjen och skiljer sig till en viss del från server processorerna. Denna nya generationens processor är utvecklad med stort fokus på överföringshastigheter och parallellisering har ökats för att t.ex. främja träningen av neurala nätverk inom maskininlärning. Processorn består av 28 kärnor som är



sammankopplade i ett nätverk. Nätverks arkitekturen är ett stort framsteg eftersom det möjliggör friare kommunikation inom och utanför processorn. Tidigare arkitekturer har bestått av kärnor som är sammankopplade i en ring där resursfördelningen per kärna varit mycket mindre. I nätverket är resurserna fördelade väldigt jämlikt med hjälp av att varje kärna har interna caching och minneskontroll egenskaper. Flera processorer kan samverka genom att utnyttja denna arkitektur och bilda större nätverk av kärnor och minne. Systemet kan därför göras väldigt skalbart och flexibelt i och med att konfigurationen kan modifieras enligt specifika krav. Kommunikationen mellan processorerna sker med en överföringshastighet på 10.4 GT/s.
Ordlängden i Xeon Scalable processorerna är 64-bitar och expanderar upp till 512 bit med hjälp av SIMD instruktioner i avancerade vektor extensions enheten. Dessa instruktioner innebär att en och samma instruktion utförs parallellt med en mängd data. Expansionen av ordlängden till 512 bitar innebär en fördubbling av databehandlingskapaciteten till 2 teraflops.
Varje Xeon Scalable processor kan kopplas till sex minnen med 2666 MT/s överföringshastighet. Låg nivå minnet har utvidgats inom Scalable processorerna så att kärnorna kan fungera mera självständigt.

Utöver CPU lösningar erbjuder Intel också mera specialiserad hårdvara för AI. Programmerbara grindmatriser (FPGA) används inom hårdvaruutveckling för tillämpnings specifika lösningar. Grindmatrisen programmeras till att behandla problem inom vissa ramar väldigt effektivt och är därför lämpliga för AI. Denna flexibilitet är värdefull eftersom samma krets kan då återanvändas i olika tillämpningar. Ytterligare fördelar med jämförelse till CPU baserade acceleratorer är energieffektivitet och ökad parallel processering.
En FPGA krets består i allmänhet av en mängd enheter som kan sammankopplas. Dessa enheter består av minne, logiska enheter och digital signalprocesserings enheter.
Stratix 10 och Arria 10 är Intels kraftigaste av FPGA enheter och tillverkas av Altera, som tillhör Intel sedan 2015. Arkitekturen inom dessa produkter följer en modern uppläggning som innehåller utöver själva programmerbara grindmatrisen också en

mikroprocessor och minne. För att maximera prestandan för AI acceleratorer kan tillämpningen ändå kräva både externt minne och CPU som huvudnod. Detta underlättas av att enheterna är kompatibla med Xeon Scalable processorn och högbandbreddsminne upp till 1TB/s. Nätverks infrastrukturen i kretsen möjliggör sändtagar kommunikation upp till 58 GB/s med en bandbredd av 8 TB/s vilket möjliggör stort dataflöde. Intel Stratix 10 presterar upp till 9.2 teraflops med 32-bitars flyttals format. Ordlängden kan varieras enligt behov inom FPGA kretsar vilket gör systemet användbart till både träning och slutledning av modeller inom maskin inlärning.

Intel Nervana och Movidius är företagets hårdvaruplatformer som är tillverkade från början till slut med tanke på AI. Båda produkterna härstammater från skilda företag som 2016 blev en del av Intel. Nervana och Movidius har olika tillämpningsområden och arkitektur vilket gör Intels utbud av AI specifik hårdvara bred och varierande. Nervana är en neuronnäts processor (NNP) i form av en ASIC, och fungerar som ett alternativ åt GPU baserade neuronnäts lösningar. Utvecklingen av hårdvaran tog 3 år och introducerar marknaden med nya lösningar för precision och numeriska operationer. Nervana använder en 16+5-bitars heltalsaritmetik som Intel kallar för Flexpoint format. Det innebär att för varje 16-bitars operation finns det 5-bitar med data som innehåller en exponent som höjer precisionen av operationerna. Jämfört med 16-bitars flyttalsoperationer har Nervanas Flexpoint format en högre precision och dynamisk område, samtidigt som heltalsaritmetiken medför lägre energikonsumtion och högre prestanda inom slutledning. Precisionen inom träning av neuronnät har bevisats vara jämförbart med den som en processor med 32-bitars flyttalsaritmetik åstadkommer.

ASIC kretsen är i dett fall i högt beroende av huvudnoden eftersom utöver instruktioner och data måste den också updatera exponenten för att hålla operationerna inom ramar som bestäms av ordlängden. För att hindra problem med exponenten och ordlängden, används Autoflex algoritmen för att justera exponenten enligt tidigare data. Detta innebär att systemet har hög komplexitet med nya tekniker på kretsen. Forskning inom neuronnät träning med låg precisions processorer är i ett tidigt stadie och utvecklingen av Nervana är pågår. Systemets databehandlingskapacitet är uppskattningsvis 55 teraflops och har stöd för

högbandbreddsminne med 1TB/s bandbredd. Kretsen implementerar rumslig arkitektur där varje processerings element har över 2 MB minne som ytterligare ökar minneshanteringens effektivitet.

Intel Movidius är så kallde “System-on-a-chip” (SoC) kretsar som tillämpas i uppgifter inom datorseende. Kretsarna är utvecklade för slutledning av neuronnäts modeller, vilket innebär att träningen måste utföras externt. Enheterna är tillverkade med tanke på låg energikonsumtion och portabilitet, i och med att de kan användas genom en USB port med en storlek av en generell minnessticka. Processeringen i kretsarna sköts av CPU och programmerbara vektor processor enheter (VPU). Vektor processorer är specialiserade i vektor aritmetik och fungerar ofta med hjälp av SIMD instruktioner vilket gör processorn typen populär inom signalbehandling och AI. Arkitekturen i Movidius VPU processorerna kallas för “Streaming Hybrid Architecture Vector Engine” (SHAVE). Processorerna har flexibel ordlängd och aritmetik, och utför MAC operationer. Movidius Myraid X är den senaste enheten och innehåller 16 SHAVE kärnor och 2 Leon CPU med stöd för ordlängd mellan 8- och 128-bitars heltal. Totala databehandlingskapaciteten är 4 tops och 1 tops inom slutledningsoperationer. Kretsen har 2.5 MB SRAM minne med 450GB/s bandbredd, och har stöd för internt och externt DDR4 DRAM. Upp till 8 sensorer med HD upplösning kan kopplas till kretsen med hjälp av en 16 filer bred MIPI interface. Detta interface används vanligen i LCD paneler och definierar kommunikationsprotokollet och databussen. Behandlignskapaciteten av sensor datat är 700 miljoner pixlar i sekunden.

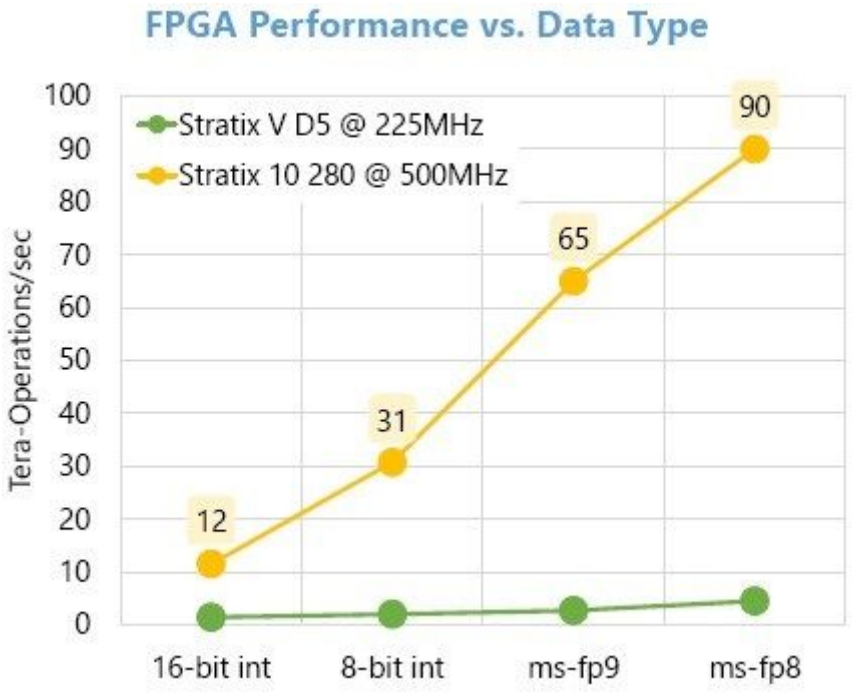
Systemet är utvecklar vänligt i och med att det finns ramverk och dokumentation för hårdvaru modifikationer som kan tillämpas av Assembler, C, och C++ kod.

Intels stora satsningar inom AI teknologi har gett konsumenter och företag tillgång till specialiserad hårdvara. Produkterna är tillverkade för att användaren ska kunna välja tillämpningsområdet fritt utan beroende av externa tjänster som t.ex. en molnplattform. Produkterna stöder olika ramverk för utveckling och Intel erbjuder också egna moln verktyg för användning av deras Xeon Scalable baserade datormoln.

2.5 Microsoft - Project Brainwave

Microsoft förespråkar tillämpning av AI inom företag och organisationer som nödvändigtvis inte är väldigt tekniska. Deras Azure platform erbjuder en mängd tjänster var utvecklare kan utnyttja datormolnet för olika sorters problemlösning. Det finns färdiga algoritmer för t.ex. kognitiva tillämpningar, men också stöd för egenutvecklade lösningar.

Hårdvaran för AI tjänsterna har bestått av en mängd tredje partens produkter, men i Augusti 2017 publicerades Project Brainwave som Microsoft utvecklat tillsammans med Intel. Brainwave är tillverkad för maskininlärning och är specialiserad för att utföra slutledning i neuronnät i realtid. Träningen av neuronnät modellerna utförs på andra enheter eftersom realtids aspekten medför vissa egenskaper som gör Brainwave mindre optimal för det. Acceleratorn används i Microsofts datormoln tillsammans med CPU huvudnoder. Arkitekturen i datacenter omgivningen är sådan att accelerator enheterna är sammankopplade i ett nätverk var de kan användas och hanteras även självständigt. Detta kan utnyttjas när en maskininlärnings modell inte ryms på endast en krets och kan då för att spara tid dela upp modellen och arbetet inom nätverket. Brainwave enheterna använder Intel Stratix 10 FPGA för processering och därmed kan inlärnings modellerna sparas in i de logiska enheterna var beräkningarna sker. Detta innebär att modellerna måste programmeras i varje FPGA i nätverket innan användning, men resultatet blir ett system som kan behandla stora mängder data för realtidskrav.

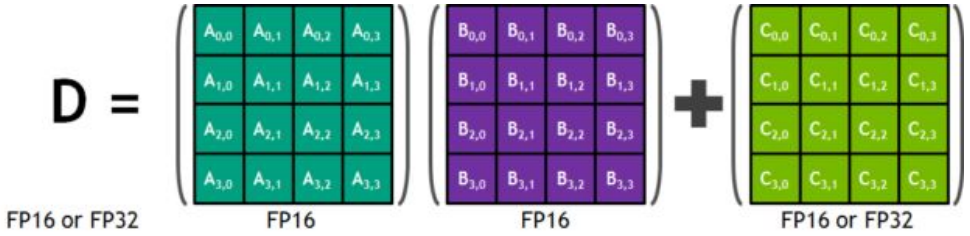


Acceleratorerna har två 40GB/s ethernet portar som är används för huvudnod och omkopplare till andra acceleratorer. Huvudnoder är anslutna till acceleratorerna också med PCIe bus.
Brainwave utnyttjar Stratix 10 för att utföra matrisberäkningar med flyttalsaritmetik som Microsoft själv utvecklat. Det baserar sig på 16-bitars flyttal som omvandlas till något som Microsoft kallar MS-FP. Utanför matris beräknings enheterna används vanliga 16-bitars flyttal. Omvandlingen av talen ökar komplexiteten men har en stor inverkan på prestandan. Med en SIMD instruktion åstadkommer Stratix 10 med MS-FP9 formatet 130000 flyttalsoperationer per klock cycel, och den totala databehandlingskapacitetetn är 90 teraops med MS-FP8 formatet. Inom lågnivå minnet i Stratix 10 är bandbredden 20TB/s. Även om Stratix 10 har kapacitet för ytterligare bandredd, är det nuvarande tillräckligt för att systemt ska fungera i realtid. Flexibiliteten i Stratix 10 och distribuerade realtids arkitekturen är egenskaper som gör Brainwave konkurrenskraftigt inom marknaden. Acceleratorn kan användas via Microsoft Azure, var en redan tillämpas inom Bing sökmotorns tjänster.

2.6 Nvidia

Grafikkortets betydelse och användning har blivit bredare i takt med utvecklingen. Nvidia är stark inom marknaden och har etablerat sig även inom AI teknologin. Grafikkort är specialiserade i parallell processering vilket gör det naturligt att tillämpa dem inom AI. De är ofta en effektivare lösning jämfört med CPU lösningar, som istället ofta har prestanda på sin sida. Nvidias AI acceleratorer används bland annat brett inom bil industrin och i datacenter. Utöver själva processorerna har företaget också utvecklat arkitektur för matris beräkningar och produkter inom dataöverföring som också avancerat framgången inom AI.
Nvidia Tesla V100 är företagets ledande accelerator och används främst till träning och slutledning av neuronnät inom maskininlärning. Acceleratorn använder Volta

GV100 GPU som processor. Den är indelad i sex GPU kluster, som tillsammans bildar 84 multiprocessorer för strömning (SM). Inom varje SM finns kärnor för varierande aritmetik och ordlängd, och 8 tensor kärnor. Teknologin i tensor kärnorna



gör produkten unik och ökar prestandan avsevärt jämfört med tidigare versioner. Kärnorna utför MAC operationer med 4x4 matriser i en blandning av 16- och 32-bitars flyttal. Matris multiplikationen sker alltid med 16-bitars flyttal, medan ackumulatorn och resultatet kan variera. Beräkningarna förblir i 4x4 matriser fastän precisionen ökar till 32-bitar, vilket resulterar i fler flyttalsoperationer per klockcykel. Processorn innehåller 640 Tensor kärnor, och tillsammans har de en databehandlingskapacitet på 125 teraflops, medan den totala databehandlingskapaciteten för konventionella 32-bitars flyttal är 15.7 teraflops.
Tesla V100 har stöd för 16GB högbandbreddsminne med 900GB/s bandbredd, och kopplas till andra enheter med NVlink teknologin som möjliggör en bandbredd upp till 300 GB/s. Nvidia anser att PCIe bussarna som vanligtvis används inom processor kommunikation är en flaskhals i systemet, som sedan gett upphov till NVlink. Utöver att bussen används för kommunicering mellan GPU används de också till CPU och minne. Med hjälp av NVlink ska prestandan öka med 31% jämfört med samma konfiguration med PCIe. Eftersom PCIe är en standard inom industrin är problemet att kunna utnyttja hastigheten som NVlink har kapacitet till. För tillfället är IBM den enda tillverkaren som stöder NVlink i två av sina CPU produkter.
Nvidia har utvecklat Tesla V100 och NVlink för att användas i datacenter omgivning där de kan bilda större helheter. Med DGX-2 systemet kopplas 16 Tesla V100 enheter med varann och ger upphov till 2 petaflops prestanda. Systemet har 512GB högbandbreddsminne och styrs av 2 Intel Xeon CPU. Eftersom Intel Xeon inte stöder NVlink, används PCIe bussen för kommunikation mellan accelerator enheterna och huvudnoderna. Alla accelerator enheter är kopplade till varann genom NVswitch

omkopplaren, som är en ASIC krets tillverkad för GPU till GPU kommunikation med NVlink. Det totala genomflödet av data blir med hjälp av NVswitch 2.4 TB/s.
Nvidias acceleratorer kan användas med deras GPU cloud platform, eller genom att köpa produkterna via återförsäljare.

2.7 Jämförelse

Acceleratorer och hårdvara i allmänhet har en tillämpning och målgrupp. Denna undersökning har behandlat acceleratorer som används av sina målgrupper genom tre sätt.

Facebook och Google använder sina acceleratorer huvudsakligen internt, och därmed utvecklar de hårdvaran för att få nytta genom användning. Företagen har båda gemensamt att de har tillgång till stora mängder data som kan användas till att skapa värde med hjälp av AI. Eftersom TPU 2 och Big basin är respektive företagens andra generations accelerator finns det ett tydligt behov för denna sortens tillgång. Google tog sin utveckling ett steg längre när de bestämde sig att utveckla egna ASIC kretsar. Facebook tog däremot en lättare utväg i och med att de använder sig av allmänt tillgängliga lösningar.

Molnbaserade datortjänster ebjuder avändare tillgång till hårdvara som kan vara kostsamt och komplicerat att tillämpa internt. Utav de undersökta acceleratorerna är det Microsoft och Google som är de mest välkända inom denna teknologi. Även om Microsoft också använder deras hårdvara internt är deras Azure platform av större vikt eftersom det är anslutet till deras mjukvaruutvecklings ramverk som används väldigt brett. Project Brainwave används genom Azure som genererar vinst genom olika medlemskaps abonnemang.
Google erbjuder tillgång till sina TPU system genom Google cloud platformen. Användningen av platformen har underlättats genom att de öppnat källkoden till deras eget utvecklingsverktyg som heter Tensorflow. Detta ramverk används inom

maskininlärning, och stöds även av Azure platformen. Liksom Microsoft erbjuder också Google sina tjänster genom medlemskaps abonnemang.

Intel och Nvidia är hårdvarutillverkare som får nytta av AI teknologin genom försäljning av deras acceleratorer. Båda företagen erbjuder också molnbaserade datortjänster och utvecklingsramverk som underlättar utvecklare att introduceras till produkterna. Utöver detta har företagen väldigt lite gemensamt i deras produktutbud eftersom Nvidia erbjuder GPU baserade lösningar medan Intel erbjuder CPU och FPGA baserade lösningar. Nvidias GPU acceleratorer tillverkas med för maximal prestanda och effektivitet inom en smal domän. Målgruppen består i huvudsak av företag med höga prestationskrav vilket har resulterat i ett stadigt fotfäste inom bilindustrin. Intel har ett bredare utbud av acceleratorer för olika syften som t.ex. Movidius VPU enheten som är den enda produkten riktad åt konsumenter.

3 Avslutning/diskussion

4 Källor

http://www.fpa-trends.com/article/ai-machine-learning-impact-fpa
Artificial Intelligence: A Modern Approach (3rd Edition)
http://news.mit.edu/2017/building-hardware-next-generation-artificial-intelligence-1201
https://www.youtube.com/watch?v=J-GOkwiwg4c
http://www.rle.mit.edu/eems/wp-content/uploads/2017/11/2017_pieee_dnn.pdf
https://www.nextplatform.com/2017/07/14/technology-requirements-deep-machine-learning/
https://ai.intel.com/flexpoint-numerical-innovation-underlying-intel-nervana-neural-network-processor/
https://cloud.google.com/tpu/
https://drive.google.com/file/d/0Bx4hafXDDq2EMzRNcy1vSUxtcEk/view
http://learningsys.org/nips17/assets/slides/dean-nips17.pdf
https://research.fb.com/publications/applied-machine-learning-at-facebook-a-datacenter-infrastructure-perspective/
https://software.intel.com/en-us/articles/intel-xeon-processor-scalable-family-technical-overview

https://www.microway.com/knowledge-center-articles/detailed-specifications-of-the-skylake-sp-intel-xeon-processor-scalable-family-cpus/
https://www.nextplatform.com/2017/03/21/can-fpgas-beat-gpus-accelerating-next-generation-deep-learning/
https://www.altera.com/solutions/technology/transceiver/overview.html
https://ai.intel.com/intel-nervana-neural-network-processor-architecture-update/
https://ai.intel.com/flexpoint-numerical-innovation-underlying-intel-nervana-neural-network-processor/
https://www.nextplatform.com/2016/08/08/deep-learning-chip-upstart-set-take-gpus-task/
https://newsroom.intel.com/news/intel-unveils-neural-compute-engine-movidius-myriad-x-vpu-unleash-ai-edge/
https://www.nextplatform.com/2017/08/24/drilling-microsofts-brainwave-soft-deep-leaning-chip/
http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf
https://www.nextplatform.com/2018/03/27/nvidia-memory-switch-welds-together-massive-virtual-gpu/