Helium får Cortex-M85

Av Eldar Sido, Renesas Electronics

Eldar Sido arbetar med produktmarknadsföring av Arms styrkretsar med särskilt fokus på implementationer av AI.

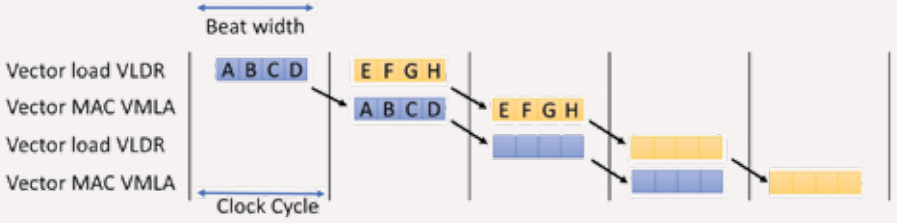

Bild 1. Beatwise-exekvering enligt Helium.

I dagens tekniklandskap finns en snabbt ökande efterfrågan på avancerad maskininlärning (ML) och digital signalbehandling (DSP) specifikt för inbyggda system med begränsad processorkraft. Dagens inbyggnadsprocessorer är en svår utmaning för dig som försöker möta en kravspecifikation genom att trycka in mer beräkningar per klockcykel. Du kanske väljer att lösa problemet genom att byta till en kraftfullare processor, som en Cortex-A? Båda valen innebär höga kostnader och långa utvecklingstider.

Räddaren i nöden kan vara Arm Cortex-M85 – den senaste styrkretskärnan från Arm. Den är inte bara designad för att klara maskininlärning (TinyML) när beräkningskraft och minne är begränsat, utan även avancerad signalbehandling.

Den använder 32-bitarsarkitekturen Armv8.1-M men dessutom har den Arms så kallade Helium-teknik vilket ger en betydande höjning av ML- och DSP-prestanda. Den har flyttalsenhet (FPU) och en omfattande instruktionsuppsättning för digital signalbehandling. Sammantaget betyder det att CM85-kärnan enkelt kan hantera komplexa utmaningar inom maskininlärning i TinyML.

MVE Helium är ett tillägg till Armv8.1-M för CM85- och CM55-kärnor. Liksom tillägget Neon i Cortex-A är Helium ett tillägg av SIMD-typ (single instruction multiple data) och den kan ersätta separata DSP-kärnor.

Istället för att bara portera Neon från Cortex-A utvecklades Helium från noll, med hänsyn tagen till den mindre chipstorleken. Den fick även stöd för datatyper och instruktioner som inte stöds på Neon, som low overhead branching och predication (mer om det senare).

Helium kan avsevärt förbättra prestanda för vanliga ML- och DSP-kärnor. Den åtgärdar typiska flaskhalsar genom att addera funktioner som:

Överlappande pipeline: Helium består av åtta vektorregister (som är återanvända 128-bitars FPU-register) där varje fjärdedel kallas en ”beat” (”taktslag”) och alltså har en bredd på 32 bitar. För att öka beräkningseffektiviteten utnyttjar Helium så kallad vektorkedjning genom att exekvera överlappande pipelines parallellt. För varje klockcykel exekverar överlappande pipelines en MAC-operation på tidigare laddat beat, medan nästa beat laddas. Cortex-M85 har en 64 bitar bred dataväg som är dual-beat, det vill säga att den kan exekvera två cykler per 128 bitar. En sådan överlappande process är känd som ”beatwise”-exekvering.

Varje beat kan delas upp i ännu mindre delar och flera dataposter av olika datatyp kan exekveras under samma klockcykel:

- två stycken Q31/int32
- fyra stycken Q15/int16
- åtta stycken Q7/int8
- två stycken fp32
- fyra stycken fp16

Datatyper: För att CM85 ska kunna användas i många sorters tillämpningar stöds många datatyper:

- Vektorer med 8-bitars heltal/fixpunktstal, som är vanliga i kvantiserade ML-modeller och normalt inte stöds av enklare DSP:er.
- Vektorer med 16-bitars heltal/fixpunktstal
- Vektorer med 32-bitars heltal/fixpunktstal
- Vektorer med 16-bitars flyttal med halv precision som används vid förbearbetning av realtidsdata, exempelvis i sensortillämpningar för att bibehålla ett stort dynamiskt område, och samtidigt halvera beräkningskraven. Detta är unikt för CM-kärnor med M8.1-arkitekturen.
- Vektorer med 32-bitars flyttal i enkel precision.

Helium har också stöd för beräkningar på komplexa tal, både heltal och flyttal, vilket används i fouriertransformer (FFT) och annan signalbehandling.

Förbättrad grenprediktion/loop-optimering: Som nämnts tidigare adderar Helium instruktioner för att accelerera DSP-beräkningar. Ett exempel är ”low-overhead branch extension” som cache-lagrar första och sista instruktionen och vid efterföljande iterationer endast exekverar loopkroppen.

”Lane predication” är en annan användbar instruktion som stöder villkorlig exekvering och hanterar specialfall som att antalet vektorbanor inte är delbart med fyra. Dessa instruktioner ger samma prestandaförbättring som en DSP med stöd för zero-overhead-loop.

Instruktioner för förbättrad minnesaccess: Extrainstruktionerna innefattar bland annat load och store med interleaving och deinterleaving vilket lyfter prestanda, särskilt med bilddata (exempelvis RGB) och ljuddata. Andra instruktioner gör adressering med scatter/gather för bit-reverserad adressgenerering för emulering av den typ av cirkulär adressering man hittar i en DSP och som är typisk för FFT och annan signalbehandling.

CM85 stöder även tätt kopplat minne (TCM) för goda realtidsresponser och har en AXI-buss för applikationer med högre minnesbandbredd, samt cache som optimerar prestanda i långsammare, icke-deterministiska applikationer.

UTÖVER ALLA DE FÖRBÄTTRINGAR som Helium MVE ger, så bidrar CM85 – den vassaste av alla Cortex M-kärnor – med en mängd imponerande funktioner.

Cortex-M85-kärnan innehåller dessutom nya och förbättrade säkerhetsfunktioner som inte finns i andra Cortex-M-kärnor, som pekarautentisering, branch target identification (PABTI), förbättringar av Arm TrustZone, eXecute Never (PXN) och debugtillägget

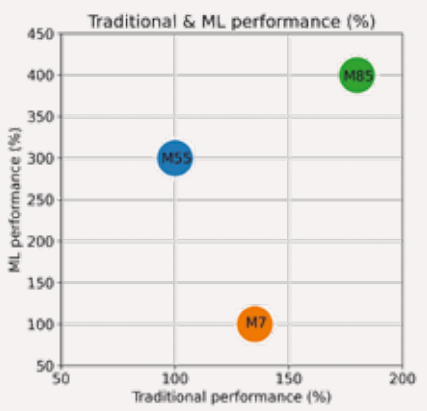

Bild 2. Prestandalyft för CM85 jämfört med CM7 och CM55. KÄLLA: ARM

att sväva

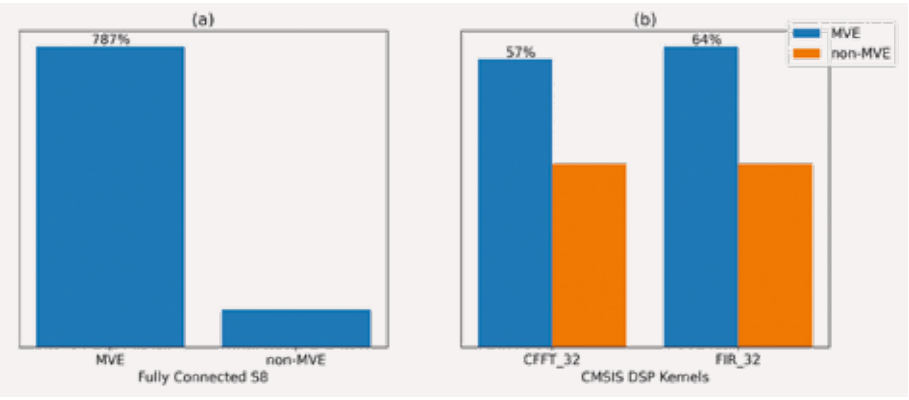

Bild 3. Benchmark för MVE relativt enheter utan MVE. (a) CMSIS-NN med Arm-kompilatorn AC6.15 – genomsnitt av mätningar över ett fullt kopplat lager. (b) CMSIS-FFT & FIR med AC6.16 med prestandan normaliserad. KÄLLA: ARM

Bild 4. Benchmark med Arms 2D-bibliotek visar att CM85 är fyra gånger bättre än CM7. KÄLLA: ARM

DUE. Det gör det enklare att klara PSA-certifiering på nivå 2. Den är därmed idealisk för applikationer som industriella styrsystem och medicinsk utrustning.

MED TANKE PÅ ALLA de funktioner som har packats in i CM85, är det inte överraskande att det finns empiriska bevis – diverse prestandatester, inklusive industristandarder – som befäster CM85:s ställning som den vassaste styrkretskärnan av dem alla.

Se exempelvis resultaten i CoreMark och DMIPS. De visar en linjär skalbarhet. Det kan vi tacka den nya mikroarkitekturen för med dess förbättrade minne, förbättrade grenprediktion och optimerade dual-issue. Integrationen av Helium gör att CM85 har fyra gånger högre prestanda än CM7 på AI/ML-beräkningar, och även tydligt klår CM55 (se bild 2).

Helium är en klar förbättring jämfört med enheter som inte använder Helium. På granulär nivå kan MVE ytterligare förbättra prestanda för individuella ML-kärnor i jämförelse med icke-MVE-enheter, som synes i bild 3a, som visar ett rejält lyft vid beräkningar på ett neuronnät med fullt kopplade skikt (fully-connected layers).

När det gäller DSP-prestanda för MVE jämfört med icke-MVE-enheter på standardkärnor, exempelvis FFT och finit impulssvar, visar bild 3b en höjning på 57 procent respektive 64 procent för flyttal. Höjningen skulle vara ännu mer signifikant för heltal på grund av MVE:s bredare inbyggda stöd för olika datatyper.

Genom de hundratals nya instruktioner som adderats till CM85, inklusive för bildbehandling, får den nästan fyra gånger bättre prestanda än CM7 på Arms 2D-bibliotek.

Till Embedded World i Nürnberg i våras hade Renesas, som är Arm:s ledande partner på CM85-kärnan, tagit fram två demonstrationer, en allmän för kameror och en för upptäckt av motorfel.

Den förstnämnda baseras på mjukvara från företaget PlumerAI som klarar av att identifiera, detektera och räkna människor från olika vinklar, under varierande ljusförhållanden och i olika miljöer. Inferenserna har fått ett lyft jämfört med att köra dem på andra Cortex-M-kärnor vilket öppnar för ännu mer avancerade användarfall som tidigare krävde en Cortex-A, exempelvis att följa personer (tracking).

Den andra demonstrationen handlade om feldetektering i motorstyrning, som kan användas i många industriella tillämpningar. Genom att titta på shuntströmmen kan modellen förutsäga hur stort felet är i upplinjeringen av motorn (genom att applicera kraft på motorkortet) och presentera resultatet med Renesas användargränssnitt.

Eftersom AI-modellen till övervägande del är baserad på faltningar var förbättringen betydligt högre i detta fall. Det visar att CM85 med Helium MVE kan användas för att ersätta lösningar med en styrkrets med DSP-kärna i låg- eller mellanklassen vad gäller prestanda.

Bild 6. Att prediktera motorfel med CM85 – en demonstration. Med samma CNN-modell hade CM85 5,39 gånger högre prestanda än CM7. För en FFT i fixpunktstal var den 3 gånger högre och för FFT i flyttal 2,07 gånger.

Bild 5. Demonstrationen på Embedded World som räknade personer. Prestandaförbättringen var 3,7 gånger jämfört med en CM7-kärna som körde samma AI-program.

SAMMANFATTNINGSVIS kan CM85 med Helium bidra till en betydande prestandahöjning för AI/ML- och DSP-uppgifter genom att den förbättrar grenprediktion, minnesåtkomst och parallellitet och lägger till nya DSP-instruktioner, nytt datatypstöd och har inbyggt stöd för komplexa tal. CM85-kärnan är också ett kraftpaket i skalär prestanda och överglänser resten av Cortex-M-kärnorna, vilket gör den till det idealiska valet för mer komplexa beräkningsuppgifter. ■