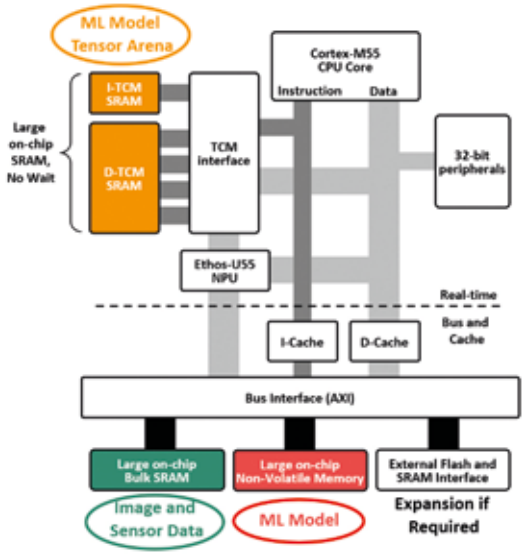
Så förvandlas en MCU till en fullvärdig AI-SoC

Artificiell intelligens kan höja värdet i slutanvändarprodukter betydligt. Ta vården som exempel. Ett instrument som patienten kan bära med sig kan med hjälp av AI upptäcka och diagnosticera allvarliga tillstånd, som förmaksflimmer. En avancerad diagnos kräver alltså inte ett sjukhusbesök utan kan ställas i patientens vardag. En hörapparat kan förvandlas från en simpel förstärkare till ett intelligent röstfilter som isolerar den du vill lyssna på, medan annat ljud dämpas eller tonas ner.

Exemplen handlar om AI i ändpunkter. AI kan höja värdet hos nästan alla typer av bärbara, batteridrivna produkter. I ändpunkter vill vi dock vanligen inte göra AI-beräkningarna i molnet. Det faller på strömförbrukning, latens, personlig integritet, trådlös räckvidd, säkerhet och kostnader. Produkter av det slaget behöver AI-kapaciteten integrerad direkt i sin egen hårdvara.

Men detta kräver att produkterna lyckas ta sig över svåra hinder gällande design och strömförbrukning. Bärbar teknik – öronsnäckor, ringar, smarta glasögon och patientmonitorer – använder små formfaktorer som rymmer bara ett fåtal komponenter och ett litet batteri.

Innan AI började dyka upp i kravbilderna kunde dessa typer av produkter – typiskt i ett större format och med enklare funktionalitet – ofta använda en vanlig mikrokontroller (MCU) eller mikroprocessor (MPU) för att implementera sin huvudfunktioner. Genom att integrera elektronikkomponenterna kunde konstruktörerna möta sina kravbilder kring utrymme och strömförbrukning, och reducera antalet komponenter och kretskortets fotavtryck.



Figur 1. Benchmarks visar den överlägsna prestandan och effektiviteten hos en NPU jämfört med en CPU, för några typiska ML-funktioner.

Av Mark Rootz, Alif Semiconductor

Mark Rootz är idag marknadschef på Alif Semiconductor och har tidigare haft liknande positioner på ST, Freescale och Renesas. Karriären startade med systemteknik för uppkopplad avionik och styrsystem för drivlinor. På halvedarområdet är idag hans fokus på styrkretsar, processorer och konnetivitet. Han är expert på inbyggda system, AI/ML, strömsnål trådlös kommunikation och extremt strömsnåla halvledare.

I AI-eran har utmaningen att integrera systemlösningen i en standard-MCU blivit ännu större. En MCU är fortfarande en attraktiv lösning för att spara plats och energi. Men en AI-MCU måste integrera ännu fler funktioner och fortsätta hålla strömförbrukningen minimal för att kunna stödja produkter med små batterier utan att offra driftstid och göra laddpauserna för täta.

En klassisk MCU-arkitektur har sina begränsningar när det gäller AI. Det går inte bara att ta den som den är och addera en extra AI-funktion. Vi grundade Alif Semiconductor 2019 med målet att ta fram en ny generation AI-kompetenta MCU:er för ändenheter. Vi har haft fördelen att kunna tänka om från grunden kring hur man ska integrera AI i en MCU. Våra idéer är ett samlat eko av hundratals diskussioner med OEM-företag som beskrivit faktorer som påverkar valet av MCU som AI-SoC för batteridrivna produkter. Här är fyra av de viktigaste läxorna de gav oss.

Måste vara tätt integrerad med CPU:n

MCU-marknadens första svar på AI-efterfrågan var mjukvaruutvecklingsverktyg (SDK:er) som lät dig köra AI och maskininlärning på samma Cortex M-CPU som hanterade systemets styrfunktioner. Men maskininlärning (ML) för ändenheter kräver i slutänden en egen processorenhet för att accelerera neuronnät, en så kallad NPU, optimerad för de matrismultiplikationer som implementerar neuronnätets inferenser. En vanlig CPU har inte tillräcklig kapacitet eftersom den måste utföra det parallella nätets inferenser sekventiellt. Det tar för lång tid och drar för mycket energi.

Figur 1 visar skillnaden i AI-prestanda mellan CPU- och NPU-kärnan i en mikrokontroller. Alif Semiconductors MCU-serie Ensemble använder en aktuell CPU-kärna, Arm Cortex M55, med en NPU som coprocessor, Arm Ethos U55. Det som mäts i tabellen är en inferens i fyra olika ML-modeller som körs på en Ensemble MCU. ML-modellerna gör nyckelordsdetektering, objektdetektering, bildklassificering och taligenkänning.

Cortex M55 är en inbyggnads-CPU av senaste snitt och presterar redan i sig fem gånger bättre på ML-beräkningar än tidigare Cortex M-generationer. Men hur bra M55 än är, blir prestandalyftet tydligt i de gula kolumnerna i diagrammet: upp till tvåhundra gånger bättre för kombinationen NPU och CPU jämfört med enbart en CPU. Om man dessutom tar hänsyn till att Cortex M55 redan presterar fem gånger bättre än äldre Cortex M-arkitekturer, betyder det rimligen ytterligare en multiplikation med fem i prestanda. För taligenkänning skulle det innebära upp till 800 gånger kortare tid och 400 gånger mindre energi per inferens jämfört med äldre Cortex M-CPU:er.

En annan viktig faktor för tätt kopplade NPU:er och CPU:er är mjukvaruutvecklingsmiljön. Det finns många alternativa NPU-kärnor att välja mellan för din SoC. OEM-företag är dock tydliga med att de inte vill behöva bygga om hela sin infrastruktur med nya verktygskedjor och nya instruktioner för att anpassa sig till en alternativ ML-arkitektur.

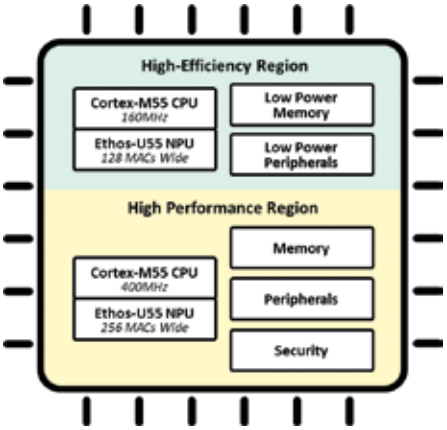
Den som redan använder Arm-ekosystemet vill fortsätta göra det även för AI och ML. Det kravet uppfylls genom användandet av Arms Ethos U-NPU.

Ethos U utgör faktiskt en sömlös coprocessor till Cortex M. Arms kompilator Vela delar automatiskt upp ML-arbetsbelastningen mellan dem. Typiskt läggs 95 procent eller mer på NPU:n. Som bonus kan Cortex M-CPU:n gå i viloläge eller utföra andra uppgifter medan ML-inferensen görs.

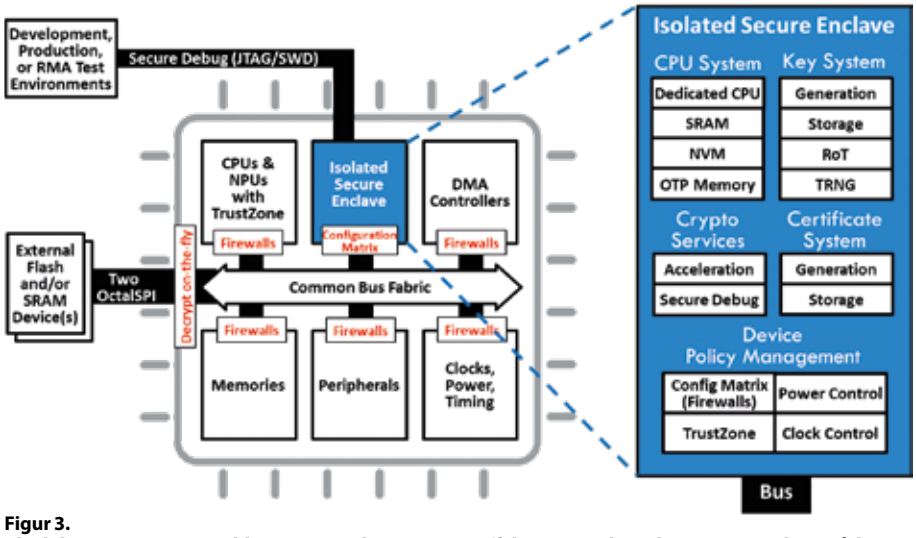
Integration måste omfatta hela systemet

NPU:n är förstås centrum för allas uppmärksamhet i en AI–ML-MCU. Men vad som integreras runt processorkärnorna, och hur integrationen görs, är avgörande. Högst på checklistan står minne och kringutrustning.

Figur 1 visar att processorkapacitet är en nyckel till prestanda och effektivitet. Men



Figur 2.
Ensemble-MCU:ns interna minnestopologi.



Figur 3.
Blockdiagram över Ensemble E3. Du ser chipets energieffektiva respektive högpresterande områden.

utan ett optimerat minnessystem som uppbackning kommer resultaten ändå inte att leva upp till förväntningarna.

En förenklad vy av minnestopologin i Ensemble visas i figur 2. Den övre halvan representerar realtidssektionen med extremt snabbt TCM (Tightly Coupled Memory) kopplat direkt till CPU- och NPU-kärnorna. För snabba inferenser måste TCM vara tillräckligt stort för att rymma tensorarenan – en buffert för ML-modellens datastrukturer, som kallas tensorer.

Den nedre delen av diagrammet visar övriga systemminnen inkopplade via en gemensam höghastighetsbuss. Ett stort delat SRAM används för att hantera sensordata, som indata från kamera och mikrofon. Ett stort icke-flyktigt minne innehåller själva ML-modellen och applikationskoden. När stora inbyggda minnen används distribuerat på detta sätt minimeras antalet kollisioner på databussen, vilket betyder att samtidiga minnestransaktioner kan ske utan problem och att flaskhalsar försvinner. Dessutom förkortas minnesåtkomsttiderna, och energiförbrukningen stannar på en nivå som kan hanteras av ett litet batteri.

Rätt uppsättning periferienheter är avgörande för en MCU i en ML-ändenhet. De arbetar ofta med de tre V:na: vision, voice, vibration. Detta kräver bildsensorer, mikrofoner, tröghetsmätare och mer, utöver standardanslutningar som seriella gränssnitt, analoga gränssnitt och bildskärmar.

En AI-ändenhet vill ha alla dessa funktioner integrerade i MCU:n.

Med hela systemet integrerat försvinner behovet av ytterligare strömförsörjning och omvandling (i till exempel en PMIC). Dessutom går det att göra mer finkornig dynamisk strömstyrning inuti chipet, vilket är den tredje önskade egenskapen hos en integrerad AI-MCU.

Adaptiv styrning förlänger batteritiden

På Alif insåg vi tidigt att lokala AI-beräkningar i ändenheter stod på tröskeln till att explodera i användning. Samtidigt minskar produkternas fysiska storlek snabbt – särskilt för bärbara enheter – vilket innebär att de måste drivas av allt mindre batterier.

Alif tog till flera metoder för att förlänga batteritiden. Två tydliga exempel:

1. Att dela upp systemet så att en del av chipet alltid är aktiv och tillhandahåller stabila beräkningsresurser – men i låg effekt. Lågeffektsdelen väcker vid behov upp högpresterande delar av chipet, som sedan kan återgå till viloläge.
2. Strömstyrningen aktiverar dynamiskt de delar av chipet som behövs, och stänger av dem när de inte längre används, allt på en finkornig nivå.

För att underlätta funktionsuppdelningen har många av Ensemble-MCU:erna dubbla uppsättningar Cortex M55- och Ethos U55-kärnor, som syns i figur 3:

- Det ena paret ansvarar för lågeffektsdelen. Det är implementerat i transistorer med låga läckströmmar. Det är alltid aktivt och tickar i upp till 160 MHz.
- Det andra paret sköter prestandadelen och klockas i upp till 400 MHz.

För att se fördelen med detta kan vi tänka på en smart övervakningskamera. Det energieffektiva kärnparet skannar kontinuerligt av ett rum i låg bildfrekvens och spanar efter intressanta händelser (som att en människa faller omkull eller gör en specifik gest). Händelsen väcker prestandaparet som exempelvis kan identifiera personen, kontrollera om utgångar är blockerade, ringa efter hjälp och så vidare.

Scenariots kamera är vaksam på ett intelligent sätt, som ger färre falsklarm och förlänger batteritiden. Man kan tänka sig liknande användningar inom vitt skilda domäner för CPU–NPU-par ac detta slag, för klassificering av ljud, röster, ord, text och sensordata.

Alla Ensemble-MCU:er använder Alifs teknik aiPM (autonomous intelligent power management) för att i realtid styra upp till 12 individuella strömdomäner i chipet efter deras aktuella arbetsuppgifter. Endast domäner som aktivt utför uppgifter är påslagna (exempelvis domäner som matar specifika processorkärnor, minnen eller kringutrustning) medan övriga domäner förblir avstängda. Allt hanteras transparent för mjukvaruutvecklaren.

Skydd för maskininlärningsmodeller

Den sista nyckelfunktionen i en AI-MCU för ändpunkter är cybersäkerhet. Detta är nödvändigt för att stå emot de olika cyberattacker som ständigt pågår. För många OEM-företag är det ännu viktigare att kunna skydda den egna IP som finns i AI-modellerna.

OEM-företag investerar stora mängder tid och pengar i att samla in träningsdata, bygga AI-modeller och utveckla och förbättra inferensalgoritmer. Detta ger oseriösa tillverkare starka incitament att försöka stjäla dyrbar IP genom att kopiera den från otillräckligt skyddade produkter.

Med hjälp av en extern säkerhets-MCU kan OEM-tillverkaren etablera root-of-trust, hantera nycklar och certifikat, säkra uppstart med mera. En extern säkerhets-MCU är en vanlig metod för att bygga in stark säkerhet i konventionella MCU-baserade konstruktioner. Däremot är det ovanligt att hitta en konventionell MCU med integrerad säkerhetsenklav med samma funktioner.

Att integrera funktionaliteten direkt i MCU:n ger dock plats- och energibesparingar – och ökad säkerhet – som särskilt bärbara, batteridrivna AI-produkter kan ha nytta av. En säkerhetsenklav är standard i alla Alif-enheter. Den utgör ett dedikerat, isolerat delsystem för hantering av viktiga säkerhetsfunktioner som säkrad nyckelhantering och lagring, säkrad start med omodifierbar root-of-trust, attestering via certifikat under körning, hårdvarukryptering, säkrad debugging, lässkydd, säkrad firmwareuppdatering och komplett livscykelhantering.

En AI-förberedd MCU-plattform

Fyra egenskaper hos en AI-MCU – tät koppling mellan NPU och CPU med standardutvecklingsverktyg, systemövergripande integrering, adaptiv strömhantering och inbyggt IP-skydd – är starkt efterfrågade av de tillverkare av batteridrivna ändpunktsenheter som Alif samarbetat med.

Konstruktörer som utvärderar Ensembleserien hittar ett brett urval skalbara, kompatibla enheter, från enkla CPU-kärnor till fyrkärnor som stöder Linux. Det gör att det går att anpassa sig till olika projekt och samtidigt återanvända mjukvara mellan dem. ■