Generativ AI ställer nya

Stora språkmodeller (LLM:er) och tjänster baserade på dessa som ChatGPT och Gemini är lysande exempel på kraften i generativ AI. Men dessa AI-modeller består av enorma kodbaser – i början av 2025 har de största mer än en biljon parametrar.

Jättelika datacenter med de mest avancerade serverkorten är nödvändiga för att tillhandahålla den beräkningskraft och de resurser som generativ AI kräver. Detta väcker frågan: hur arbetar tillverkare av inbyggda system på kanten eller i ändnoder för att skala generativ AI så att den kan fungera på deras mycket mer begränsade hårdvaruresurser?

Faktum är att tillverkarna redan arbetar med lösningar på detta pussel. Några av de tidiga erfarenheterna visar att både hårdvara och mjukvara i ändnoderna behöver anpassas för generativ AI. Arkitekturen i de styrkretsar som används är inte tillräcklig för att implementera generativ AI. Nya modeller optimerade för begränsade resurser behöver utvecklas, vilket ger vissa liknande funktioner som molnbaserad AI, men på ett annat sätt.

Alla de mjukvarufunktioner som vi kallar generativ AI är attraktiva för inbyggnadsvärlden eftersom de resulterar i system som är mer autonoma på ett intelligent sätt.

Det som kännetecknar generativa AI-system är förmågan att ”minnas” och därmed sätta nya indata i kontexten av tidigare data.

DET ÄR DETTA SOM MÖJLIGGÖR:

- Förståelse av naturligt språk och textgenerering
- Implementering av långa kommandosekvenser
- Intelligent respons på indata från flera sensorer, exempelvis en kombination av ljud, video och text

I EN KONSUMENTPRODUKT som smarta glasögon kan generativ AI användas för realtidsöversättning av utländsk text i ett skyltfönster eller på en vägskylt. Inom sektorer som medicinteknik, tillverkning eller transport ser företagen stora möjligheter med generativ AI i människa–maskin-gränssnittet – till exempel genom att införa agentliknande funktioner, eller genom att lära sig användarens beteende och självständigt fatta beslut utan att följa en förprogrammerad meny av svar.

I många av dessa fall kommer lokal AI-processning att vara avgörande på grund av den korta fördröjningen – användarna accepterar inte den fördröjning som uppstår med molnbaserade lösningar. Molnlagring av generativ AI-data är också en växande utmaning: antalet installerade IoT-enheter förväntas nå 50 miljarder år 2030, och datamängden väntas överstiga 300 zettabyte. Både kostnaden och energibehovet för att lagra den ständigt växande massan av generativ AI-data i molnet är betydande.

Av Henrik Flodell, Alif Semiconductor

Henrik Flodell är uppvuxen i Sverige och läste datavetenskap på Umeå universitet, men har bott i Bay Area sedan 2009. På Alif Semiconductor ansvarar han inte bara för partnerekosystemet utan även för produktfamiljerna Ensemble och Crescendo. Han inledde karriären i början av 90-talet med att utveckla firmware för medicinska dataloggrar.

Av dessa skäl designas produkterna för att utföra större delen, eller hela, AI-bearbetningen lokalt.

Men hur ska ett system som smarta glasögon kunna utföra uppgifter som realtidsöversättning när LLM-programvaran har ett minnesavtryck som mäts i terabyte? Även med en vanlig skalningsmetod som kvantisering är det otänkbart att dessa modeller skulle kunna krympas till mindre än flera gigabyte – och det är fortfarande en enorm beräkningsbörda för de flesta inbyggda system, än mindre för smarta glasögon.

Inbyggda system kan inte använda en LLM alls, utan andra modeller som är bättre anpassade till begränsade hårdvaruresurser. Den enda realistiska kandidaten för att utföra AI-beräkningar och styrning av systemet är styrkretsen. Den är ensam om att kunna uppfylla effekt-, storleks-, integrations- och kostnadskraven för ändnoderna.

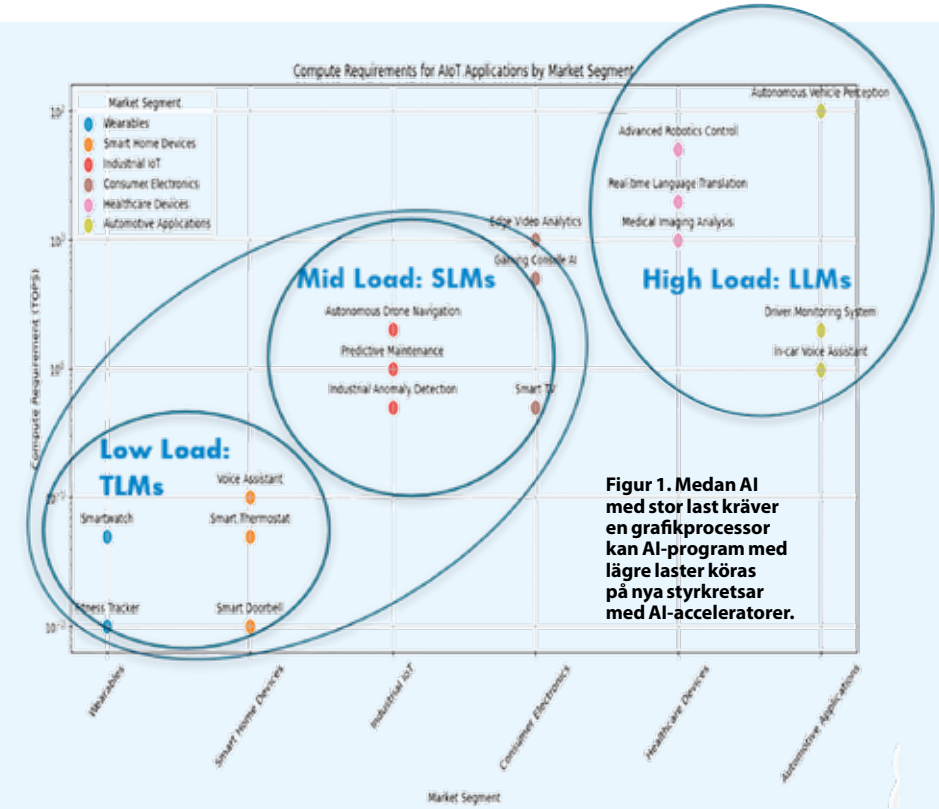
Och för mikroprocessorbaserade produkter hittar tillverkarna en balanspunkt för generativ AI genom att använda små språkmodeller (SLM:er) samt konvolutionella neurala nätverk (CNN:er) och rekurrenta neurala nätverk (RNN:er) som förstärkts med generativa AI-element. Med andra ord kommer generativ AI i ändnoden inte att implementeras med nedskalade versioner av de modeller som körs i molnet, utan med nya modeller som är optimerade för hårdvaran i inbyggda system.

Vilka krav ställer då dessa generativa AI-modeller, optimerade för ändnoder, på styrkretsen?

Dagens mest AI-kapabla styrkretsar utför i stor utsträckning operationer baserade på röst, video och rörelse, såsom ansiktsigenkänning, nyckelordsigenkänning och tillståndsbaserad övervakning av fabriksutrustning. De bästa av dessa styrkretsar klarar upp till några hundra gigaoperationer per sekund (GOPS).

Övergången till generativ AI i ändnoderna kommer att medföra att efterfrågan på rå neuralnätverkskapacitet stiger till så mycket som 10 teraoperationer per sekund (TOPS) år 2030. Detta kräver styrkretsarkitekturer som kombinerar CPU:er med neurala processorer (NPU:er). För att utföra generativ AI kommer det att behövas nya NPU:er som kan utföra de transformeroperationer som generativa AI-algoritmer är beroende av.

Men när man utvärderar hårdvaran kan man inte enbart fokusera på rå genom-



Figur 1. Medan AI med stor last kräver en grafikprocessor kan AI-program med lägre laster köras på nya styrkretsar med AI-acceleratorer.

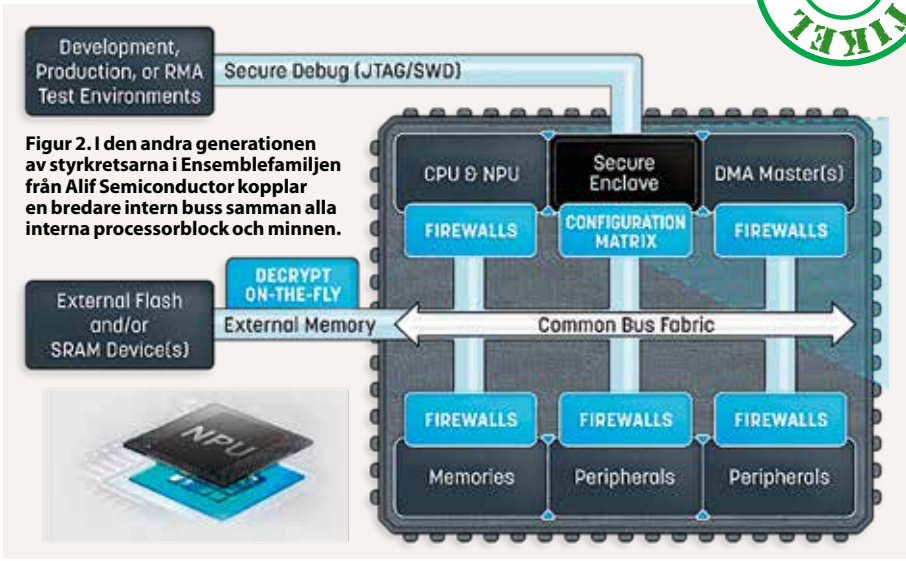
krav på styrkretsen

strömning: andra egenskaper hos arkitekturen avgör om den kan köra generativa AI-modeller eller inte.

Minneskapacitet – behovet av mycket snabb åtkomst till data är större för generativ AI än för andra typer av AI, som i sig har ett betydligt större minnesavtryck än de realtidsstyrfunktioner som konventionella styrkretsar är designade för att stödja. Interna minnesåtkomster är i grunden snabbare än externa, så vid val av styrkrets bör man lägga särskild vikt vid storlek och hastighet på det interna minnet.

Även med förbättrad intern minneskapacitet kommer många applikationer med generativ AI också att behöva ett externt minne: här är minnesgränssnittets hastighet en avgörande parameter för att minska latensen.

För att få hög prestanda med generativ-AI behöver styrkretsen koordinera olika operationer som utförs i olika funktionsblock. Dessa inkluderar inte bara neuronnätsprocessorn och CPU:n utan också hjälpprocessorer såsom en hårdvarubaserad bildsignalprocessor (ISP) för att bearbeta och förbehandla bilder innan de levereras till en neural nätverksalgoritm.



Figur 2. I den andra generationen av styrkretsarna i Ensemblefamiljen från Alif Semiconductor kopplar en bredare intern buss samman alla interna processorblock och minnen.

DENNA BLANDNING av operationer innan ett inferensresultat produceras kräver en friktionsfri förflyttning av data inom systemet och kräver generös intern bandbredd på en buss som alla funktionsblock som deltar i AI-operationer är anslutna till.

Det ligger i AI-applikationers natur, inklusive generativ AI, att en dataström kontinuerligt skannas efter relevans i ett bakgrundsläge, medan kraftfull inferenshårdvara endast används när relevant data hittas.

En styrkretsarkitektur som speglar denna dubbla natur hos generativ AI kan köra övervakningen i strömsnål och långsammare hårdvara och reservera ett högpresterande, mer kraftkrävande block för användning bara när ett snabbt och exakt inferensresultat behövs.

En strömsnål arkitektur gör det möjligt att implementera generativa AI-funktioner även i produkter som smarta glasögon eller helt trådlösa hörsnäckor som endast har plats för ett mycket litet och lätt batteri.

Hög effektivitet minskar också systemets värmeutveckling vilket hjälper konstruktören att eliminera risken för hotspots, något som är oförenligt med produkter som hörsnäckor och smarta glasögon.

Produkter som kan dra nytta av generativ AI kommer alltid att vara komplexa system. Smarta glasögon, till exempel, kan behöva integrera kameror, mikrofoner, högtalare, en display, ett batteri och mer men samtidigt vara lätta, bekväma och visuellt tilltalande. Detta innebär att komponentantalet måste minskas för att reducera systemets formfaktor. Det kräver att styrkretsen integrerar så många funktioner som möjligt som behövs för generativ AI – inte bara CPU och NPU, utan även stödjande funktioner som ISP och snabbt minne.

Företag som kan implementera generativ AI i sina produkter kommer ha immateriella rättigheter (IP) av stort värde i sina produkter. Dessa måste skyddas mot potentiella konkurrenter. Generativa AI-system som använder bilder och tal är dessutom föremål för integritetsfrågor.

Av dessa skäl är säkerhetsfunktioner en grundläggande del av ett generativt AI-system. Det är att föredra att säkerhetsfunktionerna är integrerade i styrkretsen för att förhindra att hemligheter exponeras på kretskortet och för att eliminera behovet av ytterligare säkerhetskomponenter på kortet.

Av de skäl som beskrivits ovan kan äldre styrkretsarkitekturer inte användas för att implementera generativ AI: även om en tillräckligt kraftfull NPU adderas kommer den att sakna den minneskapacitet, intern bandbredd, stöd för övervakning av dataströmmen, integration av AI-funktioner och de säkerhetsfunktioner som krävs för generativ AI i ändnoder.

Det är därför nya arkitekturer, kapabla att hantera SLM:er och andra modeller optimerade för ändnoder, dyker upp. De ger verkligt utrymme att stödja de spännande nya möjligheterna att använda generativa AI-algoritmer på video-, ljud- och rörelsedata i produkter som drivs med batteri och har mycket begränsat med plats och energi. ■