EXPERTARTIKEL

Hjärnan autonoma

Kretsen ger robotar simultankapacitet och sänker strömmen

Av Shingo Kojima, Renesas Electronics

Shingo Kojima är senior sakkunnig inom marknadsföring av mikroprocessorprodukter på Renesas, just nu med starkt fokus på artificiell intelligens. Han har drygt 35 års erfarenhet inom produktplanering, LSI-utveckling och utveckling av lösningar inom controllers och processorer. På sin första arbetsgivare, NEC, jobbade han med generella inbyggnadsprocessorer.

Renesas mikroprocessor RZ/V2H kan hantera realtidsstyrning och datorseende parallellt och har effektivt stöd för beskurna neuronnät.

En av konsekvenserna av de just nu sjunkande födelsetalen – andelen arbetande minskar och andelen äldre ökar – är att behovet av artificiell intelligens ökar.

AI kommer behövas i fabriker, inom logistik och sjukvård. Två exempel är servicerobotar och säkerhetskameror. Det kommer att användas för rörelsestyrning, för beslutsplanering, för navigation, med mera. Olika systemlösningar kommer att behöva köra avancerade AI-algoritmer i realtid för olika tillämpningar.

För att kunna reagera i realtid på förändringar i omgivningen kommer AI-systemen att behöva vara fysiskt integrerade i sina produkter. Strömförbrukningen måste vara minimal och det kommer att finnas gränser för värmeutveckling som inte får överskridas.

För att kunna möta AI-marknaden har Renesas lagt några år på utvecklingen av en accelerator med namnet DRP-AI3 (Dynamically Reconfigurable Processor for AI3). Den gör AI-inferenser och lyckas med konststycket att kombinera låg effekt med den flexibilitet som edge-enheter kräver.

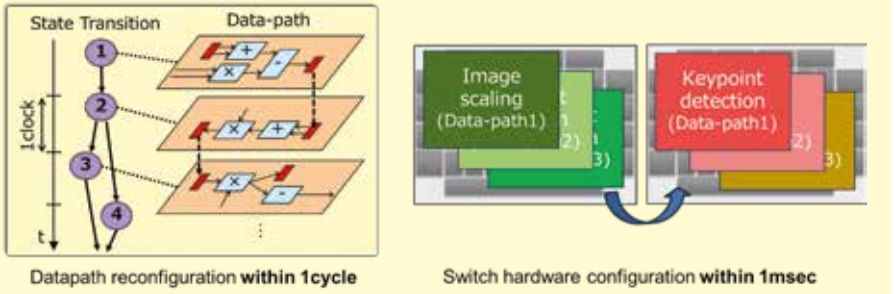

Bild 1. Så blir DRP flexibel.

Accelerator har integrerats i mikroprocessorserien RZ/V och dess nya toppmodell RZ/V2H.

RZ/V2H har en energieffektivitet som är cirka tio gånger högre än tidigare produkter i familjen. Den kommer att kunna klara den fortsatta utvecklingen av AI och i synnerhet robottillämpningarnas krav kring värmegenerering, beräkningsprestanda, realtid och strömförbrukning.

DRP-AI3 arbetar effektivt med beskurna AI-modeller

En vanlig teknik för att trimma AI-prestanda är att beskära neuronnäten, det vill säga att klippa bort delnät ur AI-modellen som gör liten skillnad för klassificeringarna.

Sådana delnät kan vara lokaliserade i det närmaste slumpmässigt i en AI-modell vilket skapar ett prestandaproblem eftersom det inte matchar så bra mot hur hårdvarans parallellitet fungerar.

För att adressera detta problem har Renesas optimerat nämnda accelerator DRP-AI för beskärning.

Vi analyserade mönstren för hur beskärningar ser ut givet olika beskärningsmetoder, och hur de påverkar klassificeringsnoggrannheten för CNN:er (Convolutional

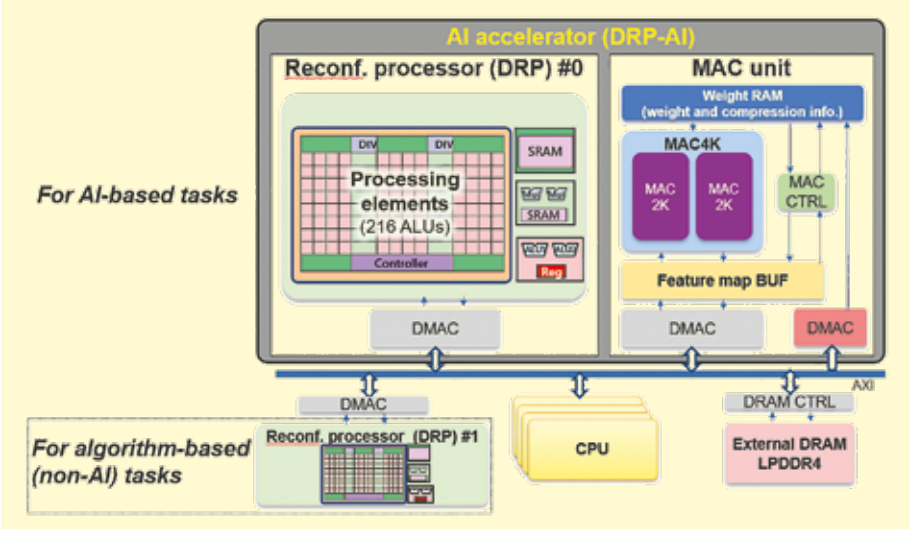

Bild 2. En konfigurering av en heterogen DRP-AI3-baserad arkitektur.

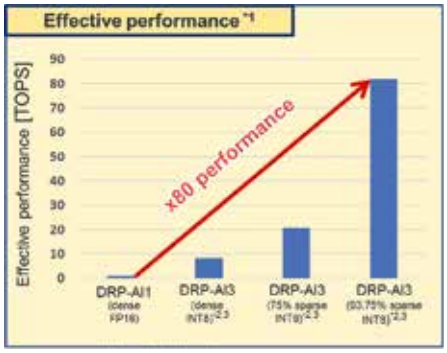

Bild 3. Jämförelse av uppmätt topprestanda för DRP-AI3.

*1) Effective performance evaluating test layer (single 3x3 conv.), Batchsize=1
*2) Difference between board power difference between DRP-AI active and DRP-AI inactive.
*3) Performance and power may change depends on AI models.

robotar längtat efter

Ceural Nets) och andra typiska AI-modeller för bildigenkänning.

Därefter lyckades vi designa en hårdvarustruktur för en AI-accelerator som både kan uppnå hög klassificeringsnoggrannhet och välja en effektiv beskärningsgrad. Den designen använde vi sedan i DRP-AI3.

Dessutom utvecklades mjukvara som minskar fotavtrycket på AI-modeller optimerade för DRP-AI3.

Mjukvaran snabbar upp AI-beräkningarna genom att konvertera en konfiguration med slumpmässigt utspridda beskärningar till högeffektiva parallella beräkningar. Detta resulterar i snabbare AI-bearbetning.

Renesas teknik för beskärning (flexibel N:M-beskärning) kan dynamiskt styra antalet beräkningscykler för att möta förändringar i graden av lokala beskärningar. Detta gör det möjligt att justera beskärningsgraden efter användarens önskemål om energiförbrukning, arbetstempo och klassificeringsnoggrannhet.

EN HETEROGEN ARKITEKTUR där DRP-AI3, DRP och CPU samarbetar:

- Flertrådad pipeline-bearbetning med hjälp av AI-accelerator (DRP-AI3), DRP och CPU
- Robottillämpningar i hög hastighet och lågt jitter med hjälp av DRP (dynamically reconfigurable wired logic hardware)

BLAND ANNAT SERVICEROBOTAR kräver avancerad AI-databehandling för att tolka sin omedelbara omgivning.

Å andra sidan kräver de även klassisk algoritmbaserad databehandling, utan AI, för att planera och styra robotens agerande.

Problemet är att dagens inbyggnadsprocessorer inte har de resurser som krävs för att göra dessa olika typer av databehandling i realtid.

Denna brist har Renesas avhjälpt genom att utveckla en heterogen arkitektur med en DRP (dynamiskt rekonfigurerbar processor) som koordinerar sitt arbete med en AI-accelerator (DRP-AI3) och en CPU.

Som visas i bild 1 kan DRP-processorn exekvera tillämpningar samtidigt som den varje klockcykel dynamiskt växlar konfiguration på chipet för hur kretsens aritmetikenheter är kopplade – allt efter vilken typ av innehåll som bearbetas.

Eftersom endast de aritmetiska kretsar används som behövs för tillfället, förbrukar DRP:n mindre ström än en CPU och kan göra beräkningarna i högre hastighet.

Jämfört med en CPU, där prestandan blir lidande av återkommande externa minnesaccesser på grund av cache-missar och annat, kan DRP dessutom sätta upp datavägar i hårdvaran i förväg. Det resulterar i mindre prestandaförluster och mindre variation i drifthastighet (jitter) från minnesaccesser.

DRP:n kan omkonfigureras dynamiskt och byta ut informationen om kretskopplingarna varje gång algoritmen förändras. Därför kan databehandling ske även med begränsade hårdvaruresurser, även i robotar som kör flera algoritmer parallellt.

▶▶

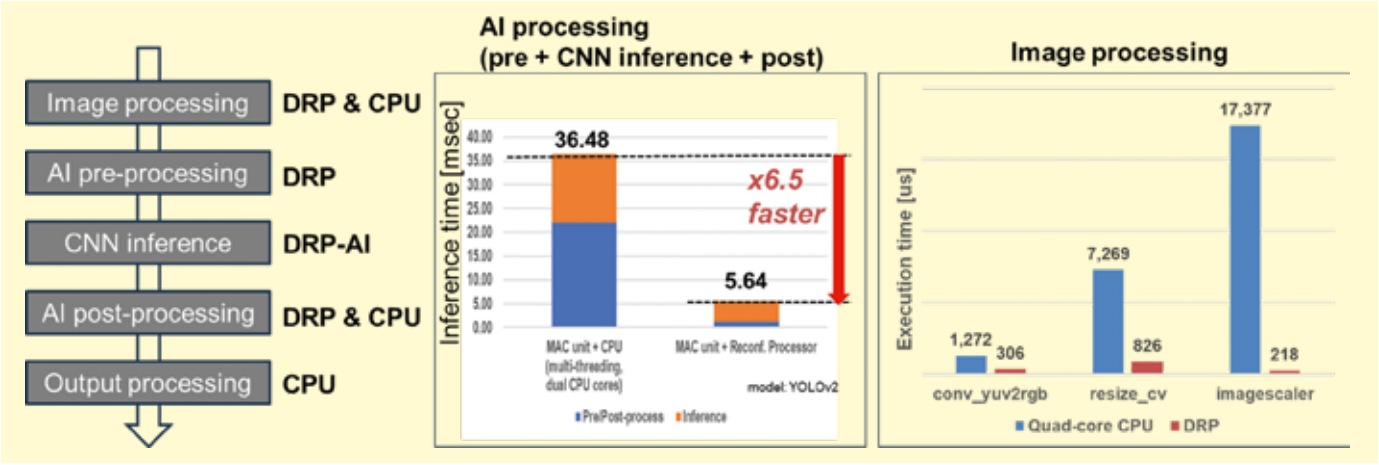

Bild 4. Heterogen arkitektur snabbar upp bearbetningen av bildigenkänning (uppmätt på ett testchip).

DRP är särskilt effektiv när det handlar om att bearbeta strömmande data, till exempel vid bildigenkänning. Parallelliseringen och pipeliningen förbättrar prestandan direkt.

Om man å andra sidan tittar exempelvis på program som planerar och styr robotbeteende sker beräkningarna samtidigt som detaljer och villkor i beräkningarna ändras i respons på förändringar i omgivningen.

Här kan en CPU vara mer lämplig än den typ av hårdvara som gör beräkningarna i en DRP. Det är viktigt att distribuera beräkningarna till rätt ställen och att göra det koordinerat. Renesas heterogena arkitektur gör det möjligt för DRP och CPU att samarbeta.

En översikt över MPU:ns och AI-acceleratorns (DRP-AI3) arkitektur finns i bild 2. I robottillämpningar används en kombination av AI-baserad bildigenkänning och icke-AI-baserade planerings- och kontrollalgoritmer.

Därför är en konfiguration med en DRP för AI-bearbetning (DRP-AI3) och en DRP för icke-AI-algoritmer något som avsevärt kan öka genomströmningen i en robottillämpning.

Resultat efter utvärdering

Utvärdering av AI-modellens prestanda

En RZ/V2H med denna teknik levererade upp till 8 Tops (åtta biljoner produktsummor per sekund) prestanda i AI-acceleratorn.

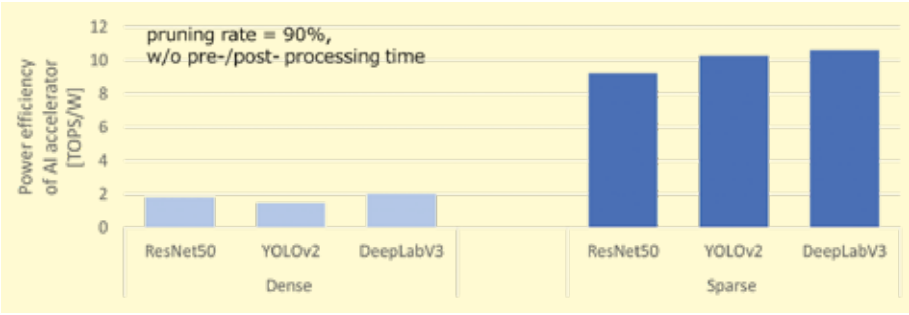

Bild 5. Energieffektivitet för verkliga AI-modeller (uppmätt på testchip).

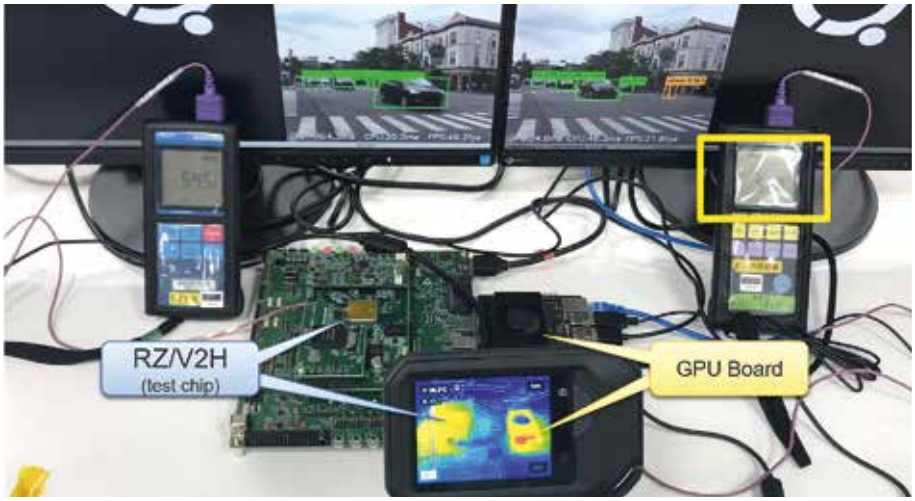

Bild 6. Jämförelse av värmeutveckling mellan ett fläktlöst RZ/V2H-kort och en GPU med fläkt.

I beskurna AI-modeller kan antalet operationer minskas proportionellt mot graden av beskärning och därmed ge en prestanda på upp till 80 Tops jämfört med obeskurna modeller.

Detta är cirka 80 gånger mer än i föregående RZ/V-produkter – en klar förbättring som är tillräckligt stor för att kunna hålla takten med den snabba AI-utvecklingen (bild 3).

I takt med att AI-beräkningarna sker allt snabbare blir å ena sidan den icke AI-baserade algoritmbaserade bildbehandlingen (exempelvis för- och efterbehandling av AI-beräkningar) relativt en flaskhals.

I AI-MPU:er avlastas en del av bildbehandlingen till DRP:n, vilket bidrar till att förbättra systemets totala prestanda (bild 4).

Vad gäller energieffektivitet har AI-acceleratorn enligt en utvärdering världens högsta sådana – cirka 10 Tops per watt vid körning av stora AI-modeller (bild 5).

Vi klarade även av att utföra AI-realtidsberäkningar på ett fläktlöst RZ/V2H-utvärderingskort vid temperaturer som var jämförbara med konkurrenters fläktutrustade produkter (bild 6).

Exempel på robottillämpningar

Slam (Simultaneous Localization and Mapping) är en vanlig tillämpning för robotar. Den har en komplex konfiguration som använder parallella processer för positionsbestämning och omgivningstolkning via AI.

Med hjälp av Renesas DRP kan roboten växla program på ett ögonblick. Parallell drift med hjälp av AI-accelerator och CPU har visat sig vara cirka 17 gånger snabbare än drift med enbart CPU. Dessutom används en tolftedel så mycket ström.

Slutsatser

AI-processorn Renesas RZ/V2H är tio gånger mer energieffektiv än sina föregångare och kombinerar ändnodernas krav på låg effekt och flexibilitet med den processorkapacitet som krävs för att arbeta med beskurna AI-modeller. ■