Systemy dla Internetu Rzeczy (43)

Układy bardzo małej mocy do przyspieszania sztucznej inteligencji

Przyspieszanie działania aplikacji sztucznej inteligencji (AI, Artificial Intelligence), uczenia maszynowego (ML, Machine Learning) i głębokiego uczenia (DL, Deep Learning) jest wciąż stosunkowo nową dziedziną. Mimo to powstaje wiele procesorów, które przyspieszają prawie każde zastosowanie sieci neuronowej. Szczególne znaczenie mają układy bardzo małej mocy.

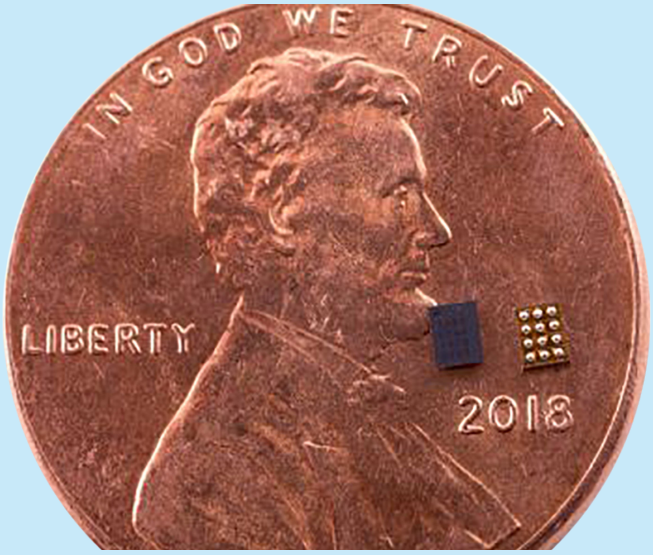
Procesor w pamięci (PIM)

Przemysł półprzewodników przez długi czas osobno projektował i konfigurował pamięć oraz procesor. Tak zwany procesor w pamięci (*PIM, Processor In Memory*), procesor w pobliżu pamięci (*PNM, Processor Near Memory*) albo obliczenia w pamięci (*IMC, In-Memory Compute*) znajduje się na chipie. Najpierw projektowany jest obwód pamięci przechowującej tablicę, a następnie obwód wykonawczy – dodawany, aby uczynić obwody przechowywania i obliczeń niemal zintegrowanymi (**rysunek 1**).

Technologię PIM zapoczątkowano w latach 90., ale nie zyskała popularności. Wraz z rozwojem sztucznej inteligencji, uczenia maszynowego i głębokiego uczenia przemysł docenił i rozwinął technologię oraz chipy PIM. Stoi za tym fakt, że obecna technologia głównego nurtu napotkała wiele wąskich gardeł w poprawie wydajności obliczeniowej. Liczba obliczeń wymaganych do głębokiego uczenia stale rośnie, szczególnie w przypadku obsługi aplikacji do autonomicznej pracy (*self-driving*).

Ze względu na rosnące zapotrzebowanie na proces głębokiego uczenia, w rozwój technologii PIM zainwestowało w ostatnich latach kilka organizacji i przedsiębiorstw, takich jak: ISAAC, Tetris, NeuroCube, Mythic, Syntiant, IBM, PRIME i PipeLayer. Wśród nich jest firma Mythic, która powstała w 2012 roku, i Syntiant, która powstała w roku 2017 [1]. Układy zarówno firmy Mythic, jak i Syntiant, oparte są na wbudowanych obwodach NOR Flash.

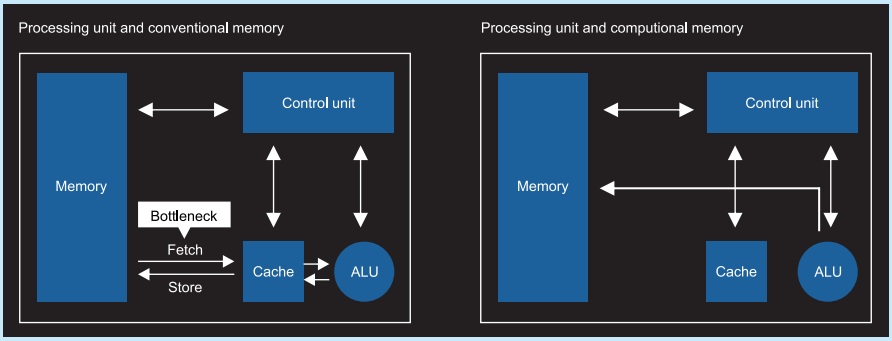
PIM nie ogranicza wyboru rodzaju pamięci: ulotnej (RAM) lub nieulotnej (NVM). Obwód pamięci jest odpowiedzialny za przechowywanie wartości wagi węzłów sieci neuronowej oraz wykonywanie obliczeń sztucznej inteligencji w procesie wnioskowania. Wartość wagi obliczana jest wcześniej, więc nie jest już zmieniana po załadowaniu systemu do pamięci RAM, w procesie bootowania (uruchamiania). Dla tych, którzy muszą często trenować model lub aktualizować model sieci neuronowej, pamięć RAM jest odpowiednia (np. NeuroCube korzysta z pamięci RAM). Obecnie technologia NOR Flash, używana przez Mythic, pochodzi z 40-nanometrowego procesu Fujitsu (obecnie UMC). Syntiant i IBM nie ujawniły informacji związanych z procesami technologicznymi [2].



Obliczenia analogowe z większą wydajnością wykonania

Operacja sieci neuronowej, zorientowana na wnioskowanie, nie wymaga dużej precyzji. Oryginalna operacja oparta na uczeniu wymaga 32- i 16-bitowych liczb zmiennoprzecinkowych. Jednak wnioski są reprezentowane przez 8- lub 4-bitowe liczby całkowite z małą precyzją. Układ przetwarzania liczb całkowitych jest znacznie łatwiejszy do implementacji niż dla liczb zmiennoprzecinkowych. Operację mnożenia z dodawaniem można zaimplementować zarówno w tradycyjnych cyfrowych układach logicznych, jak i w obwodach analogowych. Ten ostatni sposób ma większą szybkość wykonania i lepszą wydajność TOPS/Watt (*Tera-Operations per Second*).

Użycie technologii analogowej nie oznacza, że używany jest taki sam analogowy obwód implementacyjny. Przykładowo, technologia Mythic wykorzystuje przetwornik analogowo-cyfrowy (ADC),



Rysunek 1. Architektura Procesora konwencjonalnego oraz procesora PIM [2]

cyfrową symulację i przetwornik cyfrowo-analogowy (DAC), ale technologia IBM nie wymaga przetworników.

Z powodu nadmiernego przewrażliwienia firmy Intel na używanie nazwy CPU (*Central Processing Unit*), wielu dostawców chipów jest skłonnych do stosowania dla własnych układów słów związanych z PU (*Processing Unit*) i P (*Processor*). Układ PIM firmy Mythic nazywa się *Intelligent Processing Unit* (IPU), firmy IBM – *Resonator Processing Unit* (RPU), a firmy Syntiant – *Neural Decision Processor* (NDP).

Obwody analogowe i cyfrowe nadal muszą ze sobą współpracować

Układ wnioskowania PIM nie jest w całości układem analogowym. Nadal wymaga konwersji analogowo-cyfrowej i współpracy poprzez cyfrowy układ główny, w systemie interfejsu cyfrowego. Dlatego głównym modułem układu jest analogowy obwód wnioskowania, ale otaczające bloki obwodów peryferyjnych są nadal cyfrowe. Wymaga to cyfrowej jednostki sterującej (rdzeń MCU) i cyfrowej pamięci, do wspomagania jej działania.

Analogowe obwody wnioskowania mają różne metody implementacji. W zależności od różnych struktur pamięci wyrażają wartości wag na różne sposoby i implementują mnożenie, dodawanie i działanie z różnymi projektami obwodów arytmetycznych.

Oprócz wydajności, oszczędność energii jest kolejną dużą zaletą procesorów PIM. Wydajność energetyczna układów firmy Mythic wynosi obecnie około 4 TOPS/W. Z kolei firma Syntiant twierdzi, że posiada wydajność energetyczną na poziomie 20 TOPS/W. Przykładowo, dla procesora graficznego NVIDIA Volta V100 (używanego do przyspieszania) szacuje się wydajność na poziomie 0,4 TOPS/W. Firma Syntiant uważa, że jej technologia może zaoszczędzić 50 razy więcej energii niż GPU, przy tych samych wymaganiach wydajnościowych.

Istotna jest również współpraca technologiczna. Firma Lockheed Martin zamierza wdrożyć analogowy chip wnioskowania obrazu firmy Mythic na swoich dronach. Technologia wnioskowania audio firmy Syntiant uzupełnia technologię Alexa firmy Amazon. Syntiant współpracuje z firmą Infineon w zakresie rozwoju technologii wnioskowania audio. Układ NDP jest połączony z mikrofonem MEMS firmy Infineon. Nie wymaga chmury sieciowej ani cyfrowego procesora sygnałowego (DSP) do pełnej korelacji głosu.

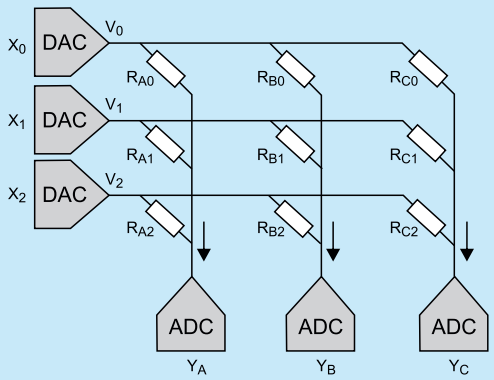
Chociaż stosowanie architektury PIM ma zalety w postaci dobrej wydajności i niskiego zużycia energii, ma obecnie swoje ograniczenia i wady. Dla przykładu: liczba warstw sieci i liczba węzłów nie jest łatwa do rozszerzenia przy dużej skali implementacji obwodów. Skończona dokładność obwodów analogowych (4-bitowy lub 8-bitowy) jest najczęściej używana do wnioskowania, a obsługa operacji uczenia sieci jest nadal ograniczona.

Układ firmy Mythic

Z powodu elastyczności i braku opłat licencyjnych firma Mythic używa rdzenia RISC-V do obsługi dedykowanego zbioru instrukcji SIMD.

Firma Mythic używa 256-rzędowych wartości analogowych rezystancji do reprezentowania 8-bitowych liczb całkowitych, które z kolei reprezentują wagi węzłów, a następnie implementuje mnożenie przez prawo Ohma (**rysunek 2**). Wartości wejściowe są wyrażane jako napięcia, wagi jako przewodnictwo (odwrotność rezystancji), a wyjście jest reprezentowane jako prąd (I=V×G). Dokładność wynosi 4 bity. Zarówno układy PIM firmy Mythic, jak i Syntiant, koncentrują się na wnioskowaniu. Liczba wag jest punktem odniesienia pomiaru wnioskowania symulacji PIM. Większa liczba wag zwykle oznacza również większą i szybszą moc obliczeniową pracy.

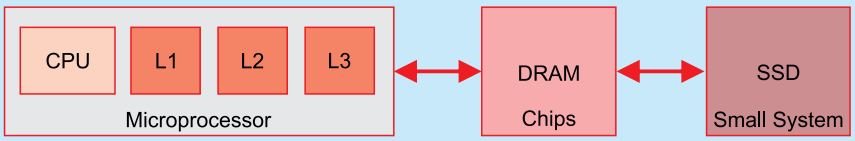
Obecnie najpopularniejsze architektury komputerowe zbudowane są na założeniach dotyczących dostępu do pamięci i jej



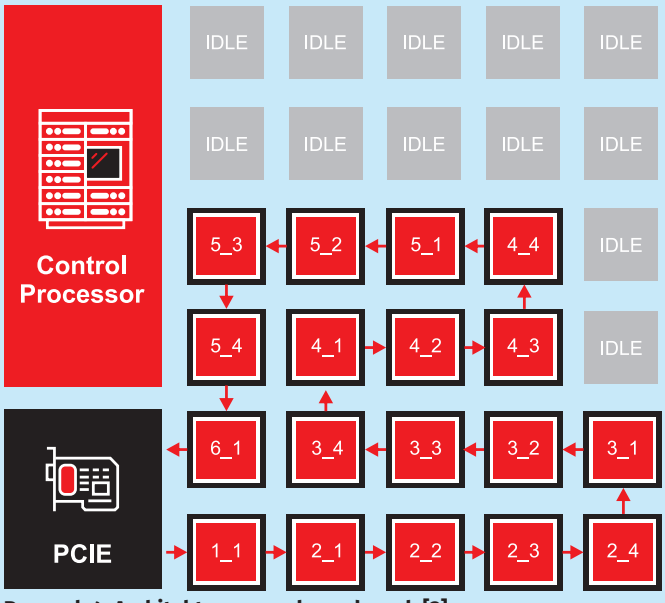
Rysunek 2. Uproszczony układ analogowy wnioskowania [2]

wykorzystania. Systemy te zakładają, że pełna przestrzeń pamięci jest zbyt duża, aby zmieścić się w chipie w pobliżu procesora i że nie wiemy, jaka pamięć będzie potrzebna w jakim czasie. Aby rozwiązać te problemy architektury, te tworzą hierarchię pamięci. Pamięć w pobliżu procesora jest mała i szybka, dlatego może obsługiwać wysoką częstotliwość użytkowania, podczas gdy DRAM i SSD są wystarczająco duże, aby przechowywać większe, mniej wrażliwe czasowo dane (**rysunek 3**).

Obliczenia w pamięci budowane są przy innych założeniach: mamy dużą ilość danych, potrzebujemy do nich dostępu i wiemy, kiedy dokładnie będziemy ich potrzebować. Te założenia są możliwe dla aplikacji wnioskowania AI, ponieważ przepływ wykonywania sieci neuronowej jest deterministyczny – nie jest zależny od danych wejściowych, jak w wielu innych aplikacjach. Korzystając z tej wiedzy, możemy strategicznie kontrolować lokalizację danych w pamięci, zamiast budować hierarchię pamięci podręcznej. Taka organizacja dodaje również lokalne obliczenia do każdej tablicy pamięci, umożliwiając przetwarzanie danych bezpośrednio obok każdej pamięci (**rysunek 4**). Mając obliczenia obok każdej tablicy pamięci, możemy mieć ogromną pamięć, która ma taką samą wydajność jak pamięć podręczna L1 (lub nawet plik rejestrów).



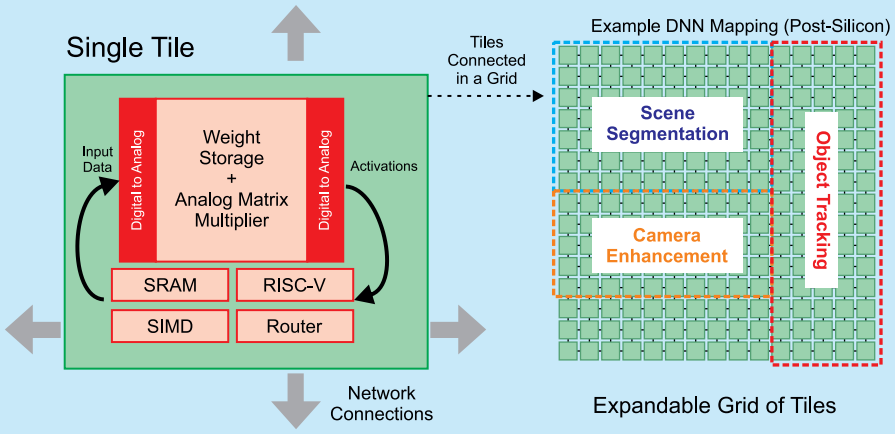
Rysunek 3. Standardowa architektura obliczeniowa [3]



Rysunek 4. Architektura przepływu danych [3]

Poprzednie części kursu i dodatkowe materiały dostępne są na stronie www.media.avt.pl

Wnioskowanie AI nie jest typową aplikacją sekwencyjną. Jest to aplikacja oparta na grafach, w której wyjście z jednego węzła grafu przepływa do wejścia innych węzłów grafu. Aplikacje z użyciem grafów zapewniają możliwości wyodrębnienia równoległości poprzez przypisanie innego elementu obliczeniowego do każdego węzła wykresu. Kiedy wyniki z jednego węzła wykresu są gotowe, przechodzą do następnego węzła wykresu, aby rozpocząć następną operację, co jest idealne dla architektury przepływu danych. W architekturze przepływu danych przypisujemy węzeł wykresu do każdej tablicy obliczeniowej w pamięci i umieszczamy dane wagi dla tego węzła wykresu, w tej tablicy pamięci. Kiedy dane wejściowe dla tego węzła wykresu są gotowe, przesyłane są następnie do właściwej lokalizacji, sąsiadującej z tablicą pamięci, a następnie wykonywane są lokalne obliczenia. Wiele aplikacji do wnioskowania wykorzystuje operacje takie jak splot, która przetwarza bity ramki obrazu naraz, zamiast całej ramki naraz.



Rysunek 5. Struktura układu Mythic [4]

Architektura przepływu danych maksymalizuje również wydajność wnioskowania, ponieważ wiele elementów obliczeniowych w pamięci działa równolegle, potokowo, przetwarzając obraz poprzez przetwarzanie węzłów sieci neuronowych (lub „warstw") równolegle w różnych częściach ramki. Architektura układów Mythic, budowana od podstaw jako architektura przepływu danych, minimalizuje ilość pamięci i obciążenie obliczeniowe, wymagane do zarządzania zależnościami, potrzebnymi do przetwarzania przepływu danych, a także zapewnia maksymalną wydajność aplikacji.

Architektura układu scalonego jest oparta na kafelkach (blokach), gdzie każdy kafelek zawiera wiele jednostek: akcelerator macierzy mnożenia (MMA), procesor RISC-V, silnik SIMD, pamięć SRAM i router Network-on-Chip (NoC) (**rysunek 5**). W zależności od produktu, będą dziesiątki do setek kafelek połączonych ze sobą w siatkę 2D i wiele chipów połączonych przez PCIe. MMA zapewnia przewagę wydajności, dzięki zastosowaniu przetwarzania analogowego, w połączeniu z wbudowaną pamięcią Flash – każda z nich zapewnia ogromną wydajność, około 250 miliardów operacji wielokrotnej akumulacji na sekundę, przy bardzo niskim koszcie energii. Jednostka SIMD zapewnia operacje cyfrowe, których MMA nie może wykonać, takie jak MaxPool lub AvgPool. SRAM przechowuje kod programu i bufory danych. RISC-V zarządza płytką.

Mało informacji jest dostępnych o układzie Mythic [4]: pamięć SRAM 5 MB, liczba wag 50 M, precyzja 1…8 i pobór mocy 1…5 W, wydajność >4 TOPS/W; 0,5 pJ/MAC, interfejs PCIe 2.1.

Układ Mythic to akcelerator wnioskowania sztucznej inteligencji ze statycznie skonfigurowaną pamięcią wagi. To znaczy, że:

1. Jest to układ znajdujący się na karcie PCIe wewnątrz stacji roboczej lub serwera albo układ znajdujący się obok mikrokontrolera, w systemie takim jak inteligentna kamera;
2. Używany jest już wyszkolony algorytm, taki jak sieci neuronowe (tylko wnioskowanie);
3. Układ jest skonfigurowany do uruchamiania jednego zestawu aplikacji naraz, a aplikacje te zmieniają się rzadko. Sporadyczne zmiany są częstym przypadkiem użycia w większości systemów, np. system bezpieczeństwa domowego, który obserwuje intruzów, nie będzie nagle musiał grać w Go.

Układy NDP100 i NDP101 firmy Syntiant

Procesory NDP100 oraz NDP101 firmy Syntiant są przeznaczone do wnioskowania uczenia maszynowego w poleceniach głosowych, w aplikacjach niewielkiej mocy. Układ oparty na architekturze PIM zużywa mniej niż 140 µW mocy czynnej i może uruchamiać modele do wykrywania słów kluczowych, wykrywania słowa budzącego, identyfikacji mówiącego lub klasyfikacji zdarzeń.

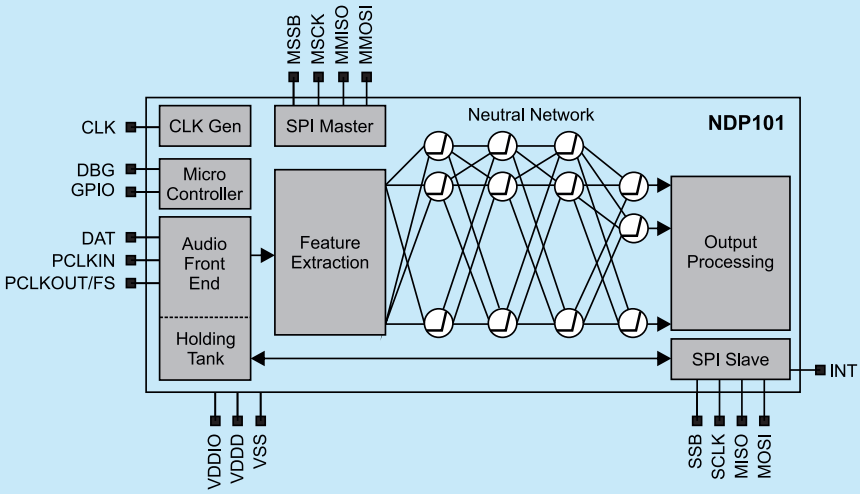
NDP10x jest przeznaczony do przetwarzania głosu w aplikacjach o bardzo niskim poborze mocy. Produkt będzie używany do obsługi urządzeń konsumenckich, takich jak słuchawki douszne, aparaty słuchowe, smartwatche i piloty zdalnego sterowania bez użycia rąk. Widok układu NDP100 pokazuje zdjęcie tytułowe.

Zgodnie z oficjalnymi wiadomościami opublikowanymi przez Syntiant, jego analogowy układ PIM może jednocześnie przechowywać 500 000 współczynników wagowych.

NDP101 osiąga przełomową wydajność, dzięki wysoce zintegrowanym obliczeniom i pamięci, wykorzystując rozległy nieodłączny paralelizm głębokiego uczenia i obliczeń, przy wymaganej precyzji numerycznej. Urządzenia łączą te elementy, aby osiągnąć około 100-krotną poprawę wydajności w porównaniu z architekturami programów przechowywanych w pamięci, takich jak mikroprocesory i procesory DSP.

Kluczowe cechy układu NDP1001 firmy Syntiant (**rysunek 6**) [5]:

- SoC w obudowie QFN32,
- nadaje się do zastosowań wymagających MCU do sterowania oraz silnika neuronowego do podejmowania decyzji,
- interfejs stereo/mono I²S, zmultipleksowany z PDM,



Rysunek 6. Schemat blokowy układu NDP101 firmy Syntiant [5]

- dezpośredni dostęp do sieci neuronowej przez SPI dla aplikacji sensorowych,
- modele wejściowe częstotliwościowe, czasowe i wsadowe,
- uniwersalny procesor Arm Cortex-M0 112 kB SRAM,
- osiem pinów GPIO, z programowalnym kierunkiem,
- obsługa zewnętrznego bootowania szeregowego z pamięci Flash,
- zintegrowany mnożnik i dzielniki zegara obsługują źródło zegara o niskiej częstotliwości lub taktowanie zewnętrzne,
- główny interfejs SPI do obsługi czujników,
- wbudowane zabezpieczenia i uwierzytelnianie oprogramowania układowego,
- English Speech Service do treningu słów kluczowych,
- Software Development Kit (SDK) integruje się z dowolnym środowiskiem oprogramowania,
- Training Development Kit (TDK) umożliwiający użytkownikowi korzystanie ze standardowych struktur, takich jak TensorFlow, do aplikacji programowanych przez klienta,
- obsługa 64 klasyfikacji wnioskowania,
- pobór mocy czynnej <140 µW podczas rozpoznawania słów.

Platforma uruchomieniowa NDP9101B0 łączy w sobie wysokowydajny cyfrowy mikrofon XENSIV IM69D120 MEMS firmy Infineon z procesorem decyzji neuronowej Syntiant NDP101 do prototypowych zastosowań, takich jak wykrywanie słów kluczowych, przetwarzanie słów budzika i identyfikacja mówcy, które zużywają tylko 140 µW. Inne możliwości rozwoju obejmują klasyfikację zdarzeń audio i środowiska oraz analizę czujników.

Zapewniając 20-krotnie większą przepustowość niż typowe rozwiązanie MCU o niskim poborze mocy, Syntiant NDP może obsługiwać do 63 lokalnych poleceń głosowych. Procesor jest również wyposażony w model słów kluczowych Amazon Alexa Voice Service. Mikrofon Infineon XENSIV IM69D120 MEMS jest przeznaczony do zastosowań, które wymagają niskiego poziomu szumów własnych, ma szeroki zakres dynamiczny, niskie zniekształcenia i wysoką odporność na przesterowanie.

Platforma NDP9101B0 jest zaimplementowana jako rozszerzenie Raspberry Pi 3B+, z łatwo konfigurowalnymi zworkami do podłączenia wielu różnych mikrofonów i czujników. Inne kluczowe funkcje obejmują 500 kB pamięci Flash, dwa wbudowane mikrofony IM69D120 Infineon o wysokim zakresie dynamiki, wstępnie skonfigurowaną kartę micro SD, programowalny układ scalony zegara Si5351A i cztery diody LED do wyświetlania IO.

Wtyczka USB NDP101B0, przydatna do zastosowań zasilanych z baterii, z funkcjami budzenia poleceniami głosowymi, jest połączona z 32-bitowym mikrokontrolerem (Atmel MCU ATSAMD21G18, kompatybilnym z Arduino) i baterią litowo-jonową (z układem ładowania). Ma zamontowany mikrofon cyfrowy o wysokim zakresie dynamiki (Infineon IM69D130) oraz 8 MB Flash, do przechowywania modeli sieci neuronowych. Została zastosowana sieć neuronowa 4-warstwowa mająca 560 K połączeń, wstępnie zaprogramowana ze słowem budzenia Alexa.

Układy ECM3531 oraz ECM3532 firmy Eta Compute

Platforma czujników neuronowych TENSAI Platform firmy Eta Compute przenosi sztuczną inteligencję do urządzeń brzegowych i przekształca dane z czujników w przydatne informacje, dotyczące między innymi głosu, aktywności, gestów, dźwięku, obrazu, temperatury, ciśnienia i biometrii. Platforma zapewnia trzy korzyści w obliczeniach brzegowych: krótszy czas odpowiedzi, większe bezpieczeństwo i wyższą dokładność [6].

Unikalna dwurdzeniowa architektura hybrydowa układów rodziny ECM353X Eta'Compute zapewnia najwyższą wydajność aplikacji uczenia maszynowego. Kontroler Arm Cortex-M3 zarządza równoległymi obciążeniami, umożliwiając optymalizację algorytmów, z wykorzystaniem zalet architektury DSP, w tym podwójnego układu MAC, dwóch banków pamięci, sprzętowej obsługi pętli i generowania adresów. Potężne oprogramowanie firmy Eta tworzy grafy zależności wykonania, uruchamia programy na dowolnym rdzeniu, który jest wolny, podczas gdy menedżer mocy automatycznie obniża napięcie zasilania niewykorzystanych zasobów.

Układy mają ciągłe dynamiczne dostrajanie skalowania napięcia i częstotliwości (CVFS, Continuous Voltage Frequency Scaling). Technologia ta umożliwia, w sposób ciągły, zmianę napięcia i częstotliwości, podczas gdy w konwencjonalnej logice synchronicznej można przełączać się tylko między kilkoma dyskretnymi opcjami napięcia i częstotliwości.

Testy porównawcze Eta Compute pokazują, że MCU Cortex działa z 10-krotnie mniejszą mocą niż konkurencyjne rozwiązania, w szerokim zakresie temperatur. Co ważniejsze, został przeprowadzony szereg testów porównawczych sieci neuronowych: rozpoznawanie obrazu, rozpoznawanie dźwięku (np. tłuczenie szkła), wykrywanie ruchu, zawsze włączone rozpoznawanie słów kluczowych i zawsze aktywne rozpoznawanie poleceń. We wszystkich przypadkach układy pracują z mocą w zakresie kilkuset mikroamperów i wykonują wielokrotne wnioskowania na sekundę (do 50 w przypadku wykrywania ruchu).

Układ ECM3531 firmy Eta Compute

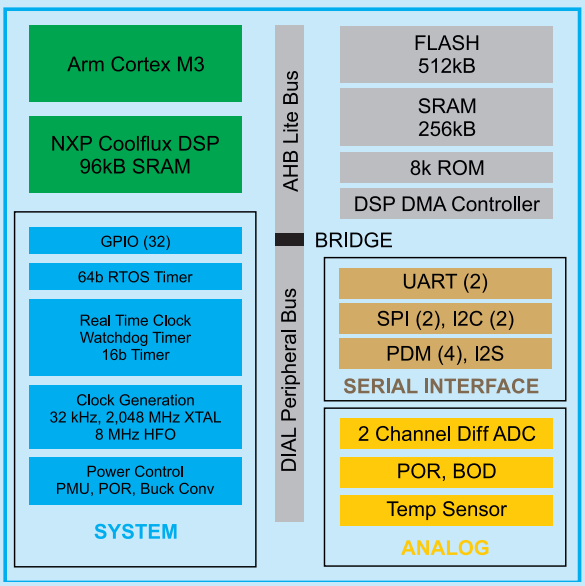
Układ ASIC oparty na rdzeniu ARM Cortex-M3. Zawiera procesor Dual MAC DSP 16 b, wysokiej wydajności interfejs czujnika przetwornika analogowo-cyfrowego (ADC) oraz wysoce wydajne obwody PMIC. Zawiera również I²C, I²S, GPIO, RTC, PWM, POR, BOD, 128 kB SRAM, 64 kB SRAM przeznaczone dla DSP i 512 kB Flash.

Układ ECM3532 firmy Eta Compute

Pierwszy układ produkcyjny ECM3532 typu SoC został zaprojektowany z myślą o akceleracji AI w projektach zasilanych z baterii lub wykorzystujących energię dla IoT. Stałe aplikacje w przetwarzaniu obrazu i fuzji czujników można osiągnąć przy budżecie mocy zaledwie 100 µW.

Układ ma dwa rdzenie: Arm Cortex-M3 i rdzeń DSP CoolFlux firmy NXP. Firma wykorzystuje zastrzeżoną technikę skalowania napięcia i częstotliwości, która dostosowuje każdy cykl zegara. Uczenie maszynowe może być przetwarzane przez dowolny rdzeń (na przykład niektóre obciążenia związane z głosem są lepiej dostosowane do procesora DSP).

Oto niektóre parametry układu ECM3532 firmy Eta Compute (**rysunek 7**) [6]:



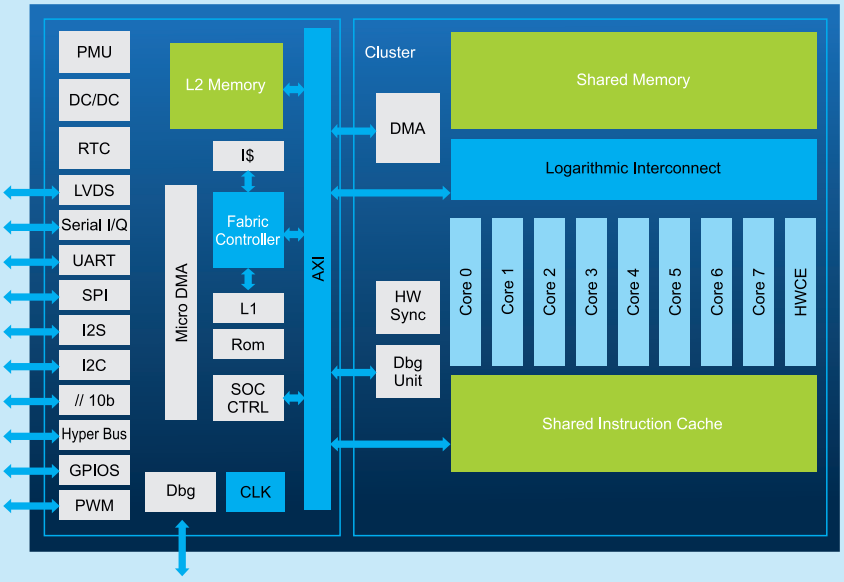
Rysunek 7. Schemat blokowy układu ECM3532 firmy Eta Compute [6]

Poprzednie części kursu i dodatkowe materiały dostępne są na stronie www.media.avt.pl

- rdzeń Arm Cortex-M3 do 100 MHz max., <5 µA/MHz, <1 µA uśpienie z pracą RTC,
- rdzeń DSP: CoolFlux DSP16 firmy NXP, dp 100 MHz, 32 kB SRAM program, 64 kB SRAM dane,
- pamięć systemowa: Flash 512 kB, SRAM 256 kB,
- wydajność 1,3 wnioskowania/s; 4,6 mW/5,6 mW z kamerą (średnio), obraz 96×96.

Układy GAP8 i GAP9 firmy GreenWaves z klasterem rdzeni RISC-V

Firma GreenWaves została założona pod koniec 2014 roku, z misją zrewolucjonizowania rynku inteligentnych czujników i urządzeń z ultraenergooszczędnymi i ekonomicznymi rozwiązaniami [7]. GAP8 firmy GreenWaves to pierwszy w branży procesor o bardzo niskim poborze mocy, który umożliwia zasilaną bateryjnie sztuczną inteligencję w aplikacjach Internetu Rzeczy. GAP8 i GAP9 zasilają nowe typy urządzeń, które łączą bardzo niskie zużycie energii z wyrafinowanym przetwarzaniem sygnału i algorytmami sieci neuronowej.



Rysunek 8. Architektura układu GAP8 firmy GreenWaves z rdzeniem RISC-V [7]

GreenWaves jest kluczowym współtwórcą platformy Open Source, opartej na równoległej platformie przetwarzania ultra niskiego poboru mocy RISC-V (PULP), która stanowi podstawę dla procesorów GAP8 i GAP9.

Procesory GAP zawierają dynamiczne skalowanie napięcia i częstotliwości oraz automatyczne bramkowanie zegara, utrzymujące w stanie czuwania tylko te elementy komponentu, które są niezbędne do obsługi bieżącego obciążenia. Niski pobór mocy w trybie czuwania oraz ultraszybkie wybudzanie i przejścia między stanami zasilania minimalizują zużycie energii w stanach uśpienia, akwizycji, przetwarzania i komunikacji.

Wielordzeniowy klaster obliczeniowy z architekturą pamięci współużytkowanej i sprzętową synchronizacją wątków umożliwia wysoce wydajną, równoległą implementację algorytmów, dając niemal optymalne przyspieszenie liniowe.

Procesory oparte na GAP RISC-V są w pełni programowalne w C/C++, co daje możliwość optymalizacji najnowszych algorytmów. Pakiet GAP SDK zawiera wszystkie narzędzia niezbędne do przyspieszenia tworzenia aplikacji, w tym w pełni zautomatyzowany łańcuch narzędzi sieci neuronowej firmy Google Tensorflow. Unikalny generator kodu GAP AutoTiler automatycznie optymalizuje przepływ danych przez chip.

Układ GAP8 zbudowano z zastosowaniem klastra ośmiu rdzeni aplikacyjnych RISC-V (175 MHz), ze współdzieloną pamięcią danych 64 kB i instrukcji 4 kB oraz akceleratorem sieci neuronowych (HWCE) (**rysunek 8**) [7]. Układ dostarcza mocy obliczeniowej 22,65 GOPS, z wydajnością 4,24 mW/GOP. Układem zarządza rdzeń RISC-V „Fabric Controller (FC)" (250 MHz), z pamięcią danych 16 kB i instrukcji 1 kB.

FC dostarcza 200 MOPS, przy 10 mW dla 1,2 V/250 MHz lub 4 mW dla 1,0 V/150 MHz. Wszystkie rdzenie i moduły układu mają kluczowanie zasilania i skalowanie częstotliwości pracy. W trybie głębokiego uśpienia układ pobiera 2 µA, a przy podtrzymaniu zawartości pamięci L2 512 kB – tylko 8 µA. Zimny start trwa 0,5 ms, start klastra 200 µs, a stop klastra 10 µs. Klaster procesorów i HWCE współdzieli dostęp do pamięci L2 RAM 512 kB oraz pamięci instrukcji. Klaster układów DMA umożliwia autonomiczny transfer danych pomiędzy obszarami pamięci, równolegle z obliczeniami i przy niskim poborze mocy.

Układ GAP9 jest znaczącym ulepszeniem poprzedniego modelu. Został wykonany w technologii 22 nm, przy zwiększeniu przepustowości pamięci, zastosowaniu kompresji danych i obliczeń zmiennoprzecinkowych [7]. Moc układu GAP9 została zmniejszona 5 razy, przy 10 razy większym rozmiarze sieci neuronowej, którą może przetwarzać. Układ GAP9 jest zbudowany z zastosowaniem klastra dziewięciu rdzeni aplikacyjnych RISC-V (400 MHz), ze współdzieloną pamięcią danych i instrukcji 128 kB oraz pamięcią nieulotną 2 MB. Zestaw instrukcji został mocno dostosowany do optymalizacji zużywanej energii. Układ dostarcza mocy obliczeniowej 150,8 GOPS, z wydajnością 0,22 mW/GOP i wydajnością przetwarzania 806 µW/ramkę/sekundę (na obrazach 160×160). Układem zarządza rdzeń RISC-V „Fabric Controller (FC)" (400 MHz).

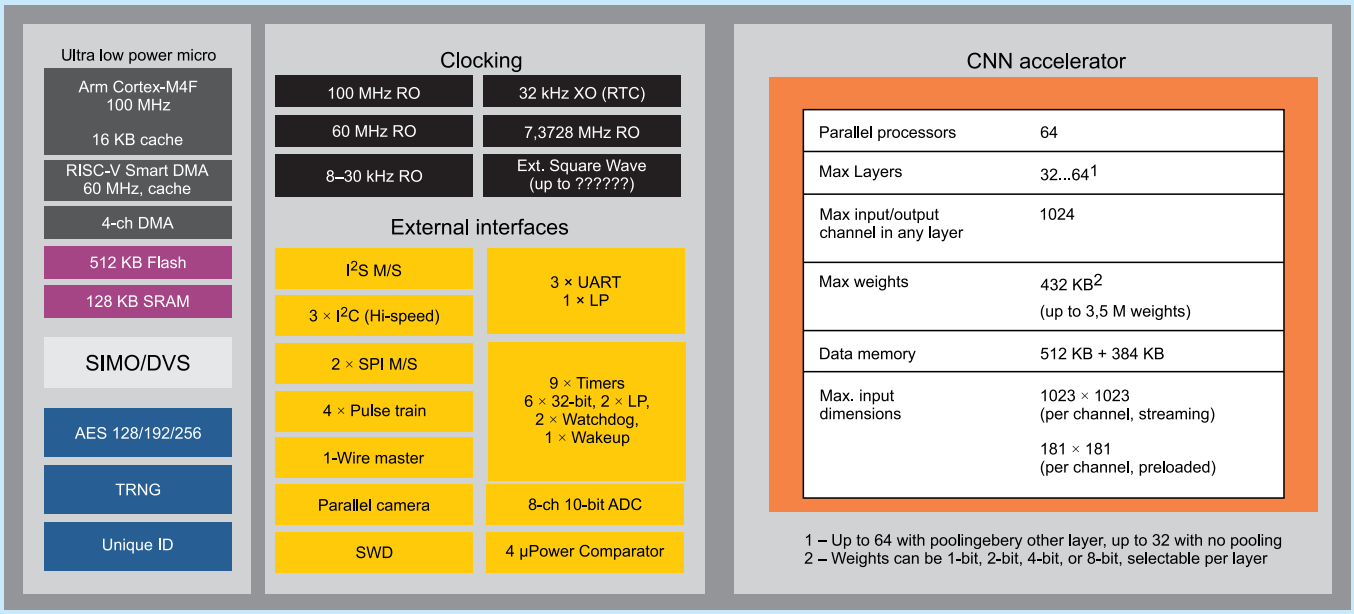
Układ MAX78000 firmy Maxim Integrated

Maxim Integrated jest znanym dostawcą elektroniki, dostarczającym wszystko, od układów mikroelektroniki, po rozwiązania do zarządzania energią. Sztuczna inteligencja (AI) i uczenie maszynowe (ML) nie były częścią tej oferty aż do teraz.

Najnowszy układ MAX78000 ma dwa rdzenie o bardzo niskim poborze mocy oraz dodany akcelerator uczenia maszynowego CNN (*Convolutional Neural Network*) [9]. Cortex-M4F jest głównym procesorem, podczas gdy RISC-V obsługuje we/wy strumienie danych do/z akceleratora CNN. Oba rdzenie mają dostęp do wszystkich urządzeń peryferyjnych. Dzięki temu programiści mogą wybrać platformę, która może najlepiej odpowiadać ich wymaganiom obliczeniowym lub użycia układu peryferyjnego. Większość programistów będzie pracować w C lub C++, więc wybór rdzenia przetwarzania zwykle nie będzie problemem. Tworzy to interesujące środowisko do kompilacji.

Akcelerator CNN jest zoptymalizowany pod kątem wysokiej wydajności i niskiego poboru mocy. Modele ML mogą obsługiwać aplikacje głosowe i wideo w czasie rzeczywistym. Chodzi o to, aby świadczyć tego typu usługi bez konieczności przechodzenia do chmury.

Akcelerator CNN wykorzystuje 64 procesory. Każdy zawiera układ z dedykowaną pamięcią, przeznaczony do łączenia (*pooling*) oraz splotu. CNN może obsługiwać obraz wejściowy do 1023×1023 pikseli na kanał, w trybie przesyłania strumieniowego lub 181×181 na kanał, gdy dane są wstępnie załadowane. Obsługuje wagi 1-, 2-, 4- i 8-bitowe, z pamięcią 432 kB zarezerwowaną do tego użytku, co oznacza obsługę 3,5 miliona wag. Poza tym obsługuje do 64 warstw w modelu, jeśli łączenie jest zastosowane na co drugiej warstwie, w przeciwnym razie limit wynosi 32 warstwy. Akcelerator CNN zaprojektowano w celu dostarczania rozwiązań o małym opóźnieniu, przy jednoczesnym zmniejszeniu zużycia energii, szczególnie w porównaniu z algorytmami działającymi tylko na rdzeniu Cortex-M4F lub RISC-V.



Rysunek 9. Architektura układu MAX78000 firmy Maxim Integrated z rdzeniem Cortex-M4F, RISC-V CPU oraz akceleratorem CNN [9]

Zastaw rdzeni i urządzeń peryferyjnych jest podobny jak w innych mikrokontrolerach firmy Maxim. Układ zawiera interfejs kamery, cztery komparatory o bardzo małej mocy i ośmiokanałowy, 10-bitowy przetwornik analogowo-cyfrowy (ADC). Chip może obsługiwać aparat cyfrowy, a także działać jako master 1-Wire.

Układ jest przeznaczony do zastosowań, takich jak środowiska zasilane bateryjnie, gdzie wymagana jest niska moc.

MAX78000 ma wiele niezwykłych cech [9] (**rysunek 9**):

- rdzeń aplikacyjny Cortex-M4F 100 MHz, 22,2 µA/MHz,
- rdzeń pomocniczy RISC-V 60 MHz,
- współdzielona pamięć 512 kB Flash i 128 kB SRAM,
- akcelerator sieci neuronowej CNN: 64 rdzenie 50 MHz, 512 kB SRAM, do 64 warstw, obliczanie splotu 1- i 2-wymiarowego, obsługa innych typów sieci (MP, Recurent NN),
- akcelerator sprzętowy AES 128/192/256, TGNG, bezpieczny rozruch,
- zasilanie: 2,0 do 3,6 V, 11,3 µA (tryb oczekiwania).

Zestaw uruchomieniowy MAX78000EVKit za 168 USD, z układem MAX78000, obsługuje przetwarzanie dźwięku i obrazu. Dostępny jest również niedrogi moduł ewaluacyjny. Akceleratory CNN obsługują biblioteki TensorFlow i PyTorch.

Podsumowanie

Na początku tego roku firma ARM zapowiedziała nowy rdzeń ARM Cortex-M55, przeznaczony dla aplikacji sztucznej inteligencji w systemach Internetu Rzeczy. Cortex-M55 umożliwia 15 razy szybsze obliczenia w algorytmach uczenia maszynowego oraz 5 razy szybsze obliczenia DSP, w stosunku do rdzeni Cortex-M poprzedniej generacji. Dodatkowo można do niego dołączyć nowy koprocesor z NPU (Neural Processing Unit) Ethos-U55, który daje kolejną 32-krotną poprawę przetwarzania uczenia maszynowego. Jednak jego dostępność jest zapowiadana na początek roku 2021.

Na podstawie informacji z raportu Farnell IoT Trends Report 2020 [8], uzyskanym z ankiet w 2015 od inżynierów IoT, okazuje się, że 49% ankietowanych korzysta już ze sztucznej inteligencji (AI) we wdrażaniu Internetu Rzeczy. Uczenie maszynowe (ML) jest najczęściej używanym typem AI z 28% popularnością, a następnie AI oparte na chmurze (19%). Jednak 51% respondentów waha się przed użyciem AI, ponieważ są nowicjuszami w tej technologii lub szukają specjalistycznej wiedzy na temat wdrażania AI.

Henryk A. Kowalski
Instytut Informatyki
Politechnika Warszawska

Literatura

[1] Top 10 Processors for AI Acceleration at the Endpoint, Sally Ward-Foxton, EETimes, April 20 2020, https://bit.ly/33pOUMv
[2] The demand for AI chips will drive the revival of this „old" architecture!, Jotrin Electronics, 2019-12-20, https://bit.ly/2USH9cY
[3] Accelerating AI that works for everyone, Technology, Mythics, https://bit.ly/393XJiw
[4] A Peek Into Software Engineering at Mythic, Dave Fick, Nov 5, 2018, https://bit.ly/35WBKZ0
[5] Always-On Voice powered by custom AI Silicon, Syntiant, https://bit.ly/3pRN54d
[6] A New Era in Edge Computing with 1000X More Energy Efficient Neural Networks, https://bit.ly/3kUjRy0
[7] The fundamentals of GAP, the IoT application processors, Green Waves, https://bit.ly/393svIp
[8] IoT Trends Report 2020, Farnell Global IoT Survey, September–-December 2019, Farnell, https://bit.ly/3nR9OvJ
[9] MAX78000, Maxim Integrated, https://bit.ly/3fvS2ew

Poprzednie części kursu i dodatkowe materiały dostępne są na stronie www.media.avt.pl