**Poniższy tekst pochodzi z Wikipedii. W tekście jest parę błędów, dodatkowo niektóre informacje są już nieaktualne. Korzystając z własnej wiedzy oraz informacji pochodzących z lokalizacji sieciowych (pamiętaj o weryfikacji informacji przynajmniej w dwóch, jak najbardziej wiarygodnych źródłach) postaraj się poprawić informacje dotyczące GPU oraz APU.**

**Możesz też opisać budowę i zasadę działania GPU i APU nie korzystając z tekstu poniżej.**

**Gotową pracę prześlij na adres m.sobocinska-malek@zsl.gda.pl**

**Procesor graficzny**

**Procesor graficzny**, **GPU** (od ang. *graphics processing unit*) – jednostka obliczeniowa znajdująca się w kartach graficznych.

**Historia**

Jako pierwsza tego terminu użyła firma [Nvidia](https://pl.wikipedia.org/wiki/Nvidia), wprowadzając na rynek karty graficzne GeForce 256 w 1999 roku. Głównym zadaniem GPU było wykonywanie obliczeń potrzebnych do uzyskania grafiki 3D, co spowodowało uwolnienie głównego procesora (CPU) od konieczności wykonywania tego zadania. W tej sytuacji mógł on zająć się innymi obliczeniami, co skutkowało zwiększeniem wydajności komputera podczas renderowania grafiki. Nowoczesne procesory graficzne wyposażone są w szereg rozkazów, których nie posiada CPU komputera.

Należy wspomnieć, że wcześniej, w połowie lat 90., na rynku konsumenckim istniały już karty graficzne wyposażone w procesory oferujące akcelerację operacji przetwarzania grafiki 3D, na przykład S3 Virge, oraz dodatkowe akceleratory graficzne (współpracujące z już zainstalowaną kartą graficzną 2D) firmy 3dfx, o nazwie Voodoo Graphics. Nvidia zadebiutowała produktem Riva z szeregiem odmian (Riva 128, Riva TNT). Kolejne generacje kart graficznych tej firmy, których powstało już kilkadziesiąt, nazywane są GeForce.

Obecnie liczącymi się producentami procesorów graficznych są firmy: Nvidia, AMD oraz Intel.

**Rodzaje GPU**

**Oddzielne karty grafiki**

Najbardziej zaawansowane procesory graficzne używane są obecnie w niezależnych urządzeniach, które nazywa się dedykowanymi kartami graficznymi. Takie karty montuje się na płytach głównych, za pomocą przeznaczonych do takich zastosowań złącz jak PCI Express ×16 lub starsze AGP charakteryzujących się odpowiednio dużą przepustowością transferu danych dla różnych kart. Procesory graficzne projektowane są do współpracy z pamięcią RAM znajdującą się na kartach graficznych (np. typu GDDR5 czy HBM). Dzięki technologiom CrossFire i SLI nowoczesne płyty główne pozwalają na współpracę wielu procesorów graficznych równocześnie na jednej platformie, zwiększając tym samym wydajność całego systemu.

**Procesory zintegrowane**

Rozwiązaniem stosowanym przez wielu producentów płyt głównych jest instalacja procesora graficznego zintegrowanego z chipsetem na mostku północnym lub bezpośrednio w APU. Jest to przede wszystkim tańsze rozwiązanie, gdyż wdrażane jest w trakcie produkcji samych płyt głównych i nie pochłania dodatkowych zasobów, ale zainstalowane w ten sposób układy graficzne charakteryzują się o wiele mniejszą wydajnością. Zintegrowane procesory graficzne wykorzystują pamięć RAM, przez co mogą zmniejszać możliwości operacyjne głównego procesora.

**Hybrydowe przetwarzanie grafiki**

Ta nowsza klasa układów GPU konkuruje ze zintegrowaną grafiką na niskich rynkach komputerów stacjonarnych i notebooków. Najczęstszymi implementacjami tego są HyperMemory ATI i TurboCache Nvidii.

Hybrydowe karty graficzne są nieco droższe niż zintegrowana grafika, ale znacznie tańsze niż dedykowane karty graficzne. Te współdzielą pamięć z systemem i mają małą dedykowaną pamięć podręczną, aby nadrobić wysokie opóźnienie pamięci RAM systemu. Technologie w ramach PCI Express mogą to umożliwić. Chociaż rozwiązania te są czasami reklamowane jako mające 768 MB pamięci RAM, dotyczy to tego, ile można udostępnić pamięci systemowej.

**Przetwarzanie strumienia i układy GPU ogólnego przeznaczenia (GPGPU)**

Coraz powszechniejsze staje się używanie procesora graficznego ogólnego przeznaczenia (GPGPU) jako zmodyfikowanej formy procesora strumieniowego (lub procesora wektorowego), pracującej na jądrach obliczeniowych. Ta koncepcja przekształca ogromną moc obliczeniową nowoczesnego potoku cieniującego akceleratora grafiki w uniwersalną moc obliczeniową, w przeciwieństwie do twardego okablowania wyłącznie do operacji graficznych. W niektórych zastosowaniach wymagających dużych operacji wektorowych, może to zapewnić wydajność o kilka rzędów wielkości wyższą niż konwencjonalny procesor. Dwóch największych dyskretnych projektanów GPU, AMD i Nvidia, zaczynają realizować to podejście z szerokim wachlarzem aplikacji. Zarówno Nvidia, jak i AMD połączyły siły z Uniwersytetem Stanforda, aby stworzyć klienta opartego na GPU dla projektu Folding@home przetwarzania rozproszonego do obliczeń zwijania białek. W pewnych okolicznościach procesor graficzny oblicza czterdzieści razy szybciej niż konwencjonalne procesory tradycyjnie używane przez takie aplikacje.

GPGPU może być wykorzystywany do wielu typów kłopotliwych zadań równoległych, w tym śledzenia promieni. Zasadniczo są one dostosowane do wysokowydajnych obliczeń typu, które wykazują równoległość danych w celu wykorzystania architektury szerokości SIMD szerokiego wektora GPU.

Co więcej, wydajne komputery oparte na GPU zaczynają odgrywać znaczącą rolę w modelowaniu na dużą skalę. Trzy z 10 najpotężniejszych superkomputerów na świecie wykorzystują przyspieszenie GPU.

GPU obsługuje rozszerzenia API do języka programowania C, takie jak OpenCL i OpenMP. Co więcej, każdy sprzedawca GPU wprowadził własny interfejs API, który działa tylko z kartami, AMD APP SDK i CUDA z AMD i Nvidii. Te technologie umożliwiają uruchamianie na procesorach strumieniowych GPU określonych funkcji nazywanych jądrami obliczeniowymi z normalnego programu C. To sprawia, że ​​programy C są w stanie wykorzystać zdolność GPU do pracy na dużych buforach równolegle, jednocześnie nadal wykorzystując CPU, gdy jest to odpowiednie. CUDA to także pierwszy interfejs API, który pozwala aplikacjom opartym na procesorze na bezpośredni dostęp do zasobów GPU w celu bardziej ogólnego przetwarzania bez ograniczeń związanych z interfejsem API grafiki.

Od 2005 r. Obserwuje się zainteresowanie wykorzystaniem wydajności oferowanej przez układy GPU do obliczeń ewolucyjnych w ogóle oraz do przyspieszenia oceny sprawności w szczególności w programowaniu genetycznym. Większość podejść kompiluje programy liniowe lub drzewa na komputerze-hoście i przenosi plik wykonywalny na GPU, który ma zostać uruchomiony. Zwykle przewaga wydajności jest uzyskiwana tylko przez równoległe uruchomienie pojedynczego aktywnego programu w wielu przykładowych problemach, wykorzystując architekturę SIMD GPU. Jednak znaczne przyspieszenie można również uzyskać, nie kompilując programów, a zamiast przekazywania ich do GPU, do interpretacji tam. Przyspieszenie można następnie uzyskać, interpretując jednocześnie wiele programów, jednocześnie uruchamiając wiele problemów przykładowych lub kombinację obu. Nowoczesny procesor graficzny może z łatwością jednocześnie interpretować setki tysięcy bardzo małych programów.

**Zewnętrzny GPU (eGPU)**

Zewnętrzny procesor graficzny to procesor graficzny umieszczony poza obudową komputera. Zewnętrzne procesory graficzne są czasem używane z laptopami. Laptopy mogą mieć znaczną ilość pamięci RAM i wystarczająco wydajną jednostkę centralną (CPU), ale często nie posiadają wydajnego procesora graficznego, a zamiast tego mają mniej wydajny, ale bardziej energooszczędny wbudowany układ graficzny. Pokładowe układy graficzne często nie są wystarczająco mocne, aby odtwarzać najnowsze gry lub inne graficznie intensywne zadania, takie jak edycja wideo.

Dlatego pożądane jest, aby móc dołączyć GPU do zewnętrznej szyny notebooka. PCI Express to jedyna powszechnie używana magistrala do tego celu. Portem może być na przykład port ExpressCard lub port mPCIe (odpowiednio PCIe ×1, do 5 lub 2,5 Gbit/s) lub port Thunderbolt 1, 2 lub 3 (PCIe ×4, do 10, 20 lub 40 Gbit/s odpowiednio). Te porty są dostępne tylko w niektórych systemach notebook.

Zewnętrzne procesory graficzne mają niewielkie oficjalne wsparcie dostawców. Jednak nie powstrzymało to entuzjastów od wdrażania konfiguracji eGPU.

**Funkcje obliczeniowe**

Współczesne układy GPU wykorzystują większość swoich tranzystorów do wykonywania obliczeń związanych z grafiką komputerową 3D. Początkowo były wykorzystywane do przyspieszania intensywnej pamięci polegającej na mapowaniu tekstur i renderowaniu wielokątów, a następnie dodawaniu jednostek w celu przyspieszania obliczeń geometrycznych, takich jak obracanie i tłumaczenie wierzchołków w różne układy współrzędnych. Najnowsze osiągnięcia w układach GPU obejmują obsługę programowalnych shaderów, które mogą manipulować wierzchołkami i teksturami za pomocą wielu takich samych operacji obsługiwanych przez procesory, techniki oversamplingu i interpolacji, aby zmniejszyć aliasing i bardzo precyzyjne przestrzenie kolorów. Ponieważ większość tych obliczeń obejmuje operacje macierzy i wektorowe, inżynierowie i naukowcy coraz częściej badali wykorzystanie układów GPU do obliczeń nie związanych z grafiką; są one szczególnie dostosowane do innych kłopotliwych problemów równoległych.

Wraz z pojawieniem się głębokiego uczenia się wzrosło znaczenie układów GPU. W badaniach przeprowadzonych przez Indigo stwierdzono, że podczas szkolenia głębokich sieci neuronowych, procesory graficzne mogą być 250 razy szybsze niż procesory. To różnica między jednym dniem treningu a prawie 8 miesiącem i 10 dniami treningu. Wybuchowy rozwój Deep Learning w ostatnich latach został przypisany pojawieniu się GPU ogólnego przeznaczenia.

Oprócz sprzętu 3D, dzisiejsze procesory graficzne posiadają podstawowe funkcje akceleracji 2D i framebuffera (zwykle z trybem zgodności VGA). Nowsze karty, takie jak AMD/ATI HD5000–HD7000, nawet nie posiadają akceleracji 2D; musi być emulowany przez sprzęt 3D.