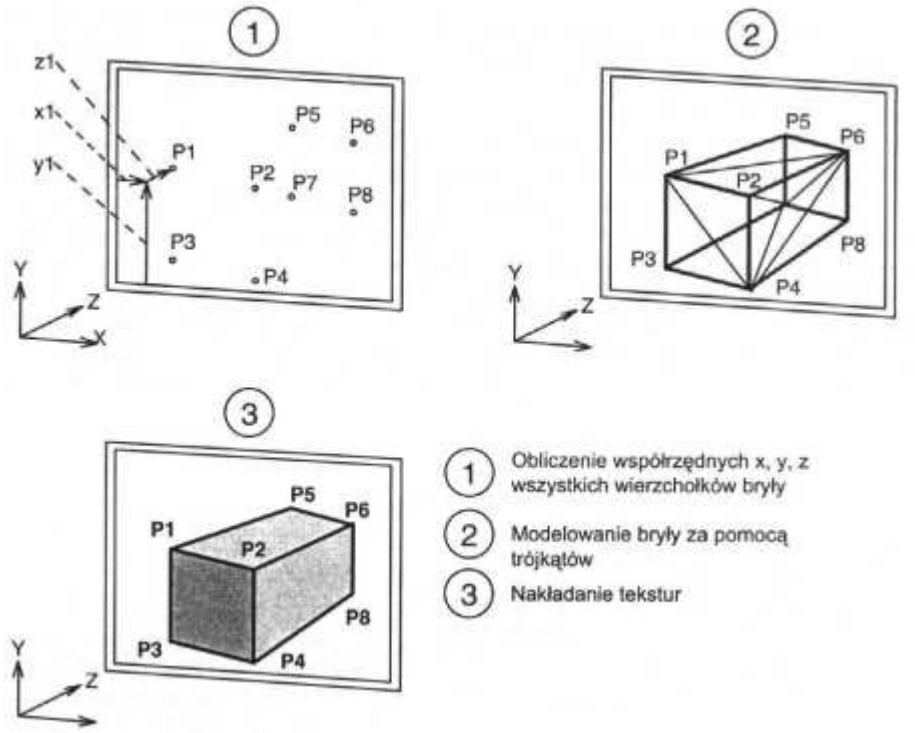
Karta akcelerowana

BUDOWA I ZASADA DZIAŁANIA

Karty graficzne poza sygnałami RGB, odpowiedzialnymi za sterownie trzema katodami kineskopu, generuje jeszcze impulsy synchronizacji poziomej HS i pionowej VS, które synchronizują pracę generatorów odchylania w monitorze ekranowym.



W grafice płaskiej (dwuwymiarowej - 2D) położenie każdego wyświetlanego na ekranie obiektu można określić za pomocą dwóch współrzędnych: x i y. Grafika trójwymiarowa 3D wymaga wprowadzenia trzeciej współrzędnej (**z**), określającej odległość obiektu od obserwatora (od ekranu). Istnienie współrzędnej **z** pozwala ustalić, który z punktów o tych samych współrzędnych x i y powinien być wyświetlony na ekranie, a który będzie niewidoczny.

Pokazana na rysunku bryła zawiera trzy ścianki widoczne (punkty leżące bliżej ekranu) i trzy ścianki niewidoczne (punkty leżące najdalej od ekranu). Wprowadzenie więc trzeciej współrzędnej (**z**) pozwala ograniczyć obliczenia wykonywane przez procesor graficzny jedynie do punktów widocznych (nie ma potrzeby obliczać położenia punktów niewidocznych). Wartość trzeciej współrzędnej punktu wyświetlanego na ekranie powinna być opisana liczbą, co najmniej 16-bitową. Współczesne procesory graficzne umożliwiają wprowadzenie współrzędnej "**z**" nawet 32-bitowej. Każdy więc punkt ekranu będzie opisany za pomocą 32-bitowej informacji o kolorze RGBA oraz za pomocą np. 32-bitowej danej określającej odległość tego punktu od obserwatora. Wartości współrzędnych "**z**" umieszczone są w wydzielonym obszarze pamięci lokalnej karty graficznej o nazwie **Z-bufor**. Przykładowo wymagana pojemność **Z-bufora** dla rozdzielczości 1024 x 768 wynosi 1024 x 768 pikseli x [4 bajty (**opisujące współrzędna Z**)] = 3072KB.

FAZY TWORZENIA GRAFIKI 3D

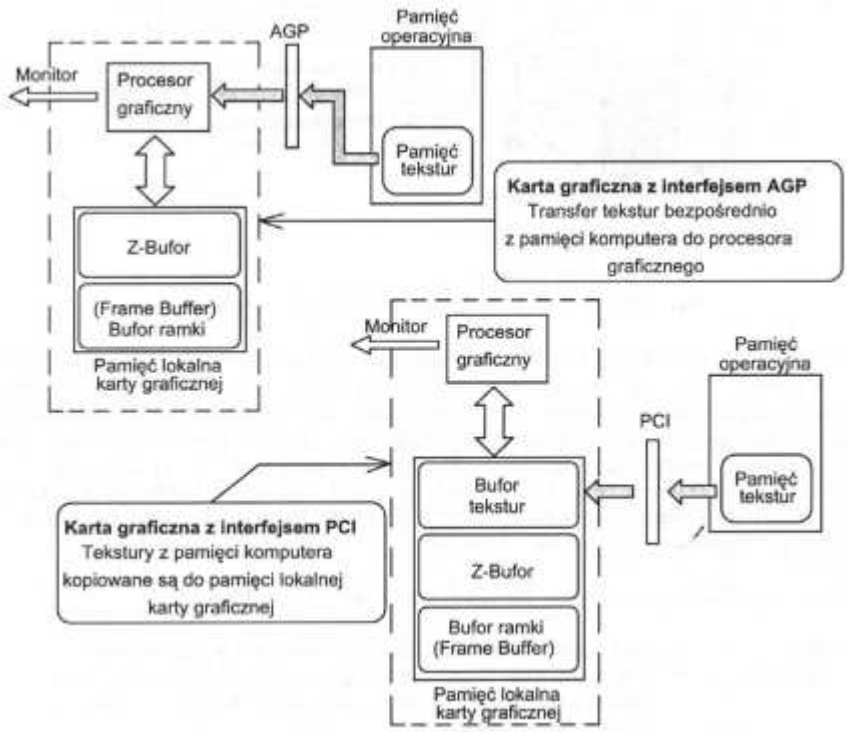
Faza 1 – obliczenie współrzędnych x, y, z wszystkich wierzchołków bryły trójwymiarowej oraz wykonanie operacji skalowania (zmiany rozmiarów bryły), rotacji (obracania bryły) i translacji (przemieszczania bryły),
Faza 2 – modelowanie bryły za pomocą trójkątów (podział wielokątów wyznaczających powierzchnie bryły na trójkąty tzw. teselacja),
Faza 3 – nakładanie tekstur (pokrywanie teksturami każdego trójkąta) - operacja ta nazywana jest renderingiem.

Po wykonaniu powyższych operacji należy dokonać rzutowania trójwymiarowych obiektów na płaski dwuwymiarowy ekran i umieszczenia tych danych w pamięci lokalnej karty graficznej (w tzw. buforze ramki).

Tekstura to dwuwymiarowy obiekt graficzny umożliwiający dekodowanie powierzchni trójwymiarowej bryły (teksturami pokrywa się wszystkie elementy trójwymiarowego obrazu). Pojedynczy punkt tekstury, wyświetlany na ekranie monitora nazywany jest Tekselem (ang. Texture Element).

Tekstury przechowywane są w wydzielonym obszarze pamięci operacyjnej komputera (w tzw. pamięci tekstur), skąd przesyłane są do karty graficznej (do jej pamięci lokalnej). Wydzielony obszar pamięci lokalnej karty graficznej, przeznaczony na tzw. **bufor tekstur** powinien być jak największy, gdyż zbyt mały obszar oznacza ***konieczność ciągłego doładowywania ich z pamięci operacyjnej***, co wiąże się ze znacznym spowolnieniem pracy karty graficznej. Dlatego **pamięć lokalna karty powinna być możliwie duża**.

Inne rozwiązanie oferuje interfejs AGP, który umożliwia transfer tekstur z pamięci komputer bezpośrednio do procesora graficznego bez konieczno ci zapisu ich do pamięci lokalnej karty. Dzięki temu pamięć lokalna karty może mieć mniejszą pojemność – patrz na rysunek.

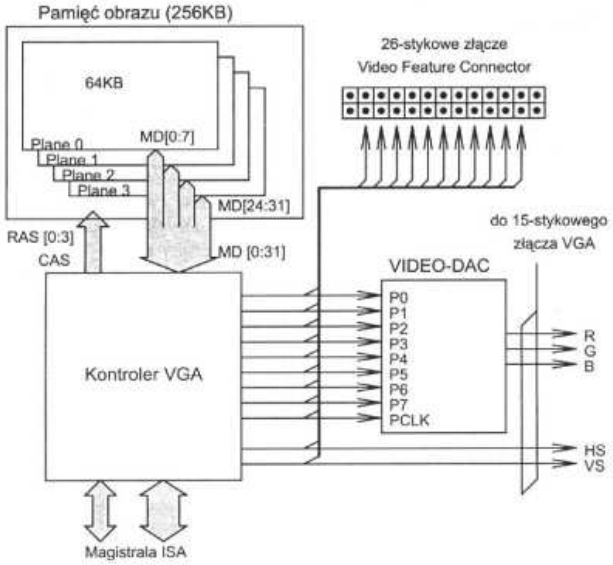


PARAMETRY

Pierwszą zaliczaną do grupy akceleratorów była karta 8514/A firmy IBM. Procesory graficzne to układy 128-bitowe z wbudowanymi konwerterami RAMDAC, współpracujące z pamięciami SDRAM i SGRAM. Obsługują grafikę dwuwymiarową (2D), tryb tekstowy i zaawansowaną grafikę trójwymiarowa (3D). Procesory te posiadają następujące parametry:

częstotliwość pracy układu	do 183 MHz
częstotliwość taktowania pamięci obrazu	do183 MHz
częstotliwość pracy zintegrowanego RAMDAC	do 300 MHz
wielkość obsługiwanej pamięci obrazu	do 32 MB
częstotliwość odchylania pionowego	od 60 do 250 Hz
częstotliwość odchylania poziomego	od 30 do 175 kHz
liczba wyświetlanych kolorów	65K (High Color), 16.7 M (True Color
osiągana rozdzielczość	1920 x 1200

W procesorach tych każdy piksel może być reprezentowany w pamięci obrazu przez 32 bity (tzw. format RGBA) - trzy bajty podstawowe kolory RGB (tryb True Color), bajt czwarty to **Alpha** - współczynnik przejrzystości, określający przejrzystość elementów obrazu przedstawiających np. szkło lub wodę. Współczynnik ten przyjmuje wartość od 0 (pełna przejrzystość) do 255 (brak przejrzystości).

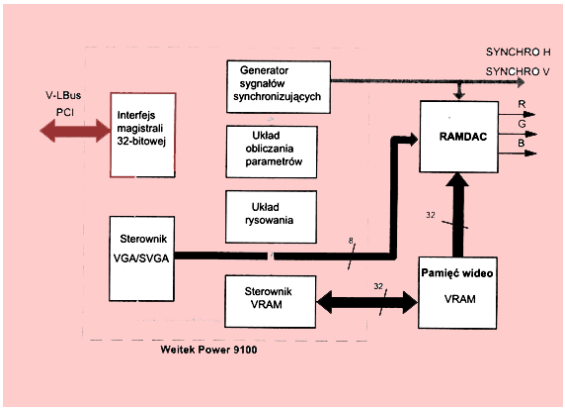


Układ RAMDAC we współczesnych kartach graficznych zintegrowany jest z procesorem graficznym. Od szybkości pracy tego układu uzależnione są rozdzielczość obrazu i częstotliwość odchylania pionowego (odświeżania obrazu).
Przykładowo przy rozdzielczości 1280 x 1024 i przy częstotliwości odchylania pionowego 100 Hz częstotliwość pracy układu RAMDAC nie może być mniejsza niż 173 MHz.
Obliczenia.
Rozdzielczość obrazu: 1280 x 1024 = 1310720 pikseli. Liczba ta pomnożona przez częstotliwość odświeżania ekranu i przez współczynnik 1,32, co daje częstotliwość pracy układu RAMDAC, równą 173 MHz.
Współczynnik 1,32 to stosunek czasu kreślenia jednej linii do czasu kreślenia widocznej na ekranie części linii.

Procesory graficzne

Procesory graficzne (**GPU**) są układami bardzo rozbudowanymi - realizują wiele funkcji, które dotychczas były wykonywane przez procesory główne. GPU zostały wyposażone w tak zwane jednostki przekształcania i oświetlania (ang. T&L - Transform and Lighting), zajmujące się operacjami przeliczania współrzędnych i oświetlania poszczególnych trójkątów. Współczesne procesory potrafią obliczyć współrzędne i wykreślić wiele milionów trójkątów w ciągu jednej sekundy.

SCHEMAT BLOKOWY GPU



Karta graficzna z układem WEITEK POWER 9100

SCHEDER

Shader – krótki program komputerowy, często napisany w specjalnym języku (*shader language*), który w grafice trójwymiarowej opisuje właściwości pikseli oraz wierzchołków. Technologia ta zastąpiła stosowaną wcześniej jednostkę T&L.

Cieniowanie pozwala na znacznie bardziej skomplikowane modelowanie oświetlenia i materiału na obiekcie niż standardowe modele oświetlenia i teksturowanie. Jest jednak dużo bardziej wymagające obliczeniowo i dlatego dopiero od kilku lat sprzętowa obsługa cieniowania jest obecna w kartach graficznych dla komputerów domowych. Wcześniej cieniowanie stosowane było w niektórych fotorealistycznych rendererach (np. Renderman), gdzie grafika nie jest generowana w czasie rzeczywistym.

W stosunku do standardowych modeli oświetlenia, stosowanych do generowania grafiki w czasie rzeczywistym, cieniowanie umożliwia uwzględnienie między innymi:

- refrakcji,
- odbić lustrzanych,
- oświetlenia HDR,
- mapy przemieszczeń (*displacement maps*),
- innych efektów takich jak rozmycie obrazu, zaszumienie, zmiana kolorów, itp.

TECHNOLOGIA CUDA

CUDA (ang. **Compute Unified Device Architecture**) – opracowana przez firmę Nvidia uniwersalna architektura procesorów wielordzeniowych (głównie kart graficznych) umożliwiająca wykorzystanie ich mocy obliczeniowej do rozwiązywania ogólnych problemów numerycznych w sposób wydajniejszy niż w tradycyjnych, sekwencyjnych procesorach ogólnego zastosowania.

CUDA (ang. **Compute Unified Device Architecture**) – opracowana przez firmę Nvidia uniwersalna architektura procesorów wielordzeniowych (głównie kart graficznych) umożliwiająca wykorzystanie ich mocy obliczeniowej do rozwiązywania ogólnych problemów numerycznych w sposób wydajniejszy niż w tradycyjnych, sekwencyjnych procesorach ogólnego zastosowania.

Integralną częścią architektury CUDA jest oparte na języku programowania C środowisko programistyczne wysokiego poziomu, w którego skład wchodzą m.in. specjalny kompilator (nvcc), debugger (cuda-gdb, który jest rozszerzoną wersją debuggera gdb umożliwiającą śledzenie zarówno kodu wykonywanego na CPU, jak i na karcie graficznej), profiler oraz interfejs programowania aplikacji.

Zalety

W stosunku do tradycyjnych metod wykonywania obliczeń inżynierskich na GPU, CUDA posiada kilka istotnych zalet:

- Język programowania oparty na jednym z najpopularniejszych języków programowania – języku C.
- Model pamięci procesora ściśle odpowiadający architekturze sprzętowej, co umożliwia świadome, efektywne wykorzystywanie dostępnych zasobów GPU, w tym pamięci współdzielonej (obecnie do 48KB). Pamięć ta jest współdzielona przez wszystkie wątki w tzw. bloku (zwykle 128-512 wątków). Można jej używać jako programowalnej pamięci typu cache.
- Kod uruchamiany na GPU może odczytywać i zapisywać dane z dowolnego adresu w pamięci GPU.
- Projekt architektury CUDA zakłada pełną kompatybilność programów – napisany dziś program wykonywalny ma w przyszłości działać bez żadnych zmian na coraz wydajniejszych procesorach graficznych posiadających coraz większą liczbę rdzeni, rejestrów, pamięci operacyjnej i innych zasobów.

Zastosowania

W grach komputerowych moc obliczeniową można wykorzystać do obliczeń fizyki w grach, ale CUDA jest również wykorzystywana do przyspieszania obliczeń w takich dziedzinach jak biologia, fizyka, kryptografia i inne obliczenia inżynierskie. Dla potrzeb tego segmentu Nvidia opracowała specjalny procesor graficzny Tesla.

- Przyspieszenie szyfrowania i kompresji oraz konwersji wideo do różnych formatów
- Symulacje fizyczne (np. w dynamice płynów) i obliczenia inżynierskie
- Obrazowanie wirtualnej rzeczywistości na podstawie tomografii komputerowej i rezonansu magnetycznego
- Efekty specjalne w grafice komputerowej, np. symulacje falujących powierzchni ubrań
- Sztuczna inteligencja

Przykłady:

Plug-in filtra do programu GIMP.

Symulacja przepływu płynów.

Aplikacja Robot Range Finder, która symuluje ruchy obiektów i reaguje na nie w czasie rzeczywistym.

Program do identyfikacji twarzy na podstawie jej cech charakterystycznych (w oparciu o obraz z kamery internetowej), oraz jej śledzenia.

PAMIĘCI NA KARTACH GRAFICZNYCH

Układ pamięci 64MB Qimonda GDDR3

GDDR 4 RAM (Graphic Double Data Rate v4) - pamięć RAM czwartej generacji wykorzystywana do produkcji kart graficznych. Ze względu na niewielkie tylko ulepszenia w stosunku do GDDR3, pamięci te były mało popularne i bardzo rzadko stosowane - układy GDDR3 zostały w nowszych konstrukcjach zastąpione od razu przez GDDR5.

Pamięci GDDR4 wykonywane były w technologii 80 nm, w kościach o pojemności 64 MB. Pracowały one nominalnie (taktowanie rzeczywiste) z prędkością 1200 MHz, a efektywnie 2400 MHz. Pamięci te uzyskiwały przepustowość na poziomie 9,6 GB/s. Były wykorzystywane tylko przez ATI. Następcą GDDR4 są pamięci GDDR5, używane obecnie w najwydajniejszych kartach graficznych.

GDDR5 (Graphic Double Data Rate v5) - typ pamięci RAM przeznaczony dla kart graficznych, następca pamięci GDDR3 (producent firma Qimonda zrezygnowała z produkcji GDDR4 z powodu niewielkich różnic wydajności pomiędzy nimi, a GDDR3), charakteryzuje się trzykrotnie wyższą wydajnością niż pamięci GDDR3.

Pierwsze próbne egzemplarze tych pamięci pojawiły się w pierwszej połowie 2008 roku, około 2011 roku pamięci GDDR5 mają docelowo zająć obecną pozycję pamięci GDDR3. Zgodnie z informacjami podawanymi przez firmę Qimonda, jednego z największych producentów pamięci graficznych, wartość sprzedaży dla samego rynku PC dla tych pamięci ma osiągnąć 1,5 miliarda dolarów.

Nowość ma pojawić się w procesie technologicznym 75 nm, z efektywną częstotliwością taktowania 3,5 GHz w górę.
Pierwszą kartę graficzną Radeon HD 4870/4870x2 opartą o GDDR5 wprowadziła firma AMD/ATI.

SGRAM (ang. ***S**ynchronous **G**raphics **RAM*** - synchroniczna pamięć RAM - rodzaj pamięci RAM stosowany głównie w starszych kartach graficznych. Umożliwia szybką zmianę zawartości pamięci - jest to ważne, ponieważ karty graficzne muszą niekiedy wyświetlać wiele klatek obrazu na sekundę, a każda z nich wymaga całkowitej zmiany danych znajdujących się w pamięci karty graficznej. Pamięć ta może jednocześnie zapisywać oraz odczytywać dane.

Kość pamięci SGRAM