

Conținutul cursului

- 1. Introducere
- 2. Metode pentru operații de I/E
- 3. Magistrale
- 4. Module de extensie pentru sisteme înglobate
- 5. Afișaje ale calculatoarelor
- 6. Adaptoare grafice
- 7. Discuri optice

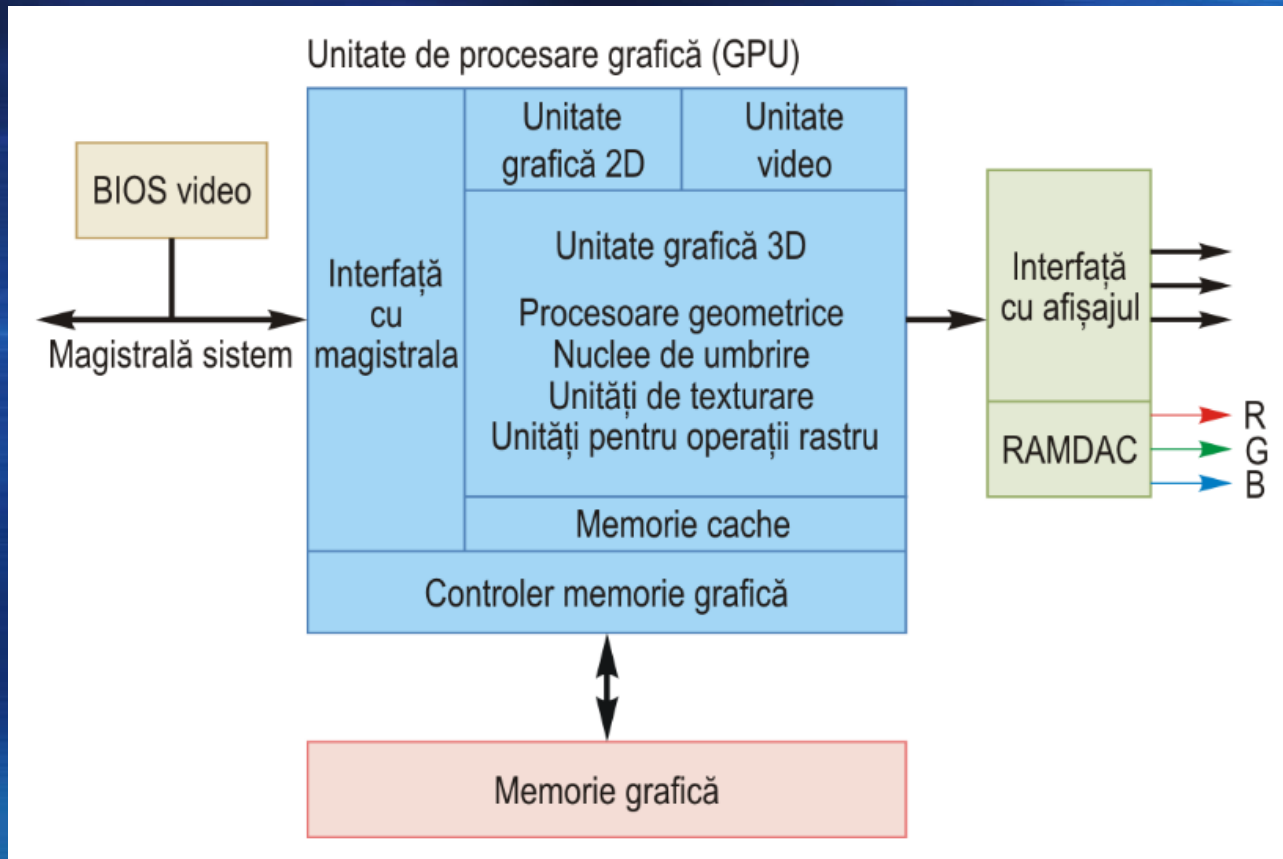
6. Adaptoare grafice

- Structura unui adaptor grafic
- Memoria grafică
- Unități de procesare grafică
- Interfețe pentru afișaje

Structura unui adaptor grafic (1)

- **Adaptor grafic:** generează imagini pentru un dispozitiv de afișare
 - **Funcții suplimentare:**
 - Accelerare pentru redarea graficii 2D și 3D
 - Decodificare și codificare video
 - Posibilitatea conectării mai multor afișaje
 - Adaptor grafic **dedicat**: implementat pe o placă de extensie
 - Adaptor grafic **integrat**: integrat într-un procesor sau circuit SoC (*System-on-Chip*)

Structura unui adaptor grafic (2)



Structura unui adaptor grafic (3)

- BIOS video
 - Inițializează adaptorul grafic
 - Conține parametri de funcționare pentru memoria grafică și procesorul grafic
 - Nu permite utilizarea tuturor facilităților adaptorului grafic
- Unitatea de procesare grafică (GPU – *Graphics Processing Unit* – GPU)
 - Implementează principalele funcții ale adaptorului grafic

Structura unui adaptor grafic (4)

- Interfața cu magistrala
 - Permite comunicarea cu UCP și memoria principală
 - Trebuie să permită transferuri în mod exploziv
- Unitatea grafică 2D
 - Implementează funcții pentru accelerarea interfețelor grafice ale SO și aplicațiilor
 - Transferuri pe blocuri de biți (BitBlt): instrucțiuni pentru mutarea datelor în bufferul de cadre
 - Desenarea figurilor geometrice
 - Umplerea poligoanelor cu o culoare specificată

Structura unui adaptor grafic (5)

- Unitatea video
 - Accelerare prin hardware pentru decodificare video, codificare video și trans-codare video
- Memoria *cache*
 - Organizată ca o ierarhie de memorii *cache*
 - Se poate utiliza ca memorie partajată de către nucleele de procesare ale GPU
 - Vor fi necesare mai puține accese la memoria grafică
- Controlerul memoriei grafice
 - Permite accesul la memoria grafică

Structura unui adaptor grafic (6)

- Unitatea grafică 3D
 - Execută operațiile necesare pentru redarea obiectelor 3D pe un afișaj 2D
 - Două etape: etapa geometrică, etapa de redare
 - Etapa geometrică: transformare, iluminare, decupare; executată de procesoare geometrice
 - Etapa de redare: rasterizare, umbrire, mixare alfa; executată de nuclee de umbrire
 - Unități de texturare: execută adresarea texturilor, filtrarea texturilor și maparea texturilor
 - Unități pentru operații rastru: creează pixelii finali

Structura unui adaptor grafic (7)

- Memoria grafică

- Conține **bufferul de cadre** care păstrează imaginea afișată pe ecran
- Mai conține: texturi, rezultate parțiale pentru nuclee de umbrire, programe compilate

- Interfața cu afișajul

- Permite transferul imaginilor grafice la un afișaj
- Poate fi o componentă discretă sau integrată în circuitul GPU

Structura unui adaptor grafic (8)

- Poate conține un circuit **RAMDAC** (*RAM Digital-to-Analog Converter*)
 - Preia imaginea digitală și o convertește în semnale analogice
 - Necesari numai pentru afișaje cu intrări analogice
- Diferite interfețe standard pentru afișaje
- **VGA** (*Video Graphics Array*)
 - Interfață analogică
 - Proiectată pentru afișaje cu tub catodic, dar utilizată și de unele afișaje cu cristale lichide

Structura unui adaptor grafic (9)

- **DVI** (*Digital Visual Interface*)
 - Interfață digitală
 - Conector **DVI-I** (semnale digitale și analogice) sau **DVI-D** (numai semnale digitale)
- **HDMI** (*High-Definition Multimedia Interface*)
 - Interfață digitală care permite transmiterea datelor video și audio pe același cablu
- **DisplayPort**
 - Interfață digitală pentru date video și audio
 - Utilizează un protocol bazat pe pachete

6. Adaptoare grafice

- Structura unui adaptor grafic
- Memoria grafică
- Unități de procesare grafică
- Interfețe pentru afișaje

Memoria grafică

- Memoria grafică
 - Tipuri de memorii grafice
 - Memoria GDDR6
 - Memoria GDDR7
 - Memoria HBM

Tipuri de memorii grafice (1)

- Pot fi cu port unic sau cu port dual
- Memorie grafică cu port unic
 - Singurul port de date este utilizat pentru a reîmprospăta ecranul și a înscrie noi date
- Memorie grafică cu port dual
 - Unul din porturi este utilizat pentru actualizarea imaginilor în memorie
 - Al doilea port este cu acces serial și este utilizat pentru reîmprospătarea imaginilor

Tipuri de memorii grafice (2)

- Exemple de memorii grafice
 - **VRAM** (*Video RAM*): DRAM cu port dual
 - **WRAM** (*Window RAM*): VRAM îmbunătățit
 - **SGRAM** (*Synchronous Graphics RAM*): variantă SDRAM → poate deschide două pagini de memorie
 - Variante **DDR**: DDR2, DDR3, DDR4, DDR5
 - **GDDR** (*Graphics Double Data Rate*)
 - Utilizează tensiuni mai reduse și frecvențe de ceas mai ridicate decât memoriile DDR
 - Mai multe generații: **GDDR2** .. **GDDR7**

Tipuri de memorii grafice (3)

- **HBM** (*High Bandwidth Memory*)
 - Interfață pentru circuite DRAM stivuite
 - Poate obține performanțe mai ridicate și consum mai redus decât o configurație 2D
 - Prima generație: **HBM** (2013)
 - Generația a doua: **HBM2** (2016)
 - Generația a doua îmbunătățită: **HBM2E** (2018)
 - Generația a treia: **HBM3** (2022), **HBM3E** (2023)
 - Generația a patra: **HBM4** (2025)

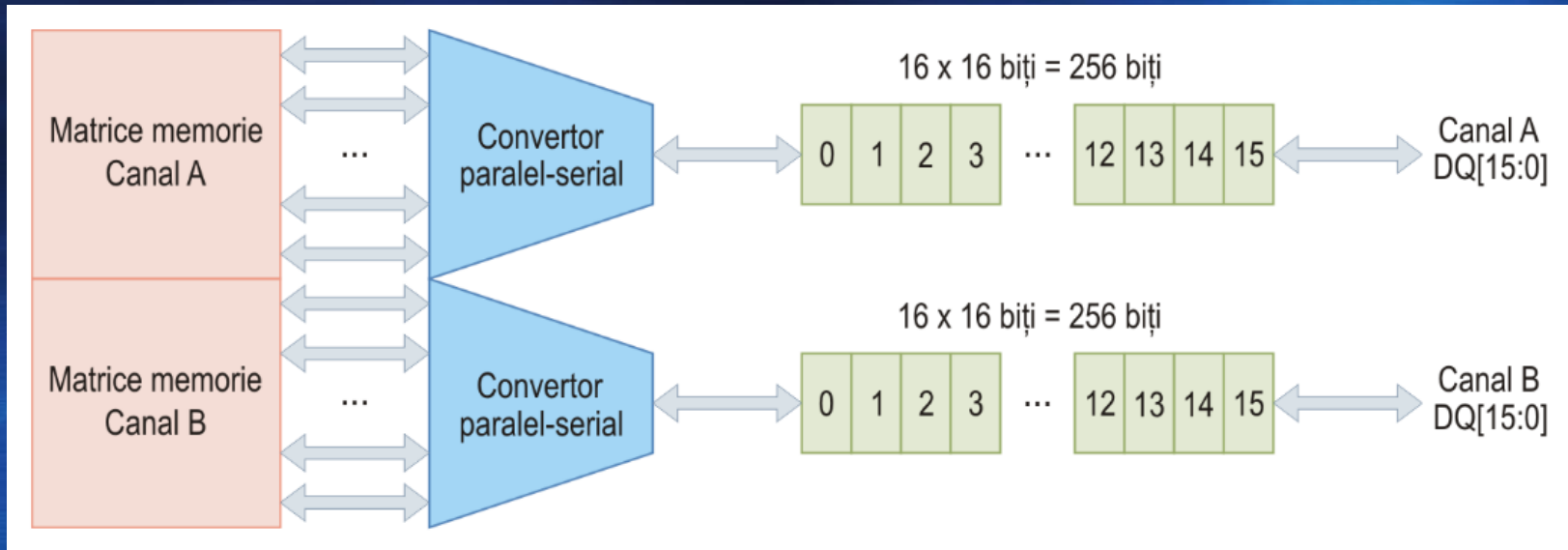
Memoria grafică

- Memoria grafică
 - Tipuri de memorii grafice
 - Memoria GDDR6
 - Memoria GDDR7
 - Memoria HBM

Memoria GDDR6 (1)

- Combină performanța ridicată cu funcționare stabilă și costuri de implementare reduse
- Organizarea memoriei: x32
- Utilizează o **preîncărcare 16n**
 - Magistrala internă de date are o lățime de 16x lățimea interfeței de I/E
- Interfața de I/E de 32 biți este divizată în **două canale independente de 16 biți**, A și B
 - Pe fiecare canal, o operație de citire sau scriere transferă $16 \times 16 = 256$ biți

Memoria GDDR6 (2)



- Rata de transfer pentru un timp de acces de 1 ns la matricea de memorie: 16 Gbiți/s pe pin
- Pentru un circuit de memorie x32: 64 GB/s
- Max.: 24 Gbiți/s pe pin; 96 GB/s pe circuit

Memoria GDDR6 (3)

- Semnale de ceas
 - Semnal diferențial de ceas pentru comenzi (CK)
 - 2 sau 4 semnale de ceas pentru scriere (WCK0_A, WCK0_B, WCK1_A, WCK1_B)
 - Doi octeți sunt aliniați la un semnal WCK, sau fiecare octet este aliniat la un semnal WCK
 - Mod QDR (*Quad Data Rate*): frecvența semnalelor WCK (f_{WCK}) este dublată intern
 - Mod DDR (*Double Data Rate*): f_{WCK} nu este dublată

Memoria GDDR6 (4)

● Tehnici utilizate

- Inversarea magistralei de date
 - Reduce numărul biților de zero transmiși
 - Indică cu un semnal **DBI#** pentru fiecare octet
 - Liniile de transmisie au terminatoare la nivel logic ridicat → se reduce puterea disipată
- Inversarea magistralei de comenzi/adrese CA
- Antrenarea semnalelor
 - Ajustarea fazei pentru semnalele de ceas, date și adrese

Memoria GDDR6 (5)

- **Antrenarea liniilor de adrese:** alinierea liniilor de adrese la semnalul de ceas **CK**
- **Antrenarea liniilor de date:** alinierea liniilor de date la semnalul **WCK** corespunzător
- Alinierea fiecărui semnal **WCK** la semnalul **CK**
- Este posibilă o re-antrenare “ascunsă” a datelor
- **Calibrarea:** îmbunătățește fiabilitatea
 - **Auto-calibrare:** curentul generat, impedanța de terminare
 - Ajustare controlată prin software

Memoria GDDR6 (6)

- Detectarea erorilor
 - Pini **EDC** (*Error Detection Code*) dedicați pentru transmiterea codurilor CRC la controler
 - Pentru fiecare canal: cod CRC de 16 biți pentru datele și semnalele **DBI#** ale unui bloc de transfer exploziv
 - **Cod CRC**: concatenarea a două coduri CRC de 8 biți, pentru prima și a doua jumătate a blocului
 - Permite detectarea completă a erorilor aleatoare de 1, 2 și 3 biți, și detectarea cu >99% a altor erori aleatoare

Memoria GDDR6 (7)

- **Gestiunea energiei consumate**

- Frecvența semnalului de ceas este scalabilă: de la 50 MHz (400 Mbiți/s) la frecvența maximă
- Mod cu consum redus pentru nucleul DRAM
- Creșterea impedanței de terminare la rate reduse ale datelor
- Moduri cu **auto-reîmprospătare** și **hibernare cu auto-reîmprospătare**
- Tensiune de alimentare redusă: 1,35 V
- Inversarea magistralei de date și de adrese

Memoria grafică

- Memoria grafică
 - Tipuri de memorii grafice
 - Memoria GDDR6
 - Memoria GDDR7
 - Memoria HBM

Memoria GDDR7

- Utilizează o arhitectură cu **preîncărcare $32n$**
 - 4 canale independente de 8 biți
- **Modulație în amplitudine a impulsurilor** (*Pulse Amplitude Modulation* – PAM)
 - Trei nivele de tensiune (**PAM3**): 3 biți/2 cicluri
 - Mai eficientă decât modulația **NRZ** (*Non-Return to Zero*) utilizată de memoria **GDDR6**
- Rata de transfer: 32 Gbiți/s pe pin (primele circuite); până la 48 Gbiți/s pe pin
 - 128 GB/s pe circuit; max. 192 GB/s pe circuit

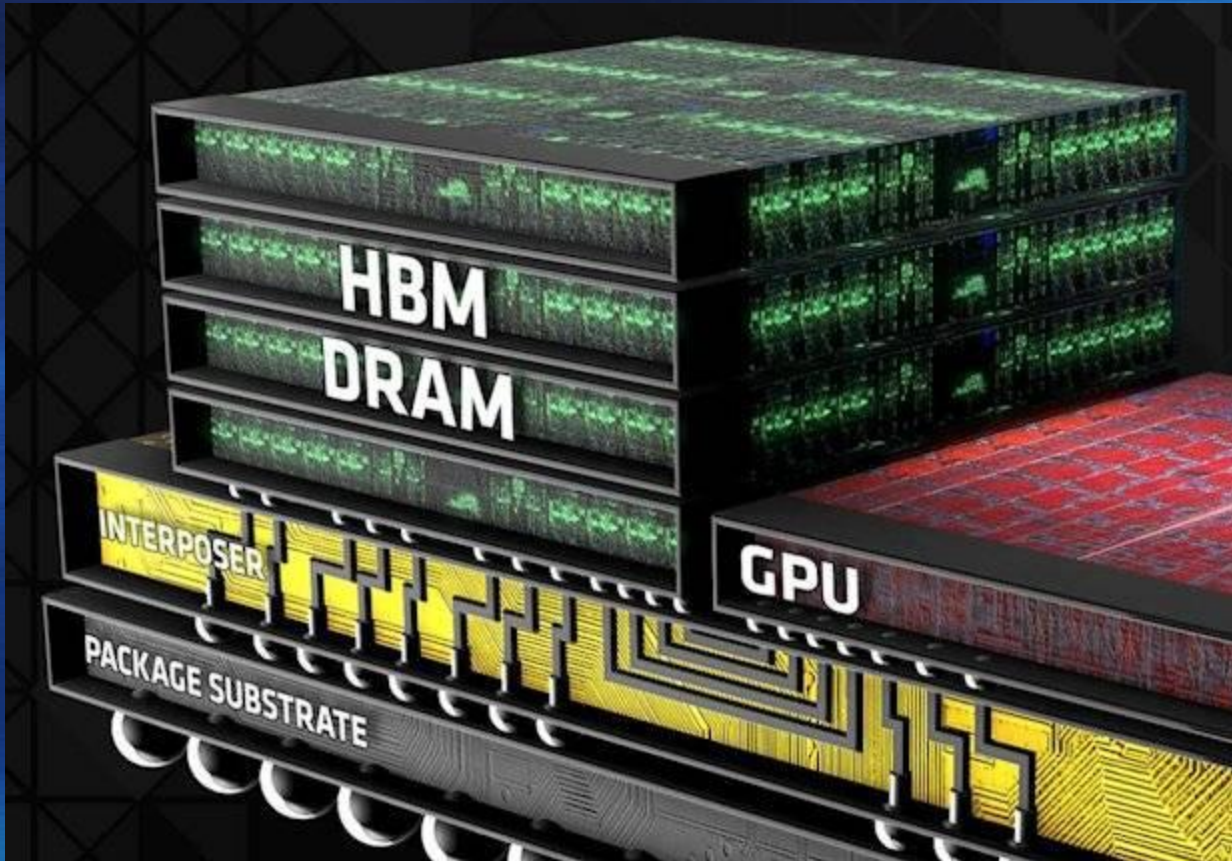
Memoria grafică

- Memoria grafică
 - Tipuri de memorii grafice
 - Memoria GDDR6
 - Memoria GDDR7
 - Memoria HBM

Memoria HBM (1)

- Este formată din una sau mai multe **stive de memorie**
 - Până la 8, 12 sau 16 straturi de circuite DRAM
 - Strat conținând un controler de memorie
 - Interconexiuni verticale
- Conectarea la UCP/GPU/SoC prin interfețe fizice și un strat de siliciu → “*interposer*”
- Stivele de memorie, UCP/GPU/SoC și stratul de siliciu sunt împachetate într-o **singură capsulă**

Memoria HBM (2)



© Advanced Micro Devices Inc.

Memoria HBM (3)

- Prima generație (HBM)
 - Funcționare la 1,3 V, 500 MHz
 - Rata de transfer: 1 GT/s pe linie (pin)
 - Fiecare circuit DRAM din stivă comunică prin canale de 128 biți
 - Opt canale pe stiva de memorie
 - Lățimea interfeței: $8 \times 128 = 1024$ biți
 - Rata: 128 GB/s pe circuit de memorie
 - Capacitatea: până la 4 GB pe circuit

Memoria HBM (4)

- Generația a doua (HBM2)
 - Funcționare la 1,2 V, 1 GHz
 - Rata de transfer: 2 GT/s pe pin
 - Rata: 256 GB/s pe circuit de memorie
 - Capacitatea: până la 8 GB pe circuit
- Generația a doua îmbunătățită (HBM2E)
 - Rata de transfer: 3,2 GT/s pe pin
 - Rata: 409,6 GB/s pe circuit de memorie
 - Până la 12 straturi, până la 24 GB pe circuit

Memoria HBM (5)

- Generația a treia (HBM3)
 - 16 canale de 64 biți ($16 \times 64 = 1024$)
 - Rata de transfer: 6,4 GT/s pe pin
 - Rata: 819,2 GB/s pe circuit de memorie
 - Până la 48 GB (12 straturi), 64 GB (16 straturi)
- Generația a treia îmbunătățită (HBM3E)
 - Rata de transfer: 8 GT/s pe pin (1 TB/s totală)
 - Micron: 9,2 GT/s (1,15 TB/s pe circuit)
 - Samsung: 9,8 GT/s (1,22 TB/s pe circuit)

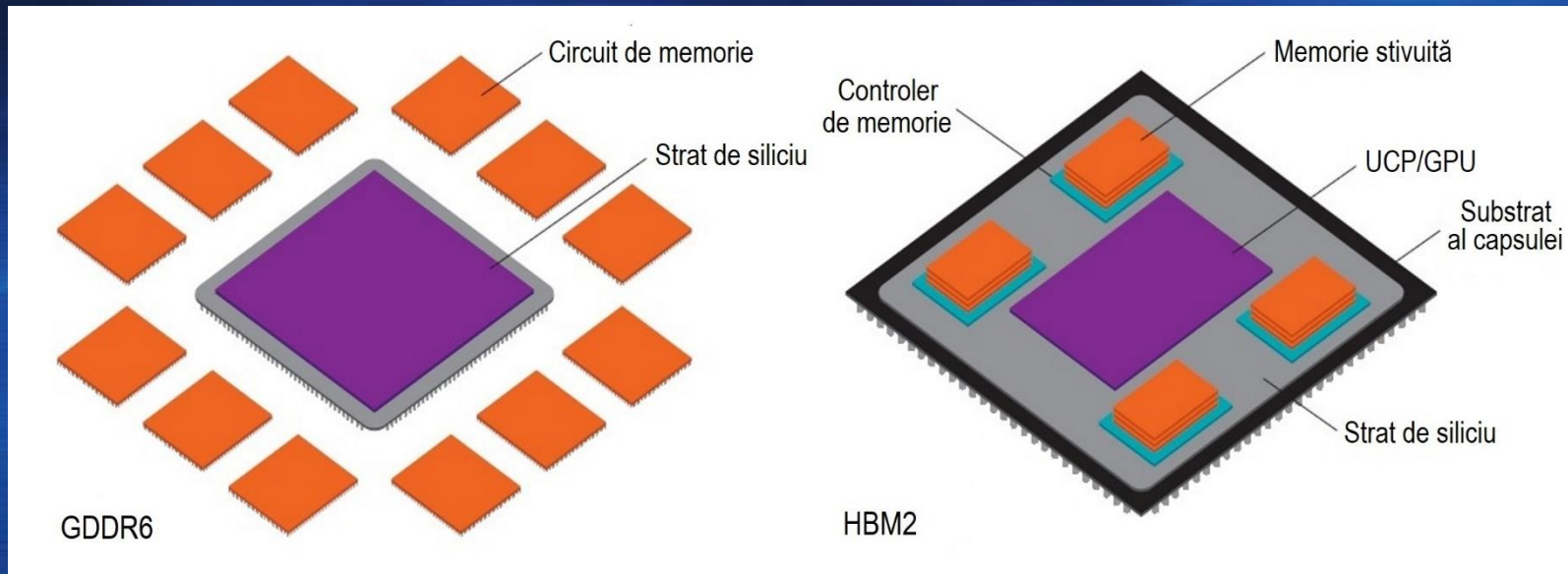
Memoria HBM (6)

- Generația a patra (HBM4)
 - Tensiunea de funcționare: 1,0 V sau 1,05 V
 - Numărul de canale pe stiva de memorie a fost dublat, la 32
 - Lățimea interfeței: 32×64 biți = 2048 biți
 - Rata de transfer: 8 GT/s pe pin
 - Rata: 2 TB/s pe circuit de memorie
 - Compatibilitate cu controlere HBM3 existente
 - 4, 8, 12 sau 16 straturi, până la 64 GB pe circuit

Memoria HBM (7)

- **Avantaje** față de o memorie **GDDR**
 - Rata de transfer mai ridicată
 - **HBM3**: de 8,5x mai ridicată față de **GDDR6**
 - **HBM4**: de 10,7x mai ridicată față de **GDDR7**
 - Permite reducerea consumului de energie
 - Necesită spațiu mai redus
 - O memorie **GDDR6** cu 4 circuite necesită o suprafață de 20x mai mare față de o stivă **HBM** cu 4 straturi
- **Dezavantaj**: cost mai ridicat

Memoria HBM (8)



Imaginea originală © Advanced Micro Devices Inc.

6. Adaptoare grafice

- Structura unui adaptor grafic
- Memoria grafică
- Unități de procesare grafică
- Interfețe pentru afișaje

Unități de procesare grafică

- Unități de procesare grafică
 - Prezentare generală
 - Etaje de prelucrare
 - Arhitectură cu paralelism de date
 - Calcul GPGPU
 - Arhitectura CUDA

Prezentare generală (1)

- GPU – *Graphics Processing Unit*
- Procesoare grafice dedicate pentru console de jocuri, calculatoare PC, telefoane
 - Utilizate inițial pentru accelerarea etapei de redare a graficii 3D (de ex., texturare)
 - Ulterior utilizate și pentru accelerarea calculelor geometrice (de ex., translatare)
- GPU conțin numeroase nuclee de umbrire, unități de texturare, unități pentru operații rastru

Prezentare generală (2)

- Nuclee de umbrire
 - Operează cu vârfuri, primitive sau pixeli
 - Execută **programe de umbrire**
- GPU din generațiile precedente conțineau nuclee de umbrire specializate
- Nuclee de umbrire pentru vârfuri
 - Utilizate pentru adăugarea efectelor speciale unor obiecte 3D
 - Operează cu datele vârfurilor obiectelor: culoare, textură, iluminare
 - Exemple: animație, deformarea suprafețelor

Prezentare generală (3)

- Nuclee de umbrire geometrică
 - Nu sunt prezente întotdeauna
 - **Intrări:** vârfurile unei primitive asamblate într-o etapă precedentă
 - Pot adăuga, elimina sau modifica primitive
- Nuclee de umbrire pentru pixeli/fragmente
 - Utilizate pentru adăugarea unor lumini, umbre și efecte de textură pixelilor individuali
 - Permit generarea scenelor complexe, realiste

Prezentare generală (4)

- Nuclee de umbrire unificate
 - Pot executa diferite operații de umbrire (vârfuri, geometrice, pixeli/fragmente)
 - GPU conțin o matrice de nuclee de umbrire unificate și o unitate de planificare dinamică
 - Nucleele de umbrire unificate permit o utilizare mai flexibilă a resurselor hardware
- Nuclee de umbrire pentru calcule
 - Pun la dispoziție resurse generale de calcul
 - Utilizează numărul mare de sisteme *pipeline* de umbrire disponibile

Prezentare generală (5)

- GPU pot fi **dedicate** sau **integrate**
- **GPU dedicate**
 - Utilizate în plăcile grafice → conțin o memorie grafică dedicată pentru GPU
 - Comunică cu UCP prin magistrala sistem, de obicei, o magistrală **PCI Express**
 - **Exemple**
 - AMD **Radeon RX 9000** (ex., RX 9070)
 - NVIDIA **GeForce RTX 50** (ex., GeForce RTX 5080)



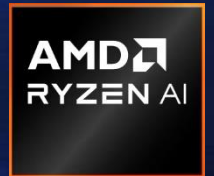
Prezentare generală (6)

● GPU integrate

- Sunt integrate într-un set de circuite sau circuit SoC
- Utilizează o parte a memoriei sistem
- Performanțe mai reduse față de GPU dedicate

● Exemple

- Intel **X^e** (ex., Iris X^e Graphics)
- AMD **Radeon 890M** din seria de circuite SoC Ryzen AI 9 HX 370
- ARM **Mali** (ex., Mali-G715 din setul de circuite Google Tensor G4)



Prezentare generală (7)

- Soluții hibride

- Plăci grafice hibride

- Partajează memorie cu sistemul, dar au și o memorie dedicată de dimensiuni mai mici
 - S-au dezvoltat tehnologii pentru accesul la memoria sistem (de ex., prin mag. PCI Express)

- Integrarea unei memorii *Embedded DRAM* (eDRAM) în aceeași capsulă cu GPU

- **Exemple:** GPU integrate Intel **Iris Graphics**, cu 64 MB sau 128 MB de memorie eDRAM

Prezentare generală (8)

- Arhitectură cu memorie unificată
 - Utilizată de circuite SoC și console de jocuri
 - GPU poate utiliza orice parte a memoriei sistem pentru diferite buffere și texturi
 - **Example:** circuite SoC Intel
 - GPU integrat are propriul set de memorii *cache* **L1** și **L2**, și o memorie *cache* **L3**
 - **Memorie *cache* de ultim nivel (LLC – Last-Level Cache):** partajată între nucleele GPU și UCP
 - Ierarhie de memorii *cache*: reduce latența memoriei și rata de transfer necesară

Prezentare generală (9)

- Proiectarea GPU a fost influențată de specificații ale bibliotecilor grafice 2D și 3D
 - Producătorii au creat implementări utilizând accelerarea hardware
 - Teste speciale calificau o implementare ca fiind conformă cu o anumită specificație
 - A rezultat o corespondență între facilitățile oferite prin accelerare hardware și cele oferite de bibliotecile grafice
 - **Exemple:** OpenGL, DirectX, Vulkan

Prezentare generală (10)

- **OpenGL** (*Open Graphics Library*)



- Specificație pentru platforme și limbaje multiple
- Întreținută de Khronos Group
- Definește funcții pentru specificarea programelor de umbrire și a datelor acestora
- **OpenGL ES**: subset pentru grafica 2D și 3D pe sisteme înglobate
- **WebGL** (*Web Graphics Library*): redare grafică într-un browser web utilizând JavaScript

Prezentare generală (11)



● DirectX

- Dezvoltat de Microsoft pentru platforme Windows și console de jocuri Xbox
- Conține interfețele de programare (API) **Direct2D**, **Direct3D** și **DirectCompute**
- **DirectX 12** (2015) reprezintă o reproiectare radicală pentru arhitecturi GPU moderne
 - Reduce întârzierea datorată driverului
 - Îmbunătățește paralelismul prin utilizarea mai multor nuclee UCP pentru grafică

Prezentare generală (12)



● Vulkan

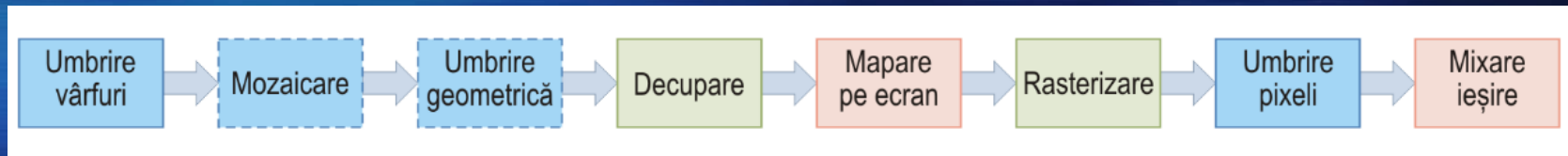
- Interfață de programare pentru grafică 3D și calcule de uz general
- Versiunea originală donată de AMD consorțiului Khronos Group
- Poate beneficia de avantajele GPU moderne și ale UCP cu nuclee multiple
- Oferă suport îmbunătățit pentru UCP cu fire de execuție multiple și reduce încărcarea UCP
- Compatibil cu majoritatea SO importante

Unități de procesare grafică

- Unități de procesare grafică
 - Prezentare generală
 - Etaje de prelucrare
 - Arhitectură cu paralelism de date
 - Calcul GPGPU
 - Arhitectura CUDA

Etaje de prelucrare (1)

- Organizate într-un sistem *pipeline* logic
- Diferite tipuri de etaje hardware
 - Etaje **programabile** (albastru deschis)
 - Etaje **configurabile** (roz)
 - Etaje cu **funcții fixe** (verde deschis)
 - Etaje opționale: mozaicare, umbrire geometrică



Etaje de prelucrare (2)

- Umbrire vârfuri

- Operează asupra vârfurilor individuale ale **primitivelor** (puncte, linii, triunghiuri)
- Utilizat pentru a modifica sau crea valori asociate cu fiecare vârf: poziție, culoare, textură
- Implementează operații de **calcul geometric**
 - **Transformare**: translatare, rotire, scalare
 - Poziționarea camerei de vizualizare în originea sistemului de coordonate
 - **Iluminare**: determinarea efectului surselor de lumină asupra materialului unui obiect

Etaje de prelucrare (3)

● Mozaicare

- Permite redarea suprafețelor curbate
- Convertește descrierea unei suprafețe de ordin superior într-un set de triunghiuri
- **Avantaje:**
 - Descrierea unei suprafețe curbate este mai compactă
 - Reduce memoria și rata de transfer necesare
 - Permite tehnici de redare scalabile → nivel de detaliere dependent de poziția camerei

Etaje de prelucrare (4)

- **Umbrire geometrică**
 - Permite conversia primitivelor în alte primitive
 - **Intrări**: vârfurile unei singure primitive
 - **Ieșiri**: zero sau mai multe vârfuri formând o singură topologie (de ex., listă de puncte)
- **Decupare**
 - Elimină primitivele și părțile primitivelor care sunt în afara volumului de vizualizare

Etaje de prelucrare (5)

- Mapare pe ecran

- Transformă coordonatele 3D ale primitivelor în coordonate 2D de pe ecran

- Rasterizare

- Determină toți pixelii care sunt în interiorul unei primitive
- **Setare triunghi:** calculează datele necesare pentru un anumit algoritm de rasterizare
- **Traversare triunghi:** determină pixelii care sunt în interiorul triunghiului → generează fragmente

Etaje de prelucrare (6)

● Umbrire pixeli

- Execută calculele de umbrire asupra pixelilor
- **Intrări:** rezultate de la etajul de umbrire vârfuri, interpolate de etajul de traversare a triunghiului
- **Ieșiri:** culoarea, valoarea opacității și valoarea adâncimii pentru un pixel
- **Texturare:** modificarea aparenței unei suprafețe utilizând o imagine sau funcție textură
- **Filtrare textură:** reduce efectele nedorite atunci când o textură este mai mică sau mai mare decât zona pixelilor care trebuie texturați

Etaje de prelucrare (7)

● Mixare ieșire

- Mixează culoarea pixelului cu culoarea curentă stocată în **bufferul de culori**
- Etaj numit și **unitate pentru operații rastru**
- Execută **teste de vizibilitate** utilizând **bufferul z**
- Pentru fiecare pixel, se stochează și o **valoare de transparență (alfa)**
- Bufferul de culori stochează o valoare $RGB\alpha$ pentru fiecare pixel (culoare și valoare alfa)
- **Mixare alfa**: mixarea culorii unui obiect transparent cu culoarea obiectului din spate

Unități de procesare grafică

- Unități de procesare grafică
 - Prezentare generală
 - Etaje de prelucrare
 - Arhitectură cu paralelism de date
 - Calcul GPGPU
 - Arhitectura CUDA

Arhitectură cu paralelism de date (1)

- Arhitectura GPU diferă de arhitectura UCP
 - Arhitectura GPU: optimizată pentru rata de prelucrare
 - Rata de prelucrare: numărul de operații executate într-un timp dat
 - GPU conține mii de nuclee de procesare simple
 - Arhitectura UCP: optimizată pentru latență
 - Latența: întârzierea de la inițierea unei comenzi până când efectul acesteia devine detectabil
 - UCP are puține nuclee de procesare
 - UCP dedică un spațiu larg pentru memorii cache

Arhitectură cu paralelism de date (2)

- UCP implementează tehnici complexe: preîncărcarea instrucțiunilor, predicția salturilor
- Unitatea de control a UCP devine complexă
- **GPU**: rată ridicată, dar latență ridicată
- **UCP**: latență redusă, dar rată redusă



Arhitectură cu paralelism de date (3)

- GPU utilizează execuția paralelă la o scară mai extinsă decât UCP
 - Nucleele de umbră prelucrează numeroase fire de execuție în paralel
 - Firele sunt în mare măsură independente
 - Tip de paralelism: **paralelism de date**
 - Arhitectura: **SIMD** (*Single Instruction, Multiple Data*), **SIMT** (*Single Instruction, Multiple Thread*)
- Un fir de execuție GPU este simplu
 - Utilizează o mică memorie locală și registre

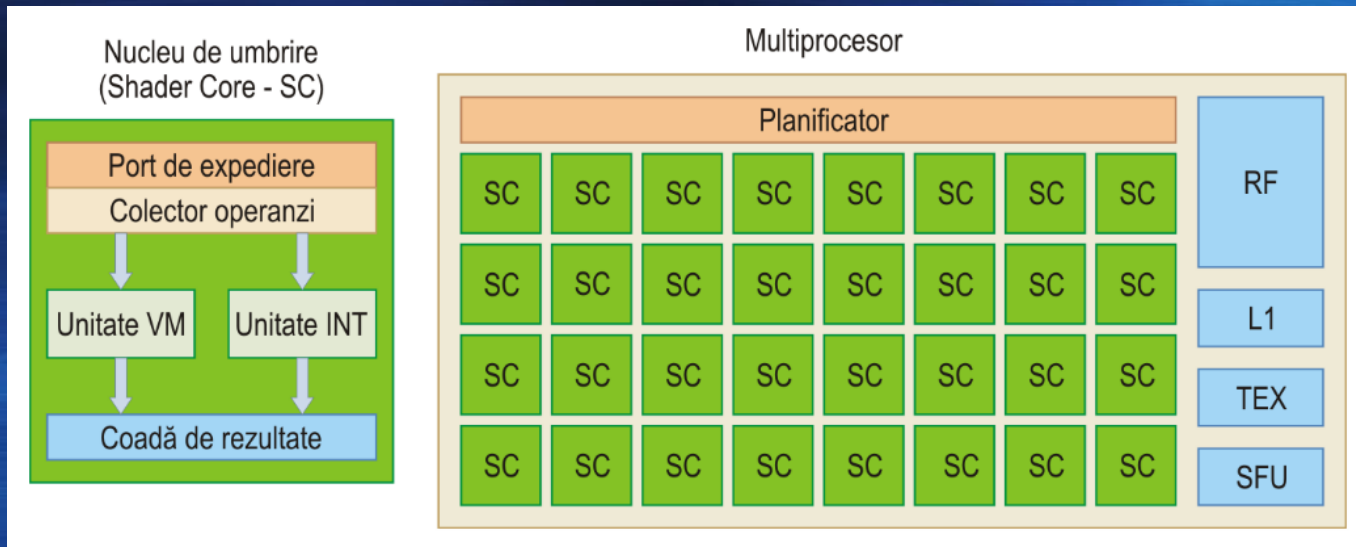
Arhitectură cu paralelism de date (4)

- Firele care execută același program de umbrire sunt grupate în **grupuri de fire** (de ex., *urzeli*)
 - O instrucțiune de **citire a memoriei** este întâlnită de toate firele dintr-un grup în același timp
 - Toate firele dintr-un grup sunt suspendate și **grupul este schimbat** cu un alt grup
 - Comutarea între două grupuri de fire este rapidă → nu sunt afectate datele niciunui fir de execuție
 - **Comutarea rapidă** între grupurile de fire ascunde latența memoriei

Arhitectură cu paralelism de date (5)

- Nucleu de umbrire unificat
 - Execută un program de umbrire
 - Componente: decodificator de instrucțiuni, colector operanzi, UAL, colector rezultat
 - Denumiri: **procesor de șir** (AMD); *pipeline* (Imagination Techn.); **nucleu CUDA** (NVIDIA)
 - Mai multe nuclee (de ex., 32) sunt grupate într-o **unitate SIMD** → **multiprocesor** (MP)
 - Denumiri: **unitate de calcul** (AMD); **unitate de execuție** (Intel), **multiprocesor de șir** (NVIDIA)

Arhitectură cu paralelism de date (6)



- **Planificator**: asignează firele nucleelor de umbră
- **RF**: set de registre
- **L1**: memorie *cache* de nivel 1
- **TEX**: unități de texturare
- **SFU**: unitate pentru funcții speciale

Rezumat (1)

- Un **adaptor grafic** generează imagini pentru un dispozitiv de afișare
 - Componente importante: unitatea de procesare grafică, memoria grafică, interfața cu afișajul
- **Memoria grafică** conține bufferul de cadre, alte buffere de date și programe de umbrire
 - Poate fi cu **port unic** sau cu **port dual**
 - Memorie cu **port dual**: actualizarea imaginilor și reîmprospătarea ecranului se pot executa în paralel
 - Exemple de tehnologii pentru memoria grafică: **GDDR, HBM**

Rezumat (2)

- Memoria **GDDR6** are facilități avansate pentru performanțe ridicate și funcționare stabilă
 - Inversarea liniilor de date și adrese; antrenarea semnalelor; calibrare; detecția erorilor
- Memoriile **HBM3** și **HBM4** au rate de transfer ridicate și necesită un spațiu redus
- **GPU** accelerează etapa **geometrică** și cea de **redare** a graficii 3D
 - Conține numeroase nuclee de umbrire, unități de texturare și unități pentru operații rastru
 - GPU actuale conțin **nuclee de umbrire unificate**

Rezumat (3)

- GPU pot fi **dedicate** sau **integrate** într-un set de circuite
- Proiectarea GPU a fost influențată de specificațiile **bibliotecilor grafice 2D și 3D**
- GPU conține **etaje de prelucrare** programabile, configurabile și cu funcții fixe
- **Arhitectura GPU** este diferită de **arhitectura UCP**
- GPU utilizează arhitecturi **SIMD/SIMT** și paralelism de date
- Firele de execuție sunt grupate și sunt comutate rapid pentru a ascunde latența memoriei

Noțiuni, cunoștințe (1)

- Structura unui adaptor grafic
- Componentele unei unități de procesare grafică
- Tipuri de memorii grafice
- Caracteristici ale memoriei GDDR6
- Tehnici utilizate de memoria GDDR6
- Facilități de gestiune a energiei consumate la memoria GDDR6
- Generații ale memoriei HBM
- Avantaje ale memoriei HBM

Noțiuni, cunoștințe (2)

- Tipuri de nuclee de umbrire ale GPU
- GPU dedicate și integrate
- Etaje de prelucrare ale GPU
- Deosebiri între arhitectura GPU și arhitectura UCP
- Paralelismul de date al arhitecturii GPU
- Structura unui nucleu de umbrire unificat
- Structura unui multiprocesor al GPU