INTERFAȚA AGP

Zoltan Baruch
Catedra de Calculatoare, Universitatea Tehnică din Cluj-Napoca
E-mail: Zoltan.Baruch@cs.utcluj.ro

Necesitatea unor rate ridicate de transfer între procesor și subsistemul video a condus la apariția magistralelor locale ale calculatoarelor personale, începând cu magistrala VL Bus (*VESA Local Bus*) și continuând cu magistrala PCI (*Peripheral Component Interconnect*). La fel cum s-a întâmplat și în cazul magistralei ISA, traficul pe magistrala PCI a calculatoarelor performante a devenit foarte intens, la acest trafic contribuind adaptorul video, discul fix și alte periferice care sunt conectate la aceeași magistrală PCI. Pentru a se evita saturarea magistralei PCI din cauza informațiilor video, *Intel* a creat o nouă interfață, proiectată special pentru subsistemul video. Această interfață este numită AGP (*Accelerated Graphics Port*). În acest articol se prezintă diferite aspecte legate de interfața AGP, incluzând principiul de funcționare, transferul datelor, maparea memoriei și aspecte software.

Principiul AGP

AGP este o nouă interconexiune pentru acceleratoarele grafice din sistemele bazate pe procesorul *Pentium II*, utilizate în special pentru grafică 3D și redarea secvențelor video. Utilizatorii calculatoarelor *PC* pot beneficia acum de tipul de grafică 3D și video disponibile în prealabil numai pe stațiile de lucru.

Procesorul *Pentium II* constă dintr-un nucleu încapsulat cu o memorie cache integrată de nivel 2 (L2). Acest procesor dispune de asemenea de o arhitectură *Dual Independent Bus* (DIB), în care două magistrale independente conectează nucleul cu memoria cache L2 și cu magistrala sistem a calculatorului. Faptul că ambele magistrale pot funcționa în același timp îmbunătățește semnificativ performanțele procesorului, deoarece procesorul poate executa instrucțiuni din memoria cache L2 și simultan poate comunica cu dispozitive externe.

Noile aplicații grafice 3D impun cerințe riguroase calculatoarelor *PC*, cuprinzând calcule geometrice mai rapide, o interpretare grafică mai sofisticată, și texturi mai detaliate. Cu toate că procesorul *Pentium II* este adaptat pentru a executa calcule geometrice sporite (cu o rată mai mare de triunghiuri pe secundă), iar generația viitoare de controlere grafice poate implementa o mare varietate de efecte grafice, dimensiunea crescută a texturilor a devenit o problemă importantă.

O problemă o reprezintă dimensiunea memoriei video utilizată de controlerele grafice. În mod tipic, această memorie are o dimensiune de 2-4 MB. Cu toate acestea, au început să apară aplicații grafice care utilizează peste 20 MB pentru o singură textură. Memoria video poate fi extinsă pentru a satisface aceste cerințe, dar o asemenea soluție este foarte costisitoare.

A doua problemă este rata de transfer permisă de magistrala PCI. Controlerele grafice trebuie să încarce în prealabil texturile din memoria sistem în memoria lor RAM locală. Deoarece dimensiunea texturilor a crescut, magistrala PCI a început să devină congestionată. Problema este chiar mai acută în cazul aplicațiilor care implică redarea secvențelor video.

Tehnologia AGP îmbunătățește performanțele sistemului punând la dispoziție o cale rapidă între controlerul grafic și memoria sistem. Această cale permite

controlerului grafic să facă acces la texturi direct în memoria sistem în timpul interpretării grafice, în loc să le încarce în prealabil în memoria video locală (Figura 1).

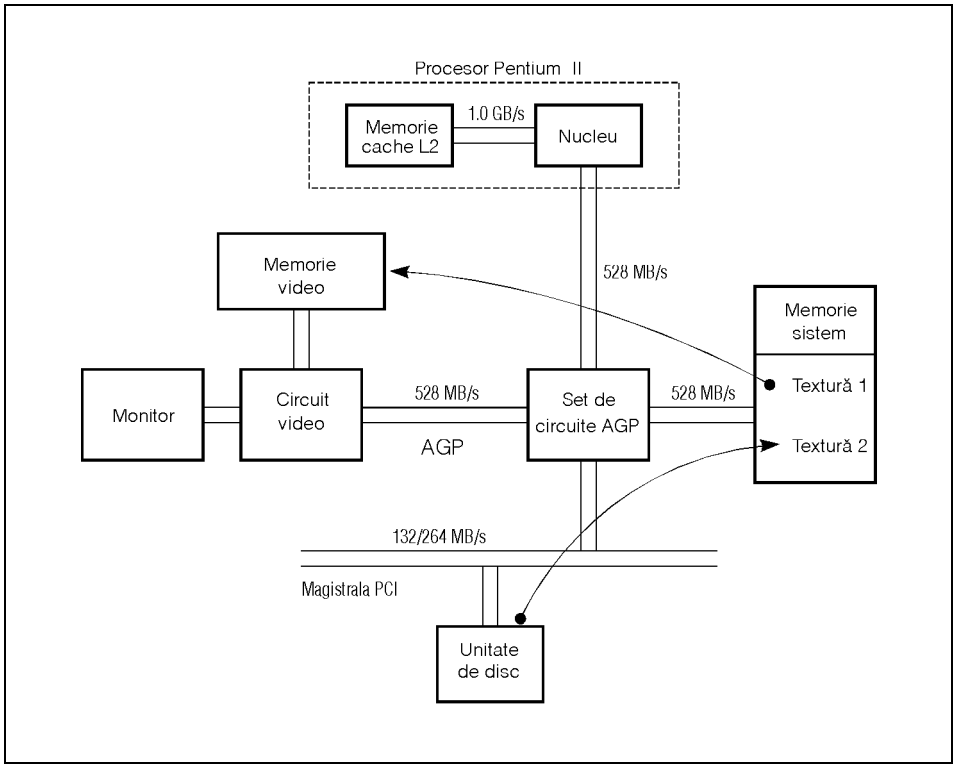

Figura 1. Transferuri de date pentru maparea texturilor la interfața AGP.

Sistemul de operare (SO) poate rezerva în mod dinamic segmente din memoria sistem, pentru a fi utilizate de controlerul grafic. Această memorie este numită memorie AGP (sau memorie video care nu este locală). Ca urmare, controlerul grafic va trebui să păstreze un număr mai mic de texturi în memoria video locală. Aceasta permite rezoluții mai mari ale ecranului, sau permite utilizarea unui *buffer Z* pentru o dimensiune dată a ecranului. Această tehnică elimină de asemenea restricția de dimensiune pe care memoria video locală o impune asupra texturilor, și deci permite aplicațiilor să utilizeze texturi de dimensiuni mult mai mari, îmbunătățind în plus realismul și calitatea imaginilor.

Mai mult, noua cale elimină de pe magistrala PCI traficul intens 3D și video. Descărcarea datelor grafice și a celor video de pe magistrala PCI permite conectarea altor dispozitive rapide pe magistrală.

AGP este un port, și nu o magistrală, deoarece la o magistrală se pot conecta mai multe dispozitive, în timp ce AGP este o conxiune punct la punct doar între adaptorul video și procesorul sistemului.

AGP este o interfață de 64 biți care poate funcționa la 66 MHz. Specificațiile AGP se bazează pe extensia de 64 biți a specificațiilor PCI 2.1, care descriu și un mod de lucru cu o frecvență de 66 MHz, care nu a fost implementat niciodată. AGP este implementat cu un conector similar celui utilizat pentru magistrala PCI, cu 32 de linii pentru adrese și date multiplexate. Există 8 linii suplimentare pentru adresarea secundară (*sideband*), descrisă în secțiunea următoare. Plăcile de bază AGP au un singur conector de extensie pentru adaptorul video AGP, și au de obicei cu un conector PCI mai puțin, în rest fiind similare cu plăcile de bază PCI.

Interfața AGP funcționează la viteza maximă a magistralei sistem, spre deosebire de magistrala PCI care funcționează la jumătatea acestei viteze. Aceasta înseamnă că la o placă de bază standard *Pentium II*, AGP funcționează la 66 MHz în locul frecvenței de 33 MHz a magistralei PCI. Astfel se dublează rata de transfer a portului. În locul limitei de 133 MB/s a magistralei PCI, în modul său cu viteza minimă

AGP are o rată de transfer de 266 MB/s. În plus, are avantajul că nu trebuie să partajeze rata de transfer cu alte dispozitive PCI.

Există mai multe cerințe pentru ca un sistem să poată utiliza avantajele AGP:

- Placă de bază cu un set de circuite AGP (de exemplu setul 440LX al *Intel* pentru procesorul *Pentium II*).
- Sistem de operare cu drivere pentru noua interfață (*Windows 98*).
- Drivere speciale ale adaptorului video pentru interfața AGP, care pot utiliza modul 2X al acesteia.

Moduri de transfer a datelor

Pe lângă dublarea vitezei magistralei, AGP a definit un mod 2X, care utilizează un protocol special pentru a permite transmiterea unui volum dublu de date prin port la aceeași frecvență de ceas. Creșterea de viteză este obținută prin transferarea datelor atât pe frontul crescător, cât și pe cel descrescător al ceasului de 66 MHz, și prin utilizarea modurilor de transfer a datelor care sunt mai eficiente. Rezultatul este că performanțele se dublează din nou, la o rată de transfer la vârf de 533 MB/s. Rata de transfer efectivă variază la diferite sisteme și aplicații, dar de obicei sistemele pot atinge în jur de 50-80% din valorile la vârf în cazul transferurilor prelungite. Există și o intenție de a implementa un mod 4X, ceea ce ar însemna o rată de transfer de 1.07 GB/s.

AGP pune la dispoziția controlerului grafic două moduri pentru accesul direct al texturilor în memoria sistem: modul *pipeline* și *adresarea secundară* (*sideband addressing*). În cazul modului *pipeline*, AGP suprapune timpii de acces ai memoriei și ai magistralei pentru o cerere *n* cu generarea cererilor următoare (*n*+1, *n*+2 etc). În cazul magistralei PCI, cererea *n*+1 nu începe până când nu se termină transferul de date al cererii *n* (Figura 2).

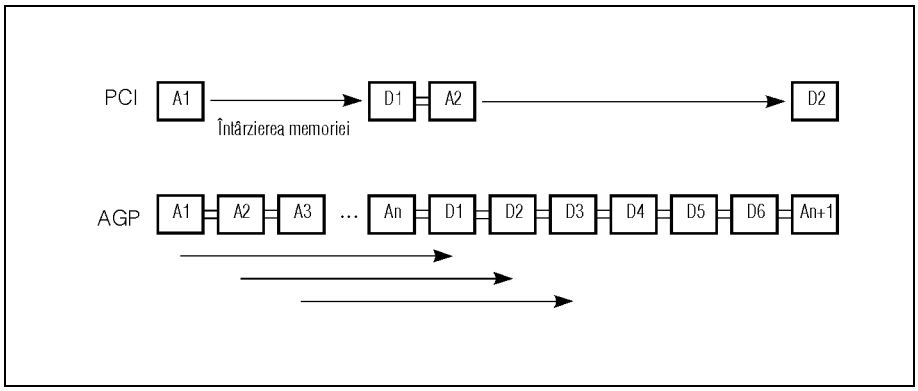

Figura 2. Cereri suprapuse la interfața AGP.

Deși atât AGP, cât și PCI permit transferuri în mod exploziv (elemente multiple de date transferate în mod continuu ca răspuns la o singură cerere), asemenea transferuri nu schimbă natura de tip *non-pipeline* a magistralei PCI.

În cazul *adresării secundare*, AGP utilizează 8 linii suplimentare de adrese, care permit controlerului grafic să transmită noi adrese și cereri simultan cu transferurile de date pe liniile principale de adrese/date ca urmare a cererilor anterioare (Figura 2).

Maparea memoriei AGP

Memoria AGP constă din zone alocate în mod dinamic ale memoriei sistem, pe care controlerul grafic le poate accesa rapid. Viteza de acces se datorează hardware-

ului încorporat în setul de circuite ale sistemului, set necesar pentru utilizarea AGP (numit uneori AGPset). Acesta translatează adresele, permițând controlerului grafic și programelor acestuia să observe un spațiu contiguu în memoria principală, în timp ce de fapt paginile sunt disjuncte. Astfel controlerul grafic poate accesa structuri de date cu dimensiuni mari, de exemplu o hartă de biți a unei texturi (de obicei de 1-128 KB), ca o singură entitate. Hardware-ul încorporat este numit GART (*Graphics Address Remapping Table*), cu funcții similare circuitelor de paginare din UCP.

Adresele virtuale liniare ale procesorului sunt translatate de circuitele sale de paginare în adrese fizice. Aceste adrese fizice sunt utilizate pentru accesul la memoria sistem. Accesurile UCP la memoria video și memoria AGP utilizează aceleași adrese ca și cele utilizate de controlerul grafic. De aceea SO setează circuitele de paginare ale UCP astfel încât să nu translateze adresele virtuale în adrese fizice pentru aceste memorii.

Pentru accesul la memoria AGP, controlerul grafic și UCP utilizează o fereastră contiguă de câțiva MB. Circuitul GART translatează însă adresele din această fereastră în diferite adrese, eventual disjuncte, ale unor pagini de 4 KB din memoria sistem. Dispozitivele PCI care fac acces la fereastra memoriei AGP (de exemplu, pentru capturarea imaginilor video) utilizează de asemenea circuitul GART.

Aspecte software

Aplicațiile existente, ca și noile aplicații care nu sunt scrise în mod special pentru AGP pot fi utilizate pe sistemele AGP, dacă sistemul de operare dispune de drivere și rutine interne pentru această interfață. Totuși, aplicațiile pot fi optimizate pentru AGP. În ambele cazuri, avantajul important al AGP este numărul mai mare de texturi detaliate, fără reducerea performanțelor în timp real.

Calculatoarele PC cu interfață AGP pot fi de trei tipuri:

- *Tipul 1:* Acest tip dispune de o interfață AGP, dar nu utilizează facilitățile interfeței legate de interpretarea texturilor, ci doar transferă datele mai rapid decât un dispozitiv PCI. Sistemul nu utilizează posibilitățile transferului pipeline sau adresarea secundară.

- *Tipul 2:* Acest tip interpretează texturile din memoria AGP, deci aplicațiile nu trebuie să transfere texturile în memoria video. Circuitele pot avea posibilitatea de interpretare a texturilor și din memoria video. Execuția poate fi mai rapidă dacă texturile nu sunt interpretate din memoria video, datorită conflictelor de acces la memoria video pentru scrierea pixelilor, reîmprospătarea ecranului, citirea elementelor de textură și a *valorilor Z*.

- *Tipul 3:* Acest tip are performanțele cele mai bune atunci când interpretarea texturilor se poate realiza atât din memoria video, cât și din memoria AGP. Texturile utilizate cel mai frecvent sau cele de dimensiuni mai mici pot fi plasate în memoria video, în timp ce texturile de dimensiuni mai mari sau cele utilizate mai puțin frecvent pot fi plasate în memoria sistem. Astfel conflictele dintre UCP și controlerul grafic vor fi minimizate.

Aplicații DOS

Interpretarea texturilor direct din memoria sistem necesită utilizarea circuitului GART, din cauza tehnicii de adresare virtuală utilizată în sistemele de operare actuale. Însă, pentru aplicațiile executate sub sisteme de operare mai vechi (de exemplu, DOS) fără adresare virtuală, circuitul GART este inutil. Aplicațiile vechi executate sub DOS vor beneficia de viteza mai mare a AGP, dar vor necesita anumite modificări ale driverelor pentru a activa posibilitatea controlerului grafic de a accesa texturile direct în memoria sistem.

Aplicații Windows

Aplicațiile *Windows* nemodificate pot beneficia de avantajele AGP, deoarece versiunile noi ale SO și biblioteca *DirectDraw* a *Microsoft* au fost actualizate pentru a permite utilizarea acestei interfețe.

Pentru implementările hardware curente, SO va marca memoria AGP (ca și o altă memorie video) pentru a nu fi încărcată în memoria cache, astfel încât nu va exista o problemă de coerență între memoriile cache ale UCP și datele utilizate de controlerul grafic. În caz contrar, accesele controlerului grafic la memoria AGP ar necesita golirea memoriilor cache ale UCP, ceea ce ar cauza întârzieri în anumite cazuri.

Alocarea memoriei de către DirectDraw

DirectDraw va aloca în mod implicit memoria pentru texturi în ordinea de mai jos, cu excepția cazului în care aplicația solicită în mod expres o altă alocare:

- Memoria video locală.
- Memoria AGP.
- Memoria sistem.

În cazul în care controlerul grafic nu poate interpreta texturile din memoria AGP, se poate împiedica alocarea de către *DirectDraw* a oricărei memorii diferite de memoria video locală pentru texturi. Driverul controlerului grafic raportează posibilitățile sale către SO și *DirectDraw*, și dacă controlerul grafic nu poate accesa direct memoria sistem, *DirectDraw* va aloca aplicației numai memorie video locală și memorie sistem. Similar, în cazul în care controlerul grafic nu poate interpreta texturile din memoria video locală, *DirectDraw* nu va aloca memorie video locală pentru texturi.

Dacă aplicația nu poate plasa toate texturile în memoria AGP alocată de *DirectDraw*, atunci aplicația trebuie să copieze texturile cerute de pe disc în memoria AGP. Aplicațiile care utilizează texturi de dimensiuni mari pot necesita încărcarea texturilor de pe disc sau din rețea în memoria AGP, indiferent de cantitatea de memorie alocată pentru acestea de *DirectDraw*.

Avantajele AGP

Principalele avantaje ale AGP sunt rezumate mai jos.

- *Rată de transfer mai ridicată*. Rata de transfer la vârf este de 2-4 ori mai mare decât cea a magistralei PCI, datorită modului pipeline, adresării secundare și a transferurilor de date care au loc atât pe frontul crescător, cât și pe cel descrescător al ceasului.
- *Interpretarea directă a texturilor din memoria sistem*. AGP permite accesul direct cu viteză ridicată la memoria sistem de către controlerul grafic, în locul încărcării prealabile a texturilor în memoria video locală.
- *Grafică de calitate mai ridicată*. Se pot utiliza texturi cu dimensiuni, și nivele de detaliere nelimitate.
- *Costuri mai reduse*. Prin minimizarea necesarului de memorie video, AGP ajută la reducerea costurilor noilor sisteme.
- *Congestie mai redusă pe magistrala* PCI. AGP funcționează concurent cu, și independent de cele mai multe tranzacții de pe magistrala PCI. Sistemele vor avea o stabilitate mai mare atunci când traficul necesar pentru imaginile grafice și cele video este eliminat de pe magistrala PCI.