Графичен порт AGP. Режими на работа

1. Причини за създаване на AGP

Когато започват да преобладават графичните потребителски интерфейси (като Windows) и особено с развитието на тримерната компютърна графика, възниква и се изостря проблемът със скоростта на обработката на видеоданните.

Входно/изходните шини ISA, MCA и EISA са относително бавни и не могат да осигурят необходимата скорост за редица периферни устройства като видеокартите и твърдите дискове. Поради това възниква концепцията за локалната шина, която се основава на идеята някои от външните устройства (най-често видеокартите) да осъществяват достъп до частта от шината, която е локална за процесора - самата процесорна шина.

Първоначално се създава локалната шина VL-Bus (VESA local bus), която по същество не е нищо друго освен чистата процесорна шина на 486. При нея към изводите на 486 процесора се свързва слот за разширителни карти. Въпреки простата конструкция и ниската си цена тази шина има кратък живот поради проблеми със съвместимостта с процесорите с различни тактови честоти и проблеми със синхронизацията, причинени от натрупване на вредните капацитети от различни карти.

Шината VLB (VESA) скоро е изместена от новата локална шина PCI. Вместо да се свързва директно с процесорната шина, PCI добавя междинен слой между процесорната шина и стандартната входно/изходна шина, като се свързва с тях чрез управляващи чипове, наричани мостове. PCI увеличава значително пропускателната способност и е достатъчна за повечето периферни устройства, с изключение на едно – видеокартите. По това време видеокартата се свързва към шината PCI, която има ограничен трансфер от 133 MB/s. Същата шина е връзка между северния и южния мост и снабдява твърдия диск и другите входно-изходни устройства. Поради това тя се превръща в тясното място на системата.

В средата на 90-те години 3D[1] игрите (като Quake и др.) изискват от системата все по-висока производителност, която не може да се осигури от PCI шината. Поради това Интел създават AGP (Accelerated Graphics Port - ускорен графичен порт), който е изцяло предназначен за графични карти.

С AGP Intel решават два основни проблема за внедряването на 3D графиката:

- обема на паметта за съхраняване на текстурните карти и z-буфера;
- скоростта на обмен между графичната подсистема и системната памет.

Всяка триизмерна графика, която се показва на компютъра се построява чрез **текстурни карти (texture map)**. Основното значение на думата „текстура” е строеж, структура на материала, от който е направено някакво тяло, например минерал, растение, кост, хартия, плат и др. В компютърната графика под текстура се разбира двумерно, плоско изображение на външният вид на повърхността, което отразява особеностите на материала, от който е направено тялото. При изобразяване на тримерно тяло компютърът взема това двумерно изображение и го обвива около определена триизмерна форма, зададена чрез своите параметри от графичната карта. Това е подобно на изработването на географския глобус, при който с хартия, на която е изобразена картата на Земята, се обвива кълбо. По този начин, чрез текстурните карти се добавя реализъм към компютърните графики.

[1] 3D- three dimensional – триизмерен

Колкото повече текстурни карти са достъпни за тримерните приложения, толкова по-добре изглежда крайният резултат. Информацията за представяне на дълбочината на изображението (третото измерение) се съхранява в т.нар. **z-буфер**. Този буфер използва при нормални обстоятелства същата памет, както и текстурите.

С други думи количеството на паметта, в която се съхраняват текстурните карти и z-буфера, оказва пряко влияние върху качеството на тримерното изображение. За увеличаване на тази памет се използват два подхода:

- увеличаване на собствената памет на графичната карта, което от своя страна оскъпява значително картата и системата като цяло и не е мащабируемо (графичната памет не може да се променя);
- използване на системната памет за съхраняване на текстурите и z-буфера.

Вторият подход предлага голям обем памет на ниска цена, но в PCI системите производителността на графичната подсистема и системната памет се ограничават от физическите характеристики на шината PCI. Освен това пропускателната способност на PCI е недостатъчна за обработка на графиката в реално време (което е важно за движението на тримерните персонажи в 3D игрите).

Поради тези причини Intel разработват ускорения графичен порт AGP.

2. Същност на AGP

AGP шината е създадена от Intel като нова шина, проектирана специално за висококачествена 3D графика в реално време и поддръжка на видео. AGP е базирана на PCI, но съдържа няколко допълнения и подобрения, и е физически, електрически и логически независима от PCI. За разлика от PCI, която е истинска шина с няколко конектора (слота), AGP по-скоро е пряка, високопроизводителна връзка от точка до точка, предназначена специално за свързване на графична карта (видеокарта) в дадена система. Чрез AGP може да се включи само един тип устройство – графична карта и то само една, тъй като може да съществува само един AGP порт.

Конекторът на AGP е подобен на PCI, но притежава допълнителни сигнали и механично и електрически е несъвместим с PCI. За да се различава от PCI, е в кафяв цвят, докато PCI слотовете са бели. Разположен е близо до северния мост, процесора и RAM.

В действителност, AGP съединява графичната подсистема с блока за управление на системната памет, делейки този достъп до паметта с централния процесор на компютъра. AGP е интегриран като мостово устройство (bridge) в северния мост на чипсета и е независим от процесора на PC, което позволява за първи път паралелна работа на процесора и графичния чип, работещ като главно устройство на шината. Както и при другите шини, при AGP има главно устройство (Master) и устройства-цели (Target). Главното устройство е графичния контролер върху картата AGP, а устройството-цел – логиката AGP, която е интегрирана в чипсета.

Портът AGP, както и PCI е паралелен, 32-битов, но работи с тактова честота 66 MHz и позволява в зависимост от версията си прехвърляне на съответно 1, 2, 4 или 8 пакета данни за един такт. С тези си характеристики той осигурява **от 2 до 16 пъти по-висока пропускателна способност от шината PCI**.

Освен повишената пропускателна способност AGP увеличава скоростта на рендиране[1] на графиката чрез подобряване ефективността на използване на системните ресурси по следните начини:

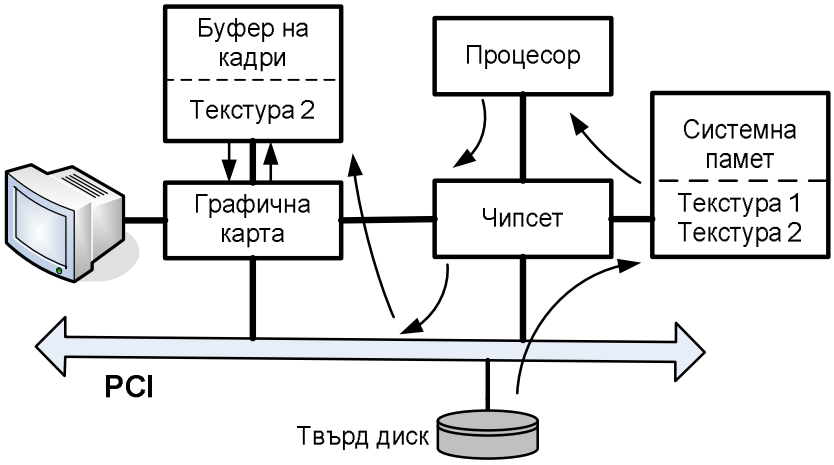
- **Специално предназначен порт**. Няма други устройства, свързани към AGP, освен графичната карта, поради което тя може винаги да работи с максималния капацитет на връзката.
- **Pipeline (конвейерно) адресиране**. PCI прави заявка за само една информация и не прави друга, докато информацията, която е изискана, не бъде прехвърлена. За разлика от това, AGP може да получава множество пакети от данни с една заявка, т.е. в рамките на един достъп до паметта. При AGP се използва сигнала PIPE#, за да се направят множество заявки. По този начин се преодолява спада на пропускливост поради непрекъснато обръщане към паметта и се увеличава производителността. Този трансфер не е съгласуван с процесорния кеш, така че четенията и записите не биват забавяни при кешовите проверки. Това намалява натовареността на шината и ускорява трансфера към графичния процесор. Например, графичната карта може да получи чрез AGP заявка за цялата информация, необходима за рендиране на определено изображение и да я изпрати наведнъж. С PCI графичната карта ще получи информация за височината на изображението и ще чака… след това дължината на изображението и ще чака…. след това широчината на изображението и ще чака … ще комбинира данните и след това ще ги изпрати.
- **Адресиране по странична шина (SBA – Side-Band Addressing)**. В този режим се разрешава използването на 8 допълнителни адресни линии освен основните 32 линии. При шината PCI няма разделение между адресните линии и линиите за данни, т.е. тя е мултиплексирана. По 32-разрядната шина отначало минава адресната информация, а след това – данните. В най-производителния „пакетен” режим шината позволява след 32-битовия адрес да се изпратят 4 двойни думи данни, за което са необходими общо 5 шинни такта. При AGP (който е базиран на PCI) възможностите на шината са разширени. Мултиплексирането не е премахнато, но е добавена допълнителна 8-битова адресна шина, която получава името странична (Side-Band Addressing – SBA). Адресната информация по тази странична шина може да се предава паралелно с данните по основната шина, така че не е необходимо да се прочете съдържанието на пакета, за да се получи адресната информация. Това позволява на графичният контролер да използва SBA шината, за да прави заявки, без да прекъсва трансфера на данни. Повишаването на производителността е до 15%. Този режим може да се включва/изключва от BIOS чрез опцията AGP Sideband Addressing.

Освен че предлага значително по-висока скорост, AGP ускорява процеса на рендиране на графиката и чрез **по-ефективно използване на системната памет**. Дава се възможност на видеокартите да осъществяват високоскоростна връзка със системната RAM памет, така че оперативната памет на дънната платка да се използва за съхраняване на текстурните

[1] Рендиране – процес на визуализация на тримерното тяло, при който чрез изчисления се получава двумерно изображение с цветове и светлосенки, създаващо илюзията за тримерност. При визуализацията се отчитат формата и разположението на тялото, ъгълът под който се гледа, светлинните източници, наложените текстурни карти и др.

карти. Това свойство на AGP се нарича DIME (**Di**rect **M**emory **E**xecute – директно изпълнение на паметта) и е една от най-важните му характеристики.

При PCI графичните карти всяка текстурна карта трябва да се съхранява два пъти (фиг. 1). Първо, текстурната карта се зарежда от твърдия диск в системната памет RAM докато бъде използвана. Когато потрябва, тя се изтегля от паметта и се изпраща на процесора за обработка. След като се обработи, се изпраща през шината PCI към графичната карта, където се съхранява отново в буфера за кадри (frame buffer) на картата. Буферът за кадри е място, където се съхранява информацията за всеки кадър след рендирането му. Тази информация представлява цветовите стойности на всеки пиксел на екрана и нейният обем зависи от разделителната способност[1] и дълбочината на цвета[2] на изображението. Цялото това съхраняване и изпращане на информация между системата и картата значително намалява общата производителност на компютъра.



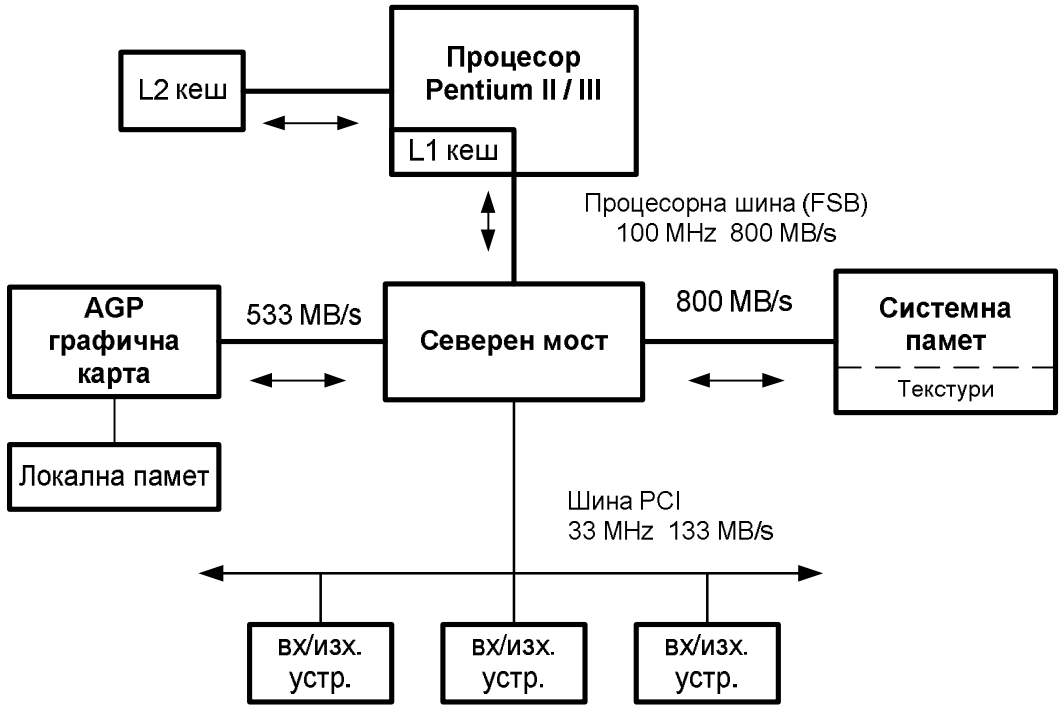
фиг. 1 При PCI графичните карти текстурните карти се зареждат от твърдия диск в системната памет, обработват се от процесора и след това се зареждат в буфера на кадри (framebuffer) на графичната карта.

За разлика от графичните карти PCI, AGP е в състояние да чете текстурите директно от системната памет, като ги съхранява само веднъж (фиг.2). Това се извършва чрез използване на GART (Graphics Address Remapping Table - таблица за преразпределяне на адресите за графика). GART преадресира частта от системната памет, която AGP заимства за съхраняване на текстурните карти. Новият адрес, осигурен от GART, кара процесора да смята, че текстурата е съхранена в буфера за кадри (frame buffer) на графичната карта. GART може реално да поставя части от текстурата из цялата системна памет, но когато процесорът се обръща към текстурата, той счита че текстурата е на адреса, посочен от GART. Така на графичната карта се позволява да има директен достъп до текстурите. Максималното количество системна памет, достъпно за AGP, се дефинира като AGP апертура (aperture – пролука, процеп). Поради по-ефективното използване на паметта се увеличава не само производителността на 3D графиката, но и на двумерната графика.

[1] Разделителната способност се измерва в брой пиксели на единица дължина.
[2] Дълбочината на цвета се измерва в брой битове за един пиксел и определя броя на възможните цветови нюанси на изображението. При 16-битова дълбочина броят на цветовете е 2^{16}= 65536

Освен по-ефективния процес на съхраняване на текстурите, чрез директния достъп до системната памет се решават още два проблема:

- като се използва повече и по-бърза системна памет, се намалява броя на текстурите, които трябва да се съхраняват в паметта на графичната карта, т.е. достатъчни са карти с по-малко собствена памет;
- размерът на текстурните карти, които компютърът е в състояние да обработи, не се ограничава от наличната памет на графичната карта.



фиг. 2 Компютърна система, използваща AGP

Използването на DIME позволява на една видеокарта да осъществява директен достъп до системната памет, като се дава възможност паметта на видеокартата да бъде разширена чрез стандартната памет от дънната платка. Това позволява както вграждането на евтини видео решения директно на дънната платка, без да е необходимо да се включва допълнителна видеопамет, така и възможност AGP картата да споделя системната памет. При високопроизводителните карти, от друга страна, се наблюдава тенденцията да включват в себе си все повече и повече собствена видеопамет, което е особено важно при изпълняването на сложни 3D видеоприложения.

Инициализацията на AGP графичната карта се извършва изцяло през PCI, преди AGP да влезе в действие. Основните AGP функции се активират не чрез BIOS, а чрез операционната система (Direct Draw).

3. Спецификации и режими на работа на AGP

Протоколът AGP има 4 режима на предаване, които се означават с 1x, 2x, 4x и 8x.

Спецификацията AGP 1.0 е публикувана от Intel през юли 1996 година. Тя дефинира 66 MHz тактова честота с 1x или 2x предаване на сигнали и използва 3,3V. Режимът 1x съответства функционално на обмена при PCI и при него на всеки такт се извършва по един

трансфер. Скоростта на предаване е 66 MHz x 4 байта x 1 трансфер = 266MB/s, което е два пъти повече от пропускателната способност на PCI. При предаване в режим 2x чрез тригер се използват двата фронта на тактовия импулс, при което на всеки такт се осъществяват по две прехвърляния, като резултатът е 533MB/s. Въпреки че най-ранните AGP карти поддържат само режима AGP 1x, повечето производители бързо преминават към режима 2x.

Версия 2.0 на AGP шината е публикувана през май 1998 година и добавя скорост на предаване на сигналите 4x, както и възможност за по-ниско работно напрежение от 1,5V. В режим 4x се работи с диференциален строб сигнал (AD, /AD), при което данните се прехвърлят четири пъти за един такт, а това се равнява па трансферна скорост 1066 MB/s.

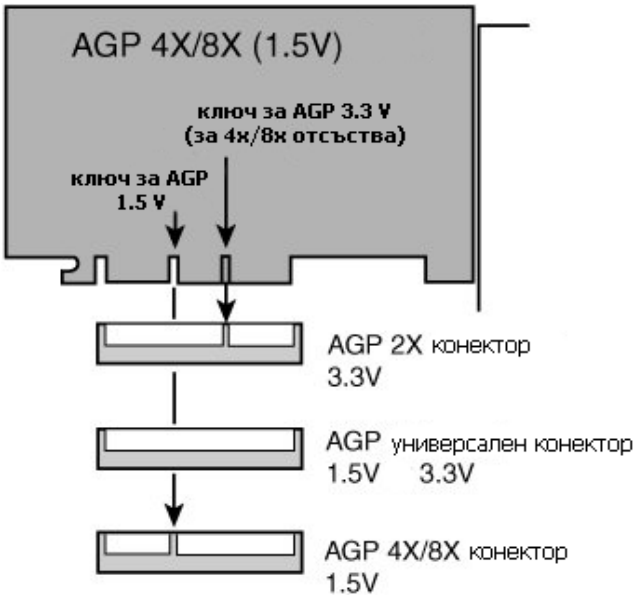
Последната версия AGP 3.0 на AGP спецификацията за PC-та е представена през ноември 2000 г. Тя дефинира режима AGP 8x, който е с 8 трансфера на такт и осигурява трансферна скорост от 2133 MB/s, която е два пъти по-голяма от тази на AGP 4x. Дава възможност за по-ниско работно напрежение - 0,8 V.

В таблица 1 са показани пропускателните способности на различните режими на AGP.

табл. 1 Режими на AGP

Режим	Честота MHz	Широчина на шината	Брой трансфери на данни за такт	Пропускателна способност MB/s
x1 AGP	**66**	**32 bits**	**1**	**266**
x2 AGP	**66**	**32 bits**	**2**	**533**
x4 AGP	**66**	**32 bits**	**4**	**1 066**
x8 AGP	**66**	**32 bits**	**8**	**2 133**

Видеокартите, които отговарят на 4x и 8x спецификацията, работят с 1.5 V. Ако някоя от тях се постави в AGP 2x слот, който е за 3.3 V, може да се повреди картата, както и дънната платка. Поради това се предвиждат специални ключове на слотовете и картите, които да предотвратяват подобни злополуки, така че 1.5 V карти да стават само на 1.5 V слотове и 3.3 V карти да стават само на 3.3 V слотове (фиг. 3). Съществуват и универсални слотове, които могат да приемат и двата вида карти.



фиг. 3 Ключове на AGP слотовете и картите против неправилно инсталиране

Изброените версии осигуряват на видеокартите мощност на захранването до 25 W. Освен тях, през август 1998 г. е представена професионална версия на AGP, наречена AGP Pro 1.0, преработена през април 1999 като AGP Pro 1.1a. Тя дефинира малко по-дълъг слот с допълнителни захранващи изводи 12 V и 3.3 V от двата края, предназначени да захранват по-големи и бързи AGP карти, които могат да консумират енергия до 110 W (фиг. 4). AGP Pro картите са предназначени за високопроизводителни графични работни станции. Слотовете AGP Pro притежават обратна съвместимост, т.е. в тях могат да се поставят стандартни AGP карти. За да се избегне неправилното поставяне на обикновени AGP карти в по-дългия слот AGP Pro, първите 20 контакта на слота, които не се използват от обикновените видеокарти, се блокират със специална капачка.

фиг. 4 Сравнение между слотове за AGP 1x/2x, AGP 4x и AGP Pro

4. Предимства на AGP пред PCI

Предимството на AGP е по-високата производителност, дължаща се на:

- по-високата тактова честота – 66 MHz срещу 33 MHz за стандартната PCI;
- способността да пренася съответно 2, 4, 8 пакета данни за един такт.
- независимост - не споделя ресурсите си с други устройства, както при PCI. Освен това, тъй като AGP е независима от PCI, използването на AGP видеокарта освобождава PC1 шината за по-традиционен вход и изход, като например IDE/ATA или SCSI контролери, звукови карти и т.н.;
- при метода за достъп до паметта **pipelining**, AGP прави множество заявки за информация в рамките на един достъп до паметта.
- възможност за използване на **sideband** адресиране (адресиране по странична шина), което означава че адресната шина и шината за данни са разделени, така че не е необходимо да се прочита съдържанието на пакета, за да се получи адресната информация. Това е направено чрез добавяне на 8 допълнителни линии, предназначени за адресирането.
- подобрено взаимодействие със системната оперативна памет (RAM) - AGP контролерът се намира в северния мост, което позволява високоскоростна комуникация между картата и системната памет. Това позволява и буферирането

на изображението (frame buffer) да се извършва в RAM паметта, а не във видео паметта.

- Използването на DIME (директно изпълнение на паметта) позволява на една видеокарта да осъществява директен достъп до системната памет, като се дава възможност паметта на видеокартата да бъде разширена чрез стандартната памет от дънната платка За разлика от графичните карти PCI, AGP е в състояние да чете текстурите директно от системната памет, като ги съхранява само веднъж

В заключение, характеристиките на AGP позволяват на видеокартата да се справи с изискванията на високоскоростното рендиране и възпроизвеждане на 3D графика в реално време, както и с възпроизвеждането на видео с кинематографично качество на PC-то.

5. Недостатъци на AGP

AGP портът има следните важни недостатъци:

- предназначен е само за видеокарти;
- може да поддържа само една видеокарта в системата.

6. Краят на AGP

Въпреки че портът AGP 8x (2133MBps) е 16 пъти по-бърз от 32-битовата 33MHz PCI шина (133MBps), AGP 8x е два пъти по-бавен от PCI-Express x16 (4000MB/s). Започвайки от средата на 2004 г., производителите на дънни платки и компютърни системи започват да заместват във високопроизводителните компютърни системи с процесори Pentium 4 и Athlon 64 порта AGP 8x с разширителни слотове PCI-Express x16. От 2006 г. повечето дънни платки във всички ценови диапазони са снабдени със слотове PCI-Express x16 вместо AGP. Някои дънни платки, създадени в периода на преход към PCI-Express притежават и двата вида слотове: AGP и PCI-Express x16.

В съвременните дънни платки вече почти не се използва AGP.

Литература:

1. Дембовски, Клаус. PC Сервизен справочник. т.3 Интерфейси и системни шини. С., Техника, 2001.
2. Мюлер, Скот. Компютърна енциклопедия. 14-то издание. С., СофтПрес, 2002.
3. Mueller, Scott Upgrading and Repairing Pcs, 17th Edition. Que. 2006.
4. Карбо, Михел. Архитектура на PC: Самоучител за всеки. – София, Егмонт България, 2003.
5. Tyson, Jeff и Robert Valdes. How AGP Works http://computer.howstuffworks.com/agp.htm