# Interface AGP e seus Sucessores

## Trabalho de Pesquisa de Arquitectura de Computadores 2

**2002/2003**

# Indicie

Introdução Histórica.....3

Interface AGP .....5

O que é o AGP? .....6

Alocação de memoria.....8

Acessos à memória através do AGP .....9

Velocidade de transferência do AGP .....10

Evolução da interface AGP.....11

Sucessores .....14

Perguntas e respostas .....20

Bibliografia .....21

# Introdução Histórica

O bus original do PC operava em 4,77 megahertz e tinha uma largura de 8 bits, significando que se podia processar 8 bits de dados em cada ciclo. Em 1982, foi melhorado para 16 bits em 8 megahertz e tornou-se oficialmente conhecida como a **Industry Standard Architecture** (ISA). Este desenho do bus é capaz de transferir dados até uma taxa de 16 MBps.

As primeiras placas gráficas, desde o **Monochrome Display Adapter** do inicio dos anos 80 até aos adaptadores **Super Video Graphics Array** (SVGA) da década de 90, ligavam-se por um slot ISA na motherboard do computador. Porque o número de cores e da definição da exposição aumentou, as placas com ligações baseadas no ISA eram simplesmente demasiado lentas. O bus do ISA não podia trasmitir os dados da imagem ao processador central rapidamente bastante.

Com o passar dos anos, os placas gráficas baseadas no ISA foram substituídas por placas gráficas **VESA Local Bus** (VL-Bus). A Video Electronics Standards Association (VESA) concordou com uma implementação padrão de SVGA que forneceu definição até 16.8 milhões de cores e uma resolução até 1280x1024. Estes placas ligavam-se por um slot especial na motherboard diferente dos ISA. O bus dos gráficos foi considerada uma bus local porque foi conectada diretamente ao processador central e teve que estar fisicamente perto dela.

A VL-Bus tinha uma largura de 32 bits e operava na velocidade da bus local, que estava normalmente à própria velocidade do processador. A VL-Bus estava ligada directamente com o processador central. Isto funcionava bem para um único dispositivo, ou até dois. Mas ligar mais do que dois dispositivos à VL-Bus introduziu a possibilidade de interferência com o desempenho do processador central. Por causa disso, a VL-Bus foi usada tipicamente somente para ligar a placa gráfica, que é um componente que beneficia realmente da elevada velocidade de acesso ao processador central.

Os cartões da VL-Bus comunicavam com o processador central à mesma velocidade que o próprio CPU. O que isto significa é que se o CPU funcionasse a 100 megahertz, a placa gráfica transferia 32 bits de dados 100 milhões de vezes por segundo. Havia dois problemas com esta aproximação:

- Os fabricantes das placas gráficas não sabiam a velocidade do sistema em que os seus produtos iam ser usados.
- Ligar directamente ao processador central podia realmente atrasar este, tendo por resultado um desempenho mais pobre.

Aparece então na década de 90 o **Peripheral Component Interconnect** (PCI), um desenho completamente novo de bus. O bus do PCI é algo de um híbrido entre ISA e VL-Bus. Fornece o acesso directo à memória de sistema para dispositivos conectados, mas usa uma ponte para se ligar ao CPU. Basicamente isto significa que é capaz de um

desempenho mais elevado do que a VL-Bus ao eliminar o potencial para a interferência com o processador central.

Rapidamente este tornou-se o padrão dominante para todo tipo de placa de expansão. O PCI trazia várias vantagens. A velocidade do bus é de 133 MB/s, contra os 8 ou 16 MB/s do bus ISA. Isso permitiu o aparecimento das placas de rede de 100 megabits, das placas SCSI de 20 megabits em diante, das placas de vídeo com suporte a mais de 256 cores, culminando no aparecimento das primeiras placas 3D, etc.

Além de mais rápido, o PCI trouxe o suporte ao plug-and-play, que acabou com as dificuladades em instalar um novo periférico. O PCI também não ocupa o processador durante as transferências de dados, como o VLB, graças ao suporte do Bus Mastering. Recursos como áudio 3D, edição de vídeo em tempo real, discos Ultra DMA, etc. seriam impossíveis sem o PCI.

São inegáveis os benefícios que o bus PCI nos proporcionou ao longo de todos estes anos, mas ao mesmo tempo as suas limitações também começam a tornar-se cada vez mais evidentes. Os 133 MB/s além de ser compartilhados entre todas as placas PCI instaladas no computador são também divididos por periféricos como as portas IDE de placa mãe, as portas USB, as portas seriais e paralelas, a drive de disquetes, enfim, dentro do PC, quase todo o fluxo de dados desagua mais cedo ou mais tarde nos 133 MB/s do PCI.

Muito já foi feito para tentar desafogar o PCI. Primeiro veio o AGP, que retirou as famintas placas de vídeo 3D do PCI e as colocou em um bus dedicado. Depois, os fabricantes passaram a cada vez mais utilizar buses próprios para interligar as pontes norte e sul (northbridge e southbridge) do chipset, como o V-Link utilizando pela Via, o HubLink utilizado pela Intel nos chipsets da série 800, incluindo o i850 do Pentium 4, o HyperTransport usado pela nVidia na série de chipsets nForce e assim por diante.

# Interface AGP

Hoje em dia, o AGP (Accelerated Graphics Port) fornece processamento de gráficos mais elevado do que o PCI. Com o AGP na terceira revisão e com uma capacidade actual de transferir perto de 2.1GB por segundo, e com o aproximar da finalização e implementação do próximo passo nos bus (PCI EXPRESS), a tecnologia continuará a evoluir a bom ritmo permitindo que as melhores imagens ainda estejam para chegar aos ecrãs.

Hoje em dia os computadores pessoais ou estações de trabalho dependem muito de gráficos, quer seja para aplicações de entretenimento quer seja para correr maior parte dos sistemas operativos que assentam num GUI (Graphical User Interface) como o Windows da Microsoft. A placa gráfica de um computador pode estar ligada de três maneiras à motherboard:

- Pode ser onboard – A placa gráfica e a memória desta estão assentes no PCB (Printed Ciruit Board) da motherboard.
- PCI – a placa liga-se a um slot PCI
- AGP – a placa está ligada pelo slot AGP

# O que é o AGP?

O AGP é uma ligação entre o controlador gráfico e o chipset da motherboard, permitindo uma melhor performance em em aplicações 3D. O AGP alivia o estrangulamento dos gráficos, mediante a adição de uma interface dedicada de alta velocidade que liga directamente o chipset e o controlador gráfico.

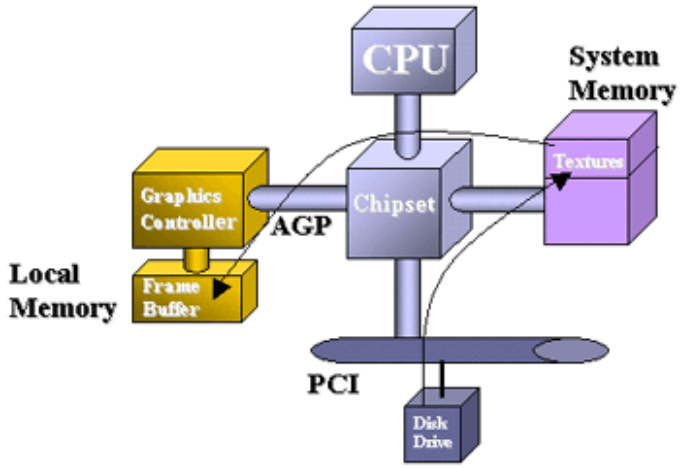



Figura 1 – Representação de um subsistema com AGP

Esta figura é uma representação de como as texturas seriam armazenadas na memória de sistema (RAM) e transportadas directamente para o frame buffer da placa gráfica. Isto remove as aplicações 3D, intensivas em largura de banda largura dos limites do bus PCI e permite-lhes a capacidade de usar mais texturas do que tipicamente pode caber na memória local para imagens de detalhe elevado e mapas de textura (texture maps). Isto reduz também a exigência de memória local na própria placa gráfica.

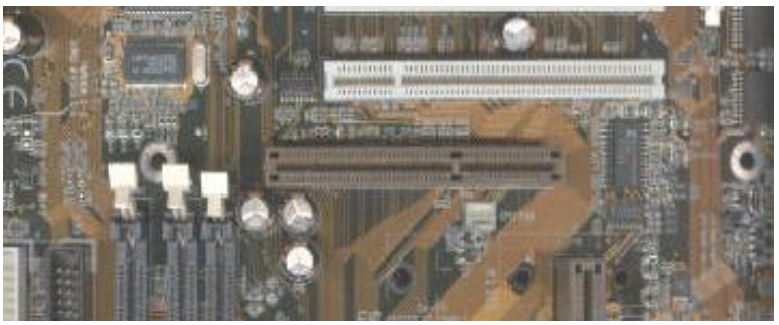


Figura 2 – Slot Agp em baixo e slot PCI em cima

Os segmentos da memória de sistema podem dinamicamente ser reservados pelo Sistema Operativo para serem usados pelo controlador de gráficos. Esta memória é denominada memória de AGP ou memória vídeo não local. O resultado líquido é que o controlador dos gráficos está obrigado a manter poucos mapas da textura na memória local. As exigências menores de memória local significam um custo mais baixo do sistema total. Esta inovação elimina também o confinamento do tamanho que a memória local dos gráficos coloca em mapas da textura, permitindo assim às aplicações usar mapas muito maiores de textura e de promover uma melhoria da qualidade da imagem e aumento do realismo.

O AGP tem 32 linhas para o endereçamento e dados multiplexed. Há 8 linhas adicionais para endereçamento do sideband. A memória vídeo local é muito mais cara do que a memória do sistema e não pode ser usada para outras finalidades pelo SO quando não está a ser ocupada pelos gráficos das aplicações que estão a correr. O controlador gráfico necessita de rápido acesso à memória vídeo local para fazer o refresh do ecrã, e vários elementos do pixel incluindo Z-buffers, double buffering, overlay planes e texturas. Por estas razões, os programadores podem sempre esperar ter mais memória de texturas disponível através da memória de sistema via AGP. Manter texturas fora do buffer de frame permite uma definição de ecrã maior, ou permite fazer o Z-buffering para um tamanho de ecrã maior. Porque a necessidade dos gráficos continua a aumentar nas aplicações intensivas em gráficos, a quantidade de texturas armazenadas na memória de sistema aumentará. O sistema AGP transporta estas texturas da memória de sistema ao controlador dos gráficos com velocidade suficiente para fazer a memória de sistema viável como espaço secundária para guardar as texturas.

# Alocação de Memória

Durante a iniciação da memória de AGP, o SO designa páginas de 4Kbytes da memória de AGP na memória (física) principal. Estas páginas não são geralmente contíguas. O controlador dos gráficos necessita que a memória seja contígua. Um mecanismo da tradução chamado GART (Graphics Address Remapping Table), faz com que a memória não contígua apareça como contígua traduzindo endereços virtuais em endereços físicos na memória principal através de uma tabela remapping.

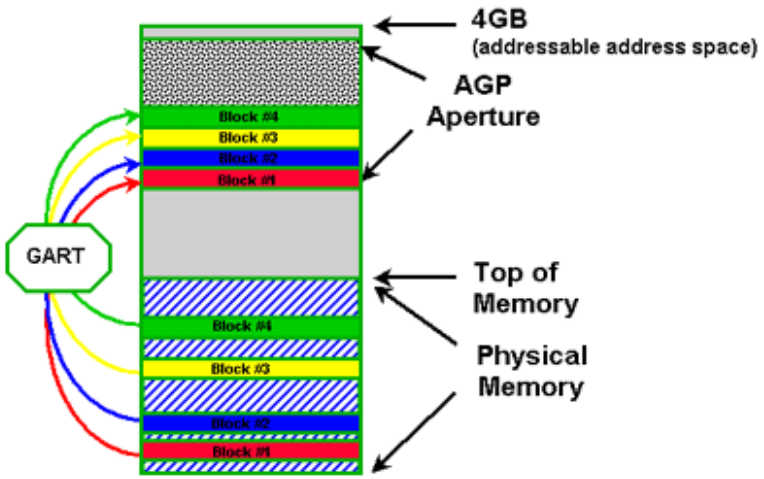



Figura 3 – Mecanismo GART

Um bloco de espaço de memória contíguo, chamado Aperture (abertura) é alocado acima do topo da memória. A placa gráfica acede a esta abertura como se fosse memória principal. O GART pode então fazer um mapeamento destes endereços virtuais para os endereços físicos na memória principal. Estes endereços virtuais são usados para aceder a memória principal, o local frame buffer, e a memória de AGP.

Para acessos à memória de AGP, o controlador de gráficos e o processador central usam uma abertura contígua, ajustada geralmente no BIOS entre 64MB e 256MB. Placas PCI que usam a abertura da memória de AGP (para por exemplo, a captação de vídeo em tempo real) também atravessam o GART.

# Acessos à memória através do AGP

O AGP fornece dois modos para o controlador de gráficos aceder directamente aos mapas de textura na memória de sistema. Usando pipelining e sideband addressing (endereçamento lateral);

- Usando o modo de Pipelining, AGP sobrepõe os tempos da memória ou de acesso do bus para um pedido (“n”) com a introdução dos pedidos seguintes ("n+1"... "n+2"... etc.). No bus do PCI por exemplo, o pedido "n+1" não começa até que a transferência de dados do pedido "n" termine. Enquanto que o AGP e o PCI podem enviar múltiplos blocos de dados continuamente em resposta a um único pedido), isto unicamente alivia em parte a natureza non-pipelined do PCI. A profundidade do pipelining do AGP depende da implementação, e continua transparente ao software de aplicação. O pipelining não é suportado na especificação AGP3.0.

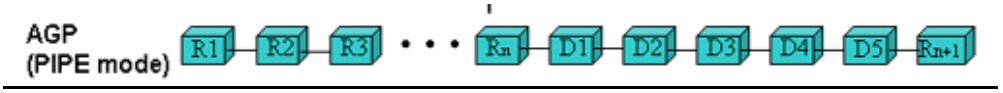



Figura 4 – Exemplo de pipelining

- Com o endereçamento lateral (SBA), O AGP usa 8 linhas extra, sideband, de endereçamento que permitem que o controlador dos gráficos emita endereços e pedidos novos simultaneamente enquanto os dados continuam a mover-se dos pedidos precedentes nas linhas do tubo principal de 32 linhas de date e endereçamento. Usar a modalidade de SBA melhora a eficiência e reduz latências.

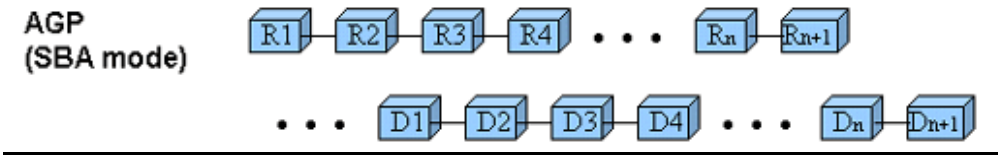



Figura 4 – Exemplo de endereçamento lateral

# Velocidade de transferência do AGP

Para calcular a taxa de transferência de um dado bus basta seguir a seguinte fórmula:

frequência de operação X número de bits por transferência / 8

O bus do PCI típico funciona a 33Mhz com palavras de 32bits. Logo a sua taxa de transferência máxima teórica é 33000000 x 32 / 8 = 132Mb/s. No caso do bus AGP, são transferidos dados de 32 bits a 66 MHz. Logo a sua taxa de transferência é de 66.000.000 x 32 / 8 = 264 MB/s. Esta é a taxa de transferência básica do bus AGP, chamado AGP1X.

O AGP pode trabalhar com taxas de transferência mais elevadas, de 528 MB/s (AGP2X) e 1 GB/s (AGP4X) e 2GB/s (AGP8X). No caso do AGP2X, esta taxa de transferência mais elevada é feita através do hardware: em vez de se fazer o sincronismo somente na subida do ciclo de relógio (quando este passa de "0" para "1"), a transferência é feita tanto na subida quanto na descida do ciclo de relógio. Com isto, a taxa de transferência é dobrada: em cada ciclo de relógio, em vez de ser feita a transmissão de somente uma palavra de 32 bits, é feita a transmissão de duas palavras de 32 bits (ou seja, 64 bits por ciclo de relógio, sendo 32 bits de cada vez).

# Evolução da interface AGP

Assim como quase tudo na Informática, o AGP também teve o seu processo de evolução. O primeiro modelo de AGP, chamado de AGP1X, permitia transferir até 266 MB de dados por segundo. O padrão seguinte foi o AGP2X, que é duas vezes mais rápido, permitindo que a placa de vídeo receba ou transmita até 528 MB de dados por segundo.

Actualmente o modo mais usado é o AGP 4x, que permite transferências de dados de até 1056 MB por segundo. Estando a ser introduzido desde o ano passado o AGP8X tem vindo a ganhar terreno e será a ultima etapa da interface AGP podendo transferir dados até 2112MB/s.

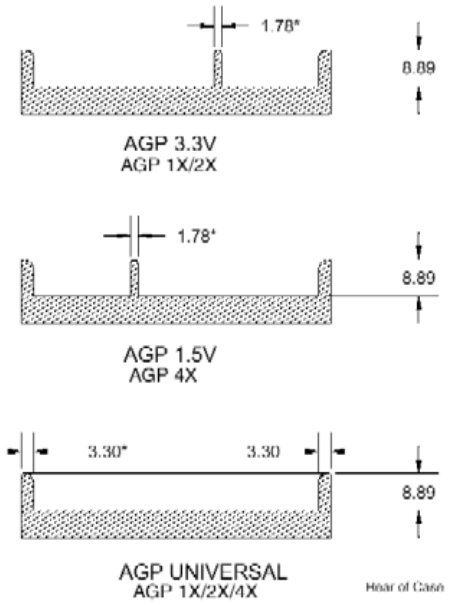
| | AGP1.0 | AGP2.0 | AGP3.0 |
|---|---|---|---|
| Signaling | 3.3V signaling | 1.5V Signaling | New 0.8V Signaling |
| Protocol | Pipelined transactions + Source synchronous clocking | AGP1.0 + Fast Writes | AGP2.0 + Some enhancements – some deletions (See Sec 1.5.1) |
| Speeds | 2X, 1X | 4X, 2X, 1X | 8X, 4X |
| Connector | 3.3V keyed | 1.5V keyed, Universal | 1.5V keyed, Universal |

Figura 5 – Diferenças entre as diferentes versões do AGP

Dependendo do chipset utilizado, o slot AGP da placa mãe pode suportar apenas um certo tipo de placas AGP. Os chipsets são na sua maioria retro compatíveis mas não suportam as novas interfaces devido a diferenças na implementação.

Uma maneira de saber quais os tipos de placas AGP que são suportadas por uma placa mãe é observar a posição do pino central do slot AGP, como ilustrado nas ilustrações a seguir. Se o pino estiver próximo do fundo do gabinete, do lado da fonte, então são suportadas apenas placas de vídeo AGP 1x ou 2x, (esta é a configuração mais comum em placas mãe antigas); se o pino estiver na posição contrária, mais próximo da frente do gabinete (como na ilustração central), então são suportadas apenas placas de vídeo AGP 4x.

Mas, se não existir pino algum (como na ilustração de baixo) então temos um slot AGP universal, onde podem ser encaixadas placas de vídeo de qualquer um dos três padrões. Esta é a configuração mais comum nas placas mãe actuais.




Naturalmente, assim como muda o encaixe na placa mãe, também muda o formato do conector da placa de vídeo. Veja nas fotos abaixo a diferença entre os conectores de uma placa de vídeo AGP 2x e de outra AGP 4x:

AGP 1x/2x

Universal (pode ser usada em qualquer placa mãe com slot AGP)

Apesar de permitir uma largura de banda larga o suficiente para satisfazer as placas de vídeo 3D mais poderosas, os slots AGP 4x possuem um grave problema, que dificulta a produção de placas de vídeo mais exigentes.

O problema é que, como no caso dos processadores, quanto mais poder de processamento um chipset de vídeo possuir, mais transístores ele deverá ter. Quanto mais transístores, maior é o consumo eléctrico.

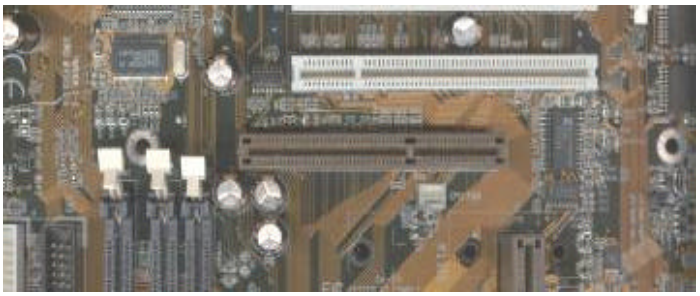
Um slot AGP 4x comum, não é capaz de fornecer estavelmente mais de 20 ou 25 Watts de corrente, o que limita bastante o potencial das placas de vídeo.

Surgiu então o AGP Pró, que é na verdade um slot AGP 4x com 48 contactos a mais, 20 de um lado e mais 28 do outro. Estes contactos adicionais são usados para aumentar a capacidade de fornecimento eléctrico do slot.

Existem dois tipos de slots AGP Pro: o AGP Pro50 e o AGP Pro110. O nome indica a capacidade de fornecimento eléctrico de ambos os padrões: o **AGP Pro50** é certificado para fornecer até **50 Watts**, enquanto o **AGP Pro110** pode fornecer até **110 Watts**.

O formato do encaixe não diz se o slot é Pro50 ou Pro110, apenas mostra quais tipos de placas são suportadas. O que muda do Pro50 para o Pro110 são os capacitadores de alimentação da placa mãe, que devem ser capaz de manter uma corrente maior, e não o formato do encaixe.

Veja nas fotos a seguir a diferença de tamanho entre um Slot AGP tradicional e um slot AGP Pro:



AGP tradicional

AGP Pro

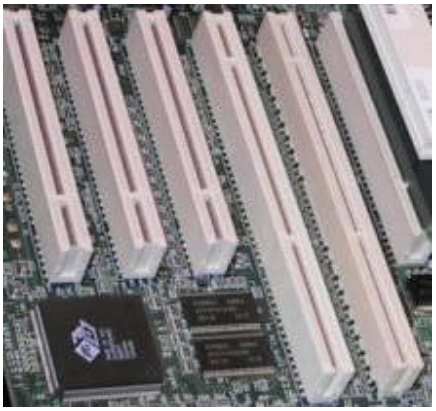
# Sucessores

Apesar dos buses novos e do AGP, o PCI tradicional ainda deixa a desejar em muitas situações. Se se tiver muitos periféricos rápidos no PC, como por exemplo uma placa RAID com 4 discos rápidos, ou uma placa SCSI também com vários discos, várias placas de rede de 100Mb/s, uma placa de vídeo 3D PCI, uma placa de captura e edição de vídeo, e assim por diante, podemos ter a certeza que em muitas situações, quando vários destes periféricos são usados simultaneamente o bus do PCI fica saturado e o PC perde desempenho.

O PCI também já não é suficiente para uma série de dispositivos a operarem sozinhos, entre eles as placas SCSI de 160 MB/s e as placas de rede Gigabit Ethernet, já que apenas uma destas placas poderia facilmente saturar o barramento. Sem falar das placas 10 Gigabit Ethernet, que em breve começarão a ser usadas em larga escala.

Para estes dispositivos surgiram duas novas versões do PCI, o PCI de 64 bits, que dobra o número de linhas de dados para atingir 266 MB/s e o PCI de 64 bits e 66 MHz, que dobra também a frequência de operação, atingindo 533 MB/s, a velocidade de um slot AGP 2X.

Os slots de 64 bits são bem maiores que os tradicionais e são encontrados apenas em placas mãe mais caras, destinadas a servidores, onde existe mais necessidade de velocidade. A maioria das placas Ultra Wide SCSI e Gigabit Ethernet utilizam slots PCI de 64 bits, para evitar que o seu desempenho seja sub aproveitado pelo bus de 132 MB/s dos slots de 32 bits.



Slots PCI de 64 bits ao
lado de slots comuns

Mas, por que então todos os PCs não passam a trazer slots PCI de 64 bits, já que além de mais rápidos eles podem manter compatibilidade com as placas antigas?

O problema de não se passar a usar os slots PCI de 64 bits resume-se a uma única razão: o seu custo. As placas mãe com slots PCI de 64 bits são mais caras por que

possuem mais linhas de dados. Isto não só aumenta muito os custos de produção das placas, mas também seus custos de projecto.

Mesmo que fossem adoptados em todas as novas placas, o padrão jamais chegaria a custar o mesmo que o PCI actual.

Já existem muitos exemplos de buses mais modernos, rápidos, mais flexíveis e em muitos casos até mesmo (teoricamente) mais baratos que o PCI. Em alguns casos não existe nem mesmo limitações quanto ao uso de periféricos externos, o bus pode se estender por uma distância suficiente para se interligar a outros PCs ou a outros periféricos, como se fosse uma arquitectura de rede. Na verdade, existem mais de 60 buses de alta velocidade que poderiam ser utilizados como substitutos do PCI, está-se a ver neste momento uma verdadeira guerra para a aceitação generalizada de um padrão.

Os mais populares no seio da comunidade são:

## HyperTransport

Esta é uma tecnologia desenvolvida pela AMD entre 1998 e 2001, que já é utilizada em alguns produtos, como a X-Box, os chipsets nForce da nVidia, o novo processador Opteron o prestes a lançar Athlon64, e em vários sistemas de comunicação. Assim como outros projectos actuais o HyperTransport baseia-se na ideia de uma alta-frequência de operação combinada com poucas linhas de dados. O padrão inicial, utiliza dois pares de linhas e trabalha a 800 MHz. Com isto, o bus de dados é de 1.6 Gb/s (ou seja 200 MB/s) em cada direcção, o que não chega a impressionar tanto em relação aos 133 MB/s do PCI de 32 bits.

Em compensação, o HyperTransport oferece uma grande possibilidade de expansão, tanto na frequência de operação, quanto no aumento no número de pares de fios. A 1 GHz a velocidade por par sobe para 2 GB/s (250 MB/s) e a 2 GHz dobra para 4 GB/s (500 MB/s). Também é possível aumentar a largura do barramento de dois para até 32 bits (16 pares de fios) em cada direcção, o que elevaria a taxa de transferência para até 6.4 GB/s em cada direcção a 800 MHz, ou 8 GB/s em cada direcção a 1 GHz. O padrão suporta tanto o uso de filamentos tradicionais de cobre, quanto o uso de fibra óptica.

Mas, ao aumentar o número de pares de fios, o custo de produção também aumenta na mesma proporção. Apesar de ser tecnicamente possível, é improvável que os padrões de 16 e 32 bits do HyperTransport cheguem a ser usados em larga escala num futuro próximo. A tecnologia é muito cara.

Em compensação, os padrões de dois e quatro pares parecem ter um grande futuro pela frente, principalmente como meio de comunicação entre a ponte norte e ponte sul do chipset e entre outros periféricos da placa mãe, como temos hoje no nForce no nforce2 e no recentemente lançado nforce3, onde são utilizados quatro pares de trilhas, resultando num barramento bidireccional de 800 MB/s.

A AMD pretende lançar novos padrões do HyperTransport, operando a até 5 GHz num futuro próximo, o que tornará o barramento ainda mais competitivo.

## RapidIO

Ao contrário do HyperTransport, o RapidIO se destina a um mercado específico, para ser mais exacto, o mercado de dispositivos embedded e pequenos dispositivos de rede. A principal vantagem é o baixo custo, que surge devido à simplicidade do padrão.

O RapidIO pode ser usado tanto para interligar os componentes da placa mãe e placas de expansão como para interligar dispositivos próximos. Esta é uma possibilidade que também existe no HyperTransport e, em teoria, também no 3GIO que veremos a seguir.

Existem dois padrões de RapidIO, com bus de 8 ou 16 bits de largura. Em ambos os casos a frequência de operação é 1 GHz, que resulta num barramento de dados de respectivamente 4 e 8 GB/s, uma velocidade impressionante, que chega a rivalizar com os padrões mais rápidos do HyperTransport.

Outra característica peculiar do RapidIO é o protocolo de comunicação usado, que se baseia em camadas e no envio de pacotes, com um bom sistema de retransmissão de pacotes e correcção de erros, um sistema que lembra muito o sistema utilizado nas redes Ethernet.

## PCI-X

Esta é uma evolução do PCI de 64 bits, que dá mais um passo adiante em termos de velocidade, mantém a compatibilidade com as placas PCI actuais, mas em compensação não soluciona o problema do custo dos slots de 64 bits.

Na verdade, as evoluções do PCI-X se limitam ao nível lógico e a uma maior frequência de operação. O número de trilhas e o formato físico dos slots continua o mesmo.

Existem duas versões do PCI-X, que operam a 100 e a 133 MHz, sempre com 64 bits por ciclo de clock. A 133 MHz a taxa de transferência atinge respeitáveis 1064 MB/s, o mesmo barramento de dados permitido pelo AGP 4X. O PCI-X é um sucessor natural para o PCI de 64 bits nos servidores.

O PCI-X 2.0, que engloba dois novos padrões do PCI-X capazes de operar a 266 e 512 MHz está prometido para o final de 2002, início de 2003. Os novos padrões serão

capazes de transferir respectivamente 2128 e 4256 MB/s, mas é pouco provável que cheguem a equipar placas destinadas a PCs domésticos, devido ao seu elevado custo.

## InfiniBand

O InfiniBand ainda está em desenvolvimento e promete várias novidade para o futuro. Ao contrário dos que citei anteriormente, o principal objectivo é interligar servidores e dispositivos de armazenamento localizados a curtas distâncias, servido como uma opção mais rápida às redes Ethernet. Usando o InfiniBand um servidor de bancos de dados poderia aceder a um dispositivo de armazenamento externo, sem nenhum congestionamento, como se fosse um dispositivo local, o que abre muitas possibilidades nos servidores de alto desempenho, clusters e server farms.

O InfiniBand é um bus serial, que oferece 2.5 Gigabits (312 MB/s) por segundo por par de cabos, onde um envia e o outro recebe dados. Como a comunicação é bidireccional, temos 312 MB/s em cada sentido, totalizando um barramento total de 625 MB/s, mas que poderia ser utilizado plenamente apenas caso ambos os dispositivos transmitissem grandes quantidades de dados ao mesmo tempo, um cenário semelhante ao que temos numa transmissão full-duplex numa rede Ethernet.

Também é possível aumentar a largura de banda usando mais cabos, A especificação original fala em ligações com até 12 pares, que permitiria ligações de até 3.75 GB/s em cada sentido, muito mais do que as redes Gigabit Ethernet (125 MB/s) e 10 Gigabit Ethernet (1.25 GB/s) são capazes de oferecer. Em compensação, as redes Ethernet já estão aí, enquanto o InfiniBand é apenas uma promessa para o futuro.

É bem provável que a especificação final do InfiniBand permita que o bus seja utilizado também para interligar componentes internos dos PCs, substituindo o PCI ou funcionando como um bus complementar, o que aumentaria bastante a flexibilidade. Mesmo sem o padrão final, já existem alguns produtos proprietários com o InfiniBand, principalmente servidores de alta densidade.

## 3GIO

Com tantos concorrentes, nada está decidido, mas o 3GIO (PCI-EXPRESS) aparece como o sucessor mais provável para o PCI, pois mantém compatibilidade com o padrão anterior, tem um custo de implementação relativamente baixo, oferece boas taxas de transferência e conta com o apoio da Intel e de várias parceiras.

O aspecto físico de um slot 3GIO lembra bastante o dos antigos slots VLB, que eram na verdade uma extensão dos slots ISA, permitindo que se utilizasse tanto uma placa ISA antiga quanto uma placa VLB.

No caso do 3GIO temos um conector PCI convencional, que mantém a compatibilidade com as placas PCI actuais, auxiliado por um slot extra, que serve as placas que precisarem de mais velocidade:

Slot PCI-EXPRESS

A versão inicial do 3GIO será capaz de transmitir apenas 2.5 gigabits por segundo, ou 312 MB/s, pouco mais que o dobro dos slots PCI actuais. Este primeiro padrão começará a ser utilizado em 2004 segundo os planos da Intel. O padrão seguinte entrará em operação em 2005 e será 4 vezes mais rápido, atingindo 10 gigabits por segundo.

Ambos os padrões conviverão por algum tempo, mas felizmente serão inter compatíveis. Uma placa 3GIO de 10 gigabits poderá trabalhar num slot de 2.5 gigabits (embora a performance possa ser prejudicada) e vice-versa. As placas PCI continuarão sendo suportadas durante muito tempo, pelo menos até o lançamento do próximo padrão. Lembre-se que as placas ISA demoraram quase 10 anos para deixarem de ser suportadas nas placas novas depois do surgimento do PCI.

Apesar de parecer apenas um “remendo” do PCI, o 3GIO elimina toda a carga de legado do bus antigo. O slot PCI foi mantido, mas toda a parte lógica foi muito modificada. Juntos, os slots do 3GIO utilizam apenas 40 linhas de dados, contra nada menos que 84 linhas do PCI tradicional, 150 linhas do PCI de 64 bits e 108 linhas do AGP. Sem dúvida uma economia expressiva.

Mais uma característica importante do 3GIO é a sua topologia ponto a ponto. Ao contrário do PCI, onde todos os dispositivos compartilham o mesmo bus e apenas um pode transmitir de cada vez, o 3GIO utiliza um switch para garantir que cada dispositivo disponha de uma ligação exclusiva com o chipset e os demais componentes do PC. Graças a isto, vários dispositivos podem transmitir ao mesmo tempo e dispor do bus a qualquer instante.

Isto é especialmente efectivo quando dois dispositivos ligados ao bus 3GIO precisam trocar dados entre si, como por exemplo dados que vão de uma placa de rede para a outra. Estas transferências podem ser feitas dentro do próprio bus, sem ocupar a ponte sul do chipset, nem o processador.

Está anunciada ainda uma versão do 3GIO destinada a notebooks, que substituirá os slots PC-Card utilizados actualmente, que são uma extensão do bus PCI. Mas, ainda não foi divulgado se o novo padrão manterá compatibilidade com o actual. As placas PC-Card actuais por exemplo, não podem ser instaladas em muitos notebooks antigos, com slots PCMCIA que são baseados no barramento ISA. O encaixe é o mesmo, mas placas não funcionam.

Apesar de mais caros, os notebooks ainda são mais descartáveis visto não terem grandes possibilidades de upgrades que os desktops, por isso os fabricantes não se preocupam tanto em manter compatibilidade com os padrões anteriores.

Para completar, está previsto que o 3GIO permitirá também a conexão de dispositivos externos, mantendo a mesma velocidade de transferência de dados, sem dúvida um grande avanço sobre os 400 megabits do USB 2.0 e do Fireware, mas que será aproveitado por poucos periféricos.

Seja o 3GIO, ou o HyperTransport, ou o InfiniBand, o fato é que o PCI será em breve substituído, uma mudança sem dúvida para melhor.

# Perguntas e respostas

1. **Pergunta:** Identifique e explique sucintamente quais os modos que o AGP fornece para o controlador de gráficos aceder directamente aos mapas de textura na memória de sistema.

   **Resposta:** Os modos fornecidos pelo AGP para que o controlador de gráficos possa aceder directamente aos mapas de textura na memória de sistema são: pipelining e sideband addressing (endereçamento lateral)**.**

   Pipelining - usando o modo de Pipelining, AGP sobrepõe os tempos da memória ou de acesso do bus para um pedido (“n”) com a introdução dos pedidos seguintes ("n+1"... "n+2"... etc.). No bus do PCI por exemplo, o pedido "n+1" não começa até que a transferência de dados do pedido "n" termine. Enquanto que o AGP e o PCI podem enviar múltiplos blocos de dados continuamente em resposta a um único pedido), isto unicamente alivia em parte a natureza non-pipelined do PCI. A profundidade do pipelining do AGP depende da implementação, e continua transparente ao software de aplicação.

   Endereçamento lateral - com o endereçamento lateral (SBA), O AGP usa 8 linhas extra, sideband, de endereçamento que permitem que o controlador dos gráficos emita endereços e pedidos novos simultaneamente enquanto os dados continuam a mover-se dos pedidos precedentes nas linhas do tubo principal de 32 linhas de date e endereçamento. Usar a modalidade de SBA melhora a eficiência e reduz latências.

2. **Pergunta:** Apesar de permitir uma largura de banda larga o suficiente para satisfazer as placas de vídeo 3D mais poderosas, os slots AGP 4x possuiam um grave problema, que dificultava a produção de placas de vídeo mais exigentes. Qual é esse problema? Como foi resolvido o problema?

   **Resposta:** O problema é que, como no caso dos processadores, quanto mais poder de processamento um chipset de vídeo possuir, mais transístores ele deverá ter. Quanto mais transístores, maior é o consumo eléctrico. Para resolver esse problema, surgiu então o AGP Pro, que é na verdade um slot AGP 4x com 48 contactos a mais, 20 de um lado e mais 28 do outro. Estes contactos adicionais são usados para aumentar a capacidade de fornecimento eléctrico do slot. Existem dois tipos de slots AGP Pro: o AGP Pro50 e o AGP Pro110. O nome indica a capacidade de fornecimento eléctrico de ambos os padrões: o **AGP Pro50** é certificado para fornecer até **50 Watts**, enquanto o **AGP Pro110** pode fornecer até **110 Watts**.

# Bibliografia

www.intel.com/developer - white papers sobre AGP e PCI-Express e o tutorial sobre AGP.

www.clubedohardware.com.br – artigo sobre AGP

www.guiadohardware.net – artigo sobre os sucessores do PCI, artigo sobre duvidas sobre AGP

www.pcguide.com – artigo sobre AGP

# AUTORES

**Marco Paulo Pereira de Melo Tinoco 501000933**

**Rui Paulo Fonseca Lousã 501000954**