As imagens que você vê na tela do seu computador são feitas de minúsculos pontos chamados Pixels. Na configuração mais comum de resolução, uma tela apresenta mais de um milhão de pixels e o computador deve decidir o que fazer com cada um para criar uma imagem. Para fazer isto, ele necessita de um conversor – algo que pegue os dados binários do CPU e transforme na imagem que você possa visualizar. É o papel da placa de vídeo.

A placa de vídeo funciona da seguinte forma: A CPU, trabalhando em conjunto com as aplicações de softwares, envia informação à placa de vídeo sobre a imagem que deseja desenvolver, que então define como utilizar os pixels na tela para criar a imagem. Isso então envia aquela informação ao monitor através de um conector (cabo), tornando a informação inicial visível aos olhos.

Para isso, uma placa primeiro cria uma moldura de linhas retas. Então ela preenche os pixels remanescentes, adicionando luzes, texturas e cores. Para jogos que exigem alto poderio gráfico, o computador precisa fazer este procedimento cerca de 60 vezes por segundo. Sem uma placa específica para executar os cálculos necessários, um computador precisaria de muito trabalho.

**Muito além dos jogos**

A razão para isso está na maneira como são projetadas as GPUs. Ao contrário das CPUs, elas possuem um número bem maior de núcleos de processamento, criados para desempenhar muitas tarefas paralelamente.

Um supercomputador atual pode ser usado para simular, por exemplo, o comportamento de grandes quantidades de fluidos, como a água represada em alguma barragem. A dinâmica do fluido precisa ser estudada com grande precisão para que engenheiros possam determinar os limites da estrutura e o planejamento de manutenção do complexo possa ser executado com maior precisão. Imagine o quão complexo é simular o comportamento de bilhões de litros de água e as forças que interagem entre o líquido e as estruturas de contenção

A tarefa seria ingrata mesmo para um grupo de engenheiros bem capacitados. O supercomputador, munido das GPUs que foram criadas para renderizar milhões de pixels por segundo, pode calcular esse tipo de informação com muito mais velocidade e precisão

Esse mesmo nível de simulação já é possível em algumas outras áreas, como a de estudo do clima. Nesse caso, outro fluido é interpretado a partir da alta capacidade de GPUs em fazer cálculos paralelos, como na análise do ar e das correntes que moldam as variações climáticas da superfície terrestre.

As placas de vídeo realizam esta tarefa utilizando 4 componentes:

– A conexão da placa mãe para os dados e energia;  
– O processador para decidir o que fazer com cada pixel na tela;  
– Memória para manter as informações sobre cada pixel e para armazenar temporariamente as imagens concluídas;  
– A conexão de uma tela para visualizar o resultado.

Processador e memória gráfica?

Como uma placa mãe, a placa de vídeo é um painel de circuito impresso que abriga um processador e memória RAM. Um processador da placa de vídeo, chamado de Unidade de Processamento Gráfico (GPU), é similar a CPU de um computador, no entanto, a GPU é projetada especificamente para realizar os complexos cálculos matemáticos e geométricos necessários para a renderização de gráficos. Algumas das GPUs possuem mais transistores do que a média de CPU. Importante lembrar que a GPU produz uma grande quantidade de calor, por isso geralmente está localizado sob ventoinhas.

À medida que a GPU cria imagens, ela precisa de um lugar para armazenar as informações e as imagens completas. Ela utiliza a RAM da placa para isto, armazenando dados sobre cada pixel, tal como sua cor e sua localização no ecrã.

**Caracteristicas de uma GPU**

Para que as imagens possam ser geradas, a GPU trabalha executando uma sequência de etapas, que envolvem elaboração de elementos geométricos, aplicação de cores, inserção de efeitos e assim por diante. Essa sequência, de maneira bastante resumida, consiste no recebimento pela GPU de um conjunto de vértices (o ponto de encontro de dois lados de um ângulo); no processamento dessas informações para que elas obtenham contexto geométrico; na aplicação de efeitos, cores e afins; e na transformação disso tudo em elementos formados por pixels (um pixel é um ponto que representa a menor parte de uma imagem), processo conhecido com rasterização. O passo seguinte é o envio dessas informações à memória de vídeo (frame buffer) para que então o conteúdo final possa ser exibido na tela.

As GPUs podem contar com vários recursos para a execução dessas etapas, entre eles:

**- Pixel Shader:** shader é um conjunto de instruções utilizado para o processamento de efeitos de renderização de imagens. Pixel Shader, portanto, é um programa que trabalha com a geração de efeitos com base em pixels. Esse recurso é amplamente utilizado em imagens 3D (de jogos, por exemplo) para gerar efeitos de iluminação, reflexo, sombreamento, etc;

**- Vertex Shader:** semelhante ao Pixel Shader, só que trabalha com vértices em vez de pixels. Assim sendo, Vertex Shader consiste em um programa que trabalha com estruturas formadas por vértices, lidando, portanto, como figuras geométricas. Esse recurso é utilizado para a modelagem dos objetos a serem exibidos;

**- Render Output Unit (ROP):** basicamente, manipula os dados armazenados na memória de vídeo para que eles se "transformem" no conjunto de pixels que formará as imagens a serem exibidas na tela. Cabe a essas unidades a aplicação de filtros, efeitos de profundidade, entre outros;

**- Texture Mapping Unit (TMU):** trata-se de um tipo de componente capaz de rotacionar e redimensionar bitmaps (basicamente, imagens formadas por conjuntos de pixels) para aplicação de uma textura sob uma superfície.

**Descobrindo o poderio de uma placa**

Um medidor geral adequado de uma placa é a taxa de quadros, (frame rate) determinada em quadros por segundo (frames per second, FPS). Esse número indica quantas imagens completas uma placa pode exibir por segundo. O olho humano pode processar cerca de 25 frames por segundo, mas jogos de ação rápida podem exigir um mínimo de até 60 FPS para fornecer animação e transição suaves.

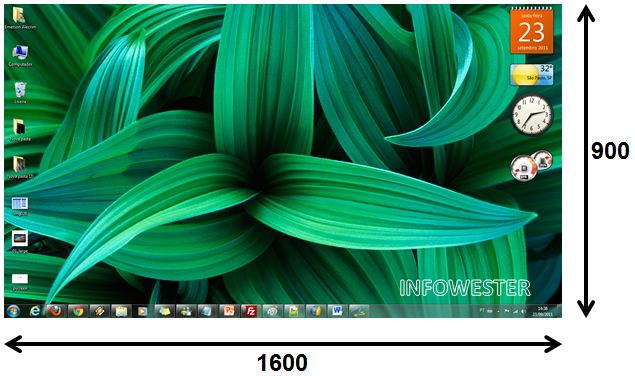
Clock da GPU

Se a GPU é um tipo de processador, então trabalha dentro de uma determinada frequência, isto é, de um clock. De maneira geral, o clock é um sinal de sincronização. Quando os dispositivos do computador recebem o sinal de executar suas atividades, dá-se a esse acontecimento o nome de "pulso de clock". Em cada pulso, os dispositivos executam suas tarefas, param e vão para o próximo ciclo de clock.

A medição do clock é feita em hertz (Hz), a unidade padrão de medidas de frequência, que indica o número de oscilações ou ciclos que ocorre dentro de uma determinada medida de tempo, no caso, segundos. Assim, quando um dispositivo trabalha à 900 Hz, por exemplo, significa que é capaz de lidar com 900 operações de ciclos de clock por segundo.

Resolução e cores

Ao comprar uma placa de vídeo, uma característica importante que geralmente é descrita nas especificações do dispositivo é a sua resolução máxima. Quando falamos deste aspecto, estamos nos referindo ao conjunto de pixels que formam linhas horizontais e verticais na tela. Vamos tomar como exemplo uma resolução de 1600x900. Esse valor indica que há 1600 pixels na horizontal e 900 pixels na vertical, como exemplifica a imagem:



Por muito tempo, a combinação de informações referentes a resoluções e cores indicava o padrão utilizado pela placa de vídeo. Eis os padrões mais comuns:

**MDA (Monochrome Display Adapter):** padrão utilizado nos primeiros PCs, indicando que a placa era capaz de exibir 80 colunas com 25 linhas de caracteres, suportando apenas duas cores. Utilizado em uma época onde os computadores trabalhavam, essencialmente, com linhas de comando;

**CGA (Color Graphics Adapter):** padrão mais avançado que o MDA e, portanto, mais caro, suportando geralmente resolução de até 320x200 pixels (podendo alcançar 640x200) com até 4 cores ao mesmo tempo entre 16 disponíveis;

**EGA (Enhanced Graphics Adapter):** padrão utilizado no então revolucionário PC AT, suportando geralmente uma resolução de 640x350 com 16 cores ao mesmo tempo dentro de 64 possíveis;

**VGA (Video Graphics Adapter):** padrão que se tornou amplamente conhecido com o sistema operacional Windows 95, trabalhando com resolução de 640x480 e 256 cores simultaneamente ou 800x600 com 16 cores ao mesmo tempo;

**SVGA (Super VGA):** tido como uma evolução do VGA, o SVGA inicialmente indicava a resolução de 800x600 pixels e, posteriormente, de 1024x768. Na verdade, a partir do SVGA, as placas de vídeo passaram a suportar resoluções ainda mais variadas e milhões de cores, portanto, é tido como o padrão atual.

GDDR - memória de vídeo

Outro item de extrema importância em uma placa de vídeo é a memória. A velocidade e quantidade desta pode influenciar significantemente no desempenho do dispositivo. Esse tipo de componente não difere muito das [memórias RAM](https://www.infowester.com/memoria.php) que são tipicamente utilizadas em PCs, sendo inclusive relativamente comum encontrar placas que utilizam chips de memória de tecnologias DDR, DDR2 e DDR3. No entanto, placas mais avançadas e atuais contam com um tipo de memória específica para aplicações gráficas: memórias **GDDR** (Graphics Double Data Rate).

As memórias GDDR são semelhantes aos tipos de memória DDR, mas são especificadas de maneira independente. Basicamente, o que muda entre essas tecnologias são características como voltagem e frequência.

Trabalhando com tensão de 2,5 V e frequência de até 500 MHz, a memória GDDR1 (que é praticamente o padrão DDR) foi até que foi bastante utilizada, mas logo perdeu a preferência para memórias GDDR3 (ao menos nas placas mais avançadas). O tipo GDDR2 teve pouca utilização, pois embora trabalhe com frequências de até 500 MHz, sua tensão é de 2,5 V, resultando em um grande problema: calor excessivo.

As memórias GDDR3 surgiram como uma solução para essa questão porque trabalham com tensão de 1,8 V (podendo em alguns casos trabalhar também com 2 V) e aliam isso com maior velocidade, pois trabalham com 4 transferências de dados por ciclo de clock (contra duas dos padrões anteriores). Sua frequência, em geral, é de 900 MHz, mas pode chegar a 1 GHz. Tecnicamente, são semelhantes às [memórias DDR2](https://www.infowester.com/memddr2.php).

Memórias GDDR4, por sua vez, são parecidas com [memórias DDR3](https://www.infowester.com/memddr3.php), lidam com 8 transferências por ciclo de clock e utilizam tensão de apenas 1,5 V. Além disso, empregam tecnologias como DBI (Data Bus Inversion) e Multi-Preamble para diminuir o "delay", isto é, o tempo de espera existente na transmissão de dados. No entanto, sua frequência permanece, em média, na casa dos 500 MHz, devido a sua "suscetibilidade" a problemas de ruído (interferências). Por esse motivo, essa tecnologia de memória GDDR tem aceitação baixa no mercado.

Em relação às memórias GDDR5, esse tipo suporta frequências semelhantes e até ligeiramente maiores que as taxas utilizadas pelo padrão GDDR3, porém trabalha com 8 transferências por ciclo de clock, aumentando significantemente seu desempenho. Além disso, também conta com tecnologias como DBI e Multi-Preamble, sem contar a existência de mecanismos que oferecem melhor proteção contra erros.

Barramentos

Para fazer com que a placa de vídeo se comunique com o computador, é necessário utilizar uma tecnologia padronizada de comunicação ou, mais precisamente, barramentos. Existem várias tecnologias para isso, sendo algumas exclusivas para placas de vídeos.

No início de 1990 surgiu o barramento **PCI** (Peripheral Component Interconnect), cuja capacidade de lidar com 32 bits por vez e o seu clock de 33 MHz resultava na possibilidade de permitir taxas de transferências de até 132 MB por segundo (houve também uma versão de 64 bits e clock de 66 MHz, mas que foi pouco utilizada pela indústria). É um padrão já em desuso, embora não seja difícil encontrar placas-mãe relativamente recentes que ainda o suportam. Uma grande variedade de modelos de placas de vídeo utilizou essa tecnologia. Tal como o ISA, também foi utilizado para outros tipos de dispositivos, como placas de modem e placas de rede.

O fato é que a evolução das GPUs é uma constante e, logo, o padrão PCI se mostrou incapaz de lidar com a quantidade de dados utilizada pelos chips gráficos. A solução para esse problema veio em 1996, com o lançamento do barramento **AGP** (Accelerated Graphics Port), criado especificamente para placas de vídeos.

A primeira versão do AGP trabalha a 32 bits e com frequência de 66 MHz, resultando em uma taxa de transferência máxima de 266 MB por segundo, podendo ser dobrada com a capacidade da tecnologia de permitir transferência de 2 dados por ciclo de clock (modo 2X). A última versão do AGP, a 3.0, é capaz de trabalhar com até 8X, resultando em taxas de transferência de até de 2.133 MB por segundo.

Apesar de permitir grandes avanços, o AGP não resistiu à tecnologia **PCI Express**, que se destaca por estar disponível em vários modos, como 1X, 2X, 8X e 16X. Dispositivos que trabalham com taxas menores de transferência de dados podem utilizar o PCI Express 1X, por exemplo, pois seu slot é muito menor. Placas de vídeo, no entanto, trabalham com PCI Express 16X, que permite taxas de transferência de cerca de 4 GB por segundo. Com o PCI Express 2.0, apresentado em 2007, esse valor pode dobrar.

Novas Placas de vídeo e Ray Tracing

O ray tracing é um tipo de renderização de gráficos 3D por computador. A técnica promete ser um diferencial para placas de vídeo ([GPU](http://www.techtudo.com.br/dicas-e-tutoriais/noticia/2014/03/o-que-e-gpu-e-como-funcionam-os-modelos-existentes-no-mercado.html)) de nova geração e games que buscam oferecer imagens impressionantes.

[Desenvolvido por Nvidia](https://www.techtudo.com.br/noticias/2018/03/nvidia-apresenta-tecnologia-rtx-para-graficos-com-qualidade-de-cinema.ghtml), [AMD](http://www.techtudo.com.br/tudo-sobre/amd.html) e [Microsoft](http://www.techtudo.com.br/tudo-sobre/microsoft.html), a implementação da tecnologia tem origem na criação visual do cinema, em efeitos de computação gráfica. O ray tracing se compromete a reproduzir luzes e sombras mais fiéis à realidade nas cenas de jogos. Há a possibilidade de que o recurso permita jogos com qualidade de imagem concorrente aos efeitos do cinema e da TV.

O que é ray tracing?

Ray tracing é um método de sintetização (também entendido como renderização) de gráficos tridimensionais por computadores. Essa técnica busca criar imagens do ambiente 3D de um jogo e reproduzir, na tela ou em óculos de realidade virtual, uma simulação do trajeto feito pelos raios de luz dos objetos até olhos do jogador. A ideia é se aproximar da realidade e da interação física entre o usuário e o local.

O método pode ser entendido por uma técnica de renderização de imagens com o objetivo de apresentar, com a maior naturalidade possível, fenômenos de iluminação, sombra e contraste. Essas características só são observadas em espaços reais, fora dos games.

Novidade entre os produtos de fabricantes como AMD e [Nvidia](http://www.techtudo.com.br/tudo-sobre/nvidia.html), o conceito de ray tracing não é inédito. Esse artifício é usado por estúdios de cinema na criação de efeitos por computação gráfica e em grandes produções de TV há alguns anos. Por isso, uma das formas de entender o impacto dessa tecnologia em jogos do futuro é compará-lo com a alta qualidade dos recursos visuais feitos por Hollywood.

Entenda como funciona



O recurso age por meio de uma simulação da trajetória feita pela luz sobre a superfície de cada pixel em direção aos olhos do gamer — algo que explica o nome do mecanismo, traduzido por “raios de luz”. Assim, características do comportamento da luz no cenário podem ser levadas em consideração durante o desenvolvimento de um jogo. O resultado é a capacidade de simular efeitos de luz, sombra e contrastes de forma mais precisa e possibilitar um cenário mais imersivo e próximo da realidade.

Jogos com qualidade de computação gráfica de cinema

A razão pela chegada tardia na informática de games é o alto processamento exigido para renderizar imagens com essa técnica. Nos filmes, as cenas que se valem da método precisam ser sintetizadas apenas uma vez por computadores especializados. Enquanto nos jogos, os cenários e objetos requerem renderização de 30 vezes por segundo — no mínimo.

Por exemplo, em uma tela de 1920 x 1080 pixels, mais de 2 milhões de pixels precisam ser atualizados 30 vezes por segundo (maioria dos games exigem o dobro), enquanto o jogo está em execução. No caso de uma cena renderizada via ray tracing, são mais de dois milhões de raios de luz que precisam ser calculados e representados dezenas de vezes por segundo para gerar uma imagem de boa qualidade. O processamento gráfico pode ser ainda mais dispendioso em telas 4K.