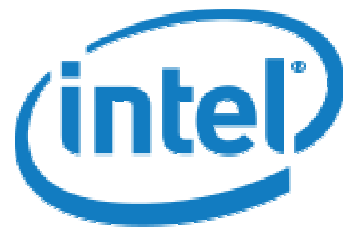


# GPUs for Noobs

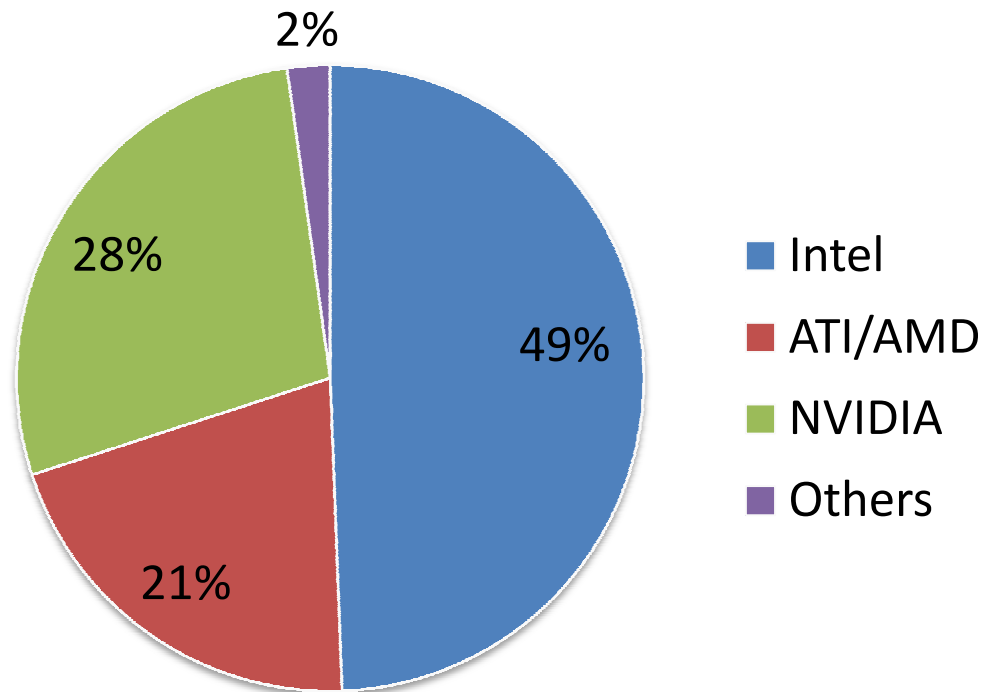


# What is a GPU ?

- A GPU (Graphics Processing Unit) is a dedicated graphics rendering device for a personal computer, workstation, or game console.
- Is a processor attached to a graphics card dedicated to calculating floating point operations and the like.
- Incorporates custom microchips which contain special mathematical operations commonly used in graphics rendering. The efficiency of the microchips therefore determines the effectiveness of the graphics accelerator.
- Implements a number of graphics primitive operations in a way that makes running them much faster than drawing directly to the screen with the host CPU.
- Modern GPUs are very efficient at manipulating and displaying computer graphics.
- Highly parallel structure makes them more effective than general-purpose CPUs for a range of complex algorithms.

# GPU makers

Market Share of Major GPU makers



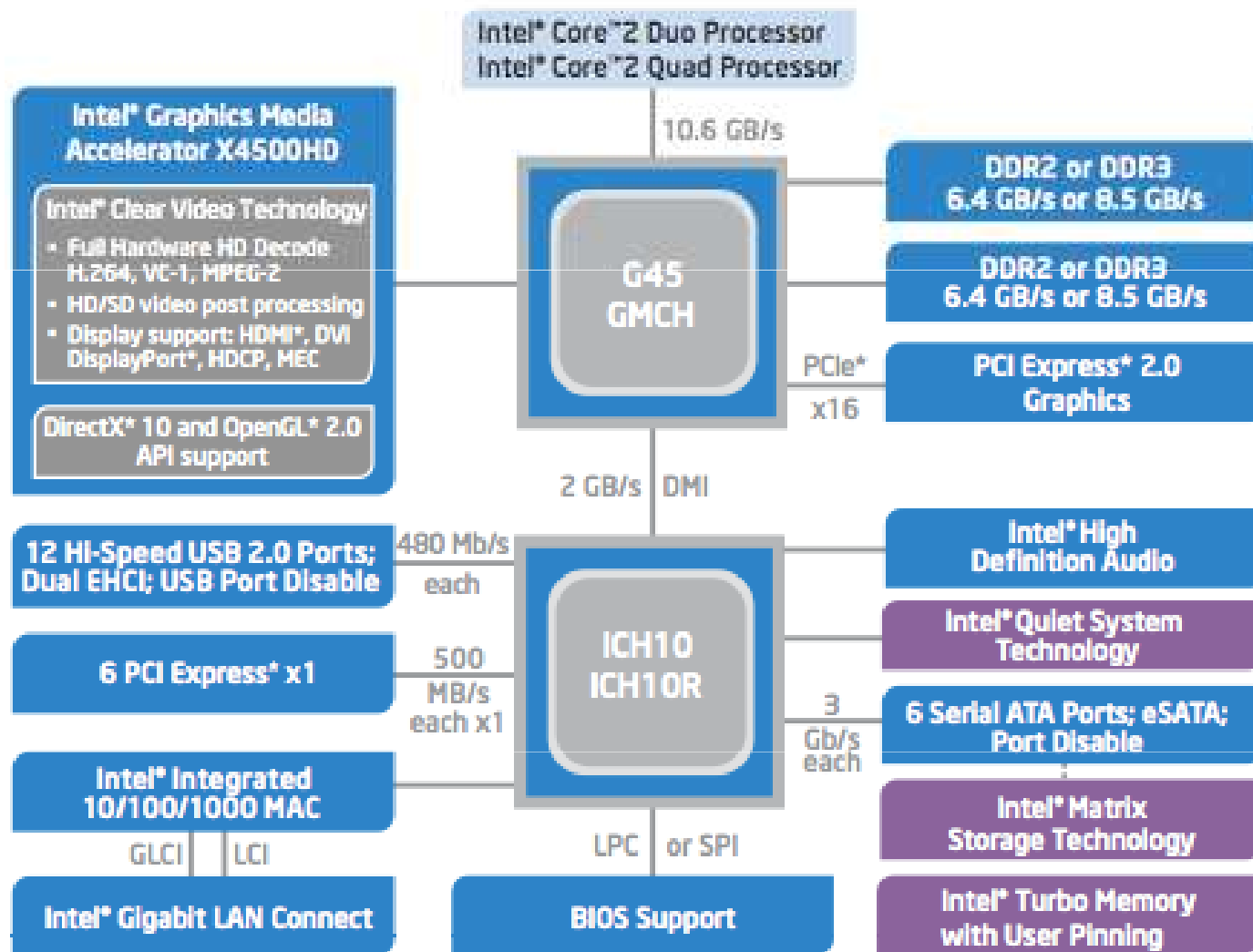
- Numbers include Intel's very low-cost, less powerful integrated graphics solutions as GPUs.
- Discounting those numbers, NVIDIA and AMD control nearly 100% of the market.
- Other makers are VIA Technologies/S3 Graphics and Matrox .

# Types of GPUs

- Integrated
  - Utilize a portion of a computer's system RAM rather than dedicated graphics memory.
  - Reduce cost, power consumption and noise, but are less capable than a discrete GPU.
  - An integrated solution finds itself competing for the system RAM with the CPU as it has minimal or no dedicated video memory. System RAM may be 2 Gbit/s to 12.8 Gbit/s, but dedicated GPUs enjoy between 10 Gbit/s to over 100 Gbit/s of bandwidth.



Intel GMA X3000



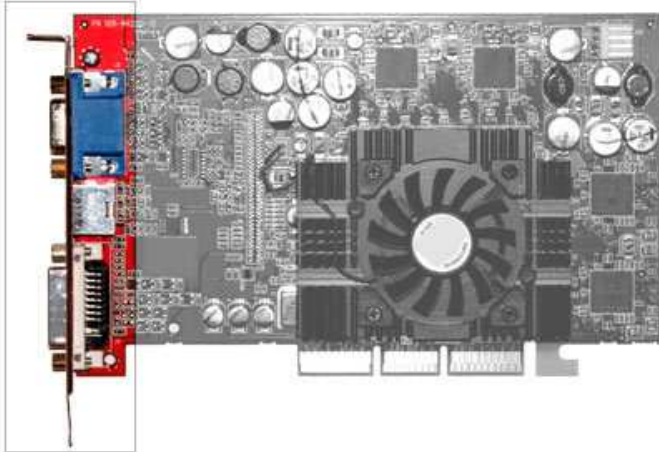
Intel G45 architecture diagram

# Types of GPUs

- Dedicated (Graphics Cards)
  - The most powerful class of GPUs.
  - Interfaced with the motherboard by means of an expansion slot (PCIe, AGP), can usually be replaced or upgraded with relative ease.
  - The term "dedicated" refers to the RAM that is dedicated to the card's use (unlike Integrated GPUs).
  - Multiple GPUs can be used to draw single image, increasing the processing power.



# The Basic Parts



Output



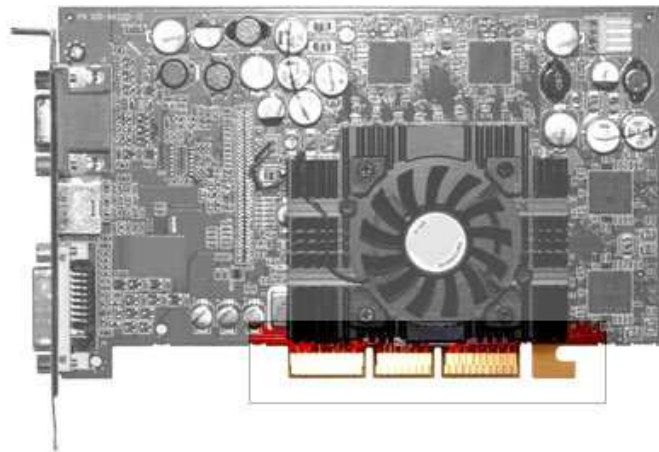
VGA: Analog-based standard adopted in the late 1980s designed for CRT displays.



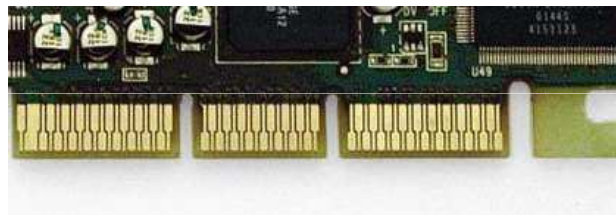
DVI: Digital-based standard designed for flat-panel displays and projectors.



S Video: Allows the connection with televisions, DVD players, video recorders etc.



Interface

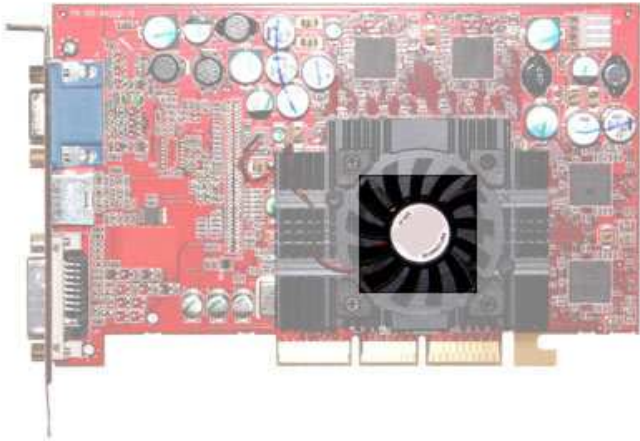


AGP: Created in late 90s, peak b/w 2 GBps. Has been replaced by PCI-E.

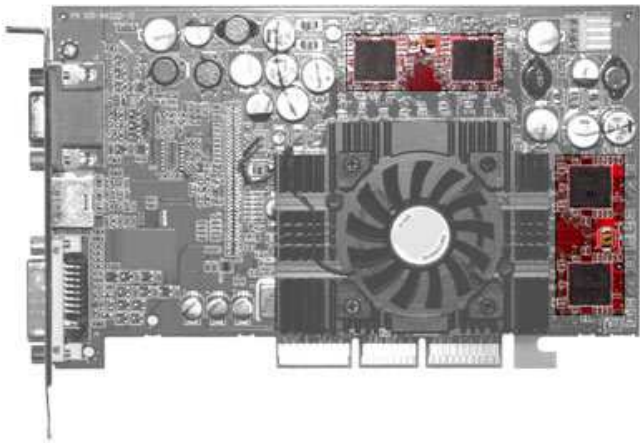


PCI-E: Current graphics interface, peak b/w 8GBps.

# The Basic Parts



The Graphics Processor: The single most important part of the graphics card. Almost all hardware specifications, such as pixel shaders, vertex shaders, pipelines, and component clock speeds refer to the architecture and capabilities of the graphics processor.



The Video Memory: Ranges from 128 MB to 4 GB. Most modern cards have a 128/256/512 bit wide memory bus. The effective memory clock rate in modern cards are generally between 400 MHz and 3.8 GHz. Used for storing screen images and other data as well (e.g Z-Buffer).

# Some Basic Graphics Terms

- Vertex: A point in 3D space with X, Y, Z coordinates.
- Texture: Is simply a 2D image, the size of which varies, that is applied to a 3D object to simulate its surface.
- Shader: Are programs to calculate rendering effects. There are 2 forms of shaders currently.
  - Vertex Shaders: They deform or transform 3D elements.
  - Pixel Shaders: Calculate the color of individual pixels, are typically used for scene lighting and related effect.

Pixel shader programs allow graphics cards to process spectacular effects like this shimmering water in "Elder Scrolls: Oblivion."



# Graphics Processor Architecture: Features

- Vertex Processors (a.k.a. Vertex Shader Units): Vertex processors are components on the graphics processor designed to process shaders that affect only vertices. Are important in 3D scenes with many or complex 3D objects. Less relevant to overall performance than pixel shaders.
- Pixel Processors (a.k.a. Pixel Shader Units): A pixel processor is a component on the graphics chip devoted exclusively to pixel shader programs. Because pixels represent color values, pixel shaders are used for all sorts of impressive graphical effects.
- Unified Shaders: The upcoming DirectX 10 specification calls for a unified shader architecture i.e the vertex geometry and pixel shader code structures will be functionally similar but have dedicated rolls. Found in Xbox 360 that was developed by ATI for Microsoft.

# Graphics Processor Architecture: Features

- Texture Mapping Units (TMUs): TMU's job is to apply texture operations to pixels. Works in conjunction with pixel and vertex shader units.
- Raster Operator Units (a.k.a. ROPs): The raster operation processors are responsible for writing pixel data to memory. The speed at which this is done is known as the fill rate.
- Pipelines: Pipeline is a term used to describe the graphics card's architecture, and it provides a generally accurate idea of the computing power of a graphics processor.

# Graphics Processor Architecture: Technology

- **Manufacturing Process:** Refers to the structural size and precision of the manufacturing process used to create an integrated circuit. The smaller the size, the smaller and more advanced the manufacturing process. e.g 45nm, 65nm, 90nm. Shortening of distances, lowered voltages and other advantages allow smaller process products to have higher clock frequency speeds.
- **Graphics Processor Clock Speed:** Has a direct effect on the performance of the graphics processor. If the processor is technically identical, a clock speed boost translates into higher performance. But a technically superior GPU can easily perform better at lower clock speed.

# Graphics Processor Architecture: Technology

- Local Graphics Memory:
  - Graphics Memory Size: The amount of RAM has a very small impact on performance when compared to other considerations like clock speed and the memory interface. Ranges from 128 MB to 4 GB.
  - Memory Bus: Is one of the most important aspects of memory performance. Can range from 64 bits to 512 bits.

# Multi-Card Solutions

- Relatively modern approach towards high performance Graphics rendering. (2000- )
- Has unmatched performance outputs for high end cards.
- It is economically costlier than a single card system, requires specific architecture, consumes more power.
- The 2 available multi GPU solutions are NVIDIA's SLI and ATI/AMD's Crossfire.

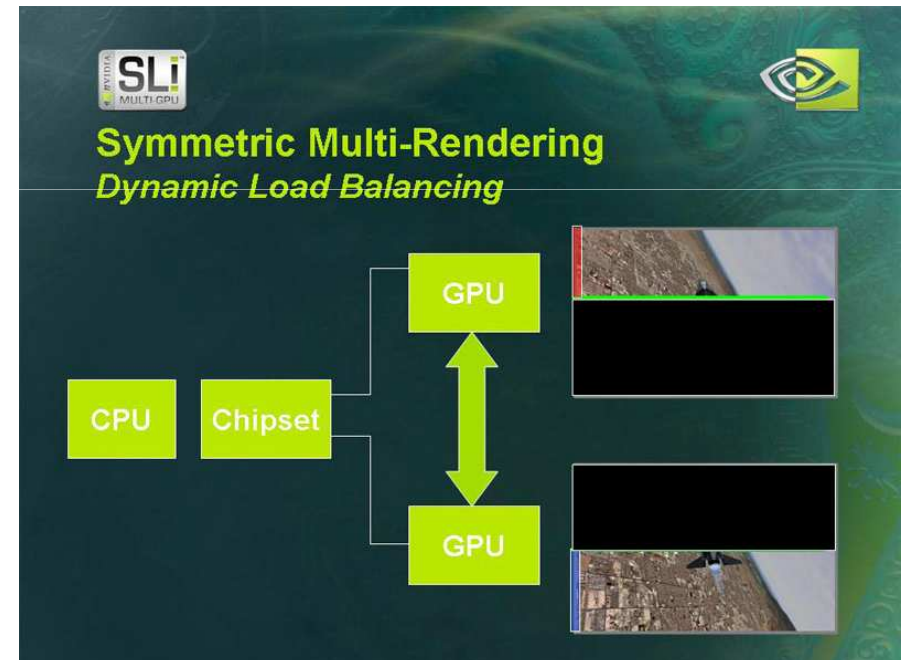


A 2 GPU system using NVIDIA SLI.

# SLI and Crossfire



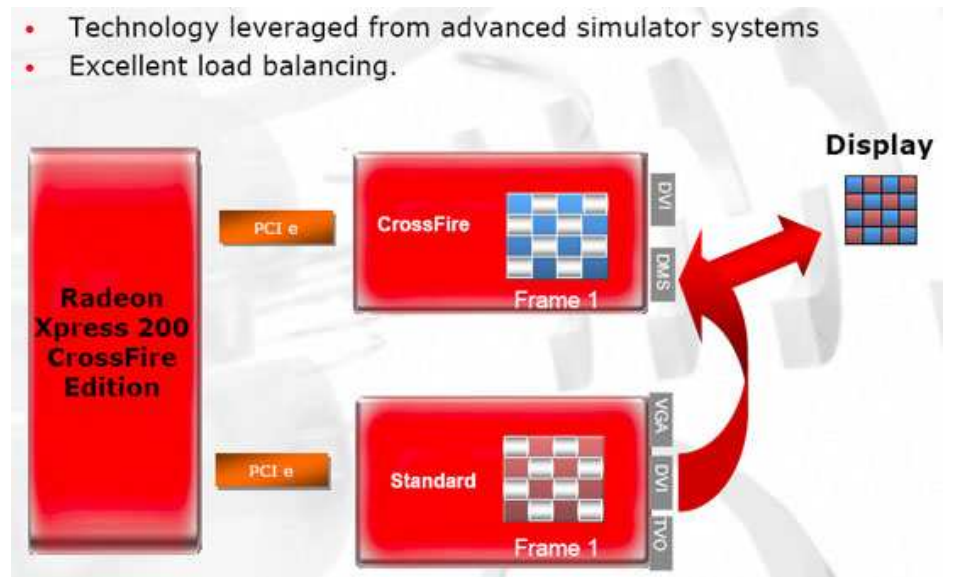
- NVIDIA SLI (Scalable Link Interface)
  - Works by evenly distributing the workload to each GPU and assigning a fraction of the total rendering to be done on the screen.
  - The area assigned to each GPU is continuous. (the area divided by the green line can be seen in the figure).
  - Uses a SLI bridge to interface the 2 cards. The 2 cards must be identical.



# SLI and Crossfire



- ATI/AMD Crossfire X
  - Uses dynamic load balancing and Super-Tiling to distribute the work load to each GPU.
  - Different GPUs can be used in a Crossfire X system.
  - Uses a ribbon-like connector attached to the top of each graphics adapter.



# GPGPU

- A new concept is to use a modified form of a stream processor to allow a general purpose graphics processing unit.
- Turns the massive floating-point computational power of a modern graphics accelerator's shader pipeline into general-purpose computing power.
- In certain circumstances the GPU calculates forty times faster than the conventional CPUs traditionally used in such applications.
- NVIDIAs API extension to the C programming language called CUDA ("Compute Unified Device Architecture"), allows specified functions from a normal C program to run on the GPU's stream processors.
- This makes C programs capable of taking advantage of a GPU's ability to operate on large matrices in parallel, while still making use of the CPU where appropriate.

# Thanks for Bearing :)

Siddharth Singh 2K6/COE/179  
Yashwant Bisht 2K6/COE/192

Sources: <http://en.wikipedia.org>  
<http://www.tomshardware.com>