# CSE 591: GPU Programming

## Introduction

Klaus Mueller

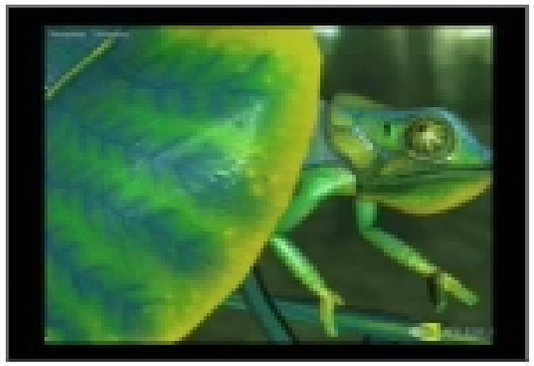
Computer Science Department

Stony Brook University

## Entertainment Graphics: Virtual Realism for the Masses

Computer games need to have:

- realistic appearance of characters and objects
- believable and creative shading, textures, surface details
- realistic physics and kinematics
- effects need to be customizable and interactive



## High Performance Computing on the Desktop

PC graphics boards featuring GPUs:

- NVidia FX, ATI Radeon
- available at every computer store for less $500
- set up your PC in less than an hour and play

the latest board:
Nvidia GeForce GTX 280

## "Just" Computing

Compute-only (no graphics): NVIDIA Tesla 1060

True GPGPU
(General Purpose Computing using GPU Technology)

4 GB memory per card

Bundle up to 4 cards: 960 processors, 16 GB memory

## Incredible Growth


- 1.3 TFLOPS (GTX 480) 500$

Performance gap GPU / CPU is growing

- currently 1-2 orders of magnitude is achievable (given appropriate programming and problem decomposition)

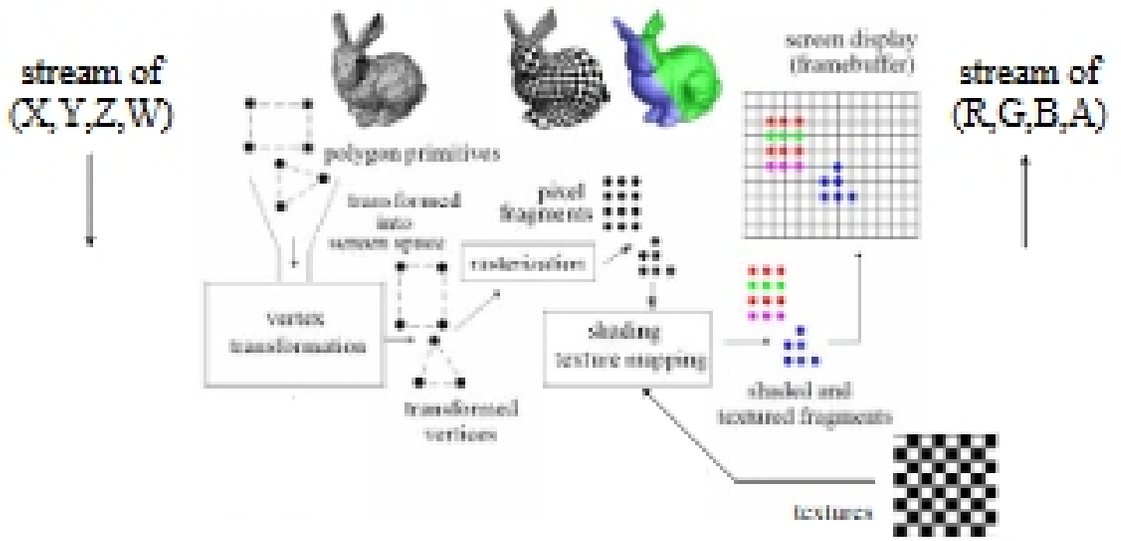
## History: Accelerated Graphics

1990s: rise (and fall) of Silicon Graphics (SGI)

- in 1991 pioneers the first graphics accelerator
- developed OpenGL
- evolving 2D to 3D graphics capabilities through mid-90s
- desktop to high-end performance workstations (O2, Octane, Onyx)
- #1: expensive
- #2: non-programmable

## The Graphics Pipeline

Old-style, non-programmable:




## History: Cheap Consumer Graphics

Late 1990s: rise of consumer graphics chips

- 1994: Voodoo graphics chip (and later the popular Voodoo 2)
- chips still separate from memory
- other chips soon emerge: ATI Rage, NVIDIA Riva

End 1990s: consumer graphics boards with high-end graphics

- transform, lighting, setup, and rendering all on a single GPU
- the world's first GPU: NVIDIA GeForce 256 (NV 10)
- → #1: inexpensive
- #2: non-programmable

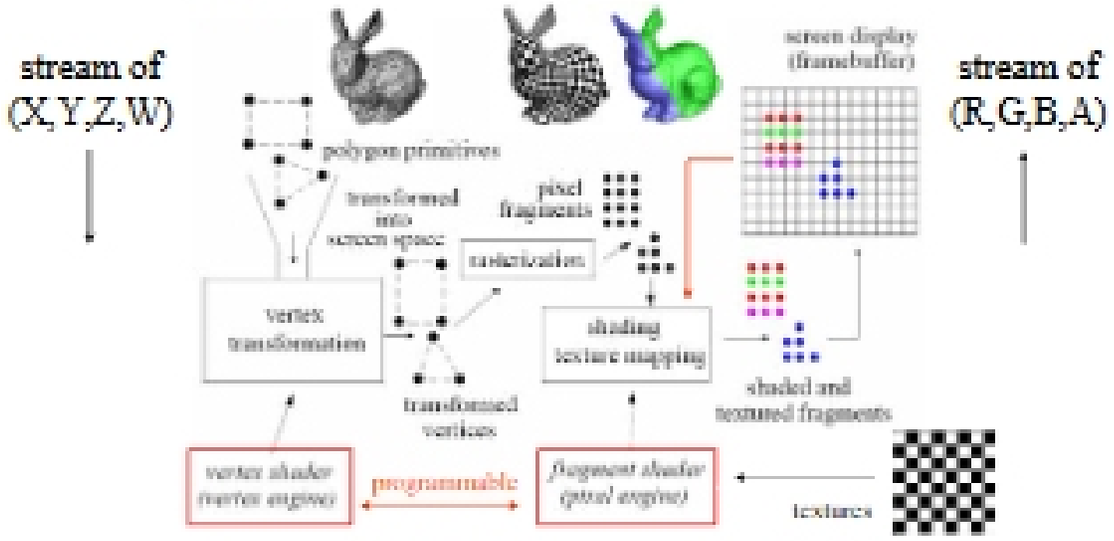
## History: Programmability, GPGPU

2000s: emergence of programmable consumer graphics hardware

- programmable vertex and fragment shaders
- graphics cards: NVIDIA GeForce 3, ATI Radeon 9700
- evolving capabilities for floating point, loops, if's
- enabled GPGPU
- HW programming languages: CG, GLSL, HLSL
- SW graphics API: OpenGL, DirectX
- → #1: inexpensive
- → #2: programmable

## The Graphics Pipeline

Modern, programmable:




## History: Focus Parallel Computing

2006: parallel computing languages appear

- address the need to provide dedicated SDK and API for parallel high performance computing (GPGPU)
- CUDA (Compute Unified Device Architecture)
  - developed by NVIDIA
- OpenCL (Open Computing Language)
  - initially developed by Apple
  - now with the Khronos Compute Working Group
- specific GPGPU boards: NVIDIA Tesla, AMD FireStream
- other parallel-computing chips: Intel Larabee, IBM Cell BE

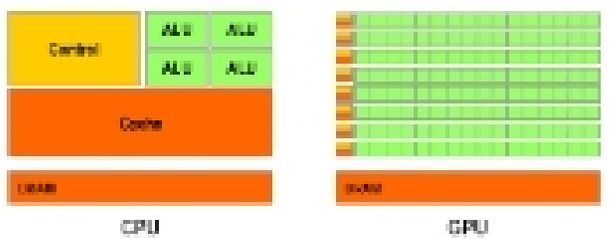
## Hardware Architecture

GPU

- Compute-intensive
- Highly data parallel
- Original SIMD

Programming language

- Expose the parallel capabilities of GPUs.




ALU: Arithmetic logic unit