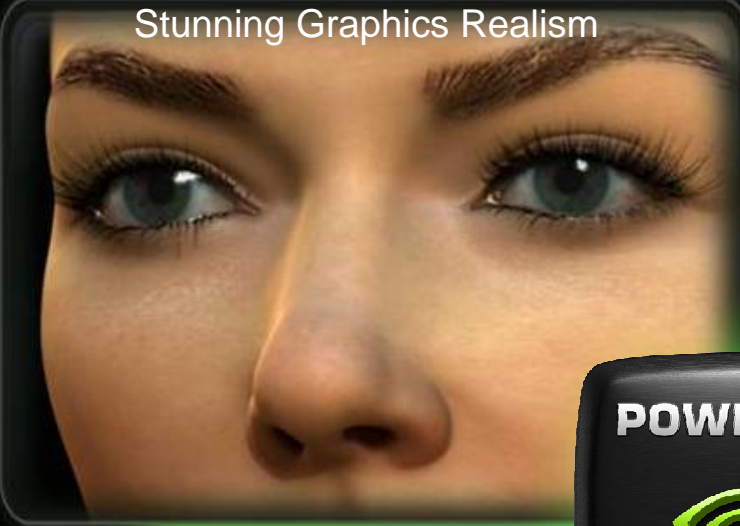


**NVIDIA**

**GeForce 8800 & NVIDIA CUDA**

**A New Architecture for Computing on the GPU**

Stunning Graphics Realism



Lush, Rich Worlds



POWERED BY



NVIDIA

Crysis © 2006 Crytek / Electronic Arts

Incredible Physics Effects



Core of the Definitive Gaming Platform



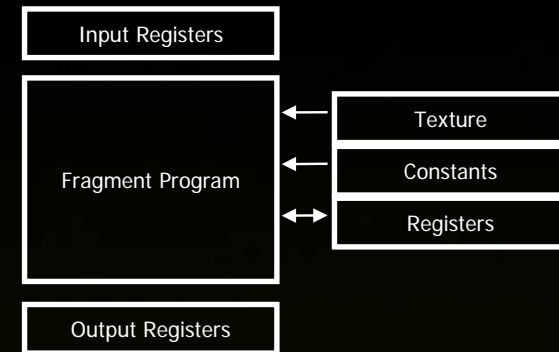
© NVIDIA Corporati

Hellgate: London © 2005-2006 Flamingo Studios, Inc. Licensed by NAMCO BANDAI Games America, Inc.

Evil Spectrum Warrior: Ten Hammers © 2006 Pandemic Studios, LLC. All rights reserved. © 2006 THQ Inc. All rights reserved.

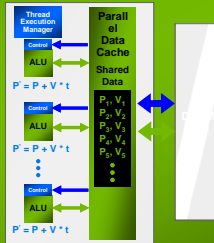
# Current Constraints

- **Graphics API**
- **Addressing modes**
  - Limited texture size/dimension
- **Shader capabilities**
  - Limited outputs
- **Instruction sets**
  - Integer & bit ops
- **Communication limited**
  - Between pixels
  - Scatter  $a[i] = p$



# CUDA & GPU Computing

## New Architecture for Computing

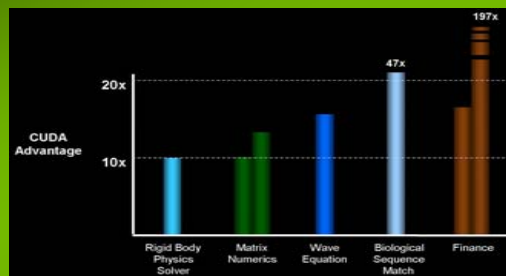


## Standard C Programming

```

dim3 DimGrid(100, 50); // 5000 thread blocks
dim3 DimBlock(4, 8, 8); // 256 threads per block
size_t SharedMemBytes = 64; // 64 bytes of shared memory
KernelFunc<<< DimGrid, DimBlock, SharedMemBytes >>>(...);
  
```

## Unprecedented Performance



## New Applications

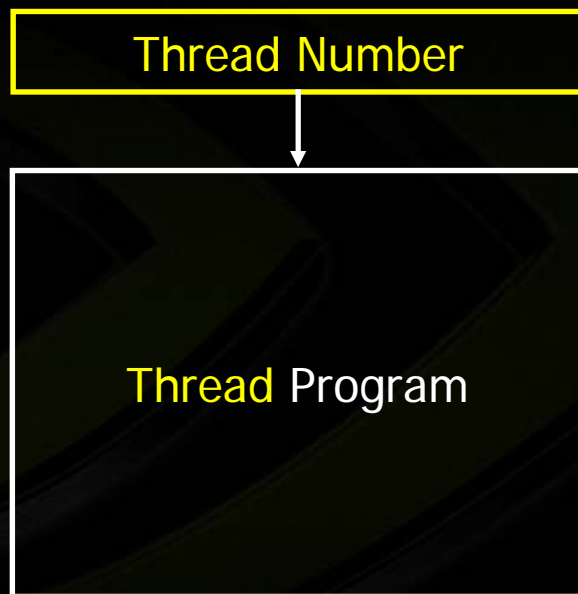


# CUDA & GPU Computing



- **CUDA is a completely new architecture and programming model for general-purpose computation on GPUs**
- **Hardware and Software designed together**
  - **NOT a new driver for old GPU architectures**
- **Data-parallel computing with thousands of threads**
- **Parallel Data Cache helps increase arithmetic intensity for massive speedups**
- **Program in C**
- **BLAS and FFT libraries**

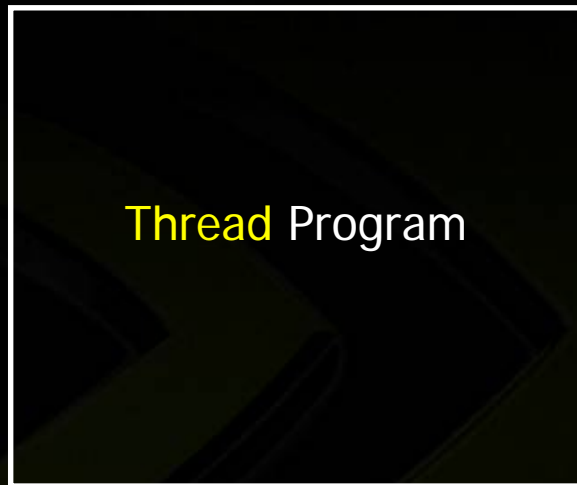
# Thread Programs



## Features

- Millions of instructions
- Full Integer and Bit instructions
- No limits on branching, looping
- 1D, 2D, or 3D thread id allocation

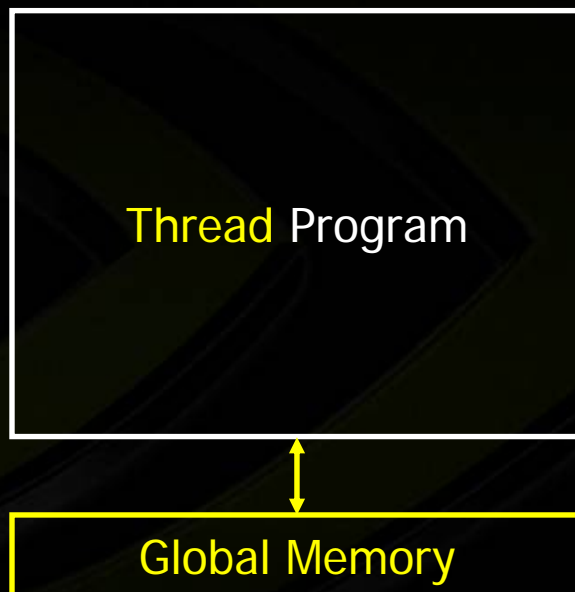
# Parallel Data Cache



Parallel Data Cache

- **Dedicated on-chip memory**
- **Shared between threads for inter-thread communication**
- **Explicitly managed**
- **As Fast Registers**

# Global Memory



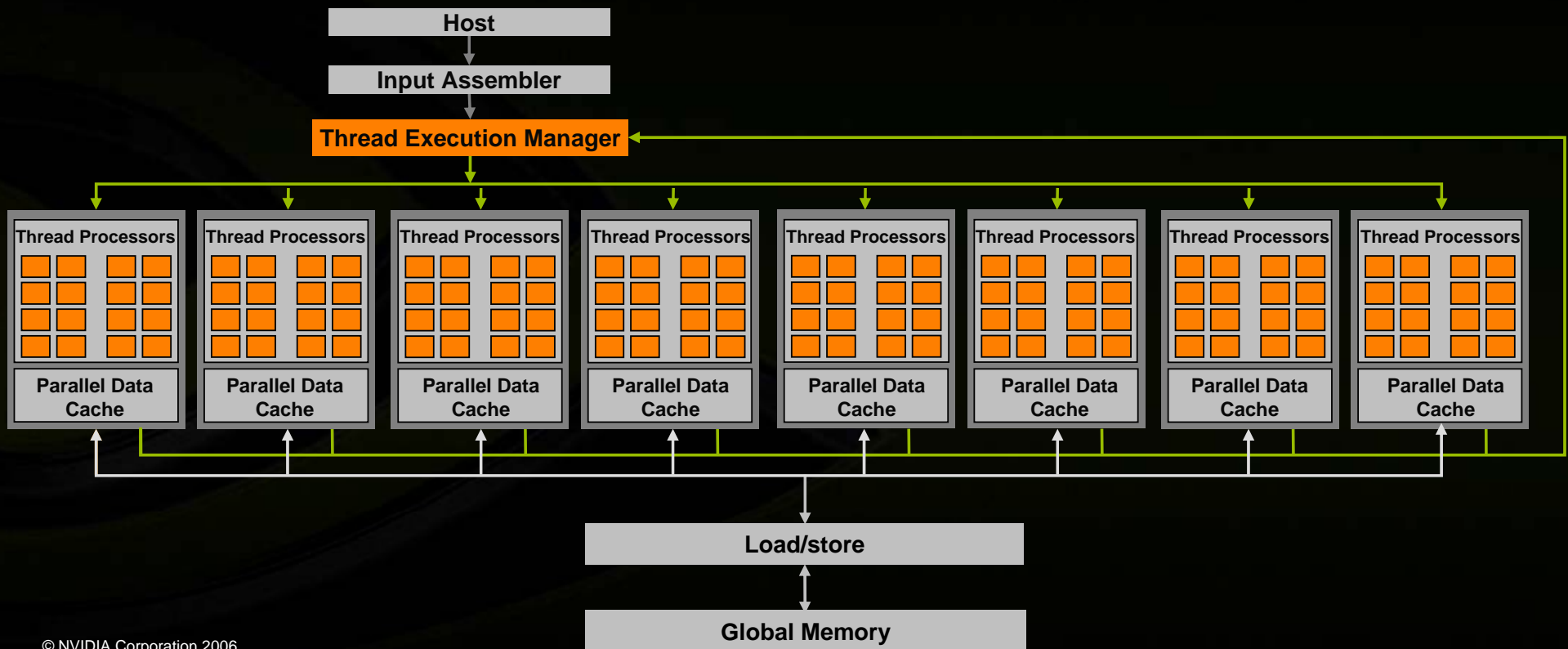
## Features

- Fully general load/store to GPU memory: Scatter/Gather
- Programmer flexibility on how memory is accessed
- Untyped, not limited to fixed texture types
- Pointer support

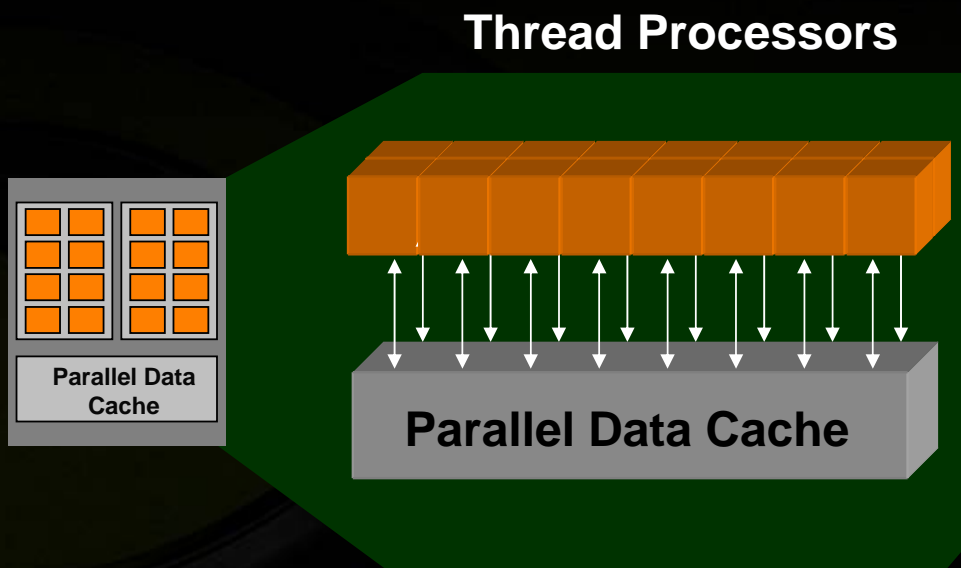
# GeForce 8800 GPU Computing



- Processors execute computing threads



# Thread Processor



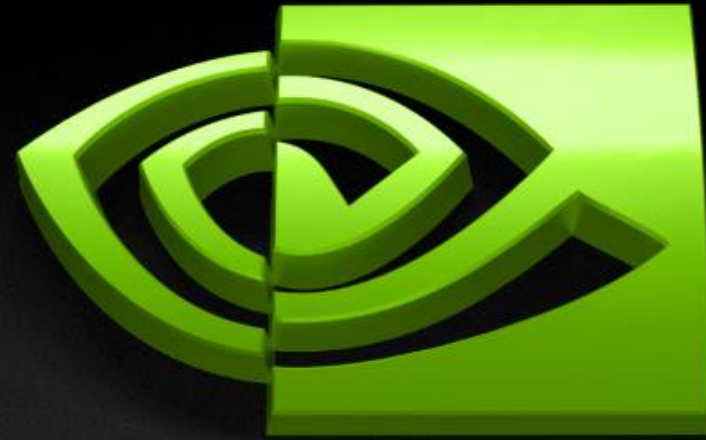
- 128, 1.35 GHz processors
- Parallel Data Cache per cluster 16KB
- Scalar architecture
- IEEE 754 Precision
- Full featured instruction set

# C on the GPU

- Provide a simple, explicit programming language solution
- Extend only where necessary

```
__global__ void KernelFunc(...);  
__device__ int GlobalVar;  
__shared__ int SharedVar;
```

```
KernelFunc<<< 500, 128 >>>(...);
```



**NVIDIA**

**Full presentation at workshop:**

**General-Purpose GPU Computing: Practice and Experience**

**Salon I (Marriott)**

**Monday 1:50-2:30**