Case Study: NVIDIA GeForce 8800 Series

gf8-1

References

gf8-2

gf8-3

Important Features

More General

Single set of processors used for all shaders.

Full integer support.

Shaders have normal memory access capabilities.

Still GPU Specialized

Rasterization hardware.

Texture address and fetch units.

FB update (ROP) hardware.

Interesting Fact

Used as reference GPU when designing DirectX10.

gf8-4

Major Parts

- Host Interface
- Data Assembler
- Multiprocessor.
- FB Update.

gf8-5

Multiprocessor

Consists of 8 **stream processors** operated in SIMD mode.

gf8-5

gf8-6

Stream Processor

Note: Name also used as a kind of computer organization.

Two-way superscalar.

Highly multithreaded.

Part of a SIMD array.

gf8-7

Stream Processor Data Types

Full integer support. (32-bit integers)

IEEE 754 singles.

No vector data types.

gf8-7

gf8-8

Stream Processor Instruction Highlights

Based on vague descriptions, no ISA available.

Full set of integer instructions.

No vector instructions.

Probably no divide instruction.

Reciprocal, reciprocal square root instructions.

gf8-8

gf8-9

Stream Processor Instruction API

API: `NV_gpu_program4` (less reflective of microarchitecture).

Much less reflective of true ISA than GF 3 and GF 6 APIs.

More like a compiler intermediate language.

Differences with true instructions:

- Instructions operate on vectors, even though machine lacks them.
- True ISA lacks transcendental instructions such as cos, log.
- No way to address registers directly.
- API instructions translated into true instructions.

gf8-10

Stream Processor Storage - Direct Access

1024 32-bit registers (total, not per thread).

Constant memory: 64 kiB, 8 kiB cache per MP.

Cached r/w memory: 16 kiB (L1, per MP).

Also can r/w uncached memory, with 400-600 cycle latency.

gf8-11

Stream Processor Storage - Mediated Access

Mediated Access means access through some other processor.

RO (read-only) access to texture through texture fetch unit.

RMW (read-modify-write) access to FB via ROP processors.

gf8-12

Unknown about GeForce 8800

Is there a special datapath and storage for streamed data?

Are the constants accessed as other cached data?

gf8-13

Multiprocessor & Thread Sequencing

Based on CUDA, assumed true for graphics use.

MP consists of 8 SPs.

Unit of control is a set of 32 threads called a **warp**.

Entire warp resident on one MP.

A single PC is used for an entire warp.

Takes 4 cycles to issue one insn of warp (even if ¡ 24 threads).

Executes one or both sides of branch, as necessary.

gf8-14

Multiprocessor & Thread Resources

Maximum program size: 2 million instructions (CUDA).

Maximum number of resident threads: 768.

gf8-15

Threads and Registers and Latency

How many threads? Let n denote number of threads per MP.

Registers: $8192/n$. At maximum: 10 registers per thread.

At maximum (768): 10 registers per thread.

At minimum (32): 256 registers per thread.

Maximum latency (CUDA 5.1.2.5) 24 cycles or 192 threads.