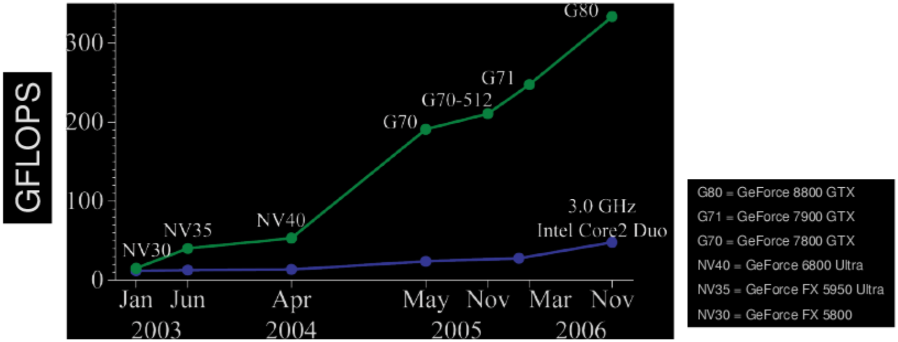
Multi- and Many-core Architectures for Scientific Programming

F. S. Schifano[1]

[1]University of Ferrara and INFN-Ferrara

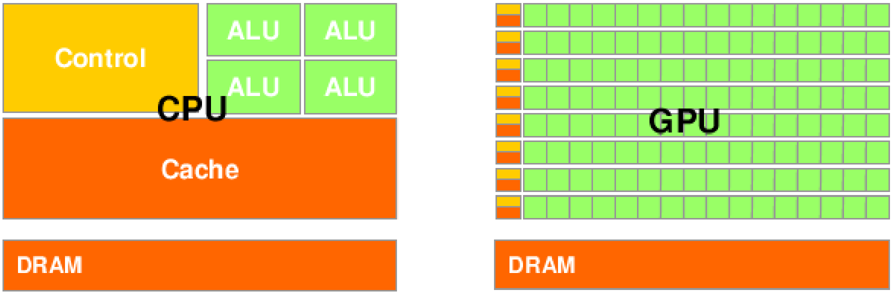
Workshop Commissione Calcolo e Reti INFN 2009
May 11 – 15 2009, Palau Olbia

Motivation: GPU Evolution



- GPUs evolve much faster in terms of raw-computing power
- Fast-growing video-game market forces innovation

GPUs vs CPUs



- GPUs specialized for highly data-parallel and intensive computation (exactly what rendering is about)
- more transistors devoted to **data-processing** rather than **data caching** and **flow-control**

NVIDIA GT200 Specs

Processor:

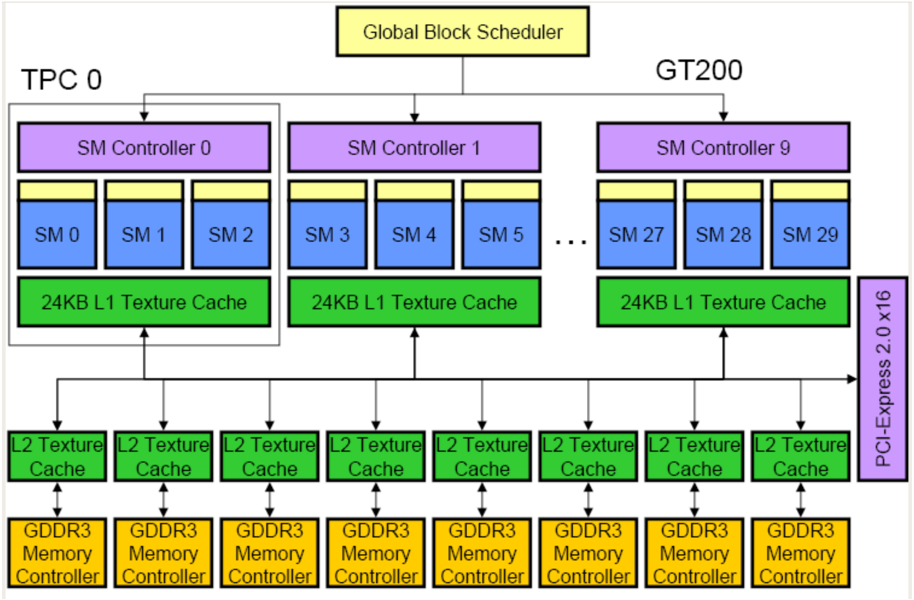
- 1.3 GHz, **936 Gflops SP**, 78 Gflops DP
- 10 **Texture Processor Cluster** (TPC)
- 1 TPC includes 3 **Streaming Multiprocessor** (SM) + 1 Texture Memory
- 1 SM include 8 **Streaming Processor** (SP)
  loosely corresponding to a modern CPU-core with 8-way SIMD computing-unit
- a total of $10 \times 3 \times 8 = 240$ **threads**

Memory:

- 4GB Global Memory, 512-bit, 102.4 GB/s, 400 . . . 600 cycles of latency
- Constant memory 64 KB (RO)
- Texture memory 256 KB (RO, two dimensional locality)

160 Watt (typical, w/Memory), $<$ 6 Gflops/Watt (SP, w/MUL), 1.5 K €

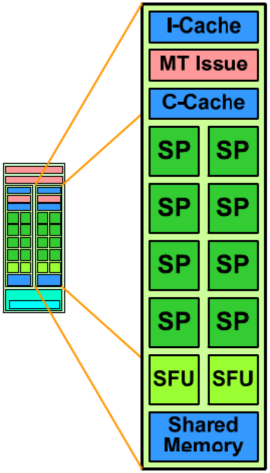
Architecture of GP-GPU: NVIDIA GT200



NVIDIA GT200 SM Specs

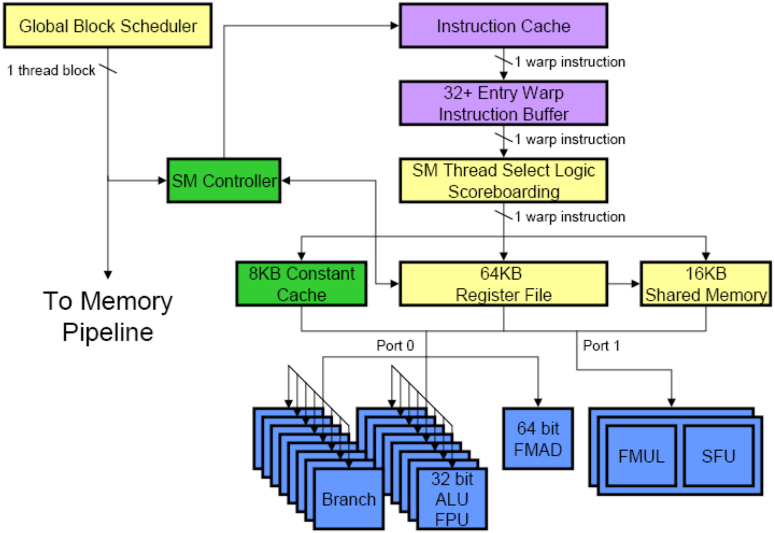
Each SM includes:

- 64KB register file (2KB for each SP)
- 16KB shared memory
- 8KB constant cache
- 8 32-bit ALU/FPU
- 1 64-bit FMAD
- 8 branch units

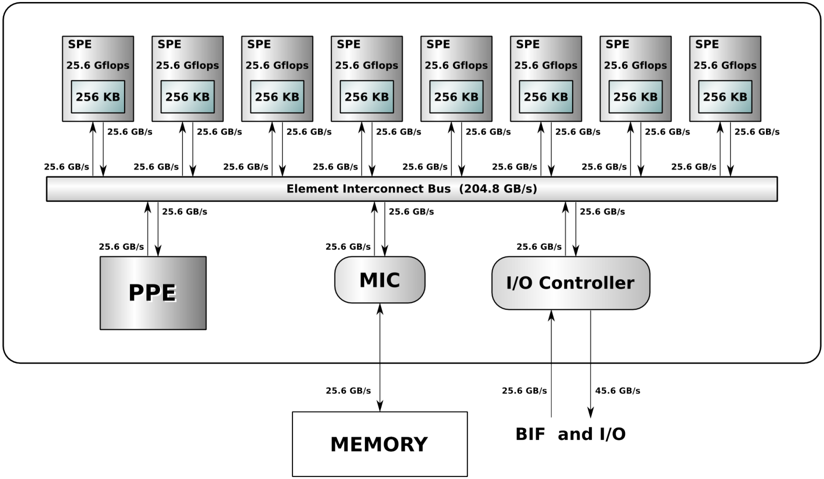


1 SM executes in parallel up to 8 thread and manages up to 1024 threads

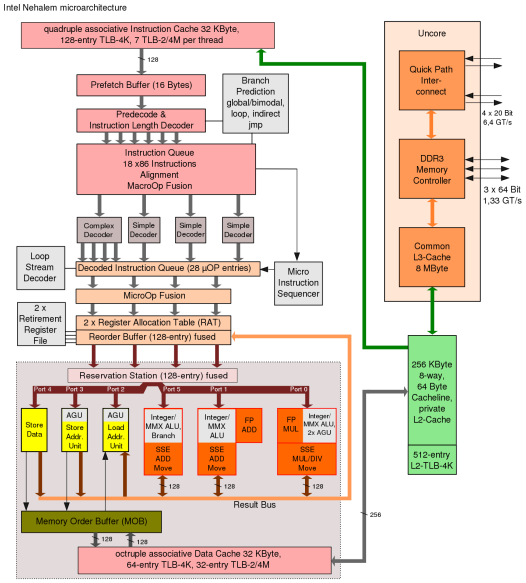
The SM Architecture



The Milticore Cell-BE Architecture



The Multicore Nehalem Architecture


Intel Nehalem microarchitecture
quadruple associative Instruction Cache 32 KByte, 128-entry TLB-4K, 7 TLB-2/4M per thread
128
Prefetch Buffer (16 Bytes)
Branch Prediction global/bimodal, loop, indirect jmp
Predecode & Instruction Length Decoder
Instruction Queue 18 x86 Instructions Alignment MacroOp Fusion
Complex Decoder
Simple Decoder
Simple Decoder
Simple Decoder
Loop Stream Decoder
Decoded Instruction Queue (28 µOP entries)
Micro Instruction Sequencer
MicroOp Fusion
2 x Retirement Register File
2 x Register Allocation Table (RAT)
Reorder Buffer (128-entry) fused
Reservation Station (128-entry) fused
Port 4
Port 3
Port 2
Port 5
Port 1
Port 0
Store Data
AGU Store Addr. Unit
AGU Load Addr. Unit
Integer/ MMX ALU, Branch
SSE ADD Move
Integer/ MMX ALU
SSE ADD Move
FP ADD
FP MUL
Integer/ MMX ALU, 2x AGU
SSE MUL/DIV Move
128
Result Bus
Memory Order Buffer (MOB)
octruple associative Data Cache 32 KByte, 64-entry TLB-4K, 32-entry TLB-2/4M
256
Uncore
Quick Path Inter-connect
4 x 20 Bit 6,4 GT/s
DDR3 Memory Controller
3 x 64 Bit 1,33 GT/s
Common L3-Cache 8 MByte
256 KByte 8-way, 64 Byte Cacheline, private L2-Cache
512-entry L2-TLB-4K

GT/s: gigatransfers per second

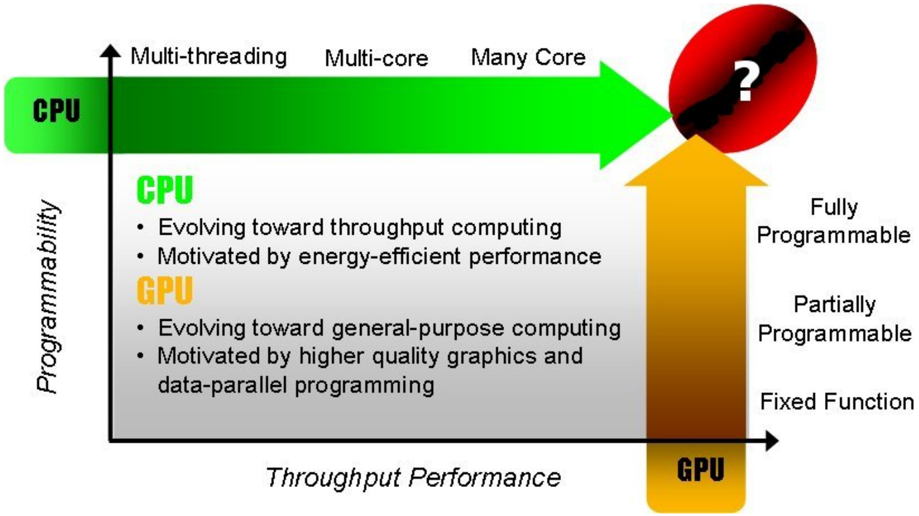
Multicore Performance for Spin-Glass Simulations

		Binary model					Gaussian model				
L	C	V	w	Mem	SUT	GUT	V	w	Mem	SUT	GUT
16	8	8	16	local	0.83	0.052	4	1	local	1.00	0.250
16	16	8	16	local	1.17	0.073	4	1	local	1.34	0.335
32	8	16	8	local	0.40	0.050	4	1	local	0.65	0.162
32	16	16	8	local	0.26	0.032	4	1	local	0.42	0.105
48	8	8	16	local	0.48	0.030	4	1	global	1.65	0.412
48	16	8	16	local	0.25	0.016	4	1	local	0.42	0.105
64	8	32	4	local	0.29	0.072	4	1	global	1.71	0.427
64	16	32	4	local	0.15	0.037	4	1	global	2.23	0.557
80	8	8	16	local	0.82	0.051	4	1	global	1.59	0.397
80	16	8	16	local	1.03	0.064	4	1	global	2.10	0.525
96	8	16	8	global	0.42	0.052	-	-	-	-	-
96	12	16	8	global	0.41	0.051	-	-	-	-	-
128	8	64	2	global	0.24	0.120	-	-	-	-	-
128	16	64	2	local	0.12	0.060	-	-	-	-	-

Perfomance are 1 order of magnitude far from that of the

Janus system for which we have a SUT of 0.016 ns/spin

Where are we going ?



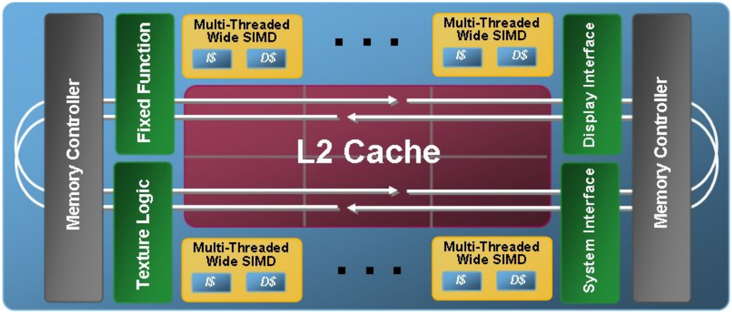
CPU

- Evolving toward throughput computing
- Motivated by energy-efficient performance

GPU

- Evolving toward general-purpose computing
- Motivated by higher quality graphics and data-parallel programming

The Larrabee Processor



- **in-order** cores, SIMD-16
- 512-bit SIMD vector unit, 32-, 64-bit operations
- 32-bit vector registers
- 8 MB L2 cache shared across the cores

2 Tflops SP, 2GHz, 32-cores, **300 W** (?)

How to go Beyond Petaflops ?

- currently scientific project attack Petaflop perfomance connecting commodity processors by a 3D-torus custom-network (a la APE):
  - QPACE: German project based on Cell
  - Aurora: Italian project based on Nehalem
- for the future one possible option would be to use 2-4 Tflops GP-GPU interconnected by a high-scalable 3D-torus network