Parallel & Distributed Processing II: *parallel processing on manycore chips*

Manycore Architectures

Eric Aubanel
Winter 2010, UNB Fredericton

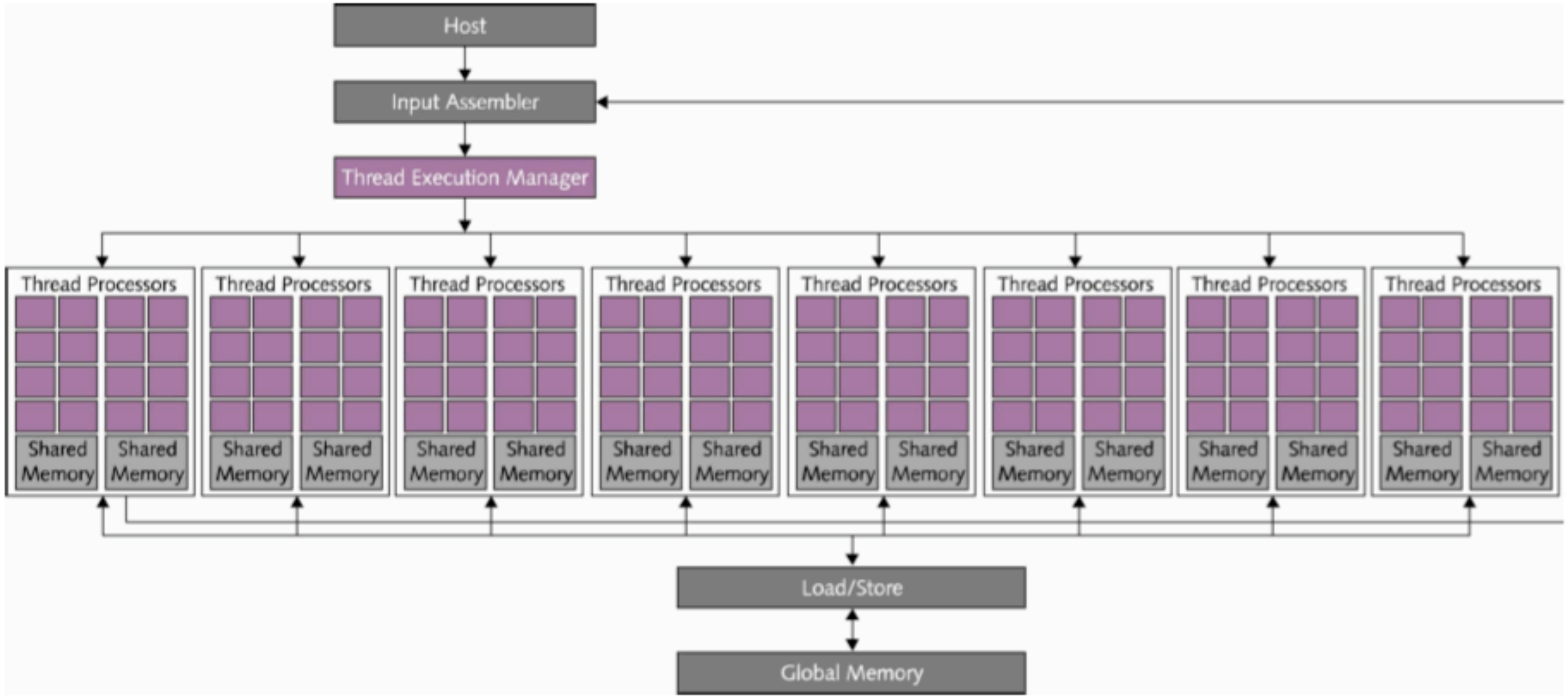
Focus on 4

- NVIDIA
- AMD
- Intel Larrabee
- STI Cell BE

GPUs: AMD and NVIDIA

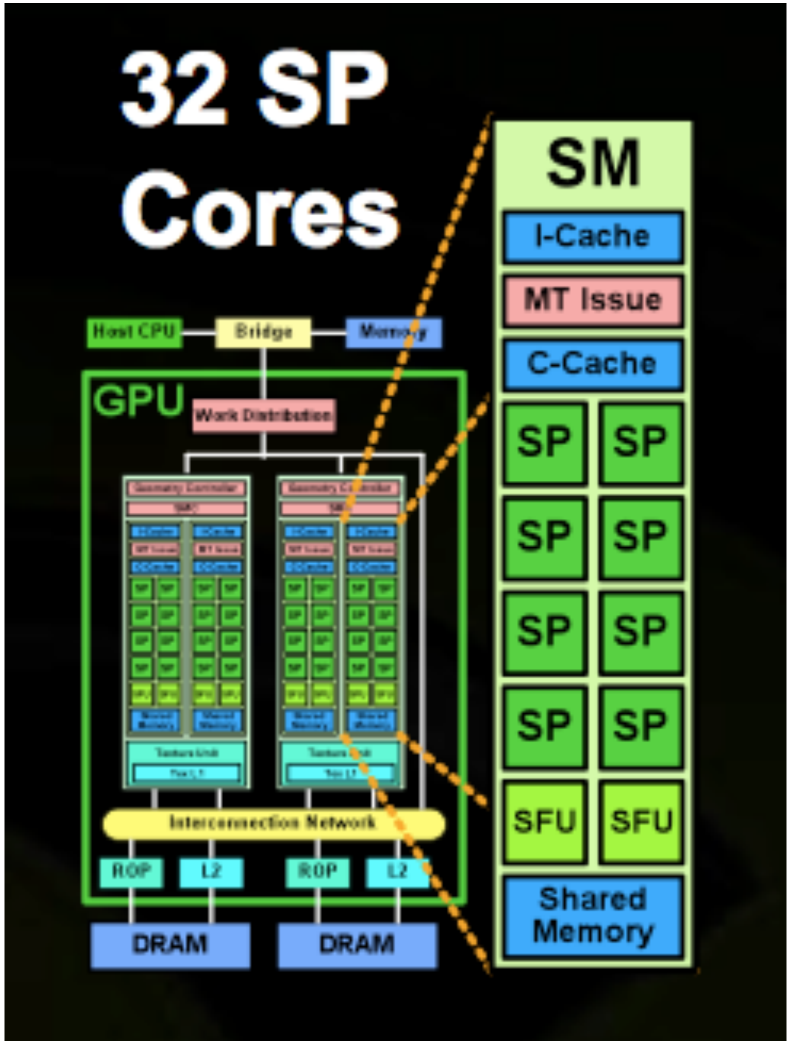
- Programmable processors
  - SIMD hardware
  - Can alternate among tasks in graphics pipeline
  - Can be used for general purpose applications
- SIMT execution model
  - Large number of threads to mask memory latency
    - The infamous memory wall
    - E.g. NVIDIA GeForce 285 GTX: 7 flops for every byte transferred to/ from off-chip memory
    - CPUs use large caches for this
  - Scheduled in groups ("warps" for NVIDIA, "wavefronts" for AMD)
  - Threads can be masked to allow conditional execution
    - As in conventional SIMD processor arrays

NVIDIA Tesla



Source: http://www.nvidia.com/docs/IO/55972/220401_Reprint.pdf

Streaming Multiprocessor

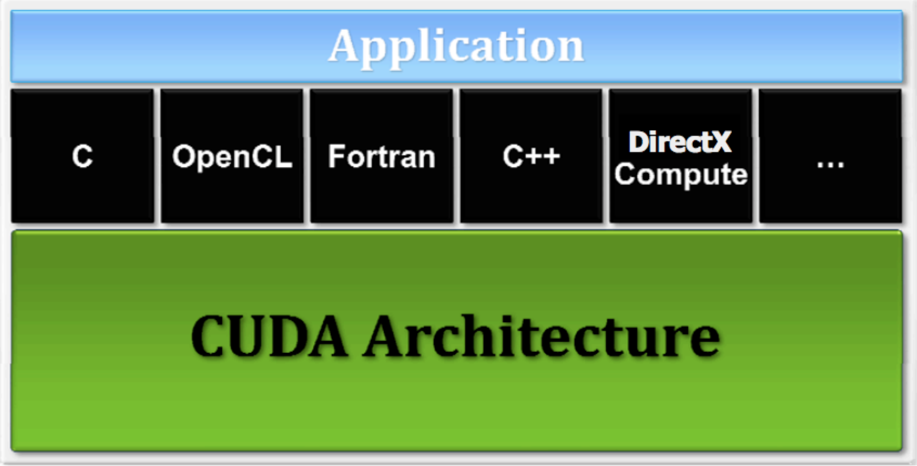


Source: Patrick LeGresley, NVIDIA

NVIDIA GeForce GTX 285

- 240 cores
- 1 Tflop peak single-precision
- 1 GB on-board memory
- 124 GB/s memory bandwidth

CUDA

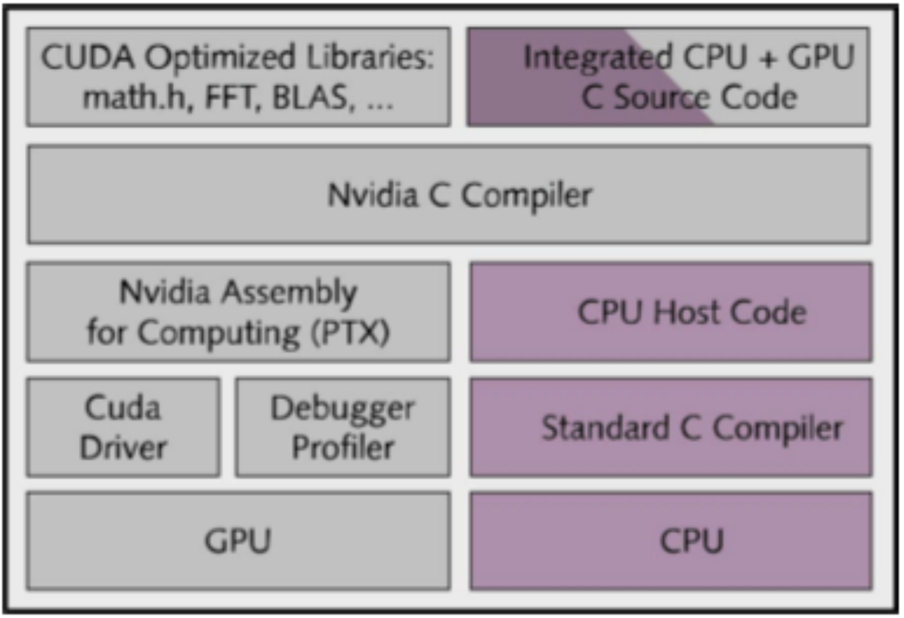


Application

C | OpenCL | Fortran | C++ | DirectX Compute | ...

CUDA Architecture

Source: NVIDIA CUDA Programming Guide 2.2.1



CUDA Optimized Libraries: math.h, FFT, BLAS, ...

Integrated CPU + GPU C Source Code

Nvidia C Compiler

Nvidia Assembly for Computing (PTX)

CPU Host Code

Cuda Driver

Debugger Profiler

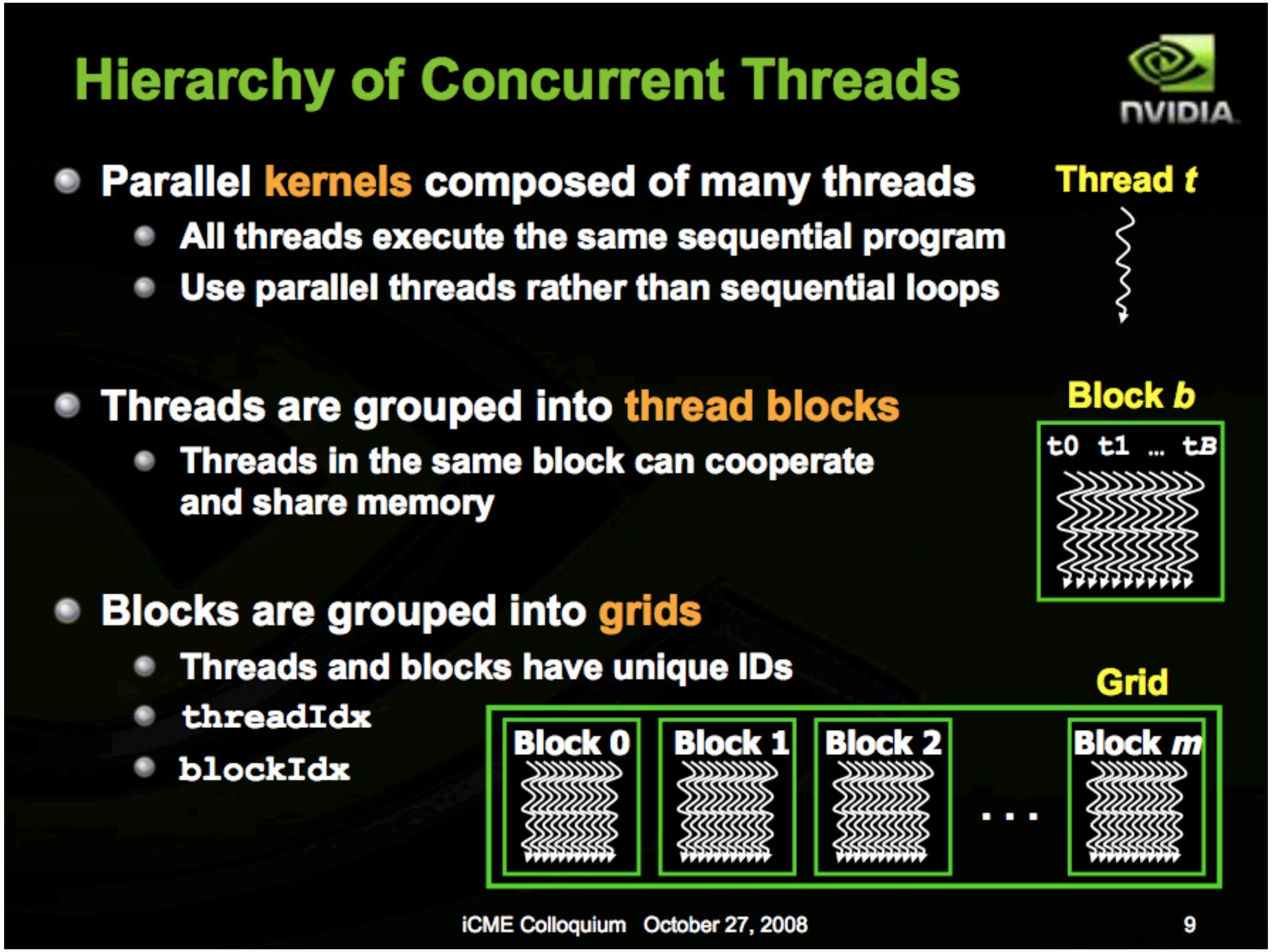
Standard C Compiler

GPU

CPU

Source: http://www.nvidia.com/docs/IO/55972/220401_Reprint.pdf

CUDA



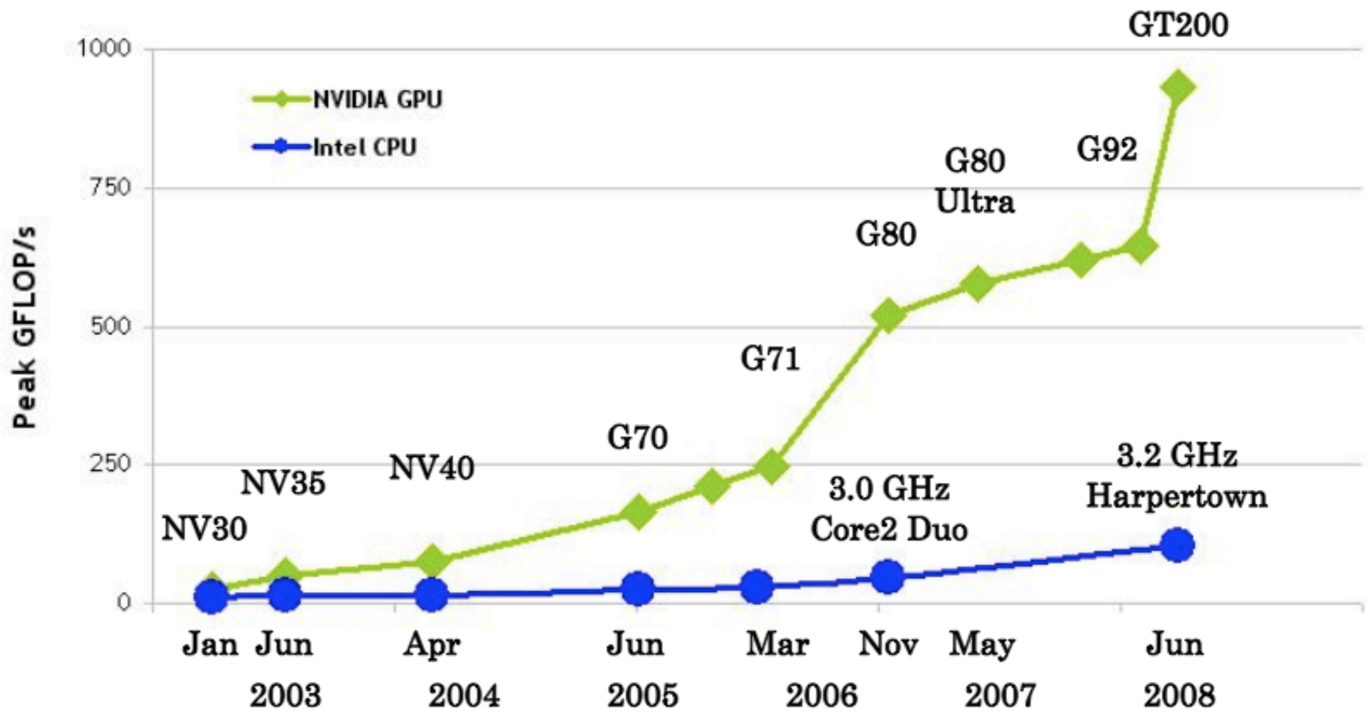
Hierarchy of Concurrent Threads

- Parallel **kernels** composed of many threads
  - All threads execute the same sequential program
  - Use parallel threads rather than sequential loops
- Threads are grouped into **thread blocks**
  - Threads in the same block can cooperate and share memory
- Blocks are grouped into **grids**
  - Threads and blocks have unique IDs
  - `threadIdx`
  - `blockIdx`

iCME Colloquium October 27, 2008 9

Source: Patrick LeGresley, NVIDIA

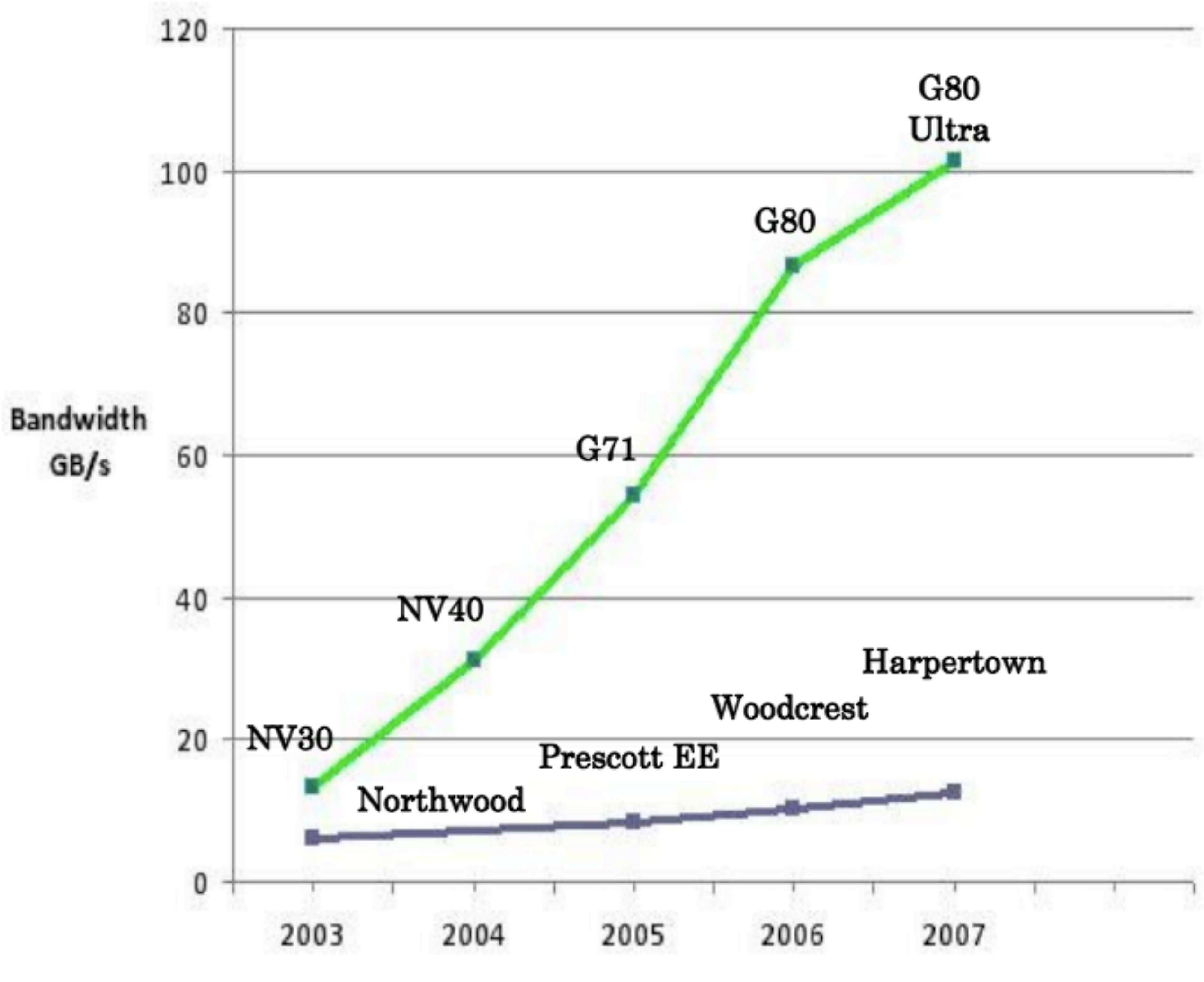
NVIDIA Performance: *flops*



GT200 = GeForce GTX 280	G71 = GeForce 7900 GTX	NV35 = GeForce FX 5950 Ultra
G92 = GeForce 9800 GTX	G70 = GeForce 7800 GTX	NV30 = GeForce FX 5800
G80 = GeForce 8800 GTX	NV40 = GeForce 6800 Ultra	

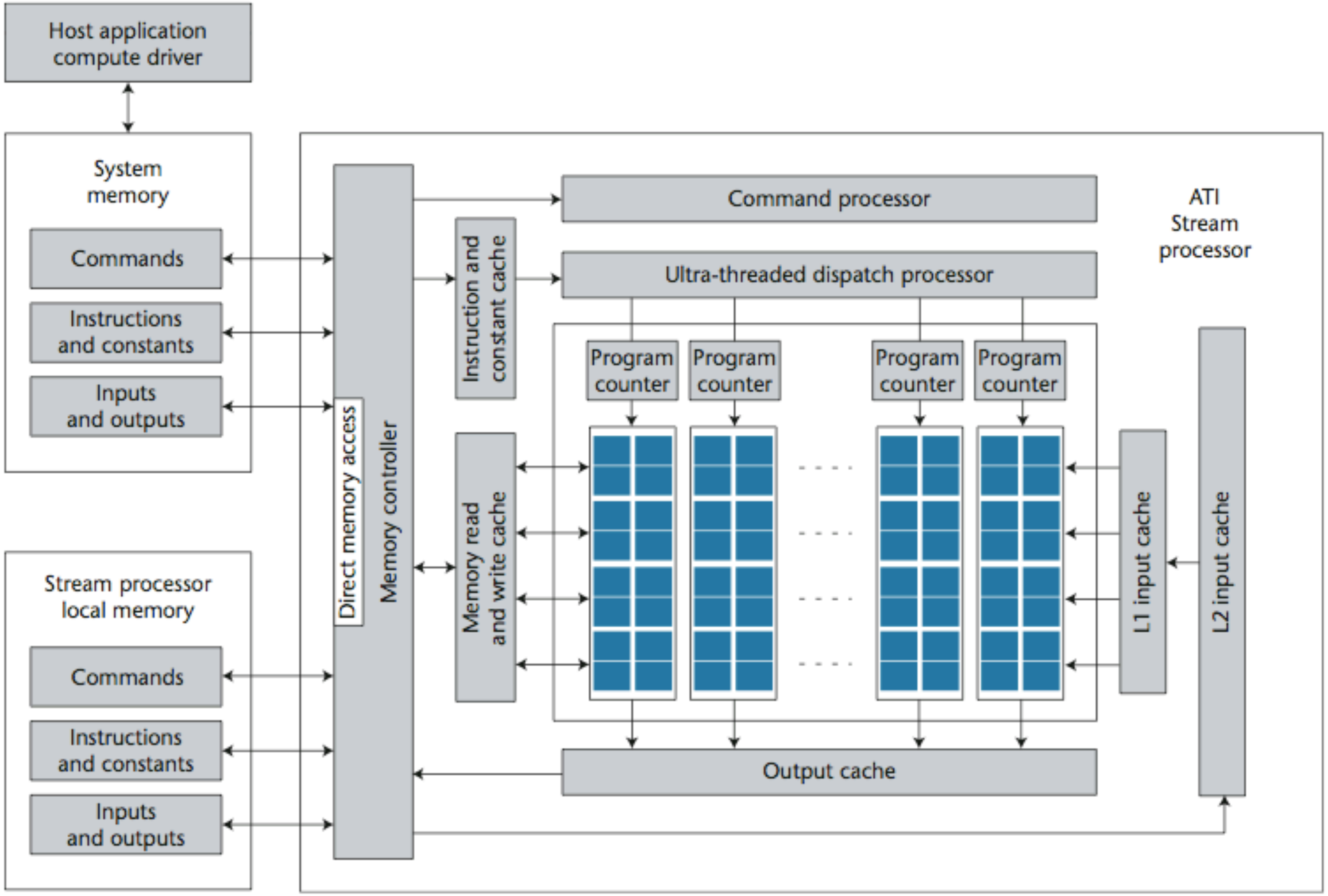
Source: NVIDIA CUDA Programming Guide 2.2.1

NVIDIA Performance: *bandwidth*



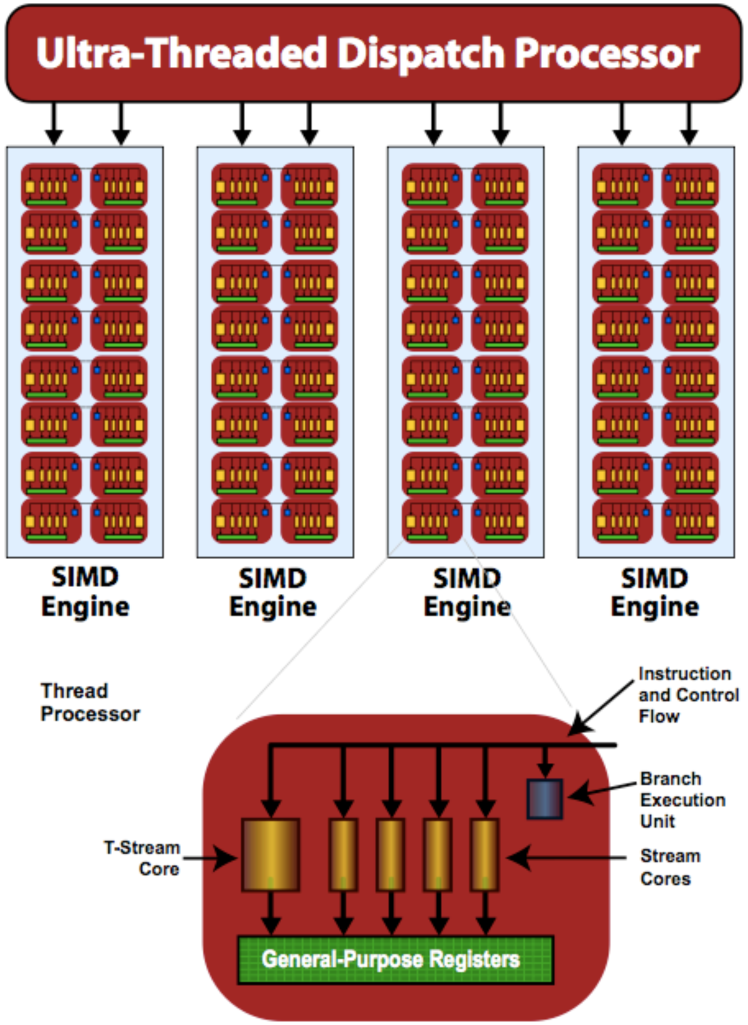
Source: NVIDIA CUDA Programming Guide 2.2.1

ATI/AMD: FireStream 9270 GPU



Source: Computing in Science & Engineering, Nov.-Dec. 2009

AMD Stream Processor



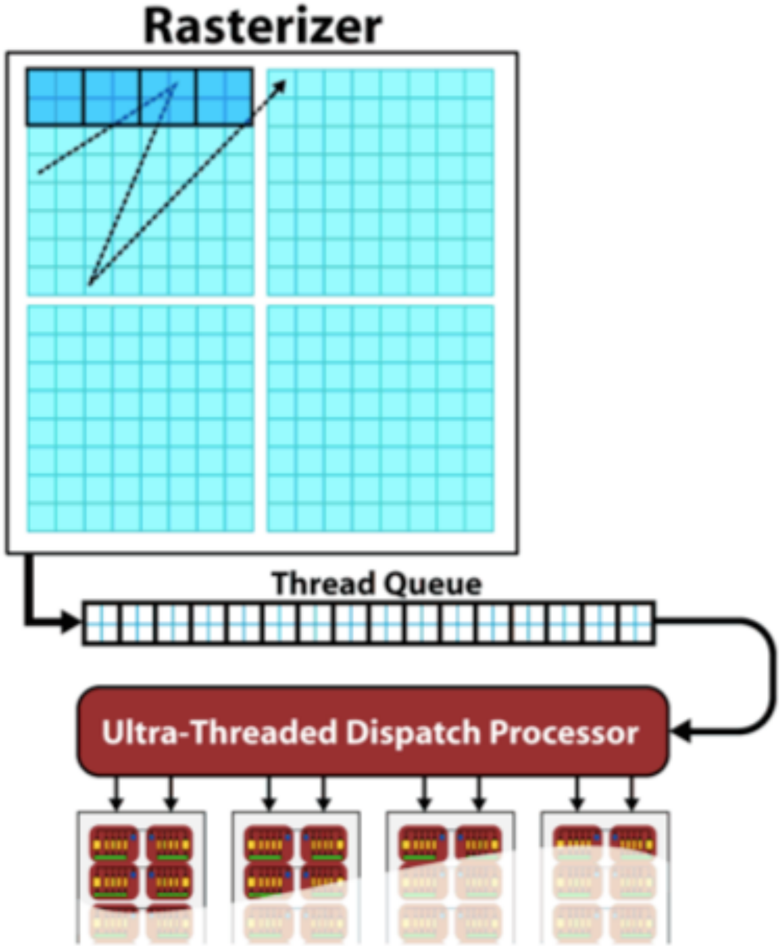
Source: http://developer.amd.com/gpu_assets/Stream_Computing_Overview.pdf

AMD Stream Processor

- Made up of thread processors
  - 5-way VLIW processor
  - One core can handle transcendental functions
    - Like NVIDIA's SFU
  - All thread processors in a SP execute same instruction each cycle
  - Up to four threads can issue four VLIW instructions over four cycles
    - To hide latency
  - Threads scheduled as wavefronts
    - Threads within wavefront subject to divergence during conditional execution

Wavefronts

- Size of wavefront differs on different stream processors
- Composed of quads



Source: http://developer.amd.com/gpu_assets/Stream_Computing_Overview.pdf

Wavefront Divergence

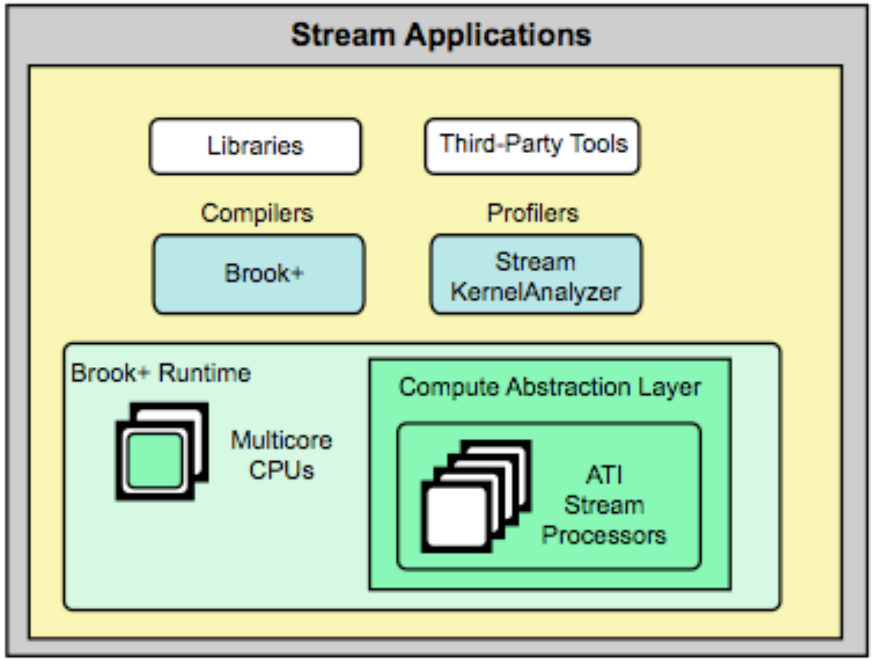
- If two branches take the same amount of time *t* to execute over a wavefront, then the total time if any thread diverges is …
  - 2*t*
- If a single loop iteration takes time *t*, and within a wavefront all threads execute the loop one time except one thread that executes the loop 100 times, then the total time is …
  - 100*t*

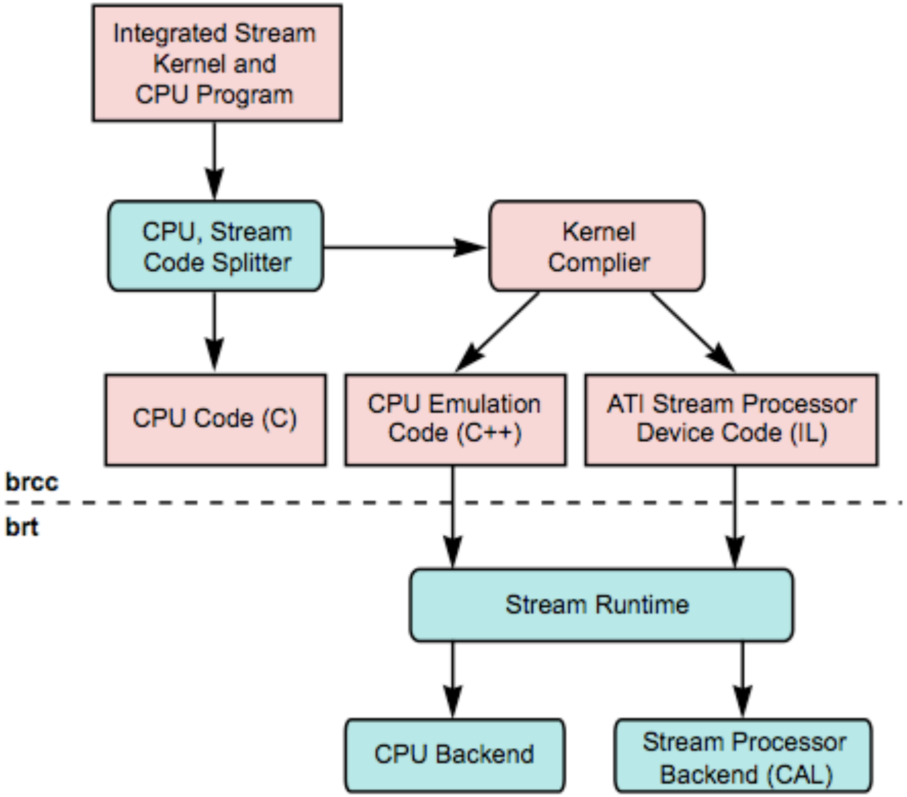
From http://developer.amd.com/gpu_assets/Stream_Computing_Overview.pdf

FireStream9250

- 800 SIMD superscalar processors
- Supports SSE-like vec4 operations
- IEEE single/double precision
- 1 TFLOP peak single precision
- 200 GFLOPS peak double-precision
- 1 GB GDDR3 on-board memory
- 108.8 GB/s Peak memory bandwidth

Stream SDK

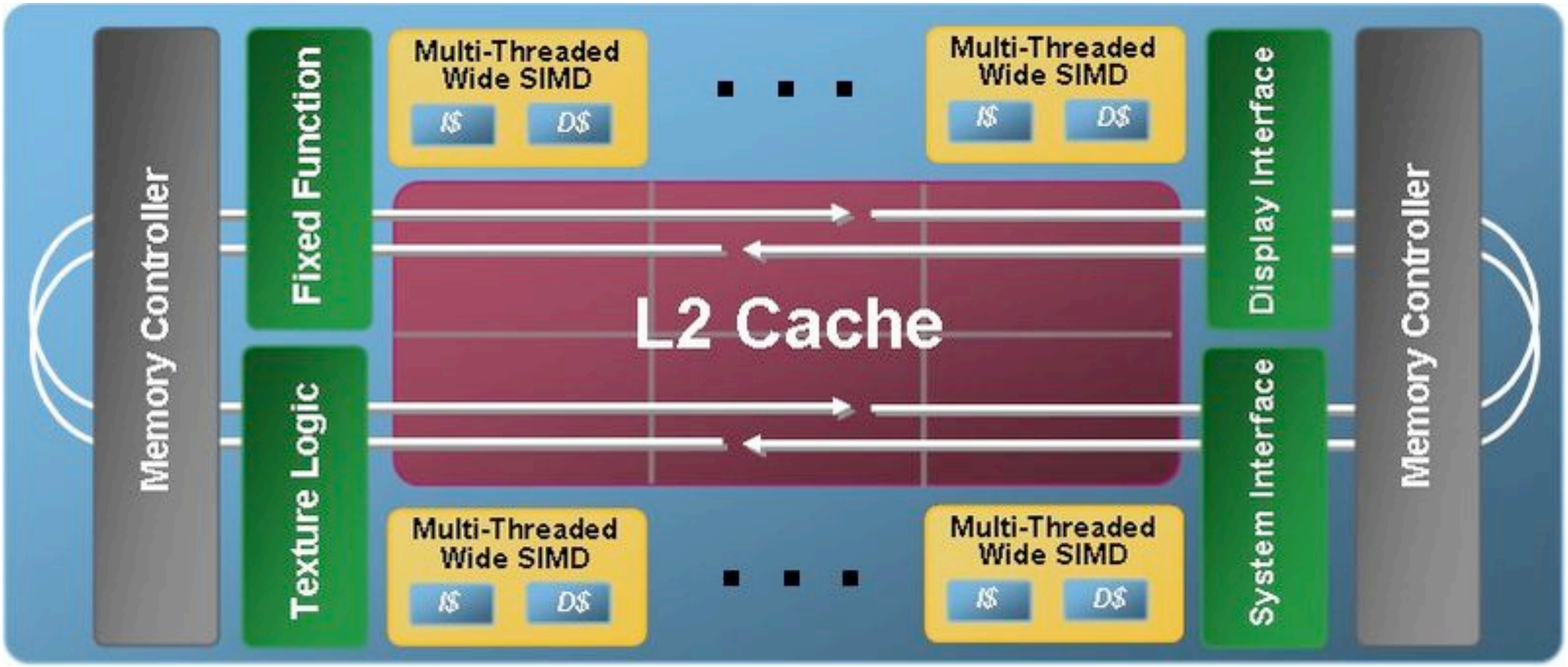



Integrated Stream Kernel and CPU Program
CPU, Stream Code Splitter
Kernel Complier
CPU Code (C)
CPU Emulation Code (C++)
ATI Stream Processor Device Code (IL)
brcc
brt
Stream Runtime
CPU Backend
Stream Processor Backend (CAL)

Source: http://developer.amd.com/gpu_assets/Stream_Computing_Overview.pdf

Intel Larrabee

- Hybrid multicore CPU/GPU
- First revealed in 2007
- Release as consumer graphics card cancelled in Dec. 2009

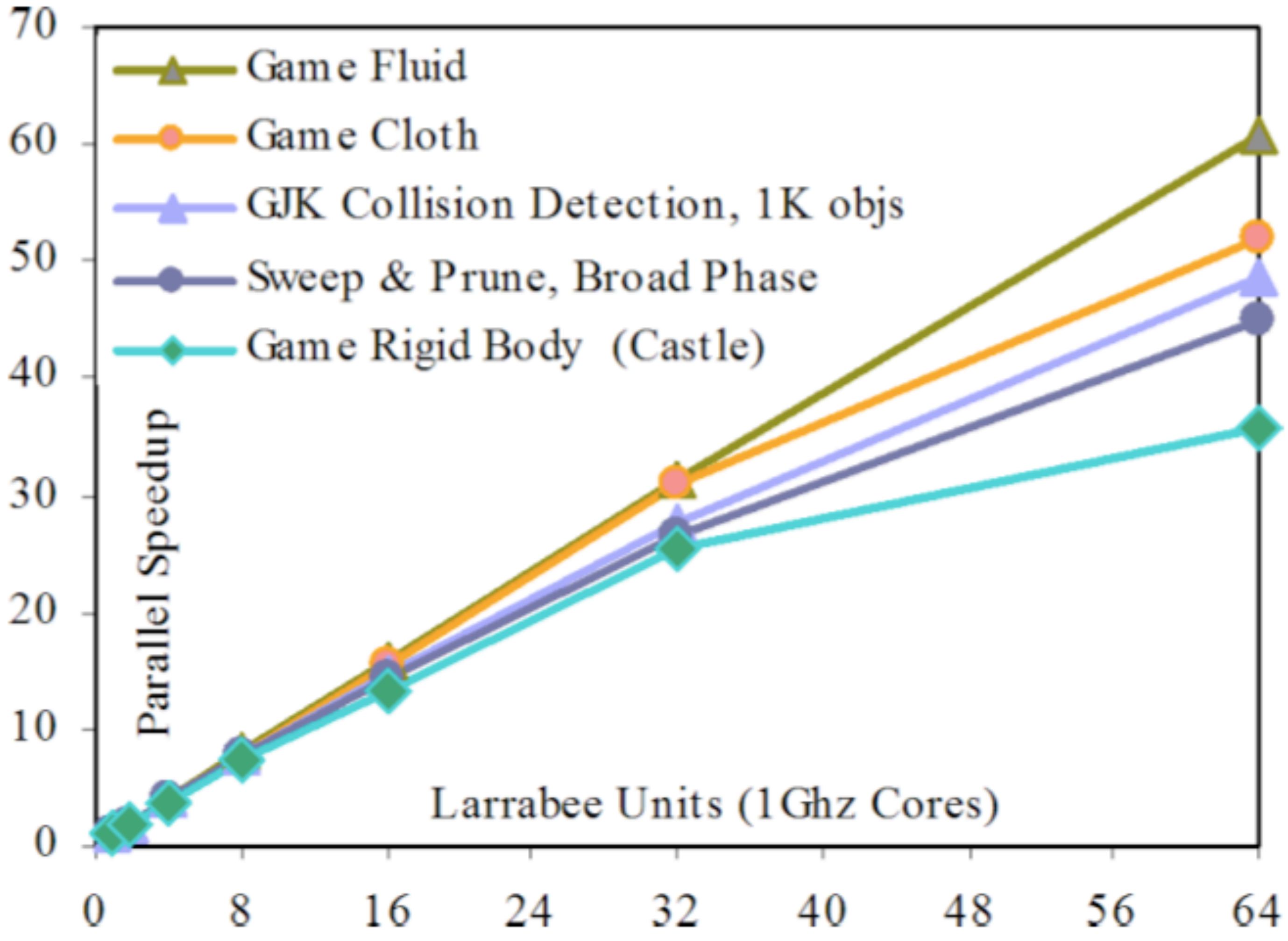


Source: Wikipedia

Intel Larrabee

- Enhanced x86 cores
  - Simplified: such as in-order execution only
  - Vector processing capability: 16 single-precision operations at a time
    - 4 times wider than most x86 processors
  - Initial release was to have 32 cores
- L2 cache with cache coherence
- Each core supports 4-way simultaneous multithreading
- Future?
  - May still be developed as accelerator card for HPC

Larrabee Performance



Source: E. Lindholdm et al., SIGGRAPH 2008

Source: http://www.hpcwire.com/features/Compilers-and-More-A-Computing-Larrabee-67032637.html

Which is better?

Larrabee	Nehalem	
32	4	cores
64 KB	64 KB	L1 cache/core
256 KB	256 KB	L2 cache/core
	2048 KB	L3 cache/core
16	4	Vector width (single prec.)
4	2	Multithreading width
2	4	Instruction issue width

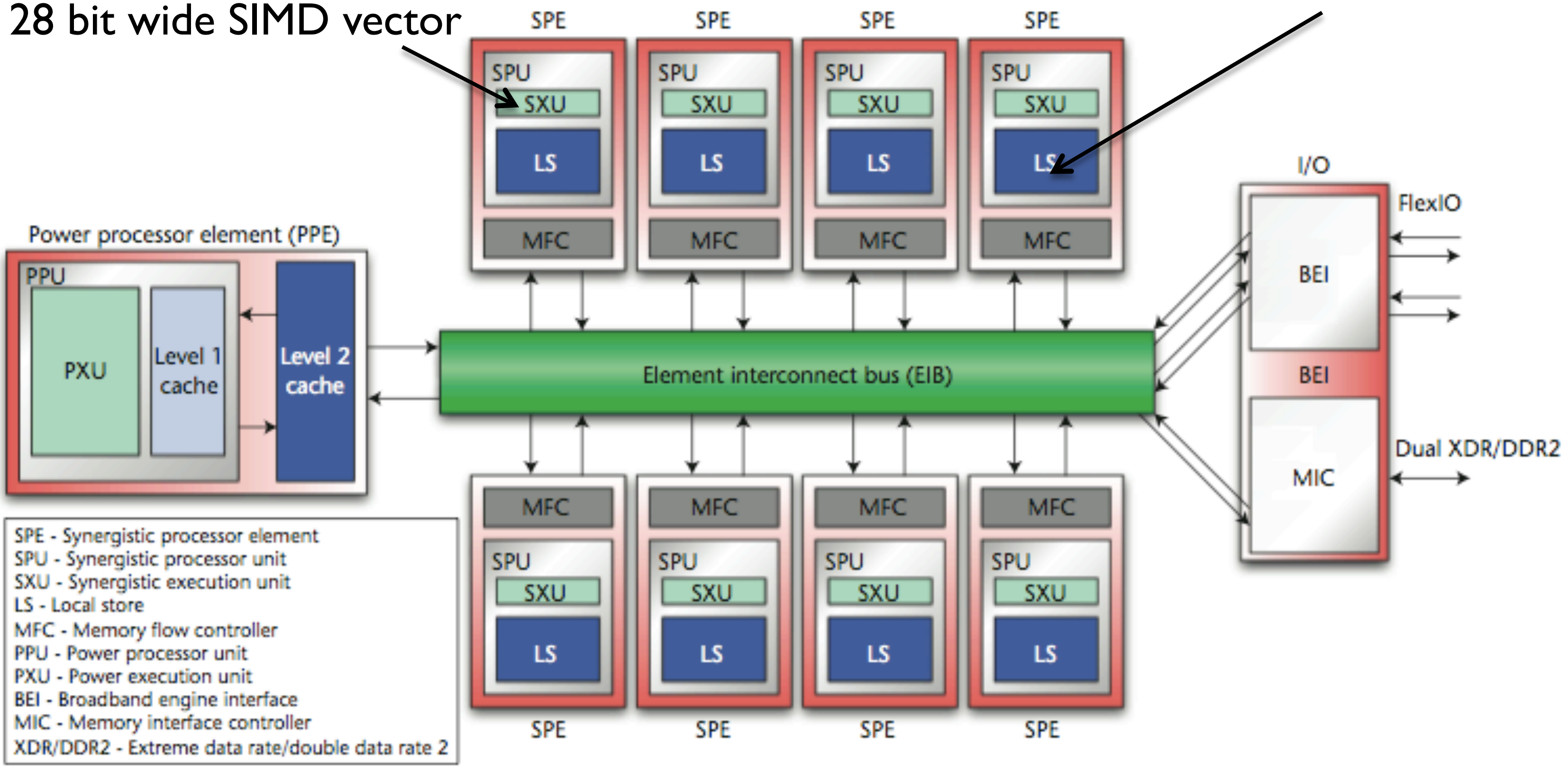
STI Cell Broadband Engine

- Jointly developed by Sony, Toshiba, and IBM (STI) in early 2000s
- First commercialized in Sony PlayStation 3
- As with GPUs, favors throughput (bandwidth) over latency
- Software developer's kit includes an SPE management for accessing and managing SPEs.
- Unique architecture has proved challenging to application developers
- Nov. 2009: further development cancelled

STI Cell Broadband Engine

local store is explicitly managed
by software (compiler or programmer)

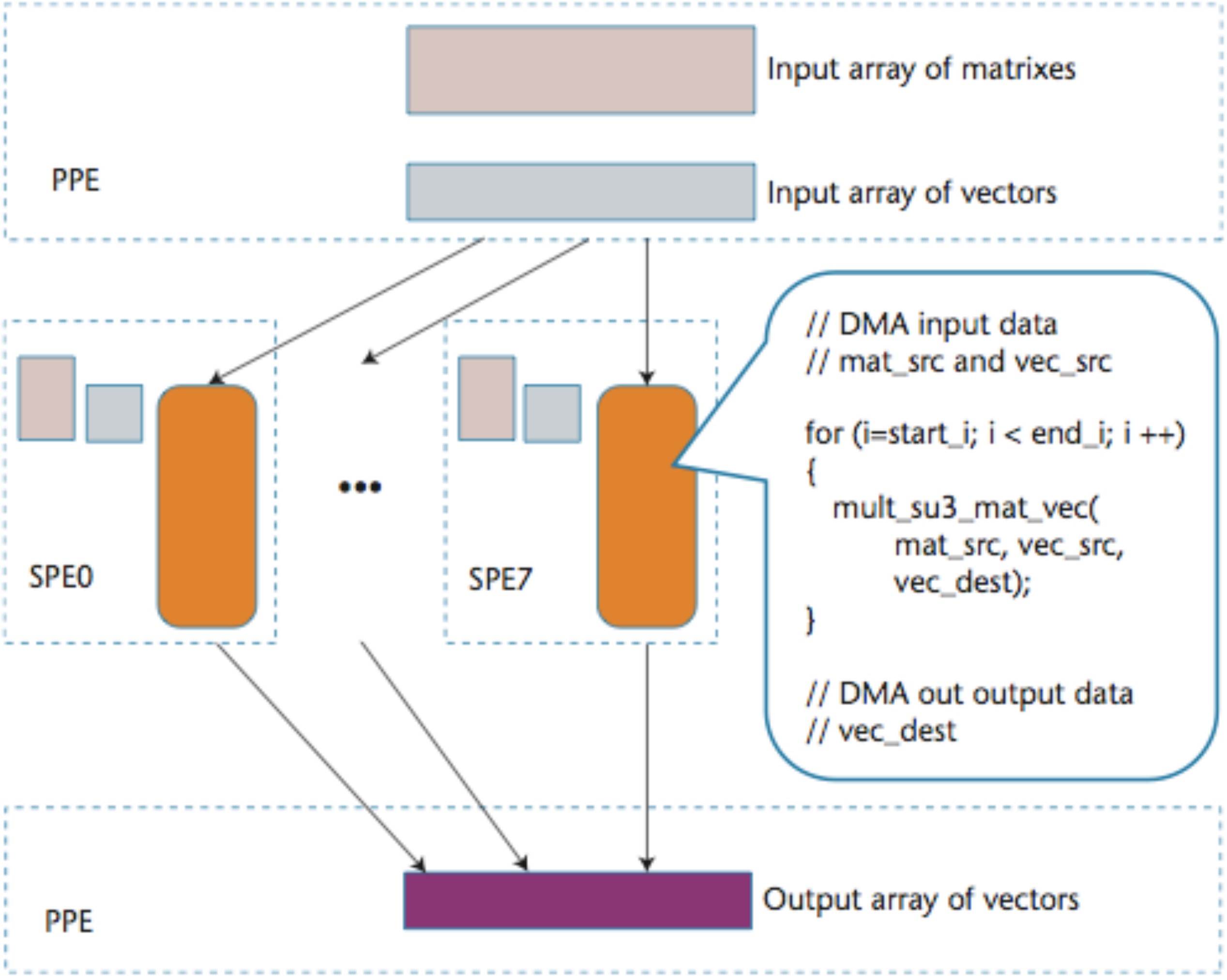
128 bit wide SIMD vector



SPE - Synergistic processor element
SPU - Synergistic processor unit
SXU - Synergistic execution unit
LS - Local store
MFC - Memory flow controller
PPU - Power processor unit
PXU - Power execution unit
BEI - Broadband engine interface
MIC - Memory interface controller
XDR/DDR2 - Extreme data rate/double data rate 2

Source: Computing in Science & Engineering, Jan./Feb. 2010

Cell Example Execution

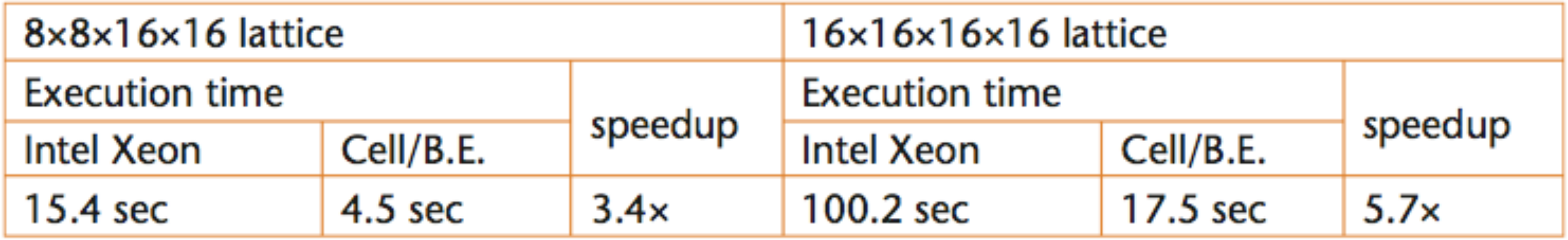


```
// DMA input data
// mat_src and vec_src

for (i=start_i; i < end_i; i ++)
{
  mult_su3_mat_vec(
        mat_src, vec_src,
        vec_dest);
}

// DMA out output data
// vec_dest
```

Source: Computing in Science & Engineering, Jan./Feb. 2010

STI Cell & HPC

- PowerXCell 8i: peak of about 12.8 GFLOPS/SPE, for total of 102.4 GFLOPS for eight SPEs.
- IBM Roadrunner supercomputer
  - 12240 PowerXCell 8i processors
  - 6562 AMD Opteron processors.
- Application example: Quantum Chromodynamics



8×8×16×16 lattice			16×16×16×16 lattice		
Execution time		speedup	Execution time		speedup
Intel Xeon	Cell/B.E.		Intel Xeon	Cell/B.E.	
15.4 sec	4.5 sec	3.4×	100.2 sec	17.5 sec	5.7×

Source: Computing in Science & Engineering, Jan./Feb. 2010