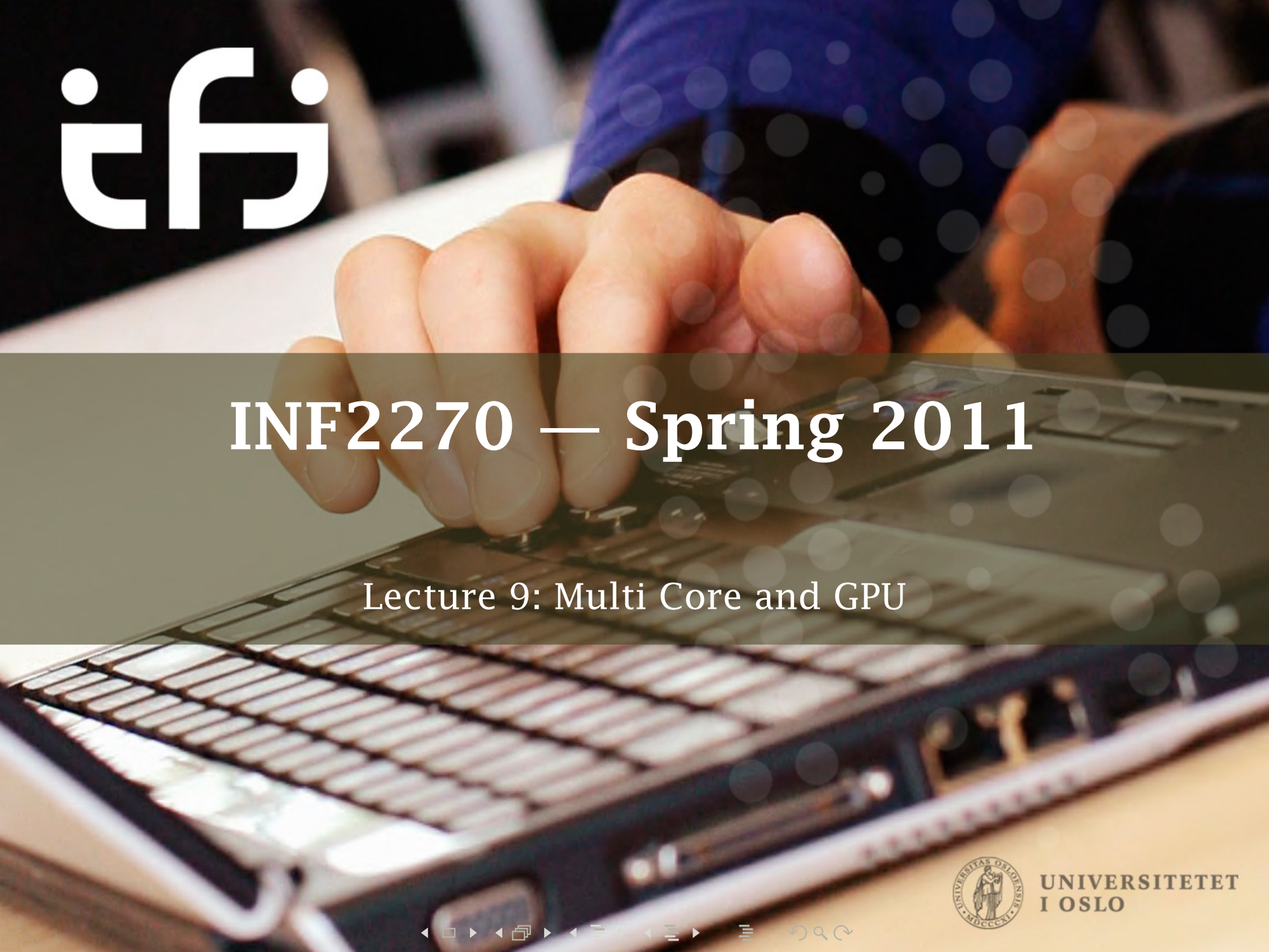


INF2270 — Spring 2011

Lecture 9: Multi Core and GPU

UNIVERSITETET I OSLO

content

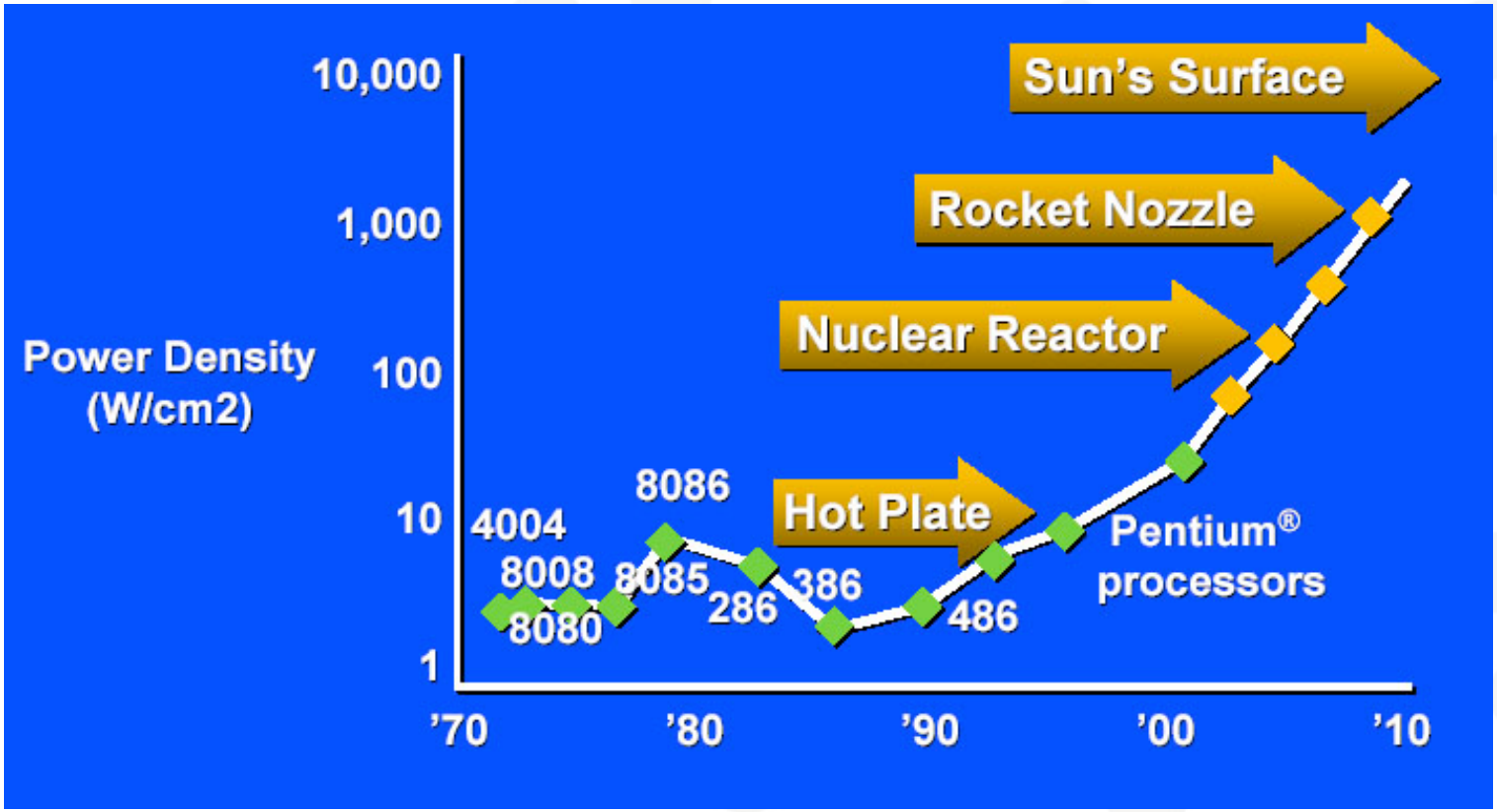
Multi Core Architecture

Graphics Processing Unit

Progress shift from speed to parallelism

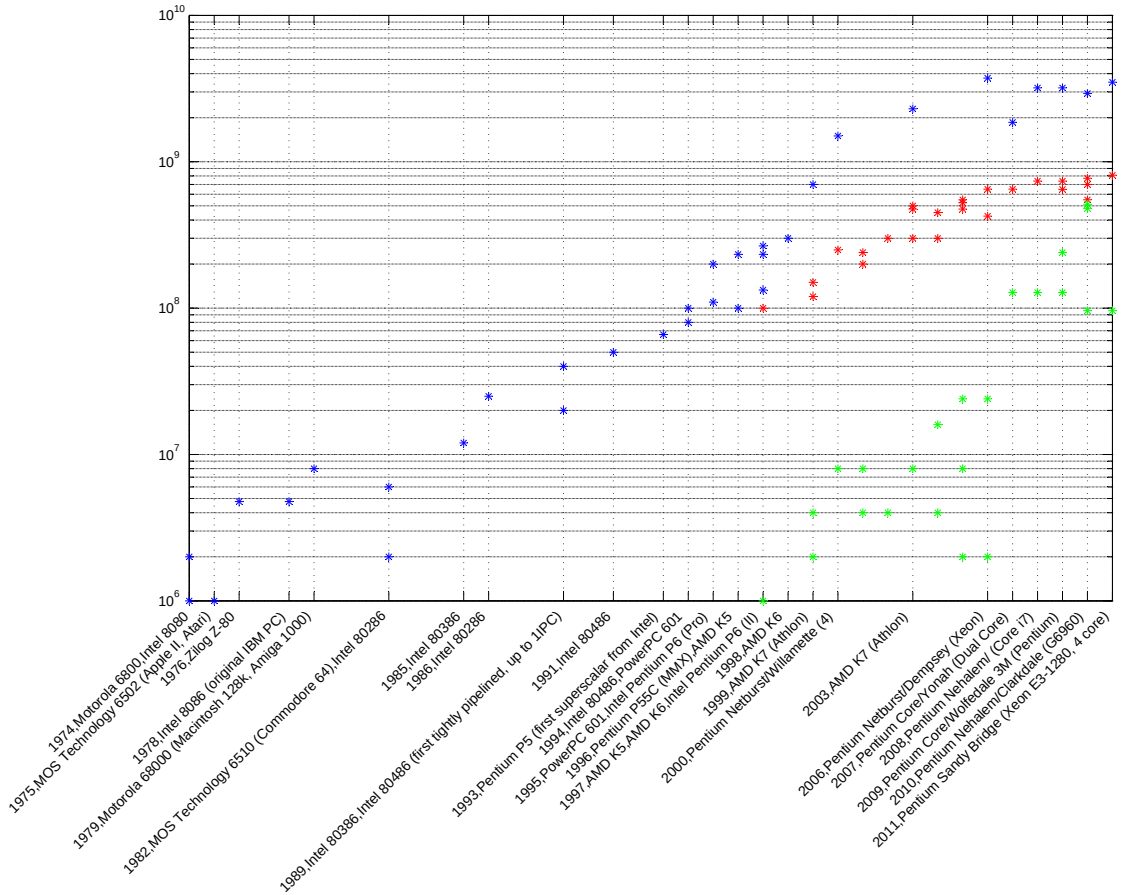
Wheras the clock frequency has been the measure of the progress in CPU design, this trend has now come to a stop. Heat disipation per surface area is the limiting factor. The new major quantitative sign of progress is the number of cores on a single chip. In contrast to earlier trends to parallelism, higher integration allows to place multiple cores on a single chip, vastly improving inter-core data communication speeds.

Surface Heat Dissipation



Source: Pat Gelsinger, ISSCC 2001

History of Clockrate

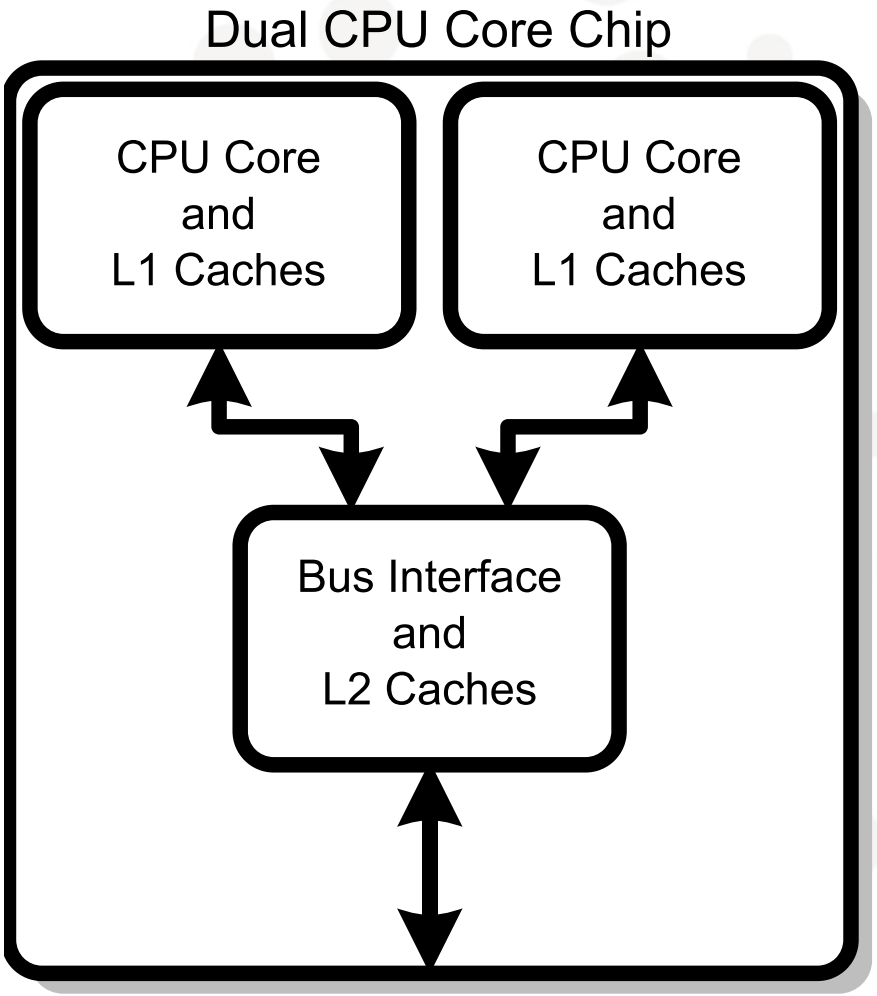


content

Multi Core Architecture

Graphics Processing Unit

Multi Core CPU



Multi Core and Parallel Applications

From a software point of view, the OS starts on core 0 and sends interrupts with the address of the starting instructions to other cores. Inter core communication and synchronization is handled purely by the OS, i.e. it's neigh impossible for an application programmer to exploit parallelism.

content

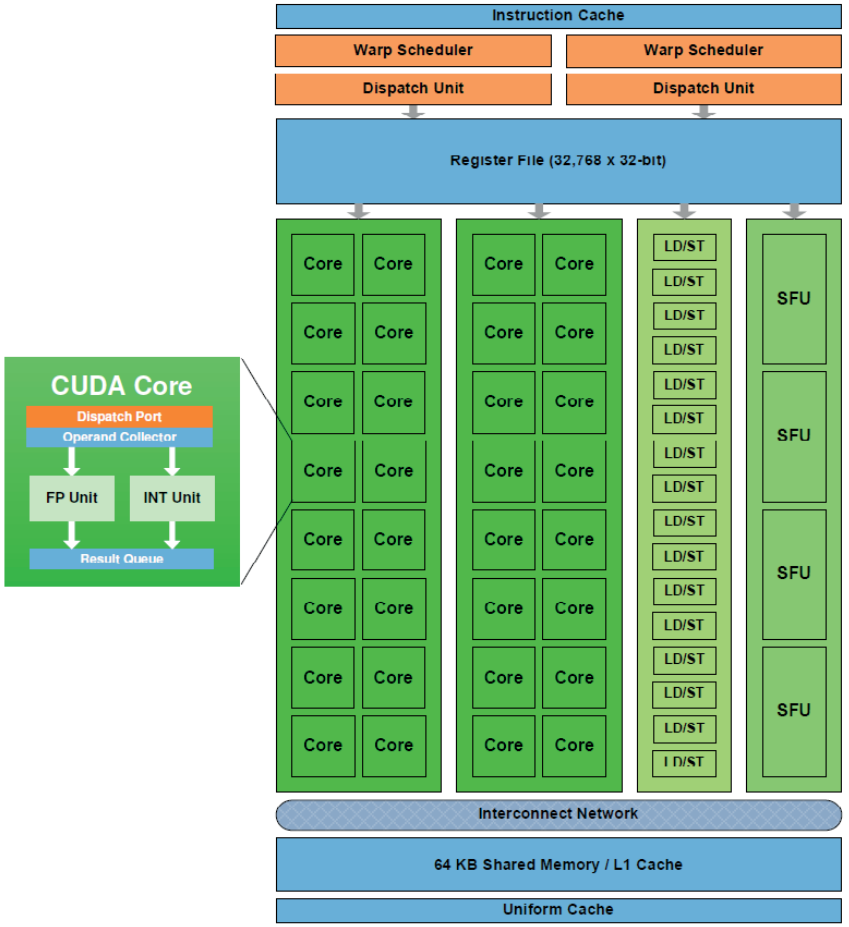
Multi Core Architecture

Graphics Processing Unit

Graphics Processing Unit

If indeed the major sign of progress is the number of cores on a single chip, CPUs are no longer the driving force of this progress but GPUs are.

NVIDIA Fermi Streaming Multiprocessor

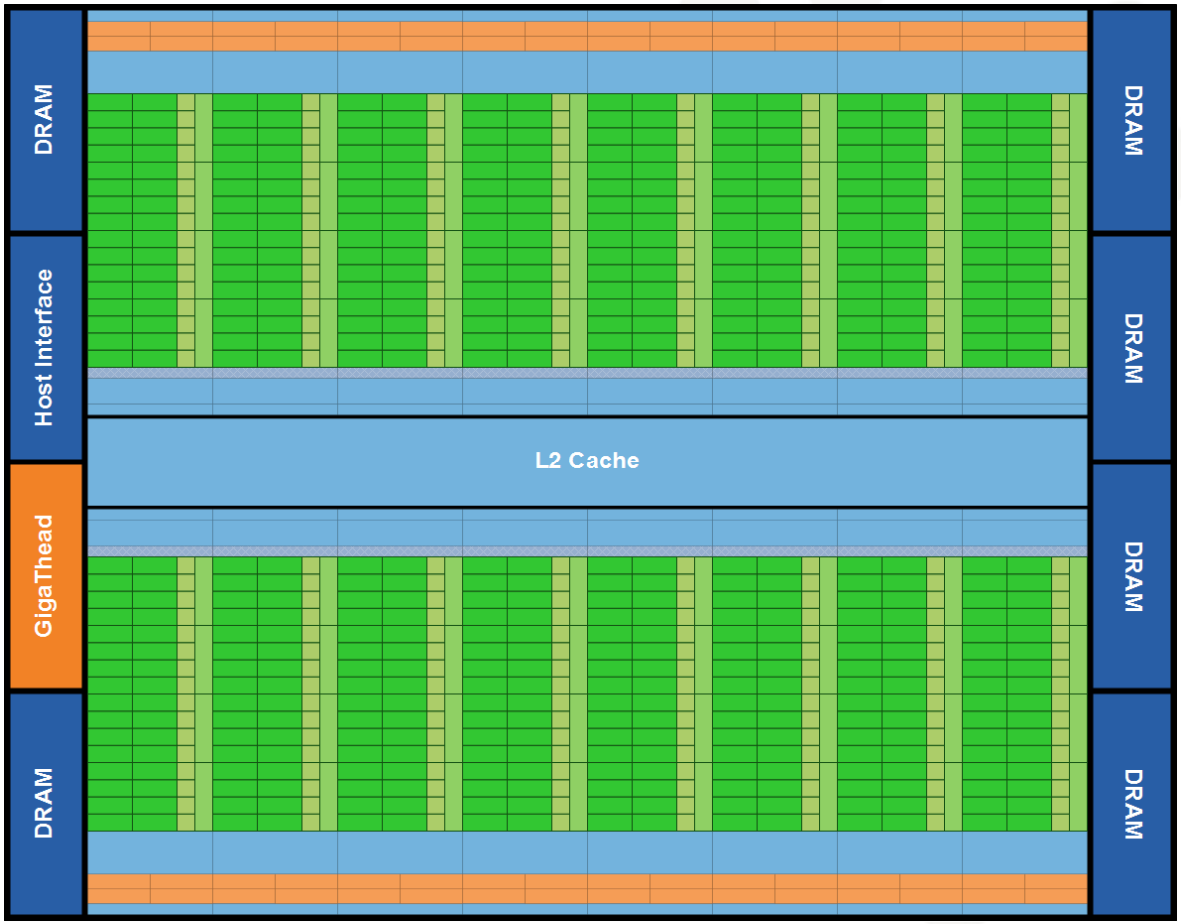


Fermi Streaming Multiprocessor (SM)

Source:

Whitepaper: NVIDIA's Next Generation CUDA Compute architecture FERMI^{tm}

NVIDIA Fermi GPU Block Diagram



Source:

Whitepaper: NVIDIA's Next Generation CUDA Compute architecture FERMI^{tm}

Maximum GOps

Estimated (!)

NVIDIA GF100			Intel GP6960 (2core)		
inst	INT32	FP64	inst	INT32	FP64
FMA	395	197	ADD/MUL	17.58	11.72
			SSE	70.32	35.16

Reduced Yield in CPU/GPU production

Number of transistors:

NVIDIA GF100	Intel Westermere-EX (Xeon E7 10 core)
$3e9$	$2.6e9$

Increasing single chip complexities and sheer numbers of transistors increase the problem of limited yield in ASIC production: even extremely small probabilities of single transistor failure result in considerable probability of chip failure. Thus, nVIDIA actually markets the GF100 also as products with only 15 or 14SMs. (Easier for them than for Intel if one out of 10 cores fails!)

GPU and Parallel Applications

Recently, GPUs are used for numerous non-graphical parallelized algorithms (General Purpose GPU (GPGPU)). NVIDIA is covering that niche and provides a C-precompiler (Compute Unified Device Architecture, CUDA) for easy application development. See INF3380, INF5063.