Innovations in the Fermi GPU Architecture :

- Fermi introduces several innovations to bring GPUs much closer to mainstream system processors than Tesla and previous generations of GPU architectures:
- Fast Double-Precision Floating-Point Arithmetic-Fermi matches the relative double-precision speed of conventional processors of roughly half the speed of single precision versus a tenth the speed of single precision in the prior Tesla generation. The peak double-precision performance grew from 78 GFLOP/sec in the predecessor GPU to 515 FLOP/sec when using multiply-add instructions.
- Caches for GPU Memory- . Fermi includes both an L1 Data Cache and LI Instruction Cache for each multithreaded SIMD Processor and a single 768 KB L2 cache shared by all multithreaded SIMD Processors in the GPU. As mentioned above, in addition to reducing bandwidth pressure on GPU Memory, caches can save energy by staying on-chip rather than going off-chip to DRAM. The L1 cache actually cohabits the same SRAM as Local Memory. Fermi has a mode bit that offers the choice of using 64 KB of SRAM as a 16 KB LI cache with 48 KB of Local Memory or as a 48 KB L1 cache with 16 KB of Local Memory.
- 64-Bit Addressing and a Unified Address Space for All GPU Memories- This innovation makes it much easier to provide the pointers needed for C and C++.

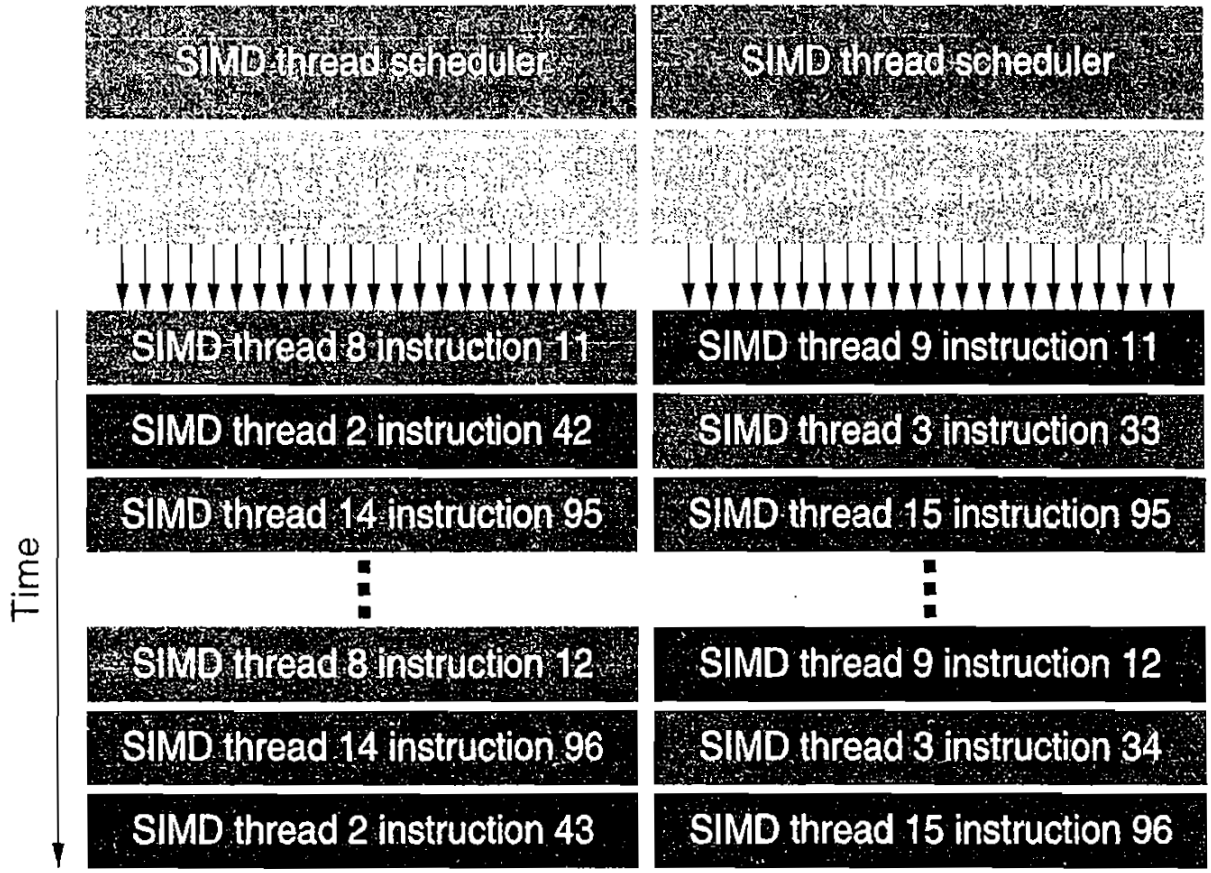
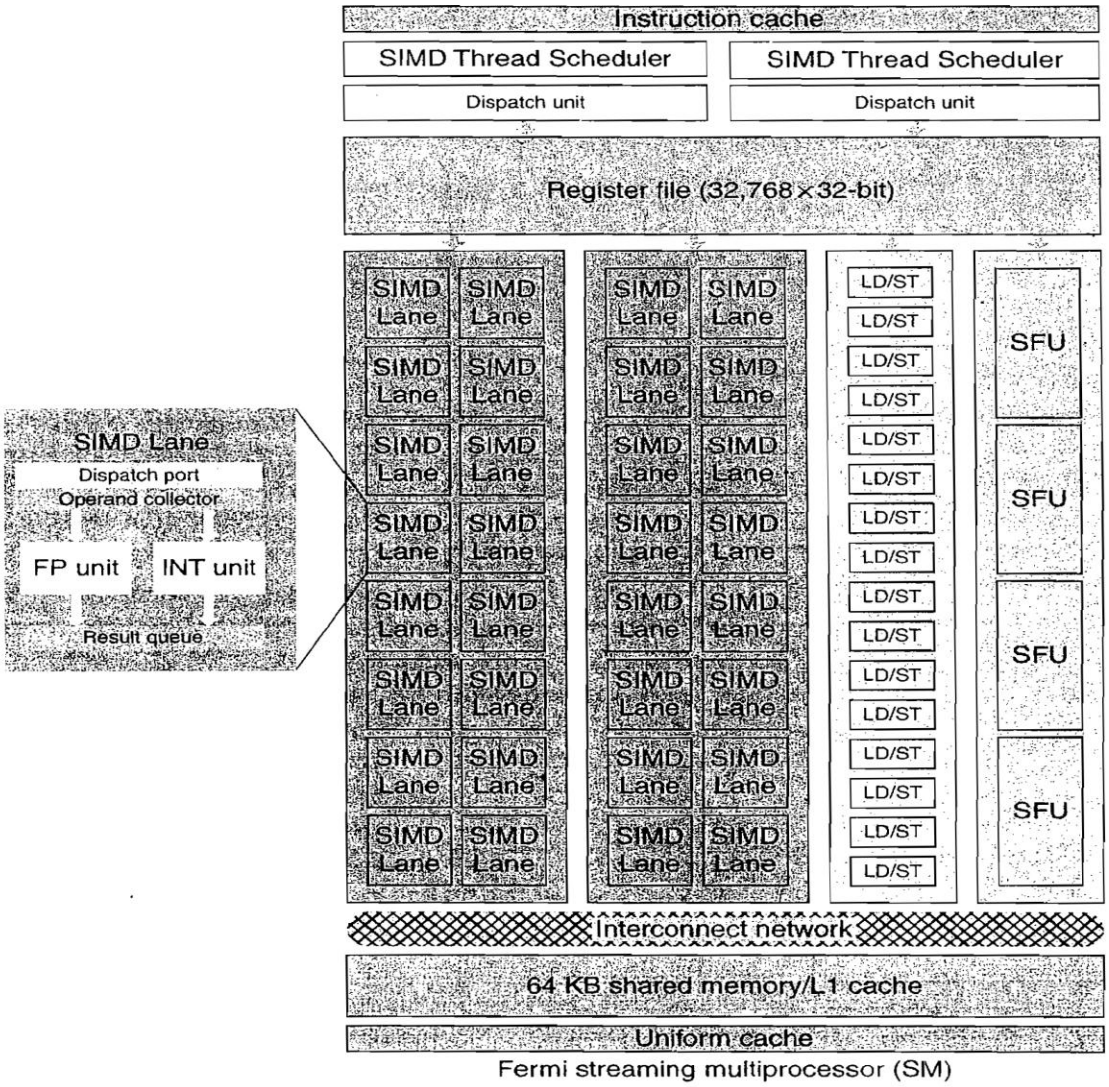


Fig: Fermi's Dual SIMD Thread Scheduler.

- Error Correcting Codes to detect and correct errors in memory and registers (see Chapter 2)-To make long-running applications dependable on thousands of servers, ECC is the norm in the datacenter.
- Faster Context Switching-Given the large state of a multithreaded SIMD Processor, Fermi has hardware support to switch contexts much more quickly. Fermi can switch in less than 25 microseconds, about I Ox faster than its predecessor can.
- Faster Atomic Instructions-First included in the Tesla architecture, Fermi improves performance of Atomic instructions by 5 to 20x, to a few microseconds. A special hardware unit associated with the L2 cache, not inside the multithreaded SIMD Processors, handles atomic instructions.



Fermi streaming multiprocessor (SM)

Fig: multithreaded SIMD Processor of a Fermi GPU

- Each SIMD Lane has a pipelined floating-point unit, a pipelined integer unit, some logic for dispatching instructions and operands to these units, and a queue for holding results. The four Special Function units (SFUs) calculate functions such as square roots, reciprocals, sines, and cosines.