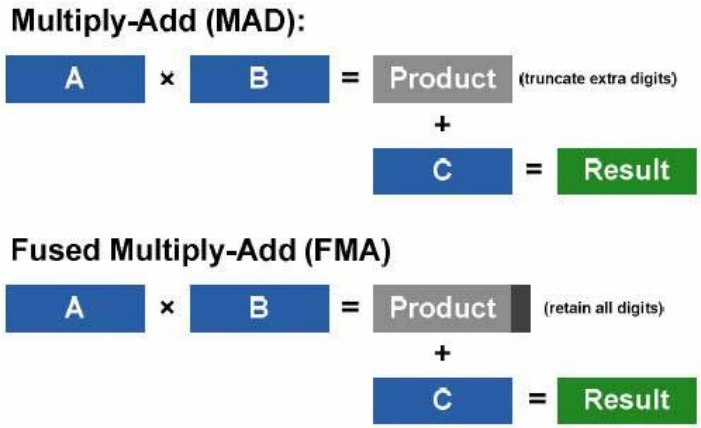


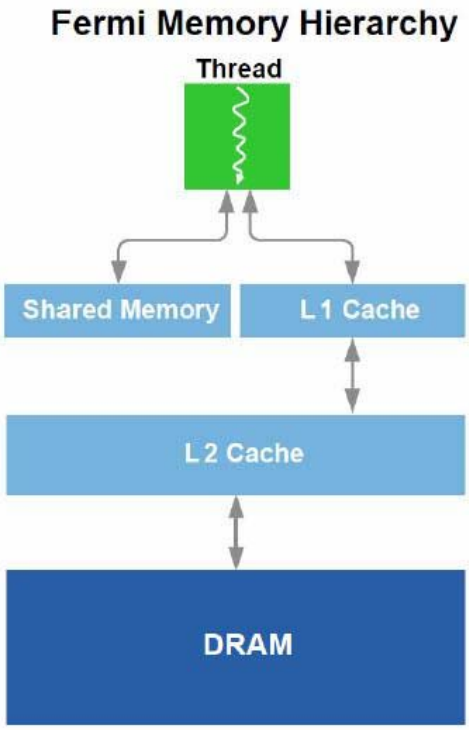
Improved Conditional Performance through Predication

In the Fermi ISA, the native hardware predication support used for divergent thread management is now available at the instruction level. Predication enables short conditional code segments to execute efficiently with no branch instruction overhead.

Memory Subsystem Innovations

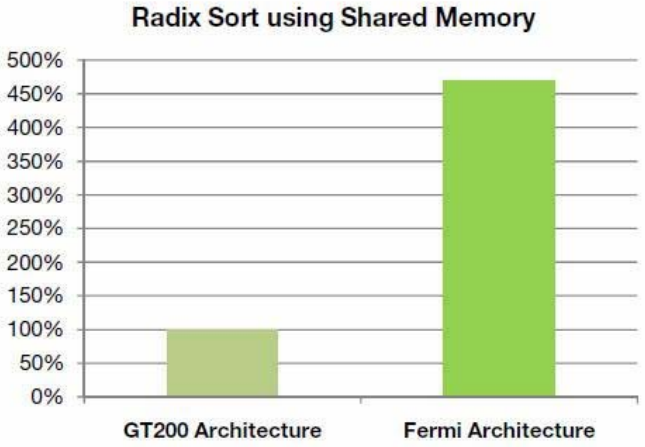
NVIDIA Parallel DataCache™ with Configurable L1 and Unified L2 Cache

Working with hundreds of GPU computing applications from various industries, we learned that while Shared memory benefits many problems, it is not appropriate for all problems. Some algorithms map naturally to Shared memory, others require a cache, while others require a combination of both. The optimal memory hierarchy should offer the benefits of both Shared memory and cache, and allow the programmer a choice over its partitioning. The Fermi memory hierarchy adapts to both types of program behavior.

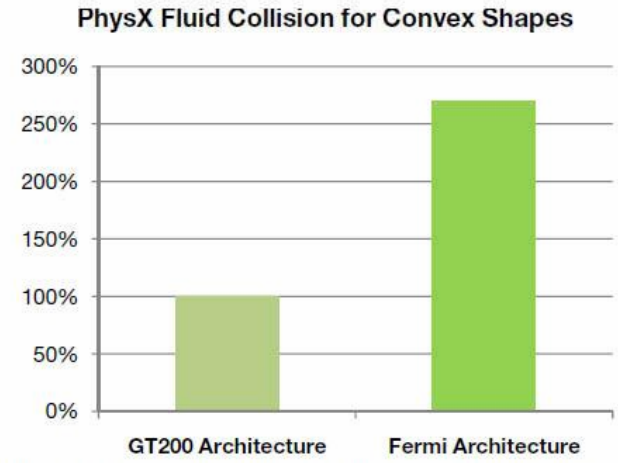


Adding a true cache hierarchy for load / store operations presented significant challenges. Traditional GPU architectures support a read-only "load" path for texture operations and a write-only "export" path for pixel data output. However, this approach is poorly suited to executing general purpose C or C++ thread programs that expect reads and writes to be ordered. As one example: spilling a register operand to memory and then reading it back creates a read after write hazard; if the read and write paths are separate, it may be necessary to explicitly flush the entire write / "export" path before it is safe to issue the read, and any caches on the read path would not be coherent with respect to the write data.

The Fermi architecture addresses this challenge by implementing a single unified memory request path for loads and stores, with an L1 cache per SM multiprocessor and unified L2 cache that services all operations (load, store and texture). The per-SM L1 cache is configurable to support both shared memory and caching of local and global memory operations. The 64 KB memory can be configured as either 48 KB of Shared memory with 16 KB of L1 cache, or 16 KB of Shared memory with 48 KB of L1 cache. When configured with 48 KB of shared memory, programs that make extensive use of shared memory (such as electrodynamic simulations) can perform up to three times faster. For programs whose memory accesses are not known beforehand, the 48 KB L1 cache configuration offers greatly improved performance over direct access to DRAM.



When using 48 KB of shared memory on Fermi, Radix Sort executes 4.7x faster than GT200.



Physics algorithms such as fluid simulations especially benefit from Fermi's caches. For convex shape collisions, Fermi is 2.7x faster than GT200.

In either configuration, the L1 cache also helps by caching temporary register spills of complex programs. Prior generation GPUs spilled registers directly to DRAM, increasing access latency. With the L1 cache, performance scales gracefully with increased temporary register usage.

Fermi features a 768 KB unified L2 cache that services all load, store, and texture requests. The L2 provides efficient, high speed data sharing across the GPU. Algorithms for which data addresses are not known beforehand, such as physics solvers, raytracing, and sparse matrix multiplication especially benefit from the cache hierarchy. Filter and convolution kernels that require multiple SMs to read the same data also benefit.

First GPU with ECC Memory Support

Fermi is the first GPU to support Error Correcting Code (ECC) based protection of data in memory. ECC was requested by GPU computing users to enhance data integrity in high performance computing environments. ECC is a highly desired feature in areas such as medical imaging and large-scale cluster computing.

Naturally occurring radiation can cause a bit stored in memory to be altered, resulting in a soft error. ECC technology detects and corrects single-bit soft errors before they affect the system. Because the probability of such radiation induced errors increase linearly with the number of installed systems, ECC is an essential requirement in large cluster installations.

Fermi supports Single-Error Correct Double-Error Detect (SECDED) ECC codes that correct any single bit error in hardware as the data is accessed. In addition, SECDED ECC ensures that all double bit errors and many multi-bit errors are also be detected and reported so that the program can be re-run rather than being allowed to continue executing with bad data.

Fermi's register files, shared memories, L1 caches, L2 cache, and DRAM memory are ECC protected, making it not only the most powerful GPU for HPC applications, but also the most reliable. In addition, Fermi supports industry standards for checking of data during transmission from chip to chip. All NVIDIA GPUs include support for the PCI Express standard for CRC check with retry at the data link layer. Fermi also supports the similar GDDR5 standard for CRC check with retry (aka "EDC") during transmission of data across the memory bus.

Fast Atomic Memory Operations

Atomic memory operations are important in parallel programming, allowing concurrent threads to correctly perform read-modify-write operations on shared data structures. Atomic operations such as add, min, max, and compare-and-swap are atomic in the sense that the read, modify, and write operations are performed without interruption by other threads. Atomic memory operations are widely used for parallel sorting, reduction operations, and building data structures in parallel without locks that serialize thread execution.

Thanks to a combination of more atomic units in hardware and the addition of the L2 cache, atomic operations performance is up to 20× faster in Fermi compared to the GT200 generation.