

| GPU  | G80               | GT200               | Fermi                       |
|--|-------------------|---------------------|-----------------------------|
| Transistors                                | 681 million       | 1.4 billion         | 3.0 billion                 |
| CUDA Cores                                 | 128               | 240                 | 512                         |
| Double Precision Floating Point Capability | None              | 30 FMA ops / clock  | 256 FMA ops /clock          |
| Single Precision Floating Point Capability | 128 MAD ops/clock | 240 MAD ops / clock | 512 FMA ops /clock          |
| Special Function Units (SFUs) / SM         | 2                 | 2                   | 4                           |
| Warp schedulers (per SM)                   | 1                 | 1                   | 2                           |
| Shared Memory (per SM)                     | 16 KB             | 16 KB               | Configurable 48 KB or 16 KB |
| L1 Cache (per SM)                          | None              | None                | Configurable 16 KB or 48 KB |
| L2 Cache                                   | None              | None                | 768 KB                      |
| ECC Memory Support                         | No                | No                  | Yes                         |
| Concurrent Kernels                         | No                | No                  | Up to 16                    |
| Load/Store Address Width                   | 32-bit            | 32-bit              | 64-bit                      |