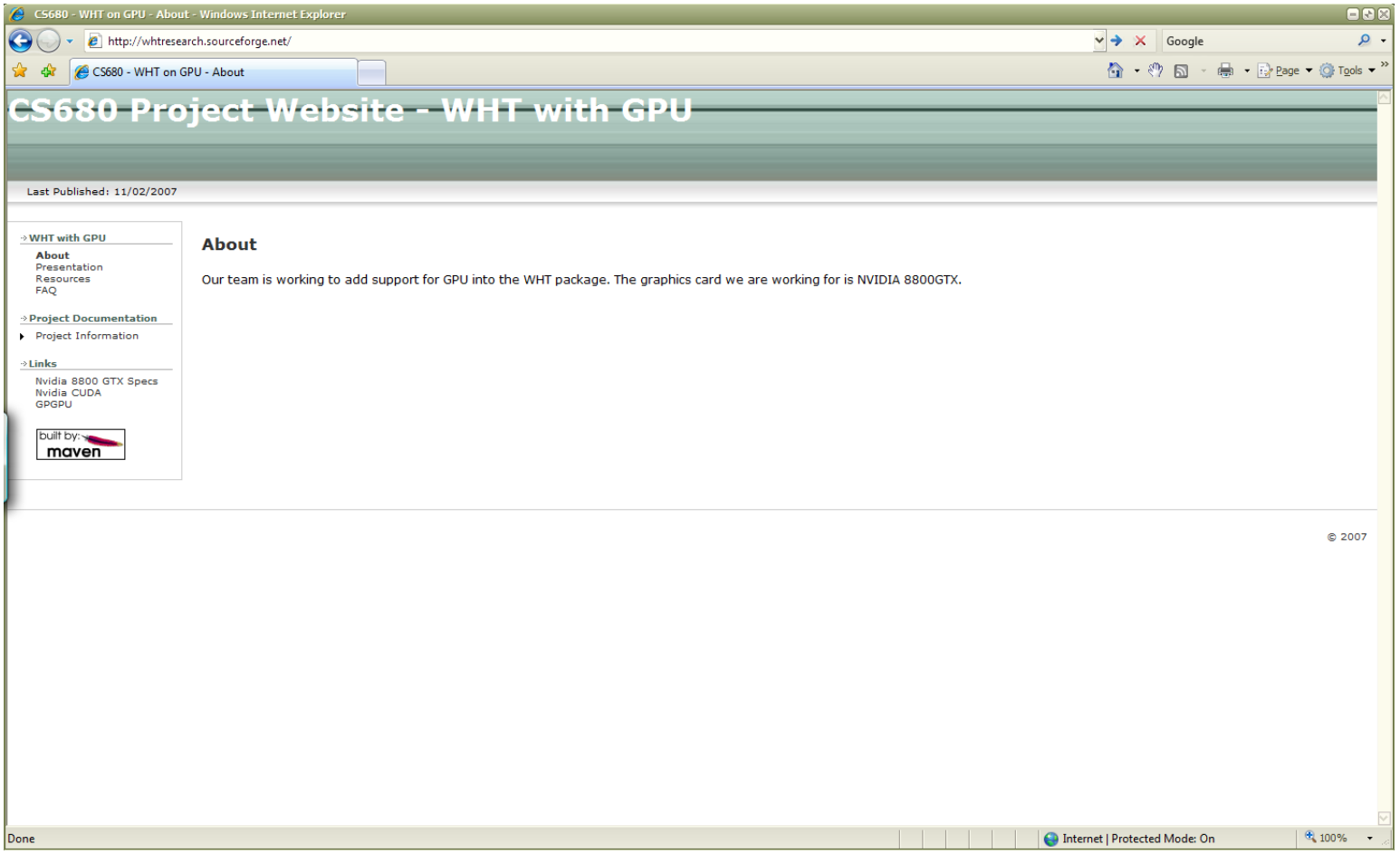
Project Website

- http://whtresearch.sourceforge.net



CPU and GPU

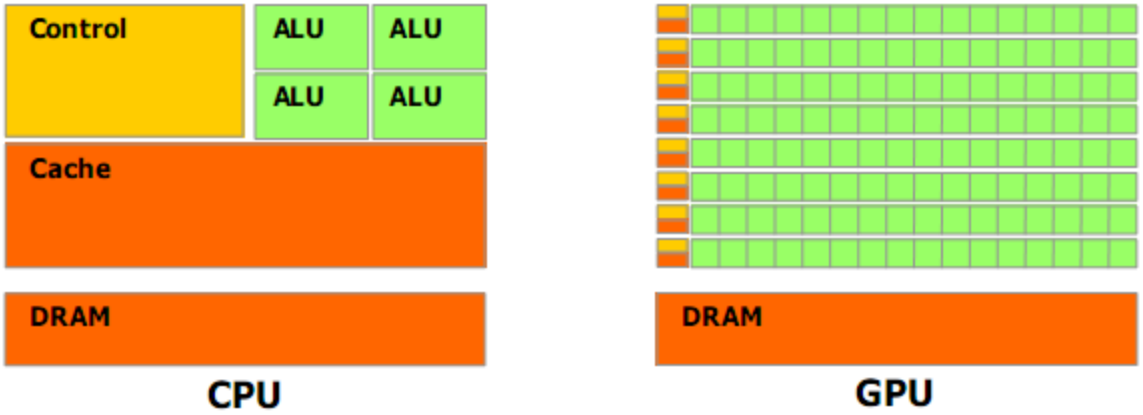
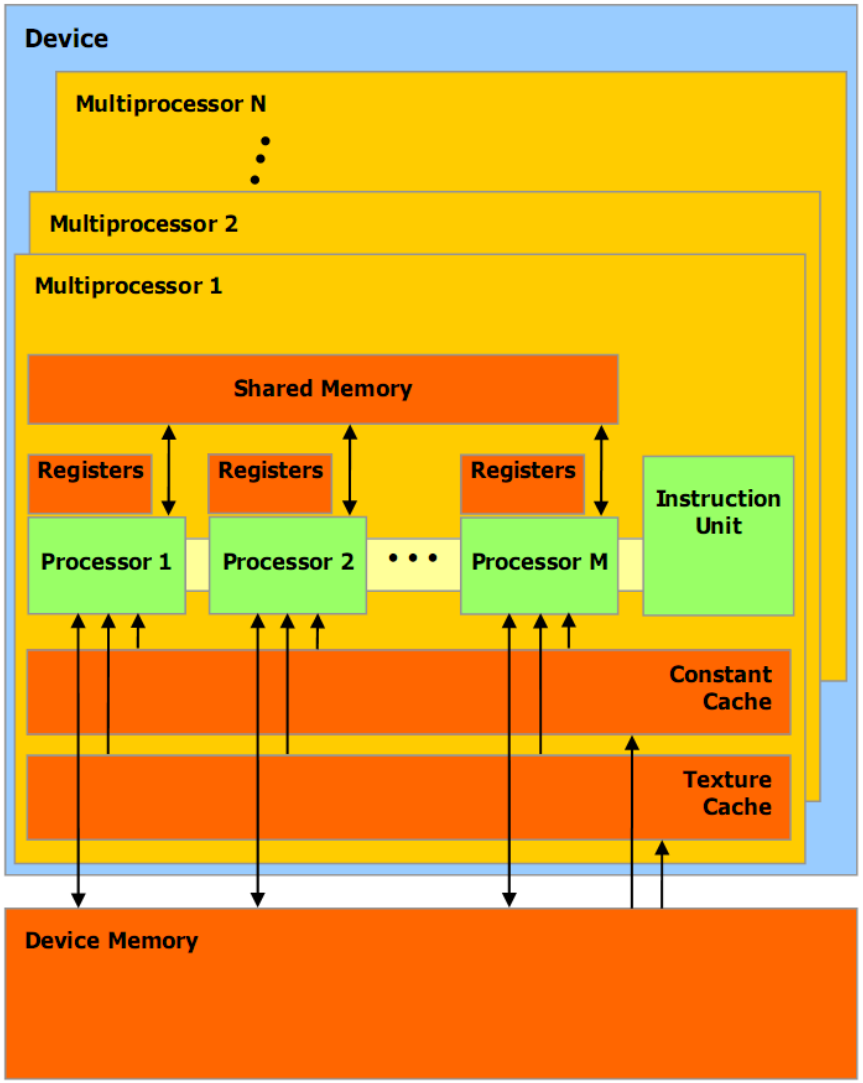


Figure 1-2. The GPU Devotes More Transistors to Data Processing

References: CUDA Programming Guide By Nvidia. http://developer.nvidia.com/object/cuda.html

Nvidia 8800 GTX Hardware Architecture

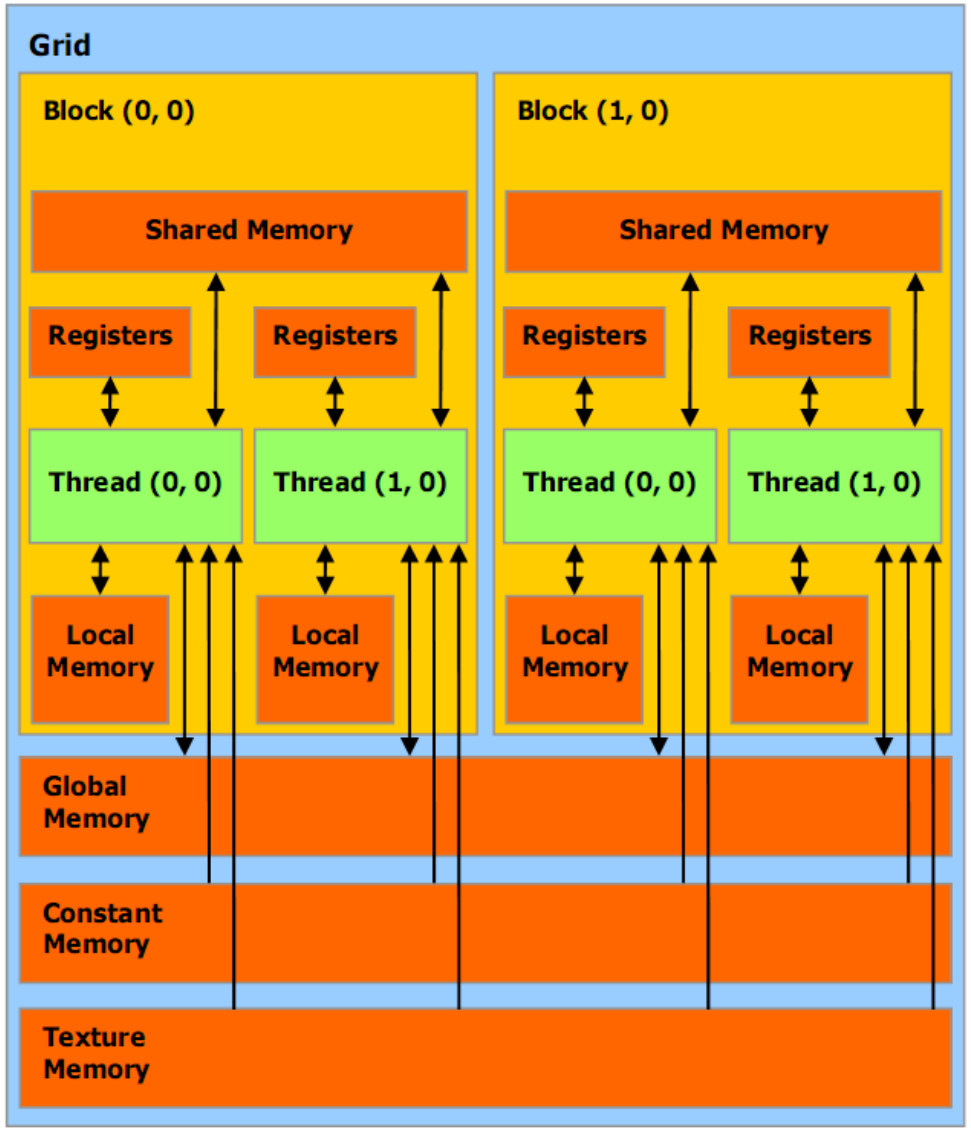
- SIMD Multiprocessor Architecture
  - At any given clock cycle, each processor of the multiprocessor executes the **same instruction** but operates on **different data.**



References: CUDA Programming Guide By Nvidia. http://developer.nvidia.com/object/cuda.html

On-chip memory

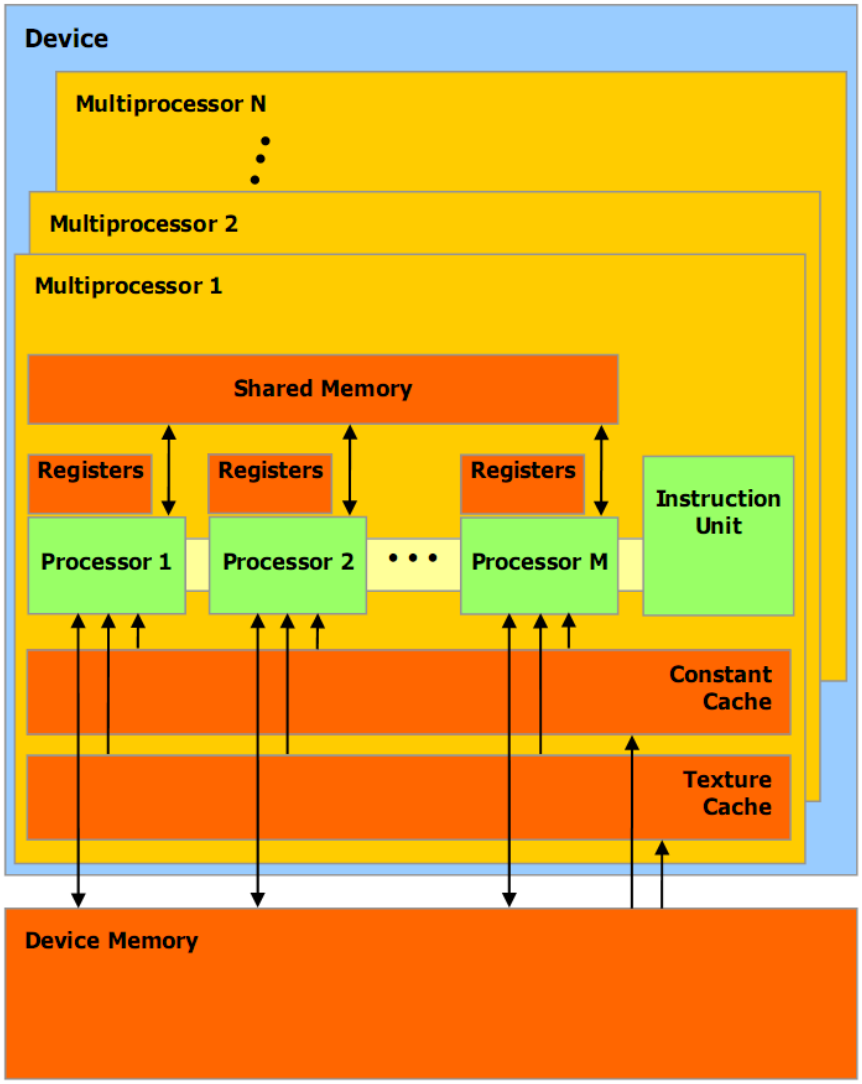
- **Each Multiprocessor has on-chip memory of four types:**
  - a set of local 32-bit registers per processor
  - a parallel data cache or shared memory that is shared by all the processors within a multiprocessor
  - read only constant cache and texture cache
- **Execution Model**
  - a grid of blocks is executed by executing one or more blocks on each multiprocessor Each block is split into groups call warps, each warp is executed on one multiprocessor.



References: CUDA Programming Guide By Nvidia. http://developer.nvidia.com/object/cuda.html

Technical Specs

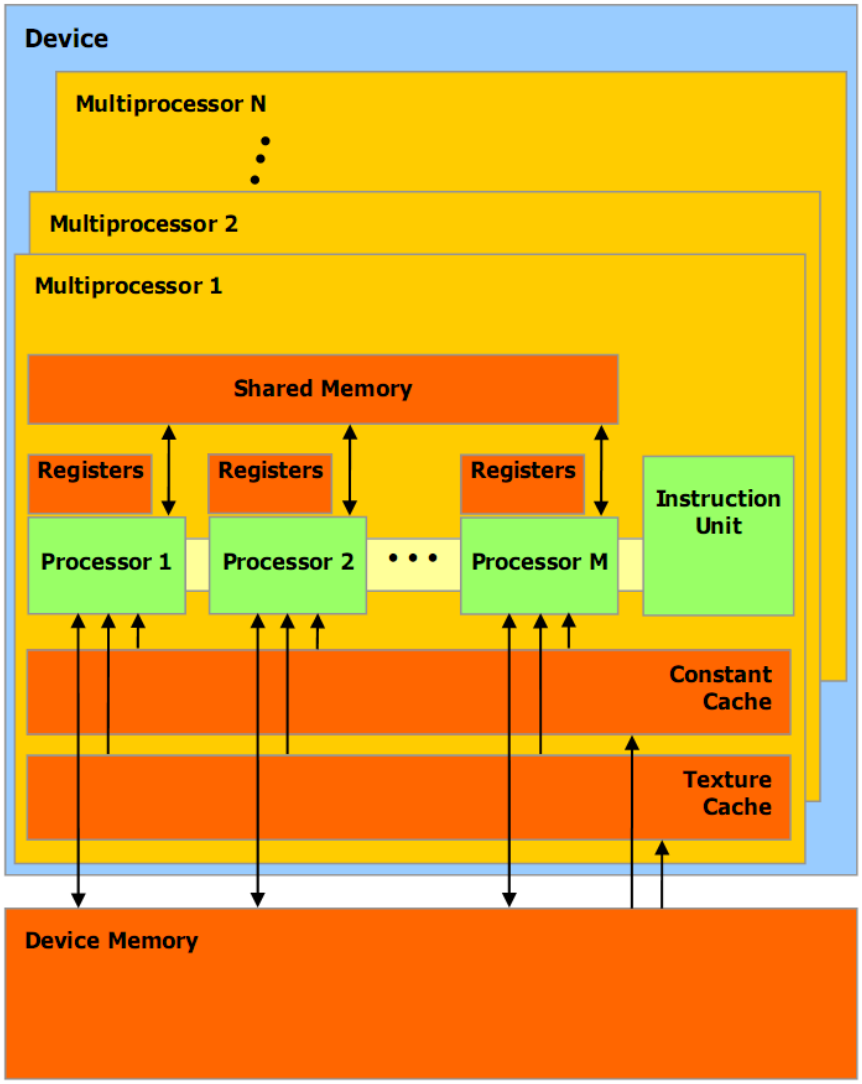
- Number of Multiprocessors: **16**
- Max number of thread per block: **512**
- Warp Size: **32 (32 processors)**
- number of registers per multiprocessor: **8192**
- Amount of shared memory available per multiprocessor: **16KB organized into 16 banks (Blocksize 1KB).**



References: CUDA Programming Guide By Nvidia. http://developer.nvidia.com/object/cuda.html

Technical Specs 2

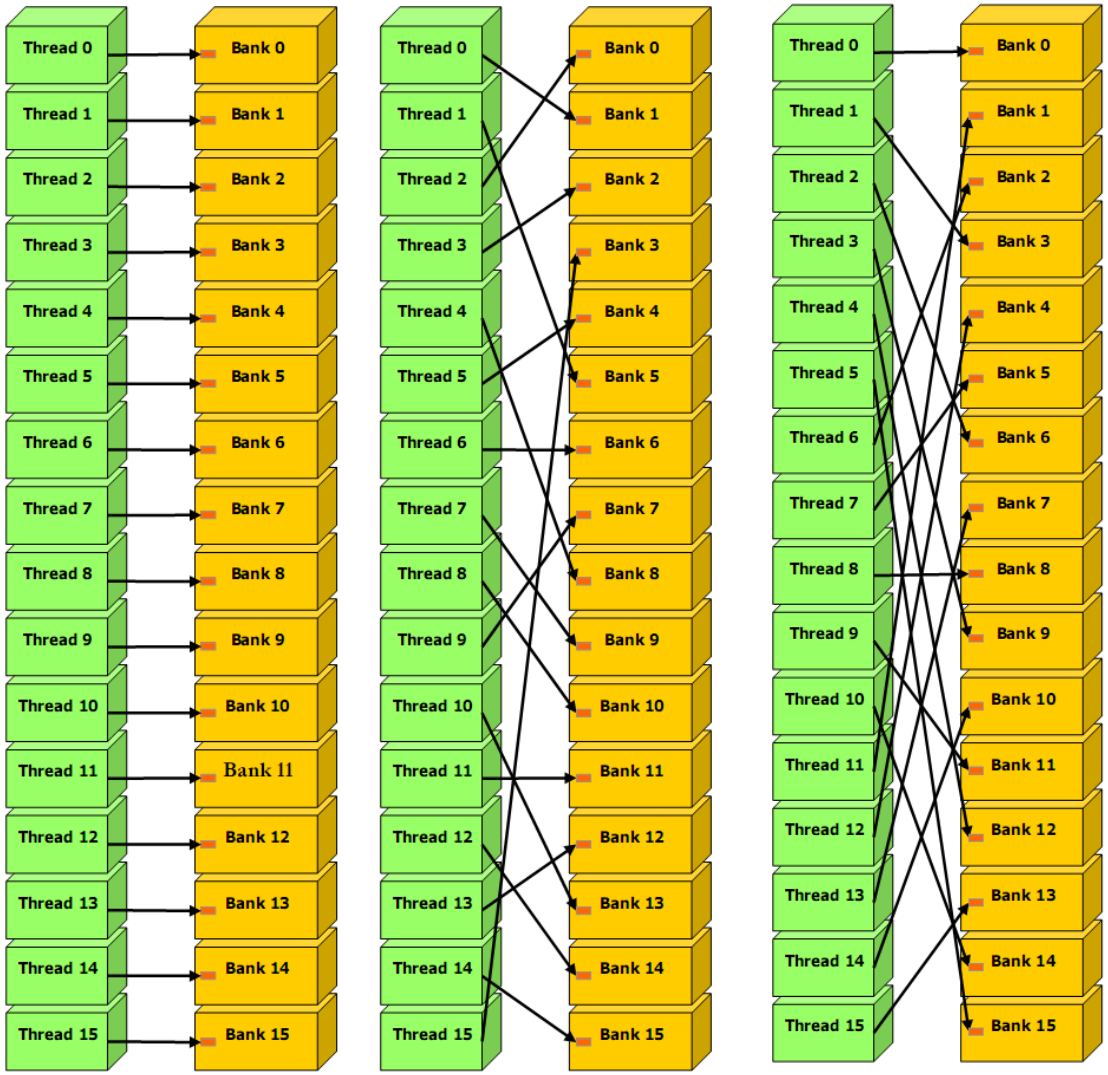
- The cache working set for one-dimensional textures is 8 KB per multiprocessor;
- The maximum number of blocks that can run concurrently on a multiprocessor is 8
- The maximum number of warps that can run concurrently on a multiprocessor is 24;
- The maximum number of threads that can run concurrently on a multiprocessor is 768;
- For a texture reference bound to a one-dimensional CUDA array, the maximum width is 213;
- For a texture reference bound to a two-dimensional CUDA array, the maximum width is 216 and the maximum height is 215;
- For a texture reference bound to linear memory, the maximum width is 227;
  The limit on kernel size is 2 million of native instructions;
- Each multiprocessor is composed of eight processors, so that a multiprocessor is able to process the 32 threads of a warp in four clock cycles.



References: CUDA Programming Guide By Nvidia. http://developer.nvidia.com/object/cuda.html

Shared Memory Access Pattern

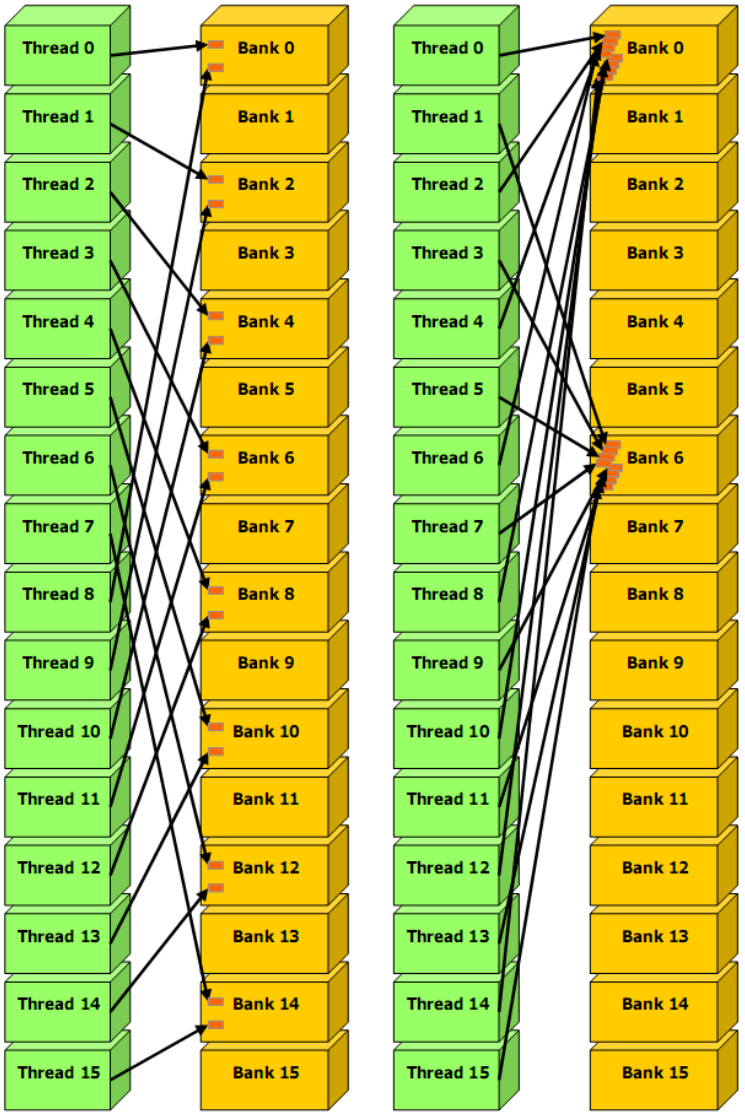
- No Conflicts
  - linear addressing with a stride of one 32-bit word.
  - random permutation
  - Linear addressing with a stride of three 32-bit words.



References: CUDA Programming Guide By Nvidia. http://developer.nvidia.com/object/cuda.html

Shared Memory Access Pattern

- Bank Conflicts
  - Linear addressing with a stride of two 32-bit words causes 2-way bank conflicts
  - Linear addressing with a stride of eight 32-bit words causes 8-way bank conflicts



References: CUDA Programming Guide By Nvidia. http://developer.nvidia.com/object/cuda.html

CUDA Example Program

- **DEMO**

References: CUDA Programming Guide By Nvidia. http://developer.nvidia.com/object/cuda.html