Experiences of porting C to CUDA

Daniel Goodman
Daniel.Goodman@cs.man.ac.uk

Single Precision

- Single precision can be slower than double precision. This is due to poorer quality libraries.

Use the Profiler

- Libraries can hide extremely slow code.
- This includes libraries shipped with the CUDA SDK.

Finding Parallelism

- The same parallelism does not need to be used for every part of a computation.

- Time can be used to add additional opportunities for parallel computation.

Increasing the Concurrency

- The arguments to NVCC can be used to restrict the number of registers per thread
  - This can often be used to reduce the number of registers further without forcing values to be pushed to the main memory
  - To do this on a kernel by kernel basis each kernel must be in a separate file
  - It is not clear that the required registers reported to the GPU is per kernel not per file.

CUDA Context

- Additional information is held on a per CPU thread basis and hidden from the user
- This information prevents:
  - Passing page locked memory directly between threads
  - Multiple threads using the same page locked memory
  - Pointers to locations within the page locked memory being used for streaming or Zero copying