LSU EE 7722 **Homework 3** **Due: 28 March 2016**

Problem 1: Read about the microarchitecture of NVIDIA GPUs from the following resources. Focus on the material needed to understand warp scheduling and instruction issue, and in particular to answer the next problem in this assignment.

NVIDIA has published whitepapers describing the microarchitecture of their GPUs. They describe a variety of features, including those for graphics. When reading them focus on the operation of the MPs (or SM's or SMX's). The whitepapers below describe three generations of GPUs, Fermi (2.X), Kepler (3.X), and Maxwell (5.X). NVIDIA uses a two-letter, three-digit name to describe the microarchiteture. The first letter is always g, the second letter indicates the generation (f, k, m), and the number indicates something like a version.

For Fermi read `http://www.ece.lsu.edu/gp/refs/gf100-whitepaper.pdf`, for Kepler read `http://www.ece.lsu.edu/gp/refs/NVIDIA-Kepler-GK110-Architecture-Whitepaper.pdf`, and for Maxwell read. `http://www.ece.lsu.edu/gp/refs/GeForce-GTX-980-Whitepaper-FINAL.pdf`.

Additional information can be found in Section 4 of the CUDA C Programming Guide: `https://docs.nvidia.com/cuda/cuda-c-programming-guide`.

Problem 2: In the matrix/vector multiply code used in class we noticed that when we assigned an entire matrix/vector multiply to a thread the FMADD instructions could access matrix elements directly. But when a thread only operated on one output vector component the compiler emitted `LDC` instructions.

Based on the descriptions above, would such `LDC` instructions be a barrier to saturating FP capability (assuming sufficiently large input)?

Answer your question with diagrams showing how instructions are issued and indicate where in the references the illustrated behavior is supported.