Accelerating Floating Point DGEMM on FPGAs

Martin Langhammer
Tom VanCourt
Altera Corp.

Approved for public release;
distribution is unlimited.

Floating Point on FPGAs

- "But FPGAs can't ..."
  - But they ***CAN***!
- FP operation:
  1) Normalize operands
     Add/sub/mul/div
     Normalize result
  2) Normalize operands
     Add/sub/mul/div
     Normalize result
  3) Normalize operands
     Add/sub/mul/div
     Normalize result
  4) ...

- FP compiler
  - Create fused data paths
  - Insert guard bits to reduce normalizations
  - Selection context-sensitive operation blocks
- Improved performance
  - Logic reduction: to 40%
  - Latency reduction: to 40%
  - Clock rates: to 200 MHz

Matrix Multiplication

- Decompose arrays into blocks
  - Large (M144K) RAMs hold column values
  - Small (M9K) RAMs present new row every cycle
  - Configurable to 128 DP values – 8Kb per cycle
- Launch new dot product every cycle
  - Pipelined: 128 mult + 127 add = 255 FLOP/cycle
  - (255 FLOP/cycle) * (~200M cycle/sec) = ~50G FLOP/s
- Data rate sustained until throttled by system bus
  - Operation concurrency: 100s of dedicated multipliers
  - Data concurrency: 100s of independently addressable RAMs
- Competitive with Xeon, GPGPU
  - In sustained performance and MFLOP/s per Watt