Quad Vectorization

- Bifrost uses quad-parallel execution
  - Four scalar threads executed in lockstep in a “quad”
  - One quad at a time executes in each pipeline stage
  - Each thread fills one 32-bit lane of the hardware
  - 4 threads doing a vec3 FP32 add takes 3 cycles
  - Improves utilization

- Quad vectorization is compiler friendly
  - Each thread only sees a stream of scalar operations
  - Vector operations can always be split into scalars