Exercise 4.9 a)

ANS

This code reads four SP floats of 16 bytes

writes two SP floats of 8 bytes for every six FLOPs.

So arithmetic intensity= 6/24

= 0.25 FLOP per byte of data accessed

Exercise 4.10

ANS

For the Vector Processor,

Computation Time = Computation Time for scalar execution+ Memory Access time

= 400 ms+ (200 + 100) MB/(30 GB/ sec)

= 410 ms

For the Hybrid Processor,

Computation Time = Computation Time for scalar execution+ Memory Access time + Transfer time between host and local memory +Memory latency

= 400ms+(200+100)MB/(150 GB/sec)+(200+100)MB/(10 GB/sec) + 10ms

= 442 ms

So the vector processor achieves better performance than the hybrid processor.

Exercise 4.13

a)

ANS

GFLOPs/sec = 1.5 x 10 x 0.8 x 0.85 x 0.7 x (32/4)

= 57.12

b)

ANS

1)

If we increase the lanes by 16, GFLOPs/sec = 1.5 x 10 x 0.8 x 0.85 x 0.7 x (32/2)

= 114.24

Speedup = 114.24/57.12

= 2

2)

If we increase the number of SIMDs to 15, GFLOPs/sec = 1.5 x 15 x 0.8 x 0.85 x 0.7 x (32/4)

= 85.68

Speedup = 85.68/57.12

= 1.5

3)

If we increase the issue rate to 0.95, GFLOPs/sec = 1.5 x 15 x 0.8 x 0.95 x 0.7 x (32/4)

= 63.84

Speedup = 63.84/57.12

= 1.12

Exercise 4.16

ANS

The clock rate of a hypothetical GPU is 1.5 GHz, exists 16 SIMD processors, each processor contains 16 single-precision floating point units and off-chip memory bandwidth is 100 GB/sec.

For this GPUs the peak single-precision floating-point throughput is,

core frequency x number of cores x number of operations per clock =1.5 x 16 x 16

=384 GFLOP/sec

Assuming the each single precision operation required 4 Byte 2 operands and output one four byte result, sustaining would required the memory bandwidth

= 12 Bytes/ Flop x 384 GFLOPS/sec

= 4.608 TB/s

Throughput is not sustainable because 4.608 TB/sec >100GB/sec

But still can be achieved in short bursts when using on-chip memory