ELE 455/555
Computer System Engineering

Section 4 – Parallel Processing

Class 4 – Performance

Parallel Processing

Performance

- Benchmarks
  - Targeted at various aspects of parallel programs
  - Linpack:
    - Matrix linear algebra
  - SPECrate:
    - Parallel run of SPEC CPU programs
    - Job-level parallelism
  - SPLASH: Stanford Parallel Applications for Shared Memory
    - Mix of kernels and applications, strong scaling
  - NAS (NASA Advanced Supercomputing) suite
    - Computational fluid dynamics kernels
  - PARSEC (Princeton Application Repository for Shared Memory Computers) suite
    - Multithreaded applications using Pthreads and OpenMP

Parallel Processing

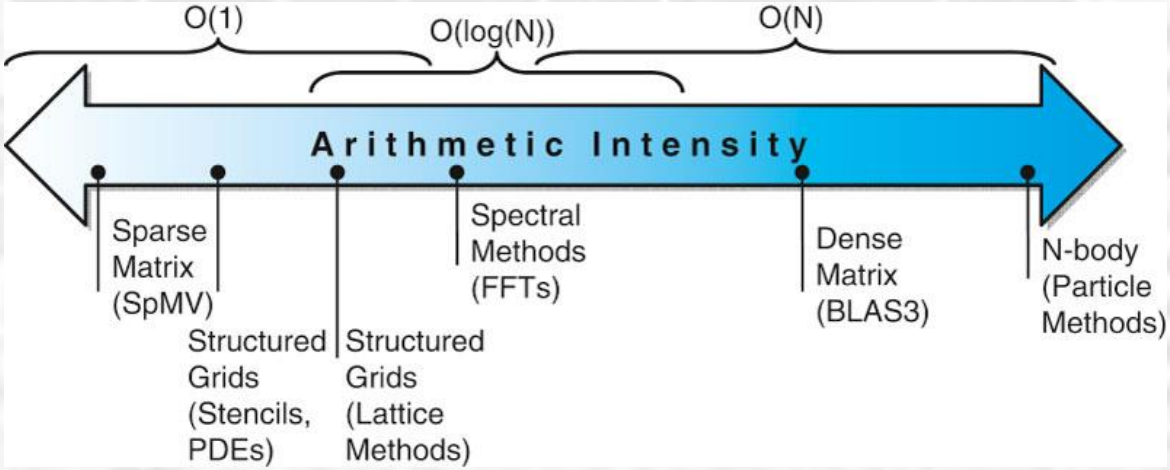
Performance

- Benchmarks
  - Parallelism makes comparing systems even harder
    - Scale the data?
    - Algorithm changes
      - Might you attack the problem differently based on your resources
  - UCB approach
    - Identify the design patterns that will be part of near future applications
    - Implement them any way you want

Parallel Processing

Performance

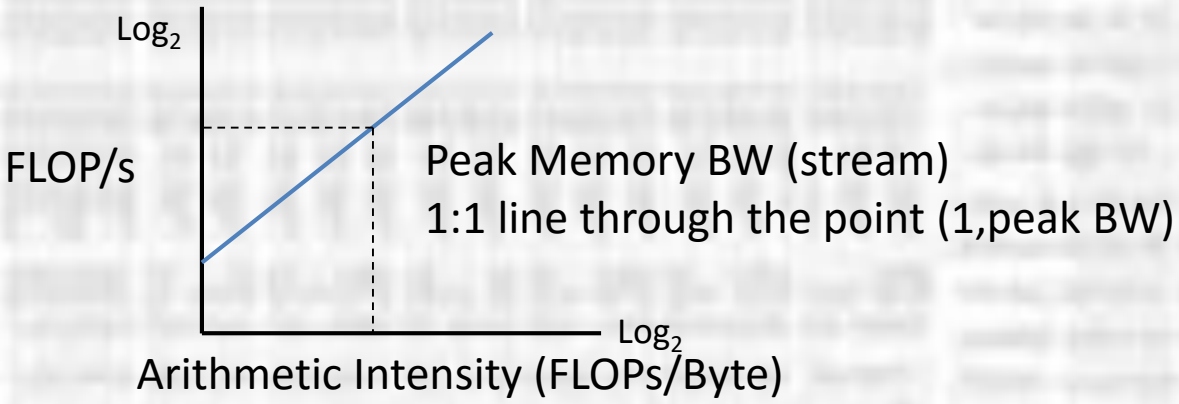
- Models
  - Floating point operations are part of many implementations
  - Arithmetic Intensity
    - Ratio of FLOPs to memory accesses (bytes)
  - Relative order of Arithmetic Intensity



Parallel Processing

Performance

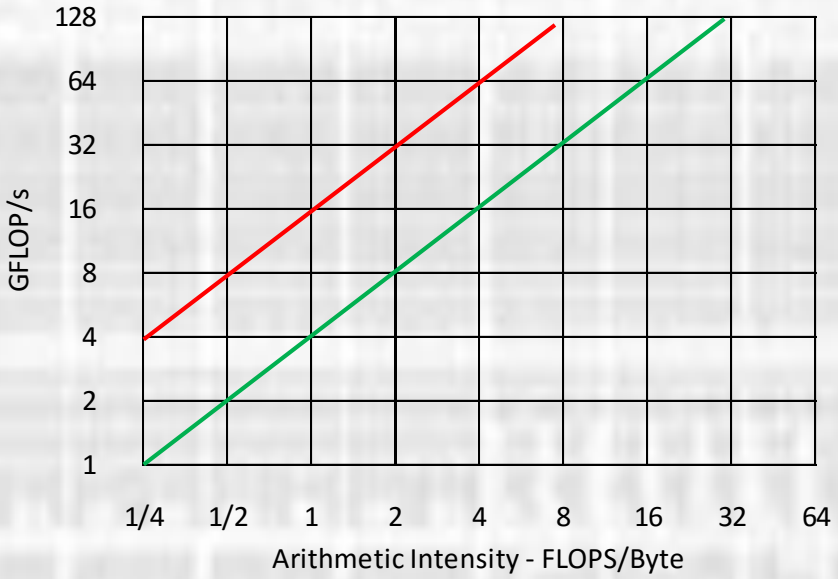
- Models
  - Stream Benchmark
    - Measures the memory performance for large data structures that do not fit in the cache
    - A good measure for our multiprocessing systems
    - Measures peak memory performance



Parallel Processing

Performance

- Models
  - Stream Benchmark



Peak memory BW = 16GB/s

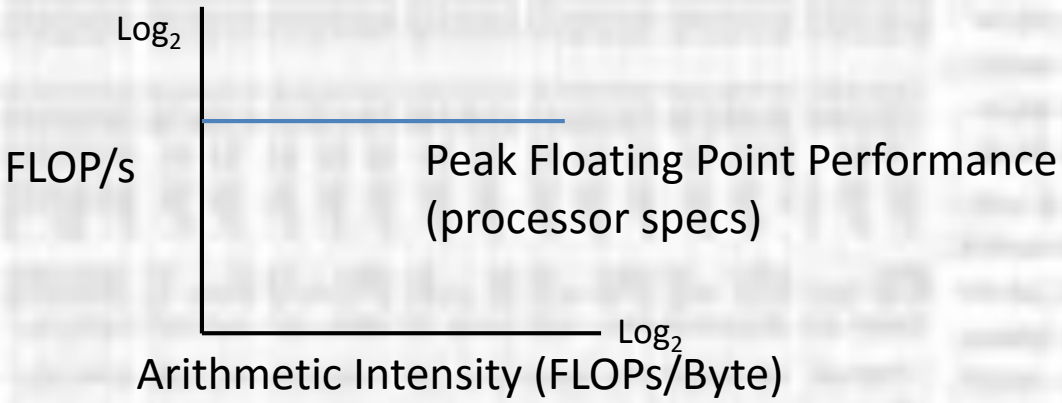
Peak memory BW = 4GB/s

If your arithmetic intensity is 4 FLOPS/Byte
AND
Your peak memory BW = 4GB/s
THEN
Your potential FLOPS/s = 16GFLOPS/s

Parallel Processing

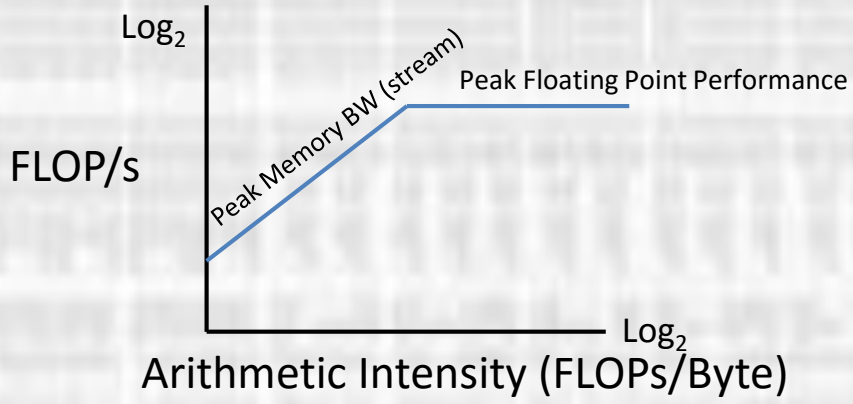
Performance

- Models
  - Peak Floating Point Performance
    - Processor dependent
      - max clock rate
      - no stalls



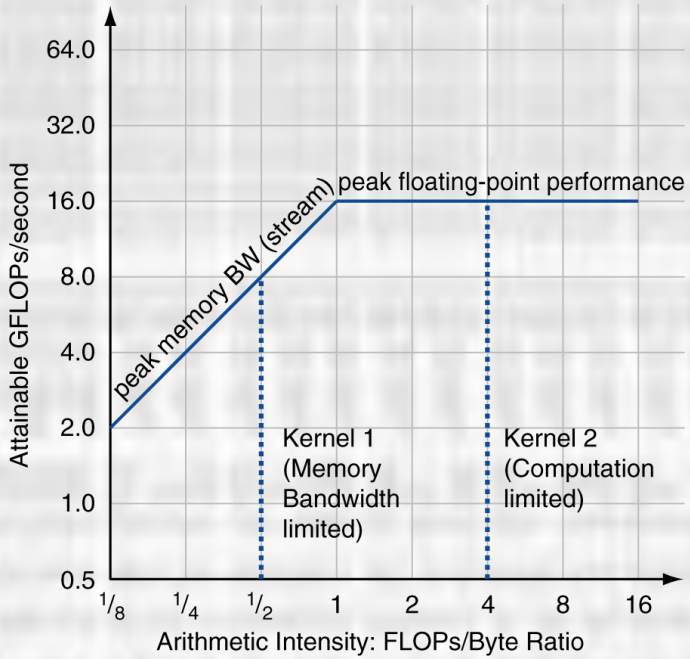
Parallel Processing
Performance

- Models
  - Roofline Model
    - Peak Memory and Floating Point performance plotted together



Parallel Processing
Performance

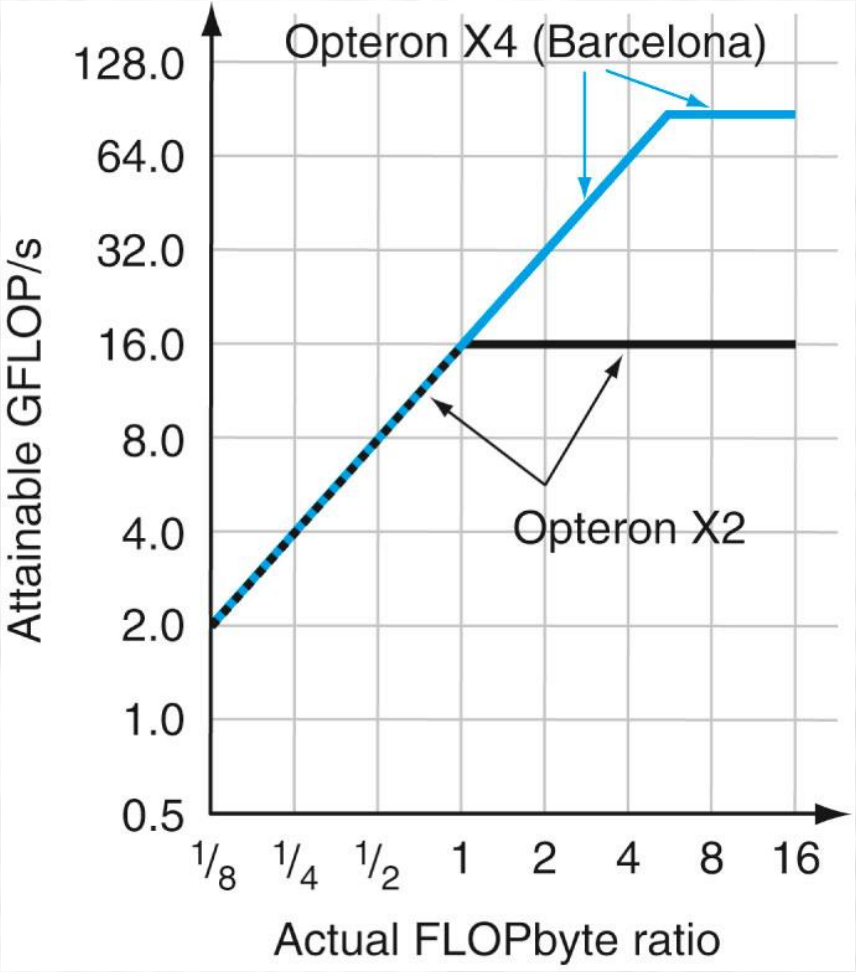
- Models
  - Roofline Model
    - Peak Floating Point Performance vs. Arithmetic Intensity



Parallel Processing

Performance

- Example
  - Opteron X2
    - 2 cores
    - 1 FLOP/core
    - 2.2GHz
  - Opteron X4
    - 4 cores
    - 2 FLOP/core
    - 2.3GHz
  - Expect 4x peak performance
    - x4 has an L3 cache



Parallel Processing

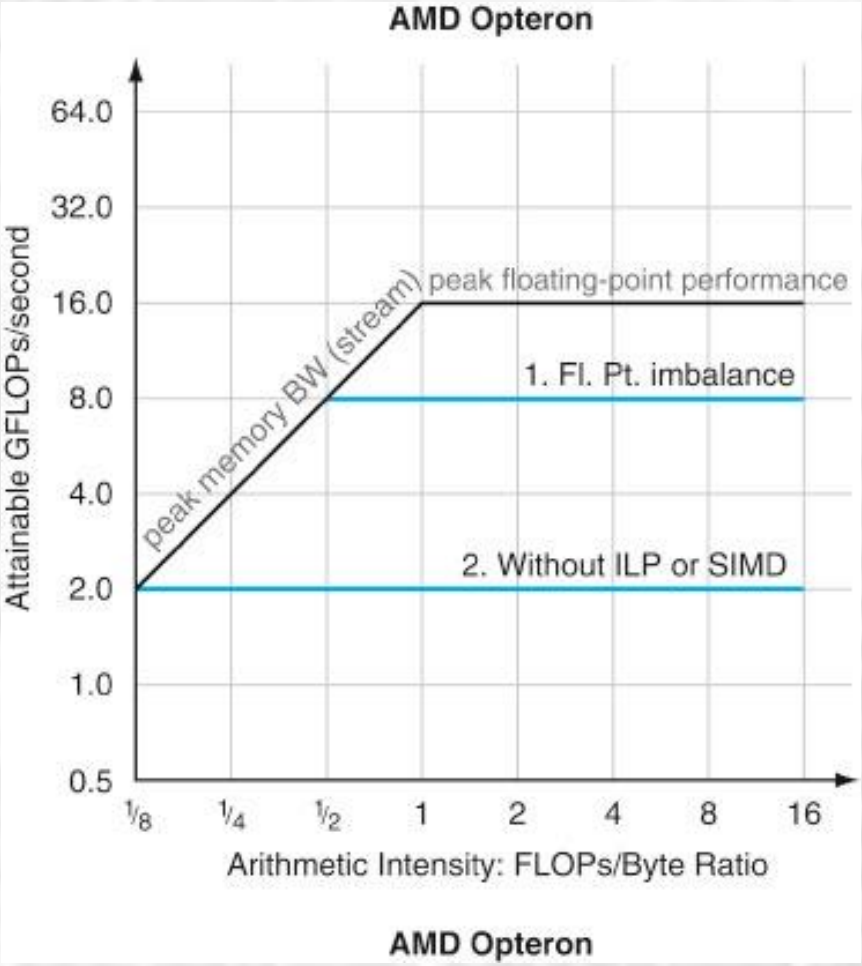
Performance

- Kernel Optimizations
  - Roofline is a maximum possible, you may get less
  - Computational Bottlenecks
    - Floating Point Operations mix
      - Imbalance in Floating point instructions vs. adds
      - Pipeline has balanced FP and Add structures
      - Need balance to fully utilize the ALU
    - Improve Instruction Level Parallelism (ILP)
      - Superscalar architectures need multiple instructions to leverage multi-issue resources

Parallel Processing

Performance

- Kernel Optimizations
  - Computational Bottlenecks
    - Floating Point Operations mix
      - potential 2x improvement
    - Improve ILP
      - potential 4x improvement
  - ILP first, then mix



Parallel Processing

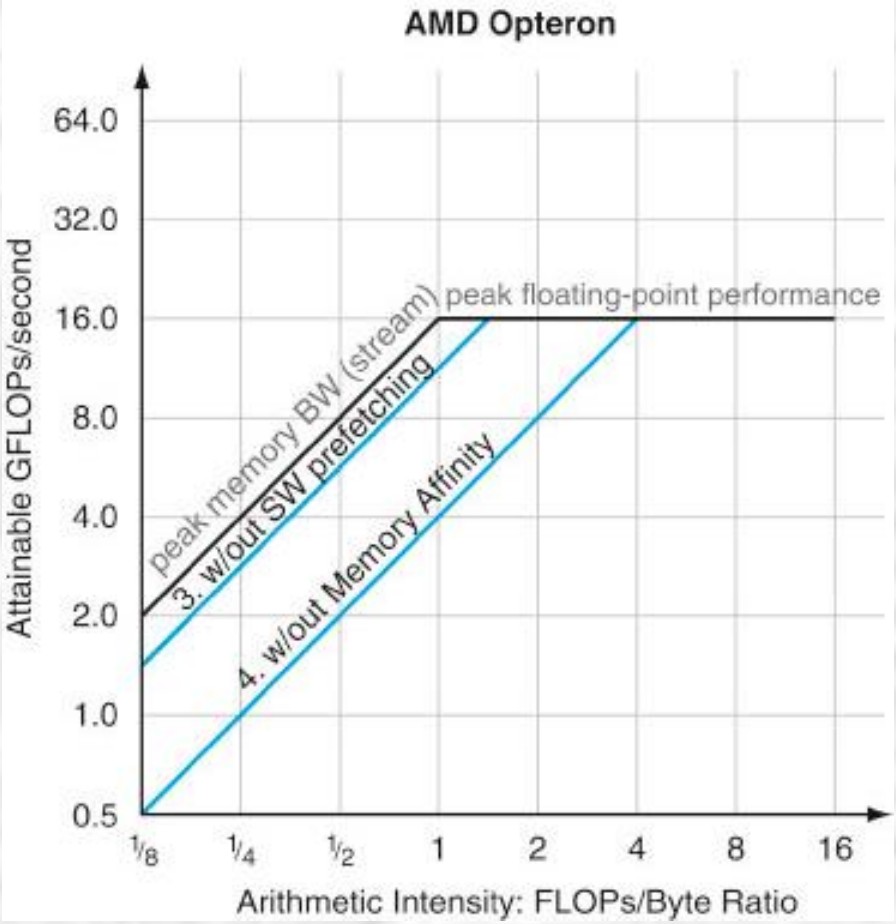
Performance

- Kernel Optimizations
  - Roofline is a maximum possible, you may get less
  - Memory Bottlenecks
    - Software Prefetching
      - Need to predict data needs to reduce stalls
      - Special instructions required to load data into the cache before it is needed
    - Memory Affinity
      - Leverage NUMA characteristics
      - Tie threads to processor / closest memory pairs
      - Try to keep a processor accesses to the lowest latency memory

Parallel Processing

Performance

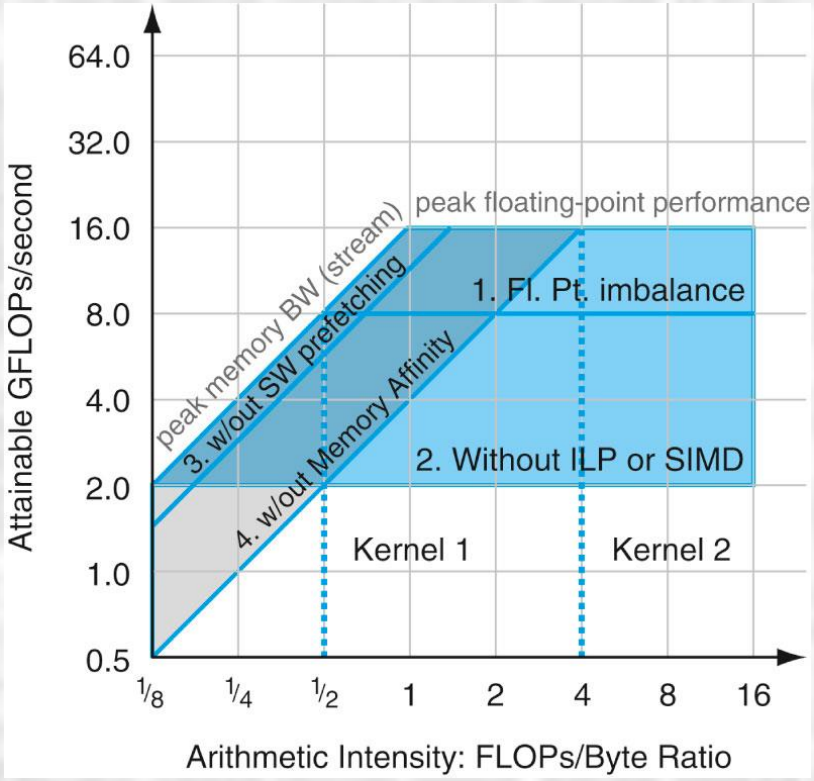
- Kernel Optimizations
  - Memory Bottlenecks
    - Software Prefetching
      - potential 50% improvement
    - Memory Affinity
      - potential 2.66x improvement
  - Affinity first, then pre-fetching



Parallel Processing

Performance

- Kernel Optimizations
  - Operating Region determines where to optimize



Parallel Processing

Performance

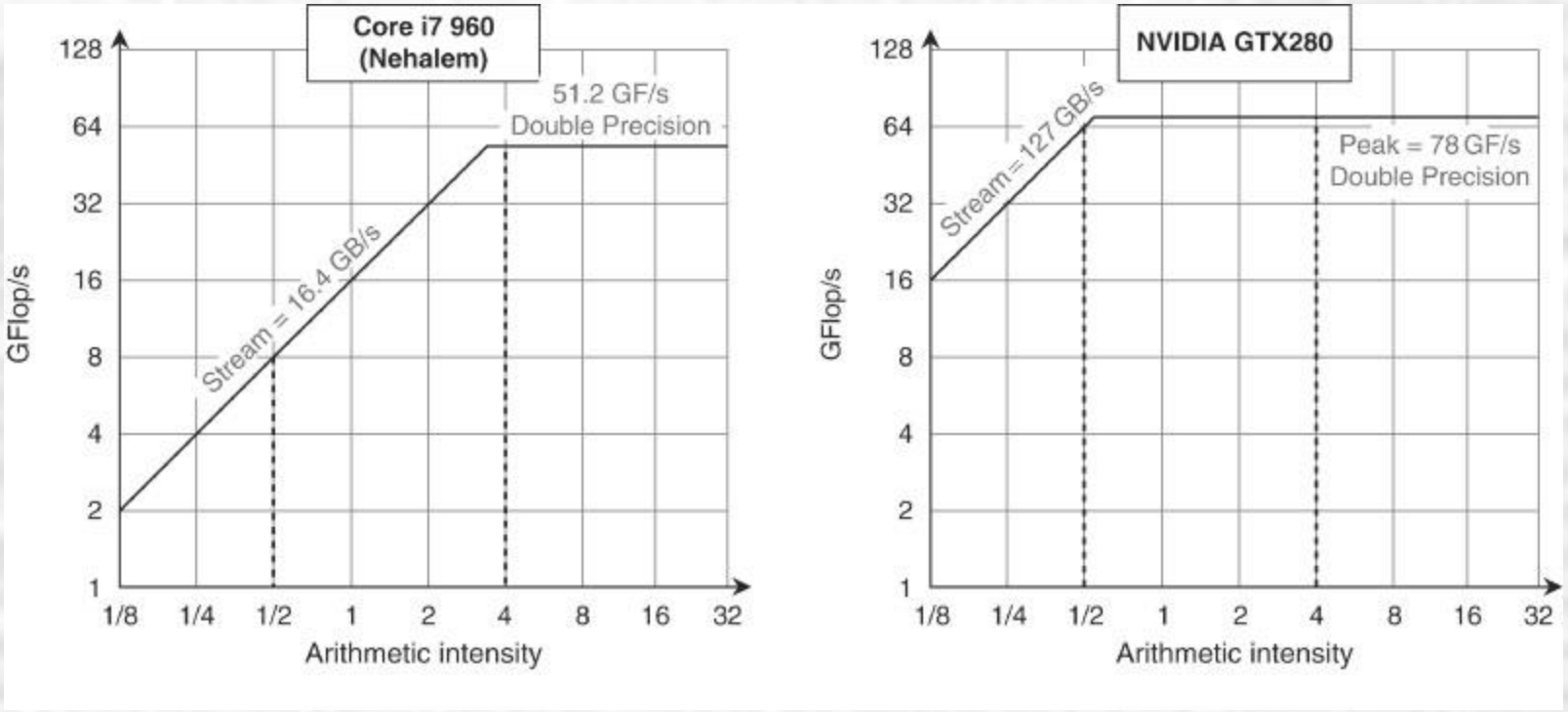
- Example
  - Core I7
  - GTX 280

	Core i7-960	GTX 280	Ratio 280/i7
Number of processing elements (cores or SMs)	4	30	7.5
Clock frequency (GHz)	3.2	1.3	0.41
Die size	263	576	2.2
Technology	Intel 45 nm	TSMC 65 nm	1.6
Power (chip, not module)	130	130	1.0
Transistors	700 M	1400 M	2.0
Memory brandwith (GBytes/sec)	32	141	4.4
Single-precision SIMD width	4	8	2.0
Double-precision SIMD width	2	1	0.5
Peak Single-precision scalar FLOPS (GFLOP/sec)	26	117	4.6
Peak Single-precision SIMD FLOPS (GFLOP/Sec)	102	311 to 933	3.0–9.1
(SP 1 add or multiply)	N.A.	(311)	(3.0)
(SP 1 instruction fused multiply-adds)	N.A.	(622)	(6.1)
(Rare SP dual issue fused multiply-add and multiply)	N.A.	(933)	(9.1)
Peal double-precision SIMD FLOPS (GFLOP/sec)	51	78	1.5

Parallel Processing

Performance

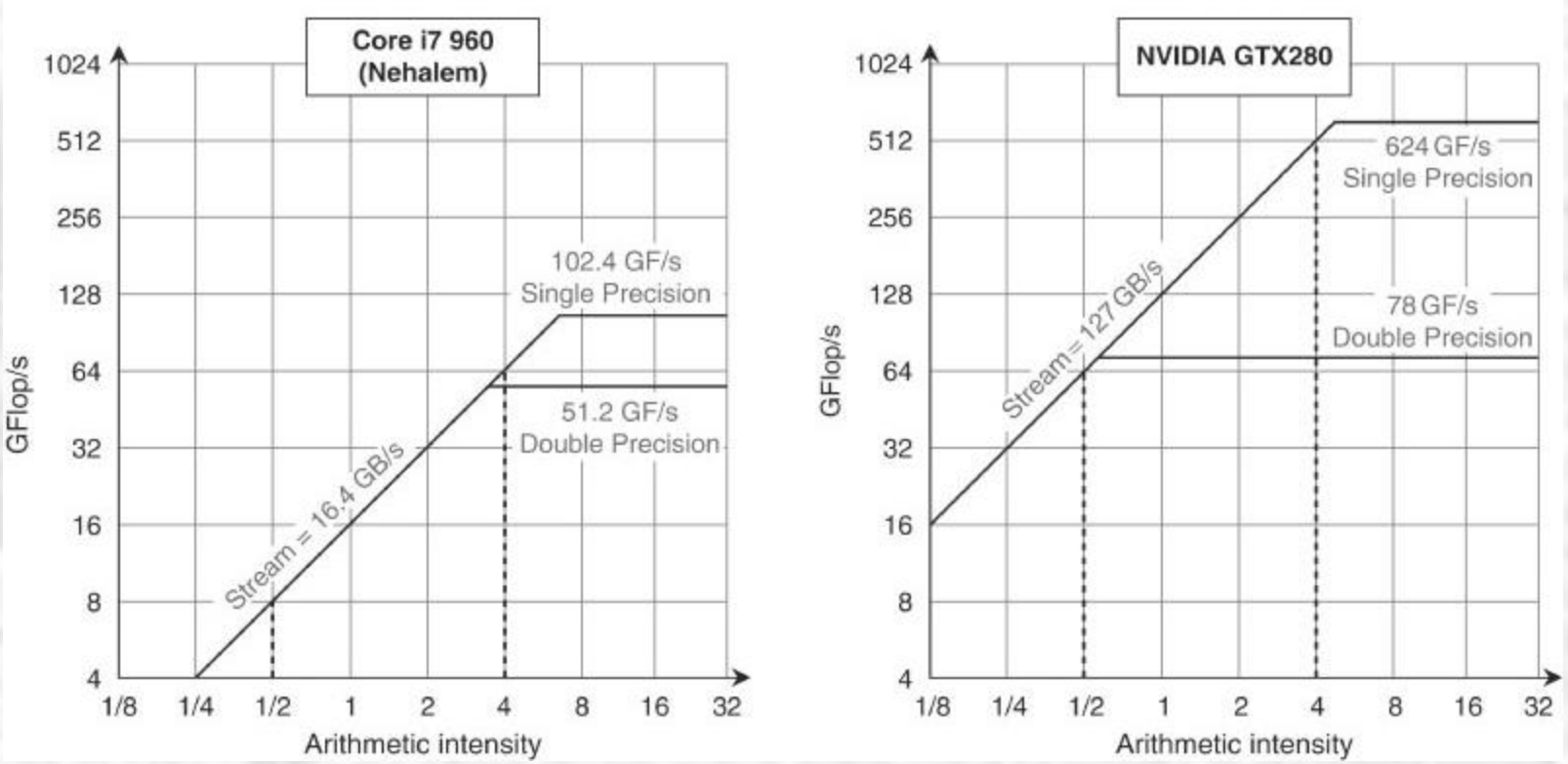
- I7-960 vs. GTX280
  - Double Precision FP



Parallel Processing

Performance

- I7-960 vs. GTX280
  - Single Precision FP



Parallel Processing

Performance

- I7-960 vs. GTX280

Kernel	Units	Core i7-960	GTX 280	GTX 280/ i7-960
SGEMM	GFLOP/sec	94	364	3.9
MC	Billion paths/sec	0.8	1.4	1.8
Conv	Million pixels/sec	1250	3500	2.8
FFT	GFLOP/sec	71.4	213	3.0
SAXPY	GBytes/sec	16.8	88.8	5.3
LBM	Million lookups/sec	85	426	5.0
Solv	Frames/sec	103	52	0.5
SpMV	GFLOP/sec	4.9	9.1	1.9
GJK	Frames/sec	67	1020	15.2
Sort	Million elements/sec	250	198	0.8
RC	Frames/sec	5	8.1	1.6
Search	Million queries/sec	50	90	1.8
Hist	Million pixels/sec	1517	2583	1.7
Bilat	Million pixels/sec	83	475	5.7

Significant single precision Floating Point content

Large data sets (memory BW)

High synchronization demands

Scattered data

Transcendentals (Native in GTX)

ELE455/555

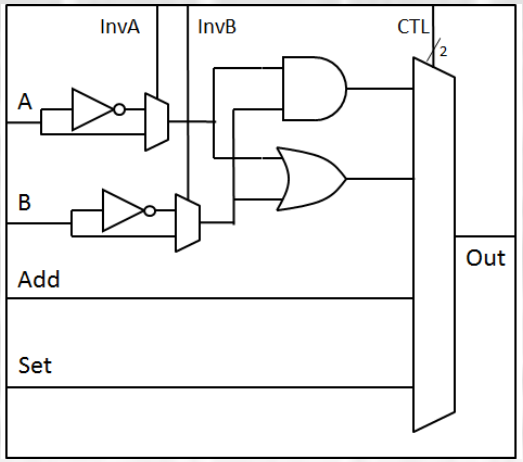
Semester Review

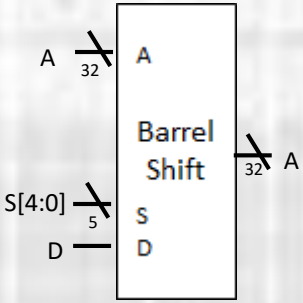
Quick Review of the Semester

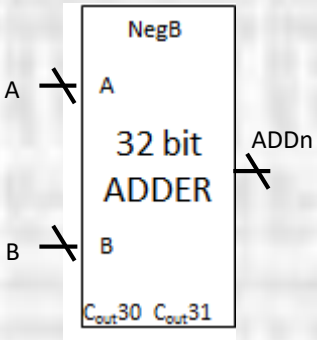
ELE 455/555

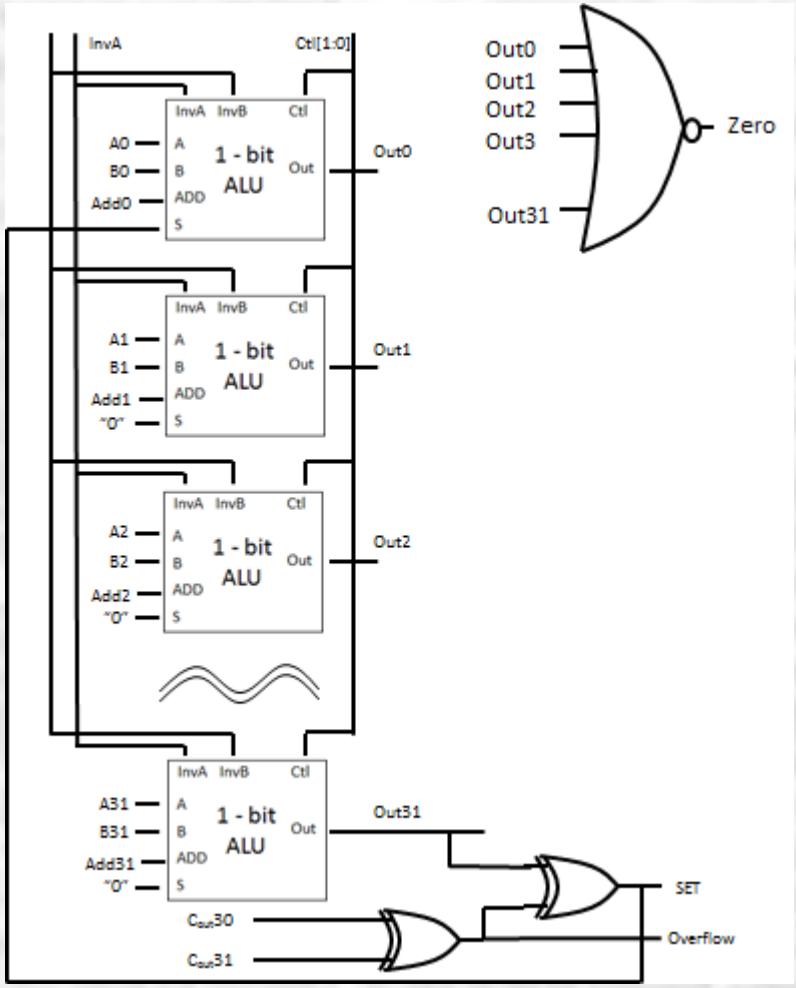
Semester Review

- ALU – Implementation









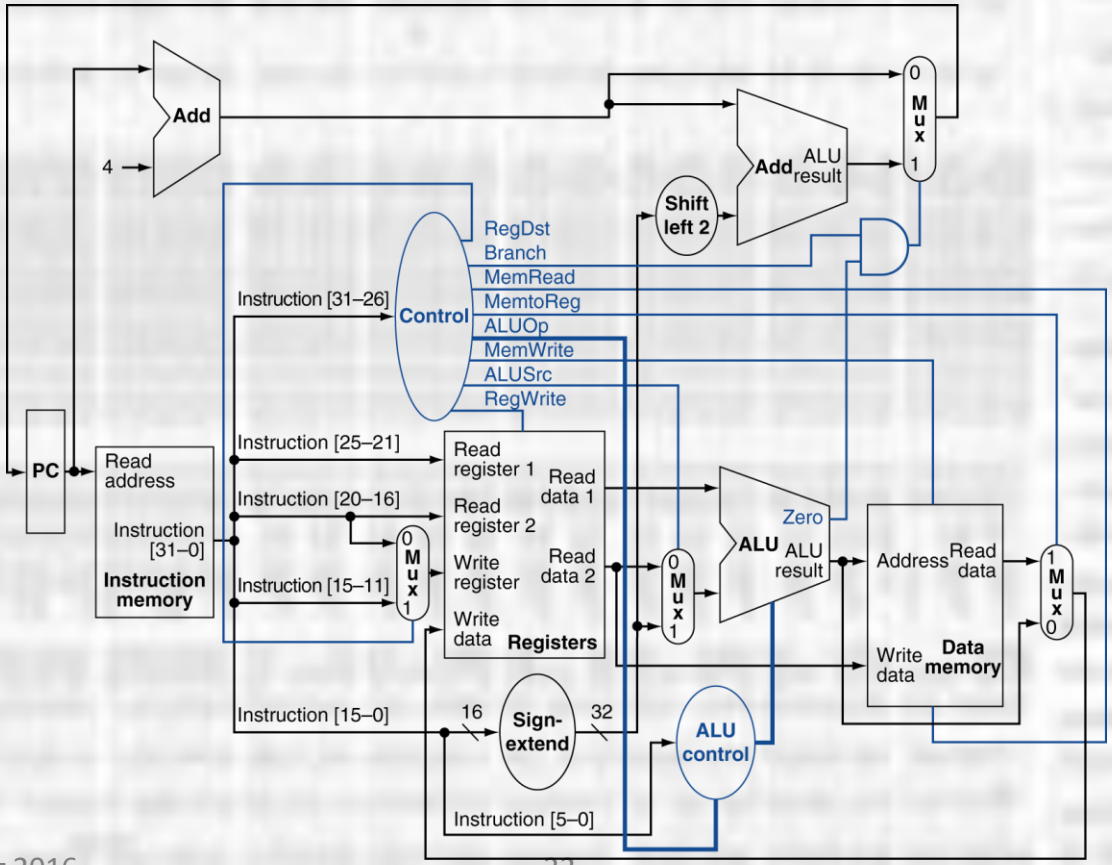
ELE 455/555

Semester Review

- Datapath Control – BEQ

op	rs	rt	constant or address
6 bits	5 bits	5 bits	16 bits

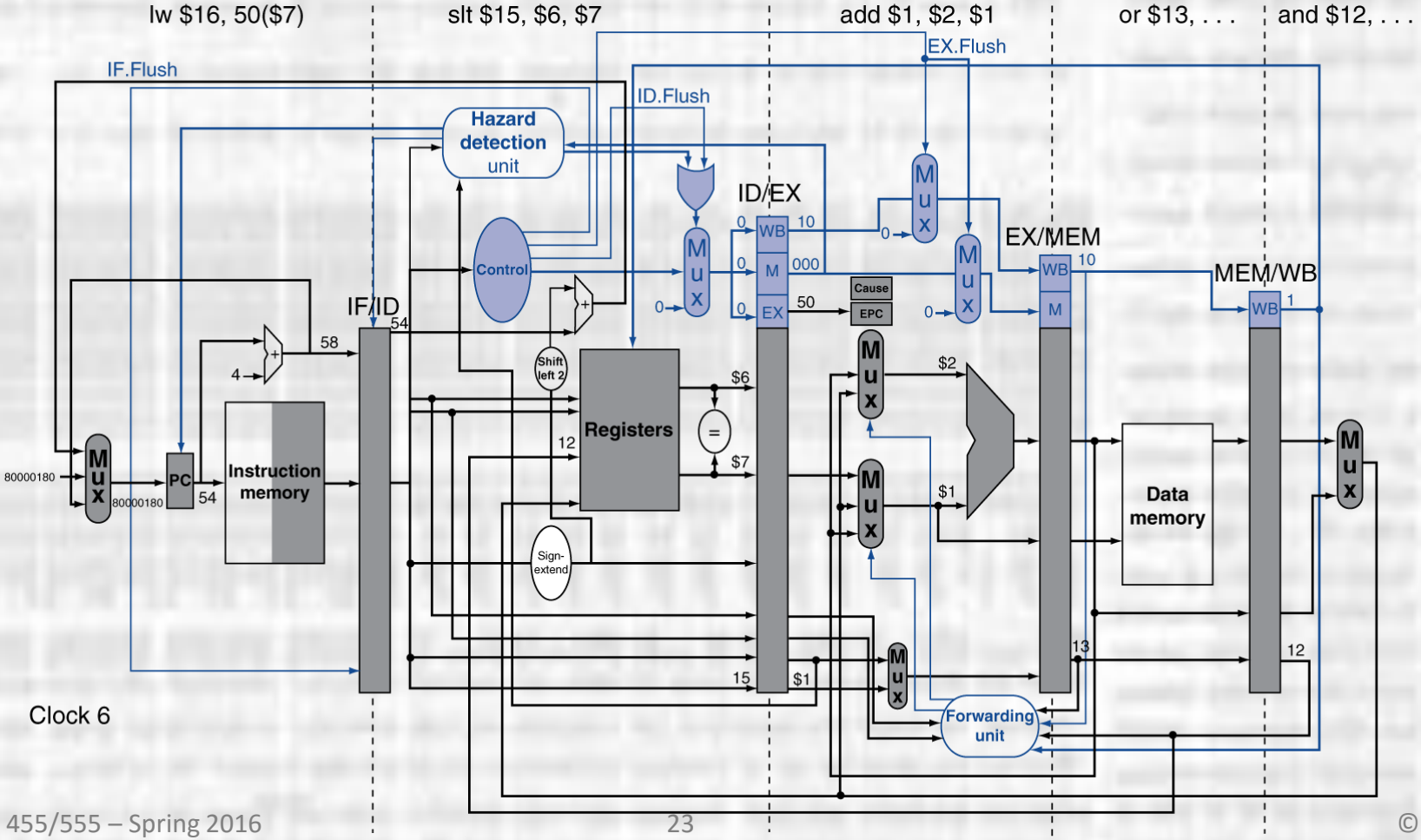
Instruction	RegDst	ALUSrc	MemtoReg	RegWrite	MemRead	MemWrite	Branch	ALUOp1	ALUOp0
LW	X	0	X	0	0	0	1	0	1



ELE 455/555

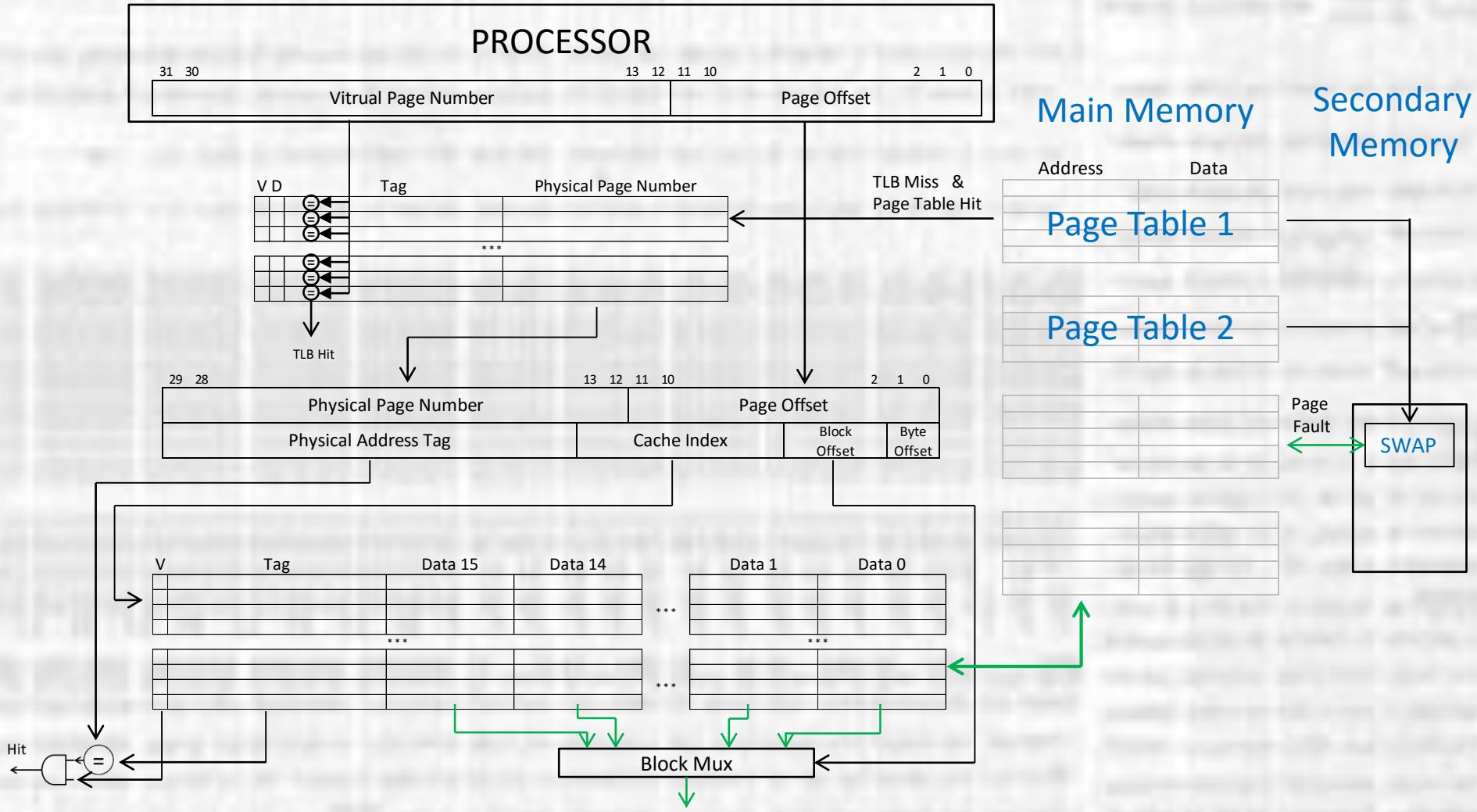
Semester Review

- Exceptions in a Pipeline - example



ELE 455/555

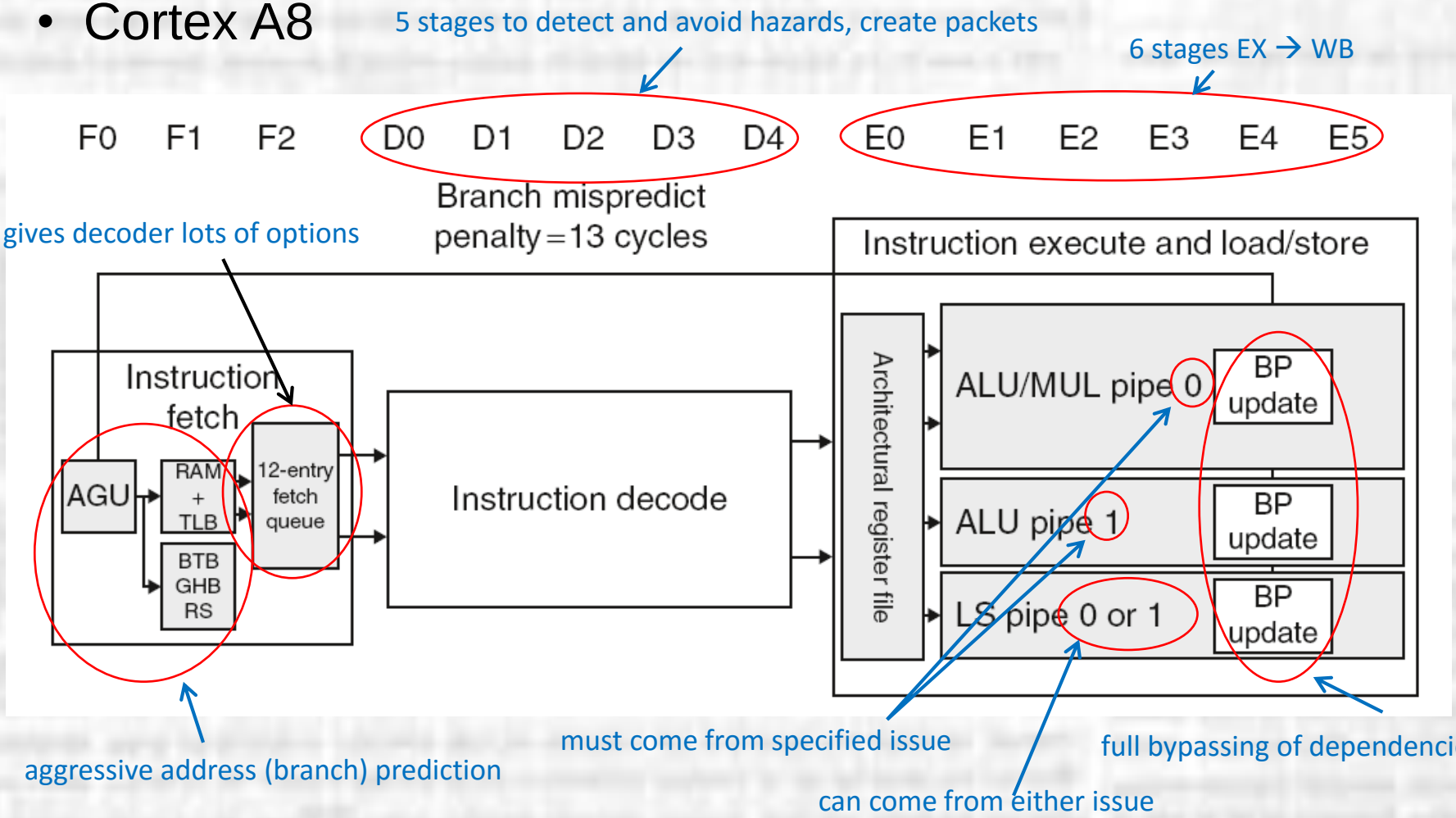
Semester Review



ELE 455/555

Semester Review

- Cortex A8



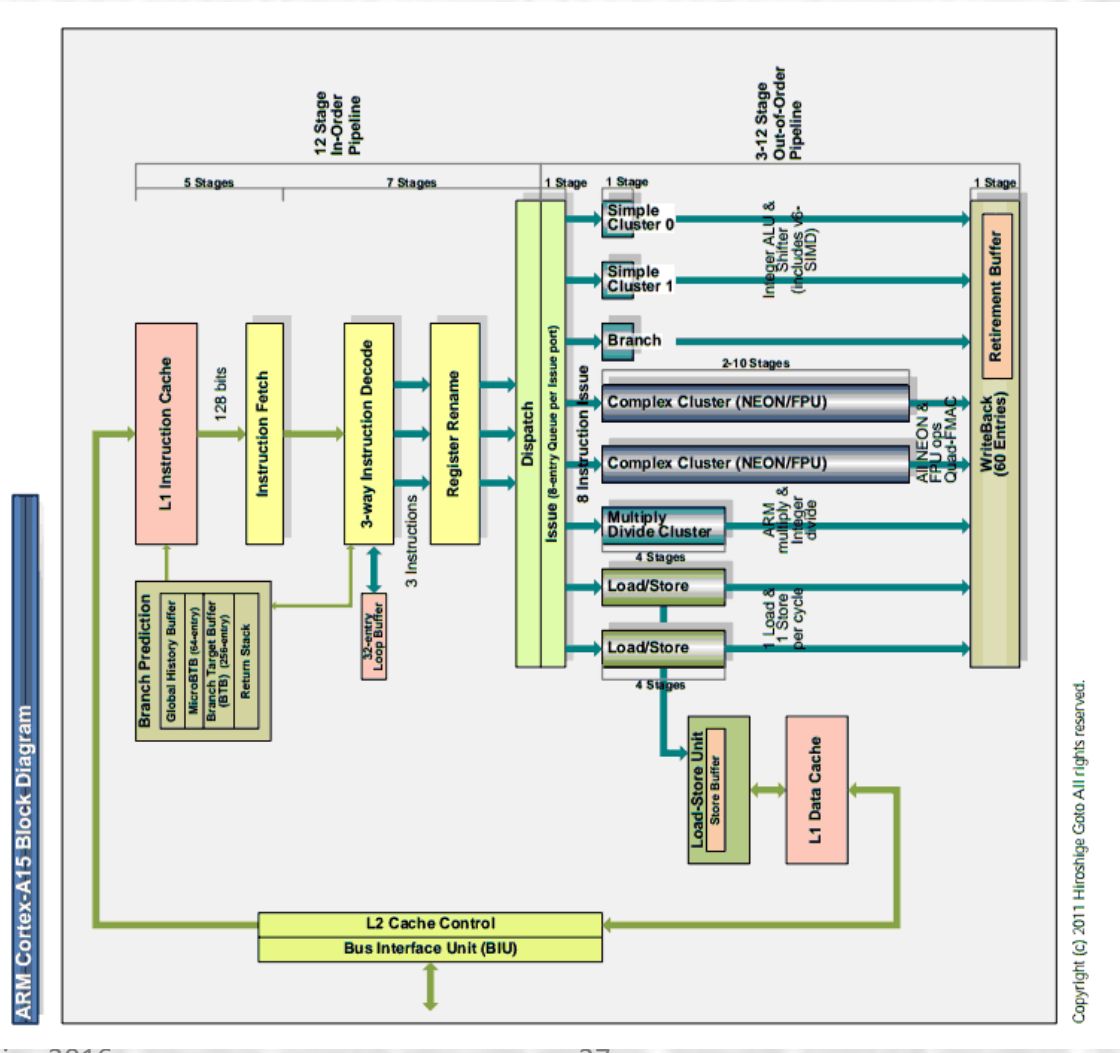
ELE 455/555

Semester Review

- Nvidia Kepler
  - 192 cores

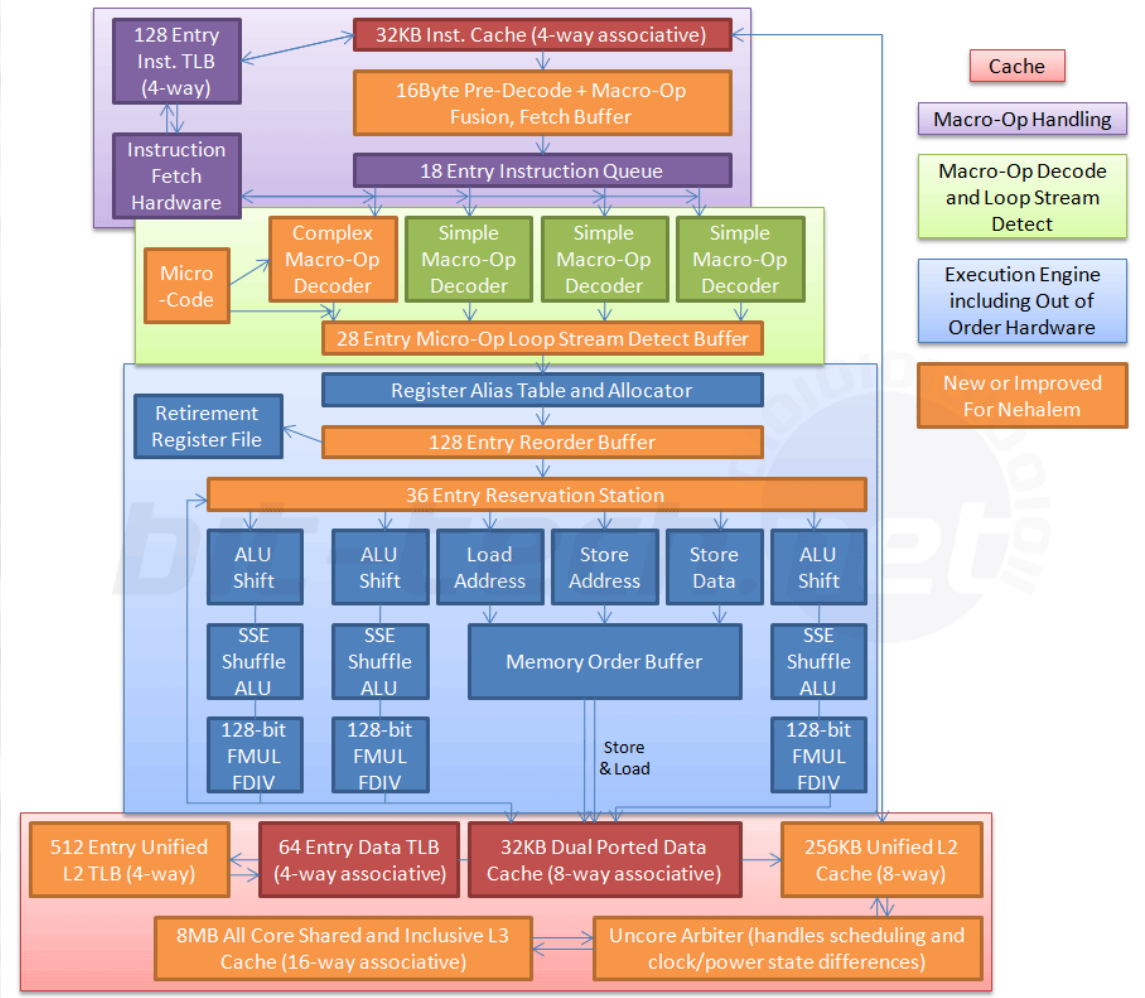
ELE 455/555

Semester Review



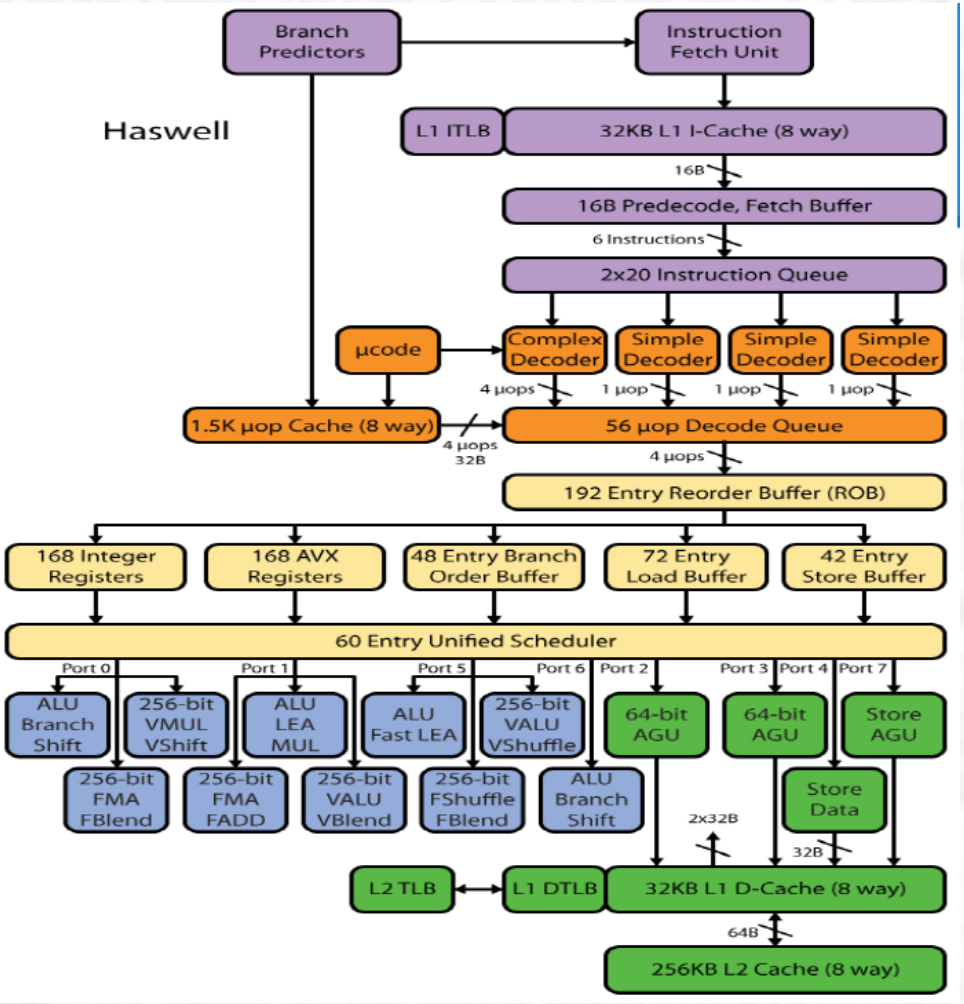
ELE 455/555

Semester Review



ELE 455/555

Semester Review



ELE 455/555

Semester Review