Homework T2: ex 1 on mem hierarchy

Consider the following study case:

- execution of a piece of code in the SeARCH node with the Xeon Skylake (Gold 6130); detailed info on this Intel microarchitetcure in https://en.wikichip.org/wiki/intel/microarchitectures/skylake_(server) ;
- execution of the same 2 instructions (that are already in the instruction cache) in all cores of a single chip: load a double in register, followed by a multiplication by another double in a memory distant location;
- the Skylake cores are 6-way superscalar and each core has 2 load units;
- these instructions are executed with a cold data cache.

Compute:

a) The max required bandwidth to access the external RAM when executing these 2 instructions (to simplify, consider clock rate = 2 GHz).

b) The aggregate peak bandwidth available in this Xeon device to access the installed DRAM-4 (using all memory channels).

Resolution

Exercise 1 on mem hierarchy

Consider the following case study:

- … code in the SeARCH node with the Xeon Skylake …
- … same 2 instructions … in all cores of a single chip…
- … cores 6-way superscalar … 2 load units/core … cold data cache.

Compute:

a) The max required bandwidth to access the external RAM …

b) The aggregate peak bandwidth … DRAM-4 (w/ all memory channels).

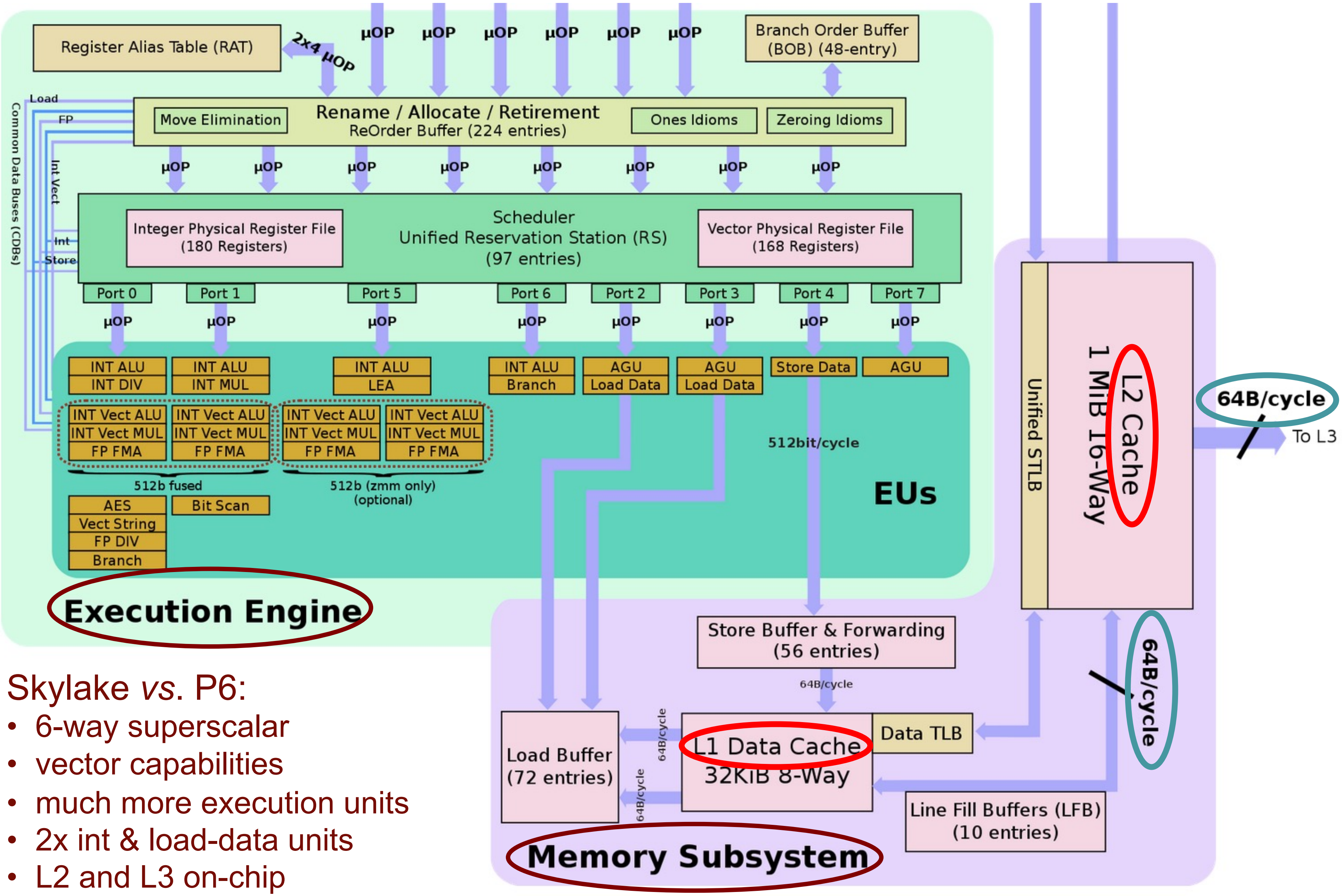
- each clock cycle needs 2 mem accesses to fetch 2 doubles
- **max required bandwidth** to fetch a cache line for each double (cache is cold & doubles are far away):

  ???? GB/s

  note: the following 7 pairs of doubles are already in cache
- RAM in each Skylake Gold 6130: 6x DDR4-2666 (6x8 GiB)
- **peak bandwidth** of 6x DDR4-2666 in 6 memory channels:

  ???? GB/s

Partial view of a Skylake core (server)

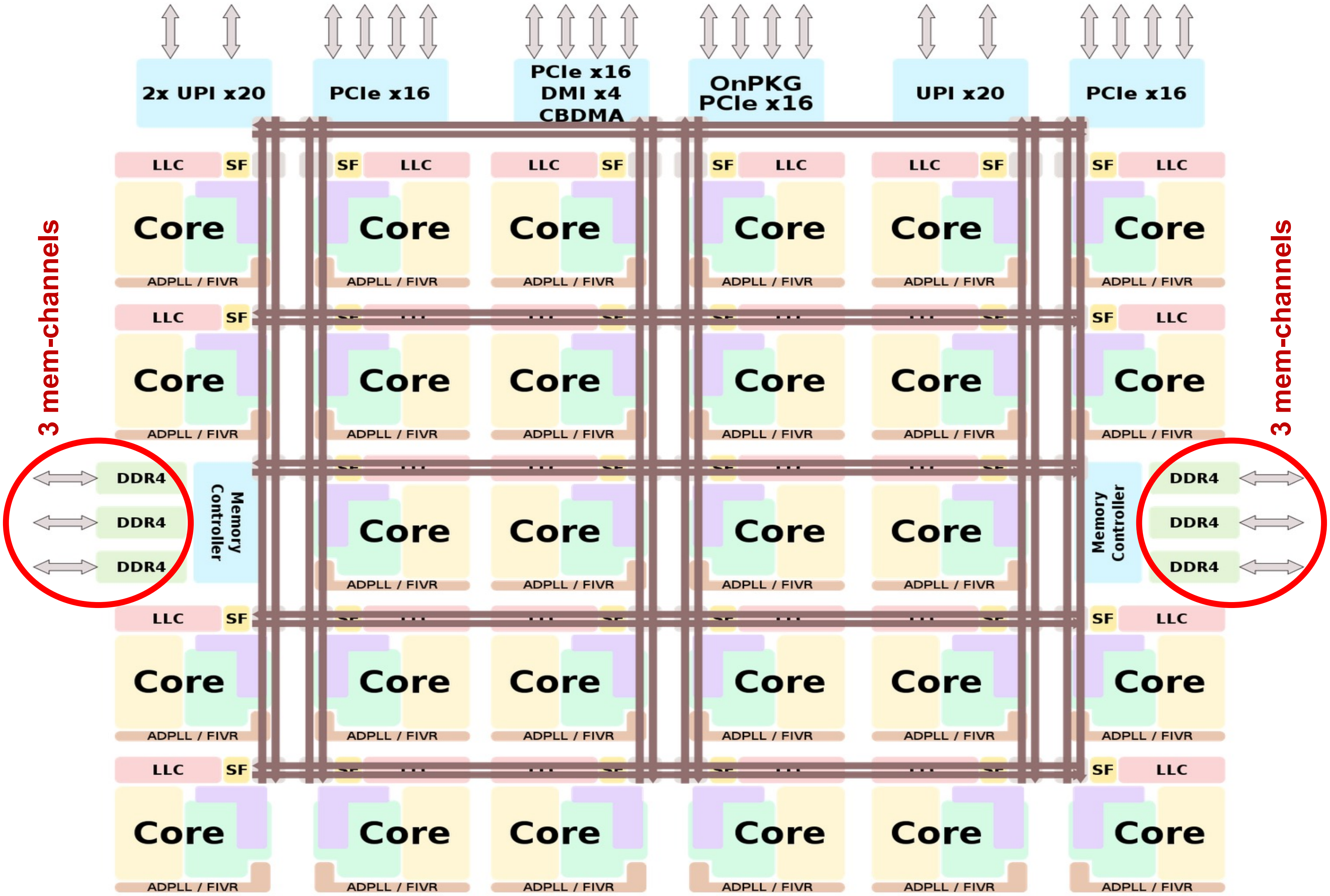


https://en.wikichip.org/wiki/intel/microarchitectures/skylake_(server)

Skylake *vs*. P6:

- 6-way superscalar
- vector capabilities
- much more execution units
- 2x int & load-data units
- L2 and L3 on-chip

Architecture of a 28-core Skylake (server)



3 mem-channels

3 mem-channels

https://en.wikichip.org/wiki/intel/microarchitectures/skylake_(server)

Resolution

Exercise 1 on mem hierarchy

- each clock cycle needs 2 mem accesses to fetch 2 doubles
- **max required bandwidth** to fetch a cache line for **each** double (cache is cold & doubles are far away):
  (16 cores x 2 lines x 64 B/line) x clock_frequency =
  2048 B x 2 GHz = **4096 GB/s**
  Note: the following 7 pairs of doubles are already in cache
- RAM in each Skylake Gold 6130: 6x DDR4-2666 (6x8 GiB)
- **peak bandwidth** of 6x DDR4-2666 in 6 memory channels:
  6 mem_chan x 2.666 GT/s x 64 b/chan = **128 GB/s**

Homework T2: ex 2 on mem hierarchy

Similar to problem 1 (same node/chip in the cluster), but consider now:

- execution of code taking advantage of the AVX-512 facilities;
- execution of the same 2 vector instructions (that are already in the instruction cache) in all cores: load in register a vector of doubles followed by a multiplication by another vector of doubles in memory;
- the Skylake cores are 6-way superscalar and 2-way MT, and each core supports 2 simultaneous vector loads;
- the Skylake 6130 base clock rate with **AVX-512** code is **1.3 GHz**;
- these instructions are executed with a cold data cache.

Compute/estimate:

- **The** max required bandwidth to access the external RAM when executing these 2 vector instructions.
  Compare with the peak bandwidth computed before.

* https://en.wikichip.org/wiki/intel/xeon_gold/6130

Resolution

Exercise 2 on memory hierarchy

- each clock cycle needs 2 mem accesses to fetch 2 vectors with 8 doubles each (512 bits)
- **max required bandwidth** to fetch a cache line for each vector with 8 doubles (cache is cold):

  ???? GB/s
- note: same max required bandwidth as exercise 1, but this **mem access is required at each clock cycle**
- RAM in each Skylake Gold 6130: 6x DDR4-2666 (6x8 GiB)
- **peak bandwidth** of 6x DDR4-2666 in 6 memory channels: **??? GB/s**

Resolution

Exercise 2 on memory hierarchy

- each clock cycle needs 2 mem accesses to fetch 2 vectors with 8 doubles each (512 bits)
- **max required bandwidth** to fetch a cache line for each vector with 8 doubles (cache cold, AVX-512 clock rate lower*):
  (16 cores x 2 lines x 64 B/line) x clock_rate =
  2048 B x **1.3** GHz = **2662.4 GB/s**
- note: same max required bandwidth as exercise 1,
  but this **mem access is required at each clock cycle**
- RAM in each Skylake Gold 6130: 6x DDR4-2666 (6x8 GiB)
- **peak bandwidth** of 6x DDR4-2666 in 6 memory channels:
  6 mem_chan x 2.666 GT/s x 64 b/chan = **128 GB/s**

* https://en.wikichip.org/wiki/intel/xeon_gold/6130

Homework T2: ex 3 on cache performance

- Given
  - I-cache miss rate = 2%
  - D-cache miss rate = 4%
  - Miss penalty = 100 cycles
  - Base CPI (ideal cache) = 2
  - Load & stores are 36% of instructions
- Miss cycles per instruction
  - I-cache: ?? x ?? = ??
  - D-cache: ?? x ?? x ?? = ??
- Actual CPI = 2 + ?? + ?? = **??**

Resolution

Exercise 3 on cache performance

- Given
  - I-cache miss rate = 2%
  - D-cache miss rate = 4%
  - Miss penalty = 100 cycles
  - Base CPI (ideal cache) = **2**
  - Load & stores are 36% of instructions
- Miss cycles per instruction
  - I-cache: 0.02 x 100 = 2
  - D-cache: 0.36 x 0.04 x 100 = 1.44
- Actual CPI = 2 + 2 + 1.44 = **5.44**

Homework T2: ex 4 on multilevel cache

- Given
  - CPU base **CPI = 1**, clock rate = 4GHz
  - Miss rate/instruction = 2%
  - Main memory access time = 100ns
- With just primary cache
  - Miss penalty = 100ns/0.25ns = 400 cycles
  - Effective **CPI = 9** (= 1 + 0.02 × 400)
- Now add L-2 cache …
  - Access time = 5ns
  - Global miss rate to main memory = 0.5%
- CPI = 1 + ?? × ?? + ?? × ?? = **??**
- Performance ratio = 9 / ?? = **??**

Resolution

Exercise 4 on multilevel cache

- **CPU**: base CPI = 1, clock rate = 4GHz
- **L1 cache**: L1 miss rate/instruction = 2%
- **L2 cache**: access time = **5ns**, L2 miss rate/instruction = 25%,
  global miss rate = 2% x 25% = **0.5%**
- **Main memory**: access time = **100ns**

- **With just primary cache**
  - Miss penalty = 100ns / 0.25ns = 400 cycles
  - Effective CPI = 1 + 0.02 × 400 = **9**

- **With L1 & L2 cache**
  - L1 miss penalty, L2 hit = **?? cycles**
  - L2 miss penalty = **?? cycles**

- **CPI** = 1 + **2%** × **??** cycles + **0.5%** × **???** cycles = **???**
- **Performance ratio** = 9 / **???** = **???**

Resolution

Exercise 4 on multilevel cache

- **CPU**: base CPI = 1, clock rate = 4GHz
- **L1 cache**: L1 miss rate/instruction = 2%
- **L2 cache**: access time = **5ns**, L2 miss rate/instruction = 25%,
  global miss rate = 2% x 25% = **0.5%**
- **Main memory**: access time = **100ns**
- **With just primary cache**
  - Miss penalty = 100ns / 0.25ns = 400 cycles
  - Effective CPI = 1 + 0.02 × 400 = **9**
- **With L1 & L2 cache**
  - L1 miss penalty, L2 hit = 99.5% × 5ns / 0.25ns ≈ **20 cycles**
  - L2 miss penalty = 100ns / 0.25ns = **400 cycles**
- **CPI** = 1 + **2%** × **20** cycles + **0.5%** × **400** cycles = **3.4**
- **Performance ratio** = 9 / **3.4** = **2.6**

Homework T2: ex 5 on multilevel performance

Characterize the memory system of Xeon Skylake Gold 6130:

1. **L1 I-cache**
   - size **?** KiB/core, **?**-way set associative, **?** sets, line size **?** B, hit time **?** cycles, **?** B/cycle on transfer bandwidth L1 to the instruction fetch unit

   L1 D-cache
   - size **?** KiB/core, **?**-way set associative, **?** sets, line size **?** B, hit time **?** cycles, **?** B/cycle on load bandwidth L1 to load buffer unit
2. **L2 cache**
   - size **?** KiB/core, **?**-way set associative, **?** sets, line size **?** B, hit time **?** cycles, **?** B/cycle on load bandwidth L2 to L1
3. **L3 cache**
   - size **?** KiB/core, **?**-way set associative, **?** sets, line size **?** B, hit time **?** cycles, **?** B/cycle on load bandwidth L3 to L2
4. **DRAM**, DDR4-2666
   - up to **?** GT/s, bandwidth **?** GB/s per channel, **?** mem channels, aggregate bandwidth **?** GB/s , **?** B/cycle on peak load bandwidth DRAM to L3, NUMA-local latency **?** ns, NUMA-remote latency **?** ns

Resolution

Exercise 5 on multilevel performance

Characterize the memory system of Xeon Skylake Gold 6130:

1. **L1 I-cache**
   - size **32** KiB/core, **8**-way set associative, **64** sets, line size **64** B, hit time **?** cycles, **16** B/cycle on transfer bandwidth L1 to the instruction fetch unit

   L1 D-cache
   - size **32** KiB/core, **8**-way set associative, **64** sets, line size **64** B, hit time **4** cycles, **2x64** B/cycle on load bandwidth L1 to load buffer unit
2. **L2 cache**
   - size **1** MiB/core, **16**-way set associative, **1024** sets, line size **64** B, hit time **14** cycles, **64** B/cycle on load bandwidth L2 to L1
3. **L3 cache**
   - size **1.375** MiB/core, **11**-way set associative, **2048** sets, line size **64** B, hit time **50-70** cycles, **64** B/cycle on load bandwidth L3 to L2
4. **DRAM**, DDR4-2666
   - up to **2.666** GT/s, bandwidth **21.33** GB/s per channel, **6** mem channels, aggregate on peak load bandwidth DRAM to L3 **128** GB/s,, NUMA-local latency **80** ns, NUMA-remote latency **120-140** ns

Homework T2: ex 6 on multilevel performance

Similar to problem 1 (same node/chip in the cluster, code), but consider now:

- execution of scalar code in a **2 GHz** single-core (already in L1 I-cache);
- code already takes advantage of all data cache levels (**L1, L2 & L3**), where 50% of data is placed on the RAM modules in the memory channels of the other PU chip (**NUMA architecture**);
- remember: the Skylake cores are **6-way superscalar** and **2-way MT**, and each core supports **2 simultaneous loads**;
- **cache latency time on hit**: take the average of the specified values;
- **memory latency**: 80 nsec (**NUMA local**), 120 nsec (**NUMA remote**);
- **miss rate per instruction** :
  - at **L1: 2%;** at **L2: 50%;** at **L3: 80%** (these are not global values!).

Compute/estimate:

1. **The miss penalty** per instruction at each cache level.
2. **The average memory stall cycles** per instruction that degrades CPI.

Resolution

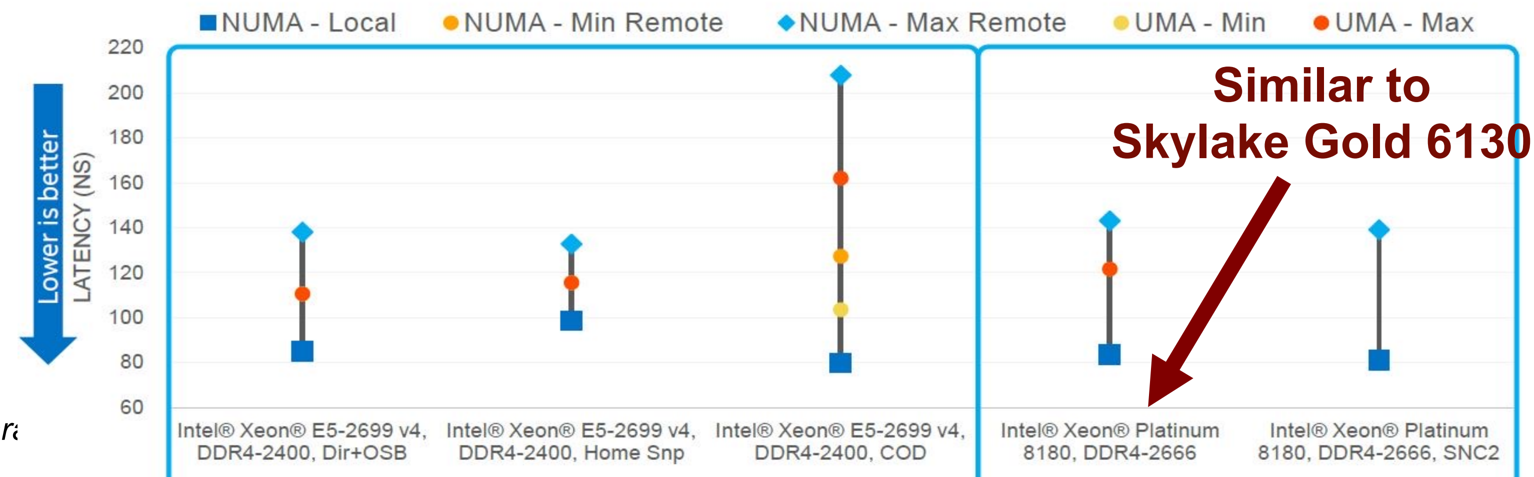
Exercise 6 on memory hierarchy

- **PU**: base CPI = 1, clock rate = 2 GHz
- **L1 cache**: L1 miss rate/instruction = 2%;
- **L2 cache**: access time = **14** cycles, global miss rate = 2% x 50% = **1%**
- **L3 cache**: access time = **60** cycles, L3 miss rate = 80%,
  global miss rate = 1% x 80% = **0.8%**
- **Main memory**: NUMA local access time = **80ns**, NUMA remote = **120ns**
  average memory access = ((80ns+120ns)/2) / 0.5ns = **200 cycles**

Memory Performance

Core to Memory Latency

- **CPI?**



Similar to Skylake Gold 6130