

Memory Hierarchy

CS 1541

Wonsun Ahn

Using the PMU to Understand Performance

Experiment on kernighan.cs.pitt.edu

- The source code for the experiments are available at:
https://github.com/wonsunahn/CS1541_Spring2022/tree/main/resources/cache_experiments
- Or on the following directory at linux.cs.pitt.edu:
/afs/cs.pitt.edu/courses/1541/cache_experiments/
- You can run the experiments by doing 'make' at the root
 - It will take a few minutes to run all the experiments
 - In the end, you get two plots: IPC.pdf and MemStalls.pdf

Four benchmarks

- `linked-list.c`
 - Traverses a linked list from beginning to end over and over again
 - Each node has 120 bytes of data
- `array.c`
 - Traverses an array from beginning to end over and over again
 - Each element has 120 bytes of data
- `linked-list_nodata.c`
 - Same as linked-list but nodes have no data inside them
- `array_nodata.c`
 - Same as array but elements have no data inside them

Code for linked-list.c

// Define a linked list node type with no data

```
typedef struct node {  
    struct node* next;    // 8 bytes  
    int data[30];         // 120 bytes  
} node_t;
```

...

// Create a linked list of length items

```
void *create(void *unused) {  
    for(int i=0; i<items; i++) {  
        node_t* n = (node_t*)malloc(sizeof(node_t));  
        if(last == NULL) { // Is the list empty? If so, the new node is the head and tail  
            head = n;  
            last = n;  
        } else {  
            last->next = n;  
            last = n;  
        }  
    }  
}
```

Code for linked-list.c

```
#define ACCESSES 1000000000
```

```
// MEASUREMENT BEGIN
```

```
// Traverse list over and over until we've visited `ACCESSES` nodes
```

```
node_t* current = head;
```

```
for(int i=0; i<ACCESSES; i++) {
```

```
    if(current == NULL) current = head;    // reached the end
```

```
    else current = current->next;          // next node
```

```
}
```

```
// MEASUREMENT END
```

- Note: executed instructions are equivalent regardless of list length
- So we expect performance to be same regardless of length. **Is it?**

Code for array.c

// Define a linked list node type with no data

```
typedef struct node {  
    struct node* next;    // 8 bytes  
    int data[30];         // 120 bytes  
} node_t;
```

...

// Create a linked list but **allocate nodes in an array**

```
void *create(void *unused) {  
    head = (node_t *) malloc(sizeof(node_t) * items);  
    last = head + items - 1;  
    for(int i=0; i<items; i++) {  
        node_t* n = &head[i];  
        n->next = &head[i+1]; // Next node is next element in array  
    }  
    last->next = NULL;  
}
```

Code for array.c

```
#define ACCESSES 1000000000
```

```
// MEASUREMENT BEGIN
```

```
// Traverse list over and over until we've visited `ACCESSES` nodes
```

```
node_t* current = head;
```

```
for(int i=0; i<ACCESSES; i++) {
```

```
    if(current == NULL) current = head;    // reached the end
```

```
    else current = current->next;          // next node
```

```
}
```

```
// MEASUREMENT END
```

- Note: same exact loop as the linked-list.c loop.
- So we expect performance to be exactly the same. **Is it?**

Code for array.c

```
#define ACCESSES 1000000000
```

```
// Define a linked list node type with no data
```

```
typedef struct node {  
    struct node* next; // 8 bytes  
    int data[30];      // 120 bytes  
} node_t;
```

```
...
```

```
// Traverse array over and over again until we've visited `ACCESSES` elements
```

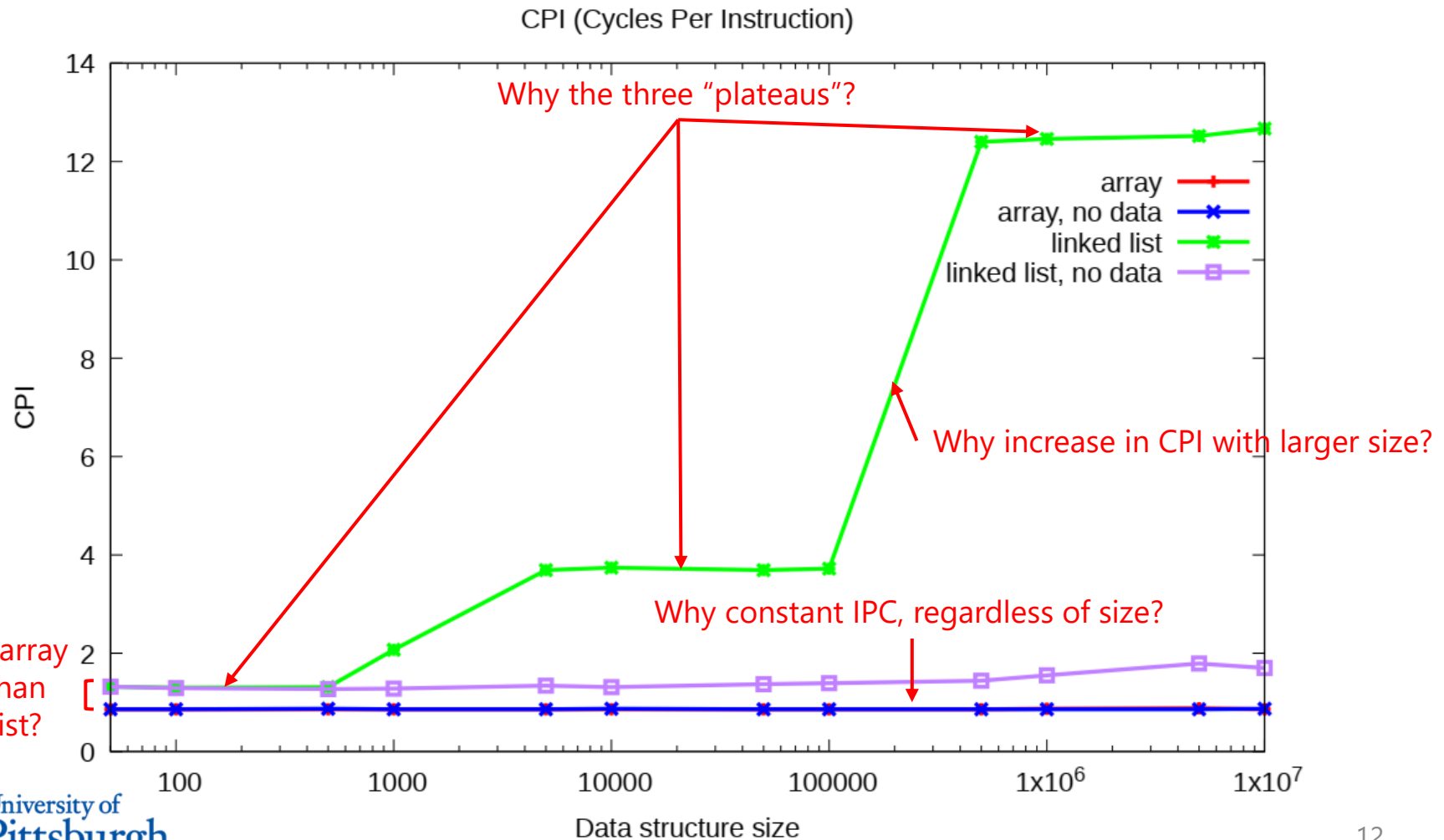
```
node_t* current = array;  
for(int i=0; i<ACCESSES; i++) {  
    if(current == array + items) current = array; // reached the end  
    else ++current; // next element  
}
```

- Two CPU sockets. Each CPU:
 - Intel(R) Xeon(R) CPU E5-2640 v4
 - 10 cores, with 2 threads per each core (SMT)
 - L1 i-cache: 32 KB 8-way set associative (per core)
 - L1 d-cache: 32 KB 8-way set associative (per core)
 - L2 cache: 256 KB 8-way set associative (per core)
 - L3 cache: 25 MB 20-way set associative (shared)
- Memory
 - 128 GB DRAM
- Information obtained from
 - "cat /proc/cpuinfo" on Linux server
 - "cat /proc/meminfo" on Linux server
 - https://en.wikichip.org/wiki/intel/xeon_e5/e5-2640_v4

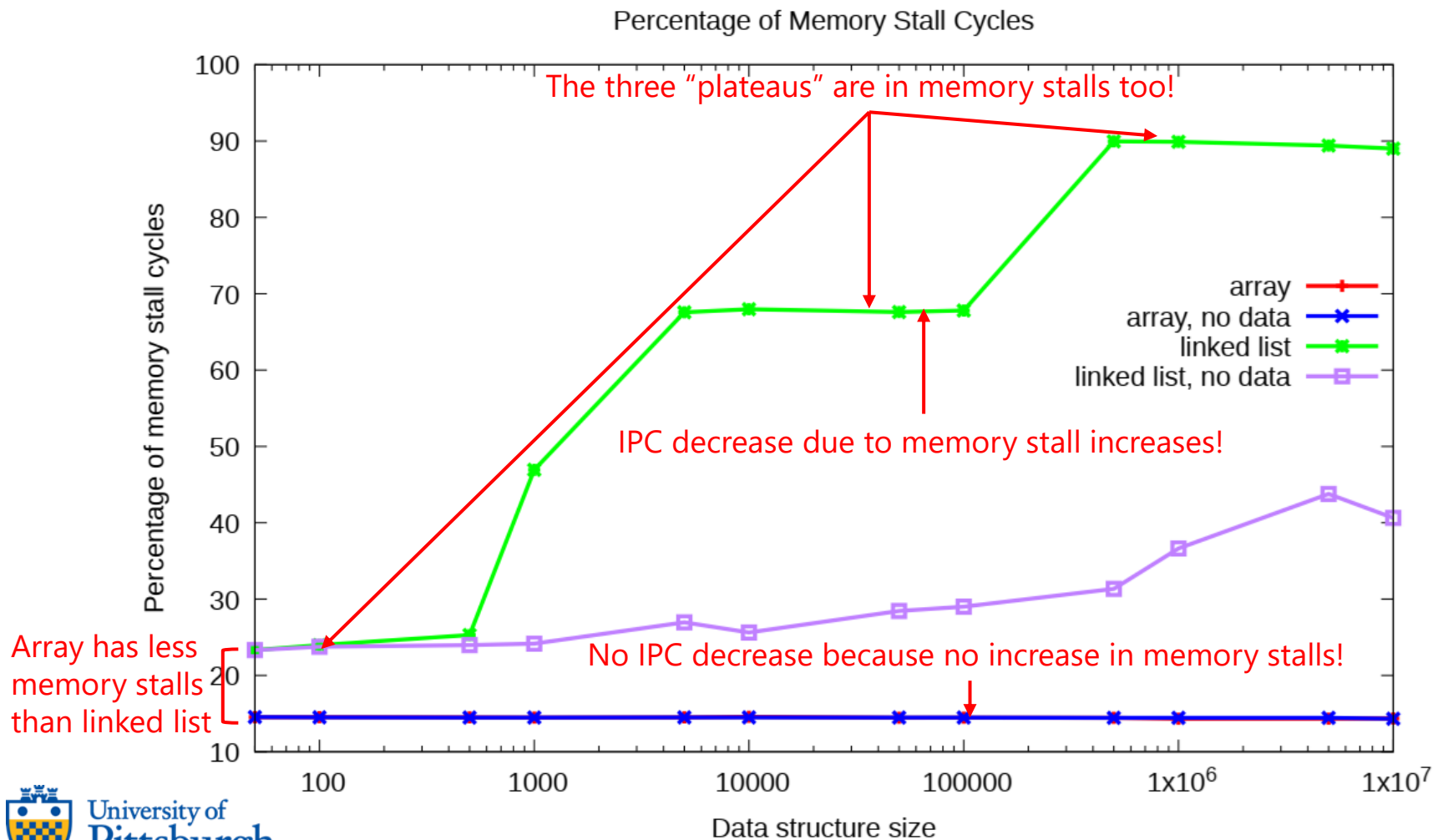
Experimental data collection

- Collected using CPU **Performance Monitoring Unit (PMU)**
 - PMU provides performance counters for a lot of things
 - Cycles, instructions, various types of stalls, branch mispredictions, cache misses, bandwidth usage, ...
- Linux **perf** utility summarizes this info in easy to read format
 - <https://perf.wiki.kernel.org/index.php/Tutorial>

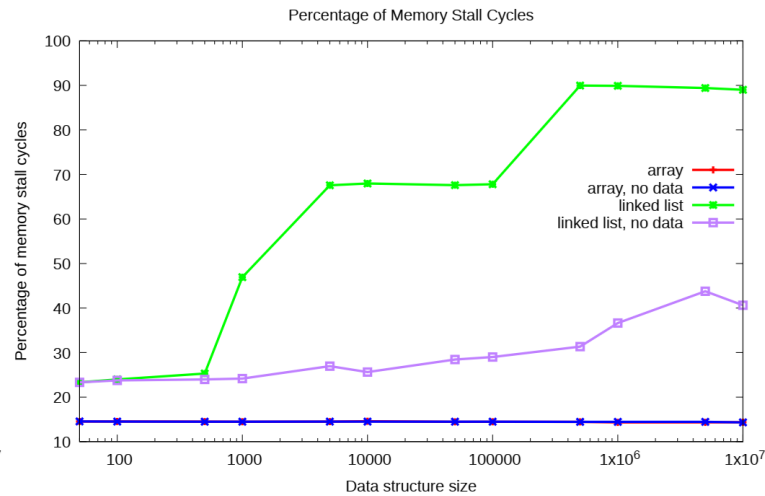
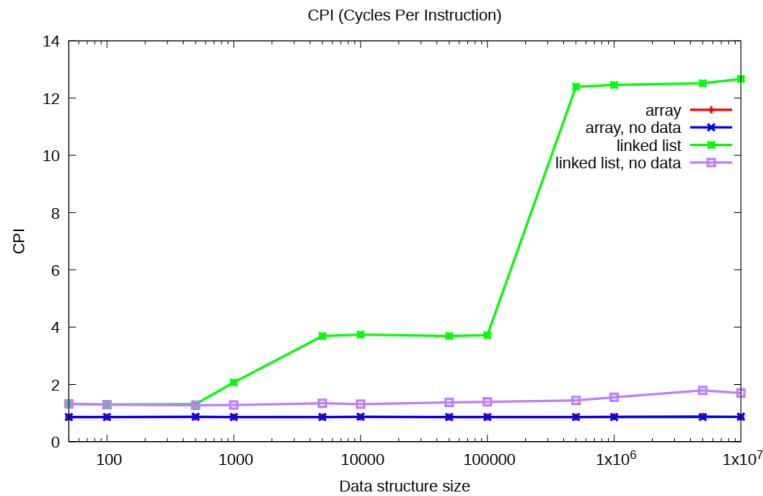
CPI (Cycles Per Instruction) Results



Memory Stall Cycle Percentage



Data Structure Performance \propto Memory Stalls

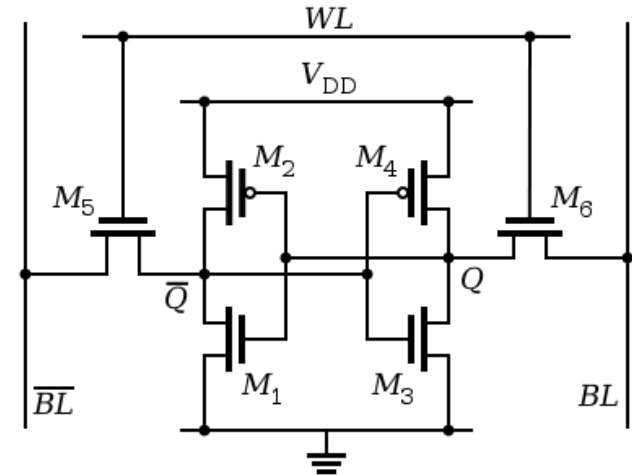
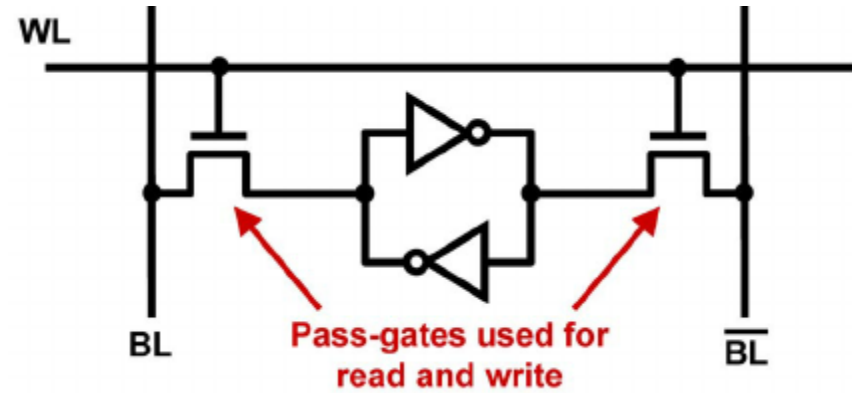


- **Data structure performance is proportional to memory stalls**
 - Applies to other data structures such as trees, graphs, ...
- In general, **more data** leads to **worse performance**
 - But why? Does more data make MEM stalls longer? (Hint: yes)
 - And why is an array not affected by data size? (I wonder ...)
- You will be able to answer all these questions when we are done.

Memory Technologies

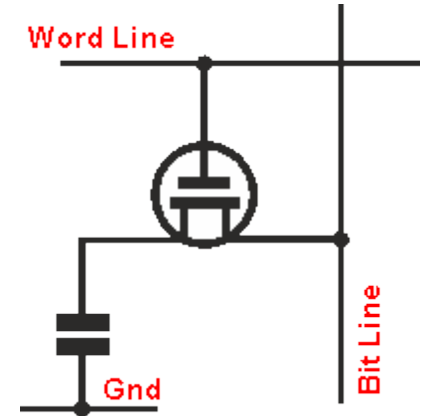
Static RAM (SRAM)

- SRAM uses a loop of NOT gates to store a single bit
- This is usually called a 6T SRAM cell since it uses... 6 Transistors!
- **Pros:**
 - **Very fast** to read/write
- **Cons:**
 - Volatile (loses data without power)
 - Relatively many transistors needed
 - > **expensive**



Dynamic RAM (DRAM)

- DRAM uses **one** transistor and **one** capacitor
 - The bit is stored as a charge in the capacitor
 - Capacitor leaks charge over time
 - > Must be periodically recharged (called **refresh**)
 - > During refresh, DRAM can't be accessed
 - Accesses are slower
 - > Small charge must be amplified to be read
 - > Also after read, capacitor needs recharging again
 - Reading a DRAM cell is slower than reading SRAM
- **Pros:**
 - Higher density -> less silicon -> **much cheaper than SRAM**
- **Cons:**
 - Still volatile (even more volatile than SRAM)
 - **Slower access time**



Spinning magnetic disks (HDD)

- Spinning platter coated with a ferromagnetic substance magnetized to represent bits
 - Has a mechanical arm with a head
 - Reads by placing arm in correct cylinder, and waiting for platter to rotate
- **Pros:**
 - **Nonvolatile** (magnetization persists without power)
 - Extremely cheap (1TB for \$50)
- **Cons:**
 - **Extremely slow** (it has a mechanical arm, enough said)



Other technology

- Flash Memory
 - Works using a special MOSFET with “floating gate”
 - **Pros:** nonvolatile, much faster than HDD
 - **Cons:**
 - Slower than DRAM
 - More expensive than HDDs (1TB for \$250)
 - Writing is destructive and shortens lifespan
- Experimental technology
 - Ferroelectric RAM (FeRAM), Magnetoresistive RAM (MRAM), Phase-change memory (PRAM), carbon nanotubes ...
 - In varying states of development and maturity
 - Nonvolatile *and* close to DRAM speeds



Memory/storage technologies

	Volatile		Nonvolatile	
	SRAM	DRAM	HDDs	Flash
Speed	FAST	OK	SLOW	Pretty good!
Price	Expensive	OK	Cheap!	Meh
Power	Good!	Meh	Bad	OK
Durability	Good!	Good!	Good!	OK
Reliability	Good!	Pretty good!	Meh	Pretty good!

I'm using **Durability** to mean "how well it **holds data** after repeated use."

I'm using **Reliability** to mean "how resistant is it to external shock."

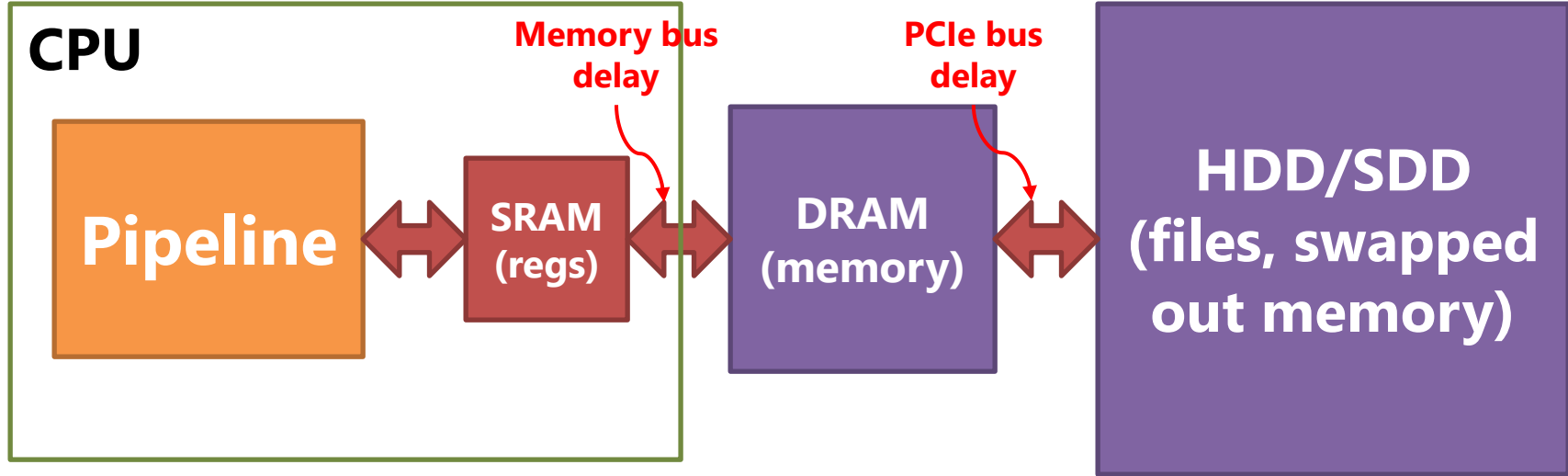
Do you notice a trend?

- The **faster** the memory the more **expensive** and **lower density**.
- The **slower** the memory the **less expensive** and **higher density**.
- There exists a hierarchy in program data:
 - Small set of data that is accessed very frequently
 - Large set of data that is accessed very infrequently
- Thus, memory should also be constructed as a hierarchy:
 - Fast and small memory at the upper levels
 - Slow and big memory at the lower levels

The Memory Hierarchy

System Memory Hierarchy

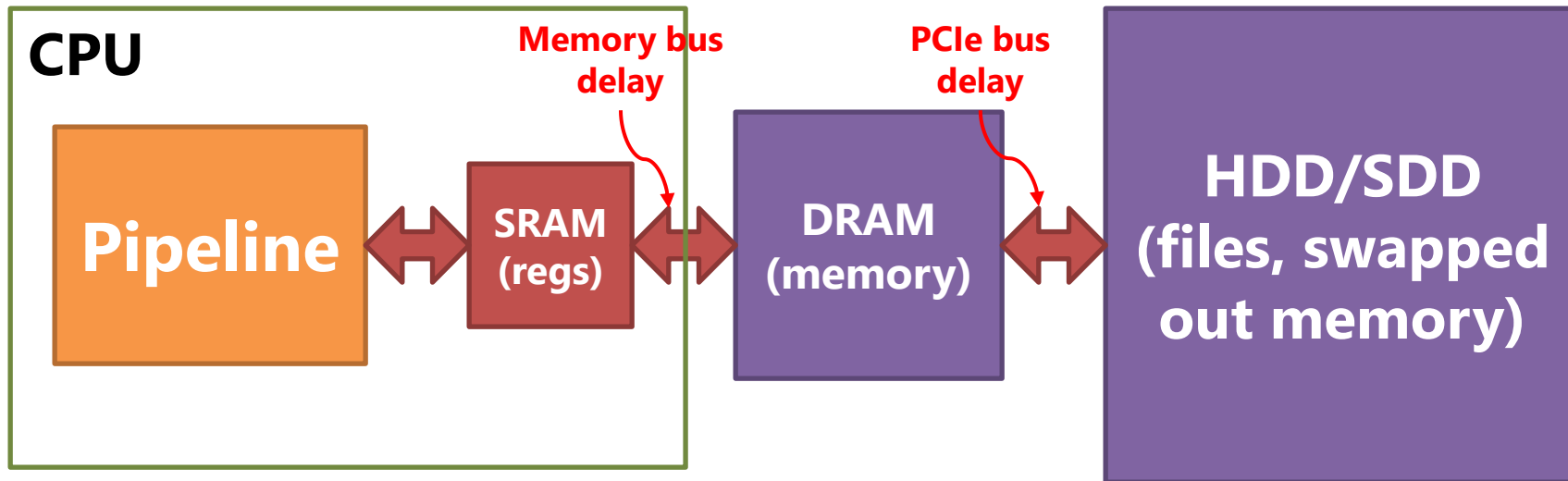
- Use fast memory (SRAM) to store frequently used data inside the CPU
- Use slow memory (e.g. DRAM) to store rest of the data outside the CPU



- Registers are used frequently for computation so are stored in SRAM
- Memory pages used frequently are stored in DRAM
- Memory pages used infrequently are stored in HDD/SDD (in swap space)
- Note: Memories outside CPU suffers from bus delay as well

System Memory Hierarchy

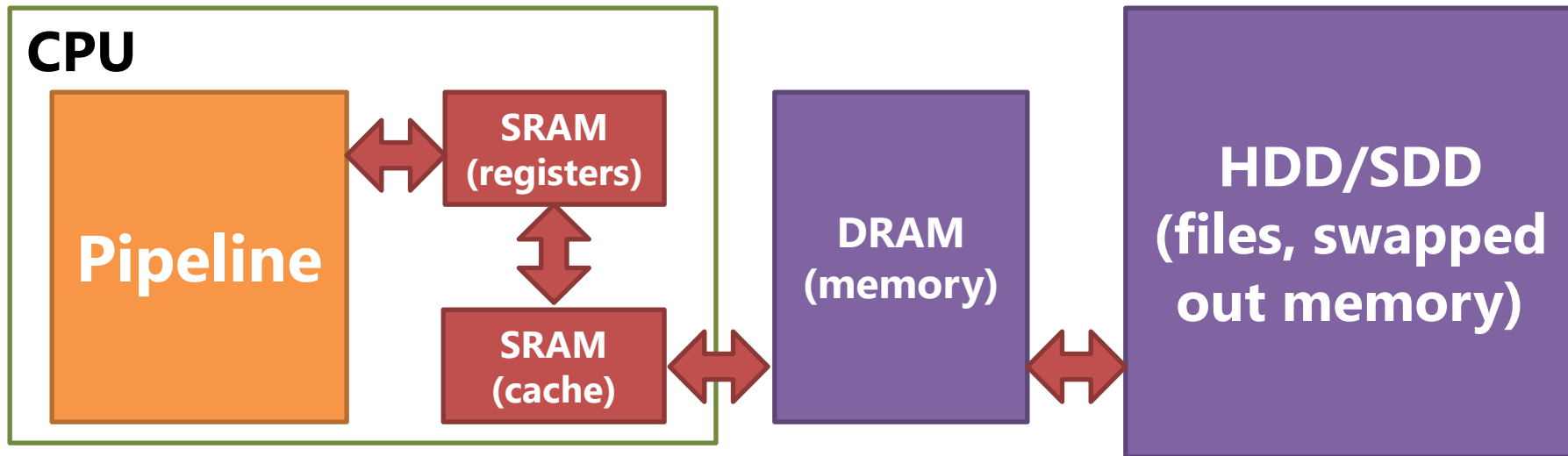
- Use fast memory (SRAM) to store frequently used data inside the CPU
- Use slow memory (e.g. DRAM) to store rest of the data outside the CPU



- Drawback: Memory access is much slower compared to registers
- Q: Can we make memory access speed comparable to register access?

System Memory Hierarchy

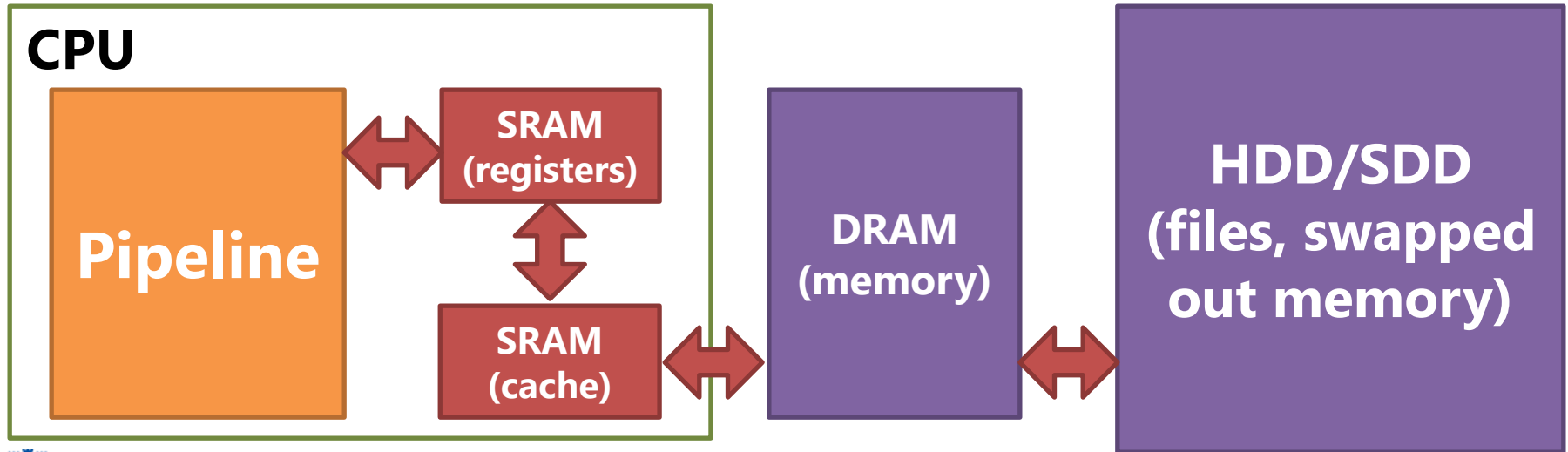
- Use fast memory (SRAM) to store frequently used data inside the CPU
- Use slow memory (e.g. DRAM) to store rest of the data outside the CPU



- Drawback: Memory access is much slower compared to registers
- Q: Can we make memory access speed comparable to register access?
 - How about storing frequently used memory data in SRAM too?
 - This is called **caching**. The hardware structure is called a **cache**.

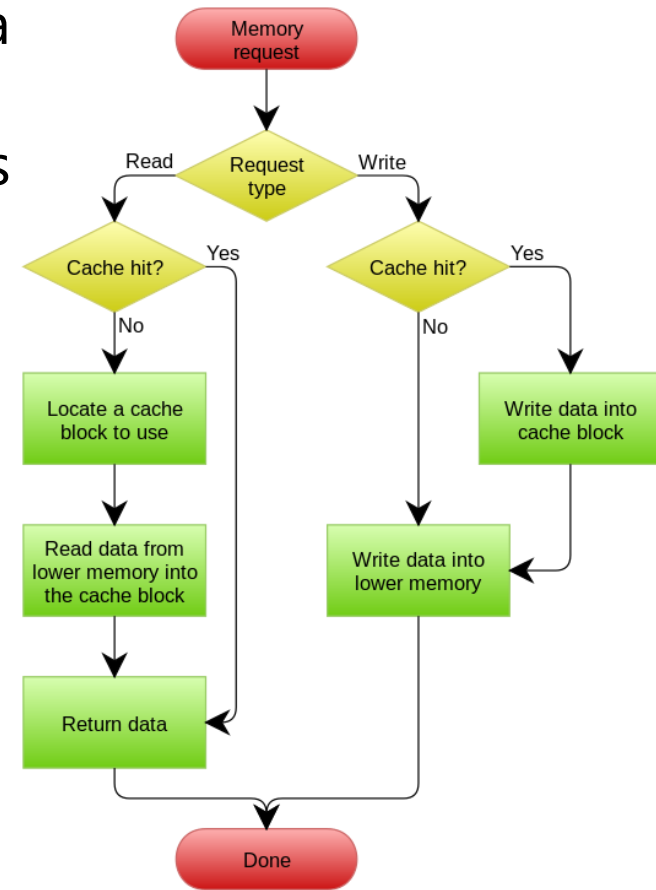
Caching

- **Caching**: keeping a temporary copy of data for faster access
- DRAM is in a sense also caching frequently used pages from swap space
 - We are just extending that idea to bring cache data inside the CPU!
- Now instructions like **lw** or **sw** never directly access DRAM
 - They first search the cache to see if there is a **hit** in the cache
 - Only if they **miss** will they access DRAM to bring data into the cache



Cache Flow Chart

- **Cache block:** unit of data used to cache data
 - What **page** is to memory paging
 - Cache block size is typically multiple words (e.g. 32 bytes or 64 bytes. You'll see why.)
- **Good: Memory Wall** can be **surmounted**
 - On cache hit, no need to go to DRAM!
- **Bad:** MEM stage has **variable latency**
 - Typically only a few cycles if cache hit
 - More than a 100 cycles if cache miss!
(Processor must go all the way to DRAM!)
 - Makes performance very **unpredictable**



Cache Locality: Temporal and Spatial

- Caching works because there is **locality** in program data accesses
 - **Temporal locality**
 - Same data is accessed many times in succession
 - **1st access** will **miss** but following accesses will **hit** in the cache
 - **Spatial locality**
 - Many data items that are spatially close are accessed together
 - E.g. fields of the same object, elements in an array, ...
 - Access to **1st item** will **miss** but bring in an entire **cache block**
 - Accesses to **other items** within same cache block will **hit**
 - Locality, like ILP, can only be exploited if it is there
- So does that mean having larger cache blocks is always better?
 - No, if block size is larger than spatial locality present in program
 - Then the space to bring in the extra unused data is wasted

Cold Misses and Capacity Misses

- **Cold miss** (a.k.a. **compulsory miss**)
 - Miss suffered when data is accessed for the **first time** by program
 - Cold miss since cache **hasn't been warmed up** with accesses
 - Compulsory miss since there is no way you can hit on the first access
 - Subsequent accesses will be hits since now data is fetched into cache
 - Unless it is replaced to make space for more frequently used data
- **Capacity miss**
 - Miss suffered when data is accessed for the **second or third times**
 - This miss occurred because **data was replaced** to make space
 - If there had been more capacity, miss wouldn't have happened
 - Capacity decides how much **temporal locality** you can leverage

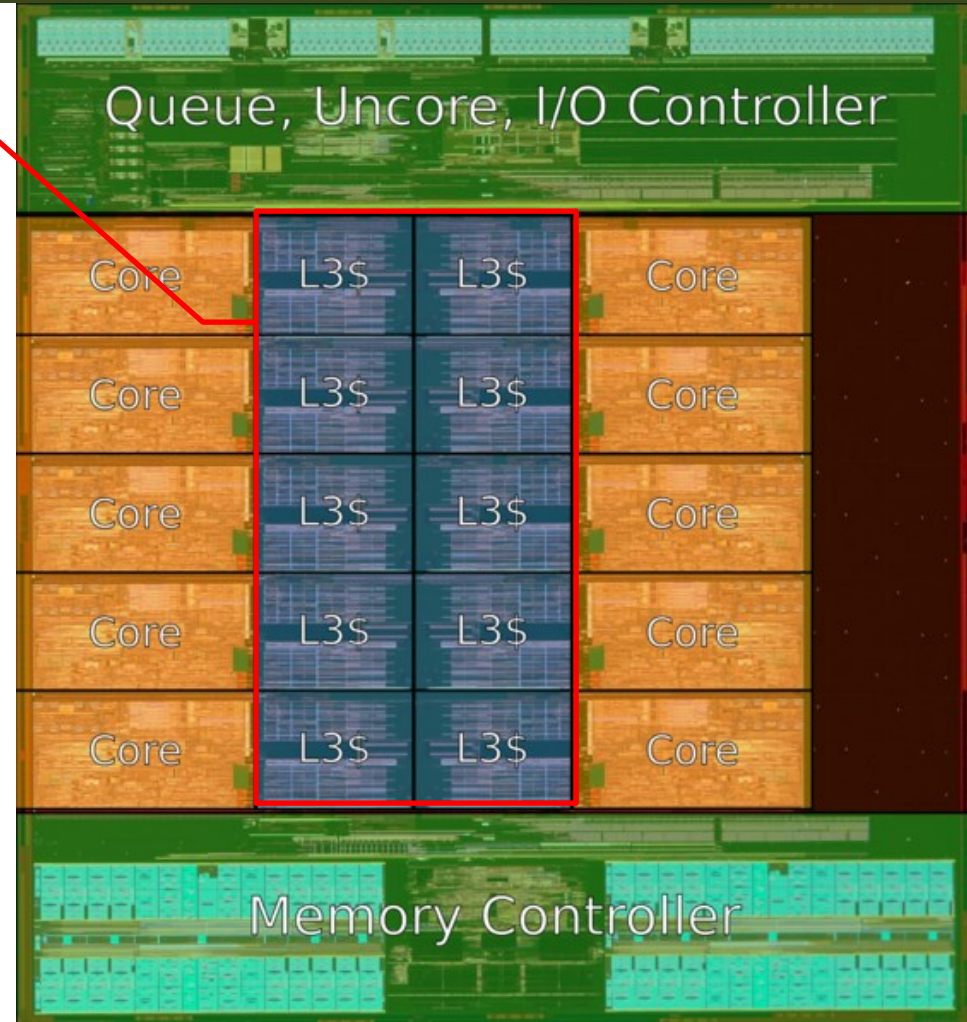
Reducing Cold Misses and Capacity Misses

- Reducing capacity misses is straightforward
 - Increase capacity, that is cache size!
 - But how do you reduce cold misses? Is it even possible?
 - Yes! By taking advantage of **spatial locality**.
 - Have a large cache block so you bring in other items on a miss
 - Those other items may be accessed for the first time but will hit!
 - **Large cache blocks** can ...
 - Potentially **increase capacity misses**
(with no spatial locality, can store less data items in same capacity)
 - Potentially **reduce cold misses**
(given some spatial locality, can bring in more data on a miss)
- Each program has a sweet spot. Architects choose a compromise.

So how big do we want the cache to be?

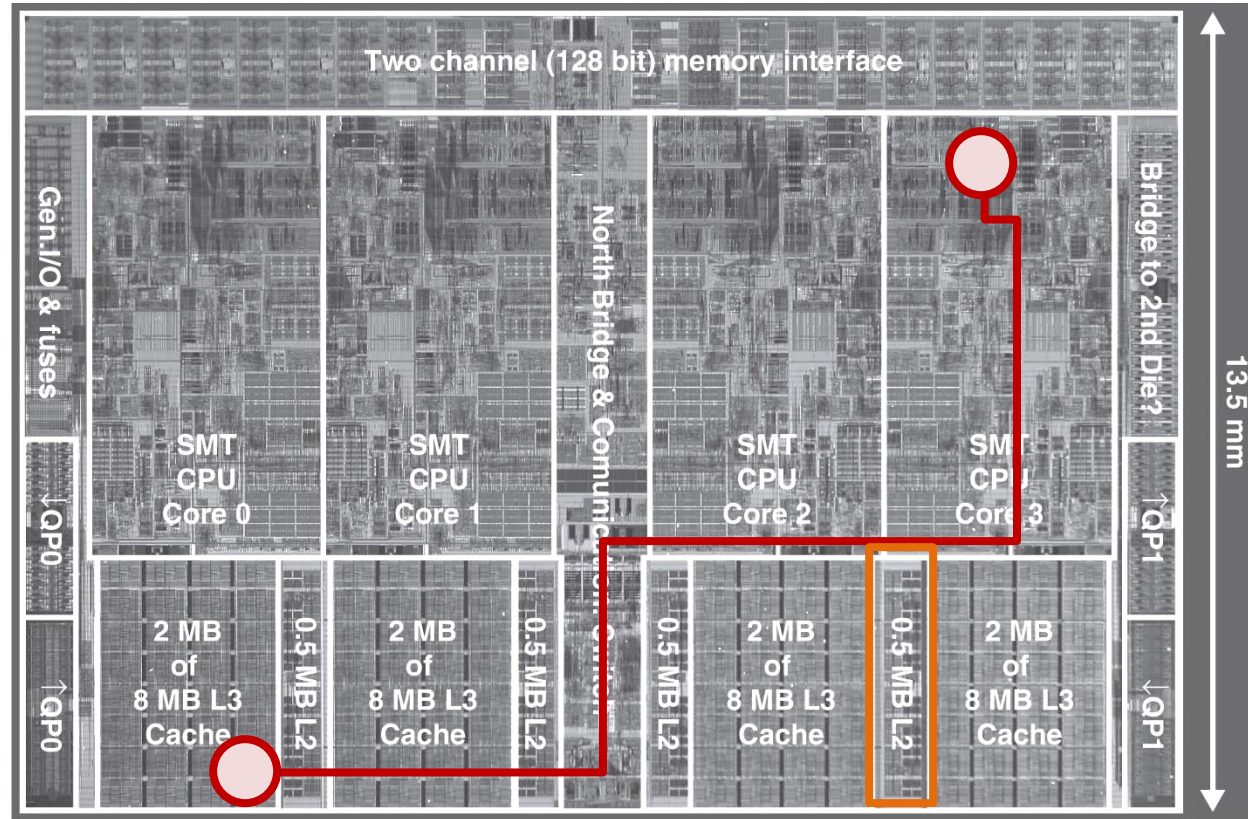
- Uber big!
- On the right is a diagram of the Xeon Broadwell CPU used in kernighan.cs.pitt.edu.
 - Caches take up almost as much real estate as cores!
 - A cache miss is that painful.
- But having a big cache comes with its own set of problems
 - Cache itself gets slower

Caches



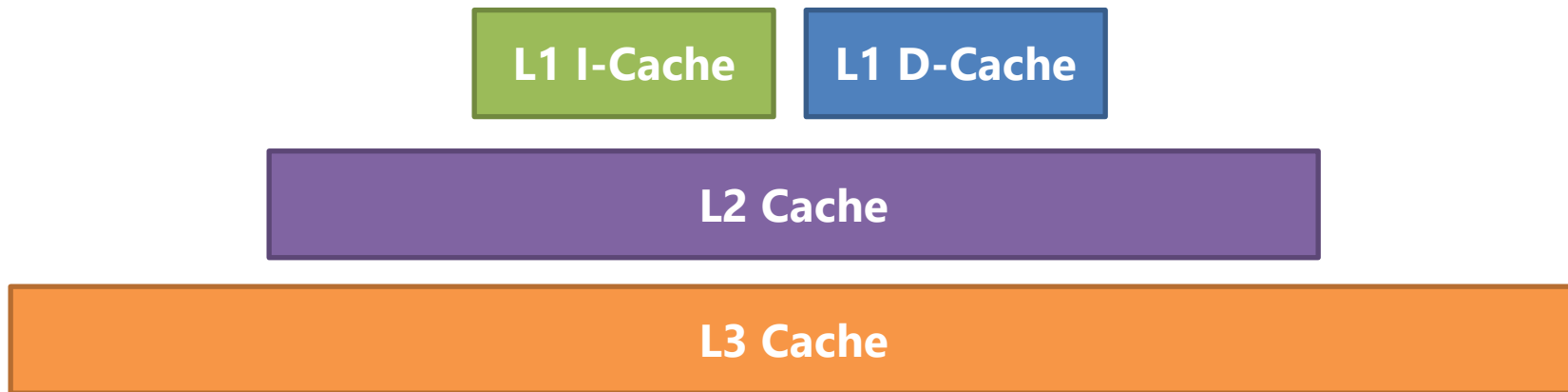
Bigger caches are slower

- Below is a diagram of a Nehalem CPU (an older Intel CPU)
- How long do you think it takes for data to make it from here...
- ...to here?
- It must be routed through all this.
- Can we cache the data in the far away "L3 Cache" to a nearby "L2 Cache"?



Multi-level Caching

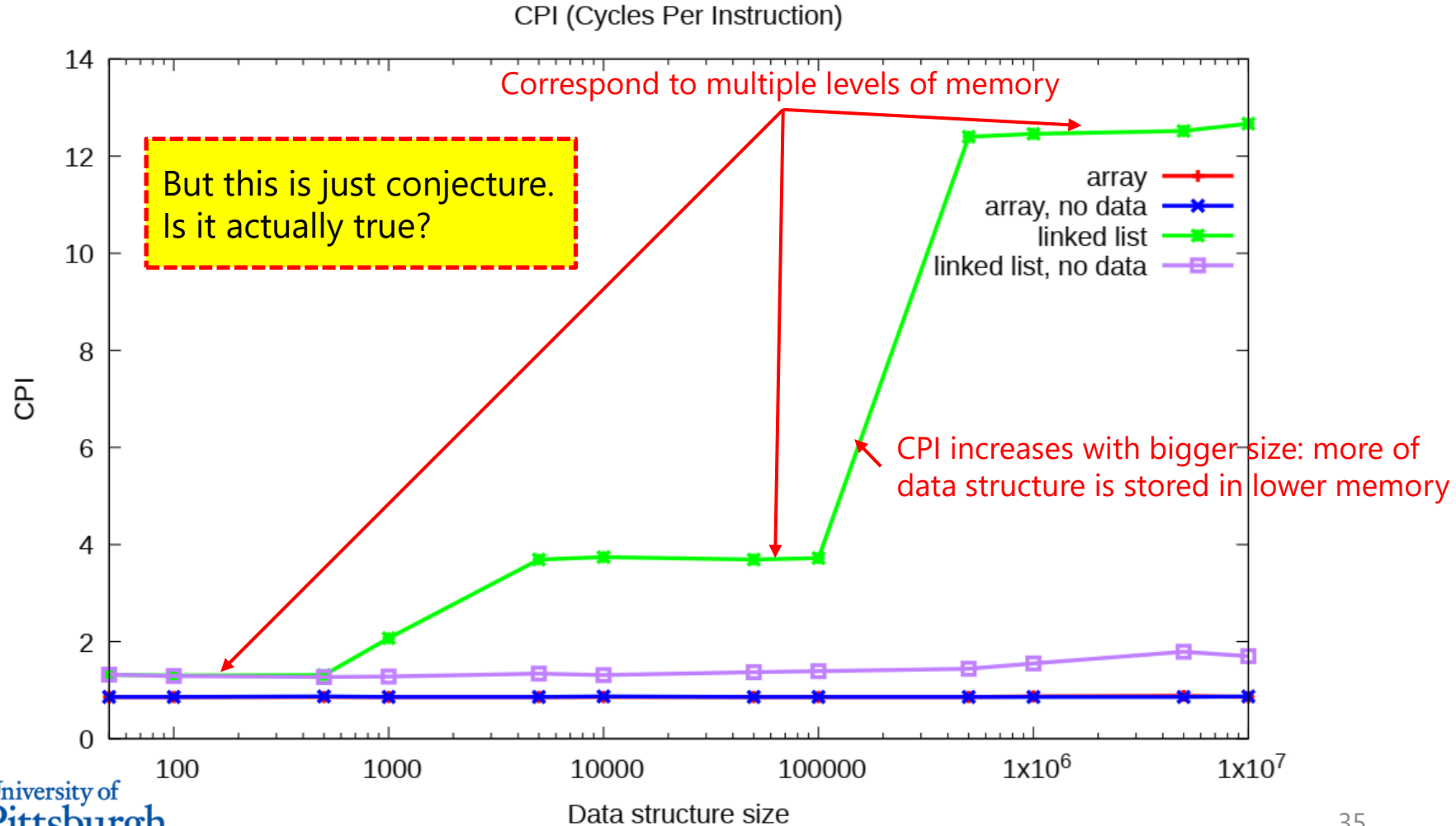
- This is the structure of the kernighan.cs.pitt.edu Xeon CPU:



- L1 cache: Small but fast. Interfaces with CPU pipeline MEM stage.
 - Split to i-cache and d-cache to avoid structural hazard
- L2 cache: Middle-sized and middle-fast. Intermediate level.
- L3 cache: Big but slow. Last line of defense against memory access.
- Allows performance to degrade gracefully

Revisiting Our Experiments

Revisiting our CPI Results with the new perspective



kernighan.cs.pitt.edu cache specs

- On a Xeon E5-2640 v4 CPU (10 cores):
 - L1 i-cache: **32 KB** 8-way set associative (per core)
 - L1 d-cache: **32 KB** 8-way set associative (per core)
 - L2 cache: **256 KB** 8-way set associative (per core)
 - L3 cache: **25 MB** 20-way set associative (shared)Ref: https://en.wikichip.org/wiki/intel/xeon_e5/e5-2640_v4
- Access latencies (each level includes latency of previous levels):
 - L1: ~**3 cycles**
 - L2: ~**8 cycles**
 - L3: ~**16 cycles**
 - DRAM Memory: ~**67 cycles**Ref: https://www.nas.nasa.gov/assets/pdf/papers/NAS_Technical_Report_NAS-2015-05.pdf

Cache Specs Reverse Engineering

- Why do I have to refer to a NASA technical report for latencies?
 - Ref: https://www.nas.nasa.gov/assets/pdf/papers/NAS_Technical_Report_NAS-2015-05.pdf
 - Because Intel doesn't publish detailed cache specs on data sheet
- In the technical report (does the step function look familiar?):

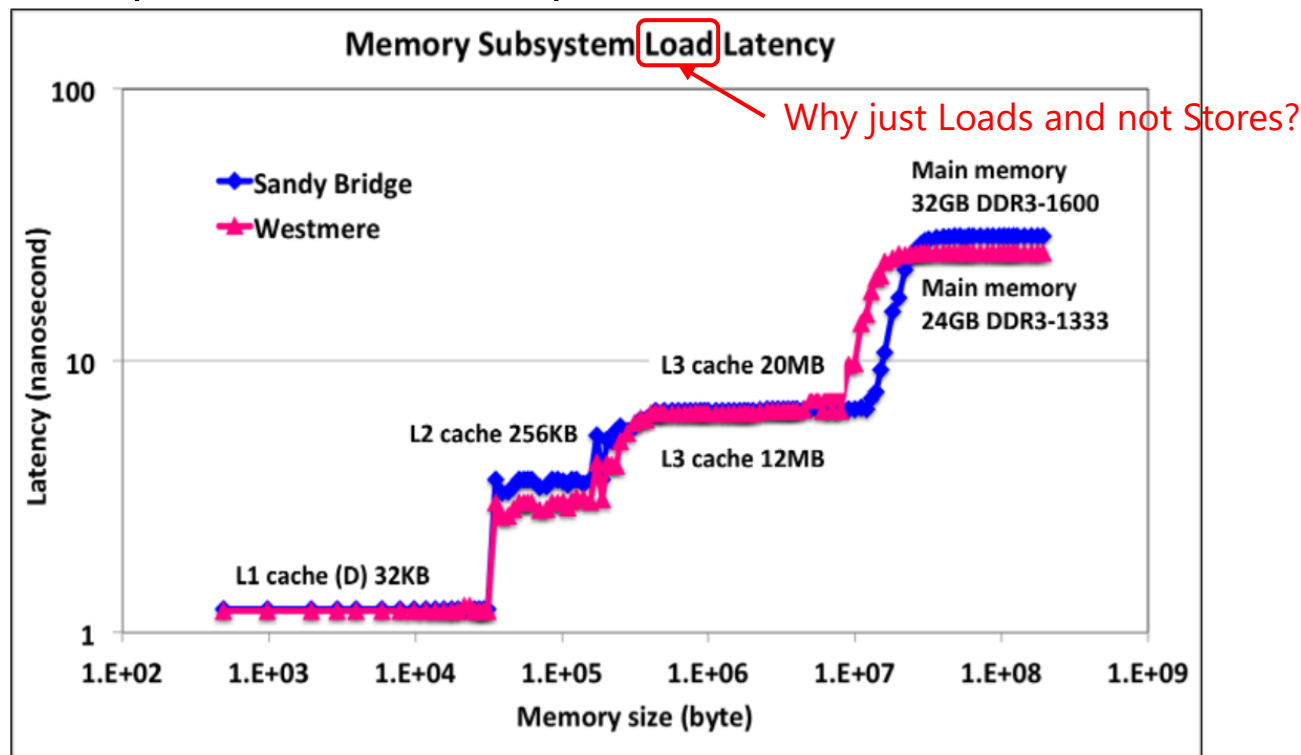


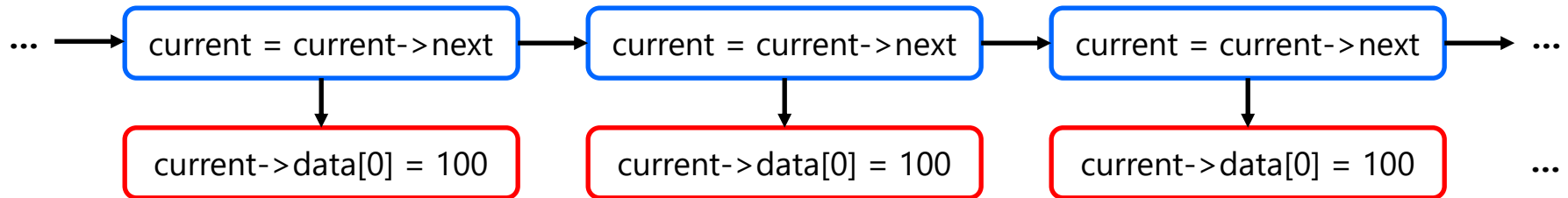
Figure 2. Memory latency of Westmere and Sandy Bridge.

Loads have more impact on performance

- Suppose we added a store to our original loop:

```
node_t* current = head;
for(int i=0; i<ACCESSES; i++) {
    if(current == NULL) current = head;           // reached the end
    else {
        current->data[0] = 100;                  // store to node data
        current = current->next;                 // load next node address
    }
}
```

- Which would have more impact on performance? The load or the store?
 - A: The load because it is on the **critical path**.

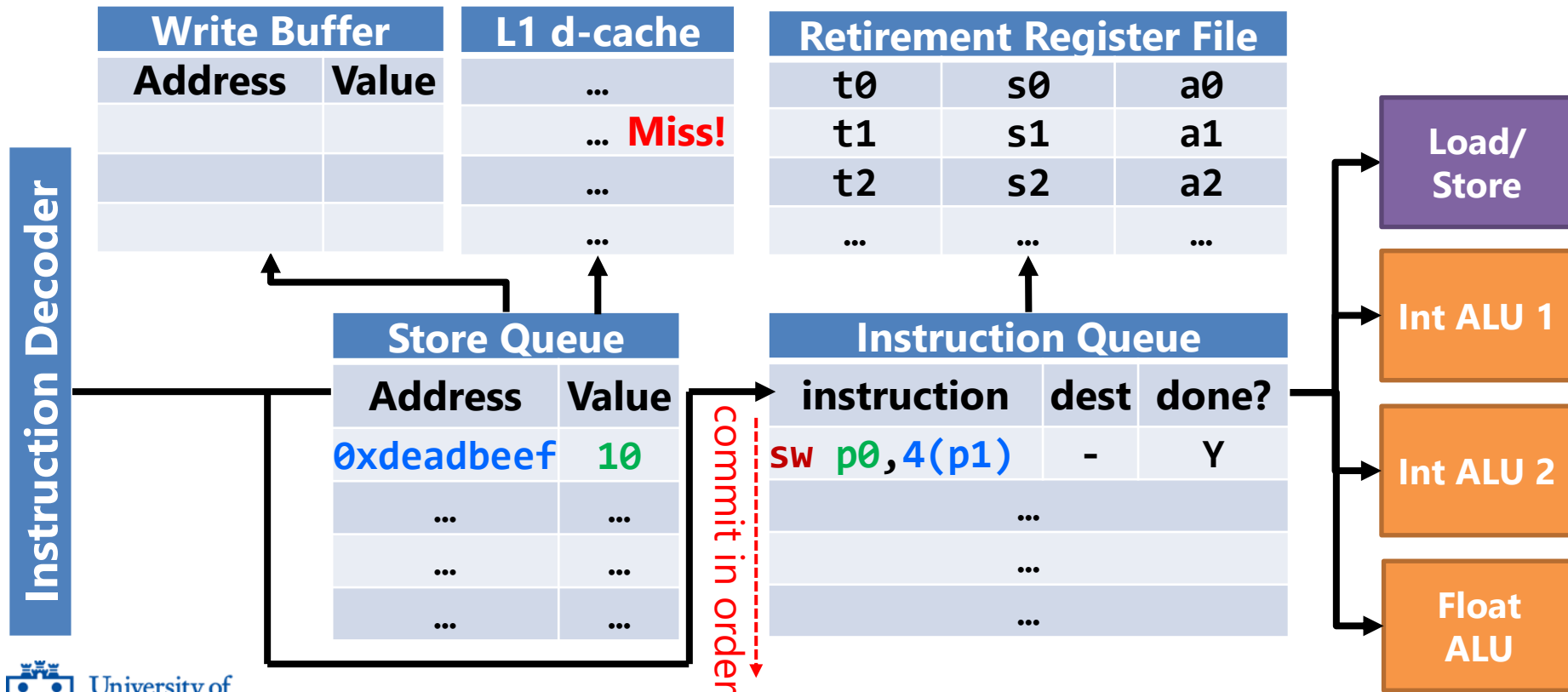


Loads have more impact on performance

- Loads produce values needed for computation to proceed
 - **Stalled loads delays computation** and possibly the critical path
- Stores write computation results back to memory
 - As long as the results are written back eventually, no big hurry
 - If store results in a cache miss,
store is marked as “pending” and CPU moves on to next computation
 - Pending stores are maintained in a **write buffer** hardware structure
- What if the next computation reads from a pending store?
 - First check the write buffer and read in the value if it's there
 - Again, performing the store is not on the critical path

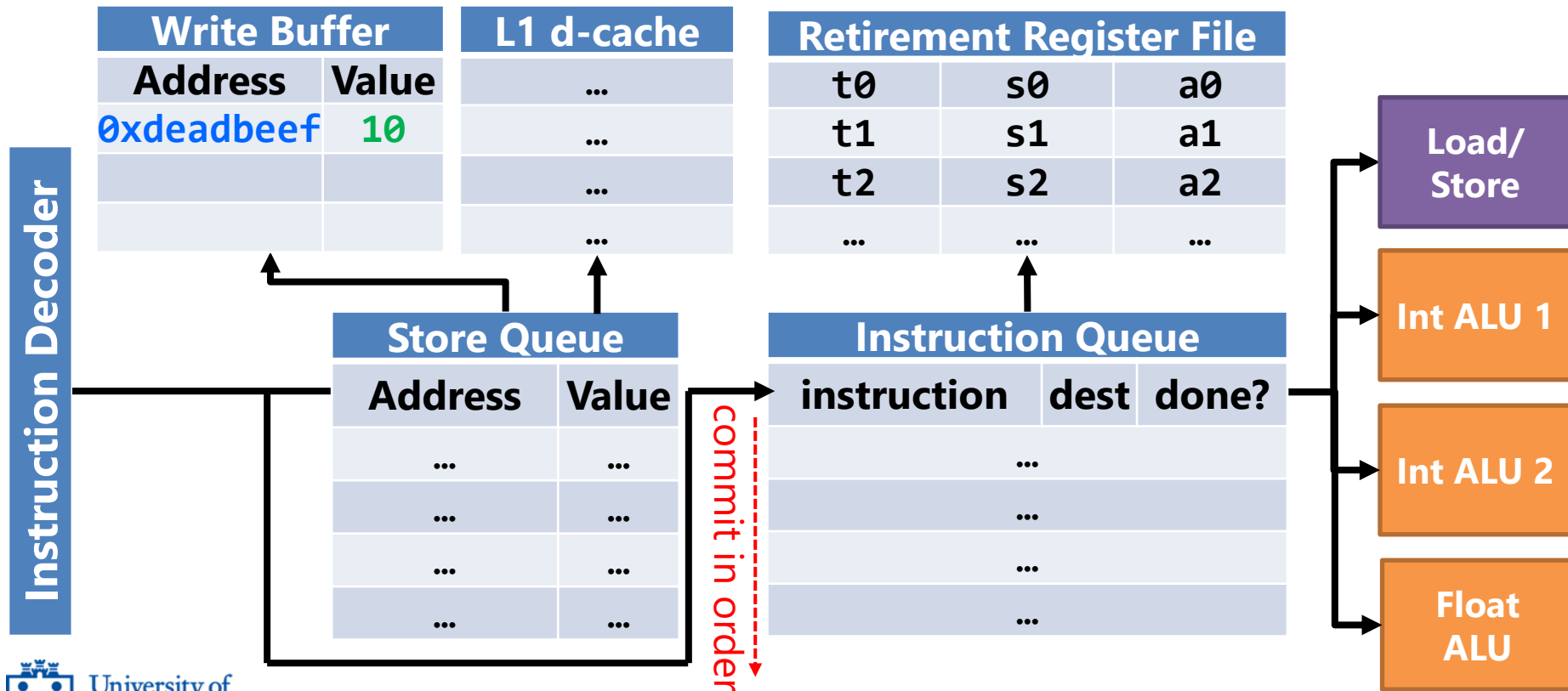
How Write Buffer Maintains Pending Stores

- **sw** p0,4(p1) is about to commit. 4(p1) == 0xdeadbeef, p0 == 10
- Unfortunately, address 0xdeadbeef is not in the L1 d-cache and it misses



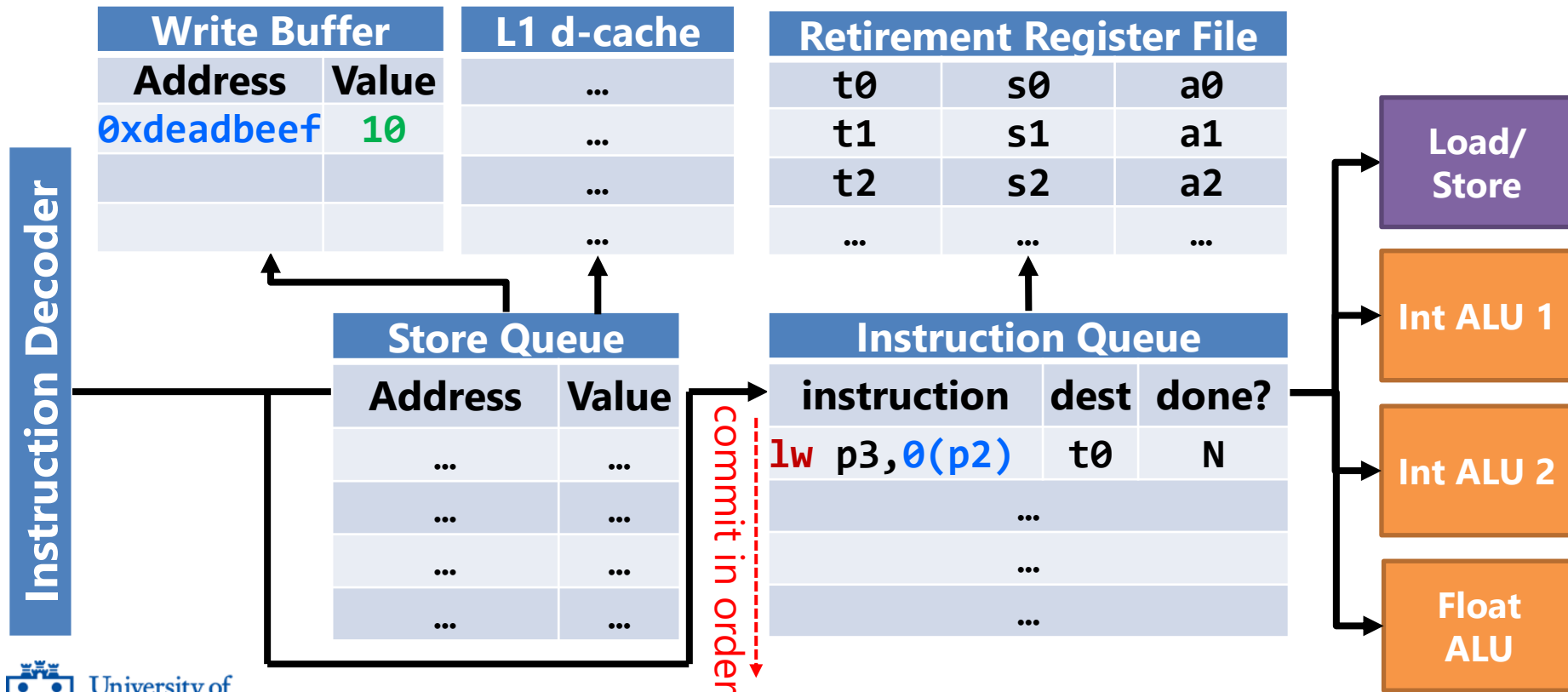
How Write Buffer Maintains Pending Stores

- **sw p0, 4(p1)** commits successfully anyway
- The store is moved to the Write Buffer and stays there until store completes



How Write Buffer Maintains Pending Stores

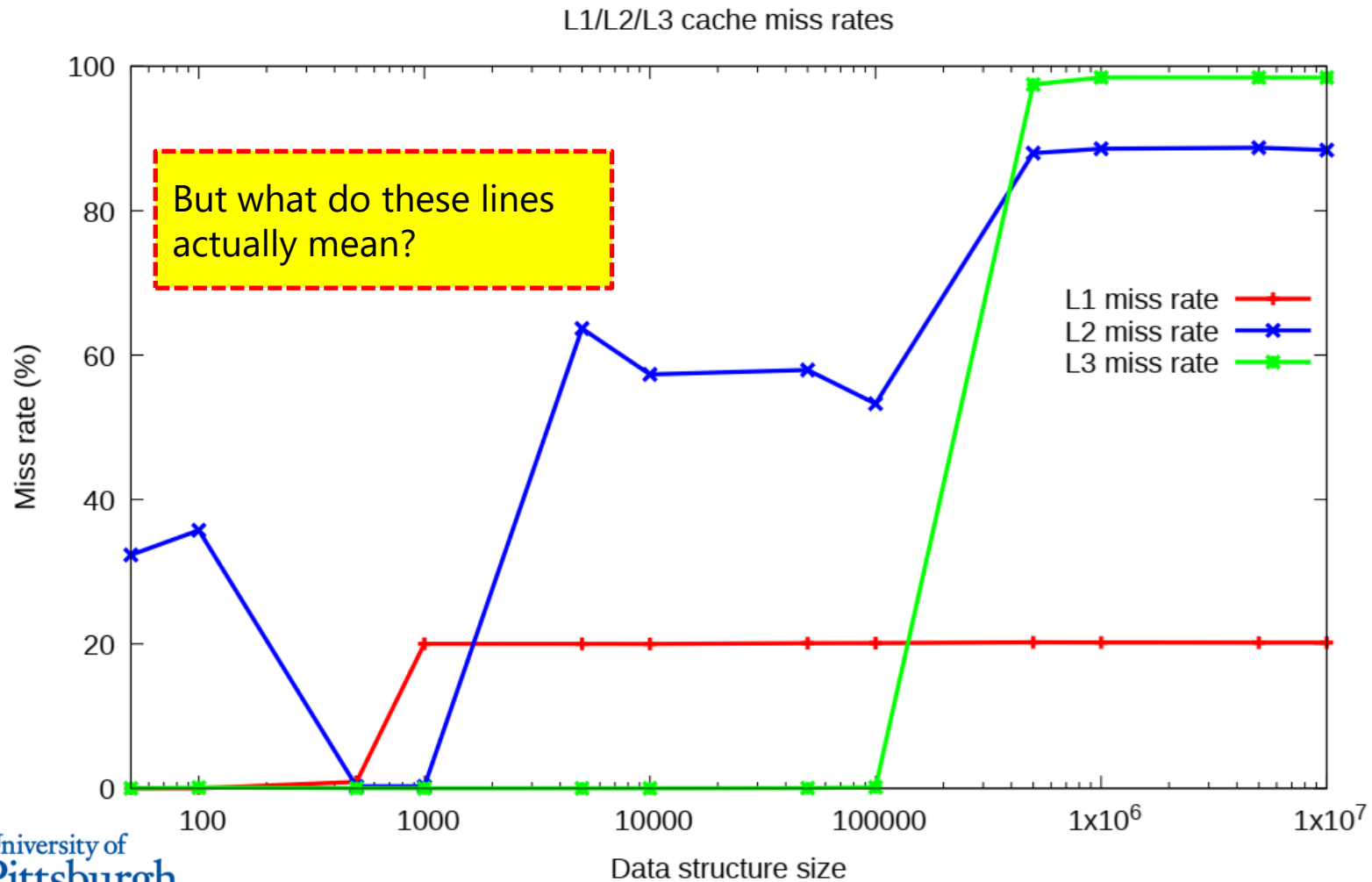
- Later, when **lw** p3, **0(p2)** comes along, it checks Write Buffer first
- If **0(p2)** == **0xdeadbeef**, Write Buffer provides value instead of memory



So are stores never on the critical path?

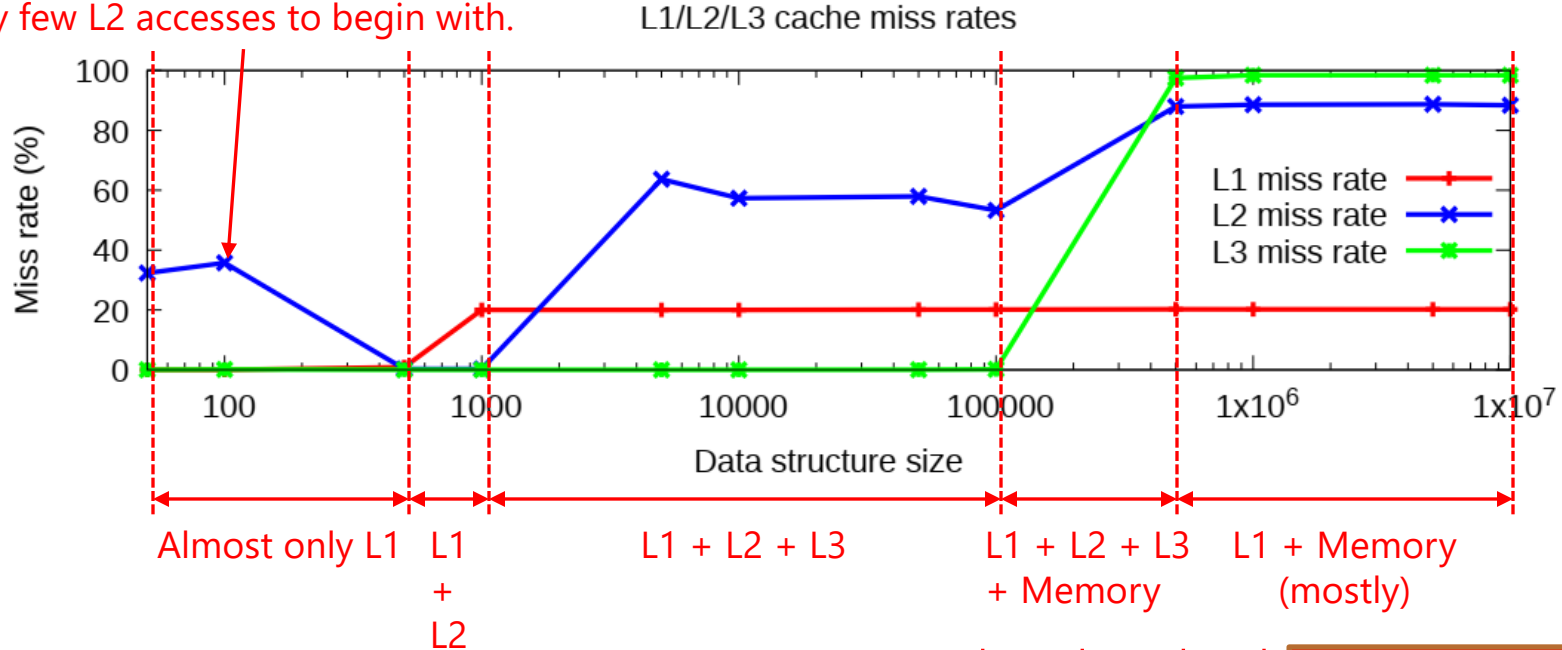
- If we had an infinitely sized write buffer, no, never.
- In real life, write buffer is limited and can suffer structural hazards
 - If write buffer is full of pending stores, you can't insert more
 - Will prevent a missing store from committing from i-queue
 - That will prevent all subsequent instructions from committing
 - ... and can eventually stall the entire pipeline
- But with ample write buffer size, happens rarely
- Hence, we will also focus on loads to analyze performance

Linked List Cache Load Miss Rates

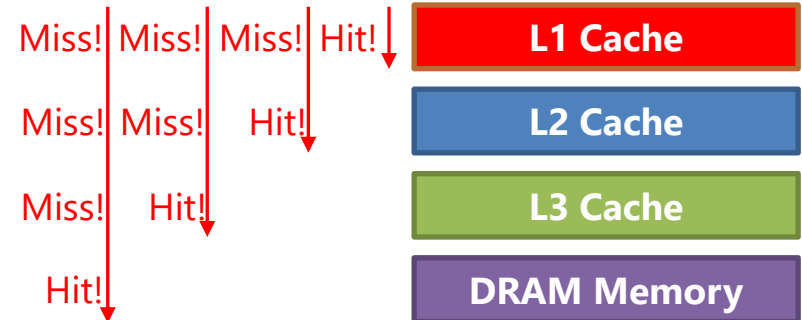


Linked List Cache Load Miss Rates

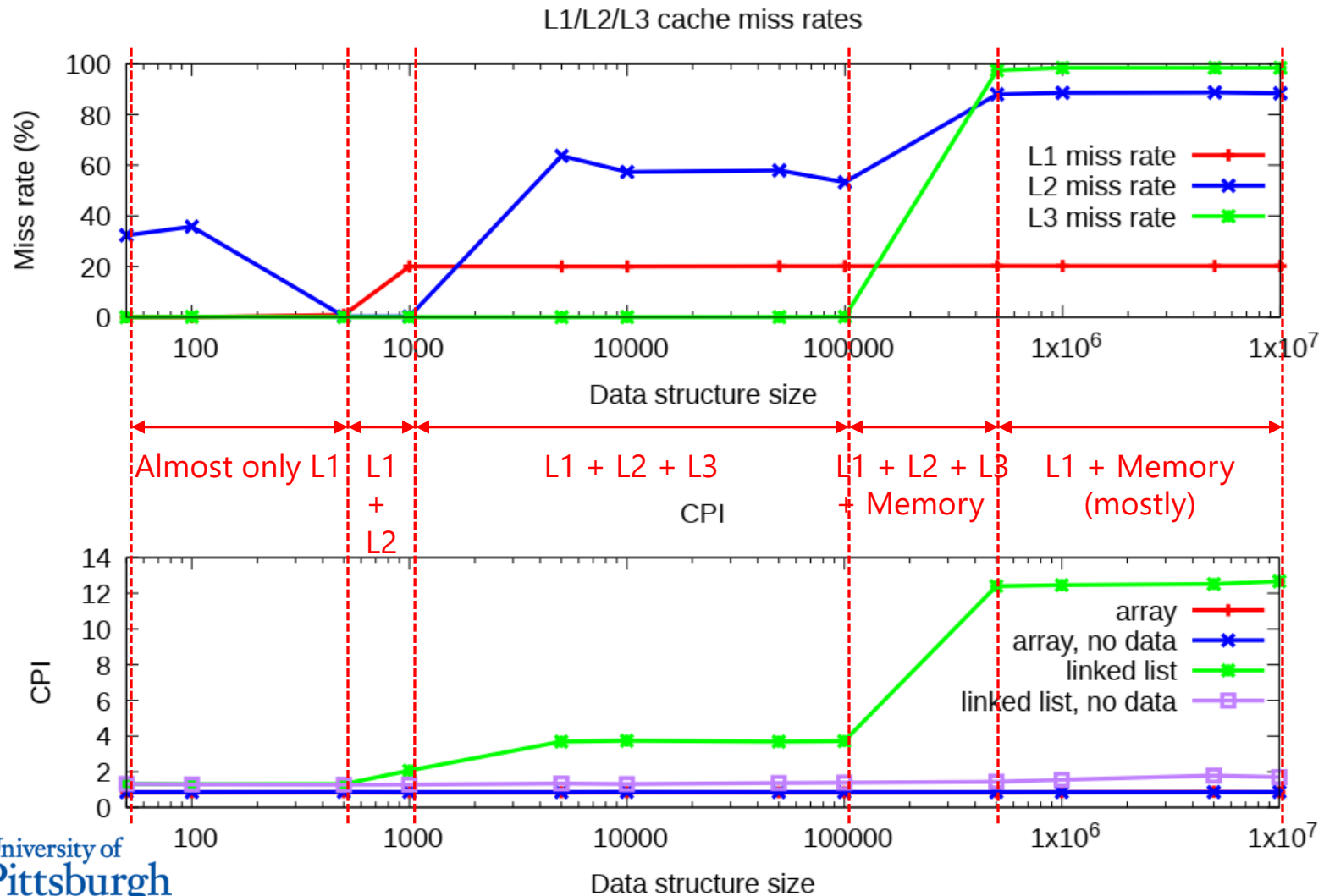
A few L2 cold misses result in high miss rate.
But very few L2 accesses to begin with.



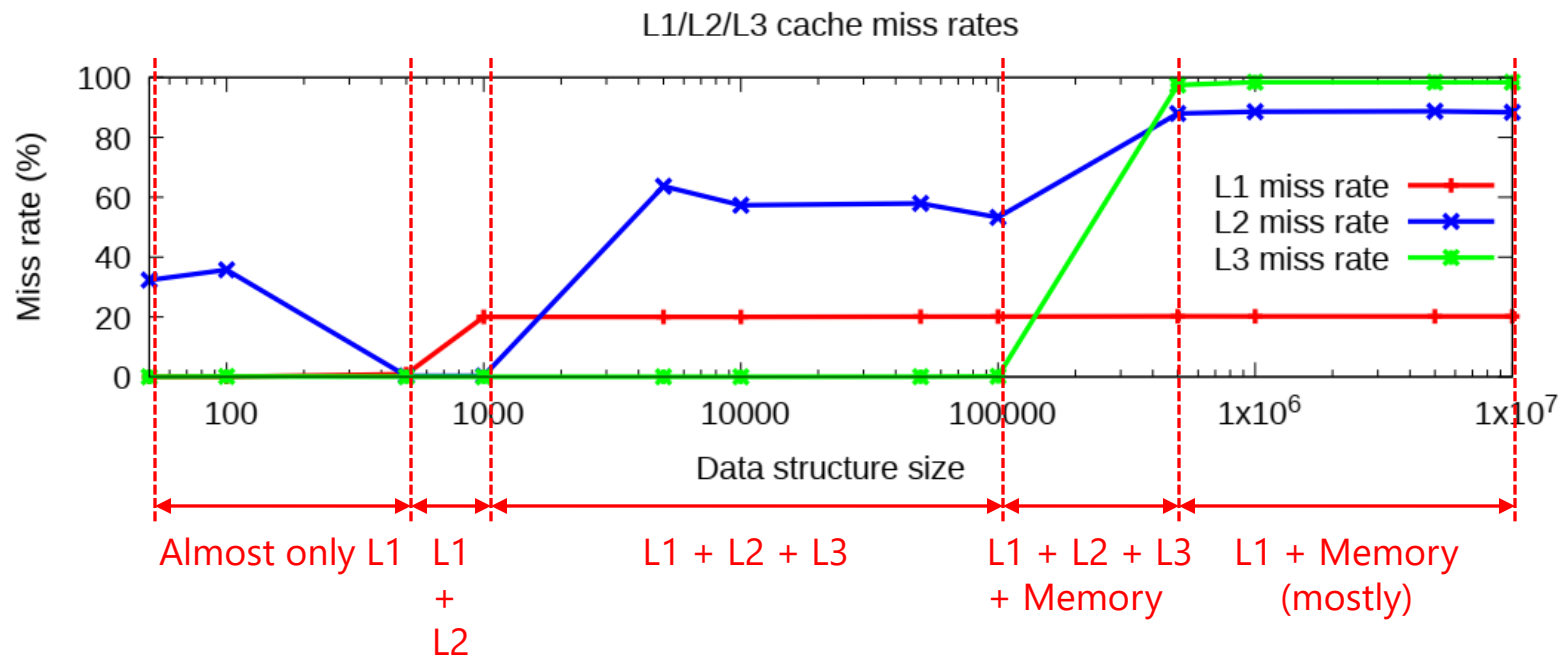
Do the miss rates correspond to CPI results?
Let's compare with our own eyes!



Linked List Cache Load Miss Rates vs CPI



Linked List Cache Load Miss Rates – Other Questions



- Why the step up in **L1** cache misses between **500 – 1000** nodes?
- Why the step up in **L2** cache misses between **1000 – 5000** nodes?
- Why the step up in **L3** cache misses between **100 k – 500 k** nodes?
- Also, why do cache miss increases look like **step functions** in general?

Working Set Overflow can cause Step Function

- The size of a node is 128 bytes:

```
typedef struct node {  
    struct node* next;    // 8 bytes  
    int data[30];         // 120 bytes  
} node_t;
```

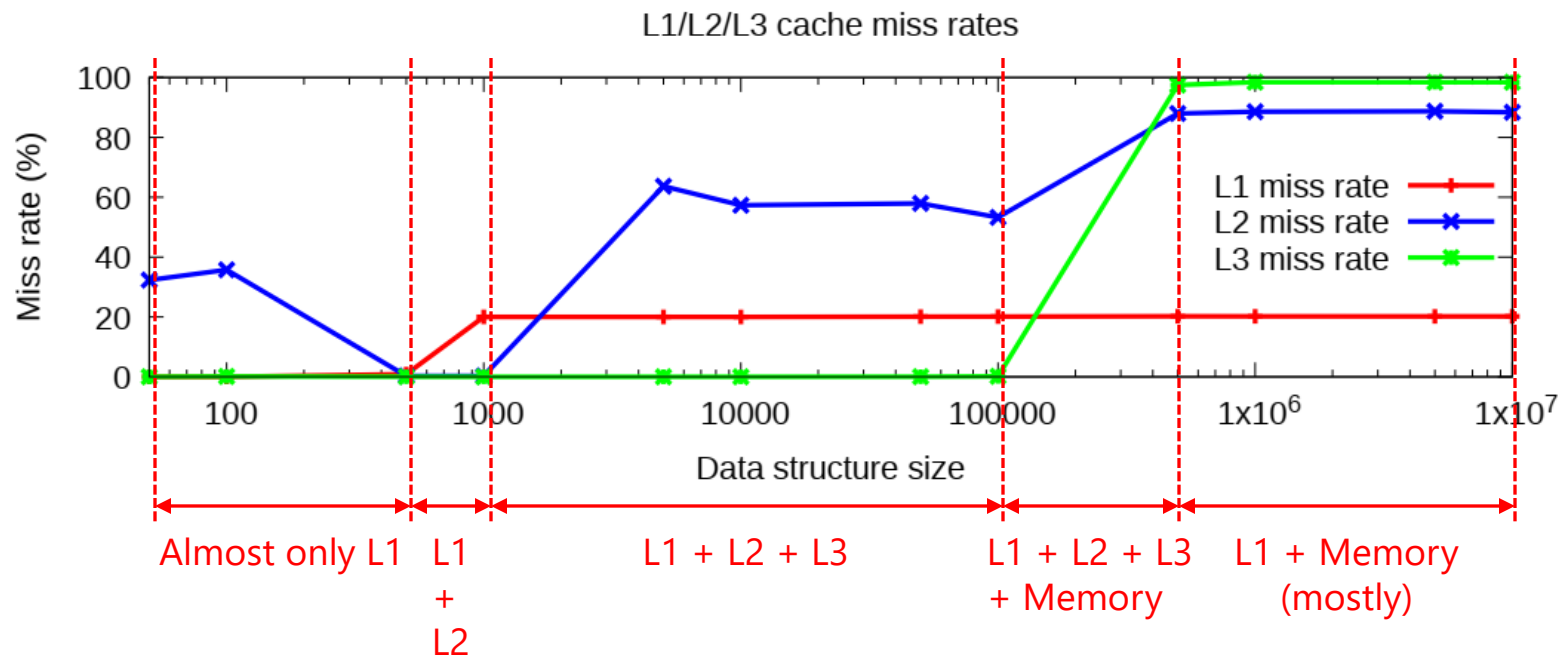
- **Working set:** amount of memory program accesses during a phase
 - For linked-list.c, working set is the entire linked list
 - Program accesses entire linked list in a loop over and over again
 - If there are 8 nodes in linked list, working set size = $128 * 8 = 1 \text{ KB}$
- When working set **overflows** cache capacity, start to see **cache misses**
 - Miss increase can be drastic, almost like a **step function**
 - Suppose cache size is 1 KB and nodes increase from 8 \rightarrow 9
 - When **8** nodes: **always hit** (since entire list is contained in cache)
 - When **9** nodes: **always miss** (if least recent node is replaced first)

Linked List Cache Load Miss Rates – Other Questions

- Why the step up in L2 cache misses between 1000 – 5000 nodes?
 - L2 cache size is 256 KB
 - Number of nodes that can fit = $256 \text{ KB} / 128 = 2048$
- Why the step up in L3 cache misses between 100 k – 500 k nodes?
 - L3 cache size is 25 MB
 - Number of nodes that can fit = $25 \text{ MB} / 128 \approx 200 \text{ k}$
- Why the step up in L1 cache misses between 500 – 1000 nodes?
 - L1 d-cache size is 32 KB
 - Number of nodes that can fit = $32 \text{ KB} / 128 = 256$
 - So, in theory you should already see a step up at 500 nodes
 - Apparently, CPU doesn't use least-recently-used (**LRU**) replacement
 - According to another reverse engineering paper, Intel uses PLRU

Ref: "CacheQuery: Learning Replacement Policies from Hardware Caches" by Vila et al.
<https://arxiv.org/pdf/1912.09770.pdf>

Linked List Cache Load Miss Rates – Other Questions



- Why did **L1** cache miss rate saturate at around **20%**?
 - Shouldn't it keep increasing with more nodes like **L2** and **L3**?

Linked List Cache Load Miss Rates – Other Questions

[linked-list.c]

```
void *run(void *unused) {
    node_t* current = head;
    for(int i=0; i<ACCESSES; i++) {
        if(current == NULL) current = head;
        else current = current->next;
    }
}
```

Within a typical iteration in for loop:

4 blue loads that hit in L1:

- 2 loads each of local vars `current`, `i`
- Read frequently so never replaced

1 red load that misses in L1:

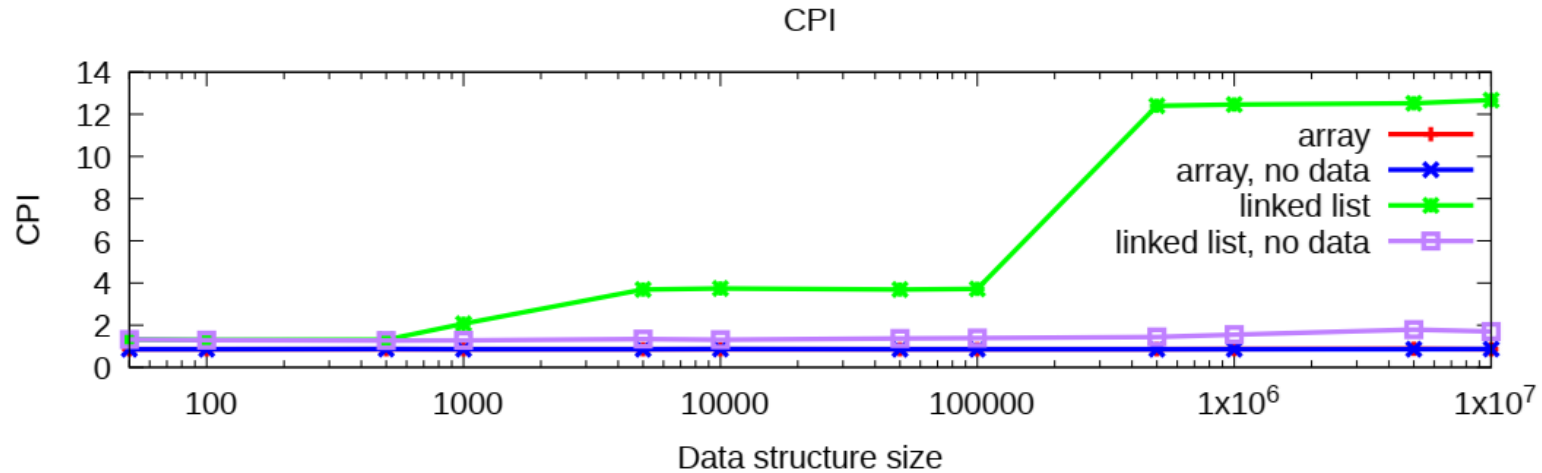
- `current->next` (next field of node)
- Node may not be in cache and miss (e.g. due to a capacity miss)

☛ 1 miss / 5 loads = 20% miss rate

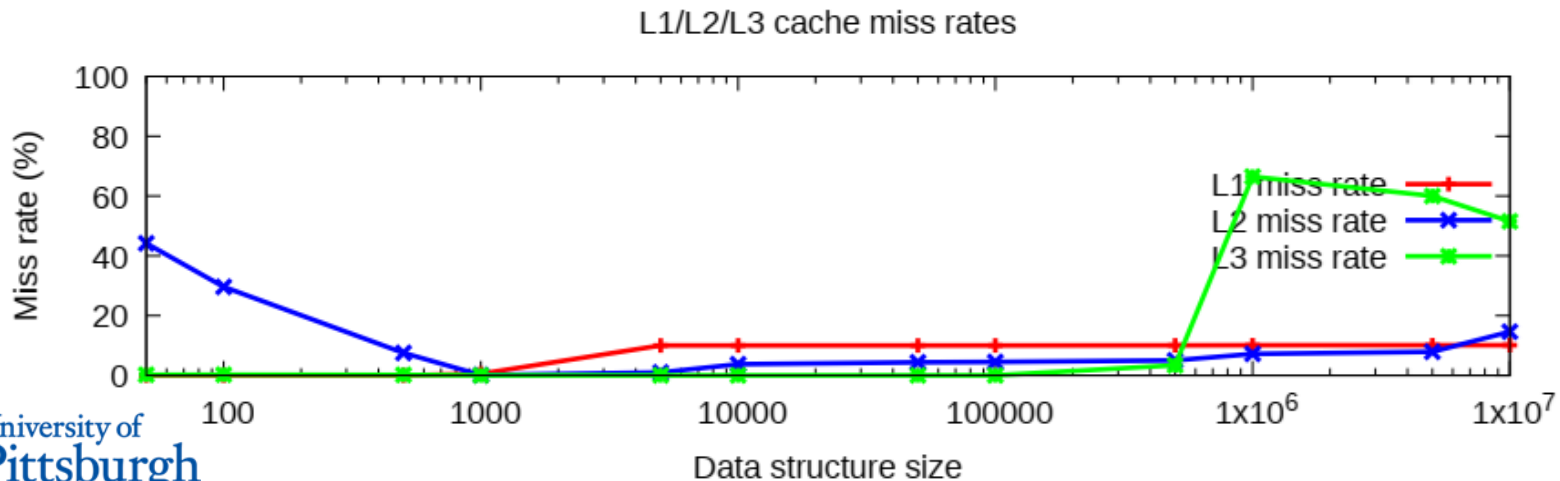
[objdump -S linked-list]

```
0000000000400739 <run>:
...
400741: mov 0x200920(%rip),%rax # %rax = head
400748: mov %rax,-0x8(%rbp) # current = %rax
40074c: movl $0x0,-0xc(%rbp) # i = 0
400753: jmp 400778 # jump to i < ACCESSES comparison
400755: cmpq $0x0,-0x8(%rbp) # current == NULL?
40075a: jne 400769 # jump to else branch if not equal
40075c: mov 0x200905(%rip),%rax # %rax = head
400763: mov %rax,-0x8(%rbp) # current = %rax
400767: jmp 400774 # jump to end of if-then-else
400769: mov -0x8(%rbp),%rax # %rax = current
40076d: mov (%rax),%rax # %rax = current->next
400770: mov %rax,-0x8(%rbp) # current = %rax
400774: addl $0x1,-0xc(%rbp) # i++
400778: cmpl $0x3b9ac9ff,-0xc(%rbp) # i < ACCESSES ?
40077f: jle 400755 # jump to head of loop if less than
```

How about Linked List with No Data?

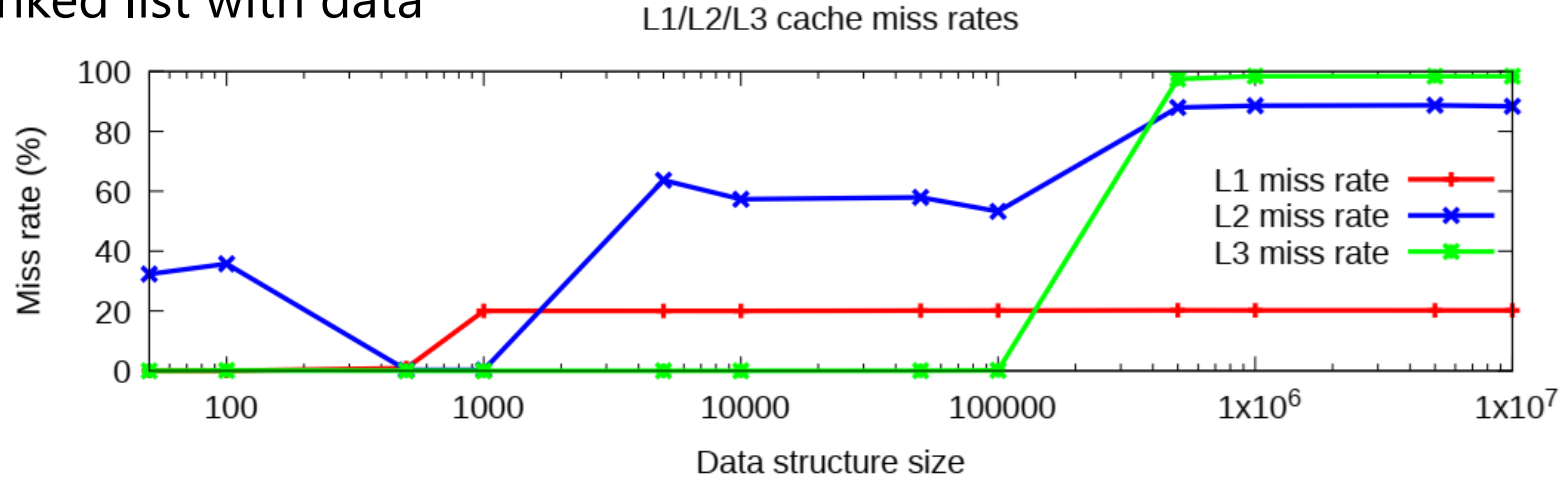


- Linked list, no data suffered almost no CPI degradation. Why?

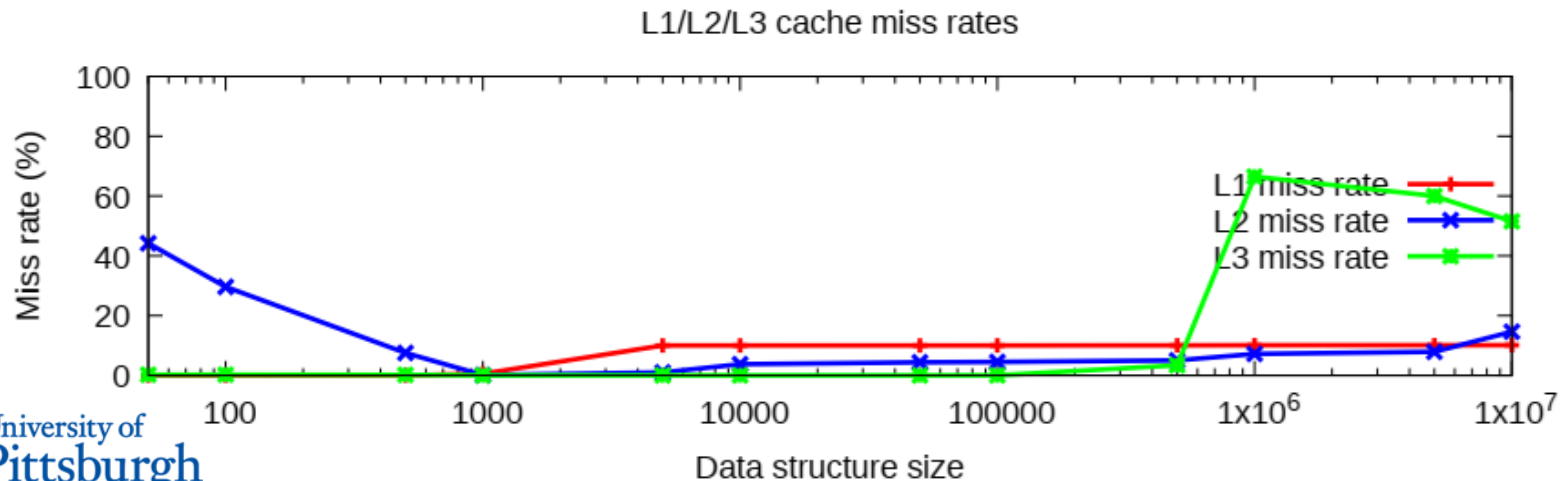


Linked List w/ Data vs. w/o Data

- Linked list with data



- Linked list with no data



Linked List w/ Data vs. w/o Data. Why?

- The size of a node with no data is only 8 bytes:

```
typedef struct node {  
    struct node* next;    // 8 bytes  
    // no data  
} node_t;
```

- Compared to 128 bytes with data, can fit in 16X more nodes in cache
 - **Temporal locality**: More likely that a node will be present in cache
- How about L1 cache miss rate that hovers around 10% instead of 20%?
 - By 10^7 nodes, there is no temporal locality with respect to the L1 cache
 - **Spatial locality** must be responsible for the reduction in miss rate

Linked List w/ Data vs. w/o Data. Why?

- Nodes of the linked list are malloced one by one in a loop:

```
for(int i=0; i<items; i++) {  
    node_t* n = (node_t*)malloc(sizeof(node_t));  
    ...  
}
```

 - I have no idea where glibc malloc decides to allocate each node
- But knowing each cache block is 64 bytes long in the Xeon E5 processor
 - Let's say multiple nodes are allocated on same cache block:

node 1	meta-data	meta-data	node 37	meta-data	meta-data	node 23	meta-data
---------------	------------------	------------------	----------------	------------------	------------------	----------------	------------------

 - Then even if access to node 1 misses, due to a capacity miss, accesses to nodes 37 and 23 that soon follow will hit!
 - This is assuming there is some **spatial locality** in how malloc allocates

Data structure with most spatial locality: Array

- Elements of an array are guaranteed to be in contiguous memory:

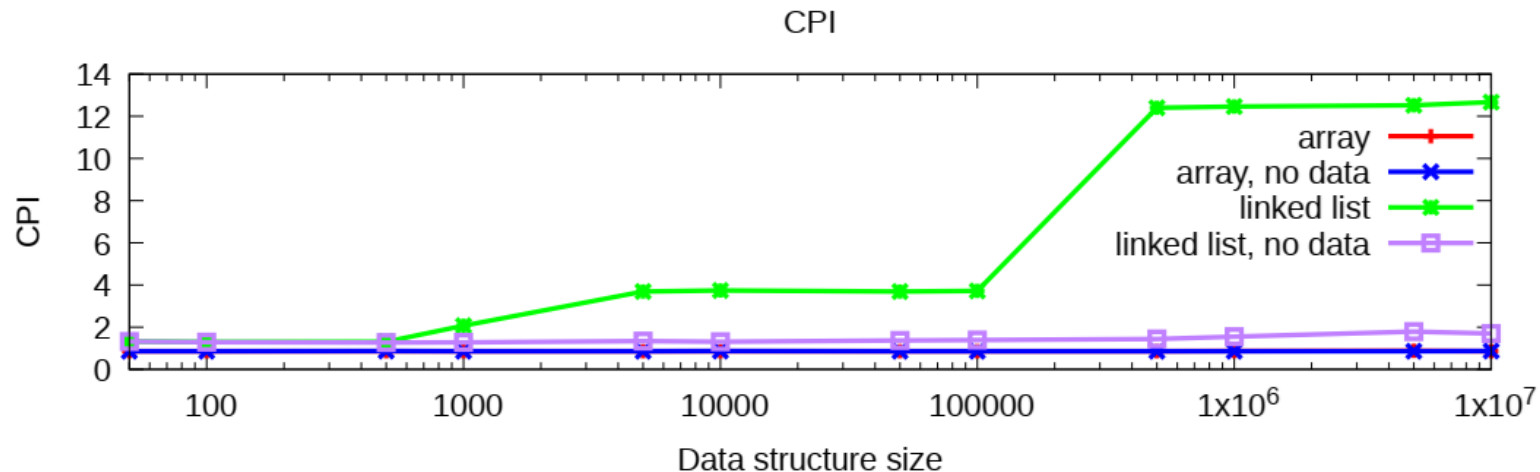
```
void *create(void *unused) {  
    head = (node_t *) malloc(sizeof(node_t) * items);  
    ...  
}
```

- Each cache block is 64 bytes long in the Xeon E5 processor
 - Now 8 elements are guaranteed to be on same cache line, in order:

node 0	node 1	node 2	node 3	node 4	node 5	node 6	node 7
---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------

- Even with cold cache, only 1 out of 8 nodes miss ($1/8 = 12.5\%$ miss rate)
- Assuming that nodes are accessed sequentially
 - If accessed in random order, no spatial locality, even with array
- True regardless of capacity (even if cache contained only a single block)

Let's look at the CPI of arrays finally



- **Array, no data** did very well as expected
 - The most spatial locality of the four benchmarks (contiguous nodes)
 - Smallest memory footprint so can also leverage temporal locality the best
- **Array** performed the same as **array, no data**. How come?
 - No spatial locality since each node takes up two 64-byte cache blocks
 - Has much larger memory footprint compared to **array, no data**
 - This is the real mystery. We will learn more about it as we go on. 😊