LAST TIME

- Examined several memory technologies:
  - **SRAM** – volatile memory cells built from transistors
    - Fast to use, larger memory cells (6+ transistors per cell)
  - **DRAM** – volatile memory cells built from capacitors
    - Slower to use, smaller memory cells, can make very large
  - **Magnetic disk storage (HDDs)** – nonvolatile memory
    - Very slow to use, compared to SRAM/DRAM/CPU!
    - Can make extremely large disks
  - **Solid State Drives (SSDs)** – also nonvolatile
    - Faster and more reliable than HDDs, but more expensive

- Also discovered two important principles:
  - Small memories tend to be faster than large ones, due to physical limitations
  - Can make larger memories from denser technologies (but addressing and accessing are more expensive)
STORAGE TECHNOLOGIES

- Modern computers use SRAM, DRAM, solid state drives (SSDs) and magnetic disks (HDDs)

Comparison:

<table>
<thead>
<tr>
<th>Technology</th>
<th>Access Time</th>
<th>Cost (2017 numbers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRAM</td>
<td>1-10ns</td>
<td>$5/MB</td>
</tr>
<tr>
<td>DRAM</td>
<td>30-100ns</td>
<td>$5/GB</td>
</tr>
<tr>
<td>SSD</td>
<td>10-100µs</td>
<td>$0.30/GB</td>
</tr>
<tr>
<td>Hard disk</td>
<td>8-30ms</td>
<td>$0.018/GB</td>
</tr>
</tbody>
</table>

- Compare these to processor performance!
  - 3GHz processor: 1 clock = 0.3ns

- If 1 clock were 1 second:
  - SRAM access would take 3-30 seconds
  - DRAM access would take 1½ to 5 minutes
  - Hard disk access would take 1-3 years!
THE PROCESSOR-MEMORY GAP

- Processor is significantly faster than memory...
- ...but it’s actually worse than that:

Source: Hennessy and Patterson 2011
The Processor-Memory Gap (2)

- Processor speed has been growing by ~60%/year
- DRAM speed has been growing by ~8%/year

- The gap has been growing by >50%/year!
COMPUTER MEMORY SYSTEM: GOAL

- Would like to have memory the size of DRAM, but performance of SRAM or faster
  - Ideally, even performance of registers
- We will never get it with this design:

  **Idea:** Introduce a cache between the processor and main memory
  - Use a faster storage technology than DRAM
  - Use the cache as a staging area for information in main memory
CACHING

- When CPU reads a value from main memory:
  - Read an entire block of data from main memory
  - As long as subsequent accesses are within this block, just use the cache!
  - If an access is outside the cached data, need to retrieve and cache a new block of data from memory

Today:
- What program behaviors maximize benefit of caching?
- How are these caches designed? What are the trade-offs?
**Caching and Data Access Patterns**

- When CPU reads a value from main memory:
  - Read neighboring values from main memory as well

- Not all programs will benefit from this approach!
- Only programs that access data in specific patterns will benefit from the cache
  - Most accesses end up being served from the cache
- Programs that usually access data in a widely varied manner will not benefit from cache at all
CACHING AND LOCALITY

- Programs must exhibit good locality if they are going to utilize the cache effectively
- Spatial locality:
  - Program accesses data items that are close to other data items that have been recently accessed
- Temporal locality:
  - Program accesses the same data item multiple times
- Well-written programs will exhibit good locality
  - (“well-written” in terms of cache-friendliness)
  - ...and the computer hardware can run them faster!
- Poorly-written programs have poor locality
  - Program can’t take advantage of system caches
Locality: Examples

- Frequently can achieve good locality very easily: programs tend to access data in regular patterns
- Vector-add code from before:
  ```c
  int * vector_add(int *a, int *b, int length) {
    int i;
    int *result =
      (int *) malloc(length * sizeof(int));
    for (i = 0; i < length; i++)
      result[i] = a[i] + b[i];
    return result;
  }
  ```
- Elements of input and output arrays are accessed in sequential order – works well with caching
LOCALITY: EXAMPLES (2)

Still *extremely* valuable to understand locality as a programmer!
- Simple choices can have a profound impact on program performance

Molecular dynamics example from lecture 1

```c
#define N_ATOMS 10000
#define DIM 2
/* Array of data for each atom being simulated. */
double atoms[N_ATOMS][DIM][DIM];
```

Version 1:
```c
for (i = 0; i < DIM; i++)
    for (j = 0; j < DIM; j++)
        for (n = 0; n < N_ATOMS; n++)
            atoms[n][i][j] = ... ;
```

This code has poor data locality, so it runs slower
LOCALITY: EXAMPLES (3)

- Memory layout of our atoms array:
  ```
  double atoms[N_ATOMS][DIM][DIM];
  ```

- Version 1:
  ```
  for (i = 0; i < DIM; i++)
      for (j = 0; j < DIM; j++)
          for (n = 0; n < N_ATOMS; n++)
              atoms[n][i][j] = ... ;
  ```

  - Code accesses every 4\textsuperscript{th} array element, because of the way the loops are arranged
  - Rearrange loops to access each element in sequence
Locality: Examples (4)

- Memory layout of our atoms array:

  ```
  double atoms[N_ATOMS][DIM][DIM];
  ```

  ![Memory layout diagram]

- Version 2:

  ```
  for (n = 0; n < N_ATOMS; n++)
      for (i = 0; i < DIM; i++)
          for (j = 0; j < DIM; j++)
              atoms[n][i][j] = ... ;
  ```

- This code accesses each array element in sequence
- A simple change, and results in a significant performance improvement!
LOCALITY AND STRIDE

- When a program accesses array elements sequentially, called a stride-1 reference pattern
  - (Consider stride in terms of processor word-size)
  - Vector-sum program has a stride-1 reference pattern
  - Updated version of molecular dynamics program also has a stride-1 reference pattern

- When a program accesses every $k^{th}$ element in sequence, called a stride-$k$ reference pattern
  - Original version of molecular dynamics program has a stride-4 reference pattern

- Generally, as stride increases, program locality decreases

- With 1D arrays, pretty easy to achieve stride-1
- With multidimensional arrays, it can become much trickier to achieve stride-1
EXTENDING OUR CACHING IDEA

- Added an external SRAM cache between the CPU and main memory (DRAM)
- Accessing the SRAM cache is still slower than using registers...
  - Can access a register in 0 clocks (i.e. same clock that instruction is executed in)
  - SRAM cache can take e.g. 1-30 clocks, depending on size
- Also, disk access is horribly slow!
  - 20 million clocks or more!
  - Accessing DRAM is much faster than accessing the disk...
  - Could exploit data locality with disk accesses as well, by using DRAM as a cache for the disk

Why not apply our caching technique in other places?
THE MEMORY HIERARCHY

L0: Registers
L1: On-Chip SRAM Cache
L2/L3: Off-Chip SRAM Cache
L4: Main Memory (DRAM)
L5: Local Secondary Storage (Local Disks)

Each level provides caching for the levels below it.

CPU registers hold and manipulate data values retrieved from L1 cache.

L1 cache holds blocks of data retrieved from L2/L3 cache.

L2/L3 caches hold blocks of data retrieved from main memory.

Main memory holds disk blocks read from local disks.

Local disks hold files retrieved from remote servers.

Smaller, faster, and more expensive per byte of storage.

Larger, slower, and cheaper per byte of storage.
THE MEMORY HIERARCHY (2)

- This is a very typical memory hierarchy used in modern computers, but by no means the only one!
- Many places where caching is employed within levels of the hierarchy
- Example: hard disks also employ caches
  - Disk buffer, often 2MB-64MB
    - Caches data prefetched from disk, pending writes to disk
  - Non-volatile RAM buffer (less common)
    - Disk writes are stored to this buffer until it fills up, then written to the disk itself
    - If power fails, this non-volatile memory retains its state
  - Finally, the magnetic disk storage itself
- Some systems use solid-state drives (SSDs) to cache data from traditional spinning hard-disk
CACHE MANAGEMENT

- Each level $k$ provides caching for level $k + 1$
- Some caches are managed entirely by hardware
  - L1 (on-chip SRAM), L2/L3 (off-chip SRAM) caches
  - Performance is absolutely critical, so hardware is designed to manage these caches
- Other caches are managed entirely by software
  - L4 (DRAM main memory) is managed extensively by the operating system
  - e.g. the operating system caches disk blocks from L5 into L4 to improve disk IO performance
  - CPU registers (L0 cache) are manually assigned by compiler to minimize need to access memory (L1+)
CACHES AND MEMORY BLOCKS

- Level $k$ caches data from level $k + 1$:
  - Memory at level $k + 1$ is partitioned into fixed-size blocks
  - Level $k$ stores these blocks in its cache
  - Anytime data needs to be transferred between levels $k$ and $k + 1$, this block size is used

- Actual block size depends on characteristics of levels $k$ and $k + 1$
  - Between L0 and L1, block size is 1 word
  - Between L1 – L4, block is 8 to 16 words (64B)
  - Between L4 and L5, block is up to several KB

- Lower levels have longer access times…
  - Also usually designed to read/write larger blocks of data from storage in one shot...

- Amortize read/write cost over a larger amount of data
Hardware caches manage blocks of data from the next level...

- Clearly need more details than just the data itself!

**Cache lines hold:**

- A flag indicating whether the line currently holds valid data
- A tag that uniquely identifies the block
  - Taken from the address where the block is actually stored
- The block of cached data itself

**Pictorially:**

<table>
<thead>
<tr>
<th>1 bit</th>
<th>$t$ bits</th>
<th>$2^b$ bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>Tag</td>
<td>Block</td>
</tr>
</tbody>
</table>
Cache Operation

- CPU makes requests to main memory

- Main memory divided into blocks of $B$ bytes ($B = 2^b$)
- Cache lines are slightly larger than $B$ bytes
  - Line also includes the tag and a “valid” flag, as well as block

- Example: CPU requests a word in block 10
  - Cache returns value directly, since block 10 is cached
  - This is called a **cache hit** – the requested item was contained within the cache
**Cache Misses**

- Next, CPU requests a word in block 6

  - Cache doesn’t contain block 6
    - This is called a cache miss
    - Cache must load block 6 from main memory before providing requested value to the CPU
    - CPU incurs performance cost of accessing main memory

- What else must the cache do?
  - Figure out where to store block 6...
  - Need to *replace* (or *evict*) an existing block in the cache
REPLACEMENT POLICIES

- CPU requested a value in block 6

- Caches have a replacement policy:
  - When loading a new block into a full cache, which existing block should be replaced?

- Example replacement policies:
  - Least Recently Used (LRU) policy: evict the block that was accessed furthest in the past
  - Random replacement policy: randomly pick a block to evict
**Cold Caches**

- Will there be a cache miss in this situation?
  - All cache lines have valid flag = 0
  - Duh, of course!
- Caches have different kinds of cache misses
- This is called a *cold cache*
  - No accesses yet, so cache isn’t populated with data!
  - Cache is “warmed up” by accessing memory and populating the cache lines
  - Misses during this phase are called *cold misses*
  - Also called “compulsory misses” since they are unavoidable
**Cache Placement Policy**

- CPU requests a word in block 5
  - The biggest part of a cache’s work?
    - Determining whether the requested block actually appears in the cache, and if so, where is it?!
- Caches implement a **placement policy**, specifying where new blocks are placed in the cache
- Most flexible placement policy is random
  - Blocks from level $k + 1$ can be stored anywhere in level $k$
- Also the most expensive placement policy!
  - Very costly to track down blocks in the cache
  - Hardware caches usually cannot implement this policy
Cache Placement Policy (2)

- Hardware caches use a much stricter placement policy
  - Blocks from level $k + 1$ can only be stored into a subset of locations in level $k$
  - Makes it much faster and easier to determine if a block is already in the cache

- Cache lines are grouped into cache sets
  - Every block from level $k + 1$ maps to exactly one set in the cache at level $k$
  - If cache set contains multiple cache lines, a block may be stored into any cache line in the set

- Two extremes for our cache organization:
  - $S$ cache sets, each of which contains exactly one line
  - One cache set which contains all $E$ cache lines
Cache Placement Policy (3)

- Cache lines are grouped into cache sets
  - Every block from level $k + 1$ maps to exactly one set in the cache at level $k$

- Direct-mapped caches:
  - $S$ cache sets, each of which contains exactly one line
  - Fast and easy to determine if a block is in the cache
  - Find the cache set associated with the block, and look at the one cache line in the set

- Fully associative caches:
  - One cache set which contains all $E$ lines
  - Much more expensive to find if a block is in the cache
  - Need to look at the tags from all cache lines, to see if the block is in the cache
**Direct-Mapped Caches and Misses**

- Direct-mapped caches map each memory block to a single cache line
  - Each block only appears in one cache set, and each cache set only contains one cache line
- Can lead to new kinds of cache misses
- Example: our cache from before
  - Four cache sets, each of which contains one line
  - A block \( i \) is mapped to cache-set \( (i \mod 4) \)

What if a program alternately accesses data only in blocks 9 and 13?
**DIRECT-MAPPED CACHES AND MISSES (2)**

- **Example:** direct-mapped cache
  - A block \(i\) is mapped to cache-set \((i \mod 4)\)
  - Program alternately accesses data only in blocks 9 and 13

Each time CPU makes a request, the cache doesn’t contain the associated block!

These are called **conflict misses**
- Cache is large enough to hold the requested blocks...
- …but due to placement policy constraints, cache has to keep reloading blocks from main memory
DIRECT-MAPPED CACHES AND MISSES (3)

- Example: direct-mapped cache
  - A block \(i\) is mapped to cache-set \((i \mod 4)\)
  - Program alternately accesses data only in blocks 9 and 13

- If a program generates a large number of conflict misses like this, it is called **thrashing**
  - For our example, program thrashes between blocks 9 and 13

- **A major issue!**
  - Program can have great locality, but still runs horribly slow!
  - If a program thrashes, the correct adjustments to make are often very subtle
CACHE ORGANIZATION

- Cache is designed to work against a main memory with $M$ bytes
  - $m = \log_2(M)$ bits in addresses to main memory
- Caches have several important parameters
  - $B = 2^b$ bytes to store the block in each cache line
  - $S = 2^s$ cache sets
  - $E$ cache lines per set
  - Both $S$ and $B$ are powers of 2
- The cache stores $B \times E \times S$ bytes of data from main memory
  - (Don’t forget: cache lines also include a tag and a valid flag, which require additional space)
Mapping Cache Blocks

- Main memory with $M$ bytes:
  - $m = \log_2(M)$ bits in addresses to main memory

- Cache parameters:
  - $B = 2^b$ bytes to store the block in each cache line
  - $S = 2^s$ cache sets
  - $E$ cache lines per set

- Given a specific memory address, the cache must:
  - Map the address to a memory block
    - *(the cache works with blocks, not individual values)*
  - Figure out which cache set the block would live in
  - Figure out the tag that uniquely identifies the block
  - Figure out the offset of the address within the block

- Take $m$ address bits, map them to these things
Mapping Cache Blocks (2)

- Relevant parameters for main memory and cache:
  - \( m = \log_2(M) \) bits in addresses to main memory
  - \( B = 2^b \) bytes to store the block in each cache line
  - \( S = 2^s \) cache sets in the cache

- When CPU accesses a data value:

  - Bottom-most \( b \) bits of the address specify offset of data value within the block
  - Middle \( s \) bits specify the cache set where the block resides
  - Remaining topmost bits constitute the block’s tag
    - (Must be able to uniquely identify all blocks that can be cached)
Mapping Cache Blocks (3)

- Why use middle bits for set index?
  - Why not topmost bits?

- If topmost bits identify the cache set, will cause long runs of addresses to map to same cache set
- Example: $M = 16$ bytes, $S = 4$ cache sets, $s = 2$
- If topmost bits select cache set, programs with good locality won’t map accesses to different sets

- If middle bits select cache set, programs with good locality use all cache sets much more evenly
CACHE ACCESS OPERATIONS

- When cache receives a memory access, it must:
  - Figure out the cache set where the block goes
  - Figure out which cache line matches the block *(if any)*
  - Access the specific value within the cached block

- Given our mapping from memory addresses to cache details, these steps become pretty easy:

<table>
<thead>
<tr>
<th>m-bit address:</th>
<th>( t = m - (s + b) ) bits</th>
<th>( s ) bits</th>
<th>( b ) bits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \text{tag} )</td>
<td>( \text{set index} )</td>
<td>( \text{block offset} )</td>
</tr>
</tbody>
</table>

- Cache set’s index is the middle \( s \) bits
- Use tag bits to identify the cache line within the set
- Bottom-most \( b \) bits are the access’ offset within the block
DIRECT-MAPPED CACHES

- Direct-mapped caches have $S$ sets, but each set contains only one cache line ($E = 1$)

$m$-bit address: $t = m - (s + b)$ bits

- Example: direct-mapped cache with 4 sets
  - 2 bits in set index
- Very fast to map an address to a cache set
- Very fast to determine if a block is in the cache
  - If the tag doesn’t match, block isn’t in the cache!

<table>
<thead>
<tr>
<th>Set</th>
<th>Valid</th>
<th>Tag</th>
<th>Block</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Fully Associative Caches**

- Fully associative caches have only one set, which contains all cache lines
  - $S = 2^s = 1 \Rightarrow s = 0$. No bits used for set index!

- Example: fully-associative cache
- Still very fast to map an address to a cache set
- More complicated to find if a block is in the cache
  - Need to examine all cache-line tags
  - Also, the tag is larger than in a direct-mapped cache