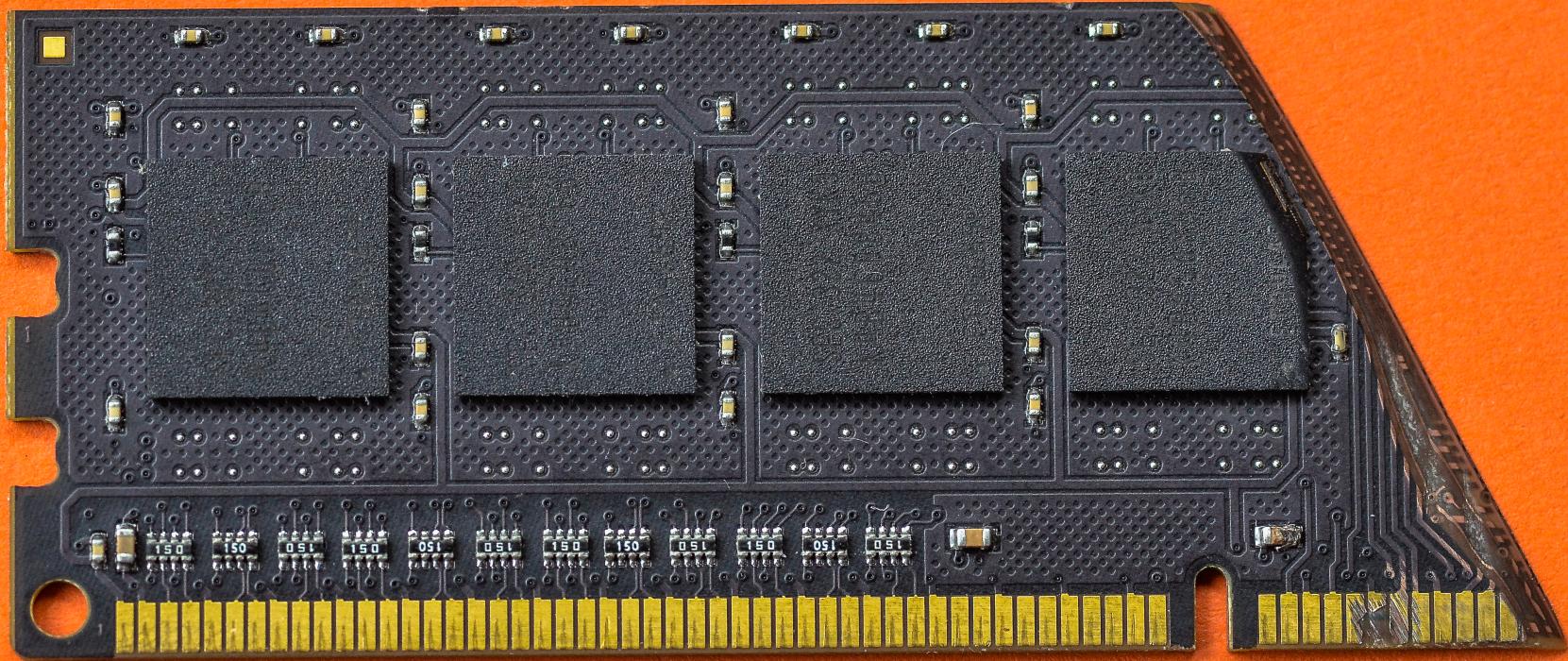


Memory



Memory

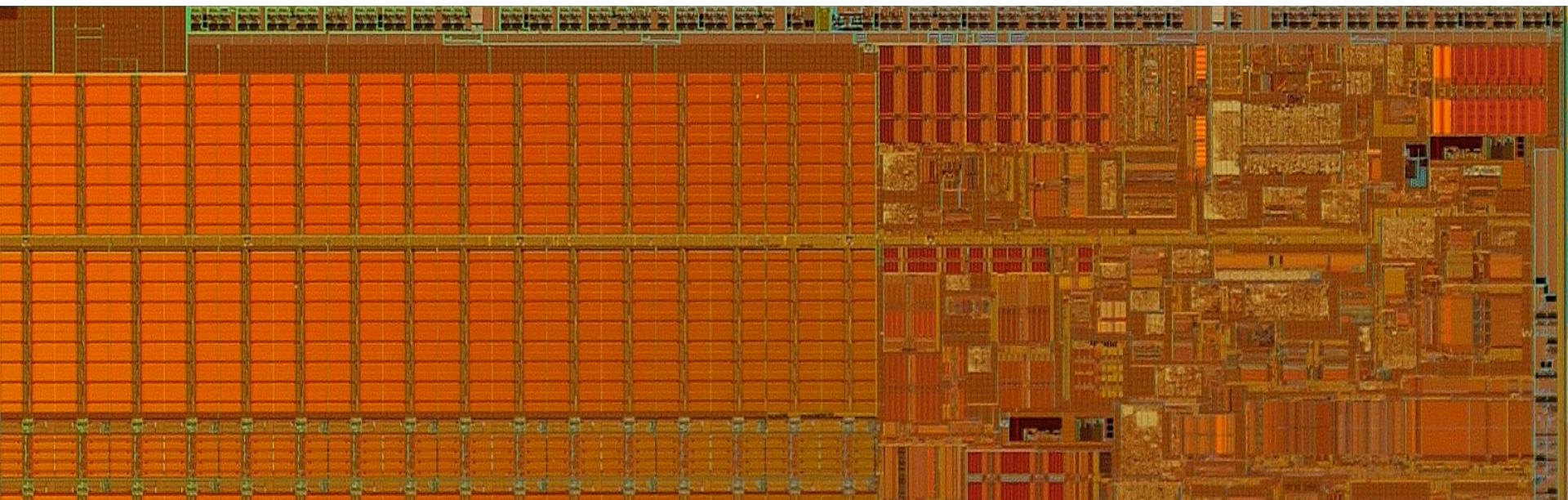
On completion of this lecture you will be able to:

Explain the origins and consequences of the memory performance gap

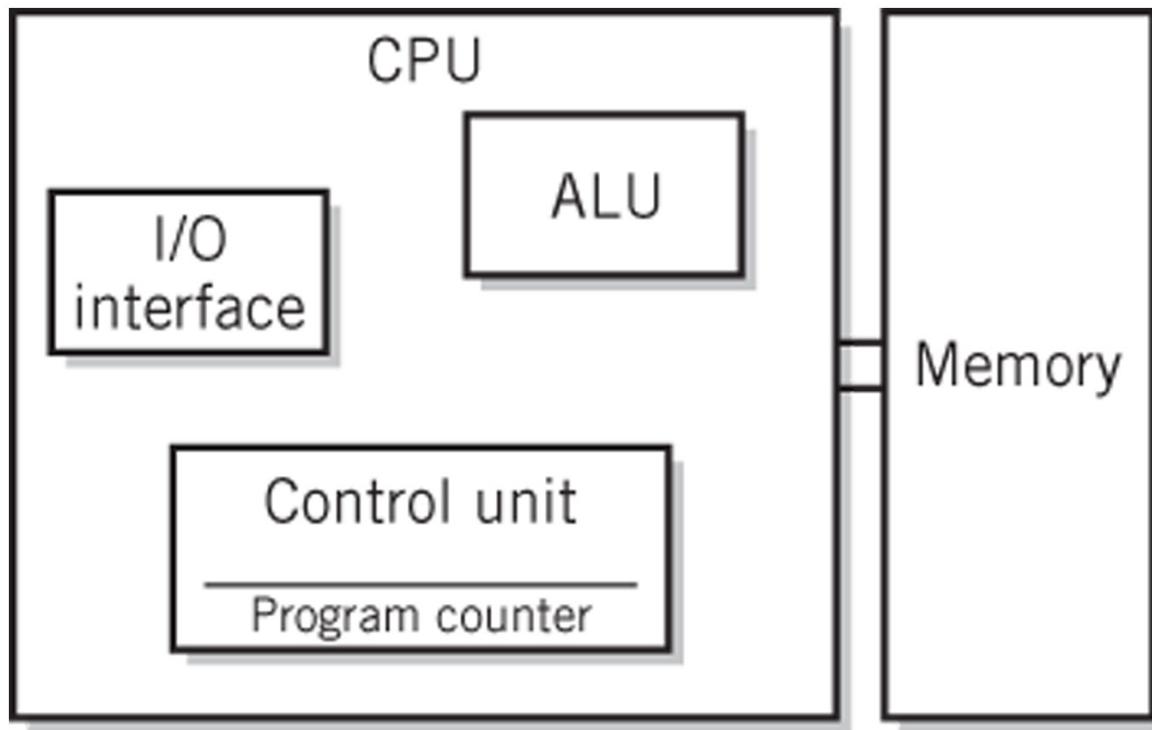
Explain the different levels in the **memory hierarchy**

Explain the concept of **locality** as it pertains to memory access

Quantify the impact of **cache hits and misses**



CPU & Primary Memory



Quantities

1 bit (b)

1 nibble = 4 b = 2^2 b

1 Byte (B) = 8 b = 2^3 b

1 KiloByte (KB) = 1,024 B = 2^{10} B

1 MegaByte (MB) = 1,048,576 B = 2^{20} B

1 GigaByte (GB) = 1,073,741,824 = 2^{30} B

1 TeraByte (TB) = 1,099,511,627,776 = 2^{40} B

What comes next?

For the curious, read more about the origin of these prefixes:

https://en.wikipedia.org/wiki/International_System_of_Units



Memory Metrics

Cost

Capacity (MB)

Maximum number of bytes of data that can be stored.

Access time (ns).

Delay between request and first data.

Memory bandwidth (Gb/s).

Sustained data transfer rate between memory and CPU.

- Width of interface (bits)
- Frequency of interface (MHz)
- Number of channels

Volatile or non-volatile: Data lost when memory is not powered

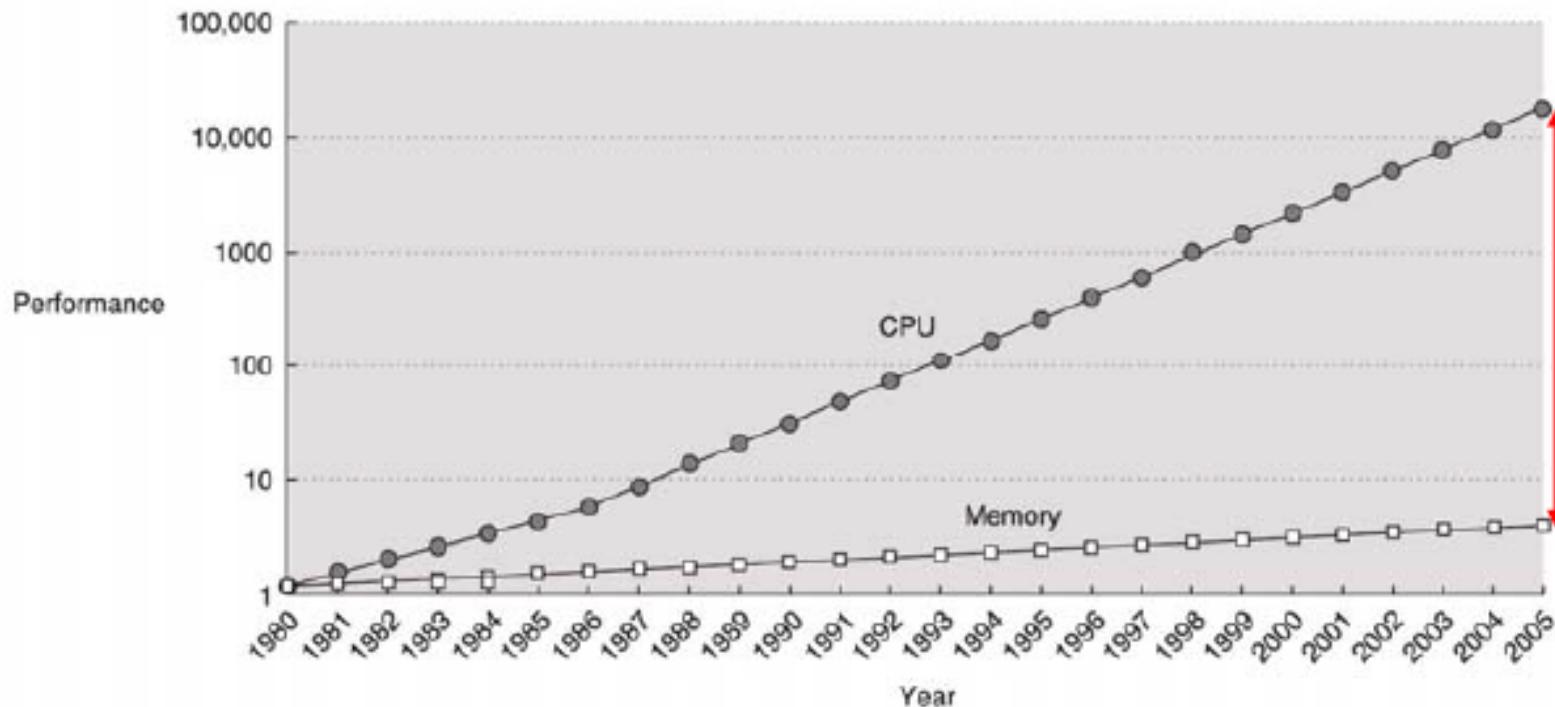


Memory performance gap

CPU speeds increase 25%-30% per year

DRAM speeds increase 2%-11% per year

https://en.wikipedia.org/wiki/Dynamic_random-access_memory



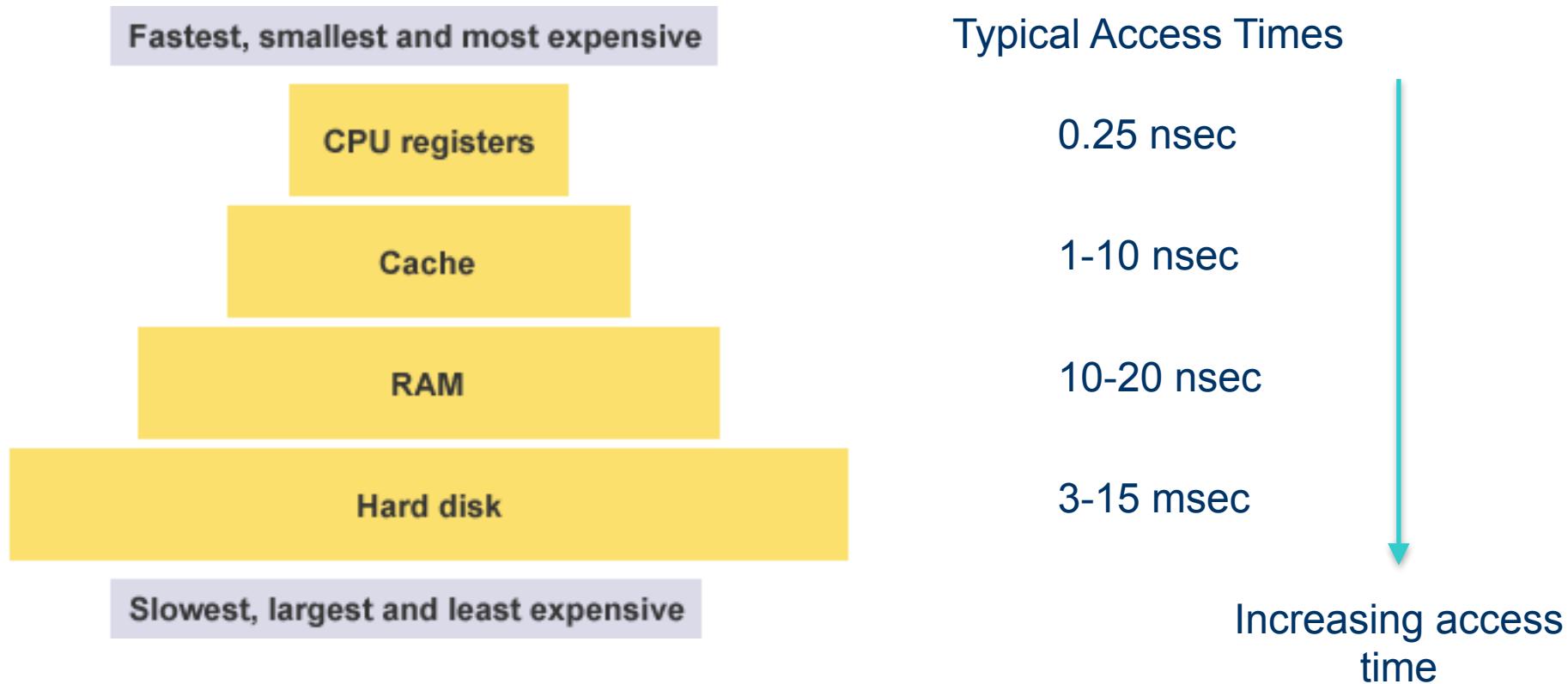
Memory Hierarchy

Ideally one would desire an **indefinitely large memory capacity** such that any particular ... word would be immediately available. ... We are ... forced to recognize the possibility of constructing **a hierarchy of memories**, each of which has **greater capacity** than the preceding but which is **less quickly accessible**.

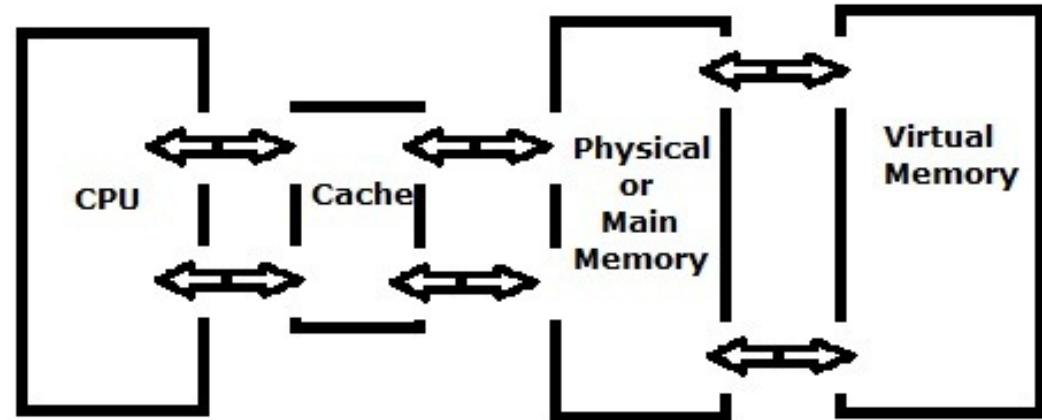
A. W. Burks, H. H. Goldstine, and J. von Neumann
Preliminary Discussion of the Logical Design of an Electronic Computing Instrument (1946)



4 types of primary memory



Aside: Virtual Memory



OS X and iOS include a fully-integrated virtual memory system.

Up to 4 GB of addressable space per 32-bit process.

OS X provides approximately 18 exabytes of addressable space for 64-bit processes.

Single process rarely gets this much RAM

- Even when it is available

Backing store (AKA Swap):

System **pretends** to a process that 4GB (or 18EB) is available

Data not in current use stored on disk (backing store)

As this data is needed it is swapped in to RAM.

This really belongs in the OS module
but we need to mention it here.

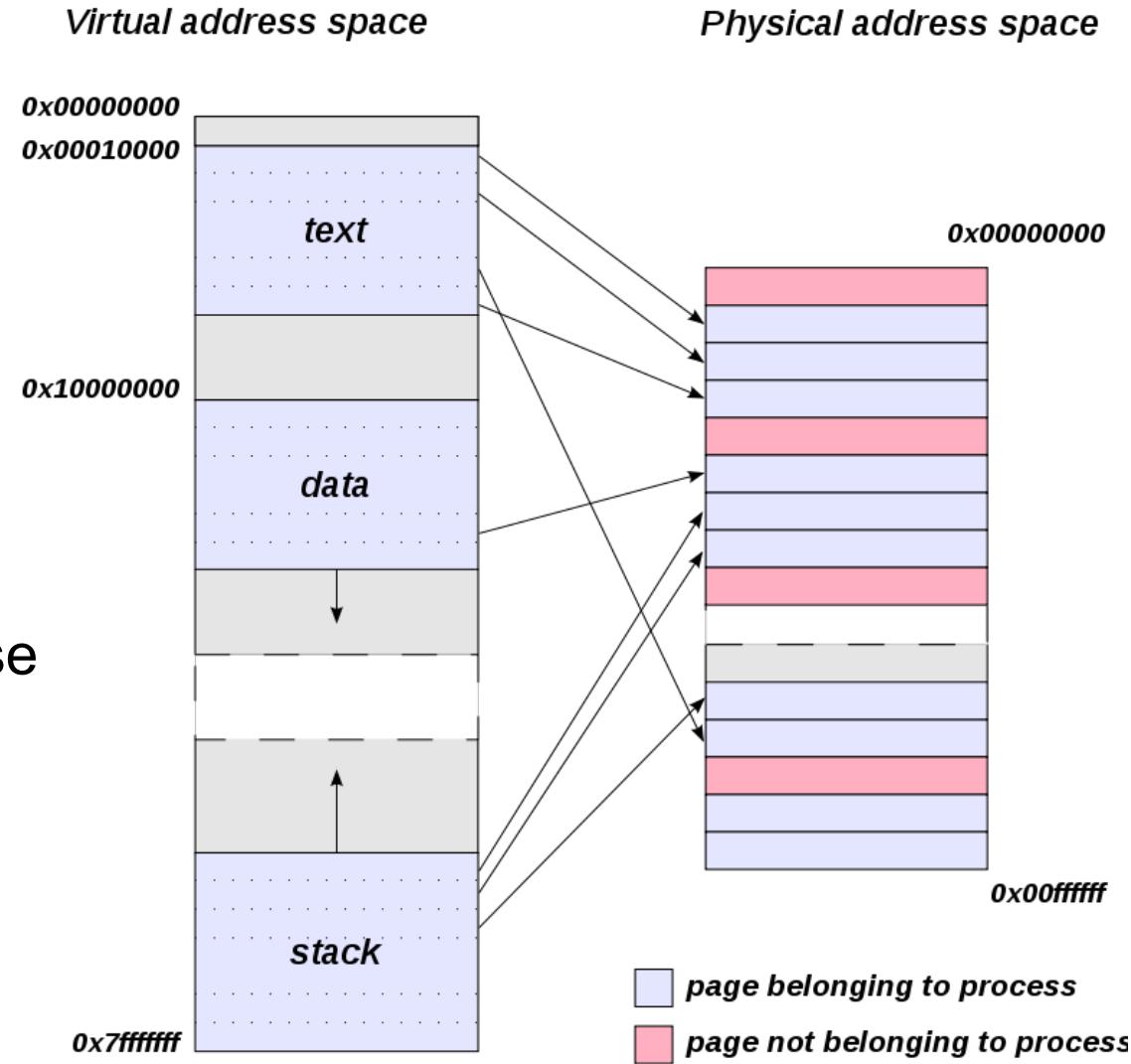
Aside: Virtual Memory

Text
program code

Data

Stack
remembering state
returning from function calls

Not all address space in use
need not be in RAM



MMU (Memory Management Unit)

Mac OS example

Activity Monitor

File Edit View Window Help

Activity Monitor (All Processes)

CPU Memory Energy Disk Network

Process Name % CPU CPU Time Threads Idle Wake Ups PID User

Process Name	% CPU	CPU Time	Threads	Idle Wake Ups	PID	User
sysmond	31.2	15.97	6	12	217	root
WindowServer	12.1	3:27:47.42	9	83	140	_windowserver
hidd	8.0	37:58.68	7	1	97	_hidd
kernel_task	6.3	1:59:41.74	162	182	0	root
Activity Monitor	5.4	1.73	5	2	38808	ladymead
expressvnd	1.1	3:25.02	17	69	68977	root
Google Chrome Helper	0.9	14.29	16	2	38586	ladymead
powerd	0.9	2:25.85	2	0	56	root
bluetoothd	0.9	5:36.82	3	0	96	root
Google Chrome	0.7	17:42.99	46	2	30736	ladymead
Google Chrome Helper	0.6	2:31.26	16	4	36175	ladymead
cfprefsd	0.6	1:24.93	3	1	243	ladymead
SystemUIServer	0.5	1:12.45	5	1	258	ladymead
Google Chrome Helper	0.5	15:44.01	15	33	30740	ladymead
Google Chrome Helper	0.4	4:49.74	13	5	30741	ladymead
Google Chrome Helper	0.4	24.83	14	5	30755	ladymead
Google Chrome Helper	0.3	2:01.50	16	3	33977	ladymead
Google Chrome Helper	0.3	49.06	16	3	37686	ladymead
UserEventAgent	0.2	1:46.65	3	0	244	ladymead
Finder	0.2	8:01.37	8	1	259	ladymead
Google Chrome Helper	0.2	29.05	16	3	32623	ladymead
Dropbox	0.2	34:49.81	187	1	37547	ladybox
systemsoundserverd	0.2	1.98	6	0	264	root
Dock	0.2	1:04.97	4	0	257	ladymead
fsevents	0.1	3:03.11	11	1	47	root
coreauthd	0.1	1:20.27	5	2	239	root
mds	0.1	6:07.34	8	2	66	root
trustd	0.1	5:57.40	4	0	252	ladymead
sharingd	0.1	2:37.92	4	1	315	ladymead
Google Chrome Helper	0.1	12.34	15	1	30961	ladymead
logd	0.1	2:02.57	3	1	59	root
notifyd	0.1	2:10.94	3	0	101	root
com.apple.PerformanceAnaly...	0.1	9.17	3	0	237	root
wirelessproxd	0.1	1:09.88	2	0	299	root

Activity Monitor (All Processes) Google Chrome (30736)

Parent Process: launchd (1) User: ladymead (501)

Process Group: Google Chrome (30736)

% CPU: 0.66 Recent hangs: 0

Memory Statistics Open Files and Ports

Real Memory Size: 277.6 MB

Virtual Memory Size: 5.16 GB

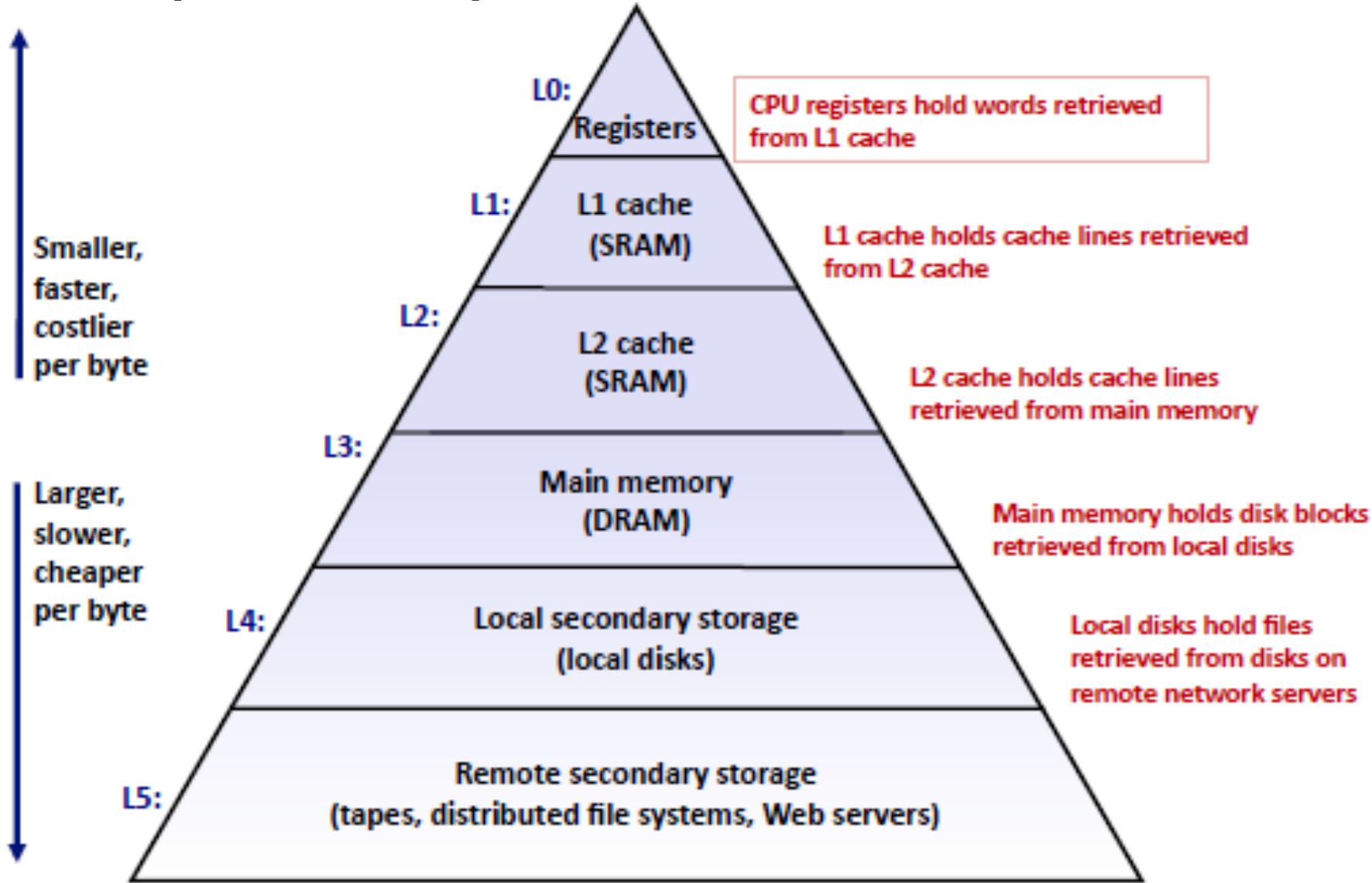
Shared Memory Size: 211.2 MB

Private Memory Size: 115.0 MB

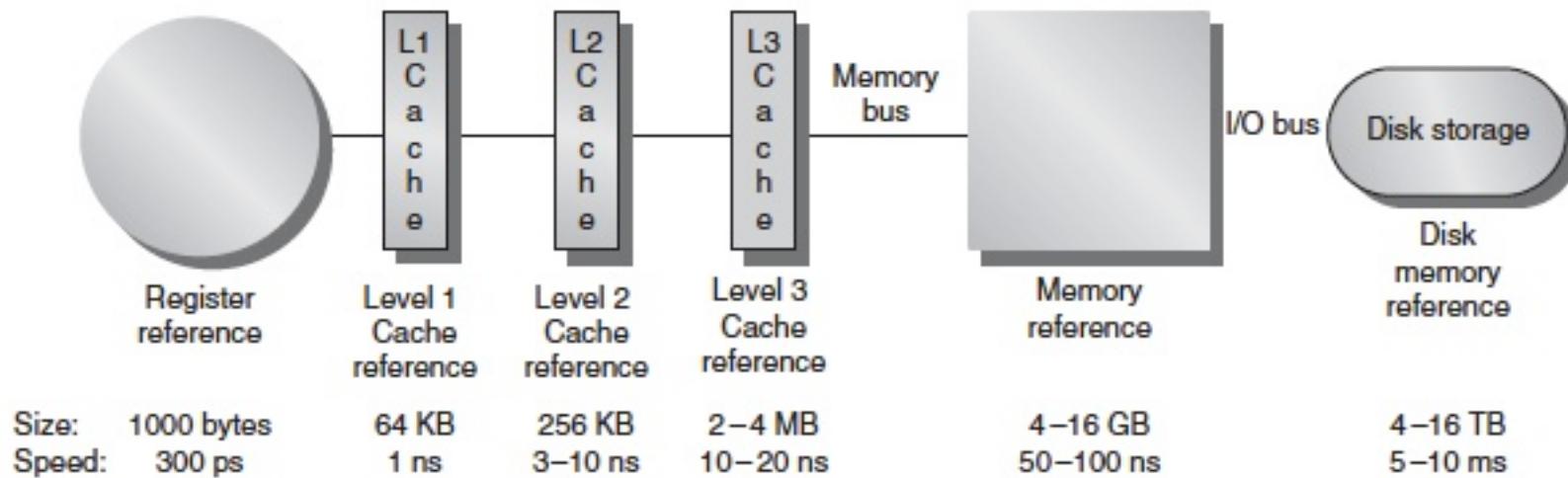
Sample Quit



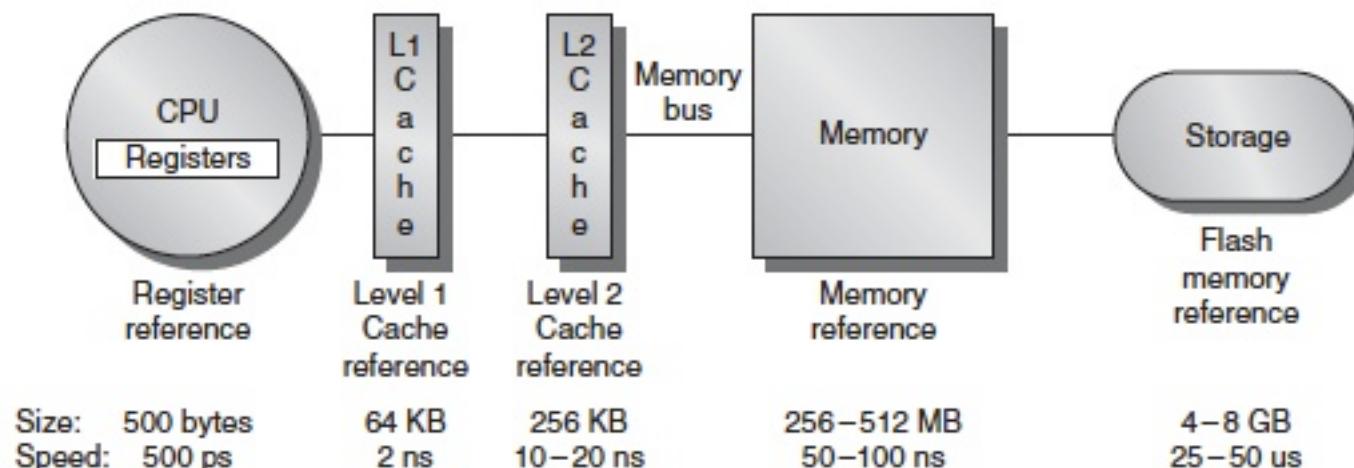
Memory Hierarchy



Memory Hierarchy



(a) Memory hierarchy for server



(b) Memory hierarchy for a personal mobile device

Examples

Intel® Core™ i7-5600U Processor – 2 cores

Level 1 cache: 2 x 32 KB instruction caches, 2 x 32 KB data caches

Level 2 cache: 2 x 256 KB caches

Level 3 cache: 4 MB shared cache

Apple A11 (iphone 8 and X) - 6 cores

L1 - 32 KB data and 32 KB instructions per-core

L2 – 8 MB shared

L3 – none

https://en.wikipedia.org/wiki/Apple_A11



Memory Hierarchy

To access a particular piece of data, the CPU first sends a request to its nearest memory.

If the data is not in that memory, then the next level in the memory hierarchy is queried.

If the data is not in main memory, then the request goes to disk.

Once the data is located, then the data and a number of its nearby data elements (a **Block**) are fetched into cache memory.

Blocks of memory from RAM(Disk) are **Mapped** into Block areas in the Cache area.



Cache

Cache used by the CPU to reduce the average time to access data.

Smaller, faster memory which stores copies of the data from frequently used main memory locations.

Caching in general – storing a copy of data to make future access requests faster

e.g. web cache



Locality

Cache memories are effective by virtue of the Principle of Locality.

Temporal locality. Recently-accessed data elements tend to be accessed again.

Spatial locality. Accesses tend to cluster; Nearby locations tend to be accessed soon.

Sequential locality. Instructions tend to be accessed sequentially.



Locality Example

```
total = 0;  
for (i = 0; i < n; i++)  
    total += a[i];  
return total;
```

Temporal: Variable total referenced on each iteration

Spatial: Array a[] accessed



Cache Memory Definitions (1 of 2)

A **Hit** is when data/instruction is found in Cache.

A **Miss** is when a reference to data/instruction is not in Cache.

The **Hit-ratio/Hit-rate** is the percentage of time data/instruction are found in Cache.

The **Miss-ratio/Miss-rate** is the percentage of time data/instructions are not in Cache.

$$\text{Miss-ratio} = 100 - \text{Hit-ratio}$$



Cache Memory Definitions (2 of 2)

The **Hit-time** is the time required to access Cache.

The **Miss-penalty** is the time required to process a miss, including the time that it takes to replace a block of memory plus the time it takes to deliver the data to the processor.



Average Memory Access Time

$$\text{AMAT} = \frac{\text{time for a hit}}{} + \text{miss rate} \times \text{miss penalty}$$

Calculate AMAT:

1GHz processor

time for a hit = (cache access time) 1 clock cycle

miss penalty = 20 clock cycles

miss rate = 5% (0.05)

Calculate AMAT if miss rate = 10%

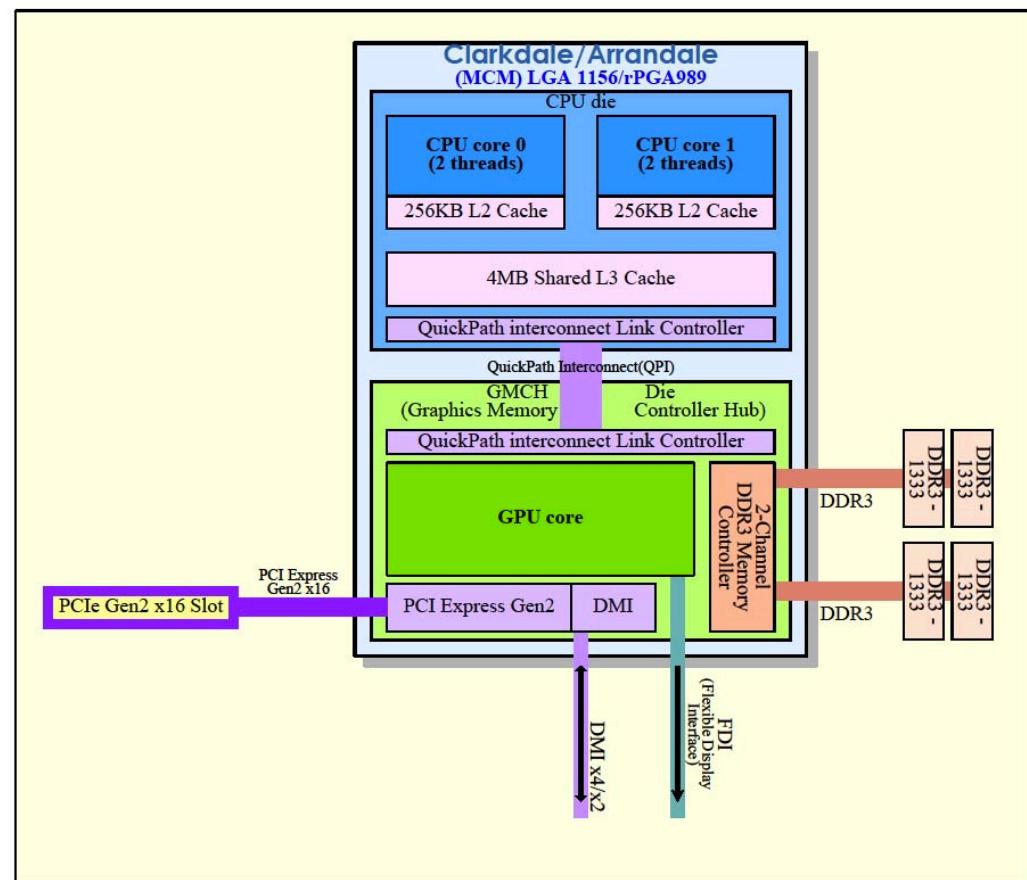
What miss rate will give an AMAT of 1.5ns?



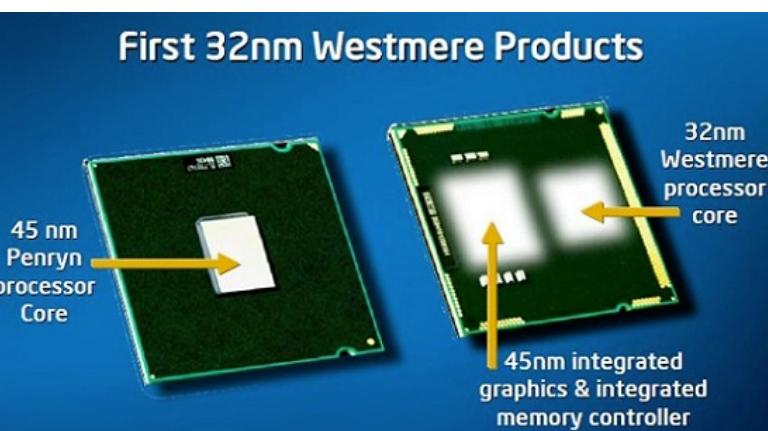
Dell Vostro 1540

Processor: Intel Core i3-370M, 2.4GHz
 Memory: 3GB 1,333MHz DDR3 RAM
 Graphics: Intel HD Graphics integrated
 Hard disk: 320GB
 Display: 15.6in 1,366x768, LED-backlit screen
 Features: 1.3 megapixel webcam, microphone
 Connectivity: 802.11g/n Wi-Fi, Gigabit Ethernet, Bluetooth 3.0
 Ports: 3 x USB2, 1 x VGA, 1 x HDMI, 1x 3.5mm headphone output, 1x 3.5mm microphone input, SD/MMC/MS card reader
 Dimensions: 376 x 260 x 33mm (WxDxH)
 Weight: 2.4kg

Clarkdale/Arrandale Architecture



Copyright (c) 2009 Hiroshige Goto All rights reserved.



Memory

On completion of this lecture you will be able to:

Explain the origins and consequences of the memory performance gap

Explain the different levels in the memory hierarchy
from registers and caches to virtual memory

Explain the concept of **locality** as it pertains to memory access

Quantify the impact of cache hits and misses

