Edinburgh Napier UNIVERSITY

Main (DRAM) Memory

Contents:

- DRAM organization
- Read timing
- Latency
- DRAM types (DDR and variants)

Introduction

Edinburgh Napier UNIVERSITY

- The main memory in a PC consists of DRAM (Dynamic Ram). This is memory that has to be refreshed at regular intervals.
- A number of chips are typically assembled onto a module (e.g. a DIMM: Dual In-line Memory Module). 8 chips, each providing 8 bits, will give the 64 bits needed by a Pentium.
- A memory controller (which probably also talks to the graphics system) interfaces to the memory.

Questions

Edinburgh Napier UNIVERSITY

- In the Apple II computer, the DRAM was clocked faster than the processor. Since then, DRAM clock rates have increased by about 9% per year, while processor clocks increased by about 50%. Why the difference?
- Processors have been tuned for speed, memories for capacity. Little attempt has been made to improve memory speed.
- What is the purpose of DRAM (where does its data go?)
- Cache (which needs a line at a time)

Edinburgh Napier UNIVERSITY

DRAM types

- The RAM in a PC is now a major bottleneck, working at a much slower rate than the processor. A lot of different types of RAM have been introduced to try and improve this situation.
- Cache is made from fast SRAM, but is too expensive for the PC's main memory, so DRAM is still used.
- It is organised so that the address is split into two halves: row and column.

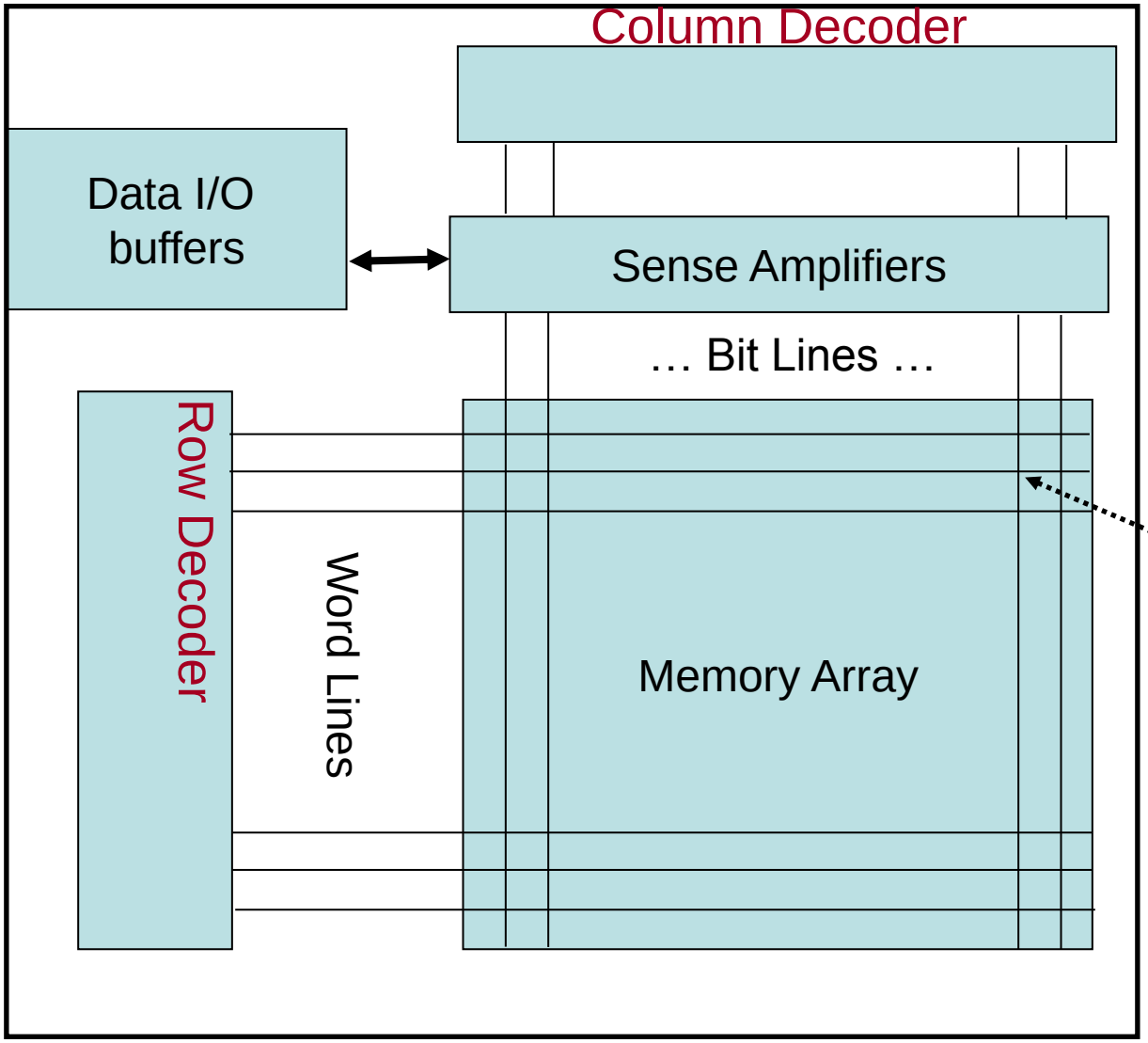
Edinburgh Napier UNIVERSITY

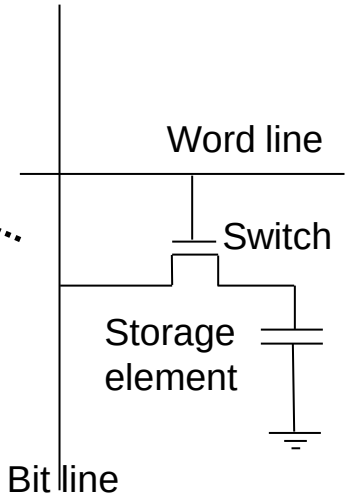
Row & Column

- A 64 M chip needs a 26 bit address. To save on pins, this is presented as two halves: a row, then a column address.
- Internally this is gated to the Row Decoder, then the Column Decoder.
- In a basic DRAM this cycle of Row, then Column address is repeated for every cycle.

Edinburgh Napier UNIVERSITY

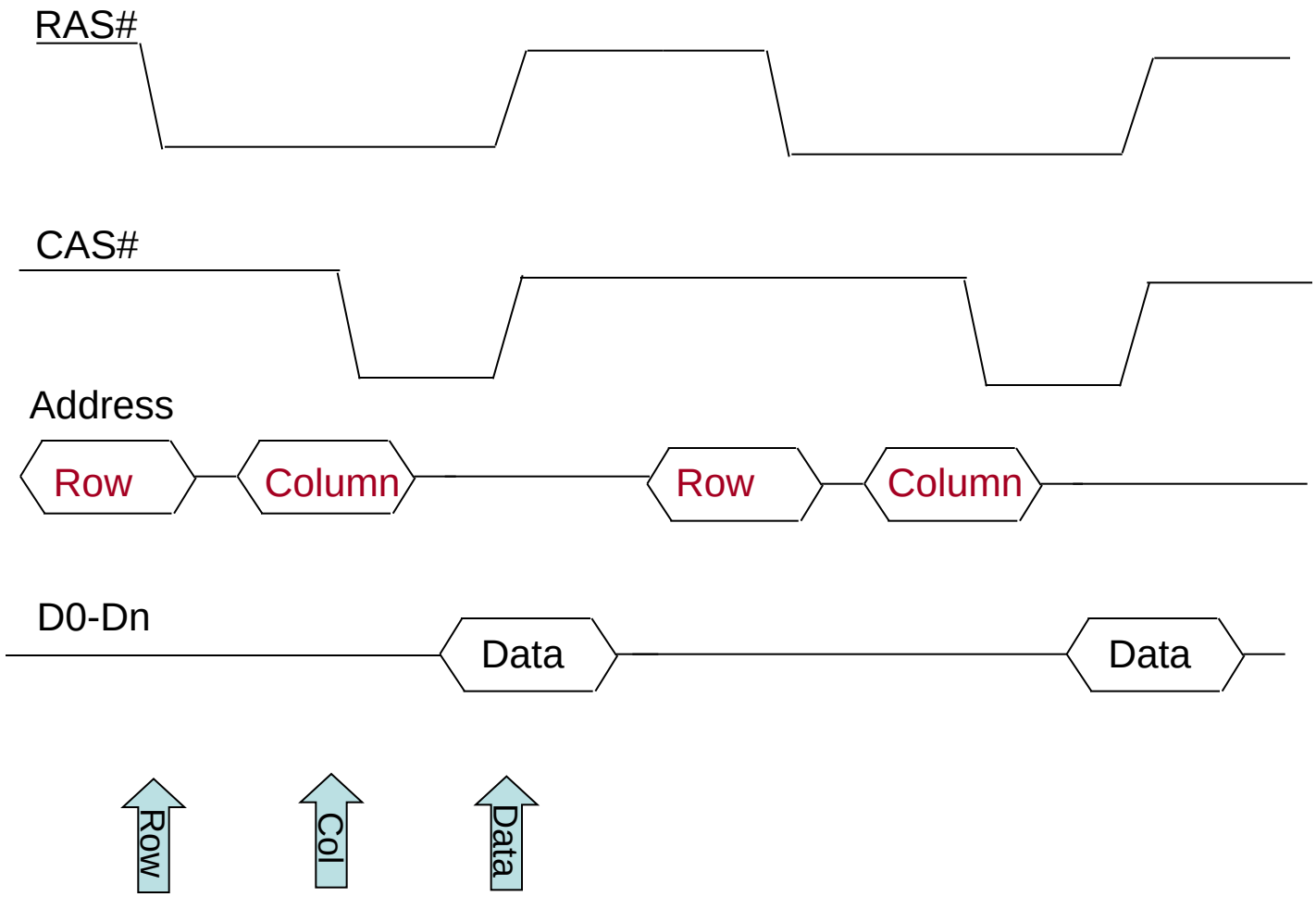
DRAM Organisation





Read timing for conventional DRAM

Edinburgh Napier UNIVERSITY



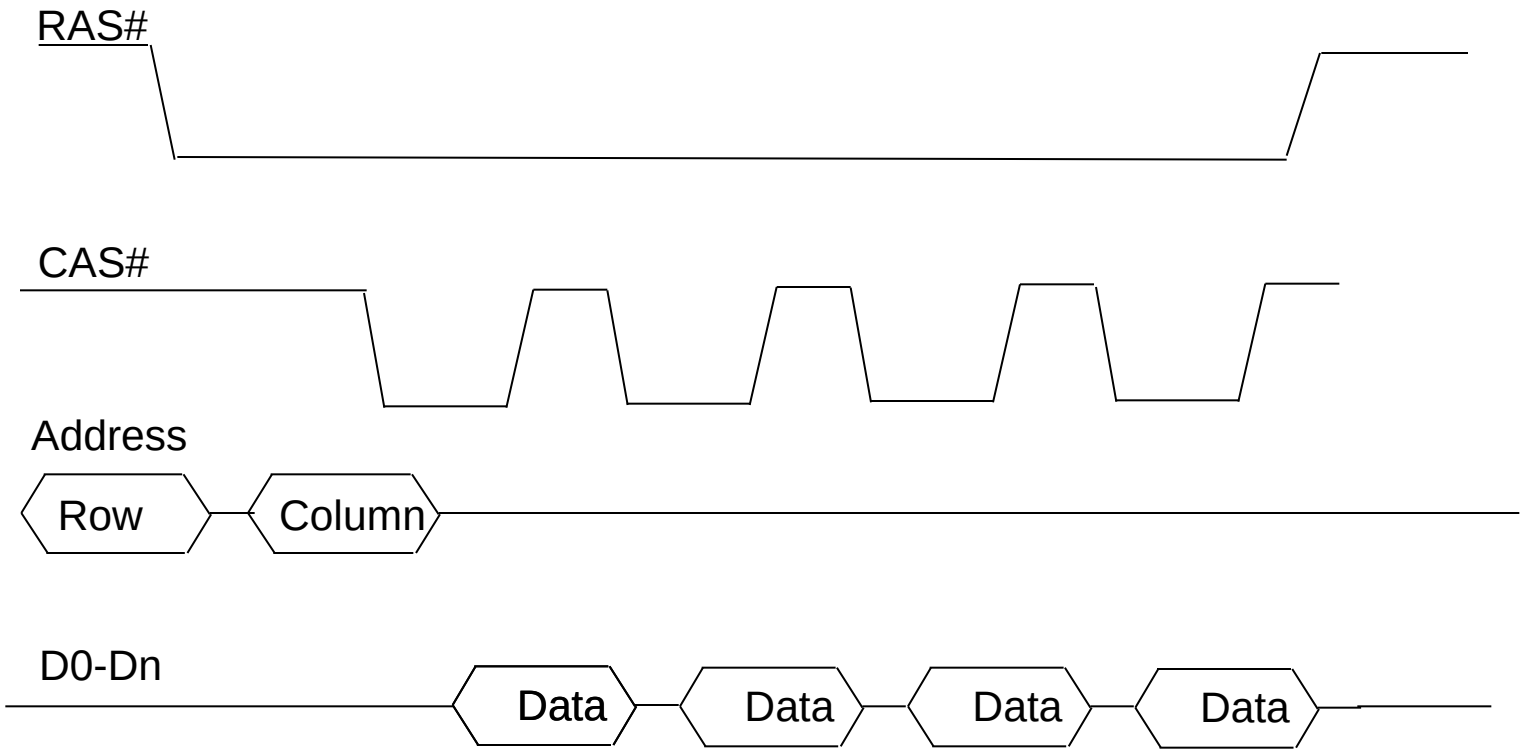
Extended Data Out

Edinburgh Napier UNIVERSITY

- Originally, EDO DRAMs were developed to speed things up by providing one row address, then a series of column addresses, one for each block of data.
- This was further developed for burst, then pipelined, EDO. This uses one column address, then pulses the Column Address Strobe (CAS#) line low to get subsequent blocks of data.
- These burst modes assume that data is located at consecutive addresses. This works fine for filling cache lines.

Pipeline Burst EDO timing

Edinburgh Napier UNIVERSITY



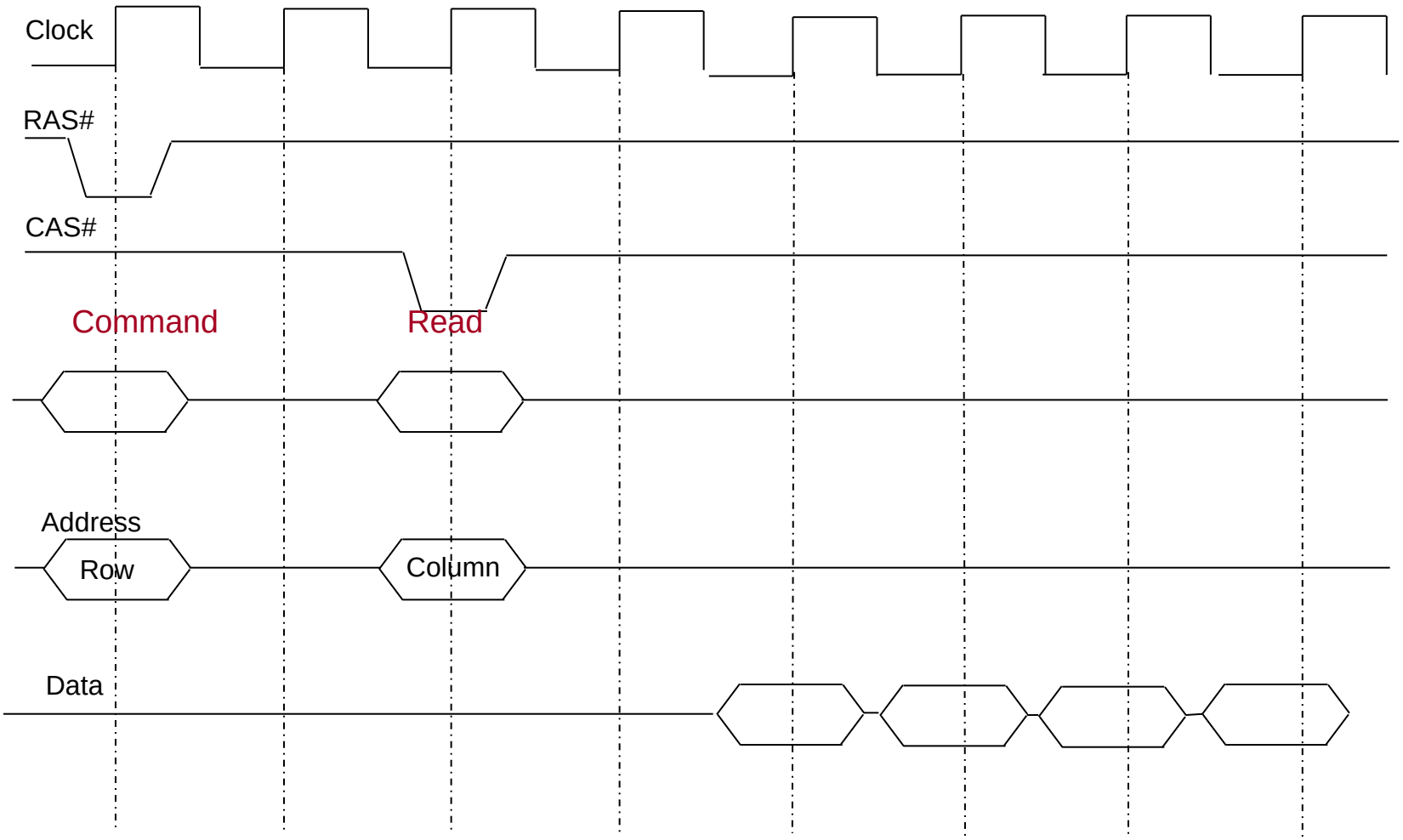
Edinburgh Napier UNIVERSITY

SDRAM

- The next major shift was to synchronise the transfer to a clock: Synchronous DRAM.
- The SDRAM chip is normally classified by the clock rate it can operate at in MHz (e.g. 800MHZ). So, a typical memory module might be a “2GB DDR2 800MHz” (£43 from Silicon Group, July 2010).

Synchronous DRAM read timing

Edinburgh Napier UNIVERSITY



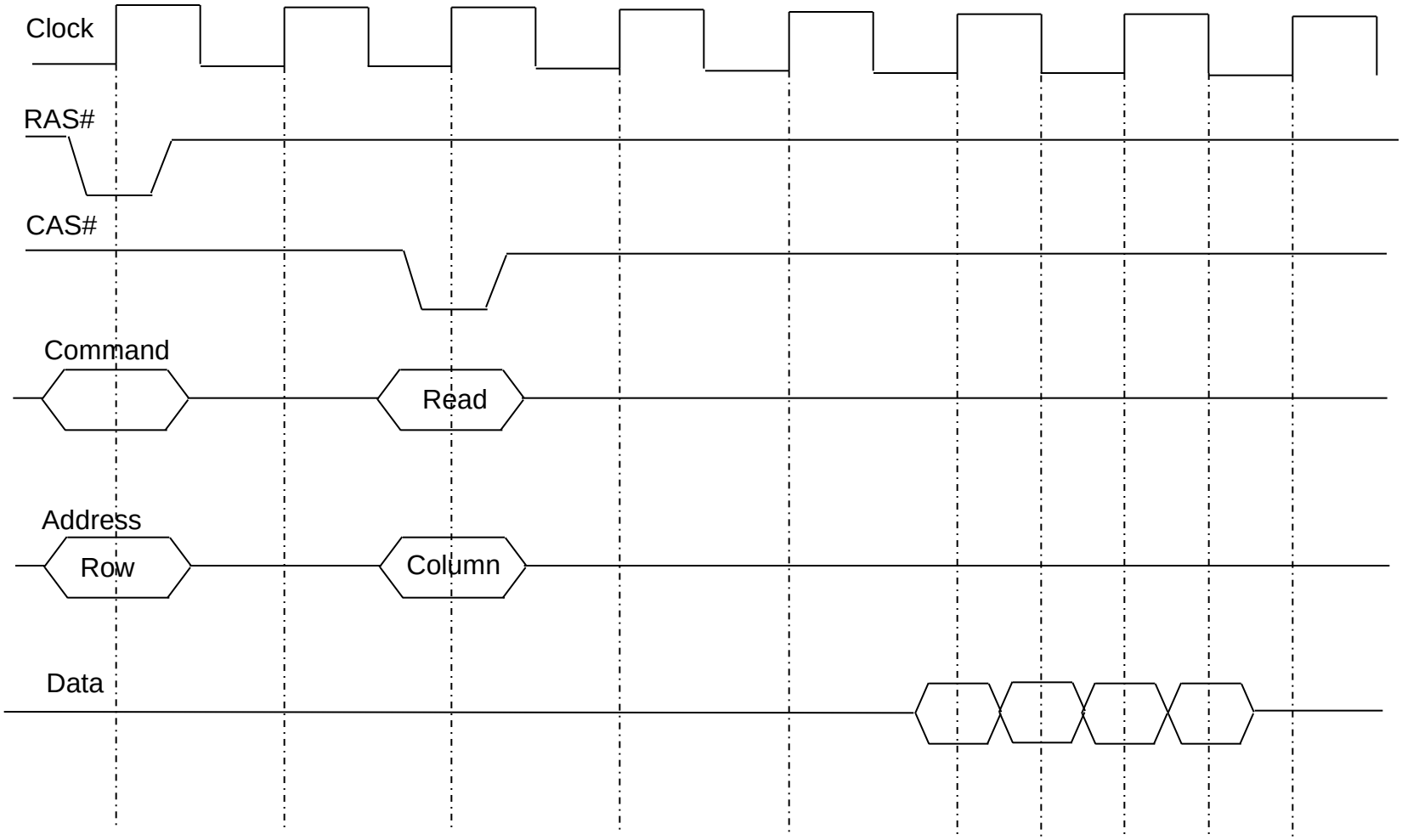
DDR: Double Data Rate

Edinburgh Napier UNIVERSITY

- Current versions of SDRAM can send data to the processor twice per clock cycle: ‘Double Data Rate’ or DDR SDRAM.
- Both edges of the clock cycle are used. So, ‘DDR667’ would be used with a 333 MHz basic clock.
- The naming scheme is different. Now it refers to the bandwidth: PC6400 means 6400 MB/sec (800 M transfers/sec * 8 bytes).
- You need to remember the 64 bit bus width, i.e. a maximum of 8 bytes per transfer.

DDR DRAM read timing

Edinburgh Napier UNIVERSITY



Latency

Edinburgh Napier UNIVERSITY

- It is not quite that case that the clock rate equals the rate that data can be read out.
- There is always a latency caused by the need to provide the row, then column, addresses. It may be quoted as ‘CAS latency’ (but actually there are several latencies to take into account).
- In a normal burst access, the first access will take perhaps 4 clocks. After that data is read on every clock.
- A typical burst might be 4-1-1-1, meaning the first block of data took 4 clocks, the next three each take one clock. (for the PC100, 4 blocks of data in 7 clocks = 70 nS).
- The CAS latency is not the only delay, but it is usually the most significant. There is also a precharge delay and a RAS to CAS delay (Trp & Trcd).
- Note that on the newest DDR2 memories, there may actually be more clock cycles in the latency, but the time is usually less because of the higher clock. A good memory now has a latency of about 10 nS.

Generation	Type	Data rate	Bit time	Command rate	Cycle time	CL	First word	Fourth word	Eighth word
SDRAM	PC100	100 MT/s	10 ns	100 MHz	10 ns	2	20 ns	50 ns	90 ns
	PC133	133 MT/s	7.5 ns	133 MHz	7.5 ns	3	22.5 ns	45 ns	75 ns
DDR SDRAM	DDR-333	333 MT/s	3 ns	166 MHz	6 ns	2.5	15 ns	24 ns	36 ns
						3	15 ns	22.5 ns	32.5 ns
	DDR-400	400 MT/s	2.5 ns	200 MHz	5 ns	2.5	12.5 ns	20 ns	30 ns
						2	10 ns	17.5 ns	27.5 ns
DDR2 SDRAM	DDR2-667	667 MT/s	1.5 ns	333 MHz	3 ns	5	15 ns	19.5 ns	25.5 ns
						4	12 ns	16.5 ns	22.5 ns
	DDR2-800	800 MT/s	1.25 ns	400 MHz	2.5 ns	6	15 ns	18.75 ns	23.75 ns
						5	12.5 ns	16.25 ns	21.25 ns
						4.5	11.25 ns	15 ns	20 ns
						4	10 ns	13.75 ns	18.75 ns
	DDR2-1066	1066 MT/s	0.95 ns	533 MHz	1.9 ns	7	13.13 ns	15.94 ns	19.69 ns
						6	11.25 ns	14.06 ns	17.81 ns
						5	9.38 ns	12.19 ns	15.94 ns
						4.5	8.44 ns	11.25 ns	15 ns
						4	7.5 ns	10.31 ns	14.06 ns
DDR3 SDRAM	DDR3-1066	1066 MT/s	0.9375 ns	533 MHz	1.875 ns	7	13.13 ns	15.95 ns	19.7 ns
	DDR3-1333	1333 MT/s	0.75 ns	666 MHz	1.5 ns	9	13.5 ns	15.75 ns	18.75 ns
						6	9 ns	11.25 ns	14.25 ns
	DDR3-1375	1375 MT/s	0.73 ns	687 MHz	1.5 ns	5	7.27 ns	9.45 ns	12.36 ns
	DDR3-1600	1600 MT/s	0.625 ns	800 MHz	1.25 ns	9	11.25 ns	13.125 ns	15.625 ns
						8	10 ns	11.875 ns	14.375 ns
						7	8.75 ns	10.625 ns	13.125 ns
						6	7.50 ns	9.375 ns	11.875 ns
	DDR3-2000	2000 MT/s	0.5 ns	1000 MHz	1 ns	10	10 ns	11.5 ns	13.5 ns
						9	9 ns	10.5 ns	12.5 ns
						8	8 ns	9.5 ns	11.5 ns
						7	7 ns	8.5 ns	10.5 ns

DDR2

- DDR2 produces data at double the clock rate, just as with DDR memory.
- The modules have a different number of pins, can run at faster clock rates and use less power (they run from 1.8V, DDR runs from 2.5V).
- DDR goes up to about DDR400, DDR2 starts at DDR2-400 and goes up to about DDR2-1300.

Edinburgh Napier UNIVERSITY

DDR3

- This is the new version of DDR. It still uses Double Data Rate, but this is clocked at higher rates and runs using less power.
- It covers data clocking at up to 1600 MHz (from a clock running at 800 MHz). This compares with DDR2's maximum of 1066 MHz.
- Latencies are high: can be up to 15 clocks for the CAS Latency.
- The voltage used (and therefore power requirement) has reduced to 1.5 Volts or less in DDR3.

Edinburgh Napier UNIVERSITY

Clock Rates/Bandwidth

- The memory is now referred to, either by the DDR clock rate (DDR800), or by the maximum throughput, (PC6400). As 8 bytes are transferred at a time, the value is 8 times the 'DDR' clock rate.
- Note that, because of latency, the actual throughput will be different. Remember also that the 'DDR' clock rate is double the actual clock rate. For instance, DDR1600 uses an 800 MHz clock.
- As an example, a memory module could be classed as DDR3-1600. This means the peak transfer rate is 1600MHz, running from an 800 MHz clock. This might also be called PC3-12800 (8x the DDR figure, as 8 bytes are transferred at a time. If you want to work out the latency, (say 9 cycles) remember to use the 800MHz clock

Typical Specifications

- Part #: CT2KIT25664AC667 • DDR2 PC2-5300 • CL=5 • Unbuffered • NON-ECC • DDR2-667 • 1.8V • 256Meg x 64

- Part #: CT2KIT51272AB80E • DDR2 PC2-6400 • CL=5 • Registered • ECC • DDR2-800 • 1.8V • 512Meg x 72

- Part #: CT2KIT12864BA1339 • DDR3 PC3-10600 • CL=9 • Unbuffered • NON-ECC • DDR3-1333 • 1.5V • 128Meg x 64

(NB the number of clock delays are the real clock cycles. So, for the middle memory, CL = 5 refers to 5 cycles of the 400MHz clock)

Edinburgh Napier UNIVERSITY

Rambus (RDRAM)

- Rambus is a very high-performance, but expensive technology.
- Intel and others (e.g. HP) have used it for high end servers. It has also been used in the PS2.
- Intel have recently dropped it from their plans. It will continue to be used in specialised applications. The PS3 uses a successor the Rambus: XDR DRAM (eXtreme Data Rate DRAM).

Edinburgh Napier UNIVERSITY

GDDR

- Graphics cards often use specialised DRAM chips: GDDR2, GDDR3 or GDDR4. These are similar to DDR2 and DDR3.
- They are not identical. For instance, they aim to run faster, and may use a higher voltage to achieve this.
- This can lead to higher power consumption. Also they are generally more expensive.

Edinburgh Napier UNIVERSITY

A latency example

- A memory module is rated as being suitable for PC5333. The sheet specifies a CL (CAS Latency) of 9. What access time does this equate to?
- PC5333 is the data rate in MB/sec. This is the same as DDR666.
- This uses a 333MHz clock.
- T= 1/f or 1/333M = 3nS.
- There are 9 clocks in CL, so 27 nS for Tcl.

Edinburgh Napier UNIVERSITY

Cache Memories

Contents:

- Cache organization
- Cache read and write strategies
- Cache placement strategies (associativity)

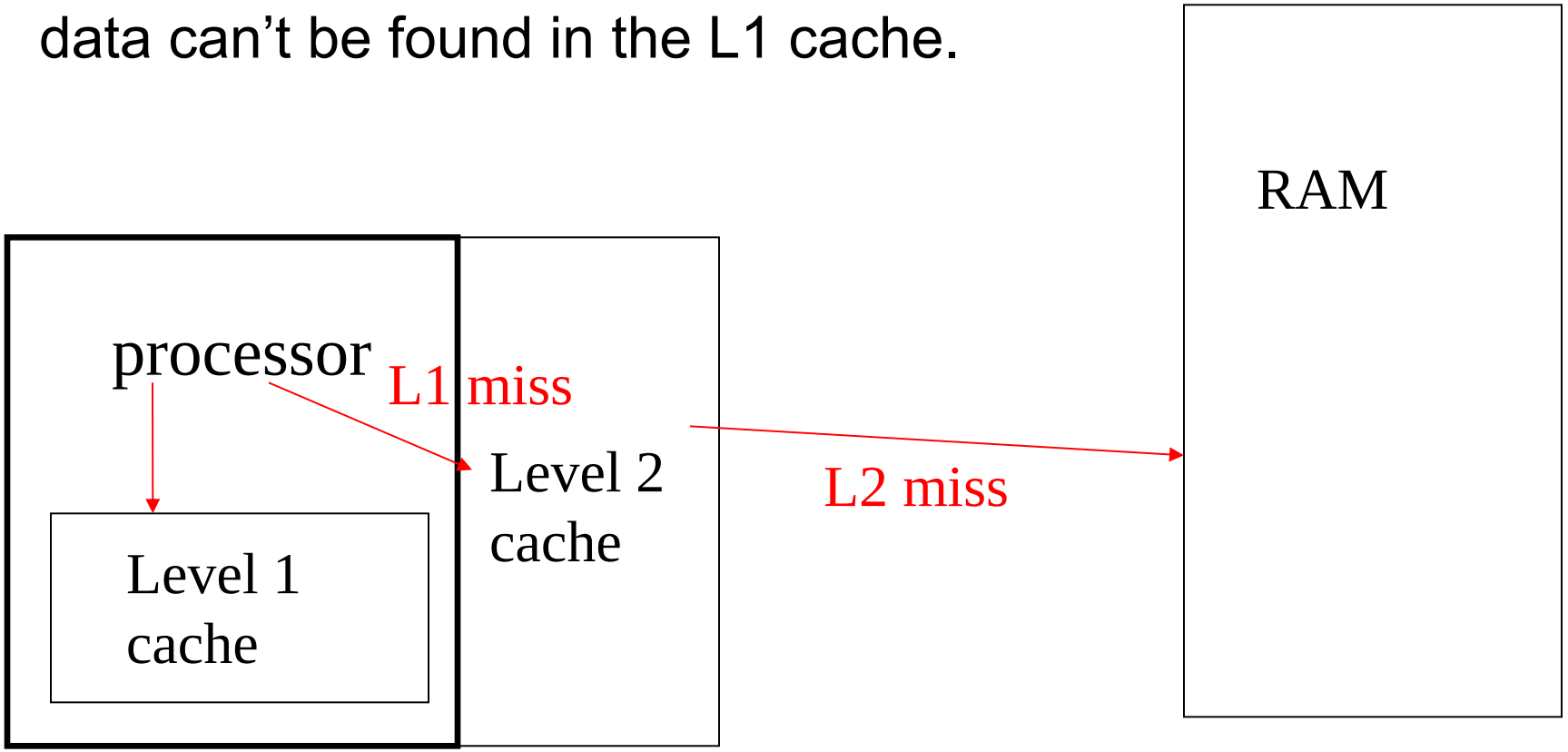
Edinburgh Napier UNIVERSITY

Introduction to cache

- Cache memory sits between the processor and Ram. With current processors, RAM is often too slow to work at the full clock speed of the processor.
- Many of the principles of cache memory are similar to those of virtual memory (principle of locality, replacement strategies).
- The key difference is that cache has to work in something like 5-10 nanoseconds. There is no time to run software; everything must be done in hardware. So, unlike virtual memory, the operating system is not involved.

Edinburgh Napier UNIVERSITY

PC cache is usually organised in 2 layers:
L1 Cache is inside the processor core; it is searched first
L2 cache is at the side of the processor; it is searched if the data can't be found in the L1 cache.



- Cache memory is relatively expensive SRAM.
- Typically the cache is one or two percent of the size of the main RAM. The principle of locality ensures that this is sufficient for at least 80-90% of the memory accesses to go successfully to cache.
- On a high-end processor, there may be a small amount (perhaps 64 Kbytes) in the processor chip, with a larger amount (1 Mbytes or so) as a second-level cache. This may be built into a module with the processor, or on the same chip as the processor. It used to be on the motherboard beside the processor, but the delays of sending signals down copper tracks have become relatively too large.

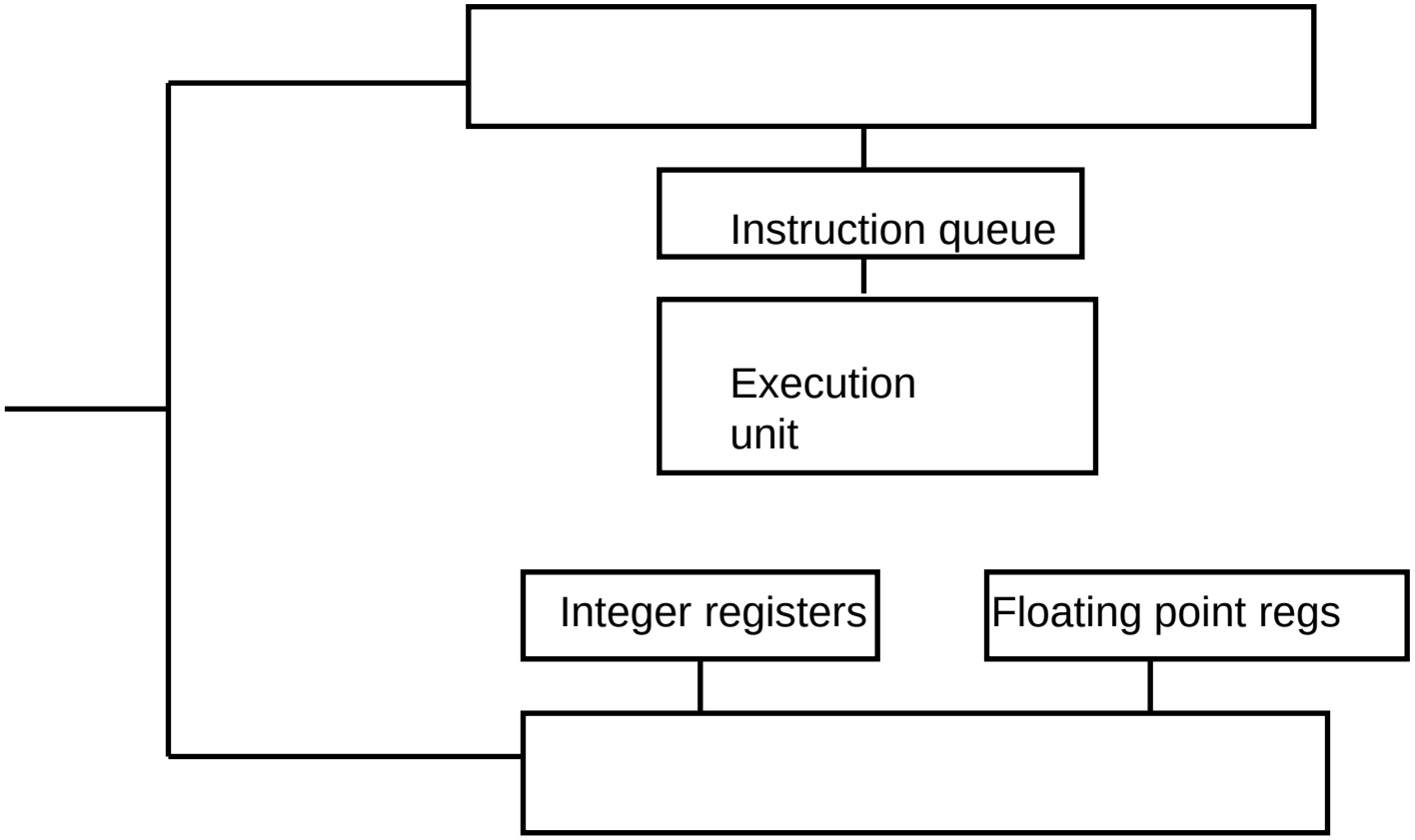
How fast does an electrical signal get along a pcb track in 1 nS? (hint: approx speed about 1/3rd of the speed of light).

$(1/3 \text{ X } 3.10^{8}).10^{-9} = 10^{-1}$ m = 10 cm

- The cache built into a processor is often split into separate sections for data and code; the processor can then fetch from both caches at the same time (Harvard architecture).
- Second level cache is usually organised as one block that will store code and data.
- Which block is I and which D in the following diagram?

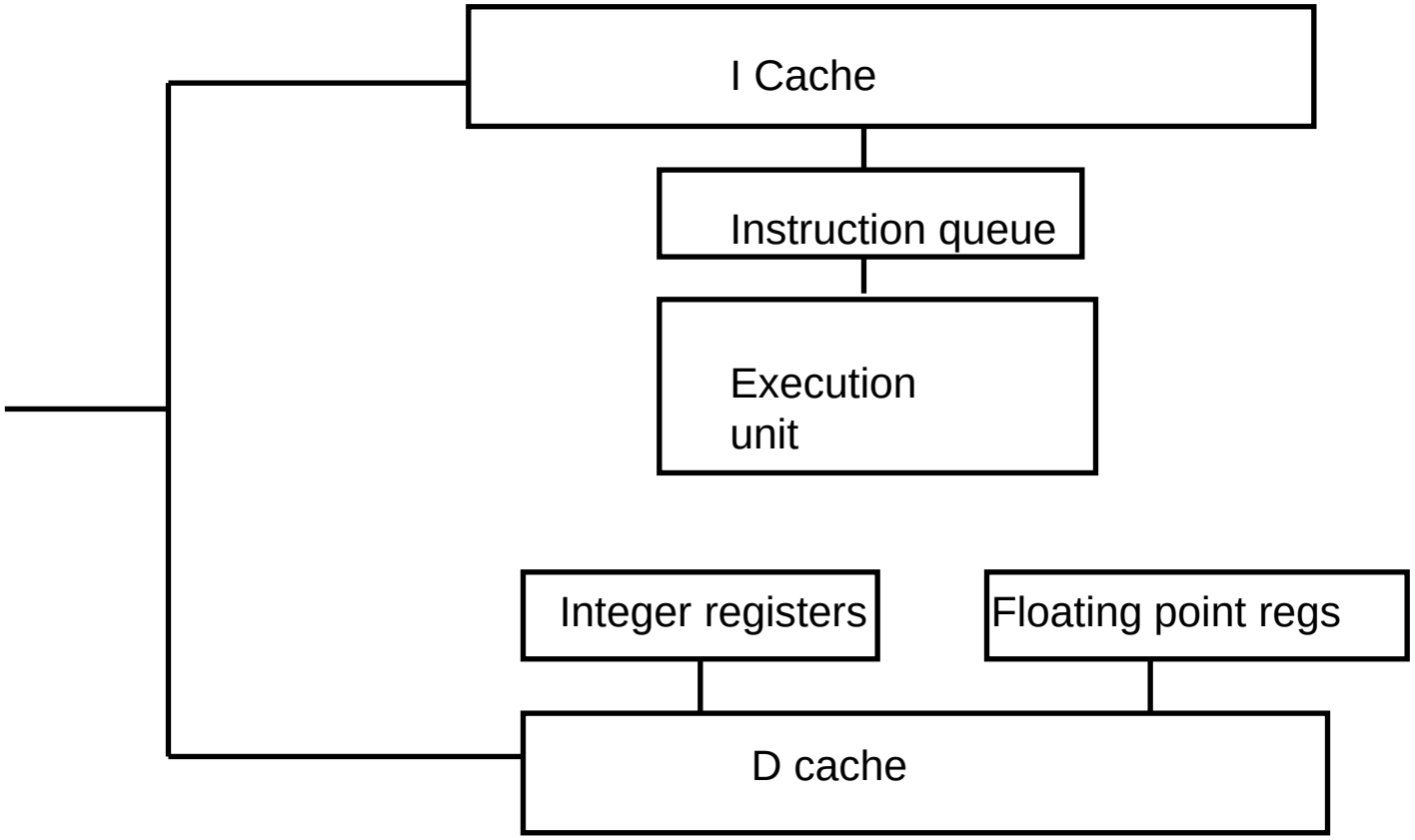
Cache inside a Pentium

Edinburgh Napier UNIVERSITY



Cache inside a Pentium

Edinburgh Napier UNIVERSITY



Edinburgh Napier UNIVERSITY

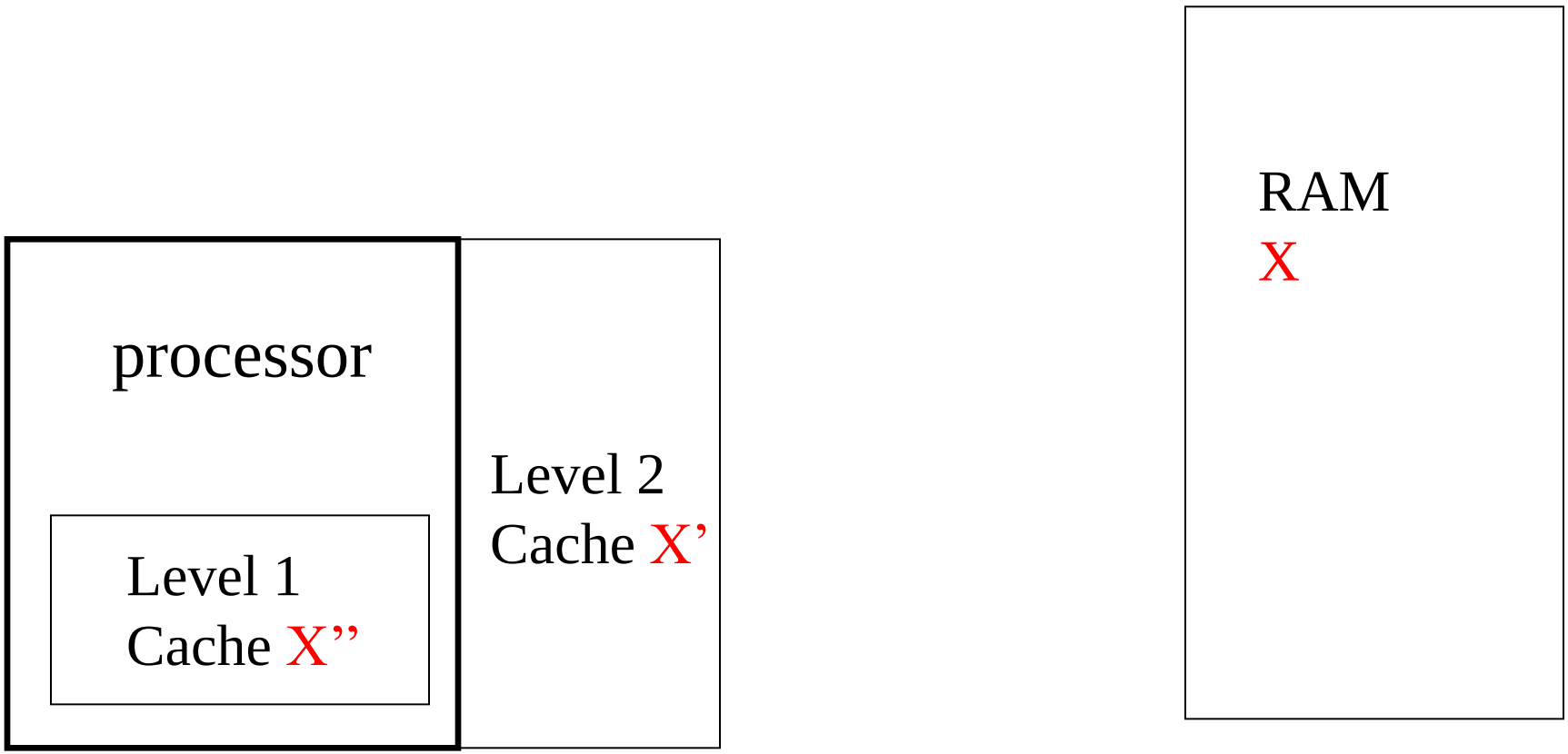
Cache reads and writes

- When the processor needs to read data, it first tries the cache.
- If it is there (a hit), the data can be read in quickly.
- If the data is not there (a miss) the processor has to get the data from the main RAM instead (obviously this slows the processor down). The data will be copied into the cache so that it has an up to date copy.

Cache writes

- When data is written to the cache, there are several ways of handling the transfer to the main RAM.
- The simplest to implement is to “write-through” the data, i.e. write the data to the RAM every time the data is written to the cache.
- This means that the RAM is kept up-to-date with the cache (handy if there are other processors using the bus that might need to see the contents of RAM).

- The disadvantage of this is that the RAM may be written to more than is strictly necessary.
- For instance, there may be a whole sequence of changes to a variable or flag that the RAM doesn’t need to know about if they are all happening to the variable/flag when it is in the cache.


processor
Level 1
Cache X''
Level 2
Cache X'
RAM
X

- The other technique, “write-back” or “copy-back”, only writes data to RAM if the cache is being emptied to make room for new data.
- This is more complicated to implement, but gives better performance.
- It does lead to possible problems if there are other processors (e.g. a DMA controller) that use the memory. They will need to know if the contents of RAM are up-to-date.
- In practice, the problem is not too great, since there are always more reads than writes. Question: why?

Cache organisation

Edinburgh Napier UNIVERSITY

- Cache is usually organised into lines/blocks, typically containing 8, 16 or 32 bytes.
- A line is written to/read from memory as one unit.
- Since cache is smaller than main memory, only some of the contents of the main memory are present in cache at any one time.
- Tag bits are used to identify what section of main memory is present in cache at any time. There are actually several distinct ways of organising the cache.

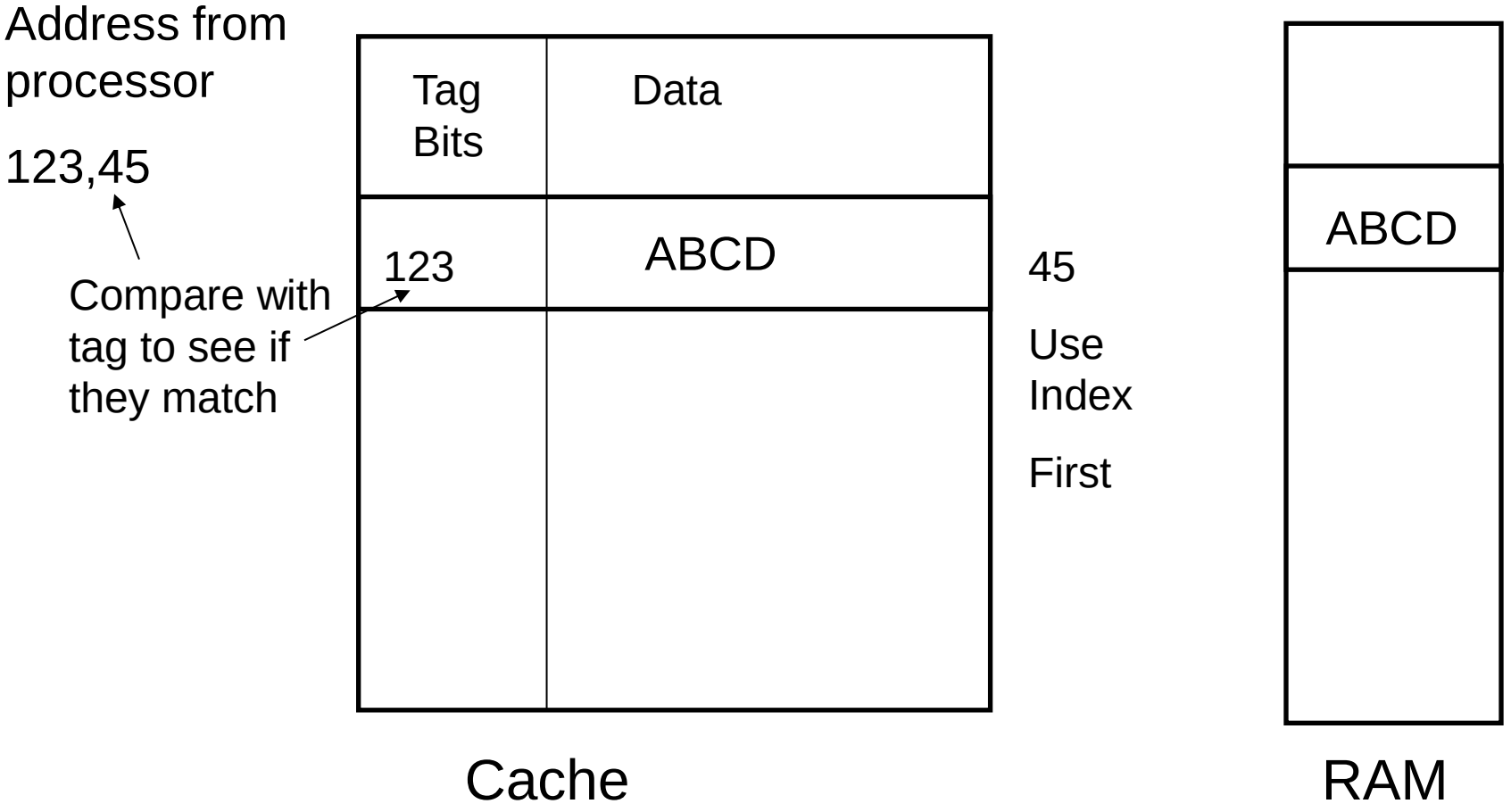
Edinburgh Napier UNIVERSITY

Direct-mapped cache

- This is the simplest way of organising cache: use some of the address to define exactly where the line goes.
- Each block of main memory can be mapped into only one line of the cache.
- The bottom address bits will be the same in main memory and the cache.

- Suppose an address of 12345 is generated. This will cause line 45 to be checked.
- This line of the cache could contain data from a number of locations (e.g. 11145, 22245, 32145 etc.) so the tag bits for line 45 are checked.
- If there is a match, then the data is available in the cache. If the tag bits do not match, then the data must be fetched from RAM .

Direct-mapped cache



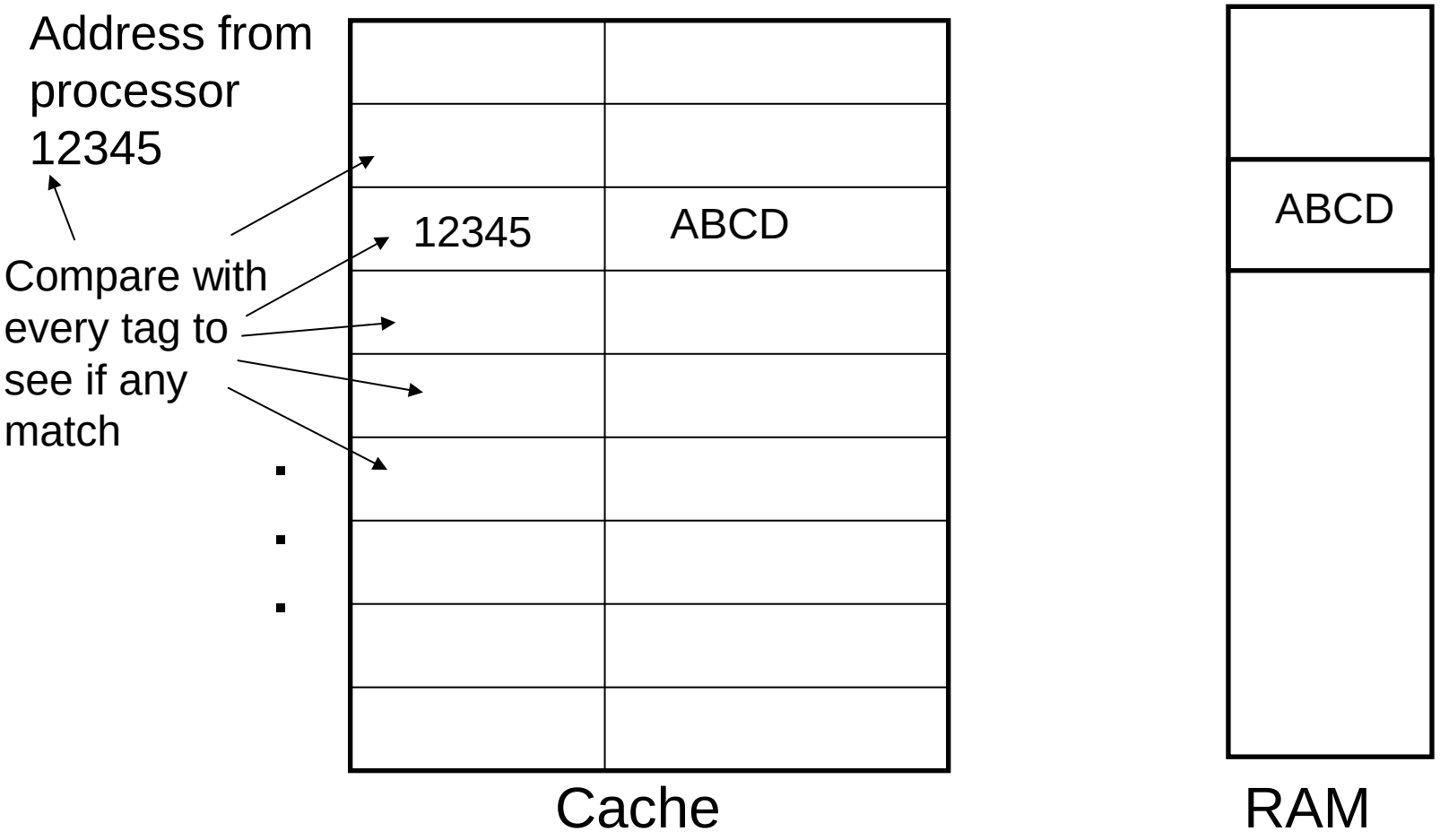
- Although this technique is simple, it does have an obvious disadvantage.
- If data from address 12345 is in the cache, then there is no where to store data from any other address that ends in 45.
- With there only being one possible place to store any block of data, the system is inflexible and the hit rate can be relatively low since there is a greater chance that the block that is about to be used will have been replaced with a different one.

Edinburgh Napier UNIVERSITY

Fully-Associative cache

- If data could be stored anywhere in cache, some of these problems would disappear.
- However, if data could be anywhere in cache, then the whole of cache must be searched simultaneously in order to find it.
- This is known as fully-associative mapping, and is a form of Content Addressable Memory

Fully associative cache



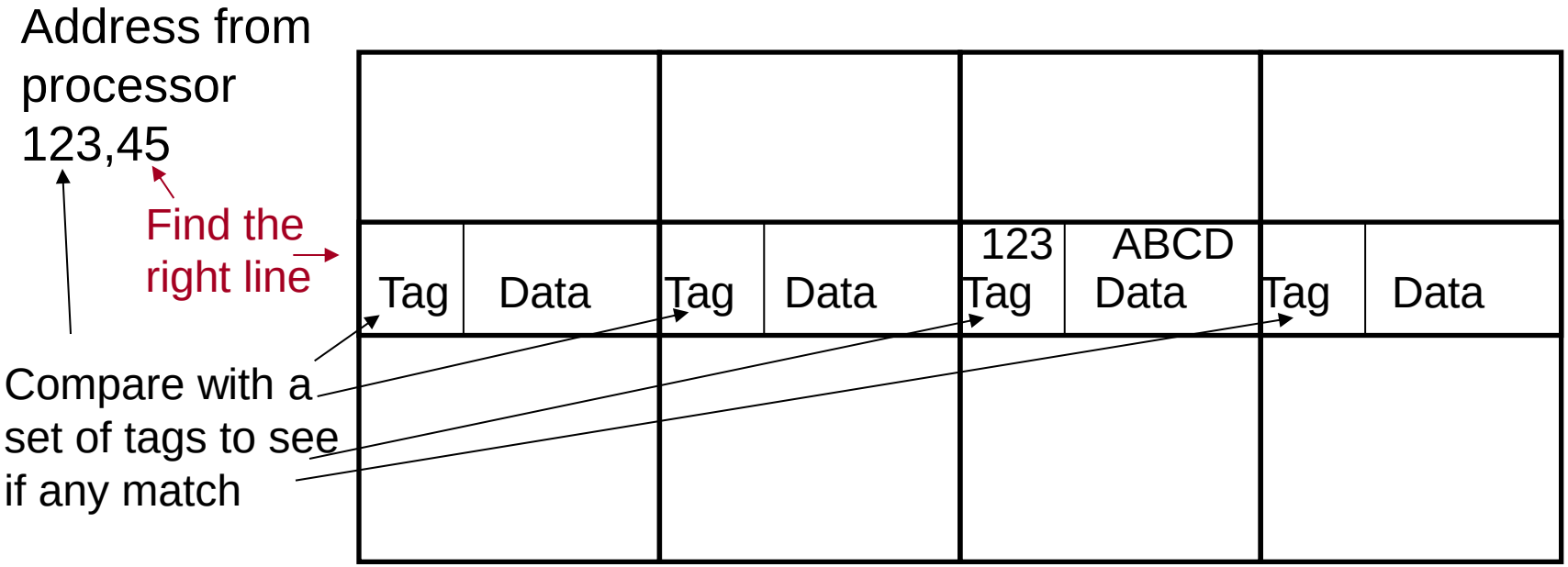
- In effect, the whole address is used for tag bits.
- The need for parallel hardware to search the whole cache simultaneously restricts this to small cache.
- Some microprocessors of a few years ago (Z8000, Z280) used this when cache sizes were smaller, typically 16 blocks of 16 bytes.
- A disadvantage is that there needs to be some way of implementing a replacement algorithm, since data can go anywhere. With Direct-mapped caches, data can only go in one location, so no replacement algorithm is needed.

Edinburgh Napier UNIVERSITY

Set-Associative cache

- This is a useful compromise that is now commonly used.
- There is more than one possible location for each block of data, but searching is kept simple by having a limit on the number of places data can be stored.

Set-Associative cache



- This example has four possible locations for each block of data.
- The index (e.g. 45 in the address 12345) is used to locate a ‘set’.
- This set has four blocks associated with it, so it could hold the data at 12345 as well as at AAA45 and 11145 and 32145.
- These four blocks have to be checked simultaneously to see if any of the tag bits match (i.e. to see if 123 can be found).

- Set-associative mapping is simpler to implement than fully associative mapping (only a few locations need to be searched, and a simpler replacement algorithm is needed).
- As there are more than one possible locations for data, there are fewer clashes than with direct-mapping.
- The more ‘ways’ there are, the fewer clashes there should be.
- In practice having a large number of ways adds little to the performance, and four/eight/sixteen ways are often used. The TLB in the Intel family uses four-way mapping.

Replacement Strategies

Edinburgh Napier UNIVERSITY

- Cache replacement is very similar to virtual memory replacement.
- When a read miss occurs, data has to be fetched from RAM. This has to be copied into cache, as the Principle of Locality suggests it is likely to be needed again soon.
- But, if cache is already full, some data will have to be over-written.
- If the replacement strategy is chosen well, then the data that is overwritten is the data that is least likely to be needed. Of course it is impossible to predict this exactly, so an algorithm based on age or usage is used.

- There is one difference between cache and virtual memory replacement. Cache replacement has to happen in nanoseconds (not milliseconds). This means that only simple hardware can be used.
- In practice, this restricts replacement algorithms to simple ones, such as FIFO or approximations to LRU. As these will be discussed in relation to virtual memory systems in the OS section, they are not further dealt with here.
- There are many similarities between cache and virtual memory. The basic principle is: look in the smaller, faster memory first, then go to the slower, larger memory only when you have to.