COMS12200 lecture: week #7

- ▶ Problem #1: we have latch (or flip-flop) based registers, but they aren't addressable.
 - An address (or index) allows dynamic rather than static reference to some stored datum.
 - ▶ By rough analogy to a C program, we have the left-hand side

```
Listing (C)

1 int A0, A1, A2, A3;

2 3 A0 = 0;
4 A1 = 0;
5 A2 = 0;
6 A3 = 0;
```

```
Listing (C)

1 int A[ 4 ];
2 3 A[ 0 ] = 0;
4 A[ 1 ] = 0;
5 A[ 2 ] = 0;
6 A[ 3 ] = 0;
```

but want the right-hand side.

▶ Problem #2: latches and flip-flops need a (relatively) large number of transistors per-bit; to support large capacities, it can make sense to use different components.

Dan Page (page@cs.bris.ac.uk) © 2014-5 Computer Architecture (new format)

Slide 1 of 31

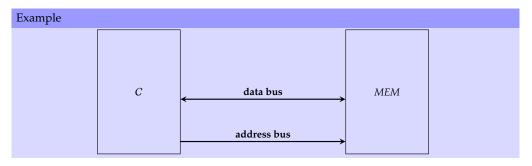


COMS12200 lecture: week #7

- ► There are various ways to classify a given memory component, e.g.,
- 1. Volatility
 - volatile, meaning the content is lost when the component is powered-off, or
 - non-volatile, meaning the content is retained even after the component is powered-off.
- 2. Access type
 - random versus constrained (e.g., sequential access to content),
 - Random Access Memory (RAM) which we can read from and write to, and
 - ▶ **Read Only Memory (ROM)** which, as suggested by the name, supports reads only.
- 3. Interface type
 - synchronous, where a clock or pre-determined timing information synchronises steps, or
 - asynchronous, where a protocol synchronises steps.

Notes:	
Notes:	

▶ We're (mainly) interested in a how a (volatile, synchronous) RAM component works:



► Goal: a concrete implementation of *MEM* with capacity for $n = 2^{n'}$ addressable words each of w bits (where $n \gg w$).

Dan Page (page@cs.bris.ac.uk) © 2014-5 Computer Architecture (new format)

Slide 3 of 31

University of BRISTOL

An Aside: History



- ► The EDSAC used **delay line** memory, where the rough idea is:
 - ► Each "line" is a tube of mercury (or something else in which sound waves propagate fairly slowly).
 - Put a speaker at one end to store sound waves into the line, and a microphone at the other to read them out.
 - Values are stored in the sense the corresponding waves take time to propagate; when they get to one end they are either replaced or fed back into the other.
- ► This is **sequential access** (cf. **random access**): you need to *wait* for the data you want to appear!

Notes:		

Notes:

An Aside: History



- ► The Whirlwind used magnetic-core memory, where the rough idea is:
 - The memory is a matrix of small magnetic rings, or "cores", which can be magnetically polarised to store values.
- Wires are threaded through the cores to control them, i.e., to store or read values.
- ► The magnetic polarisation is retained, so core memory is non-volatile!
- You might still hear main memory termed core memory (cf. core dump) which is a throw-back to this technology.

http://en.wikipedia.org/wiki/File:Project_Whirlwind_-_core_memory,_circa_1951_-_detail_1.JPG

Dan Page (page@cs.bris.ac.uk) © 2014-5
Computer Architecture (new format)

Slide 5 of 31

University of BRISTOL

Low-level Implementation (1) – SRAM and DRAM Cells

Definition (SRAM)

Static RAM (SRAM) is

- manufacturable in lower densities (i.e., smaller capacity),
- more expensive to manufacture,
- ► fast(er) access time (resp. lower access latency),
- easy(er) to interface with,
- ▶ ideal for use in **register files** and **cache memory**.

Definition (DRAM)

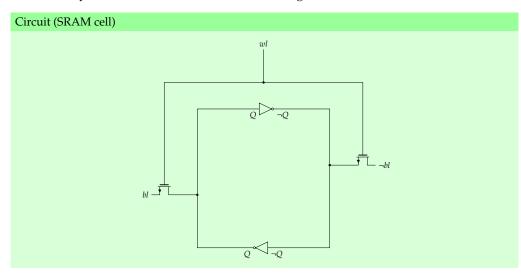
Dynamic RAM (DRAM) is

- manufacturable in higher densities (i.e., larger capacity),
- less expensive to manufacture,
- ▶ slow(er) access time (resp. higher access latency),
- hard(er) to interface with,
- ideal for use as main memory.

Notes:		
Notes:		

Low-level Implementation (2) – SRAM and DRAM Cells

► Abstractly, an **SRAM cell** resembles two NOT gates ...



• ... concretely, we can fill in the NOT gates with the transistor-level equivalents; each "6T SRAM cell" requires 6 transistors (cf. ~ 30 or 40 for a flip-flop).

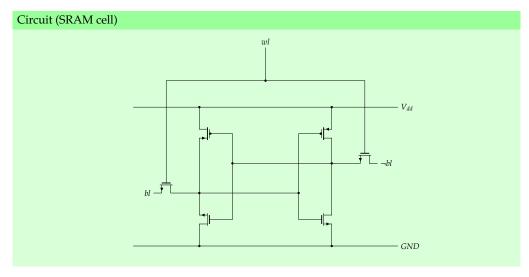
Dan Page (superior, brits, actual) © 2014-5

Computer Architecture (new format)

Slide 7 of 31

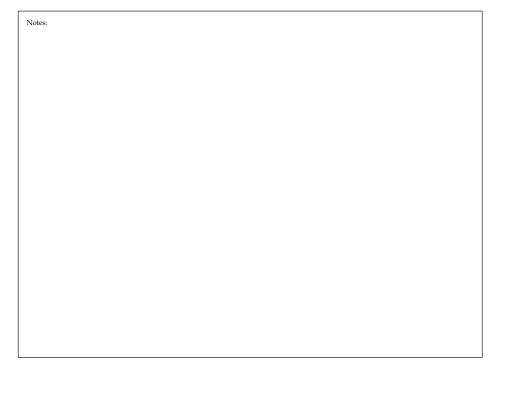
Low-level Implementation (2) – SRAM and DRAM Cells

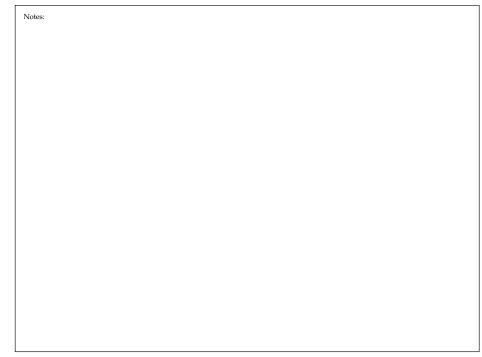
► Abstractly, an **SRAM cell** resembles two NOT gates ...



• ... concretely, we can fill in the NOT gates with the transistor-level equivalents; each "6T SRAM cell" requires 6 transistors (cf. ~ 30 or 40 for a flip-flop).

University of





Low-level Implementation (3) – SRAM and DRAM Cells

▶ Note that:

- ► The initial NOT-based circuit might look odd, but clearly has two stable states: either Q = 1 and $\neg Q = 0$, or Q = 0 and $\neg Q = 1$.
- ► The transistors re-enforce each other; the state is maintained as long as the cell is powered-on.
- \triangleright bl and \neg bl are termed the **bit lines**, wl is the **word line** which controls access to the state.
- ► The pre-charging steps are managed by extra **bit line conditioning** logic which we gloss over from here on.

► So

- ▶ to read the cell we pre-charge bl = 1 and $\neg bl = 1$ then set wl = 1, after which $\neg bl$ (resp. bl) is discharged if state is 1 (resp. 0), or
- to write x into the cell we pre-charge bl = x and $\neg bl = \neg x$ then set wl = 1, after which the state matches x.

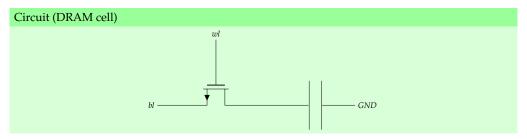


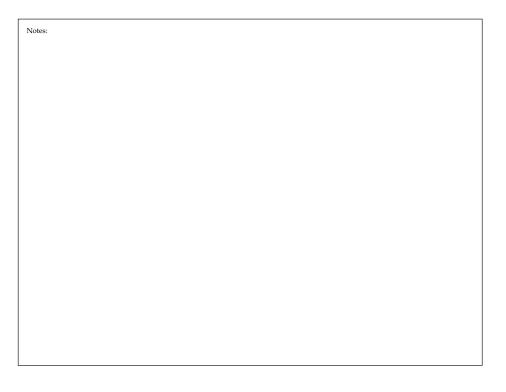
Slide 8 of 31



Low-level Implementation (4) – SRAM and DRAM Cells

▶ A **DRAM cell** is constructed using 1 transistor and a capacitor:





_		
١,	Notes:	
1		

Low-level Implementation (5) – SRAM and DRAM Cells

▶ Note that:

- The state decays when the cell is read, and also over time (even if it's not read) due to leakage; this implies a need to refresh the cell periodically.
- ► The capacitor holds a tiny charge: this must be amplified to use as a driver in whatever circuit uses the cell.
- ▶ The speed at which the capacitor charges and discharges is (relatively) slow.

► So

- to read the cell we set wl = 1, after which current flows (resp. does not flow) on bl if the capacitor is charged (resp. not charged) meaning the state is 1 (resp. 0), or
- by to write x into the cell we set bl = x then wl = 1, after which the capacitor charges (resp. discharges) and the state matches x.

Dan Page (page@cs.bris.ac.uk) © 2014-5

Slide 10 of 31

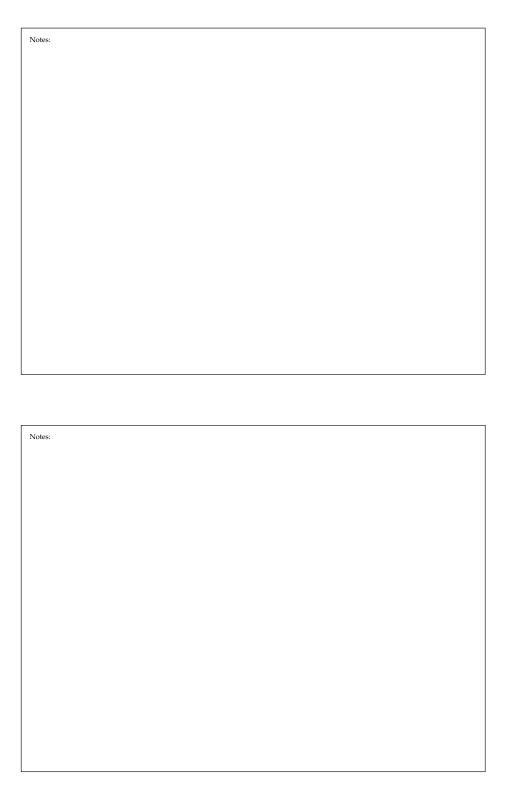


High(er)-level Implementation (1) – Cells → Device

- ► A **memory device** is constructed from (roughly) three components
- 1. a memory array (or matrix) of replicated cells with
 - r rows, and
 - c columns

meaning a $(r \cdot c)$ -cell capacity.

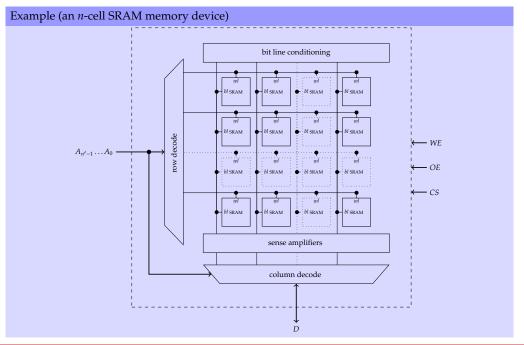
- 2. a row decoder which given an address (de)activates associated cells in that row, and
- 3. a **column decoder** which given an address (de)selects associated cells in that column plus additional logic to allow use (depending on cell type), e.g.,
- 1. **bit line conditioning** to ensure the bit lines are strong enough to be effective,
- 2. **sense amplifiers** to ensure output from the array is usable.

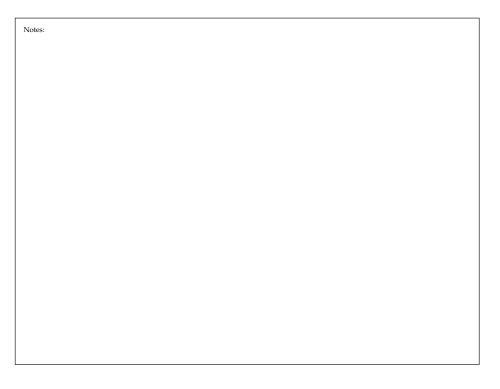


- A 1-bit **SRAM device** with $n = 2^{n'}$ cells has a (somewhat) standard physical interface:
- 1. auxiliary pin(s) for power and so on,
- 2. D, a single 1-bit **data pin** (sometimes split into two separate D_{in} and D_{out} pins),
- 3. A_i a collection of n' address pins where A_i is the i-th such pin,
- 4. a Chip Select (CS) pin, which enables the device,
- 5. a Output Enable (OE) pin, which signals the device is being read from, and
- 6. a Write Enable (WE) pin, which signals the device is being written to.



High(er)-level Implementation (3) – Cells → Device, SRAM







Algorithm (SRAM-READ)

Having performed the following steps

- 1. drive the address onto *A*,
- 2. set WE =false, OE =true and CS =true

1-bit of data is read and made available on D, then we set CS =**false**.

Algorithm (SRAM-WRITE)

Having performed the following steps

- 1. drive the address onto A,
- 2. drive the data onto *D*,
- 3. Set WE = true, OE = false and CS = true

1-bit of data is written, then we set CS =**false**.

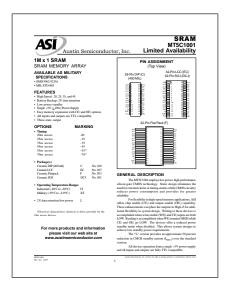
Oan Page (page@cs.bris.ac.uk) © 2014-

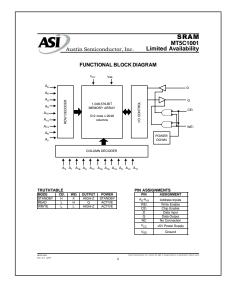
Slide 14 of 31



University of BRISTOL

High(er)-level Implementation (5) – Cells → Device, SRAM





Notes:

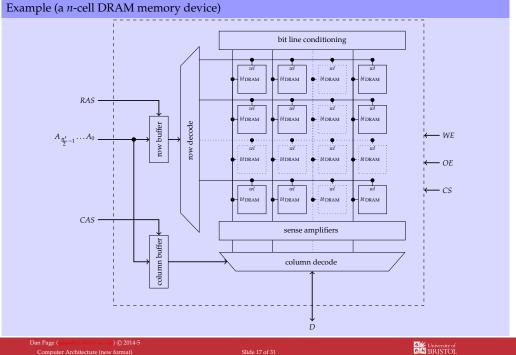


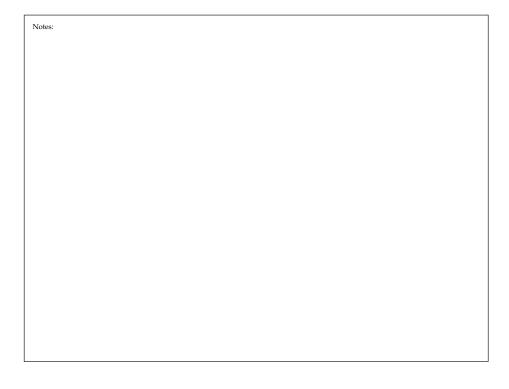
- A 1-bit **DRAM device** with $n = 2^{n'}$ cells has a (somewhat) standard physical interface:
 - auxiliary pin(s) for power and so on,
- 2. D, a single 1-bit **data pin** (sometimes split into two separate D_{in} and D_{out} pins),
- 3. *A*, a collection of $\frac{n'}{2}$ **address pins** where A_i is the *i*-th such pin,
- 4. a Chip Select (CS) pin, which enables the device,
- 5. a **Output Enable (OE)** pin, which signals the device is being read from,
- 6. a Write Enable (WE) pin, which signals the device is being written to,
- 7. a Row Address Strobe (RAS), which controls the row buffer, and
- 8. a **Column Address Strobe (CAS)**, which controls the column buffer.
- Note there are typically *half* the number of address pins versus the same sized SRAM device:

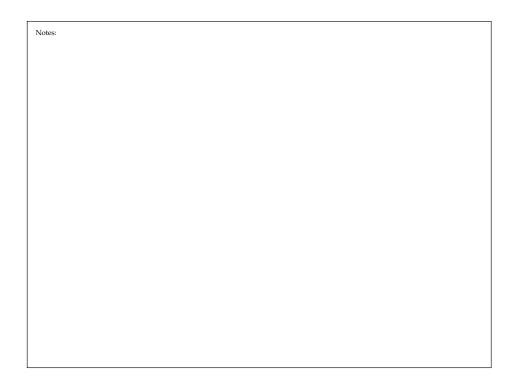
University of BRISTOL

- the pins are multiplexed to form a full address,
- since DRAM is more dense, this acts to manage the number of pins required.

High(er)-level Implementation (7) – Cells → Device, DRAM



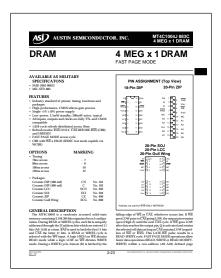


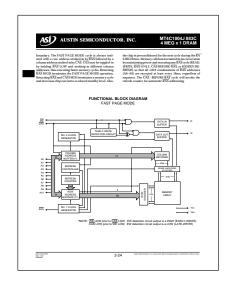


Algorithm (DRAM-READ)	Algorithm (DRAM-WRITE)
Having performed the following steps	Having performed the following steps
1. Drive the row address onto addr.	1. Drive the row address onto <i>addr</i> .
2. Set <i>RAS</i> = true , which latches row address.	2. Set $RAS = true$, which latches row address.
3. Drive the column address onto <i>addr</i> .	3. Drive the column address onto <i>addr</i> .
4. Set <i>CAS</i> = true , which latches column address.	4. Set <i>CAS</i> = true , which latches column address.
5. Set $WE = $ false, $OE = $ true and $CS = $ true.	5. Drive the data onto <i>data</i> .
1-bit of data is read and made available on <i>D</i> , and we	6. Set $WE = $ false, $OE = $ true and $CS = $ true.
set CSRAS = CAS = false.	1-bit of data is written, and we set CSRAS = CAS =
	false.

- ▶ Plus we need logic to implement DRAM refresh:
 - ▶ To cope with decay of cell content, we periodically read *all* the cells.
 - ▶ This amounts to activating each row in turn; the latency of refresh can therefore be optimised by selecting smaller r (resp. larger c) for the array.

High(er)-level Implementation (9) – Cells → Device, DRAM





Note	es:			



Externally, the configuration of a device is described as something like

$$\delta \times \omega \times \beta$$

(plus maybe some timing information) where

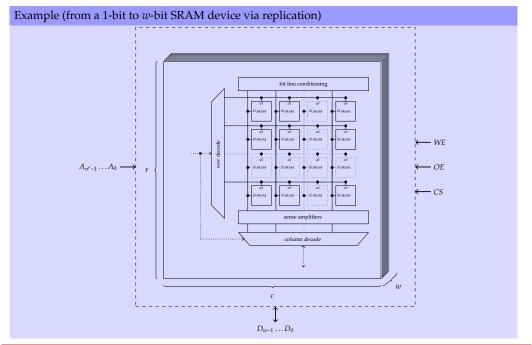
- ightharpoonup δ relates to capacity, usually measured in (large multiples of) bits,
- $\triangleright \omega$ describes the width of words, measured in bits,
- \triangleright β is the number of internal **logical banks**.
- ► Internally, this implies some organisational choices:
- 1. for $\omega > 1$, we replicate the memory device internally to give ω arrays (each copy relates to one bit of a ω -bit word),
- 2. for the arrays, *r* and *c* can be selected to match physical requirements (e.g., to get "square" or "thin" arrays), and
- 3. for $\beta > 1$, each array is split into logical banks

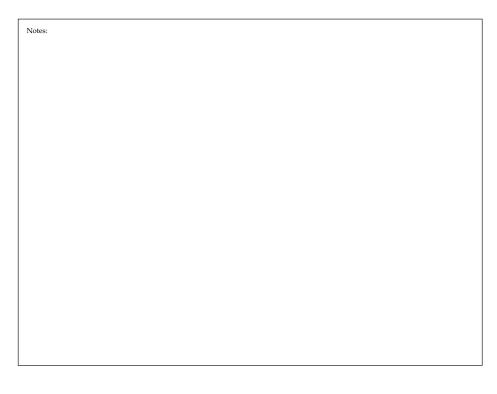
where in the latter case, the goal is to distribute the range of addresses across multiple smaller banks.

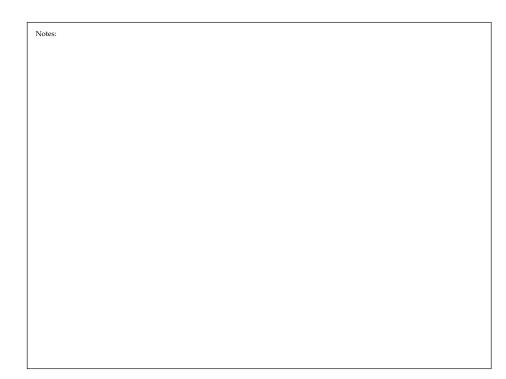
Dan Page (nage@cs.bris.ac.uk) © 2014-5

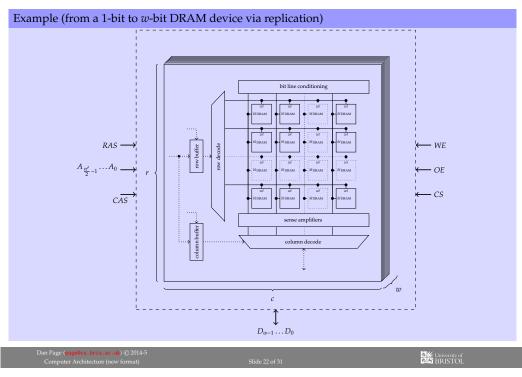
Computer Architecture (new format) Slide 20 of 31 BRISTOL

High(er)-level Implementation (11) – Cells → Device

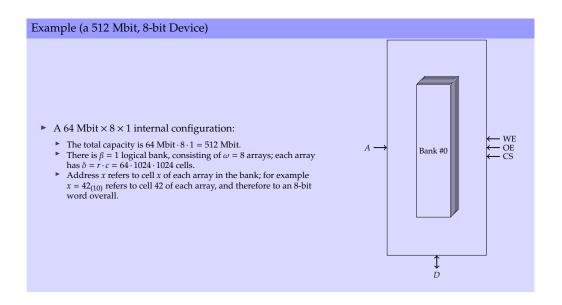






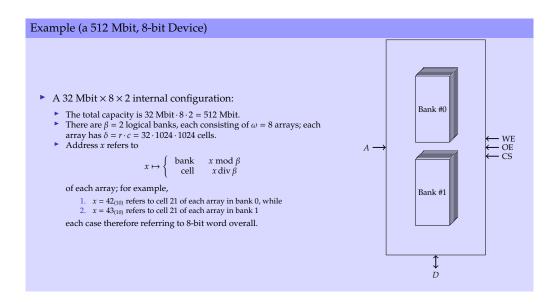


High(er)-level Implementation (13) – Cells → Device





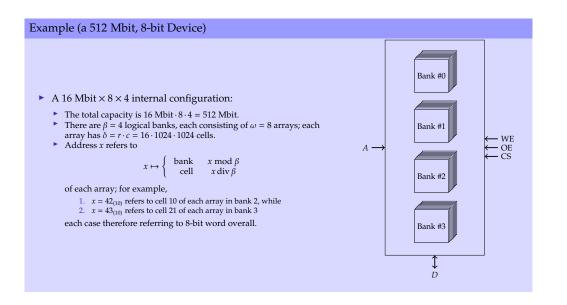




Dan Page (pagedcs.bris.ac.uk) © 2014-5

Computer Architecture (new format) Slide 23 of 31 □ BRISTOL

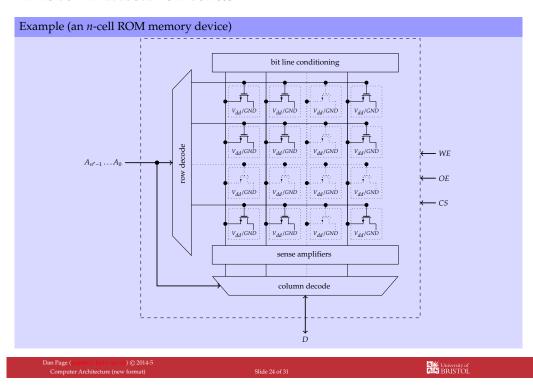
High(er)-level Implementation (13) – Cells → Device







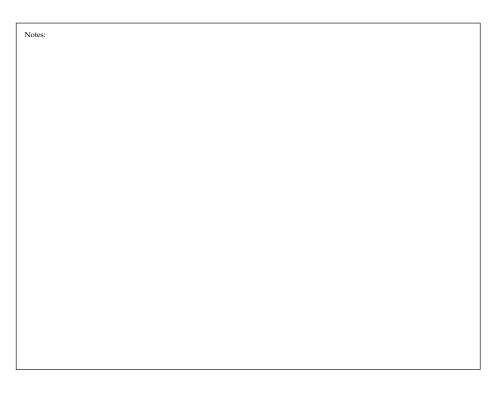
An Aside: "what about ROM devices?"



High-level Implementation (1) – Device → Module

- ► To make devices easier to use (or integrate), they are typically packaged into a **memory module**:
- 1. one or more memory devices, and
- 2. an interface which controls access.
- ► There are lots of package types; two dominate
 - 1. **Single Inline Memory Module (SIMM)**, which is (roughly) 1-sided, has less pins and a narrower word size, and
- 2. **Dual Inline Memory Module (DIMM)**, which is (roughly) 2-sided, has more pins and a wider word size

and are capable of housing different device types (e.g., EDO *or* SDRAM devices in a given DIMM package).



Notes:		

High-level Implementation (2) – Device → Module

▶ Externally, the configuration of a module is described as something like

$$\Delta \times \Omega$$

(plus maybe some timing information) where

- Δ relates to capacity, usually measured in (large multiplies of) bytes,
- \triangleright Ω describes the width of words, measured in bits.
- ▶ Internally, some organisational choices exist
- 1. usually $\Omega > \omega$ so the module is "filled" using multiple devices to form one **physical bank**, and
- 2. depending on the module type, the devices are organised into one or more ranks

where in the former case, the goal is to distribute content relating to a single address across multiple smaller devices.

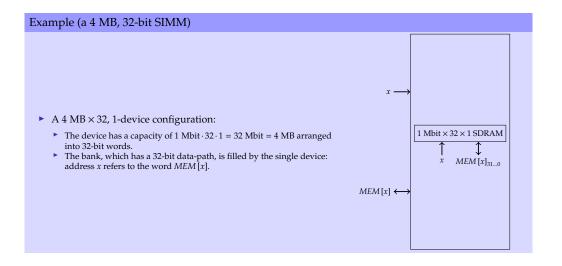


Slide 26 of 31



University of BRISTOL

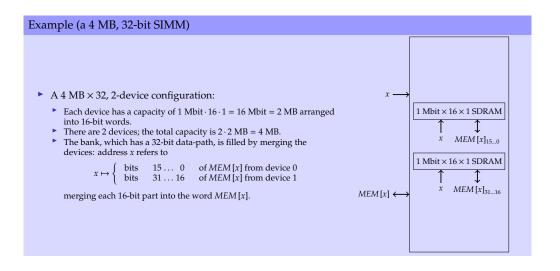
High-level Implementation (3) – Module → Bank





Notes:		

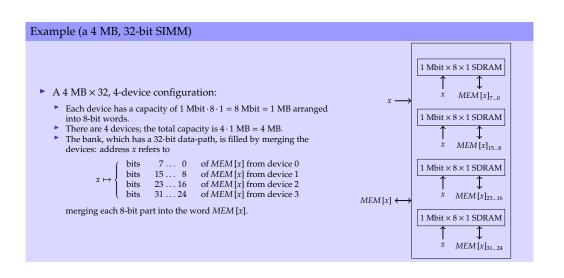
High-level Implementation (3) – Module → Bank

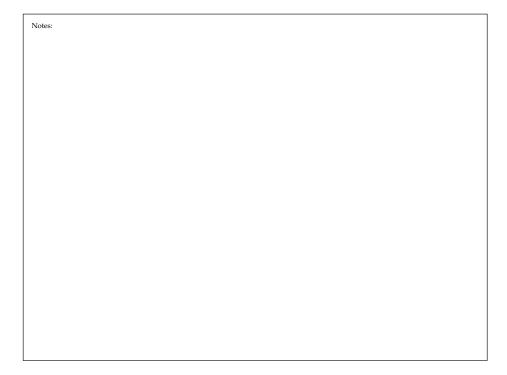


Dan Page (page@cs.bris.ac.uk) © 2014-5

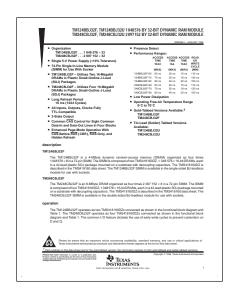
Computer Architecture (new format) Slide 27 of 31 € BRISTOL

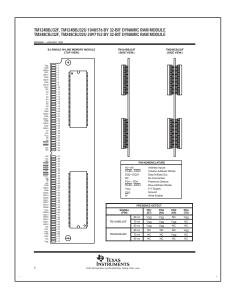
High-level Implementation (3) – Module \sim Bank





Notes:			





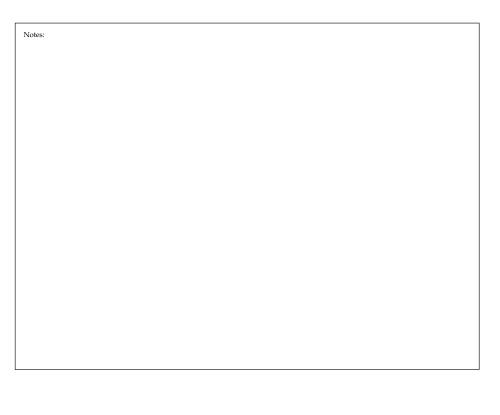
Dan Page (page@cs.bris.ac.uk) © 2014-5 Computer Architecture (new format)

Slide 28 of 31

University of BRISTOL

Conclusions

- ► Take away points:
- 1. The initial goal was an *n*-element memory of *w*-bit words; the final solution is motivated by divide-and-conquer, i.e.,
 - 1.1 one or more channels, each backed by
 - 1.2 one or more physical banks, each composed from
 - 1.3 one or more devices, each composed from
 - 1.4 one or more logical banks, of
 - 1.5 one or more arrays, of
 - 1.6 many cells
- 2. The major complication is a large range of increasingly detailed options:
 - lots of parameters mean lots of potential trade-offs (e.g., between size, speed and power consumption),
 - need to take care of detail: there are so many cells, any minor change can have major consequences!
- 3. Even so, there is just one key concept: we have some cells, and however they are organised we just need to identify and use the right cells given some address.



Notes:	

References and Further Reading

[1] Wikipedia: Computer memory. http://en.wikipedia.org/wiki/Category:Computer_memory.

[2] Wikipedia: Dynamic random access memory. http://en.wikipedia.org/wiki/Dynamic_random-access_memory.

[3] Wikipedia: Memory geometry. http://en.wikipedia.org/wiki/Memory_geometry.

[4] Wikipedia: Static random access memory. http://en.wikipedia.org/wiki/Static_random_access_memory.

[5] U. Drepper. What every programmer should know about memory. http://www.akkadia.org/drepper/cpumemory.pdf.

[6] D. Page.
 Chapter 8: Memory and storage.
 In A Practical Introduction to Computer Architecture. Springer-Verlag, 1st edition, 2009.

[7] W. Stallings.
 Chapter 5: Internal memory.
 In Computer Organisation and Architecture. Prentice-Hall, 9th edition, 2013.

[8] A.S. Tanenbaum.
 Section 3.3.4: Memory organisation.
 In Structured Computer Organisation [11].

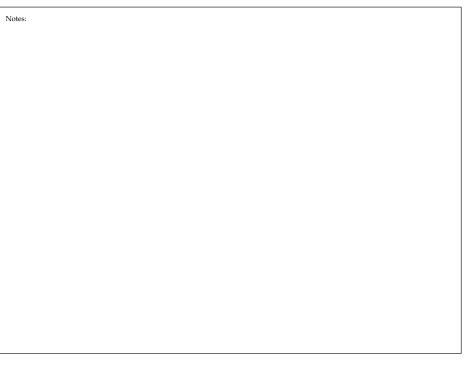
Dan Page (page@cs.bris.ac.uk) © 2014-Computer Architecture (new format)

Slide 30 of 31



References and Further Reading

- [9] A.S. Tanenbaum.Section 3.3.5: Memory chips.In Structured Computer Organisation [11].
- [10] A.S. Tanenbaum. Section 3.3.6: RAMs and ROMs. In Structured Computer Organisation [11].
- [11] A.S. Tanenbaum. Structured Computer Organisation. Prentice-Hall, 6th edition, 2012.



Notes:			