ESE534: Computer Organization

Day 5: February 1, 2010
Memories

Previously…

- Arithmetic: addition, subtraction
- Reuse:
  - pipelining
  - bit-serial (vectorization)
  - shared datapath elements
- FSMDs
- Area/Time Tradeoffs
- Latency and Throughput

Today

- Memory
  - features
  - area/delay intuition and modeling
  - design
  - technology

Preclass 1

When is:

$$2000\sqrt{N} + N < 100N$$

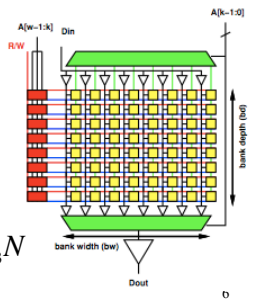
Preclass 2

Find *m* to minimize:

$$\frac{C_1 N \log_2(N)}{m} + C_2 m + C_3 N$$

Preclass 2

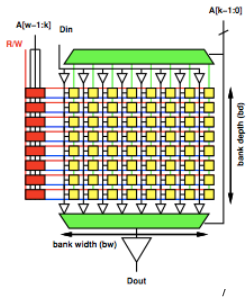
- Related?



$$\frac{C_1 N \log_2(N)}{m} + C_2 m + C_3 N$$

Preclass 3, 4

- Delay
  - Scale with bd?
  - Scale with bw?

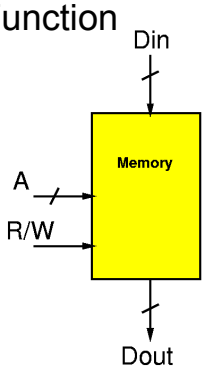


Memory

- What's a memory?
- What's special about a memory?

Memory Function

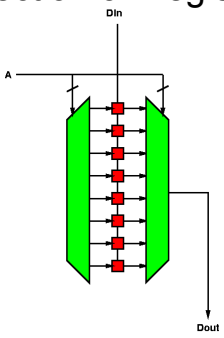
- Typical:
  - Data Input Bus
  - Data Output Bus
  - Address
    - (location or name)
  - read/write control



Memory

- Block for storing data for later retrieval
- State element
- What's different between a memory and a collection of registers like we've been discussing?

Collection of Registers



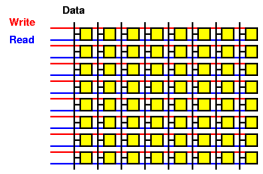
Memory Uniqueness

- Cost
- Compact state element
- Packs data very tightly
- At the expense of sequentializing access
- Example of Area-Time tradeoff
  - and a key enabler

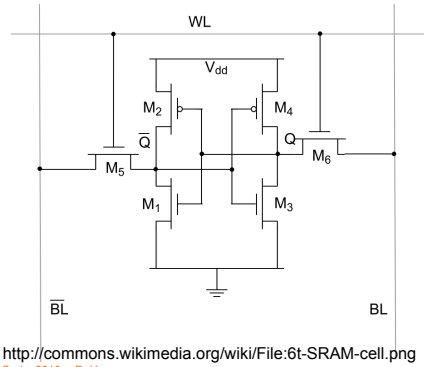
Memory Organization

- **Key idea**: sharing
  - factor out common components among state elements
  - can have big elements if amortize costs
  - state element unique → small

Memory bit cell

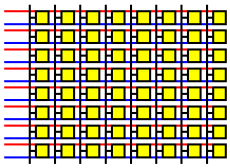


SRAM Memory bit

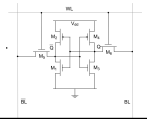


Source: http://commons.wikimedia.org/wiki/File:6t-SRAM-cell.png

Memory Organization

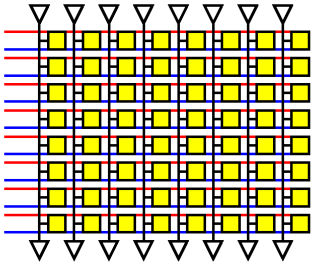


- Share: Interconnect
  - Input bus
  - Output bus
  - Control routing
- **very** topology/wire cost aware design
- Note: local, abuttment wiring



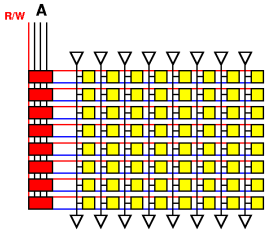
Share Interconnect

- Input Sharing
  - wiring
  - drivers
- Output Sharing
  - wiring
  - sensing
  - driving

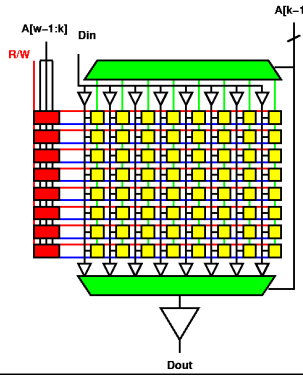


Address/Control

- Addressing and Control
  - an overhead
  - paid to allow this sharing

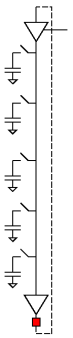


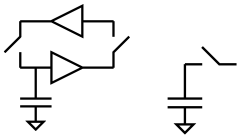
Memory Organization



Dynamic RAM

- Goes a step further
- Share refresh/restoration logic as well
- Minimal storage is a capacitor
- "Feature" DRAM process is ability to make capacitors efficiently





Some Numbers (memory)

- Unit of area = λ^2 (F=2λ)
  - [more next week]
- Register as stand-alone element $\approx 4K\lambda^2$
  - *e.g.* as needed/used last lecture
- Static RAM cell $\approx 1K\lambda^2$
  - SRAM Memory (single ported)
- Dynamic RAM cell (DRAM process) $\approx 100\lambda^2$
- Dynamic RAM cell (SRAM process) $\approx 300\lambda^2$

DRAM Layout

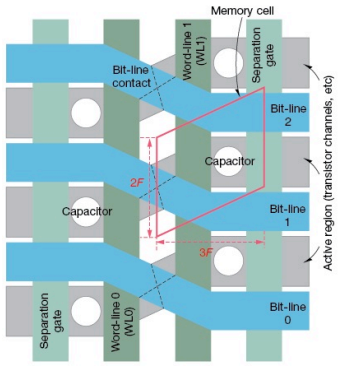
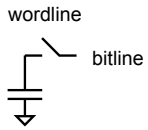
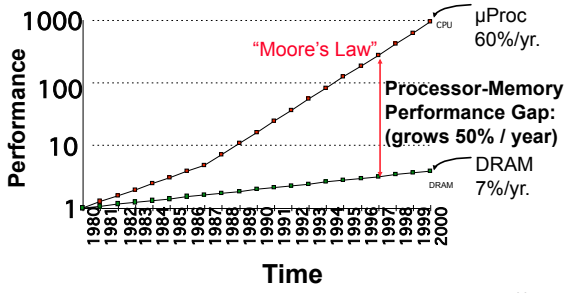


Fig 3 $6F^2$ Memory Cell Layout The $6F^2$ cell layout used by Micron Technology has active regions used for transistor channels, etc, formed diagonally. Layout presented at the 2004 Symposium on VLSI Technology.



Source: http://techon.nikkeibp.co.jp/article/HONSHI/20071219/144399/

DRAM Latency



[From Patterson et al., IRAM: http://iram.cs.berkeley.edu/]

Contemporary DRAM

- 1GB DDR3 SDRAM from Micron
  - http://www.micron.com/products/dram/ddr3/
  - 96 pin pakage
  - 16b datapath IO
  - Operate at 500+MHz
  - 37.5ns random access latency

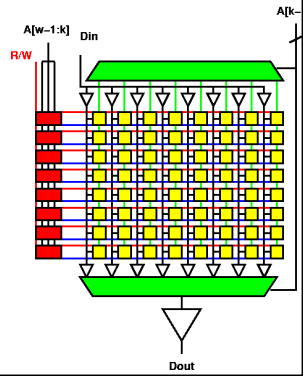
Table 1: Key Timing Parameters

Speed Grade	Data Rate (Mb/s)	Target RCD/RP/CL	^{t}RCD (ns)	^{t}RP (ns)	CL (ns)
-25E	800	5-5-5	12.5	12.5	12.5
-25	800	6-6-6	15	15	15
-187E	1,066	7-7-7	13.1	13.1	13.1
-187	1,066	8-8-8	15	15	15
-15E	1,333	9-9-9	13.5	13.5	13.5
-15	1,333	10-10-10	15	15	15

Options	Marking
• Configuration	
– 64 Meg x 16 (8 Meg x 16 x 8 banks)	64M16
– 128 Meg x 8 (16 Meg x 8 x 8 banks)	128M8
– 256 Meg x 4 (32 Meg x 4 x 8 banks)	256M4
• FBGA package (lead-free)	
– x4, x8; 86-ball FBGA (9mm x 15.5mm)	BY
– x16; 96-ball FBGA (9mm x 15.5mm)	LA
• Timing – cycle time	
– 2.5ns @ CL = 6 (DDR3-800)	-25
– 2.5ns @ CL = 5 (DDR3-800)	-25E
– 1.87ns @ CL = 8 (DDR3-1066)	-187
– 1.87ns @ CL = 7 (DDR3-1066)	-187E
– 1.5ns @ CL = 10 (DDR3-1333)	-15
– 1.5ns @ CL = 9 (DDR3-1333)	-15E

Memory Access Timing

- RAS/CAS access
- Optimization for access within a row



Contemporary DRAM

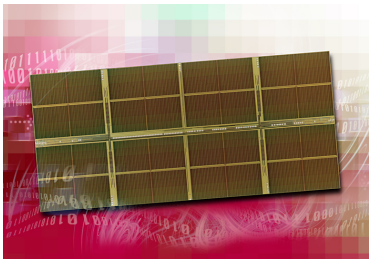
- 1GB DDR3 SDRAM from Micron
  - http://www.micron.com/products/dram/ddr3/
  - 96 pin pakage
  - 16b datapath IO
  - Operate at 500+MHz
  - 37.5ns random access latency

Table 1: Key Timing Parameters

Speed Grade	Data Rate (Mb/s)	Target RCD/RP/CL	tRCD (ns)	tRP (ns)	CL (ns)
-25E	800	5-5-5	12.5	12.5	12.5
-25	800	6-6-6	15	15	15
-187E	1,066	7-7-7	13.1	13.1	13.1
-187	1,066	8-8-8	15	15	15
-15E	1,333	9-9-9	13.5	13.5	13.5
-15	1,333	10-10-10	15	15	15

Options	Marking
• Configuration	
– 64 Meg x 16 (8 Meg x 16 x 8 banks)	64M16
– 128 Meg x 8 (16 Meg x 8 x 8 banks)	128M8
– 256 Meg x 4 (32 Meg x 4 x 8 banks)	256M4
• FBGA package (lead-free)	
– x4, x8; 86-ball FBGA (9mm x 15.5mm)	BY
– x16; 96-ball FBGA (9mm x 15.5mm)	LA
• Timing – cycle time	
– 2.5ns @ CL = 6 (DDR3-800)	-25
– 2.5ns @ CL = 5 (DDR3-800)	-25E
– 1.87ns @ CL = 8 (DDR3-1066)	-187
– 1.87ns @ CL = 7 (DDR3-1066)	-187E
– 1.5ns @ CL = 10 (DDR3-1333)	-15
– 1.5ns @ CL = 9 (DDR3-1333)	-15E

1 Gigabit DDR2 SDRAM



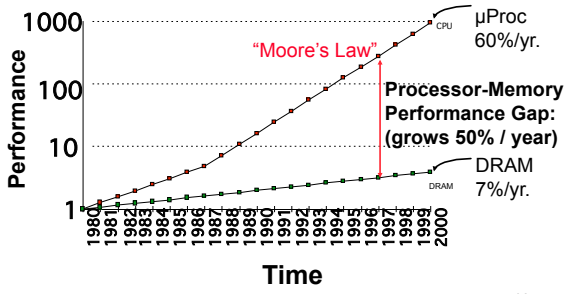
[Source: http://www.elpida.com/en/news/2004/11-18.html]

Contemporary DRAM

- 4GB DDR3 SDRAM from Micron
  - http://download.micron.com/pdf/datasheets/modules/ddr3/ksf16c256_512x64hz.pdf
  - Operate at 800MHz
  - 50ns random access latency

Speed Grade	Industry Nomenclature	Data Rate (MT/s) CL = 11	CL = 10	CL = 9	CL = 8	CL = 7	CL = 6	CL = 5	tRCD (ns)	tRP (ns)	tRC (ns)
-1G6	PC3-12800	1600	1333	1333	1066	1066	800	667	13.125	13.125	48.125
-1G4	PC3-10600	–	1333	1333	1066	1066	800	667	13.125	13.125	49.125
-1G1	PC3-8500	–	–	–	1066	1066	800	667	13.125	13.125	50.625
-1G0	PC3-8500	–	–	–	1066	–	800	667	15	15	52.5
-80C	PC3-6400	–	–	–	–	–	800	800	12.5	12.5	50
-80B	PC3-6400	–	–	–	–	–	800	667	15	15	52.5

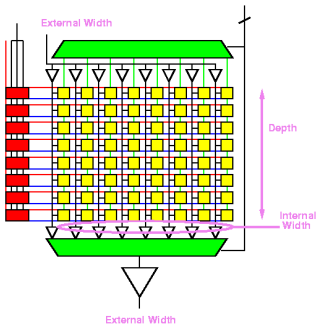
DRAM Latency



[From Patterson et al., IRAM: http://iram.cs.berkeley.edu/]

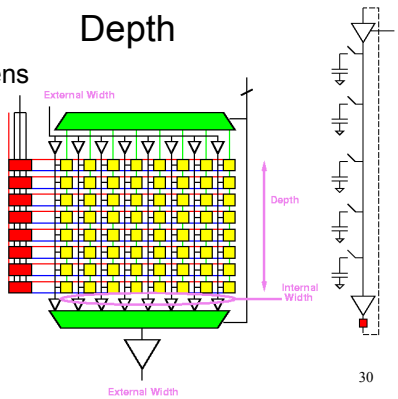
Basic Memory Design Space

- Width
- Depth
- Internal vs. External Width
- *Banking*
  - *To come*



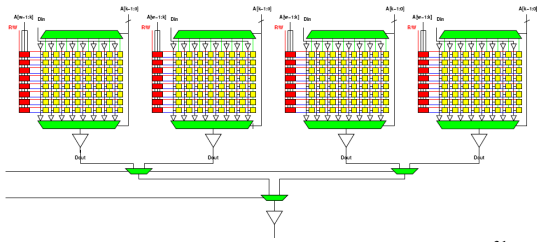
Depth

- What happens as make deeper?

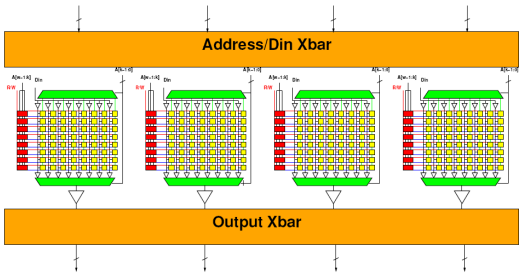


Banking

- Tile Banks/memory blocks



Independent Bank Access



Memory

- **Key Idea**
  - Memories hold state compactly
  - Do so by minimizing key state storage and amortizing rest of structure across large array

Yesterday vs. Today (Memory Technology)

- What's changed?

Yesterday vs. Today (Memory Technology)

- What's changed?
  - Capacity
    - single chip
  - Integration
    - memory and logic
    - dram and logic
    - embedded memories
  - Room on chip for big memories
    - And many memories…
  - Don't have to make a chip crossing to get to memory

Important Technology Cost

- IO between chips << IO on chip
  - pad spacing
  - area vs. perimeter (4s vs. s^2)
  - wiring technology
- **BIG** factor in multi-chip system designs
- Memories nice
  - very efficient with IO cost vs. internal area

On-Chip vs. Off-Chip BW

- Use Micron 1Gb DRAM as example

Table 1: Key Timing Parameters

Speed Grade	Data Rate (Mb/s)	Target RCD/RP/CL	tRCD (ns)	tRP (ns)	CL (ns)
-25E	800	5-5-5	12.5	12.5	12.5
-25	800	6-6-6	15	15	15
-187E	1,066	7-7-7	13.1	13.1	13.1
-187	1,066	8-8-8	15	15	15
-15E	1,333	9-9-9	13.5	13.5	13.5
-15	1,333	10-10-10	15	15	15

Options — **Marking**

- Configuration
  - 64 Meg x 16 (8 Meg x 16 x 8 banks) — 64M16
  - 128 Meg x 8 (16 Meg x 8 x 8 banks) — 128M8
  - 256 Meg x 4 (32 Meg x 4 x 8 banks) — 256M4
- FBGA package (lead-free)
  - x4, x8; 86-ball FBGA (9mm x 15.5mm) — BY
  - x16; 96-ball FBGA (9mm x 15.5mm) — LA
- Timing – cycle time
  - 2.5ns @ CL = 6 (DDR3-800) — -25
  - 2.5ns @ CL = 5 (DDR3-800) — -25E
  - 1.87ns @ CL = 8 (DDR3-1066) — -187
  - 1.87ns @ CL = 7 (DDR3-1066) — -187E
  - 1.5ns @ CL = 10 (DDR3-1333) — -15
  - 1.5ns @ CL = 9 (DDR3-1333) — -15E

On Chip BW:
assume 8 1024b banks?
assume 8×16 1024b banks?
assume 1024×1024b banks?

Costs Change

- Design space changes when whole system goes on single chip
- Can afford
  - wider busses
  - more banks
  - memory tailored to application/architecture
- Beware of old (stale) answers
  - their cost model was different

What is Importance of Memory?

- Radical Hypothesis:
  - Memory is simply a very efficient organization which allows us to store data compactly
    - (at least, in the technologies we've seen to date)
  - A great engineering trick to optimize resources
- Alternative:
  - memory is a **primary**
- State is a primary, but state≠memory

Admin

- Added paper on modeling memories as supplemental reading for today
- Reading for next Monday on web
  - Classic paper on MOS scaling
- Reading for next Wednesday in Blackboard

Big Ideas [MSB Ideas]

- Memory: efficient way to hold state
- Resource sharing: key trick to reduce area
- Memories are a great example of resource sharing

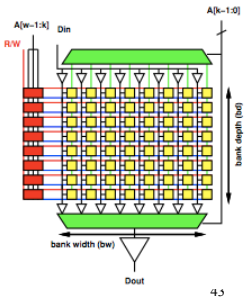
Big Ideas [MSB-1 Ideas]

- Tradeoffs in memory organization
- Changing cost of memory organization as we go to on-chip, embedded memories

Preclass 2

$$\frac{C_1 N \log_2(N)}{m} + C_2 m + C_3 N$$

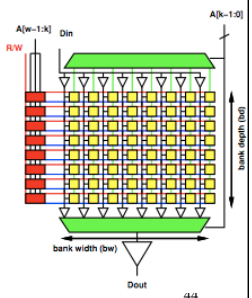
- Related?
- m=bw
- bw*bd=N
  - So bd=N/m



Preclass 2

$$C_1 bd \log_2(N) + C_2 bw + C_3 N$$

- C_3 is for area of single memory cell
- $\log_2(n)$ is for decoder
  - In red
- C_1 is for gates in decoder
- C_2 is for amps at bottom of row and drivers at top



Preclass 1

$$\frac{C_1 N \log_2(N)}{m} + C_2 m + C_3 N$$

When we solved for N, we found m is proportional to √N

$$C_1 \sqrt{N} \log_2(N) + C_2 \sqrt{N} + C_3 N$$

If we approximate $\log_2(N)$ as a constant, we get

$$C_4 \sqrt{N} + C_3 N$$

$$2000\sqrt{N} + N$$