2. Link and Memory Architectures and Technologies

2.1 Links, Thruput/Buffering, Multi-Access Ovrhds

2.2 Memories: On-chip / Off-chip SRAM, DRAM


Manolis Katevenis and
A. Psathakis, G. Passas, S. Lyberis

CS-534 – Univ. of Crete and FORTH, Greece


2.2 Memories: On-chip / Off-chip SRAM, DRAM

Table of Contents:

• 2.2.1 On-Chip SRAM blocks
  – Area, Power Consumption, Cycle Time; 1 or 2 ports
  – Power cons. per unit throughput: SRAM, pin transceivers

• 2.2.2 Off-Chip SRAM technologies
  – Address-Read-Data Pipelining
  – Separate Unidirectional versus Unified Bidirectional Data Lines

• 2.2.3 DRAM Chips and their Pin Interface
  – Row Access versus Column Access
  – Interleaved accesses to the internal DRAM banks
2.2.1 On-Chip SRAM

Read Cycle Includes:
- Precharge bit lines
- Decode row address
- Activate word line
  - faster when narrow
- Discharge bit lines
  - faster when short
- Sense amplifiers
  - don’t wait for full discharge before telling the result
- Column multiplexors
  - use column address

Sense Amplifiers: Role, Consequences

- Sense amplifiers significantly speed up read access time
  - sense 0-contents soon after bit-line discharge has started
- Sense amplifiers (SA) are large in size
  - can fit only one SA per 2 to 8 columns
  - analog multiplexors before SA select columns to be read
  - digital multiplexors after SA needed for narrow port widths – they result in large blocks being slower when port is too narrow
- Sense amplifiers consume significant energy when activated
  - only activate the block when read data are actually needed
  - power consumption is proportional to access frequency
  - power consumption is proportional to number of sense amp’s (increases with port width, or with bit capacity of SRAM)
On-chip SRAM blocks example (45 nm CMOS): Area

Area per Mbit vs Capacity, 45 nm, LSTP, Vdd: 1.1V, Vth: 0.502V

<table>
<thead>
<tr>
<th>Ports:1</th>
<th>32bit</th>
<th>64bit</th>
<th>128bit</th>
<th>256bit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Area per Megabit: Comments

- Values are $(\mu m)^2$/bit = $(mm)^2$/Mbit
- CACTI estimates, using ITRS 2010 roadmap
- Large blocks are more area-efficient than small ones
  - peripheral overhead (address decoders, column multiplexors, sense amplifiers, power ring) amortized over a larger core
- Port width costs a lot for small blocks
  - more sense amplifiers needed, possibly non-square aspect ratio
  - large blocks need many SA’s, for either narrow or wide ports
- Two-port blocks: one read-only port and one write-only port
- Two-port area is about 2x to 3x the area of one-port SRAM
- Blocks include ECC overhead
- No power ring included in the quoted area numbers
Core Area as percent of Total Area (45 nm CMOS)

Area Efficiency vs Capacity, 45 nm, LSTP, Vdd: 1.1v, Vth: 0.502v

Block aspect ratio vs Capacity, 45 nm, LSTP, Vdd: 1.1v, Vth: 0.502v
On-chip SRAM (45 nm): **Dynamic Power Consumpt’n**

Dyn. Power per port for R/W accesses vs Cap., 45 nm, LSTP, Vdd: 1.1v, Vth: 0.502v

<table>
<thead>
<tr>
<th>Ports:1</th>
<th>32bit</th>
<th>64bit</th>
<th>128bit</th>
<th>256bit</th>
<th>Ports:2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>12</td>
<td>18</td>
<td>24</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>24</td>
<td>36</td>
<td>48</td>
<td>60</td>
</tr>
<tr>
<td>16</td>
<td>24</td>
<td>48</td>
<td>72</td>
<td>96</td>
<td>60</td>
</tr>
<tr>
<td>64</td>
<td>48</td>
<td>96</td>
<td>144</td>
<td>192</td>
<td>60</td>
</tr>
<tr>
<td>256</td>
<td>60</td>
<td>120</td>
<td>180</td>
<td>240</td>
<td>60</td>
</tr>
</tbody>
</table>

**Comments of Dynamic Power Consumption**

- Dynamic power is consumed when nodes change state ⇒ proportional to access frequency: \( mW / GHz \)
- Consumption increases with block size (squ. root of capacity) due to increasing word-line and bit-line capacitance
- Consumption increases with port-width (number of SA’s)
- Two-port blocks: quoted consumption is *per-port*
- Two-port total consumption ≈ 2x to 3x consumption of 1-port
- Two-port blocks: one *read-only* port and one *write-only* port
- CACTI estimates, based on ITRS 2010 roadmap for 45 nm
- Low leakage power process assumed: \( V_{th} = 0.5 \text{ V} \)
- Typical-case consumption quoted; \( V_{DD} = 1.1 \text{ V}, 60^\circ \text{C} \)
  - all cycles active, all address and data bits switching
Comments of Static (Leakage) Power Consumption

- Static power is consumed all the time, independent of activity, by leaky transistors that should be OFF but are not fully so
- Measured in micro-Watts, for the entire block
- Consumption is proportional to the number of transistors \( \Rightarrow \) proportional to block capacity (Kbits)
- Consumption almost unaffected by port-width (not sure why)
- Two-port blocks: quoted consumption is for the entire block
- Two-port (total) consumption \( \approx \) + 5 to 10% relative to 1-port (not sure why so little)
- CACTI estimates, based on ITRS 2010 roadmap for 45 nm
- Low leakage power process assumed ("LSTP": low-standby power): \( V_{th} = 0.5 \text{ V} \); typical-case cons'ptn: \( V_{DD} = 1.1 \text{ V}, 60^\circ \text{C} \)
On-chip SRAM (45 nm CMOS): Cycle Time

Cycle time vs Capacity, 45 nm, LSTP, Vdd: 1.1v, Vth: 0.502v

Ports: 1
- 32bit
- 64bit

Ports: 2
- 128bit
- 256bit

Cycle Time (1/AccessRate): Comments

• *Small is Fast*: small blocks are faster than large blocks
  – bit-line (and word-line) capacitance increases with length
  – for large capacities, beyond about 64 Kb, it is faster to use
    multiple small blocks, perhaps with external data-out mux
    after them, than to use a single large block
• Speed is almost independent of port width
  – except for small blocks that are excessively wide
• Two-port SRAM is ≈ 20% slower than 1-port
• CACTI estimates, based on ITRS 2010 roadmap for 45 nm
• Low-leakage-power process assumed: $V_{th} = 0.5$ V
• High-performance process would give 2x to 4x higher speed
• Typical-case speed quoted; $V_{DD} = 1.1$ V, 60°C
On-Chip SRAM Buffer Example 1 of 2: 40-Byte wide

- **Width** = 1 min-size IP packet = 40 Bytes = 320 bits = 5 blocks × 64 bits/block
- **One-Port**, 2048 packets × 40 B/pck = 80 KB = 640 Kb
- 45 nm CMOS, 1.1 Volt, low-leakage (static) power process
- **Area** = 5 banks × 128 Kb/bank × 0.44 mm²/Mb = 0.64 Mb × 0.44 mm²/Mb ≈ **0.3 mm²**
- **Throughput** = 320 bits × 0.54 Gaccesses/s ≈ **170 Gb/s**
- **Dynamic Power Consumption** = 5 banks × 17.5 mW/GHz × 0.54 GHz = **47 mW**
- **Static Power** = 5 banks × 0.0015 mW/bank = negligible
  (would be ~50 mW in a high-performance process!)

On-Chip SRAM Buffer Example 2 of 2: 256-Byte wide

- **Width** ≈ 1 average-size IP packet = 256 Bytes = 2048 bits = 64 blocks × 32 bits/block
- **Two-Port**, 2048 packets × 256 B = 512 KB = 4 Mb
- 45 nm CMOS, 1.1 Volt, low-standly-power process
- **Area** = 64 banks × 64 Kb/bank × 0.9 mm²/Mb = 4 Mb × 0.9 mm²/Mb ≈ **3.5 mm²**
- **Throughput** = 2 ports × 2048 b/port × 650 MHz ≈ **2.6 Tb/s**
  (1300 Gb/s writes + 1300 Gb/s reads)
- **Power Consumption** = 64 banks × 2 ports × 11 mW/GHz × 0.65 GHz ≈ **0.9 W**
- **Conclusion**: “no problem” on-chip, except for short packets
### Power Cons./Throughput (1 of 2): on-chip SRAM

- Consider some “usual, medium-size” SRAM’s (45nm, LSTP):
  - 1-port, ×32: ≈ 10 mW/GHz = 10 mW / 32 Gbps ≈ 0.31 mW/Gbps
  - 1-port, ×64: ≈ 16 mW/GHz = 16 mW / 64 Gbps ≈ 0.25 mW/Gbps
  - 1-port, ×128: ≈ 30 mW/GHz = 30 mW /128 Gbps ≈ 0.23 mW/Gbps
  - 2-port, ×32: ≈ 12 mW/GHz = 12 mW / 32 Gbps ≈ 0.38 mW/Gbps
  - 2-port, ×64: ≈ 20 mW/GHz = 20 mW / 64 Gbps ≈ 0.31 mW/Gbps

- Conclusion: **0.2 to 0.4 mW/Gbps** power consumption for on-chip buffer memories

### Power Cons./Throughput (2 of 2): Chip I/O

- High-speed serial off-chip transceiver ≈ **12 to 35 mW/Gbps**
  - differential pair, 8b/10b encoding
  - e.g. Xilinx Virtex 7 (28 nm CMOS): 260 mW for 12.5 GBaud transceiver i.e. 10 Gbps xmit + 10 Gbps rcv; or 200 mW for 6.25 Gbaud (5+5 Gbps); or 170 mW for 3.125 Gbaud (2.5+2.5 Gbps)

  ⇒ **Conclusion:** chip-to-chip communication costs **one to two orders of magnitude more** than on-chip buffering, in term of power consumption!

- Total chip power consumption (limited to ≈ 10 to 30 Watts) limits total chip throughput to **about 1 Tbps/chip** or less
2.2.2 Off-Chip SRAM Technologies

- Large on-chip throughput, owing to parallelism of accesses
- Gradual improvements in pin-interface protocols (late 90’s):
  1. Clock-synchronous, pipelined address/data communication
  2. Double-Data Rate (DDR) data-pin timing (see §2.1)
  3. Source-synchronous clocking
     - clock signal propagating in the same direction as data (or address) signals – normally implies two separate clocks
  4. Separate, unidirectional Write-Data and Read-Data buses
     - avoids bus turn-around overhead, but
     - requires 50% writes – 50% reads for full utilization
  5. Write-data timing similar to read-data timing
     - first send the address, later send the data, so that address-bus to data-bus time-offset stays fixed for reads & writes

Clock-Synchronous RAM: Pipelined Communication

"Flow Through": old timing
- no overlapping between SRAM operation and communication

"Synchronous" Registered Interface
- pipelined SRAM operation and chip-to-chip communication
2.2 RAM Technologies

Source-Synchronous Data Clocking

...further increasing the throughput of chip-to-chip communication:

- When the clock frequency rises, the chip-to-chip (speed-of-light) delay becomes non negligible w.r.t pulse width
- clk3 is a delayed version of clk1, i.e. has (exactly) the same frequency, but its delay (phase shift) may vary (slowly) with time

SRAM Data I/O Paths

Separate D(in) & Q(out) Paths

Datapath underutilization when imbalanced (≠ 50 – 50 %) read-write transactions

Shared DQ Data Bus

Bus turn-around overhead: Databus underutilization when frequently switching between read and write transactions
“QDR” (Quad Data Rate) SRAM

Modern SRAM chip technology w. separate D(in) & Q(out) paths

Example QDR SRAM (2007): CY7C1545V18

- 72 Mbits = 4 M × 18 bits (width = 2 Bytes + parity/ECC)
- ≤ 375 MHz clock ⇒ cycle = 2.67 ns; bit-time = 1.33ns (DDR)
- Burst-of-4 words ↔ simple (non-DDR) address timing
- Peak Write Throughput:
  375 MHz × 2 (DDR) × 16 bits = 12 Gb/s/chip = 1.5 GB/s
- Peak Read Throughput = (similarly) 12 Gb/s
- Peak Total throughput for balanced (50%-50%) read-write:
  12 + 12 = 24 Gb/s = 3 GB/s
- Power consumption ≈ 2.4 W (typical) @ 375 MHz, 1.8 Volt
  ⇒ Power per throughput ≈ 2.4 W / 24 Gbps = 100 mW/Gbps
2.2  RAM Technologies

2.2  - U.Crete - M. Katevenis - CS-534

Shared “DQ” Data Bus Timing

Naïve Timing

“ZBT” (Zero Bus Turn Around) Timing

Underutilization on every read-to-write transition

D1 has not yet been written at M[A1] when reading from M[A2] starts… need to bypass mem. when A2==A1

Example Shared-Bus SRAM (2007): CY7C1550V18

- 72 Mbits = 2 M × 36 bits (width = 4 Bytes + parity/ECC)
- ≤ 375 MHz clock ⇒ cycle = 2.67 ns; bit-time = 1.33ns (DDR)
- Peak Throughput = 375 MHz × 2 (DDR) × 32 bits = 24 Gb/s
- “NoBL” (No Bus Latency) = “ZBT” (Zero Bus Turn-Around, ala Micron)
- Although NoBL/ZBT, one clock cycle is lost every time the bus direction changes from read to write (bus turn-around)
  ⇒ throughput with alternating read/writes ≈
  ≈ 2/3 × peak throughput = 16 Gb/s
- Power consumption ≈ 2.4 W (typical) @ 375 MHz, 1.8 Volts
  ⇒ Power per throughput ≈ 2.4 W / 24 Gbps ≈ 100 mW/Gbps
2.2.3 Dynamic RAM Chips and their Pin Interface

- Highest density and longest internal latency RAM chips
- Huge internal parallelism, when addresses are *favorable*:
  - multiple banks – memory interleaving
  - per-bank: entire row (hundreds of bits) accessed in parallel
- Pin Interface: advanced techniques to increase throughput
  - pins synchronized to a high-speed clock (Synchronous DRAM)
  - 100’s of bits piped thru 10’s of data pins during several clocks
  - internal RAM access is independent of clock – multiple cycles
- Three-step internal accesses – each bank independently
  - row access: activate a row in a bank, copy into sense amp’s
  - column access: read/write multiple bits in selected row
  - precharge: get this bank ready for activating another row
- Address pins time-shared: row – column addr; multiple banks

Example DDR3 SDRAM (2007): MT41J64M16

- 1 Gbit = 64 M × 16 bits = 8 banks × 8 Mw/bank × 16 b/w
- ≤ 800 MHz clock
- Bidirectional data pins, DDR timing ⇒ up to 1.6 Gbps/pin
- Internal latencies specified as absolute times:
  - row-addr. to column-addr. ≥ 14 ns
  - column-addr. to read-data ≥ 14 ns
  - bank-cycle time ≥ 48 ns; precharge time ≥ 14 ns
- Translated to # of clock cycles by user @ boot time
  - e.g. at 800 MHz: row-acc ≥ 11~, col-acc ≥ 11~, bnk-cycle ≥ 38~

- (Remaining slides are for a much older chip (~2001)…)
DRAM Basics:
Row Address, Column Address, Precharge

Multiple accesses within same row are faster

Row Address Decoder
Sense Amplifiers
Column Addr Decoder

Data

Addr
RA
CA
RA

Data
D

Addr
RA
CA1
CA2
CA3

Data
D1
D2
D3

Fast DRAM Example (Intel)
Micron M146 V2 M32
DDR SDRAM (Synchronous DRAM)
- 32-bit (shared DRAM) data bus, DDR timing 4 ns 2 words × 32 bits each per clock cycle peak data bus throughput
- 64 Mbits = 2 M × 32 bits = 512 K × 32 bit × 4 banks
- 200 MHz max clock frequency
- 8x1 watt of peak power rate, using one bank only, 2.5 V only (no need to fire up all banks)

Row Address to Column Address: t_{RD} ≥ 20 ns (typical: 4 ν)
Column Address to Read Data (CAS latency): C_{L} ≥ 15 ns (typical: 3 ν)
Write Recovery Time (write data to precharge): t_{WR} ≥ 2 ns
Precharge Time: t_{RP} ≥ 20 ns (typical: 4 ν)
Cycle Time (same bank): t_{R_C} ≥ 60 ns (typical: 12 ν)
Bank-to-Bank Activation (other bank Row-to-Row): t_{RRD} ≥ 2 ν
Read-to-Write bus turnaround: fast cycles: 3 ν
Write-to-Read same bank fast cycles (write recovery time): 2 ν
Write to Read other bank fast cycles: 8 ν
### Single-Bank Read Access

- **ACT** = Activate
- **Ra** = Row Address
- **Ca** = Column Address
- **Ba** = Bank #a
- **Ca** = Column Address #Ca

- **RD** = Read (for predefined burst size)
- **WR** = Write (for predefined burst size)

- **CL** = Command latency
- **tRCD** = Row cycle delay
- **tWR** = Write recovery time
- **tRAS** = Row access time
- **tRDP** = Row decode/precharge period

- **tRCD, tRAS, tRDP** vary based on the burst size.

### Single-Bank Write Access

- **ACT** = Activate
- **WR** = Write (for predefined burst size)
- **Ra** = Row Address
- **Ca** = Column Address

- **Ba** = Bank #a

- **RD** = Read

- **tRCD, tRAS, tRDP** vary based on the burst size.

---

2.2 - U.Crete - M. Katevenis - CS-534

31

32
Multiple Accesses to Different Columns in the same Row of a Bank.

All transactions shown are to the same bank #a, and to the same activated row Ra in that bank. The transactions shown are:
- Read From Column Ca to a1, a2
- Read From Column Cb to b1, b2
- Write c1, c2 at Column Ce
- Read From Column Ce to e1, e2

---

Multi-Bank Operation: Memory Interleaving

- burst length set to 8; each successive READ command interrupts the preceding burst, resulting in net bursts of 6.