2.1 Buffer Memory Technology

• Memory Blocks On-Chip
  – On-chip SRAM area, power consumption, access rate

• Power Consumption for chip-to-chip communication

• Memory Chips (commercially available)
  – Chip periphery interface: communication standards to memory chips and their off-chip throughput
  – DRAM chips, internal banks, Bank Interleaving
On-Chip SRAM

Memory blocks inherently provide, on-chip, very high throughputs, owing to their inherent parallelism: an entire row is accessed at once. This high throughput is available on-chip, due to the feasibility of very wide datapaths, running at high clock rates.

(Very wide or very large memories are made of several smaller memory blocks, to reduce capacitive loading on word lines and bit lines.)

Example layout: 16 Kbit = 2K x 8
On-Chip SRAM block Area (130nm CMOS)

$sq\text{-mm per M bit}$, 1-port

<table>
<thead>
<tr>
<th>SRAM block capacity (K bits)</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>256</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-bit</td>
<td>14</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>16-bit</td>
<td>12</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>32-bit</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1.5</td>
</tr>
<tr>
<td>64-bit</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1.5</td>
<td>1.2</td>
</tr>
</tbody>
</table>

2-port (1rd+1wr)

<table>
<thead>
<tr>
<th>block capacity (K bits)</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-bit</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>16-bit</td>
<td>12</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

- power ring not included
- values are $(\mu m)^2$/bit $\approx (mm)^2$/Mbit

parameter: port width – 8 to 64 bits
On-Chip SRAM Power Consumption (130nm)

**worst-case mW / MHz, 1-port**

- Parameter: port width – 8 to 64 bits
- 1.2 Volts, 130 nm (year 2005)
- Blocks compiled for performance
On-Chip SRAM block Cycle Time (130nm CMOS)

wrst-case cycle time (ns), 1-port

ns, 2-port (1rd+1wr)

SRAM block capacity (K bits)

- 8-bit
- 16-bit
- 32-bit
- 64-bit

parameter: port width – 8 to 64 bits

- 1.2 Volts, 130 nm (year 2005)
- blocks compiled for performance
On-Chip SRAM block Cost, Performance

Area per Kbit:
- Area efficiency increases with block capacity: peripheral overhead (address decoders, column multiplexors, sense amplifiers) grows slower than core
- Port width costs significantly for small memories (more sense amp’s, non-square aspect ratio)
- Two-port area $\approx 2 \times$ one-port area
- Power ring: add 25 $\mu$m on each side of the block given in the above charts (width and height increase by 50 $\mu$m each)
- 1 sense amp / 8 col., usually
- Quoted blocks have write-byte enable signals, except 8-bit ones

Power Consumption per MHz:
- Dominated by port-width for small mem’s (sense amp. consumption)
- Dominated by block size for large mem’s (word- & bit- line consum.)
- $P_{\text{two-ports}} \approx 2 \times P_{\text{one-port}}$

Access Rate ($=1/$cycle-time): 
- Large blocks are quite slower than small ones, for sizes beyond the “knee” of the curve
- For large blocks, narrow ports reduce the speed, because of extra mux’es after sense amp’s
- Two-port speed $\approx$ speed of 1-port block with twice the num. of bits
2-port vs. Dual-port Area (square-mm / Mbit)

2-port (1rd + 1wr)

SRAM block capacity (K bits)

Dual-port (2 x rd/wr)

SRAM block capacity (K bits)

8-bit — 32-bit
2-port vs. Dual-port Power (worst-case mW / MHz)

2-port (1rd + 1wr)

Dual-port (2 x rd/wr)

SRAM block capacity (K bits)

8-bit

32-bit

CS-534, Copyright Univ. of Crete
2-port vs. Dual-port Cycle Time (ns, worst-case)

2-port (1rd + 1wr)

Dual-port (2 x rd/wr)

SRAM block capacity (K bits)

8-bit  32-bit
On-Chip SRAM Buffer Example (i): 40-Byte wide

- **Width** = 1 min-size IP packet = 
  
  = 40 Bytes = 320 bits = 5 blocks × 64 bits/block

- **One-port**, 2048 packets × 40 B = 80 KB = 640 Kb

- 130 nm CMOS, 1.2 Volts

- **Area**: 5 banks × 128 Kb/bank × 3 mm²/Mb =
  
  = 0.64 Mb × 3 mm²/Mb ≈ 2 mm²

- **Throughput**: 320 bits × 300 Macc/s ≈ 100 Gb/s

- **Power Consumption**:
  
  5 banks × 0.11 mW/MHz × 300 MHz = 165 mW
On-Chip SRAM Buffer Example (ii): 256-Byte wide

- **Width**: \( \approx 1 \) average-size IP packet =
  \[
  = 256 \text{ Bytes} = 2048 \text{ bits} = 64 \text{ blocks} \times 32 \text{ bits/block}
  \]
- **Two-port (1rd+1wr)**, 2048 packets \( \times \) 256 B = 512 KB = 4 Mb
- **130 nm CMOS, 1.2 Volts**
- **Area**: \( 64 \times 64 \text{ Kb} \times 6.1 \text{ mm}^2/\text{Mb} = 4 \text{ M} \times 6.1 \approx 25 \text{ mm}^2 \)
- **Throughput**: 2 ports \( \times \) 2048 b/port \( \times \) 240 MHz \( \approx 1 \text{ Tb/s} \)
  
  (500 Gb/s writes + 500 Gb/s reads)
- **Power Consumption**: 64 banks \( \times \) 2 ports \( \times \) 0.08 mW/MHz \( \times \) 240 MHz \( \approx 2.4 \text{ W} \)
- **Conclusion**: “no problem” on-chip, except for small packets
Power Consumption / Throughput: on-chip SRAM

• (1) On-Chip Buffer Memories:

• 130 nm CMOS, “usual, medium” SRAM block sizes:
  – 1-port, ×16:  ≈ 0.03 mW/MHz = 0.03 mW / 16 Mbps ≈ 2.0 mW/Gbps
  – 1-port, ×32:  ≈ 0.05 mW/MHz = 0.05 mW / 32 Mbps ≈ 1.6 mW/Gbps
  – 1-port, ×64:  ≈ 0.10 mW/MHz = 0.10 mW / 64 Mbps ≈ 1.6 mW/Gbps
  – 2-port, ×8:   ≈ 0.02 mW/MHz = 0.02 mW / 8 Mbps ≈ 2.5 mW/Gbps
  – 2-port, ×32:  ≈ 0.06 mW/MHz = 0.06 mW / 32 Mbps ≈ 2.0 mW/Gbps

• Conclusion: **1.5 to 2 mW / Gbps** on-chip buffer memories
Power Consumption / Throughput: Chip I/O

• (2) Chip-to-Chip I/O Pin Power Consumption:
  • both directions of a high-speed serial off-chip transceiver (without equalization –which consumes considerably)
  • 130 nm CMOS: 10 to 25 mW / Gbps chip-to-chip comm
  • copper cable power consumption is very small, by comparison
  ⇒ Chip-to-chip communication costs an order of magnitude more than on-chip buffering, in terms of power consumption
• Total chip power consumption (up to few tens of Watts) limits total chip throughput to about 1 Tbps/chip or less
Off-Chip Memory — or other networking/I/O chips:
How to Increase Chip-to-Chip Communication Throughput?

Old SRAM Read ("flow through"):

1. Pipelined Reads
(Synchronous, Registered Interface)
Further increasing the data pin through part of chip-to-chip communication:

(2) DDR (Double Data Rate) Timing

Traditional Synchronous Interface:

- Transmit and receive with a positive-edge-triggered register

DDR Interface:

- Transmit with: two registers:
  - one positive-edge-triggered register
  - one negative-edge-triggered register

Receive with: two registers:
... further increasing the data pin throughput of chip-to-chip communication...

(3) **Source-Synchronous Data Clocking**

When the clock frequency rises, the chip-to-chip (speed of light) delay becomes non-negligible w.r.t. pulse width.

---

**Synchronization - clock domain crossing**

ck3 is a delayed version of ck1, 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) and Q (out) Paths:

- time-mux'ed write & read addresses

\[ \text{Addr} \rightarrow \text{SRAM} \rightarrow \text{Q} \]

\[ \text{D in write data} \rightarrow \text{SRAM} \rightarrow \text{Q out read data} \]

- data path underutilization when imbalanced (\( \neq 50\% - 50\% \)) read/write transactions

Shared "DQ" Data Bus:

\[ \text{Addr} \rightarrow \text{SRAM} \rightarrow \text{DQ} \rightarrow \text{all data} \]

- bus turn-around overhead: data bus underutilization when frequently switching between read & write transactions
modern SRAM chip technology w. separate D(in) & Q(out) paths:

"QDR" (Quad Data Rate) SRAM

Addr

burst-of-2
DDR in

D

DDR addresses

QDR SRAM

burst-of-2
DDR out

Q

ck

Other Version:

"burst-of-4":

- addr. path is plain (NOT DDR)
- each addr. refers to 4 data words.

D

W1

W2

X1

X2

Y1

Y2

A

W

B

X

C

Y

read from A

read from B

Q

A1

A2

B1

B2

burst of 2 words
read from address A

burst of 2 words
read from address B
Example QDR SRAM (2001) Micron's MT54V512 #18

9 Mbits = \(512 \times 18\) bits

Clock freq. up to 167 MHz

\(T=6\,\text{ns}\) pulse width \(\approx 3\,\text{ns}\)

Peak write throughput = 167 MHz \(\times 2 \times 18\) bits = 6 Gb/s /chip

Peak read throughput = 167 MHz \(\times 2 \times 18\) bits = 6 Gb/s /chip

Peak total throughput, when fully balanced 50-50 reads/writes = \(6 + 6 = 12\) Gb/s /chip

2.5 Volt power supply; Power consumption \(= 1\) Watt @ 167 MHz

\[\text{Power per throughput} = \frac{1\,\text{W}}{12\,\text{Gb/s}} \approx 0.08\,\text{Watt/Gb} \]
Shared "DQ" Data Bus Timing:

Naive Timing:

"ZBT" (Zero Bus Turn-around) Timing:

Underutilization on every read-to-write transition.

D3 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

Micron’s MT57V256 #36

9 Mbits = \( \frac{256K \times 36 \text{ bits}}{\text{clock freq. up to 300 MHz}} \) (!)

T \geq 3.3 \text{ ns}, \text{ bit pulse width} \geq 1.6 \text{ ns}

Burst-of-4 accesses only

(one address every 2 clock cycles)

Peak Throughput = 300 MHz \times 2 \times 36 \text{ bits} = 21.6 \text{ Gb/s}

Throughput with alternating read/writes

\[ \frac{2}{3} \times \text{peak} = 14.4 \text{ Gbps/chip} \]

2.5 Volts Power Supply; Consumption = 1.6 W

\[ \Rightarrow \sim 0.1 \text{ Watt/Gbps} \]

Although the ZBT concept is used, due to the high clock frequency and the unavoidable bus turnaround overhead (multiple drivers on the same wire, each using its own clock (source-synchronous timing)), 1 to 2 clock cycles (=2 to 4 word burst) are lost on every read-to-write transition.
DRAM Basics: Row Address, Column Address, Precharge

- Row Address Decoder
- Column Address Decoder
- Sense Amplifiers
- Read Data
- Column Access Time
- Row Access Time
- Cycle Time
- Multiple accesses within same row are faster:
  - Addr: RA CA1 CA2 CA3
  - Data: D1 D2 D3

Addr: RA CA RA'

Read Data: D
Fast DRAM Example (2001)

Micron MT46 V2 M32

**DDR SDRAM**

(Synchronous DRAM)

- 32-bit (shared DQ) data bus, DDR timing:
  - 2 words x 32 bits each per clock cycle
  - Peak data bus throughput

- Row Address to Column Address: \( t_{RCD} \geq 20 \text{ns} \) (at \( 200 \text{MHz} \): \( 4 \approx \))
- Column Address to Read Data (CAS latency): \( t_{CL} \geq 15 \text{ns} \) (at \( 200 \text{MHz} \): \( 3 \approx \))
- Write Recovery Time (write data to precharge): \( t_{WR} \geq 2 \approx \)
- Precharge Time: \( t_{RP} \geq 20 \text{ns} \) (at \( 200 \text{MHz} \): \( 4 \approx \))
- Cycle Time (same bank): \( t_{RC} \geq 60 \text{ns} \) (at \( 200 \text{MHz} \): \( 12 \approx \))
- Bank-to-Bank Activation (other bank Row-to-Row): \( t_{RRD} \geq 2 \approx \)
- Read-to-Write bus turn-around lost cycles: \( 3 \approx \)
- Write-to-Read same bank lost cycles (write recovery time): \( 2 \approx \)
- Write-to-Read other bank lost cycles: \( 0 \approx \)

- \( 200 \text{ MHz} \) max. clock frequency
- \( 64 \text{ Mbits} = 2M \times 32 \text{ bits} = 512K \times 32 \text{b} \times 4 \text{ Banks} \)
- \( \approx 1 \text{ Watt} \) at peak access rate, using one bank only, 2.5 Volt.
  (No number given for multibank op.)
Single-Bank Read Access

ACT = Activate
Ba = Bank #a
Ra = Row ≠ Ra Address
RD = Read (the predefined burst size)
Ba = from the active Row within Bank #a
Ca = at Column Address #Ca

trCD ≥ 20 ns (4υ @ 200 MHz)
CL ≥ 15 ns (3υ @ 200 MHz)
Burst Duration tRP
PRECHARGE PERIOD
≥ 20 ns (4υ @ 200 MHz)

trC same-bank cycle time
≥ 60 ns (12υ @ 200 MHz)
Single-Bank Write Access

ACT_Ba = Activate Bank #a
WR_Ba = into the active Row of Bank #a
Ra = Row Address Ra
Ba = Bank #a
Ca = at Column Address Ca

Da_i = i-th word of burst destined to Ba, Ra, Ca

t_{RC} = same bank cycle time
≥ 60 ns (12ns @ 200 MHz)

t_{PRE} = PRECHARGE PERIOD
≥ 20 ns (4ns @ 200 MHz)

Write Recovery
nxt CK

Burst Duration
nxt CK

t_{WR} = 2ns

≥ 20 ns (4ns @ 200 MHz)

t_{TRCD} = 20 ns (4ns @ 200 MHz)
Multiple Accesses to Different Columns in the same Row of a Bank.

Read-to-Write:
- 3 cycles lost due to bus turn-around

Write-to-Read:
- 2 cycles lost when same bank
- 0 cycles lost when different 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 → a1, a2
- Read from column Cb → b1, b2
- Write c1, c2 at column Cc
- Write d1, d2 at column Cd
- Read from column Ce → e1, e2

\[ t_{RCD} = 20 \text{ ns} \]

(4x @ 200 MHz)
burst length set to 8; each successive READ command interrupts the preceding burst, resulting in net bursts of 6.