2.1 Buffer Memory Technology

- Memory Blocks On-Chip
  - On-chip SRAM area, access rate, power consumption
- 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 Area per Kbi
(0.18-micron CMOS technology example - 2002)

On-Chip SRAM Power Consumption
(0.18-micron, 1.8 Volt CMOS technology example - 2002)
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**
- **0.18-µm CMOS**
- **Area**: 5 banks × 128 Kb/bank × 7.5 mm²/Mb =
  - 0.64 Mb × 7.5 mm²/Mb = 4.8 mm²
- **Throughput**: 320 bits × 200 Macc/s = **64 Gb/s**
- **Power Consumption**:
  - 5 banks × 0.7 mW/MHz × 200 MHz = **0.7 W**
On-Chip SRAM Buffer Example (ii): 256-Byte wide

- **Width**: 1 average-size IP packet = 
  
  \[ 256 \text{ Bytes} = 2048 \text{ bits} = 32 \text{ blocks} \times 64 \text{ b/bl} \]

- **Two-port**, 1024 packets \( \times \) 256 B = 256 KB = 2 Mb

- **0.18-µm CMOS**

- **Area**: \( 32 \times 64 \text{ Kb} \times 15 \text{ mm}^2/\text{Mb} = 2 \text{ M} \times 15 \) = \( 30 \text{ mm}^2 \)

- **Throughput**: 2 ports \( \times \) 2048 b/port \( \times \) 260 MHz \( \approx \)

  \[ 1 \text{ Tb/s} \] (I) (500 Gb/s writes + 500 Gb/s reads)

- **Power Consumption**:

  32 banks \( \times \) 2 ports \( \times \) 0.25 mW/MHz \( \times \) 260 MHz \( \approx \)

  \[ 4.2 \text{ W} \]

- **Conclusion**: “no problem” on-chip, except for small packets

---

Power Consumption to Throughput Ratio (1 of 2)

- **(1) On-Chip Buffer Memories**:

  - **0.18-µm CMOS**:

    - 1-port, \( \times 16 \): \( \approx \) 0.15 mW/MHz = 0.15 mW / 16 Mbps = 9.4 mW/Gbps
    - 1-port, \( \times 32 \): \( \approx \) 0.25 mW/MHz = 0.25 mW / 32 Mbps = 7.8 mW/Gbps
    - 1-port, \( \times 64 \): \( \approx \) 0.40 mW/MHz = 0.40 mW / 64 Mbps = 6.2 mW/Gbps

- **Conclusion**: 5 to 10 mW / Gbps on-chip buffer memories

- (2-port memories seem to offer lower consumption/Gbps)
Power Consumption to Throughput Ratio (2 of 2)

- (1) Chip-I/O Pin Power Consumption:
- both directions of a high-speed serial off-chip transceiver (without equalization, which consumes considerably)
  - 0.18-µm CMOS: 25 to 40 mW / Gbps chip-to-chip comm
  - 0.11-µm CMOS: ≈ 15 to 25 mW / Gbps
- copper cable power consumption is very small, compared
  ⇒ Chip-to-chip communication costs much more than buffering
- Total chip power limit (few tens of Watts) limits throughput!

Off-Chip Memory - or other networking I/O chips:
How to Increase Chip-to-Chip Communication Throughput?

Old SRAM Read (flow through):

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

(2) DDR (Double Data Rate) Timing

Traditional Synchronous shift:

Transmit and receive with a positive-edge-triggered register

DDR Interface:

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

...further increasing the data pin through part 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. Use 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\(\text{in}\) and Q\(\text{out}\) Paths:**
- Time-mux'd write & read addresses
- Addr → in write data → SRAM → out read data → Q

**Shared "DQ" Data Bus:**
- Addr → all addresses
- DQ → all data

- Data path underutilization when imbalanced (\(\neq 50\% - 50\%\))
- Read/write transactions

- Bus turn-around overhead:
- Data bus underutilization when frequently switching between read & write transactions

---

**Modern SRAM Chip Technology vs. Separate D\(\text{in}\) & Q\(\text{out}\) Paths:**

"QDR" (Quad Data Rate) SRAM

- Addr → DDR addresses
- Burst of 2 DDR in
- QDR → Q
- Other versions: "burst-of-4":
  - Address path is plain (NOT DDR)
  - Each addr. refers to 4 data words

---

CS-534, Copyright Univ. of Crete
Example QDR SRAM (2001) Micro's MT54V512H18

9 Mbits = 512 K x 18 bit
Clock freq. up to 167 MHz
T=6ns pulse; bit and #3ms DDR

Peak write throughput = 167 MHz x 2 x 18 bits = 6 Gbps/chip
Peak read throughput = 167 MHz x 2 x 18 bits = 6 Gbps/chip
Peak total throughput, when fully balanced 50-50 reads/writes = 6 + 6 = 12 Gbps/chip

2.5 Volt power supply; Power consumption 2.1 Watt @ 167 MHz
Power per throughput = \( \frac{1}{12 \text{Gbps}} \approx 0.08 \text{ Watt/Gbps} \)

Shared "DQ" Data Bus Timing:

Naive Timing:

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

D1 has not yet been written at M[4], when reading from M[2] starts...
... need to bypass row when A2=A1
Example Shown Bus SRAM at the top current performance (2001):

Micron's MT57V256H36

- 9 Mbits = 256 K x 36 bits
- Clock freq. up to 300 MHz (2)
- T = 33 ns, bit pulse width = 3.6 ns
- Burst-of-4 accesses only (one address every 2 clock cycles)

Peak throughput = 300 MHz x 2 x 361 = 216 Gbps

Throughput with alternating read/write access:
\[
\frac{2}{3} \times \text{peak} = \frac{144}{\text{chip}} \quad \text{Gbps}
\]

2.5 Volts Power Supply; Consumption = 1.6 W

\( \Rightarrow \text{Vout} \approx 0.1 \text{ Gbps} \)

Although the 2BT concept is used, due to the high clock frequency and the unavoidable bus turn around overhead (multiple drivers on the same wire), each using its own clock (source-synchronous timing)), 1 to 2 clock cycles (~24 ns and 24) are lost on every read-to-write transition.

---

DRAM Basics: Row Address, Column Address, Precharge

Row Address Decoder

Multiple accesses within same row are faster:

Add. RA CA1 CA2 CA3

Data: D1 D2 D3 D4
Fast DRAM Example (2001)

- Micron M76 V2 M32
- DDR SDRAM (Synchronous DRAM)
- 32-bit (shared DQ) data bus, DDR timing
- 6 words x 32 bits each per clock cycle
- 64 Mbits = 2M x 32 bits = 512K x 32 x 4 banks
- 200 MHz max clock frequency
- Only 1 byte at peak access rate, using one bank only, 2.5 V 64
- No clock gating for multi-bank op.
- Row Address to Column Address: \( t_{RC} \geq 200 \text{ns} \) (mean: 4\( \mu \text{sec} \))
- Column Address to Read Data (CAS latency): \( t_{CL} \geq 150 \text{ns} \) (mean: 3\( \mu \text{sec} \))
- Write Recovery Time (write data 4 precharge): \( t_{WR} \geq 2 \mu \text{sec} \)
- Precharge Time: \( t_{RP} \geq 200 \text{ns} \) (mean: 4\( \mu \text{sec} \))
- Cycle Time (same bank): \( t_{RC} \geq 60 \text{ns} \) (mean: 12\( \mu \text{sec} \))
- Bank-to-Bank Activation (other bank Row-to-Row): \( t_{RRD} \geq 2 \mu \text{sec} \)
- Read-to-Write bus turn-around last cycles: \( 3 \mu \text{sec} \)
- Write-to-Read same bank last cycles (write recovery time): \( 2 \mu \text{sec} \)
- Write-to-Read other bank last cycles: \( 0 \mu \text{sec} \)

Single-Bank Read Access

<table>
<thead>
<tr>
<th>Event</th>
<th>Address Bus</th>
<th>Data Bus</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>RA</td>
<td>DP</td>
</tr>
<tr>
<td></td>
<td>Bank 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ra Address</td>
<td></td>
</tr>
<tr>
<td>RD</td>
<td>)4 (68 bits)</td>
<td>C (512 bits)</td>
</tr>
<tr>
<td></td>
<td>)56 (512 bits)</td>
<td></td>
</tr>
<tr>
<td>CL</td>
<td>Burst Duration</td>
<td></td>
</tr>
<tr>
<td>( t_{RP} )</td>
<td>PRECHARGE PERIOD</td>
<td></td>
</tr>
<tr>
<td>( t_{RC} )</td>
<td>Burst Duration</td>
<td></td>
</tr>
<tr>
<td>( t_{RP} )</td>
<td>Some-bank cycle time</td>
<td></td>
</tr>
<tr>
<td>( t_{RC} )</td>
<td>( 60 \mu \text{sec} ) (512 bits)</td>
<td></td>
</tr>
<tr>
<td>( ACT )</td>
<td>( 2 \mu \text{sec} ) (512 bits)</td>
<td></td>
</tr>
<tr>
<td>( RB )</td>
<td>Row = Ra Address</td>
<td></td>
</tr>
<tr>
<td>( RD )</td>
<td>Read (RA)</td>
<td></td>
</tr>
<tr>
<td>( RB )</td>
<td>Bank 0</td>
<td></td>
</tr>
<tr>
<td>( Ca )</td>
<td>Column Address</td>
<td></td>
</tr>
</tbody>
</table>

CS-534, Copyright Univ. of Crete
burst length set to 8; each successive READ command interrupts the preceding burst, resulting in net bursts of 6.