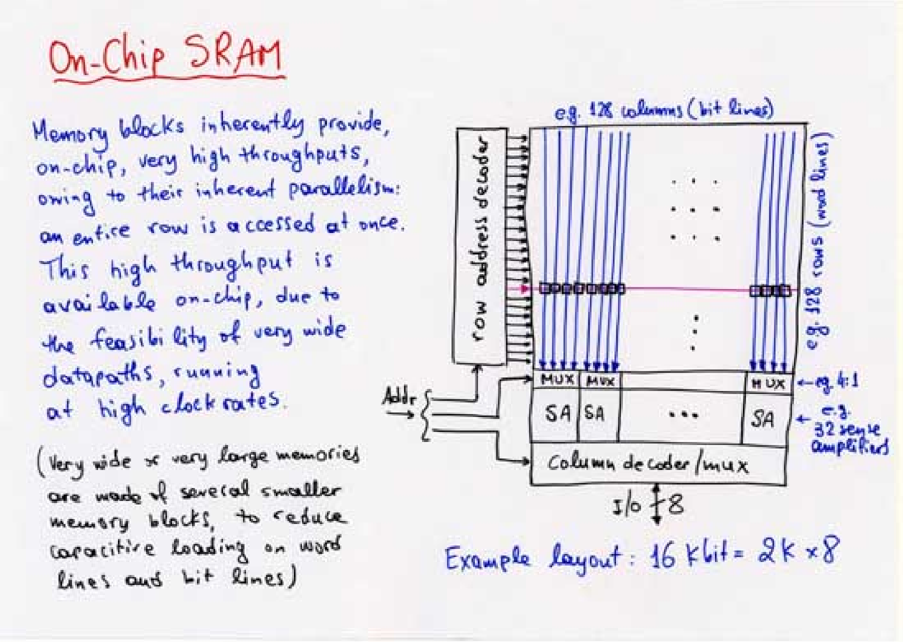
2.1 Buffer Memory Technology

- 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



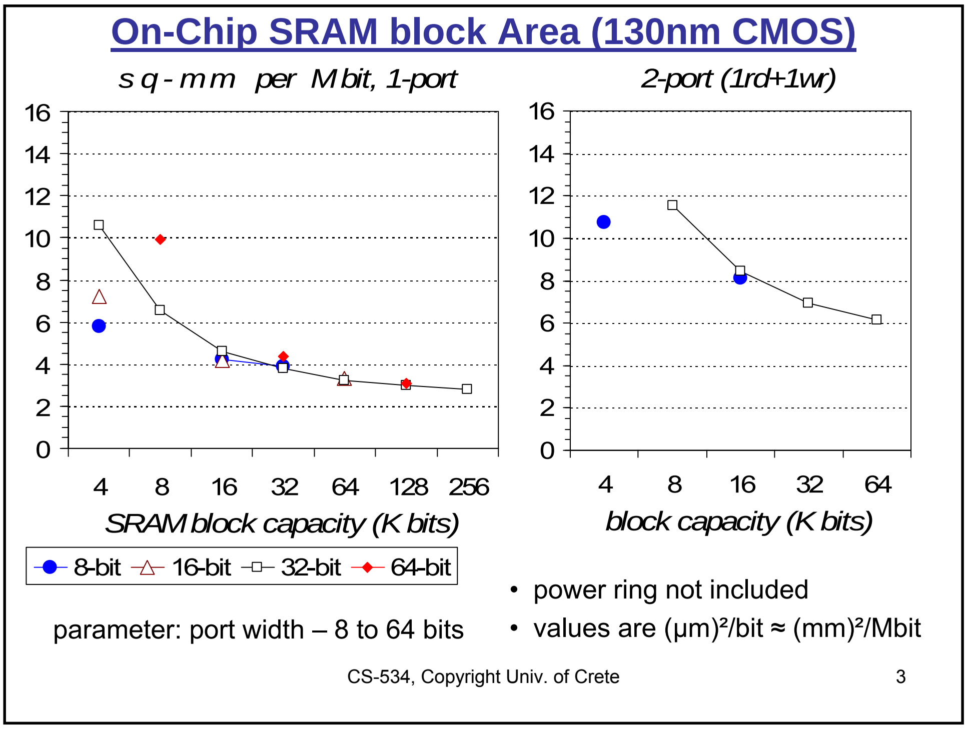
On-Chip SRAM

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 & very large memories are made of several smaller memory blocks, to reduce capacitive loading on word lines and bit lines)

e.g. 128 columns (bit lines)
row address decoder
e.g. 128 rows (word lines)
Addr
MUX
MUX
MUX
← e.g. 4:1
SA
SA
...
SA
← e.g. 32 sense amplifiers
Column decoder / mux
I/O
8

Example layout: 16 Kbit = 2K × 8

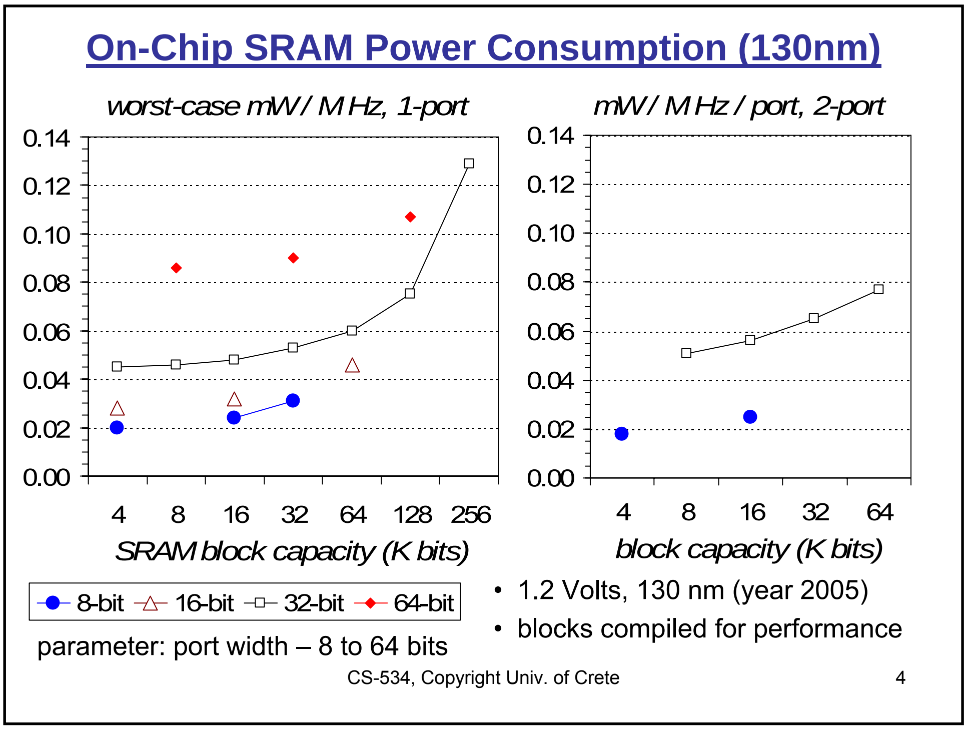


On-Chip SRAM block Area (130nm CMOS)

parameter: port width – 8 to 64 bits

- power ring not included
- values are (μm)²/bit ≈ (mm)²/Mbit

CS-534, Copyright Univ. of Crete 3

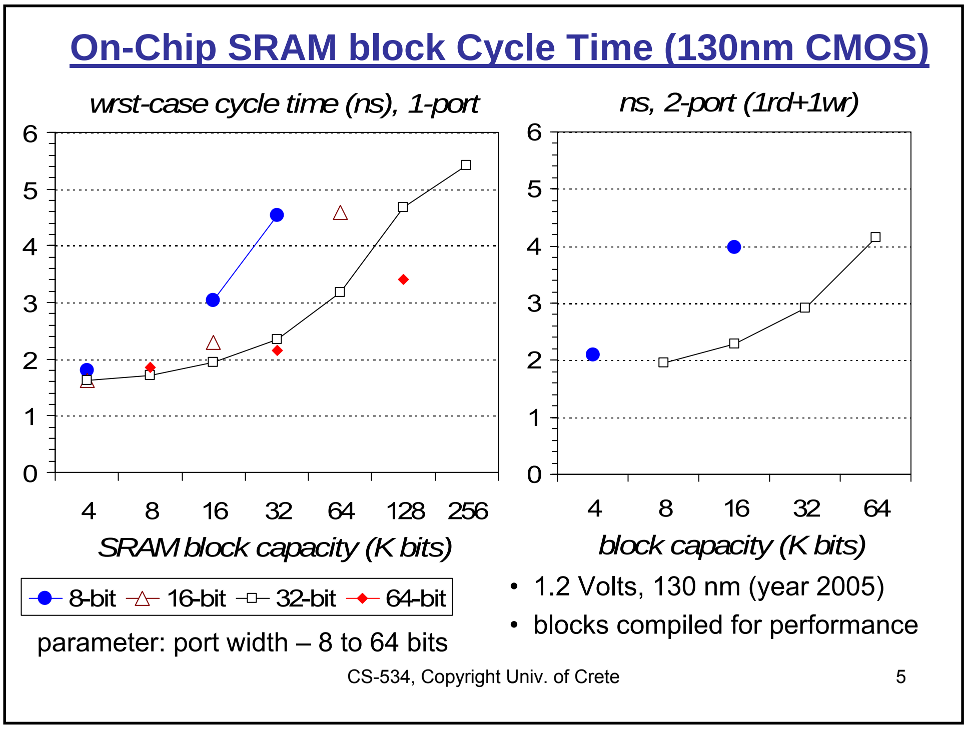


On-Chip SRAM Power Consumption (130nm)

- 1.2 Volts, 130 nm (year 2005)
- blocks compiled for performance

parameter: port width – 8 to 64 bits

CS-534, Copyright Univ. of Crete 4

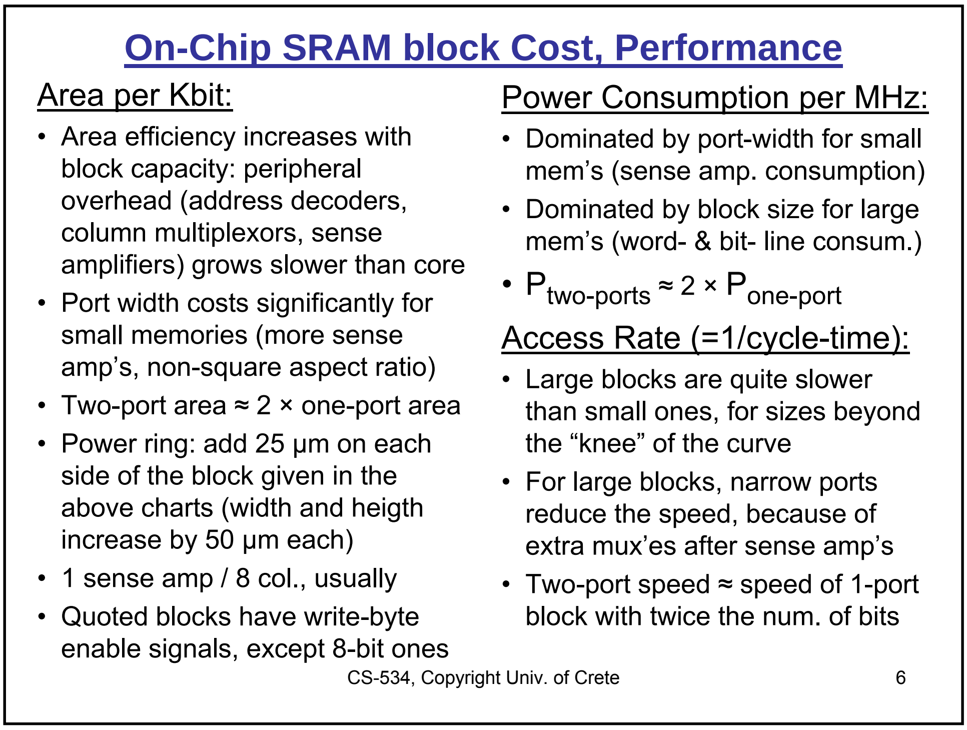


On-Chip SRAM block Cycle Time (130nm CMOS)

parameter: port width – 8 to 64 bits

- 1.2 Volts, 130 nm (year 2005)
- blocks compiled for performance

CS-534, Copyright Univ. of Crete 5



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 ≈ 2 × one-port area
- Power ring: add 25 μm on each side of the block given in the above charts (width and heigth increase by 50 μm each)
- 1 sense amp / 8 col., usually
- Quoted blocks have write-byte enable signals, except 8-bit ones

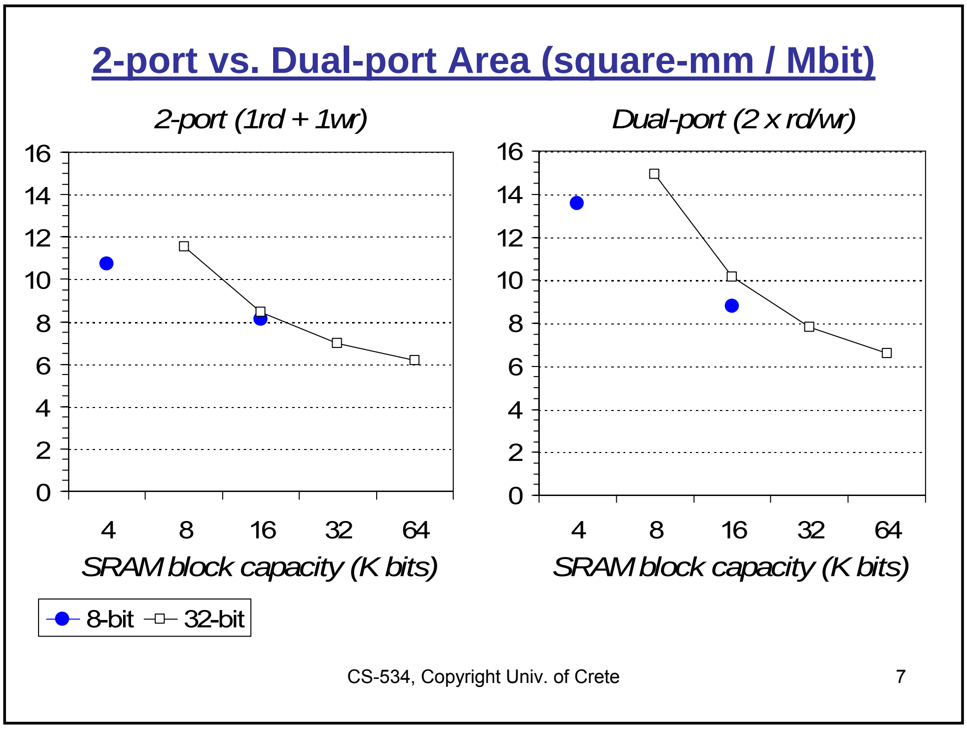
Power Consumption per MHz:

- 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):

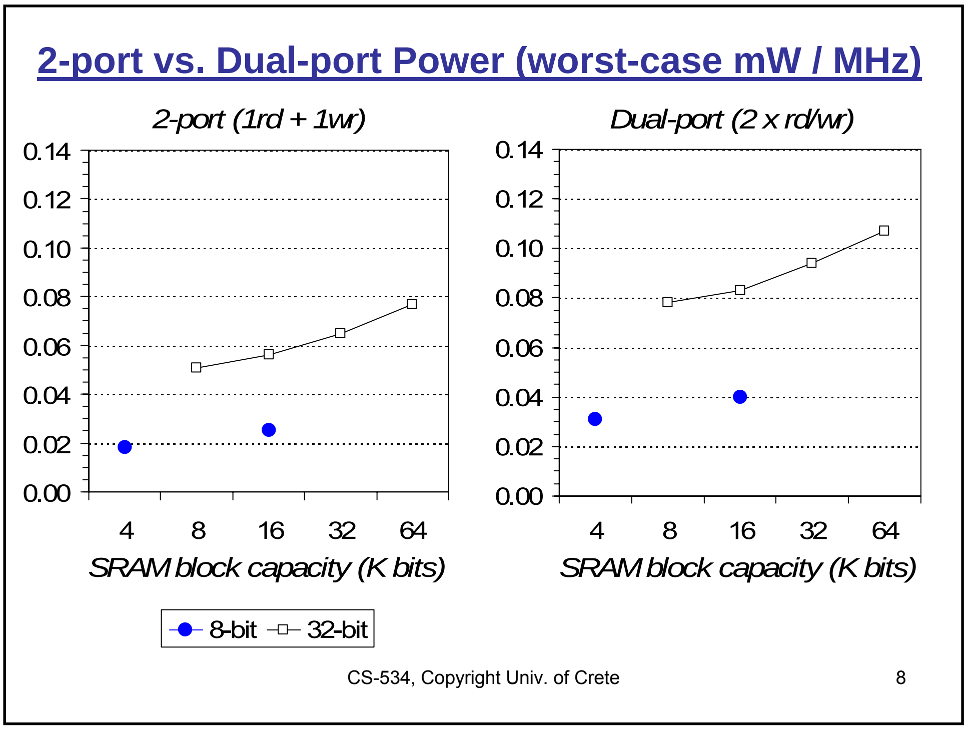
- Large blocks are quite slower than small ones, for sizes beyond the "knee" of the curve
- For large blocks, narrow ports reduce the speed, because of extra mux'es after sense amp's
- Two-port speed ≈ speed of 1-port block with twice the num. of bits

CS-534, Copyright Univ. of Crete 6



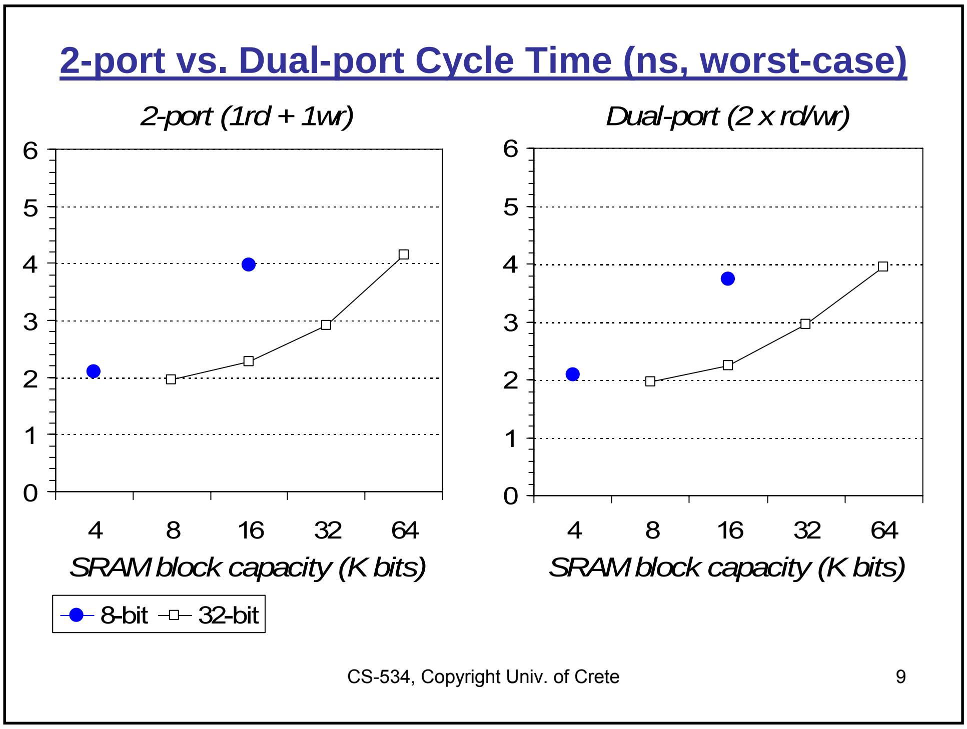
2-port vs. Dual-port Area (square-mm / Mbit)

CS-534, Copyright Univ. of Crete 7



2-port vs. Dual-port Power (worst-case mW / MHz)

CS-534, Copyright Univ. of Crete 8



2-port vs. Dual-port Cycle Time (ns, worst-case)

[intentionally left blank]

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 mm^2/Mb =
  = 0.64 Mb × 3 mm^2/Mb ≈ **2 mm^2**
- 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 ≈ 1 average-size IP packet =
  = 256 Bytes = 2048 bits = 64 blocks × 32 bits/block
- Two-port (1rd+1wr), 2048 packets × 256 B = 512 KB = 4 Mb
- 130 nm CMOS, 1.2 Volts
- Area: 64 × 64 Kb × 6.1 mm^2/Mb = 4 M × 6.1 ≈ **25 mm^2**
- Throughput: 2 ports × 2048 b/port × 240 MHz ≈ **1 Tb/s**
  (500 Gb/s writes + 500 Gb/s reads)
- Power Consumption:
  64 banks × 2 ports × 0.08 mW/MHz × 240 MHz ≈ **2.4 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

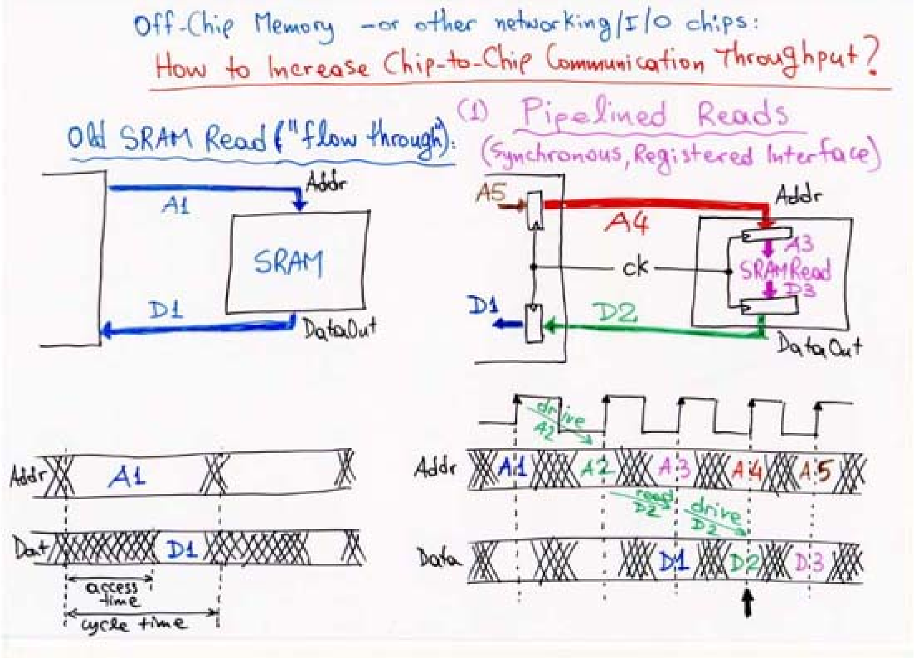
- Total chip power consumption (up to few tens of Watts) limits total chip throughput to about 1 Tbps/chip or less

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)



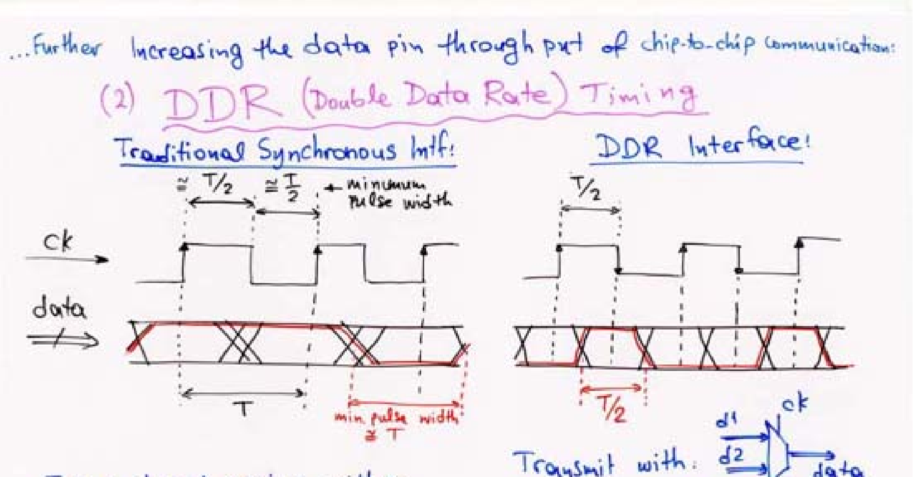
CS-534, Copyright Univ. of Crete 15

...Further increasing the data pin throughput of chip-to-chip communication:

(2) DDR (Double Data Rate) Timing

Traditional Synchronous Intf:

DDR Interface:

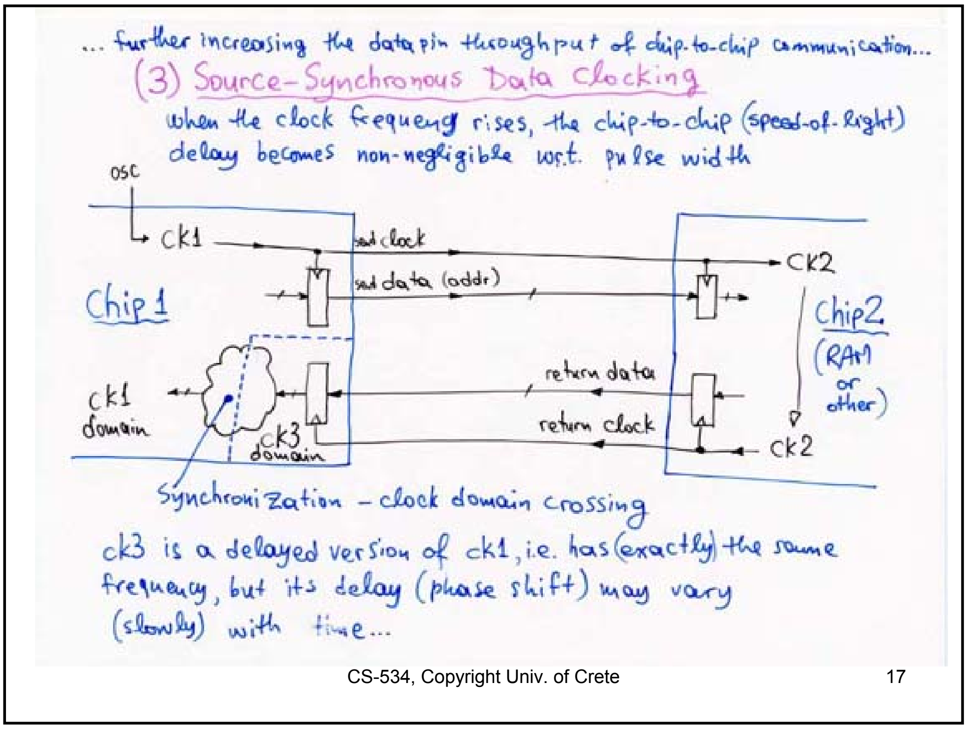


Transmit and receive with a positive-edge-triggered register

Transmit with:

Receive with: two registers:
- one positive-edge-tr. register
- one negative-edge-tr. register

CS-534, Copyright Univ. of Crete 16



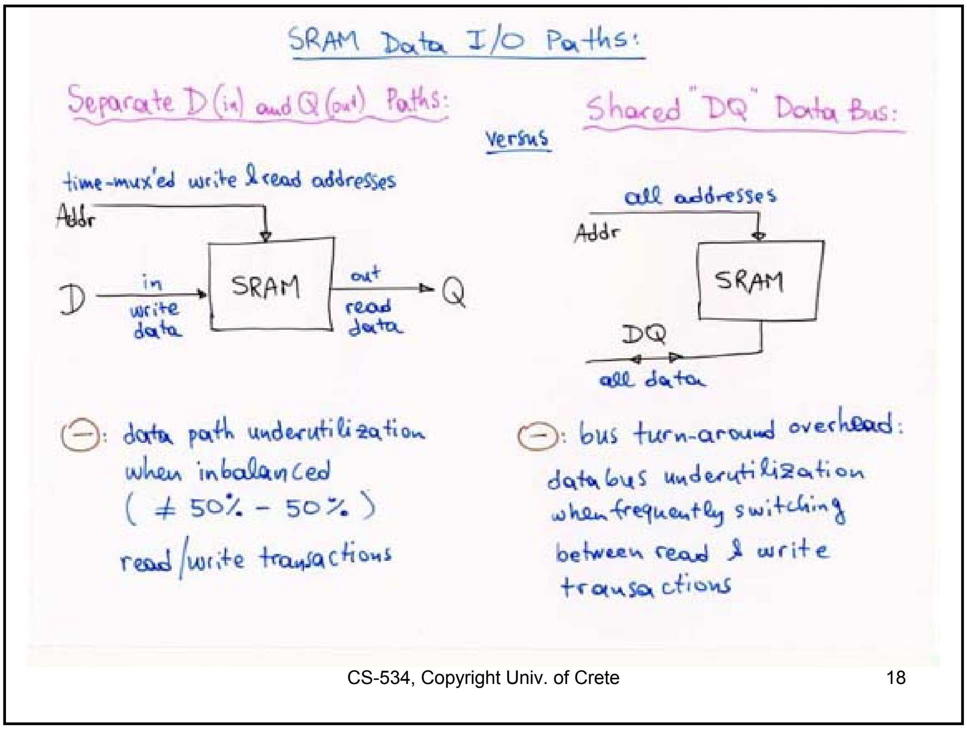
... further increasing the data pin throughput of chip-to-chip communication...

(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...

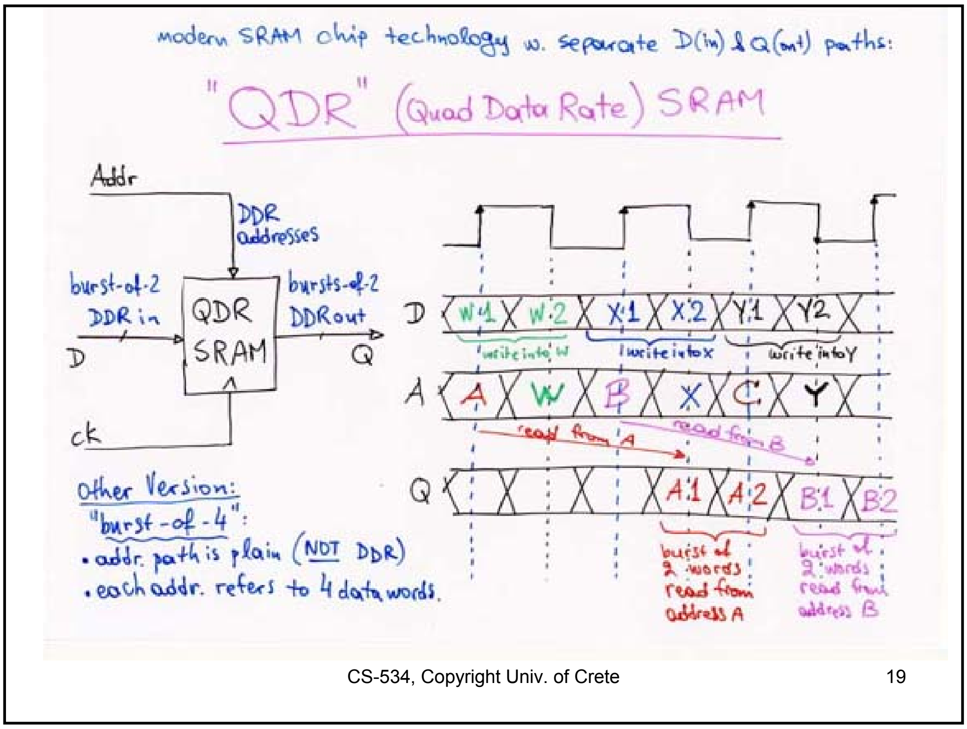


SRAM Data I/O Paths:

Separate D(in) and Q(out) Paths: versus Shared "DQ" Data Bus:

⊖: data path underutilization when inbalanced (≠ 50% - 50%) read/write transactions

⊖: bus turn-around overhead: data bus underutilization when frequently switching between read & write transactions

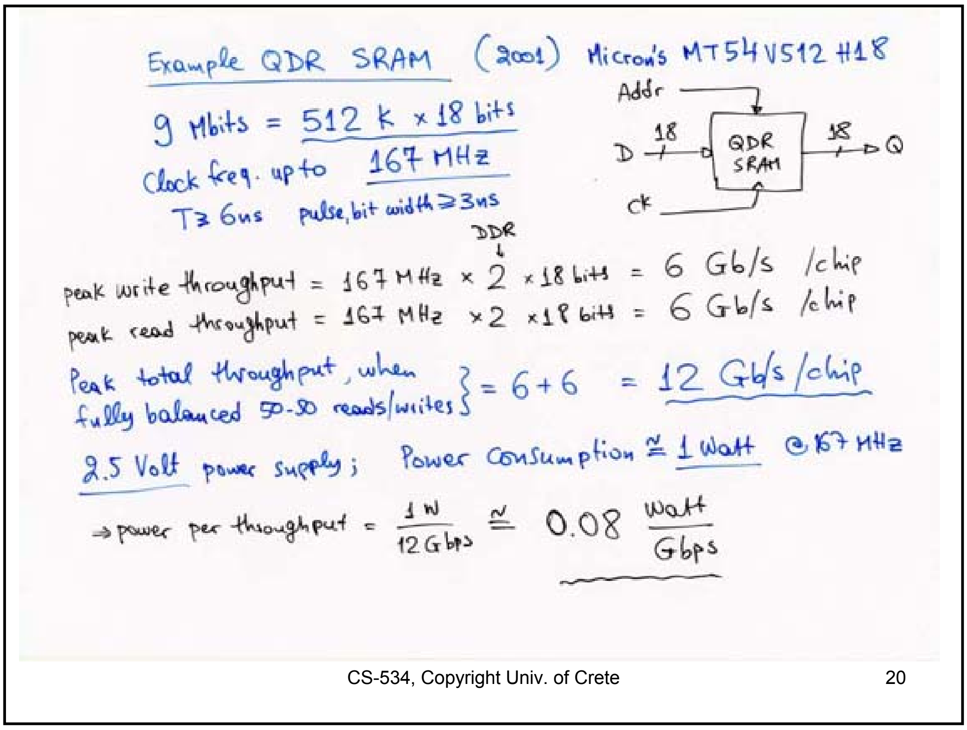


modern SRAM chip technology w. separate D(in) & Q(out) paths:

"QDR" (Quad Data Rate) SRAM

Other Version:
"burst-of-4":
- addr. path is plain (NOT DDR)
- each addr. refers to 4 data words.

CS-534, Copyright Univ. of Crete 19



Example QDR SRAM (2001) Micron's MT54V512H18

9 Mbits = 512 K × 18 bits

Clock freq. up to 167 MHz

T ≥ 6ns pulse, bit width ≥ 3ns

Addr
D
18
QDR SRAM
18
Q
ck

peak write throughput = 167 MHz × $\overset{\text{DDR}}{2}$ × 18 bits = 6 Gb/s /chip

peak read throughput = 167 MHz × 2 × 18 bits = 6 Gb/s /chip

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

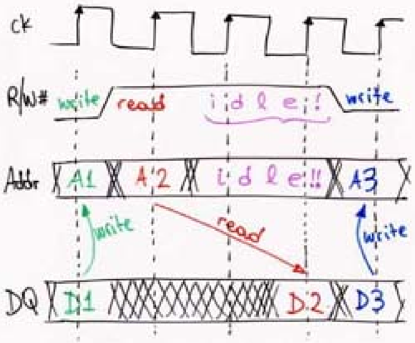
2.5 Volt power supply; Power Consumption ≅ 1 Watt @167 MHz

$\Rightarrow$ power per throughput = $\frac{1\text{ W}}{12\text{ Gbps}} \cong 0.08 \frac{\text{Watt}}{\text{Gbps}}$

CS-534, Copyright Univ. of Crete 20

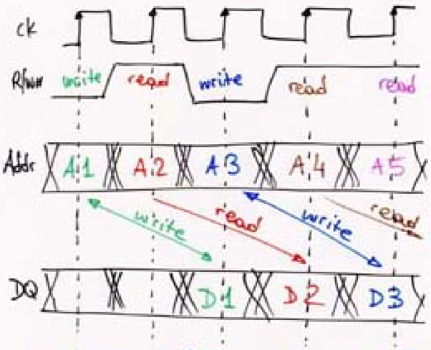
Shared "DQ" Data Bus Timing:

Naive Timing:



Underutilization on every read-to-write transition

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



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 at the top current performance (2001):

Micron's MT57 V256 H36 DDR SRAM

9 Mbits = 256K x 36 bits

Clock frequ. up to 300 MHz (!)

T ≥ 3.3 ns, bit pulse width ≥ 1.6 ns

Burst-of-4 accesses only

(one address every 2 clock cycles)

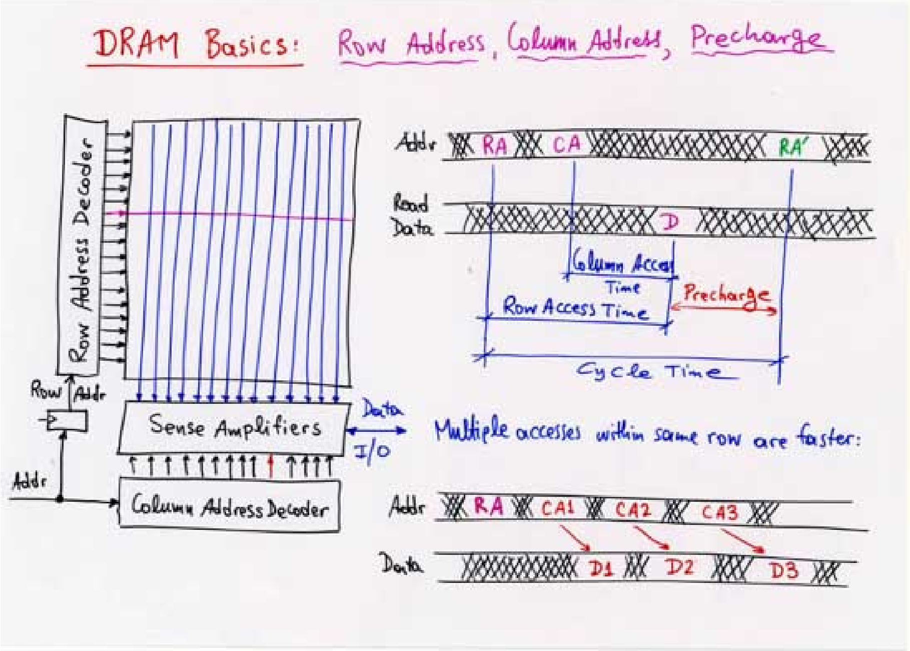
Peak Throughput = $300\ \text{MHz} \times \overset{DDR}{2} \times 36\text{b} = 21.6\ \text{Gb/s}$

Throughput with alternating read/writes $\Big\} = \frac{2}{3} \times \text{peak} = 14.4\ \frac{\text{Gbps}}{\text{chip}}$

2.5 Volts Power Supply; Consumption = 1.6 W

$\Rightarrow \sim 0.1\ \frac{\text{Watt}}{\text{Gbps}}$

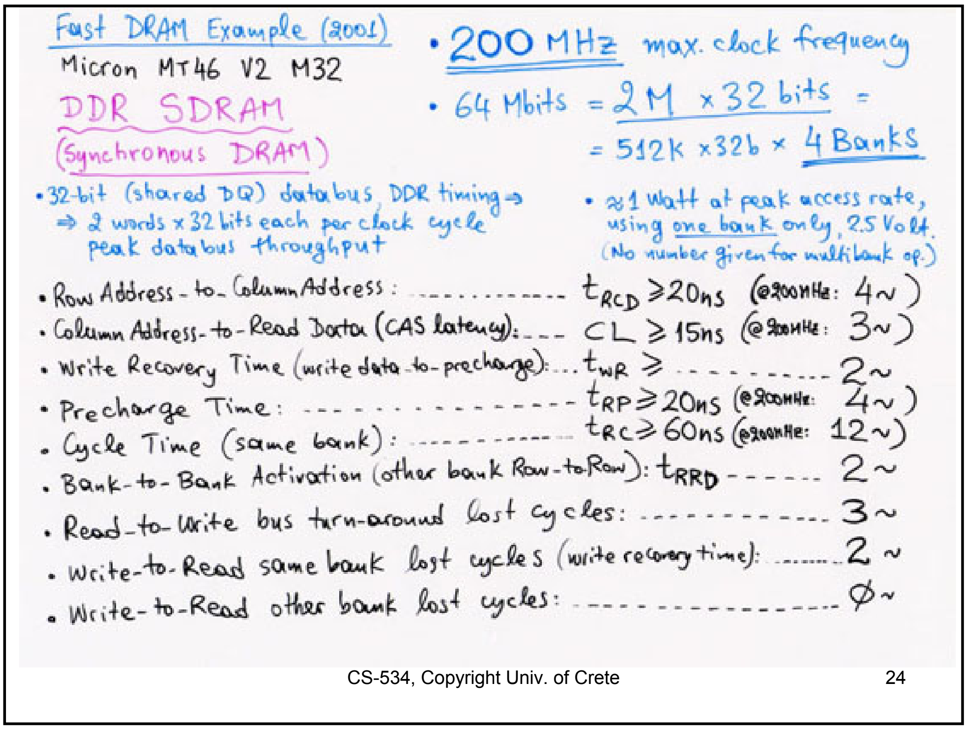
⇒ Although the ZBT concept is used, due to the high clock freqnency and the unavoidable bus turnaround overhead (multiple drivers on the same wire, each using its own clock (source-synchronous timing)), 1 to 2 clock cycles (= 2 to 4 word burst) are lost on every read-to-write transition.



DRAM Basics: Row Address, Column Address, Precharge

Multiple accesses within same row are faster:

CS-534, Copyright Univ. of Crete 23



Fast DRAM Example (2001)

Micron MT46 V2 M32

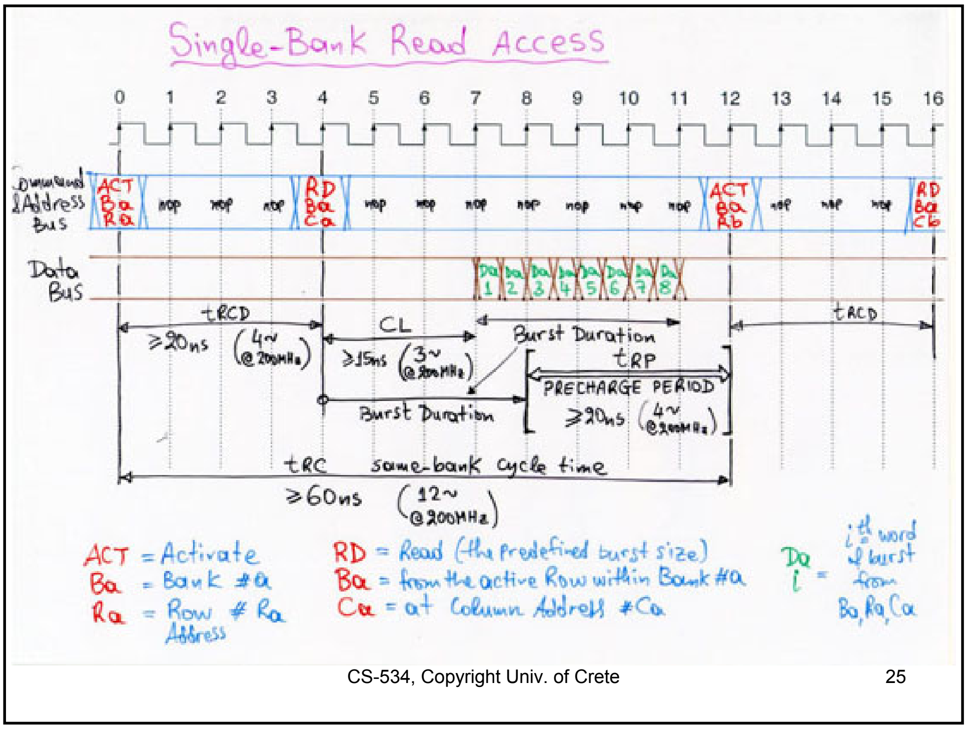
DDR SDRAM

(Synchronous DRAM)

- **200 MHz** max. clock frequency
- 64 Mbits = 2M × 32 bits =
  = 512k × 32b × 4 Banks
- 32-bit (shared DQ) databus, DDR timing ⇒
  ⇒ 2 words × 32 bits each per clock cycle
  peak databus throughput
- ≈1 Watt at peak access rate, using one bank only, 2.5 Volt.
  (No number given for multibank op.)

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

CS-534, Copyright Univ. of Crete 24



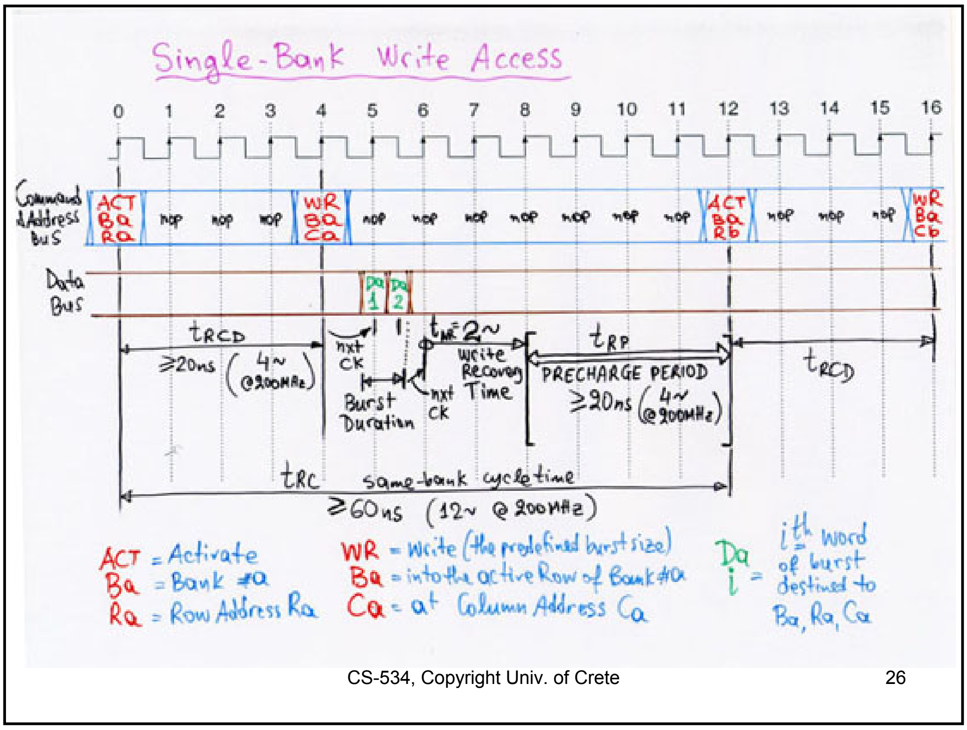
Single-Bank Read Access

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

Da i = ith word of burst from Ba, Ra, Ca

CS-534, Copyright Univ. of Crete 25



Single-Bank Write Access

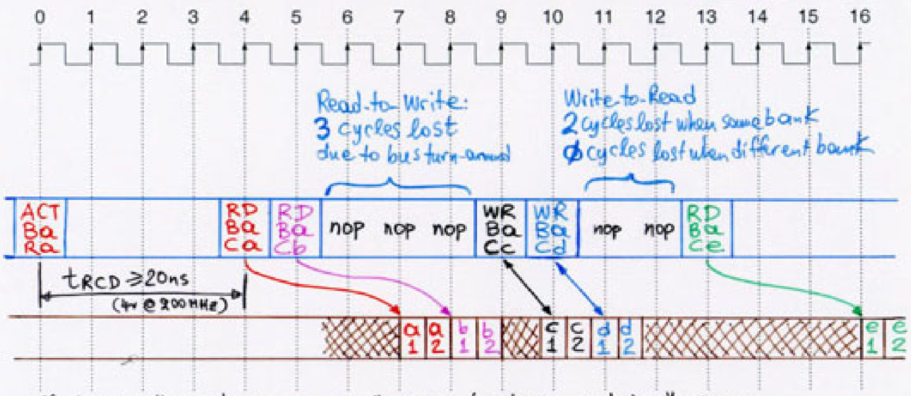
ACT = Activate
Ba = Bank #a
Ra = Row Address Ra

WR = Write (the predefined burst size)
Ba = into the active Row of Bank #a
Ca = at Column Address Ca

Da i = ith word of burst destined to Ba, Ra, Ca

CS-534, Copyright Univ. of Crete 26

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

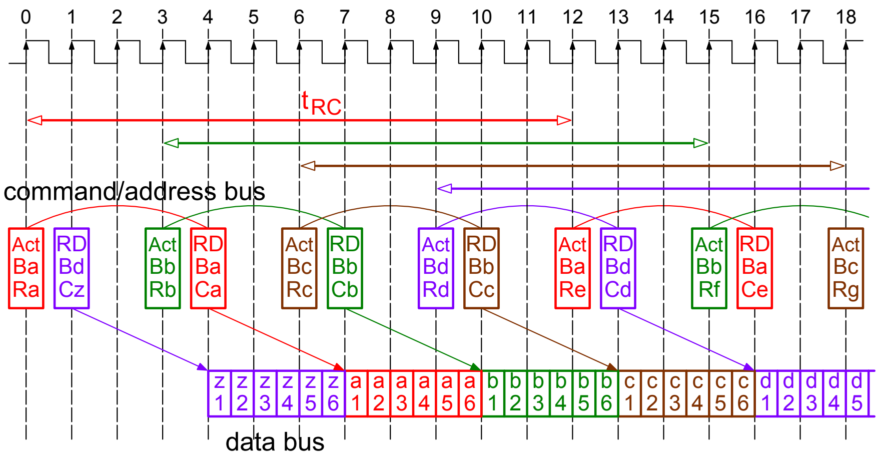


All transactions shown are to the same bank #a, and to the same activated row Ra in that bank.
The transactions shown are:

- Read from column Ca → 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

CS-534, Copyright Univ. of Crete 27

Multi-Bank Operation: Memory Interleavin



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

CS-534, Copyright Univ. of Crete 28