EECS 151/251A
Spring 2019
Digital Design and Integrated Circuits

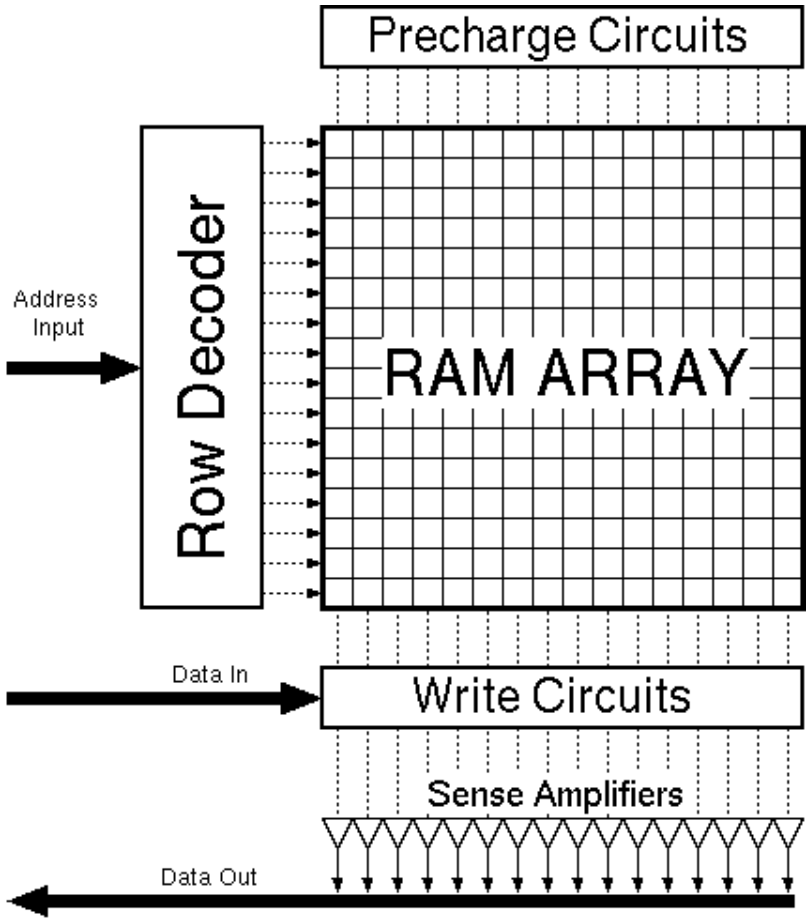
Instructor:
John Wawrzynek

Lecture 18

Memory Blocks

- Multi-ported RAM
- Combining Memory blocks
- FIFOs
- FPGA memory blocks
- Memory block synthesis
- Caches

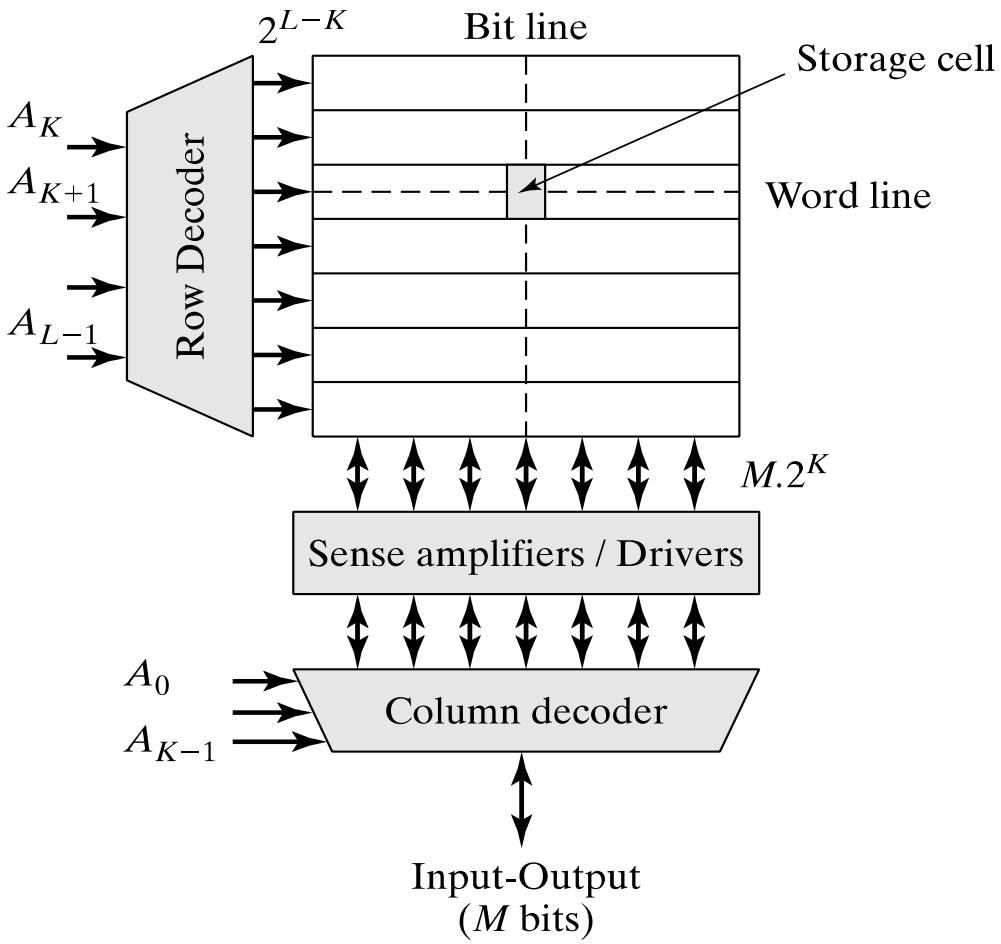
SRAM Block



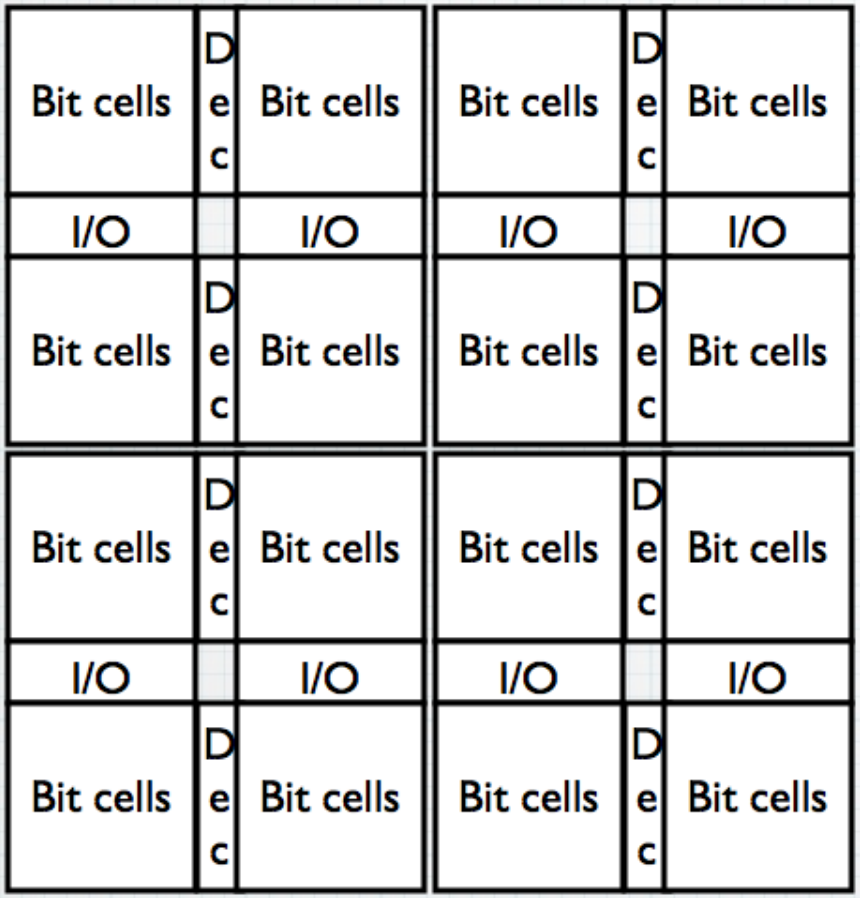
- Extra circuitry and timed control signals needed
  - Periphery circuits add a “fixed” area overhead
  - Row select, sensing, precharge must be sequenced, based on input clock signal
  - Read operation needs a clock: “synchronous read”

Memory Architecture Overview

- **Word lines** used to select a row for reading or writing
- **Bit lines** carry data to/from periphery
- **Core** *aspect ratio* keep close to 1 to help balance delay on word line versus bit line
- **Address bits** are divided between the two decoders
- **Row decoder** used to select word line
- **Column decoder** used to select one or more columns for input/output of data



Building Larger Memories

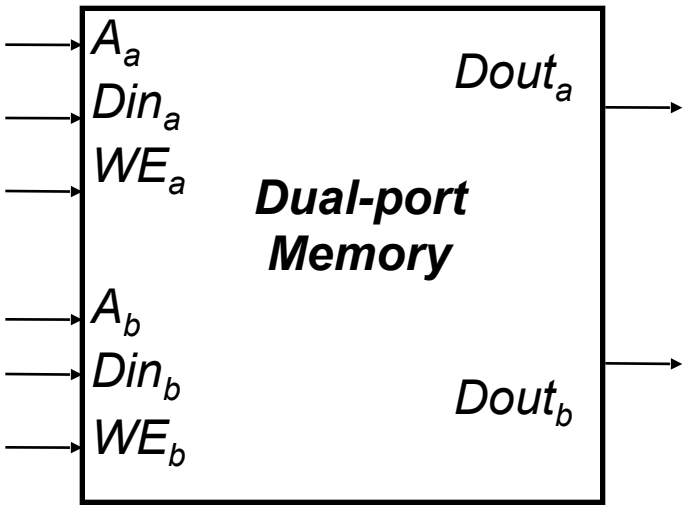


- Large arrays constructed by tiling multiple leaf arrays, sharing decoders and I/O circuitry
  - e.g., sense amp attached to arrays above and below
- Leaf array limited in size to 128-256 bits in row/column due to RC delay of wordlines and bitlines
- Also to reduce power by only activating selected sub-bank
- In larger memories, delay and energy dominated by I/O wiring

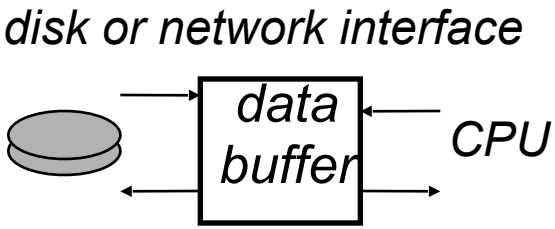
Multi-ported memory

Multi-ported Memory

- Motivation:
  - Consider CPU core register file:
    - 1 read or write per cycle limits processor performance.
    - Complicates pipelining. Difficult for different instructions to simultaneously read or write regfile.
    - Common arrangement in pipelined CPUs is 2 read ports and 1 write port.



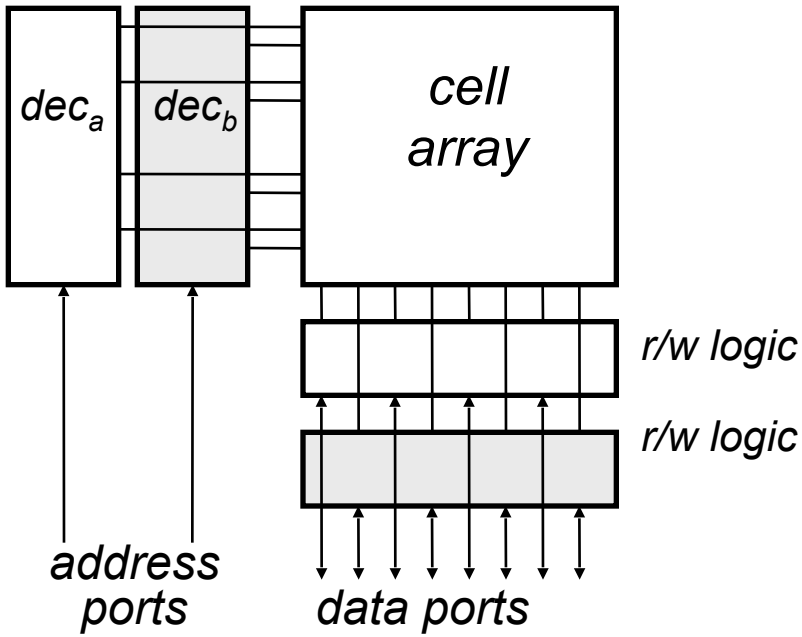
- I/O data buffering:



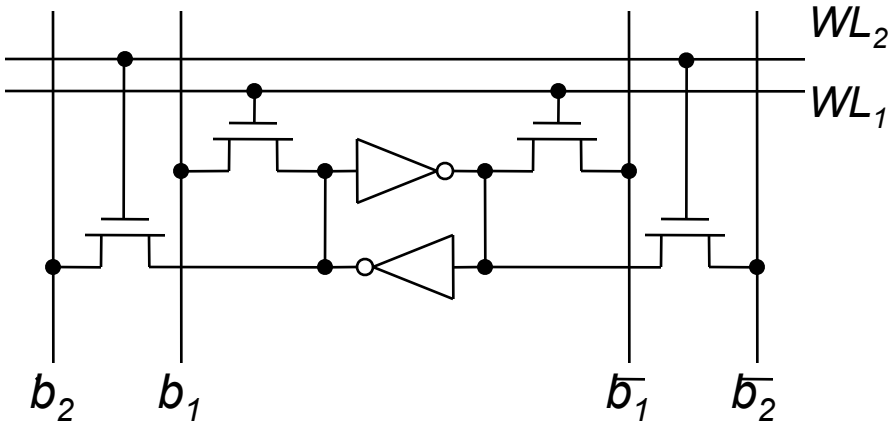
- dual-porting allows both sides to simultaneously access memory at full bandwidth.

Dual-ported Memory Internals

- Add decoder, another set of read/write logic, bits lines, word lines:



- *Example cell: SRAM*



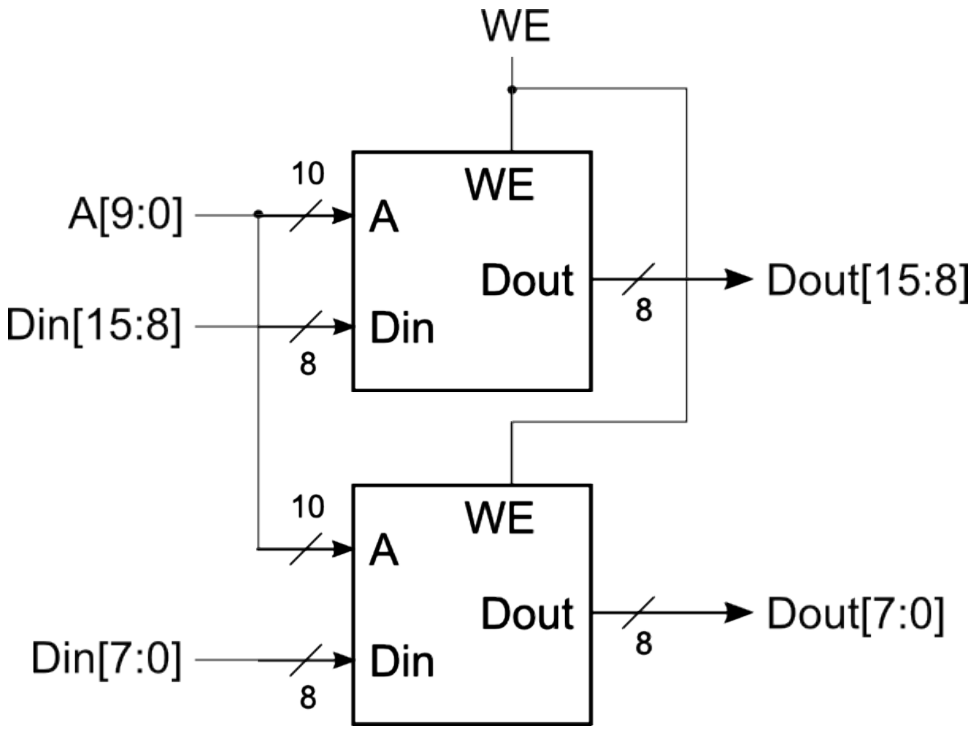
- *Repeat everything but cross-coupled inverters.*
- *This scheme extends up to a couple more ports, then need to add additional transistors.*

Cascading Memory Blocks

Cascading Memory-Blocks

How to make larger memory blocks out of smaller ones.

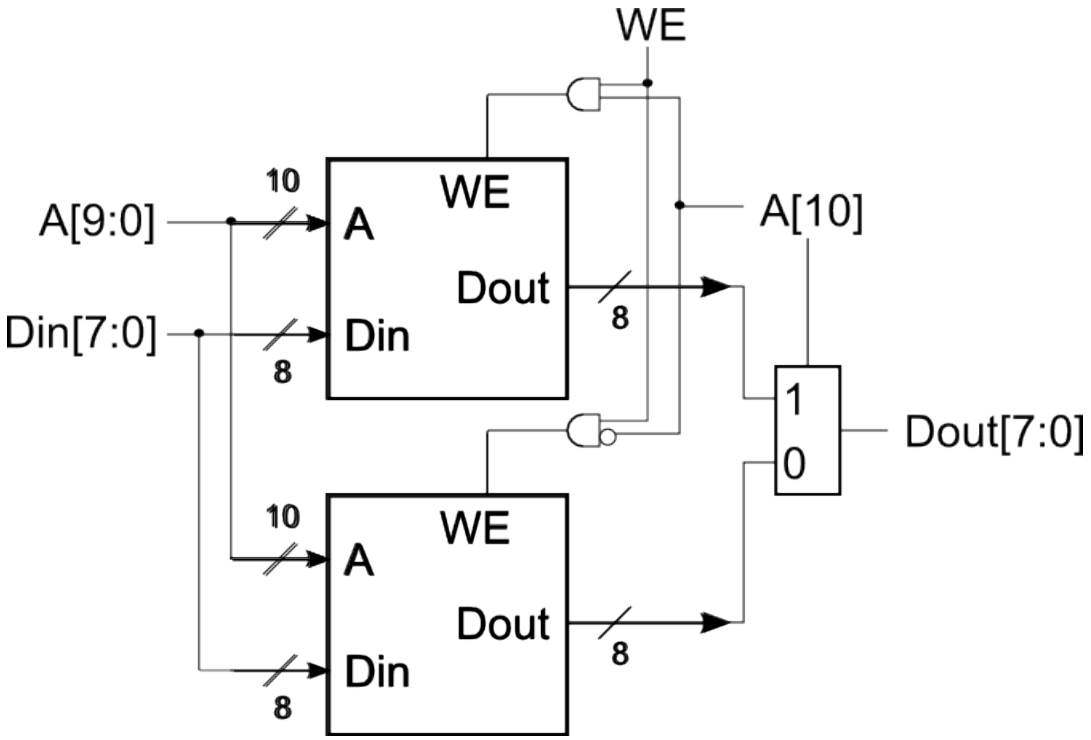
Increasing the width. Example: given 1Kx8, want 1Kx16



Cascading Memory-Blocks

How to make larger memory blocks out of smaller ones.

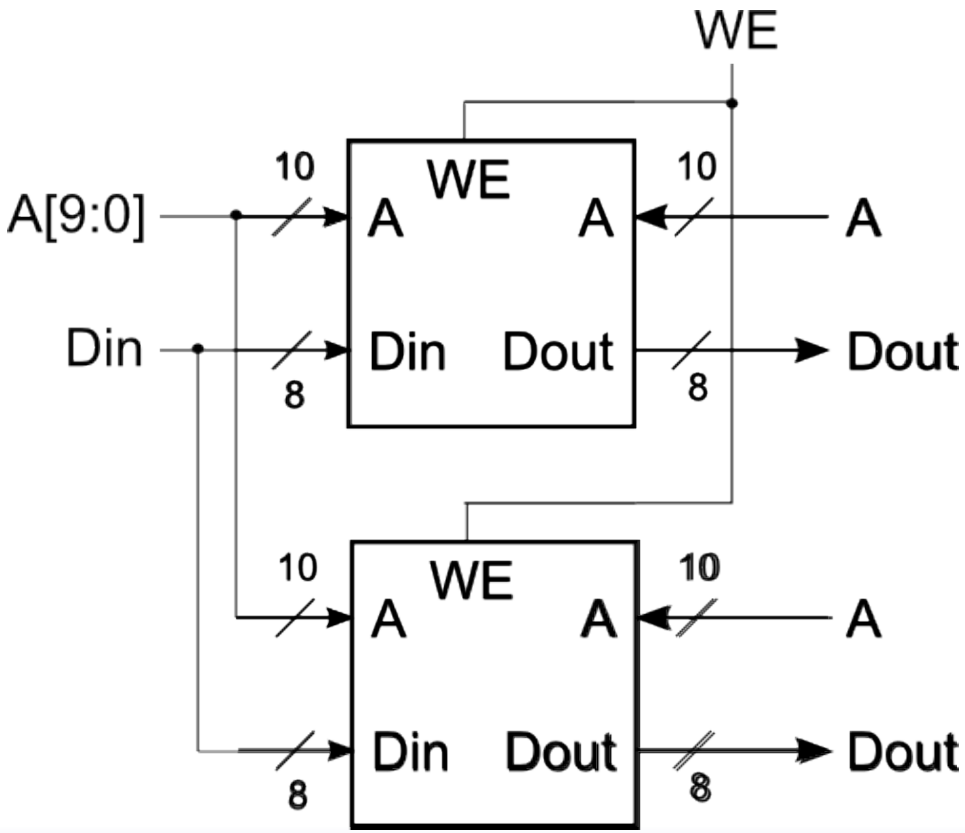
Increasing the depth. Example: given 1Kx8, want 2Kx8



Adding Ports to Primitive Memory Blocks

Adding a read port to a simple dual port (SDP) memory.

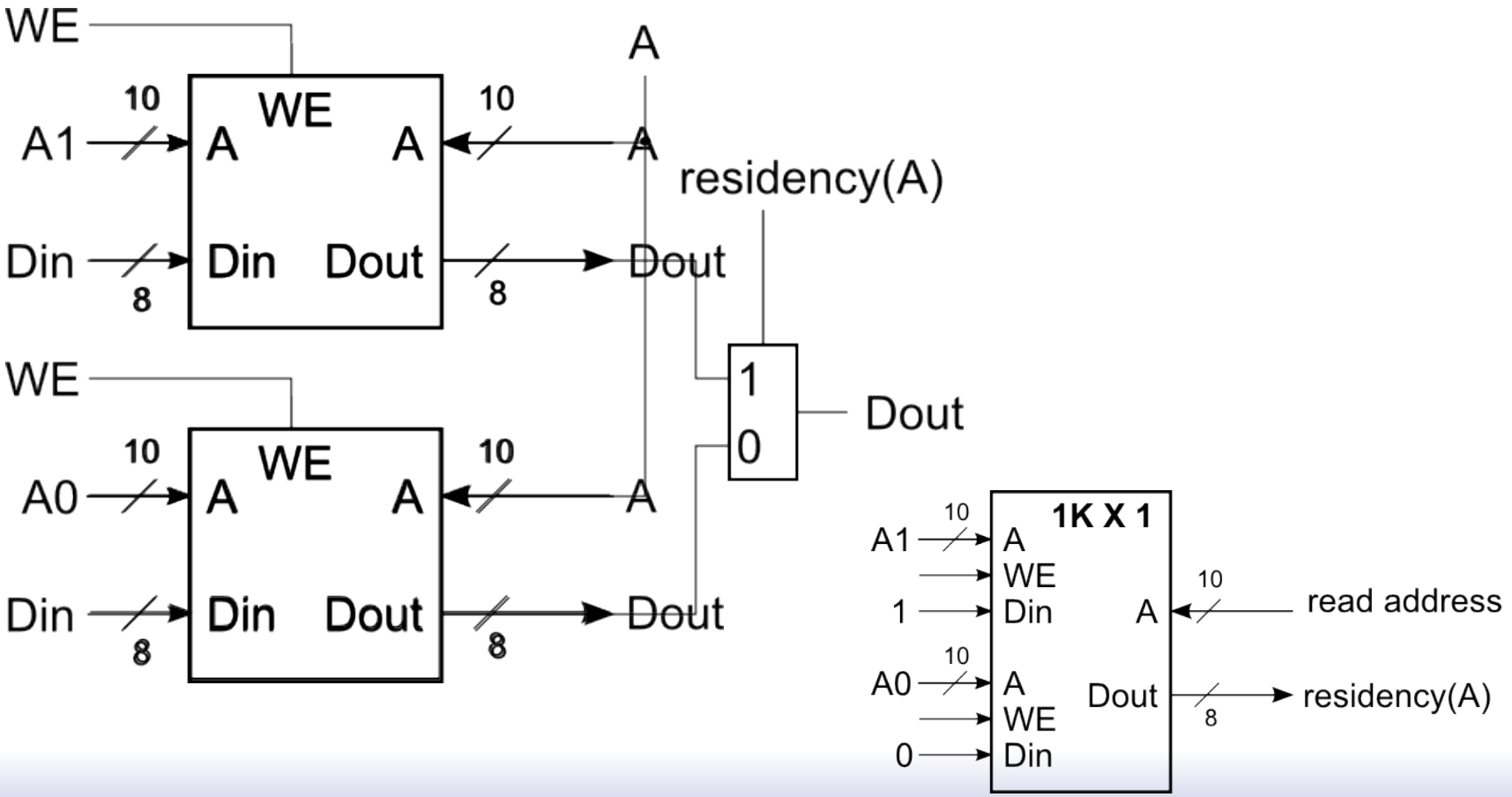
Example: given 1Kx8 SDP, want 1 write & 2 read ports.



Adding Ports to Primitive Memory Blocks

How to add a write port to a simple dual port memory.

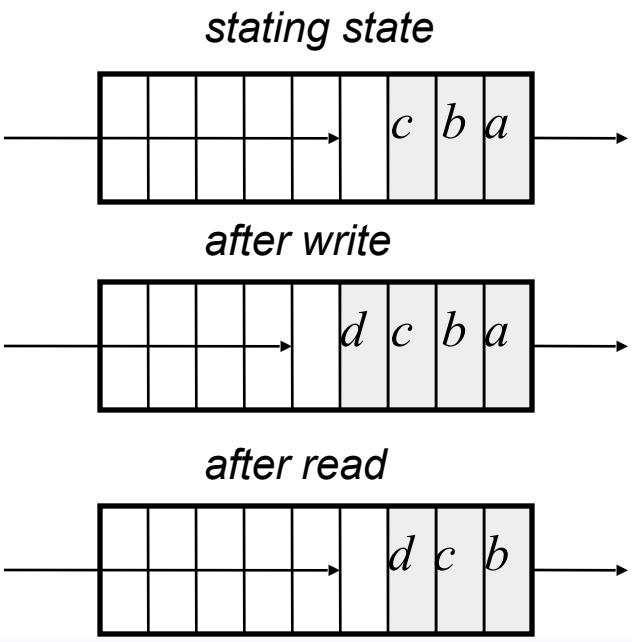
Example: given 1Kx8 SDP, want 1 read & 2 write ports.



FIFOs

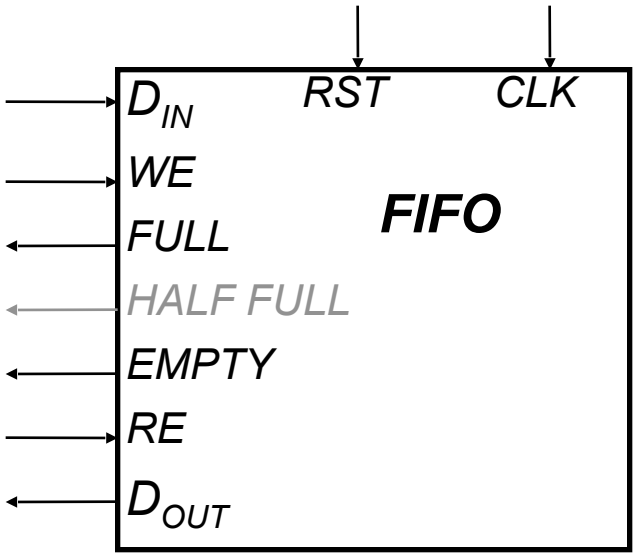
First-in-first-out (FIFO) Memory

- Used to implement *queues*.
- These find common use in computers and communication circuits.
- Generally, used to "decouple" actions of producer and consumer:



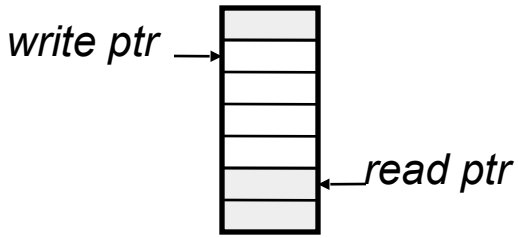
- *Producer can perform many writes without consumer performing any reads (or vis versa). However, because of finite buffer size, on average, need equal number of reads and writes.*
- *Typical uses:*
  - *interfacing I/O devices. Example network interface. Data bursts from network, then processor bursts to memory buffer (or reads one word at a time from interface). Operations not synchronized.*
  - *Example: Audio output. Processor produces output samples in bursts (during process swap-in time). Audio DAC clocks it out at constant sample rate.*

FIFO Interfaces

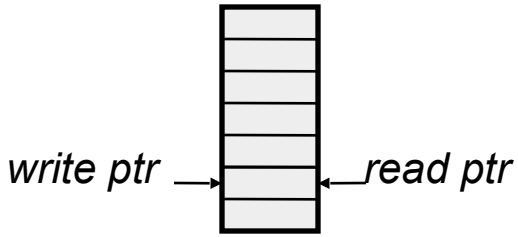


- After write or read operation, FULL and EMPTY indicate status of buffer.
- Used by external logic to control own reading from or writing to the buffer.
- FIFO resets to EMPTY state.
- HALF FULL (or other indicator of partial fullness) is optional.

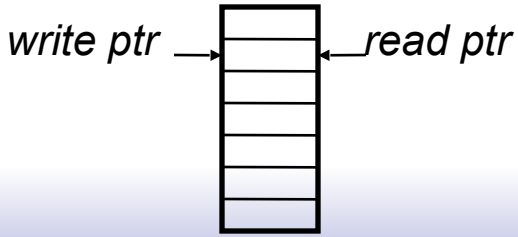
- *Address pointers are used internally to keep next write position and next read position into a dual-port memory.*



- *If pointers equal after write ⇒ FULL:*



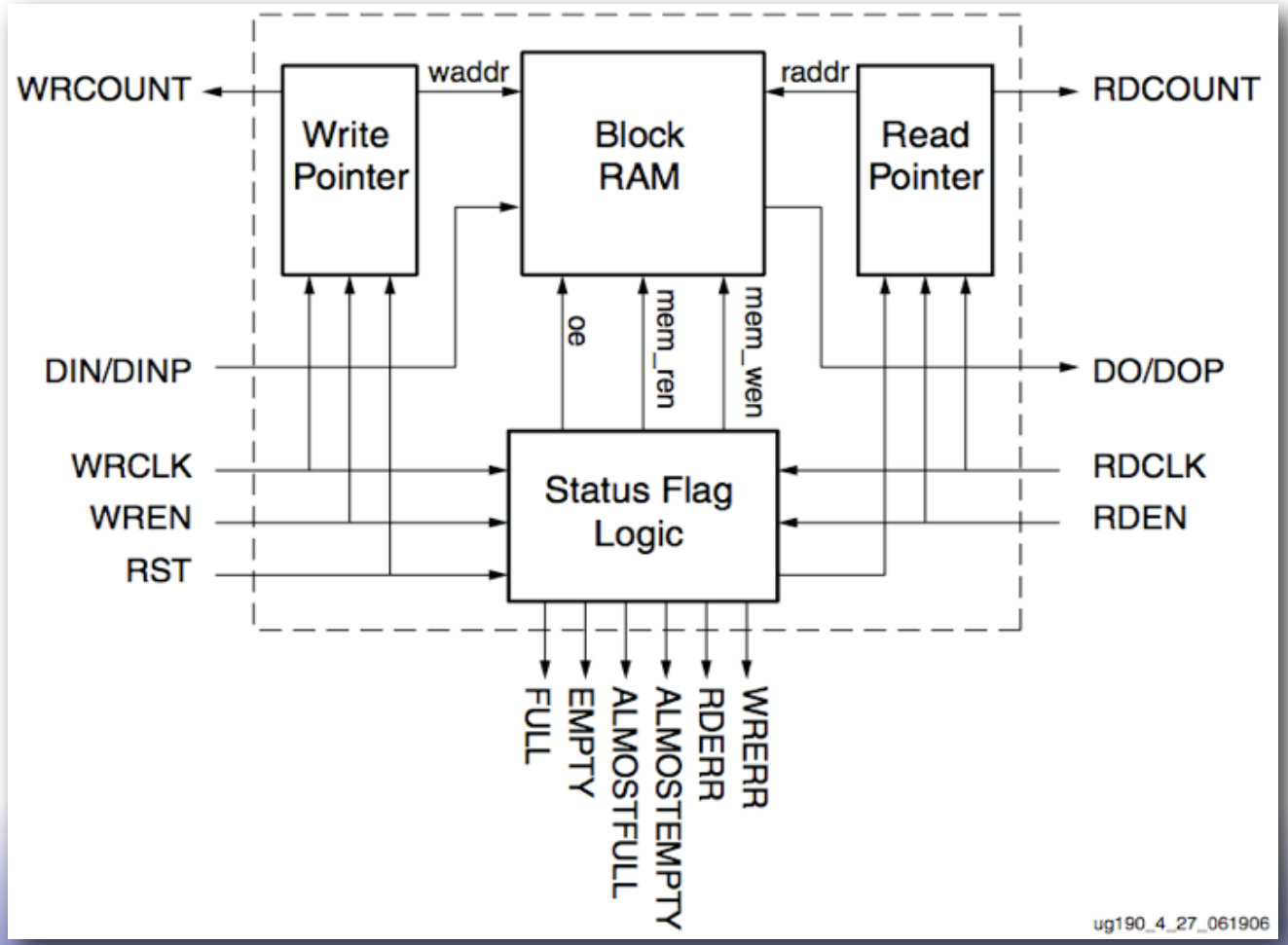
- *If pointers equal after read ⇒ EMPTY:*



Note: pointer incrementing is done "mod size-of-buffer"

Xilinx Virtex5 FIFOs

- Virtex5 BlockRAMS include dedicated circuits for FIFOs.
- Details in User Guide (ug190).
- Takes advantage of separate dual ports and independent ports clocks.

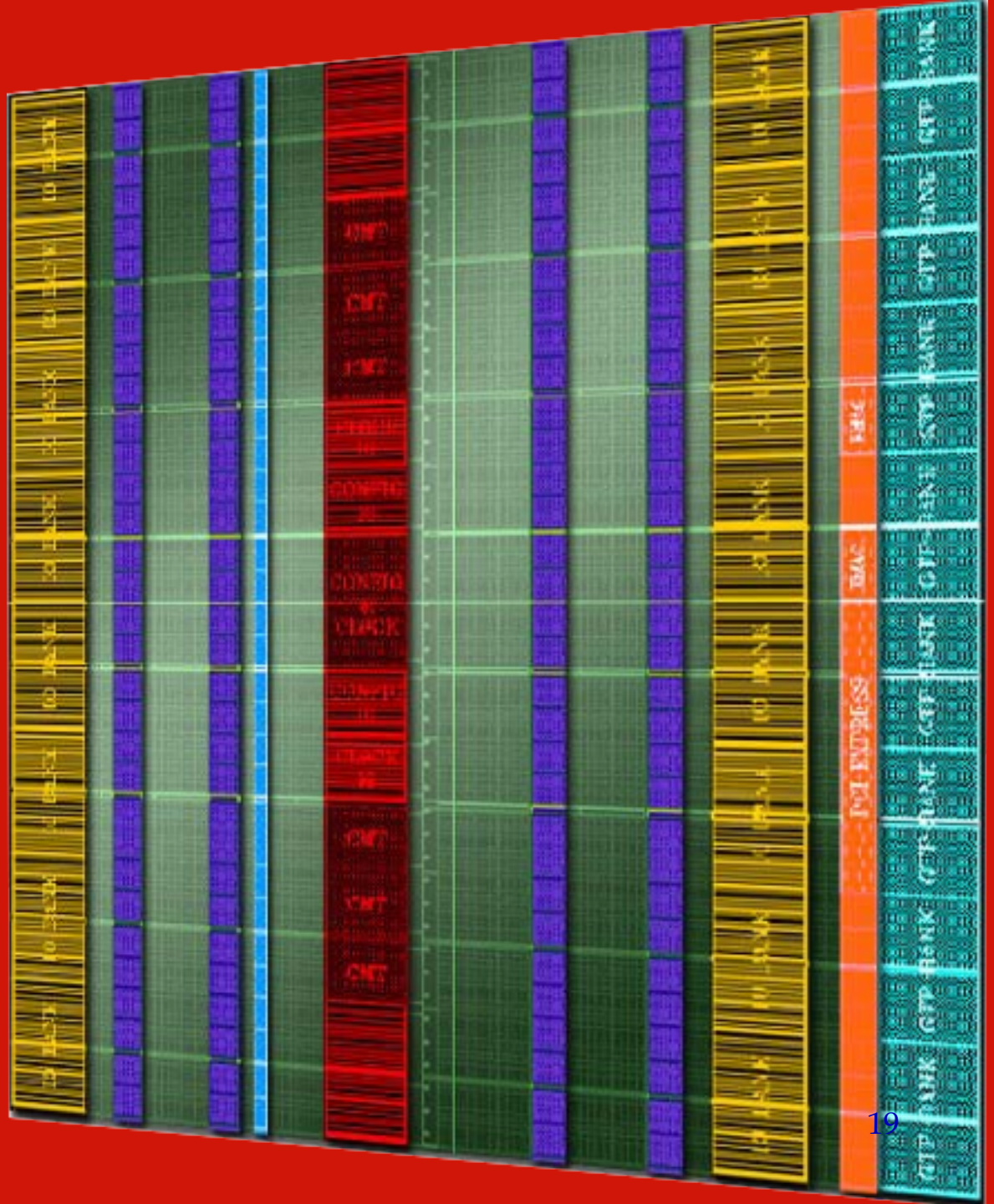


Memory on FPGAs

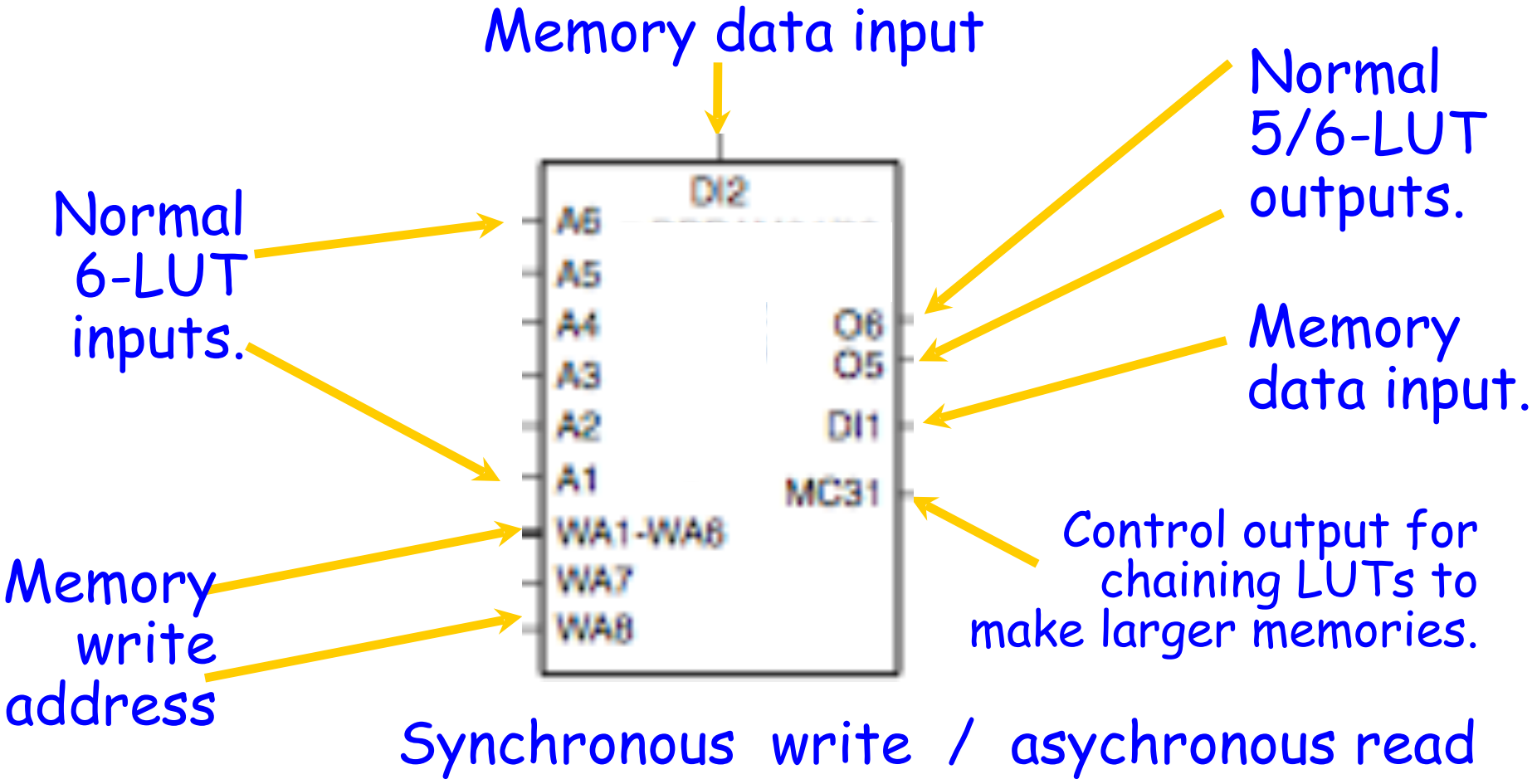
Virtex-5 LX110T memory blocks.

Distributed RAM using LUTs among the CLBs.

Block RAMs in four columns.



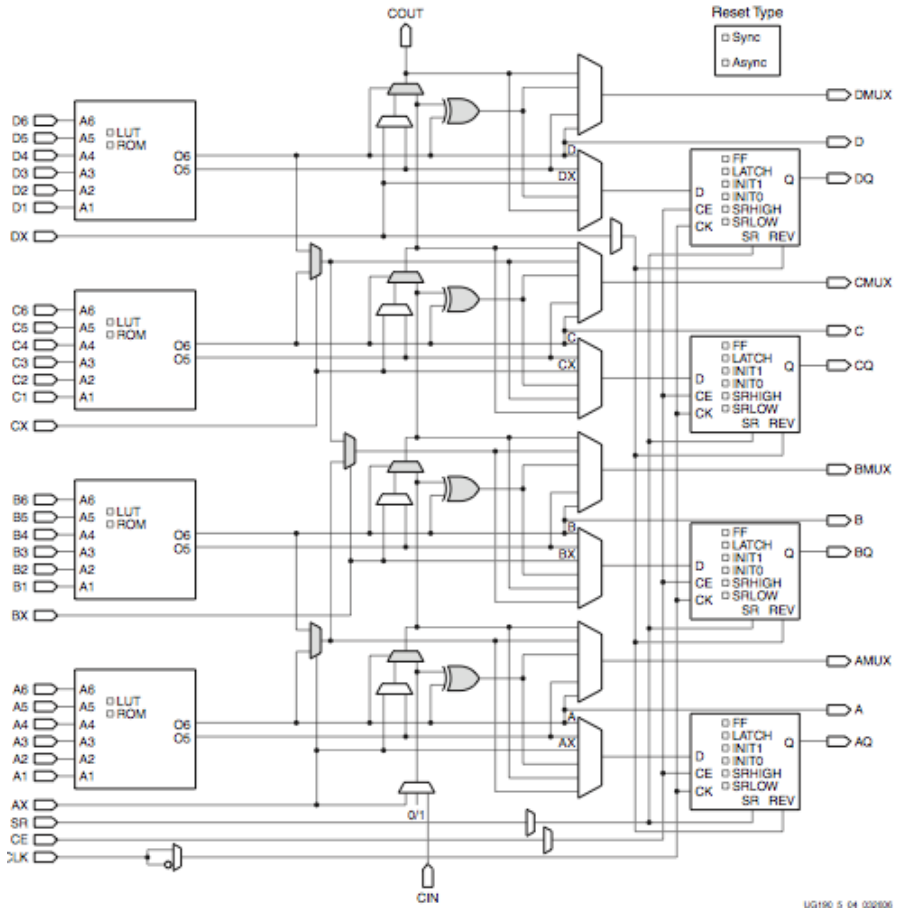
A SLICEM 6-LUT ...



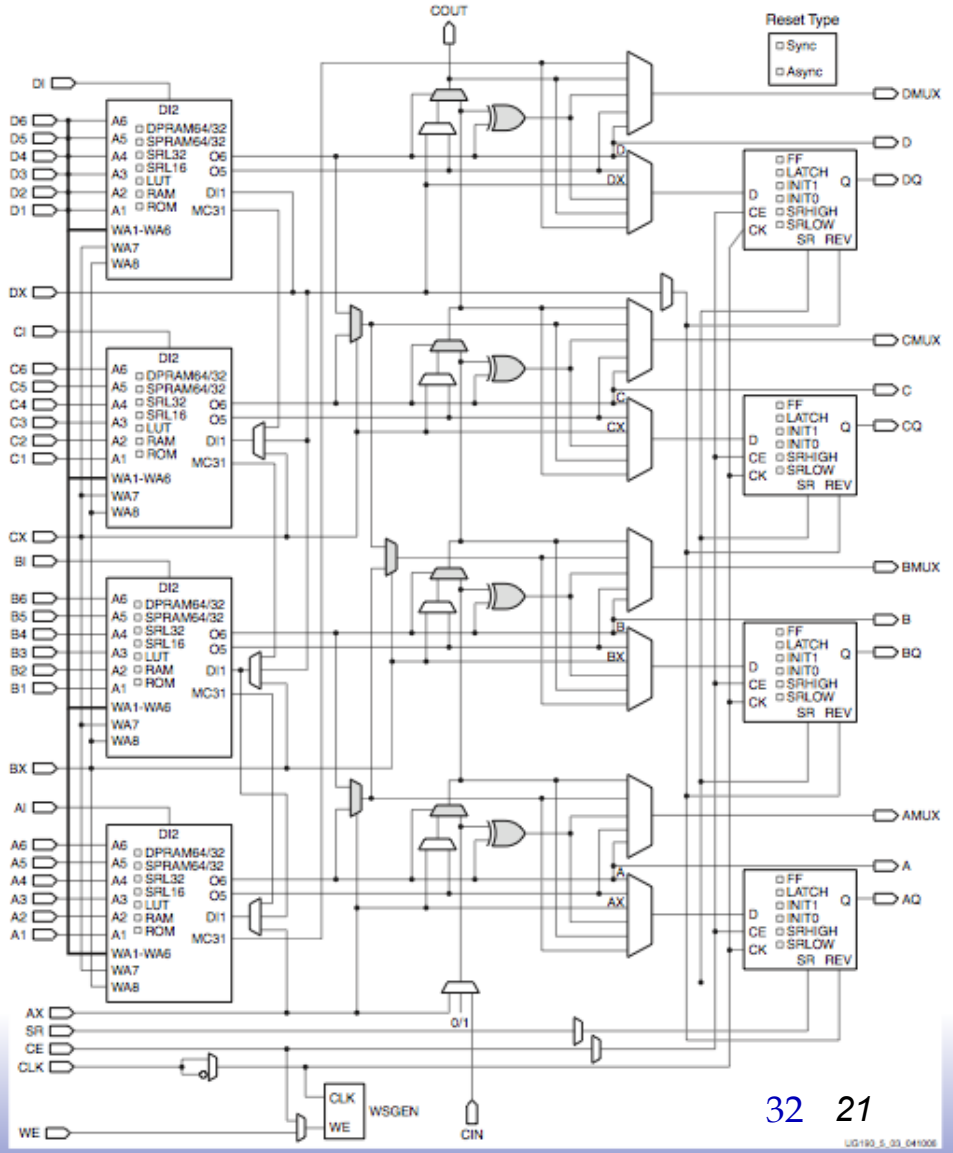
A 1.1 Mb distributed RAM can be made if all SLICEMs of an LX110T are used as RAM.

SLICEL vs SLICEM ...

SLICEL



SLICEM



SLICEM adds memory features to LUTs, + muxes.

Example Distributed RAM (LUT RAM)

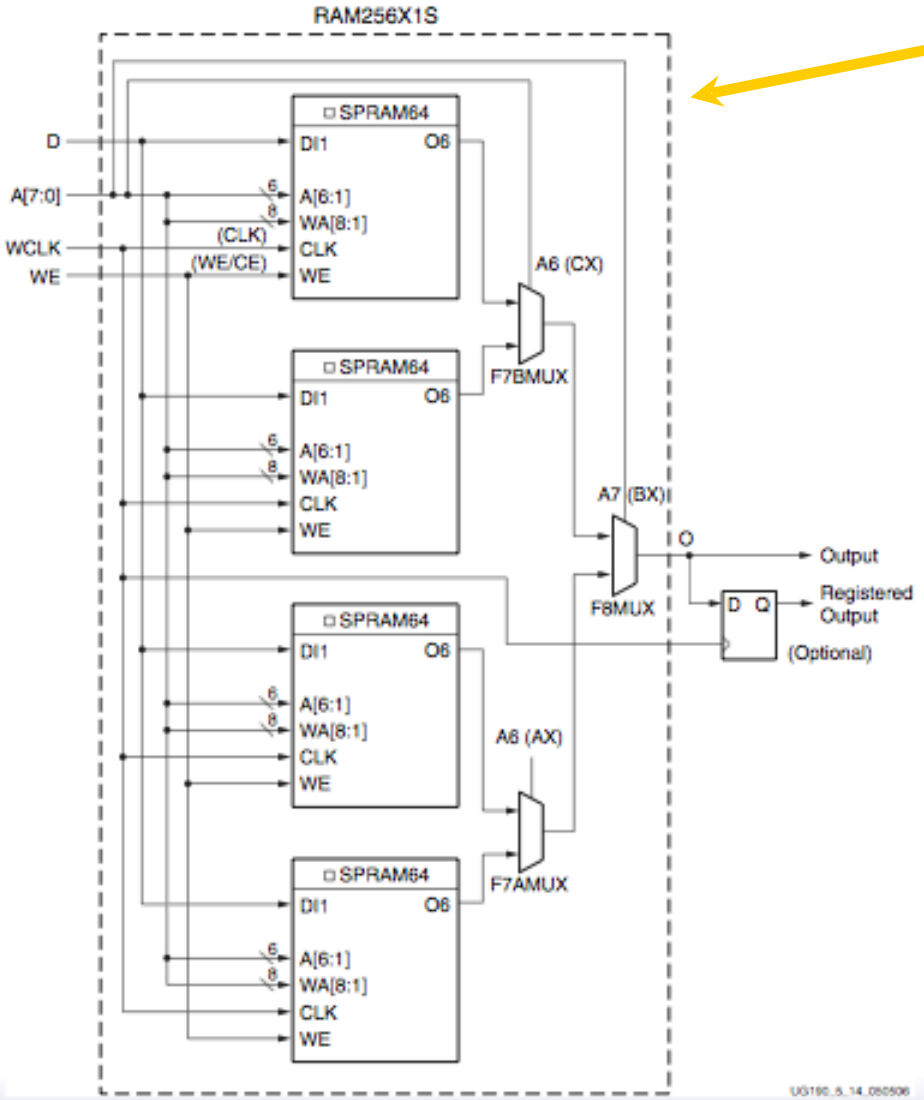
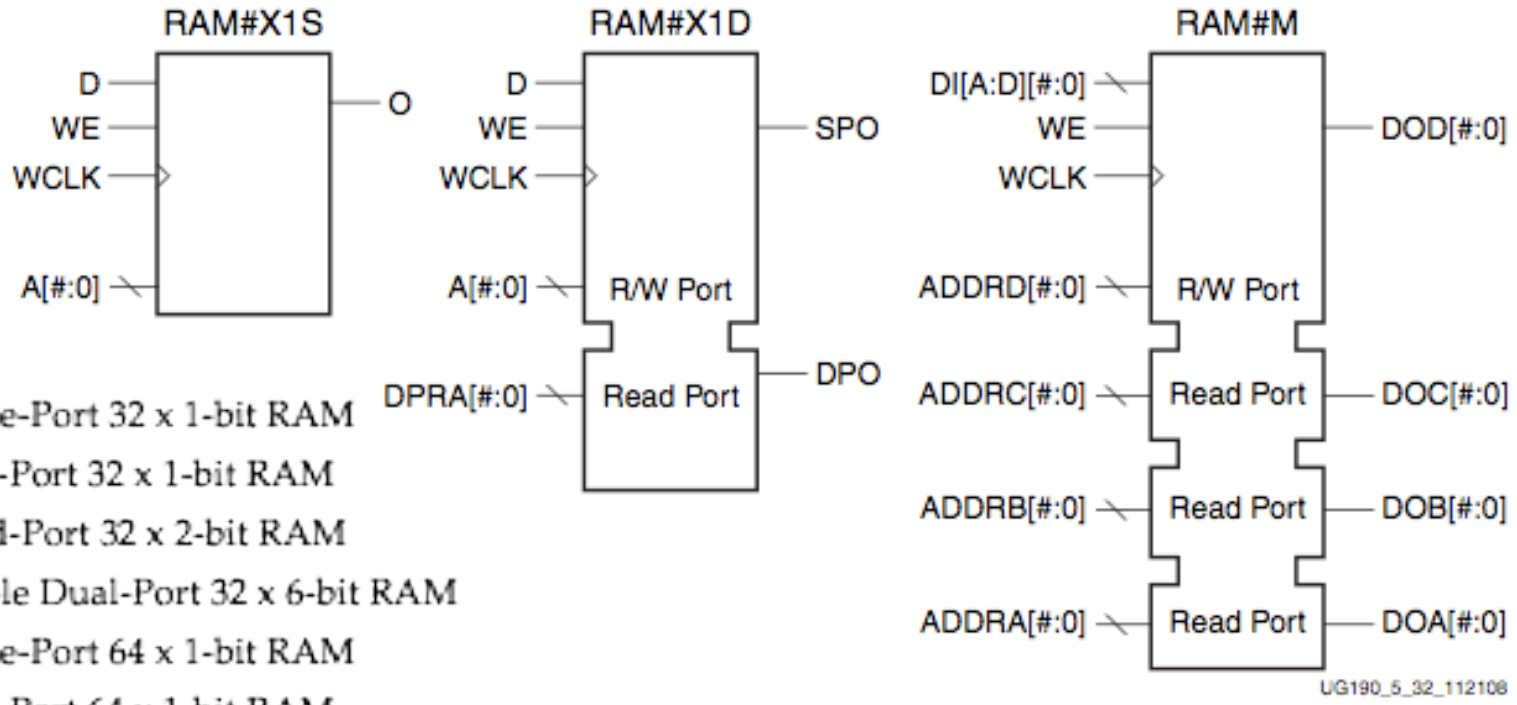


Figure 5-14: **Distributed RAM (RAM256X1S)**

Example configuration:
Single-port 256b x 1,
registered output.

Distributed RAM Primitives

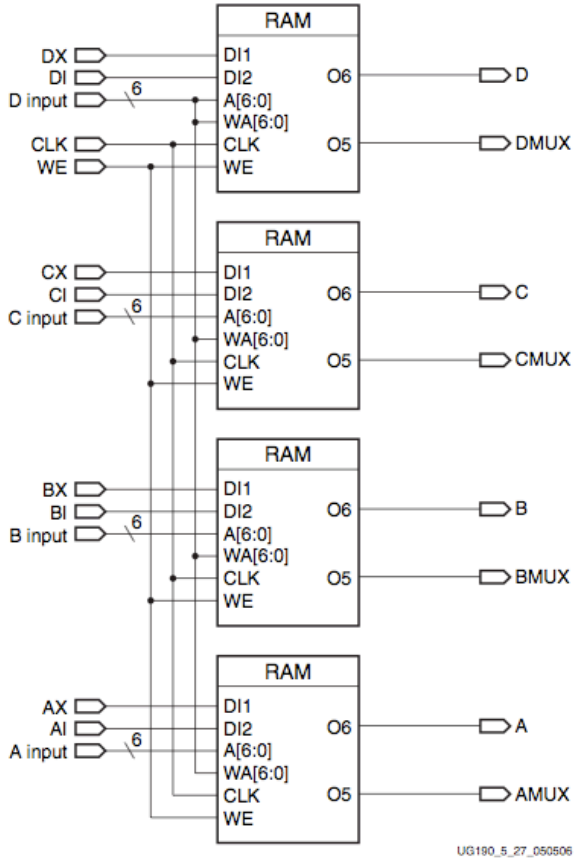


- Single-Port 32 x 1-bit RAM
- Dual-Port 32 x 1-bit RAM
- Quad-Port 32 x 2-bit RAM
- Simple Dual-Port 32 x 6-bit RAM
- Single-Port 64 x 1-bit RAM
- Dual-Port 64 x 1-bit RAM
- Quad-Port 64 x 1-bit RAM
- Simple Dual-Port 64 x 3-bit RAM
- Single-Port 128 x 1-bit RAM
- Dual-Port 128 x 1-bit RAM
- Single-Port 256 x 1-bit RAM

All are built from a single slice or less.

Remember, though, that the SLICEM LUT is naturally only 1 read and 1 write port.

Distributed RAM Timing



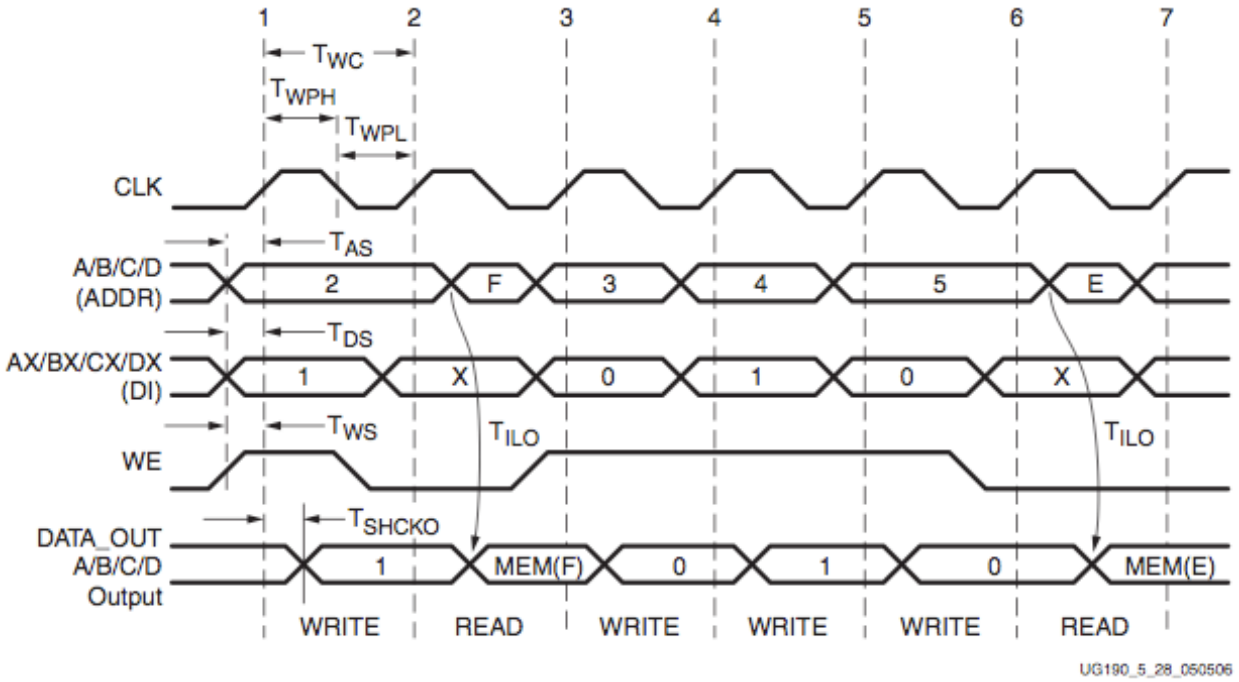
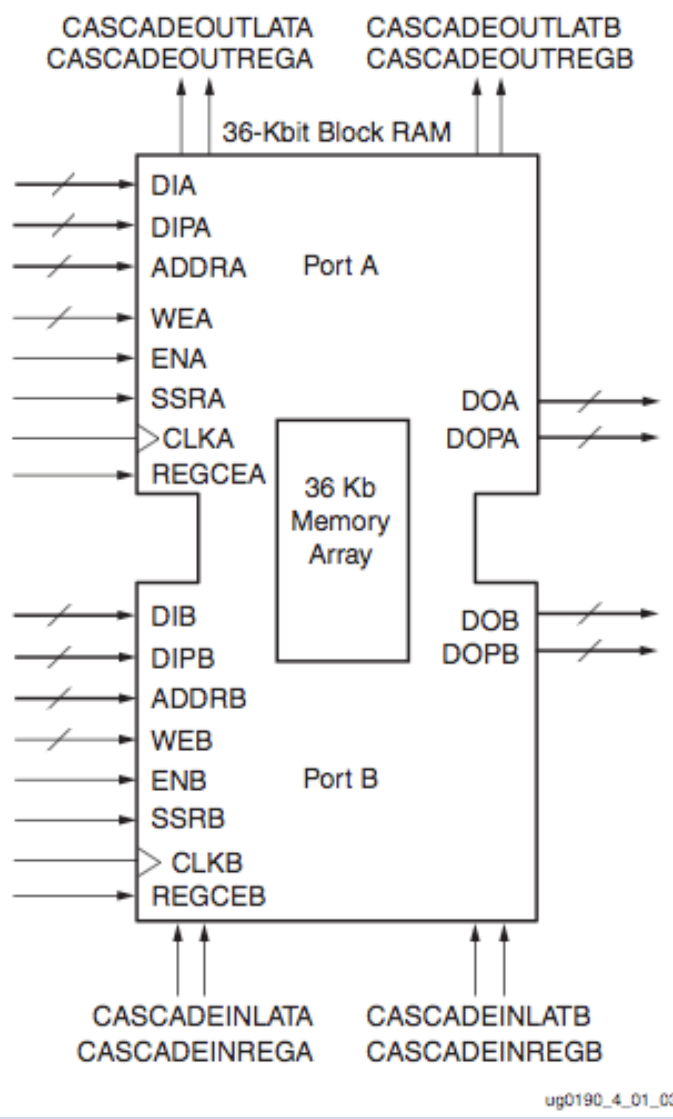


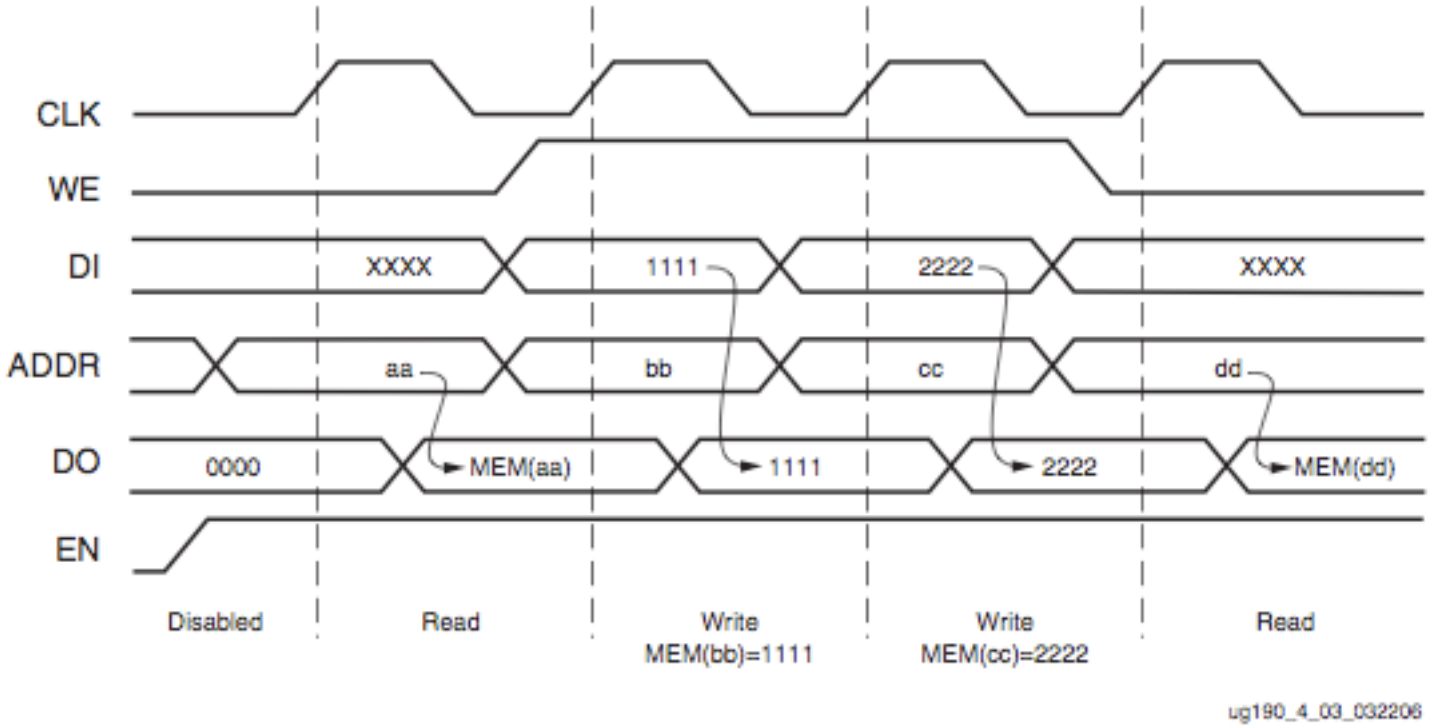
Figure 5-27: **Simplified Virtex-5 FPGA SLICEM Distributed RAM**

Block RAM Overview



- 36K bits of data total, can be configured as:
  - 2 independent 18Kb RAMs, or one 36Kb RAM.
- Each 36Kb block RAM can be configured as:
  - 64Kx1 (when cascaded with an adjacent 36Kb block RAM), 32Kx1, 16Kx2, 8Kx4, 4Kx9, 2Kx18, or 1Kx36 memory.
- Each 18Kb block RAM can be configured as:
  - 16Kx1, 8Kx2, 4Kx4, 2Kx9, or 1Kx18 memory.
- Write and Read are synchronous operations.
- The two ports are symmetrical and totally independent (can have different clocks), sharing only the stored data.
- Each port can be configured in one of the available widths, independent of the other port. The read port width can be different from the write port width for each port.
- The memory content can be initialized or cleared by the configuration bitstream.

Block RAM Timing



- Optional output register, would delay appearance of output data by one cycle.
- Maximum clock rate, roughly 400MHz.

State-of-the-Art - Xilinx FPGAs

Virtex Ultra-scale

Device Name	VU3P	VU5P	VU7P	VU9P	VU11P	VU13P	VU27P	VU29P	VU31P	VU33P	VU35P	VU37P
System Logic Cells (K)	862	1,314	1,724	2,586	2,835	3,780	2,835	3,780	962	962	1,907	2,852
CLB Flip-Flops (K)	788	1,201	1,576	2,364	2,592	3,456	2,592	3,456	879	879	1,743	2,607
CLB LUTs (K)	394	601	788	1,182	1,296	1,728	1,296	1,728	440	440	872	1,304
Max. Dist. RAM (Mb)	12.0	18.3	24.1	36.1	36.2	48.3	36.2	48.3	12.5	12.5	24.6	36.7
Total Block RAM (Mb)	25.3	36.0	50.6	75.9	70.9	94.5	70.9	94.5	23.6	23.6	47.3	70.9
UltraRAM (Mb)	90.0	132.2	180.0	270.0	270.0	360.0	270.0	360.0	90.0	90.0	180.0	270.0
HBM DRAM (GB)	–	–	–	–	–	–	–	–	4	8	8	8
HBM AXI Interfaces	–	–	–	–	–	–	–	–	32	32	32	32
Clock Mgmt Tiles (CMTs)	10	20	20	30	12	16	16	16	4	4	8	12
DSP Slices	2,280	3,474	4,560	6,840	9,216	12,288	9,216	12,288	2,880	2,880	5,952	9,024
Peak INT8 DSP (TOP/s)	7.1	10.8	14.2	21.3	28.7	38.3	28.7	38.3	8.9	8.9	18.6	28.1
PCIe® Gen3 x16	2	4	4	6	3	4	1	1	0	0	1	2
PCIe Gen3 x16/Gen4 x8 / CCIX[1]	–	–	–	–	–	–	–	–	4	4	4	4
150G Interlaken	3	4	6	9	6	8	6	8	0	0	2	4
100G Ethernet w/ KR4 RS-FEC	3	4	6	9	9	12	11	15	2	2	5	8
Max. Single-Ended HP I/Os	520	832	832	832	624	832	520	676	208	208	416	624
GTY 32.75Gb/s Transceivers	40	80	80	120	96	128	32	32	32	32	64	96
GTM 58Gb/s PAM4 Transceivers							32	48				
100G / 50G KP4 FEC							16 / 32	24 / 48				
Extended[2]	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3
Industrial	-1 -2	-1 -2	-1 -2	-1 -2	-1 -2	-1 -2	-1 -2	-1 -2	–	–	–	–

Ultra-RAM Blocks

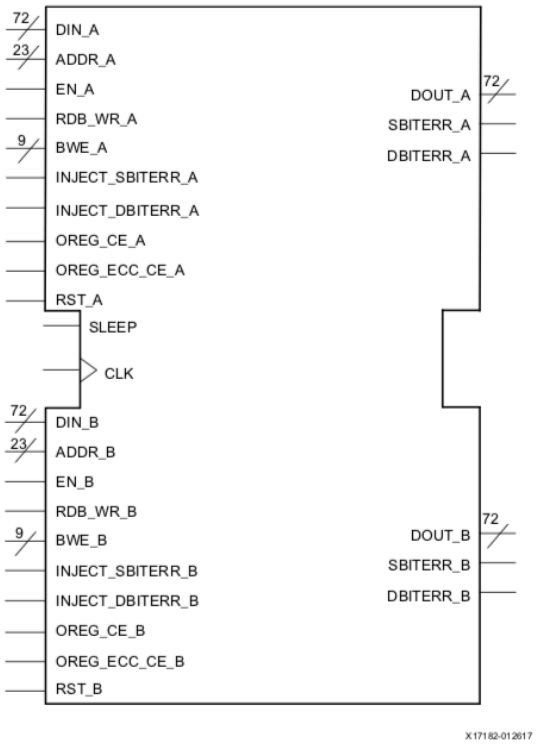


Figure 2-1: **UltraRAM URAM288_BASE Primitive**

Table 2-1: **Block RAM and UltraRAM Comparison**

Feature	Block RAM	UltraRAM
Clocking	Two clocks	Single clock
Built-in FIFO	Yes	No
Data width	Configurable (1, 2, 4, 9, 18, 36, 72)	Fixed (72-bits)
Modes	SDP and TDP	Two ports, each can independently read or write (a superset of SDP)
ECC	64-bit SECDED Supported in 64-bit SDP only (one ECC decoder for port A and one ECC encoder for port B)	64-bit SECDED One set of complete ECC logic for each port to enable independent ECC operations (ECC encoder and decoder for both ports)
Cascade	• Cascade output only (input cascade implemented via logic resources) • Cascade within a single clock region	• Cascade both input and output (with global address decoding) • Cascade across clock regions in a column • Cascade across several columns with minimal logic resources
Power savings	One mode via manual signal assertion	One mode via manual signal assertion

Memory Synthesis

Verilog RAM Specification

```
//
// Single-Port RAM with Asynchronous Read
//
module ramBlock (clk, we, a, di, do);
    input  clk;
    input  we;            // write enable
    input  [5:0] a;       // address
    input  [7:0] di;     // data in
    output [7:0] do;     // data out
    reg    [7:0] ram [1048575:0];  // 8x1Meg
    always @(posedge clk) begin    // Synch write
        if (we)
            ram[a] <= di;
    assign do = ram[a];            // Asynch read
endmodule
```

What do the synthesis tools do with this?

Verilog Synthesis Notes (FPGAs)

- Block RAMS and LUT RAMS all exist as primitive library elements. However, it is much more convenient to **use inference**.
- Depending on how you write your Verilog, you will get either a collection of block RAMs, a collection of LUT RAMs, or a collection of flip-flops.
- The synthesizer uses size, and read style (synch versus asynch) to determine the best primitive type to use.
- It is possible to force mapping to a particular primitive by using synthesis directives. Ex: (* ram_style = "distributed" *) reg myReg;
- The synthesizer has limited capabilities (eg., it can combine primitives for more depth and width, but is limited on porting options). Be careful, as you might not get what you want.
- See **XST User Guide** for examples.
- CORE generator memory block has an extensive set of parameters for explicitly instantiated RAM blocks.

Processor Design Considerations (FPGA Version)

- **Register File: Consider distributed RAM (LUT RAM)**
  - Size is close to what is needed: distributed RAM primitive configurations are 32 or 64 bits deep. Extra width is easily achieved by parallel arrangements.
  - LUT-RAM configurations offer multi-porting options - useful for register files.
  - Asynchronous read, might be useful by providing flexibility on where to put register read in the pipeline.
- **Instruction / Data Caches : Consider Block RAM**
  - Higher density, lower cost for large number of bits
  - A single 36kbit Block RAM implements 1K 32-bit words.
  - Configuration stream based initialization, permits a simple “boot strap” procedure.

Verilog Synthesis Notes (ASICs)

- Depending on how you write your Verilog, you will get either a collection of flip-flops or latches.
- Dense RAM arrays are *not* inferred and must be explicitly instantiated.
- Fab house supplied design kits and cell libraries typically come with parameterized RAM block generators (or, at least, a set of predesigned blocks).

Artisan ARM Physical IP

sram_sp_hdc_svt_rvt_hvt Compiler
TSMC cln65gplus 65nm Process, 0.525um2 Bitcell

GENERIC PARAMETERS

Instance Name	sram_sp_hdc_svt_rvt_hvt
Number of Words	2048
Number of Bits	16
Frequency <MHz>	1
Multiplexer Width	☐ 4 ☐ 8 ☑ 16 ☐ 32
Pipeline	☐ on ☑ off
Word-Write Mask	☐ on ☑ off
Write-thru	☑ on ☐ off
Top Metal Layer	☑ m5-m9
Power Type	☑ ArtiGrid
Redundancy	☐ on ☑ off
BIST MUXes	☐ on ☑ off
Soft Error Repair (SER)	☑ none ☐ 1bd1bc ☐ 2bd1bc
Back Biasing	☐ on ☑ off
Power Gating	☐ on ☑ off
Retention	☑ on

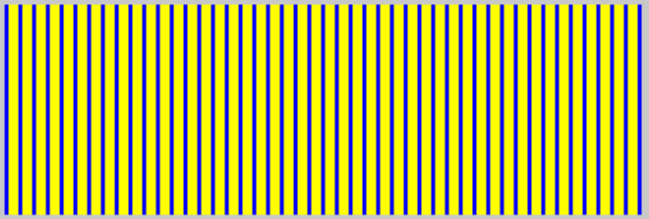
Default Update

VIEWS

Verilog Model

- PostScript Datasheet
- ASCII Datatable
- Verilog Model
- Synopsys Model
- VCLEF Footprint
- GDSII Layout
- LVS Netlist
- TetraMax Model

RELATIVE FOOTPRINT



ASCII DATATABLE

name	tt_1p00v_...	ss_0p90v_...	ss_0p90v_...	ff_
geomx	317.790	317.790	317.790	31
geomy	111.350	111.350	111.350	11
volt	1.000	0.900	0.900	1.
temp	25.000	-40.000	125.000	12
tcyc0	0.984	1.658	1.581	0.
tcyc1	1.042	1.761	1.684	0.
tcyc2	1.124	1.941	1.826	0.
tcyc3	1.150	1.990	1.874	0.
tcyc4	999.000	999.000	999.000	99
tcyc5	999.000	999.000	999.000	99
tcyc6	999.000	999.000	999.000	99
tcyc7	999.000	999.000	999.000	99
ta0	0.736	1.213	1.142	0.
ta1	0.794	1.316	1.244	0.
ta2	0.876	1.496	1.387	0.
ta3	0.901	1.545	1.435	0.
ta4	999.000	999.000	999.000	99
ta5	999.000	999.000	999.000	99
ta6	999.000	999.000	999.000	99
ta7	999.000	999.000	999.000	99
tas	0.226	0.389	0.356	0.
tah	0.133	0.206	0.213	0.
tcs	0.152	0.221	0.230	0.
tch	0.121	0.209	0.194	0.
tws	0.121	0.235	0.224	0.
twh	0.160	0.262	0.264	0.

Processor Design Considerations (ASIC Version)

- **Register File: use synthesized RAM**
  - At this size (1k bits) synthesized is competitive with dense RAM block
  - Latch-based instead of flip-flop-based will save on area.
  - Asynchronous read, might be useful by providing flexibility on where to put register read in the pipeline.
- **Instruction / Data Caches : Use generated dense Block RAM**
  - Higher density, lower cost for large number of bits
  - We will provide for you

Inferring RAMs in Verilog (FPGA)

```
// 64X1 RAM implementation using distributed RAM

module ram64X1 (clk, we, d, addr, q);
input clk, we, d;
input [5:0] addr;
output q;

   reg [63:0] temp;
   always @ (posedge clk)
     if(we)
       temp[addr] <= d;
   assign q = temp[addr];

   endmodule
```

Verilog reg array used with "always @ (posedge ... infers memory array.

Asynchronous read infers LUT RAM

Dual-read-port LUT RAM (FPGA)

```
//
// Multiple-Port RAM Descriptions
//
module v_rams_17 (clk, we, wa, ra1, ra2, di, do1, do2);
    input  clk;
    input  we;
    input  [5:0] wa;
    input  [5:0] ra1;
    input  [5:0] ra2;
    input  [15:0] di;
    output [15:0] do1;
    output [15:0] do2;
    reg    [15:0] ram [63:0];
    always @(posedge clk)
    begin
        if (we)
            ram[wa] <= di;
    end
    assign do1 = ram[ra1];
    assign do2 = ram[ra2];
endmodule
```

Multiple reference to same array.

Block RAM Inference (FPGA)

```
//
// Single-Port RAM with Synchronous Read
//
module v_rams_07 (clk, we, a, di, do);
    input  clk;
    input  we;
    input  [5:0] a;
    input  [15:0] di;
    output [15:0] do;
    reg    [15:0] ram [63:0];
    reg    [5:0] read_a;
    always @(posedge clk) begin
        if (we)
            ram[a] <= di;
        read_a <= a;
    end
    assign do = ram[read_a];
endmodule
```

Synchronous read (registered read address) infers Block RAM

FPGA Block RAM initialization (FPGA)

```
module RAMB4_S4 (data_out, ADDR, data_in, CLK, WE);
   output[3:0] data_out;
   input [2:0] ADDR;
   input [3:0] data_in;
   input CLK, WE;
   reg [3:0] mem [7:0];
   reg [3:0] read_addr;

   initial
     begin
       $readmemb("data.dat", mem);
     end

   always@(posedge CLK)
     read_addr <= ADDR;

   assign data_out = mem[read_addr];

   always @(posedge CLK)
     if (WE) mem[ADDR] = data_in;

   endmodule
```

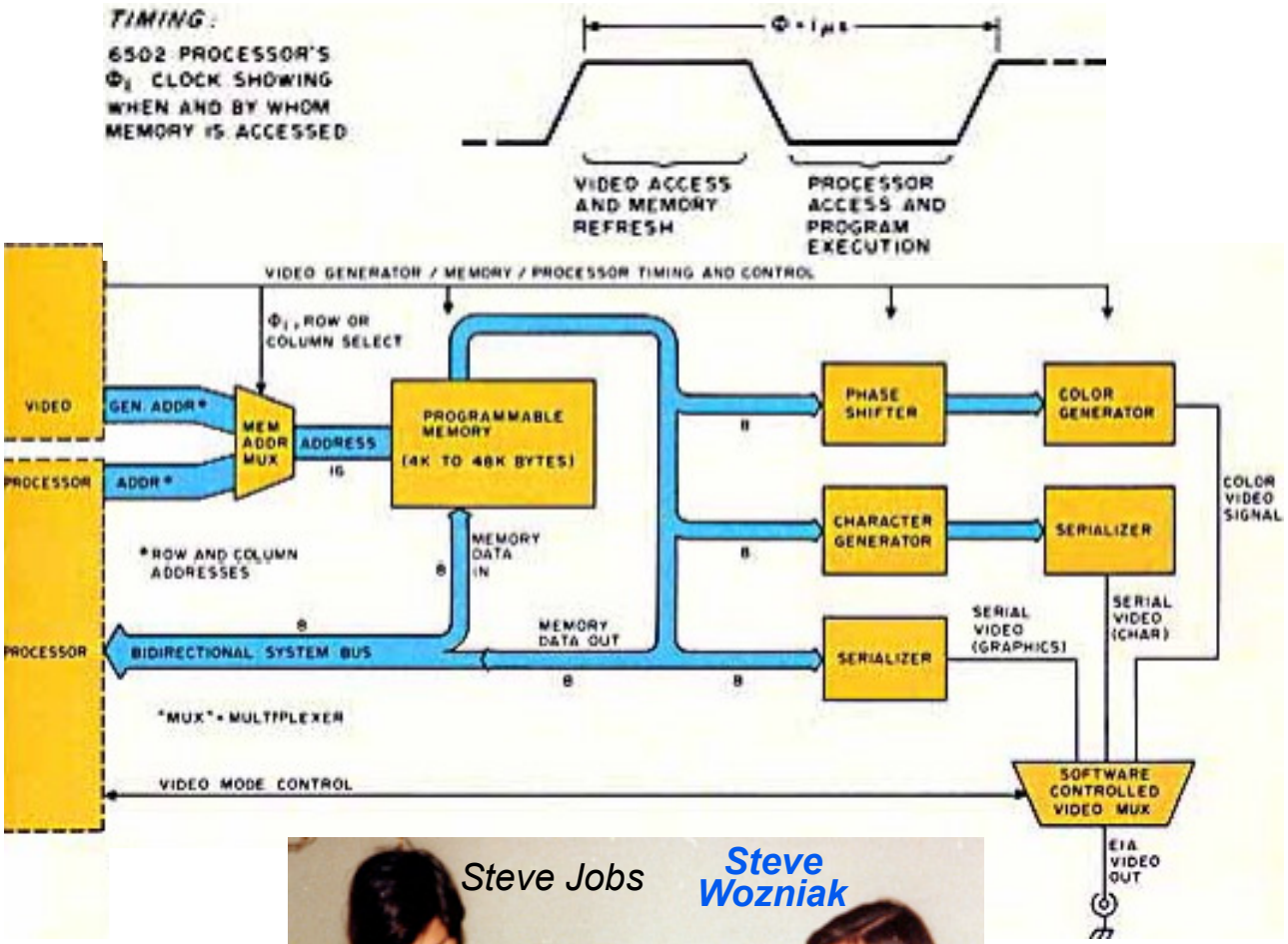
"data.dat" contains initial RAM contents, it gets put into the bitfile and loaded at configuration time. (Remake bits to change contents)

Dual-Port Block RAM (FPGA)

```
module test (data0,data1,waddr0,waddr1,we0,we1,clk0, clk1, q0, q1);

   parameter d_width = 8;  parameter addr_width = 8; parameter mem_depth = 256;

   input [d_width-1:0] data0, data1;
   input [addr_width-1:0] waddr0, waddr1;
   input we0, we1, clk0, clk1;

   reg [d_width-1:0] mem [mem_depth-1:0]
   reg [addr_width-1:0] reg_waddr0, reg_waddr1;
   output [d_width-1:0] q0, q1;

   assign q0 = mem[reg_waddr0];
   assign q1 = mem[reg_waddr1];

   always @(posedge clk0)
     begin
       if (we0)
         mem[waddr0] <= data0;
       reg_waddr0 <= waddr0;
     end

   always @(posedge clk1)
     begin
       if (we1)
         mem[waddr1] <= data1;
       reg_waddr1 <= waddr1;
     end

   endmodule
```

Caches

1977: DRAM faster than microprocessors

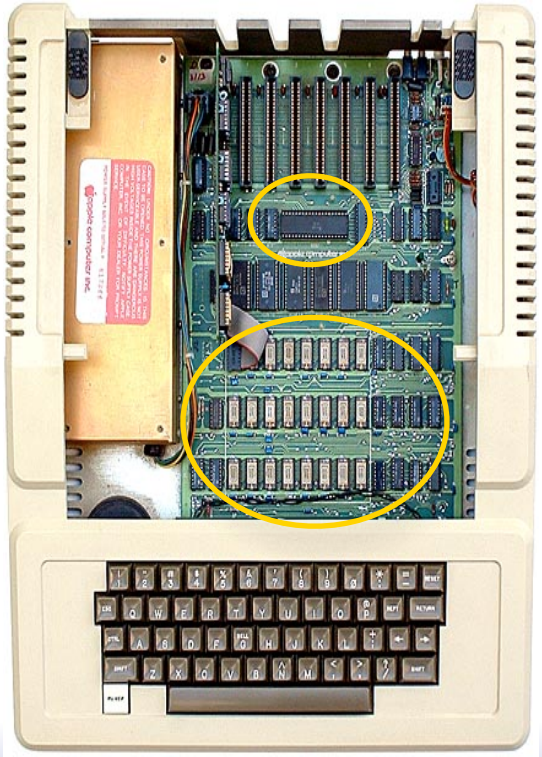


RAM Complement	Apple II System
4K	$ 1,298.00
48K	2,638.00

Apple II (1977)

CPU: 1000 ns

DRAM: 400 ns

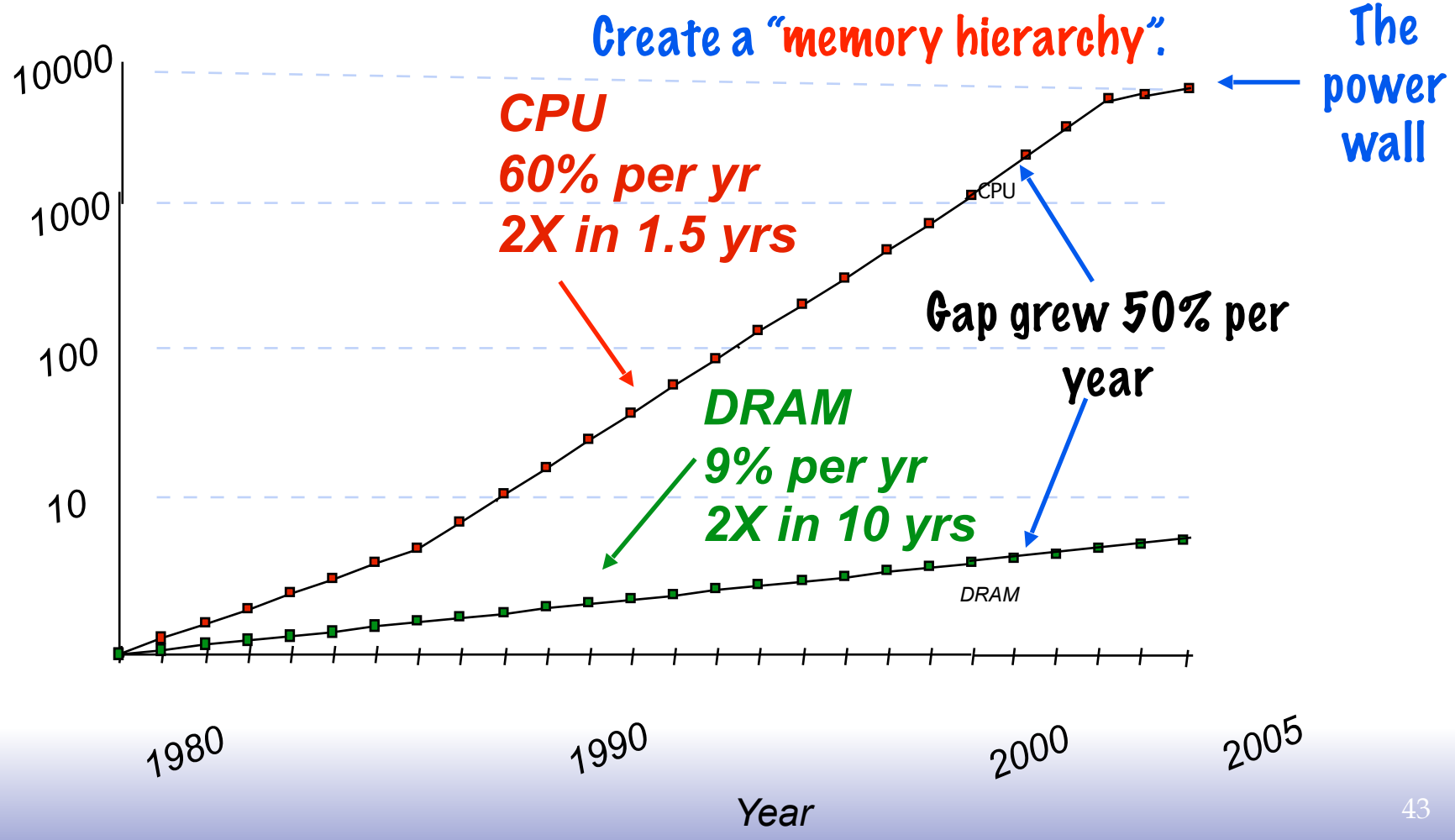


1980-2003, CPU speed outpaced DRAM ...

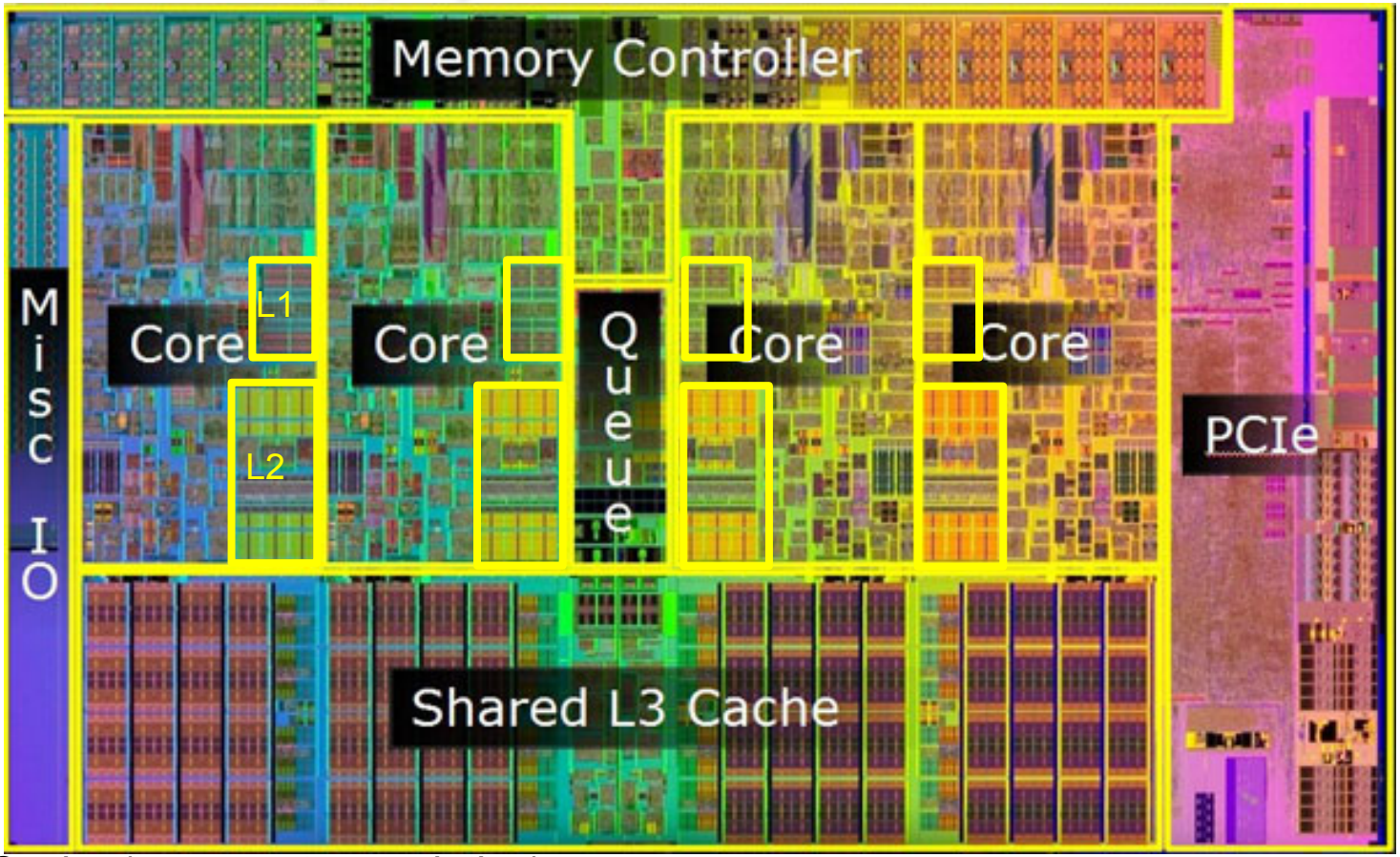
Q. How do architects address this gap?

A. Put smaller, faster "cache" memories between CPU and DRAM.

Create a "memory hierarchy".



Nahalem Die Photo (i7, i5)

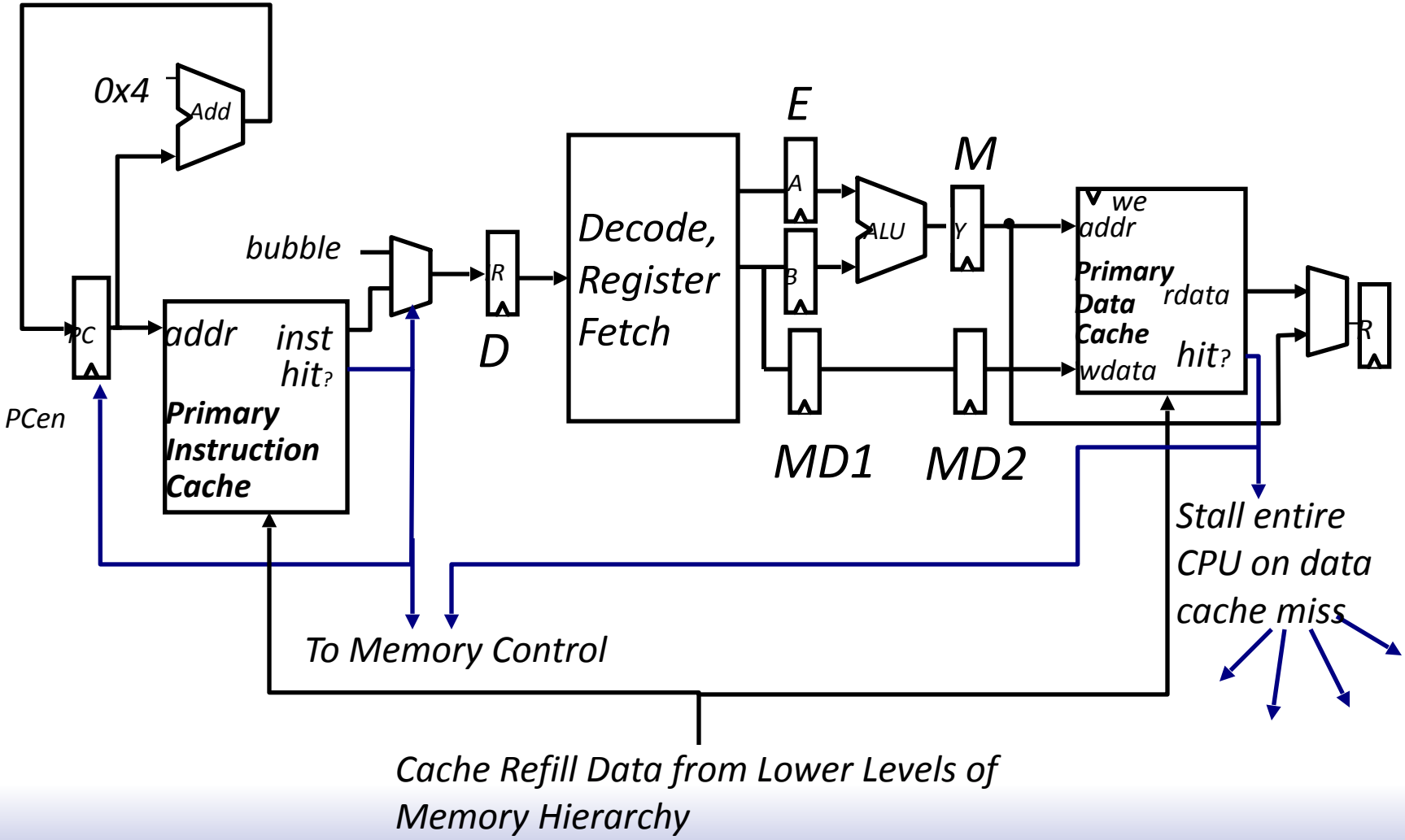


- Per core:
  - 32KB L1 I-Cache (4-way set associative)
  - 32KB L1 D-Cache (8-way set associative)
  - 256KB unified L2 (8-way SA, 64B blocks)
  - Common L3 8MB cache
- Common L3 8MB cache

Characteristic	Intel Nehalem	AMD Opteron X4 (Barcelona)
L1 cache organization	Split instruction and data caches	Split instruction and data caches
L1 cache size	32 KB each for instructions/data per core	64 KB each for instructions/data per core
L1 block size	64 bytes	64 bytes
L1 write policy	Write-back, Write-allocate	Write-back, Write-allocate
L1 hit time (load-use)	Not Available	3 clock cycles
L2 cache organization	Unified (instruction and data) per core	Unified (instruction and data) per core
L2 cache size	256 KB (0.25 MB)	512 KB (0.5 MB)
L2 block size	64 bytes	64 bytes
L2 write policy	Write-back, Write-allocate	Write-back, Write-allocate
L2 hit time	Not Available	9 clock cycles
L3 cache organization	Unified (instruction and data)	Unified (instruction and data)
L3 cache size	8192 KB (8 MB), shared	2048 KB (2 MB), shared
L3 block size	64 bytes	64 bytes
L3 write policy	Write-back, Write-allocate	Write-back, Write-allocate
L3 hit time	Not Available	38 (?)clock cycles

CPU-Cache Interaction

(5-stage pipeline)



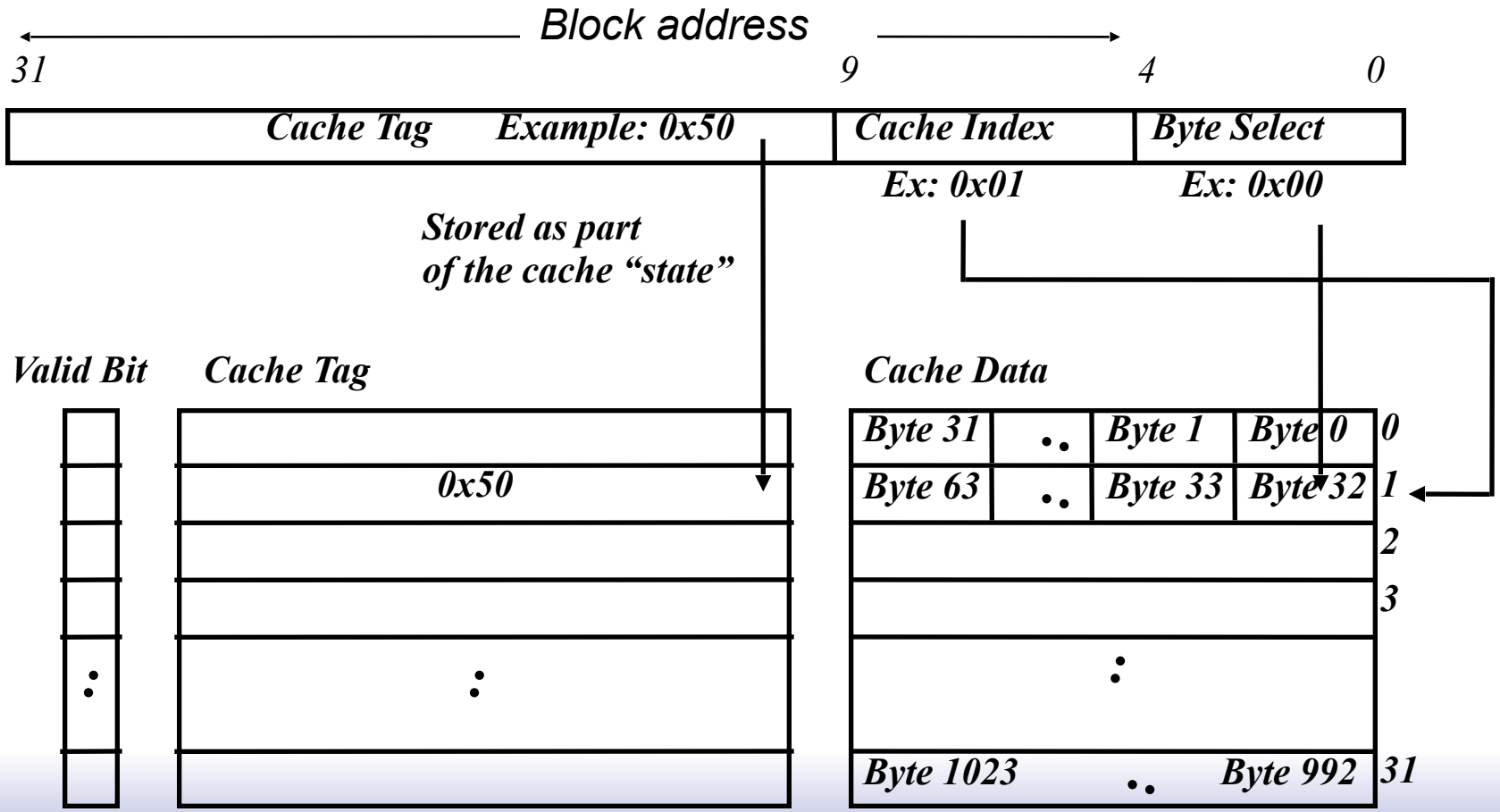
Review from 61C

- Two Different Types of Locality:
  - Temporal Locality (Locality in Time): If an item is referenced, it will tend to be referenced again soon.
  - Spatial Locality (Locality in Space): If an item is referenced, items whose addresses are close by tend to be referenced soon.
- By taking advantage of the principle of locality:
  - Present the user with as much memory as is available in the cheapest technology.
  - Provide access at the speed offered by the fastest technology.
- DRAM is slow but cheap and dense:
  - Good choice for presenting the user with a BIG memory system
- SRAM is fast but expensive and not very dense:
  - Good choice for providing the user FAST access time.

Example: 1 KB Direct Mapped Cache with 32 B Blocks

For a 2^N byte cache:

- The uppermost (32 - N) bits are always the Cache Tag
- The lowest M bits are the Byte Select (Block Size = 2^M)

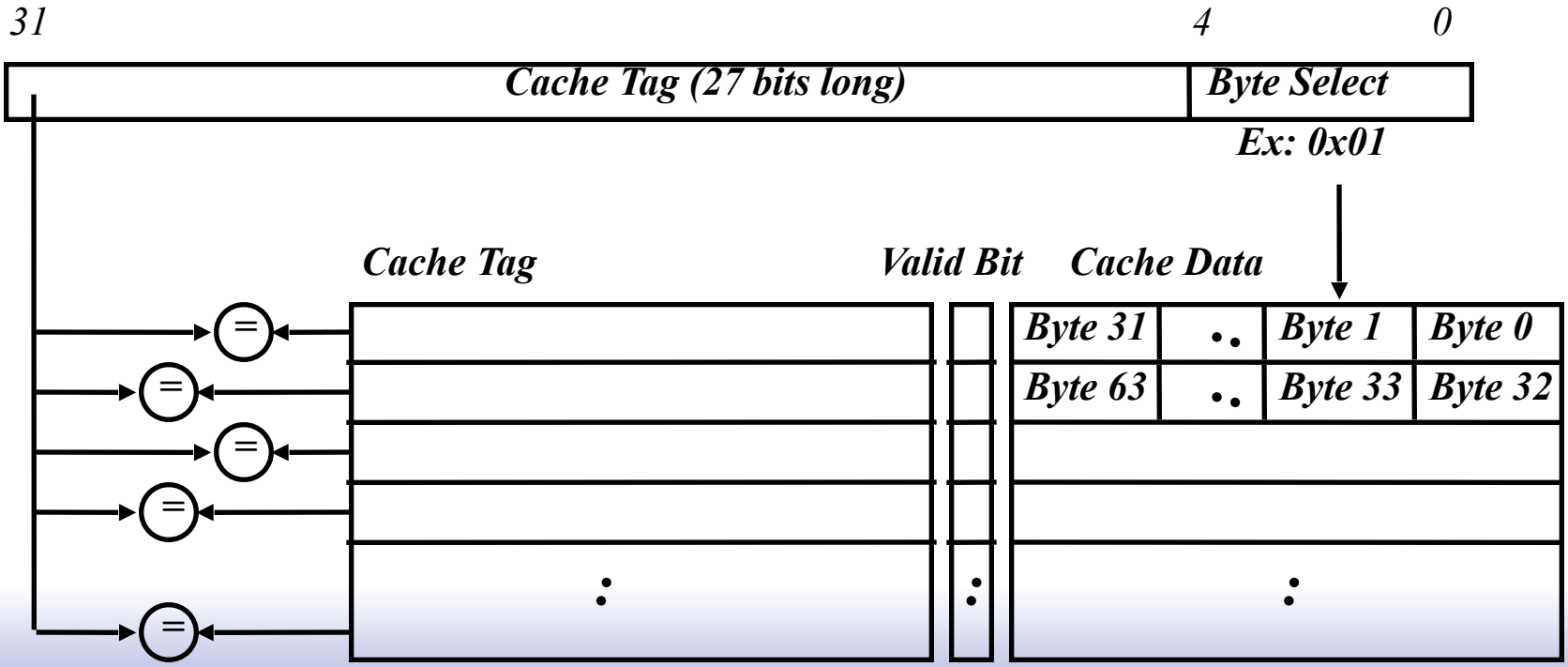


Extreme Example: Fully Associative

Fully Associative Cache

- Forget about the Cache Index
- Compare the Cache Tags of all cache entries in parallel
- Example: Block Size = 32 B blocks, we need N 27-bit comparators

By definition: Conflict Miss = 0 for a fully associative cache



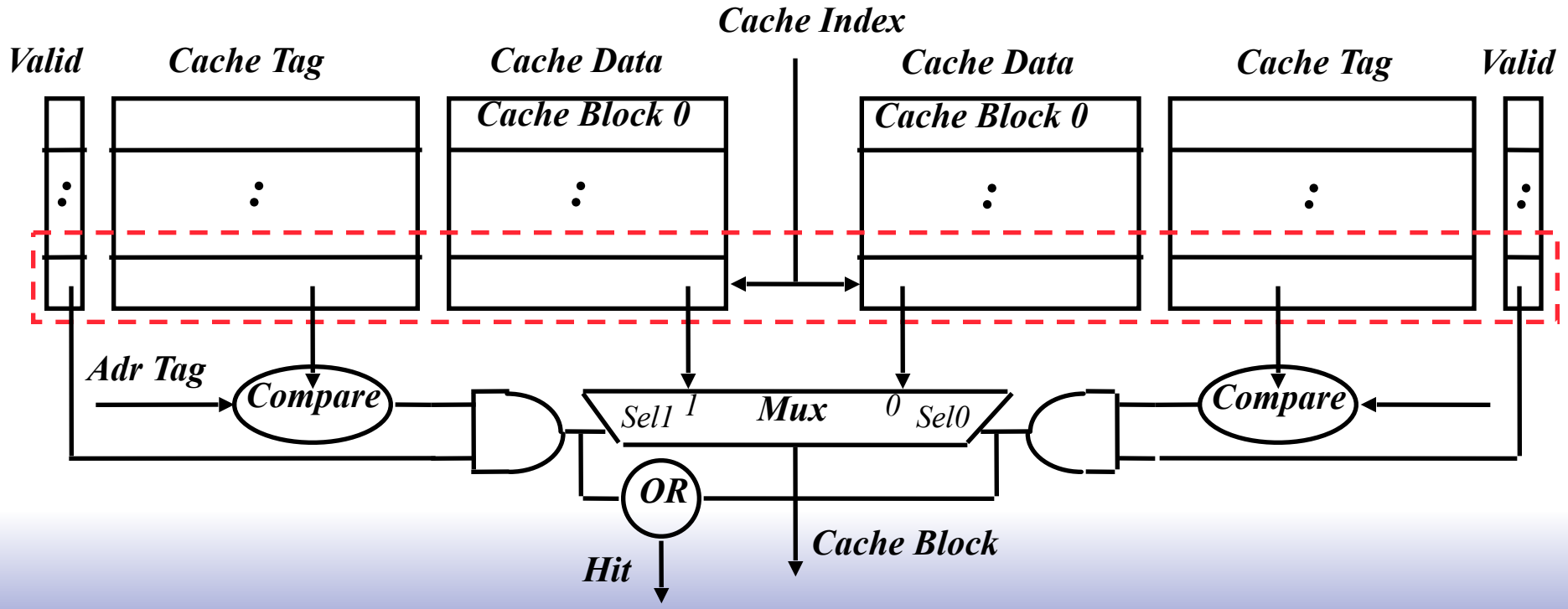
Set Associative Cache

N-way set associative: N entries for each Cache Index

- N direct mapped caches operates in parallel

Example: Two-way set associative cache

- Cache Index selects a "set" from the cache
- The two tags in the set are compared to the input in parallel
- Data is selected based on the tag result



Disadvantage of Set Associative Cache

N-way Set Associative Cache versus Direct Mapped Cache:

- N comparators vs. 1
- Extra MUX delay for the data
- Data comes AFTER Hit/Miss decision and set selection

In a direct mapped cache, Cache Block is available BEFORE Hit/Miss:

- Possible to assume a hit and continue. Recover later if miss.