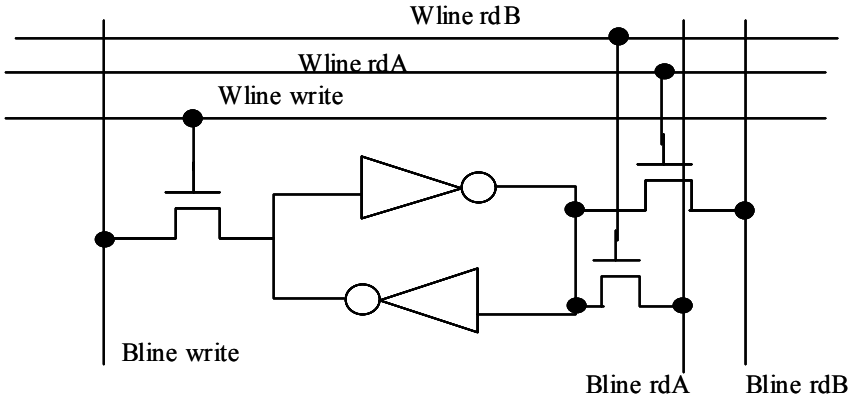
EE 8273 Test 2 - Fall '01 Solutions – Reese

1. (5 pts) I would like an SRAM register file that will support 1 write operation and 2 read operations, simultaneously. How many word lines will I need for the cell?. Draw a diagram of your design for the SRAM cell.



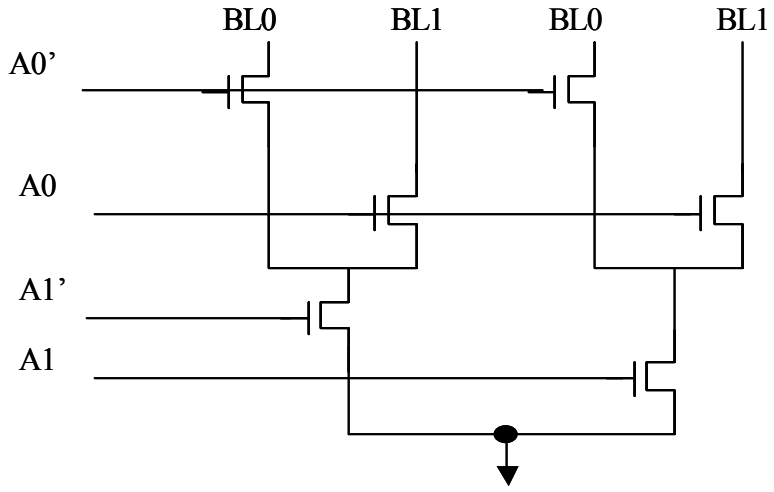
2. (15 pts) Answer the following questions about a 64K x 16 SRAM.
   a. (5 pts) How many memory cell loads would the word line driver be connected to if a single, square (# of rows = # of columns) memory array was used?

64K x 16 = 2^{16} x 2^4 = 2^{20} bits. Square = 2^{10} x 2^{10}, so 1024 rows, 1024 columns.

   b. (5 pts) Assume a hierarchical decoding scheme in which the bits are split into two identical 'planes', with the final word line loading limited to 64 RAM cells (64 columns) and 32 memory blocks in each 'plane'. How many rows would be in each sub-block?

2 planes = 2^1. Blocks = 2^5. Columns = 2^6. So $2^{20} / (2^1 x 2^5 x 2^6) = 2^8$ rows = 256 rows.

   c. (5 pts) For the 64 column block above, 4 bit lines are decoded to one Sense-Amp input using NMOS-only pass transistor decoding using two address lines. Draw a diagram with pass transistors that show this decoding.



3. (10 pts) One goal in SRAM design is to limit the bit line voltage swing.

   a. Why is limited bit line swing useful? What could be a problem with limiting bitline swing?

   Limiting the bitline swing will reduce power consumption, but it will make the bitlines more sensitive to injected noise, and require more complex sense amps.

   b. One method of limiting bit line swing was a word line kill circuit. The bitline swing circuit was used in a memory that had 128 rows x 64 columns. A word line kill was generated for every 16 rows via a bit line in the dummy bit column at the end of the array (bit line had 1/16 normal loading). Would the bit line swing increase or decrease if the word line kill was generated for every 32 rows instead 16 rows – explain your answer.

   This will increase the bitline swing because the loading on the dummy bit line will increase (go from $1/16^{th}$ normal loading to $1/8^{th}$ normal loading), which will delay the generation of wordline kill which will increase the bitline swing.

4. (10 pts) In a Domino logic pipeline, latches between stages can be eliminated by using overlapping clocks.

   a. What is the advantage of eliminating the latches?
   The latency of the latch represented by the C2Q propgation delay will be removed.

   b. What is the advantage of using more than two clocks in the domino logic pipeline?
   This gives more time for clock borrowing and skew tolerance.

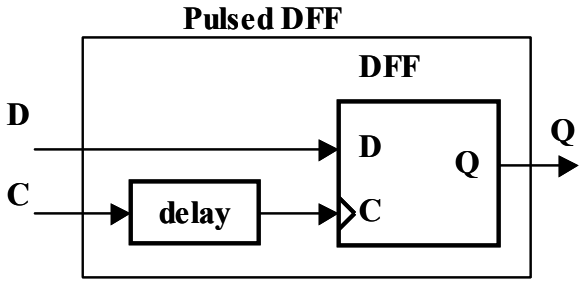
5. (5 pts) What does time borrowing mean in a pipelined system and why is it nice to be able to do this?
   The evaluation time of one clock phase spills over into the next clock phase thus 'borrowing' some time from the next clock phase. This allows us to tolerate mismatches in critical delay paths between the stages.

6. (5 pts) Assuming that we are using edge triggered, static logic, what is needed in order to take advantage of time borrowing?

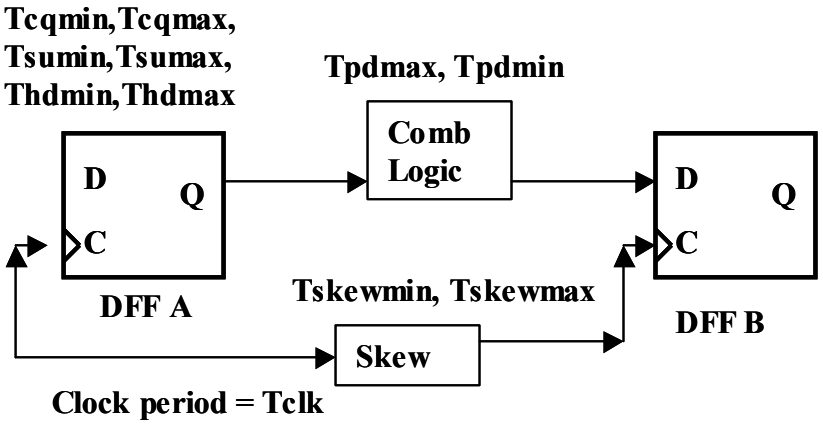
Need FFs with negative setup time.

7. (5 pts) A simple design for a pulsed DFF is shown below. The DFF has a setup time of 75 ps, and the delay is equal to 120 ps. What is the setup time of the pulsed DFF?

Setup = 75-120 = -45ps.



8. (10 pts) Timing verification is important for any VLSI system. Look at the diagram below. The timing parameters for the edge triggered DFFs are Thdmax, Thdmin (hold time), Tsumin, Tsumax (setup), and Tcqmin,Tcqmax (clock to q delay). The delay through the combinational logic is Tpdmax (maximum prop delay) and Tpdmin (minimum prop delay). The clock skew is in the range of Tskew min to Tskew max. The clock period is Tclk



a. Write an equation using the above timing parameters that should be used to check for setup time violations.

Without skew: $Tcqmax + Tpdpmax + Tsu\ max < Tclk$

With skew $Tcqmax + Tpdpmax + Tsu\ max < Tclk - Tskew\ min$ *(skew actually helps us in this case, we need to use the minimum skew).*

b. Write an equation using the above timing parameters that should be used to check for hold time (race) violations.

Without Skew $Tcqmin + Tpdmin > Thdmax$

With Skew $Tcqmin + Tpdmin > Thdmax + Tskewmax$

Skew hurts us in this case. We need to add the maximum value.

9. (5 pts) The folded bitline architecture for DRAMs was designed for improved noise suppression. Explain the key factor in the folded bitline architecture that provided improved noise suppression. Draw a diagram to illustrate your point.

   See notes. The folded bitline architecture makes the wordlines cross both bitlines so that the injected noise from wordline transistions is injected into both bitlines, which turns it into common mode noise which can be rejected by the sense amps.

10. (5 pts) What problem is the transposed bit line approach aimed at and how does it accomplish its goal?

    The transposed bitlines have the bit lines cross each other so that the noise from neighboring bitlines is coupled into both bitlines, which which turns it into common mode noise which can be rejected by the sense amps.

11. (5 pts) With SRAMs, the access time (time from address to data) is the same as cycle time (time for an entire read or write operation). However, with DRAMs the cycle time is longer than the access time. What basic aspect of DRAMs forces this?

    The read of a DRAM cell is destructive, so recovery time is needed to restore the value of the DRAM cell after read.

12. (5 pts) The Alpha 21164 was able to reduce clock skew by about ½ over the Alpha 21064. How did the clock structure of the Alpha 21164 accomplish this?

    The Alpha 21164 used two main clock buffers located at ¼ and ¾ along the chip axis, so the total distance between clock buffers was cut in half over the Alpha 21064 which used just one clock buffer located in the center of the chip.

13. (5 pts) What is the basic architectural difference between the Alpha 21264 clock distribution network and the Alpha 21164 clock network?

The Alpha 21264 use a hierarchy of clocks – global clock to major clock to local clocks (both conditional and unconditional). The Alpha 21164 only had a global clock.

14. (5 pts) The Alpha 21264 clock network saved a considerable amount of power over the Alpha 21164 clock network. Why? What features promoted these power savings?

The conditional clocks allowed the clocks to some functional units to be turned off when not in use, saving power. Also, the clock hierarchy meant the global clock grid did not have to be nearly as dense as it would have needed to be without a clock hierarchy, which saved power in the global clock grid.

15. (5 pts) Compare and contrast the Itanium and Alpha 21264 clock distribution networks. Compare and contrast means to give both similarities and differences between the two clock distribution networks.

Both were hierarchical clock networks. Both had conditional clocks. Both used shielded the clock grids with Vdd/Gnd.

The Itanimum deskewed the major clocks with respect to the global clocks, while the Alpha did not.

16. (5 pts) What is the key difference between the Itanium and the Alpha 21264 clock networks and why was it introduced by the Itanium designers?

The Itanimum deskewed the major clocks with respect to the global clocks using the active deskewing circuits, while the Alpha did not.

17. (5 pts) In an ASIC tool flow, what tool is used to get from an RTL representation to a gate level representation of a design?

    A logic synthesis tool.

18. (5 pts) In an ASIC tool flow for a submicron process, once the initial layout is generated, what steps are needed for timing verification?

    Need to extract parasitics, compute delays based on these, and back annotate the netlist. If the timing verification with back annotation does not meet timing specs, then use incremental synthesis and in-place layout optimization in order to adjust the design, and continue this loop until the timing specs are met.

19. (5 pts) What does timing driven layout mean?

    The layout tool attempts to generates routing that meet timing constraints.

20. (5 pts)What does back annotation mean?

    Delays computed from extracted parasitics from the physical layout are included in the timing simulation.

21. (5 pts) How would Formal verification save time in an ASIC tool flow?

    When a transformation of the netlist is made from one representation to another (such as RTL to gate level), a simulation would not need to be run to check the correctness since formal verification would prove the correctness.

22. (5 pts) The Itanimum designers used repeater stations to reduce the wire delay of global wires. Assume a wire fringing capacitance to substrate of 23.0 af/um (af = 1e-18 f), and wire to substrate capacitance of 8.0e-18 af/um^2, an ohms/sq value of 0.08, and a repeater delay of 50 ps. For a wire with length = 10,000 um and width = 0.4um, how many repeater stations should be added to the wire to reduce the wire delay to a minimum? Use the equation "0.9 * Rtotal * Ctotal" as an estimated wire delay. When computing the fringing capacitance contribution, assume the given capacitance value includes both sides of the wire. Show your work.

No buffers. Ctotal = Cfringe + Csub = 23.0 af/um * 10000 + 10000 * 0.4 * 8.0 af/um2 = 0.262 pf
Rtotal = 10000/0.4 * 0.08 = 2000
Delay = 0.9 * 2000 * 0.262 pf = 471 ps

With one buffer, wire length is one-half, so Ctotal = 1/2, Rtotal =1/2 , so wire delay = ¼
Total delay = 2 segments * 471/4 + 50 ps = 286 ps

With two buffers: wire segment delay = original delay / 9
Total delay = 3 segments* 471/9 + 2*50 ps = 257 ps

With three buffers wire segment delay = original delay / 16
Total delay = 4 segments * 471/16 + 3 * 50 = 268 ps

Two buffers are optimal.

23. (5 pts) Assume you are creating a standard cell library for a submicron process that has 5 metal layers. How many metal layers would you allow use of in the design of the standard cell and why?

You want to use as few metal layers as possible so that the other layers are available inter-cell routing and inter-block routing. Typically, only poly and M1 are allowed in a standard cell design, with only limited use of M2 (if M2 and M2 are reserved for inter-cell routing, the usage of M2 inside the standard cell will block over the cell routing with M2).