University of Pennsylvania
Department of Electrical and System Engineering
Computer Organization

ESE534, Spring 2010 — Assignment 4: Energy and Scaling — Monday, Feb. 8

Due: Wednesday, February 17, 12:00PM
(originally Monday, Feburary 15; change to accomodate canceled classes 2/15)

We will build up an energy model for memory and explore tradeoffs. We will use absolute and scaling numbers from the 2009 edition of the ITRS `http://www.itrs.net/Links/2009ITRS/Home2009.htm`.

In making this assignment reasonable, it may oversimplify in places. You are welcome to point out any oversimplifications you are uncomfortable with along with your answers.

You may want to use a spreadsheet with this assignment—particularly in cases where we ask you to explore the impacts of changing parameters and to plot tradeoff curves.

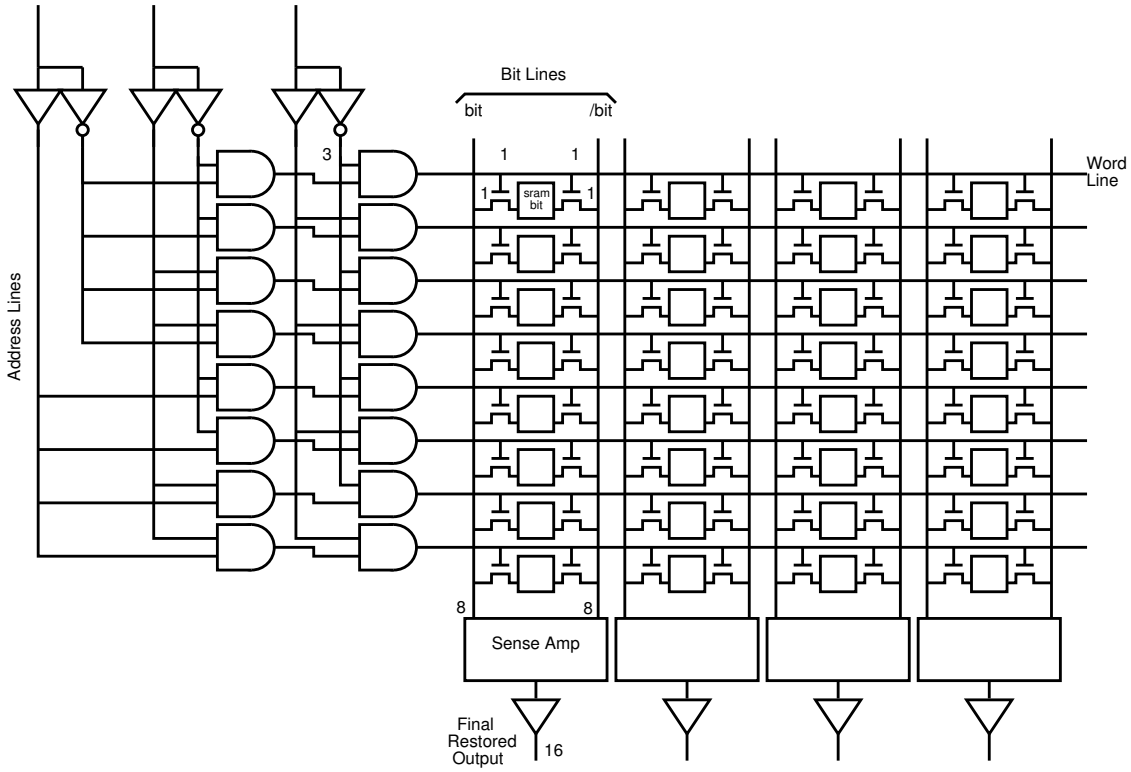

Figure 1: Memory Bank Organization

1. Develop an energy model for a single memory bank (Figure 1). We will focus on switching energy.[1]

   - We will assume no multiplexing on the output of this memory bank—that is, all the bitlines that come out of the bank are restored by sense amps and buffers.
   - Consider the following lines as switching: address lines, word line, bit lines, final restored output.
   - Assume the memory uses complementary bit lines (bit and /bit), so there are two bitlines per output bit read.
   - You should think about how many word lines switch on each memory operation.
   - Since we focus on switching energy, the key thing to watch is the load capacitance that must be switched.
     - Assume minimum feature size transistors for the transistor connected to the bit and word lines. This means each bit places a load of $2\times$ the mimimun transistor gate width on word line. The bit lines are actually connected to the drain of a transistor rather than the source; nonetheless, model the load each memory bit places on each of the bit lines as $1\times$ the minimum transistor gate load.[2]
     - Assume the final output load switched by the memory is a transistor $16\times$ the minimum transistor gate width in the process.
     - Assume the sense amp places a capacitive load on each of its two inputs (bit and /bit) that is $8\times$ the minimum transistor gate width. This is a very crude (and possibly inaccurate) characterization of a sense amp. Understanding sense-amp design is beyond the scope of this course.
     - Assume the address decoders are built out of a collection of AND2 gates, placing a load on each input of $3\times$ the minimum transistor gate widths. You may assume all the inputs to all the gates in each row decoder will switch.[3]
     - It is typical to run both polarities of each address bit into the address decoders, so that the inverters to produce the negations of the address bits are only performed once. For further simplicity, you are **not** considering the switching energy on the inputs of these inverters (and the buffers for the non-inverter address inputs), but you are considering the switching of the capacitive load they are driving (the Address Lines).
     - Our model for the capacitance contributed by a gate will be:

       $$C = (\text{number-of-minimum-transistor-widths}) \times F \times C_{g,total} \qquad (1)$$

       F is the feature size of the technology node (*e.g.* 45nm technology, has $F = 45\text{nm}$). $C_{g,total}$ is a technology-dependent, per unit gate width value we can find in the ITRS.

[1] Leakage energy is very important and often dominate for items like memory cells that do not switch at high rates. Nonetheless, one of the first simplification we make is to look only at switching energy.

[2] Another simplification...

[3] In practice, many may not switch.

- Problem 3 asks you about scaling, so it would be worthwhile to express your equations as functions of the ITRS parameters to allow you to easily recompute for different technology nodes. This is one place where you may want to use a spreadsheet.

(a) Write an equation for the energy of a read operation. This should include variables for the logical organization (w-bit wide, d-words deep), the capacitance of the relevant components (capacitance of each cell imposed on wordline, on bitline, ...), and the operating voltages, V_{dd}.

(b) Write equations for the capacitances appearing in part (a) in terms of the technology parameter $C_{g,total}$.

(c) Using parts (a) and (b) write your equation for the energy required for each read operation as a function of the technology parameters ($C_{g,total}$, V_{dd}) and the logical parameters (w,d).

2. Extend the energy model to a banked memory (Figure 2) and determine the area-energy tradeoff and energy-optimal bank size.

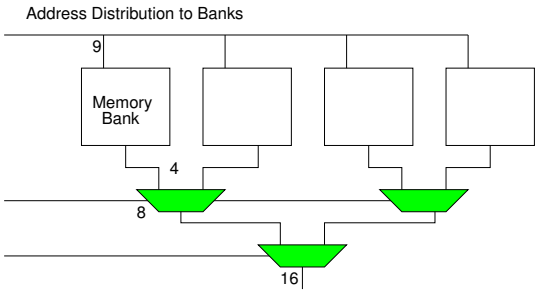

Figure 2: Banked Memory with $b = 4$ Banks

- Consider breaking an N-word memory into b banks ($b \times d = N$).
- Assume you only have to switch a single bank at a time.
- Assume the bank outputs are combined by a mux-tree of 2-input multiplexers. Model each 2-input multiplexer as placing a load of $8\times$ the minimum transistor width on its control inputs and $4\times$ the minimum transistor width on its data inputs.
- There is a mux that will need to be controlled for each bit of the datapath.
- We will assume the logic to select which bank to energize is small compared to the other terms and ignore it.
- It is necessary to drive the addresses to all of the banks. Assume each of the addresses sees the load of $9\times$ the minimum transistor width for each bank for this distribution. This loading is for buffering (isolating) the bank address load that you calculated in the previous problem. These buffers are also partially responsible for disabling the non-active banks.

(a) Write an equation for the energy per read operation for the banked memory similar to the one developed for the previous problem. Again, this should be reduced to a function of $C_{g,total}$ and V_{dd} as well as the organization parameters (b, d, w).

(b) What b will minimize the energy per read operation? (This is a symbolic questions. Your answer will be an equation.)

(c) Write an equation for the area of the banked memory, building on the area model from HW.3 (below):
- Area(2-input gate) = 2
- Area for a bank: d-entry, w-bit wide memory = $d(2\log_2(d)+w)+10w$ (assume this includes the sense amps)
- Compute the area of the 2-input mux based on its gate-level implementation.

(d) For the 45nm node, identify $C_{g,total}$ and V_{dd} in the ITRS (Table PID3B in the FOCUS C Tables set—footnotes explained at bottom PIDS3A table).

(e) Assuming $N = 2^{14}$, $w = 32$, compute the area and energy for b being powers of 2 from 2^0 to 2^{14}. If your answer to (b) suggests a non-power of two bank count, b, also calculate area and energy for that b. Plot the resulting area-energy tradeoff curve (note that b is **not** an axis in this graph; the axes are energy and area).

3. Determine the impact of scaling on energy and power. For concreteness, we ask you to evaluate this at four ITRS technology nodes: 45nm, 21nm, and 11.9nm, and 7.5nm. Assume area scales proportional to F^2.

   (a) Using your equations from Problem 2(b) and considering $N = 2^{14}$, $w = 32$, what is the energy-minimizing bank size at each technology node?
   (b) At the energy-minimizing bank size and using ITRS PIDS3B parameters appropriate to each technology, what is the energy per read in each technology?
   (c) Assuming these operate at the same frequency,[4] what is the power density associated with each technology.

   Summarize these results in a table with a row for each technology.

[4] A likely oversimplification...