CS228A: Hardware Models
Continued

George Varghese

January 14, 2002
Energy and Power

• Power \( P = CV^2 f \). New processes shrink voltage levels and capacitance but higher speed (OC-192, OC-768) speed circuits must increase clock frequency.

• Similarly, parallelism (e.g., pipelining) implies more capacitors being charged at a time.

• High speed chips dissipate a lot of heat, requiring non-trivial cooling techniques such as heat sinks. ISPs and co-location facilities are among the largest consumers of power.

• Some practical limits today are 30 watts per square centimeter on a single die, and 10,000 watts per square foot in a data center.
Priority Encoder Timing

- Input $I$ and outputs $O$ are $N$-bit vectors such that $O[j] = 1$ if and only if $I[j] = 1$ and $I[k] = 0$ for all $k < j$. Find-first one, represent in one-hot notation.

- $O[j] = \overline{I[1]} \ldots \overline{I[j-1]} I[j]$ for $j > 0$.

- **Design 1**: Implement directly using $N$ AND gates where each gate takes 1 to $N$ inputs. $O(N^2)$, appears to take $O(1)$ time.

- **Design 2**: Every output bit requires the AND of the complement of the first $j - 1$ input bits. Construct recursively using $P[j] = P[j-1] \overline{I[j]}$ using $N$ two-input AND gates in series. $O(N)$ transistors but takes $O(N)$ time.

- **Design 3**: Tradeoff design, more area efficient than 1 but faster than 2. Uses parallel-prefix. Build a balanced binary tree to compute $P(N)$ and then combine partial results.
Raising Design Abstraction Level

Hand designing 1 million transistors is infeasible.

- *Random Logic*: Compact form of software lookup table. Rewrite function as OR of product of complemented ANDs. Uses an AND plane to get products (chosen by connections) and an OR plane. More restrictive but simpler form called PAL.

  
  - A *decoder* converts a log $N$ bit binary value to an $N$ bits in unary
  
  - A *barrel shifter* shifts an input $I$ by $s$ positions to the left or right.
  
  - Mux connects one of $n$ input bits $I_j$ to the output $O$ if a log $n$ bit select signal $S$ encodes the value $j$ in binary. Demux does opposite.
With standard cells, game is reduction

Example: Divide and conquer reduction. Build a 4-input multiplexer from 2-input multiplexers. We

- Start by choosing one of $I_0$ and $I_1$ using a 2-input mux, and choosing one of $I_2$ and $I_3$ by another 2-input mux. How to combine using a third mux.

- Use select signal appropriately.
More satisfying example: Crossbar Scheduler

- PPE: like a PE but with rotating priority. Find first 1 beyond pointer position \( P \).
- Arises in switch arbitration. Several input links wish to send a packet to output \( L \) in time slot. Use PPE to grant to input \( I \) and set \( P = I \) for fairness. Used in Tin Tera and by Abrizio.
- Simple divide and conquer ideas: Use a barrel
shifter? 3 times faster design using thermometer encoding. Tested on TI library.
Memories

- Router forwarding done using logic but packets and forwarding state are stored in memories.
- Since memory access times are significantly slower than logic delays, memories form major bottlenecks in routers and endnodes.
- Different subsystems within a router require different memory characteristics. 200 msec worth of packet buffering (DRAM) versus forwarding table (SRAM)?
- Thus need simple models for different memory technologies. (registers, SRAMs, DRAMs, and interleaved memory technology.)
Registers

- Need to store a bit such that in the absence of writes and power failures, the bit stays indefinitely.
- Output of inverter leaks away via “parasitic”. So use feedback after flipping twice to get polarity right. Flip-flop. Register is ordered collection of flip-flops.
SRAM versus DRAM

- $N$ registers addressed by $\log N$ address bits $A$. Self-refreshing. Needs a decoder to translate $A$ to unary so add decode delay. On-chip SRAM 0.5 nsec, 32 Mbits: off-chip 5-10 nsec

- **Dynamic RAM**: SRAM flip-flop needs 5 transistors. Less dense than SRAM using a transistor plus capacitor. Leakage fixed by refresh. Slower because output not driven by the power supply as in SRAM. 40-60 nsec.
Form follows function: Page Mode DRAMs

- DRAM chips appear to quadruple in capacity every 3 years [?] and are heading towards 1 Gigabit on a single chip.
- Direct decoding of say 20 bits is hard. Use divide-and-conquer. Decode in two stages. First decode upper address bits to get row and then decode column using CAS. Reduces decoder from $O(N)$ to $O(\sqrt{N})$ gates.
- If accesses have spatial locality, access within rows without doing a row decode (RAS). Page mode.
Getting the best of both worlds: Interleaved Memory

- Way to increase memory throughput (bandwidth). Single DRAM with word size of 32 bits and cycle time of 100 nsec has 32 bits every 100 nsec. Can string together multiple banks on a chip to increase bandwidth.

- Instead of idling during this 100 nsec delay, the user can then place a second address for Bank 2, a third for Bank 3, and so on.

- Prominent examples include SDRAM with two banks, and RDRAM with 16 banks.
Example 4: Pipelined FlowId Lookups


- SRAM infeasible. DRAM too slow (20 DRAM accesses). So use interleaved memory and pipeline using a single lookup chip. Direct RDRAM runs at a 800 Mhz, read access of 60 nsec.

- $2^{16}$ is too small. So use RAMBUS page-mode and retrieve two 96-bits keys in one 256 bit access, which allows $3^{16}$ or 43 million flow IDs.
Pin count limitations matter

- Example: Router with five 10 Gb/sec links. Overall buffering 200 msec * 50 Gb/sec which is 10 Gigabits.

- Want to use DRAM, memory bandwidth needed = 2 * 50 = 100 Gbit (packets in and out) and overhead for headers for a total of 200 Gbits.

- Using a single direct RDRAM with 16 banks, specifications show peak memory bandwidth of 1.6 Gbyte/sec or 13 Gb/sec.

- Accessing each RDRAM requires 64 interface pins for data and 25 other pins for address and control, for a total of 90 pins. Thus 200 Gbps memory bandwidth requires 16 RDRAMs which totally require 1440 pins.

- A conservative upper bound on the number of pins on a chip is around 1000. Thus even with a very fast forwarding chip one needs at least one
more chip to drive data in and out of the RAMBUS packet buffers.
In the Real World: Chip Design Process

- Box architect partitions functions between chips, design team writes specification and then logic designers write RTL using Verilog or VHDL. Estimate block sizes using crude floorplan.

- In *synthesized* design, apply synthesis tools to generate hardware circuits. In *custom* design, on the other hand, the designer can design individual gates or drag-and-drop cells from a standard library.

- If the chip does not meet timing, the designer must change the design. Finally, the chip “tapes out”, is manufactured, and the first yield is inspected.

- Easy to add features before RTL is written. Second “spins” are expensive and better to work around.