High Speed CMOS Circuit Design

Lecture 10: Summary

1.0 Designing Fast CMOS Circuits

This course has covered the fundamentals of high speed CMOS VLSI design.

The first step in designing fast circuits is to be able to predict circuit speed. We began by modeling transistors as RC circuits to easily estimate the delay of circuits and optimize circuits without resorting to simulation. From this RC delay model, we developed the principles of Logical Effort, which let the designer quickly size circuits for highest performance, rather than oversizing or undersizing. Nevertheless, simulation is still an important tool, so we explored simulation techniques.

As processes continue to scale, gates are becoming so fast that wires dominate the performance of many circuits. We developed wire RC delay models and looked at noise and delay effects from capacitive coupling as well as design techniques to reduce wire delay. Eventually, even wire inductance becomes important; we described the potential effects of inductance on the most susceptible structures: wide busses, low-resistance signal lines, and the power network.

Then we returned to gate design. We found that static CMOS, pass-transistor logic, and domino circuits are the three most important families for practical high speed design. Pseudo-NMOS and other specialized circuits see occasional use., but most other families from the literature are ignored because of either indifferent performance or circuit pitfalls.

The next stage of the class explored assembling gates into pipelined systems with minimum overhead. Static pipelines can be efficiently constructed with either two-phase transparent latches or with pulsed latches; pulsed latches achieve slightly better performance at the expense of min-delay restrictions. Domino pipelines are most efficiently built by eliminating all the explicit latches and overlapping clocks to tolerate clock skew. Generating and distributing clocks with low skew is important to achieve good performance and prevent circuit failures. Asynchronous circuits eliminate clocks and clock skew, but in their place introduce control signals which may also be skewed.

The final portion of the class examined execution unit design. We looked at a variety of adder designs, from simple ripple carry adders through lookahead structures to logarithmic designs. We also looked at arrays including SRAMs, ROMs, and PLAs. We covered the principles of SRT division and observed that circuit design is more important than architecture for divider performance as long as a reasonable architecture is selected. We concluded with low-power design techniques. Power consumption is a steadily increasing problem and may eventually limit system performance; unfortunately, there is very little a circuit designer can do besides common-sense strategies of turning off unused blocks and making conscious choices about hardware which increases performance slightly at the expense of much larger power consumption.

2.0 Future Challenges

High speed circuit design is an exciting field because as each challenge is overcome, the demand for even higher speed raises new challenges. The two forces toward increased clock frequency are process scaling and reducing the number of gate delays per cycle. Let us look at the implications of each trend, then at some of the challenges which we will encounter in the GHz region and beyond.

Process scaling provides faster transistors with each generation. Unfortunately, wires are actually slowing down per unit length, so wires now dominate the delay of many paths. This means that planning chips for locality is increasingly important. Eventually, microarchitectures themselves will have to adapt to multiple clock latencies across a chip, just as they now recognize multiple clocks of latency between chips. Moreover, increasing frequencies and switching capacitances means that power consumption is steadily rising. Eventually we will reach the point that, like in the NMOS days, performance is limited not by circuit design, but rather by tolerable power consumption. At this point, microarchitecture and circuits will have to be simplified, discarding features that provide only a few percent of performance improvement at the cost of large amounts of power.

Cycle time is also decreasing on account of fewer gates per cycle. Cycle times were once well over 30 fanout-of-4 (FO4) inverter delays. As microarchitecture offers diminishing returns, faster cycle times are necessary to improve throughput. Cycle times were first reduced by good pipelining. They are now approaching the 16-20 range, requiring use of domino circuits in datapaths and dictating careful attention to clocking overhead which was once a negligible portion of the cycle. With extensive use of domino in datapath and control circuits along with microarchitectural simplifications, cycle times of 12 FO4 delays are probably achievable. Drastic rethinking of the pipelining, including multi-cycle ALU operations, can allow cycle times as short as 8 FO4 delays, at which point hiding clocking overhead is extremely difficult.

These reductions in raw gate delay and number of gates per cycle will enable operating frequencies of a GHz and beyond. Let us look at the implications of such design on microarchitecture, logic design, and circuit design:

2.1 Microarchitecture

As we have seen, high performance microarchitectures will need to explicitly recognize locality and tolerate latency between widely separated blocks. This may reverse some of the trend toward ever more complex machines because the small performance increase of the extra complexity will be overwhelmed by the extra wire delay as well as the power consumption.

Another important microarchitectural challenge is that machines no longer can be stopped on a dime; stalls take more than a single cycle to broadcast across a chip. Therefore, machines will require a FIFO-based decoupled design in which certain portions can stop for lack of resources without requiring all other portions to instantaneously stop as well.

2.2 Logic Design

High frequency design requires close interaction between logic and circuit design; logic designers no longer can specify a machine, then "throw it over the wall" to synthesis tools or circuit engineers for implementation. RTL must be written recognizing the very limited number of gate delays in a cycle. Consciously taking advantage of the efficient wide NOR and AOI structures provided by domino logic will produce better logic.

Moreover, logic designers must be concerned with power and testability. They should design units to shut down when they are not being used. Since probing low-level metal signals is now almost impossible, scan, self-test, and other design-for-testability features should be built into the machine. Power and test features are already widely used, and will only increase.

2.3 Circuit Design

As the number of gates per cycle decreases, overhead that was once just a small fraction of the cycle time becomes important. Therefore, conventional flip-flops are no longer acceptable. Static designs should use pulsed latches or two-phased latches. Domino designs should overlap clocks and eliminate latches using a scheme like skew-tolerant domino. Since most functions are non-monotonic, this pushes designers to use mostly dual-rail domino logic.

The clock network will be stressed more with each generation of process. Clock distribution will require lower skews to hold skew at a reasonable fraction of the cycle time. However, skews from wire delays and component mismatches are getting more severe. Asynchronous circuits could eliminate clocks and clock skew, but introduce difficult problems of their own. Clock domains will offer a practical solution, allowing greater skew at the cost of lower bandwidth over longer communication lines.

The power network will also become steadily more important. Both increasing power consumption and dropping supply voltages lead to higher supply currents. The IR drop they produce is also a larger fraction of the smaller supply. Thus, power supplies need much of the upper layers of metal for distribution. Good supply networks are also important to provide return paths to reduce inductance and to shield critical networks from capacitive coupling. The burden of power distribution may increasingly move to the package.

3.0 Conclusions

None of these challenges seem insurmountable in the near future. Creative designers, driven by a market hungry for ever-faster machines, will reach several GHz in the coming years. The biggest difficulties will be increasing design effort and power consumption. Exponentially increasing design complexities will force small design teams out of the mainstream unless they can achieve exponentially increasing productivity. Exponentially increasing power consumption will partially counter the complexity trend by forcing designers to select only features which have a good power/performance tradeoff.