

---

# Digital Integrated Circuit Design I

## ECE 425/525

### Chapter 4

*Professor R. Daasch*

Department of Electrical and Computer Engineering  
Portland State University  
Portland, OR 97207-0751  
(daasch@ece.pdx.edu)

<http://ece.pdx.edu/~ecex25>

---

## Chapter 4

- Many CMOS designs meet the logical requirements
- Fewer CMOS designs meet all requirements
  - delay, power, manufacturable, reliable
- CMOS design can be realized meet requirements from extensive trial and error simulation
- Initial guesses far from final solution
  - Improvements are slow
  - Lower reliability and difficult to reach yield targets
- Design choices that require a rapid estimate
  - What circuit topology?
  - How many stages between flip-flops
  - Size of transistors

## Chapter 4

- A simple delay model addresses these and other questions
  - Are all gates different, the delay model says no
  - Does every design decision need spice, the delay model says no
  - Can the decision be technology independent, the delay model says most of the time
- Delay models
  - $\text{delay} = \text{parasitic delay} + \text{stage delay}$
  - Stage delay
    - How much current can this gate deliver to the load compared to a predetermined base (typically an inverter)
    - What is the fanout of the gate output
  - Parasitic delay

## Chapter 4

- What would be the delay without an external load
  - Parasitic delay — aka internal delay is the internal delays of the (primarily) diffusion capacitance
- Logical effort is the ratio of the input capacitance to inverter with the drive same current
  - Major contributions to *RC* delay are
    - Gate delay — transistor sizing, circuit topology
    - Interconnect — wiring
    - Delay estimates are particularly important for the critical path
  - *RC* time constant first level delay estimate
  - Effective estimates early in design means improvements are structured for efficient design intervals

## Chapter 4

- Estimates are not for perfection but for reliability
    - Good estimates get 80%-90% of the final detailed solution
    - Common to expend design effort (and time) on small corrections when the overall solution is still unknown
    - $R_{eff}$  is the equivalent replacement of the average current through FET
    - Such an estimate is accurate for only a brief moment in the logic transition
    - $C_{eff}$  is the nominal load such that the transition time
- $$R_{eff} C_{eff} = \frac{t_{plh} + t_{phl}}{2}$$
- Contributions to  $C$ 
    - Diffusion — drain terminal of transistor at output of gate (ie parasitic delay)

## Chapter 4

- Gate oxide — gate terminals of transistors controlled by output of gate (ie logical effort)
  - Interconnect — the wire, metal, poly, contacts, vias linking drain to gate
- Contributions to  $R$ 
    - Series transistor network — effective increases  $L$  of combined logic transistor
    - Parallel transistor network — effective increase of  $W$  of combined logic transistor
  - Naturally propagation delay accumulates from gate to circuit, circuit to the system
  - There are a few (< 10%) of the gate paths that limit performance, the critical path
  - Rapid delay estimate within 10% key element to designer productivity

## Chapter 4

- A RC delay estimate includes all major design decisions
- Effective resistance of single FET

$$R_{\text{eff}} \propto \frac{L}{W} \frac{1}{V_{DD} - V_t}$$

- Internal capacitance of the gate
- $C_{\text{eff}} \propto W + \text{Fanout}$
- Logical effort — ratio of logic gate input capacitance ( $C_{\text{ox}}WL$ ) to standard inverter input capacitance
- Common gates NANDn
- $= (n + 2)/3$
- NORn
- $= (2n + 1)/3$

## Chapter 4

- Xor depends on network topology with different  $g$  values for different inputs
- The inverter delay is a convenient delay normalization
- Each inverter drive is scaled from minimum sizing
- Tracks each technology shift
- Inverter drive current proportional to input capacitance (transistor  $W$ )
- Inverter current combines all major contributions of process ( $t_{\text{ox}}$ ,  $N_A$ ,  $\Delta W$ ,  $\Delta L$  etc.)
- Logical effort is the slope of the linear delay model
- Electrical effort is the independent variable
- Unit analysis
- Normalized delay time is a scaled time constant

## Chapter 4

$$d = \frac{RC}{\text{inverter}}$$

- Delay capacitance is gate output capacitance
- Delay resistance is logic gates transistor sourcing or sinking current
- Logical effort is inverse of drive (not to be confused with output current)

$$\text{Drive} \equiv \frac{C_{in}}{\text{logicaleffort}}$$

Logical effort indicates how much worse a gate is at producing output current compared to the inverter, Weste p. 166

Logical effort measures the effects of transistor sizing on the input transition compared to the inverter input

## Chapter 4

- Logical effort favors NANDs over NORs
- Identical functions may have different logical efforts
- Different inputs of the same gate may have different logical efforts
- First estimate parasitic delay
- Common gates have same internal delay (pullup and pulldown networks are duals)
- NANDs and NORs have the same output diffusion capacitances
- Second method parasitic delays consider internal capacitances second (Elmore) model

$$= R(3nC) + \sum_i^{n-1} iRC = \left(\frac{n(n-1)}{2} + 3n\right)RC$$

## Chapter 4

- As the series transistor stack grows the delay grows by  $n^2$
- Internal parasitic capacitance favors NAND gates
- A chain of identical inverters each block has  $g=1$ ,  $h=1$ ,  $p=1$   
$$d = g \cdot h + p = 1 \cdot 1 + 1 = 2$$
- Two units of normalized delay for each inverter in the chain  
normalized frequency =  $1/d$
- Logical effort for single gates
- Evaluate transistor topology options for a specific Boolean function
- Transistors sized for equivalent gate-to-gate performance for varying output load
- Logical effort for multiple stage logic functions

## Chapter 4

- Expands beyond a single logic gate design
- Select topology (as before)
- Select the number of levels in the logic
- Size transistors for best (optimum) overall performance (internal gate-to-gate load and output load)
- Method for computing the logical effort
- Care must be used to identify the path in multi-path circuits
- Selecting logic gate (NAND, NOR etc) determines the logical effort
- Inverter of some size normalizes the effects of increasing output current
- Output and number of load sets transistor sizing
- 1) Compute the path effort  $F = GBH$

## Chapter 4

- 2) Estimate the stages  $N = \log_4 F$
  - 3) Sketch path for N
  - 4) Estimate least delay  $D = NF^{1/N} + P$
  - 5) Determine stage effort  $\hat{f} = F^{1/N}$
  - 6) Find gate widths (sizes)  $C_{in,i} = g_i C_{out,i} / \hat{f}$
- Step 5 key conclusion from constrained sums and products
    - Minimize the sum of  $f_1 + f_2 \dots$  and limit the product  $f_1 \cdot f_2 \dots$  to a constant
$$f_i = constant = F^{1/N}$$
  - Sets the number of stages for driving specific output load
- Roughly is best with effort delays of 4 nominal inverters
  - Smaller gates (inverters and NANDs) best for large output loads

## Chapter 4

- Limitations of the linear logic effort model
  - G and stages are dependent on each other
  - No rise and fall times of the inputs (see laboratory)
  - Interconnect adds a significant delay
  - Optimization speed not power or area
- CMOS power dissipation is limited to dynamic power, a transient short circuit power and a small, static leakage (at or near the transistor)
- Static active area reverse, diode leakage is small and generally constant
- Active area leakage smaller contribution to static power each generation
- Power is an equal partner in the design metric list

## Chapter 4

- Energy replaces power as a metric

$Power \cdot delay = Energy$  = aka Power-delay product

or

$Power \cdot delay^2$  aka Energy-delay product

- The other are area, timing and testability
- Dynamic power is the charging and discharging of the output loads at fixed frequency and supply voltage
- Not all capacitance changes state every clock period (there is one notable exception)

$$Power = CloadV_{DD}^2F$$

$$Power = \alpha CloadV_{DD}^2F$$

where  $\alpha$  tracks the average switching rate (toggle rate)

## Chapter 4

- Simple ideas to reduce power
  - Cut down on active average switching rate
  - Reduce the supply voltage generally means scale the frequency too
  - Dynamic power ranges from  $< 1W$  to  $100W$
  - Chips with power  $150W$  or so have packaging and other system limitations
  - Chip a  $1W$  impose limitations on the best battery powered devices
  - Dynamic power is a decreasing fraction of the total power

$$P_{total} = P_{dyn} + P_{other}$$

- Power-delay (energy) is a better metric at  $100W$  because the requirement for  $100W$  may be for a few  $10s$  of microseconds

## Chapter 4

$$100W \cdot 10 \mu\text{seconds} = 1000 \mu J = 1mJ$$

- Clocks and other periodic circuit nodes (sorted by size of capacitive load) are a primary focus for reducing dynamic power
- Increasing the stage effort, reduces transistor size (ie capacitance) increases delay
- Transient leakage is the connection between  $V_{DD}$  and GND when pullup and pulldown networks are connected to the load
- Transient leakage is a function of input slew rate
  - Poor matching between output load capacitance and sizing pullup and pulldown drivers
  - Slower the transitions means longer times when both networks are conducting
  - Typical rise/fall times the transient power is a few percent of the total

## Chapter 4

- Static leakage is the sum of subthreshold, gate to source, active area and potentially ratios of pseudo-NMOS or diode connected topologies

$$P_{static} = I_{static} V_{DD}$$

- Large range from 10mW to 10W

$$\frac{100mW}{100} \text{ MHz} = 100nJ$$

where the 100MHz is from the clock rate of the dynamic power

- Two components are increasing with technology scaling
  - Charge tunneling from gate to source or gate to channel is significant only in deep-submicron
  - Statistically and operationally Gate-oxide of 20 Ang or less result in charge tunneling

## Chapter 4

- Subthreshold currents are exponentially controlled by  $V_t$  and increase with decreasing  $V_t$
- Subthreshold drain-source current
- Subthreshold leakage is a double problem in battery systems
- Lower supply voltage to reduce dynamic power requires a smaller  $V_t$
- Most 130nm and below processes have two, three or more different transistor thresholds
- Use the body-effect attached to a second supply rail
- Add dummy transistors to short stacks
- Interconnect is more common than transistors
- About 2 wires for transistors 1 million to 10 million transistors

## Chapter 4

- Transistors live under the wires
- Transistors are manufactured in the “front-end” process and wires are the process “back-end”
- 3 to 8 layers of wire
- Interconnect contributes to all design metrics speed, power, signal integrity (noise)
- Design of wire is described by their pitch (width plus separation or spacing)
- Manufacture of wire has to consider the thickness
- Wires near transistors are thin and get thicker at the top of the chip
- Thick wires are good for power distribution
- Wires have resistance and capacitance and are also scaled by a normalization (Ohms per square)

## Chapter 4

- All metals are not the same (best is low resistivity and low diffusivity)
- Other semiconductor layers (poly, well, active) are lousy wires
- Each wire change requires a via (design rules etc) and adds 1 Ohm per series contact
- Wire capacitances are all to all
- Cross-talk is active wire to active (or passive) wire
- Cross-talk as a concern for noise is not a major concern with static CMOS
- Similarly power supply noise, leakage and noise feed-through affect delay and potential disrupt operations
- Capacitive noise and leakage affect the reliability of the chip
- Reliability is a lesson in practical probability

## Chapter 4

- The goal is to guess the lifetime of what at first is a perfect part
- Reliability is often measured by “failures in time” (FIT)

$$FIT = 10^9 \cdot \frac{\text{failure rate}}{\text{hour}}$$

- “mean time to failure”

$$MTBF = \frac{\text{Number of chips} \cdot \text{Operating hours}}{\text{number of failure}}$$

- Some parts fail the first moment you turn them on
- Some parts fail shortly after (infant mortality)
- Combining the two is a controlled stress called burn-in parts fail only under voltage, temperature and current stress

## Chapter 4

- Some parts work forever
- Some parts fail when they get old (hours of operation not calendar years)
- Common modes of failure can be circumvented by process, design or both
- Design rules an obvious example of design for reliability
- Major sources of reliability risk include
  - Electromigration
  - Self-heating
  - Hot-carriers
  - Latchup
  - Overvoltage failures
- Electromigration — electron flow is a charge wind and a force for movement of individual metal atoms

## Chapter 4

- DC currents are the major concern
- Step by step the continuous metal wire forms voids around grain boundaries which grow and induce a general material flow in direction of current
- For most metal wires limit current density to 1-2mA per  $\mu m^2$
- Upper metals tend to carry larger loads and so are thicker
- Cu wires increase the limit 5x
- Extreme limit is an open circuit analogous to a blown fuse
- Significant dependence on temperature
- Self-heating is the AC effect
  - The dielectric (silicon dioxide) is a thermal insulator surrounding every wire

## Chapter 4

- Higher temperatures raises the resistance to create positive feedback loop
- AC currents with sufficient self-heating compounded by electromigration
- More resistive wires increase the delays in the circuit
- Better thermal management (heat sinks or heat conductive substrates) lowers the probability of reliability failure
- Hot-carriers
  - Charge carriers in channel have high kinetic energies
  - Scattered carriers can be “frozen” into the substrate
  - For fixed  $V_{gs}$  nMOS (holes don't do this) additional oxide trapped negative charge reduces the channel charge
  - Reduced current leads to slower transitions and long-term failures

## Chapter 4

- Oxide thickness,  $V_{DD}$  and doping (LDD) all reduce effects of hot-carriers
- Slow rise times means longer periods of large current more chances for electrons to embed in oxid
- Latchup is a active switching problem leading to runaway DC currents
  - Only occurs in CMOS
  - Problem of parasitic bipolar transistors
    - NPN between nwell, p-substrate (base) and  $n^+$  active (emitter)
    - PNP between p-substrate, nwell (base) and  $p^+$  active (emitter)
    - Base and collector are shared between two devices
  - Need small shunt resistors to reduce or eliminate chance that both transistors are forward-active

## Chapter 4

- Generally less of a problem in current technologies
- Key effect is the switching lowering the nwell potential (forward bias PNP) or increasing the substrate voltage (forward bias NPN)
- Additional guard-ring charge collector around potentially sensitive areas
- Overvoltage occur from poor handling (electrostatic discharge), Oxide breakdown and arcing
- Time-dependent oxide breakdown as a result of small dimension thin-oxide wears out from tunneling
- Static CMOS is just a vulnerable as other logic families
- Robust and reliable designs generally limit the performance
- Design rules and other static checks can not detect active failures

## Chapter 4

- Probability of failures is combined with the probability of circuits manufactured
- Design margins measures the effects of expected variability of the process
- Each parameter is not a single unvarying number but distribution of values
- Parameters for each transistor, wire, contact etc. is a unique combination selected from the distributions
- Within the chip operation operating conditions are not constant
- Local resistances cause non-uniform voltage drops across the chip
- Local activity (recall dynamic power) results in a local temperature gradient

## Chapter 4

- Selecting which combination of parameters a design is most sensitive to is a art
- Some general rules can be used
  - Slow, Fast and nominal transistors tend to vary slowly across the chip
  - A particular gate will have nearly the same nFETS and another nearly constant set of pFETS
  - The nFETS and the pFETS may come from different parts of their respective distributions
  - Some circuits like diode-connected or pseudo-nmos are more affected by the Fast/Slow or Slow/Fast combinations
  - Some circuits such as delay sensitive circuits are more affected by the Fast/Fast or Slow/Slow combinations
- Mismatching of like transistors can be controlled

## Chapter 4

- Timing and other critical global signals are more sensitive than other local signals
- All transistors have source and drain in the vertical direction
- Dummy polysilicon or other materials can be added to the physical design to maintain constant density of materials