On a small chip, the clock distribution network is just a wire
And possibly an inverter for clkb
On practical chips, the RC delay of the wire resistance and gate load is very long
Variations in this delay cause clock to get to different elements at different times
This is called clock skew
Most chips use repeaters to buffer the clock and equalize the delay
Reduces but doesn’t eliminate skew
Skew comes from differences in gate and wire delay
- With right buffer sizing, $clk_1$ and $clk_2$ could ideally arrive at the same time
- But power supply noise changes buffer delays
- $clk_2$ and $clk_3$ will always see RC skew
Ideally full cycle is available for work

Skew adds sequencing overhead

Increases hold time too

\[ t_{pd} \leq T_c - (t_{setup} + t_{pcq} + T_{skew}) \]

sequencing overhead

\[ t_{cd} \geq t_{hold} - t_{ccq} + t_{skew} \]
- Flip-flops are sensitive to skew because of **hard edges**
  - Data launches at latest rising edge of clock
  - Must setup before earliest next rising edge of clock
  - Overhead would shrink if we can soften edge
- Latches tolerate moderate amounts of skew
  - Data can arrive any time latch is transparent
2-Phase Latches

\[ t_{pd} \leq T_c - (2t_{pdq}) \]

sequencing overhead

\[ t_{cd1}, t_{cd2} \geq t_{hold} - t_{ccq} - t_{nonoverlap} + t_{skew} \]

\[ t_{borrow} \leq T_c/2 - (t_{setup} + t_{nonoverlap} + t_{skew}) \]

Pulsed Latches

\[ t_{pd} \leq T_c - \max(t_{pdq}, t_{pcq} + t_{setup} - t_{pw} + t_{skew}) \]

sequencing overhead

\[ t_{cd} \geq t_{hold} + t_{pw} - t_{ccq} + t_{skew} \]

\[ t_{borrow} \leq t_{pw} - (t_{setup} + t_{skew}) \]
- Static circuits are slow because fat pMOS load input
- Dynamic gates use precharge to remove pMOS transistors from the inputs
  - Precharge: $\phi = 0$, output forced high
  - Evaluate: $\phi = 1$, output may pull low
Domino Circuits

- Dynamic inputs must monotonically rise during evaluation
  - Place inverting stage between each dynamic gate
  - Dynamic/static pair called domino gate
- Domino gates can be safely cascaded
Domino Timing

- Domino gates are 1.5 – 2x faster than static CMOS
  - Lower logical effort because of reduced $C_{in}$
- Challenge is to keep precharge off critical path
- Look at clocking schemes for precharge and evaluate
  - Traditional schemes have severe overhead
  - Skew-tolerant domino hides this overhead
Traditional Domino Circuits

- Hide precharge time by ping-ponging between half-cycles
  - One evaluates while other precharges
  - Latches hold results during precharge

\[ t_{pd} = T_c - 2t_{pdq} \]
Clock Skew

- Skew increases sequencing overhead
  - Traditional domino has hard edges
  - Evaluate at latest rising edge
  - Setup at latch by earliest falling edge

\[ t_{pd} = T_c - 2t_{pdq} - 2t_{skew} \]
Logic may not exactly fit half-cycle
- No flexibility to borrow time to balance logic between half cycles
- Traditional domino sequencing overhead is about 25% of cycle time in fast systems!
Relaxing the Timing

- Sequencing overhead caused by hard edges
  - Data departs dynamic gate on late rising edge
  - Must setup at latch on early falling edge
- Latch functions
  - Prevent glitches on inputs of domino gates
  - Holds results during precharge
- Is the latch really necessary?
  - No glitches if inputs come from other domino
  - Can we hold the results in another way?
Use overlapping clocks to eliminate latches at phase boundaries

Second phase evaluates using results of first

No latch at phase boundary
After second phase evaluates, first phase precharges
Input to second phase falls
  - Violates monotonicity?
But we no longer need the value
Now the second gate has a floating output
  - Need full keeper to hold it either high or low
Time Borrowing

- Overlap can be used to
  - Tolerate clock skew
  - Permit time borrowing
- No sequencing overhead

\[ t_{pd} = T_c \]
Multiple Phases

- With more clock phases, each phase overlaps more
  - Permits more skew tolerance and time borrowing
Clock Generation

en clk

\[ \phi_1 \]

\[ \phi_2 \]

\[ \phi_3 \]

\[ \phi_4 \]
Opportunistic Time Borrowing

U. S. Patent no. 5517136 (Harris et al., May 14, 1996, assigned to Intel Corporation)
Pipelined domino logic allowing a slow stage to “borrow” from the time normally allocated to a faster stage

[Diagram of domino logic with labeled inputs and outputs]

To High Skew

Vcc

A
B
CLK
D1

To High Skew

Vcc

A
B
CLK
D1K
Clocking of Time-Borrowing Pipeline

- Delayed falling edges on clocks allow evaluation to continue into subsequent half cycle
  - Time delay $t_d$ should be greater than or equal to the hold time of the domino logic gate plus any global clock skew
- Can generate the clocks by a local reference driven by the chip’s global reference clock signal
Example of an OTB Pipeline

Half-cycles 1 and 3 evaluate when CLK is high, half-cycle 2 when CLK is low
Flip-flop delay versus data arrival time
Flip-Flop Setup and Hold Times – Different Data Values

Input = 0
D
\[ t_{\text{setup0}} \]
\[ t_{\text{hold0}} \]
\[ t_{\text{pcq0}} \]
\[ t_{\text{ccq0}} \]
\[ \phi \]
\[ Q \]

Input = 1
D
\[ t_{\text{setup1}} \]
\[ t_{\text{hold1}} \]
\[ t_{\text{pcq1}} \]
\[ t_{\text{ccq1}} \]
\[ \phi \]
\[ Q \]
Blue Elements use $V_{DDL}$
Metastability

(a)

(b)

(c)

(d)

Metastable state in static latch
Metastable Transients and Propagation Delay

(a) Diagram showing metastable states and propagation delay with parameters $t_m$, $\phi$, $V_m$, and $Q$.

(b) Graph depicting $t_{\phi}$ as a function of $t_{DC} - t_m$ with parameters $h$ and $t_{pdq}$. 

ECE Department, University of Texas at Austin

Lecture 21. Skews, Scaling

Jacob Abraham, November 10, 2020
Simple Synchronizer

Diagram of a simple synchronizer showing the input (D) and output (Q) signals, with intermediate states X, F1, and F2. The diagram includes timing notations such as $T_c$, $t_{\text{setup}}$, and $t_{\text{pcq}}$. The metastable time is indicated by the blue line.
Handshake protocols

(a) Four-Phase

(b) Two-Phase
Wave Pipelining

(a) Waveforms for clock skew.

(b) Waveforms for clock scaling.

(c) Waveforms for both skew and scaling.
Chip Densities increase with Scaling

- In 1965, Gordon Moore predicted the exponential growth of the number of transistors on an IC (Moore’s Law)
- Transistor count doubled every year since invention
- Predicted $> 65,000$ transistors by 1975!
- Growth limited by power
The only constant in VLSI is constant change
- Feature size shrinks by 30% every 2-3 years
  - Transistors become cheaper, and faster
  - Wires do not improve (and may get worse)
- Scale factor $S$ (typical technology nodes) $S = \sqrt{2}$
Scaling Assumptions

- What changes between technology nodes?
- **Constant Field Scaling**
  - All dimensions \((x, y, z \Rightarrow W, L, t_{ox})\)
  - Voltage \((V_{DD})\)
  - Doping levels
- **Lateral Scaling**
  - Only gate length \(L\)
  - Often done as a quick gate shrink \((S = 1.05)\)
### Table 4.15 Influence of scaling on MOS device characteristics

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sensitivity</th>
<th>Constant Field</th>
<th>Lateral</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scaling Parameters</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length: $L$</td>
<td>1/$S$</td>
<td>1/$S$</td>
<td>1</td>
</tr>
<tr>
<td>Width: $W$</td>
<td>1/$S$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Gate oxide thickness: $t_{ox}$</td>
<td>1/$S$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Supply voltage: $V_{DD}$</td>
<td>1/$S$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Threshold voltage: $V_{in}, V_{tp}$</td>
<td>1/$S$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Substrate doping: $N_A$</td>
<td>$S$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>Device Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta$</td>
<td>$\frac{W}{L \cdot t_{ox}}$</td>
<td>$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>Current: $I_{ds}$</td>
<td>$\beta(V_{DD} - V_t)^2$</td>
<td>1/$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>Resistance: $R$</td>
<td>$\frac{V_{DD}}{I_{ds}}$</td>
<td>1</td>
<td>1/$S$</td>
</tr>
<tr>
<td>Gate capacitance: $C$</td>
<td>$\frac{W L}{t_{ox}}$</td>
<td>1/$S$</td>
<td>1/$S$</td>
</tr>
<tr>
<td>Gate delay: $\tau$</td>
<td>$RC$</td>
<td>1/$S$</td>
<td>1/$S^2$</td>
</tr>
<tr>
<td>Clock frequency: $f$</td>
<td>$1/\tau$</td>
<td>$S$</td>
<td>$S^2$</td>
</tr>
<tr>
<td>Dynamic power dissipation (per gate): $P$</td>
<td>$C V^2 f$</td>
<td>1/$S^2$</td>
<td>$S$</td>
</tr>
<tr>
<td>Chip area: $A$</td>
<td></td>
<td>1/$S^2$</td>
<td>1</td>
</tr>
<tr>
<td>Power density</td>
<td>$P/A$</td>
<td>1</td>
<td>$S$</td>
</tr>
<tr>
<td>Current density</td>
<td>$I_{ds}/A$</td>
<td>$S$</td>
<td>$S$</td>
</tr>
</tbody>
</table>
Observations

- Gate capacitance per micron is nearly independent of process
- But ON resistance \( \times \) micron improves with process
- Gates get faster with scaling (good)
- Dynamic power goes down with scaling (good)
- Current density goes up with scaling (bad)
- Velocity saturation makes lateral scaling unsustainable

Solution

- Gate capacitance is typically about 2 fF/\( \mu \)m
- The FO4 inverter delay in the TT corner for a process of feature size \( f \) (in nm) is about 0.5f ps
- Estimate the ON resistance of a unit (4/2 \( \lambda \)) transistor
  \[ FO4 = 5 \tau = 15 RC \]
  \[ RC = (0.5f)/15 = (f/30) \text{ ps/nm} \]
  \[ \text{If } W = 2f, \ R = 8.33 \text{ k}\Omega \]

Unit resistance is roughly independent of \( f \)
Scaling Assumptions

- Wire thickness
  - Hold constant vs. reduce in thickness
- Wire length
  - Local/scaled interconnect
  - Global interconnect
  - Die size scaled by $D_c \approx 1.1$
### Table 4.16 Influence of scaling on interconnect characteristics

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sensitivity</th>
<th>Reduced Thickness</th>
<th>Constant Thickness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scaling Parameters</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Width: $w$</td>
<td>$\frac{1}{w}$</td>
<td>$S^2$</td>
<td>$S$</td>
</tr>
<tr>
<td>Spacing: $s$</td>
<td>$1/s$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thickness: $t$</td>
<td>$1/t$</td>
<td></td>
<td>$1$</td>
</tr>
<tr>
<td>Interlayer oxide height: $h$</td>
<td>$1/h$</td>
<td></td>
<td>$1$</td>
</tr>
<tr>
<td><strong>Characteristics Per Unit Length</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wire resistance per unit length: $R_w$</td>
<td></td>
<td>$S^2$</td>
<td>$S$</td>
</tr>
<tr>
<td>Fringing capacitance per unit length: $C_{uf}$</td>
<td></td>
<td>$1$</td>
<td>$S$</td>
</tr>
<tr>
<td>Parallel plate capacitance per unit length: $C_{up}$</td>
<td>$\frac{w}{h}$</td>
<td>$1$</td>
<td>$1$</td>
</tr>
<tr>
<td>Total wire capacitance per unit length: $C_w$</td>
<td>$C_{uf} + C_{up}$</td>
<td>$1$</td>
<td>between $1, S$</td>
</tr>
<tr>
<td>Unrepeated RC constant per unit length: $\tau_{uw}$</td>
<td>$R_w C_{uw}$</td>
<td>$S^2$</td>
<td>between $S, S^2$</td>
</tr>
<tr>
<td>Repeated wire RC delay per unit length: $\tau_{wr}$ (assuming constant field scaling of gates in Table 4.15)</td>
<td>$\sqrt{RCR_w C_w}$</td>
<td>$\sqrt{S}$</td>
<td>between $1, \sqrt{S}$</td>
</tr>
<tr>
<td>Crosstalk noise</td>
<td>$\frac{1}{s}$</td>
<td>$1$</td>
<td>$S$</td>
</tr>
</tbody>
</table>
### Table 4.16 Influence of scaling on interconnect characteristics

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sensitivity</th>
<th>Reduced Thickness</th>
<th>Constant Thickness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scaling Parameters</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Width: $w$</td>
<td></td>
<td></td>
<td>$1/S$</td>
</tr>
<tr>
<td>Spacing: $s$</td>
<td></td>
<td></td>
<td>$1/S$</td>
</tr>
<tr>
<td>Thickness: $t$</td>
<td></td>
<td>$1/S$</td>
<td>$1$</td>
</tr>
<tr>
<td>Interlayer oxide height: $h$</td>
<td></td>
<td></td>
<td>$1/S$</td>
</tr>
<tr>
<td><strong>Local/Scaled Interconnect Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length: $l$</td>
<td></td>
<td>$1/S$</td>
<td></td>
</tr>
<tr>
<td>Unrepeated wire RC delay</td>
<td>$P_{t_{wu}}$</td>
<td>1</td>
<td>between $1/S$, 1</td>
</tr>
<tr>
<td>Repeated wire delay</td>
<td>$l_{t_{wr}}$</td>
<td>$\sqrt{1/S}$</td>
<td>between $1/S$, $\sqrt{1/S}$</td>
</tr>
<tr>
<td><strong>Global Interconnect Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length: $l$</td>
<td></td>
<td>$D_c$</td>
<td></td>
</tr>
<tr>
<td>Unrepeated wire RC delay</td>
<td>$P_{t_{wu}}$</td>
<td>$S^2D_c^2$</td>
<td>between $SD_c^2$, $S^2D_c^2$</td>
</tr>
<tr>
<td>Repeated wire delay</td>
<td>$l_{t_{wr}}$</td>
<td>$D_c\sqrt{S}$</td>
<td>between $D_c$, $D_c\sqrt{S}$</td>
</tr>
</tbody>
</table>
Observations

- Capacitance per micron is remaining constant
  - About 0.2 fF/μm
  - Roughly 1/10 of gate capacitance
- Local wires are getting faster
  - Not quite tracking transistor improvement
  - But not a major problem
- Global wires are getting slower
  - No longer possible to cross chip in one cycle
<table>
<thead>
<tr>
<th>Table 4.17</th>
<th>Predictions from the 2002 ITRS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Feature size (nm)</strong></td>
<td>130</td>
</tr>
<tr>
<td>$V_{DD}$ (V)</td>
<td>1.1–1.2</td>
</tr>
<tr>
<td>Millions of transistors/die</td>
<td>193</td>
</tr>
<tr>
<td>Intermediate wire pitch (nm)</td>
<td>450</td>
</tr>
<tr>
<td>Interconnect dielectric constant</td>
<td>3–3.6</td>
</tr>
<tr>
<td>I/O signals</td>
<td>1024</td>
</tr>
<tr>
<td>Clock rate (MHz)</td>
<td>1684</td>
</tr>
<tr>
<td>FO4 delays/cycle</td>
<td>13.7</td>
</tr>
<tr>
<td>Maximum power (W)</td>
<td>130</td>
</tr>
<tr>
<td>DRAM capacity (Gbits)</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Technology Roadmaps

- Many steps needed to produce an IC
- Each step requires specialized (and expensive) equipment produced by different vendors
- Roadmaps give equipment manufacturers an idea what equipment would be used, and when the capability would be needed (example, scaling factor of $\sqrt{2}$)

ITRS established in 2013

- Scaling projections to 2028
- With Moore’s law coming to an end, the final roadmap was issued in 2016
- Through IEEE’s Rebooting Computing initiative, the IRDS was started
Some of the Focus Team Topics in the International Roadmap for Devices and Systems

- Application Benchmarking
- Systems and Architectures
- More Moore
- Beyond CMOS
- Packaging Integration
- Outside System Connectivity
- Factory Integration
- Lithography
- Metrology
- Emerging Research Materials
- Environment, Safety, Health, and Sustainability
- Yield Enhancement
- Cryogenic Electronics and Quantum Information Processing (added in 2018)
Scaling Implications

- Improved Performance
- Improved Cost
- Interconnect Woes
- Power Woes
- Productivity Challenges
- Physical Limits
We can’t send a signal across a large fast chip in one cycle anymore.

But the microarchitect can plan around this.

Just as off-chip memory latencies were tolerated.

Globally Asynchronous, Locally Synchronous (GALS)
VLSI Economics

- Selling price $S_{\text{total}}$
  - $S_{\text{total}} = C_{\text{total}}/(1 - m)$
- $m =$ profit margin
- $C_{\text{total}} =$ total cost
  - Nonrecurring engineering cost (NRE)
  - Recurring cost
  - Fixed cost

NRE

- Engineering cost
  - Depends on size of design team
  - Include benefits, training, computers
  - CAD tools:
    - Digital front end: $10K$
    - Analog front end: $100K$
    - Digital back end: $1M$

- Prototype manufacturing
  - Mask costs: $500K – 1M$ in 130 nm process
  - Test fixture and package tooling
Recurring and Fixed Costs

Recurring costs

- Fabrication
  - Wafer cost/(Dice per wafer × Yield)
  - Wafer cost: $500 - $3000
  - Dice per wafer:
    \[ N = \pi \left[ \frac{r^2}{A} - \frac{2r}{\sqrt{2A}} \right] \]
  - Yield: \[ Y = e^{-AD} \]
    - For small A, \( Y \approx 1 \), cost proportional to area
    - For large A, \( Y \to 0 \), cost increases exponentially

- Packaging

- Test

Fixed costs

- Data sheets and application notes
- Marketing and advertising
- Yield analysis
You want to start a company to build a wireless communications chip. How much venture capital must you raise?

Because you are smarter than everyone else, you can get away with a small team in just two years:

- Seven digital designers
- Three analog designers
- Five support personnel
Digital designers
- $70k salary
- $30k overhead
- $10k computer
- $10k CAD tools
- Total:
  $120k × 7 = $840k

Analog designers
- $100k salary
- $30k overhead
- $10k computer
- $100k CAD tools
- Total:
  $240k × 3 = $720k

Support staff
- $45k salary
- $20k overhead
- $5k computer
- Total:
  $70k × 5 = $350k

Fabrication
- Back-end tools: $1M
- Masks: $1M
- Total: $2M/year

Summary:
- 2 years at $3.91M/year
- $8M design and prototype
New chip design is fairly capital-intensive

Maybe you can do it for less?