L15: VLSI Integration and Performance Transformations

- Moore’s Law and Integration
- IC Design
  - ASIC Design
  - Clocks
  - Test
- Performance Transformations
- Trends

Acknowledgements:
- Lecture prepared by Professor Anantha Chandrakasan
- Curt Schurgers
Cost of Transistor

Gordon Moore, Keynote Presentation at ISSCC 2003
In 1965, Gordon Moore was preparing a speech and made a memorable observation. When he started to graph data about the growth in memory chip performance, he realized there was a striking trend. Each new chip contained roughly twice as much capacity as its predecessor, and each chip was released within 18-24 months of the previous chip. If this trend continued, he reasoned, computing power would rise exponentially over relatively brief periods of time.
Evolution of Transistor Integration

Moore’s Law: transistor density doubles every 1.5 - 2 years
Layout 101

3-D Cross-Section

N-channel MOSFET

P-channel MOSFET

Circuit Representation

- Follow simple design rules (contract between process and circuit designers)
Itanium has 6 integer execution units like this

Die photograph of the Itanium integer datapath

Bit-slice Design Methodology

- Hand crafting the layout to achieve maximum clock rates (> 1Ghz)
- Exploits regularity in datapath structure to optimize interconnects
The ASIC Approach

Most Common Design Approach for Designs up to 500Mhz Clock Rates
Standard Cell Example

Each library cell (FF, NAND, NOR, INV, etc.) and the variations on size (strength of the gate) is fully characterized across temperature, loading, etc.

<table>
<thead>
<tr>
<th>Path</th>
<th>1.2V - 125°C</th>
<th>1.6V - 40°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>In1→tpLH</td>
<td>0.073+7.98C+0.317T</td>
<td>0.020+2.73C+0.253T</td>
</tr>
<tr>
<td>In1→tpHL</td>
<td>0.069+8.43C+0.364T</td>
<td>0.018+2.14C+0.292T</td>
</tr>
<tr>
<td>In2→tpLH</td>
<td>0.101+7.97C+0.318T</td>
<td>0.026+2.38C+0.255T</td>
</tr>
<tr>
<td>In2→tpHL</td>
<td>0.097+8.42C+0.325T</td>
<td>0.023+2.14C+0.269T</td>
</tr>
<tr>
<td>In3→tpLH</td>
<td>0.120+8.00C+0.318T</td>
<td>0.031+2.37C+0.258T</td>
</tr>
<tr>
<td>In3→tpHL</td>
<td>0.110+8.41C+0.280T</td>
<td>0.027+2.15C+0.223T</td>
</tr>
</tbody>
</table>

3-input NAND cell (from ST Microelectronics):
C = Load capacitance
T = input rise/fall time

Power Supply Line (V_DD)  Delay in (ns)!!
Ground Supply Line (GND)
2-level metal technology

Current Day Technology

- With limited interconnect layers, dedicated routing channels between rows of standard cells are needed
- Width of the cell allowed to vary to accommodate complexity
- Interconnect plays a significant role in speed of a digital circuit
module adder64 (a, b, sum);
    input [63:0] a, b;
    output [63:0] sum;
    assign sum = a + b;
endmodule
256×32 (or 8192 bit) SRAM Generated by hard-macro module generator

- Generate highly regular structures (entire memories, multipliers, etc.) with a few lines of code
- Verilog models for memories automatically generated based on size
For 1Ghz clock, skew budget is 100ps. Variations along different paths arise from:

- **Device**: $V_T$, W/L, etc.
- **Environment**: $V_{DD}$, °C
- **Interconnect**: dielectric thickness variation
Analog Circuits: Clock Frequency Multiplication (Phase Locked Loop)

- **VCO**: produces high frequency square wave
- **Divider**: divides down VCO frequency
- **PFD**: compares phase of ref and div
- **Loop filter**: extracts phase error information

Used widely in digital systems for clock synthesis (a standard IP block in most ASIC flows)

**Courtesy M. Perrott**

Intel 486, 50Mhz

F_{out} = N \cdot F_{ref}
Scan Testing

Idea: have a mode in which all registers are chained into one giant shift register which can be loaded/read-out bit serially. Test remaining (combinational) logic by:

1. In “test” mode, shift in new values for all register bits thus setting up the inputs to the combinational logic.
2. Clock the circuit once in “normal” mode, latching the outputs of the combinational logic back into the registers.
3. In “test” mode, shift out the values of all register bits and compare against expected results.

Courtesy of C. Terman and IEEE Press
There are a large number of implementations of the same functionality
- These implementations present a different point in the area-time-power design space
- Behavioral transformations allow exploring the design space a high-level

**Optimization metrics:**

1. **Area** of the design
2. **Throughput** or sample time $T_S$
3. **Latency**: clock cycles between the input and associated output change
4. **Power** consumption
5. **Energy** of executing a task
6. ...
**Fixed-Coefficient Multiplication**

### Conventional Multiplication

\[ Z = X \cdot Y \]

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>X3</td>
<td>Y3</td>
<td>Z7</td>
</tr>
<tr>
<td>X2</td>
<td>Y2</td>
<td>Z6</td>
</tr>
<tr>
<td>X1</td>
<td>Y1</td>
<td>Z5</td>
</tr>
<tr>
<td>X0</td>
<td>Y0</td>
<td>Z4</td>
</tr>
</tbody>
</table>

### Constant multiplication (become hardwired shifts and adds)

\[ Z = X \cdot (1001)_2 \]

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>X3</td>
<td>Y3</td>
<td>Z7</td>
</tr>
<tr>
<td>X2</td>
<td>Y2</td>
<td>Z6</td>
</tr>
<tr>
<td>X1</td>
<td>Y1</td>
<td>Z5</td>
</tr>
<tr>
<td>X0</td>
<td>Y0</td>
<td>Z4</td>
</tr>
</tbody>
</table>

\[ Y = (1001)_2 = 2^3 + 2^0 \]

![Diagram of shifts using wiring](attachment:image.png)
Canonical signed digit representation is used to increase the number of zeros. It uses digits {-1, 0, 1} instead of only {0, 1}.

Iterative encoding: replace string of consecutive 1’s

\[ 2^{N-2} + \ldots + 2^1 + 2^0 \]

\[ 2^{N-1} - 2^0 \]

Worst case CSD has 50% non zero bits
Algebraic Transformations

Commutativity

\[ A + B = B + A \]

Distributivity

\[ (A + B) C = AB + BC \]

Associativity

\[ (A + B) + C = A + (B+C) \]

Common sub-expressions
Transforms for Efficient Resource Utilization

Time multiplexing: mapped to 3 multipliers and 3 adders

Distributivity

Reduce number of operators to 2 multipliers and 2 adders
A Very Useful Transform: Retiming

Retiming is the action of moving delay around in the systems

- Delays have to be moved from ALL inputs to ALL outputs or vice versa

**Cutset retiming:** A cutset intersects the edges, such that this would result in two disjoint partitions of these edges being cut. To retime, delays are moved from the ingoing to the outgoing edges or vice versa.

**Benefits of retiming:**
- Modify critical path delay
- Reduce total number of registers
Retiming Example: FIR Filter

Symbol for multiplication

\[ y(n) = h(n) \otimes x(n) = \sum_{i=0}^{K} x(n-i) \cdot h(i) \]

Direct form

Transposed form

Note: here we use a first cut analysis that assumes the delay of a chain of operators is the sum of their individual delays. This is not accurate.
Pipelining, Just Another Transformation
(Pipelining = Adding Delays + Retiming)

Contrary to retiming, pipelining adds extra registers to the system.

How to pipeline:
1. Add extra registers at all inputs
2. Retime
The Power of Transforms: Lookahead

\[ y(n) = x(n) + A y(n-1) \]

Try pipelining this structure

How about pipelining this structure!

\[ x(n) \rightarrow + \rightarrow y(n) \]
\[ D \rightarrow A \rightarrow 2D \]

\[ y(n) = x(n) + A[x(n-1) + A y(n-2)] \]

loop unrolling

\[ D \rightarrow + \rightarrow A \rightarrow 2D \]

precomputed
Key Concern in Modern VLSI: Variations!

Technology Node (nm)

Random Dopant Fluctuations

Deterministic design techniques inadequate in the future

With 100b transistors, 1b unusable (variations)

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Temp Variation & Hot spots

Mean Number of Dopant Atoms

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay

Probability

Due to variations in: $V_{dd}$, $V_t$, and Temp

Delay
Trends: “Chip in a Day”
(Matlab/Simulink to Silicon…)

Map algorithms directly to silicon - bypass writing Verilog!

Courtesy of R. Brodersen
Fingerprinting is a technique to deter people from illegally redistributing legally obtained IP by enabling the author of the IP to uniquely identify the original buyer of the resold copy.

The essence of the watermarking approach is to encode the author's signature. The selection, encoding, and embedding of the signature must result in minimal performance and storage overhead.

same functionality, same area, same performance
watermark of 4768 bits embedded
(courtesy of G. Qu, M. Potkonjak)
Evolution of Transistor Integration

Moore’s Law: transistor density doubles every 1.5 - 2 years
Processor performance follows Moore’s Law - doubles every 2 years
Clock Frequency of Microprocessors

Doubles every 2 years

S. Borkar
Power Consumption Trends

Power per gate goes down but total power ...
Six layers of Cu metallization

- Lower layers are finer and are used for “local” interconnection between cells
- Middle layers are wider and are used for global interconnection between blocks
- Upper layers are wider and are used for clocks, ground and power distribution
- Oxide is the Inter Metal Dielectric (etched)
Interconnect Metallization

SEM cross-section

Copper layer 1
Copper layer 2
Copper layer 3
Copper layer 4
Copper layer 5
Copper layer 6

Local tungsten interconnect