L15: VLSI Integration and Performance Transformations

Acknowledgement:

Materials in this lecture are courtesy of the following sources and are used with permission.

Curt Schurgers

Follow simple design rules (contract between process and circuit designers)
Custom Design/Layout

- Handcrafting the layout to achieve maximum clock rates (> 1GHz)
- Exploits regularity in datapath structure to optimize interconnects

Die photograph of the Itanium integer datapath

Itanium has 6 integer execution units like this

Bit-slice Design Methodology

From register files / Cache / Bypass

To register files / Cache

Adder stage 1
Adder stage 2
Adder stage 3
Sum Select
Shifter
Wiring

LU: Logical Unit

SUMSEL
REG
SUMGEN + LU
LU
CARRYGEN

From  register files / Cache / Bypass

Multiplexers

Die photograph of the
Itanium integer datapath


The ASIC Approach

Most Common Design Approach for Designs up to 500Mhz Clock Rates
Each library cell (FF, NAND, NOR, INV, etc.) and the variations on size (strength of the gate) is fully characterized across temperature, loading, etc.
Standard Cell Layout Methodology

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

Cell-structure hidden under interconnect layers
module adder64 (a, b, sum);
  input [63:0] a, b;
  output [63:0] sum;
  assign sum = a + b;
endmodule
Macro Modules

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
Clock Distribution

For 1Ghz clock, skew budget is 100ps. Variations along different paths arise from:

- **Device:** \( V_T, W/L, \) etc.
- **Environment:** \( V_{DD}, ^\circ C \)
- **Interconnect:** dielectric thickness variation

IBM Clock Routing

Clock skew

Image removed due to copyright restrictions.
The IR-drop problem causes internal power supply voltage to be less than the external source.

Used with permission.
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 Michael Perrott. Used with permission.
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.

![Diagram of Scan Testing](image-url)
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

\[ Z = X \cdot Y \]

\[
\begin{array}{cccc}
X_3 & X_2 & X_1 & X_0 \\
Y_3 & Y_2 & Y_1 & Y_0 \\
X_3 \cdot Y_0 & X_2 \cdot Y_0 & X_1 \cdot Y_0 & X_0 \cdot Y_0 \\
X_3 \cdot Y_1 & X_2 \cdot Y_1 & X_1 \cdot Y_1 & X_0 \cdot Y_1 \\
X_3 \cdot Y_2 & X_2 \cdot Y_2 & X_1 \cdot Y_2 & X_0 \cdot Y_2 \\
X_3 \cdot Y_3 & X_2 \cdot Y_3 & X_1 \cdot Y_3 & X_0 \cdot Y_3 \\
\end{array}
\]

\[
\begin{array}{cccc}
Z_7 & Z_6 & Z_5 & Z_4 \\
Z_3 & Z_2 & Z_1 & Z_0 \\
\end{array}
\]

Constant multiplication (become hardwired shifts and adds)

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

\[
\begin{array}{cccc}
X_3 & X_2 & X_1 & X_0 \\
X_3 & X_2 & X_1 & X_0 \\
1 & 0 & 0 & 1 \\
\end{array}
\]

\[
\begin{array}{cccc}
Z_7 & Z_6 & Z_5 & Z_4 \\
Z_3 & Z_2 & Z_1 & Z_0 \\
\end{array}
\]

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

\[ X \rightarrow \leftarrow + \rightarrow Z \]

shifts using wiring
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

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

Worst case CSD has 50% non zero bits

\[ 01101111 \]

\[ X \]

\[ 10010001 \]

\[ Z \]

\[ \text{Shift translates to re-wiring} \]
Algebraic Transformations

Commutativity

A + B = B + A

(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
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

Direct form

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

Transposed form

T_{clk} = 22 \text{ ns}

T_{clk} = 14 \text{ ns}

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

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

How about pipelining this structure!

\[ \text{loop unrolling} \]

\[ \text{distributivity} \]

\[ \text{associativity} \]

\[ \text{retiming} \]

\[ \text{precomputed} \]
Key Concern in Modern VLSI: Variations!

With 100b transistors, 1b unusable (variations)

Deterministic design techniques inadequate in the future

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

Path Delay

Temp Variation & Hot spots

Courtesy of Shekhar Y. Borkar (Intel). Used with permission.
Trends: “Chip in a Day”
(Matlab/Simulink to Silicon…)

Map algorithms directly to silicon - bypass writing Verilog!
(Courtesy of R. Brodersen. Used with permission.)
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.