1. Introduction

Dr. Paul D. Franzon

Outline
1. Types of Custom Digital Circuits
2. Digital circuit characteristics, design objectives & issues
3. Tradeoffs
4. Transistor device trends
5. Interconnect trends

References
• D&P, CH. 1, 2
Module Objectives and Motivation

Learning Objectives

- Understand different scenarios in which full custom digital design is used, why, and what are the major design drivers
- Be able to describe typical design objectives for digital circuits
- Be able to describe on-going trends in technology (transistor, interconnect, packaging)

Motivation

- Describes when and why full custom digital circuit design is still used
- Puts overall course in the “big picture” from devices through to systems
Why Custom Digital Circuits?

Custom digital circuits are used in applications where their performance premium justifies their high design cost.

Standard Cells

- Why custom design?

- Design Objectives:

  - Area Yield
  - Speed
  - Robustness
  - Min. Variation in timing
  - Power Consumption
  - Utility as synthesis target

  - Right mix of cell types and drive strengths
## Standard Cells

- **Design Objectives:**

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robustness</td>
<td>Good noise immunity despite wide variations in input drive and output load</td>
</tr>
<tr>
<td>Area</td>
<td>Good tight layout, that ...</td>
</tr>
<tr>
<td>Yield</td>
<td>Does not compromise yield (this is getting harder)</td>
</tr>
<tr>
<td>Timing</td>
<td>Fast but ...</td>
</tr>
<tr>
<td>Power Consumption</td>
<td>Several cell families: High-speed, low-power, low-standby-power, different Vdd's</td>
</tr>
<tr>
<td>Variations in timing</td>
<td>Want tight min/max timing (this is getting harder)</td>
</tr>
<tr>
<td>Good mix of cell types and strengths</td>
<td>Function and drive strength (different strengths for different loads)</td>
</tr>
</tbody>
</table>
Custom Circuits

Microprocessor Datapaths

- Why?
  - Today:
  - Tomorrow:

- Design Objectives:

  \[
  \text{Area} \quad \{ \quad \text{So as to maximize \# of yielded die per wafer} \\
  \text{Yield} \quad \} \quad = \quad \# \text{ of die sites per wafer (roughly wafer area/die area)}^1 \quad * \quad \text{die yield}
  \]

Notes: 1. Not exactly equal due to edge wastage

©2011, Dr. Paul D. Franzon, www.ece.ncsu.edu/erl/faculty/paulf.html
... Custom Circuits

... Microprocessor design objectives

Speed: Sufficient to maximize performance/chip or performance/$. Used to be “maximize Instructions Per Cycle * t_cycle”.

Power: Maximize performance within power budget. Power budget determined by temperature control, cooling constraints, battery life, and/or available wall power. In era of multi-core, will lead to SLOWER clock rates.

Resiliency: Ability to complete computation despite permanent or temporary noise/timing upset. Main component today: Transient single bit errors due to noise.
... Custom Circuits

Embedded and Standalone Memory Chips

- Why?

- Types:
  - Stand-alone:
    - Commodity: DRAM, GRAM, SRAM, NVRAM
    - Specialized: CAM, Specialized DRAM (e.g. Low latency, XDR)
  - Embedded:
    - SRAM, cache, sometimes DRAM, NVRAM

1. Non Volatile, e.g. Flash
2. Extended Data Rate, e.g. As used in Playstation3
... Custom Circuits

... Memory Circuits Design Objectives:

Cost per bit: Critical in commodity RAMs. Layout & design rules, use of spare columns/rows are highly optimized.

Latency and Bandwidth: Often not highly optimized, except in specialty memories. Often has to meet a standard, e.g. DDR

Power consumption: Active, and Standby. Usually not optimized but becoming more important.

Retention time: DRAM and NVRAM

Soft-Error Rate: After application of Error Correction Code

Failure Rate: Hard failure rate
… Custom Circuits

Field Programmable Gate Arrays (FPGAs)

- Why?

- Design Objectives:

  Density: Want to maximize effective density (#gates/area after programming). Highly specialized layout and design rules.

  Yield. High end parts command a premium, so yield can be sacrificed some.

  Speed. Valuable but hard to do well due to programmable interconnect.
... Custom Circuits

...FPGA Design Objectives:

Power Consumption: becoming more important, but usually not optimized for.

Reliability: Noise immunity and control.
I/O, especially high-speed I/O

- Why?

- Design Objectives:
  
  Data Rate: I/O speed per line
  
  Signal Integrity: Bit Error Rate
  
  Power: Energy per bit. Clock and Data Recovery is a significant portion of this.
  
  Area: To fit within available form factor.
  
  Largely dominated by synthesized overhead functions.
Generalized Design Objectives

- **Performance**
  - Delay or throughput

- **Power Consumption**
  - Minimize or meet a constraint

- **Yield maximization**
  - % of product without logic defect
  - % of product that meets spec, and/or
  - % of product in each speed-bin

- **Reliability**
  - Against Failure (bit-error – hard or soft)
    - Failure due to process related issues
    - Failure due to noise

- **Area**

- **Design Cost**
Example

Clock

A

B

n1

n2

n3

Out

Speed Determined By: n1, n2, n3

Noise Immunity improved by: p2, n4/p4

High power consumption

©2011, Dr. Paul D. Franzon, www.ece.ncsu.edu/erl/faculty/paulf.html
Performance

Combinational Logic

- Circuit Delay
- How to measure?
  - Need a fixed reference point so that:
    \[ t_{P12} = t_{P1} + t_{P2} \]
  - For static CMOS, met by using \( 0.5 \times V_{swing} \) as a reference point

\[ t_{PH} = \text{Propagation Delay for a low to high transition on OUTPUT} \]

\[ t_{TL} = \text{Transition Time Low to High} \]

\[ t_{r} = \text{Rise Time} \]

\[ t_{f} = \text{Fall Time} \]
**Performance**

Flip-flops and related sampled data circuits

- Throughput = Maximum bit-rate consistent with safe operation

![Diagram showing bit period, skew, jitter, aperture time, setup time, hold time, Rx sensitivity, and Rx offset voltage.](image)
**Power Consumption**

**Power** = \( \Sigma \) all circuits

- \( I_{DC} \) * Vdd +
- Dynamic power
  - \( N_{\text{switch}} \) * \( f \) * \( V_{\text{swing}}^2 \) * \( C_{\text{load}} \) +
  - crowbar current +
- \( I_{\text{leakage}} \) * Vdd:
  - Sub-threshold
  - Reverse leakage
  - Parasitic bipolar (if turned on)
  - Gate leakage
  - Band to band tunneling

\( N_{\text{switch}} \) = % of cycles in which node toggles 010 or 101

\( f \) = clock frequency

\( V_{\text{swing}} \) = voltage swing on Cload
Sidebar: Data Center Power Consumption

10 MW

Power Delivery → 2 MW → Computation → Cooling (AC)

4.7 MW  → Memory 1 MW  → Storage 0.3 MW

©2011, Dr. Paul D. Franzon, www.ece.ncsu.edu/erl/faculty/paulf.html
It's not just Computation

<table>
<thead>
<tr>
<th>Conversion Step</th>
<th>Efficiency</th>
<th>Delivered</th>
<th>Dissipated</th>
<th>Delivered</th>
<th>Dissipated</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC In</td>
<td></td>
<td>1.00W</td>
<td></td>
<td>2.06W</td>
<td></td>
</tr>
<tr>
<td>Uninterruptible Power Supply (UPS)</td>
<td>88%</td>
<td>0.88W</td>
<td>0.12W</td>
<td>1.81W</td>
<td>0.25W</td>
</tr>
<tr>
<td>Power Distribution Unit (PDU)</td>
<td>93%</td>
<td>0.82W</td>
<td>0.06W</td>
<td>1.69W</td>
<td>0.13W</td>
</tr>
<tr>
<td>In Rack Power Supply Unit (PSU)</td>
<td>79%</td>
<td>0.65W</td>
<td>0.17W</td>
<td>1.33W</td>
<td>0.35W</td>
</tr>
<tr>
<td>On Board Voltage Regulator (VR)</td>
<td>75%</td>
<td>0.49W</td>
<td>0.16W</td>
<td>1.00W</td>
<td>0.33W</td>
</tr>
<tr>
<td>Target Logic</td>
<td></td>
<td>0.49W</td>
<td>0.49W</td>
<td>1.00W</td>
<td>1.00W</td>
</tr>
</tbody>
</table>

Table 4.1: Power distribution losses in a typical data center. Source: Intel

40% - 50% lost just in power delivery!
## Energy per Operation - samples

<table>
<thead>
<tr>
<th>Component</th>
<th>Energy per Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDR3</td>
<td>4.8 nJ/word</td>
</tr>
<tr>
<td>Optimized DRAM core</td>
<td>128 pJ/word</td>
</tr>
<tr>
<td>MIPS 64 core*</td>
<td>0.4 nJ/cycle</td>
</tr>
<tr>
<td>45 nm 0.8 V FPU</td>
<td>38 pJ/Op</td>
</tr>
<tr>
<td>SERDES I/O</td>
<td>1.9 nJ/Word</td>
</tr>
<tr>
<td>20 mV I/O</td>
<td>128 pJ/Word</td>
</tr>
<tr>
<td>LPDDR2</td>
<td>512 pJ/Word</td>
</tr>
<tr>
<td>On-chip/mm</td>
<td>7 pJ/Word</td>
</tr>
<tr>
<td>TSV I/O (ESD)</td>
<td>7 pJ/Word</td>
</tr>
<tr>
<td>TSV I/O (no ESD)</td>
<td>2 pJ/Word</td>
</tr>
</tbody>
</table>

Without better solutions, memory and interconnect power will dominate computing.

* At 90 nm. Includes 40 kB cache, no FPU
Memories

DDR3 energy budget
- Total energy/access
- ~ 550 pJ / bit

Energy to read a DRAM cell
- ~ 30 fJ

Efficiency = 1/20,000
- i.e. 0.005%
Computation

FP Multiply in 45 nm technology:
- 45 pJ/FLOP

pJ/cycle in a 45 nm RISC processor
- 1500 pJ/cycle

Computation is 4% efficient (at best – ignores “overhead” instructions)
Putting it all together

For every 1 MJ of energy going into a data center
- 20 kJ goes into core computation (2%)
- 5 J goes into accessing memory cells

The rest is “wasted”

On current scaling trends, a 1000 PetaFlop machine would require 80 MW of power just for the computer
Leakage Sources

(Off current)
Sub-threshold $\propto e^{-Vt/(S/\ln 10)}$
(increased by Drain Induced Barrier Lowering)

Reverse Diode $\propto e^{-V_{bias}/kT}$

Gate Oxide Tunneling

Fig. 7. Calculated (lines) and experimental (dots) results for tunnel currents from inversion layers through thin oxides. Adapted from Lo et al. [24].

Fig. 9. Plot of band-to-band $p$-$n$ junction tunneling current versus electric field for 1 V reverse bias. Adapted from [27].

©2011, Dr. Paul D. Franzon, www.ece.ncsu.edu/erl/faculty/paulf
Yield Maximization

- Transistor Parameters as process control gets harder:
  - Threshold voltage
  - Gate $\beta$
  - Interconnect R, C

- Types
  - Inter-die
  - Intra-die

Fig. 2: Experimental results for typical threshold voltage distributions of small and big transistors.

Source: Gyvez 2004

Source: Kong 2004
Yield Maximization

Issues:

- Statistical design across realistic process corners
- Increased variation between on-chip circuits
- Increased variation between chips
- % of product falling within required timing margins in fabrication
- Or % of product in each speed bin
Reliability

Hard failures:

- i.e. Permanent failure or degradation
- E.g. Electromigration; oxide punch-through due to electro-static discharge; hot-electron effects
- Managed through design rules + some structures
- Measured as FITs or “Failures in Time” or in $10^9$ (billion) hours
  - Includes hard (permanent) and soft failures
  - Related to MTBF (Mean Time Between Failures)

Soft Failures:

- Incorrect logic state due to noise
- Incorrect sampling point due to noise-induced jitter
- Incorrect logic state due to alpha-particle (package radiation)
- Measured as a Single Event Upset/Transition (SAU/T) rate (SER)
Digital Circuits

Unlike Analog circuits, digital circuits reject noise.

Logic restoration

Permits limitless scaling with correct “function”.

How much noise can a circuit reject?
How Digital Ckts reject noise

Via non-linear gain:
DC Noise Margin

- $V_{OL}$ $V_{output\_low}$
- $V_{OH}$ $V_{output\_high}$
- $V_{IH}$ $V_{input\_high}$
- $V_{IL}$ $V_{input\_low}$

Define at points where transfer characteristic has unity slope.

- $NM_H = V_{OH} - V_{IH}$
- $NM_L = V_{IL} - V_{OL}$
**Why This Definition?**

What happens in each case?

For Gain < 1:
- Noise pulse attenuated

For Gain > 1:
- Noise pulse amplified!
  - Potential Logic Error!
**AC Noise Margin**

Circuit can often reject larger narrow pulses

Fail = incorrect bit transition or output dips below VIH, depending on logic type

Source: Ding 2004

(0.18 μm dynamic logic)
Noise Sources

Power Supply Noise

- Noise on local ground and power nodes on chip due to IR drop and \( \frac{L}{dt} \)
- E.g.:

\[
V = \frac{L}{dt}
\]

\[\text{Power}\]

\[\text{Ground}\]
Other Noise Sources

Crosstalk noise
  • Between wires
  • Between wires and device terminals

“Charge sharing noise”
  • In dynamic logic

Reflection noise
  • In transmission line structures

Substrate noise
  • Power and ground noise transmitted through substrate

Device Noise
  • Flicker noise, 1/f noise, shot noise
Tradeoffs

These objectives can not be simultaneously optimized e.g. Threshold voltage as affected by body bias

\[ \Delta V_T = \gamma \left( \sqrt{2 \psi_B} - \sqrt{2 \psi_B - V_{BS}} \right) \]

NMOS $I_{DS}$-$V_{GS}$ varying $V_{BS}$

Lower $Vt$: Faster

Higher $Vt$: Lower Leakage, Better Noise Margin
Tradeoffs

To capture fundamental improvements, rather than just different points in the tradeoff curve, some metrics are commonly used:

- Area.delay product
  - As often they can be traded to give overall performance
- Power.delay product
- Energy.delay product (≡ power.delay²)
  - Less susceptible to “gaming” than power.delay = energy, as the latter can be minimized by just lowering Vdd.
**Power/Energy vs. Speed tradeoffs**

Generally, there is a tradeoff between power/energy-per-op and speed

- The most efficient processor is generally not at the fastest operating point
Transistor Device Trends

Bulk FET → Partially Depleted SOI → Fully Depleted SOI → Double Gate FET
**Bulk FET**

**Scaling issues:**

- Band-to-band tunneling increases as fields get higher
- Harder to fabricate low leakage gate oxide
- Impact of dopant variations
- Likely limit: Around 25 nm gate length

![3-D perspective plot of the dopant atoms in a 25-nm MOSFET. Darker dots are donors and lighter dots are acceptors. From [79].](image-url)
Silicon On Insulator

- Transistors “built” on top of oxide
- Eliminates several leakage paths and reduces capacitance by 30% - 50%
- Partially Depleted:
  - Channel does not fill silicon region
- Fully depleted:
  - Channel “fills” silicon region
- Floating Body effects, esp. in PDSOI
  - E.g. Hysterisis
  - Body ties
**Double Gate FET**

Or “FINFET”

- Excellent electrostatics
- Can scale to smaller sizes without excessive fields
- + all the advantages of SOI
- Scalable to ~ 8nm gate length
  - Leakage issues
- Not clear how to make at this point!
# Intl. Tech Roadmap for Semiconductors (ITRS)

## Table 47a  High-performance Logic Technology Requirements—Near-term

<table>
<thead>
<tr>
<th>Year of Production</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology Node</td>
<td>hp90</td>
<td>hp90</td>
<td>hp90</td>
<td>hp90</td>
<td>hp90</td>
<td>hp90</td>
<td>hp90</td>
</tr>
<tr>
<td>DRAM 1/2 Pitch (nm)</td>
<td>100</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td>65</td>
<td>57</td>
<td>50</td>
</tr>
<tr>
<td>MPU/ASIC Metal 1 (M1) 1/2 Pitch (nm)</td>
<td>120</td>
<td>107</td>
<td>95</td>
<td>85</td>
<td>76</td>
<td>67</td>
<td>60</td>
</tr>
<tr>
<td>MPU/ASIC 1/2 Pitch (nm)</td>
<td>107</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td>65</td>
<td>57</td>
<td>50</td>
</tr>
<tr>
<td>MPU Printed Gate Length (nm)</td>
<td>65</td>
<td>53</td>
<td>45</td>
<td>40</td>
<td>35</td>
<td>32</td>
<td>28</td>
</tr>
<tr>
<td>MPU Physical Gate Length (nm)</td>
<td>45</td>
<td>37</td>
<td>32</td>
<td>28</td>
<td>25</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td>Physical gate length high-performance (HP) (nm) [1]</td>
<td>45</td>
<td>37</td>
<td>32</td>
<td>28</td>
<td>25</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td>EOT: equivalent oxide thickness (physical) for high-performance (nm) [2]</td>
<td>1.3</td>
<td>1.2</td>
<td>1.1</td>
<td>1.0</td>
<td>0.9</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Saturation threshold voltage (V) [7]</td>
<td>0.21</td>
<td>0.20</td>
<td>0.20</td>
<td>0.21</td>
<td>0.18</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td>Nominal high-performance NMOS sub-threshold leakage current, I_d,leak (at 25°C) (µA/µm) [8]</td>
<td>0.03</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Nominal high-performance NMOS saturation drive current, I_d,sat (at V_d, at 25°C) (mA/µm) [9]</td>
<td>980</td>
<td>1110</td>
<td>1090</td>
<td>1170</td>
<td>1510</td>
<td>1530</td>
<td>1590</td>
</tr>
<tr>
<td>Required &quot;mobility/transconductance improvement&quot; factor [10]</td>
<td>1.0</td>
<td>1.3</td>
<td>1.3</td>
<td>1.4</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>High-performance NMOS intrinsic delay, τ = C_{gate} * V_{dd} / I_{d,sat} (ns) [16]</td>
<td>1.20</td>
<td>0.95</td>
<td>0.86</td>
<td>0.75</td>
<td>0.64</td>
<td>0.54</td>
<td>0.48</td>
</tr>
<tr>
<td>Relative NMOS intrinsic switching speed, 1/τ, normalized to 2003 [17]</td>
<td>1.00</td>
<td>1.26</td>
<td>1.39</td>
<td>1.60</td>
<td>1.86</td>
<td>2.20</td>
<td>2.49</td>
</tr>
<tr>
<td>NMOSFET power-delay product (µm) [19]</td>
<td>1.41E-15</td>
<td>1.27E-15</td>
<td>1.03E-15</td>
<td>9.66E-16</td>
<td>1.07E-15</td>
<td>8.33E-16</td>
<td>7.66E-16</td>
</tr>
<tr>
<td>NMOSFET static power dissipation due to drain and gate leakage (W/µm) [20]</td>
<td>3.96E-07</td>
<td>6.60E-07</td>
<td>6.05E-07</td>
<td>6.05E-07</td>
<td>8.47E-07</td>
<td>7.70E-07</td>
<td>7.70E-07</td>
</tr>
</tbody>
</table>

©2011, Dr. Paul D. Franzon, www.ece.ncsu.edu/erl/faculty/paulf.html
**Chip Level Interconnect**

Increasing impact on design performance and power

Source: Sakurai 00
Chip Level Interconnect

Why?

- Interconnect Delay $\sim$ RC
- As dimensions scale, R gets worse while C stays the same

- So instead, try to keep RC constant
- Meanwhile transistor C goes down
Chip Level Interconnect

Trends:

- Al $\rightarrow$ Cu
- SiO$_2$ $\rightarrow$ low-k dielectric
- Increasing layer count
- Interconnect delay and power becoming a greater % of path delay and chip power
- Increasing variability
- Increasing impact of “inductive effects”
System Level Interconnect

Package – Board – Connector – Backplane
I/O Scaling

Source: Intel
3DIC with Through Silicon Vias

Technology set:

- Insulator deposition
- Deep RIE etching
- TSV (X ray image)
- Top and back bump formation
- High aspect Cu plating
- Underfill
- Wafer Thinning

S. Denda, Nagano Prefectural Institute of Technology.
©2011, Dr. Paul D. Franzon, www.ece.ncsu.edu/erl/faculty/paulf.html
3-Tier 3DIC Cross-Section

Second DARPA Multiproject Run (3DM2)

Two Digital & One RF 180-nm 1.5V FDSOI CMOS Tiers

3DM2 Process Highlights
- 11 metal interconnect levels
- 1.75-μm 3D via tier interconnect
- Stacked 3D vias allowed
- Tier-2 back-metal/back-via process
- 2-μm-thick RF back metal
- Tier-3 W gate shunt
- Tier-3 silicide block

©2011, Dr. Paul D. Franzon, www.ece.ncsu.edu/erl/faculty/paulf.html
System Level Interconnect

Trends:

- Multi Mbps $\Rightarrow$ multi-Gbps on every pin and wire
- Cost-effective structures and materials have little room for improvement
  - Standard high end of low-cost: 100 mil thick glass-epoxy FR4 board
  - Need new low-cost materials or better circuit ideas
- Noise management becoming very complex
- Chip I/O bandwidth scales at a slower rate than on-chip performance and bandwidth
Review Questions

What applications did we identify that justify full custom digital design?

What are some of the design objectives common to these applications?
Review Questions

What is the main source of static leakage power?

Why does leakage get worse with technology scaling?

How is DC Noise margin defined?
Review Questions

Why does interconnect power increasingly dominate with technology scaling?

What is the impact of process control limitations/difficulties on circuit design with technology scaling?

Will bulk CMOS transistors ever be replaced?