The VLSI Interconnect Challenge

Avinoam Kolodny
Electrical Engineering Department
Technion – Israel Institute of Technology

VLSI Challenges

• System complexity
• Performance
• Tolerance to digital noise and faults
• More challenges...

The Dominant Challenge is
Power dissipation!

Interconnect is the crux of the problem

"Old view" of VLSI:
• Speed and power are dominated by logic gates
• Wires are “ideal”

"New view":
• Logic is fast and virtually free
• Speed and power are limited by wires
Outline of this talk

- Background of the VLSI interconnect challenge
- Implications for energy-efficient computing
- Research directions
Chips are Like Cities: Complexity is Shown in Connectivity

- In each generation of technology:
  - More transistors
  - More interconnect wires

Technology Scaling: Faster Transistors, Slower Wires

Note: Distances across a full chip are virtually constant: "Global wires" do not scale!

Technology Scaling: Faster Transistors, Slower Wires

Note: Distances across a full chip are virtually constant: "Global wires" do not scale!

Non-uniform scaling

1. "Old wires": ground capacitance is dominant
2. "New wires": line-to-line capacitance is dominant

Trying to Keep Wire Resistance in Check Leads to Larger Capacitances

2) Using bigger drivers, inserting repeaters

More Capacitance → More Power!
If Bits Were Cars….

Energy per task = \sum_{i} (\# \text{messages}) \cdot \frac{C_i V_i^2}{2}

The Nature of Design for Low-Power

- No critical root cause
  - Because power is cumulative
  - Need power-saving efforts at all levels!

Types of Improvement

- 1 Reduce waste of energy
- 2 (Optimal) Tradeoff power with delay (or other metric)
- 3 Change the algorithm or computational task

Wire layout optimization:

Wire Widths and Spaces in a Wire Bundle

A is a fixed constraint
Wire layout optimization: Finding Optimal Wire Widths and Spaces under Delay Constraints

Wire-Spacing is precious real-estate!


Optimal Ordering Theorem for Power

- Given an interconnect channel with wires of uniform width $W$, use ‘Symmetric Hill’ ordering according to activity factors of the signals

Activity Factor


Circuit optimization: Optimal Power-Delay Tradeoff for Logic Paths

For 2.5% delay increase, get $2x2.5=30\%$ energy reduction!

For 20% delay increase, get $2x20=40\%$ energy reduction


3-D Integrated Circuit Technology

Reduce interconnect distances by building vertically

Through silicon vias (TSV)

Most Power Savings Can be Made at High Abstraction Levels

Pollack's Rule on Power Efficiency of Uniprocessors

- Power-efficiency requires many parallel local computations
  - Chip Multi Processors (CMP)
  - Thread-Level Parallelism (TLP)

\[
\text{Performance} = \frac{1}{\text{Area}}
\]

\[
\text{Power} = a_2(\text{Area})
\]

\[
\text{Power} = a_2\sqrt{\text{Area}}
\]

Pollack, MICRO 1999
Borkar, DAC 2007

Processor Architectures: Uni-core, Symmetric multicore, Asymmetric (Heterogeneous)

Processor System Evolution to CMP

- Power = \(a_2(\text{Area})\)
- Performance = \(a_1\sqrt{\text{Area}}\) [Pollack]

Architecture optimization:

CPU
Cache
Dual Core
CPU1 Cache
CPU2 Cache
Quad Core
CPU1 Cache
CPU2 Cache
CPU3 Cache
CPU4 Cache

Chip Multi Processors
**Architecture optimization:**

**Asymmetric Multi-Core Performance**

ACCMP Performance Vs. Power  (25% serial code)

- Symmetric Upper Bound
- Asymmetric ($a=1, \beta=4$)
- Asymmetric ($a=0.33, \beta=6$)

**Future VLSI systems**

- Classes of Replicated cores
  - Standard modules (Processors, Accelerators, Cache banks, ...)
  - Network on Chip (NoC)
- Power management
  - Different clocks
  - Different operating voltages
  - "dark silicon"


**If a chip is like a city, Network-on-Chip (NoC) is like a subway system**

- A new paradigm: Network instead of dedicated wires and buses
  - Inherently parallel
  - Efficient sharing of wires
- Scalable, cost effective bandwidth
Issues Addressed by NoC

1) Global wire design
   (delay, power, noise, scalability, reliability issues)
2) System integration productivity
   (key to modular design)
3) Multi-Core Processor Systems
   (key to power-efficient computing)

Interconnect-aware and NoC-aware Architectural Research

Accessing On-Chip Cache Banks through a NoC

Where to Store the Shared Data?

A small number of lines, shared by many processors, is accessed numerous times

What can be done better?
- Bring shared data closer to all processors
- Preserve vicinity of private data
This Has Been Addressed Before

Overview of Nahalal cache organization
Aerial view of Nahalal cooperative village

This Has Been Addressed Before

Overview of Nahalal cache organization
Aerial view of Nahalal cooperative village

Memory Bottleneck

CPU
Control Unit
Arithmetic/Logic Unit
Memory

Reducing Distances by Embedding Memory in Execution Units
**Memristor Devices**

Decrease resistance

\[ \text{Current} \rightarrow P_{\text{OFF}} \rightarrow P_{\text{ON}} \rightarrow \text{Voltage [V]} \rightarrow \text{Current [mA]} \]

\[ \text{ON} \rightarrow \text{OFF} \]

**Sea of Memory**

- Dense and fast

**Towards Memory-Intensive Machines**

**Throughput**

- Bandwidth

- "Memory-intensive" machines

- "Bandwidth Demons"

- Constant-throughput-curves

- Increase on-die-memory!

**Logic within the Memory**

Beyond von Neumann Architecture

- CPU
- Control Unit
- Arithmetic/Logic Unit
- Memory

- TP1
- TP2
- TP3
- TP4
Summary

- VLSI power is dominated by interconnect!
- New architectures are driven by interconnect distances/latencies/power