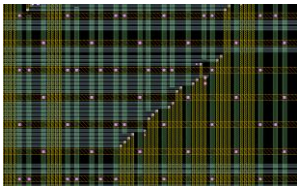
The VLSI Interconnect Challenge



Avinoam Kolodny

Electrical Engineering Department
Technion – Israel Institute of Technology

VLSI Challenges

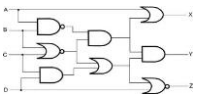
- System **complexity**
- **Performance**
- Tolerance to **digital noise** and **faults**
- More challenges…

The Dominant Challenge is

Power dissipation!

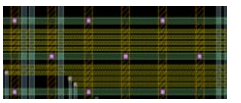
Interconnect
is the crux of the problem

Interconnect
is the crux of the problem



"Old view" of VLSI:

- **Speed and power are dominated by logic gates**
- **Wires are "ideal"**



"New view " :

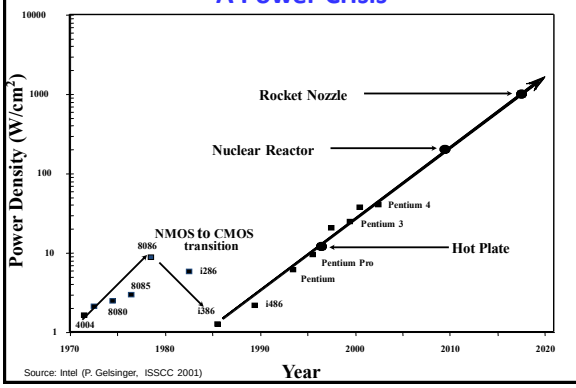
- **Logic is fast and virtually free**
- **Speed and power are limited by wires**

Outline of this talk

- Background of the VLSI interconnect challenge
- Implications for energy-efficient computing
- Research directions

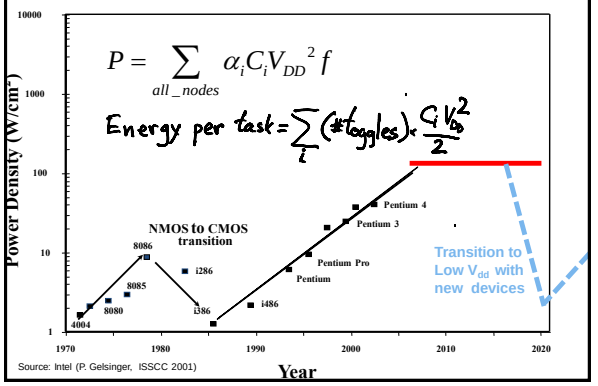
5

2001 - Extrapolation Towards A Power Crisis



Source: Intel (P. Gelsinger, ISSCC 2001)

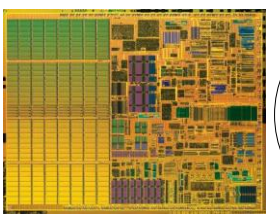
VLSI hits the power wall

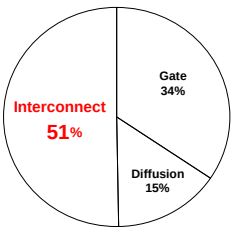


Source: Intel (P. Gelsinger, ISSCC 2001)

Interconnect Power: A case study - 2004

- Intel's Pentium-M, low-power microprocessor, 0.13 micron CMOS
- Bit-Transportation energy is larger than computation energy!



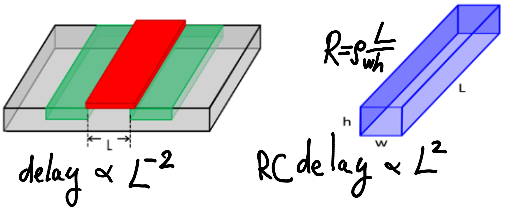


* N. Magen, A. Kolodny, U. Weiser and N. Shamir, "Interconnect-Related Energy dissipation in a Low-Power Microprocessor", Proc. SLIP, 2004.

Chips are Like Cities: Complexity is Shown in Connectivity

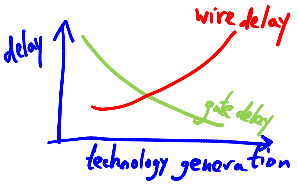
- In each generation of technology:
  - More transistors
  - More interconect wires

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

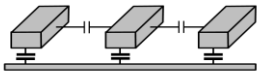


Note: Distances across a full chip are virtually constant:
"Global wires" do not scale!

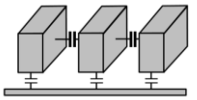
Trying to Keep Wire Resistance in Check Leads to Larger Capacitances

1) Non-uniform scaling

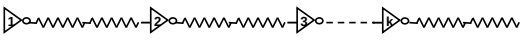
"Old wires": ground capacitance is dominant



"New wires": line-to-line capacitance is dominant



2) Using bigger drivers, Inserting Repeaters



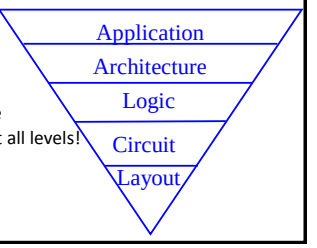
More Capacitance → More Power!

If Bits Were Cars….

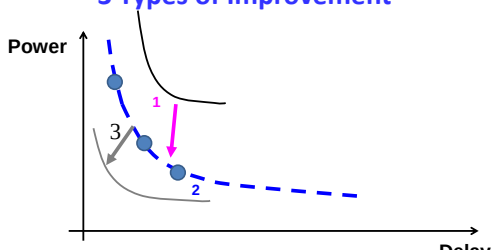
$$\text{Energy per task} = \sum_i (\#\text{toggles})_i \frac{C_i V_{DD}^2}{2}$$

The Nature of Design for Low-Power

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



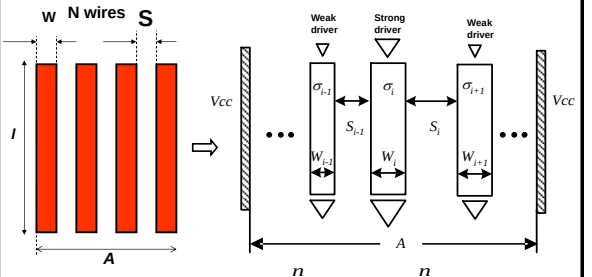
3 Types of Improvement



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

* Mark A. Horowitz, Vladimir Stojanovic, Borivoje Nikolic, Dejan Markovic, Robert W. Brodersen, "Methods for true power minimization," *iccad* , pp. 35-42, 2002.

Wire layout optimization:
Wire Widths and Spaces in a Wire Bundle

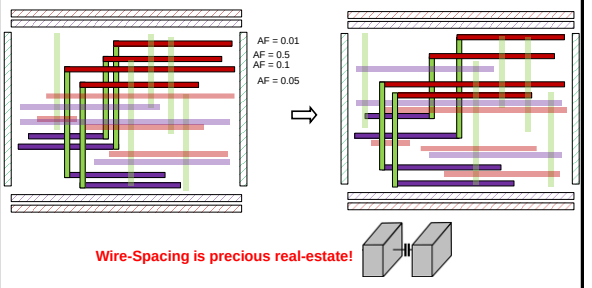


* S. Wimer, S. Michaely, K. Moiseev and A. Kolodny, "Optimal Bus Sizing in Migration of Processor Design", *IEEE Transactions on Circuits and Systems – I*, 2006.

$$\sum_{j=1}^{n} W_i + \sum_{j=0}^{n} S_i = A$$

***A* is a fixed constraint**

Wire layout optimization:
Finding Optimal Wire Widths and Spaces under Delay Constraints

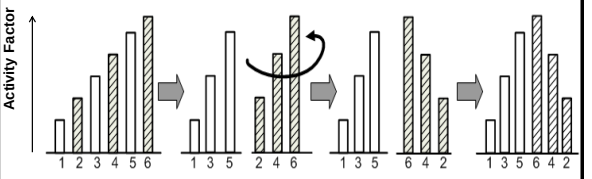


Wire-Spacing is precious real-estate!

* K. Moiseev, S. Wimer and A. Kolodny, "Timing-constrained Power Minimization in VLSI Circuits by Simultaneous Multilayer Wire Spacing," *Integration, 2014.*

Wire layout optimization:
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

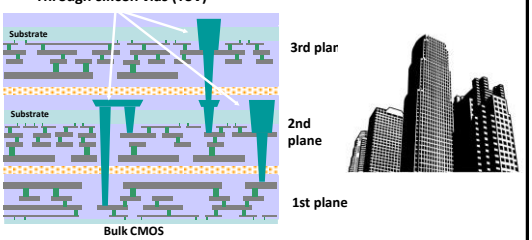


* K. Moiseev, S. Wimer and A. Kolodny, "Timing Optimization of Interconnect by Simultaneous Net-Ordering, Wire Sizing and Spacing," *INTEGRATION, 2007.*

Wire layout optimization:
3-D Integrated Circuit Technology

Reduce interconnect distances by building vertically

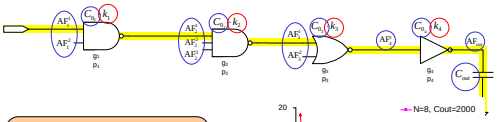
Through silicon vias (TSV)



*R. J. Gutmann *et al.*, "Three-Dimensional (3D) ICs: A Technology Platform for Integrated Systems and Opportunities for New Polymeric Adhesives," *Proceedings of the Conference on Polymers and Adhesives in Microelectronics and Photonics*, pp. 173-180, October 2001

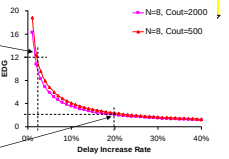
31

Circuit optimization:
Optimal Power-Delay Tradeoff for Logic Paths



For 2.5% delay increase, get 12x2.5=30% energy reduction!

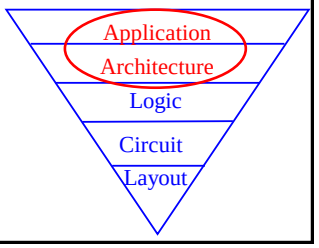
For 20% delay increase, get 2x20=40% energy reduction



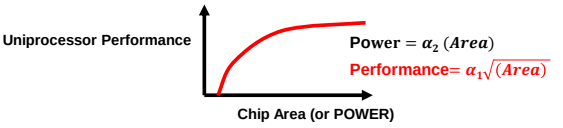
32

* Y. Aizik and A. Kolodny, "Finding the Energy Efficient Curve: Gate Sizing for Minimum Power under Delay Constraints," *VLSI Design*, 2011.
* A. Morgenshtein, E. G. Friedman, R. Ginosar and A. Kolodny, "Unified Logical Effort – A Method for Delay Evaluation and Minimization in Logic Paths with RC Interconnect," *IEEE Transactions VLSI*, 2010.

Most Power Savings Can be Made at High Abstraction Levels



Pollack's Rule on Power Efficiency of Uniprocessors



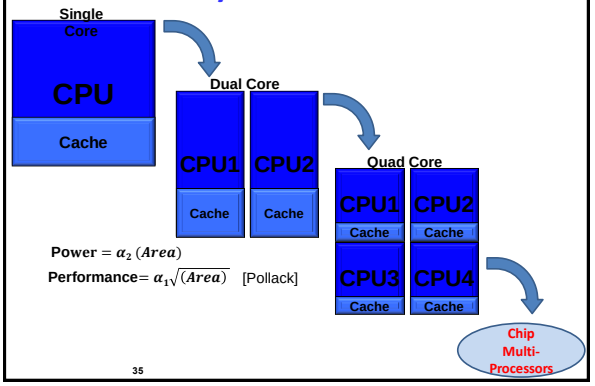
- Power-efficiency requires many parallel **local computations**
  - Chip Multi Processors (CMP)
  - Thread-Level Parallelism (TLP)

* F. Pollack, MICRO 1999.
* S. Borkar, DAC 2007.

34

Architecture optimization:

Processor System Evolution to CMP

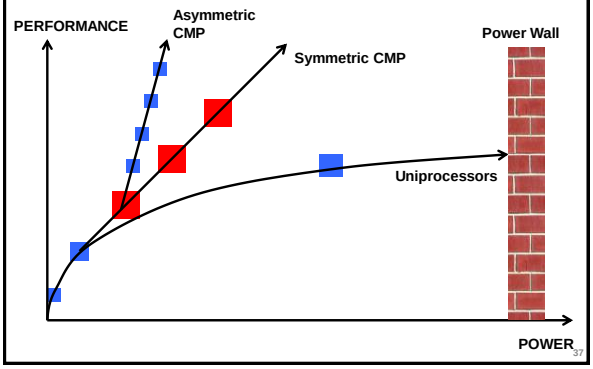


Power = $\alpha_2 (Area)$
Performance= $\alpha_1\sqrt{(Area)}$ [Pollack]

35

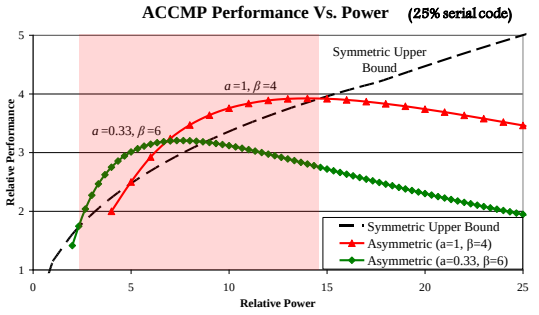
Processor Architectures:

Uni-core, Symmetric multicore, Asymmetric (Heterogeneous)



37

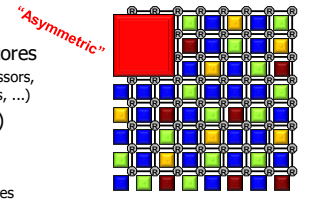
Architecture optimization:
Asymmetric Multi-Core Performance



* T. Morad, U. Weiser, A. Kolodny, M. Valero and E. Ayguade, IEEE Computer Architecture Letters, vol. 4, 2005.
* M. Hill and R. Marty, *Computer*, July 2008.

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"

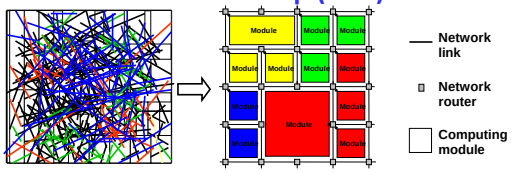


* I. Walter et al., NoCArc 2010.

43

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

Architecture optimization:
Network on-Chip (NoC)



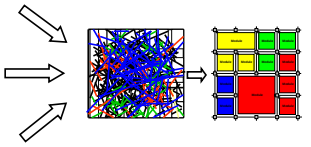
- **A new paradigm:**
  Network instead of dedicated wires and buses
  - **Inherently parallel**
  - **Efficient sharing of wires**
- **Scalable, cost effective bandwidth**

45

Issues Addressed by NoC

1) **Global wire design**
(delay, **power**, noise, scalability, reliability issues)

2) **System integration productivity**
(key to modular design)



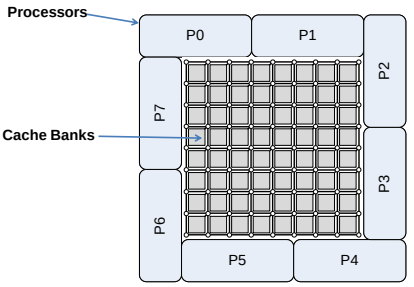
3) **Multi-Core Processor Systems**
(key to power-efficient computing)

46

Interconnect-aware and NoC-aware Architectural Research

47

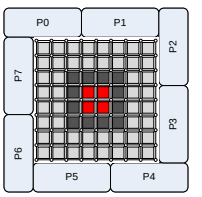
Accessing On-Chip Cache Banks through a NoC



48

Where to Store the Shared Data?

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



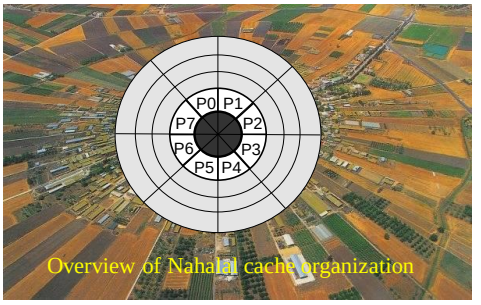
What can be done better?

- Bring shared data closer to all processors
- Preserve vicinity of private data

This Has Been Addressed Before

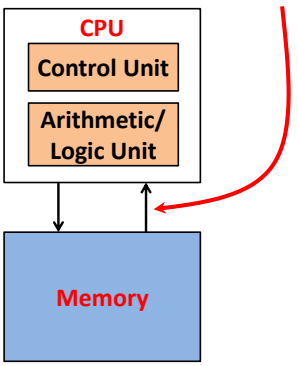
Aerial view of Nahalal cooperative village

This Has Been Addressed Before



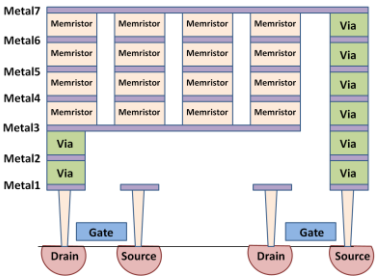
Aerial view of Nahalal cooperative village

Memory Bottleneck



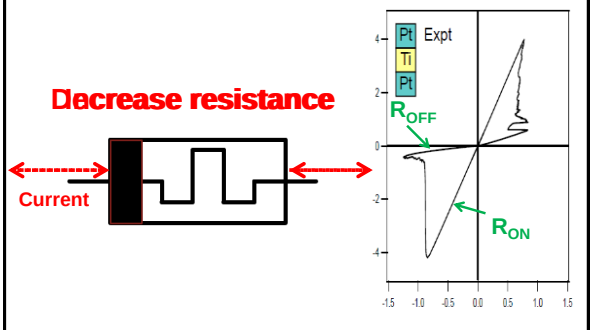
54

Reducing Distances by Embedding Memory in Execution Units



55

Memristor Devices

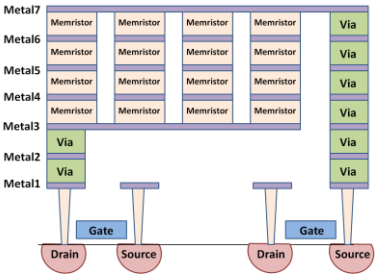


* D.B. Strukov et al, "The Missing Memristor Found," Nature, 2008

56

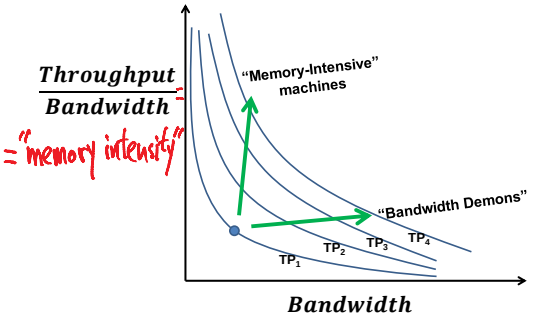
Sea of Memory

- Dense and fast



57

Towards Memory-Intensive Machines

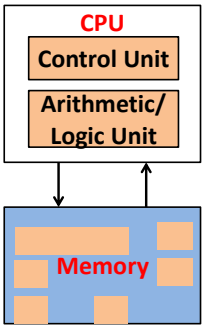


- Constant-throughput-curves
- ➔ increase on-die-memory!

58

Logic within the Memory

Beyond von Neumann Architecture



59

* S. Kvatinsky, G. Satat, N. Wald, E. G. Friedman, A. Kolodny, and U. C. Weiser, "Memristor-based Material Implication (IMPLY) Logic: Design Principles and Methodologies," IEEE Trans. on VLSI (to appear)

Summary

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