

Costs

André DeHon

<andre@cs.caltech.edu>

Wednesday, June 19, 2002



CBSSS 2002: DeHon

Key Points

- Every feature in our computing devices has a **cost**
 - Is something physical
 - Takes up space, has delay, consumes energy
- Cost structure varies with technology
- Optimal allocation/organization varies with cost structure

CBSSS 2002: DeHon

Costs

CBSSS 2002: DeHon

Physical Entities

- **Idea:** Computations take up space
 - Bigger/smaller computations
 - How fit into limited space?
 - Size → resources → cost
 - Size → distance → delay

CBSSS 2002: DeHon

Comment

- Experience from VLSI
 - Primarily 2D substrate
- Will want to generalize as appropriate for other substrate
 - Use concretes from VLSI

CBSSS 2002: DeHon

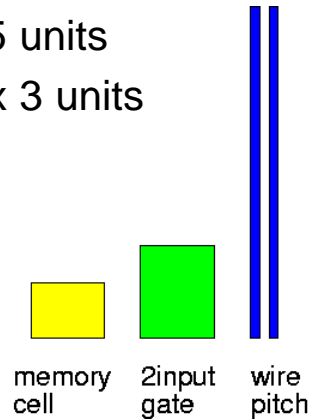
Area Components

- Gates -- compute
- Memory Cells -- state
- Wires -- interconnect

CBSSS 2002: DeHon

Typical VLSI

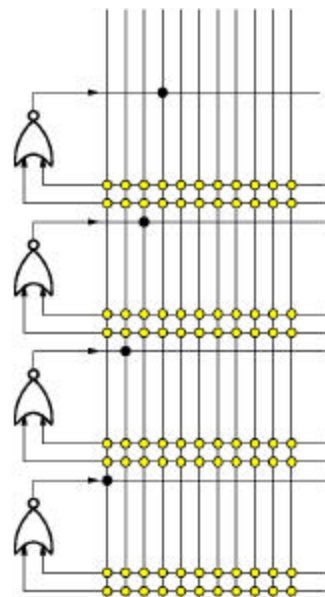
- Wires – normalizer – pitch 1 unit
- 2-input gate – maybe 4 x 5 units
- Memory Cells – maybe 4 x 3 units



CBS55 2002: DeHon

Structure Area

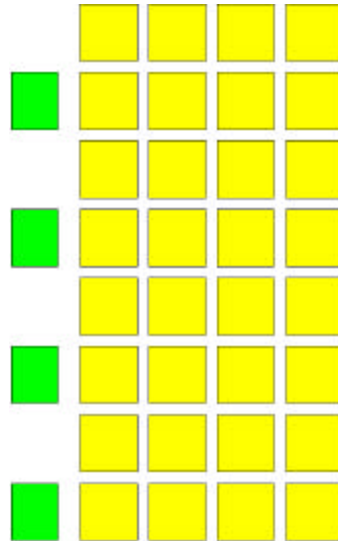
- Example: **nor2**-crossbar architecture
 - Crosspoint: about 2x memory cell
 - 5x5 units



CBS55 2002: DeHon

nor2-crossbar

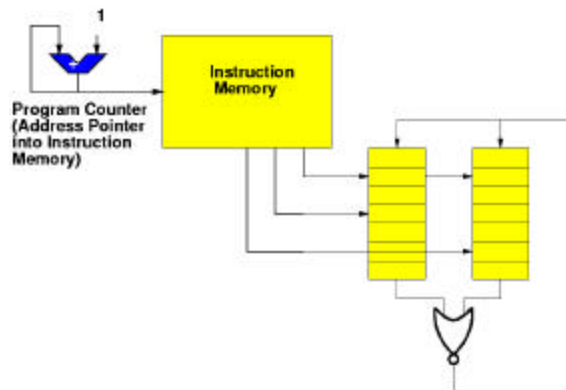
- N tall
 - Two crosspoints per NOR gate
 - Height/gate~10
- N wide
 - Width/xpoint~5
- Area= $50 \times N^2$



CBS55 2002: DeHon

Structure Area

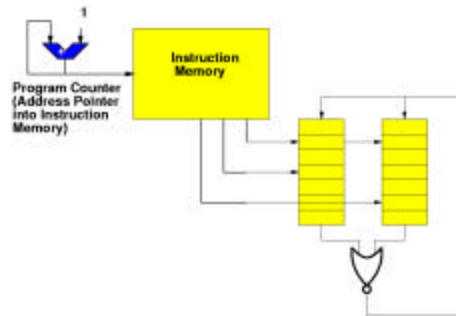
- Example 2: **nor2**-processors



CBS55 2002: DeHon

Components

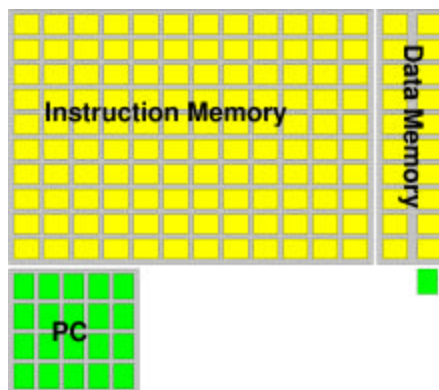
- Gate: 1
- Data Memory:
 - $2N$ memory cells
 - (underestimate)
- Instruction Memory:
 - $3 \log_2(N) \times N$ memory cells
- Counter:
 - $\log_2(N) \times 5$ gates/bit



CBS55 2002: DeHon

Components

- Gate: 1
- Data Memory:
 - $2N$ memory cells
 - (underestimate)
- Instruction Memory:
 - $3 \log_2(N) \times N$ memory cells
- Counter:
 - $\log_2(N) \times 5$ gates/bit



CBS55 2002: DeHon

nor2-processors

- Area:
 - $12(2N+3 \log_2(N) N) + 20(5 \log_2(N))$
 - $100 \log_2(N) + 24N + 36 \log_2(N) N$

CBS5 2002: DeHon

Area Compare

- crossbar processor
- 10: 5000 2080
- 100: 500,000 30,000
- 1000: 50M 380,000
- 10,000: 5G 15M
- (processor does Nx less calculations at a time)

CBS5 2002: DeHon

Area Comments

- When need to fit in limited area
 - Processor (temporal) version beneficial
 - Why processors preferred in early VLSI (pre-VLSI)
 - Physical space limited
 - Problems large
- In VLSI
 - State/description smaller than active
 - Largely because of compact memory

CBS55 2002: DeHon

Area Comments

- Can do better than crossbar for interconnect
 - ...next time

CBS55 2002: DeHon

Key Costs

- In VLSI:
 - Area, delay, energy
- Often, not simultaneously optimized
 - Give rise to tradeoffs
 - Previous is crude example of area-delay

CBS55 2002: DeHon

Costs Vary

CBS55 2002: DeHon

VLSI World

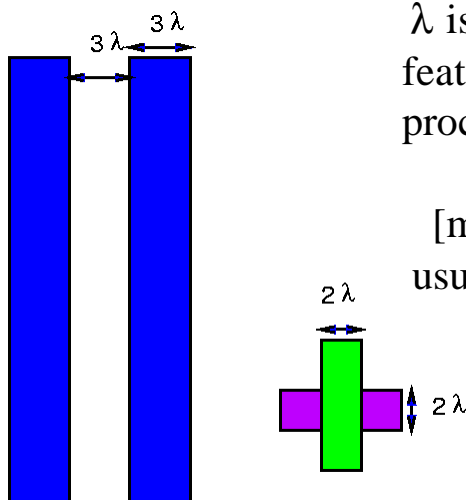
- Technology largely defined by precision in fabrication
 - Minimum feature size
 - A physical limit
 - On our ability to build and transfer patterning
 - Do so precisely

CBS5 2002: DeHon

Feature Size

λ is half the minimum feature size in a VLSI process

[minimum feature usually channel width]



CBS5 2002: DeHon

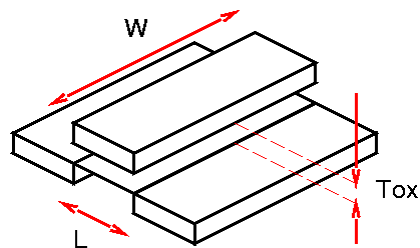
Predictable Variation

- Feature Sizes have been shrinking
 - As we get control over physical dimensions
- Feature Size shrink
 - Changes size limits
 - Shifts costs

CBSSS 2002: DeHon

Scaling

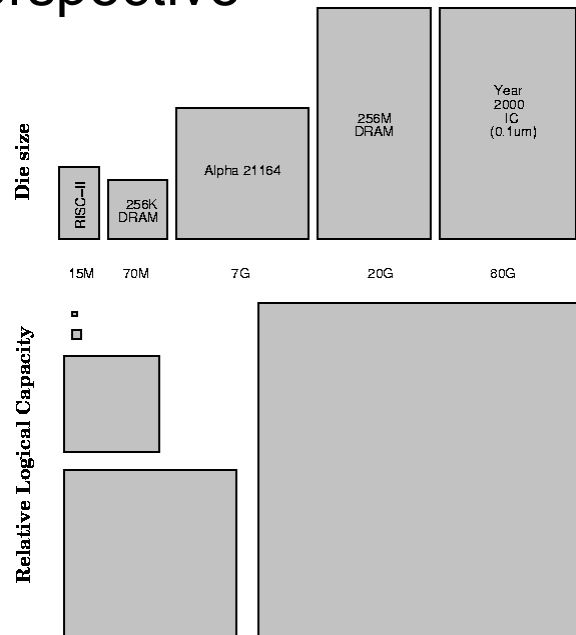
- Channel Length (L) λ
- Channel Width (W) λ
- Oxide Thickness (T_{ox}) λ
- Doping (N_a) $1/\lambda$
- Voltage (V) λ



CBSSS 2002: DeHon

Area Perspective

[2000 tech.]
18mm×18mm
0.18 μ m
60G λ^2



CBSSS 2002: DeHon

Capacity Growth

- Things which were not feasible a 5—10 years ago
 - Very feasible now
- Designs which **must** be done one way (e.g. temporal)...
 - now have many new options

CBSSS 2002: DeHon

Effects of Ideal Scaling?

- Area $1/\kappa^2$
 - Capacitance $1/\kappa$
 - Resistance κ
 - Threshold (V_{th}) $1/\kappa$
 - Current (I_d) $1/\kappa$
 - Gate Delay (τ_{gd}) $1/\kappa$
 - Wire Delay (τ_{wire}) 1
 - Power $1/\kappa^2 \rightarrow 1/\kappa^3$
- Delay shifts from gates to wires
 - Distance becomes a bigger factor in delay than gates

CBS55 2002: DeHon

VLSI Scaling Forward

- Can't scale forward forever
- Depend on bulk effects, large numbers of atoms
 - ...but approaching atomic scale
- Conventional VLSI feeling this pain
- Andrew Kahng will share the industry roadmap with us tonight

CBS55 2002: DeHon

Beyond VLSI

- Even w/in VLSI Scaling
 - Changing costs effect our designs
- Effect more pronounced moving between substrates
 - Memory not compact?
 - Memory and switches in 1x1 wire pitches?
 - Unit resistance wires?
 - Three dimensional wiring?
 - Three dimensional active device layout?

CBS5S 2002: DeHon

Beyond Silicon

- Don't know what the key costs and limits are
 - Unique/identifiable proteins or match addresses?
 - Length of binding domains?
 - Number of qbits?
- But, understanding them
 - Will be key to understanding how to engineer efficient structures

CBS5S 2002: DeHon

Cost Optimization Example

LUT Size

CBS55 2002: DeHon

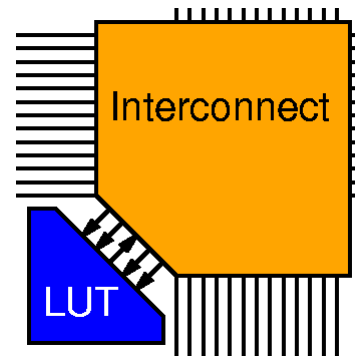
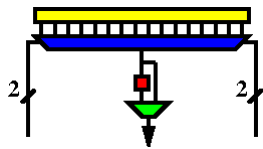
From Last Time

- Could build a large Lookup-Table
 - But grows exponentially in inputs
- Could interconnect a collection of programmable gates
 - How much does interconnect cost?
- How complex (big) should the gates be?

CBS55 2002: DeHon

LUTs with Interconnect

Alternative to one
big LUT



CBS55 2002: DeHon

Question Restated

- How large of a LUT should we use as the basic building blocking in a set of programmably interconnected gates?

CBS55 2002: DeHon

Qualitative Effects

- Larger LUTs
 - Reduce the number needed
 - Capture local interconnect, maybe cheaper than paying interconnect between them
 - Are less and less efficient for certain functions
 - *E.g. xor* and addition mentioned last time

CBS55 2002: DeHon

Qualitative Effects

- Smaller LUTs:
 - Pay large interconnect overhead
 - Overhead per gate less than exponential
 - Some functions take small numbers of gates
 - ...but other functions still require exponential gates (net loss)

CBS55 2002: DeHon

Memories and 4-LUTs

- For the **most complex** functions an M-LUT has $\sim 2^{M-4}$ 4-LUTs
- SRAM 32Kx8 $\lambda=0.6\mu\text{m}$
 - $170M\lambda^2$ (21ns latency)
 - $8 \cdot 2^{11} = 16\text{K}$ 4-LUTs
- XC3042 $\lambda=0.6\mu\text{m}$
 - $180M\lambda^2$ (13ns delay per CLB)
 - 288 4-LUTs
- Memory is 50+x denser than FPGA
 - ...and faster

CBS5 2002: DeHon

Memory and 4-LUTs

- For “regular” functions?
- 15-bit parity
 - entire 32Kx8 SRAM
 - 5 4-LUTs
 - (2% of XC3042 $\sim 3.2M\lambda^2 \sim 1/50\text{th}$ Memory)
- 7b Add
 - entire 32Kx8 SRAM
 - 14 4-LUTs
 - (5% of XC3042, $8.8M\lambda^2 \sim 1/20\text{th}$ Memory)

CBS5 2002: DeHon

Empirical Approach

- Look at trends across benchmark set of “typical” designs
 - Partially a question about typical regularity
 - Much of computer “architecture” is about understanding the **structure** of problems
- Use algorithm for covering with small LUTs
- How many need?
- How much area do they take up with interconnect?

CBSSS 2002: DeHon

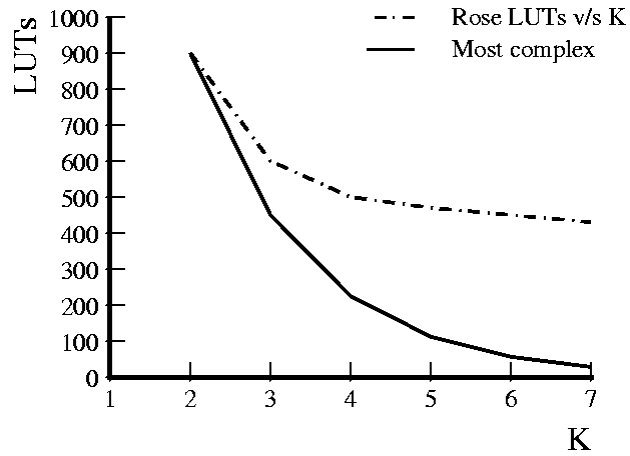
Toronto Experiments

- Pick benchmark set
- Map to K-LUTs
 - Vary K
- Route the K-LUTs
- Develop area/cost model
- Compute net area
 - Minimum?

[Rose et. al. JSSC v25n5p1217]

CBSSS 2002: DeHon

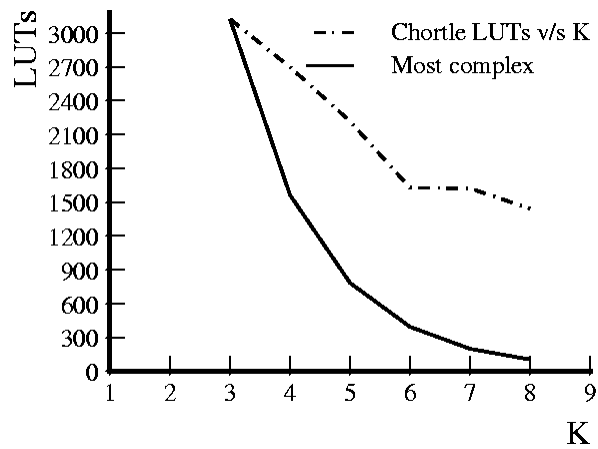
LUT Count vs. base LUT size



CBS2002: DeHon

LUT vs. K

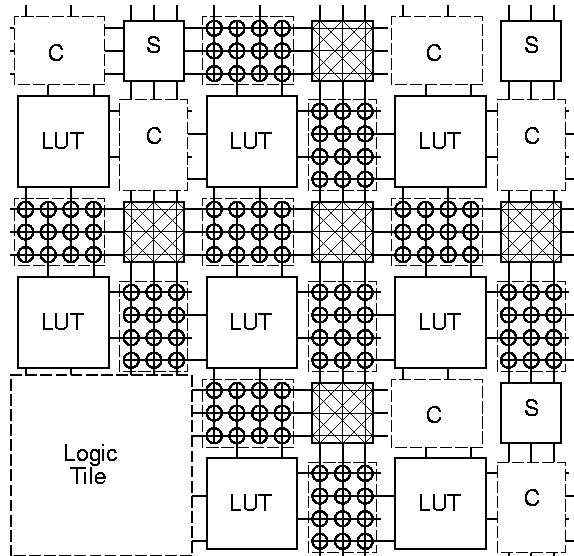
- DES MCNC Benchmark
– moderately irregular



CBS2002: DeHon

Toronto FPGA Model

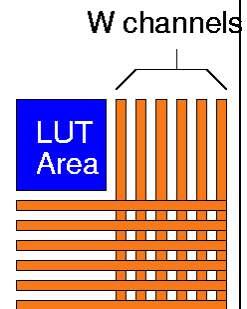
Connect
FPGAs
In Mesh
(hopefully,
less than
crossbar)



CBS 2002: DeHon

Toronto LUT Size

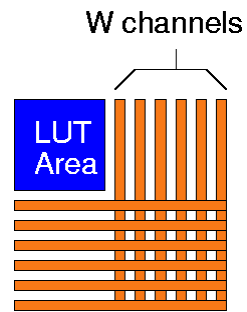
- Map to K-LUT
 - use Chortle
- Route to determine wiring tracks
 - global route
 - different channel width W for each benchmark
- Area Model for K and W



CBS 2002: DeHon

LUT Area

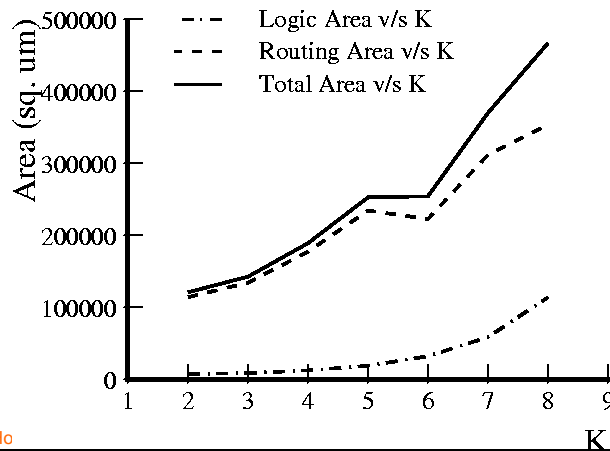
- K-LUT: $c + \text{memcell} * 2^K$
- Switches: linear in W
 - E.g. $\text{Area} = 12 * W * \text{switches}$
 - *How does W grow with N ?*
 - (for next time)
- Interconnect in fixed layers:
 - $W^2 * \text{pitch}^2$
 - (but assume switched dominate)



CBS55 2002: DeHon

LUT Area vs. K

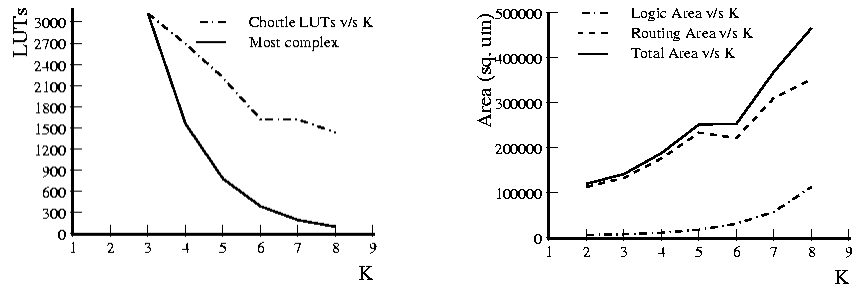
- Routing Area roughly linear in K



CBS55 2002: DeHo

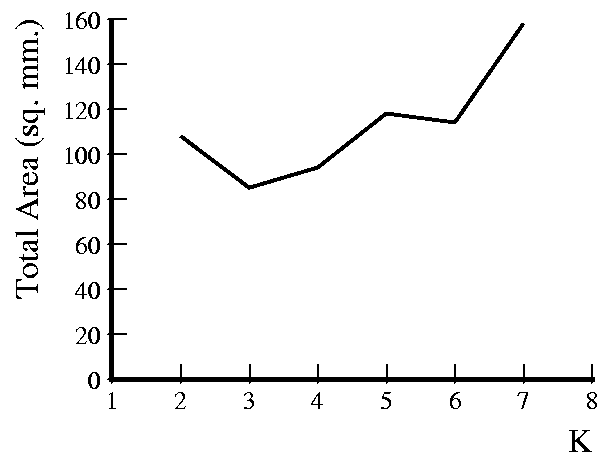
Mapped LUT Area

- Compose Mapped LUTs and Area Model



CBSSS 2002: DeHon

Mapped Area vs. LUT K



N.B. unusual case minimum area at K=3

CBSSS 2002: DeHon

Toronto Result

- Minimum LUT Area
 - at $K=4$
 - Important to note minimum on previous slides based on particular cost model
 - robust for range of switch sizes

CBSSS 2002: DeHon

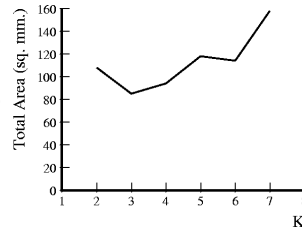
Implications

- For this cost model,
 - Efficient to interconnect small LUTs
 - **Even** though it may mean most of the area in wiring
 - Need wiring to exploit structure of problems

CBSSS 2002: DeHon

General Result

- This kind of result typical
 - Understand competing factors
 - Cost (area per K-LUT)
 - Utility (unit reduction w/ K-LUT)
 - Understand variations
 - Find minimum for cost and variation model



CBSSS 2002: DeHon

Wrapup

CBSSS 2002: DeHon

Key Points

- Every feature in our computing devices has a **cost**
 - Is something physical
 - Takes up space, has delay, consumes energy
- Cost structure varies with technology
- Optimal allocation/organization varies with cost structure

CBS5 2002: DeHon

Coming Attractions

- Change and limits in VLSI
 - Andrew Kahng, this afternoon (4:30pm)
- Interconnect requirements and optimization
 - Tomorrow
- **No 10:30am lecture today**

CBS5 2002: DeHon