ESE680-002 (ESE534): Computer Organization

Day 14: February 28, 2007

Interconnect 2: Wiring Requirements and Implications

Previously

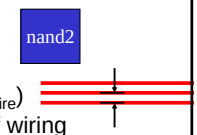
- Identified need for Interconnect
- Saw why simplest things don't work
  - Bus
  - Crossbar
- Identified need to understand/exploit structure in our interconnect problem

Today

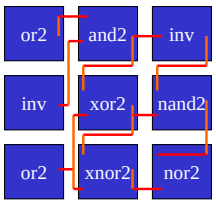
- Wiring Requirements
- Rent's Rule
  - A model of structure
- Implications

Wires and VLSI

- Simple VLSI model
  - Gates have fixed size (A_{gate})
  - Wires have finite spacing (W_{wire})
  - Have a small, finite number of wiring layers
    - *E.g.*
      - one for horizontal wiring
      - one for vertical wiring
  - Assume wires can run over gates



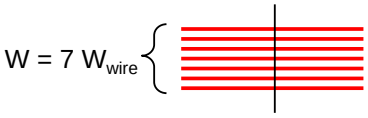
Visually: Wires and VLSI



Important Consequence

- A set of wires
- crossing a line
- take up space:

$$W = (N \times W_{wire}) / N_{layers}$$



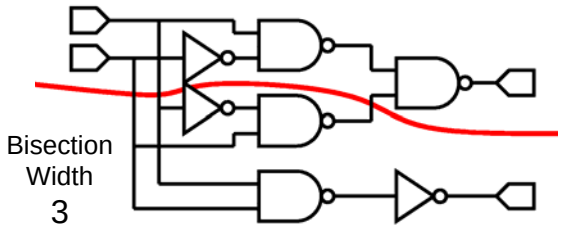
Thompson's Argument

- The minimum area of a VLSI component is bounded by the larger of:
  - The area to hold all the gates
    - $A_{chip} \geq N \times A_{gate}$
  - The area required by the wiring
    - $A_{chip} \geq N_{horizontal}\ W_{wire} \times N_{vertical}\ W_{wire}$

How many wires?

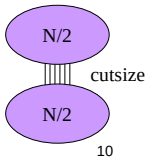
- We can get a **lower bound** on the total number of horizontal (vertical) wires by considering the **bisection** of the computational graph:
  - Cut the graph of gates in half
  - Minimize connections between halves
  - Count number of connections in cut
  - Gives a lower bound on number of wires

Bisection



Next Question

- In general, if we:
  - Cut design in half
  - Minimizing cut wires
- How many wires will be in the bisection?



Arbitrary Graph

- Graph with N nodes
- Cut in half
  - N/2 gates on each side
- **Worst-case:**
  - Every gate output on each side
  - Is used somewhere on other side
  - Cut contains N wires

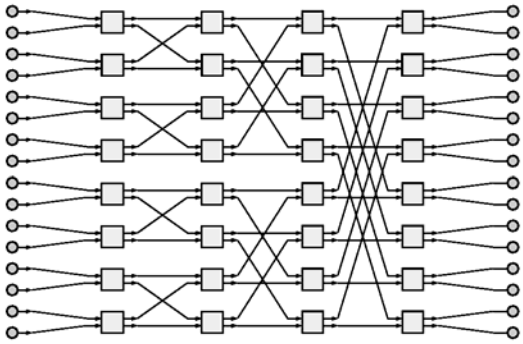
Arbitrary Graph

- For a random graph
  - Something proportional to this is likely
- That is:
  - Given a random graph with N nodes
  - The number of wires in the bisection is likely to be: $c \times N$

Particular Computational Graphs

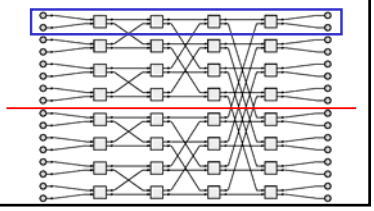
- Some important computations have exactly this property
  - FFT (Fast Fourier Transform)
  - Sorting

FFT



FFT

- Can implement with N/2 nodes
  - Group row together
- Any bisection will cut N/2 wire bundles
  - True for any reordering



Assembling what we know

- $A_{chip} \geq N \times A_{gate}$
- $A_{chip} \geq N_{horizontal}\ W_{wire} \times N_{vertical}\ W_{wire}$
- $N_{horizontal} = c \times N$
- $N_{vertical} = c \times N$
  - [bound true recursively in graph]
- $A_{chip} \geq cN\ W_{wire} \times cN\ W_{wire}$

Assembling …

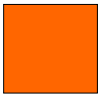
- $A_{chip} \geq N \times A_{gate}$
- $A_{chip} \geq cN\ W_{wire} \times cN\ W_{wire}$
- $A_{chip} \geq (cN\ W_{wire})^2$
- $A_{chip} \geq N^2 \times c'$

Result

- $A_{chip} \geq N \times A_{gate}$
- $A_{chip} \geq N^2 \times c'$
- Wire area grows faster than gate area
- Wire area grows with the square of gate area
- For sufficiently large N,
  - Wire area dominates gate area

Intuitive Version

- Consider a region of a chip
- Gate capacity in the region goes as area (s^2)
- Wiring capacity into region goes as perimeter (4s)
- Perimeter grows more slowly than area
  - Wire capacity saturates before gate



Result

- $A_{chip} \geq N^2 \times c'$
- Wire area grows with the square of gate area
- Troubling:
  - To **double** the size of our computation
  - Must **quadruple** the size of our chip!

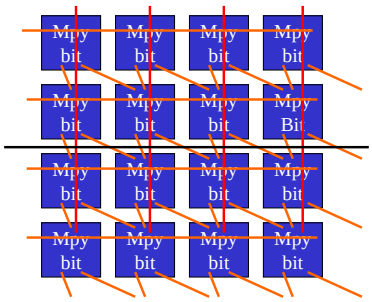
So what?

What do we do with this observation?

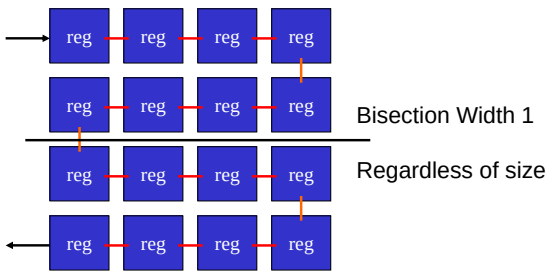
First Observation

- Not all designs have this large of a bisection
- What is typical?

Array Multiplier



Shift Register

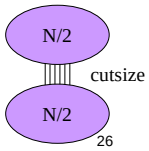


Architecture ⇔ Structure

- Typical architecture trick:
  - exploit expected problem structure
- What structure do we have?
- Impact on resources required?

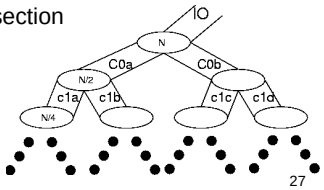
Bisection Bandwidth

- Bisection bandwidth of design
  - →lower bound on network bisection bandwidth
  - important, **first order** property of a design.
  - Measure to characterize
    - Rather than assume worst case
- Design with more locality
  - → lower bisection bandwidth
- Enough?



Characterizing Locality

- Single cut not capture locality within halves
- Cut again
  - → recursive bisection

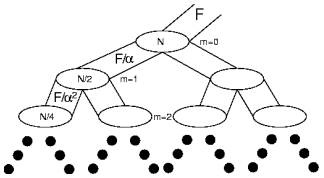


Regularizing Growth

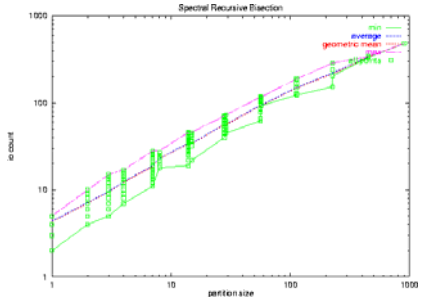
- How do bisection bandwidths shrink (grow) at different levels of bisection hierarchy?
- Basic assumption: Geometric
  - 1
  - $1/\alpha$
  - $1/\alpha^2$

Geometric Growth

- (F,α)-bifurcator
  - F bandwidth at root
  - geometric regression α at each level



Good Model?



Log-log plot → straight lines represent geometric growth

Rent's Rule

- In the world of circuit design, an empirical relationship to capture:

$$IO = c\,N^p$$

- $0 \le p \le 1$
- p – characterizes interconnect richness
- Typical: $0.5 \le p \le 0.7$
- "High-Speed" Logic p=0.67

Rent's Rule

- In the world of circuit design, an empirical relationship to capture:

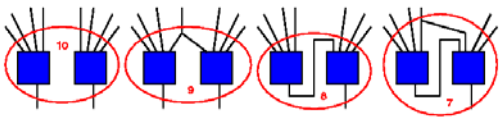
$$IO = c\,N^p$$

- compare (F,α)-bifurcator

$$\alpha = 2^p$$

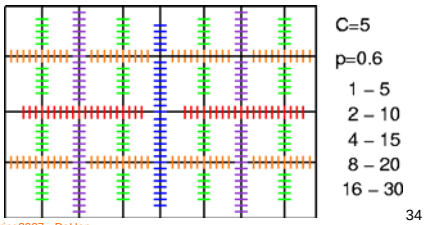
Rent and Locality

- Rent and IO capture/quantifying locality
  - local consumption
  - local fanout



What tell us about design?

- Recursive bandwidth requirements in network



As a function of Bisection

- $A_{chip} \ge N \times A_{gate}$
- $A_{chip} \ge N_{horizontal}\, W_{wire} \times N_{vertical}\, W_{wire}$
- $N_{horizontal} = N_{vertical} = IO = cN^p$
- $A_{chip} \ge (cN)^{2p}$
- If p<0.5

$$A_{chip} \propto N$$

- If p>0.5

$$A_{chip} \propto N^{2p}$$

In terms of Rent's Rule

- If p<0.5, $A_{chip} \propto N$
- If p>0.5, $A_{chip} \propto N^{2p}$

- **Typical** designs have **p>0.5**

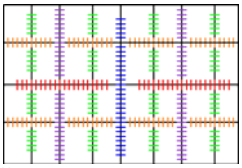
→ interconnect dominates

What tell us about design?

- Recursive bandwidth requirements in network
  - **lower bound** on resource requirements
- N.B. **necessary** but not **sufficient** condition on network design
  - *I.e.* design must also be able to *use* the wires

What tell us about design?

- Interconnect lengths
  - Intuition
    - if p>0.5, everything cannot be nearest neighbor
    - as p grows, so wire distances

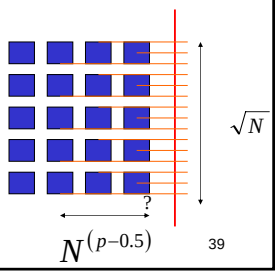


Can think of p as dimensionallity:
p=1-1/d

Interconnect Lengths

- Side is $\sqrt{(N)}$
- IO crossing it is N^p
- What's minimum length for longest wires?

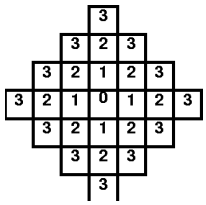
$$BW = N^P$$



What tell us about design?

- Interconnect lengths (more general)
  - IO=$(n^2)^P$ crossing distance n
  - end at exactly distance n:

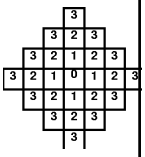
$$\frac{\partial(IO)}{\partial n}$$



What tell us about design?

- E(length) =
  - 1×(number at length 1)+2 ×(number at length 2) + 3 ×(number at length 3) …

$$E(length) = \int_0^{n=\sqrt{N}} n \times (Number(n))\partial n$$



$$Number(n) = \frac{\frac{\partial(IO)}{\partial n}}{n^2}$$

Assume iid sources; Equally likely to originate at any point in area.

Math

$$\frac{d(IO)}{dn} = \frac{d\left(cn^{2p}\right)}{dn} = 2pcn^{2p-1}$$

$$\int_0^{\sqrt{N}} \left(\frac{n \times \frac{d(IO)}{dn}}{n^2} \right) dn$$

$$\int_0^{\sqrt{N}} \left(\frac{n \times 2pcn^{2p-1}}{n^2} \right) dn$$

Math continued

$$\int_0^{\sqrt{N}} \left(\frac{n \times 2pcn^{2p-1}}{n^2}\right) dn$$

$$\int_0^{\sqrt{N}} \left(\frac{2pcn^{2p-1}}{n}\right) dn$$

$$\int_0^{\sqrt{N}} \left(2pcn^{2p-2}\right) dn$$

Math continued

$$\int_0^{\sqrt{N}} \left(2pcn^{2p-2}\right) dn$$

$$\left(\frac{2pcn^{2p-1}}{2p-1}\right)\Bigg|_0^{\sqrt{N}}$$

Math Continued

$$\left(\frac{2pcn^{2p-1}}{2p-1}\right)\Bigg|_0^{\sqrt{N}}$$

$$\left(\frac{2pc}{2p-1}\right)\left(N^{0.5}\right)^{2p-1}$$

$$\left(\frac{2pc}{2p-1}\right)\left(N^{p-0.5}\right)$$

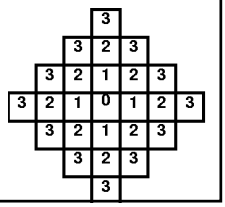
What tell us about design?

- Interconnect lengths
  - IO=$(n^2)^p$ cross distance n
  - at exactly distance n: $\frac{\partial(IO)}{\partial n}$

$$E(length) = \int_0^{n=\sqrt{N}} n \times \left(\frac{\frac{\partial(IO)}{\partial n}}{n^2}\right) \partial n$$

  - E(length)=$\Omega(N^{(p-0.5)})$
    - p>0.5

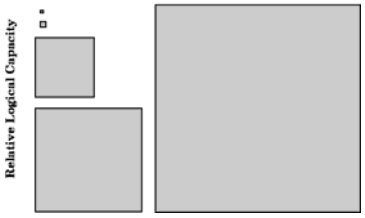
True even with multiple metal layers.



Recall from Day 6

Delays

- Logical capacities growing
- Wirelengths?
  - No locality→κ
  - Rent's Rule
    - L →$n^{(p-0.5)}$
    - [p>0.5]



Capacity

- Rent: IO=C*N^p
- p>0.5
- A= C*N^{2p}
- N=$(A/C)^{(1/2p)}$
- Logical Area →κ^2
- N'=$((\kappa^2 A)/C)^{(1/2p)}$
- N'=$(A/C)^{(1/2p)} \times (\kappa^2)^{(1/2p)}$
- N'=N ×$(\kappa^2)^{(1/2p)}$
- N'=N ×$(\kappa)^{(1/p)}$

- Sanity Check
  - p=1
  - $N_2 = \kappa N$
  - p~0.5
  - $N_2 \sim \kappa^2 N$

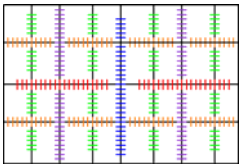
Capacity (alternate)

- Rent: IO=C*N^p
- p>0.5
- A= C*N^{2p}
- Logical Area →κ^2
  κ^2 A= C*N_2^{2p}
  κ^2 C*N_1^{2p} = C*N_2^{2p}
  κ^2 N_1^{2p} = N_2^{2p}
  κ N_1^p = N_2^p
  $N_2 = \kappa^{(1/p)} N_1$

- Sanity Check
  - p=1
  - $N_2 = \kappa N$
  - p~0.5
  - $N_2 \sim \kappa^2 N$

What tell us about design?

- IO∝N^P
- Bisection BW∝N^P
- side length ∝N^P
  - N if p<0.5
- Area ∝N^{2p}
  p>0.5
- Average Wire Length ∝ $N^{(p-0.5)}$
  p>0.5



N.B. 2D VLSI world has "natural" Rent of P=0.5 (area vs. perimeter)

Rent's Rule Caveats

- Modern "systems" on a chip -- likely to contain subcomponents of varying Rent complexity
- Less I/O at certain "natural" boundaries
- System close
  - (Rent's Rule apply to workstation, PC, PDA?)

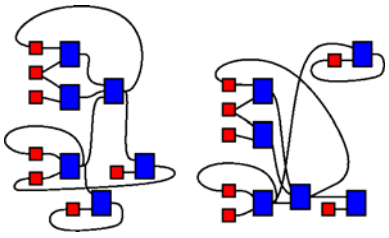
Area/Wire Length

- Bad news
  - Area ~ $\Omega(N^{2p})$
    - faster than N
  - Avg. Wire Length ~ Ω $(N^{(p-0.5)})$
    - grows with N
- Can designers/CAD control p (locality) once appreciate its effects?
- *I.e.* maybe this cost changes design style/criteria so we mitigate effects?

What Rent didn't tell us

- Bisection bandwidth purely geometrical
- No constraint for delay
  - *I.e.* a partition may leave critical path weaving between halves

Critical Path and Bisection



Minimum cut may cross critical path multiple times.
Minimizing long wires in critical path → increase cut size.

Rent Weakness

- Not account for path topology
- ? Can we define a "Temporal" Rent which takes into consideration?
  - Promising research topic

Big Ideas [MSB Ideas]

- Rent's rule characterize locality
  Fixed wire layers:
  →Area growth Ω (N^{2p})
  → Wire Length Ω ($N^{(p-0.5)}$)
- p>0.5→ interconnect growing faster than compute elements
  - expect interconnect to dominate other resources