ESE534: Computer Organization

Day 17: March 29, 2010
Interconnect 2: Wiring Requirements and Implications

Previously

- Identified need for Interconnect
- Seen that interconnect can be expensive
- Identified need to understand/exploit structure in our interconnect design

Today

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

Wires and VLSI

- Simple VLSI model
  - Gates have fixed size ($A_{\text{gate}}$)
  - Wires have finite spacing ($W_{\text{wire}}$)
  - 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

Preclass 1

- How many $40F \times 40F$ gates in $25,000F \times 25,000F$ region?

- How many wires can go in and out?

- Ratio?
Important Consequence

- A set of wires
- crossing a line
- take up space:
  \[ W = \frac{N \times W_{\text{wire}}}{N_{\text{layers}}} \]

Thompson’s Argument

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

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

- 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

Next Question

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

Assembling what we know

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

Assembling ...
Result

- $A_{\text{chip}} \geq N \times A_{\text{gate}}$
- $A_{\text{chip}} \geq N^2 \times c'$
- 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

Preclass 2

- How does ratio change for 100,000 F×100,000 F region?

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

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

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 wire crossings
  - 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

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

Rent’s Rule

- In the world of circuit design, an empirical relationship to capture:
  \[ \text{IO} = c \ N^p \]
- \(0 \leq p \leq 1\)
- \(p\) – characterizes interconnect richness
- Typical: \(0.5 \leq p \leq 0.7\)
- “High-Speed” Logic \(p=0.67\)

Rent and Locality

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

What tell us about design?

- Recursive bandwidth requirements in network
As a function of Bisection

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

In terms of Rent’s Rule

- If $p<0.5$, $A_{\text{chip}} \propto N$
- If $p>0.5$, $A_{\text{chip}} \propto N^{2p}$
- 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 dimensionality: $p = 1-1/d$

Preclass 3

- 25,000 F side, 40F × 40 F gates
- Wire length?
Generalizing Interconnect Lengths

- \( P > 0.5 \)
- Side is \( \sqrt{N} \)
- IO crossing it is \( N^p \)
- What’s minimum length for longest wires?
  - Implication:
    - Wire lengths grow at least as fast as \( N^{(p-0.5)} \)

\[
BW = N^p
\]

Delays

- Logical capacities growing
- Wirelengths?
  - No locality \( \propto \kappa \)
  - Rent’s Rule
    - \( L \propto N^{(p-0.5)} \)
    - \([p>0.5]\)

Recall from Day 7

Capacity

- Rent: \( IO = C \times N^p \)
- \( A = C \times N^{2p} \)
- \( N = \left(\frac{A}{C}\right)^{1/(2p)} \)
- Logical Area \( \propto \kappa^2 \)
- \( N = \left(\frac{\kappa^2 A}{C}\right)^{1/(2p)} \)
- \( N = \left(\frac{\kappa^2}{C}\right)^{1/(2p)} \)
- \( N = N \times (\kappa)^{1/p} \)

- Sanity Check
  - \( p=1 \)
  - \( N_2 = \kappa N \)
  - \( p-0.5 \)
  - \( N_2 = \kappa^2 N \)

What tell us about design?

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

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, MP3 player, Smart Phone?

Preclass 4

- Depth 20 circuit, 2-input gates
  - Maximum number of gates?
    - Topology?
    - Rent \( p \)?
  - Minimum distance?
    - Lower bound maximum length
- Depth 24 circuit
  - Lower bound maximum length?
Area/Wire Length

• Bad news
  – Area $\sim \Omega(N^{2p})$
    • faster than $N$
  – Avg. Wire Length $\sim \Omega(N^{p-0.5})$
  • 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 $\rightarrow$ increase cut size.

Original Memo

• Current Issue (Winter 2010, v2n1) of IEEE Solid-State Circuits Magazine
• Retrospect on IBM 1401 and E. F. Rent
  – Including original memos
• Added link to reading

Admin

• HW5 graded
• HW8 out – due April 12th
• Reading for Wed. on web

Big Ideas

[MSB Ideas]

• Rent’s rule characterizes locality
  Fixed wire layers:
    $\rightarrow$ Area growth $\Omega(N^{2p})$
    $\rightarrow$ Wire Length $\Omega(N^{p-0.5})$
• $p>0.5 \rightarrow$ interconnect growing faster than compute elements
  – expect interconnect to dominate other resources