On-Chip Interconnects and Wire Problems

ACAL Group Seminar

Raj Parihar
Outline

- Wire Problem
- Wire Metrics
- Leveraging Wire Properties
- Wire Aware Microarchitectures
- Summary
- References
Future of Wires

- Scaling and Current Scenario
  - Increasing Chip Complexity (More functionality)
    - Local wire % grows exponentially
    - CAD tool should account this increased wire density
  - Global On-Chip Communication Cost
    - Wire performance, relative to gates will continue to worsen
    - Aggressive use of repeaters and buffers results into increased chip area and power consumption
Wire Problem: What is it?

- With Technology Scaling
  - Gate delay decreases
    - less area and capacitance
  - On chip Local wire delay decreases
    - Wires geometry is scaled down too
  - Global wire delay scales upward
    - These wires do not scale down

- Increasing difference in “global wire delay” w. r. t. gate delays is popularly known as “Wire Problem”.

- On an average 50% power is consumed in on-chip interconnects.
Metrics for Gates & Wires

- How does a wire affect Circuit Performance?
  - Capacitance
    - Adds load to driving gate
  - Capacitance, Resistance and Inductance
    - Add signal delay (i.e. RC delay)
  - Capacitance, Inductance
    - Add signal noise by coupling to neighbors
Wire Characterizes

- RC Delay Models
  - Inductance effect is avoided for various reasons

- Resistance Modeling
  - R wire = p / (Thickness – barrier) . (Width – 2 barrier)

- Capacitance Modeling
  - C wire = 2. C vert + 2. C horz + C fringe (vert, horz)
Leveraging Wire Properties

- Heterogeneous – Hybrid Wire Architecture
  - Low Latency Wires
  - High Bandwidth Wires
  - Low Power Wires

- The idea is to use appropriate sets of wire for a particular kind of transfer
  - Improves performance (Roughly 10%)
  - Saves overall energy (Roughly 60%)
Wire Aware Microarchitecture # 1

- A Directory-based Cache Coherence Implementation with hop imbalance
- Example: Write Miss by P1
Wire Aware Microarchitecture # 2

- Accelerating Cache Access in Highly Associative caches
- Optimize the Hit Latency of Tag Array

Diagram:
- Low Latency L wire
- High BW/ Low Power wire
- Tag Array
- Data Array
Wire Aware Design: Guidelines

- Narrow Width Messages (Low Latency Wire)
  - ACK, NAK and other short messages
  - Invalidation in Cache Coherence Protocols
  - Checks and Control Signal Data

- High Bandwidth Messages (High BW Wire)
  - Data read from L1 or L2 cache
  - Inter-processor data sharing in MP

- Low Power Communication (Low Power Wires)
  - Write backs due to cache eviction
L Wire: Implementation & Deployment

- Low latency wire implementation
  - Fat RC wire with higher aspect ratio
  - High Speed Optical Links
  - Differential Twisted Pairs
  - Electromagnetic Waveguides
  - Low Swing Differential Pairs

- Deployment of L wires
  - Traffic characterization
  - *Latency – Bandwidth* trade-off
Leveraging Wires Further… (Ideas)

- Traffic Classification
  - Latency Critical
  - Bandwidth Critical
  - Power Aware Transfer

- Decoupled Architecture
  - The idea is to decouple the whole architecture into two parts
    - Latency Critical / Bandwidth Relaxed
    - Latency Relaxed / Bandwidth Critical
Wire Problem: Implications

- **Architecture**
  - Communication centric rather than computation centric

- **CAD Tools**
  - Accurate wire modeling and inductance extrication

- **Circuit**
  - High speed serial links and on-chip deployment

- **Manufacturing**
  - Feasibility of packaging and manufacturability
Summary

- On-Chip Wires: No more a second class citizens. They need attentions now.

- Heterogeneous – hybrid, optimized for specific wire architecture would be most efficient in terms of Power, Area, Delay.

- Adding the support for modeling into CAD tools is interesting and challenging. Avoid “Wire Exceptions” using low Latency wires.

- On-chip traffic study, characterization & Exploitation
References


