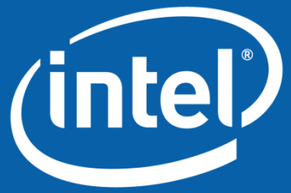


Impact of Local Interconnects on Timing and Power in a High Performance Microprocessor

Rupesh S. Shelar

Low Power IA Group

Intel Corporation, Austin, TX

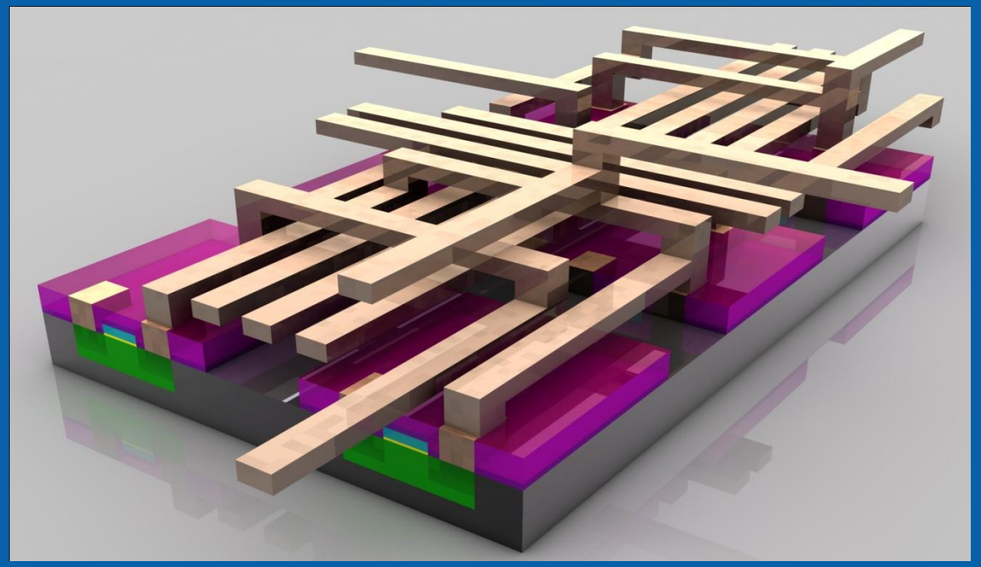
December 13th 2011

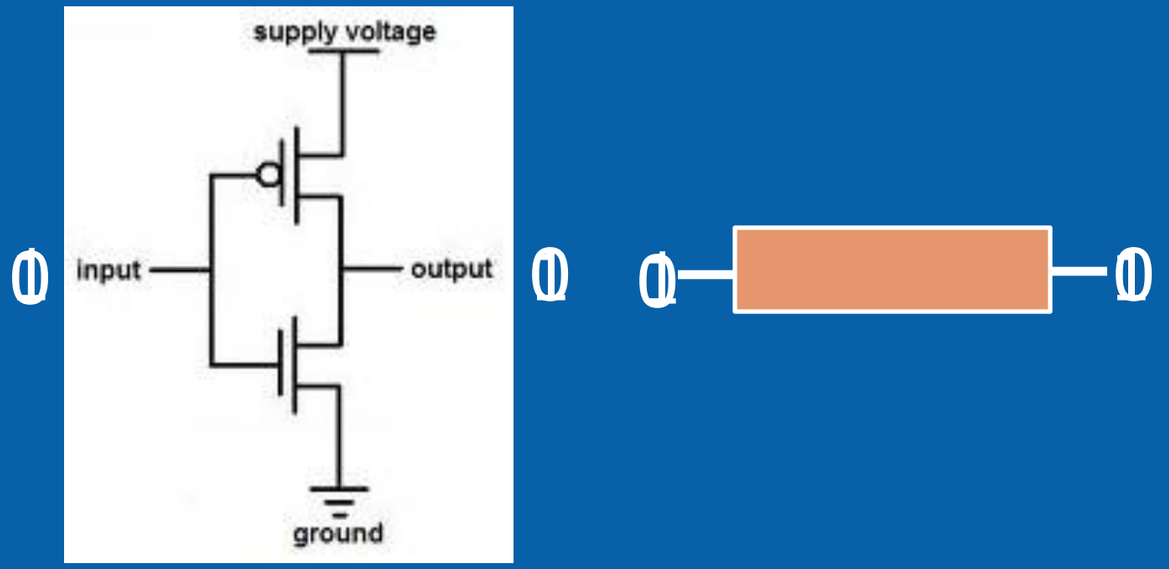
UT ICS/IEEE Seminar, UT Austin

Agenda

- Introduction
- Impact on Timing
- Impact on Power
- Conclusions

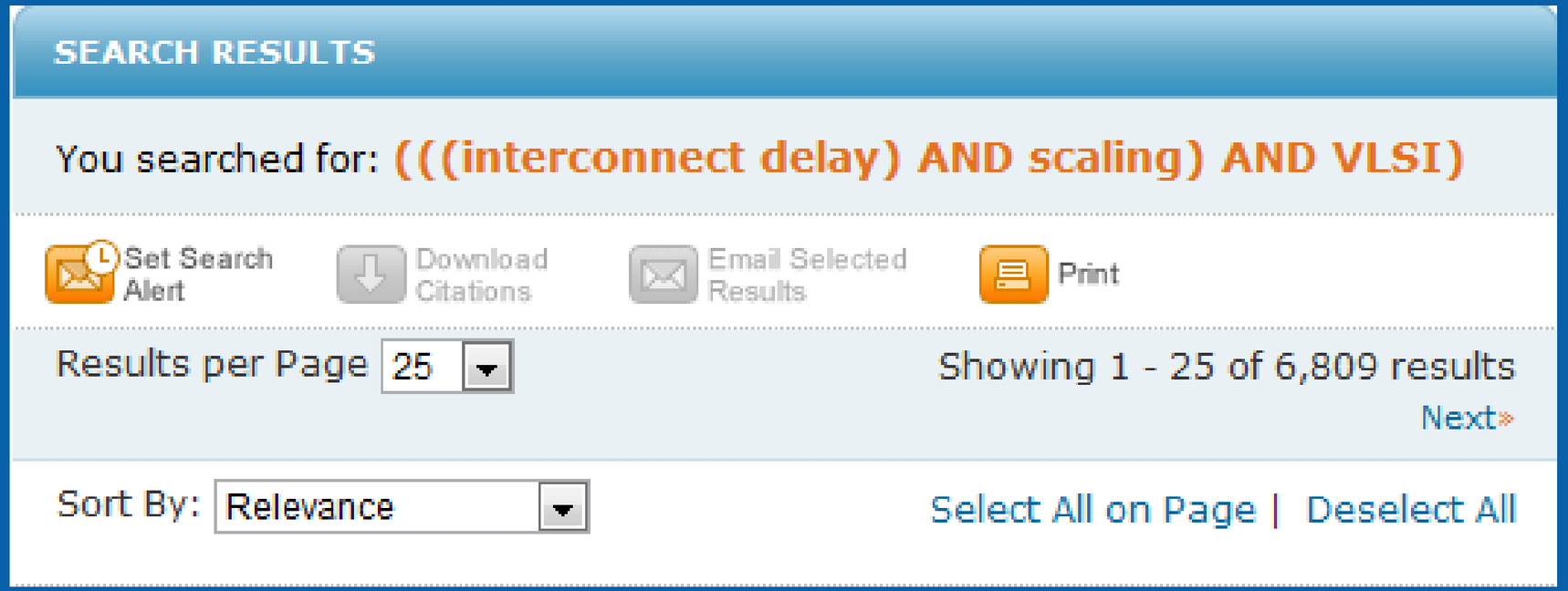
Why Look at Interconnects Closely





- Unlike transistors, interconnects
  - do not perform any computation
  - merely transfer information
- Paying power/timing cost for wires yields nothing

Motivation: Interconnect Delay & Power



- Global interconnects known to contribute significantly to path delays
- For local interconnects in intra-block paths, exact numbers probably not known, as these vary depending on the block-size, design style
- Relatively less attention paid to interconnect power dissipation
- Many academic studies exist: most based on small data

About Data

- Delay/power data from blocks in a high performance microprocessor core in 45 nm technology
- Blocks implemented using RTL-to-Layout Synthesis (RLS) design style
  - Mostly automated (using vendor/in-house tools); write RTL, partition, and run tools/flows
  - Design quality determined by algorithms, tools, flows, parameters; supposedly poor utilization, or sparse layouts
- Local interconnects: implemented mostly in min-width M2 to M5 layers
- Delay/power impact due to interconnects inside standard cells is considered as cell-delay/-power contribution in this study

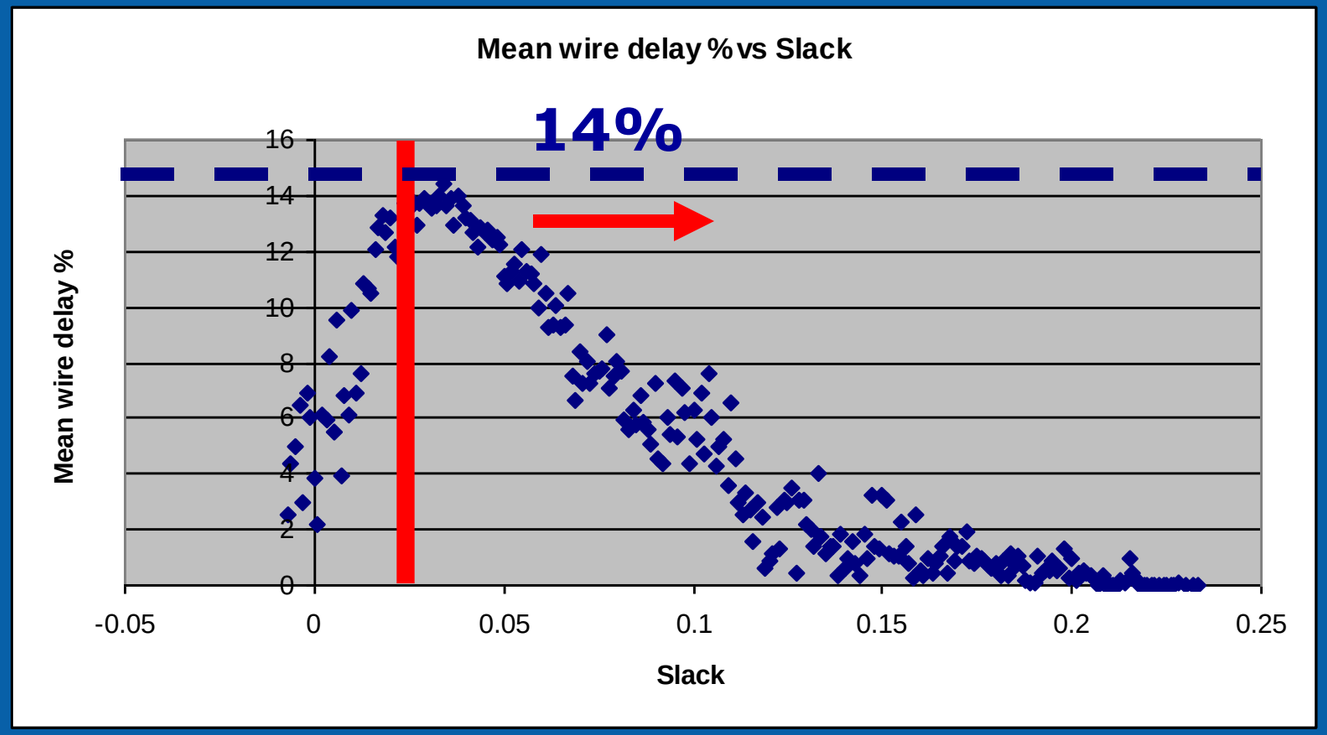
Agenda

- Introduction
- Impact on Timing
- Impact on Power
- Conclusions

Impact of Interconnects on timing

- For max timing, interconnects contribute in terms of
  - Wire delay
  - Slope degradation (slows down receivers)
  - Cell-delay degradation (slows down driver)
  - Cumulative effect of above 3 on path delays
  - Delays due to repeaters (inserted for timing/slope/noise)
- Chose 3 metrics on the worst internal paths:
  - Wire delay
  - Interconnect impact (obtained by setting R=C=0)
  - Repeater delay
- Why internal paths: should exclude the effect of timing constraints on primary i/os on results due to synthesis flows (RLS)
- Why worst paths: determine operating frequency

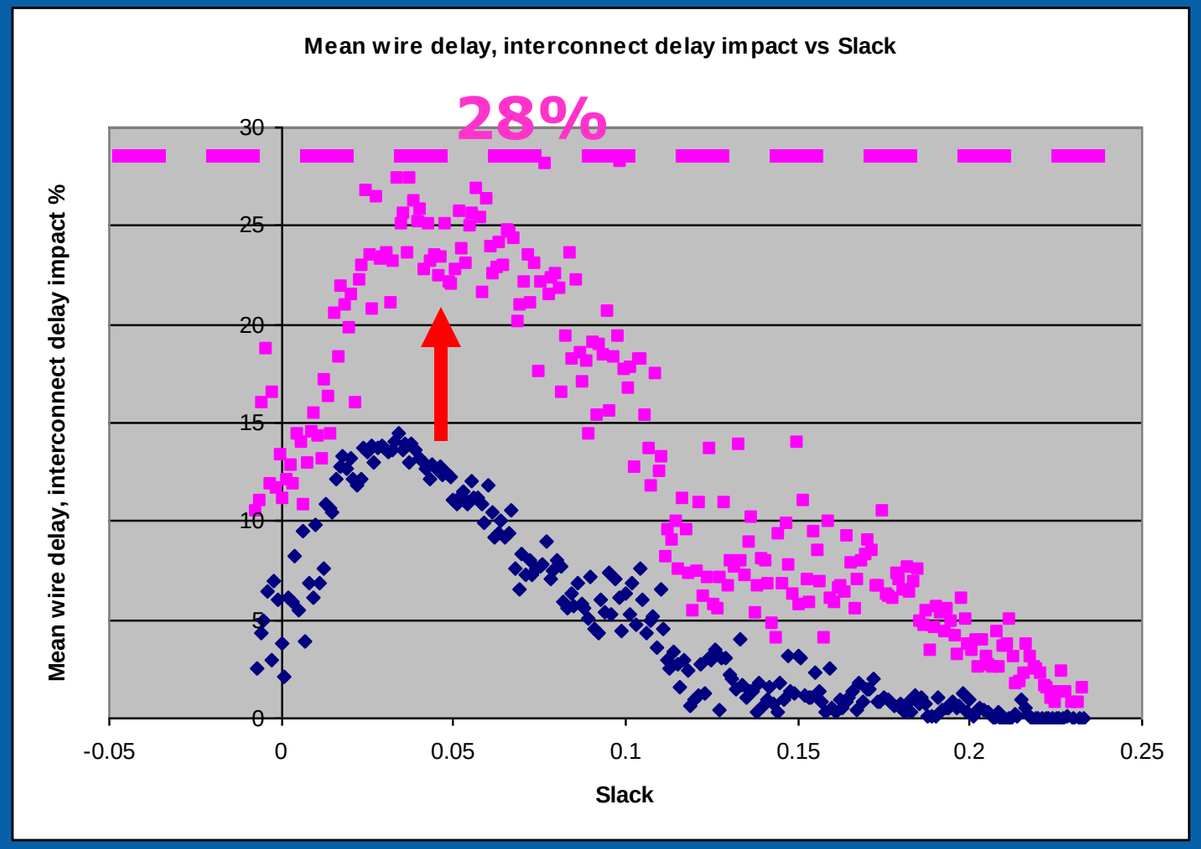
A Close Look at One Block: Wire Delay



Mean wire delay % vs slack for worst internal paths between unique pairs of sequentials in a ~40 K cell block with ~4 K sequentials

- Wire delay increases as slack decreases
- Timing wall due to sizing/ll-insertion because of emphasis on power also
- Interconnect delay impact won't change without power optimization

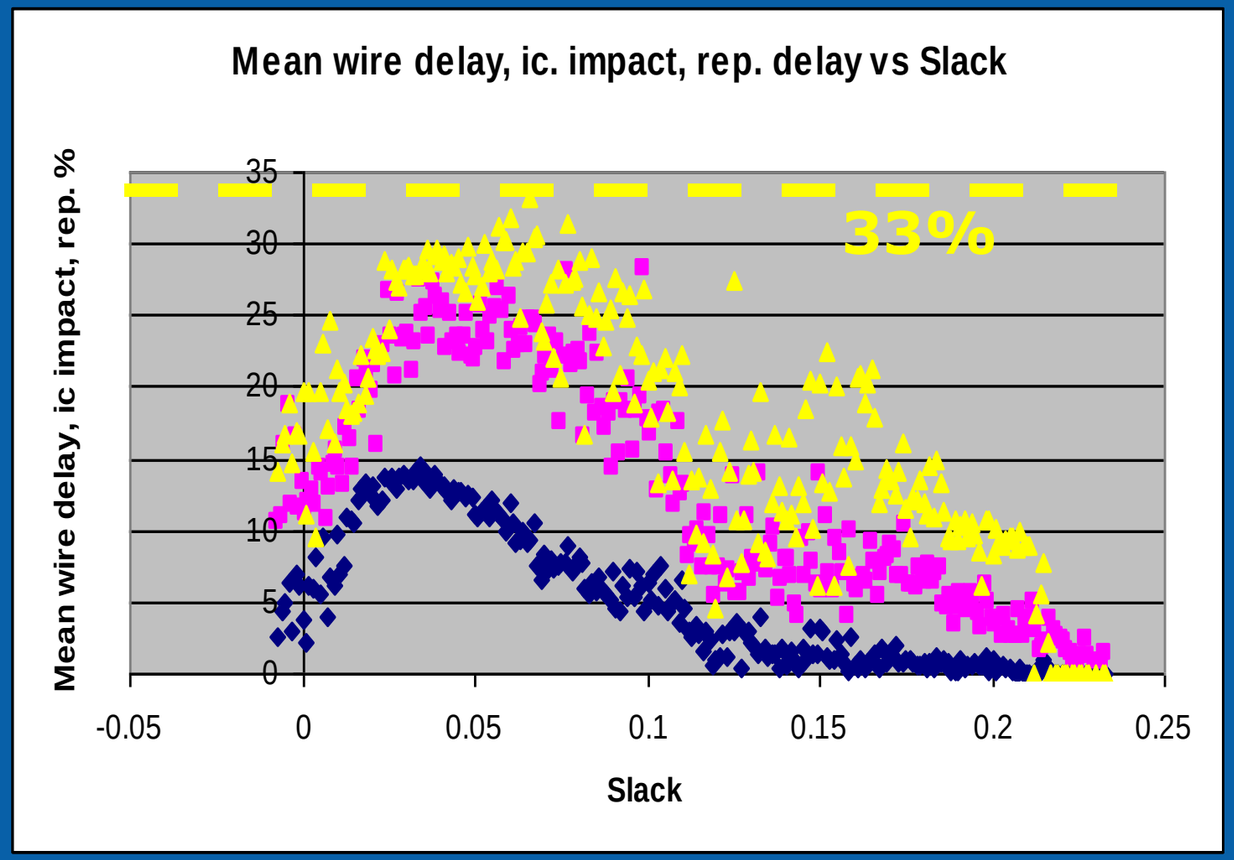
A Close Look: Slope-/Cell-delay Degradation



Mean wire delay & impact vs slack for worst internal paths between unique pair of sequentials

- Slope-/cell-delay degradation contribute as much as wire delay
- Secondary effect not second order

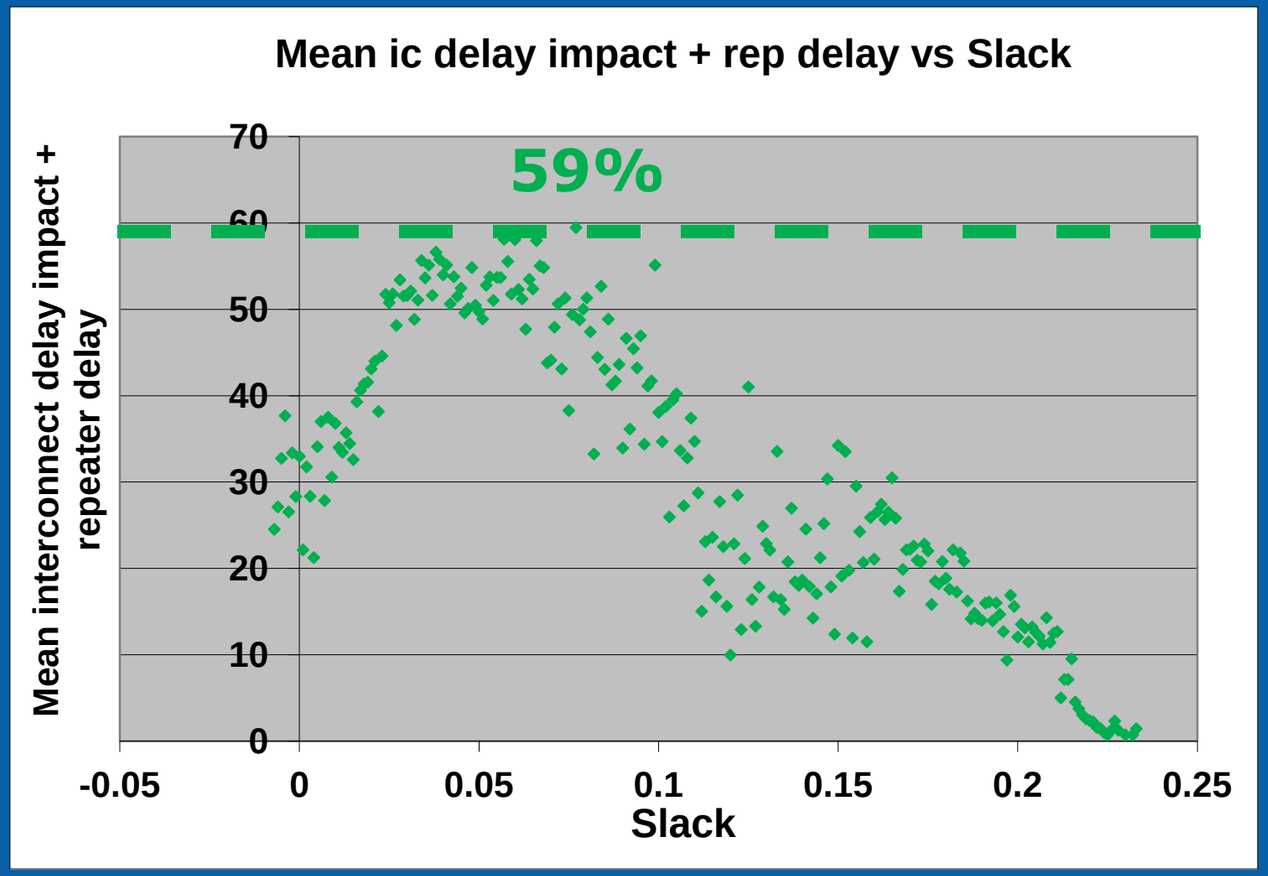
A Close Look: Repeater Delay



Mean wire delay, interconnect impact, repeater delay vs slack for worst internal paths

- Repeater = inverter or buffer
- On critical path, most inverters/buffers are repeaters
  - Cell library is granular
- Repeater delay same as interconnect delay impact

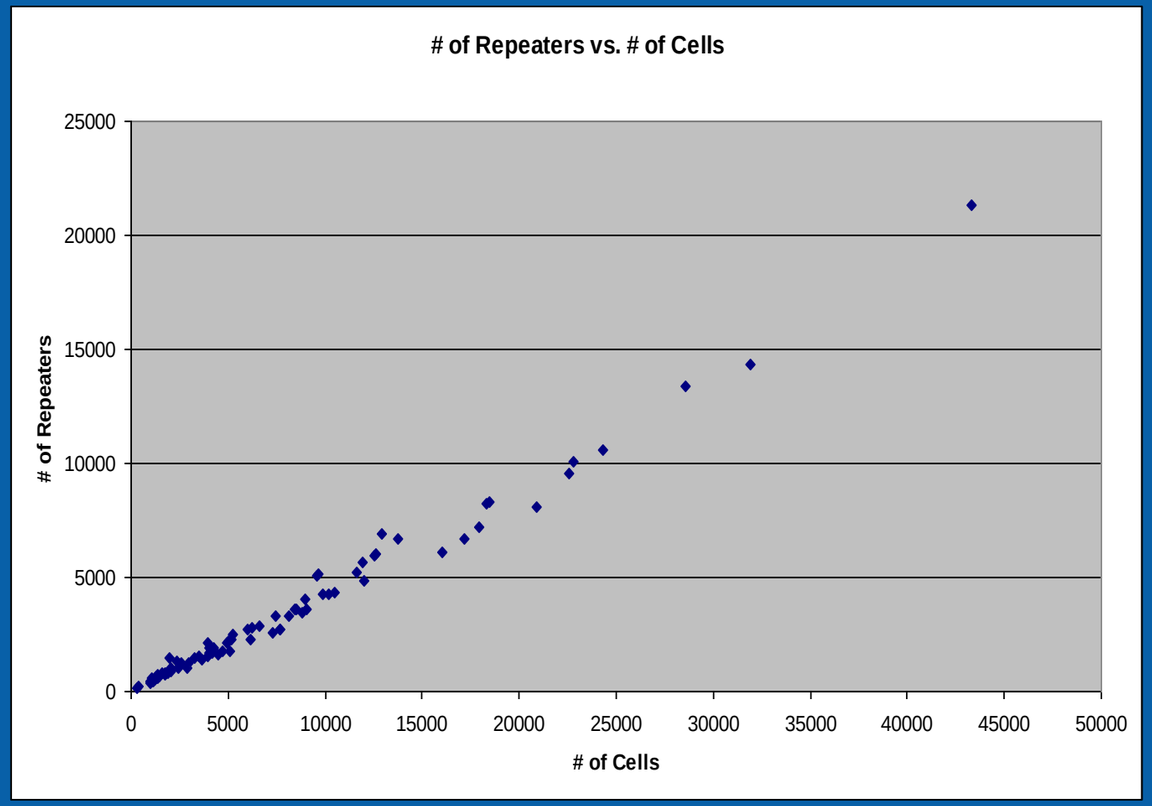
A Close Look: Adding all 3



Overall interconnect delay impact, including repeater delay vs slack for worst internal paths

- Average overall impact: 30%
- Similar behavior for smaller block sizes also
  - Same quality: repeaters are indicators of synthesis quality
- One had hoped for better!

Repeater Count in RLS blocks



- Varies almost linearly with block-size
- Tools/flows used in the linear region

Summary of Observations so far

- Interconnect delay dominance regardless of design style
- Secondary effects as big as primary effect, the wire delay
- Repeater count more than 40% and linear in the size of blocks
- Repeater delays contribute as much as wires

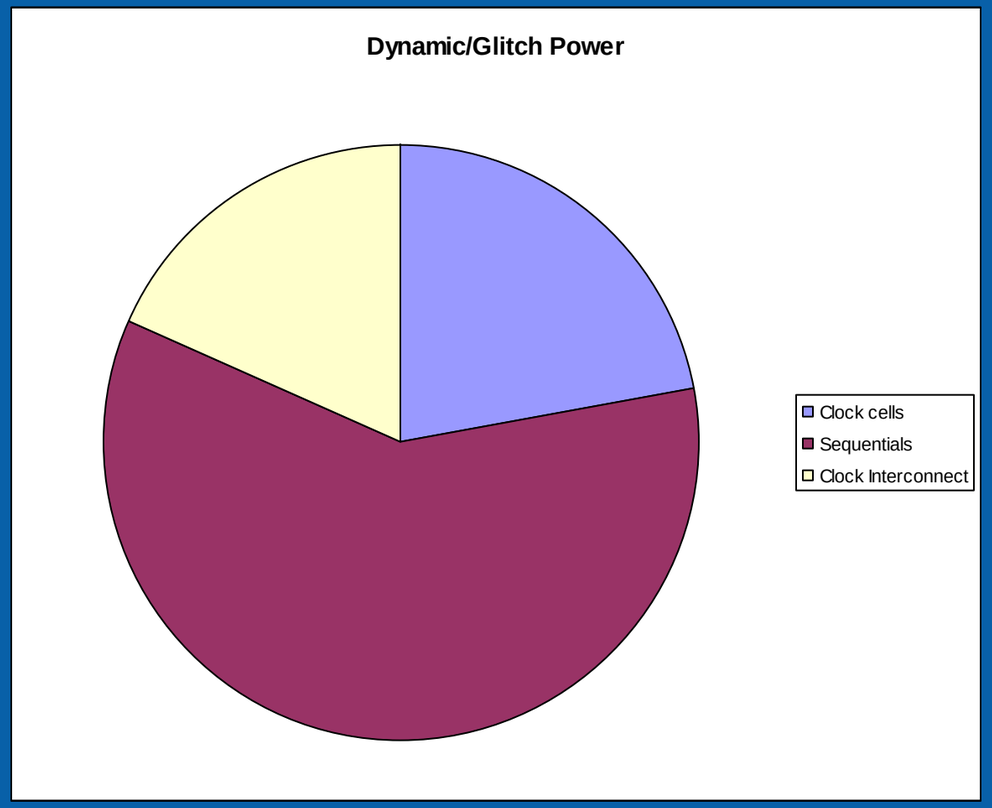
Agenda

- Introduction
- Impact on Timing
- Impact on Power
- Conclusions

Power Dissipation in RLS blocks

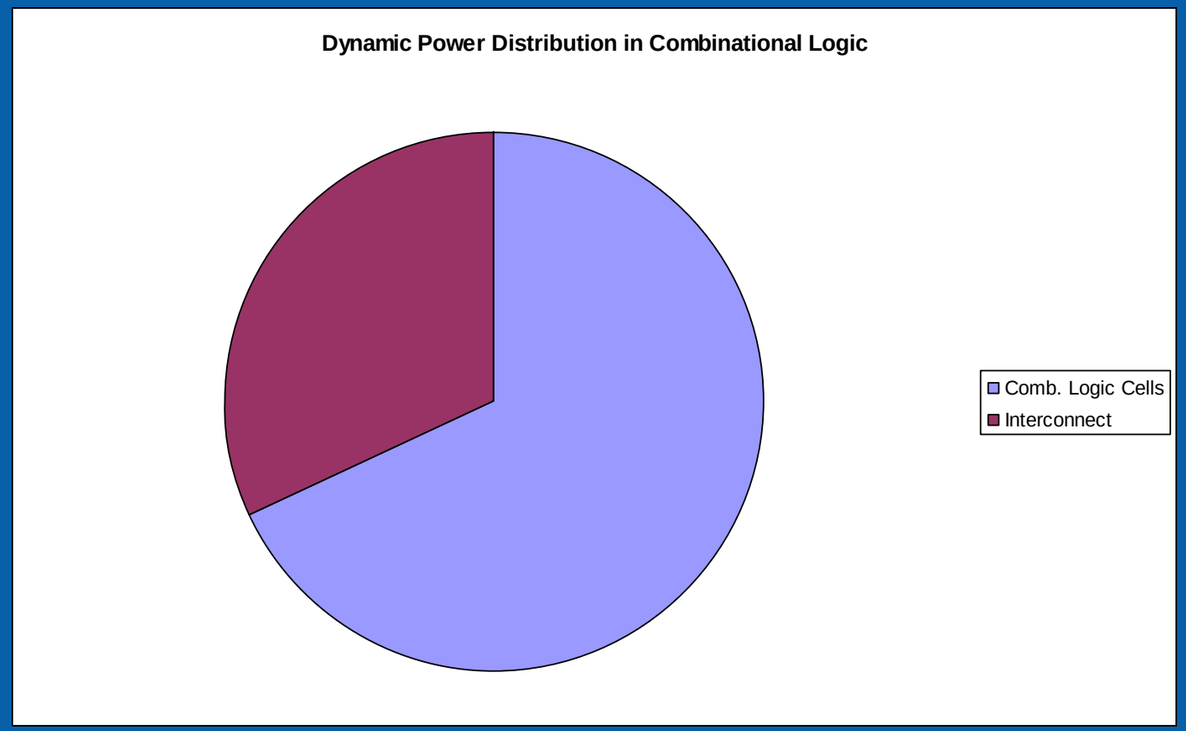
- Typical power dissipation distribution in high speed microprocessors: 60%/10%/30%: Dyn./S. Ckt./Lkg.
- Leakage contained by
  - High-k metal gate transistors with strain
  - High percentage of low-leakage/high-vt devices
  - Power gates
- High use of clock gating reduces the dynamic power in combinational logic
- Synthesized logic blocks consume nearly 30%

Clock Interconnect Power in RLS blocks



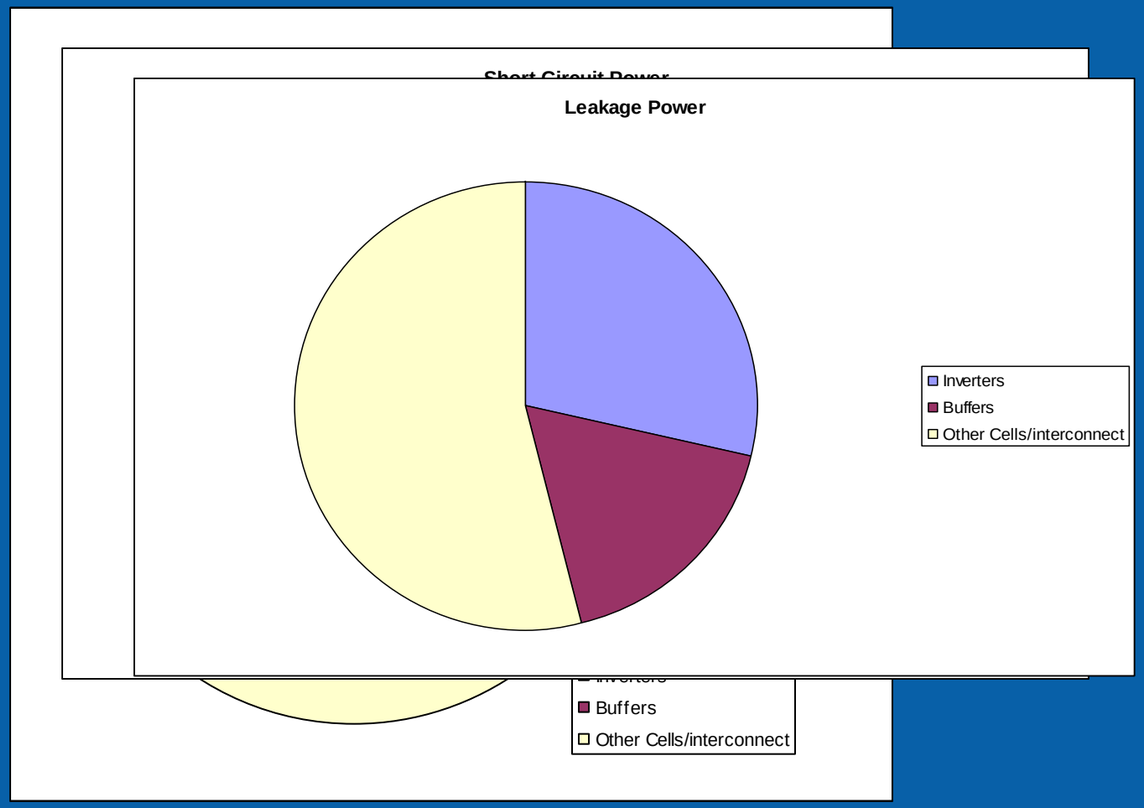
- Interconnects contribute to 18% of dynamic/glitch power in clocks
- Clock tree (including sequentials) contribute to 71% of dynamic power
  - # of sequentials contribute roughly to 1/5th of cell count in RLS
- Out of total dynamic/glitch power in RLS blocks
  - Clock cells contribute 16%
  - Clock interconencts contribute 13%
  - Sequentials contribute 42% of dynamic power in RLS

Interconnect Power in Combinational Logic in RLS blocks



- 32% of dynamic/glitch power in combinational logic; 8% of dynamic/glitch power in RLS

Repeater Power in RLS blocks



- Dynamic power in combinational logic: 27% of dynamic power in RLS
  - Inv./buf. contribute 30% to that; somewhat low, given 44% of cell count, since activity factors for combinational logic are lower than those in clock tree
- SC power in combinational logic: 50% of SC power in RLS
  - Inv./buf. contribute 65% to that; high since no transistors for stacking
- Lkg power in combinational logic: 71% of leakage in RLS
  - Inv./buf. contribute to 46% to that; can be explained by 44% repeater count

Agenda

- Introduction
- Impact on Timing
- Impact on Power
- Conclusions

Impact of Interconnects on Timing/Power

- Avg. impact of interconnect on timing: 30% of cycle time
- Dynamic Power dissipated by interconnects: ~30%
  - ~21% by wires and ~8% by repeaters
- Thus, impact on speed and power: nearly 1/3rd
- Avg. repeater count: 44%
  - Makes layout/timing convergence difficult
- Overall, pose severe challenges to high-speed design

Implications

• [Bohr 95] “Interconnect Scaling – The Real Limiter to High Performance ULSI”

•Would have been true, had we kept doubling the frequency and not moved to Cu

•Pushing speed
  – Microprocessors? Cores already run at 3.2 GHz
  – Processors in netbooks/smartphones
  – Graphics processors

•Technology scaling:
  – Transistors improve; Wire R /um increases; Wire C /um stays the same
  – RC stays the same, assuming ideal length scaling
  – Interconnect impact component likely continue to increase

Possible Solutions

- From technology side:
  - 3 D?
  - Al → Cu → ? Low k?
  - Not in sight for next few years?
- From CAD:
  - Placement, routing, physical synthesis running out of steam
    - "don't know what the opportunities are" – ISPD 2010
  - Logic synthesis/tech. mapping doesn't help, where it is used: serves the purpose of creating a netlist from RTL
    - "The Death of Logic Synthesis" – ISPD 2005
- How about incremental logic re-synthesis after global routing

Logic Re-Synthesis After Global Routing

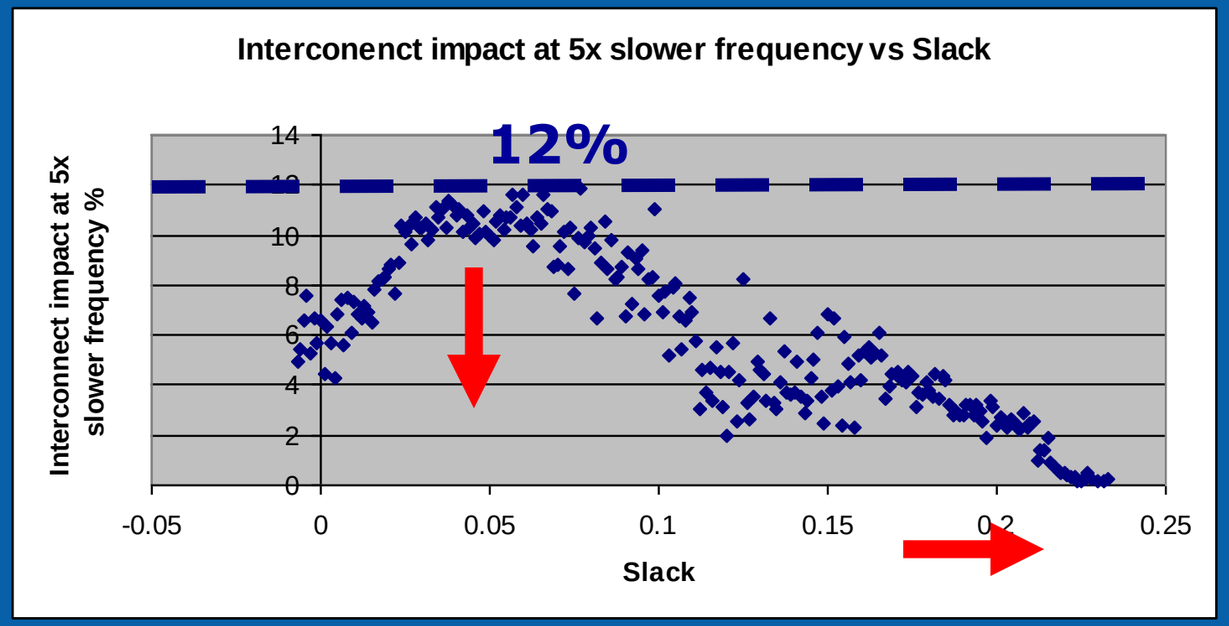
- Why?
  - Routing picture known after placement/CTS/global route
  - Only then we know the real impact of interconnects on delay
    - Dependence on topology, layers, vias, repeaters, detours, congestion
  - Logic synthesis/technology mapping powerful transformations, but…
- Challenges:
  - Using placement/routing information
  - Requires more memory/computation: faster/better/multi-core CPUs
  - Polynomial time algorithms performing simultaneous optimizations
    - An example: simultaneous mapping/placement

Acknowledgments

- Marek Patyra, Intel
- Noel Menezes, Intel
- Xinning Wang, Intel
- Wei-kai Shih, Intel
- Andy Carle, Intel
- … many from EMG/TMG, Intel

Q&A

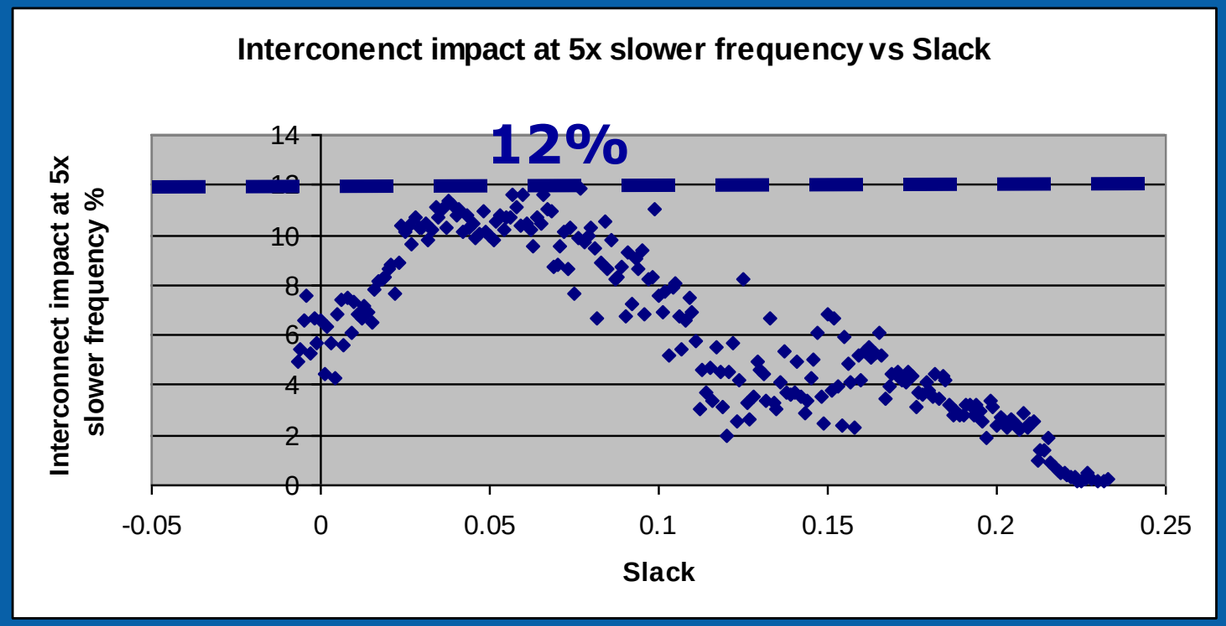
Low Frequency (high 100s of MHz)/Low Power Designs



Projected* interconnect delay impact for 5x slower design (could be much lower)

- Processor running at 5X slower frequency consumes 5x lower dynamic power
  - Interconnect delay impact as percentage of cycle time reduces by same factor
- Additional quadratic power savings due to supply voltage reduction
  - Slower gates, but interconnect component stays roughly the same
  - Overall interconnect impact on delay goes down further
  - Doesn't require as many repeaters
  - Critical paths gate-delay dominated

Low Frequency (high 100s of MHz)/Low Power Designs



Projected* interconnect delay impact for 5x slower design (could be much lower)

- Effect of re-pipelining on delay
  - Less sequentials ➔ Less clock buffers/nets ➔ More routing resources for signals ➔ Better routing ➔ Lower interconnect impact
- Problems for low power/high speed not the same!
- 1 Million cell placement for 600 MHz != 200 K cell placement for 3 GHz
- What if we want to run a processor in both the modes