Interconnection technologies

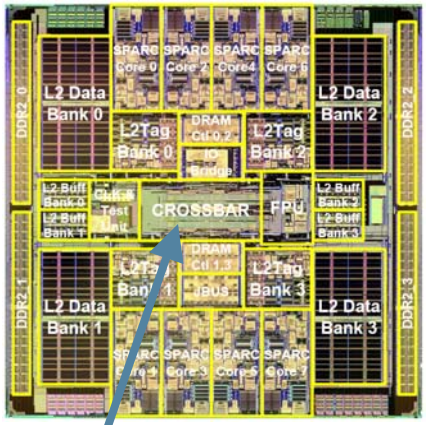
Ron Ho
VLSI Research Group
Sun Microsystems Laboratories

Acknowledgements

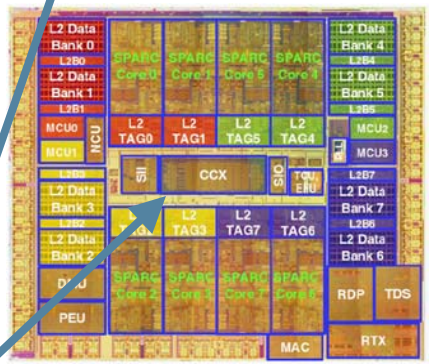
- Many contributors to the work described here
  - Robert Drost, David Hopkins, Alex Chow, Tarik Ono, Jo Ebergen
  - … and the entire Sun Labs VLSI Research Group
  - Danny Cohen

Multicore chips today

- Niagara: 8 cores
  - > 32 threads (4 threads/core), 1 shared FP
  - > 90nm CMOS, 1.2GHz, 63W, 380mm^2
  - > High-volume production now



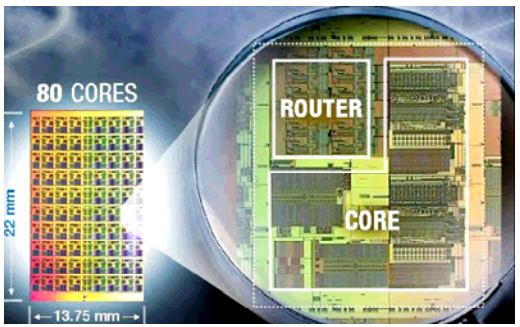
- Niagara2: 8 (improved) cores
  - > 64 threads (8 threads/core), 1 FP per core
  - > 65nm CMOS, 342mm^2
  - > Shipping systems in 2H07



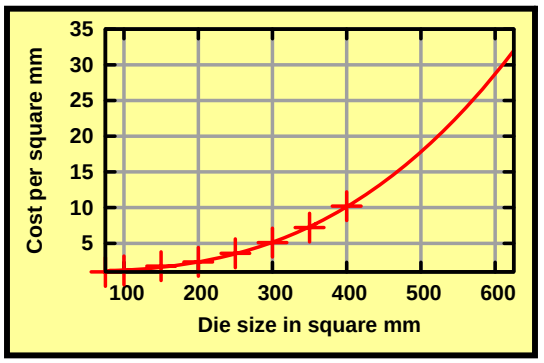
Interconnect networks: an 8x9 xbar

Multicore chips tomorrow?

- How do you scale up to tens or hundreds of cores?
  - …Assuming you want to (more total power for high performance)
  - …And on a single chip: avoid the costs of chip-to-chip IO
    - Bumps are big (180μm), links are power-hungry (20pJ/bit = 20mW/Gbps)
- You need to do some combination of:
  - Making the cores very small
    - But you lose functionality: what kind of programming models do you want?
  - Making the die very big
    - But you lose the yield/cost battle very quickly
  -



Source: 2006 IDF

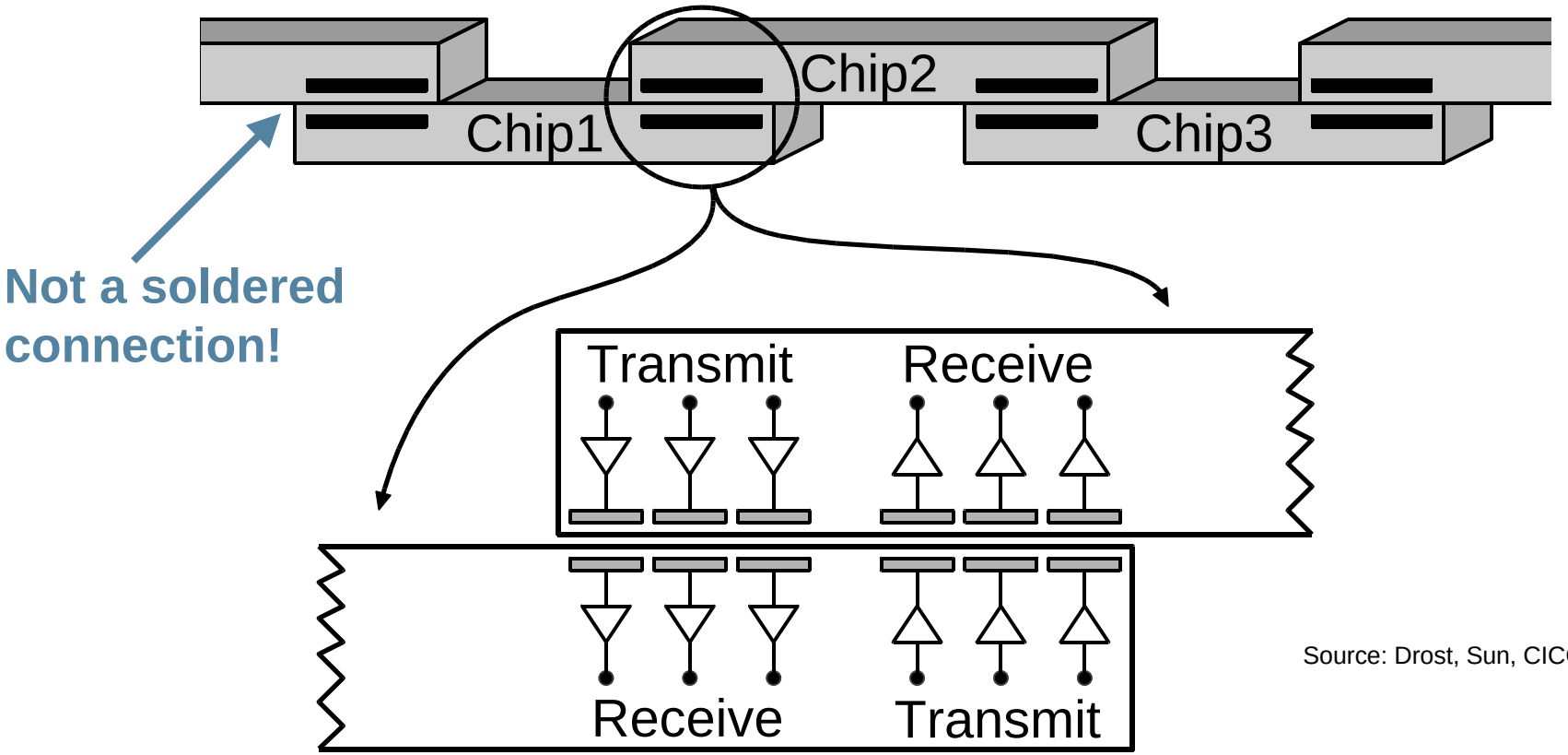


A different direction…

- We can rethink the “single-chip” requirement…
  - If we reduce (eliminate) the costs of chip-to-chip communication
    - Power, area (bandwidth density), latency, known-good-die
- Break a multi-core chip into a many-chip-system
  - Smaller chips lead to higher yields and lower cost
  - Different chips lead to system adaptability and reconfigurability
  - Aggregate systems of chips effectively break the reticle limit
- Enable a broad set of interconnection explorations

Proximity I/O

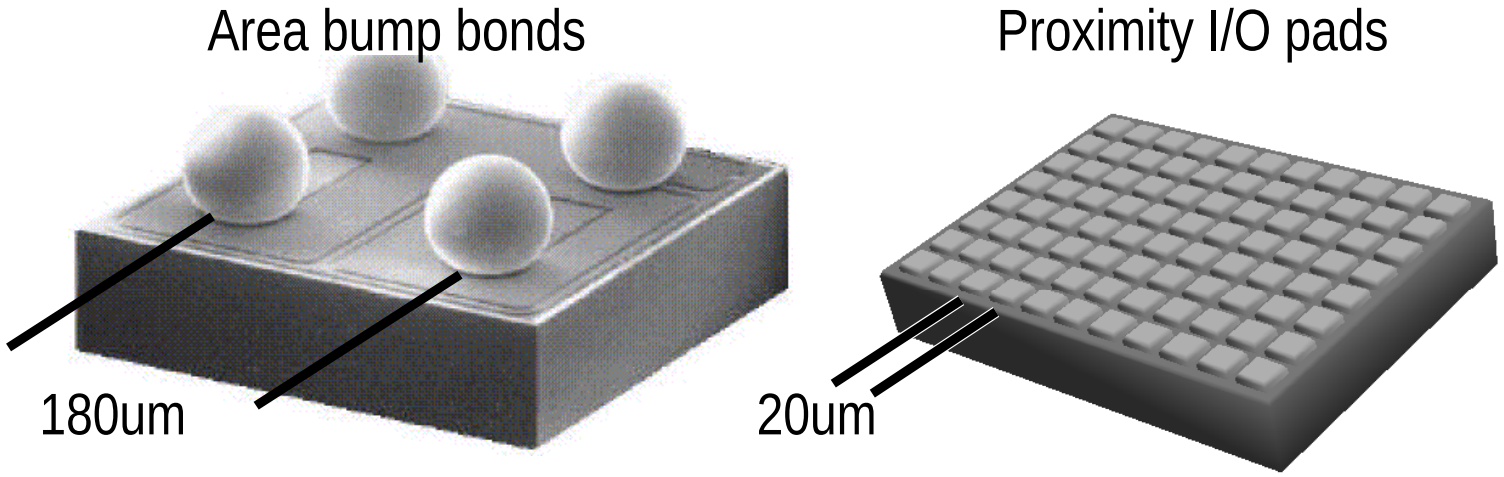
- Pioneered by Ivan Sutherland at Sun Labs
- The big idea:



Source: Drost, Sun, CICC 2003

Proximity I/O benefits

- Bandwidth density 60x-100x greater than balls

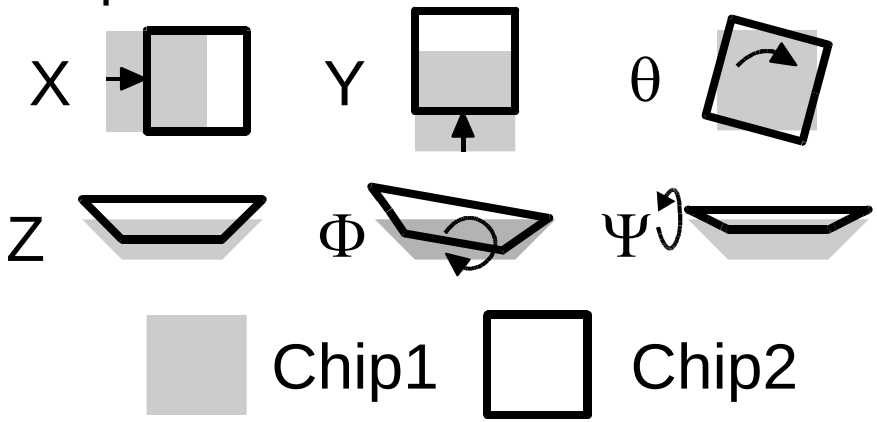


Source: Drost, Sun, CICC 2003

- Much lower power
  - No ESD required
  - Use wide parallel links instead of narrow SerDes links
  - Very small Tx and Rx circuits

Proximity I/O challenges

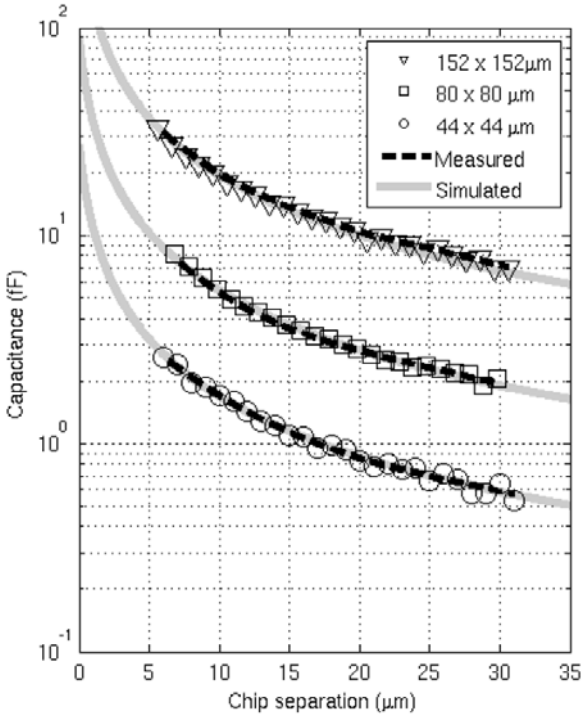
- Misalignment is the proverbial monkey's wrench
  - > Initial imperfect chip placement
  - > Dynamic chip movement from vibration or thermal expansion

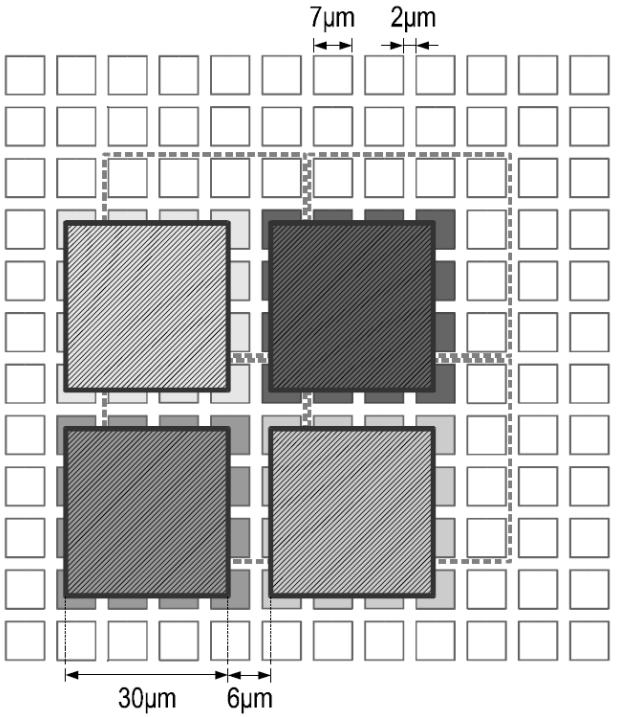


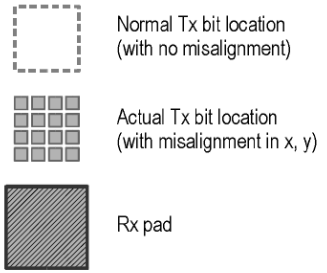
- A robust solution is (at least) two-pronged
  - > Combination of specialized packaging and custom electronics
  - > Here, discuss some of the electronic/circuit strategy

Fixing misalignment

- Detect misalignment using Vernier-like arrays
  - > Measure capacitance between chips to sub-fF resolution
- Correct in-plane misalignment using data steering



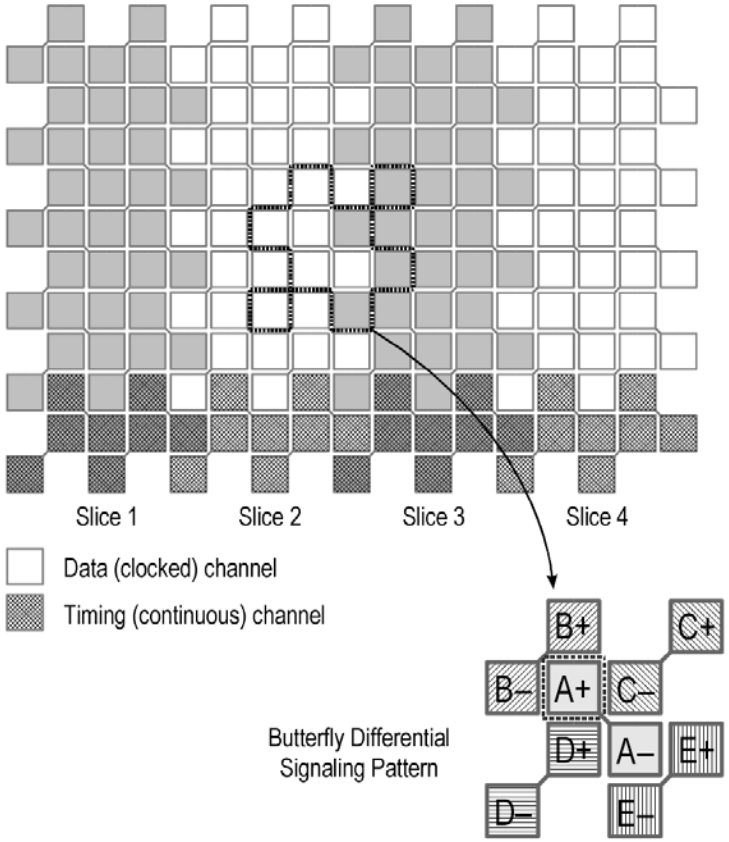


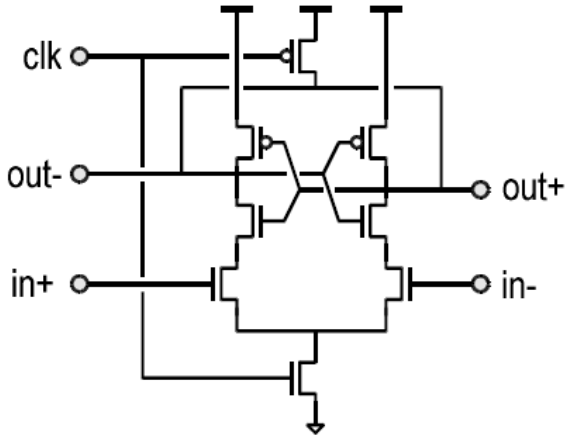


Source: Hopkins, Sun, ISSCC 2007

Dealing with noise

- Receivers are differential sense-amplifiers
  - > “Butterfly” scheme rejects noise (receivers are offset-limited)
  - > We employ a clock—it uses unclocked receivers with larger pads



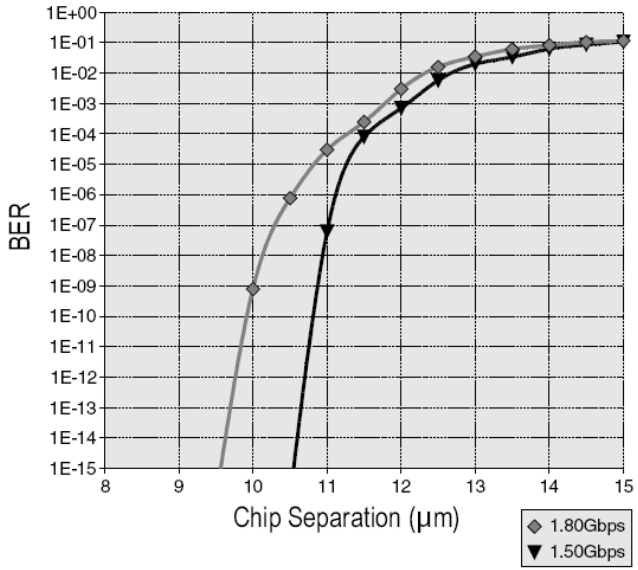


Data channel receiver
(clocked)

Source: Hopkins, Sun, ISSCC 2007

Silicon measurements

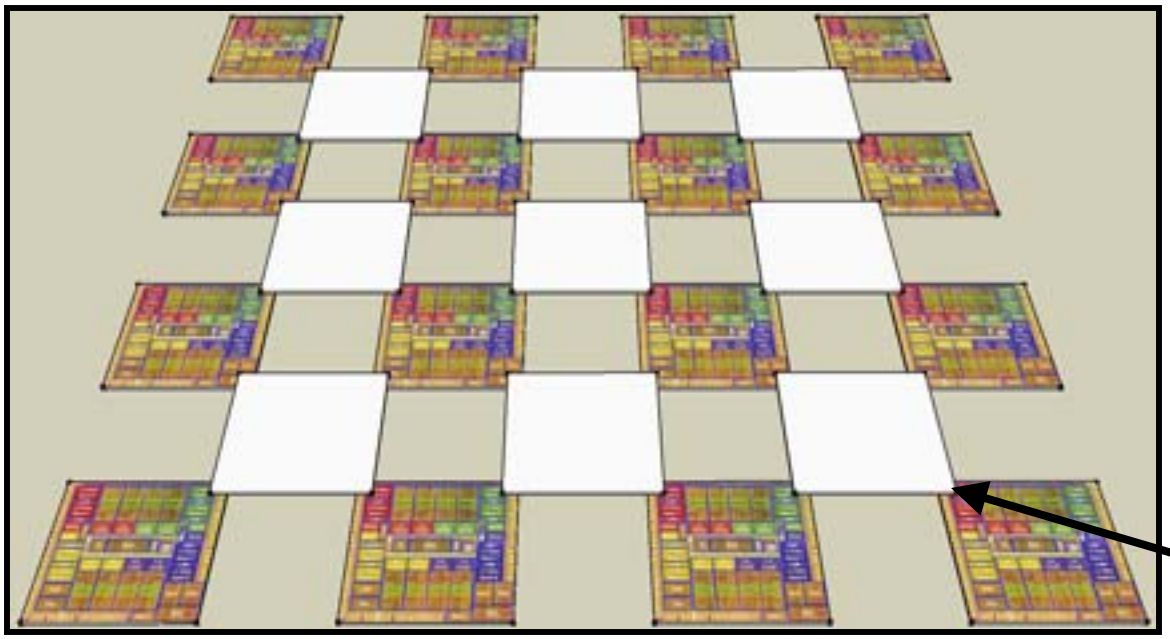
- 144b Proximity I/O datapath, 180nm TSMC chip
  - 1.8Gbps per pin for 260Gbps in 0.5mm^2
  - Measured BER<10^{-15}
  - 3pJ/bit at a 1.8V power supply
  - > 0.7UI timing margin, > 200mV voltage margin at speed
  - Measured sensitivity to chip separation

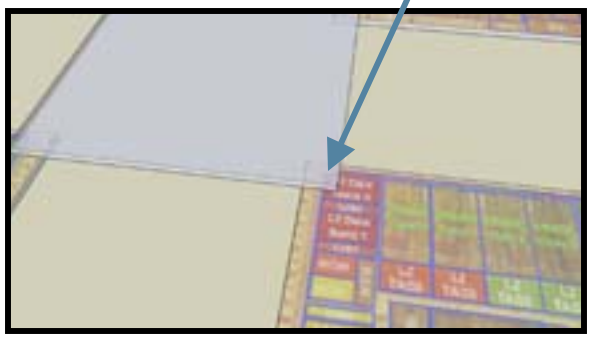


Source: Hopkins, Sun, ISSCC 2007

Proximity I/O as an enabler

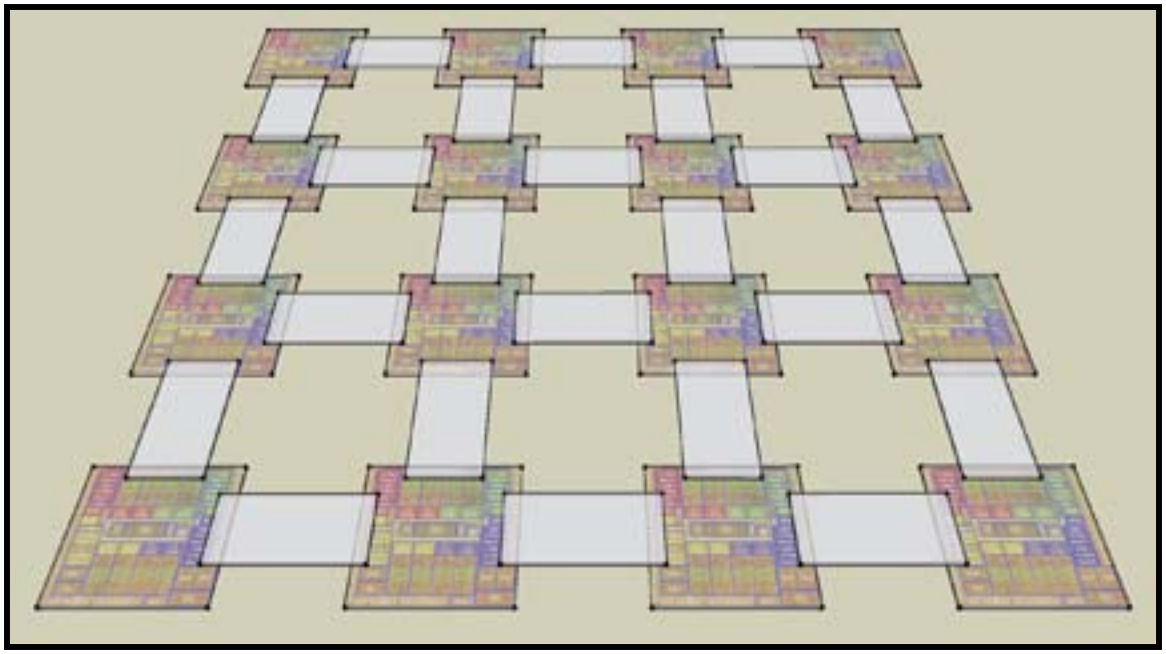
- Low-cost chip-to-chip communication
  - > Off-chip I/O looks like an extension of on-chip wires
  - > Many chips look like a (big) single chip
  - > What kinds of interconnect networks should we consider?





Proximity I/O-based grids

- Another, perhaps more realistic grid
  - All big (high-power) chips are all face-up on a cold plate
  - Face-down chips merely transmit data

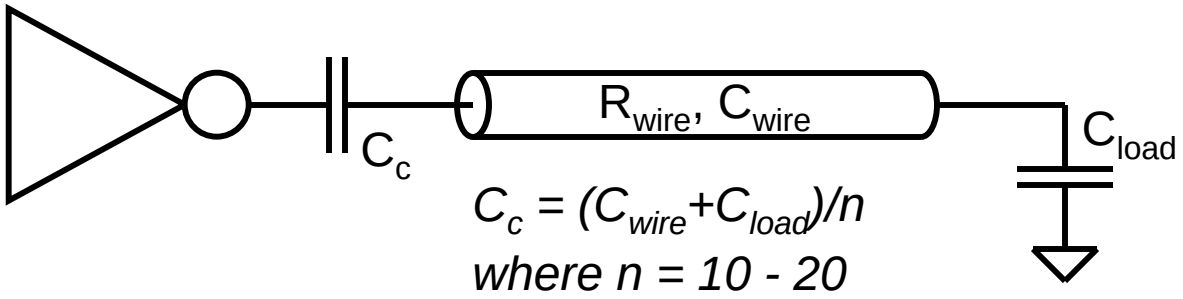


- Natural extension to various network topologies

A red flag?

- Such a system will have lots of VLSI wires
  - The entire interconnection network consists of on-chip wires
- Latency and bandwidth characteristics well-known
  - …And so are the energy costs: $E=C*V*\Delta V$
    - Cap is about 0.45pF/mm (incl. repeaters), and not really scaling
    - Voltage is about 1V, and not really scaling
  - Therefore, energy is about 0.45pJ/bit/mm of linear length
  - So 100 64b buses at 4GHz, 10% activity over 20mm: 24W!
- Need an efficient wiring system

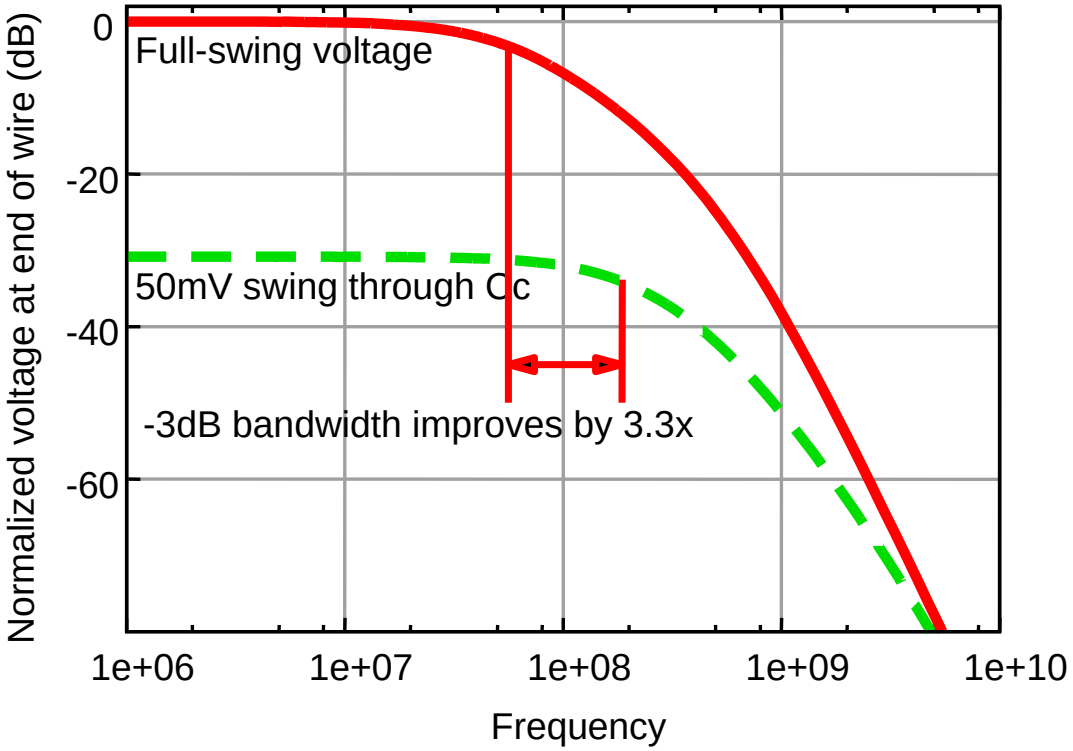
Efficient capacitively-driven wires



- Reduced power: low voltage swing on wires
  - Swing is $V_{dd}/(n+1)$ without requiring a second power supply
- Reduced power: small load seen by driver
  - Allows use of a 1/n-th sized driver, saving power and area
- Reduced latency: capacitor pre-emphasizes edges
  - Also, charge distribution effectively cuts wire delay in half

Pre-emphasis extends bandwidth

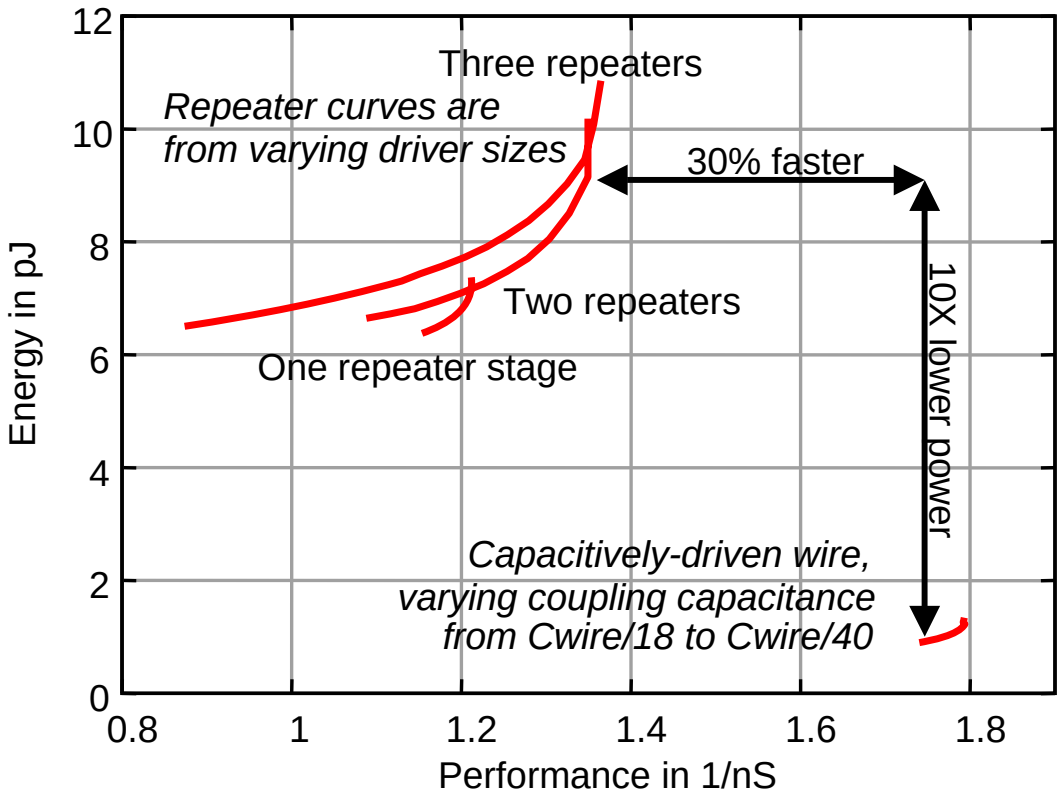
- The inline capacitor acts as a high-frequency short
  - Example of a 14mm minimum width wire (simulated)
    - Bandwidth RC-limited to 50MHz
    - Capacitor (of 1/20th total capacitance) extends it to 180MHz



Source: Ho, Sun, ISSCC 2007

Making a better repeater

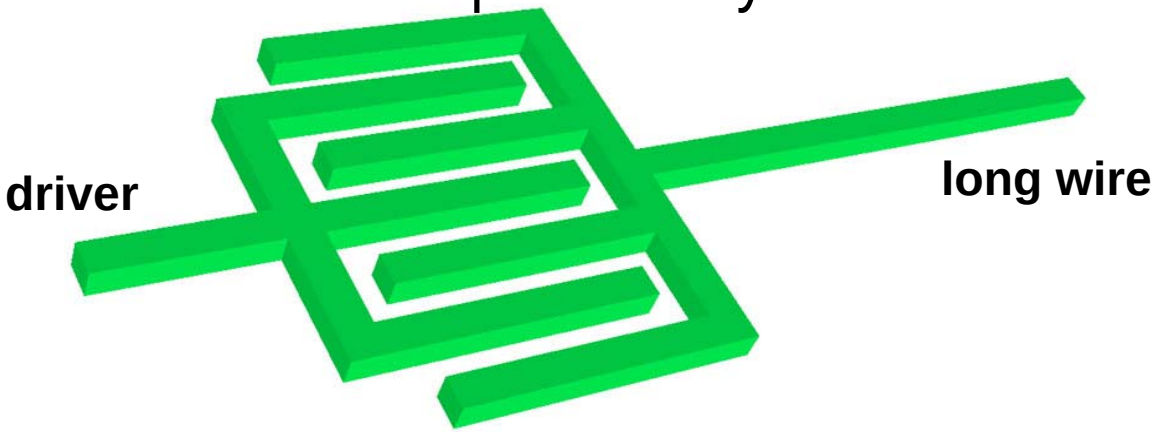
- Analysis of energy-efficiency shows benefits
  - Compare against repeaters for a 10mm wire in a 180nm process



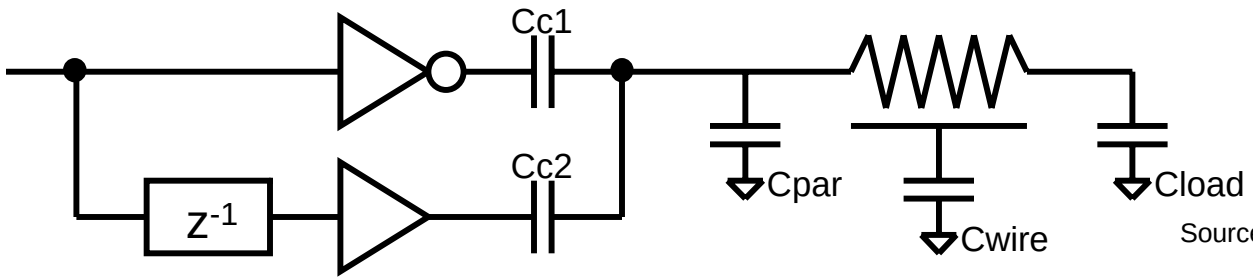
Source: Ho, Sun, ISSCC 2007

Building pitchfork capacitors

- Exploit a “problem” of wires: sidewall capacitance
  - Can use more tines or multiple wire layers



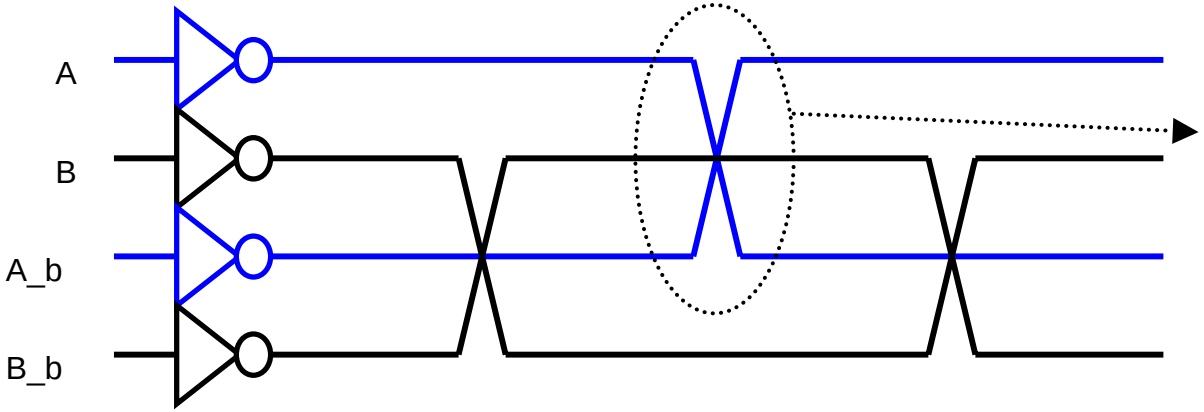
- Coupling caps ideal for summing junctions
  - Use them to build a cheap FIR filter

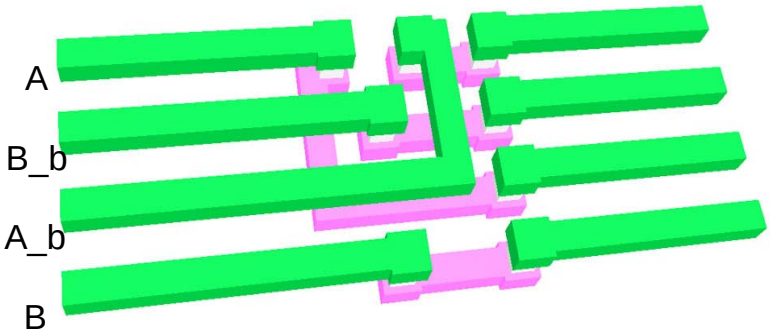


Source: Ho, Sun, ISSCC 2007

Some costs

- Differential wires cost area, power; twisted for noise



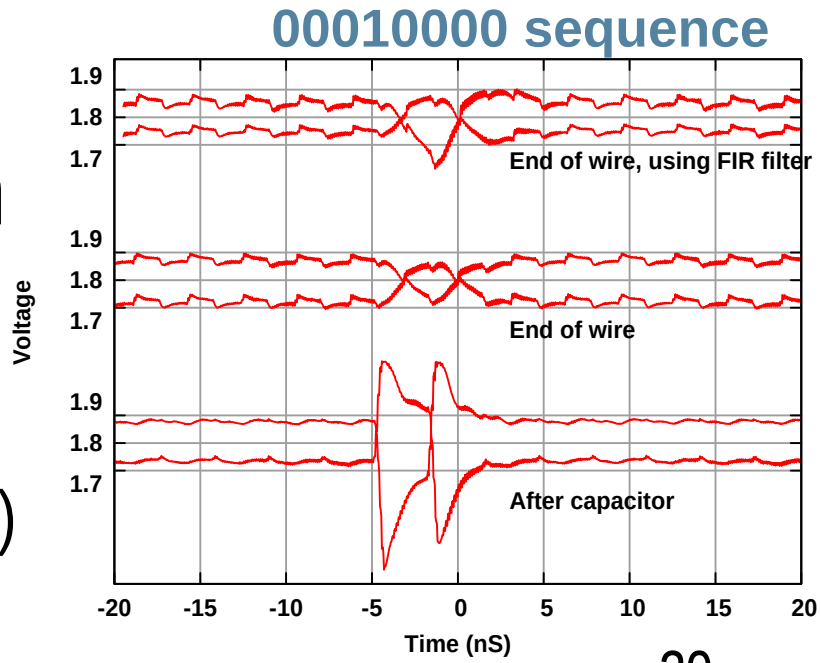


- Sense-amps have offset and biasing requirements
  - Biasing can use refresh, with (n+1) channels for an n-bit bus

- Verification requires some care
  - Correctness comes from being near, not from being connected!

Silicon measurements

- Multiple datapaths on a 180nm TSMC 1.8V chip
  - > Measured 10x less energy at a 50mV swing (max. savings 18x)
  - > Measured 4x bandwidth extension from pre-emphasis
  - > Bit error rates < 10^{-11} (limited by test time)
  - > 50% UI eye opening on each experiment
- 14mm, unrepeated, min.width
  - > RC-limited to 55MHz
  - > Capacitor pushes it to 200MHz
  - > Sub-optimal 2-tap FIR (wrong delay)



Source: Ho, Sun, ISSCC 2007

Look at new system architectures

- A pair of complementary enabling technologies
  - Multi-chip grids connected using Proximity I/O and efficient wires
  - Chip-to-chip latency, bandwidth, power equal to on-chip wires
  - Long on-chip wires can be lower power and higher performance
- Multi-chip grids look like a big “virtual” chip
  - With (re)configurability, cost, and “reticle-breaking” benefits
- A question: how to best interconnect these chips?
  - Or: how to best arrange these chips for an interconnect network?

Interconnection technologies

Ron Ho, Ph.D.
ron.ho@sun.com