A Wave-Pipelined On-Chip Interconnect Structure for Networks-on-Chips

Jiang Xu and Wayne Wolf
Dept. of ELE Princeton University
Aug. 2003

Outline

- Challenges in the Near Future
- Networks-on-Chip
- Wave Pipeline
- Design Description
- Simulation Results
- Design Methodology
- Conclusions
- Future Work

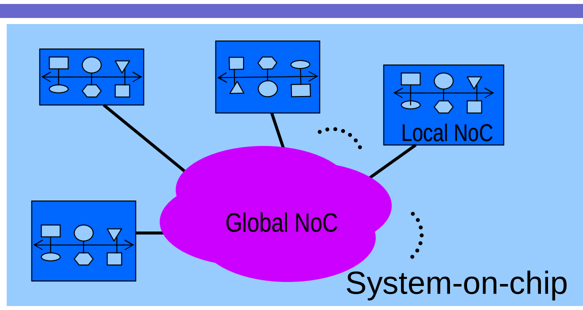
Challenges in the Near Future

ITRS (International Technology Roadmap for Semiconductors) 2002

Year	2003	2004	2005	2006	2007	2010
Process (nm)	107	90	80	70	65	45
Clock (MHz)	3,088	3,990	5,173	5,631	6,739	11,511
Chip size (mm^2)	572	572	572	572	572	572
Power voltage (V)	1.0	1.0	0.9	0.9	0.7	0.6
Transistor (M)	810	1020	1286	1620	2041	4081

- Global interconnect becomes bottleneck
  - Multiple cycles to cross a chip
- Crosstalk noise
- Multi-billion transistors on a single chip
- Interconnect defect decreases yield

Networks-on-Chip

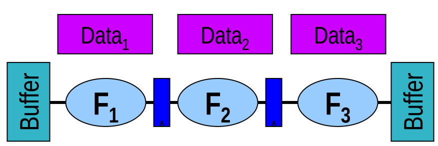


- Physical level
  - A collection of physical interconnects [CON01]
- Architecture level
  - Links between function units [DAL01]
- System level
  - Communications between processes [SAN00]

"On-Chip": Requirements for NoC

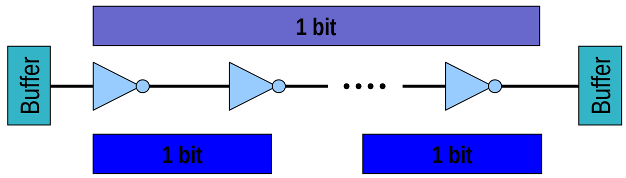
- Working at the chip frequency (GHz) and low latency
  - Connect processor cores to memory and other "on-chip" function units
- Full or partial parallel (32~256bit)
  - Reduce latency, relatively large amount of wiring resource
- Low power and small metal area
  - A lot of power is used to drive on-chip interconnect
- Globally asynchronous and locally synchronous

Wave Pipeline (I)



- L. Cotton introduced wave pipeline in 1969
- Compute, transmit, and store at the same time
  - "Maximum pipelining"
- Functional block, memory, processor
  - Rambus
- Simple, but difficult to design
  - Multiple data paths

Wave Pipeline (II)

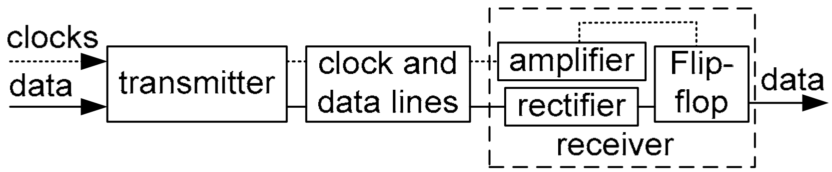


- More efficient than latches
  - Transistor size, power, and latency
- Regular structure simplifies implementation
  - Identical data paths
- Well suited to globally asynchronous design

Outline

- Challenges in the Near Future
- Networks-on-Chip
- Wave Pipeline
- Design Description
- Simulation Results
- Design Methodology
- Conclusions
- Future Work

Design Overview



- TSMC 0.25µm Al
- 10mm interconnect
- 1.2µm wide in 3.2µm pitch [KAH98]
- Bidirectional 16-bit interconnect
- Explicit clock signals without using PLL
  - Low latency
- Cadence Spectre

Delay Analysis

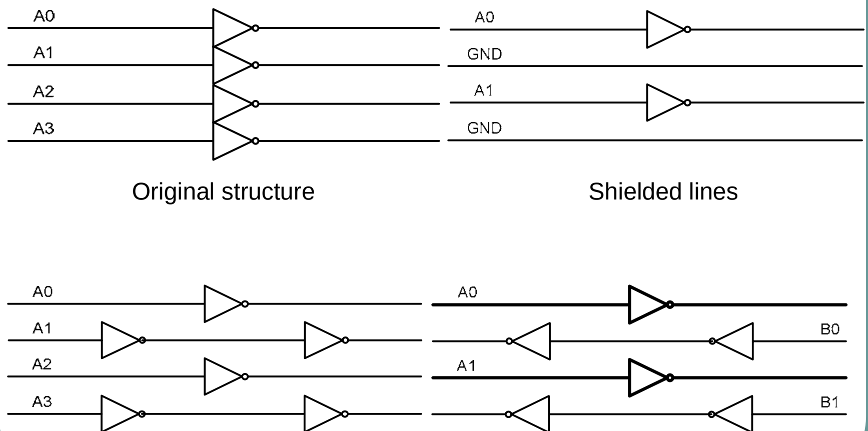
- Delay variation limits the performance

$$T_c > (D_{max} - D_{min}) + 2\Delta + T_s + T_h$$

T_c --- clock period
D_{max} --- worst-case delay
D_{min} --- best-case delay
Δ --- one-side clock skew
T_s --- setup time
T_h --- hold time

- Crosstalk causes delay variation

Buffering Alternatives



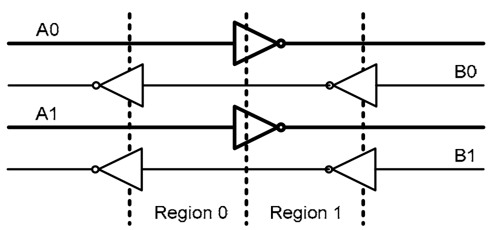
Original structure

Shielded lines

Inverter offset

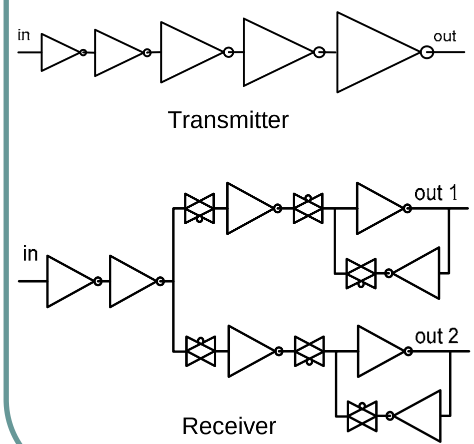
Interleaved lines

Interleaved Lines



- Up-link and down-link are not active at the same time
  - Shielded lines
- Active at the same time
  - Inverter offset
- Reduce delay variation from 140ps to 29ps (79%)
- Clock lines have the same structure as data lines
  - Match delay in design and fabrication

Transmitter and Receiver



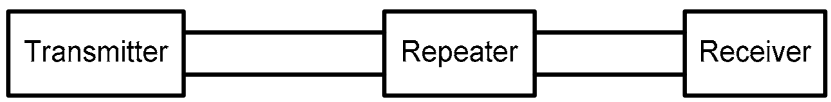
- Transmitter
  - A chain of inverters
- Rectify signal
- Dual flip-flop
  - Work at different clock edges
  - 3.45GHz at 0.25μm technology
- Simplify the first stage
  - Improve sensitivity and response time
  - Add extra clock edge

Clock and Signaling

- Two opposite phase clock signals
  - Half of the data frequency
  - Clock lines are the critical paths
    - Most switching activity
    - Clock driving power, latency
- Clock signals are only sent with data
  - Embedded with control signals
  - Save power

Architecture Decisions

- Explicit clock distribution
- Voltage mode signaling

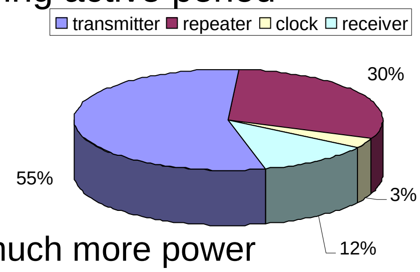


Simulation Results (I)

- 3.45GHz input data
  - 290ps wide with 20ps rising and falling time
  - 55.2Gbps for a 16-bit link
- 50% delay time
  - 1167ps for up-link and 1290ps for down-link
- Effective pipeline length
  - 4.4 pipeline stages for down-link
- Delay variation is 29ps

Power and Area

- 1039mW for up-link during active period
  - 18.8 pJ/bit
- Clocks use 12%
  - save power when idle
- Repeaters use 30%
  - Using latches needs much more power
- Transmitters $0.067mm^2$, receivers $0.012mm^2$
  - ARM9E-S $2.7mm^2$ and PowerPC405 $2.0mm^2$



Conclusions

- High throughput in 0.25µm technology
- Small area and low power compare using latches
- New interconnect structure reduces crosstalk
- Asynchronous mode without using PLL

Future Work

- Optimize the inverter offset
- Wave-pipelined asynchronous bus
  - Different pipeline stages on the same bus
- Use it in the Smart Camera SoC
- Fabricate and test the wave-pipelined interconnect

Reference

[BUR98] Wayne P. Burleson, Maciej Ciesielski, Fabian Klass, Wentai Liu, "Wave-pipelining: A tutorial and research survey", IEEE Transactions on VLSI systems, 6(3): 464-474, 1998.

[CON01] Jason Cong, “An interconnect-centric design flow for nanometer technologies”, Proceedings of the IEEE, 89(4): 505-528, 2001.

[DAL01] William J. Dally, Brian Towles, “Route packets, not wires: on-chip interconnection networks”, Proceedings of the 38th Design Automation Conference, 2001.

[KAH98] Andrew B. Kahng, Sudhakar Muddu, Egino Sarto, Rahul Sharma, “Interconnect Tuning Strategies for High-Performance ICs”, Proceedings of DATE ’98.

[SAN00] A. Sangiovanni-Vincentelli, M. Sgroi, L. Lavagno, “Formal models for communication-based design”, Proceedings of CONCUR 2000, 29-41, 2000.