Routing Chip Set for Intel Paragon™ Parallel Supercomputer

Roger Traylor
Dave Dunning
Intel Corporation

- The Paragon System
- Paragon Network Fabric
- Network Interface Chip (NIC)
- High Speed Signaling
- Mesh Routing Chip (MRC)
- Summary
The Paragon System

A scalable parallel supercomputer

5-300 MFLOPS
66-4096 processors
Internode communication bandwidth > 200 MBytes/s

Paragon Network Fabric

- **Topology**
  - 2D Mesh
  - All data channels 16 bits wide
  - Bidirectional 4-way transfer
  - Interface to processor via NIC

- **Performance**
  - All channels > 200 MBytes/s
  - Bisection bandwidth 12.8 GBytes/s max.
Paragon Network Fabric

- 2D Mesh Motivation
  - Physically easy to build
  - Easy to expand
  - Proven in Touchstone Delta
  - Short, point-to-point electrical connections
  - Fast

- Self-timed logic
  - Precludes all high speed clock distribution issues
  - Scalable to any practical size mesh

Network Interface Chip (NIC)

- What is it?
  - Interface between nodeboard data bus and MRC
  - Data funnel
    - 64 bits <---> 16 bits
  - Protocol conversion
    - Synchronous <---> Self-timed
  - Data integrity via CRC and parity
  - Rate buffering via FIFO buffers
Network Interface Chip (NIC)

Where does it fit?

Network Interface Chip (NIC)

A look under the hood
Network Interface Chip (NIC)

A look under the hood

Network Interface Chip (NIC)

Synchronous <-> Self-timed Interfaces

- Use FIFO flags as basis of interface
- Synchronizing flags ("Doctor, it hurts when I do this ...")
  - Use clean flags (grey code counters)
  - Synchronize infrequently (only at boundaries)
  - Synchronize thoroughly with fast flip flops

intel®
Network Interface Chip (NIC)

Self-timed Pipeline Methodology
- Internally, 2 cycle interlocked handshake
- Externally, interlocked or streaming

Request  
Request  
Request  
Data    
Data    
Data

Acknowledge  
Acknowledge  
Acknowledge  
Pipe 0  
Pipe 1  
Pipe 1

Request and data are "bundled" in time

Request  
Request  
Data

Data  
Data  
Data

Acknowledge  
Acknowledge  
Acknowledge  
F

intel®
Network Interface Chip (NIC)

Self-timed Pipeline Methodology (cont.)

- **Physical implementation**
  - User defined hard macros and careful placement maintain isochronous regions
- **Performance**
  - 72 bit pipelines can run at 1.2 GBytes/s
  - External streaming mode
    > 300 MBytes/s
Network Interface Chip (NIC)

Process and Technology

- 1.0 mm CMOS standard cell
- 299 pin CPGA
- 3W max
- Rail-to-rail I/O switching
- Die 15 x 15 mm

High Speed Signaling

- 75 Ω traces/wires were chosen
  - Good match for CMOS
  - Lower power required
- Multiwire circuit boards
  - Good wire length matching, minimal signal skews
  - Very tight impedance control
- NIC signals are source terminated with discrete resistors
- MRC signal impedances matched by tunable strength output drivers
High Speed Signaling

Impedance Tunable Output Drivers

- MRC is Pad limited, large Pads do not affect die size
- PBIAS, NBIAS control pull up and pull down strengths respectively

Mesh Routing Chip (MRC)

Physical Description

- 325 pin ceramic PGA
  - Interstitial pins, 70 mil. spaced
  - 1.75 inches per side
  - 0.8 mm Intel process
  - Full custom die, 320 mils/side
  - 2 watts max.
- Completely self-timed component
Mesh Routing Chip (MRC)

Architecture

- 5 input ports, 5 output ports
- Each port contains 16 data bits, 2 parity bits, 1 tail, 1 request, 1 acknowledge
- Displacement based addressing
- Routes X before Y
- Hardware broadcast in rows, columns or rectangles
- Some diagnostic/error detection capabilities

Mesh Routing Chip (MRC)

Block Diagram

B = Broadcast cell, A = Arbiter cell, S = Stripper cell, D = Decrementer cell
Mesh Routing Chip (MRC)

Block Diagram Description

- **5 major functional blocks**
  - Broadcast cells: Allow messages to select a path or fork in multiple directions (broadcast).
  - Arbiter cells: Arbitrate between two messages that require one path (fair arbitration).
  - Stripper cells: Strip off the first flit of messages that pass through it.
  - Decrement cells: Decrement the first flit of messages that pass through it.
  - Pipe stages: Fall through FIFO stages that buffer the flits of the messages (not drawn in diagram).
Mesh Routing Chip (MRC)

Performance

- 200+ MByte/sec at all ports simultaneously
- 40 nsec input to output latency (no change in dimension)
- 70 nsec input to output latency (changing X to Y)
- High speeds are achieved by KISS principle:
  - Keeping each pipe stage simple
  - Minimal number of buffers per stage
- I/O is fast due to:
  - Data streaming
  - Careful attention to between chip analog issues

Mesh Routing Chip (MRC)

Interlocked vs. Streaming Handshakes

- Data streaming allows for much higher data rates
  - No waiting for return acknowledge
  - Independent of physical space (propagation delay) between chips
  - Can be extended to any streaming depth
Summary

Hindsight

- Low voltage signaling; 3.3 volts? 1.0 volts?
- Faster slew-rate output buffers
- Error correction code in NIC
- Variable streaming depth
- Adaptive routing?
- Virtual channels

Summary

- High speed self-timed logic was implemented using standard cells and standard vendor tools (NIC).
- Off-the-shelf technologies were made to run fast using self-timed techniques.
- Two generic simple chips allow for high speed scalable interconnect networks.