

# Scalability Port

## A Coherent Interface for Shared Memory Multiprocessors

Mani Azimi, Faye Briggs,  
Michel Cekleov\*, Manoj Khare\*,  
Akhilesh Kumar, Lily P. Looi

Enterprise Platforms Group  
Intel Corporation



\* Author made contribution while at Intel



## Outline

- Design Goals
- Architecture
- Implementation
- Summary



Hot Interconnects 10  
Aug 21-23, 2002

2

## Scalability Port (SP) Goals

- **Designed for mid-range shared memory multiprocessors**
  - Point-to-point interface with
    - Good scalability for mid-range systems and ability to extend to high-end systems
    - Enable cost-effective system architecture
  - **Shared buses not cost-effective beyond limited number of processors**
    - Limited speed due to signaling challenges
    - Proximity of devices on a bus causes thermal and mechanical challenges
    - Hierarchical approaches limit scalability



Hot Interconnects 10  
Aug 21-23, 2002



3

## SP Goals

- **Allow flexible system architecture**
  - Enable cost-optimal small systems to scalable high-end systems
  - Enable system vendors with proprietary system interconnects and components to use Intel building blocks
- **Modular architecture**
  - Enhancements in orthogonal increments
  - Reduced verification complexity



Hot Interconnects 10  
Aug 21-23, 2002



4

## SP Architecture Overview

- **Layered Architecture**
  - Physical, Link and Protocol Layers
- **Parallel interface with cut-through routing to minimize latency**
- **Robust error detection and recovery**
  - Link layer retry + end-to-end ECC
- **Protocol tolerant of unordered network and resource constraints at targets**
  - Allows flexibility in implementation choices



Hot Interconnects 10  
Aug 21-23, 2002



5

## Physical Layer

- **Simultaneous Bi-Directional (SBD) signaling**
  - Pin efficient, full-duplex interface
  - 800 Million Transfers per Sec per Direction
- **Clocking**
  - Single clock source
  - Source synchronous interface
- **3.2 GB/Sec/Direction per SP**
  - 42 bit wide interface with 32 data, 2 link layer control, 2 SSO and 6 error control bits
  - Additional signals for strobe, reference voltage, etc.
- **20"-25" trace on FR4 with up to 2 connectors**
- **Support for hot-plug**

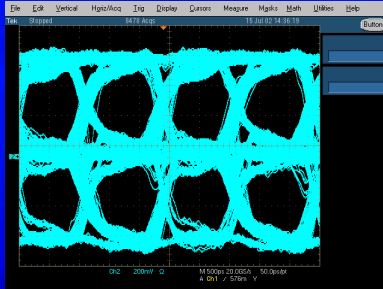


Hot Interconnects 10  
Aug 21-23, 2002



6

## Physical Layer



Data Rate	800Mb/sec/direction SBD
Process	0.18um CMOS
Supply Voltage	1.3V
Voltage and Timing Margins	>10%



Hot Interconnects 10  
Aug 21-23, 2002



7

## Link Layer

- Request and response virtual interconnect
  - Two virtual channels with VC identifier per flit (168 bits)
  - Network routing restricted to avoid deadlock
- Flow Control
  - Credit based flow control at flit granularity
- Reliable Transmission
  - Parity for error detection, link level retry for recovery
  - Sliding window protocol without sequence number



Hot Interconnects 10  
Aug 21-23, 2002



8

# Link Layer

## Flit Format



- Link layer control
  - Flit type, VC identifier, Credit, Packet delimiter, Ack



Hot Interconnects 10  
Aug 21-23, 2002

9

# Protocol Layer

- Packetized interface with multiplexed request, response and data
  - Packet header allows up to 32 nodes, 50 bit address space and 64 outstanding transactions per node
- No reliance on ordered fabric
  - Can work with other unordered fabric
  - Improved performance due to available concurrency
- Event driven protocol without fixed timing dependencies
- Transaction retry for flow control and conflict resolution



Hot Interconnects 10  
Aug 21-23, 2002

10

# Coherency Protocol

- Invalidation protocol with MESI states
- Allows separation between directory agent and home memory
  - Speculative memory accesses to hide latency
  - Building blocks designed for cost-sensitive systems can be reused in scalable systems
- Optimization suitable for commercial workloads and for accesses from I/O devices
  - Cache to cache transfers without memory update
  - Cache line ownership without data fetch
  - Coherent data fetch without altering cache states



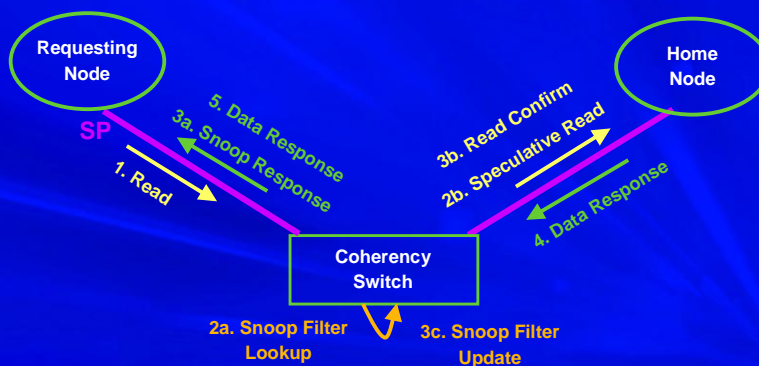
Hot Interconnects 10  
Aug 21-23, 2002



11

# Coherency Example

## Read to a Remote Clean Line



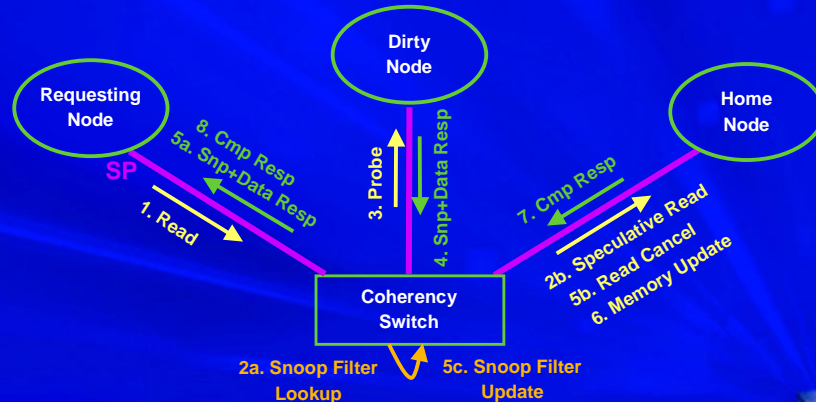
Hot Interconnects 10  
Aug 21-23, 2002

SNC = Syst Node Cntrl  
IOH = IO Hub  
SPS = SP Switch  
SF = Snoop Filter

12

# Coherency Example

## Read to a Dirty Line



Hot Interconnects 10  
Aug 21-23, 2002

SNC = Syst Node Cntrl  
IOH = IO Hub  
SPS = SP Switch  
SF = Snoop Filter

13

# Coherency Protocol

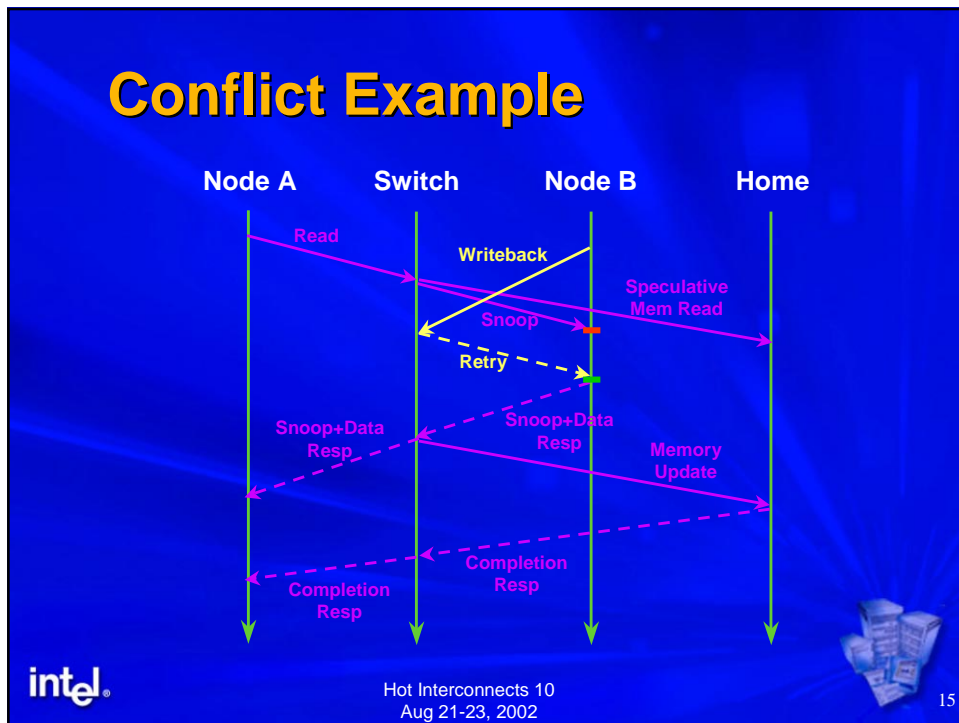
- Distributed conflict resolution
  - Directory controller determines the access order between concurrent accesses
    - Retry response on conflict at directory agent
  - Caching agents resolve races due to network reordering
    - Addresses of outgoing requests and incoming probes are compared
    - Incoming probes blocked on conflict and released on completion or retry of outgoing requests



Hot Interconnects 10  
Aug 21-23, 2002

14

## Conflict Example



## Coherency Protocol

- **Deadlock Prevention**
  - Separate request and response virtual network
  - Request retry on resource unavailability
    - Not applicable to forwarded requests
  - Nodes provide buffering for blocked forwarded requests
- **Starvation Avoidance**
  - Source detects starving transactions
    - Starving transactions completed before accepting new transactions
  - Target must be fair to all sources
    - Resource reservation per source or prioritization of sources with rejected requests

intel.

Hot Interconnects 10  
Aug 21-23, 2002

16

## Other Protocol Features

- Message based interrupt delivery
- TLB consistency protocol
- Support for pipelined writes to MMIO
- Error Handling
  - End-to-end data protection with ECC
  - Data poisoning for error containment
  - Error isolation at component and link level
  - Response status to indicate transaction failure



Hot Interconnects 10  
Aug 21-23, 2002

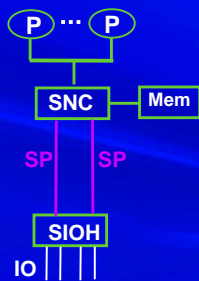


17

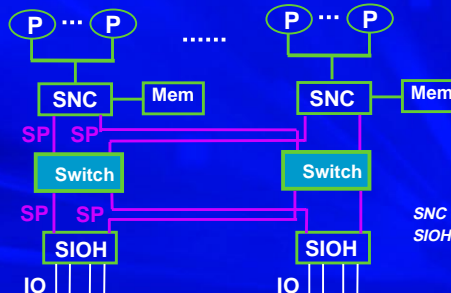
## SP Implementation

- E8870/E8870SP chipset for Itanium<sup>®</sup>2 processor
  - 1-4P without coherency switch
  - 8P and above using coherency switch

Two Node Direct Connect



Multi-Node with SPS



SNC = Scalable Node Ctr  
SIOH = Server I/O Hub



Hot Interconnects 10  
Aug 21-23, 2002



18

## Comparison

	Point-to-point (E8870SP System)	Hierarchical bus (Sun Fire 6800)
<b>Processors</b>	16 Itanium®2	24 UltraSPARC-III
<b>Max Data BW</b>	25.6 GB/sec	9.6 GB/sec
<b>Max Snoop Rate</b>	532 million/sec with 2 switches	150 million/sec
<b>Complexity</b>	4 SNC, 2 IOH and 2 switch components 2 SP links per CPU/memory board	>32 Data Switches, >8 Address Repeaters 288 bit data + address/control per CPU/memory board

\*Other names and brands may be claimed as the property of others.



Hot Interconnects 10  
Aug 21-23, 2002



19

## SP Summary

- **Scalable System Interface**
  - Point-to-point interface with simultaneous bi-directional signaling
  - Modular layered architecture
  - Interface for scalable building blocks
  - Enhanced RAS support
- **E8870/E8870SP Chipset Implementation**
  - Cost-effective and scalable architecture
  - Building block with other architectures



Hot Interconnects 10  
Aug 21-23, 2002



20