# *The Intel® 870 Family of Enterprise Chipsets*

*Fayé Briggs, Michel Cekleov*, Kai Cheng, Ken Creta, Manoj Khare, Steve Kulick, Akhilesh Kumar, Lily Looi, Chitra Natarajan, Linda Rankin*

*Enterprise Products Group*
*Intel® Corporation*
*Hot Chips XIII*

*Michel Cekleov is now at 3PARdata.

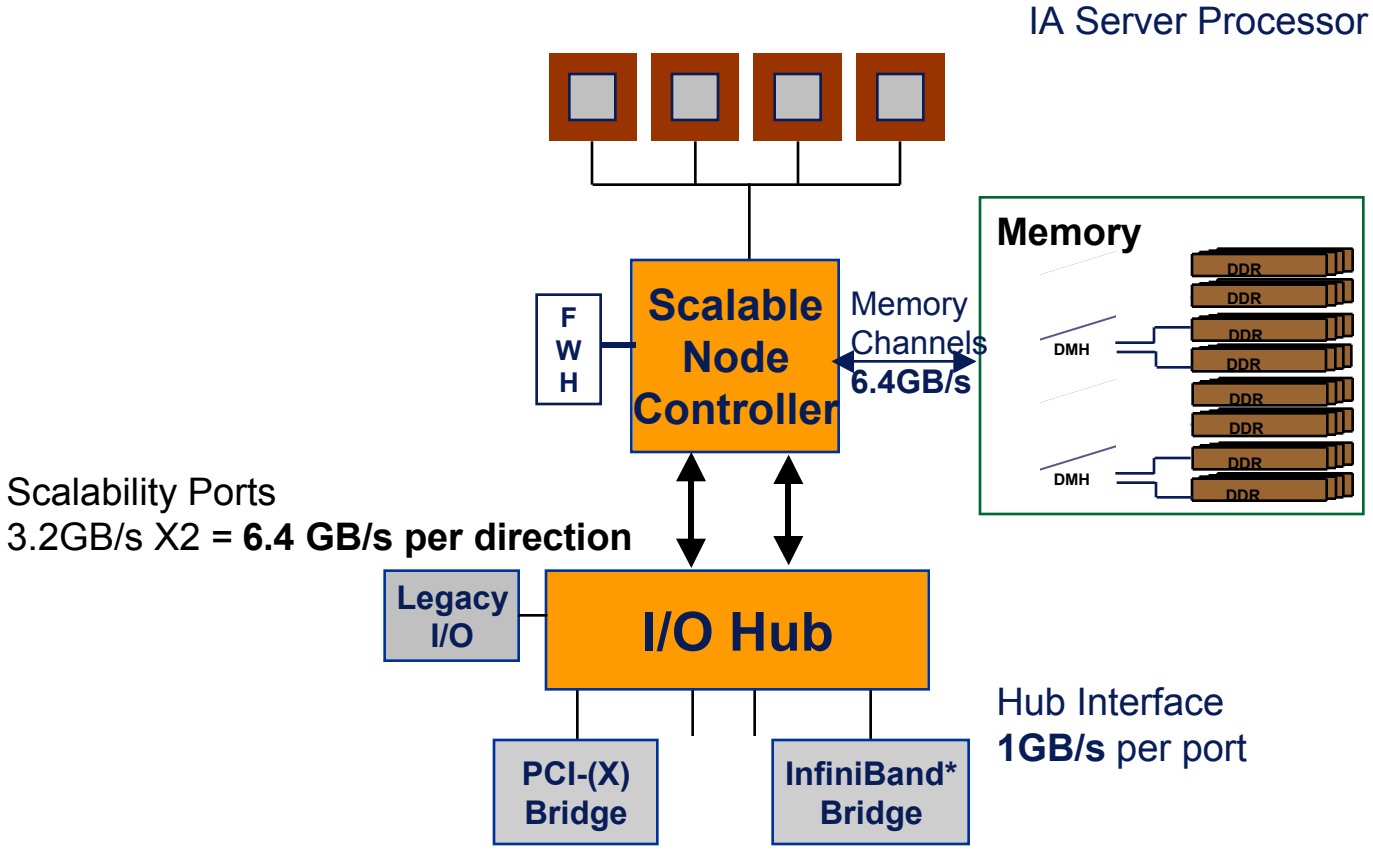
# *Introduction*

- Scalable building blocks
  - 2P to large-scale multi-node (distributed shared memory platforms)
  - 3 major subsystems (CPU/memory, IO, coherent switch)
  - Supports next generation Itanium Processor Family(IPF) & IA32 server CPUs
- Key technologies
  - Scalability Port – physical interconnect, coherency protocol
  - Scalability Port Switch – multi-node coherency, router
- Distributed Shared Memory architecture
- High End Features
  - Caching I/O hub
  - Domain partitioning and node hot plug
  - Extensive Error handling support

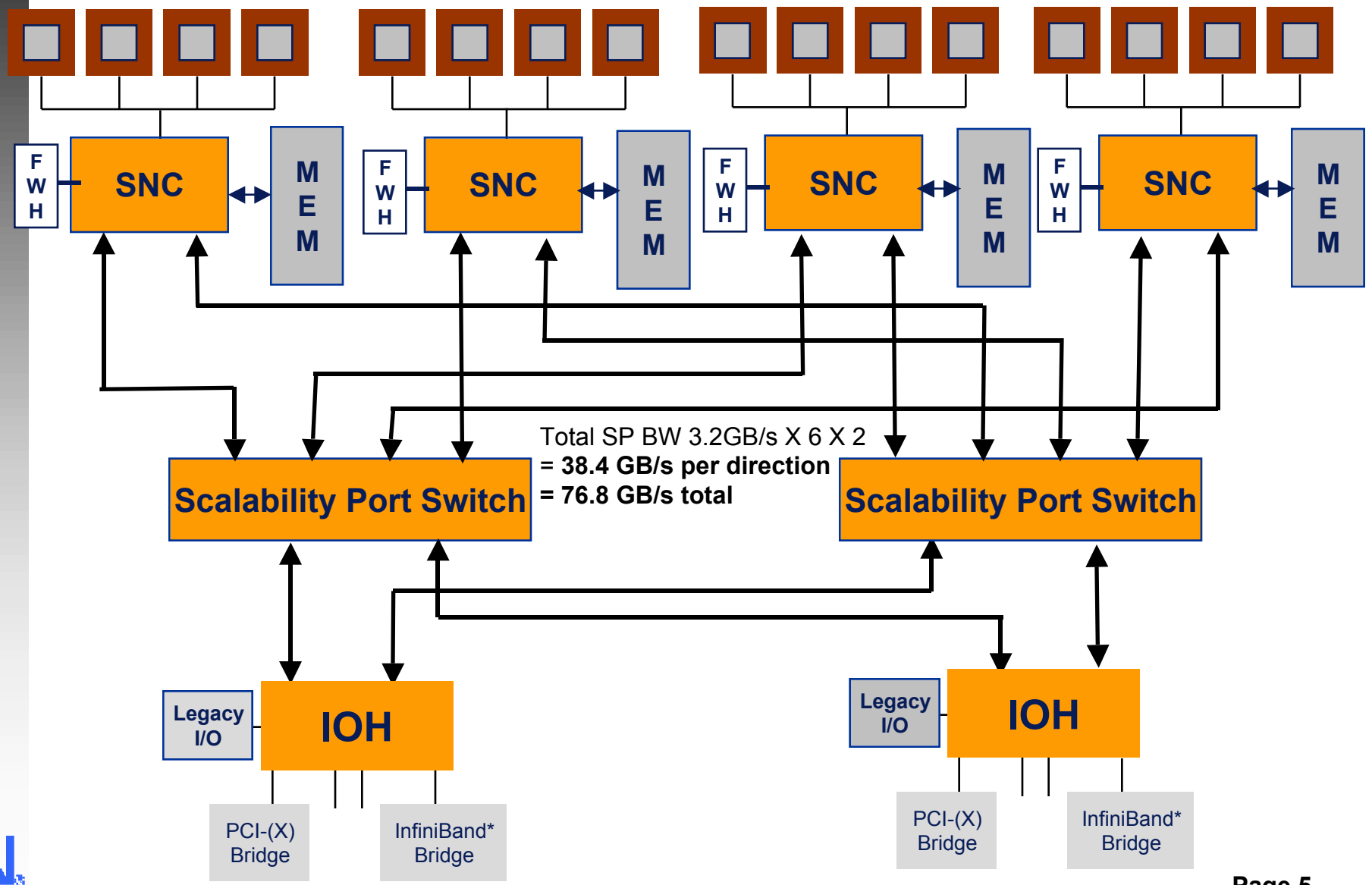
# *Motivation*

- Provide a family of scalable enterprise chipsets for next generation Itanium Processor Family(IPF) & IA32 server processors
  - Common building blocks to scale from 2P to 256P servers
- Extensive RAS features for the enterprise
- Provide building blocks for OEMs to differentiate and scale beyond 16P with persistent interface

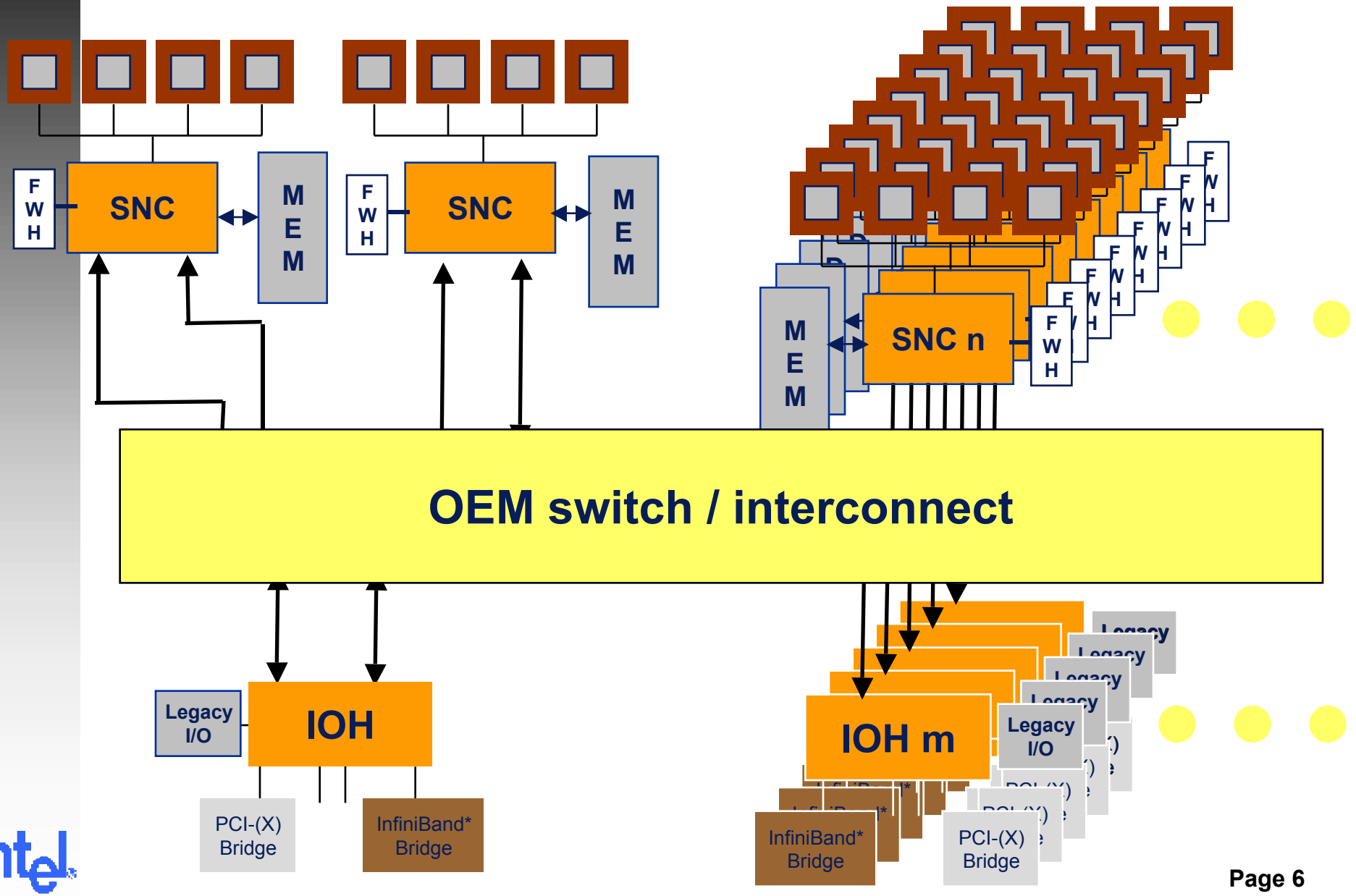
# Intel® 870 2 - 4 Way Server



# Intel® 870 16 Way Server

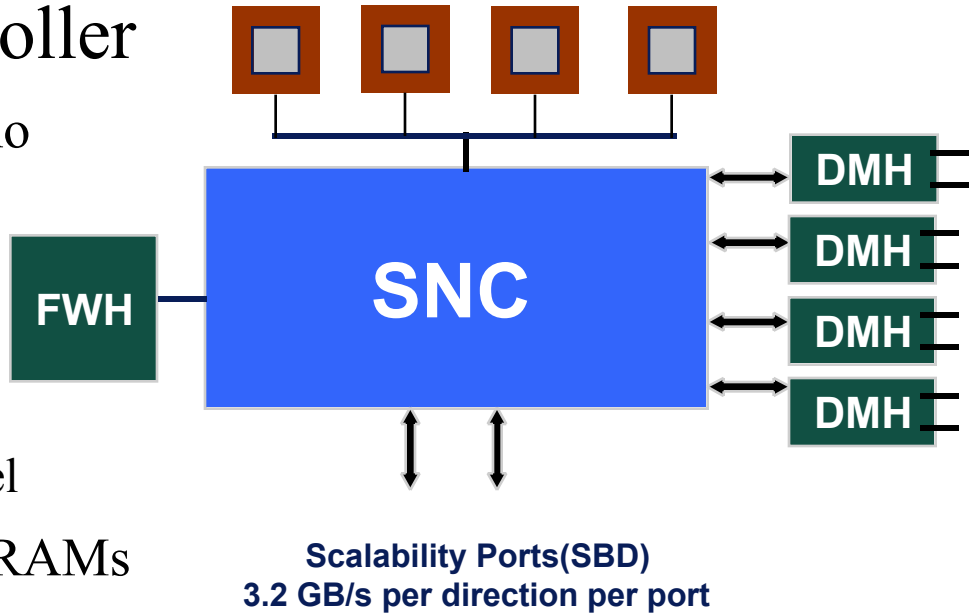


# Greater than 16-Way Server



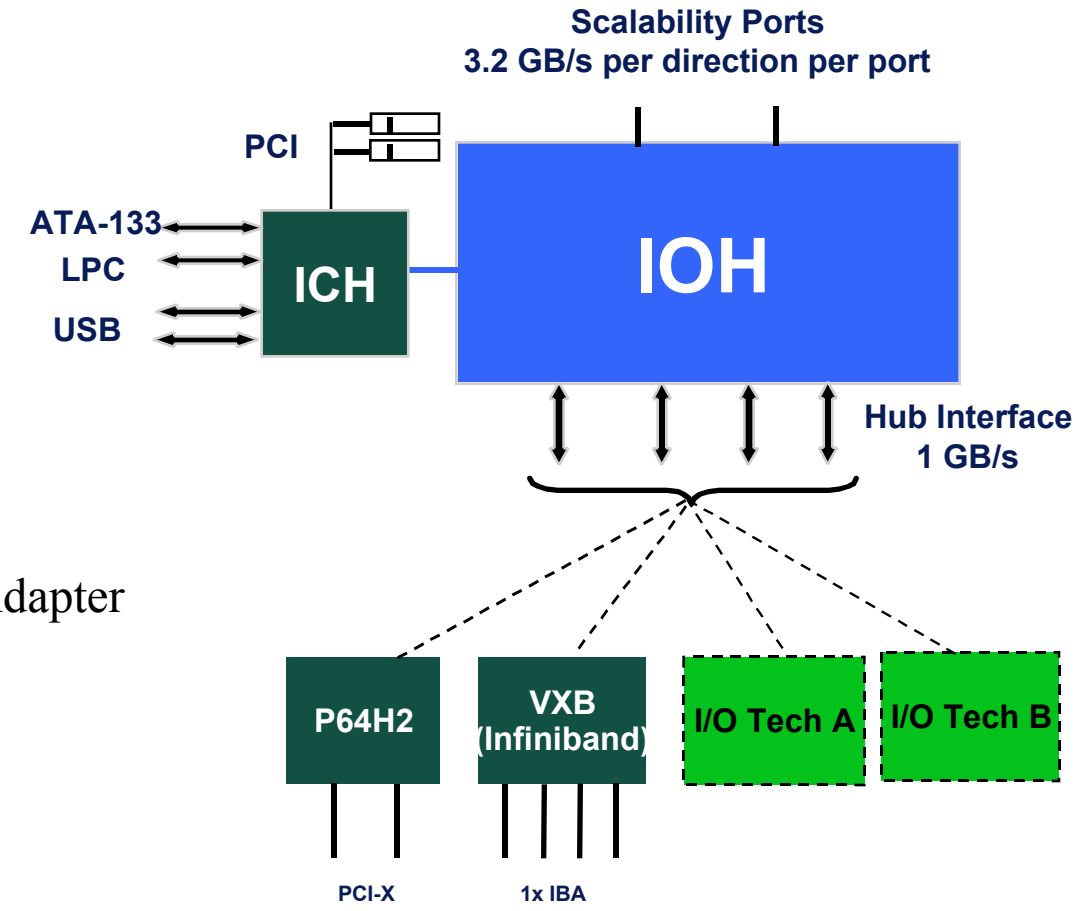
# 870 Processor/Memory Subsystem

- Scalable Node Controller
  - Supports Memory-Only (no processor) option
- DDR Memory Hub
  - Supports 2 DDR channels
    - 4 DIMMs on each Channel
  - 128GB/SNC with 1Gb DRAMs
- FWH
  - Firmware for each Processor bus supports parallel initialization & Boot



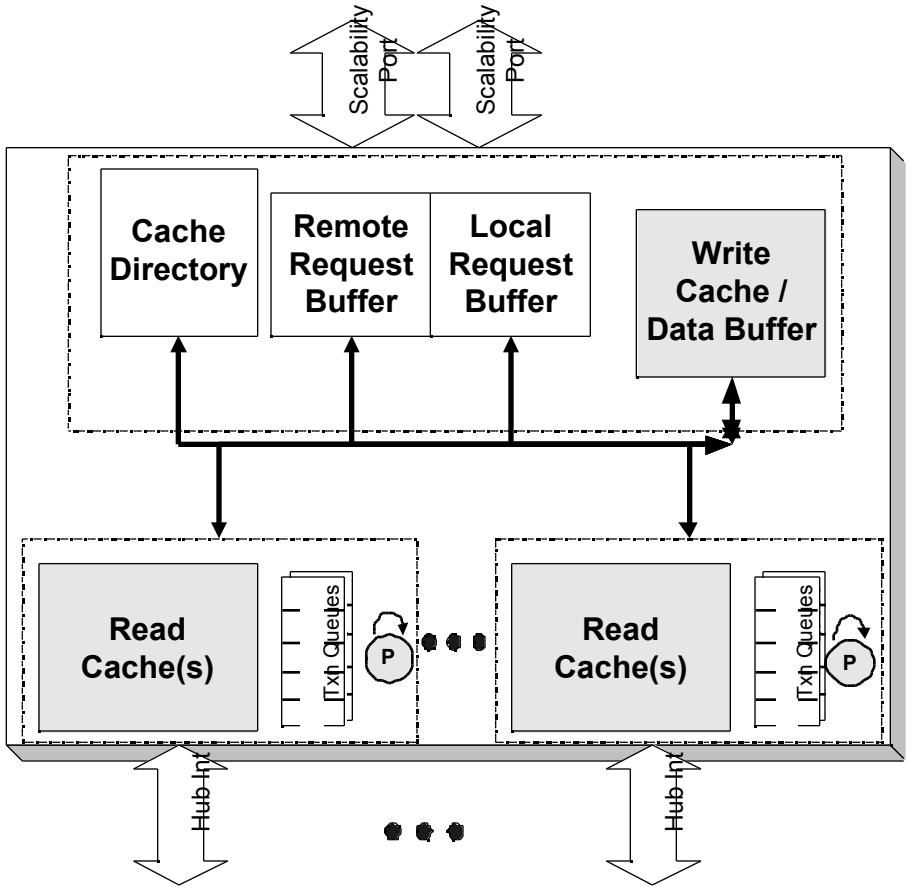
# 870 I/O Subsystem

- 870 I/O Hub
  - Multi-stream caching hub
  - Adaptive Prefetch Logic
- P64H2
  - Dual PCI-X bridge
  - Hotplug PCI
- VXB
  - Infiniband Host Channel Adapter
  - Four 1x ports (2.5 Gbps)
- ICH
  - Legacy I/O bridge
- Others Possible...



# 870 I/O Hub

- Common Write Cache
  - Promotes combining for partials
- Read Cache per Hub Interface
  - Least-recently allocated replacement policy
  - Data storage shared across one or many streams
- Cache Directory
  - Tracks the lines inside the IOH
- Multiple Transaction queue
  - Relaxes Ordering wherever possible

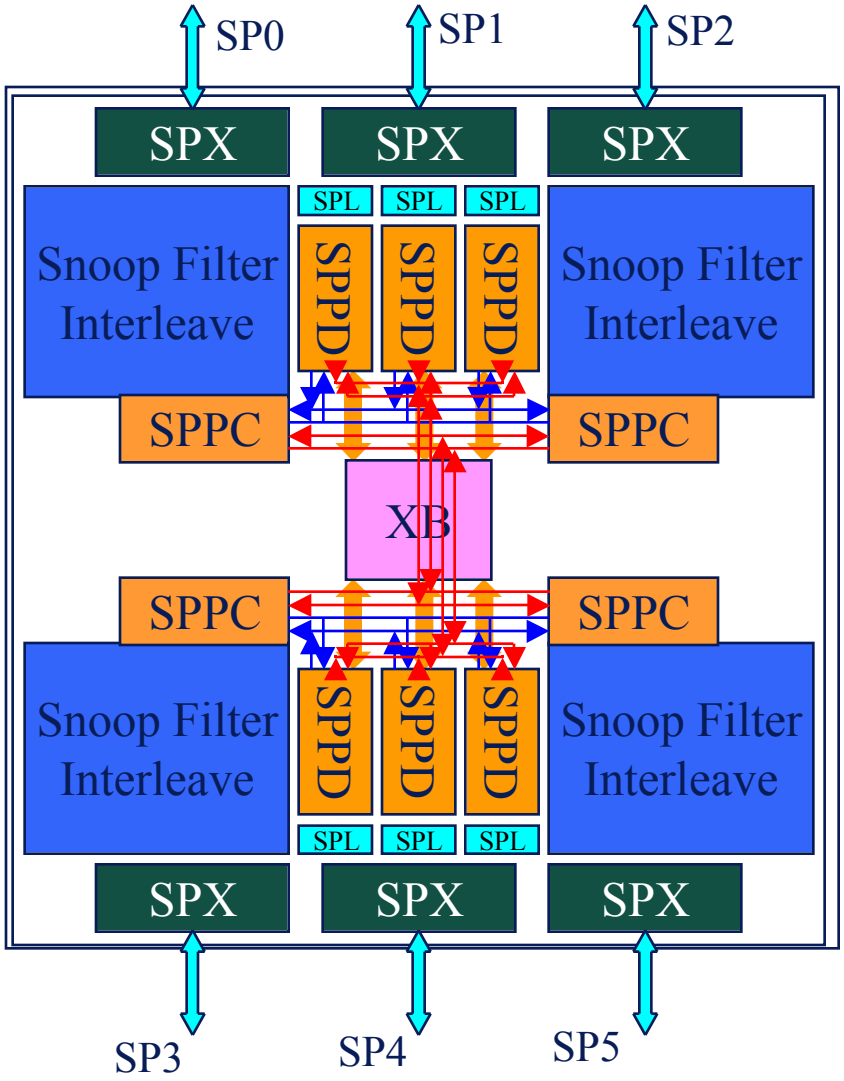


# 870 IOH Prefetch Logic

- Speculative Prefetching
  - Useful for PCI where read length is *not* specified
  - Hides long memory latencies and provides data streaming
  - Priority given to received PCI bridge requests over speculative pre-fetches
  - Dynamically detects and distinguishes between streams using address patterns
  - Dynamically adapts pre-fetching by throttling pre-fetch per stream based on the current number of active streams
- Non-speculative Prefetching
  - Useful for PCI-X and Infiniband where read length is specified
  - I/O bridges can be designed for low latencies
  - IOH prefetches up to indicated amount and locally buffers for subsequent reads

# Coherent Switch - (SPS)

- Six identical scalability ports
  - Each supports up to 3.2 GB/s peak bandwidth each direction
  - Each port contains the 3 layers of the Scalability Port (SPX/SPL/SPP)
  - Protocol layer further partitioned into distributed (SPPD) and centralized functions (SPPC)
- Interconnect includes a crossbar and bypass buses for critical coherent traffic
- Enables Central Snoop Filter Architecture
  - Minimizes snoops to remote nodes



# *SPS Coherency*

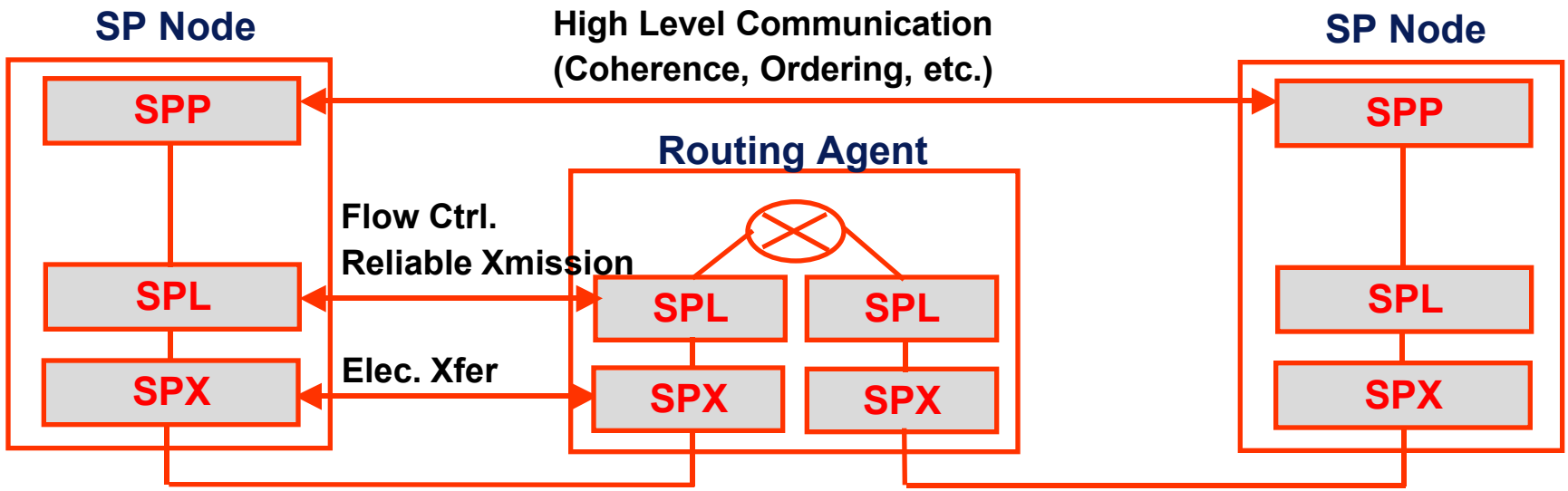
- Distributed SP Protocol (SPPD)
  - Address/request decoding determines how a packet should be routed in the SPS.
  - Controls data transfers between ports including modified data transfers
- Centralized coherency protocol divided into four interleaves
  - Interleaved to improve throughput
  - Includes Snoop Filter (SF) and Centralized SP Protocol (SPPC)
  - SPPC contains programmable protocol engine. Processes requests and responses and spawns transactions as needed.
  - Handles global ordering.
  - Contains anti-starvation logic to guarantee fairness between nodes.

# *Scalability Port*

- Point to Point Coherent Interconnect – 6.4 GB/s
- Supports both IA32/IPF processors
- Efficient I/O transfer support
- High bandwidth, Low Latency
- Enhanced RAS support
- Scalable, Pin-efficient Architecture
  - Layered architecture
  - Packet based protocol
  - No fixed timing, buffer size relations
  - Muxed Request/Response/Data

# *Layered Architecture*

- Three Layers
  - Physical Layer (SPX)
  - Link Layer (SPL)
  - Protocol Layer (SPP)
  - Benefits: Modularity/Longevity, Efficient routing



# *SPS Snoop Filter*

- Stores address tags/state for all system caches
  - Supports multiple line sizes
  - ECC coverage –single-bit errors corrected, double-bit errors detected

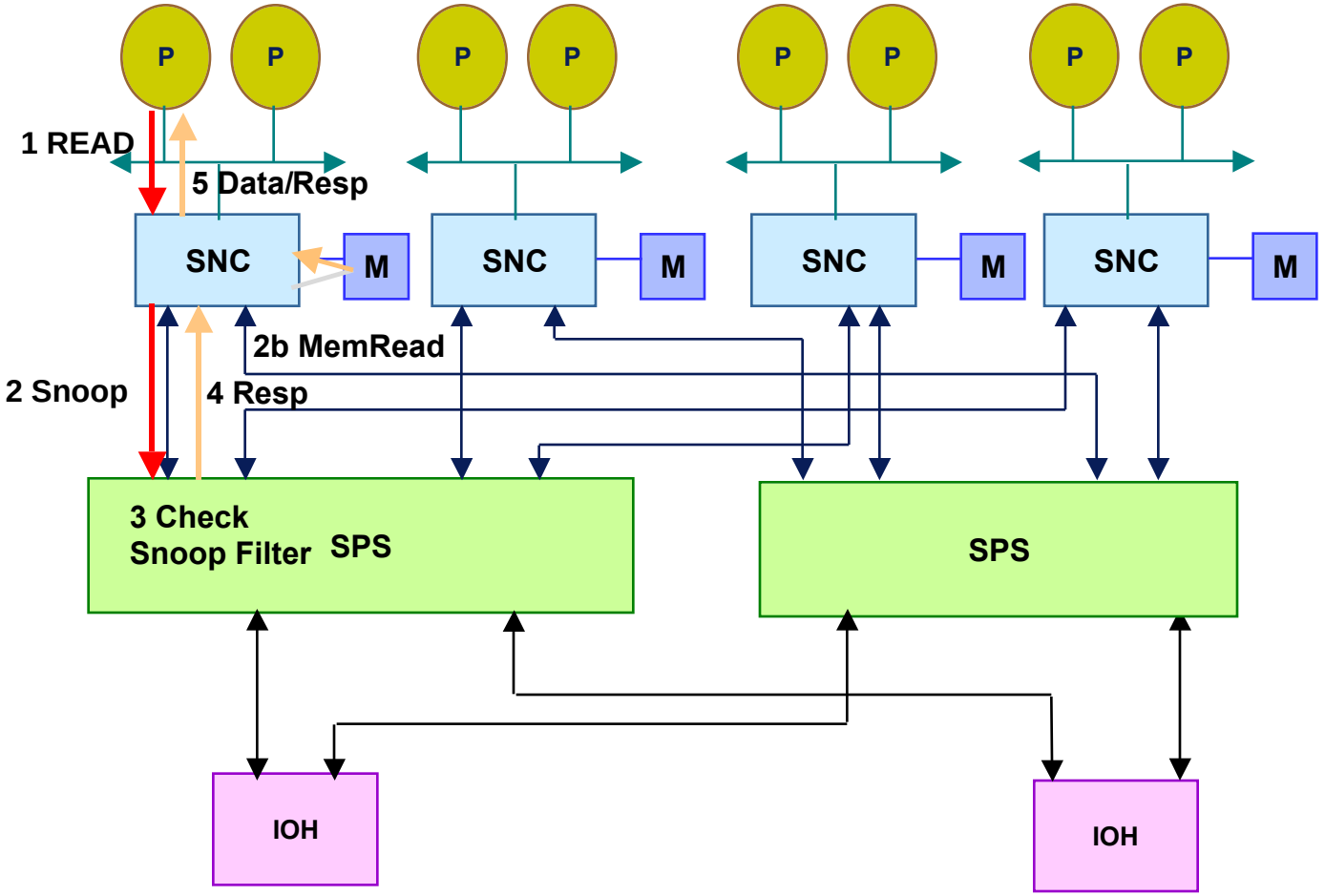
| Bits | |
|---|---|
| [35:29] | ECC check bits |
| [28] | State of the cache line (M/E, S, I) |
| [27:22] | Presence vector |
| [21:0] | Tag Portion of the address |

**SF entry**

- Snoop Filter size is ~1 MB
  - Can maintain state of ~200K cache lines per SPS
  - 12-way set associative array partitioned into 4 interleaves
  - Pseudo-Least-Recently-Used (PLRU) replacement algorithm
  - Snoop filter operates at 400 MHz. Maximum throughput of 266M LUU/s (lookup-update/s) per SPS.

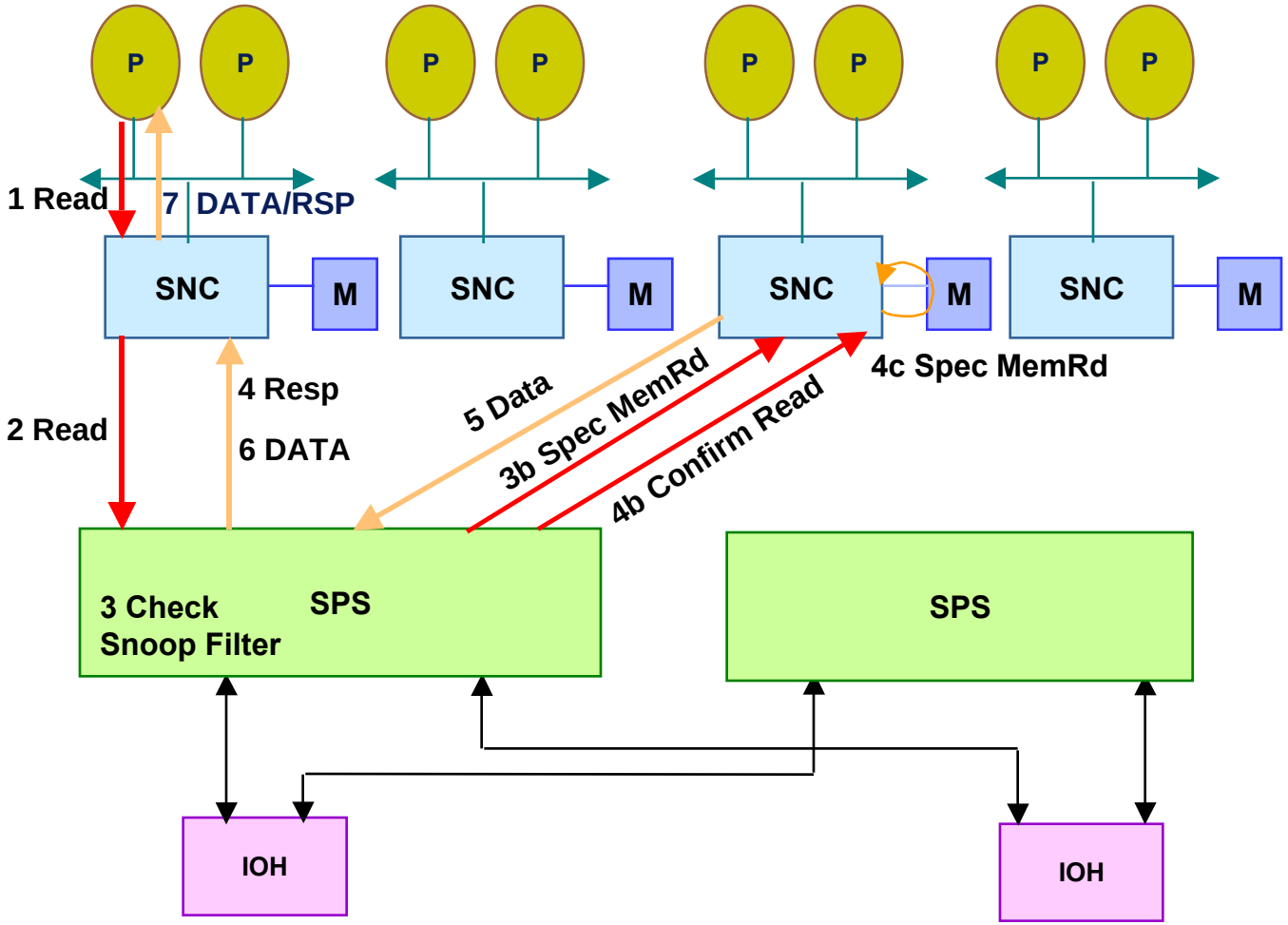
# Transaction Flow Example #1

- Clean local read



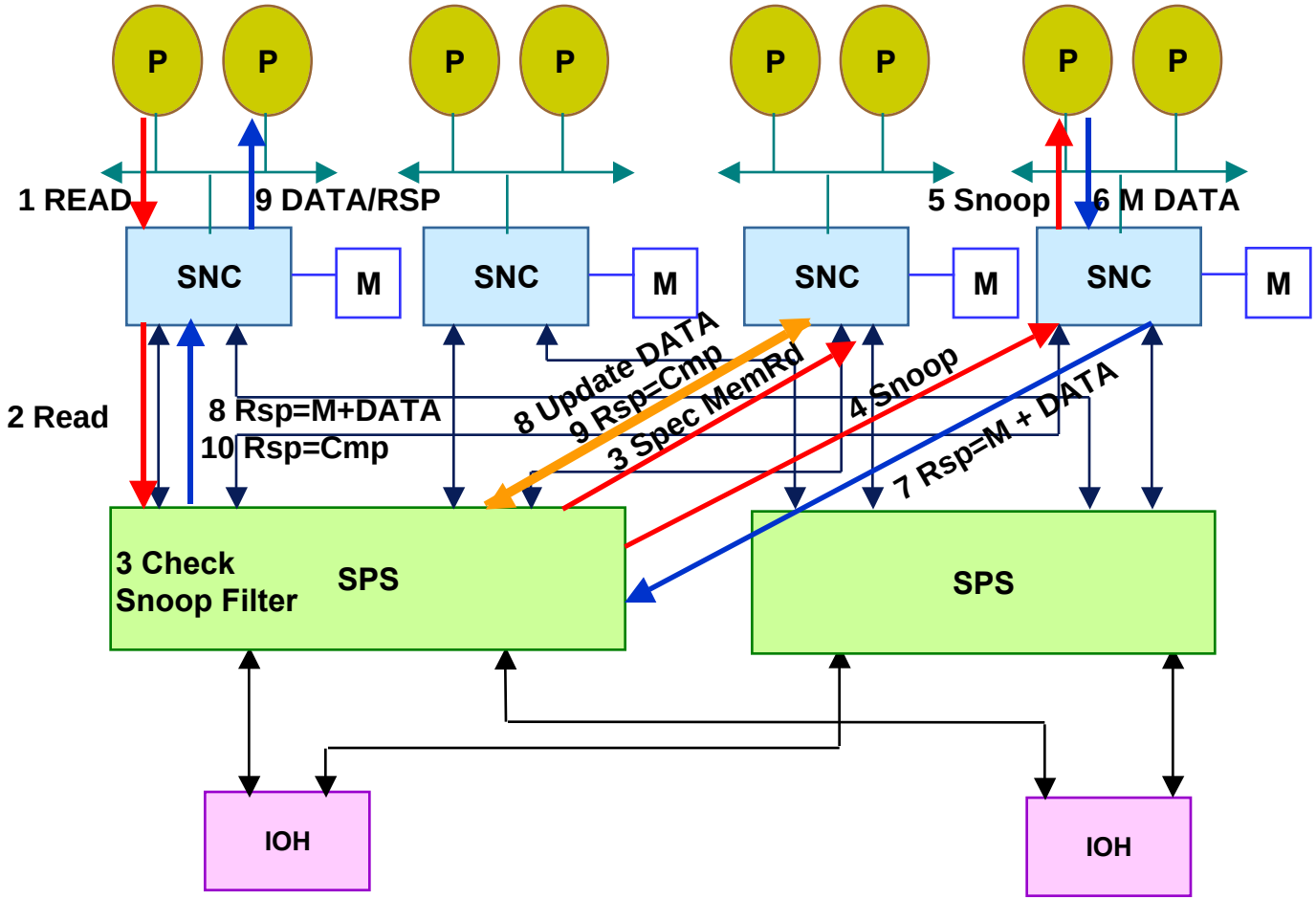
# Transaction Flow Example #2

- Clean remote read



# Transaction Flow Example #3

- Remote read with modified (HITM) data



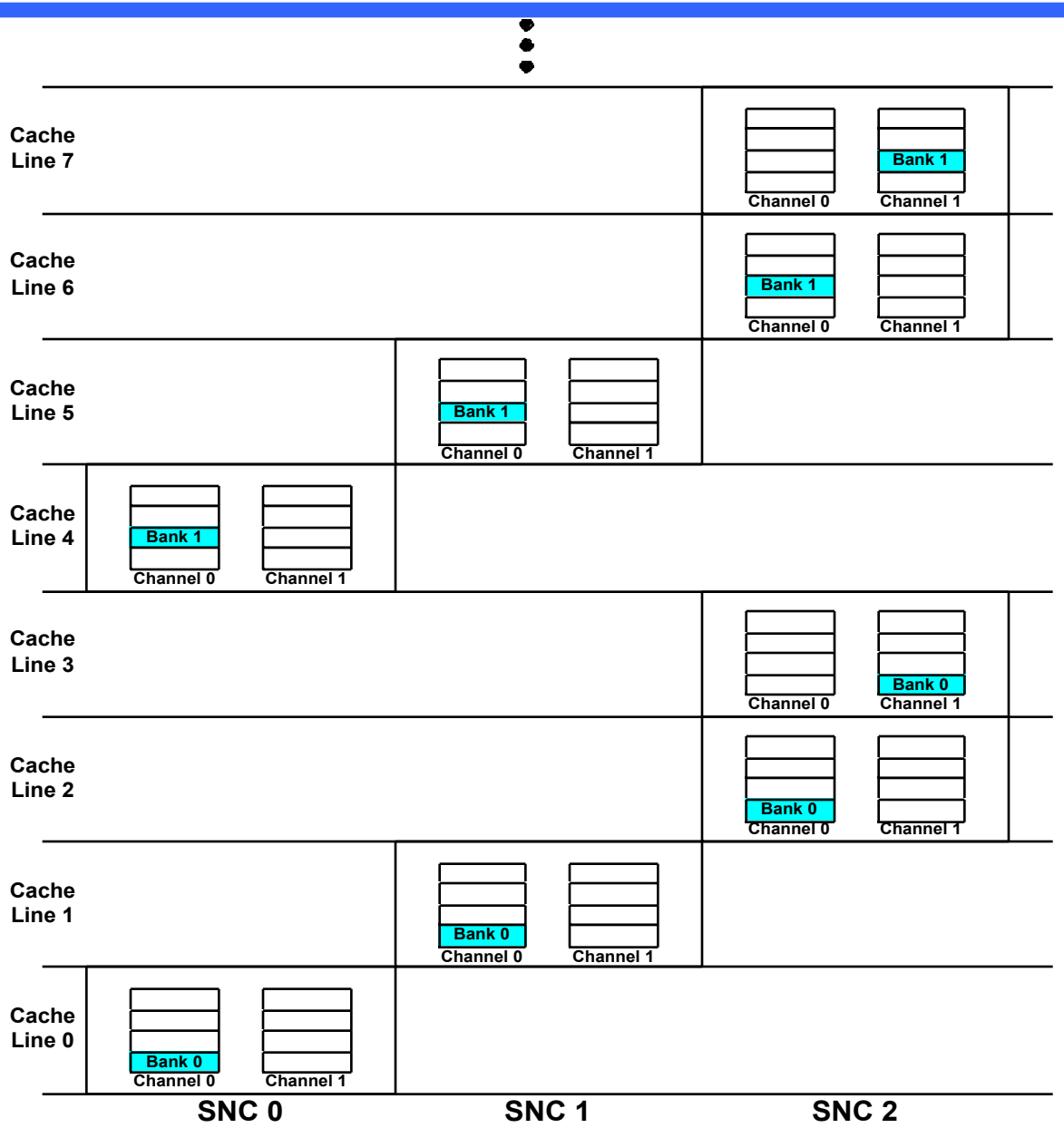
# *Distributed Memory Interleaving*

- Supports Interleaving 4 ways across:
  - 2, 3, or 4 SNCs.
  - Arbitrary numbers of SNCs for >16 way
  - Asymmetric Configurations: odd numbers of DIMM rows, different DIMM sizes and DRAM densities
  - Block mode for NUMA optimized systems and cache-line interleave for non-NUMA.
  - 64B (IA32) cache lines and 128B (IPF) cache lines

# *Hierarchical Decode*

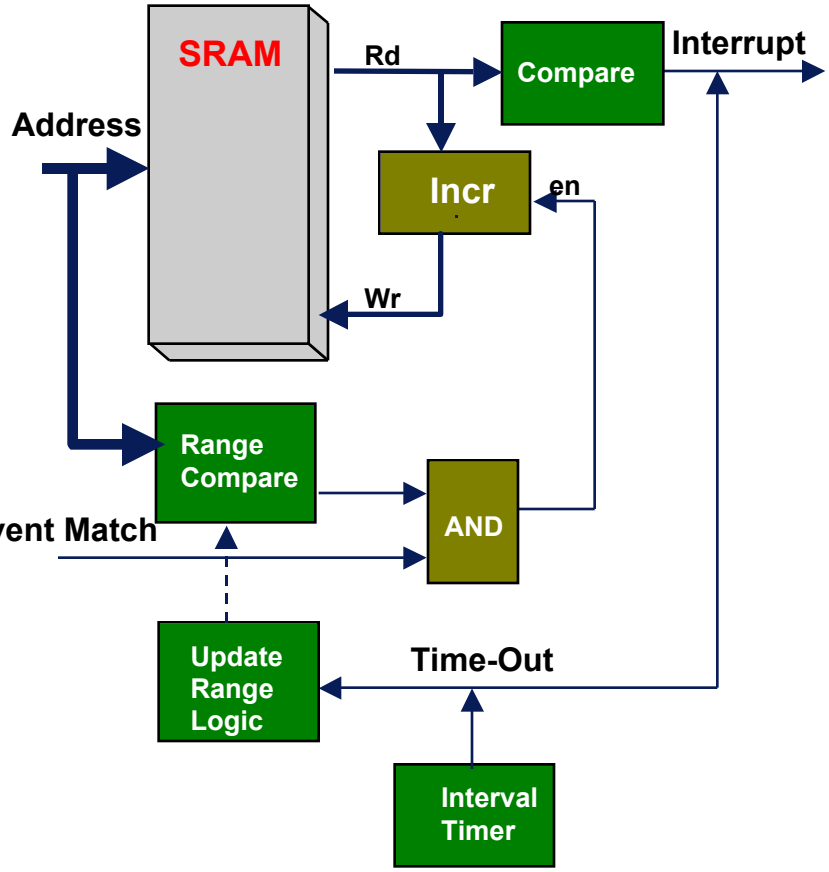
- Global Interleaving (across SNCs)
  - Lines may be interleaved up to 4 ways (A[8:7] = 00,01,10,11)
  - Memory Map is divided into 6 Global Interleave Ranges (in SPSs)
    - Each interleave range has 4 ways which may be assigned to memory on a different SNC, of a different DIMM size, DRAM density, etc.
    - OEM switches implement as many Global Ranges as required by their topology.
- Local interleaving (across DRAMs)
  - SNC address bit permuting rotates accesses to consecutive lines across a sequence of 32 DIMM Channels, DIMM Rows, banks, sides, etc.
  - These local interleaves can be combined on or across SNCs to form global interleaves.

# Interleaving Example: 3 SNCs



# HotPage For Software Tuning

- Real-time event collection for host bus transactions
  - Address granularity (64B,4KB, 8KB, 64KB, 256MB)
  - Event qualifiers (reads, writes, etc.)
  - Collection qualifiers (sample interval, threshold)
  - SRAM with indexed access
  - Autorange
  - Ease of use: compatible with all event logic
- Application
  - Histogramming (maximum counts per range)
  - Scanning with event triggered on threshold
  - At 8KB resolution, 0.5 sec sample periods, 32 GBs of memory scanned in 17 minutes



# *RAS Feature Summary*

## Detection

- ECC/parity on buses
- Memory ECC
- Memory scrubbing
- Control/operational errors
- > 50 unique errors detected

## Containment

- Correction, Data poisoning
- Transaction error response

## Status/Signaling

- Error typing
- Error Masking
- First error/Next error status

## Logging

- Error logs (control/data)
- Multi-node error trail

## Serviceability

- Memory failure correction/isolation
- PCI hotplug
- CPU/memory, IOH node hotplug

## Multi-node

- Multi-pathing/redundancy
- Domain partitioning

# *Node Hot Plug*

- Add/remove/replace processor/memory node, or I/O node while OS is running
- SP is the hot plug interface
  - Physical Layer Support
    - Tri-state control based on connection (SP_Pres)
  - Link Layer Support
    - Connection/Initialization handshake sequence
  - Software controllability/observability
    - Enable/disable control
    - Signaling and status on SP connection/initialization events
    - Register storage available to store hot plug sequencing states
    - Observability/controllability of SP related GPIOs
    - Connection status (SP_Pres pin)

# Multi-pathing/Redundancy

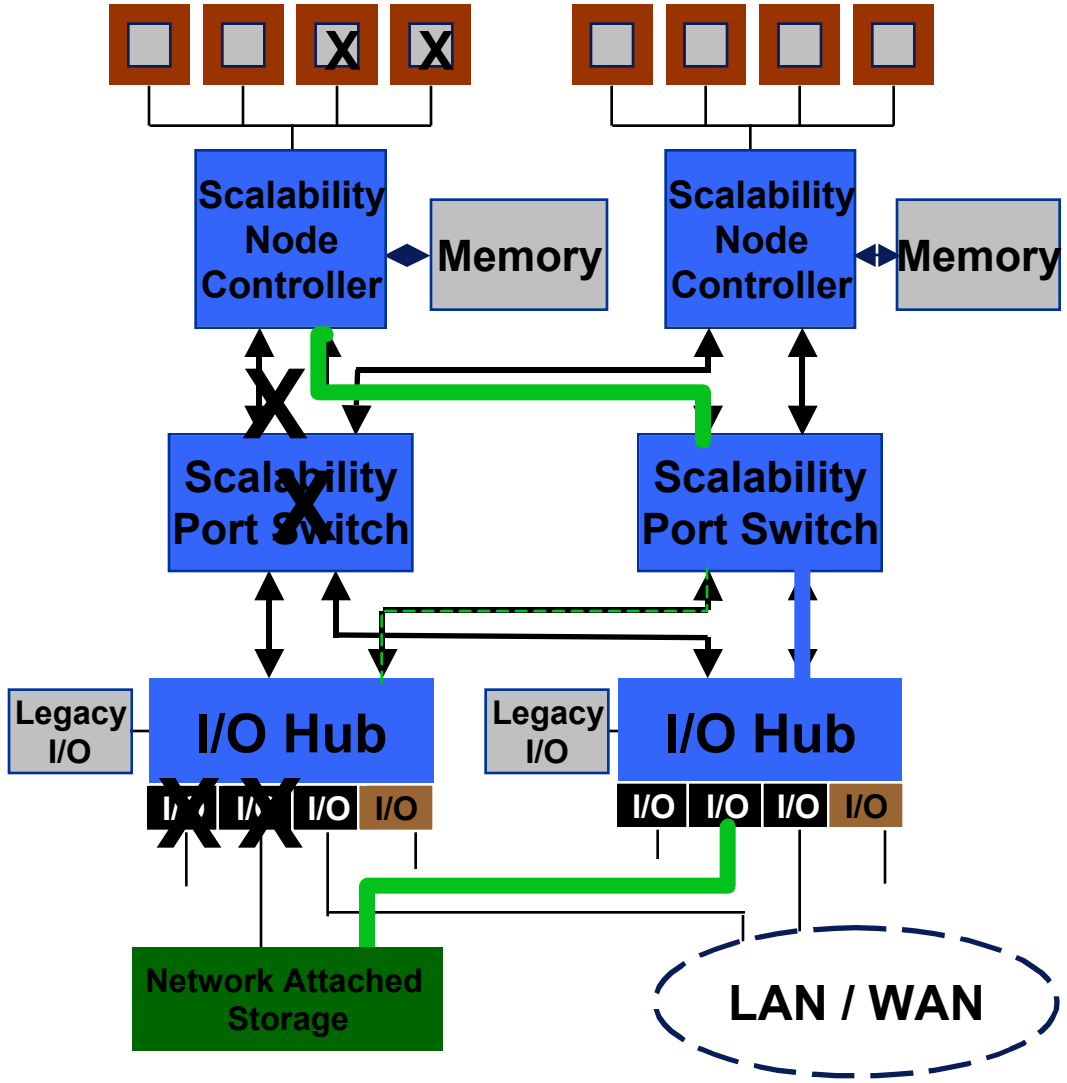
## Redundancy

- Processors
- SNC / memory
- Scalability Port / SPS
- IOH
- I/O Bridges

## Multi-pathing

- Around SPS
- Configurable I/O

## Fast reset and reconfiguration



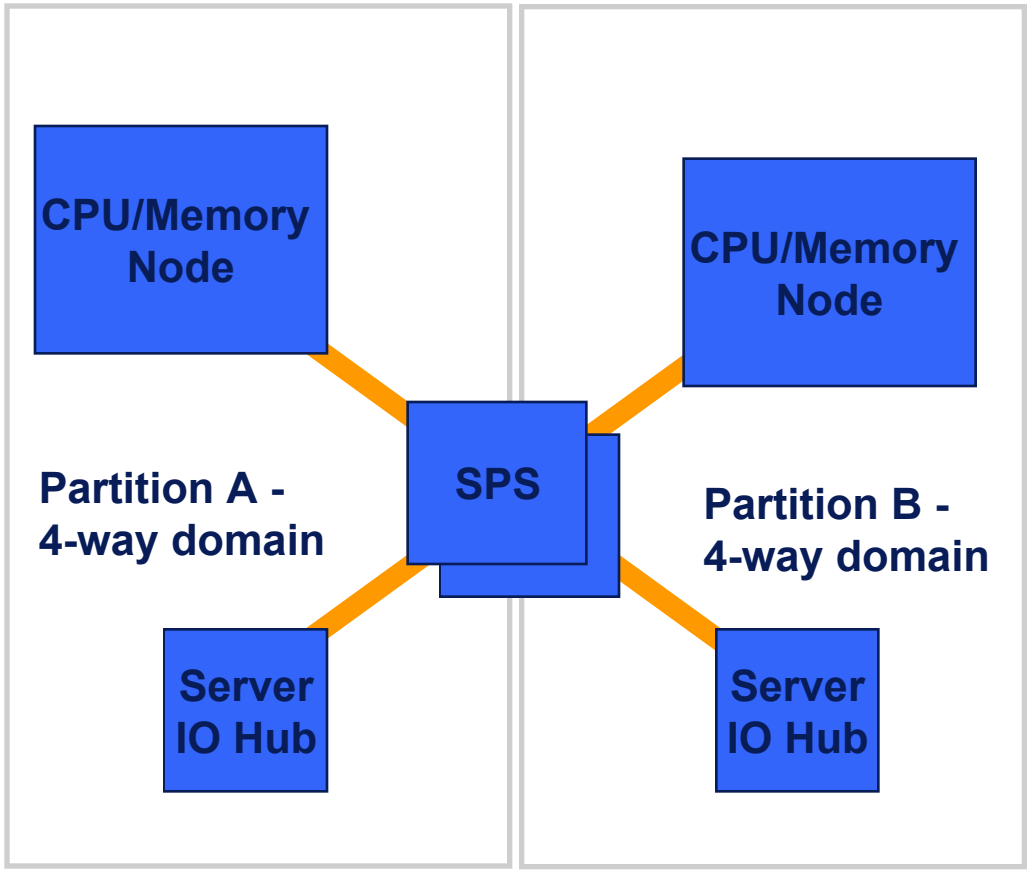
# SPS : Domain Partitioning

**SPS partitions**

- _ of protocol core
- _ of snoop filter
- Any 2 or more SP ports
- Independent reset
- Independent error status/signaling
- Independent SP interrupt output
- Domain write protection of registers

**Each domain supports**

- Node Hot Plug
- Redundancy
- Multi-Pathing
- Reset



# *Summary*

- The Intel® 870 chipset enables Enterprise-level features for two Intel Architectures and multiple system topologies up to 16P.
- The Intel® 870 chipset also provides building blocks for scaling beyond 16P through the scalability port.