HORUS

Large Scale SMP for Opterons

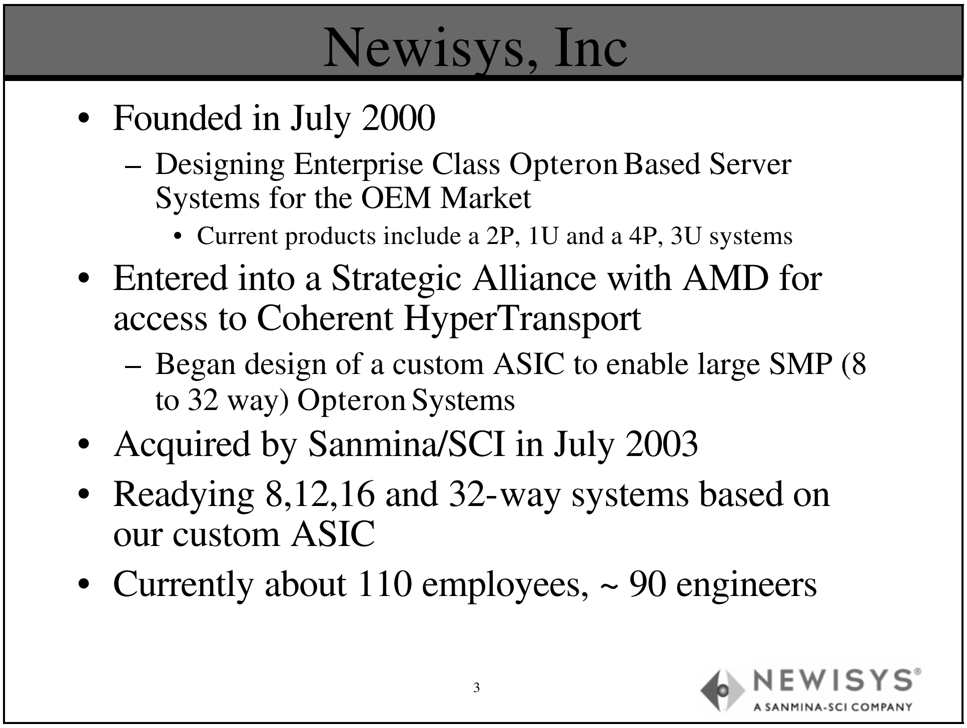
Rich Oehler

Rajesh Kota

23 August 2004

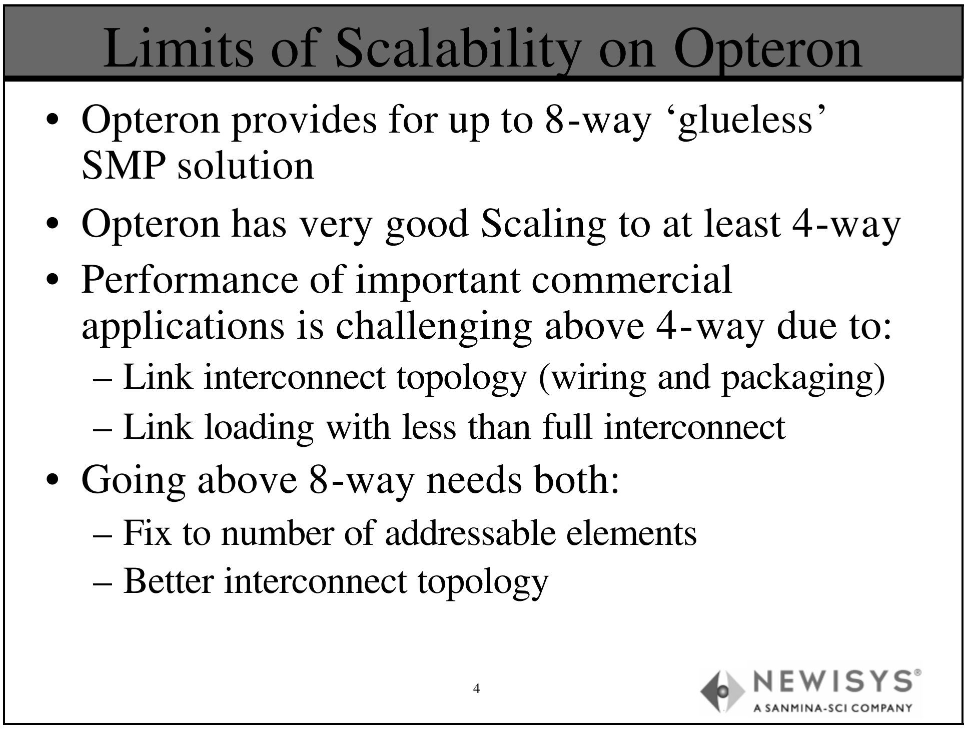
Outline

- Newisys, Inc. A Sanmina-SCI company
- Limits of Scalability on Opteron
- Horus – Our Custom ASIC
- System Management around Horus
- Summary
- Horus Team



Newisys, Inc

- Founded in July 2000
  - Designing Enterprise Class Opteron Based Server Systems for the OEM Market
    - Current products include a 2P, 1U and a 4P, 3U systems
- Entered into a Strategic Alliance with AMD for access to Coherent HyperTransport
  - Began design of a custom ASIC to enable large SMP (8 to 32 way) Opteron Systems
- Acquired by Sanmina/SCI in July 2003
- Readying 8,12,16 and 32-way systems based on our custom ASIC
- Currently about 110 employees, ~ 90 engineers

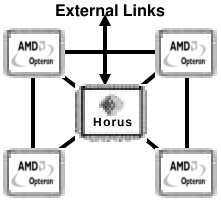


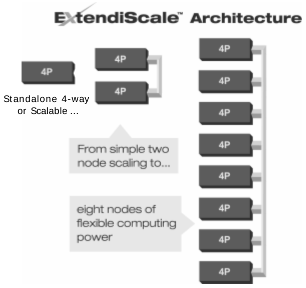
Limits of Scalability on Opteron

- Opteron provides for up to 8-way ‘glueless’ SMP solution
- Opteron has very good Scaling to at least 4-way
- Performance of important commercial applications is challenging above 4-way due to:
  - Link interconnect topology (wiring and packaging)
  - Link loading with less than full interconnect
- Going above 8-way needs both:
  - Fix to number of addressable elements
  - Better interconnect topology

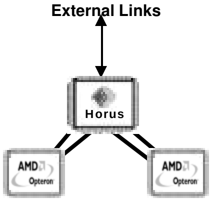
Newisys ExtendiScale™ Architecture

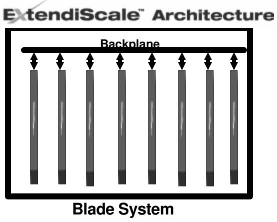
The Ultimate Answer to Performance Headroom




ExtendiScale™ Architecture
4P
Standalone 4-way or Scalable ...
From simple two node scaling to...
eight nodes of flexible computing power

- ***Pay as you go flexibility with the ability to change server resources as real-time IT needs change***
- Enables modular systems
  - Traditional 4-32 way SMP(64 way with dual core)
  - Blade frame 2-16 way SMP(32 way with dual core)
- The ExtendiScale Architecture delivers:
  - Pay as you grow budget flexibility
  - Low system cost derived from use of industry standard parts
  - Mission Critical ready: Availability, Manageability, Reliability

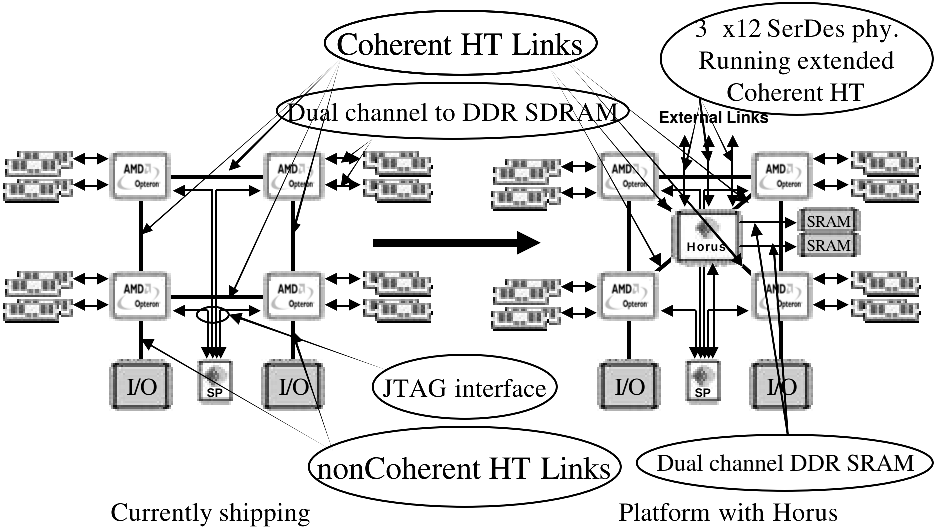




NEWISYS®
A SANMINA-SCI COMPANY

5

4 Socket (Quad) Opteron System extended with Horus



Currently shipping
Newisys 4300 platform

Platform with Horus

6

NEWISYS®
A SANMINA-SCI COMPANY

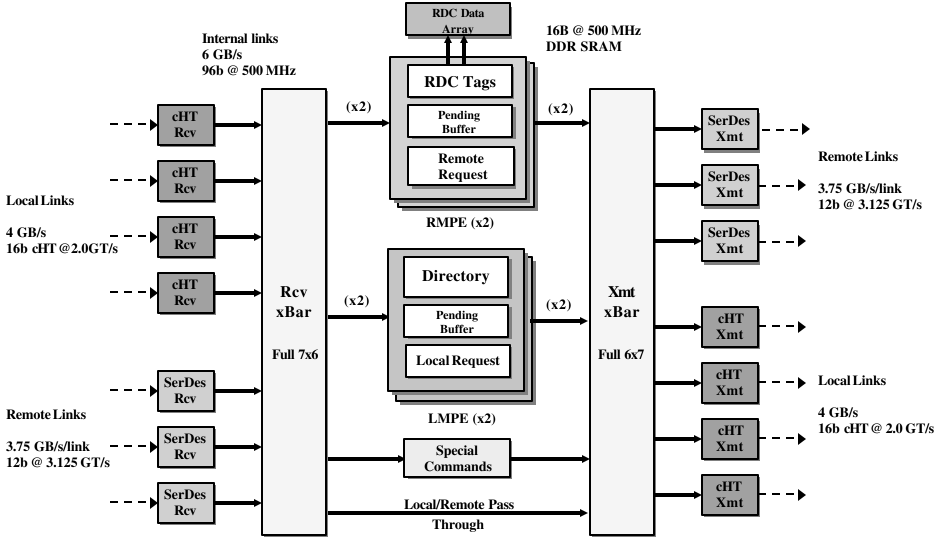
HORUS – Our custom ASIC

- Horus solves
  - – Scalability (to 32 sockets) using Coherent HT links
  - – Remote memory access latency (RDC, 64MB)
  - – Local memory access latency (DIR, 50% sparcity)
  - – Remote link bandwidth usage (RDC & DIR)
- Horus provides RAS features equivalent to those present in large RISC/UNIX systems using Opteron industry standard servers
- Horus extends every glue-less SMP feature of Opterons to multiple quads.
  - – MMIO, PCI-Conf., PCI-IO, Locks, System Management, Interrupts and SEM

7

NEWISYS®
A SANMINA-SCI COMPANY

Horus Command Path



RDC Data Array

16B @ 500 MHz
DDR SRAM

Internal links
6 GB/s
96b @ 500 MHz

Local Links

4 GB/s
16b cHT @2.0GT/s

cHT Rcv

cHT Rcv

cHT Rcv

cHT Rcv

Remote Links

3.75 GB/s/link
12b @ 3.125 GT/s

SerDes Rcv

SerDes Rcv

SerDes Rcv

Rcv xBar

Full 7x6

(x2)

RDC Tags

Pending Buffer

Remote Request

RMPE (x2)

(x2)

Directory

Pending Buffer

Local Request

LMPE (x2)

Special Commands

Local/Remote Pass Through

(x2)

(x2)

Xmt xBar

Full 6x7

SerDes Xmt

SerDes Xmt

SerDes Xmt

Remote Links

3.75 GB/s/link
12b @ 3.125 GT/s

cHT Xmt

cHT Xmt

cHT Xmt

cHT Xmt

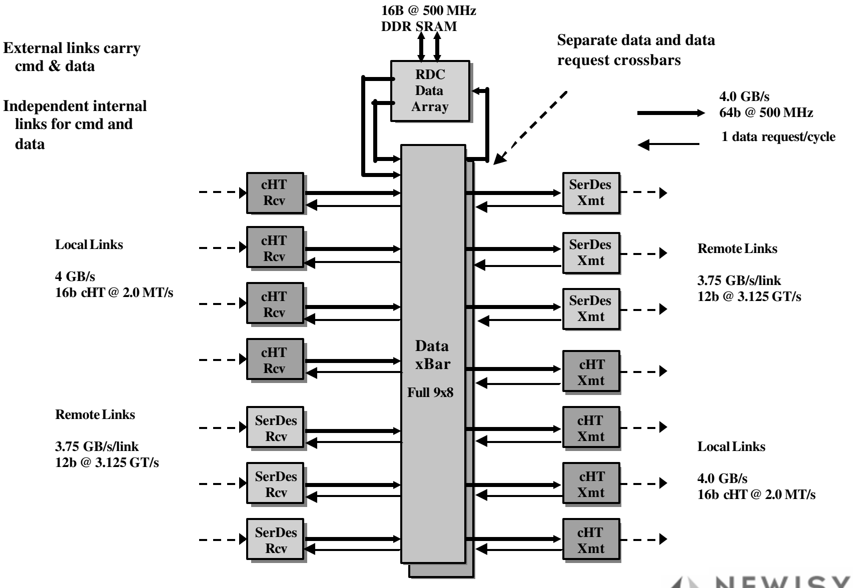
Local Links

4 GB/s
16b cHT @ 2.0 GT/s

8

NEWISYS®
A SANMINA-SCI COMPANY

Horus Data Path



External links carry cmd & data

Independent internal links for cmd and data

Separate data and data request crossbars

Other system features in HORUS

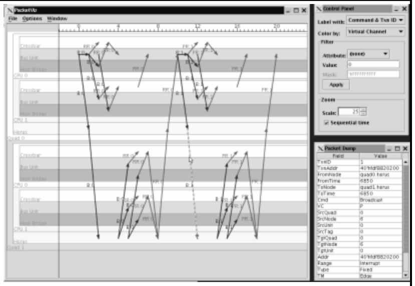
- Partitioning
- Programmable protocol engine
- Highly configurable with configuration control registers
- Reliability features
- Machine check features
- JTAG Mailbox
- Performance counters

Vital Statistics

- Technology: 130nm, TSMC, LVOD
- Core frequency: 500MHz
- Die size: 19mm x 18.3mm
- Gate count excluding memory: ~ 10 Million
- Transistor count excluding memory: ~ 38 Million
- On-chip repairable SRAM size: 3.75 MB (from Virage)
- Verilog LOC: ~ 115K + 50K (auto generated) + 20K (verilog libraries)
- Verification LOC: > 700K (Vera, Java, C++)
- IO pin count: 730
- P&G pin count: 479
- Expected power consumption: 35 to 45 W
- Hardmacros: HT, SerDes, PLLs (from Artisan)
- Current Status: Taped out and in TSMC fab
    - » Bring up and system validation in Fall 2004

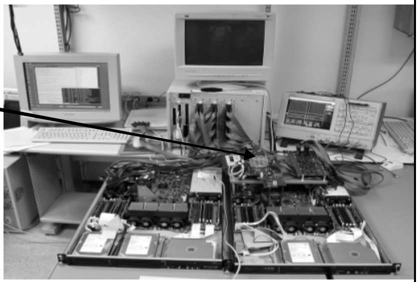
Horus Design Verification

- Multiple Strategy Design Verification
    - C++ / SystemC for original architecture modeling
    - Detailed dynamic performance model using SystemC
    - Object-oriented Synopsys Vera / VCS for RTL simulation environment.
    - Behavioral and Opteron RTL CPU model stimulus
        - Co-simulation with Opteron Northbridge RTL and AMD's whackers / checkers
    - FPGA prototype of single protocol-engine Horus combining multiple Newisys 2100 servers into single Coherent system.
    - Test chip fabricated and tested with memory interface, HyperTransport Interface, and SERDES

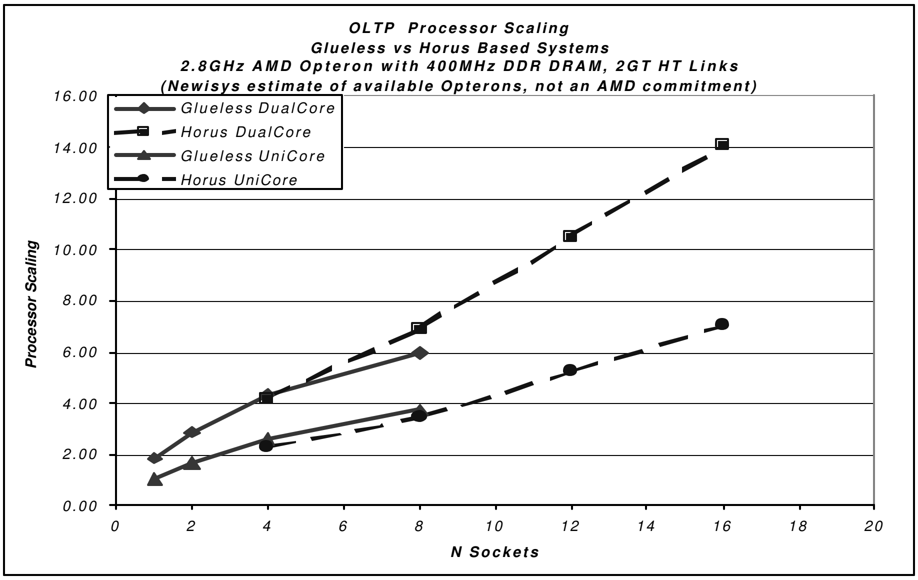


Label with: Command & Txn ID
Color by: Virtual Channel
Filter
Attribute: (none)
Value: 0
Scale: 25
☑ Sequential time

Field	Value
TxnID	1
TxnAddr	40'hfdf88202 00
FromNode	quad0.horus
FromTime	6850
ToNode	quad1.horus
ToTime	6850
Cmd	Broadcast
VC	P
SrcQuad	0
SrcNode	6
SrcUnit	0
SrcTag	0
TgtQuad	0
TgtNode	6
TgtUnit	0
Addr	40'hfdf88202 00
Range	Interrupt
Type	Fixed
TM	Edge



Performance Projections



13

NEWISYS®
A SANMINA-SCI COMPANY

System Management

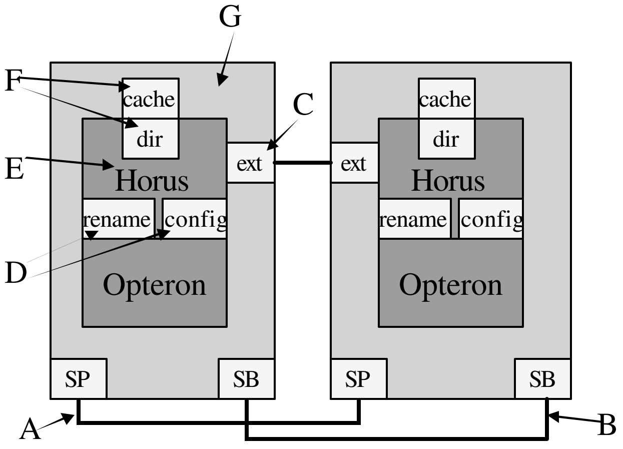
- Horus provides Coherent memory interconnect building blocks, but a complete solution to the single SMP system requires more:
  - Embedded Service Processor and special interconnect hooks
  - Two Service Processors with independent System Management code
    - one primary and one redundant in each system.
  - System Management code deals with configuration control, partitioning, various RAS issues and managing the various hardware hooks for Power On/Off, Reset, Hard and Soft IPL, HT Stopping and Restarting, etc.

14

NEWISYS®
A SANMINA-SCI COMPANY

Horus IP Classification

(37 patents filed)



A. SP roles, failover/takeover, partitioning (3)
B. SouthBridge – config, resets and LDTSTOP (2)
C. External Links – cHT mapping, RAS (3)
D. Rename & Config – including interrupts (7)
E. Horus Implementation (16)
F. RDC & RDIR cHT optimizations (4)
G. Future Optimizations (2)

Summary

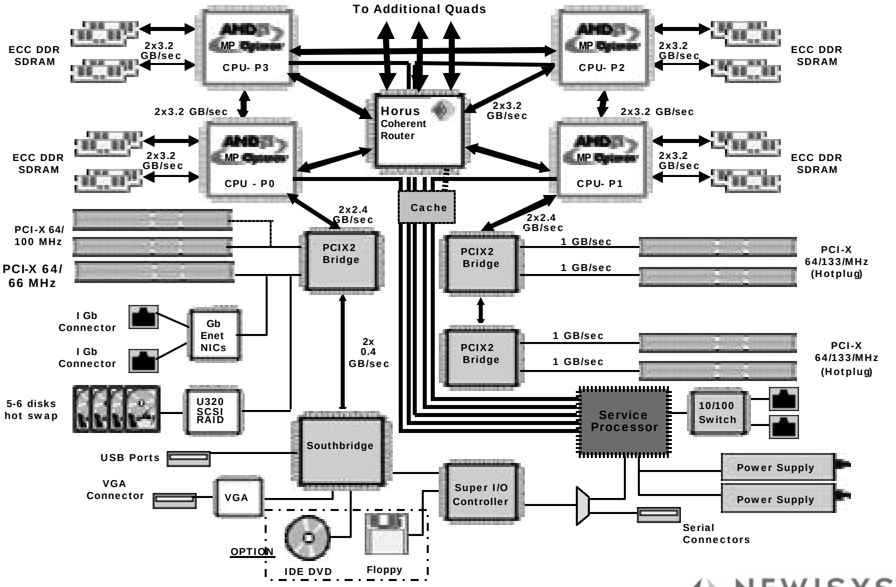
- Horus is based on AMD’s Coherent HT protocol
  - Relies on Home based, single point line synchronization
- Horus extends AMD’s protocol by
  - Significantly increasing the size of the largest systems
  - Introducing a Remote Data Cache for rapid presentation of cached data
  - Adding a Remote Directory for probe filtering
  - Providing RAS at the level expected of enterprise class servers
- Horus remote link technology allows distinct quad implementations, using industry standard parts, to be coupled into very large SMP implementations
- Newisys is building systems based on Horus

Horus Team

Q & A

Back Up

Horus validation platform



To Additional Quads

ECC DDR SDRAM

2x3.2 GB/sec

CPU- P3

CPU- P2

CPU - P0

CPU- P1

2x3.2 GB/sec

Horus Coherent Router

Cache

2x2.4 GB/sec

PCI-X 64/ 100 MHz

PCI-X 64/ 66 MHz

PCIX2 Bridge

1 GB/sec

PCI-X 64/133/MHz (Hotplug)

I Gb Connector

Gb Enet NICs

2x 0.4 GB/sec

5-6 disks hot swap

U320 SCSI RAID

Southbridge

Service Processor

10/100 Switch

USB Ports

VGA Connector

VGA

Super I/O Controller

Power Supply

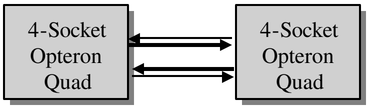
Serial Connectors

OPTION

IDE DVD

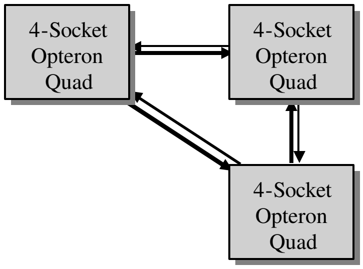
Floppy

Building Larger Configurations


4-Socket Opteron Quad

4-Socket Opteron Quad

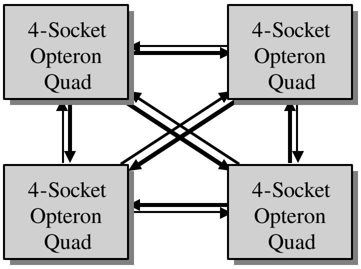
Typical 8-way


4-Socket Opteron Quad

4-Socket Opteron Quad

4-Socket Opteron Quad

Typical 12-way


4-Socket Opteron Quad

4-Socket Opteron Quad

4-Socket Opteron Quad

4-Socket Opteron Quad

Typical 16-way

Up to 32 Sockets (8 quads) possible

Scalability

- One Horus chip in each box. And each box can have upto 4 Opteron sockets (aka Quad)
- Horus uses Coherent HT protocol to talk to Opterons
- Using Horus and IB cables, different quads with independent clock and power domains are connected via remote links
- The protocol extensions used on remote links enables us to run Coherent protocol on cables
- Horus looks just like an Opteron to other Opterons in the quad and it abstracts all other Opterons (CPUs, MCs and IOs) in remote boxes
- Horus has the protocols necessary to maintain coherency across all quads

Remote memory access latency

- Horus supports 64MB of Remote Data Cache (RDC). Requests to Memory lines that hit in RDC will result in the transaction completing 4.7x faster than not having RDC
- The cache is implemented using off-chip SRAM (500MHz or 250MHz DDR)
- The tags for the RDC are on-chip
- Only data whose home is located in remote quads is cached in RDC

Local memory access latency

- Horus also has a Directory that keeps track of the state of local memory lines
- For each local memory line that is cached remotely Horus maintains its state (Shared, Owned, Modified) and Occupancy Vector
- Directory is sparse and will cause eviction of memory lines from remote quads if needed
- For requests from local CPU accessing local memory a miss in Directory will cause the transaction completing 3x faster than not having Directory

Improves bandwidth usage

- More hits in RDC means fewer requests on remote links
- More misses in DIR means fewer probes on remote links
- Due to DIR, remote probes are not broadcasted but directed to specific quads
- With RDC and DIR combined there is significant reduction in bandwidth usage on remote links

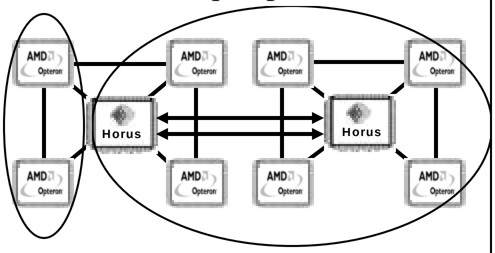
Miscellaneous Features

- Apart from handing transactions targeted at DRAM address, Horus handles transactions to local and remote MMIO, PCI-Config, PCI-IO, Locks, System Management, Interrupts and SEM
- All functions supported by the Opterons for glue-less SMP are extended by Horus across multiple boxes

Partitioning

- Hardware hooks present in Horus to allow dynamic partitioning on remote links. Support is required in OS to achieve dynamic partitioning
- Full hardware support present in Horus to hot plug and unplug remote links. Horus can interrupt SP on hot plug, unplug and errors on remote links

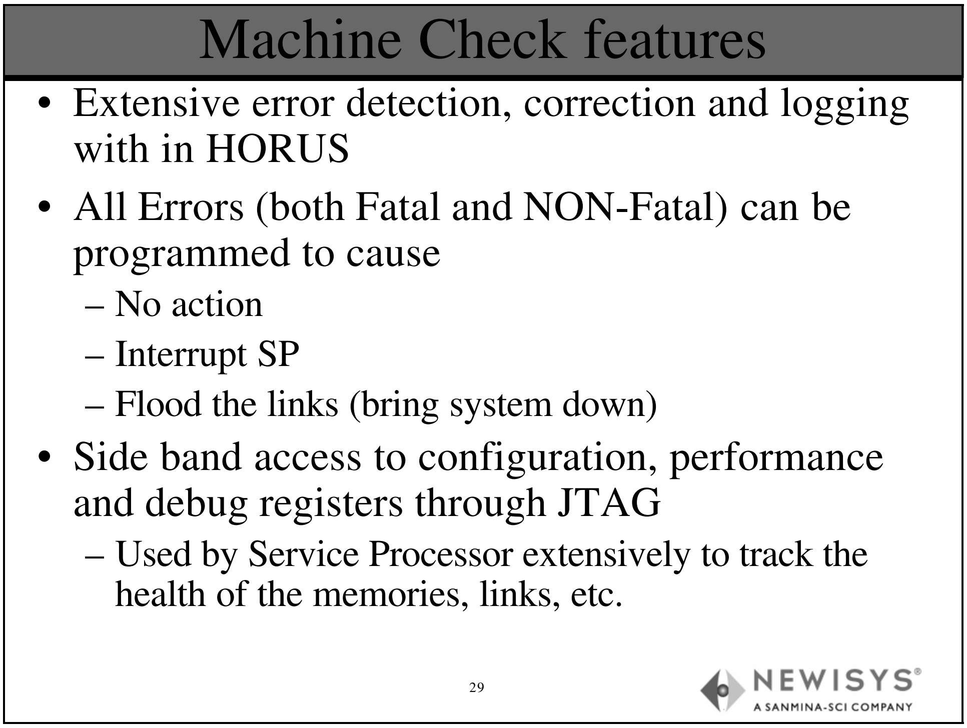
Two unequal partitions



- Static partitioning can be done on both local and remote links. BIOS can program HORUS to enable/disable different remote links
- Horus can't stride multiple partitions. Horus and Opterons have features to fence one partition from another partition

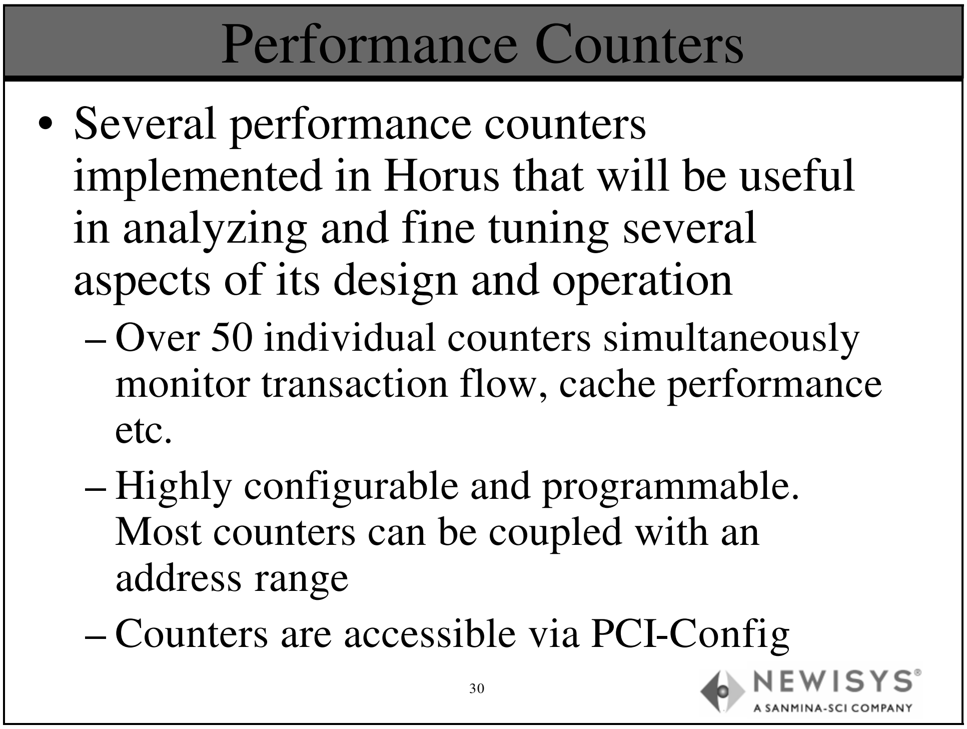
Reliability Features

- ECC on all on-chip and off-chip SRAMs (double bit detect, single bit correct)
- Scrubbing on all on-chip and off-chip SRAMs
- Guaranteed exactly once delivery protocol on remote links
  - All soft errors are recoverable (Disparity, Out of Band, LOS, FIFO overflow on physical layer. CRC mismatch, Loss of packet, Seq ID mismatch, illegal packet)
  - Re-initialization of remote links without box reset.
  - BOX ID exchange on link up



Machine Check features

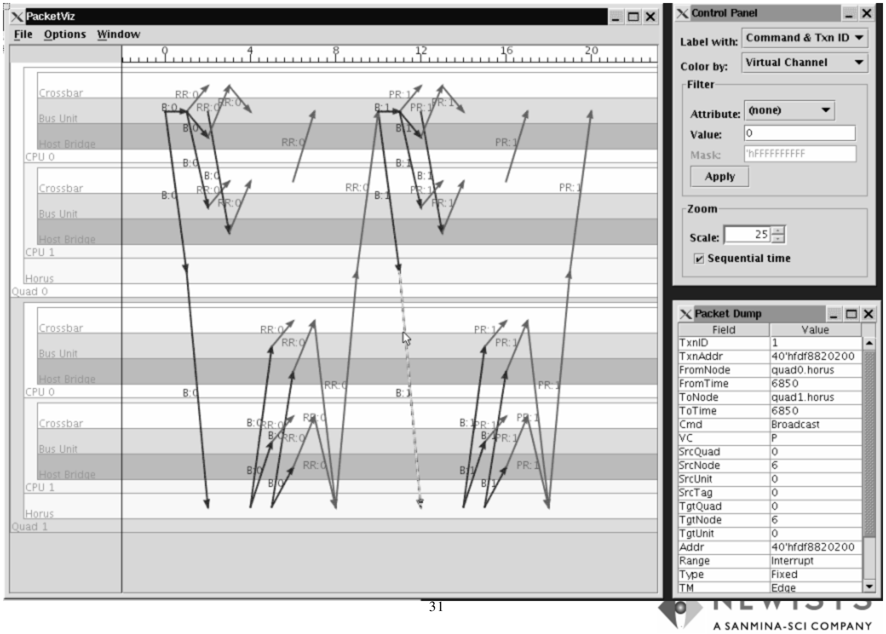
- Extensive error detection, correction and logging with in HORUS
- All Errors (both Fatal and NON-Fatal) can be programmed to cause
  - No action
  - Interrupt SP
  - Flood the links (bring system down)
- Side band access to configuration, performance and debug registers through JTAG
  - Used by Service Processor extensively to track the health of the memories, links, etc.



Performance Counters

- Several performance counters implemented in Horus that will be useful in analyzing and fine tuning several aspects of its design and operation
  - Over 50 individual counters simultaneously monitor transaction flow, cache performance etc.
  - Highly configurable and programmable. Most counters can be coupled with an address range
  - Counters are accessible via PCI-Config

Transaction Visualization Tools



Control Panel

Label with: Command & Txn ID

Color by: Virtual Channel

Filter

Attribute: (none)

Value: 0

Mask: 'hFFFFFFFFFF

Apply

Zoom

Scale: 25

☑ Sequential time

Packet Dump

Field	Value
TxnID	1
TxnAddr	40'hfdf8820200
FromNode	quad0.horus
FromTime	6850
ToNode	quad1.horus
ToTime	6850
Cmd	Broadcast
VC	P
SrcQuad	0
SrcNode	6
SrcUnit	0
SrcTag	0
TgtQuad	0
TgtNode	6
TgtUnit	0
Addr	40'hfdf8820200
Range	Interrupt
Type	Fixed
TM	Edge

A SANMINA-SCI COMPANY