Ideas for a Dependable ‘Industry Standard Architecture’ Platform

Newisys, Inc.

Rich Oehler

27 January 2005

Outline

- Our Company - Newisys
- Our Current Products – 2100 and 4300
  - Under-development - Horus
- Industry Standard Architecture Products
  - Attributes
    - Weaknesses
- Dependable Systems
  - Attributes
- Achieving Dependable System Structures
  - Scaling (both Up and Out)
  - I/O Connectivity and Configuration
  - Systems Management
- Performance Projections
- Summary

Newisys, Inc

- Founded in July 2000
  - Designing Enterprise Class, Rack Mounted, Opteron Based Server Systems for the OEM Market
- Entered into a Strategic Alliance with AMD for access to coherent HyperTransport
  - Began design of a custom ASIC (Horus) to enable large SMP (8 to 32 socket) Opteron Systems
- Acquired by Sanmina/SCI in July 2003
- Bringing up systems based on our custom ASIC
- Currently about 110 employees, ~ 90 Eng/PGM
  - Located in Austin TX

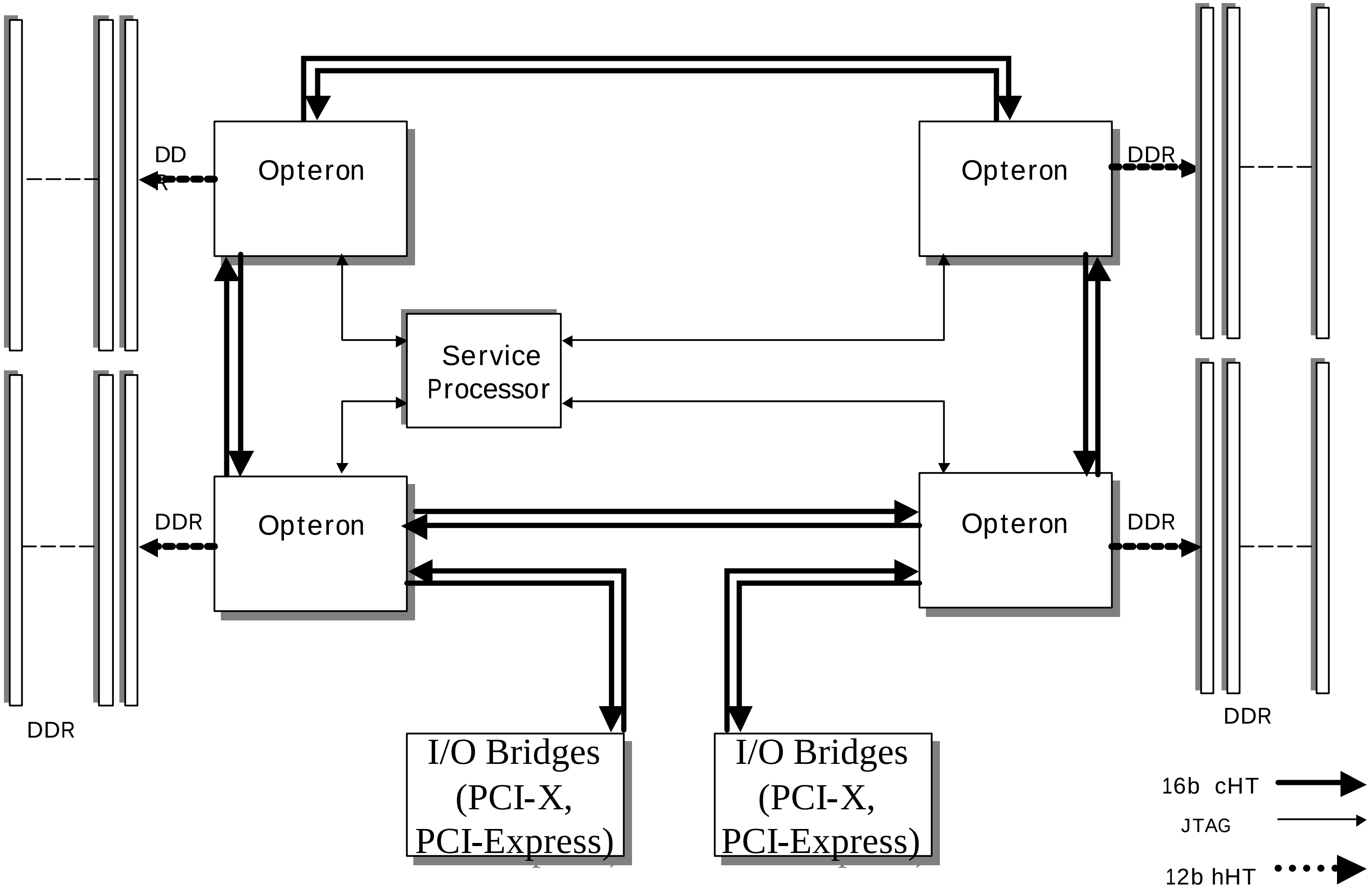
Why Opteron?

- AMD radically changed the system architecture of Industry Standard platforms
- Opteron has 3 point to point links (HyperTransport) on each chip
  - Each link can be used to connect to other Opterons (coherent) or to I/O (non-coherent)
- Opteron has a direct memory interface on each chip

Results:

- Glueless SMP – up to 8 sockets
- Adding Opterons greatly improves scalability
  - More memory capacity and bandwidth
  - More coherency bandwidth
  - More I/O bandwidth

Typical 4 Socket (Quad) Opteron System



DD
R

Opteron

Opteron

DDR

Service Processor

DDR

Opteron

Opteron

DDR

DDR

DDR

I/O Bridges (PCI-X, PCI-Express)

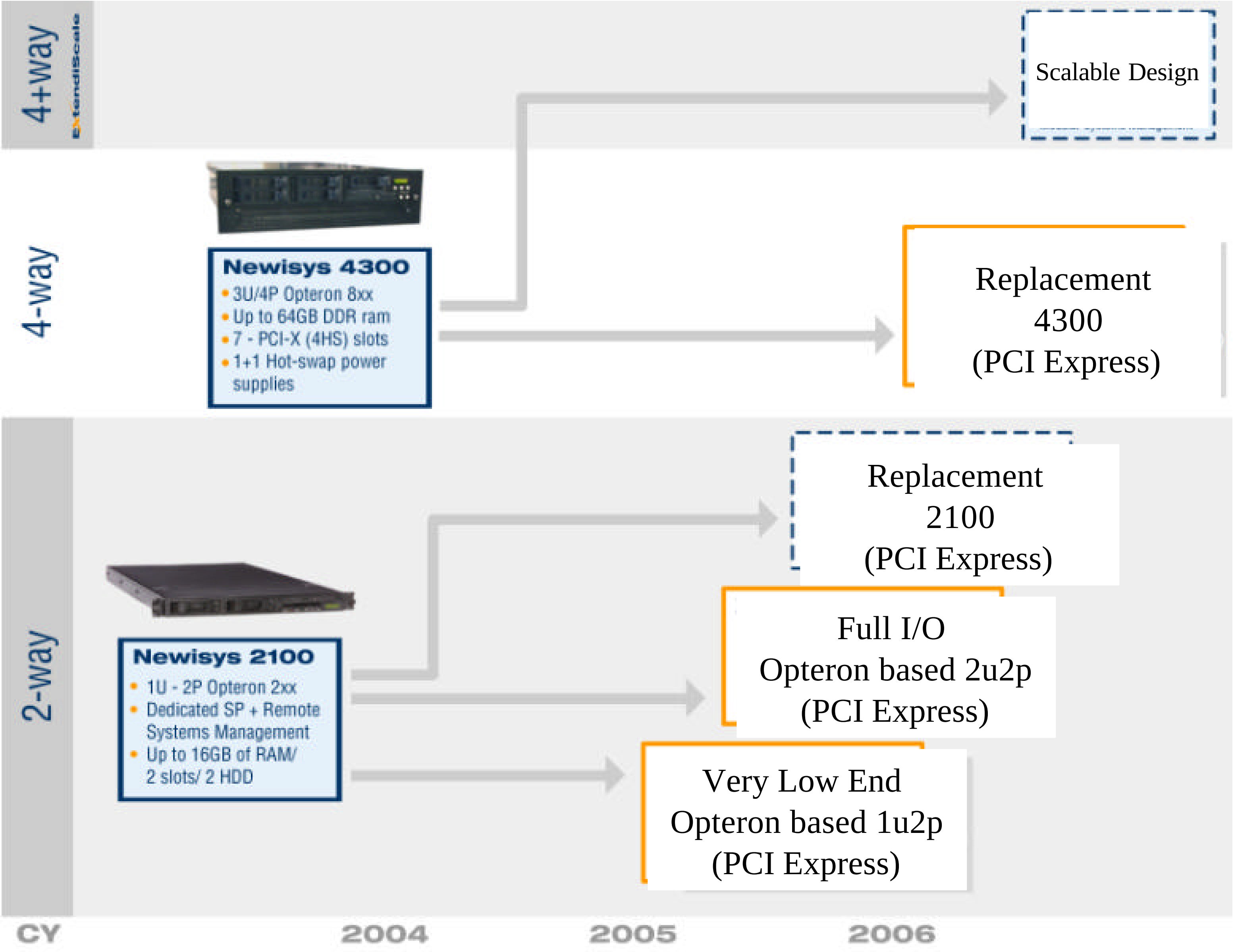
I/O Bridges (PCI-X, PCI-Express)

16b cHT

JTAG

12b hHT

Newisys Product Roadmap



Newisys 4300

- 3U/4P Opteron 8xx
- Up to 64GB DDR ram
- 7 - PCI-X (4HS) slots
- 1+1 Hot-swap power supplies

Replacement
4300
(PCI Express)

Newisys 2100

- 1U - 2P Opteron 2xx
- Dedicated SP + Remote Systems Management
- Up to 16GB of RAM/ 2 slots/ 2 HDD

Replacement
2100
(PCI Express)

Full I/O
Opteron based 2u2p
(PCI Express)

Very Low End
Opteron based 1u2p
(PCI Express)

Limits of Scalability on Opteron

- Opteron provides for up to 8-socket ‘glueless’ SMP solution
- Opteron has very good Scaling to at least 4-socket
- Performance of important commercial applications is challenging above 4-socket due to:
  - Link interconnect topology (wiring and packaging)
  - Link loading with less than full interconnect (even less than 3 links)
- Going above 8-socket needs both:
  - Fix to number of addressable sockets
  - Better interconnect topology
- Ever larger Coherency Fabric will increase delays (loading/queuing) and become the major obstacle to good SMP scaling

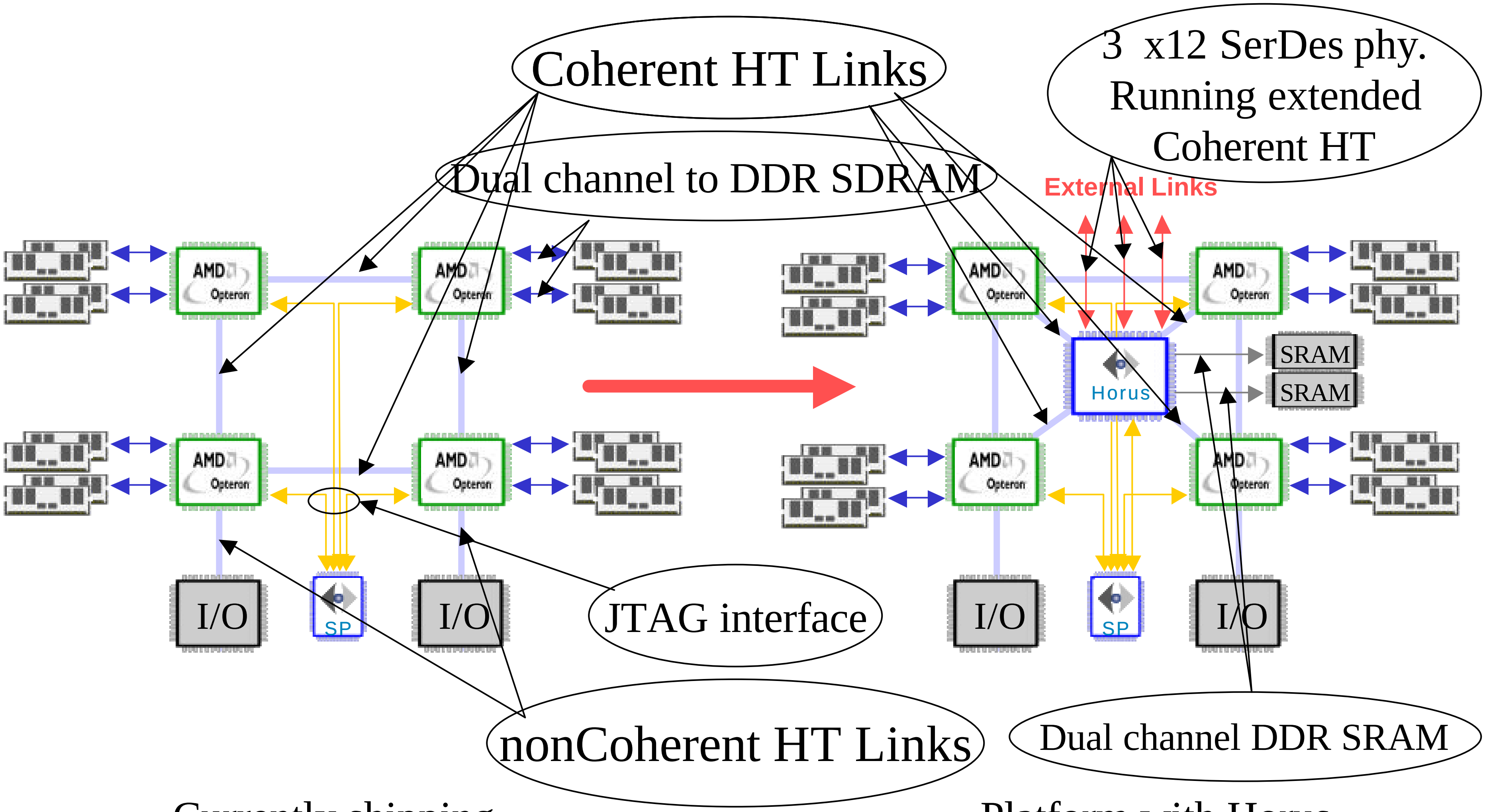
Solving the Fundamental Problem

- Combine multiple four socket quads into a larger coherent domain…
- But local quads have no knowledge of “remote quads” (CPUs, I/O or Memory) outside of the their own local space
- So our approach is to add into each quad a “fifth” socket that abstracts all of the remote quads
  - Acts as a “cache” for local request probing
  - Acts as a “memory controller” for requests to remote memory space and from remote CPUs
  - Acts as a “CPU” for requests from remote nodes
- And to place in all of the Opteron sockets an abstraction of all of the remote resources

Horus – Newisys Custom ASIC

- Defines a coherence mechanism to support two or more 4-socket AMD Opteron quads
  - Built into our standard 4 socket rack building block
  - Industry Standard Servers (Industry Standard Pricing)
- Acts as a Distributed Router in the coherency domain
  - Multiple Horus are connected by an extension of coherent HyperTransport
  - Direct connect (cut through) to non adjacent quads
- Adds facilities to reduce coherency traffic
  - Remote Directory, Remote Data Cache
- Provides a management point and performance optimization point
  - Partitioning between/among quads

Scalable Newisys Opteron Systems



Coherent HT Links

3 x12 SerDes phy. Running extended Coherent HT

Dual channel to DDR SDRAM

JTAG interface

nonCoherent HT Links

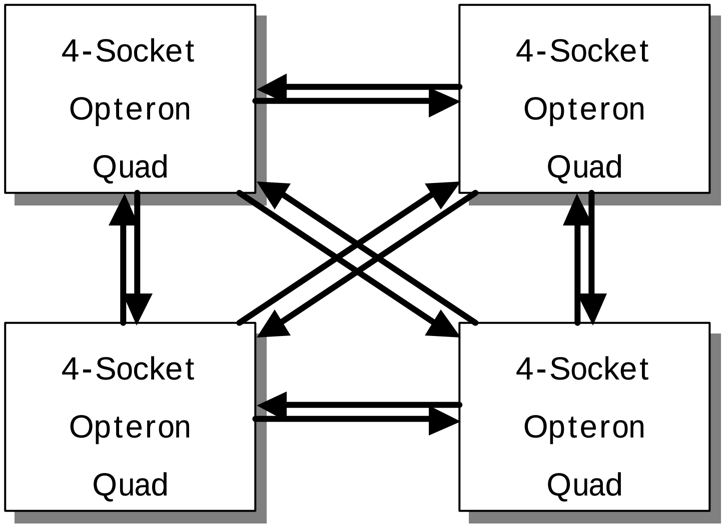
Dual channel DDR SRAM

Currently shipping
Newisys 4300 platform

Platform with Horus

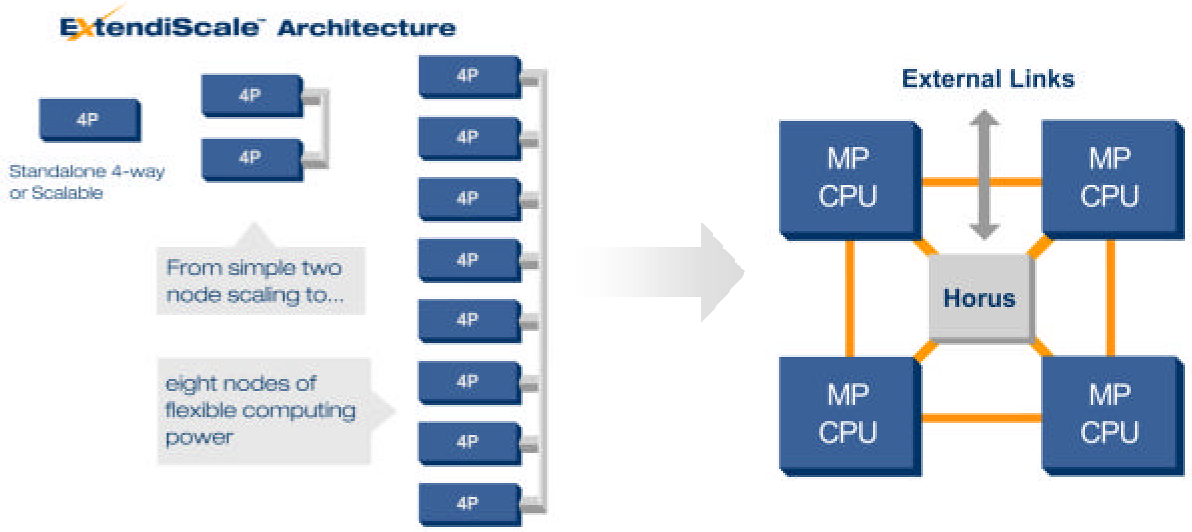
Building Larger Configurations

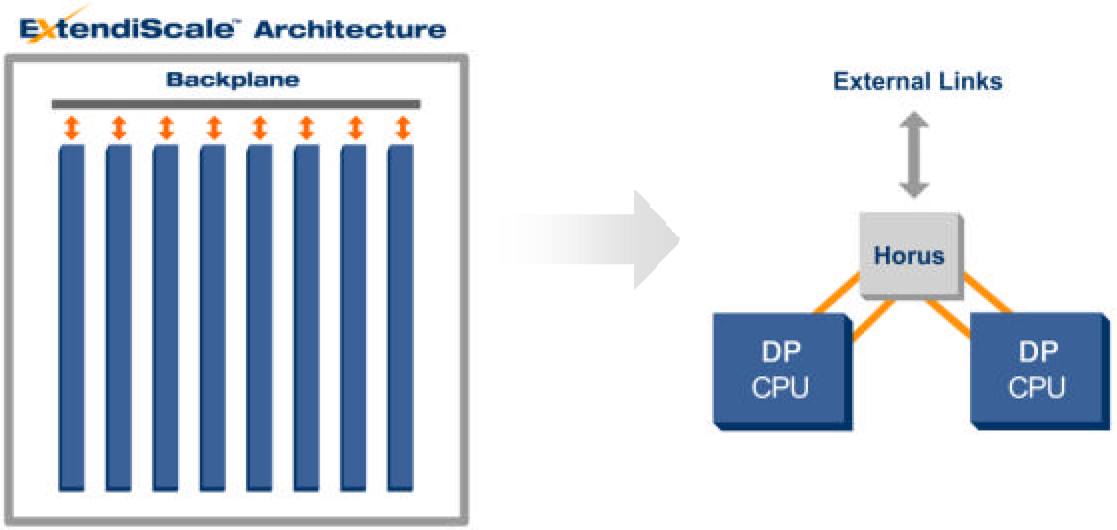
Typical 16-way



Up to 32 Sockets possible

Newisys ExtendiScale Architecture


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



- Exceptional performance headroom
- Enables modular systems
  - Traditional 8-64 way CC SMP (Dual core)
  - Blade frame 2-32 way CC SMP (Dual core)
- The ExtendiScale Architecture delivers:
  - Pay as you grow budget flexibility
  - RISC/UNIX replacement at a fraction of the cost
  - Mission Critical ready: Availability, Manageability, Reliability

What makes hardware dependable?

- Hardware that never fails; or if it does, self heals; has no loss of data or incorrect results; or if it does, contains and identifies the error; adjusts to workloads without bogging down; or if it does, can apply additional or spare resources; …
  - – Typically (Very?) expensive
  - – Certainly custom design
- Are there different design points for dependability? Can Industry Standard Servers be made dependable enough?
  - – Certainly lower cost
  - – How much dependability is required / sufficient?
    - Software can make up for many hardware deficiencies
      - – At what cost? Performance?

Acceptability of Industry Standard Servers

- Industry Standard Servers suffer from
  - silent failures, catastrophic failures, lock up failures
- Newisys is building enterprise class servers out of Industry Standard parts.
  - Our hardware systems are much more reliable than those produced by Taiwan Inc. (better engineering)
  - Our incremental cost is marginal
- Our System Management with an out of band Service Processor fixes even more problems not solved in current Industry Standard parts

Focus on Newisys Opteron Blades

Disclaimer – not currently on our road map

- Built around 2 socket CPU Blades and I/O Blades
- Coherency Fabric connects all CPU Blades
  - Used to configure larger than 2 socket SMP systems
  - Each CPU Blades also develop at least 2 connections to an I/O Fabric based on PCI-Express
- I/O Fabric connects all I/O Blades with connections to each CPU Blade
  - I/O Fabric contains a switch (two for redundancy)
    - Based on Advanced Switching or more specialized solutions
  - I/O Blades can be dedicated or shared

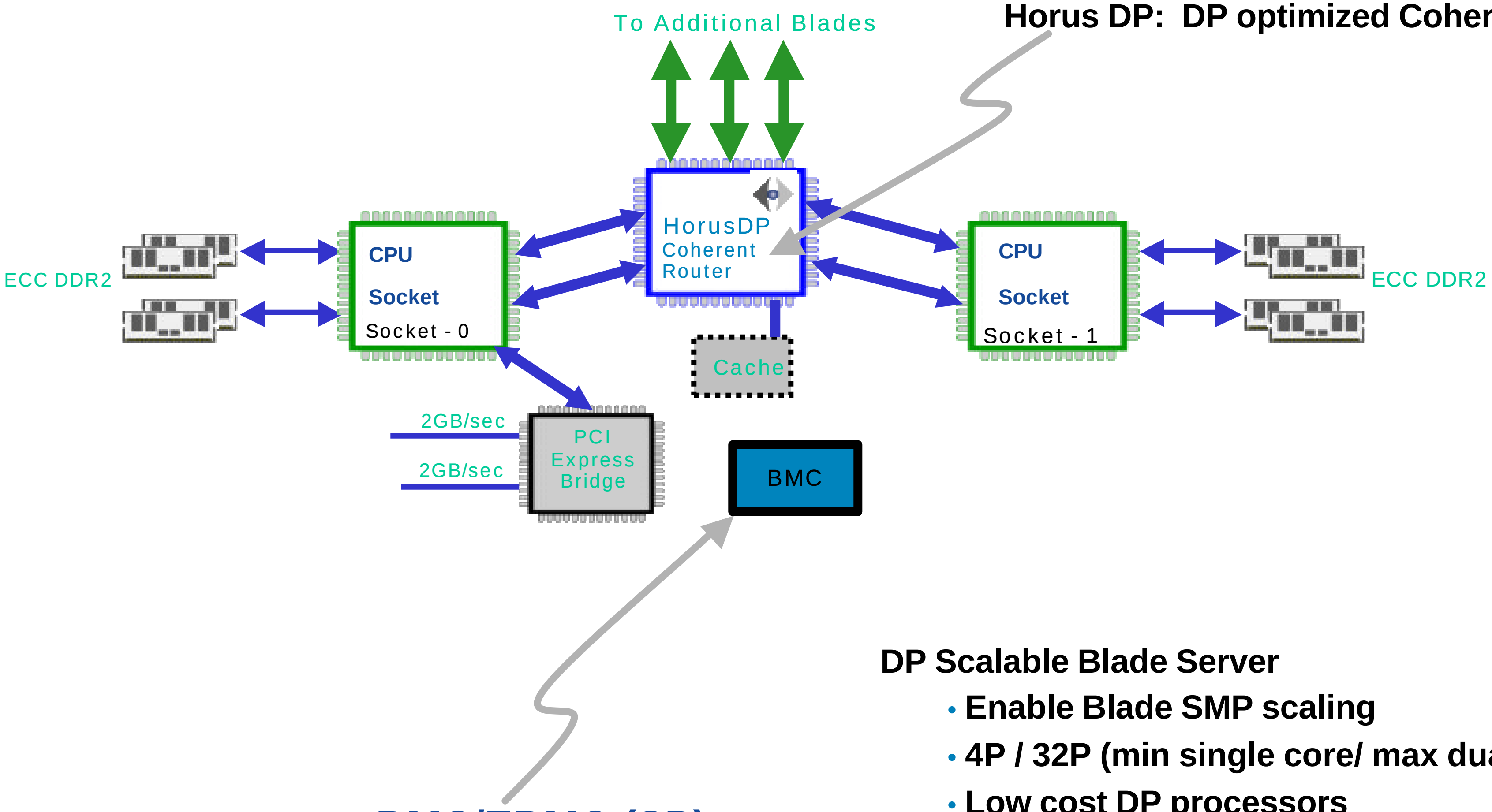
Why Blades?

- Blades are not about power packaging and cooling (although these problems are hard and getting harder and must be solved)
- Blades are not scaled down systems
  - Large and Powerful systems can be built as Blades
- Blades are about defining a uniform set of structures over which many problems are solved in a systematic way
  - Provisioning
  - Configuration (including partitioning)
  - Recovery (including hot swap, fail over, …)
  - Maintenance and Repair
  - Alignment of hardware boundaries with application boundaries
  - …

Why Scale Up?

- For many web applications scale out is the best answer
  - Especially near the edge of the net (tier 1 and 2)
- But for many tier 3 applications, the answer is not obvious
  - Lots of existing large monolithic databases and their associated applications
  - Some problems/applications just don’t partition well
    - Pieces are too small, synchronization cost too high
- Newisys Blades can do both scale up and scale out
  - Can be configured/controlled to go from scale out to scale up and back as needed by policy, workload, …

Scalable DP Blade



Horus DP: DP optimized Coherent Router

BMC/EBMC (SP)

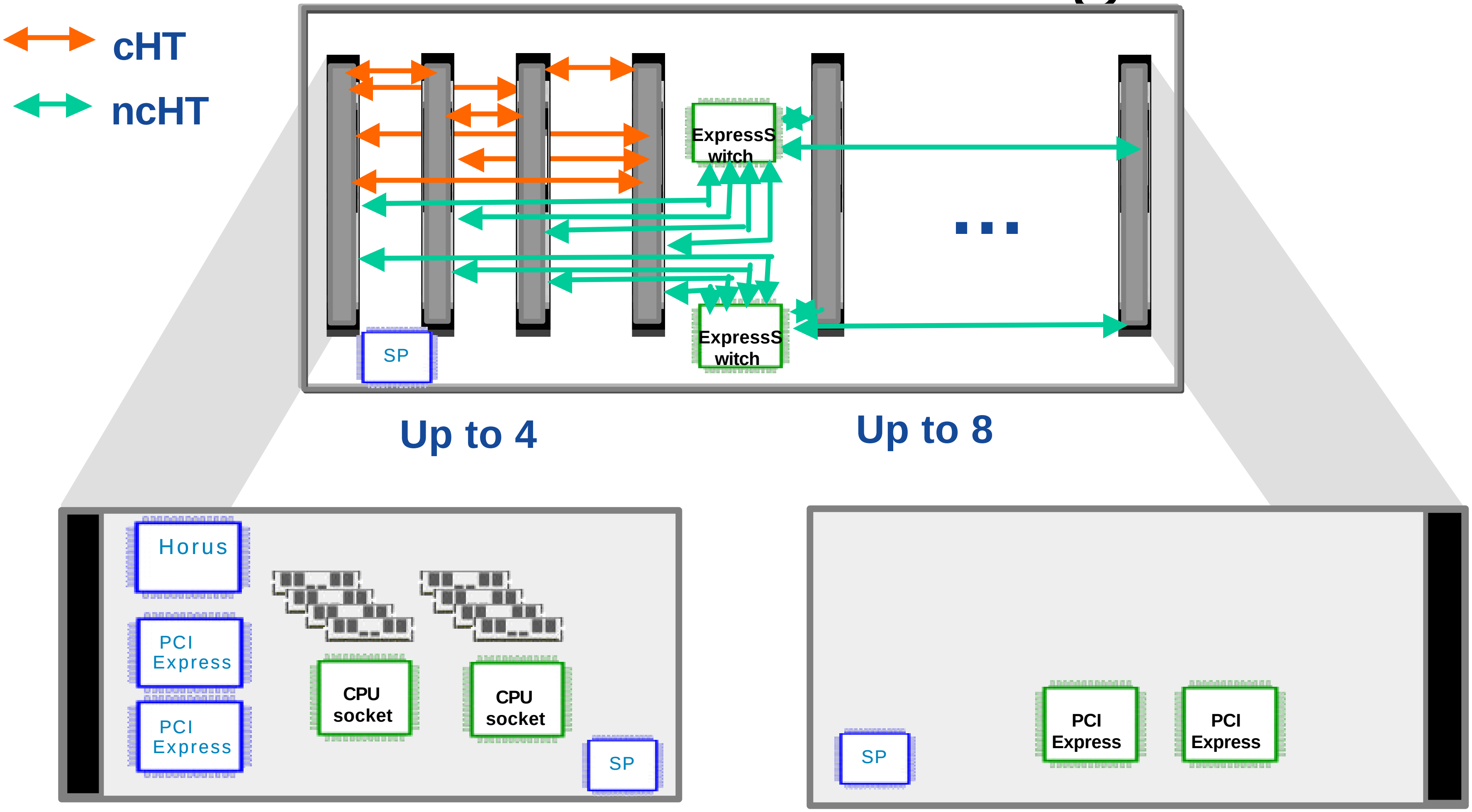
DP Scalable Blade Server

- **Enable Blade SMP scaling**
- **4P / 32P (min single core/ max dual core)**
- **Low cost DP processors**
- **Up to 32GB per Blade**

PCI Express Attributes

- Aggregated very high speed I/O lanes
  - Each lane can be 2Gb/second (today)
  - 16, 24, 32 lanes can be bundled together
- ‘Advanced Switching’ Technology exists today
  - Defined to map up to and down from PCI Express
- Several Startups working on direct PCI Express switching
- Controllers / adapters can be
  - Dedicated (1 to 1) with a system
    - Examples: today’s storage, network controllers (HBA)
  - Shared (1 to n) with multiple systems
    - Examples: shared 10Gb Ethernet adapter, shared FC adapter

Blade Mid-Plane Diagram



Up to 4

Up to 8

CPU Blade

I/O Blade

Shared

Dedicated

Virtualization and Hardware Partitioning

- Virtualization (creating many virtual machines / environments) works really well
- When is it not better to virtualize on a really big system
  - Depends on structure of the really big system
  - If virtualized resources don't correspond to equivalent hardware resources, performance issues may result
    - Many of today's OSs can not match physical resources with virtual resources
  - Again, if no correspondence, hardware failure boundaries may impact many virtual environments (sometimes significantly more)
- Matching real system resources with program resource needs leads to
  - Better performance with dedicated resources
  - More robust execution when errors occur

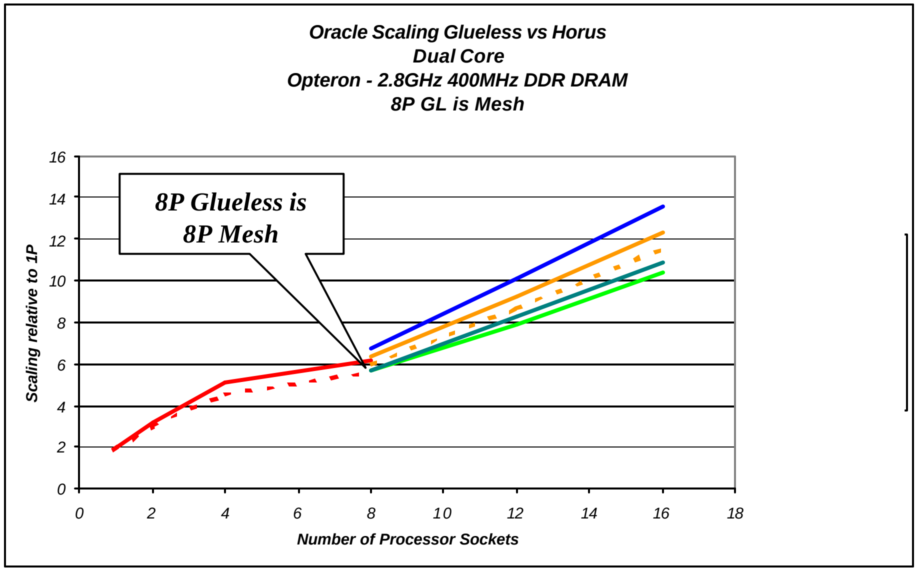
Role of System Management

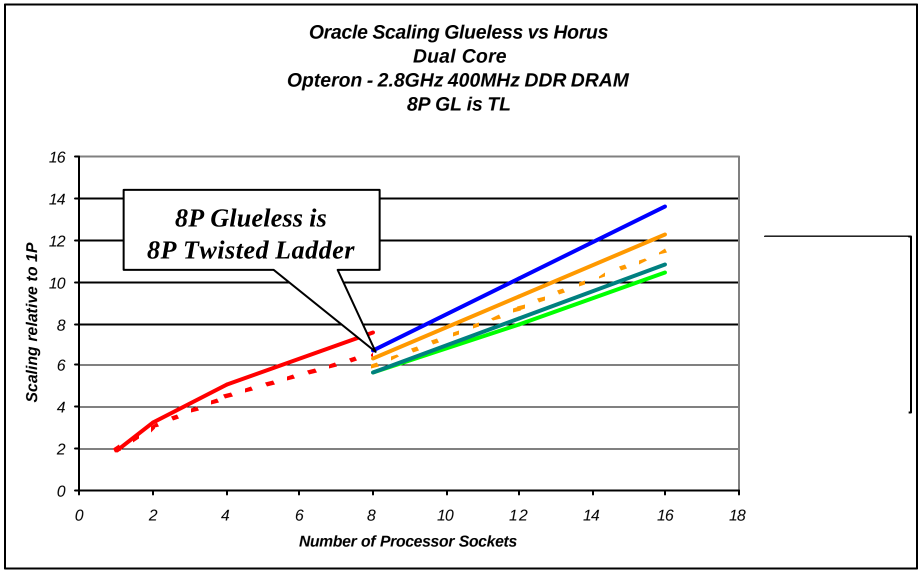
- Separate, out of band management required
- At Several Levels
  - CPU card and I/O card
    - Used for standard environmental controls
    - Also acts as a surrogate during provisioning, configuration and initialization, error detection and recovery
    - Can provide local performance monitoring and local power management
  - At Switch (coherent and non-coherent)
    - Configuration control and performance monitoring
  - At Frame/Rack
    - Overall complex view

Newisys Systems Management

- Horus provides building blocks not a complete solution for a single SMP system
- We use an onboard but independent Service Processor and special interconnect hooks to provide the rest
- There are at least two Service Processors and their system management code, one primary and one fall back in each complex system.
- The system management code deals with configuration control, including partitioning, various RAS issues including watch dog timers and managing the various hardware hooks for Power On/Off, Reset, Hard and Soft IPL, HT Stopping and Restarting, etc.

Scaling – Dual Core





Summary

- Newisys is building robust Industry Standard Servers as well as a Scalability ASIC
- Blades can be built out of Newisys parts that offer
  - SMP scaling through Horus
  - I/O scaling through PCI Express switching
- Newisys Systems Management offers a level of RAS in Industry Standard Serves previously only achievable in RISC/Unix servers
- Dependable Systems can be built out of Newisys building blocks