


ORACLE®



Bixby: the Scalability and Coherence Directory ASIC in Oracle's Highly Scalable Enterprise Systems

Thomas Wicki and Jürgen Schulz
Senior Principal Hardware Engineers, Microelectronics

Hot Chips 25 – August 25-27, 2013



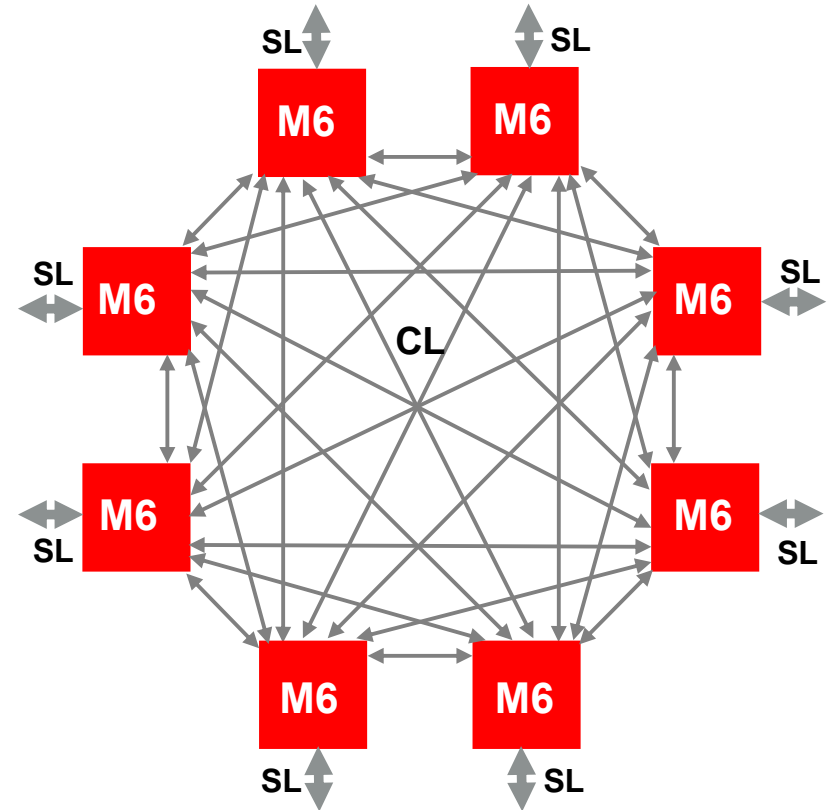
The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Outline

- Motivation and Design Objectives
- M5 System and Beyond
- System RAS Features
- Implementation Details
- Debug and DFT Features
- Summary

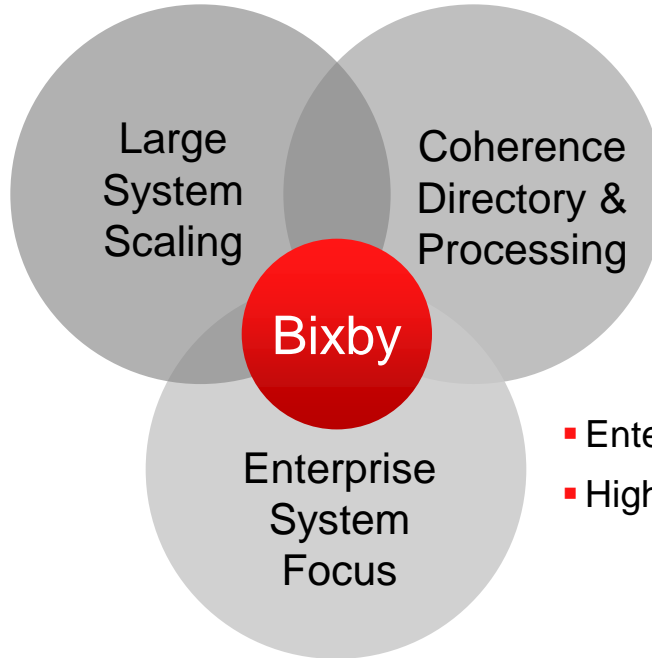
Motivation

- M5 and M6's direct interconnects scale up to 8 processors using Coherence Links (CL) (Glueless system)
- To enable systems to scale beyond 8 processors:
 - Scalability Links (SL) were added to M5 and M6
 - Bixby ASICs are needed (Glued system)



Bixby Design Objectives

- Scalable up to 96 processors
- Communication switch between 8-processor SMPs



- Directory for L3 caches of all processors
- Multi-generation support
- Enabling mixed processor systems
- Enterprise-Class RAS feature set
- High bandwidth, low latency

Challenges and Trade-Offs

	Challenge	Solution
Directory Size	Large directory size requirement	Scale up number of Bixbys with system size
Directory Width	Massive number of L3 cache ways x number of processors per look-up	Pipeline look-ups
Switch Size	24 x 24 crossbar efficiency	Overprovision switching bandwidth
Shared Resources	Some resources shared by multiple hardware domains	Associate errors with single domain and clean up shared resources after error

Outline

- Motivation and Design Objectives
- **M5 System and Beyond**
- System RAS Features
- Implementation Details
- Debug and DFT Features
- Summary

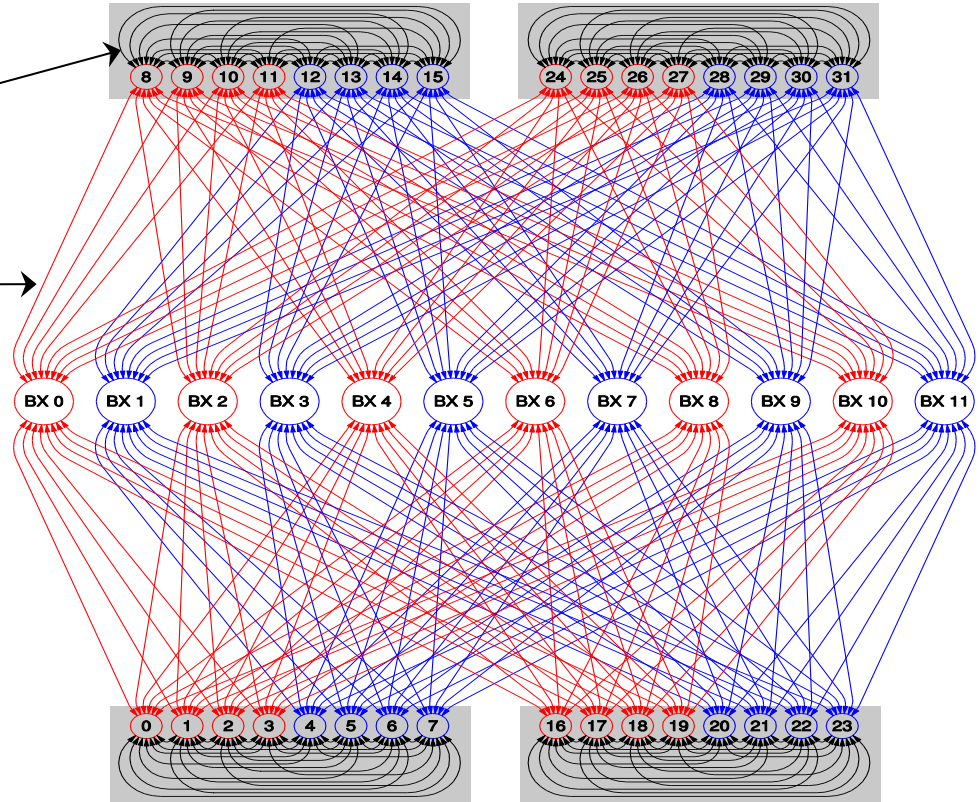
Oracle's M5-32 System

- 32 M5 SPARC processors
- 12 Bixbys
- 4 physical (hardware) domains
- 3.1TB/s payload coherence bandwidth
- 1.5TB/s payload scalability bandwidth

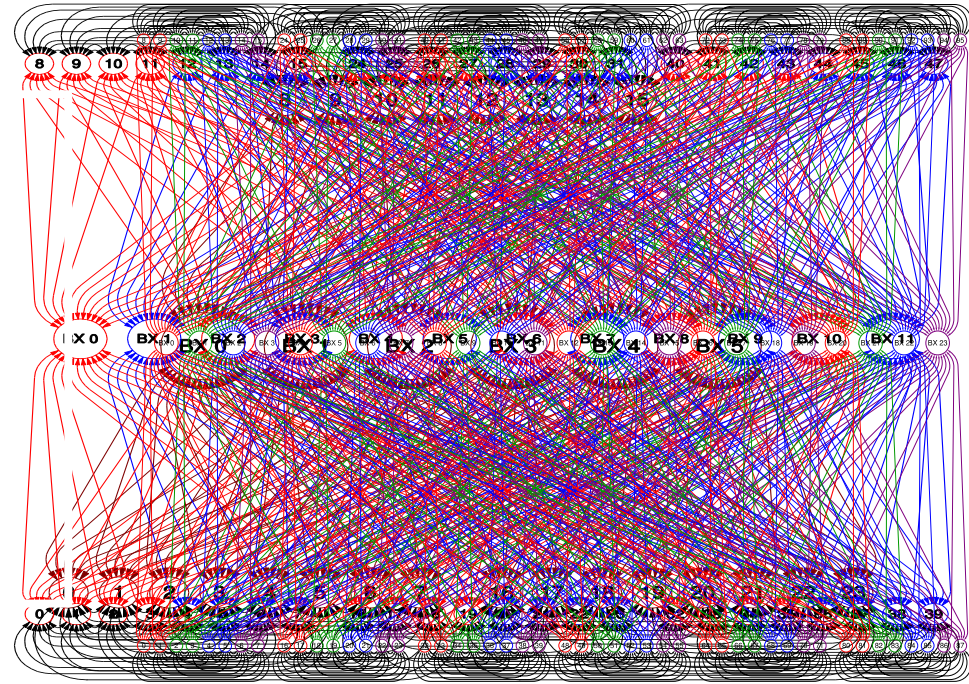
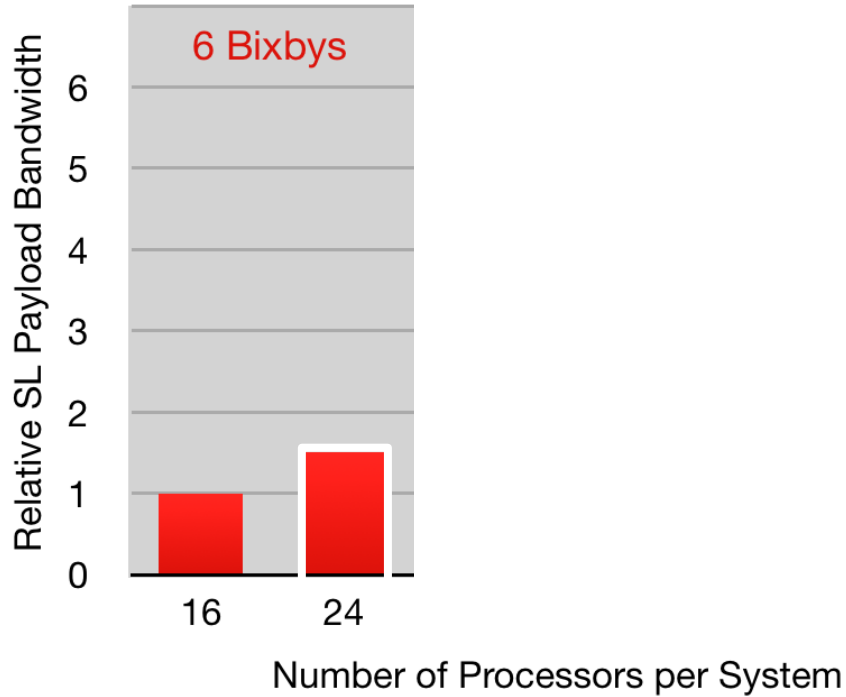


M5-32 System Coherence Interconnect

- Coherence Links (CL)
 - 12 lanes per direction
- Scalability Links (SL)
 - 4 lanes per direction
- 12Gbps per lane
- 7 CLs + 6 SLs per processor
- 16 SLs per Bixby



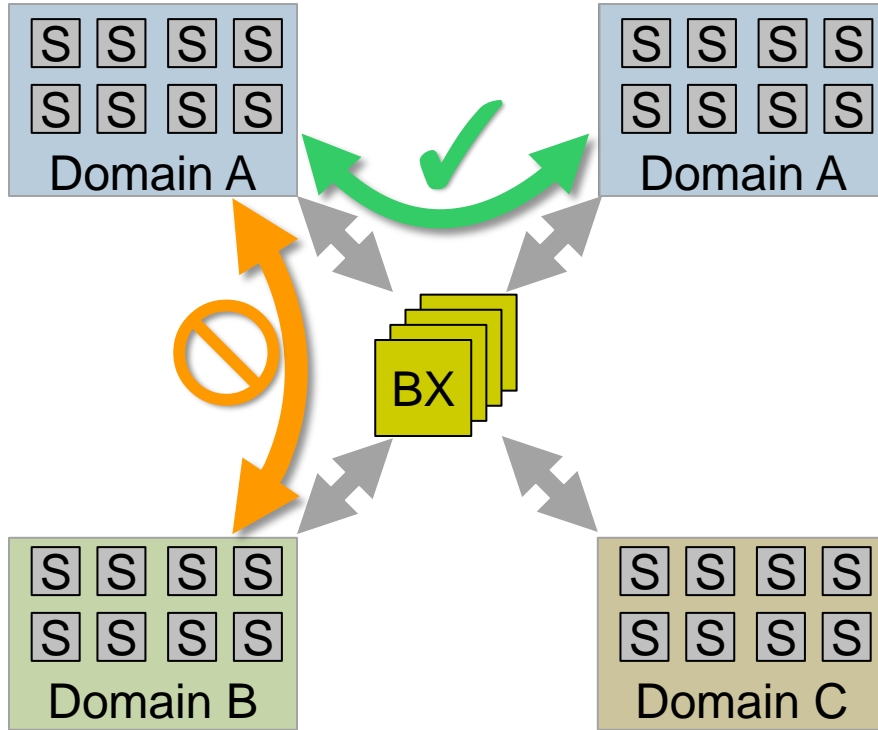
Scalability Link Bandwidth



Outline

- Motivation and Design Objectives
- M5 System and Beyond
- **System RAS Features**
- Implementation Details
- Debug and DFT Features
- Summary

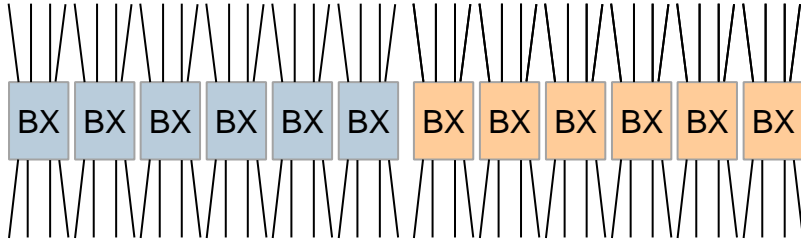
Physical (Hardware) Domain Support



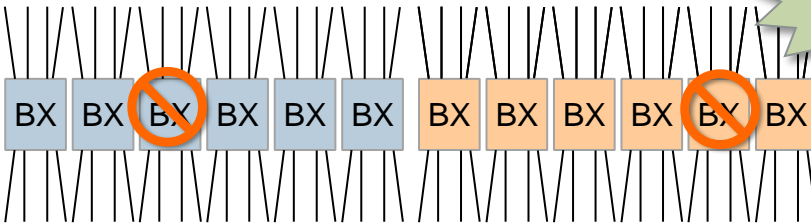
- Up to 12 physical domains
- Dynamically configurable by Service Processor
- Packet filtering and Physical Address fencing
- Errors resolved to physical domain
- Per-domain Cease Operation support

5-of-6 Redundancy Mode

Normal configuration:

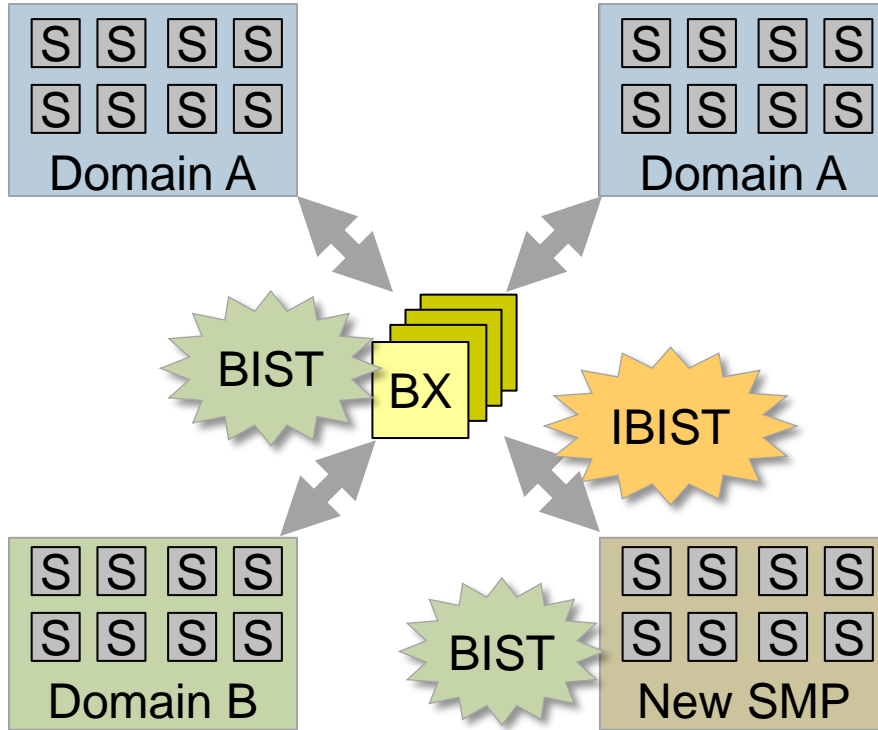


Failover configuration:



- System can boot with any 5 out of each group of 6 Bixbys
- Increases availability since system can be used until service is performed

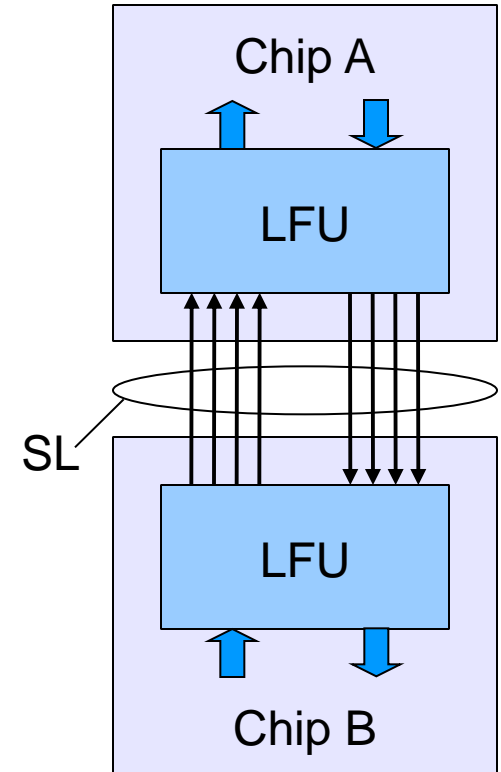
Hot Maintenance Support



- In a running system, failing Bixby or SMP can be:
 - Replaced
 - Tested
 - Re-integrated

Link Protection

- CRC check and auto retry
 - Replay, if CRC error detected
 - Guaranteed lane failure detection
- Built in PRBS testing during link training
- Auto link re-initialization
 - Re-training link, if Replay unsuccessful
 - No Service Processor intervention required
- Auto single lane failover (per direction)
 - Based on PRBS testing
 - No Service Processor intervention required



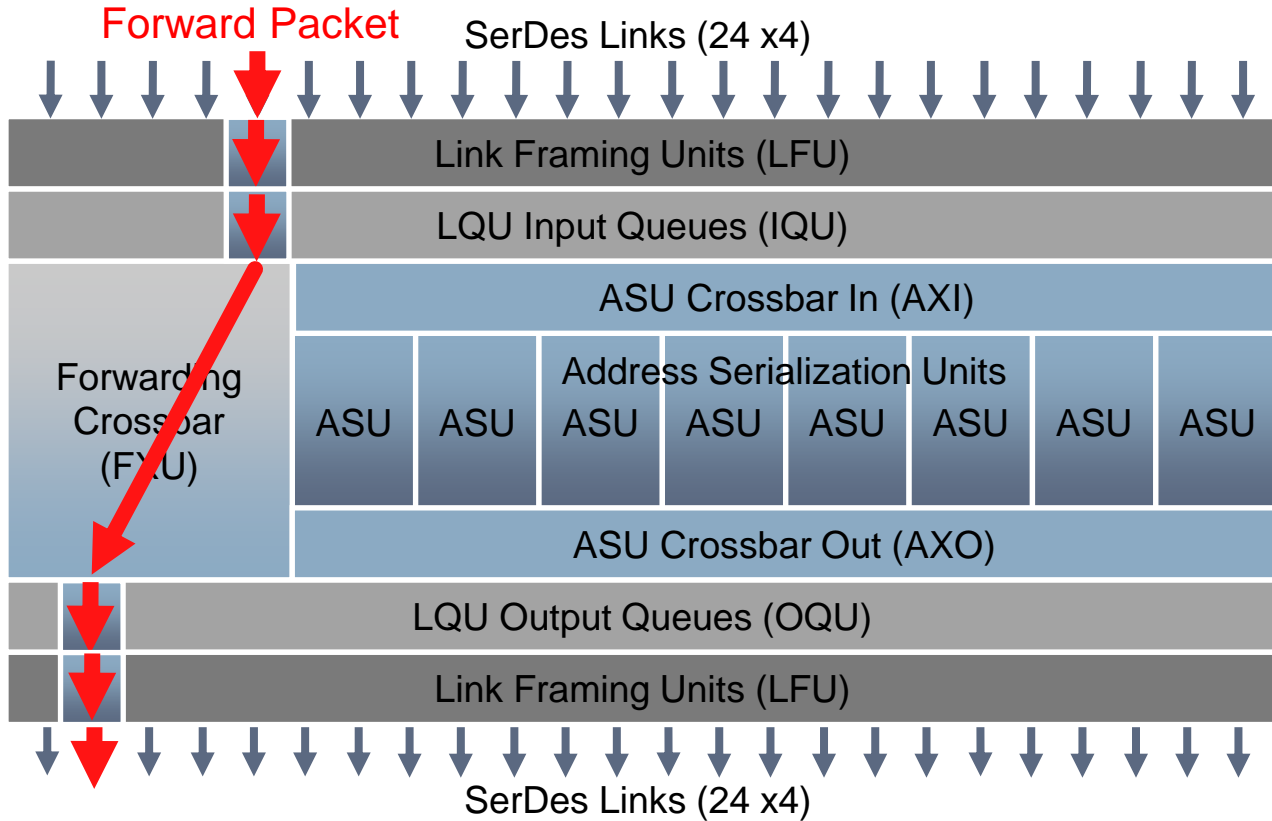
Outline

- Motivation and Design Objectives
- M5 System and Beyond
- System RAS Features
- **Implementation Details**
- Debug and DFT Features
- Summary

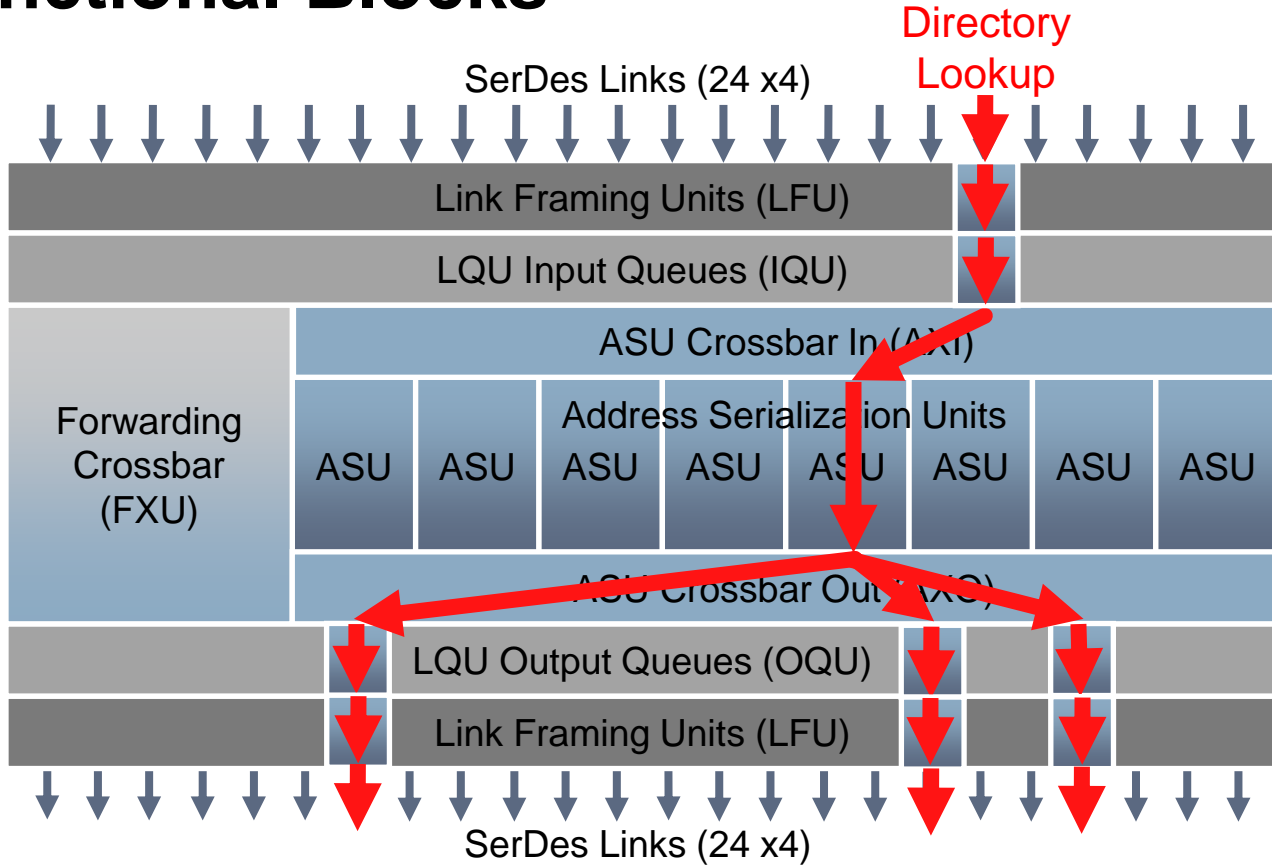
Implementation Details

- 96 Tx + 96 Rx 16Gb/s Long-Reach AC coupled SerDes
- Package: 45mm x 45mm 1677-pin FPBGA (~500 signal IO)
- Process: 28nm 10 layer metal 0.85V ASIC
- ~160 Mbits SRAM (~20MB Tags)
- ~70M Gates (nand2 equivalent)

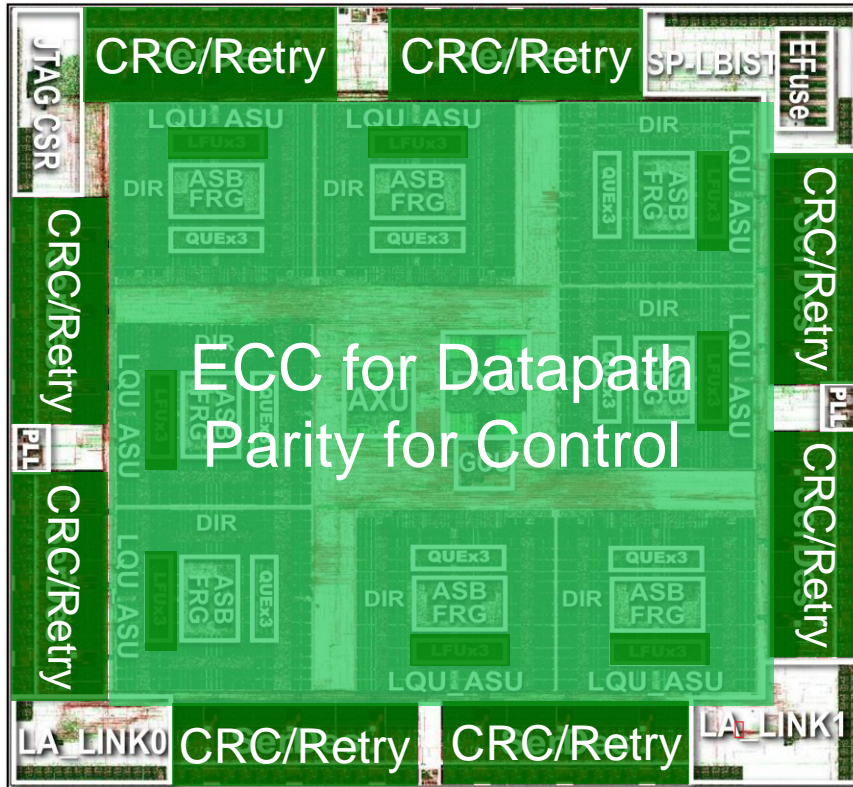
Functional Blocks



Functional Blocks

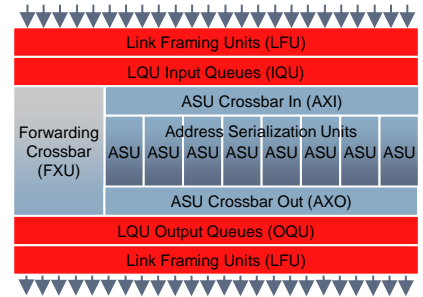


Floorplan



- SEC-DED on all major datapaths
- Parity on control signals
- Custom top level wires on top two routing layers
 - Critical nets implemented by Buffer on route
 - Faster ps/mm
- PVT invariant clock distribution

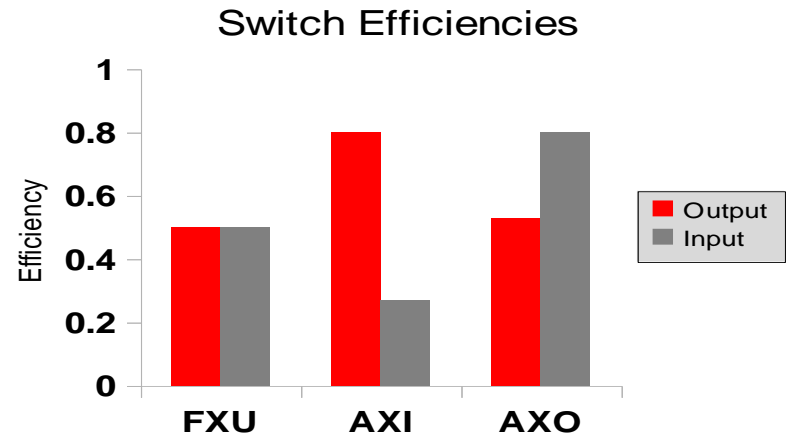
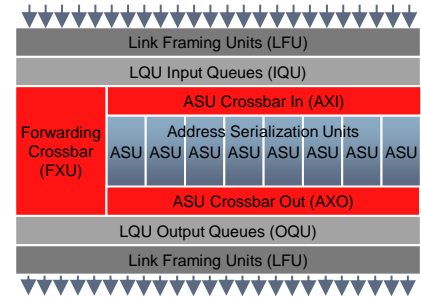
Link Queuing Unit (LQU)



- Each manages an x4 Scalability Link
- Provides queuing support for multiple Virtual Channels
- Each LQU is part of a single physical domain
- SEC-DED on Link FIFOs (RAM based)

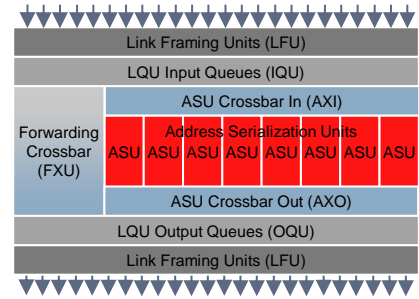
Cross Bar Units (XBU)

- A separate data path forwards traffic and is sized to account for any Head-of-Line blocking inefficiencies
 - FXU: 24in x 24out (2-cycle packet)
 - AXI: 24in x 8out
 - AXO: 16in x 24out
 - Switch fabrics implemented as custom layout hard macros
- Bixby fully sustains mixed request and data traffic at full line rate
- FXU is single domain, AXI/AXO are multi-domain logic
- Flow through SEC-DED, parity on routing control



Address Serialization Unit (ASU)

- Partitioned into eight parallel units
- Each directory unit can compare and process up to 22,656 bits per cycle (total 181,248 bits per chip per cycle)
- 0.5 request lookups per cycle (total 4 per chip)
- Flow-through correction on incoming packets and tag directory contents
- Retry on directory tag staging flops
- Supports up to 12 hardware domains with error steering
- Per domain Built-In Self Initialization (BISI)
- Tag RAM scrubber



Outline

- Motivation and Design Objectives
- M5 System and Beyond
- System RAS Features
- Implementation Details
- Debug and DFT Features
- Summary

Debug and DFT Features

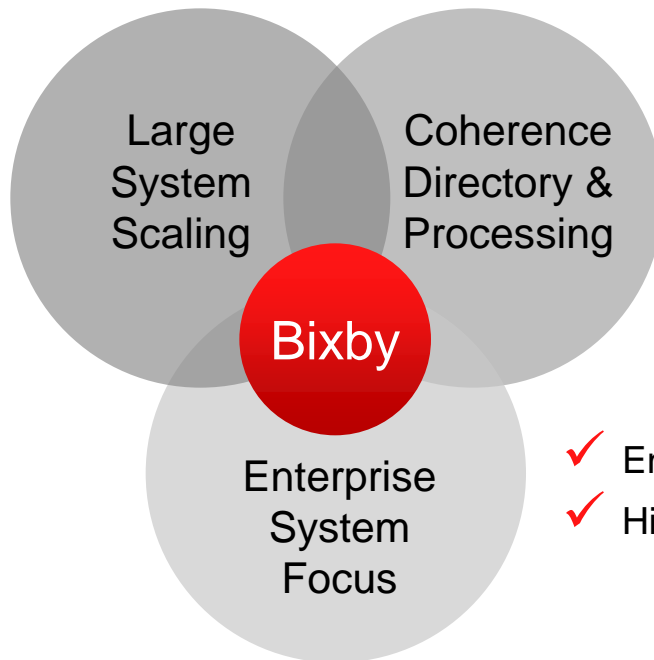
- Monitoring link at full signaling speed is challenging
 - Two internal rings to allow capturing packet flow in ingress or egress direction
 - Internal triggering logic and RAM to store packet flow
 - External DDR interface to allow capturing packet flow on Logic Analyzer
- In-system test features:
 - MemBIST
 - InterconnectBIST
 - ASU tag RAM can be read, written or read-modify-write

Outline

- Motivation and Design Objectives
- M5 System and Beyond
- System RAS Features
- Implementation Details
- Debug and DFT Features
- Summary

Bixby Design Objectives Accomplished

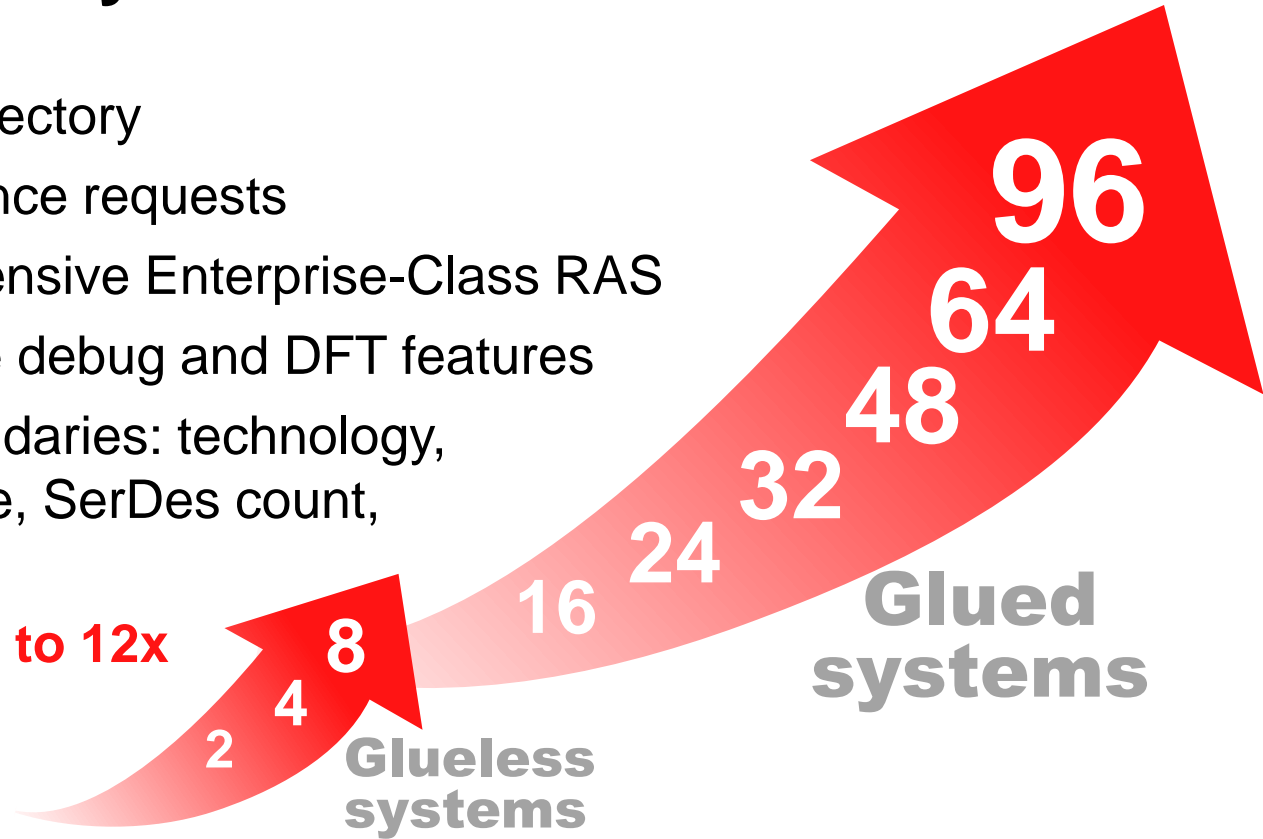
- ✓ Scales to 96 processors
- ✓ Provides communication switching between 8-processor SMPs



- ✓ Directory for L3 caches of all processors
- ✓ Multi-generation support
- ✓ Enabling mixed processor systems
- ✓ Enterprise-Class RAS feature set
- ✓ High bandwidth, low latency

Bixby Scalability ASIC

- ✓ Hosts L3 cache directory
- ✓ Processes coherence requests
- ✓ Includes comprehensive Enterprise-Class RAS
- ✓ Provides extensive debug and DFT features
- ✓ Pushes ASIC boundaries: technology, complexity, die size, SerDes count, power, ...
- ✓ Flexible scaling **up to 12x**



Q&A

References

- White Paper:
 - SPARC M5-32 Server Architecture
<http://www.oracle.com/technetwork/server-storage/sun-sparc-enterprise/documentation/o13-024-m5-32-architecture-1920556.pdf>
- Data Sheets:
 - SPARC M5-32 Server
<http://www.oracle.com/us/products/servers-storage/servers/sparc/oracle-sparc/m5-32/sparc-m5-32-ds-1922642.pdf>
 - SPARC M5 Processor
<http://www.oracle.com/us/products/servers-storage/servers/sparc/oracle-sparc/m5-32/m5-processor-ds-1922646.pdf>

Glossary

- BISI – Built-In Self Initialization
- BIST – Built-In Self Test
- BX – Bixby ASIC
- CL – Coherence Link
- CRC – Cyclic Redundancy Check
- IBIST – Interconnect Built-In Self Test
- MemBIST – Memory Built-In Self Test
- PRBS – Pseudo-Random Binary Sequence
- PVT – Process Voltage Temperature
- RAS – Reliability Availability Serviceability
- SEC-DED – Single-bit Error Correction - Double-bit Error Detection
- SL – Scalability Link
- SMP – Shared Memory Processor
- SPARC - Scalable Processor ARChitecture

Hardware and Software

ORACLE®

Engineered to Work Together

ORACLE®