Barcelona: a Fibre Channel Switch SoC for Enterprise SANs

Nital P. Patwa

Hardware Engineering Manager/Technical Leader
Agenda

- Introduction to Fibre Channel Switching in Enterprise SANs
- Barcelona Switch-On-a-Chip Architecture
- Advanced Features
- Implementation and Lab Testing
- Summary
- Q&A
Fibre Channel Switch

- Storage Area Networks (SAN) based on Fibre Channel protocol revolutionized the way storage was connected to the servers
- FC defined both Arbitrated Loop & Switch topologies
- Modern day SANs are switch based, with arbitrated loop being used in storage arrays
- SAN Address Space: 3-byte address dynamically assigned at login time
Fibre Channel Switch

- Port Types: F(switch), FL(arbitrated loop) or E (ISL)
- 1G/2G/4G  (similar to 10/100/1000),  Also 10G and 8G
- All links have flow control based on buffer to buffer credit scheme
- Each credit allows transmission of a frame up to its max size (~2200 bytes)
- No WRED drop. Only timeouts or errors cause drop
- Amount of buffering required is much smaller than in ethernet switches
- Requires elaborate Congestion Control scheme
Barcelona-based 32-port Switch

Diagram showing the layout of the Barcelona-based 32-port switch with various components such as TCAM, SRAM, FIB, ACL, and ADJ, connected by PCI and DDR buses.
Barcelona Switch-On-A-Chip

- 16 FC Ports and 16 proxy ports to form a 32-port non-blocking switch with 2 Barcelonas
- 320 Buffers – allowing 16 credits for F and 60 for E
- 64 Gb Shared-Memory Switching Fabric
- Cut-through and Store-Forward modes
- Virtual SAN Support
- Zoning ACL Support
- FC-AL intelligent frame ordering
Major Blocks

- Fwd Engine
  - MC/Span Repl
  - Shared Memory
  - Fabric
  - 320 buffers

- Inter-chip Arbitration
- Arbiter
- FL Reorder

- Fabric Ctl
  - Shared Memory Fabric
  - 320 buffers

- FCC
- CPU
- Port
- PCI
- DDR
- DDR0
- DDR15

- SFP
  - srds
  - mac
  - port

- SFP
  - srds
  - mac
  - port

- FC Ports 0..15
Switching across Barcelona

- Requests from Barcelona B0 go to Arbiter in Barcelona B1
- Grants are sent to the fabric in B0 and B1
- Fabric in B0 will schedule the packet to arrive on DDR port
- Fabric in B1 will see the remote descriptor
- Fabric in B1 will schedule a bypass from the DDR port to the FC port
Switching Fabric

- Shared Memory Fabric with TDM access for 33 ports
- Dual-ported header ram doubles the bandwidth for short packets
- Line buffers and Pods simplify physical design

**Diagram:**
- **Header RAM**
  - Dual Port
  - 250MHz
- **Packet RAM**
  - Single Port
  - 250Mhz
- **8 Line Buffers**
- 512 bits
- 64 bits
- **9 Pods**
- 4-ports sharing a pod
Fabric BW Analysis

Per-Port Bandwidth for a 2.125G link

- **bw needed**
- **bw available**

**Gbps**

**Packet Size**

- **bw needed**
- **bw available**

- 36, 40, 52, 56, 116, 120, 244, 248, 504, 508, 1000, 2292
Advanced Features

Director-class features in 1 RU form-factor

- 4K Virtual SANs with inter-VSAN routing support
- Multipath Forwarding and Port Channels
- Spanning with fine-grain filter
- Fibre Channel to MPLS tunneling for remote span
- Fibre Channel congestion detection and avoidance
- Fibre Channel Trace Route
Virtual SANs (VSANs)

- Overlay isolated virtual fabrics on same physical infrastructure
  - Each VSAN contains zones and separate (replicated) fabric services
  - VSAN membership determined by port
- Eliminates costs associated with separate physical fabrics
- VSANs for availability
  - Isolate virtual fabrics from fabric-wide faults/reconfigurations
- Security
  - Complete hardware isolation
- Scalability
  - Replicated fabric services
  - Thousands of VSANs per storage network
Intelligent Network Services—
Multipath Forwarding and PortChannel

- Optimize use of fabric
- Port Channel: Bundle up to 8 links for aggregate of 16Gbps
- Multipath: Utilize up to 16 equal-cost paths
- Hardware-based intelligent load distribution
Spanning – A Diagnostic Tool

• Spanning gives capability to intelligently observe traffic
Span Implementation in Barcelona

- Only Descriptors need to be replicated – packets are already in the shared-memory fabric
- Span filter rules define what to observe
- Span Replication Engine sends replicas to fabric
- Fabric forwards packet to span observation port
- MPLS tunneling allows Remote Span
An inter-switch link (ISL) carrying multiple flows could be under B2B flow control when any one of the flows is congested.

Once ISL is under B2B flow control, all flows operate at the rate determined by the slowest of the flows.

Cisco invented FCC – Fibre Channel Congestion Control.
Congestion Detection and Avoidance in Barcelona

• Congestion monitoring on-chip
• Quench Packet Generation
  – Triggered by receipt of packet on congested Output Queue
  – Triggering packet supplies SA/DA
• Quench Packet sent towards the source
• Source-facing F port on last hop executes Quench and limits packet arrival rate via credit pacing
Implementation

- 13M Gates, 10.5 Mb Memory
- 16 SRDS on-chip
- 16.5x15.7 mm die in 0.13µ Technology
- 250 MHz clock rate
- 1517 pin package - 900 signal pins with interfaces to TCAM, DDR SRAM, PCI Bus, Debug Bus
- Extensive debug features
- Implemented using ASIC flow – no custom circuit in this implementation
- Hierarchical Placement, Flat Routing
Die Photo

- SRDS x8
- FC Port
- Forwarding Engine
- Arbiter
- Shared Memory Fabric
- SRDS x8
Lab Results

- Chip demonstrated switching packets with internal packet generators within first 4 hours!
- Scripts allowed port-login with Agilent SAN Tester without need for driver
- Full 32-port SNAKE demonstrated at line rate within first week
Performance Testing With Snake Configuration

Fiber Connection

Forwarded by switch

RX

TX

SAN Tester

0 1 2 3 4 5 6 7
Summary

- Barcelona Switch-On-a-Chip for Enterprise SAN Applications has been described
- Barcelona-based fabric-class FC Switch offers industry’s most advanced features
- It’s unique architecture has permitted implementation using ASIC flow
- Modular implementation allows easy upgrade path