A 400Gbps Multi-Core Network Processor

James Markevitch, Srinivasa Malladi
Cisco Systems

August 22, 2017
THE INFORMATION HEREIN IS PROVIDED ON AN “AS IS” BASIS, WITHOUT ANY WARRANTIES OR REPRESENTATIONS, EXPRESS, IMPLIED OR STATUTORY, INCLUDING WITHOUT LIMITATION, WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.
Acknowledgements

• The architecture, design, and implementation described in this presentation was done by a diverse team
  • Chip and processor architects, designers, DV engineers, physical designers, software engineers
  • 7 geographic locations
• Credit and thanks to the entire team
Background: Distributed Router Architecture

- Line: interfaces (such as 10 Gigabit Ethernet) used for customer connections
- Fabric: interconnect internal to the router to transfer data between different components of the router, such as network processors
Overview

- 800Gbps (400Gbps full-duplex) network packet processor
- 672 general purpose processors
- > 6.5Tbps serdes I/O bandwidth
- External DRAM for large data structures and packet buffering
- External TCAM for large data structures
- Integrated Ethernet MACs from 10GE to 100GE
- Integrated traffic manager
- Most logic in 1GHz and 760MHz domains
Overall Chip Architecture

- **MACs**: Handles queuing and implements features such as quality of service.
- **Packet Storage**: Entire packets stored on-chip during processing.
- **Processor Array**: Processor array can access and modify entire packet content.
- **Hardware Accelerators**: Networking specific hardware accelerators. Examples: packet ordering, prefix lookup, global ALU operations, counters.
- **Traffic Manager**: External DRAM
- **Fabric Interface**: 400Gbps Fabric Line
- **400Gbps Fabric**: Processor array can access and modify entire packet content.
Processing Flow – 400Gbps Line to Fabric

1. Packets received from line-side serdes

2. Entire packet stored in on-chip memory

3. Packet descriptors sent to processor thread for processing

4. Packet is processed by one of the general purpose threads

4a. External DRAM accessed for large data structures

4b. Hardware accelerators assist certain features

5. After processing, packets are sent to fabric in proper order

MACs — Packet Storage — Processor Array — Fabric Interface

400Gbps Line — 400Gbps Fabric — External DRAM
Processing Flow – 400Gbps Fabric to Line

1. Packets received from fabric serdes

2. Entire packet stored in on-chip memory

3. Packet descriptors sent to processor thread for processing

4. Packet is processed by one of the general purpose threads

4a. External DRAM accessed for large data structures

4b. Hardware accelerators assist certain features

5. After processing, packets are sent to traffic manager in proper order

5a. Packets stored in external DRAM

6. Packets sent to line from traffic manager using configured policies

400Gbps Line

400Gbps Fabric

External DRAM
Processing Model

- 672 processor cores (2688 total threads)
- Run-to-completion model
  - A single thread “owns” a single packet throughout its processing life
  - Different packets may require different features and therefore have different processing times
  - Contrast with: feature-pipelined architecture, systolic arrays, others
- Programmable in C and assembly language
  - Support for traditional stack
Processor Core Architecture

- **Reg File**
- **L0D$**
- **MMU**
- **Request Generator**
- **TLU**

- Separate per-thread L0D$ 2-way set associative
- Software managed MMU
- Round robin thread switching on a cycle by cycle basis
- 8-stage non-stalling pipeline
- Network-specific instructions added to ISA
- Per-thread MMU entries
- Request generator for L0D$ miss, explicit requests
- Separate per-thread L0I$
- 2-way set associative
- Round robin thread switching on a cycle by cycle basis
- 8-stage non-stalling pipeline
- Network-specific instructions added to ISA
- Separate per-thread L0I$

© 2017 Cisco and/or its affiliates. All rights reserved.
Processor Cluster Architecture

- 16 Processor cores
- 4-way set associative L1I$
- L1D$ 4-way set associative
- Messages sent to processor interconnect

16 Processor cores
L1I$
L2I$
Core
Table Lookup Unit accelerates data structure accesses
Interconnect
L1D$
Processor Array Architecture

Cluster 0

Cluster

L2I$

Cluster

L2I$

Cluster

Cluster 41

- 1GHz processor frequency
- Entire processor array “hand”-placed using custom in-house tools

Interconnect

Packet storage

Accelerators

On-chip memories

DRAM controllers

Message-based interconnect between processor cores/caches and accelerators/memories

Bi-partite In-order > 9Tbps bandwidth

4-way set associative L2I$ shared by multiple clusters
Hardware Accelerators (not a complete list)

- **Prefix look-up**
  - Looks up addresses, such as IPv4 and IPv6 (e.g. 192.168.0.123)
  - These address spaces in the Internet are largely unstructured and not hierarchical

- **TCAM, hashing, range compression**
  - Access control lists, quality of service, and other features map to a variety of data structures

- **Statistics counters, rate monitors**
  - Some applications require a large number of counters for network management and customer Service Level Agreements are met

- **Packet ordering**
  - Special hardware to ensure that packets within a flow do not leave the router out of order
Traffic Manager

- Hierarchical queuing structure
  - Flexible levels of queueing hierarchy
  - Minimum rate guarantees
  - Maximum rate limits
  - Weighted sharing
- 256k queues
- Packet data stored in off-chip memory for large buffering

Example of a traffic scheduling hierarchy
Processor Die

- 672 processors (2688 threads)
- 9.2 billion transistors
- 343 megabits of SRAM
- 276 serdes
- 643 mm$^2$ die
- 22nm process
- Hybrid COT design flow
External Memory System Motivation

• Portions of the network require more buffering than can be accommodated on a CMOS logic die

• Portions of the network require larger table sizes than can be accommodated on a CMOS logic die
  • Rough rule of thumb is that each feature requires one or more memory accesses
  • Examples of features: access control, quality of service, link aggregation, statistics for service level agreements
  • Proprietary architecture details are typically used to help reduce access rate (one example could be caching)
Serial-Attached memory

- 12.5Gbps serdes (up to 28 links)
- Proprietary serial protocol optimized for networking
- > 1 billion random accesses per second
- > 300Gbps data transfer rate
  - After removing overhead for commands, addresses, etc.
Serial-Attached Memory

- Multi-chip module
  - 28nm logic die (serdes, protocol controller, BIST)
  - 30nm DRAM die
- Parallel I/O interface between logic die and DRAM die
  - 0.85V @ 1250Mbps
Thank you!

Questions?
CISCO  TOMORROW starts here.