Architecting a multi-core server SoC for the cloud

Barry Wolford
Sr. Director of Technology, Chief SoC Architect
Qualcomm Datacenter Technologies, Inc.

Linley Processor Conference  October 4-5, 2017
Agenda

• Seeding the Cloud...

• Qualcomm Centriq™ 2400 Overview

• Selected Architecture Features

• Summary
The shift to the cloud...

Traditional enterprise
- Monolithic
- Stateful
- OS or VM bound
- Scale up
- Silo’d

Cloud environments
- Microservices
- Mix of stateless / stateful
- Containerized
- Scale out
- Devops
- Multi-tenant

More than 50 percent of servers sold by 2020 will be deployed for cloud computing services.

Source: IDC
...driving new requirements for datacenter infrastructure...

Cloud environments

- Microservices
- Mix of stateless / stateful
- Containerized
- Scale out
- Devops
- Multi-tenant

- Throughput scalability
- Performance at scale
- Power efficiency
- Workload-optimized infrastructure tiers
- Application level redundancy
- Efficient resource pooling
…driving new considerations in processor architecture…

Among those are…

- Aggregate Performance
  - Throughput, concurrency, parallelism
- Thread density
  - VM-hosting, multi-instance/multi-tenant
- Thread isolation
  - Reliable performance, SLAs
- Quality of service
  - SLAs, “noisy neighbors”, tail latencies
- Power efficiency
  - Performance/Watt

Qualcomm Centriq 2400
Purpose-built for the Cloud

- Qualcomm® Falkor™ CPU
  - 5th-generation custom core design
  - ARMv8-compliant
- High core count
  - Up to 48 cores single-thread CPUs
- Highly integrated server SoC
  - Distributed architecture
  - Single chip platform-level solution
  - ARM SBSA Level 3-compliant
- Cloud and throughput-oriented workloads
  - Virtualization and containerization
  - Multi-instance and multi-tenant

Qualcomm Falkor is a product of Qualcomm Datacenter Technologies, Inc.
SoC Overview

- 48 Falkor Cores @ >2 GHz
- 60 MB L3 cache
- Coherent ring interconnect
- 6 Channel DDR4 w/ECC
- 32 PCIe 3.0 lanes
- 8 SATA Gen 3 ports
- 2 GbE ports
- 4 SD/SPI ports
- Integrated management controller
- Robust peripheral mix
  - SPIs, UARTs, I2Cs, USBs, GPIO
- 55mm x 55mm FCLGA (Socketed)
- ARM SBSA Level 3
Foundational Elements

- 24 Falkor Duplex clusters
  - 512 KB Shared L2 cache/cluster
- Distributed L3 cache
  - 12 x 5MB
- Multi-channel DDR
  - 6 x DDR4
- Scalable coherent ring bus
  - Cache and IO coherent
- Integrated expansion/IO
  - 32 PCIe 3.0
  - 8 SATA Gen 3
Falkor Core Duplex

- Two Falkor CPUs
  - ARM v8 compliant
  - AArch 64 Only
  - EL2, EL3, AES, SHA1, SHA2-256
  - Performance/Power Optimized
- Shared 512KB L2 Cache w/ECC (SEC/DED)
  - 128-byte line size, 8-Way
  - 128-byte interleaved
  - Unified between I-side & D-side
  - Inclusive of L1 D-cache
- Shared System Bus Interface
  - Qualcomm® System Bus (QSB)
  - 128-byte interleaved
  - 32-bytes/clk/direction/interleave

Qualcomm System Bus is a product of Qualcomm Technologies, Inc.
LLC & Memory

L3 cache
- Distributed 60 MB L3 w/ECC (SEC/DED)
- 20-way, Non-inclusive/Non-exclusive
- 128B line; 128B Interleave
- Integrated L2 Snoop Filter
- QoS: Way-based partitioning
- Line and way locking support
- Standard cache or victim mode

Memory
- x64 DDR4 memory controller w/ECC (SEC/DED)
- 6 Channels; Interleaved
- Inline bandwidth compression
- Up to quad rank 3DS
- 16-128 GB/Channel - 768 GB maximum
- Up to 2667 MT/s per channel
  - 128 GB/s peak aggregate bandwidth
- x4 devices (and wider)
- RDIMM/LRDIMM
On-Chip Interconnect

- Qualcomm proprietary protocol
- Bi-directional segmented ring bus
- Multi-ring architecture (scale/performance)
  - Fully coherent (cache and IO)
  - Shortest path routing
  - Multicast on read
- 128B interleaved
  - Even and odd interleave ring segments
- >64 GB/s/ring/direction @ >2 GHz
  - >256 GB/s aggregate bandwidth
Distributed LLC & DDR

L3 cache - 12 x 5MB
- Memory address hashed across all L3s
  - Distribute accesses
  - "Smooth" hot/cold sets
- Independent, interleaved ring ports per L3 (24)
- Up to 4 loads + 4 stores concurrently per port

DDR - 6 Controllers
- Full out-of-order execution
- Memory address hashed across all DDRs
- 128B interleaved
- Independent ring port per DDR controller
  - Up to 1 load + 1 store concurrently per port
Distributed PoC & Snoop Filter

Point of Coherency/Serialization
- Orders simultaneous requests to same coherence granule
- Co-located with L3 - 24 instances
  - Up to 24 concurrent operations per clock (SoC)
  - Up to 32 outstanding snoop ops per instance
  - Maximize parallelism/concurrency

Snoop Filter
- \textit{Precise} copy of L2 tag arrays
  - Optimized power and latency vs. \textit{Statistical}
- Co-located with L3 - 24 instances
  - Maximize parallelism/concurrency
  - Enables non-inclusive L2/L3 hierarchy
  - Up to 24 concurrent operations per clock (SoC)
Distributed IOMMUs

IO Memory Management Units
- Address translation and access control
- Shared/Distributed virtual memory support
- Dedicated instances per major IO function
  - 24 instances - PCIe, DMA, SATA, Etc.
  - Eliminates source of resource contention
  - Enables concurrent page table lookup and translation for maximum IO throughput and concurrency
L3 Quality of Service (QoS) Extensions

Shared Resource Contention
- Distributed L3 Cache
- Limited/No Allocation Policy Enforcement

QoS Extensions:
- Hardware Abstracted QoS Domain Identifier
  - Per Client (Core/Virtual Machine, IO/Virtual Function)
- Per-Resource Monitoring and Way-based Allocation
  - Monitor Utilization per QoSID per L3
  - Policy Enforcement per QoSID per L3
    - Instruction/Data Granularity
      - Fine-Tune Cache Allocation per Thread or Class of Threads

Improved cache utilization and per-workload performance (lower application latency) for critical workloads.....
Memory Bandwidth Compression

**Constrained Memory Bandwidth**
- Channel limited peak MT/s
- Limited number of DDR Channels

**Bandwidth Compression:**
- Proprietary algorithm
- Inline compression w/in Memory Controllers
  - Fully transparent to software
- Compress 128B line to 64B when possible
- ECC is encoded with compression bit
- Very low latency decompression
  - 2 – 4 cycles
- Effective on compressible bandwidth intensive workloads

**Increased effective memory bandwidth and reduced power for compressible workloads.....**
Summary

- World’s first 10nm server processor
- Purpose-built for the Cloud
- Architected for scalability and high throughput
- Designed for performance, optimized for power
- Delivers critical server-class features for cloud environments