Key-Value Store Acceleration with OpenPower

Michaela Blott, Principal Engineer
Xilinx Research
Agenda

- Background
- Acceleration of KVS with FPGAs and data flow architectures
- OpenPower
- Conclusions
Background

- Power has become a key limiting factor for modern systems
- Performance scaling problematic on multicores
- Heterogeneous compute is picking up
- New platforms are emerging such as
  - IBM OpenPower

Join the conversation at #OpenPOWERSummit
What FPGAs Can Offer

### COM计算机
- **Image Search**: 8x throughput (4x over GPU)
- **Video Transcode**: 20x throughput
- **Image Processing**: 50x throughput

### 网络
- **Secure socket**: Line rate 100% used, Latency sub 5us, Encryption rate 10x
- **TCP endpoint**: Line rate 100% used, Latency sub 2us, 10x virtual circuits
- **Packet switch**: ASSP line rates, Latency sub 100ns, Protocol choices

### 存储
- **Hybrid memory**: Latency hiding, 10x power saving, +cost+density
- **Key-Value Stores**: 36x RPS/Watt, 10x-100x latency reduction
- **Compression/Encryption**: Customize algorithms, Latency sub 5us, Encryption rate 10x

Join the conversation at #OpenPOWERSummit
Key-Value Stores

- Common middleware application to alleviate access bottlenecks on databases
- Most popular and most recent database contents are cached in main memory of a tier of server platforms
- Used by many well-known websites
- Current server-based implementations are limited, cannot keep up with 10Gbs network speed and won’t scale with more cores
- Investigated using dataflow architectures on FPGAs to dramatically increase performance and lower power and latency

Join the conversation at #OpenPOWERSummit
Standard Implementations

- **Best published numbers for x86 - fully TCP/IP memcached compliant:**

| Intel Xeon (8cores)* | 1.34MRps, 200-300usec, 7KRPS/Watt |

- **Significantly below 10Gbps line rate (13MRPS)**
- **Bottlenecks:**
  - TCP/IP is CPU intensive, interrupt intensive, too large to fit into instruction cache (up to 160 MPKI)
  - Synchronization overhead
    - Threads stall on memory locks, serializing execution for x86s
  - Last level cache ineffective due to random-access nature of the application (miss rate 60% - 95% on x86)
    - Multithreading can’t effectively hide memory access latencies
    - Causes considerable power waste
Dataflow Architecture

Streaming architecture:
Flow-controlled series of processing stages which manipulate and pass through packets and their associated state

FPGA

Request Parser
Hash Table
Value Store
Response Formatter

DRAM Controller

Standardized interface: Key, value, meta-data

DRAM
Hash Table
Value Store
Results

Performance for GET operations as a function of network packet size

Demonstrator:
Up to 36x in performance/power demonstrated
52MSps
Scalability to higher rates possible

Line-rate maximum response rate achieved by Xilinx FPGA accelerator
CAPI provides coherent access to host memory (up to 1 Terabyte) with
- Low latency
- High bandwidth

CAPI simplifies and speeds up host integration
- User programs directly interact with accelerator with one shared virtual memory space
- Saves 800 lines of code/13.5k instructions

Flexible partitioning of functionality
- CAPI enables easy migration of code between host and accelerators
- Easy extension of functionality on host leveraging base platform
Prototype System

Xilinx CAPI Accelerator
ADM_PCIE_7V 3
(V7690T, 16G DRAM)

Power8 Processor
(3GHz, 64G DRAM)

OpenPower Server
(8427-21L)

Join the conversation at #OpenPOWERSummit
Towards 1TB KVS with CAPI

- Request Parser
- Hash Table
- Value Store
- Response Formatter
- UDP/TCP Offload Engines
- Hybrid Memory Controller
- Cache coherent interface to host memory via CAPI (PSL)
- DRAM Controller
- DRAM
  - Hash Table
  - Value Store
- Host Memory
  - Value Store
- 1TB
- 2GB
- 14GB

Network interface (10GE)
Design Environment

- Leverages OpenPower’s accelerator API for host integration in combination with C-programs and HLS for accelerators
- Full development and simulation environment is C-based

*No RTL required!*
Conclusions

- Demonstrated sustained line rate processing for 10GE (13MRPS) from host memory
  - 60Gbps 2TB with 2U OpenPower Server possible

- OpenPower enables
  - Expanding memory capacity to Terabytes
  - Flexible partitioning of functionality
  - Fully C-based design environment
  - Simplifies & speeds-up host integration (14k inst.)

**CHECK OUT OUR DEMO!**