Memory Use Cases in FPGA-enabled Systems

Oct, 2014
Richard Shaw
Sr. Manager, Product Planning
“Stratix V FPGAs… increase ranking throughput in a production search infrastructure by 95% at comparable latency to a software-only solution. The added FPGA compute boards only increased power consumption by 10% and did not exceed our 30% limit in the total cost of ownership of an individual server, yielding a significant overall improvement in system efficiency”

Source: Microsoft paper, *A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services*
- Parallel IO memory: DDR4, DDR3, QDR IV, QDR II+, RLDRAM3
- Serial memory: HMC, MoSys BE
# FPGA Replaces Traditional ASICs and ASSPs

## Device Comparative Snapshot

<table>
<thead>
<tr>
<th></th>
<th>CPLDs</th>
<th>FPGAs</th>
<th>ASICs</th>
<th>ASSPs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Configurability</strong></td>
<td>Via Software and Hardware</td>
<td>Via Software and Hardware</td>
<td>Via Software Only</td>
<td>Via Software Only</td>
</tr>
<tr>
<td><strong>Total Cost of Ownership for System Designs</strong></td>
<td>Low</td>
<td>Low, Volume Dependent</td>
<td>High, Volume Dependent</td>
<td>Contextually Determined: Lowest Cost for Fixed Function Within System</td>
</tr>
<tr>
<td><strong>Time to Market</strong></td>
<td>Fast</td>
<td>Fast</td>
<td>Slowest</td>
<td>Slowest for New ASSP/ Fastest for Established ASSP*</td>
</tr>
<tr>
<td><strong>Design Flexibility that Lowers Risk</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

* If the ASSP is available, you are later than the competition to market despite fastest design time. If it’s not, you have high risk, no flexibility, no differentiation.
Altera’s Target Markets & Industries

Industrial and Automotive
- Automation and Process Control: PLC and I/O Modules, Motion and Motor Control, Industrial Networking, Sensor/Encoder Interfaces
- Building Control and Security: Video Surveillance, Access Control, HVAC Control
- Automotive: Displays, Infotainment, Driver Assistance
- Smart Energy: Smart Grid/Meter, Energy Management, Power Distribution

Military and Aerospace
- Intelligence: Deep Packet Inspection, Data Analysis, High Performance Computing, Acceleration, Access
- EW/Radar: Counter-IED, Jammers, Decoys, Early Warning Radar; Airborne, Ship-Borne and Stationary Radar
- Secure Communications: In-Line Network Encryptors; Airborne, Vehicular, Tower and Tactical Radios
- Guidance & Control: Aircraft, Missile, Vehicle and Robot Guidance and Control, Instrumentation Clusters

Communications
- Networking: Switches, Routers
- Wireline: Optical Metro Access
- Wireless: Remote Radio Head, Basestations, Wireless LAN
- Broadcast: Studio, Satellite, Broadcasting

Computing, Consumer, Storage, Test, and Medical
- Computer and Storage: Servers, RAID, High Performance Computing, Flash Storage, MFP
- Consumer: Displays, Set-Top-Boxes
- Test: IP Video Testers, Protocol Testers
- Medical: CT Equipment, Ultrasound
Application Case 1: Data Center

- **FPGAs used for search acceleration**
- **Two dual-rank DDR3-1600 SO-DIMMs**
  - 8GB @ DDR3-1333 or 4GB single-rank @ DDR3-1600
- **DRAM to store models**
  - Models loaded to FPGA M20K RAM during run-time
  - Model Reload takes up to 250us, much slower than processing
- **Increased memory bandwidth needed**
  - Insufficient physical space to add additional DRAM channels
- **Food for thought**
  - Could HMC or 2.5D DRAM be better solution in the future?

Source: Microsoft paper, *A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services*
### Application Case 2: Memory Intensive Networking

- **Backplane Switch (FIC)**
- **Traffic Manager (TM)**
- **Packet Processing (PP)**
- **Front End Optics (& Processing)**

#### PP Function Memories Used

<table>
<thead>
<tr>
<th>PP Function</th>
<th>Memories Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parsing</td>
<td>M20K*</td>
</tr>
<tr>
<td>Packet Store</td>
<td>M20K, DDR</td>
</tr>
<tr>
<td>Classification</td>
<td>TCAM</td>
</tr>
<tr>
<td>Packet Editing</td>
<td>M20K, QDR, RLD</td>
</tr>
<tr>
<td>Statistics</td>
<td>M20K, DDR, RLD</td>
</tr>
<tr>
<td>Policing</td>
<td>M20K, QDR, RLD</td>
</tr>
<tr>
<td>Forwarding</td>
<td>DDR</td>
</tr>
</tbody>
</table>

#### TM Function Memories Used

<table>
<thead>
<tr>
<th>TM Function</th>
<th>Memories Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free List</td>
<td>M20K, QDR, RLD</td>
</tr>
<tr>
<td>Linked List</td>
<td>M20K, QDR, RLD</td>
</tr>
<tr>
<td>Queue &amp; Buffer Management</td>
<td>QDR, DDR, RLD</td>
</tr>
<tr>
<td>nQ, dQ (head,tail ptrs)</td>
<td>QDR, RLD</td>
</tr>
<tr>
<td>Congestion Mgt.</td>
<td>QDR, RLD</td>
</tr>
<tr>
<td>Scheduler</td>
<td>QDR, RLD</td>
</tr>
</tbody>
</table>

* M20K: Distributed embedded SRAM in Altera FPGA
Packet Buffering DRAM Requirement

FPGA not enough IO for 200+G system
HMC & HBM meets BW requirement

Note: * Assume 70% DRAM controller efficiency
Existing control plane memory is port and IO constrained for forward looking applications.
FPGA IO & packaging solution will be challenged to meet system-level power & performance requirements. Inflection point at 200G.
Application Case 3: Flash Storage

- FPGA used as bridge between flash memory and CPU
- FPGA is a DDR3 slave to the x86 CPU
- FPGA also implements flash controller
Breakthrough Advantage with Generation 10

- TSMC 20 nm process
- 15% higher performance than current high-end with 40% lower midrange power
- 5x higher customer commitment dollar value at time of launch
- Dual-core 32-bit ARM Cortex-A9 processor

- Intel 14 nm Tri-Gate process
- 2x performance increase
- 70% power savings
- Quad-core 64-bit ARM Cortex-A53 processor
- 3D-capable for integrating SRAM, DRAM, ASIC

Reinventing the Midrange

Delivering Unimaginable Performance
Summary

- FPGA’s flexibility and versatility enable wide usage in different industries
- Altera FPGAs have broad memory technology support
  - DDR4, DDR3, QDR IV, RLDRAM 3
  - Flash, MRAM
  - HMC, MoSys BE
  - 2.5D / 3D memory
- Altera well positioned to support target markets and applications
Thank You