Architecting High-Performance SoC Video Processors

Sorin C. Cismas

http://www.quarc.com
Outline

• Digital Video and the SoC Challenges
• Architecture, Design Methodology and Tools
• Videris™ HD - MPEG2 4:2:2@HL Video Decoder
  – Multi-Threaded MPEG Decoder
  – Fused Multiply/Add/Subtract DCT
  – Tile based Super-Scalar Memory Controller
  – Videris™ HD Statistics
• Conclusions
Outline

• Digital Video and the SoC Challenges
  • Architecture, Design Methodology and Tools
  • \textit{Videris}™ HD - MPEG2 4:2:2@HL Video Decoder
    – Multi-Threaded MPEG Decoder
    – Fused Multiply/Add/Subtract DCT
    – Tile based Super-Scalar Memory Controller
    – \textit{Videris}™ HD Statistics
  • Conclusions
More Performance @ Low Power

• Higher Resolution, Higher Bitrate - **Video quality**
  – Video CD  360 x 240 @  1-3 Mbit/sec.
  – DVD       720 x 480 @  4-10 Mbit/sec.
  – HDTV      1920 x 1080 @ 20-40 Mbit/sec.
  – Digital Cinema  4096 x 3072 @  ??? Mbit/sec.

• Multiple Streams, Multiple Standards - **Flexibility**
  – MPEG2 (DirecTV, DVD, DVB, ATSC, ISDB)
  – MPEG4, J2K, MJ2K
  – JPEG, DV

• Wireless and Portable - **Low power**
  – Limited and variable bandwidth
  – Scalable performance
SoC (Systems not Chips)

- System = Hardware + Software + Application
- Hardware/Software partitioning is crucial
- The SoC Challenges
  - reuse and easy integration
  - faster time-to-market and smaller circuit geometry
  - high performance and low power
  - all of the above @ low cost
- The Solution - Innovative Architectures, Design Methodologies and Tools
  - they will drive the SoC revolution
  - handcrafting to squeeze the last picoseconds and the last thousands gates will be a thing of the past
Outline

• Digital Video and the SoC Challenges
• Architecture, Design Methodology and Tools
  • Videris™ HD - MPEG2 4:2:2@HL Video Decoder
    – Multi-Threaded MPEG Decoder
    – Fused Multiply/Add/Subtract DCT
    – Tile based Super-Scalar Memory Controller
    – Videris™ HD Statistics
• Conclusions
QuArc Architecture and Methodology

• “divide et impera”
  – Functional partitioning in manageable objects
  – Well defined interfaces
  – Independently testable objects, easy to integrate, and reuse
  – Automate most of the design, verification, and synthesis process
  – Enable engineers to work on the creative and fun stuff

• Key Features
  – Encapsulates algorithms in self-contained data-driven objects
  – No need for master controller or scheduler
  – Synchronous but self-timed (variable schedule, elastic pipelines)
  – Adapts to instantaneous variations in processing load
  – Split memory transactions
  – Stall tolerant - works well in systems with shared memory
  – Minimal interfaces to simplify the wiring complexity
QuArc Objects

• Any design is a collection of Objects
  – Atoms: leaf objects (indivisible)
    • two global signals: clock and reset
    • one or more Input Interfaces
    • one or more Output Interfaces
  – Molecules: collection of Atoms and/or Molecules
  – Interfaces:
QuArc Interfaces

- Minimal set of signals
- Synchronous and uni-directional
- One transmitter and one or more receivers
- Sustained one token/cycle throughput

Token bus:
- At the physical level, like any other data bus
- At the logical level, equivalent to a C-language data structure
- Can be a collection of sub-busses, each with its own syntax

Handshake signals:
- Simple rdy/req handshake protocol
- One rdy/req pair for each receiver
- Data is handed over when both rdy and req are asserted
- Transmitter and Receivers can stall the transaction in any cycle
QuArc Pipestages (Qpipes)

- Contains one or more registers, sometimes a memory
- Has it’s own controller that keeps track of how many tokens are in the pipeline
- Atoms knows when valid data is in the pipeline
- Atoms can independently shut down the clock to save power
- Library of Qpipes
  - Hides variable schedule complexity
  - Simplifies design task
  - Designers can focus on algorithms, not on low level control
QuArc Design Language

• Every Object has a .qdl file (Object Spec.)
  – What the Object is
  – How to use other Objects to build a system

• QDL files have four parts
  – Parameters - customizes Atoms and Interfaces based on the system requirements
  – Interfaces - describes their properties
    • Input/Output
    • Interface Type (Class)
      – the syntax is described in a QDL library
      – only Interfaces of the same type can be connected together
    • Prefix - to uniquely identify an Interface if an Object has more than one of the same Type
  – Instantiations (for Molecules only)
  – Register Description (for Atoms only)
Automatic Configuration Tool

Verilog

QDL

C Model

Automatic Configuration Tool (ACT)

Verilog Objects

Netlist

Chip Layout

Verilog Wrapper

Synthesis Driver

Test Bench

C Model Wrapper

Simulation (RTL or Gate Level)
Design Style Rules for Easy Design Reuse

- Positive-edge triggered flip-flops, no latches
- Single clock
- Reset can be synchronous or asynchronous
- Control registers are always reset
- Low input set-up time (<25% of the cycle time)
- Low output delay time (<25% of the cycle time)
- Output Data comes directly from registers
- Input Data goes directly to registers
- No combinational paths from inputs to outputs
- Simple internal RAM (1-port or 1 read/1 write port)
Outline

• Digital Video and the SoC Challenges
• Architecture, Design Methodology and Tools
• Videris™ HD - MPEG2 4:2:2@HL Video Decoder
  – Multi-Threaded MPEG Decoder
  – Fused Multiply/Add/Subtract DCT
  – Tile based Super-Scalar Memory Controller
  – Videris™ HD Statistics
• Conclusions
MPEG2 4:2:2@HL Video Decoder

- 300 Mbits/sec (30 x DVD)
- All non-scalable profiles at all levels
- Dedicated, hard-wired units to guarantee high-performance at low power and low cost
- High level decisions and error recovery is under software control (few MIPS and not real-time critical)
Multithreaded MPEG Decoder

• Video processors will evolve from dedicated, one-application-at-a-time to multithreading
  – In the ‘90ties, single stream decoders
  – Context switching and multithreading is a must in future visual communication and entertainment devices

• MPEG multithreading
  – Most HD decoders have the processing power to decode several SD bitstreams. After a picture is completely decoded, a new picture from a different bitstream can be decoded, but context switching takes far too many cycles and reduces performance
  – Videris™ HD can process up to 16 bitstreams simultaneously, each with its own different bitrate, resolution and frame rate, without any penalty in stall cycles
  – This is an important feature when many and relatively small MPEG textures need to be mapped on objects, as in games, multimedia and visual communications
Video Context Switching

• Context stored in special QuArc Pipestages distributed over the whole design
  – too much overhead to store/retrieve it to/from memory
  – very long pipeline compared to general purpose processors
  – context switch happens at different times in different objects
  – at any time, Videris™ HD can be processing several contexts

• Context switching is supported in QDL

• ACT configures the QuArc Pipestages for the requested number of contexts

• Independent Bit Buffer Control and Decode/Display Interlock for each bitstream
Multiply/Add/Subtract IDCT Algorithm

- Based on a DCT Algorithm by Elliot Linzer and Ephraim Feig (26 Multiply/Add)

- Our Algorithm (16 Multiply/Add/Subtract)

```
\[ x_0 \rightarrow z_0 \rightarrow y_0 \rightarrow y_0 \rightarrow x_0 \]
\[ x_0 \rightarrow z_1 \rightarrow y_1 \rightarrow y_1 \rightarrow x_1 \]
\[ x_0 \rightarrow z_2 \rightarrow y_2 \rightarrow y_2 \rightarrow x_3 \]
\[ x_0 \rightarrow z_3 \rightarrow y_3 \rightarrow y_3 \rightarrow x_2 \]
\[ x_0 \rightarrow z_4 \rightarrow y_4 \rightarrow y_4 \rightarrow x_7 \]
\[ x_0 \rightarrow z_5 \rightarrow y_5 \rightarrow y_5 \rightarrow x_4 \]
\[ x_0 \rightarrow z_6 \rightarrow y_6' \rightarrow y_6 \rightarrow x_6 \]
\[ x_0 \rightarrow z_7 \rightarrow y_7' \rightarrow y_7 \rightarrow x_5 \]
```

Legend:
- __________ +
- __________ -
- \( k_1 \)
- \( + k_1 \times \)
- \( k_2 \)
- \( - k_2 \times \)
MAS Operation:
- \( \text{tmp} = k \times b \)
- \( s_0 = a + \text{tmp} \)
- \( s_1 = a - \text{tmp} \)
Multiply/Add/Subtract IDCT Features

- Minimal hardware: 2 MAS, 9 registers and 5 muxes
- Throughput: 1 DCT coefficient/cycle - no stall cycles (94MHz for HD)
- Latency: 12 cycles - 2 or 3 IDCTs are processed at any time
- Narrow busses: 16 bits on the interfaces, 18 bits internally
- HD requires two IDCTs plus a 64 x 16 transpose memory
- Exceeds IEEE 1180 requirements
- Extra bit in data path to guard against large quantization noise

<table>
<thead>
<tr>
<th>Precision</th>
<th>IEEE 1180</th>
<th>QuArc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak Error</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pixel Mean Error</td>
<td>0.015</td>
<td>0.0043</td>
</tr>
<tr>
<td>Overall Mean Error</td>
<td>0.0015</td>
<td>0.0001</td>
</tr>
<tr>
<td>Pixel Mean Square Error</td>
<td>0.06</td>
<td>0.0226</td>
</tr>
<tr>
<td>Overall Mean Square Error</td>
<td>0.02</td>
<td>0.0185</td>
</tr>
</tbody>
</table>
Memory Bandwidth

- Biggest limitation for high performance (microprocessors, 3D and video)
- Higher memory density makes them cheaper and deeper, but not faster
- Shared memory is a must for SoC
- MMU becomes the bottleneck that distributes bandwidth to many clients
- Caches do not necessarily help for video (random accesses for small blocks of data)
- Tile based memory organization, multiple banks and memory transactions reordering do help
**Microprocessor vs. SDRAM Pipeline**

- **Microprocessor Pipeline**
  - Branches
  - Load delay

- **SDRAM Pipeline**
  - Variable length
  - Shared command bus
  - Write to Read penalty = CAS Latency

### Microprocessor Pipeline

<table>
<thead>
<tr>
<th>Stage</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF</td>
<td>DEC</td>
</tr>
<tr>
<td>EX</td>
<td>MEM</td>
</tr>
<tr>
<td>WB</td>
<td></td>
</tr>
</tbody>
</table>

### SDRAM Pipeline

- Variable length
- Shared command bus
- Write to Read penalty = CAS Latency

**Diagram: Microprocessor vs. SDRAM Pipeline**

**Hot Chips 2000**
8/14/00
Super-Scalar Memory Controller

- Complex Memory Transactions (array with implicit or variable stride) are translated to Simple Slice Instructions (access to consecutive addresses within the same bank and page)

Diagram:
- Input QPipe
- Memory Transaction Slicer
  - Slice 0
  - Slice 1
  - Slice 2
  - Slice 3
  - Slice 4
  - Slice 5
- SDRAM Command Scheduler
  - Bank 0
  - Bank 1
  - Bank 2
  - Bank 3
  - Refresh
- SDRAM Command Arbiter
- Output QPipe

Complex Transactions
Simple Slice Instructions
6-wide Instruction Window
Out of Order Commands
(4+1) wide super-scalar
In Order Data
SDRAM Commands
Pipeline Operation

Memory Transactions:
- \( T_0 \) - read 3 x 4 words (array)
- \( T_1 \) - read 2 x 1 words (linear)
- \( T_2 \) - read 2 x 1 words (linear)
- \( T_3 \) - read 1 x 1 words (linear)

- Out of Order Commands: \( T_2 \) (Slice 5) is executed before \( T_1 \) (Slice 4)
- Stall cycle can be eliminated with Out of Order Data
  (requires on-chip memory to reorder the data for the clients)
Memory Controller Performance

• 90 - 94% memory utilization
• Needs memory transaction size longer than $t_{RC}$ (Active to Active command period)
• MPEG prediction size is 3 x 9, 3 x 8, 3 x 5, and 3 x 4 words on a 64-bit bus
• This high efficiency enables Videris HD to use a 32-bit SDRAM or 16-bit DDR at 133MHz for HDTV
• 93 - 96% memory utilization with Out of Order Data
  – on-chip memories are needed for both read and write to reorder the data for the clients)
  – longer latencies
### Videris™ HD Statistics

<table>
<thead>
<tr>
<th>Videris™ HD Object</th>
<th>Base Area</th>
<th>Inc. Area</th>
<th>Base RAM</th>
<th>Inc. RAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video Parser</td>
<td>0.137 mm²</td>
<td>0.004 mm²</td>
<td>128 bytes</td>
<td>128 bytes</td>
</tr>
<tr>
<td>Inverse Quantizer</td>
<td>0.073 mm²</td>
<td>0.005 mm²</td>
<td>96 bytes</td>
<td></td>
</tr>
<tr>
<td>IDCT</td>
<td>0.195 mm²</td>
<td></td>
<td>128 bytes</td>
<td></td>
</tr>
<tr>
<td>Motion Vectors</td>
<td>0.087 mm²</td>
<td>0.005 mm²</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motion Compensation</td>
<td>0.148 mm²</td>
<td></td>
<td>432 bytes</td>
<td></td>
</tr>
<tr>
<td>Bit Buffer Control</td>
<td>0.154 mm²</td>
<td>0.019 mm²</td>
<td>512 bytes</td>
<td></td>
</tr>
<tr>
<td>Decode/Display Interlock</td>
<td>0.045 mm²</td>
<td>0.003 mm²</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Videris™ HD Total</td>
<td>0.839 mm²</td>
<td>0.035 mm²</td>
<td>1,296 bytes</td>
<td>128 bytes</td>
</tr>
<tr>
<td>Memory Management Unit</td>
<td>0.340 mm²</td>
<td>0.030 mm²</td>
<td>1,280 bytes</td>
<td></td>
</tr>
<tr>
<td>SDRAM Controller</td>
<td>0.193 mm²</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Videris™ HD + Mem. Cntr.</strong></td>
<td>1.372 mm²</td>
<td>0.065 mm²</td>
<td>2,576 bytes</td>
<td>128 bytes</td>
</tr>
</tbody>
</table>

- Logic Area based on TSMC 0.18um, 150MHz, worst-case operating conditions (V, T, P)
- up to 200MHz worst-case operating conditions (logic area increases by approx. 20%)
- Base Area (routing overhead not included) and Base RAM are for one context
- For every additional context, add Inc. Area and Inc. RAM
- Total size, including routing overhead for one HD + two SD (or eight SD) is 3-4 mm²
Outline

• Digital Video and the SoC Challenges
• Architecture, Design Methodology and Tools
• Videris™ HD - MPEG2 4:2:2@HL Video Decoder
  – Multi-Threaded MPEG Decoder
  – Fused Multiply/Add/Subtract DCT
  – Tile based Super-Scalar Memory Controller
  – Videris™ HD Statistics

• Conclusions
Conclusions

• Pure Hardware or pure Software are not the right answer for the future visual communication and entertainment devices

• SoC is a big challenge for all semiconductor companies

• Moore’s Law can not be sustained just by the future progress in semiconductor processes

• Architecture, Design Methodology and Tools will drive the semiconductor industry