Details of a New Cortex Processor Revealed

Cortex-A9

ARM Developers’ Conference
October 2007
ARM Pioneering Advanced MP Processors

- August 2003  ARM shows first synthesisable MP processor
- May 2004    ARM11 MPCore™ quad-core launched
- August 2004  AMD launches x86 dual-core processor
- April 2005   Intel launches x86 dual-core processor
- June 2007   More than ten ARM11 MPCore licences
               Five in the top 11 semiconductor companies
               Four public – NEC, NVIDIA, Renesas, Sarnoff Corp.

October 2007  Announcing the next generation Cortex-A9
Multi-function and Convergence

In the Home
- ARM® Cortex™ CPU
- Smart Cards for protected content
- Location free TV
- Wireless LCD TV
- Gaming displays
- In-home security
- HD STB
- Bluetooth
- Home automation
- Memory cards >10 GB
- Terabyte storage for PVR

On the Go
- Mali™ Graphics processor
- GPS
- Stereo Headset
- Bluetooth/UWB
- NFC / RFID
- TV out
- PC / Mac
- WiMax (10 Mbit down)
- 3.5G (HSDPA)
- Biometrics
- MEMS
- 20 GB HD
- 512 MB Memory card
CMC Requirements Example

Multicore processing technology extends what is possible for next-generation Connected Mobile Computers (CMC)

**Better Processor and System Performance**
- Rich Operating Systems
- Advanced browsers
- Advanced security
- General applications
- 3D graphics & gaming
- Java & Execution Environments
- High bandwidth networks
- Multi-format audio
- Video Recorder / Player
- Voice/Video over IP

...and these pictures are from today’s ARM technology
## Market-Driven Requirements

<table>
<thead>
<tr>
<th>Next-Generation Devices</th>
<th>Performance Drivers</th>
</tr>
</thead>
</table>
| Mobile Handsets         |  ➢ Large-screen mobile compute devices can leverage large performance increases to meet computing needs  
| Connected Mobile Computers | ➢ Main-stream smart handsets require increased performance but within existing low-power budgets  
|                         | ➢ Low-cost devices require reduced cost and longer battery life at current performance levels |
| Consumer and Auto-infotainment | ➢ Higher quality media is driving higher bandwidth connections and increased system complexity  
|                         | ➢ Device consolidation driving function integration requiring higher performance and compatibility |
| Networking/Home Gateways | ➢ Scaling enterprise performance to meet the needs of latest high-speed networks  
|                         | ➢ Consumer devices adding significant new features while routing higher bandwidth needing scalable performance |
| Embedded                | ➢ Higher image quality, more mega-pixels,  
|                         | ➢ More advanced image processing driving performance |

### Common Requirements

- Increased power efficiency with higher performance
- Increased top-end performance for demanding applications
- Ability to share software investments across multiple devices
The ARM Cortex-A9 Processors

- ARM has developed a completely new processor design and is releasing it as two product variants within the common ARM® Cortex™ application and OS architecture.

- **ARM Cortex-A9 MPCore™ processor**: A *multicore processor* delivering the next generation of the ARM MPCore technology for increased performance scalability and increased control over power consumption. Ideal for high-performance mobile handsets, networking and auto-infotainment devices.

- **ARM Cortex-A9 processor**: A traditional *single core processor* for simplified design migration in high-performance, cost-sensitive markets such as mobile handsets and other embedded devices reducing time-to-market and fully maintaining existing software investments.
Cortex-A9 Processor Highlights

✓ First synthesizable ARM processor able to deliver more than 8000 aggregate DMIPS for demanding embedded applications

✓ First ARM processor to combine Cortex-A architecture with MP capabilities for increased performance within the tight power limitations of high-performance mobile devices

- Improved second-generation superscalar pipeline design
  - Superscalar, dynamic multi-issue technology, on an efficient 8-stage pipeline
    - Early resolution of branches evaluated asynchronously to instruction fetch
    - Continuous fetch & decoding of 2 instructions per clock cycle
    - Common instructions take 9 cycles, while complex instructions take up to 11 cycles
  - Out-of-order (OoO) instruction dispatch and completion
    - Leverages OoO without traditional power/resource hungry reorder buffers
    - Supports dispatch of 4 instructions and completion of 7 instruction per clock cycle
    - Provides optimal performance from binaries compiled for previous ARM processors
  - Multiple outstanding memory transactions with out of order completion
    - Low load-use penalty of a single clock cycle optimizes benefit from OoO dispatch
  - Optimized for high performance and low cost
Cortex A9 Microarchitecture (single core variant)
Offering Design Scalability

- Delivering over 8000 peak aggregate DMIPS through more than 2.0 DMIPS/MHz
- Higher performance and lower power consumption for mobile devices
- Providing software portability across multiple devices
- Synthesis flexibility to target from 500MHz to 1GHz in TSMC 65nm processes

A single processor that provides a single product that can be applied across multiple markets, sharing a common software platform.
Cortex-A9 Companion Products

- **Cortex-A9 Floating-Point Unit (FPU):**
  - Delivering high performance single and double precision floating-point unit for accelerated 2D/3D, imaging and scientific computation

- **Cortex-A9 NEON Processing Engine (NPE):**
  - Enhancing the capabilities of the FPU to include the ARM NEON™ Advanced SIMD (Single Instruction, Multiple Data) support for accelerated media and signal processing computation

- **Advanced L2 Cache Controller:**
  - The PL310 L2 controller was developed alongside the Cortex-A9 processor to ensure high data throughput and maximum performance

- **Cortex-A9 Program Trace Macrocell (PTM):**
  - Providing ARM CoreSight™ compatible program trace capabilities
  - An associated CoreSight Design Kit is also available to enable advanced multicore trace with source correlation and time stamping
Cortex-A9 Floating-Point Unit

- Cortex-A9 FPU implements the Cortex VFPv3-D16 Floating-Point architecture
  - Targeting high performance from compiled code
    - Existing hand-optimized vector code supported by software exception
    - Software simplification from removal of trapped exceptions
  - Register bank optimized to 16 double-precision registers for size and power considerations and compatibility with VFPv2 and ARM11 software

- Single precision and double precision format are supported.
  - Emphasis was first put on fast single precision support.
  - Higher performance single/double precision over to VFP11 and Cortex-A8
  - New instructions for FP16 data type conversion as used in 3D graphics

- Flexible modes of operations
  - Flush-to-Zero, Default NaN and full compliance with IEEE-754 standard
  - All rounding modes are supported.
  - Full hardware support of denormal numbers (no need for support code).
Cortex-A9 NEON™ Media Processing Engine

- Extends Cortex-A9 single and multicore processors
  - ARM NEON advanced SIMD instruction support
    - 64-bit and 128-bit registers supporting SIMD operations on:
      - 8, 16 and 32-bit integer data formats
      - 32-bit floating-point data
  - Enlarges register file available to FPU
    - Increases design to support 32 double-precision registers
  - Support fused data types to remove packing/unpacking overheads
  - Structured load/store capabilities eliminate shuffling of data between algorithm-format to machine-formats

- Whilst retaining:
  - Existing Cortex-A9 FPU 32/64-bit scalar floating-point performance
  - Existing Cortex-A9 core integer performance
PL310 L2 Cache Controller

- Capable of supporting over double the sustained throughput of today's high performance ARM L220 AXI cache controller
  - Supporting multiple outstanding AXI transactions on each interface
  - Support for parity and ECC RAMS
  - Support for per-master way lockdown when shared between multiple CPU or falcon coherent accelerators

- Advanced SoC bus interface
  - Support for synchronous ½ clock ratios to reduce latencies on high speed processor designs
  - Ability to address-filter second master AXI interface for split-domain designs and fast access to on-chip scratch memories

- Supporting Cortex-A9 specific optimizations
  - Auto-preload on instruction miss
  - Operate an “exclusive” allocation policy with processor L1 for increased cache utilization
## Cortex-A9 – Market-Driven Solutions

<table>
<thead>
<tr>
<th>Next-Generation Devices</th>
<th>Cortex-A9 Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mobile Handsets</strong></td>
<td><strong>Next-generation high-end devices (1500-3000DMIPS)</strong></td>
</tr>
<tr>
<td><strong>Connected Mobile Computers</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2-3 core processor with IEM™ technology and adaptive shutdown</td>
</tr>
<tr>
<td></td>
<td>32K Instruction and Data caches, 256-512K shared L2 cache using PL310, partitioned AXI</td>
</tr>
<tr>
<td></td>
<td>NEON technology-based Media Processing Engine, coherent GPU</td>
</tr>
<tr>
<td><strong>Consumer and Auto-infotainment</strong></td>
<td><strong>Mid-range, cost reduction, (900-1500DMIPS)</strong></td>
</tr>
<tr>
<td></td>
<td>Single core processor with NEON or FPU</td>
</tr>
<tr>
<td></td>
<td>16K or 32K instruction and data caches</td>
</tr>
<tr>
<td></td>
<td>128-256K L2 cache using PL310, single AMBA AXI bus</td>
</tr>
<tr>
<td><strong>Networking / Home Gateways</strong></td>
<td><strong>Feature-rich mass market (600-900DMIPS)</strong></td>
</tr>
<tr>
<td></td>
<td>Single core processor with FPU</td>
</tr>
<tr>
<td></td>
<td>16K instruction and data caches, single AXI</td>
</tr>
<tr>
<td><strong>Embedded</strong></td>
<td><strong>Embedded media and imaging (800-2000DMIPS)</strong></td>
</tr>
<tr>
<td></td>
<td>2x multicore utilizing coherent accelerators</td>
</tr>
<tr>
<td></td>
<td>32+32K instruction and data with 256K shared L2 cache</td>
</tr>
<tr>
<td></td>
<td>FPU for postscript and image manipulation and enhancement</td>
</tr>
<tr>
<td></td>
<td>Code migration through selective AMP/SMP deployments</td>
</tr>
<tr>
<td><strong>Enterprise market (4000-8000DMIPS)</strong></td>
<td><strong>Consumer devices (800-1500DMIPS)</strong></td>
</tr>
<tr>
<td></td>
<td>3-4 core performance optimized implementation</td>
</tr>
<tr>
<td></td>
<td>32K+64K instruction and data cache</td>
</tr>
<tr>
<td></td>
<td>512K-2MB L2 cache, dual 64 bit AMBA AXI interfaces</td>
</tr>
<tr>
<td><strong>Connected Mobile Computers</strong></td>
<td><strong>Networking / Home Gateways</strong></td>
</tr>
<tr>
<td></td>
<td>1x or 2x multicore utilizing coherent accelerators</td>
</tr>
<tr>
<td></td>
<td>32+32K instruction and data, with 256-512K shared L2 cache</td>
</tr>
<tr>
<td></td>
<td>NEON or VFP when offering media gateway or services</td>
</tr>
<tr>
<td><strong>Mid-range, cost reduction, (900-1500DMIPS)</strong></td>
<td><strong>Feature-rich mass market (600-900DMIPS)</strong></td>
</tr>
<tr>
<td></td>
<td>Single core processor with NEON or FPU</td>
</tr>
<tr>
<td></td>
<td>16K or 32K instruction and data caches</td>
</tr>
<tr>
<td></td>
<td>128-256K L2 cache using PL310, single AMBA AXI bus</td>
</tr>
<tr>
<td><strong>Networking / Home Gateways</strong></td>
<td><strong>Embedded</strong></td>
</tr>
<tr>
<td></td>
<td>2x multicore utilizing coherent accelerators</td>
</tr>
<tr>
<td></td>
<td>32+32K instruction and data with 256K shared L2 cache</td>
</tr>
<tr>
<td></td>
<td>FPU for postscript and image manipulation and enhancement</td>
</tr>
<tr>
<td></td>
<td>Code migration through selective AMP/SMP deployments</td>
</tr>
<tr>
<td><strong>Consumer and Auto-infotainment</strong></td>
<td><strong>Networking / Home Gateways</strong></td>
</tr>
<tr>
<td></td>
<td>1-4 core processors giving design scalability across family of devices</td>
</tr>
<tr>
<td></td>
<td>32K instruction and data caches with 0-512K L2 cache</td>
</tr>
<tr>
<td></td>
<td>NEON technology for advanced media and DSP processing</td>
</tr>
<tr>
<td></td>
<td>Advanced bus interface unit for high-speed memory transfers between on-chip 3D engines and network interface MACs</td>
</tr>
<tr>
<td></td>
<td>AMP configurations using separate CPU for real-time RTOS</td>
</tr>
</tbody>
</table>
Enhanced ARM MPCore Technology

- ARM11™ MPCore processor support maturing since 2005
  - Devices completing device integration before mass market availability
  - Growing support of operating systems and software middleware

- Summary of ARM’s proven MPCore technology
  - Pre-integrated and validated scalable multicore processing
  - Integrated interrupt distribution and inter-processor communications
  - Advanced Snoop Control Unit supporting enhanced cache coherence

- Cortex-A9 MPCore further enhances MPCore technology
  - Accelerator Coherence Port (ACP)
  - Advanced Bus Interface Unit
  - Multicore ARM TrustZone® technology with interrupt virtualization
  - Generalized Interrupt Controller (GIC)
Cache Coherence: Embedded Challenge

**Snooping Techniques**
- Each cache checks the bus for accesses to memory locations that it already has cached
- Fails to scale due to every request going to every client, reducing processing efficiency
- Consumes unproductive system power since snoop may be unnecessary

**Directory**
- Each cache operation needs to be exported to a central directory to hold a global state
- More scalable since only caches holding shared data are ‘snoop’ed, but slower due to higher latencies, also higher power solution

Both techniques act a communication mechanism for a coherence protocol. E.g.

- **M.** Cache line has been modified
- **E.** Cache line is held exclusively
- **S.** Cache line is shared with another CPU
- **I.** Cache line is invalidated
Optimized MESI Snooping

**Increased power efficiency and increased scalability by avoiding system accesses:**

- **Duplicated Tag RAMs acting as a local Directory**
  - Stored in Snoop Control Unit of the multicore processor
    - Checks if data is in cache without interrupting CPU
    - Filters access to only CPU that are sharing data
    - Keeps power lower since directory is local
  - Allows independent tasks to run at full single thread performance resulting in linear scalability

- **Direct Data Intervention (cache-2-cache transfer)**
  - Copy **clean** data from one CPUs cache to another
  - Removing need for main memory accesses reducing associated power

- **Migratory Lines**
  - Move **dirty** data between CPUs and skips shared state.
  - Avoids power and latency associated with write back

- **Read/Write cache allocation**
  - With adaptive back-off for temporally inappropriate

**ARM MPCore Snoop Control Unit**

- Clocked at CPU frequency for lower latency lookups and filtered access to CPU
- Keeps data within processor permitting lower power consumption than if time-sliced on a uniprocessor
- Power aware allowing per CPU logic and cache shutdown for advance power management.
Cortex-A9 MPCore Processor Structure

- **Multicore trace and debug**
- **Multicore Scalability**
- **ARM Coresight Multicore Debug and Trace Architecture**
- **FPU/NEON**
- **PTM I/F**
- **Cortex-A9 CPU**
- **Instruction Cache**
- **Data Cache**
- **Cache-2-Cache Transfers**
- **Snoop Filtering**
- **Timers**
- **Accelerator Coherence Port**
- **Advanced Bus Interface Unit**
  - **Primary AMBA 3 64bit Interface**
  - **Optional 2nd I/F with Address Filtering**

- **ARM Coresight**
- **Multicore Scalability**
- **High memory bandwidth**
- **ARM MPCore Technology**

**Diagram Details**:
- **L2 Cache Controller (PL310)**
- **Generalized Interrupt Control and Distribution**
- **Snoop Control Unit (SCU)**
Processor Coherence

- Memory Management Unit Maintenance Operations
  - When a process spans more than 1 processor virtual memory map must be maintained so as to provide a symmetric view of memory
  - Traditional solution by broadcasting an interrupt to all CPU requesting update and blocking until all processors respond
  - Cortex-A9 multicore processors contain hardware to ensure coherence of MMU within the appropriately defined memory regions

- Cache Maintenance Operations
  - For non-coherent devices, it is necessary to ‘flush’ the appropriate cache regions prior to the device access the associated memory
  - Traditional solution also uses a broadcast interrupt based notification
  - Cortex-A9 provides hardware to synchronize cache operation for defined regions of memory
Accelerator Coherence Port

- Sharing benefits of the ARM MPCore optimized coherency design
- Accelerators gain access to CPU cache hierarchy, increasing system performance and reducing overall power
- Uses AMBA® 3 AXI™ technology for compatibility with standard un-cached peripherals and accelerators

![Diagram of Accelerator Coherence Port]

- **Cortex-A9 MPCore (1-4 CPUs)**
  - CPU
  - D$ I$
  - MPCore Technology / SCU
  - ACP
- **L2 Cache**
  - shared, with per-master lockdown to limit high-throughput master flooding
- **Main Memory / SoC**
- **“Event pulse” enabling next cycle notifications**
- **Local Coherence Bus (no snooping on bus)**
  - AMBA 3 AXI
- **DMA**
  - Example: Write
  - Writes clean and invalidates L1 lines if necessary
  - optionally allocating into the shared L2 cache
- **Crypto**
  - Example: Read
  - May hit and resolve in CPU’s L1 cache
  - else may hit in shared L2 cache
  - else read from main memory
Advanced Bus Interface Unit

- Increased tolerance to memory latencies increasing processor capacity
  - Up to 16 outstanding bus transactions per processor
    - Including 4 cache line-fill requests to unblock pipeline due to memory stalls
  - Out of order execution enabling speculative execution of more than 30 instructions
  - Multicore L1 cache-2-cache transfer capabilities reducing system power
  - Exclusive storage between L1 and L2 cache to maximize cache effect

- Processor-speed partitioning of address map enabling highest performance access to key system components
  - With support for low-latency synchronous half-clock bus ratios providing higher memory performance and further flexibility in low-power design

- Ability to fully utilize the bandwidth of dual 64-bit AMBA 3 AXI bus interfaces in data streaming scenarios
  - Load balance memory requests with the primary interface to offer more than 8GBytes/s to the main memory system
  - Optimizing throughput and power efficiency for key networking and connected computing in mobile phones, gateways and image processing devices
Generic Interrupt Controller and TrustZone

- Framework for enhanced software protection and security services
  - Any processor can run normal or secure software at any moment
- Memory is coherent for both normal and secure memory
- Integrated and standardized interrupt controller with security
  - Routes normal interrupts to IRQ and secure interrupts to FIQ
  - Provides normal and secure inter-processor interrupts and virtualization
- Architecture scales to support both single processor and SMP operating systems in normal and secure domains
Improved Paravirtualization

- **Today’s Paravirtualization solutions**
  - Provide the ability to run multiple independent OS/RTOS on a single processor
  - Requires a notable modification of a OS port to appropriately defer all privileged operations to a virtual machine manager (VMM) managing processor resource sharing

- **Cortex-A9 MPCore multicore processor**
  - Enables concurrent execution of multiple paravirtualized operating systems
  - Providing improved real-time response and dynamic load balancing

- **ARM TrustZone Architecture**
  - Allows the open OS to maintain their User and Privilege states and run the VMM in the privilege contexts of the secured software domain
  - Provides an AMBA architecture mechanism to signal TrustZone context accesses to peripherals and any system based memory protection units
  - Cache state is maintained and secured between the open and TrustZone OS context during virtualization traps and request forwarding

- **ARM GIC Architecture**
  - Allows the VMM to manage and arbitrate access to drivers for each open OS

- Together providing an accelerated and simplified paravirtualization solution
Cortex-A9: Summary

The **Cortex-A9 single core processor** and **Cortex-A9 MPCore multicore processor** are poised to transform embedded computing.

- Offering unprecedented performance within tight power and cost constraints
  - Performance scalability through the enhanced MPCore technology
  - Utilizing an unique superscalar, dynamic-length, multi-issue, out of order, speculative-execution microarchitecture
- Providing application specific acceleration though FPU and NEON units
- Supported by the ARM ecosystem and the existing SMP and OS solutions

- Performance and power scalability from 600 to 8000 DMIPS providing the solution for various market requirements
  - ARM MPCore technology providing advanced controls to further lower power
  - Enabling higher performance for today’s battery-operated devices
  - Fully maintaining software investments and Cortex architectural compatibility