QorIQ Portfolio: What’s New with Power Architecture Cores

Jeffrey Ho
Business Development
r26191@freescale.com
Agenda

• Power Arch Cores Overview
  - e500mc, e5500, e6500 core brief
• e6500 core details
  - Shared L2 Cache
  - Multithreading
  - Memory Subsystem Enhancements
  - MMU Enhancements
  - AltiVec Technology
  - Enhanced Power Management
  - General Improvements
### e500mc
- E500v2 enhanced for multi-core
- Hypervisor, MMU TLB w/64 variable size
- DP FPU (classic), decorated L/S instructions
- CoreNet Coherency Fabric (2x snoop BW), L3 cache
- Up to 8 cores, 1.5GHz nominal

#### e500mc Diagram
- L2 Cache
- e500mc
- L3Cache
- DDRC
- P40xx, P30xx, P204x

### e500v2
- 32b PPC-E
- SS, OOO, DI, 7st
- SPE & EFPU
- ECM Bus & Platform L2
- Up to 2 cores, 1.5GHz nominal

#### e500v2 Diagram
- e500 v2
- L2 Cache
- DDRC
- MPC85xx, P10xx, P202x

### e6500
- E500mc plus
- 64b ISA 2.06, 36b (64GB) per process
- Extended Branch Predictor
- Full speed DP FPU (classic)
- Larger L2
- 2.2GHz in 45SOI

#### e6500 Diagram
- L2 Cache
- e5500
- L3Cache
- DDRC
- P5040, P5021, P5020, T1040

#### e6500 Features
- Dual SMT Thread
  - 2 load/store per core (1 per thread)
  - 4 issue (2 decode/completion per thread)
- 2x wide fetch
- AltiVec
- Banked L2
- 2.0GHz in 28nm many core
e500mc Core

- Embedded Hypervisor support
  - Hypervisor privilege level
  - Allows SMP/AMP operation without OS modifications
- Cache Enhancements
  - 32-Kbyte Data and Instruction Caches w/ tag and data parity
  - Recovery from single-bit L1 errors w/ Write shadow mode
  - Back Side L2 w/ tag parity and ECC data
  - 64-byte cache lines
  - Cache intervention, stashing and MESI cache coherency
  - Cache line locking supported
- Classic Double Precision FPU
- Expanded interrupt model
  - HW interrupt ACKs
- Interprocessor Signaling APU
  - Mqsgnd instruction and Doorbell interrupts
- Decorated Storage APU
  - Lockless Atomic updates of shared counters and statistics
- Enhanced Debug Capabilities
  - NEXUS Level II ++
    - Data trace lite
    - Enhanced trace filtering
  - Performance Monitor
    - Watchpoint triggering

- Power Architecture compatible
- Superscalar dual-decode, quad-issue: out-of-order execution/in-order completion
- 64 TLB SuperPages, 512-entry 4K Pages, 36-bit Physical Address
- Branch unit with a 512-entry, 4-way set associative Branch Target/History
- Three integer units: 2 Simple, 1 Complex for integer Multiply, Divide
**e5500 Core**

- 64-bit Power Architecture
- Embedded hypervisor support
- Caches
- 32-Kbyte Data and Instruction Caches w/ tag and data parity
  - Recovery from single-bit L1 errors w/ write shadow mode
  - 512KB Back Side L2 w/ tag parity and ECC data
  - 64byte cache lines
  - Cache intervention, stashing and MESI cache coherency
  - Cache line locking supported
- Classic Double Precision FPU
  - Faster FPU: 2X faster SP, 4X faster DP over e500mc
- New Power ISA v.2.06 Instr
  - instructions for byte- and bit-level acceleration: Parity, population count, bit permute, compare bytes, FPU convert to/from 64-bit integer
- Enhanced Debug Capabilities
- 45-nm Technology
- 2 GHz Implementation point
- 20% DMIPS/MHz improvement over e500mc

- Superscalar dual-decode, quad-issue: out-of-order execution/in-order completion
- 64 TLB SuperPages, 512-entry 4K Pages, 36-bit Physical Address
- Branch unit with a 512-entry, 4-way set associative Branch Target/History, Link stack
- Three integer units: 2 Simple, 1 Complex for integer Multiply, Divide
e6500 Core Complex

- 64-bit Power Architecture®
- 28nm technology
- Superset of e5500
- Two threads per core (SMT)
- Dual load / store units, one per thread
- Shared L2 in cluster of 4 cores (8 threads per cluster)
  - 2048KB 16-way, 4 banks
  - High-performance eLink bus between Ld/St and instruction fetch units
- Power
  - Drowsy core and caches
  - Power Mgt Unit
  - Wait-on-reservation instruction
- Enhanced MP Performance
  - Accelerated atomic operations
  - Optimized barrier instructions
  - Fast intra-cluster sharing
- AltiVec SIMD Unit
- CoreNet BIU
  - 256-bit Din and Dout data busses
- 36/40-bit Real Address
  - 1 Terabyte physical addr. space
- Hardware Table Walk
- LRAT
  - Logical to Real Addr. Translation mechanism for improved hypervisor performance

- Each thread: superscalar, seven-issue, out-of-order execution / in-order completion, branch units with a 512-entry, 4-way set associative branch target / history/Link stack
- Execution units: 1 load / store unit per thread, 2 simple integer per thread, 1 complex for integer multiply & divide, 1 floating-point unit, AltiVec technology
- 64 TLB superpages, 1024-entry 4K pages, 36 or 40-bit physical Address
e6500 Pipeline
Observation:
• Quad decode / dispatch (2 per thread)
• Quad completion (2 per thread)
• 14 issue (7 per thread)
• 12 execution units
Shared L2 Cache
Shared L2 Cache

- Each L2 cache in the system has 4 e6500 cores attached
  - The cores “share” the L2 cache and are connected by a synchronous interface called eLink
  - The combination of the L2 cache and the cores, which are connected to it, are called a core cluster or sometimes just a cluster.
- Each cluster contains a single CoreNet port which connects the cluster (and the cores) to the rest of the system
- Each cluster also contains other core-specific logic for the cores connected to it:
  - Reservations for each core thread
  - Logic specific to power management to awaken sleeping cores when messages are received (from msgsnd)
Shared L2 Cache (cont’d)

- L2 cache is inclusive of L1 caches
- L1 data caches are write-through
- Coherency handled in L2 cache
- L2 cache can be partitioned
- Each bank can process requests simultaneously
- The L2 cache and its cores are called a cluster
- There can be multiple clusters connected via CoreNet fabric
- Coherency among cores in a cluster is faster (lower latency)
- L2 cache protected by ECC on tags and data
- L1 caches protected by parity on tags and data
- Single bit errors are fully recoverable in any cache
Core Cluster L2 Cache Characteristics

- 2MB per cluster, 16-way associative, in 4 banks
- Each bank can operate in parallel with the other banks
- 1 CoreNet port (core clusters are connected to the rest of the system and other core clusters via CoreNet fabric)
- 64-byte cache line size
- Dynamic harvard tagging of cache lines (N bit)
- The L2 cache is the point of cache coherency in the system:
  - Core L1 data caches are write-through in that they do not contain any modified data (similar to “write shadow" mode from e5500/e500mc)
  - L2 cache is inclusive for all cores L1 data caches in the cluster
  - If a data line is evicted from the L2 cache it will be invalidated in any core’s L1 data cache that contains the line
- The L2 cache is ECC protected (single bit error correct, double bit error detect)
L2 Cache Programming Model Changes

• In general, the programming model is the same as e5500 L2 cache except that the registers are not SPRs, but are accessed through memory mapped registers
  – Not that big of an issue since most control and configuration is done at initialization time
  – EREF 2.1 defines the memory mapped interface and defines the registers for control and configuration
  – Additional registers are provided for L2 cache partitioning controls similar to how partitioning is done in the CPC

• Cache locking is supported. However, if a line is unable to be locked, that status is not posted. Cache lock query instructions are provided for determining whether a line is locked
Multithreading
Multithreading – General Overview

- Each core contains 2 threads (thread 0 and 1)
- The implementation is more similar to fused cores than traditional multithreading, leading to much better performance scaling than most industry implementations. Each thread performs almost as well as a separate core:
  - Scaling for the threads is between .8 and 1.0, making the 2 threads equivalent to 1.6 – 2.0 cores, with most applications in the higher range
  - With the additional performance improvements, each thread is more powerful than an e500mc core
  - Many of the internal resources of the core are replicated for better performance (each thread has its own load/store unit and its own simple integer units, as well as instruction issue queues and completion unit)
- Threads can be disabled or enabled. If disabled, they do not fetch or execute instructions
Multithreading – Programming Model

• In the documentation (EREF and e6500 Core Reference Manual) the threads are denoted as processors:
  - When something is unique to the core itself, the term core or multithreaded processor
  - In general, system software treats a multithreaded processor with threads almost the same as multiple single-threaded cores (i.e. an SMP programming model)

• Other than some shared resources owned by the hypervisor privileged level, the programming model is essentially the same as what it would be for multiple single-threaded cores
  - Linux® will treat each thread as it would a separate core except for some very special cases that involve modifying shared state
Multithreading – Private vs. Shared State

- All the states in a multithreaded processor are either shared or private:
  - For registers and other architected states that are shared between threads:
    ▪ There is only one copy of the register or architected state
    ▪ A change in one thread affects the other thread if the other thread reads it
  - Registers and other architected states that are replicated per thread are private to the thread:
    ▪ There is one copy per thread of the register or architected state
    ▪ A change in one thread does not affect the other thread if the thread reads it
- EREF 2.1 and e6500 Core Reference Manual denote whether each register is private or shared (a register is private if it is not denoted as shared)
Multithreading – Private State

- Most resources are private
  - GPRs, FPRs, VRs, etc. are all private (and associated renames)
  - Most SPRs are private. If it's something that can be written by guest supervisor or user it is always private
  - Almost anything involved in normal computational software will be private
  - All interrupts are private (i.e. each thread has its own set of signals, registers, etc.)
  - PIR, processor ID register, is private and contains a different value for each thread on reset (the value is unique among all threads on all cores in the entire SoC)
  - TCR, timer control register, and TSR, timer status register, are private, making the timers that these registers control private as well (watchdog timers are private - a final timeout from either thread will send such on the WRS, watchdog reset status, signal pins, which may result in the entire core being reset)
Multithreading – Shared State

- L1 data cache (and associated configuration and control registers) is shared, although each cache line has a valid bit per thread.
- IVPR, interrupt vector prefix register, and IVORs, interrupt vector offset registers, are shared. GIVPR and GIVORs are private. This is because the hypervisor runs on both threads, but different guests can run in different threads.
- Any registers that are accessed through the memory map (like the L2 registers) are shared.
- MMU control and configuration registers (MMUCFG, MMUCSR0, TLBnCFG0, LRATCFG, LRATPS, TLB0PS, TLB1PS, EPTCFG) are shared – note that MAS registers, PID, LPIDR, and external PID registers are private.
- TLB entries are shared, although there are private data L1 MMUs per thread.
Multithreading – New Registers

• TEN register (SPR, hypervisor only):
  - Used to enable and disable threads
  - Each bit in TEN corresponds to a thread
  - There are different SPR numbers for performing different actions:
    ▪ TENS - setting bits (enabling)
    ▪ TENC – clearing bits (disabling)
    ▪ TENSR – status bits (when enabling and disabling, TENSR represents that the operation is complete)
  - The process of enabling a disabled thread is context synchronizing in the enabled thread
    ▪ This feature is used to synchronize the effects of updating shared resources
• TIR, thread identification register (SPR, hypervisor read only):
  - A read-only register that returns the current thread number
  - The value of TIR will match the thread field of PIR at reset
Multithreading – New Registers (cont’d)

- Thread Management Registers
  - Like SPRs and PMRs, except they have a separate address space
  - New instructions, mttmr and mftmr
  - Currently all TMRs are hypervisor only
  - Initial next instruction address (INIA_n) – allows a thread to write another thread’s NIA, only when the target thread is disabled
  - Initial machine state register (IMSR_n) – allows a thread to write another thread’s MSR, only when the target thread is disabled
  - Thread priority register (TPRIn) - allows a thread to write another thread’s priority
    - Thread priorities currently have no effect
  - Thread management configuration register – information about how threads are implemented by the multithreaded processor
Multithreading – Debug

• Debug
  – Almost all resources are private. Internal debug works as if they are separate cores
  – External debug is private per thread. An option exists to halt both threads when one thread halts
Multithreading – Power Management

- Power Management
  - Power management control is per-thread (and the associated SoC programming model will be per-thread)
  - Actual power management will only occur when both threads reach the same power management state
  - For example, when wait occurs on one thread, fetching stops for that thread, but we don’t go into lower power states until both threads execute wait
  - PMGMTCR0 is a shared register used to configure and control power management capabilities
Memory Subsystem Enhancements
New Memory Subsystem Instructions

- Additional sizes for load and reserve and store conditional instructions (like `lwarx` and `stwcx`):
  - `lbarx`, `lharx` – load and reserve [byte,halfword] instructions for atomic updates on byte and halfword data types
  - `stbcx`, `sthcx` – store conditional [byte,halfword] instructions for atomic updates on byte and halfword data types

- Just like `lwarx` and `stwcx`. except for byte and halfword sizes

- Programmer should not mix sizes between load and reserve and store conditional
  - For example, doing `lbarx` (byte) and `stwcx` (word) is not guaranteed to work because the sizes are different

- The new sizes are useful for doing atomic updates to byte and halfword sizes more efficiently
Memory Subsystem Enhancements – miso

• **miso** is a hint to the processor that performance may improve if previous stores are sent to the coherency point
• On e6500, **miso** causes previous stores that are sitting in the store gather buffer to be sent to the L2 cache
  • If another core is waiting to see the results of a store, it will see it faster

```
<table>
<thead>
<tr>
<th>Core 0</th>
<th>Core 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ori r5,r0,1</td>
<td>lwz r6,A</td>
</tr>
<tr>
<td>stw r5,A</td>
<td>cmpwi r6,1</td>
</tr>
<tr>
<td>miso</td>
<td>bne loop</td>
</tr>
</tbody>
</table>
```

• **miso** – “make it so”
• **miso** completes before the stores are pushed so it cannot be used to determine when the stores are performed.
• **miso** executes as a NOP on older processors
New Memory Subsystem Instructions

- Elemental memory barriers perform faster memory barriers with explicit barriers between loads and stores for memory ordering:
  - There are 4 new elemental barriers, which apply only to memory accesses, that are not Cache Inhibited and not Write Through Required (i.e. normal cacheable write-back memory):
    - Barrier previous loads with future loads (load with load, $E = 0b1000$)
    - Barrier previous loads with future stores (load with store, $E = 0b0100$)
    - Barrier previous stores with future loads (store with load, $E = 0b0010$)
    - Barrier previous stores with future stores (store with store, $E = 0b0001$)
• The elemental barriers are specified with a new $E$ operand to
  the *sync* instruction: *sync L,E*
  - The $E$ operand is a 4-bit field where each bit corresponds to the
    elemental barriers from the previous slide
  - The barriers that are enforced are the total of all the bits in the $E$
    field which are set. That is, any combination of all 4 barriers can be
    specified with one instruction. (E.g. *sync L,0b1001* creates a barrier
    for loads with loads and stores with stores)
• An extended mnemonic, *esync E* is provided to specify the
  barriers
  - The assembler will set the value of L in the *sync* instruction so that
    previous implementations, such as e500mc and e5500, will still
    perform the requested barrier correctly (although the barrier may be
    stronger than what is specified)
Memory Subsystem Enhancements – Cache Lock Query

- 2 new instructions to determine whether a cache line is locked:
  - dcblq. \(CT, rA, rB\) – queries whether the line addressed by \(rA\) and \(rB\) is locked in the data cache specified by \(CT\)
  - icblq. \(CT, rA, rB\) – queries whether the line addressed by \(rA\) and \(rB\) is locked in the instruction cache specified by \(CT\)

- If the line is locked, CR0 is set to 0b001 || XER[SO]
- If the line is not locked, CR0 is set to 0b000 || XER[SO]

- Unable to lock status bits in L1CSR0 and L1CSR1 are no longer present

- CT can only be 0 (for primary caches) or 2 for the L2 cache
  - Other CT values will be added in the future (L3 cache for example)
Other Memory Subsystem Enhancements

- The e6500 has a larger store queue than e5500
- The load store unit contains store gather buffers to collect stores to cache lines before sending them on eLink to the L2 cache
- There are no more Line Fill Buffers (LFB) associated with the L1 data cache
  - These are replaced with Load Miss Queue (LMQ) entries for each thread
  - They function in a manner very similar to LFBs
- Note there are still LFBs for L1 instruction cache
MMU Enhancements
TLB Enhancements

- MMU changes (MMU architecture version 2)
  - TLB1 (variable size pages) supports power of 2 page sizes (previous cores used power of 4 page sizes)
    - 4KB to 1TB page sizes
  - TLB0 increased to 1024 entries, 8-way associativity (from 512, 4 way)
    - Fixed 4KB page sizes
    - Supports HES (hardware entry select) when written to with `tlbwe`
- LRAT (logical to real address translation)
  - The LRAT converts logical addresses (addresses guest operating system thinks are real) and converts them to true real addresses
  - Translation occurs when guest executes `tlbwe` and tries to write TLB0 or during hardware tablewalk for a guest translation
  - Does not require hypervisor to intervene unless the LRAT incurs a miss
  - 8 entry fully associative supporting variable size pages from 4KB to 1TB
TLB Enhancements (cont’d)

• PID register is increased to 14 bits (from 8 bits)
  – Now the operating system can have 16K simultaneous contexts
• Real address increased to 40 bits (from 36 bits)
Hardware Tablewalk Overview

- The e6500 hardware tablewalk is not like the Power Architecture classic hardware tablewalk. It is **not** a reversed hashed page table. It is more friendly to Linux
  - The definition comes from Power ISA 2.06
  - It defines a page table structure and PTE (page table entry) in-memory format which is compatible with what native Linux uses

- This is MMU architecture version 2 (V2) which is denoted by bits in the MMUCFG register
  - Different formats for all the “config” registers (TLBnCFG for example)
  - There are new config registers for TLB page size (TLBnPS) and LRAT page size (LRATPS)
AltiVec Technology
Altivec SIMD Numeric Acceleration

- Altivec – Vector instructions that operate on quad-word operands
- Vector (Altivec) A 128-bit wide SIMD attacks parallel data-oriented compute application
  - 32, 128 bit vector registers
  - 16 x 8, 8 x 16, 4 x 32 bit integer operations per clock
  - 4 x 32-bit IEEE floating point operations per clock
  - Powerful ‘permute’ unit (splats, shifts, rotates)
AltiVec Technology

• AltiVec technology for the e6500 core is essentially the same as AltiVec technology from the e600 core, except for the following:
  
  - Adds new instructions for computing absolute differences
    
    - vabsdub – absolute differences (byte)
    - vabsduh – absolute differences (halfword)
    - vabsduw – absolute differences (word)
    - These speed up in the inner loop of motion estimation video processing
  
  - Adds new instructions for dealing with misaligned vectors more easily
    
    - lvtx[l], stvfx[l] – load/store vector to/from left [LRU]
    - lvrtx[l], stvrfx[l] – load/store vector to/from right [LRU]
    - lvswx[l], stvswx[l] – load/store vector with left/right swap [LRU]
AltiVec Technology (cont’d)

- AltiVec technology for the e6500 core is essentially the same as AltiVec technology from the e600 core, except for the following:
  - Adds new instructions for dealing with elements of vectors
    - lvexbx, stvebx – load/store vector element indexed byte
    - lvexhx, stvehx – load/store vector element indexed halfword
    - lvexwx, stvewx – load/store vector element indexed word
    - These allow loading/storing of arbitrary elements to arbitrary addresses
- Instructions for moving data from GPRn to vector register
  - mvidsplt – move to vector from integer double word and splat
  - mviwsplt – move to vector from integer word and splat
AltiVec Technology - Power Management

• AltiVec execution units will go drowsy (reducing static power) when not used
  - Software must bring it out of drowsy using the CDCSR0 register
  - Software will take an AltiVec unavailable interrupt when drowsy and the hypervisor or operating system can turn it back on
  - All AltiVec state is retained when the unit is drowsy
Enhanced Power Management
Enhanced Power Management

• Drowsy core
  - Core will go drowsy when executing **wait**, retaining state and reducing static power
    ▪ Can go drowsy after so many wait cycles or immediately after executing wait if commanded to (second operand to wait instruction)
    ▪ Wakes up from drowsy state and starts execution when a cache stash occurs, an interrupt occurs or a store to the threads reservation address occurs (if **waitrsv** was the cause of the wait)
  ▪ AltiVec unit can go drowsy as well

• New form of **wait** instruction, “wait on reservation” (**waitrsv**)
  - **waitrsv** will **wait** until an interrupt occurs or the reservation for the thread is clear:

    ```assembly
    loop:    lwarx r6,A       // set reservation for address A
            cmpwi r6,0       // if it’s the value we want go process it
            beq gotit
            waitrsv         // wait at low power until some other processor
            bu loop         // stores to A
    gotit:  
    ```