

Implementing the Thread Programming Model on Hybrid FPGA/CPU Computational Components

David Andrews, Douglas Niehaus, Razali Jidin
Information Technology and Telecommunications Center
Department of Electrical Engineering and Computer Science
University of Kansas
{dandrews,niehaus,rjidin}@ittc.ukans.edu}

Abstract

Designers of embedded systems are constantly challenged to provide new capabilities to meet expanding requirements and increased computational needs at ever improving price/performance ratios. Recently emerging hybrid chips containing both CPU's and FPGA components have the potential to enjoy significant economies of scale, while enabling system designers to include a significant amount of specialization within the FPGA component. However, realizing the promise of these new hybrid chips will require programming models supporting a far more integrated view of the CPU and FPGA components than provided by current methods. This paper describes fundamental synchronization methods we are now developing for supporting a multi-threaded programming model that provides a transparent interface to the CPU and FPGA based component threads.

1. Introduction

Designers of embedded and real time systems are continually challenged to provide increased computational capabilities to meet tighter system requirements at ever improving price/performance ratios. Best practice methods have long promoted the use of commercial off the shelf (COTS) components to reduce design costs and time to market. Creating COTS components that can be reused in a wide range of real-time and embedded applications is a still a difficult challenge, in part, because it requires the simultaneous satisfaction of apparently contradictory design forces: generalization and specialization. Systems designers are all too familiar with the tension caused by these opposing forces in trying to balance cost versus performance.

Recently emerging hybrid chips containing both CPU and FPGA components are an exciting new

development that promise COTS economies of scale, while also supporting significant hardware customization. For example, Xilinx [4] offers the Virtex II Pro which combines up to four Power PC 405 cores with up to approximately 4 million free gates, while Altera [5] offers the Excalibur, which combines an ARM 922 core with approximately the same number of free gates. Designers now have the freedom to select a set of FPGA IP to create a specialized System-on-a-Chip (SoC) solution. These capabilities allow the designer to enjoy the economies of scale of a COTS device but based on a selected set of IP that produces a design tailored for their specific requirements. Additionally, the free FPGA gates may also be used to support customized application specific components for performance critical functions. While the performance of an FPGA based implementation is still lower than that of an equivalent ASIC, the FPGA based solution often provides acceptable performance but with a significantly better price/performance ratio.

Tapping the full potential of these hybrid chips presents an interesting challenge for system developers. Specifying custom components within the FPGA requires knowledge of hardware design methods and tools, which dangles the full potential of these hybrids tantalizingly out of reach for the majority of system programmers. Researchers are seeking solutions to this barrier by investigating new design languages, hardware/software specification environments, and development tools. Projects such as Ptolemy [2], Rosetta [9], and System-C [6] are investigating system level specification capabilities that can drive software compilation and hardware synthesis. Other projects such as Streams-C [3] and Handel C [7] are focused on raising the level of abstraction at which FPGAs are programmed from one of gate-level parallelism to that of modified and augmented C syntax. System Verilog [8] and a newly evolving VHDL standard [10] are also now being designed to abstract away the distinction between the two sides of the traditional low level

hardware/software interface into a system level perspective. Although these approaches differ in the scope of their objectives, they all share the common goal of raising the level of abstraction required to design and integrate hardware and software components.

A good question then, is if high level programming language capabilities for FPGA's continue maturing at current rates, will this be sufficient to allow software engineers to apply their skills across the FPGA/CPU boundaries? Unfortunately, current hybrid programming models are still immature, generally treating FPGA's as independent accelerators with their computations outside the scope of the programs running on the CPU. Communications between the FPGA based and CPU based computations are generally through simple input/output queues, and the immaturity of the hybrid model does not provide for synchronizing the execution of the component computations: a critical capability in distributed and parallel computation. The addition of a high level programming model is needed that abstracts the FPGA/CPU components, bus structure, memory, and low level peripheral protocols into a transparent system platform [12]. In general, programming models provide the definition of software components as well as the interactions between these software components (see Lee [1] for a discussion of software frameworks for embedded systems). Message passing and shared memory protocols are two familiar forms of component interaction mechanisms in use today. Both have been successfully used in the embedded world and practitioners enjoy debating the relative merits of their personal choice.

This paper presents work being performed in extending the multithreaded programming paradigm across hybrid architectures. This capability, coupled with recent advancements in hardware synthesis from high level languages will open up the potential of the reconfigurable hardware to system programmers.

2. Multithreaded Programming Model

The multi-threaded programming model is convenient for describing embedded applications composed of concurrently executing components that synchronize and exchange data. Under such a multi-threaded programming model, applications are specified as sets of threads distributed flexibly across the system CPU and FPGA assets. The familiar multi-threaded programming model can greatly reduce design and development costs as the computational structure of hybrid applications at the highest level of abstraction remains familiar. Whether the threads implementing a

computation are CPU-based or FPGA-based can become just one more of the available design and implementation parameters with resource use and application performance implications. How to perform this partitioning to best support the needs of an application or system is yet another challenging problem currently being investigated.

At this point it is appropriate to draw a distinction between policy and mechanism. The policy of the multi-threaded model is fairly simple: To allow the specification of concurrent threads of execution and protocols for accessing common data and synchronizing the execution of independent threads. On a general purpose processor, the mechanisms used to achieve this policy include the definition of data structures that store thread execution state information, and the semantics of how thread synchronization interacts with the operating system thread scheduler. Unlike general purpose CPU's, FPGA's provide no a priori computational model. Although the lack of an existing computational model at first glance seems to be a liability, instead it is an asset because it presents an opportunity to create efficient mechanisms for implementing FPGA threads and for supporting thread synchronization within the FPGA and across the CPU/FPGA boundary.

A key aspect of the multithreaded programming model is the ability for independent threads to co-ordinate execution and share data using standard synchronization primitives. We present our work in designing hybrid thread synchronization mechanisms that can be used in embedded systems that contain both hardware and software threads. A significant aspect of our approach is our ability to achieve the base synchronization semantics, but without relying on traditional assembly language conditional instructions, or additional memory coherence mechanisms. This is critical point as new evolving computational models being developed for hardware threads will not need to adhere to these processor family dependent instructions and memory coherence system capabilities, thus simplifying system design.

3. Hybrid Thread Synchronization

Semaphore implementations on general purpose CPU's are based an atomic read and (conditional) write of a shared variable. In modern multiprocessor implementations, these operations occur as dependent pairs of conditional instructions, such as load linked and store conditional [11]. These instructions have evolved from the original test and set, and compare and swap type instructions in order to eliminate the

need to lock the system bus for long periods of time. These instructions require additional control logic within the CPU that interfaces into the memory coherency policy of the system. While semantically correct, these existing mechanisms introduce significant complexity in the system design that is not easily portable when creating parallel hardware threads. Instead of replicating these mechanisms, we use the FPGA to implement more efficient mechanisms that are CPU family independent, and require no additional control logic to interface into the system memory coherence protocol. As such, our new mechanisms are easily portable across shared and distributed memory multiprocessor configurations. The basic mechanisms defined use a standard write of a thread id into a memory mapped request register. We define a simple control structure within the FPGA that conditionally accepts or denies the request. The thread requesting the write then performs a read operation of an "owner" register to see if its thread_id has been accepted as the new owner of the lock. This basic policy is the core upon which more complex synchronization mechanisms are constructed. The design of binary and counting semaphores, both busy wait and blocking, are presented below.

3.1 Binary Spin Lock Semaphores

The block diagram for a binary spin lock semaphore is shown in Figure 1.

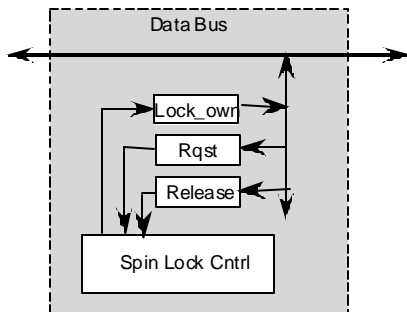


Figure 1. Spin Lock Binary Semaphore

The basic semantics of all API's for accessing the lock are implemented identically in both hardware and software threads, and are made available as library routines to the system developer. To request the semaphore, the API first writes the thread_id into the request register.

After the thread_id has been written, the API then reads back the lock_owr register and compares to see if its thread_id is now the lock owner. To release the

semaphore, the thread writes its thread_id into the release register.

When a thread_id is written into the request register, if the semaphore is free, then the control logic implemented as a state machine within the semaphore IP updates the lock_owr register. If the semaphore is currently locked, then the control logic performs no update. After the first access, the lock is only freed when a thread writes into the release register.

3.2 Spin lock counting semaphore

The block diagram of the counting semaphore is shown in Figure 2. The thread first gains access to the counting semaphore registers by accessing the binary spin lock. The binary spin lock protects the next two instructions that first write the requested number of resources and then reads back a status. A requesting thread writes its request for a number of resources into the rqst_num register. The semaphore IP then checks to see if sufficient resources are available and sets the grant register. If insufficient resources are available, then a boolean value of 0 remains in the grant register. If sufficient resources are available, then the boolean value of 1 is written into the grant register. In either event, the thread reads the result of the request from the grant register.

A thread can release any number of resources by writing into the rel_num register. Note that no accessing of the spin lock is necessary for releasing resources.

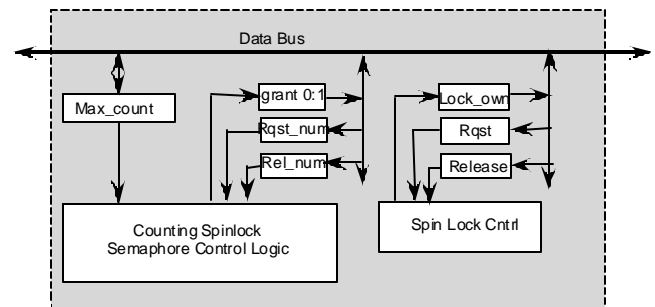


Figure 2. Counting Semaphore

The count register is initialized by the operating system during system initialization and is not needed further by running threads. However, the count register can be reset at any point without requiring a system reboot. After a thread gains access to the controlling spin lock, two ordered operations take place. First, the requested number of resources are written into the request register which is then subtracted from the current count register. After the number or requested resources have been latched, the

grant register is updated with a Boolean value reflecting if the subtraction resulted in a value greater than or equal to zero, or less than zero. If the result was less than zero, then insufficient resources were available, and no further action is required. If the result was greater than or equal to zero, then the grant flag is set to one and the subtracted value is loaded into the count register. The semaphore IP logic will clear the grant register value (set it back to 0) upon a read request to the address of the register.

To release resources, the thread writes the number of resources to be released into the release register. No access of the binary semaphore used by the requesting threads is needed to release resources. Obviously, only one thread will be allowed access to the semaphore IP registers during any bus cycle. Assuming the requesting thread is the current bus owner, the request value is latched at the end of the cycle and will be available to the logic circuits the next clock cycle. If an immediate request for resources during the next clock cycle, the semaphore IP performs the subtraction and sets the grant flag. That flag will stay valid for the request and will not change until the requestor performs the reading of the grant flag. No other requester can cause the recalculation to occur as the request/grant check pair are protected by the spin lock.

3.3 Blocking Semaphores

Blocking semaphores allow threads that cannot gain access to a semaphore to be queued and suspended thus providing more efficient usage of the computing resources and decreasing congestion on the system bus. Our basic mechanism includes queue structures to be associated with each blocking semaphore to hold onto thread ids that are suspended as shown in Figure 3. The release of a blocking semaphore by the current owner is an event trigger from the semaphore IP that may cause granting of the released semaphore to a new owner if a thread is suspended on that particular semaphore. The event trigger must interface to both hardware threads and the operating system. If no thread is suspended, then obviously no change of ownership is required. The semaphore IP itself has the capability of choosing from a queued list of threads whom to grant the semaphore to next. With queues residing (conceptually) within each semaphore the control logic can issue only one (achieving a traditional queuing semaphore), or multiple threads (implementing generic sleep wakeup) when the lock is either freed in the binary case, or resources are returned in the counting semaphore case. In both cases, the centralized interface structure shown below is provided to simplify the interface between the CPU and hardware thread components, and the individual

semaphores. This structure simplifies the interface logic needed for the blocking semaphores to interact with the CPU and hardware thread components, and also reduces the complexity of the interrupt service routine running on the CPU required to query the multiple semaphore components. The structure provides an initial separation between hardware and software threads to eliminate unnecessary context switching on the CPU when a hardware thread is being awakened. The structure forms a basic framework to support our longer term goal of migrating system scheduling decisions from the CPU into hardware.

3.3.1 Blocking Binary Semaphores

The API writes a thread id into the request register and then follows up with checking the owner register similar to the binary spin lock. If the thread did not receive the lock, then the API puts the thread to sleep. Once wakened, the thread will again submit a request to the request register and check for ownership. The lock is released by writing the thread_id into the release register.

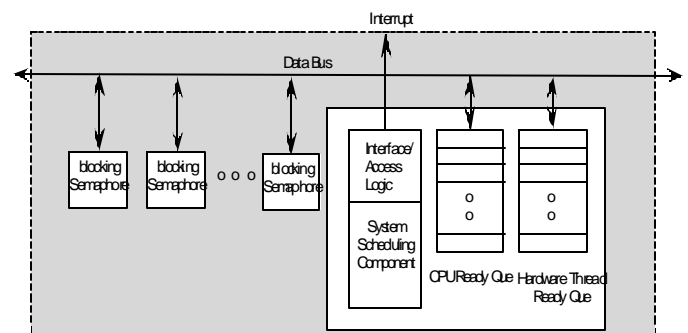


Figure 3. Blocking Semaphore and Interface Framework

The depth of the request queue for our prototype is a system design parameter set at design time. The block diagram of blocking binary semaphore system components is shown in Figure 3. The block diagram of a blocking binary semaphore is shown in Figure 4.

Each thread requests the lock by writing to the request register. If the lock is free, the control logic will update the lock owner register in a single cycle following the request. If the lock is owned by another thread, the control logic queues the request in the next clock cycle. No race conditions for the lock can occur with this policy.

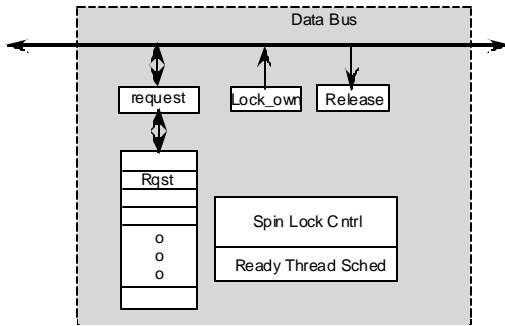


Figure 4. Blocking Binary Semaphore

A release is initiated by the writing of the thread id into the release register. If the queue has entries, the control logic will invoke the ready thread scheduler logic to select which queued request or requests show be signaled to the system as available for rescheduling and signal the event to the system. The event signaling method from within each semaphore consists of writing out thread_ids to ready to run queues: one for the CPU and one for the hardware threads.

The ready thread logic is a generic module that can be tailored to support any particular scheduling algorithm based on specific system requirements. For our prototype, we have implemented a basic protocol that copies the thread_id/@command_register to one of two global memory mapped ready to run queues, one for the CPU, and one for hardware threads. Writing into these queues causes an appropriate mechanism to occur that alerts the receiving component (either the CPU or a hardware thread) that new ids have been queued and are ready to be rescheduled.

3.3.2 Blocking Counting Semaphores

The block diagram of blocking counting semaphore is shown in Figure 5. With the blocking counting semaphores if the resources are available, the thread continues to run. However, if insufficient resources are available, then the thread will be queued and suspended. The semaphore IP will consider issuing queued thread_ids for rescheduling only when resources are released. Implementing blocking semaphores within the hardware provides significant latitude in creating flexible system scheduling policies. We discuss two sample approaches to selecting thread_id's from the suspend queue for scheduling. In the first approach, one or multiple threads that are requesting resources equal to or less than the number that are available will be considered, and in the second approach all threads queued, independent of the number of resources re-allocated, will be considered. The first approach will reduce the number of context

switches and contention for the semaphore, and supports a system scheduling policy of giving priority to tasks that request fewer resources. The second approach considers all queued thread_ids, and allows the scheduler(s) maximum flexibility in the scheduling decision. Our semaphore IP logic provides a scheduling component that can be tailored to specific system policies.

To implement the first approach, the semaphore IP needs to hold onto the number of resources requested, and additionally needs to queue associated thread_id/@command_registers. To implement the second approach, the queue holding the requested number of resources is removed. If the thread needs to be blocked, only the thread_id is queued.

Just as in the spin lock protocol, writing a value into request number register causes a Boolean value to set in the grant register. If insufficient resources are available, then the thread_id needs to be sent before the spin lock is released. The semaphore IP will associate the next thread_id written with the request number that is kept latched in the request_num queue.

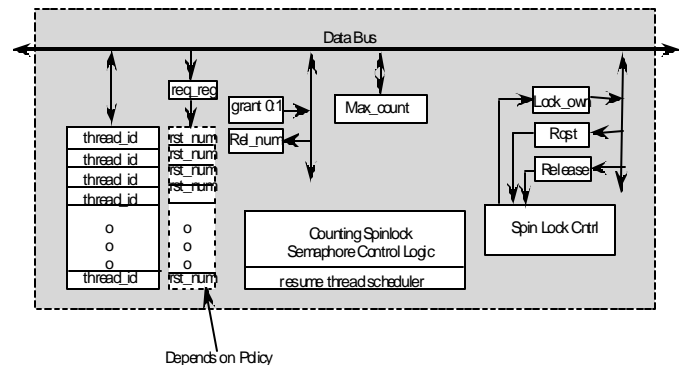


Figure 5. Blocking Counting Semaphore

4. Performance Analysis

In addition to functional tests, tests were performed to quantify the performance of the hardware threads. This was in response to our concern that the speed of the hardware thread requests could lead to blocking and starvation of competing software threads. We performed our analysis on an experimental board that contained a Virtex II Pro PPC/FPGA chip. The Virtex II Pro contained a 100 MHz Power PC microprocessor embedded within the FPGA fabric. The system configuration included a CoreConnect[13] bus with 100 MHz PLB and OPB busses. The test sequence for each thread was a semaphore request, status check, and release. The software thread average access time was

153 clock cycles, compared to 22 clock cycles for the hardware thread, yielding a 7:1 average performance access ratio. We next ran a series of experiments with both hardware and software threads competing for binary spin lock semaphore. Figure 6 summarizes the results of the experiments. The y-axis shows the ratio of the number of hardware to software accesses, and the x-axis normalizes the total number of achieved requests to the maximum number of requests possible by a hardware thread with no contention. We included a delay loop within the hardware thread that was used to variably delay requests. The last data point on the graph represents a zero delay. This point shows that even though the absolute request rate between hardware and software thread accesses was 7:1, when a hardware and software thread were allowed to compete within the system for a semaphore, the observed ratio was 23:1, and the aggregate number of accesses was approximately 85% of the total number that could be achieved from a hardware thread operating with no competition. This indicated that the access efficiency of the hardware thread has definite potential to starve a software thread. In order to achieve a better balance and a fairer competition for accessing the semaphore, the remaining data points show the relative ratio's and normalized percentage of accesses with non zero delays. From this graph, it becomes clear that system designers must consider the speed advantage of a hardware thread in order to achieve a balanced system capability.

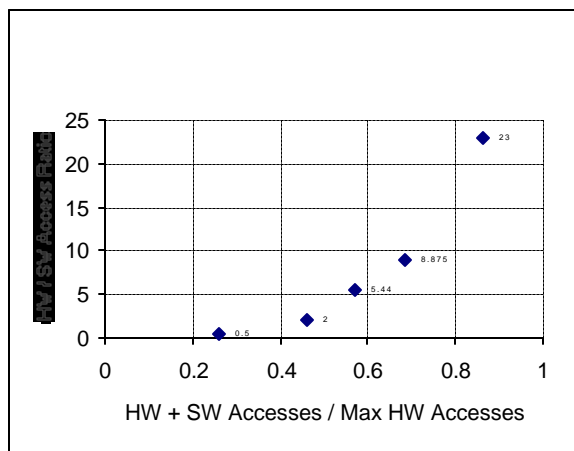


Figure 6. HW/SW Access Results

5. Conclusion

Significant advances in fabrication technology are providing new commercial off the shelf components that combine a general purpose CPU and reconfigurable logic gates (FPGAs). These new devices are a significant step toward realizing a single

component that can support both the generalization of commercial off the shelf components and specialization required for individual embedded applications. Work is currently underway in developing unified programming models that allow computations within and across the hybrid hardware components to be expressed using the familiar multi-threading programming paradigm. Creating a system level multi-threaded programming capability requires new hardware/software co-design approaches to supporting operating system and application functions.

When complete, this capability will enable these new devices to be accessible by a much broader community of system programmers and provide increases in operating system performance. Enabling the multi-threaded model across the hybrids components will ultimately provide shorter design times and lower development costs. We have presented the design of new mechanisms for supporting the synchronization primitives of the multithreaded model. We are currently performing analysis of these mechanisms and will present performance metrics for the designs. The work is partially sponsored by NSF EHS contract # CCR-0311599

6. Bibliography

- [1] Lee, Edward, "Whats ahead for Embedded Software?", IEEE Computer, Sept 2000, pp. 18-26
- [2] Lee, Edward, Overview of the Ptolemy Project, Technical Memorandum, March 6, 2001, UCB/ERL M01/11 University of California
- [3] Maya B. Gokhale and Janice M. Stone and Jeff Arnold and Mirek Kalinowski, Stream-Oriented FPGA Computing in the Streams-C High Level Language, Proceedings of the Eight Annual IEEE Symposium on Filed-Programmable Custom Computing Machines (FCCM), April 2000, pp. 49-56
- [4] www.xilinx.com
- [5] www.altera.com
- [6] www.systemc.org
- [7] www.celoxica.com
- [8] www.eda.org/sv-cc/
- [9] Perry Alexander and Cindy Kong, Rosetta: Semantic Support for Model Centered Systems Level Design, IEEE Computer, November, 2001, pp. 64-70
- [10] www.eda.org/vhdl-200x/
- [11] Hennessey J.L., and Patterson, D. A., "Computer Architecture: A Quantitative Approach", 3rd Edition, Morgan Kaufmann, 2003
- [12] Andrews, D.L., Neihaus, D., Ashenden, P. " Programming Models for Hybrid FPGA/CPU Computational Components:", IEEE Computer, January 2004
- [13] www-3.ibm.com/chips/products/coreconnect/