<u>**Preliminary Draft of SPAA'99 Submission**</u>

A System-Level Specification Framework for I/O Architectures[1]

Mark D Hill, Anne E Condon, Manoj Plakal, Daniel J Sorin
Computer Sciences Department
University of Wisconsin - Madison
`{markhill,condon,plakal,sorin}@cs.wisc.edu`

Contact Author: Mark D. Hill, `markhill@cs.wisc.edu`

Abstract

A computer system is useless unless it can interact with the outside world through input/output (I/O) devices. I/O systems are complex, including aspects such as memory-mapped operations, interrupts, and bus bridges. Often I/O behavior is described for isolated devices without a formal description of how the complete I/O system behaves. The lack of an end-to-end system description makes the tasks of system programmers and hardware implementors more difficult to do correctly.

This paper proposes a framework for formally describing I/O architectures called Wisconsin I/O (WIO). WIO extends work on memory consistency models (that formally specify the behavior of normal memory) to handle considerations such as memory-mapped operations, device operations, interrupts, and operations with side effects. Specifically, WIO asks each processor or device that can issue k operation types to specify ordering requirements in a k x k table. A system obeys WIO if there always exists a total order of all operations that respects processor and device ordering requirements and has the value of each "read" equal to the value of the most recent "write" to that address.

This paper then illustrates WIO with a directory-based system with a single I/O bus. We describe this system's ordering rules and protocol in detail. Finally, we apply our previous work using Lamport's logical clocks to show that our example implementation meets its WIO specification.

Keywords: input/output, memory consistency, cache coherence, verification

1. This work is supported in part by the National Science Foundation with grants MIP-9225097, MIPS-9625558, CCR 9257241, and CDA-9623632, a Wisconsin Romnes Fellowship, and donations from Sun Microsystems and Intel Corporation.

1 Introduction

Modern computer hardware is complex. Processors execute instructions out of program order, non-blocking caches issue coherence transactions concurrently, and system interconnects have moved well beyond simple buses that completed transactions one at a time in a total order. Fortunately, most on this complexity is hidden from software with an interface called the computer's "architecture." A computer architecture includes at least four components:

1) The *instruction set architecture* gives the user-level and system-level instructions supported and how they are sequenced (usually serially at each processor).

2) A *memory consistency model (e.g.,* sequential consistency, SPARC Total Store Order, or Compaq Alpha) gives the behavior of memory.

3) The *virtual memory architecture* specifies the structure and operation of page tables and translation buffers.

4) The *Input/Output* (*I/O*) *architecture* specifies how programs interact with devices and memory.

This paper examines issues in the often-neglected I/O architecture. The I/O architecture of modern systems is complex, as illustrated by Smotherman's venerable I/O taxonomy [10]. It includes, at least, the following three aspects. First, software, usually operating system device drivers, must be able to direct device activity and obtain device data and status. Most systems today implement this with *memory-mapped operations*. A memory-mapped operation is a normal memory-reference instruction (e.g., load or store) whose address is translated by the virtual memory system to an uncacheable physical address that is recognized by a device instead of regular memory. A device responds to a load by replying with a data word and possibly performing an internal side-effect (e.g., popping the read data from a queue). A device responds to a store by absorbing the written data and possibly performing an internal side-effect (e.g., sending an external message). Precise device behavior is device specific. Second, most systems support *interrupts* whereby a device sends a message to a processor. A processor receiving an interrupt may ignore it or jump to an interrupt handler to process it. Interrupts may transfer no information (beyond the fact that an interrupt has occurred), include a "type" field, or less-commonly include one or more data fields. Third, most systems support *direct memory access* (DMA). With DMA, a device can transfer data into or out of a region of memory (e.g., 4Kbytes) without processor intervention.

An example that uses all three types of mechanisms is a disk read. A processor begins a disk read by using memory-mapped stores to inform a disk controller of the source address on disk, the destination address in memory, and the length. The processor then goes on to other work, because a disk access takes millions of instruction opportunities. The disk controller obtains the data from disk and uses DMA to copy it to memory. When the DMA is complete, the disk controller interrupts the processor to inform it that the data is available.

A problem with current I/O architectures is that behavior of disks, network interfaces, frame buffers, I/O buses (e.g., PCI), system interconnects (e.g., PentiumPro bus and SGI Origin 2000 interconnect), and bus bridges (that connect I/O buses and system interconnects) is usually specified in isolation. This tendency to specify things in isolation makes it difficult to take a "systems" view to answer system-level questions, such as:

- What must a programmer to do (if anything) if he or she wants to ensure that two memory-mapped stores to the same device arrive in the same order?
- How does a disk implementor ensure that a DMA is complete so that an interrupt signalling that the data is in memory does not arrive at a processor before the data is in memory?
- How much is the system interconnect or bus bridge designer allowed to reorder transactions to improve performance or reduce cost?

This paper proposes a formal framework, called *Wisconsin I/O* (WIO), that facilitates the specification of systems aspects of an I/O architecture. WIO builds on work on memory consistency models that formally specifies the behavior of loads and stores to normal memory. Lamport's sequential consistency (SC), for example, requires that "the result of any execution is the same as if the operations of all the processors were executed in some sequential order, and the operations of each individual processor appear in this sequence in the order specified by its program

[5].” WIO, however, must deal with several issues not included in most memory consistency models: (a) processor can perform more operations (e.g., memory-mapped stores and incoming interrupts), (b) devices perform operations (e.g., disks doing DMA and sending interrupts), (c) operations can have side effects (e.g., a memory-mapped load popping data or an interrupt invoking a handler), and (d) it may not be a good idea to require that the order among operations issued by the same processor/device (e.g., memory-mapped stores to different devices) always be preserved by the system.

To handle this generality, WIO asks each processor or device to provide a table of ordering requirements. If a processor/device can issue *k* types of operations, the required table is *k* x *k*, where the *i*,*j*-th entry specifies the ordering the system should preserve from an operation of type *i* to an operation of type *j* issued later by that processor/device (e.g., a disk might never need order to be preserved among the multiple memory transactions needed to implement a DMA). A system with *p* processors and *d* devices obeys WIO if there exists a total order of all the operations issued in the system that respects the subset of the program order of each processor and device, as specified in the *p*+*d* tables given as parameters, such that the value of each “read” is equal to the value of the most recent “write” to that address[1].

This paper is organized as follows. In Section 2, we discuss related work. Section 3 presents the model of the system we are studying. Section 4 explains the orderings that are used to specify the I/O architecture, and Section 5 defines Wisconsin I/O consistency based on these orderings. Section 6 describes a system with I/O that is complex enough to illustrate real issues, but simple enough to be presented in a conference paper. In Section 7, we prove that the system described in Section 6 obeys Wisconsin I/O. Finally, Section 8 summarizes our results.

We see this paper as having three contributions. First, we present a formal framework for describing system aspects of I/O architectures. Second, we illustrate that framework in a complete example. Third, we use our verification technique (which uses Lamport’s logical clocks, and which has been applied in previous work[11, 7, 2]) to show that our example implementation meets its specifications.

2 Related Work

The publicly available work that we found related to formally specifying the system behavior of I/O architectures is sparse. As discussed in the introduction, work on memory consistency model is related [1]. Prior to our current understanding of memory consistency models, memory behavior was sometimes specified individually by hardware elements (e.g., processor, cache, interconnect, and memory module). Memory consistency models replaced this disjoint view with a specification of how the system behaves on accesses to main memory. We seek to extend a similar approach to include accesses across I/O bridges and to devices.

Many popular architectures, such as Intel Architecture-32 (x86) and Sun SPARC, appear not to formally specify their I/O behavior (at least not in the public literature). An exception is Compaq Alpha, where Chapter 8 of its specification [9] discusses ordering of accesses across I/O bridges, DMA, interrupts, etc. Specifically, a processor accesses a device by posting information to a “mailbox” at an I/O bridge. The bridge performs the access on the I/O bus. The processor can then poll the bridge to see when the operation completes or to obtain any return value. DMA is modeled with “control” accesses that are completely ordered and “data” accesses that are not ordered. Consistent with Alpha’s relaxed memory consistency model, memory barriers are needed in most cases where software desires ordering (e.g., after receiving an interrupt for a DMA completion and before reading the newly-written memory buffer). We seek to define a more general I/O framework than the specific one Alpha chose and to more formally specify how I/O fits into the partial and total orders of a system’s memory consistency model.

3 System Model

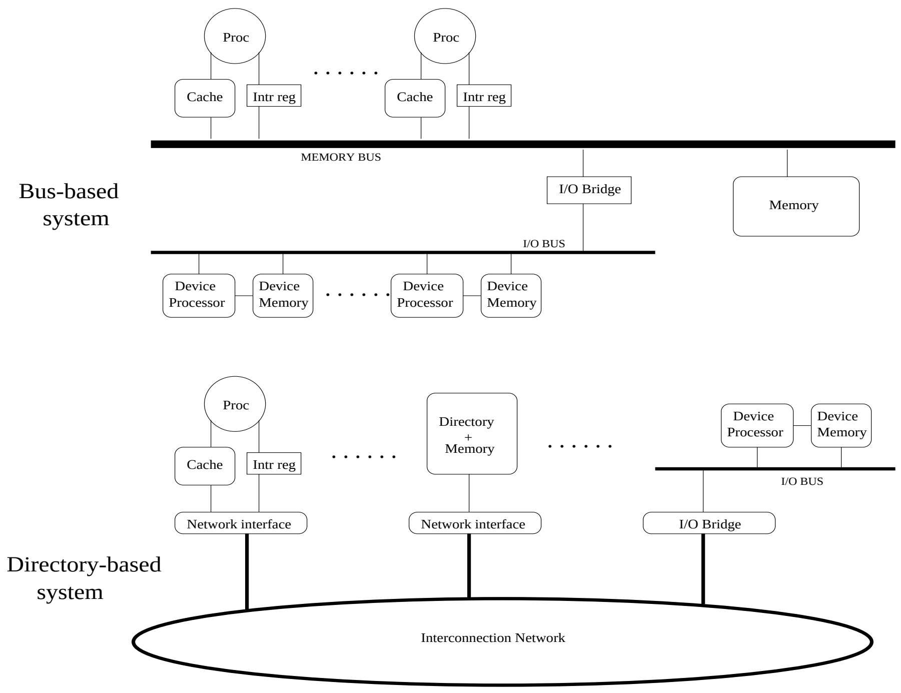
We consider a system consisting of multiple processor nodes, device nodes, and memory nodes that share an interconnect. Figure 1 shows two possible realizations of such a multiprocessor system, where shared memory is implemented using either a broadcast bus or a point-to-point network with directories [3]. The addressable memory

1. The same table can be re-used for homogeneous processors and devices. We precisely define “read” and “write” in later sections.

space is divided into ordinary cacheable memory space and uncacheable I/O space. We now describe each part of the system.

Processor Nodes: A processor node consists of a processor, cache, network interface, and interrupt register. Each processor "issues" a stream of operations, and these operations are listed and described in Table 1. We classify operations based on whether they read data (ReadOP) or write data (WriteOP). If the cache cannot satisfy an operation, it initiates a transaction (these will be described in Section 6) to either obtain the requested data in the necessary state or interact with an I/O device[1]. In addition, the processor (logically) checks its interrupt register, which we consider to be part of the I/O space, before executing each instruction in its program, and it may branch to an

FIGURE 1. System Organizations



interrupt handler depending on the value of the interrupt register.

TABLE 1. Processor Operations

Operation	Class	Description
LD	ReadOP	Load - load word from ordinary memory space
ST	WriteOP	Store - store word to ordinary memory space
LDio	ReadOP	Load I/O - load word from I/O space
STio	WriteOP	Store I/O - store word to I/O space

Device Nodes: We model a device node as a device processor and a device memory. Each device processor can issue operations to its device memory. In addition, it can also issue operations which lead to transactions across the I/O bridge (via the I/O bus). These requests allow a device to read and write blocks of ordinary cacheable memory (via DMA) and to write to a processor node's interrupt register. The list of device operations is shown in Table 2.

1. Note that the cache could also "proactively" issue transactions (e.g., it could prefetch blocks into the cache).

A request from a processor node to a device memory can "cause" the device to "do something useful." For example, a read of a disk controller status register can trigger a disk read to begin. This is modeled by the device processor executing some sort of a program (that specifies the device behavior) which, for example, makes it sit in a loop, checking for external requests to its device memory, and then do certain things (e.g., manipulate physical devices) before possibly an operation to its device memory or to ordinary memory. The device program will usually be hard-coded in the device controller circuits, while the requests from processor nodes will be part of a device driver that is part of the operating system.

TABLE 2. Device Operations

Operation	Class	Description
LDio	ReadOP	Load I/O - load word from device memory (I/O space)
STio	WriteOP	Store I/O - store word to device memory (I/O space)
INT	-	Interrupt - send an interrupt to a processor node
LDblk	ReadOP	Load Block - load cache block from ordinary memory
STblk	WriteOP	Store Block - store cache block to ordinary memory

Memory nodes: Memory nodes contain some portion of the ordinary shared memory space. In a system that uses a directory protocol, they also contain the portion of the directory associated with that memory. Memory nodes respond to requests made by processor nodes and device nodes. Their behavior is defined by the specific coherence protocol used by the system.

Interconnect: The interconnect consists of the network between the processor and memory nodes, and the I/O bridge. This could either be a broadcast bus or a general point-to-point interconnection network. The I/O bridge is responsible for handling traffic between the processor and memory nodes, and the device nodes.

4 Processor and Device Ordering

In a given execution of the system, at each processor or device there is a total ordering of the operations (from the list LD, ST, LDio, STio, INT, LDblk, and STblk) that can be issued by that processor or device. Call this *program order* and denote it by $<_p$.

Let *necessary order* be any relaxation of program order at a processor or a device processor. For example, let $<_n$ be the necessary order that respects program order with respect to operations to the same address and also satisfies the constraints of Tables 3 and 4, where entries in these tables use the following notation:

A: OP1 $<_n$ OP2 always

-: no ordering constraint on OP1, OP2 (if not to the same address)

D: OP1 $<_n$ OP2 if the addresses of OP1 and OP2 refer to the same device

TABLE 3. Necessary Ordering at a Processor

		Operation 2			
		LD	ST	LDio	STio
Operation 1	LD	A	A	A	A
	ST	A	A	A	A
	LDio	A	A	D	D
	STio	-	-	D	D

TABLE 4. Necessary Ordering at a Device Processor

		Operation 2				
		LDio	STio	INT	LDblk	STblk
Operation 1	LDio	D	D	A	A	A
	STio	D	D	A	A	A
	INT	-	-	D	-	-
	LDblk	-	-	A	-	-
	STblk	-	-	A	-	-

The entries in the table reflect the behavior of current systems e.g., in most systems, STios to multiple devices are not guaranteed to be ordered in any particular way. It is important to realize that a programmer who wishes to enforce ordering between operations that are not guaranteed to be ordered can create an ordering through transitivity. For example, a programmer can order a processor's LD after a STio by inserting a LDio to the same device as the STio between the two operations. Since STio $<_n$ LDio and LDio $<_n$ LD, we have STio $<_n$ LD (for this particular sequence of three operations).

5 System Ordering: Wisconsin I/O Consistency

Using the definition of necessary ordering, we can now define a system ordering which we call Wisconsin I/O ordering. The definition of Wisconsin I/O (WIO) ordering takes as a parameter an n-tuple of necessary orderings, such as the 2-tuple specified by Tables 3 and . Let $<_W$ be a total ordering of all LD, ST, LDio, STio, INT, LDblk, STblk operations of an execution of the system. Then $<_W$ satisfies Wisconsin I/O ordering with respect to a given necessary ordering if:

1. $<_W$ respects the necessary ordering, and

2. the value read by every ReadOP operation is the value stored by the most recent WriteOP operation to the same address in the $<_W$ order.[1]

In the next two sections, we will present a system protocol specification and show that it obeys WIO.

6 A System Protocol Specification

In this section, we describe a protocol for a directory-based system consisting of the components described in Section 3. This description builds upon the directory protocol described in Plakal et al. [7]. The description is divided into descriptions of the processor nodes, interconnect, I/O devices, bridge and memory nodes.

<u>Processor nodes:</u> The cache receives a stream of LD/ST/LDio/STio operations from the processor and, if it cannot satisfy a request, it issues a transaction[2]. The complete list of transactions, including block transfer transactions (Rblk/Wblk) that can only be issued by devices and which will be discussed later, are shown in Table 5. Cache coherence transactions (GETX/GETS/UPG/WB) are directed to the home of the memory block in question (i.e., the memory node which contains the directory information for that block). I/O transactions (Rio/Wio) are directed to a specific I/O device and also contain an address of a location within the memory of the device (and, if Wio, the data to write as well). The granularity of access for an I/O transaction is one word. Wios do not generate any reply

1. This definition is in the spirit of SC, but there is an analogous definition for other consistency models where a ReadOP does not necessarily have to get the value of the most recent WriteOP (e.g., SPARC TSO).

2. As noted earlier, caches can also proactively issue transactions without receiving an operation from their processors.

messages from the target device, while Rios generate a reply message from which the cache extracts a register value and passes it to the processor.

TABLE 5. Transactions

Transaction	Description
GETX	Get Exclusive access
GETS	Get Shared access
UPG	Upgrade (Shared to Exclusive) access
WB	Write Back
Rio	Read I/O - read word from I/O space
Wio	Write I/O - write word to I/O space
Rblk	Read Block - read cache block from ordinary memory
Wblk	Write Block - write cache block to ordinary memory

Processor nodes must conform to the list of behavior requirements specified in Section 2.4 of Plakal et al. [7] (e.g., a processor node maintains at most one outstanding request for each block). They must also conform to the ordering restrictions laid out in Table 3. For example, they cannot issue a LD/ST until all LDios preceding it in program order have been "performed" (i.e., the reply has been written into the register by the cache).

A processor node's network interface sends all transactions from the cache into the interconnection network. In addition, the network interface will pass a Wio coming from the network to the processor's interrupt register. It also will pass all replies to transactions to the cache.

Interconnect: The network ensures point-to-point order between a processor node and a device node, and it ensures reliable and eventual delivery of all messages.

Bridge: The I/O bridge performs the following functions: it receives Rio/Wios from processor nodes and broadcasts them on the I/O Bus (this has to be done in order of receipt on a per-device basis); sends Wio replies from device memory to processor nodes; sends Wios (to interrupt registers) from device processors to processor nodes; participates in Rblk/Wblk transactions (discussed below) and broadcasts completion acknowledgments on the I/O bus. The I/O bridge must obey certain rules. It provides sufficient buffering such that it does not have to deny (negative acknowledgment or NACK) requests sent by processors or devices. It also handles the re-try of its own NACKed requests (to memory nodes). No order is observed in the issue/overlap of Rblk/Wblk transactions.

Device Nodes: Each device processor can issue LDio/STios to its device memory and INTs to processor interrupt registers. INT operations are converted to Wio transactions by the I/O bridge. These are directed to a specific processor's interrupt register and do not generate reply messages. In addition, a device can also issue LDblk and STblk requests, and these operations are converted to Rblk and Wblk transactions by the bridge and are directed to the home node. The data payload for both requests is a processor cache line (equal to a block of memory at a memory node, which is equal to the coherence unit for the entire system). Both requests generate acknowledgments (ACKs) on the I/O bus (from the bridge) and, in the case of the Rblk, the ACK contains the data as well. A Wblk request carries the data with it.

Each device memory receives a stream of LDio/STios from its device processor. In addition, it also receives a stream of Rio/Wios from the bridge (via the I/O bus) which it logically treats as LDio/STios. These two streams are interleaved arbitrarily by the device memory. For each incoming Rio, the device memory sends (via the bus and the bridge) the value of that location back to the node that sent the Rio. LDio/STio operate on device memory like a processor's LD/ST operate on its cache.

The device processor must obey the ordering rules specified in Table . For example, an INT is not issued until all LDblk/STblks preceding it in "device program order" have been performed (i.e., an ACK has been received from the bridge for the corresponding Rblk/Wblk).

Memory Nodes: Memory nodes operate as described in Plakal et al.[7] (with respect to directory state and transactions) with the following modifications for handling Rblk/Wblk transactions. Protocol actions depend on the state of the block at the home node for both transactions.

Rblk:

- *Idle* or *Shared*: the home sends the block to the bridge, which broadcasts an ACK with the data on the I/O bus.
- *Exclusive*: the home changes state to *Busy-Rblk*, removes the current owner's ID from CACHED, and forwards the request to the current owner. The owner sends the block to the bridge, invalidates the block in its cache, and sends an update message (with the block) to the home, which changes the state to *Idle* and writes the block to memory. The bridge receives the block and broadcasts an ACK along with the data on the I/O bus.
- *Busy-Any*: the home NACKs the request.

Wblk:

- *Idle*: the home stores the block to memory and sends an ACK to the bridge. The bridge sends an ACK to the device (via broadcast on the I/O Bus).
- *Shared*: the home stores the block to memory, sends invalidations to all shared copies, sends a count of the copies to the bridge and changes the state to *Busy-Wblk*. The bridge waits until it receives all ACKs for the invalidations before broadcasting the transaction completion ACK on the I/O Bus. The bridge also then sends an ack to the home which enables it to change its state to *Idle*.
- *Exclusive*: the home stores the block to memory, sends an invalidation to the (previous) owner, sends an ACK to the bridge, and changes the state to *Busy-Wblk*. The former owner invalidates its copy and sends an ack to the bridge, which then sends an ACK to the device and to the home (which then changes its state to *Idle*).
- *Busy-Any*: the home NACKs the request.

Note that we now have two new "busy" home states, *Busy-Rblk* and *Busy-Wblk*, which serve similar roles as the busy states used in the original directory protocol. These modifications make some formerly impossible situations possible. In particular, *Writeback* requests may find the home busy. One solution is to modify the transaction cases:

- *Writeback* on home *Busy-Rblk* or *Busy-Wblk*: This is the same as when the home is *Busy-Shared*.

7 Proof that an Implementation Satisfies WIO

In this section, we will demonstrate that the implementation described in the previous section satisfies the definition of WIO. We will use a verification technique based on Lamport's logical clocks that we have successfully applied to systems without I/O [11, 7, 2]. The technique relies on being able to assign timestamps to operations in a system and then proving that the ordering induced by the timestamps has the properties required of the implementation. Section 7.1 provides background to our verification technique, Section 7.2 describes the timestamping scheme for our implementation and Section 7.3 provides the proof of correctness of the implementation. Both the timestamping scheme and proof are intended to be modest extensions of those presented in our previous work [7].

7.1 Background to Lamport Clocks[1]

Our previous work on using Lamport Clocks to verify shared-memory multiprocessor systems[7,11] proved that implementations (without I/O) using a SGI Origin 2000-like [6,3] directory protocol and a Sun Gigaplane-like [8] split-transaction bus protocol both implement SC. Both implementations use three-state invalidation-based coherence protocols. We have also extended this research to use Lamport clocks to prove that systems obey two relaxed memory consistency models, SPARC TSO and Compaq Alpha [2].

1. This summary is similar to the summary we present in Section 2.1 of Condon et al. [2].

Our reasoning method associates logical timestamps with loads, stores, and coherence events. We call our method *Lamport Clocks*, because our timestamping modestly extends the logical timestamps Lamport developed for distributed systems [4]. Lamport associated a counter with each host. The counter is incremented on local events and its value is used to timestamp outgoing messages. On message receipt, a host sets its counter to one greater than the maximum of its former time and the timestamp of the incoming message. Timestamp ties are broken with host ID. In this manner, Lamport creates a total order using these logical timestamps where causality flows with increasing logical time.

Our timestamping scheme extends Lamport's 2-tuple timestamps to three-tuples: <**global . local . node-id>**, where **global** takes precedence over **local,** and **local** takes precedence over **node-id** (e.g., 3.10.11 < 4.2.1). Coherence messages, or transactions, carry global timestamps. In addition, global timestamps order LD and ST operations relative to transactions. Local timestamps are assigned to LD and ST operations in order to preserve program order in Lamport time among operations that have the same global timestamp. They enable an unbounded number of LD/ST operations between transactions. Node-ID, the third component of a Lamport timestamp, is used as an arbitrary tiebreaker between two operations with the same global and local timestamps, thus ensuring that all LD and ST operations are totally ordered.

7.2 Timestamping Scheme for Our Implementation

Before we present the timestamping scheme, we would like to define some concepts and make some changes which will make the timestamping and the proof simpler to express and understand.

First, we split up Rblk and Wblk transactions into two steps: RBlk-Start/End and WBlk-Start/End, respectively. The reasoning behind this is as follows: cache coherence transactions (e.g., a GETX) will bring a block into a processor cache where it can be accessed until it is removed via another transaction (e.g., a WB or an incoming invalidation generated by another GETX). On the other hand, RBlk/Wblk transactions access a cache block but they do not give the device permission to do more than one operation (LDblk/STblk). It is as if the LDblk/STblk was immediately followed by a transaction that removed the device's access to the block. Breaking RBlk and WBlk into Start and End transactions unifies cache coherence and DMA transactions into one framework and simplifies the timestamping and the proof. This was not done earlier (in Section 6) to avoid confusing the reader with extra detail. The changes to the protocol are minimal: every RBlk/WBlk transaction is now regarded as a RBlk/WBlk-Start transaction. After such a transaction succeeds, a device node is now capable of performing a LDblk/STblk operation. The Rblk-End/Wblk-End is considered to occur when the transaction is complete.

Consistent with our previous work [7], we introduce the notion of a per-block *A-state* (address-state) at a node to describe the home node's view of that node's access to that block of memory. The A-state can be one of A_I (*Idle*), A_S (*Shared*), or A_X (*Exclusive*). The A-state of a block at a node changes as it participates in transactions for that node (either initiated by it or forwarded to it by the home). The A-state is set to A_I when the node receives an invalidation or a forwarded *Get-Exclusive*, or an acknowledgment for its own *Writeback* request. The A-state is set to A_S when the node receives a downgrade, or a response to its own *Get-Shared* request. Finally, the A-state is set to A_X when the node receives a response to its own *Upgrade* or *Get-Exclusive* request, along with all associated invalidation acknowledgments. In addition, we now define the A-state of a device node for a block B of memory to change to A_S or A_X when it performs a RBlk-Start or WBlk-Start, and that it change to A_I on a RBlk-End or WBlk-End. Similarly, after a RBlk/WBlk-Start transaction, the home node's A-state will change to A_I or A_S according as the final home state for that block is *Idle* or *Shared* respectively. After a RBlk/WBlk-End transaction, the home node's A-state will change to A_X if the final home state for that block (after the corresponding RBlk/WBlk-Start) was *Idle*.

We assign timestamps to the operations and transactions defined in Tables 1, 2 and (with RBlk and WBlk split up as described above). The rules listed in Tables 6, and 8 below indicate the assignment of the global and local components of the timestamp for each kind of operation/transaction. Note that transactions do not need a local timestamp and could be assigned some arbitrary local timestamp (e.g., zero so that a transaction gets ordered before operations with the same global timestamp).

Conceptually, each node (processor/memory/device) maintains a global and local clock which get updated in real time for operations and transactions. To do this in a well-defined manner, we define a *timestamping order* which is a per-node total order which decides the order in which operations and timestamps get assigned timestamps. Operations enter the timestamping order of a node at the point in real time when they are retired (i.e., they cannot be undone due to mis-speculation handling), and operations are retired in a real time order that is consistent with program order. If more than one operation is committed at the same point in real time, they can be ordered arbitrarily in the timestamping order. Transactions enter the timestamping order of a node at the point in real time when the corresponding A-state changes occur at that node[1].

The timestamping rules given below also determine the maintenance of the per-processor clocks in that a node updates its global and local clocks to equal the corresponding timestamp of each operation/transaction it timestamps in timestamping order. Any increase in the global clock value causes the local clock to be reset to zero before it is updated as specified by the rule. There are a few cases where a transaction originating at a node is timestamped elsewhere (e.g., the Wio at a device corresponding to an INT). The assignment of this timestamp causes the local node's global clock to get incremented (if necessary). For purposes of timestamping, we consider a bridge to be part of each device node, and all transactions in which a bridge participates on behalf of a device node will update the clocks of that device node.

Processor nodes: Let P-UP be a transaction that causes an increase in coherence permissions (upgrade) at processor node p_i (GETX, GETS, or UPG by p_i), and let P-DOWN be a transaction that causes a decrease in coherence permissions (downgrade) at p_i (WB by p_i, GETX by p_j for a block that p_i has Shared or Exclusive, UPG by p_j for a block that p_i has Shared, GETS by p_j for a block that p_i has Exclusive, or Rblk/Wblk by a device for a block that p_i has Shared or Exclusive). Then the processor node timestamping rules are as shown in Table 6.

TABLE 6. Processor node timestamping

Operation/ Transaction	**Global Timestamp**	**Local Timestamp**	**Node ID**
LD, ST	current global clock	1 + current local clock	processor
LDio	global timestamp of corresponding Rio (sent)	1	device
STio	global timestamp of corresponding Wio (sent)	1	device
P-UP	1 + max {global clock, timestamps assigned to P-UP by all other nodes that downgrade as a result of P-UP}	0	processor
P-DOWN	1 + global clock	0	processor
Rio (sent)	*only timestamped at device*		
Wio (sent)	*only timestamped at device*		
Wio (recv)	1 + max {global clock, global timestamp of device when Wio was sent}	0[a]	processor

a. Timestamp is 0, but the clock is set to 1. This ensures that LDio/STios issued by a processor get a local timestamp of 1, while those issued by a device get a local timestamp of 2 or greater.

Memory nodes: Let M-UP be a transaction that causes an increase in permissions at memory node m_i (WB by p_i), and let M-DOWN be a transaction that causes a decrease in permissions at m_i (GETS, GETX, or UPG by p_i). With these definitions of M-UP and M-DOWN, the timestamping rules for memory nodes are as shown in Table 7. The

1. There is the exceptional case of *Get-Shared* transactions at the home for a *Shared* block. In this case, we consider the timestamp to be assigned at the point that the home sends the block to the requester, i.e., when the A-state "changes" from A_S to A_S.

memory node timestamps transactions in the real-time order in which they are processed. In the case of transactions that involve transient Busy states, the “current global clock” corresponds to the global clock at the time the Busy state is entered, while the timestamp of the transaction is assigned when the memory enters a non-transient state (*Idle*, *Shared*, *Exclusive*).

TABLE 7. Memory node timestamping

Transaction	Global Timestamp
M-UP	1 + max {current global clock, timestamps assigned to M-UP by the nodes that downgrade as a result of M-UP}
M-DOWN	1 + current global clock
Rblk-Start	1 + max {current global clock, global timestamp of device when Rblk-Start was sent, global timestamp assigned to Rblk-Start by Exclusive node that downgrades as a result of Rblk-Start (if any)}
RBlk-End	1 + current global clock
Wblk-Start	1 + max{current global clock, global timestamp of device when Wblk-Start was sent, global timestamp assigned to Wblk-Start by all nodes that downgrade as a result of Wblk-Start (if any)}
WBlk-End	1 + current global clock

<u>Device nodes:</u> A device node timestamps operations and transactions as shown in Table 8.

TABLE 8. Device node timestamping

Operation/ Transaction	Global Timestamp	Local Timestamp	Node ID
LDio, STio	current global clock	1 + current local clock	device
INT	global timestamp of corresponding Wio	1	processor
LDblk	global timestamp of corresponding Rblk-Start	1	memory
STblk	global timestamp of corresponding Wblk-Start	1	memory
Rio (recv)	1 + max{global clock, global timestamp of sender when Rio was sent}	0[a]	device
Wio (recv)	1 + max {global clock, global timestamp of sender when Wio was sent}.	0[a]	device
Wio (sent)	*only timestamped at processor*		
Rblk-Start	*only timestamped at memory*		
RBlk-End	*only timestamped at memory*		
Wblk-Start	*only timestamped at memory*		
WBlk-End	*only timestamped at memory*		

a. See Footnote a under Table 6.

7.3 Proof of Correctness of Our Implementation

To prove WIO, it is sufficient to show that there is a total order of operations such that the orderings in Tables 3 and 4 are respected and such that every Read-OP gets the value of the most recent Write-OP. The timestamping scheme ensures the total order and, combined with the protocol specification, ensures that Tables 3 and 4 are respected. LDios and STios to device memory are ordered at the device in the order in which they are performed, so a LDio must get the value of the most recent STio. Now we will prove that every LD/LDblk gets the value of the most recent ST/STblk.

The proof that we provide here is very similar in structure to the proof that we provided in our previous work [7]. In what follows, we first outline how definitions from our previous work can be extended to the implementation presented in this paper. We then summarize the claims and lemmas that are used in the main theorem. The changes in the statements of these results (relative to our previous work in SPAA'98 [7]) are emphasized in underlined bold.

The consistency model is established using the concept of *coherence epochs*. An epoch is an interval of logical time during which a node has read-only or read-write access to a block of data. In the rest of the paper, we assume a *block* to be a fixed-size, contiguous, aligned section of memory (usually equal to the cache line size). Also, LDs and STs operate on *words*, where we assume that a word is contained in a block and is aligned at a word boundary. Our scheme could be extended to handle LDs and STs on sub-units of a word (half-words or bytes) which need not be aligned. However, this makes the specification of the memory models very tedious without any gain in insight or clarity.

Transactions on a given block are serialized by the block's directory. Hence, we can speak about a sequence of transactions on the same block where the ordering is implied by their serialization at the directory. For each node N, a sequence of t transactions on block B (where the order among transactions is seen at the Home) defines a unique sequence $A_{(1)}$, $A_{(2)}$,..., $A_{(t)}$ of associated A-states for N, given some initial A-state value at N. If $A_{(i)}$ is not equal to $A_{(i-1)}$ for some $i \geq 1$, we say that the i^{th} transaction in the sequence "affects" N and that the transaction "implies that N's A-state for block B change from $A_{(i-1)}$ to $A_{(i)}$". For example, consider a single block of memory and three nodes: N_1 (processor), N_2 (device) and N_3 (memory). Suppose that both N_1 and N_2 start out with an initial A-state of A_I and N_3 starts with A_X. Let the sequence of transactions at N_3 be N_1's *Get-Exclusive*, N_2's *RBlk-Start* and N_2's *RBlk-End*. Then the sequence of A-states for N_1, N_2 and N_3 are A_I, A_X, A_I, A_I; A_I, A_I, A_X, A_I and A_X, A_I, A_I, A_X respectively. The *Get-Exclusive* affects N_1 and N_3, while the *RBlk-Start/End* affect N_2 and N_3. In the special case that a node is the directory, we say that it is also affected by all transactions resulting from *Get-Shared* requests, even though no change in the A-state at the directory may be implied by such a transaction.

Each transaction implies an "upgrade" of A-state (i.e. change from state A_I to A_S, from A_I to A_X, or from A_S to A_X) at exactly one node. For example, a RDblk-Start causes an upgrade at the device, a downgrade at memory, and possibly a downgrade at a processor. Also, each transaction implies a "downgrade" of A-state (i.e. change from A_X to A_S, from A_X to A_I, or from A_S to A_I) at zero or more nodes. In the special case that node N is the directory, we say that N's A-state "downgrades" as a result of every *Get-Shared* transaction, even though its A-state may not be changed by the transaction. On each transaction, exactly one node upgrades and zero or more nodes downgrade.

The definitions of "affects" and "implies" in the previous two paragraphs depend only on the sequence of transactions on block B at B's directory. In Claim 2 below, we show that the protocol specification ensures that, at every node, the actual sequence of changes to the A-state for block B occurs in the order implied by the serialization of the transactions at B's directory, even though messages on successive transactions may be received out of order by a node.

Claim 1: For each transaction T, a message is sent to every processor affected by T. Also, if **processor** N upgrades as a result of T, exactly those nodes that are affected by transaction T (other than N) send a message to N.

Claim 2: The sequence of A-state changes on block B at a node occurs in real time in the order implied by the serialization of transactions on block B at its directory.

Claim 3: For a transaction T on block B,

(a) The timestamps of the downgrades associated with T are less than or equal to the timestamp of the upgrade associated with T.

(b) The timestamp of the upgrade associated with T is less than the timestamp of the upgrade associated with any transaction T' on block B occurring after T in the serialization order at the directory, so long as one of T or T' is a Get-Exclusive or Writeback **or WBlk-Start**.

Claim 4: Every LD/ST **or LDblk/STblk** operation on block B at processor p_i is bound[1] to the most recent (in Lamport time at p_i) transaction on block B that affects p_i.

By construction, the Lamport ordering of LDs and STs within any processor is consistent with program order. Therefore, to prove sequential consistency, it is sufficient to show that the value of every load equals the value of the most recent store.

Recall that a coherence epoch is simply a Lamport time interval [t1,t2) during which a node has access to a block. All operations that have global timestamp t where $t_1 \leq t < t_2$ are contained in epoch $[t_1,t_2)$. A shared or exclusive epoch for block B at node N starts at time t_1 if a transaction with timestamp t_1 (at N) implies that N's A-state for block B changes to A_S or A_X respectively. The epoch ends at time t_2, where t_2 is N's timestamp of the next transaction on B that implies a change in A-state at N.

Lemma 1 shows that two processors cannot have "conflicting" permission to the same block at the same (Lamport) time. Lemma 2 states that processors do operations within appropriate epochs. Finally, Lemma 3 shows that the "correct" block value is passed among processors and the directory between epochs.

Lemma 1: Exclusive epochs for block B do not overlap with other exclusive or shared epochs for block B in Lamport time.

Lemma 2:

(a) Every LD/ST, **LDblk/STblk** operation on block B at p_i is contained in some epoch for block B at p_i and is bound to the transaction that caused that epoch to start.

(b) Furthermore, every ST **or STblk** operation on block B at p_i is contained in some exclusive epoch for block B at p_i and is bound to the transaction that caused that epoch to start.

Lemma 3: If block B is received by node N at the start of epoch $[t_1,t_2)$, then each word w of block B equals the most recent ST **or STblk** to word w prior to t_1 or the initial value in the directory, if there is no store to word w prior to global time t_1.

The proof of the Main Theorem shows how WIO follows from the lemmas.

Main Theorem: The value of every **LD or LDblk** equals the value of the most recent ST **or STblk** or the initial value, if there has been no prior store.

8 Conclusions

Although I/O devices are integral parts of computer systems and having clean I/O architectures would offer benefits, the commercial systems we are familiar with tend to use ad hoc, complex, and undocumented interfaces. In this paper, we have proposed a framework called Wisconsin I/O for formally describing I/O architectures. WIO is an extension of research on memory consistency models that incorporates memory-mapped I/O, interrupts, and device operations that cause side effects. WIO is defined through ordering requirements at each processor and device, and a system is considered to obey WIO if there exists a total order of all operations that satisfies these ordering requirements such that the value of every read is equal to the value of the most recent write. We showed how to use Lamport clocks to prove that an example system that we specified satisfies its WIO specification.

The framework presented here for specifying and analyzing systems with I/O can be generalized in several ways that were not presented earlier in order to simplify the discussion. First, it can be extended to handle systems with more components than the system described in Section 6. Systems may have multiple I/O devices and possibly even multiple I/O bridges, and the framework we have developed here can be extended to these systems. Second, unlike in Section 6, we can model I/O bridges that do not have enough buffering to ensure that they can sink all incoming requests. Third, the definition of Wisconsin I/O consistency is parameterized by a n-tuple of necessary

1. In our previous work [7], we had defined the notion of LDs/STs being *bound* to the coherence transaction that brought the corresponding block into the cache. Similarly, we can think of LDblk/STblk operations being bound to their corresponding RBlk/WBlk-Start transactions.

orderings and is therefore easily generalized to handle an arbitrary set of local ordering rules. In the extreme case, each processor and each device would have its own table of necessary ordering rules.

References

[1] Sarita V. Adve and Kourosh Gharachorloo. Shared Memory Consistency Models: A Tutorial. *IEEE Computer*, pages 66–76, December 1996.

[2] Anne E. Condon, Mark D. Hill, Manoj Plakal, and Daniel J. Sorin. Using Lamport Clocks to Reason About Relaxed Memory Models. In *Proceedings of High Performance Computer Architecture*, January 1999.

[3] David Culler, Jaswinder Pal Singh, and Anoop Gupta. *Parallel Computer Architecture: A Hardware/Software Approach*. Morgan Kaufmann, 1998.

[4] Leslie Lamport. Time, Clocks and the Ordering of Events in a Distributed System. *Communications of the ACM*, 21(7):558–565, July 1978.

[5] Leslie Lamport. How to Make a Multiprocessor Computer that Correctly Executes Multiprocess Programs. *IEEE Transactions on Computers*, C-28(9):241–248, September 1979.

[6] James P. Laudon and Daniel Lenoski. The SGI Origin: A ccNUMA Highly Scalable Server. In *Proceedings of the 24th International Symposium on Computer Architecture*, Denver, CO, June 1997.

[7] Manoj Plakal, Daniel J. Sorin, Anne E. Condon, and Mark D. Hill. Lamport Clocks: Verifying a Directory Cache-Coherence Protocol. In *Proceedings of the 10th Annual ACM Symposium on Parallel Architectures and Algorithms*, Puerto Vallarta, Mexico, June 28–July 2 1998.

[8] A. Singhal, D. Broniarczyk, F. Cerauskis, J. Price, L. Yuan, C. Cheng, D. Doblar, S. Fosth, N. Agarwal, K. Harvey, E. Hagersten, and B. Liencres. Gigaplane: A High Performance Bus for Large SMPs. *Hot Interconnects IV*, pages 41–52, 1996.

[9] Richard L. Sites, editor. *Alpha Architecture Reference Manual*. Digital Press, 1992.

[10] Mark Smotherman. A Sequencing-Based Taxonomy of I/O Systems and Review of Historical Machines. *Computer Architecture News*, 17(5):10–15, September 1989. See also URL http://www.cs.clemson.edu/~mark/io.ps.

[11] Daniel J. Sorin, Manoj Plakal, Mark D. Hill, and Anne E. Condon. Lamport Clocks: Reasoning About Shared-Memory Correctness. Technical Report CS-TR-1367, University of Wisconsin-Madison, March 1998.