PC-ATOMIC

Joseph D. Touch[1]
USC / Information Sciences Institute
touch@isi.edu

ABSTRACT: PC-ATOMIC is a PC interface for the ATOMIC LAN. PC-ATOMIC is implemented as a VL-Bus (VESA) short-form card for Intel i486 PCs, providing an interface for low-cost workstations to a 640 Mbps LAN. This document describes the PC-ATOMIC interface, its design, capabilities, and performance. The board design is public, and a small number of boards are available as government-furnished equipment for research projects.

1.0 Introduction

PC-ATOMIC is an Intel i486 PC VL-Bus host interface for the ATOMIC LAN [5] [6]. It provides an inexpensive interface for inexpensive hosts to a high-speed (640 Mbps) LAN. It provides a read bandwidth of 85 Mbps, and a write bandwidth of 135 Mbps, both using programmed I/O. The board architecture can support DMA if reprogrammed at the PLD-level. It also provides a zero-overhead, maskable IP-checksum in hardware. This board can be used point-to-point, or together with Myricom's Myrinet switches and SPARC S-Bus host interfaces. PC-ATOMIC is compatible at the hardware link-layer with Myricom equipment.

This document describes the design and use of the PC-ATOMIC board. It also reviews its capabilities, and reports some preliminary performance measurements. The board design is public, and small numbers of the board are available as government-furnished equipment for research purposes.

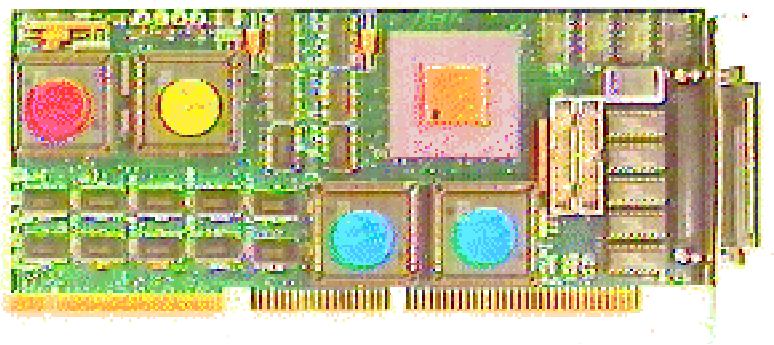

FIGURE 1. PC-ATOMIC board photo

1.1 Background

The following is a brief history of ATOMIC and Myricom, and their relationship to PC-ATOMIC.

1. This research was sponsored by the Advanced Research Projects Agency through Ft. Huachuca Contract No. DABT63-91-C-0001, entitled "Gigabit Network Communication Research". The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Army, the Defense Advanced Research Projects Agency, or the U.S. Government.

ATOMIC is a source-routed, cut-through packet-switched LAN based on the components of the CalTech Mosaic supercomputer [4]. The Mosaic supercomputer uses custom VLSI 16-bit processors with integrated 2-dimensional communication channels, arranged in a mesh. The Mosaic channels are 8-bits wide, at 100 Mhz, for a channel rate of 800 Mbps. Packet routing is performed in simple on-chip hardware. ISI developed ATOMIC to use Mosaic chips to implement an inexpensive high-speed LAN. The preliminary ATOMIC LAN components were designed by CalTech, and programmed by ISI.

In 1994, members of ISI's ATOMIC group and CalTech's Mosaic group left their respective organizations to form Myricom to develop the ATOMIC LAN as a commercial product called Myrinet [9]. Myrinet components are not compatible with the prototype ATOMIC LAN, but are based on the same design principles. Myrinet is 8 bits wide at 80 Mbps per line, for a channel rate of 640 Mbps. Myricom currently produces host interfaces for Sun SPARC workstations, cables, and switches. PC-ATOMIC is compatible with Myrinet hardware, and can be programmed for use in a production Myricom LAN. Myrinet is based on the LANai chip, a descendant of the Mosaic chip.

The PC-ATOMIC network interfaces are designed to be compatible with the emerging Myricom hardware. In addition, the PC-ATOMIC board design examines some general host-interface design issues, such as zero-cost IP checksum hardware, DMA and board control by both on- and off-board processors, and interrupt signalling. The result is a host interface whose programmed I/O rate exceeds that of comparable prototypes from Myricom.

2.0 Description

PC-ATOMIC is a VL-bus host interface based on the Myricom LANai 1.2 interface processor. The LANai 1.2 has no DMA or IP checksumming capabilities. The PC-ATOMIC board is based on the following design criteria:

- *Use available and commercial off-the-shelf (COTS) parts wherever possible*
- *Make the hardware general and reconfigurable*
- *Make the hardware fast*

PC-ATOMIC uses programmable logic devices (PLDs), rather than custom VLSI, to reduce the design time and provide a more flexible design. PLDs were chosen because they are easier to program than arbitrary programmable logic arrays (PLAs, e.g., Xylinx). The LANai 1.2 processor was used because the LANai 2.3 was not available at the time of this design.

General and fast hardware requires Direct Memory Access (DMA) transfer capability. The PC-ATOMIC card is DMA-capable, although time constraints did not permit complete testing of the current PLD programming. DMA is possible via firmware upgrades. In addition, the board has the capability to allow the LANai to access the DMA and board-level registers in a future firmware release. All board-level interrupts are maskable, and some can be triggered under register control.

There is a zero-overhead IP checksum, that is capable of 1.2 Gbps. The LANai 1.2 supports clock speeds up to 20 Mhz, but in the PC-ATOMIC card it is clocked at 1/2 the VL-Bus frequency, i.e., at 17.5 Mhz. Using a half-rate VL-Bus clock simplifies the design significantly.

2.1 Board-level hardware

The PC-ATOMIC board is memory-mapped, making control and OS integration simple. Board control is managed by a set of board-level registers and interface processor registers, as well as dual-access shared RAM [7].

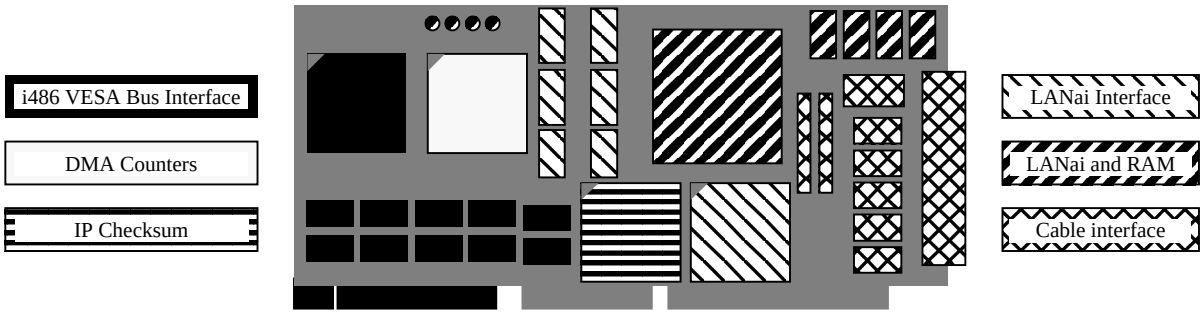

FIGURE 2. PC-ATOMIC board component layout

The board has a set of configuration jumpers, one for setting the board's hardware base address, and the other for setting the Myricom link interface [7]. The link interface jumpers are specified in the Myricom literature, and come pre-configured by ISI. The address jumpers specify a block of 2^24 bytes (16 Mbytes), in the 00xx xxxx - 0Fxx xxxx range (i.e., the lower nibble of the high byte of the 32-bit address). The board is preconfigured to a base address of 0A00 0000.

The board consists of 128 K bytes of dual-access RAM, configured as 32K of 32-bit words, and a set of eight 32-bit board-level registers that repeats in the next 128 K byte range. The overall space of 256 K bytes repeats throughout the 16 M byte block. The host and on-board interface processor share access to this RAM on alternate on-board clock cycles, emulating dual-ported RAM. This RAM is accessed through the on-board processor, which maps the top 64 bytes of the space as a shadow of its internal processor registers, and the bottom 8 K bytes are write-accessible only off-board. The base of RAM is the start location for the on-board processor following a reset.

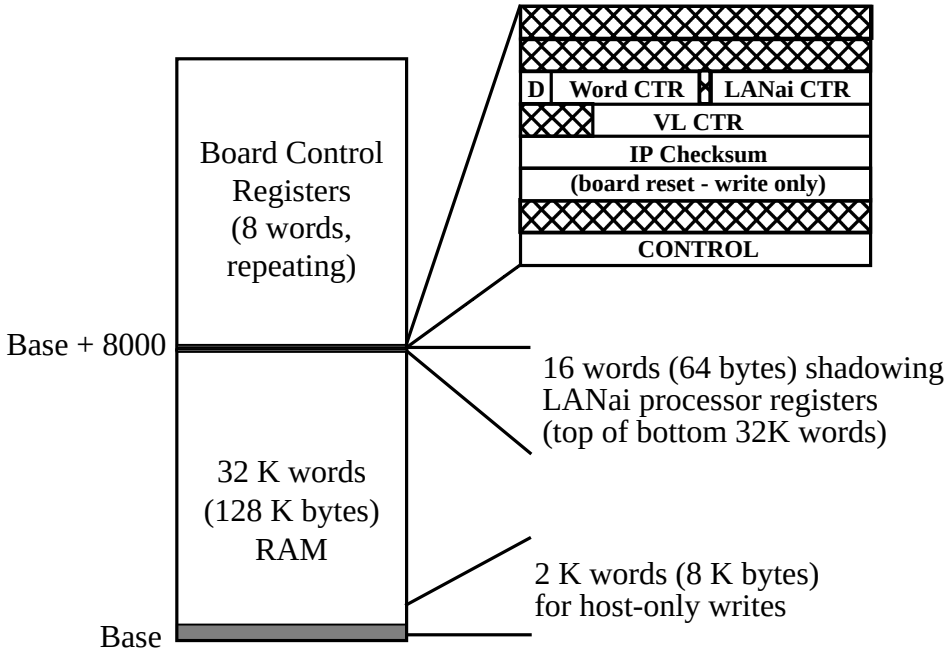

FIGURE 3. Board memory layout

2.1.1 Board-level configuration

The board contains a LANai 1.2 subsystem, nearly identical to that on the prototype Myricom 1.2 Sun SPARC SBus host interfaces [7]. The LANai uses a separate on-board crystal to control the link clocking, which is set to 40 Mhz for compatibility with the current Myrinet switches.

There are 4 programmable LEDs, connected directly to the LANai processor, as documented in the Myricom literature. The board also contains 4 AMD MACH 435 programmable PLDs. These are socketed for firmware upgrades.

The board also has two sets of jumpers. One set of three, located near the link crystal, is used to specify the link clock offset, as specified in the Myricom literature. The other set of jumpers are used to specify the base address of the board. They indicate the values of bits 27 through 24, i.e., the low nibble of the high byte of the board base address.

This board also has loopback capability. The loopback can be performed after the cable drivers, or before. When used after the cable drivers, a standard Myricom DB-37 D-connector loopback is used on the Myricom cable interface. Use of pre-driver loopback requires removal of the AT&T 41MM cable driver chips, and installation of a half-twisted ribbon connecting two 26-pin headers in the pre-driver header sockets.

2.1.2 Board registers

There are five main board registers which provide for Internet checksum, board-level reset, interface-processor reset-and-hold/release, maskable interrupts, and DMA control (DMA can be supported in future firmware releases) [8].

The control register provides general board-level management. Bits in the register can be used to:

- *Suspend the DMA engine*
- *Enable/disable IP checksumming (on both PI/O and DMA)*
- *Reset/release the LANai reset*
- *Mask and set interrupts*

The DMA engine can be suspended via a single bit in the mask. This bit is usually set before the DMA registers are loaded, and released to indicate that DMA can commence. The IP checksum can similarly be suspended via a bit. When enabled, the checksum automatically sums every RAM data access, whether DMA or programmed I/O, and whether read or written. The LANai processor has a reset bit, so that the operation of the processor can be suspended while the RAM is being loaded with the LANai control program. This bit is also automatically set when the entire board is reset. There are also bits to set and mask various interrupts to the host and LANai.

The IP checksum register maintains a partial Internet checksum [1]. All data accesses are incrementally summed into this register when checksumming is enabled (as per a bit in the control register). This includes data reads and writes, and both programmed I/O and DMA are included. The checksum value is maintained as a pair of 16-bit ones-complement sums, one each for high and low half-words. The sum can be read at any time, and its halves folded to yield the actual IP checksum. The register is cleared by writing any value to it.

The two DMA registers are used to initiate DMA transfers to and from the host. One register contains a bit indicating the direction of the data transfer, the LANai-side transfer base address, and the number of 32-bit words to be transferred. The other register indicates the host-side base address. DMA operation has not been fully tested at this time. DMA capability has been disabled in the PLDs on the board, as distributed.

The entire board can be reset by writing to a phantom “reset” register. A reset re-initializes the board-level registers, and places the LANai processor in stasis to allow for the host to load the LANai control program into the board RAM.

2.2 Software

The PC-ATOMIC board is distributed with limited test software, and “include” files. The software was implemented and compiled for NetBSD 0.9 on an i486 VL-Bus platform, and requires the Myricom LANai 1.2 compiler and its associated documentation[1]. Test software includes RAM,

register, and loopback tests. “Include” files are provided for easy access to board-level registers, and for loading LANai control programs. Some performance test programs are also included.

3.0 Performance

The PC-ATOMIC card has been tested in point-to-point mode to other PC-ATOMIC cards, and to Myricom 1.2 and 2.3 Sun SPARC SBus host interfaces [5]. The PC-ATOMIC card supports programmed I/O at 88 Mbps. The testing found zero errors over 10 million packets. The tests were:

- *i486 on-board loopback (pre-driver and post-driver)*
- *i486 to i486*
- *i486 to Myricom S-Bus interface.*

In addition, some experiments have been performed regarding the programmed-I/O bandwidth capabilities of the PC-ATOMIC boards, including a comparison to Myricom's 2.x SPARC S-Bus interface cards. Figure 4 shows write (from host to interface) and read (from interface to host) programmed I/O bandwidth between host memory and the board (“bcopy” performance). The graph indicates that the PC-ATOMIC card is comparable to the Myricom SBus card in read bandwidth, which is the limiting factor for both cards.

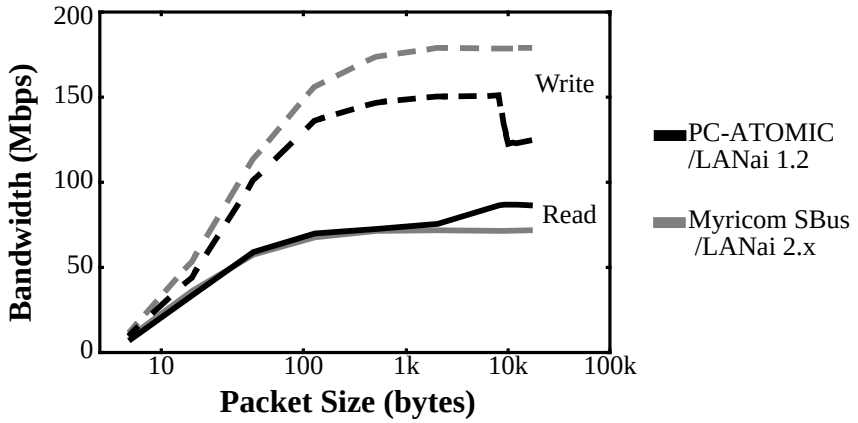

FIGURE 4. Interface memory to host memory bandwidth (bcopy)

All graphs have been computed with 90% confidence intervals which are near +/- 2 Mbps for each data point. The i486 PC-ATOMIC system performance degrades at 8K byte packets, due to page-boundary crossing.

In comparing the PC-ATOMIC and Myricom interfaces, it is useful to keep the following information in mind:

Host	Host Processor	Backplane	Interface	LANai Processor
VL-Bus PC	i486 @ 66 MHz	VL @ 33 MHz	PC-ATOMIC	v1.2 @ 16.5 MHz
Sun 10/51	SPARC @ 40 MHz	SBus @ 20 MHz	Myricom SBus	v2.0 @ 20 MHz

Table 1. Comparison of host configurations

1. Myricom software is not included in the PC-ATOMIC software release. Use of the PC-ATOMIC software requires a Myricom nondisclosure agreement, due to proprietary information in the code.

Table 1 indicates that the PC-ATOMIC tests use a faster backplane than the Sun SBus tests, but use a slower CPU[1] and slower LANai processor. Even so, Figure 4 indicates that the local RAM access performance is nearly identical for the bottleneck read rate. This may be an indication that the backplane speed plays a critical role in the performance of the host interface.

The application-application bandwidth was also measured, both between PC-ATOMIC interfaces, and to Myricom interfaces. Figure 5 indicates the memory-memory bandwidth for native ATOMIC packets sent directly from a user process. For reference, the Myricom SBus PI/O bcopy performance and kernel-based TCP measurements are also included. These performance graphs indicate that the PC-ATOMIC interface performs as well as the Myricom interface for programmed I/O.

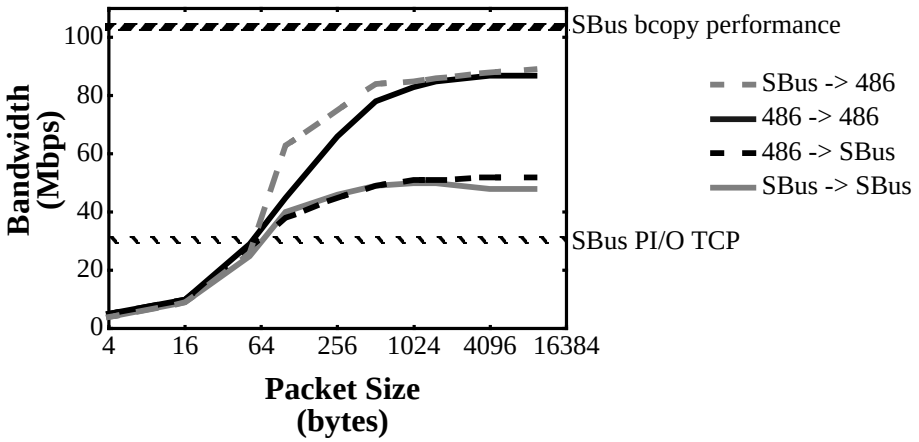

FIGURE 5. Host memory to host memory bandwidth

4.0 Observations

The PC-ATOMIC design encountered some common host-interface issues, as well as discovering some new ones.

The PC is a little-endian host, but network-standard byte-order is big-endian. This problem is compounded by the use of shared-memory for communication, and a big-endian 16-bit off-board processor (the LANai). We considered the use of a dual address space, where byte-order conversion was performed in the data path via wire routing from two sets of data buffers. This was not implemented in the PC-ATOMIC interface due to space and complexity limitations. We are not sure of the utility of such a mechanism.

The Internet checksum was implemented in a very inexpensive part at gigabit rates (Appendix A). This provided a zero-overhead checksum during any DMA and P I/O operations. Subsequent research considered the replacement of this checksum with an IPv6 header authentication algorithm [10].

DMA capability is part of the PC-ATOMIC design. It has not been fully tested, due to OS limitations. The PLDs can be reprogrammed to enable DMA, and other data paths that are part of the board design (Appendix B). This includes LANai control of the board-level registers.

1. Even though the i486 processor runs at a faster clock rate than the SPARC, it is less powerful. SPEC benchmarks indicate that the i486 at 66 Mhz is capable of 32 SPECInts, but the SPARC at 40 Mhz is capable of 53.

The choice of interface bus has proven the major limitation to the PC-ATOMIC interface. At the time the project was initiated, the VL-Bus and PCI bus were in development. The VL-Bus hosts were available at the time the project was underway, and the bus specification was stable enough to implement an interface. There were no standard interface chips for the VL-Bus at the time. The PCI bus as subsequently become the *de-facto* standard for PC host platforms. PCI interface chips are now available, making host interface development much simpler.

5.0 Conclusions

PC-ATOMIC is a high-speed VL-Bus PC host interface to an inexpensive high-performance local-area network technology. It achieves programmed I/O rates in near those of its commercial counterpart for the Sun SPARC S-Bus. It also currently supports IP checksum in hardware, and is capable of supporting DMA with a firmware upgrade.

5.1 Acknowledgments

The PC-ATOMIC project was initiated by Danny Cohen, Robert (Bob) Felderman, and Greg Finn. Bob wrote the original specification, from which this design evolved. PC-ATOMIC was designed with the assistance of ISI's Integrated Systems Lab, including Jeff LaCoss, Mike Gorman, and Bruce Parham. The software and bandwidth measurements were performed with the assistance of ISI's ATOMIC project members, including Hong Xu, Annette DeSchon, and Ted Faber. This project also benefited from the additional assistance of Jan Brooks and Vickie McCorkendale, and especially the diligence of Celeste Anderson.

6.0 References

[1] Braden, R., Borman, D., and Partridge, C., "Computing the Internet Checksum," RFC-1071, September 1988.

[2] Boden, N., et. al, "Myrinet - A Gigabit-per-Second Local-Area Network," *IEEE Micro*, Vol. 15, No. 1, Feb. 1995, pp. 29-36.

[3] Felderman, R., DeSchon, A., Cohen, D., and Finn, G., "ATOMIC: A High-Speed Local Communication Architecture," *Journal of High Speed Networks*, Vol. 3, No. 1, 1994, pp. 1-29.

[4] Information Sciences Institute, ATOMIC Web site, http://www.isi.edu/div7/atomic.

[5] Information Sciences Institute, PC-ATOMIC Web site, http://www.isi.edu/div7/pcatomic

[6] Information Sciences Institute, "PC-ATOMIC (overview),", part of the PC-ATOMIC Software Release, Nov. 1994, available separately via ftp://ftp.isi.edu/pub/hpcc-papers/touch/pca_overview.txt.

[7] Information Sciences Institute, "PC-ATOMIC Board Information," part of the PC-ATOMIC Software Release, in docs/board.info, Nov. 1994.

[8] Information Sciences Institute, "PC-ATOMIC Register Information," part of the PC-ATOMIC Software Release, in docs/register.info, Nov. 1994.

[9] Myricom, Inc., Myricom Web site, http://www.myri.com

[10] Touch, J., “Performance Analysis of MD5,” to appear in Sigcomm ‘95.

[11] Touch, J., and Parham, B., “Computing the Internet Checksum in Hardware,” (paper in progress).

Appendix A : IP checksum

The PC-ATOMIC interface required an inexpensive and fast implementation of the Internet Checksum in hardware [1]. Various designs were considered, including MSI 16-bit fast-adders. The final solution used a $40 AMD MACH 435 PLD to implement the checksum at 1.23 Gbps, with one 32-bit word accumulated every 26 ns [11].

The Internet checksum is computed as a 16-bit ones-complement sum. A ones-complement sum is equivalent to the twos-complement sum, where carries are summed back into the accumulation. It can also be designed ‘natively’ as a twos-complement adder where every bit includes the carry-in of the ring of bits to its right (wrapped around to the left, stopping at the bit to its left). It is this “toroidal” native property that we exploit.

The PC-ATOMIC Internet Checksum is computed as a pair of 16 ones-complement sums, over the high and low half-words of the data. The pair of partial sums are folded together in a single ones-complement sum, which is then inverted to result in the Internet Checksum.

The implementation of this checksum in the AMD MACH 435 PLD uses input latching of 32-bit words, one per clock. The data is then summed into the accumulator on the next clock, such that the latch is pipelined. The summation is composed of groups of 2- and 3-bit fast carry-lookahead adders with pipelined carries between the adder stages in a ring (Figure 6). The carries are propagated during all clocks, and when data is not present on the latch, a zero is added-in (e.g., as a null operation). The resulting pipeline settles in 6 clock cycles.

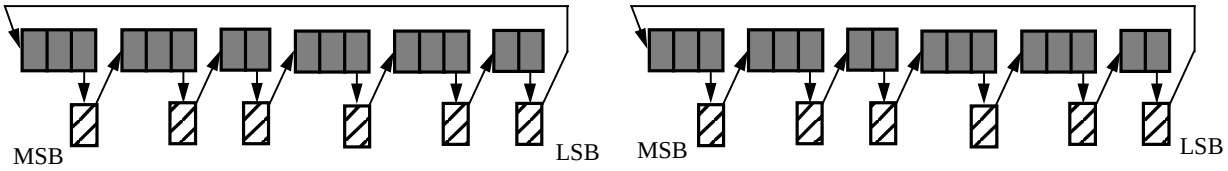

FIGURE 6. Toroidal-pipelined ones-complement adders

Appendix B : Board data flows

The PC-ATOMIC host interface incorporates a Myricom LANai 1.2 processor and communication subassembly, and an interface with registers and control (Figure 7). The LANai interface is very similar to that on the Myricom Sun SBus interface [9]. Logically, the entire LANai processor and communication subassembly appears as RAM to the interface, and thus to the host. The LANai uses dual-access RAM for communication between the network and host.

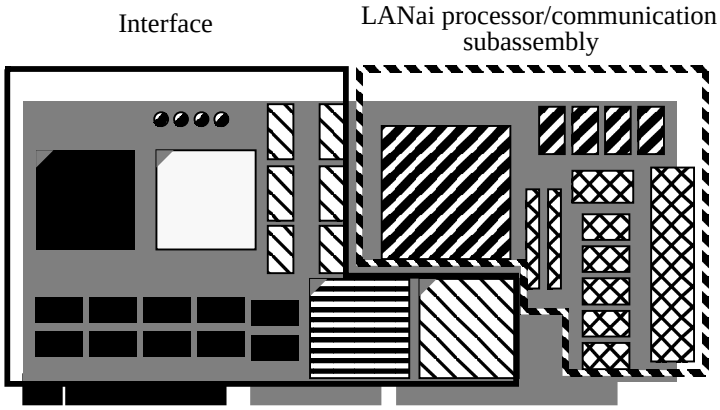

FIGURE 7. Board subassemblies

The remainder of the interface converts between VL-Bus and LANai interface signals, and provides board-level registers (Figure 8).

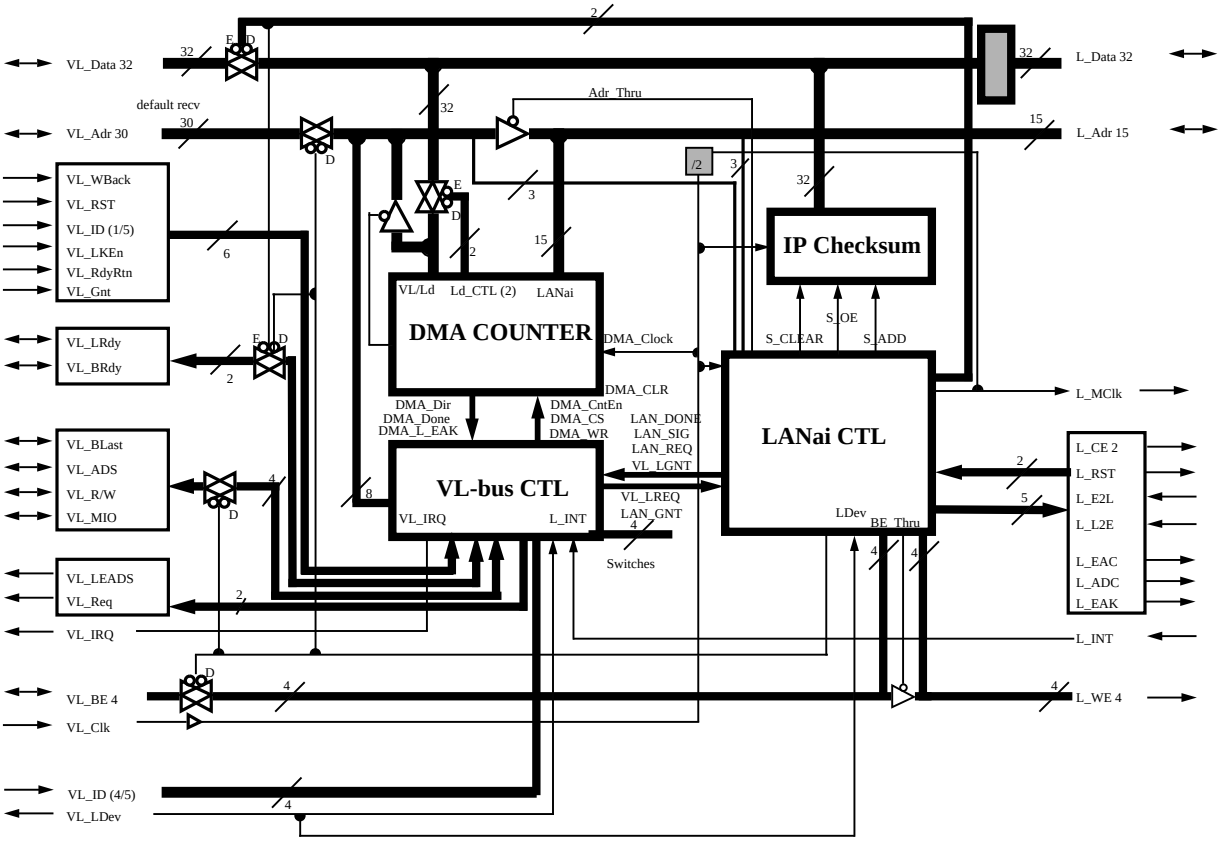

FIGURE 8. Board interface, with registers and control

6.1 Flow diagrams

The PC-ATOMIC board interface supports a total of 9 board flows, supporting host access of shared RAM and board registers, LANai access of both, DMA, and an idle flow[1]. The Idle flow is shown larger to indicate control, address, and data path names, as well as register locations (grey rectangles) (Figure 9). In the Idle flow, VL-Bus control is monitored, and address is decoded into the board-level registers.

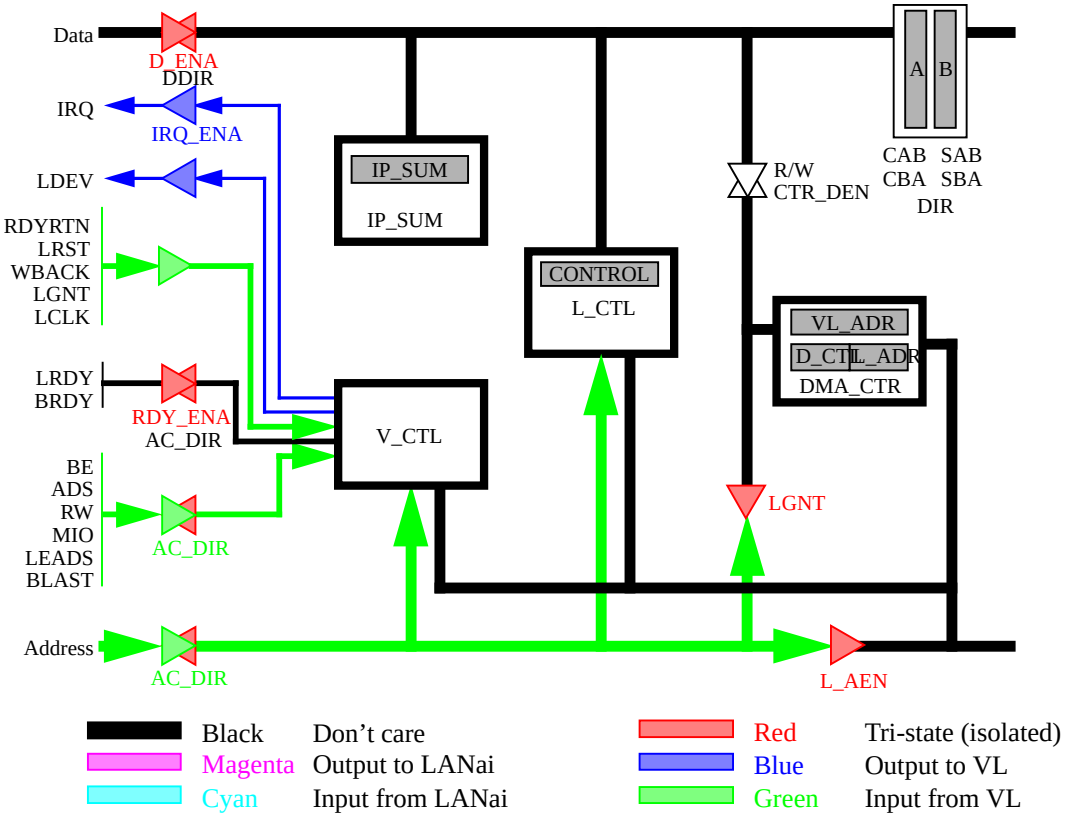

FIGURE 9. Idle

The host can access the board level registers by recognizing a VL-Bus access, decoding the board-level register address, and enabling the appropriate register and data path and direction (Figure 10).

The host can access the LANai dual-access RAM via dual-ported data registers on the interface assembly (Figure 11). These data registers hold data to adjust the clocking between the board (at 33 Mhz) and LANai subsystem (17.5 Mhz). This also permits the RAM accesses to occur during the appropriate phase of the LANai subsystem clock. For host access of LANai RAM, the address is passed on through the interface. Data is asynchronously propagated through the registers when received from the LANai, because the LANai is clocked more slowly. Data sent to the LANai must be clocked through the data port registers to latch it long enough for the LANai.

DMA operation is permitted by using the board-level interface registers to drive the VL-bus and LANai addresses independently as the data is clocked through the data port register (Figure 12). Data written to the LANai is clocked through the data port registers, as in host access of LANai RAM. Data read from the LANai is similarly unclocked (as in host access), because data is held stable longer over the slower LANai clock.

1. These diagrams require color. For a color copy, see <http://www.isi.edu/div7/pcatomic/flows.html>.

The LANai is permitted to access the board-level registers by passing LANai addressing through, and using the data port registers as a data bus (Figure 13). The data port registers use asynchronous signal propagation in both directions, because the LANai read and write cycles are synchronous, where the data is stable until acknowledged.

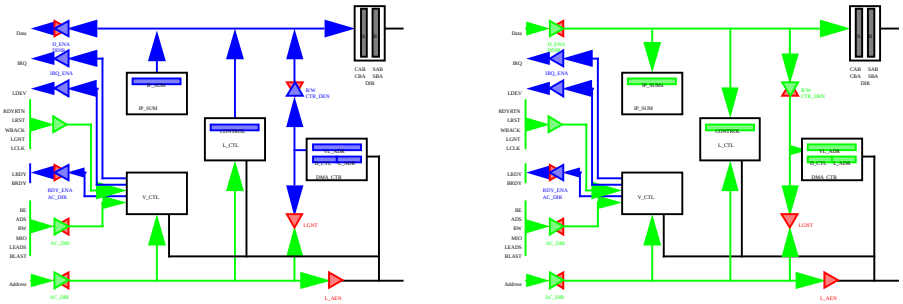

FIGURE 10. Host read register / write register

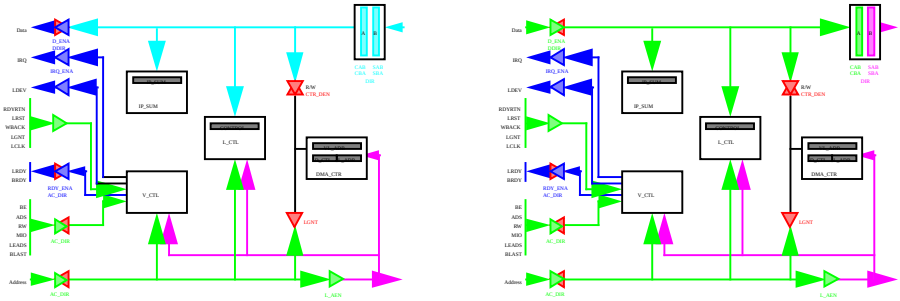

FIGURE 11. Host read memory / write memory

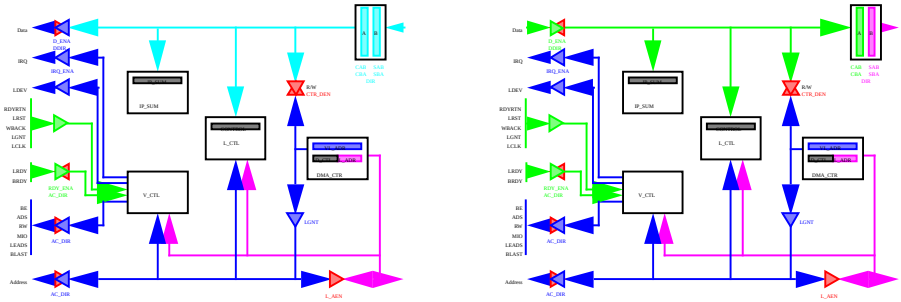

FIGURE 12. DMA out / in

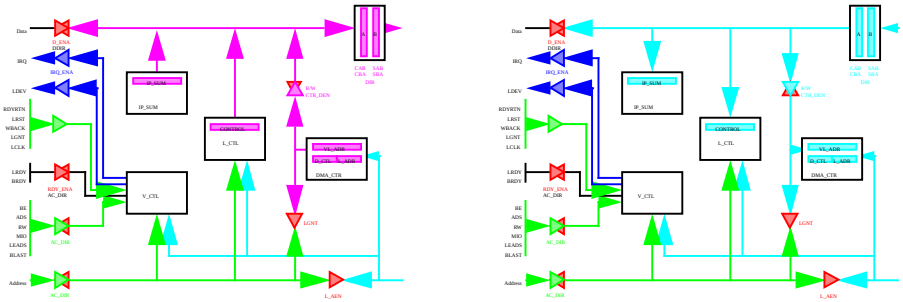

FIGURE 13. LANai read register / write register