Integrated Cluster Bus performance for the IBM S/390 Parallel Sysplex

by C. L. Rao
G. M. King
B. A. Weiler

As the speed of the S/390® processors has increased at a rapid rate over the years, it has become extremely challenging to be able to connect them in a Parallel Sysplex® without compromising system performance. The cost of synchronous accesses to the coupling facility will build up progressively unless coupling link technology keeps pace with the development of the processors. Beginning with the G5 processors, IBM has introduced the ICB coupling link technology to enhance the performance of closely integrated processor clusters. With a high data transmission rate complemented by reduced hardware and microcode path lengths, the ICB provides excellent coupling efficiency. This paper emphasizes the need for the ICB and discusses the superior performance delivered by this latest coupling link technology.

Introduction

As processors become increasingly powerful, it is essential to improve link technology in order to sustain high efficiency in a parallel sysplex cluster. Thus, enhancements to the coupling links are of paramount importance in the overall improvement of the coupling technology. The connectivity of the IBM Parallel Sysplex* is based on intersystem channel (ISC) architecture. So far, IBM has developed two generations of ISC link hardware, which are currently known as ISC-1 and ISC-2 links. (The ISC-2 links are also frequently referred to as HiPerLinks, signifying the high performance of the coupling links.) In addition to these two, IBM has introduced Integrated Cluster Bus (ICB) coupling links, which can coexist with the ISC-1 and ISC-2 but are unique to the processors of the G5 and G6 generations.

Following a brief description of the S/390* Parallel Sysplex, this paper describes the components that constitute synchronous access using the coupling facility. The cost of synchronous access is crucial in determining the efficiency of a Parallel Sysplex. The impact of coupling link speed on the performance of the Parallel Sysplex is considered next, followed by a description of IBM efforts in the development of intersystem channels and the Integrated Cluster Bus. The excellent performance of the ICB is reflected in the coupling facility access times in a G5 Parallel Sysplex, as discussed in the section on the

©**Copyright** 1999 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

0018-8646/99/$5.00 © 1999 IBM

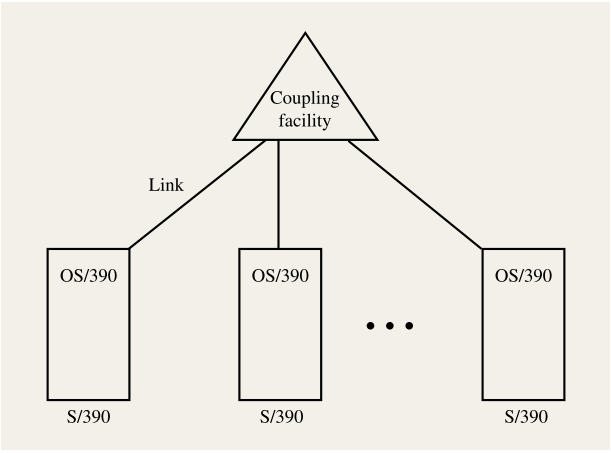

Figure 1

Major elements of the S/390 Parallel Sysplex.

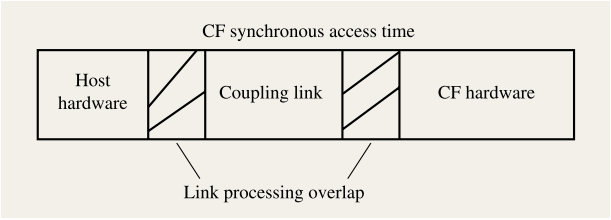

Figure 2

Components of a coupling facility synchronous access.

performance measurements. Finally, the performance benefit of the ICB for customers is mentioned, followed by a comment on the advantages of the IBM S/390 Parallel Sysplex over HDS and Millennium processor complexes.

S/390 Parallel Sysplex

The S/390 Parallel Sysplex allows up to 32 S/390 servers to be clustered together and appear as a single system image to the end user and applications. Databases are fully shared and are accessible to each server in the cluster. Transactions and jobs are dynamically distributed among the servers on the basis of available capacity. Servers can be temporarily removed for planned hardware and software maintenance or upgrade while the remaining servers continue to process the work. In this manner, the Parallel Sysplex provides a highly scalable, highly available enterprise computing system. References [1–3] provide a thorough exposition of the architecture, facilities, and performance aspects of a Parallel Sysplex.

The major components of a Parallel Sysplex are illustrated in **Figure 1**. The "heart" of the configuration is the coupling facility (CF). To facilitate data sharing and workload management, various lock, cache, and list structures are placed in the coupling facility and are accessed by software components executing under OS/390* on each S/390 host server. The coupling facility itself consists of hardware resources and microcode running in a partition on an S/390 server. Each S/390 host is connected to the coupling facility by coupling links.

Coupling facility

The efficiency of a Parallel Sysplex is affected by the cost of an access to the coupling facility. To avoid the penalty associated with interrupts and task switching, most accesses to the coupling facility are synchronous in nature; that is, the processing engine which initiates the request will "dwell" (wait in a busy state) while the request is processed. The amount of time a processing engine spends dwelling for a coupling facility access represents lost capacity to that processor. Therefore, the time required to process a coupling facility request is critical, and must stay in balance with the speed of the processing engine in order to preserve a high level of efficiency.

- *Components of a coupling facility access*

The processing of an access to the coupling facility consists of three major components; these are illustrated in **Figure 2**.

Host hardware

When a processing engine initiates a request to the coupling facility, the request is first processed by microcode and hardware on the host S/390 server before it is sent across the coupling link. This "host hardware" time is affected by the speed (cycle time) of the host processor, the path length in the microcode, and the efficiency of the adapter that interfaces with the coupling link.

Coupling link

The request next proceeds across the link to the coupling facility. The time spent on the link is affected by the speed of the link, the amount of data, the distance to the coupling facility, and the amount of overlap that occurs with other processing associated with the request at the host or coupling facility ends of the link.

Coupling facility hardware

Once the request reaches the coupling facility, it is processed by the hardware and microcode on that server. As in the host hardware component, the time spent in the coupling facility is affected by the speed (cycle time) of the processor, the path length of the microcode used to perform the requested coupling structure operation, and

the efficiency of the adapter that interfaces with the coupling link.

Upon completion of the request in the coupling facility, the data to be returned is sent back over the coupling link to the host server. The host hardware then finishes processing the request. Finally, the processing engine that initiated the request leaves its dwelling state and returns to actively executing instructions.

- *Importance of coupling link speed*

As the speed of the S/390 host server increases, the time required to process a coupling facility request must decrease proportionately, or the capacity penalty due to dwelling time will grow. Consider the following example: If a 50-MIPS (millions of instructions per second) processing engine were to dwell for 100 μs while a coupling facility request was being processed, the capacity cost of that request would be $50 \times 100 = 5000$ instructions (that is, the equivalent of 5000 instructions of processing capacity would be lost while dwelling). Suppose that an advance in server technology increases the speed of the engine to 100 MIPS but the dwelling time remains at 100 μs; the capacity cost of a coupling facility operation would now be doubled: $100 \times 100 = 10\,000$ instructions. To preserve the coupling efficiency, the dwelling time for this faster server would have to be reduced to 50 μs.

The components of a coupling facility access were described earlier. **Figure 3** illustrates the relative cost of a coupling facility access as the speed of the S/390 host processor increases and improvement is made to each component. The steepest curve shows the increasing cost when the host processor speed increases but nothing is done to improve the other components of coupling facility access time. The middle curve reflects changing the coupling facility server to be equal in speed to the host server (that is, the processing-engine speed is the same for both servers). The lower curve additionally includes speeding up the coupling link and associated link hardware. The graph clearly shows that improvements in coupling link technology must keep pace with improvements in host server technology in order for the cost of a coupling facility access to remain flat.

Evolution of the Parallel Sysplex coupling links

To offset the performance degradation in a Parallel Sysplex caused by increases in processor speed, IBM has been continually searching for solutions to use in launching better link technology to significantly reduce latencies and provide greater bandwidth (data transmission rate). The evolution of the coupling links ISC-1 and ISC-2 is considered in this section. Another development, the ICB, is discussed in the following section.

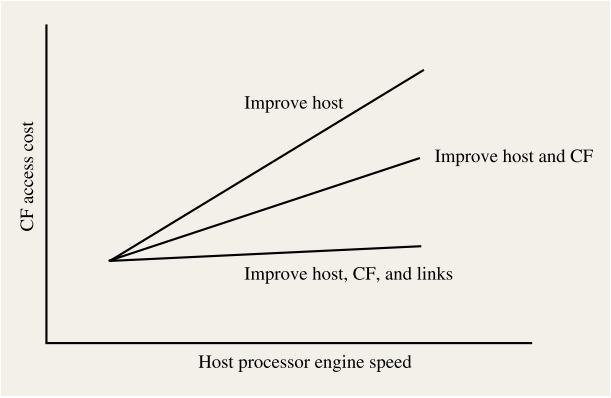

Figure 3

Cost of a coupling facility access vs. host processor speed.

- *ISC-1*

The ISC-1 is the coupling link hardware that was initially offered at the onset of IBM coupling technology in 1994. It is a full-duplex fiber optic link capable of transmitting either 531 or 1062 megabits of data per second. The 531-Mb link is a low-cost multiple-mode optical link that could be used up to a distance of half a kilometer. On the other hand, the 1062-Mb link is a single-mode optical link which could be extended up to 20 kilometers. With 8/10-bit data encoding, the multiple-mode and single-mode links provide data bandwidths in excess of 50 and 100 Mb/s, respectively. The ISC-1 links have been used extensively with the G1 and G2 CMOS-based systems, as well as the earlier H5 bipolar processors.

The CF command transmission over the link from the host or sender processor begins with the sending of a message control block (MCB) to the coupling facility. Additionally, for a CF write operation, data blocks may be sent over the link to be written into the structures in the CF. (Similarly, data may be read from the CF in a CF read operation.) Since the data block size cannot exceed the size of the link buffers, the CF sends an acknowledgment back to the sender at the end of each data block transfer, in the case of multiple blocks of data transfer. Finally, a message response block (MRB) is sent back to the sender by the CF, indicating the completion of CF command processing.

- *ISC-2*

The ISC-2 (HiPerLinks) hardware is the second generation of link hardware developed for S/390 processors and was first used in the G3. HiPerLinks improve channel efficiency and reduce latency in the processing of requests to the coupling facility. The ISC-2

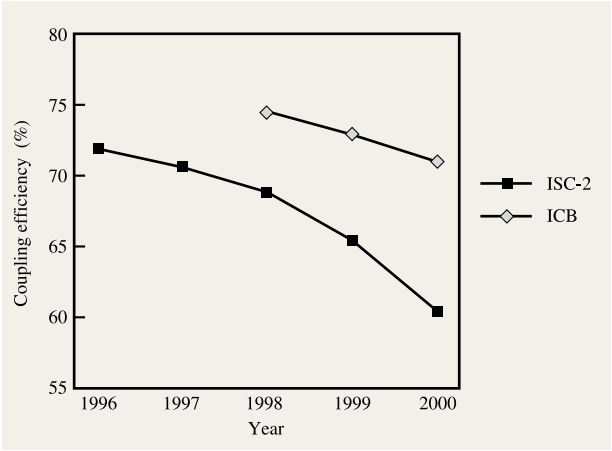

Figure 4

ICB and ISC-2 coupling efficiency over time.

links are capable of transmitting CF commands and data at the rate of 100 MB/s.

As stated earlier, most coupling facility requests are accessed in synchronous mode. On the other hand, when a CF command is issued asynchronously, the requesting processor frees itself to do other useful work until the response is received from the coupling facility. However, asynchronous processing requires additional software path length to perform task switching as well as to process the completion of the asynchronous CF command. For CF commands that can be issued in either mode, the operating system (OS/390) generally selects the mode that provides the best coupling efficiency (overall system throughput in a data-sharing environment). The improvements provided by ISC-2 benefit both modes of CF commands.

The superior performance of ISC-2 is accomplished by a redesign of the host adapter hardware that interacts with the intersystem channel function. This design differs from the design of ISC-1 primarily in two respects—the addition of a list processor and the inclusion of hardware that prepares the address of the MRB in advance. The list processor significantly improves the performance of multiple data transfers to and from the CF by providing direct control of the transfers by the host adapter and ISC. Earlier, in the design of ISC-1, the processing of multiple data transfers included an interaction between the system assist processor (SAP) and the coupling facility control code (CFCC), resulting in inefficient data transfers. The list processor also accommodates a high-performance scatter/gather function in the coupling facility.

The preparation-for-MRB function facilitates improved send-message performance by allowing the processor in the sender to perform the front-end processing of the CF command in both synchronous and asynchronous modes.

More discussion on the host adapter, list processor, and preparation for MRB can be found in [4, 5]. The coupling efficiency of a Parallel Sysplex may increase by one to two percent when the previous-generation coupling channels are replaced with HiPerLinks. Coupling facility link capacity may also improve by 20% with the use of HiPerLinks. Data-sharing applications that frequently transfer blocks of data greater than 4 KB to and from the coupling facility see considerable reduction in elapsed time with the use of HiPerLinks. Also, HiPerLinks reduce XCF (the intersystem messaging facility used by many OS/390 subsystems) signaling latency, which becomes comparable to the channel-to-channel (CTC) communication time, when the XCF structures reside in the coupling facility. The CF commands pertaining to large data transfers (greater than 4 KB) and the XCF signaling are issued asynchronously in the sender processor. Because of considerable improvement in the latency of the asynchronous processing of CF commands with the implementation of HiPerLinks, significant coupling performance gains can be realized.

Integrated Cluster Bus (ICB)

While the performance of ISC-2 is impressive on the G3 and G4 processors, the coupling efficiency of a Parallel Sysplex cluster would gradually wear away over time with the introduction of the faster G5 and later S/390 processors, unless enhancements to the coupling links are contemplated. Extensive studies of the coupling performance of future processors at IBM have led to a requirement for a substantial increase in the data transmission rates of the coupling links. The Integrated Cluster Bus (ICB) design, with a sustained data transmission rate of 250 MB/s, has emerged as a viable option that could provide the desired levels of coupled-system performance in a cluster of G5 or G6 processors. However, the ICB connectivity mandates that the systems that are coupled be situated within seven meters of each other.

Figure 4 depicts the coupling efficiency associated with synchronous execution of the coupling facility commands in the evaluation of the ISC-2 design across current and future IBM processors. The workload modeled here comprises IMS shared-message queues (SMQ) as the transaction manager and DB2 as the database manager, with full data sharing across the Parallel Sysplex. In this benchmark, message blocks are written into and read out of the SMQ list structure in the coupling facility by the shared system. Also, the IMS logger writes data blocks of varying sizes into the CF logger structure and reads out 64 KB of data at a time. The CF locking and caching

characteristics are derived from the IBM Relational Warehouse Workload (IRWW) benchmark that is used by the Poughkeepsie and Santa Teresa laboratories to evaluate DB2* data-sharing performance.

For the selected benchmark, it is evident from Figure 4 that the drop in coupling efficiency is projected to be rather steep with the use of ISC-2, as processors evolve through the time period from 1996 to 2000. Even though the workload used in this evaluation is somewhat stressful, much more than would normally be expected from a typical customer configuration, the message is strikingly clear. In order to sustain coupling performance, it is imperative to improve coupling link technology at least every two years.

Figure 4 also shows the coupling performance of the ICB design with a data transmission rate of 250 MB/s for the years 1998 and 1999, the period during which the G5 and G6 processors were introduced. The improvement in coupling performance provided by the ICB links is exemplary, and the coupling efficiency attains values that were never realized before.

In addition to its capability to transmit data at a high rate, the superior performance of the ICB is also attributable to its reduced hardware and microcode path lengths. The ICB is an implementation of the S/390 coupling communication architectures over the self-timed interface (STI). The ICB consists of a relatively inexpensive copper connection between the STI ports of the S/390 host processor and the attached coupling facility, which provides direct and fast communication between the coupled systems. Comprehensive details of the ICB design can be found in [4]. The overall benefit of the ICB design is to significantly reduce latency in accessing the coupling facility. The measured response times of the coupling facility using ICB links are given in the next section.

- *Measured performance*

The ICB is able to transfer data substantially faster than its predecessor link, the HiPerLink. The technology differences stated above allow for this faster delivery of data to or from the coupling facility. Experiments have been conducted to determine the amount of performance improvement for coupling facility accesses achievable by using the faster ICB instead of the relatively slower HiPerLink. How this performance affects requests to the coupling facility that contain varying amounts of data is of particular interest. Any benefit for requests requiring no data transfer, or only small amounts of data transfer, is also of interest. The improvement achieved by replacing a HiPerLink with an ICB has been measured, and the results are presented here.

A number of measurements were designed to determine the degree to which the ICB would improve the performance of various data-transferring requests to the coupling facility. Requests to the coupling facility were issued for the various coupling facility commands at varying data lengths. These included read and write requests (which transfer data) to list and cache structures located in the coupling facility. The list and cache requests were issued with data lengths ranging from 256 bytes to 64 KB of data transfer. These requests were timed from the issuance of the request by the hardware until the response was received from the coupling facility. This time includes the time required for the request to be sent, along with any data transfer to the coupling facility via the link (HiPerLink or ICB), the time required for the coupling facility to process the request, and the time required for the response and any associated data to make its way back to the requester via the link. This time is referred to as the coupling facility response time.

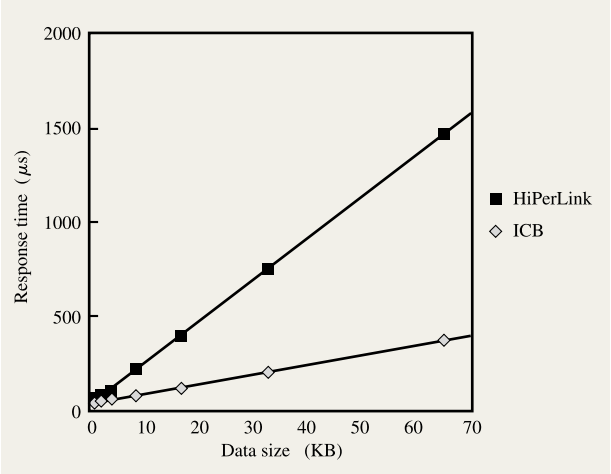

Figure 5

Coupling facility response times for requests with varying data sizes for a G5 configuration with HiPerLinks and with ICBs.

These times are for the hardware only, and do not include any of the software cost associated with executing these commands in a customer environment. No operating system or subsystem was involved in the experiments. For this reason, the times here represent only the hardware portion of the actual times observed in customer environments. One would expect the times in a customer environment to be higher than those that appear here.

The tests described above were run on a Generation 5 (G5) CMOS configuration featuring HiPerLinks, and then repeated on the same configuration using ICBs. Results of the experiments were compared in several ways. **Figure 5** illustrates the improvement in coupling facility response times for varying data lengths with the use of ICBs. Notice that the performance improvement for the ICB increases

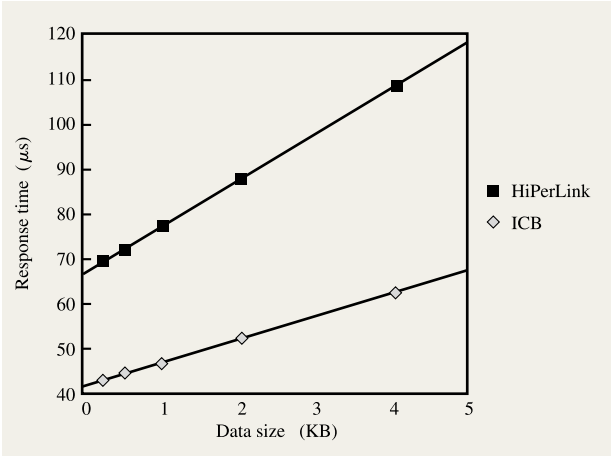

Figure 6

Response times for requests with 4096 bytes of data or less for a G5 configuration with HiPerLinks and with ICBs.

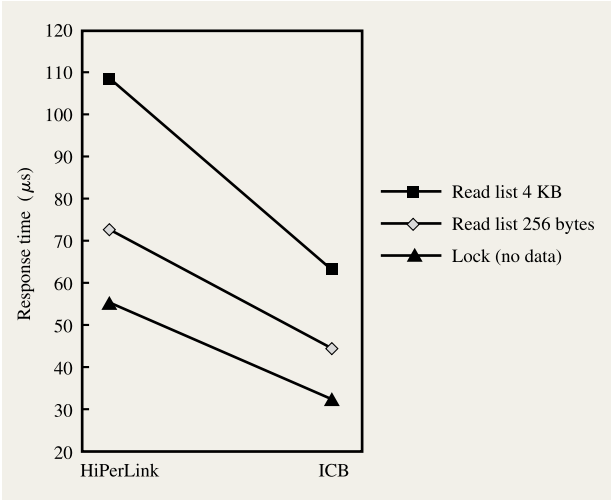

Figure 7

Response time improvement for ICB with little or no data transferred.

with the amount of data being transferred. That is, the difference between the response time for the HiPerLink and the response time for the ICB increases as the amount of data transferred increases. These response time "deltas" represent an improvement of up to 74% in coupling facility response time for the requests which transfer the most data, 64 KB.

The smaller data sizes are more prevalent in customer environments, and it is important to note that they also show significant performance improvement for coupling facility accesses when ICBs are used. **Figure 6** is a magnification of the section of Figure 5 dealing with data lengths up to 4096 bytes. Here, the average improvement in coupling facility response time is approximately 44%, a rather significant improvement.

Looking closely at the trend line in Figure 6, it appears that one would also expect substantial improvement for coupling facility accesses which do not involve data movement. In fact, the zero-data case does achieve significant benefit from the ICB. Elimination of the link adapters on each end of the link, combined with the faster link, provides the improvement for the zero-data case. Although no data is being transferred, the request and response information will gain minimal benefit from the faster speed on the ICB. **Figure 7** illustrates the improvement seen in the coupling facility response time for lock requests which do not transfer data, in contrast to the response time for a 256-byte data transfer request and a 4096-byte transfer request. The key point to notice here is that the improvement curves for little (256 bytes) data and no data (lock) have the same slope.

The improvement in coupling facility response times due to the faster ICB link shown in these measurements allows the performance of the Parallel Sysplex to remain efficient when using the faster processors that are now available for the host and coupling facility. The improvement becomes greater with larger data transfers but is still significant in cases where little or no data is transferred.

- *Customer value*

Many customers have realized the value of a Parallel Sysplex by moving their mission-critical applications to data sharing. These workloads produce a variety of accesses to the coupling facility. The vast majority of these accesses are synchronous and contain data sizes up through 4 KB. The rate of access to the coupling facility in these heavy data-sharing production sysplexes tends to fall in the range of 8 to 10 accesses per million instructions. With prior generations of CMOS processors and coupling technology, this load resulted in a coupling cost of about 10%; that is, about 10% of the capacity of the sysplex was used to process accesses to the coupling facility. (Note: This cost includes both hardware and software components.) Since the G5 processors are more than twice as fast as previous generations, these customer workloads would have experienced increasing coupling costs without the ICBs.

Table 1 shows the coupling costs for customer workloads that are primarily data-sharing as a function of host processor speed and coupling technology. It begins with G3 processors and the corresponding C04 coupling technology, which results in a 10% cost. One can see that a G5 processor using the C04 coupling technology would

Table 1 Effect on coupling cost of processor speed and coupling technology (in percent of host capacity).

Coupling technology	*Host*				
	G3	*G4*	*G5*	*Skyline**7*	*Millennium7*5*
C04	10	11	16	19	17
C05 HL	9	10	14	16	16
R06 HL	9	9	12	14	14
ICB	—	—	9	—	—

experience a significantly higher cost (16%). Upgrading to the R06 (G5 coupling facility) but staying with a HiPerLink would lower the cost to 12%. Introducing the ICB technology reduces the cost to 9%. Thus, the ICB improves the capacity of a sysplex by 3%.

- *Competitive advantage*

Table 1 also includes several PCM processors. The Hitachi Data Systems† Skyline**7 series is similar in speed to the G5 on an individual engine basis. However, the Skyline**7 coupling links are somewhat less efficient than IBM HiPerLinks, as can be seen by the 14% cost with the R06-HL coupling technology versus the 12% cost for the G5. This is also the lowest cost achievable by the Skyline**7. Thus, the Skyline**7 processors will incur more than 50% higher coupling costs (14% versus 9%) compared to a G5 with ICB coupling technology. The Amdahl** Millennium7*5 processors fall between the G4 and G5 processors in speed, and also suffer from less efficient coupling links. Even with R06-HL coupling technology, the Millennium7*5 will incur a coupling cost of 14%, more than 50% higher than a G5 with ICBs. Thus, the ICBs give the G5 a significant competitive advantage in customer sysplexes.

Conclusion

The Parallel Sysplex cluster technology is an integral part of the S/390 processor platform. Development of high-speed, low-latency coupling link technology is crucial for the continued success of Parallel Sysplex cluster technology in the server market. Since the advent of Parallel Sysplex clusters, IBM has continuously improved the performance of the coupling links commensurate with the speeds of the S/390 processors. Along with fiber optic ISC links, the development of the high-performance ICB suitable for the connectivity of processors in close proximity is an example of IBM's efforts in this direction.

*Trademark or registered trademark of International Business Machines Corporation.

†Trademark or registered trademark of Hitachi, Ltd. or Amdahl Corporation.

References

1. J. M. Nick, B. B. Moore, J. Y. Chung, and N. S. Bowen, "S/390 Cluster Technology: Parallel Sysplex," *IBM Syst. J.* **36,** No. 2, 172–201 (1997).
2. N. S. Bowen, D. A. Elko, J. F. Isenberg, and G. W. Wang, "A Locking Facility for Parallel Systems," *IBM Syst. J.* **36,** No. 2, 242–283 (1997).
3. G. M. King, D. M. Dias, and P. S. Yu, "Cluster Architectures and S/390 Parallel Sysplex Scalability," *IBM Syst. J.* **36,** No. 2, 221–241 (1997).
4. T. A. Gregg, "S/390 CMOS Server I/O: The Continuing Evolution," *IBM J. Res. Develop.* **41,** No. 4/5, 449–462 (1997).
5. T. A. Gregg, K. M. Pandey, and R. K. Errickson, "The Integrated Cluster Bus for the IBM S/390 Parallel Sysplex," *IBM J. Res. Develop.* **43,** No. 5/6, 795–806 (1999, this issue).

Received January 11, 1999; accepted for publication June 2, 1999

Chitta L. Rao *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (clrao@us.ibm.com).* Dr. Rao is with the S/390 hardware performance group, working on Parallel Sysplex performance. He received an M.S. degree in electrical engineering from McGill University, Montreal, Canada, and a Ph.D. degree in theoretical nuclear physics from the University of Tennessee. Prior to joining IBM, Dr. Rao was engaged in research in nuclear physics, digital filters, and digital image processing. He joined IBM to work on large-system scientific and engineering processor design and development. Since 1991 he has been involved in the performance analysis of coupling facility design alternatives and Parallel Sysplex performance.

Gary M. King *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (garyk@us.ibm.com).* Mr. King is a Senior Technical Staff Member in the S/390 Division, consulting on all aspects of system performance. He joined IBM in 1974 and has been involved in the design and evaluation of system resource managers, most notably in the area of storage management. For the past nine years, his efforts have focused on clustered system performance, particularly the S/390 Parallel Sysplex and its exploiters. He holds five U.S. patents and has received six Outstanding Technical Achievement and Outstanding Innovation Awards in a variety of areas, including storage management, data compression, and performance analysis. Mr. King received a B.S. degree in mathematics from the University at Albany, State University of New York, in 1972 and an M.S. degree in computer science from the Pennsylvania State University in 1974.

Barbara A. Weiler *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (bajala@us.ibm.com).* Ms. Weiler is an Advisory Software Engineer; she has been working in the S/390 performance area since joining IBM in 1981. Her current responsibilities include performance evaluation of the S/390 Parallel Sysplex, including coupling facility control code (CFCC) and various coupling options. Her previous experience includes processor design analysis and performance evaluation, including Vector Facility performance and FORTRAN performance evaluation. Ms. Weiler received a B.S. degree in mathematics from Marist College in 1974 and an M.S. degree in secondary mathematics education from the State University of New York at New Paltz in 1978.