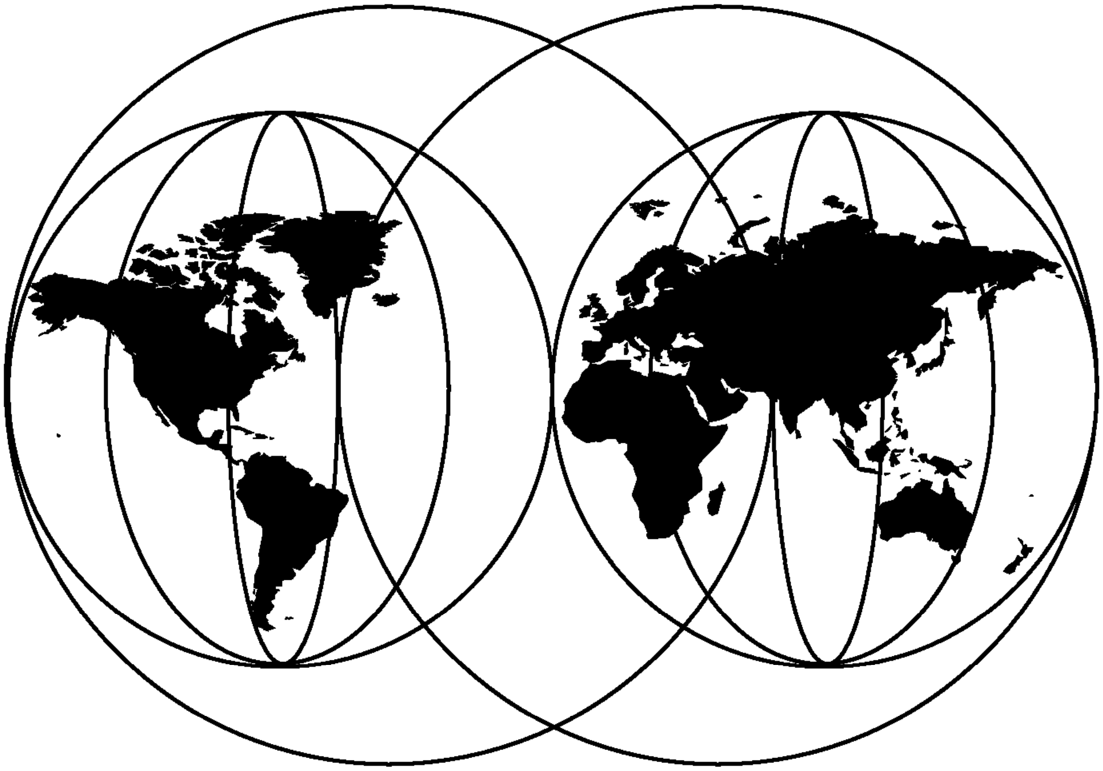
System/390 Parallel Sysplex Performance



International Technical Support Organization

http://www.redbooks.ibm.com

SG24-4356-03

IBM

International Technical Support Organization

SG24-4356-03

System/390 Parallel Sysplex Performance

December 1998

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix E, “Special Notices” on page 191.

Fourth Edition (December 1998)

This edition applies to Version 5 Release 2 of MVS/ESA System Product (5655-068 or 5655-069) or any release of OS/390 (5645-001) running in a Parallel Sysplex.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. HYJ Mail Station P099
522 South Road
Poughkeepsie, New York 12601-5400

Contents

Figures

Tables

Preface

This redbook is intended to help customer technical staff understand the performance characteristics of Parallel Sysplex. The Large System Performance Evaluation team in the IBM System/390 Division measured various aspects of Parallel Sysplex performance in several different configurations. Shared database test cases used IMS DL/I, DB/2 or VSAM/RLS.

This document presents an overview of the concepts involved, performance tips collected during the measurements and details of performance in specific configurations.

Products measured in a Parallel Sysplex environment include the 9021, 9672, IMS/VS, CICS/VS, DB2, JES2, system logger, CICS temporary storage and XCF signalling.

The Team That Wrote This Redbook

This document is the result of many people′s efforts.

The fourth edition was updated by the following people:

Dave Clitherow	ITSO Poughkeepsie
Clive Druett	IBM UK
Steve Grabarits	IBM USA
Mike Kenley	IBM USA
Jock Marcks	IBM Australia

The following people from System Performance Evaluation and Performance Design were responsible for the individual studies in this document, and are the authors of the individual sections.

- Andy Abbey
- Judi Bank
- Peter Bunk
- Seewah Chan
- Joe Fontana
- Kathy Grabarits
- Steve Grabarits
- Joan Kelley
- Gary King
- Rose Manz
- Brian Murphy
- Faith Pacifico
- Jim Perlik
- Judy Richtmyer
- Russ Smith
- Clarisse Taaffe-Hedglin
- Mark Wisniewski

The authors would like to thank the following people who contributed to the completion of the Parallel Sysplex Performance Evaluation effort.

- Dave Bliss
- John Burgess
- John Campbell
- Cedric Chen
- Sinmei DeGrange
- Judy Dibbell
- George Dillard
- Tim Dunn
- Tim Fohner
- Jim Furlong
- Russ Heald
- Jeff Jones
- Gopal Krishnan
- Phil Lahue
- Diana Magel
- Max Maurer
- John Morris
- Manfred Olschanowsky
- Tim Pickrell
- Chitta Rao
- Steve Schneider
- Ollie Simpson
- Jim Strickland
- Brian Thomas
- Dave Thomas
- Sherry Whitney
- Joe Winkelbauer
- Chung Wu
- Larry Zaino

This redbook was produced by the ITSO Poughkeepsie. The project to produce the first edition was coordinated by Tom Russell from the ITSO and Joan Kelley from the System Performance Evaluation group in the IBM S/390 Division.

The second and third editions were coordinated by Tom Russell from the ITSO and Kathy Grabarits from the Performance Design group in the IBM S/390 Division.

The fourth edition was coordinated by Dave Clitherow from the ITSO and Steve Grabarits from the Performance group in the IBM S/390 Division.

Comments Welcome

Your comments are important to us!

We want our redbooks to be as helpful as possible. Please send us your comments about this or other redbooks in one of the following ways:

- Fax the evaluation form found in "ITSO Redbook Evaluation" on page 201 to the fax number shown on the form.
- Use the online evaluation form found at `http://www.redbooks.ibm.com/`
- Send your comments in an Internet note to `redbook@us.ibm.com`

Chapter 1. The Evolution of Parallel Sysplex Performance

In April 1994, IBM announced Parallel Sysplex, the parallel clustering architecture for IBM System/390. Since the original announcement, IBM has made many significant enhancements to OS/390, the software subsystems, the microcode and the hardware infrastructure of a Parallel Sysplex. Many of these enhancements have been in the performance area, enabling Parallel Sysplex to keep pace with the rapid year-on-year growth of faster CMOS microprocessors from IBM. Furthermore, Parallel Sysplex has gained very wide acceptance in the high-end Systems/390 marketplace, with many customers now running Parallel Sysplex, thus providing a wealth of practical performance data.

In this redbook we look at four areas related to Parallel Sysplex performance:

IBM Benchmarks These are benchmark workloads IBM has run on Parallel Sysplex, from the original CICS and IMS DBCTL data published in 1994 through to the latest IMS Version 6 benchmarks.

Customer Data The benchmarks however only tell part of the story, depicting a very stressed data sharing environment. With the depth of customer data now available, this redbook provides many customer examples of the performance they have achieved in practice, and relate this to the benchmarks.

New Technology To show how Parallel Sysplex has been keeping pace with hardware performance, the third area we address is to describe the factors that affect performance and how the implementation has evolved.

Tuning Tips The fourth topic covered will be how to tune your Parallel Sysplex to gain optimal performance.

1.1 Summary of Enhancements

Since April, 1994, IBM has announced and delivered significant enhancements for Parallel Sysplex in both hardware and software, as follows:

- IBM CMOS technology uniprocessor speed has increased from approximately 14 MIPS for Generation 1 in 1994 to over 100 MIPS for the Generation 5 family of CMOS processors announced May 7, 1998. This improved uniprocessor speed has not only provided faster systems, but also much faster coupling facilities.
- Coupling Facility Control Code has evolved from the initial CFCC Level 0 in 1994 to CFCC Level 6 with the announcement of the Generation 5 CMOS processors.
- ISC link performance has been improved with Hiperlinks in 1997, and Integrated Clustering announced May 7, 1998.
- IMS and DB2 have both evolved through several new versions, with significant reduction in locking rates.

Taken together, these enhancements have kept the increased load for data sharing to under 10% for most customer environments as the following data shows.

1.2 Comparison of Customer and IBM Benchmark Data

IBM invests a great deal of time and effort in benchmarking Parallel Sysplex and very openly publishes the results. The purpose of these tests are to intentionally stress the Parallel Sysplex data sharing environment. We want to establish the extremes of performance, when every access to data is a shared data access with minimal application logic and very high locking rates. It is only by this method that we can observe just the path length of the Parallel Sysplex functions. What do we find from this? We can make three observations:

- There is a cost, a fairly fixed cost we call *multisystems management*, for running systems in a Parallel Sysplex. Customers would typically expect to see a similar cost; it is of the order of 3-4%, which is similar to moving from MVS Version 4 to MVS Version 5.
- Once the multisystems management cost has been paid, it is very linear as the number of systems grows, since each system adds less than 0.5%. This is true for both the benchmarks and customer data.
- There is also a variable cost, which depends on the percentage of the total data being shared and the rate at which the shared data is accessed. This in turn determines the locking rate. Research has shown that customers typically share less than 50% of their data, and because they execute considerable application logic with each access, the locking rate is much lower than in the IBM benchmarks. This means that the benchmarks are typically at least twice as intensive as the customer workloads.

Figure 1 summarizes the results from four years of IBM benchmarking of different DB/DC subsystems and versions with the data from the analysis of a number of customers′ production RMF reports.

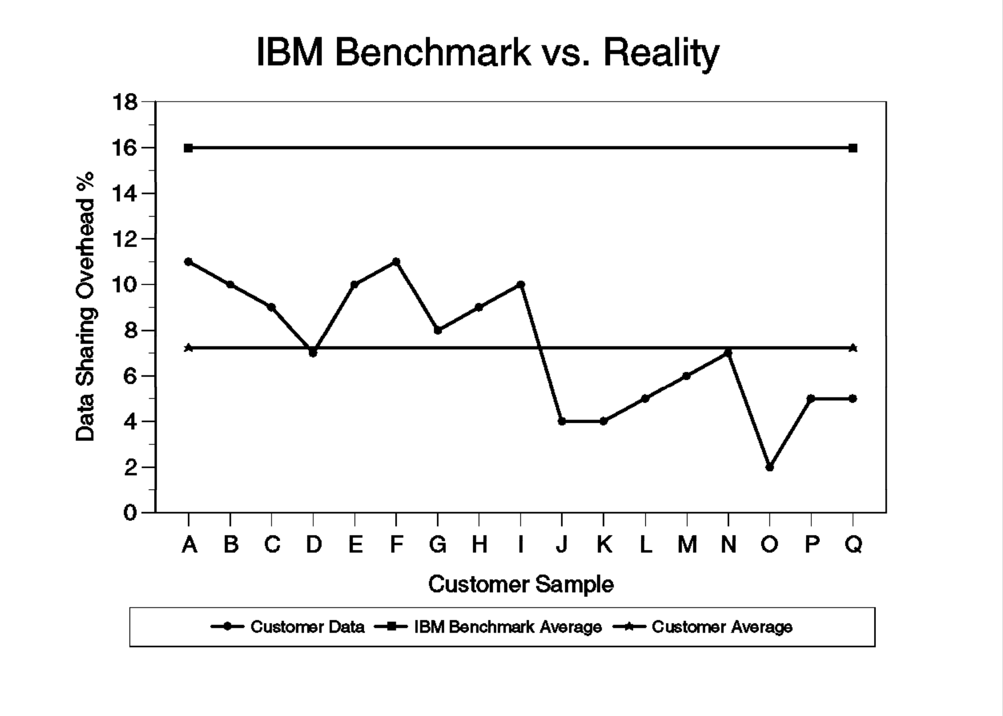


Figure 1. Customer Production Data Sharing Overhead

As can be seen, the customer analysis shows data sharing adds between 5% and 11% to the workload.

For a detailed look into the customer environments that make up the data in Figure 1, refer to Chapter 3, “Customer Data Sharing Experiences” on page 53.

1.3 Parallel Sysplex Functions and Performance

This section gives you a high-level view of Parallel Sysplex functions and their effect on performance. An understanding of these functions and the extent of their use is a critical part of evaluating the performance of the Parallel Sysplex.

1.3.1 Coupling Technology

Coupling technology extends the concept of the sysplex to up to 32 MVS Version 5 or OS/390 images while allowing all systems to share data. To provide this capability on a mixed platform of processor technologies with no application code impact, multisystems management and data sharing functions were significantly enhanced, with performance as a primary objective. Additional S/390 instructions, optimized for performance, have been architected. The coupling facility (CF) has been defined to provide hardware assistance to the management of global resources and workloads. Coupling links connect the Central Electronic Complexes (CECs) to the CFs, providing balanced bandwidth for years to come.

1.3.1.1 Coupling-Capable Hardware

The additional S/390 Central Processing Unit (CPU) instructions have been designed to enhance performance. They have been introduced to access the CF and the information provided by the CF to the OS/390 images. The following IBM hardware platforms support these instructions and are defined as coupling capable:

- All 9672 models from Generation 1 to Generation 5
- The 9021 711-based models
- The 9121 511-based models

1.3.1.2 Coupling Facility

The coupling facility facilitates system level functions in support of sharing data, distributing work, and balancing system resources within the sysplex by providing a means of managing multiple systems while preserving the integrity of the data.

The coupling facility provides a common memory for the sysplex that is dynamically partitioned to hold lock, cache, and list *structures.* These structures are typically used to hold status information required for intersystem coherency and provide a serialization mechanism for multiple systems. In addition, the cache structure can be used as a buffer for storing shared data with common read/write access.

A coupling facility can consist of one or more of the following platforms used in conjunction with the Coupling Facility Control Code (CFCC) Licensed Internal Code:

- An LPAR on the IBM 9674 or 9672-R06 standalone coupling facility
- An LPAR on any 9021 711-based and 9672 models, including those exploiting ICF engines on a 9672
- The Integrated Coupling Migration Facility (ICMF) in the case of a single coupling capable model

Since the CF became available, the CFCC µcode has been enhanced several times. The original CFCC code (CFLEVEL 0) contained all of the base functions needed to enable a Parallel Sysplex with the subsystem software levels of that time. Since then, CFLEVELs have delivered additional functional and performance enhancements:

- CFLEVEL 1 provided list structure enhancements, providing storage and *performance benefits* for the system logger.

  Also, dynamic structure reconfiguration was added to make recovery more granular by allowing structures to expand or contract without requiring OS/390 to deallocate the structure first.

  This level was also the vehicle for introducing single-mode CF link support, and improved tracing and debugging facilities.

- CFLEVEL 2 provided enhanced commands to register interest in a cache entry (register name list support). This support is aimed at providing *performance benefit* for DB2 data sharing, and to a lesser extent IMS data sharing, by reducing the number of trips to the CF needed at transaction commit points.

  Also, CICS/VSAM RLS benefits from the support and in fact had this level as the minimum required CFLEVEL. Support to unlock a list of locks held in a lock structure (unlock resource list) was added to *minimize the overhead* of releasing all locks individually.

  In addition, some known constraints were removed by allowing more connectors to a cache structure (up to 255) and by allowing more structures in a CF (up to 1023).

- CFLEVEL 3 primarily contains function for shared message queue support in IMS/ESA V6.1 through transition notification on subsidiary lists within a list structure (list structure sublist monitoring). IMS shared message queue support requires this minimum level of the code.

  Concurrent patch for CFCC on IBM 9672 G3 (9674C04) and IBM 9672 G4 (9674C05) hardware was added at this level.

- CFLEVEL 4 goes beyond its predecessor by adding additional function that benefits IMS in particular, including dynamic structure expansion and contraction (dynamic change of structure sizes).

- CFLEVEL 5 adds support for DB2 duplexing of cache structures. This support is a key customer requirement for DB2 continuous availability and is targeted for a future release of DB2.

1.3.1.3 Coupling Links

CF LPARS not using ICMF or the Internal Coupling (IC) channel (available in 1Q 1999) require coupling links to connect coupling-capable processors together. Three types of links exist for the following:

- The 250 MB/second Internal Coupling Bus (ICB) for distances up to 7 meters
- The 50 MB/second multimode links for distances up to 1 km
- The 100 MB/sec single mode links for distances up to 10 km

Note: The ICB link is classified as an *integrated* link, whereas the single and multimode links are called *external* links.

If ICMF is used in a single system configuration, the links are emulated, and no real links are required.

IC channels are internal links, from the memory of the sender LPAR to the memory of the receiver or CF LPAR. For this reason, they are capable of delivering very high performance (700 GB/second). In addition, unlike ICMF, where you can only use emulated links, using IC channels allows the use of internal, integrated and external CF links all to the same CF LPAR.

The IBM 9672 G4 models introduced HiPerLinks in 1997. These provide better performance through the enhanced link adapter technology. The IBM 9672 G3 models may also be upgraded to use HiPerLinks. The performance benefit of HiPerLinks are achieved through a higher degree of parallelism in the processing of CF requests. This in turn increases the link capacity.

In laboratory measurements, HiPerLinks have shown response time improvements ranging from 15% to 40% for asynchronous CF requests. The larger the data transfer, the greater the improvement. For synchronous CF requests, HiPerLinks have shown response time improvements of 10% to 15%.

On average, the link capacity may be improved by 20%, and data sharing overhead may be reduced by up to 10%. For example, if the overhead were 10%, it would drop to 9% with HiperLinks. Gains in response time achieved by upgrading to HiPerLinks are close to the response time gain that is achieved by replacing an IBM 9674-C02 with an IBM 9674-C04 CF. HiPerLinks permit XCF to use CF structures for signalling with performance equivalent to CTC connections.

The following XCF exploiters are likely to benefit from HiPerLinks:

- GRS (ring)
- CICS MRO
- SMSQ (especially for large messages)
- BatchPipes
- DB2 (especially when using 32 K buffers)
- VSAM RLS (with large CI sizes supported)
- CICS temporary store support
- VTAM high performance routing (HPR)

Note: With Hiperlinks, the nominal speed of the link is still 50 or 100 MB/sec, depending on whether single-mode or multimode fiber is used.

1.3.2 Multisystem Management

Multisystem management functions are performed by coupling facility exploiters to ensure optimal usage of resources and assist in balancing the workload across systems. Customers may require exploitation of new functions as they migrate to the Parallel Sysplex based on their current environment. A customer, for example, may have previously defined a CICS application in an MRO-type configuration and a base MVS 4.1 sysplex using CICS Intersystem Communications (ISC). Communications, routing, and load balancing functions are needed to effectively manage the Parallel Sysplex.

1.3.2.1 Communications

A shared intermediate memory scheme is used to communicate between systems in the sysplex. By indicating through a central control that a given system has a message to deliver to another system, a "star" of communication paths is created. The hub of the "star" is the coupling facility (see Figure 2).

Other communication options tend to break down as more systems are added. The any-to-any mechanism for communication, such as channel-to-channel protocols (CTCs), may prove to be too complex to manage when the number of connections increase. A "ring" alternative leads to continually increasing performance costs.

The coupling facility, in conjunction with the cross-system coupling facility (XCF) component of MVS, provides the standard communication mechanism for MVS system applications and reduces the complexity of system management compared to CTCs.

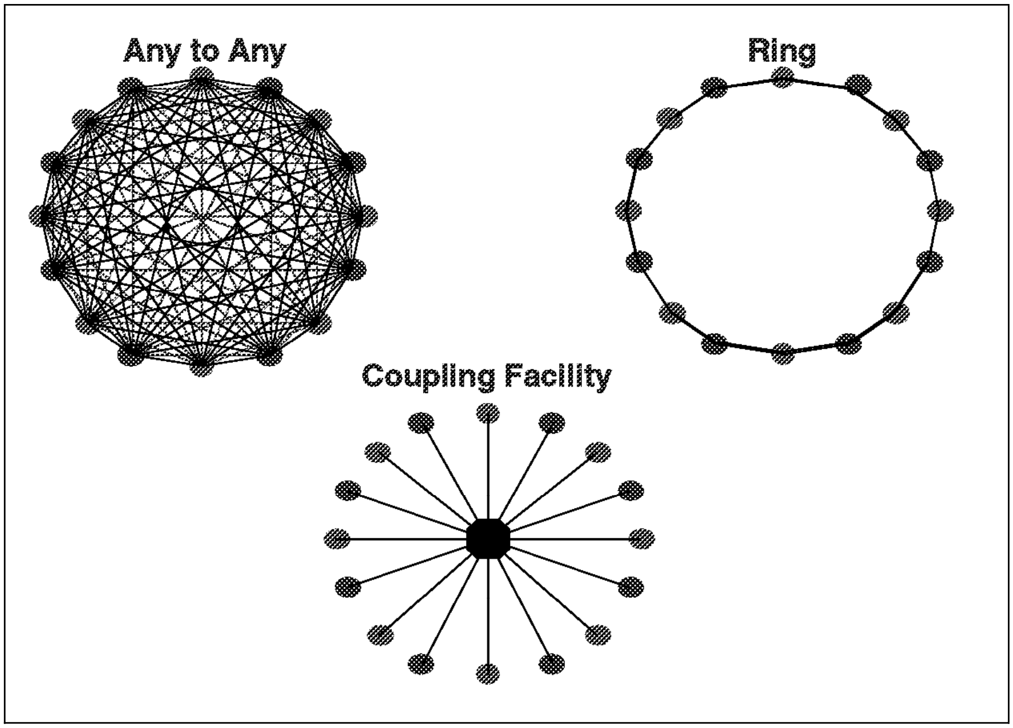


Figure 2. Communication Options

In addition to communication between systems using XCF, information required by all systems can be placed in the coupling facility for central access. Some examples are Job Entry Subsystem 2 (JES2) checkpoint data and Resource Access Control Facility (RACF) profiles.

1.3.2.2 Routing

Functions in the Parallel Sysplex, combined with CICS/VS Version 4, provide an extremely flexible method of routing transactions. With previous versions of CICS, the customer must define to CICS the *connections* between the Terminal Owning Region (TOR) and the Application Owning Regions (AORs). These connections use a cross-memory technique if the TOR and AOR are on the same MVS system, but they use a VTAM LU6.2 connection if the AOR is not on the same MVS. The cross-memory technique is considerably more efficient than the VTAM connection.

Starting with CICS Version 4, connections defined as cross-memory are supported if the TORs and AORs are in the same *sysplex*. This means that the customer gets the flexibility to run the various CICS regions anywhere in the Parallel Sysplex, with no changes to the CICS specifications or JCL. Connections that are on the same MVS will use the same cross-memory technique as before; connections to other systems in the sysplex use XCF. Using XCF for this connection yields a significant performance advantage over the previously required VTAM link.

Routing is also used to preserve an end-user view of continuous availability. In the CICS environment, improved availability is achieved by having transactions routed to cloned AORs of a given AOR that has been quiesced for a scheduled outage.

1.3.2.3 Load Balancing

Parallel processing in a sysplex is the ability to simultaneously process a mixture of work with various characteristics across many systems, without incurring changes or inhibiting access to data. This work could be defined either as many small but distinct units of work (such as transactions), or as a larger unit of work (such as a batch job). The choice of CPU size becomes a function of single work unit CPU requirements as opposed to application requirements. Reduced end-user response times or increased throughput result from parallel processing.

There are two basic workload distribution and data access methodologies that are used in the industry for parallel implementations: partitioned data and shared data. In a partitioned data implementation, the database and the workload are divided between the various systems in such a way that each part of the data is accessed and updated by only one system. This is the approach that most vendors in the industry have used to implement parallel processing. It requires human-intensive skills to closely monitor the statically defined workload distribution and to frequently alter the data organization to achieve a proper balance of resources.

The parallelism of the S/390 Parallel Sysplex offers a shared data approach to online transaction processing (OLTP), batch, and query processing by giving all processors equal access to all data managed by IMS/DBCTL and DB2 and VSAM Record Level Sharing for CICS. Dynamic workload balancing can be used to spread the work according to processor load rather than data location. Load balancing is achieved initially during user logons and, later, with dynamic transaction routing driven by the transaction managers. This allows all requests to have access to every part of shared data in the sysplex, thus utilizing resources more efficiently. Most transactions will run in the S/390 Parallel Sysplex environment without end-user changes.

1.3.3 Data Sharing

Figure 3 on page 8 illustrates the components of a system in a Parallel Sysplex. The multisystems management components were discussed in the previous section. The other types of Parallel Sysplex functions are those associated with running a database manager that supports applications that update the same database from multiple MVS systems. We refer to this as *data sharing*. A customer can choose to share databases across systems based on business and application requirements.

The percentage of applications that use shared databases is an important consideration in evaluating the costs of running a Parallel Sysplex.

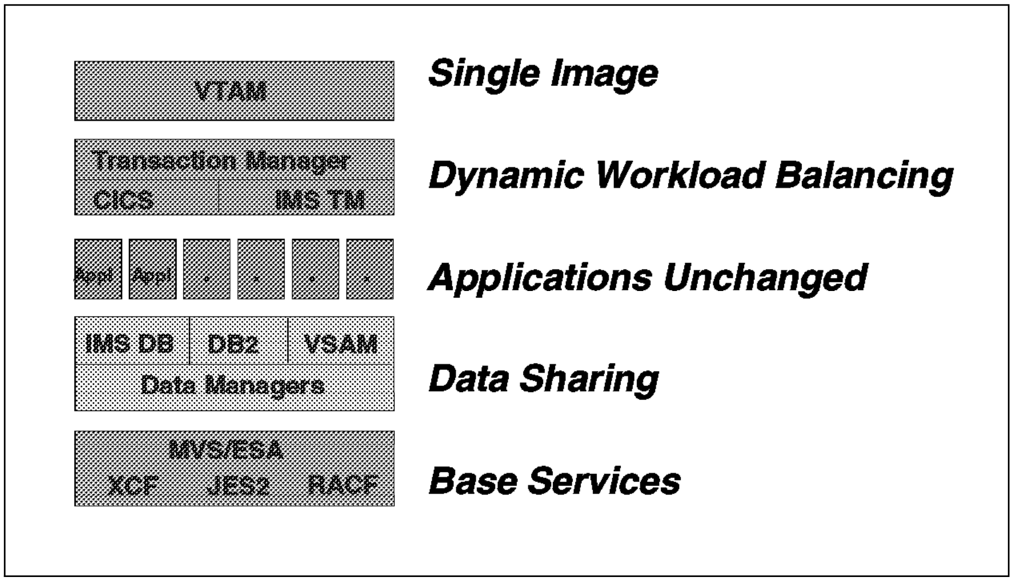


Figure 3. Multisystems Management and Data Sharing

1.3.3.1 Locking

The lock structure supplies shared and exclusive locking capability for serialization of shared resources down to a very small unit of data. The coupling facility provides the mechanism to grant locks as they are requested. The transaction will be suspended when the lock cannot be granted because of contention for the resource.

In a single system environment, locking occurs within the system. In a Parallel Sysplex, global locking occurs via a synchronous access to the lock structure in the coupling facility (see Figure 4 on page 9). The *speed* and *rate* of these accesses have an effect on the overall sysplex performance. An IMS Resource Lock Manager (IRLM) lock structure is used in conjunction with the new optimized S/390 instructions to maintain data integrity in the IMS-TM/DB2 and CICS/DBCTL environments. The single, central lock structure for CICS/VSAM Record Level Sharing (RLS), IGWLOCK00, is provided via the lock manager internal to the DFSMS product, and is also used in conjunction with the new optimized S/390 instructions to maintain data integrity for the CICS/VSAM data sharing environment. This lock structure provides sysplex-wide locking at a record level; that is, control interval (CI) locking is not used.

When a sysplex using GRS STAR issues an ENQ, DEQ, or RESERVE request for a global resource, the request will be converted by global resource serialization to a lock request against the ISGLOCK lock structure. Global resource serialization uses the ISGLOCK lock structure to coordinate the requests to ensure proper resource serialization across all systems in the complex.

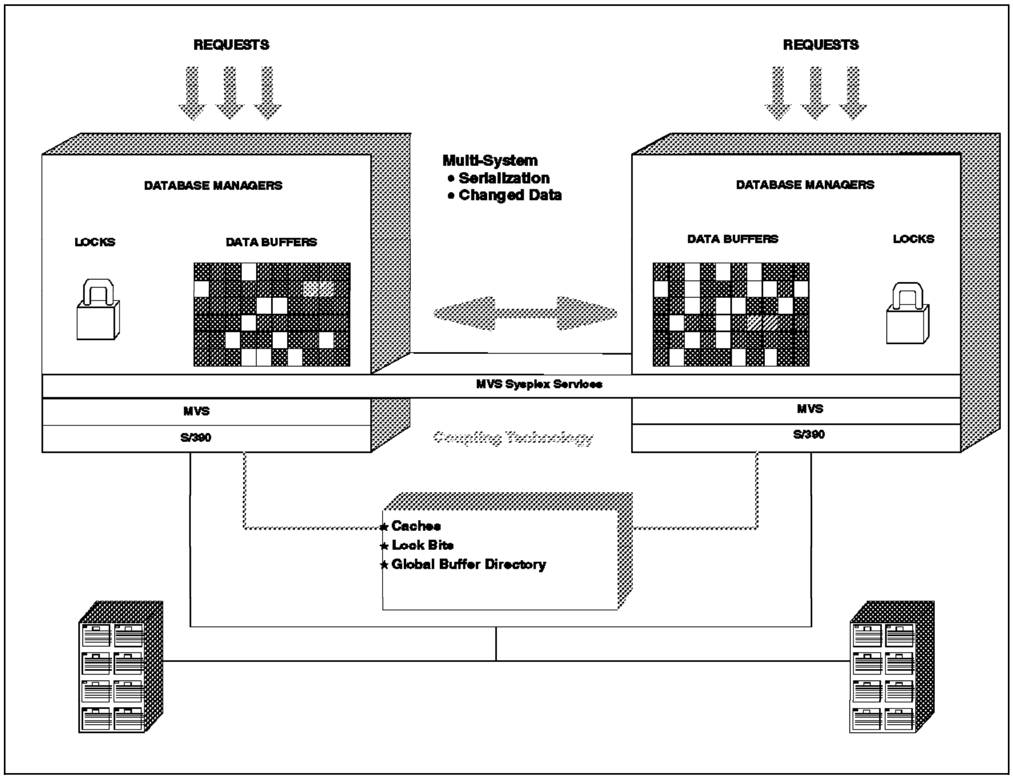


Figure 4. Data Sharing in a Parallel Sysplex

1.3.3.2 Caching Data in the Coupling Facility

Some data sharing subsystems cache data in the coupling facility. For example, both RACF and CICS/VSAM RLS use the coupling facility as a store-through cache mechanism. Data is written to the coupling facility and to DASD. The coupling facility provides high-speed access and eliminates the necessity of using hardware RESERVE/RELEASE for serialization.

In DB2 Version 4, a coupling facility cache structure is used to store changed data when multiple DB2 subsystems have read/write interest in a pageset. Optionally, unchanged data can also be stored in the coupling facility. DB2 uses buffer registration and invalidation to coordinate access to shared read/write data. In IMS Version 6, a coupling facility cache structure may be used for storing OSAM or shared VSO DEDB data.

1.3.3.3 Buffer Invalidation

Cache structures supply a mechanism called *buffer invalidation* to ensure consistency of cached data. A global buffer directory in the CF maintains which systems have a local copy of the data in their buffer pools. The CF can then request changes to the status flags to indicate that buffer contents are invalid as various systems update the buffers (see Figure 4).

Data integrity is preserved by requiring the buffers to be registered, by creating a directory entry in the CF, and by having the systems check a local copy of the status of the data they wish to use. The local copies of the status are kept in bit vectors located in the Hardware Storage Area (HSA) and are updated by the coupling facility without MVS intervention. New high-performance instructions allow MVS to interrogate the HSA bit vector contents without the need to communicate with the coupling facility.

If the contents of an IMS buffer are needed and have been invalidated by another system, reading the data again from DASD and registering it in the CF is required. Two cache structures, used as directories, are defined for OSAM and VSAM buffer invalidation and provide data integrity for the IMS databases. In IMS Version 6, however, coupling facility cache structures may be used for storing OSAM and VSO DEDB data, thus removing the requirement to read the data again from DASD.

If the database is DB2 and an application requests data on a system whose buffer has been invalidated, the local DB2 can read the data in from the Group Buffer Pool structure in the coupling facility, avoiding the additional I/O to DASD. DB2 uses a store-in cache structure for buffers that have been changed.

1.4 Parallel Sysplex Costs

As stated in 1.3.2.3, “Load Balancing” on page 7, IBM′s Parallel Sysplex strategy provides a shared data approach to parallel processing. With dynamic workload balancing, I/O contention and workload skew are avoided, while more consistent response times and higher processor utilizations are achieved. The resulting “single system image” management of resources is more efficient than partitioned data implementations. So, the personnel and performance costs associated with a Parallel Sysplex implementation are smaller than many of the alternatives.[1]

Multisystems management and data sharing represent two different types of functions provided by the Parallel Sysplex. Each type has a slightly different effect on performance based on the implementation of Parallel Sysplex. The customer data, for various workloads, as discussed in the introduction, is analyzed in Chapter 3, “Customer Data Sharing Experiences” on page 53. The IBM benchmark data Appendix A, “CICS/DBCTL Workload Details and Coupling Efficiency Details” on page 171 provides a detailed analysis of performance costs for the CICS/DBCTL benchmark. Appendix B, “IMS-TM/DB2 Workload Details and Coupling Efficiency Details” on page 177 provides a detailed analysis of performance costs for the IMS-TM/DB2 benchmark. Appendix C, “CICS/VSAM Workload Details and Coupling Efficiency Details” on page 183 provides a detailed analysis of performance costs for the CICS/VSAM benchmark. Chapter 11, “IMS/TM Version 6 Performance Study” on page 127 provides a detailed analysis of performance costs for the IMS shared message queue (SMQ) benchmark. General observations based on the measurements are summarized in this section. Detailed measurement results and analysis of these benchmarks are summarized in the next chapters.

[1] Additional reading for data sharing approaches that may be of interest to the reader are the following articles:

- Rahm, Erhard, Empirical performance evaluation of concurrency and coherency control protocols for database sharing systems, ACM Transactions on Database Systems, Vol.18, No.2, June 1993, pages 333-377
- Yu, P.S., et al, On coupling multisystems through data sharing, Proc. IEEE 75, 5 (1987), pages 573-587

1.4.1 Multisystems Management Costs

Each MVS image in a sysplex incurs some costs related to the fact that it is in a sysplex. The costs associated with managing a base sysplex or a Parallel Sysplex depend on the customer′s current environment, but are typically 3% to 4% of total capacity, as shown in Figure 5.

For example, in the case of a customer migrating from an MVS 4.3 sysplex to a Parallel Sysplex, the initial multisystems management costs have already been incurred in the migration to a MVS 4.3 sysplex. Thus, there may be no additional performance costs. Conversely, a migration from a single system to a Parallel Sysplex typically costs 3% to 4% of a customer′s total capacity. This is similar to the version-to-version migration cost of, for example, migrating from MVS Version 4 to Version 5.

Once a Parallel Sysplex is implemented, adding more systems increases the multisystems management cost by less than 0.5%.

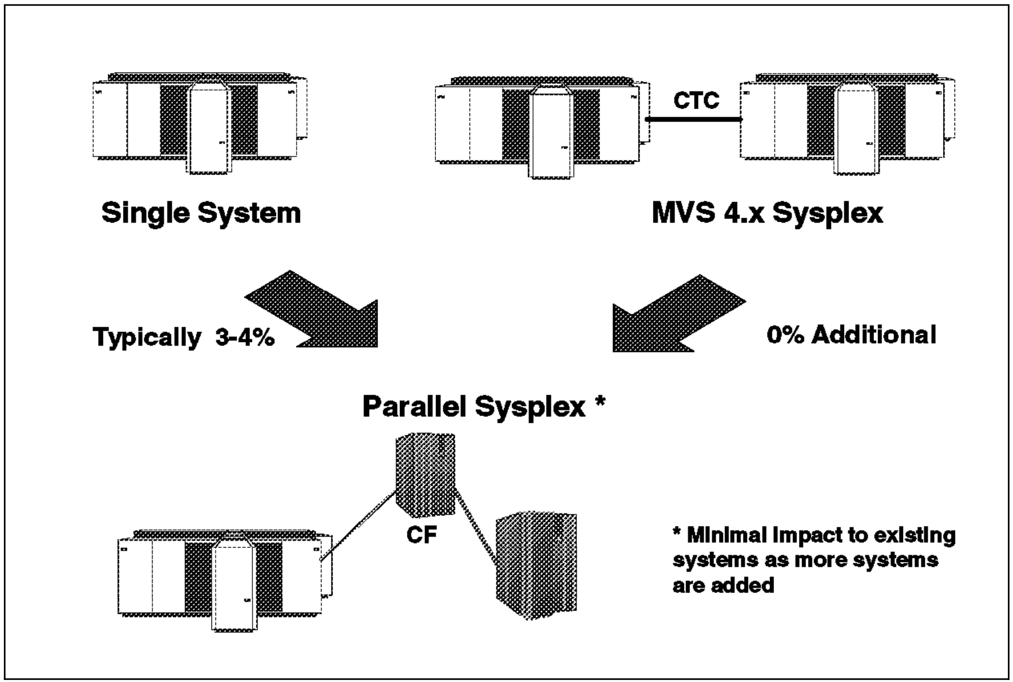


Figure 5. Multisystems Management and Data Sharing

1.4.2 Cost of Data Sharing

The performance costs that arise from exercising the Parallel Sysplex data sharing functions are variable. Hardware configurations and workload characteristics, such as the amount of sharing required by applications, are all factors in the evaluation of the performance of customer environments. Business considerations will help determine the best configuration for a Parallel Sysplex to gain the most value from data sharing, availability, and overall capacity and workload management.

The access rates to the coupling facility for locking, caching, and local buffer invalidation are a function of the amount of work requiring shared data and the intensity of the access to the shared data by the workloads. Some applications may be totally contained for business reasons, while others may be spread across multiple systems to maximize availability. This may translate into some transactions needing shared data while others may not.

The data sharing cost also varies based on the type of processors participating in data sharing. The difference in performance is primarily driven by the differences in uniprocessor power of the various systems and the coupling facility. Faster systems, relative to the coupling facility, effectively lose more potential capacity during synchronous processing of requests by the coupling facility than do systems of the same speed as the CF. As a general recommendation, therefore, it is best to keep the technology and hence the performance of the coupling facility and processor the same to give optimal performance.

Finally, the cost to a system initially joining a Parallel Sysplex is greater than the subsequent performance effects it may incur when additional systems are joining the Parallel Sysplex. In fact, the cost of adding each additional system into the Parallel Sysplex is about 0.5%. This minimal incremental cost allows the Parallel Sysplex to achieve maximum growth with minimal impact as more systems are added to the Parallel Sysplex.

1.4.3 Additional Workload Considerations

In most installations, a particular processor runs several different types of work. Some of these applications will want to take advantage of the new coupling technology; others will not. The percentage of the total available processing power that is being used for applications sharing data is a good indicator of how much the new functions will be exercised.

The measurements in this document represent the most stressful case of *100% data sharing*, where all the workload accesses data in shared databases. For most environments, only part of a system′s overall resources are used for the data sharing functions. Batch, system monitoring, TSO, and read-only database work usually coexist with OLTP processing against shared data. The net effect is that the access rates to the coupling facility may be much less than any of the benchmarks measured in this redbook.

Large transactions may not necessarily have high cost running in a shared database environment. It is the *intensity* with which the application accesses shared data that dictates the costs when Parallel Sysplex functions are added.

Depending on the workload, additional I/O requests may be generated as the number of CECs in the Parallel Sysplex increase. For example, when an IMS system updates its local copy of shared data, local copies of that data on other systems are invalidated. Thus, the next future reference to the data results in I/O to retrieve the updated copy. The magnitude of this effect will depend on the customer′s particular database reference pattern. Enlarging the local VSAM and OSAM buffer pools may offset some of the I/O growth. Remember, with IMS Version 6, the OSAM buffers may be placed in the coupling facility cache structure. Note that enlarging a buffer pool will require a corresponding increase in the size of the buffer validation structure as well.

DB2 uses Group Buffer Pools to cache changed data in the coupling facility. DB2 often retrieves this data from the coupling facility instead of DASD. Therefore, it is less likely that I/O will increase significantly as the number of DB2 subsystems in the sysplex increases.

1.5 Measurement Metrics

The cost of coupling varies with the sysplex configuration (including number of systems and processor speeds), and data sharing workload characteristics (such as amount of data sharing and transaction sizes). These are the principal elements in the evaluation of the Parallel Sysplex performance. A consistent methodology and common metrics are needed to compare measurements. It is also important to remember that transactions are workload-unique and should not be compared to other types of transactions.

1.5.1 Throughput

From the smallest uniprocessor to the largest collection of systems working in parallel, the basic law of performance applies: the fundamental measure of capacity can be expressed as the ratio of how many units of work can be completed to how many units of time they take to complete. External Throughput Rate (ETR) measures units of work divided by elapsed time to complete the work. ETR is particularly useful in evaluating the performance of batch jobs and long-running transactions. For many measurements, processor utilizations, defined as the ratio of processor busy time to elapsed time, will vary. A way to remove the effects of slightly unequal utilizations when comparing measurements is to normalize the measured ETR to full processor utilization. The resulting value is the Internal Throughput Rate (ITR) and can be calculated as units of work divided by CPU busy time. In OLTP environments, it is often expressed as number of transactions per CPU second.

The Quick Sizer, described in Appendix D, “Capacity Planning Tools” on page 189, may be used to compare the performance of Parallel Sysplex on different systems.

1.5.2 Internal Response Time

A second metric typically tracked is the transaction′s internal response time, which is affected by CPU busy times and I/O times. The response time or elapsed time of a transaction or batch job consists of an I/O component and a CPU component.

I/O Component Since the I/O architecture is the same for all S/390, processors and assuming the I/O layout would remain unchanged when moving a particular type of to a Parallel Sysplex, the I/O component would remain relatively unchanged. In cases with similar work running on multiple systems, I/O operations could be delayed when accessing shared data.

CPU Component The CPU time and CPU queueing time of a transaction are a function of effective single engine processor power.

Network time and other components of end-user response times will change to the extent that the CPU components change. The average transaction response time was less than a quarter of a second for all our measured OLTP environments.

1.5.3 Coupling Efficiency

In a coupled environment, the measure of effective throughput compared to single system capacity is often used to describe the multisystem management and data sharing cost that may be incurred. We define coupling efficiency as the effective throughput of the Parallel Sysplex divided by the sum of uncoupled throughputs of all the individual systems that comprise the sysplex.

In mixed workload environments and configurations, the coupling costs associated with each system in the sysplex may differ and the concept of coupling efficiency of the Parallel Sysplex becomes a function of the various efficiencies of the systems that make up the Parallel Sysplex.

1.6 Parallel Sysplex Performance Benefits

The measurement environments chosen for this book demonstrate the capacity and scalability aspects of the S/390 Parallel Sysplex. The measurements in the following chapters use 9021 711-based models and 9672 R models coupled together with 9674 coupling facilities. The measurement workloads, configured to take full advantage of data sharing capabilities, utilize either an IMS transaction manager or a CICS transaction manager, and IMS DBCTL databases derived from the standard CICS and IMS workloads used for the LSPR publication, DB2 databases, or VSAM data sets. Additional information on the hardware configuration, software levels and workloads can be found in each measurement chapter. This section focuses on the general performance characteristics of the Parallel Sysplex.

1.6.1 Availability

In a Parallel Sysplex environment, if a system fails, the remaining systems maintain access to all the shared data, and the workload from the failed system is routed automatically to the other systems. The reserve capacity needed in case of a system failure can be spread out evenly. The benefits of this are that systems can run at higher utilization, and users do not experience significant performance impacts when the work of a failed system is redistributed.

MVS images can be dynamically added to a properly configured coupling facility. Customers can then grow horizontally without disruption. Multiple coupling facilities with redundant links can also be planned so that outages are transparent to the end user.

1.6.2 Intermixing of Technology Levels

Parallel Sysplex supports multiple technology levels with nearly 10-fold variability in engine size. Customers may therefore have many different technologies in their Parallel Sysplex. This requires a different approach to traditional capacity planning techniques. The different size of building block allows for horizontal and vertical growth with maximum flexibility. Exceptional scalability makes horizontal growth an attractive option.

1.6.3 Parallel Sysplex Scalability

Figure 6 depicts what is the effective system capacity as system size continues to increase. The *Ideal* line shows what the performance characteristics would be in an environment that does not need to share any information across participating members as new members are added. On this line, the full capacity of each system is added to the existing system.

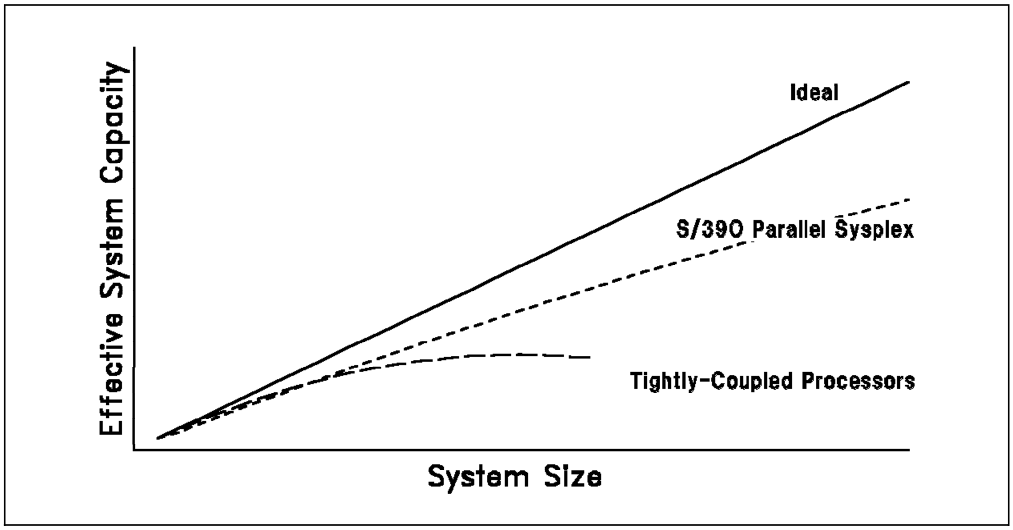


Figure 6. Parallel Sysplex Scalability

With the *Tightly-Coupled Processors* line, the point of diminishing returns occurs rather quickly due to current hardware and software designs. This line is based on current tightly coupled processor ratios, as well as an understanding of current hardware and software internals. This line tends to flatten and eventually degrade as more CPs are tied together in a tightly coupled configuration. Improving this line would require additional hardware and software research and development costs, which would raise the price of the tightly coupled processor.

The *S/390 Parallel Sysplex* line shows that the design and use of data sharing across multiple CECs is such that additional systems may be added with minimal impact. For the 100% data sharing measurements detailed in Chapter 4, “CICS/DBCTL Data Sharing Scalability Study” on page 65, Chapter 5, “IMS-TM/DB2 Data Sharing Scalability Study” on page 73, and Chapter 6, “VSAM RLS Data Sharing Scalability Study” on page 81, less than half a percent degradation was observed. For the IMS shared message queue environment, measurements showed a slight improvement in ITR when adding more systems. This can be seen in detail in 11.1, “Shared Message Queue (SMQ) Study” on page 127. Although the building block for the Parallel Sysplex is a tightly coupled system, joining multiple systems together through the use of a coupling facility allows the S/390 sysplex to achieve a new level of effective single system image capacity.

1.6.3.1 Industry Leading Scalability

The results for the 9672 environment indicate that adding systems to the Parallel Sysplex does not generate significant overhead. Thus, nearly the full power of the additional systems may be applied to transaction processing. This excellent scalability characteristic of the S/390 Parallel Sysplex is in contrast to previous message-passing parallel architectures.

Figure 7 on page 16 is an example of the scalability when the system or Parallel Sysplex hardware capacity increases in size. The 9021-821 machine has two CPs, while the 9021-982 machine has eight CPs. The 9021 machines are not doing data sharing and utilize the tightly coupled design. Based on LSPR data for the IMS/DS workload, the ITR ratio between the two is 3.4.

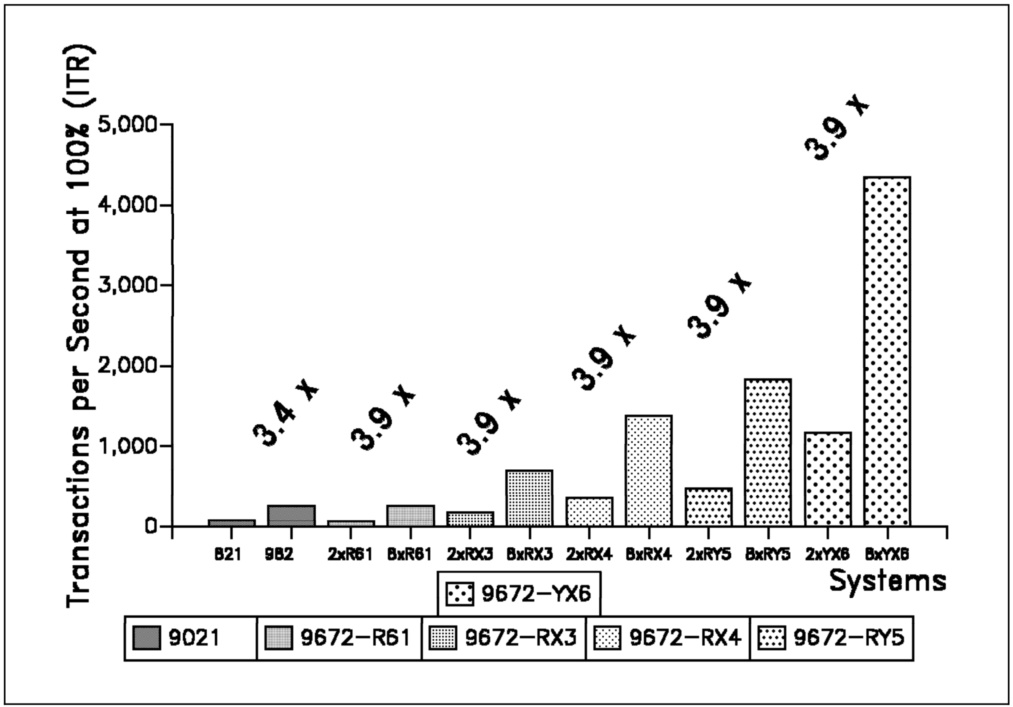


Figure 7. Industry-Leading Scalability

Comparing two 9672-R61 CECs sharing data using the Parallel Sysplex design to eight 9672-R61 CECs sharing data using the Parallel Sysplex design yields an ITR ratio of 3.9. Comparing two 9672-RX3 CECs sharing data using the Parallel Sysplex design to eight 9672-RX3 CECs sharing the Parallel Sysplex design yields the same ITR ratio.

In addition, comparing two 9672-RX4 CECs sharing data using the Parallel Sysplex design to eight 9672-RX4 CECs sharing the Parallel Sysplex design yields the same ITR ratio. Furthermore, if you compare other processors in the same manner, the ratio will be the same. Similar results are seen with the CICS/VSAM RLS and the IMS-TM/DB2 environments. This shows that the Parallel Sysplex delivers capacity more efficiently than the tightly coupled design.

1.6.3.2 Scalability by Design

Why is adding a CEC to a S/390 Parallel Sysplex much more efficient than adding an engine to a tightly coupled MP? The answer to this question focuses on two areas:

- Differences exist between a hardware architecture that follows certain rules with respect to atomicity, instruction sequencing, and control block level locking, and a database manager that has an understanding of the data and the relationships between data.

  The hardware architecture must always use the same rules to access data, while the database manager can be designed to have significant flexibility in serialization techniques, locking levels, buffer management algorithms, and so forth.

As a result, the granularity of managing data integrity is much finer on a tightly-coupled MP than on a Parallel Sysplex. This results in more locking and cross-invalidation activity on the MP.

- One copy of MVS must manage the larger workload on an MP while an "extra" copy of MVS comes with the added CEC on a Parallel Sysplex. Each MVS will be less stressed by workload size in a Parallel Sysplex than a single MVS trying to manage the entire workload.

For a tightly coupled MP, when an engine is added and more work is processed, what happens to MVS and the original engines on the system?

- I/O interrupts and task switches increase. This results in increased software overhead and lowers the efficiency of hardware caches, leading to reduced effective capacity.
- Number of locks per second on commonly used control blocks, such as control blocks for address spaces, storage frames, and others, goes up. This causes lock contention to increase (often by percents) which can lead to significant increases in "spinning" for locks.
- Queues of control blocks get longer, leading to increased path length to manage them.

For the Parallel Sysplex, when a CEC is added and more work is processed, what happens to the original CECs in the sysplex?

- Locking rate for I/O blocks stays the same. Lock contention may increase slightly since more transactions may request the same lock, but this increase is generally measured in hundredths of a percent.
- I/O rate may go up slightly due to local buffers that were invalidated from updates on the new CEC.
- Transaction routing may increase to balance the new workload across the sysplex (of course, this can usually be minimized with tuning, as was done in our benchmarks).
- Message traffic among the CECs may increase slightly for GRS and MVS workload manager.

Although the S/390 uses messaging for locking and buffer coherency, it does so far more effectively than a software-driven I/O message scheme. Specifically:

- Lock requests can be granted by the CF upon the CF′s checking of the lock table for contention and that no interrogation of the other connected images is required.
- If lock contention is detected, messages are sent only to the systems that are needed to resolve the contention.
- CF buffer cross-interrogation signals are sent in parallel and do not cause a S/390 interruption on the targeted MVS image, avoiding a context switch in the cache and the Translation Lookaside Buffer (TLB).
- A buffer manager can, after acquiring a lock, assess the validity of a buffer frame with no communication to the coupling facility. Since buffer checking is done very frequently, this architectural aspect has very significant performance advantages.

This last point about message traffic is a key distinction that separates the IBM S/390 Parallel Sysplex implementation (specifically, the use of the coupling facility for locking and data) from other clustering architectures that rely on messaging for locking or I/O access. These latter clusters are prone to larger degradations as more nodes (CECs) are added and messaging among the nodes increases significantly.

1.6.4 Largest Single System Image

The S/390 coupling technology discussed in this document can provide capacity growth beyond the configurations measured. Up to 32 MVS or OS/390 images on ES/9000 and 9672 models can be coupled together using multiple 9674 coupling facilities.

The 9672 environment was chosen to illustrate a customer migration from a current 9021 model to a 9672 Parallel Sysplex. Figure 8 provides a projected view of capacity growth as the numbers of coupled 9672 models increase. Note that the capacity extends far beyond the capacity available on the largest ES/9000 models.

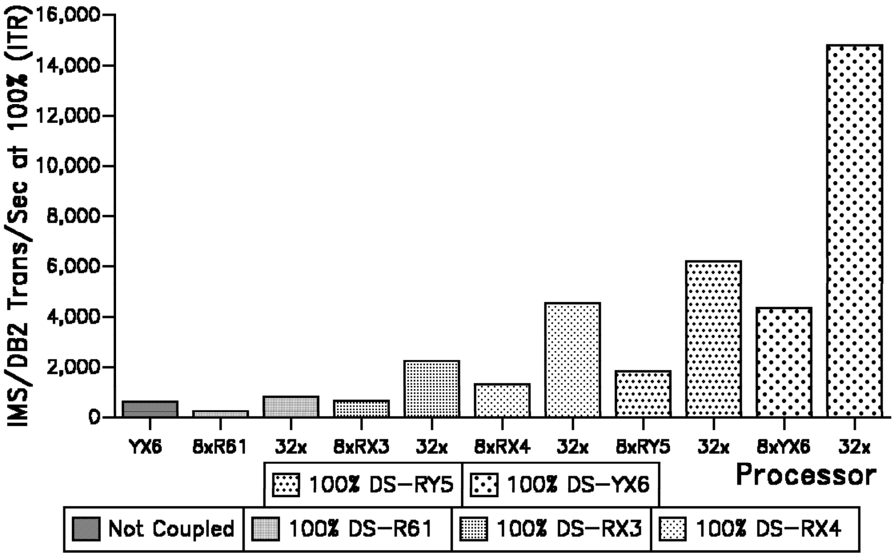


Figure 8. Largest Single System Image

This graph illustrates that the S/390 Parallel Sysplex provides the largest single system image in today′s online transaction processing marketplace. It depicts IBM′s largest single image processor, the 9672-YX6 (10 CPs joined together in a tightly coupled environment). By comparing the IMS-TM/DB2 transaction per second of a 9021-YX6 to a 9672 8-way sysplex and multiples of the 9672 8-way sysplex, the Parallel Sysplex configurations show that a single image can still be maintained above and beyond current tightly coupled designs. Similar results can be expected for the CICS/VSAM and CICS/DBCTL environments. This serves to illustrate the growth potential of the S/390 Parallel Sysplex, as well as show the superior linear growth that is exhibited by the 9672 Parallel Sysplex. Note that the Parallel Sysplex results in this chart assume that 100% of the workload running on the 9672 is sharing data with the other MVS systems.

1.7 Summary

We have discussed the Parallel Sysplex from many aspects throughout this chapter including the following:

1. The coupling technology that has performance implications for multisystem management as well as data sharing. These performance effects, for the most part, will be *less than 10% of the CPU capacity* within the complex, as the customer data has shown.

2. Multiple migration choices available to customers as they exploit the parallel architecture and CMOS processors and some of the effects that need to be understood through the use of capacity planning tools.

3. Parallel Sysplex measurements that expand the range of the S/390 line of tightly coupled processors to the Parallel Sysplex environment, thus providing industry leading scalability and the “Largest Single System Image.”

Chapter 2. Tuning Recommendations

In the process of planning, tuning, and measuring the various experiments described in this document, we did several things to improve the performance of our systems. This information is collected in this chapter. Additional examples from other test systems have also been included. It is not a comprehensive list of all possible things that can be done to tune a Parallel Sysplex. Rather, it is a set of recommendations to make your initial tuning efforts a bit easier.

2.1 MVS Tuning

We start with the tuning changes we made to MVS to enable the Parallel Sysplex environment. Some of the changes were a direct result of combining single systems into a sysplex. Other changes were needed when we increased the size of the sysplex.

2.1.1 Placement of Couple Data Sets

As data sets are shared by an increasing number of systems in the sysplex, minimizing I/O response time to those data sets becomes more critical.

Couple data sets (CDS) are shared data sets used to coordinate activities across the sysplex. Non-volatile data (that is, data which must be preserved across an IPL) is written to these data sets. For recovery, each CDS consists of a primary and alternate data set. An update to the couple data set is not complete until both the primary and the alternate data sets have been updated. Therefore, a delay to the alternate serves to delay the primary as well.

Long response times from the couple data sets can result in delays during initialization and recovery, slowdowns in sysplex communication, and even GRS ring disruptions.

The following things can be done to improve response time to the couple data sets:

- CDSs should be placed on a volume behind a separate cached control unit with the DASD fast write feature. This is highly desirable for all size sysplexes, especially those with greater than 12 systems. This recommendation is primarily for the sysplex CDS and the coupling facility resource management (CFRM) CDS, though it certainly does not hurt to have the other types on such a device as well (which they would be if the following recommendation is followed).
- CDSs should not be placed on a volume subject to reserve/release contention, or significant I/O contention from other, non-CDS related sources, even if that I/O contention is periodic (that is, comes in high, though infrequent, bursts).
- Another way to improve response time is judicious placement of the couple data set so different data sets on the same volume are accessed at different points in the process.
  - The sysplex CDS is updated during changes in the configuration (IPL, partitioning, joining or leaving a group). It is also accessed for system status updates, which occur frequently.

– The coupling facility resource management (CFRM) CDS contains the policy for allocating coupling facility resources. It is referenced whenever the connection to a structure is established or dropped, during rebuild and recovery processing, and when display commands and queries are issued. Although these events do not occur that often, when they do occur, the access rate to this CDS can be fairly high.

– The workload manager (WLM) CDS contains the MVS WLM policy. It is referenced at IPL and during policy changes.

– The sysplex failure management (SFM) CDS defines recovery policies. It is accessed during initialization, policy changes, and failure processing. The number of accesses during failure processing can be sizeable.

– The automatic restart management (ARM) CDS describes how MVS should manage restarts for specific batch jobs and started tasks that are registered as elements of ARM. It is accessed during initialization, policy changes, and failure processing.

– The system logger (LOGR) CDS describes log stream or structure definitions. It is accessed during initialization, policy changes, and failure processing.

See *Setting up a Sysplex* for information about setting up these various CDSs.

Based on the frequency of accesses to these data sets, we recommend the following data set placement:

- All primary CDSs can be placed on the same volume, and all alternate CDSs can be placed on the same (different from primary) volume, with one exception:. The primary sysplex CDS should be on a different volume than the primary coupling facility resource management (CFRM) CDS. A sample configuration is shown:

Volume A	**Volume B**
Primary sysplex CDS	Alternate sysplex CDS
Alternate CFRM CDS	Primary CFRM CDS
Primary SFM CDS	Alternate SFM CDS
Primary WLM CDS	Alternate WLM CDS
Primary ARM CDS	Alternate ARM CDS
Primary LOGR CDS	Alternate LOGR CDS

For a large sysplex or one with aggressive recovery criteria, you may want to place the primary and alternate sysplex couple data sets and primary and alternate CFRM data sets on four separate volumes.

If CICS is using the logger function and there are several CICS address spaces accessing the LOGR CDS during initialization and recovery, you may want to put these couple data sets on separate volumes.

2.1.2 SRM Changes

We made two changes to MVS that affected SRM for this environment.

2.1.2.1 Transaction Response Times

To study the response time of the transactions for the workloads using CICS as the transaction manager, we wanted to track the response time of the transactions in the CICS TOR. These transaction times include the time spent in the CICS AOR (on the local or remote system) and thus represent the total internal response time for the transaction.

To obtain these response times, we had to change the SRM parmlib members IEAIPSxx and IEAICSxx to place the TOR transactions in their own report performance group. We used the same parmlib members on all the systems. By assigning the TORs to the same performance group on each system, we could look at the same RPGN on each system to get response times.

```
SUBSYS=STC,PGN=23
 TRXNAME=CICSDTA1(1),PGN=11,RPGN=400        /* CICS TOR on SY#A    */
 TRXNAME=CICSDTB1(1),PGN=11,RPGN=400        /* CICS TOR on SY#B    */
```

We also had to update two parameters in the CICS system initialization table, DFHSIT. We set MN=ON to enable CICS monitoring and MNEVE=ON to activate SYSEVENT recording. The SIT table is described in the *CICS/ESA V4R1 System Definition Guide*, SC33-1164 for CICS/ESA Version 4 Release 1 and in *CICS Transaction Server for OS/390 V1R1 System Definition Guide*, SC33-1682 for CICS Transaction Server.

With these changes, we could use the RMF workload manager report to observe the CICS TOR response time.

2.1.2.2 Dispatching Priorities

The experiments described in this document were run with MVS workload manager (WLM) in compatibility mode (MODE=COMPAT).

We had to ensure that the address spaces that managed multiple transactions had higher dispatching priorities than those address spaces that were running single tasks.

As an example, for the CICS/DBCTL workload, we used the following priority scheme:

1. VTAM, CPSM, IRLM
2. DBRC, CICS
3. IMS Control Regions and DLISAS

For the IMS-TM/DB2 workload, we used the following priority scheme:

1. VTAM, IRLM
2. IMS Control Regions, DB2 Systems Services, DB2 Database Services
3. IMS Dependant Regions

For the CICS/VSAM RLS workload, we used the following priority scheme:

1. SMSVSAM
2. VTAM, CPSM

3. CICS

If your system is running in Goal Mode, do NOT classify system address spaces which normally run with high dispatching priorities. Let them default to SYSTEM/SYSSTC.

For more information on performance of MVS WLM in GOAL mode, see *Workload Manager Performance Studies*, SG24-4352.

2.1.3 GRS Resource Name Lists

We made the following updates to the GRS resource name lists (RNLs):

- SYSTEMS Exclusion RNL

  We wanted to avoid GRS serialization of resources that were already serialized by some other means and resources that did not require serialization. This was done by adding the resource names to the SYSTEMS Exclusion RNL.

  - All IMS data set names

    IMS DBRC and the IRLM already provide serialization for the IMS data sets, so GRS serialization is redundant. Depending on the naming conventions at your installation, you may be able to include all the IMS data set names with one generic entry.

  - CICS data sets

    All the CICS run time system data sets, with the exception of the CSD, were unshared and therefore excluded from GRS management.

- RESERVE Conversion RNL

  For resources that are not included in the RESERVE Conversion RNL, reserves issued by IMS will result in physical reserves to the DASD. If there is no contention for other data sets on the device, the hardware reserves are faster.

  We recommend that you do not include the DBRC RECON or the OLDS and WADS names in the RESERVE conversion RNL.

More detailed information can be found in *IMS V5 System Administration Guide* and *MVS/ESA SP V5 Planning: Global Resource Serialization*, GC28-1450 and *OS/390 MVS Planning: Global Resource Serialization*, GC28-1759.

For the CICS/VSAM RLS workload, we added IGDCDSXS to the RESERVE conversion RNL as a generic resource. This resource is used to serialize the SMS control data sets. Converting this reserve to a global ENQ minimizes delays due to contention for resources and prevents deadlocks associated with the VARY SMS command.

2.1.4 XCF Tuning

The key to ensuring good performance for the XCF signalling service is to provide sufficient signalling resources, namely, message buffers, signalling paths and the message buffer space, and to control access to those resources by defining transport classes. *Setting up a Sysplex* gives a detailed description of all the factors involved in tuning XCF. This information is also discussed in the redbook *MVS/ESA Version 5 Sysplex Migration Guide*.

After experimenting with several different transport class definitions, we assigned all the groups to two transport classes:

- One class with the default length handled most of the traffic. We assigned one structure and four CTCs to this class.
- A second class was defined to handle the larger messages. Since it had much less traffic, we assigned one structure to this class.

There are three RMF XCF reports that provide data on XCF signalling resource utilization. Examples of these reports are in Figure 9, Figure 10 on page 26, and Figure 11 on page 27.

2.1.4.1 RMF XCF Usage by System Report

```
                                                       X C F   A C T I V I T Y
                                                                                                                      PAGE    1
            OS/390                   SYSTEM ID JCO              DATE 10/14/1996            INTERVAL 30.00.000
            REL. 01.02.00            RPT VERSION 1.2.0          TIME 20.00.00              CYCLE 1.000 SECONDS

                                                        XCF USAGE BY SYSTEM
----------------------------------------------------------------------------------------------------------------------------------
                                            REMOTE SYSTEMS                                                          LOCAL
---------------------------------------------------------------------------------------------------------     -----------------
                          OUTBOUND FROM JCO                                       INBOUND TO JCO                   JCO
---------------------------------------------------------------------------   ---------------------------     -----------------
                                        ----- BUFFER -----      ALL
TO        TRANSPORT  BUFFER        REQ   %    %    %    %     PATHS     REQ   FROM            REQ     REQ     TRANSPORT     REQ
SYSTEM    CLASS      LENGTH        OUT SML  FIT  BIG  OVR   UNAVAIL  REJECT   SYSTEM           IN  REJECT     CLASS      REJECT
JAO       DEFAULT    16,316      1,355 100    0    0    0         0       0   JAO          74,705       0     DEFAULT         0
          DEFSMALL      956     58,177   0  100    0    0         0       0                                   DEFSMALL        0
JBO       DEFAULT    16,316      1,963 100    0    0    0         0       0   JBO          40,860       0
          DEFSMALL      956     38,836   0  100    0    0         0       0
   :
                            ----------                                                 ----------
TOTAL                          318,657                                        TOTAL       334,512
```

Figure 9. Example of RMF XCF Usage by System Report

This report includes XCF statistics for each system in the sysplex. The key indicators on this report are:

- Buffer length defines the smallest buffer size that will be used for a particular transport class.
  - %BIG indicates that the buffer size is too small. XCF will dynamically expand the buffer size, but this requires additional resources, so the %BIG should be small or 0.
  - %SML indicates that the buffer size is too big. If the buffer size is much larger than needed, storage is being wasted.

    If there are large numbers in the SML or BIG columns (or both), you can segregate message traffic by size using transport classes. This requires defining unique PATHINs and PATHOUTs. You can determine the effect of changing class lengths by defining temporary transport classes via the SETXCF command, then using RMF to investigate the appropriateness of that class length.
- The message buffer space is fixed real storage used for XCF signalling. MAXMSG places an upper limit of amount of storage that can be used for this purpose. The length of the message buffer space can be specified in the COUPLExx member of parmlib by using MAXMSG(nnnnnn), where nnnnnn is the number of 1 KB blocks of storage. It specifies a value that XCF uses to

determine the allotment of message buffers when the MAXMSG parameter is not specified on any one of the following:

- The CLASSDEF statement
- The PATHIN statement
- The SETXCF START,CLASSDEF command
- The SETXCF START,PATHIN command

The default value for MAXMSG is 500 for OS/390 R1 and prior releases of MVS, and 750 for OS/390 R2. If this value is too small you will see nonzero values in the REQ REJECT fields. In this case, you may wish to increase the size of the message buffer.

- In this report, ALL PATHS UNAVAIL should be low or 0.

  Counts in this field are outbound message requests that were migrated to a signalling path in another transport class because there was no operational signalling path connected to the intended remote system and assigned to the selected transport class. This is usually caused by an error in the path definition.

- Total outbound messages should be about equal to total inbound messages. If more messages are sent than received, consider providing more outbound paths than inbound paths.

2.1.4.2 RMF XCF Path Statistics

```
                                                      X C F  A C T I V II T Y
                                                                                                                        PAGE   15
           OS/390                   SYSTEM ID JC0              DATE 10/14/1996            INTERVAL 30.00.000
           REL. 01.02.00            RPT VERSION 1.2.0          TIME 20.00.00              CYCLE 1.000 SECONDS

OTAL SAMPLES = 1,791                                   XCF PATH STATISTICS
---------------------------------------------------------------------------------------------------------------------------------
                            OUTBOUND FROM JC0                                                      INBOUND TO JC0
--------------------------------------------------------------------------------       ------------------------------------------
        T FROM/TO                                                                               T FROM/TO
TO      Y DEVICE, OR      TRANSPORT        REQ     AVG Q                               FROM     Y DEVICE, OR         REQ  BUFFERS
SYSTEM  P STRUCTURE       CLASS            OUT     LNGTH    AVAIL   BUSY   RETRY       SYSTEM   P STRUCTURE           IN  UNAVAIL
JA0     S IXCPLEX_PATH1   DEFAULT        1,355      0.00   1,355       0       0       JA0      S IXCPLEX_PATH1    2,240        0
        S IXCPLEX_PATH2   DEFSMALL       5,580      0.01   5,580       0       0                S IXCPLEX_PATH2    8,915        0
        C C600 TO C624    DEFSMALL      17,571      0.02  17,571       0       0                C C620 TO C604    20,456        0
        C C601 TO C625    DEFSMALL      19,509      0.02  19,508       1       0                C C621 TO C605    23,433        0
        C C602 TO C626    DEFSMALL      16,979      0.02  16,979       0       0                C C622 TO C606    23,780        0
JB0     S IXCPLEX_PATH1   DEFAULT        1,963      0.00   1,963       0       0       JB0      S IXCPLEX_PATH1    2,945        0
        S IXCPLEX_PATH2   DEFSMALL      11,070      0.01  11,070       0       0                S IXCPLEX_PATH2    1,980        0
        C C610 TO C624    DEFSMALL       8,386      0.02   8,386       0       0                C C620 TO C614    14,065        0
        C C611 TO C625    DEFSMALL       8,358      0.02   8,357       1       0                C C621 TO C615    13,738        0
        C C612 TO C626    DEFSMALL      13,930      0.02  13,930       0       0                C C622 TO C616    13,818        0
   ⋮
                                    ----------                                                                 ----------
TOTAL                                  344,571                                       TOTAL                        386.943
```

Figure 10. Example of RMF XCF Path Statistics Report

The TYP field indicates that the connection is a CF structure (S) or CTC link (C).

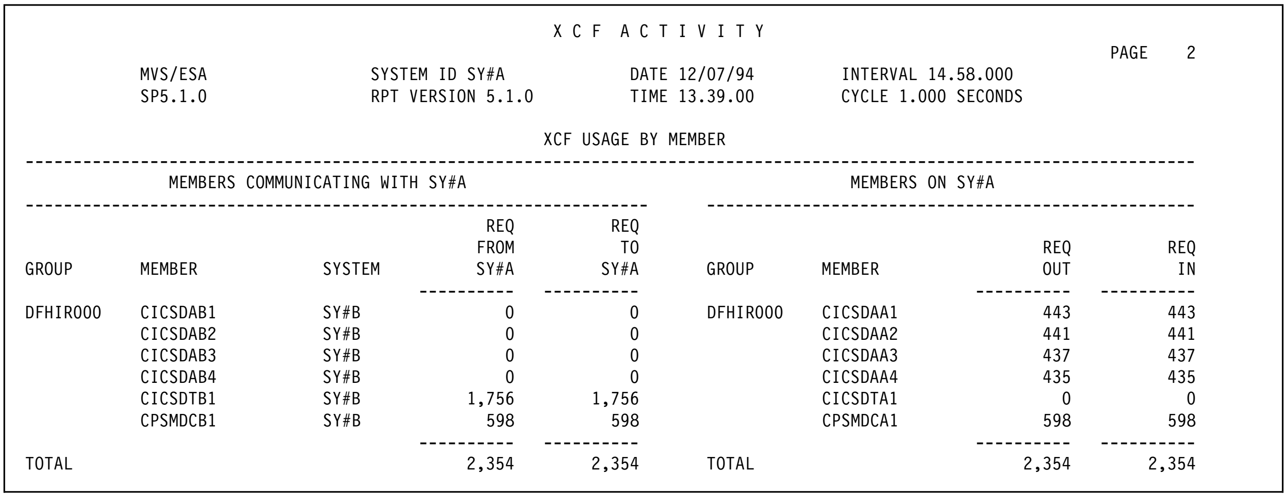
The key indicators on this reports are:

- AVG Q LNGTH show the number of requests queued for transfer over each outbound path. If this number is greater than one, more paths may be needed for this transport class.

- RETRY counts should be low relative to REQ OUT for a class. A nonzero count indicates that a signal has failed and was resent. This is usually indicative of a hardware problem.
- AVAIL counts should be high relative to BUSY counts.
- If BUSY count of an outbound path is high, check to see if the inbound signalling path on the remote system has a high BUFFER UNAVAIL value. If so, consider increasing the amount of message buffer space for the inbound path.

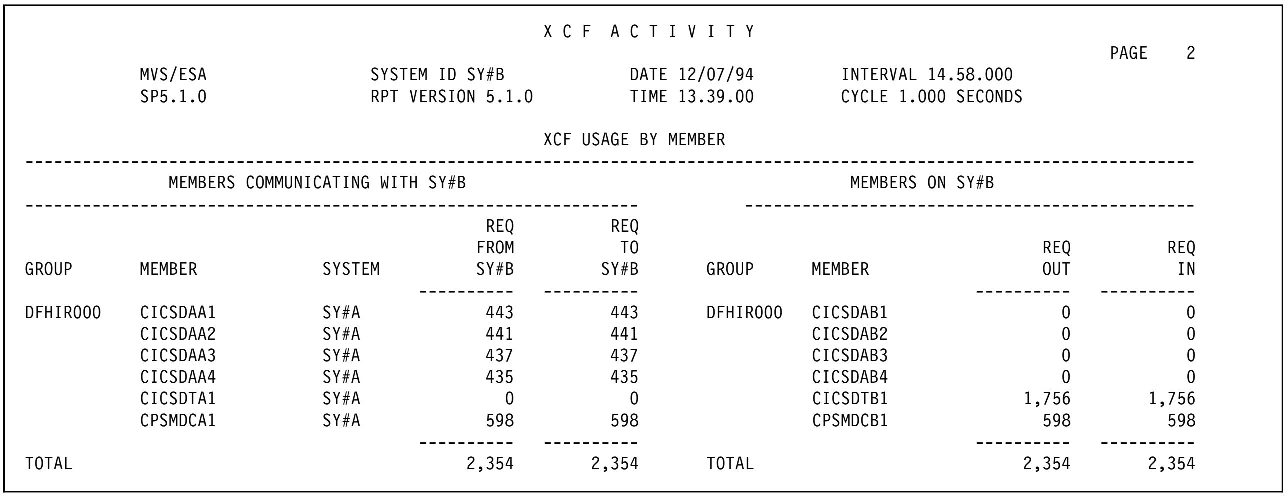
2.1.4.3 RMF XCF Usage by Member

GRS is an example of this type of group. If you know the routing characteristics of an application, you can use this report to verify it is behaving as expected.



```
                                                     X C F   A C T I V I T Y
                                                                                                                   PAGE    2
            MVS/ESA                 SYSTEM ID SY#A            DATE 12/07/94         INTERVAL 14.58.000
            SP5.1.0                 RPT VERSION 5.1.0         TIME 13.39.00         CYCLE 1.000 SECONDS

                                                      XCF USAGE BY MEMBER
-------------------------------------------------------------------------------------------------------------------------
               MEMBERS COMMUNICATING WITH SY#A                                          MEMBERS ON SY#A
-----------------------------------------------------------------      --------------------------------------------------
                                                     REQ          REQ
                                                    FROM           TO                                          REQ          REQ
GROUP       MEMBER             SYSTEM               SY#A         SY#A      GROUP       MEMBER                  OUT           IN
                                              ----------   ----------                                   ----------   ----------
DFHIR000    CICSDAB1           SY#B                    0            0      DFHIR000    CICSDAA1                443          443
            CICSDAB2           SY#B                    0            0                  CICSDAA2                441          441
            CICSDAB3           SY#B                    0            0                  CICSDAA3                437          437
            CICSDAB4           SY#B                    0            0                  CICSDAA4                435          435
            CICSDTB1           SY#B                1,756        1,756                  CICSDTA1                  0            0
            CPSMDCB1           SY#B                  598          598                  CPSMDCA1                598          598
                                              ----------   ----------                                   ----------   ----------
TOTAL                                              2,354        2,354      TOTAL                             2,354        2,354
```

Figure 11. Example of RMF XCF Usage by Member Report for SY#A



```
                                                    X C F   A C T I V I T Y
                                                                                                                   PAGE    2
            MVS/ESA                 SYSTEM ID SY#B            DATE 12/07/94         INTERVAL 14.58.000
            SP5.1.0                 RPT VERSION 5.1.0         TIME 13.39.00         CYCLE 1.000 SECONDS

                                                      XCF USAGE BY MEMBER
-------------------------------------------------------------------------------------------------------------------------
               MEMBERS COMMUNICATING WITH SY#B                                          MEMBERS ON SY#B
----------------------------------------------------------------          -----------------------------------------------
                                                     REQ          REQ
                                                    FROM           TO                                          REQ          REQ
GROUP       MEMBER             SYSTEM               SY#B         SY#B      GROUP      MEMBER                   OUT           IN
                                              ----------   ----------                                   ----------   ----------
DFHIR000    CICSDAA1           SY#A                  443          443      DFHIR000   CICSDAB1                   0            0
            CICSDAA2           SY#A                  441          441                 CICSDAB2                   0            0
            CICSDAA3           SY#A                  437          437                 CICSDAB3                   0            0
            CICSDAA4           SY#A                  435          435                 CICSDAB4                   0            0
            CICSDTA1           SY#A                    0            0                 CICSDTB1               1,756        1,756
            CPSMDCA1           SY#A                  598          598                 CPSMDCB1                 598          598
                                              ----------   ----------                                   ----------   ----------
TOTAL                                              2,354        2,354      TOTAL                             2,354        2,354
```

Figure 12. Example of RMF XCF Usage by Member Report for SY#B

If we want to look at the routing of CICS transactions, we find the CICS default XCF group, DFHIR000, on the RMF XCF report. In the sample RMF report shown above, CICSDAA1 to CICSDAA4 are the four CICS AORS defined on SY#A. CICSDTB1 is the TOR on SY#B.

In Figure 11, we can see that the TOR on SY#B is routing requests to the other system, SY#A. In Figure 12, we can see the AORs on SY#A are receiving signals from SY#B.

Other indicators of routing are shown in the CICSSTAT reports shown in Figure 18 on page 34.

2.2 IMS Tuning

This is not a comprehensive list of all the IMS tuning required to do multisystem IMS data sharing. It is a list of a few items that had to be changed as we increased the number of systems sharing data.

2.2.1 RECON Data Sets

The IMS RECON data set is the single recovery point that records status for all subsystems and databases in the data sharing group. Contention on the RECON data set grows with an increase in the number of systems and number of databases in the data sharing group. This contention can be very intense during system initialization, and OLDS switch time.

The following recommendations improve the I/O response to the RECON data set:

1. In the VSAM definitions for the RECON data set, do not specify WRITECHECK. WRITECHECK indicates that each time a record is written to a device it must be read back. The default for this value is NOWRITECHECK.
2. Increase the number of RECON LSR buffers. We used 12 and 48 index and data buffers, respectively. The defaults are 6 and 12. These values can be changed by using the DSPBUFFS CSECT described in the *IMS/ESA Customization Guide*.
3. Place the RECON data set on a volume behind a separate cached control unit with the DASD fast write feature.

   This may not be necessary for the smaller configurations, but it is recommended for larger configurations. The time to initialize the RECON data set decreased by a factor of two when we implemented this change for the 8x9672-R61 configuration.
4. Place RECONs on separate devices, controllers, and channels. Isolating the RECONs eliminates hardware reserve contention caused by access to other data.
5. Place each RECON in a separate user catalog with the BCS for the catalog on the same pack as the RECON. This is recommended to avoid shared DASD deadlocks.
6. Assuming that you have followed the recommendations for RECON placement given above, add DSPURI01, the resource name for the RECON data set, to the SYSTEMS Exclusion RNL. Do not include the DBRC RECON in the RESERVE conversion RNL. By not including them in this RNL, the reserves issued by IMS will cause physical reserves to DASD.
7. If /DBRs need to be issued against a group of full function databases, place as many database names as possible in each /DBR command. (One /DBR command with 10 databases will process faster within DBRC than 10 /DBR commands with one database.)

8. Increase the RECON logical record size as the number of systems in the sysplex grows.

2.2.2 DMB Pool Size

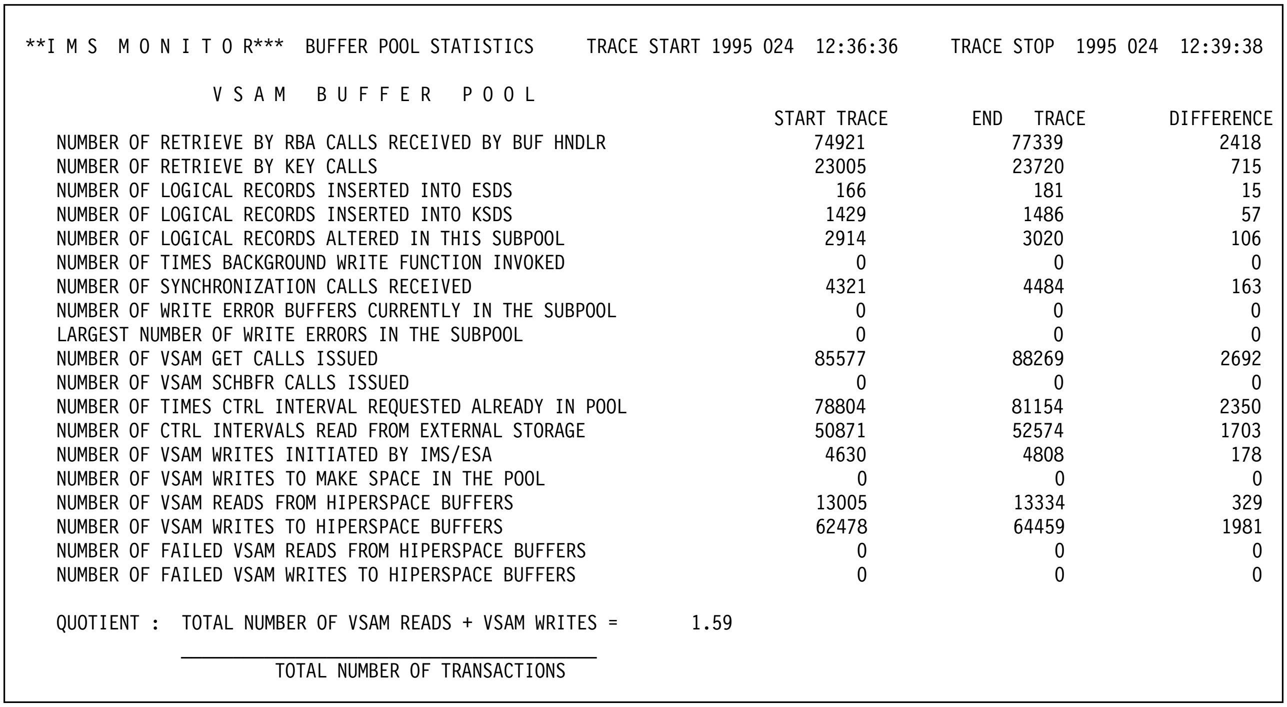
Running an IMS workload on a Parallel Sysplex or adding systems to a sysplex should not require a change to the DMB pool size. For the experiments in this document, as we increased the number of systems in the Parallel Sysplex, we also increased the number of shared databases. This necessitated a corresponding increase in the DMB Pool.

Since a growing sysplex may include new applications, this is just a reminder that you should specify a DMB pool size (DMB=) large enough to contain the working set of nonresident DMBs. Additional RECON I/O (and contention) can result from this pool being too small. Use /DIS POOL DMBP to check whether the pool size should be increased.

2.2.3 Key Information in the IMS Monitor Report

The IMSMON reports can provide additional information about IMS databases. Excerpts from some of the reports follow.

To validate the structure size of the VSAM and OSAM buffer cross invalidate structure, we can look at data from the VSAM and OSAM buffer pool reports.



```
**I M S  M O N I T O R***  BUFFER POOL STATISTICS     TRACE START 1995 024  12:36:36     TRACE STOP  1995 024  12:39:38

                  V S A M   B U F F E R   P O O L
                                                                    START TRACE        END   TRACE        DIFFERENCE
   NUMBER OF RETRIEVE BY RBA CALLS RECEIVED BY BUF HNDLR                74921              77339              2418
   NUMBER OF RETRIEVE BY KEY CALLS                                      23005              23720               715
   NUMBER OF LOGICAL RECORDS INSERTED INTO ESDS                           166                181                15
   NUMBER OF LOGICAL RECORDS INSERTED INTO KSDS                          1429               1486                57
   NUMBER OF LOGICAL RECORDS ALTERED IN THIS SUBPOOL                     2914               3020               106
   NUMBER OF TIMES BACKGROUND WRITE FUNCTION INVOKED                        0                  0                 0
   NUMBER OF SYNCHRONIZATION CALLS RECEIVED                              4321               4484               163
   NUMBER OF WRITE ERROR BUFFERS CURRENTLY IN THE SUBPOOL                   0                  0                 0
   LARGEST NUMBER OF WRITE ERRORS IN THE SUBPOOL                            0                  0                 0
   NUMBER OF VSAM GET CALLS ISSUED                                      85577              88269              2692
   NUMBER OF VSAM SCHBFR CALLS ISSUED                                       0                  0                 0
   NUMBER OF TIMES CTRL INTERVAL REQUESTED ALREADY IN POOL              78804              81154              2350
   NUMBER OF CTRL INTERVALS READ FROM EXTERNAL STORAGE                  50871              52574              1703
   NUMBER OF VSAM WRITES INITIATED BY IMS/ESA                            4630               4808               178
   NUMBER OF VSAM WRITES TO MAKE SPACE IN THE POOL                          0                  0                 0
   NUMBER OF VSAM READS FROM HIPERSPACE BUFFERS                         13005              13334               329
   NUMBER OF VSAM WRITES TO HIPERSPACE BUFFERS                          62478              64459              1981
   NUMBER OF FAILED VSAM READS FROM HIPERSPACE BUFFERS                      0                  0                 0
   NUMBER OF FAILED VSAM WRITES TO HIPERSPACE BUFFERS                       0                  0                 0

   QUOTIENT :  TOTAL NUMBER OF VSAM READS + VSAM WRITES =       1.59
               ________________________________________
                        TOTAL NUMBER OF TRANSACTIONS
```

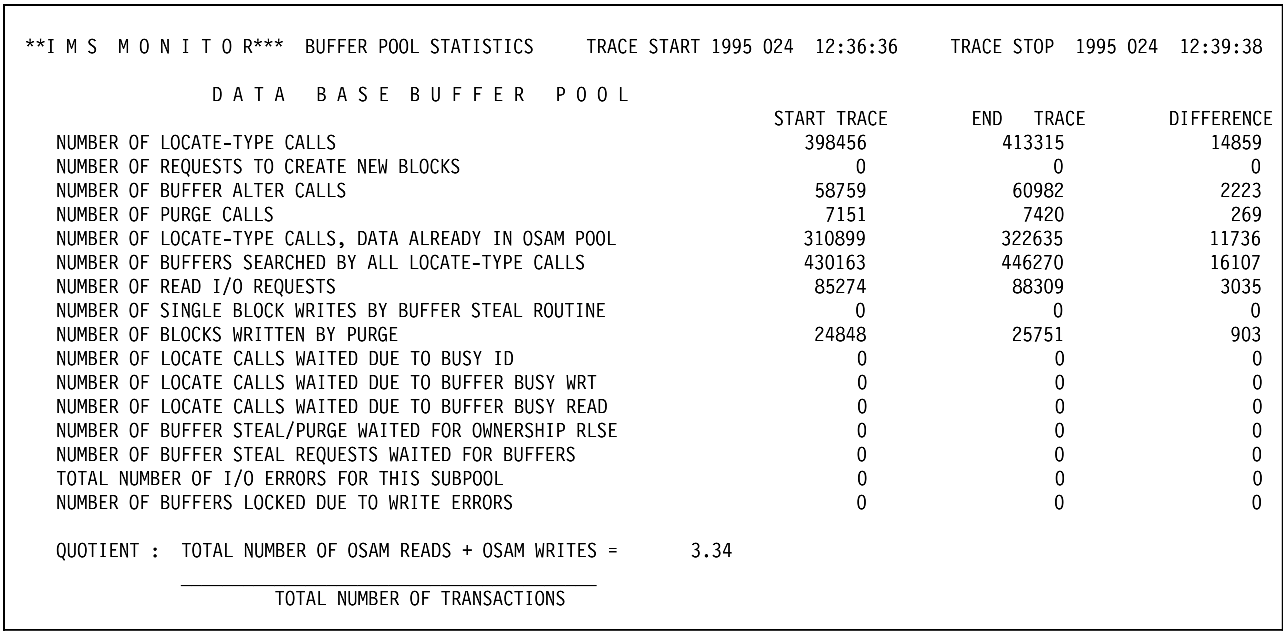
Figure 13. Example of VSAM Buffer Pool Report

- Database I/Os per transaction and/or per buffer locate call (hit ratio) can be determined from the database Buffer Pool reports.

  The hit ratio can be derived from the NUMBER OF TIMES CTRL INTERVAL REQUESTED ALREADY IN POOL and NUMBER OF CTRL INTERVAL READ FROM EXTERNAL STORAGE.

  For example, the cache hit ratio in the report example in Figure 13 is:

$$\textit{Hit ratio} = \frac{2350}{(2350 + 1703)} = 0.58$$



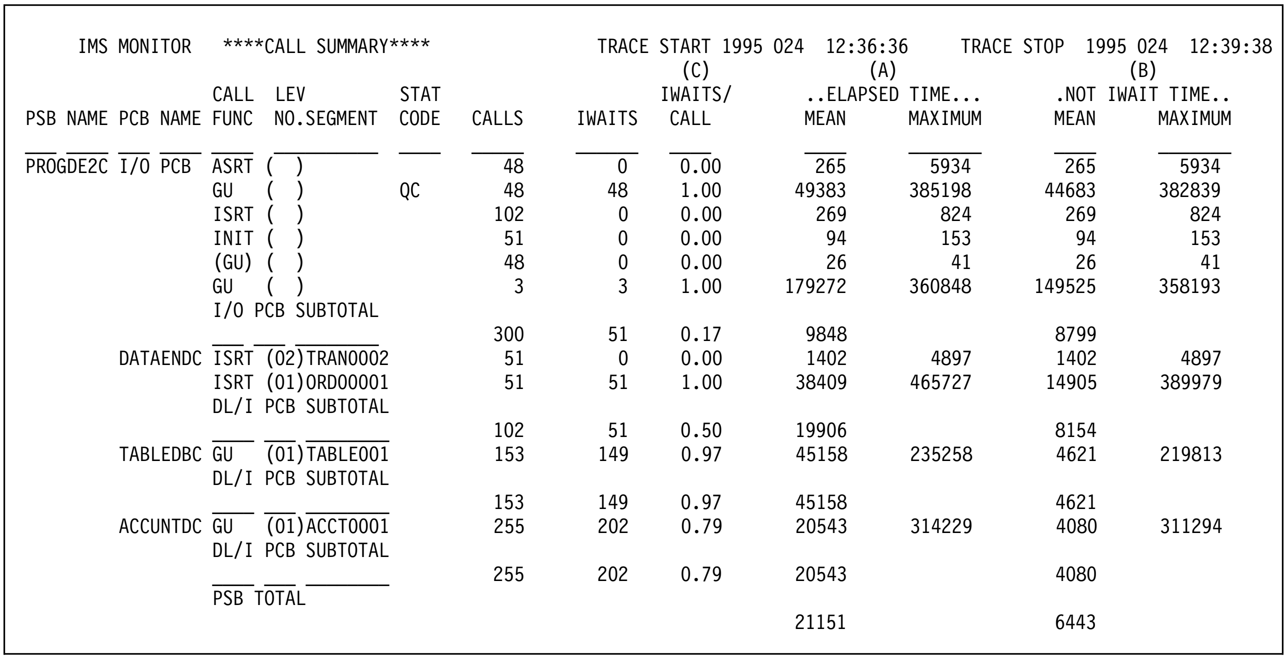
```
**I M S  M O N I T O R***  BUFFER POOL STATISTICS     TRACE START 1995 024  12:36:36     TRACE STOP  1995 024  12:39:38

                D A T A   B A S E  B U F F E R   P O O L
                                                                            START TRACE        END   TRACE         DIFFERENCE
   NUMBER OF LOCATE-TYPE CALLS                                                   398456             413315              14859
   NUMBER OF REQUESTS TO CREATE NEW BLOCKS                                            0                  0                  0
   NUMBER OF BUFFER ALTER CALLS                                                   58759              60982               2223
   NUMBER OF PURGE CALLS                                                           7151               7420                269
   NUMBER OF LOCATE-TYPE CALLS, DATA ALREADY IN OSAM POOL                        310899             322635              11736
   NUMBER OF BUFFERS SEARCHED BY ALL LOCATE-TYPE CALLS                           430163             446270              16107
   NUMBER OF READ I/O REQUESTS                                                    85274              88309               3035
   NUMBER OF SINGLE BLOCK WRITES BY BUFFER STEAL ROUTINE                              0                  0                  0
   NUMBER OF BLOCKS WRITTEN BY PURGE                                              24848              25751                903
   NUMBER OF LOCATE CALLS WAITED DUE TO BUSY ID                                       0                  0                  0
   NUMBER OF LOCATE CALLS WAITED DUE TO BUFFER BUSY WRT                               0                  0                  0
   NUMBER OF LOCATE CALLS WAITED DUE TO BUFFER BUSY READ                              0                  0                  0
   NUMBER OF BUFFER STEAL/PURGE WAITED FOR OWNERSHIP RLSE                             0                  0                  0
   NUMBER OF BUFFER STEAL REQUESTS WAITED FOR BUFFERS                                 0                  0                  0
   TOTAL NUMBER OF I/O ERRORS FOR THIS SUBPOOL                                        0                  0                  0
   NUMBER OF BUFFERS LOCKED DUE TO WRITE ERRORS                                       0                  0                  0

   QUOTIENT :  TOTAL NUMBER OF OSAM READS + OSAM WRITES =       3.34
               ________________________________________
                        TOTAL NUMBER OF TRANSACTIONS
```

Figure 14. Example of OSAM Buffer Pool Report

Similarly, on the OSAM Buffer Pool Report, the hit ratio can be derived from NUMBER OF LOCATE-TYPE CALLS, DATA ALREADY IN OSAM POOL and NUMBER OF LOCATE-TYPE CALLS.

For example, the cache hit ratio in the report example in Figure 14 is:

$$\textit{Hit ratio} = \frac{11736}{14859} = 0.79$$



```
     IMS MONITOR   ****CALL SUMMARY****             TRACE START 1995 024  12:36:36     TRACE STOP  1995 024  12:39:38
                                                                (C)            (A)                         (B)
                  CALL  LEV           STAT                    IWAITS/     ..ELAPSED TIME...        .NOT IWAIT TIME..
PSB NAME PCB NAME FUNC  NO.SEGMENT    CODE    CALLS     IWAITS  CALL          MEAN     MAXIMUM        MEAN     MAXIMUM
________ ________ ____  __________    ____    _____     ______  ____          ____     _______        ____     _______
PROGDE2C I/O PCB  ASRT (  )                      48          0  0.00           265        5934         265        5934
                  GU   (  )           QC         48         48  1.00         49383      385198       44683      382839
                  ISRT (  )                     102          0  0.00           269         824         269         824
                  INIT (  )                      51          0  0.00            94         153          94         153
                  (GU) (  )                      48          0  0.00            26          41          26          41
                  GU   (  )                       3          3  1.00        179272      360848      149525      358193
                  I/O PCB SUBTOTAL
                  ___  ___ ________             300         51  0.17          9848                    8799
         DATAENDC ISRT (02)TRAN0002              51          0  0.00          1402        4897        1402        4897
                  ISRT (01)ORD00001              51         51  1.00         38409      465727       14905      389979
                  DL/I PCB SUBTOTAL
                  ____ ___ ________             102         51  0.50         19906                    8154
         TABLEDBC GU   (01)TABLE001             153        149  0.97         45158      235258        4621      219813
                  DL/I PCB SUBTOTAL
                  ____ ___ ________             153        149  0.97         45158                    4621
         ACCUNTDC GU   (01)ACCT0001             255        202  0.79         20543      314229        4080      311294
                  DL/I PCB SUBTOTAL
                  ____ ___ ________             255        202  0.79         20543                    4080
                  PSB TOTAL
                                                                             21151                    6443
```

Figure 15. Example of IMSMON DL/I Call Summary

- Time per DL/I call per PCB per PSB can be determined from the DL/I call summary. Use elapsed time with NOT IWAIT TIME. NOT IWAIT TIME is the elapsed time of the call minus the time spent in doing I/O. Time spent waiting for locks is included in the NOT IWAIT TIME. One way to find calls waiting on lock requests is to find calls that have high NOT IWAIT TIME. Other wait time included in NOT IWAIT TIME is waiting for logging, DBRC RECON access, dispatching, and paging.

  You can use the RMF DASD report to check for any increase in I/O response time.

- Also look at the IWAITs per DL/I call. This does not measure the I/O durations, just average number of IWAITs. An increase in the number of IWAITS could be caused by I/O to refresh invalidated buffers.

- Look for NO POOL SPACE FAILURES; there should be none. Also look at the RMF paging and virtual storage reports for any storage problems. There can be unexpected constraints, for example, if the number of database buffers is substantially increased.

```
     IMS MONITOR   ****PROGRAM SUMMARY****              TRACE START 1995 024  12:36:36     TRACE STOP  1995 024  12:39:38
                                                               (A)........(B)........         (A)........(B).....
                                                        I/O   TRAN.       CPU       .        ELAPSED    SCHED.TO
          NO.   TRANS.              CALLS      I/O   IWAITS  DEQD.       TIME    DISTR.        TIME    1ST CALL     TIME
PSBNAME  SCHEDS.   DEQ.      CALLS  /TRAN    IWAITS  /CALL  /SCH.      /SCHED.     NO.       /SCHED.    /SCHED.   /TRANS.
_______  _______  ____      _____  _____    ______  _____  _____      _______    ___       _______    _______   _______

PROGIT8C      44    59       1091   18.4       595    0.5    1.3       104398  2073A,B      709343     676348    529001
PROGPS3D      46    56       2164   38.6       383    0.1    1.2       105960  2084A,B      370523     743402    304358
PROGOE4C      25    27        644   23.8       456    0.7    1.0       110148  2090A,B      724124     624625    670485
PROGOE5D      20    24        712   29.6       311    0.4    1.2       129630  2094A,B      900892     458428    750744
PROGOE2C      26    30        292    9.7        37    0.1    1.1        65454  2102A,B      278207     656925    241113
   ⋮
```

Figure 16. Example of IMSMON Program Summary

- The program summary from different monitor runs can be compared to study workload uniformity. A valid throughput comparison requires that the transaction mix be similar.

2.3 Avoiding IRLM Deadlocks

An IRLM deadlock occurs when one system owns lockA and requests lockB, and simultaneously, a second system owns lockB and requests lockA. While this situation may occasionally occur, these deadlocks should be kept to a minimum. The following parameters can be used to minimize deadlocks:

- IRLM DEADLOK parameter

  DEADLOK is a parameter that is specified on the IRLM procedure. It specifies the local deadlock-detection interval (in seconds), and number of local cycles that are to occur before a global detection is initiated.

  Frequent local deadlock detection, using a low value for the first of the DEADLOK values, increases IRLM processing devoted to local deadlock detection. Global deadlock detection also increases processing when it is performed and it requires IRLM-to-IRLM communication. However, frequent deadlock detection reduces the length of time that applications are left waiting in a deadlock. In our OLTP test environments we used a DEADLOK parameter of (1,1). If your environment includes a batch job that does a checkpoint call every two or three seconds, the IMS 2.1 default of (5,1) might be a good choice.

- Decrease IRLM MAXUSRS from 32 to the actual number of users (IRLM subsystems).

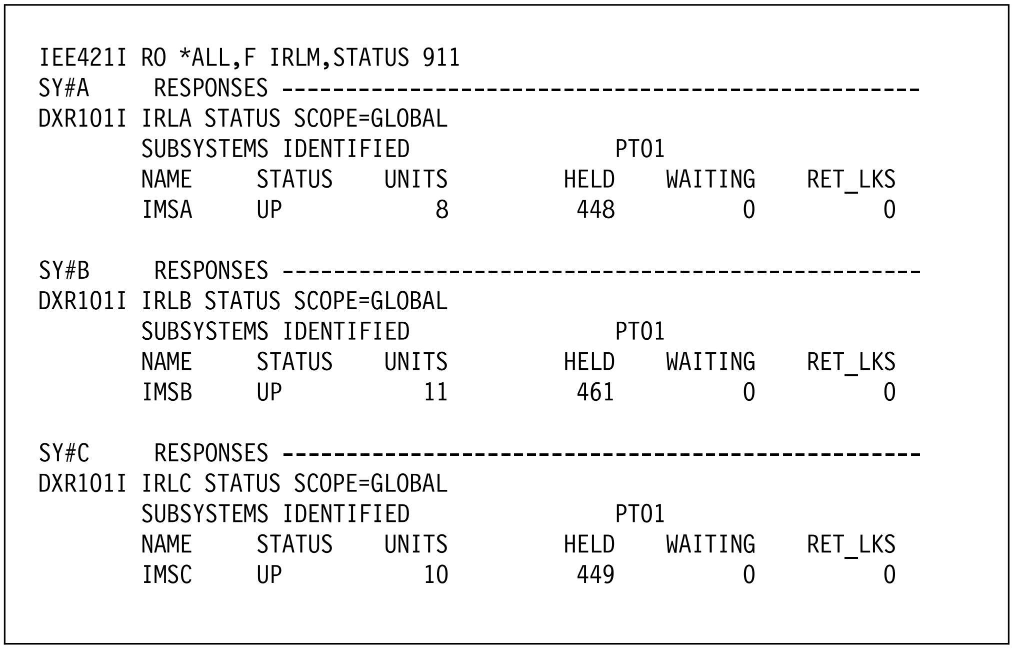
  MAXUSRS is a parameter on the IRLM procedure. It specifies the number of users of the global lock structure. The users are online systems and any additional DLI batch jobs that could connect to the global lock structure. By reducing MAXUSRS to the number of actual users, the size of each lock table entry may be reduced, allowing more locks in the table and hence less false lock contention. For seven or fewer MAXUSRS, the lock entries are two bytes. For 8 to 23 MAXUSRS, the lock entries are four bytes. For 24 to 32 MAXUSRS, the lock entries are eight bytes.

  If you intend to add additional systems during this instance of the sysplex, they should be included in the number of systems specified.

- Dispatching priority

  For the CICS/DBCTL workload, we gave IRLM a higher dispatching priority than DBRC, which had a higher priority than the IMS control regions.

You can use the F IRLM,STATUS command to display IRLM waiters. This is most interesting during peak periods.



```
IEE421I RO *ALL,F IRLM,STATUS 911
SY#A     RESPONSES ---------------------------------------------------
DXR101I IRLA STATUS SCOPE=GLOBAL
        SUBSYSTEMS IDENTIFIED              PT01
        NAME     STATUS    UNITS        HELD    WAITING    RET_LKS
        IMSA     UP            8         448          0          0

SY#B     RESPONSES ---------------------------------------------------
DXR101I IRLB STATUS SCOPE=GLOBAL
        SUBSYSTEMS IDENTIFIED              PT01
        NAME     STATUS    UNITS        HELD    WAITING    RET_LKS
        IMSB     UP           11         461          0          0

SY#C     RESPONSES ---------------------------------------------------
DXR101I IRLC STATUS SCOPE=GLOBAL
        SUBSYSTEMS IDENTIFIED              PT01
        NAME     STATUS    UNITS        HELD    WAITING    RET_LKS
        IMSC     UP           10         449          0          0
```

Figure 17. Example of IRLM STATUS Command

The UNITS are the current number of work units (threads or dependent regions) holding locks. The HELD column shows the number of IRLM locks currently held. WAITING shows the number of tasks that are waiting because they cannot obtain a lock. This is the number you want to minimize.

2.4 CICS and CICSPlex SM Tuning

Most of the CICS changes resulted from the introduction of CICSPlex System Manager (CP/SM) and the addition of systems to the sysplex.

2.4.1 MAXTASKS

The MAXTASKS parameter limits the number of concurrent tasks. Tuning this parameter is documented in *CICS/ESA V4R1 Performance Guide* for CICS/ESA Version 4 Release 1 and *CICS Performance Guide*, SC33-1699 for CICS Transaction Server. If the peak number of tasks in your system is close to MAXTASKS and you are installing CP/SM, you should increase MAXTASKS by 20 for each CICS region.

2.4.2 Shared CSD on Cached DASD

We placed the shared CSD on a volume behind a separate cached control unit with the DASD fast write feature to improve CICS initialization time. This became more important as we added additional systems into the sysplex.

2.4.3 Dispatching Priority

The CAS and CMAS should have the same priority, and that priority should be greater than the managed MASs. For more information, see *CICSPlex SM Setup and Administration* in the chapter titled, “Preparing to start a CMAS.”

2.4.4 CICS Routing

VTAM generic resources can be used to connect a terminal to the CICS Terminal Owning Region (TOR) on a certain system. Once the terminal is connected to a TOR, CP/SM will give recommendations for routing to the TOR. The TOR will route transactions to an Application Owning Region (AOR) on the local system or send them to a remote system.

Routing CICS transactions to other systems involves additional processing. Unnecessary routing can be minimized by:

1. Defining a TOR on every system

   In most cases, the transaction from a TOR will be run on the AORs on the same system. If there are systems with AORs, but no TORs, all transactions processed by these systems will have to come from another system, which means there is 100% routing on these systems.

2. Identifying transactions with affinity

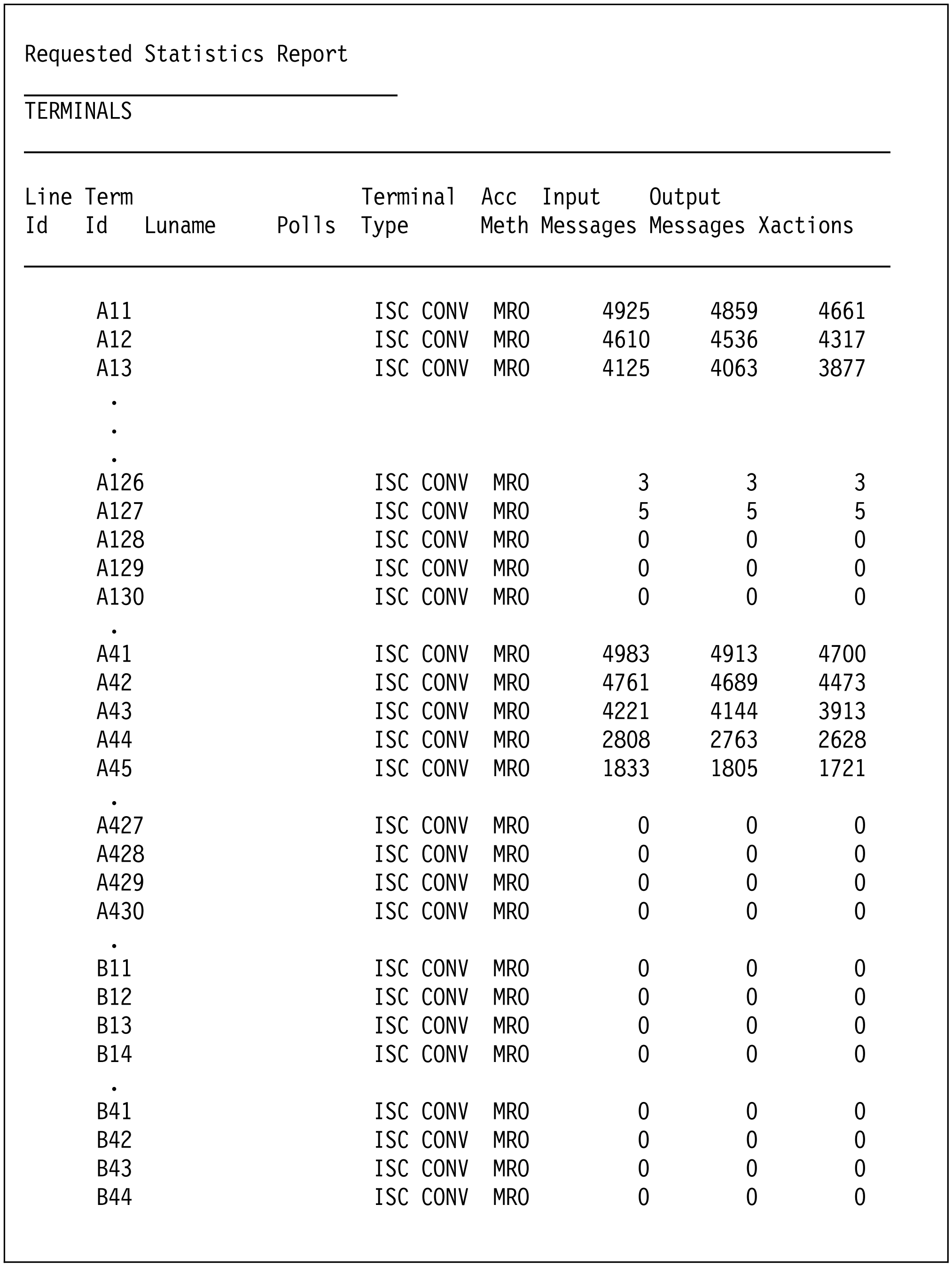
   Ensure that transactions with affinity are run on the same processor. An example could be a transaction that uses a CICS temporary storage object created by a previous transaction. CICSPlex SM will ensure that for transactions defined to have affinity, the subsequent transactions will be routed to the system where the first one was scheduled.

   There is a product, *IBM CICS Transaction Affinity Utility*, (5695-587), that can be used to identify the transactions with affinity and create an input list to CP/SM, which will force affinities to the same system. This utility is now part of the CICS Transaction Server package.

3. Choosing the best routing algorithm

> We used the “Queue” algorithm in CP/SM. The “Goal” algorithm is also available if you are running all AORs and TORs in WLM goal mode. You should select the appropriate algorithm for your installation.

In the CICS/DBCTL workload, we tuned the systems so that more than 90% of the transactions were processed on the local system.



```
Requested Statistics Report
_______________________________
TERMINALS
________________________________________________________________________

Line Term                    Terminal  Acc  Input    Output
Id   Id   Luname     Polls  Type      Meth Messages Messages Xactions
________________________________________________________________________

      A11                     ISC CONV  MRO      4925     4859     4661
      A12                     ISC CONV  MRO      4610     4536     4317
      A13                     ISC CONV  MRO      4125     4063     3877
       .
       .
       .
      A126                    ISC CONV  MRO         3        3        3
      A127                    ISC CONV  MRO         5        5        5
      A128                    ISC CONV  MRO         0        0        0
      A129                    ISC CONV  MRO         0        0        0
      A130                    ISC CONV  MRO         0        0        0
       .
      A41                     ISC CONV  MRO      4983     4913     4700
      A42                     ISC CONV  MRO      4761     4689     4473
      A43                     ISC CONV  MRO      4221     4144     3913
      A44                     ISC CONV  MRO      2808     2763     2628
      A45                     ISC CONV  MRO      1833     1805     1721
       .
      A427                    ISC CONV  MRO         0        0        0
      A428                    ISC CONV  MRO         0        0        0
      A429                    ISC CONV  MRO         0        0        0
      A430                    ISC CONV  MRO         0        0        0
       .
      B11                     ISC CONV  MRO         0        0        0
      B12                     ISC CONV  MRO         0        0        0
      B13                     ISC CONV  MRO         0        0        0
      B14                     ISC CONV  MRO         0        0        0
       .
      B41                     ISC CONV  MRO         0        0        0
      B42                     ISC CONV  MRO         0        0        0
      B43                     ISC CONV  MRO         0        0        0
      B44                     ISC CONV  MRO         0        0        0
```

Figure 18. Example of CICSSTAT Requested Statistics Report

The CICSSTAT Requested Statistics reports can be used to determine the extent of routing between AORS. The example below displays the viewpoint of the TOR on SYS#A. There are four AORs on SY#A; each has 30 sessions. The first AOR has sessions A11 through A130. Also shown are some AORs on another CEC, SY#B.

In studying this report, we want to verify that the routing to AORs on SY#B is less than 10%. In this case all the messages to the AORs on SY#B (B1n through B4n) are 0, so no routing is occurring.

We also want to verify that we have established enough sessions for each AOR. Sessions are reused from the lowest number to the highest number. In this example, we see the usage dropping as we go to higher session numbers, with A128 through A130 being zero. This is an indication that we have enough sessions.

Other indicators of an adequate number of sessions can be found in the System and Mode Entries report.

```
ISC/IRC SYSTEM AND MODE ENTRIES
_______________________________
  System Entry
  ____________
  Connection name                               :     DAA1
  Aids in chain                                 :        0
  Generic aids in chain                         :        0
  ATIs satisfied by contention losers           :        0
  ATIs satisfied by contention winners          :        0
  Peak contention losers                        :        0
  Peak contention winners                       :       21
  Peak outstanding allocates                    :        0
  Total number of allocates                     :    19360
  Queued allocates                              :        0
  Failed link allocates                         :        0
  Failed allocates due to sessions in use       :        0
  Maximum queue time (seconds)                  :        0
  Allocate queue limit                          :        0
```

Figure 19. Example of CICSSTAT System and Mode Entries Report

In this example, the field “Failed allocates due to sessions in use” is 0. The sum of the fields “Peak contention winners” (21) and “Peak contention losers” (0) for LU6.1 should equal the number of send and receive sessions used. This is less than the session count of 30.

2.5 CICS/VSAM RLS Tuning

Two functions of CICS/VSAM RLS that impact performance are the lock management and buffer management.

2.5.1 Locking

VSAM RLS provides three read integrity options. They are:

- NRI - no read integrity
- CR - consistent read
- CRE - consistent read explicit (or RR - repeatable read)

For NRI, VSAM does not obtain a lock on the record. However, VSAM does test validity of the buffer containing the record. If the buffer is invalid, VSAM discards the record and accesses the data set to obtain a new copy of the control interval

(CI). This insures that a sequential reader does not miss records that are moved to new CIs by CI split and CA split.

For CR, VSAM obtains a lock on the record when processing the GET request. After moving a copy of the record to the caller′s area, the lock is released.

For CRE, VSAM obtains a lock on the record when processing the GET request. The lock remains held until end of transaction when CICS makes an explicit "release locks" request to VSAM. The CRE option is only available to CICS applications. At the CICS API, this option is called repeatable read (RR).

The general recommendation is that applications use the NRI option. This avoids locking overhead while returning a current copy of the record. It is true that for a recoverable file, the returned record could be the new version of an updated record and the update could subsequently be backed out. While the CR option would not return the uncommitted version of a record, the fact that the record lock is released before the application processes the returned record means the record may be changed or deleted before it is processed by the application. The CRE option insures consistency of the record across the life of the transaction that issued the GET CRE (CICS repeatable read) request.

The decision to use NRI, CR, or CRE (RR) is an application design consideration. When considering NRI or CR, look closely at the differences between the two. In cases where NRI is sufficient, it offers a performance advantage.

Resolution of deadlocks is controlled by the DEADLOCK_DETECTION parameter in the IGDSMSxx parmlib member. As the level of read increases, this parameter becomes more important. The default is (15,4), but, based on our IMS IRLM experience, we ran with (1,1) to minimize the time required to detect a deadlock.

The size of VSAM RLS lock structure, IGWLOCK00, can be determined using the information in *DFSMS/MVS V1R3: DFSMSdfp Storage Administration Reference*, SC26-4929. For large sysplexes, you may get a more accurate size estimate by using RLSLKSZ on MKTTOOLS. To use this tool, you will need data from the CICSSTAT report (before RLS implementation) or the SMF 42.15 record (after RLS implementation). Like IRLM, CICS/VSAM RLS bases the size of the lock table entry on the number of systems in the sysplex. It uses the MAXSYSTEM value specified on the couple dataset. Overspecification of this variable can result in wider lock entries and therefore fewer lock entries for a given structure size.

D SMS,CFLS displays lock information. An example of this command is shown in Figure 20 on page 37. Locks in this display are only the OBTAIN locks. In the RMF CF Structure Activty Report, both the OBTAINS and RELEASEs are reported.

2.5.2 Caching

The performance of VSAM RLS is dependant on its buffer management. The following considerations apply:

- Ideally, the sum of all the RLS CF cache structure sizes should equal the sum of the local VSAM shared resources (LSR). At a minimum, they should be large enough to hold all the directory entries. The ratio of directory entries to data elements was changed by APAR OW23008 which dynamically adjusts the ratio.

```
D SMS,CFLS
IGW320I 23:46:08 Display SMS,CFLS 862
STRUCTURE NAME:IGWLOCK00 VERSION:ADC1A9748CF89A02 SIZE:35072K
RECORD TABLE ENTRIES:129892 USED:159
System   Interval   LockRate   ContRate  FContRate   WaitQLen
W1       1 Minute      188.0      2.249      2.059      54.38
W1        1 Hour       155.3      0.601      0.482      87.71
W1        8 Hour       200.4      0.451      0.326      20.91
W1         1 Day  ---------- ---------- ---------- ----------
  (06)   1 Minute      134.6      1.349      1.266      47.39
  (06)    1 Hour       120.7      0.298      0.305      61.07
  (06)    8 Hour       152.6      0.280      0.205      15.12
  (06)     1 Day  ---------- ---------- ---------- ----------
***************** LEGEND ******************
 LockRate = number of lock requests per second
 CONTRATE = % of lock requests globally managed
 FCONTRATE = % of lock requests falsely globally managed
 WaitQLen = Average number of requests waiting for locks
```

Figure 20. Sample Output from D SMS,CFLS Command

- A parameter, RLS_MAX_POOL_SIZE has been added to the IGDSMSxx parmlib member to allow the installation to limit the size of the local buffer pool. For the workloads described in Chapter 6, “VSAM RLS Data Sharing Scalability Study” on page 81 and Chapter 9, “VSAM RLS Asymmetric Configuration Performance Study” on page 109, this parameter was set to twice the sum of the local buffers on a single system. Setting the RLS_MAX_POOL_SIZE too low might result in unnecessarily degrading the local hit rate. You can use the data in the SMF42.19 record to evaluate the local buffer hit rates.

2.6 DB2 Tuning

To achieve the best DB2 performance in the data sharing environment, it is essential to control DB2 locking rates. We recommend the following to minimize lock requests:

- Plan for frequent commits in long running applications to ensure a minimum of global lock conflicts, to improve the effectiveness of DB2′s global lock avoidance algorithms, and to release resources in IRLM, in XES, and in the coupling facility.

- Consider the use of isolation level uncommitted read (UR) for query applications that can tolerate reading data that is potentially uncommitted. This will reduce the number of global locks.

- Use an isolation level of cursor stability (CS) and select CURRENTDATA(NO) on package and plan bind operations. This will maximize the effectiveness of DB2′s global lock avoidance.

- Bind application plans and packages with RELEASE(DEALLOCATE). Where you have thread reuse, and there is a low level of concurrent access against a tablespace, this will reduce the number of global locks sent to the coupling facility and reduce global contention.

- Use Type 2 indexes with page level locking (LOCKSIZE PAGE) to significantly reduce the number of global locks sent to the coupling facility.

- Avoid all use of the isolation level repeatable read (RR) with Type 2 indexes to avoid next-key locking on SQL insert processing.

Some other general suggestions are:

- Row level locking may provide additional concurrency for some applications. However, it may increase the number of locks that must be obtained, which in turn increases the CPU cost of locking in data sharing environments. We recommend judicious use of this function.
- Partitioned tables are usually beneficial in a DB2 data sharing environment. Their use allows DB2 to consider the use of CPU and I/O parallelism to improve query elapsed time. Their use also allows you to run utility processes such as tablespace reorganizations in parallel.
- Long running or batch units of work with high CPU usage and time-critical requirements should be analyzed with regard to their behavior on processors with small engines.
- A single group buffer pool is sufficient in test environments or where the total number of DB2 buffers is less than 10,000. Otherwise, consideration should be given to using multiple group buffer pools to provide better information for making tuning decisions.

Various performance improvements were made to DB2 performance through maintenance to MVS 5.2, DB2, and IRLM. To obtain these improvements, the coupling facility must be CFLEVEL=2.

Figure 21 on page 39 is an example of the DB2 display command that can be used to obtain real-time information about DB2 GBP activity. The data reported is a delta from the last display command issued.

See *DB2 for MVS/ESA Version 4 Data Sharing Performance Topics* for a detailed explanation of all the fields in this and other displays, as well as examples of data from DB2 PM. There are a few fields which are related to the definition of the DB2 cache structure:

- An increase in RECLAIMS FOR DIRECTORY ENTRIES could indicate a structure that is too small or a problem with the directory-to-data ratio.
- CURRENT DIRECTORY TO DATA RATIO should indicate that there are more directory than data entries. If this is not the case, directory entries for unchanged data will be reclaimed, causing unnecessary I/O to DASD, and CPU consumption caused by notification of other systems.

Some of this data is now reported by RMF (APAR OW13536). See Figure 23 on page 46 for an example. This APAR also stores additional cache related data in the SMF 74, subtype 4 record.

2.7 Tuning for GRSSTAR Mode

GRS in STAR mode uses a lock structure to serialize system resources. Each instance of contention on this lock structure represents contention to a resource. Typical contention on this structure is 1%. If you see higher contention, you may want to check the ENQUEUE activity. In our case, we had a lot of contention on IGDCDSXS, the resource used to serialize the SMS COMMDS dataset. By increasing the INTERVAL in the IGDSMSxx parmlib member, we were able to reduce contention to less than 1%.

```
@DBB1 DIS GBPOOL(GBP0) GDETAIL(INTERVAL)
DSNB750I @DBB1 DISPLAY FOR GROUP BUFFER POOL GBP0 FOLLOWS
DSNB755I @DBB1 DB2 GROUP BUFFER POOL STATUS
             CONNECTED                                 = YES
             CURRENT DIRECTORY TO DATA RATIO           = 5
             PENDING DIRECTORY TO DATA RATIO           = 5
DSNB756I @DBB1   CLASS CASTOUT THRESHOLD                   = 10%
             GROUP BUFFER POOL CASTOUT THRESHOLD       = 50%
             GROUP BUFFER POOL CHECKPOINT INTERVAL     = 8 MINUTES
             RECOVERY STATUS                           = NORMAL
DSNB757I @DBB1 MVS CFRM POLICY STATUS FOR DSNDB1G_GBP0     = NORMAL
             MAX SIZE INDICATED IN POLICY              = 21600 KB
             ALLOCATED                                 = YES
DSNB758I @DBB1     ALLOCATED SIZE                          = 21760 KB
               VOLATILITY STATUS                       = VOLATILE
DSNB759I @DBB1     NUMBER OF DIRECTORY ENTRIES             = 21606
               NUMBER OF DATA PAGES                    = 4321
               NUMBER OF CONNECTIONS                   = 9
DSNB782I @DBB1 INCREMENTAL GROUP DETAIL STATISTICS SINCE 11:30:55
                                                     NOV 15, 1995
DSNB784I @DBB1 GROUP DETAIL STATISTICS
             READS
               DATA RETURNED                           = 7
DSNB785I @DBB1     DATA NOT RETURNED
                 DIRECTORY ENTRY EXISTED               = 6
                 DIRECTORY ENTRY CREATED               = 2
                 DIRECTORY ENTRY NOT CREATED           = 3, 0
DSNB786I @DBB1   WRITES
               CHANGED PAGES                           = 118
               CLEAN PAGES                             = 0
               FAILED DUE TO LACK OF STORAGE           = 0
             CHANGED PAGES SNAPSHOT VALUE              = 3
DSNB787I @DBB1   RECLAIMS
               FOR DIRECTORY ENTRIES                   = 0
               FOR DATA ENTRIES                        = 0
             CASTOUTS                                  = 70
DSNB788I @DBB1   CROSS INVALIDATIONS
               DUE TO DIRECTORY RECLAIMS               = 0
               DUE TO WRITES                           = 4
```

Figure 21. Example of DB2 Display Group Buffer Pool Command

2.8 Planning for Coupling Facilities

In prior systems, capacity planning generally consisted of sizing three resources: processor size, processor storage, and I/O. The Parallel Sysplex introduces another element to be sized, namely, the coupling facility.

Sizing a coupling facility involves estimating the number of processors and the amount of storage needed on the coupling facility. It also includes estimating the number CF links that will be needed.

The tools mentioned earlier in Appendix D, “Capacity Planning Tools” on page 189 can be used to size the coupling facilities for your installation.

The following sections describe the initial sizing done for the workloads described in this document. After the Parallel Sysplex was implemented, measurement data about the coupling facility was used to refine this sizing.

2.8.1 Coupling Facility Size

The number and size of the coupling facilities needed for any configuration is highly dependent on the selection of functions that use the coupling technology. For availability, we recommend that you install more than one coupling facility. For best CF performance, you will have to determine the number of engines (or CPs) on each coupling facility and the amount of storage needed.

2.8.1.1 CF Storage

The storage required is the sum of the storage for the individual structures residing on the CF, an allocation for CF dump space, and some unallocated storage for structures that may be rebuilt on this CF in a recovery situation. This sum should be rounded up to the nearest power of 2 (for example, 256 MB, 512 MB).

Structure Size: In the Parallel Sysplex environment, there are several functions that can use the coupling facility: XCF signalling, JES2 checkpoint, VTAM generic resource management, IRLM locking, IMS OSAM and VSAM buffer cross invalidation, DB2 buffer caching, system logger, automated tape switching, RACF sysplex data sharing, GRS Star, CICS logging, and VSAM RLS. To use these functions, you must define the associated structures in the coupling policy.

Calculating the correct size for each structure is the first step in determining the amount of storage and the number of coupling facilities that will be needed. The size of each structure is dependent on various factors, which vary widely from installation to installation. There is a document (INFOSYS Q662530) that has lookup tables for each structure. Using these tables and the factors appropriate for your installation, you can determine the size of the structures you intend to use.

If the structures defined are too small, various problems can occur:

- If the IRLM lock table is too small, false lock contention (the mapping of two or more different lock requests to the same lock table entry) can occur.
- If the XCF structure is too small, XCF will not initialize. If it is a little larger (but not large enough), XCF may initialize, but it will not allow additional systems to join the sysplex.
- If the cache structures are too small, directory entries may be purged. This will cause invalidation of local buffers that may not have been changed and, ultimately, more I/Os and CPU overhead caused by notification of other systems. An increase in the number of reclaims (as reported by RMF with OW13536 installed) is indication that this is occurring.
- If the DB2 cache structures are too small, the castout threshold for changed pages will be reached more frequently. Castout processing can become a continuous background process, leading to increased CPU and I/O consumption.
- If the LOGR structure is too small, the system logger will not be able to offload log data in the structure before the structure is filled. Once the structure is full, system logger will not accept write messages until the offload processing is completed.

Structure Placement: Once you have determined the size of the structures you intend to use, you must decide where they should reside. We will assume that you have two coupling facilities. In some of the smaller configurations, one coupling facility will have sufficient capacity, but two coupling facilities are recommended for availability.

INFOSYS Q662530 contains some recommendations for structure placement. For the experiments in this document, we used the RMF CF Summary Report to observe structure sizes and request rates. Then we distributed the structures to balance the sizes and request rates.

2.8.1.2 CF Processors

The number of CPs required in the CF depends on the speed of and number of the sending processors, the extent to which the coupling technology is used by each sending processor and, in the PR/SM environment, the definition of the CF partition.

The CICS/DBCTL workload made extensive use of the coupling technology, both in use of CF functions and in the workload devoted entirely to data sharing transactions. Based on the access rates to the CF that we observed for this workload, the following formula can be used to calculate the minimum number of CPs needed on a CF:

$$MinimumCPs = \frac{1}{12} \times \sum_{n} \frac{LSPR_n}{LSPR_{coupling\ facility}} \times CP_n$$

$$\frac{LSPR_n}{LSPR_{coupling\ facility}} \equiv$$ the LSPR ratio of a single CP for the sending CEC of processor family n to the single CP of the CF processor family

$CP_n \equiv$ the number of CPs for all the sending CECs of type n

This number should be rounded to the next largest whole number.

For example, for the largest configuration described in Chapter 7, “CICS/DLI Asymmetric Configuration Performance Study” on page 91, we had a 9021-821 and eight 9672-R61s coupled to IBM 9674-C01 CFs. The formula yields:

Ratio of a single CP of a 9021 to a single CP of a 9672 = 5
Number of 9021 CPs $(1 \times 2) = 2$
Ratio of a single CP of a 9672 to a single CP of a 9672 = 1
Number of 9672 CPs $(6 \times 8) = 48$

$$\begin{aligned} Min\ CPs &= \frac{1}{12} \times (5 \times 2 + 1 \times 48) \\ &= \frac{58}{12} \rightarrow 5\ CPs \end{aligned}$$

These could be divided between two CFs; two CPs on one CF, three CPs on the other CF. You should add additional CPs to handle recovery situations

The LSPR ratios are found in *Large Systems Performance Reference*.

This is a general guideline that can be adjusted up or down, depending on the number of functions that are using structures on the coupling facility. It can also be adjusted for the percentage of the processor that is using the coupling technology.

For the configurations and workloads described in 4.1.5, “Workload Description” on page 68, we used the coupling facility for system management functions (sysplex communication via XCF and JES2 checkpoint data set) and data sharing functions (IRLM locking and IMS OSAM and VSAM buffer cross invalidation). We used 9674s that were defined as a single LPAR partition with all resources (including CPs) dedicated to that partition.

If the coupling facility is a PR/SM partition of a sending CEC, this formula assumes that the CF CPs are defined as dedicated.

As part of a testing scenario, you may partition part of the sending processor to be used as a coupling facility. You can use the sizing formula to determine the number of CPs required in this partition. In this case, you do not have to round to the nearest whole number since you can define a nonintegral portion of the total available CP capacity by using shared logical partitions and assigning processing weights to them.

In a test environment, access rates to the coupling facility will probably not match the CICS/DBCTL workload, so this number may not be valid.

2.8.2 CF Links

In planning for CF links, you have to determine the required speed of the links, the length of the links, and the number of links.

2.8.2.1 Link Speed

IBM currently provides two fiber optic link types for connectivity. The first one is a multimode link with a speed of 50 Megabytes/second, and the second type is a single-mode link with a speed of 100 Megabytes/second. In general, the effect on the performance due to speed of the link depends on the workload sharing data and the processing speeds of the sender CECs and the CF.

For the CICS/DBCTL workload, the performance improvement if a 50 MB link were replaced with a 100 MB link would be negligible. This is primarily because these benchmarks contain CF operations with small data transfers to and from the CF. For this workload, a faster link would have little effect on performance, even in the case of ES/9000 711-based processor CECs connected to the 9674 coupling facility. For the IMS-TM/DB2 and CICS/VSAM RLS workloads, we have a higher frequency of requests and larger data transfers to the CF. In this case, we would recommend single-mode links.

For distances greater than 1km, the higher speed single-mode link is required.

2.8.2.2 Link Distance

Link distance (the length of the Fiber Optic CF link) connecting the sender CEC and the coupling facility is another factor that determines the performance of the individual sending CECs in the a Parallel Sysplex. As a general rule, the hardware component of time spent in the coupling facility (the CF service time shown on the RMF reports) increases 10 microseconds for each kilometer added to the length.

Based on the CICS/DBCTL workload, we would expect to see the following impact on the throughput of the sending CECs as link distances to those CECs increases:

- For the S/390 9672-R1 processors, the throughput on the sending CEC drops by 0.2% per km (up to 10 km) of the link distance.
- For the S/390 9672-R2, -R3, or -R4 processors, the throughput on the sending CEC drops by 0.5% per km (up to 10 km) of the link distance.
- For the ES/9000 711-based processors, the throughput on the sending CEC drops by 1.0% per km (up to 10 km) of the link distance.

The impact on the performance due to link distances should be estimated on a case-by-case basis, as customer workloads may vary.

2.8.2.3 Number of Links

The number of links from the sending processor to the CF is dependent on the size of the sending processor relative to the CF. Based on the CICS/DBCTL workload and a 9674 CF with six CPs, we feel that the following guidelines can be used:

- One link per 9672 Model 1 CEC
- One link per 9121 CEC
- One link for every four CPs on a 9672 Model 2 or 3 CEC
- One link for every three CPs on a 9672 Model 4 CEC
- One link for every two CPs on a 9021 711-based CEC

For the IMS-TM/DB2 and CICS/VSAM RLS workloads, we would double these guidelines.

This is a general guideline that can be adjusted up or down depending on the number of functions that are using structures on the coupling facility. It can also be adjusted for the percentage of the processor that is using the coupling technology. Additional links may also be required for availability considerations.

2.8.2.4 Integrated Coupling Migration Facility

When the Parallel Sysplex consists only of MVS images in a single system, the Integrated Coupling Migration Facility (ICMF) can be used rather than fiber optic links to provide the communication between the MVS images and the CF. The performance characteristics of ICMF are very similar to CF links.

In each of experiments described, the number of CF links is listed. In any case, we recommend that you use two coupling facility links from each sender for availability.

To determine if we have planned the CF resources correctly, we will look at the RMF coupling facility reports.

2.9 Collecting RMF Sysplex Reports

RMF reports on sysplex wide data. The RMF XCF reports were previously available. The RMF coupling facility reports were introduced in RMF 5.1.0. RMF Monitor III collects the data for these reports. In order to combine the data from all the systems, RMF must be synchronized on all systems. The RMF post processor is used to give composite reports. There are several steps needed in this process.

On each system:

- Synchronize RMF monitors with the RMF monitors on the other systems by using the SYNC parameter in the ERBRMFxx parmlib member. You can specify that they be synchronized to the RMF interval or the SMF interval.

  If RMF Mon I is not active, Mon III will synchronize to the SMF interval.

- Collect the RMF SMF records (types 70-78) by using the TYPE parameter in the SMFPRMxx parmlib member.
- Activate RMF monitor III.
- At the end of the interval, dump type RMF SMF records using the SMF dump program, IFASMFDP.

On one system:

- Sort the RMF SMF records from all systems by date and time.
- Run the RMF post processor against the merged data, specifying SYSRPTS(CF) to get the coupling facility reports and REPORTS(XCF) to get the XCF reports.

As an alternative, you can obtain information for the most recent intervals by using the in-storage wrap-around SMF buffer. To do this, you must:

1. Specify SMFBUF in the PARM parameter of the RMF cataloged procedure or specify it as an option on the START or MODIFY command. The default is NOSMFBUF, so specify SMFBUF with the SPACE and/or the RECTYPE subparameter. The default is:

   ```
   SMFBUF(SPACE(32M),RECTYPE(70:78))
   ```

   This means the buffer will contain up to 32 MB of data, and RMF SMF records will be collected.

2. Ensure that the security product on your system has granted you the postprocess authority to access the in-storage buffers —. See DOC APAR II08404.
3. Run the post processor, but do not specify MFPINPUT. RMF will use data from the in-storage buffers of each system in the Parallel Sysplex.

2.10 Tuning Coupling Facilities

The primary objective of tuning coupling facilities is to ensure that we have allocated sufficient quantities of the various resources needed by the coupling technology. The starting point is the RMF Structure Activity report. An example is in Figure 22 on page 45 and Figure 23 on page 46.

In these figures, we see several different kinds of structures:

- IXCPLEX_PATH1 is a LIST structure used for XCF signalling. All requests to this structure will be asynchronous requests.
- COUPLE_CKPT1 is a serialized list structure used for the JES2 checkpoint data.

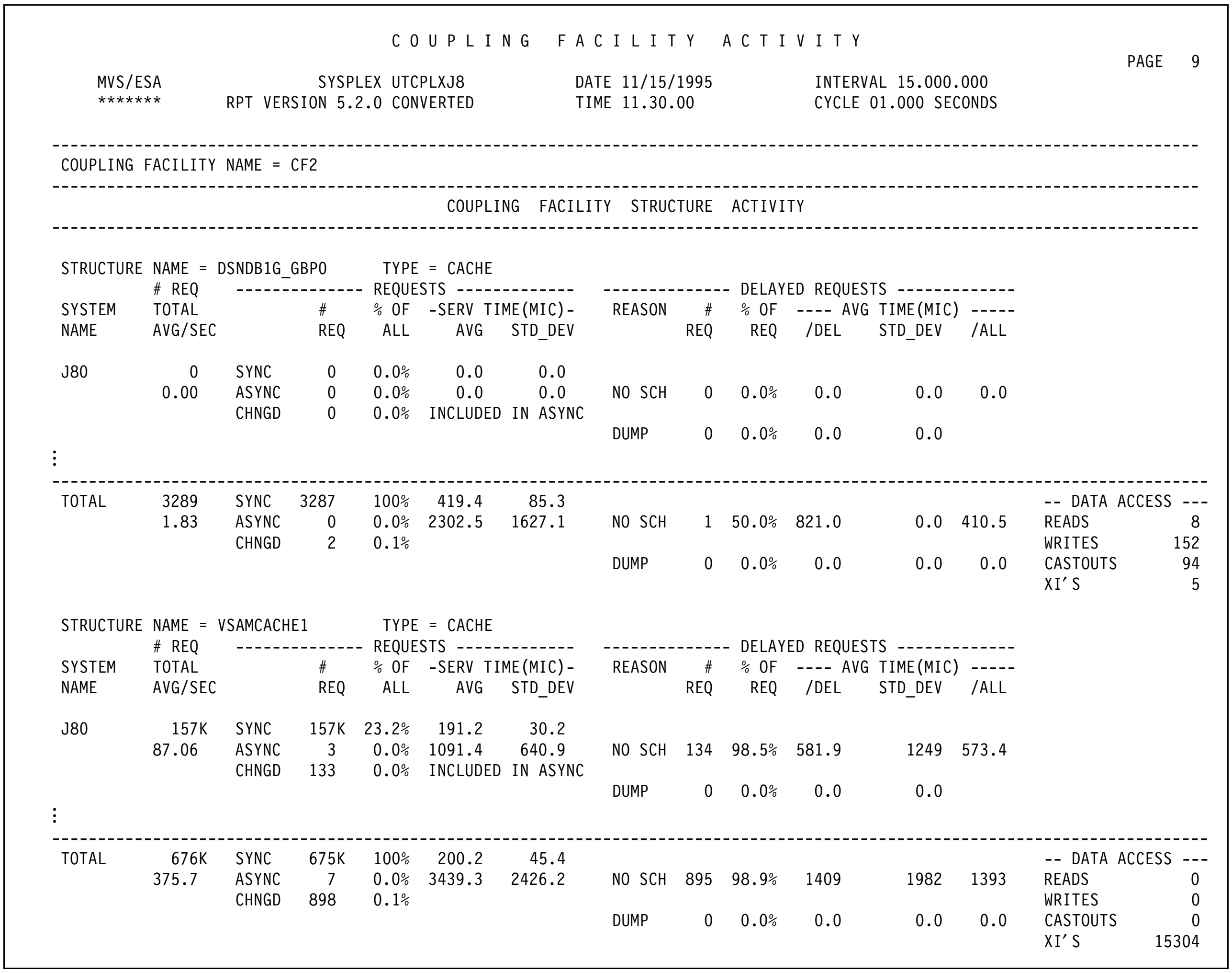
```
 STRUCTURE NAME = IXCPLEX_PATH1     TYPE = LIST
           # REQ    -------------- REQUESTS -------------   -------------- DELAYED REQUESTS -------------
 SYSTEM    TOTAL            #     % OF  -SERV TIME(MIC)-    REASON    #    % OF  ---- AVG TIME(MIC) -----
 NAME      AVG/SEC          REQ    ALL     AVG   STD_DEV            REQ    REQ   /DEL    STD_DEV   /ALL

 J80       28347    SYNC      0    0.0%     0.0      0.0
           15.75    ASYNC    28K  11.0%  2092.3   1573.0    NO SCH 1315   4.6%   1737       1416   80.6
                    CHNGD     0    0.0%  INCLUDED IN ASYNC
                                                            DUMP      0   0.0%    0.0        0.0
⋮
-----------------------------------------------------------------------------------------------------------------------------
 TOTAL       258K   SYNC      0    0.0%     0.0      0.0
           143.1    ASYNC   258K   100%  2817.5    12722    NO SCH   14K  5.3%   7122      66947  378.8
                    CHNGD     0    0.0%
                                                            DUMP      0   0.0%    0.0        0.0    0.0



 STRUCTURE NAME = COUPLE_CKPT1      TYPE = LIST
           # REQ    -------------- REQUESTS -------------   -------------- DELAYED REQUESTS -------------
 SYSTEM    TOTAL            #     % OF  -SERV TIME(MIC)-    REASON    #    % OF  ---- AVG TIME(MIC) -----   EXTERNAL REQUEST
 NAME      AVG/SEC          REQ    ALL     AVG   STD_DEV            REQ    REQ   /DEL    STD_DEV   /ALL   CONTENTIONS

 J80        8463    SYNC   2927    5.8%   306.6    127.8                                                  REQ TOTAL     8339
            4.70    ASYNC  5535   11.0%  1502.1   1263.2    NO SCH 1240  22.4%  508.4      597.7  113.9   REQ DEFERRED   123
                    CHNGD     1    0.0%  INCLUDED IN ASYNC
                                                            DUMP      0   0.0%    0.0        0.0
⋮
-----------------------------------------------------------------------------------------------------------------------------
 TOTAL     50370    SYNC     17K  34.3%   342.2    124.5                                                  REQ TOTAL       50K
           27.98    ASYNC    33K  65.5%  557551G     0.0    NO SCH 8797  26.6%   1047       1245  278.4   REQ DEFFERED   497
                    CHNGD    69    0.1%
                                                            DUMP      0   0.0%    0.0        0.0    0.0



 STRUCTURE NAME = IRLMLOCK1         TYPE = LOCK
           # REQ    -------------- REQUESTS -------------   -------------- DELAYED REQUESTS -------------
 SYSTEM    TOTAL            #     % OF  -SERV TIME(MIC)-    REASON    #    % OF  ---- AVG TIME(MIC) -----   EXTERNAL REQUEST
 NAME      AVG/SEC          REQ    ALL     AVG   STD_DEV            REQ    REQ   /DEL    STD_DEV   /ALL   CONTENTIONS

 J80         384K   SYNC    384K  22.9%   189.5     33.0                                                  REQ TOTAL     383K
           213.2    ASYNC     0    0.0%     0.0      0.0    NO SCH    0   0.0%    0.0        0.0    0.0   REQ DEFERRED   170
                    CHNGD     0    0.0%  INCLUDED IN ASYNC                                                -CONT          170
                                                                                                          -FALSE CONT     39
⋮
-----------------------------------------------------------------------------------------------------------------------------
 TOTAL      1676K   SYNC   1676K   100%   198.1     44.7                                                  REQ TOTAL    1674K
           930.9    ASYNC     0    0.0%     0.0      0.0    NO SCH    0   0.0%    0.0        0.0    0.0   REQ DEFERRED   906
                    CHNGD     0    0.0%                                                                   -CONT          906
                                                                                                          -FALSE CONT    225
```

Figure 22. Example of RMF CF Structure Report

- IRLMLOCK1 is a LOCK structure used to serialize the sharing of the IMS database. All requests to this structure will be synchronous requests.
- VSAMCACHE1 is a CACHE structure used to cache the directories for the IMS databases. It is used as part of IMS buffer cross invalidation to insure currency of data.
- DSNDB1G_GBP0 is a CACHE structure used to cache DB2 data. Some requests to CACHE structures are synchronous; others are asynchronous.

2.10.1 Lock Contention

The first indicator we want to check in this report is the lock contention fields at the far right of the LOCK table structure. Contention is defined as the number of requests delayed (REQ DEFERRED) divided by the number of lock requests (REQ TOTAL).



```
                                     C O U P L I N G   F A C I L I T Y   A C T I V I T Y
                                                                                                                     PAGE   9
     MVS/ESA                 SYSPLEX UTCPLXJ8            DATE 11/15/1995           INTERVAL 15.000.000
     *******       RPT VERSION 5.2.0 CONVERTED           TIME 11.30.00             CYCLE 01.000 SECONDS

-----------------------------------------------------------------------------------------------------------------------------
 COUPLING FACILITY NAME = CF2
-----------------------------------------------------------------------------------------------------------------------------
                                           COUPLING  FACILITY  STRUCTURE  ACTIVITY
-----------------------------------------------------------------------------------------------------------------------------

 STRUCTURE NAME = DSNDB1G_GBP0      TYPE = CACHE
           # REQ    -------------- REQUESTS -------------   -------------- DELAYED REQUESTS -------------
 SYSTEM    TOTAL             #     % OF  -SERV TIME(MIC)-    REASON    #   % OF  ---- AVG TIME(MIC) -----
 NAME      AVG/SEC           REQ    ALL     AVG   STD_DEV            REQ    REQ   /DEL    STD_DEV   /ALL

 J80           0    SYNC      0    0.0%     0.0      0.0
            0.00    ASYNC     0    0.0%     0.0      0.0     NO SCH    0   0.0%    0.0        0.0    0.0
                    CHNGD     0    0.0%  INCLUDED IN ASYNC
                                                             DUMP      0   0.0%    0.0        0.0
⋮
------------------------------------------------------------------------------------------------------------------------------
 TOTAL      3289    SYNC   3287    100%   419.4     85.3                                                    -- DATA ACCESS ---
            1.83    ASYNC     0    0.0%  2302.5   1627.1     NO SCH    1  50.0%  821.0        0.0  410.5    READS           8
                    CHNGD     2    0.1%                                                                     WRITES        152
                                                             DUMP      0   0.0%    0.0        0.0    0.0    CASTOUTS       94
                                                                                                            XI′S            5

 STRUCTURE NAME = VSAMCACHE1        TYPE = CACHE
           # REQ    -------------- REQUESTS -------------   -------------- DELAYED REQUESTS -------------
 SYSTEM    TOTAL             #     % OF  -SERV TIME(MIC)-    REASON    #   % OF  ---- AVG TIME(MIC) -----
 NAME      AVG/SEC           REQ    ALL     AVG   STD_DEV            REQ    REQ   /DEL    STD_DEV   /ALL

 J80         157K   SYNC    157K  23.2%   191.2     30.2
           87.06    ASYNC     3    0.0%  1091.4    640.9     NO SCH  134  98.5%  581.9       1249  573.4
                    CHNGD   133    0.0%  INCLUDED IN ASYNC
                                                             DUMP      0   0.0%    0.0        0.0
⋮
------------------------------------------------------------------------------------------------------------------------------
 TOTAL       676K   SYNC    675K   100%   200.2     45.4                                                    -- DATA ACCESS ---
           375.7    ASYNC     7    0.0%  3439.3   2426.2     NO SCH  895  98.9%   1409       1982   1393    READS           0
                    CHNGD   898    0.1%                                                                     WRITES          0
                                                             DUMP      0   0.0%    0.0        0.0    0.0    CASTOUTS        0
                                                                                                            XI′S        15304
```

Figure 23. Example of RMF CF Structure Report (cont.)

Note: The lock requests displayed on the RMF CF report include both lock obtains and lock releases except in the case of false contention, where only lock obtains are counted. Higher lock contention can result in an increase in CPU utilization and a reduction in throughput.

The field REQ DEFERRED is the subset of the number of lock requests. It represents the number of requests that were unable to be immediately completed beneath the request issuer′s thread. These include any requests that required additional processing to complete.

The number of requests with contention (CONT) is a subset of the number of request delays (REQ DEFERRED). It represents the number of requests that were in fact delayed due to contention on the lock (whether that is true resource contention or false contention). Except for some unusual exception cases, CONT will be the same as REQ DEFERRED. For the various workloads, we have different expectations for true contention: for the CICS/DBCTL, the CICS/VSAM RLS and the GRSSTAR workload, we had less than 1% contention. For the IMS-TM/DB2, the contention was less than 2%.

- If the CONT count is too high, you may want to check the other applications that are running on the systems. In some cases, batch applications that share the databases with online applications hold locks for a much longer time. The time that the lock is held by the batch program can be shortened by taking more frequent checkpoints.

- False lock contention occurs when the hashing algorithm hashes to the same lock table entry (hash value) for two different locks. The requests with false contention (FALSE CONT) are a subset of the requests with contention (CONT). As guidelines, we recommend less than 0.1% contention for the CICS/DBCTL, the CICS/VSAM RLS and the GRSSTAR workload and 1% for the IMS-TM/DB2. False contention can be decreased by increasing the size of the lock structure.

2.10.2 Serialized List Contention

A serialized list structure contains an associated lock structure that is completely under the control of the exploiter. If this lock is held by some other accessor of the structure, XES will queue the request on a multisystem FIFO queue. The number of these requests that are deferred is reported as REQ DEFERRED.

2.10.3 Service Time

Service time is accumulated from the time MVS issues a CF command to the coupling facility until the return from the command is recognized by MVS. It includes the time spent on the coupling facility links, the processing within the coupling facility, and time for MVS to recognize that the command has completed. This service time is one of the key indicators used to determine if the coupling facility is tuned.

Service time in microseconds is recorded for each structure used by each system. The service times for synchronous requests that have been changed to asynchronous requests are included in the asynchronous request service times.

Guidelines for 9674 CF:

- For synchronous requests (SYNCH) of 0 to 4 KB and 9672-R1 sender CECs, the service times should range from 250 to 350 microseconds, depending on the amount of data transferred. Lock requests transfer very little data, so their service times should fall into the low end of the range.
- For asynchronous requests (ASYNCH) of 0 to 64 KB and 9672-R1 sender CECs, the service times should range from 1500 to 5000 microseconds, depending on the amount of data transferred.

  See Appendix A, "CICS/DBCTL Workload Details and Coupling Efficiency Details" on page 171 for details about how to adjust these times for different types of sending CECs and CFs.

2.10.4 CHNGD Requests

Even if the service times fall within the guidelines, there is one additional indicator to check.

Asynchronous requests come from a variety of sources. Some requests are issued as asynchronous commands. If MVS determines that a synchronous request will be delayed (probably because the subchannels are busy), it will change the request to an asynchronous request. It is these requests that are counted in the CHNGD field of the RMF Coupling Facility Structure Activity report.

Based on our experience, we recommend that the total number of requests CHNGD be less than 10% of the total requests for a given structure. For the DSNDB1G_GBP0 and VSAMCACHE1 structures in Figure 22 on page 45, the % OF ALL CHNGD REQUESTS is less than 10% of the total requests.

If this field is greater than 10% you may need additional CF links. See 2.10.4.2, “Additional CF Links Needed?” on page 50.

Note: CF requests that transport more than 4 KB of data are converted to asynchronous requests before they are issued. These are not included in the CHNGD counts.

Note: The CHNGD field should be 0 for XCF LIST structures. It should also be 0 for LOCK structures on 9674 coupling facilities.

If the service times are too high or too many synchronous requests are being changed to asynchronous requests, you may need additional CF resources or additional CF links on the sending CECs.

2.10.4.1 Additional CF Resources Needed?

If both the asynchronous and synchronous service times for the various structures exceed the recommended values, the next thing to check is the coupling facility utilization. The RMF Coupling Facility Usage Summary shows the average utilization of all the processors in the coupling facility.

Note: If SMF records from only one system in a sysplex are used as input to the RMF post processor, data about requests is only for that system. Most other fields on this report give data for the entire sysplex.

For the CICS/DBCTL workload, we observed CF utilizations as high as 50% with minimal effect on service time. It may be possible to run at higher CF utilizations, but we have not verified this.

Note: In an ICMF environment, the correct CF utilization is obtained from the RMF Partition Data report.

If the utilization increases seem to be elongating service time, there are various actions that can be taken, depending on the environment:

- Verify that the number of logical processors defined is correct.

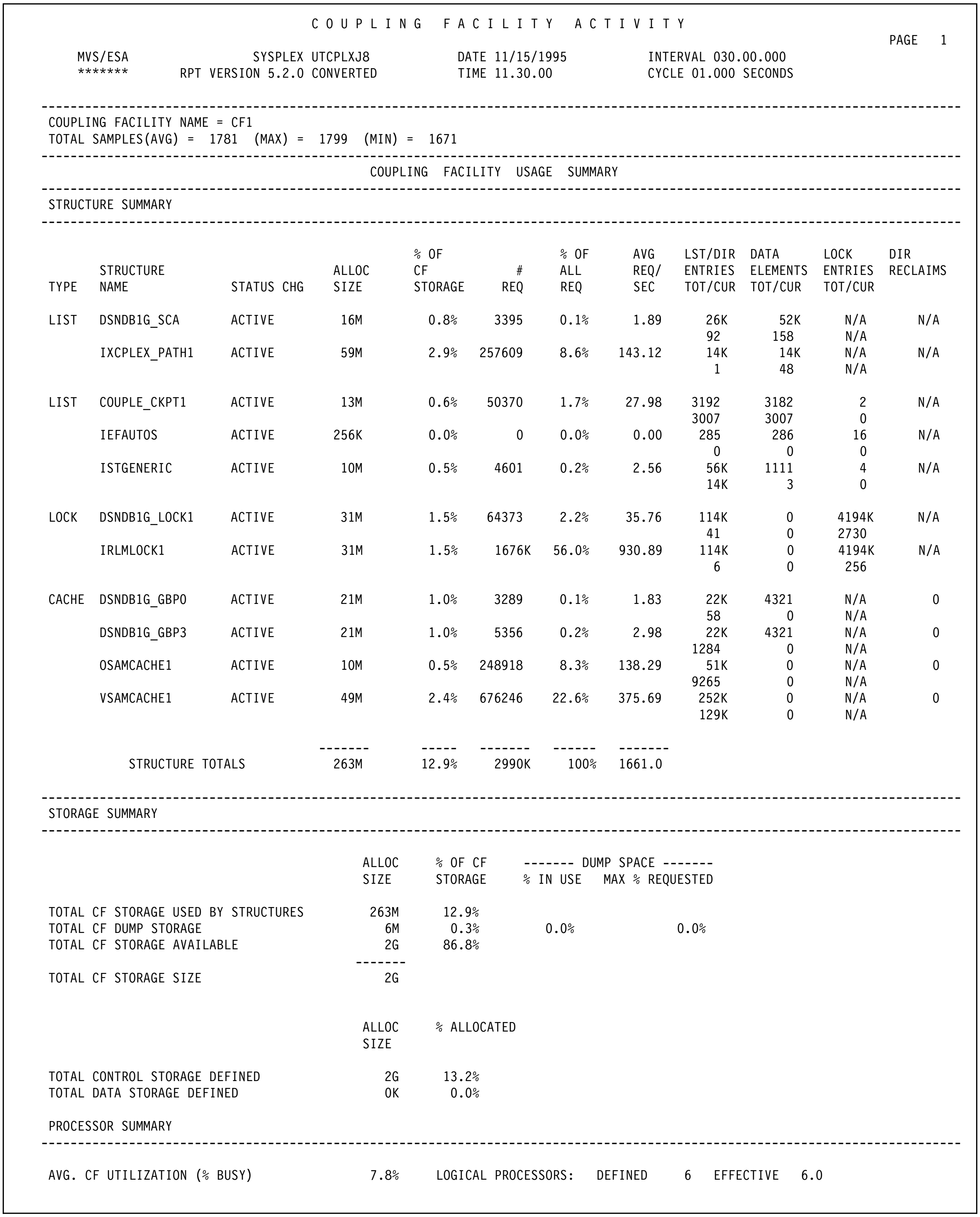
  There have been instances when the CF was configured to one logical processor during service and was not reset.

- Verify that the processor resource allocated to the CF is what you expected.

  If this is an LPAR environment and the CF partition is sharing CPs, the processor resource will be limited by the definition of the WEIGHT and CAP parameters and contention from the other partitions. The amount of processor resource available to the CF is reported as LOGICAL PROCESSOR EFFECTIVE.

  Note: The CF LIC code runs in an “active-wait” polling loop, so it is always active. If multiple CFs share logical processors, each will use as much processor resource as it can get, even if it has no work to process. So a test CF that shares logical processors with a production CF can divert processor resource from the production CF. If there are only CF partitions on this CEC, we recommend that you dedicate the CPs.

  In an LPAR environment with MVS images and CF partitions on the same CEC, you will get the best response time by dedicating the CPs to the CF



```
                                     C O U P L I N G   F A C I L I T Y   A C T I V I T Y
                                                                                                                   PAGE   1
     MVS/ESA                 SYSPLEX UTCPLXJ8           DATE 11/15/1995           INTERVAL 030.00.000
     *******       RPT VERSION 5.2.0 CONVERTED          TIME 11.30.00             CYCLE 01.000 SECONDS

-----------------------------------------------------------------------------------------------------------------------------
 COUPLING FACILITY NAME = CF1
 TOTAL SAMPLES(AVG) =  1781  (MAX) =  1799  (MIN) =  1671
-----------------------------------------------------------------------------------------------------------------------------
                                             COUPLING  FACILITY  USAGE  SUMMARY
-----------------------------------------------------------------------------------------------------------------------------
 STRUCTURE SUMMARY
-----------------------------------------------------------------------------------------------------------------------------

                                                   % OF               % OF      AVG    LST/DIR  DATA     LOCK      DIR
        STRUCTURE                       ALLOC      CF            #    ALL       REQ/   ENTRIES  ELEMENTS ENTRIES   RECLAIMS
 TYPE   NAME              STATUS CHG    SIZE       STORAGE     REQ    REQ       SEC    TOT/CUR  TOT/CUR  TOT/CUR

 LIST   DSNDB1G_SCA       ACTIVE         16M         0.8%     3395     0.1%      1.89      26K       52K      N/A       N/A
                                                                                           92       158      N/A
        IXCPLEX_PATH1     ACTIVE         59M         2.9%   257609     8.6%    143.12      14K       14K      N/A       N/A
                                                                                            1        48      N/A

 LIST   COUPLE_CKPT1      ACTIVE         13M         0.6%    50370     1.7%     27.98    3192      3182         2       N/A
                                                                                         3007      3007         0
        IEFAUTOS          ACTIVE        256K         0.0%        0     0.0%      0.00     285       286        16       N/A
                                                                                            0         0         0
        ISTGENERIC        ACTIVE         10M         0.5%     4601     0.2%      2.56      56K     1111         4       N/A
                                                                                           14K        3         0

 LOCK   DSNDB1G_LOCK1     ACTIVE         31M         1.5%    64373     2.2%     35.76     114K        0     4194K      N/A
                                                                                           41         0      2730
        IRLMLOCK1         ACTIVE         31M         1.5%     1676K   56.0%    930.89     114K        0     4194K      N/A
                                                                                            6         0       256

 CACHE  DSNDB1G_GBP0      ACTIVE         21M         1.0%     3289     0.1%      1.83      22K     4321      N/A         0
                                                                                           58         0      N/A
        DSNDB1G_GBP3      ACTIVE         21M         1.0%     5356     0.2%      2.98      22K     4321      N/A         0
                                                                                         1284         0      N/A
        OSAMCACHE1        ACTIVE         10M         0.5%   248918     8.3%    138.29      51K        0      N/A         0
                                                                                         9265         0      N/A
        VSAMCACHE1        ACTIVE         49M         2.4%   676246    22.6%    375.69     252K        0      N/A         0
                                                                                          129K        0      N/A

                                      -------       -----   -------   ------  -------
            STRUCTURE TOTALS            263M        12.9%     2990K     100%   1661.0

-----------------------------------------------------------------------------------------------------------------------------
 STORAGE SUMMARY
-----------------------------------------------------------------------------------------------------------------------------

                                            ALLOC     % OF CF     ------- DUMP SPACE -------
                                            SIZE      STORAGE     % IN USE   MAX % REQUESTED

 TOTAL CF STORAGE USED BY STRUCTURES         263M      12.9%
 TOTAL CF DUMP STORAGE                         6M       0.3%         0.0%              0.0%
 TOTAL CF STORAGE AVAILABLE                    2G      86.8%
                                           -------
 TOTAL CF STORAGE SIZE                         2G


                                            ALLOC     % ALLOCATED
                                            SIZE

 TOTAL CONTROL STORAGE DEFINED                 2G      13.2%
 TOTAL DATA STORAGE DEFINED                    0K       0.0%

 PROCESSOR SUMMARY
-----------------------------------------------------------------------------------------------------------------------------

 AVG. CF UTILIZATION (% BUSY)                7.8%     LOGICAL PROCESSORS:   DEFINED    6   EFFECTIVE   6.0
```

Figure 24. Example of RMF CF Usage Report

partition. If this is not possible, consider increasing the capacity of the CF partition by insuring that it is not capped and by increasing its processing weight.

Note: For a more detailed explanation of LPAR considerations in a Parallel Sysplex Environment, see WSC FLASH 9609.

- If this is a 9672-R1, -R2, or -R3, and the processor is being used solely as a coupling facility, check the Customize Operating Environment Panel to verify that the CF MODE has been specified (rather than PROCESSOR INTENSIVE or I/O INTENSIVE).
- If you have multiple stand-alone CFs in your configurations, check the distribution of the structures.

  If one CF has much higher utilization than the other, redistribute the structures based on ALLOC SIZE and # REQ. For an example of redistributing structures, look at Figure 24 on page 49. We might want to distribute these structures evenly across two CFs. Based on the size of the structures and the AVG REQ/SEC, we would probably move the two LOCK structures and the SCA to another CF. This would balance the utilization on the two CFs.
- If all else fails, you can increase the capacity of the CF by adding CPs or obtaining an additional CF.

2.10.4.2 Additional CF Links Needed?

Another possible cause for higher service times and excessive queuing may be a shortage in the number of CF links defined from the sending CECs.

For every CF link, there are two CF subchannels where data is placed to be sent across coupling facility links. When you define a coupling facility link in the IOCDS, HCD automatically defines the two subchannels.

Figure 25 on page 51 shows a sample RMF subchannel report. Examine this report for the following:

- Number of subchannels

  The number of subchannels configured (CONFIG) tells you how many subchannels were defined in the IOCDS (SCH GEN), how many are currently being used (SCH USE), how many could be used if they were available given the current pathing configuration (SCH MAX), and how many coupling facility links are currently connected (PTH). You should check this information and verify that:

  – The correct number of subchannels are connected.

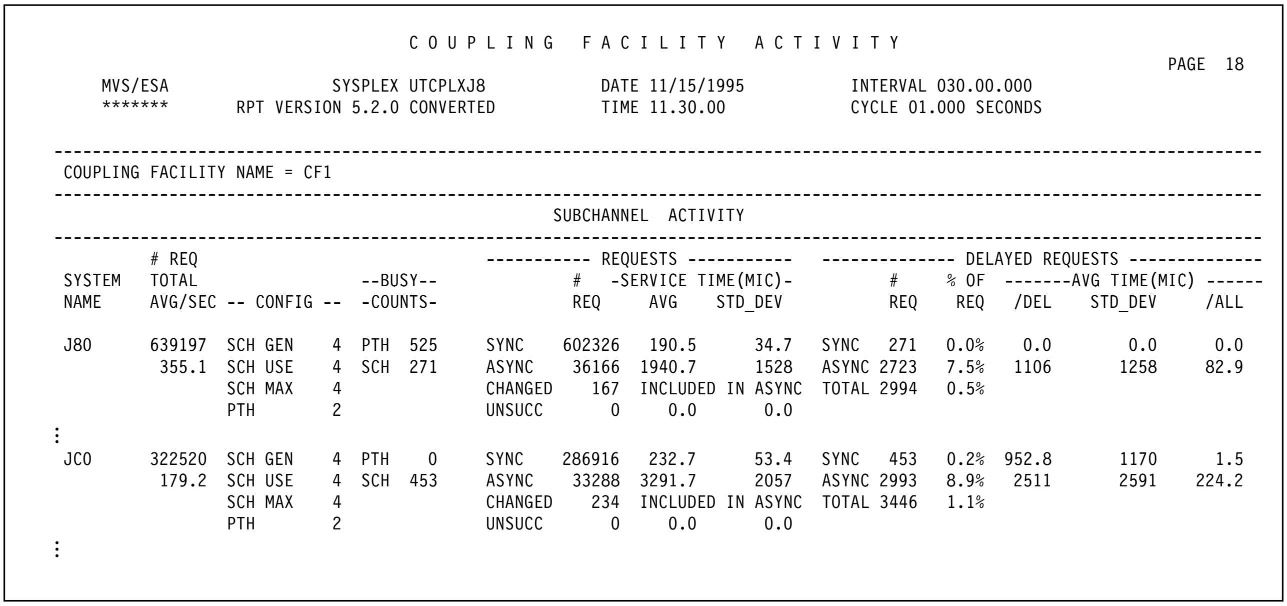
    If the number of subchannels currently being used are fewer than the number of subchannels defined, verify that you have not lost connectivity on some CF links.

  – The correct number of subchannels have been defined.

    If more subchannels have been defined than you intend to use, you could reduce the number generated and save a small amount of storage.

- Subchannel busy

  For each CF link, there are two subchannels for each MVS image. The data to be sent to the CF is loaded into these subchannels. If no subchannels are available, ASYNC requests are queued. Non-immediate SYNC requests are changed to ASYNC requests (which are then also queued). Immediate SYNC requests (like locks) “spin,” waiting for the next available subchannel. Data about the delayed requests is reported under the SYNC and ASYNC DELAYED REQUESTS and summarized on the TOTAL line.



```
                                     C O U P L I N G   F A C I L I T Y   A C T I V I T Y
                                                                                                                   PAGE  18
     MVS/ESA                 SYSPLEX UTCPLXJ8            DATE 11/15/1995          INTERVAL 030.00.000
     *******       RPT VERSION 5.2.0 CONVERTED           TIME 11.30.00            CYCLE 01.000 SECONDS

-----------------------------------------------------------------------------------------------------------------------------
 COUPLING FACILITY NAME = CF1
-----------------------------------------------------------------------------------------------------------------------------
                                                    SUBCHANNEL  ACTIVITY
-----------------------------------------------------------------------------------------------------------------------------
          # REQ                              ----------- REQUESTS -----------   -------------- DELAYED REQUESTS --------------
 SYSTEM   TOTAL                 --BUSY--              #   -SERVICE TIME(MIC)-          #    % OF  -------AVG TIME(MIC) ------
 NAME     AVG/SEC -- CONFIG --  -COUNTS-              REQ     AVG    STD_DEV           REQ   REQ   /DEL    STD_DEV     /ALL

 J80      639197  SCH GEN    4  PTH  525     SYNC    602326   190.5      34.7   SYNC   271  0.0%    0.0        0.0      0.0
           355.1  SCH USE    4  SCH  271     ASYNC    36166  1940.7      1528   ASYNC 2723  7.5%   1106       1258     82.9
                  SCH MAX    4               CHANGED    167  INCLUDED IN ASYNC  TOTAL 2994  0.5%
                  PTH        2               UNSUCC       0     0.0       0.0
⋮
 JC0      322520  SCH GEN    4  PTH    0     SYNC    286916   232.7      53.4   SYNC   453  0.2%  952.8       1170      1.5
           179.2  SCH USE    4  SCH  453     ASYNC    33288  3291.7      2057   ASYNC 2993  8.9%   2511       2591    224.2
                  SCH MAX    4               CHANGED    234  INCLUDED IN ASYNC  TOTAL 3446  1.1%
                  PTH        2               UNSUCC       0     0.0       0.0
⋮
```

Figure 25. Example of RMF CF Subchannel Activity Report

For those requests that experience a delay, the duration of the delay is reported separately (/DEL). The subchannel delay for each type of request is amortized over all the requests of each type and reported in /ALL. You can assess the impact of these delays by adding this to the service time for that type of request.

As a guideline, we recommend that sum of the SYNC and ASYNC requests delayed (TOTAL – % OF REQ) be less than 10% of all requests. If the percentage of requests queued is greater than 10%, you should consider adding another CF link on the related sending CEC. If the coupling facility is totally configured for coupling facility paths, you may want to consider moving a structure to another coupling facility or adding another coupling facility.

- Path busy

When a CF request obtains a subchannel, in most cases, it will proceed down the path to the coupling facility and complete the processing with no further delays. However, if this is a PR/SM environment with multiple MVS images sharing coupling facility links, the request could encounter a busy path. If this is a SYNC request, the request is immediately retried until it obtains a path. The time spent spinning for the path is accumulated in the SYNC service time. If this is an ASYNC request, the request is returned and goes back through the process of obtaining another subchannel, which may include requeuing for a subchannel.

The number of times a “Path Busy” condition is encountered is reported in the BUSY COUNTS — PTH. To limit the additional overhead incurred in processing requests deferred for Path Busy, we recommend that the percentage of requests encountering this delay be limited to 10%. If it exceeds this amount, you may want to consider dedicating the CF links to each MVS image or adding additional CF links.

2.11 Additional Information

Additional information about the data in the RMF reports can be found in the following publications:

- *MVS/ESA Analyzing Resource Measurement Facility Version 5 Reports* (LY33-9178-04) gives a detailed description of each field on the report as well as tuning suggestions.
- *MVS/ESA Analyzing Resource Measurement Facility Version 5 —. Getting Started on Performance Management* (LY33-9176-00) is a top-down discussion of key fields used in tuning.
- *DB2 for MVS/ESA Version 4 Data Sharing — Planning and Administration* (SC26-3269-01) gives additional tuning information for the DB2 environment.
- *Washington System Center Flash 9609, CF Reporting Enhancements to RMF 5.1* gives a more detailed explanation of the various fields on the RMF CF reports.
- *Washington System Center Flash 9609, LPAR Performance in a Parallel Sysplex Environment* discusses considerations for the LPAR environment.

Chapter 3. Customer Data Sharing Experiences

With over 300 production data sharing Parallel Sysplex implementations worldwide, there is a wealth of performance data available, clearly showing the effects of data sharing on performance.

This chapter looks at the reality of Parallel Sysplex performance in a selection of these data sharing environments. Over the last 18 months, the S/390 Performance team, in conjunction with the ITSO in Poughkeepsie, have analyzed sets of customer data from production data sharing customers. The results of this analysis are presented in this chapter.

Of the customers studied, the results from seven of the customers are presented in detail in 3.4, "Detailed Results" on page 55. The customers involved in this chapter come from many different countries and cover the majority of the major business areas including:

- Banking
- Manufacturing
- Insurance
- Transport
- Telecommunications
- Distribution
- Financial consultancy

As can be seen in Table 1 on page 54, there is a spread of customer size in the data presented, varying from 1700 MIPS down to only 150 MIPS, so the data represents a wide range in the size of customer Parallel Sysplex implementation, in addition to a geographical spread. Additionally, all of the major IBM data sharing subsystems are represented in the customers surveyed.

3.1 Introduction

The analysis was done using a data sharing overhead methodology developed by the IBM S/390 Division, located in Poughkeepsie, New York.

Most of the cost of data sharing may be accounted for in three components: global locking (use of a coupling facility lock structure to prevent transactions running on different images from updating the same record), global lock contention (use of XCF messaging among local lock managers to resolve the case when transactions on different images want to update the same record at the same time), and local buffer coherency (use of a coupling facility cache structure to insure updated records are reflected in all images′ local buffer pools).

It is straightforward to estimate these costs for a system that is currently data sharing. The RMF Coupling Facility Activity Report gives the frequency and hardware cost for global locking (from the lock structure activity report), local buffer coherency (from the cache structure activity report) and the frequency of global lock contention (from lock structure activity report).

A software cost (path length within the lock managers, buffer managers and MVS services) for each function has been determined through laboratory measurements. Thus, the cost of data sharing can be estimated by multiplying the frequency of each operation times the software plus the hardware cost, and summing for all operations.

IBM benchmarks were run in order to calibrate the model, which was then used to analyze the customer workloads.

This methodology based on coupling facility activity can be applied to the “after” case without concern for whether the application is the only one running, or whether the workload changed from the “before” case. Of course, as the workload continues to change after the data sharing migration (or more applications are migrated to data sharing), the calculations may be repeated to keep the overhead estimation current.

For some worked examples of the use of this methodology see Appendix A, “CICS/DBCTL Workload Details and Coupling Efficiency Details” on page 171, Appendix B, “IMS-TM/DB2 Workload Details and Coupling Efficiency Details” on page 177 and Appendix C, “CICS/VSAM Workload Details and Coupling Efficiency Details” on page 183.

Attention

You will find quoted in this chapter, in various places, a processor utilization value in MIPS (Million Instructions Per Second). For the purpose of this document, the term MIPS relates to the published IBM Large Scale Processor Performance Report (LSPR) processor ratios using a base value of 63 MIPS for an IBM 9672 R15.

3.2 Summary

The results of applying the methodology to the customer production data is summarized in Table 1. Detailed data for samples I through to Q are provided in Table 2 on page 56 through to Table 10 on page 64 in 3.4, “Detailed Results” on page 55.

Table 1 (Page 1 of 2). Customer Data Sharing Overhead Experience

	Industry	# Images	Sender CPCs	MIPS Used	% Total MIPS	Exploiter TM/DB	Over -head %
A	Banking	4	711, 9672-G1 G2 G3	650	66	CICS/IMS	11
B	Manufacturing	3	711, 9672-G2 G3	700	78	CICS/DB2	10
C	Banking	8	711, 9672-G1 G2 G3	750	63	CICS/IMS	9
D	Banking	2	711	800	82	IMS/IMS	7
E	Insurance	9	9672-G2 G3	1500	74	CICS/IMS	10
F	Banking	4	711, 9672-G3	1500	99	IMS/IMS+DB2	11
G	Transport	3	711	1400	96	CICS/DB2	8
H	Banking	2	711	800	91	IMS/IMS+DB2	9
I	Banking	7	711, 9672-G3 G4	1020	60	CICS/IMS	10
J	Banking	7	711, 9672-G3 G4	1250	70	CICS/IMS	4
K	Banking	2	9672-G4	280	30	CICS/DB2+IMS+RLS	4
L	Telecom	3	9672-G3	150	15	CICS/DB2+RLS	5

Table 1 (Page 2 of 2). Customer Data Sharing Overhead Experience

	Industry	# Images	Sender CPCs	MIPS Used	% Total MIPS	Exploiter TM/DB	Over -head %
M	Distribution	3	711, 9672-G2 G4	340	75	CICS/DB2	2
N	Banking	2	711, 9672-G3	390	62	CICS/DB2	7
O	Banking	2	711, 9672-G3	450	72	CICS/DB2	2
P	Consultant	4	9672-G4	1700	95	CICS/DB2	5
Q	Distribution	2	711, 9672-G3	530	60	IMS/IMS+DB2	5

3.3 Conclusions

One of the principal reasons behind IBM performance benchmarks is to establish the absolute service times and associated costs for specific system-related functions.

In the case of Parallel Sysplex, this has enabled us to calibrate a model that, when applied to data from customer workloads, shows a realistic view of Parallel Sysplex data sharing performance.

The results from the customers analyzed show an average data sharing cost or overhead of 7% of the total capacity used to process the customers′ workload. This overhead is approximately half of that shown in the IBM-run benchmarks described in this book.

The key reasons for the production customer data showing significantly lower overheads than the IBM benchmark data are described in 1.2, “Comparison of Customer and IBM Benchmark Data” on page 2. In brief, the reasons are based on the percentage of data being shared, typically less than 50% in customers and 100% in the benchmarks, and also in the CF access rates which are typically lower in the customer cases than the IBM benchmarks, due to significantly greater business logic in the applications.

3.4 Detailed Results

This section contains some additional details for seven of the customers analyzed. Information such as the following is provided:

- CF configuration
- CF access rates
- RMF report information, and
- Processor types and percent busy

This information is provided as an illustration of the values seen in these customer environments. They are sensitive to the workload, as you might expect, so direct comparison between one set of data and another is unlikely to be of great value, since no two customer workloads are the same.

Sample I and J represent prime shift and batch processing from the same customer. This is also the case for Samples N and O.

Note: In some cases, only one 15-minute interval was provided by the customer, representing their data sharing environment. In such cases, it was confirmed that the sample provided was a good representation of their system.

3.4.1 Customer Sample I Details

- RMF report from 04/15/98 at 00:00 am for 120 minutes
- Data sharing TM/DB was CICS/IMS
- Other CF exploiters were RACF and VTAM GR
- CF LPAR 1 was on a 6-way C01 and was 19.2% busy
- CF LPAR 1 had 75 MB (out of 231 MB) allocated for structures
- CF LPAR 2 was on a 6-way C01 and was 1.5% busy[2]
- CF LPAR 2 had 42 MB (out of 231 MB) allocated for structures
- Total CF access rate was 11741 per second[3]
- Total Data Sharing CF access rate was 11377 per second
- Data Sharing CF access rate, was 11.5 per second per used MIP
- Data Sharing Overhead was calculated to be 9.4%, of total used capacity

Table 2. Customer I Data Sharing Details

	SYS1	**SYS2**	**SYS3**	**SYS4**	**SYS5**	**SYS6**	**SYS7**
ISC links	2	2	2	2	2	2	2
Processor	9021-982	9672-R44	9672-R34	9672-RX4	9672-R34	9672-R45	9672-R55
Busy %	94.1	34.0	21.3	56.1	21.3	88.2	49.9
MIPS used	385	58	28	191	28	183	132
Overhead MIPS	16.9	10.7	1.4	31.2	1.2	0.4	32.9
Overhead %	4.4	18.4	4.9	16.3	4.2	0.2	24.8
IMS LOCK rate	1217	1010	81.2	4028	69	26.8	3442
Contention count	1401	618	424	4708	547	398	3759
VSAM CACHE rate	455	224	51.8	387	47.0	13.9	326

[2] CF LPAR1 is significantly more busy than LPAR2 because the IMS LOCK structure CFIRLM000 accounts for more than 90% of all CF accesses across both CF LPARs.

[3] The total CF access rate is ***all*** accesses to the CF, including those that are not related to data sharing. Examples of CF accesses that are not related to data sharing include JES2 checkpoint, RACF, System operlog, GRS STAR, and CICS logging.

3.4.2 Customer Sample J Details

- RMF report from 04/15/98 at 09:00 am for 120 minutes
- Data sharing TM/DB was CICS/IMS
- Other CF exploiters were RACF and VTAM GR
- CF LPAR 1 was on a 6-way C01 and was 12.7% busy
- CF LPAR 1 had 75 MB (out of 231 MB) allocated for structures
- CF LPAR 2 was on a 6-way C01 and was 5.0% busy
- CF LPAR 2 had 42 MB (out of 231 MB) allocated for structures
- Total CF access rate was 5606 per second[4]
- Total data sharing CF access rate was 4374 per second
- Data sharing CF access rate was 11.5 per second per used MIP
- Data sharing overhead was calculated to be 3.4% of total used capacity

Table 3. Customer J Data Sharing Details

	SYS1	SYS2	SYS3	SYS4	SYS5	SYS6	SYS7
ISC links	2	2	2	2	2	2	2
Processor	9021-982	9672-R44	9672-R34	9672-RX4	9672-R34	9672-R44	9672-R54
Busy %	98.1	58.3	55.3	72.6	58.2	56.7	73.8
MIPS used	402	99	73	247	77	118	196
Overhead MIPS	5.6	5.9	3.4	9.8	3.4	6.1	6.8
Overhead %	1.4	6.0	4.7	4.0	4.4	5.2	3.5
IMS LOCK rate	467	411	212	971	214	377	474
Contention count	115	671	606	1528	384	1083	1289
VSAM CACHE rate	111	211	130	243	119	230	205

[4] The total CF access rate is ***all*** accesses to the CF, including those that are not related to data sharing. Examples of CF accesses that are not related to data sharing include JES2 checkpoint, RACF, System operlog, GRS STAR, and CICS logging.

3.4.3 Customer Sample K Details

- RMF report from 03/14/98 at 10:00 am for 60 minutes
- Data sharing TM/DB was CICS/DB2, IMS(VSAM) and RLS
- Other CF exploiters were JES2, OPERLOG and RACF
- CF LPAR 1 was on a 1-way C05 and was 5.4% busy
- CF LPAR 1 had 492 MB (out of 949 MB) allocated for structures
- CF LPAR 2 was on a 1-way C05 and was 6.1% busy
- CF LPAR 2 had 225 MB (out of 949 MB) allocated for structures
- Total CF access rate was 1625 per second[5]
- Total data sharing CF access rate was 1336 per second
- Data sharing CF access rate, was 4.9 per second per used MIP
- Data sharing overhead was calculated to be 3.8% of total used capacity

Table 4. Customer K Data Sharing Details

	SYS1	**SYS2**
ISC links	2	2
Processor	9672-RY5	9672-RY5
Busy %	38.0	24.1
MIPS used	170	108
Overhead MIPS	2.7	7.8
Overhead %	1.6	7.2
DB2 LOCK rate	71.75	104.3
Contention count	10485	10902
GBP0 CACHE rate	0.463	0.295
GBP1 CACHE rate	2.42	46.9
GBP2 CACHE rate	41.6	84.2
IMS LOCK rate	61.5	73.4
Contention count	13	16
VSAM CACHE rate	44.93	54.46
RLS LOCK rate	58.8	490
Contention count	292	392
VSAM1 CACHE rate	9.04	19.7
VSAM2 CACHE rate	58.1	114

[5] The total CF access rate is ***all*** accesses to the CF, including those that are not related to data sharing. Examples of CF accesses that are not related to data sharing include JES2 checkpoint, RACF, System operlog, GRS STAR, and CICS logging.

3.4.4 Customer Sample L Details

- RMF report from 04/15/98 at 22:00 pm for 30 minutes
- Data sharing TM/DB was CICS/DB2 and RLS
- Other CF exploiters were JES2, GRS STAR and CICS LOGSTREAM
- CF LPAR 1 was on a 6-way C04 and was 1.0% busy
- CF LPAR 1 had 779 MB (out of 2000 MB) allocated for structures
- CF LPAR 2 was on a 6-way C04 and was 0.2% busy
- CF LPAR 2 had 448 MB (out of 2000 MB) allocated for structures
- Total CF access rate was 334 per second[6]
- Total data sharing CF access rate was 281 per second
- Data sharing CF access rate was 1.9 per second per used MIP
- Data sharing overhead was calculated to be 4.5% of total used capacity

Table 5. Customer L Data Sharing Details

	SYS1	**SYS2**	**SYS3**
ISC links	2	2	2
Processor	9672-RX4	9672-RX4	9672-R94
Busy %	15.4	15.2	15.5
MIPS used	50.4	49.7	48.7
Overhead MIPS	2.41	2.14	2.20
Overhead %	4.79	4.31	4.52
DB2 LOCK rate	38.0	29.4	27.4
Contention count	4670	5559	5638
GBP0 CACHE rate	0.61	0.78	0.57
GBP1 CACHE rate	0.17	0.17	0.17
GBP2 CACHE rate	20.8	20.3	19.0
GBP3 CACHE rate	43.3	37.8	41.3
RLS LOCK rate	0.04	0.07	0.07
Contention count	0	19	26
VSAM1 CACHE rate	0.20	0.07	0.07
VSAM2 CACHE rate	117.3	177.7	176.0

[6] The total CF access rate is ***all*** accesses to the CF, including those that are not related to data sharing. Examples of CF accesses that are not related to data sharing include JES2 checkpoint, RACF, System operlog, GRS STAR, and CICS logging.

3.4.5 Customer Sample M Details

- RMF report from 03/19/98 at 10:00 am for 30 minutes
- Data sharing TM/DB was CICS/DB2
- Other CF exploiters were JES2
- CF LPAR 1 was on a 1-way C02 and was 7.9% busy
- CF LPAR 1 had 175 MB (out of 492 MB) allocated for structures
- CF LPAR 2 was on a 1-way C02 and was 1.6% busy
- CF LPAR 2 had 70 MB (out of 492 MB) allocated for structures
- Total CF access rate was 625 per second[7]
- Total data sharing CF access rate was 560 per second
- Data sharing CF access rate, was 1.75 per second per used MIP
- Data sharing overhead was calculated to be 1.8%, of total used capacity

Table 6. Customer M Data Sharing Details

	SYS1	**SYS2**	**SYS3**
ISC links	2	2	2
Processor	9021-831	9672-R35	9672-R63
Busy %	89.9	55.8	71.8
MIPS used	152	92.1	85.1
Overhead MIPS	1.4	4.35	0.08
Overhead %	0.93	4.73	0.01
DB2 LOCK rate	95.2	115	4.74
Contention count	115	87.5	215
GBP0 CACHE rate	5.33	1.22	0.6
GBP2 CACHE rate	18.5	255	0.13
GBP3 CACHE rate	12.6	49.4	0.12
GBP4 CACHE rate	0.14	0.09	
GBP5 CACHE rate	0.34	0.11	
GBP6 CACHE rate		0.13	
GBP7 CACHE rate	0.1		
GBP8 CACHE rate	0.11	0.1	1.2
GBP9 CACHE rate			0.47

[7] The total CF access rate is ***all*** accesses to the CF, including those that are not related to data sharing. Examples of CF accesses that are not related to data sharing include JES2 checkpoint, RACF, System operlog, GRS STAR, and CICS logging.

3.4.6 Customer Sample N Details

- RMF report from 03/14/98 at 12:45 pm for 15 minutes
- Data sharing TM/DB was CICS/DB2
- Other CF exploiters were JES2, Tape Allocation
- CF LPAR 1 was on a 1-way C04 and was 15.0% busy
- CF LPAR 1 had 181 MB (out of 949 MB) allocated for structures
- CF LPAR 2 was on a 1-way C04 and was 5.0% busy
- CF LPAR 2 had 70 MB (out of 949 MB) allocated for structures
- Total CF access rate was 1850 per second[8]
- Total data sharing CF access rate was 1810 per second
- Data sharing CF access rate was 4.9 per second per used MIP
- Data sharing overhead was calculated to be 6.3% of total used capacity

Table 7. Customer N Data Sharing Details

	SYS1	**SYS2**
ISC links	2	2
Processor	9021-962	9672-R84
Busy %	85	38
MIPS used	274	114
Overhead MIPS	14.3	10.3
Overhead %	5.2	9.1
DB2 LOCK rate	420.9	419.5
Contention count	28491	25572
GBP0 CACHE rate	7.53	6.76
GBP1 CACHE rate	306.9	268.8
GBP2 CACHE rate	98.8	123.7
GBP3 CACHE rate	81.4	33.6
GBP4 CACHE rate	18.0	7.47

[8] The total CF access rate is ***all*** accesses to the CF, including those that are not related to data sharing. Examples of CF accesses that are not related to data sharing include JES2 checkpoint, RACF, System operlog, GRS STAR, and CICS logging.

3.4.7 Customer Sample O Details

- RMF report from 03/14/98 at 22:00 pm for 15 minutes
- Data sharing TM/DB was CICS/DB2
- Other CF exploiters were JES2, Tape Allocation
- CF LPAR 1 was on a 1-way C04 and was 12.0% busy
- CF LPAR 1 had 181 MB (out of 949 MB) allocated for structures
- CF LPAR 2 was on a 1-way C04 and was 1.0% busy[9]
- CF LPAR 2 had 70 MB (out of 949 MB) allocated for structures
- Total CF access rate was 740 per second[10]
- Total data sharing CF access rate was 690 per second
- data sharing CF access rate was 1.64 per second per used MIP
- Data sharing overhead was calculated to be 1.9% of total used capacity

Table 8. Customer O Data Sharing Details

	SYS1	**SYS2**
ISC links	2	2
Processor	9021-962	9672-R84
Busy %	93.2	48.3
MIPS used	301	145
Overhead MIPS	8.0	0.5
Overhead %	2.6	0.4
DB2 LOCK rate	59.6	14.2
Contention count	1536	1554
GBP0 CACHE rate	3.41	1.22
GBP1 CACHE rate	519.5	27.0
GBP2 CACHE rate	53.2	1.44
GBP3 CACHE rate	8.11	0.29
GBP4 CACHE rate	237.7	211.8

[9] CF LPAR1 is significantly more busy than LPAR2 because the DB2 Global Buffer Pool GBP1 accounts for nearly 60% of all CF accesses across both CF LPARs.

[10] The total CF access rate is ***all*** accesses to the CF, including those that are not related to data sharing. Examples of CF accesses that are not related to data sharing include JES2 checkpoint, RACF, System operlog, GRS STAR, and CICS logging.

3.4.8 Customer Sample P Details

- RMF Report from 03/23/98 at 10:00 am for 60 minutes
- Data sharing TM/DB was CICS/DB2
- Other CF exploiters were JES2, RACF, OPERLOG, LOGREC, GRS STAR, and VTAM GR
- CF LPAR 1 was on a R24 ICF and was 37.3% busy
- CF LPAR 1 had 430 MB (out of 1013 MB) allocated for structures
- CF LPAR 2 was on an R24 ICF and was 17.0% busy
- CF LPAR 2 had 448 MB (out of 1013 MB) allocated for structures
- Total CF access rate was 4450 per second[11]
- Total data sharing CF access rate was 4234 per second
- Data sharing CF access rate was 3.0 per second per used MIP
- Data sharing overhead was calculated to be 4.1% of total used capacity

Table 9 (Page 1 of 2). Customer P Data Sharing Details

	SYS1	SYS2	SYS3	SYS4
ISC links	2	3	3	2
Processor	9672-RY5	9672-RY5	9672-RY5	9672-RY5
Busy %	100	95	85	100
MIPS used	447	425	380	447
Overhead MIPS	10.3	21.8	12.6	25.4
Overhead %	2.3	5.1	3.3	5.7
DB2 LOCK rate		173.3	120.4	13.0
Contention count		25902	31131	1301
GBP0 CACHE rate		25.9	19.2	2.3
GBP1 CACHE rate		504.3	216.2	29.3
DB2 LOCK rate		308.5	122.5	147.7
Contention count		20896	10880	13549
GBP0 CACHE rate		2.5	0.1	0.4
GBP1 CACHE rate		292.1	134.4	202.6
DB2 LOCK rate	32.3			93.4
Contention count	6432			2695
GBP0 CACHE rate	8.1			51.2
GBP1 CACHE rate	669.6			646.3
DB2 LOCK rate	27.6			51.6
Contention count	2593			1720
GBP0 CACHE rate	11.2			12.8
GBP1 CACHE rate	4.7			1.6
DB2 LOCK rate	409.3			81.5
Contention count	8325			14149
GBP0 CACHE rate	20.6			60.5

[11] The total CF access rate is ***all*** accesses to the CF, including those that are not related to data sharing. Examples of CF accesses that are not related to data sharing include JES2 checkpoint, RACF, System operlog, GRS STAR, and CICS logging.

Table 9 (Page 2 of 2). Customer P Data Sharing Details

	SYS1	SYS2	SYS3	SYS4
GBP1 CACHE rate	28.2			16.9
DB2 LOCK rate	3.4			2.1
Contention count	314			15
GBP0 CACHE rate	0.5			0.1
GBP1 CACHE rate	0.3			0.0

3.4.9 Customer Sample Q Details

- RMF report from 04/16/98 at 07:59 am for 15 minutes
- Data sharing TM/DB was IMS/IMS, DB2
- Other CF exploiters are not known
- CF LPAR 1 was on a 2-way C04 and was 3.0% busy
- CF LPAR 2 was on a 2-way C04 and was 4.5% busy
- Total data sharing CF access rate was 2240 per second
- Data sharing CF access rate was 5.37 per second per used MIP
- Data sharing overhead was calculated to be 3.9% of total used capacity

Table 10. Customer Q Data Sharing Details

	SYS1	SYS2
ISC links	Not known	Not known
Processor	9672-R45	9021-9X2
Busy %	56	87
MIPS used	116	422
Overhead MIPS	1.7	19.1
Overhead %	1.7	4.5
DB2 LOCK rate	31.6	167.8
Contention count	1011	14000
GBP0 CACHE rate	0.28	0.39
GBP1 CACHE rate	73.13	67.87
GBP2 CACHE rate	1.19	250.2
GBP3 CACHE rate	0.24	3.41
GBP4 CACHE rate	26.81	108
IMS LOCK rate	34.68	787.7
Contention count	66	69
VSAM CACHE rate	4.92	672
OSAM CACHE rate	0	8.52

Chapter 4. CICS/DBCTL Data Sharing Scalability Study

The purpose of this study is to evaluate the performance of Parallel Sysplex environment using 9672-R61 CMOS processors, and exploiting the new coupling facility technology.

The following scenarios were measured:

1. A single image 9672-R61 environment with no sharing of data

   The objective of this test case is to determine a baseline for comparison to all other coupling measurements.

2. Two 9672-R61s in a sysplex, with 100% of the workload sharing data

   The objective of this test case is to evaluate the initial cost to enter the data sharing environment.

3. Eight 9672-R61s in a sysplex, with 100% of the workload sharing data

   The objective of this test case is to evaluate the scalability of this solution.

4. Sixteen 9672-R61s in a sysplex, with 100% data sharing

   The objective of this test case is to evaluate the scalability of this solution up to 16 MVS images.

4.1 Environment Overview

Figure 26 on page 66 illustrates the configuration used for these tests:

I/O connectivity: Each 9672 had two paths to the shared IMS databases. The processors were connected to 9032 ESCON Directors that in turn were connected to 3990 Model 6 cached control units. All IMS databases were allocated on 3390 Model 3 triple capacity volumes.

CF connectivity: There was a single coupling facility link from each 9672 to each of the 9674 coupling facilities.

MVS sysplex: The master catalog was shared by all systems. JES2 was run in a MAS with the checkpoint residing on the coupling facility. XCF signalling was achieved via the coupling facility, with two structures defined. Each image in the sysplex was identical; cloning techniques were used to achieve this. MVS was run with WLM in COMPAT mode. See Chapter 2, "Tuning Recommendations" on page 21 for a detailed description of the dispatching priorities used in this environment.

Workload balancing: CICSplex SM (CP/SM) was used for workload balancing. The CP/SM configuration was such that each CMAS had the capability to communicate with any other CMAS in the CICSPlex. The "QUEUE" algorithm was employed for these measurements.

Transaction routing: CICS 4.1 MRO/XCF functions were used for dynamic transaction routing. All XCF communications were via the coupling facility; therefore, any inter-CEC MRO messages sent between CICS regions exploited XCF and the coupling facility. The CP/SM workload specification was defined so that all TORs in the configuration could route to any AOR in the sysplex.

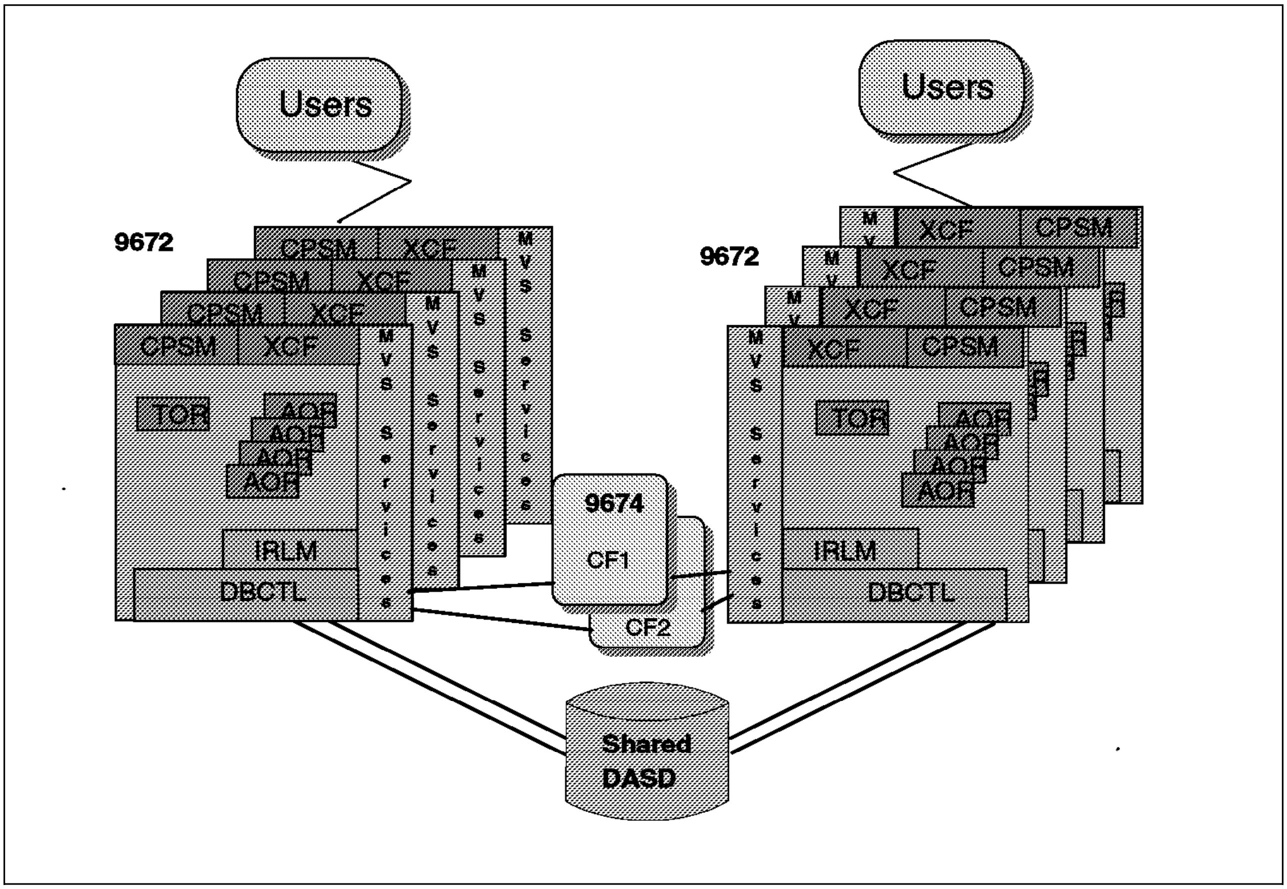


Figure 26. CICS/DBCTL Test Sysplex Configuration

IMS data sharing: The sysplex was configured with 100% connectivity to all IMS data. All IMS DBCTL regions were in a single data sharing group managed by IRLM. All IMS systems shared all data.

4.1.1 Hardware Resources

Table 11 lists the hardware resources used in the study.

Table 11. Hardware Resources — CICS/DBCTL Scalability Tests

	9672-R61s	**9674-C01**
# of processors/CEC	6	6
Central storage/CEC	512 MB	2048 MB[12]
Expanded storage/CEC	0 MB	0 MB
# of CF sender links/CEC to CF	1	N/A
# of CF receiver links on CF	N/A	16

The 9674s were at MEC level D57264.

[12] See Table 12 on page 67 and Table 13 on page 68 for actual storage usage.

4.1.2 Software Levels

The following software products were used at the specified level:

- CICS 4.1
- CICSplex SM 1.1.1 with APAR AN65633
- IMS 5.1 (PI level)
- IRLM 2.1 (PI level)
- JES2 5.1.0
- MVS 5.1 at Service Level 9406
- VTAM 4.2

4.1.3 Measurement and Reporting Tools

- RMF 5.1
- CICS Statistics (DFHSTUP)

4.1.4 Coupling Facility Exploitation

The following subsystem functions exploited the coupling facility:

- JES2 checkpoint
- IRLM lock table
- XCF signalling structures
- OSAM and VSAM buffer invalidation structures

One 9674-C01 coupling facility was used for the 2x9672-R61 and 8x9672-R61 measurements.

Table 12. Structure Size/Placement for 2x9672-R61 and 8x9672-R61 Configuration

Structure Description	Structure Name	CF	Structure Type	Structure Size
XCF signalling	IXCSIG1	CF1	List	69 MB
XCF signalling	IXCSIG2	CF1	List	69 MB
IRLM lock table	IRLMLOCKTBL2	CF1	Lock	64 MB
JES2 checkpoint	JES2CKPT	CF1	List	13 MB
OSAM buffer invalidation	OSAMSESXI	CF1	Cache	64 MB
VSAM buffer invalidation	VSAMSESXI	CF1	Cache	64 MB
Total CF storage used:[13]				343 MB

Two 9674-C01s were used for the 16x9672-R61 measurements. Structures were divided between the two 9674 CFs to achieve both reliability and equal CF utilization.

Table 13. Structure Size/Placement for 16x9672-R61 Configuration

Structure Description	Structure Name	CF	Structure Type	Structure Size
IRLM lock table	IRLMLOCKTBL2	CF1	Lock	64 MB
XCF signalling	IXCSIG2	CF1	List	69 MB
XCF signalling	IXCSIG1	CF2	List	69 MB
JES2 checkpoint	JES2CKPT	CF2	List	13 MB
OSAM buffer invalidation	OSAMSESXI	CF2	Cache	64 MB
VSAM buffer invalidation	VSAMSESXI	CF2	Cache	64 MB
Total CF storage used:[13]				343 MB

4.1.5 Workload Description

The CICS/DBCTL workload was used for these tests. The workload consisted of a CICS front end that acts as the transaction manager. The database manager was an IMS DBCTL subsystem. The workload consisted of a mix of light to moderate transactions from CICS applications covering diverse business functions, including order entry, stock control, inventory tracking, production specifications, hotel reservations, banking, and teller system. There were 17 unique transactions that accessed a total of 35 unique databases.

The workload has been modified so that there are no transaction affinities. This removed dependencies between transactions, and allowed any transaction to be routed to, and run on any other system.

The workload accessed both VSAM and OSAM databases, with VSAM indexes (primary and secondary). DLI HDAM and HIDAM access methods were used. The workload had a moderate I/O load.

A summary of the CICS/DBCTL workload characteristics is as follows:

Table 14. CICS/DBCTL Workload Characteristics

	9672-R61	2x9672-R61
CPU milliseconds per tran	52.12	63.35
VSAM+OSAM DB I/O per tran	3.78	3.94
IRLM lock table req/tran to CF	N/A	6.28
OSAM cache req/tran to CF	N/A	0.93
VSAM cache req/tran to CF	N/A	3.01
Lock contention	N/A	0.05%

[13] Although each 9674 coupling facility was configured with 2048 MB of central storage, only 343 MB of the storage was actually utilized.

4.1.6 Measurement Description

The measurements taken for this study are discussed in detail below.

4.1.6.1 Measurement Methodology

To ensure that the subsystems processed smoothly with no unnecessary points of contention, database tuning was done.

In order to keep the contention for databases and I/O to database volumes constant, the number of copies of the workload was adjusted as the configuration grew. This kept the access rate per copy of the workload constant throughout all configurations.

The local buffer pools were tuned to the largest configuration (16x9672-R61) and remained constant across all measurements.

All structure sizes were tuned to the largest configuration (16x9672-R61).

There was a simulated remote network attached to each CEC. For all measurements in this performance study, a constant, predefined number of users logged on to each TOR on each CEC. The users then executed the transaction scripts. Routing the transaction to an AOR was done by the CP/SM “QUEUE” algorithm. All AORs in the configuration were eligible to run the transactions.

The workload was run in steady-state at a predefined processor utilization. Given a constant number of users on each CEC, the processor utilization was controlled by the user think time. A fifteen minute RMF interval was captured for each measurement. The average processor utilization of 80% was achieved on each CEC.

4.1.6.2 Initial Cost of Data Sharing

Objective: The first test case focused on understanding the costs of data sharing as we take a single system CICS/DBCTL workload, and introduce it into a data sharing environment on the 2x9672-R61.

Methodology: A 9672-R61 measurement was taken as a baseline for comparison for all other measurements. The measurement was run in a noncoupled environment. The system was run with a sysplex configuration of XCFLOCAL and GRS of NONE. There was no data sharing, and the IRLM subsystem was not started. IMS Program Isolation (PI) was used for data integrity on the single system. TOR to AOR communication was via cross memory services. The CP/SM “QUEUE” algorithm was used for intra-CEC workload balancing.

For the 2x9672-R61 runs, MVS was run in a sysplex. A single 9674-C01 coupling facility was used. All coupling facility exploitations discussed in 4.1.4, “Coupling Facility Exploitation” on page 67 were used. The CICS/DBCTL workload was cloned to the second system, and the IMS subsystems on both systems were in the data sharing group sharing all data.

4.1.6.3 Scalablity When Adding CECs to the Sysplex

Objective: The majority of the cost when entering a data sharing environment is taken when going from a single system (with no sharing) to a configuration with two systems sharing data. It is then expected that the additional cost per CEC as the sysplex increases will scale appropriately. In order to show scalability as the configuration grows, configurations were sized at increments between two and sixteen CECs within the sysplex.

Methodology: A 8x9672-R61 sysplex was measured. A single 9674-C01 coupling facility was used. All coupling facility exploitations discussed in 4.1.4, "Coupling Facility Exploitation" on page 67 were used. All IMS subsystems were in the data sharing group sharing all IMS databases.

A 16x9672-R61 sysplex was measured. Two 9674-C01 coupling facilities were used. All coupling facility exploitations discussed in 4.1.4, "Coupling Facility Exploitation" on page 67 were used. All IMS subsystems were in the data sharing group sharing all IMS databases.

4.2 Results

Following are the detailed results of all test cases discussed in this performance study. Details regarding the analysis of overheads discussed can be found in Appendix A, "CICS/DBCTL Workload Details and Coupling Efficiency Details" on page 171.

4.2.1 Data

Table 15. 9672-R61 Performance Study

	9672-R61	**2x9672-R61**	**8x9672-R61**	**16x9672-R61**
Transactions/sec (ETR)	92.5	151.2	587.74	1106.6
CPU activity	80.4%	79.8%	79.8%	78.9%
Transactions/sec at 100% (ITR)	115.11	189.4	736.9	1402.4
CPU milliseconds per tran	52.1	63.4	65.1	68.7
Internal CICS response time	0.195	0.206	0.187	0.199
CF utilization	N/A	CF1 4.4%	CF1 19.5%	CF1 27.4% CF2 25.0%
Coupling Facility Rates (Requests/Second)				
IXCSIG1	N/A	82	220	1116[14]
IXCSIG2	N/A	57	215	1154[14]
IRLMLOCKTBL	N/A	949	3903	7349
OSAMSESXI	N/A	143	566	1066
VSAMSESXI	N/A	453	2012	3723

[14] Much of the increase in signalling traffic between the 8x9672-R61 and the 16x9672-R61 runs was due to internal-to-XCF management messages unrelated to subsystems (such as GRS or CICS) exchanging messages. We believe that further tuning of XCF transport classes would have eliminated this unnecessary traffic.

4.2.2 9672 Results with 100% Data Sharing

Figure 27 on page 72 contains a graphic view of the results from this experiment.

4.2.3 Observations/Conclusions

Initial Cost of Data Sharing: The initial cost to go from a single system, nondata sharing environment to a multisystem, data sharing environment can be seen in the comparison between the 9672-R61 and 2x9672-R61 measurements.

Comparison of internal throughput shows a cost of 17.7% when entering this multisystem shared data environment.

When entering the data sharing environment, IRLM is introduced. Management of data and locks across the sysplex is required. For a detailed breakdown of the overheads when entering the multisystem data sharing environment, see Appendix A, “CICS/DBCTL Workload Details and Coupling Efficiency Details” on page 171.

The internal CICS transaction response times increased slightly between the 9672-R61 and 2x9672-R61 measurements. However, this increase is negligible.

Scalability When Adding CECs to the Sysplex: The cost of going from a single system nondata sharing environment (9672-R61) to a 8x9672-R61 is 20.0%.

The cost of going from a single system nondata sharing environment (9672-R61) to a 16x9672-R61 is 23.8%.

In looking at the additional overhead as we add systems to the 2x9672-R61, we can calculate the coupling overhead. The coupling overhead can be calculated as the difference between the cost of going from one to two systems, and the cost of going from one to eight or one to sixteen systems. This number is then normalized by the number of additional systems, 6 or 14 respectively. For example:

Coupling overhead from 9672-R61 to 2x9672-R61 = 17.7%

Coupling overhead from 9672-R61 to 16x9672-R61 = 23.8%

Total additional overhead when adding 14 more systems to the sysplex is (23.8 – 17.7) = 6.1%. Therefore, the cost per additional system is $\frac{6.1}{14}$ = 0.44% additional overhead per system added.

Cost Per Additional CEC

This shows that as the number of systems in the sysplex grows, the cost of adding CECs is scalable at roughly 0.4% per CEC.

The internal CICS transaction response times remained level when going from the 2x9672-R61 to the 16x9672-R61.

For a detailed breakdown of the overheads when entering the multisystem shared data environment, see Appendix A, “CICS/DBCTL Workload Details and Coupling Efficiency Details” on page 171.

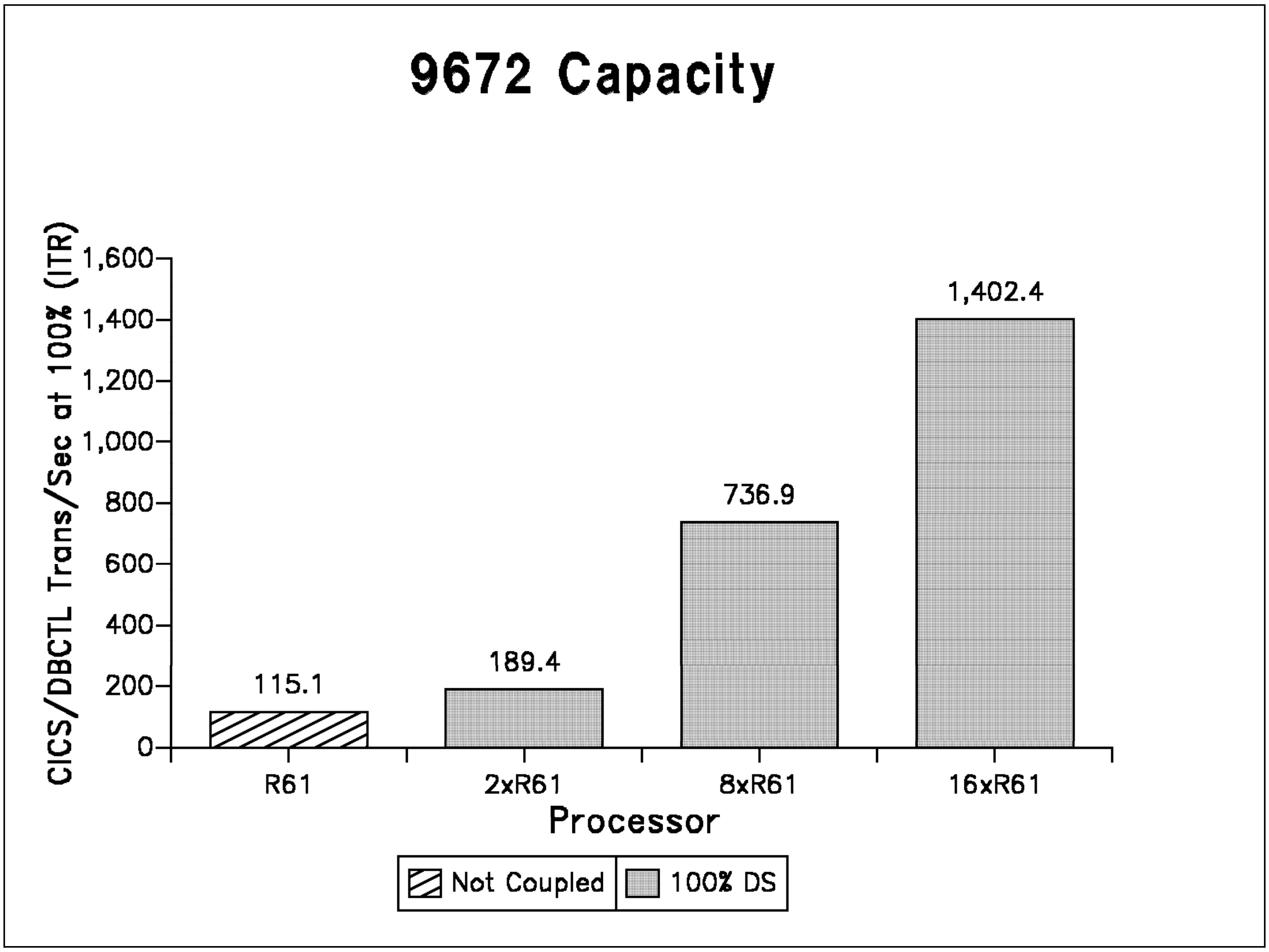


Figure 27. CICS/DBCTL Scalability Test ITR Comparison

As discussed in 1.4.3, "Additional Workload Considerations" on page 12, the cost of data sharing is very workload-dependent. Our benchmark results reflect a stressed data sharing environment. Customer experiences have typically seen the cost to move to a sysplex data sharing environment to be about half that seen in our benchmarks.

Chapter 5. IMS-TM/DB2 Data Sharing Scalability Study

The purpose of this study is to evaluate the performance scalability of DB2 data sharing using 6-way 9672-R61 CMOS processors and exploiting the coupling facility technology.

The following scenarios were measured:

1. Base 9672-R61 environment with no sharing of data

   The objective of this test case is to determine a baseline for comparison to all other coupling measurements.

2. Two 9672-R61 in a sysplex, with 100% of the workload sharing data

   The objective of this test case is to evaluate the initial cost to enter the data sharing environment.

3. Eight 9672-R61 in a sysplex, with 100% of the workload sharing data

   The objective of this test case is to evaluate the scalability of this solution.

5.1 Environment Overview

Figure 28 on page 74 illustrates the configuration used for these tests:

I/O connectivity: Each 9672 had four paths to the shared DB2 databases. The processors were connected to 9032 ESCON Directors that in turn were connected to 3990 Model 3 cached control units. All DB2 databases were allocated on 3390 Model 3 triple capacity volumes.

CF connectivity: There were dual coupling facility links from each 9672 to each of the 9674 coupling facilities.

MVS sysplex: The master catalog was shared by all systems. JES2 was run in a MAS with the checkpoint residing on the coupling facility. XCF signalling was achieved via the coupling facility, with two structures defined and through CTCs. Each image in the sysplex was identical; cloning techniques were used to achieve this. MVS was run with WLM in COMPAT mode. See Chapter 2, “Tuning Recommendations” on page 21 for a detailed description of the dispatching priorities used in this environment.

The sysplex was configured with 100% connectivity to all DB2 data. All DB2s systems shared all data.

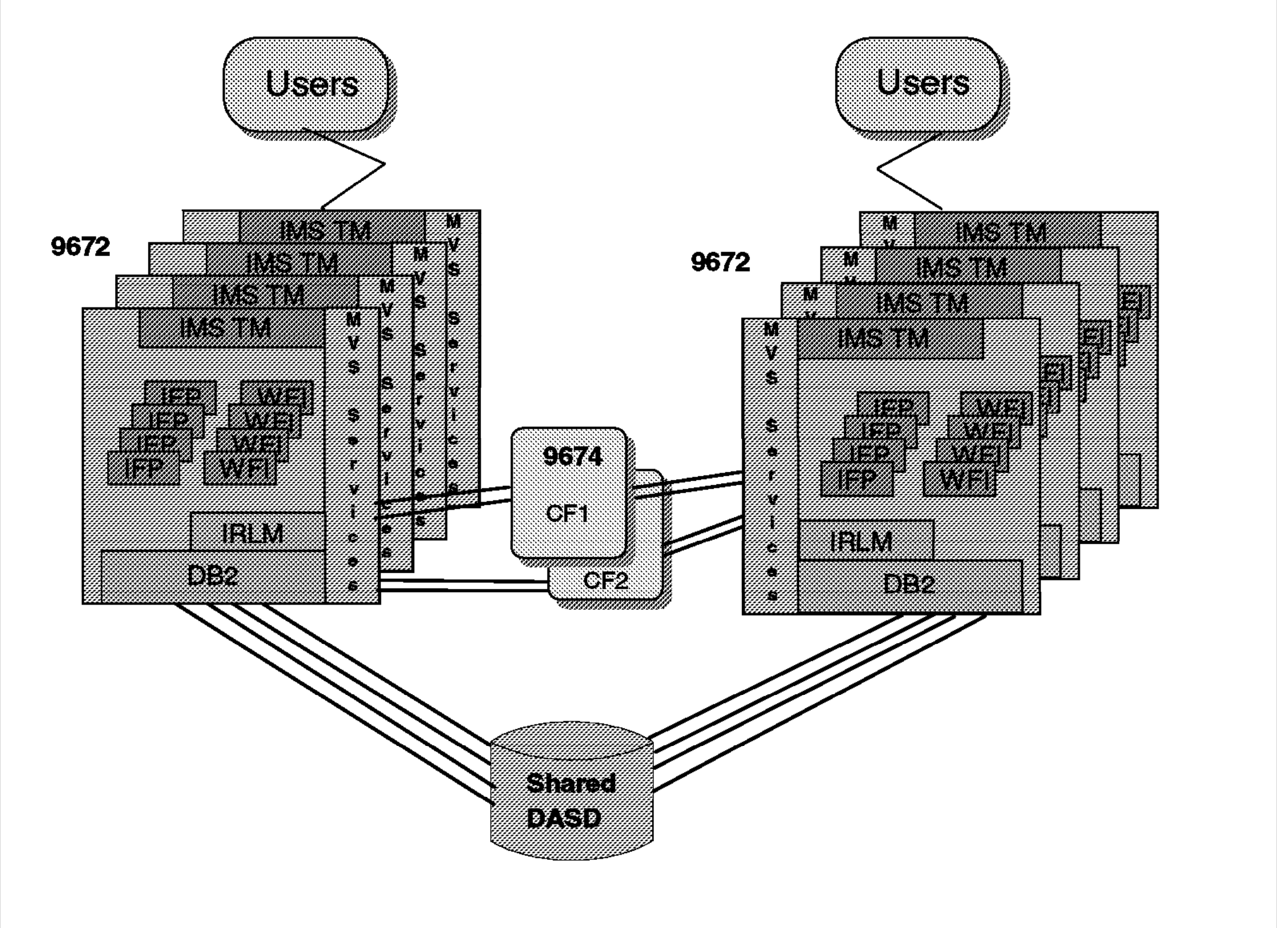


Figure 28. IMS-TM/DB2 Test Sysplex Configuration

5.1.1 Hardware Resources

Table 16 lists the hardware resources used in the study.

Table 16. Hardware Resources — IMS-TM/DB2 Scalability Tests

	9672-R61	9674
# of processors/CEC	6	6
Central storage/CEC	2 GB	2 GB[15]
Expanded storage/CEC	0M	0M
# of CF sender links per CEC to each CF	2	N/A
# of CF receiver links on each CF	N/A	16

The 9674 E/C level was D79533 with MCLs 058, 059, 060 for DR44 installed.

[15] See Table 17 and Table 13 on page 68 for actual storage usage.

5.1.2 Software Levels

The following software products were used at the specified level:

- DB2 V4.1
- IMS 5.1
- IRLM 2.1
- JES2 5.2.0
- MVS 5.2.0
- VTAM 4.2

5.1.3 Measurement and Reporting Tools

The tools used for measuring the system were:

- RMF 5.2.0
- DB2 PM V4

5.1.4 Coupling Facility Exploitation

The following subsystem functions exploited the coupling facility:

- DB2 group buffer pools
- DB2 SCA
- DB2 lock
- JES2 checkpoint

Two 9674 coupling facilities were used for the 9672-R61 and 8x9672-R61 measurements.

Table 17. Structure Size/Placement for the Configuration

Structure Description	Structure Name	CF	Structure Type	Structure Size
DB2 group buffer pools	DSNDB0G_GBP0 — GBP8	CF1	Cache	1121 MB
JES2 checkpoint	COUPLE_CKPT1	CF2	List	13 MB
DB2 SCA	DSNDB0G_SCA	CF2	List	49 MB
DB2 lock	DSNDB0G_LOCK1	CF2	Lock	586 MB

5.1.5 Workload Description

The following describes the workload and its environment for our measurements.

The IMS-TM/DB2 workload consisted of:

- An IMS front end that was the transaction manager.
- The database manager was DB2 V4.
- Degree of data sharing was 100%.
- Random scheduling of transactions and keys across the members of the DB2 data sharing group.

The workload was made up of several different transactions varying in characteristics from very read intensive to update intensive as well as simple to complex SQL. The workload has multithread capability as well as a broad spectrum of SQL functionality. The workload has many application and database design features which are typical of customer applications. There were eight unique transactions that accessed a total of nine unique tables.

The workload has been modified so that there are no transaction affinities. This removed dependencies between transactions, and allowed any transaction to be routed to and run on any other system.

A summary of the IMS-TM/DB2 workload characteristics appears in appendix B.1, “IMS-TM/DB2 Workload Characteristics for a Single MVS Image” on page 177.

5.1.6 Measurement Description

The measurements taken for this study are discussed in detail in this section.

5.1.6.1 Measurement Methodology

To ensure that the subsystems processed smoothly with no unnecessary points of contention, database tuning was done.

In order to keep the contention for data pages and I/O to database volumes constant, the number of copies of the workload was adjusted as the configuration grew. This kept the access rate per copy of the workload constant throughout all configurations.

The local buffer pools were tuned to the largest configuration (8x9672-R61) and scaled across all measurements.

All structure sizes were tuned to the largest configuration (8x9672-R61) and scaled across all measurements.

There was a simulated remote network attached to each CEC. For all measurements in this performance study, a constant, predefined number of users logged on to each IMS/DB2. The users then executed the transaction scripts. Transaction routing was done via the TPNS scripts. All DB2 subsystems in the configuration had access to the data for each transaction.

The workload was run in steady-state at a predefined processor utilization. Given a constant number of users on each CEC, the processor utilization was controlled by the user think time. A twenty minute RMF interval was captured for each measurement. An average processor utilization of at least 70% was achieved on each CEC.

5.1.6.2 Initial Cost of Data Sharing

Objective: The first test case focused on understanding the costs of data sharing as we take a single system IMS-TM/DB2 workload and introduce it into a data sharing environment on the 9672-R61.

Methodology: A 9672-R61 measurement was taken as a baseline for comparison for all other measurements. The measurement was run in a noncoupled environment. The system was run with a sysplex configuration of XCFLOCAL and GRS of NONE. There was no DB2 data sharing.

For the 2x9672-R61 runs, MVS was run in a sysplex. Two 9674 coupling facilities were used. All coupling facility exploitations discussed in 5.1.4, “Coupling Facility Exploitation” on page 75 were used. The IMS-TM/DB2 workload was cloned to the second system, and the DB2 subsystems on both systems were in the data sharing group sharing all data.

5.1.6.3 Scalability When Adding CECs to the Sysplex

Objective: The majority of the cost when entering a data sharing environment is taken when going from a single system (with no sharing) to a configuration with two systems sharing data. It is then expected that the additional cost per CEC will scale appropriately as the sysplex increases. In order to show scalability as the configuration grows, configurations were sized at increments between two and eight CECs within the sysplex.

Methodology: An 8x9672-R61 sysplex was measured. Two 9674 coupling facilities were used. All coupling facility exploitations discussed in 5.1.4, “Coupling Facility Exploitation” on page 75 were used. All DB2 subsystems were in the data sharing group sharing all DB2 tables.

5.2 Results

Following are the detailed results of all test cases discussed in this performance study. Details regarding the analysis of overheads discussed can be found in B.1, “IMS-TM/DB2 Workload Characteristics for a Single MVS Image” on page 177.

5.2.1 Data

Table 18. IMS-TM/DB2 Data Sharing Performance Study

	9672-R61	**2x9672-R61**	**8x9672-R61**
Transactions/sec (ETR)	35.1	54.5	219.6
CPU activity	75.9%	71.3%	74.3%
Transactions/sec at 100% (ITR)	46.2	76.4	295.6
CPU milliseconds per tran	130	157	162
CLASS 2 ELAPSED time in sec	1.0	0.8	0.7
CF1 utilization	N/A	6.6%	31.5%
CF2 utilization	N/A	2.9%	8.2%
Coupling Facility Rates (Requests/Second)			
COUPLE_CKPT1	N/A	3.0	12.9
DSNDB0G_SCA	N/A	0.4	1.7
DSNDB0G_LOCK1	N/A	824	3490
DSNDB0G_GBP0 — GBP8	N/A	1081	4945

5.2.2 IMS-TM/DB2 Results with 100% Data Sharing

Figure 29 on page 79 contains a graphic view of the results from this experiment.

5.2.3 Observations/Conclusions

Initial Cost of Data Sharing: The initial cost to go from a single system nondata sharing environment to a multisystem data sharing environment can be seen in the comparison between the 9672-R61 and 2x9672-R61 measurements.

Comparison of internal throughput shows a cost of 17.3% when entering this multisystem shared data environment.

When entering the data sharing environment, management of data and locks across the sysplex is required. For a detailed breakdown of the overheads when entering the multisystem data sharing environment see Appendix B, "IMS-TM/DB2 Workload Details and Coupling Efficiency Details" on page 177.

The internal IMS/DB2 transaction response times (CLASS 2 ELAPSED times) decreased slightly between the 9672-R61 and 2x9672-R61 measurements.

Scalability When Adding CECs to the Sysplex: The cost of going from a single system nondata sharing environment (9672-R61) to an 8x9672-R61 is 20.0%.

In looking at the additional overhead as we add systems to the 9672-R61, we can calculate the coupling overhead. The coupling overhead can be calculated as the difference between the cost of going from one to two systems and the cost of going from one to eight. This number is then normalized by the number of additional systems, that is, six systems. For example:

Coupling overhead from 9672-R61 to 2x9672-R61 = 17.3%

Coupling overhead from 9672-R61 to 8x9672-R61 = 20.0%

Total additional overhead when adding six more systems to the sysplex is (20.0 − 17.3) = 2.7%. Therefore, the cost per additional system is $\frac{2.7}{6} = 0.45\%$ additional overhead per system added.

Cost Per Additional CEC

This shows that as the number of systems in the sysplex grows, the cost of adding CECs is scalable at roughly 0.5% per CEC.

The internal IMS/DB2 transaction response times (CLASS 2 ELAPSED times) decreased slightly between the 2x9672-R61 and 8x9672-R61 measurements.

For a detailed breakdown of the overheads when entering the multisystem shared data environment, see B.1, "IMS-TM/DB2 Workload Characteristics for a Single MVS Image" on page 177.

As discussed in 1.4.3, "Additional Workload Considerations" on page 12, the cost of data sharing is very workload dependent. Our benchmark results reflect a stressed data sharing environment. Customer experiences have typically seen the cost to move to a sysplex data sharing environment to be about half that seen in our benchmarks.

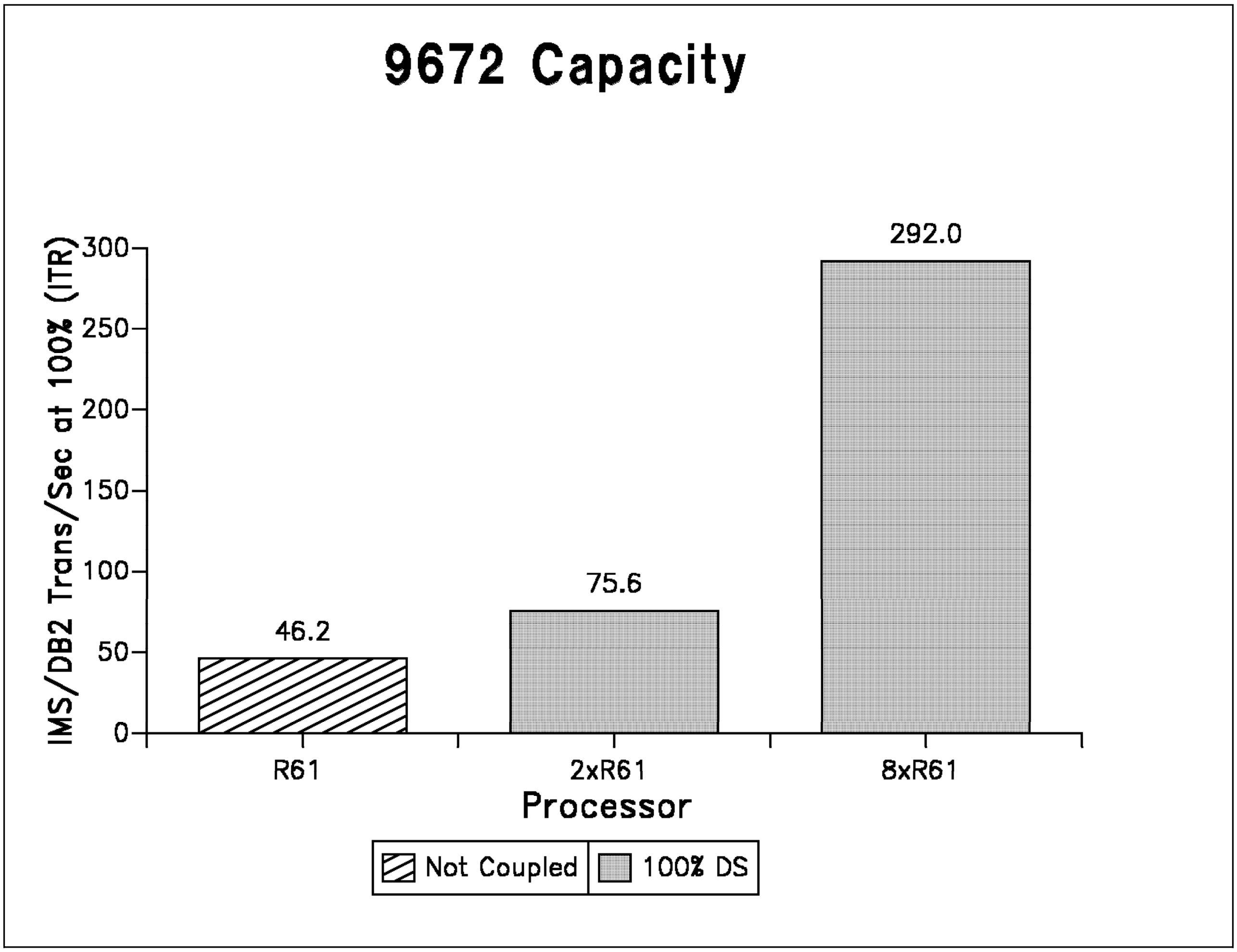


Figure 29. IMS-TM/DB2 Scalability Test ITR Comparison

Chapter 6. VSAM RLS Data Sharing Scalability Study

The purpose of this study is to evaluate the performance scalability of the CICS/VSAM RLS Solution using 9672-R63, 9672-R61, and 9672-R64 CMOS processors with the exploitation of the coupling facility technology.

The following scenarios were measured:

1. Base environment with CICS/ESA 4.1 MRO Function Shipping on a single dedicated 3-way logical partition on a 9672-R63

   The objective of this test case was to determine a baseline for comparison to the 9672-R61 and 9672-R63 coupling measurements.

   The objective of this test case was to determine a baseline for comparison to the 9672-R64 coupling measurement.

2. Two dedicated 3-way logical partitions on one 9672-R63 in a Parallel Sysplex, with 100% of the workload using RLS for the files

   The objective of this test case was to evaluate the initial cost to enter the RLS environment on a 9672-R63.

3. Four dedicated 3-way logical partitions on two 9672-R63 plus four 9672-R61s in a Parallel Sysplex, with 100% of the workload using RLS for the files

   The objective of this test case was to evaluate the scalability of the RLS solution up to 8 MVS images.

6.1 Environment Overview

Figure 30 on page 82 and Figure 31 on page 83 illustrate the configurations used for these tests:

I/O connectivity: Each 9672 had four paths to the shared CICS/VSAM RLS databases. The processors were connected to 9032 ESCON Directors that in turn were connected to 3990 Model 6 cached control units. All CICS/VSAM RLS databases were allocated on 3390 Model 3 triple capacity volumes.

CF connectivity: There were dual coupling facility links from each 9672 to the 9674 coupling facility. Multimode fiber was used for the links; these support a maximum distance of 1 km at a speed of 50 MB/sec.

MVS sysplex: The master catalog was shared by all systems. JES2 was run in a MAS with the checkpoint residing on the coupling facility. XCF signalling was achieved through CTCs. Each image in the sysplex was identical; cloning techniques were used to achieve this. MVS was run in WLM compatibility mode. See Chapter 2, “Tuning Recommendations” on page 21 for a detailed description of the dispatching priorities used in this environment.

The sysplex was configured with 100% connectivity to all CICS/VSAM RLS data. All of the CICS AORs in the sysplex shared all of the data.

Workload balancing: CICSplex SM (CP/SM) was used for workload balancing. The CP/SM configuration was such that each CMAS had the capability to communicate with any other CMAS in the CICSPlex. The QUEUE algorithm was used for these measurements.

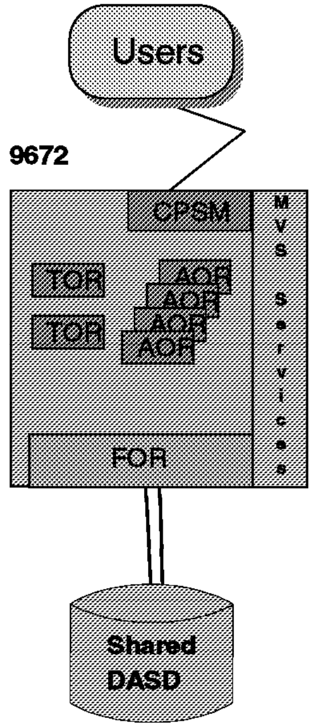


Figure 30. CICS/VSAM RLS Base Case Configuration

Transaction routing: CICS 4.1 and CICS TS 1.1 MRO/XCF functions were used for dynamic transaction routing. All CICS regions were part of the XCF group DFHIR000. Any inter-CEC MRO messages sent between CICS regions exploited XCF. The CP/SM workload specification was defined so that all TORs in the configuration could route to any AOR in the sysplex. More than 90% of the transactions were routed to the local AORs.

CICS/VSAM RLS: The sysplex was configured with 100% connectivity to all VSAM data. All CICS regions were in a single data sharing group. All CICS AORs shared the VSAM data.

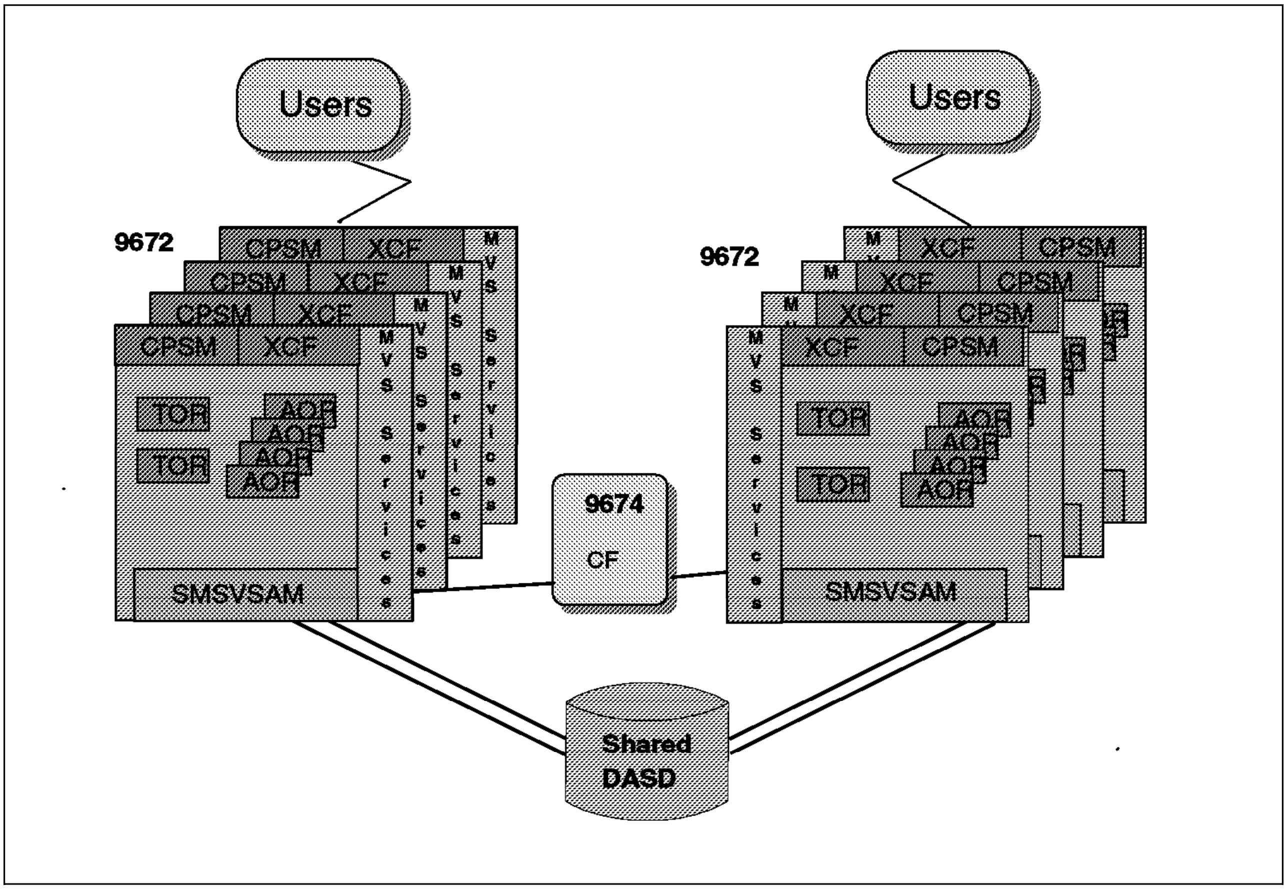


Figure 31. CICS/VSAM RLS Test Sysplex Configuration

6.1.1 Hardware Resources

Table 19 lists the hardware resources used in this study.

Table 19. Hardware Resources

	9672-R61	9672-R63[16]	9674-C03
# of CECs	4	2	1
# of processors/CEC	6	6	5
Central storage/CEC	1 GB	1 GB	2 GB
Expanded storage/CEC	0	0	0
# of CF sender links per CEC to each CF	2	2	N/A
# of CF receiver links on each CF	N/A	N/A	16

The coupling facilities used in these experiments were dedicated 9674s. Each CF was defined as a single logical partition which had all hardware resources dedicated to it.

[16] Two dedicated 3-way logical partitions were defined on each of the 9672-R63s for a total of four images.

6.1.2 Software Levels

The following software products were used at the specified level:

- CICS TS 1.1 + APAR PN87305
- CICS/ESA V4.1 for comparison
- CICSplex SM 1.2.0
- JES2 5.2.0
- MVS 5.2.2 + APARs OW20060 and OW15588
- DFSMS 1.3 + APAR OW19918
- VTAM 4.3

6.1.3 Measurement and Reporting Tools

The following tools were used to measure the system.

- RMF 5.2.0
- CICS Statistics (DFHSTUP)

6.1.4 Coupling Facility Exploitation

The following subsystem functions exploited the coupling facility:

- DFSMS lock table
- DFSMS cache
- LOGGER
- JES2 checkpoint
- VTAM generic resource

For the 2x9672 R63 and 8x9672 R63 measurements, a single 9674-C03 coupling facility was used. Table 20 lists the structures used in this configuration.

Table 20. Structure Size/Placement for the Configuration

Structure Description	Structure Name	CF	Structure Type	Structure Size
DFSMS lock	IGWLOCK00	CF1	Lock	32 MB — 128 MB[17]
DFSMS cache	SMSCACHE_STR1 — SMSCACHE_STR2	CF1	Cache	100 MB — 800 MB[18]
LOGGER	DFHLOG_STR1 — DFHLOG_STR4[19]	CF1	List	64 MB
LOGGER	DFHSHUNT_STR	CF1	List	8 MB
JES2 checkpoint	COUPLE_CKPT1	CF1	List	13 MB
VTAM generic resource	ISTGENERIC	CF1	List	10 MB

[17] The size of the DFSMS lock structure was scaled across the measurements. A 32 MB lock structure was used for the 2x9672 measurements and a 128 MB lock structure was used for the 8x9672 measurements.

[18] The sizes of the DFSMS Cache structures were scaled across the measurements. For the 2x9672 measurements, two 100 MB structures were defined for a total size of 200 MB. For the 8x9672 measurements, two 800 MB structures were defined for a total of 16000 MB.

6.1.5 Workload Description

The CICS/VSAM RLS workload was used for these tests. The workload consisted of a mix of light-to-moderate transactions from CICS applications which simulated order entry, stock control, inventory tracking, production specifications, hotel reservations, banking, and teller system functions. The applications were written in COBOL II. There was an average of 6 VSAM calls per transaction. The workload was a mix of 28% Read, 48% Browse, 9% Readupdate, 9% Update, 5% Add, and 1% Delete.

There were twenty-four unique transactions which accessed 20 unique files. The files consisted of 18 KSDSs, 1 RRDS, and 1 ESDS. The files were defined with a CI size of 4096. The KSDSs had an index size of 2048.

The workload was modified so that there were no transaction affinities. This removed dependencies between transactions and allowed any transaction to be routed to any system. Thus, the transactions were able to be executed on all systems.

One LOGGER structure was defined for every two images in the Parallel Sysplex. Each structure contained twelve log streams (that is, one for each CICS region). Staging datasets were not used in our environment.

A summary of the CICS/VSAM RLS workload characteristics appears in appendix C.1, “CICS/VSAM Workload Characteristics for a Single MVS Image” on page 183.

6.1.6 Measurement Description

The measurements taken for this study are discussed in detail below.

6.1.6.1 Measurement Methodology

To ensure that the subsystems processed smoothly with no unnecessary points of contention, database tuning was done.

In order to keep the contention for databases, and the I/O to database volumes constant, the number of copies of the database was adjusted as the configuration grew. This kept the access rate per copy of the workload constant throughout all configurations.

The size of the local buffer pool and the DFSMS Cache structures were tuned to the largest configuration (8x9672) and scaled across all measurements. The local buffer pool size was specified via the RLS_MAX_POOL_SIZE parameter in the IGDSMSxx member in SYS1.PARMLIB. The size of the lock structure, IGWLOCK00, was scaled across all measurements in order to keep the false contention rate constant throughout all configurations.

TPNS was run outboard to simulate a remote network. For all measurements in this performance study, a predefined number of users logged on to the TORs on each CEC. The users then executed the transaction scripts. The CP/SM QUEUE algorithm was used to route transactions to the AORs. All AORs in the sysplex were eligible to run all of the transactions. More than 90% of the transactions

[19] Two 64 MB Logger structures were used for the 2x9672 measurements for a total of 128 MB. Four 64 MB Logger structures were used for the 8x9672 measurements for a total of 256 MB. One structure was defined for every two systems used. Each structure contained 12 log streams (that is, 1 log stream per CICS region).

were processed on the local system. All CICS AORs in the configuration had access to the data for each transaction.

The workload was run in a steady-state with a fixed number of users on each CEC and a fixed user think time. This was done to ensure a constant ETR for all measurements. A fifteen minute RMF interval was captured for each measurement.

Table 21 shows the local buffer pool sizes which were used for the various configurations. This size was specified on the RLS_MAX_POOL_SIZE parameter in the IGDSMSxx member in SYS1.PARMLIB.

Table 21. Local Buffer Pool Sizes for the Configuration

Configuration	RLS_MAX_POOL_SIZE
2x9672	100 MB
8x9672	400 MB

6.1.6.2 Initial Cost of Data Sharing

Two test cases were executed to show the initial cost of RLS in the 9672 environment.

Objective: The test case focused on understanding the costs of data sharing as we took a single system CICS/VSAM RLS workload and introduced it into the RLS environment on the 9672-R63.

Methodology: A measurement of a single dedicated 3-way logical partition on a 9672-R63 was taken as a baseline for comparison for all other measurements. The workload was run using CICS 4.1 MRO Function Shipping to access the files from a single FOR. All AORs shared all of the data owned by the FOR. There were no files defined local to the AORs. All file accesses were achieved via Function Shipping. Journaling was active. The files were defined in the CSD with backout only recovery (that is, RECOVERY(BACKOUTONLY)). Local Shared Resources (LSR) were used.

For the 2x9672 measurement, MVS was run in a sysplex with CICS TS 1.1 with 100% RLS. The FOR which was present in the single system configuration was removed. The files were defined with No Read Integrity (that is, READINTEG(UNCOMMITTED)). Backout only recovery was enabled. LOG(UNDO) was specified on the IDCAMS DEFINE CLUSTER statement.

One 9674-C03 coupling facility was used. All coupling facility exploitations discussed in 6.1.4, “Coupling Facility Exploitation” on page 84 were used. The CICS/VSAM RLS workload was cloned to the second system, and the CICS subsystems on both systems were in the data sharing group which shared all data.

6.1.6.3 Scalability When Adding CECs to the Sysplex

Objective: The majority of the cost when entering a data sharing environment is taken when going from a single system to a configuration with two systems sharing data. It is then expected that the additional cost per CEC as the sysplex increases will scale appropriately. In order to show scalability as the configuration grows, configurations were sized at increments between two and eight CECs within the sysplex.

Methodology: An 8x9672 sysplex was measured. A single 9674-C03 Coupling facility was used. All coupling facility exploitations discussed in 6.1.4, "Coupling Facility Exploitation" on page 84 were used. All CICS subsystems were in the same data sharing group and shared all of the VSAM data.

6.2 Results

Table 22 contains the detailed results of the test case discussed in this performance study. Details regarding the analysis of overheads discussed can be found in C.1, "CICS/VSAM Workload Characteristics for a Single MVS Image" on page 183.

6.2.1 Data

Table 22. CICS/VSAM RLS Data Sharing Performance Study

	9672 MRO	**2x9672**	**8x9672**
Transactions/sec (ETR) per CEC	95.55	96.46	97.2
CPU activity	73.3%	89.4%	93%
Transactions/sec at 100% (ITR) per CEC	130.37	107.91	104.53
CPU milliseconds per tran	23.01	27.8	28.7
TOR response time (sec)	.083	.212[20]	.441[20]
CF1 utilization	N/A	2.3%	10.3%
Coupling Facility Rates (Requests/Second) per CEC			
COUPLE_CKPT1	N/A	1.6	1.6
IGWLOCK00	N/A	144	138.4
SMSCACHE_STR1 - SMSCACHE_STR2	N/A	393.8	434.8
DFHLOG_STR1 - DFHLOG_STR4	N/A	98	96.4
ISTGENERIC	N/A	0	0

6.2.2 CICS/VSAM RLS Results with 100% Data Sharing

Figure 32 on page 88 contains a graphic view of the results from this experiment.

6.2.3 Observations/Conclusions

Initial Cost of Data Sharing: The initial cost to move from a single system MRO Function Shipping environment to a multisystem RLS environment can be seen in the comparison between the 9672 and 2x9672 measurements.

Comparison of internal throughput shows a cost of 17.2% when entering the multisystem RLS environment. This is for a 9672-R63; the cost would be 19% if it were a 9672-R64. However, both costs would be reduced by approximately 3%

[20] The increase in CICS TOR response time warranted further investigation. Subsequent measurements showed that the increase was due to increasing CPU queuing effects at higher CPU utilizations, most notably at CPU utilizations exceeding 90%. A 2x9672 measurement taken at a lower CPU utilization similar to the MRO base of 73.3% showed the TOR response to be approximately 0.08.

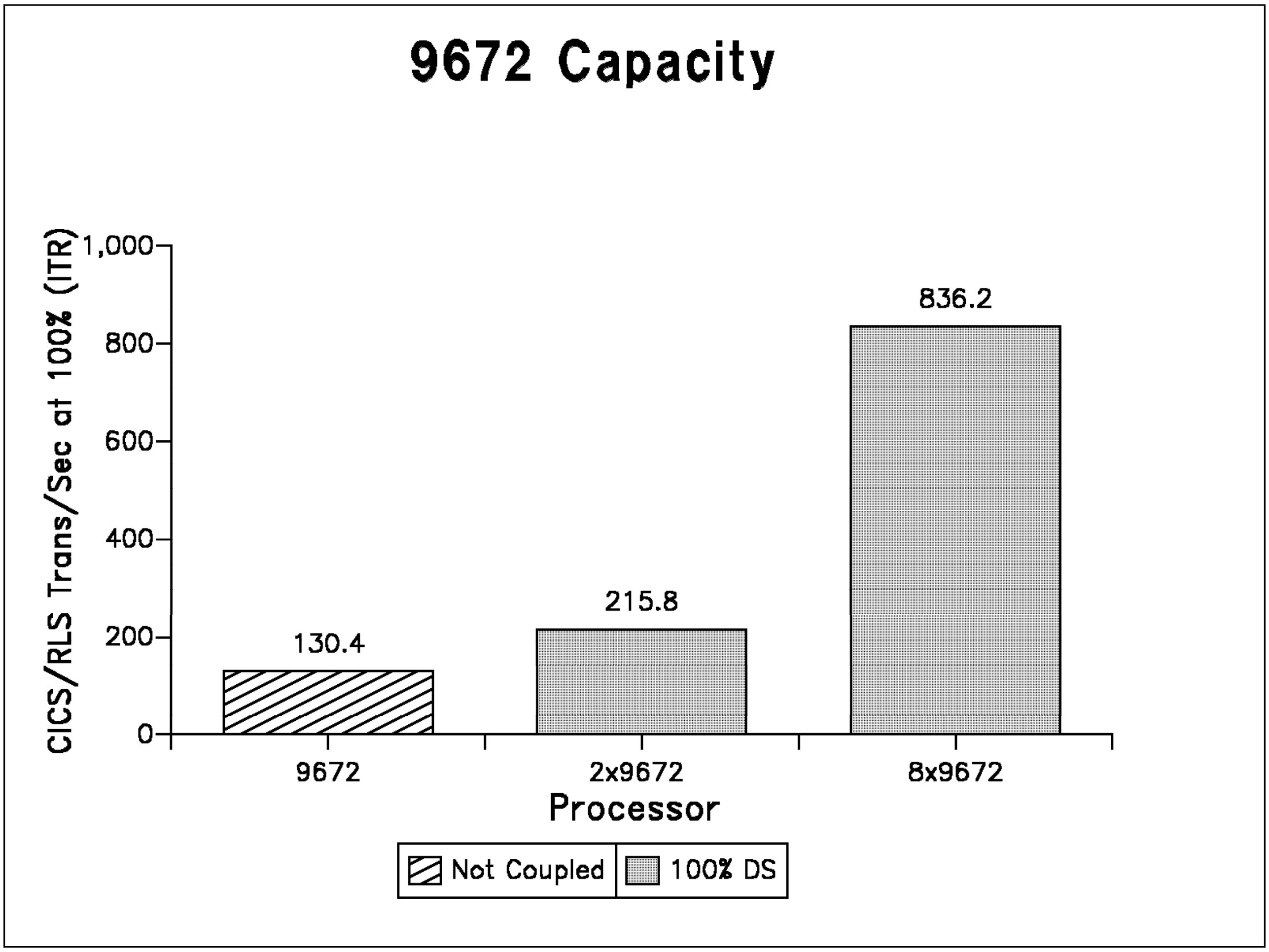


Figure 32. CICS/VSAM RLS Scalability Test ITR Comparison

with the performance enhancements provided by OW25820, OW25821, and OW29302.

This can be attributed to two things; the cost of multisystems management and the cost of data sharing. In the RLS environment, coupling facility accesses are made to the DFSMS lock and cache structures to enable data sharing across the Parallel Sysplex. For a detailed breakdown of the overheads when entering the multisystem data sharing environment, see Appendix C, “CICS/VSAM Workload Details and Coupling Efficiency Details” on page 183.

The internal CICS transaction response times increased slightly between the 9672 and 2x9672 measurements.

Scalability When Adding CECs to the Sysplex: The cost of going from a single system MRO Function Shipping environment 9672 to an 8x9672 RLS environment is 19.8%.

In looking at the additional overhead as we add systems to the 9672, we can calculate the coupling overhead. The coupling overhead can be calculated as the difference between the cost of going from one to two systems and the cost of going from one to eight. This number is then normalized by the number of additional systems (that is, six systems). For example:

Coupling overhead from 9672 to 2x9672 = 17.2%

Coupling overhead from 9672 to 8x9672 = 19.8%

Total additional overhead when adding six more systems to the sysplex is (19.8 − 17.2) = 2.6%. Therefore, the cost per additional system is $\frac{2.6}{6}$ = 0.43% additional overhead per system added.

Cost Per Additional CEC

This shows that as the number of systems in the sysplex grows, the cost of adding CECs is scalable at roughly 0.5% per CEC.

For a detailed breakdown of the overheads when entering the multisystem shared data environment, see C.1, “CICS/VSAM Workload Characteristics for a Single MVS Image” on page 183.

As discussed in 1.4.3, “Additional Workload Considerations” on page 12, the cost of data sharing is very workload-dependent. Our benchmark results reflect a stressed data sharing environment. Customer experiences from other data sharing environments have typically seen the cost to move to a sysplex data sharing environment to be about half that seen in our benchmarks.

6.3 MRO/CTC Special Study

A special study was made comparing the transaction rates and response times of MRO/CTC on a 2-way sysplex, versus a 2-way RLS datasharing environment. The study was made with the same hardware resources, software levels, tools and workload described earlier in this chapter. The 2-way RLS measurement data is also the same as was used earlier.

For the MRO/CTC environment, one of the systems was set up exactly like the MRO environment previously described. The other system of the MRO/CTC sysplex was also the same, except that it had no FOR; 100% of its file requests were remotely function-shipped to the other system.

The end result was that one system had 100% of its file requests function shipped locally and the other system had 100% of its file requests function shipped remotely. The target average transaction rate (ETR) to fully utilize each system was about 96 transactions per second.

Under these circumstances, the MRO/CTC sysplex was unable to achieve the external transaction rate of the 2-way RLS sysplex. This was primarily due to the fact that the single FOR was over 100% utilized. The transaction rate had to be reduced from the target of 96 to 66 transactions/sec to achieve reasonably acceptable FOR utilization (that is 82%) as shown in Table 23 on page 90. Note that the 2x9672 RLS TOR response time is higher than the Unconstrained MRO/CTC run since the RLS run was made at a much higher ETR, thus causing the CPU ulitization to reach 90%.

The other objective of the study was to scale the percentage of MRO file requests function-shipped between CECs using XCF. Results showed that at about 75% ships over XCF/CTCs, the average MRO cost per transaction was comparable to that of RLS, regardless of the number of systems involved.

6.3.1 Data

Table 23. CICS/VSAM RLS MRO/CTC Performance Study

	2x9672 MRO/CTC FOR-Constrained	2x9672 MRO/CTC Reduced ETR	2x9672 RLS
Avg Transactions/Sec (ETR)	86.70	66.06	96.46
CPU Activity - system 1	59.1%	46.6%	90.1%
CPU Activity - system 2	90.8%	74.0%	88.8%
FOR Utilization	103%	82%	N/A
TOR Response Time (sec)	.25	.09	.21

Chapter 7. CICS/DLI Asymmetric Configuration Performance Study

The purpose of this study is to examine the performance of data sharing processor complexes of unequal capacity, or more importantly, processors of unequal single engine speed, in a Parallel Sysplex environment.

In order to quantify the performance characteristics of asymmetric configurations in a Parallel Sysplex environment, the following three test cases were established:

1. Base 1-way (sysplex) 9021-821 with no sharing of data

   The objective of this test case is to determine a baseline for comparison to all other coupling measurements.

2. Adding two 9672-R61s to this base creates a 3-way (2+1) asymmetric configuration with 100% of the workload sharing data

   The objective of this test case is to evaluate the initial cost to enter the Parallel Sysplex environment in an asymmetric configuration.

3. Adding eight 9672-R61s to this base creates a 9-way (8+1) asymmetric configuration with 100% of the workload sharing data

   The objective of this test case is to evaluate the scalability of the Parallel Sysplex solution in an asymmetric configuration.

These three test cases revealed several important results:

- The initial cost of data sharing for 9021-711 bipolar-based machines when adding 9672-R61 CMOS-based technology
- The linearity of growth in asymmetric configurations
- Effects on response time due to processors of unequal processing power or more appropriately, single engine speed

7.1 Environment Overview

Figure 33 on page 92 illustrates the configuration used for these tests:

I/O connectivity: The 9021-821 had four channel paths to the shared IMS databases, while each of the 9672-R61 had two. The processors were connected to 9032 ESCON Directors, which in turn were connected to 3990 Model 6 cached controllers. All IMS databases were allocated on 3390 Model 3 triple capacity volumes.

CF connectivity: The 9021-821 has two CF sender links configured to each 9674 coupling facility. Each 9672-R61 has one CF sender link configured to each 9674.

MVS sysplex: The master catalog was shared by all systems. JES2 was run in a MAS with the checkpoint residing on the coupling facility. XCF signalling was achieved via the coupling facility, with two structures defined. Each image in the sysplex was identical; cloning techniques were used to achieve this. MVS was run with WLM in COMPAT mode. See Chapter 2, “Tuning Recommendations” on page 21 for a detailed description of the dispatching priorities used in this environment.

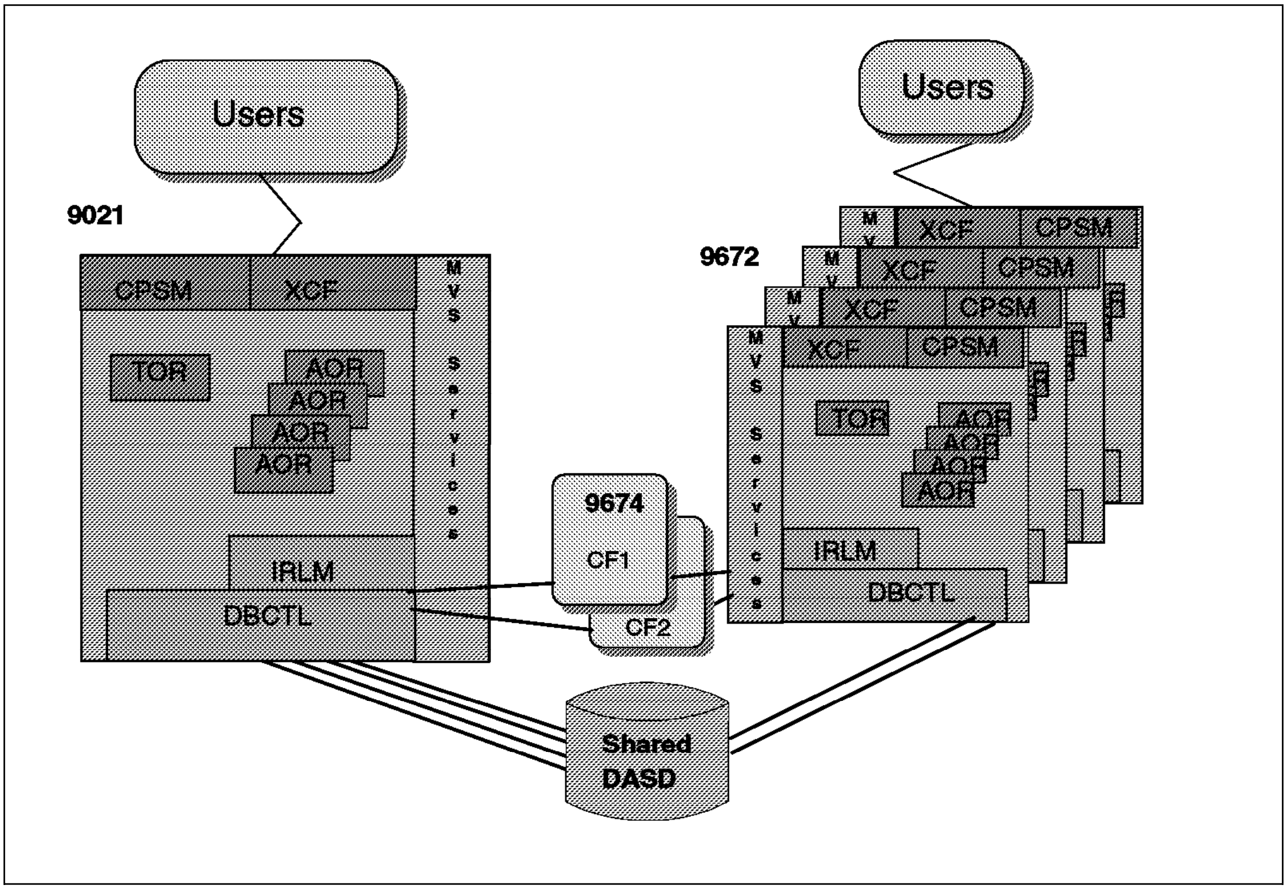


Figure 33. Asymmetric Configuration — CICS/DBCTL Test Case

Workload balancing: CICSplex SM (CP/SM) was used for workload balancing. The CP/SM configuration was such that each CMAS had the capability to communicate with any other CMAS in the CICSPlex. The QUEUE algorithm was employed for these measurements.

Transaction routing: CICS 4.1 MRO/XCF functions were used for dynamic transaction routing. All XCF communications are via the coupling facility; therefore, any inter-CEC MRO messages sent between CICS regions exploited XCF and the coupling facility. The CP/SM workload specification was defined so that all TORs in the configuration could route to any AOR in the sysplex.

IMS data sharing: The sysplex was configured with 100% connectivity to all IMS data. All IMS DBCTL regions were in a single data sharing group managed by IRLM. All IMS systems shared all data.

7.1.1 Hardware Resources

7.1.1.1 Processors

Table 24 lists the hardware resources used in the study.

Table 24. Hardware Resources — CICS/DBCTL Asymmetric Configuration

Processor Model	**9021-821**[21]	**2x9672-R61**	**8x9672-R61**	**9674**
# of CECs	1	2	8	1
# of processors/CEC	2	6	6	6
Central storage/CEC (installed)	512 MB	512 MB	512 MB	2048 MB[22]
Expanded storage/CEC (installed)	2048 MB	0 MB	0 MB	0 MB
# of CF sender links/CEC to CF	2	1	1	N/A
# of CF receiver links on CF	N/A	N/A	N/A	16

The 9674s were at MEC level D57264.

7.1.1.2 Coupling Facilities

The coupling facilities used in these experiments were dedicated 9674s. Each CF was a single LPAR partition with all hardware resources dedicated to that partition.

7.1.2 Software Levels

The following software products were used in all measurement test cases:

- CICS 4.1
- CICSplex SM 1.1.1 with AN65633 PTF applied
- IMS 5.1 (PI level)
- IRLM 2.1 (PI level)
- JES2 5.1.0
- MVS 5.1.0 at Service Level 9406
- VTAM 4.2

7.1.3 Measurement and Reporting Tools

The following tools were used to measure the experiment.

- RMF 5.1
- CICS Statistics (DFHSTUP)

[21] 9021-821 created by physically partitioning a 9021-942.

[22] See Table 25 on page 94 and Table 26 on page 94 for actual storage usage.

7.1.4 Coupling Facility Exploitation

The following function′s subsystems exploited the coupling facility:

- JES2 checkpoint
- IRLM lock table
- XCF signalling structures
- OSAM and VSAM buffer invalidation structures

For the 3-way (2+1) asymmetric configuration, a single 9674 coupling facility (CF1) was used containing the following structures:

Table 25. Structure Size/Placement for 3-Way (2+1) Configuration

Structure Description	**Structure Name**	**CF**	**Structure Type**	**Structure Size**
XCF signalling	IXCSIG1	CF1	List	69 MB
XCF signalling	IXCSIG2	CF1	List	69 MB
IRLM lock table	IRLMLOCKTBL2	CF1	Lock	64 MB
JES2 checkpoint	JES2CKPT	CF1	List	13 MB
OSAM buffer invalidation	OSAMSESXI	CF1	Cache	12 MB
VSAM buffer invalidation	VSAMSESXI	CF1	Cache	12 MB
Total CF storage used[23]:				239 MB

For the 9-way (8+1) asymmetric configuration, two (2) 9674 coupling facilities (CF1/CF2) were used containing the following structures:

Table 26. Structure Size/Placement for 9-Way (8+1) Configuration

Structure Description	**Structure Name**	**CF**	**Structure Type**	**Structure Size**
IRLM lock table	IRLMLOCKTBL2	CF1	Lock	64 MB
OSAM buffer invalidation	OSAMSESXI	CF1	Cache	12 MB
XCF signalling	IXCSIG1	CF2	List	69 MB
XCF signalling	IXCSIG2	CF2	List	69 MB
JES2 checkpoint	JES2CKPT	CF2	List	13 MB
VSAM buffer invalidation	VSAMSESXI	CF2	Cache	12 MB
Total CF storage used[23]:				239 MB

[23] Although the 9674 coupling facilities were configured with 2048 MB of central storage each, only 239 MB total of the installed CF storage was used for these configurations.

7.1.5 Workload Description

The workload used for the following asymmetric configurations is the same CICS/DBCTL workload used in the Chapter 4, “CICS/DBCTL Data Sharing Scalability Study.” Refer to 4.1.5, “Workload Description” on page 68 for a full description of the workload.

7.1.6 Measurement Description

The measurements taken for the asymmetric configuration study are discussed in detail as follows.

7.1.6.1 Measurement Methodology

To ensure that the subsystems processed smoothly with no unnecessary points of contention, database tuning was done.

In order to keep the contention for databases and I/O to database volumes constant, the number of copies of the workload was adjusted as the configuration grew. Since the 9021-821 is roughly twice the processing power of a single 9672, double the number of databases were used for the 9021-821. This kept the access rate per copy of the workload constant throughout all configurations.

There was a simulated remote network attached to each CEC. For all measurements in this performance study, a constant, predefined number of users logged on to each TOR on each CEC. Since the 9021-821 is roughly twice the processing power of a single 9672, double the number of users were logged onto the 9021-821. The users then executed the workload transaction scripts. Routing the transaction to an AOR was done by the CP/SM QUEUE algorithm. All AORs in the configuration were eligible to run any transaction routed from any TOR.

The workload was run in steady-state at a predefined processor utilization. Given a constant number of users on each CEC, the processor utilization was controlled by the user think time. A fifteen minute RMF interval was captured for each measurement. The average processor utilization of 90% was achieved on each CEC.

7.1.6.2 Initial Cost of Data Sharing

Objective: The first test case focused on understanding the costs of data sharing as we took a single system CICS/DBCTL workload running on a 9021-821, and introduced it into a data sharing environment with two 9672-R61s.

Methodology: A 9021-821 measurement was taken as a baseline for comparison for all other measurements. The measurement was run in a noncoupled environment. The system was run with a sysplex configuration of XCFLOCAL and GRS of NONE. There was no data sharing, and the IRLM subsystem was not started. IMS Program Isolation (PI) was used for data integrity on the single system. TOR to AOR communication was via cross-memory services. The CP/SM “QUEUE” algorithm was used for intra-CEC workload balancing.

For the 3-way (2+1) asymmetric run, MVS was run in a sysplex. A single 9674 coupling facility was used. All coupling facility exploitations discussed in 7.1.4, “Coupling Facility Exploitation” on page 94 were used. The CICS/DBCTL workload was cloned to the second and third systems, and the IMS subsystems on all systems were in a data sharing group, which allowed sharing of all data.

7.1.6.3 Scalability When Adding CECs to the Sysplex

Objective: The majority of the cost when entering a data sharing environment is taken going from a single system with no sharing to two systems data sharing. It is then expected that the additional cost per CEC as the sysplex increases will scale appropriately. In order to show scalability as the asymmetric configuration grows, six additional 9672-R61s were added to the sysplex.

Methodology: The 9021-821 plus 8x9672-R61 sysplex was measured. Two 9674 coupling facilities were used. All coupling facility exploitations are discussed in 7.1.4, "Coupling Facility Exploitation" on page 94. All IMS subsystems were in a data sharing group sharing all IMS databases.

7.2 Results

Following are the detailed results of the test cases discussed in this performance study.

7.2.1 Data

Table 27 shows the performance metrics for the base 1-way 9021-821. Since no coupling facilities were used nor any data sharing exploited, no coupling facility rates are shown.

Table 27. Asymmetric Base 1-Way

	9021-821
Transactions/sec (ETR)	195.4
CPU utilization	93.2%
Transactions/sec at 100% (ITR)	209.7
TOR response time	0.105

Table 28 shows the performance metrics for the aggregate of the systems in the two 9672-R61s as well as the 9021-821. Since the 9672-R61 systems are running symmetrically, dividing the rates in the 2x9672-R61 column by two will produce the rates for a single 9672-R61 in an asymmetric data sharing environment.

Table 28 (Page 1 of 2). Asymmetric 3-Way (2+1) Configuration

<table>
<tr><th colspan="2"></th><th>2x9672-R61</th><th>9021-821</th><th>Total</th></tr>
<tr><td colspan="2">Transactions/sec (ETR)</td><td>162.5</td><td>130.5</td><td>293.0</td></tr>
<tr><td colspan="2">CPU activity</td><td>86.5%</td><td>84.7%</td><td></td></tr>
<tr><td colspan="2">Transactions/sec at 100% (ITR)</td><td>187.8</td><td>154.1</td><td>341.9</td></tr>
<tr><td colspan="2">TOR response time</td><td>0.194</td><td>0.107</td><td></td></tr>
<tr><td>CF utilization</td><td colspan="3">CF1</td><td>8.5%</td></tr>
</table>

Table 28 (Page 2 of 2). Asymmetric 3-Way (2+1) Configuration

	2x9672-R61	**9021-821**	**Total**
Coupling Facility Rates (Requests/Second)			
IXCSIG1	62.0	50.2	112.2
IXCSIG2	56.5	37.4	93.9
JESCKPT	0.8	0.5	1.2
IRLMLOCKTBL2	1049.5	849.0	1898.5
OSAMSESXI	154.0	123.3	277.3
VSAMSESXI	504.9	409.5	914.4

Table 29 shows the performance metrics for the aggregate of the eight 9672-R61 systems as well as the 9021-821. Since all eight 9672-R61 systems are running symmetrically, dividing the rates in the 8x9672-R61 column by eight will produce the rates for a single 9672-R61 in an asymmetric data sharing environment.

Table 29. Asymmetric 9-Way (8+1) Configuration

		8x9672-R61	**9021-821**	**Total**
Transactions/sec (ETR)		647.5	122.3	769.8
CPU activity		88.6%	82.3%	
Transactions/sec at 100% (ITR)		730.8	148.6	879.4
TOR response time		0.230	0.123	
CF utilization	CF1			18.4%
	CF2			9.7%
Coupling Facility Rates (Requests/Second)				
IXCSIG1		302.3	54.0	356.4
IXCSIG2		288.9	54.6	343.5
JESCKPT		3.3	0.5	3.8
IRLMLOCKTBL2		4273.6	809.8	5083.4
OSAMSESXI		1089.7	117.7	1207.4
VSAMSESXI		2356.0	447.0	2803.0

Table 30 shows a side-by-side comparison of the performance metrics of the three measurement test cases:

Table 30 (Page 1 of 2). Asymmetric Configuration Comparison

Configuration		**1**	**2+1**	**8+1**
Transactions/sec (ETR)		195.4	293.0	769.8
Transactions/sec at 100% (ITR)		209.7	341.9	879.4
CF utilization	CF1	N/A	8.5%	18.4%
	CF2	N/A	N/A	9.7%

Table 30 (Page 2 of 2). Asymmetric Configuration Comparison

Configuration	**1**	**2+1**	**8+1**
Coupling Facility Rates (Requests/Second)			
IXCSIG1	N/A	112.2	356.4
IXCSIG2	N/A	93.9	343.5
JESCKPT	N/A	1.2	3.8
IRLMLOCKTBL2	N/A	1898.5	5083.4
OSAMSESXI	N/A	277.3	1207.4
VSAMSESXI	N/A	914.4	2803.0

7.2.2 Asymmetric Configurations and Data Sharing

Figure 34 on page 99 contains a graphic view of the results from this experiment.

7.2.3 Observations/Conclusions

General: By calculating the requests per transaction for the OSAM and VSAM buffer invalidate structures, one will note an increase in the requests per transaction. This increase could be eliminated by better tuning of VSAM and OSAM local buffer pools when configuring additional systems to the data sharing group.

Although two 9674 coupling facilities were used in the 9-way (8+1) asymmetric configuration, one 9674 would have been adequate for that experiment.

The difference in transaction (TOR) response time between the 9021-821 and a single 9672-R61 can be attributed primarily to the difference in single engine speed of the each system. Given the CPU component of an OLTP (On-line Transaction Processing) transaction is a small percentage of the total transaction response time, OLTP type work lends itself nicely to the Parallel Sysplex solution. Customers should evaluate the response time components of their workloads to determine which work is appropriate to be offloaded to a Parallel Sysplex environment. Refer to 1.5.2, “Internal Response Time” on page 13.

The asymmetric configurations exhibited no “race” conditions; the faster 9021 711-based processor did not dominate the smaller, slower CMOS-based 9672 class processors. Coupling facility service times for 9672 class machines in an asymmetric environment were consistent with those in a symmetric 9672 environment, discussed in Chapter 4, “CICS/DBCTL Data Sharing Scalability Study” on page 65.

Initial Cost of Data Sharing: The initial cost of moving from a single system environment with no shared data, to a multisystem, asymmetric data sharing environment can be seen by comparing the base 9021-821 1-way measurement with the 3-way (2+1) measurement. Comparison of internal throughput shows a cost of 26.5% when entering the multisystem, data sharing environment. When entering the data sharing environment, IRLM is introduced. Management of data and locks across the sysplex is required. For a detailed breakdown of the overheads when entering the multisystem data sharing environment, see A.3, “Coupling Overhead Breakdown” on page 174.

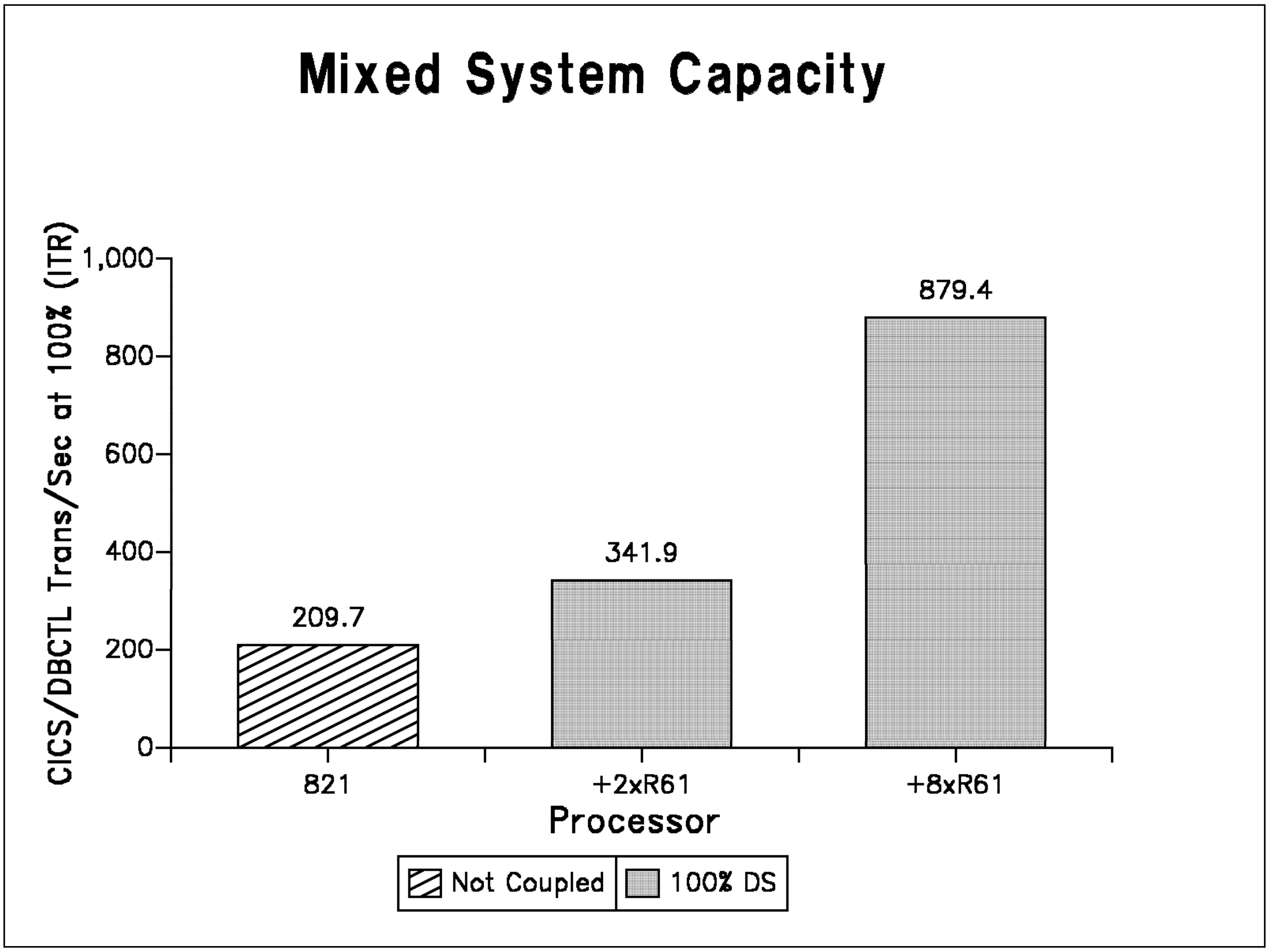


Figure 34. CICS/DBCTL Asymmetric Configuration ITR Comparison

The internal CICS transaction response times (TOR) remained constant on the 9021-821 when adding two 9672-R61s.

Linearity of Cost in Adding CECs to the Sysplex: The 9021 cost of going from a 9021 711-based single system environment with no shared data, to a 3-way (2+1) asymmetric configuration is 26.5%

The 9021 cost of going from a 9021 711-based single system environment to a 9-way (8+1) asymmetric configuration is 29.1%

In looking at the additional overhead as we add systems to the two 9672-R61s, we can calculate the coupling efficiency. The coupling efficiency can be calculated as the difference between the cost of going from one to two systems, and the cost of going from three (2+1) to nine (8+1) systems. This number is then normalized by the number of additional systems added (6). For example:

9021 Coupling overhead from 9021-821 to 9021-821 plus 2x9672-R61 = 26.5%

9021 Coupling overhead from 9021-821 to 9021-821 plus 8x9672-R61 = 29.1%

Total additional 9021 overhead when adding six more 9672 systems to the sysplex is (29.1 − 26.5) = 2.6%; therefore, the cost per additional 9672 system is $\frac{2.6}{6}$ = 0.43% additional overhead per 9672 system added, which is consistent with the 9672 overhead results found in Chapter 4, “CICS/DBCTL Data Sharing Scalability Study” on page 65.

Minimal Cost Per Additional CEC

This shows that as the number of 9672 systems in the sysplex grows, the cost of adding 9672 CECs is scalable at roughly 0.4% per CEC.

The internal CICS transaction response times (TOR) increased slightly when going from the 3-way (2+1) to the 9-way (8+1) asymmetric configuration, but the response times are still subsecond.

For a detailed breakdown of the overheads when entering the multisystem data sharing environment, see Appendix A, “CICS/DBCTL Workload Details and Coupling Efficiency Details” on page 171.

As discussed in 1.4.3, “Additional Workload Considerations” on page 12, the cost of data sharing is very workload-dependent. Our benchmark results reflect a stressed data sharing environment. Customer experiences have typically seen the cost to move to a sysplex data sharing environment to be about half that seen in our benchmarks.

Chapter 8. IMS-TM/DB2 Asymmetric Configuration Performance Study

The purpose of this study is to examine DB2 data sharing performance using 9021 711-based bipolar technology and 9672-R2/3 CMOS technology in an environment consisting of processor complexes of unequal capacity and unequal single engine speed.

In order to quantify the performance characteristics of asymmetric configurations in this environment, consider the following test cases:

1. Base single image 9021-941 with no data sharing

   The objective of this test case is to determine a basis for comparison to other coupling measurements.

2. Base single image 9672-RX3 with no data sharing

   As above, the objective of this test case is to determine a basis for comparison to other coupling measurements.

3. Adding two 9672-RX3s to the 9021-941 base environment creates a 3-way (2+1) asymmetric configuration with 100% of the workload sharing data

   The objective of this test case is to evaluate the cost to enter the data sharing environment in a 3-way asymmetric configuration.

These three test cases revealed several important results:

- The cost of data sharing for 9021 711-based bipolar processors and 9672-R2/3 CMOS processors
- The effects on response time due to processors of unequal capacity and unequal single engine speed
- The effects on internal throughput as processors initially run in a single system environment are integrated into a sysplex environment

8.1 Environment Overview

Figure 35 on page 102 illustrates the configuration used for these tests:

I/O connectivity: All processor complexes had four paths to the shared DB2 databases. The channel paths from each processor were connected to 9032 ESCON Directors which in turn were connected to 3990 controllers. All DB2 databases were allocated on 3390 Model 3 triple capacity volumes.

CF connectivity: The 9021-941 had four CF sender links configured to the 9674-C03 coupling facility. Each 9672-RX3 CEC had four CF sender link configured to the 9674.

MVS sysplex: The master catalog was shared by all systems. JES2 was run in a MAS with the checkpoint residing on DASD. XCF signalling was achieved via serial CTCs. Each image in the sysplex was identical. MVS was run with WLM in compatibility mode. See Chapter 2, “Tuning Recommendations” on page 21 for a detailed description of the dispatching priorities used in this environment.

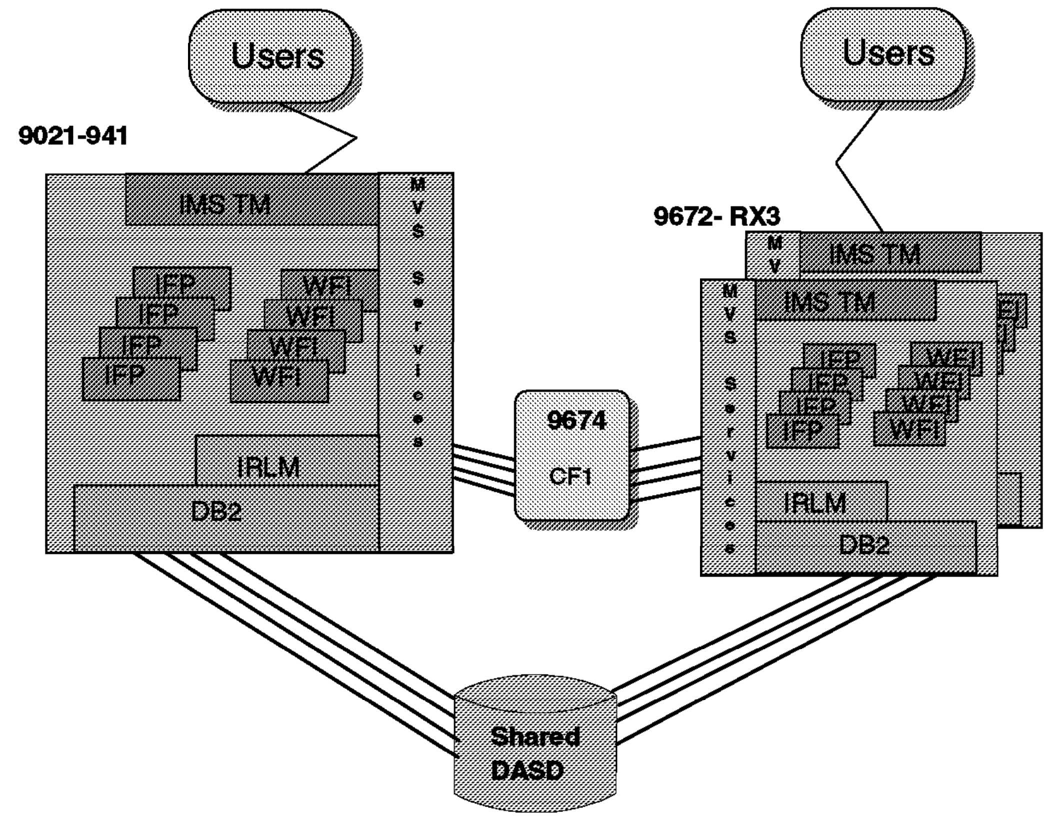


Figure 35. Asymmetric Configuration — IMS-TM/DB2 Test Case

Workload balancing: The workload was evenly distributed across the members of the sysplex. The number of IMS message processing regions (MPRs) was tuned to ensure no unnecessary points of contention existed.

Transaction routing: The workload did not contain any transaction affinities. All transactions were run on all systems.

DB2 data sharing: The sysplex was configured with 100% connectivity to all DB2 data. All DB2 members shared all data. Thus, there was 100% data sharing.

8.1.1 Hardware Resources

8.1.1.1 Processors

Table 31 lists the hardware resources used in this test.

Table 31. Hardware Resources — IMS-TM/DB2 Asymmetric Configuration

Processor Model	**9021-941**[24]	**9672-RX3**	**9674-C03**
# of CECs	1	2	1
# of processors/CEC	4	10	10
Central Storage/CEC (Installed)	1024 MB	2048 MB	1024 MB[25]
Expanded Storage/CEC (Installed)	1024 MB	0 MB	0 MB
# of CF Sender Links/CEC to CF	4	4	N/A
# of CF Receiver Links on CF	N/A	N/A	12

The 9674 E/C level was E45568 with MCL 097 for DR66 installed.

8.1.1.2 Coupling Facility

The coupling facility used in these experiments was a dedicated 9674-C03. The CF was a single LPAR partition with all hardware resources dedicated to that partition.

8.1.2 Software Levels

The following software products were used in all measurement test cases:

- MVS 5.2.0
- JES2 5.2.0
- VTAM 4.2.0
- IMS 4.1
- DB2 4.1
- IRLM 2.1

8.1.3 Measurement and Reporting Tools

The following tools were used for this measurement.

- RMF 5.2.0
- DB2 PM 4.1

[24] 9021-941 created by physically partitioning a 9021-982.

[25] See Table 32 for actual storage usage.

8.1.4 Coupling Facility Exploitation

The following functions′ subsystems exploited the coupling facility:

- DB2 group buffer pools
- DB2 shared communication area
- IRLM lock structure

For the 3-way (2+1) asymmetric configuration, the 9674 coupling facility contained the following structures, as shown in Table 32.

Table 32. Structure Size/Placement for 3-Way (2+1) Configuration

Structure Description	Structure Name	CF	Structure Type	Structure Size
DB2 Group Buffer Pools	DSNDB0G_GBP0 — DSNDB0G_GBP8	CF1	Cache	941 MB
DB2 Shared Communication Area	DSNDB0G_SCA	CF1	List	20 MB
IRLM Lock Structure	DSNDB0G_LOCK1	CF1	Lock	125 MB

8.1.5 Workload Description

The workload used for the following asymmetric configurations is the same IMS-TM/DB2 used in the Chapter 5, “IMS-TM/DB2 Data Sharing Scalability Study.” Refer to Appendix B, “IMS-TM/DB2 Workload Details and Coupling Efficiency Details” on page 177 for a full description of the workload.

8.1.6 Measurement Description

The measurements taken for the asymmetric configuration study are discussed in detail below.

8.1.6.1 Measurement Methodology

To ensure that the subsystems handled work smoothly with no unnecessary points of contention, database tuning was done.

Large database tables with corresponding indexes were partitioned, with up to 60 partitions. These partitions were spread across the DASD volumes to minimize DASD contention. Nine local buffer pools were used. The assignment of tables to buffer pools, and the size of the buffer pools were based on the frequency of reference and reuse characteristics of the tables.

In order to keep the contention for databases and I/O to database volumes constant, three replicates of the database were used for all runs.

There was a simulated remote network attached to each CEC. For all measurements in this performance study, a constant, predefined number of users logged on to each IMS subsystem. The users then executed the workload transaction scripts. All users in the configuration are eligible to run any transaction against any database replicate. Binds for all transactions were done on the 9021-941.

The workload was run in steady-state at a predefined processor utilization. Given a constant number of users on each CEC, the processor utilization was controlled by the user think time. A 20 minute RMF interval was captured for each measurement. The average processor utilization of at least 70% was achieved on each CEC.

8.1.6.2 Cost of Data Sharing

Objective: This test case focused on understanding the costs of data sharing as we take a single system IMS-TM/DB2 workload and introduce it into a data sharing environment.

Methodology: Two separate measurements were taken as a basis for comparison to other measurements. The first was done on a 9021-941. The second was done on a 9672-RX3. Each measurement environment was single image, with no data sharing.

For the 3-way (2+1) asymmetric run, MVS was run in a sysplex. A 9674-C03 coupling facility was used with the exploitations discussed earlier in 8.1.4, "Coupling Facility Exploitation" on page 104. The IMS-TM/DB2 workload was replicated to the second and third systems, and the DB2 members on all systems were in a data sharing group, which allowed sharing of all data. Thus, there was 100% data sharing.

8.2 Results

Detailed results of the measurements in this performance study follow.

8.2.1 Data

Table 33 shows the performance metrics for the base single image 9021-941. Since the coupling facility was not used and no data was shared, no coupling facility rates are shown.

Table 33. Asymmetric Base Single Image 9021-941

	9021-941
Transactions/sec (ETR)	114.3
CPU utilization	76.1%
Transactions/sec at 100% (ITR)	150.2
CPU milliseconds per transaction	27
Class 2 elapsed time, seconds	0.4

Table 34 shows the performance metrics for the base single image 9672-RX3. As above, the coupling facility was not used and no data was shared. Thus, no coupling facility rates are shown.

Table 34. Asymmetric Base Single Image 9672-RX3

	9672-RX3
Transactions/sec (ETR)	95.0
CPU utilization	76.3%
Transactions/sec at 100% (ITR)	124.5
CPU milliseconds per transaction	80
Class 2 elapsed time, seconds	0.6

Table 35 shows the performance metrics for the aggregate of the systems in the 9672-RX3 as well as the 9021-941. Since both systems in the 9672-RX3 are running symmetrically, dividing the rates in the 9672-RX3 column by two will approximate the rates for a single 9672-RX3 in an asymmetric data sharing environment.

Table 35. Asymmetric 3-Way Configuration

	9021-941	**9672-RX3**	**Total**
Transactions/sec (ETR)	82.2	154.3	236.5
CPU activity	73.3%	75.4%	
Transactions/sec at 100% (ITR)	112.2	204.7	316.9
CPU milliseconds per transaction	36	98	
Class 2 elapsed time, seconds	0.4	0.6	
CF Utilization	CF1		12.5%
Coupling Facility Rates (Requests/Second)			
DSNDB0G_GBP0-DSNDB0G_GBP8	1556	3259	4815
DSNDB0G_SCA	2	4	6
DSNDB0G_LOCK1	1280	2382	3662

8.2.2 Asymmetric Configurations and Data Sharing

Figure 36 on page 107 contains a graphic view of the results from this experiment.

8.2.3 Observations/Conclusions

General: The difference in Class 2 Elapsed time between the 9021-941 and a single 9672-RX3 can be attributed primarily to the difference in single engine speed of each system. Similarly, the effect of the smaller CMOS engine is seen in the difference in CPU milliseconds per transaction between the two processors. However, moving into the data sharing environment had no effect on response time for either processor.

The asymmetric configurations exhibited no “race” conditions; the faster 9021 711-based processor did not dominate the 9672-R2/3 processors.

Cost of Data Sharing: The cost of moving from a single system environment with no shared data, to a multisystem asymmetric data sharing environment can be seen for both the 9021-941 and the 9672-RX3. This is done by comparing a processor′s internal throughput from its base measurement to the same processor′s internal throughput within the sysplex.

Thus, for the 9021-941, the cost to enter the multisystem data sharing environment is 25.3%.

For the 9672-RX3, the cost to enter the multisystem data sharing environment is 17.8%.

In addition, when entering the data sharing environment, management of data and locks across the sysplex is required. For a detailed breakdown of the overheads involved, see B.3, “Coupling Overhead Breakdown” on page 180.

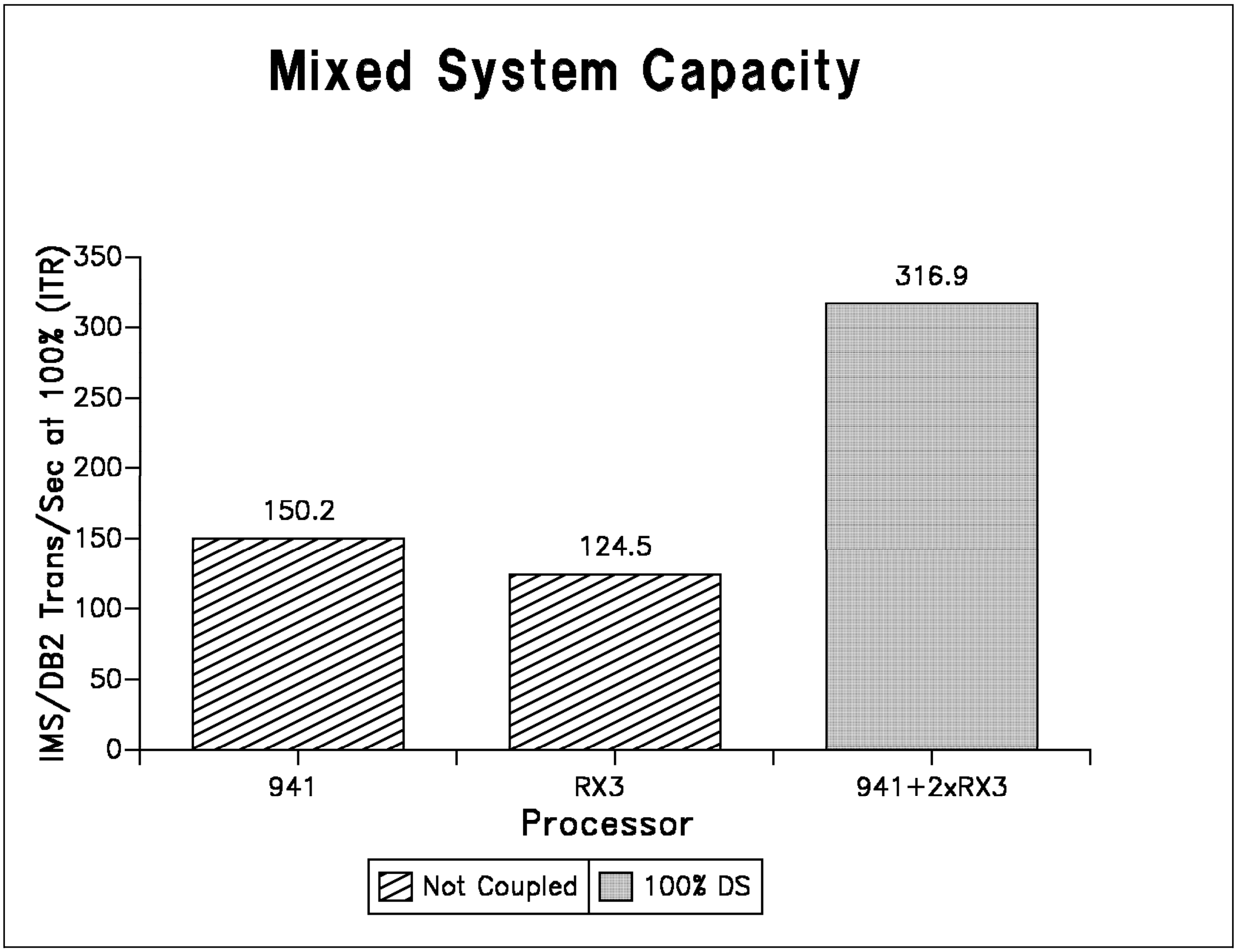


Figure 36. IMS-TM/DB2 Asymmetric Configuration ITR Comparison

As discussed in 1.4.3, “Additional Workload Considerations” on page 12, the cost of data sharing is very workload dependent. Our benchmark results reflect a stressed data sharing environment. Customer experiences have typically seen the cost to move to a sysplex data sharing environment to be about half that seen in our benchmarks.

Chapter 9. VSAM RLS Asymmetric Configuration Performance Study

The purpose of this study was to examine CICS/VSAM RLS performance using ES/9000 9021 model 821 bipolar technology and ES/9000 9672 model R63 CMOS technology in an environment consisting of processor complexes of unequal capacity and unequal single engine speed.

In order to quantify the performance characteristics of asymmetric configurations in this environment, consider the following test cases:

1. Base single image ES/9000 9021-821 with CICS/ESA 4.1 MRO Function Shipping.

   The objective of this test case was to determine a basis for comparison to other coupling measurements.

2. Base single MVS image running in a 3-way logical partition with dedicated CPs on a ES/9000 9672-R63 using CICS/ESA 4.1 MRO Function Shipping.

   As in the previous case, the objective of this test case is to determine a basis for comparison to other coupling measurements.

3. Adding two MVS images each running in a logical partition with three dedicated CPs on an ES/9000 9672-R63 to the ES/9000 9021-821 base environment creates a three-system sysplex with 100% Record Level Sharing.

   The objective of this test case was to evaluate the cost to enter the RLS environment in a 3-way asymmetric configuration.

These three test cases revealed several important results:

- The cost of data sharing for ES/9000 9021-821 bipolar processors and ES/9000 9672-R63 CMOS processors
- The effects on response time due to processors of unequal capacity and unequal single engine speed
- The effects on internal throughput as processors initially run in a single system environment were integrated into a sysplex environment

9.1 Environment Overview

Figure 37 on page 110 illustrates the configuration used for these tests:

I/O connectivity: All processor complexes had four paths to the shared CICS databases. The channel paths from each processor were connected to 9032 ESCON Directors; which in turn were connected to 3990 controllers. All CICS databases were allocated on 3390 Model 3 triple capacity volumes.

CF connectivity: The ES/9000 9021-821 had two CF sender links configured to the ES/9000 9674 model C03 coupling facility. Each ES/9000 9672-R63 CEC had two CF sender links configured to the ES/9000 9674 model C03 coupling facility.

MVS sysplex: The master catalog was shared by all systems. JES2 was run in a MAS with the checkpoint residing on CF. XCF signalling was achieved via serial CTCs. Each image in the sysplex was identical. MVS was run with WLM in compatibility mode. See Chapter 2, “Tuning

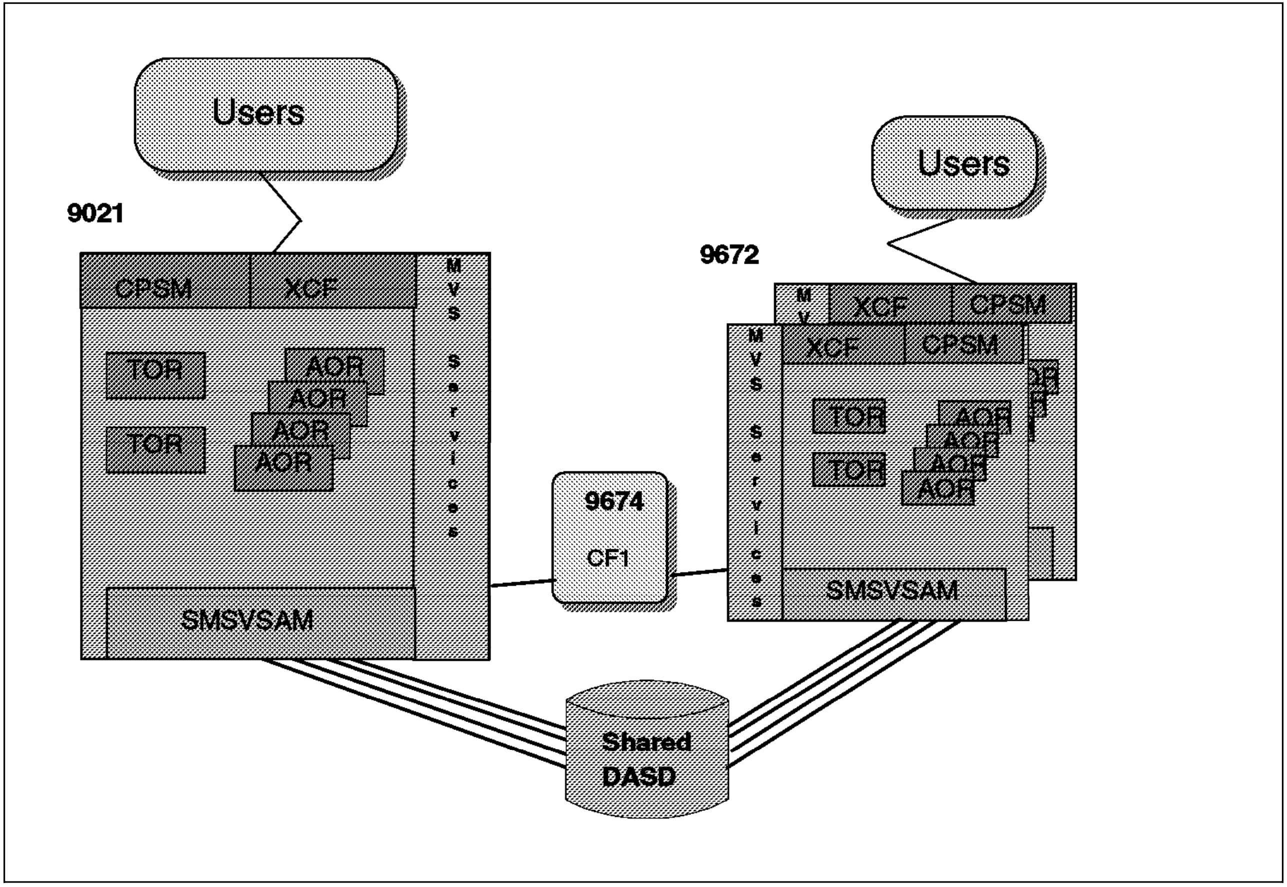


Figure 37. Asymmetric Configuration — CICS/VSAM RLS Test Case

Recommendations" on page 21 for a detailed description of the dispatching priorities used in this environment.

Workload balancing: CICSPlex SM (CP/SM) was used for workload balancing. The CICSPlex SM configuration was such that each CMAS had the capability to communicate with any other CMAS in the CICSPlex. The QUEUE algorithm was used for these measurements.

Transaction routing: CICS 4.1 and CICS TS 1.1 MRO/XCF functions were used for dynamic transaction routing. All CICS regions were part of the XCF group DFHIR000. Any inter-CEC MRO messages sent between CICS regions exploited XCF. The CICSPlex SM workload specification was defined so that all TORs in the configuration could route to any AOR in the sysplex.

CICS/VSAM RLS: The sysplex was configured with 100% connectivity to all CICS data. All CICS members shared all data. Thus, there was 100% RLS data sharing.

9.1.1 Hardware Resources

9.1.1.1 Processors

Table 36 lists the hardware resources used in this test.

Table 36. Hardware Resources — CICS/VSAM RLS Asymmetric Configuration

Processor Model	**9021-821**	**9672-R63**[26]	**9674-C03**
# of CECs	1	1	1
# of processors/CEC	2	6	10
Central Storage/CEC (Installed)	1024 MB[27]	1024 MB	2048 MB[28]
Expanded Storage/CEC (Installed)	0 MB[27]	0 MB	0 MB
# of CF Sender Links/CEC to CF	2	2	N/A
# of CF Receiver Links on CF	N/A	N/A	16

9.1.1.2 Coupling Facility

The coupling facility used in these experiments was a dedicated ES/9000 9674-C03. The CF was a single logical partition with all hardware resources dedicated to that partition.

9.1.2 Software Levels

The following software products were used in all measurement test cases:

- CICS TS 1.1 + APAR PN87305
- CICS/ESA V4.1 for comparison
- CICSplex SM 1.2.0
- JES2 5.2.0
- MVS 5.2.2 + APARs OW20060 and OW15588
- DFSMS 1.3 + APAR OW19918
- VTAM 4.3

9.1.3 Measurement and Reporting Tools

The following tools were used for the measurement of this test.

- RMF 5.2.0
- CICS Statistics (DFHSTUP)

[26] Two dedicated 3-way logical partitions were defined on the 9672-R63.

[27] For the 9021-821 base MRO measurement, 512 MB of Central storage and 512 MB of expanded storage were used.

[28] See Table 37 on page 112 for actual storage usage.

9.1.4 Coupling Facility Exploitation

The following functions′ subsystems exploited the coupling facility:

- JES2 checkpoint
- DFSMS CACHE
- VTAM Generic Resource
- LOGGER

For the 3-way (2+1) asymmetric configuration, the 9674 coupling facility contained the following structures, as shown in Table 37.

Table 37. Structure Size/Placement for 3-way (2+1) Configuration

Structure Description	Structure Name	CF	Structure Type	Structure Size
LOGGER	DFHLOG_STR1 — DFHLOG_STR2	CF	LIST	128 MB
LOGGER	DFHSHUNT_STR	CF	LIST	8 MB
JES2 checkpoint	COUPLE_CKPT1	CF	LIST	13 MB
VTAM generic	ISTGENERIC	CF	LIST	10 MB
DFSMS lock	IGWLOCK00	CF	LOCK	32 MB
DFSMS cache	SMSCACHE_STR1 — SMSCACHE_STR2	CF	CACHE	790 MB

9.1.5 Workload Description

The workload used for the following asymmetric configurations is the same CICS/VSAM RLS used in the Chapter 6, “VSAM RLS Data Sharing Scalability Study.” Please refer to Appendix C, “CICS/VSAM Workload Details and Coupling Efficiency Details” on page 183 for a full description of the workload.

9.1.6 Measurement Description

The measurements taken for the asymmetric configuration study are discussed in detail in the following sections.

9.1.6.1 Measurement Methodology

To ensure that the subsystems handled work smoothly with no unnecessary points of contention, database tuning was done. In order to keep the contention for databases, and the I/O to database volumes constant, the number of copies of the database was adjusted as the configuration grew.

There was a simulated remote network attached to each CEC. For all measurements in this performance study, a constant, predefined number of users logged on to each TOR on each CEC. The same number of users were logged on to each of the dedicated 3-way logical partitions on the ES/9000 9672-R63. Twice as many users were logged onto the ES/9000 9021-821 to compensate for the difference in processor capacity. The users then executed the workload transaction scripts. Routing the transaction to an AOR was done by the CP/SM QUEUE algorithm. All AORs in the configuration were eligible to run any transaction routed from any TOR. More than 90% of the transactions were processed on the local system.

The workload was run in a steady-state with a fixed number of users on each CEC and a fixed user think time. This was done to ensure a constant ETR for all measurements. A fifteen minute RMF interval was captured for each measurement.

The RLS_MAX_POOL_SIZE parameter in the IGDSMSxx parmlib member was set to 200 MB.

9.1.6.2 Cost of Data Sharing

Objective: This test case focused on understanding the costs of RLS as we took a single system CICS/VSAM RLS workload and introduced it into a RLS environment.

Methodology: Two separate measurements were taken as a basis for comparison to other measurements. The first was done on a ES/9000 9021-821. The second was done on a single dedicated 3-way logical partiton on a 9672-R63. Each measurement environment was single image, with CICS 4.1 MRO Function Shipping. All files were owned by a single FOR, and they were accessible from all of the AORs. The files were defined with backout-only recovery. Journaling was active. Local Shared Resources were used. For the ES/9000 9021-821, hiperspace buffers were defined.

For the 3-way asymmetric run (the 9021-821, and two dedicated 3-way logical partitions on a 9672-R63), MVS was run in a sysplex with CICS TS 1.1 with 100% RLS. All AORs shared all of the files. Backout-only recovery was enabled. The files were defined with no read integrity.

A ES/9000 9674-C03 coupling facility was used with the exploitations discussed earlier in 9.1.4, “Coupling Facility Exploitation” on page 112.

9.2 Results

Detailed results of the measurements in this performance study follow.

9.2.1 Data

Table 38 shows the performance metrics for the base single image ES/9000 9021-821.

Table 38. Asymmetric Base Single Image ES/9000 9021-821

	9021-821
Transactions/sec (ETR)	190
CPU utilization	70%
Transactions/sec at 100% (ITR)	271.4
CPU milliseconds per transaction	7.37
Transaction response time	.049

Table 39 shows the performance metrics for the base single dedicated 3-way logical partition on an ES/9000 9672-R63.

Table 39. Asymmetric Base Single Image 9672

	9672
Transactions/sec (ETR)	95.6
CPU utilization	73.3%
Transactions/sec at 100% (ITR)	130.35
CPU milliseconds per transaction	23.01
Transaction response time	.083

Table 40 shows the performance metrics for the aggregate of the systems in the ES/9000 9672-R63 as well as the ES/9000 9021-821. Since both systems in the ES/9000 9672-R63 are running symmetrically, dividing the rates in the ES/9000 9672-R63 column by two will approximate the rates for a single dedicated 3-way logical partition on a ES/9000 9672-R63 in an asymmetric data sharing environment.

Table 40. Asymmetric 3-way Configuration

		9021-821	**2x9672**	**Total**
Transactions/sec (ETR)		191.3	194	385.3
CPU activity		89.7%	89.5%	
Transactions/sec at 100% (ITR)		213.4	216.4	429.8
CPU ms per transaction		9.37	27.73	
Transaction response time		.095[29]	.330[29]	
CF utilization	CF1			4.5%
Coupling Facility Rates (Requests/Second)				
IGWLOCK00		286.1	282.4	568.5
SMSCACHE_STR1 - SMSCACHE_STR2		780.9	764.2	1545.1

9.2.2 Asymmetric Configurations and Data Sharing

Figure 38 on page 115 contains a graphic view of the results from this experiment.

9.2.3 Observations/Conclusions

General: The differences between the ES/9000 9021-821 and the single dedicated 3-way logical partition can be attributed primarily to the difference in single engine speed of each system. Similarly, the effect of the smaller CMOS engine is seen in the difference in CPU milliseconds per transaction between the two processors.

Cost of Data Sharing: The cost of moving from a single system environment with MRO Function Shipping, to a multisystem asymmetric RLS environment can be seen for both the ES/9000 9021-821 and the ES/9000 9672-R63. This is done by

[29] The increase in CICS TOR response time warranted further investigation. Subsequent measurements showed that the increase was due to increasing CPU queuing effects at higher CPU utilizations, most notably at CPU utilizations exceeding 90%. A 2x9672 measurement taken at a lower CPU utilization similar to the MRO base of 73.3% showed the TOR response to be approximately 0.08.

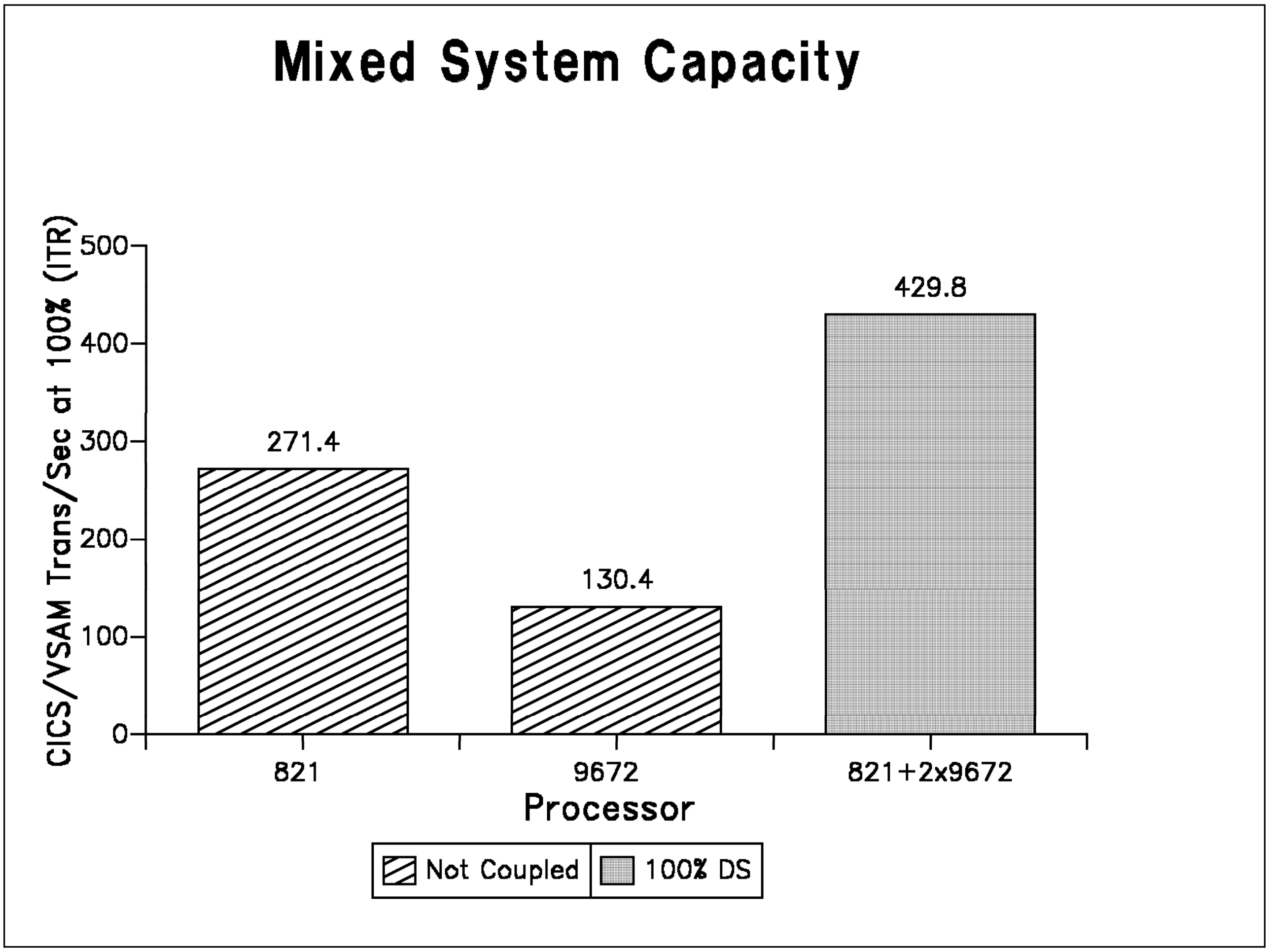


Figure 38. CICS/VSAM RLS Asymmetric Configuration ITR Comparison

comparing a processor′s internal throughput from its base measurement to the same processor′s internal throughput within the Parallel Sysplex.

Thus, for the ES/9000 9021-821, the cost to enter the multisystem RLS environment is 21.4%.

For the ES/9000 9672-R63, the cost to enter the multisystem RLS environment is 17%. The cost would be 19% if it were a 9672-R64. However, both costs would be reduced by approximately 3% with the performance enhancements provided by OW25820, OW25821, and OW29302.

For a detailed breakdown of the overheads involved, see C.3, "Coupling Overhead Breakdown" on page 186.

As discussed in 1.4.3, "Additional Workload Considerations" on page 12, the cost of data sharing is very workload-dependent. Our benchmark results reflect a stressed data sharing environment. Early customer experiences have typically seen the cost to move to a sysplex data sharing environment to be about half that seen in our benchmarks.

Chapter 10. IMS/TM DLI Data Sharing Performance Study

Prior to the availability of Parallel Sysplex, IMS data sharing between two systems could be done using a technique where lock requests were bundled and routed by one IMS Resource Lock Manager (IRLM) to the partner IRLM using a VTAM link. This was known colloquially as a “pass-the-buck” (PTB) scheme. This data sharing was limited to two MVS systems. With Parallel Sysplex, it is now feasible to share IMS workloads across several systems with reasonable response times by using the coupling facility (CF).

This study compared the two methods of sharing data between two systems (PTB/CTC versus CF) and measured the throughput of a Parallel Sysplex as the number of systems increased. Both studies used a similar workload.

10.1 IMS Data Sharing Workload

The IMS workload consisted of light-to-moderate transactions from DLI applications covering diverse business functions, including order entry, stock control, inventory tracking, production specifications, hotel reservations, banking, and teller functions. The IMS workload contained sets of 17 unique transactions, each of which had different transaction names and IDs, and used different databases. Conversational and wait-for-input transactions were included in the workload.

The number of copies of the workload and the number of message processing regions (MPRs) configured was adjusted to ensure that the IMS subsystem was processing smoothly, with no unnecessary points of contention. No batch message processing (BMP) was run. MVS WLM was running in COMPAT mode. IMS address spaces were nonswappable.

The IMS workload accessed both VSAM and OSAM databases, with VSAM indexes (primary and secondary). DLI HDAM and HIDAM access methods were used. The workload had a moderate I/O load.

Performance data collected consisted of IMSPARS, and the usual SMF data, including type 30 records (workload data), and RMF data.

IMS was measured by logging on a predetermined number of terminals, each of which executed scripts consisting of end-user actions. Once the logons were complete, the average think time was adjusted to provide a transaction rate that caused the processor to reach the target utilization level. After appropriate stabilization periods, measurements were made at approximately 90% processor busy. IMS was measured as a steady-state system over an elapsed period deemed adequate to produce a repeatable sample of work.

Table 41 on page 118 is a summary of the characteristics of the workload running on the various 9021 configurations.

Table 41. Workload Characteristics

	One 9021	Two 9021s
CPU milliseconds per tran	14.9	19.8
VSAM+OSAM DB I/O per tran	5.7	5.8
IRLM lock table req/tran to CF	N/A	8.5
OSAM cache req/tran to CF	N/A	1.6
VSAM cache req/tran to CF	N/A	4.3
Lock contention	N/A	0.3%

10.2 PTB/CTC versus CF Data Sharing Environment

This test was designed to compare the two methods of sharing data between two systems (PTB/CTC versus CF). Figure 39 on page 119 illustrates the configuration used for these tests.

10.2.1 Hardware Environment

Table 42 lists the hardware resources used in this test.

Table 42. Hardware Resources — IRLM Tests

Processor Model	**9021-711**[30]	**9674**
# of CECs	2	1
# of processors/CEC	1	6
Central Storage/CEC (Installed)	512 MB	2048 MB[31]
Expanded Storage/CEC (Installed)	2048 MB	0 MB
# of CF Sender Links/CEC to CF	2	N/A
# of CF Receiver Links on CF	N/A	16

10.2.2 Software Environment

The following software was used in this test.

- IMS 5.1 (PI level)
- MVS 5.1 at Service Level 9406
- JES2 5.1.0
- IRLM 2.1 (CF)
- IRLM 1.5 (PTB/CTC)
- VTAM 4.2

[30] The two 9021-711s were created by physically partitioning a 9021-842.

[31] See Table 43 on page 119 for actual storage usage.

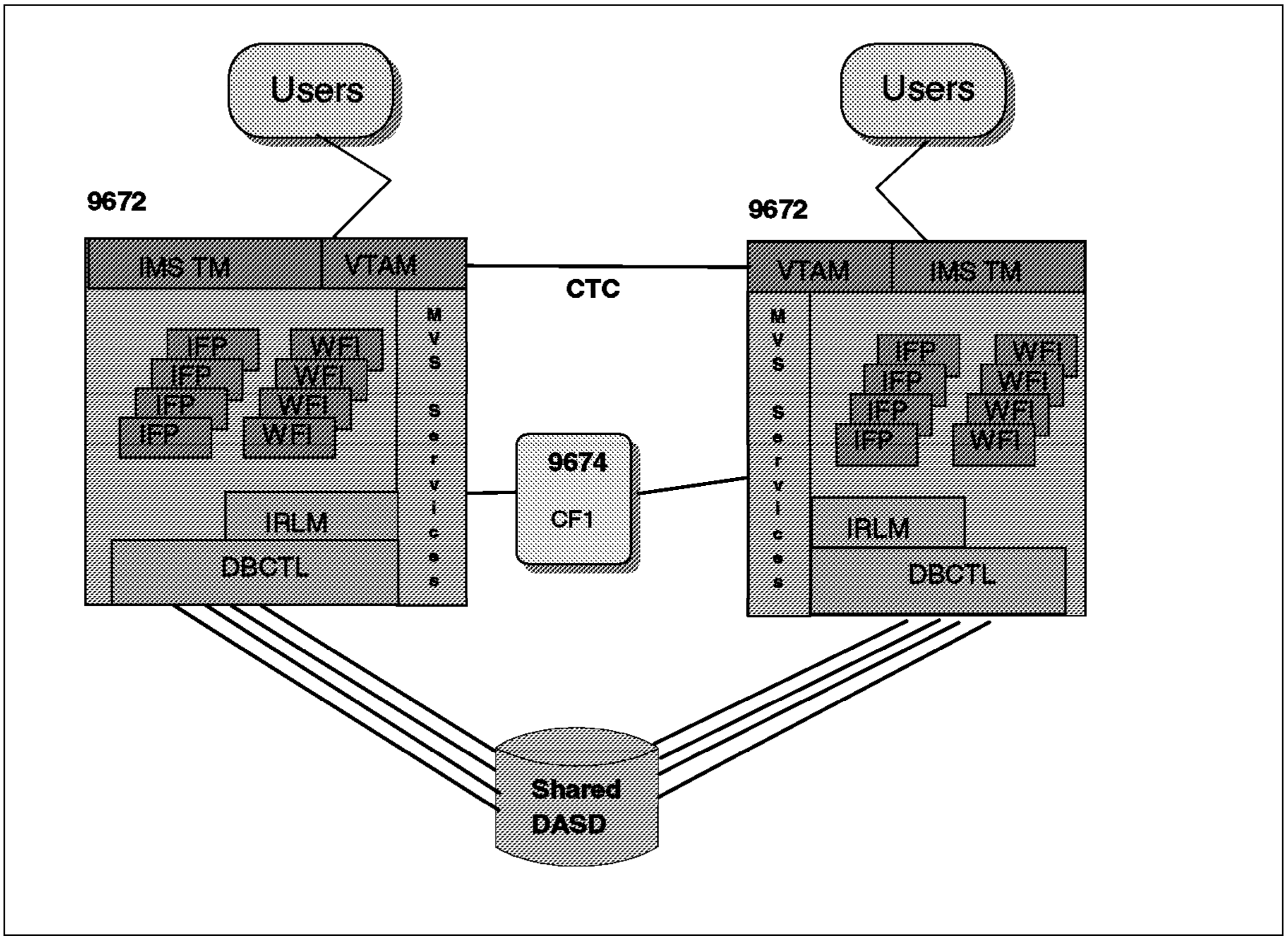


Figure 39. IMS Data Sharing Configuration

10.2.2.1 Structures on the CF

For the two 9021-711s, a single 9674 coupling facility (CF1) was used which contained the structures listed in Table 43.

Table 43. Structure Size/Placement for 2 9021-711 Configuration

Structure Description	Structure Name	CF	Structure Type	Structure Size
OSAM buffer invalidation	OSAMESXI2	CF1	Cache	12 MB
VSAM buffer invalidation	VSAMESXI2	CF1	Cache	12 MB
IRLM lock table	IRLMLOCKTBL2	CF1	Lock	64 MB
Total CF storage used:[32]				88 MB

[32] Although the 9674 coupling facility was configured with 2048 MB of central storage, only 88 MB, or 4.3%, was used for this configuration.

10.2.3 Measurement and Reporting Tools

For these runs the main tools used for reporting were:

- RMF 5.1
- IMSPARS

10.2.4 Methodology

The objective of these runs was to evaluate the performance differences between using a PTB/CTC technique for locking, and using the CF in a two-way data sharing environment.

In order to quantify the performance impact of using the coupling facility for data sharing we:

- Ran two ES/9000 9021 Model 711s data sharing via PTB/CTC
- Ran two ES/9000 9021 Model 711s data sharing with the CF

Each measurement was run for approximately 15 minutes after the workload stabilized. Eight replicates of the IMS workload were used for each measurement.

10.3 PTB/CTC versus CF Data Sharing Results

Based on the above tests at 100% data sharing the following results were obtained.

The cost of locking with the CF is a constant, remaining the same regardless of the lock request rate. This is not true with IRLM 1.5. The two IRLMs communicate by passing a VTAM message (the “buck”) back and forth between each other. The overhead per lock is a function of how many lock requests are contained in the buck. The number of locks per buck is a variable depending on lock rate and an IRLM tuning parameter, COMCYCL. COMCYCL specifies the length of time, in milliseconds, for the IRLM to delay before processing its inter-IRLM requests. Indirectly, COMCYCL influences the frequency of the buck being sent between the two systems sharing data. Given a constant lock request rate, a low COMCYCL value would yield a low number of locks per buck. Likewise a high COMCYCL value would yield a high number of locks per buck at the expense of prolonged transaction response time.

10.3.1 Observations/Conclusions

From the graph in Figure 40 on page 121, where the “Y” axis of the graph is a measure of processor busy time, it can be seen that when coupling 9672s, the 9674 CF should always provide equal or better performance than PTB/CTC. Note that the cost per lock for the 9672 and the 9021 are both constant, and the cost per lock for the PTB method drops as more locks are requested in a single VTAM message.

The story is effectively reversed when coupling with 9021s; the magnitude of the difference is very sensitive to the average number of locks per buck. For some workloads, depending on response time requirements and locking request rate, the cost of PTB locking can be less than using the CF. The PTB scheme is, however, limited to only two systems.

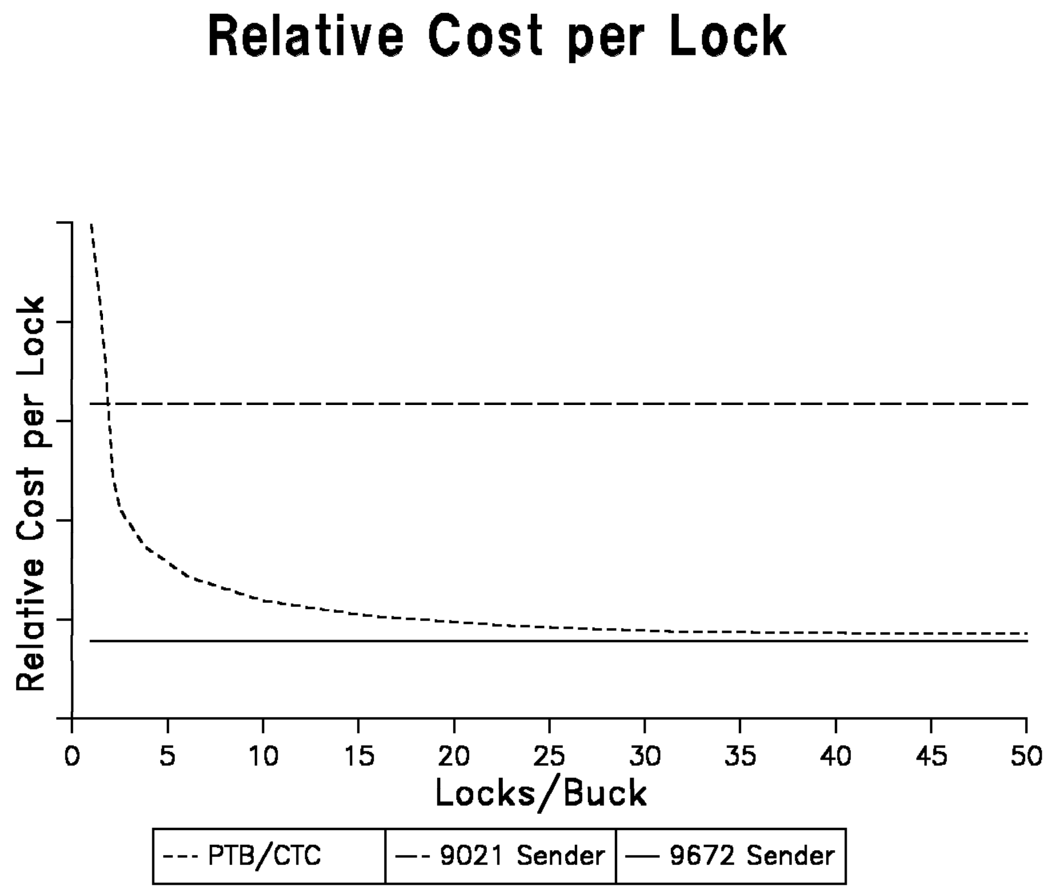


Figure 40. Comparison of PTB/CTC versus CF

10.4 Data Sharing Scalability Environment

This test measured the throughput of a Parallel Sysplex as the number of systems increased.

10.4.1 Hardware Environment

Table 44 lists the hardware resources used for this test.

Table 44 (Page 1 of 2). Hardware Resources — IMS TM Data Sharing Tests

Processor Model	**9021-821**[33]	**9674**
# of CECs	4	1[34]
# of processors/CEC	2	6
Central storage/CEC (installed)	512 MB	2048 MB[35]

Table 44 (Page 2 of 2). Hardware Resources — IMS TM Data Sharing Tests

Processor Model	**9021-821**[33]	**9674**
Expanded storage/CEC (installed)	2048 MB	0 MB
# of CF sender links/CEC to CF	2	N/A
# of CF receiver links on CF	N/A	16

The 9674s were at MEC level D57264.

10.4.2 Software Environment

The following software was used in this test:

- IMS 5.1 (PI level)
- MVS 5.1 at Service Level 9406
- JES2 5.1.0
- IRLM 2.1
- VTAM 4.2

10.4.2.1 Structures on the CF

For the 2 9021-821s, a single 9674 coupling facility (CF1) was used, which contained the structures listed in Table 45.

Table 45. Structure Size/Placement for Two 9021-821 Configuration

Structure Description	**Structure Name**	**CF**	**Structure Type**	**Structure Size**
OSAM buffer invalidation	OSAMESXI2	CF1	Cache	12 MB
VSAM buffer invalidation	VSAMESXI2	CF1	Cache	12 MB
IRLM lock table	IRLMLOCKTBL2	CF1	Lock	64 MB
Total CF storage used[36]:				88 MB

For the four 9021-821s, two 9674 coupling facilities (CF1 and CF2) were used, which contained the structures listed in Table 46 on page 123.

[33] Four 9021-821s were created by physically partitioning a Model 842 and a Model 982.

[34] Two 9674s were used to measure the environment with four 9021-821s.

[35] See Table 45 and Table 46 on page 123 for actual storage usage.

Table 46. Structure Size/Placement for Four 9021-821 Configuration

Structure Description	Structure Name	CF	Structure Type	Structure Size
OSAM buffer invalidation	OSAMEXSI2	CF2	Cache	24 MB
VSAM buffer invalidation	VSAMESXI2	CF2	Cache	24 MB
Total CF storage used[36]:				48 MB
IRLM lock table	IRLMLOCKTBL2	CF1	Lock	128 MB
Total CF storage used[36]:				128 MB

10.4.3 Measurement and Reporting Tools

The following tools were used for this test:

- RMF 5.1
- IMSPARS

10.4.4 Methodology

The initial performance cost to enter a Parallel Sysplex environment occurs with the migration from nondata sharing to 2-way data sharing. Part of the initial performance cost is a result of the required use of the IRLM in the IMS data sharing environment. IRLM is optional in the IMS nondata sharing environment, and most customers use IMS “Program Isolation” (PI) locking, which has better performance characteristics than IRLM.

Beyond two CECs, the cost will grow minimally for each CEC added. The runs consisted of:

- One ES/9000 9021 model 821 without data sharing and without IRLM
- Two ES/9000 9021 model 821s data sharing with the CF
- Four ES/9000 9021 model 821s data sharing with two CFs

An IMS workload modified to be more representative of our customers′ workloads was used on these experiments. The database replicates were scaled to the total processing capacity of the systems. For the 1-, 2-, and 4-way sysplexes, we used four, eight, and 16 database replicates were used respectively. The lock and cache structure sizes were also scaled with the number of database replicates. One CF was used for the 2-way and two CFs were used on the 4-way.

10.5 Data Sharing Scalability Results

The 9021s were measured at a utilization of approximately 90% while maintaining response times less than 0.25 of a second. Figure 41 on page 124 illustrates the results from this test.

[36] Although both coupling facilities were configured with 2048 MB of central storage, only the storage listed was used in these configurations.

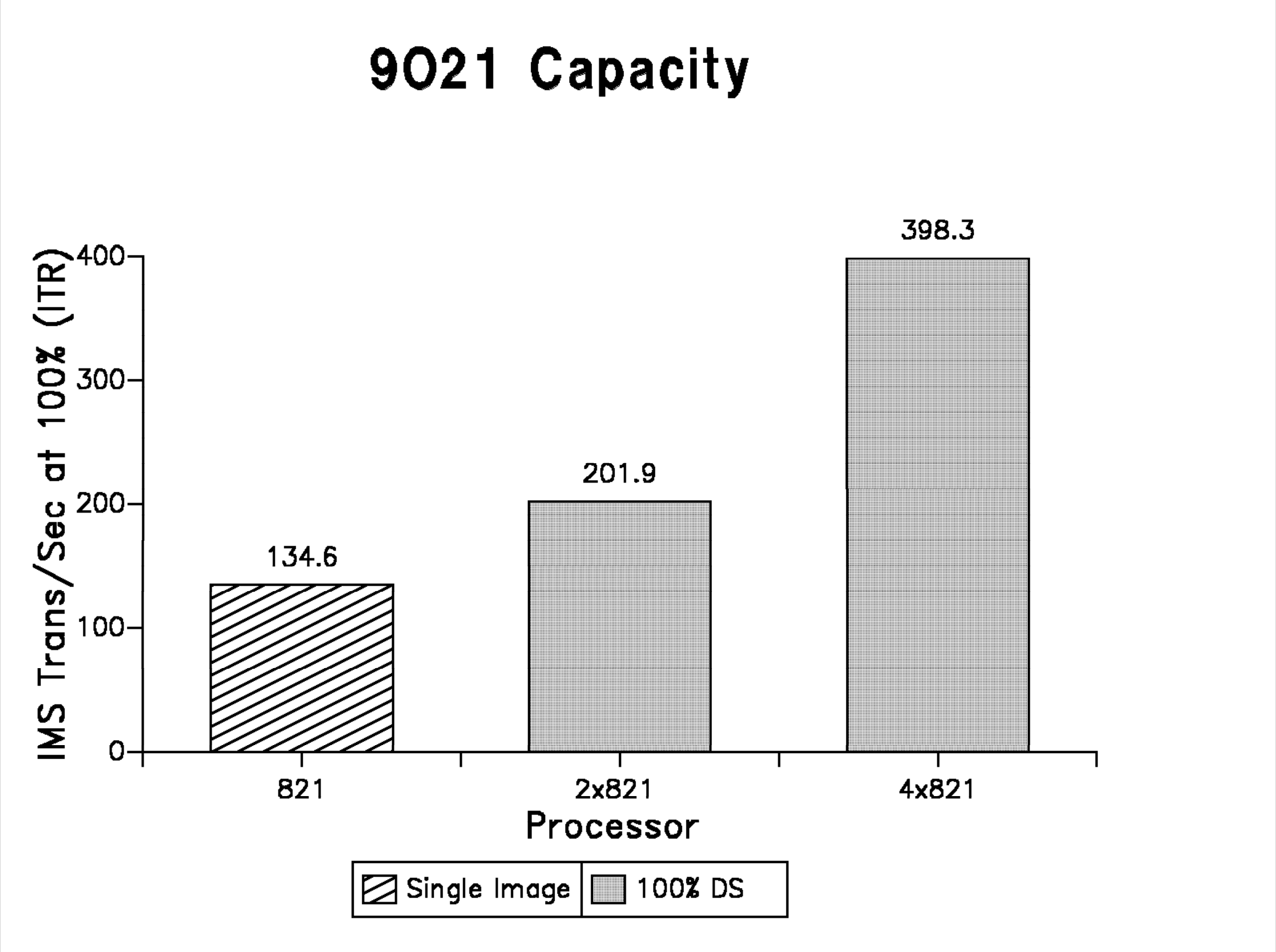


Figure 41. Comparison of 1- to 2- to 4-Way Data Sharing

10.5.1 Data

Table 47 contains the results for the IMS scalability tests.

Table 47. IMS Data Sharing

	One 9021-821	**Two 9021-821s**	**Four 9021-821s**
Transactions/sec (ETR)	125.3	178.8	357.7
CPU activity	93.1%	88.6%	89.8%
Transaction/sec at 100% CPU util (ITR)	134.6	201.9	398.3
CPU milliseconds/trans	14.8	19.8	20.1
Response time	0.128	0.168	0.201
CF utilization	N/A	CF1 5.7%	CF1 6.9% CF2 4.7%
Coupling Facility Rates (Requests/Second)			
IRLMLOCKTBL2	N/A	1479.8	3068.4
OSAMESXI2	N/A	285.4	565.2
ISAMESXI2	N/A	770.0	1506.0

10.5.2 Observations/Conclusions

A comparison of ITR for the 1-way system to the ITR of the 2-way Parallel Sysplex shows that the initial cost of joining a Parallel Sysplex in this configuration of 100% data sharing is 25.0%. A 1-way to 4-way comparison shows a cost of 26%.

Cost of Additional CEC

The cost of adding each additional CECs (beyond two) is 0.5% per CEC.

As discussed in 1.4.3, “Additional Workload Considerations” on page 12, the cost of data sharing is very workload-dependent. Our benchmark results reflect a stressed data sharing environment. Customer experiences have typically seen the cost to move to a sysplex data sharing environment to be about half that seen in our benchmarks.

Chapter 11. IMS/TM Version 6 Performance Study

With IMS Version 5, each IMS subsystem has its own queues for both input and output messages, and its own Expedited Message Handler for Fast Path messages. The IMS subsystem that receives a message processes it, unless that IMS is set up to send the message to another IMS subsystem (using MSC, Message Requeuer, or some other means).

With IMS Version 6, all of the IMS subsystems in a sysplex can share common sets of input/output queues, one set for full function messages and one set for Fast Path messages. A message placed on a shared queue can be processed by any IMS subsystem that has access to the shared queue and is capable of processing the message.

There are three main benefits in using shared queues:

- Automatic workload balancing across all IMS systems in the sysplex can be achieved by using shared queues. In prior releases of IMS, workload balancing was a user responsibility, requiring the use of methods such as network balancing, multiple system coupling (MSC), intersystem communication links (ISC), advanced program-to-program communication (APPC) and the workload router product. With shared queues, no single IMS will remain underutilized while other subsystems are saturated. Overall throughput is optimized to the full capability of the sysplex.
- Incremental growth can be achieved using shared queues. As workload increases, new IMS systems can be easily added to the sysplex to process the extra workload. This approach supports peak processing periods. Processors can also be removed from the sysplex when the extra capacity is no longer needed.
- Reliability, availability, and failure isolation can be increased using shared queues. If any IMS system in the sysplex fails, any of the remaining IMS systems can process the workload in the shared queues. If one or more of the IMS systems requires a cold start, the contents of the shared queues are not affected.

The objectives of this study are:

1. To measure shared message queue impact on a IMSplex running 100% data sharing, including CPU usage and transaction response time, in 1-, 2-, and 4-system sysplex configurations
2. To measure shared expedited message handler queue impact in an IMSplex Fast Path environment

11.1 Shared Message Queue (SMQ) Study

In this study, we explore the cost associated with implementing full function Shared Message Queues in an environment previously needing no message routing. This migration would be experienced by someone who now requires message routing for any number of reasons, or by someone wanting to take advantage of:

- Automatic workload balancing
- Incremental growth

- Reliability, availability, and failure isolation

11.1.1 Environment Overview

The following describes the workload and the environment for the SMQ measurements.

11.1.1.1 IMS/TM Full Function Workload

The IMS full function workload used for this study was a modified version of the Data System Workload (DSW). The DSW Workload consists of seventeen transactions covering diverse business functions:

- Teller system
- Data entry/accounting banking
- Order entry
- Stock control
- Hotel reservation
- Inventory tracking
- Production specification

There is a mix of transactions including conversational/non-conversational, inquiry only, response/non-response, and wait-for-input (WFI). These applications are written in COBOL, PL/I, and Assembler. The databases accessed by these transactions are a mixture of HIDAM, HISAM, and HDAM, with secondary index and logical relationships. Both OSAM and VSAM databases are used. The workload has a moderate I/O load. The network consists of SLU2-type terminals and SLU1-type printers.

The number of copies of the workload and the number of Message Processing Regions configured are adjusted to ensure that the IMS subsystem is processing smoothly, with no unnecessary points of contention. No Batch Message Processing (BMPs) are run.

The modification to the DSW workload made for the shared message queue study consisted of increasing the message sizes and the application program content to better match the profiles of the early support customers. Table 48 on page 129 lists the characteristics of the measured workload.

Table 48. Full Function Workload Details

Transaction DL/I calls DB	30.5
Transaction DL/I calls DC	4.0
Transaction DL/I calls SYS	.7
Transaction path length	2000 K
Message sizes Input(01)	1.8 K
Message sizes Output(03)	1.2 K
03:01 ratio	2:1
Messages per transaction	2.5
Multisegment messages	3%
Messages per Million Instructions	1.3
IRLM CF Structure accesses per transaction	18
OSAM CF Structure accesses per transaction	6
VSAM CF Structure accesses per transaction	10
IMS Logging per transaction	6 K
Database I/O per transaction	14-16

11.1.1.2 Hardware Environment

Figure 42 on page 130 illustrates the configuration used for these tests. The hardware environment consisted of:

- Up to 4 9672-RX3s
- Two 9674-C03s with 2 coupling facility links to each 9672
- 9394 RAMAC Array Subsystem for IMS system datasets and the IMS databases
- 3990 Model 6 with 3390 DASD with DASD fast write enabled for MVS logger data sets

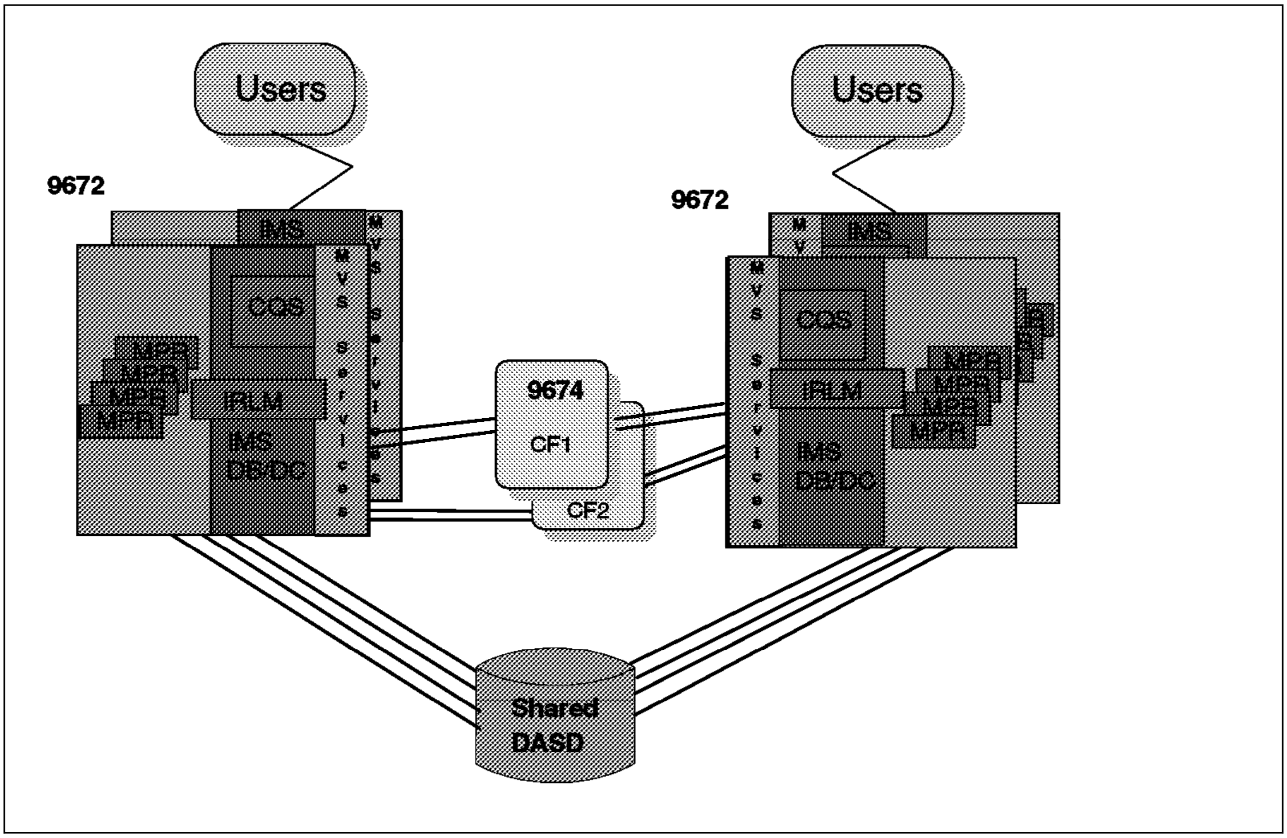


Figure 42. IMS SMQ Configuration

11.1.1.3 Structures on the CFs

Table 49 lists the structures on the CFs.

Table 49. Structure Size/Placement for the Configuration

Structure Description	Structure Name	CF	Structure Type	Structure Size
Logger	M52LOGMSGQ01	CF1	List	98 MB
IMS Shared Message Queue	I61MSGQ01	CF2	List	64 MB
IRLM lock	IRLMLOCKTBL1	CF2	Lock	128 MB
IMS OSAM	OSAMSESXT1	CF2	Cache	64 MB
IMS VSAM	VSAMSESXT1	CF2	Cache	64 MB
XCF message	IXCSTRUC	CF2	List	3 MB

11.1.1.4 Software Environment

The following software was used in this test:

- IMS 6.1
- IRLM 2.1
- MVS 5.2.2
- VTAM 4.3

11.1.1.5 Measurement and Reporting Tools

The following measurement and reporting tools were used for this study:

- RMF
- TPNS Response Time Utility
- IMS log reduction tools

11.1.1.6 Methodology

The full function workload was run with the same number of terminals and the same user think time for all test scenarios to ensure a constant transaction rate so that comparisons could be made with the base and the shared queues environments with regard to CPU and response time effects.

Teleprocessing Network Simulator (TPNS) was used to simulate end-user terminals. The users were evenly split across all systems via pre-determined scripted logon requests. The VTAM generic resource function was not used in this particular study. The systems were measured once a stable, steady state environment was reached.

11.1.2 SMQ Results

The first test case focused on understanding the costs of the shared message queue support. A system that was already defined for global data sharing was changed to exploit shared message queues. Even though one system was used in this test case, the costs for shared message queues (and global data sharing) are present since all of the functions were enabled.

Table 50 lists the details of the SMQ vs non-SMQ performance study.

Table 50 (Page 1 of 2). SMQ vs Non-SMQ

	Non-SMQ	**SMQ**
ETR	49.5	49.4
CPU Busy %	66.1	78.8
ITR	74.9	62.7
ITR delta		-16%
CPU milliseconds per tran	133.5	159.5
TPNS terminal response time	.26	.33
IMS Logging OLDS number of cylinders	276	337
SSCHs/second to logger offload DASD	0	8.6
CPU busy per system (RMF APPL%)		
IMS CTL	38	64
MPR	569	598
CQS	0	50
IXGLOGR (Offload)	0	6

Table 50 (Page 2 of 2). SMQ vs Non-SMQ

	Non-SMQ	SMQ
Coupling facility rates (total requests/second)		
Shared MsgQ structure	0	483•
Logger structure	0	263•
IRLMLOCKTBL	899	900
OSAM structure	293	295
VSAM structure	520	522
IXCSTRUC	3.8	3.8
Note: •Each message has an average of 3.9 MsgQ structure accesses, of which 2.6% are async.		
Note: •Each message has an average of 2.1 Logger structure accesses, of which 3.2% are async.		

Table 51 lists the details of the SMQ scalability measurements with a sysplex of 1-system, 2-system and 4-system configurations.

Table 51. 1/2/4-Systems

	1-System	2-Systems	4-Systems
ETR	49.4	98.7	197.4
Avg. CPU busy %	78.8	78.6	78.1
ITR	62.7	125.6	252.9
ITR per system	62.7	62.8	63.2
CPU milliseconds per tran	159.4	159.3	158.2
TPNS terminal response time	.33	.32	.33
CPU busy per system (RMF APPL%)			
IMS CTL	64	64	65
MPR	598	594	588
CQS	50	50	51
IXGLOGR	6.0	5.4	5.5
Coupling facility rates (total requests/second)			
Shared MsgQ structure	483	972	1958
Logger structure	263	523	1083
IRLMLOCKTBL	900	1797	3593
OSAM structure	295	596	1175
VSAM structure	522	1031	2066
IXCSTRUC	4	100	216

11.1.3 Observations/Conclusions

In the comparison of a system not using Shared Message Queues to a system exploiting Shared Message Queues, a 16% processing cost was incurred in the measurement environment.

The significant components of the Shared Message Queue cost are comprised of the following:

1. 3-5 shared message queue structure accesses per message. This processing involves the accesses to the coupling facility and the software processing to manage the shared message activity. This additional software processing is spread across the IMS Control Region, Common Queue Server (CQS), and MPR dispatchable units and includes the use of the MVS cross-system extended services (XES) to interface with the coupling facility. The accesses to the structure generally consist of:

   - Writing an IMS message
   - Reading the message for processing/delivery
   - Deleting the message on the shared queue at completion
   - Notification logic for detecting the presence of a message on a queue (queue transition from empty to non-empty)
   - Potentially an additional read to detect the emptiness of a queue (queue transition from non-empty to empty)

   In the case of the same IMS that originally received an input message across the network and having the system resources available to process it, some activity is eliminated related to reading the message from the shared queue and the notification logic. This is known as a local optimization benefit. In addition, costs are spared when the system is busy and the queues are full. The queue transition state activity would become minimal.

2. 2-3 accesses per IMS message to the shared system log structure on the coupling facility and the software processing (CQS dispatchable units including use of MVS logger and XES services) associated with this activity. The accesses consist primarily of logging the events of writing and reading a shared message, as well as the message delete and additional message write log records which are batched.

3. Additional logging cost is incurred both in IMS and in MVS. IMS will log more data to its own log data sets in the shared queues environment. Also, the MVS system logger will have additional processing incurred with offloading its log records from the coupling facility to DASD log data sets.

4. There may be additional IMS pseudo-scheduling costs with multiple systems potentially being notified with the presence of a message on a shared queue.

In the scalability measurement involving 1-system, 2-system and 4-system sysplexes, no measurable increase in processing costs were seen as systems were added. In addition, no response time differences were seen as systems were added. The conclusion is the expense of exploiting shared message queues is seen with the first system migrating to this environment, and the processing is very scalable and consistent as more systems are added to the sysplex.

The benchmark results shown were generated in a stable laboratory environment and the degree of any costs are very dependent on the workload characteristics. Key workload factors that will affect your results consist of the following:

- Messaging intensity
  - Messages per transaction
  - Transaction rate and transaction path length size
- Host and coupling facility configuration

- Message size
- Percent of input messages processed locally

Recent benchmarks with the same lab workload in a G4 configuration have shown costs for the shared message queues function falling in the 12-14% ITR range. Some of this improvement is due to advances in the CMOS technology, as well as some software pathlength reductions.

11.2 Shared EMH Queue Study

The Shared Expedited Message Handler (EMH) Queues support for Fast Path transactions allows different options for sysplex exploitation. The following options are available at the program definition level (balancing group level):

- Local first (default)- if IFP regions are available, process message locally. Otherwise, put the message on the global shared queue.
- Local only - handle message locally, never use the global queue.
- Global only - all messages are put on the global queue, even if a local IFP region is available.

If a message is handled locally, there is no activity and no overhead associated with the shared queues on the coupling facility. If a message is handled on the global queue, then the processing will be similar to the basic shared message queue activity outlined for full function messages in 11.1.3, “Observations/Conclusions” on page 132. The facilities of CQS will be used, and in turn, the MVS logger and XES services will be exploited. This processing will be driven by the accesses required to put, retrieve, delete, and log the shared message activity on the global structures in the coupling facility. As a result, the absolute cost for this activity for a fast path message to be put on the global queue will be similar to the costs for a full function shared message. However, since a fast path transaction generally is much smaller than a full function transaction, the relative overhead associated with a global fast path message will be significantly higher than that of a full function message. Overall, the CPU usage of a fast path workload will increase as the proportion of global transactions increases, and region occupancy time of dependent regions and IRLM lock contention may be greater when transactions are processed globally.

With the Local First option, the capability to exploit workload balancing can be retained without paying any cost during those periods when it is not actually needed. This result is shown in Table 52 with a 2-system data sharing sysplex running only a pure fast path workload in a G4 configuration. No additional cost was measured with the Local First option and no global routing.

Table 52. Shared EMH Queue 2-System Sysplex

	Local Only Option	**Local First (no global trans)**
ETR	275	276
ITR	1250	1255
Avg. CPU Busy %	22	22
Average Transit Time in Seconds	0.132	0.131

When a stressed system capacity is exceeded, a transaction input message will be placed on the shared EMH Queue and can be processed by other IMS systems which are not as heavily loaded. This approach allows more work to flow through the systems. In the extreme setup of all input messages handled globally (global-only definition), the following worst-case cost results were seen in a measurement in a G4 configuration. Table 53 lists the results of a global-only Shared EMH Queue measurement.

Table 53. Shared EMH Queue Global-Only

	Local First (no global trans)	**Global Only**
ETR	383	377
ITR	1260	556
Avg. CPU busy %	30	69
Average transit time in seconds	0.156	0.184

The global-only option is recommended to be chosen judiciously for only those transactions that always need the availability provided with shared queues because of the costs. The local first option would be preferred in most scenarios because of the absence of cost when local capacity is sufficient to handle the activity of all fast path transactions, and the overall throughput benefit gained when the local capacity is exceeded. The local-only option would be chosen for those fast path transactions and workloads that can never tolerate any increase in processing time.

Chapter 12. IMS/DB Version 6 Performance Study

IMS Version 6 provides three database sharing enhancements for the Parallel Sysplex environment:

- Block-level sharing of Virtual Storage Option (VSO) DEDB areas
- Block-level sharing of sequential dependent (SDEP) control intervals of DEDBs
- OSAM database caching of data in the CF

The benefits of these enhancements are:

- Block-level sharing of VSO data entry database (DEDB) areas allows multiple IMS subsystems to concurrently read and update VSO DEDB data. This provides potentially greater availability for those applications needing to access VSO DEDB areas.
- Shared Sequential Dependent control intervals remove the existing restriction that precludes data sharing of SDEPs so n-way sharing can be fully exploited, potentially resulting in application availability improvements.
- OSAM caching can reduce I/O delays in block-level data sharing systems where OSAM data is being updated and concurrently accessed by other IMS subsystems. This is accomplished by having updated blocks available from the coupling facility instead of refreshing the buffer from DASD. This reduction of I/O delay could result in increased performance for applications requiring access to shared OSAM data.

The objectives of this study are:

1. To assess the system performance cost from non-block-level data sharing VSO areas to block-level data sharing VSO areas
2. To demonstrate the performance benefit of shared VSO over normal shared DEDB areas
3. To observe the performance effects while running with shared SDEPs
4. To assess the impact of OSAM CF Caching on reducing database I/O and improving transaction elapsed time

12.1 Fast Path Data Sharing Enhancements

The following sections provide a technical overview of the new shared VSO and SDEP support.

12.1.1 Shared Virtual Storage Option

Shared VSO is more clearly named block-level sharing of VSO DEDB areas. In the IMS Version 5 Fast Path environment, Virtual Storage Option (VSO) allows a main storage database (MSDB) that was converted to a data entry database (DEDB) to perform as well as an MSDB for a single IMS subsystem (such as IMS/BATCH, DBCTL, or CICS). VSO accomplishes this using MVS data spaces as a local cache storage area for DEDB data. The data is read and written faster in this cache area. Therefore, this feature is used for DEDB data that is highly active (read and updated frequently) and that requires very good performance.

Using IMS Version 5, the DBRC options SHARELVL(2) and SHARELVL(3) allow multiple IMS subsystems to read and update data from any database *except* VSO DEDB data areas (and DEDBs with SDEPS).

This data sharing is called *block-level* sharing because multiple IMS subsystems can read and update the same DEDB. In IMS Version 5, however, block-level sharing of VSO DEDB data is not allowed because IMS subsystems could not share MVS data spaces. Block-level sharing of VSO DEDB areas in IMS Version 6 allows multiple IMS subsystems to concurrently read and update VSO DEDB data. Each VSO DEDB area is represented in the coupling facility by one (or two, for backup) cache structure.

IMS Version 6 provides special private buffer pools for Shared VSO areas. Each pool can be associated with an area, a DBD, or a specific group of areas. These private buffer pools are only used for Shared VSO data. Using these private buffer pools, the customer can request buffer lookaside for the data. The new keywords LKASID or NOLKASID, when specified on the DBRC commands INIT.DBDS or CHANGE.DBDS, indicate whether to use this lookaside capability or not.

12.1.2 Sequential Dependent Sharing

Before IMS Version 6, the sequential dependent (SDEP) segment function provided the user with a time sequential insert capability for a portion of an area that has SDEPs defined. An SDEP for a DEDB is a segment that is chained off the root segment and inserted in a last-in first-out manner into the last part of a DEDB area. The SDEP buffer is kept in main storage and is written by Fast Path synchronization point processing after it is filled to capacity; at the same time, the next CI location is used to start the next SDEP buffer.

The set of SDEP segments is a historical log of events and is presented in time sequence from the oldest to the newest when the segments are removed by the SCAN utility. Utilities provide a way to sequentially remove SDEP segments from an area and to delete a range of inserted segments. The segments cannot be deleted or replaced by an application program. The control structure is part of the local construct for the area and is not written to DASD very often. Therefore, it offers limited, but fast, local access to one IMS subsystem.

With IMS Version 6, the shared SDEP enhancement changes the method of allocating SDEP CIs. Each IMS subsystem image in a sysplex data-sharing environment (IMS partner) now manages its own SDEP CIs. A time stamp is being added to each SDEP segment. This allows SDEPs to be shared without changing time-of-insert processing. The bulk of the existing SDEP support is not being changed by this enhancement. The Fast Path SDEP function has been changed to support a shared database environment which utilizes the current DBRC and IRLM components.

Because of the Shared SDEP enhancement changes, the DEBD SDEP Scan utility JCL must contain sortwork data definition statements to allocate sufficient sort disk space. The sortwork DD statements can be dynamically allocated by the SORT product or by the SYSOUT DD statement. The default for the scan utility is to sort. The user can, however, choose to modify this requirement by using either the new NOSORT or SORTSETUP parameters to tailor their JCL.

12.1.3 Environment Overview

The following describes the workload and the environment for the measurements.

12.1.3.1 IMS Fast Path Workload

A pure fast path workload consisting of a single transaction accessing multiple databases was used in the measurements. This workload had a heavy write characteristic in which a write to the database follows every read.

All IMS systems in the sysplex shared all data. In the conversion to exploit the features of IMS Version 6, one of the DEDBs of the workload was converted to shared VSO. This DEDB was small in size and heavily updated. The result was approximately one-third of the workload DB calls accessing the shared VSO areas. In addition, another of the DEDBs had SDEP segments.

Shared EMH queues were not utilized in this particular study.

12.1.3.2 Hardware Environment

Figure 43 illustrates the configuration used for these tests. The hardware environment consisted of:

- A 9672-RX5 partitioned into multiple LPARs. Up to 2 LPARs were used. Five logical processors per LPAR were used in some of the studies, and two logical processors per LPAR were used in the rest.
- Two 9674-C05s with 2 coupling facility links to each 9674.
- A 9394 RAMAC array subsystem for IMS system datasets.
- IMS databases were on 3390s.

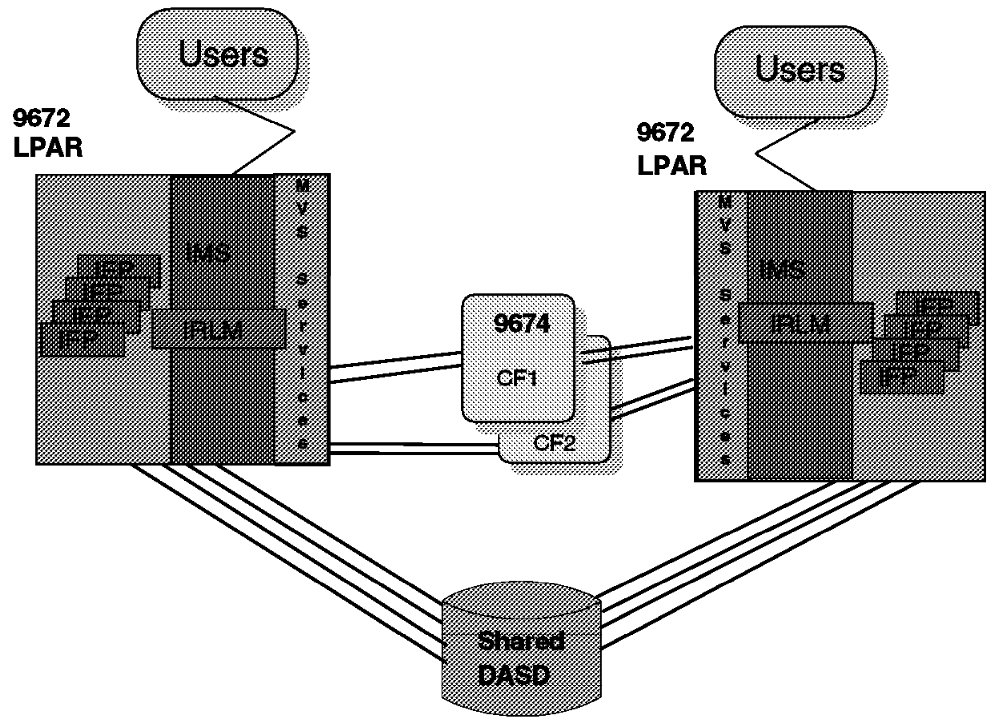


Figure 43. IMS VSO Configuration

12.1.3.3 Structures on the CFs

Table 54 lists the structures on the CFs.

Table 54. Structure Size/Placement for the Configuration

Structure Description	Structure Name	CF	Structure Type	Structure Size
12 VSO areas	WHxxSTR1	CF1	Cache	2MB x 12
IRLM lock	IRLMLOCKTBL1	CF2	Lock	128 MB
XCF message	IXCSTRUC	CF2	List	3 MB

12.1.3.4 Software Environment

The following software was used in this test:

- IMS 6.1
- IRLM 2.1
- OS/390 V1R3 containing:
  - VTAM 4.4
  - JES2 1.3.0

12.1.3.5 Measurement and Reporting Tools

The following measurement and reporting tools were used for this test:

- RMF 1.3.0
- Log Analysis Utility DBFULTA0

12.1.3.6 Methodology

In all the measurements, the DASD configuration and the number of databases was kept constant. This configuration was the same regardless of the number of systems in the sysplex, which led to side-effects of more I/O and locking contention as more systems were added and more total transactions were processed.

A DEDB with SDEP segments was always used in the measurements. Scenarios are not illustrated converting shared DBs with SDEPs since this activity had well-understood data sharing processing characteristics (that is, global locking) in the on-line environment. However, the effects of using shared SDEPs are present in all the VSO measurement variations.

In the first VSO migration scenario, one system was measured with all of the fast path databases defined with global data sharing, except for the VSO areas. The VSO areas were defined with SHARELVL(0). Then, the VSO areas were changed for global sharing as SHARELVL(3). The VSO PRELOAD option was specified for all of the areas. The cost of sharing the VSO areas was analyzed with this change.

The second migration scenario consisted of 1-system and 2-system sysplex measurements in which all of the DEDBs were specified with global sharing. Then, one DEDB was migrated from a shared normal DEDB to a DEDB with shared VSO areas. The benefits of this change were analyzed.

The CI size of the VSO areas was 4 K and one CF structure was defined for each VSO area in the measurements. The LKASID option was used in the shared VSO measurements as LKASID is likely to be more common than NOLKASID.

12.1.4 Shared Virtual Storage Option Results

Table 55 details the VSO data sharing performance cost comparing (non-shared) VSO and shared VSO. These measurements were made with systems with five logical processors.

Table 55. VSO Data Sharing Performance Cost (First Migration Scenario)

	Non-shared VSO	**Shared VSO**
ETR (transactions per second)	180.0	180.1
CPU %	15.69	17.02
ITR (transactions per busy second)	1147	1058
ITR delta	----	-7.8%
CPU milliseconds per tran	4.36	4.73
Transaction transit time (milliseconds)	63	64
Input queue time, processing time, output queue time (milliseconds)	1, 24, 37	2, 24, 38
Total DASD rate	775	776
Coupling facility rates (requests/second)		
WHxxSTR1 total	N/A	182
IRLMLOCKTBL	741	1099
IXCSTRUC	4	0

Table 56 details the VSO data sharing performance benefit when comparing a configuration of shared normal DEDBs to the same configuration containing a DEDB converted to shared VSO areas. The results are from a one system migration environment with two logical processors.

Table 56 (Page 1 of 2). VSO Data Sharing Performance Benefit 1-System (Second Migration Scenario)

	Shared Normal DEDB	**Shared VSO**
ETR (transactions per second)	100.6	100.5
CPU %	27.40	26.59
ITR (transactions per busy second)	367.2	378.0
ITR delta	----	+2.9%
CPU milliseconds per tran	5.45	5.29
Transaction transit time (milliseconds)	79	77
Input queue time, processing time, output queue time (Milliseconds)	(0, 24, 53)	(0, 21, 55)
Total DASD rate	634	434

Table 56 (Page 2 of 2). VSO Data Sharing Performance Benefit 1-System (Second Migration Scenario)

	Shared Normal DEDB	**Shared VSO**
Coupling facility rates (requests/second)		
WHxxSTR1 total	N/A	102
IRLMLOCKTBL	604	606
IXCSTRUC	4	4
Note: Sysplex ITR is the total ITR of all IMS partners participating in data sharing in the sysplex.		

Table 57 details the VSO data sharing performance benefit comparing a configuration of shared normal DEDBs to the same configuration containing a DEDB converted to shared VSO areas. The results are from a two system sysplex environment with two logical processors defined for each system.

Table 57. VSO Data Sharing Performance Benefit 2-Systems (Second Migration Scenario)

	Shared Normal DEDB	**Shared VSO**
ETR (transactions per second)	201.8	202.1
CPU %	29.41	29.29
ITR (transactions per busy second)	686.0	690.1
ITR delta	----	+0.6%
CPU milliseconds per tran	5.83	5.80
Transaction transit time (milliseconds)	82	80
Input queue time, processing time, output queue time (Milliseconds)	(0, 27, 54)	(0, 23, 55)
Total DASD rate	1268	874
Coupling facility rates (requests/second)		
WHxxSTR1 total	N/A	317
IRLMLOCKTBL	1206	1206
IXCSTRUC	229	227
Note: Sysplex ITR is the total ITR of all IMS partners participating in data sharing in the sysplex.		

12.1.5 Observations/Conclusions

In the first migration scenario, converting from non-shared VSO areas to shared VSO areas showed 7.8% measurement cost. The system cost of using shared VSO areas is comprised of the following:

1. Additional global locking was incurred to maintain the integrity of the shared VSO data. This cost results when defining the databases with SHARELVL(3), as was done in this migration scenario, even though only one system was currently accessing the data.
2. Maintaining local buffer pool coherency with data update activity.

3. Transferring data which previously was always kept in data spaces and now exists in the cache structure in the coupling facility. To minimize this cost, local VSO look-aside buffers were used to avoid read accesses to the coupling facility. (Not using the look-aside buffers resulted in 2-4% additional ITR cost as measured in other tests.)

The first two factors, global locking and buffer coherency, are the processing requirements of all Parallel Sysplex data sharing implementations (that is, DB2, DL/I). The third cost factor is a unique component for this environment because in the base VSO case, the data always resided in fast processor storage and was not accessed from DASD. Now, the data needs to be maintained in coupling facility cache structures in order to be shared across systems.

In the measurement, requests to the IRLM lock structure grew from 741 to 1099 per second to support the global locking of VSO data. The activity to maintain the local buffer coherency and manage the global data in the coupling facility is represented by the accesses to the cache structures WHxxSTR1. Reading of information in the cache structure usually results in a synchronous call to the coupling facility, and the write activity results in an asynchronous coupling facility call.

As in all the data sharing implementations, the data sharing cost for the VSO workload is a function of the access frequency to the lock and cache structures. The resulting total relative overhead relates to the data access intensity to the shared VSO data. In this measurement environment, the shared VSO content of the workload and the resulting costs were pronounced to understand the effects of this support.

In the second migration scenario, converting a shared DEDB to a shared DEDB with VSO areas, some processing benefits were observed. The use of shared VSO areas was slightly more efficient in CPU time and response time. DASD I/O to access data was replaced with the following:

- References to the local VSO look-aside buffers to access data and the processing to maintain buffer coherency.
- Faster coupling facility calls to read data, if not in the local buffers, and to write data. Reading of VSO data in the cache structure usually results in a synchronous call to the coupling facility and the write activity results in an asynchronous coupling facility call.

The locking cost in the second migration scenario was equivalent since the data was always globally shared.

12.2 OSAM CF Caching

The OSAM Database Coupling Facility Caching enhancement is an extension of sysplex data sharing introduced in IMS Version 5. This function allows OSAM database buffers to be read from or written to a coupling facility cache structure.

The writing of data to the coupling facility cache structure is enabled by a caching option that allows the caching of all data, or the caching of changed data only. *Cache all data* means that if an application program requests data that is not already in a subpool or in the coupling facility, that data is either read from DASD or copied from the SB buffer pool into the subpool, and then written to the coupling facility. This option includes writing changed data to the coupling facility. *Cache changed data only* means that if an application program modifies data in a subpool, that data is first written to DASD and then written to the coupling facility.

Data is always backed up on DASD so that data integrity will not be lost by coupling facility failures. The caching option is specified on a subpool basis. Users can associate each subpool with a caching option and can assign databases or database data sets to specific subpools defined with or without a caching option.

12.2.1 Environment Overview

The workload, hardware configuration, software configuration, and measurement and reporting tools used are the same as the environment documented in 11.1.1, "Environment Overview" on page 128.

12.2.1.1 Structures on the CF

Table 58 lists the structures on the CFs.

Table 58. Structure Size/Placement for the Configuration

Structure Description	Structure Name	CF	Structure Type	Structure Size
XCF messaging	IXCSTRUC	CF1	List	3 MB
IRLM lock	IRLMLOCKTBL1	CF1	Lock	128 MB
IMS OSAM	OSAMSESXI1	CF1	Cache	720 MB
IMS VSAM	VSAMSESXI1	CF1	Cache	24 MB

12.2.2 Methodology

For the database measurement, the CUSTOMER and PARTS databases (out of 14 databases) were cached. One-third of the transactions in the transaction mix were affected. The key ranges were narrowed to the first 10% in randomization to force referencing. The number of updates was increased by lowering the weights of the readonly transactions in the mix.

The systems were targeted to less than 40% CEC busy to insure that no hardware bottlenecks existed. TPNS used transaction drivers on another separate CEC to insure that the TPNS activity was isolated from the measurements.

The OSAM caching option of caching changed data only is illustrated in Table 59 on page 145 as this is expected to be the typical use of OSAM caching. The data cached in the CF had 4 K blocksizes, so all data was accessed synchronously from the CF.

12.2.3 OSAM CF Caching Results

Table 59 on page 145 shows OSAM CF Caching results with cached DASD.

Table 59. OSAM Caching, 2-System Sysplex

	No CF Data Caching	**Cache Changed Data**
ETR (transactions per second)	53.1	55.8
CPU %	37.0	38.6
ITR	143.5	144.6
ITR delta		+.8%
Transaction elapsed time in milliseconds	153	127
Elapsed time delta		-17%
Database I/Os per transaction	16.3	15.1
OSAM structure synchronous access time (microseconds)	176	244
VSAM structure synchronous access time	156	155
Lock structure synchronous access time	149	147
Coupling facility rates (requests/second)		
OSAM structure	637	675
VSAM structure	244	262
lock structure	950	999
IXCSTRUC	112	112

12.2.4 Observations/Conclusions

Where there is re-referencing, OSAM CF caching can reduce database DASD I/O. Significant reductions in database DASD I/O have been measured. As a result, reduced transaction elapsed time will be a benefit of OSAM CF caching.

There are some minimal costs associated with OSAM CF caching (if caching changed data). Caching data can elongate the OSAM structure access time due to an increase in data transfer and busier links. This cost of the additional data transfer to the coupling facility is recovered by the referencing of this changed data later on by the multiple systems of the sysplex.

Specifying caching of all data has additional costs of extra writes of read data from DASD to the coupling facility. This cost of the additional data transfer and additional CF writes to the coupling facility can only be recovered if the re-referencing frequency of the data from the CF cache (beyond finding the data in the local OSAM buffers) warrants the use of this type of caching.

Chapter 13. XCF Signalling Study

Communication between members of the sysplex is done by the cross system coupling facility (XCF) component of MVS. It provides the standard communication mechanism for MVS system applications. XCF can use different methods to accomplish this:

- Each member system of the sysplex can be connected to every other member with CTC links in both directions. As the number of systems in the sysplex grows, this becomes more difficult to manage.
- A shared intermediate memory scheme can be used to communicate between systems in the sysplex. Messages from one system are sent through a central control to any other system, thus creating a "star" of communication paths. The hub of the star is the coupling facility.

This study compared the performance characteristics of XCF signalling using the coupling facility to XCF signalling using CTC connections. The first experiment was done in a typical environment where XCF signalling was one of many functions occurring. The second experiment was done in an "XCF driver" environment where XCF activity was the only significant function being exercised.

Figure 44 on page 148 illustrates the configuration used in the test.

13.1 OLTP Environment

This environment was the fully utilized data sharing workload. XCF was primarily used for GRS ring processing, CP/SM, and transaction routing.

13.1.1 Hardware Environment

Table 60 lists the hardware resources used in this test.

Table 60. Hardware Resources — XCF/MRO Tests

Processor Model	**4x9672-R61**	**9674**
# of CECs	4	1
# of processors/CEC	6	6
Central storage/CEC (installed)	512 MB	2048 MB[37]
Expanded storage/CEC (installed)	0 MB	0 MB
# of CF sender links/CEC to CF	1	N/A
# of CF receiver links on CF	N/A	16

The 9674s were at MEC level D57264.

[37] See Table 61 on page 149 for actual storage usage.

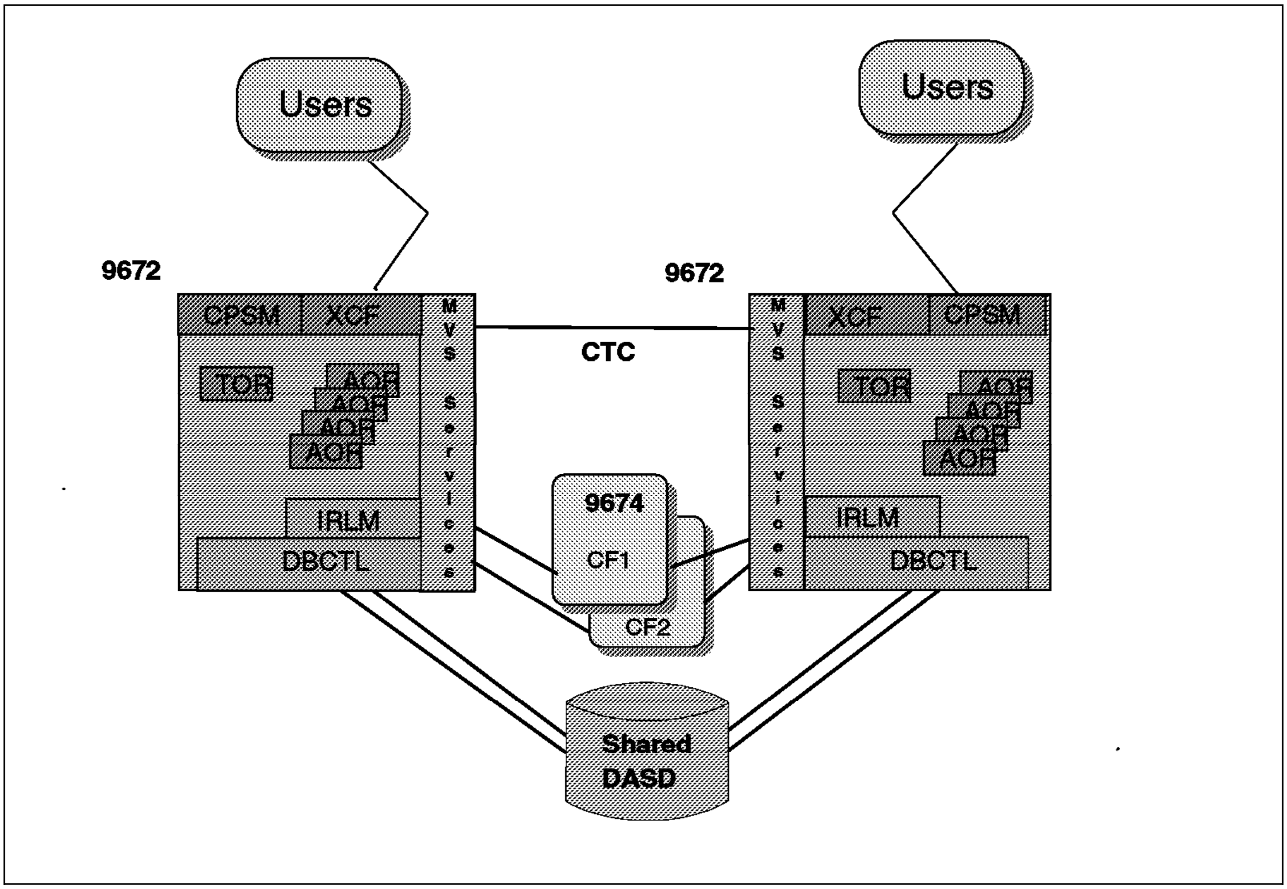


Figure 44. Environment for the XCF Test Cases

13.1.2 Software Environment

The following software products were installed:

- MVS 5.1.0 at Service Level 9406
- CICS 4.1.0
- CP/SM 1.1.1 October driver
- IMS 5.1.0 at the PI level
- IRLM 5.1.0 at the PI level
- RMF 5.1.0

On each processor, we defined one CICS TOR and four AORs. We used the CP/SM QUEUE algorithm.

13.1.2.1 Structures on the CF

The following structures were defined on the coupling facility as shown in Table 61 on page 149.

Table 61. Structure Size/Placement for 4x9672-R61 Configuration

Structure Description	Structure Name	CF	Structure Type	Structure Size
XCF signalling	IXCSIG1	CF1	List	69 MB
XCF signalling	IXCSIG2	CF1	List	69 MB
IRLM lock table	IRLMLOCKTBL2	CF1	Lock	64 MB
JES2 checkpoint	JES2CKPT	CF1	List	13 MB
OSAM buffer invalidation	OSAMSESXI	CF1	Cache	64 MB
VSAM buffer invalidation	VSAMSESXI	CF1	Cache	64 MB
Total CF storage used[38]:				343 MB

13.1.3 Measurement and Reporting Tools

The data shown was taken from the RMF Monitor I Workload Activity report and the RMF Monitor III Coupling Facility Reports.

13.1.4 Workload

The workload used was the CICS/DBCTL data sharing workload described in Chapter 4, "CICS/DBCTL Data Sharing Scalability Study" on page 65.

13.1.5 Description of Experiment

Theoretically, a direct connection should be faster than one that involves an intermediate step. So we would expect the CTC connection to be somewhat faster than a connection through the coupling facility. However, there is a trade-off between the speed of the direct connection and the simplified management of the star configuration using the coupling facility. We wanted to determine if using the coupling facility instead of CTC connections would have a noticeable effect on the final transaction response time in a fully loaded system.

We brought up a 4x9672-R61 Parallel Sysplex using CTCs for XCF signalling. We activated enough transactions to make each processor over 85% busy. We collected data for a 15 minute interval.

We repeated the experiment replacing the CTC connections with XCF structures on the coupling facility. Two XCF structures were defined for redundancy.

13.2 OLTP Results

Table 62 on page 150 shows the results from the OLTP XCF test.

[38] Although the 9674 coupling facility was configured with 2048 MB of central storage, only 343 MB of the installed CF storage was used for these configurations.

13.2.1 Data

Table 62. XCF Via CF versus XCF Via CTC OLTP Study

	Using CTC	Using CF
	4x9672-R61	4x9672-R61
Transactions/sec (ETR)	324	324
CPU activity	84.7%	85.8%
Transactions/sec at 100% (ITR)	382.8	377.7
TOR response time	.209 secs	.211 secs
CF utilization	N/A	9.4%
Req/sec IXCSIG1	N/A	155
Req/sec IXCSIG2	N/A	145

Response time is the internal response time at the TOR level.

13.2.2 Observations/Conclusions

The internal response time at the TOR level is equivalent in both cases. Although sending an XCF message through the coupling facility may have taken longer than over a CTC (see study 13.3, “XCF Driver Environment”), the frequency of XCF messages per transaction was small in this environment, thus no change in transaction response time was noted.

13.3 XCF Driver Environment

A batch job was used to generate XCF signals to compare the elapsed time to send a message between systems via CTCs versus the coupling facility in a heavily-loaded signalling environment.

13.3.1 Hardware Environment

Table 63 lists the hardware resources used in the test.

Table 63. Hardware Resources — XCF Driver Test

Processor Model	**9021-711**[39]	**9674**
# of CECs	3	1
# of processors/CEC	1	6
Central Storage/CEC (Installed)	256 MB	2048 MB[40]
Expanded Storage/CEC (Installed)	512 MB	0 MB
# of CF Sender Links/CEC to CF	1	N/A
# of CF Receiver Links on CF	N/A	16

The 9674s were at MEC level D57264.

[39] 9021-711s created by logically partitioning a 9021-982.

[40] See Table 64 on page 151 for actual storage usage.

13.3.2 Software Environment and Coupling Facility Structures

This test was done on MVS 5.1.0 at Service Level 9406. Table 64 shows the structure used in this test.

Table 64. Structure Size/Placement for 9021-711 Configuration

Structure Description	Structure Name	CF	Structure Type	Structure Size
XCF signalling	IXCSIG1	CF1	List	10 MB
Total CF storage used:[41]				10 MB

13.3.3 Measurement and Reporting Tools

The following measurement tools were used in this test.

- RMF
- SYSLOG

13.3.4 Workload

A batch job was created to send 100,000 messages from system "X" to system "Y" via XCF. The job would not send message n+1 until an acknowledgement was received for message n. Therefore, the job would actually cause 100,000 messages to be sent from X to Y, and 100,000 acknowledgement messages to be sent from Y to X.

13.3.5 Experiment Description

The three 9021-711s were configured as systems A, B, and C in a Parallel Sysplex. Three copies of the batch job were run on systems A and C and all used system B as the "target" system. All jobs were started in parallel, and the elapsed time for all to complete their 100,000 messages sent was measured. In total, 1,200,000 messages were sent among the systems in the sysplex (six jobs times 100,000 messages sent from X to Y and 100,000 acknowledgement messages sent from Y to X). However, since all six jobs were sending their messages in parallel, the elapsed time of a measurement actually reflects the time to send 100,000 and receive 100,000 messages in a heavily loaded environment.

Two measurements were made in each of the XCF via CTC and XCF via CF configurations. The first used a message size of 1 KB and the second used a message size of 5 KB.

13.4 XCF Driver Results

Table 65 on page 152 shows the results from the XCF driver tests.

[41] Although the 9674 coupling facility was configured with 2048 MB of central storage, only 10 MB (0.5%), of the installed CF storage was used for these configurations.

13.4.1 Data

Table 65. XCF Via CF versus XCF Via CTC Study

	Using CTC	**Using CTC**	**Using CF**	**Using CF**
Message size	1 KB	5 KB	1 KB	5 KB
Elapsed time	309 sec	568 sec	654 sec	674 sec
Average message delivery time	1.5 ms	2.8 ms	3.3 ms	3.4 ms

Average message delivery time is calculated by taking the elapsed time and dividing by 200,000 messages.

13.4.2 Observations/Conclusions

In this heavily-loaded XCF environment, message delivery time varied from 1.5 ms to 2.8 ms using CTCs, and from 3.3 ms to 3.4 ms using the coupling facility. Thus, messages were delivered somewhat faster when using CTCs, although the delivery time through the coupling facility was less sensitive to message size.

To send a message from X to Y via the coupling facility, system X issues an asynchronous write of the message to the XCF list structure on the coupling facility; then system Y issues an asynchronous read of the message from the list structure. Thus, each XCF message results in two accesses to the coupling facility. These accesses, and the delays in MVS scheduling these accesses, are the major contributors to the observed message delivery times.

As improvements are made to the speed of the coupling facility, message delivery times will improve accordingly.

Chapter 14. JES2 Checkpoint Performance Study

The purpose of this study is to examine the performance characteristics of JES2 checkpoint processing in a Parallel Sysplex environment.

Specifically the performance characteristics of the checkpoint data set placed on DASD versus the coupling facility (CF) are compared in three separate Multi-Access Spool (MAS) environments:

- A 3-way MAS on a 9021 in LPAR mode
- An asymmetric MAS that included a 9021 and two 9672-R61s systems
- An 8-way MAS consisting of eight 9672-R61 systems

The following is a detailed description of the environments and methodology used and conclusions derived from this study. The measurements described in these sections should not be used to compare anything other than the effect of using a coupling facility for the checkpoint versus using DASD.

For a detailed discussion of JES2 in a Parallel Sysplex, including installation, customizing and performance, see the Washington System Center Technical Bulletin, *JES2 Multi-Access Spool in a Sysplex Environment*.

14.1 Environment

The 9021 Model 982 used in the 3-way MAS experiment was physically partitioned to create a Model 941. Then with the use of the PR/SM LPAR feature, three logical partitions were created on this 941, each with one dedicated CP. The 3-way MAS created using these three dedicated logical partitions was equivalent to one made with three 9021-711 systems.

The asymmetric environment was composed of one 9021 and two 9672s running in a 3-way MAS. The 9021 Model 942 used in this experiment was physically partitioned to create a Model 821. The 9021 was joined in the MAS by two 9672-R61s. This configuration was chosen because the computing power of the 9021 roughly matches that of the combined 9672s.

The 8-way environment was composed of eight 9672s. Each system in this sysplex was identical in its configuration.

14.1.1 Hardware Resources

Table 66, Table 67 on page 154, and Table 68 on page 154 list the hardware resources we used in these tests.

Table 66. Hardware Resources — 3-Way MAS

Processor Model	**9021-711**
# of processors/CEC	1
Central storage/CEC	512 MB
Expanded storage/CEC	1024 MB
# of CF sender links/CEC to CF	1
# of SPOOL DASD	32

Table 67. Hardware Resources — Asymmetric MAS

Processor Model	**9021-921**	**9072-R61**
# of processors/CEC	2	6
Central storage/CEC	512 MB	512 MB
Expanded storage/CEC	1024 MB	0 MB
# of CF sender links/CEC to CF	1	1
# of SPOOL DASD	32	32

Table 68. Hardware Resources — 8-Way MAS

Processor Model	**9672-R61**
# of processors/CEC	6
Central storage/CEC	512 MB
Expanded storage/CEC	0 MB
# of CF sender links/CEC to CF	1
# of SPOOL DASD	32

The 9674s were at MEC level D57264.

14.1.2 Software Levels

The software used in the test was at the following levels:

- MVS 5.1 at Service Level 9406
- JES2 5.1.0

14.1.3 Coupling Facility Exploitation

The measurements that included a coupling facility were taken with the use of one 9674.

Table 69. Structure Size/Placement for All Configurations

Structure Description	**Structure Name**	**CF**	**Structure Type**	**Structure Size**
JES2 checkpoint	COUPLE_CKPT1	CF1	List	13 MB
Total CF storage used:				13 MB

Note: Although the 9674 coupling facility was configured with 2048 MB of central storage, only 13 MB (0.7%) of the installed CF storage was used for these configurations. For the JES2 checkpoint on DASD measurements, the size of the checkpoint data set was 20 cylinders.

14.1.4 Measurement and Reporting Tools

The following tools were used for measurement of this test.

- RMF 5.1

  Data collected included standard system performance indicators such as CPU, and device utilization.

- JES2 Trace 17 Records and reduction program (J2TR17A in SYS1.SAMPLIB).

14.1.5 Workload Description

A JES2 checkpoint intensive workload was used for this study. The objective of the JES2 checkpoint intensive workload was to create an environment in which there was a very high demand for JES2 services.

This workload consisted of the following types of work, which generated a heavy demand for checkpoint processing:

- TSO SUBMIT, STATUS and CANCEL commands
- JES2 commands to change the writer and initiator classes (which caused JES2 to check every job it is currently managing).
- A print job (IEBGENER)
- Several NJE jobs that entered the MAS on one system and executed on another system

The number of jobs run in an experiment varied depending on the number and types of the systems in the MAS and the highest CPU busy percent that was achievable.

The SYSAFF parameter on the JES2 JOBPARM statement was used to ensure the conversion and execution of the jobs were done on a specific system. SYSAFF=* was specified on all the jobs, which indicates the system that reads the job will complete these phases.

14.1.6 Description of Experiments

The following describes the tests performed.

14.1.6.1 Methodology

A start-to-finish methodology was selected to determine the impact on performance when only the checkpoint data set hardware was changed from DASD to CF.

For each measurement, the following steps were followed for each system in the MAS:

1. Start the measurement tools (RMF and TRACE 17).
2. Start the workload via JES2 internal readers.
3. Stop the measurement tools when the measurement completes (that is, when the last workload execution job completes).

All output from the jobs was targeted to an output class defined to cause the output to be automatically purged. This was done to alleviate JES2 resource shortage problems (such as JOES, JNUM, and so forth), since the experiments were not set up to print large amounts of output.

Two measurements were made (with checkpoint on DASD, and then with checkpoint on CF) for each of the three MAS environments (3-way MAS, asymmetric MAS and 8-way MAS).

The JES2 checkpoint DUPLEX mode configuration was used in each of the measurements. Each DASD checkpoint data set (primary or secondary) was placed on a dedicated 3390 DASD device that was attached to a dedicated 3990-3 control unit with DASD Fast Write (DFW) enabled.

Note: It is recommended that you use MODE=DUAL with both checkpoints on DASD. However, these DASD experiments used MODE=DUPLEX to eliminate another variable when comparing DASD and CF measurements, since the mode must be DUPLEX if a checkpoint data set resides on a coupling facility structure.

Runs with the coupling facility had CKPT1 on the CF and CKPT2 on DASD. CKPT1 is the primary data set copy of the JES2 checkpoint information, and CKPT2 is used for the DUPLEX (backup) data set copy.

All DASD (spool and checkpoint) used DASD Fast Write (DFW).

XCF signalling for all measurements was done using CTCs. The objective was to isolate the effects of having the checkpoint data set on the coupling facility.

The asymmetric experiment included two sets of measurements: a controlled environment (in which each member was allowed limited and regulated checkpoint data set access) and a contention-driven environment (in which the members competed for checkpoint data set control).

14.1.6.2 Metrics

The metrics used in the analysis included:

- CPU busy percent from RMF

  This refers to the percentage of time the processor was busy; that is, not in a wait state.

- External throughput rate (ETR) measured

  This refers to the number of jobs that complete per elapsed second.

- Elapsed time

  This refers to the amount of time, in seconds, that it took the members of the MAS to complete executing the workload.

- JES2 cycle time

  This refers to the total time it took for each processor in the MAS to obtain ownership of the JES2 checkpoint data set, hold it, release control, and obtain ownership again.

  A target cycle time in the range of 2 to 3 seconds was used to provide somewhat consistent internal throughput across all runs.

14.1.6.3 Methodology Tuning

The JES2 Trace 17 data contains information that was used to tune the HOLD and DORMANCY values. Trace 17 data reports the amount of time required to complete the different I/Os within the checkpoint cycle. The useful and non-useful time could be calculated based on the Trace 17 data collected. The non-useful time was calculated by adding READ1 + READ2 + FINAL WRITE time.

For the 3-way MAS runs (both LPAR and asymmetric), the hold and dormancy values used to calculate cycle time where derived using the following formulas:

$$H_1 = \frac{Ct - (NU_1 + NU_2 + NU_3)}{W_1 + W_2 + W_3} \times W_1$$

$D_1 = (H_2 + NU_2) + (H_3 + NU_3) + (0.5 \times NU_1)$

Where:

$Ct \equiv$ target checkpoint cycle time
$NU_n \equiv$ non-useful time for processor n
$W_n \equiv$ weight factor for processor n
$H_n \equiv$ HOLD value for processor n
$D_n \equiv$ DORMANCY value for processor n

Refer to the "Checkpoint Data Set Definition and Configuration" section of the *MVS/ESA SP V5 JES2 Initialization and Tuning Guide* for a discussion on the checkpoint cycle.

When calculating the HOLD value for the asymmetric environment, a weight factor of two was used for the 9021, as compared to a weight factor of one for the 9672s. This was done to reflect the relative processing power of the two processors used in this environment and the amount of JES2 work going on in the system; in this case the ratios are about the same.

When calculating the HOLD value in the 8-way environment, an initial HOLD value of 10 was selected. This value was derived by starting with a hold value from a smaller tuned environment, and then factoring in the number of systems in the MAS. This method was used to get a first approximation instead of using the above formula because, in our environment, the formula for calculating HOLD generates a negative number. Then the DORMANCY value was adjusted until the desired cycle time was obtained.

Due to resource constraints, once the desired cycle time was achieved for both DASD and CF measurements, the experiments were completed. Ideally, the checkpoint cycle time should have been further tuned to minimize checkpoint idle time.

In the 3-way MAS experiment, the total number of jobs was held constant at 16842, and through the use of JOB SYSAFF, each member executed 5614 jobs. This resulted in elapsed times over 10 minutes and a CPU busy percent over 60.

In the asymmetric MAS JES2 checkpoint experiment, the total number of jobs was held constant at 8022 (4011 per 9672). The number of jobs run on the 9021 was kept constant at 8021. These numbers also reflect the relative processing power of the 9021 over each of the 9672s. Another factor that was involved in deriving these numbers was elapsed time. It turns out that increasing the total number of jobs would have increased the amount of time it took the sysplex to run the jobs. The values selected resulted in elapsed times over 15 minutes and CPU busy percent values over 20.

The initial concern with the asymmetric experiment was that the larger/faster processors would lockout the smaller/slower processors, thus decreasing the overall sysplex-wide throughput. The strategy was to create a realistic environment that would highlight any advantage the 9021 might have in gaining access to the JES2 checkpoint.

In the 8-way MAS JES2 checkpoint experiment, the total number of jobs was held constant at 16048 (2006 per 9672). The number of jobs run was based on CPU busy percent and elapsed time. It turns out that increasing the total number of jobs running on the system did not greatly increase the processors CPU busy percent; however, it did drastically increase the amount of time it took the sysplex to run the jobs. These values resulted in elapsed times over 25 minutes and CPU busy percent values over 10.

14.2 Results

The measurement data in Table 70, Table 71, and Table 72 indicates that as the checkpoint is migrated to the coupling facility, there are minor increases in CPU busy percent and decreases in elapsed seconds to complete the workload resulting in a slight increase in External Throughput Rate (ETR). This increase in ETR, with this particular workload, was noticed for all measurements attempted.

The sysplex-wide matrix consists of CPU busy and ETR values that are representative of all the members of the sysplex. For example, the ETR value is calculated by adding together the number of batch jobs from SY#A, SY#B and SY#I, and then dividing the result by the workload elapsed seconds from each of these systems. The elapsed seconds value is the time that expired until the last job in the sysplex completed. The number of jobs is an aggregate value for the sysplex.

Table 70. JES2 3-Way LPAR Checkpoint Study — Sysplex-Wide Matrix

	Checkpoint on DASD	**Checkpoint on Coupling Facility**	**% Delta**
CPU busy	61.84	67.72	9.5%
Elapsed seconds	743	714	-3.9%
ETR	7.480	7.923	5.92%
Num. of Jobs	16815	16815	0

Table 71. JES2 Asymmetric Checkpoint Study — Sysplex-Wide Matrix

	Checkpoint on DASD	**Checkpoint on Coupling Facility**	**% Delta**
CPU busy	23.93	24.83	3.76%
Elapsed seconds	1744	1711	-1.89%
ETR	3.424	3.510	2.51%
Num. of Jobs	16043	16043	0

Table 72. JES2 8-Way Checkpoint Study — Sysplex-Wide Matrix

	Checkpoint on DASD	**Checkpoint on Coupling Facility**	**% Delta**
CPU busy	10.39	11.07	6.54%
Elapsed seconds	1857	1756	-5.43%
ETR	1.097	1.160	5.74%
Num. of Jobs	16048	16048	0

14.2.1.1 Asymmetric Observations — Round-Robin versus Contention Mode

The asymmetric measurement data in Table 73 and Table 74 show the cycle times to be nearly equivalent for all the systems in the sysplex, regardless of machine type. This is true for both the DASD and CF measurements.

Table 73. JES2 Asymmetric Checkpoint Study — Checkpoint on DASD

System	**SY#A (9672)**	**SY#B (9672)**	**SY#I (9021)**
Cycle time	2576	2576	2613
Average held	579	580	327
Initial hold	400	400	200
Average not held	1996	1993	2286
Initial dormancy	1800	1800	2200
Elapsed seconds	1737	1744	1203

Table 74. JES2 Asymmetric Checkpoint Study — Checkpoint on CF

System	**SY#A (9672)**	**SY#B (9672)**	**SY#I (9021)**
Cycle time	2568	2568	2619
Average held	587	593	353
Initial hold	400	400	200
Average not held	1982	1976	2259
Initial dormancy	1800	1800	2200
Elapsed seconds	1711	1706	1152

The asymmetric measurements were attempted again to see what effect a contention-driven environment would have on access to the JES2 checkpoint data set.

The cycle and average held times in Table 75 and Table 76 on page 160, as with the previous controlled environment asymmetric measurement, indicate that no one system dominated the access of the JES2 checkpoint data set.

Table 75. JES2 Asymmetric Checkpoint Study — Checkpoint on DASD

System	**SY#A (9672)**	**SY#B (9672)**	**SY#I (9021)**
Cycle time	844	1036	672
Average held	338	333	65
Initial hold	0	0	0
Average not held	504	705	608
Initial dormancy	0	0	0
Elapsed seconds	771	974	958

Table 76. JES2 Asymmetric Checkpoint Study — Checkpoint on CF

System	SY#A (9672)	SY#B (9672)	SY#I (9021)
Cycle time	860	864	653
Average held	312	312	63
Initial hold	0	0	0
Average not held	546	550	590
Initial dormancy	0	0	0
Elapsed seconds	804	805	929

14.2.2 Observations/Conclusions

Performance of the JES2 checkpoint on the coupling facility is equivalent to the performance of the checkpoint on the DASD.

The Trace 17 data, in Table 75 on page 159 and Table 76, demonstrates that JES2′s use of the coupling facility provides for a more equitable sharing of the checkpoint among all members. This is accomplished by JES2′s use of the first-in, first-out (FIFO) method of queuing accesses to the checkpoint. This was observed in both symmetric and asymmetric environments.

Some increases in JES2 SRB time were noticed. This was attributed to XES (MVS cross-system extended services) processing in support of the checkpoint on the coupling facility. Note this increase was observed in a laboratory environment using a JES2 intensive workload that stresses JES2 processing at a much higher level than the average JES2 workload does. Therefore, the impact in a given customer environment will vary depending on the amount of JES2 processing (as measured in JES2 CPU busy).

Chapter 15. CICS Temporary Storage Performance Study

Temporary Storage Data Sharing (TSDS) is a new feature introduced in CICS TS that uses the Coupling Facility (CF) as a repository for Temporary Storage queues (TSQs), and by this means, allows sharing of TSQs across multiple CPCs and MVS images within a sysplex.

This will be of considerable benefit and interest to CICS customers who plan to migrate toward a parallel processing environment, since it removes another of the sources of “transaction affinity” which currently act as a barrier toward an easy migration of existing customer workloads.

15.1.1 Objective

The intent of this study was to measure the performance of temporary storage data sharing (TSDS) with CICS TS. The performance is compared to a configuration where TS queue requests are Function Shipped (FS), via a Multi-Region Operation (MRO) connection to a TS queue-owning region (TSQOR), this being the currently recommended method of sharing TS queues within a single MVS image.

15.1.2 Method

15.1.2.1 Comparison Configurations

The base configuration for comparison was a TOR/AOR which had its TS table defined such that TS operations were Function Shipped via MRO (cross-memory) to a TSQOR. The TS queues in the TSQOR were defined in main storage, and TS was cold started for each run.

The TSDS configuration consisted of a TOR/AOR accessing shareable TS queues located in a shared TS pool (a list structure) defined in the Coupling Facility. The shared TS pool was accessed via the required TS Server address space associated with the specific pool.

Access to the shared TS queues was via the required TS Server address space, running in the same MVS image.

An important factor in the performance of the TSDS function is the availability of queue index buffers in this TS server address space.

TS queues having an overall length less than 32 KB can be contained completely within the queue index entry of a list structure, and the design for the TS server provides the ability to store entire queue index entries in 32 KB buffers defined in the server address space, thus reducing the numbers of accesses to and from coupling facility storage. Such buffers are controlled by a Least Recently Used (LRU) algorithm which will discard the oldest buffer if a new buffer is required and no free buffer storage is available.

The testing described was conducted with a more than adequate number of buffers. (This was confirmed by examination of the TS Server statistics during the evaluation).

15.1.2.2 Temporary Storage Workload

A new artificial workload designed specifically for TS performance evaluation was created for this test.

A transaction was written to perform the following logic:

1. The transaction attempts to read a queue with a specific name.
2. If the queue does not exist, the transaction creates it by writing ten equal length records, and then returns control to CICS (the transaction ends).
3. If the queue does exist on the initial read, the transaction performs a further nine read operations, deletes the queue, and then returns control to CICS.

- Iterations of the transactions for a particular queue name therefore run in pairs. The first iteration creates the queue by writing it, having established that it doesn′t exist by receiving a queue ID error on the initial read, and the second iteration reads and deletes the queue. Queues remain in existence for an amount of time randomly determined by the TPNS network operation parameters.
- TS queue names are of the form TSD0xxxx, where the xxxx value is taken from the TPNS simulated terminal ID.

  One hundred simulated terminals are used on the TPNS network, so a total of 100 possible queue names will be created and subsequently deleted.
- The record length is hard coded at 150 bytes; the total queue length is therefore 1.5 KB.
- All queues are non-recoverable (TSDS does not currently support recoverable queues) and those defined in the TSOR for the MROFS comparison configuration are defined to be in main storage.

This workload was not intended to fully simulate a customer application, but to provide a clear comparison of the relative costs of the two TS access methods, by concentrating solely on them.

15.1.2.3 Test Environment

Tests were performed on a 9672-R21, together with a 9674-C01 coupling facility with two CPs.

CICS TS was used for both configurations tested, the operating system was MVS 5.2 in both cases, and all other system and subsystem software was identical for the two sets of measurements.

Care was taken to ensure that:

- The test configurations were defined to be as similar as possible (for example, by using similar CICS system initialization options where possible).
- Test configurations were not constrained by storage or I/O.

Results from the throughput runs were gathered using the Monitor 1 component of RMF Version 5.

15.1.3 Results and Conclusions

The following reports two sets of results.

- Figure 45 is a graph showing a comparison of CPU percentage versus External Throughput Rate for the two configurations compared. The CPU percentages are based on the two engines of the 9672-R21 processor used.
- Figure 46 on page 164 is a graph showing average transaction response times versus ETR.

15.1.3.1 Workload Results

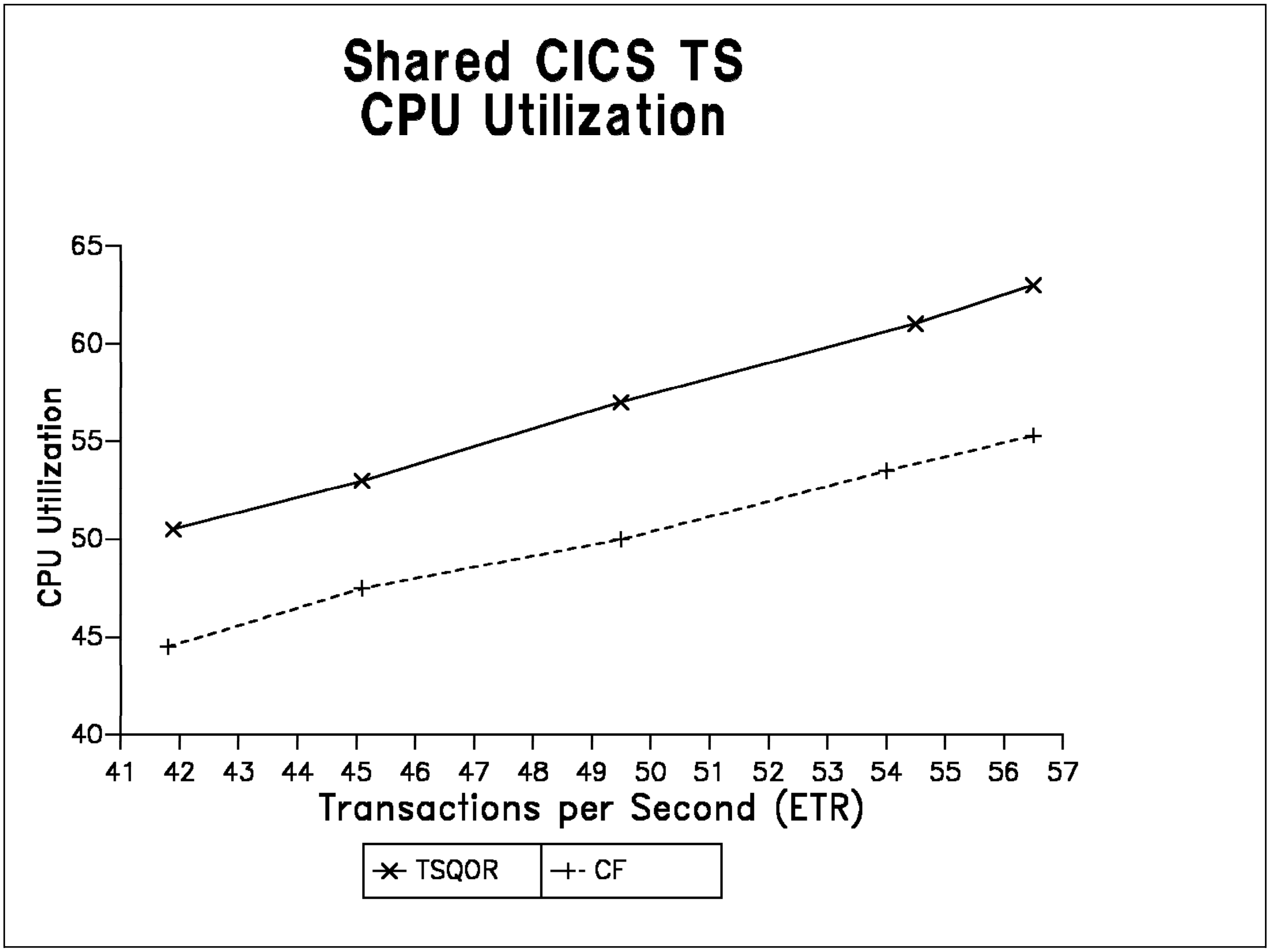


Figure 45. Comparative CPU% versus ETR Results

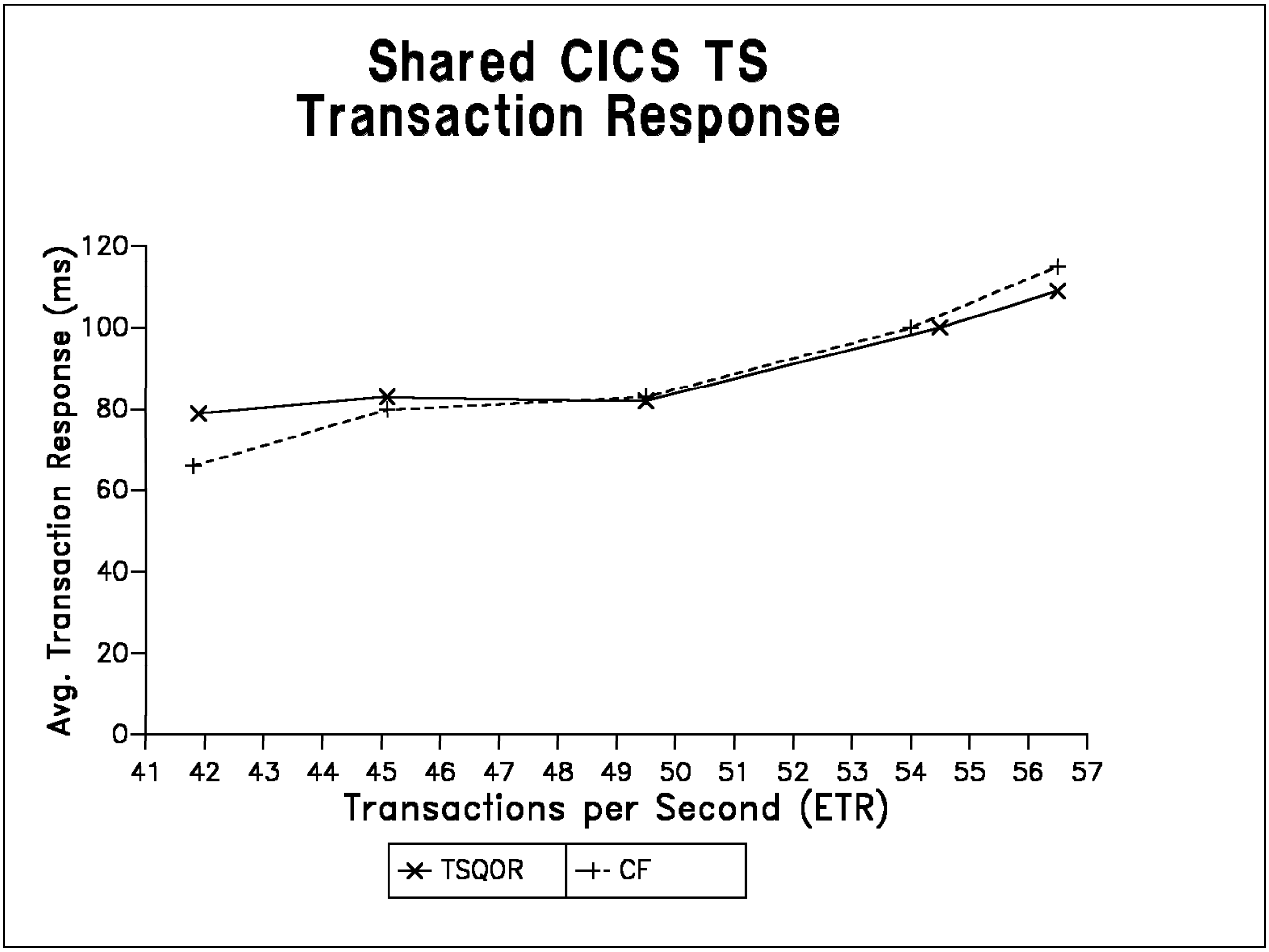


Figure 46. Comparative Response Time Results

15.1.3.2 Discussion of Results

For this workload, where all the accesses to the coupling facility were synchronous, TS Data Sharing performance is significantly better than the MRO base in terms of CPU useage and ITR, and transaction response times are either equivalent (allowing for repeatability), or better.

For applications that create or delete longer temporary storage objects, the access to the coupling facility will be asynchronous, and so it is expected that the response time for TS data sharing will be increased from the results reported here. However, in a multi-system sysplex, operations using function shipping as an alternative to TS data sharing, will require the use of MRO-XCF to ship the requests between the CECs to the single TS queue-owning region. This will increase the response time of this solution.

Chapter 16. GRS Ring/Star Special Study

GRS STAR introduces a significant change in the method GRS uses to coordinate ENQ/DEQ global resource serialization requests in a sysplex environment.

16.1 Introduction

In the past, GRS used a concept referred to as a GRS RING, where the complex is viewed as a ring of systems. Each request for global resources (ENQ or DEQ) is circulated around each system in the ring before actually granting the request. In this approach, all systems are peers, with each maintaining a local copy of the entire queue of global resource requests.

With the introduction of the STAR support in OS/390 R2, GRS will now use the contention detection and management capability of the coupling facility. Key highlights of the new STAR support are as follows:

1. Rebuild time is significantly reduced because the participating systems are not inter-dependent on each other to perform ENQ/DEQ processing. Also, in the case of the GRS RING method of processing, if a system is lost, the RING is disrupted and global resource processing on all systems is impacted. With STAR, if any participating system fails, none of the other systems are impacted.
2. GRS real storage consumption across the systems in the sysplex remains relatively constant as the size of the sysplex increases. Unlike GRS RING, which maintains a queue of all global resource serialization requests on all systems in the complex, STAR support maintains only a queue of requestors from the local system.
3. Both the processing capacity and responsiveness of GRS is significantly improved. As such, the new GRS STAR support provides installations with significant growth capability beyond what is possible with the RING.

16.2 Performance Test Environment

Measurements were made on an logically partitioned ES/9000 9021 model 952, with each partition using 1 dedicated CP and running the standard LSPR TSO workload at about 55% CPU. Global ENQ rate was about 30/sec per CEC. Variations include 2-way through 5-way OS/390 Release 1 GRS RING, 2-way through 4-way OS/390 Release 2 GRS RING, and 2-way through 5-way OS/390 Release 2 GRS STAR.

For the RING Measurements, RESMIL was set to 2 on the 2-way and 3-way, 1 on the 4-way and 0 on the 5-way. ACCELSYS was set to 2. XCF signalling via CTCs was used for GRS RING communication. For the STAR a 4CP C03 CF was used, with structures defined for XCF LIST and GRS LOCK, each of which were 10 MB. Contention on the GRS lock structure was kept to below 0.5%.

16.3 Results

This section includes our measurements of ENQ response time, ITR, and real storage consumption.

16.3.1 GRS Global ENQ Response Time

Figure 47 and Table 77 show the response time numbers measured by the GRS GENQRESP tool.

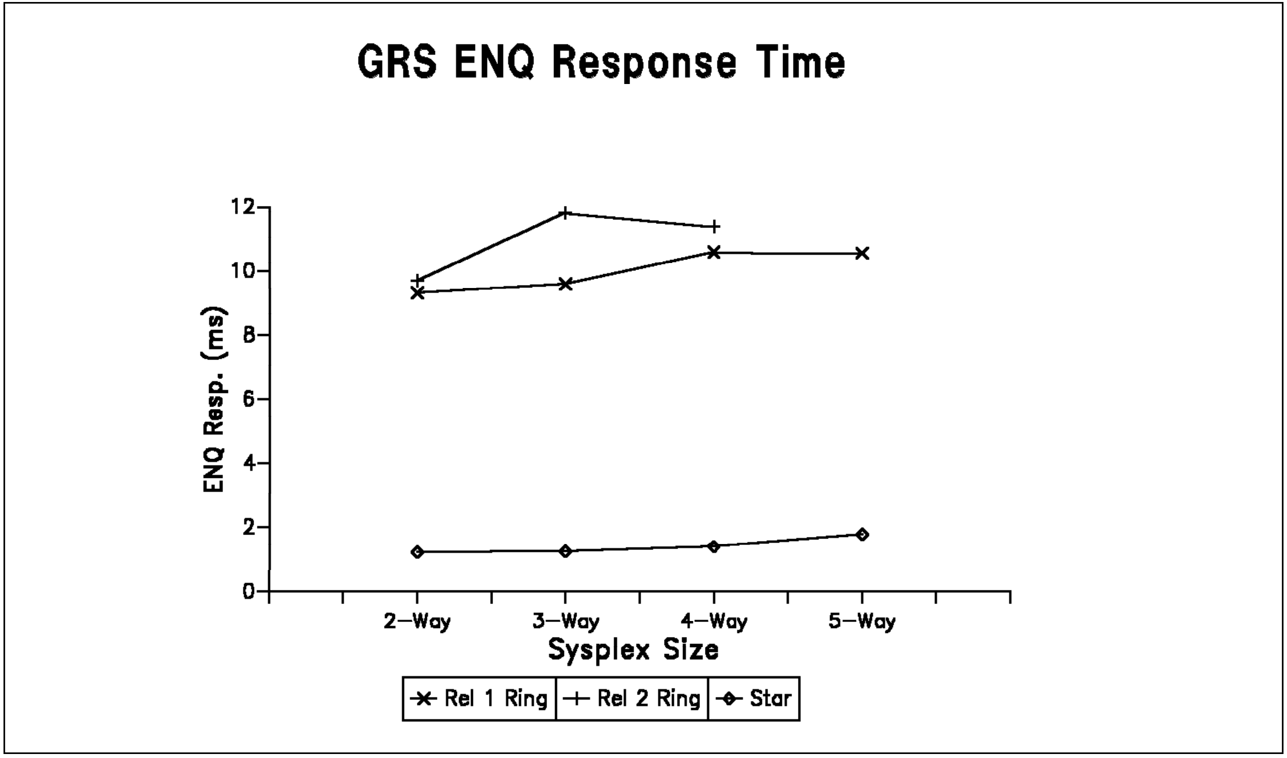


Figure 47. ENQ Response Time Comparison between OS/390 R1, R2 and GRS STAR

Table 77. GRS ENQ Response Time (ms)

GRS Environment	2-way	3-way	4-way	5-way
RESMIL	2	2	1	0
Release 1 ring respsonse time	9.32	9.61	10.60	10.56
Release 2 ring response time	9.71	11.83	11.40	N/A
STAR response time	1.23	1.26	1.40	1.79

As can be seen in Figure 47, the time it takes to turn around a GRS ENQ has dropped considerably with the new STAR implementation, such that all GRS RING sysplex sizes should experience some benefit. More importantly, STAR ENQ time remains fairly flat as the size of the sysplex increases, whereas it had continued to grow substantially with the RING implementation. The net effect on total system response time and CPU utilization will depend on current ENQ rates, and just how constrained GRS ENQ performance is in the RING. CF speed will affect GRS lock service times, which in turn will affect GRS ENQ times.

There is some variability in the results from the GRS GENQRESP tool. It is expected that GRS RING ENQ response times will continue to increase as the number of CECs in the RING increases, whereas GRS STAR ENQ response times will remain fairly flat. Note that the RESMIL value was decreased as we added more images to the sysplex. This helps the response time for the ring measurements. Note however, that at the 5-way sysplex we had reduced RESMIL to the minimum, and so response will continue to increase for the RING as the sysplex grows beyond five systems.

TSO period 1 and overall response time varied within measurement repeatability across release 1 and 2 RINGs and in the STAR.

16.3.2 System ITR

Figure 48 and Table 78 contain a comparison of the ITR between OS/390 R1, R2 and GRS STAR. The numbers are a ratio of the ITRs measured compared to the 2-way OS/390 Release 1 ring measurements.

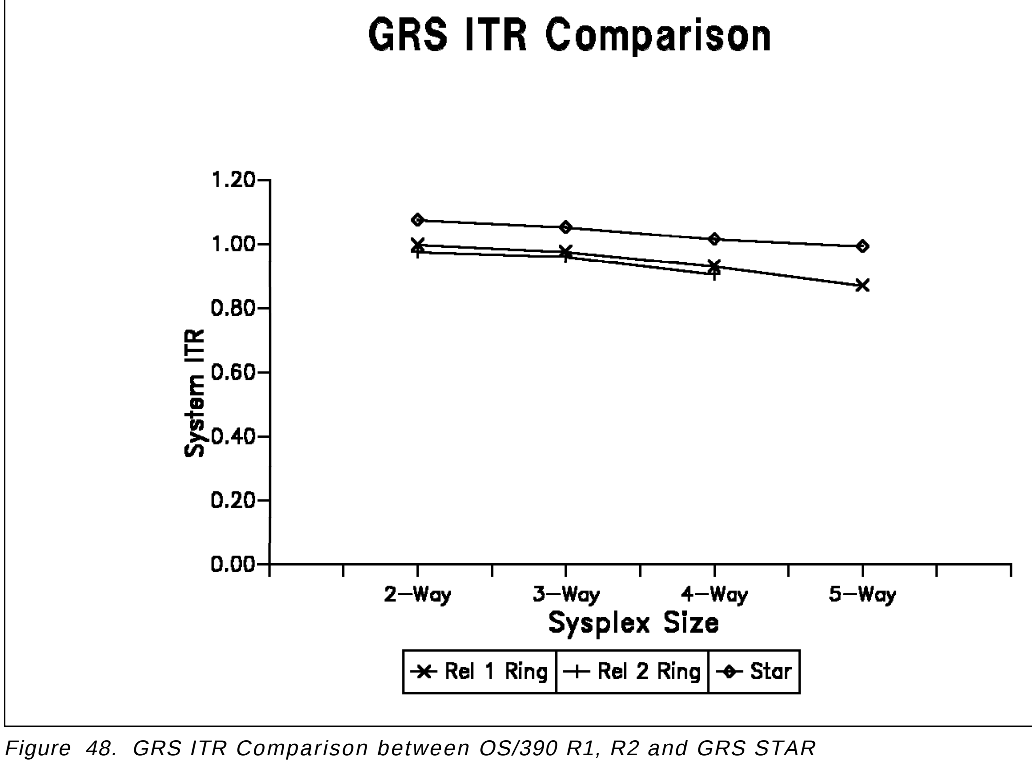


Figure 48. GRS ITR Comparison between OS/390 R1, R2 and GRS STAR

Table 78. GRS System ITR Ratio

GRS Environment	2-way	3-way	4-way	5-way
Release 1 Ring	1.000	0.979	0.932	0.873
Release 2 Ring	0.977	0.963	0.907	N/A
STAR	1.078	1.055	1.017	0.995

System internal throughput (ITR) also improved in the GRS STAR environment, due mostly to the fact that we no longer need to pass an RSA around the sysplex.

16.3.3 GRS Real Storage

Figure 49 and Table 79 report the number of storage frames used by GRS. The numbers were taken from the RMF Workload Activity report.

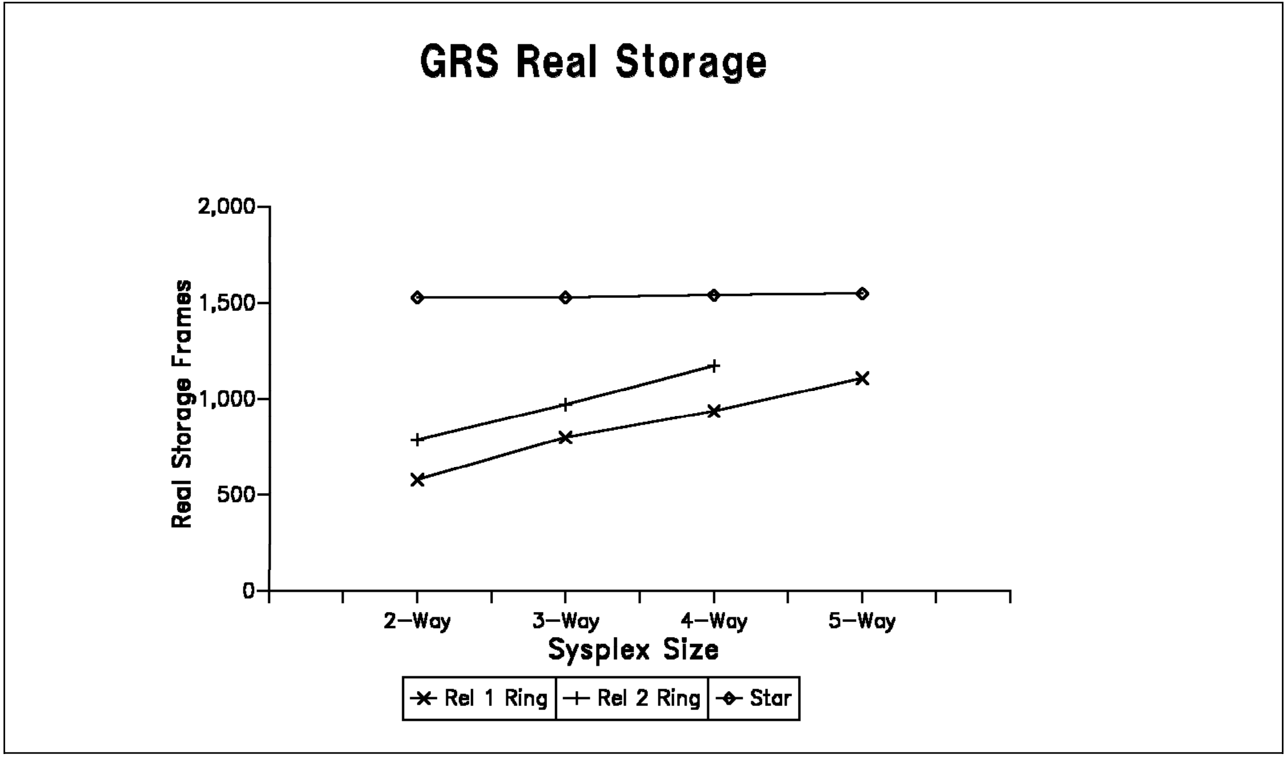


Figure 49. GRS Storage Comparison between OS/390 R1, R2 and GRS STAR

Table 79. GRS Real Storage (frames)

GRS Environment	2-way	3-way	4-way	5-way
Release 1 Ring	580	800	940	1110
Release 2 Ring	790	970	1170	N/A
STAR	1530	1530	1540	1550

There is some initial cost in terms of GRS real storage (central and expanded) moving from RING to STAR, which is mostly due to control blocks that are used in accessing the CF lock structure. However, GRS STAR real storage use remains fairly flat as the number of CECs increases, whereas GRS RING real storage use continues to grow. This is because STAR maintains a queue of only local system requestors, while in a RING, each system keeps a queue of all requests on all systems in the RING. With this workload, it is expected that the storage cost of release 1 RING will be equivalent to that of STAR in an 8-way Parallel Sysplex.

16.4 Miscellaneous

GQSCAN performance is much more expensive in a STAR environment. This is because each system no longer maintains a view of sysplex-wide ENQ activity, as was the case with GRS RING. Now, an initiating system must communicate with other systems in the STAR via XCF signalling to get a complete picture of sysplex-wide ENQs. All of the extra signalling and queue building results in higher GQSCAN costs in the STAR. However, GQSCAN frequency should be fairly low, being primarily issued by performance monitors.

The rate to the CF lock structure for GRS STAR (ISGLOCK) averaged about 61/sec per CEC in our studies. The GRS XCF signalling rate in the STAR environment, which primarily reflects GQSCAN activity, averaged about 0.2/sec; in the RING environment, the outbound rate generally averaged about 250/sec per CEC. The GRS STAR lock rate is roughly equal to the global ENQ (and DEQ) rate in the GRS RING.

The size of the GRS lock structure is primarily determined by the peak number of outstanding global ENQs. It should be large enough so that contention is kept to a minimum (that is, to less than 1%, as determined by RMF Coupling Facility reports).

Appendix A. CICS/DBCTL Workload Details and Coupling Efficiency Details

As discussed in section 1.3, "Parallel Sysplex Functions and Performance" on page 3, the CPU time cost of coupling comes from multisystem management functions and data sharing. This cost may be broken into its component and subcomponent parts through the use of a workload characterization that can be extracted from RMF and subsystem monitor data, coupling facility access rates and response times that can be found in RMF, and an understanding of data sharing effects. This breakdown would allow a knowledgeable analyst to then tailor each component or subcomponent to a particular customer′s workload to provide a better estimate of the capacity effects of coupling for that customer.

The CP90 and SNAPSHOT capacity planning tools incorporate (and automate) much of the discussion presented here. It is expected that most capacity planners will take advantage of these tools. This section is provided for those who want a more detailed understanding of the data.

A.1 CICS/DBCTL Workload Characteristics for a Single MVS Image

A detailed breakdown of the CPU time of the CICS/DBCTL workload as run on a single MVS image (9672-R61) with no IRLM is presented below. The data is drawn from the 9672-R61 run described in Chapter 4, "CICS/DBCTL Data Sharing Scalability Study" on page 65.

Table 80. CICS/DBCTL Workload: CPU Time Detail

	Milliseconds Per Transaction		
	Total	Subtotal	Sub-subtotal
Total CPU Time	52.12		
Time Charged to the TOR		10.56	
CICS Time in the TOR			8.18
Other Time in the TOR			2.38
Time Charged to the AORs		31.62	
CICS Time in the AORs			18.25
Other Time in the AORs			13.37
Other Captured Time		4.31	
Uncaptured Time		5.63	

Total CPU Time: Since this measurement was dedicated to transaction processing, the total CPU time may be divided by the total transactions to yield the total CPU time per transaction.

Time Charged to the TOR: CPU time charged to the TOR address space as reported in the TOR performance group (PGN) TCB+SRB time in the RMF Workload Activity Report. This includes time within CICS and time spent in other address spaces that MVS charges to the CICS TOR.

CICS Time in the TOR: CPU time within the TOR as recorded by CICS and reported in a CICS STATISTICS report. This will not include time spent outside the CICS TOR address space.

Other Time in the TOR: Calculated by subtracting *CICS Time in the TOR* from *Time Charged to the TOR*.

Time Charged to the AORs: CPU time charged to the AOR address spaces as reported in the AOR PGNs TCB+SRB time in the RMF Workload Activity Report. This includes time within CICS and time spent in other address spaces that MVS charges to the CICS AORs (for example, most of the time spent in the DBCTL address space on behalf of CICS will be charged here).

CICS Time in the AORs: CPU time within the AORs as recorded by CICS and reported in a CICS STATISTICS report. This will not include time spent outside the CICS AOR address spaces.

Other Time in the AORs: Calculated by subtracting *CICS Time in the AORs* from *Time Charged to the AORs*.

Other Captured Time: Calculated by subtracting the sum of *Time Charged to the TOR* and *Time Charged to the AORs* from the total captured time as reported in the RMF Workload Activity Report (TCB+SRB from ALL ALL ALL ALL). This reflects how much time is spent in other address spaces, generally in support of the CICS transactions (for example, VTAM, CAS, CPSM), but not charged to the TORs or AORs.

Uncaptured Time: Calculated by subtracting the total captured time as reported by RMF from the *Total CPU Time*. This reflects how much time is spent, generally in support of the CICS transactions, but not charged to any address space (for example, I/O interrupt processing, and parts of storage management).

A detailed breakdown of the data accesses of the CICS/DBCTL workload as run on a single MVS image (9672-R61) with no IRLM is presented in Table 81. The data is drawn from the 9672-R61 run described in Chapter 4, “CICS/DBCTL Data Sharing Scalability Study” on page 65.

Table 81. CICS/DBCTL Workload: Data Access Detail

	Data Access Average Count Per Transaction		
	Total	Subtotal	Sub-subtotal
VSAM+OSAM Requests to Buffer Manager	14.92		
Buffer Hits		11.54	
Buffer Misses		3.38	
Total I/O	5.66		
VSAM+OSAM		3.78	
Read I/O			3.38
Write I/O			.40
Other I/O		1.88	

Requests to Buffer Manager: The total number of requests for data to the VSAM and OSAM buffer managers taken from an IMS buffer pool statistics report. This value in relation to the CPU time breakdown provides an indication of the data access intensity of the application. The higher the number of buffer manager requests per unit of CPU time, the more data intense the application. Since a significant contributor to coupling overhead is how intensely an application references shared data, this can be a key value to analyze.

Buffer Hits: The number of times a request to the buffer manager was satisfied by data already existing in a buffer (thus no I/O was required) as reported by an IMS buffer pool statistics report.

Buffer Misses: The number of times a request to the buffer manager was not satisfied from the buffer pool (thus an I/O was required) as reported by an IMS buffer pool statistics report.

Total I/O: The total number of DASD I/Os taken from the RMF DASD Activity Report.

VSAM+OSAM I/O: The number of DASD I/Os to the VSAM and OSAM database volumes taken from the RMF DASD Activity Report (may also be calculated as the sum of the *Read I/O* and *Write I/O* given below).

Read I/O: The number of VSAM and OSAM read I/Os as reported by an IMS buffer pool statistics report.

Write I/O: The number of VSAM and OSAM write I/Os as reported by an IMS buffer pool statistics report.

Other I/O: The number of DASD I/Os other than those to the VSAM and OSAM database volumes. For this workload, these are mostly CICS Journal and IMS WADS and OLDS I/Os.

A.2 CICS/DBCTL Workload Characteristics for a Parallel Sysplex

Due to the effects of multisystem management and data sharing, the characteristics of the CICS/DBCTL workload change in a Parallel Sysplex environment. These changes include adding accesses to the coupling facility to support global locking (lock structure for IRLM), local buffer coherency (cache structures for VSAM and OSAM), and intersystem communication (list structures for XCF) to send messages on behalf of GRS, CPSM, WLM, IRLM, and XES. Other changes include additional I/O and an increase in CPU time per transaction.

These changes are summarized in Table 82 on page 174 and reflect the CICS/DBCTL workload as run on two different Parallel Sysplex configurations: 2x9672-R61 and 16x9672-R61. These runs are described in Chapter 4, "CICS/DBCTL Data Sharing Scalability Study" on page 65. Note the 9672-R61 data is repeated for comparison.

Table 82. CICS/DBCTL Workload: Summary

	Per Transaction		
	9672-R61	**2x9672-R61**	**16x9672-R61**
Total CPU Time (ms)	52.12	63.35	68.65
Total I/O	5.66	6.12	6.72
CF Accesses			
IRLM Lock	-	6.28	6.64
VSAM+OSAM Cache	-	3.94	4.33
XCF List	-	.92	2.05
JES2 Chkpt List	-	.01	.01

Total CPU Time: Defined earlier.

Total I/O: Defined earlier. The increase in I/O per transaction is generally due to invalidated local buffers and the spread of transactions across multiple systems, both contributing to a reduction in local buffer pool hits.

CF Accesses: A count of the number of accesses to the various structures on the coupling facility as reported in the RMF Coupling Facility Report.

IRLM Lock: The number of lock and unlock accesses to the lock structure used by IRLM for global locking. The increase in accesses per transaction is related to the I/O increase.

VSAM+OSAM Cache: The number of accesses to the cache structures for VSAM and OSAM. The increase in accesses per transaction is related to the I/O increase.

XCF List: The number of reads and writes to the list structures used by XCF. The increase reflects increased messaging activity between various members to resolve lock contention and to broadcast updated workload management statistics. Also, in the 16x9672-R61 measurement, there were additional XCF-management messages that could have been reduced with tuning.

JES2 Chkpt List: The number of reads and writes to the list structure used for the JES2 checkpoint. It is very small due to the lack of JES2 activity in this workload.

A.3 Coupling Overhead Breakdown

Coupling overhead refers to the percentage of a Parallel Sysplex′s capacity that is used for multisystem management and data sharing. See 1.4, “Parallel Sysplex Costs” on page 10 for a more thorough description.

As an example calculation using some CICS/DBCTL workload measurements, an 9672-R61 (Single Image) achieved an ITR of 115.1 transactions/sec while an 2x9672-R61 (2-way Sysplex) delivered 189.4. The percentage of capacity lost due to coupling may be calculated as:

$$\frac{(theoretical - actual)}{theoretical}$$

Thus, in this case we have:

$$\frac{(2 \times SI - 2way)}{(2 \times SI)} = \frac{(2 \times 115.1 - 189.4)}{(2 \times 115.1)} = 17.7\%$$

Since capacity and CPU cost per transaction have an inverse relationship, to yield a 17.7% loss in capacity, the CPU cost per transaction must have increased by:

$$\frac{1}{1 - 0.177} - 1 = 21.5\%$$

Referring to the workload description previously discussed, the total CPU time per transaction on the 9672-R61 was 52.12, which grew to 63.35 on the 2x9672-R61. This is an increase of $\frac{(63.35 - 52.12)}{52.12} = 21.5\%$, which cross-checks with the above expectation.

The coupling overhead comes from the following areas:

- Multisystem management functions
- Data sharing
  - Processing accesses to the coupling facility
  - Processing additional I/Os

A.3.1.1 Multisystem Management

The multisystem management functions include such things as GRS ring processing, JES2 shared checkpoint processing and workload management. These basic functions generally cost approximately 3-4%, plus an additional 0.1% to 0.2% per MVS image in the sysplex. To the basic functions, additional costs may be incurred for transaction routing and/or function shipping. Generally these additional costs will be reflected and accounted for through XCF activity. XCF activity to the coupling facility is asynchronous in nature, meaning once the read or write to the list structure is started, the CPU is free to process other work. The CPU overhead on a 9672-R61 for an XCF-related access (includes XCF processing, MVS processing and a small amount of hardware processing) is about 1.5 milliseconds. To scale to other configurations, this time would simply scale with the MVS image processor′s single-engine speed.

A.3.1.2 Data Sharing

The large variable in coupling overhead is the cost associated with data sharing. Shared data may be read and updated by any system, thus global locking and local buffer pool coherency functions must be applied to it. The cost is a function of the data access intensity to shared data and is directly related to the access frequency to the lock and cache structures on the coupling facility. See 1.3.3, “Data Sharing” on page 7 for a discussion of the various overheads. This cost can be 0% for sysplexes with no shared data.

An access to a lock or cache structure on the coupling facility is generally synchronous in nature, meaning the CPU processing the request will remain busy until it completes. Thus, the CPU time for a synchronous access includes time spent in the subsystem software (IRLM or IMS), time spent in MVS, and time spent while waiting for the response from the coupling facility. For a 9672-R61 connected to a 9674-C01 coupling facility, the total CPU time per synchronous access to a lock or IMS DB cache structure is in the range of 0.6 to 0.7 ms; the hardware component (as given in the RMF Coupling Facility Report) is usually a little less than half this total, and the software component is split roughly 50/50 between the subsystem and MVS. To scale to other configurations, the software time and about 40% of the hardware time would scale to the MVS-image processor′s single-engine speed; the rest of the hardware time would improve relative to the coupling-facility processor′s single-engine speed. The amount of improvement here would scale to approximately one-half the increase in single-engine speed. For example, if the single-engine speed of the coupling facility is 1.8 times faster, this component of access time would reduce by a factor of approximately 1.4. The full increase is not applied since some of the subcomponents of access time, namely the link-transit and link-hardware time, are insensitive to the coupling facility engine speed.

Also, the length (distance) of a coupling link can affect the time of a synchronous access; approximately 0.01 ms will be added for each kilometer of distance.

An additional component of data sharing overhead is related to growth in I/O rate.[42] We expect I/O per transaction to grow by about 1% to 5% per CEC added to a sysplex. The approximate CPU time cost of an IMS DB I/O on a 9672-R61 is 1.5 ms worth of software instructions and cache damage from the associated task switch.

A.3.1.3 Example Application to the CICS/DBCTL Benchmark

We can apply the above to our CICS/DBCTL benchmark measurements (refer to Table 82 on page 174). The CPU time per transaction on the 9672-R61 was about 52.1 ms. To calculate the cost of going to an 2x9672-R61, we would first apply about a 4% overhead for multisystem management (.04 × 52.1 = 2.1 ms). Next, we would account for the accesses to the coupling facility. There were 10.2 accesses to the lock and cache structures (10.2 × 0.65 = 6.6 ms) and 0.9 accesses to the XCF structure (0.9 × 1.5 = 1.4 ms). Finally, there were an additional 0.5 I/Os per transaction (0.5 × 1.5 = 0.8 ms). Thus, by component we have 2.1 ms for multisystem management and 8.8 ms for data sharing for a total of 10.9 ms; this is comparable to the measured growth of 11.3 ms (63.4-52.1).

We can repeat this exercise for the 16x9672-R61 measurement. Now the multisystem cost is up to around 7% (effects of the 0.1% to 0.2% per CEC variable) yielding 0.07 × 52.1 = 3.6 ms. There were 11 accesses to the lock and cache structures (11 × 0.70 = 7.7 ms), 2.1 accesses to the XCF structure (2.1 × 1.5 = 3.2 ms), and 1.1 additional I/Os (1.1 × 1.5 = 1.7 ms). Summing these gives a total of 16.2 ms, which is comparable to the measured growth of 16.6 ms (68.7-52.1).

Note the largest component of the coupling overhead in our benchmark was the data sharing component. This is due to our benchmark′s shared data access intensity. Since all data was shared, the access intensity is reflected by the buffer manager requests per transaction and the corresponding CPU time per transaction. If this ratio was half as large (with a lighter workload, for example), then the data sharing overhead component would likely be halved also, leading to a much lower coupling overhead. This, in fact, has tended to be the case with the early customer workloads that have migrated to a Parallel Sysplex where the data sharing overhead has been observed to be half that seen in our benchmark.

[42] The reasons for I/O growth in a Parallel Sysplex were briefly discussed in 1.3.3.3, “Buffer Invalidation” on page 9.

Appendix B. IMS-TM/DB2 Workload Details and Coupling Efficiency Details

As discussed in section 1.3, “Parallel Sysplex Functions and Performance” on page 3, the cost of CPU time for coupling results from multisystem management functions and data sharing. This cost may be broken into its component and subcomponent parts through the use of a workload characterization that can be extracted from RMF and subsystem monitor data. The cost is a function of coupling facility access rates and response times that can be found in RMF, and an understanding of data sharing effects. This breakdown would allow a knowledgeable analyst to then tailor each component or subcomponent to a particular customer′s workload to provide a better estimate of the capacity effects of coupling for that customer.

The CP90 and SNAPSHOT capacity planning tools incorporate (and automate) much of the discussion presented here. It is expected that most capacity planners will take advantage of these tools. This section is provided for those who want a more detailed understanding of the data.

B.1 IMS-TM/DB2 Workload Characteristics for a Single MVS Image

A detailed breakdown of the CPU time of the IMS-TM/DB2 workload as run on a single MVS image (9672-R61) is presented in Table 83. The data is drawn from the 9672-R61 run described in Chapter 5, “IMS-TM/DB2 Data Sharing Scalability Study” on page 73.

Table 83. IMS-TM/DB2 Workload: CPU Time Detail

	Milliseconds Per Transaction (Commit)		
	Total	Subtotal	Sub-subtotal
Total CPU Time	130		
DB2 Class 1 Time		63	
DB2 Class 2 Time			55
DB2 Service Address Spaces		19	
System Services			3
Database Services			15
IRLM			1
Other Captured Time		28	
Uncaptured Time		20	

Total CPU Time: Since this measurement was dedicated to processing the IMS-TM/DB2 workload, the total CPU time may be divided by the total transactions (commits) to yield the total CPU time per transaction.

DB2 Class 1 Time: The class 1 CPU time as reported by DB2 Accounting. This includes time within DB2 and that portion of time spent within IMS and the application while the DB2 thread was active.

DB2 Class 2 Time: The class 2 CPU time as reported by DB2 Accounting is the time spent within DB2.

DB2 Service Address Spaces: The sum of the CPU time spent in the DB2 service address spaces as reported by DB2 statistics.

System Services: The CPU time spent in the DB2 system services address space as reported by DB2 statistics.

Database Services: The CPU time spent in the DB2 database services address space as reported by DB2 statistics.

IRLM: The CPU time spent in the IRLM address space as reported by DB2 statistics.

Other Captured Time: Calculated by subtracting the sum of *DB2 Class 1 CPU Time* and *DB2 Service Address Spaces* from the total captured time as reported in the RMF workload activity report (TCB+SRB from ALL ALL ALL ALL).

Uncaptured Time: Calculated by subtracting the total captured time as reported by RMF from the *Total CPU Time*. This reflects how much time is spent, generally in support of the DB2 transactions, but not charged to any address space (for example, I/O interrupt processing and parts of storage management).

A detailed breakdown of the DB2 data accesses of the IMS-TM/DB2 workload as run on a single MVS image (9672-R61) is presented in Table 84. The data is drawn from the 9672-R61 run described in Chapter 5, "IMS-TM/DB2 Data Sharing Scalability Study" on page 73.

Table 84. IMS-TM/DB2 Workload: Data Access Detail

	DB2 Data Access Average Count Per Transaction (Commit)	
	Total	Subtotal
Getpages	71	
Synchronous Reads	8	
Pages Read Asynchronously	38	
Buffer Updates	15	
Pages Written	5	
SQL Statements	21	
Selects		5
Fetches		9
Insert+Update+Delete		7

Getpages The number of Getpages as reported by DB2 Statistics. Each getpage is a logical request to the DB2 buffer manager for an index or data page. This value in relation to the CPU time breakdown provides an indication of the data access intensity of the application. The higher the number of getpages per unit of CPU time, the more data intense the application. Since a significant contributor to coupling overhead is how intensely an application references shared data, this can be a key value to analyze.

Synchronous Reads: As reported by DB2 Statistics, this reflects the number of times a request to the DB2 buffer manager resulted in a single-page read I/O from a DB2 table.

Pages Read Asynchronously: As reported by DB2 Statistics, this reflects the number of pages that were read from a DB2 table via some form of pre-fetch (sequential, list, or dynamic). These pages are generally blocked from 8 to 32 pages per I/O and reflect the amount of sequential data access in the workload.

Buffer Updates: As reported by DB2 Statistics, this reflects the number of times index entries or data rows were updated.

Pages Written: As reported by DB2 Statistics, this reflects the number of pages that were written back to a DB2 table.

SQL Statements: As reported by DB2 Accounting, the total number of SQL statements. **Selects** and **fetches** reflect read access to data, **inserts+updates+deletes** reflect changes to data.

B.2 IMS-TM/DB2 Workload Characteristics for a Parallel Sysplex

Due to the effects of multisystem management and data sharing, the characteristics of the IMS-TM/DB2 workload change in a Parallel Sysplex environment. These changes include adding accesses to the coupling facility to support global locking (lock structure for IRLM), local buffer coherency and globally buffered data (cache structures for DB2 called Group Buffer Pools). These and additional processing in various system components (such as GRS, JES2, and WLM) may result in an increase in CPU time for some transactions. Note that unlike the IMS data sharing environment, there should be little increase in I/O rate for a DB2 data sharing workload. Since DB2 places the updated data in the coupling facility, systems with invalidated local copies of the data may refresh their copy from the coupling facility rather than from DASD. Thus, in our benchmark measurements, there was virtually no change in I/O rate.

The changes in the workload characteristics are summarized in Table 85 reflecting the IMS-TM/DB2 workload as run on two different Parallel Sysplex configurations: 2x9672-R61 and 8x9672-R61. These runs are described in Chapter 5, “IMS-TM/DB2 Data Sharing Scalability Study” on page 73. Note the single-image 9672-R61 data is repeated for comparison.

Table 85. IMS-TM/DB2 Workload: Summary

	Per Transaction (Commit)		
	9672-R61	**2x9672-R61**	**8x9672-R61**
Total CPU Time (ms)	130	157	162
CF Accesses			
IRLM Lock	-	15.1	15.9
DB2 Group Buffer Pool Caches	-	19.8	22.5
Lock Contention Events	-	0.3	0.5

Total CPU Time: Defined earlier.

CF Accesses: A count of the number of accesses to the various structures on the coupling facility as reported in the RMF Coupling Facility Report.

IRLM Lock: The number of lock, unlock and change accesses to the lock structure used by IRLM for global locking.

DB2 Group Buffer Pool Caches: The number of accesses to the cache structures used for DB2 Group Buffer Pools. The increase in accesses per transaction is related to the increased invalidation of local buffers.

Lock Contention Events: The number of times a lock request resulted in contention. When two systems request a lock with incompatible states (for example, exclusive and shared), lock contention occurs. This results most often when one system wishes to read data and another system wishes to update the same data.

B.3 Coupling Overhead Breakdown

Coupling overhead refers to the percentage of a Parallel Sysplex′s capacity that is used for multisystem management and data sharing. See 1.4, “Parallel Sysplex Costs” on page 10 for a more thorough description.

As an example calculation using some IMS-TM/DB2 workload measurements, an 9672-R61 (Single Image) achieved an ITR of 46.2 transactions/sec while an 2x9672-R61 (2way Sysplex) delivered 76.4. The percentage of capacity lost due to coupling may be calculated as:

$$\frac{(theoretical - actual)}{theoretical}$$

Thus, in this case we have:

$$\frac{(2 \times SI - 2way)}{(2 \times SI)} = \frac{(2 \times 46.2 - 76.4)}{(2 \times 46.2)} = 17.3\%$$

Since capacity and CPU cost per transaction have an inverse relationship, to yield an 17.3% loss in capacity, the CPU cost per transaction must have increased by:

$$\frac{1}{1 - 0.173} - 1 = 20.9\%$$

Referring to the workload description above, the total CPU time per transaction on the 9672-R61 was 130, which grew to 157 on the 2x9672-R61. This is an increase of $\frac{(157 - 130)}{130} = 20.8\%$, which cross-checks (given rounding effects) with the above expectation.

The coupling overhead comes from the following areas:

- Multisystem management functions
- Data sharing
  - Processing accesses to the coupling facility
  - Processing lock contention

B.3.1.1 Multisystem Management

The multisystem management functions include such things as GRS ring processing, JES2 shared checkpoint processing, and workload management. These basic functions generally cost approximately 3-4%, plus an additional 0.1% to 0.2% per MVS image in the sysplex. To the basic functions, additional costs may be incurred for transaction routing and/or function shipping. Generally these additional costs will be reflected and accounted for through XCF activity. Note that for the IMS-TM/DB2 workload, there was no XCF activity for transaction routing or function shipping since IMS was used as the transaction manager. This tended to keep the multisystem management effects for this workload at the lower end of the range.

B.3.1.2 Data Sharing

The large variable in coupling overhead is the cost associated with data sharing. Shared data may be read and updated by any system, thus global locking and local buffer pool coherency functions must be applied to it. Additionally, in the DB2 data sharing environment, updated data is stored and retrieved from global buffer pools to minimize the effect of invalidating local buffers. The cost is a function of the data access intensity to shared data and is directly related to the access frequency to the lock and cache structures on the coupling facility. See 1.3.3, "Data Sharing" on page 7 for a discussion of the various overheads. This cost can be 0% for sysplexes with no shared data.

An access to a lock or cache structure on the coupling facility is generally synchronous in nature, meaning the CPU processing the request will remain busy until it completes. Thus, the CPU time for a synchronous access includes time spent in the subsystem software (IRLM or DB2), time spent in MVS, and time spent while waiting for the response from the coupling facility. For a 9672-R61 connected to a 9674-C01 coupling facility, the cost in CPU time to access the IRLM lock structure (above what a locally-accessed lock would cost) is about 0.4 ms; the hardware component (as given in the RMF Coupling Facility Report) is usually a little more than half this total, and the software component is mostly in MVS.

DB2 sends a variety of requests to its group buffer pool cache structures on the coupling facility. Some of the accesses have no data, some do have data, and some contain lists of pages to be processed. For a 9672-R61 connected to a 9674-C01 coupling facility, the average cost in CPU time across all the different cache accesses made by the IMS-TM/DB2 workload was approximately 0.7 to 0.8 ms; the hardware component (as given in the RMF Coupling Facility Report) is usually about half this total, and the software component includes time spent in DB2 and MVS.

To scale the CPU time for lock or cache accesses to other configurations, the software time and about 40% of the hardware time would scale to the MVS-image processor′s single-engine speed; the remainder of the hardware time would improve relative to the coupling-facility processor′s single-engine speed. The amount of improvement here would scale to approximately one half the increase in single-engine speed. For example, if the single-engine speed of the coupling facility is 1.8 times faster, this component of access time would reduce by approximately 1.4. The full increase is not applied since some of the subcomponents of access time, namely the link-transit and link-hardware time, are insensitive to the coupling facility engine speed.

Also, the length (distance) of a coupling link can affect the time of a synchronous access; approximately 0.01 ms will be added for each kilometer of distance.

An additional component of data sharing overhead is related to lock contention. The reasons for lock contention were discussed earlier. The approximate CPU time cost to resolve a lock contention event on a 9672-R61 is 7 ms worth of software processing.

B.3.1.3 Example Application to the IMS-TM/DB2 Benchmark

We can apply the above to our IMS-TM/DB2 benchmark measurements which reflect a 100% data sharing case (refer to Table 85 on page 179). The CPU time per transaction on the 9672-R61 was 130 ms. To calculate the cost of going to an 2x9672-R61, we would first apply about a 3% overhead for multisystem management (0.03 × 130 = 3.9 ms). Next, we would account for the accesses to the coupling facility. There were 15.1 accesses to the lock structure (15.1 × 0.4 = 6.0 ms) and 19.8 accesses to the cache structure (19.8 × 0.75 = 14.9 ms). Additionally, there were 0.3 lock contention events (0.3 × 7 = 2.1 ms). Thus, by component we have 3.9 ms for multisystem management and 23.0 ms for data sharing for a total of 26.9 ms; this is comparable to the measured growth of 27 ms (157-130).

We can repeat this exercise for the 8x9672-R61 measurement. Now the multisystem cost is up to around 4% (note the multisystem cost tends to be lower in the IMS-TM/DB2 workload than in the CICS/DBCTL workload due to less transaction routing and function shipping) yielding 0.04 × 130 = 5.2 ms. There were 15.9 accesses to the lock structure (15.9 × 0.4 = 6.4 ms), 22.5 accesses to the cache structure (22.5 × 0.75 = 16.9 ms), and there were 0.5 lock contention events (0.5 × 7 = 3.5 ms). Summing these gives a total of 32.0 ms, which is comparable to the measured growth of 32 ms (162-130).

Note the largest component of the coupling overhead in our benchmark was the data sharing component. This is due to our benchmark′s high shared data access intensity. Since this is a transaction processing workload, the main factors contributing to the high intensity were 100% data sharing, negligible application processing, and the use of IMS IFP and WFI regions. (Note that access intensity can also be high for some batch and query workloads). Since all data was shared, the shared data access intensity is reflected by the DB2 getpage requests per transaction and the corresponding CPU time per transaction. If this ratio was half as large (due to, for example, more application processing or less than 100% data sharing), then the data sharing overhead component would likely be halved also, leading to a much lower coupling overhead. This, in fact, has tended to be the case with the early customer workloads that have migrated to a Parallel Sysplex where the data sharing overhead has been observed to be half that seen in our benchmark.

Appendix C. CICS/VSAM Workload Details and Coupling Efficiency Details

As discussed in section 1.3, “Parallel Sysplex Functions and Performance” on page 3, the CPU time cost of coupling comes from multisystem management functions and data sharing. This cost may be broken into its component and subcomponent parts through the use of a workload characterization that can be extracted from RMF and subsystem monitor data, coupling facility access rates and response times that can be found in RMF, and an understanding of data sharing effects. This breakdown would allow a knowledgeable analyst to then tailor each component or subcomponent to a particular customer′s workload to provide a better estimate of the capacity effects of coupling for that customer.

The CP90 and SNAPSHOT capacity planning tools incorporate (and automate) much of the discussion presented here. It is expected that most capacity planners will take advantage of these tools. This section is provided for those who want a more detailed understanding of the data.

C.1 CICS/VSAM Workload Characteristics for a Single MVS Image

A detailed breakdown of the CPU time of the CICS/VSAM workload as run on a single MVS image (9672-R33) with MRO is presented in Table 86. The data is drawn from the 9672-R33 run described in Chapter 6, “VSAM RLS Data Sharing Scalability Study” on page 81.

Table 86. CICS/VSAM Workload: CPU Time Detail

	Milliseconds Per Transaction		
	Total	Subtotal	Sub-subtotal
Total CPU Time	23.02		
Time Charged to the TORs		5.88	
CICS Time in the TORs			4.93
Other Time in the TORs			1.33
Time Charged to the AORs		6.72	
CICS Time in the AORs			6.55
Other Time in the AORs			.17
Time Charged to the FOR		5.44	
Other Captured Time		3.34	
Uncaptured Time		1.64	

Total CPU Time: Since this measurement was dedicated to transaction processing, the total CPU time may be divided by the total transactions to yield the total CPU time per transaction.

Time Charged to the TORs: CPU time charged to the TOR address spaces as reported in the TOR performance group (PGN) TCB+SRB time in the RMF Workload Activity Report. This includes time within CICS and time spent in other address spaces that MVS charges to the CICS TORs.

CICS Time in the TORs: CPU time within the TORs as recorded by CICS and reported in a CICS STATISTICS report. This will not include time spent outside the CICS TOR address spaces.

Other Time in the TORs: Calculated by subtracting *CICS Time in the TORs* from *Time Charged to the TORs*.

Time Charged to the AORs: CPU time charged to the AOR address spaces as reported in the AOR PGNs TCB+SRB time in the RMF Workload Activity Report. This includes time within CICS and time spent in other address spaces that MVS charges to the CICS AORs (for example, most of the time spent in the VSAM address space on behalf of CICS will be charged here).

CICS Time in the AORs: CPU time within the AORs as recorded by CICS and reported in a CICS STATISTICS report. This will not include time spent outside the CICS AOR address spaces.

Other Time in the AORs: Calculated by subtracting *CICS Time in the AORs* from *Time Charged to the FORs*.

Time Charged to the FORs: CPU time charged to the FOR address spaces as reported in the FOR PGNs TCB+SRB time in the RMF Workload Activity Report.

Other Captured Time: Calculated by subtracting the sum of *Time Charged to the TORs* and *Time Charged to the AORs* from the total captured time as reported in the RMF Workload Activity Report (TCB+SRB from ALL ALL ALL ALL). This reflects how much time is spent in other address spaces, generally in support of the CICS transactions (for example, VTAM, CAS, CPSM) but not charged to the TORs or AORs.

Uncaptured Time: Calculated by subtracting the total captured time as reported by RMF from the *Total CPU Time*. This reflects how much time is spent, generally in support of the CICS transactions, but not charged to any address space (for example, I/O interrupt processing, and parts of storage management).

A detailed breakdown of the data accesses of the CICS/VSAM workload as run on a single MVS image (9672-R33) with MRO is presented in Table 87. The data is drawn from the 9672-R33 run described in Chapter 6, “VSAM RLS Data Sharing Scalability Study” on page 81.

Table 87. CICS/VSAM Workload: Data Access Detail

	Data Access Average Count Per Transaction		
	Total	Subtotal	Sub-subtotal
VSAM Requests to Buffer Manager	9.29		
Buffer Hits		7.82	
Buffer Misses		1.47	
Total I/O	3.68		
VSAM		2.75	
Read I/O			1.47
Write I/O			1.28
Other I/O		.93	

Requests to Buffer Manager: The total number of requests for data to the VSAM buffer manager taken from an CICS LSR pool statistics report. This value in relation to the CPU time breakdown provides an indication of the data access intensity of the application. The higher the number of buffer manager requests per unit of CPU time, the more data intense the application. Since a significant contributor to coupling overhead is how intensely an application references shared data, this can be a key value to analyze.

Buffer Hits: The number of times a request to the buffer manager was satisfied by data already existing in a buffer (thus no I/O was required) as reported by an CICS LSR pool statistics report.

Buffer Misses: The number of times a request to the buffer manager was not satisfied from the buffer pool (thus an I/O was required) as reported by an CICS LSR pool statistics report.

Total I/O: The total number of DASD I/Os taken from the RMF DASD Activity Report.

VSAM I/O: The number of DASD I/Os to the VSAM database volumes taken from the RMF DASD Activity Report (may also be calculated as the sum of the *Read I/O* and *Write I/O* given below).

Read I/O: The number of VSAM read I/Os as reported by an CICS LSR pool statistics report.

Write I/O: The number of VSAM write I/Os as reported by an CICS LSR pool statistics report.

Other I/O: The number of DASD I/Os other than those to the VSAM database volumes.

C.2 CICS/VSAM Workload Characteristics for a Parallel Sysplex

Due to the effects of multisystem management and data sharing, the characteristics of the CICS/VSAM workload change in a Parallel Sysplex environment. These changes include adding accesses to the coupling facility to support global locking (lock structure for VSAM), local buffer coherency (cache structures for VSAM), and intersystem communication (list structures for XCF) to send messages on behalf of GRS, CPSM, WLM, and XES. Other changes include an increase in CPU time per transaction.

These changes are summarized in Table 88 on page 186 reflecting the CICS/VSAM workload as run on a Parallel Sysplex configuration consisting of two dedicated 3-way logical partitions on a 9672-R63. These runs are described in Chapter 6, “VSAM RLS Data Sharing Scalability Study” on page 81. Note the data for the single 3-way dedicated logical partition is repeated for comparison.

Table 88. CICS/VSAM Workload: Summary

	Per Transaction	
	9672	**2x9672**
Total CPU Time (ms)	23.01	27.80
Total I/O	3.68	1.93
CF Accesses		
VSAM Lock	-	1.46
VSAM Cache	-	4.02
JES2 Chkpt List	-	.02
CICS Log	-	1.02

Total CPU Time: Defined earlier.

Total I/O: Defined earlier. The decrease in I/O from MRO to RLS is due to the MRO Journal I/O now going to the CF with CICS log structure on the CF.

CF Accesses: A count of the number of accesses to the various structures on the coupling facility as reported in the RMF Coupling Facility Report.

VSAM Lock: The number of lock and unlock accesses to the lock structure used by VSAM for locking.

VSAM Cache: The number of accesses to the cache structures for VSAM.

JES2 Chkpt List: The number of reads and writes to the list structure used for the JES2 checkpoint. It is very small due to the lack of JES2 activity in this workload.

CICS Log List: The number of reads and writes to the list structure used for the CICS logging.

C.3 Coupling Overhead Breakdown

Coupling overhead refers to the percentage of a Parallel Sysplex′s capacity that is used for multisystem management and data sharing. See 1.4, “Parallel Sysplex Costs” on page 10 for a more thorough description.

As an example calculation using some CICS/VSAM workload measurements, a single image dedicated 3-way logical partition on a 9672-R63 achieved an ITR of 130.4 transactions/sec while a 2-way sysplex of 3-way dedicated logical partitions delivered 215.8. The percentage of capacity lost due to coupling may be calculated as:

$$\frac{(theoretical - actual)}{theoretical}$$

Thus, in this case we have:

$$\frac{(2 \times SI - 2way)}{(2 \times SI)} = \frac{(2 \times 130.4 - 215.8)}{(2 \times 130.4)} = 17.3\%$$

Since capacity and CPU cost per transaction have an inverse relationship, to yield a 17.3% loss in capacity, the CPU cost per transaction must have increased by:

$$\frac{1}{1 - 0.173} - 1 = 20.9\%$$

Referring to the workload description above, the total CPU time per transaction on the single-image base case was 23.01, which grew to 27.81 on the 2-way sysplex. This is an increase of $\frac{(27.81 - 23.01)}{23.01} = 20.9\%$, which cross-checks with the earlier expectation.

The coupling overhead comes from the following areas:

- Multisystem management functions
- Data sharing
  - Processing accesses to the coupling facility

C.3.1.1 Multisystem Management

The multisystem management functions include such things as GRS ring processing, JES2 shared checkpoint processing, and workload management. These basic functions generally cost approximately 3-4%, plus an additional 0.1% to 0.2% per MVS image in the sysplex. To the basic functions, additional costs may be incurred for transaction routing and/or function shipping. Generally these additional costs will be reflected and accounted for through XCF activity.

C.3.1.2 Data Sharing

The large variable in coupling overhead is the cost associated with data sharing. Shared data may be read and updated by any system, thus global locking and local buffer pool coherency functions must be applied to it. The cost is a function of the data access intensity to shared data and is directly related to the access frequency to the lock and cache structures on the coupling facility. See 1.3.3, “Data Sharing” on page 7 for a discussion of the various overheads. This cost can be 0% for sysplexes with no shared data.

An access to a lock or cache structure on the coupling facility is generally synchronous in nature, meaning the CPU processing the request will remain busy until it completes. Thus, the CPU time for a synchronous access includes time spent in the subsystem software (VSAM), time spent in MVS, and time spent while waiting for the response from the coupling facility. For the dedicated 3-way logical partition connected to a 9674-C03 coupling facility, the total CPU time per synchronous access to a lock or VSAM cache structure is in the range of 0.45 to 0.65 ms; the hardware component (as given in the RMF Coupling Facility Report) is usually a little less than half this total, and the software component is split roughly 50/50 between the subsystem and MVS. To scale to other configurations, the software time and about 40% of the hardware time would scale to the MVS-image processor′s single-engine speed; the rest of the hardware time would improve relative to the coupling-facility processor′s single-engine speed. The amount of improvement here would scale to approximately one half the increase in single-engine speed. For example, if the single-engine speed of the coupling facility is 1.8 times faster, this component of access time would reduce by a factor of approximately 1.4. The full increase is not applied since some of the subcomponents of access time, namely the link-transit and link-hardware time, are insensitive to the coupling facility engine speed.

Also, the length (distance) of a coupling link can affect the time of a synchronous access; approximately 0.01 ms will be added for each kilometer of distance.

The CICS Log use of the CF generally costs more than CICS Journalling, but is also in the 0.45 to 0.65 ms range.

C.3.1.3 Example Application to the CICS/VSAM Benchmark

We can apply the above to our CICS/VSAM benchmark measurements. Please refer to Table 82 on page 174. The CPU time per transaction on the base single-image case was about 23.0 ms. To calculate the cost of going to a 2-way sysplex, we would first apply about a 4% overhead for multisystem management (.04 × 23.0 = 0.9 ms). Next, we would account for the accesses to the coupling facility. There were 6.5 accesses to the lock, cache, and log structures (6.5 × 0.55 = 3.6 ms). Thus, by component we have 0.9 ms for multisystem management and 3.6 ms for data sharing for a total of 4.5 ms; this is comparable to the measured growth of 4.8 ms (27.8-23.0).

Note the largest component of the coupling overhead in our benchmark was the data sharing component. This is due to our benchmark′s shared data access intensity. Since all data was shared, the access intensity is reflected by the buffer manager requests per transaction and the corresponding CPU time per transaction. If this ratio was half as large (with a lighter workload for example), then the data sharing overhead component would likely be halved also, leading to a much lower coupling overhead. This, in fact, has tended to be the case with the early customer workloads that have migrated to a Parallel Sysplex where the data sharing overhead has been observed to be half that seen in our benchmark.

Appendix D. Capacity Planning Tools

Various tools have been updated and made available to IBM representatives to help customers understand capacity planning issues. These include but are not limited to:

- CP2000
- SNAPSHOT
- QUICKSIZER
- BWATOOL

CP2000 supports "S/390 Parallel Sysplex" environments as targets in traditional capacity planning scenarios. Workloads may be designated to be distributed across a group of system images defined as a Parallel Sysplex.

CP2000 is available from HONE for downloading to a PC.

SNAPSHOT services are part of the End-to-End Capacity Planning and Design Services Installation Service Offering (ISO). This ISO provides premier tools and services to IBM′s customers in the area of capacity planning, performance analysis, and design. For more information on this service, contact SNAPSHOT at DALVM41B.

The *Quick-Sizer* tool is available, both in the host-based CP2000, and as a PC-based OS/2 application. Minimal input is required, describing current workloads and the portion of the workloads targeted for Parallel Sysplex implementation. Results provide a high-level estimate for the Parallel Sysplex configuration required to support the designated workload (number of processors, number of links, and storage size). The size and number of processors required for the coupling facility is also estimated. The PC-based "CP90 Quick-Sizer" can be obtained from CPSTOOLS at WSCVM using the either of the following commands:

```
OMNIDISK CPSTOOLS GET SPSSZR PACKAGE
EXEC TOOLS SENDTO WSCVM TOOLS CPSTOOLS GET SPSSZR PACKAGE
```

In addition, two new tools have been developed to help pull together the data required for Quick-Sizer for a DB2 and a CICS/VSAM environment. *CVRCAP* is available from MKTTOOLs and is essential to capacity planning for CICS/VSAM RLS. *DB2PARA* is also available from MKTTOOLS, and provides another option to gathering data for the DB2 environment.

For additional information on CP90 capacity planning tools, please send a note to IBM′s Capacity Planning Services (CP90ID at DALVM41B).

The *BWATOOL* Batch Workload Analysis Tool evaluates response times for batch jobs and can be obtained from the LSCD S/390 Parallel Center by phone at (914) 435-3752 or by contacting VNET ID: CALLS390 at PKEDVM9.

Appendix E. Special Notices

This publication is to help customer technical staff understand the performance characteristics of Parallel Sysplex. Since not all measurements were made with generally available hardware and software, some adjustments were made to estimate the effects of moving to these levels.

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM′s product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM′s intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to the IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer′s ability to evaluate and integrate them into the customer′s operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

Any performance data contained in this document was determined in a controlled environment, and therefore, the results that may be obtained in other operating environments may vary significantly. Users of this document should verify the applicable data for their specific environment.

Reference to PTF numbers that have not been released through the normal distribution process does not imply general availability. The purpose of including these reference numbers is to alert IBM customers to specific information relative to the implementation of the PTF when it becomes available to each customer according to the normal IBM PTF distribution process.

The following terms are trademarks of the International Business Machines Corporation in the United States and/or other countries:

CICS
CICS/ESA
CICSPlex
DATABASE 2
DB2
DFSMS
DFSMS/MVS
Enterprise System/9000
Enterprise Systems Architecture/390
Enterprise Systems Connection Architecture
ES/9000
ESA/390
ESCON
Hiperspace
IBM
IMS
IMS/ESA
MVS/ESA
OS/390
Parallel Sysplex
S/390
S/390 Parallel Enterprise Server
Sysplex Timer
System/390
S/390 Parallel Enterprise Server

The following terms are trademarks of other companies:

C-bus is a trademark of Corollary, Inc.

Java and HotJava are trademarks of Sun Microsystems, Incorporated.

Microsoft, Windows, Windows NT, and the Windows 95 logo are trademarks or registered trademarks of Microsoft Corporation.

PC Direct is a trademark of Ziff Communications Company and is used by IBM Corporation under license.

Pentium, MMX, ProShare, LANDesk, and ActionMedia are trademarks or registered trademarks of Intel Corporation in the U.S. and other countries.

UNIX is a registered trademark in the United States and other countries licensed exclusively through X/Open Company Limited.

Other company, product, and service names may be trademarks or service marks of others.

Other trademarks are trademarks of their respective companies.

Appendix F. Related Publications

The publications listed in this section are referenced elsewhere in this document. At the time of printing this redbook some of the manuals mentioned here were not yet available.

F.1 International Technical Support Organization Publications

For information on ordering these ITSO publications see "How to Get ITSO Redbooks" on page 195.

- *CICS and VSAM Record Level Sharing: Planning Guide*, SG24-4765
- *DB2 for MVS/ESA Version 4 Data Sharing Performance Topics*, SG24-4611
- *MVS/ESA Version 5 Sysplex Migration Guide*, SG24-4581
- *Workload Manager Performance Studies*, SG24-4352

F.2 Redbooks on CD-ROMs

Redbooks are also available on CD-ROMs. **Order a subscription** and receive updates 2-4 times a year at significant savings.

CD-ROM Title	Subscription Number	Collection Kit Number
System/390 Redbooks Collection	SBOF-7201	SK2T-2177
Networking and Systems Management Redbooks Collection	SBOF-7370	SK2T-6022
Transaction Processing and Data Management Redbook	SBOF-7240	SK2T-8038
Lotus Redbooks Collection	SBOF-6899	SK2T-8039
Tivoli Redbooks Collection	SBOF-6898	SK2T-8044
AS/400 Redbooks Collection	SBOF-7270	SK2T-2849
RS/6000 Redbooks Collection (HTML, BkMgr)	SBOF-7230	SK2T-8040
RS/6000 Redbooks Collection (PostScript)	SBOF-7205	SK2T-8041
RS/6000 Redbooks Collection (PDF Format)	SBOF-8700	SK2T-8043
Application Development Redbooks Collection	SBOF-7290	SK2T-8037

F.3 Other Publications

These publications are also relevant as further information sources. Books with an "LY" prefix are available to IBM-licensed customers only.

- *CICS Transaction Server for OS/390 V1R1 System Definition Guide*, SC33-1682
- *CICS/ESA Performance Guide*, SC33-1183
- *CICS/ESA System Definition Guide*, SC33-1164
- *CICSPlex SM Setup and Administration*, SC33-0784
- *DB2 for MVS/ESA Version 4 Data Sharing — Planning and Administration*, SC26-3269
- *DFSMS/MVS Version 1 Release 3: DFSMSdfp Storage Administration Guide*, SC26-4920
- *DFSMS/MVS Version 1 Release 3: DFSMSdfp Storage Administration Reference*, SC26-4929
- *ES/9000 PR/SM Planning Guide*, GA22-7123
- *IMS/ESA Version 5 Administration Guide: System*, SC26-8013

- *IMS/ESA Version 5 Customization Guide*, SC26-8020
- *JES2 Multi-Access Spool in a Sysplex Environment*, GG66-3263
- *Large Systems Performance Reference*, SC28-1187
- *MVS/ESA Analyzing Resource Measurement Facility Version 5 — Getting Started on Performance Management*, LY33-9176
- *MVS/ESA Analyzing Resource Measurement Facility Version 5 Reports*, LY33-9178
- *MVS/ESA Resource Measurement Facility Version 5 User′s Guide*, GC33-6483
- *MVS/ESA SP Version 5 Release 1 Performance Studies*, GG66-3258
- *MVS/ESA SP V5 JES2 Initialization and Tuning Guide*, SC28-1453
- *MVS/ESA SP V5 Planning: Global Resource Serialization*, GC28-1450
- *MVS/ESA SP V5 Setting up a Sysplex*, GC28-1449
- *OS/390 Parallel Sysplex Application Migration*, GC28-1863

How to Get ITSO Redbooks

This section explains how both customers and IBM employees can find out about ITSO redbooks, redpieces, and CD-ROMs. A form for ordering books and CD-ROMs by fax or e-mail is also provided.

- **Redbooks Web Site** http://www.redbooks.ibm.com/

  Search for, view, download or order hardcopy/CD-ROMs redbooks from the redbooks Web site. Also read redpieces and download additional materials (code samples or diskette/CD-ROM images) from this redbooks site.

  Redpieces are redbooks in progress; not all redbooks become redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

- **E-mail Orders**

  Send orders via e-mail including information from the redbook order form to:

	IBMMAIL	**Internet**
In United States:	usib6fpl at ibmmail	usib6fpl@ibmmail.com
In Canada:	caibmbkz at ibmmail	lmannix@vnet.ibm.com
Outside North America:	dkibmbsh at ibmmail	bookshop@dk.ibm.com

- **Telephone Orders**

United States (toll free)	1-800-879-2755	
Canada (toll free)	1-800-IBM-4YOU	
Outside North America	(long distance charges apply)	
(+45) 4810-1320 - Danish	(+45) 4810-1220 - French	(+45) 4810-1270 - Norwegian
(+45) 4810-1420 - Dutch	(+45) 4810-1020 - German	(+45) 4810-1120 - Spanish
(+45) 4810-1540 - English	(+45) 4810-1620 - Italian	(+45) 4810-1170 - Swedish
(+45) 4810-1670 - Finnish		

This information was current at the time of publication, but is continually subject to change. The latest information for customers may be found at http://www.redbooks.ibm.com/ and for IBM employees at http://w3.itso.ibm.com/.

IBM Intranet for Employees

IBM employees may register for information on workshops, residencies, and redbooks by accessing the IBM Intranet Web site at http://w3.itso.ibm.com/ and clicking the ITSO Mailing List button. Look in the Materials repository for workshops, presentations, papers, and Web pages developed and written by the ITSO technical professionals; click the Additional Materials button. Employees may also view redbook, residency and workshop announcements at http://inews.ibm.com/.

IBM Redbook Fax Order Form

Fax your redbook orders to:

United States (toll free)	1-800-445-9269
Canada	1-403-267-4455
Outside North America	(+45) 48 14 2207 (long distance charge)

Please send me the following:

Title	**Order Number**	**Quantity**

First name Last name

Company

Address

City Postal code Country

Telephone number Telefax number VAT number

- Invoice to customer number ________________________________

- Credit card number ________________________________

Credit card expiration date Card issued to Signature

We accept American Express, Diners, Eurocard, Master Card, and Visa. Payment by credit card not available in all countries. Signature mandatory for credit card payment.

List of Abbreviations

ACEE	access control environment element
ALLOC	allocate virtual storage
AOR	application owning region
APAR	authorized program analysis report
CDS	couple data set
CEC	central electronics complex (synonym for CPC, central processing complex)
CF	coupling facility
CFRM	coupling facility resource management
CICS	customer information control system (IBM)
CICS/ESA	customer information control system/enterprise systems architecture (IBM)
CMAS	CICS managing address space
CMOS	complementary metal oxide semiconductor
CP	central processor
CPC	central processing complex
CPU	central processing unit
CTC	channel to channel
DASD	direct access storage device
DB	database
DBCTL	database control subsystem
DBRC	database recovery control (IMS)
DEQ	dequeue
DFW	DASD fast write
DL/I	data language 1
DMB	data management block (IMS)
ESCON	enterprise systems connection (architecture, IBM System/390)
ESDS	entry sequenced data set (VSAM)
FIFO	first in/first out
GRS	global resource serialization (MVS)
GU	get unique (IMS)
H/W	hardware
HDAM	hierarchic direct access method
HIDAM	hierarchic indexed direct access method
I/O	input/output
IBM	International Business Machines Corporation
IEBGENER	utility program (MVS)
IMS	information management system
IMS/ESA	information management system/enterprise systems architecture
IO	input/output
IOCDS	I/O configuration data set
IRLM	IMS/VS resource lock manager
IS	Information Systems
ISC	intersystem coupling
ISRT	insert (IMS)
ITR	internal throughput/transaction rate
ITSC	International Technical Support Center (IBM)
ITSO	International Technical Support Organization
JES	job entry subsystem
KSDS	key sequenced data set
LSPR	large systems performance reference (IBM, Washington Systems Center)
MAS	multiple access spool
MB	megabyte, 1,000,000 bytes (1,048,576 bytes memory)
Mb	megabit, 1,000,000 bits (1,048,576 bits memory)
MEC	manufacturing engineering change
MIPS	million instructions per second
MP	multiprocessor
MRO	multiregion operation
MVS	multiple virtual storage (IBM System 370 & 390)

MVS/ESA	multiple virtual storage/enterprise systems architecture (IBM)
N/A	not applicable
OLDS	online log data set (IMS)
OLTP	online transaction processing
OSAM	overflow sequential access method (IMS)
PCB	program communication block (IMS)
PGN	performance group number
PI	performance index
PR/SM	processor resource/systems manager (IBM)
PROCLIB	procedure library (IBM System/360)
PSB	program specification block
PSBNAME	program specification block name
PTF	program temporary fix
PTS	parallel transaction server
PUT	program update tape
RACF	resource access control facility
RBA	relative byte address
RECON	recovery control (data set)
RMF	resource measurement facility (MVS)
RNL	resource name list
S/W	software
SCH	subchannel
SCHED	schedule
SMF	system measurement facility
SMQ	shared message queue
SMSG	special message (VM)
SP	system product
SPOOL	simultaneous peripheral operation online
SRM	system resources manager
SUP	service update package/process
SUSP	suspend
SYSPLEX	systems complex
TOR	terminal owning region
TSO	time sharing option
USA	United States of America
VNET	virtual node exchange transmission
VSAM	virtual storage access method (IBM)
VTAM	virtual telecommunications access method (IBM) (runs under MVS, VM, & DOS/VSE)
XCF	cross-system coupling facility (MVS)
XES	cross-system extended services (MVS)

Index

ITSO Redbook Evaluation

System/390 Parallel Sysplex Performance
SG24-4356-03

Your feedback is very important to help us maintain the quality of ITSO redbooks. **Please complete this questionnaire and Fax it to: USA International Access Code + 1 914 432 8264 or:**

- Use the online evaluation form found at http://www.redbooks.ibm.com
- Send your comments in an Internet note to redbook@us.ibm.com

Which of the following best describes you?
__Customer __Business Partner __Solution Developer __IBM employee
__None of the above

Please rate your overall satisfaction with this book using the scale:
(1 = very good, 2 = good, 3 = average, 4 = poor, 5 = very poor)

Overall Satisfaction ____________

Please answer the following questions:

Was this redbook published in time for your needs? Yes____ No____

If no, please explain:
____

____

____

____

What other redbooks would you like to see published?
____

____

____

Comments/Suggestions: **(THANK YOU FOR YOUR FEEDBACK!)**
____

____

____

____

____

System/390 Parallel Sysplex Performance

SG24-4356-03

SG24-4356-03
Printed in the U.S.A.