FAT TOPICS

An OS/390 Newsletter - Issue 3

August 2000

From the editors ...

As our resident curmudgeon once observed on hearing that time flies when you're having fun: "Time flies whether you're having fun or not!" Well, six months have passed very quickly and we DID have fun, once again, putting Issue 3 of the Newsletter together. And a hefty one it is! We even thought about changing the title to FAT TOPICS!

Our articles this time focus on helping you make your information technology installation a more dynamic place. What do we mean by that? More caffeine in your coffee? Free energy pills? Of course not. We mean maximizing your systems' ability to manage increasing volumes of workloads in the most efficient way. And we believe you achieve this by allowing your systems, dynamically, to make workload balancing and resource management decisions based on real-time conditions. So, we're featuring articles about Parallel Sysplex® and 'goal mode' in WLM as our theme this time, articles that showcase the underlying technologies, demonstrating the value and use of dynamic resource and workload management.

Our supplement adds to the theme by containing articles on related topics such as tools, hints and tips, and level-set information for those just getting started. We suggest you take a quick look at the table of contents to see what you might be interested in, then turn to the article immediately following this one for an overview of our theme and an introduction to the articles supporting it.

In addition to the gratitude we feel toward all the authors, who put in much time and effort to provide you the benefit of their expertise, a special note of thanks is due our guest editors, Madeline Nick, Jose Castano, and Kris Perry from the Parallel Sysplex Product Development Team. They solicited, read, reviewed, and coordinated the many articles, and in the case of Madeline, wrote a few.

In this issue we also bring you edification and amusement (we hope) by offering a crossword puzzle whose clues and answers revolve around OS/390 system concepts, especially those that are WLM and Sysplex related. The crossword puzzle is in the centerfold and answers on the following page.

Finally, in the next issue we want to begin a 'letters to the editor' column. If you have a question or a comment that you feel has general interest or applicability, please send us an e-mail, and we'll publish what you have to say and respond to your comment.

The Editors,
newsletr@us.ibm.com

Find out What's Hot!

Table of Contents

Introduction to the theme: Dynamic workload balancing: "The magic inside the S/390 box"

Madeline Nick

Madeline is a Senior Software Engineer in the Parallel Sysplex Product Development team in Poughkeepsie, NY. Before that she worked in MVS Design and Performance for six years. She has worked on Parallel Sysplex since its announcement in 1994.

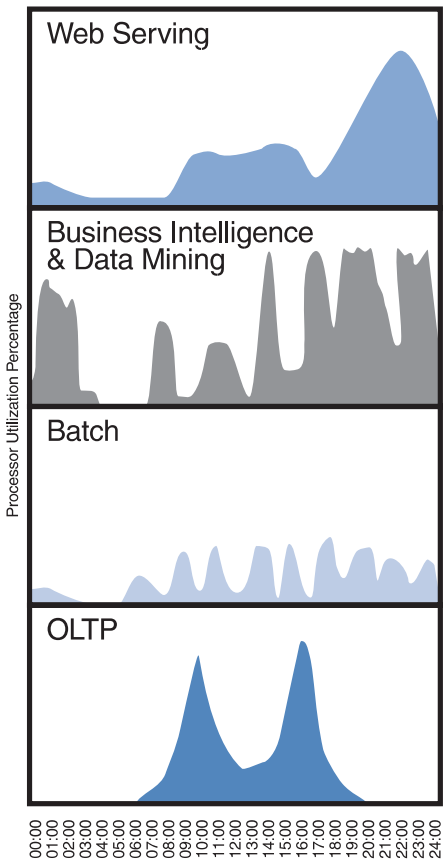
The IBM S/390® server platform is generally regarded by industry analysts and leading companies as the defining standard in enterprise computing — especially in the age of e-business. The platform's scalability, availability, security, interoperability, ability to manage mixed workloads, and its low total cost of ownership are the pillars on which enterprise-wide solutions are built. Although the qualities of service of S/390 are generally accepted as best-of-breed, the underlying technologies or "magic" that delivers those qualities are not as widely understood. This Newsletter reveals the secrets behind the "magic".

One of the most critical components of the magic is: *Dynamic workload Balancing:* the capability to dynamically direct work requests associated with a multi-system workload to run on an OS/390® image within a Parallel Sysplex that has sufficient server resources required to meet customer-specified workload goals. A related capability is *Dynamic Resource Management:* the ability to dynamically allocate and/or re-distribute server resources, such as CPU, I/O, and memory across a set of workloads based on their workload goals and relative priorities within an OS/390 images. Dynamic resource management occurs at every level of the integrated S/390 platform.
One of OS/390's greatest strengths as an operating system, certainly one of its leading differentiating values, lies in its ability to manage multiple workloads efficiently within a single operating system image. With Workload Manager, physical CPU, I/O and memory resources are flexibly distributed across mixed workloads in accordance with the priorities and goals of each of the managed workloads. Since different workloads have different resource utilization characteristics over time, peaks in resource demand by one workload can be met by applying resource from another workload which is not currently running at peak capacity, or is lower in priority. This would not be possible if each workload was running on a separate physical server. Assuming each system is configured to meet the peak capacity demand of the workload that is deployed on that system, when the workload is not at peak demand any excess resource capacity is simply wasted.
As one OS/390 customer, Dan Kaberon from Hewitt Associates put it: "You can't put these unused cpu cycles in your pocket and apply them later!" OS/390 customers deploy multiple workloads such as batch, TSO, CICS®, DB2® within the same single OS/390 server today in order to take advantage of these rich flexible resource management characteristics.

In an e-business environment, where the standard deviation between average resource consumption and peak utilization is even more dramatic, this capability becomes increasingly important.

Benefits of Workload Aggregation

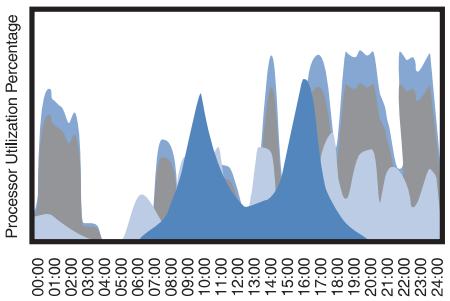
Physically separate servers
Unused capacity is wasted



S/390 optimizes capacity usage via resource management across one or more images

Same 4 workloads run concurrently on single S/390 server

- Prime shift average utilization - 70%
- Peak utilization - 100%
- Requires significantly less total capacity than equivalent number or physically separate servers
- S/390 workload Manager maximizes throughout and minimizes "white space"



In this Newsletter, we cover the key components that make up this part of the S/390 "magic." These components include hardware and microcode features that provide dynamic management of server resources, such as Processor Resource/Systems Manager (PR/SM)™ and Capacity Upgrade on Demand (CUoD). We also discuss core OS/390 resource management software functions like the I/O Subsystem Manager (IOS), the Workload Manager, Communications Server, and major S/390 subsystems. We further explore the synthesis of these functions in a Parallel Sysplex, delivering true dynamic application workload balancing across multiple system images.

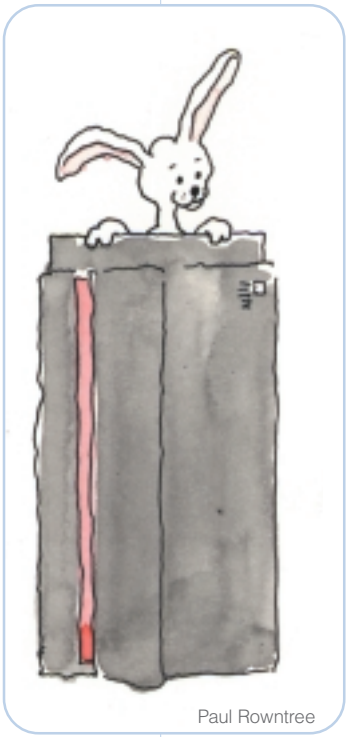
This ability to dynamically allocate physical CPU, I/O, and memory resources on demand based on priorities is not a capability limited at the OS/390 operating system level. The S/390 platform can also re-distribute resources across multiple operating system logical images within a physical processing system via the PR/SM hipervisor predefined system weights. These weights are adjusted dynamically based on your workload today, but watch this space for future enhancements in this area.

* In September 1999, according to Gartner analysts, Andrew Butler and Derek Prior, "Dynamic hard partitioning is the 'Holy Grail' of Unix vendors; only mainframes offer truly dynamic partitioning. Domains can be built, resized, merged and deleted under application control. CPUs can be shared on a moment-to-moment basis, memory can be moved in smaller increments (independently of CPUs or I/O channels) and I/O channels can be moved one by one or even shared." To put this in IBM S/390 terms: Dynamic Hard Partitioning is S/390 Logical Partitions, mainframe is S/390 and domains are S/390 Logical Partitions (LPARS).

The following topics are all discussed in detail in the Newsletter. Here's a quick preview:

Processor Resource/Systems Manager

Processor Resource/Systems Manager is a component of S/390, since 1988, that provides Logical Partitions (LPARs) of your S/390 server. Each S/390 server can have up to 15 LPARs, each running their own copy of the OS/390 operating system and each having some share of the processors, memory, and I/O of the server. Resources can by assigned/managed on a LPAR basis as defined by the customer's needs. PR/SM does such a great job of dividing the hardware resources and isolating the environment that, as far as OS/390 is concerned, it is running on a separate server. Each workload is isolated from failures that occur in other OS/S390 images on different Logical Partitions. S/390 PR/SM is considered to be the most powerful and flexible partition solution in the industry today.* See "Taking the myth out of PR/SM" by Jeff Kubala for more details.



Paul Rowntree

OS/390 Workload Manager

OS/390 Workload Manager, introduced at the same time as Parallel Sysplex, is a key part of the overall magic of OS/390 dynamic resource balancing. Workload Manager works with the S/390 subsystems and the OS/390 Communications Server to *dynamically balance diverse workloads at both the OS/390 image level and across the S/390 Parallel Sysplex.* Using real-life business goals, WLM continuously adapts systems resources within an LPAR to actual workload demands, without human intervention within a S/390 Parallel Sysplex. Be sure to check out our articles about WLM for more information.

S/390 G5/G6 Server (Capacity Backup and Capacity Upgrade on Demand) CBU and CUoD

S/390 G5/G6 Server (Capacity Backup and Capacity Upgrade on Demand) CBU and CUoD provide outstanding flexibility to add capacity at short notice when you need it, with *no* outage to the running applications. This technology provides dynamics capabilities to upgrade your server from as little as 1 processor all the way up to 12 processors, depending on your business need. Have the capacity available for when you need it, but only pay for it when you activate it! Together with S/390 Parallel Sysplex, these features are important ingredients in your availability toolkit. For customers with the highest availability requirements, CBU support is integrated into GDPS™, providing the ability for automated CPC upgrades in case of a disaster. See Frank Kyne's article on this topic for more information.

S/390 Dynamic I/O Management in the Channel Subsystem

S/390 I/O instrumentation data allows the OS/390 I/O Supervisor and Workload Manager components to gather I/O performance statistics and contention data on all the systems in the sysplex. This information is used to determine if I/O bottlenecks are contributing to the reason for a workload to fall short of its goals. WLM uses this data with the Enterprise Storage Server™ (Shark) to adjust I/O priorities and PAV-alias assignments across the sysplex to help the workload achieve the workload goals while reducing the specialized skills required of the customer. See Harry Yudenfriend's article for more details.

Communication Server for OS/390's dynamic session balancing for both SNA and TCP/IP

Communication Server for OS/390 uses the sysplex coupling technology to assist it in providing session balancing across replicated applications within a Parallel Sysplex according to goals defined within the S/390 work load manager. Dynamic session balancing is available in both SNA and IP environments. The article in the newsletter will cover the aspects of this dynamic session balancing in an SNA environment and take you through the many IP options you may have. Finally, this article will give information on some new and exciting functions, Sysplex Distributor, etc. in V2R10 that will enhance the dynamic capabilities of an IP environment and allow Policy and QoS factors to be brought into the distribution decision. See Mac Devine's article on SNA and TCP/IP for more scoop.

S/390 Parallel Sysplex Dynamic workload balancing

Parallel Sysplex technology enables multiple S/390 systems to behave as a single logical resource image to customer applications. OS/390 enhanced components, such as Communications (XCF), workload management (WLM), locking, logging, automatic restart, are all examples of components that assist in enabling multiple OS/390 systems to function and appear as one single logical resource. These components along with Communications server for OS/390 and the subsystems like IMS®, DB2, CICS, along with a growing portfolio of new middleware vendors (SAP® R3, Oracle®, PeopleSoft®, etc.) enables work requests associated with a single workload to be dynamically distributed for parallel execution across nodes in the Sysplex cluster based on available processor capacity anywhere in the Parallel Sysplex. The power of the Parallel Sysplex comes in its transparency to business applications. Refer to the articles on the "Life of a transaction," clustering, and the Parallel Sysplex overview in the Supplement.

IBM® is one of the very few vendors that possesses it all: hardware, software, middle ware: all working in concert to provide the best possible solution for today's business applications .

The life of a transaction

Angelo Corridori

Angelo currently works in the Parallel Sysplex PDT where he is responsible for Parallel Sysplex requirements and plans. He has been involved with Parallel Sysplex since its inception.

Background

Resource Management - we do it every day. In response to ever-changing circumstances that surround us, we manage resources to achieve maximum benefit from their deployment. Whether it is time, money, equipment, or other resources, we change the use of resources in response to new information, requirements, and situations as they occur. For example, no matter how well we have planned our daily calendar, a "beep" cell phone call, or brief conversation with a manager in the hallway, can quickly change the day's planned activities. In short, most people employ dynamic resource management in their daily lives.

Can the same be said for computer systems that manage thousands of end users and millions of work requests every day? Surprisingly, the answer is "no!" In many cases, workloads are assigned to servers in a static fashion. If a server is overloaded, work suffers because new work requests cannot be easily routed to another server that is less busy (or resources cannot easily be moved to the overloaded server). And, if a server is underutilized, resources are wasted, as there is usually no means to move more work or other work to the server to take advantage of the excess resources. Imagine the consequences if we ran our daily lives this way ..... "I always go to the gym at 5PM - yes, even if there's a major earthquake today, I'm going to the gym!"

S/390 and OS/390 are the rare exception to static resource management. Over the years, as Parallel Sysplex technology has matured, OS/390 and S/390 have become tightly integrated and very adept at dynamic resource management. Resource management is the assignment of resources (CPs, storage, I/O, peripherals, etc.) to the consumers of resources (web requests, end user sessions, batch jobs, transactions, complex queries, object requests, logical partitions or servers, etc.). While dynamic resource management can be achieved within a single OS/390 image, dynamic resource management is even more effective and flexible when used in a Parallel Sysplex with two or more closely cooperating OS/390 images.

Dynamic Resource Management

Let's take a look at some examples of S/390 and OS/390 dynamic resource management capabilities. When an end user wants to establish a new session, which of several possible servers should be chosen to accept the new session? OS/390 has a Workload Manager (WLM) component that understands the state of servers in the Parallel Sysplex and their responsiveness. The networking component of OS/390, the Communication Server, has function called generic resource (for SNA sessions) and WLM DNS Balancing (for TCP/IP sessions) that can consult with WLM to determine which server is best positioned to accept a new session at the time the session is requested. The generic resource function is supported by the major subsystems such as CICS, IMS/TM, TSO, and APPC, and WLM/DNS balancing is supported by CICS, DB2, and TN3270 so that sessions for these subsystems can be seamlessly and transparently (to the end user) established with a Parallel Sysplex.

Once a session has been established, the end user will submit work requests for processing. The work request might be a classic transaction (for example, show a checking account balance) that needs to run quickly. Just as with the establishment of a session, WLM, in conjunction with subsystems like CICS and IMS that support dynamic transaction routing, can make a decision as to where the transaction will receive the most responsive processing at the time the transaction is initiated. CICS uses dynamic transaction routing algorithms in the CICSplex System Manager (CPSM) product to route work to an application owning region (AOR). CPSM can route work based on a relatively simple algorithm such as "route to AOR with shortest queue". CPSM can also be instructed to consult with WLM for a recommendation as to which AOR should to chosen to run the transaction to most likely meet the transaction response time objective. In most cases, work units run on the server where they are submitted. If need be, CICS transactions can be transparently routed to the server and application instance with the resources that can best service them. In the case of IMS transactions, IMS provide a means to gather the incoming transactions on a central work queue. Then, when application regions are free to process new transactions, the application can select a new transaction from the central queue, thus implicitly balancing work, and thereby sharing resources like processing capability, within the Parallel Sysplex configuration.

Once a transaction is running, it will need to access data. The database managers, IMS/DB, DB2 and VSAM, have the ability to dynamically manage access to the data which resides in a shared database. Sharing data provides numerous benefits to the end user of a Parallel Sysplex, most notably near continuous availability.

In order to share data effectively, with good performance and integrity, the database managers must effectively manage the data under their control. This is done by using global locks and managing buffers using the coupling facility technology. Global locks, representing data in the database, are recorded in the coupling facility to allow other transactions that may want to access the data a quick way to tell if the data is in use. For fine granularity data, an individual data item (for example, a database record) will not be in use most of the time. When data is updated in the memory of a local processor, the database manager, in conjunction with the coupling facility, ensures that any other local copies of the data that may reside in local processor memory within the Parallel Sysplex are marked as being "down level". This is done to ensure that subsequent updates are applied to the most recent copy of the data.

Traditional transaction processing workloads often have related batch workloads. So, let's look also at an end user who submits a job that has unique resource requirements like four tape drives and a software license. OS/390 not only has the ability to recognize and manage real resources, but it can also manage abstract resources as well. An abstract resource is anything the installation wants to define it to be, and in the case of our batch job it might be a license to run a program. That license may exist on only one server in the Parallel Sysplex. If a job has been submitted that needs an abstract resource, such as the license, WLM can ensure it gets to the server where the resource exists. This takes the burden of directing the job to the appropriate server away from the end user who submitted the job, or the operations and system programming staff. As for the tapes requested by the job, they can be assigned from a pool of devices shared, and dynamically allocated, among servers. Finally, one of the key resources needed by a job to run, an initiator, can itself be managed dynamically by WLM as workloads of various job classes start and stop. The overall effect is to balance the batch workload within the Parallel Sysplex configuration.

Work units that are large and can consume large quantities of resource are especially troublesome in terms of their management. A complex query is a good example of a work request that can consume a large percentage of processor, storage and I/O resources. OS/390 and DB2 have the ability to break a long running complex query into several smaller pieces that can run in parallel, thus reducing the overall elapsed time for the query to complete. But when this one large work unit spawns many smaller work units that may run on the same or other servers, how are the resources required by the query allocated and managed? What is the impact to other work running in the Parallel Sysplex? OS/390 and WLM have functions for exactly this type of work - enclaves. Enclaves allow the multiple independent work units to be treated as related work units. Further, the importance and responsiveness of such large long running units of work, as represented by its many subcomponents, can be established so that the hardware and software can effectively allocate resource to "the query" while avoiding negative impact to other work in the system.


Paul Rowntree

Finally, let's examine resource management associated with e-business and a Web application and/or server. Web workloads are notoriously erratic in terms of resource consumption patterns. It is extremely difficult to predict the workload peaks and valleys associated with what could be an essentially undefined end user population. This makes effective and dynamic workload management all the more important. When a surge of Web work arrives, you need the ability to apply resources (capacity, storage, web servers, etc.) to service it as it arrives. If there is a lull in Web related activity, you want the resources deployed to other productive tasks, not sitting idle. These are exactly the kinds of capabilities S/390 and OS/390 bring to the world of Web serving.

The examples above are based on the ability to move work to a resource available in the Parallel Sysplex configuration. OS/390 and S/390 hardware also have the unique capability to move resources to the workload that needs them. In particular, the S/390 Processor Resource /System Manager, PR/SM, has had the ability to share CPs between logical partitions (or OS/390 images which can be thought of as logical servers) for many years. PR/SM allows fine levels of granularity in the allocation of resources such as CPU, I/O, and storage. Resource management on a single processor complements the multi-system movement of work to resources to make the OS/390 software and S/390 hardware combination the undisputed champion of dynamic, flexible, resource management.

Conclusion

Why is resource management so important? For inexpensive resources or those available in abundant quantity, it is not important. But the computing resources (processors, peripherals, data, applications, etc.) of a business usually represent a large investment. And collectively, in more and more instances, they are the business. So, effective management makes good business sense. But, why is dynamic resource management important? Computing resources have been managed by people since their inception. However, in today's skill constrained environment and with ever escalating people costs, it makes good economic sense to turn over the resource management of computing systems to the computers. Even beyond economics, people just cannot do effective resource management in a timely manner. Even the simplest processors are capable of millions of operations per second. People just cannot react quickly enough to unfolding scenarios of resource imbalances and shifting priorities. In short, you need to manage computer resources at computer speeds, not people speed.

S/390 and OS/390 have evolved their cooperative dynamic management of resources to the point where they can provide unprecedented benefits to businesses. A business can run multiple competing workloads in a single compute resource "pool" secure in the knowledge that multiple competing priorities and responsiveness requirements will be honored from end user through the network, application, and operating system right down to an individual I/O operation. When investing in S/390 computing equipment, a business can feel secure knowing that they will be getting maximum utilization from those resources as they are managed on a minute by minute basis. A properly configured Parallel Sysplex has the ability to dynamically manage the never ending ebb and flow of workload variations. And all of this, with less burden on the people, operators and systems programmers, responsible for the computing environment.

PR/SM - Taking the myth out of logical partitioning

Jeff Kubala

Jeff is a Senior Technical Staff Member who has been with IBM for more than 19 years. He is currently the lead designer for the PR/SM team.

Mergers, acquisitions, ISPs, ASPs, e-business. These are all driving constant change in workload requirements on company IT departments. Wouldn't it be great if your enterprise server could be partitioned into many logical servers to *effectively* run all of your heterogeneous workloads? What if you could consolidate workloads from multiple servers onto a single platform and still maintain isolation and independence? What if...

There is a lot of discussion these days about the dawning of "new" ideas such as logical partitioning: "Split your server into multiple, smaller servers." However, many implementations of the "new dawn", are really a *physical* partitioning of servers; allocating memory, processors and I/O in coarse increments to workloads with no sharing of the resources. This allows for consolidation of footprints but gives no real benefits in terms of capacity management of the workloads in aggregate. You need to purchase total, physical capacity to handle workload peaks for all of your workloads.

Fortunately, there is a product that delivers on the promise of effective consolidation and management of workloads IBM's *Processor Resource/Systems Manager* (PR/SM). PR/SM supports the creation of up to 15 logical partitions today. Each of these logical partitions is fully capable of running an operating system independently. OS/390 (with or without Parallel Sysplex), VM®, TPF, UNIX® applications under US system services, and, yes, Linux® are all supported in a logical partition.

PR/SM allows granular levels of allocation of resources. Storage can be assigned to logical partitions in increments on the order of megabytes (1-64) rather than being on the order of gigabytes. I/O channel paths can be shared by logical partitions to fully utilize the bandwidth of ESCON® and Fibre channels. PR/SM's dynamic reconfiguration capability offers the opportunity to allocate additional resources to partitions with demanding workloads without a disruption to the workloads in the logical partitions.

Paul Rowntree

The real power of PR/SM is in its management of central processor (CP) resources. Although CPs can be dedicated to logical partitions like all the partitioners-come-lately, that's not really interesting. The real power is in the effective use of *shared processors*. With shared processors, when a workload in one partition goes idle, the available processing time will automatically be redistributed to other partitions with no intervention.

Logical partition processor weights are a priority policy for logical partitions based on your dispatching priorities, and can be specified when allocating your machine's resources to your workloads. Using processor weights with shared processors, you can use sub-processor granularity for allocating your machine resources to your workloads. Each logical partition is treated as a separate workload that is managed against the processor weight policy. Processor weights define relative priorities of logical partitions for determining which logical partition gets the resource when there is *contention* for the resource. When there is no contention for the resource (i.e. all logical partitions are not, at the moment, trying to use all the CP resource they are entitled to), the other logical partitions will *automatically* fill that "white space" and use that excess capacity. Even the redistribution of the "white space" is done in accordance with the processor weight policy. As soon as the logical partition that was not trying to use all of its entitled resources needs to use resources again, the resource moves back to it in accordance with the processor weights. The policy can be dynamically updated with changes taking effect immediately.

OS/390 running in a logical partition gives yet another level of heterogeneous workload management that is world class. Multiple workloads within an OS/390 logical partition are given even a finer granularity of workload management via the Workload Manager (WLM) component of OS/390. The priority of the logical partition with respect to other logical partitions is managed by PR/SM. The priority of the individual workloads/applications

within the logical partition is managed by WLM. This capability is unmatched in the industry outside of the S/390 platform. The only thing that could be better is capitalizing on the synergy of WLM and PR/SM in managing priority of individual workloads across logical partitions.

But what does this buy me? Isolated workloads that are now sitting on small boxes, surrounding my corporate database server, running only 20-50% utilized can be brought onto the S/390 enterprise server in a logical partition and use shared excess capacity of the S/390 to support multiple workloads. The S/390 box can be effectively driven and utilized at upwards of 90-95%. The natural peaks and valleys of diverse workloads can be taken advantage of to move all your work through, according to your priorities, without cluttering the floor!

Need more processor capacity? The IBM S/390 enterprise server can be upgraded dynamically and the processor resources are automatically used to back the logical partitions that are sharing CPs. If more logical CPs are needed in a logical partition to exploit this additional capacity, these can be configured online dynamically as well.

The best part of the PR/SM feature? It's probably in your IT shop right now. It's a standard feature of all IBM S/390 enterprise servers. With a wealth of experience over years of continuous refinements, PR/SM offers the premier logical partitioning solution for your IT needs.

Get into goal mode! An overview of Workload Management

Steve Hamilton and John Arwe

Steve is an Advisory Information Developer in the OS/390 organization. He has supported Workload Management since 1996. Steve is also an Edgar-award winning mystery author.

John, a Senior Software Engineer in Poughkeepsie, NY, has been a key member of the Workload Management team since 1990. He is a frequent speaker at SHARE, CMG, IDUG, and the MVS Expo.

Managing system performance is like managing an electric utility: nobody even thinks about it until they lose power. It's easy to think everything is fine when resources are plentiful — the real test comes when everybody turns on their air conditioners at once — or to put it into I/S terms, when Marketing starts a new promotion and suddenly triples the number of hits per minute on your web site.

Of course, your servers don't just serve Web pages. They also perform back-end fulfillment functions, generate reports, and support other parts of your business. Are your systems prepared to concentrate valuable resources on the revenue-generating functions when traffic spikes? Or will they doggedly continue to generate reports rather than process new orders?

Workload Manager (WLM), a component of OS/390, lets you run all of this work concurrently, allocating the resources to the most important work first. When traffic gets heavy, your important work will still get what it needs, while the other work gets what's left.

When you run OS/390 in goal mode, you define goals for each type of work. WLM will use these goals to determine how much resource (CPU, storage, and I/O) should be given to that work to achieve its goal. WLM constantly monitors the system, automatically adjusting the resource allocation as necessary.

Workload Manager allows you to categorize your system work into separate "service classes." For each service class, you assign a performance goal. These goals are developed based upon real life service objectives, similar to the objectives found in service contracts. Each performance goal has an importance level and a type. The three types of goals are:

Paul Rowntree

- **Response Time** — This can be expressed in one of two ways: either "Average Response Time" (for example, all transactions in this service class should complete in an average of .5 seconds), or "Response Time and Percentile" (for example, 80% of all transac tions in this service class should complete within .5 seconds).

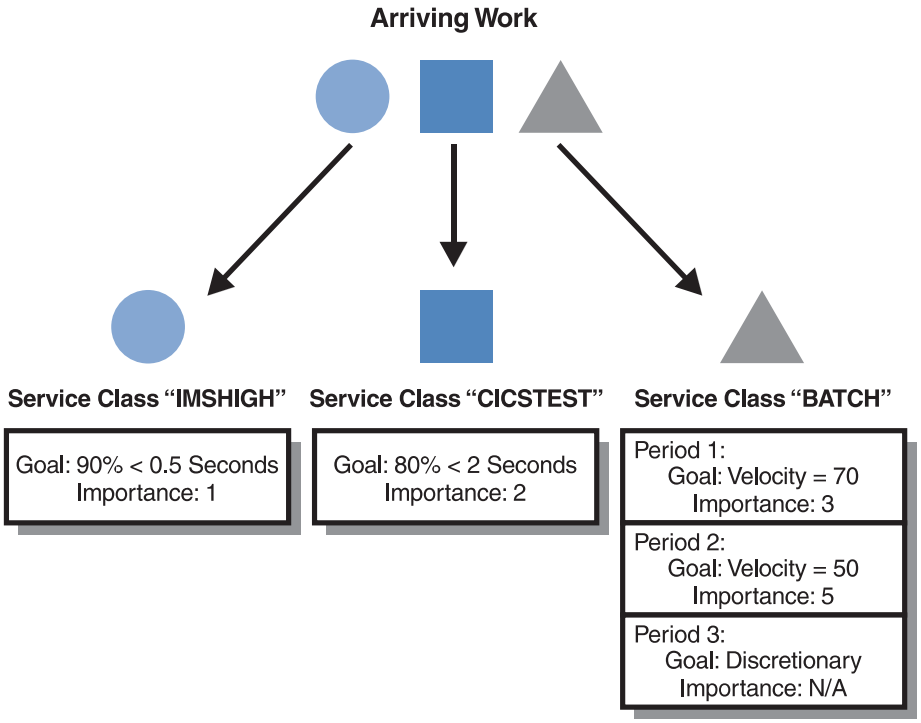
- **Discretionary** — You assign this goal to work that can be run whenever the system has extra resources to spare. These could be low-importance batch jobs, for instance, that run when only when the system load is light and all other work is meeting its goals.

- **Execution Velocity** — Because response time is not an appropriate measure of performance for all transactions, especially long-running transactions with no definite beginning and end, a formula was defined to measure a transaction's "velocity." It's a number between 1 and 99, and the best way to think of it is as a measure of how much time a transaction spends waiting for system resources (either CPU, storage, or I/O services). A transaction with an execution velocity of 90 is moving along extremely well, rarely waiting for system resources. A transaction with an execution velocity of 10 is moving slowly, spending most of its time in a dead stop, waiting for resources.

  The importance level is a number between 1 and 5, representing how important it is for work in that service class to meet its goal. For a service class with an importance level of 1, Workload Management will make every attempt to ensure that the work is meeting its goal. Then it will take care of the importance level 2 work, and so on down to level 5. A service class with a discretionary goal is not assigned an importance level at all — you can think of it has having an importance level of 6.

A service class can also be broken up into more than one "service period." As certain work consumes more and more resources, it may "age" from one period into another, at which point it can be assigned a new goal and a new importance level. For instance, in a service class

to which batch jobs are assigned, you could have a velocity goal of 50 and an importance level of 3 for the first 2500 Service Units (the units of time that Workload Management uses to define periods). Work that takes longer than this would go into a second period, with a velocity goal of 20 and an importance level of 5. You could put a time limit on this period, as well, and define a third period with a discretionary goal (and of course no importance level at all). In this way you can give work less stringent goals, and less importance, as it takes longer and longer to execute.



WLM collects performance data for each service class four times per second. At least once every ten seconds, WLM will make resource adjustments, as necessary. If a service class is "underachieving" its goal, WLM will attempt to give that service class more of the specific resource it needs: CPU, storage, or I/O services. As several service classes may be competing for the same resource, however, WLM must perform a constant balancing act, making the appropriate trade-offs based on the business importance and goals of the different types of work.

All of the service classes you define, along with the performance goals, periods, and importance levels, are part of the "service definition" you define for your entire sysplex. As part of that service definition, you define one or more "service policies." A service policy lets you dynamically change the performance goals for your service classes, without having to reinstall the entire service definition.

You may have one service policy in effect for the nine to five workday, for instance, with aggressive response time goals defined for your high-importance CICS work. You can activate a new policy at the end of the business day, with adjusted goals for offshift work — and yet another policy on the weekends.

Once you've defined your service definition, you need to install it on the WLM couple data set, activate a service policy, and then you can switch one or more systems to goal mode!

There's plenty of help available to you when you're defining your service definition for the first time. The Goal Mode Migration Aid, for instance, is a semi-automated tool to help you migrate your system performance goals from compatibility mode to a service definition for goal mode. It does this by:

- Helping you select performance data for the migration process by analyzing RMF data.
- Mapping your existing performance definitions from the IPS/ICS parmlib members, as closely as possible, to a service definition.
- Maintaining your reporting structure by generating report service classes for all report and control performance groups in your ICS definition.
- Assisting you in documenting your migration to goal mode.
- Supporting the creation of service classes for CICS and IMS transactions.
- Creating a service definition on your OS/390 system which can be immediately examined, corrected, enhanced, and then activated from the WLM administrative application.

You can download this tool, and find other valuable information on the WLM Website:

www.ibm.com/s390/wlm

Top ten benefits of Workload Manager in goal mode

By Kris Perry

Kris has been part of the Parallel Sysplex team since its announcement. She focuses on technical marketing and support programs to assist the field and customers in implementing new technologies.

Customers tell us they went to goal mode because of the number of jobs, the complexity of the systems and constantly changing demands were outstripping their ability to manage their system resources in any other fashion. While the reasons for migrating to goal mode may differ on a case by case basis, the benefits can be universal:

1. Real Life Goals - Performance goals are specified in business terms rather than technical parameters.

2. Improved Workload Balancing - The control of the system changes as the units of work change, ensuring that work receives appropriate resources to meet its goals.

3. Better Control of Unpredictable Workloads - This is particularly valuable for new web based workloads.

4. Better Control of Inconsistent Workloads - WLM manages peaks and valleys to meet business objectives.

5. Consistent Response Times - Online systems maintain a consistent response time and high quality of service, even while responding to complex, resource consuming transactions such as DB2 queries.

6. Easier Management - Simplified external definitions, improved LPAR management, standardization by using a single policy, hands off monitoring.

7. Better Information - Detailed information on how workloads are performing and meeting their objectives. Problems or bottlenecks are identified. Clearer indication of SLA attainment. Data is available to support capacity management.

8. Better Performance and Improved batch throughput - WLM goal mode drives the systems harder through dynamic adjustments.

9. Speed! - WLM goal mode is faster than any systems programmer.

10. The Future! - Goal mode will continue to be supported and enhanced to match the technology innovations in the future, and to deliver to the customer the most complete and efficient means to manage their systems resources - Workload Manager in goal mode.

Paul Rowntree

The goal mode experience

Cheryl Watson

Cheryl Watson, president of Watson & Walker, Inc., is an internationally known consultant who has worked with IBM mainframes since 1965. Since 1991 she has been the author of Cheryl Watson's TUNING Letter, a journal of MVS ®and OS/390 advice now read by over 10,000 analysts around the world. She was recently given the prestigious A. A. Michelson Award for technical excellence and professional contribution. She has served both as a Director of the Computer Measurement Group's national organization and as chair of the SHARE Enterprise Wide Capacity and Performance group, and is a popular speaker at these user conferences. Her articles have appeared in such publications such as Enterprise Systems Journal, Technical Support Magazine and CMG Proceedings.

Now that, from my observations, more than 50% of OS/390 sites are running at least one goal mode system, we can look at the experiences that people had while making that migration. Much of this article is based on information I received from a survey that I conducted in November. You can see the full survey and results at **www.watsonwalker.com/archives.html** in Cheryl's List #33.

Survey Results

Interestingly enough, the survey indicated that almost half the sites which chose not go to goal mode did so because of lack of staff or time. But of the sites that did migrate to goal mode, 46% made the migration within two weeks. To get the first system into goal mode:

18% - took less than a week; some said hours or one day (especially when using my Quickstart Policy!)
13% - one week
16% - >1 to 2 weeks
26% - >2 to 4 weeks
11% - >4 to 8 weeks
17% - >2 months to 12 months

I don't think that a week or two is too much time to expend in order to gain the benefits of goal mode (see below). Some people said that only a few hours after they downloaded my Quickstart Policy they were running goal mode on their test system. (Most respondents used my Quickstart Policy for their migration. See 'Tools' below.) Once you have one test policy working, the production systems are fairly easy. To get the next system into goal mode:

45% - took less than one week; most were hours, not days
28% - 1 to 2 weeks
11% - >2 to 4 weeks
15% - over 4 weeks

According to the survey, the most time-consuming part of goal mode conversion is the effort to change reporting programs and capacity planning systems (see my note below). Once a test system is in goal mode, however, you can then take your time about getting the reporting systems working well. During that time you can also learn more about goal mode.

Some sites had to back out of goal mode temporarily, but in most cases it was because they had not set their goals correctly or had not applied current maintenance. Both are relatively easy problems to resolve. See the "Protecting your 'critical regions'" for more information.

What People Like About Goal Mode

People saw many different benefits of goal mode. Here are some replies, in descending frequency, from our survey question "what is the largest benefit that goal mode gives you?"

- Takes less management time than IPS/ICS.
- It's easier to add new work.
- Manages the system better than compat mode.
- Better performance and turn around. For one site, the night batch completed 30 to 60 minutes earlier. For other sites, the ability to maintain good online performance while still pushing batch work through the system was most important.
- Better throughput (as one reader says "pumps CICS transactions through"; another said that they regained 10% of the processor).
- Provides better and more consis tent response times for online transactions (CICS, IMS, TSO).
- Easier to understand what's going on in the system (i.e. who is meeting goals, who is not, why they aren't). Goal mode provides much more information on why workloads are being delayed. For example, if you aren't able to achieve sub-second response in CICS, you can tell from WLM statistics whether the problem is due to CPU, storage, I/O, or higher importance work. I believe this is the most significant reason to move to goal mode. I person ally would go to goal mode just for the improved reporting and statistics available!
- Ability to specify business goals rather than technical IPS parameters.
- Better workload balancing. Most sites find that WLM balances the workload across multiple systems than they could manage manually or with auto mation products. Batch initiator management provides a big part of that balancing, but people are also finding excellent balancing occurring with their online systems as well.
- Better able to manage over-committed systems and perfor mance spikes. The ability of WLM to adjust working set sizes and dispatch priorities dynami cally allows more work to complete on the system. Most people are surprised (and happy) to find that WLM can respond so quickly to spikes

of unexpected workload activity.
- Resource groups allow better management. WLM resource capping provides an excellent mechanism to either guarantee a certain amount of capacity to loved ones or to restrict resource hogs.
- Managing many LPARs is much easier in goal mode.
- Standardization by using a single policy.
- Much better management of DDF enclaves.
- More flexible classification.
- Necessary for new subsystems. Many of the new workloads, such as Webserver and Corba, need to respond to radical changes in capacity requirements. They require WLM goal mode as a method to manage these spikes of resource need. Therefore, using goal mode when implementing these new workloads is extremely important. We can expect more new workloads in the future to require goal mode.

What Problems Did People Have?

There are some additional considerations you want to be aware of before going to goal mode. Over 20% of the sites indicated that they had NO problems at all going to goal mode. When asked "What was the largest problem in going to goal mode?" here were the responses.

Conversion of capacity planning and performance reports which were based on performance groups (this was the most common response and the item that takes the most time). If you have little reporting, then this won't be a major problem. If you have daily or weekly reports based on performance groups, you'll need to change them to base off service classes and report classes. The biggest problem you'll have is converting monthly or yearly trend charts that are now based on performance groups. Installations have taken one of two approaches, with the first one being weeks easier!

1. Start fresh. Dispense with history and start from scratch. This is especially useful when there is no easy comparison between PGNs and service classes. I'd recommend this method if at all possible.
2. Change programs to convert PGNs to SCs or vice versa. Perhaps the easiest way to handle this is to ensure that there is a report class for every PGN. Then you can easily change the program to continue tracking your old PGNs, but by using report class data instead. There are a few differences with report classes, however.

Paul Rowntree

Check with your software vendors. Some current reporting products provide a simple way of producing trend reports with combinations of performance groups and service classes.

- Political resistance and fear of the unknown (and this was the second most common response).
- Paranoia (every problem became a goal mode issue) - another very common response.
- Operator and user training needed. (I personally believe this is fairly minor.)
- DB2 distributed work needed fine-tuning (i.e. getting definitions for DDF correct).
- Memory and CPU thrashing (early SP 5 systems; but now resolved).
- Problems with third party products (primarily for early users of goal mode, although a few people complained about the inability of modeling programs to use report classes).
- Figuring out how to define CICS and IMS. (My recommendation is to start with velocity goals - they're easier to figure out.)
  - Understanding WLM (classes or conferences help out here).
  - Trying to merge many (some times 50) different IPS/ICS members into a single policy.
  - Converting auto ops packages to use service classes rather than PGNs.
  - Lower importance work receiving higher DP than higher importance work (resolved with PTFs).
  - Didn't like maximum of 25 service class periods with non-discretionary goals.
  - Amount of planning time.
  - Needing to reformat couple data set with new releases (this is required only if you want to exploit new facilities).
- Determining proper velocities.

Tools

If you haven't made the move to goal mode, here are two tools that may reduce your migration time.

Quickstart Policy

Because there were no conversion programs available in the early days of goal mode, I proposed and wrote about a generic goal mode policy, called my Quickstart Policy. It's a policy that I believe will work in most installations, with some slight modification (e.g. defining test batch classes and the region names of your online systems).

A description of the policy can be found on our Web site at **www.watsonwalker.com/qsp.html**. I wrote the original article in 1995 and then modified it in 1999. At the time of the modification, I also created a downloadable policy that you can start with (it's at the same Web site). Over 1400 people have downloaded the policy and many have said that they were running in goal mode on their test system the next day. You can too!

RMF Migration Tool

Robert Vaupel of IBM OS/390 Development was helping customers and fellow IBMers migrate to goal mode. He found that the biggest time element for many sites was entering their large ICS members into the WLM panels. It takes a LOT of time to enter a six-page ICS one line at a time into the panels. So Robert wrote a remarkable semi-automated conversion tool to help RMF™ customers get to goal mode even more quickly. This Goal Mode Migration Aid is distributed as a free, "AS-IS", tool on the RMF Web site, but does not have any official support. Robert provides support as he can, although he says it hasn't been a problem keeping up.

Robert uses the CPU and Workload Activity reports generated by the RMF Postprocessor. The data is downloaded to a PC and converted to spreadsheets with the help of the RMF Spreadsheet Reporter. The Goal Mode Migration Aid tool reads the converted RMF reports in order to define initial goals for corresponding service classes. The output of the tool is a PDS that can be brought into the WLM application as a ready policy. Refer to "Get into goal mode" for more information about this tool and its Web site.

You can obtain the instructions and the tool (along with the downloadable RMF Spreadsheet Reporter) at **www.s390.ibm.com/rmf**. Click on Tools.

Summary

So should you convert to goal mode? Obviously, my recommendation is that you do. It often takes only a few days to create a policy to run on your test system, and then you can migrate to production when you feel comfortable with it.

Here are some of my recommendations:

- Learn about goal mode. You can get this information from published articles, from the description of my Quickstart Policy (and a quick checklist for creating a monoplex) on our Web site, reading the IBM WLM: Planning manual (GC28-1761), reading the IBM Redbook: OS/390 WLM Implementation and Exploitation (SG24-5326), taking a class on WLM, or attending user conference presentations on WLM.
- Apply maintenance before starting; especially for relatively new WLM facilities, such as batch initiator management.
- Plan for the time to change your reporting programs to use service or report classes rather than performance groups. This is the most time-consuming step.
- Understand that you'll never be able to thoroughly test your policy on a test system. As an example, response goals are only effective when you have a high enough volume for WLM to manage using the goals. (On my single-user P/390 system, every goal is a velocity goal because I don't generate the minimum volume of 3 transactions every 20 minutes!)
- I now recommend that people start in goal mode by using velocities for their online systems (CICS and IMS) rather than transaction goals. It's much easier to convert to goal mode using velocities, and you can later take the time to figure out the more difficult transaction response goals.
- Expect that every small problem that occurs as soon as you go to goal mode will be seen as "your" fault. It will take time to show that the problem isn't a goal mode problem.

Finally, get ready to REALLY enjoy better control and the wealth of new data that goal mode makes available to you!

Protecting your "critical regions" when running goal mode

Steve Hamilton and Gail Whistance
Gail is an Advisory Programmer on the OS/390 Workload Manager team responsible for algorithm verification of new releases and for customer support. She often presents at customer conferences on WLM topics. She joined IBM in 1972, and has held various positions in programming and management. For more information about Steve, see the article titled "Get into Goal Mode!"

OS/390 customers have identified several concerns that have made them reluctant to run their production systems in WLM goal mode. These concerns are particularly serious for customers running critical CICS and IMS regions. Here are some specific steps IBM is taking to help protect those regions:

Long-Term Storage Protection

If a CICS or IMS production region experiences a period of low activity, WLM may stop managing it as a server. At this point, the region's pages are more likely to be stolen by competing workloads. By assigning long-term storage protection to these regions, you can make sure that this won't happen. WLM will take away storage only for work of greater importance, or for work of equal importance that needs the storage "more."

This long-term storage protection can be assigned to a specific region, or to a transaction service class. By protecting a transaction service class, you ensure that any region serving that transaction service class will inherit the long-term storage protection.

Long-Term CPU Protection

For critical CICS or IMS production regions, you can make sure that work of lower importance will never interfere with that region's access to CPU resources. By assigning long-term CPU protection, you guarantee that work of lesser importance will always have a lower dispatch priority.

You can assign long-term CPU protection to a service class — if you protect a CICS or IMS transaction service class, then any region serving that transaction service class will inherit the long-term CPU protection.

Exemption from Transaction Server Management

Normally, a CICS or IMS region will be managed according to the response time goals of the transactions it is processing. In certain situations, however, you may wish to exempt one or more regions from this type of management. You may wish to migrate one region at a time, for instance. Or you may simply wish to run low-activity test regions with velocity goals instead of response time goals. When you declare a region exempt from transaction server management, it will no longer be managed to the response times of the transactions. Instead, it will be managed to whatever performance goal you specify for the service class assigned to that region.


Paul Rowntree

Assignment of Goals Based on System Name

Customers have long expressed their desire to assign performance goals based simply on the name of the system in which the work is executing. IBM is responding with new qualifiers for both system name and system name group. If you are consolidating workloads into one sysplex, for example, it may be difficult to classify and separate similar kinds of work. By simply running one workload on one system and another workload on another system, you can use the system name qualifier to make sure they have different goals.

While we're at it, here are three more new qualifiers to help you assign performance goals: sysplex name, scheduling environment name, and subsystem collection name. For JES2 and JES3, the subsystem collection name is the XCF group name.

Capacity Backup and Capacity Upgrade on Demand

Frank Kyne

Frank is a project leader in the ITSO in Poghkeepsie, NY, responsible for production of all Parellel Sysplex-related Redbooks. He has worked for the ITSO for two years, prior to which he was a system programmer with IBM Ireland, supporting both IBM's own data center and those of customers.

Two of the challenges faced by companies in the modern business world are having the capacity to handle volatile workloads without buying more capacity than currently needed, and being able to change the available capacity with little or no impact to the running applications. Traditionally, there were two ways to add capacity to an existing environment:

- Upgrade an existing processor, however this required a planned outage and would impact applica tions that did not support data sharing.
- If you have a Parallel Sysplex, add an additional processor to the sysplex. This was potentially less disruptive than upgrading one of the existing processors, but there is a limit to how many processors you want to have.

One example of how quickly your capacity requirements can change is the recent activity on the stock market. One S/390 customer decided, based on news reports over the weekend, that additional capacity was going to be required to handle the volume of trading on the Monday morning. Fortunately for that customer, they had an IBM 9672 processor that supported Capacity Upgrade on Demand and Capacity Backup. Using these features, they were able to add the additional capacity non-disruptively and were ready for the additional workload when the markets opened on Monday morning.

Capacity Upgrade on Demand and Capacity Backup are two of the many features announced for the IBM G5 and G6 9672 processors that build on the S/390 and Parallel Sysplex characteristics of flexibility, availability, and non-disruptive growth. In this article, we will compare and explain these features, describe what they can do for you, and provide some implementation information.

The first thing to understand is that Capacity Upgrade on Demand and Capacity Backup are *not* the same thing - they are related, but not the same. Capacity Upgrade on Demand (CUoD) is a capability in G5 and G6 9672 Central Processing Complexes (CPCs) at current microcode levels to *non-disruptively* add new Central Processors (CPs) to an existing configuration. Those new CPs can be added permanently via an MES or temporarily using Capacity Backup. There is no charge for having the CUoD capability on a CPC.

Paul Rowntree

Capacity Backup (CBU) is a chargeable feature that allows you to quickly add additional CPs to an existing CPC, for a period of up to 90 days. CBU is aimed primarily at disaster recovery situations, where you have a smaller CPC available that you need to temporarily upgrade to take on additional workloads. It is possible to upgrade a CPC in as little as 10 minutes using CBU. In fact, IBM's GDPS offering has been specifically enhanced to automatically make use of this facility in a failover situation. A description of GDPS is included in the Supplement.

The other option for upgrading a CPC is to order a processor upgrade in the usual manner. Once the upgrade has been agreed and contracts signed, an MES will be shipped to the customer. In most cases, the customer engineer can install the MES non-disruptively using the capabilities of CUoD. The main differences between using an MES and CBU are:

- Elapsed time - a CBU upgrade can be in place in minutes. An MES has order, process, delivery, and installation time to consider which may span several days.
- Permanence - a CBU upgrade is intended to be a temporary upgrade, whereas an MES upgrade is permanent.

The crucial advantage that *CUoD* provides is the ability to upgrade a CPC without having to schedule an outage. As availability requirements relentlessly increase, it is becoming more and more difficult to schedule planned outages. CUoD adds CPC upgrades to the list of changes you can implement without requiring an outage.

The advantage that *CBU* provides is speed. If you have a disaster, which could be the loss of a site or just the loss of a single CPC, it is vital to replace the lost CPC capacity as quickly as possible. CBU provides you with this capability, and allows you to add CPs to a CPC without even having to involve any IBM personnel.

The CBU feature can be ordered on all on G5 and G6 9672s that have the current microcode levels installed. However, in order to activate the upgrade non-disruptively, CUoD must be available - CUoD is also available on all G5

and G6 9672s that have the latest microcode levels installed. For simplicity, we will assume that your CPC is using the latest Driver levels available at the time this document in published.

CUoD provides the ability to predefine additional CPs in the Image Profile. These additional CPs are known as "reserved" CPs. If you have defined reserved CPs for an LPAR, those additional CPs will become available to be varied online to the LPAR as soon as the new CPs are brought online by CUoD. Reserved CPs can be defined for LPARs that use dedicated CPs as well as those that use shared CPs. If you do not define any reserved CPs, the additional CPs are added to the pool of CPs used to dispatch shared logical CPs, but each LPAR will still see the same number of logical CPs as they had before the additional CPs are activated.

CUoD only provides the capability to *add CPs* non-disruptively. CPC storage can not be added non-disruptively, and CPs can not be removed non-disruptively. There are also a limited number of CPC upgrades that can not be implemented non-disruptively with CUoD:

- The first ICF can not be added using CUoD.
- The technology used by the CPs can not be changed - for example, you can't use CUoD to upgrade from an X77 to a Z87.
- The number of SAPs can not be increased using CUoD.

When you order CBU, you have to specify the target configuration. For example, you might specify that you need the ability to upgrade your Z77 to a ZZ7. Your machine will be configured with this information. When you then activate CBU, it will know that it must add 5 CPs.

Some additional characteristics of CBU are:

- CBU comes with the ability to have one CBU test per year for a contract period of 5 years.
- Removing the extra CPs added by CBU is disruptive. You must do a POR of the CPC in order to revert back to the previous configuration.
- You can not upgrade the CPC with an MES while CBU is active (that is, the extra CPs have been added and are in use).
- Five days before the end of the 90-day maximum CBU period, you will be provided with a series of messages indicating that the period is coming to an end. When the period actually runs out, the number of CPs is not reduced, however, the speed of each CP is reduced, severely impacting the performance of any work still running on the CPC. Performing the "UNDO Tempo rary Upgrade (CBU off)" via the SE and doing a POR will remove the added CPs and bring each CP back to its original speed.

When additional CPs are added using CUoD, the CPU version number, as updated by the STORE CPU ID (STIDP) instruction, does not reflect the additional CPs until the next POR is done. However, the new STORE SYSTEM INFORMATION (STSI) instruction *is* able to see the additional CPs. If you issue a D M=CPU command at this point, you will be provided with two configurations - the original one as seen by STIDP, and the new upgraded configuration as seen by STSI. Any products that require information about the actual configuration they are running on (maybe for licensing purposes) should be upgraded to add the STSI support. This consideration applies to any CPs added by CUoD - whether they were provided by an MES or by CBU. A POR of the CPC is required in order to allow STIDP see the upgraded CPC configuration. You should apply the PTFs for APARs OW37091, OW38489, and OW37254, which add CUoD support to OS/390.

To get the maximum benefit from CBU and CUoD, it is vital to do careful planning in advance, to ensure that you end up with the configuration you want when you add the extra CPs. This planning should cover how many LPARs you will need. For example, on a CPC used for disaster recovery, you may wish to start additional LPARs to run the new workload.

Further information about CUoD is available on the Web at: http:/www.s390.ibm.com/pes/apg/. There is also a *CBU Users Guide*, SC28-6804, which is being updated to reflect the latest enhancements at the time of writing For additional information, contact your IBM representative.

These LPARs, along with the number of CPs you want in each LPAR, both now and in the target configuration, must be defined in advance. Don't forget that to get the maximum benefit from the additional CPs, the CPC should have enough channel capacity and memory to provide a balanced configuration. Not only should you have the channel *capacity* that you need, but you also have to make sure that they are *connected* to all the devices you will need access to in a disaster. You should also document the procedure required to end up with the mix of dedicated and shared CPs that you need.

To summarize, CBU and CUoD provide outstanding flexibility to add capacity on short notice with *no* outage to the running applications. While CBU and CUoD have no sysplex requirement, together with Parallel Sysplex, these features are important ingredients in your availability toolkit. For customers with the highest availability requirements, CBU support is integrated into GDPS, providing the ability for completely automated CPC upgrades in case of a disaster.

OS/390 I/O resource management

Harry Yudenfriend

Harry is a Senior Technical Staff Member in the OS/390 Software Design Organization. He is responsible for the design and development of S/390 I/O functions.

Over the years, S/390 has built into the system architecture a number of features that have allowed customers to efficiently manage their systems and exploit the availability, scaling and workload management capabilities of OS/390 to efficiently run multiple workloads at the same time. For I/O, these S/390 features include the Channel Path Measurement Facility (CPMF), and Extended Channel Path Measurement Facility, to gather performance data on channel path resources, the channel Monitoring Mode that allows the creation of Channel Measurement Blocks (CMB) that gather I/O resource usage and contention statistics to individual device granularity, plus a number of other S/390 machine facilities that allow resource monitoring products such as RMF to provide detailed reporting on other I/O statistics and resource contention for capacity planning and problem analysis. These I/O facilities also provide the customer with the ability to accurately do accounting and billing of applications for the consumption of I/O resources. S/390 control units and devices also provide mechanisms to gather and report I/O statistics outboard from the central processor complex (CEC). For example, S/390 DASD controllers provide a number of statistics that allow reporting on things like cache hit ratios and subsystem utilization.

Less understood by the server market is that OS/390's Workload Manager (WLM) component has been constantly evolving to exploit the S/390 I/O instrumentation data. It is OS/390's objective to create systems that are self tuning, require fewer specialized skills in order to plan and configure, and to maximize the efficient utilization of all the I/O resources to provide the maximal business value to the customer.

I/O Priority

In OS/390 Version 1 Release 3, WLM introduced the capability to prioritize I/O requests. The objective of this support was to introduce sysplex-wide, goal-oriented management of I/O priorities, driven by WLM knowledge of goals for work and the business importance of those goals. The I/O priority for each request must be divorced from the dispatching priority associated with the requesting dispatchable unit in order to allow independent algorithmic adjustments. This implies that:

- I/O priorities be dynamically determined, based upon the business value of that I/O to the customer
- I/O priorities be maintained at a WLM service class period level, and be synchronized across the systems in a Parallel Sysplex
- The I/O priority associated with each I/O request be passed to the disk storage systems, for example, the Enterprise Storage Server or "Shark" for efficient management of the resources it controls outboard
- Monitoring and reporting extensions be provided to externalize relevant performance data.

I/O Priority improves the ability of the system to manage contention at the device when it occurs.

Paul Rowntree

This results in more effective utilization of shared devices, improving the WLM ability to satisfy service level objectives for work. The addition of I/O delay monitoring and reporting assists customer I/T personnel in the identification of workloads impacted by I/O delays and/or contention, providing additional systems management benefits.

With S/390 exploitation of Fibre Channel technology (FICON)™ the I/O priority assignment made by WLM will be even more effective in achieving the customer goals because of the frame multiplexing characteristics of the Fibre Channel (FC) architecture. Long running I/O requests that transfer larger amounts of data will no longer be able to lock out shorter transactional style I/O requests with higher priority. Additionally, the FC architecture also continues to evolve through standards committees. Future enhancements to the FC architecture include the ability for I/O priorities to be carried through the FC fabric where standard FC switches can exploit it.

S/390 I/O instrumention data allows the OS/390 Workload Manager to gather I/O performance statistics and contention data on all systems in the sysplex and determine if I/O bottlenecks are contributing to any missed goals. S/390 will continue to evolve and to exploit its ability to assign I/O priorities on behalf of workloads to other aspects of the system implementation.

Shark and I/O parallelism

Harry Yudenfriend

Enhancing I/O Parallelism with the Enterprise Storage Server

The Enterprise Storage Server (ESS or "Shark") implements two new features that significantly enhance disk storage system performance, Multiple Allegiances (MA) and Parallel Access Volumes (PAV). The Multiple Allegiance feature enables the disk storage system to concurrently execute I/O operations for the same logical volume across multiple sharing systems. Concurrently executing channel programs are allowed to occur for a logical volume as long as the extents on the volume that are being written to do not overlap the same extents on that volume that are being read. The disk storage system automatically determines these extent collisions and serializes the requests when required. To increase the probability that no extent conflicts do not occur, software changes were necessary to make sure to specify as granular an extent range as possible in the channel programs used by the various access methods and I/O drivers.

The Parallel Access Volume feature (PAV) extends on the MA feature by allowing concurrent execution of channel programs to the same logical volume from the same operating system! Prior to PAVs a disk volume was seen to the operating system as a single serially reusable resource, represented by a single Unit Control Block (UCB) to the software and a single subchannel to the S/390 channel subsystem. Only one I/O operation could be permitted to execute at a time. As volume sizes grow larger and larger, more and more I/O operations may need to execute against the volume at the same time. Thus, I/O queuing delays become a concern. PAVs allow simultaneous access to a logical volume by multiple users or jobs from the same OS/390 system. As with MA, both read and write operations may execute concurrently, as long as they are not accessing the same area on the disk (extents).

The combination of PAV and MA complement each other by increasing the amount of parallelism at the logical volume.

WLM PAV-Alias Tuning for the Enterprise Storage Server

The Enterprise Storage Server (ESS or "Shark") exploitation of WLM derived I/O priorities was just one of the first synergies between OS/390 and the disk storage system for efficient resource management. WLM's exploitation of Parallel Access Volumes goes even further to create a self tuned system requiring fewer specialized skills to manage effectively.

Exploiting PAV in OS/390 was accomplished by creating multiple UCBs and subchannels per logical volume. The PAV-base UCB represents the logical volume to the application. It is defined with the device number that is used to represent the logical volume to the applications, operating system and operations staff. Applications wishing to access the volume may allocate by the PAV-base device number. The operating system will use the PAV-base device number for error messages and EREP records. The operations staff uses the base device numbers for all reconfiguration and display commands, such as the VARY device and path commands, and the display units command. In this way, the implementation of PAV is essentially invisible to applications and operations staff.

PAV-alias UCBs are defined in order to provide the additional concurrent access to the logical volume. Their existence is transparent to applications and the operations staff. These PAV-alias UCBs and subchannels can have their association with a specific logical volume non-disruptively changed to any logical volume within a logical subsystem (LSS) on Shark. This change can be manually done via controls that manage the Shark (ESS Specialist) or automatically done by OS/390 Version 2, Release 7.

In OS/390 Version 2 Release 7, while operating in WLM goal mode, the system programmer can choose to let WLM and IOS dynamically manage the PAV-aliases. Dynamic management of PAVs means that WLM will determine which workloads in the sysplex are not meeting their goals due to IOS queue time delays. With this information WLM will optimize which PAV-base UCBs should have PAV-alias devices added to it in order to increase the I/O parallelism and eliminate IOS queuing delays.

Two algorithms are used to make adjustments, the Goal Mode Algorithm and the Efficiency Algorithm. The Goal Mode algorithm's intent is to help a service class to meet its defined goals. WLM first identifies service classes that are missing their goals due to IOS queue time. If the LSS does not appear to be overly constrained, for each such service class, WLM considers adding a PAV-alias to the PAV-base causing the most IOS queue delay to work in the service class. The first choice is to use a PAV-alias that is not logically assigned to a PAV-base in OS/390. These devices are ones that are associated with a PAV-base that was taken off-line for some reason. If there are no unassigned PAV-alias devices then one is taken from a device that is executing lower priority work. A PAV-alias will never be taken from equal or higher importance work.

The second algorithm is referred to as the Efficiency Algorithm. Its purpose is to try and reduce overall IOS queuing by moving aliases to the most utilized volumes within an LSS. WLM will periodically look for devices with significant IOS queuing on PAV devices. If there is a PAV-alias available from another device in the LSS and taking the PAV-alias will not cause an increase in queuing on the donor device, then it is moved to help reduce the queuing on the target device. The Efficiency Algorithm is designed to be very conservative in its choices in order to minimize any chance of thrashing between devices. The Goal Mode algorithm takes precedence over the Efficiency algorithm.

Both the Goal Mode and the Efficiency algorithms make their decisions on the measurement data collected by all the systems sharing the devices in the same sysplex. The changes that WLM makes apply to all systems sharing the device.

Having WLM manage the PAV-alias assignments has quite a number of benefits. The systems programmer does not have to perform detailed analysis on where to place data sets and where to assign aliases. Workloads are constantly changing and the systems programmer can at best only statically configure for an average access pattern for the LSS. But the average access pattern and I/O rate is almost always sub optimal. With WLM Dynamic Alias Tuning data does not need to be moved around to avoid hot spots. This improves application availability and reduces management costs. All that the system programmer need do is to assign enough PAV-aliases in the LSS to meet the overall aggregate workload requirements. WLM can adjust the PAV-alias requirements as the workload changes.

WLM Dynamic Alias Tuning helps to reduce the hardware costs. This occurs because by dynamically moving PAV-aliases around the average response time to the LSS is decreased. This has the equivalent effect as to increasing the cache size of the control unit, without the added hardware cost.

Allowing OS/390 to manage the PAVs facilitates larger volume sizes. Multiple I/O requests arriving for the same logical volume will not necessarily block each other from executing. Having larger volume sizes has the added benefit of reducing virtual storage constraints below the 16MB storage line. Instead of running with three 3390-3 DASD volumes, each with a UCB below the line, the installation can choose to define a single 3390-9 volume with the PAV-base still below the 16MB line and the PAV-alias devices in 31-bit storage. The PAV-aliases are invisible to the applications, but continue to provide the I/O parallelism achieved by 3 separate devices.

Paul Rowntree

Finally, with WLM Dynamic Alias Tuning the installation does not need to dedicate as many unit addresses (subchannels) in the LSS for use as PAV-aliases. Thus, more data can be addressed in a single LSS by using more of the available unit addresses to define PAV-base devices. The fewer number of PAV-aliases can be adjusted as needed by the work load.

S/390 I/O instrumentation data allows the OS/390 Workload Manager to gather I/O performance statistics and contention data on all systems in a Parallel Sysplex and determine if I/O bottlenecks are a contributing factor if a workload fails to meet its stated goals. WLM uses this data with the Enterprise Storage Server or Shark to adjust I/O priorities and PAV alias assignments across the sysplex to help the workload achieve its specified goals while reducing the specialized skills required of the customer.

What are clusters? Why are all server platforms moving to support them?

Madeline Nick

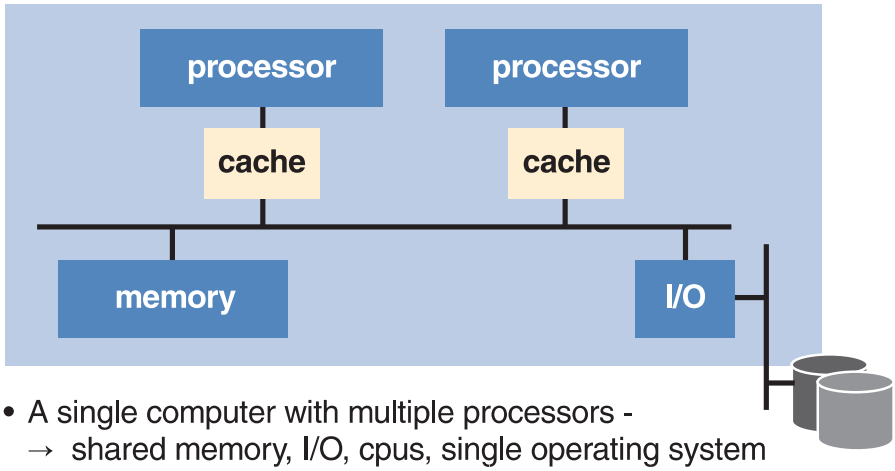
For information about Madeline, see her previous article.

There is one constant in business computing: workload requirements will always, at some point, exceed the capacity of even the largest single server. Additionally, in today's ".com" business environment, availability, has become exceedingly important. This article describes how clusters can address these two major IT requirements and which type of cluster addresses one requirement better than the other.

Before we dive deeply into the definition of a cluster and the different types of clusters, lets start with a definition of an SMP (Symmetric Multi-processor), which is the building block of clusters. An SMP is a set of processors that access a single common memory, and typically operate under the control of one operating system. In an SMP, work units can be dispatched on any processor with common access to shared resources such as virtual memory and I/O channels. To the applications running on an SMP, the underlying physical system presents a Single Logical Resource Image, as defined in the previous sentences.

In an SMP, some processing power must be used to provide serialization and cache coherency across the processing engines, which reduces the additional computing power achieved by adding each successive processor (1-3%). In fact, with very large single image SMPs, in an Online Transaction Processing environment, adding that next processor can even reduce the total system throughput!

Symmetric Multi-Processor (SMP)



- A single computer with multiple processors -
  → shared memory, I/O, cpus, single operating system
- ...and those processors are equal ("symmetric")
- Single Logical Resource Image
- De facto standard in parallel computing
- Limited scalability
- Not highly available - single points of failure

Traditionally, most companies have dealt with the challenge of scalability by leveraging large symmetric multiprocess or (SMP) servers. But, today, as part of the migration to e-business enablement, corporations are developing competitive strategies which have an increasing dependence on Information Technology (IT) for both internal and external core business processes. This movement has created heightened demand for ever increasing scale and availability from the IT organization while holding systems management (people) costs flat. As capacity and availability requirements continue to grow beyond the capabilities of single large SMP systems, more and more companies are moving towards clusters of servers to provide the attributes required for their demanding e-business growth.

In fact, for many businesses of today, a single system, no matter how large and powerful, will never meet all the requirements for enterprise computing: scalability, availability, security, handling mixed workloads and new applications, all at and a low total cost of computing.

e-Business: What does it mean to IT?

- e-business = growth without limits
- e-business is more than a state of mind. It is about leveraging IT for business competitive advantage. It is about connecting commercial assets to the Web: applications and data.
- e-business personifies unpredictability
- *e-business* DEMANDS:
  - → *Scalable capacity* for workloads beyond the limits of a single SMP
  - → *Non-disruptive Growth* in the face of explosive capacity demand
  - → *Consistent response times* for web-clients even during peak demand
  - → *Continuous Availability* for planned and unplanned outages
  - → *Systems Management simplification* (proliferation of mid-tier nodes is becoming a management nightmare)
  - → *Disaster Recovery* measured in minutes of down-time

These are the metrics by which e-Business companies measure their IT technology suppliers.

These are the qualities of service delivered by S/390 Parallel Sysplex.

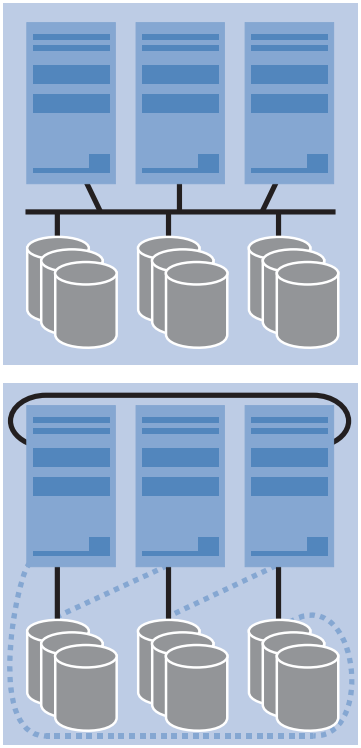
Cluster

A cluster is a group of server nodes (SMPs or uni-processors) that work together to provide greater total capacity and availability than can be delivered by a single SMP, with better systems management and single points of control (SPOC) than is typically delivered by network-attached distributed processors.

A number of vendors offer clustered systems that provide scalability as well as availability, using parallel middle-ware and high-speed proprietary interconnects between system nodes. All clusters are proprietary; there are no open standards. In fact, the amount of effort a vendor puts into his interconnect, particularly for scaling purposes, varies widely, but has a significant affect on the quality of the resulting product. IBM has spent a significant amount of effort on its coupling facility. We have 50 patents related to our coupling technology.

Cluster

- A collection of whole computers...
  - → processor(s), memory, I/O, OS
- Natural high availability: a spare for everything
- Scalability typically provided through data partitioning and message passing
- Separate “islands of compute resource”
  - → Data affinity binds work to a given server node
  - → Can’t dynamically move work based on cpu/memory/IO capacity need



While there are many types of clusters, and some clusters may span multiple models, for the purpose of this article, let us divide clusters into three broad groupings:

High-Availability Clusters

High-availability clusters are used to achieve higher availability than single systems by means of a core set of fail over services. *Heartbeat* monitoring via signaling between systems is typically used to determine the health of systems operating in the cluster. Some form of system isolation functions (initiated manually or automatically) is required to prevent the system determined to be failing from accessing any resources once removed from the cluster.

Manual or automated switch over of workloads or the operating system to another designated system in event of failure provides improved application availability. Quick "fail over" is very important for some users. In such cases a backup server in the cluster is normally idle, ready to take over restart of the work as soon as the primary server fails. Examples of clustering technology for fail over include RS/6000 HACMP and Microsoft ® Wolfpack.

Paul Rowntree

Parallel Clusters

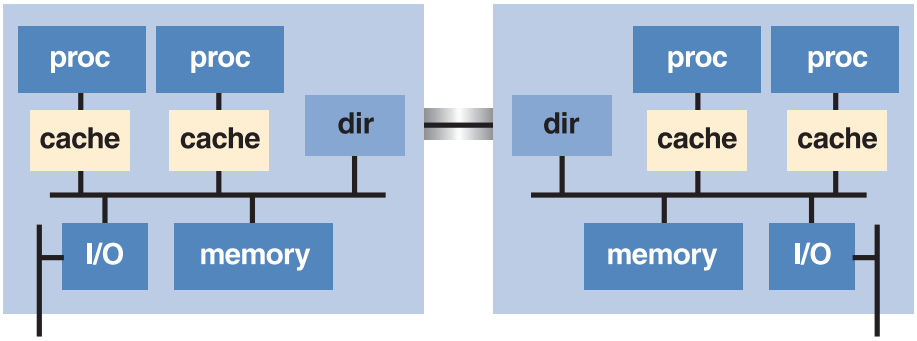
In a parallel cluster, multiple servers are used to provide scalability for an application. Typically, a means is provided to split the workload among the different servers so that it can be processed in parallel. For efficiency, each server processes its own work and then the results are combined. In many of these implementations the database is partitioned between the servers so that each server provides read/write access to a portion of the total database. This model is also referred to as a "shared nothing" model. Data requests originating on one server directed at data "owned" by another server in the cluster is sometimes provided using a high speed interconnect. Dependent on the number of remote accesses to data, this could get very expensive. But, keep in mind that, because of the partitioning technology, while there is an overhead for accessing data on a remote note, the scalability characteristics of this Highly Parallel Cluster are excellent. Local access to data is very efficient, remote access is costly, but does not increase with the number of servers, hence it is highly scalable through the addition of nodes. Engineering/Scientific and Business Query kinds of applications fit this model well.

Single System Image Clusters

In a single system image cluster, all nodes appear to the client application as a single logical resource image system, wherein any dispatchable work unit can be scheduled for execution on any node in the cluster with concurrent access to shared databases, without sacrificing caching of data in local processor memory. In this approach, each node has direct read/write access to all data, and any transaction can run on any node. This model is known as a "shared data" model.

Workload balancing is critical in this model to ensure an even distribution of work. There is a single point of system administration. Users can continue to access an application even if one server fails, and the combined power of all servers can be used on a single application to provide aggregate throughput substantially greater than any single SMP can provide, without requiring database partitioning or application splitting across nodes. Scalability is the challenge for this form of

NUMA (CC-NUMA)



- Cache-Coherent Non-Uniform Memory Access
  → distributed shared memory
  → One memory space, one OS; not a cluster
- Low-cost hardware scalability building blocks
- Single Logical Resource Image
- Availability like SMP.
  → Really! For the same reasons.
- Workload scalability depends on significant Operating System and subsystem re-structuring

clustering, requiring highly-optimized system interconnects and/or multi-system cache coherency and serialization technology to deliver near linear scalability. (The S/390 Parallel Sysplex is an example of this form of clustering).

Other servers aiming for increased scalability, have adopted variations on the traditional SMP system structure that provide for a single operating system structure that can support a large number of processor engines, such as, Non-Uniform Memory Architecture (NUMA)® systems. Here, processors, I/O channels and memory are distributed across a collection of physical "sub-nodes" which are combined through a cache coherency fabric to provide a single, large, shared virtual memory system. As with the S/390 Parallel Sysplex, the many physical processors provide a large, single, logical resource image.

Thus, NUMA provides a manageable, large scale compute capability, but, with a single operating system and shared virtual memory, NUMA systems do not provide continuous availability.

In conclusion, cluster solutions, in general, address two key challenges facing enterprise information systems: scalability and availability. S/390 Parallel sysplex goes beyond ... The unique, S/390 Parallel Sysplex data-sharing approach combines an SMP-like Single Logical Resource Image with the availability strengths of clustered systems and NUMA-like scalability. IBM S/390 offers a robust cluster solution by integrating the strength of S/390 enterprise-class servers as the building blocks for clusters with workload balancing software to provide true dynamic resource/application management .
Please refer to the "Parallel Sysplex Overview" article in the Supplement and the "Life of a Transaction" article for additional information .

A bit about batch

David Raften

David has been in the Parallel Sysplex Competency Center since 1994. He comes there with a background in performance testing, DB2, IMS, and batch workloads.

A lot of the attention devoted to Parallel Sysplex has focused on online systems, and the technicalities of data sharing. However, a sizable portion of workloads is still made up of batch work. This article describes how you can run and manage your batch workload in a Parallel Sysplex.

As was (almost) quoted in "The Treasure of the Sierra Madre:"

"Batch?" ... Well, see the cartoon ...

Contrary to this (pseudo) quote, batch was, and continues to be, an integral part of the modern workload. It is the most efficient means to read files and produce reports. It is also under more pressure than ever to perform quickly, efficiently, to maintain high availability as the amount of data that needs to be processed increases and the amount of time to do it in decreases.

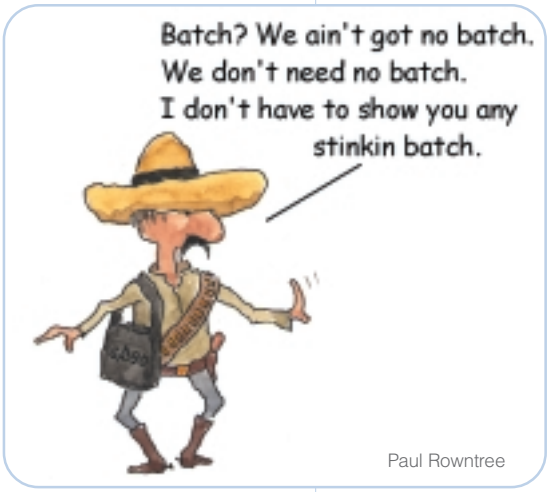
There are two questions that users will ask about their critical batch job:
1. Did the job run successfully?
2. Did the job complete within the deadline?

The first question is related to the availability of a system and subsystem with access to the user's data. The second is dependent on the elapsed time of the job, assuming there was a subsystem available to service it. Features of OS/390, including the Workload Manager and the data base managers, work together to help insure that the answer to both of these questions are "Yes."

The goals of getting the job to run quickly is done with the dynamic workload balancing capabilities for batch workloads. Depending upon the job and the data that it accesses (DB2, DLI, VSAM, etc.), this is done in different ways.

JES2 and JES3 both support a multi-JES environment. JES2 does this with a Multi-Access Spool (MAS), and JES3 with a Global/Local environment. This allows batch jobs to run on any system in the "JESPlex". If the JESPlex spans the Parallel Sysplex, the batch workload can use the entire resources of the Parallel Sysplex to meet its needs.

When jobs get submitted by either a TSO user or by a job scheduler, they are placed on a JES input queue in SPOOL. For JES2, any JES2 system that has access to the shared SPOOL can pick up and run the job. For JES3, the Global directs the job to run on one of the Local systems. For the job to run on a system,
1. The system needs an available JES initiator.
2. The initiator is able to run the job class specified in the JCL. Job classes are typically defined according to resource requirements of tape, CPU time, location of specific software, etc.
3. No system affinity was specified in the JCL. If one was, then only the system that matches the SysID specified can pick up the job. SysIDs are typically there due to specific hardware or software requirements that are on one OS/390 image. Data base manager support for Parallel Sysplex can help remove the need for the system affinity to be specified.



Using this model, there was a simple form of workload balancing. All else being equal, jobs running on a processor with more capacity will finish sooner than on a processor with less capacity. This frees up the initiator so the initiator can pick up the next job to run. Further, since jobs are mostly I/O-bound, more initiators should be defined on the OS/390 image with the most capacity.

A problem with this environment is that "all else" is not always equal. Different OS/390 images have different loads at different times. Bursts of activity from another workload, or different mixes of batch jobs affect the overall characteristics of the OS/390 image in terms of CPU, storage, and I/O usage. In addition, the "normal" workload is constantly changing as applications are modified and added, LPARs move, and hardware resources are reallocated. This requires continuous tuning of the number of initiators, the job classes that they pick up, and the individual job priorities.

WLM Managed Initiators

Workload Manager, starting with OS/390 R4 for JES2, OS/390 R6 for JES3, automates the initiator and scheduling management with WLM managed initiators and Resource Scheduling Environment.

Job initiators for WLM managed job classes are controlled dynamically by the workload manager. WLM adjusts the number of initiators on each system based on:

- The queue of jobs awaiting execution in WLM managed classes,
- The performance goals and relative importance of this work,
- The current Performance Index (PI) of the service classes associ ated with those jobs,
- The capacity of each system to do more work.

Initiator management is best for jobclasses that have a steady supply of arriving jobs. This allows WLM to accumulate history on the queue of jobs and determine at what point the number of initiators should increase or decrease. Other jobclasses that have very low arrival rate or just a few long-running jobs at a time, may not lend themselves to dynamic initiator management. JES-managed and WLM-managed initiators coexist nicely.

Put together, this works to guarantee that the high priority jobs get an initiator to run on with the workload getting balanced automatically among all the systems. If it is desired that a hot batch job needs to run quickly on a system with a particular resource, it is no problem.

Effect of WLM-managed Initiators

Performance tests were run with WLM managed initiators. These tests provided the most balanced processor utilization and consis-tently provided the shortest total elapsed time, from the submission of the first job of the suite to the end of the last job. WLM reacted very quickly to changes in the number of jobs waiting to start and also to system utilization. As would be expected, the elapsed time for some jobs increased, as there was now more competition for CPU resource on some of the systems; however, the overall elapsed time was always better, even when time consuming efforts had been made to manually balance the workload.

Scheduling Environments

Function has been added to WLM which makes it possible to define and control arbitrary resource names, mapped against real system resources, for example, systems in a particular DB2 data sharing group or systems dedicated to online production workloads or those with cryptographic capabili-ties. JES provides scheduling mechanisms that only initiate jobs on systems where the required resources are available.

Resource scheduling is intended to provide a method for an end user or system programmer to declare which resources are required for a job and have the job scheduled only on systems having the required resources. It is a way in which you can cause batch work to be initiated:

- Where a job's resources exist
- Where jobs are desired to execute
- When jobs are desired to execute
- For a combination of the above

This eliminates the need for an end user to specify system affinities, and may eliminate the need to specify a particular or a system specific job class. The scheduling environment provides scheduling and controls that ensure that jobs are scheduled only on the systems having access to the required resources, such as:

- Devices
- Hardware features such as encryption
- Specific products
- Specific subsystems

Once the Scheduling Environment is specified, the specific resources and availability of resources can then be dynamically changed without further JCL adjustments.

For example, you can define a scheduling environment such as:

Scheduling Environment	Resource Names	Required State	Description
DB2ANIGHT	DB2A	ON	DB2A data sharing group
	PRIME	OFF	Not Prime Shift

Any job specifying "SCHENV=DB2ANIGHT" will only run on the system or systems where all required resources are ON.

Batch and Data Base Managers

For optimal performance, the batch workload should be balanced among all available systems in the Parallel Sysplex to speed up the nightly job stream. A factor when balancing the batch jobs is require-ments that the job has to run on the same OS/390 image as the data base manger (DB2, IMS) for the data bases it references. The data base manager subsystem ID also needs to be specified within the batch job's JCL. Unless action was taken by the data base manag-ers, it would be necessary to manually restrict where a job can run. The IBM data base managers support for the Parallel Sysplex technology provides the flexibility to allow jobs to run on any member in the data sharing group and makes dynamic workload balancing for batch a non-issue.

DB2's Group Attach Name and IMS's IMSGROUP name provide this needed flexibility. They can be substituted for the subsystem name in the batch job (and utility) JCL. By knowing this data sharing group name, it is not necessary to write user routines to determine which specific subsystem name is running on the LPAR and dynamically modify JCL. Whether you have one or thirty two subsystems, the job is freed to run on the system where it

run to best meet your requirements.

Summary

Whether you are running “generic” batch, or batch that requires specific data base managers to run, the capability is there to easily schedule the job on the system where it can run and with the most available capacity. High availability and dynamic workload balancing is not just for transactions, but also for the most crucial and underrated workload you have: Batch.
With OS/390 simplifying the effort of batch management, it is no wonder that so many forget about it!

More information on running Batch in a Parallel Sysplex can be found in SG24-5329 *Batch Processing in a Parallel Sysplex*

CICS and CPSM

Mike Bredenkamp

Mike has worked for IBM since 1981 and has spent most of that time with Transaction Processing systems on S/390 and its predecessors. During this time he has worked both in the field and in the CICS Development Laboratory. Currently, Mike is at the Dallas Systems Center's Advanced Technical Support Group.

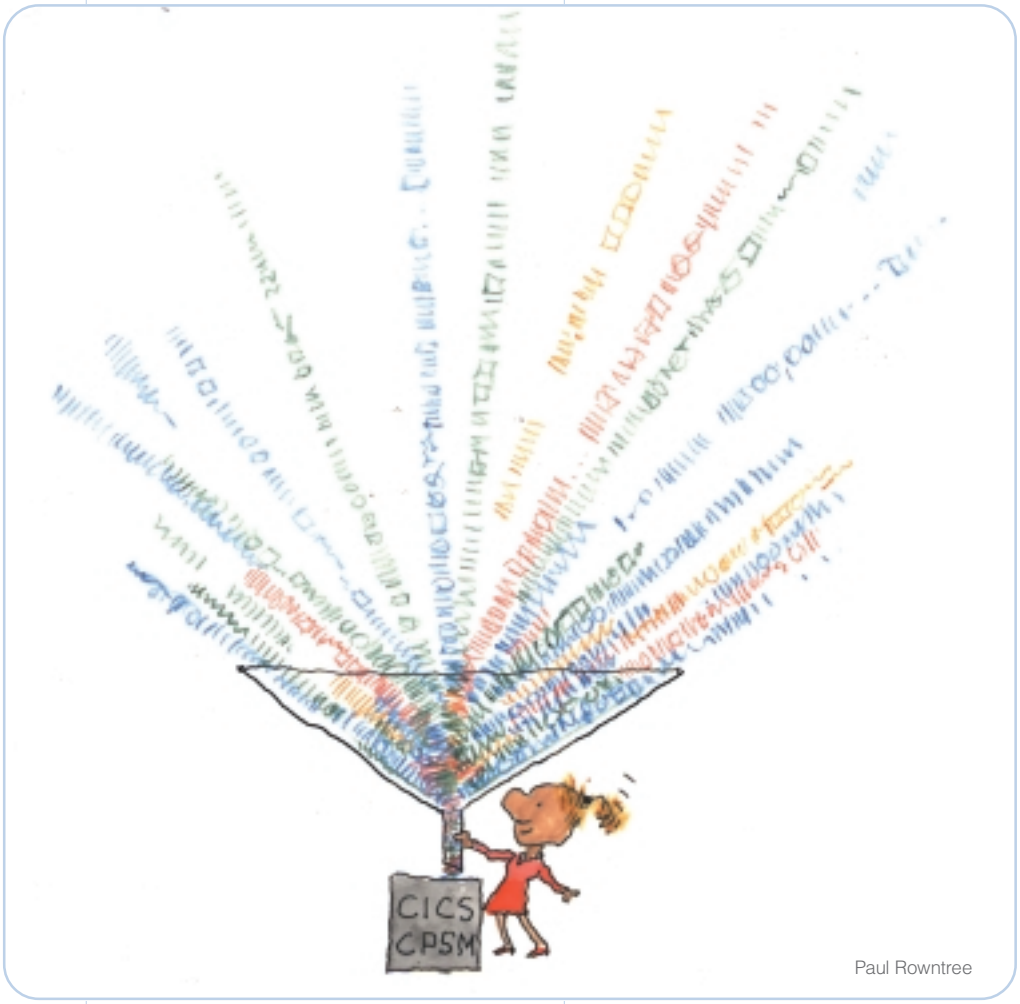
CICS is one of the most popular transaction processors in the world and handles vast volumes of transactions on S/390. The CICS internal architecture is well suited to exploit multiple processors regardless of their location within a Parallel Sysplex. With the addition of CICSPlex® Systems Manager (CPSM), the operational and single-image aspects of workload management becomes ideal to exploit S/390 Parallel Sysplex for large transaction volumes. Both these products are packaged as CICS Transaction Server.

CICS is also integrated with many of the advanced facilities of Parallel Sysplex, such as WLM, ARM, XCF, VTAM Generic Resource or TCP/IP VIPA to ensure optimal availability and resource usage when running in a Parallel Sysplex.

CICS Architecture

A CICS workload consists of many *transactions* that originate from outside the system, typically from an end-user or another computer system connected via a network. Popular networks are either SNA or TCP/IP.

Each transaction requires the use of CICS facilities as well as data from a database or file system, such as DB2, VSAM, or HFS. The transaction may require cooperative processing from more than a single CICS *region* and may generate additional workload in the form of non-terminal or *background* transactions that will process asynchronously from the originating transaction. In order to handle large workloads, multiple CICS regions may be configured similarly — this is often referred to as *cloned regions*. Other CICS regions may be configured for a specific function, such a terminal or network connection, file or database management, or for specialized processing.



Paul Rowntree

Typically, these regions are referred to as Terminal Owning Regions (TORs), Application Owning Regions (AORs) or File Owning Regions (FORs). Also in common terminology are Resource-, Printer- or Data Owning Regions (ROR, POR or DOR, respectively). A number of CICS regions configured to process one or more workloads is commonly referred to as a CICSplex. Do not confuse the term CICSplex with the product CICSPlex Systems Manager or CPSM.

Session Balancing

Before a transaction can be initiated, a session needs to be established between the end-user and the CICSplex. A typical CICS system may have 100s to 100,000s of concurrent sessions, only some of which will be active at any given time.

This point of entry into the Sysplex is the first point where a number of very important actions take place.

Here it is possible to mask the underlying topology of the CICSplex and individual availability of components in the CICSplex from the end-user. When the end-user attempts to establish a session to CICS, it is possible for the end-user to establish the connection to a generic VTAM® name or a virtual IP address that in both cases will translate into a specific VTAM application or IP address. Whilst normal session establishment will ensure connection to an active instance of the application, involvement of WLM can assist in making load-related choices between various instances of the application. Thus, for processors of equal power, the networking software in conjunction with WLM can balance the number of sessions over the number of available application instances, without the end-user being aware of the specific instance of the application that will process his requests. Should the processors have disparate processing capabilities or differing workloads, WLM will attempt to route more sessions to the processor that would seem to have capacity to process the additional workload.

The session is typically established with the front-end region or TOR in a CICSplex. In the case of a Sysplex with multiple MVS images, there may be one or more TORs on each MVS image. The ability to connect to any one of a multiple of regions without knowing the exact location of that region in the Sysplex is referred to as *single-image* and facilitates the end-user's ability to successfully connect to the CICSplex without being aware of the specific name or availability of an instance of the CICS region. At the same time, the ability to balance sessions between various processor images or CICS regions, allows for a basic form of workload distribution. This is achieved through the use of VTAM Generic Resources and by IP Connection Optimization for TCP/IP sessions.

The lifetime of a session is typically long and during this time many transactions with various characteristics may be submitted to the CICS regions. To further facilitate workload separation and workload management, the first CICS region that creates the transaction (typically the TOR), will interact with WLM to classify the transaction for the purposes of reporting on its flow through the various CICS and subsystem regions. This information is made available to WLM for sampling and for various resource adjustments by WLM to meet the transaction's stated objectives within the overall goals for the system. The transaction might not be executed in the TOR, but merely be routed to one of a group of similar AORs connected to the TOR that have the capability to process the transaction. In making a route selection, CICS has an ideal opportunity to interact with CPSM and include in the selection criteria the particular goal or Service Level defined for that transaction.

Target Selection

CPSM fulfills a critical role in a CICSplex. Not only does it provide for a single-system operational interface to the multitude of CICS regions in the CICSplex and for monitoring the status of the systems, but, more pertinent to this discussion, it also acts as an intelligent routing mechanism to route work to the most appropriate AOR in the CICSplex.

CPSM has 2 algorithms for target selection; queue-based or QUEUE (Join Shortest Queue) and Goal-based. We will briefly look at the Queue methods before discussing goal-mode with reference to WLM.

Using the Queue algorithms, CPSM selects a target region to route a transaction based on the following criteria. It will select the region that:

- Is the healthiest
- Has the least queue depth (or load)
- Has the fastest CICS link from the routing region
- Has the least transaction abend probability.

For Goal-mode routing, the following considerations apply: CPSM will select the region that:

- Is the healthiest
- Has the least load
- Has the fastest CICS link from the routing region
- Has the least transaction abend probability, when calculated
- Is the most likely to allow the transaction to meet the response time goal set for it and other transactions in the MVS workload management class.

Note that the primary consideration in selecting a target is the ability to perform work, only in the final instance, if multiple candidate AORs are found capable to execute the transaction, does CPSM factor in a Service Level.

It is also worth noting that a routing decision is made to favor the fastest (cross memory vs. XCF) link. Given that session balancing has already been involved to balance the sessions over the available processors, this will ensure that all processors participate to some degree in the execution of the workload.

Whilst most installations currently use the Queue algorithm with good results, Goal-mode routing presents the best opportunity for managing Service Levels. Queue-mode also function in those instances where Goal-mode is not possible, for example, when the CICSplex spans several different Sysplexes, or some components of the CICSplex are

not on MVS.

Goal Definition

For those systems that are in WLM goal mode, CICS transactions can be associated via MVS WLM to a service class. For CPSM, the only goal that is recognized is the average response time goal; specifying any other goal will cause CPSM to assume that the goal is being met. CPSM will manage at the service-class level and will allocate a service class to a set of target regions, thus minimizing the amount of resource reallocation by MVS WLM. This also has the effect of grouping those transactions meeting their goals separately from those transactions that does not meet their goals, thus allowing WLM to reallocate resources appropriately.

The goal-mode algorithms work best if the number of AORs per MVS are symmetrical, with a number of service classes comparable to the number of target regions in a given MVS image.

In order for Goal mode algorithm to be used, all participating CICS regions must be in the same Sysplex and WLM must manage these regions to the average response-time goal as mentioned previously. Switching any image back to SRM mode cause CPSM to revert to QUEUE algorithm.

Considerations

The user should bear in mind that WLM will allocate resources to a CICS region. All work within that region will be dispatched by the CICS dispatcher according to the CICS priorities, without reference to the Service objectives of the individual transactions. An inappropriate specification of the CICS priorities, or a large number of transactions with different objectives in the same CICS region will thus reduce the effectiveness of WLM's efforts to provide adequate resources to meet the goals of critical transactions.

The inclusion of the "fastest CICS link" in the selection criteria for selecting a region has the result that CPSM will tend to route somewhat more work to AORs on the same MVS image as the TOR. This is to avoid the slightly higher cost of routing over XCF links as opposed to a direct cross-memory route. The effect of this can usually be offset because of session balancing over TORs on each MVS image.

When a CICSplex with workload routing is implemented, the systems programming staff commonly queries the apparent lack of balance between various AORs in that the work is not evenly distributed across the available AORs. It should be borne in mind that the objectives of CPSM and WLM is not to spread the work around, but to ensure execution where the workload has a high probability of successful execution within the Service Level. There is no "round-robin" algorithms implemented. From a WLM perspective, it is easier to dispatch a single CICS region that successfully execute the work, rather than a multitude of regions. This has no detrimental effect on the workload, and with an increase in the workload, the load across all available CICS regions might appear more balanced.

Conclusion

CICS Transaction Server is designed to integrate with a Sysplex and to exploit the features and facilities of the Sysplex to achieve optimum throughput, response time and availability to the end user.

IMS is plex-ible

Bill Stillwell

Bill's IMS experience goes back to 1974 when he was with the US Army. In 1982, he went to the Dallas Systems Center, where he has been since, providing IMS technical support to some of IBM's largest customers.

IMS, that "grande old dame" of transaction managers and database managers (would you believe IMS/360 DB came out in 1968), has kept pace with its host software and now exploits many of the best features of the Parallel Sysplex. IMS has been architected and re-architected over these 30+ years to meet the ever increasing needs of the "bet your business" user who relies on IMS to be there when it is needed and to do however much work is needed; to take advantage of new technologies when they are available and useful; and to do so with little or no impact (the goal is always NO impact) on existing applications.

Paul Rowntree

Today's IMS is a far cry from what we had in the 70s, 80s, and even the early 90s. In the early days most leading edge IMS users were entering perhaps dozens of transactions per second from 3270-type terminals to a single IMS image which they hoped was available and they hoped could meet the demands of peak loads. Major trauma occurred when either of those hopes was not realized.

Other transaction managers and database managers were or came on the scene, and the word went out that IMS was in its death throes. Processing strategies changed from batch or directly connected terminal networks to (ooooh) client server, and then the (aaaaaah) internet. Where would IMS fit in all this, and why wouldn't it just go away and leave us alone. Well, it didn't go away. It adapted to the changing demands of its faithful. It will act as a server if that's what you need, it can be accessed from the Internet if that's your e-bag, and those hopes for availability and capacity have become near certainties.

How did it do it? Well, one reason is the S390 Parallel Sysplex (remember, at one time we thought maybe that would just go away too, along with its antiquated (booooo) maaaain-fraaaame). By exploiting the S390 Parallel Sysplex, users of all sizes and shapes can know that when IMS is needed for legacy or new applications, it will be there just begging for more work.

This article discusses three benefits that IMS systems gain from the parallel sysplex:
1. Increased capacity (bring on the work)
2. Increased availability (I'm OK, you're OK)
3. Increased flexibility (what would you like to do today)

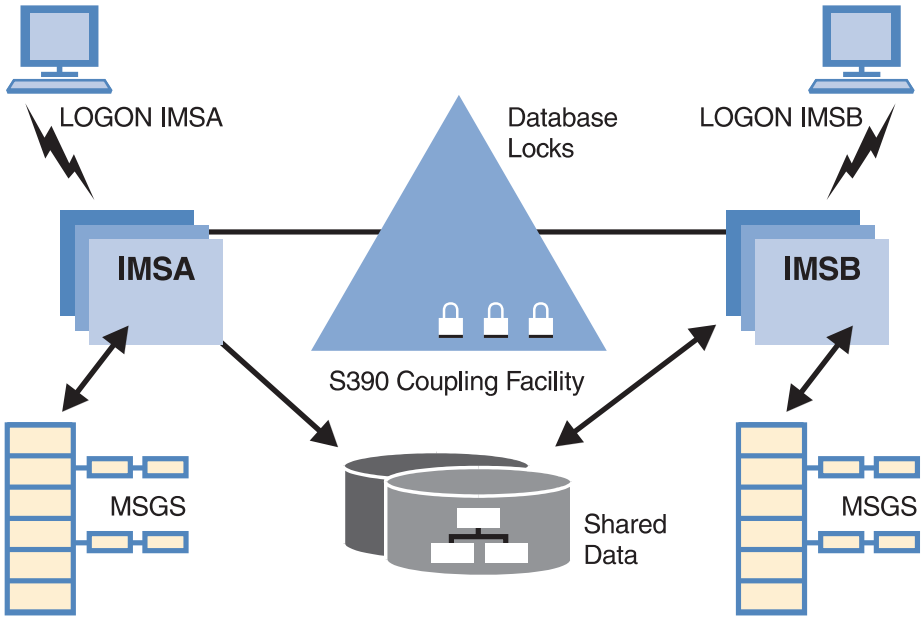
This is done by using the capabilities of the Parallel Sysplex to distribute work across multiple MVS or OS390 platforms, presenting a single image view to the end user. The first step in allowing work to process in any of several images was data sharing support. Although IMS first supported data sharing with IMS/VS 1.2, it was not until IMS/ESA V5 that IMS made its first use of the Parallel Sysplex by exploiting lock structures and cache structures in a coupling facility to maintain data integrity (see the'IMS Version 5' graphic on the next page). Not only did this increase the "n" in n-way data sharing to 32, but it improved the performance over the earlier IRLM "pass the buck" processing which used VTAM to communicate between lock managers. Work could now be distributed to up to 256 IMS systems on up to 32 S/390 platforms.

However, the distribution of this workload was still an "exercise left to the user." Many techniques were used, including the simplest one of asking (for example) half of your users to log on to IMSA and the other half to IMSB. More sophisticated means became available, including the use of a VTAM USERVAR exit to distribute the logons across multiple IMSs, and the use of the Workload Router - a program which uses the IMS Multiple Systems Coupling facility (MSC) to route transactions to different sharing IMSs based on a user provided algorithm.

Although these helped, they were basically "push down" systems - that is, the work was sent to one of the IMSs regardless of whether it was very busy already. In some cases, if one of several IMS systems were not available, then attempts to log on to that system failed and the user had to either know of an alternate active IMS to log on to, or wait for the original system to be restarted.

And then along came IMS V6. Data sharing was enhanced and additional capabilities were introduced:
1. Data caching
2. Support for VTAM Generic Resources
3. Support for Shared Queues

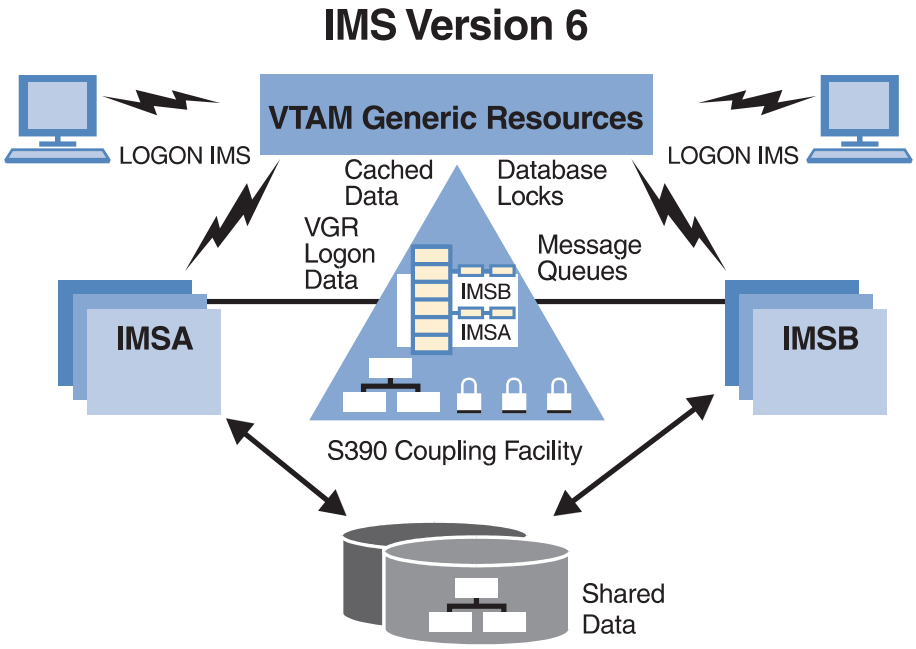
IMS Version 5



IMS users of the OSAM data access method can choose to have data cached in the OSAM store-through-cache structure in the coupling facility. When used judiciously, this can greatly improve the performance of that database (or databases) by reducing DASD I/O. Secondly, use of store-in cache structures allow users of the Fast Path DEDB with the Virtual Storage Option (VSO) to share data while getting additional performance benefits by placing entire areas of the database in dedicated store-in cache structures in the coupling facility. Both of these features substitute coupling facility I/O for DASD I/O.

VTAM Generic Resources (VGR) addresses the problem of a failed (or otherwise inactive) IMS which is (of course) not available when the user logs on. Each IMS, when it initializes and opens its VTAM ACB, joins a "generic resource group" managed by VTAM. VTAM, using list structures in the coupling facility, keeps track of which IMSs are active, and how many users are currently logged on to each. A logon request using a generic name is routed by VTAM to the active IMS with the fewest number of logged on users, thereby balancing the network workload across the active IMSs, and relieving the end user of the need to know which IMSs are available. In fact, if a user is logged on to an IMS which fails, a subsequent logon will be directed by VGR to one the of remaining active IMSs.

And finally, there is IMS Shared Queues. The concept behind this is that, even when the logons are equally distributed across all active IMSs, the current, instantaneous workload may not be balanced. It would be nice if a user could enter a transaction on his or her IMS, and if that IMS is really really busy, have that transaction execute on another, less busy IMS. To do this requires a common work queue which functions in a "pull down" manner. That is, when there are available resources to process a transaction, IMSx will "pull down" the next piece of work to be done. This is implemented in IMS V6 using Shared Queues, which exploits List Structures in the coupling facility (see the'IMS Version 6' graphic on the next page) to hold input messages to be processed. Each IMS in the "shared queues group" then functions as a server for these queues. Regardless of which IMS processes the transaction, the response is placed back on the shared queue list structure for delivery to the end user by the IMS that received the input message.



Getting back to our original three benefits:
1. Capacity - the user can start as many IMSs as necessary to handle the workload. This can change from month-to-month, day-to-day, or even hour-to-hour. When the workload increases (say around Christmas time), then additional IMSs can be started. After the Christmas rush, those IMSs can be stopped.

2. Availability - when an IMS in a parallel sysplex shared queues/data sharing/VTAM generic resource group fails, the end user can easily switch to a surviving IMS to continue working by simply logging on again to the generic resource name. Frequently, work submitted to the failing IMS can be processed and delivered to the user by another surviving IMS.

3. Flexibility - IMS systems can be added to or removed from the processing group at will. With careful planning there need be no impact on the end user. Prior to a planned shutdown of an IMS system, users can be migrated to other IMSs so that when IMS is shut down, no users are even aware of it.

With each new release of IMS, additional Parallel Sysplex exploitation is added. In Version 7, for example, IMS will exploit VTAM SNPS or MNPS to provide for "rapid network reconnect" to speed up the reconnection of users following an IMS failure and restart.

You know, it's not just having big hummers that gives us lots of power, and it's not just having bunches of them that gives us availability. It's being able to exploit them as they are needed. Both IMS and OS390 have been architected over the years to excel in managing their resources. Work is delivered to wherever it can be processed. When one component is down, others step in to pick up the slack. The end user, whether it is our old friend still sitting at a 3270, some hot-shot mouseketeer that can't type - only point and click- or the "we don't need no stinkin' batch" nemesis
that refuses to go away, IMS and the S390 Parallel Sysplex can handle anything you can throw at it.

Please see, in this same issue, the article "A bit about batch" by Dave Raften, to see how even that oldest of processing techniques excels in the parallel world.

S/390 and DB2 Query Parallelism: Delivering peak parallel performance

Bryan Smith and Caryn Meyers

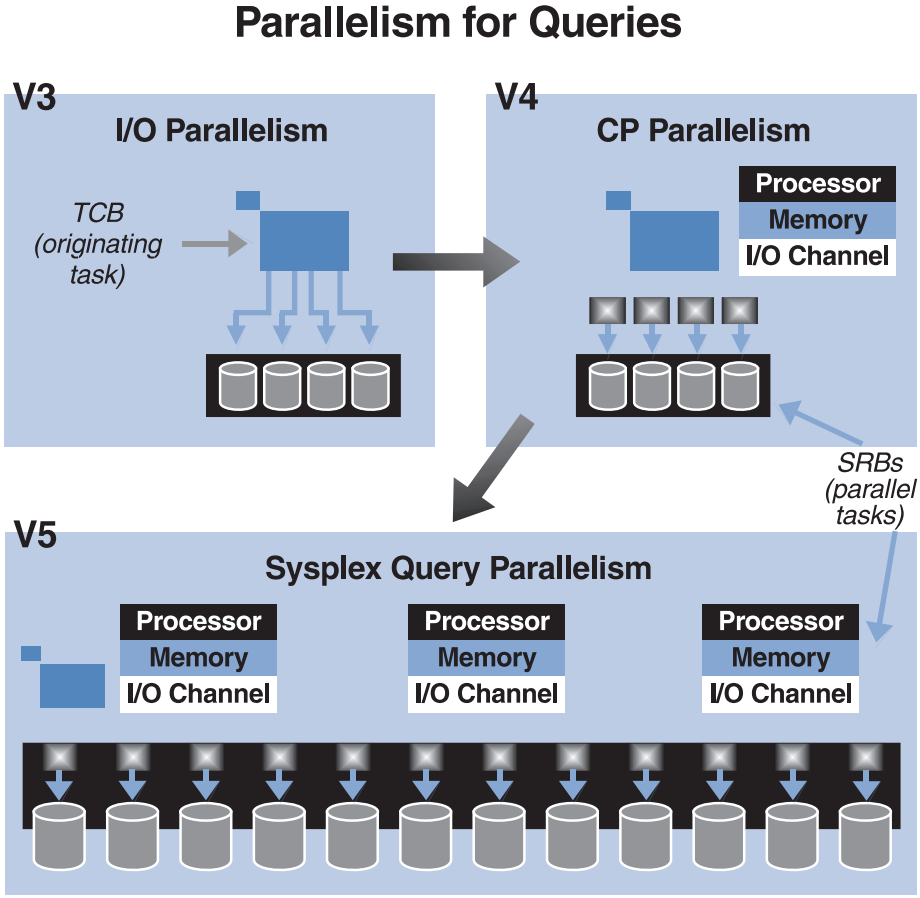
Bryan works as a developer on the DB2 for OS/390 product at the IBM Santa Teresa Laboratory. Bryan formed the Data Technology Institute for S/390 last year, and this year he is on assignment in Poughkeepsie, NY, at the IBM Design Center for e-transaction Processing, representing DB2 across all hardware platforms.

Caryn Meyers has been in Large System Sales and Technical Support for the last 20 years. The last few years she has specialized in Business Intelligence on the DB2 for OS/390 Platform.

Faster response time is key to any business application today. When working with DB2, improving the execution time for a query can be paramount. Often, DBA's turn to tuning a subsystem by adjusting and separating buffer pools, adding indexes and summary tables to the database. However, DB2 offers another alternative, and that is query parallelism.

Query parallelism occurs when the database manager splits a single query into smaller parts, and executes those parts simultaneously. The goal of query parallelism is to reduce the elapsed time of queries by giving those queries access to resources in parallel, such as multiple I/O paths or multiple processors. If the DBMS can assign more processing and I/O power to a query, its overall elapsed time is reduced.

This is equivalent to organizing a team of people to build a house. Sure, one person could do it, but it would take forever. By splitting the construction up into separate discrete components, it can be done in significantly less time. The same can be said for query parallelism.



Types Of Parallelism

The above graphic traces the evolution of query parallelism in DB2 for OS/390. When DB2 Version 3 was introduced, it offered the first form of parallelism with I/O parallelism. Queries, which existed as single executable units of work initiated multiple I/O streams, thereby reducing the time necessary to retrieve large volumes of data from disk. This data was pulled in to the many DB2 buffers, however, a query was still limited to the use of a single engine to read all that data.

This was addressed in Version 4 of DB2, where multiple units of work for each query enabled any number of available engines within a system, to process the data that is retrieved for each engine. This allows all the resources of the system, from the processors, to the I/O channels and memory, to be used in parallel to shorten the execution time of a query. This is illustrated in the above graphic.

Finally, this parallelism was expanded yet again with the delivery of DB2 Version 5. In this latest enhancement, DB2 is able to spread the segments (or subqueries) of a query, throughout all the engines available within a Parallel Sysplex. This enables enormous resources for the execution of a large query. To date, few queries demand CPU resources beyond that found within a single system. However, should a workload demand that level of resource, it is available to the DB2 subsystem through a data sharing environment.

While individual queries may not need to span multiple systems at this time, we have any number of customers currently running large query workloads in a Parallel Sysplex. One large customer,

for instance, has a large workload consisting of thousands of queries accessing over 1 Terabyte of information stored in a central data warehouse. They recognized the need for a dedicated information system for its Customer Relationship Management system, and they choose to install a S/390 environment. Because they partitioned the database to take advantage of DB2 parallelism, their users experienced exceptional response time and reliability for the warehouse system. The S/390 Parallel Sysplex environment met the users needs for accessibility and performance.

How DB2 Determines the Degree of Parallelism

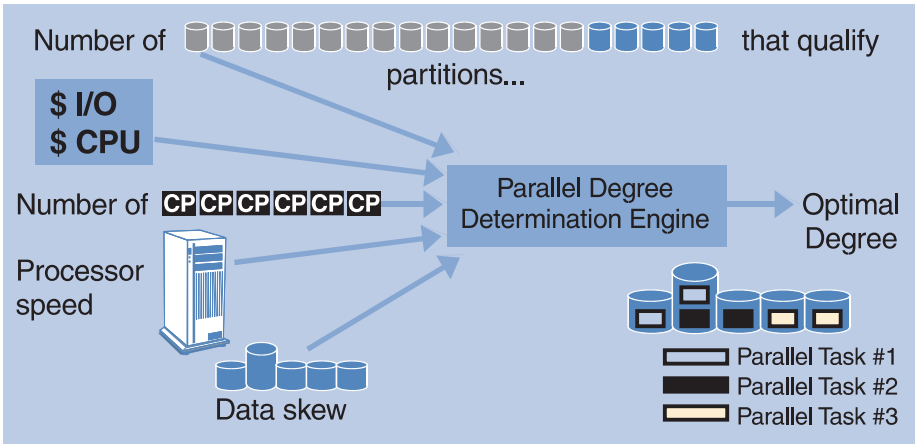
Parallelism is key to performance in a query environment. DB2's cost-based optimizer determines the degree of parallelism for a query with one specific goal: to minimize overhead while achieving the maximum reduction in execution time. It is helpful to compare DB2's approach to degree determination with that offered by other database management systems.

Many other DBMSs choose a degree of parallelism based upon the number of ways the data is partitioned, or they require you to select a degree of parallelism when you define the table or execute a query.

With DB2 for OS/390's approach to degree determination, DB2 considers the following factors when it determines how-many parallel operations are optimal for a query or part of a query:

- The number of partitions
- The estimated I/O cost of the largest physical partition.
- Estimated processing cost based on: the complexity of the SQL; the speed of the central processor; and the number of CPs that are online.

Parallelism Degree Determination



With the shared data model, DB2 UDB for OS/390 has the flexibility to choose the degree of parallelism.

As a simple example, assume that a table space has ten partitions in which the data is evenly balanced. DB2 determines that the I/O cost to read each partition is five seconds, and that the processor cost for the entire query is 60 seconds. If the system is configured with four processors, there is no way to reduce the query elapsed time below 15 seconds, because the query will then be bound by the number of available processors. Setting the parallel degree at more than four does not reduce the overall elapsed time; in fact, it will actually increase the overhead for the query without the benefit of further reducing elapsed time. Recognizing this, DB2 selects a degree of parallelism of four for the query.

DB2 looks at estimated I/O costs and processor costs when determining the parallel degree. The more I/O intensive the query is (scans for example), the more likely that the degree of parallelism will approach the number of partitions. This is because DB2 assumes that each partition is supported by a separate I/O path. For example, assume that you have a table space with 40 partitions on a system with ten processors either within a single

SMP machine, or as the sum total of the processors in a data sharing group.

The above graphic reflects the improvement achieved by implementing query parallelism. In each case, a query was run without parallelism, then, with parallelism enabled. Depending on the I/O and processing requirements of the query, the improvements ranged from 20 times faster execution time, to 114 times faster execution time with query parallelism! Clearly, the execution time improvements can be dramatic!

For an extremely I/O intensive query, DB2 chooses a high degree of parallelism that approaches the number of partitions; for an extremely processor-intensive query, such as one that applies many predicates or requires a sort, DB2 chooses a degree of parallelism that approaches the number of processors. When the degree of parallelism is some where in-between, it means that the query is somewhat balanced in its I/O and processors needs. In the cases illustrated above, there were significant improvements in the execution time of a query through parallelism, for all queries.

Enclave Management of DB2 Queries

To gain access to processors, DB2 coordinates the work of all the parallel tasks spawned from an originating task. Parallel tasks run as one of two types of preemptable SRBs: a client SRB or an enclave SRB. Preemptable SRBs are part of OS/390's enclave services, available with MVS 5.2 and subsequent releases. With preemptable SRBs, the MVS dispatcher can interrupt a task at any time to run more critical other work on the processor.

A client SRB is an SRB that inherits the importance of the address space from which the query originated. For example, if a parallel query originates in a TSO address space, the parallel tasks that are created are managed to the importance of the TSO address space.

Paul Rowntree

An enclave is an entity that can contain multiple executable units and is reported on and managed as a unit. These executable units can be TCBs or SRBs, but DB2 currently uses only enclave SRBs.

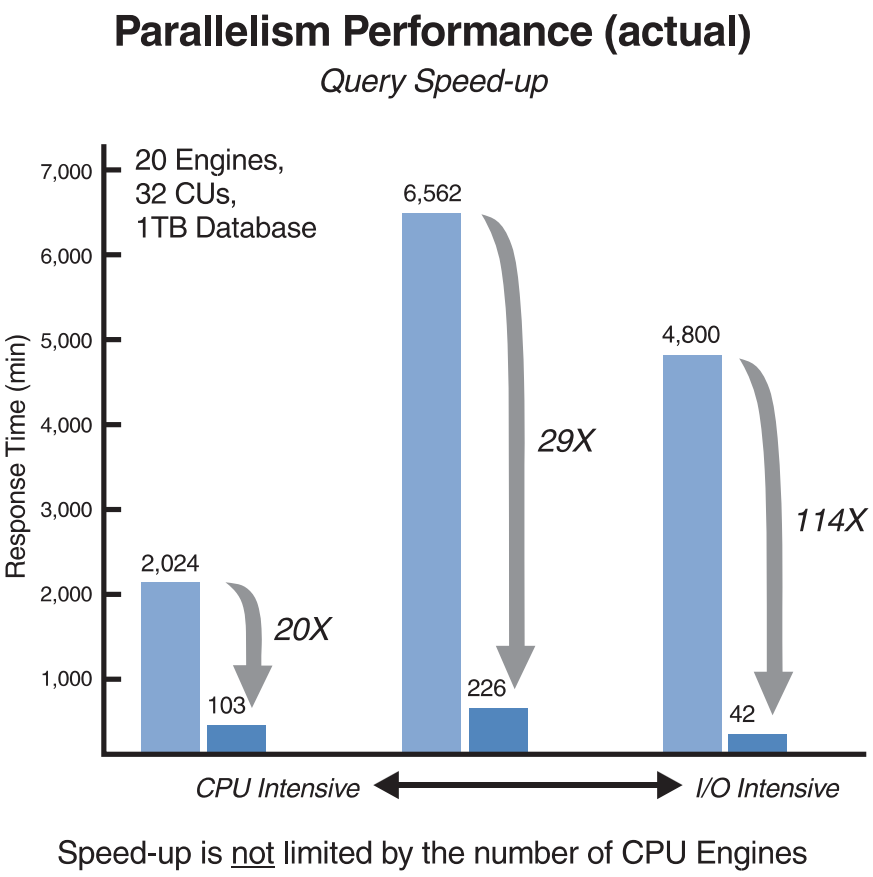
Enclaves are used in DB2 for DDF threads. Using enclaves, you can assign importance to individual pieces of DDF work, rather than having that work managed to the importance of the DDF address space. If DDF work spawns parallel tasks, the SRBs for the parallel tasks are created under that enclave and are managed to the enclave's importance. DB2 also uses enclave SRBs for parallel tasks that are running on an assisting DB2 for Sysplex query parallelism. This allows work to be carry the correct priority, when it is broken into smaller segments for execution.

Workload Management of Query Environments

Enclaves permit DB2 queries and their parallel tasks, to be managed by the OS/390 Workload Management. Workload Manager (WLM)has been imitated on a number of UNIX platforms, however they cannot deliver the efficient capabilities to control the flow of work through a system that is found with OS/390's Workload Manager, particularly in a DB2 query environment.

When a central Data Warehouse or mart environment is enabled to a large population of users, it quickly becomes necessary to differentiate and prioritize queries within the system. When executives submit a query, they need to receive optimum response time. Only OS/390 Workload Manager can guarantee that a user will receive superior response time.

The graphic on the next page illustrates the ability of WLM to deliver consistent throughput, even through the system is filled to capacity. Note this graph consists of a series of lines. Each line in the center graph, represents a different query mix that this customer experienced throughout the week. Each line was drawn from a series of data points, each representing a test case on a S/390 system. Each point represents the number of queries that completed in a 20 minute test, with the specific workload mix for the line. Note the arrows on the graph that indicate when the system reached 100% utilization. Even at system saturation, note that the lines level off delivering consistent throughput for the system. Workload Manager is able to continue to complete work at a consistent level, thereby allowing you to derive maximum value from the system. Notice the smallest blue bar in the lower right corner of the graph. This reflects a similar graph for the UNIX system the customer had been using. Once the system reached its saturation point (noted with the arrow), the throughput drops off dramatically in their system. The system begins to loose the ability to complete work, as it consumes resources internally managing the overload of work, Compare that with the center bar. The query completion on the S/390, never dropped off as dramatically as the alternative platform. This is a direct result of the efficient nature of the S/390 Workload Manager.



Speed-up is not limited by the number of CPU Engines

Summary

The introduction of enclaves was important to allow Workload Manager to identify and manage parallel segments of queries. It is through query parallelism, that you are able to minimize the query execution time for queries submitted by local as well as remote users. This enables a system to perform queries in exceptionally short execution times. These developments have enabled OS/390 environments to deliver similar performance enhancements as those found in other, partitioned environments.

DB2 for OS/390, and OS/390 itself, continue to evolve to deliver one of the most robust query environments available today. The evolution of query parallelism and its ability to capitalize on the Parallel Sysplex environment, as well as the ability of queries to take advantage of the unique strengths within the OS/390 operating system, such as with Workload Manager, highlight the synergy between these products. As Business Intelligence and query workloads become more critical to the fundamental re-engineering of business processes, the strengths of this environment will continue to grow.

Be sure to see the article on "Meeting Expectations in a Business Intelligence Environment" in the Supplement.

Install MPC+ on RS/6000® Gateway for Higher TCP/IP-over-ESCON Throughput

Tom Moore

Either CLAW or MPC+ can be used as the DLC (Data Link Control) on an ESCON-attached RS/6000 gateway to the TCP/IP stack within OS/390. Throughput tests comparing these options indicate that MPC outperformed CLAW by as much as 92% (binary FTP transfers). The throughput gains for MPC+ are largely attributable to the larger IP MTU supported by the MPC+ device driver. Workloads which involve heavy S/390 to RS/6000 bulk data transfers will receive the most benefit.

A summary of the performance experiments, along with CLAW to MPC+ migration tips, can be found under "Tips" at the Communications Server Support site: **http://www-4.ibm.com/software/network/commserver/support/.**

SNA and IP workload balancing in a Parallel Sysplex

Mac Devine

Mac has worked on a variety of APPN/HPR® functions and was the Chief Programmer of several major releases. He is currently a member of the Communications Server for OS/390 Strategy and Design Group. He served as the Chief Designer for Communications Server for OS/390 V2R7.

While SNA and IP application servers are very different, the requirements for effective workload distribution among a clustered group of replicated servers are consistent across both application protocols. These requirements go well beyond finding a simple methodology for the spreading of workload. More importantly, they include the ability to maximize the overall availability of the network, make efficient use of network resources, allow for non-disruptive growth, and balance workload in accordance with business goals. The attributes of the S/390 Parallel Sysplex and its exploitation by the Communication Server for OS/390 are the keys to reaching these goals.

SNA Workload Balancing

The Generic Resources function is the key to effective workload balancing in an SNA environment. Generic Resources allow replicated SNA applications to be known by a single generic name. Communication Server for OS/390 uses the Parallel Sysplex Coupling Facility to provide a data structure that is accessible to all sysplex images so that real-time information about the applications registering under a particular generic name can be shared. This allows sessions to be balanced across all the applications within a generic resource group according to goals defined within the S/390 Workload Manager (WLM). In addition, many SNA applications like CICS and IMS use shared message queues in the coupling facility, which allow them to also balance transaction workload across the replicated applications. Therefore, in many ways, Generic Resources is more important as an availability function than it is as a workload balancing function. In order to fully appreciate the availability benefits of Generic Resources, some concepts need to be understood.

Role of APPN/HPR

Generic Resources relies on the dynamics and networking capabilities of APPN to ensure that sessions can be balanced across a multi-instance application with the assurance that sessions are never distributed to an unavailable instance and without any dependencies on the client or network. Therefore, Generic Resources requires at least one network node (NN) residing in the Parallel Sysplex. As in any good sysplex design, it is important to remove any single points of failure, so it is wise to have two NN's. For all sessions originating from the network destined for a generic resource in the sysplex, the NNs in the sysplex will access the coupling facility structure to retrieve the registered application instances, interface with the Workload Manager to determine the best application instance, and handle all of the necessary routing and directory services needed to complete session setup to the best application instance. While HPR itself is not required for session balancing with Generic Resources, it is a very valuable since it will allow non-disruptive path switches of the session around intermediate failures (e.g. Link). This additional availability can be significantly increased by extending High Performance Routing (HPR) further out into the network so networking solutions like Enterprise Extender (i.e. HPR/IP) are also valuable additions for Generic Resources.

Paul Rowntree

Role of the Application Instance

It is also important to understand that the functions provided by Generic Resources are shared by the Communications Server for OS/390 and the multi-instance application itself. The responsibilities of each application may only include the registering and de-registering of the generic name across its VTAM API or it may also include managing a session affinity. This session affinity is used for certain application protocols to ensure a particular end user will be reunited with a particular application instance on a re-logon attempt following a session failure. While APPN/HPR provides significant availability for Generic Resources, it does not address outages of an application itself or its underlying subsystems (i.e. VTAM, MVS, or hardware failure).

If the SNA session is broken due to such an outage, then the success of a re-logon attempt by the end user has a direct relationship with the presence of a session affinity in the Generic Resources structure of coupling facility. If the affinity is owned by the application instance, which is the case for LU 6.2 sessions, then it survives the outage and therefore a restart of the application instance, either manually or via the automatic restart manager, is required for re-logon success. However, if the affinity is owned by the Communications Server for OS/390 (e.g. SNA 3270 and TN3270), the end user is allowed to logon to another available instance of the application since the affinity

goes away as soon as the original session was broken. This usually means that the session can be restarted in seconds instead of the 10 or more minutes it usually takes to recover the application instance.

IP Workload Balancing

The task of workload balancing in a IP environment is not nearly as straightforward as in an SNA environment due to the wide variety of IP workload distribution choices. To choose the best method for your network, it is important to understand your current and future network requirements as well as your network configuration.
For example, a cluster IP address method for workload distribution, like Cisco's Multi-Node Load Balancer or IBM's Interactive Network Dispatcher, uses a single IP address to represent the replicated application servers and therefore requires that all clients have the same access to the node advertising the cluster IP address. In contrast, a DNS method for workload distribution, like S/390's DNS/WLM, uses a single name to represent the replicated application servers and therefore has no dependency on the network access of the clients. However, it does depend on the clients (and other DNS nodes) honoring the time-left value of zero returned on the resolution (i.e. no caching is allowed). While each of these methods supply workload balancing based on input from the S/390 Workload Manager, they do not factor client/server specific policy or quality of service into the workload distribution decision. These additional factors are vital in a service provider environment where differentiated services for different client groups and adherence to service level agreements are required. To alleviate any dependence on network access or specific client behavior as well as address the service provider environment, the Communication Server for OS/390 has developed the Sysplex Distributor in OS/390 V2R10.

Sysplex Distributor

The Sysplex Distributor function is actually shared among the TCP/IP stacks in the Parallel Sysplex by utilizing R7's XCF dynamics support for inter-sysplex communication and R8's dynamic VIPA support for configuration and recovery. The role of each stack is established by configuring a dynamic VIPA which has been defined with a distribution server list for a particular port or ports. When the 'ALL' keyword is specified in the distribution server list, any TCP/IP stack on an existing or new sysplex image, automatically becomes a candidate for workload distribution. This can reduce the administrative burden significantly in a rapidly changing environment (e-Commerce, for example) by allowing the complete flexibility to move application servers or add new application servers instances to new locations within the sysplex and still be considered part of the same single system image application group. Once a dynamic VIPA becomes a sysplex wide VIPA, workload can be distributed to multiple server instances without requiring changes to clients or networking hardware and without delays in connection setup, thus the data center and the customer's business are allowed to grow non-disruptively.

Configuration

The stack defined as the primary owner of a dynamic VIPA (via a VIPADEFINE statement in the TCP/IP profile) receives all IP datagrams destined for that particular VIPA. If this dynamic VIPA definition includes a distribution server list for a particular port or ports, then the Sysplex Distributor code running in the primary owning stack is activated so that it can distribute connection setup requests destined for that particular VIPA and port(s). The stacks identified as candidates for workload distribution for particular port(s) require no configuration since they are notified of their role via inter-sysplex communication with the primary owner. In order to avoid the primary owning stack being a single point of failure, it is recommended that one or more backup stacks are defined (via a VIPABACKUP statement in the TCP/IP profile). The backup stack(s) can inherit the distribution server list via inter-sysplex communication with the primary owner or can specify an entirely different distribution server list to be used if the primary owner experiences an outage. It is also possible to specify a distribution server list at the backup stack so that distribution only occurs during an outage of the primary owner. This allows the additional workload originally destined to the primary owner to be spread across multiple servers thus lessening the overall impact to the data center during the outage.

Current Functionality

Because the Sysplex Distributor resides in the Parallel Sysplex itself, it has the ability to factor "real-time" information regarding policy, quality of service (QoS) and application status into the distribution decision. By combining these "real-time" factors with CPU utilization information, the Sysplex Distributor has the unique ability to ensure that the best destination server instance is chosen for a particular client connection while ensuring that client/server specific service level agreements are maintained.

Unlike other workload distribution methods, the Sysplex Distributor uses a variety of sources to obtain its distribution decision criteria. In addition to using information

obtained from WLM, it also uses information from the Communications Server for OS/390's Service Policy Agent and information directly obtained from the target stacks themselves. While it is very desirable to factor in the CPU utilization supplied by WLM to understand the workload on a particular system, it's not enough since it does not consider the network performance (i.e. Quality of Service) in its workload balancing algorithm. Network performance is often the bottleneck in the overloaded Internet/ISP network and is a critical factor in the end-to-end response time. Also, enterprise networks often have alternate paths to address network availability and reliability, and yet they're not taken into consideration in the optimization of end-to-end response time and availability and reliability. This makes it difficult for the service provider to adhere to service level agreements.

For example, it may be desirable to route more incoming connection requests to a more loaded server (higher CPU utilization) with better network performance than to a less loaded server with much worse network performance. Therefore the Service Policy Agent will inform the Sysplex Distributor whenever a particular server instance is not adhering to the QoS specified in its service level agreement.
The Sysplex Distributor has also been updated to include policy concerning the clients' characteristics into the workload distribution decision. This policy can include the IP characteristics of the client (IP address and port, IP Subnet, etc.), time of day, day of week, as well as any other policy rule supported by the Service Policy Agent. An example of the usefulness of this function is in application hosting or Internet service provider marketplace where clients, accessing the same application, can be assigned to different servers having different capability and/or connectivity.

The target stacks also assist the Sysplex Distributor in making the best distribution decision possible by supplying immediate server status information via inter-sysplex communication. Whenever an application server binds and listens to a port on a target stack being serviced by the Sysplex Distributor, the target stack sends a notification via inter-sysplex communication to the primary owner indicating that an application server exists and is ready to accept connections. When the application terminates or closes the listening socket, a corresponding notification is sent to the primary owner so that no additional connection requests will be distributed to this stack.
The Sysplex Distributor has up-to-date information on available servers on target stacks so there is no need for application-specific advisors to issue periodic null application requests to determine existence of functioning servers, as is the case with many other workload distribution methods.

In addition to providing workload distribution, the Sysplex Distributor also enhances the availability provided by the dynamic VIPA support. The dynamic VIPA support in R8 allowed a backup stack to takeover the VIPA in cases where the primary stack experienced a system outage. It did not, however, allow non-disruptive movement of the VIPA during normal operations. There was no way to preserve existing connections while relocating an application server which was using the VIPA via BIND or IOCTL DVIPA or while recovering the VIPA ownership at the primary stack upon its recovery from the outage. The non-disruptive VIPA takeover function, which is available in R10, allows for this freedom of movement by maintaining the active connections established with the backup stack and allowing the VIPA to move immediately back to the primary owning stack.
The primary owner will then be allowed to accept all new connections requests and internally use the Sysplex Distributor function to forward the IP datagrams to the backup stack for connections which were active at the time of the non-disruptive takeover. This ensures minimal impact for planned or unplanned system outages since workload can be redirected without affecting existing connections.

Future Functionality

Watch this space for more exciting functions involving the Sysplex Distributor, coming your way in the not so distant future. The best is yet to come!

Learning Services education: WLM and Parallel Sysplex

Ree Howard

For more information about our courses and/or to enroll, call 1-800-IBM-TEACH (426-8322) or visit: **ibm.com/services/learning/spotlight/s390.html**

1. Courses Available for WLM - goal mode:

H4013 - OS/390 Workload Manager Goal Mode Migration - 3 days - $1145
Learn about the OS/390 Workload Manager (WLM) and its associated RMF reports. Lectures include: why move to WLM, the conversion process, identifying workload categories, determining service goals for work, and designing the classification rules. The course features details of the migration process so you can immediately begin the conversion process.

ES540 - OS/390 Workload-Based Performance Management - 4.5 days - $1895
Are you new to performance management? Do you need to know how to establish a practical performance management program for your OS/390 system? As a new performance analyst, learn to work with the Workload Manager (WLM) in goal mode. This course is designed to guide you in learning concepts and performance management in the OS/390 system using the WLM. We will discuss both monoplex and sysplex environments. Learn how to analyze Resource Monitoring Facility (RMF) reports and implement service definitions via the WLM Interactive System Production Facility (ISPF) application. Case studies are used to reinforce the concepts and techniques discussed in lecture. Hands-on lab projects may be done in teams depending on the number of attendees and location.

2. Courses Available for Parallel Sysplex:

H3910 - Parallel Sysplex Overview and MVS/ESA V5 Update - 2 days - $750
This course introduces the S/390 Parallel Sysplex environment. Highlights include: the S/390 Coupling Facility; the software and hardware that support the parallel sysplex data sharing solution; system and subsystem exploiters of the Parallel Sysplex; and the additional functions provided by MVS/ESA® Version 5.

ES290 - Planning for Consoles and Operations in a Sysplex - 2.5 days - $1495
In a sysplex environment, each system monitors all consoles and their routing codes in the entire sysplex, including the system that consoles are attached to. This console monitoring makes it possible to direct messages to the appropriate systems and consoles. While this environment provides a single-system image and a single-point of control support for the sysplex, it can also make operations in a sysplex seem complicated. This course covers sysplex console definition (including NetView® console considerations), command and message management, and automation in the sysplex environment. The course also presents system initialization and recovery management scenarios.

H3995 - S/390 Parallel Sysplex Planning & Implementation for OLTP - 2 days - $845
This course prepares you to take advantage of the S/390 parallel sysplex for data sharing in the Online Transaction Programming (OLTP) environment. You explore the coupling technology that provides the availability benefits of data sharing with improved performance and with reduced cost of computing. Find out how to implement the systems management enhancements to improve multisystem management and increase productivity in a sysplex environment.

H3996 - S/390 Parallel Sysplex Planning & Implementation for MVS/ESA and OS/390 - 3 days - $1045
Learn how to take advantage of the enhancements that MVS/ESA SP 5 and OS/390 bring to the S/390 product. These enhancements address key cost of computing, Client/Server, and open enterprise customer concerns. You will explore the coupling technology that provides the availability benefits of data sharing with improved performance and with reduced cost of computing. Find out how to implement the systems management enhancements to improve multisystem management, and increase productivity in a sysplex environment.

ES420 - Parallel Sysplex Implementation Workshop - 4.5 days - $3850
Learn to lead and assist in the performance activities required for the implementation of the OS/390 Parallel Sysplex environment. Explore the S/390 Coupling Facility technology that provides the availability benefits of data sharing with improved performance and with reduced cost of computing. Implement the systems management enhancements that improve multisystem management and increase productivity in a S/390 Parallel Sysplex environment. Hands-on lab projects may be done in teams depending on the number of attendees and location.

CF350 - DB2 for OS/390 Data Sharing Implementation Workshop - 4.5 $3225
Learn to implement DB2 for OS/390 Data Sharing. Lectures and hands-on labs will teach you how to set up a naming convention, enable a DB2 data sharing group on a parallel sysplex, verify data sharing function, and move a database to

the data sharing group. Lab teams of two or three people go through the implementation steps of data sharing on their own parallel sysplex.

CF360 - DB2 for OS/390 Data Sharing Recovery/Restart Workshop - 4 days - $4595

Active 1999 Education Cards and Extended IBM Education Cards may be used for this course. Learn to plan and implement recovery procedures for a DB2 for OS/390 Data Sharing installation.

CI750 - Introduction to CICSPlex System Manager - 1 day - $495

This course introduces the capabilities, functions, and services provided by CICSPlex System Manager (SM). The definition of the topology for CICSPLEXes and CICSPlex SM is also discussed.

Paul Rowntree

CI760 - CICSPlex System Manager, Administration - 3.5 days - $1745

Learn to install, configure, and use functions of the CICSPlex System Manager (SM). In teams of two, you define a CICSPLEX and a policy for each CICSPlex SM component for Work Load Manager (WLM), Real Time Analysis (RTA), and Monitor (MON). You use Business Application Services (BAS) to define and install resources.

ES810 - Monitoring Parallel Sysplex Performance and Capacity - 2.5 days - $1345

With the exciting growth of the Parallel Sysplex supporting our large system needs, there is a requirement to verify correct operation of the system. This course discusses the coupling facility and XCF sizing and performance issues related to the Parallel Sysplex. Upon completion of this course, you will be able to estimate the resource needs to support the coupling facility and the necessary sysplex connections. Additionally, you will be able to validate correct performance of the features supporting the parallel sysplex such as the coupling facility and the sysplex connections (XCF) after installation.

H4057 - Parallel Sysplex Operations and Recovery - 5 days - $3645

Learn how to operate in the 9672/9674 parallel sysplex environment. Through lecture and hands-on exercises, you will learn setup and operating procedures for the 9672 Hardware Management Console (HMC), gain in-depth problem determination skills, practice enhanced sysplex operations including the Coupling Facility, and utilize recovery capabilities provided by the S/390 parallel transaction server and the 711 based ES/9000®.

ES900 - Advanced Parallel Sysplex Operations and Recovery - 4.5 days - $3495

This course will help you understand and setup advanced recovery techniques in the S/390 CMOS Parallel Sysplex environment. You will have the opportunity to experience various problems in a S/390 Parallel Sysplex via extensive hands-on labs and take the appropriate recovery actions to maintain high availability of the Parallel Sysplex.

Downward compatibility in Language Environment - Problem solved!

Eric Busta and Kershaw Mehta

Language Environment® for OS/390 V2R10 is now downward compatible. This means that with required restrictions and programming guidelines observed, applications developed on higher levels of OS/390 will be able to run on execution platforms with lower release levels of OS/390.

Did you ever get an ABENDFCC (ABENDSFCC) because you linked an application with any OS/390 Version 2 Language Environment release, but ran it on any release of OS/390 Version 1 Language Environment? And the application didn't even take advantage of any new function of OS/390 Version 2, did it?

Unfortunately, this used to happen because a common practice many MVS mainframe customers follow. Many use their application development machine as a testing environment for the newest level of the operating system. The application programmers are not only developing and maintaining applications, but invariably testing the newest level of the operating system. However, these applications have a different release schedule to the production environments than the operating system.What ends up happening is that customer applications are compiled and linked on a higher level OS/390 system while they are executed on a lower level of OS/390. In those days you would have incompatibilities a plenty, better known as "downward compatibility" problems. But not anymore. We've felt your pain and came up with an antidote!

Prior to the solution we provide in OS/390 V2R10, IBM had prescribed a remedy which involved saving the older level of OS/390 Language Environment libraries. While we knew this was a tactical solution, we provided instructions (as documented in INFO APAR II11316) to allow system programmers to keep the lowest level of the Language Environment libraries currently executing in the enterprise, on the development machine. To prevent any downward incompatibilities, application developers were then required to use these saved libraries during coding, link-editing, and testing before deploying the application on the production machines.

Kerry Collia

See the related XPLINK article on page 50

But now that our strategic solution is available, customers no longer need to save around lower levels of the Language Environment data sets. We have eliminated this step altogether with our unique design. We implemented our solution mostly through toleration PTFs. These toleration PTFs not only correct the incompatibilities but also provide diagnosis assistance when new functions are used in applications; functions which were available on the higher level of OS/390 on the development machine, but not available on the lower OS/390 level of the production environment. For C/C++ programmers, we have enhanced the TARGET compiler option to prevent new functions to be coded in the application, specifically those unavailable on the targeted system. Changes to our system headers and to the C/C++ compiler allow this to work.

The Language Environment downward compatibility support is available to all applications written in COBOL™, PL/I, C/C++, and Fortran. Details of the diagnosis assistance, programming guidelines, and restrictions can be found documented in the OS/390 V2R10 level of *Language Environment Programming Guide*. A list of toleration PTFs are provided in the OS/390 V2R10 PSP bucket.

To avoid confusion, you should be aware of what "downward compatibility" does not mean. "Downward compatibility" does not mean that new functions are rolled back to old releases. Applications developed with the intention of being downwardly compatible must not use Language

Environment function that is unavailable in previous releases. It also does not change Language Environment's upward compatibility. Upward compatibility has always been a part of Language Environment and will continue to be in OS/390 V2R10. This means that applications coded and link-edited on one release of OS/390 will continue to work on later releases.

So what does one need to do start taking advantage of this new function? For system programmers it means getting OS/390 V2R10 installed on the development machine as soon as possible. In addition, toleration PTFs will need to be installed on production machines. Application developers on the other hand,will need to ensure that they are using this new level of OS/390 Language Environment when compiling and link-editing their applications. The rest of our support is transparent to the development community, which means that **no** analysis nor any redevelopment of current applications is necessary.

Finally, what you can take away from all of this is that OS/390 V2R10 Language Environment is both upward and downward compatible. IBM is committed to provide this support with future releases of the operating system and Language Environment.

Using XPLINK to improve C/C++ program performance

Dave Sudlik and Barbara Neumann

Do you want to reduce overhead for C and C++ function linkage in your programs? Using Extra Performance Linkage (XPLINK) with the OS/390 C/C++ compiler and Language Environment for OS/390 V2R10 will help you do just that.

XPLINK is a new call linkage between programs that can significantly improve your program's performance when your environment contains frequent calls between small programs. XPLINK makes function calls as fast and efficient as possible by removing all nonessential instructions from the main path. Think of it as a performance power-pack to boost your program's processing time.

Applications written in the C++ language tend to involve many small function calls. Also, modular applications written in C may also have high linkage overhead. XPLINK, by reducing linkage costs by up to 50%, can improve the performance of these applications. The performance improvement for a given application depends on how much time it currently spends in function call overhead.

Kerry Collia

XPLINK significantly speeds up the linkage for C and C++ routines by using a stack that grows from high addresses to low addresses in memory (that is, a downward-growing stack) and by passing parameters in registers. It reduces the size of the program that is loaded into memory, allowing you to load more functions into memory. It supports reentrant and non-reentrant code and calls to functions in Dynamic Link Libraries (DLLs). When all functions are compiled with XPLINK, you can use function pointers without restrictions, which assists you in porting new applications to S/390. Compatibility between XPLINK and non-XPLINK programs is provided. This allows XPLINK programs to communicate with programs built with the standard linkage using, for example, a DLL call. In this mixed linkage environment, some of the performance gains of an XPLINK application may be offset by the cost of calling legacy programs until they are rebuilt using the new linkage conventions. The maximum performance improvement can be achieved by recompiling an entire application with XPLINK.

The DFSMS 1.6 Binder is used to produce XPLINK executables that must reside in a partitioned data set extension (PDSE) or the hierarchical file system (HFS).

XPLINK enhances the OS/390 platform for e-Commerce applications. XPLINK is available for applications running under OS/390 batch, TSO/E, and OS/390 UNIX System Services. You can also use IBM Debug Tool or the OS/390 UNIX dbx debugger to debug applications that use the XPLINK linkage conventions.

While XPLINK can provide a significant performance enhancement to the right kind of application, it can also degrade the performance of an application that is not suitable for XPLINK. An instance of this is an application that is predominantly coded in COBOL or PL/I.

Attention: Language Environment Run-Time Options

The ABTERMENC run-time option has a new default sub-option. It's been changed from RETCODE to ABEND. So if you're migrating to Release 10 and want to retain RETCODE, you'll have to code it. But we've always recommended the ABEND sub-option and if you've selected this in the past, relax. You have one less thing to do when migrating. We've done it for you. This support was also made available with an APAR in Release 9.

Searching for a better way to locate message explanations for OS/390?

Jim Steipp

Step right up ladies and gentleman. Don't crowd. Have I got a deal for you! Remember the LookAt message help currently available on OS/390 hosts? It's the help facility that doesn't expect you to know anything about the library or how to construct a search argument. You only need to know the number of the message that you want to find.

Well, we're prepared to go one better. Today and today only, to readers of this issue of Hot Topics, we are able to offer this special deal. Starting June 30, IBM will make available, on our OS/390 Web site, a version of LookAt that will show you OS/390 message information using just your browser. Just point at:

http://www.s390/ibm.com/os390/bkserv

Click on "LookAt Messages," and, at the risk of repeating ourselves, take a look at LookAt. You'll like this one.

So, what gets XPLINK started?

By Bruce McLaughlin

The first thing to do is compile as much of your application as possible with the C/C++ XPLINK compiler option.

If any parts of your application are compiled with the C/C++ XPLINK compiler option, you need to have the XPLINK(ON) run-time option either implicitly or explicitly specified. If your main() function is compiled with XPLINK, XPLINK(ON) is implicitly set and there is nothing else to do. Otherwise, you enable XPLINK at run time by explicitly specifying XPLINK(ON).

The STACK and THREADSTACK run-time options control the allocation of both the standard linkage upward-growing and XPLINK downward-growing stacks.

The THREADSTACK run-time option replaces the NONIPTSTACK run-time option. The NONIPTSTACK run-time option is deprecated but remains for compatibility, but without the XPLINK downward-growing stack support.

SUF can benefit your service strategy

Jim Steipp

Are you suffering over servicing your host systems? Let IBM take the SUF out of it and make your job a lot easier.

IBM's S/390 Service Update Facility (SUF) can enhance your ability to place and receive orders for OS/390, VM, and VSE service. It lets you order both preventive and corrective service using your Web browser. Optionally, you can chose to have orders delivered either over the Internet, on tape, or by satellite if you have an Advanced Digital Delivery receiving workstation. However, VSE preventive service (that is, system refreshes) can only be delivered on tape. Of course, you must be an IBM customer and have a vaild service agreement in place to use SUF and you will need to register to validate you eligibility, but you can do that over the Web too.

What Do You Have to Do?

You have to set up a Customer Application Server on a Windows NT® or OS/2® workstation server or on an OS/390 UNIX System Services (host) server. The SUF Customer Application Server, using the IBM Internet Connection Server, receives the service packages that you ordered if you chose the Internet delivery option. When the packages arrive, they are uploaded to the hosts for which they were ordered.

To place service orders from your workstation, you must be sure that your workstation can access the SUF Customer Application Server through TCP/IP. You must also have a supported Web browser available on your workstation. Your hosts must also be connected to the Customer Application Server in one of the following ways:

For OS/390 and VM:
TCP/IP
For VSE:
APPC, LU2, or TCP/IP

You must be sure that the SUF OS/390 and VM host programs are installed and configured on each host that is to receive the service packages that you order. VSE does not use a host program.

What Will SUF Do for You?

When you place an order through SUF, it creates and submits to IBM an OS/390 consolidated software inventory (CSI or bitmap) or a VSE History File that describes the software that is actually installed on your OS/390 or VSE hosts. IBM uses this inventory or history file along with your order to send you only the updates you need minimizing the amount of service that needs to be applied. With each order, IBM includes the latest enhanced ++HOLDDATA. You can also use SUF to obtain just the latest OS/390 Enhanced HOLDDATA from the IBM HOLDDATA Web site.

How Do You Order Service?

With SUF installed and running, you can begin ordering service for your hosts. In general, here's how:
1. Specify the host that is to receive the service.
2. Specify the type of service.
3. Select the preferred and alternate delivery methods.
4. Submit your order.

You can review the status of your orders while they are being processed and you are waiting for them to arrive. Here's how IBM handles the order for you:
1. Your order is received at an IBM Service Center.
2. The order is prepared and packed.
3. Depending upon the delivery method you selected or the size of the package, the SUF Customer Application Server retrieves the order from IBM over the Internet or it is delivered by alternate means.
4. The SUF Customer Application Server automatically uploads your to order to the host you chose. Depending on your host, the order is received and installed.

SUF can save you time and make servicing your hosts much less complicated. You do not have keep track of the service that your hosts need, SUF can get it for you. You will only receive the service that's needed. SUF automatically uploads the service to your host. It's quick; it's easy, and it's available at no cost.

Want to Get More Information?

For more information about SUF, how to register, and how to download the code, go to the S/390 Service Update Facility Web site at:
www.s390.ibm.com/suf

Need time for a coffee break? Use S/390 Service Update Facility!

Cheryl Loughlin

Wouldn't it be nice if you had a quick and easy way to order and receive service for your OS/390, VM®, and VSE® systems? Well, with S/390 Service Update Facility (SUF), you will have just that! When you use SUF, you'll actually have more time to be productive doing other important tasks, such as taking a much needed coffee break!

SUF is an Internet-based S/390 software service tool that lets system programmers order both corrective and preventive service over the Internet and receive it in any of three ways: over the Internet, on standard physical media, or by Advanced Digital Delivery (satellite) where available. When using SUF, system programmers will benefit from a simplified process for S/390 software service maintenance, increased system stability through use of recommended service levels, and productivity gains. See the following article for more details.

We have developed an interactive multimedia CD-ROM titled *S/390 Service Update Facility: A day in the life of a system programmer.* With video and animation, this CD-ROM highlights the many benefits of SUF and provides details about how to install and use SUF. You can get your own copy of this CD-ROM and ask our SUF representatives any questions you may have when you visit the SUF booth at the following trade shows this year:

SHARE: July 23–28, 2000, Boston, MA

MS390: October 2–6, 2000, Orlando, FL

OS/390 Expo and Performance Conference: October 23–27, 2000, Washington, DC

WAVV: October Oct 6–10, 2000, Colorado Springs, CO

For more information about SUF, grab a cup of coffee and check us out on the Web at **http://www.ibm.com/s390/suf/**

A new Redbook!

Andrew Schmidt

This past fall, the IBM PartnerWorld for Developers S/390 team funded a project to produce a Redbook that deals with the planning and implementation required to deliver a ported product using Java for OS/390. The Poughkeepsie, New York-based project was led by Alex Louwe Kooijmans of the International Technical Support Organization (ITSO). Redbooks, as you might know, are produced by the ITSO, and you can find their summaries, read, and order them from the ITSO website **http://www.redbooks.ibm.com**. This particular Redbook, entitled "Experiences Moving A Java Application to OS/390" (SG24-5620) was written on behalf of our ISV (Independent Software Vendor) community that my group supports. But it should have interest and relevance to those programmers and systems programmers that are involved with Java for OS/390, one of today's more interesting, and fairly new, technologies on the S/390 platform. Although the book was written from a porting point of view of a Java-application from a non-S/390 platform to the S/390 platform, many of the challenges of the project and the technologies and features used and described in the redbook should be of interest to you even if developing a Java application from scratch.

The Redbook opens with a discussion of the overall suitability of the application to Java for S/390. Despite the claim of Java's "write once, run everywhere" model, there are still some 'gotchas' in Java implementation on many platforms. In S/390, for example, two areas to be concerned with are those applications that deal with, expect, or are sensitive to ASCII character encoding or code that contains calls to native methods that are part of the application. Redbook "Porting Applications to the OpenEdition MVS Platform" (GG24-4473) (although now a little dated - see the website http://www.s390.ibm.com/oe for more recent documentation) contains guidance in this area. Additionally, guidance is offered to ensure that the programming and operational environments in which the application will ultimately be deployed are considered, such as ensuring that any classes or packages that you need to use must be verified as to being available on S/390 as well.

The actual coding project is then described in detail, beginning with obtaining the latest JDK for OS/390, validating it, and installing a development application environment on a client, if needed (we chose IBM's VisualAge for Java, natch!).

Despite a great anticipation that your project will go forward without problems, that may not be the case, and an entire chapter is devoted to a discussion of the debugging facilities available to you, such as the JDK debugger tool, the debugger tool in IBM Visual Age for Java, and the IBM Distributed Debugger. An area of importance with large-scale applications in general, especially when dealing with Java, is how well the application ultimately will perform when being delivered from a typically small development environment to the S/390 production implementation. An entire chapter in the book is therefore dedicated to Performance Analysis and Tuning. Hints and tips concerning coding practices, variable and API class usage, available Java facilities, and application packaging techniques are discussed. Additionally, information on performing analysis of the code is provided as well by using some of the tools available on S/390 for this purpose.

In summary, if you're interested in exploring Java as a new technology on S/390, this book of 'how-tos' and real-world examples of how OS/390 plays in this arena should be part of your bookshelf and required reading!

Softcopy printing news

John Sefcik

In OS/390 Release 10, the Softcopy Print element was removed from OS/390. Why is that you say? Well, the Softcopy Print element was for printing BookManager books that resided on the OS/390 host to AFP™ printers. This function was included in OS/390 back in Release 2, when we were just starting to develop an OS/390 Web site, many fewer customers had Internet access, BookServer was not available on OS/390, and there were no PDF files of OS/390 books yet. With all of the advancements in softcopy searching, viewing, management, and printing over the past four years, as well as the additional exploitation by customers of the Internet and intranets, the Softcopy Print element is no longer a strategic method for printing.

Some options that you can use for printing are:
Print sections of BookManager® books (served by BookServer) from the OS/390 Internet Library (**http://www.ibm.com/s390/os390/bkserv/**) using the print function of your browser. Use PDF files of the OS/390 books for printing. You can obtain the PDFs from the PDF Library Collection CD-ROM (SK2T-6718) or the OS/390 Internet Library. For those of you who wish to continue using the Softcopy Print function to print BookManager books on AFP printers, see the OS/390 Planning for Installation book for information on the products that you would need to obtain, if you do not already have them.

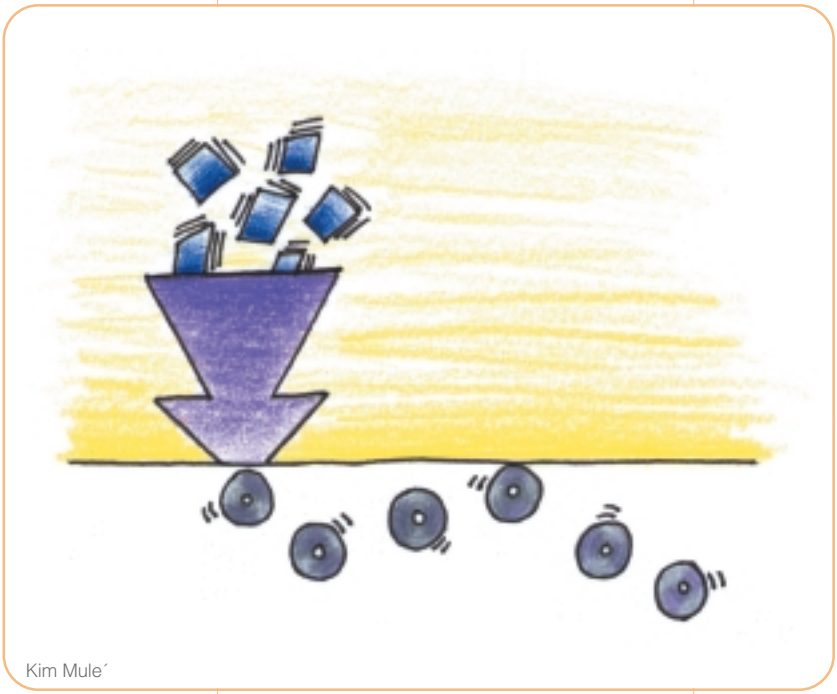
As an environmentally friendly tip, search the softcopy to find what you want, and then print the sections that you prefer to have on paper.

Where have all the hardcopy collection books gone?

Shirley Swenson

When you receive a copy of the OS/390 Collection, September 2000 edition, you undoubtedly notice the new streamlined packaging. Yes, streamlined — no more hardcopy books ship with the CD-ROMs.

Now don't get excited! We know that many of you relied on the collection catalog book and used it as an index into the humongous OS/390 Online Library Collection. We haven't taken this book away entirely, only the hardcopy. We continue to include four softcopy versions of the catalog (index) on each CD in the collection, and the softcopy versions are better than hardcopy because you can search them, link directly to books if the correct CD is in your drive, or even print a high quality hardcopy if you still need that security. But try the softcopy formats first. We think you'll agree that they are even better than the hardcopy.



Kim Mule´

The following softcopy formats of the index are on each CD in the collection:

- HTML for viewing with any Web browser, including links that enable you to open the softcopy files for the libraries and manuals listed, if you have inserted a CD containing the book
- BookManager for viewing and searching using any IBM reader or supported BookManager READ product
- PDF for viewing or printing a high quality hardcopy of the catalog using one of the Adobe Acrobat Readers
- ASCII (text) for viewing using any text editor, if you don't have a browser or softcopy reader installed

So, how do you find the softcopy index?

- All formats are in the root directory on each CD: The files are named SCINDEX.HTM, SCINDEX.BOO,SCINDEX.PDF, and SCINDEX.TXT.
- The BookManager format is also available in its own bookshelf, Booklet for OS/390 Collection (filename EZ239Sxy), on each CD-ROM.
- If you use the new autorun program for installing or updating softcopy tools provided on the tools disc, disc 1 of the Septem ber 2000 edition, you can create a shortcut on your Windows desktop to the collection index.

Insert any CD from a September 2000 collection in your CD-ROM drive, click on the shortcut, and the HTML version of the collection index is displayed.

Likewise, the other hardcopy books previously shipped with the CDs, *Installing, Managing, and Using the Online Library* and the *OS/390 Collection Summary*, continue to be available in softcopy on the CDs. Look for the various formats of these books in the root directory of the CDs: INSTALL.HTM, INSTALL.PDF, INSTALL.TXT, SUMMARY.PDF, and SUMMARY.TXT. BookManager formats are also available in bookshelves on the CDs. The installing book is also accessible from a "Read about..." option of the softcopy installation program on disc 1.

And for those who really need the hardcopy and don't want to print their own copies, you can now order the hardcopy separately:

Collection booklet: GC23-3292
Collection Summary: SA22-7455
Installing, Managing, and Using the Online Library: GC31-8311

Upfront access to softcopy!

Shirley Swenson

Not only have we streamlined the CD-ROM packaging by eliminating the hardcopy books that accompanied previous editions, we've added a tools disc, disc 1, to the collection and a **new** autorun program for installing and updating softcopy tools for Windows users. The installation program can help you:

Tracy Bondi

- Check the levels of installed softcopy tools, such as softcopy readers and the IBM SoftCopy Librarian (SCL), and recommend what should be installed
- Install the recommended softcopy tools or a list of tools that you modified
- Place a shortcut to the collection catalog (index) on your Windows desktop
- Find specific books and technical information more easily than with a hardcopy catalog of the collec tion contents
- Obtain online help and general information about softcopy and tools so hardcopy documentation is not required

In future editions of Online Library collections (except single-disc collections), disc 1 will be the tools disc. This means that the actual softcopy book files now start on disc 2. The second disc includes the current OS/390 release books—OS/390 Version 2 Release 10 books in September 2000. Other discs include books for Parallel Sysplex, OS/390 related software products, and the next most current OS/390 release (currently Release 9).

Note that the September 2000 edition includes only the two latest releases of OS/390 books, instead of four releases. Books for the older releases are unchanged and still available on prior editions:

Release 1 and 2 books: SK2T-6700-07 (December 1997)
Release 3 books: SK2T-6700-11 (December 1998)
Release 4 books: SK2T-6700-13 (June 1999)
Release 5 books: SK2T-6700-15 (December 1999)
Release 6, 7, and 8 books: SK2T-6700-17 (June 2000)

You can access all OS/390 libraries on the World Wide Web by clicking on "Library" at http://www.ibm.com/s390/os390/.

An overview OS/390 R10

The following new enhancements make OS/390 Release 10 one of the best ever:

- Ease in porting C and C++ applications to S/390 through Extra Performance Linkage
- Application development flexibility through downward compatibility of Language Environment
- Better security in the network operating environment and the TCP/IP environment
- Increased flexibility in data storage and management
- Standard workstation access to SAM, PDS(E), and VSAM files
- Easier systems management through enhanced performance reporting
- Wizard technology, multimedia instructional animations, and Automatic Alter support to advance ease-of-use

For details, check out the OS/390 Web site and look for the May16 Announcement letter or the *OS/390 V2R10 Introduction and Release Guide.*

OS/390 Enhanced Custom Build Product Delivery Offering, 5751-CS3, replaces "stand alone" product media offerings.

Dave Weaving and Stephen Saroka

Starting in September, 2000, IBM will offer software products via "customized offerings" instead of the "stand alone" product media options available today.

The added benefits to you are the most current product service available, ensuring product stability, shortened product installation time and new options simplifying selection of service for only the products or selected releases of OS/390 you ordered.

CBPDO currently provides service for all releases of each product licensed under your customer number, as well as non-integrated service for the ordered products. The following new options will be available when ordering the enhanced CBPDO, allowing you to customize the level of service provided as well as reduce the overall size of the deliverable:

Product-only Selective Service (for MVS SREL orders only)

If you choose the Product-only option, your order will include the products you selected as well as all the non-integrated service applicable to these products. Your order will not include service for any other products licensed under your customer number.

If you choose the Selective Service option, your order will include service only for those releases of OS/390 you select, as well as all releases of the non-OS/390 products licensed under your customer number. In addition the order will include all the non-integrated service for any products you selected.

Kerry Collia

The OS/390 UNIX Configuration Wizard

Elizabeth Holland Kern

For best results, you know you should read every word when installing and customizing ... but who has the time to read every word? The OS/390 UNIX Configuration Assistant is designed to reduce the time and decisions required to set up OS/390 UNIX in "full function mode" and start testing your system. It's a response to customers who have said:

- "Why do I have to read so many books?"
- "This is all new to me. I don't have any clue about what value to pick."
- "Prompt us in dialogs — don't just say 'Add this to that parmlib member!'"

The tool is intended for system programmers, system administrators, and security administrators who are doing a ServerPac installation and, for the first time, are configuring OS/390 UNIX System Services. It assumes the audience is familiar with OS/390 installation, but not necessarily with OS/390 UNIX or RACF® customization. Even if you have already configured OS/390 UNIX before, you can use the tool to compare and verify some of your configuration settings.

There are four interview topics: your system environment, OS/390 UNIX processing values, file systems, and initial security setup. Based on your responses, the tool generates "best practice" settings that you could have determined by reading through various product manuals.

The tool builds:

- A BPXPRMxx member with parameters that define your system settings
- A second BPXPRMxx member with parameters that define the file systems
- A sample job that contains the initial RACF commands you need to start your OS/390 UNIX system in full-function mode. If you are using a non-IBM security product, you can use our definitions as guidance.
- Sample HFS files
- Checklists of follow-on actions with links to the appropriate documentation in OS/390 UNIX Planning and OS/390 Communi cations Server: IP Configuration .

After reviewing the output, you can upload the parmlib members, JCL jobs, and HFS files to your S/390 system, install them and begin testing. Before you can fully use OS/390 UNIX, further customization steps are required, and they are outlined in a checklist that the Configuration Assistant produces.

The Release 10 version of the Configuration Assistant will be available by GA of OS/390 V2R10.

So test drive the OS/390 UNIX Configuration Assistant (also known as a "wizard") today ... at **http://www.ibm.com/s390/os390/wizards/**
and send us your feedback (elizab@us.ibm.com)!

Here's where you can find Hot Topics

- Hardcopy of the Newsletter is included in each "OS/390 Installation Planning Kit" that you get free of charge when a new release of OS/390 is made generally available (twice a year).
- We will continue to distribute hardcopy at user group meetings like SHARE.
- It is available from the OS/390 home page at **http://www.s390.ibm.com/os390/**
- It is also on each copy of *OS/390 PDF Library Collection* (SK2T-6718).
- It is also available to order hardcopy through your regular mechanisms at a nominal fee.

An OS/390 Newsletter - Issue 3 August 2000

Executive Editor ..... Bob Ward
Managing Editor ..... Dick Wagenaar
Creative Director ..... Gene Posca

© International Business Machines Corporation, 1999, 2000

Produced in the United States of America
8-2000
All Rights Reserved

The OS/390 Hot Topics Newsletter is published twice yearly as a service to OS/390 customers, providing articles of interest to the OS/390 community.

The information in this Newsletter is presented as is and is based on the best knowledge of the authors. No warranty is provided (neither expresssed nor implied).

IBM may not offer the products, features or services discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM representative for information on the products or services available in your area.

® IBM, BookManager, DB2, ESCON, Language Environment, OS/2, OS/390, Parallel Sysplex, RS/6000, S/390, VTAM, CICS, PR/SM, FICON, AFP, GDPS, Enterprise Storage Server, IMS, NUMA, CICSPlex, APPN, MVS/ESA, NetView, ES/9000, VM, VSE, and RACF are registered trademarks of International Business Machines Corporation in the United States, other countries, or both.

™ MVS, RMF are trademarks of International Business Machines Corporation.

Java, and all Java-based trademarks are trademarks of Sun Microsystems, Inc., in the United States, or other countries, or both.

Microsoft, Windows, Windows NT are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark in the United States and other countries licensed exclusively through The Open Group.

Other company, product, and service names may be trademarks or service marks of others.

GA22-7431-02

OS/390 TOPICS supplement
August 2000

What the Heck is Parallel Sysplex Clustering Technology Anyway?

Madeline Nick and James J. Guilianelli

IBM's® Parallel Sysplex® Clustering Technology is a mouthful - but what it offers you is simple: the highest standard for multisystem performance and availability at the lowest cost.

Parallel Sysplex Clustering Technology - or Parallel Sysplex for short - builds on the strengths of the S/390 platform, its robustness, reliability, and security. Through a unique combination of hardware and software, Parallel Sysplex provides a way to cluster systems and support multisystem resource sharing and read/write data sharing resulting in efficient **dynamic application workload balancing**. The benefit of this data sharing and workload balancing is that, in a Parallel Sysplex environment, work can be directed/re-directed to an available server in the sysplex "on the fly." The Parallel Sysplex by its very nature permits servers to be dynamically added to the sysplex cluster without requiring costly downtime AND without having to split applications or databases across multiple servers. Therefore, processing power can be added when you need it, without having to make significant investments in standby capacity. Additionally, the beauty of this "Single System Image" clustering technology, is that when you do add capacity non-disruptively, you do not need to add staff or new systems management tools to manage that added capacity; your environment is already set up for this scalability.

Parallel Sysplex Parallel Sysplex Parallel Sysplex Parallel Sysplex

Did you know?

In fact, you can scale that processing power by adding up to 32 S/390 servers, each with more than 1,600 MIPS, to the Parallel Sysplex. With that kind of scalability and flexibility, the Parallel Sysplex can handle any unpredicted capacity demand caused by increased volumes or a failure - and it grows as your workloads grow.

Moreover, with its multisystem database sharing, the Parallel Sysplex™ provides industry leading availability: five nines of availability, which means only five minutes of down time a year.

If a system fails, other systems can pick up the work so that unplanned outages are virtually eliminated. Additionally, you can even set up a multi- site Parallel Sysplex for either disaster recovery or for a planned site reconfiguration with a Globally Dispersed Parallel Sysplex or GDPS™. For more on GDPS, refer to the article contained in this Supplement.

Now that you know what a Parallel Sysplex can provide you when you fully exploit data sharing, let's step back and take a look at what you need to build one! We will look at S/390 hardware and how it has improved over time to integrate components for simplification and reduced cost. Then a quick look at the software and how even the new application environments can take advantage of Parallel Sysplex's power. Finally, we will end with S/390® resource sharing, for those customers who are not ready to jump right into data sharing but want to use the OS/390® software and S/390 hardware to its full potential. Additionally, S/390 Resource Sharing will position you to take advantage of the numerous enhancements coming in the near future!

The Hardware

The centerpiece of the Parallel Sysplex is advanced (patented) hardware that makes read/write data sharing among multiple systems possible: a Coupling Facility. Using S/390 Parallel Enterprise Servers, you create a configuration by linking the processors to a Coupling Facility through high-speed Coupling Facility links. Sysplex Timer hardware completes the picture by synchronizing time-of-day clocks between multiple processors in the cluster.

The diagram to the right demonstrates some advancements IBM has made on the physical topology of a Parallel Sysplex via integration of hardware technology with the G5/G6 servers. On the left we have the pre - G5/G6 technology and on the right, the G5/G6 technology.

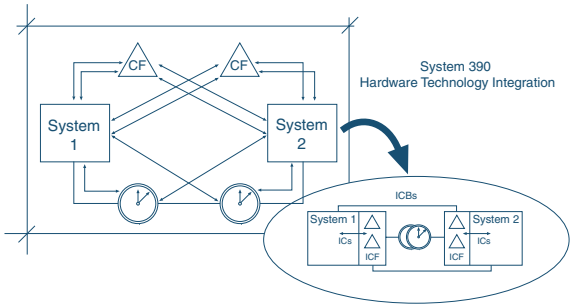
The coupling facility can be a:

~ 9674 Coupling Facility Model C05

~ 9672 R06

~ or an S/390 Parallel Enterprise Server (latest is a G6) with an Internal Coupling Facility (ICF)

~ or a Parallel Enterprise Server with a logical partition defined as a coupling facility

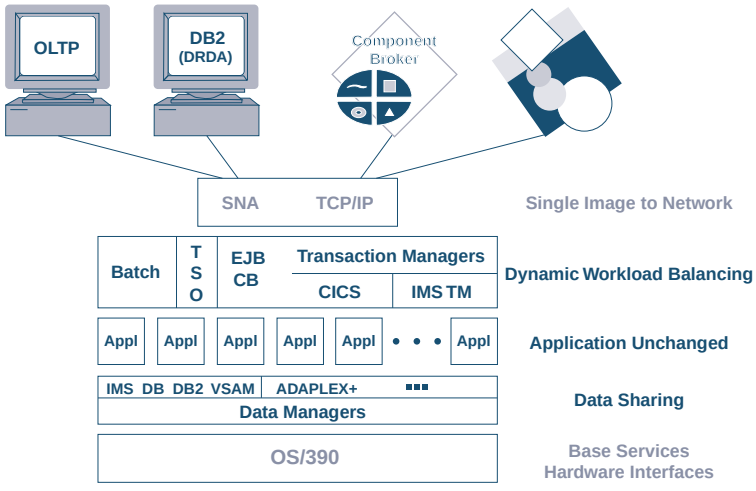


G5/G6 Exclusives (ICBs/ICs) for an optimized Resource Sharing Environment

What this provides is flexibility - in your choice of processor, Coupling Facility model, and links especially with the new G5/G6 technology. IBM allows the links to be integrated in the hardware (linkless configurations) or shorter distance copper links that are almost 3 times the speed of the latest Hiperlinks. For all the gory details on Coupling Facility options and their advantages, go to our Web site, **www.ibm.com/s390/pso**, and pull off the paper on "Coupling Facility Alternatives: A positioning paper".

The Software

The S/390 software community has been focused on insulating your applications and giving you the benefits of Parallel Sysplex clustering technology as transparently as possible. This is accomplished by delivering exploitation of the Parallel Sysplex technology in the OS/390® base , S/390 software subsystems and Communication Server. How such exploitation provides transparent value to the customer business applications is best understood through examination of the OS/390 software stack.



Parallel Sysplex Software Structure

At the base operating system layer, OS/390 provides physical resource management of the Coupling Facility resources in the Parallel Sysplex configuration. Multi-system data-sharing is provided by the database managers on OS/390 - DB2®, IMS®, VSAM, and OEM vendors such as Software AG's ADAPLEX+®. These subsystems exploit the CF transparently to their clients and enable concurrent sharing of database resources across systems.

Dynamic workload balancing of work requests across systems based on available capacity is the responsibility of the transaction managers such as CICS®, IMS and workload manager (WLM) on OS/390. Customer applications which stay within the execution shell delivered by these subsystems (accessing system resources only via subsystem-provided APIs), are able to grow seamlessly across systems without application re-engineering.

At the network layer, the systems in the Parallel Sysplex are collectively viewed as a single large system to clients. SNA sessions and TCP/IP socket connections are dynamically bound to systems in the sysplex through load-balancing software provided by VTAM® and the OS/390 TCP/IP stack. (Refer to Mac Devine's article in the OS/390 HOT TOPICS Newsletter for additional information on some new enhancements in this area.)

Through this structured layering of the OS/390 system and subsystem components with leverage of the Coupling Facility technology embedded in each layer of the stack, well-behaved applications are able to transparently inherit the scalability, availability, and throughput benefits of S/390 Parallel Sysplex technology. (See Angelo Corridori's article in the OS/390 HOT TOPICS Newsletter for additional information). On the other hand, applications which were written to cache data in processor memory themselves, or run in application execution environments which have not yet been modified to exploit the Parallel Sysplex technologies do not derive these benefits (i.e. , UNIX® applications but note that S/390 Parallel Sysplex technology does provide value as a DATA server for workloads such as SAP R3® in a 3-tier configuration and new with OS/390 Release 9 with Shared HFS).

New Application Environments

Another major Parallel Sysplex technology extension for client/server computing previewed (200-075) in 2000 with MQSeries® Shared Queues support via the Coupling Facility. With this capability, customer applications will be able to share MQ message queues concurrently across all systems in the Parallel Sysplex for scalable capacity and availability gains.

Probably the most significant OS/390 platform extensions for generating new business applications with S/390 qualities of service are contained in the Enterprise Java Beans and Component Broker Websphere™ support currently under development. Here, the rich OS/390 environment will be differentiated in the marketplace through transparent Parallel Sysplex enablement for portable applications

written using industry-standard, platform-independent application generation tools. Customer business application objects can be developed on one platform and then deployed on S/390 systems.

S/390 Resource Sharing

A new and significant value proposition for S/390 Parallel Sysplex has emerged as exploitation of the technology has evolved over time. A number of base OS/390 components have discovered that the Coupling Facility shared storage provides an excellent medium for sharing component state information for the purpose of multisystem resource management. Through CF exploitation, these system components have enabled multiple systems to share physical resources, such as files, tape drives, consoles, catalogs, etc. with significant improvements in cost, performance and simplified systems management. This is not to be confused with Parallel Sysplex data-sharing by the database subsystems, which requires significant application migration planning even when application re-engineering is not required. The following graphic will help you better understand the immediate benefits of this new value proposition called "S/390 Resource Sharing".

S/390 Resource Sharing Function/Value

XCF Star
Multisystem signaling
- ~ Simplified systems definition
- ~ Improved performance
- ~ Reduced cost
- ~ Channel constraint relief

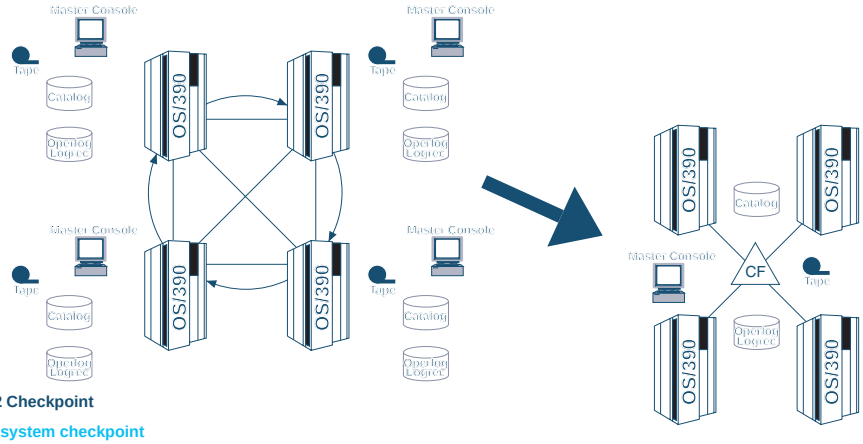
Shared Tape
Multisystem tape sharing
- ~ Eliminate duplication
- ~ Reduced cost

JES2 Checkpoint
Multisystem checkpoint
- ~ Simplified systems definition
- ~ Improved performance
- ~ Reduced cost
- ~ Channel constraint relief

RACF - Security Server
Multisystem shared security profiles
- ~ Improved manageability
- ~ Systems Management simplification

GRS Star
Multisystem resource serialization
- ~ Highly scalable
- ~ Rapid recovery
- ~ Improved performance

Operlog/Log Rec
Multisystem merged log
- ~ Improved Systems Image
- ~ Single System Image



High value and easy transition
No stand-alone CF requirement
Optimized with G5/G6 servers

Shared Catalog
Shared Master and User catalogs
- ~ Systems Management simplification
- ~ Improved performance

As you can see in the diagram, each OS/390 image or system requires basic hardware and software resources to function For example: tape drives, data sets, consoles, log data, catalogs, JES2 checkpoints - any of a number of system resources that run on an OS/390 system. But as your business grows, so does the number of OS/390 systems and the corresponding resources you need, not to mention the cost and complexity.

Whether you have these OS/390 images run on separate servers or a single server footprint with logical partitions in the Parallel Sysplex cluster, you get the benefits of resource sharing: simplified systems management, reduced complexity and cost, and increased performance .

Using the coupling facility, you can define structures that allow you to share information for these resources. So instead of having to define a tape device, for example, on every system, you can define a structure in the coupling facility that allows you to share the tape resource and thereby eliminate redundant tape drives.
For complete details, see our Web site: **www.ibm.com/s390/pso**

In conclusion, Parallel Sysplex clustering technology addresses many of your IT business requirements for today and the future as you deploy new and exciting workloads . The value of the technology is delivered as you enable it. You can start by enabling S/390 resource sharing : all you need for this is a G3 server or above and OS/390 Rel 2 or above. In fact, you can just use our Web- based tools to help you configure your environment and then automate your Parallel Sysplex environment with Systems Automation for OS/390. (see articles on these topics in this Supplement). Then move at your own pace to enable your applications to exploit datasharing as your need for increased availability and scalability grow. Keep in mind that IBM continues to build more and more functionality into the Parallel Sysplex and while you move with us, you will be better positioned to get the immediate benefits of all new functions as we deliver them .

For more on clustering, see Madeline Nick's article in the OS/390 HOT TOPICS Newsletter.

GRS: Star of the Enterprise

Scott Fagen

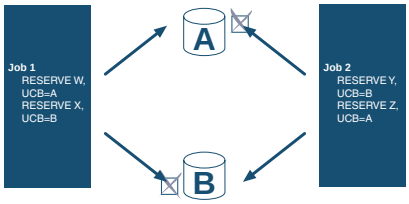
GRS: History

Global Resource Serialization (GRS) has been a key element of the MVS and OS/390 operating systems since MVS version 1. GRS was originally conceived as a mechanism to enable multiple MVS™ images, running on separate hardware, to be able to share DASD resources at an individual file level far more granular than an entire DASD device. Before GRS, IBM provided only a hardware mechanism, known as RESERVE, to enforce mutually exclusive access to shared DASD.

The RESERVE mechanism, although fast and efficient, would often cause deadlocks between jobs on multiple systems for the following reasons:

The unpredictability of where files would be placed on DASD: with the advent of SMS (Systems Managed Storage), and later, HSM (Hierarchical Storage Manager), the location of files is not under the control of the programmers submitting the workload. If jobs on different systems, requiring files on the same DASD, execute simultaneously, it is possible for a resource deadlock to occur.

As can be seen in the previous example, the two jobs are contending for different data sets (W, X and Y, Z) which happen to be allocated on the same volumes (A and B). Because the requests arrive in different orders, both of the jobs will be in a "deadly embrace," which will remain in effect until one of the jobs is canceled.



Shared DASD

The "unfairness" of the RESERVE implementation: The software implementation of RESERVE is such that, when a system obtains a RESERVE on a device, it will continue to hold that device until all requests for that device have drained off a system. The potential exists that workload on a system will always require some data set on that device, thus locking out any other system from access to that device.

To alleviate these problems, IBM introduced Global Resource Serialization in MVS/SP 1.3.6.

GRS Ring

Using GRS, the installation configured a group of MVS systems into a "GRS Complex." The systems were connected to each other via CTC connections and the GRS software arranged those systems into a token ring. The first GRS Complexes were generally 2-4 systems, although there are legendary installations that constructed rings of as many as 20.

The token that is passed between the systems is known as the RSA (Ring Systems Authority). While a system has the RSA, it is allowed to process serialization requests, known as enqueue (ENQ) and dequeue (DEQ). What the RSA contains is the set of changes to the resource allocations managed by GRS that have occurred on the previous loop through the GRS complex.

While GRS is waiting for the RSA to arrive, it accepts requests from the jobs running on that system and suspends the requesters.

When the RSA arrives, the system will:

1. Remove the last set of requests it placed on the RSA,
2. Process all the requests that were placed in the RSA from the other systems in the complex,
3. Add the new requests that arrived since the last time the RSA stopped on the system.

In this methodology, each system in the complex retains a complete view of the allocation of global resources in the complex.

Along comes Sysplex

In MVS/ESA™ SP4.1.0, the operating system was enhanced to join multiple systems into a Sysplex (Systems Complex). The Cross-system Coupling Facilities component (XCF) was added to the system to enable messaging between the systems and to manage the systems as they joined and left the topology. Other components, such as Console Services, were enhanced to simplify management of the sysplex environment.

With Sysplex, GRS is required to be activated. This fulfills the need that many MVS components and other program products have for a generalized distributed lock manager. Some of the frequent users of GRS include Console Services (Comm Task), Automatic Tape Sharing (IEFAUTOS), Dump Analysis and Elimination (DAE), UNIX System Services, DFSMS®, RACF®, and Workload Manager (WLM).

Prior to Sysplex, users of GRS were required to manage the orderly transition of systems in and out of the complex. If a system failed or a communications link failed, GRS required assistance from an operator to reactivate GRS. With sysplex technology, GRS is able to use the services provided by XCF to manage communications between the systems and the topology of the ring, greatly simplifying the configuration and operation of a GRS complex.

With sysplex, the following operations are automated:

1. Systems leaving the configuration:
Given enough lead time, GRS can cleanly remove a system from the ring without causing a ring disruption.

2. Automatic recovery:
If a system leaves the sysplex unexpectedly or a communications path fails, GRS will take a "ring disruption" to recover from the failure. Prior to sysplex, many of these actions required operator intervention to rebuild the ring. With sysplex, these operations are fully automatic; the GRS complex will recover without any operator intervention.

Additionally, since XCF manages inter system communication in a sysplex, the need to define CTC connectivity specifically for GRS was removed, simplifying installation and systems management. Another systems management and availability improvement with Sysplex is the ability to dynamically modify the Resource Name Lists with a single command. This enables an installation to modify how GRS handles certain ENQ and DEQ requests without having to restart the entire complex.

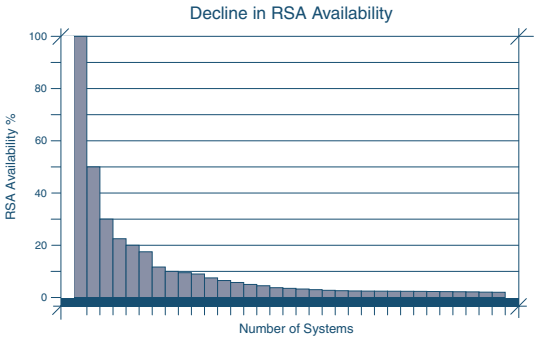
Parallel Sysplex: A Star is Born

In MVS/ESA SP5.1.0, a new technology, Parallel Sysplex, was introduced to allow better interaction between applications running on different images within a Sysplex. The Coupling Facility (CF), the hub of Parallel Sysplex technology, enables applications to communicate through the use of structures defined within a Coupling Facility image. To enhance availability and scalability, the limit of systems in a Parallel Sysplex was increased from 8 to 32. This change enabled customers to define a larger number of sharing systems to alleviate problems caused by system outages. If a customer took a S/390 CP and partitioned it (via PR/SM)™ into multiple images, the effects of a software outage to the user community was reduced, since the entire processing capacity would not be lost, only the failing partition.

What was perceived as a net gain in availability and scalability for a Parallel Sysplex does not translate well for a GRS Ring, for a number of reasons:

1st

First, the ability to have workload proceed is based on the availability of the RSA. As the number of systems in a sysplex is increased, the availability of the RSA is reduced by the number of systems in the complex. For example, if you have a two system sysplex, the opportunity for either system to have the RSA (ignoring the time the RSA is in transit between systems) is 50%. Increase the number to three systems and the availability is reduced to 33%, at eight systems, the availability is 12.5%, reducing all the way down to nearly 3% at 32 systems.



At the same time the availability of the RSA is decreasing, the amount of time between RSA arrivals increases. These two effects conspire to greatly reduce GRS availability, slowing down workload and increasing the opportunity for faults.

2nd

Second, since the Ring architecture requires that each system maintain the complete view of resource allocation, several other effects become quite noticeable in a large sysplex.

1. GRS address space storage utilization:

Each system in the sysplex has some number of requests for global resources, which are reflected in every system in the sysplex. As more systems are added to the sysplex, the overall amount of storage used by the GRS address space increases, generally in a linear fashion.

2. GRS address space CPU utilization:

As the number of systems in the sysplex increases, the overall number of ENQs and DEQs processed by the sysplex increases. Since all of the global ENQ and DEQ traffic in the sysplex is processed by each system in the sysplex, each request is actually processed one time for every system in the sysplex, resulting in linear growth in CPU utilization across the sysplex.

An especially harmful effect of this implementation occurs when systems of greatly varying capacity are in the same sysplex. Since larger systems have larger workload, implying greater numbers of GRS requests to be processed, the smaller systems can be overrun just by having to reprocess the requests locally.

3rd

Third, the results of the first two effects are magnified when systems join and/or leave the sysplex. When a system joins the sysplex, it needs to obtain the view of global resources in the sysplex. Since the size of this data is related to the number of the systems in the sysplex, the CPU cost (and delay time for the rest of the sysplex) grows as the number of systems in the sysplex increases.

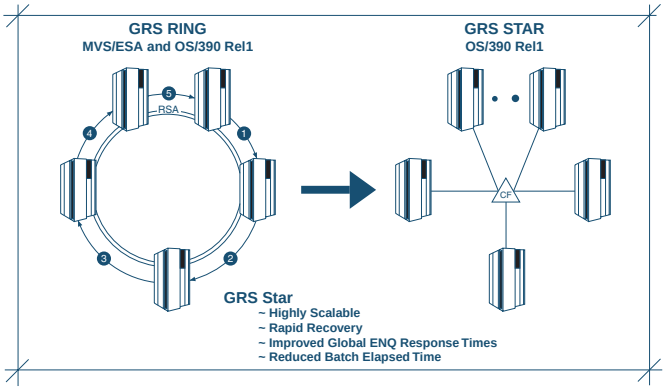
It became clear that the ring technology was not going to be able to handle the diversity in machine technology, sysplex size and LPAR weights. It was also clear that a radical change in the methodology used to implement GRS would need to occur. With that, IBM implemented GRS Star.

GRS Star: Removing the shackles on batch throughput

Like its predecessor, GRS Ring, GRS Star operates as a set of peers in a parallel sysplex. However, that is where the similarity ends. In solving the scalability problems of the GRS ring, GRS Star represents a complete departure from the ring philosophy by making each system as

independent as possible, while still cooperating to provide fair, first-in first-out, allocation for global resources across the sysplex. Programs that use GRS do not need to be modified to run in the Star environment.

GRS Star operates by placing global resource allocation information into a lock structure (called ISGLOCK) in the coupling facility. Contention for global resources is handled through the management provided by the XES locking function.



S/390 Resource Sharing - GRS Star

To quote one customer, "GRS=STAR was like winning the lotto for our users … we decided to move to GRS=STAR. This was accomplished over one weekend. I saw TSO logons go from minutes to seconds and a daily batch jobstream go from 14 hours to less than 2 hours..."*

That being said, what really are the advantages of GRS Star over GRS Ring?

1. ENQ/DEQ Response Time:

In GRS Star, ENQ response time (for resources that are not in contention) is improved in terms of *orders of magnitude* over GRS Ring. On the best S/390 hardware, coupling facility locking operations can complete in under 40 microseconds. This time does not degrade as the number of systems in the sysplex increases. Conversely, in GRS Ring, the best ENQ response times are measured in the 1-5 millisecond range, with the time increasing linearly as the number of systems in the sysplex increases.

This tremendous benefit is derived from the way global resources are processed in Star mode. Resource requests are reflected in the CF by way of optimized locking instructions, where, in the ring, there is latency waiting for the RSA to arrive and then for the other systems in the sysplex to unpack and process the request.

In fact, it is now possible with GRS Star to consider converting all RESERVE requests, including those for CATALOG. GRS Star ENQs are now processed faster than the RESERVE I/O.

2. Reduced CPU utilization:

In GRS Star, the CPU utilization of a particular image in the sysplex is directly proportional to the ENQ/DEQ workload on that system. If one system in the sysplex has a high number of GRS requests, that system will accrue higher GRS CPU cost. Conversely, if one system in the sysplex has little or no GRS activity, the cost in GRS CPU is minimal. Compare this to GRS ring, where *every system* in the sysplex will process every global resource request.

3. Reduced storage utilization:

Similar to CPU utilization, the storage utilization of a particular system in GRS Star is directly proportional to the ENQ/DEQ workload on that system. More ENQs means more storage. In a GRS ring, every system maintains the complete view of GRS resources, so storage used for global requests on every system represents the aggregate GRS workload.

4. Improved Startup and Shutdown Time:

Since the systems in a GRS Star do not maintain aggregate GRS resource allocation information, there is no cost associated with IPLing a new image into the sysplex. In GRS Ring, however, each system maintains that aggregate view, so during startup, that system must receive a copy of the information to join the ring. As the number of systems increases, the volume of data increases. As the volume of data increases, the time to transfer and build the local copy increases, increasing the amount of time it takes for each additional system to join the sysplex. Also, during the join process, new ENQ requests, from any system in the sysplex, are not honored until the join has completed.

Similarly, when a system leaves the sysplex, the members of a GRS Star do not have to update their view of the global resources, since that view only contains information local to that system. In a GRS Ring, all the systems must process DEQs for all resources held by the system that has left.

5. No Ring Disruptions:

In a GRS Ring, when a system unexpectedly fails or a communications link is lost, the continuity of transfer of the RSA breaks down, resulting in a GRS Ring Disruption. To recover, the remaining systems negotiate to build a new ring. This process is serial, that is only one system joins the new ring at a time. The more systems that are participating in the ring, the longer this process takes. Like the regular join process, no ENQ requests are honored until the ring has been rebuilt.

In a GRS Star, if a system fails, or loses connectivity to the CF, only that system and the global resources that it owns are affected. Workload in the rest of the sysplex, that does not have dependencies on those global resources, continues unabated. Recovering the ISGLOCK structure, via a structure rebuild, is a parallel process. Each system is able to restore the contents of the structure in parallel with the others.

GRS Star: The “No Brainer” choice

As you can see, GRS Star is the clear choice for your parallel sysplex. It is more available, it scales, it performs better and it uses less overall resources. Many installations are implementing parallel sysplex simply to enjoy the benefits of GRS Star. **Why don’t you?**

**Dean Tesar, Hewitt Associates*

System Logger 101

Doug Zobre and Mary Anne Hogarty

The MVS System Logger is an integral part of both the base sysplex and the Parallel Sysplex configurations. The MVS logger function allows its exploiters to merge log records between address spaces within an MVS image or on different MVS images within the same Parallel Sysplex. More and more subsystems and applications are using the MVS System Logger services to achieve their logging needs.

Who is Using System Logger?

Some of the current users are transaction and database services such as CICS and IMS, MVS OPERLOG (syslog) and Logrec, along with the Resource Recovery Services (RRS) and APPC/MVS Protected Conversations. The System Logger sysplex-wide log merge function allows VSAM record level data to be logged from any MVS image in the Parallel Sysplex. Even WebSphere application server uses System Logger for its logging needs.

How Does System Logger Work?

System Logger puts data in primary storage for the log writers, duplexes the primary storage, and when necessary, offloads (moves) primary storage data to secondary storage (hardened to DASD log data sets). The duplexing of the data prevents a single point of failure from inhibiting recovery of the log data. Log data in primary storage is vulnerable to loss due to system or sysplex failure because it has not yet been hardened to DASD log data sets. Logger recovery support is provided in the event of application, system, sysplex, or coupling facility structure failures.

In OS/390 Release 10, System Logger provides easier and faster access to log data.

How Does System Logger Maintain Log Data?

MVS System Logger provides two types of configurations for maintaining log data: coupling facility log streams and DASD-only log streams. The major differences between these two types of log stream configurations are the storage medium System Logger uses to hold interim log data and how many systems can use the log stream concurrently.

~ In a coupling facility log stream, interim storage for log data is in coupling facility list structures.

~ In a DASD-only log stream, interim storage for log data is contained in local storage buffers on the MVS system. Local storage buffers are data space areas associated with the System Logger address space, IXGLOGR.

Which Option Should I Choose?

The considerations for choosing either coupling facility structure based log streams or DASD-only based log streams are:

- The location and concurrent activity of writers and readers to a log stream's log data
- The volume (data flow) of log stream log data

CF structure based log streams are required when:

- There needs to be more than one concurrent log writer and/or log reader to the log stream from more than one system in the sysplex.
- There are high volumes of log data being recorded to the log stream. (Since DASD-only log streams always use staging data sets, high volume writers of log data may be throttled back by the I/O required to record each record sequentially to the log stream's staging data sets.

DASD-only log streams can be used when:

- There is not a need to have more than one concurrent log writer and/or log reader to the log stream from more than one system in the sysplex.
- There are low volumes of log data being recorded to the log stream.

Note: A DASD-only log stream is single system in scope. This means that even though there can be multiple connections to it from a single system in the sysplex, there cannot be multiple systems connected to the logstream at the same time.

Why Do I Need SMS?

System Logger requires that you have SMS installed and its address space active at your installation on each system where System Logger is expected to run. This is true even if you do not use SMS to **manage** your volumes and data sets. SMS must be active because System Logger uses VSAM linear data sets. If the SMS address space is not active when you attempt to use a System Logger application, System Logger issues system messages that indicate allocation errors and the application will not be able to use the logger function.

How Do I Set Up System Logger?

Your installation can use just coupling facility log streams, just DASD-only log streams, or a combination of both types of log streams. Most of the requirements and preparation steps for the two types of log streams are very similar, but coupling facility log streams and multi-system configurations require some additional steps.

COMMON SETUP ACTIVITIES	ADDITIONAL CF LOG STREAM STEPS
Sysplex Environment	
~ Need to be in a sysplex environment (PLEXCFG) ~ Set up sysplex, LOGR couple data sets and LOGR policy ~ Parmlib members - IEASYSxx, COUPLExx	~ Need to be in a Parallel Sysplex environment ~ Add coupling facility ~ Set up CFRM couple data sets and policy
Log stream Configuration	
Plan ~ Types of log streams ~ How many log streams ~ Log stream recovery	**Map log streams to CF structures** ~ Determine structure sizes ~ Plan structure placement ~ Plan peer connector recovery
Naming Conventions	
~ Log stream names ~ Log stream DASD data set names ~ Add log stream DASD data set names to GRSRNLxx inclusion list	~ CF structure names
DASD Space	
~ Allow for DASD expansion, DSEXTENTs ~ Plan use of retention period for DASD log stream data ~ Plan log stream staging data set size and access ~ Plan SMS data set characteristics for log stream data sets ~ If multisystem sysplex: Need shared catalog Access to DASD volumes Serialization mechanism SHAREOPTIONS(3,3)	~ Select duplexing method for log stream data
Security Authorization	
~ IXGLOGR address space ~ Administrator of LOGR couple data set ~ Exploiters of log streams	**CFRM couple data set and structures** ~ For IXGLOGR address space ~ Administrator of CFRM couple data set ~ Exploiters of log stream structures
Activate SMS	
Activate LOGR	
Establish SMF88 Reporting	

How Can I Find More Information about System Logger?

The primary source for System Logger information is *OS/390 MVS Setting Up a Sysplex*. There are also several web sites that will help you set up your System Logger configuration.

Parallel Sysplex Home Page: **http://www.s390.ibm.com/pso**

Configuration Assistant: **http://www.s390.ibm.com/pso/psotool/ps_intro.html**

CF Sizer: **http://www.s390.ibm.com/cfsizer/**

There's more to come, but we're loggin' off for now!

Maintenance in a S/390 Parallel Sysplex Environment

Barbara Bryant

IBM products developed for the S/390 Parallel Sysplex environment follow a comprehensive maintenance strategy designed to maximize availability and reduce problems caused by known software defects. An important goal of the maintenance strategy is to reduce unplanned outages to critical applications caused by product defects for which fixes are available. In a Parallel Sysplex environment, where critical applications can run on multiple systems using data sharing, planned outages for maintenance upgrades can also be reduced. Maintenance can be installed one system at a time, minimizing the impact to the applications.

Understanding the key elements of the S/390 maintenance and service process can help customers optimize their maintenance practices.

IBM provides both corrective and preventive maintenance for OS/390 based products. Corrective maintenance includes Program Temporary Fixes (PTFs) that can be ordered separately to resolve specific defects described by Authorized Program Analysis Reports (APARs). Customers can call the IBM Support Center or they can access the IBM databases electronically to search APARs and order PTFs. Preventive maintenance is available through several offerings which provide service packages on a regular basis. Currently, the most common preventive service packages are the Enhanced Service Offering (ESO) and the Customer Built Product Delivery Offering (CBPDO). However the latest, and greatest, service delivery tool is the S/390 Service Update Facility (SUF) which provides both corrective and preventive maintenance through the internet. Based on the customer's order, SUF interfaces with the customer's SMP/E (System Modification Program/Extended) database to build a bit map of the installed products and PTFs in order to build a customized package with only the requested and requisite maintenance. The SUF order is automatically received into the customer's SMP/E libraries. (Read more about SUF in the OS/390 HOT TOPICS Newsletter.)

Notification of High Impact Maintenance

There are 2 classes of high impact maintenance, HIPER (High Impact/ Pervasive) and PE (PTF in Error) APARs, which are specifically identified for customer awareness. It is recommended that HIPER and PE APARs be reviewed weekly and those affecting the customer

environment should be installed monthly, or as soon as possible.

HIPER is used to indicate that an APAR has a high impact on availability. The following symptom flags are used to identify the type of impact the HIPER problem may cause:

~ loss of data
~ system outage
~ loss of major function or subsystem
~ severe performance impact

Customers running in a Parallel Sysplex should pay particular attention to HIPER APARs which include the keyword SYSPLXDS. A PE APAR is used to identify an error caused by a PTF. The PE designation is intended to prevent a customer from unknowingly installing an error PTF. PE APARs that cause high impact symptoms are also marked HIPER.

SMP/E ++HOLD statements are generated for both HIPERs and PEs and distributed with preventive maintenance packages, such as ESO and CBPDO. There are several other mechanisms for a customer to learn about HIPER and PE maintenance:

~ Preventive Service Planning (PSP) buckets are available through the IBM Support Center or electronically via ServiceLink.
~ ALERT/ASAP, available through fee based offerings, notifies customers electronically through ServiceLink whenever a HIPER or PE is identified by the IBM Support Center.
~ Enhanced HOLDDATA is latest option available from the worldwide web at: **http://service.boulder.ibm.com/390holddata.html** and provides the easiest method of identifying HIPER and PE which should be reviewed for installation. OS/390 Enhanced HOLDDATA is an SMP/E installable file containing ++HOLD statements for HIPERs and PEs. After the file is received, an SMP/E REPORT will indicate the HIPERs and fixes for PEs not installed on the system. The report includes the HIPER symptom flags to easily identify the impact and special keyword that may be relevant, such as SYSPLXDS. Enhanced HOLDDATA on the web is updated daily and can be pulled as often as necessary. Weekly reviews of enhanced HOLDDATA is recommended.

What Preventive Maintenance Should be Installed?

Although IBM makes all PTFs available with preventive maintenance packages, it is strongly recommended that only the recommended preventive maintenance be installed for OS/390. The recommended maintenance is identified as RSU (Recommended Service Upgrade).

PTFs for low impact problems which are not likely to affect customers are not recommended. These PTFs can be installed as corrective service, if necessary. The main reason for RSU is to reduce the volume of maintenance for customers to install, thereby reducing the work involved and the risk of change associated with installing maintenance.

RSU maintenance is identified monthly on ESO and CBPDO preventive service deliverables and is the only preventive maintenance available via SUF. RSU maintenance is identified using SMP/E ++ASSIGN statement for easy installation.

The criteria for automatic inclusion of a PTF in the RSU is based on the APAR it resolves, as follows:

~ Severity 1 and 2 APARs
~ HIPERs
~ Special Attention APARs
~ PE APARs
~ Security/Integrity APARs

In addition to these criteria, if and when an APAR is identified as resolving three or more customer reported problems at the IBM Support Center, the PTF will be included in the current RSU.

RSU maintenance is installed and tested monthly for OS/390 in a production-like Parallel Sysplex environment. The testing begins the first Monday after the RSU is generally available via the preventive service deliverables.

For high availability, IBM suggests that RSU maintenance be installed every three months. HIPERs and PEs should be monitored weekly and installed monthly, or as soon as possible. between the preventive upgrades.

What are Special Attention APARs?

Special Attention APARs are low impact APARs that are recommended by IBM to install with preventive maintenance. PTFs for APARs marked Special Attention are included in the RSU. Special Attention APARs fall into one of the following categories:

~ Pervasive (but not High Impact) problem
~ New Function Support
~ Serviceability
~ Installability
~ XSYSTEM
~ Product Specific Keyword: _____

The XSYSTEM flag is used for cross system toleration and is intended to identify maintenance needed for toleration or coexistence of multiple product or maintenance levels in a sysplex environment. The product specific keyword SYSPLXDS is used to identify APARs specifically related to the Parallel Sysplex environment. Although these APARs are not high impact problems, the PTFs should be installed when convenient.

Maintenance Reduction

Along with the goal of reducing unplanned outages for known defects, the S/390 maintenance strategy focuses on reducing the amount of maintenance that customers need to manage and install.

RSU is one example of reduced maintenance. Installing only RSU maintenance reduces the volume of PTFs that need to be managed by 20% to 30%. Reducing the volume of maintenance reduces the work associated with research and installation and also reduces the risk of change being introduced to the system.

For low impact APARs, where no immediate PTFs are required, fixes are deferred to future releases. The rationale is that better testing can be done during the product test cycles than during the service process. If the customer agrees that the fix can wait, the APAR describing the problem will indicate that the fix will be provided in a future. These APARs are referred to as FIN (Fixed if Next release) APARs. If another customer experiences the same problem as described by a FIN APAR but requires an immediate fix, a new APAR is created and a PTF is provided.

It is important for all S/390 customers to have a maintenance plan that suits their particular environment. IBM recommends having a well-defined process that is followed regularly. While rigorous maintenance reviews and upgrades take time and resources, availability in a Parallel Sysplex environment depends on a proactive process. A proactive maintenance process reduces time spent analyzing problems that could have been avoided.

System Automation for OS/390 Version 2 Release 1

Ronald Northrup

System Automation for OS/390 Version 2 Release 1 has undergone significant architectural changes that logically remove the boundaries between systems within a Parallel Sysplex. At the heart of this re-architecture, is the newly patented manager/agent design which leverages the MQSeries transport mechanisms for intercommunication.

In version 1 of System Automation for OS/390, the automation function is completely self contained within the System Automation application that runs under a Netview address space on each system that is being automated. This is frequently referred to as the Automation Engine.

In version 2 of System Automation for OS/390, elements of the Automation Engine have been separated. Those that observe, react and do, remain within the Netview address space. This portion is known as the **Automation Agent** and must be present on every system to be automated. The coordinating, decision making and controlling elements have been grouped into a single new address space known as the **Automation Manager**. A primary Automation Manager needs to be active in the Parallel Sysplex and our recommendation is for at least one backup or secondary Automation Manager to be started as well. The primary Automation Manager is loaded with a model of all the automated resources defined across the entire Parallel Sysplex when it initializes. It then communicates with the automation agents on each system, receiving updates about the status of resources in its model and sending orders out to the agents as various conditions within the model are encountered. The emphasis has also switched from a purely command driven automation to a goal driven automation. Automation programmers define the default behavior of the systems and application components in terms of dependencies, triggering conditions and scheduled requests and the Automation

Manager will strive to keep systems in line with these goals.
The real beauty of this design is that multiple systems can be automated from a shared automation policy that is maintained in a central location. This further supports the concept of a single point of control and introduces a true single S/390 automation instance for the first time.

Complete Application Automation

The version 2 design is particularly powerful in a Parallel Sysplex; data sharing applications can be automated as a whole, without regard to the number of resources or where these resources are actually owned. Resources can have complex dependencies inside and outside of the application, can be started as required due to workload pressures and be moved to other systems as part of planned or unplanned interruptions. The SA OS/390 Automation Manager uses its awareness of status, dependencies, location of resources, prioritized operator requests and policy goals (specified by the automation administrator), to decide what resources need to be made available or unavailable as well as when and where. Once implemented, this goal_driven automation can dramatically simplify operations. Operational complexity is managed and reacted to in a faster, predetermined manner by automation logic. The operator merely indicates what is wanted and automation takes care of the dependencies and resolves affected or conflicting goals. Sysplex-wide automation can also remove the need to specify additional automation configurations for contingency purposes. By using cross-system dependencies in conjunction with server and system goals, the Automation Manager is able to decide the best alternative location for an application should the need arise.

Resource Grouping Reduces Complexity

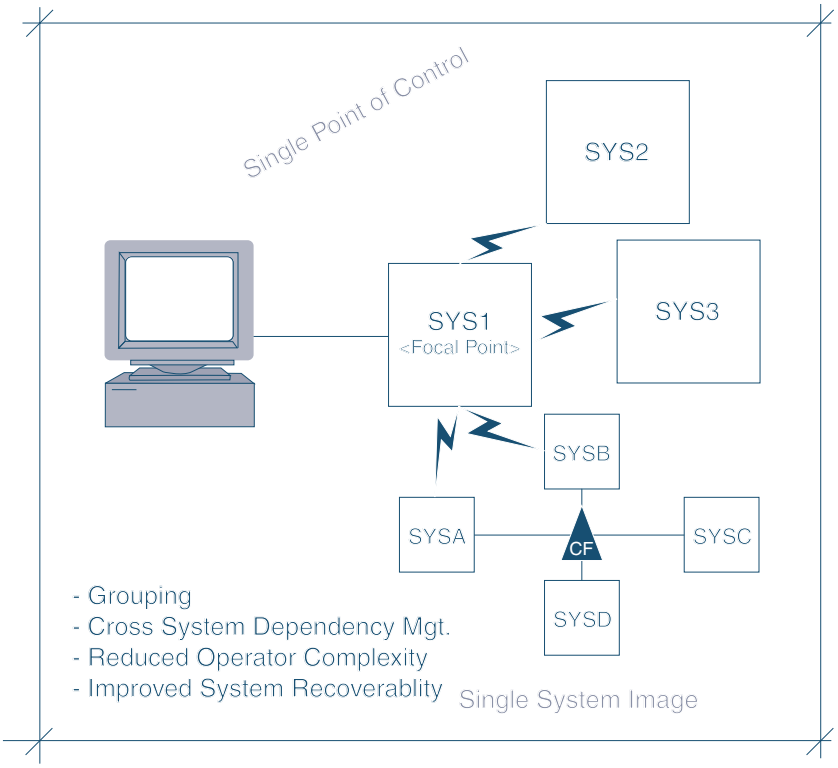
The ability to group resources and definitions by business application simplifies the automation definition and operational activities. Resources are any instance of independently automated or monitored entities that can be defined; applications and application groups all generate resources when they are associated with specific systems. This grouping removes the necessity for operators to understand details of the various components that make up the application. Grouping simplifies operations by showing the aggregated status and by using them for actions like startup or shutdown. A group is a collection of various resources which can be distributed across multiple systems within a Parallel Sysplex. A group can be a part of any dependency or other group. Resources can be a member of multiple groups and are referenced by a unique sysplex or system-wide name.

Parallel Sysplex Automation Routines

Customers have told us that they need more help handling day-to-day operational situations, particularly recovery events that are encountered less frequently and are considerably more complicated. Many of them write their own versions of automation exploiting code to achieve control over specific system events. For the less sophisticated

installation this is virtually impossible to accomplish. Our intention is to supply code to deal with specific OS/390 system related situations. Customers unable to develop their own code or who have lost key automation skills and are consequently running with unsupported code are particularly interested in this concept.

The objective is to reduce *operator complexity*, create *greater operational awareness* and improve *system recoverability*. Routines have been and are continuing to be developed that will assist operators in quiescing and removing active elements from the Parallel Sysplex and later on reintroducing them when scheduled work on them has been completed. An example of this is when a coupling facility needs to be upgraded; structures allocated in it need to first be moved to another coupling facility, prior to the hardware being taken offline. A panel known as the Parallel Sysplex Operations Center is started using the DISPPLEX command. Operators will find that many of the long, complicated commands necessary to display Parallel Sysplex resources can be issued by simply typing a character in the appropriate field on the panel. The results are returned in a scrollable format. This reduces the complexity and makes better use of the information that is returned.



System Automaion for OS/390 Version 2.1

Usability Improvements

Faster automation initialization time, solid manager/agent communication and the ability to dynamically change and activate automation policy, results in improved application availability. Individual monitoring granularity means that critical resources can be monitored more closely than those of lesser importance.

Operators and automation can choose or preset how an application will be started through the use of a startup type, such as, hot or cold or any other valid start option. Both the pre-start and post-start phases allow the preparation of an application startup; for example, by activating a network resource or by executing the necessary commands before or after the startup command is processed.

The capability of having multiple automation tables offers greater flexibility and will make it easier to separate IBM supplied rules from customer written rules, even if the same message is automated.

The customization dialogs have been enhanced to take advantage of many standard ISPF functions which result in faster navigation shortcuts and stacked commands. More detailed policy input fields mean that automation commands have been increased in length to allow specification of more complex commands thereby eliminating the need to write REXX CLISTs. Building the automation policy is significantly faster. Coexistence with System Automation for OS/390 V1.3 and migration functions allow for a smooth, efficient transition. The automation features for IMS, CICS and OPC are now integrated functions.

In Conclusion

These advances in OS/390 automation provide new and innovative ways of controlling operational tasks within the installation. This is particularly true for systems within a Parallel Sysplex; subsystems, applications and resources are all managed through a single administrative interface. This means less administration but most importantly that the applications and resources necessary for processing the business workload can be controlled across the entire Parallel Sysplex. This results in increased availability and reduced operational complexity.

Hot Enhancements for DB2 for OS/390 Data Sharing

Chris Munson

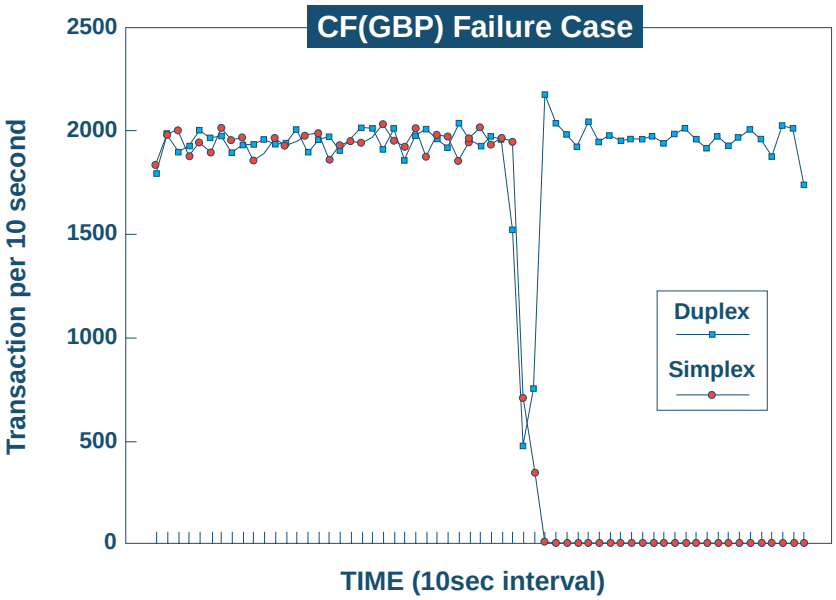
In 1995, DB2 introduced with Version 4 the capability of enabling a subsystem for data sharing. A data sharing system runs in a Parallel Sysplex and provides unparalleled (pun intended) scalability, availability and price/performance.

Since its release in Version 4, data sharing has gone through a number of enhancements, further making it the ideal solution for customers who require near continuous availability or who have outgrown the biggest baddest box on the market. This article discusses some of the recent and more exciting enhancements to DB2 data sharing and focuses upon Version 6 and the newly announced Version 7 product. Note that some of these enhancements are available in V6 base code and others via APARs in the service stream.

Version 6 Enhancements

By far one of the biggest V6 data sharing enhancements was duplexing of group buffer pools. In fact it was such an important enhancement that it was retrofit back into Version 5. The idea is to have 2 copies of the group buffer pool (GBP) structure, each in a different coupling facility. Changed pages are then written to both structures (overlapped for good performance). In the event of 100% loss of connectivity, a coupling facility or structure failure, DB2 switches to use the surviving copy of the group buffer pool without a big interruption to applications. Without duplexing the data in the group buffer pool would have to be recovered from the DB2 logs (done automatically with AUTOREC(YES) setting of the GBP) and the data would be unavailable during the recovery. The time required to recover from a simplexed GBP failure depends on several factors. Most notable are the size of the GBP and the amount of data that was in the GBP at the time of the failure. This time typically ranges from 10's of minutes to hours. Using GBP duplexing, this outage can be reduced to seconds.
The following chart depicts 2 workloads, one running with a duplexed

GBP and the other with a simplexed GBP with IBM's IRWW workload. Note that the transaction rate dips for both the simplexed and duplexed cases. The transaction rate is almost immediately returned to normal for the duplexed case, whereas in the simplexed case the transaction rate drops to nothing. This occurred because the transactions that were running were 100% dependent on GBP dependent objects (worst case scenario for a failure case such as this).



In addition to GBP duplexing, DB2 V6 also offers new GBP caching options for group buffer pools. First for the group buffer pool itself you can now specify GBPCACHE YES or NO. With GBPCACHE NO, no data is cached in the group buffer pool. Instead of writing changed data to the group buffer pool, it is written synchronously to disk. Its primary use is for applications that write large quantities of data to the group buffer pool with very little chance of re-reference. The default is GBPCACHE YES. Also offered is the ability to change the caching option at the page set/partition level. In addition to GBPCACHE CHANGED or ALL you can now specify NONE or SYSTEM. GBPCACHE NONE is similar to the GBPCACHE NO group buffer pool option, but for a specific page set/partition. GBPCACHE SYSTEM only caches system pages (i.e., space map pages). GBPCACHE SYSTEM is primarily to be used for LOB table spaces and is used to prevent large quantities of data (as is usually found with LOBs) from being written to the GBP.

Another option which has impact to the group buffer pools are new page sizes of 8K and 16K. These new page sizes can be used when 4K is not enough but 32K is too much. This can prove to be a data sharing benefit because data sharing costs include writing changed data to the coupling facility. The less data that is written the lower the data sharing cost for the application. For the case of a single row update (row size of 5000 for example), an 8K page size would write less data than a 32K page and therefore less data sharing overhead. There are 10 new virtual pools to manage 8K data (BK8K0 - BK8K9) and 10 for managing 16K data (BK16K0 - BK16K9).

DB2 also provides more efficient MVS message handling for CF or structure failures which will hopefully simplify the structure recovery operation.

DB2 Version 6 offers a new option on the STOP DB2 command to speed up shutdown. By specifying STO DB2,CASTOUT(NO), data in the group buffer pool is not castout to disk and remains in the GBP. Connections to the GBP for a CASTOUT(NO) stop of DB2 remain FAILED-PERSISTENT and when reconnected DB2 will find the data as it existed before the shutdown (assuming no other DB2 member performed the castout). This is an excellent availability feature for customers wishing to recycle DB2 to pick up new maintenance as it has been shown to have a 2X improvement in shutdown elapsed time in laboratory results (your mileage may vary!). Please note that since the latest version of changed data may still reside in the GBP and that using CASTOUT(NO) to establish a recovery point across disk volumes is not recommended. Also note that if the structure is forced or damaged when all DB2's are stopped, GBP recovery by the first DB2 that attempts to connect to the structure on a subsequent restart will occur.

Several performance related features were added to V6 that affect DB2 data sharing. The first one mentioned is space map update tracking, a DDL option available by specifying TRACKMOD YES or NO on the CREATE TABLESPACE statement. By specifying TRACKMOD NO, DB2 does not keep track of the 'dirty bit' in space map pages. The dirty bit is set when a page is changed and is used by incremental image copies to know which pages to copy. Since a page p-lock is used to serialize this space map update in a data sharing environment, heavy update to the same space map page by different members can cause a 'ping ponging' of the page p-lock from member to member. While TRACKMOD NO can improve data sharing update performance, it will degrade incremental image copy performance as a tablespace scan will now be needed.

Data sharing performance can also be improved by the V6 Fast log apply (FLA) feature. FLA uses up to 10MB to cache and sort log records in object order, then employs list prefetch and multiple read engines to read the pages needing modified by the soon to be applied log records. While not only a data sharing enhancement, DB2 data sharing can use FLA for GRECP/LPL recovery. Faster GRECP/LPL recovery means better availability of user data.

Another V6 performance feature that helps data sharing systems is called Identity Columns. This feature attempts to replace the 'ever increasing dumb number' keys used by customers today. The 'ever increasing dumb number' is a key value that is obtained by getting the current highest key value for a table, adding one to it and using that value for the next key. A one row table with the 'next available key value' or a SELECT MAX(key) FROM table can be used, both of which can cause a 'hot spot' for data access. This hot spot can degrade performance and availability in a data sharing environment. Using Identity Columns you specify during the CREATE TABLE that a numeric column of your choice is defined 'AS IDENTITY'. You can also specify starting value, increment and several other options of the new column. The column can be defined as GENERATED ALWAYS (DB2 always generates the next value) or GENERATED BY DEFAULT where DB2 will only generate the value if one is not provided. GENERATED BY DEFAULT can be used in the child table of a referential integrity table set. Identity Columns offer significant performance improvements over conventional user managed values as no locking is done by the user program to obtain/maintain the identity value. The IDENTITY_FUNCTION_VAL function can be used to retrieve the newly inserted identity column value.

So you don't have enough trace records you say?

Well, a last performance option added by V6 for data sharing is IFCID 329. This new trace record attempts to quantify the time spent waiting for asynchronous GBP requests. This can help explain 'NOT ACCOUNT' for time on a DB2PM or equivalent report.

Several usability enhancements are available in DB2 V6 for data sharing. It has been long recommended that BP0 be used only for the catalog and directory. Prior to V6, DB2 did not give a good way of enforcing this as objects automatically defaulted to BP0 when not specified on the CREATE DATABASE, TABLESPACE or INDEX

statements. In V6, two new system parameters (one for table spaces the other for indexes) were added to give the systems programmer the ability to assign a virtual pool when the buffer pool was omitted in the CREATE DDL statements. Additionally INDEXBP was added to the CREATE DATABASE statement to allow a user to specify a default virtual pool for indexes.

Other data sharing usability enhancements include adding a SCOPE(GROUP) keyword to the DISPLAY THREAD, DISPLAY PROCEDURE, and DISPLAY FUNCTION SPECIFIC commands. With this group scope designation, DB2 will issue the command on all active members, collect the results and display in one report. Group scope has also been added to the IFI interface enabling users to start and stop traces on all members with one IFI command.

A last version 6 usability enhancement to DB2 data sharing were changes to the DISPLAY BUFFERPOOL command output. As seen below in output from a DIS BPOOL(BP10) LIST command, a V6 data sharing system now reports information by partition (previously this was at the table space/index space level) which includes indication of group buffer pool dependency, members who have interest in the page set/partition, their p-lock state, use count and who the castout owner is.


```
Pre V6
DSNB450I : TABLESPACE = DSNDB04.STAFF, USE COUNT = 0, GBP-DEP = N
DSNB450I : TABLESPACE = DSN8D61A.DSN8S61R, USE COUNT = 0, GBP-DEP = N
DSNB451I : INDEXSPACE = DSNDB01.DSNLLX02, USE COUNT = 0, GBP-DEP = N

V6 (if data sharing)
DSNB460I @
- - - - - - - - PAGE SET/PARTITION LIST INFORMATION- - - - - - - -
- - - - - - - - - - - - - - - - - - - - DATA SHARING INFO- - - - - - - -
                              TS GBP MEMBER CASTOUT  USE   P-LOCK
DATABASE SPACE NAME PART IX DEP  NAME   OWNER  COUNT  STATE
======== ========== ==== == === ====== ======= =====  =====
DSNDB01  DSNLLX02        IX  Y   V61A     Y        0   IX
                                 V61B              0   IX
DSN8D61A DSN8S61E   001  TS  Y   V61A     Y        0   IX
                                 V61B              0   IX
                    002  TS  N   V61A              0   S
                                 V61B              0   S
```

In addition to the above format change, a parameter was added to the LSTATS keyword with values of * (all page sets/partitions are to be displayed) or ACTIVE (only active page sets/partitions displayed). The need for this keyword is driven by the fact that LIST and LSTATS output is now separated (previously interwoven in one report). Two new filters were also added to limit the report output. GBPDEP(Y) indicates that the report is to include only the group buffer pool dependent page sets/partitions as known by the DB2 member executing the command. CASTOWNR(Y) was also added to indicate that the report is to include only the page sets/partitions for which the DB2 member that is executing the command is the castout owner for.

Version 7 Enhancements

The following enhancements are not described in as much detail as the final preparations for the V7 code release is still underway at the time of writing this article. The enhancements can be broken down into several categories:

Availability

Continued restart performance enhancements will help make members of a data sharing group active sooner, thus improving availability of resources. New keyword LIGHT(YES) will soon be available on the STA DB2 command to request a 'restart light'. This is intended for a cross system restart (DB2 restarted on another OS/390 image due to a CEC or OS/390 failure) when insufficient resources are available on the new system to support a fully functional DB2. Using a restart light will provide quicker access to data as DB2 will restart with a small storage footprint far enough to free the retained locks that it can, after which terminating normally. Retained locks prevent access to data so the sooner these are released the better.

Also aiding availability is a no backout option on the CANCEL THREAD command. As is widely known, backout for a long running update UR can take longer than the initial update time. Using this no backout option DB2 terminates the thread and does not perform the backout to the beginning of the UR. Instead the object is put in 'refresh pending' and would require re-population via LOAD or INSERTs. This option would be used if, for example, it would take less time to refresh the data than to wait for the recovery to occur.

Usability

In the area of usability enhancements, DB2 V7 offers several new features. First group attach processing has been modified to include the following new features:

1. First, a *startecb* will now be honored for group attach processing and will be 'posted' by the first member to start in the group on the OS/390 image where the IDENTIFY request originated. A *startecb* is a parameter passed on IDENTIFY requests which ask that the requesting agent be suspended if a member of the group is not active on the OS/390 image and 'posted' (woke up) when a member is started.

2. A way of overriding group attach processing will be provided for CAF and RRSAF interfaces. Currently there is no way to avoid group attach processing. This is of use when the group attach name matches a DB2 member name and this name is used on an IDENTIFY request.

3. DL/I group attach support is now provided in non-usermod form.

4. Pre-conditioning has been added to support group attach for CICS and IMS. Each of these components will need to add code to complete the function.

Also available in V7 will be 'skip level' release coexistence/migration and fallback. You will now be able to migrate from V5 directly to V7 without first migrating to V6. Additionally you can fall back from V7 to V5 and can have coexistence between V5 and V7 members. No support will be available for V5, V6 AND V7 coexistence.

Another V7 change will be that DB2 structure size changes will remain persistent across the recycle of DB2 and the rebuilding of the structures. Prior to this enhancement a structure size changed via the SETXCF START,ALTER command may be 'forgotten' when DB2 is shut down or the structure is de-allocated or rebuilt.

A final option added in regards to structures is the exploitation of an OS/390 function called Name Class Queues. This function, when enabled automatically by DB2 with the correct DB2, MVS and CFCC software levels, will dramatically decrease the cost of finding and deleting group buffer pool pages (as done sometimes during the pseudo-close interval or DB2 shutdown). This decrease will help flatten CF utilization and can reduce the cost of DB2 shutdown. With Name Class Queues enabled, the CFCC organizes the directory entries in a more efficient manner to enable the finding and deleting of the GBP data more efficient.

As was mentioned in the V6 section, Identity Columns were added to enhance performance for the 'ever increasing dumb number' scenario. In V7 the Identity Column function was enhanced to 1) allow wrapping of the identity column value and 2) allow user defined min and max values. In V6 when the highest identity value was generated the next attempt to get the next value would result in a -359 SQLCODE indicating that all values have been exhausted. In V7 the value will wrap and 'start over'. The user defined min and max values give the user greater control over the range of values that can be assigned to the Identity Column.

That wraps up the V6 and V7 data sharing enhancements to date. Another hot item on the horizon that is worth mentioning is that DB2 is planning on exploiting a new OS/390 function (when available) called system managed duplexing for the lock and SCA structures. A duplexed lock or SCA works similarly to a duplexed GBP, except that MVS is managing the duplexing instead of DB2. Lock and SCA duplexing eliminates the concern of use of an Internal Coupling Facility (ICF) for these structures when a 'double failure' occurs. A double failure is, for example, when a CEC holding a ICF with a lock and or SCA and a member of a DB2 group who relies on that lock or SCA fails. In this example the entire DB2 group fails followed by a group restart on the next restart of a DB2 member.

As you can see IBM has made a significant investment into DB2 data sharing functionality for V6 and V7. Using data sharing and exploiting these new features customers can enjoy near continuous availability and superior scalability than in non-data sharing systems.

Meeting Expectations in a Business Intelligence Environment

Chris Panetta and Caryn Meyers

Business Intelligence workloads often start with small, homogeneous groups of users accessing localized data marts, and quickly grow to include diverse populations with varying requirements and needs. To satisfy these diverse groups, systems require the ability to differentiate between individuals in the population. Managing workflow in any system is crucial when you have a number of people sharing a centralized system to process work. Workload Manager is a component internal to the OS/390 operating system, that evolved to deliver a fully mature system resource management capability. No other platform has a heritage of being designed to handle multiple workloads concurrently, or offers as robust and efficient a method of controlling multiple units of work in a system.

In a Business Intelligence environment, Workload Manager (WLM) provides sophisticated industry-unique features for managing complex query workloads without the need to employ external work flow control mechanisms, such as governors, or requiring a separate server to act as a queuing manager. In OS/390, Workload Manager provides the crucial ability to:

~ manage resource consumption based on established business objectives
~ run diverse mixed query and other workloads concurrently
~ prevent monopolization of system resources by "killer" applications or queries
~ dynamically alter 'existing' work priorities as business needs demand
~ fully utilize all existing capacity
~ avoid the need to pre-identify short and long running queries
~ provide more consistent response times for end users
~ isolate high priority work to insure its timely completion.

Queries that originate from local users, as well as those that are submitted through websites via Net.Data, can all be managed to achieve the business goals in place for the user. This flexibility is allows diverse populations scattered in remote locations to access information on centrally located systems.

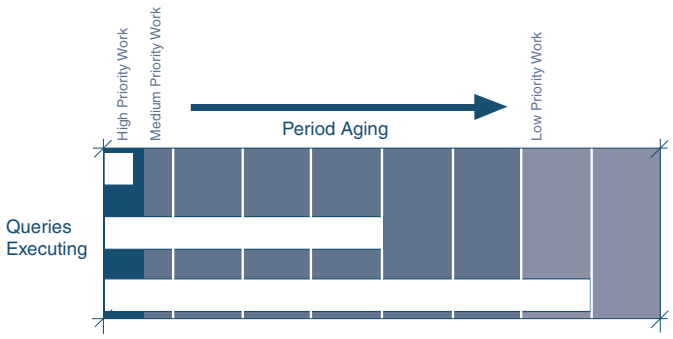
Managing Resource Consumption for Mixed Workloads

Managing resource allocation for diverse workloads in a key feature in the OS/390 environment. This unique feature called "Period Aging" allows OS/390's Workload Manager to allocate system resources based

on priority, where all work enters the system at equal priority. WLM dynamically lowers the priority of work that remains in the system for extended periods of time, thus reducing resource allocation to excessively long running work.

Over time, WLM will segregate short, medium, and long running queries into separate classes of work. This technique prevents bottlenecks by allowing longer queries to remain in the system, while still permitting additional smaller work to enter, complete, and exit the system. This natural prioiritization is accomplished dynamically by Workload Manager, without requiring the pre-identification of short and long running work PRIOR to entering the system.

WLM: Business Importance and Favoring Small Queries



All queries begin as high priority work, the longer they run, their priority is lowered relative to other queries in the system.

Favoring Selected Users within a Workload

In addition to period aging, WLM can differentiate between multiple groups of users, prioritizing their work at different levels. This allows work from critical users to execute at a high priotity than all other work in the system. Often there may be a group of end-users, such as executives or critical business users, who require a higher priority in the system, regardless of the size of the query. WLM offers the ability to identify critical work, and to deliver a consistent, reliable response time for that work.

Ease of Use

Establishing the controls for this environment is exceptionally easy. By stating the goals of the system in business terms, for example, "90% of the CEO class queries should complete in one minute or less", the system is able to interpret this information and manage the work flow to meet the stated objectives. Data is entered into the system through simple ISPF panels.

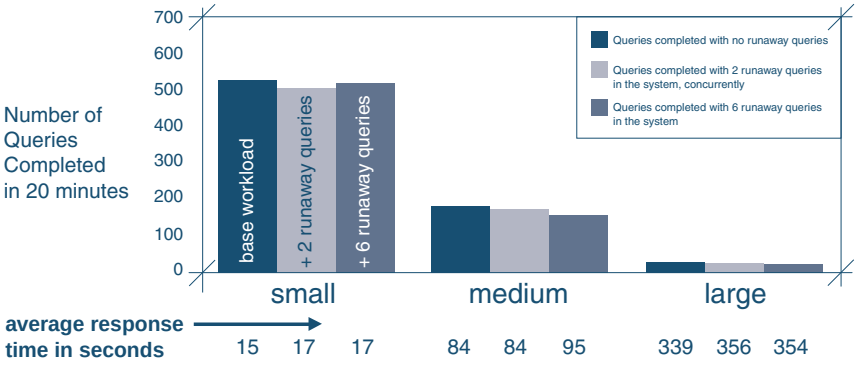
WLM also has the ability to define distinct reporting classes, used to monitor resource consumption and how well business objectives are being met for the separate workloads and users. This automated capability is unique in the industry, and is important when managing large diverse workloads on a single system.

Concurrency Capabilities

As important as single query speeds are, a more critical aspect of system performance is the server's flexibility dealing with high levels of query concurrency, fluctuating query volumes, and system effects at

saturation points. Today's parallel processing technologies place extremely heavy demands on all system resources. It is not uncommon that a single query, alone, could drive system resources to 100% utilization. Therefore, it becomes an extremely desirable attribute of any large scale data warehousing system to be able to deal effectively and efficiently with such issues without constant reconfiguration or external control mechanisms. OS/390's dynamic workload management technologies lead the industry in the ability to maintain high levels of concurrency and throughput in this demanding environment.

Query Throughput with Overload
30/40/30 Workload mix with 40 users
AND Runaway Queries



Results of 3 separate tests, each test reflected by 3 bars of the same color.
***Runaway queries cannot monopolize system resources.*

Conclusion

OS/390's capability to control query work from local and remote users, provides customers with a system that will deliver consistent, reliable query performance to business users, allowing IT to meet all service levels required by the business community. The OS/390 system is self-regulating, and can identify problem queries, fencing them to avoid impacting more critical work. Having a self-regulating system that can dynamically manage work flow in the server avoids severe management issues as the application moves into a production environment, and as a system is made accessible to increased groups of users. While other systems are best suited to dedicated departmental systems, S/390 offers the ability to scale to meet your companies needs today, and in to the future.

Read more about OS/390's Workload Manager and what consulants think about this leading technology in the data warehouse environment: "Business Intelligence Scalability or Real Workloads: The IBM System/390 Workload Manager", a white paper authored by Judith Davis and Richard Winter of Winter Corporation. (February 2000.) Ask your IBM sales representative, to print a copy of this white paper from the following IBM Internal website: w3.ibm.com/server/sales (select: Americas and s390; select: Solutions; select: BI; select White papers).

Geographically Dispersed Parallel Sysplex:

The Ultimate Application Availability Solution for any e-business

Noshir Dhondy

Introduction

How would a shutdown of your S/390 system affect your business? Do you put off system maintenance and upgrades to avoid system down-time? What about a site disaster? Are your business-critical applications and data protected from a site disaster? Table 1: Financial Impact of an Outage, lists the potential average hourly impact a business may incur due to an outage .

Type of Business	Average Hourly Impact
Retail Brokerage	$6,450,000
Credit Card Sales Authorization	$2,600,000
Home Shopping Channel	$113,750
Catalog Sales Centers	$90,000
Airline Reservations Centers	$89,500
Cellular Service Activation	$41,000
Package Shipping Service	$28,250
On-line Network Connect Fees	$25,250
ATM Service Fees	$14,500

Source: Contigency Planning Research (CPR) division of Eagle Rock Alliance, LTD (12/95)

Table 1:
Financial Impact of an Outage

IBM's S/390 multisite application availability solution, the **Geographically Dispersed Parallel Sysplex (GDPS)** is a multisite management facility that is a combination of system code and automation that utilizes the capabilities of Parallel Sysplex technology, storage subsystem mirroring and databases to manage processors, storage and network resources. It is designed to minimize and potentially eliminate the impact of any failure including disasters, or a planned site outage. It provides the ability to perform a controlled site switch for both planned and unplanned site outages, with no data loss, maintaining full data integrity across multiple volumes and storage subsystems and the ability to perform a normal Data Base Management System (DBMS) restart - not DBMS recovery - at the opposite site. GDPS is application independent and, therefore, covers the customer's complete application environment.

GDPS has been successfully installed at several customer locations. . They have experienced significant reductions to the recovery time window with no data loss when running Disaster Recovery (D/R) drills. For example, with GDPS, a simulated site disaster at a customer site caused no data loss and the recovery window was reduced from 12 hours to 22 minutes. Additionally, a user-defined planned site switch from one of the sites to the second site took 42 minutes. Reference 2 describes the GDPS experiences of several reference customers.

As illustrated in Figure 1 on page 36: Geographically Dispersed Parallel Sysplex, GDPS consists of a base or Parallel Sysplex cluster spread across two sites (known as site 1 and site 2 in this paper) separated by up to 40 kilometers (km) – approximately 25 miles – with one or more OS/390 systems at each site. The multisite Parallel Sysplex cluster must be configured with redundant hardware (for example, a Coupling Facility (CF) and a Sysplex Timer® in each site) and the cross-site connections should be redundant. All critical data resident on storage subsystem(s) in site 1 (the primary copy of data) is

mirrored to site 2 (the secondary copy of data) using the open, synchronous Peer to Peer Remote Copy (PPRC).

GDPS consists of production systems, standby systems and controlling systems. The **production** systems execute the mission critical workload. The **standby** systems normally run expendable work which will be displaced to provide processing resources when a production system or a site is unavailable. There must be sufficient processing resource capacity, such as processor capacity, main and expanded storage, and channel paths, available that can quickly be brought on-line to restart a system's or site's critical workload (typically by terminating one or more systems executing expendable (non-critical) work and acquiring its processing resource). A significant cost savings is provided by the S/390 9672 Capacity BackUp feature, which provides the ability to increment capacity temporarily, when capacity is lost elsewhere in the enterprise. The CBU function adds Central Processors (CPs) to a shared pool of processors and is activated only in an emergency. GDPS-CBU management automates the process of dynamically adding reserved Central Processors, thereby minimizing manual customer intervention and the potential for errors. The outage time for critical workloads can be reduced from hours to minutes. The **controlling** system coordinates GDPS processing. By convention all GDPS functions are initiated and coordinated by one controlling system.

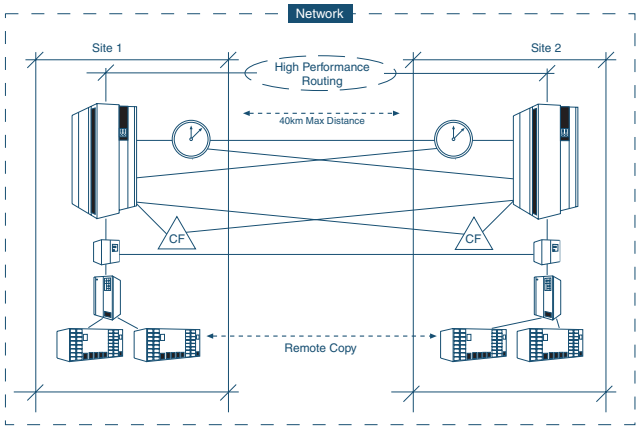


Figure 1: Geographically Dispersed Parallel Sysplex

All GDPS systems are running GDPS automation based upon Tivoli® NetView® for OS/390 and System Automation for OS/390. Each system will monitor the Parallel Sysplex cluster, Coupling Facilities, and disk storage subsystems and maintain GDPS status. GDPS automation can coexist with an enterprise's existing automation product.

Need for Data Consistency - the Freeze Function

Data consistency across all primary and secondary disk volumes spread across any number of storage subsystems is essential in maintaining data integrity and the ability to do a normal database restart in the event of a disaster. The main focus of GDPS automation is to make sure

that, whatever happens in site 1, the secondary copy of the data in site 2 is data consistent (the primary copy of data in site 1 will be data consistent for any site 2 failure). **Data consistent** means that, from an application's perspective, the secondary disks contain all updates until a specific point in time, without anything missing, and no updates beyond that specific point in time.

The fact that the secondary data image is data consistent means that applications can be restarted in the secondary location without having to go through a lengthy and time-consuming data recovery process. Since applications only need to be restarted, an installation can be up and running in less than an hour, even when the primary site has been rendered totally unusable. Data recovery involves restoring image copies and logs to disk and executing forward recovery utilities to apply updates to the image copies - a process measured in hours or days.

GDPS uses a combination of storage subsystem, Parallel Sysplex technology, and environmental triggers to capture, at the first indication of a potential disaster, a data consistent secondary site copy of the data, using the new, recently patented PPRC **freeze function**. The freeze function will freeze the image of the secondary data at the very first sign of a disaster, even before any database managers will be aware of I/O errors. This prevents the logical contamination of the secondary copy of data that would occur if any storage subsystem mirroring were to continue after a failure that prevents some but not all secondary volumes from being updated. This optimizes the secondary copy of data to perform *normal* restarts (instead of performing database manager recovery actions). This is the essential design element of GDPS in minimizing the time to recover the critical workload in the event of a disaster at the primary site.

GDPS Functions

GDPS provides the following functions:

~ PPRC configuration management —
automates the management of the remote copy infrastructure.

~ Planned reconfiguration support —
automates operational tasks from a single point of control.

~ Unplanned reconfiguration support —
automates recovery from an OS/390, software subsystem, processor, coupling facility, storage subsystem, or site failure.

PPRC Configuration Management

PPRC configuration management simplifies the storage administrator's remote copy management functions by managing the remote copy configuration rather than individual remote copy pairs. This includes the initialization and monitoring of the PPRC volume pairs based upon policy and performing routine operations on installed storage subsystems.

Planned Reconfigurations

GDPS planned reconfiguration support automates procedures performed by an operations center to simplify operations. These include standard actions to: (a) quiesce a system's workload and remove the system from the Parallel Sysplex cluster (e.g., stop the system prior to a change window); (b) IPL a system (e.g., start the system after a change window); and (c) quiesce a system's workload, remove the system from the Parallel Sysplex cluster, and re-IPL the system (e.g., recycle a system to pick up SW maintenance). The standard actions can be initiated against a single system or group of systems. Additionally, user defined actions are supported (e.g., planned site switch in which the workload is switched from processors in site 1 to processors in site 2).

Unplanned Reconfigurations

GDPS unplanned reconfiguration support not only automates procedures to handle site failures, but will also minimize the impact and potentially mask an OS/390, software subsystem, processor, coupling facility, or storage subsystem failure. Parallel Sysplex cluster functions along with automation are used to detect OS/390 system, processor, or site failures and to initiate recovery processing to help minimize the duration of the recovery window. If an OS/390 system fails, the failed system will automatically be removed from the Parallel Sysplex cluster, re-IPLed in place if possible, and the workload restarted. If a processor fails, the failed system(s) will be removed from the Parallel Sysplex cluster, re-IPLed on another processor, and the workload restarted.

If there is a site failure, GDPS provides the ability to perform a controlled site switch with no data loss, maintaining full data integrity across multiple volumes and storage subsystems. The next section describes this in more detail.

Unplanned Site Reconfiguration for a Multiple Site Workload

GDPS supports two configuration options:

~ Single Site Workload — the single site workload configuration is intended for those enterprises that have production in site 1 and expendable work (e.g., system test platform, application development, etc.) in site 2.

~ Multiple Site Workload — the multiple site workload configuration is intended for those enterprises that have both production and expendable work in site 1 and site 2. This configuration has the advantage of utilizing the resources available at the second site to provide workload balancing across sites for production work. GDPS provides the operational simplification to manage resources in a multiple site environment.

An unplanned site reconfiguration for a multiple site workload is discussed below. For a detailed description of planned and unplanned reconfigurations for both workloads, see Reference 3.

The multiple site workload configuration is shown in Figure 2: Multi-site Workload Configuration During Normal Processing. The production systems, SYSA, SYSB, SYSC in site 1, and SYSD, SYSE, SYSF in site 2 execute the mission critical workload. SYST, SYSU, and SYSV are standby systems that provide processing resources when a production system is unavailable or a site is unavailable. The 9672-X37 in site 2 has the CBU feature and can be expanded to a 9672-XZ7 during an emergency. The controlling system, 1K, coordinates GDPS processing. The primary copy of data (P) in site 1 is mirrored to the secondary copy of data (S) in site 2. The online transaction processing and data sharing related CF structures reside in CF1 in site 1 and other structures reside in the ICF in site 2.

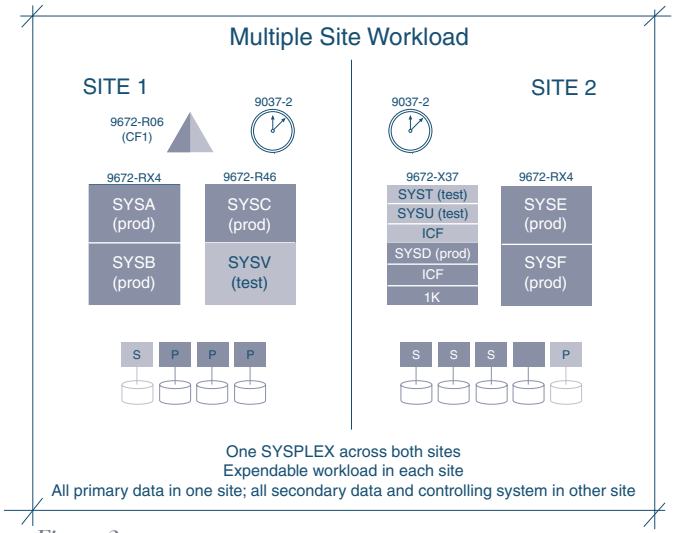


Figure 2:
Multi-site Workload Configuration During Normal Processing

When site 1 experiences a failure or disaster, referring to Figure 3: Multi-site Workload Configuration After a Site 1 Failure, GDPS will freeze the secondary copy of the data at the first indication of a problem to maintain data consistency. . When GDPS detects the last system in site 1 is no longer functional, it will initiate a site takeover. The secondary storage control units will be reconfigured to simplex mode; the CF and the Couple Data sets will be reconfigured; automatic CBU activation will expand the X37 to a XZ7 thus providing processing resources needed to execute the mission critical workload. SYSA, SYSB, SYSC, SYSD, SYSE,and SYSF will be re-IPLed, the CF structures will be rebuilt in the ICF, and finally, the mission critical workload will be restarted.

When site 2 experiences a failure, remote copy processing suspends and one of the systems in site 1 assumes the controlling system role while site 1 continues to execute the mission critical workload. If cloned, data sharing applications are active across all production systems, site 1 users will continue to execute with no impact and former site 2 users can re-logon onto site 1. Automatic CBU activation will expand the R46 to a RX6 thus providing processing resources needed to restart any mission critical site 2 workload in site 1. SYSD, SYSE, and SYSF will be re-IPLed in site 1, the CF structures will be rebuilt in CF1, and finally, the mission critical workload running on these systems will be restarted.

GDPS Service and Prerequisites

IBM has developed a set of tailored services for customers who have the requirements for evolution to a GDPS environment.
There are three GDPS services offered by IBM Global Services - **Remote Copy Management Facility (RCMF)**; **Parallel Sysplex Management Facility (PSMF)**; and **Geographically Dispersed Parallel Sysplex (GDPS)**. For a detailed description go to Reference 1.

For a list of hardware and software prerequisites required to implement GDPS, refer to the GDPS white papers (References 2 and 3).

Summary

In summary, GDPS provides not only all the resource sharing, workload balancing, and continuous availability benefits of a Parallel Sysplex, but it significantly enhances the capability of an enterprise to recover from disasters and other failures, as well as manages planned exception conditions. GDPS allows a business to achieve its own continuous availability and disaster recovery goals.

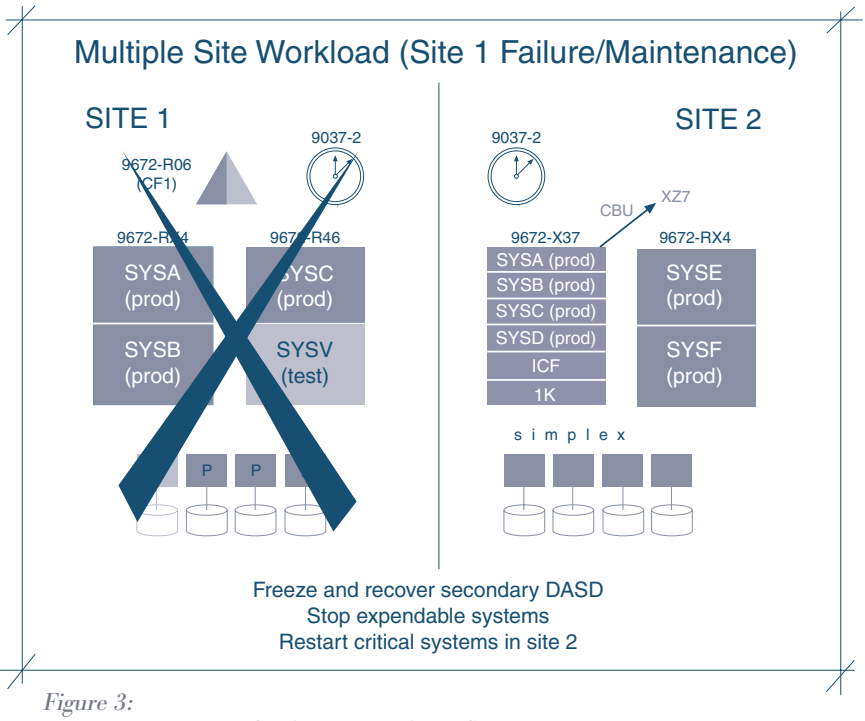


Figure 3:
Multi-site Workload Configuration After a Site 1 Failure

References

IBM Installation Services for GDPS http://www.as.ibm.com/asww/offerings/oww62b1.htm

GDPS white paper, Executive Summary http://www.s390.ibm.com/marketing/gf225114.html

GDPS white paper, detailed version http://www.s390.ibm.com/marketing/gf225063.html

Additional Information

Contact your IBM representative or e-mail gdps@us.ibm.com for additional information on GDPS.

The Ultimate Application Availability Solution in a Nutshell

Eugene Sale

With IBM's Parallel Sysplex Clustering technology enabling resource sharing and dynamic workload balancing, companies are now able to dynamically manage workloads across multiple sites to achieve high-levels of availability. Geographically Dispersed Parallel Sysplex (GDPS) compliments a multisite Parallel Sysplex by providing a single, automated solution to dynamically manage storage subsystem mirroring, processors, and networks, to allow a business to attain "near continuous availability", without data loss. It is designed to minimize and potentially eliminate the impact of any failure including disasters or planned outages. In addition, GDPS can exploit S/390's Capacity BackUp (CBU) feature to dynamically expand processors to handle the failed site's workload in the event of an emergency. Through proper planning and exploitation of IBM's Parallel Sysplex and Geographically Dispersed Parallel Sysplex technologies, companies are now able to fully protect their critical business applications from any unplanned or planned outage event.

S/390 Parallel Sysplex Performance:

Efficient, Scalable and Maybe Even Free!?

Gary King

The S/390 Parallel Sysplex provides a highly-efficient, highly-scalable cluster technology, and it may even come for free! When moving applications into a resource sharing and/or data sharing environment, can an increase in host capacity requirement be expected? The answer, of course, is "it depends!" It depends on: how much data sharing the application is doing, the type of host processor running the application, the number of systems in the sysplex, and the type of coupling technology that has been deployed. These topics will be discussed here, but for those that can't wait to get to the "bottom line," you can expect an application moving to full data sharing and running on the latest coupling technology to require around a 10% increase in capacity. As systems are added to the sysplex, this costs grows at only 0.25% to 0.5% per system added. This provides industry-leading scalability, particularly when contrasted to large Nway SMPs where the overhead often increases at 2% to 3% per added engine, and other clustered-system designs which can suffer 30% to 50% overheads for production applications.

"But wait," you ask, "if my application does incur a 10% overhead, how can that possibly come for free?"

As data sharing applications are rolled out, the transaction managers and workload manager can provide automatic balancing of the workload across the processors in the Parallel Sysplex. This allows all processors to run at high utilizations as workload is automatically shifted to wherever there is spare capacity. Without data sharing and workload management, workload balancing must be done "by hand" by physically moving workloads among processors usually resulting in some processors being underutilized for significant periods of time before the next physical reshuffling. After implementing data sharing and workload balancing, many customers have found more work flowing through the same processor configuration resulting in statements such as, "What overhead? We're running more work through the same hardware!" Thus, through better management of existing resources, the capacity cost of exploiting a Parallel Sysplex can actually come for free!

The efficiency of a Parallel Sysplex is most affected by the cost of an access to the coupling facility. To avoid the penalty associated with interrupts and task switches, most accesses to the coupling facility are

synchronous in nature, that is, the host processing engine which initiates the request will "dwell" (wait in a busy state) while the request is processed. The amount of time a processing engine spends dwelling for a coupling facility access represents lost capacity to that processor. Thus, the frequency of access to the coupling facility and the time required to process each request become the dominant factors in determining the capacity increase seen by applications running on a Sysplex.

The frequency of requests to a coupling facility varies with how an application exploits various sysplex functions A convenient measure of the intensity of coupling facility traffic is coupling facility operations per million instructions (CF ops/mi). This is simply calculated by summing the rates to all coupling facilities and dividing by the sum of the used mips across all processors in the sysplex. The exploitation of resource sharing functions generally results in relatively low rates - on the order of 1 or 2 CF ops/mi or less. For production applications that exploit data sharing, the rates are higher, typically falling in the range of 5 to 10 CF ops/mi.

The processing time of a coupling facility request consists of four major components: host software, host hardware, coupling link, and coupling facility hardware. The host software includes instructions spent in a subsystem (for example, DB2) and OS/390 to setup for and complete the request. The host hardware is involved in executing the specific coupling facility instruction and moving data in and out its link buffers. The coupling link moves the request to and from the coupling facility. The coupling facility hardware manages its link buffers and executes the instructions to process the specific operation (for example, a DB2 lock request). All components must be kept in balance to insure good performance. For example, as the speed of the host processor increases, the speed of the coupling facility and the coupling links should also be increased to keep pace.

Applications running on a Parallel Sysplex configured with well balanced coupling technology will experience an increase in capacity requirements of approximately 1% for each CF ops/mi (factoring in both software and hardware costs). Therefore, for these configurations, the net impact is generally 1% to 2% for resource sharing exploitation, and 5% to 10% for data sharing. However, if the coupling technology should lag behind the host processor technology, the capacity impact may grow to 1.5% to 2% per CF ops/mi. For these unbalanced configurations, the resource sharing cost could approach 4% while the data sharing cost could approach 20%. Thus, particularly for data

sharing exploitation, it is important to stay current with coupling technology. For example, a DB2 data sharing application running on a balanced Parallel Sysplex configuration consisting of G4 host processors, G4 coupling facilities, and fiber-optic based coupling links (hiperlinks) may experience a cost of 10%. Should the G4 host processors be upgraded to G6 processors but no change made to the coupling technology, the capacity impact will grow to 19%. Upgrading the G4 coupling facilities would lower the impact to 14%. Upgrading the coupling links from hiperlinks to integrated cluster buses would return the impact to 10%. As this example illustrates, IBM continues to evolve all components of the coupling technology in order to preserve the efficiency of a Parallel Sysplex. Other vendors may have a difficult time matching these advances.

There are cases where the capacity impact of data sharing appears to be higher than suggested above, however, these are usually just accounting artifacts. Most of the cost of the coupling facilities' accesses are charged to the data base manager's CPU time since most accesses are related to insuring the integrity of the data. The cost as seen from a before/after comparison of this metric may be 30%. However, in a typical system, only about 30% of the total CPU time is spent in the data base manager; the rest of the CPU time is spent in the application, in the transaction manager, in the monitors, etc., where there is a minimal data sharing effect. Thus, the net effect on system capacity is a 30% overhead on 30% of the CPU time and a near 0% overhead on the other 70% of CPU time - yielding a system-level effect of 9% which is consistent with the discussion above.

Lastly, how can you track the impacts, measure the effectiveness of workload balancing, and know when it's time to add or upgrade technology? Full monitoring and reporting functions are available through RMF and subsystem monitors. At minimal cost, these provide all the data necessary to maintain a well-running, well-balanced Parallel Sysplex.

Sysplex Failure Management

Riaz Ahmad

Customer experience has shown that the ability to detect and recover from component failures is critical for providing high availability in clustered systems like OS/390 Parallel Sysplex. Several availability studies have demonstrated that a percentage of Parallel Sysplex outages could have been avoided or mitigated if Sysplex Failure Management (SFM) were in effect This and the fact that SFM has been enhanced with APAR OW30926 , should be an impetus to consider (or reconsider) implementing SFM in your Parallel Sysplex.

SFM is integrated into the OS/390 base and it is recommended that all installations implement SFM when they are configured in either a basic or a parallel sysplex. To take advantage of the full range of failure management capabilities that SFM offers, a coupling facility must be configured in the sysplex.

SFM allows you to define a sysplex-wide policy that specifies the actions OS/390 is to take when certain failures occur in the sysplex.

XCF Signalling Connectivity Failures

All systems in the sysplex must have full signaling connectivity at all times. Loss of signaling connectivity between sysplex members can result in one or more systems being removed from the sysplex so that the remaining members in the sysplex retain full signaling connectivity to one another.

Status Update Missing Condition

Each system periodically updates its own status and monitors the status of other systems in the sysplex. The status of the systems is maintained in XCF Couple Data Set. A ***status update missing condition*** occurs when a system in the sysplex does not update its status information within the Failure Detection Interval, specified on the INTERVAL keyword in COUPLExx, and appears dormant. SFM allows you to specify how a system is to respond to this condition. System isolation allows a system to be removed from the sysplex as a result of status update missing condition, without operator intervention, while ensuring that the data integrity in the sysplex is preserved.

Additionally, with APAR OW30926, which is available for OS/390 Release 2 or higher, and SFM active with ISOLATE specified, the operator will be prompted if a *status update missing condition* occurs for a system, but XCF signalling is still active with that system. For instance, when temporary conditions occur in which the system status cannot be updated, messages IXC427A *and* IXC426D will now be issued if XCF signalling is active. This will eliminate the "false partitioning" due to

The type of failures that are handled by SFM are:

~ XCF signaling connectivity failures in the sysplex

~ Status update missing condition

~ PR/SM Reconfiguration

DASD on which the XCF couple datasets reside become inaccessible, but the systems in the Parallel Sysplex are otherwise healthy. The operator can then evaluate the state of the system targeted in the messages, and determine whether to remove the system from the Parallel Sysplex.

PR/SM Reconfiguration

After a system running in PR/SM partition is removed from the sysplex, SFM allows a remaining system in the sysplex to reconfigure processor central and expanded storage for use by the remaining system.

The Sysplex Failure Management Policy

For a sysplex to take advantage of SFM, the policy must be active in the sysplex. If any system loses access to the SFM couple data sets, the policy becomes inactive in the sysplex. If that system regains access to the SFM couple data set, SFM automatically becomes active again in the sysplex.

The ability of SFM to determine which system to remove from the sysplex in a connectivity loss situation is the fact that a ***weight*** has been assigned to each system according to it's importance in the sysplex.

SFM determines which systems to keep, which to remove from the sysplex and attempts to implement that decision by system isolation. If a system that is being removed is connected to a coupling facility that is also connected to an active system, SFM can reconfigure the sysplex by isolating and removing the system without operator intervention.

SFM and Structure Rebuilds

When there is a loss of connectivity between a coupling facility structure and its connector, it makes sense to have the structure that has suffered a connectivity loss rebuild in another available coupling facility thereby allowing the structure user to connect to the structure in the new location and continue processing. We must understand that not all structures support this rebuild processing, and it is also advantageous to allow an installation to control which structures will be rebuilt in a failure situation. The way an installation can influence which structures should be rebuilt is to use the REBUILDPERCENT keyword in the CFRM policy when defining the structures. The value specified for this keyword is used as part of a formula that also incorporates the system weights.

If the rebuild value calculated by the system is equal to or greater than the value specified by the installation on the REBUILDPERCENT keyword in the CFRM policy then and only then will the structure rebuild operation commence.

Some structures, such as XCF structures are so important to the integrity of the sysplex that they will always initiate a rebuild themselves as opposed to system initiating the action regardless of whether or not there is an SFM policy active. With OS/390 Release 3 and higher systems and fix for APAR OW30814 installed, the loss of CF connectivity processing has changed such that in the event there is no SFM policy active in the sysplex and there is a loss of CF connectivity, a structure rebuild will be initiated by OS/390 R3 and higher systems as long as the structure users support the structure rebuild.

Enhancements have been made to improve the CF selection process used during structure allocation by factoring in the SFM system weights. Also, structure rebuild processing was improved by insuring that the rebuilt structure has better connectivity to the systems in the sysplex than the old structure for a rebuild due to a loss of connectivity. Rebuilds started for other reasons, connectivity to the new structure must be equal to or better than the old structure. If no SFM policy is active, however, then all systems are treated as having equal weights when determining the suitability of a coupling facility for initial structure allocation.

Benefits

It is self evident that if you are running in parallel sysplex, you ***should*** have an SFM policy active and you should specify appropriate weights for your systems according to their importance in the sysplex. If there is a loss of signalling connectivity, Sysplex Failure Management will help you isolate the image(s) out of the sysplex according to your SFM policy and without the operator involvement. In the case where the status update missing condition has been detected, SFM will help isolate the image using the coupling facility Fencing Services without operator intervention. This is even more critical when you are in a data sharing environment. With the help of Automatic Restart Manager (ARM), restart the failing workload on other images within the sysplex for a swift recovery.

Having implemented Sysplex Failure Management will provide the ability to detect and recover from component failures in a timely fashion and therefore providing near continious availability of your parallel sysplex.

Easy Does It with These S/390 Parallel Sysplex Enhancements

Madeline Nick and Jim Guilianelli

Ok. You discovered that S/390 Parallel Sysplex Technology is no longer just for the "big guys", or maybe you realized that there's a gold mine inside the OS/390 base, functions like GRS star, Enhanced Catalog Sharing, XCF signalling, shared tape, and more, that you're just not utilizing.

So now you want to set up an S/390 Parallel Sysplex environment and exploit those new resource sharing functions you've been hearing about.

What do you do?

With two easy-to-use wizards on the web, - the Parallel Sysplex Configuration Assistant and the Coupling Facility Sizer - IBM can help you configure your S/390 Parallel Sysplex resource sharing environment. It's as simple as that.

BUT FIRST . . . MORE ABOUT IBM'S ON-LINE WIZARDS

As first announced in Release 8, the S/390 team has developed a set of internet-based wizards. The wizards provide an innovative approach for simplifying a set of OS/390 planning and configuration tasks, including setting up a Parallel Sysplex configuration. The initial wizards customize information based on your individual requirements. They can also simplify your planning and configuration needs by exploiting recommended values and providing checklists that reduce the number of steps and number of information sources you need to refer to. (For the URL paths to these wizards, see "Where on the Web" at the end of this article.)

The Parallel Sysplex Configuration Assistant Wizard

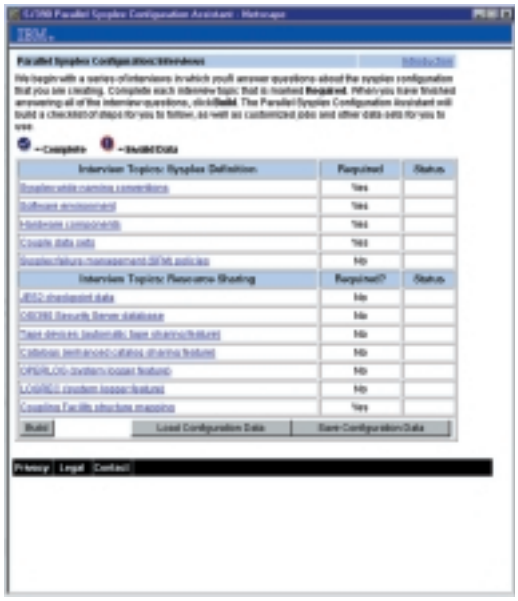
By using the **Parallel Sysplex Configuration Assistant wizard**, you can define the basic parmlib members, data sets, policies, and JCL needed for your Parallel Sysplex resource sharing environment.

Through a series of panels that present interview topics, this wizard allows you to define:

~ Naming conventions for the Parallel Sysplex:
sysplex name, couple data set names, and parmlib suffixes

~ Couple data sets, XCF signalling, and policies:
CFRM, SFM, ARM, WLM, and LOGR

~ Resource sharing for GRS, JES2 checkpoint, Security Server database, catalogs, tape devices, OPERLOG, and LOGREC



A wizard example

You select what you want and the tool guides you through the process, building the necessary jobs, JCL, and parmlib statements. When you're done, you can upload these definitions to your host system, or cut and paste what you need to the appropriate sources. Your configuration work sessions can be portable so you can save your data at any time to a disk location and re-load the data whenever you want and on whatever workstation has a connection to the web. This portability makes it easy to schedule your resource sharing configuration whenever it suits you, so you can define GRS star and basic setup to start, then return later to configure Logger or Enhanced Catalog Sharing (OS/390 Rel 7 and later).

For example, you've heard great things about GRS star but you haven't migrated to it. The Assistant lets you define structures for GRS star, then builds the necessary parmlib members, defines the policy statements in CFRM policy for the structure, sizes the structure for you, and creates the JCL for the job - all based on your answers to a few simple questions.

You can do the same for XCF signalling that provides the heartbeat for the Parallel Sysplex, coordinating timing events among the member systems, and for other kinds of resource sharing. Resource sharing in the Parallel Sysplex includes basic communication and management functions like sharing RACF databases, JES2 checkpoint data sets, sharing system logs, tape sharing, and enhanced catalog sharing as well as simplifying OPERLOG and LOGREC setup. Based on your input, the Assistant does the work for you - defining the coupling facility structures that make sharing possible, and providing the right fit by calculating the structure sizes that your configuration needs.

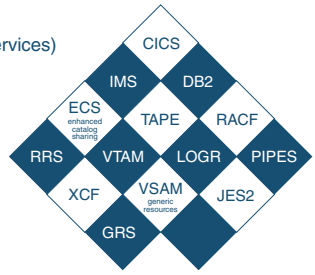
The great thing about the Assistant is that the parameter definitions are based on best practices of customers and IBM recommendations.

The Coupling Facility Structure Sizer

But if you're really interested in **sizing IBM structures** (for modification or planning purposes), the **Coupling Facility Structure Sizer** wizard is the way to go.

With a few input values that you provide, the Sizer takes the hassle out of sizing any IBM coupling facility structures.. Here's a list of the IBM environments or functions you can have structures sized for you:

ECS (enhanced catalog sharing)	VTAM
XCF signalling	BatchPipeplex
Tape sharing	RRS (resource recovery services)
JES2 checkpoint data set	DB2
Operlog	CICS
Logrec	VSAM RLS
RACF	CICS Logs
GRS	IMS



For example, say you want to define a structure size for Enhanced Catalog Sharing. Using 5 active catalogs (the default), the Sizer produces the following output:

ECS Structure Sizing Results

Structure and Sizing Results				
Function	Type	Name	INITSIZE	SIZE
ECS	CACHE	SYSIGGCAS_ECS	256K	256K

```
Sample CFRM policy statements

//STEP20 EXEC PGM=IXCMIAPU
//SYSPRINT DD SYSOUT=A
//SYSABEND DD SYSOUT=A
//SYSIN DD*
DATA TYPE(CFRM) REPORT(YES)
DEFINE POICY NAME(CFRMPOL1) REPLACE(YES)

CF NAME(CF01)
TYPE(009672)
MFG(IBM)
PLANT(02)
SEQUENCE(123456789012)
PARTITION(1)
CPCID(00)
SIDE(0)
DUMPSPACE(5 percent of total Coupling Facility Space)
```

Both the Parallel Sysplex Configuration Assistant and the Coupling Facility Structure Sizer wizards provide HELP along the way. The Assistant also lets you link to a task roadmap if you want to find out more about a product or function.

OS/390 Release 10 and Parallel Sysplex - More Good Stuff

Be sure to check out these OS/390 enhancements (Ease-of-Use: Simplifying Your IT Environment) in Release 10. They are making Parallel Sysplex and S/390 easier to implement and manage.

~ AutoAlter: adds support for *automatic tuning* of CF structure size and ratios of structure objects in response to changing structure object usage. Logger enhancements

~ SDSF Parallel Sysplex and wizard enhancements

~ System automation for OS/390 2.1

~ DFSMS enhancements - Transactional VSAM beta

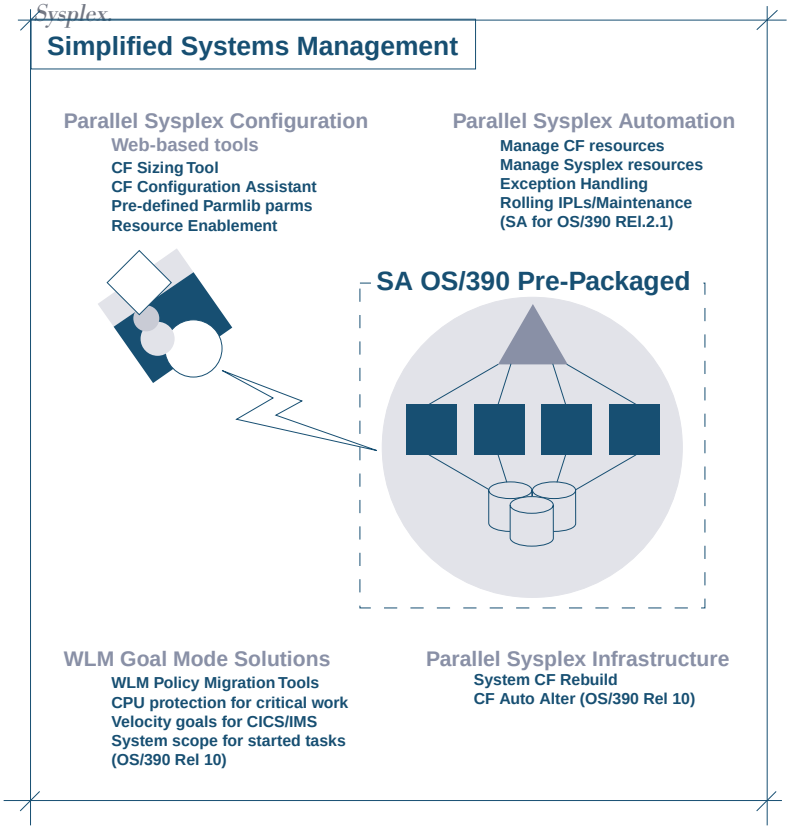
~ Enhanced Catalog Sharing - User rebuild support

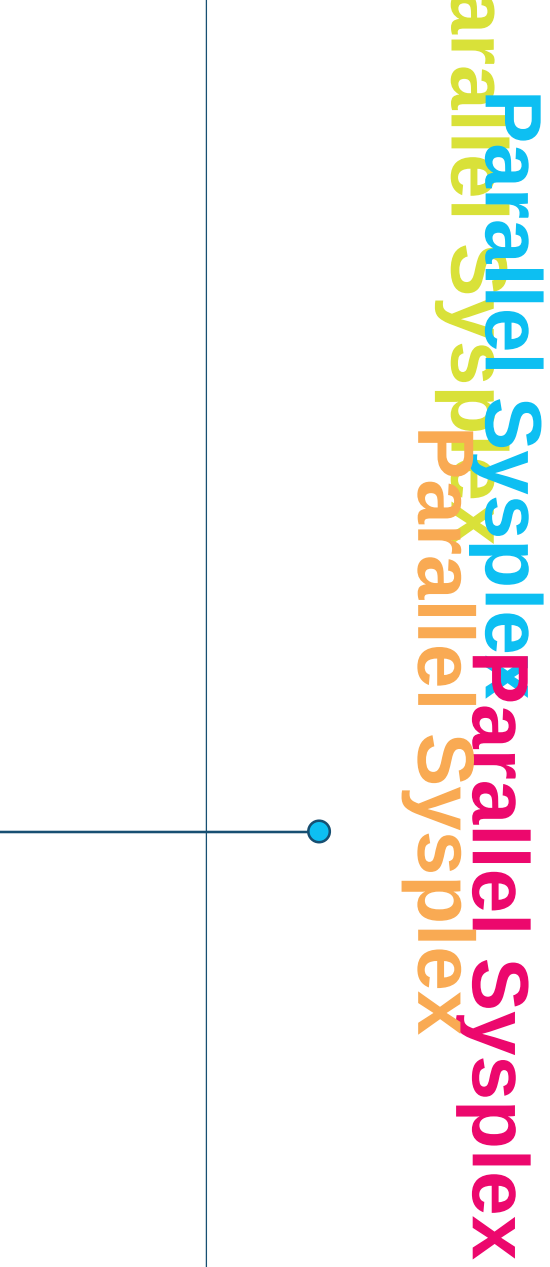
Where on the Web

To access the OS/390 wizards page, visit the following S/390 site:
http://www.s390.ibm.com/os390/bkserv/wizards

To access the Parallel Sysplex web page, visit the following site. (It also contains direct links to the Parallel Sysplex Configuration Assistant and the Coupling Facility Sizer wizards):
http://www.ibm.com.com/s390/pso

The following diagram summarizes usability enhancements to the Parallel Sysplex.





The End

Here's where you can find Hot Topics

- Hardcopy of the Newsletter is included in each "OS/390 Installation Planning Kit" that you get free of charge when a new release of OS/390 is made generally available (twice a year).
- We will continue to distribute hardcopy at user group meetings like SHARE.
- It is available from the OS/390 home page at: **http://www.s390.ibm.com/os390/**
- It is also available on each copy of *OS/390 PDF Library Collection* (SK2T-6718).
- It is also available to order hardcopy through your regular mechanisms at a nominal fee.

IBM

HOT TOPICS

An OS/390 Newsletter

for Release 10 - August 2000

Executive Editor
..... Bob Ward

Managing Editor
..... Dick Wagenaar

Art Director
..... Gene Posca

Graphic Designer and Illustrator
..... Tracy Bondi

© International Business Machines Corporation, 2000

Produced in the United States of America
8-2000
All Rights Reserved

The OS/390 Hot Topics Newsletter is published twice yearly as a service to OS/390 customers, providing articles of interest to the OS/390 community.

The information in this Newsletter is presented as is and is based on the best knowledge of the authors. No warranty is provided (neither expresssed nor implied).

IBM may not offer the products, features or services discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM representative for information on the products or services available in your area.

® IBM, DB2, CICS, DFSMS, IMS, MQSeries, NetView, OS/390, Parallel Sysplex, RACF, S/390, Software AG's ADAPLEX+, SysplexTimer, VTAM are registered trademarks of International Business Machines Corporation in the United States, other countries, or both.

™ GDPS, Globally Dispersed Parallel Sysplex, MVS, PR/SM, Websphere are trademarks of International Business Machines Corporation in the United States, other countries, or both.

Java, and all Java-based trademarks are trademarks of Sun Microsystems, Inc., in the United States, or other countries, or both.

Microsoft, Windows, Windows NT are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark in the United States and other countries licensed exclusively through The Open Group.

Other company, product, and service names may be trademarks or service marks of others.