RAS design for the IBM eServer z900

by L. C. Alves
M. L. Fair
P. J. Meaney
C. L. Chen
W. J. Clarke
G. C. Wellwood
N. E. Weber
I. N. Modi
B. K. Tolan
F. Freier

The IBM eServer zSeries™ Model 900, or z900, has been designed with major enhancements for hardware reliability, availability, and serviceability (RAS) in support of the zSeries RAS strategy, the eServer self-management technologies, and the z900 design objective of continuous reliable operation. The eServer self-management technologies enable the server to protect itself, to detect and recover from errors, to change and configure itself, and to optimize itself, in the presence of problems and changes, for maximum performance with minimum outside intervention. From the RAS perspective, the longstanding RAS strategy for the IBM S/390® and now the zSeries has provided an excellent foundation for self management. This paper describes the z900 RAS enhancements and how they strengthen the RAS strategy building blocks and provide a basis for autonomic computing.

Introduction

The purpose of the zSeries* RAS strategy is to enable delivery to our customers of servers which are capable of continuous reliable operation (CRO). The RAS strategy building blocks and the concept of CRO were described in Volume 43, Number 5/6 of the *IBM Journal of Research and Development* [1], which was devoted to the design of the S/390* G5/G6 servers. The two elements of CRO, *continuous* and *reliable*, require the server to run the customer's operation without interruption caused by errors, maintenance, or change in server hardware or Licensed Internal Code (LIC), while ensuring error-free execution and data integrity. The seven building blocks of this strategy, which are intended to support the drive to CRO, are error prevention, error detection, recovery, problem determination, service structure, change management, and measurement and analysis [1]. These RAS building blocks are intended to be independent of the operating system (OS), so that no particular OS is required to reconfigure the logical partition (LPAR), or to enable processor unit (PU), memory, or input/output (I/O) port sparing. No particular OS is required to enable Capacity Upgrade on Demand (CUoD) or Capacity Backup (CBU). No particular OS is required to provide service or remote support. All of these RAS functions are built into the structure of the hardware. Therefore, when a new OS (e.g., Linux**) is introduced, zSeries RAS is already at work. The RAS functions are operational whether Linux is the only OS running or Linux is sharing the server with a traditional OS (e.g., z/OS*).

Major enhancements in RAS design, concurrent upgrade, and concurrent repair for the z900 have been made in the processor, storage, I/O, power/cooling, service, support, and LIC subsystems, as well as for the Parallel Sysplex*. These enhancements strengthen the RAS building blocks and support the self-protecting, self-healing, self-configuring, and self-optimizing capabilities of the z900.

By definition, CRO implies the capability of a server to prevent or tolerate errors, eliminate outage, ensure error-

0018-8646/02/$5.00 © 2002 IBM

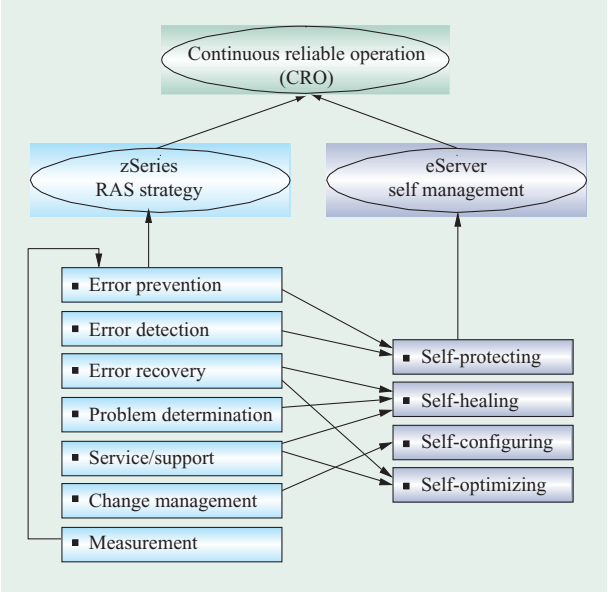

Figure 1

zSeries RAS strategy building blocks and eServer self management.

free operation, and enable the maximum possible operating capacity in the presence of errors, service, or change. This is also the essence of self management. The interrelationship between the RAS strategy building blocks and eServer self management is shown in **Figure 1**.

Processor subsystem

The z900 is based on the proven G5/G6 dual cluster system structure, described in [2], consisting of PU, storage controller control (SCC), storage controller data (SCD), memory bus adapter (MBA), memory system controller (MSC), external timer reference (ETR), clock (CLK), cryptographic coprocessor element (CCE), oscillator (OSC), and memory cards (**Figure 2**). The z900 utilizes two multiple-chip module (MCM) designs; one contains 35 chips (20 PU + 2 SCC + 8 SCD + 4 MBA + 1 CLK), and the other contains 23 chips (12 PU + 2 SCC + 4 SCD + 4 MBA + 1 CLK). Both MCMs are cooled by modular cooling units (MCUs) and operate at a nominal junction temperature of 0°C, improving server performance and semiconductor reliability.

The z900 processor subsystem continues the robust RAS design of the G5/G6. The PUs are not shipped with a fixed function assignment, but are assigned during first power-on reset (FPOR) following server power-on. The first assignments are for the system assist processors (SAPs), followed by the central processors (CPs), the Integrated Coupling Facilities (ICFs), and the Integrated Facilities for Linux (IFLs). Making these assignments dynamically allows the power-on reset (POR) code to respond to changes in the configuration due to upgrades or failures. The PUs are assigned alternately between the two clusters of the Level 2 (L2) cache (SCC and SCD). This allows all of the PUs to survive an SCD failure and half to survive an SCC failure.

The increase in the maximum number of PUs to twenty allows for more nontraditional processors, such as ICFs and IFLs, while keeping some PUs available as spares. The PUs are tightly coupled through a binodal L2 cache. The PUs can survive intermittent failures on the PU–L2 interface, or the z900 can completely fence (logically remove from the active configuration) a PU. Fencing a failing PU shifts the execution to a spare PU. If the failing PU is a "master" SAP and a spare PU is not available, an active CP is reassigned as the "master" SAP. This dynamic PU sparing process is transparent to the OS. Each PU protects itself with dual instruction/execution engines. The two engines execute each instruction in lockstep. The output of the engines is compared at checkpoints, and any mismatch causes the PU to retry from the previous checkpoint. Continued failure results in PU clockstop and a dynamic PU sparing, as described in [1]. The on-chip Level 1 (L1) cache can survive an array failure by purging and deleting the cache line or compartment. The L1 cache "fuse" relocation technology allows the defective cache line to be relocated (L1 cache-line sparing) at the next FPOR. All PUs, SCCs, and SCDs participate in the logic built-in self-test (LBIST) and the array built-in self-test (ABIST).

The z900 provides RAS enhancements in MCM reliability and in CCE serviceability, and introduces a redundant design for the system oscillator.

MCM reliability

To enhance error prevention, the z900 introduces a number of MCM stress test enhancements, increasing the already phenomenal MCM reliability even further. The improvements are as follows:

1. A −20°C temperature test to screen failures at low temperature. This test reduces or eliminates failures in manufacturing (shipped product quality level—SPQL) and in the field.
2. A port test to reduce defective single-cell array failures. Under this test, V_{dd} (drain voltage) is reduced to 0.7 V for 80 ms. Under these conditions, the defective single-cell failures are screened out.
3. A 1.2-V V_{dd} low-voltage test to screen cell failures which occur under low-voltage conditions.
4. Respective increases in ac and dc test coverage to 98.77% and 99.86% to improve SPQL fallout in

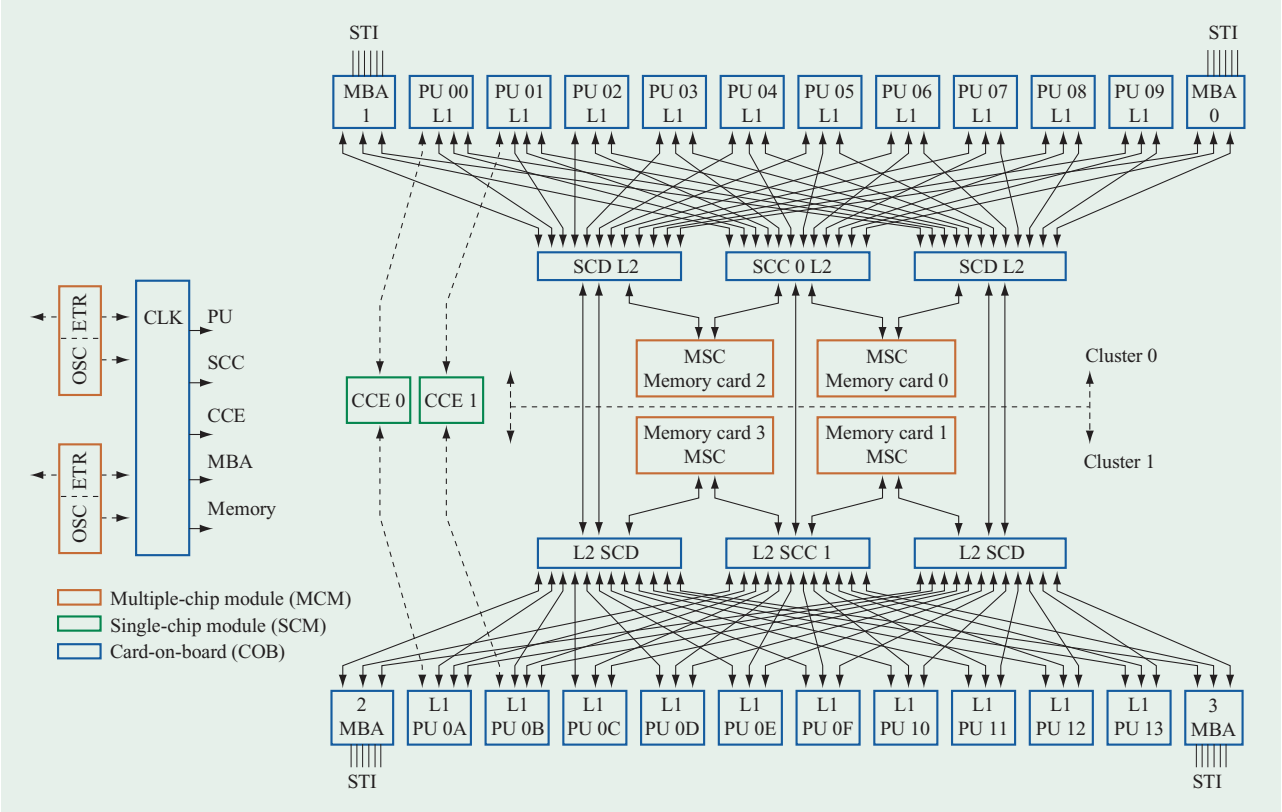

Figure 2

z900 system structure (20-PU MCM).

manufacturing and field reliability. Analysis techniques and TestBench[1] procedures are also improved.

5. Reduction of the nominal junction temperature of z/900 MCM chips to 0°C from 15°C for G6 to improve server cycle time. A side benefit of this change is a significant improvement in chip reliability due to lower junction temperature.
6. Addition of three hours of on-product clock generator post-burn-in stress testing for the PU, SCC, and SCD chips. This helps reduce the number of ac defects that escape to system manufacturing from the supplier.

Cryptographic coprocessor element

The CCE chips inherit their design from the G5/G6 servers, thus continuing the strong cryptographic RAS characteristics. The CCE chips are each "twin-tailed," with one active interface and one "hot" standby interface to the processor subsystem (Figure 2). CCE0 has interfaces to physical PU 00 and PU 0A chips; CCE1 has interfaces to PU 01 and PU 0B. PUs 00 and 01 are attached to the storage subsystem through L2 cache chip SCC0, while PUs 0A and 0B are attached through L2 cache chip SCC1. Attaching to different physical PU chips which are themselves attached to different L2 chips allows the CCE chips to be accessed after the loss of either a PU chip or an L2 chip. Each CP is mapped to a physical PU during FPOR. This mapping is biased in favor of those PUs with CCE chips attached. The first two CPs mapped, CP 00 and CP 01, are mapped to those PUs with CCE chips attached; conversely, the other two PUs with CCE chips attached are mapped last (i.e., they remain available as spare PUs, if possible). In addition to the mapping, the PU sparing algorithms are similarly biased in favor of the CCE chips.

The CCEs have error checking on all data paths and on all arrays, including the arrays containing keys. Sequence, invalid states, or duplicated logic is used on all logic state machines. The Data Encryption Standard (DES) engines are duplicated. Adders and arithmetic logic units (ALUs) have parity-predicting logic and carry checking. The modular exponentiation engines use residue checking, including the arrays containing keys. The CCEs are

[1] TestBench is a UNIX**-based set of test design automation tools developed by IBM for internal use. It was made commercially available in 1994 by the IBM Microelectronics Division.

included in the LBIST and ABIST routines, which are run during FPOR.

The z900 enhances the availability of the CCEs by repackaging them as separate single-chip modules (SCMs). Because the CCEs function independently, the failure of one CCE does not affect the other. Therefore, when one CCE fails, the other continues to function and maintain its encryption keys. The failing CCE is replaced during a scheduled service action. When the server is brought back online, the nonfailing CCE is available immediately. The new replacement CCE SCM is then loaded with the keys and begins processing.

Redundant oscillators

The availability of the z900 is enhanced by providing two independent system oscillators, which are housed on the OSC/ETR cards. One of these oscillators is active, to supply the clock (CLK) chip with the oscillator frequency. The other oscillator is the standby. The status of each oscillator is maintained in a register in the CLK.

One oscillator is selected as active during server power-on and during POR by the self-test function. If both oscillators are functional, the hardware selects oscillator 0 as active and oscillator 1 as standby. If oscillator 0 is functional and oscillator 1 is defective, the hardware selects oscillator 0 as active. If oscillator 0 is found to be defective, oscillator 1 is selected. In all cases, the status register is updated to reflect the status of the two oscillators. The defective oscillator is fenced to avoid being selected at the next server power-on and during POR. A failure of the active oscillator during run-time results in a system checkstop, requiring a server power off/on sequence to select the standby oscillator. A failure in the standby oscillator requires no recovery action. The failure analysis routines analyze the information in the status register and isolate the failure to the defective field-replaceable unit (FRU). A system reference code (SRC) is created and a call-home is generated, reporting the problem back to IBM.

Concurrent upgrade

The z900 continues the evolution of concurrent PU upgrades, utilizing the Licensed Internal Code configuration control (LICCC) technology introduced in G3 [3]. Capacity Upgrade on Demand (CUoD) improves availability by allowing the customer to obtain permanent additional capacity (CP, ICF, IFL) without taking the server out of production. Capacity Backup (CBU) improves availability by allowing the customer to rapidly obtain additional capacity for a defined period of time, to compensate for a disaster that disrupts a portion of his computing power.

The z900 introduces concurrent CBU downgrade. CBU increases the number of CPs available to the OS. The OS and applications determine the new configuration using the *store system information* instruction (or, alternatively, applications may use OS services). The customer uses the temporary increase in capacity for either a test or a real emergency. CBU downgrade is the process by which the server is restored to its original configuration following the emergency period. The customer selects "undo temporary upgrade" from the "perform model conversion" panel on the Hardware Management Console (HMC). The server is then restored to its original model. For G5/G6, this action was disruptive.

Concurrent repair

The z900 supports dynamic replacement of a defective PU (provided that a spare PU is available), fully transparent to the customer's operation. With the introduction of the redundant oscillator design, a defective oscillator card is repaired concurrently.

Storage subsystem

The z900 storage (cache and memory) design is an extension of the storage design used for G5/G6 servers with significant new RAS features.

Cache

As with G5/G6, the Level 2 (L2) cache and directory both have defective cache-line relocation capability. The L2 cache and directory both have error-correction code (ECC) protection. The cache uses a (72, 64) ECC, while the directory has two fields, the address field with a (25, 19) ECC, and the ownership field with an (11, 5) ECC. This provides for single-error correction and double-error detection. The L2 cache uses its directory ownership field to delete any combination of individual lines (in contrast to the L1 cache, which has compartment delete and up to one line delete controlled by registers).

When a PU fetches a double line of 256 data bytes from the L2 cache, the ECC logic corrects correctable errors (CEs) and returns the data to the PU. If the data has an uncorrectable error (UE), the PU is notified that the data is bad by a blocked "data valid" indicator. For either CEs or UEs, the failing location in the cache is trapped in a register which is later logged out for analysis. Also, the hardware initiates an L2 cache-line purge, which forces the failing line out of the cache. If the line was read-only or unchanged from main storage (L3), it is simply invalidated. If the line was changed, it is sent back to L3 and the data is either corrected (for CE cases) or marked as permanently bad (for UE cases). When the PU refetches the line, the data either will be clean (in the case of invalidated lines) or will still contain the UEs placed in L3. UEs cause the hardware to notify the OS that a failing storage address has occurred. Some OSs use

the failing address to take the failing page offline and restart the operations that used this page.

When a CE is detected, the trapped information is compared to information previously trapped for that given area of the cache. If the new failure is in the exact same bit as the previous failure, the line is not only purged but is also deleted from the cache so that it will not be used again. G5/G6 purged on the first occurrence of the UE and purged/deleted on the second. The z900, unlike its predecessors, purges/deletes on the first UE in order to minimize the propagation of UEs caused by L2 cache errors; UEs from memory or other sources are never allowed to enter the cache.

The z900 enhances its use of the L2 cache fuse information. When an L2 cache-line delete occurs, the fuse information for the deleted line is used to schedule a repair of the array for the next POR. When the next POR occurs, the fuse is applied instead of the delete. Therefore, ABIST runs successfully, using the fuse repairs, thereby verifying the array again prior to its use.

Memory

The memory (L3) consists of up to four cards per server. Each card has a memory controller. The memory card contains up to eight rows of 144 synchronous DRAM chips. Data is stored into one row at a time, two bits per chip, and is organized as two 144-bit data words. To protect the data, z900 uses a (140, 128) ECC with 128 data bits and 12 check bits. The code corrects any single-bit failure as well as any single-symbol failure (i.e., 2-bit failure within the same chip). Therefore, if a DRAM is completely broken and the bits coming from that chip are unpredictable, the hardware is able to correct the bits and calculate the proper data without replacing the chip. If two of the 72 DRAMs in the same row/same data word are broken, the ECC logic is able to detect errors in the data fetched from these broken chips. Since there are only 140 bits in an ECC data word and there are 144 bits in the bus, the four additional bits are stored in two spare chips. These chips can be used to spare any two of the 70 chips normally used for the data. There are up to 32 spare chips per card as compared to four for G5/G6.

A (144, 132) code has been designed, using the techniques described in [4], to minimize the number of circuits for the ECC implementation. Since the code can accommodate up to 132 data bits and only 128 data bits are required for storage, four excess data bits can be assigned for other usage. In z900, two of the excess bits are used for memory address protection and the other two are used for failure isolation.

To prevent data from being fetched from an erroneous location, two memory address parity bits are treated as two additional data bits in the ECC check-bit generation. However, the memory address parity bits are not stored into the memory. If data is fetched from a memory address with incorrect parity, the ECC decoding logic will detect this as a UE.

When encountering data errors in memory, it is important to isolate the location of the original failure. Therefore, when data is known to have a UE but is required to be stored into memory (e.g., an L2 cache failure on changed data that must be cast out, or a memory error that is being stored as part of a move-page operation), the data is encoded as a special UE pattern. Data that results from a cache or other nonmemory failure is encoded as a "cache special UE." Data that results from a memory error is encoded as a "memory special UE." Data that results from a memory store interface error is encoded as an "interface special UE." Two of the unused data bits in the (144, 132) code are used to represent these three special UE events during the generation of ECC check bits. When the data fetched from storage is a special UE, the special UE is identified by its unique syndrome.

Although the ECC can correct two bits per DRAM, DRAMs with many defects are spared. Sparing the DRAM prevents these defects from lining up in the same ECC data word with other defects (which may fail in the near future) and causing a UE. DRAM sparing can occur either at POR or dynamically while the server is running, as was done for G5/G6. During FPOR, the memory controller runs self-test. It stores and fetches fixed and random patterns while accumulating error counts on a DRAM basis. Some patterns are used across the entire 144 bits so that simultaneous errors from multiple chips can be identified and counted by comparing actual and expected data. Other patterns use the (140, 128) ECC along with the spare chips, to help exercise the ECC logic. The use of unique pseudorandom data for every data word location verifies the addressing function within the DRAM. DRAMs with a high failure count are spared. Spare DRAMs that have high failure counts themselves are not used as spares. A spare DRAM that has been brought into use via sparing can itself be spared out, should it ever reach the high failure count.

The dynamic form of sparing is performed via background scrubbing, a process of error avoidance. This was used on G5/G6 and continues for z900. Scrubbing involves fetching and storing data from all addresses in memory while correcting errors with ECC. By correcting errors and storing good data back into the memory periodically, soft errors (those that occur temporarily due to alpha particles and other perturbations and not due to broken hardware) are corrected. This removes soft errors from memory before they line up with other errors and become uncorrectable. Error counts are accumulated while scrubbing, and DRAMs with high counts are spared. To spare the defective DRAM, the data is first written

to both the original and the spare DRAM. After the scrubbing pass is complete, the spare DRAM contains the corrected data from the original and is then used in place of the original.

Key storage

The G5/G6 operation to request key data resulted in a parallel access of an A and a B copy of the keys from a selected group of four key arrays. If at least one copy had good parity, the key data from the good copy was returned. Having a parity error on both copies resulted in a key UE response being returned. Sophisticated code routines were needed in order to attempt to recover the lost data. The z900 adds another copy, copy C, to each of the four key array groups. When a key is read from memory, the data from all three copies of a key array group are compared. If a mismatch occurs, one of the two copies that match is used for the key access. If one copy has a parity check, the data from the other two copies is compared, and if it matches, the key access is successful. If two copies have a parity error, the data from the third copy is used for the access. Only when there is a parity error from all three copies, or the data from all three copies is a mismatch, is a key UE response returned instead of key data.

During FPOR, the memory controller runs self-test on the key arrays. It stores and fetches fixed patterns of data while it accumulates single-bit error counts on a chip basis. There is a counter for each of the three key copies in each of the four array groups. If a counter exceeds the threshold, the corresponding key copy within the group is disabled. Two copies within a group can be disabled. A trap register monitors the key copies for multiple-bit errors. If a multiple-bit error is detected, the corresponding key copy within the group is disabled.

The status of the key arrays is determined by examining the parity checkers and data comparators for the three key copies. This yields either all good key copies, a single-bit error in a key copy, a single-bit error in each of two key copies, a multiple-bit error in a key copy, or a UE.

Configuration array

The configuration array provides the translated physical address for the absolute address used to access the array. Since this array is seldom written during normal server operation, the data is seldom refreshed. The z900 configuration array is scrubbed to reduce the possibility of data errors. The entire array is scrubbed within 8 to 16 thousand memory controller cycles. That is, the data and ECC check bits are read from the array and written back into the array unless a UE is detected. Should a CE be detected, the data is corrected and the modified data is written back into the array. Similarly, a CE detected while reading the array will force a writeback to the array with the corrected data. The ECC used to protect the data in the configuration array is capable of correcting all single-bit errors and detecting all double-bit errors in the data word consisting of ten data bits and six check bits. Also, there is a pair of valid bits (for redundancy) associated with each entry in the array. Each valid bit has an associated parity bit. If both pairs of valid and parity bits match and have good parity, no CE or UE condition exists. If one valid bit and associated parity has a parity error, a CE is indicated and the valid bit on the good parity is used. If both pairs have a parity error or both pairs contain good parity but do not match one another, a UE is indicated.

L2 cache—memory controller interface

The command and address interfaces from the L2 cache controller to the memory controller are protected by parity bits. A command-response protocol keeps the L2 and memory controllers synchronized. Receipt of a command and address without errors returns a "command accept" status. A request with an address and/or command parity error returns a "command check" status. In contrast to G5/G6, no further processing of the faulty command, either store or fetch, takes place within the memory controller. No clock-stop action is necessary to protect against a data integrity problem, since all operations are terminated on the interface. The request is then reissued by the L2 cache controller.

The command bus utilizes a nonzero null pattern. This pattern is such that any one bit can change during the periods when the command bus is idle; the resulting encoded value is not confused with a valid command and thus can be ignored. Therefore, no unexpected "command check" response is returned to the L2 cache controller, resulting in a clock-stop synchronization error.

The L2 cache controller controls the direction of the bidirectional (bidi) data bus shared between the L2 cache and the memory. The L2 cache controller can request that the direction be changed, but it must wait for an acknowledgment from the memory controller before the direction change can take place. At that time, both sides must change direction such that both sides are driving the bidi bus for one overlapped cycle. The line from the L2 cache controller requesting the change in bidi bus direction is also protected by the command bus parity bit. In the presence of a command parity error (which could be caused by the bidi direction change bit), the bidi bus direction is maintained and no acknowledgment is returned. The acknowledgment signal is protected by the status bus parity bit.

Concurrent upgrade

The z900 extends concurrent upgrades to the memory. This provides improved flexibility and availability to the z900. Memory size is increased using a LICCC diskette

similar to CUoD. Upgrades can be ordered through the Internet by the IBM marketing representative. Many of the orderable memory configurations are delivered with excess (dormant) memory. For example, a z900 with the 20-PU MCM, ordered with 10 GB of memory, is shipped with 16 GB installed. The memory is "dialed down" via LICCC to the ordered configuration of 10 GB. The customer may take advantage of the installed dormant memory and concurrently add all or part of the remaining 6 GB. Concurrent memory upgrades are supported only under the Processor Resource/Systems Manager* (PR/SM*) operating mode [5]. Memory downgrades and temporary upgrades are not supported.

Concurrent repair

Through the memory sparing function, the z900 continues to provide dynamic replacement of defective memory chips fully transparent to the customer's operation. With up to 32 spare DRAMs per memory card, a memory card will rarely have to be replaced because of DRAM failure. In case of catastrophic failure in the memory logic, the partial memory restart function allows the z900 to be configured with half of the memory. A nonconcurrent memory card replacement is then scheduled.

I/O subsystem

The z900 contains a new I/O subsystem and infrastructure delivering high performance and an unprecedented level of RAS. The channel subsystem (CSS) supports three types of operations:

- Channel command word I/O (ESA/390, ESAME architecture).
- Coupling I/O (message architecture).
- Open system adapter I/O (queued direct I/O architecture).

The CSS structure shown in **Figure 3** consists of CPs, SAPs, MBAs, primary, primary extended, and secondary self-timed interface (STI) ports, the new I/O cage supporting the new higher-density I/O cards, and the compatibility I/O cage supporting some I/O cards from previous-generation servers.

The new I/O cage supports the Enterprise Systems Connection (ESCON-16), Fiber Connection (FICON*), Open System Adapter Express (OSA-E) providing Asynchronous Transfer Mode (ATM), Fast Ethernet (FENET), Gigabit Ethernet (GbE) connection, Intersystem Channel 3 (ISC-3) and the PCI cryptographic coprocessor (PCI-CC) cards. The PCI-CC card, though not a channel, is defined in the I/O control program (IOCP) or in the hardware configuration dialog (HCD) and, for completeness, is included here in the I/O subsystem section. The compatibility I/O cage supports the existing ESCON-4, and the Open System Adapter 2 (OSA-2) Token Ring (TR) and Fiber-Distributed Data Interface (FDDI) network connection cards [6].

The new I/O design provides major RAS enhancements in the area of flexible channel path identifier (CHPID) management, CHPID swapping, CUoD for ESCON-16 and ISC-3, port sparing for ESCON-16, concurrent installation, concurrent repair, and RAS support for the Intelligent Resource Director (IRD) I/O functions.

Flexible CHPID management

The flexible CHPID management function relieves the CHPID assignment constraint that existed in previous servers, in which a fixed mapping of CHPIDs to addressable ports was based on the physical I/O card location, regardless of the number of ports that were actually used. The introduction of denser I/O cards and the enablement of individual ports via LICCC for ESCON-16 and ISC-3 cards exacerbates the situation in which an I/O configuration might be constrained by the 256-CHPID architectural limitation. The default CHPID assignment and mapping are performed by the service element (SE) when new hardware is detected. Once created, this assignment and mapping remains unchanged until modified via flexible CHPID management by the service representative under the customer's direction. Flexible CHPID capabilities do not extend to the PCI-CC, since it is not considered to be a channel.

Flexible CHPID management provides the capability to keep I/O definitions the same across servers with different I/O card placements in the I/O cages and/or CHPID number assignments. As with the previous servers, the z900 supports the "channel-swapping" feature to assist the service representative in isolating difficult types of channel failures (usually intermittent) to the source of the problem. The channel-swapping function builds on the new channel CHPID assignment function and supports all channel types except for the internal coupling (IC) channels. The channel-swapping function in the SE guides the service representative to place the CHPIDs to be swapped into service mode and to configure the CHPIDs offline to the OS and swap the I/O cables. Each channel swap must be of the same channel type. To facilitate problem determination, up to four channels can be swapped at the same time. New for z900, once the source of the problem is identified, the service representative can replace any defective channel card concurrently while the cards are in the swapped state.

ESCON-16 port sparing

Although the ESCON-16 channel card is concurrently replaceable, the impact of losing up to 15 channel ports during a repair scenario is significant. Since the most likely failure mode is that of a single port, the capability

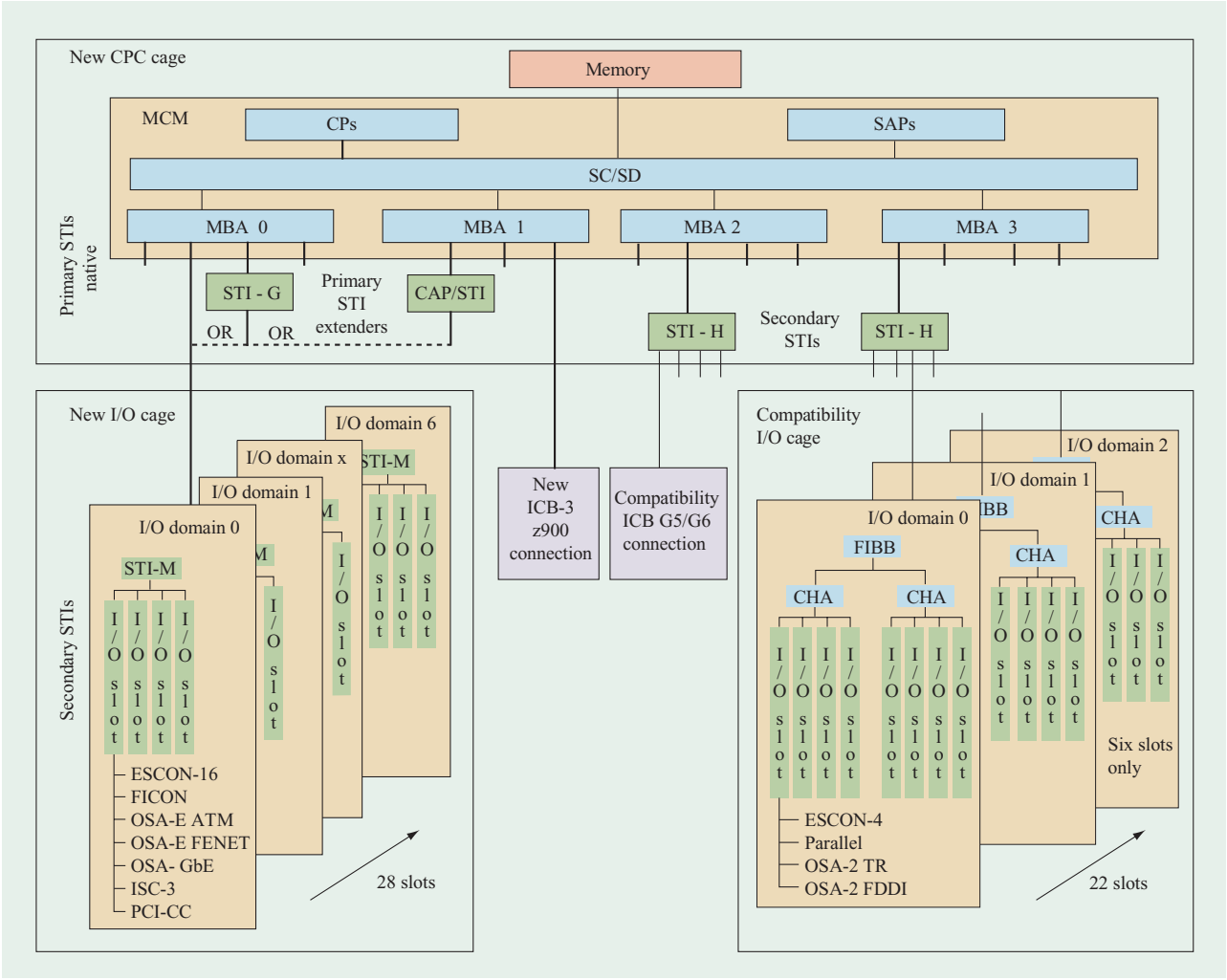

Figure 3

Channel subsystem structure.

of sparing an ESCON* port within the card was introduced. The port-sparing action is activated by the service representative using the repair and verify (R&V) procedures in the service element. The R&V function directs the service representative to ensure a safe state during this process. The SE blinks the port light-emitting diode (LED) of the defective port to indicate that it is offline and the I/O cable can be safely unplugged. It then blinks the spare port LED to indicate where the I/O cable is to be replugged (**Figure 4**). Once this function is completed, the spare port is addressed with the same CHPID value as the failed port, and the CHPID mapping table in the SE is updated accordingly.

ESCON port sparing is not limited to the initial dedicated spare port. In the unlikely event of multiple port failures, the other "reserved" ports can also be used as replacements. Vital product data (VPD) for the channel card is updated to reflect the number of remaining reserved ports and is transmitted to IBM. When an upgrade is ordered, this information is used by manufacturing to determine whether the order can be satisfied by existing hardware or additional cards must be installed.

Concurrent upgrade

The higher technology and packaging density of the new I/O cage and cards has necessitated the implementation of LICCC control for ESCON-16 cards to ensure a higher level of availability.

The ESCON-16 channel cards contain 15 ports which are "usable" by the customer and one port which is

dedicated as a spare and may be used only as a replacement in the event that a used port becomes defective. The ESCON-16 channel cards are always installed in groups of two, and the minimum orderable increment is four ESCON ports. Therefore, in most cases, the number of physically installed ports is greater than the number of ports ordered. These "dormant" ports are controlled by LICCC and are reserved for CUoD. The dedicated spare port is also controlled by LICCC but is not available to be used for upgrade. For example, an order for 32 ESCON channel ports would be delivered with four ESCON-16 cards installed in the new I/O cage, each card with eight ports enabled, leaving seven dormant ports reserved for future upgrade and one spare port per card.

When CUoD is requested, the new LICCC is delivered to the server via diskette or download, reflecting the new enabled configuration. The SE guides the service representative to the card(s) affected by the upgrade by turning on the LED on the slot indicator to identify the card slot in the new I/O cage. The SE blinks the port LEDs to identify the ports that can be used to attach the I/O cables. This automated process ensures that the correct ports are being plugged (Figure 4).

Additional I/O cards can be installed concurrently in empty I/O cage slots to fulfill an upgrade request that cannot be accomplished via CUoD. In these situations, the service representative is guided by the hot-plug function in the SE. As in the CUoD scenario, the SE turns on the LED on the slot indicator to identify the card slot in the I/O cage (empty in this case) to ensure that the new cards are plugged into the correct location. All cards in the new I/O cage can be installed concurrently, and no special consideration is necessary as long as there are sufficient slots available in the existing I/O cage(s) to satisfy the order. All I/O cards, with the exception of the fast internal bus (FIB) and the channel adapter (CHA) in the compatibility I/O cage, are concurrently installable. To address the cases in which additional FIB/CHA cards may have to be installed, a "plan-ahead" feature is available and provides additional FIB/CHA cards in the initial configuration in preparation for future I/O upgrades which exceed the capability of the existing FIB and CHA cards. The I/O cards (ESCON-4, Parallel, OSA-2) in the compatibility I/O cage can also be "uninstalled" and concurrently replaced with higher-bandwidth I/O cards in the new I/O cage. This capability is particularly useful for installations in which the I/O configuration approaches the architectural limit of 256 CHPIDs.

Concurrent repair

All cards in the new I/O cage and all I/O cards except for the FIB and CHA in the compatibility I/O cage can be replaced concurrently. The R&V function in the SE guides the service representative through each step in the procedure to ensure that the ports are offline, the defective card is properly identified, and the correct replacement is installed. For replacement of defective ESCON-16 and ISC-3 daughter cards, the correct LICCC data for the replacement card, matching that of the card being replaced, is automatically downloaded and verified as part of the R&V process. The concurrent installation and repair capabilities for all I/O cards are shown in **Table 1**.

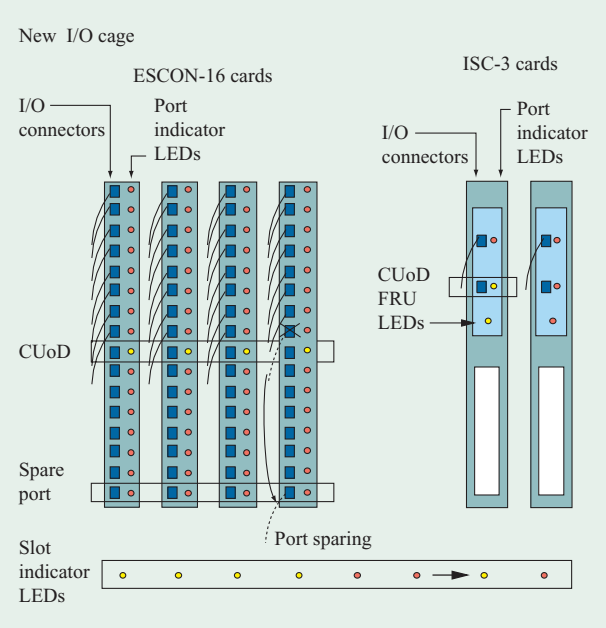

Figure 4

CUoD and ESCON-16 port-sparing example.

Power/cooling subsystem

The power/cooling subsystem is composed of several major hardware elements, each critically important in achieving the high availability requirements of the z900:

1. Central bulk power assembly (BPA) with integrated distribution and controls.
2. Point of load direct current assembly (DCA).
3. Air-moving devices (AMDs).
4. Modular cooling units (MCUs).

The BPA is duplicated, and each of the two BPAs is powered from its own ac line cord. The BPA converts the ac power of the three-phase utility line into the universal 350-V distribution bus. Two such buses are produced by the two BPAs, and each of these buses is capable of powering the entire server. The mode of redundancy is

Table 1 Concurrent installation and repair data for I/O cards.

Card type	*New CPC cage*		*New I/O cage*		*Compatibility I/O cage*	
	Install	*Repair*	*Install*	*Repair*	*Install*	*Repair*
STI-H	Yes	Yes				
STI-G	Yes	Yes				
CAP/STI	N/A	No				
STI-M			Yes	Yes		
ESCON-16			Yes	Yes		
FICON			Yes	Yes		
OSA-E ATM			Yes	Yes		
OSA-E FENET			Yes	Yes		
OSA-E GbE			Yes	Yes		
ISC-3			Yes	Yes		
PCI-CC			Yes	Yes		
ESCON-4					Yes	Yes
Parallel					Yes	Yes
OSA-2 TR					Yes	Yes
OSA-2 FDDI					Yes	Yes
FIB					No	No
CHA					No	No

"active"; that is, both sides share the power load of the server, and if one side fails or is taken offline, the remaining side will seamlessly supply the full load. In addition to duplication of the BPAs, additional RAS robustness is provided by active phase switching within each BPA. If a utility phase is lost for any reason, circuitry within the BPA automatically switches to single-phase operation, and, for most configurations, the server will run indefinitely in this mode, even if one of the BPAs is offline. Since most power losses are single-phase transient dips, the server is impervious to almost all power-line disturbances, even without the benefit of a battery or uninterruptible power supply (UPS) backup.

The BPA also distributes the 350-V bus to all power supplies and cooling devices. They are energized from both buses, and this "cross-coupled redundancy" makes the server highly fault-tolerant, requiring exact coincidence of specific faults to disrupt server operation. Besides the redundancy aspects of the single high-voltage dc distribution bus, two additional properties of this architecture enhance availability. First, because a single duplicated internal voltage bus powers the entire server, an internal battery feature which is fully integrated within the power subsystem is possible. When installed, this feature permits the server to ride through complete power outages for extended periods without resorting to external equipment. Second, the bus voltage of 350 V allows relatively low branch currents to feed the various power and cooling FRUs. This makes it possible to use solid-state branch-protection devices, which are extremely fast-acting, instead of slower fuses or electrical–mechanical devices.

Power is supplied to the DCAs, AMDs, and MCUs in the same manner, employing redundant feeds from both 350-V dc buses. The DCAs produce the precise voltages required by the various logic functions. They are $N + 1$-redundant with regard to their current capacity; i.e., each power boundary contains one more DCA than is necessary to support its load current. The mode of redundancy is also "active," since the DCAs share the load equally under normal conditions.

The AMDs are basically blowers (called motor/scroll assemblies, or MSAs) with attached intelligent speed-control devices (called motor drive assemblies, or MDAs). These devices provide air cooling to the cages and provide the necessary heat removal for everything except the MCM, which is cooled by refrigeration. The logic cages (CPC and I/O) each contain a pair of blowers employing parallel air flow, while the power cage (BPA) contains a pair of blowers in series. In either case, when a blower failure is detected, the speed of the remaining blower is increased to maintain airflow sufficient to prevent malfunction of the devices cooled by the blowers. This mode of redundancy is also "active," but this is not full functional redundancy, since the temporarily sped-up blower will increase the acoustic noise beyond the specification for a fault-free server.

The MCU is the refrigeration unit which provides cooling to the MCM. The MCU is composed of a modular refrigeration unit (MRU) containing a compressor/motor; a condenser; an intelligent controller (MDA-R); an

integral cooling fan (MSA-R); an evaporator; an evaporator cavity humidity sensor (HUSN) to warn against potentially damaging moisture buildup at the MCM; a board heater block to prevent the cold MCM temperatures from spreading to other areas of the CPC board and causing condensation; and a desiccant container to maintain dry air within the evaporator cavity. The MCUs are duplicated but operate separately in order to lengthen compressor service life. Each MCU runs for half the service time, and a scheduled switchover is performed every 160 hours. The mode of redundancy is "standby," since only one MCU is running at a time. If a running MCU fails and thus is no longer capable of maintaining temperature control, an automatic recovery switchover to the standby unit is performed. The evaporator, attached to the MCM copper hat, is internally constructed of two independent loops of copper tubing, one connected to each MCU, through which a liquid refrigerant is passed at low pressure. The evaporation of this liquid within the evaporator FRU removes heat. Control of refrigerant flow, via closed-loop feedback, maintains the MCM chip junction temperatures at approximately 0°C.

Concurrent upgrade

The z900 is shipped with full $N + 1$ power supplies and cooling as standard. The I/O power sequence controller (PSC) card is an optional feature and is concurrently installable.

Concurrent repair

The DCA, MDA, and MDA-R are responsible for fault isolation and error logging for logic load failures, fan motor failures, and MCU failures, respectively. A failed unit is removed and replaced while redundant hardware maintains server operation. The microcode which controls this operation must perform several critical functions in order to guarantee that there is no immediate or eventual disruption due to the repair:

1. Before any removal of hardware, a check is performed to ensure that the redundant hardware can support server operations. If this check fails, there are multiple and coincident faults, and the repair must be scheduled. It cannot proceed concurrently.
2. The repair must be directed such that the correct FRU is replaced.
3. A check is made to ensure that the FRU has been replaced.
4. The microcode level of the FRU is checked and updated if necessary ("auto code download").
5. A check is made that the failure has been removed and that no new problems have resulted from the repair.

The BPA is composed of nine different FRUs: three high-voltage power supplies, three power-distribution units, one power-control unit, one cooling fan, and the power enclosure. All are concurrently replaceable, as are the DCAs. The MDAs can be removed without removing the MSA, but not vice versa. In either case, replacement is concurrent. The IBF and PSC cards are also concurrently replaceable. All of the MCU FRUs, with the exception of the evaporator, support concurrent repair.

Service subsystem

The service subsystem acts as the backbone of the z900 server for control, maintenance, and change management. It includes the new power and service control network (PSCN), the SE, the hardware management console (HMC), and the multi-tiered IBM remote support structure. This section describes the RAS design characteristics for the service network, the auto SE switchover, and the remote support structure.

Service network

Two of the efforts toward providing CRO are to eliminate sources of server outage and to increase concurrent repair capabilities. The introduction of the PSCN on the z900, as described in [7], advances both of these pursuits.

The intra-server communication method used on previous servers was implemented through the parallel printer port of the SE to a universal power controller (UPC) in the main logic cage. This UPC then distributed the control information through its internal network to the UPCs in the other cages and also to the bulk power controller (BPC) in the power cage. If a problem occurred with the main UPC or with the communication path between the SE and the main UPC, communication was lost between the SE and the entire server, including the power cage. Another limitation of the UPC network was that it used a nonstandard interface for communications.

In the z900, the BPA houses the communications hub for the diagnostic, control, and service subsystem (PSCN). This hub is duplicated (one per BPA) and operates using standby redundancy (master/slave with automatic "failover"). It provides the communication path (Ethernet) between the logic cages, the BPA cage itself, and the SEs. The criticality of this communications function is such that the BPAs cross-power one another so that failure of the input power to a BPA will not disturb the communications hub within that BPA, should it happen to be the master.

In the z900, the single UPCs in each cage were replaced by dual cage controllers (CCs). Each individual cage (CPC, I/O, and BPA) is controlled by a cage controller, which is responsible for building the server configuration, scanning logic, providing FRUs with information critical to their operation, and running all cage-related serviceability microcode. For the CPC and I/O cages, the cage

controllers reside in the DCAs. This was a natural location, since they were required to be redundant; placed within the DCAs, which were already redundant, they take up no critical board space. The cage controllers operate in a master/slave arrangement of standby redundancy with automatic failover.

The new redundant PSCN provides numerous RAS and non-RAS benefits over the UPC network. The dual network and dual cage controllers eliminate potential sources of outage in the communication path between the support element and the server. The Ethernet network allows direct communication between the SE and each cage. With this direct communication and dual path, FRU isolation capability is enhanced. With direct, multi-path communication with each of the cages, the SE is better able to isolate the source of any problems. Finally, using an Ethernet network as the communication path allows for use of the standard TCP/IP protocol for communications.

Alternate SE auto switchover

An alternate SE was introduced on G5/G6, which served as a backup for the primary SE. While the alternate SE automatically kept a mirrored copy of the primary SE, a manual intervention was required to effect a switchover. This involved flipping a physical switch on the front panel of the server. For the z900, this switchover has been fully automated. Both the primary and alternate SEs can now communicate with the server and each other at the same time, a benefit of the new PSCN.

On G5/G6, the SE that had communication with the server assumed the role of the primary SE, while the other was the alternate. Once the physical switch was thrown and the primary SE lost communication, it would switch roles and become the alternate. The alternate, on the other hand, would then detect that it had communication with the server, and it too would switch roles. In z900, both SEs have communication with the server, and therefore both have the capability of assuming the primary role. Upon SE initialization, the two SEs go through an algorithm to ensure that only one becomes the primary [8].

The communication between the SEs allows for the alternate SE to be able to detect when there is a problem with the primary SE. When a problem is detected, the alternate SE informs the primary and any attached HMCs that it will attempt to take over as the primary SE. A service call is placed to repair any problems with the former primary SE. The former primary SE is fenced to prevent any communication collisions, and the alternate SE assumes the role of the primary.

A "soft" switch on the HMC has also been introduced, allowing a remote operator to initiate the SE switchover. It is no longer necessary to have someone in physical contact with the server to take advantage of the redundancy of the SEs.

Concurrent upgrade

The z900 is shipped with the redundant service network and SEs as standard. Additional HMCs are installed concurrently.

Concurrent repair

Prior to z900, since there was only a single UPC in each cage, along with the fact that the CPC cage UPC was the server master, concurrent repair of the UPCs was not possible. In the z900, the cage controllers are packaged in the DCAs and use the cabling of the power network. This packaging, on a concurrently maintainable FRU, along with the dual cage controllers for each cage, enables concurrent replacement of what were previously nonconcurrently replaceable FRUs. In the z900, as in the previous servers, the replacement of a defective SE is concurrent.

Remote support subsystem

The automatic capture, retrieval, and transmission of critical error data at the time of a server failure is essential in providing an effective remote support structure. In addition, the ability to retrieve informational and event log files to perform predictive maintenance is an important factor in today's remote support environment. The z900 continues to enhance the extensive remote support capabilities in both of these areas.

First error data capture filter

The ability to gather and preserve pertinent data related to a failure in the server, and do it at the earliest possible time after the failure, is known as first error data capture (FEDC). This is essential to identifying the root cause of the problem, enabling the server to take the correct recovery action and initiating the proper service action if service is required.

An important element of problem isolation is retrieving the correct error-log files from the server and transmitting them to the support team. Development engineering determines which files are required for each type of problem encountered. This list of files is then stored in the HMC or the SE. Each time a particular error is encountered, the associated list of files is transmitted to the IBM Remote Technical Assistance Information Network (RETAIN), an IBM-internal database of technical-support resource information. The z900 introduces an FEDC filter in which the list of files can be changed as new information about a given failure is learned. Files may be added to or removed from the list. This allows the flexibility to retrieve only those files absolutely required for effective problem isolation. It

optimizes file retrieval, transfer, and storage in order to speed problem resolution. The FEDC filter resides in the z900; each time a new version of the filter is released to RETAIN, it is automatically downloaded to the server on the next connection to RETAIN.

Get other files easily, remotely

Problems occur for which additional data is required for remote analysis. Even with the FEDC filter, certain files will be required on occasions that do not warrant updating the FEDC filter. A new function called "get other files easily, remotely" (GOFER) has been introduced on the z900. This function gives product engineering and the service support centers the ability to define specific files from specific servers for retrieval. The party requiring the file(s) submits a request to a VM service machine. The service machine creates a unique file for that specific server serial number and sends the file to RETAIN. The next time that z900 connects to RETAIN, the data file is retrieved and forwarded to the requesting party. This enhancement reduces the need to send a service representative to the server to transmit specific files for analysis.

Multiple phone servers

The server's call-home effectiveness has been improved for z900. Each HMC now has the responsibility to act as a "phone-home" server. Previously, a single HMC acted as the phone server for a given zSeries server. If that HMC was unable to connect to RETAIN, a problem call could be lost. With z900, each server's SE may now access multiple HMCs, as required, to achieve the phone-home connection. If the first HMC cannot connect to RETAIN, that information is passed back to the requesting SE, and the SE requests another HMC phone server to place the problem call to RETAIN. This continues until the problem call is placed to RETAIN or until every HMC phone server has attempted to place the problem call and failed. In that unlikely event, the problem call is queued for later transmission once the phone connections are restored. This process enhances call-home effectiveness.

Transmit system availability data

The transmit system availability data (TSAD) function is a regularly scheduled transmission of data from the customer's server to the IBM support structure. The data includes power status, SE status, CBU status, microcode level (MCL) apply status, and recovery information. This function is set up by the customer as a scheduled operation. Each week, at the appointed day and hour, the server connects to RETAIN and transmits the data. This data is used for multiple purposes. RAS engineers use the data to verify that the servers can connect to RETAIN on a weekly basis, thus ensuring that the RETAIN connection works. For z900, RAS engineers automatically open a problem management hardware (PMH) record in RETAIN for any server that has connected three weeks in a row (showing that scheduled operations are set and working) and then misses two weeks (showing that scheduled operations have been interrupted). This alerts the service representative that the connection to RETAIN may no longer be working.

TSAD data is also used to verify $N + 1$ power status, second SE operation status, and CBU expiration dates. If problems are found, a PMH record is opened, if one is not already open, in order to alert the service representative to the problem. This data is also sent to Resource Link to be available for customer access. The TSAD data process has the flexibility to add or delete files as required.

Virtual RETAIN

A virtual RETAIN has been introduced for z900 that stores all files relevant to a server problem on that server's SE. In the event that the connection to RETAIN fails, these files, which would normally be sent to RETAIN, are preserved on the virtual RETAIN. Once the RETAIN connection is restored, the files are transmitted. This ensures that important failure data is not lost.

Resource Link

A key feature of remote support is the presentation of server-related data to those persons requiring the information. z900 expands the ability to view specific server information remotely. Resource Link is a Lotus-Notes**-based server used to display the information obtained from the vital product data (VPD) files, TSAD, and MCL status. The intended users are the customers, service representatives, marketing representatives, and business partners.

Resource Link enables customers to access server data from a central location. Resource Link has a server outside the IBM firewall to allow external access. Any user can access Resource Link from a web browser and view specific server data after being registered to access that data. Registration can be based on an individual server type and serial number, a customer number (group of servers in a single customer location), or an enterprise number (group of customer numbers for a given corporation or enterprise.)

The engineering change/microcode level (EC/MCL) feature of Resource Link displays the latest EC/MCL status of each individual server. EC/MCL information allows the user to quickly determine how many and which MCLs have been received (REC), activated (ACT), and accepted (ACC) from all of the MCL fixes available in RETAIN (MAX). The example of a particular server's

Table 2 MCL status example.

EC	*Name*	*REC*	*ACT*	*ACT time*	*ACC*	*MAX*
H25113	HMC/SE MVS polite/SUF	1	1	2000/12/30 21:59:39	1	1
E26931	SE base power control	2	2	2000/12/30 21:59:39	2	2
E26939	SE cage LIC	1	1	2000/12/30 21:59:39	1	1
E25106	SE coupling facility code	1	1	2000/12/30 21:59:39	1	1
H25108	SE channel code	1	1	2000/12/30 21:59:39	1	2
H25101	SE Channel I diagnostics	0	0	2000/12/30 21:59:39	0	0
H25104	PCI crypto channel	0	0	2000/12/30 21:59:39	0	0
H25105	SE FCS code	1	1	2000/12/30 21:59:39	1	1
E26929	SE MISR data (self-test)	0	0	2000/12/30 21:59:39	0	0
H25103	SE hydra code	1	1	2000/12/30 21:59:39	1	1
H25110	I390/PU millicode	0	0	2000/12/30 21:59:39	0	0
H25107	SE LPAR code	0	0	2000/12/30 21:59:39	0	0
H25129	SE C-disk	2	2	2000/12/30 21:59:39	2	2
H25109	SE OSA/Flash/ROM	0	0	2000/12/30 21:59:39	0	0
H25102	SE processor exerciser	0	0	2000/12/30 21:59:39	0	0
E26932	SE PSCN LIC	1	1	2000/12/30 21:59:39	1	1
E26928	SE SOS code	0	0	2000/12/30 21:59:39	0	0
H25117	SE D-disk	1	1	2000/12/30 21:59:39	1	1

MCL information in **Table 2** shows that all MCL fixes have been installed except for H25108. Additional detail for each MCL shows whether it is a high-impact or pervasive problem (HIPER) MCL and whether the installation of the MCL is disruptive or nondisruptive. This enables the customer to optimize the scheduling of microcode service. Once a day, a Lotus Notes agent reviews all new MCLs that were released the previous day. If any were released as HIPER MCLs, the agent scans every MCL record, looking for any server which may need that HIPER MCL installed. If a customer or service representative is registered for that server, a Lotus note is sent stating that a HIPER has been released and the registered server may have to have the HIPER MCL installed. This checking is based on the EC streams installed and MCLs activated on each server. Each server which has the affected EC stream and does not have the MCL activated is notified.

Resource Link also provides information on the date/time of the call-home and the operating status of the $N + 1$ power subsystem and the alternate SE. A Lotus note defining the problem is sent to each registered user of every server that has a power fault detected, an error on the alternate SE, or a failure to call home in the last 15 days.

The CHPID management tool is another significant feature of Resource Link. It provides the flexibility to modify the default CHPID assignment (described in the section above on CSS) to meet the customer's own individual requirement. This may be done to match an existing configuration or to establish a new configuration. The availability-mapping function within the CHPID management tool allows a maximum-availability CHPID assignment mapping to be planned. Resource Link creates a new CHPID assignment file. This file is saved on a diskette and is used later by the service representative to modify the default configuration in the z900.

Parallel Sysplex

Parallel Sysplex continues to provide the highest level of continuous reliable operation. z900 introduces the Intelligent Resource Director (IRD), which provides unparalleled flexibility to the sysplex. The flexibility and automation provided by the IRD give the customer new options for managing the sysplex and handling failures. z900 also introduces dynamic ICF expansion to provide temporary capacity growth to an ICF image. z900 enhances the serviceability of the ISC-3 cards and the ETR/OSC cards, thereby improving availability.

Intelligent Resource Director

OS/390* introduced the workload manager (WLM), which enabled the operating system to manage workloads across multiple images of OS/390. z900 introduces the Intelligent Resource Director (IRD), which gives z/OS, through its new WLM, more tools and flexibility with which to manage workloads [9].

The IRD allows z/OS to move CP resources among logical partitions on the z900. It also allows z/OS to change the relative weights of the logical partitions (weights are used by PR/SM during dispatching of logical partitions [5]). The WLM was designed to handle failures. The additional tools provided by the IRD are used by the WLM to improve recovery. Individual CP failures are handled by load balancing (the WLM responds to resource demands from the z/OS images, whether caused by workload demands or failures). The resources associated with a failed logical partition are made available to the

other partitions when the failing partition is deactivated. This is done by reapportioning the remaining weights. When the partition is reactivated, the initial weights defined in the LPAR profile are used. The IRD I/O functions, dynamic channel path management (DCM) and channel subsystem I/O priority queueing (CSS IOPQ), optimize the channel resource utilization across logical partitions within a single server assigned to a Parallel Sysplex.

In addition to providing improved I/O performance by dynamically adjusting the available channel bandwidth to the place where it is most needed, the DCM function enhances the reliability and availability of ESCON and FICON channel paths to control units. The channel paths from the server to an ESCON/FICON director are considered to be a pool of resources accessing any of the control units attached to the director. The DCM component automatically bypasses failed or "hung" paths and provides additional and reliable paths to that control unit by avoiding failure points in the new I/O path.

ICF expansion

The z900 introduces the capability to concurrently divert capacity to a coupling facility (CF) LPAR with dedicated ICFs from a z/OS LPAR with shared CPs. This new function allows the ICF image to expand to meet peak workloads or a disaster situation. The dynamic CF-dispatching feature must be enabled in the CF LPAR image profile.

Intersystem coupling

ISC-3 consists of two hot-pluggable ISC-3 daughter cards, containing two links, plugged into an ISC-3 mother card. The minimum orderable increment for ISC-3 is one link; for high availability, the coupling link connections are spread across multiple mother and daughter cards, thus also resulting in "dormant" ISC-3 links. For example, an order for two coupling links is shipped with two mother cards, each with one daughter card with a single link enabled. The two remaining "dormant" ISC-3 links are reserved for CUoD.

External time reference

The external time reference (ETR) function is packaged on two separate ETR/OSC cards, which are shared with the system oscillator. Each card houses one physical ETR port and provides the optical-to-electrical conversion for this port. The CLK chip on the MCM contains port selection logic and the clock logic (G5/G6 had this logic on the MBA chip).

The ETR network was designed with redundancy. The sysplex timer function is normally installed with two IBM 9037 timer units for redundancy. Each unit provides a separate fiber to the z900 through the ETR/OSC cards.

Concurrent upgrade

G5/G6 saw the marriage of PU sparing with CBU in the Parallel Sysplex. Capacity Backup–Geographically Dispersed Parallel Sysplex* (CBU–GDPS*) enables z/OS to do at the Parallel Sysplex level what LIC does with PU sparing in a single server. When the OS recognizes a failure (a permanent loss of resources), it can request that the remaining servers activate their CBU to restore the lost resource. Effectively, z/OS is able to add CPs on a different server, possibly at a geographically separated site. CBU-GDPS gives the customer a powerful, reliable, multi-server, multi-site recovery tool.

The dormant ports in the ISC-3 cards are enabled via CUoD similarly to the ESCON-16 CUoD described earlier (Figure 4). Additional ISC-3 cards can be installed concurrently in empty I/O cage slots to fulfill an upgrade request that cannot be accomplished via CUoD.

Concurrent repair

The ISC-3 mother and daughter cards are concurrently replaceable. Since the ETR function shares the ETR/OSC card with the system oscillator, the status of the system oscillator, active or standby, determines whether the ETR/OSC card can be replaced concurrently (described in the preceding section on the processor subsystem).

Licensed Internal Code subsystem

The "beating heart" of a zSeries server is the Licensed Internal Code (LIC). This microcode manages and controls the hardware for processing, communicating, error handling and recovery, configuring, environmental monitoring, service notification, and service execution. It is critical that this microcode be of the highest possible quality and be flexible to changes to support the hardware (and itself), be monitored closely in the field, and be serviceable in a rapid and nondisruptive manner.

The zSeries microcode development process governs the design and test of all LIC. Key steps in that process include the following:

- The microcode work breakdown structure, a set of tools for project planning and execution.
- A design-requirements process leading to a documented specification, with reviews and approvals.
- An approved test plan.
- Unit-level design reviews, both high-level and detail-level.
- Unit-level microcode reviews.
- Unit-level testing to ensure base function and error handling.
- Development testing to include multiple units within the subsystem.

Table 3 z900 RAS and zSeries RAS building blocks/eServer self-management attributes for continuous reliable operation.

zSeries RAS strategy	*z900 RAS enhancement*	*eServer self-management attributes*
Error prevention	MCM testing	Self-protecting
Service/support	Cryptographic coprocessor element Single-chip module	Self-optimizing
Error recovery Service/support	Redundant system oscillators	Self-healing
Change management	Concurrent CBU downgrade	Self-configuring
Error recovery	L2 cache delete on first UE	Self-healing
Error recovery	L2 cache cache-line "fuse" apply	Self-healing
Error prevention	Multiple memory module sparing	Self-protecting
Error prevention	Key store triple redundancy	Self-protecting
Error prevention	Configuration array scrubbing	Self-protecting
Error recovery	Memory interface—command check	Self-healing
Change management	Concurrent memory upgrade	Self-configuring
Change management	Flexible CHPID management	Self-configuring
Problem determination	Service with channels swapped	Self-healing
Change management	CUoD for ESCON-16 and ISC-3	Self-configuring
Error recovery	ESCON-16 port sparing	Self-healing
Service/support	Concurrent repair—I/O, oscillator, ETR	Self-optimizing
Error prevention Problem determination Service/support	Redundant PSCN network	Self-protecting Self-healing Self-optimizing
Error recovery	Alternate SE autoswitchover	Self-healing
Problem determination	FEDC file filter	Self-healing
Problem determination	GOFER file retrieval	Self-healing
Service/support	Multiple HMC phone servers	Self-healing
Service/support	RETAIN connectivity Virtual RETAIN	Self-healing
Change management Service/support	Resource Link	Self-configuring Self-healing
Error recovery	CF structure duplexing	Self-healing Self-configuring
Change management	Intelligent Resource Director	Self-configuring Self-optimizing
Error recovery	CBU–GDPS	Self-healing
Error prevention	LIC simulation tools	Self-protecting

- Joint system test to engage the test organization with development for familiarization and preparation for full system test.
- System test to include full subsystem, multiple subsystems, customer environments, error injection, recovery, and serviceability.

Problems are carefully tracked. Targets are established and measured as to problem quantity, severity, discovery rate, and resolution rate. Databases are in place to track problems and fixes by assigned ownership.

Throughout the design and test phases, simulation is used to greatly speed up the process. Where actual hardware or other related microcode modules are not yet available, those roles are simulated. Thus, virtually every possible action/reaction between the microcode being tested and the simulated hardware/microcode can be exercised on the basis of the design specifications [10–12].

The server's early support program gives the first "real-customer" view of microcode quality and ensures microcode serviceability. Prior to general availability, the microcode problem assessment process in product

engineering and development is up and running. In this process, every field microcode problem is assessed as to potential severity and pervasiveness. Fixes are prioritized on the basis of this assessment.

Microcode fixes undergo thorough review and testing in development and then are packaged into MCLs as single or small groups of fixes. MCLs are delivered to the field via RETAIN. Service representatives and customers are made aware of the availability of the fixes by communication from product engineering and the service support center, and by information on Resource Link. MCLs may be automatically downloaded from RETAIN to the server.

The design objective is that all MCLs be nondisruptive. That is, the MCLs can be activated concurrently. The track record for S/390 and zSeries is that well over 90% of all MCLs may be activated concurrently. This is a major contributor to high server availability. The established high-quality standards of this LIC design process and the recent enhancements enable the z900 to sustain the zSeries objective to continuous reliable operation.

Summary of z900 enhancements

The z900 RAS enhancements described in this paper clearly advance the zSeries server objective of continuous reliable operation. Each of the RAS enhancements delivers improved overall RAS performance and supports the zSeries RAS building blocks and eServer self-management attributes (**Table 3**).

Conclusion

From the perspective of reliability, availability, and serviceability (RAS), and continuous reliable operation (CRO), the z900 monitors and manages itself to prevent errors, detects and reacts effectively to problems with a minimum of outside intervention, and survives those problems with minimum capacity loss and minimum disruption. Beyond the problem/recovery/service scenario, CRO means having the capability to change and grow, again with minimum disruption and intervention. These objectives, which form the basis for autonomic computing, have been enhanced in each generation of CMOS servers. They have served well to achieve for the zSeries its reputation as the "gold standard" for high availability in eBusiness computing. Now, for the z900, the RAS capability has once again been significantly enhanced.

Acknowledgments

The authors would like to recognize the other members of the zSeries RAS Council for their support and contribution in pursuing/delivering the highest RAS standard in the industry (T. Franklin, J. Li, S. Swaney, D. Cole, C. Harshberger, M. Mueller, W. Fischer).

References

1. M. Mueller, L. C. Alves, W. Fischer, M. L. Fair, and I. Modi, "RAS Strategy for IBM S/390 G5 and G6," *IBM J. Res. & Dev.* **43,** No. 5/6, 875–888 (1999).
2. P. R. Turgeon, P. Mak, M. A. Blake, M. F. Fee, C. B. Ford, P. J. Meaney, R. Siegler, and W. W. Shen, "The S/390 G5/G6 Binodal Cache," *IBM J. Res. & Dev.* **43,** No. 5/6, 661–670 (1999).
3. J. Probst, B. D. Valentine, C. Axnix, and K. Kuehl, "Flexible Configuration and Concurrent Upgrade for the IBM eServer z900," *IBM J. Res. & Dev.* **46,** No. 4/5, 551–558 (2002, this issue).
4. C. L. Chen, "Symbol Error Correcting Codes for Memory Applications," *Proceedings of the 26th International Symposium on Fault-Tolerant Computing*, June 1996, pp. 200–207.
5. IBM Corporation, *PR/SM Users Guide*, Order No. SB10-7033-00; available through IBM branch offices.
6. D. J. Stigliani, Jr., T. E. Bubb, D. F. Casper, J. H. Chin, S. G. Glassen, J. M. Hoke, V. A. Minassian, J. H. Quick, and C. H. Whitehead, "IBM eServer z900 I/O Subsystem," *IBM J. Res. & Dev.* **46,** No. 4/5, 421–445 (2002, this issue).
7. F. Baitinger, H. Elfering, G. Kreissig, D. Metz, J. Saalmueller, and F. Scholz, "System Control Structure of the IBM eServer z900," *IBM J. Res. & Dev.* **46,** No. 4/5, 523–535 (2002, this issue).
8. B. D. Valentine, H. Weber, and J. D. Eggleston, "The Alternate Support Element, a High-Availability Service Console for the IBM eServer z900," *IBM J. Res. & Dev.* **46,** No. 4/5, 559–566 (2002, this issue).
9. W. J. Rooney, J. P. Kubala, J. Maergner, and P. B. Yocom, "Intelligent Resource Director," *IBM J. Res. & Dev.* **46,** No. 4/5, 567–586 (2002, this issue).
10. S. Koerner, M. Kuenzel, and E. McCain, "z900 Server System Microcode Simulation: The Virtual Power-On Process—An Innovative Approach for Microcode Verification," *IBM J. Res. & Dev.* **46,** No. 4/5, 587–595 (2002, this issue).
11. J. von Buttlar, H. Böhm, R. Ernst, A. Horsch, A. Kohler, H. Schein, M. Stetter, and K. Theurich, "z/CECSIM: An Efficient and Comprehensive Microcode Simulator for the IBM eServer z900," *IBM J. Res. & Dev.* **46,** No. 4/5, 607–615 (2002, this issue).
12. J. Kayser, S. Koerner, and K.-D. Schubert, "Hyper-Acceleration and HW/SW Co-Verification as an Essential Part of IBM eServer z900 Verification," *IBM J. Res. & Dev.* **46,** No. 4/5, 597–605 (2002, this issue).

Received September 21, 2001; accepted for publication March 18, 2002

Luiz C. Alves *IBM eServer Group, 2455 South Road, Poughkeepsie, New York 12601 (alves@us.ibm.com).* Mr. Alves is a Senior Engineer working in the zSeries system design group. He graduated from New York University in 1975 with a B.S. degree in electrical engineering and received his M.S. degree in electrical engineering in 1977 from the Polytechnic Institute of New York. He joined IBM in 1977 working in the advanced system manufacturing engineering organization, where he held various technical and managerial positions. In 1985 he was named 3090 field quality assurance manager, and in 1987 he became the RAS manager for the 9021 processor families. He is currently responsible for defining the RAS requirements for future products.

Myron L. Fair *IBM eServer Group, 2455 South Road, Poughkeepsie, New York 12601 (mlfair@us.ibm.com).* Mr. Fair is a Senior Engineer in the zSeries systems RAS group. He graduated from the University of Illinois in 1967 with a B.S. degree in mathematics. He joined IBM that same year in the Systems Development Division, where he held various technical positions related to field and development RAS, and in special contracts technical assessment. In 1980, he transferred to the group staff headquarters and became manager of product RAS analysis in 1984. He is currently the team leader for RAS requirements and objectives for zSeries servers.

Patrick J. Meaney *IBM eServer Group, 2455 South Road, Poughkeepsie, New York 12601 (meaney@us.ibm.com).* Mr. Meaney received a B.S. degree in electrical and computer engineering from Clarkson University in 1986 and an M.S. degree in computer engineering from Syracuse University in 1991. He is a Senior Engineer in eServer zSeries custom hardware design. He is the SCE RAS and recovery leader responsible for design for fault tolerance, error avoidance, error correction, recovery, and design for test and debug. He was also the SCE timing leader for the G4, G5, G6, and z900 servers. He is responsible for definition of SCE RAS and recovery features of future zSeries CMOS systems as well. Since joining IBM Poughkeepsie in 1986, he has held design and timing leadership positions on the ES/9021 bipolar-based servers as well as the S/390 G4, G5, G6, and z900 CMOS systems. Mr. Meaney holds fifteen U.S. patents and has nine patents pending. He has received several awards, including IBM Outstanding Technical Achievement Awards for H5, G4, G6, and z900 and an IBM Outstanding Innovation Award for the G5 design.

C. L. (Jim) Chen *IBM eServer Group, 2455 South Road, Poughkeepsie, New York 12601 (clchen@us.ibm.com).* Dr. Chen is a Distinguished Engineer in the zSeries systems development group. He received a B.S. degree from the National Taiwan University and a Ph.D. degree in electrical engineering from the University of Hawaii. Prior to joining IBM, he was a Research Assistant Professor of the University of Illinois at Urbana–Champaign. Dr. Chen has worked in the areas of error-correcting codes, computer reliability, and digital data cryptography. He has received seventeen IBM Invention Achievement Awards, five IBM Outstanding Innovation Awards, and an IBM Corporate Award. Dr. Chen is a Fellow of the Institute of Electrical and Electronics Engineers and a member of the IBM Academy of Technology.

William J. Clarke *IBM eServer Group, 2455 South Road, Poughkeepsie, New York 12601 (wjclarke@us.ibm.com).* Mr. Clarke is an Advisory Engineer in the zSeries product engineering group. He graduated from Rutgers University in 1982 and joined IBM in Poughkeepsie that same year. He worked on various chip designs for the system control element for the 3090 processor family. He also held development responsibility for the recovery design, error detection, fault isolation, and service interface. He worked in chip design on the 9021 family in the Level 2 cache. Mr. Clarke worked on server bringup for the 9021 and continued to hold RAS responsibility for the storage subsystem. He worked in engineering systems test in recovery and serviceability, joining product engineering in 1994 on the processors and storage team.

George C. Wellwood *IBM eServer Group, 2455 South Road, Poughkeepsie, New York 12601 (wellwd1@us.ibm.com).* Mr. Wellwood joined IBM in 1963 in the thin-film memory development area. In 1969 he joined the S/360 Model 195 development team and has since remained in mainframe development. He was a designer on the S/370 Models 3032 and 3033. On the 9021 servers, he was a designer on the L3 memory controller. On the G4, G5, and G6 CMOS servers, he was a designer on the L2 cache and memory subsystem. For the z900, he was a designer on the L3 memory controller. Mr. Wellwood is an Advisory Engineer in zSeries custom SCE design. He received IBM Outstanding Technical Achievement Awards for his work on S/390 G4 L2 cache development in 1997 and S/390 G5 cache and memory subsystem development in 1998. He holds two U.S. patents and has four publications.

Norman E. Weber *IBM eServer Group, 2455 South Road, Poughkeepsie, New York 12601 (nweber@us.ibm.com).* Mr. Weber is an Advisory Engineer working in the zSeries RAS group. He joined IBM as a Customer Engineer in Chicago in 1967, working on mid-range 360 and 370 processors. In 1980 he moved to Endicott, New York, to work in service planning supporting 4300 processors and developing service strategies for 9370 systems. Mr. Weber was promoted to his current position as an Advisory RAS Engineer in 1990 and moved to Poughkeepsie in 1995.

Indravadan (Dan) N. Modi *IBM eServer Group, 2455 South Road, Poughkeepsie, New York 12601 (modidan@us.ibm.com).* Mr. Modi is an Advisory Engineer in the zSeries systems RAS group. He graduated from Gujarat University, India, in 1966 with a B.S. degree in electrical engineering and received his M.S. degree in electrical engineering in 1968 from Utah State University. He joined IBM in Poughkeepsie, New York, that same year working in the power system design area, where he held various technical positions. He moved to IBM East Fishkill in 1982 as a manager of electrical analysis in the MLC area, responsible for substrate electrical characteristics along with delta I and coupled noise analysis. Mr. Modi returned to Poughkeepsie in 1986 to join the system assurance group, where he was responsible for system technology assurance. In 1990, he joined the RAS group, where he was responsible for technology reliability, SPQL, and development of system test and acceptance specifications. In 1998 he joined the custom microprocessor design group, where he was responsible for chip/MCM reliability, SPQL, and the system test and acceptance specifications for zSeries. In 1999, he returned to the systems RAS group, with his original responsibilities.

Brian K. Tolan *IBM eServer Group, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (tolan@us.ibm.com).*
Mr. Tolan is an Advisory Engineer working in the zSeries Licensed Internal Code development group. He graduated from Columbia University with a B.S. degree in electrical engineering in 1981. He joined IBM at Endicott, New York, that same year, working in the first level packaging test development group. He has worked in system serviceability engineering development since 1988.

Fritz Freier *IBM eServer Group, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (freier@de.ibm.com).* Mr. Freier is an Advisory Engineer working in the support element application development group. He joined IBM in 1973, working as a Customer Engineer. He has held various positions in product assurance test and microcode development and is currently the team leader for system serviceability engineering in Boeblingen.