Packaging the IBM eServer z990 central electronic complex

J. C. Parrilla
F. E. Bosco
J. S. Corbin
J. J. Loparco
P. Singh
J. G. Torok

The z990 eServer™ central electronic complex (CEC) houses four multichip-module-based processor units instead of one, as in the previous-generation z900 eServer. The multichip module (MCM) input/output pin density in z990 processor units is more than twice that of the MCMs in z900 processor units. This increase in packaging density and the consequent tripling of the current drawn by the processor units were accommodated by the first-time use of land grid array (LGA) MCM-to-board interconnections in an IBM zSeries® eServer. This was done by using innovative refrigeration cooling of the MCM with air cooling as backup, and by a new mechanical packaging and power distribution scheme. This paper describes the mechanical engineering of the CEC cage, the LGA MCM-to-board interconnection scheme, and the mechanical isolation of the MCM evaporator–heat-sink mass from the LGA contacts. The paper also describes the electrical power and the cooling solutions implemented to meet the more demanding requirements of the denser CEC package.

Introduction

The IBM z990 eServer*, an evolution of the z900 eServer [1], consists of two frames, A and Z, as shown in **Figure 1**. Frame Z houses the bulk power assemblies (BPAs) which convert the utility power into 350 Vdc. The 350 Vdc is distributed throughout the system to point-of-load converters, called distributed converter assemblies (DCAs), which convert the 350 Vdc to dc low voltages required by the input/output (I/O) and the CEC cages. The BPAs also feed the 350 Vdc power to the drive circuits for the compressors in the modular refrigeration units (MRUs) and the blowers in the air-moving assemblies (AMAs). The Z frame also houses up to two I/O cages and the associated AMAs. The A frame houses the CEC cage, an I/O cage, two MRUs, and two AMAs.

A CEC is a specialized device that performs all of the high-speed computing associated with a computer. The z990 CEC cage, while not much changed in overall dimensions compared with that in the z900 eServer, contains up to four processor unit (PU) books, each PU book containing one multichip module (MCM). In comparison, the z900 CEC cage houses only one MCM. The z990 MCMs have 2.3 (= 0.6/0.26) times higher I/O density than the z900 MCMs, as shown in **Table 1**. As a consequence, the z990 cage requirement for electric current drawn is triple that of the z900 cage.

This paper describes the z990 eServer power, packaging, and cooling solutions that overcome the challenge of quadrupling the number of MCMs in a CEC cage, more than doubling the MCM I/O density and tripling the current draw by the CEC cage compared with its z900 eServer predecessor.

PU books and the CEC cage

To meet the on-demand needs of today's business environment, the z990 eServer provides a significant increase in system scalability over its z900 predecessor by taking an MCM processor-in-book design approach with a CEC cage housing up to four PU books. From a processor and memory perspective, the z990 PU book design

0018-8646/04/$5.00 © 2004 IBM

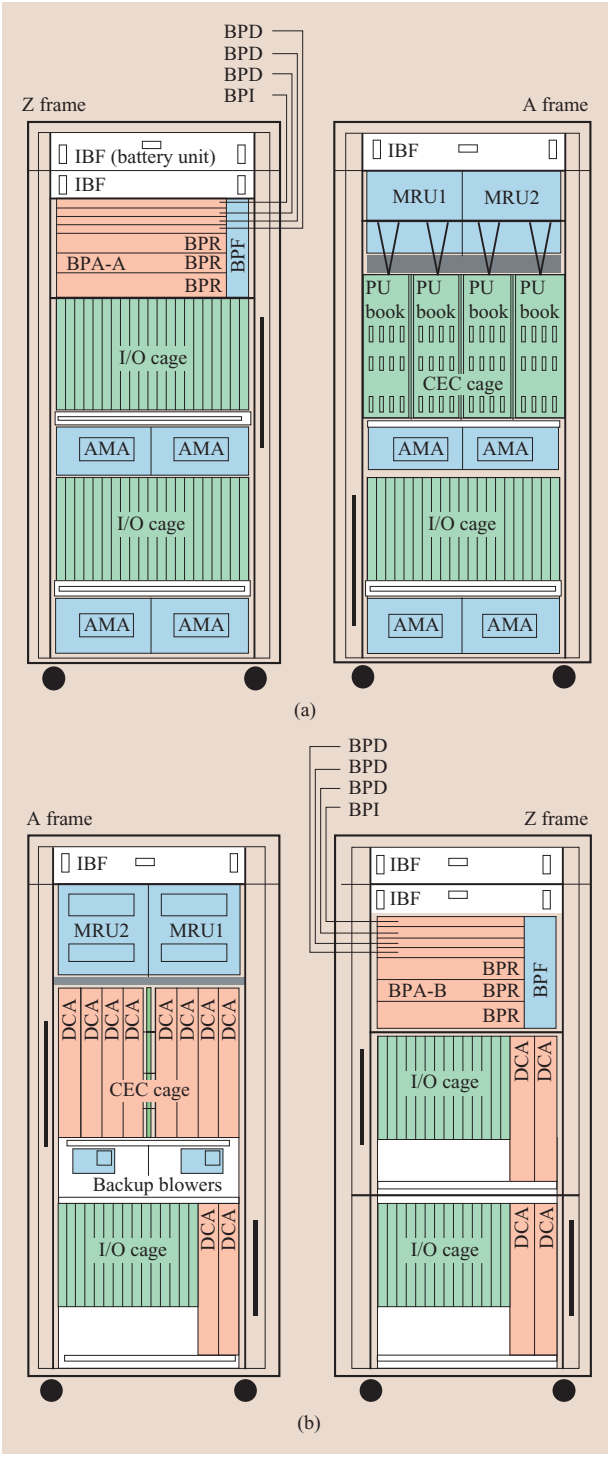

Figure 1

(a) Front view and (b) rear view of IBM z990 server showing the subsystem building blocks: Bulk power assembly (BPA); power distributor (BPD); bulk power fan (BPF); bulk power interface (BPI); bulk power regulator (BPR); integrated battery function (IBF); central electronic complex (CEC) cage; I/O cage; air-moving assemblies (AMAs); and backup blowers in the AMAs.

Table 1 Comparison of I/O count in z990 and z900 eServers.

IBM eServer	*Connector style*	*MCM size* (mm × mm)	*MCM I/O count*	*I/O per mm²*
z900	Zero insertion force pin	127 × 127	4,224	0.26
z990	Land grid array	93 × 93	5,184	0.6

approach yields four times the performance of the z900 in a comparable 24-in.-wide (610-mm-wide) rack. Each book contains an MCM, two memory books, each with up to 64 GB of memory, and up to 12 new high-performance enhanced self-timing interconnects (eSTIs) for communication with peripheral devices. To achieve efficient packaging density and performance, a hybrid refrigeration cooling system was developed that features a specially designed evaporator–finned-heat-sink combination attached to the MCM hat [2]. The PU book is fully surrounded by a plated steel sheet enclosure for effective immunity from electromagnetic interference. The PU book (**Figure 2**) delivers balanced system performance in the new on-demand era of e-business.

The CEC cage provides the physical structure and the interconnections to support up to four PU books at the front portion of the enclosure with their associated DCAs at the rear. The general design and packaging approach was governed by many factors, including functional density, processor timing, system cooling, electromagnetic compatibility, shipping shock and vibration, cost, and ease of assembly and service. To manage these packaging considerations, the CEC enclosure, with its modular design, has been mounted in a standard 24-in.-wide (610-mm-wide) rack. The z990 CEC cage mechanically supports separate modules for the PU and the DCA books and occupies 13 standard Electronic Industries Association (EIA) vertical units. Each EIA unit equals 44.5 mm. The cage also contains air inlet and exhaust ducts. The PU and DCA book modules are merged with a center midplane circuit-board assembly. The center midplane circuit-board assembly consists of a mother board mechanically supported between two aluminum stiffeners that mechanically stabilize the individual modules into a cohesive structure. The design allows interconnection access to both sides of the midplane board, provides manageable modules from a human-factors perspective, and minimizes the time required for attaching components and circuit-board assembly while providing a robust structure for the CEC components. An AMA located under the CEC cage houses the blowers. In contrast to previous air-cooling approaches, the AMA is in a completely separate enclosure occupying four EIA vertical

units. By making the AMA an independent and separate enclosure, it was possible to optimize the design from a cost and weight perspective. Moreover, the separate-AMA-enclosure approach provided future cooling design enhancements without necessarily affecting the basic CEC structure. The separate AMA enclosure, located under the CEC cage, also acts as a structural pedestal that has proved to be effective in helping to mechanically support a fully populated CEC weighing ~250 kg during shipping and earthquake scenarios. The AMA houses two front-located primary blowers to cool the memory and a memory bus adapter within the PU books at the front of the CEC, as well as the DCAs at the rear. Also housed in the rear of the AMA are two backup blowers. Should the modular refrigeration unit (MRU) cooling subsystem fail, the two backup blowers are activated to provide adequate air flow for cooling the processors until the refrigeration system can be serviced.

A fully populated z990 CEC, with its AMA, occupies a total of 17 vertical EIA units (756 mm). The z990 CEC cage supports up to four PU books, eight DCA books, two external time reference (ETR) books, and two oscillator (OSC) books. The pluggable PU books provide protective shells for the printed-circuit cards, as well as mechanical features to guide and dock the Very High Density Metric** (VHDM**) connectors from Teradyne used to connect the books to the midplane board. Docking of the PU book connector system is especially challenging owing to the high contact density, the high number of contacts (1,080 signal contacts with ground shields and 26 power modules), and the relatively high nominal plug force (approximately 11 kN) required to seat the connectors. To accomplish the blind docking of the PU book connector system, the z990 CEC utilizes a series of custom-designed guidance features that work in conjunction with the connector lead-in features. These guidance features, coupled with stiffening features in the rear of the CEC enclosure and on the rear side of the midplane stiffener, provide repeatable connector docking with minimal board deflection. At the heart of the docking system are plastic guide rails with an integral lifting ramp that pre-aligns the PU book with the guide pins on the midplane board. These design elements allow the fine-alignment features of the connector system to engage properly. Final seating is accomplished by actuating the PU book cam latches that provide the required docking insertion force.

Packaging the PU book

In order to fit four PU books into a CEC cage, narrow PU books with low-profile MCM assemblies had to be developed. The low-profile MCM assemblies were achieved by developing a new evaporator with an integrated finned heat sink, an LGA interconnection between the MCM substrate and the printed-circuit mother board, and a low-profile actuation mechanism to mate the MCM and the printed-circuit board.

To avoid moisture condensation on microprocessor chips operating below the dew-point temperature, the MCM has to be housed in a thermally insulated, moisture-impervious enclosure. In the absence of adequate space in the narrow PU book for a thermally insulated and moisture-impervious enclosure, the MCM temperature had to be kept above the dew point to avoid moisture

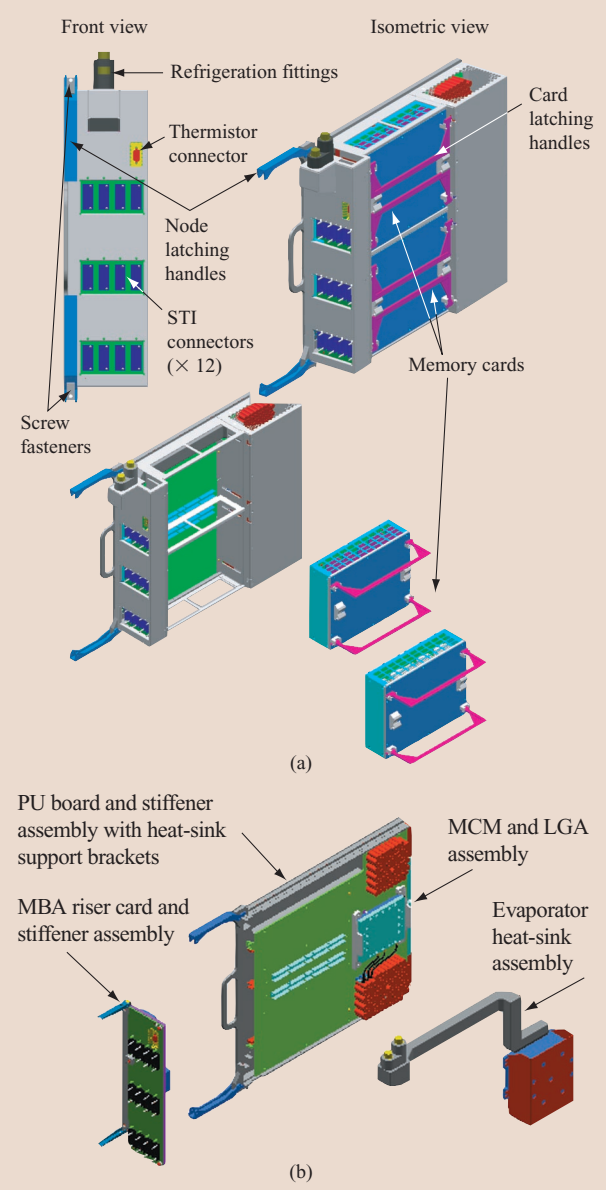

Figure 2

(a) PU book with memory books installed (top) and with memory books removed. (b) Exploded view of the PU book showing the MCM and the evaporator heat-sink assemblies (covers removed).

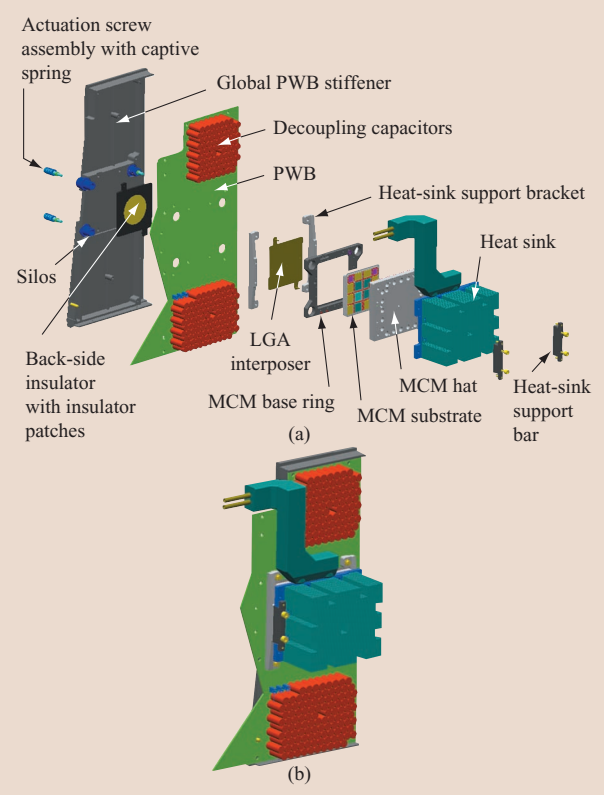

Figure 3

(a) Exploded and (b) assembled view of the MCM-to-PWB connection.

condensation on the microprocessor chips. In addition, since the high packaging density did not allow redundant MRUs, a hybrid cooling approach was utilized in which the primary mode of cooling was by an MRU; in the unlikely event of an MRU failure, a backup blower cooled the MCM via the MCM evaporator with integral finned heat sink.

To achieve a low-profile MCM hardware, an actuation mechanism was developed to apply uniform mechanical loading to the MCM–LGA interconnect, and a new mechanical heat-sink support was developed to isolate the dynamic effects of the heavy copper heat sink from the LGA interconnect contact surfaces.

Land grid array (LGA) connector

The z990 eServer represents the first zSeries* application of LGA technology to connect an MCM to a printed wiring board (PWB). As noted in Table 1, the z990 MCM represents an overall I/O density increase of ~230% over the z900, achieved by utilizing a modified version of the IBM p690 LGA interposer [3, 4]. A four-point, on-corner actuation design approach was used which specifically allotted 25 mm of back-side printed-circuit-board (PCB) distance for the MCM LGA actuation hardware. The final hardware design utilized a four-coil-spring approach in which a significant portion of the required actuation hardware was housed within the LGA assembly, thereby minimizing space and eliminating any adverse impacts to other entities such as the evaporator assembly [5].

The actuation components include an MCM with an LGA interposer, a global PWB stiffener with threaded silos, a local LGA insulator with customized contact patches, and four actuation screw assemblies with captive compression springs. These components are illustrated in exploded and assembled views of the MCM-to-PWB connection in **Figure 3**. In addition, the MCM contains a base ring with clearance holes to allow the external diameters of the silos to protrude into the MCM base ring, as well as a customizable shim between the base ring and hat to minimize variations in the actuation force caused by variations in the substrate thickness. The MCM is mounted on the PWB by placing the MCM on the PWB, applying a pre-load across the MCM and PWB global stiffener, and torquing the four actuation screws into threaded holes in the MCM hat.

LGA contact load control

Evaluations of the LGA actuation systems require both analytical and empirical assessments to ensure that adequate contact load control is established and maintained [4, 6]. The analytical assessment consists of two evaluations: The first is an analysis of the variation in the bulk load developed by the actuation load delivery system; the second is the response of the total LGA system to the bulk load developed, including component tolerances that affect contact load variation within the array. The bulk load variation analysis begins with a definition of the nominal and three standard deviation (3σ) contact loading range that is attained given the specified actuation distance and a statistical assessment of the component dimensional tolerances. On the basis of experience, a 90-g-per-contact nominal loading was selected for the technology class of LGA used in this application. A 3σ tolerance assessment gave an average load variation (total bulk load variation divided by the number of contacts in the array) of ± 20 g per contact. The load variation range was predominantly governed by the specified tolerance for the high-spring-rate coil spring.

A finite element analysis was completed to quantify the structural deflection characteristics of the actuated assembly. A four-point actuation system has a tendency to create deformations in the MCM structure and in the PWB that can exacerbate the gap variation between the two, increasing load variations across the LGA array. The structural deformation under load, along with other key component variations, was evaluated using a Monte Carlo

technique [4] to estimate the contact load variation within the array. Key component variables included the MCM substrate camber, PWB thickness variation, applied bulk load variation, PWB stiffener and insulator camber and thickness variation, and individual LGA contact-free height and stiffness characteristics. **Figure 4(a)** represents the resulting cumulative LGA contact load variations across the MCM diagonal.

Additionally, an individual factor sensitivity analysis identified the three main factors contributing to contact load variation to be substrate camber, PWB thickness variation, and structural deflection of the actuated assembly. The contact load variation due to structural deformation is a systematic response that, for a given mechanical configuration, depends only upon the bulk load being delivered by the LGA actuation system. Hence, a contour plot of the minimum contact load (3σ band) as a function of the two primary statistically varying parameters, module camber and PWB thickness variation, was created as shown in **Figure 4(b)** to demonstrate the performance tradeoffs of these variables. Given the reliability requirement of a minimum load of 20–30 g per contact and PWB thickness variations of ± 0.05 mm, the analyses concluded that the substrate camber must have a 15–75-μm convex shape. The convex substrate camber is difficult to maintain and has the potential to produce a severe impact on substrate yield.

A study using pressure-sensitive film was done to validate the analytical work. The contact fading from two substrates, one with 39-μm concave shape and the other with 53-μm convex shape, agreed with the Monte Carlo analyses.

The necessity of a dual-sided (i.e., convex or concave) substrate camber specification required improvement in the actuation hardware. Circular concentric insulating discs were stacked and adhered to the back-side local insulator [7]. The effect of adding the insulating discs was to compensate for the contact load variation resulting from structural deformation, essentially providing a local pre-convex shape to the PWB stiffener. This pre-convex effect essentially centers the contact load contour plot with respect to the substrate camber, and hence allows both convex and concave substrate camber without deviating from the required minimum contact load. The effectiveness of the insulating discs was verified via Monte Carlo analysis and pressure-sensitive film studies. A Monte Carlo analysis assessed the statistical impact of this design modification, as shown in **Figure 5(a)**. From **Figure 5(b)**, it is evident that the insulating discs shifted the acceptable contact load to a substrate camber region including both convex and concave shapes. With the attachment of the insulating discs to the back-side local insulator, the reliability qualification activities were successfully completed, and a specification allowing the use of both

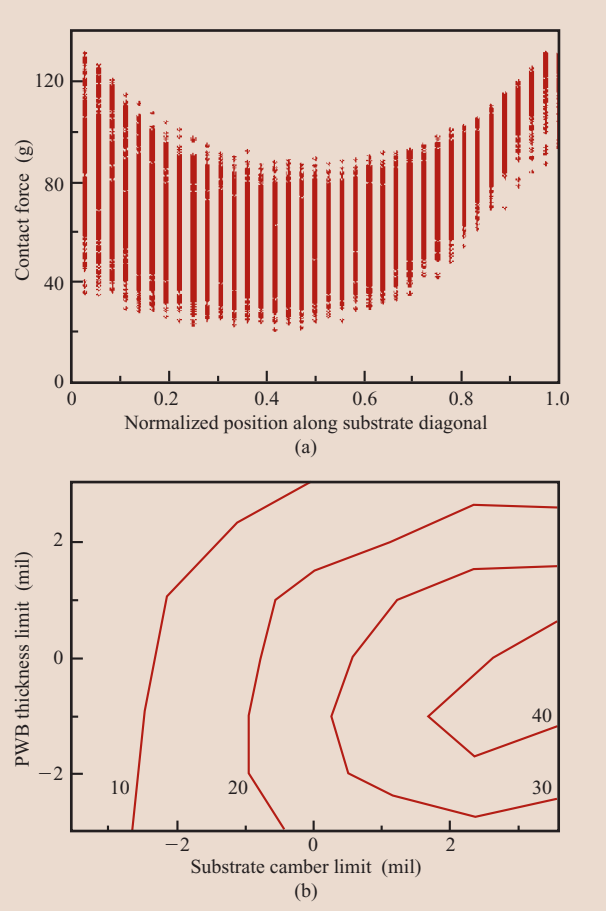

Figure 4

(a) Contact force variation along LGA quadrant diagonal when no insulating discs are attached to the back-side local insulator. (b) Contour plot of the contact force (g) as a function of the substrate camber and the card thickness when insulator discs are attached to the back-side local insulator (1 mil = 25.4 μm).

convex and concave 50-μm substrates was implemented which effectively minimized the MCM yield issues.

Isolation of the MCM heat-sink mass

In the z990 hybrid cooling scheme, refrigeration cooling is accomplished by the evaporator copper plate bolted to the MCM hat. The backup cooling, in case of refrigeration cooling failure, is accomplished by air flowing through gaps in copper fins soldered to the evaporator plate. If left uncorrected, the mass of the copper evaporator with its soldered copper fins has the potential of inducing fretting corrosion of the LGA contacts due to mechanical vibrations during shipping. One potential solution would have been the application of an additional MCM pre-load to minimize the sprung-mass effects of the MCM assembly. Unfortunately, given the high magnitude of

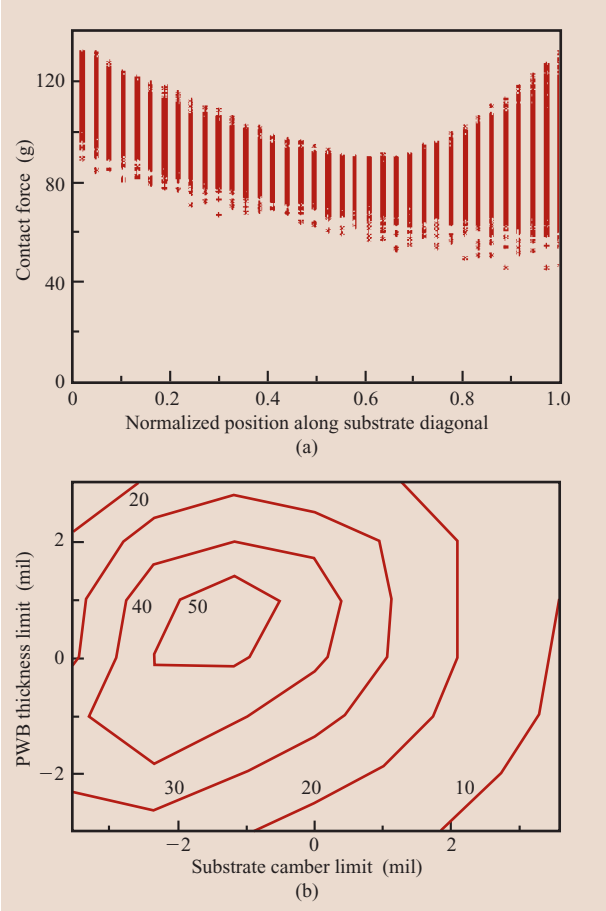

Figure 5

(a) Monte Carlo analysis of the LGA contact forces with insulator discs attached to the back-side local insulator. (b) Contour plot of the contact force (g) as a function of the substrate camber and the PWB thickness with insulator discs attached to the back-side local insulator (1 mil = 25.4 μm).

the mass involved, the required increase in actuation force would have mechanically overstressed the MCM substrate and the PWB. Thus, the challenge was to develop an approach that did not stress the MCM hardware and yet restricted the sprung mass of the evaporator heat sink from lateral motion parallel to the PWB plane. The approach taken was to use a pair of aluminum heat-sink support brackets and bars straddling the MCM, as shown in Figure 3. The restriction of lateral motion was accomplished by pins in the bars engaging holes in the evaporator heat-sink base. The following assembly steps help explain the mass-isolation scheme: The heat-sink support brackets are bolted onto the PWB stiffener. The evaporator heat sink is then loosely attached to the MCM hat. The support bars are placed over the brackets with the bar pins inserted in holes in the evaporator heat-sink base. The bars are then bolted tightly to the brackets. Finally, the evaporator heat sink, now free to move only in the direction vertical to the PWB plane, is bolted to the MCM hat. The bar and bracket arrangement does not impart the mechanical load perpendicular to the LGA contact, and thus does not restrict the LGA actuation springs from maintaining the desired LGA contact force during thermal excursions or normal operation [8].

Powering the CEC cage

The z990 power subsystem evolution is now in its seventh generation. The subsystem architecture, introduced in 1995, is based on redundant bulk power regulators (BPRs) that convert utility power into regulated 350 Vdc, which is distributed via a BPD or a BPI to the cooling units and point-of-load power supplies (DCAs) throughout the system. The high-voltage (350-Vdc) bus was selected to significantly reduce source currents compared with the more widespread 48-V bus voltage which is a prerequisite for achieving the high functional density of the zSeries eServers. High-efficiency, high-capacity redundant blowers reduce the number of required blowers. Point-of-load power supplies (DCAs) are plugged directly into the functional midplane. This simple but flexible architecture has permitted the total integration of the refrigeration units that cool the MCMs. The resulting 40° to 60°C cooler MCM operation results in unprecedented performance and reliability. The z990 eServer is shown in simplified form in Figure 1, where power elements are indicated in red, cooling elements in blue, and eServer functions in green. **Figure 6** shows a simplified block diagram of the interconnection of these units within the z990 server. Active redundancy is employed for power supplies and air-moving devices [9].

The z990 introduced two new power requirements. First, the increased function of the PU books in the CEC and the introduction of higher-performance I/O adapters increased the z990 eServer total system power for a maximum configuration by ~33% over that of the z900 eServer. Second, the combination of increasing the number of MCMs in the system from one to four and lowering the processor voltage from 1.5 V to 1.2 V tripled the required current delivery to the CEC cage.

Bulk power assembly

The BPA for the z900 eServer was designed with extendability in mind with regard to both power capacity and number of switched 350-Vdc ports. This was possible because the BPA is designed around a fixed dc bus which does not vary from one eServer generation to another. Since the capacity of the BPA from the z900 was

Table 2 Comparison of DCA counts for Alternatives 2 and 3. The DCA current output remained unchanged at ~1000 A.

No. of PU books installed in the CEC	*No. of DCAs powering the CEC*	
	Alt. 2	*Alt. 3*
1	2	2
2	3	4
3	4	6
4	5	8

adequate, the design changes made to the BPA were confined to redefinition of switched ports to accommodate the unique arrangement of power and cooling units within the z990 eServer and increasing the current capacity of certain branch protection circuits. The BPR units, which convert utility voltage to the common 350-Vdc internal power bus, were unchanged.

Processor power

The maximum total current supplied to the z900 CEC is 1200 A. For the z990 multi-PU-book CEC structure, the current has been increased to 3400 A. This near-tripling of the required current presented two challenges. First, if the z900 model for plugging the DCAs into the midplane had been adopted for the z990, the midplane copper planes would have to handle three times more current. Second, the PU book structure of the z990 eServer requires the current to pass through an additional connector to supply the MCM within the PU book. In contrast to the z900, which has its sole MCM plugged directly into the midplane board, the z990 has its MCMs housed in PU books that are plugged into the midplane board. Three alternatives were considered to handle the tripling of current in the z990 eServer:

1. The DCAs embedded within the PU book according to the common industry practice.
2. A single set of $n + 1$ DCAs plugged into the midplane board to supply the entire CEC, as in the z900 eServer, where n is the minimum number required to power the fully populated CEC cage.
3. A unique set of $n + 1$ DCAs dedicated to each PU book, where n is the minimum number required to power each of them.

The first alternative would overcome both of the challenges described above: A high-voltage, low-current source would supply the PU book, and high currents at tightly regulated low voltages would be generated within the PU book. However, this solution was discarded for

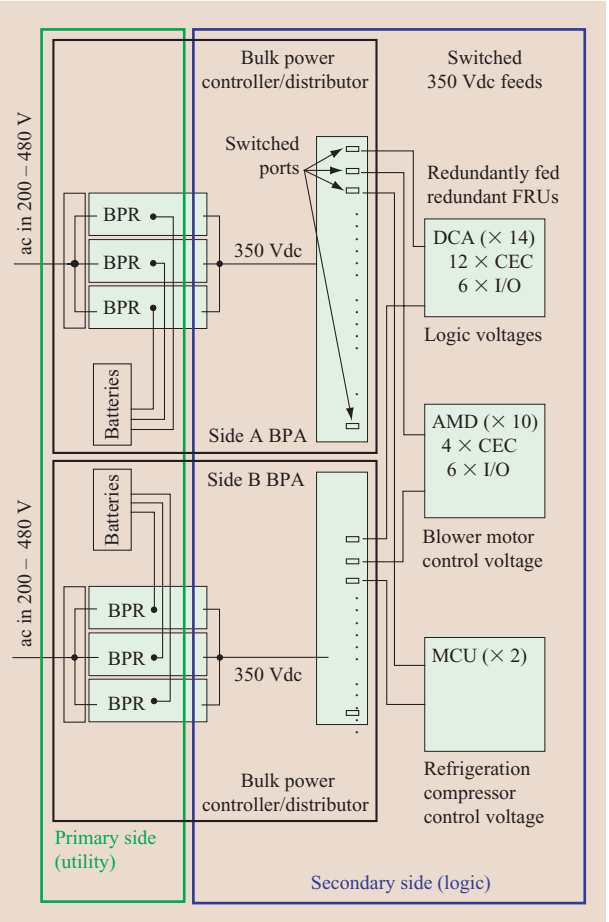

Figure 6

Simplified power block diagram for the z990 server.

two reasons. The power supplies (DCAs) would have to be redundant, so that a malfunction would not disrupt the PU book operation, and the space required for two DCAs in a PU book would have been excessive. Also, even if the redundant DCAs could be packaged within the PU book, a power supply could not be replaced concurrently with PU book operation. The second alternative is the most economical from the standpoint of DCA utilization. **Table 2** compares Alternatives 2 and 3 above with regard to DCA counts.

While Alternative 2 requires fewer power supplies for all configurations containing more than a single PU book, it has three problems. The midplane board would have a single-processor voltage boundary, requiring its cross section to accommodate the current requirement for the entire CEC cage. Also, with a single boundary, faults in any PU book could potentially disrupt the function of all PU books in the CEC cage. Finally, individual PU books

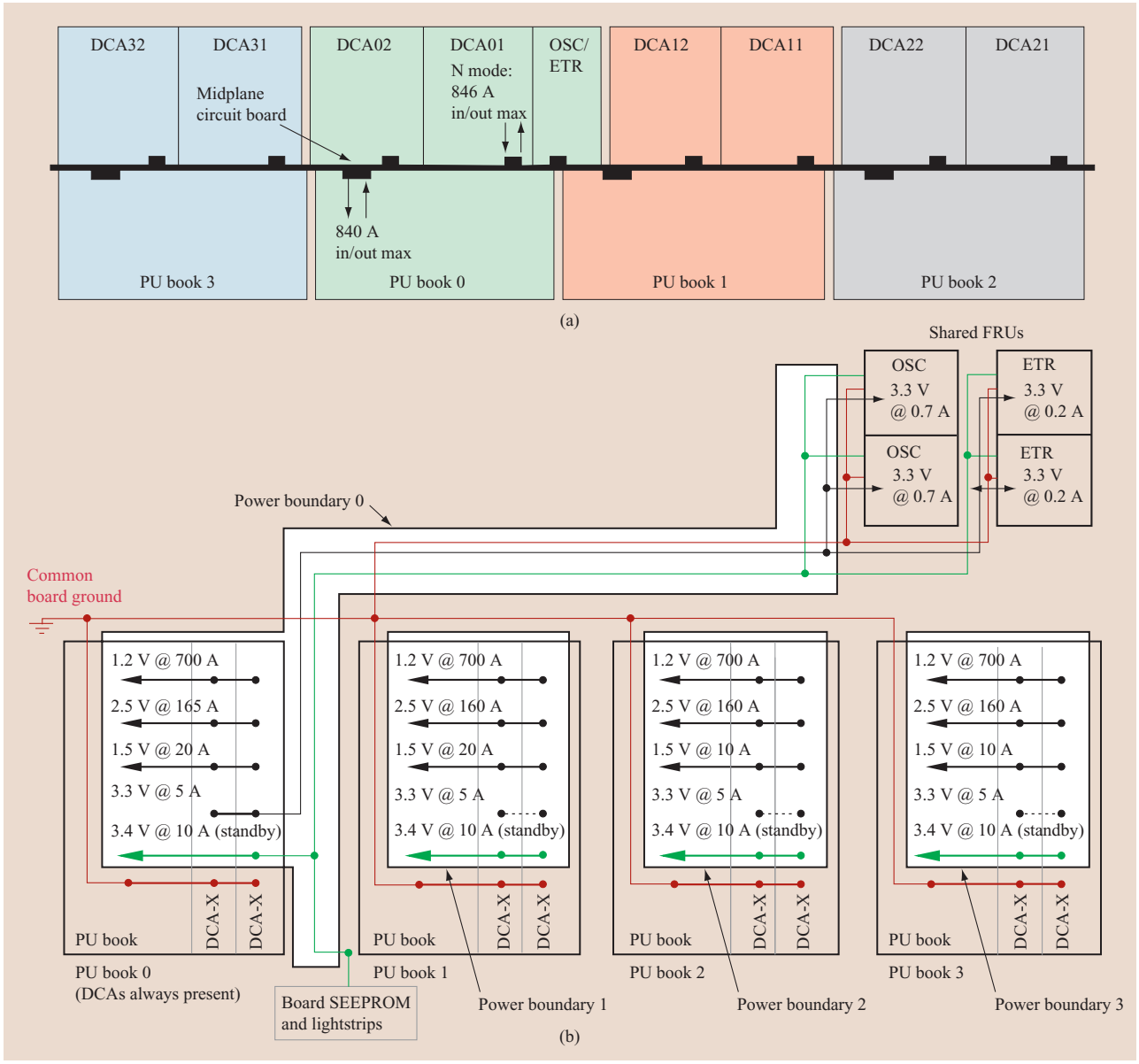

Figure 7

(a) Top view of the CEC cage, showing physical relationships between the PU books and their respective DCAs for Alternative 3 from Table 2. (b) Power distribution on z990 CEC midplane.

must be powered on separately. Alternative 3, shown in **Figure 7(a)**, was adopted in the final design.

Individual power boundaries for each PU book require that the power planes within the midplane board be split into four sections for each required voltage level. However, it was necessary to maintain the ground return planes continuously throughout the midplane so that the high-speed signals traveling between PU books would have continuous high-frequency return paths [10]. This segmenting of the power in the midplane concentrated current flow to such an extent that the required midplane current capacity is no greater than that of the z900 eServer. The power distribution on the midplane is depicted in **Figure 7(b)**. Note that the functions common to all PU books (system oscillator and external time reference) are powered from the PU book 0 power boundary, which is always populated with DCAs. The physical relationship between the PU books and their

respective DCAs is shown in Figure 7(a) in a top view of the CEC cage. It is evident that the shortest possible current paths are achieved while maintaining independent power-on and power-off capability for each PU book, PU-book-to-PU-book fault isolation, and DCA maintenance concurrent with full-system operation.

While the above arrangement of PU books and DCAs solved the problem of midplane board current capacity, connectors of sufficient capacity were still required to connect the DCAs and PU books to the midplane board.

DCA connector

A DCA connector is required to supply ~850 A in and out of the midplane board. Because the DCAs also contain the processors which control the power, service, initialization, and vital product data functions of the CEC cage, the connector must contain enough signal I/Os to provide high-speed serial interfaces to the PU books and (in the case of the PU book 0 position), to the oscillator and the external time reference functions as well. Eighty signal I/Os are required to provide these interfaces with sufficient grounds for signal integrity, as well as the voltage and thermal sense lines required for protection and voltage regulation. This combination of power delivery and signal I/Os has been a requirement of the DCAs since the inception of the zSeries eServers. The Winchester HD+** series of connectors has met these requirements from the beginning, with evolutionary enhancements introduced periodically to increase the linear current density. The capacities of the connector used on the z990 DCA are 7 A/mm and 1.6 I/Os per mm, which permit a connector length of 360 mm including charge blades and guide modules at each end. A simplified block diagram of the DCA connector is shown in **Figure 8**.

PU book connector

The PU book connector is also required to supply ~850 A in and out of the PU book; however, in this case the connector must also provide for 1,080 I/Os, most of which are extremely high-speed signals. The ring connecting the PU books runs at 600 Mb/s, and all of the ring signals pass through the PU book connector. The VHDM series of connectors, initially utilized in the z900 in a 6-pin-wide configuration, was considered for this application. A total connector length of 465 mm could be accommodated, limited by the midplane board dimensions permitted on the large panel board manufacturing line. Using the 8-pin-wide configuration of the VHDM connector modules, the required I/O count used 296 mm of the available length, including the guide modules at each end. This left 169 mm for power delivery. The linear power capacity of the 8-row

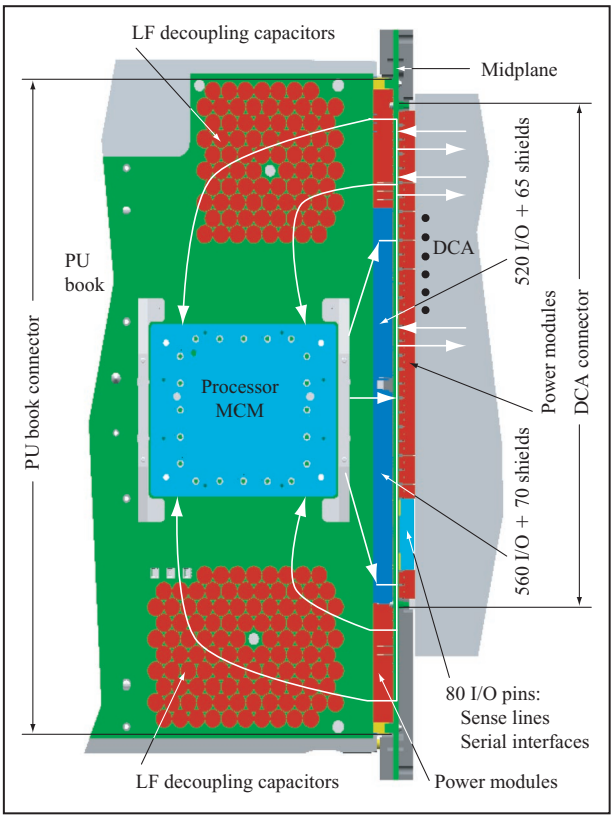

Figure 8

z990 PU book and DCA connections to the midplane board. The white arrows illustrate current flow.

power module is 5 A/mm, allowing 845 A capacity in the available length. Since this did not allow for return current, the signal shields between each row of I/O pins had to carry all of the return current. The number of signal return shields available for return current is 1,080/8 = 135. Teradyne rated the capacity at 3 A per shield, giving a total return current of 405 A, which was not nearly sufficient. Initial testing of this shield using a representative number of midplane board layers demonstrated a much higher current capacity for the shield than the Teradyne rating. Subsequent testing performed at IBM facilities using Teradyne-built test fixtures established a capacity of at least 7 A/mm in the z990 application. The current enters the top and bottom of the connector in an approximately symmetrical pattern and exits the center as shown in Figure 8. Although the current paths in and out of the connector itself are not interlaced or physically adjacent, the overall loop inductance is very small because of the short current paths and the tight coupling of the board power planes. The PU book connector was established in this manner.

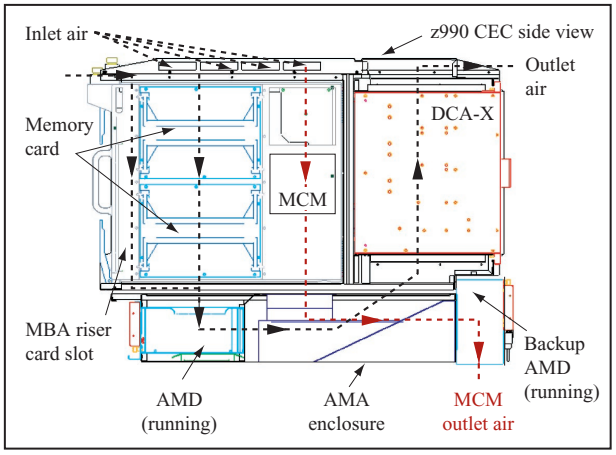

Figure 9

Cooling the CEC. Black dashed line only: air cooling when refrigeration cooling is operational. Black dashed line + red dashed line: air cooling when refrigeration cooling is not operational.

Voltage regulation and low-frequency decoupling

In order to support the extremely high signal propagation rates within and between the z990 PU books, a tight voltage regulation tolerance must be achieved to account for both static and dynamic effects. In order to achieve the smallest possible voltage deviations, remote sensing is used on all voltage levels. The DCA provides a static voltage regulation tolerance of ±0.5% at the sense point, including set point accuracy, end-of-life drift, and static load variation from no load to maximum load. The sense points are chosen to be as close as possible to the center of the electrical load. In the case of the processor level, the sense point was actually positioned within the voltage planes of the processor module in order to eliminate any variations in the LGA connector contacts. All other levels are sensed on printed-circuit-board planes.

Low-frequency decoupling provides voltage stabilization in response to step load changes, which are of two types. First, a point-of-load step change can be caused by operating conditions with the processor or memory. This load change is unpredictable, but experience shows that it has an upper limit of ~30% of maximum load. Second, a source step load change is caused by a DCA failure, forcing the remaining DCAs to adapt to the new load condition. This load change is simply $1/(n + 1)$ times the maximum load, where n is the minimum number of DCAs required to support the load. When this load change is considered, a conservative estimate of the required number of low-frequency decoupling capacitors is obtained. For z990, $n + 1 = 2$, so the source step load change is 50% of maximum load for each level. The target voltage deviation at this load change is 3% of nominal.

The number of decoupling capacitors is calculated as follows:

$$\Delta V/\Delta I = (R/n),$$

where R is the equivalent series resistance (ESR) of each capacitor and n is the number of capacitors. The capacitor selected is a computer-grade aluminum electrolytic with ESR = 0.025 Ω and C = 4.7 mF. Approximately 190 capacitors are required. These are placed in the open area above and below the processor module, as shown in Figure 8. Measurements made on a running PU book showed the voltages to be well within the specified range. Reference [10] discusses dc losses from the sense point to the circuits and mid- and high-frequency decoupling.

Cooling the CEC cage

The cooling architecture of zSeries eServers provides independent cooling for each I/O and CEC cage. Cooling for the I/O cage and the bulk power assembly (BPA) is basically unchanged from the z900 eServer. The z990 CEC, like the four previous generations of zSeries eServers, is cooled by a combination of forced air and refrigeration. The MCM in each PU book is cooled by refrigeration under normal operation. The rest of the CEC hardware (memory, I/O drivers, DCAs, oscillator, and ETR) is air-cooled. The design point established for the CEC, prior to development of the z990, was ~700 W removed by refrigeration and 1,150 W removed by forced air. For the z990, the necessity to provide refrigeration cooling for up to four MCMs compared with only one in all prior-generation machines required a change in the basic structure from fully redundant refrigeration to an air-cooling backup system (i.e., the failure of refrigeration cooling results in a backup mode involving forced-air cooling and slowing of the processor clock speed). The method by which power is controlled and system operation is maintained in backup mode is referred to as "cycle steering" [2]. Several unique design features were required to support this type of redundancy:

1. The air-moving assembly (AMA) located below the CEC in the z990, instead of being a simple plenum, was modified into a complex of isolated tunnels to provide two independent air-flow paths.
2. Two backup blowers were added to the rear of the AMA to turn on and draw air through the MCM finned heat sink in the event of refrigeration failure. It should be noted that the AMA also contains the primary blowers that air-cool all of the components in the CEC cage, with the exception of the MCM, at all times.

The cooling scheme during normal operation is illustrated in **Figure 9**. The view is from the side of the CEC, with the air inlet on the left and the outlet on the

right. The MCM in this case is cooled by the modular cooling unit (MCU) and the evaporator bolted onto the MCM hat [2]. There is no air flow through the MCM finned heat sink because the backup blowers are not running and the chimney structure in which the MCM is mounted is sealed by a hinged door at the bottom. Air flow through the MCM finned heat sink is blocked because warm room air would actually inhibit the cooling effect of refrigeration by adding to the heat load. The primary air-moving devices (AMDs) are running and drawing air down through the memory and the I/O driver modules in the PU books and then pushing the air up through the DCAs to the outlet at the top rear of the cage. The blower speed is dependent on the number of installed PU books. In the event of an MCU failure, the MCM temperature, which is redundantly sensed by a thermistor probe screwed into the MCM hat, begins to rise. "Cycle steering" is invoked, and the backup AMDs are turned on. The back-pressure within the AMA tunnel forces the hinged door at the bottom of the MCM cavity to open, and forced air is drawn through the MCM finned heat sink. As shown in Figure 9, the air flow through the MCM finned heat sink is completely separate from the cooling air for the rest of the PU books. The primary AMDs continue to run at normal speed. When the defective MRU is replaced, the refrigeration cooling is restored, the backup blowers are turned off, and "cycle steering" automatically returns the system to normal-mode operation. In this manner, the processor chips are maintained at low junction temperature and full performance is maintained for virtually 100% of the operating life of the PU book. The system performs at a slightly degraded level for a short period in the unlikely event of an MCU failure.

Summary

The z990 eServer central electronic complex houses four multichip-module (MCM) -based processor units compared with one in the previous-generation z900 eServer. The processor unit input/output pin density in the z990 MCM is more than twice that of the z900 MCM. This increase in packaging density and the consequent tripling of the current drawn by the processor units have been achieved by significant changes in the z990 CEC hardware.

The CEC structure was changed to include a distinct set of DCAs for each PU book. The new design enables the midplane board to handle the increased current flow while allowing independent power on/off capability for each PU book, PU-book-to-PU-book fault isolation, and DCA maintenance concurrent with full-system operation.

The structure of the CEC and AMA was changed from the fully redundant refrigeration cooling of the z900 MCM to z990 hybrid MCM cooling, in which the primary cooling mode is by refrigeration and the backup cooling mode is by high-pressure air flow through copper fins soldered to the MCM evaporator. The MCM temperature is kept above the dew point to eliminate the space that would have been taken up by the insulators necessary to mitigate moisture condensation.

The need for low-profile MCM hardware led to the use of the land grid array MCM-to-board interconnection system. A unique design feature involving contact pads was developed to obtain uniform LGA contact force across the whole LGA area. The design allowed the use of substrates with both convex and concave camber.

The copper evaporator and copper-finned heat sink required for the MCM hybrid cooling increased the mass overhanging the PU circuit board stiffener enough to cause potential contact instabilities and fretting corrosion of the LGA contacts if left uncorrected. A novel isolation scheme was developed to isolate the sprung mass from causing micro motion at the LGA contacts by restricting the sprung-mass motion in the plane parallel to the PU circuit board.

Fitting more CMOS microprocessors into a condensed volume provides considerable performance advantages. However, this technique demands solutions to deal with ever-increasing I/O density, required electric current, and the associated heat dissipation. The ingenuity of power, packaging, and cooling engineers in designing, developing, and manufacturing hardware with decreasing cost and ever-increasing performance, availability, quality, and reliability will continue to be challenged in the foreseeable future.

References

1. P. Singh, S. J. Ahladas, W. D. Becker, F. E. Bosco, J. P. Corrado, G. F. Goth, S. Iruvanti, M. A. Nobile, B. D. Notohardjono, J. H. Quick, E. J. Seminaro, K. M. Soohoo, and C. Wu, "A Power, Packaging, and Cooling Overview of the IBM eServer z900," *IBM J. Res. & Dev.* **46,** No. 6, 711–738 (November 2002).
2. G. F. Goth, D. J. Kearney, U. Meyer, and D. W. Porter, "Hybrid Cooling with Cycle Steering in the IBM eServer z990," *IBM J. Res. & Dev.* **48,** No. 3/4, 409–423 (May/July 2004, this issue).
3. J. U. Knickerbocker, F. L. Pompeo, A. F. Tai, D. L. Thomas, R. D. Weekly, M. G. Nealon, H. C. Hamel, A. Haridass, J. N. Humenik, S. N. Reddy, R. A. Shelleman, K. M. Prettyman, B. V. Fasano, S. K. Ray, T. E. Lombardi, K. C. Marston, P. A. Coico, P. J. Brofman, L. S. Goldmann, D. L. Edwards, J. A. Zitz, S. Iruvanti, S. L. Shinde, and H. P. Longworth, "An Advanced Multichip Module (MCM) for High-Performance UNIX

Servers," *IBM J. Res. & Dev.* **46,** No. 6, 779–804 (November 2002).
4. J. S. Corbin, C. N. Ramirez, and D. E. Massey, "Land Grid Array Sockets for Server Applications," *IBM J. Res. & Dev.* **46,** No. 6, 763–778 (November 2002).
5. J. G. Torok, G. F. Goth, K. D. Waddell, and J. J. Loparco, "Compression Connector Actuation System," U.S. Patent 6,845,311, November 26, 2002.
6. Mark K. Hoffmeyer, John L. Colbert, John G. Torok, John S. Corbin, and William L. Brodsky, "System Packaging for Robust LGA Interconnect Technology in High Performance Computing Applications," presented at the International Symposium on Microelectronics (IMAPS), Boston, November 16–20, 2003.
7. J. G. Torok, B. D. Notohardjono, J. J. Loparco, W. P. Kostenko, and J. S. Corbin, "Method and Apparatus for Providing Positive Contact Force in an Electrical Assembly," U.S. Patent Application POU920030012US1, May 12, 2003, patent pending.
8. J. G. Torok, B. D. Notohardjono, G. F. Goth, J. A. Hickey, and J. J. Loparco, "Support Means for Isolating Thermal Device Mass Effects on LGA Interconnections," *IBM Tech. Disclosure Bull.* (December 2002); published on *IP.com* June 20, 2003.
9. 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, and F. Freier, "RAS Design for the IBM eServer z900," *IBM J. Res. & Dev.* **46,** No. 4/5, 503–521 (July/September 2002).
10. T.-M. Winkel, W. D. Becker, H. Harrer, H. Pross, D. Kaller, B. Garben, B. J. Chamberlin, and S. A. Kuppinger, "First- and Second-Level Packaging of the z990 Processor Cage," *IBM J. Res. & Dev.* **48,** No. 3/4, 379–394 (May/July 2004, this issue).

Received November 1, 2003; accepted for publication April 15, 2004; Internet publication May 27, 2004

Juan C. Parrilla *IBM Systems and Technology Group, 2455 South Road, Poughkeepsie, New York 12601 (parrilla@us.ibm.com).* Mr. Parrilla is an Advisory Engineer in the Product, Power, Packaging, and Cooling Architecture and Business Management Department. He received a B.S. degree in mechanical engineering from the Massachusetts Institute of Technology in 1982. That same year he joined the IBM Systems Development Division in Endicott, New York, helping to design the 9309 monocoque frame which was used on S/390, AS/400, and RS/6000 products. Mr. Parrilla joined the Power, Packaging, and Cooling Group in Poughkeepsie, where he has served as project leader on various bipolar and CMOS zSeries products from the 9121 and 9672 to the z900. He currently serves as project leader on the zSeries z990 and pSeries 7039.

Frank E. Bosco *IBM Systems and Technology Group, 2455 South Road, Poughkeepsie, New York 12603 (bosco@us.ibm.com).* Mr. Bosco received a B.S.E.E. degree from Manhattan College in 1964 and an M.S.E.E. degree from Syracuse University in 1975. In 1964 he joined the IBM Systems Development Division in Poughkeepsie, where he worked on the early development of monolithic integrated circuits. In 1973, he worked on a team which was attempting to implement a full wafer memory package; following this, he worked on circuit designs for the first implementations of cryptographic algorithms. He worked in power-supply development from 1974 to 1978. From 1979 to 1981, Mr. Bosco worked on the hardware and software algorithms for automating personal identity verification. He joined the memory area in Kingston in 1981 and managed the group which developed the first L4 storage subsystem. In 1985 he joined the Corporate Development staff, where he led a task force which resulted in the introduction of built-in self-test (BIST) into CMOS logic arrays. Mr. Bosco rejoined the Kingston power group in 1988; since 1993, he has been involved in the development of the power, packaging, and cooling subsystems for zSeries CMOS mainframes. He is now a power, packaging, and cooling subsystem architect for future pSeries and zSeries eServers.

John S. Corbin *IBM Systems and Technology Group, 11400 Burnet Road, Austin, Texas 78759 (jcorbin@us.ibm.com).* Mr. Corbin received his B.S. and M.S. degrees in mechanical engineering from the University of Texas at Austin in 1972 and 1974, respectively. He joined IBM in 1974 and spent ten years in printer development, working primarily in the area of motion control. This was followed by six years in the System Technology Division Packaging Laboratory in Austin, applying finite element techniques in the area of mechanical-packaging reliability. He is currently a Senior Engineer in the eServer Group in Austin, Texas, where he has spent ten years in AIX workstation and eServer mechanical packaging development. Mr. Corbin is a Registered Professional Engineer in the state of Texas.

John J. Loparco *IBM Systems and Technology Group, 2455 South Road, Poughkeepsie, New York 12601 (loparco@us.ibm.com).* Mr. Loparco is a Senior Designer Specialist in the High End Mechanical Design and Integration Department in the IBM eServer Group. He received his

associate degree from Mohawk Valley Community College in 1981, joining IBM that same year. Since then, he has been involved in high-end large-system development. Mr. Loparco holds nine U.S. patents and has received an IBM Outstanding Technical Achievement Award for his mechanical development leadership role on the S/390 G5 product.

Prabjit Singh *IBM Systems and Technology Group, 2455 South Road, Poughkeepsie, New York 12601 (pjsingh@us.ibm.com).* Dr. Singh is a Senior Engineer in the Materials and Processes Engineering Department in the IBM eServer Group, Poughkeepsie, New York. He received the B.Tech. (Honors) degree in metallurgical engineering from the Indian Institute of Technology, Kharagpur; the M.S. degree in microelectronic manufacturing from the Rensselaer Polytechnic Institute, Troy, New York; and the M.S. and Ph.D. degrees in metallurgy from the Stevens Institute of Technology, Hoboken, New Jersey. He has received an IBM Outstanding Technical Achievement Award for his contributions to the first IBM multichip refrigeration unit, more than 15 patents, nine IBM Invention Achievement Plateau awards, and four IBM Publication Achievement Awards. He is a recipient of the 2003 IBM Poughkeepsie Master Inventor Award. Dr. Singh is an Adjunct Professor of Electrical Engineering at the State University of New York, New Paltz. He is a member of the industry advisory board of the Department of Electrical Engineering and Computer Science, University of Illinois, Chicago, and the past chairman of the Electronic Materials Division of the ASM International.

John G. Torok *IBM Systems and Technology Group, 2455 South Road, Poughkeepsie, New York 12601 (jtorok@us.ibm.com).* Mr. Torok is a Senior Engineer in the Product, Power, Packaging, and Cooling Development Group. He graduated from the State University of New York in 1985 with a B.S. degree in applied physics. He is currently pursuing an M.S. degree in mechanical engineering at the National Technological University. Prior to joining IBM, he worked for 17 years in a variety of engineering capacities for Hubbell, Inc., developing connectors and vibration-suppression devices for the power utility industry. Since joining IBM in 1998, he has provided packaging integration development leadership for the 2029 DWDM Fiber Saver, the ET4 Emulator product, and the z990 series eServer. Recently, he has worked on the development of a variety of mechanical actuation systems for the land grid array attachment of multichip modules. He holds four patents, has achieved the second IBM invention plateau, and has received an IEEE Standards Development Award. Mr. Torok is a member of ASME.