A power, packaging, and cooling overview of the IBM eServer z900

by 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
C. Wu

This paper provides an overview of the power, packaging, and cooling aspects of the IBM eServer z900 design. The semiconductor processor chips must be supported and protected in a mechanical structure that has to provide electrical interconnects while maintaining the chip junction temperature within specified limits. The mechanical structure should be able to withstand shock and vibrations during transportation or events such as earthquakes. The processor chips require electrical power at well-regulated voltages, unaffected by the ac-line voltage and load current fluctuations. The acoustical and electromagnetic noise produced by the hardware must be within the limits set by national regulatory agencies, and the electronic operations must be adequately protected from disruption caused by electromagnetic radiation. For high availability, the power, packaging, and cooling hardware must have redundancy and the ability to be maintained while the system is operating. This paper first overviews the packaging hardware, followed by a description of the first- and second-level packaging, which includes the mother board and the multichip module. Thermal management is discussed from the point of view of both the multichip module and the overall system. Power conversion, management, and distribution are presented next. Finally, the design aspects involved with meeting the requirements of electromagnetic compatibility, acoustics, and immunity to shock, vibration, and earthquakes are discussed.

Introduction

The disciplines of power, packaging, and cooling are essential to the design and manufacture of a commercial computer. The conversion of semiconductor processing chips into functioning components requires packaging of the chips, placing the packaged chips on circuit boards with the appropriate electrical interconnections, delivering well-regulated electrical power to the processor chips, removing the heat generated by the chips, and providing mechanical hardware to support this infrastructure. These

0018-8646/02/$5.00 © 2002 IBM

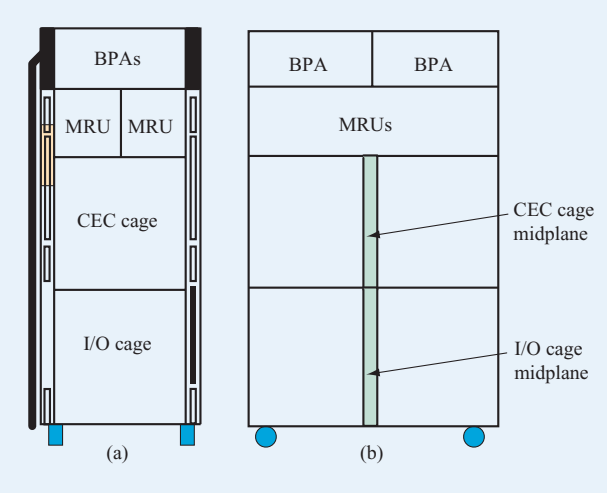

Figure 1

(a) Front view and (b) side view of a z900 server showing the four subsystem building blocks: bulk power assemblies (BPAs); modular refrigeration units (MRUs); central electronic complex (CEC) cage; and input/output (I/O) cage.

tasks are essential to creating a computer that has high reliability, essentially zero downtime, minimal floor space and power requirements, the ability to be powered by a wide range of ac- or dc-line voltages, no objectionable electromagnetic radiation or acoustical noise, and the ability to withstand electromagnetic radiation, mechanical shock, and vibration within specified limits. While a truly extraordinary product is one that employs state-of-the-art components, it also requires a premier packaging, power, and cooling design.

Using the IBM eServer z900 as an example, the aim of this paper is to describe the following portions of the development of a computing system: the electronic package, the power conversion and delivery subsystem, the cooling subsystem, and the mechanical support infrastructure.

System package overview

The IBM eServer z900 comprises four subsystem building blocks packaged in a steel frame, as shown in **Figure 1**. The four subsystems are the bulk power assembly (BPA), the central electronic complex (CEC) cage, the input/output (I/O) cage, and the modular refrigeration unit (MRU). They are powered and controlled through a network of cables called universal power input cables (UPICs)[1] and a service control Ethernet-based network of hardware and software referred to as the power service control network (PSCM).[2] The system architecture supports both single- and dual-frame configurations, depending on the number of I/O cages included. The single-frame configuration consists of a frame referred to as the "A" frame, while the dual-frame configuration adds an expansion frame, referred to as the "Z" frame, to house additional I/O cages. In a dual-frame configuration, power to the complete system is usually provided by two BPAs in the top section of the A frame. However, when system power requirements exceed the power available from the two BPAs in the A frame, the dual-frame system is reconfigured such that the top section of the A frame and its BPA contents are eliminated, and power to the complete system is supplied by two larger BPAs installed in the space above the I/O cages in the Z frame. The front and rear views of a two-frame system are shown in **Figure 2**.

The system enclosure provides roughly 10 dB of acoustical noise attenuation, primarily through the use of specially designed front and rear acoustical doors. The enclosure is also designed to meet FCC regulations and limits for electromagnetic compatibility (EMC) and the containment of electromagnetic interference (EMI).

The system requires $n + 1$ redundancy across all subsystems, where n is the number of subsystems needed for operation of the system. A failed subsystem is field-replaceable, or repairable during concurrent operation of the system. This requirement necessitates the use of two BPA subsystems, which are mirrored about the midplane of the server. Each can independently power the entire server. There are two BPA subsystem designs, one with up to 13 kW and the other with up to 19.5 kW of power delivery capability. Dual redundant line cords bring in 200–480 Vac power, which is converted to 350 Vdc by the BPAs and distributed to point-of-load dc–dc converters and motor drives. The point-of-load dc–dc converters are housed in distributed converter assemblies (DCAs). The DCAs, plugged directly into the midplane boards of the CEC and the I/O cages, convert 350 Vdc to the precise voltages required by the logic and memory circuitry. The motor drives are special regulators that power the motors used in the various air-moving devices (AMDs).

The CEC cage subsystem, shown in **Figure 3**, provides the physical structure and the interconnections for the processors and the memory. The arrangement of components is governed to a large extent by processor

[1] A UPIC cable connects a power distribution unit to a DCA or a motor drive and contains both the branch circuit power feed and the communications link.

[2] Power Service Control Network (PSCN): The power and cooling subsystems of the z900 server are controlled through a fully redundant dual-Ethernet communications network. The 100Mb network provides for communication to all field-replaceable units and hierarchic control through a mirrored system of control cards and IP addresses. Each cage contains a master and slave control card. These cards, referred to as "cage controllers," provide the interface points with the PSCN. The PSCN provides a means for subsystems to communicate and control the dynamic parameters of system operation. The PSCN also supports error detection and correction of both the internal hardware and software systems.

timing, power, and thermal management issues. The design of the CEC cage is centered around a midplane circuit board assembly [Figure 1(b)] consisting of the planar mother board mechanically supported between two aluminum stiffeners (not shown) [1]. The cage is designed to allow interconnect access to both sides of the midplane circuit board. The multichip module (MCM), consisting of up to 16 processors, is plugged directly into the midplane circuit board. The rest of the hardware that plugs into the midplane circuit board is housed in self-contained sheet-metal enclosures called "books" because of their similarity in size and shape to a large book. A book provides a protective package to the printed-circuit card, while providing latches for mechanical attachment to the cage structure. Very High Density Metric (VHDM**) connectors connect the printed-circuit cards in the books to the midplane circuit board [2]. Plugging the books into the board is aided by mechanical guides built into the cage structure. The CEC cage is powered by three DCA books, one of which is shown in **Figure 4**, which house the point-of-load dc–dc converters used to convert the 350 Vdc distributed from the BPAs to the precisely regulated low dc voltages required by the CEC. The CEC cage DCAs have $n + 1$ redundancy by including one more DCA than is required to power the CEC cage. For BPA redundancy, each DCA receives power and control signals through UPIC connectors and cables from the two BPAs. The CEC cage supports as many as four memory books, two plugged vertically into each side of the midplane. A fully

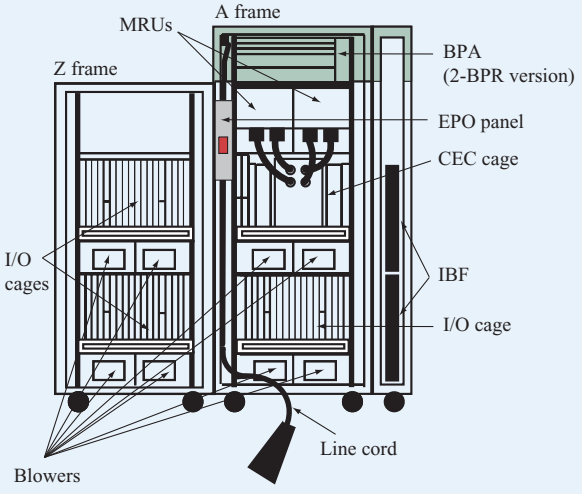

Figure 2

(a) Front view and (b) rear view of a two-frame system. The shaded section of the A frame is the removable top section containing the BPAs. BPR = bulk power regulator; EPO = emergency power-off. IBF = integrated battery function.

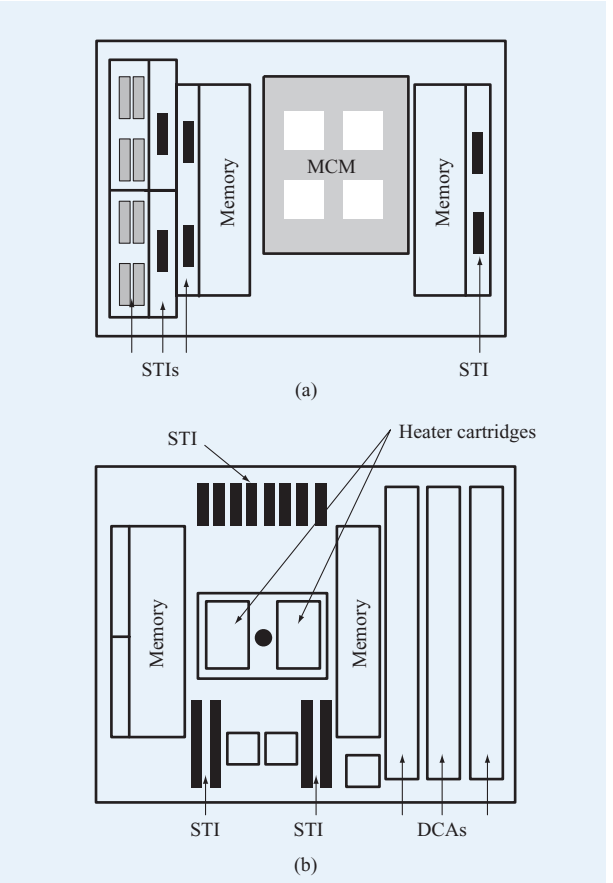

Figure 3

Central electronic complex (CEC) cage: (a) Front view; (b) rear view.

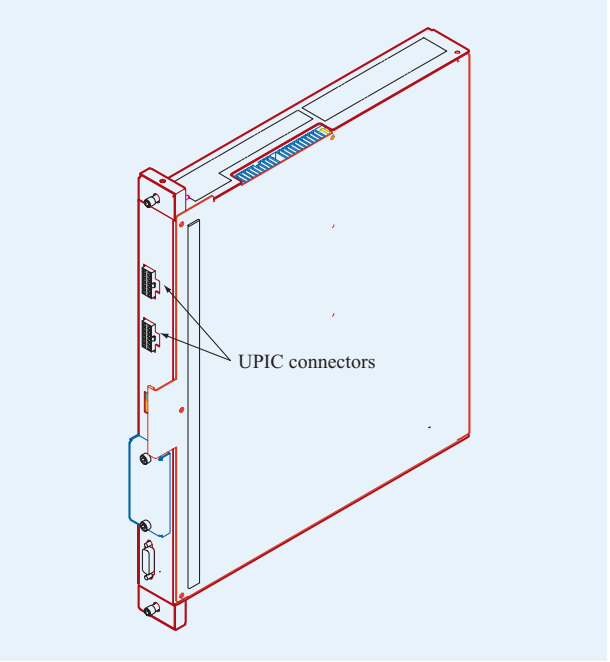

Figure 4

DCA book containing the point-of-load dc–dc converters. The face opposite the face that contains the UPIC connectors, not shown above, contains the Very High Density Metric (VHDM) connector that plugs into the backplane circuit board.

configured z900 server can support up to 64 gigabytes of memory. The memory books and the DCAs are cooled by two fully redundant blower assemblies located in the lower section of the CEC cage. Each of these AMDs is driven by its own motor drive card and is capable of independently cooling the cage.

The I/O cage has a mid-planar design similar to that of the CEC cage. Its function is to provide standardized slot locations for a variety of I/O books. The z900 server supports more than ten different I/O book types. Each I/O cage may contain as many as 36 I/O books powered by two DCAs. As with the CEC, each I/O cage has two centrifugal blowers that provide cooling with full redundancy and concurrent maintainability. The modular design of the cage allows the cage to become its own independent subsystem, requiring only a 350-Vdc power feed from the two BPA subsystems and an Ethernet connection to the PSCN. I/O cages are divided into seven domains, and each domain can be supported with multiple data channels called self-timed interface (STI) links [3]. These STI cable interconnects provide the pipeline for data flow between the I/O devices in the cage and the processors in the CEC. The single-frame z900 server configuration has one I/O cage at the bottom of the A frame. In the larger two-frame configuration, the addition of the expansion Z frame provides system scalability by supporting two additional I/O cages. I/O cages are packaged in the lower section of a frame to allow easy access to the mass of interconnect cables typically routed under the raised floor.

The z900 server contains two fully redundant freon-based MRU subsystems [**Figure 5(a)**], each packaged in its own sheet-metal enclosure that contains a compressor, condenser, drive electronics, and an AMD. Each MRU is designed to be a fully redundant field-replaceable unit. The function of the MRU is to cool the MCM and maintain a nominal processor junction temperature of 0°C. The evaporator for the refrigeration system, shown in **Figure 5(b)**, is a copper cold plate with dual independent channels through which refrigerant flows. The evaporator cold plate is bolted directly to the MCM in the CEC cage. Each channel in the cold plate is supplied refrigerant, under pressure, from its own dedicated MRU via insulated, flexible metal hoses. In this tethered evaporator design, the MRUs are located just above the CEC cage to minimize the length of the refrigerant hoses. Quick-connect hose couplings attached to the ends of the evaporator hoses provide convenient means of connection to the two MRUs without loss of refrigerant.

The mechanical skeleton of the z900 server is the steel frame. Industry standards for frames have been established for two key metrics: width and height. The z900 server is packaged in a 24-in. (610-mm)-wide frame, as opposed to the 19-in. (483-mm) racks commonly found in the telecommunications industry. Frame height is measured with a standardized unit of length known as an EIA U, or simply U, as defined by the Electronic Industries Alliance [4]. One U is equivalent to 1.75 in. (44.45 mm). U is also the unit of measure for rack- and frame-mounted component heights. The z900 server is housed in a 24-in.-wide, 42-U frame. Within the frame, subsystems are mounted on pairs of standardized "EIA rails" that are part of the frame structure. The structural integrity of a frame is a key metric of frame quality and is closely monitored by a design team. Each z900 A frame supports a total sustained static load of well over a ton. When it is subjected to shock, drop, or earthquake testing, the mechanical stresses on the frame become considerable. The loading capacity of a frame is often specified as the number of pounds per U. The structural design of the IBM server frames allows them to be shipped fully assembled, with no assembly required at the customer site.

First-level electronic package

The first-level package in a z900 server consists of a glass-ceramic multichip module on which are mounted microprocessors and supporting chips for second-level cache and communication with peripheral devices. The

second-level package consists of a multilayer fire-retardant epoxy (FR4) midplane circuit board with 36 metal layers containing the signal wiring and power distribution planes [1]. The single-chip S/390* microprocessor was introduced in 1996 in the IBM G3 server, in which a single 127.5-mm × 127.5-mm MCM contained 12 microprocessors plus the required second-level cache and I/O chips that made up the heart of the computer CEC. The technology and design aspects of these MCMs are described in detail in [5] and [6]. The tradeoffs that led to this choice of technology are discussed in [7].

Table 1 lists the attributes of the MCMs. All of these MCMs were built using 127.5-mm × 127.5-mm ceramic substrates with four to six layers of thin-film technology on top, taking advantage of the MCM technology and design infrastructure already developed for the S/390 mainframes [5, 6]. With each generation of CMOS technology, the circuit density and device performance improved, more function was integrated into the MCM, and faster cycle times allowed higher computational performance. The G4 server, with its uniprocessor capable of 63 million instructions per second (MIPS) and its 10-way shared multiprocessor (SMP) configuration capable of 450 MIPS, became the first CMOS machine to match the performance of the last IBM bipolar mainframe (ES/9000*). The SMP configuration includes multiple microprocessors with a coherent, shared second-level cache in a single operating system image [8]. The 150-MIPS uniprocessor in the G5 server yielded more than 1000 MIPS in a 10-way SMP configuration. The G6 server provided 200 MIPS in a single uniprocessor and more than 1600 MIPS in a 12-way SMP. The z900 server introduced the 64-bit S/390 uniprocessor and a 16-way SMP capable of 250 and 2700 MIPS, respectively. The MCM provides the chip packaging density and chip I/O density to allow a high-bandwidth, nonblocking-crossbar second-level cache architecture which provides the scalability of performance from a uniprocessor to a full SMP that is expected in a S/390 server [7].

Table 1 shows that from the G4 server to the z900 server, the number of processor chips increased from 12 to 20, the frequency of the processor chips doubled, and the amount of second-level cache memory increased tenfold. These increases were due to improvements in CMOS technology. Table 1 also shows how the MCM technology has improved with each server generation to provide connections with faster signal propagation, more wiring capacity, and higher signal pin density. The MCM technology has been introduced so that the performance per unit cost is improved with each generation [7]. The nominal processor temperature, also provided in Table 1, is the mean transistor junction temperature of all the transistors on the processor chips as determined by thermal modeling based on the calculated power of the chip. Lower temperatures result in processor chips that run more reliably and at higher frequencies. The section on the modular cooling unit provides details on the cooling technology.

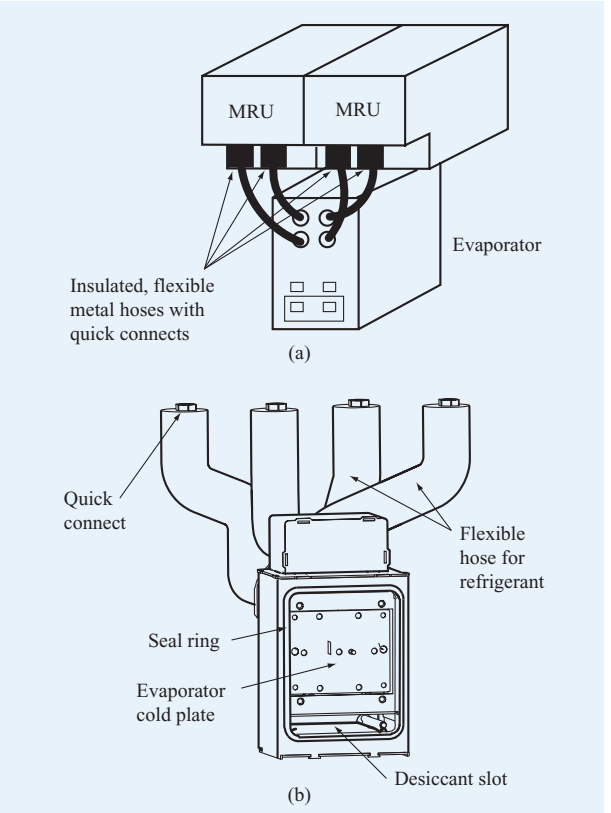

Figure 5

(a) z900 server refrigeration system. The evaporator attaches to the MCM cold plate; thus, heat from the MCM chips is removed by the freon flowing through the evaporator channels. (b) Some details of the evaporator design. The evaporator cold plate contains two independent channels through which refrigerant flows. Each channel is connected to an MRU by two flexible hoses.

As shown in Table 1, the first-level packaging was changed from alumina in the G4 server to glass-ceramic for the MCMs of the G5 to z900 servers. The glass-ceramic provides faster signal propagation and less loading capacitance on the interconnects, which enhances the performance per unit cost of the G5 to z900 systems [7]. The glass-ceramic substrate feature dimensions were reduced for the z900 server to provide the wiring capacity needed for the 1816-signal I/O chip and the 920 meters of wire.

The higher circuit densities provided by enhanced CMOS technology resulted in an increase in the number of smaller processor chips on an MCM substrate.

Table 1 Attributes of S/390 processor MCMs.

Year—server	1997—G4	1998—G5	1999—G6	2000—z900
Number of chips	30	29	31	35
Number of processor unit chips	12	12	14	20
Processor frequency (processor voltage)	370 MHz (2.7 V)	500 MHz (2.0 V)	637 MHz (2.0 V)	770 MHz (1.7 V)
L2 cache (MCM voltage)	3 MB (2.7 V)	8 MB (2.6 V)	16 MB (2.0 V)	32 MB (1.7 V)
MCM power capacity (W)	1050	800	900	1400
Processor nominal[†] *junction temperature* (°C)	45	20	15	0
MCM technology 127.5 mm × 127.5 mm	Alumina/thin-film redistribution	Glass-ceramic/ thin-film wire	Glass-ceramic/ thin-film wire	Glass-ceramic/ thin-film wire
Total number of C4s/wire length	83,000/460 m	80,000/600 m	85,000/640 m	101,000/920 m
Max. number of signal C4s per chip	1244	1244	1244	1816
Ground rules for the thin-film (TF) layers	Four layers/56 μm	Six layers/45 μm	Six layers/45 μm	Six layers/33 μm
Glass-ceramic (GC) ground rules (Note: G4 used alumina.)	87 layers/450 μm	75 layers/450 μm	87 layers/450 μm	101 layers/396 μm
MCM I/O technology	3528 total pins/ 1764 signal	4224 total pins/ 2450 signal	4224 total pins/ 2450 signal	4224 total pins/ 2489 signal
Nominal characteristic impedance, Z_0 (Ω)	50	TF–39 GC–60	TF–39 GC–60	TF–43 GC–55
Nominal signal delay (ps/mm)	11.5	TF–6.4 GC–7.8	TF–6.4 GC–7.8	TF–6.4 GC–7.8
Nominal signal line resistance (Ω/mm)	0.035	TF–0.2 GC–0.022	TF–0.2 GC–0.022	TF–0.24 GC–0.022
Signal line cross section (μm)	90 × 30	TF–18 × 6 GC–70 × 25	TF–18 × 6 GC–70 × 25	TF–16 × 4.5 GC–70 × 25
Dielectric thickness (fired) (mm)	0.2/0.15	TF–0.01 GC–0.111	TF–0.01 GC–0.111	TF–0.01 GC–0.092
On-MCM capacitors	100 nF/cap 220 capacitors	200 nF/cap 177 capacitors	200 nF/cap 202 capacitors	300 nF/cap 366 capacitors

[†]The term *nominal* refers to the expected value of the respective parameters.

Improved refrigeration has continued to reduce the nominal chip temperatures to the calculated values listed in Table 1, as indicated by the increasing performance and power-handling capability of the MCM. Nominal processor junction temperatures are the mean of the junction temperatures of all of the transistors on the processor chips. This temperature is determined by thermal modeling based on the calculated power of the chip, and it is one of the parameters for doing system timing and reliability analyses. (See the section below on the modular cooling unit.)

MCM net delay timing and package performance

Meeting the signal timing requirements is a key factor in the design of the MCM package. With each generation of CMOS processors, the processor clock frequency has been increased. The design constraint is that the MCM net delay must support a frequency of one-half the processor frequency. The components that make up the total signal delay of the interconnects are shown in **Figure 6**. The electronics signal delay occurs in the on-chip portion of the path, including the latch, set-up time, off-chip driver, and receiver. The wire delay is the propagation time on the MCM interconnect. The clock skew includes clock distribution variations and phase-locked-loop (PLL) jitter [7]. The noise delay is the impact of crosstalk and power-supply switching noise on the signal delay.

In Figure 6, one can see the relative decrease in MCM net (path) signal delay from system to system. The decrease in electronics delay does not scale with the

expected speed-up of CMOS devices because the slew rate of the drivers is limited to control of the crosstalk noise on the interconnects. The large decrease in MCM net delay between G4 (1997) and G5 (1998) servers shows the significant impact of introducing a glass-ceramic material with a relative dielectric constant of 5.3 to replace alumina with a relative dielectric constant of 9.5. The reduced clock skew in the G5 server was also due to better power-distribution noise and reduced PLL jitter. For the z900 server, placement of chips and routing of interconnects provided shorter connections and significantly reduced wire delay to meet the cycle-time requirements of the machine.

The off-MCM nets are more challenging than the on-MCM nets because of the longer interconnect lengths, but they are still required to operate at one-half the processor clock frequency to provide the required latency and bandwidth between the second-level cache and main memory. For the G4 server, the cycle time was achieved on the off-MCM nets using a moderate amount of cycle stealing, delaying the receiver clock, and launching the driver early. In the G5 server, the clock on the off-MCM chips was run at a 90° offset from the MCM chips [5]. In the G6 server, a 180° offset was used to achieve the cycle time. For the z900 server, a source-synchronous interface similar to the interface described in [9] has been introduced. This interface has a differential clock bundled with each off-MCM data bus, and the timing impact of process and environmental variations is minimized by electronically aligning the clock and data during the power-on sequence of the z900 system.

Crosstalk noise

Noise must be limited to the noise tolerance of the receivers in the critical timing windows for proper operation of the computer. To achieve this, the noise is budgeted at the time of the high-level design of the MCM, as described in [7] and used in [5]. The components are identified, the noise estimated, and a wiring strategy developed to meet that budget.

Once the design is completed, a post-route analysis of all of the connections is performed as described in [10]. This methodology identifies the individual MCM net segments introducing crosstalk noise, calculates the amplitude of the crosstalk noise created and the time during the cycle when it is created, transmits the noise to the receiver circuit, and sums the crosstalk noise of all of the coupled sections at the input of the receiver. For off-MCM nets, it is not enough to analyze each package level separately. The interconnects at each package level (MCM, printed-circuit board, memory card, and the connectors between the levels of packaging) are stitched together using customized software so that the crosstalk

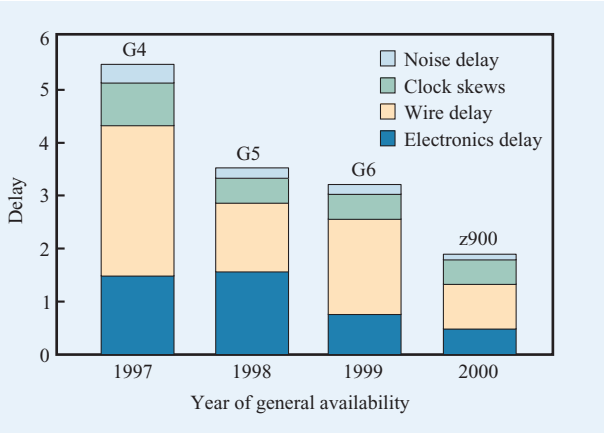

Figure 6

MCM net delay components in picoseconds.

noise for the complete connection is calculated at the receiver inputs. The high-frequency switching noise waveform (simultaneous switching noise) from the power supplies is added to the summed crosstalk noise to determine the total noise amplitude at the receiver input. The total noise is compared against the allowable noise tolerance in the window of vulnerability, where the receiver is susceptible to excessive delay or latching to an incorrect logic level. If the noise envelope fails the criteria, the net is rerouted so that the noise does not exceed the noise tolerance of the receiver.

Power distribution and decoupling capacitors

The power distribution provides the foundation for the proper operation of the computer. Sufficient power and ground layers are included to maintain the resistive voltage drops at acceptably low levels so that the voltage tolerance specifications are met. Assignment of the power and ground plane layer provides the proper return paths for the high-frequency signals interconnecting the chips. The vias in the MCM, the pins on the MCM, and the card and board connectors are designed to have signal-to-power ratios of one to one; that is, for every signal via or pin, there is either a voltage or ground via or pin to provide crosstalk shielding. The voltage and ground assignment of the vias and pins provides continuous current return paths from the reference planes of one package level to the reference planes on the next package level.

Once this power distribution is provided and the voltage requirements are defined, a power-distribution voltage-variation budget is created and a decoupling strategy is defined. For the z900, the voltage-regulation set point on the second-level package is 1.75 V. On average, there is 50 mV of voltage reduction from the set point to the

Table 2 Decoupling capacitors for the processor supply in the G4 to z900 servers.

Server generation	G4	G5	G6	z900
Chip voltage, V_{dd} (V)	2.7	2	2	1.7
Total current on processor unit supply (A)	400	200	400	700
Total thin oxide capacitance on processor unit chip (nF)	100	200	200	200
On-MCM ceramic decoupling capacitors (μF)	220–0.1	90–0.2	120–0.2	275–0.3
On-board ceramic decoupling capacitors (μF)	440–2.2	920–1 250–10	920–1 250–10	2900–1 670–10
Electrolytic capacitors for bulk decoupling (μF)	100–560	200–560	200–560	320–1800

circuit terminals on the processor chips. The average voltage across the CMOS circuits is 1.7 V. For frequencies up to 100 MHz, the voltage at the processor chip is allowed to vary ±100 mV from 1.7 V. Of this 100 mV, there is a budgeted ac variation of ±50 mV (3% of the voltage). The electrolytic capacitors and on-board ceramic capacitors shown in **Table 2** are placed so that this budget is attained. For frequencies about 100 MHz, the allowed voltage variation is ±5% and the on-chip thin-oxide capacitance and on-MCM ceramic capacitors shown in Table 2 are needed to meet that requirement.

The ac variation limits are met by the proper use of decoupling capacitors. The number of capacitors is important, but their performance is optimized by appropriate placement and by limiting their series inductance and series resistance. The decoupling capacitors used in the G4 to z900 servers are listed in Table 2. Decoupling capacitor technology and the design strategy for using decoupling capacitors have improved significantly as systems have advanced from the G4 to newer z900 servers. Improvements in the capacitor technology have provided lower-impedance capacitors. The design improvements include better capacitor placement relative to the active components and designing the MCMs and boards for lower series inductance from the capacitor to power-distribution layers by reducing the voltage-to-ground via inductance.

Second-level electronic package

The complexity of IBM computer systems increases with the addition of new features, increased performance, and smaller sizes. In turn, this increases the demands on the second-level electronic package, which is more complex and dense. Reducing the size of the devices on a chip and module packages with increased connections is no longer only an integrated circuit (IC) technology issue, but is also becoming a second-level packaging concern.

The second-level electronic packaging refers to the enclosure for the electronic logic and its associated hardware, which includes the midplane circuit board, multichip modules, passive devices, and supporting integrated-circuit chips. The processes involved in producing this hardware are also considered a part of the second-level electronic packaging technology. The costs associated with packaging these complex, high-performance products are rising faster than the integrated-circuit costs.

Today, the conventional boundaries between individual components (first-level packaging) and their interconnections at the board assembly level (second-level packaging) have all but disappeared. The effectiveness with which a computer system performs its electrical functions, as well as the reliability and cost of the system, is strongly determined not only by the electrical design, but also by how each component is packaged into the assembly. The complexity of an electronic package is determined by the type of tradeoffs made between packaging constraints and system performance requirements.

Major advances in the IBM single-chip and multichip first-level device technologies have had a dramatic effect on the way each functional element within a system is partitioned. Greater function and tighter integration of each new generation of IBM and industry devices allow greater flexibility in developing denser second-level packages. For example, **Figure 7(a)** shows six I/O logic board assemblies that are required to provide sixteen IBM Enterprise Systems Connection (ESCON*) I/O channels. **Figure 7(b)** shows that advances in first- and second-level IBM packaging technologies have made it possible to package the same function in a single logic board assembly.

The choice of the printed wiring board (PWB) material and structure is important, because the PWB provides the electrical and mechanical support upon which the logic structure is built. It provides not only the component mounting surfaces but also the medium for component-to-component interconnections, electromagnetic interference (EMI) shielding, and paths for off-board power and signal connections through Very High Density Metric (VHDM) connectors to the midplane circuit board.

The dense array of wiring connections required to package the logic in Figure 7(b) was achieved by reducing the PWB hole size, pad diameter, and wiring trace widths to allow for an increase in wiring density between pads. Additional PWB layers were added to prevent electrical interaction between the critical high-speed logic. This resulted in a controlled-impedance PWB design that met all of the electrical requirements for the interconnections, and provided good manufacturing yields at low costs.

Next, electronic and mechanical components were selected that not only met the electrical design criteria but were compatible with the PWB material and board-assembly manufacturing and test processes. The PWB design was required to support ceramic ball grid array (CBGA) modules, pin-in-hole (PIH) components, optical and logical modules, active and passive surface-mount components, and the VHDM connector. The package also had to support light-emitting-diode indicators attached on the far side of the board assembly. After selection of components was complete, the components on the PWB were arranged and positioned to provide the shortest electrical wiring paths and to allow sufficient space between them to provide adequate cooling. Capacitors were positioned close to the ICs to be electrically effective while providing for the ease of board manufacturing steps such as placement, soldering, and repair and testing of these and other components. In today's high-performance designs, capacitors are critical components that provide filtering of noise from the electrical interfaces and store electrical energy. Finally, a protective envelope enclosed the completed board assembly within the dimensional limits of the cage and frame assembly. The sheet-metal enclosure had to provide thermal control, EMI shielding, electrostatic discharge (ESD) protection, and positive retention when installed in the midplane. The package was also required to have a design that provided ease of maintenance in the field during system operation, including plugging fiber optic cables into the optic modules, and a level of structural integrity that supported the board assembly during shipment. The package also provided thermal management for the sixteen small-form-factor (SFF) fiber optic modules [11] that are outside the normal airflow path. The heat-sink subassembly shown uses a combination of forced-air convection and thermal conduction materials that allows fiber optic module heat to be dissipated to the cover of the board assembly enclosure and also into the normal airflow within the package.

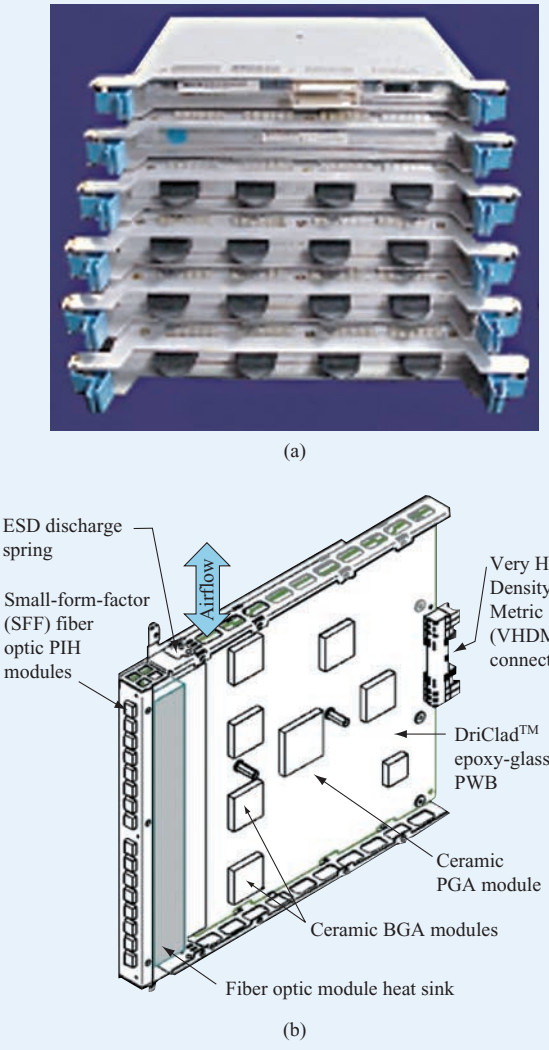

Figure 7

(a) Six I/O logic board assemblies required to provide sixteen IBM Enterprise Systems Connection I/O channels within large IBM servers. (b) Logic package (shown with cover removed) which is equivalent in performance to the six packages shown in Figure 7(a). BGA = ball grid array; PGA = pin grid array.

The trend in electronic packaging toward more sophisticated first-level packaging with more I/O connections is making, second-level design considerations, design tradeoffs, and selection of materials more complex.

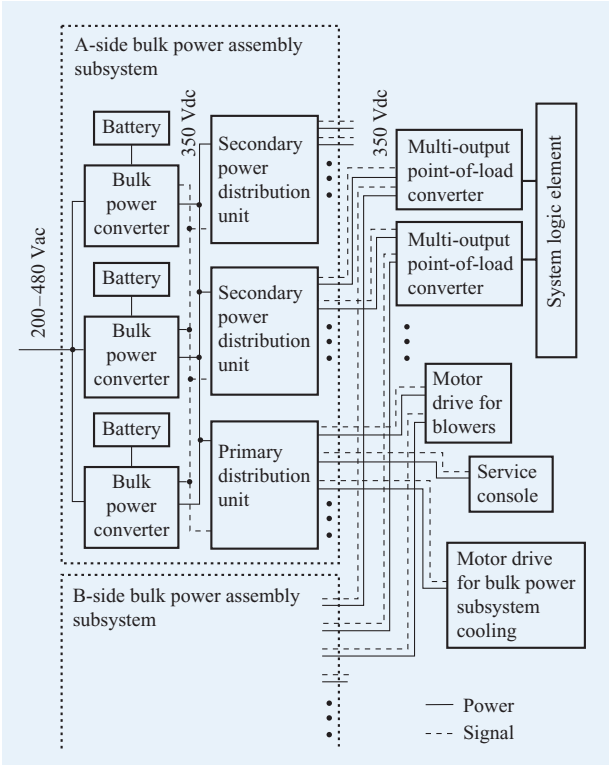

Figure 8

High-level block diagram of the power system in the Z frame. The multi-output point-of-load converters are the DCAs.

The power system

The z900 server power system architecture provides high packaging density for the processor complex while delivering very high availability. High availability is achieved through reliability, redundancy, and concurrent maintainability. Concurrent maintainability is the ability to replace a redundant, failed unit while the system is in operation. It is achieved by a distributed microprocessor control system. The control system flags the failed component, which can then be concurrently replaced.

An overview of the power system is given in [12]. A 200–480-Vac input is converted in the BPAs to 400 Vdc using a single-phase, power-factor-correction, buck+boost converter, and then to 350 Vdc using a dc–dc converter. The 350 Vdc is distributed throughout the server and is locally transformed by the point-of-load dc–dc converters, housed in the distributed converter assemblies (DCAs), to the voltages required by the loads being served.

The entire power system is fully redundant, from the customer ac line inputs to the precision-regulated low-voltage feed to the CEC and I/O cages. As shown in **Figure 8**, a z900 server power system contains two BPA subsystems, one on the A side (the front side) and the other on the B side (the back side). In a one-frame system, each BPA has two bulk power converters (BPCs) that convert the 200–480-Vac input into 350 Vdc. Battery backup is provided to each BPC. In a two-frame system, the additional power requirement of the Z frame may dictate the need for higher-power-rated BPAs; in this case, the BPAs in the A frame are eliminated and larger BPAs, each with three BPCs, are installed in the top portion of the Z frame. The outputs of all of the BPCs, in each BPA, are diode-ORed to form a common 350-Vdc bus feeding the power distribution units (PDUs), which in turn feed the DCAs, the motor drives, and the service console. The PDUs also serve as communications hubs for all of the downstream hardware they feed. Upstream communication to the service console (a laptop computer) is managed by the primary PDU. The DCAs plug directly into the CEC and the I/O midplane circuit boards. DCAs are $n + 1$ redundant. Each DCA has a 350-Vdc power input and a communications connection, via UPIC, from two PDUs, one in the A-side bulk power assembly subsystem and the other in the B-side BPA. Either PDU is capable of powering and controlling the DCA. The power feeds from the two PDUs are diode-ORed inside the DCAs. Each of the two power-line cords that attach to the z900 server is capable of powering the system entirely by itself. This permits continuous operation in the event that service must be performed on the facilities power system or, if there is a complete loss of power, on one power-line cord.

Bulk power assembly (BPA) subsystem

The bulk power converters (BPCs) in z900 servers are rated at 6.5-kW continuous output power. Power conversion within a BPC is accomplished in two stages. As shown in **Figure 9(a)**, the input power stage takes wide-ranging ac line voltage and converts it to regulated 400 ± 12 Vdc while ensuring less than 5% total harmonic distortion of the input line current [13, 14(a)]. The output stage of the BPC [**Figure 9(b)**] is a dc–dc converter with active clamp that transforms 400 Vdc into 350 Vdc, which is then fed to the point-of-load dc–dc converters in the DCAs [15]. The input power stage of the BPC consists of a full-wave bridge rectifier followed by a buck+boost converter which can step up or step down the input voltage to achieve a near-constant output voltage. If the ac line voltage peak is greater than the desired output voltage, the buck+boost converter will buck (step down) during the portion of the cycle when the input voltage is greater than the desired output voltage and boost (step up) during the portion of the cycle when the input voltage is less than the desired output voltage. If the ac line voltage peak is less than the desired output voltage, the buck+boost converter will

always boost (step up) the input voltage. The full-wave bridge rectifier followed by a buck+boost converter achieves a power factor near unity by a using a controller that forces the input current to follow the input voltage; that is, the input power stage behaves as a resistive load, which is the most desirable load from the power utility point of view [13, 14(a)]. The buck+boost topology is preferred over the more conventional boost topology because of the wide range of ac line voltages (200–480 Vac) over which the buck+boost converter can operate [13]. A boost topology alone would have required the output voltage to be, typically, 20% greater than the peak of the ac line voltage, making the minimum output voltage for a boost converter running off a 480-Vrms ac line to be 680 V ($= 480\sqrt{2}$). High output voltage severely limits component selection. The buck+boost converter, on the other hand, allows the output voltage of the converter to be arbitrarily selected to be larger or smaller than the ac line voltage. The output voltage was selected as 400 Vdc because components in this voltage range are readily available, and 400 Vdc is high enough to permit low switch currents for the BPC output stage power dc–dc converter [Figure 9(b)]. The insulated gate bipolar transistor (IGBT) buck switch [Figure 9(a)] also provides the ability to directly limit the initial surge in current during start-up as well as under power-line transient conditions.

The IGBT buck switch of the buck+boost converter is exposed to the full magnitude of the potentially noisy input voltage waveform. The requirement of the buck switch to withstand high voltages led to the selection of a 1200-V IGBT. MOSFETs cannot be used for this application because they have limited blocking voltage. Also, the on-resistance of the high-voltage MOSFETs is quite high, making conduction losses unacceptable for operation at input line voltages of 240 Vac and below, where the line currents are high. Since the main drawback of the IGBT was its slow switching speed, a converter-switching frequency of 50 kHz was determined to be the optimal tradeoff between efficiency and size for the converter. The boost switch of the buck+boost converter is never exposed to voltages greater than the output voltage (400 Vdc), which makes a MOSFET the preferred choice for this application.

The 400-Vdc intermediate bus in the bulk power converter (BPC) provides power to the output-stage dc–dc converter of the BPC [Figure 9(b)]. The intermediate bus is a convenient point for connection of the backup battery unit. The battery-stack voltage is selected as 350 V to allow charging from the 400-Vdc intermediate bus. The battery stack is connected to the intermediate bus via a power MOSFET. The body diode of the power MOSFET in the return leg of the battery stack permits a natural battery-discharge path should the 400-Vdc bus dip because

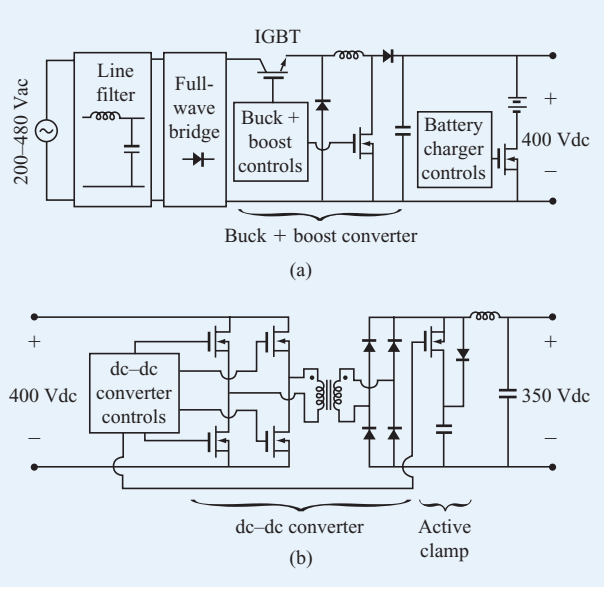

Figure 9

Bulk power converter: (a) Input stage; (b) output stage.

of line outage. A battery-control circuit connected to the gate of the MOSFET acts as a charger when line voltage is present.

The main purpose of the dc–dc converter output stage of the BPC is to buffer the line transients that occur due to input line incidents and during battery discharge. During these situations, the 400-Vdc intermediate bus may dip as much as 30%. The regulation of this variability allows the point-of-load dc–dc converters in the DCAs to be optimized around a much smaller input voltage range. The dc–dc converter output stage of the BPC is a phase-shifted, full-bridge, zero-voltage transition converter with active clamp [13]. This topology was selected for its high efficiency and because its fixed frequency (200 kHz) allows more predictable behavior over large load variations. The typical efficiency of this power-conversion stage is 95%. The control scheme for load sharing of parallel bulk power converters is the forced-droop method [16].

For high-voltage-output-stage dc–dc converters, the ringing between the leakage inductance of the transformer and the capacitance of the full-bridge rectifier diodes imposes a severe voltage stress on the diodes [Figure 9(b)]. This problem is addressed by using an active clamp circuit that eliminates the ringing in a nondissipative manner [13]. The voltage stress on the diodes is significantly reduced, making possible the use of faster, lower-voltage rectifier diodes.

Inside each of the BPCs is a microprocessor subsystem. The microprocessor monitors the intermediate bus voltage,

output voltage, input current, and internal temperature of the converter. In addition, the microprocessor monitors the status of the battery stack. Diagnostic information and commands are communicated to the primary power-distribution unit via an RS422 interface.[3]

The power-distribution units (PDUs) act as both the power and communication hubs for the server. Each DCA or blower motor drive has a dedicated branch circuit formed within a PDU. The branch circuit is protected by a static circuit breaker, which plays a key role in ensuring the integrity of the 350-V bus during failures of DCAs or motor drives. The breaker is designed to regulate the current in the event of a fault and disconnect the fault if it is not cleared within 200 μs. Current limiting and load disconnection are performed electronically. Each PDU has four high-current branch circuit feeds and four low-current feeds. The high-current feeds are provided for the DCAs, and the low-current feeds are for the blower motor drives. A microprocessor unit in each of the PDUs monitors the static breakers and identifies their status to the primary power-distribution unit via the same RS422 bus connection to the bulk power converters.

The Ethernet provides communication paths through the PDUs to the DCAs, and the RS422 interface provides communication to the motor drives. An Ethernet bus was selected because the DCAs act not only to supply power to the logical elements, but also as the high-speed service interface to the logical elements. The Ethernet bus originates in the system control console and connects to the primary power-distribution unit, where an Ethernet repeater is provided which extends the bus to the secondary power-distribution units. Each of the secondary power-distribution units contains an Ethernet repeater that extends the bus to the DCAs. A single UPIC cable connects a power-distribution unit to a DCA or a motor drive and contains the branch-circuit power feed as well as the communication link.

The primary power-distribution unit performs all of the functions of the secondary distribution units as well as acting as the link to the service console for all system entities via the Ethernet connection. The RS422 buses that are used to communicate with bulk power converters and secondary power-distribution units originate in the microprocessor unit in the primary power-distribution unit. The primary power-distribution unit also contains a motor drive used for the blower that cools the bulk power assembly subsystem cage. The motor drive is similar in design to those described later in this paper.

A customer-accessible emergency power-off (EPO) switch is provided in the server. This switch is cabled to the primary power-distribution unit, from which it is routed to each of the bulk power converters. When operated, the EPO switch disconnects power to the control circuits of the bulk power converters, ensuring that all power is removed from system entities. The primary power-distribution unit also provides power to a laptop computer that is the system console.

Choice of 350-Vdc power distribution

As already mentioned, the DCAs provide regulated power directly to the CEC and the I/O midplane circuit boards, and the motor drives provide regulated power directly to the fans and blowers. The power source for the DCAs and the motor drives is ultimately the utility power. Several choices are available for distributing this utility power to its points of use throughout the system.

Direct distribution of the ac-power-line voltage to the DCAs and motor drives was the universal method employed in large computing systems as recently as a decade ago. In the IBM z900 system, this has been entirely supplanted by centralized ac-to-dc conversion and dc voltage distribution to the point-of-load power regulators, such as DCAs and motor drives. Advances in the density and efficiency of point-of-load power regulators since the advent of surface-mount technology have permitted packaging of the regulators close to and even integrated with the system logic, eliminating large bus bars and complex discrete sense networks. Moreover, new requirements imposed on front-end power conversion, such as power factor correction and harmonic current suppression, are an added incentive for centralizing ac-to-dc conversion. Because the functions associated with ac-to-dc power conversion are performed once, the local point-of-load power regulators, freed from the requirements imposed on front-end power converters, are thus simplified and smaller in size. In summary, dc voltage distribution, through the more efficient packaging of power conversion and the associated elimination and miniaturization of hardware, has contributed much to the compacting of the system footprint since the ascendancy of CMOS in mainframe computers.

A number of dc distribution voltages are favored in the industry. The standard distribution voltage in the telecommunication sector is 48 Vdc because it is safe, extra-low voltage (SELV), for which hardware designs need not consider high-voltage spacing requirements. However, the currents must be seven times higher for a given load power, compared to a 350-Vdc distribution approach. For the G-series and z900 servers, the distribution voltage has been standardized to 350 Vdc. The distributed loads in these servers are generally greater than 2 kW. Using a 48-Vdc distribution voltage would require the distribution cables and branch protection devices to carry currents greater than 50 A. This fact

[3] RS422 is an Electronic Industries Alliance specification that deals with data communication. Data rates of up to 100Kb/s and distances up to 4000 ft can be accommodated with RS422. RS422 is also specified for multidrop (party-line) applications in which only one driver is connected to, and transmits on, a "bus" of up to ten receivers.

alone ruled out the use of 48 Vdc or any other low voltage, since the physical bulk of the required components and cables would necessarily have enlarged the system footprint. An additional advantage of the 350-Vdc distribution system is that it becomes a relatively simple matter to integrate other types of equipment into the power-distribution and -control system. This proved especially important when redundant, high-capacity refrigeration units had to be incorporated into the higher-performance models. All z900 processor logic is cooled by refrigeration to a chip-junction temperature calculated to be 0°C. This would have been impractical in a system employing a distribution voltage in the SELV range.

Distributed converter assembly (DCA)

DCAs transform 350 Vdc to the regulated low voltages required by the CEC and the I/O midplane circuit boards. Redundancy is achieved by having $n + 1$ converters, each of which is powered by and has communication links to two BPCs, one in the A-side bulk power assembly subsystem and the other in the B-side bulk power assembly subsystem, as shown in Figure 8. The overall connection configuration of DCAs is shown in **Figure 10**. The Ethernet connections into the converter, coming from the PDUs, terminate at the cage controller. The cage controller is the service interface into the CEC and I/O logic boards. It also provides a communication path to the power-control microprocessor subsystem within the converter, which continuously monitors the status of each of the power stages in the converter. The reference voltages for the power converters are located in the power-control microprocessor, permitting the output voltage of any of the multiple power stages to be adjusted via a digital-to-analog converter. Hardware protection thresholds such as overvoltage may also be set via the power-control microprocessor.

The DCAs contain two types of dc–dc converters: One type has output current less than 60 A; the other type has output current greater than 60 A. The circuit selected for output currents below 60 A is the dual-switch forward converter switched at 300 kHz [14(b)]. It provides good power density and is cost-effective in the current range below 60 A. The power trains use conventional wire-wound magnetics mounted on printed-circuit boards using pin-in-hole technology.

The circuit selected for output currents above 60 A is a 300-kHz phase-shifted full-bridge converter with a current doubler on the transformer secondary [17]. As in the bulk power converter, the phase-shifted-bridge, zero-voltage transition topology [**Figure 11(a)**] was selected for its high efficiency and predictable performance over a wide range of loads. The current-doubler output section lends itself well to a planar transformer design, since it permits a single flex circuit to be used as the transformer secondary

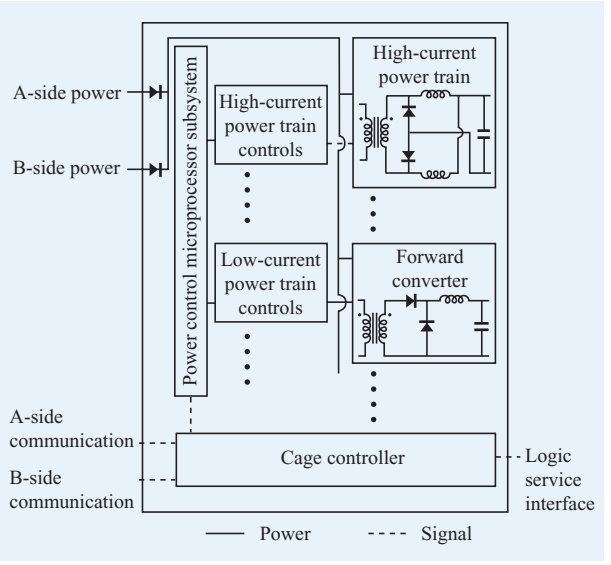

Figure 10

DCA block diagram. A DCA book (Figure 4) can house multiple phase-shifted, full-bridge high-current power trains and low-current forward converters.

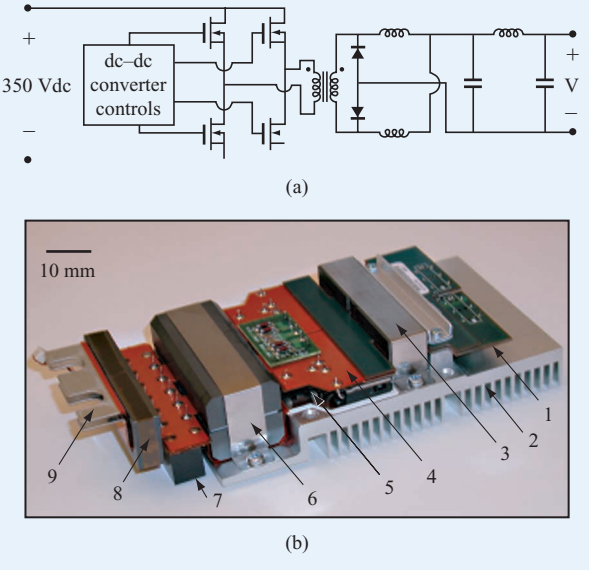

Figure 11

DCA power train: (a) Circuit schematic. The output voltage V is fed to the CMOS logic, I/O, and memory circuits. The last output capacitor shown is not on the power train; it is on the card on which the power train is mounted. (b) 1) Transformer primary; 2) heat sink; 3) transformer core; 4) transformer secondary; 5) rectifiers; 6) first-stage inductors; 7) filter capacitor; 8) second-stage filter inductor; 9) output terminal.

and also as the conductor for the output filters. The current doubler also makes greater use of the output rectifiers, which gives it a slight increase in efficiency over a conventional half-wave output section. Efficiency is typically 67% for the 1.7-V level provided to the z900 server processor.

The DCAs are plugged directly into the CEC midplane board, where space is at a premium. The thickness of the DCAs (Figure 4) had to be limited to the available space. For this reason, the high-current power train shown in **Figure 11(b)** is planar in design. The power transformer primary is contained on a pair of four-layer polyimide circuit boards. The high temperature of the transformer necessitates the use of polyimide circuit boards with a high glass-transition temperature. The transformer secondary is a 1-mm-thick copper flex circuit that forms a single turn sandwiched between the two primary boards. The overall height of the finished transformer assembly is 18 mm. The transformer is attached to a heat sink, along with the output rectifiers, making the height of the completed power train 38 mm. The transformer secondary flex circuit, which forms the positive-output conductor of the power train, is used to attach standard TO-247 rectifiers to the transformer secondary as well as filter capacitors for the regulator. The negative-output terminal is a second flex circuit that originates at the anode of the output rectifiers and continues through the output filter. The output inductors are single-turn and are formed by placing gapped ferrite cores around the flex circuits. The high-current power train, which operates at output currents to 250 A, connects to the MOSFETs that drive the primary via a board-to-board connector; the high-current flex circuits are bolted to the output power board. Each of the low- and high-current power trains contains a dedicated control card packaged as a separate assembly. This creates a more modular design and increases testability during manufacturing of these complex power supplies. For the CEC cage, power-train redundancy is achieved using three converters; for the I/O cage, power-train redundancy is achieved by using two converters.

Current sharing for the point-of-load dc–dc converters is accomplished by the forced-droop method [16]. The amount of voltage droop permitted from zero to full load is designed to be 0.8% of the nominal output voltage. The control circuits for each power train emulate a small but very tightly controlled dc output resistance and Thevenin equivalent voltage. When tied in parallel, the output currents of the power trains naturally share current, since they are all operating on the same load line. The microprocessor subsystem ensures accurate current sharing by making minor adjustments to the regulation reference of the control circuits.

The cooling system motors of all z900 servers run off the 350-Vdc bulk power bus. For both the blowers and the modular refrigeration unit, the motors are three-phase ac motors driven by a pulse-width-modulated motor control subsystem. The motor control subsystem contains a microprocessor that can receive commands and report status back to the primary power-distributor unit via an RS422 communication link. All cooling hardware is fully redundant; should one of the blowers fail, the remaining blowers can be speeded up to ensure that proper cooling is maintained for the load.

The power thermal subsystems of z900 servers have fulfilled their objective of providing very high levels of availability. An average server can expect an outage due to the power thermal subsystem less than once every 1500 years.

MCM internal cooling

As integration at the chip level continues, with large increases in the number of circuits, circuit density, and functionality, effective heat removal from the devices is of paramount importance in ensuring device performance and reliability. The critical parameter for device performance and reliability is the chip junction temperature, T_j, which is related to the heat-transport resistance (R_{int} and R_{ext}) and the chip (P_{chip}) and module (P_{module}) powers as follows [18]:

$$T_j = \Delta T_{cj} + \Delta T_{fh} + P_{chip}(R_{int}) + P_{module}(R_{ext}) + T_{amb},$$

where ΔT_{cj} is the temperature drop from the junction to the back of the chip, ΔT_{fh} is the temperature adder due to ambient fluid heating at the module, and T_{amb} is the ambient temperature of the cooling fluid.

The term *MCM internal cooling* refers to the heat path from the chip to the mechanical housing, or hat, of the MCM; it typically includes a number of resistances in series and/or parallel that add up to R_{int}. The hat spreads the heat from the module components to the external cooling means, such as a cold plate, a finned heat sink, or an evaporator of a refrigeration system. The hat also provides support to the next-level assembly and protects the module components from handling damage and corrosive elements of the ambient. By definition, R_{ext} includes the resistance from the hat to the ambient, including the interface resistance between the hat and the external cooling means. The following discussion pertains to module internal cooling designs and therefore to R_{int}.

Within IBM, continual development over decades has resulted in the adoption of various cooling schemes to meet the changing demands of the MCM designs. In the 1970s, low-powered (about 9 W) MCMs were cooled by air–water heat-exchanger schemes (IBM 3033 system) [19]. For high-powered (300 W) MCMs, a direct-immersion cooling scheme ("liquid encapsulated module," or LEM)

was qualified which involved immersing the chips in a fluorocarbon liquid, utilizing liquid boiling to transport heat from the chips to the internal fins on the hat, and then exchanging the heat with chilled water circulating through an externally mounted cold plate [20].

In 1981, a mechanical design called the thermal conduction module (TCM) was introduced in the IBM 3081 system [18, 21]. In the TCM design, individually spring-loaded aluminum pistons contacted the back of each chip, and the multiple pistons and springs were housed in cylindrical holes provided in the aluminum hat. The springs behind the pistons allowed for differences in joined chip height and accommodated chip tilt. Heat was conducted away from the back of the chip through an interface fluid layer to the piston and then, through another interface fluid layer, to the hat. To minimize the interface resistance [$R = gap/(conductivity \times area)$] from chip to piston and piston to hat, the gaps were made small, and a high-thermal-conductivity gas (helium) or an oil was employed as the thermal interface medium. A hermetic seal was provided. The pistons, in the case of aluminum pistons, or the hat, in the case of copper pistons, was anodized to maintain electrical isolation of the chips. Details of the TCM design and a one-dimensional thermal analysis are available in [18] and a three-dimensional numerical thermal analysis and empirical verification are described in [21].

The TCMs in 1990, in the IBM S/390 system, were larger and contained 127.5-mm × 127.5-mm substrates with 6.5-mm × 6.5-mm chips. These TCMs incorporated many changes such as copper pistons, narrower gaps, and oil as the interface medium in order to meet the higher-performance design challenges [22]. The IBM mainframes incorporating this design included the lower-priced air-cooled ES/9000 Model 320 computer and the higher-performance water-cooled ES/9000 Model 900. The air-cooled TCM was capable of cooling 10 W at chip level and 600 W at module level. The water-cooled TCM in ES/9000, introduced in 1990, was capable of cooling 27 W at chip level and 1800 W at module level and had an internal thermal resistance, R_{int}, of 1.75°C/W on 6.5-mm × 6.5-mm chips.

In the 1990s, as IBM began to phase in the CMOS-based servers, the cooling demands changed again. The chips became larger, 9.5 to 18.3 mm^2 in size, and more closely spaced on the substrate. The large chip size and gaps as narrow as 1 mm between the chips made the piston-in-hole design impractical. The changing marketplace also put more emphasis on design simplification, shortened development cycles, and cost reduction, leading to another shift in the cooling design. A thermal-paste-based cooling design (flat-plate cooling, or FPC) with a reduced number of interfaces and parts was introduced in 1995 in the G2 server and has been used in all subsequent G-series servers, including the z900 server. The design uses a layer of high-thermal-conductivity paste between the chips and a flat-plate hat. Heat is conducted away from the back of the chip through a thin layer of the paste to an aluminum or copper hat. The compliant paste accommodates chip tilt and differences in joined chip height. The thermal path (R_{int}) is shortened, and the cooling hardware is simplified. An air-cooled finned heat sink, or an evaporator of a refrigeration system, is employed to extract the heat from the module hat. A photograph of a sectioned flat-plate cooling module is shown in **Figure 12**.

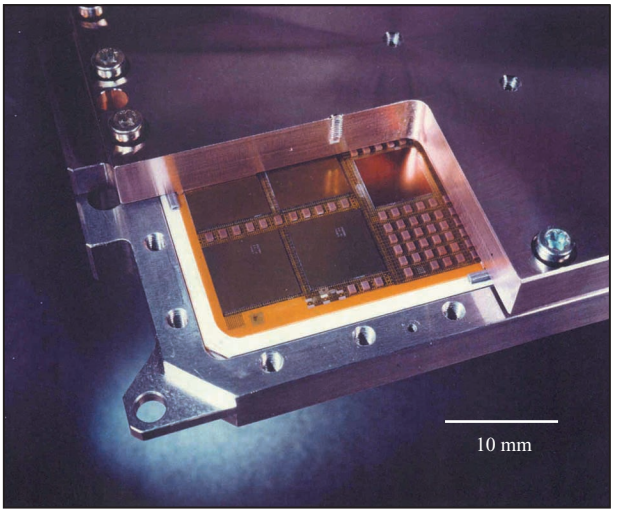

Figure 12

Flat-plate cooling module.

In the CMOS MCM designs, the chips are often "brick-walled," with very little spacing between them. Also, many decoupling capacitors are introduced between the chips. The flat-plate cooling (FPC) design allows flexibility of chip size, shape, layout, and spacing without drastic changes in the cooling hardware. This design flexibility is very important, as new G-series servers have been introduced every 12 to 18 months with different chip sizes and layouts. Earlier TCM designs, if at all practical, would have required a new set of cooling hardware for every design cycle. Also, since the FPC design did not require small (micrometer-wide) gaps, fine-tolerance machining and particle contamination concerns, with associated clean-room assembly requirements, were eliminated. Hermetic C-ring sealing from earlier TCM designs was retained. The module could easily be disassembled and the thermal paste cleaned with common solvents to allow module rework. Individual chip cooling could be customized easily, if needed, by varying the paste gap through pedestal height differences on the hat, or by using

Table 3 TCM and FPC MCM design comparisons.

	IBM 3081	*ES/3090*	*ES/9000 Model 900*	*G2*	*G3*	*G4*	*G5*	*G6*	*z900*
Year	1981	1985	1990	1995	1996	1997	1998	1999	2000
Module power (W)	300	600	2000	209	283	1100	650	800	1260
Chip power (W)	4	7	27	3.6–12.8	5.1–11.6	15–54	10–33	10–36	4–39
Chip size (mm)	4.5	4.5	6.5	9.5–14.9	15.75	14.6–17.3	12.7–17	12.7–16.4	12–18
No. of chips	118	121	121	36	34	30	29	31	35
Internal cooling	Aluminum pistons, He	Aluminum pistons, He	Copper pistons, oil	ATC 2.8	ATC 3.8	ATC 3.8	ATC 3.8	ATC 3.8	Low temp. ATC 4.5
Hat material	Aluminum	Aluminum	Copper	Aluminum	Aluminum	Copper	Copper	Copper	Copper
External cooling	Water	Water	Water	Air	Air	Air or refrig.	Air or refrig.	Refrig.	Refrig.
Substrate	90-mm alumina	127-mm alumina	127-mm glass-ceramic	127-mm alumina	127-mm alumina	127-mm alumina	127-mm glass-ceramic	127-mm glass-ceramic	127-mm glass-ceramic

thermal pastes of different conductivities. The chip electrical isolation requirement was met by the electrically insulative thermal paste.

Through the G6 server, the modules used a standard IBM-developed and -manufactured advanced thermal compound (ATC), called ATC 3.8, with a nominal thermal conductivity of 3.8 W/m·K. The thermal compound with high-solids content is a highly viscous, phase-stable solid–liquid dispersion that is qualified for junction temperatures to 105°C and second-level packaging temperatures to 240°C. However, the z900 server module is cooled by refrigeration and has an evaporator temperature of −20°C. To ensure compliance of the thermal paste at these operating temperatures, a new low-temperature advanced thermal compound with a nominal thermal conductivity of 4.5 W/m·K was developed.

The G-series server CMOS modules have ranged in chip power from 3.6 W to 45 W and in module power from 200 to 1300 W. The chip sizes have ranged from 9.5 to 18.3 mm^2. The internal thermal resistance, R_{int}, ranged from 0.12 to 0.45°C/W. **Table 3** compares the design parameters of different TCM and FPC modules.

In the FPC design, thermal-paste-gap control is important, and various stack-up tolerances must be taken into account. These tolerances are accommodated by using shims or an adjustable spacer to achieve the desired paste gap [23]. The critical challenges in a successful FPC design are the ability to retain the thermal compound in the gap between the chip and the hat and maintaining good chip coverage. These challenges were met by controlling factors such as paste adhesion, voids in the paste, and paste "pumping" resulting from the power-on–off cycling.

Modular cooling unit

With the change from bipolar to CMOS microprocessor circuit technology, the S/390 system changed from water-cooled mainframes supporting only the OS/390* operating system to air-cooled servers supporting OS/390, AIX*, and Linux**. The large servers are valued for their performance, reliability, serviceability, and bandwidth. The goal of retaining these attributes, while gaining the cost, power, and space efficiencies offered by CMOS technology, was achieved by integrating the cooling technology with the overall CMOS-based server design. Over the past six years and six generations of mainframe servers, from G2 through z900, cooling has been an integral part of the success of CMOS technology in S/390 servers, enhancing both the reliability and the processor performance.

The original S/390 servers were 19-in. racks with limited capability. Starting with the G2 server, the I/O rack width was increased to 24 in., while the CEC function was retained in 19-in. racks. During this period of the G2 and G3 server generations, lasting until 1996, customers relied on bipolar mainframes for many high-end enterprise-wide applications, with the high-end promise of CMOS technology still in the future. Air cooling in G2 servers was achieved by using 8.5-in.-wide forward-scrolling blowers. In the G3 and G4 servers, the CEC racks were widened to 24 in. The widened CEC racks made it possible to accommodate the modular refrigeration units (MRUs) introduced in G4 to cool the CMOS chips down

to 45°C (calculated), thereby providing CMOS technology for the first time with performance exceeding that of the fastest bipolar mainframes.

Cooling in the CMOS microprocessor chip is achieved by flip-chip-mounting the chips on a 127-mm × 127-mm MCM substrate capped with a metal hat. A custom thermal paste bridges the submillimeter gap between the silicon chips and the hat. Thermistors surface-mounted on the substrate provide thermal protection in the event of a loss of cooling. The z900 127-mm × 127-mm MCMs generate up to 1400 W of heat, which is removed by a copper evaporator.

The G2 and G3 servers had only air-cooled models. The G4 and G5 servers had both air-cooled and refrigeration-cooled models. From G6 onward, all server MCMs have been refrigeration-cooled. The air-cooled-only versions of the G2 through G5 servers had the entire contents of the CEC cage cooled by a single pair of 10.5-in. forward-scrolling blowers that used the 350 Vdc directly from the bulk power assembly subsystem. The blowers were capable of both high pressure drop and high airflow rates. Care was taken to ensure that the logic and memory received the coolest available air, as they are fundamental to the speed and reliability of the system, whereas the DCAs received the exhaust air from the front-side bulk power assembly. As shown in Figure 2, the 10.5-in. 350-Vdc blowers were physically located beneath the logic and front-side memory, midway between the components being cooled, to minimize air leakage and gain two decibels (dB) noise attenuation. The use of just two blowers to cool the entire CEC and power system contrasts sharply with the banks of fans used in many competing server systems. Speed control is used to adjust airflow on the basis of room ambient and whether any of the power or cooling components have failed, thus delivering reasonably steady temperatures without excessive noise. Directly below the CEC cage is the I/O cage, which is cooled by two forward-scrolling blowers of the same design and part number as those in the CEC cage. With lower power levels in this cage, the blowers typically run at slower speeds than in the CEC cage, resulting in a quieter machine. To further reduce noise levels, the server frame covers have been designed with acoustic attenuation of roughly 10 dB while keeping the air-pressure drop below 0.05 in. of water. As in the CEC, the blowers in the I/O cage are equipped with recirculation flaps to enable concurrent repair while the machine is running.

In the G4, the first IBM server to be cooled by refrigeration, the MRU utilized a small, efficient rotary compressor driven by the 350-Vdc output of the bulk power assembly subsystem. The compressor developed adequate refrigeration over the full range of environmental conditions of 10 to 35°C ambient at 0 to 7000 ft elevation above sea level. Problems encountered in the prototype G4 server, including "hunting and seeking" instabilities from the MRU expansion valve, developing a suitable evaporator to attach to the MCM with thermal feedback controls, and vibration resonance in the MRU copper tubing, were quickly resolved. More details of the MRU developed for the G4 server can be found in [24]. Like other functional components of power, packaging, and cooling, the MRU has $n + 1$ redundancy. Two MRUs were placed between the CEC cage and the I/O cage below, increasing the A-frame height to 42 U. By utilizing the exhaust air from the I/O cage to cool the condenser in the MRU, savings were achieved in both space and cost.

The G4 server lowered the chip-junction temperature to a calculated value of 45°C. With each succeeding-generation server, IBM lowered the junction temperature, thereby improving system performance and reliability. CMOS circuits can be switched ~1.4% faster for each 10°C lowering of the junction temperature.[4] For instance, consider the current z900 series MCM. If the lowest possible cycle time of a hypothetical air-cooled z900 server MCM is compared with that of an identical MCM that is MRU-cooled, the refrigerated z900 server MCM posts a 10% performance advantage in a 22°C room ambient.

In addition, because the chips in the z900 server MCM normally operate at 0°C, MRU-cooled circuits can be operated reliably at higher voltages. For the z900 server chip technology, the processor voltage for air-cooled chips is 1.5 V, but when cooled to 0°C, the processor chips can be operated reliably at 1.7 V. Since cycle time is nearly inversely proportional to voltage, a typical MCM at a given temperature would run more than 10% faster at 1.7 V than at 1.5 V [5].

For four server generations, IBM has successfully applied refrigeration technology to enhance its chip technology and deliver faster, more reliable systems. Whether the cooling paradigm shift that occurred in G4 will continue in future IBM enterprise servers will depend on system requirements. Across the server industry, power densities are increasing at all packaging levels. At the chip level, the power density is currently 40 W/cm^2 and going higher with time. At the module level, power densities are climbing because of both the increasing chip power and the opportunity to reduce substrate costs by using smaller, more densely packed substrates. At the frame level, more function and more processors provide the customer more value in a given footprint. Higher chip and module power densities favor refrigeration (or water-cooling) solutions, whereas higher frame power densities favor the smaller space needs of air-cooling components.

Cost is another key factor in deciding between refrigeration and air cooling. Refrigeration hardware is an

[4] A. Sutcliffe, IBM, Poughkeepsie, New York, private communication, 1998.

order of magnitude more expensive than the blowers used for air cooling. However, refrigeration possesses two advantages over air cooling: Refrigeration can reduce logic cycle time, and it can remove more heat flux. In many applications, the additional cost of refrigeration over air cooling is the least expensive way of enhancing system performance.

Another factor in deciding between refrigeration and air cooling is the issue of single-chip modules (SCMs) compared to MCMs. Air cooling is more readily distributed across numerous SCMs, with heat sinks placed on each module and airflow properly directed. In G4 through z900 MCMs, most of the logic function has been contained on a single large (127-mm × 127-mm) MCM. This has enhanced the ease of designing and controlling the evaporators, especially those that operate at temperatures as low as −20°C in the z900. At such temperatures, significant design and control code effort is required to eliminate condensation concerns, and having only a single large MCM to cool has been helpful. Condensation in the z900 is avoided by sealing the MCM hardware and its evaporator in a metal enclosure that contains desiccant [25]. The midplanar board surface opposite the side on which the MCM is plugged is heated to avoid condensation. All evaporators contain two separate refrigerant loops, one for each MRU, which provide redundant cooling. Developing this type of system to cool numerous modules with different heat loads would have been difficult, especially at temperatures below the dew point. Finally, power consumption favors air cooling. The z900 server MRU consumes about 10% of the total system power.

Electromagnetic compatibility

Electromagnetic compatibility (EMC) requires that a system be properly designed in order to perform reliably in its intended environments without causing interference with other systems. Electromagnetic compatibility is achieved by limiting conducted and radiated emissions and by improving the system immunity to radiated electric field, electrical fast transients and burst, electrostatic discharge, power-line transients including lightning strike, and conducted radio frequencies. Some commonly used EMC terms and effects are described below.

1. Conducted emission limits govern the maximum permitted noise voltage level that the system is allowed to conduct to the outside through the ac-line cords over certain specified frequency ranges. This requirement brings about the need for front-end EMC filtering upstream of the bulk power converters, miscellaneous intermediate converter-stage filtering, and shielding of the internal power-distribution cables against system noise pickup.
2. Radiated emission limits place an upper bound on the electric field strength that the system is allowed to emit through the cover seams and the external I/O cables into the airwaves over specific frequency ranges. Naturally, this requirement has a strong influence on the system shielding objective as well as internal noise-source suppression schemes.
3. The radiated electromagnetic susceptibility standard is an immunity requirement specifying minimum system tolerance against electric field bombardments over a specified frequency range. This requirement raises the need to consider internal shielding at the cage and signal cable levels, since with the system in maintenance mode the covers are generally fully open, making the system most vulnerable to electromagnetic interference (EMI).
4. The electrical fast transient and burst immunity standard specifies the minimum noise pulse level of a given duration and repetition rate that a system must be able to tolerate from utility company load switching or other power-relaying events. The transient noise pulse can be conducted either directly through the ac-line cord or via indirect coupling on the signal cable shields or conductors. Areas that require special EMC consideration are the ac input filtering, power-supply internal layout, cable shielding for specific sensitive applications, and software recovery.
5. Electrostatic discharge (ESD) can have a direct effect on the system reliability. ESD-sensitive circuits must be protected with shielding, proper cable selection, and software recovery in the event of failure.
6. The term *conducted immunity* refers to the minimum radio frequency (rf) level that a system must be able to tolerate without degradation. Conducted immunity is determined by subjecting the system to power-line-conducted noise of a given voltage over a specified frequency range. The test involves both ac lines and signal cables. Power-supply filtering and signal cable selection have a direct bearing on the outcome of the test.

Proper EMC design requires careful analysis of all potential problems, with particular focus on known problem areas. For the z900, the known problem areas were as follows:

1. Dense I/O packages increase noise coupling problems. For instance, the ESCON book in the z900 server contains 16 channels compared to four for the previous design.
2. Higher I/O speed introduces a higher radiation potential, since the electric field amplitude is generally proportional to the square of the frequency [26]; the intersystem channel (ISC) card runs at twice the speed

of its predecessor at over 2 Gb/s. Although the ISC card does not have external copper cables, its harmonic frequencies can couple to adjacent I/O adapter cards with copper cables.
3. Higher data-transfer rates on ISC and other I/O adapters result in greater sensitivity to ESD and other transient noise coupling.
4. Higher EMI is also expected of the system clock harmonic frequencies, since the STI cables that distribute clock signals to the I/O boards contain harmonic frequencies well over 1 GHz.
5. Higher power consumption results in the need for more power-supply and air-moving-device distribution cables, which act as noise-source antennas, increasing the broadband radiation potential.
6. Implementation of the cage controller Ethernet links inside the power supplies pushes the broadband noise floor even higher; the data-switching activities of the Ethernet links produce a noise spectrum roughly coinciding with that of the power supplies.

z900 EMC design strategy

From the above list, one can see that power-supply noise and high-frequency I/O and system clock noise are the areas requiring special attention. The only way to control I/O and system-clock harmonic frequencies of 1 GHz and above is through source suppression, because there is no easy way to eliminate these frequencies once they appear at the I/O adapter tailgate and cables as common-mode current. (The term *tailgate* is commonly used for the I/O interface port area.) Therefore, source suppression at the logic card and board level, described below, is the approach used to reduce EMI.

- *Printed-circuit card and board Faraday cages*: The card- and board-level EMI is trapped within Faraday cages formed in the printed-circuit cards and boards using "peripheral stitching." A Faraday cage is formed by making the outermost copper planes the ground planes and connecting ground planes along the entire perimeter of the printed-circuit card or board using copper-plated vias. The "stitching" of the ground planes using peripheral vias embeds a shielding structure within the printed-circuit card or board that attenuates the EMI emission level and also protects the internal circuit nets from outside interference.
- *Creation of high-frequency resonance structures in printed-circuit cards and boards*: The card and board resonant frequencies are shifted to much higher values (>10 GHz) so that they do not interfere with the system and I/O clock frequencies. This is accomplished by the placement of at least one ground-plane-connecting via per square inch of area everywhere on the printed circuits. This internal-ground-plane-stitching scheme eliminates any long ground structure that may resonate at the system or I/O clock harmonic frequencies.
- *EMC evaluation plan*: The z900 server is the first large-scale system with all of its logic cards and boards qualified using the EMC evaluation plan (EMCEP). EMCEP is a knowledge-based tool that consists of EMC checking rules based on a large collection of EMC design experience obtained from many IBM sites. The tool checks for clock module, I/O module, and critical net placement violations; it also checks for net termination problems, differential pair mismatches, and many other features related to wiring-rule violations. Even with the aid of the EMCEP tool, tradeoffs are often made in cases of less-critical EMC rule violations to avoid the havoc associated with redesign.
- *External clock line EMC treatment*: Selecting the proper grounding scheme for the system and I/O clock distribution cables is extremely important in preventing radiated EMI problems. STI cables utilize the VHDM connectors, which provide 360° termination to the logic books, thus realizing a low-impedance ground path to the chassis. Slight imperfections in the logic book sheet-metal housing next to the VHDM connectors for the STI cables have shown up during an early evaluation as a 15-dB problem.
- *EMC problems with power supplies and related cables*: Power-supply broadband noise has been an increasing threat to radiated EMI compliance with the introduction of each server. The z900 server is no exception, since it consumes more power than its predecessors. An effort to internally filter and clamp the broadband noise sources such as the power MOSFETs and the IGBTs used as the power-supply switches was offset by the introduction of cage-controller Ethernet links. The EMI emission causing broadband noise was reduced by as much as 14 dB by designing a new UPIC cable with full braid over both power and signal wires. The added shielding effectiveness provided by the new improved UPIC cables helped to keep the broadband EMI level in compliance despite the increased number of power supplies and air-moving devices.
- *Cover-level EMC*: Large servers have relied on the external covers as the ultimate EMI suppression mechanism because of the number of signal and power cables inside the servers. A z900 server relies on its external covers for compliance to a lesser extent than its predecessors because of the direct I/O cable connections to the logic cage which bypass the conventional tailgate enclosures found in the earlier systems. Without the tailgate enclosure separating the server from the outside world, any common-mode current induced on the outbound copper I/O cables by the internal noise will cause EMI emission directly to the outside, bypassing the external covers. For this reason, broadband EMI

noise suppression on the UPIC cables by means of cable-shield termination at both ends is extremely important. If broadband noise escapes through the UPIC cables, the cables can also function as receiving antennas, bringing unwanted external EM disturbance to the power electronics and the miscellaneous analog sensors inside the power supplies. The design objective for shielding effectiveness by the external covers is a minimum of 20 dB at 1 GHz, and test results showed that the covers with the latest nickel copper gasket delivered the required shielding.

- *Cage- and frame-level EMC*: The frames of the z900 server are coated with a conductive paint (<50 Ω/□). The paint, used since the mid-1980s, delivers greater than 40 dB shielding at 1 GHz. The cover system, consisting of frames, new gaskets, and the external covers, provides a minimum of 40 dB attenuation at 1 GHz. The CEC and the I/O cages rely on the formation of a Faraday cage around the individual logic books; each logic book in its metal container contacts an adjacent book through a conductive metallic cloth gasket. The gasket delivers greater than 60 dB attenuation at 1 GHz. Overall, the shielding at the frame and cage levels is quite good. The remaining EMI escape path is the opening between the system frame bottom and the raised metal floor. To prevent EMI emission through this large gap, a set of "EMC skirts" was designed to block the EMI of the I/O cables. The EMC skirts were first introduced in the G2 server to combat broadband power-supply noise. They are equally effective for discrete frequencies and, together with the external covers, form the last line of defense against EMI emission.

Acoustical noise control

Acoustics and noise control is an integral part of packaging a high-performance server or workstation. Since most of the noise in modern information technology products is associated with the air-moving devices needed to cool the electronics, acoustics helps determine the speed at which these devices can operate and still meet the appropriate noise limits. Balancing the conflicting requirements of thermal design with those of acoustical design has become more difficult over the past few years, since heat loads have grown nearly exponentially and the required amounts of airflow have increased accordingly. Furthermore, combining one or more modular refrigeration units with several relatively large air-moving devices makes the volume associated with the noise-generating cooling components a significant percentage of the overall volume of the high-end server. Noise-control engineering is always applied first to the sources of noise themselves, but for large IBM servers such as the z900, acoustical treatments to the package are also necessary. Special acoustical front and rear doors allow the servers to pass their acoustical specifications, but these add even more volume to the package and may affect customer floor space requirements.

Acoustical noise has several adverse effects on people. On the physiological side, such effects include hearing damage and hearing loss. On the psychological side, they include annoyance, interference with speech communications, impairment of performance, and stress. Although information technology equipment may have been loud enough in the distant past to cause hearing damage, the primary concern to IBM today in its noise-control efforts is to prevent customer annoyance. IBM is striving to lower the noise levels of its products while at the same time studying and identifying the psychological aspects of particular noises that most contribute to annoyance.

Measurement and rating of noise

Acoustical measurements are taken for two principal reasons: to characterize a source and to characterize an environment. The distinction between noise source and environment must be kept in mind to avoid confusion in interpreting or defining noise-level requirements and specifications. A particular noise source emits a certain amount of sound power; and standardized acoustical measurements are taken to determine its sound power level. The sound *power* level characterizes the source. On the other hand, sound *pressure* level measurements may be taken at a point in an ordinary room, such as the user's position in front of a workstation, to determine the level at that point in the environment. The sound pressure level characterizes the environment. It is not a good measure of the noise level of a product itself, because 1) there may be other noise sources in the room contributing to the sound pressure level at the measurement point; and 2) the boundaries and obstacles in the room reflect the sound emitted from the product back to the measurement point. Manufacturers want to lower the sound *power* levels of their products (regardless of the environment for which they are destined), and customers or employees want to lower the sound *pressure* levels to which they are exposed, regardless of the source or sources of the noise.

In order to characterize the spectrum of a particular noise, the levels are usually specified or reported in terms of octave-band or one-third octave-band levels. Alternatively, a single-number rating is commonly used to express the overall level of the noise. A standardized frequency weighting, called *A-weighting*, has been internationally agreed upon and incorporated into almost all sound-level-measuring instruments. This weighting reflects the nonuniform sensitivity of the human ear by attenuating very low and very high frequencies relative to the middle frequencies. A measurement made using

A-weighting will yield the A-weighted sound pressure level, L_{pA}, in decibels (dB), or the A-weighted sound power level, L_{WA}, in either bels (B) or decibels (dB).

With regard to standardization and regulation in the information technology industry, there are many national and international standards specifying noise measurement methods, both for sound power levels of machines and sound pressure levels in the environment [27, 28]. On the other hand, there are only a few standards or regulations specifying actual noise limits; the one most important to the IT industry is the Swedish Technical Standard 26:3 [29].

Sources of acoustical noise

Air-moving devices, such as centrifugal blowers, motorized impellers, and axial fans, represent the principal sources of noise in large servers. In addition to broadband random noise, air-moving devices may also radiate periodic discrete tones, usually associated with the blade passage frequency and its harmonics. Such tones are most prevalent in axial fans and can be very annoying, especially when several fans of the same type and speed are used. For this reason, the air-moving devices in the IBM z900 server are either centrifugal blowers or motorized impellers (blowers in the CEC and I/O cages; impellers in the BPA and MRU). There are standardized methods for evaluating discrete tones in noise emissions [27, 28]. IBM acoustical standards require that "discrete tone penalties" be applied to the measured noise-emission level when a tone is evaluated and classified as "prominent." Tone evaluations were conducted for the z900 according to international and national standards [27, 28], and the results showed no prominent discrete tones. In terms of the broadband noise, an empirical rule that generally holds true for air-moving devices is that the noise level increases as the fifth power of rotational speed, $50 \log_{10}(N_2/N_1)$ dB, where N_1 is the initial speed in rpm, and N_2 the final speed; that is, if the speed is halved ($N_2/N_1 = 1/2$), the noise level will decrease by 15 dB. Even a 20% reduction in speed results in a 5-dB reduction in noise level. Clearly, there is great benefit acoustically in slowing down air-moving devices whenever feasible.

Modular refrigeration units, recent additions to the acoustical noise sources in high-end IBM servers, have two primary sources of noise: a compressor unit and an air-moving device cooling the condenser. The compressor may also be a major source of vibration. If the compressor is not isolated properly, its vibrations can couple easily into the cage or covers and radiate as low-frequency noise, which is difficult to attenuate.

Control of acoustical noise

Noise-control engineering treats the problem as a system having three components: the *source*, the *path*, and the *receiver*. The receiver is important in defining noise specifications, of course, but from a product noise-control standpoint, efforts are focused on the source and the path. Receiver noise control may involve isolating the listener from the noise through the use of hearing protectors or special screens and enclosures.

Noise-control efforts are most effective when applied directly to the source, thereby reducing at the outset the amount of sound energy generated. Source noise control always starts with fully characterizing and understanding the mechanisms of noise generation in the system, but from there various approaches may be followed, depending on factors such as the cost of the treatment, the expected noise reduction, and even the time available for study, analysis, and experimentation. For the z900 server, the following source noise-control techniques were applied:

1. *Source selection*: Several decibels of "free" noise control can often be realized by an informed selection of the air-moving device at the outset, including the proper sizing for a particular application in terms of airflow rate and static pressure and the operation at its so-called point of maximum efficiency.
2. *Source location*: Prudent selection of board-level, cage-level, and frame-level layouts with respect to the primary noise sources is also a relatively low-cost approach to noise control. For instance, a design that embeds the air-moving devices deep within a cage or subpackage may result in noise levels several dB lower than a design in which the air-moving devices are located near the front or back faces of the frame. **Figure 13** shows the one-third octave-band sound power levels of a pair of 10.5-in. (267-mm) centrifugal blowers, running at 2000 rpm against roughly equivalent aerodynamic loads, in one case freely exposed to the room environment (mounted on a special test plenum) and in the other packaged well within the IBM z900 server CEC cage. Attenuations of 5, 10, and 15 dB are realized depending on frequency, with an overall A-weighted attenuation of about 6 dB.
3. *Unobstructed inlets for air-moving devices*: Nonuniform airflow and turbulence caused by inlet obstructions can greatly increase noise levels and accentuate the levels of annoying discrete tones. For the z900, the goal was to allow between one and two fan radii of free clearance at the inlet of the air-moving devices.
4. *Air-moving-device speed control*: As already mentioned, the noise from an air-moving device generally increases as the fifth power of rotational speed. However, because of thermal requirements, it is not usually an option to lower fan speeds arbitrarily to produce lower noise levels. One approach that has been used successfully by IBM, and on the z900 server, is a form of speed control in which the air-moving-device speeds

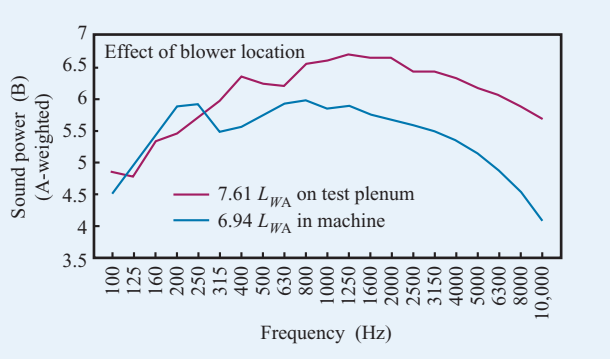

Figure 13

Measured sound power levels of a side-by-side pair of 10.5-in. centrifugal blowers running at 2000 rpm. In the upper curve, the blowers are fully exposed to the sound field; in the lower curve, they are installed in their application, deep within an IBM z900 server CEC cage. Note the natural attenuation afforded by proper source location. A measurement made using A-weighting yields the A-weighted sound pressure level L_{pA}, in decibels (dB), or the A-weighted sound power level L_{WA}, in either bels (B) or decibels.

are set at their minimum values for nominal operation and allowed to speed up for atypical conditions on the basis of input from various sensors and control circuitry. These high-speed conditions may include excessively hot rooms, very high altitudes, and certain temporary failure conditions within the electronics, e.g., loss of a power supply or an air-moving device. This approach allows most customers, those with the nominal conditions, to be exposed to the lowest noise levels.

5. *Vibration isolation and vibration damping*: Isolating the noise source from the structure on which it is mounted and applying damping treatments to the structure are sometimes necessary to minimize the transmission of mechanical energy. The radiation of structure-borne noise can be quite efficient, since the radiating structural element may couple better to the air than to the original source of vibration.

In addition to source noise control, the path from a particular noise source in the machine to the receiver can be modified in many ways to attenuate the sound. The principal elements of path noise control are acoustical barriers, acoustical enclosures, acoustical ducts, sound-absorptive materials, and acoustical covers.

1. Acoustical barriers are panels or other solid materials that prevent a portion of the sound energy on one side from being transmitted to the other. Barrier performance is rated in terms of the transmission loss, in decibels—the higher the transmission loss, the better. The solid external covers as well as the panels around internal components in the z900 server function as acoustical barriers.
2. Acoustical enclosures are a combination of barriers constructed around one or more noise sources in an attempt to contain the acoustical energy within the enclosure. Acoustical materials are usually needed within the enclosures to prevent a reverberant buildup of energy.
3. Acoustical ducts are used for redirecting rather than containing the acoustical energy. Redirection introduces further attenuation into the path of the sound transmission through the use of absorptive materials, bends, added length, changes in cross section, plenums, or barriers.
4. Sound-absorptive materials are used extensively in IBM server products, including the z900. The term *absorption* refers to the ability of a material to convert airborne acoustical energy into another form of energy, usually heat. The performance of an acoustical material is given as a function of frequency in terms of its absorption coefficient, a real number between 0 and 1, where 0 means no absorption (everything is reflected) and 1 means total absorption (nothing is reflected). The most common material used for sound absorption, and the one used in the z900 server, is urethane foam. Its absorptive characteristics are good; it is relatively inexpensive; and recent advances have reduced its flammability and toxicity and enhanced its long-term stability. However, in order to achieve high absorption at lower frequencies (often needed because of the spectra from the air-moving devices), the material must be relatively thick. In the covers of the z900 server, acoustical polyurethane foam is used in thicknesses up to 140 mm.
5. Acoustical covers are probably the most effective means of controlling the transmission of noise from the inside to the outside of the machine and lowering the overall sound power level of the product. The front and rear doors in the z900 server are examples of such acoustical covers. Acoustical covers are essentially a combination of the above elements, but because of the presence of needed openings for airflow, they are separate entities from a design standpoint. As opposed to the transmission loss from a barrier, the performance of an acoustical door is rated in terms of its sound power-level attenuation. This is defined as the difference in sound power level of the computer frame measured without the doors and with the doors. **Figure 14** shows the one-third octave-band sound power-level attenuation (in decibels) of the front and rear acoustical covers for the IBM z900 server. For comparison, the cover attenuation values for two of the IBM previous-generation servers (the G5 and G6) are also shown,

along with the "attenuation" (nearly zero) provided by a standard, nonacoustical, perforated-metal door such as that used on the IBM RS/6000* SP supercomputer. The number in parentheses is the average attenuation rating (AAR) for each cover set [30]. Note particularly how the demands for cover attenuation have increased steadily over the years because of the increased heat loads and attendant increases in airflow noise. With covers, each product measured about the same overall sound power level and met its acoustical specifications, but without covers the z900 server would be much noisier.

Acoustical noise control for high-end servers will become more and more challenging as processor speeds and thermal heat loads continue to increase in the future. Although there are some promising new thermal approaches on the horizon, the use of air-moving devices and refrigeration components is not expected to diminish in the near future. It is hoped that there will be breakthroughs in the acoustical materials arena to allow good absorption at low frequencies in a compact package. The future will also see a greater focus on psychoacoustics and sound quality as we attempt to better understand what customers want and expect in terms of low-noise products.

Shock and vibration

Shock and vibration testing and analysis of subassemblies and the entire system frame consist of endurance and ruggedness testing, shipping and packaging testing, operational shock and vibration testing, compliance testing, and stress screening.

Endurance and ruggedness testing of subassemblies comprises two tests: the thermal shock test and the shock and vibration test. These tests are performed on subassemblies to ensure the shippability of the entire system frame. The thermal shock test determines the susceptibility of subassemblies to thermal shocks associated with shipping. Subassemblies, removed from their packages, are subjected to five cycles of thermal shock, from −40°C to 60°C, to evaluate the effect of thermal-coefficient-of-expansion mismatch between the subassembly components. The thermal shock testing is followed by shock and vibration testing. The subassemblies should be able to withstand 100 g (gravitational constant), 3 ms, two half-sine pulses of shock along all three axes. Random vibration testing is also done to ensure the robustness of the subassemblies. The subassemblies should be able to withstand the random vibration test profile, consisting of vibration with a fixed amplitude of 0.01 g^2/Hz and frequency randomly varying in the range of 10–500 Hz for one hour per axis. These test conditions are chosen to

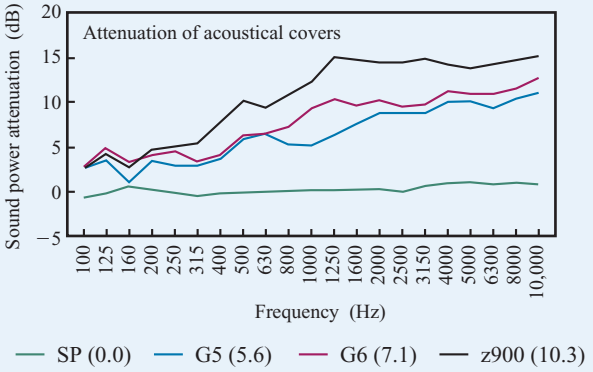

Figure 14

Sound power level attenuation provided by front and rear acoustical doors on the last three generations of IBM Enterprise servers, G5, G6, and z900, compared to a standard nonacoustical perforated-metal door such as that used on the RS/6000 SP supercomputer. Numbers in parentheses indicate average attenuation rating.

cover the worst-case shipping environment and are severe enough to uncover any manufacturing or workmanship problem such as cold solder joints and loose fasteners. The random-vibration test profile selected has proven effective in ensuring zero defects for a normal shipping environment. For critical components that show a high amount of displacement in the vibration testing, diagnostic testing is performed, using a half-hour sine sweep test at 0.5 g per axis from 10 to 500 Hz, to determine the transmissibility (output level/input level) of the components inside the subassemblies. If the transmissibility of a component is greater than 5, design enhancement effort is initiated.

- *Subassembly shipping test*: For a subassembly that can be shipped directly to a customer, either as a replacement part or an upgrade, the shipping shock and vibration test is conducted to ensure that the product will function properly at the customer location. For example, for a packaged product weighing less than 20 lb (9 kg), the packaged product must be able to withstand free fall of 36 in. (0.9 m) on all sides, corners and edges; and it should be able to withstand random vibration at the 1.04 g rms test profile shown in **Figure 15** for 15 minutes.
- *Frame-level shipping test*: The z900 server has a shipping environment unique to the industry. Within the USA, the systems are shipped on casters; for overseas shipments, pallets are used. Consequently, both palletized and nonpalletized versions must be tested. The IBM common-carrier random-vibration test and the

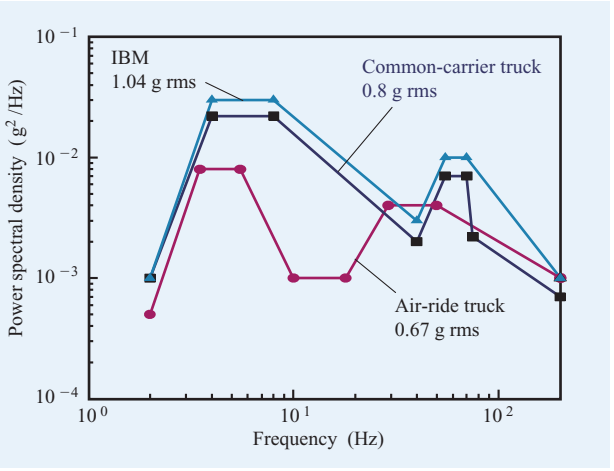

Figure 15

IBM random-vibration test profiles.

air-ride truck test are the primary tests to ensure the robustness of populated-frame products. The profiles for these tests were developed on the basis of actual shipping environment data. These tests have proven effective in ensuring zero defect for a normal shipping environment. The occurrence of shipping damage has been low, approximately two to three incidents per year, and is typically related to systems tipping over. Before the advent of the IBM common-carrier test, two endurance tests were in vogue: the sine sweep test and the sine dwell test. However, neither of these tests represents actual shipping environments. The sine sweep test is still useful for determining transmissibility, and it can therefore be beneficial during the early stages of design development. It uses criteria and test profiles similar to those of the endurance test. The sine dwell test can be damaging to the product, and it has been the experience within IBM that system frames and subassemblies designed to withstand the sine dwell test are often overdesigned from a structural perspective. With the availability of the random-vibration controller, the motivation for doing sine dwell and sine sweep tests disappeared.

- *Subassembly operational shock and vibration test*: All subassemblies, especially the hard drives, and their mountings are subjected to an operational shock and vibration test. The subassembly being tested is turned on and its functions monitored while it is being subjected to external vibration.
- *Frame-level operational shock and vibration test*: As part of the final test prior to customer general availability, compliance testing is performed. The system is powered on, and its functionals are monitored continuously while the system is subjected to low-level shock and vibration (lower than the subassembly test level) that simulates the actual operating environment. There have been no field failures of z900 servers attributable to external system vibrations such as the effect of rock blasting and train-passby vibrations. During hot-plugging of subassemblies such as power supplies and MRUs, high magnitudes of shock level (higher than 30 g), for short duration (less than 1 ms) can occur. There has been no problem related to mechanical shocks arising from hot-plugging.
- *Stress screening during manufacturing*: An application of shock and vibration testing technology that falls outside the design function is stress screening during manufacturing. Stress screening may involve a battery of stress tests such as thermal cycling, random vibration, and burn-in while the function of the subassembly is being continuously monitored. This test is effective for subassemblies. A frame-level test is not effective, since the frame will attenuate the vibration level being experienced by the subassemblies being tested. The manufacturing stress screening is conducted at the subassembly manufacturing locations. It has been proven statistically that by subjecting at least 207 consecutive production subassemblies to manufacturing screening tests, it is possible to detect latent component defects and design deficiencies inherent in the subassemblies [31]. All z900 power supplies and switcher subassemblies are subjected to manufacturing screening. This process has been quite effective in speeding up the bring-up of vendor manufacturing processes by quickly revealing manufacturing workmanship defects.

Earthquake simulation test

The shipping test evaluates the effect of vertical vibration on the system and its subassemblies and components. Seismic events can have a significant horizontal vibration component in addition to the vertical component. Seismic events lead to low-frequency vibration in the 1–50-Hz range horizontally. Fortunately, most critical components such as hard drives, microprocessors, and memory units have resonance frequencies higher than 100 Hz and therefore will not resonate during an earthquake. Seismic events will not affect system availability as long as the supporting frames, which have low resonance frequencies, can sustain the vibration energy imparted by the seismic events. Many IBM customers require adequate retention of rack-mounted systems to avoid personal injury and system damage. All high-end IBM systems have been designed and tested to meet internal IBM specification and external specification such as NEBS (Network Equipment Building Specification) for both raised-floor and non-raised-floor facilities [32–34].

Summary and conclusions

Challenges faced in powering, packaging and cooling CMOS microprocessors which continue to increase in circuit density and performance are discussed. The computer must operate within the acoustic and electromagnetic noise limits set by the various regulatory agencies, and it must be able to withstand mechanical shock and vibration, including rare events such as earthquakes. In addition to the performance challenges, there must be continuous improvement in quality and reliability.

The MCM technology chosen for the IBM servers provides dense and efficient packaging of CMOS processor chips with a high-bandwidth nonblocking crossbar second-level cache architecture. As the circuit speed and transistor density of the CMOS technology have increased, the MCM technology has kept pace by providing continuously improved performance as well as performance per unit cost. The second-level design and materials tradeoffs are becoming more complex because of the trend in electronic packaging toward more sophisticated first-level packaging containing more I/O connections. The density of the array of wiring and connections required to package the more complex logic is being achieved by reducing the dimensions of the printed-wire-board features and by adding additional PWB layers to prevent electrical crosstalk noise.

The uniqueness of the power subsystem of the IBM large servers, to which the z900 belongs, lies in its ability to accept a very wide range of ac-line voltages, its compactness, and its redundancy. The bulk power assembly accepts 200–480 Vac and converts it to a well-regulated 350 Vdc that is distributed via the UPIC cables to all of the point-of-load converters in the system. The UPIC cables can afford to have narrow cross sections because of the high voltage (350 Vdc) they carry. The UPIC cables also contain the signal wires used for the microprocessor control of the entire system. The point-of-load dc–dc converters plug directly into the midplane board they feed, eliminating bus bars. The power hardware has redundancy and is concurrently maintainable. On the average, a server can expect an outage due to a power thermal subsystem fault less than once every 1500 years.

Refrigeration improvements in each generation have lowered the processor chip-junction temperature. Refrigeration challenges have included the packaging of redundant MRUs without excessively affecting the server footprint and while avoiding moisture condensation. For each 10°C lowering of the chip temperature, for a given chip voltage, the processor performance increased. Lowering the chip temperature permits the chip voltage to be raised without impact on reliability, thus further enhancing the processor performance. Refrigeration also increases the power-handling capability of the MCM.

In the z900 system, the major EMC and EMI challenges were in the areas of power-supply noise and the high-frequency I/O and system clock noises. The power-supply broadband noise has been an increasing threat to radiated EMI compliance with each introduction of servers demanding more power. The broadband EMI, originating in the main power MOSFET switches and radiating from the UPIC cables, was kept within compliance level by improving the shielding around the cables. The I/O and the system clock noises were kept in check by source suppression because there is no easy way of eliminating them once they appear at the I/O adapter tailgate and cables as common-mode current. Source suppression included the use of a Faraday cage structure, the shifting of the card and board resonant frequency to higher values to avoid interference with the system and I/O clock frequencies, and the use of an EMC evaluation plan, which is a knowledge-based tool consisting of EMC checking rules based on a large collection of EMC design experience gathered at IBM over many decades.

Moving higher volumes of air to meet the new cooling demands has an unfortunate side effect: higher acoustical noise levels. Thus, another challenge in packaging a high-end server such as the z900 was to be able to cool it successfully but still have it meet its acoustical requirements. This was achieved through judicious component design, selection, and placement, coupled with acoustical treatment of the frame covers and doors.

Designing a structure that can sustain normal shipping, operational, and seismic events involves comprehensive shock and vibration testing and analysis of the components and the entire frame structure. The tests described in this paper have proven effective in avoiding mechanical defects and ensuring robust structural design.

Acknowledgments

This paper would not have been possible without the support and encouragement of Vincent Cozzolino, who was the director of the IBM Power, Packaging, and Cooling organization during the period from 1996 to 2000.

References

1. J. G. Davis, C. R. LeCoz, and J. J. Tomaine, "Circuit Board Repair and Engineering Change for BGA," *Proceedings of the Technical Program, Surface Mount International Advanced Electronics Manufacturing Technologies*, Surface Mount Technology Association, September 1996, Vol. 1, pp. 181–187.

2. G. Patel and T. Cohen, "Shielded High Performance Interconnect Technology Dramatically Increases Real Signal Density," *Proceedings of the WESCON Conference*, Los Angeles, 1997, IEEE Cat. No. 97CB36148, pp. 140–145.
3. J. M. Hoke, P. W. Bond, T. Lo, F. S. Pidala, and G. Steinbrueck, "Self-Timed Interface for S/390 I/O Subsystem Interconnection," *IBM J. Res. & Dev.* **43,** No. 5/6, 829–845 (September/November 1999).
4. *Racks, Panels and Associated Equipment*, Document No. EIA-310, Revision D, 1992, American National Standards Institute, Washington, DC 20036.
5. G. A. Katopis, W. D. Becker, T. R. Mazzawy, H. H. Smith, C. K. Vakirtzis, S. A. Kuppinger, B. Singh, P. C. Lin, J. Bartells, Jr., G. V. Kihlmire, P. N. Venkatachalam, H. I. Stoller, and J. L. Frankel, "MCM Technology and Design for the S/390 G5 System," *IBM J. Res. & Dev.* **43,** No. 5/6, 621–650 (September/November 1999).
6. G. A. Katopis, W. D. Becker, H. H. Smith, and H. Stoller, "MCM C/D Design for the CMOS Implementation of the S/390 System," *Proceedings of the 47th Electronic Components and Technology Conference*, IEEE Cat. No. 97CH36048, May 1997, pp. 479–485.
7. G. A. Katopis and W. D. Becker, "S/390 Cost Performance Considerations for MCM Packaging Choices," *IEEE Trans. Components, Packaging, Manuf. Technol*., *Part B: Adv. Packaging* **21,** No. 3, 286–297 (August 1998).
8. C. F. Webb, "S/390 Microprocessor Design," *IBM J. Res. & Dev.* **44,** No. 6, 899–908 (November 2000).
9. E. Cordero, F. Ferriaolo, M. Floyd, K. Grower, and B. McCredie, "A Synchronous Wave-Pipeline Interface for POWER4," presented at HOT CHIPS, Stanford University, August 15–17, 1999.
10. H. Smith, S. Kuppinger, P. Venkatachalam, and W. Becker, "Noise Verification Across Three Levels of Packaging Hierarchy for the IBM G5/G6 Mainframes," *Proceedings of the 50th Electronic Components and Technology Conference (ECTC)*, Las Vegas, May 21–24, 2000, pp. 754–749.
11. C. DeCusatis, J. Trewhella, and J. Fox, "Performance Comparison of Small Form Factor Fiber Optic Connectors," *IEEE Trans. Components, Packaging, Manuf. Technol*., *Part B: Adv. Packaging* **23,** No. 2, 188–196 (July 2000).
12. "Setting the Standard—IBM Power Systems," *Switching Power Magazine* **1,** No. 1, 12–17 (July 2000).
13. R. Ridley, S. Kern, and B. Fuld, "Analysis and Design of a Wide Input Range Power Factor Correction Circuit for Three Phase Application," *Proceedings of the IEEE Applied Power Electronics Conference and Exposition*, 1993, pp. 299–305.
14. R. W. Erickson and D. Maksimovic, *Fundamentals of Power Electronics*, Second Edition, Kluwer Academic Publishers, New York, 2001; (a) pp. 638–640; (b) pp. 1154–1159.
15. J. A. Sabate, V. Vlatkovic, R. B. Ridley, and F. C. Lee, "High-Voltage, High-Power, ZVS, Full-Bridge PWM Converter Employing an Active Snubber," *Proceedings of the IEEE Applied Power Electronics Conference and Exposition*, 1991, pp. 158–163.
16. B. T. Irving and M. M. Jovanovic, "Analysis, Design, and Performance Evaluation of Droop Current-Sharing Method," *Proceedings of the IEEE Applied Power Electronics Conference and Exposition*, 2000, pp. 235–241.
17. N. H. Kutkut, "A Full Bridge Soft Switched Telecom Power Supply with a Current Doubler Rectifier," *Proceedings of the 19th International Telecommunications Energy Conference (INTELEC 97)*, 1997, pp. 344–351.
18. R. C. Chu, U. P. Hwang, and R. E. Simons, "Conduction Cooling for an LSI Package: A One-Dimensional Approach," *IBM J. Res. & Dev.* **26,** No. 1, 45–54 (January 1982).
19. V. W. Antonetti and A. L. Pascuzzo, "Cooling Large Scale Computer Systems," *ASHRAE J.* **13,** 25–30 (1971).
20. N. G. Aakalu, R. C. Chu, and R. E. Simons, "Liquid Encapsulated Air Cooled Module," U.S. Patent 3,741,292, June 26, 1973.
21. S. Oktay and H. C. Kammerer, "A Conduction-Cooled Module for High-Performance LSI Devices," *IBM J. Res. & Dev.* **26,** No. 1, 55–66 (January 1982).
22. J. U. Knickerbocker, G. B. Leung, W. R. Miller, S. P. Young, S. A. Sands, and R. F. Indyk, "IBM System/390 Air-Cooled Alumina Thermal Conduction Module," *IBM J. Res. & Dev.* **35,** No. 3, 330–341 (May 1991).
23. P. A. Coico, "Mechanical and Thermal Design Aspects of Large MCM-C Packages for IBM S/390 Servers," *Proceedings of the International Systems Packaging Symposium*, International Microelectronics and Packaging Society, Reston, VA, 1997, pp. 315–320.
24. R. R. Schmidt, M. J. Ellsworth, R. C. Chu, and D. Agonafer, "Refrigeration Cooled Computers: Application and Review," *Proceedings of the International Mechanical Engineering Congress and Exposition*, Orlando, FL, November 5–10, 2000, pp. 277–283.
25. S. Kang, V. Mahaney, R. Schmidt, P. Singh, and H. Victor, "Dehumidified Cooling Assembly for IC Chip Module," U.S. Patent 6,233,959, May 22, 2001; G. Goth, J. Loparco, and P. Singh, "Sealed Multi-Chip Module Cooling System," U.S. Patent 6,192,701, February 27, 2001.
26. D. R. J. White and M. Mardiguian, "Electromagnetic Shielding," *A Handbook Series on EMI and Compatibility*, Vol. 3, Interference Control Technology, Inc., Gainsville, VA, 1988, pp. 1, 2.
27. *Acoustics—Measurement of Airborne Noise Emitted by Information Technology and Telecommunications Equipment*, Second Edition, ISO 7779, International Organization for Standardization, Geneva, Switzerland, 1999.
28. *American National Standard Measurement of Sound Pressure Levels in Air*, ANSI S1.13, American National Standards Institute, Washington, DC, 1995.
29. *Noise of Computer and Business Equipment*, Version 3, Statskontoret 26:3, Teknisk NORM No. 26, Swedish Agency for Administrative Development, 1993.
30. J. A. Shaw, B. Donald, and M. A. Nobile, "Method for Evaluating the Attenuation of Acoustical Covers Used in High-End Computer Systems," *Proceedings of the 2000 National Conference on Noise Control Engineering on CD-ROM*, paper 1pNSc3, Institute of Noise Control Engineering of the USA, Inc., Washington, DC, December 2000.
31. A. Duncan, *Quality Control and Industrial Statistics*, Fourth Edition, Irwin Publishing Co., Homewood, IL, 1974.
32. B. Notohardjono, J. Wilcoski, J. Gambill, D. Porter, U. Jourdan, D. Linkstrom, and S. McIntosh, "Design of Earthquake Resistant Server Computer Structures," *Proceedings of the American Society of Mechanical Engineers (ASME) Pressure Vessels and Piping (PVP) Conference*, July 22–26, Atlanta, 2001, pp. 101–111.
33. B. D. Notohardjono, J. S. Corbin, S. J. Mazzuca, S. C. McIntosh, and H. Welz, "Modular Server Frame with Robust Earthquake Retention," *IBM J. Res. & Dev.* **45,** No. 6, 771–782 (November 2001).
34. Bellcore, *Network Equipment—Building System (NEBS) Requirements: Physical Protection*, GR-63-Core, 1995; available through Telcordia Technologies, Inc., 8 Corporate Place, Piscataway, NJ 08854.

Received September 18, 2000; accepted for publication February 14, 2002

Prabjit Singh *IBM Server 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 Server Group. He received his B.Tech. degree (with honors) in metallurgical engineering from the Indian Institute of Technology, Kharagpur, his M.S. degree in microelectronics manufacturing from Rensselaer Polytechnic Institute, and his M.S. and Ph.D. degrees in metallurgy from the Stevens Institute of Technology. Dr. Singh received an IBM Outstanding Technical Achievement Award for his contributions to the first IBM multichip refrigeration unit. He holds more than ten patents and has received seven IBM Invention Achievement Plateau Awards and three IBM Publication Achievement Awards. He is a past chairman of the Electronic Materials Division of ASM International.

Steven J. Ahladas *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (ahladas@us.ibm.com).* Mr. Ahladas is a Senior Engineer in the Product Power, Packaging, and Cooling Architecture Department in the IBM Server Group. He received his B.S. degree in electrical engineering from the University of Massachusetts at Amherst in 1984. He has received one IBM Outstanding Technical Achievement Award and has been granted four U.S. patents. Mr. Ahladas is a registered Professional Engineer in the state of New York.

Wiren D. (Dale) Becker *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (wbecker@us.ibm.com).* Dr. Becker is a Senior Technical Staff Member in the IBM Server Group. He received his B.E.E. degree from the University of Minnesota at Minneapolis, his M.S.E.E. degree from Syracuse University, and his Ph.D. degree from the University of Illinois at Urbana–Champaign. He is currently a Senior Technical Staff Member leading the MCM design team that integrates and implements the multiprocessor design for IBM S/390 platforms. He has received IBM Outstanding Technical Achievement Awards for the design and development of G4, G6, and z900 MCM packaging and an IBM Outstanding Innovation Award for the G5 package development. He has authored or co-authored more than fifty journal articles and conference papers and has achieved the first invention plateau at IBM. Dr. Becker's current interests focus on the electrical design of the components that comprise a high-frequency CMOS processor system. He specializes in the application of electromagnetic numerical methods to the issues of signal integrity and simultaneous switching noise in electronic packaging, the measurement of these phenomena, and the verification of the models. Dr. Becker is a member of IEEE and IMAPS.

Frank E. Bosco *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (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; he subsequently worked on circuit designs for the first implementations of cryptographic algorithms. Mr. Bosco worked in power-supply development from 1974 to 1978. From 1979 to 1981, he 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 servers.

Joseph P. Corrado *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (jpcorrad@us.ibm.com).* Mr. Corrado is an Advisory Engineer and Project Manager in the IBM Server Group. He received his master's degree in computer science from Union College in 1993 and his B.S. degree in aerospace space engineering from the State University of New York at Buffalo in 1985. He holds several U.S. patents and has extensive experience in DFMA, plastic part design, and optical instrument design. In 1994 Mr. Corrado received a Business Week magazine IDEA Gold Award, and he recently received an IBM Program Management Excellence Award for his management role in the development of the z900 server.

Gary F. Goth *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (gfgoth@us.ibm.com).* Mr. Goth is a Senior Technical Staff Member in the Thermal Engineering Department in the IBM Server Group. He received his bachelor's degree from Princeton University and master's degrees from Rensselaer Polytechnic Institute and Union College. Mr. Goth has received an IBM Outstanding Invention Award and three IBM Outstanding Technical Achievement Awards; he holds 14 U.S. patents. He has been the lead thermal engineer of G2 through G6, zSeries, and other high-end servers, and has been responsible for air and refrigerant cooling of numerous logic systems. He is a member of ASME.

Sushumna Iruvanti *IBM Technology Group, East Fishkill facility, Hopewell Junction, New York 12533 (iruvanti@us.ibm.com).* Dr. Iruvanti is a Senior Engineer in the Interconnect Products Development group. He received his B.Tech. degree from Madras University, India, his M.S. degree from Clarkson University, and his Ph.D. degree from the State University of New York at Buffalo, all in chemical engineering. He has received an IBM Corporate Award and an IBM Outstanding Innovation Award for the development of advanced thermal compounds and flat-plate cooling technologies. Dr. Iruvanti holds more than twenty patents, and he has received seven IBM Invention Achievement Plateau Awards. He is a member of AIChE and IMAPS.

Matthew A. Nobile *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (nobile@us.ibm.com).* Dr. Nobile is a Senior Engineer in the IBM Server Group, responsible for acoustics and noise-control engineering for zSeries products. He received his B.S. degree in electrical engineering from Cornell University, his M.S. degree in acoustical engineering from Pennsylvania State University, and his Ph.D. degree in acoustics, also from Pennsylvania State University. He is a Fellow of the Acoustical Society

of America and a Board Certified Member of the Institute of Noise Control Engineering; he serves as Technical Director of the NVLAP-accredited IBM Hudson Valley Acoustics Laboratory. He is active on ANSI and ISO standards committees and ITI industry groups, and is currently on the Board of Directors of INCE. Dr. Nobile has reached the third-level plateau of the IBM Publication Achievement Awards.

Budy D. Notohardjono *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (budy@us.ibm.com).* Dr. Notohardjono is a Senior Engineer in the Mechanical Analysis Shock and Vibration Department in the IBM Server Group. He received his Ph.D. degree in mechanical engineering from the University of Wisconsin at Madison in 1984, and his M.B.A. degree in finance from New York University in 1997. He has received two IBM Outstanding Technical Achievement Awards and holds several U.S. patents. He is a registered Professional Engineer in the state of New York.

John H. Quick *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (jhquick@us.ibm.com).* Mr. Quick is a Senior Engineer responsible for high-end eServer I/O logic card development; he has more than thirty years in the electronic packaging field. He received a B.S. degree in technical management from the State University of New York at New Paltz. Prior to joining IBM in 1981, he developed electronic packaging for military and commercial products. His previous IBM assignments in the Kingston and Poughkeepsie laboratories include roles as design team leader responsible for ESCON director and channel card designs. Mr. Quick has received an IBM Outstanding Technical Achievement Award for his work in packaging z900 card designs.

Edward J. Seminaro *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (seminaro@us.ibm.com).* Mr. Seminaro is the Director of Server System Design and Architecture and a Distinguished Engineer in the IBM Server Group. He and his team are responsible for establishing and executing the system design of the IBM iSeries and pSeries UNIX product family. He received a B.S. degree in electrical engineering from Rutgers University and has done graduate work in electrical engineering at Syracuse University. Mr. Seminaro has been involved in the design, development, and manufacturing of S/390 and UNIX servers for more than 19 years in both management and technical roles. He has detailed technical expertise in the areas of power, packaging, and cooling. Mr. Seminaro has received five IBM Outstanding Technical Achievement Awards; he holds eight patents and has received two IBM Invention Achievement Plateau Awards.

Kwok M. Soohoo *IBM Server Group, 2455 South Road, Poughkeepsie, New York 12601 (ksoohoo@us.ibm.com).* Mr. Soohoo is a Senior Engineer responsible for high-end pSeries and zSeries server product EMC development; his previous assignments in the Poughkeepsie Electromagnetic Compatibility Laboratory include roles as compliance test team leader responsible for worldwide compliance sign-off, and principal EMC development engineer for S/370, S/390, and PPS systems for more than twenty years. He received a B.S. degree in physics from Fordham University, pursued graduate studies in geophysics and computer science at Columbia University, and received an M.S. degree in electrical engineering from Union College. He is a Senior Member of the IEEE, Technical Session Chairman for the 2000 International IEEE EMC Symposium in Washington, DC, Technical Session Chairman for the 2002 International IEEE EMC Symposium in Beijing, and a Senior NARTE Certified EMC and ESD Control Engineer.

Chang-yu Wu *IBM Corporate Division, Route 100, Somers, New York (changywu@us.ibm.com).* Dr. Wu is a Senior Technical Staff Member and Program Director in charge of Asia Pacific Regional EMC/Telecom/Safety Regulatory affairs. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from Taiwan University, the University of Missouri, and Syracuse University, respectively, joining IBM in 1966. Dr. Wu was a visiting professor at Union College and has lectured extensively on electromagnetic compatibility. He is an IEEE Fellow and an Academic Advisor to the Chinese Electronics Standardization Institute.