First- and second-level packaging for the IBM eServer z900

by H. Harrer
H. Pross
T.-M. Winkel
W. D. Becker
H. I. Stoller
M. Yamamoto
S. Abe
B. J. Chamberlin
G. A. Katopis

This paper describes the system packaging of the processor cage for the IBM eServer z900. This server contains the world's most complex multichip module (MCM), with a wiring length of 1 km and a maximum power of 1300 W on a glass-ceramic substrate. The z900 MCM contains 35 chips comprising the heart of the central electronic complex (CEC) of this server. This MCM was implemented using two different glass-ceramic technologies: one an MCM-D technology (using thin film and glass-ceramic) and the other a pure MCM-C technology (using glass-ceramic) with more aggressive wiring ground rules. In this paper we compare these two technologies and describe their impact on the MCM electrical design. Similarly, two different board technologies for the housing of the CEC are discussed, and the impact of their electrical properties on the system design is described. The high-frequency requirements of this design due to operating frequencies of 918 MHz for on-chip and 459 MHz for off-chip interconnects make a comprehensive design methodology and post-routing electrical verification necessary. The design methodology, including the wiring strategy needed for its success, is described in detail in the paper.

1. Introduction

The IBM S/390* platform has seen a new revitalization with the movement to complementary metal oxide semiconductor (CMOS) servers which began in 1993 because of the reduced hardware costs, high integration density, excellent reliability, and lower power of CMOS compared to bipolar technology. **Table 1** shows the development of the S/390 CMOS servers for the last four machine generations. From 1998 to 2000, the symmetric multiprocessor (SMP) MIPS number tripled in two years. This improvement in MIPS performance was achieved by using a faster processor cycle time (due to chip technology scaling), improved cycles per instructions (CPI), and an increase from 12 to 20 in the number of central processor units (CPUs) per system. The 20 CPUs allow the implementation of a 16-way node with four service processors for the z900 server. Table 1 also shows the continued increase of CPU electrical power during the last four years, which has led to the significant challenge of cooling a 1300-W multichip module.

In order to achieve the extremely high system performance for the z900 server, an elaborate hierarchical system design had to be followed. The basic strategy was

©**Copyright** 2002 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

0018-8646/02/$5.00 © 2002 IBM

Table 1 Development of the zServer from 1998 to 2002.

Year (Machine)	*Uni MIPS*	*SMP MIPS*	*Processors per MCM*	*Processor power* (W)	*Chip technology* (μm)	*Processor cycle time* (ns)	*Package cycle time* (ns)
1998 (G5)	127–152	901–1069	12	31–36	0.25	2.4–2.0	4.8–4.0
1999 (G6)	178–205	1441–1644	14	25–31	0.22	1.8–1.57	3.6–3.14
2000 (z900)	250	2694	20	32	0.18	1.3	2.6
2002 (z900+)	>250	>2694	20	38	0.18	1.09	2.18

to package the zServer core chips consisting of the processor, second-level (L2) cache, system control, memory bus adapter, and memory storage control chips on a single MCM (first-level package). Here, the short interconnect lengths with well-defined electrical behavior allowed a 1:2 cycle time ratio between the processor and the shared L2 cache. This approach has been used in previous S/390 server designs [1]. The 20 processors required about 16000 interconnections. For this number of interconnections, the MCM technology was the only cost-effective packaging solution for supporting the required bus widths, which are essential for the performance of the zSeries* SMP node. (For MCM cost/performance issues, the reader is referred to [2], which is also valid for this design.) This MCM technology also enabled the use of an easily implementable and reliable refrigeration system for the CEC chips by using a redundant cooler scheme. This scheme achieves a low-temperature operating point for the MCM chips at which the chip junction temperature is 0°C. The multichip module was interconnected to the rest of the system elements (e.g., memory and I/O) using an advanced printed-wiring-board-based technology.

Section 2 gives a detailed overview of the logic system structure of the z900 server. In the first system, released in 2000, the chip set had a cycle time of 1.3 ns using the IBM 0.18-μm technology. However, the MCM was designed to support a CEC chip set that operates at 1.09 ns, allowing the processor chips to be upgraded to a faster technology in 2002. This aggressive package design approach enables an easy upgrade at the customer's site by simply replacing the MCM with another MCM containing new processor chips. It also minimized the package development costs by using a single MCM design for two product offerings.

A Hitachi/IBM partnership enabled us to have two suppliers for the MCM used in the same z900 server product. Specifically, there are two types of multichip modules used. One, manufactured by the IBM Microelectronics Division (MD), uses glass-ceramic technology with a thin-film wiring plane pair. The other MCM, which is functionally equivalent and is manufactured by Hitachi, uses a glass-ceramic technology with tighter ceramic ground rules and no thin-film wiring. However, since these two designs have the same mechanical dimensions and are functionally equivalent while employing the same bottom side connector to the processor planar board, they can be used interchangeably. This is the first time that such a complex design (total wiring length of 1000 m connecting 16000 nets, in which 80% of all connections must operate at 459 MHz) has been implemented in two different technologies in a completely transparent manner and in record design time. Although many factors contributed to this achievement, the primary ones were excellent team cooperation and an efficient and effective design/verification system. The two MCM technologies mentioned earlier are compared with respect to cross section and electrical properties in Section 3.

To achieve a better granularity and cost reduction for the mid-range system, a simpler MCM has been designed which contains 12 processor chips (instead of 20) and connects to the memory subsystem with two buses instead of four. In addition, some cost reduction was achieved through the use of an alumina material instead of a glass-ceramic material for the MCM substrate and a reduced number of ceramic layers. However, this cost-reduced MCM maintained the same plug-in form factor in order to be able to use the same connector system and CEC board to reduce development expense.

Section 4 describes the processor subsystem and the second-level packaging. Specifically, the MCM is plugged into a board that contains power supplies, four memory cards with a maximum capacity of 64 GB, two cryptographic coprocessors, and self-timed interface (STI) bus connectors to the I/O subsystem. The high-speed STI connections provide a bandwidth of 24 GB/s to the I/O subsystem.

The IBM parallel processed printed wiring board (PWB), or P3, technology has been introduced for the z900 processor board in 2002. The building block for this technology is a three-layer core, featuring a reference plane sandwiched between two signal planes. This construction allows buried vias to pass through sandwiched reference planes for better wirability, an advantage which could not be achieved by the standard technology [1]. The enhanced buried-via technology provides a balanced triplate structure that eliminates all coupling between the x and y signal planes. Furthermore, it increases the

effective wiring density of each signal plane, as discussed in Section 4.

The design of a high-frequency packaging structure requires a well-controlled process for physical design, electrical analysis, and verification. The high current demand (630 A) of the 2002 multichip module requires a decoupling strategy to avoid malfunctions due to power distribution noise. The Fourier transform of this power distribution noise has three distinct components, which occur in the low-frequency, mid-frequency, and high-frequency ranges. The low-frequency noise is caused by changes in the power-supply load current and is filtered by two decoupling capacitor cards plugged into the processor subsystem board. The mid-frequency noise is dominated by inductive parasitic packaging elements in the power-supply path between the on-MCM and on-board decoupling capacitors. It affects primarily phase-locked-loop (PLL) circuitry, and it can be controlled by decoupling capacitors with low inductive paths on the multichip module and on the processor subsystem board. The high-frequency noise is dominated by a large synchronous on-chip switching and must be controlled by on-chip capacitors.

Approximately 80% of the nets on the MCM are source-terminated and operate at a clock frequency of 459 MHz. Each net has a wiring rule to define its allowable wiring length. Short nets had to be time-delay padded to avoid latching of a signal from the previous cycle (an early-mode fail) due to clock skew and PLL jitter between driver and receiver chips. The high wiring density on the packaging components also required a carefully coupled noise control between interconnection nets. The design methodology, described in considerable detail for S/390 G5 servers in [1], was followed for the IBM eServer z900 design. In this paper, we present the timing and noise results obtained by following this methodology for the z900 system, which confirm that this system meets its performance specifications. Section 5 gives details of the design methodology, including the decoupling strategy for low-, mid-, and high-frequency noise for both the first- and second-level packaging. In addition, timing analysis results and signal integrity results for all interconnections across all the package levels are disclosed.

2. Logical system structure and chip technology

The major change in the zSeries system has been the implementation of a 64-bit CEC architecture. For the second-level cache interface to the processor chips, we were able to continue using the same double-quadword bus introduced in the S/390 G5 server, but now feeding 20 instead of 14 processors in the z900 server. This increase in the number of processors allows us to achieve the desired multiprocessor SMP performance, but it has produced a significant increase in the number of interconnects among the chips in the CEC.

Figure 1 shows the high-level logical structure of the z900 system. The 20 processor chips are traditionally arranged in a binodal structure, in which ten processors are fully connected within an L2 cache cluster of four chips. In addition, each node contains two memory bus adapter (MBA) chips and one system control chip. A binodal core structure consists of two nodes, in which all CPUs are fully connected to the L2 cache within a node and can operate independently of the other node. This results in the excellent reliability, availability, and serviceability (RAS) features which are the hallmark of all S/390 mainframes and Enterprise zSeries servers. Only the clock (CLK) chip, which has a small number of circuits and uses a mature CMOS technology to minimize its probability of failure, is singular in the CEC.

Each single-core processor is implemented on a 9.9-mm $\times$ 17.0-mm chip in 0.18-μm technology and operating at a cycle time of 1.3 ns in the initial Year 2000 technology, ultimately operating at 1.09 ns in 2002. This is an 18% cycle-time improvement over the S/390 G6 processor. A single-core processor chip design point is chosen because it results in a relatively small chip size (170 mm^2) and provides satisfactory manufacturing chip yield.

The processor chip is connected with a 16-byte bidirectional bus to each L2 cache chip within each cluster of the binodal structure. This connection achieves the required memory bus performance, aided by an L1 cache on the processor chip that contains 256 KB of data and 256 KB of instruction capacity. The L2 cache size of each chip is double that of the G6. The eight 4MB L2 cache chips provide a total of 32 MB on the MCM. The cache chip is the largest chip in the CEC chip set, measuring 17.6 mm by 18.3 mm. The interconnection between the two nodes in the CEC is provided through the cache and system control chips. Specifically, every pair of corresponding L2 cache chips on the two nodes is connected by means of an 8-byte unidirectional store bus and an 8-byte unidirectional fetch bus. The large data bandwidth between processor and L2 cache is unique in IBM systems and has been achieved by the use of the dense glass-ceramic MCM packaging technology. It allows the operation of the interface to the L2 cache at twice the processor cycle time, which is crucial for the zSeries multiprocessor performance. In comparison to using a switch for connecting processor cards as in the Sun Microsystems Fireplane System [3], this structure allows a higher bandwidth and minimizes the latency between the processor and the L2 cache chips.

Each of the four memory cards contains a memory storage controller (MSC). The physical implementation has one MSC connected to two L2 cache chips with 8-byte buses to each. This bus interface is very critical for system performance, and it is required to meet the 2:1 frequency ratio with respect to the processor operating frequency.

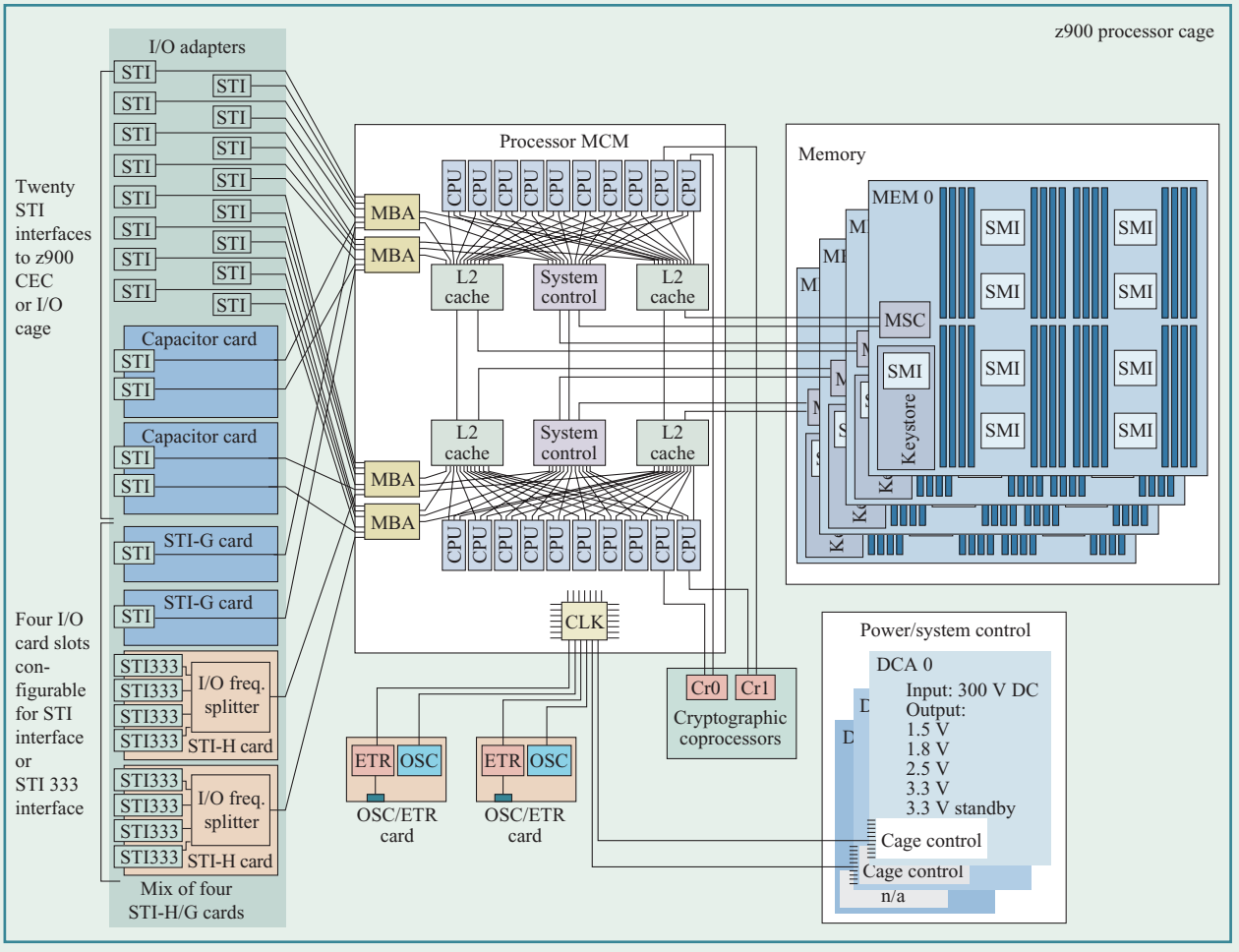

Figure 1

System structure of the IBM eServer z900. The core consists of the 20 processors, the L2 cache, and the system control chips arranged within a binodal structure. Four buses connect 64 GB of total memory. The clock chip synchronizes the external time reference from other servers; it connects to the three power-supply cards for the power-on and power-off sequence. (OSC/ETR: oscillator/external timing reference; SMI: storage memory interface.)

Therefore, in order to achieve the required 459-MHz operation for this bus, a nondifferential source-synchronous interface (called an elastic interface) has been used in a zSeries system for the first time [4]. In a source-synchronous interface, the clocks are transferred together with the data, and multiple bits of data are stored on the bus during the transfers. To initialize this interface, state machines are implemented on the MSC chip, the L2 cache chip, and the system control chip. Experimental data on a test-vehicle MCM substrate have confirmed that an on-MCM bus speed of 1 ns and an off-MCM bus speed of 1.4 ns can be achieved [5] using this elastic interface.

Each pair of cache chips is also connected, via an 8-byte bidirectional bus, to an MBA chip. Since these buses are not critical for the overall system performance, their operating frequency was allowed to be one fourth of the processor chip operating frequency (i.e., a gear ratio of 4:1). The MBA chips also contain the high-speed STIs [6] to the I/O cages, running at 1 GB/s per differential pair. The MBA chip contains six STI ports, each of which supports 1GB/s unidirectional links in both directions. Together with the transmitted clock signals, this requires a bus width of 240 lines for each MBA chip, resulting in a total I/O bandwidth of 24 GB/s. The 24 STI ports connect to the I/O cage or to the memory bus adapters of other

Table 2 Comparison of technology for chips of the central electronic complex with a central processor speed of 1.09 ns.

	No. of chips	*Technology* (μm)	V_{DD} (V)	*Used I/Os*	*Size* (mm)	*Power* (W)
Processor	20	0.18	1.7	636	9.9 × 17.0	38
L2 cache	8	0.18	1.7	1607	17.6 × 18.3	32
System control	2	0.18	1.7	1666	16.8 × 16.8	34
Memory bus adapter	4	0.22	1.95	700	11.4 × 10.6	23
Clock	1	0.22	1.95	767	12.2 × 12.2	4
Cryptographic coprocessor	2	0.22	1.95	133	7.45 × 7.52	6
Memory storage controller	4	0.18	1.7	755	11.0 × 11.0	14

z900 servers. Here, 20 STI ports are fixed as a direct 1GB/s link, and four STI ports are configurable either as a direct-feedthrough 1GB/s link or as four 333MB/s links going to the I/O cage of the previous system generation. In this case, the 333MB/s links provide the maximum data rate per channel, but the bandwidth is limited by the 1GB/s feed. The channel separation is controlled by a bus protocol converter chip.

The system control chip is the only I/O-limited chip in the system. It requires a maximum number of 1666 signal I/Os, leading to a chip size of 16.8 mm × 16.8 mm. Compared to the last machine generation, this is an increase in signal I/O count of 40%. This chip controls the operation of a single cluster and must therefore be connected to all processor and L2 cache chips in the same node as well as to the corresponding system control chip on the other node, which results in this extremely high signal I/O number. In order to reduce the number of I/Os, the closest L2 cache chips have been connected with two drop nets (a driver connected to two receivers) still operating at a 2:1 cycle time ratio with respect to the processor chip.

The clock chip receives a reference clock from either of the two oscillator cards. The reference clock is buffered and distributed to all processor, cache, system control, memory bus adapter, memory storage controller, and cryptographic chips. Except for the elastic interface buses to the memory cards, and the self-timed interfaces to the system I/O, which are source-synchronous, the CEC system connections are common-clock-synchronous. The reference clock is increased on-chip by a programmable 8:1 multiplier to the ultimate processor cycle time of 1.09 ns. The clock chip also contains the logic for reading the internal machine status by shifting bit chains out of the system. In this design, the external time reference (ETR) interface was implemented on the clock chip in this design (instead of the MBA chip, as was done in the past) in order to achieve chip size optimization for the CEC chip set. It synchronizes up to 32 servers, all working together as a single system image. The clock chip also connects to the three power supplies and provides the interface to the system for the power-on/off sequence, which is controlled by the service processor on the cage controller.

Two cryptographic coprocessors (one for each node) have been connected to the processors using a 4-byte bidirectional bus operating at five or six times the processor cycle time via a programmable clock gear ratio. To achieve better RAS, two processor chips have been connected to one cryptographic coprocessor chip with a two-drop net to provide the required redundancy at the system level. In addition, up to 16 cryptographic coprocessors are supported within the I/O cage.

Table 1 summarizes the chip technology. The highest power per chip (38 W) is consumed by the processor chips. Please note that the power values in **Table 2** describe the nominal case for a central processor (CP) cycle time of 1.09 ns. The power-supply domains are 1.7 V for the 0.18-μm CMOS processor, L2 cache, system control, and memory storage controller chips, which have higher performance and higher integration density, while the slower memory bus adapter, clock, and cryptographic coprocessor chips use a 1.95-V supply. The latter are implemented in an older 0.22-μm CMOS technology, leading to a significant cost reduction. It is possible to use this lower-cost technology because the cycle time that these chips must support is four times larger than the corresponding processor cycle time.

3. First-level packaging of the central electronic complex

With the z900 first-level packaging, IBM and Hitachi have designed and manufactured the most complex multichip modules in the world. In order to achieve the performance requirements of the system structure, 11000 nets out of a total 16000 nets for the logical structure shown in Figure 1 had to be embedded in these MCMs and operated at a cycle time of 2.18 ns. This results in a capacity bandwidth of 5 terabits per second for the processor, L2 cache, and

Table 3 Comparison of the 127-mm × 127-mm MCM in IBM zSeries servers from 1998 to 2002.

Machine (Year)	*No. of chips*	*Processors on MCM*	*MCM cycle* (ns)	*Power* (W)	*Thin-film layers*	*Thin-film ground rule* (μm)	*Ceramic layers*	*Ceramic ground rule* (μm)	*MCM wiring* (m)
G5 (1998)	29	12	4.0	800	6	45	75	450	645
G6 (1999)	31	14	3.2	900	6	45	87	450	689
z/900 (2000)	35	20	2.6	1100	6	33	101	396	997
z/900+ (2002)	35	20	2.18	1300	6	33	101	396	997

system control chip connections on the 127-mm × 127-mm substrate. The other 5000 nets operated at cycle times of 4.36 ns, 5.45 ns, and 8.72 ns.

In addition to the number of nets and the operating frequency, the total number of I/Os (signal and power pads for a chip site on the MCM) and the electrical power have significantly increased in the last two years, as shown in **Table 3**. The MCM size of 127 mm × 127 mm has remained unchanged throughout all of these system generations because it provides the desired integration density for the CEC chip set and maintains the MCM manufacturing tooling.

The total wiring length (including the via lengths in the MCM) doubled from 500 m to 1000 m within three years because of the growing number of processor chips in the binodal structure. For the z900 server, this resulted in a total of 35 chips on the module, with a total C4 count (number of chip I/Os) of 101 000 compared to 83 000 in the MCM of the G6 generation. This density required a tighter interstitial C4 footprint of 280-μm pitch (compared to 318 μm for the MCM of the G6 generation). The thin-film pitch was decreased from 45 μm to 33 μm to account for the tighter C4 ground rules of the chips. In order to meet the large wiring demand, a tighter wiring pitch of 396 μm compared to 450 μm was necessary in the ceramic part of the substrate [1]. Even with this tight pitch, 101 glass-ceramic layers were required to wire the interconnects, in addition to the one plane pair (PP) of thin-film (TF) wiring. Initially, a wirability study of the nets was performed with an IBM proprietary tool to estimate the number of plane pairs based on the Manhattan lengths (shortest possible wiring between two points) of the chip I/Os from the MCM floorplan. In the MCM, the signal-via to power-via ratio of 1:1 has been maintained for all chip sites to ensure a low inductive power path and to minimize the signal via-to-via coupling.

The high bus frequency of 459 MHz for the processor-to-cache and system control chips has been achieved by careful chip placement and strict wiring rules for each net. **Figure 2** shows the floorplan in which the symmetry of the binodal structure is obvious. The multichip module contains 20 processor chips (CPs), eight L2 cache chips (SCDs), two system control chips (SCCs), four memory bus adapter chips (MBAs), and the clock chip (CLK). The arrangement has been chosen to minimize the wiring lengths within the two CP–SCD–SCC clusters. A maximum length of 95 mm has been achieved for all of the time-critical nets that operate at 459 MHz. Since a net consists of three different types of wire, each having a different electrical behavior (e.g., the on-chip wire that runs from the driver/receiver to the C4, the thin-film wire, and the glass-ceramic wire), an optimization procedure was applied to achieve the shortest path delay. This procedure also determined the C4 assignment within the boundaries of the chip floorplanning and the assignment of the vias from the thin film to the glass-ceramic. It is to be noted that a complete thin-film wiring of the long nets would have led to worse results despite the faster time-of-flight for the thin-film material. This is because the high resistance of long thin-film lines slows down the signal and yields a slower voltage slew rate at the receiver input [7].

The routing challenge of the three-point connections between the SCC and a pair of SCD chips in each cluster operating at 459 MHz was met by using a near-end star connection between the driver and the two receivers. This means that the net is wired with two legs running from the driver to the two receivers using identical lengths for the two wiring segments. This avoided the typical pedestals caused by reflections by maintaining a length difference of less than 4 mm between the two legs of each net. This net topology, combined with the short interconnection distances of 55 mm for each wire leg, allowed us to meet the performance requirements. The clock chip connections were run with a frequency gear ratio of 4:1 or 8:1 with respect to the rest of the nest chips (e.g., L2 cache, system control chip, etc.) and were not particularly difficult to implement.

All of the on-MCM critical interconnections met the cycle time target of 1.8 ns, which provides the design with an 18% safety margin at normal system operation and guarantees that the package will not limit the system performance under any circumstances. A strict electrical verification of all nets has been applied, and the results are given in Section 5.

This MCM contains a total of 367 decoupling capacitors, each of which provides 200 nF at room

temperature. It is to be noted that this capacitance value is reduced by 20% when the MCM is cooled such that the chip junction temperature is 0°C. The decoupling capacitor allocation for each power-supply domain is 275 capacitors for 1.7 V and 47 capacitors for 1.95 V. The 2.5-V voltage is used only for the analog PLL circuit, and it has a separate filter capacitor at each chip site [the clock chip and memory bus adapter (MBA) chip each have two PLLs and two filter decoupling capacitors]. Five additional decoupling capacitors minimize the PLL jitter between the chips for the 2.5-V domain.

A pin grid array (PGA) connector based on the IBM high-density area connection (Harcon) technology was used at the bottom of the MCM as in the previous zSeries generation. Of the 4224 connector pins, 1735 pins are power and ground pins. To minimize pin coupling on the timing-critical CEC-to-memory interconnections, a signal-pin-to-power-pin ratio of 1:1 has been used.

Special emphasis had to be placed on the design of the differential pairs for the reference clocks and the self-timed interface, with blank tracks maintained next to each pair of wires in order to reduce the line-to-line coupling, and with the maximum length difference between the wires of each differential pair constrained to 3 mm in order to reduce signal skew.

The cooling of these multichip modules is achieved by two refrigeration units. Six thermistors continuously monitor the MCM temperature. They are used to shut down the system in case an over-temperature condition arises. Because of the low-temperature (0°C) operation of the z900 CEC chips on these MCMs, heating of the board from the back side of the MCM is required to maintain the board temperature above the dew point in the area where the seal-ring of the MCM separates the low-temperature environment from the rest of the processor subsystem environment.

The IBM-fabricated MCM substrate cross section is shown on the left side of **Figure 3**. It has a six-layer thin-film structure, which is identical to that of the last server generation, as described in [1]. The top thin-film layer contains the C4 pads and a repair structure which enables rerouting of defective signal lines in the substrate and increases the MCM yield [8]. The next layer is a ground mesh for a low inductive path of the decoupling capacitors; it provides excellent shielding properties for electromagnetic compatibility (EMC). These two layers are followed by two signal layers for routing and fan-out from an interstitial 396-μm chip footprint (280-μm minimum distance). In the IBM MCM, 30% of the wiring has been placed within this thin-film plane pair. The routing is done in orthogonal channels to minimize the line coupling. To achieve optimum shielding and a good decoupling, a V_{DD} 1.7-V mesh plane is placed at the bottom of the thin-film structure to obtain a triplate-like

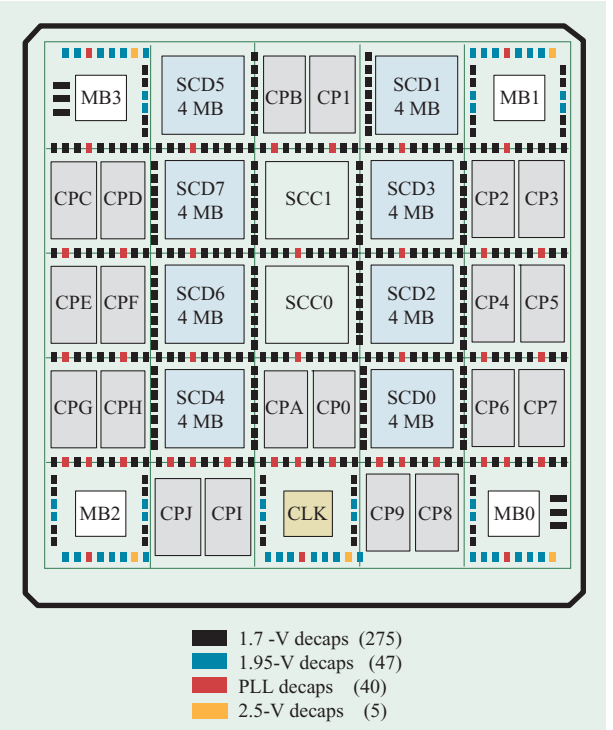

Figure 2

Floorplan of the IBM/Hitachi 20-processor MCM. The 35 chips are arranged symmetrically for the binodal structure. The 367 decoupling capacitors (decaps) are placed around the chip sites. Each decap provides 200 nF at room temperature. Six thermistors are placed at the top left corner and the bottom right corner to monitor the temperature of the MCM.

structure. The last layer contains the capture pads, which connect to the glass-ceramic substrate.

The glass-ceramic substrate begins with planarization layers, which were needed because of the large substrate thickness of 10.2 mm. After the signal wire jogging layers (which are fan-out layers with a very small wiring length), there are four voltage mesh planes, which achieve an excellent decoupling and a low inductive current path. The 2.5-V and 1.95-V mesh planes were not implemented in the thin film because the major current (450 A) flows on the 1.7-V domain.

The nets are wired within triplate-like structures in which two orthogonal wiring planes are always embedded between a ground and a V_{DD} mesh plane. This layer arrangement comes closest to a true triplate structure with respect to its electrical behavior and results in an impedance of 55 Ω. The shielding through the mesh planes is acceptable for such a high-complexity design, if a careful post-physical design verification is performed across the full system. The reference clocks are wired

Figure 3

Cross section of 20-processor MCM fabricated by (a) IBM and (b) Hitachi. The major difference is the six-layer thin-film structure on top of the IBM version, which includes a thin-film wiring plane pair, while the Hitachi MCM has only two layers of thin film for repair purposes. The planarization layers are to ensure flatness and parallelism. The jog- and via-transition layers are for via integrity. The redistribution layers fan out the 280-μm chip footprint on the 396-μm ceramic wiring grid. The wiring in the glass-ceramic is done within an x and y signal plane, with the wires within the x plane running orthogonal to the wires within the y plane. A voltage or ground mesh plane is placed above and below an x/y plane pair to ensure the closest return current and minimize coupling between adjacent lines.

within a single plane pair to minimize tolerances for the on-MCM nets. The multichip substrate contains 26 plane pairs for routing of the signals. One plane pair has been reserved for nets on which noise violations occurred from the electrical verification of the physical design of this substrate. Even if there were no bona fide noise violators, it was observed that by rerouting a very small number of nets (about 30), the maximum total noise was reduced by 200 mV. This rerouting was implemented in order to enhance the robustness of the MCM electrical design. At

the bottom of the MCM, a symmetric decoupling structure of all three V_{DD} planes and the GND mesh appears before the jogging and planarization layers end at the pin pads of the PGA Harcon connector.

The partnership with Hitachi allowed IBM to create a logically and electrically equivalent design using an alternative MCM technology without thin-film wiring. The wiring demand of this MCM design was satisfied in the absence of the thin film because of the superior ceramic via pitch that was available in the Hitachi technology.

Because of the lack of thin film, the design objective for the Hitachi MCM had to be increased from 1.8 ns to 2.0 ns for the most critical nets. While the floorplan of the chips is identical for the two MCM technologies, the cross section of the Hitachi MCM shown on the right side of Figure 3 has been slightly modified because of the absence of thin-film power mesh planes. It shows only a two-layer thin-film structure on the top of the C4 pads. The second thin-film layer contains the capture pads for the glass-ceramic substrate, in which a repair line can be wired. Five redistribution layers in the glass-ceramic were necessary to fan out the signals from the 396-μm interstitial chip footprint to a regular 297-μm grid substrate via pitch, with a constant power-to-signal via ratio of 1:1 in the whole substrate. Special care was invoked in the fan-out structure for the differential reference clock and the differential STI signal wiring.

Table 4 shows some important wirability attributes for this substrate. Please note that the average layer utilization is lower than the 50% stated in some textbooks [9]. This difference is attributed to the noise control applied to the interconnections of this MCM; i.e., some wiring channels are deliberately left vacant to minimize line-to-line coupled noise. It is interesting to note the high wiring utilization characteristics of the Hitachi MCM shown in Table 4. In Section 5 it is shown that the CEC cycle-time improvement achieved with the incorporation of a six-layer thin-film structure that contains one wiring plane pair is only 170 ps, or roughly 8% of the cycle time. However, special attention had to be paid to the layout of the redistribution layers to avoid a large amount of wrong-way wiring (layer-to-layer overlapped wires) for the performance-critical signals and to achieve a 1:1 power-to-signal via ratio in the substrate.

The total wiring length of the Hitachi MCM is 1004 m, which is nearly identical to the 998 m required by the IBM MCM. The electrical properties of these two MCMs are compared in **Table 5**. While the impedance of 55 Ω is

Table 4 Total wiring length including vias and average utilization for the IBM and Hitachi MCMs containing 20 and 12 processor chips. The average utilization is defined as the total length of available wiring channels divided by the length of the wiring channels actually used.

	Total length (m)	*Average utilization* (%)
IBM 20-CPU thin film	315	38.2
IBM 20-CPU glass-ceramic	682	34.2
IBM 20-CPU total	997	35.4
IBM 12-CPU thin film	237	28.7
IBM 12-CPU alumina	195	26.1
IBM 12-CPU total	432	27.5
Hitachi 20-CPU glass-ceramic	1006	41.0
Hitachi 12-CPU glass-ceramic	448	37.0

Table 5 Comparison of the electrical properties of the IBM and Hitachi MCM technologies for the horizontal (XY) and vertical (via) connections. The thin-film vias are too short to have a significant contribution and thus have been neglected.

	Z0 XY (Ω)	*T0 XY* (ps/mm)	*R_DC XY* (Ω/mm)	*Z0 via* (Ω)	*T0 via* (ps/mm)	*R_DC via* (Ω/mm)
IBM 20-CPU MCM thin film	43	7.2	0.28	—	—	—
IBM 20-CPU MCM glass-ceramic	55	7.8	0.021	47	8.7	0.04
Hitachi 20/12-CPU MCM glass-ceramic	55	8.3	0.038	45	8.7	0.04
IBM 12-CPU MCM alumina	48	11.5	0.06	41	13.8	0.058

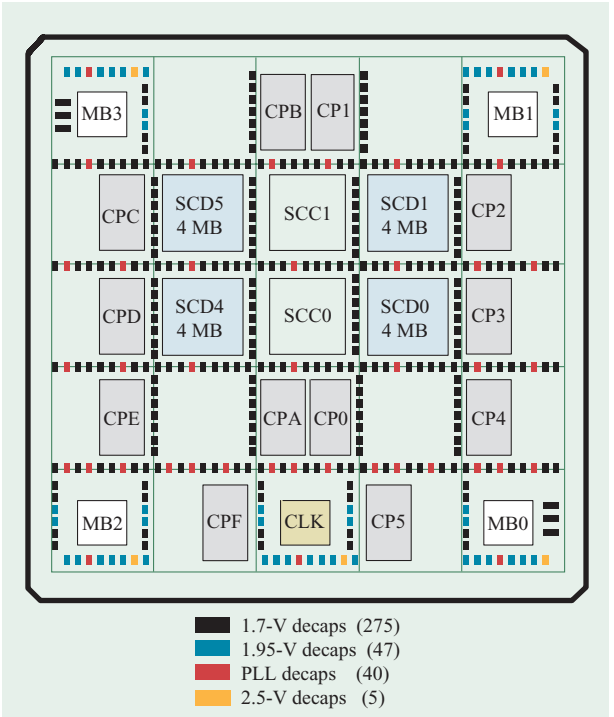

Figure 4

Floorplan of the IBM/Hitachi 12-processor MCM. It contains only half of the L2 cache, but all four memory bus adapter chips. The 367 200-nF decoupling capacitors have not been changed from the 20-processor version.

identical for the two MCMs, the tighter wiring pitch and the necessary adjustments in signal width, height, and dielectric layer thickness led to a slight increase in the time-of-flight (T_0) from 7.8 ps/mm to 8.3 ps/mm for the Hitachi MCM. Even without thin-film wiring, the tighter ceramic pitch of this MCM (297 μm vs. 396 μm) reduced the number of plane pairs from 26 (+1 PP of TF wiring) to 23 (+0 PP of TF wiring).

To achieve a more cost-effective entry-level product offering for the z900 system, a different ceramic substrate has been designed using a lower-cost alumina ceramic with a 396-μm line pitch and with one plane pair of thin-film wiring. This MCM contains a similar binodal structure (**Figure 4**), but with only 12 processors and four L2 cache chips supporting only two out of the original four available memory buses. The 12-processor MCM supports the full I/O bandwidth via all STI ports from the four MBA chips. Since optimization of the cost, not performance, was the primary objective for this design, the high-end floorplan and fixed-layer design of the voltage/ground mesh planes were reused to minimize the MCM design time. The use of the thin-film structure was mandatory for the fan-out and escape of the high-end chips in the IBM MCM technology. The alumina ceramic and the reduction from 26 to 10 plane pairs resulted in a significant cost reduction. Otherwise, the cross sections of both the thin-film and ceramic structures are identical to that of the high-end MCM implemented in glass-ceramic material. This design supports a maximum data switching time of 2.2 ns for all nets, with a fixed nest-to-processor-cycle-time gear ratio of 1:2. This implies that the alumina ceramic results in a performance impact for the nest interconnections of 400 ps, or 18% of the nest cycle time compared to the glass-ceramic with thin-film wiring substrate. The total wiring length for the interconnections embedded in this MCM was 432 m. However, the longest nets, with a 1:2 cycle-time gear ratio, were placed on the thin film for this design because of the lower crosstalk noise and faster time-of-flight compared to the alumina ceramic. Table 5 shows the utilization attributes of this design. It should be noted that the utilization in this alumina substrate was smaller than in the glass-ceramic. This is the direct result of the superior noise performance of the glass-ceramic structures compared to alumina and our requirement to achieve designs with the same robustness with respect to system switching noise.

The cooperation with Hitachi resulted in the design of another mid-range MCM which utilizes glass-ceramic material and contains only two thin-film layers for repair, like the corresponding high-end 20-processor (20-CPU) MCM. Because of the smaller pitch of the Hitachi ceramic technology, thin-film wiring was not required in order to maintain the chip-site fan-out, as was the case with the 20-processor design. In addition, the ceramic wiring plane pairs were reduced from 23 to 12 plane pairs by rerouting the 12-processor version. This result confirms the functional wirability advantage of the smaller ceramic line pitch in the Hitachi glass-ceramic technology that was also observed in the high-end MCM designs.

4. Second-level packaging for the central electronic complex

This section describes the physical partitioning of the logical structure shown in Figure 1. This physical partitioning is chosen to meet all of the zSeries system requirements, such as performance, connectivity, granularity, and the capacity of the system to be upgraded and hot-plugged.[1] The resulting processor subsystem is integrated in a mechanical structure known as a CEC cage whose backbone is a planar printed wiring board (PWB) that contains the following components: the CEC (i.e., the processor MCM already described), the memory subsystem, the cryptographic subsystem, I/O connectivity

[1] "Hot-pluggability" is the ability to plug and unplug cables without shutting off power.

ports to connect to I/O cages and other processor cages, decoupling cards, oscillator/external time reference (OSC/ETR) cards, the cage control (CC) structure, the logic voltage generators (distributed converter assemblies, or DCAs), and the cooling control system. A short description of the functional characteristics of each of these card designs follows.

- *Memory cards*
  The maximum amount of main memory is 64 GB. It is partitioned into four field-pluggable memory cards as in previous zSeries (S/390) machine generations. Each 16GB memory card contains 16 soldered dual inline memory modules (DIMMs), nine storage memory interface (SMI) control chips, store-protect memory chips, and a memory storage control (MSC) chip. The memory card is treated as a field-replaceable unit (FRU) and can be replaced after power-off; i.e., it is not hot-pluggable.
- *Self-timed interface (STI) I/O connectors*
  Each of the four MBA chips on the MCM drives six STI interfaces, each comprising ten differential pairs (nine data and one clock). All 24 STI interfaces are wired to three different kinds of I/O connections. Sixteen of these ports have a high-speed connection directly attached on the z900 processor board. They are designed with minimum board wiring lengths to connectors to allow a maximum cable length of 10 m for coupling to other zSeries processor clusters. An additional four STI ports are placed on two decoupling cards, while another four STI ports are wired to dedicated I/O adapter card slots. Each I/O slot is designed to hold either a feedthrough STI card with a 1GB/s bus to drive a 3-m cable or a converter card which splits the 1GB/s STI link into four ports, each with a data rate of either 333 MB/s or 500 MB/s. This is accomplished by using a custom-designed chip housed on this card. All STI cables and the associated cards are hot-pluggable.
- *Oscillator/external time reference (OSC/ETR) cards*
  Two OSC/ETR cards are plugged into the z900 processor board. One card is always active, while the other one is a backup. Each card provides two functions: the clock signal generator for various master clocks and the external time reference (ETR) optical receiver function for the clock chip on the CEC module. Each card contains oscillators, PLLs, drivers, and optical receivers/drivers for fiber cables. For enhanced system reliability, the clock chip has two independent inputs; during the system power-on process, the clock chip selects which OSC/ETR card becomes the master and which one is the backup after completion of the power-on sequence. The timing synchronization of multiple central electronic complexes coupled together via a sysplex is achieved by the ETR electronics on the OSC/ETR card. This allows the use of up to 32 systems, each one having a 16-way zSeries node, which results in a maximum sysplex of 512 processors within a single system image.
- *Capacitor (CAP) card*
  Two decoupling cards located as close as possible to the MCM and memory cards are used to satisfy the power noise constraints at low frequencies. Because of their high capacitance content of 211 mF ($176 \times 1200\ \mu$F), these cards cannot be hot-plugged.
- *Logic voltage generator/cage controller (DCA/CC) cards*
  Three DC–DC adapter (DCA) cards are plugged into the z900 processor board to provide $n + 1$ redundant power supplies for all logic voltages (1.7 V, 1.95 V, 2.5 V, 3.3 V, and 3.4 V standby). Two cards are required by the electrical load, while the third one provides the required redundancy. All three cards are hot-pluggable. Two of the three DCAs host the cage controller (CC) cards. The CC cards control the cage infrastructure; i.e., they read vital product data (VPD) and configuration data, generate reset signals, boot/load the processor through a high-speed interface into the CLK chip, and run the DCAs. Each FRU contains a FRU gate array (FGA) to control the logic on each card and a system electrically erasable programmable read-only memory (SEEPROM) to read any important VPD data.

Processor board description

The CEC board is the backbone of the processor subsystem; it is shown in **Figure 5** with all of the subsystem components that it can contain. The CEC, described in Section 3, enables the use of a passive board, which is another consequence of the MCM technology that leads to the cost-effectiveness of the overall processor cage. The only active component on the back plane is a VPD on a little cardlet to identify the cage serial number. The z900 processor board is inseparable from the processor cage structure, but all of the other subsystems are field-pluggable, with some even being hot-pluggable because of system serviceability requirements.

Since the memory and the cryptographic interface are the most critical interfaces on the z900 processor board, both the memory cards and the cryptographic modules are positioned as close as possible to the processor MCM. Two memory-card pairs are placed on either side of the MCM, with each pair having one card located on the front and one on the rear side of the board. For the same reason, the two cryptographic single-chip modules (SCMs) are located below the MCM. This has produced the board component topology shown in Figure 5, with the MCM in the middle of the board. This physical implementation minimizes the average wiring length to all of the adapters and realizes the performance requirements for all of the interconnections in the processor cage.

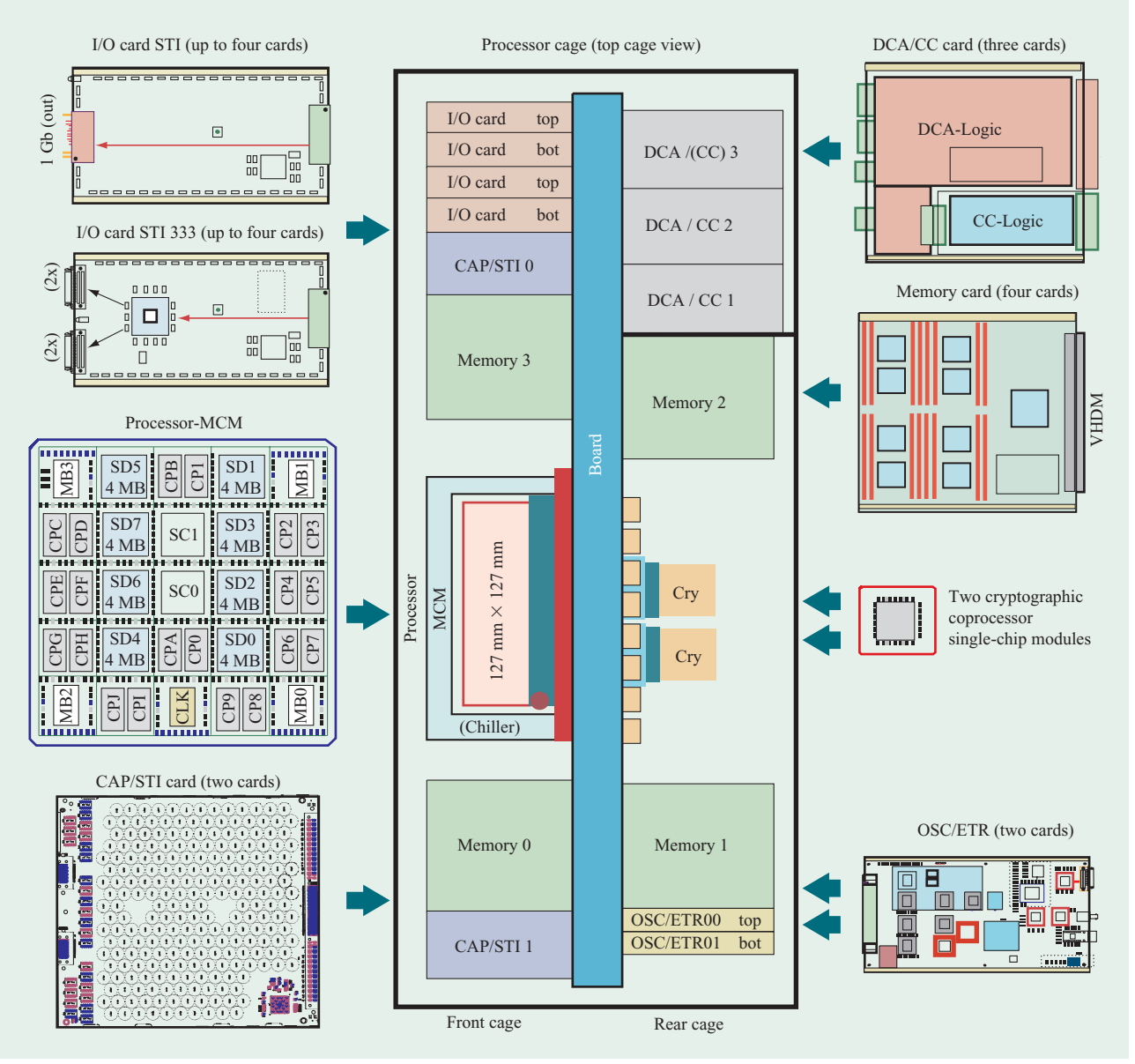

Figure 5

Center of figure: Top view of central electronic complex (CEC). The parts plugged into the CEC cage are shown at the left and right.

The processor board is 553 mm wide and 447 mm high, and has a thickness of 5.53 mm (217 mil). The card connectors are mainly Very High Density Metric (VHDM**) from Teradyne, Inc. Each of these six-row high-speed connectors has matched impedance (50 Ω) through the signal paths, with shield plate contacts for an excellent high-frequency return ground path. The DCA card-to-board electrical paths are through high-power connectors (HD4+2 by Winchester) which are used to carry 630 A total current into the system board. All connectors are pressed into plated-through holes (PTHs).

The MCM PGA Harcon zero-insertion-force (ZIF) connector is a side-actuated connector system which transmits zero force to the board, eliminating the need for separate support plates and reducing the concern for board laminate damage. For the Harcon, the board contains 4224 soldered bifurcated springs which mate to the MCM pins. The contact springs are divided into four quadrants with 1056 contacts each. The interstitial pitch

Table 6 Components or cards used in the z900 CEC cage.

Component type		*Signal pins*	*Power/ground pins*	*Pins/ component*	*Total PTHs*
Processor MCM	1	2,489	1,735	4,224	4,224
Memory cards	4	270	225	495	1,980
I/O cards	4	120	124	244	976
OSC/ETR cards	2	120	124	244	488
CAP/STI cards	2	180	433	613	1,226
STI cable connectors	16	42	24	66	1,056
Cryptographic SCM	2	200	345	545	1,090
SEEPROM card	1	8	8	16	16
DCA card	3	120	348	468	1,404
Resistor 470 Ω SMT1206	1	1	1	2	2
Decaps 1-μF SMT0805 double-sided	4,318	0	2	2	8,636
Decaps 10-μF SMT1210 double-sided	748	0	2	2	1,496
Total PTHs in processor board					22,594

is 2.2 mm by 2.4 mm, with a gap of 6 mm between the quadrants.

The surface mount technology (SMT) capacitors are soldered on both sides of the z900 processor board. This board assembly comprises 35 large components and 5066 decoupling capacitors. The board contains a total of 3516 nets with 19788 signal and power pins. The nets are mostly point-to-point (i.e., a two-pin topology). The average pin density is about 52 pins/in.^{2}, while the pin density under the MCM is 228 pins/in.^{2} and 250 pins/in.^{2} under the VHDM connector. **Table 6** gives an overview of the components with their signal/power pins.

Printed wiring board technology

Choosing an appropriate board technology is one of the challenges in a complex system. Different aspects must be considered when defining a board cross section. A few of the main considerations are manufacturing design, power distribution design, and signal layer design.

The manufacturing design is limited by board size, because of its impact on board yield, as well as by board thickness, because of drilling and copper plating limitations. The maximum board thickness is a function of the minimum plated-through-hole diameter, which is defined by the wirability requirements of the system and the connector technologies. There are two distinct factors that limit the amount of copper that can be connected to a given PTH. The PTH drilling limitations are strongly linked to the amount of copper being drilled through. Increased drill breakage and poor hole quality can increase board costs and reduce reliability if excess copper is connected to a PTH. With increasing amounts of copper, soldering and rework for pin-in-hole components will be more challenging because of the heat-sinking effect. This can drive up costs by reducing the assembly yield.

Power planes built with 1-oz copper foils provide the signal reference in the processor board. The minimum linewidth is 3.3 mils, and the line spacing is kept at or above 4 mils. This reflects the optimization of electrical requirements and the number of signal layers.

In general, the technology selection was completed with a continuous interaction between development and manufacturing to ensure that no parameters were unduly increasing manufacturing costs.

The main design considerations with respect to power distribution were the number of voltage planes and the dc voltage drop. The number of voltage planes is determined by the number of different voltages used on the board, the amount of copper required per voltage domain, the maximum copper thickness, the number of reference planes required by the signal planes, and any restriction on split voltage planes. Since the board-resistive heating was not a concern in this design, the voltage drop limit was set by the voltage variation tolerated by the active components. This defined the total amount of copper required per voltage level.

Further considerations were based on the design of the signal layer. In the board, the number of signal layers was determined by the wiring channels needed for the signals to escape from beneath the MCM area. This is a typical situation with complicated MCM designs. Equally stringent were the requirements for wiring channels in the connector area, which depended upon the maximum allowed signal coupling, the required signal line impedance, and the maximum allowed signal attenuation. Furthermore, a quasi-triplate structure consisting of two orthogonal signal planes between reference planes, similar to the one existing in the MCM, was required for a controlled characteristic impedance to provide an adequate high-frequency return path and to avoid discontinuities and thus signal reflections.

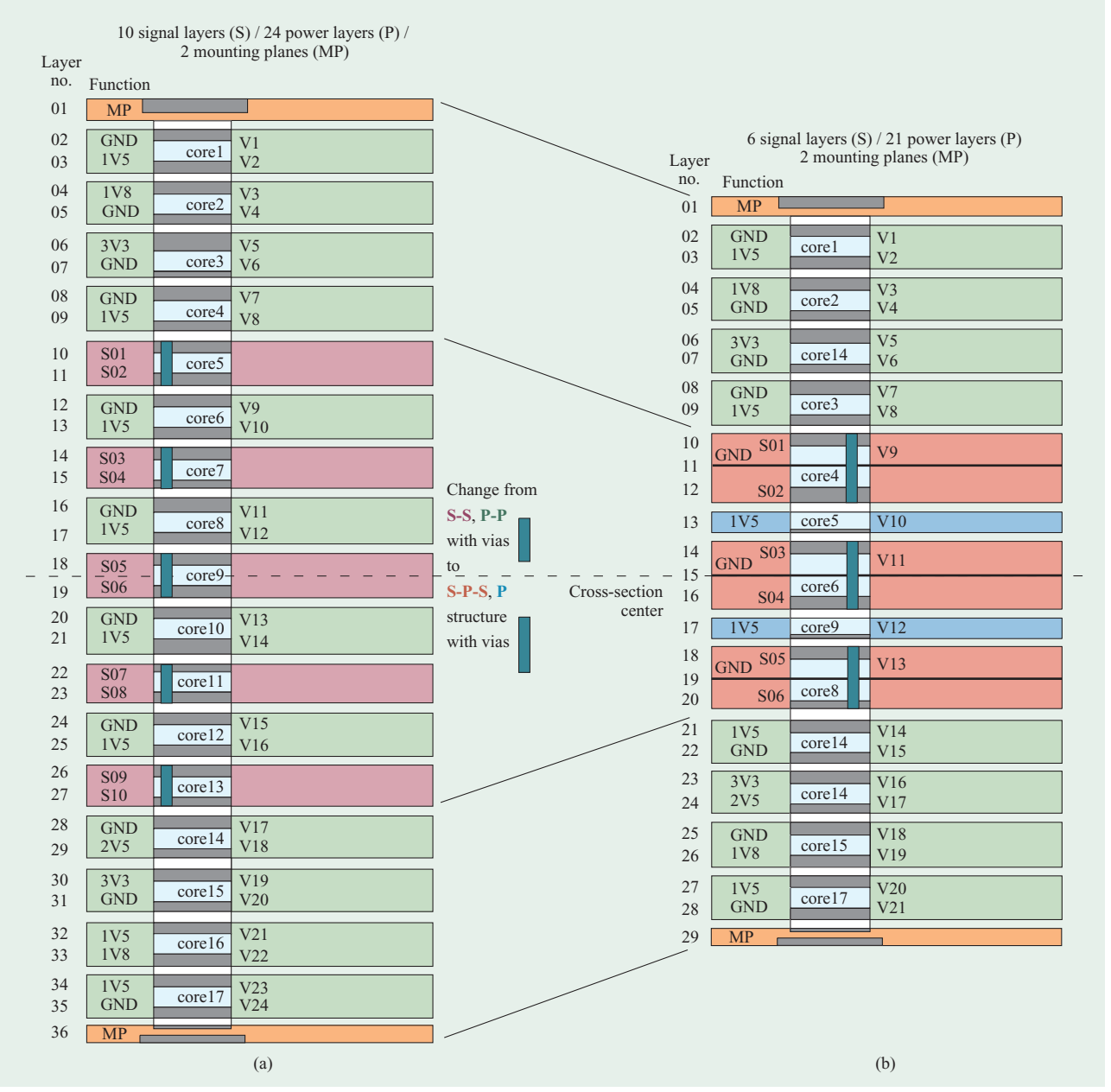

Figure 6

Comparison of z900 processor board cross sections for (a) standard and (b) enhanced buried-via technology. An improvement can be achieved by a reduction from ten to six signal layers.

All of these considerations resulted in a board with a quasi-triplate structure for ten signal planes, 24 power/GND planes, and two mounting planes, as shown on the left-hand side of **Figure 6**. The term *quasi-triplate* in this case means that two signal planes were placed between two power/GND planes, with the additional constraint that parallel wiring (vertical coupled lines) in the adjacent signal layers is minimized to avoid additional signal noise. For the signal lines, a linewidth of only 3.3 mils, with a 4-mil spacing for the areas under the various components and a 12-mil spacing for the component-free board area, was chosen. This choice provides the best tradeoff for minimizing the coupled noise and maximizing the wirability of the board. The signal layers were

fabricated using a 0.5-oz copper foil (0.7 mil) plated to the final thickness of 1.4 mils.

Although 100% wiring efficiency, with no wiring restrictions, would allow fanning out the MCM in only five signal planes, noise magnitude and length restrictions on the interconnections made the wiring of this board in ten signal planes challenging. Specifically, constraints on the coupled-noise magnitude allowed a maximum of two wires between two adjacent module pins. With an average signal-pin-to-power-pin ratio of 1.6:1 and a pin quadrant depth of 24 pins, at least 15 signals had to be brought out in one vertical set of wiring channels across all signal layers. To meet the system wiring demand in ten signal planes requires an optimized MCM pin assignment and the standard buried-via core technology which allows buried vias between signal lines within a signal–signal core. Using these vias, two signal lines in adjacent signal layers could easily be connected with a much smaller electrical discontinuity for the signal line than the standard IBM plated-through-via approach. A further advantage of a buried via is that it does not block wiring channels in other signal layers. Also, the buried vias use significantly less space than a through-hole because the via diameter is smaller and the adjacent spacing can be less. This results in improved wirability for the signal layers. On the other hand, the disadvantage of buried vias is the existence of two closely spaced adjacent signal layers, resulting in additional constraints on the noise between lines on these layers and reducing the wiring utilization of all signal planes. Controlling this noise requires careful checking of each interconnection at the completion of the design and before the design is committed to manufacturing.

As shown in Figure 6, most of the power planes were placed at the top and at the bottom of the board. In the center of the board, signal plane pairs were separated by 1.7-V/GND plane pairs. This approach was successfully used in many previous S/390 generations for two main reasons: 1) The 1.7-V/GND plane pairs in the center of the center board and adjacent to the signal plane pairs ensure a good high-frequency (HF) signal return path for the 1.7-V signals; and 2) the 1.7-V/GND planes on the top and bottom of the board minimize the via length between these planes and the decoupling capacitors mounted on the board surface and improve the effectiveness of the decoupling capacitors.

The total wiring length for all of the interconnections in the z900 processor board is 19132 in. (753 m), and 10316 buried vias are used to achieve the final wiring length. Each on-board connection is almost equal to its Manhattan length.

The limiting factor for the wirability in the standard IBM buried-via core technology was the blocking of wiring channels in adjacent signal layers in order to limit the noise magnitude. This problem was eliminated with the alternative cross section shown on the right-hand side of Figure 6, which uses real triplate structures with only one signal layer between power/GND layers. Since the vertical coupling among signal lines was eliminated, the signal-layer wirability increased, and the number of signal layers was reduced to six. In addition, the number of voltage/GND layers required for signal return paths has been reduced, but the number of copper ounces must remain the same for each voltage domain. Because of this latter requirement, the resulting board cross section does not provide any advantage in the number of power planes used. This board cross section is now available in the latest board technology from IBM. It achieves a real triplate structure with buried-via support by allowing the fabrication of these vias between two adjacent signal layers with one power or GND plane between. Thus, the wirability of each signal layer is maximized, and signal reflections due to discontinuities are minimized. A further advantage of this new IBM board technology over the IBM standard PWB technology with buried vias (MGF) is the reduced impedance tolerance for the signal wiring (from $50\ \Omega \pm 9\ \Omega$ to $50\ \Omega \pm 5\ \Omega$).

The design of the z900 processor board in the enhanced buried-via core technology resulted in a board cross-section change from ten standard buried-via cores to six signal layers, 21 power/ground layers, and two surface mounting planes. Increases in wiring length (20059 in., or 790 m) and buried-via count (12805) were observed because of the denser wiring. A complete comparison of the geometric and electrical data for these two PWB technologies is given in **Table 7**. It should be noted that the reduction in stacked-up layers afforded by the new IBM board technology resulted in a proportional decrease in cost for the board structure used in the z900 processor subsystem compared to the standard buried-via PWB technology.

The assembly concerns due to the total amount of copper in the cross section and the amount of copper connected per plated-through-hole were reduced through the use of thermal vents and more effective thermal breaks. The results can be seen in **Figure 7**.

5. Electrical design aspects

The electrical design for every component of the second-level package including the CEC MCM followed the philosophy and approach described in [1]. Specifically, the high-level design includes a large number of circuit simulations and extensive package structure modeling in order to establish the appropriate physical structures, the power distribution for the various voltage domains, the

Table 7 Comparison of the parameters of the standard buried-via and enhanced buried-via printed-wiring technology used for the z900 processor board.

	Line width/ line spacing/ line thickness (mil)	*Dielectric thickness to reference planes* (mil)	*Impedance Z_0 with tolerance* (Ω)	*Resistance* (mΩ/mm)	*Time-of-flight* (ps/mm)	*Effective dielectric constant, E_r*	*Buried via diameter* (mil)	*Dielectric loss, tan d*
Standard buried-via core technology	3.3/4.0/1.4	2.9 and 8.8	50 ± 9	7.4	6.5	3.9	8	0.018
Enhanced buried-via core technology	3.0/4.0/1.4	3.1 and 3.1	50 ± 5	8.2	6.1	3.4	6	0.010

determination of noise budgets, the timing and noise requirements for each class of interconnections, and the required high-frequency return paths for the signals. The design verification phase included the checking of the timing of every interconnection across all of the package boundaries, as well as the associated noise level at every input port of the chips in the processor subsystem.

Because of the significant magnitude of switching currents (a delta of 140 A) in the processor subsystem, special attention was paid to the decoupling approach and implementation. Specifically, a hierarchical decoupling strategy was used to ensure an acceptable voltage drop and ground bounce on all packaging levels.

To meet the timing requirements of the interconnections in this subsystem, a detailed pre-physical-design (PD) timing and coupling noise analysis was performed on all buses. After the physical design, a post-PD timing and noise verification was performed for each net to guarantee correct switching times and no failing signal condition due to the coupling or power noise. It is to be noted that all MCM versions required a one-pass design, because otherwise the long MCM manufacturing turnaround time would have delayed product general availability (GA) or time-to-market by several months. To this end, both the pre-PD detailed electrical analysis and the post-PD detailed verification of the timing and noise quantification of this design were necessary.

The hierarchical decoupling strategy

Uncontrolled power noise is one limiting factor of any system performance. The z900 system uses the latest CMOS chip technology requiring a power-supply level that is 15% lower than the power supply of the previous system generation. The lower operating voltage level results in reduced noise margin for the logic circuits. At the same time, the increased current demand requires more decoupling capacitors in order to meet the noise tolerance limits of the logic circuits. The overall increase in system complexity and component density requires more efficient decoupling capacitors while it imposes more severe placement challenges for these capacitors. The problem is compounded by the long reaction time of the internal power regulation, which is 20 to 50 μs after the detection of power-supply voltage drops. Since this reaction time is too long to ensure continuous system performance, a decoupling strategy capable of controlling the power noise at chip, MCM, and board levels at all frequency ranges necessitates the hierarchical decoupling design approach employed for this design. The details in this section are limited to the processor voltage level, but similar analysis and design are applied to the power distribution on the other voltage levels.

Three main frequency ranges were considered on the basis of our past experience [10]: the high-frequency (HF) range (above 100 MHz), the mid-frequency (MF) range (1 MHz–100 MHz) and the low-frequency (LF) range (below 1 MHz). Every frequency range requires the choice of a special capacitor type and unique analysis techniques to maximize accuracy and reduce design time. Not only the type but also the placement of the capacitor is essential for its decoupling effectiveness. The HF power noise can effectively be reduced only by using on-chip capacitors such as n-well and thin-oxide capacitors. These capacitor types are characterized by a small response time that allows the capacitors to deliver their stored charge as quickly as possible. Because of the impedance of the on-chip power/GND lines and the relatively low capacitance, the effective frequency range of the HF capacitor is limited. The total on-chip capacitance depends on the chip size, circuit utilization, and chip technology. Typically, values of 100–300 nF are possible with the chips and technology used in the z900 system.

To control the MF voltage drop, a second stage of decoupling capacitors is needed. Analysis of the system operation led to the requirement that the MF stage of

capacitors had to handle 22% of the 630 A of current on the processor power supply—a current step of 140 A. The second capacitor stage comprises 275 170-nF low-inductance capacitor array (LICA) capacitors on the MCM, as well as 2886 1-μF and 670 10-μF surface-mount technology (SMT) capacitors on the board. The 46.75-μF on-module capacitance is required in order to provide a low impedance close to the chips, but it does not have enough capacitance to accommodate 140 A of current for the 80-ns MF duration. Therefore, 2886 1-μF capacitors (type 0805) were placed closely around the MCM to provide the charge storage to reload the on-module capacitors. The 10-μF capacitors (1210 body size) were placed a bit farther away from the MCM and served to replenish the 1-μF capacitors. In addition, some 1-μF capacitors were placed around all card connectors to improve the HF return path for the signal lines and around the board edge to reduce plane resonances and reduce electromagnetic interference (EMI) levels.

The effectiveness of the on-board capacitors was maximized by using a special design for the capacitor pad and via connection (**Figure 8**) that significantly reduced the connection inductance in series with the capacitor compared to previous designs. This design integrates the vias inside the pad to minimize the distance between the vias and the capacitor body. A further reduction of the series inductance for capacitor groups can be achieved by connecting adjacent power pads and adjacent GND pads. Thus, the total series inductance for the 1-μF capacitors was reduced from 0.87 nH in the 1999 G6 system implementation to 0.16 nH in the z900 implementation [11]. This reduction of the series inductance also reduced the parallel resonance of the capacitors.

To estimate the power noise magnitude in a complex system such as the processor subsystem of the z900 server, the SPEED 97 program from Sigrity, Inc. was used to perform the required simulations [12]. This tool uses a full-wave approach to accurately predict the noise in all frequency ranges for systems consisting of chips on an MCM, a board, and capacitors mounted on different packaging levels. The results were derived by using a combined model of the chips and the first- and second-level package including the main electrical connections and all capacitors. The target was to keep the on-MCM power noise to less than 65 mV in the mid-frequency range and less than 10 mV at the board level. The SPEED 97 simulations confirmed that the power noise at the different packaging levels met this very aggressive goal. Details of the analysis, simulations, and verification by system measurements are documented in [12].

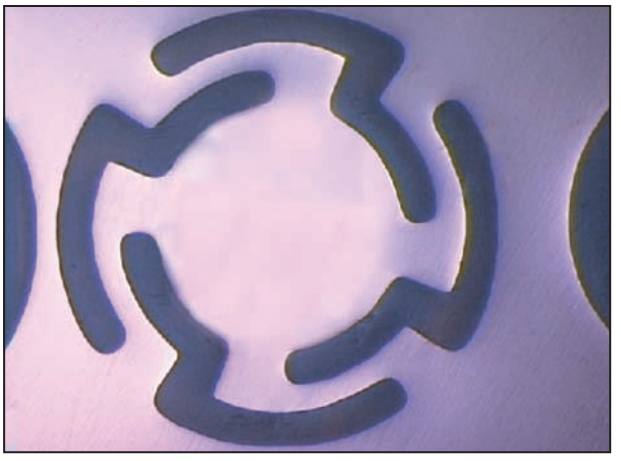

Figure 7

Photograph of 18-mil plated-through-hole (PTH) internal power connection. For thermal stress, this connection shows superior behavior compared to a solid power plane connection.

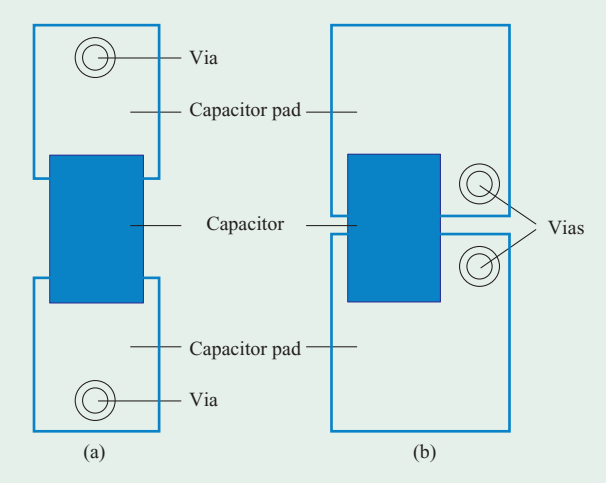

Figure 8

Top view of capacitors mounted on pads: (a) Earlier pad/via design; (b) new pad/via design with reduced parasitic inductance.

In the LF (less than 1 MHz) range, the system could produce a current change of one third of the 630 A for the case in which one of the redundant DCAs would fail. The LF power-supply noise magnitude was specified to be less than 40 mV. This was achieved by using 320 1.2-mF electrolytic capacitors on two capacitor cards and 144 1.2-mF electrolytic capacitors on the three power-supply cards (DCAs). The relatively long response time associated with the LF capacitors allows their placement at the edge of

Table 8 Decoupling strategy for the 1.7-V power domain, giving current and timing assumptions for high-, mid-, and low-frequency noise.

	Noise amplitude (mV)	*Placement of capacitors*	*Type of capacitors*	*Number of capacitors*	*Amount of decoupling 1.7 V*
High-frequency chip		Chip	Thin oxide		120 nF CP chip
(>100 MHz) ΔI = 20 A ΔT = 0.4 ns	100	MCM	1.6 mm A VX C4 ceramic	275*	47 μF on MCM at 0°C
Mid-frequency MCM		MCM	1.6 mm A VX C4 ceramic	275*	47 μF on MCM at 0°C
(1–100 MHz) ΔI = 140 A ΔT = 2.5 ns	60	Board	1 μF ceramic 10 μF ceramic	2886 670	2886 μF on board 6700 μF on board
Low-frequency board (1 kHz–1 MHz) ΔI = 235 A ΔT = 1 μs	40	Subsystem	1200 μF electrolytics	320	384,000 μF on cap cards

*Same capacitors are used for both mid- and high-frequency power noise decoupling.

Table 9 Packaging limiting-path delay for the IBM 20-processor MCM.

Delay	*Time* (ps)	*Percentage of total* (%)
On-chip	439	25
Packaging	846	47
Total adders	501	28
Clock adder	260	
Noise	77	
Clock PD skew	100	
Guardband (4%)	64	

the board, as long as the series resistance is low enough to meet the specified noise magnitude. **Table 8** summarizes the decoupling assignment and power-supply components in this system so that direct comparisons to a similar table in [1] can be made.

Timing analysis

The cycle time of the synchronous interconnects between chips is determined by the time-of-flight on the package interconnect, the on-chip delay including the driver and receiver, the timing differences in the clock tree (clock skew), the long-term PLL jitter for the driver/receiver chips, the impact of coupling and power noise on the signal arrival times, and a safety margin for the engineering testing [1]. The safety margin is required to ensure that the system has enough margin to run to end-of-life. In fact, each system is run at a 4% lower voltage and 4% faster cycle time before it is shipped.

Table 9 shows the packaging limiting-path delay of the z900 system for the IBM 20-processor MCM. It is a processor-to-L2 cache interconnection with a net length of 95 mm on the MCM. The package delay was minimized for the high-end MCM by forcing the routing into the top ceramic plane pairs and optimizing the thin-film lengths for the chip fan-out. The clock adder of 260 ps includes the PLL jitter as well as the impact from the process tolerances on the clock tree and on the reference clock of the different chips. **Table 10** summarizes the most critical net delay for the different buses wired on the MCM. The net delay of the SCC-SCD-SCD two-drop nets (driver on the system control chip and two receivers on the L2 caches) is about 50 ps shorter than the net delay of the longest point-to-point connection, which makes these nets not timing-critical.

For the long off-MCM nets to the memory cards, a source-synchronous clocking scheme is used (elastic interface or wave-pipelined interface) [4]. 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. This significantly reduces the timing tolerances and allows 459-MHz operation on these nets. One of the key design requirements of this interface is to minimize the delay differences between the bits in a bus so that the total

Table 10 Net delays for the different buses for the IBM and Hitachi 20-processor MCM. The rows give the packaging net delay and the minimum supported processor cycle time. The two-CPU cycle column contains the slowest net of all point-to-point nets, operated at twice the processor cycle time. The SCC-SCD-SCD nets are the multidrop nets operated at twice the CPU cycle time. The MBA interface operates at four times the processor cycle, and the nets to the cryptographic coprocessor are operated at five times the CPU cycle. The memory bus delay skew describes the wiring differences for the elastic interface. The last column shows the fastest net in the system (early mode), which must have a minimum delay of 500 ps to be latched at the correct cycle.

	Two-CPU cycle (ns)	*SCC-SCD-SCD* (ns)	*MBA* (ns)	*Cryptographic coprocessor* (ns)	*Memory bus delay skew* (ns)	*On-MCM early mode* (ns)
Hitachi 20-CPU MCM	1.93	1.76	3.05	5.11	<0.2	0.58
Hitachi 20-CPU CP cycle	**0.97**	0.88	0.76	1.03 0.86	n/a	n/a
IBM 20-CPU MCM	1.80	1.75	3.00	5.25	<0.2	0.56
IBM 20-CPU CP cycle	**0.90**	0.88	0.75	1.05 0.88	n/a	n/a

skew is less than one fourth of a cycle time, or 550 ps. On the Hitachi and IBM 20-processor MCMs, a nominal delay difference of 202 ps is achieved for all of these connections across all of the packaging components including on-chip wire between the driver or receiver circuit and the C4 pad, MCM, board, memory card wiring lengths, connectors, and SCM redistribution lengths. The addition of this delay difference to the 250 ps of on-chip skew yields a path skew less than 500 ps, the interface design skew specification.

In a synchronous interface, the signal could be latched in the previous cycle for very short interconnect lengths because of clock skew or clock jitter between the driving and receiving chips. Since the delay of the driving and receiving circuits was reduced more than the clock skew from previous systems, additional padding of very short lines was required to guarantee a minimum package delay for a net. The z900 server shows a minimum path delay of 556 ps (Table 10), which satisfies the requirements for early-mode skew and PLL jitter. Since this problem exists only for shorts on MCM nets, the wiring strategy was to place them in the lower plane pairs and maximize their wire delay. The autorouter used for the MCM wiring must be capable of assigning a minimum and maximum length to each net as well as satisfying other wiring constraints. Since no commercial autorouter is available to handle these constraints and the complexity in an approximately 35-chip MCM technology, IBM and Hitachi use different proprietary autorouters.

It is interesting to note that the wiring strategy was different for the MCM design manufactured by IBM for the high-end offering and mid-range systems. This was due to the significantly larger noise level that is generated in an alumina MCM compared to a glass-ceramic-based MCM design. This also explains the lower utilization of the alumina-ceramic layers shown in Figure 3, and the cost-effectiveness of the similar MCM built by Hitachi in glass-ceramic technology. In short, for the mid-range IBM MCM, all of the long nets were placed on thin film to minimize their exposure to coupled noise in the alumina ceramic. The latter was judged to be more dangerous than the increased attenuation because of the resistive attributes of the thin-film lines. The post-physical-design (PD) timing verification of all nets for all MCM designs was analyzed by an IBM proprietary tool called SLAM [1], which guaranteed correct operation for early- and late-mode conditions. This analysis applied delay equations to each net in the MCM to guarantee that the short-path and long-path delays were met on each net. In addition, the skew between bits on the elastic interface was also verified using SLAM. Since delay equations were used, a SPICE[2]-like circuit simulation program (AS/X) was used on the bounding nets to ensure the accuracy of the SLAM analysis.

[2] SPICE: Software Process Improvement and Capability Emulation, an ISO standard simulation process.

Table 11 Comparison of coupling noise in the IBM and Hitachi glass-ceramic technology. The comparison shows the far-end (FE) and near-end (NE) noise of 100-mm-long lines or 10-mm-long vias for a constant dv/dt = 5 V/ns and a voltage step of 1 V.

	Type	*Length* (mm)	*V_FE* (mV)	*V_NE* (mV)
IBM glass-ceramic	hor	100	40	8
	vert	100	39	7
	diag	100	20	2
	via	10	11	13
Hitachi glass-ceramic	hor	100	73	14
	vert	100	34	5
	diag	100	22	2
	via	10	9	13

Table 12 Comparison of the simulated 3σ worst-case values for the IBM and Hitachi 20-processor MCMs.

Net type	*Two-CPU cycle* (mV)	*Four-CPU cycle* (mV)	*Memory* (mV)	*STI* (mV)
Hitachi 20-CPU MCM	699	618	819	84
IBM 20-CPU MCM	645	606	862	120
Noise budget	930	920	920	150

Noise analysis

While the coupling noise can be well constrained within the triplate-like structures of a board and card technology with solid reference planes, there is a concern with the long nets on an MCM because the vertical and diagonal couplings through mesh planes are a significant contributor to the total noise. **Table 11** shows the MCM coupling for IBM and Hitachi glass-ceramic material of a 100-mm-long line from the first horizontal neighbor, the vertical coupling through a mesh plane from the closest vertical neighbor, and the contribution from the first diagonal neighbor. The coupling lines are terminated through a 50-Ω resistor, and the driver has a 5-V/ns slew rate. There is no coupling between lines on adjacent signal planes within one plane pair because the signal channels are strictly orthogonal. While the noise contribution from a single horizontal neighbor is 40 mV on the far end, the vertical coupling through the mesh plane is still 90% of this value, with another 50% contributed by the first diagonal neighbor. This results in a constellation of eight significant contributors for a coupling line when second-order effects are neglected. Table 11 compares coupled noise values for the IBM and Hitachi glass-ceramic structures. The Hitachi technology yields higher noise value for the horizontal coupling because of the tighter line pitch (297 μm compared to 396 μm). Even though all MCM designs have a strict 1:1 signal-to-power via ratio to reduce via noise, the Hitachi MCM produces smaller far-end noise because of the better shielding effects of the closely located power and ground vias (Table 11).

For cycle times below 1.8 ns, noise resonances are observed on the longest lines. The resonances are caused by reflected far-end noise from the current cycle, which is superimposed on the peak noise of the following cycle, creating an even larger noise peak. This can produce a noise increase of more than 50–80%. While frequency-dependent attenuation has nearly no impact on ceramic MCM nets, the high-frequency noise pulses are significantly attenuated on long board wires because of the dielectric losses of the board material. **Table 12** summarizes the highest noise contributors for the worst-case coupling nets. The typical dynamic noise pulse width of 350 ps results in a 930-mV worst-case statistical noise margin for the 1.7-V domain receivers. These receiver circuits have a dc hysteresis of about 160 mV. The highest noise peaks were observed on the memory nets where the coupling was dominated by the MCM/SCM line/via coupling and connector coupling. This is because the board/card line coupling contributions were less than 150 mV with the appropriate board/card line pitch.

Table 12 shows the results from a complete statistical noise analysis for all nets in the design based on the IBM proprietary tools SXTALK and GXTALK [13]. A typical noise value distribution from a first-pass MCM wiring run that includes the chip power switching noise component did not uncover any interconnections violating the noise limits, but identified a small number of nets with high noise values. By rerouting only 20 nets on the reserved plane pair for noise violators, the maximum noise voltage level on any interconnection was reduced by about 200 mV, or more than 20%. Because of this result, we believe that the allocation of an empty plane for noise violators is a worthwhile resource investment. The values in **Table 13** show the noise budget estimate from the pre-PD analysis for the on-MCM nets and the most critical memory nets. The numbers include reflections and multicycle effects. While the long on-MCM nets have the worst contribution from adjacent line coupling, the major contributor of coupled noise for the memory nets is the coupling of the vias in the MCM and the connectors between different package levels, because MCM wiring lengths were kept to less than 50 mm.

6. Summary and conclusions

The hierarchical package design of the processor subsystem for the z900 server comprises two parts, the MCM and the PWB. The highest wiring density providing the highest data rate is implemented on the MCM. The CEC contains the processor–L2 cache structure in a binodal implementation with a clock frequency of 459 MHz for the majority of its connections. In the MCM, the bandwidth for switching signals between the processor chips and the L2 cache/system control chips on a 127-mm $\times$ 127-mm carrier is 5.0 terabits per second, a new world record for packaging technology. This high data bandwidth combined with minimum latency between processor and L2 cache is essential for the zSeries multiprocessor performance. The MCM technology is the only solution that could support the large chip I/O count and the dense wiring.

The new multi-sourcing strategy of the IBM Enterprise Systems Group and the cooperation between IBM and Hitachi provided a unique opportunity to quantify the flexibility of the CEC package design and the robustness of the verification tools used. In fact, the most important conclusion is that the existing design tool set [1] is robust and flexible enough to permit single-pass error-free designs for the most complicated MCM structures using different packaging technologies, with no differences in design time between the two. One corollary to this conclusion is that the physical implementation of the zSeries does not require an IBM proprietary packaging technology.

Table 13 Determination of the noise budget for an on-MCM net and a memory net. The maximum acceptable budget values have been separated into MCM, board, and card coupling noise. The values also include power-switching noise.

Noise comp.	*On-MCM* (mV)	*Memory* (mV)
MCM		
xy wiring	600	130
vias	220	110
power noise	100	50
Board		
connectors	—	200
xy wiring	—	100
vias	—	50
Card		
xy wiring	—	85
SCM/power noise	—	185
Total noise	920	910

The MCM ceramic technology, even with the tighter ceramic ground rules needed to eliminate the use of thin-film wiring, still appears to be the more cost-effective solution for this application. This is because the layer yield loss associated with the tighter ceramic ground rules is more than offset by the reduction in the number of ceramic layers and thin-film layers. Because of the existence of the electrically and physically equivalent IBM and Hitachi designs, the advantage of thin-film wiring on system performance has been quantified in an unambiguous and consistent manner for the first time. The MCM-D has an advantage of 170 ps, which is 8% of the 2.2-ns interconnection cycle time. On the basis of the zSeries performance attributes, the advanced thin-film structures provide a 3–4% system performance advantage over the MCM-C. The size of such a performance advantage can be used to derive the cost target of the thin-film technology in order to make it cost/performance-competitive.

An equally important conclusion on the cost/performance of MCM structures was derived from comparison of the mid-range MCM designs. We estimate that 15% more layers are needed in alumina than in glass-ceramic to maintain the same crosstalk noise level. Higher crosstalk in alumina forces the lines to be spaced farther apart, resulting in reduced wiring efficiency. We suggest that this percentage should be used as a rule of thumb to estimate the required number of layers for an alumina-based MCM solution and, hence, the cost of the resulting MCM.

If the CEC is the brain that provides the expected functional performance, the PWB is the backbone holding all of the required components together. It was fortunate

that we had the opportunity to compare the new IBM enhanced buried-via wiring board technology with our standard buried-via board technology for the z900 processor board. This new printed wiring board technology offered buried vias through the power and ground planes, signal lines with lower resistance and dielectric loss, tighter signal-line impedance tolerance, and the elimination of signal coupling noise between traces on adjacent layers. The significant reduction in the number of signal planes for this design enabled by this technology proved the potential of the enhanced buried-via technology for significant cost reduction of complex second-level packages.

A key element in avoiding potential intermittent failures of the system was the definition of a realistic constraint for the total voltage power-supply variation at the circuit level (100 mV). To meet this constraint, it was necessary to apply the hierarchical decoupling concept with the proper selection and placement of decoupling capacitors to accommodate 140 A of current variation on the main power supply. Furthermore, it was once more verified that a detailed and accurate analysis of the power-distribution system is a requirement for determining the required number and proper placement of decoupling capacitors when the avoidance of intermittent system field failures is a design imperative, as it is for the zSeries. Decoupling capacitors were the primary means of controlling the power-distribution noise in CMOS-technology-based systems. Analyses and simulations with commercial tools and verification by system measurements confirmed that power noise at different packaging levels met the very aggressive goals for the zSeries servers.

The physical design of both the first- and second-level packaging structures of this processor subsystem was accomplished using autorouters with stringent net length control based on the results of detailed electrical analysis. A post-physical-design electrical verification of the timing and signal integrity was applied to every interconnection in the subsystem to guarantee its functionality, even under extreme conditions for the chip process and system environmental parameters; this resulted in a robust 3σ design.[3] This concept follows the tradition of the past mainframe machine generations that the package should never limit the system performance or cause design-related intermittent fails during system operation.

Acknowledgment

The authors would like to emphasize that this work could only have been accomplished by the outstanding teamwork of the members of the IBM packaging departments in Poughkeepsie, Boeblingen, East Fishkill, and Endicott, and of the Hitachi packaging department in Hadano.

[3] The term *3σ design* means that the design-limited yield allows one out of a thousand machines to fail during their lifetimes. In practice, voltage and frequency guardbanding during manufacturing final test eliminates this one possible fail.

*Trademark or registered trademark of International Business Machines Corporation.

References

1. G. Katopis, W. D. Becker, T. R. Mazzawey, 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).
2. 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).
3. A. Charlesworth, "The Sun Fireplane System Interconnect," presented at the IEEE Supercomputing 2001 Workshop, Denver, November 2001.
4. E. Cordero, F. Ferriaolo, M. Floyd, K. Grower, and B. McCredie, "A Synchronous Wave-Pipeline Interface for POWER4," presented at the IEEE Computer Society HOT CHIPS Workshop, Stanford University, August 15–17, 1999.
5. M. F. McAllister, H. Harrer, and J. Chen, "Measurements of Signal Transmissions Using a Source Synchronous Wave-Pipeline Interface Across Multiple Interface Structures," presented at the Fourth Workshop on Signal Propagation of Interconnects, Magdeburg, Germany, May 2000.
6. 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, 829–846 (September/November 1999).
7. H. Harrer, D. Kaller, T. W. Winkel, and E. Klink, "Performance Comparison and Coupling Impact of Different Thin Film Structures," *VDE/VDI Proceedings of the Mikroelektronik 97*, Munich, March 1997, pp. 333–338.
8. H. Stoller, S. Ray, E. Perfecto, and T. Wassik, "Evolution of Engineering Change (EC) and Repair Technology in High Performance Multichip Modules at IBM," *Proceedings of the 48th IEEE Conference on Electronic Components and Technology*, Seattle, May 1998, pp. 916–921.
9. R. R. Tummala and E. J. Rymaszewski, Eds., *Microelectronics Packaging Handbook*, Van Nostrand Reinhold, New York, 1989.
10. T. W. Winkel, E. Klink, R. Frech, H. Virag, S. Böhringer, B. Chamberlin, D. Becker, and W. M. Ma, "Method and Structure for Reducing Power Noise," Patent No. DE8-1999-0040, IBM Boeblingen, March 1999.
11. B. Garben and M. F. McAllister, "Novel Methodology for Mid-Frequency Delta-I Noise Analysis of Complex Computer System Boards and Verification by Measurements," *Proceedings of the 9th IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, Scottsdale, AZ, 2000, pp. 69–72.
12. Sigrity, Inc., *SPEED 97*; available online at *http://www.sigrity.com*.
13. H. H. Smith and G. A. Katopis, "Multireflection Algorithm for Timed Statistical Coupled Noise Checking," *IEEE Trans. Components, Packaging & Manuf. Techn.* **19,** 503–511 (August 1996).

Received September 21, 2001; accepted for publication March 6, 2002

Hubert Harrer *IBM Deutschland Entwicklung GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (hharrer@de.ibm.com).* Dr. Harrer is a Senior Engineer working in the IBM Server Group. He received his Dipl.-Ing. degree in 1989 and his Ph.D. degree in 1992 from the Technical University of Munich. In 1993 he received a DFG research grant to work at the University of California at Berkeley. Since 1994 he has worked for IBM in the Packaging Department at IBM Boeblingen, leading the IBM MCM design team for the z900 server. In 1999 he was on international assignment at IBM Poughkeepsie, New York. Dr. Harrer's interests currently focus on the development of a new timing and noise-checking methodology for the IBM Server Division. He has published multiple papers and holds four patents in the area of first-level and second-level packaging.

Harald Pross *IBM Enterprise Server Group, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (hpross@de.ibm.com).* Mr. Pross is an Advisory Engineer working in the Enterprise Server Group. After joining IBM in 1975, he studied applied physics at the Fachhochschule Heilbronn from 1979 to 1984, graduating with a B.S. degree in physics in 1984 and joining the IBM Boeblingen laboratory that same year. He held various technical positions in S/390 processor packaging design in Boeblingen. In 1993 he moved to Rochester, Minnesota, to lead the card/board packaging design of the new high-end AS/400 processor. Returning to Boeblingen in 1994, he worked on various printed wiring board designs for S/390 processors. In 1997 Mr. Pross became the project leader for the Boeblingen processor card and board designs. Since 2000 he has led the overall packaging project for the zSeries 900 processor cage design and is responsible for the packaging implementation for future zSeries processors.

Thomas-Michael Winkel *IBM System/390 Division, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (winkel@de.ibm.com).* Dr. Winkel received his Diploma in electrical engineering in 1989 and his Ph.D. degree in 1997 from the University of Hannover, Germany. His research activities covered the area of characterization and modeling of on-chip interconnects using high-frequency measurements. In 1996 he joined the IBM development laboratory in Boeblingen, Germany. He is currently a Staff Engineer in the IBM Server Group, leading the electrical design team for the second-level packaging for the z900 CEC cage. Dr. Winkel's current focus is electrical packaging design with respect to high-frequency signal distribution and power noise. He is also interested in high-frequency on-chip measurements and modeling of on-chip signal as well as power and ground lines.

Wiren D. (Dale) Becker *IBM System/390 Division, 2455 South Road, Poughkeepsie, New York 12601.* Dr. Becker received his B.E.E. degree from the University of Minnesota, his M.S.E.E. degree from Syracuse University, and his Ph.D. degree from the University of Illinois. He is currently a Senior Technical Staff Member in the IBM Server Group. He leads the MCM design team that integrates and implements the multiprocessor design for the IBM S/390 platforms. Dr. Becker 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 IBM invention plateau. 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 the IEEE and IMAPS.

Herb I. Stoller *IBM Microelectronics Division, East Fishkill facility, Route 52, Hopewell Junction, New York 12533 (stollerh@us.ibm.com).* Mr. Stoller is a Senior Technical Staff Member with the Microelectronics Division, responsible for application engineering for high performance and special MCMs. He has authored and coauthored numerous papers on MCMs and MCM technology. Mr. Stoller holds 12 patents and has reached the Fifth Invention Plateau. He holds a B.S. degree from City College of New York and an M.S. degree from Rutgers University, both in physics.

Masakazu Yamamoto *Hitachi Enterprise Server Division, Horiyamashita 1, Hadano, Kanagawa 259-1392, Japan (masakazu.yamamoto@itg.hitachi.co.jp).* Mr. Yamamoto received the B.S. degree in physics from Kyoto University in 1977 and the M.S. degree in applied physics from Osaka University in 1979. He joined the Central Research Laboratory at Hitachi Ltd., where he worked on the development of high-speed packaging systems for large-scale computers. Since 1989, he has worked in the Enterprise Server Division of Hitachi Ltd. and engaged in the development of the packaging technology for large-scale computers such as the MP5800 and the Super Technical Server SR8000. He is currently a General Manager in the Enterprise Server Division of Hitachi Ltd., responsible for the hardware technologies of server and network systems.

Shinji Abe *Hitachi Enterprise Server Division, Horiyamashita 1, Hadano, Kanagawa 259-1392, Japan (shinji.abe@itg.hitachi.co.jp).* Mr. Abe received B.E. and M.E. degrees in mechanical engineering from the University of Tokyo in 1986 and 1988, respectively, and the M.S. degree in electrical engineering from Stanford University in 1999. He joined the Enterprise Server Division of Hitachi Ltd. in 1988 and has been working in the Hardware Technology Development Department. He is currently a Senior Engineer, and is responsible for MCM design.

Bruce J. Chamberlin *IBM Microelectronics Division, 1701 North Street, Endicott, New York 13760 (chamberb@us.ibm.com).* Mr. Chamberlin is a Senior Engineer in the Organic Packaging Development group at IBM Endicott. He holds a B.S. degree in mechanical engineering from Clarkson University. He has been the product engineer for Clark raw boards and for the Clark board assembly. He is currently the program manager for development and qualification of high-end PWBs supporting zSeries machines.

George A. Katopis *IBM System/390 Division, 2455 South Road, Poughkeepsie, New York 12601 (katopis@us.ibm.com).* Mr. Katopis is a Distinguished Engineer in the IBM ESG Server Division, responsible for the technology selection and

packaging strategy of CMOS servers. He has authored more than fifty papers on the subject of net design and switching-noise prediction and containment in the digital server engines. He holds three patents on switching-noise reduction and has coauthored chapters in three books on the electrical design of electronic packages. Mr. Katopis received an M.S. degree and an M.Ph. degree from Columbia University. He is an IEEE Fellow, and an industrial mentor to the electrical engineering departments of Cornell University and the University of Arizona at Tucson.