MCM technology and design for the S/390 G5 system

by G. A. Katopis
W. D. Becker
T. R. Mazzawy
H. H. Smith
C. K. Vakirtzis
S. A. Kuppinger
B. Singh
P. C. Lin
J. Bartells, Jr.
G. V. Kihlmire
P. N. Venkatachalam
H. I. Stoller
J. L. Frankel

The multichip module (MCM) that contains the central electronic complex (CEC) of the S/390® G5 system is described in this paper. The glass-ceramic module, topped with six layers of polyimide full-field thin-film wiring for chip-to-chip interconnection, represents IBM's most advanced packaging technology. This MCM provides a large wiring capacity, with 595 meters of routed interconnection; it supports the highest synchronous interconnection performance in the industry at 300 MHz; and it allows for cooling flexibility at the system level—either a heat sink for air-cooled systems or a cooling "hat" for systems using refrigeration cooling. The physical and electrical characteristics of this packaging technology, necessary to support the aggressive system performance goals (1040 MIPS) of the IBM G5 Enterprise Servers, are presented here. In addition, the approach used to produce a robust electrical and physical design is described.

1. Introduction

One of the fundamental premises for the design of the S/390* system packages is that the on-package interconnects must support two machine generations without limiting system cycle times. This premise applies to both the first-level package, which consists of ceramic multichip modules (MCMs) for all of the S/390 CECs, and the second-level package, which consists of multilayer FR4 boards (32 to 36 layers). For S/390 CMOS servers, it was found that with the appropriate ratio of L1 and L2 cache sizes, the off-chip interconnects must operate at half the processor clock frequency. Thus, for processors operating at 500 MHz, the on-MCM system bus operates at 250 MHz. This enables the G5 server to meet the single-image system performance goal of 1000 MIPS in a cost-effective manner. As already stated, this MCM, called GEMI (pronounced "gem" I) in the following, must support two machine generations and is therefore designed to support chip-to-chip on-MCM interconnects at 300 MHz for the 0.25-μm CMOS chip technology described elsewhere in this issue.

The off-chip interconnect performance is determined by many variables, namely, the chip technology, the package technology, the system CEC structure and content, and the interconnect strategy and topology. Since the chip technology is described elsewhere in this issue, we provide only a summary of the electrical characteristics of the buffer and latch circuits in this paper. However, we describe in detail the rest of the variables that affect the performance of the GEMI MCM and allow us to achieve

0018-8646/99/$5.00 © 1999 IBM

300-MHz off-chip interconnect performance using synchronous on-MCM connections and avoiding the overhead in latency and chip area of source-synchronous or asynchronous interconnections [1].

The CEC content, in particular the number and chip size of the processors that are included in a node, the number of system data (SD) chips required to provide the appropriate L2 cache size, and the width of the buses connecting the processor chips to the cache chips, determines the physical distances between the chips comprising the CEC, and directly affects the maximum off-chip performance that can be achieved. Naturally, this performance can be modulated by the MCM material's dielectric constant, but this is not a first-order factor for maximum off-chip interconnect performance. The system structure is the most important factor in off-chip performance. Therefore, Section 2 provides a concise but detailed overview of the structure of the G5 system, which offers at least 12 processors with eight L2 cache data chips supporting 8 MB of memory and denoted as SD in the following. As we discuss in this section, the system architecture consists of two closely linked nodes, providing us with the flexibility to increase the number of processors and double the size of the L2 cache when advances in CMOS technology permit us to do so. The optimized topological arrangement of these chips provides the minimum possible off-chip interconnection length that will limit off-chip performance. However, this performance can be affected by the selection of the MCM material to a significant degree (up to 25%, as was shown in [2]). Therefore, in Section 3 a detailed description of the GEMI substrate technology is given. It has to be noted that full-field thin-film wiring technology (e.g., a whole layer of thin film is exposed with one mask and contains chip-to-chip interconnection across the substrate) was used for the first time in the packaging industry at the 127-mm $\times$ 127-mm dimension of the GEMI MCM. This thin-film structure was combined with 75 layers of glass-ceramic material that provided the medium for the bulk of the interconnections (383 m). Note that there is an additional 212 meters of wire in the thin-film plane pair to reach the total of 595 meters of routed wire mentioned earlier. The choice of glass-ceramic over alumina material was based on cost–performance considerations that are described in [2]. For a robust package design one should, from the beginning, establish a strategy so that the noise magnitude at the three frequency ranges of interest [3] stays within the noise budgets required for the reliable operation of the chips on this MCM under the switching activity expected during system operation. This subject is presented in Section 4 of this paper. Since the only efficient way to control the switching noise of the CMOS circuits is through decoupling capacitors, this section presents techniques and approaches to estimate the required amount and type of decoupling for the various frequency ranges of interest. In addition, the noise budget used for the GEMI MCM is highlighted.

With the switching noise contained by design within acceptable limits, the interconnection strategy and physical design methodology are the next most important factors to be improved in the off-chip interconnect performance. This is described in Section 5 and is followed by the timing analysis of the nets in Section 6. Although this paper addresses only the first-level package of the G5 server, in Section 6 we cover both the on-MCM and off-MCM connections, specifically the CEC-to-main-memory connections, that have a direct impact on the system performance and hence influence the on-MCM wiring strategy.

Given the exceptional—and recently used as an example by competitive servers—RAS (reliability, availability, and serviceability) and scalability features of the S/390 system, we do not consider adequate the "noise avoidance by design" philosophy with subsequent verification of the noise containment and net timing by checking only a select number of nets. Our design philosophy is founded on checking the noise magnitude of every individual net on the MCM and the board. To this end, very efficient and precise proprietary tools have been developed and exercised for the noise verification of the GEMI module. The record time in the delivery of this module with no noise problems constitutes the fundamental justification of this approach. In Section 7 the results of the noise verification for the GEMI are presented, and the sensitivities of these results to design variables are discussed and quantified. Finally, we provide our conclusions for the extendability of this MCM and the corresponding design methodology in Section 8.

2. G5 system overview

The latest IBM S/390 G5 server design contains twelve microprocessors, as did the G4 design [4], but there are significant system structure differences between these two system families which reduce the microprocessor (denoted as CP) to L2 bus contention and increase the flexibility of the design to accommodate as many processors as one can fit on the GEMI MCM (as we will see in the following). This is accomplished by implementing the concept of a nonblocking crossbar switch for the L2 cache that services the microprocessors of this system. In addition, to minimize the number of I/Os required for the CP–L2 connections, a binodal structure is architected, as shown in **Figure 1**. From this figure one can see that all of the six CPs in each node are connected to all four SD chips belonging to this node with 72 bits of unidirectional buses that include ECC. Each SD chip contains 1 MB of L2 cache memory. However, the CPs in one node can access information from the main memory that is connected to the SD chips of the other

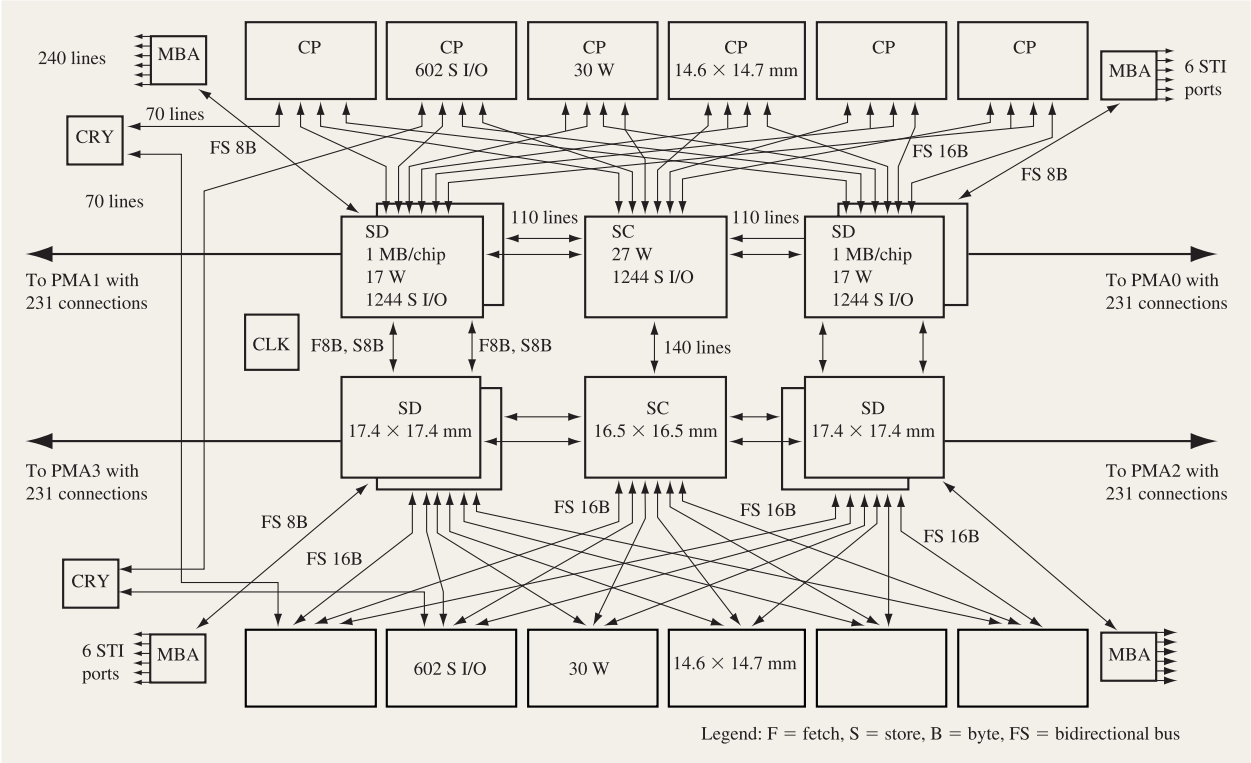

Figure 1

Block diagram of the G5 binodal architecture structure.

Table 1 G5 CEC technology.

Chip name	*Physical dimensions Litho/L_{eff}* (μm)	*Physical dimensions* (mm)	*Signal/power I/Os*	*Minimum cycle time* (ns)	*Power* (W)	*Power supplies* (V)
CP	0.22/0.15	14.6 × 14.7	600/1087	2	30	2
SD	0.3/0.18	17.4 × 17.4	1086/1519	4	16	2.6
SC	0.3/0.18	16.5 × 16.5	1189/1367	4	27	2.6
MBA	0.3/0.18	12.9 × 12.9	629/839	4	25	2.6, 3.3
Cryptographic coprocessor	0.5/0.25	12.7 × 12.7	200/839	8	13	2.6
Clock	0.3/0.18	12.9 × 12.9	748/839	4	7	2.0, 2.6
GEMI MCM	TF 18/45 GC 75/450	127.5 × 127.5	2630/1594	3.3	850	2.0, 2.6, 3.3

node via connections between the SD chips of the two nodes. The bus width of this connection is 72 bits with parity, and separate buses are used for data and control. In fact, in order to achieve the largest amount of L2 memory size, the controls necessary for the operation of each node are generated on two control chips, one for each node, denoted in the following as system control chips (SC). The SC chips are I/O-limited, requiring 1240 signal I/Os out of the 1244 signal I/Os that a 16.4-mm × 16.4-mm chip can provide. On the other hand, the size of SD chips is circuit-limited; in other words, we used the largest size that could be economically produced in order to maximize the size of the L2 cache. This resulted in a 17.4-mm × 17.3-mm SD chip. The characteristics of all

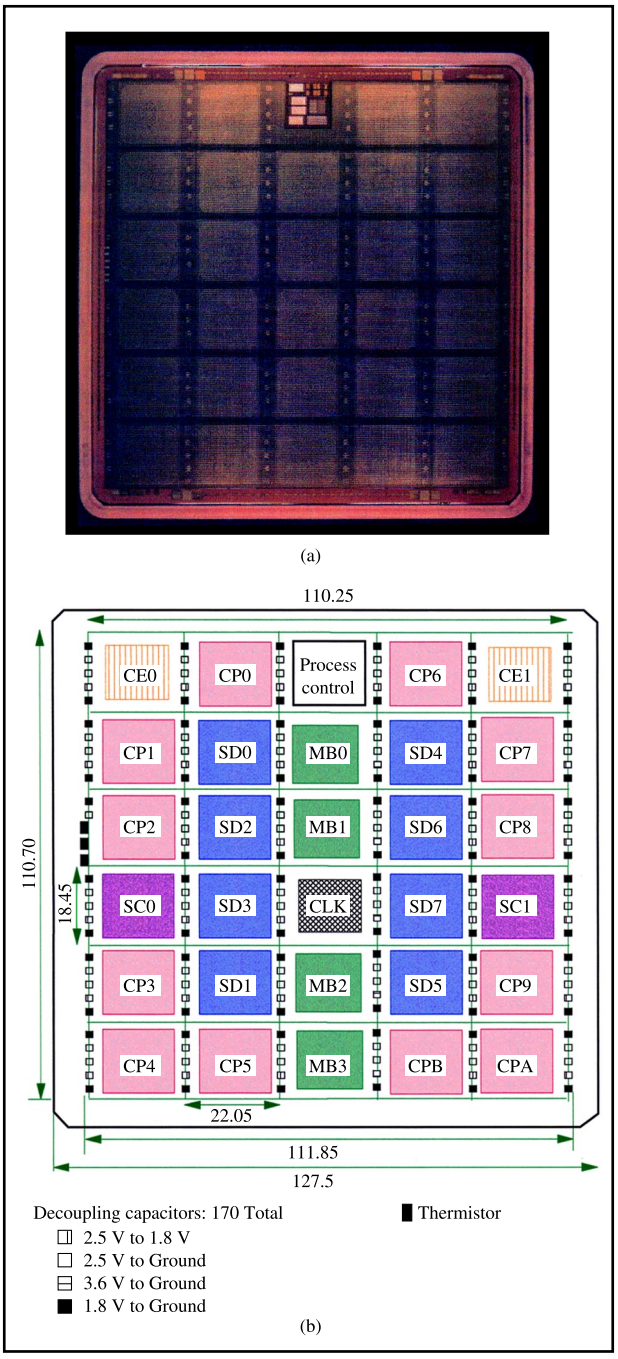

Figure 2

(a) GEMI top surface; (b) GEMI floorplan.

of the chips contained on the GEMI MCM that are important for the specification of the first-level package and its cooling requirements are shown in **Table 1**.

A fully shared architecture, where all 12 CPs are connected to all SD chips, would give a few-percent boost to the single-image system performance as compared to a binodal architecture. However, the fully shared architecture would increase the amount of required MCM wiring by 40% and would produce a proportionally more complicated module that would not meet the system's cost–performance objectives. A binodal architecture was therefore chosen for the G5 system.

Figure 1 shows that the G5 server provides for 24 self-timed interface (STI) ports for the connection of the CEC to the system I/Os coming from the four CMOS chips denoted as MBA. Each port consists of 40 wires organized in 20 differential signal pairs: eight pairs for a byte of data, one for the parity bit, and one for the clock that travels along with the data. Six of these STI macros, along with the time-of-day logic and the external reference connections, are implemented in each MBA chip, which measures 12.9 mm on a side and has 629 signal I/Os.

For the G4 system, the two cryptography coprocessors were packaged on two SCMs on the board. For the G5 server, the two coprocessors were moved onto the MCM, and each coprocessor was connected to two CPs, one in each node, so that the availability of the cryptography function was protected from any single CP malfunction. The move of the cryptography engines onto the GEMI MCM reflects the maturity of the coprocessor architecture at the time of the G5 system release, as well as our desire to reduce the overall system cost by removing the two SCMs used for packaging the coprocessors.

The clock chip was also brought onto the GEMI MCM and used to distribute the required testing control signals to the rest of the chips on the MCM. This approach increased the number of required signal I/Os for the clock chip to 746 but reduced the MCM signal I/Os by reducing the number of test lines required to be brought into the GEMI to control the testing of the CMOS chips. The technology of the clock chip and MBA chips is one generation older than the processor technology, in order to reduce the cost and also because it produced chips with an excellent balance of the area needed for the I/Os and the area required by their circuits.

The last line of Table 1 indicates the salient features of the first-level package that houses the CEC. The 127-mm × 127-mm MCM size was used because it could contain all 29 chips making up the CEC of the G5 system. The top surface of the MCM is shown in **Figure 2(a)**, and the topology of the CEC's 29 chips on that surface is shown in **Figure 2(b)**. The site on the first row denoted as "Process control" was used to monitor and control the electrical properties of the GEMI substrate. As the manufacturing of this type of MCM matures, this process control site will be used to house additional CPs. In fact, the dimensions of the G5 chip set permit the arrangement

shown in **Figure 3**, where the chip sites on the first row are different from the chip sites in the rest of the MCM. Admittedly, this arrangement will have an impact on the MCM yield, but it permits a system using 14 processors. In addition, by using more advanced CMOS technology, one could double the cache size of this system from 8 to 16 MB on the eight available SD chip sites.

The GEMI MCM or its 14-CP extension can support a cycle time that is better than the required 2:1 ratio of the CP cycle time shown in Table 1. This allows the same MCM structure to support more than one machine generation. The maximum wire distance for the cycle-time-limiting interconnections is six chip pitches, or 130 mm (a 10% reduction from the G4 generation). The common-chip C4 footprint is a 225-μm half-populated grid. This area signal array footprint provides the best signal escape capability for full area connection, and has the best electrical characteristics, since there exists at least a 1:1 signal-to-power ratio at the C4 connection. The 1:1 signal-to-power ratio is also maintained between the power and signal vias in the MCM under every chip site. However, the glass-ceramic MCM vias are on a 450-μm straight grid; therefore, the thin film on top of the glass-ceramic MCM is used both as a space transformer from the C4-to-MCM via grid, and as the on-MCM chip-to-chip interconnection medium.

The cycle time is estimated using the equation from [4]:

$$\begin{aligned} \textit{Cycle time} &= \textit{electrical delay} + \textit{driver loading} \\ &\quad + \textit{line propagation} + (\textit{clock skew} + \textit{PLL jitter}) \\ &\quad + \textit{latch setup} + \textit{noise impact}. \end{aligned} \tag{1}$$

This MCM technology for refrigerated operation yields the following net performance when the terms of Equation (1) are replaced by the numbers obtained from our timing analysis of time-critical functions for the associated packaging media and the net topologies, as described in Section 6.

$$\text{Thin film: } \textit{Cycle time} = 935 + 110 + 750 + 340 + 200 + 180 = 2515 \text{ ps.} \tag{2}$$

$$\text{Glass-ceramic: } \textit{Cycle time} = 935 + 291 + 810 + 340 + 200 + 180 + (450) = 3206 \text{ ps.} \tag{3}$$

The number in parentheses in Equation (3) represents the amount of delay padding required to meet the minimum path delay requirements for this class of point-to-point nets. This delay adder is design-specific for a given bus, and should not be used for timing comparisons of nets on thin and thick film. In fact, ignoring this factor, the performance difference between these two net types is approximately 250 ps.

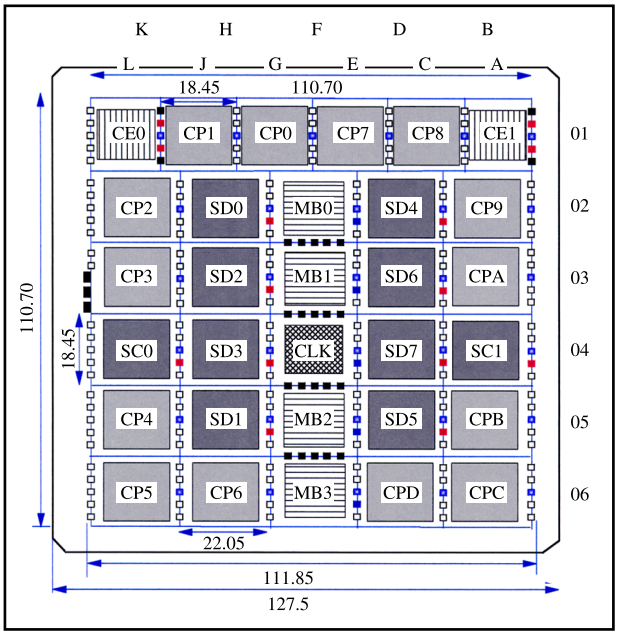

Figure 3

Future GEMI floorplan extensions for 14-processor design.

Glass-ceramic two-drop nets with no delay padding can support the following cycle time:

$$\textit{Cycle time} = 783 + 1000 + 833 + 340 + 200 + 180 = 3336 \text{ ps.} \tag{4}$$

The large driver/receiver capacitive loading is caused by reflections at the receiver nearest to the driver circuit (the first of the two receivers in this class of nets).

All of these cycle times are calculated for the chilled operating conditions of the GEMI. For air-cooled operation of the GEMI, the electronic delay component of Equation (1) increases, and the cycle times of Equations (2–4) increase by approximately 250 ps.

However, the on-MCM time is not the limiting cycle time for the performance of the G5 system. The off-MCM interconnections to the control chip on the main memory cards are the cycle-limiting connections. For this reason (and for control of mid-frequency noise magnitude), one has to consider the effect of the second-level package (board) on the electrical operation of the first-level package (MCM). **Figure 4** is an edge view of the board that houses the GEMI MCM. The ceramic capacitors for control of mid-frequency noise are mounted on the board surrounding the MCM. The low-frequency noise is minimized by electrolytic capacitors on the CAP and CAP/OSC card. The estimation approach for the number

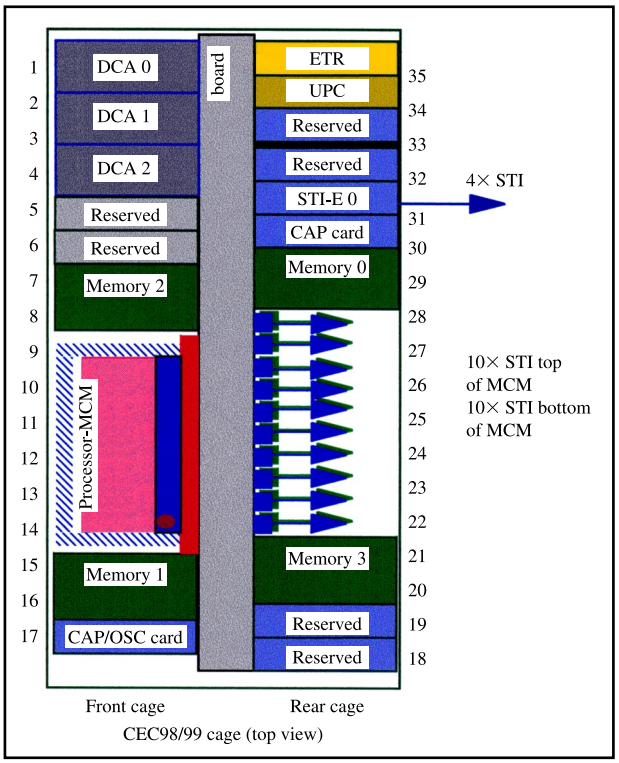

Figure 4

Floorplan of board that houses the GEMI.

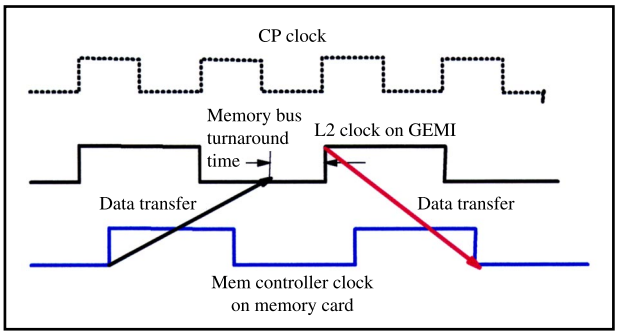

Figure 5

Time relationships for pumping the GEMI-to-memory bus.

and type of these capacitors is explained in a subsequent section. Twenty-four STI ports are available on the board: twenty ports on the back of the board, as shown in the figure, and four ports on the STI-E 0 card. DCA 0 through DCA 2 are the dc–dc power converters. Memory 0 to Memory 3 are the cards containing the main memory for this system. Each memory card contains up to 8 GB of memory, with two control chips and eight redrive chips. **Table 2** presents a technical summary of the features of this board.

The total distance of the interconnection from the SD and SC chips to the memory is minimized by appropriate tradeoffs between the board and memory card wiring lengths. The maximum on-MCM horizontal (e.g., "*x*–*y*") wire length for the memory interconnections is limited to 30 mm in order to minimize the magnitude of coupled noise on these time-critical nets. Since the per-unit-length signal propagation for the board and cards is the same, the worst-case total combined length for the memory interconnection is 250 mm. The interconnection delay to the memory card also includes the propagation delay through two connectors: the standard IBM Harcon connector under the GEMI MCM, and the right-angle AMP TBC+ card connector.

The Harcon connector contains an interstitial grid that has a main pitch of 2.2 mm × 2.4 mm. Of the 4224 total pins, 920 pins are used for the signal lines connecting the SD/SC cache chips to the controller chip on the main memory card, and 920 pins are used for power to provide the 1:1 signal-to-power ratio required for this interface. The STIs are also wired in a 1:1 signal-to-power ratio with one power pin for each differential signal, resulting in 960 signal pins (480 differential pairs) and 480 power pins. The rest of the pins are used for miscellaneous functions such as test and performance monitoring; one power pin is allocated for every two such signal pins.

The right-angle AMP TBC+ connector has 300 signal and 300 power pins arranged in a 1:1 signal-to-power ratio. A first-order timing analysis is enough to prove that the path delay from the SD to the main memory controller chip is greater than two CP cycles. Therefore, a bus-pumping scheme has been developed to overcome this deficiency. Specifically, the clocks of the memory control chips on the memory cards are generated so that they are 90° out of phase with respect to the clocks on the SD and SC chips. The use of the PLL circuits for clock generation on the G5 chip set eased the implementation of such clock phase-shifting. An inverter circuit is added in the feedback path of the PLL circuit used in the memory controller chips. **Figure 5** is a timing diagram showing the relationship between the clock edges on the CP, SD, and memory controller chips, and the timing of the data transfer from the SD chip to the memory controller chip. This bus-pumping arrangement allows us to achieve a cycle time of 80% of the interconnection delay, but it requires that the fastest interconnection be longer than 25% of the cycle time in order to avoid early-mode problems (the data arriving one cycle too early). Naturally, this is accomplished through judicious layout of the wiring

Table 2 Attributes of G5 board on which GEMI resides.

Board characteristic	*Value (dimensions)*
Footprint	W 550 mm × H 360 mm
No. of signal planes	12 signal planes, 2 mounting planes (w/o wiring)
No. of power planes	20 power planes
Total no. of nets	3811
Total wiring length	502 m (19770 in.)
No. of signal vias	16956 component vias 7611 vias (buried vias, through vias)
Type, number, value of ceramic capacitors	SMT 0805, 1690x, 1 μF SMT 1210, 354x, 10 μF
Type, number, value of electrolytic capacitors	0, all are on the CAP cards
Dielectric constant	3.85
Characteristic impedance	50 Ω
Propagation delay per unit length	67 ps/cm
dc line resistance	0.08 Ω/cm
Line capacitance per unit length	1.4 pF/cm
Line width × thickness	85 μm × 30 μm
Line-to-line spacing	102 μm (4 mil)
Dielectric layer thickness	V-S: 73 μm (2.9 mil) V-V: 114 μm (4.5 mil) S-S: 114 μm (4.5 mil)
Buried via diameter	305 μm (12 mil)
Via diameter	460 μm (18 mil)

Table 3 Characteristics of GEMI glass-ceramic substrate.

Total ceramic layers	75
Number of signal layers	34
Substrate metallurgy	Copper
Layer thickness (fired)	0.111 mm
Signal/via pitch	0.45 mm
Signal line cross section	70 × 25 μm
Ceramic dielectric constant	5.3
Signal line resistance	0.25 Ω/cm
Line impedance	60 Ω
Signal delay	77.5 ps/cm

Table 4 GEMI thin-film characteristics.

Total no. of metal layers	6
No. of signal layers	2
Metallurgy	Copper
Dielectric	Polyimide
Polyimide layer thickness	10 μm
Signal/via pitch	45 μm
Signal line cross section	18 × 6 μm
Dielectric constant	3.5
Signal line resistance	2.50 Ω/cm
Line impedance	38 Ω
Signal delay	68 ps/cm

of these nets on the board and the memory cards. By doing so we have achieved support of on-MCM interconnection cycle times of 3.6 ns and 3.3 ns for air-cooled systems and refrigerated systems, respectively, as discussed in Section 6.

3. Description of the GEMI substrate

The G5 MCM represents the most complex MCM-D product that IBM has ever designed and shipped to a customer. As previously stated, this MCM consists of a 75-layer glass-ceramic substrate with a deposited six-level thin-film structure. This section discusses in some detail this composite structure and the design philosophy adopted. The physical characteristics of the structure are given in **Tables 3** and **4**.

- *Material selection*

The choice of glass-ceramic (GC) material for the substrate was made after exhaustive evaluation of several options. The first use of GC by IBM in MCMs was in the

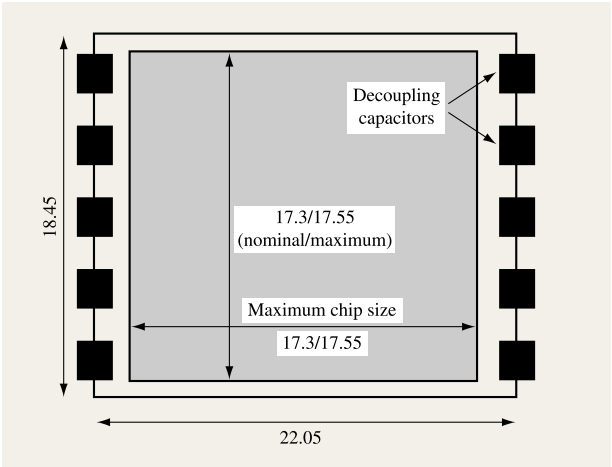

Figure 6

GEMI chip-site structure.

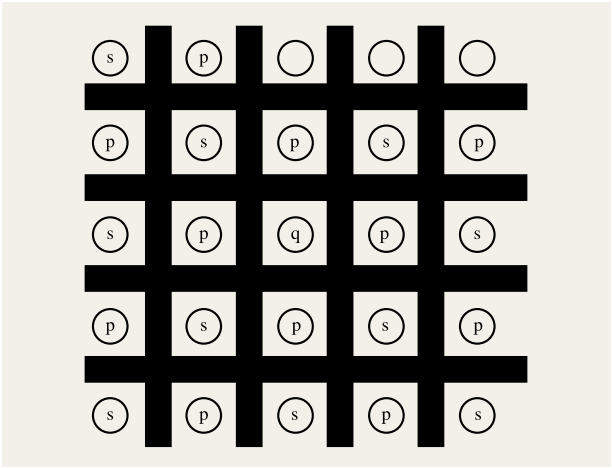

Figure 7

GEMI ceramic via pattern. A matrix of 5 × 5 vias is shown, and signal and power vias are interleaved. Twelve active signals couple into the middle quiet signal via.

ES/9000* H2/H5 bipolar machines. These substrates, known as the ARGO/Hercules MCMs, were developed during the late 1980s and shipped in 1990 [5]. Such substrates consisted of 67 layers of GC and had a two-level deposited thin-film (TF) structure. The TF structure was used for signal redistribution only; no chip-to-chip TF wiring was used. Thus, all of the intrachip wiring nets were in the GC material. GC was chosen for this application because the cycle time of the H2 CPU was in part determined by the chip-to-chip delay. GC material has a dielectric constant of 5.3, compared to a dielectric constant of 9.5 for the previously used 9211 ceramic material. This results in a time of flight (ToF) of 7.8 ps/mm for the GC, compared to 11.8 ps/mm for the 9211 material. This performance advantage of GC made possible a significant performance advantage for the H2 CPU.

However, from a manufacturing perspective, GC is a somewhat more difficult material to handle. Because the GC material does not shrink during the sintering cycle (9211 material shrinks 17%), features must be punched and screened on a tighter pitch in GC to achieve the same final feature dimensions as 9211. GC material is also not as strong a material, and special techniques are necessary for the attachment of I/O pins, providing a smooth surface for C-ring sealing, etc. All of these characteristics tend to make GC a somewhat more expensive material to use for a given application compared to 9211.

As technology evolved beyond the era of bipolar machines, the first CMOS machines were designed with "simple" (i.e., no TF) 9211 MCMs. For these machines, low cost was the primary concern. Since there were no CPU- or system-cycle-time-limiting paths on the MCM, 9211 was the best choice. As the performance and density requirements for these CMOS MCMs increased, a four-level deposited TF structure was added to the G3/G4 MCMs [4]. These MCMs had 65–70 layers of 9211 with a four-level deposited TF structure. As in the ARGO/H2 MCMs, the TF was not used for chip-to-chip wiring. All intrachip wiring was done in the 18–20-plane-pair 9211 structure. The four-level TF structure was used for signal redistribution and EC/repair operations, and to provide a low-inductance path between the chips and the decoupling capacitors.

The performance and density requirements for the G5 system were such that there would be system-cycle-time-limiting nets in the MCM. Our initial evaluation indicated that using only 9211 material would significantly affect system performance from the perspective of time of flight and system noise. IBM had previously developed a full-field TF capability that was used on a moderately sized (64-mm) MCM for an AS/400* application. In this application, a five-level TF structure was deposited on a 30-layer 9211 substrate. For this application, all chip-to-chip nets were in the TF wiring layers. The 9211 substrate served to provide the external I/O, power distribution, and mechanical structure for this application. The TF structure uses Cu for the conductive layers and polyimide as the dielectric. The particular polyimide used had a dielectric constant of 3.5, yielding a ToF of 6.8 ps/mm.

To meet the performance and density objectives of the G5 machine, we had several options:

1. A 9211 substrate with intrachip wiring AND a six-level TF structure with a plane pair of wiring with sufficient

wiring density to contain all of the "critical" nets. This would require a very high-density TF wiring pitch, which would be difficult to yield.

2. A GC substrate with four levels of TF for redistribution and repair/EC. All of the intrachip nets would be located in the GC substrate. This option would require us to manufacture a far more complex substrate than had ever before been tried and would still require four levels of TF.
3. A GC substrate with a six-level TF structure, where the wiring load would be "shared" between the TF and the GC substrate.

After carefully evaluating the cost–performance of each of these options and considering the requirements for the next several generations of systems, it was decided to pursue option 3.

- *Design*

The substrate size of 127.5 mm was chosen because it is compatible with our manufacturing tool set, it could meet the system requirements for the number and size of chips, it could support the number of I/O pins needed, and we could use the same hermetic sealing technology developed for ARGO. The top-surface layout with chip assignments is shown in Figure 2(b); Figure 2(a) is a photograph of the substrate (without chips).

One of these chip sites is reserved as a process monitor site, used to ensure the integrity of the TF structure as it is being built. This process control site has chains of varying-diameter vias to assess via integrity, chains of varying-width lines/spaces to access opens/shorts, area plates of metal/polyimide to access dielectric integrity and thickness, etc.

Chip sites

Because this project was on an extremely tight schedule, it was decided to simplify the design as much as possible. Although there is a significant variation in chip size (10.5–17.3 mm), it was decided to divide the substrate into a 5 × 6 matrix of identical chip sites, as shown in **Figure 6**. This figure shows the GEMI chip site with the maximum size chip and the controlled collapse chip connector (C4) decoupling capacitors.

Each site is 22.05 mm × 18.45 mm, which yields a 49 × 41 array of ceramic vias on a 0.45-mm pitch. To minimize noise and provide a controlled impedance, the ceramic vias were allocated in the pattern shown in **Figure 7**.

Ceramic cross section

The ceramic cross section of the design is shown in **Figure 8**. The top four and bottom four layers are used for planarization to ensure flatness and parallelism. The next two layers are for paste transition, and the next two

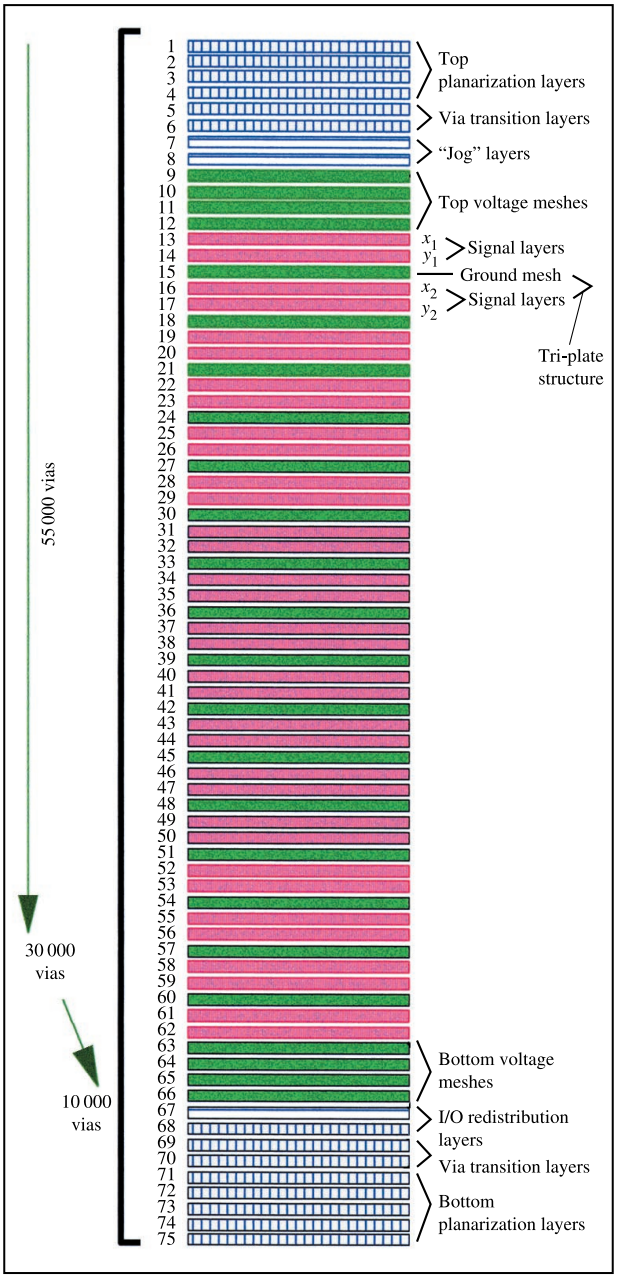

Figure 8

GEMI glass-ceramic cross section.

are "jog" layers to ensure via integrity. The next three layers are the voltage mesh layers to ensure reference plane integrity. The next 50 layers contain the *x*–*y* wiring layers and the mesh layers to create the tri-plate structure to ensure impedance and noise control.

We refer to this *x*–*y*-mesh triplet as a plane pair of wiring. For this design, each plane pair has 245 channels

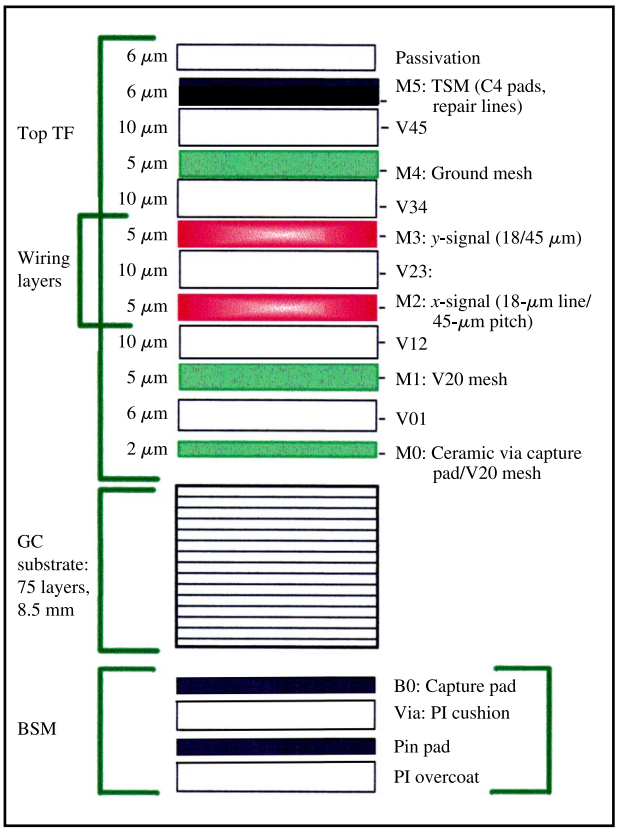

Figure 9

GEMI thin-film cross section.

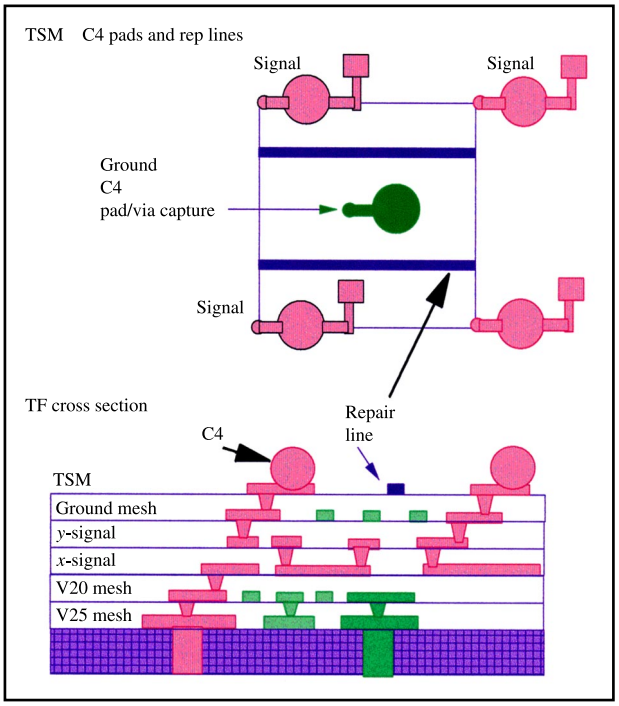

Figure 10

Detail of GEMI thin-film structure.

in each direction (x and y), with a length of 110 mm for a total wiring capacity of more than 900 meters. As is typical in "random" wiring, the actual utilization is considerably less than the available capacity, and, as discussed later, in Section 5, the actual wiring utilization for GEMI is 39%. This is due to a variety of factors: the basic "efficiency" of the autorouting, and the need to meet timing and noise constraints.

Thin-film structure

The thin-film structure shown in **Figures 9** and **10** is a nonplanarized TF structure, and so a "staggered" via technology is employed. The first Cu layer is 2 μm thick, patterned using a subtractive etching process [6]. This layer serves a dual function: It provides a "capture pad" to cover the ceramic vias and also creates a mesh layer for the 2.6-V power supply. The second Cu layer (5 μm thick) is patterned using a plate-up process. This is the V20 mesh layer. The next two layers are the x–y signal layers. On these layers, the signal lines are 18 μm wide on a 45-μm pitch. The fifth Cu layer is the ground mesh. The last layer, sometimes referred to as the TSM (top surface metallurgy), also serves two purposes: 1) to hold the C4 attach pads which connect the chip C4s to the module metallurgy, and 2) to provide a pattern of x–y lines and other features needed for repair and EC activity [7].

To enhance the reliability of the 4224-I/O pin connector system, thin-film layers are also added at the bottom of the substrate prior to pin attachment in a procedure known as the "policushion" process. These layers are labeled as the BSM (bottom surface metallurgy) in Figure 9.

The top surface and cross section of a small area of the TF are illustrated in Figure 10. Here we see a power C4 surrounded by its four nearest-neighbor signal C4s. We also see the features attached to the signal C4 pads which enable the EC/repair actions to take place. These features and structures on the TSM/M4 layer enable the creation of engineering changes (ECs) or the repair of thin-film and/or ceramic signal nets.

Synergy of technologies

The use of a composite structure consisting of both a high-density thin-film component and a high-function glass-ceramic substrate has led to many benefits: 1) It supports a very high-speed, high-bandwidth bus; 2) it supports EC and repair capability; 3) it is highly manufacturable; 4) it is readily extendable for future product requirements; and 5) the propagation delays for

long lines in thin-film and GC material are similar, which provides the MCM physical designer with the maximum flexibility for the placement of time-critical nets.

4. Noise-budget-based design

A 300-MHz bus cycle time in a robust system has been achieved, in part by minimizing the impact of noise on cycle time and removing the possibility of intermittent noise spikes causing logic errors. Planning in the early stages of design was essential to meet the requirements in a timely and cost-effective manner. The system noise budget was defined, the circuit parameters were specified, and enough packaging capacity was made available for placing capacitors and rerouting wires to minimize noise and reduce the risk of noise-induced failures.

- *Noise tolerance*

The processor chip is built using the latest 2-V CMOS technology to achieve maximum performance. The supporting logic chips are built using 2.6-V CMOS technology, which is available at a lower cost and lower risk. The communication between these chips is accomplished using bidirectional drivers. Three broad classes must be examined for noise tolerance: 2.6 V driving 2.6 V, 2.0 V driving 2.6 V, and 2.6 V driving 2.0 V. All processor-to-processor communication is done through the L2 cache chips, so there is no 2.0-V-to-2.0-V communication. The 2.0-V driving and receiving from 2.6 V is done using reduced-swing drivers and receivers on the 2.6-V chip [8]. This simplifies the design to the extent that each chip requires only one voltage level, and this is done with a single voltage distribution on each chip that supports the core logic and the off-chip driver and receiver circuits. Having the off-chip drivers share a common voltage level with the logic core simplifies power distribution design on the chip and the first-level package. It also gives us a higher-performance design, because the total noise amplitude on the core logic is reduced even when the off-chip drivers are considered.

The off-chip drivers have slew-rate control to limit the maximum crosstalk noise on the off-chip nets while minimizing the delta-I noise impact on the signal delay through these drivers. The receivers are designed with hysteresis to increase the noise tolerance for package

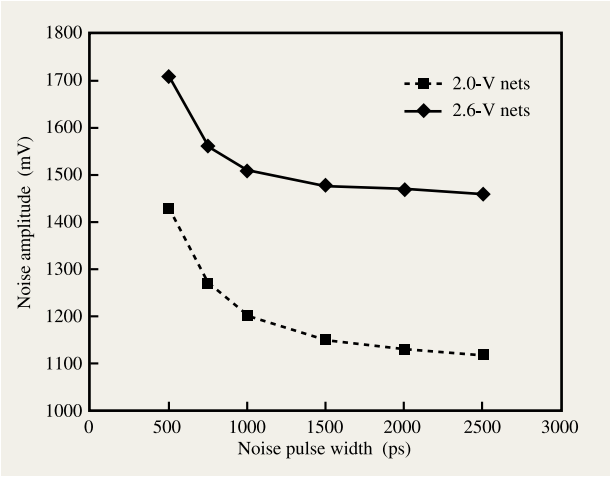

Figure 11

Nominal noise immunity curves for 2.0-V and 2.6-V receiver circuits.

interconnects. This strategy affords us interconnects that are much less sensitive to noise, with a minimal impact on total path delay. The 30-Ω drivers are designed so that their nominal dV/dT does not exceed 4.5 V/ns on a 50-Ω transmission line. The receivers are designed to accept the appropriate voltage swing of 2.0 V or 2.6 V with 150 to 200 mV of hysteresis, as summarized in **Table 5**. The ac receiver noise tolerance is shown as a function of pulse width in **Figure 11**. This choice of driver impedance, dV/dT, and hysteresis gave us the design point for which the worst-case crosstalk in a bus configuration with full coupling along the entire net is within the noise tolerance, and the performance goals of the interconnects were maintained.

Table 6 itemizes the resulting noise budget. This noise budget is deduced from the driver-voltage slew rate (dV/dT), noise tolerance, and experience from previous designs, and it accounts for the topologically worst-case values.

- *Power distribution noise*

The noise due to current fluctuations in the power distribution, a key component of the total noise,

Table 5 CEC CMOS receiver thresholds and noise tolerance in the G5 system.

Receiver type	*Threshold/hysteresis*	*Noise tolerance (750-ps pulse width)*	*dc noise tolerance*
2.0 V_{DD}, 2.0-V swing	$V_{DD}/2 \pm 150$ mV	1100 mV ± 33%	1000 mV
2.6 V_{DD}, 2.6-V swing	$V_{DD}/2 \pm 200$ mV	1450 mV ± 33%	1340 mV
2.6 V_{DD}, 2.0-V swing	$V_{DD}/2 \pm 150$ mV	1100 mV ± 33%	1000 mV

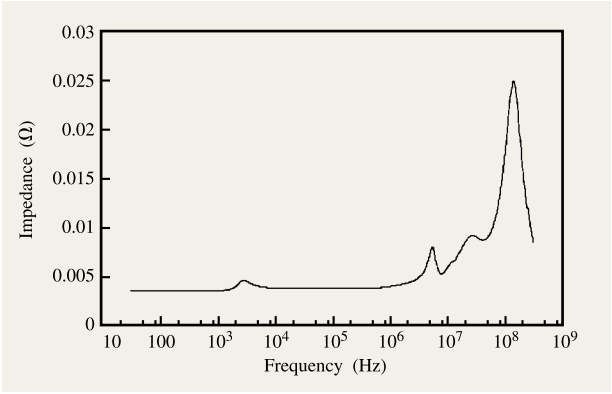

Figure 12

Power distribution vs. frequency curve for G5 system.

must be managed across the full frequency spectrum. A representative plot of the power distribution impedance, looking back into the system from a circuit on a logic chip, is shown in **Figure 12**. This plot is derived from a circuit simulation, in the frequency domain, of a model of the package components. The model consists of resistors, inductors, and capacitors representing the chips, MCM, board, and voltage regulator, as well as decoupling capacitors at all packaging levels including the chips. The noise source is a sine-wave current source of unit amplitude at each frequency, and the impedance plotted is the voltage across this current source. This model is what we call the mid-frequency model later in this section; therefore, the high-frequency response provided by this model and shown in Figure 12 is not very accurate because of the assumptions used. These assumptions were necessary in order to represent the board structure, which is essential for the accuracy of the results in the mid-frequency range (e.g., 1 to 100 MHz). The high-frequency response is also discussed in more detail later in this section.

Figure 12 shows three distinct frequency ranges where the power distribution noise must be managed: low-frequency (below 1 MHz); mid-frequency (1 MHz–100 MHz); and high-frequency (above 100 MHz).

The final decoupling capacitor implementation is summarized in **Table 7** for the 2.0-V supply for the processor chips. The logic chips using the 2.6-V supply require about half the current and produce half the current deltas that occur on the 2.0-V supply; therefore, roughly half the decoupling capacitors are needed for these chips.

Low-frequency noise

For the G5 system, the power-supply voltage can deviate from its nominal value by ±100 mV at the circuit terminals on the processor chip. We call this the voltage tolerance. This variation includes the voltage regulation tolerance, a resistive drop from the set point on the board just outside the MCM pins to the circuit, and the low-frequency ac noise. This voltage tolerance defines the range for which the worst-case noise immunity curve is generated. The low-frequency ac noise is the power-supply voltage variation due to changes in current demand on the power regulators.

The low-frequency ac noise accounts for ±50 mV of the voltage tolerance. The design of the switching regulators provides for the current demands of the system for each of the voltage levels. There are three high-current voltage levels: the processor voltage of 2.0 V, the supporting logic chip voltage of 2.6 V, and the memory voltage of 3.3 V.

Table 6 Noise budget for MCM and memory buses.

Noise component		*On-MCM nets* (mV)	*Off-MCM nets* (mV)
MCM	Horizontal wire	730	130
	Vertical wire (vias)	220	110
	Delta-I	150	150
Board	MCM connector		200
	Horizontal wire		100
	Card connector		130
Card	Horizontal wire		35
	SCM (crosstalk/delta-I)		200
Reflection and superposition			240
Total noise at receiver		1100	1305

The bulk capacitance for the regulators is supplied by electrolytic capacitors on two separate cards in the system. The two capacitor cards have 180 electrolytic capacitors each and provide the bulk capacitance for all three high-current supplies. The electrolytic capacitors chosen have a value of 560 μF and a maximum ESR (effective series resistance) of 56 mΩ.

For system availability requirements, the regulator design includes redundancy. Two regulators are required to run the system, and three regulators are available. If any one of the regulators fails, the system continues to operate on the remaining two regulators. Therefore, we needed to decouple one third of the maximum operating system current with the bulk decoupling capacitors and stay within a 50-mV ac power-supply variation. If the current deltas during the rest of the system operation can be held to less than one third of the maximum current, this is the limiting case.

Under normal operation, the maximum current variation that the S/390 processor can generate due to the number of switching circuits is 20% of the maximum current. This 20% factor has been maintained from generation to generation of S/390 CMOS processors and is less than the current variation expected when one of the three regulators fails. However, this is not the case when the system makes a transition from a low-current state to a full operating state. Therefore, the system's assist processor is programmed to stagger the starting of the clocks in the system so that during the power-on-reset sequence, the maximum current variation cannot exceed one third of the maximum current.

To determine the number of capacitors needed on the decoupling cards, a circuit model consisting of the resistances of the board, the capacitor cards, and the capacitors was built. The circuit was driven with a step of current equal to one third of the maximum current, and a simple regulator model was built that simulates the regulator responding to the change in voltage at the set point. The capacitors are required to store enough charge to run the system until the voltage regulators can respond and supply the additional current. This model is shown in **Figure 13**.

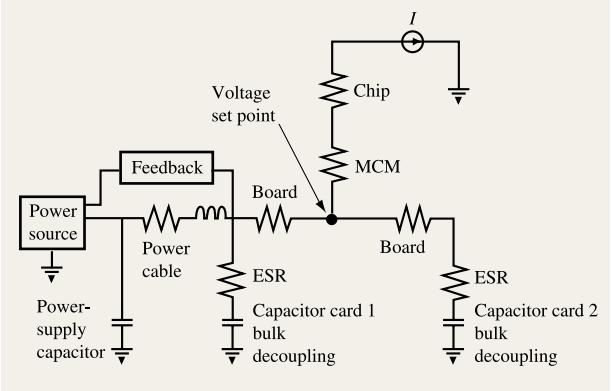

Figure 13

G5 low-frequency electrical model.

For the 2.0-V supply on the G5 system, the current delta that must be decoupled is 100 A. The board resistance in Figure 13 is 0.18 mΩ between the MCM and the capacitor cards. Each capacitor card has a total of 56000 μF each, with an effective ESR of 0.84 mΩ including 0.3 mΩ of resistance for the card connector. The output impedance of the three regulators is 1 mΩ with 80000 μF, and the feedback loop has a delay of 20 μs.

This model is coded into the circuit simulator, AS/X, along with a delta-I current source and a regulator model in order to determine the power-supply voltage variation. For sizing and design purposes, we estimate the peak voltage drop in the following way. The equivalent

Table 7 Decoupling strategy for the 2.0-V supply in the G5 system.

	Noise amplitude (mV)	*Placement of capacitors*	*Type of capacitors*	*Number of capacitors*	*Amount of decoupling at 2.0 V*
High-frequency ΔI = 5 A	100	Chip	Thin oxide	50 mm × 2	200 nF on chip
		MCM	200-nF ceramic	90	18 μF on MCM
Mid-frequency ΔI = 60 A	40	MCM	200-nF ceramic	90	18 μF on MCM
		Board	1-μF ceramic 10-μF ceramic	920 250	920 μF on board 2500 μF on board
Low-frequency ΔI = 100 A	50	Capacitor cards	560-μF electrolytics	200	112000 μF on capacitor card

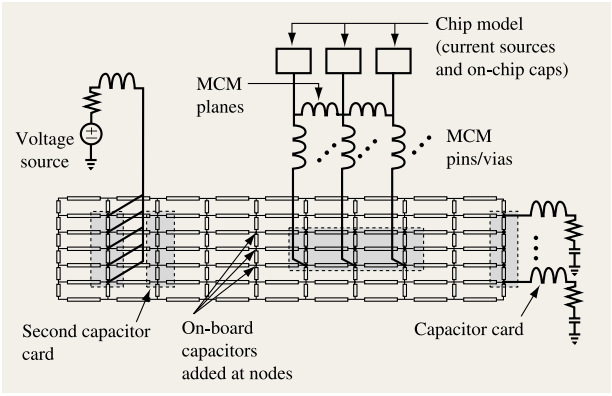

Figure 14

G5 mid-frequency circuit model.

resistance looking into the board from the set point near the MCM pins is 0.4 mΩ, so 100 A will produce a 40-mV voltage drop until the regulation recovers. In addition, the charge being drawn from the capacitors will cause the voltage of the capacitors to drop. From the equation, $I = C * \Delta V/\Delta T$, we approximate $\Delta V = I * \Delta T/C =$ 100 A * 20 μs/192000 μF = 10.4 mV. The total voltage drop is the addition of the 40-mV *IR* drop plus the 10-mV ΔV drop, or 50 mV.

For the 2.6-V and 3.3-V supplies, the objective is that the low-frequency ac variation should not exceed 3% of the supply voltage. A procedure similar to the one outlined above led to a design with 40 capacitors on each capacitor card for each of the 2.6-V and 3.3-V power supplies.

Mid-frequency noise

Mid-frequency noise, consisting of voltage variations that last for more than one cycle, also changes the operating power-supply voltage of the circuits on the chip and is factored into the voltage tolerance of these circuits. However, because of the higher-frequency components of this noise, the ESL, or effective series inductance, becomes a significant factor in the effectiveness of the decoupling capacitors. Capacitors with low ESL and ESR are required, and they must be placed close to the MCM.

The amount of current that must be decoupled is the 20% variation due to changes in circuit utilization discussed previously. The 33% low-frequency current delta does not require decoupling in this frequency range because its rise time is longer than the mid-frequency resonance time constant of the packaging structure. Furthermore, the low-frequency noise peak does not coincide with mid-frequency peak noise. If a maximum current delta occurs which collapses the voltage, an opposite-polarity current delta must occur next, and this helps the system voltage level recover. To ensure the ±100 mV total voltage variation, the mid-frequency noise should not exceed 50 mV at a delta current equal to 20% of the maximum current, or 60 A. Our design goal is 40 mV to provide some margin, because the 20% variation can occur rather frequently during the operation of the machine.

For the MCM and board of the G5 system, the fundamental resonance is near 10 MHz. This frequency is determined by the inductance of the package components and the capacitance of the decoupling capacitors. The mid-frequency noise is controlled by decoupling capacitors on the board and on the MCM. On the board, we choose to use two types of ceramic capacitors. The primary decoupling is provided by 1-μF ceramic capacitors in the 0805 body size. Because of the impedance curve shape shown in Figure 12, the number of capacitors required is estimated early in the design by calculating the impedance of a capacitor at 20 MHz, a frequency slightly above the expected mid-frequency resonance of the system. This impedance is obtained from the inductance, capacitance, and resistance of this capacitor in series with the inductance of the board via and the surface land to which the capacitor is soldered. We estimate this impedance to be 110 mΩ. The target is to place enough capacitors so that the 60-A current delta will produce a 10-mV voltage change at the vias of the decoupling capacitors. The number of capacitors required on the 2.0-V supply is then N_{cap} = (60 A * 110 mΩ)/10 mV = 660 capacitors. A similar calculation for the 2.6-V supply, assuming a 40-A current delta, results in a requirement of 440 capacitors on that supply. As seen in Table 7, this requirement is exceeded, and 920 of the 1-μF capacitors are placed on the board.

In addition, 10-μF ceramic capacitors in the 1210 body size were placed on the board to provide some bulk capacitance with a lower impedance than the electrolytic capacitors could provide at the MHz frequencies.

The mid-frequency noise response exhibited at the chip power-supply terminals on top of the MCM is affected by the inductance of the pins and vias in the MCM, and the decoupling capacitors on the top of the MCM. In comparison, the decoupling capacitors on the board have a much lower impedance. For the first time, 200-nF C4-attached decoupling capacitors were used on the MCM. The 100-nF decoupling capacitors used in the G4 system were modified by increasing the height and the number of plates in the capacitor to achieve a 200-nF value. The tighter space constraints for capacitors and the higher current demands compared to G4 made the introduction of this capacitor necessary to meet our noise targets. The above calculations provide a good rule of thumb, but a

circuit model and AS/X simulations are needed to understand the noise wave shapes and the frequency response of the overall system. (AS/X is an IBM proprietary circuit analysis simulation tool.)

With the design parameters in place, an AS/X model was built to analyze the mid-frequency noise. The key parts of the model are shown in **Figure 14** and modeled as presented in [3]. The board planes are modeled as a two-dimensional matrix of resistors, inductors, and capacitors. The high-frequency model of the MCM is simplified and connected to the board. The ceramic capacitors are placed in their relative positions, and simple models of the regulators and capacitor cards are added for completeness. This model is used to verify that the time-domain noise pulses are within the required limits. A frequency-domain analysis is also done, in order to determine the resonant frequencies and ascertain that the impedance objectives across the frequency band are met. This model can be used for noise-sensitivity analysis with respect to capacitor choice and placement. This sensitivity analysis can result in design changes.

Final verification is done by measurements on the system. A known current delta is created by starting the clocks. The resulting noise waveform is compared to the corresponding waveform obtained from model simulations. Also, the power-supply voltage variation is monitored during system operation with special test programs intended to stress the current variations in the system. With this design, we were able to keep the mid-frequency noise amplitude to 30 mV, which gives us sufficient margin to design a new MCM with higher current demands and still maintain the 40-mV objective.

As in the mid-frequency discussion of the G4 system presented in [3], providing a voltage with a minimum amount of mid-frequency noise is essential for limiting the phase error, or long-term jitter, of the PLL on the logic chips. The G5 system introduces a new noise-filtering circuit to limit the impact of mid-frequency noise on the PLL [9]. The 2.6-V power distribution is used to supply the voltage for the PLL, as in G4. A two-stage cascaded resistor and capacitor filter network is used to filter the noise. The first stage, which filters the mid-frequency noise, uses a 200-nF on-MCM capacitor for each chip, as shown in **Figure 15**.

High-frequency noise

While the low- and mid-frequency noise detracts from the dc noise immunity of the circuits, the high-frequency power-distribution noise is a portion of the ac noise on the interconnects. For off-chip nets, the delta-I noise is the high-frequency noise seen at the receiving circuits; the corresponding budget is shown in Table 6. This noise feeds through the 30-Ω source-terminated drivers and is reflected at the receivers, which have a 0.7 reflection

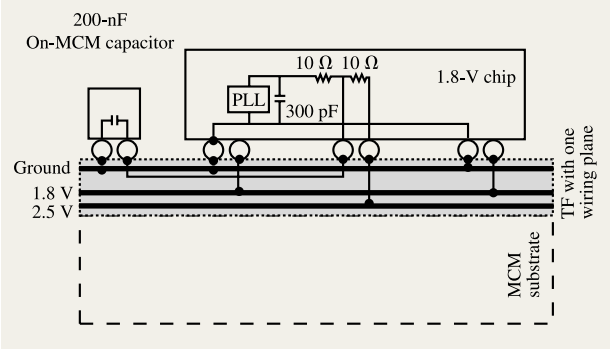

Figure 15

Filter network for power supply of PLL circuit.

coefficient. The 150 mV in the budget is the high-frequency noise that is superimposed on the crosstalk noise at the receiver input. The GEMI design, including the wiring rules, on-chip and on-module capacitor allocation, and signal-to-power via ratio, is such that this budget is met. The verification process described in Section 7 ascertains that the GEMI design meets the noise budget objectives long before any hardware is built.

High-frequency noise also affects the operation of on-chip circuits. Crosstalk checking of on-chip nets using static circuits has a noise budget as well, but this budget specifies the spatial delta-V, that is, the difference between the power-supply voltage at the driver and that at the receiver circuit with respect to the local grounds. The budget for this spatial delta-V is 150 mV peak to peak. However, for analog circuits, such as phase-locked-loop (PLL) or sense circuits, where the time-varying noise is important, this noise budget is 200 mV peak to peak on the processor chips, a variation of ±5% of the supply voltage. The noise control of on-chip interconnects is discussed in another paper in this issue [10].

The high-frequency noise is produced by the simultaneous switching of off-chip drivers and the core logic, which is fed by the same power distribution. For the 30-W processor chip, the high-frequency noise is dominated by the core switching noise; the off-chip drivers have a small impact on the noise appearing at the power-supply terminals of the core logic on a chip. The amplitude of the high-frequency noise is determined primarily by the ratio of switching capacitance to total capacitance on the chip. The design guideline is to add on-chip gate-oxide decoupling capacitors to supplement the intrinsic capacitance available from the substrate and other on-chip devices, so that there is a total capacitance equal to eight times the switching capacitance. Each switching circuit must have sufficient decoupling

capacitance within a 700-μm radius. The factor of 8 is based on the model of a voltage divider between switching capacitance and total capacitance. The 700-μm limit is derived from our measurement results [11], from which we observed that the effectiveness of the nonswitching capacitance falls off by 50% at 700 μm.

The processor chip has about 20 nF of switching capacitance and 150 nF of intrinsic on-chip capacitance. We assumed that the intrinsic capacitance was placed so that it was 75% effective in decoupling the noise. We also assumed that the added decoupling would be placed around the macros so that it would be only 50% effective. Under these assumptions, the requirement for on-chip decoupling can be approximated by the equation

$$8 * 20 \text{ nF} = (150 \text{ nF} - 20 \text{ nF}) * 75\% + C_{\text{decap}} * 50\%.$$

Solving this equation for C_{decap} results in $C_{\text{decap}} = 125$ nF. The final design of the processor chip contains 200 nF of added thin-oxide capacitance, but some of this is placed in the lightly utilized areas at the edges of the chip, where its effectiveness is expected to be less than 50% because it is farther than 700 μm from the chip areas with high switching activity.

Equally important for the minimization of high-frequency noise is to maintain the C4 leads on the chip as close to equipotential as possible. This is done by having a sufficient density of voltage and ground C4s, and by providing voltage and ground planes to supply the interconnections between these C4s. We take advantage of the thin-film technology to include robust voltage and ground mesh planes which provide a low-impedance interconnection that minimizes high-frequency noise.

The dimensions of the ceramic packages, including the 450-μm via pitch and the capacitance of the circuits on the chip, produce a resonance between 150 and 300 MHz for the power-supply impedance of the chips in G5. The use of thin-film power planes lowers the inductance of the package, shifting the resonant point slightly; more significantly, the impedance value in the resonant frequency range is reduced.

The actual power-supply delta-I noise is predicted by means of circuit simulations of models derived using the modeling approach described in [11, 12]. In short, the chip sites of the MCM are broken into a 5 $\times$ 5 array of cells. A model of the chip including the noise current sources, decoupling capacitance, and power-distribution grid is placed on the center 3 $\times$ 3 array of these cells. The MCM decoupling capacitors are placed surrounding the chip on the outer cells of the 5 $\times$ 5 array. The MCM is modeled using a three-dimensional inductance calculator. The via model is built in three sections, with an expanding area to simulate the current spreading out as it flows away from the chip. The mesh planes in the MCM are modeled to interconnect the sections of the chip and the decoupling capacitors. The MCM pins are modeled, but the board is considered to be an equipotential surface for these high-frequency simulations.

Since a 3 $\times$ 3 chip model cannot accurately reflect voltage variations on a local scale, a detailed model of the resistive on-chip power grid is used for that purpose. However, the model is used to determine the noise waveforms produced by the logic cores and the off-chip drivers at the operating frequency. The model properly represents the voltage variations and the noise attenuation as they propagate from the noise source. These waveforms and an exponential equation fit of the noise attenuation are inputs for the DELI program discussed as part of the MCM net noise-checking procedure in Section 7.

Measurements and simulations on the product in the system show that the peak-to-peak high-frequency noise is between 80 and 100 mV in the operating range of the processor, 2.0 to 2.5 ns, within the $\pm$5% variation objective.

5. Physical design methodology

The physical design (PD) methodology used to design the GEMI module was one of the keys to the success of the design. The PD problem for the module was very complex, and it required congestion optimization of the interconnect netlist between the thin-film wiring medium and the glass-ceramic, and application of net properties to the split-netlist data in order to drive the detailed wiring. After the actual wiring of the module was implemented within the wiring rule limits, proprietary signal integrity analysis tools were employed to check each net for compliance with timing and noise limitations. All of this pre-PD analysis, wiring, and post-PD checking was done with the absolute schedule requirement that only a single design release into the manufacturing facility could be done for this MCM. The key module design considerations are highlighted in the following.

- *Design complexity and manufacturing time*

The decision to design the GEMI module with two wiring media necessitated a long manufacturing time for the substrate in comparison to the silicon chip manufacturing time. The design required the definition of two separate rule sets for the detailed description of the wiring environment, and four person-months of effort. The actual manufacturing turnaround time of the GEMI spanned a period of six calendar months, creating the schedule requirement that only a single design pass and release to the factory could be allowed for this MCM. Therefore, the PD methodology employed for this design had to allow for fast completion of wiring and individual net signal integrity analysis, while guaranteeing that the design was both logically and physically correct.

• *Optimization of wiring congestion*
The effective use of the two wiring media became paramount for the design in order to meet the goal of a single, correct manufacturing release. The challenge was to allocate the GEMI's 10541 nets between thin film and thick film while still adhering to the timing and noise limitations required for proper system performance. Additionally, the previously mentioned time constraints did not permit the use of a detailed wiring tool to optimize the wiring in the two media, find the congestion areas, and then swap the connections back and forth. Therefore, an early definition of the netlist allocation between thin and thick film was attained through the use of a global routing tool. This global router makes use of sparse, early design information, yet it is still capable of giving the module physical designer a reasonably detailed split of the wires between the two media.

• *Rigorous electrical and physical verification*
The requirement for a single design release demanded an extremely rigorous checking of the design. Because the module was required to function correctly at first power-on, every single net within the module had to be checked against the appropriate physical and electrical constraints. The tool set deployed for the module design is required to provide, on a net-by-net basis, information regarding the wiring rule, the net timing, and the net noise profile, as well as manufacturing checks and logical-to-physical verification.

• *Methodology design flow*
The methodology flow employed for the GEMI design is shown in **Figure 16**. This methodology allowed the PD team to quickly and successfully address the design issues discussed previously. The methodology can be segmented into four specific sections: Section A – High-Level Design and Logic Implementation; Section B – Wiring Tool Rules Generation, Detailed Wiring Congestion Analysis, and ALLEGRO Model Build; Section C – Detailed Wiring, Wiring Rule Checking, and Signal Integrity Analysis; and Section D – Final Logical-to-Physical Verification, RIT (Release Interface Tape—the design as sent to manufacturing) dataset creation and manufacturing checking, and final RIT signoff and release to manufacturing.

Flowchart Section A: High-level design
This step involves early analysis of possible chip and I/O pin placement options against expected system performance targets. Since the specifics of early noise budgeting are covered in Section 4, and the early timing methods in Section 6, this section concentrates on the early wirability analysis that is done as part of the high-level design step.

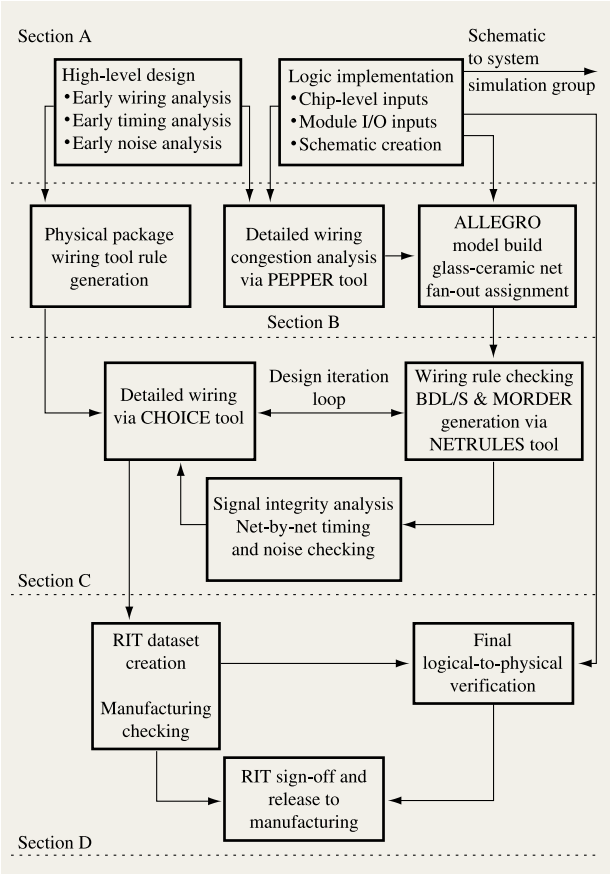

Figure 16

Flowchart of GEMI physical design process.

Early analysis of package wirability is achieved through application of the PEPPER (Program for Early Path Prediction based on Early Requirements) tool to the design point. Earlier work has established the method for this application [13]. The input to this tool is a basic technology description of the MCM, a wiring strategy, and a bundled netlist. The bundled netlist is a netlist in which all bits belonging to a single bus and emanating from a common origin are represented as a single data-line entry in the file. The basic technology description of the MCM utilizes a quantized grid cell approach to represent physical counts and location of C4 pins, module I/O pins, hybrid vias, wiring tracks, and chips. The wiring strategy classifies the netlist into detailed logic groupings in order to prioritize the most time-critical and difficult-to-wire nets (e.g., differential nets). Finally, a bundled netlist is created that allows for the fast analysis of wiring congestion (both localized and across the entire module). One can determine from the basic results of PEPPER the

Table 8 PINDATA file field description.

Property	*Usage*
Driver/receiver flag	Defines pin as driving or receiving at system level
Net logic name	Unique English name used for all package levels
Bidi/Uni flag	Indicates usage directionality mode at the system level
Signal pin	Physical pin name as defined in physical footprint model
Electrical book type	Code used to represent circuit for purposes of wiring-rule checking as well as later manufacturing testing
I/O circuit-to-signal-pin capacitance	Used during signal integrity analysis loop
x, *y* coordinates of I/O circuit relative to position of signal pin	Used during signal integrity analysis loop for delta-I analysis
I/O circuit book output or input capacitance (as appropriate)	Used during signal integrity analysis loop
Signal pin name	Nine-digit alphanumeric field used by the wiring tool to uniquely identify the signal pin

appropriate wiring density, hybrid via count, and number of plane pairs of wiring resource to be designed into the module.

Flowchart Section A: Logic implementation

Chip-level and module-level inputs are obtained through the use of a standard-format PINDATA file, which contains specific parameters that are required for both the physical design of the MCM and the noise and timing analysis that is done once the first-pass detailed wiring has concluded. The PINDATA file format allows all pertinent information for each used signal C4 on a chip to be passed to the MCM designer. Specifically, for each signal net, the PINDATA file contains the information shown in **Table 8**.

A logical schematic is then built within the Cadence COMPOSER** tool framework. COMPOSER is a design layout tool used by the chip design teams to generate their logic macro descriptions. This ensures that the MCM schematic is always completely synchronized with the chip I/O interface and that checking is completely automatic.

Additionally, the use of the COMPOSER tool allows for direct input of the module schematic information into the system simulation environment. This is another critical factor in achieving a fast, correct-the-first-time design release. While the subsequent steps of the methodology address the physical aspects of the design, the system simulation group is able to test the netlist interconnect against the design specification. Errors (if any) are fed back to the MCM design team early in the design cycle, thus preventing iterations of the wiring/signal integrity analysis phases of the MCM design.

Flowchart Section B: Physical package wiring tool rule generation

Manufacturing complexities of S/390 MCMs dictate a continued reliance on the CHOICE wiring tool (the capabilities of which are discussed in a later section), but this server-based tool requires an extensive rules environment for successful routing of signal wires. The data obtained from the high-level-design step of the methodology is used by the substrate designer as an input to the rules-development process. The substrate designer then creates a set of rules files that represent the MCM's physical parameters to the CHOICE wiring tool. These parameters include wiring plane-pair count; number and location of signal and voltage pins; number and location of signal and voltage vias; the wiring-track grid, length, and count; and representations of the I/O interfaces present between the module and the chips on the top surface as well as the interface connector on the bottom surface.

Flowchart Section B: Detailed wiring congestion analysis via the PEPPER tool

As discussed earlier, the PEPPER tool is used to analyze a "bundled netlist" in the high-level design step for the purpose of determining the required package design attributes. With the detailed netlist available, the PEPPER global router is used to perform the net allocation

between the thin-film and glass-ceramic wiring media. Multiple global router optimization passes are run, with the definition of nets allowed in thin film vs. nets allowed in the ceramic portion of the MCM refined after each pass. The result is a congestion-driven apportionment of the MCM nets between the thin-film and ceramic wiring media, which is used to drive the delivery of data to the CHOICE detailed wiring tool described below.

Flowchart Section B: ALLEGRO model build/glass-ceramic net fan-out assignment

Once the logic netlist is defined and apportioned between the glass-ceramic and thin-film wiring media, the package physical parameters have been determined, and the rules needed to support the CHOICE tool are in place, all that remains is to merge the logical schematic information into the physical description of the package. The S/390 MCM design team makes use of the Cadence ALLEGRO** tool to perform this step. ALLEGRO is an industry-standard tool for multichip packages, providing excellent design function in an easy-to-view, graphical format. The graphic nature of the tool is used to great advantage by the design team in assigning the fan-out for nets that must be wired in the glass-ceramic wiring medium.

ALLEGRO tool physical chip symbols are built based on the same PINDATA files that were used to generate the logic model in COMPOSER (this provides for logic equivalence between the COMPOSER tool and ALLEGRO tool databases). A basic ALLEGRO model is created that contains placed chip and I/O component symbols, the full logic netlist, and a "middle" component whose pins represent the location of discrete hybrid via transition points between the thin-film and glass-ceramic wiring media. These discrete transition points between the thin-film and ceramic wiring media are nicknamed "middle pins," and are referred to as such for the rest of this paper. No other physical data need be entered into the ALLEGRO model for this design methodology.

Finally, taking advantage of the aforementioned graphic capabilities of the ALLEGRO tool, internally written code routines are exercised to assign specific "middle pins" to the logic nets wired into the glass-ceramic. The assignment is done using a simple length-minimization algorithm based on the timing criticality of the nets, with timing-critical nets receiving first priority (and, hence, the shortest possible fan-out from the chip I/O to the "middle pin"), and less critical nets being assigned lesser priority. This algorithm is exercised until all required middle pins have been assigned according to the wiring rule limitations.

Flowchart Section C: NETRULES wiring rule checking

An internally developed tool, NETRULES, is used for the purpose of reading in the data extracts from the ALLEGRO model, evaluating the data against design wiring constraints, and then generating the appropriate files for input to the CHOICE wiring tool. The main feature of the NETRULES tool is a richly featured wiring rule set that allows the definition of many key constraints to the detailed wiring tool. Checking features include valid I/O circuit combinations, individual net length checks, group limits, differential net length limits, etc.

Once pre-physical design checking is complete, NETRULES is used to generate detailed logic descriptions and netlist ordering information for the CHOICE wiring tool. The ordering information breaks each net down into a sequence of point-to-point connections and assigns priority information to those segments. This data is input into the CHOICE tool, and the detailed wiring is completed. A design iteration loop is then entered in which a feedback file is passed from CHOICE to NETRULES, NETRULES generates a data model that is fed to the signal integrity analysis tools, and a net-by-net timing/noise check is performed. Failing nets are fed forward to the module designer for rerouting with the CHOICE tool, and the process is repeated. The details of net timing/noise checks, as well as the tools that perform these checks, are covered in Sections 6 and 7 of this paper.

Flowchart Section C: The CHOICE wiring tool

The CHOICE tool is an internally developed wiring and checking tool. The development of this tool was done in conjunction with the MCM manufacturing experts, and provides unmatched routing capability for "gridded" MCM designs. CHOICE provides the environment and functions necessary for wiring on complex, hybrid-medium designs (i.e., designs consisting of more than one wiring medium), using predefined discrete wiring channels in a gridded routing environment. Its functions include automatic wiring, manual or interactive wiring, and execution of physical checking against the final wired image. CHOICE runs under the MVS environment and can be run in either foreground or batch modes, which facilitates execution of routing runs on a round-the-clock basis. Not only do these outstanding characteristics provide rapid turnaround time for design iterations (the team has achieved a 24-hour design iteration cycle time), but the routing algorithm has never failed to provide a 100% automatic wiring solution on the first pass through the data. Another outstanding CHOICE feature is the first-pass avoidance of shorts, length violations, and net spacing problems, on datasets that are an order of magnitude larger than any industrial router can handle today. CHOICE provides extensive wired-image-checking capabilities, including (for example) min/max length, shorts, loops, and illegal T-junctions, as well as manufacturing checks that defend against via-to-via spacing violations and excessive wrong-way wiring.

Table 9 GEMI net types and delay predictor accuracy.

Bias supply/interface level (V)	*Net topology*	*Medium*	*Wiring rule range* (mm)	*Maximum regression error* (% of net delay)
2.0/2.0	On-MCM Point-to-point	Thin film	40–100	4
2.0/2.0	On-MCM Point-to-point	Glass-ceramic	30–140	3.6
2.6/2.6	On-MCM Point-to-point	Thin film	40–100	3.8
2.6/2.6	On-MCM Point-to-point	Glass-ceramic	30–140	2.1
2.6/2.6	On-MCM Distributed 2	Glass-ceramic	Wire 1: 30–70 Wire 2: 0–30	2.1
2.6/2.0	On-MCM Point-to-point	Thin film	40–100	4
2.6/2.0	On-MCM Point-to-point	Glass-ceramic	30–140	3.6

Table 10 Delay components for GEMI off-chip interconnections.

Net name	*Driver chip Simulated/estimated* (ns)	*Off-chip net Simulated/estimated* (ns)	*Receiver chip Simulated/estimated* (ns)	*Total net Simulated/estimated* (ns)
MBA ⇒ L2 Glass-ceramic net Wire = 95 mm	0.864/0.820	0.917/0.941	0.952/0.919	3.3/3.2
L2 ⇒ L2 Glass-ceramic net Wire = 135 mm	0.860/0.939	1.461/1.412	0.459/0.473	3.3/3.4
MBA ⇒ L2 Glass-ceramic net Wire = 115 mm	0.864/0.850	1.131/1.138	0.629/0.689	3.2/3.2
L2 ⇒ L2 Glass-ceramic net Wire= 122 mm	1.142/1.132	1.183/1.152	0.425/0.439	3.3/3.3
L2 ⇒ CP Glass-ceramic net Wire = 138 mm	1.299/1.378	1.351/1.298	0.440/0.438	3.6/3.6

Flowchart Section D: RIT dataset creation/manufacturing checks/logical–physical verification

Upon successful completion of the wiring/timing/noise design iteration loop, the RIT (release interface tape) datasets are created. These datasets are used to drive the manufacturing of the MCM. Prior to release, however, several levels of checking are done. Manufacturing checks that have been built into the CHOICE wiring tool (as discussed above) are run. Additionally, the wiring image output of the CHOICE tool is processed and compared with the original COMPOSER schematic, thereby guaranteeing that the final physical product matches the original logic input. This check, along with the logic simulation verification results, guarantees full functional compliance with the design specification on the first release of a module design.

- *GEMI module design results*

The GEMI module in product form was exercised through the physical design methodology described

above, involving a total effort of one person-year. The 12-processor design, involving 595.67 meters of wire, was contained within a single plane pair of thin-film wiring with 17 plane pairs of glass-ceramic wiring. The allocation driven by the intermediate PEPPER analysis resulted in 212.32 meters of wire in the thin-film resource (thin-film utilization for the final design was 37.29%), and 383.35 meters of wire in the less dense glass-ceramic resource (the average utilization of a glass-ceramic plane pair was 39.61%). This utilization data does not reflect the number of wiring tracks kept empty for the purpose of providing electrical shielding to sensitive nets such as differential STI nets and reference clock nets.

The module was released to manufacturing only once. Subsequent system testing revealed 100% functionality of the module. This fact provides the vindication for the rigor of the design methodology described herein as applied by the S/390 MCM design team.

6. Wiring rules and timing

As we have seen in Section 5, the GEMI MCM contains 10540 connections. The structure and performance goals are not the same for all of these connections, and consequently the methods for their timing differ as well. The on-MCM connections can be separated into three classes: the time-critical connections representing the data and control lines between SD/SC chips and the CP and MBA chips, non-time-critical multidrop connections between the on-MCM chips used for testing and other auxiliary functions, and finally the reference clock connections. The off-MCM nets can also be separated into three classes: the time-critical bus-pumped connections to the main memory controller chips, the STI differential connections to the cable connectors at the edge of the board (Figure 4), and nonswitching testing lines.

For the time-critical (either on- or off-MCM) nets, accurate prediction of the delay from the input of the output buffer (driver circuit) to the output of the input buffer (receiver circuit) is essential for accurate prediction of the cycle time that the system can support. Most of these time-critical connections were point-to-point connections with bidirectional drivers at the ends of the net. A number of two-drop near-end-distributed nets [14] were used for a small number of connections between the SC and a pair of CP chips in order to contain the number of signal I/Os of the SC chip within the bounds of the chip footprint described in Section 3. Closed-form delay predictors were developed for all of these nets using a least-squares fitting method with excellent results. **Table 9** lists type and fitting error for each of the main classes of these interconnections in both thin-film and glass-ceramic material. This table shows that the worst-case fitting error is less than 5% for the point-to-point nets, while the error for the first load on a two-drop distributed net is almost

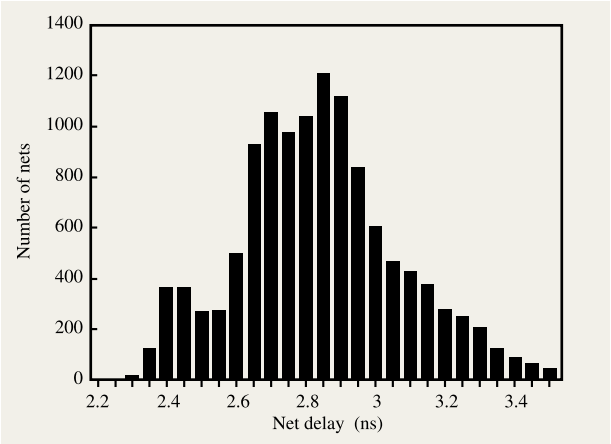

Figure 17

Off-chip net delay distribution for GEMI interconnections.

the same, namely 4%. Since these estimation errors are well within the delay tolerance of these nets that is produced by environmental and manufacturing variations, the closed-form predictors were used for the timing of the G5 system. The distribution of the cycle times supported by the nets on the GEMI is a result of the timing check of every net after the MCM physical design is completed; it is given in **Figure 17**. For a set of characteristic nets, **Table 10** provides a comparison of delays yielded by the off-chip timing methodology to that of a SPICE-like simulation.

The results in this table verify the accuracy of our timing methodology (error less than 7%) and prove that the GEMI MCM can support 300-MHz connections. In addition, one should note that the package delay of an off-chip path corresponds to 43% of the total path delay, another 40% of the total path delay is due to the CMOS circuits on the driver and receiver chip, and 17% is due to clock skew, PLL cycle-to-cycle jitter, and noise delay penalty. The combined effect of the clock elements (PLL jitter and clock skew) accounts for 10% of the total path delay. This is an inferred quantity and is not directly measurable. As such, it is expected to be pessimistic, as was shown by system measurements on the test floor that exceed the predicted cycle-time limits by a couple of hundred picoseconds. This path delay adder is computed as follows.

Since the off-chip path involves two on-chip clock distributions, and each one of them can have a clock skew of 150 ps (which represents 10% of the longest clock distribution delay), the total clock skew for the data chips is calculated as $150\sqrt{2} = 210$ ps. The skew between different drivers at the output of the clock chip is specified

at 80 ps, while the cycle-to-cycle PLL jitter is 240 ps and the noise impact 180 ps. The root-mean-square summation of the clock skews and noise delay impact is algebraically added to the PLL jitter which is always expected to occur during the machine's lifetime operation, and it results in a 527-ps adder to the path delay.

Although Table 10 depicts the delay predictions for a number of worst-case on-MCM interconnections by IBM's proprietary timing tools, at the completion of the MCM physical design, the delay of every on-MCM and off-MCM net is calculated. The technology and physical data required for these calculations are found in the following datasets or tools:

Pindata file: Provides the electrical and physical properties of the chip I/Os, including the buffer circuit type. A detailed description of its contents can be found in Table 8.
VIM database: Provides the electrical and physical properties of the interconnections on MCM, board, card, SCM, and connectors.
SLAM tool (see below): Provides hard-coded delay equations for each net type, with their corresponding coefficients and tolerance.

SLAM is a system path delay calculator that adds the off-chip net delays, which it calculates, to the delays of the latch, driver, receiver, and whatever other circuits may be in the path. The delays of these CMOS circuits exist in a set of special files generated by the on-chip path timer. The result of this addition is a "slack report" for each latch-to-latch connection analyzed by the SLAM. A slack report depicts the amount of time by which a latch-to-latch connection time may be shorter or longer than the target cycle time. In addition, SLAM provides the computed path delay and netlist in uniquely formatted files (*pio*, *netlist*) which are used to provide the required system timing and interconnection information to the noise verification tools described in Section 7. The results of the noise verification tools are translated to representative delay adders, which SLAM accepts to produce the final system timing slack report.

The second class of nets are the reference clocks. These are differential pairs emanating from the clock chip and connected to every chip on the G5 MCM and memory cards. Each chip receives one reference clock. Although the absolute arrival time of these signals is not important, their relative arrival time with respect to one another is. In particular, if T_{cm} is the time required for the clock signal to propagate from the clock chip to a logic chip on the GEMI MCM, and T_{cb} the corresponding time of the reference clock propagating from the clock chip on the GEMI MCM to any of the main memory controller chips on a memory card, it is desirable that $|T_{cm} - T_{cb}| = 0$. This implies that under nominal environmental conditions, the length for the reference clock connections should be such that its propagation time for the on-MCM interconnect is equal to the propagation time of the clock signal from the clock chip through some MCM wiring, MCM vias and pins, board wiring, card connector and wiring, and finally pins and vias of the SCMs that house the main memory controller chips, i.e., the off-MCM interconnect. Although this null relationship can be achieved with judicious selection of the on- and off-MCM lengths, it cannot be maintained at zero for all machines or even for the life of a given machine, because of the different sensitivities to environmental and manufacturing conditions of the on-MCM wires and the on-board or card wires. Therefore, the optimum lengths for the reference clock wires are obtained through trial and error from statistical circuit simulations for the structure depicted in Figure 5, until the quantity $[(T_d + |T_{cm} - T_{cb}|) - \text{cycle time}]$, where T_d is the path delay from the controller chip on the memory card to the SD/SC chips on the GEMI, is minimized. To minimize the effect of logic noise coupling onto the reference clock lines, and hence to reduce clock skew, all of the clock signals were wired on a dedicated plane pair at the bottom of the GEMI MCM and were surrounded by empty wiring tracks.

The third class of nets are the non-time-critical (cycle times greater than 20 ns) but multidrop nets. For these nets, wiring rules for the segment lengths are developed through circuit simulations so that there is no switching uncertainty after a predefined time interval. The correct implementation of these wiring rules is confirmed after physical design of the MCM through the use of a proprietary tool called NETRULES, described in Section 5.

Finally, the STI nets are differential pairs for which neither the absolute delay nor the relative delay among them is important, because of the self-timing nature of the source synchronous scheme used for the signal propagation. However, the correct sensing of the differential signal by the receiver at the end of a 10-m cable (system configuration design requirement) requires that very little noise be coupled on any of these differential pairs and that within each pair the imbalance of the lengths be very small. Consequently, the STI pairs are laid out with an empty track on either side of the differential pair; in addition, the length difference of the two wires within a differential pair is kept to less than 2 mm on the GEMI and 4 mm on its board. The proper performance of the STI nets was determined using detailed package, cable, and CMOS circuit models in large SPICE-like simulations. The results of these simulations were confirmed experimentally, and it was found that

these connections can support up to 14 m of cable and operate at 337 MHz.

7. Noise verification

For G5 and G6, both signal rise times and machine cycle time have been reduced to the point that signal integrity issues such as noise containment at the system level represent a significant challenge for the comprehensive verification of the off-chip nets. The total noise is composed of coupling noise and switching or delta-I noise. These noise sources are evaluated on all MCM and board nets to ensure that coverage is not compromised.

The noise verification process has been developed within the IBM System/390 Division over several generations of technology and machine designs [4, 15]. It is intended to provide a bounding calculation of the total noise and to identify nets which exceed their design limits for subsequent rerouting. The accuracy of such bounding calculation is a function of deterministic parameters, such as physical layouts, and statistical variations, such as switching time uncertainty. Given this objective, the noise verification process is structured as follows:

1. Electrical characterization of chip buffer circuits and package parameters from a noise perspective:
   a. Driver slew rate.
   b. Receiver noise margins.
   c. Coupling coefficients (mV/cm).
   d. Driver and receiver reflection coefficients.
   e. On-chip switching noise (both core and I/O).
2. Geometrical extraction of package layout and net topologies.
3. Use of a transmission-line algorithm to predict noise amplitude and arrival time at receiver inputs, based on driver switching times.
4. Use of a delta-I calculator to account for high-frequency on-chip power-supply noise at chip I/Os.
5. Statistical summation of the coupled and delta-I noise at every receiver based on driver switching time variations.

The resulting noise magnitude is a statistical quantity characterized by a probability function at every discretization interval of the cycle time. A net is defined as a failing net when the probability function of the corresponding noise magnitude at any discretization interval of the cycle time satisfies the following bivariate inequality:

$$P[V_n(t) > V_c] > P_c. \quad (5)$$

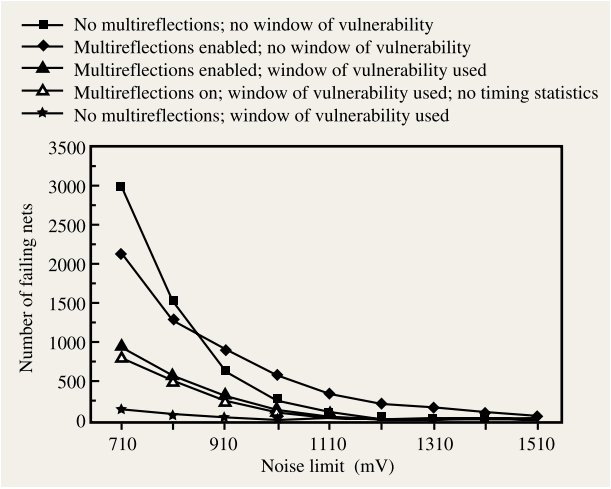

Figure 18

Noise magnitude distribution tail as a function of receiver noise tolerance.

This inequality states that if the acceptable probability of failure, P_c, is less than the probability that the noise magnitude, $V_n(t)$, is greater than the receiver's noise tolerance, V_c, the net must be rerouted. The acceptable probability of failure, P_c, is based on many factors: number of cycles during a machine lifetime, number of nets, etc. [16]. This number is typically of the order of 10^{-20}. **Figure 18** shows the results of the application of Equation (5) to the on- and off-MCM nets for the G5 systems.

• *Electrical characterization of chip and packages*

Coupling noise is a result of electromagnetic interaction from a switching or active line to a quiet one in close proximity. This electromagnetic interaction between a pair of coupled wire segments results in a forward or far-end wave traveling in the same direction as the active signal, and a backward or near-end wave that travels in the opposite direction [17]. To account for this interaction, capacitive and inductive coupling coefficients (K_c, K_l) are computed for each package level (MCM, board, card, etc.) using various field solver tools [18–20]. The outputs of these tools are the capacitance and inductance matrices used to derive the appropriate coupling coefficients such as $K_c = C_m/C_s$ and $K_l = L_m/L_s$, where

C_m = mutual capacitance between two signal conductors,
C_s = total capacitance of a conductor (mutual plus ground),
L_m = mutual inductance between two conductors,
L_s = self-inductance of a conductor.

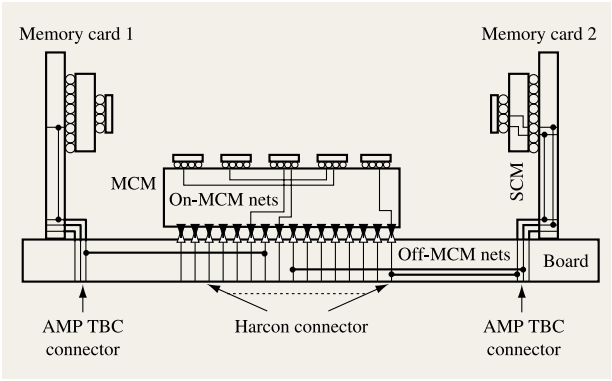

Figure 19

GEMI-to-memory-card controller chip interconnection schematic.

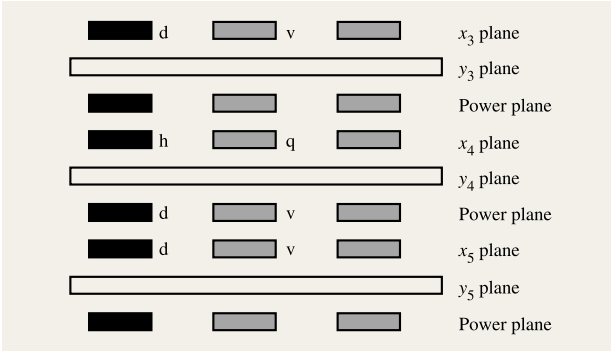

Figure 20

Cross-sectional view of three ceramic plane pairs in GEMI, showing horizontal, vertical, and diagonal wire adjacencies.

The per-unit-length coupled voltages, PULX, in mV/mm, are obtained from circuit simulations of the driver circuits and package models for the various net types that are used for the on- and off-MCM interconnections in the G5 system. The CMOS driver circuits on these nets are series-terminated. Reflection coefficients are calculated that account for the mismatches between the driver-circuit output impedance or the receiver-circuit input impedance and the characteristic impedance of the transmission lines that model the chip-to-chip interconnections. The receiver-circuit reflection coefficients derived from these simulations reflect the capacitive loading that exists at the input of the receiver due to the gate and ESD diode.

Several packaging components must be characterized before system-level noise analysis can be performed on the interconnects in the G5 system. These packaging components are the GEMI MCM, its board, the memory cards, and the complex connector structures used at package interfaces. **Figure 19** is a schematic illustration of the various package components used for the L3 memory communications to the L2 cache chips on the GEMI.

The mesh reference planes in the GEMI allow coupling between lines on either side of a power plane (e.g., vertical or diagonal coupling), due to the lack of electromagnetic shielding by such a power plane. Therefore, three contiguous wiring-plane pairs must be considered for the calculation of the PULX parameters of the *x–y* lines, as shown in **Figure 20**, where 50% other lines and vias (OLVs) are assumed [5].

The coupled noise in the via area is also important. In the via region, signal and power are interleaved in a checkerboard fashion, resulting in a 1:1 signal-to-power ratio to minimize coupling, as shown in Figure 7. The structure depicted by this figure is extended to a 5×5 arrangement of signal vias for the accurate calculation of the PULX values, as described in [21].

About 40% of the MCM nets are used for interconnections to the L3 memory cards. The board signal lines are the equivalent of strip lines because they are embedded among voltage and ground planes. Therefore, the far-end PULX value among the board signal lines is very small. Signal lines in the memory card have similar characteristics. The PULX values for these PCB packages are calculated using techniques similar to the ones used to determine the PULX of the MCM *x–y* wires.

A high-density zero-insertion-force connector known within IBM as the Harcon connector is used to connect the MCM onto the board, while a right-angle card-edge connector is used for the memory card/board interface. Connectors are sources of large coupling noise, primarily because these structures are nonhomogeneous. Because of their complex geometry and fixed lengths, the coupled-noise voltages are obtained directly from an application-specific field solver [22]. A geometrical abstraction of the Harcon connector used for its electrical modeling and coupled-noise determination is shown in **Figure 21**.

All of these noise-related quantities are stored in noise files, together with pertinent geometry and timing data, after the MCM, board, and cards are fully wired. The details of the contents of these noise files are described in the following subsections.

- *Geometrical extraction*

The determination of the position and length of all of the coupled segments associated with each net in a package component (MCM, board, or card) is the basis for the

geometrical extraction process. By treating each coupling interaction as a pairwise network, the geometrical extractor is extended to determine the physical lengths from the beginning and the end of each coupling segment to the driver/receiver pins for both the active and quiet nets [23].

Pin and wire database information is used in conjunction with PULX and time-of-flight parameters to compute noise voltages and delays for all coupling interactions associated with every net within the package. This results in the generation of a geometrical crosstalk file known as the GEO file. A typical entry within the GEO file contains the following:

1. Quiet net name.
2. Active net name.
3. Near-end coupled noise.
4. Far-end coupled noise.
5. Coupled-segment delay.
6. Delay from the active driver to the coupled segment.
7. Delay to the active receiver from the far end of the coupled segment.
8. Delay to the quiet driver from the coupled segment.
9. Delay to the quiet receiver from the far end of the coupled segment.
10. Reflection coefficient at active driver pin.
11. Reflection coefficient at quiet driver pin.
12. Reflection coefficient at active receiver pin.
13. Reflection coefficient at quiet receiver pin.

This extraction process is applied on all package components independently. For nets which are completely contained within the MCM, the required timed noise analysis can begin immediately. For off-MCM nets, i.e., memory nets, a "flattener" process is used to connect wire segments across several package boundaries and generate an equivalent GEO file across all the topologies involved.

- *Transmission-line effects*

With this information, the amplitude and arrival time of the far-end and near-end noise pulses at the receiver input can be predicted by employing transmission-line theory and taking into account the switching time of the active drivers. For each pairwise-coupled network on a victim line, the reflection coefficients at the driver output and receiver input of both the active and victim lines are employed in a multiple-reflection algorithm [23] to determine the complete noise waveform at the input of the victim receiver. This noise waveform can be viewed as the noise "signature" for a pairwise-coupled network. Since there are many such pairwise networks for a victim line, the total coupled noise is determined by the superposition of independent noise signatures within the cycle time. However, the resultant waveform is a statistical quantity, because its magnitude and shape depend on the switching times of the driver circuits on the active lines, which are statistical variables.

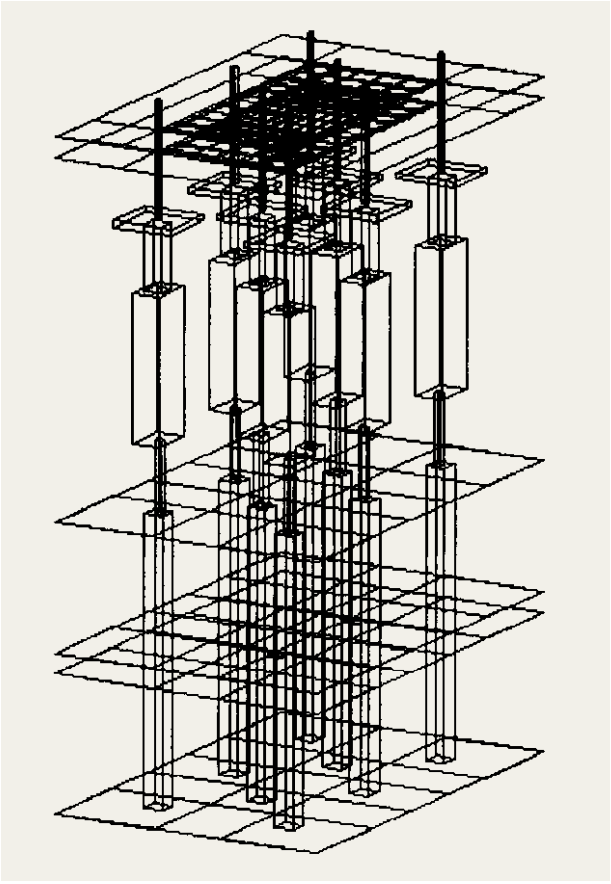

Figure 21

Connector abstraction for Harcon electrical modeling.

The reflections from one coupling interaction can either add to or cancel out the noise at the input of a receiver circuit, depending on the specific net topology. In Figure 18 a significant increase in the number of violators is observed, indicating that noise addition occurs when reflections are considered. Since the magnitude of the coupled noise generated on a net is proportional to its length, the results shown in this figure reflect the fact that the population of nets that exceed the defined noise limit is associated with the memory bus connection. For these interconnections, the reflection of the logic signal at the receiver input of an active net generates significant near-end noise from the various connectors within the path that propagates back to the input of the victim receiver and adds to the far-end noise generated by the initial logic signal on the active lines. In **Figure 22**, only on-module

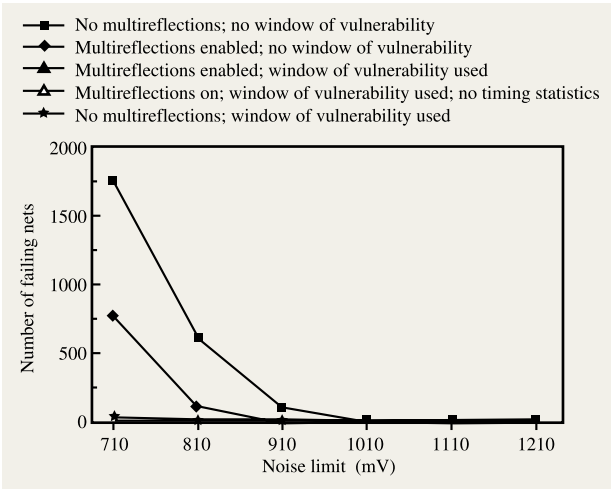

Figure 22

On-MCM net-only noise magnitude distribution tail as a function of receiver noise tolerance.

nets are considered. As can be seen in this figure, not only is the noise magnitude significantly lower because of shorter wiring lengths, but including reflections in the noise estimation reduces the resultant noise magnitude at the input of the victim receiver. This is due to cancellation of the noise caused by the reflected and incident logic signals on the active line, but this cancellation is possible only because of the relatively short lengths of these nets.

Because of the topology of the nets and the inclusion of multireflections in the noise determination, some noise components will appear at the receiver circuit input after the end of the cycle time during which all of the active lines are assumed to have switched. To handle these noise components without affecting the turnaround time of the noise verification process, we monitor the arrival time of each noise component; for those which occur beyond one cycle, we readjust their arrival time at the input of the receiver circuit by subtracting out the appropriate multiple of cycle times. This approach causes the overlapping of noise components which actually occur in different system cycles by wrapping them back into one reference system cycle.

Under the worst-case assumption that the active lines are switching in every cycle (e.g., alternating their state from cycle to cycle and creating a pulse train), their switching activity in odd multiples of the system cycle will create noise pulses with a polarity opposite to that of noise pulses generated in even multiples of the system cycle. Therefore, when a noise pulse arrives at the input of a victim line's receiver circuit after the system cycle being considered but before the end of the subsequent system cycle, its arrival time will be adjusted by subtracting one cycle, and its magnitude will be adjusted by reversing the polarity or multiplying its amplitude by -1. This ensures that noise pulses created from a driver circuit switching in one system cycle will properly add to or subtract from the noise pulses created by this driver switching in subsequent system cycles.

- *Delta-I calculator*

The delta-I calculator computes the high-frequency delta-I noise at each chip I/O on the MCM. A noise waveform is used as an input to the calculator to represent the delta-I noise created by the switching drivers, receivers, and on-chip macros. The on-chip macros may be logic blocks or arrays. The output is a sampled time waveform over the cycle for all chip I/Os which is later combined with the coupling noise to compute the total noise on the net.

The noise waveforms are generated in AS/X by placing a group of circuits on a section of the MCM high-frequency model discussed in Section 4. The noise waveform for each *type* of active circuit is created, and the amount of noise that propagates through each *type* of output circuit onto the off-chip interconnect is determined. The noise at an output due to a source somewhere on the chip is determined by

$$V_n(t, x, y) = A(t) \exp\left[-\sqrt{[\alpha(x' - x)]^2 + [\beta(y' - y)]^2}\right], \quad (6)$$

where $V_n(t, x, y)$ is the noise at a driver or receiver on the chip, $A(t)$ is the noise waveform that propagates onto the interconnect, α is the coefficient of attenuation in the x-direction of the chip, β is the attenuation in the y-direction, and (x', y') is the location of the noise source. The noise at the output of a quiet driver or receiver is simulated in AS/X for the case in which the quiet circuit is next to the switching circuits; that sampled waveform becomes $A(t)$ in the calculator. The noise amplitude attenuates as it propagates across the chip. The amplitude and the attenuation are dependent on the on-chip power grid and the amount of decoupling capacitance. The attenuation is determined from the AS/X results and fit in Equation (6) to determine α and β. By using Equation (6), the noise can be approximated at any point on the chip by summing all of the noise sources on the chip at the location of the driver and receiver. **Table 11** gives a sample of coefficients for the CP and nest chips on the G5 machine.

For active-source-terminated drivers, there are two delta-I noise pulses, the first when the driver switches and the second after a time has elapsed that is equal to twice the propagation delay to the end of the interconnection length. This second delta-I component is generated by the reflection of the signal at the receiver when it arrives at the output of the driver circuit. The delta-I calculator

determines the delay between the two pulses from the timing data discussed in Section 6 and includes it in the final noise waveform determination.

In addition, we found it convenient to include within the delta-I calculator the crosstalk of the on-chip wires from the C4 to the driver or receiver. On-chip extraction tools can calculate the coupling capacitance for the on-chip portion of the off-chip net. The calculator computes the coupling noise for this situation using mV/pF coefficients and adds this noise to the final noise waveform for each chip I/O on the MCM.

- *Statistical noise summation*

Driver switching-time variations have traditionally been determined by short- and long-path static timing analysis [16]. This is not the case for the G-series class of IBM mainframes, since all off-chip nets are latch-to-latch. Although this represents a simplification of the timing process, driver switching-time windows still exist because of manufacturing and clocking considerations. These driver switching-time windows, known as the early and late actual times (EAT and LAT), are computed for all off-chip driver pins on the MCM as described in Section 6 of this paper. This variation is treated as a Gaussian probability density function for each active driver and convolved [16] with the appropriate noise signature computed by the multireflection algorithm. This convolution results in a value for the magnitude of the noise voltage at each discretized point within the system cycle and its corresponding probability of occurrence. This calculation is performed on all active nets associated with a victim net. A statistical summation of all noise waveforms yields the statistical discrete distribution of the noise magnitude at any discretized point within the system cycle for all nets.

The smearing effect of the convolution tends to increase the overall noise levels. This is verified by the results shown in Figure 18. From this figure it is apparent that when the switching-time uncertainty of the active lines is not used in the noise estimation, some nets that violate the limit of the probability of failure will not be detected. In fact, for a given set of deterministic switching times of the active driver circuits, the number of noise violators can be 15% to 31% lower than the number obtained when the statistics of the switching-time variation of these drivers are taken into account.

Because of the nature of latch operation, there is a small time window within the system cycle time in which noise can corrupt data. The noise magnitude outside this time window is irrelevant because it will not propagate through the latch. This time window is based on the collective influence of clock arrival times, data path timings, manufacturing tolerances, etc., and is called the "window of vulnerability." The window of vulnerability is computed on a net-by-net basis and is essential for determining the real noise violators in the system by accounting for noise only within the required data arrival time period. The effect of using the window of vulnerability is shown in Figure 18. The number of nets exceeding the noise limit of 1000 mV is reduced by seven to ten times when the window of vulnerability is considered. The incorporation of this feature in the noise calculator not only yields more accurate results for the nets needing rerouting, but also has reduced the time spent in route analysis by one order of magnitude.

Table 11 Sampling of DELI coefficients used for the design of GEMI.

	Peak of A(t) off-chip driver (mV/driver)	α (mm^{-1})	β (mm^{-1})
CP chip (G5)	0.20	0.40	0.34
Nest chips	0.30	0.40	0.34
CP chip (G6)	0.18	0.58	0.58

The process of identifying nets which have a high probability of excessive noise and rerouting them is an iterative one and continues until all nets in the system are within the failing criterion. This iterative process is performed several times in conjunction with timing analysis during the electrical verification phase of the MCM design. Therefore, it is very important that all problem nets be identified quickly and accurately.

8. Conclusions

The previous discussion shows that a very complex MCM can be designed in a time comparable to that required for the design of the processor chip and with less than half the resources. Although the design regime with respect to tools and approach is rigid, it has resulted in good product with a one-pass design. Two-phase physical design (i.e., pre-PD circuit simulation and post-PD verification) has been shown to be essential in achieving this goal. Using the experimental data from the GEMI MCM, we have proved that closed-form equations for the estimation of both the timing and the noise of the wired nets provide adequate accuracy and superb execution performance.

However, the manufacturing time of such MCM packaging is extremely long, nearly three times the corresponding processor chip manufacturing turnaround time. A significant cause of this long turnaround time is the growth of thin films on the glass-ceramic material in a serial manner. Modification of the manufacturing process so that this serial approach can be changed is very desirable.

Since the MCM manufacturing turnaround time is so much longer than that of the processor chip, a great

degree of integration between the package and chip design teams is required, and this has been the case for the S/390 server design teams. This team structure partly ameliorates the constraint imposed on chip design by the requirement for a very early definition and spatial allocation of the corresponding chip I/Os.

Even with the current MCM turnaround time, the GEMI MCM provides the lowest cost–performance solution for a symmetric multiprocessor system (SMP) containing 12 to 14 processors with the associated L2 cache and system I/O chips. It is significantly cheaper than an equivalent SCM implementation of such a system on complicated card-and-board technologies, but it requires the design rigor and technical expertise in controlling noise and interconnect timing that have been described in this paper. In addition, the MCM technology provides us with the most dense configuration of chips in a two-dimensional arrangement, which facilitates refrigerated cooling. In fact, the GEMI MCM is the key contributor to a low-cost refrigeration apparatus that can enhance the performance of the system by more than 10%. In this refrigerated operation we have shown that the GEMI MCM can support 300-MHz synchronous interconnections of significant bandwidth. However, as was pointed out previously in this paper, this frequency limitation is affected by the delay of the electronic circuits associated with the interconnections. As CMOS technology advances, glass-ceramic MCMs with thin film should be able to support even higher interconnect frequencies of operation. Indeed, we believe that glass-ceramic MCMs with thin film can easily support 500-MHz off-chip interconnects with future CMOS technologies, provided that the maximum net length is 90 to 100 mm. This naturally assumes the availability of the highly sophisticated manufacturing capability which was brought to bear on the construction of the GEMI MCM, and which can deliver more than half a kilometer of wire and chill 1000 watts in an area equal to the size of the human palm.

Acknowledgments

The authors thank V. Minnick, J. Quick, and H. Pross for their input of the technical data for the board that houses the GEMI.

*Trademark or registered trademark of International Business Machines Corporation.

**Trademark or registered trademark of Cadence Design Systems, Inc., or Synopsis, Inc.

References

1. E. A. Reese, D. Nedwek, J. Jex, M. Khaira, T. Burton, P. Nag, H. Kumar, C. Dike, D. Finan, and M. Haycock, "A Phase-Tolerant 3.8 GB/s Data Communication Router for a Supercomputer Backplane," *ISSCC Digest of Technical Papers*, February 1994, pp. 296–297.
2. G. A. Katopis and W. D. Becker, "S/390 Cost Performance Considerations for MCM Packaging Choices," *IEEE Trans. Components, Packaging, & Manuf. Technol. B: Adv. Packaging* **21,** No. 3, 286–297 (August 1998).
3. W. D. Becker, H. Smith, T. McNamara, P. Muench, J. Eckhardt, M. McAllister, G. Katopis, S. Richter, R. Frech, and E. Klink, "Modeling, Simulation, and Measurement of Mid-Frequency Noise in Computer Systems," *IEEE Trans. Components, Packaging, & Manuf. Technol. B: Adv. Packaging* **21,** No. 2, 157–163 (May 1998).
4. G. A. Katopis, W. D. Becker, H. H. Smith, and H. Stoller, "MCM C/D Design for the CMOS Implementation of the S/390 System," *Proceedings of the 47th Electronic Components and Technology Conference, IEEE Cat. No. 97CH36048*, May 1997, pp. 479–485.
5. E. E. Davidson, P. W. Harding, G. A. Katopis, M. O. Nealon, and L. L. Wu, "Physical and Electrical Design Features of the IBM Enterprise System/9000 Circuit Module," *IBM J. Res. Develop.* **36,** No. 5, 277–288 (September 1992).
6. E. D. Perfecto, A. P. Giri, R. R. Shields, H. P. Longworth, J. R. Pennacchia, and M. P. Jeanneret, "Thin-Film Multichip Module Packages for High-End IBM Servers," *IBM J. Res. Develop.* **42,** No. 5, 597–606 (September 1998).
7. H. Stoller, S. Ray, E. Perfecto, and T. Wassick, "Evolution of Engineering Change (EC) and Repair Technology in High Performance Multi-Chip Modules at IBM," *Proceedings of the 48th Electronic Components and Technology Conference*, Seattle, WA, May 1998, pp. 916–921.
8. R. R. Livolsi, "Variable Voltage Driver," U.S. patent pending, Docket No. PO998062.
9. J. P. Eckhardt and K. A. Jenkins, "PLL Phase Error and Power Supply Noise," *Proceedings of the 7th Topical Meeting on Electrical Performance of Electronic Packaging*, West Point, NY, October 1998, pp. 73–76.
10. R. M. Averill III, K. G. Barkley, M. A. Bowen, P. J. Camporese, A. H. Dansky, R. F. Hatch, D. E. Hoffman, M. D. Mayo, S. A. McCabe, T. G. McNamara, T. J. McPherson, G. A. Northrop, L. Sigal, H. Smith, D. A. Webber, and P. M. Williams, "Chip Integration Methodology for the IBM S/390 G5 and G6 Custom Microprocessors," *IBM J. Res. Develop.* **43,** 681–706 (1999, this issue).
11. B. Singh, W. Becker, and M. McAllister, "Core Logic Simultaneous Switching Noise Measurements on a 500 MHz CMOS Chip on a CBGA SCM," *Proceedings of the 48th Electronic Components and Technology Conference*, Seattle, WA, May 1998, pp. 605–609.
12. B. D. McCredie and W. D. Becker, "Modeling, Measurement, and Simulation of Simultaneous Switching Noise," *IEEE Trans. Components, Packaging, & Manuf. Technol. B: Adv. Packaging* **19,** No. 3, 461–472 (August 1996).
13. David P. Lapotin, Toufie R. Mazzawy, and Marlin L. White, "Early Package Analysis: Considerations and Case Study," *IEEE Computer* **26,** No. 4, 30–39 (April 1993).
14. Rao R. Tummala, Eugene J. Rymaszewski, and Alan G. Klopfestein, *Microelectronics Packaging Handbook*, Second Edition, Part I, Chapter 3, Chapman & Hall, New York, 1997.
15. P. N. Venkatachalam, H. H. Smith, and R. Rippens, "Noise Containment in a High Wire Density Multichip Module," *Proceedings of the 2nd Topical Meeting on Electrical Performance of Electronic Packages*, October 20–22, 1993, pp. 58–60.
16. D. Rude, "Statistical Method of Noise Containment in a Synchronous System," *IEEE Trans. Components, Packaging, & Manuf. Technol.* **17,** 514–519 (November 1994).

17. A. Feller, H. R. Kaupp, and J. J. DiGiacomo, "Cross-Talk and Reflections in High-Speed Digital Systems," *AFIPS Conf. Proc. Fall Joint Computer Conference* **27,** 511–525 (1965).
18. B. J. Rubin, "An Electromagnetic Approach for Modeling High Performance Computer Packages," *IBM J. Res. Develop.* **34,** No. 4, 585–600 (July 1990).
19. W. T. Weeks, "Calculations of Coefficients of Capacitance of Multi-Conductor Transmission Lines in the Presence of a Dielectric Interface," *IEEE Trans. Microwave Theory* **MTT-18,** 35–43 (1970).
20. J. F. Janak, D. D. Ling, and H. M. Huang, "C3DSTAR: A 3D Wiring Capacitance Calculator," *Proceedings of the IEEE International Conference on Computer-Aided Design*, November 1990, pp. 530–533.
21. G. Katopis, P. Lin, and P. N. Venkatachalam, "Coupled Noise Estimation for MCM Via Structures," *Proceedings of NEPCON West 1992*, February 1992, pp. 64–76.
22. H. H. Smith and C. T. Spring, "Coupled Noise Analysis of a Complex Connector Structure Using a Full-Wave Modeling Approach," *Proceedings of NEPCON West 1992*, February 1992, pp. 29–35.
23. H. H. Smith and G. A. Katopis, "Multireflection Algorithm for Timed Statistical Coupled Noise Checking," *IEEE Trans. Components, Packaging, & Manuf. Technol.* **19,** 503–511 (August 1996).

Received April 22, 1999; accepted for publication July 14, 1999

George A. Katopis *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (katopis@us.ibm.com).*
Mr. Katopis is a Senior Technical Staff Member in the IBM S/390 Division, responsible for the technology selection and packaging strategy of all CMOS S/390 servers. Mr. Katopis has authored more than fifty papers on the subjects of net design and switching-noise prediction and containment in the digital server engines. He holds three patents on switching-noise reduction and has co-authored chapters in three books on electrical design of electronic packages. Mr. Katopis received an M.S. degree and an M.Ph. degree from Columbia University. He is a Senior Member of the IEEE, and an industrial mentor to the electrical engineering departments of Cornell University and the University of Arizona at Tucson.

W. Dale Becker *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (wbecker@us.ibm.com).*
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 Engineer in the IBM Server Group, leading the MCM design team that integrates and implements the multiprocessor design for IBM S/390 platforms. 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.

Toufie R. Mazzawy *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (tmazzawy@us.ibm.com).*
Mr. Mazzawy is currently an Advisory Engineering Manager, managing the Package Design and Logic Verification teams working on the S/390 G-Series servers. Mr. Mazzawy joined IBM in 1989, and has held many positions related to MCM package design, including MCM Design Coordinator, High Level Design, Early Package Wirability Analysis, and Design Team Leader. He has received Outstanding Technical Achievement Awards for his contributions to both the G4 and G5 MCM package designs. Mr. Mazzawy received a B.E. degree in electrical engineering from the Stevens Institute of Technology in 1989 and an M.S. degree in electrical engineering from Syracuse University in 1994.

Howard H. Smith *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (smithh@us.ibm.com).*
Mr. Smith received his B.S. and M.S. degrees in electrical engineering from the New Jersey Institute of Technology, Newark, in 1984 and 1985, respectively. He joined IBM in 1984 as an integrated circuit engineer at the semiconductor development laboratory in East Fishkill, New York, working in the area of high-performance masterslice designs. Mr. Smith is currently an Advisory Engineer in the IBM System/390 Division in Poughkeepsie, New York, where he is responsible for electrical analysis issues associated with high-density CMOS circuit technology and package-related products. Recent assignments have included the development and coordination of on-chip noise verification processes for the S/390 processor designs. His expertise lies in the area of electrical noise modeling and prediction in system-level computer operation. Mr. Smith has co-authored several papers on system-level noise prediction, on-chip interconnects, and electromagnetic characterization of connectors and antennas.

Charles K. Vakirtzis *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (vakirtzi@us.ibm.com).* Mr. Vakirtzis received the B.S.E.E. degree from the Rochester Institute of Technology in 1975; in addition, he has accumulated more than 30 graduate credits in the areas of theoretical mathematics and engineering. Joining IBM in 1977, he worked as a Manufacturing Engineer supporting the final system test process until 1985, when he joined the Department of Electrical Modeling of Packaging, focusing on interconnect performance and modeling. He is currently an Advisory Engineer working on the electrical modeling of noise and its containment for on-chip interconnects. Mr. Vakirtzis has co-authored a number of papers and holds two patents on noise-estimation techniques and an on-chip temperature-sensing system.

Scott A. Kuppinger *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (skupping@us.ibm.com).* Mr. Kuppinger is a Staff Engineer in the S/390 Custom Packaging Department at the IBM Poughkeepsie Development Laboratory. He joined IBM in 1989 after receiving a B.S. degree in electrical engineering at Clarkson University.

Bhupindra Singh *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (singhb@us.ibm.com).* Mr. Singh is an Advisory Engineer at the IBM Poughkeepsie Development Laboratory. He received his B.S.E.E. degree from Punjab University, Chandigarh, India, in 1966, and an M.S.E.E. degree from Syracuse University in 1978. He is currently working in advanced VLSI & Packaging Applications. Mr. Singh is involved in design, modeling, and noise simulation of high-performance multichip and single-chip packages.

Phillip C. Lin *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (linp@us.ibm.com).* Mr. Lin is a Staff Engineer working in high-performance packaging development and on-chip power grid design. He received an M.S. degree in electrical engineering from Columbia University. Mr. Lin is the author of one technical disclosure and one U.S. patent.

John Bartells, Jr. *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (barteljr@us.ibm.com).*

Gregory V. Kihlmire *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (gregki@us.ibm.com).* Mr. Kihlmire is an Advisory Software Engineer. He received an A.A.S. degree in electronics from Dutchess Community College in 1975. Since joining IBM in 1977, he has worked in various positions dealing with second-level packaging, including managerial responsibility for the Module/Board Packaging and Release Department for the 3090 mainframes. Mr. Kihlmire is currently the MCM Coordinator for the Alliance packaging team.

Panangattur N. Venkatachalam *IBM System/390 Division, 522 South Road, Poughkeepsie, New York 12601 (chalam@us.ibm.com).* Mr. Venkatachalam is an Advisory Engineer at the IBM Poughkeepsie Development Laboratory. He joined IBM in 1978 at East Fishkill, New York, working in MLC MCM electrical design. In 1992 he transferred to the System/390 Division and worked on signal integrity in modules, boards, and cards. Mr. Venkatachalam received an Outstanding Technical Achievement Award for S/390 G5 Package Development in 1998. He received an M.S.E.E. degree from the University of California at Berkeley.

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 Engineer with the IBM Microelectronics Division, responsible for the application engineering for all high-performance MCMs. He has co-authored numerous articles and papers on the application of MCMs and MCM technology. Mr. Stoller holds eight patents and has reached the Fourth Invention Plateau. He holds a B.S. degree from CCNY and an M.S. degree from Rutgers University, both in physics.

Jason L. Frankel *IBM Microelectronics Division, East Fishkill facility, Route 52, Hopewell Junction, New York 12533 (frankel@us.ibm.com).* Mr. Frankel graduated from Clarkson University in 1989 with a B.S. degree in electrical and computer engineering. From 1989 until the present, he has been a Package Design Engineer in the IBM Microelectronics Division. Since 1996, Mr. Frankel has designed the signal redistribution and power distribution structures of thin-film, glass-ceramic, and alumina multichip modules for S/390 servers.