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

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

This paper describes the system packaging of the processor cage for the IBM eServer z900. This server contains the world's most complex multichip module (MCM), with a wiring length of 1 km and a maximum power of 1300 W on a glass-ceramic substrate. The z900 MCM contains 35 chips comprising the heart of the central electronic complex (CEC) of this server. This MCM was implemented using two different glass-ceramic technologies: one an MCM-D technology (using thin film and glass-ceramic) and the other a pure MCM-C technology (using glass-ceramic) with more aggressive wiring ground rules. In this paper we compare these two technologies and describe their impact on the MCM electrical design. Similarly, two different board technologies for the housing of the CEC are discussed, and the impact of their electrical properties on the system design is described. The high-frequency requirements of this design due to operating frequencies of 916 MHz for on-chip and 459 MHz for off-chip interconnects make a comprehensive design methodology and post-routing electrical verification necessary. The design methodology, including the wiring strategy needed for its success, is described in detail in the paper.

1. Introduction

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

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

0018-8646/02/$5.00 © 2002 IBM

Table 1 Development of the zServer from 1998 to 2002.

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

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

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

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

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

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

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

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

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

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

2. Logical system structure and chip technology

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

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

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

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

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