High-speed interconnect and packaging design of the IBM System z9 processor cage

H. Harrer
D. M. Dreps
T.-M. Winkel
W. Scholz
B. G. Truong
A. Huber
T. Zhou
K. L. Christian
G. F. Goth

This paper describes the system packaging and technologies of the IBM System z9™ enterprise-class server. The central electronic complex of the system consists of four nodes, each housing a multichip module (MCM) with 16 chips consuming up to 1,200 W. The z9™ server doubles the multiprocessor performance of the System z990 by increasing the central processing unit (CPU) configuration and using an internally developed elastic interface to increase interconnect speed on all high-speed buses. In contrast to all previous zSeries® designs, which were running at half of the processor speed, the packaging interconnects on the multichip module run at the same speed as the processor (1.72 GHz). High frequencies and massively parallel connectivity lead to a raw packaging bandwidth of up to 1,764 GB/s between processors and cache within a single frame for a fully configured four-node z9 system.

1. Introduction

The IBM System z9* server follows the IBM System z* family, as described in [1–3]. It was developed using System z990 packaging technology [2], but now includes 90-nm chip technology, which results in higher signal frequency and leakage current growth. It reuses the modular concept, in which up to four processor nodes can be plugged into a single computing machine. This results in a maximum configuration of 64 processors per frame. The complete system consists of two racks housing the processor nodes, three I/O cages, a modular refrigeration cooling unit, and the bulk power supplies. In contrast to the z990 server, a flexible I/O and memory configuration supports the on demand business concept, in which up to eight memory cards and up to eight I/O cards can be plugged into a single node. New reliability features such as concurrent node upgrade allow the customer to add processor nodes without a system reboot, and dynamic oscillator card switching improves reliability and serviceability (RAS). The details of the system structure are described in Section 2.

Section 3 describes the multichip module (MCM) technology and design. The 375-μm glass-ceramic pitch of the previous system generation [2, 3] was reduced to 350 μm to contain the routing within 102 layers, as in the z990 system. The C4 solder bump pitch of the chips was adapted to match the via pitch of the module. The high chip power dissipation made it necessary to develop a new sort strategy. The processor chips were sorted into "low leakage current" chips and "high leakage current" chips so that the eight processor chips on the MCM would comprise a combination of high-leakage and low-leakage chips and would limit worst-case power consumption.

Section 4 compares the technology and design of the cards and boards. Despite running at higher frequencies and using net lengths for the point-to-point nets of the ring (which connects the MCM between the processor nodes) that are similar to those of the System z990, the System z9 does not use low-loss dielectric material in its circuit cards. In a two- or three-node configuration, those circuit card nets reach a line length of more than 80 cm and join to four very high-density metric (VHDM**) connectors. This was achieved by the design features of the Elastic Interface 2 (EI-2). These features include V_{ref} forwarding, where the receiver threshold is generated directly from the transmitted clock signal, drivers with de-emphasis, and restricted placement of driving and receiving circuits to minimize loss of the on-chip wiring [4].

0018-8646/07/$5.00 © 2007 IBM

Furthermore, although all high-speed signals on the z990 were bidirectional, they are unidirectional on the System z9. On the z990 the bidirectional nets had to be quiesced by sending zero data before switching the direction. Since the net topology and the net lengths did not change, the number of zero cycles would have had to be increased with the reduction of the cycle time, and this was not acceptable for the overall system performance. Because the packaging technology did not support a doubling of the I/O count, an increase in frequency was the only solution. Thus, all buses on the MCM run at the same speed as the processor, which between the processor and cache chips results in a total raw packaging bandwidth of 441 GB/s for each node. This can be achieved by improvements in the source-synchronous I/O circuits (EI-2) on the chips, as shown in Section 5.

Another major challenge was chip power delivery, described in Section 6. Leakage currents in 90-nm technology resulted in a significant power increase, especially in the air-cooled backup mode at high temperatures. New dc drop analysis methodologies were required for robust power delivery (up to 1,200 W) to the processor, cache, and bus adapter chips on the multichip module.

Cooling is handled by a modular refrigeration unit (MRU) that cools the central electronic complex (CEC) chips to 45°C. This low operating temperature enables high reliability and reduced leakage power. An air-cooled backup mode at lower chip frequencies ensures system operation in case of an MRU failure. The cooling of the MCM did require improvements in the technology of the z990 [5]; i.e., the thermal resistance between the processor chips and the hat was reduced. This was achieved by using a small-gap technology (SGT) [6], in which the MCM hat has special cooling "pistons" for each CPU, allowing reduction of the gap between chip and cooling hat, as shown in Section 7.

2. Logical system structure

The node-based server design of the System z9 accommodates up to 32 processor chips, or 64 processor cores, per system. The system memory size can be increased up to a maximum of 512 GB, and the system I/O connectivity has been enhanced to a maximum of 64 self-timed interface (STI) I/O paths, each with a capability of 2.27 Gb/s. The increase in number of processors, memory size, and connectivity allows us to achieve the desired symmetrical multiprocessor (SMP) performance, but it generates a significant increase in the total number of interconnections between chips. This results in a higher complexity for the first- and second-level package, namely the MCM and the printed wiring cards and boards.

The z9* CEC packaging top view is shown in **Figure 1(a)**, and the processor node packaging is shown in **Figure 1(b)**. The 64-processor system is divided into four processor cards (nodes). Each processor card contains 16 dual-core processors on an MCM, up to 128 GB of main memory, and 16 STI interfaces. The processor chip is connected to each cache chip by a 16-byte bus operated at 1.72 Gb/s. This results in a bandwidth of 441 GB/s on a single node.

Up to four nodes are plugged into the center board using connectors each with 1,160 signal pins. The connector holds the ring of the processor nodes and feeds the power into the processor cards. The high-speed interconnections of the EI connect the cache chips on the MCM with the cache chips on the other processor cards to ensure fast access of all cache data to each CPU. The bus speed is 0.86 Gb/s for a single line, with a total bandwidth of 124 GB/s. A jumper card closes the ring in a two- or three-node configuration.

The MCM also provides the connection to the memory via the memory storage controllers (MSCs). A maximum of eight memory cards can be plugged per node; each memory card contains four dual inline memory modules (DIMMs) and two storage memory interface (SMI) control chips. The interfaces are operated at a fixed gear ratio of 2:1 with respect to the processor cycle time.

The eight I/O slots per node are provided by the fan-out cage adapter card. They connect a high-speed GX+ bus (running at half the processor frequency) from each slot to the two MSC chips on the MCM.

The I/O card holds the memory bus adapter (MBA) module, which converts the GX+ bus into two 2.27-GB/s fast unidirectional STI links. The STI I/O connection is used to connect to zSeries* I/O cages or other zSeries systems. The two STI cable channels per card result in 16 STI links per processor card and up to 64 STI I/O paths per system.

The processor node is controlled by a system control card in the fan-out cage. This card provides the control structure for the processor and I/O cards. It controls the configuration and power-on sequence, as well as the error handling of the node. There are redundant communication paths via the serial interface from the flexible service processor (FSP) cards in the distributed converter assemblies (DCAs).

Two high-availability redundant DCAs provide redundant power supplies for each processor card. One FSP card, hosted in the DCA, controls one processor card, and the second FSP, hosted in the second DCA, is for redundancy. A maximum of eight DCA cards with eight FSP daughter cards are used in each central electronic complex (CEC) cage.

Two redundant oscillator (OSC) cards are plugged into the z9 center board. One card is always active, while

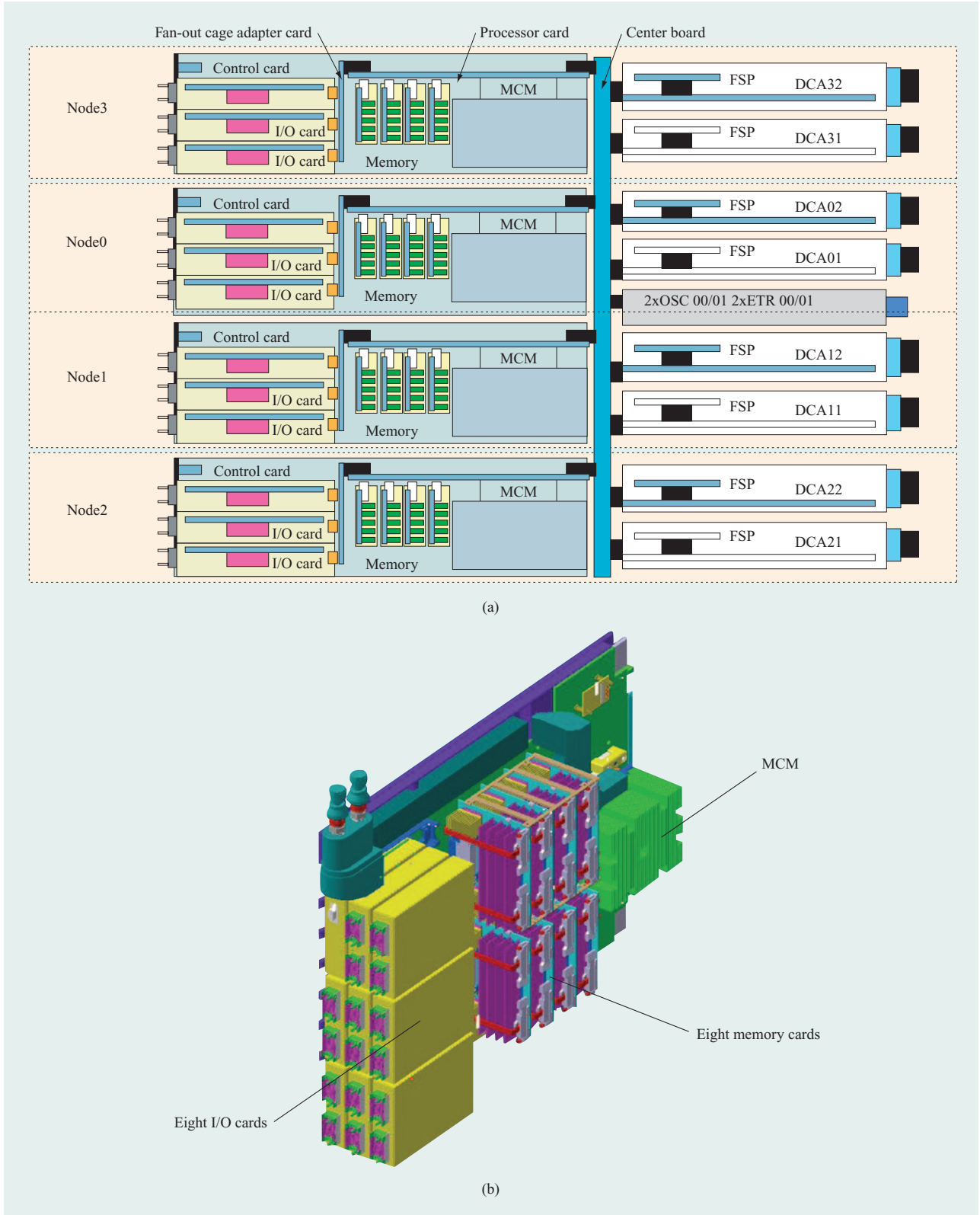

Figure 1

(a) CEC packaging (top view); (b) processor node packaging.

Table 1 Attributes of chips in the z9 CEC system. The maximum chip power is given for the normal operation of the modular refrigeration unit and the air-cooled backup mode

	CP	*SD*	*SC*	*MSC*	*CLK*	*MBA*	*SMI*	*FGA-T*
Chip size (mm × mm)	15.8 × 11.8	15.2 × 15.6	16.3 × 16.3	14.2 × 14.2	9.5 × 9.4	9.3 × 9.4	7.5 × 7.5	6.1 × 6.1
Technology (nm)	90	90	90	90	130	130	130	180
No. of C4s	2,962	3,763	4,305	3,261	1,387	624	854	413
No. of signals	603	1,600	1,768	1,344	487	283	415	183
No. of transistors (millions)	121	660	162	24	6	17	—	0.3
V_{dd}	1.2	1.2	1.2	1.2	1.5 core 1.2 I/O	1.35 core 1.2 I/O	1.5 core 1.2 I/O	1.8 core 3.3 I/O
Frequency (GHz)	1.72	0.86	0.86	0.86	0.25 0.26 ETR	0.31 1.25 STI	0.533	0.07
Packaging	GC MCM	GC MCM	GC MCM	GC MCM	GC MCM	Alumina SCM	Alumina SCM	Alumina SCM
Max. power (W)	77 101	40 62	30 51	20 30	5 6	20 20	11 11	0.5 1.0

the other is a redundant backup. A newly introduced dynamic oscillator switch function provides enhanced redundancy. The dynamic oscillator switch modules on the two OSC cards communicate with each other, switching the clock generation to the other card at a failure of the master for enhanced redundancy. Each card provides the clock signal generator for various master clocks at the clock chip in the CEC. All four processor cards are controlled by the same OSC card.

Two redundant external time reference (ETR) cards are implemented to allow the z9 to be coupled to another z9 system. One card is always active, while the other one is a redundant backup. Each card provides the ETR optical receiver function for the clock chip on the processor MCM as well as the receivers/drivers for fiber cables.

3. First-level packaging

The attributes of the chipset for the central electronic complex are summarized in **Table 1**. All chips on the multichip module besides the CLK are using a 90-nm chip technology and are therefore high contributors to leakage current. The system control chip has the largest chip size (266 mm^2). It also has the highest signal count (1,768 signals), which is caused by the control signals to the CPU chips and the ring connecting the nodes. The memory storage controller (MSC) has a significantly higher I/O count than the System z990, as the I/O buses (GX+ bus) emanate from here. This balances the overall I/O structure on the chip C4 count, since the SD and SC chips (which hosted this bus on the z990 system) cannot be further increased. The largest chip density is reached by the cache chip, with 660 million transistors, which results in a 10-MB cache for each SD chip. While all processor chips and nest chips on the MCM are operated at a power supply of 1.2 V, the clock (CLK) chip is operated at the 1.5 V required by the 120-nm technology. The I/O domain has been set to 1.2 V for all chips including MBA and SMI, so that all major buses are operated at the same voltage level. The MBA, SMI, and service element (FGA-T) are application-specific integrated circuit (ASIC) chips and are packaged on an alumina ceramic ball grid array (CBGA). Their power dissipations and frequency operation are moderate and do not require special cooling solutions. A direct heat-sink attachment was adequate. The ceramic technology was chosen in order to achieve the highest reliability and support a high-temperature, high-voltage burn-in process to identify ac defects. The nest chips (SC, SD, and MSC) are operated at half of the processor cycle time. Because the data between processor, cache, and memory storage controller is transferred at processor speed, a 1:2 bus-speed conversion was required on the nest chips.

Figure 2(a) shows the footprint of the chips, which was common to all chips on the MCM. The C4 bumps are arranged on a 350-μm interstitial grid. This enables a 350-μm rectangular signal grid with two adjacent power C4s and two adjacent ground C4s for shielding, close return current, and good power delivery. Using 100-μm C4 pads, the closest spacing between adjacent C4s is 247 μm.

Figure 2(b) shows the top-surface layout of the z9 glass-ceramic MCM, which looks similar to the z990. Eight dual-core processor chips are arranged around four cache chips to minimize the lengths for the large number of CP-to-SD interconnections. The memory storage controllers are placed at the corners and allow a good signal fan-out from the bottom of the MCM to the I/O and memory

cards. There are 217 decoupling capacitors which have very low parasitic inductances ($L < 30$ pH). Each chip has a filter capacitor for the phase-locked loop (PLL). The analog PLL voltage is delivered to the chips in a separate plane and additionally decoupled by capacitors. The MCM had to use a tighter wiring pitch than System z990 to accommodate the 12.3% increase in the number of nets from 8,252 to 9,271. The total wiring length including vias was 545 m. This was a new record for MCM wiring density (60 mm wiring per mm^2 of MCM area). The packaging attributes are summarized in **Table 2.** In the redistribution area, the pitch was 175 μm, defining the tightest ground-rule dimension for the MCM. The 102 layers consist of nine redistribution layers for fanning out the signals under the chip sites and 23 plane pairs with an orthogonal routing for completing the connections. As on previous systems, all redistribution layers and routing plane pairs are referenced by a V_{dd} and GND mesh plane in order to achieve a constant impedance of 55 Ω nominal. For System z9, the screen sheet thickness was reduced from 7 mil to 6 mil, which allowed a via diameter of 84 μm; this resulted in a better match to the nominal line impedance (52 Ω vs. 43 Ω). The land grid array connector at the bottom side of the MCM has been reused from the z990 system, providing 5,184 I/Os on four quadrants at a 1-mm pitch.

The processor chip reaches a maximum power of 77 W for worst-case leakage during standard operation and up to 101 W during air-cooled backup mode. The total MCM cooling limitation of 850 W in normal operation mode and 1,200 W in the air-cooled backup mode made it necessary to sort the chips by leakage current and limit the number of "high leakage current" chips and "low leakage current" chips on the MCM. In Figure 2(b) only four of the eight processor chips are allowed to have high leakage current because of the temperature dependency of the air-cooled backup mode, in which the chip junction temperature can rise to up to 100°C.

There are two alumina ceramic carriers designed for test and burn-in. Before the chips are assembled on the MCM, they are tested on a wafer probe through a small number of test pins. The next step is to mount the chip on a temporary chip attachment (TCA) carrier on a test board. In addition to the chip self-test and basic I/O functionality, the burn-in is done on the TCA carrier, where the chip is exposed to the highest power of up to 300 W at high temperature and high voltage to identify ac defects. The chip is then tested again before it is finally assembled on the MCM. This procedure ensures minimum chip rework on the MCM at the system test level. Because of the high I/O count, a dotting of the signal I/Os is done on the TCA carrier, and those signals are multiplexed on the tester.

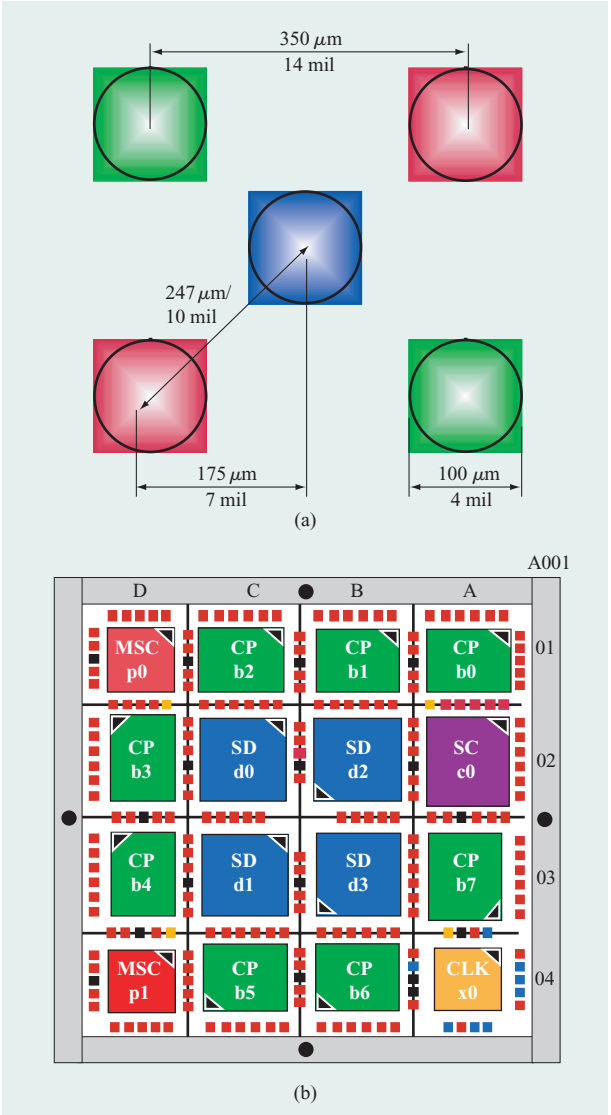

Figure 2

(a) Interstitial 350-μm footprint. A signal (blue) is surrounded by two power (red) and two ground (green) C4s. (b) System z9 multichip module consisting of eight dual-core CPUs (CP), four 10-MB cache chips (SD), one system control chip (SCC), two memory storage controllers (MSC), and the clock chip (CLK). There are 217 decoupling capacitors (200-nF LICA).

The other application-specific chips use custom-designed single-chip modules. They are alumina ceramic with a standard layer count and 1.27-mm CBGA connectors between module and card.

4. Second-level package

The special challenge for the design of the second-level package was to incorporate packaging solutions

Table 2 Key attributes of the System z9 first-level packaging designs.

	MCM	*TCA for CP, CLK*	*TCA for SC, SD, MSC*	*MBA (shared)*	*SMI*	*FGA-T*
Size (mm × mm)	95 × 95	50 × 50	50 × 50	32.5 × 32.5	25 × 25	21 × 21
Material	Glass ceramic	Alumina ceramic	Alumina ceramic	Alumina ceramic	Alumina ceramic	Alumina ceramic
No. of layers	102	39	39	21	10	7
Thickness (mil)	6	7	7	7	6	7
Ground rules (μm)	350	350	350	450	450	450
Connector	5,184 LGA 1-mm pitch	2,114 LGA 1-mm pitch	2,114 LGA 1-mm pitch	1.27-mm CBGA	1.00-mm CBGA	1.27-mm CBGA
Total no. of I/Os	5,184	2,114	2,114	624	575	255
No. of signal I/Os	2,970	1,210	1,210	270	413	183
Max. power (W)	850 (MRU) 1,200 (air)	300 (burn-in)	300 (burn-in)	20	11	1

for all system requirements and needs. For the system performance, for example, it was mandatory to maintain a speed of half of the processor cycle for the major buses. Another requirement was to deliver multiple voltages, some with very high currents, to different cards in the system. This was a special challenge for the processor unit (PU) card containing a large number of signal and power layers. For this card, the maximum tolerable card thicknesses had to be maintained in order to be producible at reasonable component costs. As a consequence of the different system requirements, extensive signal and power integrity analysis were performed. On the basis of the signal integrity analysis, we were able to use low-cost standard-loss material for the most critical components, namely the PU card and the center board (Figure 1). In the previous System z990, in contrast, a low-loss, more expensive material was used for these components. This material change was achieved by using the EI-2 and thus lowering the electrical requirements. In the second-level package, the signal speed was increased to 0.86 Gb/s for single-ended lines.

The main component of the second-level package is the PU card carrying the MCM, the eight memory cards, the I/O adapter card, and the huge connectors for the ring bus. The PU card includes ten signal layers, 21 voltage layers, and two mounting planes. In order to stay within ten signal layers for routing, buried-via technology was used. Buried vias interconnect signal lines between adjacent signal layers separated by just one voltage or ground layer. The advantage of this type of via is that it stays within this three-layer core and does not block wiring channels in other signal layers. As a result, the PU card consists of an odd number of layers.

To meet the customer's needs for flexible scaling system components such as memory cards or I/O cards, the granularity in the present z9 system was increased in comparison with the previous z990 system. Instead of just two huge memory cards, a total of eight cards can now be plugged into the PU card. A total of eight I/O cards can also be plugged into the I/O adapter card. Improved connectivity of the system was achieved by increasing the signal speed of the self-timed interface (STI) by 36%, from 1.67 Gb/s on the z990 system to 2.27 Gb/s on the z9. It was the only bus using a differential interface, because the interconnects to other systems require a maximum cable length of 10 m.

To reduce the wiring complexity in the different cards and thus the number of signal layers, a special wiring study and predefinition of connector pins was performed using a new IBM internally developed program. With this program, the number of wiring crossings was reduced by choosing an optimized connector pin assignment. The primary advantage of this program is the reduction of development time, especially in physical design and card layer reduction and thus card costs.

5. I/O interface

The I/O interface used on the z9 server consisted of the EI-2 signaling and data alignment described in [7]. A major improvement in signaling over that of previous System z990 servers includes data drivers capable of predistortion and data receivers with aggressive clamping to decrease overshoots and undershoots.

Signaling

The z9 I/O interfaces fall into three categories: those with sources and sinks on chips in the same multichip module, those with chip sources from one MCM and chip sinks on another MCM module, and miscellaneous interfaces that, regardless of link topology, run at slower speeds.

The miscellaneous I/O interfaces in z9 operated at hundreds of megabits per second. Because they are relatively slow, there was no need for exotic signaling or retiming. Data was transmitted over pure synchronous boundaries with source-terminated drivers and rail-clamped receivers. **Figure 3** shows a circuit example of the driver and receiver used in z9 miscellaneous interfaces such as JTAG and slower diagnostic buses.

The driver circuit in Figure 3(a) represents an n-FET/p-FET network in which the number of active devices controls the drive impedance. The receiver in Figure 3(b) shows a receiver comprising three tri-state inverters that can be asymmetrically enabled and disabled to control switching threshold.

For the higher-speed on-MCM interfaces (source and sink on the same module), z9 data rates reached 1.72 Gb/s per pin. As with the miscellaneous data drivers, the high-speed data drivers were source-terminated and tuned to the "point of diminishing returns" with respect to far-end edge rate. At the receiver inputs, an aggressive clamping circuit clipped overshoots and undershoots to minimize inter-symbolic interference and maximize setup and hold margins. Figure 3(c) is a schematic illustration of the z9 aggressive clamp.

The aggressive clamping circuit had threshold voltages artificially raised above ground and below voltage supply. Essentially, when the voltage on a receiver input rose above the upper threshold, a path to ground would activate, pulling the receiver input down to that upper threshold. Conversely, when the voltage at the receiver input fell below the lower threshold, a path to the voltage supply would activate, pulling the receiver input up to the lower threshold. The eye diagrams in **Figures 4(a)** and **4(b)** respectively show the data quality comparing a Thevenin-terminated receiver with an aggressively clamped receiver. The faster edge rate and jitter reduction shown by the aggressively clamped eye diagram provide greater timing margin for data capture than that for the terminated eye.

The off-MCM buses, operated at a rate of 860 Mb/s per pin. They shared the same aggressive clamping circuits as the on-MCM receivers but incorporated a Thevenin termination for backup purposes. This protected against cases in which the clamp circuit could have been overwhelmed by incoming current and would not have been able to sink enough current for effective overshoot/undershoot clipping.

For the off-MCM buses, predistortion-capable voltage-mode drivers were used. Unlike current-mode drivers, in which predistortion is achieved by varying drive current and keeping drive impedance constant, the z9 off-MCM drivers varied both drive current and drive impedance. Predistortion compensates for signaling losses, which are

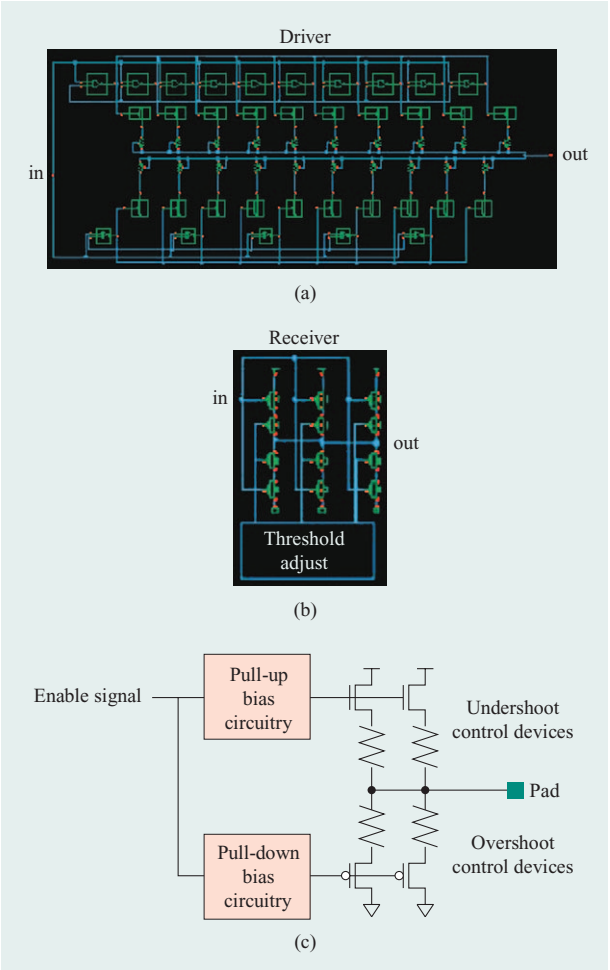

Figure 3

(a) Driver and (b) receiver circuits; (c) aggressive clamping circuit.

greater at higher frequencies and less at lower frequencies. With respect to frequency content, a transitioning signal has more high-frequency content than a steady one. To compensate for z9 frequency losses, the off-MCM drivers pushed data transitions with a lower impedance and maintained voltage levels with a higher impedance. Analysis showed that jitter incurred by a non-source-terminated driver was less than the margin gained because of predistortion. **Figures 4(c)** and **4(d)** show an eye diagram at the receiver input before and after predistortion.

Process variations are the root cause for hardware-to-simulation-model mistracking. Since the z9 was implemented in 90-nm technology, the signaling circuits were not immune to threshold shifts. For this reason, the z9 I/O interfaces implemented receivers with adjustable thresholds as well as threshold-tracking schemes. For interfaces that stayed on the MCM, the voltage-supply

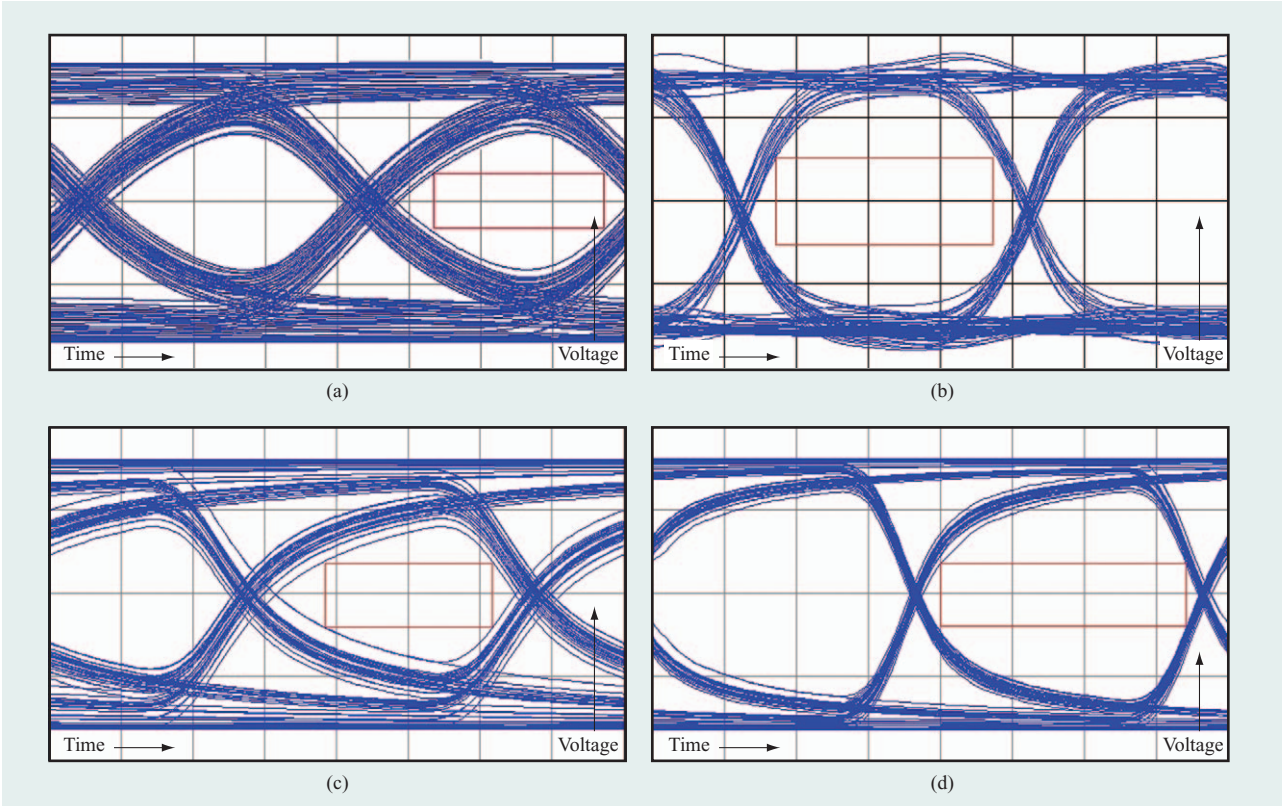

Figure 4

(a) Terminated eye opening vs. (b) aggressively clamped eye opening. Receiver input (c) before and (d) after predistortion.

voltage was assumed to be similar and warranted only minor threshold. The z9 high-speed interfaces, which ran across multiple chip and power boundaries, required a more aggressive threshold-adjust/tracking method.

Because they are source-synchronous, z9 high-speed interfaces required a sampling clock to be sent along with data. For the z9 off-MCM buses, this meant a differential clock for each data group. Because they have the same circuit topology, the clock and data drivers reacted similarly to power-supply variation, clock jitter, and process skew. On the receive end, the same commonality was maintained with respect to termination, locality, and on-chip wiring. Because of this, the common-mode voltage of the received differential, source-synchronous clock proved to be an excellent receiver (rcv) threshold for the z9 off-MCM buses. **Figure 5** shows the System z9 differential clock-termination and threshold-generation circuit. The four vertical resistors in Figure 5 represent the Thevenin termination for the source-synchronous clock "clk" and its inverse, "clk_b." The two horizontal resistors are high in value and reduce interaction between the incoming "clk" and "clk_b." The capacitors filter noise from the generated threshold, "rcv threshold."

Elastic Interface (EI) hardware verification

Early in the machine design phase, budgets are established to support the performance of the elastic interfaces. Two critical budgeted items are target setting and eye size. Since these specifications are critical to the hardware meeting the design goals of the machine, their values are established with detailed analysis and later verified on the product test floor with hardware.

The target setting reflects the bus latency impact to the performance of the machine. The target setting establishes the number of bus clock cycles used for a data transfer from the clock domain of one chip to that of the other. Although the interfaces support a wide variation in the bus latency, a number of buses have aggressive target requirements. Initial target settings are based on the analysis of clock distribution, critical chip path delay, and package delays.

Eye size bounds the minimum valid data eye of particular buses in the design at the receiving chip at the data sample point. Establishing this budget requires critical I/O circuit, signal transition, and noise analysis.

These interface budgets are critical to understanding the interface performance and thus the machine

performance. The budgets are derived as the machine plan is developed, and are set early as chip and package design are started, monitored as the design progresses, and finally verified in the product hardware. Test floor work on the machine allows the characterization of these budgeted items, providing a critical calibration of our predictions, modeling tools, and manufacturing process. Part of the machine bring-up work involves the characterization of the interchip bus characteristics. This is achieved via a test plan which exercises the hardware across the environment, frequency, and process extremes.

The chip interfaces have initialization processes, settings, and result registers which are used for characterization. During machine power-up and initialization, the "interface alignment pattern and elastic interface calibration" (IAP/EIcal) process de-skews data and centers clocks in the resulting eye. Put simply, data de-skew is achieved by delaying the earliest data bits to the delay of the latest bit of the bus. The data eye is next evaluated, and the clock delayed to be centered in the eye. Finally, guard-band registers are initialized to detect the actual eye size on the basis of random data interface transfers to further optimize the clock centering. Extracting the register data associated with these functions provides the measured horizontal eye size. Verifying the target setting involves characterizing the interface margin under different clock speed, voltage, temperature, and process.

6. Power delivery

One function of electronic packaging is to deliver power to the chips. The task is to provide sound voltage to various chip, module, and card locations while considering a variety of requirements and limitations. The required voltage boundaries must be guaranteed at various and distant locations within a power domain, even under the condition of variable load. Appropriate sense points must be defined, maximum pin currents in modules or connectors as well as maximum current densities in printed circuit boards (PCBs) must be verified, and total power dissipation must be predicted. Safety questions are also involved as dissipated power heats up the boards, connectors, and carriers. Consequently, power distribution analysis (dc analysis) is part of the entire system/packaging design flow. Depending on whether or not physical design (PD) data is available, it is referred to as PrePD or PostPD dc analysis. During the early system design, PrePD dc analysis is required to define and optimize the system accordingly to the given requirements and specifications. Since no physical design data is created at this time, this analysis requires tools that can efficiently generate PrePD simulation models [8]. Depending on the required analysis, these models are very complex, containing some

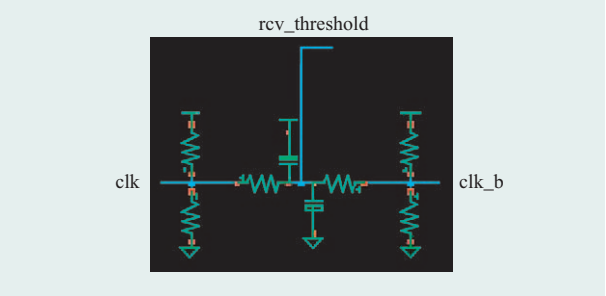

Figure 5

System z9 clock-termination and threshold-generation circuit.

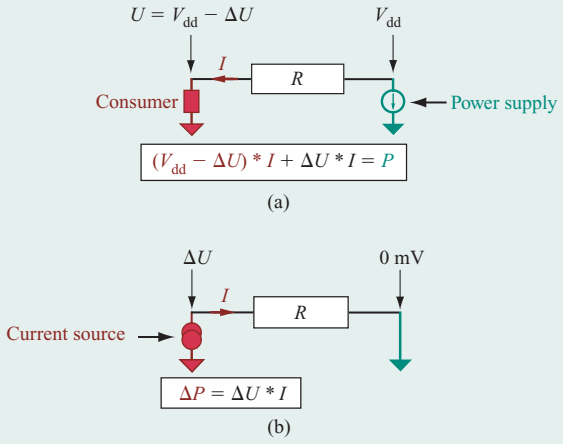

Figure 6

(a) PDS consisting of power supply, packaging, and consumer; (b) its "inverted model."

tens of thousands of sources and metallic bodies. For the PostPD dc analysis, scripts are available to extract the simulation model from the CADENCE** layout. For the analysis we use RGEN, an IBM tool for resistance calculation [9].

Modeling for dc analysis

Typical power-distribution systems (PDSs) consist of one or multiple power supplies and a packaging that distributes current to one or multiple consumers. Since consumers require a constant voltage supply, we consider a PDS with voltage sources and consumer as depicted in **Figure 6(a)**.

The term R represents the packaging that provides a (resistive) path to the consumer. The current I is the current requirement of the consumer. $\Delta U = R * I$ is the voltage drop in the packaging, and $U = V_{dd} - \Delta U$ is the

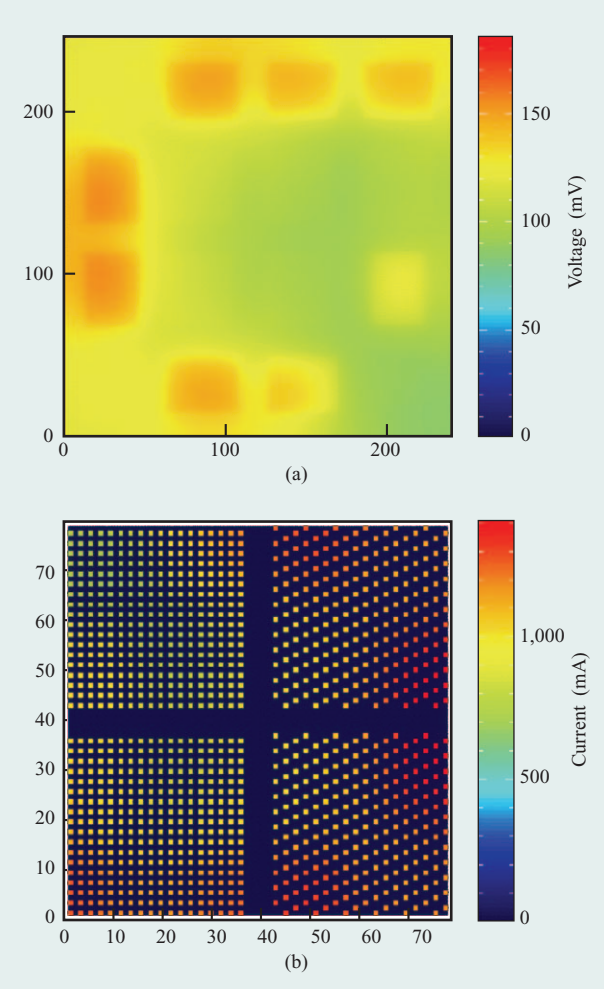

Figure 7

(a) Voltage drop in mV at the MCM. The numbers at the *x*- and *y*-axes define the location in multiples of the resolution of 0.4 mm. (b) LGA pin current distribution. The *x*- and *y*-axes give the position of the pin relative to the pin in the lower left corner in mm.

voltage available at the consumer. $P = V_{dd} * I$ is the power delivered from the supply, and the power dissipated in the packaging is $\Delta P = \Delta U * I = R * I^2$. To predict the voltage drop ΔU and the dissipated power ΔP, the "inverted model" **[Figure 6(b)]** is derived, as shown in [8]. In this model, power supplies are replaced with ideal connections to ground, and consumers are replaced with a corresponding current source. The current in the inverted model is inverted; the simulated voltage is the voltage drop. The extension of this model to systems with multiple consumers is straightforward. Since multiple voltages must be provided, this modeling approach is applied to all voltage nets separately as well as to the GND net which provides the common return.

PrePD dc analysis

During high-level design, no design data is available with which to build a model. In addition, sensitivity analysis requires a methodology that makes it possible to modify design parameters in the model. Because of the complexity of the model, a graphical user interface (GUI) is not the appropriate tool for generating the simulation input. AMOC, an IBM-developed programming language for efficiently building simulation models, enables us to assemble chip and card models using ground connections, power supplies, conductor elements, and current sources. The models contain the backplane, the node card, the MCM, and an I/O adapter card. The high-pin-count connectors between cards and module are modeled on a per-pin basis. The eight memory cards and I/O cards are simplified to equivalent current sources; this has no impact on the accuracy on the board. The on-MCM current sources are applied at the chip locations on the MCM and represent an inhomogeneous on-chip power distribution, with the power supply connected to the backplane. The resulting models contain ~100,000 elements. The RGEN simulation takes approximately one hour per net. In the following sections, some variations of the dc analysis are outlined that were used iteratively to optimize the system design.

System voltage variation and sense-point definition

This section shows the voltage drop variation across the system. This analysis was performed iteratively during system definition and design in order to ensure that each chip receives the required voltage level. The voltage variation across the system was minimized by optimizing component placement, cross sections, and split-plane configurations.

The voltage distribution on top of the MCM is shown in **Figure 7(a)**. The shown values combine the drops for V_{dd} and GND. The example shows a variation of 50 mV among the eight central processor (CP) chips. The higher drop can be seen at the two CPs on the left side of the module. In the left half of the MCM, the signal-to-power pin ratio is 2:1 because of the high signal count of the memory and GX+ buses; the right side has a 1:1 ratio, causing a higher voltage drop on the left as well as higher LGA pin currents.

Connector verification and current-balancing optimization

Connector current balancing is important for achieving equal voltage drops and verifying that no pin exceeds the maximum current per pin. **Figure 7(b)** shows the V_{dd} pin pattern and its corresponding current distribution. At the left side, the signal pin density has been increased, leaving lesser pins for power distribution. This causes the highest current load per pin in this area.

The ratio between maximum and minimum pin current is called the *current-balancing ratio* (CBR). For the land grid array (LGA), the $CBR = 2$. The analysis has been used to design and optimize the interplane connector that connects the node card to the backplane. In order to support the required current of up to 1,400 A with one connector, a $CBR < 1.25$ was required. Special design features have been introduced that reduce the initial $CBR = 1.8$ to a value of $CBR = 1.2$. This was achieved by placing gaps and slots in the power and ground planes on both cards.

Design verification for maximum DCA currents

For product safety and customer protection reasons, overload situations are considered in the design. The boards and connectors must be able to handle any current load the DCAs can provide. Therefore, the packaging is designed to make sure that connector currents are within the specified values. The specified maximum currents are derived for a 30°C temperature rise in the connector area. In addition, the power plane configuration has been optimized in order to reduce the maximum current densities in the board. This study has been done on a test vehicle on which the MCM was replaced by a multitude of resistors placed in the lower left area of the node card. **Figure 8** shows the horizontal current density distribution for the common return path GND on the test vehicle. For reasons of efficiency, the planes are represented by one equivalent sheet. At each x/y location, the square resistance represents the combined thickness of all layers. The plotted value is the current divided by the board thickness.

A commonly used copper thickness is 32 μm, which is referred to as one ounce (1 oz). With this notation, the unit used for current density is A/cm/oz. In general, current crowds at connector locations; in these areas, power planes are perforated, causing the current and power density to increase even more. However, the connector pins allow heat to dissipate from the internal board layer. In connector areas, we found that current densities of 20 A/cm/oz were not critical. Outside the connector areas, high current density can cause more local heating. Therefore, we limited the current density outside the connector areas to 5 A/cm/oz (10 mA/mil^2).

Power distribution losses

Because of the high currents, power distribution losses must be considered in the system design. The dc analysis readily provides these numbers, as shown in Figure 6. The only sources of power are the current sources, and all power is dissipated in the packaging. The maximum power distribution losses have been derived for the path from the voltage regulators to the node card and the riser card. During regular operation (with the MRU working),

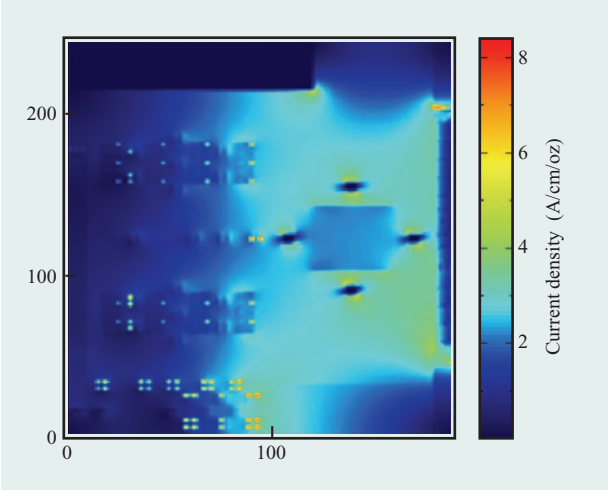

Figure 8

Current density plots on GND planes. The numbers at the x- and y-axes define the location in multiples of the used resolution, in this case 2.0 mm.

the maximum power distribution losses are 237 W per node (**Table 3**). This number assumes a total current on GND of 1,067 A.

PostPD dc analysis

In PostPD analysis, the focus switches from the system level used for the PrePD analysis to first-level packaging (i.e., modules). This is because PrePD analysis does not allow accurate prediction of local voltage distribution on first-level packaging. Corresponding to the complexity of first-level packages, the dimensions of the linear system resulting from the use of real design data were too great for PostPD MCM analysis. Thus, we had to analyze each chip site individually to keep the simulation time in a useful dimension. This introduced errors due to separation, but the global calculated MCM dc drop values were proven to be correct. The etch resistance is about five to ten times higher than the via resistance, and therefore the current flows vertically and, in general, chip site analysis still can provide accurate IR drop information. Module physical design data (allegro file) is converted to an RGEN input file, which is used to perform the dc analysis. The postPD analysis verifies the assumptions of the simplified PrePD models and allows fine adjustment of the design parameters.

There are jogging layers[1] at the top and bottom sections of the z9 MCM. Between these two jogging layers are redistribution layers and xy layers from top to

[1]In the jogging layers, all signal and power vias have a small jog in the wire, which is required by the ceramic technology in order to achieve the best reliability.

Table 3 Power dissipation (W) in the packaging. The total dissipated power per node is 237 W.

Domain	*1V0*	*1V5*	*1V8*	*3V3*	*3V4*	*GND*	*Power* (W)
Current I (A)	808	145	104	5.6	4.2	1,067	
Fan-out cage adapter card	0.10	0.50	0.10	0.19	0.03	0.10	1.0
Adapter card–connector	0.04	0.60	0.04	0.02	0.01	0.15	0.9
Processor card	26.40	11.70	6.00	0.04	0.07	19.80	64.0
Center board–connector	17.00	2.70	1.70	0.03	0.04	20.00	41.5
Center board	38.50	12.20	8.80	0.04	0.08	16.60	76.2
DCA–connector	29.00	3.60	2.40	0.02	0.01	18.50	53.5
Total dissipated power (W)	111.0	31.3	19.0	0.3	0.2	75.2	237.1

Table 4 GND and V_{dd} maximum IR drop at different CP chip sites across the MCM.

dc analysis (mV)	*CP0*	*CP1*	*CP2*	*CP3*	*CP4*	*CP5*	*CP6*	*CP7*
GND top layer maximum	52.7	49.0	65.3	74.6	70.5	65.1	49.7	48.6
V_{dd} *top layer maximum*	52.6	52.3	61.8	74.8	73.7	60.4	51.2	49.0

bottom. In the redistribution layers, the power delivering via density is higher than in the *xy* layers. From the point of view of performance, a higher GND, V_{dd} via density is wanted for the design. On the other hand, the yield of an MCM is closely related to the risk sites in the MCM. Risk sites can be identified in very dense wiring and via areas of the MCM. Increased via density in the redistribution layers and *xy* layers increases the number of risk sites in the MCM. At a certain phase of this MCM design, the redundant vias in the redistribution layers had to be removed to reduce the number of risk sites for better yield. The dc analysis mentioned previously is best for this kind of analysis. Such analysis provides a way to optimize the IR drop number and the number of risk sites, thereby optimizing performance and yield. Through the dc analysis, we also verified that it is better to have straight V_{dd} or GND vias from the top jogging layer to the bottom jogging layer for IR drop reduction. In this MCM, we have long via connections on all of the 700-μm grids from the top to the bottom jogging layer. In the MCMs of previous systems, there were more jogging layers; we later found that this induced higher IR drops in the MCM.

Up to eight processor chips can be packed onto the z9 MCM. The power delivery net GND and V_{dd} are analyzed for each processor chip site. The maximum IR drop numbers are listed in **Table 4**. We can see that owing to the plane pattern, the via pattern variation inside the MCM, and especially the connector pattern variation across the bottom layer of the MCM (we usually have four quadrants of connector pins at the bottom layer), the average IR drop is different from chip site to chip site. The PostPD dc analysis makes it possible to fine-tune the chip partitioning and BSM pin pattern to minimize the IR drop difference across the GND and V_{dd} C4s on the top layer of the MCM.

This IR drop information is also very useful when the MCM is configured to different node sizes. For instance, if the MCM is configured to a 12-way node system with four dual-core CP chips and four single-core CP chips, performance will be better if four dual-core chips are placed on positions 0, 1, 6, and 7, and the rest are used for single-core chips. If the MCM is offered as a 16-way node, the chip-sorting strategy allows the faster dual-core CPs on positions 2, 3, 4, and 5 to operate at lower voltage and the slower dual-core CPs to go to the rest positions for best performance.

A PostPD analysis that uses real design data is very helpful for identifying potential design defects such as shorts, open circuits, missing etches, and hot spots. Additionally, by providing the detailed voltage gradients across a module, the PostPD analysis can identify potential signal integrity risks due to insufficient shielding-via density. In areas with reduced shielding-via density, the current sees longer horizontal paths, which results in significant voltage gradients because of the high resistance of the horizontal wiring. Therefore, very high IR drops of some of the V_{dd} or GND C4s indicate the risk of an inadequate return path or of increased coupling for signals located in that area.

7. Cooling

The System z9 processor cage utilizes several advances in the hybrid cooling technology that was introduced in the System z990. Hybrid cooling uses direct refrigeration for normal cooling, with an air-cooling backup system for use in the event of a refrigerant failure. **Figure 9** is a schematic diagram of this concept in which one modular refrigerant unit (MRU) cools the MCMs in two nodes. Previously, in the G4, G5, G6, and z900 servers, IBM utilized fully redundant refrigeration, with two MRUs

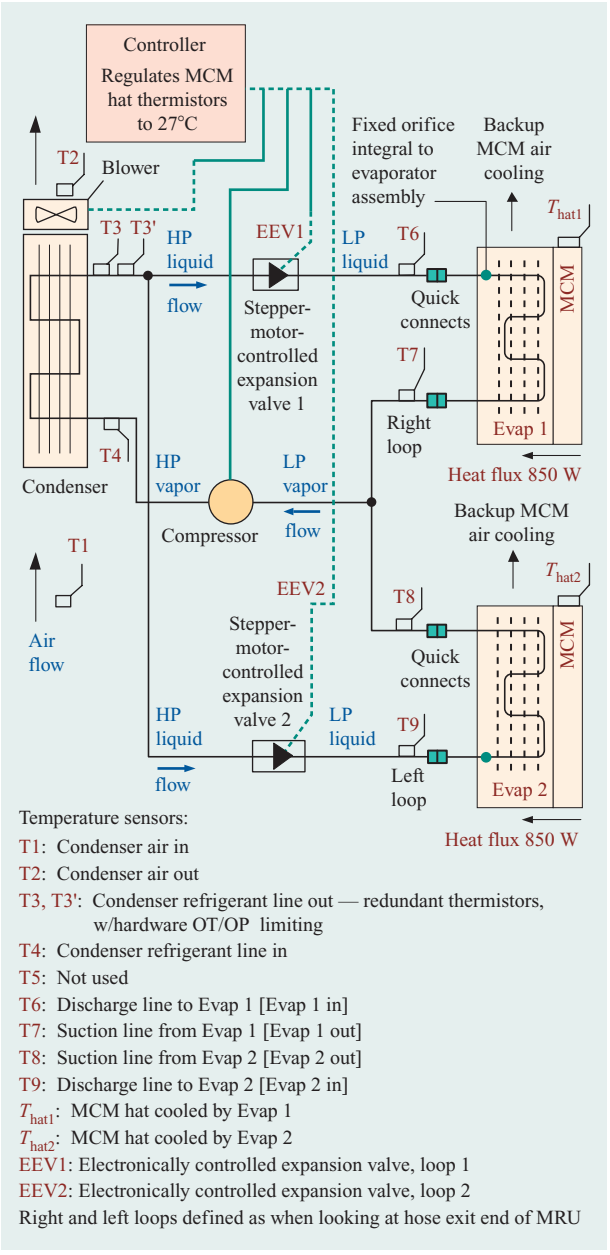

Figure 9

Schematic diagram of System z9 hybrid cooling. One cooling unit is used to cool two nodes to a nominal junction temperature of 45°C for the chips. The picture shows the high-pressure (HP) and low-pressure (LP) flows.

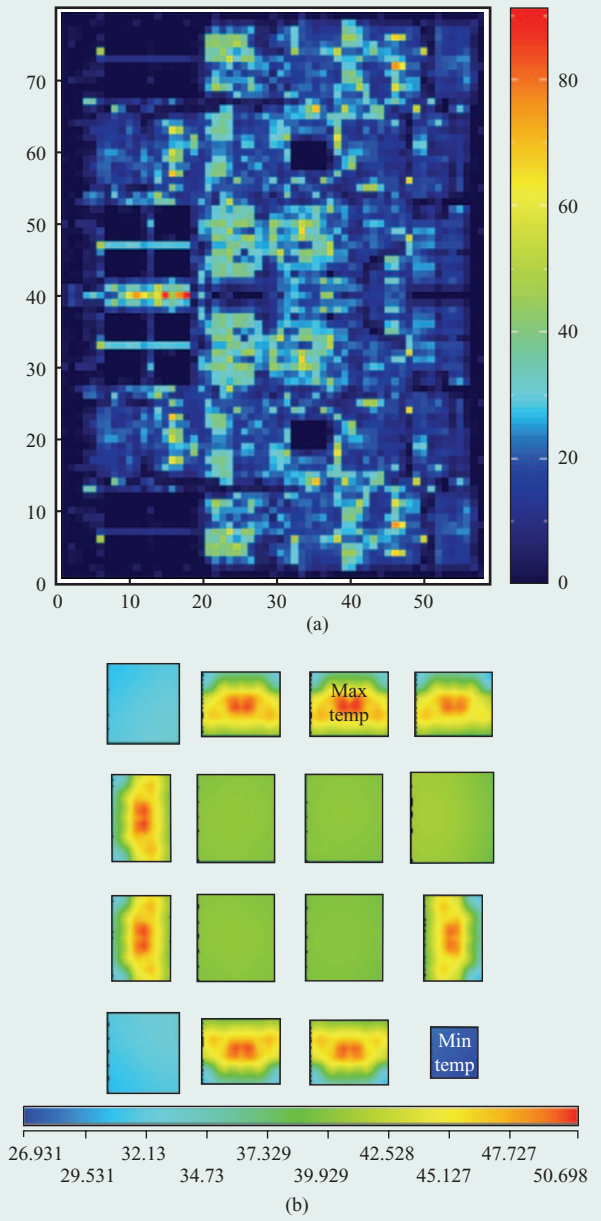

Figure 10

(a) Power and current distribution within the processor chip. Average current map in mA/cell (matrix dimensions: 59 × 79 cells). (b) Approximate nominal junction temperatures in degrees C on MCM while MRU-cooled (ANSYS** modeling by Amilcar Arvelo).

cooling a single MCM. The z9 superblade concept incorporates four MCMs on four nodes and requires the spatial, cost, and power efficiency achieved with hybrid cooling.

In normal operation, the modular refrigerant unit enables the high-power processors and cache to operate in the 30°C and 40°C ranges. **Figure 10(a)** shows the active and leakage power distributed across a typical processor chip. **Figure 10(b)** shows the approximate junction temperatures on each chip in the 16-way MCM, assuming that the power is uniformly spread in the SD, SC, MSC, and clock chips.

The lower junction temperatures enabled by refrigeration yield three benefits—faster, more reliable,

and more energy-efficient servers. First, CMOS circuits switch faster at lower temperatures and higher voltages, enabling faster clock speeds. It can be shown that at normal acoustic blower speeds, the server processor clock speed would have to be reduced by 12% to 14% if air is cooled because of temperature and voltage effects.

Second, the circuits and package function with greater reliability than if air-cooled. The refrigerant system controls the MCM hat thermistor to 27°C, slightly above room ambient. Mitigating temperature swings during power-on/off cycles increases the reliability of circuits and package, which are critical components in System z9 server reliability.

The third benefit is less intuitive. When compared with air cooling, refrigeration often saves energy at the server level. While the compressor and fan in the refrigerant system require additional power, we have selected an evaporation temperature between 5°C and 10°C at which the compressor workload is minimized and condenser fan speeds are low. Offsetting the refrigeration power is less device leakage power due to lower junction temperatures plus less CEC blower power. In Table 2 we find that despite the lower frequency and voltage applied during the air-cooled backup mode, the MCM power climbs from 850 W to 1,200 W because of increased leakage with warmer junction temperatures. The leakage power, coupled with the additional blower power when the MCM is air-cooled, may result in server-level power that is higher for the air-cooled mode than for the normal MRU-cooled mode. Faster and more reliable circuits, power, and space efficiency are attributes hybrid cooling brings to z9.

From a hardware design perspective, we found that the System z990 MRU with an optimized evaporator was able to handle the small System z9 increase in power vs. the z990 MCM at low junction temperatures. However, the additional chip leakage found in the latest circuit technology drove the air-cooling backup power much higher, to 1,200 W. Our System z9 thermal design provides sufficient air cooling in five steps:

1. Minimizing the chip power. System z9 is the first server with both reduced frequency and reduced voltage during cycle steering to limit power increases in air-cooling mode.
2. Optimized placement of the highest-leakage chips. CPs and SDs were divided into standard and maximum power part numbers and controls put on locations for the highest-powered parts to be placed on the MCM to optimize module thermal conduction paths.
3. Development of a more powerful blower, approximately doubling the pressure and airflow available for the z9 MCM heat sink in comparison with the System z990.
4. Coupling of this more powerful blower with a heat sink having 20% denser fin pitch soldered to the top of the evaporator lid; this arrangement yielded additional cooling efficiencies.
5. Design of the z9 module hat using small-gap technology (SGT), which was first utilized in the POWER5 high-end server in 2004. SGT employs copper pistons above each of the eight processor chips which are soldered in place after the thermal interface material (TIM1) gap between the silicon and piston is precisely fixed at 0.1 mm. This process decreases the nominal TIM1 resistance by a third and the maximum by more than a half.

The improved airflow, optimized heat-sink fin pitch, lower TIM1 resistance with smaller gaps between silicon and hat, and optimal placement of the highest-powered chips together allow the z9 package to support the power increase found in this latest technology.

8. Summary

The System z9 central electronic complex structure has followed the successful System z990 approach with a modular node structure housing the multichip module with the processor and cache chips, the main memory, and the I/O hub. This system has achieved significant granularity improvements by allowing concurrent node upgrade during normal system operation. It allows the customer to change the configuration with respect to memory cards or I/O cards while the other nodes are running in the system. For the increased multiprocessor performance, a balanced system design has been ensured by increasing the bandwidths on the package. This was done by the EI high-speed interface at a speed of 1.72 Gb/s on a single-ended line. It resulted in a raw packaging bandwidth of 441 GB/s per node between processor and cache chips. Because of the increased leakage current of the 90-nm technology, the power delivery and heat removal of the MCM have become a major challenge. Careful design of the first- and second-level package was required in order to guarantee the voltage stability within 100 mV for a worst-case current of 1,100 A in the air-cooled backup mode. Small-gap technology was used to accomplish the cooling of a 1,200-W multichip module.

References

1. H. Harrer, H. Pross, T.-M. Winkel, W. D. Becker, H. I. Stoller, M. Yamamoto, S. Abe, B. J. Chamberlin, and G. A. Katopis, "First- and Second-Level Packaging for the IBM eServer* z900," *IBM J. Res. & Dev.* **46**, No. 4/5, 397–420 (2002).
2. T.-M. Winkel, W. D. Becker, H. Harrer, H. Pross, D. Kaller, B. Garben, B. J. Chamberlin, and S. A. Kuppinger, "First- and Second-Level Packaging of the z990 Processor Cage," *IBM J. Res. & Dev.* **48**, No. 3/4, 379–394 (2004).
3. G. A. Katopis, W. D. Becker, and H. Harrer, "T-Rex, a Blade Packaging Architecture for Mainframe Servers," *IEEE Trans. Adv. Pkg.* **28**, No. 1, 24–31 (2005).
4. E. Cordero, D. Dreps, F. Ferraiolo, M. Floyd, K. Gower, and B. McCredie, "A Synchronous Wave Pipeline Interface for POWER4," presented at the IEEE Computer Society HOT CHIPS Workshop, Stanford University, CA, August 15–17, 1999.
5. J. C. Parrilla, F. E. Bosco, J. S. Corbin, J. J. Loparco, P. Singh, and J. G. Torok, "Packaging the IBM eServer z990 Central Electronic Complex," *IBM J. Res. & Dev.* **48**, No. 3/4, 395–407 (2004).
6. P. A. Coico, G. Messina, S. Ostrander, J. Zitz, and W. Zou, "Internal Thermal Management of IBM p-Server Large Format Multichip Modules Utilizing Small Gap Technology," *Proceedings of the ASME/PACIFIC Rim Technical Conference and Exhibition on Integration and Packaging of MEMS, NEMS and Electronic Systems*, 2005, Paper 73422, six pages.
7. A. Huber, T. Zhou, W. D. Becker, R. Weekly, and E. Klink, "Power Distribution Analysis for IBM eServer System Integration Optimization," *Proceedings of the Conference on Electrical Performance of Electronic Packaging (EPEP)*, Portland, OR, 2004, pp. 189–192.
8. IBM Electromagnetic Field Solver Suite of Tools, May 2006; see *www.alphaworks.ibm.com/tech/eip*.
9. D. M. Berger, J. Y. Chen, F. D. Ferraiolo, J. A. Magee, and G. A. Van Huben, "High-Speed Source-Synchronous Interface for the IBM System z9 Processor," *IBM J. Res. & Dev.* **51**, No. 1/2, 53–64 (2007, this issue).

Received June 21, 2006; accepted for publication August 24, 2006; Internet publication December 5, 2006

Hubert Harrer *IBM Systems and Technology Group, IBM Deutschland Entwicklung GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (hharrer@de.ibm.com)*. Dr. Harrer is a Senior Technical Staff Member working in the IBM Systems and Technology Group. He received his Dipl.-Ing. degree in 1989 and his Ph.D. degree in 1992 from the Technical University of Munich. In 1993 he received a DFG research grant to work at the University of California at Berkeley. Since 1994 he has worked for IBM in the Boeblingen Packaging Department. In 1999 he was on International assignment at IBM Poughkeepsie, New York. Dr. Harrer's interests currently focus on System z first- and second-level packaging design for the IBM Systems and Technology Group. He has published multiple papers and holds six patents in the area of first-level and second-level packaging.

Daniel M. Dreps *IBM Systems and Technology Group, 11400 Burnet Road, Austin, Texas 78758 (drepsdm@us.ibm.com)*. Mr. Dreps is a Distinguished Engineer working in the IBM Systems and Technology Group. He received his B.S.E.E. degree in 1983 from Michigan State University. During his IBM career, he has designed and developed transistor models, fiber optic links, ASIC technology custom elements, and high-speed serial links for IBM servers. His interests currently focus on high-speed serial development and applications in the entire range of IBM servers. He has published multiple papers and holds more than 30 patents in broad areas of interconnect and server design.

Thomas-Michael Winkel *IBM Systems and Technology Group, IBM Deutschland Entwicklung GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (winkel@de.ibm.com)*. Dr. Winkel received his Diploma in electrical engineering in 1989 and his Ph.D. degree in 1997 from the University of Hannover, Germany. His research activities covered the area of characterization and modeling of on-chip interconnects using high-frequency measurements. In 1996 he joined IBM Development in Boeblingen, Germany. He is currently an Advisory Engineer in the IBM Systems and Technology Group, leading the electrical design team for second-level packaging in Boeblingen. Dr. Winkel's current focus is on electrical packaging design with respect to high-frequency signal distribution, power integrity, and system aspects for the pSeries* and zSeries. He is also interested in high-frequency on and off-chip measurements and the modeling of signal as well as power and ground lines. In this area he has led several internal and external high-frequency measurement projects up to 65 GHz. Dr. Winkel has authored or coauthored more than 20 conference and journal papers; he holds eight patents. He is program chair of the 2007 IEEE European Packaging Workshop and a member of the technical program committee of the workshop on Signal Propagation on Interconnects (SPI).

Wolfgang Scholz *IBM Systems and Technology Group, IBM Deutschland Entwicklung GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (scholzw@de.ibm.com)*. Mr. Scholz is an Advisory Engineer working in the IBM Systems and Technology Group. He graduated with a B.S. degree in telecommunications in 1979 at the GHS Paderborn and joined IBM at the Boeblingen laboratory in 1984. He has held various technical positions in S/390* processor power and packaging design in Boeblingen and Poughkeepsie. Returning from Poughkeepsie to Boeblingen in 2000, Mr. Scholz joined the Boeblingen processor card and board design team. In 2002 he certified as a Project Management Professional at the PMI. In 2003 he became card project manager, leading the overall packaging project for the z9 processor cage design.

Bao G. Truong *IBM Systems and Technology Group, 11400 Burnet Road, Austin, Texas 78758 (bao@us.ibm.com)*. Mr. Truong received a B.S. degree in electrical engineering from the University of Texas at Austin in 1999. Since 1998, he has been a member of the IBM Systems and Technology Group in Austin, Texas, working in high-speed link design. His areas of specialty include interconnect analysis, circuit design, and design optimization for the IBM pSeries and zSeries platforms. He currently works on SRAM design and power optimization.

Andreas Huber *IBM Systems and Technology Group, IBM Deutschland Entwicklung GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (hubera@de.ibm.com)*. Dr. Huber is an Advisory Engineer working in the IBM Systems and Technology Group. He received his Dipl.-Ing. degree in 1996 and his Ph.D. degree in 2001 from the Technical University of Karlsruhe, Germany. Since 1999 he has worked in the Packaging Department at the IBM Boeblingen Development Laboratory, working on first- and second-level packaging. From 2003 to 2006 he was on international assignment at IBM Austin, Texas. Dr. Huber's field of interest is power distribution and power integrity. He has published a number of papers and has reached the third patent application plateau.

Tingdong Zhou *IBM Systems and Technology Group, 11400 Burnet Road, Austin, Texas 78758 (tingdong@us.ibm.com)*. Dr. Zhou received the B.S. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1992 and 1995, and the Ph.D. degree in electrical engineering from the University of Arizona at Tucson in 2002. Since 2002, he has been with the IBM Systems and Technology Group, working on the electrical design of the components that comprise a high-frequency CMOS processor system. His current interests focus on electromagnetic modeling of high-speed circuits, electrical modeling and simulation of high-speed and high-performance packages, and signal integrity issues in the design of server systems. Dr. Zhou has published more than twenty conference and journal papers.

Kenneth L. Christian *IBM Systems and Technology Group, 2455 South Road, Poughkeepsie, New York 12601 (kenchris@us.ibm.com)*. Mr. Christian received a B.S. degree from Southern Illinois University. He is an Advisory Engineer in the I/O and Package Development Department at the IBM Poughkeepsie Development Laboratory. He joined IBM in 1984 and has worked on package signal integrity and timing across module, card, and boards. Mr. Christian holds three patents; he received an IBM Outstanding Technical Achievement Award in 1999 for development of the NetRules tool for system and package design.

Gary F. Goth *IBM Systems and Technology Group, 2455 South Road, Poughkeepsie, New York 12601 (gfgoth@us.ibm.com)*. Mr. Goth received his B.S.E. degree from Princeton University in 1971, his M.S. degree in mechanical engineering from Union College in 1975, and his M.S. degree in statistics in 1978 from Rensselaer Polytechnic Institute. Since joining IBM in 1979, he has held managerial and technical positions in server group development. Currently a Senior Technical Staff Member, Mr. Goth is responsible for high-end server cooling. He holds 29 patents in thermal technologies, with nine patents pending. He has coauthored several technical publications and has received IBM Outstanding Innovation and Outstanding Technical Achievement Awards in thermal technology.