Electronic packaging of the IBM System z196 enterprise-class server processor cage

In this paper, we describe the first- and second-level system packaging structure of the IBM zEnterprise™ 196 (z196) enterprise-class server. The design point required a more than 50% overall increase in system performance (in millions of instructions per second) in comparison to its predecessor. This resulted in a new system design that includes, among other things, increased input/output bandwidth, more processors with higher frequencies, and increased current demand of more than 2,000 A for the six processor chips and two cache chips per multichip module. To achieve these targets, we implemented several new packaging technologies. The z196 enterprise-class server uses a new differential memory interface between the processor chips and custom-designed server memory modules. The electrical power delivery system design follows a substantially new approach using Vicor Factor Power™ blocks, which results in higher packaging integration density and minimized package electrical losses. The power noise decoupling strategy was changed because of the availability of deep-trench technology on the new processor chip generation.

Introduction

With its general availability announced in September 2010, the IBM System z* 196 was introduced to the family of IBM System z high-end servers [1, 2]. It replaces its predecessor, the IBM System z10* Enterprise Class (z10* EC), offers 50% higher symmetric multiprocessor (SMP) performance, and adds new features such as enhanced virtualization and new specialty engines, which were previously unavailable in System z high-end servers [3]. Similar to z10, the IBM System z196 is designed as a supernode architecture with up to four processor unit (PU) books that communicate over a common backplane.

The SMP performance increase was achieved mainly by a higher central processor (CP) frequency (5.2 GHz versus 4.4 GHz), by largely increasing the caches available on the CP and the new cache [i.e., system controller (SC)] chip, and by adding a sixth CP chip to each multichip module (MCM). With two SC chips per MCM, this results in a total of 24 quad-core CP and 8 SC chips for a fully populated system. On each PU book, the CP chips communicate with up to 30 dual inline memory modules (DIMMs), each with a maximum capacity of 32 GB. Due to memory redundancy features, this leads to a total memory availability of up to 768 GB per PU book. In contrast to its predecessor, the IBM System z196 uses a differential memory interface (DMI) at a signal speed of 4.8 Gb/s per lane. The interconnect buses between the PU books are operated at 2.6 Gb/s using a fixed gear ratio of 1:2 with respect to the CP speed.

This paper focuses on the packaging structure of the z196 server. This paper is structured as follows: The first section contains an overview of the packaging components in the central electronic complex (CEC). It describes the logical system structure and focuses on the electrical properties of the packaging elements. The second section describes the challenges that had to be overcome in the z196 system design to deliver the high electrical power of more than 3 kW per PU book to all chips without generating excessive package
power loss and heating. It also describes the architecture of the regulation loop for the electrical power supply system. The third section focuses on the frequency-dependent properties of the power supply system. The CP and SC chips of the z196 server use deep-trench capacitor cells, which vastly increase the electrical capacitance of those chips compared with previous IBM zSeries* chip generations. The effect of this change on electrical power noise is discussed, and results from simulations and measurements are shown. Details of a new methodology that was used to characterize the impedance of the z196 server electrical power supply system over a frequency range from low kilohertz up to 1 GHz are the focus of the fourth section. An introduction of the new DMI is given in the fifth section. This includes details of the architectural implementation, results obtained from simulations, and characterization measurements. During the physical design phase of the first- and second-level packaging components, several new checking methodologies have been developed and applied to ensure that all components meet the requirements defined in advance to ensure functionality. This is described in more detail in the sixth section. New digital measurement approaches allowed us to acquire large amounts of data on elastic interface buses during manufacturing tests. The measurement and characterization methodology is described in the seventh section.

**CEC components and system logic structure**

The frame of the z196 server with a water-cooled option is shown in Figure 1(a). The right-hand frame contains the CEC cage, one conventional input/output (I/O) cage that has been developed for earlier zSeries generations, and the water-cooling unit (WCU). The left frame contains the bulk power generation unit, an optional battery unit as a backup in case of external power failures, and two newly developed I/O drawers. The z196 system supports I/O cage, I/O drawer, and the PCI Express** (PCIe**) I/O drawer.

The WCU chills water to a temperature well below ambient room temperature and pumps it through hoses to the CEC cage. Water cooling is only applied to the six CP chips and two cache chips that are located on one common MCM on each of the four node cards in the CEC. Effective heat removal achieved by this solution allows the system to operate the chipset at higher power densities with cooler chip junction temperatures than was previously possible. As an alternative to the WCU, customers can also obtain the modular refrigerant unit that has been used for chip cooling over several previous zSeries generations [4].

Figure 1(b) shows the detailed assembly structure of the CEC unit, which is similar to that of the z10 EC. There is a common backplane to which a maximum of four PU books can be connected. Two oscillator (OSC) cards and two external timing reference/CEC/flexible support processor (FSP) cards are connected as well. The PU books are connected to the backplane by reusing the 14-row Amphenol Ventura** connector introduced in z10 [2].

Each PU book consists of the processor card, three distributed converter assemblies (DCAs) for power supply, 11 voltage transformation units called gearboxes, one MCM, up to 30 DIMMs, two FSP cards, up to eight I/O cards, and a high-voltage card.

The power distribution concept has been significantly changed compared with the z10 system. In the z10 server, the DCA provided the low voltage levels required to operate the chips; hence, high currents had to be delivered over large distances on the processor card, causing power distribution losses. In z196, the DCAs provide adjustable intermediate voltage levels between 40 and 48 V for the main chip power levels. Gearboxes, which are placed close to the MCM and DIMMs, convert the intermediate voltages locally to low voltages on a fixed gear ratio. Because package power distribution losses are proportional to the square of the current, this approach significantly lowers power distribution losses. As a consequence, this allowed a single board instead of the sandwich structure used in z10 where a separate power distribution board was used in addition to the logic PU board. To further improve efficiency, the main voltage of each CP is individually supplied and controlled by the DCAs.

The logical structure of the CEC unit is shown in Figure 1(c). The core of the system is the MCM, hosting six CP chips and two shared cache (i.e., SC) chips. The z196 CP is a quad-core chip with two threads per core running at 5.2 GHz [5]. It has 1.5 MB of private Level 2 (L2) cache per core and 24 MB of shared L3 cache, two coprocessors for data encryption and compression, a double-data-rate-3 (DDR3) memory controller, and an I/O bus controller. Each z196 SC chip provides 96 MB of shared L4 cache and fabric interfaces to connect between the four PU nodes. On the MCM, each CP is connected to both SCs using a 180-bit-wide single-ended cache bus. We use unidirectional generation 3 Elastic Interface (EI3) connections running at 2.6 Gb/s [6].

Three CPs are connected to fully buffered DDR3 DIMMs using the IBM custom SuperNova buffer chip. Per CP, we use a 390-pin-wide DMI, 160 pins wide for the write direction and 230 pins wide for the read direction. Compared to z10, which has four CPs using 200-pin-wide single-ended interfaces with 98 pins wide for write and 102 pins wide for read, the pin count has been increased by 50% in order to increase bandwidth by 70% and support redundant memory. The DMI runs at 4.8 Gb/s. The DIMMs are arranged in a (4 + 1) × 2 configuration. Four DIMMs are backed up by a fifth DIMM in a redundant array of independent memory (RAIM) device configuration that allows for continued system operation even in case of a complete DIMM failure. The memory DIMMs are cascaded to double available memory size. Of the 960 GB on board, the maximum amount
Figure 1

IBM z196 frame: (a) front view; (b) CEC assembly structure showing one populated node; and (c) logic description of the z196 CEC frame, comprising four nodes. (DCA: distributed converted assembly; DIMM: dual inline memory module; ECF: external timing reference/CEC/FSP; FSP: flexible support processor; HCA: host channel adapter; MCM: multichip module; OSC: oscillator.)
of available memory is 768 GB per node card due to redundancy. A maximum of eight I/O cards can be populated on one PU book to interface with host channel adapters or other peripherals. Each I/O bus connecting the MCM with an I/O card is 100 bits wide, 50 bits for each direction. The CP-to-I/O buses operate at 2.6 Gb/s. Finally, there are two control cards with FSPs on each node.

The system clocks are redundantly created on two independent OSC cards. The redundant clocks are received by one of the two SC chips on each MCM and, from there, are distributed to all other chips on this MCM. Failover logic ensures continued system operation in case of a failing OSC card.

In order to support the increased bandwidth requirements over z10, the signal referencing scheme has been revised. The DMI signals are referenced to ground (GND) and memory voltage. All other signals are referenced to GND only. In addition, the physical design checking methodologies have been improved, enabling a more carefully designed signal return path and smaller discontinuities along the signal path. An overall CEC bandwidth comparison between the z196 and the z10 system is given in Table 1.

The z196 MCM hosts the six CPUs and the two SC chips. Both chips are manufactured in 45-nm technology and use 185.6-μm C4 pitch. The z196 MCM reuses the known glass–ceramic technology from its predecessor. It is a 96 mm × 96 mm carrier with 103 layers and a 7,356-pin land-grid array connector. The CP is designed for 250–300 W and the SC for 110-W power consumption. Power sorting of the chips allows operation of the MCM within the allowed total power of 1,800 W. The MCM power supply is broken into 17 main power supplies, individually controlled by the DCAs.

Power delivery

Vicor Factorized Power

In the prior zSeries system (z10 EC), a significant improvement in efficiency and packaging density was achieved using polyphase buck converters housed in the DCA, which directly plugged into the processor book as opposed to power supplies sharing a midplane board with the books, as was the case in systems that preceded z10 EC. However, power conversion to logic levels was still totally contained within the DCA, and the DCA output connector carried the high currents at the logic voltage levels. It was determined that the current feed through the DCA connector had reached a practical limit in z10 EC and that any increase in current would result in the DCA connector requiring additional power modules and pins, thereby exceeding reasonable space on the circuit board of the processor book. Furthermore, the high currents from the DCA to the load traveled the entire distance on the circuit board, resulting in a complex heavy copper structure in order to contain voltage drops and resistive losses.

A new approach was needed where the final level of voltage and current conversion is completed as close as possible to the point of use. The approach chosen for z196 is referred to as Vicor Factorized Power*. The factors in this system are bulk conversion, regulation, and voltage transformation using building blocks supplied by Vicor Corporation. With judicious packaging of the three factors, the following advantages are realized.

- Significant reduction in the size of the power supply connector and thus the power supply itself. This was an enabling element in successfully packaging the function required in the processor book.
- Simplification of the board structure, which reduced complexity and cost.
- Reduced resistive losses in the power distribution system, improving overall power conversion efficiency.
- Maintaining full redundancy and concurrency with no single points of fail or single points of repair in the power subsystem.

A simplified block diagram of this system illustrating one voltage boundary is shown in Figure 2(a).

The constituent components that implement the aforementioned factors are the bulk converter module (BCM), preregulator module (PRM), and voltage transformation module (VTM). These are packaged in hermetic “bricks” of identical size (1.28 inches × 0.87 inches × 0.265 inches), which are designed for surface mounting on a PC board using a J-lead connector. The BCMs and PRMs are packaged within the DCA, which is N + 1 redundant (N = 2) and concurrently replaceable. The VTM is packaged on small daughter cards called gearboxes, which are positioned very close to the point of load. The gearboxes are not field replaceable and therefore must be designed to fail in place. N + 2 redundancy is used, and no action is taken if a VTM fails on a gearbox.

Table 1  Bandwidth comparison between z10 and z196.

<table>
<thead>
<tr>
<th>Bus bandwidths per node</th>
<th>z10</th>
<th>z196</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
<td>4.4 GHz</td>
<td>5.2 GHz</td>
</tr>
<tr>
<td>No. of cores</td>
<td>80</td>
<td>96</td>
</tr>
<tr>
<td>Processor to cache/processor</td>
<td>470 GB/s</td>
<td>500 GB/s</td>
</tr>
<tr>
<td>Memory</td>
<td>100 GB/s</td>
<td>170 GB/s</td>
</tr>
<tr>
<td>Memory capacity</td>
<td>380 GB</td>
<td>770 GB</td>
</tr>
<tr>
<td>I/O hub</td>
<td>140 GB/s</td>
<td>170 GB/s</td>
</tr>
<tr>
<td>Fabric</td>
<td>140 GB/s</td>
<td>250 GB/s</td>
</tr>
</tbody>
</table>

The constituent components that implement the aforementioned factors are the bulk converter module (BCM), preregulator module (PRM), and voltage transformation module (VTM). These are packaged in hermetic “bricks” of identical size (1.28 inches × 0.87 inches × 0.265 inches), which are designed for surface mounting on a PC board using a J-lead connector. The BCMs and PRMs are packaged within the DCA, which is N + 1 redundant (N = 2) and concurrently replaceable. The VTM is packaged on small daughter cards called gearboxes, which are positioned very close to the point of load. The gearboxes are not field replaceable and therefore must be designed to fail in place. N + 2 redundancy is used, and no action is taken if a VTM fails on a gearbox.

A simplified block diagram of this system illustrating one voltage boundary is shown in Figure 2(a).

The constituent components that implement the aforementioned factors are the bulk converter module (BCM), preregulator module (PRM), and voltage transformation module (VTM). These are packaged in hermetic “bricks” of identical size (1.28 inches × 0.87 inches × 0.265 inches), which are designed for surface mounting on a PC board using a J-lead connector. The BCMs and PRMs are packaged within the DCA, which is N + 1 redundant (N = 2) and concurrently replaceable. The VTM is packaged on small daughter cards called gearboxes, which are positioned very close to the point of load. The gearboxes are not field replaceable and therefore must be designed to fail in place. N + 2 redundancy is used, and no action is taken if a VTM fails on a gearbox.

A simplified block diagram of this system illustrating one voltage boundary is shown in Figure 2(a).

The constituent components that implement the aforementioned factors are the bulk converter module (BCM), preregulator module (PRM), and voltage transformation module (VTM). These are packaged in hermetic “bricks” of identical size (1.28 inches × 0.87 inches × 0.265 inches), which are designed for surface mounting on a PC board using a J-lead connector. The BCMs and PRMs are packaged within the DCA, which is N + 1 redundant (N = 2) and concurrently replaceable. The VTM is packaged on small daughter cards called gearboxes, which are positioned very close to the point of load. The gearboxes are not field replaceable and therefore must be designed to fail in place. N + 2 redundancy is used, and no action is taken if a VTM fails on a gearbo
Figure 2

Block diagrams of IBM z196. (a) Simplified diagram of the Factorized Power supply design. The diagram shows only a single voltage domain for illustration. The components in this system are BCM, PRM, and VTM. (b) Block diagram of a gearbox showing four redundant VTMs. The redundancy level is $n+2$, i.e., two VTMs can be taken offline without affecting system performance. (c) Block diagram showing the I2C data communications implemented to control the Factorized Power supply design circuits (note: only VCL0 is shown as an example, and DCA3 does not communicate with the VTMs). (DCA: distributed converted assembly; DIMM: dual inline memory module; ECF: external timing reference/CEC/FSP; FSP: flexible support processor; HCA: host channel adapter; MCM: multichip module; OSC: oscillator.)
A second failure on the same gearbox will result in a scheduled replacement of the processor book.

**Capacities and efficiency**

Each BCM converts 350-V digital current (Vdc) to 45 Vdc and has a capacity of 325 W. Sufficient BCMs are packaged within the DCA to supply more than 4 kW per processor book. PRMs provide a regulated voltage up to 55 V to satisfy the load voltage requirement and can supply up to 400 W. The processor book has voltage levels with power up to 625 W, which requires three PRMs for \( N + 1 \) redundancy. VTM have a transformation ratio of 32:1 and can supply steady-state current up to 115 A. Gearboxes with four and six VTMs are used in the processor book and supply up to 230 and 460 A, respectively, with \( N + 2 \) redundancy. Each factor is 96% or 97% efficient, and when resistive losses in the power distribution (connectors, voltage planes, etc.) are included, an efficiency value of 80% is achieved for voltage levels in the range of 1.1 V. Eleven unique voltage levels are supplied by Factorized Power in the z196 processor book.

**Gearbox**

As already stated, the gearbox containing the final voltage conversion and current multiplication is not replaceable in the field and must be designed to fail in place. This imposes a number of requirements on the design.

- A failing VTM must safely fail, i.e., it must be instantly and completely isolated from the circuit.
- The occurrence of a failure must be accurately reported to the control program so that a second failure can trigger the correct service action.
- The parameters at the VTM and load must be continuously monitored for proper error detection and fault isolation. Load voltage, load current, and VTM temperature are the parameters monitored for this purpose.

An application-specific integrated circuit (ASIC) provides the above functions. Each redundant circuit on the gearbox is composed of a VTM and an associated ASIC. The block diagram in **Figure 2(b)** shows the “four-up” gearbox. This gearbox with four VTMs is \( N + 2 \) redundant and is designed to supply 230 A in \( N \) mode. The ASIC monitors input current and output voltage and will engage the field-effect-transistor protection devices if these parameters exceed design limits. The transformation module pin monitors internal VTM temperatures and serves as the “VTM good” signal back to the power control program. This information is transmitted using the inter-integrated circuit (I2C) serial interface, which provides diagnostics and tracks VTM failures for the system. If a second failure occurs on the same gearbox, the system will signal a repair action to be scheduled.

**Diagnostic data communications**

Eleven voltage levels are implemented in Factorized Power, with a total of 48 VTMs in each PU book, all of which must communicate status back to the power control program. An I2C multiplexor is used to condense all of the serial interfaces back to the DCA where the controls reside. Because the multiplexor is not field replaceable, this function must be also designed to fail in place. Additionally, because a failure of the multiplexor will not result in a PU book failure, a duplicate multiplexor with the proper synchronization circuitry provides the necessary redundancy. The data communications system is shown in **Figure 2(c)**.

Loss of communication from any gearbox to both DCAs will trigger a repair. The dual I2C buses from each multiplexor and watchdog signals between the multiplexors serve to eliminate any metastability that might arise from the DCAs that are asynchronously operating.

**Packaging Factorized Power delivery**

Supporting the z196 CP with four cores results in a challenge for the power delivery system. In order to optimize yield, performance, and power consumption, individually controllable voltage domains are required for each processor.

This drove an increase in system voltage domains from the z10 EC. Each z196 CP requires two individual voltage domains for the processor circuitry, two I/O voltage domains, and one global standby domain. In total, the z196 CPs, SCs, DDR3 DIMMs, I/O adapters, and the system control function require 22 individually controllable system voltage domains. The MCM alone requires 17 voltage domains, as compared with 14 domains in the previous system.

One I/O voltage domain is a common boundary for the CP and SC I/Os supporting the on-MCM and node-to-node interfaces. The second I/O voltage domain supports the memory interface. A node-wide voltage boundary is used for several ASIC chips such as the InfiniBand** or PCI2 hub chips.

In addition to the increase in voltage domains, the processor chipset power also increased as a result of the added sixth CP chip. The nominal z196 CP is rated at 270 W, and the SC is rated at 170 W. This results in a total MCM power value of 2,000 W, which reached new dimensions of power densities per unit area. The mixture of the amount of voltage domains and the power densities, particularly in the MCM area, turned out to be a challenge from a power delivery point of view and drove new packaging solutions.

The PU board is populated with the MCM, 30 DIMMs, and 8 I/O cards, and current is supplied by the three DCAs directly plugged into a PU board. The 22 individually controllable voltage domains are distributed from the DCA connectors. The 11 gearboxes are placed as close as possible to the load circuits. For the high current path, only local voltage islands on the PU board are required to close the
power delivery path between the gearboxes and the circuits.

The three DCAs are connected to the bulk power supply by two redundant connections located on a small high-voltage card. As a consequence, the 350-Vdc high-voltage distribution is restricted to this small card. The complete system package is able to support a current capability of 2,600 A per PU book.

When dealing with such currents, a detailed system dc analysis already in the concept design phase is required to avoid later design issues. The system dc analysis methodology must be capable of analyzing the entire power distribution system. It is also very important to perform sensitivity studies in order to select adequate technology for connectors, printed circuit boards (PCBs), and other components.

The system dc analysis objectives are as follows:

- Define power supply sense point locations.
- Define the board cross-sectional power layer assignment for voltage distribution.
- Verify that the voltage specification of each circuit in the system design will be met.
- Verify that the current ratings of the chosen connectors will not be exceeded.
- Reduce current density to avoid local heating in the PCB to maintain reliability.
- Verify and reduce the package power dissipation to achieve power efficiency goals.

In the z10 EC system, the package power dissipation became a significant fraction of the actual circuit power and could not be neglected in the overall system power efficiency. Using the identical power delivery concept for the z196 would have resulted in more than 600 W of copper losses. Because of the Factorized Power approach, the package power dissipation is now significantly reduced to approximately 400 W, with roughly one-third being dissipated in the copper structures of the package and two-thirds being dissipated in the power conversion circuits. This demonstrates the effectiveness of the new implemented power delivery concept.

Factorized Power is a means of power conversion that maximizes power conversion efficiency, minimizes PCB complexity, and maintains a power subsystem with continuous availability. Centralizing all power conversion in the DCAs would compromise packaging and power efficiency. Placing the complete power conversion loop out on the processor book would make the elimination of single points of fail and/or single points of repair a practical impossibility.

Voltage stabilization

For the first time in zSeries server development, the IBM System z196 processor chips use large on-chip embedded dynamic random access memory (eDRAM) arrays as L3 (on CP) and L4 (on SC) caches, which are manufactured using deep-trench technology. With the use of this technology, new on-chip capacitor cell types are available, which utilize the vastly increased capacitance density of a deep-trench structure in comparison to a planar design of the same size. This allows the placement of sizable amounts of additional capacitance on the chip die, which can be utilized to minimize on-chip power noise.

On-chip power noise is due to current variations caused by sudden changes in the activity of the chip logic. In extreme cases, for example, a clock stop event, the change in current has been observed to be more than 50 A per CP within a few processor cycles. Since the package inductance prevents an instantaneous change in the incoming current, the additional current within the first nanoseconds after such an event needs to be supplied by the chips’ capacitor cells. The higher their capacitance, the less the chip voltage drops as a result of a suddenly increasing chip current demand.

Figure 3(a) shows the results from power noise simulations for two IBM System z processor chips. The gray curve corresponds to the voltage experienced by an IBM z10 EC generation processor chip with an on-chip capacitance value of approximately 400 nF. The blue curve is the simulated voltage response of an IBM z196 generation processor with roughly 15-μF on-chip capacitance on this voltage rail. In both simulations, an instantaneous 30% increase in current demand on the chip main voltage supply is simulated.

The z10 CP chip exhibits a sharp voltage drop immediately after the current step, resulting in almost 180-mV voltage compression roughly 4 ns after the chip current is increased. This feature is usually called “first droop,” and it arises from the interaction between the fairly small on-chip capacitance and the much larger capacitance of the discrete module capacitors located at the chip perimeter [7]. There is a second shallower minimum in chip voltage after approximately 40 ns (often called “second droop”), which arises from the interaction between module capacitors and the capacitor groups on the PU card around the MCM. In contrast, the z196 processor chip shows a much slower voltage change following the current step. In particular, the first droop has completely disappeared, and it takes almost 80 ns for the full voltage compression of roughly 60 mV to evolve.

The simulations indicate that, for a similar current transient, the new CP chip will exhibit much less power noise. In particular, the chip’s minimum voltage is 60 mV below the chip nominal voltage. Since the maximum operating frequency of a chip is, among other factors, given by the minimal chip voltage, the increased robustness against power noise is one of the factors that allowed us to raise the z196 chip frequency with respect to its predecessor while maintaining power efficiency.
Figure 3

(a) Simulated voltage drop after an identical instantaneous increase in chip current for (gray) a z10 CP chip and (blue and green) a z196 CP chip. The time window of the simulation shown here (x-axis) is 100 ns. (b) Voltage rise (in arbitrary units) observed after a clock-stop event for (lower graph, yellow curve) a z10 CP and (upper graph, orange curve) a z196 CP chip. The time window (x-axis) of the oscilloscope traces shown here is 500 ns. (c) Oscilloscope trace of the array voltage supply for a z196 CP chip of the (top curve) first generation (DD10) and the (bottom curve) second generation (DD20). The oscillation observed in both traces is due to the DCA voltage regulation. The spikes, which are predominantly present in the DD10 CP chip, are due to the L3 eDRAM refresh operation. In the DD20 chip hardware, these spikes are much less pronounced because of an increase in on-chip decoupling.
The simulation results are confirmed by measurements of voltage overshoots following a clock stop event. Figure 3(b) displays the measured on-chip voltage over a period of 500 ns after a clock stop command was issued. The abrupt stop of the clock tree activity generated current steps of 43 and 63 A in the z10 EC and the z196 processor chip, respectively. The chip voltage rise is much faster in the previous processor generation than in the z196 chip generation.

Another example of the effectiveness of deep-trench capacitors can be observed when comparing the power noise visible on the array voltage supply for different chip releases of the z196 processor. In Figure 3(c) (top), we show the measured chip array voltage as a function of time for the first-generation (DD10) z196 processor chip. On top of an oscillatory voltage fluctuation, sharp voltage spikes can be observed, which correspond to voltage drops of 75 mV. The origin of those spikes was determined to be related to the L3 cache refresh operation, which creates a regular 8-A current burst on the array voltage rail that occurs every 175 ns and persists for 8 ns. The 75-mV power noise exceeded initial design specifications; thus, more on-chip capacitance was placed on the array voltage rail of the second-generation (DD20) z196 processor chip. Figure 3(c) (bottom) displays the array voltage measured on such a second-generation CP chip where, for identical refresh conditions, the corresponding power noise is dramatically reduced to less than 20 mV/refresh event. As a result, the DD20 CP chip now safely meets power noise targets, and the on-chip eDRAM arrays can be operated within projected voltage and frequency targets.

With the availability of deep-trench decoupling cells for on-chip power noise decoupling, it is now possible to keep chip voltage within the voltage tolerance limits despite even higher chip power values, faster chip current transients, and lower operating voltages. Furthermore, this technology provides potential for further cost reduction because of the reduced necessity of module and board capacitors. This will be an important factor in the power design of future systems.

**Power supply impedance measurements**

With an electrical power value of more than 3 kW/PU book, the z196 server power supply currents are in the kiloamperes range and the dynamic variance of these currents can reach 600 A or more. The main power supply voltages in the z196 server are all near 1 V, and system functionality requires ac stabilizing these voltages at approximately ±5% around their nominal levels. Therefore, power supply distribution and stability are critical items that need to fulfill ambitious requirements. The power domains need to be analyzed, and the voltage tolerance limits need to be maintained over the dynamic range of the current demand. In order to provide and test the desired functionality, there is a need for a verification methodology that is able to check power supply function and robustness during early silicon tests, first module tests, and final compound operation in system environment.

Traditionally, power supply voltage integrity is measured in the time domain at the load while executing selected power demand change conditions such as program task activity switch or clock stop. The observed local power supply voltage excursions depend on the specific operation power demand change and therefore do not reveal the underlying power supply distribution properties such as power supply impedance $Z(f)$.

The power supply impedance profile $Z(f)$ at the load location is a property of the system design and is useful for analyzing and predicting the voltage response of the system caused by a power demand change. However, the practical way to determine the impedance is to impress a known current and measure the voltage response.

**Evaluation problem**

To calculate the power supply impedance $Z(f)$ at the chip, both voltage and current at this location have to be determined with high precision. Ideally, those would be measured using a device that generates a sinusoidal load stimulus with tunable frequency, and the magnitude of the impedance is then simply determined by

$$|Z(f)| = \frac{|U(f)|}{|I(f)|}. \quad (1)$$

However, approaches that are commonly used for signal line impedance determination in the frequency range using a network analyzer are not practical in this case because the impedance of interest is the equivalent of a chip with thousands of leads and capacitive loads in parallel. Compared with a typical signal line impedance of approximately 50 Ω, the expected power supply impedance values are in the range of 10 mΩ or lower. Moreover, it is not practical to design a test chip to generate a predictable periodic current with tunable frequency. Alternatively, power supply impedance $Z(f)$ can be obtained for an arbitrary chip activity from accurate time-domain measurements of both chip voltage $U(t)$ and chip current $I(t)$ [8–10]. Measuring the voltage at an on-chip location in the time domain can be accomplished by probing the $V_{DD}$ and GND rails close to the location of interest and recording the differential on-chip power supply voltage in the time domain

$$U(t) = V_{DD}(t) - GND(t). \quad (2)$$

The magnitude of the local power supply voltage spectrum $U(f)$ can be then calculated by a fast Fourier transformation (FFT)

$$|U(f)| = |\text{FFT}[U(t)]|. \quad (3)$$
The difficulty in this approach is the accurate determination of chip current $I(t)$, which is composed of many components. Digital currents are supplied from the primary power supply source in the DCA; low-frequency supply currents are supplied by charge stored in on-board capacitors; on-module capacitors are engaged for mid-frequency current demand; and finally, high-frequency supply currents are supported by the deep-trench on-chip decap cells close to the load location. There is no method available to measure these currents with the necessary precision and time resolution to transform into the frequency domain and then use this for the calculation of the power supply impedance in Eqn. (1).

The key idea is to overcome the lack of reliably measured power supply current data and to replace the current measurement by a current synthesis of a well-defined and known power demand change condition. A simple activity procedure is best suited to be correctly synthesized, and its implementation must be possible without having an impact on system functions. One design feature that pervades all on system functions. One design feature that pervades all digital systems is the treelike clock distribution. In general, the clock tree represents a significant load for the power supply. Switching clock trees periodically ON and OFF provides a simple, well-defined, and precise activity procedure that generates a transient current. In addition to the clock tree ON/OFF modulation ability, there is no other additional means that can be used.

The spectral content of such a clock modulation activity in the range between the ON/OFF modulation and clock base frequencies is completely defined by the modulation timing and its harmonics. The spectral content of the clock tree current, on the other hand, resides entirely above the clock base frequency and, therefore, does not interfere with spectrum portions below the clock base frequency.

Using clock tree modulation as a stimulus thus opens the frequency band from the modulation frequency up to clock tree frequency for power supply impedance evaluation.

The new methodology is demonstrated for a newly developed zSeries I/O ASIC chip. This chip design is manufactured in a technology without deep-trench on-chip capacitor cells. Figure 4(a) shows an oscilloscope screen shot of the local on-chip power supply voltage and GND behavior during one cycle of switching clock tree operation ON and OFF. There is a very sharp $V_{DD}$ collapse of 34 mV at the very beginning of the clock tree ON time and, simultaneously, a sharp GND bounce of 14 mV. After 100 μs, the mid-frequency/low-frequency voltage excursions have settled. The power supply voltage disturbance changes its sign when the clock tree switches OFF. The power supply impedance evaluation needs the true on-chip differential voltage $V_{DD}$-GND versus time. In this example, the measurements have been averaged 7,864 times to increase numeric voltage resolution.

Figure 4(b) shows a small portion of the sampled on-chip differential power supply voltage $V_{DD}$-GND versus time while clock tree toggling starts. Digitized voltage samples are taken at every 1 ns.

Figure 4(c) (upper graph) shows the results from a chip voltage measurement where the clock tree is switched ON and OFF at a duration period of 1 ms per event (2-ms period). The clock tree itself is toggling at 800 MHz. Measuring the power supply impedance up to a frequency of 500 MHz requires the voltage sampling rate to be at least 1 GHz to avoid aliasing effects (Nyquist theorem), which results in 2,000,000 sampling points for a 2-ms voltage trace. Modern real-time oscilloscopes provide this capability.

Figure 4(c) (middle graph) displays the corresponding synthesized chip clock current. The current scale has been determined while keeping the clock tree statically ON and OFF.

Figure 4(c) (lower graph) displays the resulting power supply impedance at the chip location in a frequency range between 500 Hz and 500 MHz. Since the clock tree was operated at 800 MHz, there is basically no influence of the clock tree spectral content in the displayed frequency interval. Unfortunately, there are some small clock tree branches that are divided down to lower frequencies and cannot be switched off. The power demand of these is not covered by the current synthesis, which results in artificial spikes in the calculated $|Z(f)|$ graph at 100, 200, 300, and 400 MHz. $|Z(f)|$ clearly shows quantitatively that the dc regulation loop starts to drop off at approximately 1 kHz. Low-frequency decoupling resonates at about 15 kHz.

There are two significant mid-frequency resonances with high-impedance values at 7 and 90 MHz.

Taking into account the high harmonic numbers to be processed and their associated small numeric values, the signal-to-noise ratio of the primary voltage measurement can be improved by averaging. Averaging more than 4,000 single ON/OFF events will boost the raw resolution of true 6 bits (out of an 8-bit analog/digital Flash converter) up to 12 bits, which leads to a useful evaluation range of about four frequency decades.

**Differential memory interface**

The z196 system uses a high-bandwidth I/O link to transfer data between the processor and memory through a memory buffer chip. The link was specifically designed for this type of application and consists of 15 or 22 differential data lanes and 1 clock lane. The interface is designed as a source-synchronous link, which means that the bus clock and data are driven from the same source clock in the driving chip and the bus clock is used to latch the data in the receiving chip [11].

The link consists of drive logic, transmitters, wiring channels, receivers, and receiver logic. The various components of the link were designed in concert to optimize power and performance for this type of application. This section of this paper describes the various pieces of the DMI with focus on the wiring channels, the logic used to verify
Figure 4

(a) Screenshot of the measured local on-chip (yellow) $V_{DD}$-voltage, (green) GND-voltage, and (white) their differential $V_{DD}$-GND during one cycle of switching clock tree toggling ON and OFF. (Two million samples have been taken and averaged 7,864 times.) (b) Enlarged detail of the sampled differential on-chip power supply voltage $V_{DD}$-GND while clock tree switches from OFF state (no clock activity) to ON state (clock tree is toggling). The maximum collapse is 48 mV, and the sampling rate is 1 GSa/s. (c) The upper graph shows a 2-ms period of measured and digitized time-domain supply voltage. The middle graph shows a 2-ms period of synthesized time-domain supply current. The lower graph shows derived power supply impedance magnitude versus frequency profile. (See text for further comments.)
Drive and receive logic
The drive and receive logic serializes and deserializes the data stream, buffers data to accommodate static and dynamic timing differences between individual lanes, initializes the interface, spares out bad data and clock lanes, dynamically centers the clock in the data eye, and provides signal-quality measurement capabilities.

A 4-bit-wide packet of data is delivered to each lane at a 1.2-GHz rate. The data is serialized four-to-one, and data is transmitted on the channel at 4.8 GHz. The data is deserialized in the receiver one-to-four, and a 4-bit-wide packet is delivered to the receiving chip core logic at a 1.2-GHz rate. Furthermore, the data is deskewed from lane to lane so that data delivered across the bus to the drive logic on one clock is delivered to the receive chip core logic on a single clock. Thus, at the chip core logic level, the interface looks like a wide data buffer with the output data being simply a delayed version of the input data.

The initialization process aligns the data and detects bad lanes that have a high error rate. It consists of a series of steps controlled by the “upstream” chip, i.e., either the processor or the memory buffer chip closest to the processor using a communication protocol across the interface. The steps select a lane to use as a clock, align the clock to the data, detect and spare out bad data lanes, align data across the lanes, and align the data on a nibble boundary as referenced above. Initialization is driven by hardware, with no software intervention required enabling high-speed dynamic reinitialization while the machine is running.

During initialization, a synchronization pattern is used to check each lane. Lanes that fail are spared out in hardware by shifting the data to unused lanes in the driver and the receiver. This sparing may be also controlled by registers in the design, enabling software to spare out lanes that are failing at a lower rate than the hardware test pattern can detect.

Beginning with initialization and continuing during operation, the logic keeps the clock centered in the data by using an edge-detect mechanism and a control loop to clock delay circuitry. Built into this logic is the capability to move the clock off center, enabling the diagnostic functions described later.

DMI channel topology and simulation
The IBM zSeries server z196 introduces with the DMI the benefits of differential data transmission. Among other benefits, the differential data transmission shows good common mode rejection, more immunity against noise and parasitic effects, and as a consequence, higher frequencies and data rates can be achieved. This part concentrates on the analysis of the electrical signal transmission of the channel, including the output and input stages by ignoring all additional logic.

The topology of the DMI channel consists of the glass–ceramic MCM, which is connected to the PCB with a land-grid array connector. The DIMM connector was chosen as a surface mounted version to minimize the parasitic effects due to vias connecting the connector and the signal lines inside the PCB. On the DIMM card, the channel is terminated by the receiving chip that is mounted on an organic package. Figure 5(b) shows the high-level topology of the DMI channel designed for system z196.

For the purpose of channel simulation, each individual part of the channel, such as MCM trace, transmission line, PCB via, and connector, is modeled in a first step by means of 2-D or 3-D full-wave solvers, resulting in separate representations of the nominal electrical characteristics of those components in S-parameter format. Manufacturing tolerances such as variations in PCB material characteristics such as dielectric constant, line spacing, ohmic resistance, and other deviations from the nominal process parameters usually lead to variations of card impedance and attenuation and may have significant effects on the electrical performance of the DMI channel. To represent these deviations properly within the simulation, each of the nominal PCB models is extended by four additional models, reflecting the worst case corner scenario by a combination of high and low impedance as well as high and low attenuation. This approach allows a parameter sweep during simulation across all impedance and attenuation corners that may occur in the hardware. A prerequisite for the usage of S-parameter models in a second step simulation is their quality with respect to passivity and causality. Therefore, all models have been individually checked against these requirements.

Within the simulation tool, these models are cascaded to build up the entire channel. An S-parameter representation of the output and input stages of the driver and receiver derived from their respective transistor-level representation is eventually connected with the channel cascade. For the system z196 design, the IBM proprietary1 simulation software high-speed SerDes/clock data recovery (HSSCDR) has been extensively employed for the first time to simulate high-speed nets such as the DMI in first- and second-level packaging.

The HSSCDR core offers a fast inverse discrete Fourier transform/convolution engine, providing deterministic and statistical analysis of the signal channel to be assessed, i.e., in particular impulse response, signal-to-noise ratio, and eye diagram with bit-error-rate curve. HSSCDR requires the channel elements to be represented by linear S-parameter models. It allows multigigabit channels such as the DMI to be simulated with several millions of random data

1 A high-speed SerDes/clock data recovery (HSSCDR) version with reduced command set is commercially available.
Figure 5

Differential memory interface (DMI). (a) Basic design showing a single lane. The z196 design uses up to 22 data lanes per clock lane. (b) DMI channel topology. (c) DMI channel eye plot generated by the HSSCDR tool. (d) Diagnostic latch in receiver giving the ability to measure error rate. The diagnostic sample offset is changed to sample diagnostic data at different horizontal points in the data eye. The data compare/counter is used to compare the diagnostic and data sample and accumulate miscompares. (e) Example of tub data output showing raw data and data centered and extrapolated. (LGA: land-grid array; MCM: multichip module; SMT: simultaneous multithreading.)
bit patterns, i.e., orders of magnitude faster than traditional time-domain solvers (SPICE derivatives).

In addition to the actual signal channel (victim), HSSCDR allows an unlimited number of aggressor channels to be simulated in parallel, providing far-end crosstalk (FEXT) and near-end crosstalk (NEXT) information essential for evaluation of noisy lines located close to the victim inside the packaging. Deterministic and statistical jitter can be added to emulate driver and receiver characteristics in a realistic way. For the purpose of data recovery, HSSCDR features a set of equalization tools [e.g., forward feedback equalization (FFE), continuous-time equalization, and decision-feedback equalization] with optionally self-optimized or fixed parameter settings.

With respect to the analysis of the simulation results, HSSCDR provides several output files comprising the channel transfer function, impulse response, and eye diagram, including statistical bit error rate among others. A predefined minimum value for the horizontal and vertical eye opening can be then compared to the simulated eye aperture based on the required receiver specifications. **Figure 5(c)** shows an example eye plot of the DMI simulation with HSSCDR.

The FFE feature has been extensively employed during DMI channel simulation in order to match the driver stage with the attenuation characteristics of the MCM and PCB wiring. The HSSCDR built-in optimization algorithm allows calculating the best FFE setting for a specific channel characteristic. This strategy has been applied across all corners of the channel. The FFE setting resulting in the worst case eye opening derived by simulation is applied to all corners of the channel to ensure that the deviation from the optimal setting for that channel does not affect the global worst case of the eye opening and to ensure that the interface specification for horizontal and vertical eye opening is met across all hardware corners.

The optimal FFE setting identified by the simulation is eventually loaded onto the server system during hardware characterization to initialize the driver stages with reasonable parameters. If this setting enables the DMI to run as reliably as expected, the parameters get stored in the machine as default. Only in the case that the simulated setting causes performance problems is a hardware internal algorithm activated to adjust the parameters accordingly.

Extensive model-to-hardware correlations have been performed after design completion to compare the predicted electrical performance with the actual hardware. The data measured stayed well within the projected boundaries, i.e., HSSCDR simulations provided increased product reliability at reduced development cost.

**Hardware measurements**
The hardware measurements relied on the ability to sample data at an offset, measuring an error rate at each offset.

**Figure 5(d)** shows a diagram of the receiver diagnostic hardware. In this particular application, the receiver design provided a diagnostic latch per lane intended to mimic the latch used in the functional data path. In addition, compare logic and counters were made available at the bus level together with controls to select a particular lane to examine.

Data was driven across the interface, and an error rate was measured across the interface at offsets from the center. The resulting data was plotted and extrapolated to compute a normalized eye width at $10^{-12}$ error rate. **Figure 5(e)** shows an example plot of data gathered using this diagnostic logic and technique. Using this data, we could verify tuning of the design, confirm that the hardware results were bounded by the simulation results, and ensure that signaling met the requirements for this application.

**Card and module electrical verification**
During the development of the z196 system, the electrical packaging integration team made the transition from a design verification methodology based on postroute simulation to one based on physical rules checking. The previous methodology relied on a proprietary simulator that yielded fast and accurate results for every high-speed net in the system [12]. This approach called for the extraction of I/O circuit models from transistor-level models and triangular impulse response waveforms from lossy interconnect models. The infrastructure associated with a parallel model library proved expensive to maintain over the long run, and the team took a new direction.

During the initial phases of architectural definition, the signal integrity team wrote a set of physical design rules for first- and second-level packaging that were consistent with the simulation and modeling assumptions used to define the design space and confirm system performance targets. The document containing these rules served as a framework for the new verification methodology that associated each physical design rule with a computer program from IBM’s “verification toolbox.” The new methodology heavily relies on the ability of engineers to thoroughly analyze the electrical limitations of silicon and translate them into a complete set of instructions for building a circuit board or chip module. There is no postroute simulation filter to catch permutations of routing that may have slipped through the preroute analysis net. Engineers must ensure that the physical design rules adequately cover the design space and then pass along those rules to the verification team for implementation into the checking process.

For many years, the Allegro* computer-aided design tool from Cadence Design Systems has handled the majority of physical design rule checking associated with high-speed processor, memory, and I/O nets. However, the high-density routing found in IBM servers often strains the limits of industry tools. In the z196 system, bandwidth...
requirements necessitate routing long wires beneath the 87 mm × 87 mm array of MCM pins, with a card thickness of more than 6 mm. These wires are susceptible to interlayer crosstalk through misaligned antipad holes caused by lateral shifting of layers during card manufacturing. To combat this problem, card physical designers carefully centered each wire segment between vias, both orthogonally and diagonally. The misalignment observed in a z196 node test card and the pattern implemented to minimize coupling between misaligned wires are displayed in Figure 6.

Since checking this unique wiring pattern was beyond the scope of available industry tools, IBM reused an existing proprietary tool designed to check organic modules and modified it to check card wiring. System Electrical Rule Checker (SERC) employs its own scripting language to enable rapid prototyping of new checks. It is not unusual for an engineer to have functional code two or three days after identifying a need for it. The quick-turn capability aids the physical design process in an environment where signal integrity analysis and rules generation often continue late into the design cycle. SERC also checks the proximity of a signal via to the nearest GND via and the spacing between vias that are part of a differential pair. The variety of possible checks is limited only by the syntactical features available in the scripting language.

In addition to single-card checking, SERC allows the user to define connections between cards and chip modules and between cards and backplanes. This feature enables the definition of a “system-level net,” i.e., an electrically continuous collection of module, card, and connector wires that connects chip pins located anywhere in the system.

The elastic interfaces that pass data between processor chips have requirements for skew between data bits within the same clock group and between data bits and their respective clock. SERC traces through each system-level net, adding up segment lengths as it goes and multiplying by the propagation delay per unit length to arrive at a total delay for that net. When it encounters a connector, it refers to a lookup table of delays versus pin number.

The ability to build a hierarchical model for each net in the system enables checking of another class of problems. For example, a chip pin may have an internal weak pull-up resistor that was undocumented in the component data sheet. A card logic designer who is unaware of the presence of the pull-up resistor on the chip may inadvertently place a weak pull-down resistor on the card. If these two resistors happen to be close to the same value, the receiver input may become biased near its threshold voltage. Using a system-level logical model for the net, SERC compares the I/O circuit characteristics of each chip on the net against each other and against the discrete components on the card to identify possible conflicts before releasing a card design to manufacturing. SERC also identifies other common logic entry errors such as conflicting drivers, floating inputs, and chips with different power supply voltages connected to the same net. These relatively simple failure mechanisms present an entirely different challenge than the high-speed interfaces that are dominated by wire losses. The number and variety of nets and lack of basic design data require a systematic approach that relies on a hierarchical model of every digital net.

**Manufacturing tests**

Before being shipped, each IBM z196 EC server component must pass a series of tests performed by manufacturing qualification testing. In these tests, the system is operated under conditions that include voltage, temperature, and clock speed settings outside of the regular system specifications. For the first time in zSeries qualification, Elastic Interface (EI) [2, 6] calibration data was collected in parallel to the
qualification tests. The collection of EI data has to satisfy several conditions.

1. It must not interfere with or add runtime to the existing manufacturing qualification testing.
2. The EI should be only exercised within valid corner (voltage, temperature, and frequency) settings and with all EI driver hardware components fully populated.
3. The data collected should provide meaningful information to track hardware performance so that shifts in process or other factors may be identified.

Collecting EI data from the machine under manufacturing test is done by a program that reads the interface status registers. The status registers are dynamically updated whenever the machine is running. Key data values are collected and stored at both the channel (bus) and lane (bit) levels.

The key data stored includes guard band current and historic minimum; select edge minimum and maximum; and learned target cycle (LTC) minimum, maximum, and target.

Guard band is a result of each receiver bit I/O circuit measuring the horizontal open eye of the data stream received. This value is dynamically updated and is used by the receiver data sample circuitry to optimize centering of the sample clock of the data. The value represents the number of delay steps, each approximately 15 ps that the data eye is open to both the left and right of the sample point within the data eye. Current guard band is the result under the machine’s current conditions. Historic minimum guard band is the lowest recorded result since the last initialization, normally, since initial machine load.

The guard band value may be impacted by many parameters such as data frequency; process shifts during manufacturing or burn in stressing; signal integrity impacts such as crosstalk, power collapse, or chip switching events; different data patterns, clock, or phase-locked loop (PLL) jitter; and poor interface connections.

Select edge is a counter in the interface of how many unit intervals (or data bit times) the bit is held to be aligned with the latest bit of the bus. Guidelines require that this value is to be 0 or 1 for any bit. Values greater than 1 indicate that the skew within the bus has exceeded the design specifications resulting in additional latency of the bus.

The LTC is a measure for the data latency from sending to receiving chip. This delay includes the package wiring and connector delays, as well as the driver and receiver circuit delays. Temperature and voltage drift, clock/PLL jitter, and clock switchover are among the primary impact factors of the resulting range between the LTC minimum and maximum.

Additional information collected with each data set allows the identification of the corresponding system under test, including all its parts such as PU nodes, I/O cards, buses, and single bits, as well as the specific running conditions applied (temperature, voltage stress, workload applied, and stage of manufacturing test). This allows the tracking of the results back to individual chips and wafers if necessary.

**Data evaluation**

Based on these measurements, the EI interface data from the qualification test is written out into Extensible Markup Language (XML)-based file reports, together with other test relevant parameters and settings. The XML files are imported into an IBM-hosted DB2* database that contains test data from earlier test levels such as MCM or wafer test level. This enables correlating of the manufacturing test data with earlier test results using standard Structured Query Language (SQL) queries.

A back-end process allows for transforming table-based query results into graphical charts. A sample output graphic is shown in Figure 7. Retrieving the data on a regular and automated basis allows for timely feedback into test and design. Using DB2 and SQL helps identify shifts in the manufacturing process over time and helps in correlating to environmental conditions or even to component and assembly changes of cards and boards involved.

**Summary**

The packaging solutions presented in this paper have demonstrated how the z196 server was designed as a successful follow-on for the z10 server, addressing the need for increased platform performance as a consequence of the continuing growth of workloads. New packaging concepts were introduced to deal with the challenges associated with higher processor speed, higher power consumption, increased memory bandwidth, cooling, and reliability aspects.
The power design of the z196 server takes advantage of a new Factorized Power delivery approach, which strongly reduces package power losses and therefore allows a further increase in the maximum electrical currents delivered to the main processor chips. The use of large amounts of on-chip decoupling cells keeps electrical power noise in narrow limits and contributes to power savings by allowing the system to operate at lower voltages. The new DMI increases memory bandwidth, and redundant DIMMs offer a new level of data reliability.

System z196 achieves a balanced system design, where the high integration level of the IBM processor and cache chips is matched by an optimized system packaging solution that allows for maximum system performance.

*Trademark, service mark, or registered trademark of International Business Machines Corporation in the United States, other countries, or both.

**Trademark, service mark, or registered trademark of PCI-SIG, Amphenol Corporation, Vicor Corporation, InfiniBand Trade Association, or Cadence Design Systems in the United States, other countries, or both.

References


Received January 15, 2011; accepted for publication August 4, 2011.

Thomas Strach IBM Systems and Technology Group, IBM Deutschland Research and Development GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (strach@de.ibm.com). Dr. Strach received his M.S. degree from McMaster University, Hamilton, Canada, in 1992 and his Ph.D. degree in physics from the University of Stuttgart, Germany, in 1997. He joined IBM in 1997 and then joined the IBM Systems and Technology Group in 2004 as a staff engineer. His main responsibility is the power decoupling design of IBM high-end servers. His current fields of interest include the simulation and experimental verification of on-chip power noise propagation as well as the design of guidelines for the effective placement of on-chip decoupling cells.

Frank Bosco IBM Systems and Technology Group, Poughkeepsie, NY 12603 USA (bosco@us.ibm.com). Mr. Bosco received the B.S.E.E. degree from Manhattan College, Riverdale, NY, in 1964 and an M.S.E.E. degree from Syracuse University, Syracuse, NY, in 1975. He joined IBM Systems Development Division in Poughkeepsie in 1964 where he worked on the early development of monolithic integrated circuits. In 1973, he worked on a team which was attempting to implement a full wafer memory package and following this, he worked on circuit designs for the first implementations of cryptographic algorithms. He worked in power supply development from 1974 to 1978. From 1979 to 1981, Mr. Bosco worked on the hardware and software algorithms for automating personal identity verification. He joined the memory area in Kingston in 1981 and managed the group which developed the first L4 storage subsystem. He joined the Corporate Development staff in 1985 where he led a task force that resulted in the introduction of built-in self-test into complementary metal-oxide semiconductor (CMOS) logic arrays. Mr. Bosco rejoined the Kingston power group in 1988 and, since 1993, has been involved in the development of the power, packaging and cooling subsystems for zSeries CMOS mainframes. He is currently Lead Architect for the power, packaging, and cooling subsystem for future mainframe servers.

Kenneth L. Christian IBM Systems and Technology Group, Poughkeepsie, NY 12601 USA (kchrist@us.ibm.com). Mr. Christian received the B.S. degree from Southern Illinois University, Carbondale. 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, received an IBM Outstanding Technical Achievement award in 1999 for development of the NetRules tool for system and package design, and received an IBM Employee Excellence award in 2007 for ensuring the highest quality deliverable to customers.

Kevin R. Covi IBM Systems and Technology Group, Poughkeepsie, NY 12601 USA (kcovi@us.ibm.com). Mr. Covi is a Senior Technical Staff Member with the IBM Systems and Technology Group in Poughkeepsie, NY. He joined IBM in 1978 after receiving
the B.S.E.E. degree from Princeton University, and he also earned an M.S.E.E. degree from Syracuse University in 1988. Since 1984, Mr. Covi has been involved with the development of fault-tolerant power systems for high-end mainframes and servers, during which time he has been granted 16 U.S. patents. He has received several IBM Outstanding Technical Achievement awards as well as an IBM Corporate Award for his work on the z900 bulk power regulator.

His most recent assignment was as lead power architect for IBM’s POWER® family of high-end servers, and he is continuing in that role for POWER8®.

Martin Eckert  IBM Systems and Technology Group, IBM Deutschland Research and Development GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (eckertm@de.ibm.com). Mr. Eckert is a Hardware Development Engineer. He received his Dipl.-Ing. (FH)/bachelor’s degree from University of Applied Sciences, Esslingen, Germany. His main responsibility is the test and characterization of input/output application-specific integrated circuit chips in IBM high-end servers and memory arrays for IBM high-end server chips. His current field of interest is the data evaluation and visualization of test and characterization results from laboratory environments and also from manufacturing test areas.

Martin Eckert  IBM Systems and Technology Group, IBM Deutschland Research and Development GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (eckertm@de.ibm.com). Mr. Eckert is a Hardware Development Engineer. He received his Dipl.-Ing. (FH)/bachelor’s degree from University of Applied Sciences, Esslingen, Germany. His main responsibility is the test and characterization of input/output application-specific integrated circuit chips in IBM high-end servers and memory arrays for IBM high-end server chips. His current field of interest is the data evaluation and visualization of test and characterization results from laboratory environments and also from manufacturing test areas.

Gregory R. Edlund  IBM Systems and Technology Group, Rochester, MN 55901 USA (gedlund@us.ibm.com). Mr. Edlund is a Senior Engineer whose primary responsibility is electrical design verification of IBM servers. In 1988, he earned a B.S. degree in physics from the Minnesota Institute of Technology. He has also worked for Digital Equipment Corporation, Cray Research, Inc., and Supercomputer Systems, Inc. His book, Timing Analysis and Simulation for Signal Integrity Engineers, reflects a long-term interest in predicting and measuring operating margins.

Roland Frech  IBM Systems and Technology Group, IBM Deutschland Research and Development GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany. Dr. Frech received his Ph.D. degree in physics from the University of Stuttgart, Stuttgart, Germany, in 1983. He joined IBM in 1984 and is a Development Engineer in the VLSI Packaging Development department. During recent years, he has been involved in signal and power supply integrity of very-large-scale-integration (VLSI) chips and packaging.

Hubert Harrer  IBM Systems and Technology Group, IBM Deutschland Research and Development GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (hharrer@de.ibm.com). Dr. Harrer is a Senior Technical Staff Member. He received his Dipl.-Ing. and Ph.D. degrees in electrical engineering, both from the Technical University of Munich. He has been the technical lead for zSeries central electronic packaging designs since 2001. His technical interests focus on packaging technology, high-frequency designs, and electrical analysis for first- and second-level packaging. He has published multiple papers and holds 12 patents in the area of packaging.

Andreas Huber  IBM Systems and Technology Group, IBM Deutschland Research and Development 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 from the Technical University of Karlsruhe, Germany, in 2001. Since 1999, he has worked in the Electronic Packaging department at IBM Boeblingen, Germany, and IBM Austin, Texas, working on first- and second-level packaging. Dr. Huber’s field of interest is power distribution and power integrity. He holds several patents.

Dierk Kaller  IBM Systems and Technology Group, IBM Deutschland Research and Development GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (dkaller@de.ibm.com). Mr. Kaller is an Advisory Engineer working in the Systems and Technology Group. He received his diploma in electrical engineering in 1995 from the University of Hannover, Germany, joining IBM that same year. Since then, he has worked on various System/390® packaging designs in packaging development. He is currently responsible for the signal integrity of the first- and second-level packaging of the zSeries and pSeries® high-end servers.

Martin Kindscher  IBM Systems and Technology Group, IBM Deutschland Research and Development GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (kindsche@de.ibm.com). Dr. Kindscher is an Advisory Engineer. He received his diploma and Ph.D. degree in electrical engineering from the Technical University Berlin, Germany. In his current role, he is responsible for the signal integrity in ceramic and organic packaging for IBM server systems. His focus is on analysis and simulation of high-speed bus nets.

Andrew Z. Muszynski  IBM Systems and Technology Group, Poughkeepsie, NY 12601 USA (amuszyns@us.ibm.com). Mr. Muszynski received his B.S.E.E. degree from the City College of New York in 1970 and his M.S. degree from Massachusetts Institute of Technology, Cambridge, in 1974. He joined IBM in 1974 in East Fishkill, New York, focusing on package modeling and signal characterization. His present work is focused on bring-up and electrical characterization of high-end systems.

Gary A. Peterson  IBM Systems and Technology Group, Rochester, MN 55901 USA (garyp@us.ibm.com). Mr. Peterson is a Senior Engineer. He received his B.S.E.E. degree from the University of Nebraska. He has developed methods to verify signal integrity and system timing in high-speed, large-scale systems through the design and hardware verification process. His current focus is efficient interface data gathering and analysis in laboratory, manufacturing, and customer environments.

Claudio Siviero  IBM Systems and Technology Group, IBM Deutschland Research and Development GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (civiero@de.ibm.com). Dr. Siviero received his diploma in electrical engineering in 2003 and his Ph.D. degree in 2007 from the Politecnico di Torino, Italy. His research activities were in the field of electromagnetic compatibility, where he worked on the macro-modeling of nonlinear circuit elements with specific application to the behavioral characterization of digital integrated circuits for the assessment of signal integrity and electromagnetic compatibility effects. In 2008, he joined IBM Development in Boeblingen, Germany. Dr. Siviero’s current focus is on electrical packaging design with respect to the signal integrity assessment for the pSeries and zSeries.

Jochen Supper  IBM Systems and Technology Group, IBM Deutschland Research and Development GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (juspper@de.ibm.com). Mr. Supper is a Development Engineer in the VLSI Packaging Development department. He received his M.S. degree in electrical engineering from the Fachhochschule Furtwangen, Germany, in 1997 and joined IBM the same year. He has recently been involved in ASIC chip design and power supply integrity of VLSI chips and packaging.

Otto Andreas Torreite  IBM Systems and Technology Group, IBM Deutschland Research and Development GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (torreite@de.ibm.com). Mr. Torreite received the B.S. degree in technical informatics from the University of applied sciences of Esslingen, Germany, in 1987. He joined IBM in 1987 to work on very-large-scale integration (VLSI) test equipment automation, inline test and then test and characterization of actual processor and input/output test chips and modules. His current interests focus on high-frequency testing and future characterization and diagnostic methodologies.
Dr. Winkel is currently an Advisory Engineer and the technical lead for the z10 midrange systems. He received both his diploma and his Ph.D. degree in electrical engineering from the University of Hannover, Germany. His current focus is the electrical packaging design with respect to signal integrity, power integrity, and system aspects. He has authored or coauthored more than 20 conference and journal papers and holds 10 patents. He is the general co-chairman of the IEEE European Systems Packaging Workshop 2009 and a technical program committee member of the Electrical Performance on Electronic Packaging and the Signal Propagation on Interconnects Workshops.