Power Grid Physics and Implications for CAD

Sanjay Pant
University of Michigan, Ann Arbor

David Blaauw
University of Michigan, Ann Arbor

Eli Chiprout
Intel

Editor's note:
This article describes a full-die dynamic model of an Intel Pentium IV microprocessor design. The authors show that transient supply noise is sensitive to nonuniform decoupling-capacitor distribution, and that supply-drop locality is a tight function of frequency and package-die resonance, leading to significant localized resonant effects.
—Kenneth M. Butler, Texas Instruments

SHRINKING DEVICE DIMENSIONS, faster switching frequencies, and increasing power consumption in deep-submicron technologies cause large currents to flow in power distribution networks (PDNs). These rapid transient currents flow through the transistors onto the power grid, charging and discharging various capacitances, then flow onto the package through the C4 (controlled collapse chip connection) bumps, and eventually make their way to the voltage regulator module (VRM). This flow of currents causes spatial and temporal voltage variation in the PDN, degrading circuit performance and reliability. Power supply verification is, therefore, a critical concern in high-performance designs.

However, PDN modeling and verification is complicated due to the presence of decoupling capacitors (decaps), on-die inductance, various resonance effects, and simply the enormous size of the PDN. The following questions must be resolved for accurate supply-drop analysis:

- How significant is the impact of on-die inductance?
- How localized are the currents as they flow outward from a device?
- Does the decap charge respond locally or globally?
- What is the impact of C4s and package inductance?
- Do resonance effects occur, and if so, how?

The answers to these questions are critical to addressing the types of models and CAD algorithms required to deal with the PDN verification and chip-package codesign. For example, if supply-drop effects are localized, then it's possible to considerably simplify the analysis by verifying several partitions of the PDN in parallel, as Chiprout has proposed.[1] Researchers have described investigations of some of these effects on large-scale industrial designs.[2–4] Previous work, however, has not comprehensively spanned the entire range of modeling parameters, from detailed PDN modeling to full-die simulation, including a package model and nonuniform decap distribution. To the best of our knowledge, ours is the first comprehensive simulation study of an entire industrial processor, covering in detail these modeling and analysis issues.

In this article, we concentrate our study on the core region; we do not cover the I/O region. We electrically model a full-core die in the highest level of detail possible within computational-power constraints, and we justify the model from the bottom up. This requires beginning with a full-wave model for a small section of the die area and progressing in steps to a full-die and package cosimulation model containing all the essential elements required to attain the desired accuracy level. Simulations of the package-die model at every step highlight the critical and noncritical elements constituting the model. We ignore the noncritical elements, which do not significantly affect simulation accuracy, to incorporate a larger die area at

0740-7475/07/$25.00 © 2007 IEEE Copublished by the IEEE CS and the IEEE CASS

the next abstraction level. This enables the analysis of a larger region of the die for the same simulation time.

Using these models, we demonstrate the following for an Intel microprocessor designed in 90-nm technology: First, popular 2D inductive models, often used to model on-die inductance,[5] overestimate the impact of on-die inductance on supply noise. High-frequency (> 5 GHz) effects, which excite on-die inductive effects, are comparatively smaller, highly localized (with a radius of a few microns) to the switching device, and transient in nature (they decay quickly). The on-die power grid behaves otherwise as an RC network. Therefore, we can ignore on-die inductance for important frequencies and scales, considerably simplifying modeling and analysis. Second, the package has a significant impact on the accuracy of on-die power grid analysis. This necessitates including an accurate package model for a CAD approach, targeting transient PDN analysis and optimization.

Third, decoupling capacitors act both globally (full chip) and locally, depending on the frequency of excitation currents. They act globally at the main resonance frequency because of their interaction with package inductance (low frequency of about 100 MHz to 200 MHz). But the impact of decaps becomes increasingly more localized at frequencies higher than this resonance frequency. This is important for CAD placement, sizing, and optimization of decoupling capacitors. Fourth, localized (about a 1,000-micron radius), mid-frequency (about 1 GHz to 2 GHz) effects are possible due to the resistive isolation of pockets of capacitors interacting with localized C4 and package inductors, and these pockets collectively act as several mini-dies. This is a new, unpublished phenomenon, as yet to be addressed in CAD literature.

Full-wave inductive effects

The PDN is usually modeled as a linear network consisting of RLC elements, nominal voltage sources, and independent time-varying currents representing the switching currents of on-die devices. Computational constraints make it infeasible to model and analyze the entire PDN in complete detail. Hence, we began with a fully detailed but smaller (500 micron × 500 micron) section of the die area, consisting of metal layers M4 through M7. We used this model to observe high-frequency inductive effects and their locality. For this purpose, we began with a full-wave modeling method known as partial-element equivalent circuit (PEEC),[6] which has been used extensively in package-level analysis. Using the grid dimensions for each layer, we broke up the grid description into detailed via-to-via metal segments, including vias for all layers. The PEEC method models every metal segment with its self-resistance, self-inductance, and capacitance to ground, as well as its capacitive and inductive coupling to every other metal segment. This results in a dense, full-wave electromagnetic model that is highly accurate but extremely CPU and memory intensive.

The PEEC capacitors, which model the dielectric and metal charge interaction effects, served only to dampen the inductive ringing, so we removed them to highlight any inductive effects. The PEEC model consisted of 67,150 electrical nodes, 84,470 R and L elements, and 12 million mutual inductors. In addition, we assumed that total intrinsic and extrinsic decoupling capacitance was a low value of 10 pF for that area. We attached the total decoupling capacitance in a uniformly distributed manner to the lowest metal layer as 5,512 individual capacitors. This low value of decoupling capacitance is useful for highlighting potential inductive effects. The simulated area contained 13 C4 bumps, each modeled as a series RL element of 0.01 ohm and 0.325 nH, representing both the C4 and package input impedance. This is, in fact, a low inductance value per C4 for the package input. But, again, we used this low value to highlight any on-die inductive effects. We attached a source of variable rise times (10 ps to 100 ps) to the lowest metal layer.

PEEC simulation results

We compared the 3D PEEC simulation results to a standard model that modeled every layer separately in two dimensions. We then discretized the per-unit-length values to via-to-via segments and stitched the 2D layers together using resistive vias. We compared the simulation results for an R model (resistive-only PEEC grid, with decaps attached to M4), an RL model (R model with self-inductance), and an RLM model (RL with mutual inductors), all of which had RL C4 models attached to M7.

Figure 1a gives the simulation results for 3D PEEC. Clearly, the PEEC model, despite all the assumptions intended to highlight inductive effects, indicates little impact on inductance. However, in the 2D model (Figure 1b), there are significant inductive differences.

We performed this same study for various uniform and nonuniform sources, different rise times (down to 10 ps), and increased C4 inductance values (conforming to actual values) or device capacitance values

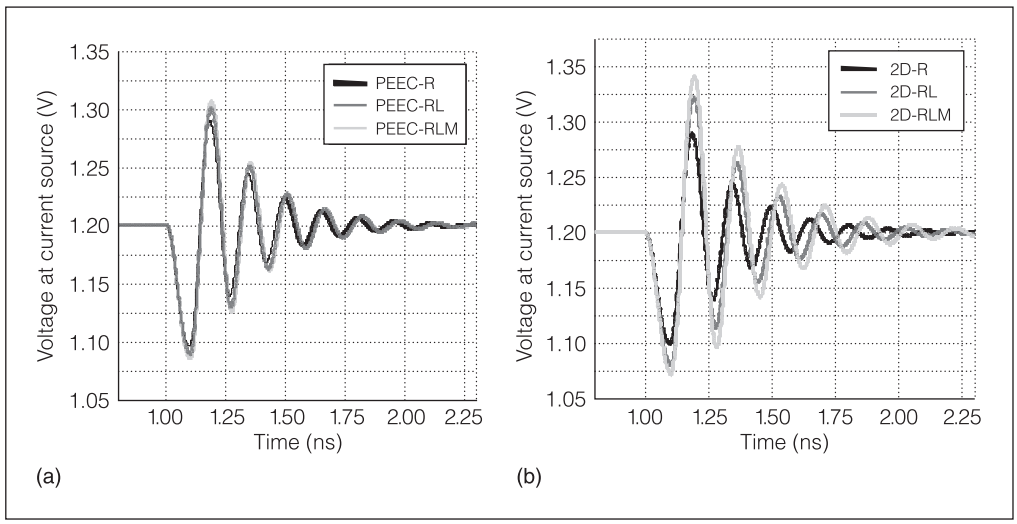

Figure 1. 3D partial-element equivalent circuit (PEEC) simulation of (500 micron × 500 micron) grid voltage response (a) compared with a 2D modeling approach (b) for R, RL, and RLM (RL with mutual inductors) models.

(conforming to actual decap densities). The results were similar to those in Figure 1.

The only difference observed in the full-wave RLC model was for very fast transients of 10 ps. These transients caused an initial high-frequency, localized transient blip (with a radius of a few microns), as Figure 2 shows. This blip quickly degenerated into a wave fully described by an RC grid model. Given that our assumptions highlighted the impact of on-die inductance, even this small localized inductive effect would be smaller than what we observed if all of the details of the localized model were in place.

High-frequency noise

The PEEC model describes all potential inductive interactions for the full dimensions of the model. However, remote potential interactions do not determine the return path, and the high-frequency currents (> 5 GHz) tend to remain extremely localized. If the model were extended to a larger area, the locality of the high-frequency supply drop would not change. There are four main reasons for this:

- High-frequency inductive current loops tend to remain small because of the high energy involved in maintaining larger current loops.
- There are many power rail vias in a microprocessor PDN, providing various return pathways.

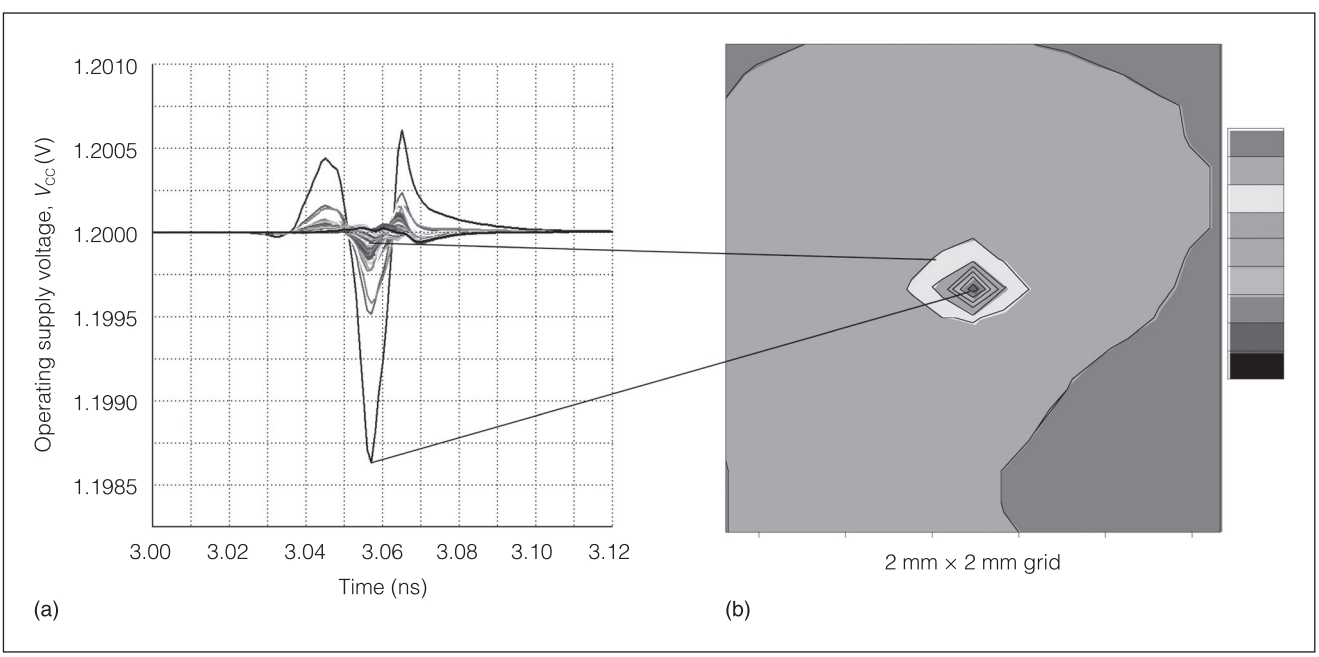

Figure 2. A fast current spike (a) injected into a power grid can excite localized inductive effects (b), which quickly dissipate in time and space and become RC-only effects. (V_{cc}: nominal supply voltage.)

- The grid is loaded with wire resistance, device and wire capacitance, and C4 and package inductance. All of these help dissipate the high-frequency energy.
- The higher the frequency of current transients, the better the response of nearby decaps, whose impedances are a function of frequency. At very high frequencies, the nearby decaps provide most of the charge to the current source, thus making the voltage drop more localized.

Although the high-frequency response is very localized, the mid- to low-frequency currents (1 GHz to 2 GHz) dissipate outward from the source to affect several gates. Therefore, when there are multiple switching sources of current, each gate's high-frequency transients have only a local impact around the gate. The mid- to low-frequency transients, on the other hand, have an additive impact at every neighboring gate, overwhelming each gate's localized high-frequency effect in amplitude. This is another significant reason for not requiring inductance in a power grid model. The model necessary for understanding the full die requires only a resistive grid with device capacitance and a C4-package model. Note that the high-frequency response is localized even for wire-bond packages. This is true because at high frequencies, most of the current comes from the nearby decaps, and not through the package, whose inductance gives it a very high impedance at those frequencies. Thus, the locality should be more pronounced for wire-bond packages due to their higher inductance as compared with flip-chip packages.

Midsize model and capacitive effects

Given the conclusions of the previous study, we eliminated inductance in our larger models, used an R-only model for the grid along with device decaps, and extended our detailed via-to-via metal segment model to 2 mm × 2 mm and to metals M2 through M7. This let us determine a larger area of interaction and understand the properties of this larger grid so that we could build a full-chip model.

We attached this die model to an RLC package model, which modeled the package from the die shadow (the package area where the die is placed) to the VRM, and which was discretized to 9 mm × 9 mm in the die shadow area. Our segment of the grid was only big enough to cover a (2 mm × 2 mm) section of that area. In the middle of our grid at M2, we placed a single-frequency-domain current source to observe its effect on the surrounding droop. There was an open question regarding the modeling of the discrete die shadow pins in the rest of the die area (outside the 2 × 2 section). As Figure 3 shows, we tried two cases:

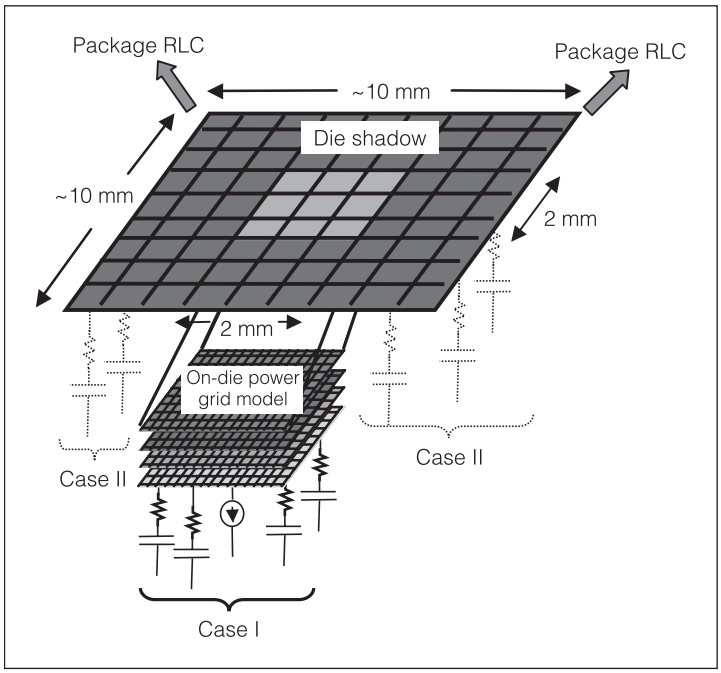

Figure 3. Microprocessor die shadow and package interface. We attached a detailed (2 mm × 2 mm) section of the on-die grid to the middle of the package die shadow, with decoupling capacitors either attached only to the 2 × 2 grid section (Case I) or also attached to the rest of the die shadow pins (Case II).

- *Case I.* Attach all the die capacitance under the 2 × 2 die model.
- *Case II.* Distribute the die capacitance evenly in the 9 × 9 die shadow, with the 2 × 2 section placed under the attached die model, and the rest directly attached to the package-die interface pins.

The frequency response of a detailed 3D package model for the design under study showed approximately a 200-MHz resonance frequency, which we also validated through silicon measurements. We also observed this 200-MHz resonance frequency for the package-die model in Case I. However, when we distributed the die capacitance outside the 2 × 2 section, another spurious frequency (60 MHz) resulted. We deduced that this spurious resonance was due to the high-impedance path from the capacitors outside

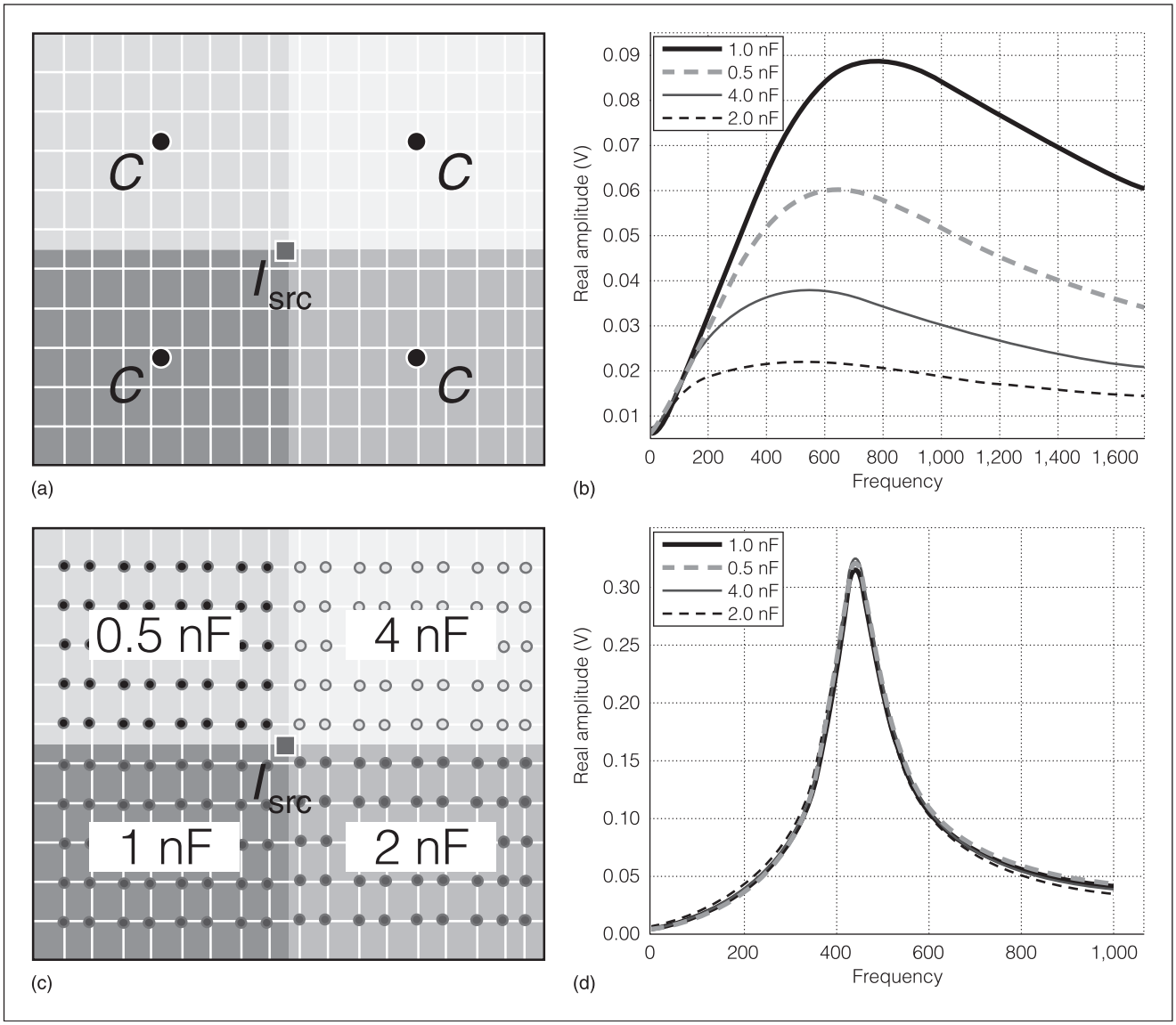

Figure 4. Frequency-domain simulations of the 2 × 2 grid section with C4 RL models and four distinct capacitors of value 1 nF, 2 nF, 0.5 nF, and 4 nF (a); the frequency response over each capacitor showing four distinct resonant frequencies (b); the same grid and capacitor values but with the capacitors randomly interspersed between the nodes in each quadrant (c); and the frequency response of the randomly interspersed four-capacitor voltages, showing the same resonant frequency (d). (I_{src} is the AC source placed at the center of the grid.)

the 2 × 2 section to those inside the section because all remote decap currents had to travel through the package without an on-die connection. This shows that a full-die grid resistance model is essential for modeling correct global die behavior. We illustrate this principle more clearly in the following simple example.

We simulated the same 2 × 2 grid section but with four individual capacitors placed in the middle of the four (1 mm × 1 mm) quadrants, with values of 1 nF, 2 nF, 0.5 nF, and 4 nF (Figure 4a). We attached RL models to the C4 pins with values equal to the input impedance of the rest of the package, and we attached a single AC source, I_{src}, to the center of the grid. When we probed the frequency response of the voltage over the capacitors, we observed four distinct resonant frequencies (Figure 4b). However, when we spread the capacitor values randomly around the four quadrants while maintaining the same total capacitance (Figure 4c), we observed only a single resonant frequency (Figure 4d). This led us to conclude that resistive

Table 1. Runtime and peak memory usage of the 2 × 2 die model before and after multigrid-based reduction.

2 × 2 model	No. of nodes	No. of elements	Runtime (s)	Peak memory usage
Original	877,259	1,249,250	1,453.75	2.2 Gbytes
2× reduction	222,475	323,874	797.83	650 Mbytes
4× reduction	57,861	85,110	140.46	300 Mbytes

isolation between capacitive regions, along with the limited number of C4-package inductors above the regions, caused the four regions to act as distinct mid-frequency resonant circuits (four mini-dies, in a sense). The fact that there could be isolated mid-frequency pockets greater than the die-package resonance was an important new effect that this analysis exposed.

Model reduction

To progress to a full-die model that fit into memory, we had to reduce the resistive grid from the level of detail contained in the 2 × 2 model. Using the same level of accuracy was not feasible for a full die. However, we needed to determine how much we could reduce the grid and still maintain the accuracy of the detailed effects we wanted to observe, especially with respect to resonance. We applied a previously proposed multigrid method for this purpose.[7] We used this method to reduce our 2 × 2 model by a factor of 2, and then by 4, to determine the accuracy of the resultant models. Table 1 shows the comparison of runtime and peak memory usage for a transient simulation of the original and the reduced grids. In our experiments, we observed that a 4× reduction allowed the RC model of the entire (10 mm × 10 mm) grid to fit into the memory without incurring significant accuracy loss.

Locality in power grids

In flip-chip power grids, the *IR* drop, analyzed using DC simulations, has the property of locality; the voltage droop from a single current source stays in the proximity of that source because of the presence of C4 sources.[1] However, it was not clear what this locality principle meant for a package-die PDN model in the time and frequency domains.

For this purpose, we attached a 4×-reduced, M2–M7 resistive grid to an RL C4-package model with a uniform capacitor distribution at M2. We placed a single frequency source in the middle, then we probed all the voltage nodes on M2 in the frequency domain and simulated them from low (DC) to mid-frequencies (about 1 GHz to 2 GHz). Figure 5 shows the results.

Each curve represents the frequency response of one node on M2 to a single source in the middle of the grid. On the DC (left) side, there is clear locality because there is a decreasing response of nodes as we move away from the source (downward movement on the *x*-axis), until there is a zero response. On the mid-frequency (right) side, there is quasi-locality as the response gets smaller with distance but never goes to zero, indicating that some diminishing capacitor currents are always supplied at a distance. At the main low-frequency package-die resonance in the middle, all locality effect is lost. This indicates that at the main resonant frequency, the die and the package are acting as one, and charge is flowing everywhere on the die. However, at other frequencies, the capacitors and decaps tend to act in a local manner, implying partial locality at mid-frequencies.

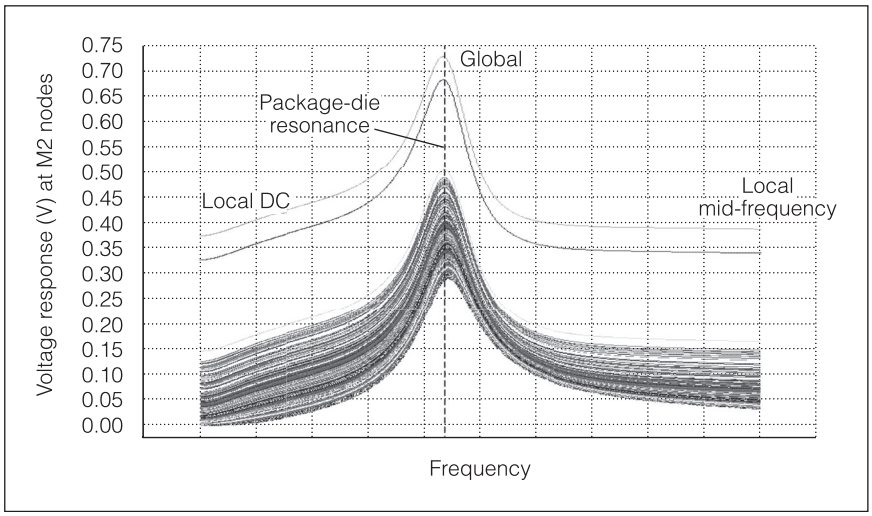

Figure 5. Frequency response of all M2 voltage nodes, illustrating locality as a function of excitation frequency in the grid.

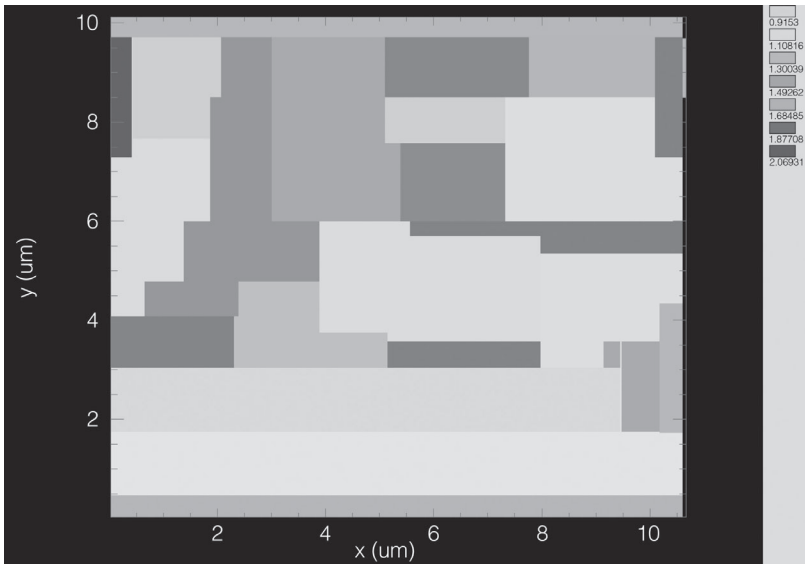

Figure 6. The nonuniform, block-based decoupling-capacitance distribution of the die.

distribution. We reduced the package model to a per-C4 input impedance. Furthermore, we took the actual nonuniform, per-design-block, full-die capacitor distribution (Figure 6) and placed it on the M2 metal nodes.

With a single 10-ps source placed in the middle of a central unit, we observed the time domain current waveforms of all nonuniformly distributed capacitors on M2 (Figure 7). As time progressed, all the capacitor currents synchronized with the global resonant frequency described by the die-package resonance. According to Figure 5, this is the stage where locality is lost and all capacitors are charge-sharing. However, in the beginning, the response to the fast transient consists of multiple frequencies higher than the global resonant one.

In Figure 7, we observe multiple resonant frequencies, some higher than the main resonant frequency. This demonstrates the presence of mid-frequency effects due to the nonuniform capacitor distribution and the resistive-grid isolation. To understand the locality of these mid-

Complete package-die model

We constructed the most realistic full-die model we could using the 4×-reduced, M2–M7, full-die grid; the package model; and a realistic nonuniform decap

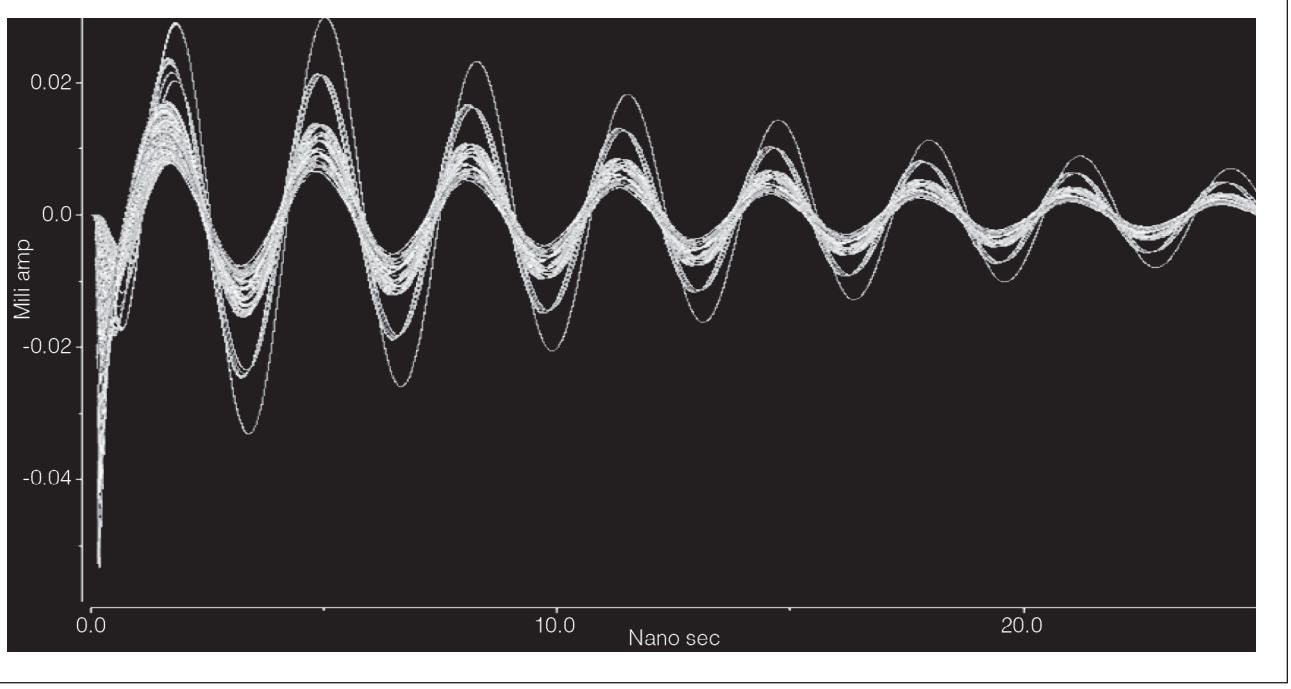

Figure 7. Transient current response of the nonuniform decaps. Initially, there are various frequencies greater than the global resonant frequency, but eventually all currents respond at the main global resonant frequency.

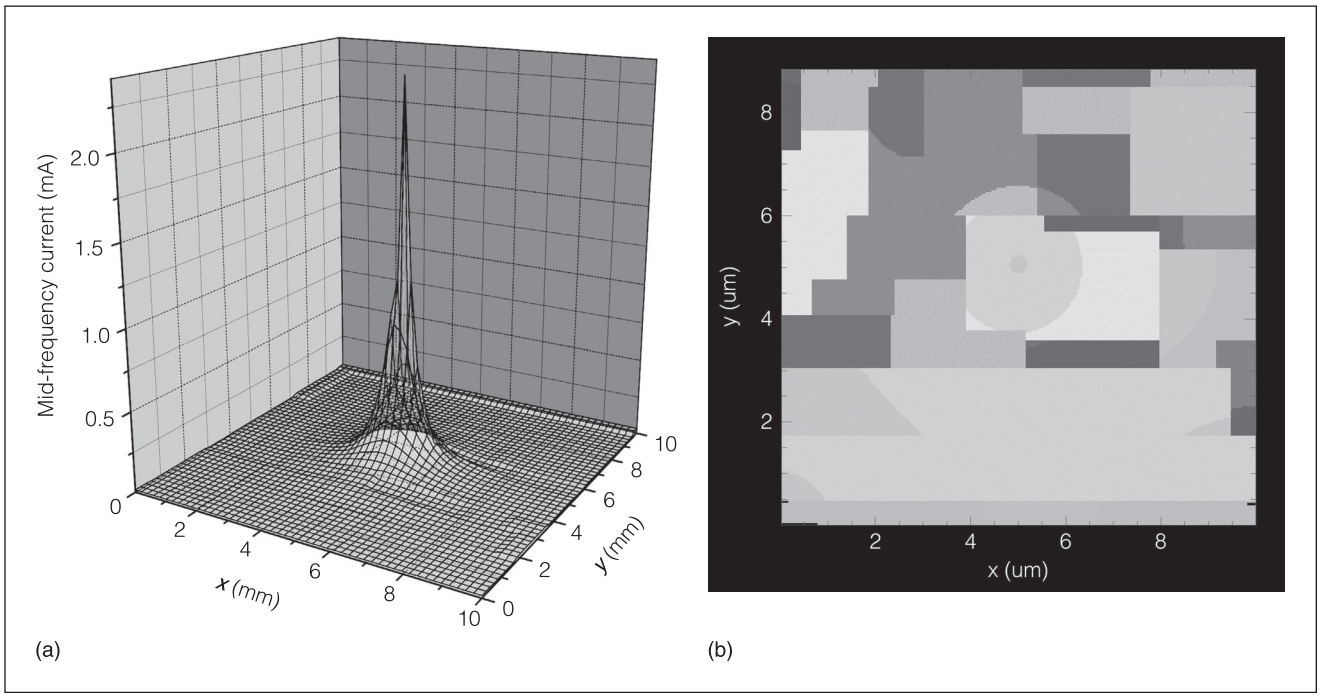

Figure 8. Locality of mid-frequency (a) and low-frequency (b) currents present in the first and second dips, respectively, of the current amplitude curves in Figure 7.

frequency transients compared with the global resonance, we plotted the amplitude of the currents at two specific time points: at the bottom of the first dip in Figure 7, where the mid-frequency effects were visible (Figure 8a), and at the bottom of the second dip in Figure 7, where the low-frequency responses almost converged to a global resonance (Figure 8b). The 3D plot in Figure 8a clearly shows that the mid-frequency effects are local to a radius of approximately 1 mm. By the second dip in Figure 7, there is almost global convergence (Figure 8b), and the capacitor currents reflect an almost perfect correlation with the full-die capacitor distribution in Figure 6. Thus, mid-frequency effects can be resonant at less than full die, but low-frequency die-package effects are global. This is a new phenomenon, not demonstrated previously. This kind of RC locality is very different from, and of a wider area than, the high-frequency locality mentioned earlier (Figure 2), when we were discussing inductive effects.

We can explain these effects as follows. When a single gate switches, it pulls in power delivery current from various sources. At high frequencies, the package with large parasitics is effectively isolated from the die. If the frequency is high enough in a small local area, it will excite on-die inductance, but this effect will be highly transient and limited to a radius of a few microns before the RC background absorbs the high-frequency energy. The high-frequency currents are immediately satisfied by capacitors nearby—either explicit decaps, nonswitching device capacitors, or wire capacitors. The farther away the capacitor, the less the current supplied, but the supply radius grows larger as the frequency decreases. At some mid-frequencies (greater than the global die-package frequency), the package comes into play, and mid-frequency currents are supplied through the C4s. However, even at these frequencies, the mid-frequency currents also continue to flow from the capacitors. If pockets of capacitors, resistively isolated (partially or completely) from other capacitors farther out, surround the gate, the local capacitors resonate only with the local C4 bump-package inductance above them, causing a mid-frequency resonance of a radius of a few hundred or more microns. This, effectively, is a small version of the total die at resonance. When the frequency is low enough (main die-package resonance), all the capacitors and all the C4s interact to produce a global resonant frequency that is full die in nature.

Because of the distributed nature of C4s in flip-chip packaging, voltage drop induced from current excitation may be limited to the vicinity of the

current source. Recently, several researchers have proposed exploiting this locality in power grids to accelerate voltage drop analysis.[1,8] However, as we demonstrated in this article, transient locality is a strong function of the excitation frequency. Although voltage drop exhibits locality for the DC and high-frequency excitations, it is global at frequencies around the resonance frequency caused by package inductance and on-die decaps. Moreover, the area of locality depends on the frequency of excitations. Thus, although locality can simplify and accelerate the static power grid analysis, as Chiprout has proposed,[1] its use for transient power grid analysis and optimization could lead to erroneous results unless integrated with these effects. ■

Acknowledgments

We are grateful for the feedback we've received from Marek Patyra, Kim Eilert, Bob Martell, Kaladhar Radhakrishnan, and several other persons at Intel.

■ References

1. E. Chiprout, "Fast Flip-Chip Power Grid Analysis via Locality and Grid Shells," *Proc. Int'l Conf. Computer-Aided Design* (ICCAD 04), IEEE CS Press, 2004, pp. 485-488.
2. A. Dharchoudhury et al., "Design and Analysis of Power Distribution Networks in PowerPC Microprocessors," *Proc. 35th Design Automation Conf.* (DAC 98), ACM Press, 1998, pp. 738-743.
3. H. Chen and D. Ling, "Power Supply Noise Analysis Methodology for Deep-Submicron VLSI Chip Design," *Proc. 34th Design Automation Conf.* (DAC 97), ACM Press, 1997, pp. 638-643.
4. A.V. Mezhiba and E.G. Friedman, "Impedance Characteristics of Power Distribution Grids in Nanoscale Integrated Circuits," *IEEE Trans. Very Large Scale Integration (VLSI) Systems*, vol. 12, no. 11, Nov. 2004, pp. 1148-1155.
5. M. Xu and L. He, "An Efficient Model for Frequency-Dependent On-Chip Inductance," *Proc. 11th Great Lakes Symp. VLSI,* 2001, ACM Press, pp. 115-120.
6. A.E. Ruehli, "Equivalent Circuit Models for Three Dimensional Multi-conductor Systems," *IEEE Trans. Microwave Theory and Technology*, vol. 22, no. 3, Mar. 1974, pp. 216-221.
7. J.N. Kozaya, S.R. Nassif, and F.N. Najm, "A Multigrid-Like Technique for Power Grid Analysis," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, no. 10, Oct. 2002, pp. 1148-1160.
8. H. Qian, S.R. Nassif, and S.S. Sapatnekar, "Power Grid Analysis Using Random Walks," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 8, Aug. 2005, pp. 1204-1224.

Sanjay Pant is pursuing a PhD in electrical engineering and computer science at the University of Michigan, Ann Arbor. His research interests include VLSI design, with emphasis on power delivery and signal integrity. Pant has a BTech in electrical engineering from the Indian Institute of Technology, Kanpur, India, and an MS in electrical engineering and computer science from the University of Michigan, Ann Arbor. He is a student member of the IEEE.

Eli Chiprout is a principal research engineer at Strategic CAD Labs, Intel. His research interests include power delivery modeling and optimization, nonlinear macromodeling, variational models, and silicon correlation. Chiprout has a BEng in electrical engineering from McGill University, Montreal, and an MEng and a PhD in electrical engineering from Carleton University, Ottawa. He is a member of the IEEE.

David Blaauw is an associate professor in electrical engineering and computer science at the University of Michigan, Ann Arbor. His research interests include VLSI design and CAD, with emphasis on circuit design and optimization for high-performance, low-power applications. Blaauw has a BS in physics and computer science from Duke University, and an MS and a PhD in computer science from the University of Illinois, Urbana. He is a member of the IEEE.

■ Direct questions and comments about this article to Sanjay Pant, University of Michigan, Ann Arbor, 4844 CSE, 2260 Hayward St., Ann Arbor, MI 48109-2122; spant@umich.edu.