Power Grid Physics and Implications for CAD

Sanjay Pant
University of Michigan, Ann Arbor, MI

Eli Chiprout
Intel Strategic CAD Labs, Hillsboro, OR

ABSTRACT

Much research has been done lately concerning analysis and optimization techniques for on-chip power grid networks. However, all of these approaches assume a particular model or behavior of the power delivery. In this paper, we describe the first detailed full-die dynamic model of an industrial microprocessor design, including package and non-uniform decap distribution. This model is justified from the ground up using a full-wave model and then increasingly larger but less detailed models with only the irrelevant elements removed. Using these models we show that there is little impact of on-die inductance in such a design, and that the package is critical to understanding resonant properties of the grid. We also show that transient effects are sensitive to non-uniform de-cap distribution and that locality is a tight function of frequency and of the package-die resonance, producing newly explained localized resonant effects. Specifically, all of these points have impact on what kind of analysis and optimization are required from CAD.

Categories & Subject Descriptors:

B.8.2: Perf. Anal & Design Aids.

General Terms: Verification, Reliability.

Keywords

IR, Ldi/dt, decap, resonance, locality, power supply networks.

1. INTRODUCTION

On-die power delivery network (PDN) physics has long been an incomplete puzzle because the effects are complicated. In a microprocessor design, rapid transient currents are generated by the transistors in various spatial and temporal distributions and are fed in and out of the PDN. If the PDN is in the form of a grid and the packaging is flip-chip, then current flows through the transistors onto the power grid, charging and discharging various capacitances and flows onto the package through the C4 bumps and eventually to the voltage regulator module (VRM). This flow of currents will cause voltage variation on the PDN, either through DC or transient currents.

Incomplete in the above picture are the following questions:

- Are there inductive effects on the power grid?
- How localized are currents as they flow outward from a gate?
- Does capacitance charge respond locally, globally or both?
- What is the transient impact of the C4 bumps and the package?
- Do resonant effects occur and, if so, how?

The answers are critical to address the kind of models and CAD algorithms are required to deal with the PDN and in chip-package co-design. For example, if effects are localized, then analysis may be considerably simplified[8].

There have been some previous investigations of some of these effects on a real design in the literature[1][2][3]. Previous works, however, have not been comprehensive to span the range from detailed models to full die simulation, including a package model and non-uniform decoupling capacitor (decap) distribution. This is the first simulation of an entire industrial processor in detail.

In this paper we do not cover the I/O region but rather concentrate our study on the core region. We electrically model a full core die in the highest level of detail possible and we justify the model from the bottom up. This requires us to begin with a full-wave model for a small region of the grid and progress in steps to a full-die and package model that contains all essential elements for the accuracy level required. Each model shows components that are important, as well as those that are not; components which are removed in order to use a model of a larger area at the next level of abstraction which allows a larger analysis region for the same CPU time. Using these models we were able to show that for an Intel Pentium® class microprocessor in 90nm technology:

- Popular 2D inductive models are over-estimating the inductance on the power grid. High frequency (>5GHz) effects, which excite inductive effects are small, highly localized (a few micron radius) to a switching gate and very transient (decay quickly). The grid is an RC phenomenon otherwise. Therefore inductance may be ignored on-die for frequencies and scales of importance, considerably simplifying a CAD approach.
- The package has a large impact on the kind of model necessary for the on-die power grid. One must include the package model for a realistic CAD approach to transient analysis/optimization of the PDN.
- Decoupling capacitors act both globally (full-chip) and locally, and when they act globally they do so at the main resonant frequency of the package and die (low frequency ~100-200MHz). This is important for CAD placement, sizing and optimization of decoupling caps.
- Localized (~1000μ radius), mid-frequency (~ 1-2 GHz) effects are possible by resistive isolation of pockets of capacitors interacting with localized C4 inductors, acting as a mini die. This is a new, unpublished phenomenon, as yet to be dealt with in the CAD literature.

We begin with an explanation of our smallest but most detailed model, progress to a 2mm X 2mm model and finally, describe our full-chip microprocessor model. We conclude with a coherent explanation of the various highlighted effects and their interaction and the implications on CAD modeling and optimization.

2. Full Wave Inductive Effects

Our PDN models were linear because, in the first order, the PDN, with gates attached, acts as a linear network [2]. Initially, we wanted to understand high-frequency effects, including inductive

DAC 2006, July 24–28, 2006, San Francisco, California, USA.

effects, and their locality. For this purpose we began with a full-wave modeling method known as Partial Element Equivalent Circuit or PEEC [4]. This method is used extensively on 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, as shown in Figure 1, 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 process results in a dense full-wave electromagnetic model (excluding speed of light effects) that is highly accurate but extremely CPU and memory intensive.

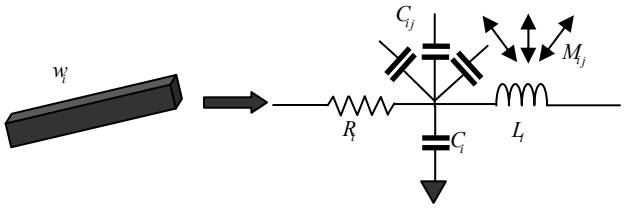

Figure 1. A PEEC model for one wire segment (assuming coupling to other nearby metal segments).

We were able to fit into memory only a 500µ X 500µ PEEC model of the power grid using metals M4 to M7. We found that the PEEC capacitors, which modeled the dielectric and metal charge interaction effects, only served to dampen the inductive ringing and thus we removed them to highlight any inductive effects. The PEEC model contained 67,150 electrical nodes, 84,470 R and L elements, and 12 million mutual inductors. In addition, we assumed that total active and passive decoupling capacitance was a low value of 10pF for that area and attached it to the lower metal as 5,512 individual capacitors. This low value is useful in order to highlight potential inductive effects. The area contained 13 C4 bumps and these were each modeled as series R-L element of 0.01ohm and 0.325nH respectively, representing both the C4 and package input impedance effects. This is, in fact, a low package input inductance value per C4 but, again, used in order to highlight any on-die inductive effects (a more accurate generation of package parasitics will be described later). A source of 10 to 100ps rise times was attached to the lowest metal layer. The model also made the following simplifications:

- It avoided discretizing the wires for skin effects
- It assumed room temperature rather than high temperature
- It did not model signals in the neighborhood of the grid (which have the effect of absorbing inductive noise)
- It injected current directly into M4 rather than model M1-M3

All of the above assumptions effectively increase inductive effects and are a worst case scenario in order to detect them. The 3D PEEC simulation results were compared to a 2D model which modeled every layer separately in 2D, discretized the per-unit-length values to via-to-via segments, and then stitched the 2D layers together using resistive vias. The simulation results were compared for an R model (resistive only PEEC grid, with de-caps attached to M4), an RL model (R model with self inductance) and an RLM model (RL model with mutual inductors), all of which had R-L C4 models attached to M7. The simulation results for PEEC 3D are given in Figure 2, left.

It is clear that the PEEC model, in spite of all the assumptions intended to highlight inductive effects, shows little impact of inductance. However, we note that in the 2D model (Figure 2, right), we see significant inductive differences. This indicates that there is a flaw in the 2D model which increases the return path inductance. Interested readers can refer to [5] for the proof. Unfortunately, it is 2D models that have often been used to model inductance on power grids, yielding subsequent tenuous conclusions based on those models. The same study above was performed for various uniform or non-uniform sources, different rise times (down to 10ps) and for increased C4 L values (conforming to actual values) or device capacitance values (conforming to actual de-cap densities). The results were similar.

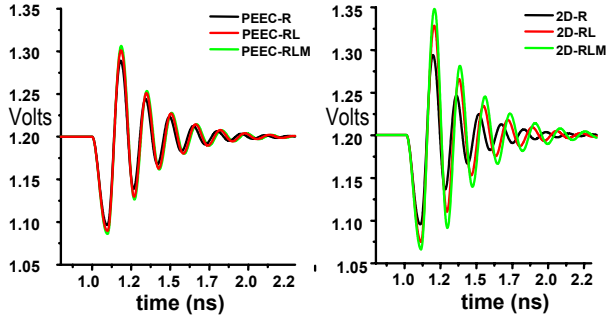

Figure 2. A 3D PEEC simulation of 500µX 500µ grid voltage response (left) compared to a 2D modeling approach (right).

The only difference observed in the full-wave RLC model was for very fast transients of 10ps, which caused an initial high frequency localized (a few micron radius) transient "blip" which quickly degenerated into a wave fully described by an RC grid model. Given that we overestimated the inductance, even this small localized inductive effect should be smaller than what we observed, if all of the details of the localized model were in place.

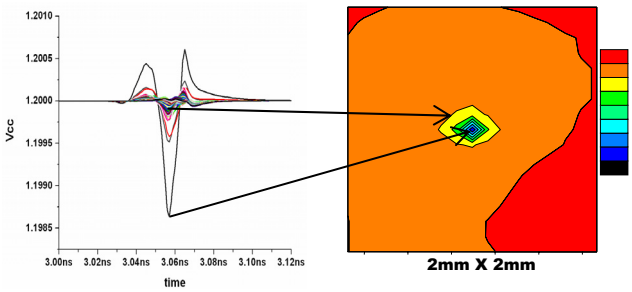

Figure 3. The fast current transients generated by a driver switching at 10ps injected into the power grid can excite localized inductive effects which quickly dissipate in time and space in a few micron radius and become "RC only" effects.

2.1 High-frequency Noise

It is important to note that 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 (>5GHz) tend to remain extremely localized. If the model were to be extended to a larger area, the locality of the high frequency would not change. This result may be explained due to three main reasons:

- High frequency L current loops tend to remain small due to the high energy involved in larger loop sizes.

- There are frequent power rail vias in a microprocessor PDN, providing frequent return pathways.
- The grid is loaded with wire resistance, device and wire capacitance, and C4 and package inductance. All of these help to dissipate the high frequency energy.

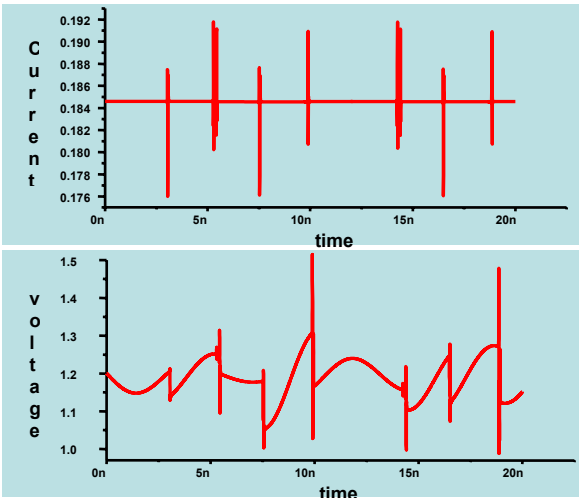

Figure 4. Fast device currents (top) injected into a PDN model with caps and C4/package model. The voltage response (bottom) shows that the high frequency energy is quickly transferred into the low frequency package/die interaction.

To further explain the impact of high frequencies we obtained the fast transient switching currents of a design unit of a few thousand gates by using a circuit simulation of random vector inputs injected into the circuit netlist (Figure 4, top). When we placed these fast transients on a detailed grid with device caps and C4 RL models, we noticed an interesting phenomenon. The high frequency currents, as noted, quickly dissipate in the time domain, and the energy is transferred to the low frequency resonance determined by die package interaction (Figure 4, bottom) which will be described next. The current injections act much like an impulse input to the slow response of the larger system. This is a well-known phenomenon in differential equations and is illustrated by someone quickly kicking their automobile. The high frequency generated by the kick will quickly dissipate due to the low pass filter of the system and the car will oscillate at a much lower resonant frequency to which the energy is transferred.

Although the high frequency response is highly localized, the mid-low frequency currents (1-2 GHz) dissipate outward from the source to affect several gates. Therefore, when there are multiple switching sources of current, the high frequency transients of each gate will only have local impact around the gate while the mid-low frequency transients will have 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.

3. Mid-size Model and Capacitive Effects

Given the previous conclusion and its explanation, we eliminated inductance in our larger models, used an R-only model for the grid together with device caps and extended our detailed via-to-via metal segment model to 2mm X 2mm and to metals M2 to M7. This allowed us to determine a larger area of interaction and to understand the properties of this larger grid in order to build a full-chip model.

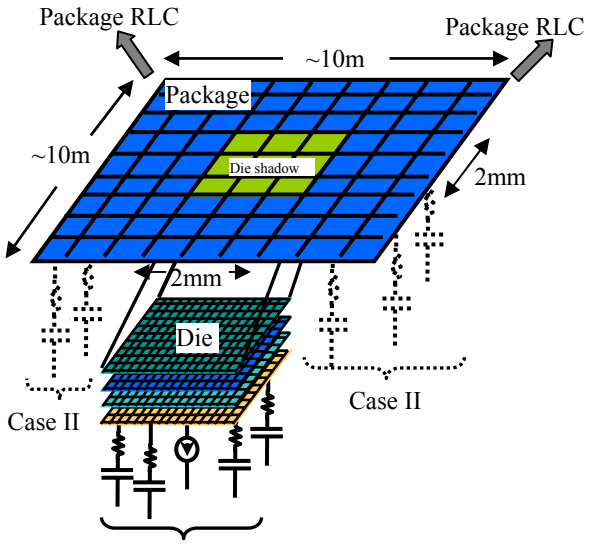

Figure 5. The microprocessor die shadow and package interface. A detailed 2mm X 2mm section of the on-die grid was attached to the middle of the package die shadow, with decoupling caps either being attached (Case I) only to the 2mm X 2mm grid section or (Case II) to both it and the rest of the die shadow pins.

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 9x9 in the die shadow area. Our segment of the grid was only big enough to cover a 2x2 section of that area. In the middle of our grid at M2 we placed a single frequency domain current source in order to observe its impact on the surrounding droop. What to do with the other discrete die shadow pins was an open question and we tried two cases: (I) attach all the die capacitance under the 2mm X 2mm die model or (II) distribute it evenly in the 9x9 die shadow, with 2x2 of the sections placed under the attached die model, and the rest directly attached to the package/die interface pins (see Figure 6). From the package perspective, a resonance frequency of ~200MHz was expected, which is a well known phenomena to package designers[5]. This was seen for case (I). However, when the die capacitance was distributed outside of the 2mm X 2mm section, another spurious frequency was observed, that of 60 MHz. We deduced that this spurious resonance was due to the high impedance path from the caps outside of the 2mmX2mm section to those inside the section, since all remote de-cap currents had to travel through the package without an on-die connection. This showed that a full-die grid resistance model is essential for a correct global die behavior. We illustrate this principle more clearly in the following simple example.

We simulated the same 2mm X 2mm grid section but with four individual capacitors placed in the middle of the four 1mm X 1mm quadrants, with values of 1nF, 2nF, 0.5nF and 4nF (Figure 6, left). We attached RL models attached to the C4 pins with values equal to the input impedance of the rest of the package. When we probed the

frequency response of the voltage over the caps, using one current source in the middle of the grid, we observed 4 distinct resonant frequencies (right). However, when we spread the cap values randomly around the 4 quadrants (while maintaining the same total amount), we observed only a single resonance frequency (Figure 7). This implied that resistive isolation between capacitive regions, together with the limited number of C4 inductors above the regions, caused them to act as distinct mid-frequency resonant circuits (four mini die, in some sense). This implied a principle of locality which is explained in the next section. The fact that one could have isolated pockets of mid-frequency (greater than the die-package resonance) was an important new effect that was exposed by this analysis.

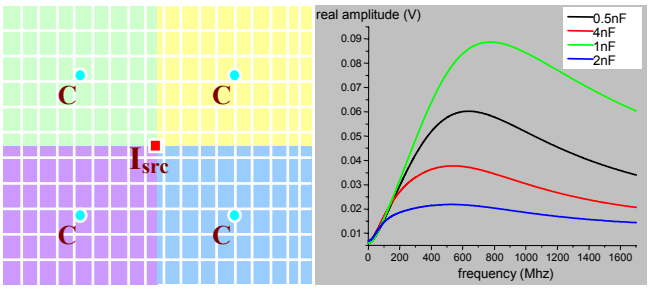

Figure 6. The 2mm X 2mm grid section with C4 RL models and 4 distinct capacitors of value 1.0, 2.0, 0.5, 4.0nF respectively (left) and the frequency response over each capacitor showing four distinct resonant frequencies.

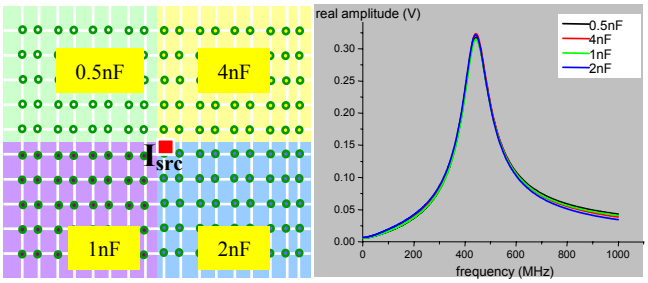

Figure 7. The same grid and cap values as in Figure 6 but with the caps randomly interspersed among the nodes in each quadrant (left) and the frequency response of the 4 capacitor voltages (right) showing the same resonant frequency.

3.1 Model Reduction

In order to progress to a full-die model that fit into memory, it was important to reduce the resistive grid from the level of detail contained in the 2mm X 2mm model. Using the same level of accuracy was not feasible for a full die. However, we needed to determine how much to reduce the grid and still maintain the accuracy of the detailed effects we wanted to observe, especially with respect to resonance. We applied the prior-proposed multi-grid method for this purpose [5]. We used this method to reduce our 2mm x 2mm model by a factor of 2 and then by 4 in order to determine the accuracy of the resultant models. When we placed unit transient currents on the original grids and the reduced grids we obtained the results in Figure 8. A 4X reduction allowed entire grid to fit in memory and not suffer much accuracy loss.

3.2 Locality in Power Grids

Flip-chip power grids in DC have been shown to have property of locality, in which the voltage droop from a single current source stays in the proximity of that source due to the C4 sources[8]. However, it was not clear what this locality principle meant for a package-die PDN model in the time and frequency domain.

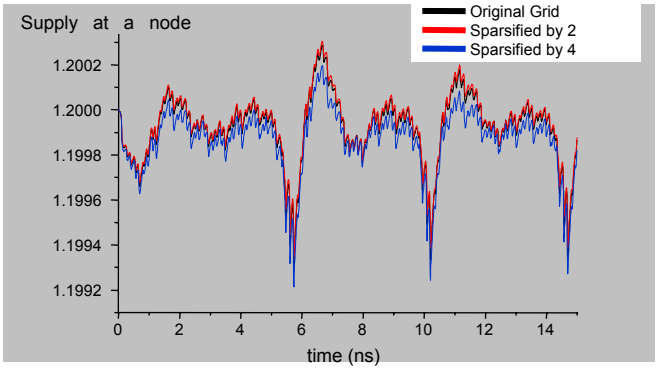

Figure 8. The voltage droop of a 2mmX2mm grid and the 2x and 4x reduced grids using the same block current excitation.

For this purpose, a 4X reduced M2-M7 resistive grid was attached to an RL C4/package model and with a uniform cap distribution on M2. A single frequency source was placed in the middle. All the voltage nodes on M2 were probed in the frequency domain and simulated across DC to mid-frequencies (~1-2 GHz). The results are shown in the information-rich graph in Figure 9.

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 one moves away from the source (downward movement on the X-axis) until there is a zero response. On the mid-frequency right side, there is a 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, both the die and package are acting as one and charge is flowing everywhere on the die. However, at other frequencies, the caps and de-caps tend to act in a local manner. This implies principles of partial locality at mid-frequencies.

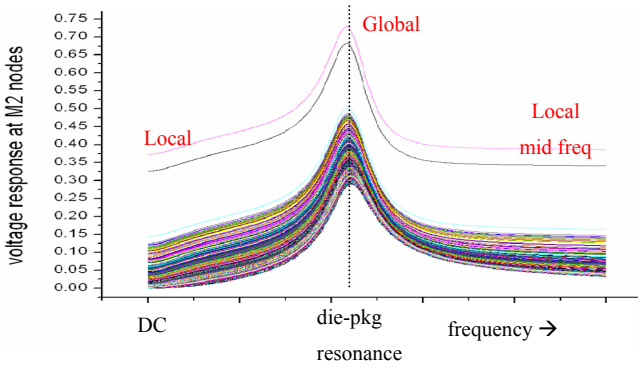

Figure 9. Frequency response of all M2 voltage nodes illustrating locality in the grid.

4. Complete Package-Die Model

The most realistic full-die model was constructed using the 4x reduced M2-M7 full-die grid, the package model and a realistic non-uniform decap distribution. We reduced the package model to a per-C4 input impedance. Further, we took the actual non-uniform per-design-block full-die cap distribution (Figure 10) and placed it on the M2 metal nodes.

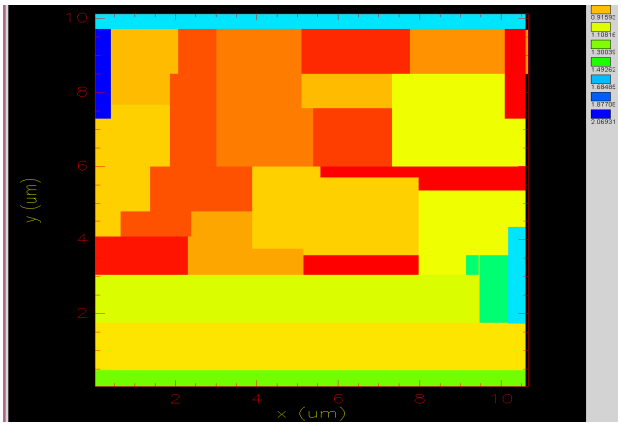

Figure 10. The non-uniform block-based decoupling capacitance distribution of the die.

With a single 10ps source placed in the middle of a central unit, we observed the time domain current waveforms of all of the non-uniformly distributed caps on M2 (Figure 11). Note that as time progresses, all the cap currents sync up with the global resonant frequency described by the die-package resonance. As understood from Figure 9, this is the stage where locality is lost and all caps are charge-sharing. However, note that in the beginning, the response to the fast transient consists of multiple frequencies much higher than the global resonant one.

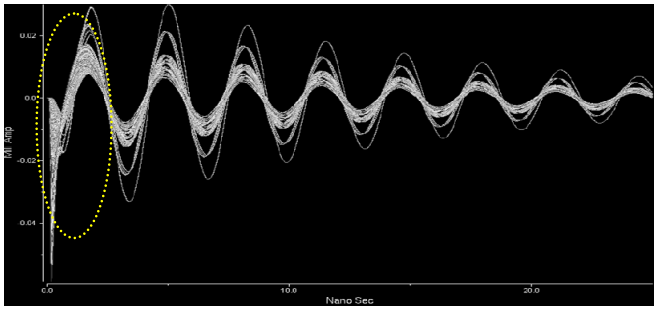

Figure 11. The transient current response of the non-uniform decaps. Initially there are various frequencies (circled, and magnified in Figure 12) greater than the global resonant frequency, after which all currents respond at the main global resonant frequency.

The highlighted waveforms of the first dip in Figure 11 are expanded in the graph of Figure 12. Here we see that the frequencies are multiple with some being higher than the main resonant frequency. This demonstrated that there were mid-frequency effects due to non-uniform cap distribution and resistive grid isolation. In order to understand the locality of these mid-frequency transients as compared to the global resonance, we plotted the amplitude of the currents at two specific time points: at the bottom of the first dip in Figure 11, where the mid-frequency effects were visible (Figure 13), and one at the bottom of the second dip where the responses have almost converged to a global resonance (Figure 14). We observe clearly in the 3D plot in Figure 13 that the mid-frequency effects are "local" to a radius of approximately 1mm. However the current magnitude plot of Figure 14 shows that by the second dip there is almost global convergence and the cap currents reflect an almost perfect correlation with the full-die cap distribution in Figure 10. Thus mid-frequency effects may be resonant at less than full-die, but low frequency die-package effects are global. This is a new phenomenon that was not demonstrated previously. Please note, that this kind of RC locality is very different and of a wider area as compared to the high frequency locality spoken of earlier in discussing inductive effects. There the locality was much smaller.

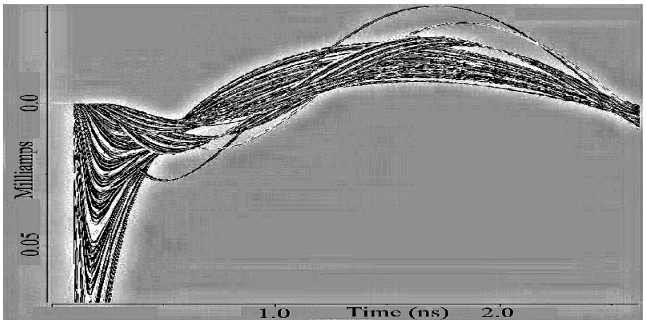

Figure 12. The various initial current responses of Figure 11, showing multiple frequencies higher than the dominant resonant frequencies (mid-frequency effects).

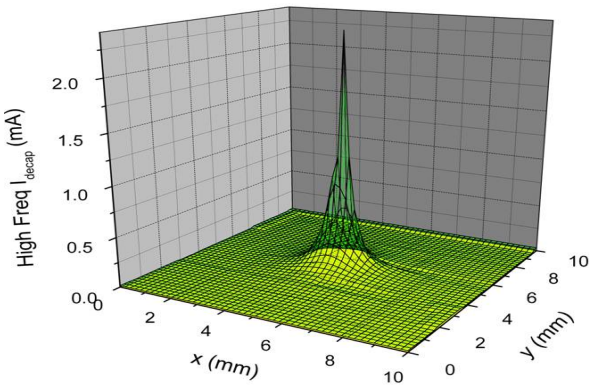

Figure 13. The locality of the mid-frequency currents

A visualized explanation of all of the above effects is shown in Figure 15. When a single gate switches (middle), 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 within a few microns radius before the RC background absorbs the high frequency energy. The high frequency currents are immediately satisfied by caps nearby, either explicit de-caps, active device caps or wire cap. The farther away the cap, the less current will be supplied but the supply radius grows larger as the frequency lowers. At some mid frequency (greater than the global die-package frequency) the package comes into effect and mid-frequency currents are supplied through the C4's. However, even

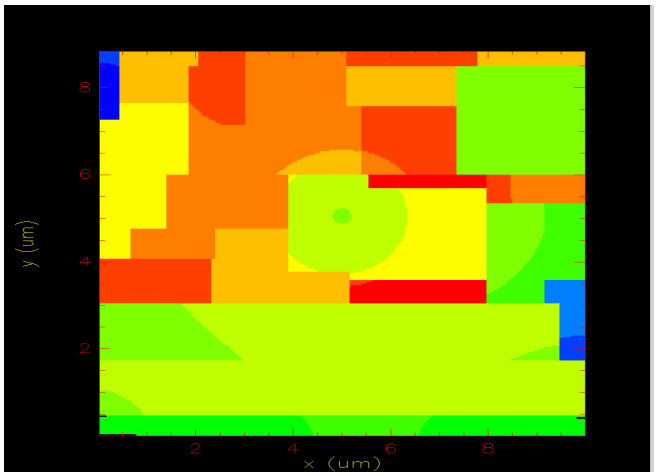

Figure 14. The distribution of currents in the second dip of the response where global convergence has almost been attained. These currents are highly correlated to the full-die capacitance distribution in Figure 10.

at these frequencies, they continue to also be supplied by the caps. If the gate is surrounded by pockets of cap that are resistively isolated (partially, or completely) from other caps further out, the local caps will 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 of the caps and all of the C4's will interact to produce a global resonant frequency that is full-die in nature.

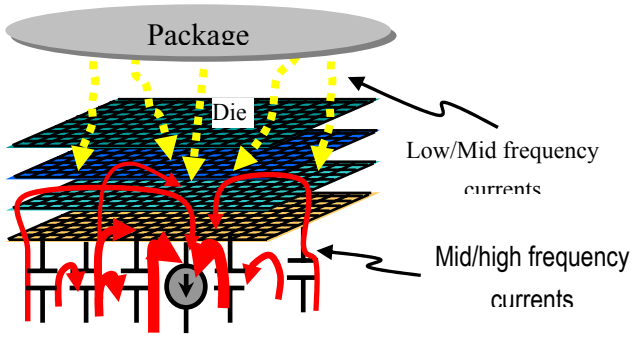

Figure 15. A visualization of the flow of currents on die.

5. Conclusions and CAD implications

In this study, we demonstrated that in an industrial microprocessor design, the package inductance overwhelms the effect of on-die inductance even in 90nm technology and the impact of on-chip inductance remains insignificant and highly localized except at frequencies > 5Hz. It is also shown that the 2-D or lumped inductance models, which are commonly used, significantly overestimate the impact of on-die grid inductance compared to 3D models. The dominance of package inductance over on-die inductive effects necessitates accurate modeling of the package and co-simulation of package and on-die power distribution networks.

The impact of decap distribution on the first and second transient voltage droops was shown in Section 4. A new mid-frequency localized effect was found to be dependent on the nearby non-uniform decap distribution, while the main resonant behavior depends on the total global decap. It was noted that resistively isolated pockets of decaps can lead to multiple resonances. Thus, accurately modeling the decap distribution is extremely important to analyze the transient behavior of power grid accurately.

Due to the distributed nature of C4s in flip-chip packaging, voltage drop induced due to a current excitation may be limited to the vicinity of the current source. Recently, several works [8][9] have been proposed to exploit this locality in power grids to accelerate voltage drop analysis. However, we demonstrated that transient locality is a strong function of the excitation frequency. We showed that although the 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 is dependent on the frequency of excitations as illustrated in Section 3.2. Thus, although locality can be used to simplify and accelerate the static power grid analysis as proposed in [8], its usage for transient power grid analysis and optimization may lead to erroneous results, unless integrated with these effects.

6. Acknowledgements

We would like to acknowledge the feedback of Prof. David Blaauw, Marek Patyra, Kim Eilert, Bob Martell, Kaladhar Radhakrishnan and several others at Intel.

7. REFERENCES

[1] Dharchoudhury A, Panda R, Blaauw D, Vaidyanathan R, Tutuianu B and Bearden D, "Design and Analysis of Power Distribution Networks in PowerPC® microprocessors", in Proc. IEEE/ACM DAC, 19985, pp. 738–743.

[2] Chen H, Ling D, "Power Supply Noise Analysis Methodology for Deep-submicron VLSI Chip Design," in Proc. IEEE/ACM DAC, 1997, pp. 638-643.

[3] Mezhiba A.V and Friedman E.G, "Impedance Characteristics of Power Distribution Grids in Nanoscale Integrated Circuits," IEEE Trans. On VLSI Systems, Vol. 12, No. 11, pp. 1148-1155, November 2004.

[4] Ruehli A.E, "Equivalent circuit models for three dimensional multi-conductor systems", IEEE Trans. Microwave Theory Tech., vol. MTT-22, pp. 216-221, Mar. 1974.

[5] Lin T, Beaftie M.W, Pileggi L.T, "On the efficacy of simplified 2D on-chip inductance models", IEEE/ACM Design Automation Conference, 2002, pp. 757 – 762.

[6] Waizman A and Chee-Yee Chung, "Resonant free power network design using extended adaptive voltage positioning (EAVP) methodology", IEEE Transactions on Advanced Packaging, Vol. 24, Issue 3, Aug. 2001, pp. 236 – 244.

[7] Kozaya J.N, Nassif S.R, Najm F.N, "A multigrid-like technique for power grid analysis", IEEE Trans. on CAD, Vol. 21, Issue 10, Oct. 2002, pp. 1148 – 1160.

[8] Chiprout E, "Fast flip-chip power grid analysis via locality and grid shells", IEEE/ACM ICCAD, 2004, pp. 485 – 488.

[9] Qian H, Nassif S.R, Sapatnekar S.S, "Power Grid Analysis using Random Walks", IEEE Trans. on CAD, Volume 24, Issue 8, Aug. 2005, pp. 1204 – 1224.

[10] Xu M, He L, "An Efficient Model for Frequency-based On-chip Inductance", GLSVLSI, March 2001.