25-27 September 2012, Budapest, Hungary

MiMAPT: Adaptive Multi-Resolution Thermal Analysis at RT and Gate Level

Mohammadsadegh Sadri, Andrea Barolini, Luca Benini
Micrel Lab, DEIS, University of Bologna
{mohammadsadegh.sadr2,a.bartolini,luca.benini}@unibo.it

***Abstract*—Tight timing/area constraints produce on-chip layouts with non-regular shapes for RTL entities. Thus, grid-like floorplans where RTL entities are abstracted as rectangular blocks for thermal simulation lead to inaccurate results. In addition, spatial and temporal variability of chip workload causes localized temperature variations. Exact localization of hotspots at gate-level necessitates an extremely detailed spatial resolution which is very computationally intensive.**

We propose MiMAPT, a tool capable of performing thermal simulation at RT and gate-level with multiple scales of spatio-temporal resolution. To demonstrate the tool advantages we run various tests for a sample chip. We show that our tool provides high level of flexibility in terms of speed vs. accuracy of results.

I. Introduction

High power densities of today's integrated circuits lead to on-chip thermal hotspots which can compromise chip functionality. As [1] describes, the volumetric power density of a 20nm device is on the order of 10 TW/cm^3. Consequently, researchers and CAD vendors are developing tools to facilitate design-time prediction and identification of thermal hazards.

Spatial and temporal variability of chip workload results in non-uniform on-chip power density and localized temperature variations which significantly affect device parameters. [2] Shows an example of this. Considering the above facts, precise localization of hotspots at gate-level is necessary for many high-performance designs. This however requires an extremely detailed spatial resolution for power/thermal simulation which is very computationally intensive and almost impossible for large designs. This is where multi-scale analysis techniques can help. Conceptually, power and temperature estimation can be done at low resolution when it provides satisfactory level of accuracy at a high level of speed. Resolution (and computational effort) should be increased only for areas of interest, where hot spots are likely. The challenge lies in avoiding false negatives (i.e. missing hot spots) while minimizing false positives to achieve significant speedups w.r.t. fine-grained (gate-level) thermal analysis.

We propose MiMAPT (Micrel Multi-scale Analyzer for Power and Temperature), a tool capable of performing power/thermal simulation at RT and gate-level with multiple scales of resolution and speed. The tool starts coarse-grained transient power/thermal analysis at RTL for a user-defined time interval. It considers non-uniform shapes of on-die units during analysis. MiMAPT switches to accurate gate-level simulation only when a likely hot-spot is suspected. At gate-level it performs iterative power/thermal simulation while refining spatial resolution just for the areas which are suspected to contain hotspots.

II. Related Work

We categorize published results from academic researchers into three areas:

- Thermal simulation platforms which estimate chip/package temperature distribution based on a given power and floorplan.
- Power estimation packages, which calculate dynamic and static power of a target architecture based on activity statistics.
- Concurrent power and temperature computation solutions, which mainly mix the capabilities of the above instruments.

For chip/package thermal simulation [3] introduces Hotspot, a boundary conditions independent compact thermal model. Hotspot package assumes that a per-floorplan-block power trace is given as an input. It performs thermal simulation based on a given floorplan and does not perform automatic mesh refinement and iterative thermal simulation based on the results. [4] and [5] introduce ISAC and NanoHeat. The two packages together create a platform capable of performing thermal simulation at different scales of spatial and temporal resolutions. The tool provides two solvers, one based on Fourier heat conduction, and the other based on Boltzman heat transfer equations (BTE)[1].

The multi-scale analysis capability of the mentioned tools is confined to the provided power trace by the user from outside. This is one important motivation that our solution, MiMAPT, performs joint power and temperature calculation at different scales of resolution.

[6] introduces Logi-Therm. The tool performs concurrent electrical and thermal simulation of standard cell ASIC circuits. It takes standard cells of digital design as basic building blocks and based on cell's power characteristics and switching activity calculates a power/thermal distribution map of the chip. Analysis done by Logi-Therm however are based on a fixed mesh resolution.

[7] introduces ICTherm, a tool developed for thermal analysis of 3D chips. The tool first performs a thermal evaluation of the target with a very high spatial resolution. Based on the results it creates a multi-granularity mesh which is used for thermal simulation. The mesh however is fixed during the rest

 ISBN:978-2-35500-022-5

of the analysis. Instead, MiMAPT changes the mesh dynamically based on the current on-die temperature distribution.

Power estimation is an extensively explored research area: Wattch [8] and SimpleScalar [9] provide basic power models for CPU cores. McPAT [10] extends this capability to wider range of architectures including multi-core clusters with network-on-chips, DRAM-controllers and high-speed interfaces. The above tools however, do not provide power at different levels of granularity. Moreover power estimation is based on pre-defined full-custom architectures thus, in practice these tools do not produce accurate enough results for newly designed hardware and for an ASIC implementation style.

[11] provides a tool for concurrent power, performance and temperature estimation. The tool however works at micro-architecture level and not RTL and so it does not provide enough accuracy for power and temperature estimation especially with new hardware blocks.

As of the author's best knowledge, no academic package, meets the following requirements together:

1- Power and temperature estimation at RT and Gate level with different scales of spatial resolution.
2- Handling non-uniform shapes of design sub-modules for thermal simulation.
3- Seamless integration into major design tool flows and compatibility with widely used library standards.
4- Compatibility and open interfaces with commercial and academic thermal analysis tools (such as Hotspot[3], 3D-ICE[12] and FloTHERM[13]).

Considering commercial software packages, Gradient Design Automation is known to be able to perform concurrent power/thermal analysis with multiple scales of temporal and spatial resolution in transient and steady state [14], [15]. Compared to Gradient's solution, MiMAPT provides wider range of analysis speed and accuracy, since it basically performs thermal analysis at RT level, which is very fast, and switches to gate-level only when higher levels of resolution are demanded.

III. MiMAPT Architecture

Our approach leverages power analysis features provided by state-of-the-art commercial tools. Recent versions of logic synthesis tools (e.g. Cadence RC® and Synopsys DC®) are capable of estimating power at RTL before doing synthesis based on switching activity obtained from functional logic simulation. RTL power estimation can be done very fast, but it is not very accurate[16], [17]. On the other hand, very accurate power at gate level can be obtained after (or during) back-end flow. The power analysis tool should be provided with the finished (placed, routed, clock tree synthesized) design and also switching activity statistics obtained from logic simulation of the finished netlist with timing delays annotated through an SDF file.

In classical thermal simulation flow the power estimation at gate-level is used to obtain per cell power values. This fine-grain power map is then used as input to fine-grain thermal simulation [5], [3]. MiMAPT instead performs power/thermal simulation at RT level with the goal of avoiding

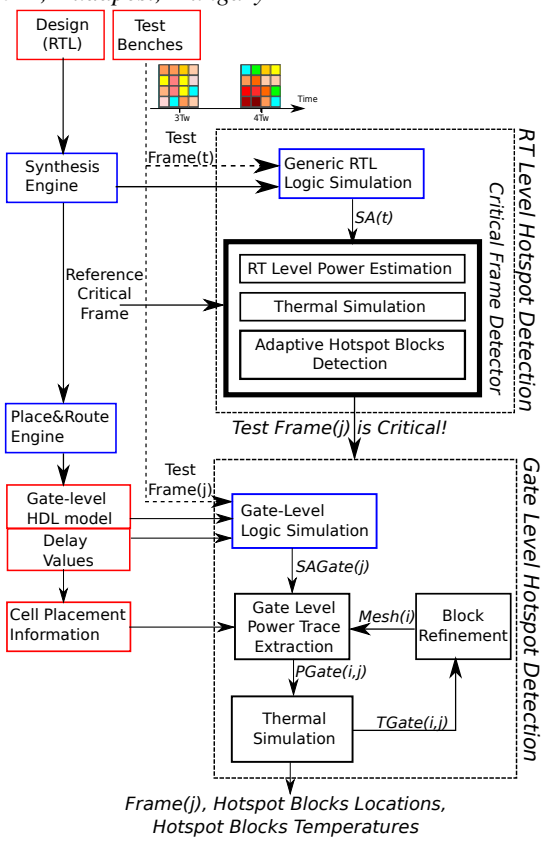

Fig. 1. MiMAPT Block Diagram

time consuming gate-level simulation when hotspots do not exist. This is achieved by using faster but less accurate, RT level power/thermal simulation to filter out non-critical (non-hotspot) portions of a die floorplan. A die area is defined as hotspot when its temperature (T) is higher than a specified threshold (TH). Figure 1 shows the basic building elements of MiMAPT which are two major parts: RTL and gate-level.

We partition the simulation input sequence in subsequences, called *test frames*, typically representing different use cases in a real design. We then dynamically switch to gate-level for any arbitrary *test frame* if needed. We use the RTL state to initialize the gate-level simulation and prepare the input patterns for the *test frame* to simulate at gate-level.

A. RT Level Hotspot Detection

We perform logic simulation for each of the *test frames* at RTL to obtain switching activity ($SA(t)$). This information are then fed to the synthesis tool to obtain power estimation for each of the design sub-modules.

We define a new *thermal floorplan* for the design which divides chip area into equally sized rectangular blocks. For each floorplan block (FB), we calculate what percentage of each sub-module is located inside this block. Based on the percentage we add a fraction of sub-module's power to the total power of the block.

Figure 2 shows this procedure in more detail for a sample design. In this figure, *(a)* shows the defined floorplan for the

 ISBN:978-2-35500-022-5

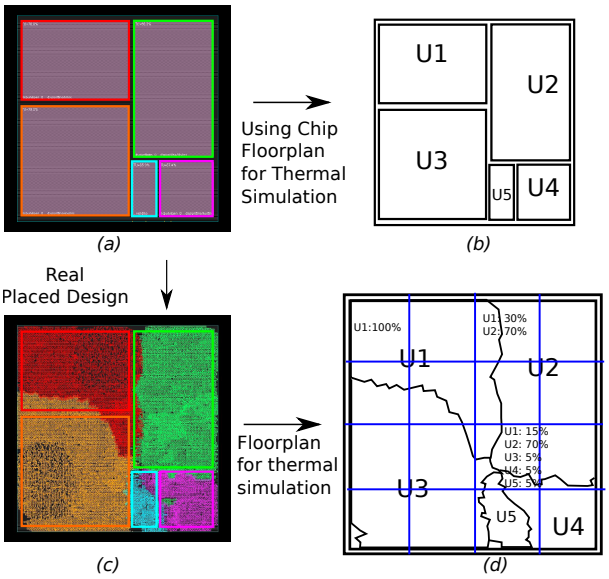

Fig. 2. Generation of power map for thermal simulation at RTL

chip in the layout tool. *(b)* shows the traditional method of creating *thermal floorplan* in which we use the dimensions defined by the default chip floorplan directly. *(c)* shows the chip after placement, as we can see, the real placement of sub-modules is different than the default floorplan thus an accurate thermal simulation is not possible using traditional methods. In *(d)* we show our method of defining a new *thermal floorplan*. Each floorplan block contains one or more design sub-modules. The synthesis tool provides us with the power consumption of each sub-module. Using chip placement information that we have, we obtain the percentage of each sub-module inside each floorplan block by counting the number of cells owned by this sub-module inside the floorplan block. We suppose that the sub-module's power is uniformly divided between its cells thus, we use the obtained percentage to add the fraction of sub-module's power to the total power of the floorplan block.

Final obtained power map is then used by the thermal simulator[3] to estimate per-block temperature map. This thermal map is then compared with a set of adaptive thresholds to identify if the *test frame* contains critical areas. If this situation is detected we trigger the gate level hotspot detection for the same frame. As shown in experimental results, RTL hotspot detection executes significantly faster than the gate-level simulation.

Estimated power at RT level is usually not equal to gate-level power since the design is not fully synthesized yet. Consequently, using a unique temperature threshold value to identify hotspot blocks at RT and gate-level may not lead to accurate detection of hotspots at RTL. Thus we use an adaptive method to detect hotspots at RTL.

We first select the *test frame* in which every design sub-module has highest level of activity and thus power, as reference. We perform thermal simulation for this *test frame* at RTL and gate-level and we mark critical floorplan blocks (FB) by comparing their temperature values at gate-level ($CMapGate$) to a threshold ($THigh$) which contains a safe margin with respect to TH and is slightly lower (e.g. 1°C) than it. We use the computed critical blocks map ($CritMatrix$) for identifying hotspots at RT level for the rest of *test frames*.

For each *test frame* we perform thermal simulation at RTL and we compare each block temperature value with an adaptive threshold. If the block is marked as critical, we act more carefully, thus we create a reduced threshold value by multiplying a coefficient A smaller than 1.0 to the reference block temperature ($CMapRTL$). If the block is not critical we use an increased threshold value by multiplying a coefficient B bigger than 1.0 to the reference block temperature to avoid un-necessary hotspot detection. The smaller values for A make detection of hotspots at RT level safer however, they decrease overall operation speed. Psudocode 1 describes this in detail.

Algorithm 1 Adaptive Hotspot Detection

Require: CMapRTL, CMapGate= Highest-workload frame, temperature map at RTL and gate-level
Require: TMapRTL= current frame temperature map
Require: THigh= threshold value for critical block
Require: A, B: Coefficients ($A < 1.0$) and ($B > 1.0$)
Require: N= Total number of thermal floorplan blocks
for $i = 1 \rightarrow N$ **do**
  $t = CMAPGate(i)$
  if $t > (THigh)$ **then**
    $CritMatrix(i) = 1$
  else
    $CritMatrix(i) = 0$
  end if
end for
for $i = 1 \rightarrow N$ **do**
  $t = TMapRTL(i)$
  if $CritMatrix(i) == 1$ **then**
    if $t > A \times CMapRTL(i)$ **then**
      This Block is hotspot!
    end if
  else
    if $t > B \times CMapRTL(i)$ **then**
      This Block is hotspot!
      $CritMatrix(i) = 1$
    end if
  end if
end for

B. Gate Level Hotspot Detection

If *test frame* (j) is detected as critical in RTL hotspot detector, it will be processed with high accuracy at gate-level to correctly estimate the hotspot position and temperature. This is done by first performing logic simulation at gate-level to obtain circuit switching activity ($SAGate(j)$) used for power estimation. The power map ($PGate(i, j)$) is then converted to temperature ($TGate(i, j)$) using an iterative multi-granularity

 ISBN:978-2-35500-022-5

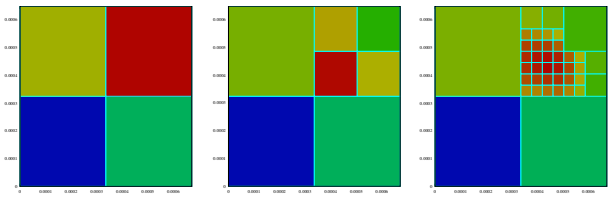

Fig. 3. Increasing spatial resolution of thermal analysis for the area of interest

meshing scheme. Starting from an initial mesh granularity ($initL$) in each iteration (i) we increase spatial resolution for on-die areas which are suspected to contain hotspots. Indeed, in each iteration, we examine the temperature map for the current thermal floorplan. For every floorplan block with a temperature value higher than TH we break the block into $M \times M$ equal sized smaller blocks. The process will continue until the finest spatial granularity ($finL$) is reached. $initL$, M and $finL$ are constants mainly defined by the user. They should be selected according to the chip die area and the desired spatial resolution and the accuracy with which detection of hotspots should be done. Figure 3 shows an example output of our multi-granularity thermal simulation for three continuous iterations.

IV. Experimental Results

We first evaluate our adaptive hotspot detection algorithm at RTL. Based on real power values, we create an ensemble of virtual gate-level and RTL power maps. RTL power maps are obtained by adding Gaussian random variables to per-floorplan block power values at gate-level. We change random variable's characteristics to simulate different situations of RTL power estimation.

Figure 4 shows total power values for 180 different virtual test cases that we have. For each power map at gate-level we create 10 different power maps at RTL by changing the mean of the Gaussian random variable ($\mu = \{-0.2, -0.1, 0.0, 0.1, 0.2\}$). We compare the output of thermal simulation for all of the gate-level and RTL power pairs. For each pair we obtain the number of hotspots and their locations at gate-level and we compare it with the output of hotspot detection methods at RTL.

Figure 5 shows the performance of our adaptive method (*A-Temp*) compared to using a unique user defined threshold value (*TH Only*). In this figure, *(a)* shows the percentage of situations that hotspots exist in the chip and gate-level simulation should be triggered, compared to percentage of situations that each of *A-Temp* and *TH Only* trigger gate-level simulation. *(b)* shows percentage of cases in which the estimated spatial location of hotspot by each of *A-Temp* and *TH Only* is different than its estimated location at gate-level. As we can see, *A-Temp* estimates the spatial location of hotspots correctly in all of the situations. Finally, *(c)* shows the percentage of detected false positives and false negatives for each of *A-Temp* and *TH Only* methods considering all of the 180 test cases. Different than *TH Only*, false-negative for our method is equal to zero which means it captures all of the hotspots completely.

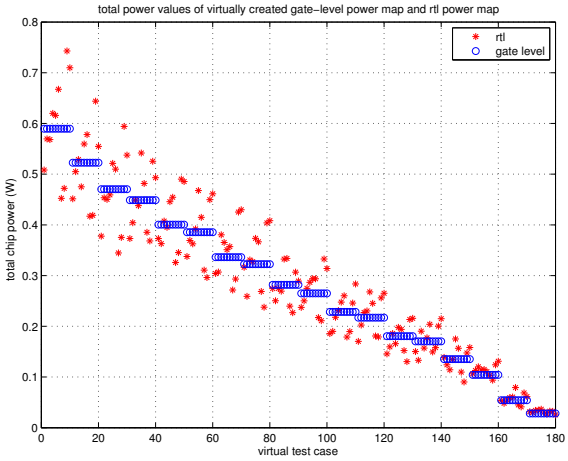

Fig. 4. virtual power values at RTL and gate-level used for evaluation of adaptive method.

TABLE I
Sample Chip : AES, FFT and FPU. (3 separate clock domains).

Block	Area(mm^2)	#Cells	FFs	ClkBuf	F(MHz)
FPU	0.2730	41477	499	13	143
FFT	0.6997	81651	42684	875	525
AES	0.4758	110758	7882	167	1328
Top	1.49	233887	51065	1055	-

To evaluate MiMAPT for a real test case, we create a sample chip containing 3 widely used digital IP blocks (AES, FPU and FFT), and fully perform synthesis, placement, CTS and routing using TSMC 65nm standard-cell library. The results are based on typical corner case of $VDD = 1.2V$ and $T = 25°C$. Table I shows chip's key specifications.

Six different *test frames* are created to evaluate MiMAPT. Figure 6 shows for each *test frame*, total power of the design

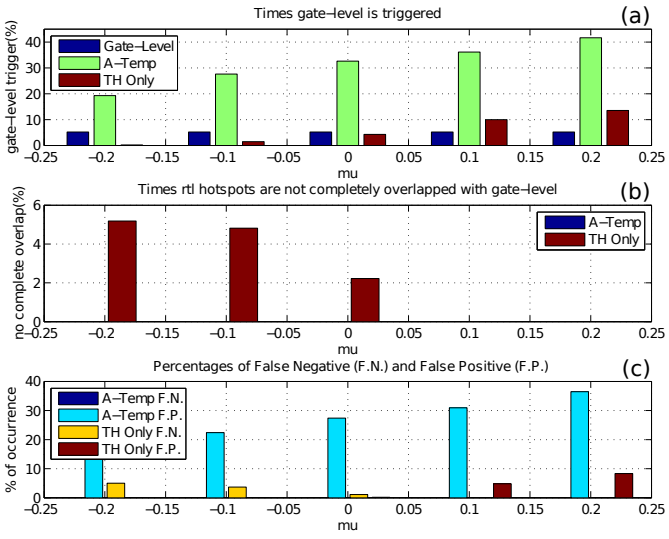

Fig. 5. hotspot detection methods comparison (Adaptive vs. TH only)

 ISBN:978-2-35500-022-5

estimated at RTL and gate-level. Duration of each *test frame* is 0.2 seconds.

We execute MiMAPT for all *test frames* and we store spatial and temporal information related to detected hotspots. We then perform power and temperature estimation of the design without MiMAPT for the same set of *test frames* at gate-level and at the finest level of granularity. We perform comparison between MiMAPT and fine-grain simulation for a sample value of $TH = 358K$.

The finest level of granularity ($finL$) is $50um$ and fine-grain floorplan contains 24x24 (total 576) blocks. RTL floorplan and initial floorplan at gate-level are 8x8 blocks ($initL$). For increasing spatial resolution of the multi-granularity mesh at gate-level, we divide each hotspot block into 3×3 equal sized smaller blocks ($M = 3$). A and B coefficients in RTL hotspot detection are 0.9 and 1.1 respectively.

We represent the results in terms of execution time and accuracy of detecting hotspot location and temperature. Among 6 available *test frames*, for 3 of them MiMAPT detects critical blocks at RTL and triggers gate-level, for the other 3, gate-level is not triggered saving time. The difference between estimated temperature by MiMAPT and fine-grain is around 0.02K. For every hotspot block at fine-grain, there exists a corresponding block in MiMAPT which has the same location and size and is announced as hotspot. As a result, the distance between location of hotspots detected by MiMAPT and fine-grain is zero. MiMAPT detects all of hotspots with a very good level of accuracy, thus there is no false negatives. When gate-level is triggered MiMAPT performs two iterations of thermal simulation to achieve required spatial resolution of $50um$. For the first iteration ($i = 1$) the thermal floorplan contains 64 and for the second one ($i = 2$) 168 blocks.

Table II shows the execution time of MiMAPT compared to fine-grain. Considering all the six *test frames*, total execution time at fine-grain is 26520 seconds, it includes required time for gate-level simulation and obtaining switching activities ($Tgate_{sim} = 1610s$) and time for fine-grain thermal simulation ($Tgate_{finthr} = 24910s$). In contrast, total execution time for MiMAPT is 1446 seconds. It contains the time for RTL logic simulation and obtaining switching activities ($Trtl_{sim} = 83s$), time for RTL thermal simulation ($Trtl_{thr} = 72s$) and time for gate-level simulation for each of the 3 detected critical test frames ($Tgate_{sim} = 908s$). Two of the *test frames* are false-positives thus they do not have hotspots at gate level and gate-level thermal simulation will be done for them only at coarse grain ($Tgate_{thrc} = 24s$). One of the test frames is critical and thus contains hotspots at gate-level and triggers MiMAPT multi-granularity engine ($Tgate_{thrm} = 335s$). In total, MiMAPT is 18.3X times faster than fine-grain at the same level of accuracy.

In order to provide a better perspective on the range of possible speedups for MiMAPT, we calculate average MiMAPT time for one *test frame* when it is either a non-critical, false positive or a critical frame. As shown in table III, for our sample chip, for an assumed experiment in which every *test frame* is non-critical, MiMAPT reaches the maximum speedup of 171X. However when every *test frame* is false-positive, the speedup decreases to 13X. Finally if every *test frame* is critical the speedup reaches a lower bound of 7X.

TABLE II
EXECUTION TIMES COMPARISON: MIMAPT VS FINE-GRAIN. (TIME IN SECONDS)

Method	RTL logic Sim	RTL Thermal Sim	Gate-level logic Sim	Gate-level thermal Sim	Total
Fine-grain	-	-	1610	24910	26520
MiMAPT	83	72	908	359	1446

TABLE III
AVERAGE MIMAPT SPEED-UP OVER FINE-GRAIN FOR A GENERIC TEST FRAME FOR DIFFERENT FRAME TYPES.

Test frame Type	MiMAPT Time RTL	MiMAPT Time Gate	Example SpeedUp
Non-critical	$Trtl_{sim}$ $+Trtl_{thr}$	-	171X
False positive	$Trtl_{sim}$ $+Trtl_{thr}$	$Tgate_{sim}$ $+Tgate_{thrc}$	13X
Critical	$Trtl_{sim}$ $+Trtl_{thr}$	$Tgate_{sim}$ $+Tgate_{thrc} + Tgate_{thrm}$	7X

V. CONCLUSIONS

We proposed MiMAPT, and described its approach to power/thermal simulation at RT and gate level. MiMAPT highly accelerates power/thermal simulation while keeping accuracy at acceptable levels. For the developed sample chip, we observed an speed-up range between 7X to 170X (with a typical value of around 18X) over classical method while providing the same level of hotspot detection accuracy.

ACKNOWLEDGMENT

This work was supported, in parts, by the EU FP7 ERC Project MULTITHERMAN (GA n. 291125) and the EU FP7 Project THERMINATOR (GA n. 248603).

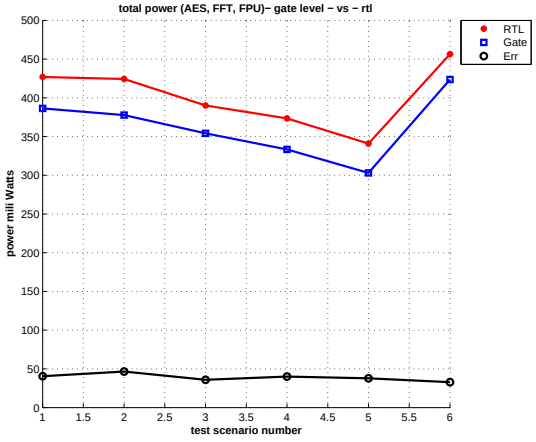

Fig. 6. Total power of each test case at RT and gate-level

 ISBN:978-2-35500-022-5

REFERENCES

[1] J. Rowlette *et al.*, "Fully coupled nonequilibrium electronphonon transport in nanometer-scale silicon fets," *IEEE TRANSACTIONS ON ELECTRON DEVICES*, vol. 55, no. 1, pp. 220–232, January 2008.

[2] M. Sadri, A. Bartolini, and L. Benini, "Single-chip cloud computer thermal model," in *Thermal Investigations of ICs and Systems (THERMINIC), 2011 17th International Workshop on*, sept. 2011, pp. 1 –6.

[3] W. Huang *et al.*, "Hotspot: a compact thermal modeling methodology for early-stage vlsi design," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 14, no. 5, may 2006.

[4] Y. Yang *et al.*, "Isac: Integrated space-and-time-adaptive chip-package thermal analysis," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 26, no. 1, jan. 2007.

[5] Z. Hassan *et al.*, "Full-spectrum spatial-temporal dynamic thermal analysis for nanometer-scale integrated circuits," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 12, dec. 2011.

[6] A. Timar *et al.*, "Electro-thermal co-simulation of ics with runtime back-annotation capability," in *Thermal Investigations of ICs and Systems (THERMINIC), 2010 16th International Workshop on*, oct. 2010.

[7] A. Fourmigue *et al.*, "Multi-granularity thermal evaluation of 3d mpsoc architectures," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2011*, march 2011.

[8] D. Brooks *et al.*, "Wattch: a framework for architectural-level power analysis and optimizations," in *Computer Architecture, 2000. Proceedings of the 27th International Symposium on*, june 2000.

[9] T. Austin, E. Larson, and D. Ernst, "Simplescalar: an infrastructure for computer system modeling," *Computer*, vol. 35, no. 2, feb 2002.

[10] S. Li *et al.*, "Mcpat: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Microarchitecture, 2009. MICRO-42.*

[11] M.-y. Hsieh *et al.*, "Sst: A scalable parallel framework for architecture-level performance, power, area and thermal simulation," *The Computer Journal*, 2011.

[12] A. Sridhar *et al.*, "3d-ice: Fast compact transient thermal modeling for 3d ics with inter-tier liquid cooling," in *Computer-Aided Design (ICCAD), 2010 IEEE/ACM International Conference on*, nov. 2010, pp. 463 –470.

[13] M. Graphics, *FloTHERM: Optimizing the Thermal Design of Electronics*, May 2011.

[14] G. Design, "Us patent 7823102, thermally aware design modification," Tech. Rep., 2011.

[15] ——, "Us patent 8019580, transient thermal analysis," Tech. Rep., 2011.

[16] Cadence, *Low Power in Encounter RTL Compiler*, Cadence Design Systems, April 2010.

[17] Synopsys, *Design Compiler User Guide*, Synopsys, December 2011.

 ISBN:978-2-35500-022-5