ANSYS White Paper

Differential Energy Analysis to Optimize Mobile GPU Power

Operating power has become one of the most important metrics for modern electronic devices. This is true for mobile and remote applications because operating power directly affects battery life; for appliances because power consumption directly contributes to cost of use; and for datacenters because the cost of cooling is now as much as 40 percent of total power cost. More generally, regulatory requirements and global warming concerns demand we reduce power consumption as much as possible. The explosion of electronic device usage with the advent of 5G, artificial intelligence (AI) and machine learning, autonomous vehicles and the Internet of Things (IoT) makes reducing power consumption an even more urgent concern. Each of these applications has more capabilities, often with higher performance, yet their power budgets are even more constrained.

Figure 1. Mobile gaming

The problem is especially noticeable in mobile devices when running "free" apps such as Angry Birds. Playing the game consumes power, but so does background location tracking, adware and other functions the app supplier adds to collect data or generate revenue. The companies that build the devices at the heart of our smartphones and tablets, such as Qualcomm Technologies, are typically the most aggressive in managing power consumption. They do this in part by using the most advanced semiconductor processes (such as FinFETs), which —together with power-island techniques — greatly reduce the leakage component of power draining. But these methods don't help with dynamic power, which has now become the dominant component of power consumption. Figuring out how to reduce this power drain requires detailed dynamic power analysis against realistic workloads.

Why Power Analysis at RTL?

Power consumption is a challenging factor to manage in device design. The most accurate pre-manufacturing estimates of power consumption are available near the end of the design process, yet at this stage there is very little that can be done to correct the design if the power target is missed. At most, you might be able to tweak the design to reduce power by a few percent at this late stage.

Big reductions in power require changes in system architecture or in micro-architecture (RTL) as early in the design process as possible. The challenge here is that the earlier the power estimation is made, the less accurate it is, because implementation details are not yet clear. Less accuracy means less certainty in the effectiveness of any power fixes you might make.

The generally agreed balance-point in this trade-off is to focus optimization at RTL. Absolute accuracy compared with premanufacturing signoff is still quite good, typically within 15 percent. Relative accuracy, which compares power in two or more very similar cases — such as the same design before and after applying a power fix — can be within 5 percent, similar to power signoff accuracy. The reason is quite simple: in comparisons, many of the implementation details will largely cancel, at least around those areas where RTL fixes are typically considered.

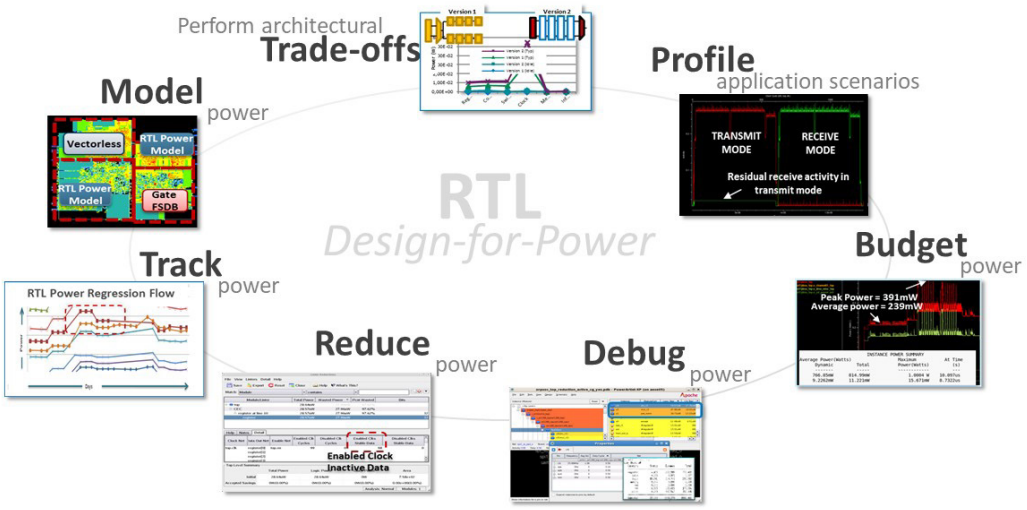

Figure 2. PowerArtist capabilities

ANSYS PowerArtist is the leading platform in the industry for power analysis at RTL; this position has been gained over many years of development and optimizing to evolving power needs. Consider for example power profiling, in which you want to study how design power varies during practical usage. Traditional power methodologies are typically limited to sampling design activity over a few microseconds, which is simply not representative of realistic use cases today. PowerArtist provides the industry's fastest per-cycle analysis of real-world use-cases (such as OS boot-up) within a few hours, orders of magnitude faster than standard approaches.

In addition to power profiling, PowerArtist supports multiple goal-oriented power analyses at RTL. These include support for comparing trade-offs (such as different micro-architectural implementations), as well as support for power budgeting and power-debug to trace power problems to root causes. The tool provides power trend analyses, useful in regression monitoring as a design progresses. This helps, for example, to detect potential problems when an otherwise innocuous functional fix suddenly blows up power. All these and other PowerArtist capabilities are supported by strong data visualization techniques and data mining features, to accelerate analysis and debug, and to help you build your custom analytics.

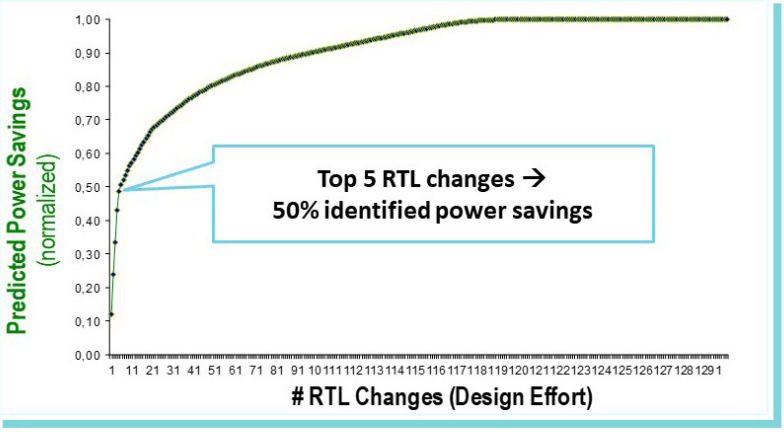

Figure 3. Rank-ordered power-reduction options typically exhibit a small number of changes with major impact.

A particular strength is PowerArtist's analysis-driven automated power reduction, which starts by identifying all wasted activity in the design of clocks, memories and data path logic. Based on this analysis, PowerArtist identifies high-impact clock, memory and data gating opportunities at hierarchical function and instance levels. These techniques look at combinational and sequential methods to identify new clock enables, redundant memory accesses and redundant activity in cones of logic. Importantly, PowerArtist can estimate the impact of these optimization techniques as a part of its analysis, before you implement the changes. You get to see exactly which changes you want to make and how much they will save before committing to changing and simulating the design again.

Reduction suggestions are based on production-proven and physically-aware analysis to ensure that identified RTL optimizations are limited to those that will minimize design impact (such as timing).

Another important capability, unique to ANSYS, is support for RTL-based power grid integrity planning. The power delivery network (PDN) must work within acceptable margins across all real use-cases. Analyzing across the extensive use-case profiles that can be handled at RTL, PowerArtist will isolate a power-critical subset and generate a unique RTL Power Model to interface with ANSYS RedHawk. This gives the PDN design team the information they need early enough to help optimize the power delivery network with confidence that use-case coverage is solid, avoiding late-stage corrections to the PDN.

All of this analysis depends on ensuring that RTL power estimation is as accurate as possible, despite the lack of implementation details. Combining the design RTL with extensive activity files in any of the standard formats, PowerArtist will perform a micro-architectural synthesis, mapping to the power parameters in standard Liberty models and proprietary PACE models for wire capacitance, mitigating the impact of early analysis on accuracy.

Clock tree power is an especially important consideration in accurate estimation, since power consumed in the clock tree typically contributes a significant percentage to overall power. PowerArtist uses advanced proprietary techniques to model the clock tree that will ultimately be synthesized in the implemented design. These methods ensure that clock tree power estimation will be comparable to logic power estimation, and that power-reduction suggestions in the clock tree will be made based on similar accuracy.

A Case Study from Qualcomm Technologies

Qualcomm Technologies presented "Differential Energy Analysis for Improved Performance/Watt In Mobile GPU" at DAC 2018. The presentation focused on their use of PowerArtist to optimize power in a mobile GPU. They opened by noting that the impact of power on heating is a big challenge in mobile GPUs. As you play a game on your phone, the temperature rises. Eventually thermal mitigation kicks in and the processor automatically drops the clock speed. This is a common trick in processors; if temperature approaches a dangerous level, the device automatically reduces the clock speed to reduce power and therefore heating. Unfortunately, your game now runs slower. This clock-speed downgrade can go through several steps, so the longer you play, the slower the game runs (down to some limit), which doesn't make for great customer satisfaction. This is why thermal-constrained performance is becoming one of most important key performance indicators (KPIs) in mobile design.

This is a dynamic power problem. Assuming the design team has done all they can to minimize leakage (through process selection and power islands), they have to direct most of their attention to minimizing redundant activity. Qualcomm emphasized that RTL is the low-cost design phase for performing design changes because they can iterate quickly on different options; the impact of changes made here is more significant than anything they can do in implementation. Qualcomm thus confirmed ANSYS' view of the importance of power analysis at RTL.

Qualcomm Technologies took an ingenious approach to finding inefficiencies in their GPU. Rather than looking directly at redundant switching, they compared the energy (power integrated over run-time) in the design with the equivalent measure for a slowed-down version of the design. They simulated slowing down by adding latencies, to mimic starvation or stalls for example, through scripted addition of NOOP operations in ALUs, modifications in the testbench or through other techniques. One clever aspect of this analysis is that they are simply comparing the design with itself — no fixes have yet been made, so concerns over the accuracy of RTL mitigation are further reduced.

In a simplified view, this analysis leads to the energy diagrams shown in Figure 4. On the left is the ideal case. The unmodified test (the blue bar) runs more quickly at some average power level. The slowed down test (the yellow bar) runs for a greater time period thanks to added latencies, but at the same clock frequency. If the design is optimally clock-gated, average power will be lower because you are running the same workload but over a longer time. Therefore total energy for the run should be the same as in the unmodified case.

However, if there are any gating inefficiencies in the design, the energy comparison will look more like the case on the right side in figure 4. Redundant toggles in the design will be active over a longer period in the modified run, and therefore the integrated energy in that run will be higher than in the original run. This is interesting both as a different way of looking at the impact of inefficiencies on the design and as an energy metric that is especially relevant in battery-powered applications like this.

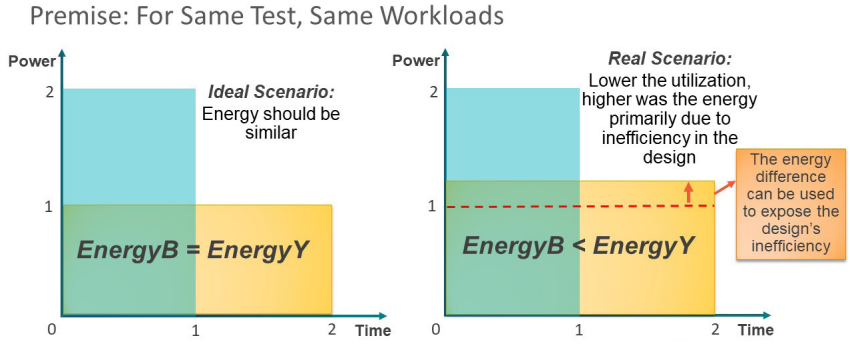

Figure 4. Energy comparison – a different way to find inefficiencies

Qualcomm Technologies took this further. PowerArtist breaks out (at each level) switching and internal energy contributions separately, in addition to total energy. Internal energy is the energy dissipated inside gates such as registers, whereas switching energy is the energy associated with the interconnect between gates. They noted that redundant data input or output toggles on a register will, in the modified case, cause an increase in both switching and internal energy, whereas redundant toggles on the clock input will increase only internal energy.

In comparison with the unmodified analysis there are 4 possibilities, as shown in Figure 5. If there is no difference in either internal or switching components, optimization is ideal. In the other cases, it is easy to determine where there must be redundant activity. They can use this analysis to drill down to find candidate blocks for more detailed analysis, including individual registers where fixes could have a big impact.

What is especially encouraging is that Qualcomm Technologies was able to reduce dynamic power by 10 percent through this approach. This is in a company (and a market) where reducing power is already an obsession. The improvement was based purely on register toggle optimization. They plan to apply similar techniques to analyzing memories, the clock tree and even combinational logic to squeeze energy down even further.

Four register's internal/switching differential energy scenarios:

ID	Internal Energy	Switching Energy	Description
1	→	→	• No extra toggles. Energy is efficient.
2	↑	→	• D pin has no extra toggles during bubbles • Extra toggles on clock pin when data is stable
3	→	↑	• Extra data toggles on D/Q pins when clock is off
4	↑	↑	• Extra toggles on both D/Q pins and clock pin

Figure 5. Four scenarios in comparing modified versus unmodified energy

Are there Risks?

Any method to reduce dynamic power in a design will involve making (small) changes to the design. These RTL-based methods do not remove the need for judgement in making those changes, but they do provide good guidance for that judgement. PowerArtist will give you a list of opportunities and suggested improvements to clock enables, ranked by estimated impact. Should you implement all of them? Almost certainly not. Many of the lower items in the list will offer only small savings, and you should remember these are all subject to the relative estimation error, so you really only want to consider those first few suggestions offering the largest net savings.

Similarly, some suggestions may offer savings but only using a complex enable signal which, if implemented, may create problems in timing closure. Some may result from not having considered important use-cases in the sample workloads. So the suggestions offered should not be implemented blindly; they should be reviewed by the design team (and perhaps the verification team) who can decide which changes they want to implement and which they will discard as potentially risky.

If you only need to squeeze design power by a couple of percent, perhaps you don't need this kind of analysis. But if you need to reduce power by 10 percent or more, you really don't have a choice. RTL-based power reduction, as provided in PowerArtist, may be your only option short of significant redesign. Qualcomm Technologies and others who depend on meaningful reductions will agree with these caveats. But even then, they still find enough value in the high-ranked suggestions to build this methodology into their design flows.

Summary
That a world-class mobile solution provider like Qualcomm Technologies could find more opportunities to squeeze out energy in an already challenging market is a good indicator of the value of power reduction at RTL. They are not alone; many companies who are pressed to reduce power for competitive or regulatory reasons are taking similar approaches because, for them, it is not an option to wait until just before manufacturing to find out whether or not they will hit their power target. Perhaps your company should take a closer look, too.

Click the link to learn more about ANSYS Power Artist:

ansys.com/products/semiconductors/ansys-powerartist

ANSYS, Inc.
Southpointe
2600 ANSYS Drive
Canonsburg, PA 15317
U.S.A.
724.746.3304
ansysinfo@ansys.com

If you've ever seen a rocket launch, flown on an airplane, driven a car, used a computer, touched a mobile device, crossed a bridge or put on wearable technology, chances are you've used a product where ANSYS software played a critical role in its creation. ANSYS is the global leader in engineering simulation. We help the world's most innovative companies deliver radically better products to their customers. By offering the best and broadest portfolio of engineering simulation software, we help them solve the most complex design challenges and engineer products limited only by imagination.
Visit www.ansys.com for more information.