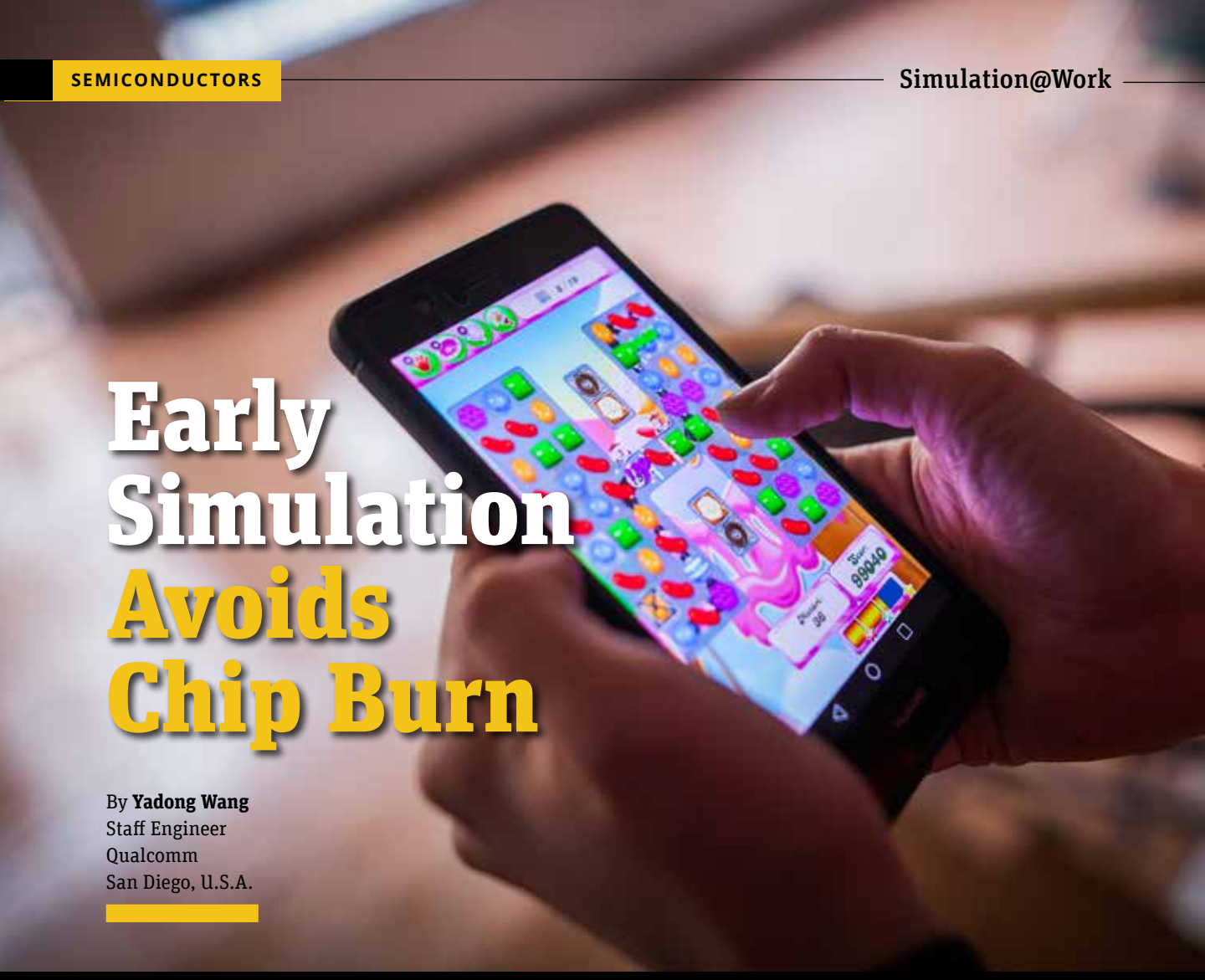


Early Simulation Avoids Chip Burn

By **Yadong Wang**
Staff Engineer
Qualcomm
San Diego, U.S.A.

Thermal constrained performance is a challenge for GPU designs. Using ANSYS PowerArtist to perform a unique differential energy analysis early in the chip design process (during RTL design), Qualcomm engineers were able to identify and fix redundant switching in their GPU to improve the power efficiency of key design blocks by 10%.

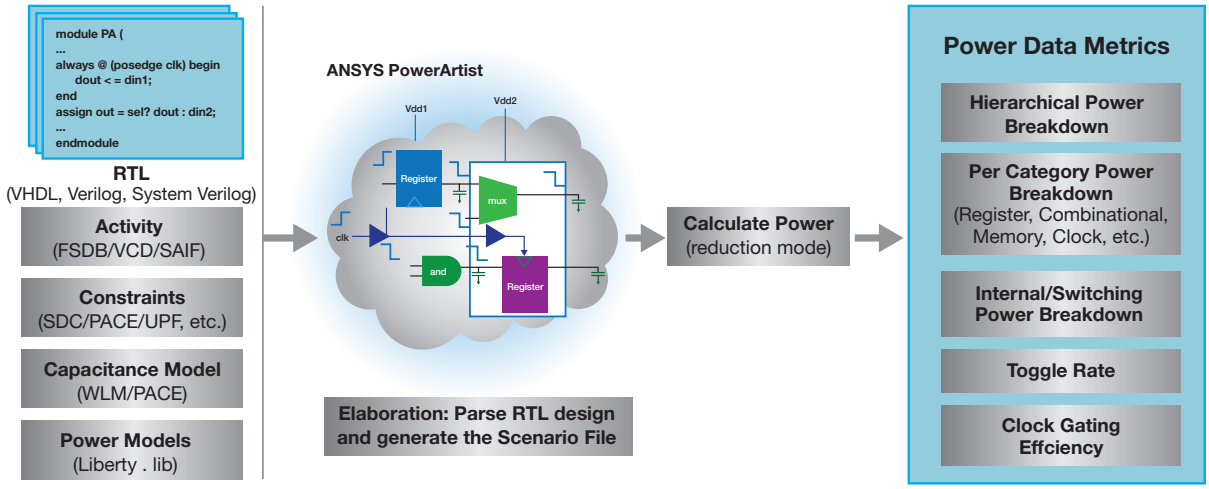
Smartphone and tablet manufacturers are continually changing their designs, searching for advantages over competitors' offerings. Each new model can do more, quicker, with longer battery life. At the same time, applications and background functions consume increasingly larger amounts of power.

Engineers at Qualcomm, a global leader in mobile technologies, are always exploring ways to improve the performance of semiconductor components in mobile devices. The graphics processing unit (GPU), in particular, is a critical component for consumer applications such as gaming. Imagine a consumer playing a video game on a phone. The faster the GPU and the longer the game goes on, the more the GPU

IO	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 stable
3	→	▲	• Extra data toggles on D/Q pins when clock is off
4	▲	▲	• Extra toggles on both D/Q pins and clock pin

Methodical approach pinpoints redundant register pin toggles by investigating four scenarios.

RTL-Based Power Flow



RTL-based power efficiency enables early and reliable design decisions.

"Qualcomm's differential energy analysis early in the design flow using ANSYS PowerArtist delivers 10% higher performance per watt."

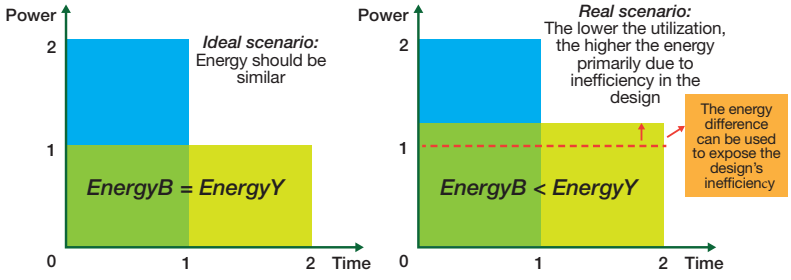
dissipates power, causing the temperature of the phone to increase. At some point, the phone automatically reduces the clock speed (within reasonable limits) to cool itself by reducing power dissipation. But this causes the game to slow down. While annoying, these slowdowns are part of the phone's design. Such thermally constrained performance is becoming a key performance indicator in GPU design.

Instead of just living with these slowdowns, Qualcomm is doing something about them. Using ANSYS PowerArtist simulations to perform differential energy analysis of GPUs early in the development process, at the register transfer level (RTL) when the microarchitecture is being determined, optimizes the power efficiency of GPUs and keeps device temperature down.

EARLY RTL POWER ANALYSIS

Qualcomm selected ANSYS PowerArtist for power analysis and reduction at RTL because of its realistic approach to evaluating power. For example, traditional power profiling only samples design activity over a few microseconds, which is too short a time to provide a realistic snapshot. Instead, ANSYS PowerArtist analyzes real-world use cases (like a high-definition video frame) to create

Premise: For Same Test, Same Workloads



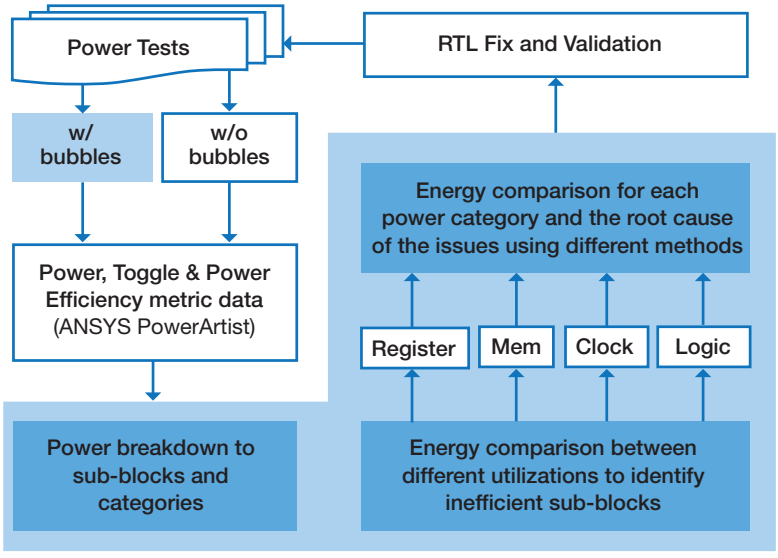
Higher energy for slower vectors exposes redundant activity.

power profiles within a few hours, which is orders of magnitude faster than standard approaches. Beyond power profiles, it allows engineers to budget power for the different parts of the design reliably at RTL through unique modeling of design implementation effects such as clock trees. It supports power efficiency analysis through quantifiable metrics, what-if power trend analysis, power-debugging for tracing problems to their roots, and power regressions, which are useful when a seemingly small fix suddenly causes a power surge elsewhere.

"ANSYS PowerArtist analyzes real-world use cases within a few hours, which is orders of magnitude faster than standard approaches."

DIFFERENTIAL ENERGY ANALYSIS

In their quest for a power-optimized design, the Qualcomm design team first minimized power leakage through process selection and power islands. Next, they focused on minimizing redundant switching activity to find dynamic power savings. They took an ingenious approach for this task: Instead of looking directly for redundant switching in the GPUs — a time-consuming, tedious process — they compared two versions of the same GPU by simulating them running at different speeds. Slower speeds were simulated by adding latencies to mimic starvation or stalls, for example. If the original design was optimally clock-gated, the number of nets switching should be the same for both runs, and the total energy for both runs should

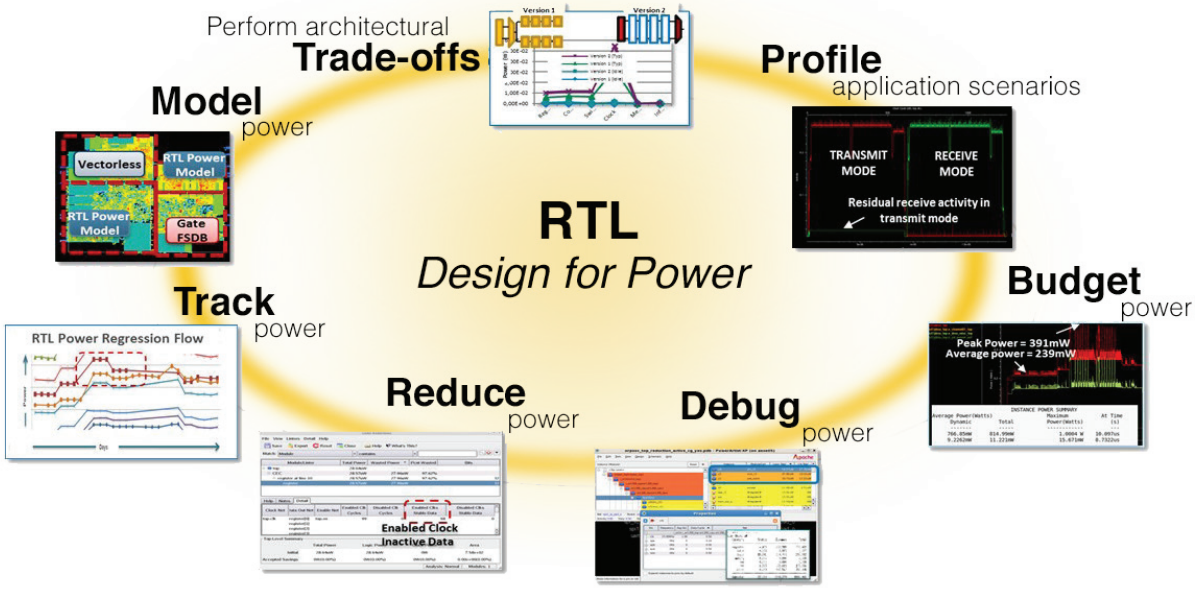


A unique methodology for differential energy and power analysis

be the same. However, if there were any gating inefficiencies in the original design, redundant switching in the design would be active over a longer period in the slower run, and therefore the integrated energy in that run would be higher than in the original run.

LOCATING REDUNDANT ACTIVITY

After discovering that the integrated energy in the slower run was higher, indicating the presence of gating inefficiencies, Qualcomm engineers took the analysis a step further in terms of dynamic power analysis. Noting that PowerArtist separates (at each level) switching and internal energy contributions, in addition to total energy, they were able to pinpoint the locations of redundant activity.



Seven steps to low-power RTL design using PowerArtist

"Differential energy analysis early in the development process, at the register transfer level (RTL), can optimize the efficiency of GPUs and keep mobile device temperature down."

Internal energy is the energy dissipated inside gates such as registers, whereas switching energy is the energy associated with the interconnect between gates. Redundant data input or output toggles on a register will, in the slower simulation run, cause an increase in both switching and internal energy, whereas redundant toggles on the clock input will increase only internal energy. There are four possible switching scenarios that help to pinpoint redundancies.

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. These include:

1. Extra toggles on the clock pin when data is stable
2. Extra toggles on D/Q pins when the clock is off
3. Extra toggles on both the D/Q pins and the clock pin

A MAJOR EFFICIENCY GAIN

Using this novel differential energy analysis methodology, Qualcomm engineers drilled down to find candidate blocks for more detailed analysis, including individual registers where fixes could have a big impact. Making these initial fixes helped reduce dynamic power consumption by 10%. This figure is significant for a company and an industry that is (and has been for years) incredibly focused on power reduction and squeezing inefficiencies out wherever possible. This increased efficiency came from register toggle optimization early in the design process, at the RTL stage. Similar analyses will be conducted to look for further improvements in the clock tree, memories and combinational logic. Qualcomm's successes with power efficiency through improvements in their GPU power and performance illustrates the value of early RTL power analysis using ANSYS PowerArtist.