POWER
RETOOLING FOR CHIPS

In a conventional approach, chip power-consumption analysis begins late in the design flow and occurs typically when the physical design is complete. At this stage, possible changes that can impact power are limited by schedule and cost considerations. AMD used ANSYS PowerArtist to apply a design-for-power approach at an early stage in the design flow, making it possible to reduce power consumption to unprecedented levels.
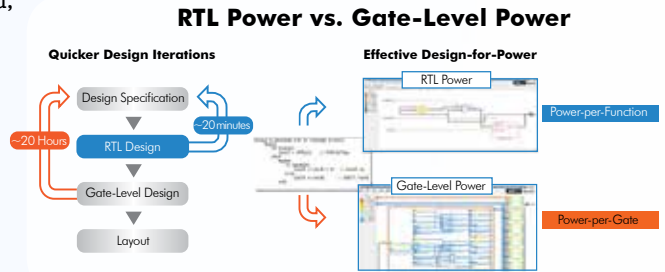
Power has been at the forefront of chip design for mobile applications and is now a key design concern for the Internet of Things (IoT), automotive, networking and other applications. Since the 1960s, the electronics industry has reliably followed Moore's Law. Gordon Moore predicted that computing power would double nearly every 18 months. Achieving this hasn't come easily, as engineers need to continually balance power, performance, reliability and cost. Modern server class processors, for example, contain billions of transistors that switch on and off at gigahertz frequencies, consume several hundred watts of power and generate significant heat. Device temperature is among the many factors that affect device performance — hotter chips run slower, become unreliable and can fail prematurely. The inability of the chip–package to dissipate heat is now becoming a performance bottleneck, limiting the ability to run chips at higher frequencies and also restricting the number of transistors per device. Therefore, reducing chip power consumption often makes it possible to increase performance while reducing the cost of powering and cooling the servers.

In the past, AMD engineers addressed power consumption using power analysis tools that operate at the gate and transistor levels. However, this approach is limited for several reasons. Any design changes at this late stage require re-synthesis of the design followed by an extensive verification process using a workflow with multiple tools. This substantially increases time to design closure. In addition, changes are limited because these tools optimize within the predetermined high-level architecture of the design. When the design is represented as a multitude of gates and transistors, it is also difficult to identify power hotspots at architectural or functional levels.

By **Mark Silla**,
Principle Member of Technical Staff,
and **Dave Larson**,
Member of Technical Staff,
AMD, Austin, USA

More recently, AMD engineers used ANSYS PowerArtist power analysis software on a processor to evaluate power consumption earlier in the design flow, and achieved an extraordinarily high level of power efficiency. By establishing a methodical approach of tracking register transfer level (RTL) power over various activity scenarios, they identified areas of significant wasted power consumption, and then addressed them through specific RTL changes.

MOVING POWER ANALYSIS UPSTREAM

Power consumption studies run on early-stage designs are limited in accuracy because key physical design elements are not completely defined. At later stages of design implemen-

tation, when power can be much more accurately estimated, changes are expensive and run the risk of delaying product introduction. AMD engineers overcame this challenge by adopting RTL as the abstraction to address power in a relative sense. RTL also provides a functional view of the design — for example, at the multiplexer and adder levels — that enables efficient power debug, in contrast to individual logic gates such as AND/OR. PowerArtist RTL power analysis runs multimillion-instance designs in minutes, which enables the designer to quickly evaluate multiple

RTL Power vs. Gate-Level Power

^ RTL Workflow

what-if scenarios. It also models physical effects such as clock-tree and wire capacitance that enable predictable accuracy for early design decisions. A typical server-class processor combines a central processing unit (CPU) with a data fabric that communicates with random-access memory (RAM). In the past, it took six to eight weeks to generate power consumption numbers based on the physical implementation for such designs, at which point the design had typically progressed to a stage where the analysis results were irrelevant. Not only did ANSYS PowerArtist trim analysis time to a single day (a reduction of 98 percent), but early findings enabled key decisions that reduced power beyond even the designer's expectations.
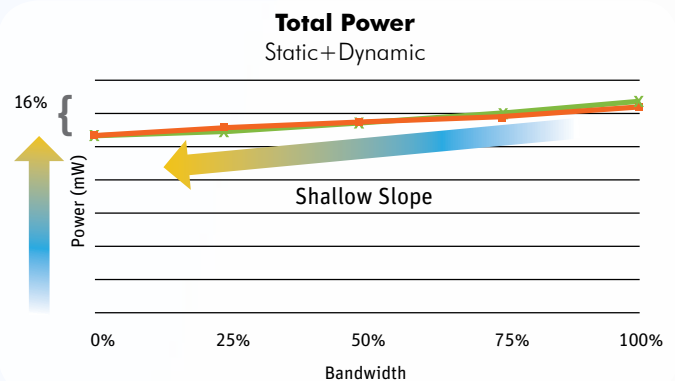
BANDWIDTH VERSUS POWER

AMD engineers established a robust methodology to address power that starts with generating the correct activity scenarios. Engineers ran RTL simulations to exercise the design

from idle to the highest bandwidth, and then examined power hotspots in various sections of the design. This unique approach allowed the team to carefully examine the relative difference in power between the low and high bandwidth scenarios.

They made two keys observations. First, PowerArtist computed the power consumption in the idle mode as only 16 percent lower than the 100 percent bandwidth mode of operation, and more than 50 percent of this power was consumed by the clock distribution network alone. Second, idle power was high, primarily because many inactive blocks were subject to clock toggling.

Total Power
Static+Dynamic

^ Initial ANSYS PowerArtist simulations showed that idle power was only 16 percent less than maximum power.

ONLY CIRCUITS DOING WORK SHOULD CONSUME POWER

AMD engineers established the goal that only circuits that were doing work should consume power. They used PowerArtist to identify design elements that were continuously supplied with a clock signal, even when they were not active, and therefore provided opportunities for improvement. The RTL owners used these clearly defined opportunities to significantly reduce power in their blocks.

For example, PowerArtist identified numerous cases in which a multiplexer was fed by several cones of logic — only one of which was active at a time — yet all of the cones remained continuously activated. The RTL design owners added clock gates to turn off power

⊕ **ANSYS PowerArtist**
ansys.com/powerartist

to the inputs and corresponding logic cones that were not active. Wakeup signals were used as another effective mechanism to alert disabled logic to capture data or to respond on the next cycle, generating a one-clock cycle latency cost-per-wakeup sequence but saving significant power.

CUTTING THE CLOCK POWER

AMD engineers explored different clock-gating architectures to determine the impact on power consumption. When the inactive logic is moved farther downstream from the clock, power is increasingly wasted in the clock distribution network. Engineers created rules to group specified classes of logic so that they could shut down sections of the clock tree as close as possible to the root to maximize power savings. In some cases, wakeup time has too great an impact on performance so the logic continues running despite the resulting power inefficiency. The team labeled these as intentional inefficiencies.
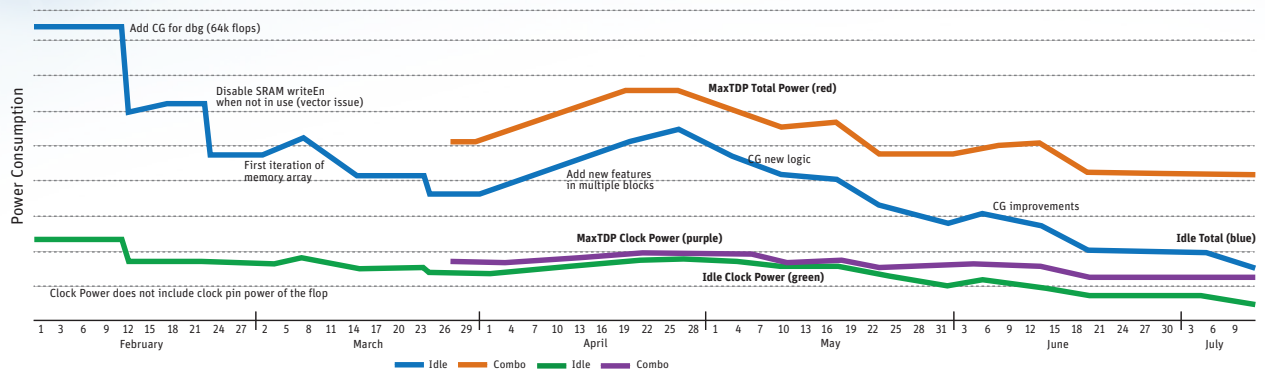
OPTIMIZING QUEUE DEPTH

AMD engineers ran experiments to determine the impact of queue depth on power consumption and performance. They concluded that if the queue was busy, increasing its size often reduced power consumption. However, if the queue was relatively inactive, then reducing its size could be beneficial. They also added logic to adjust the queue size on the fly based on utilization. The ability to adjust queue size in this way made low-bandwidth cases more efficient and demonstrated typical advantages of approximately 10 percent power reduction.
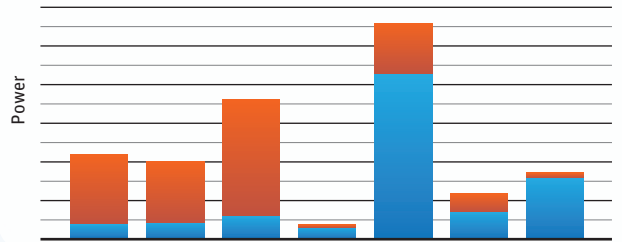
EARLY POWER ANALYSIS

Engineers reviewed the PowerArtist results, implemented the suggestions, and reran the simulations and power analysis to quickly verify the effectiveness of the suggestions. They ran weekly regressions to track power, allowing for rapid analysis and verification of modifications. Over the course of the project, the idle power was reduced by more than 70 percent. The improvements in idle power not only benefited the idle case but also created a 22 percent improvement in the maximum TDP case. The slope of the power versus bandwidth curve improved by 400 percent.

All in all, performing power analysis simulations earlier in the design flow made it possible to produce substantial reductions in power consumption, which in turn enabled performance improvements. AMD plans to integrate ANSYS PowerArtist RTL power analysis technology into its standard methods.
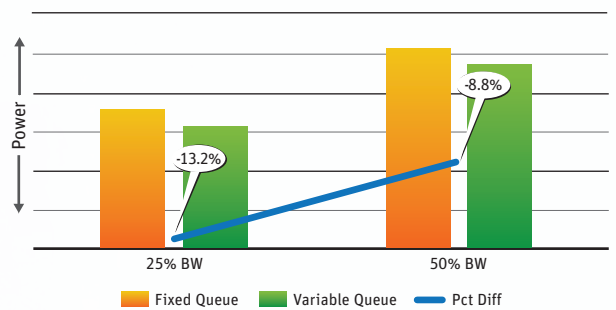
Aggressive Clock Gating

Aggressive clock gating reduced idle power by 25 percent. Red shows additional power wasted when not using aggressive clock gating and blue indicates idle power with aggressive clock gating enabled.

Power of Fixed Queue vs. Variable Queue

Variable queues provided power consumption improvements across the bandwidth spectrum.

Power consumption was steadily reduced over the course of the project.