Reaching the Limits of Low Power Design

J. S. Hobbs

Marketing & Business
Development
Synopsys, Inc.
Mountain View, CA
94043
Tel : 512-372-7514
e-mail:
josefina@synopsys.com

T.W. Williams

Implementation Group
Synopsys, Inc.
Broomfield, CO 80020
Tel : 650-584-4867
Fax : 617-600-9697
e-mail:
tww@synopsys.com

Abstract: As process technologies continue to shrink, and feature demands continue to increase, more and more capabilities are being pushed into smaller and smaller packages. But are we finally reaching the point where power density limitations make this trend no longer sustainable? What advanced techniques are in use today, and on the horizon, to address this? Are we limited only to hardware techniques, or can these power limitation issues be addressed with smarter software development? And how do we handle verification of these complex implementations? This paper explores possible methods for improving the "power capacity" of power sensitive designs.

I Introduction

In 1971, Intel introduced its first microprocessor, the 4004, which contained 2,250 transistors. More than 30 years later, the Itanium 2 processor, released in 2003, contained approximate 220 million transistors [1]. This is clear evidence that Moore's Law, the technology axiom that states that the number of transistors on a chip doubles roughly every two years, continues to hold true. And this trend applies not just to microprocessors, but to semiconductor design in general. Figure 1 shows the relative improvement to various aspects of laptop technology [2]. International Data Group (IDC), a market research and consulting firm, believes that the complexity inherent in today's mobile phone market will only increase over time, and wireless carriers, content developers, and handset manufacturers will face rising pressure to create compelling, customizable content faster than they ever have before [4].

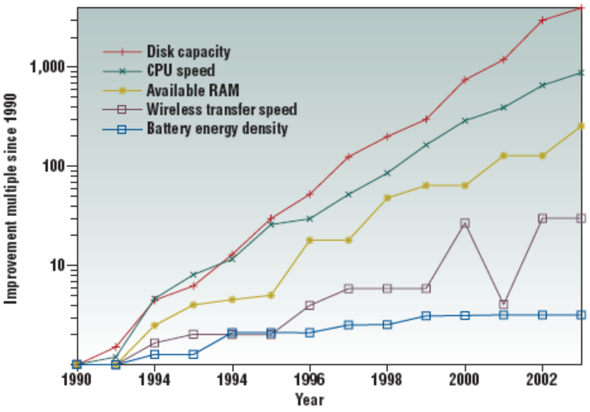

Figure 1. Relative improvement in laptop computing technology

However, as more and more transistors are packed closer and closer together, the ability to supply power to, and dissipate heat from, these circuits is becoming harder to accomplish in a reliable, efficient, and cost-effective fashion. And the problem continues to worsen, not only in terms of dynamic power as clock frequencies continue to rise, but also in terms of leakage power. Moving from one process technology to the next-generation node (e.g. from 90 nanometer (nm) to 65nm) can provide about twice the transistor density; however, leakage more than doubles. Even though the chip area is effectively halved, the increased power consumption can result in higher production costs. For example, it can force a change from less expensive plastic packages to more expensive ceramic packages with heat sinks. Excessive power consumption can also lead to non-competitive portable end-products due to poor battery life. In addition, too much power adversely affects product reliability because of electromigration.

With this power density issue becoming the rate-limiting factor in how much capability can be squeezed into an ever-shrinking amount of silicon real estate, power reduction

978-1-4244-1922-7/08/$25.00 ©2008 IEEE

is no longer a "nice to have" feature—it is essential. In fact, meeting the power budget has become one of the most important design goals for any system-on-chip (SoC) development team. Therefore, semiconductor companies are turning their focus towards evolutionary and revolutionary methods for increasing the "power capacity" of designs.

Designers of mobile and handheld devices have a particularly onerous task in addressing power, since it's not just a matter of reducing power once your timing and area goals are met, nor is it simply a matter of reducing power as much as possible. Consumers in this market space want it all: the SOC-low-power personal digital assistant (SOC-LP PDA) driver requires flat average and standby power, even as logic content and throughput continue to grow exponentially. Figure 2 shows the relationship between these contradictory goals [2].

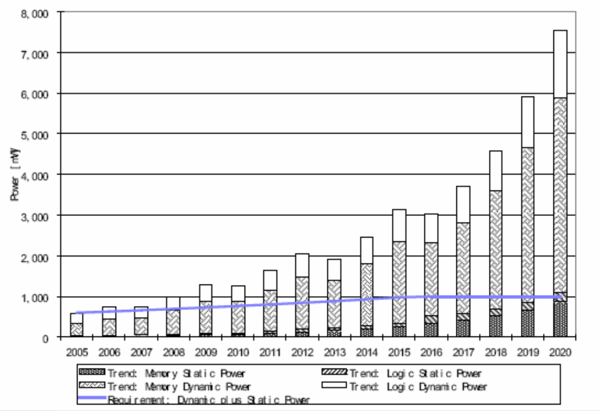

Figure 2. Predicted power consumption trend overlaid with power requirement

And at the smaller process technologies, they have to address power not just in terms of battery life, but also in terms of localized high-power consumption "hot spots" and greater sensitivity to process variation. Increasing power densities worsen thermal impact on reliability and performance, while decreasing supply voltages worsen switching currents and noise. The impact of process variability on leakage is very costly. Leakage scales inversely and exponentially with respect to the critical dimension of the transistor gate (i.e., channel length), and total leakage may vary by up to 5X to 20X across chips from the same wafer lot. Mitigation of leakage through Multi-Vt, power shutdown, or higher-level design techniques can incur area overhead and design process complexity, along with added variability (random dopant fluctuations and reduced supply voltage headroom make today's triple-Vt strategies less viable in future nodes) [3].

Therefore, the challenge is to find the optimal balance between the competing goals of timing, area, and power, such that their designs will provide the desired suite of functionality with the performance, form factor, and battery life metrics that will make them competitive in the market. To achieve this balance demands power optimizations that simultaneously exploit many degrees of freedom, including multi-Vt, multi-Tox, multi-Vdd coexisting in a single core, while guiding additional power optimizations at the architecture, operating system, and application software levels [2]. The remaining portion of this paper will step through the various levels of power optimization and management in some detail.

Process Technology

Some companies are addressing power density issues at the transistor level. Earlier this year, Intel announced new process technologies at 45nm that reportedly reduce transistor gate leakage by over tenfold, by adopting new materials for the dielectrics and metal alloys used to build their transistors [5].

Another technique for reducing leakage power is well biasing, where the transistors' gate leakage current is mitigated by adjusting the gate threshold voltage. However, with CMOS processes of 65nm and below, the benefits of well biasing on reducing leakage power are reportedly attenuating, as gate threshold leakage becomes a less dominant component of leakage power.

Gate Level

At the gate level, techniques like clock gating and multi-voltage-threshold (multi-Vt) optimization have been employed for a number of years. Clock gating reduces dynamic power by shutting off the clocks for circuits that are idle. Basic register level clock gating has been employed for decades. However, new techniques for merging, collapsing, splitting, and otherwise optimizing clock gating logic allow more sophisticated tradeoffs between clock tree power and skew, two of the most critical metrics for determining the optimal clock tree design.

Multi-Vt optimization reduces leakage power by replacing the faster but leakier low-threshold-voltage (LVT) cells with cells that run slower but have higher threshold voltages (HVT). Since HVT cells are slower, this replacement is only performed on logic paths where timing is already being met, so it is essential that the timing calculations made on these paths is accurate, such that the proper tradeoffs are made. This is another technique that has been generally available in the EDA market for quite some time, although it continues to be enhanced, to afford the user more granularity of control over what swaps are made by the tools. Note that while leakage power is generally considered to be independent of switching activity, for the best accuracy in performing this Vt swapping, it is recommended that switching activity be provided, since leakage power is state dependent, and at smaller technologies, cells with a large number of inputs can have a wide variation in leakage power characterization for those different states.

Architectural and System Level

At the chip level, as designers are writing the RTL, there is a wide range of techniques in use today, with varying degrees of adoption. Clock gating, while implemented by the tools at the gate level, is a technique that must also be considered during the architectural phase, since the synchronous load enable architecture used as the basis for clock gating must be designed in to the RTL. The chip architect must also decide the chip level clocking strategies, such as whether to buffer the clock tree before or after the clock gating logic, use a clock tree or a clock mesh, use synchronous or asynchronous clocks, etc.

Advanced low power techniques such as Multi-Vdd and power shutdown, if employed, will provide noticeably more power savings than the traditional, mainstream techniques such as clock gating and Multi-Vt. However, these techniques add another level of complexity to both the implementation and the verification aspects of the design. If power shutdown is employed and retention of the design state is required, yet another layer of complexity is added to the design flow. State retention incurs a relatively high cost of implementation, not only due to the additional placement and power routing requirements, but also because most retention registers used today are significantly larger than standard registers. Dynamic voltage scaling (DVS), dynamic voltage and frequency scaling (DVFS), and adaptive voltage scaling (AVS are further extensions to these advanced low power techniques, and are even more complex to implement and verify. It is essential that the chip architect have a clear understanding of the design requirements – and design metric priorities – prior to attempting an implementation of any of these advanced techniques. It is also recommended that a design team planning to use DVS, DVFS, or AVS should have prior experience taking a full design through tapeout using the more straightforward Multi-Vdd and power shutdown approaches.

Another technique rapidly gaining adoption is multi-core processor architecture, due in large part to the power density limitations mentioned earlier. It has become increasingly impractical to increase either the clock speed or instructions per clock of a single core. Therefore, instead of relying on the need to make any single processor faster, this architecture breaks the processing duties into parallel tasks, effectively increasing the performance of the design overall, if not an individual processor's performance.

Early insight into power consumption at an architectural and system level provides the greatest insight into potential power savings. Manual analysis has typically provided initial estimates for the most basic power budgeting purposes. Unfortunately, accurate system level power analysis is now becoming too difficult and time-consuming to perform manually. When increasingly sophisticated power management technologies such as dynamic frequency and voltage scaling, multiple power domains, and clock gating are employed, the architect needs an environment in which he/she can quickly and efficiently assess the overall efficacy of various configurations of power intent.

Early power analysis requires properly representative design abstractions that will be "good enough" in terms of accuracy, for both timing and power, such that relative comparisons are consistent and appropriate performance / power tradeoffs can be made accordingly. A system-level EDA solution should enable power modeling at the system level, providing a relative measure of power consumption on an application-by-application basis using the real applications' software and target OS. This enables the system architect to make tradeoffs to minimize power.

Low Power IP

Connectivity intellectual property (IP) for high-speed serial buses such as USB 2.0, PCI Express, and HDMI was not architected with power in mind. However, these standard interfaces are now commonly included in SoCs designed for mobile applications such as single-chip recordable DVD codecs and MP3 players. Mobile platforms and small form-factor devices must be complemented by power-conscious IP, to extend the battery life of these power sensitive products. Therefore, semiconductor designers require ultra low power derivatives of this traditionally high-performance logic IP. The challenge from the IP provider's viewpoint is to meet analog performance in a standard CMOS digital process technology that has been targeted for densely packed digital logic. Reduced supply voltages mean that architectures that once worked at 3.3V or 2.5V now need to work at 1.8V or lower without any loss in performance. One way to address this is to use a mixture of high-voltage I/O devices with the lower-voltage core devices. Performance can be maintained at 90-, 65- and even at 45nm by using a mixture of I/O and core devices. The key is to know where and how to use them, which is where the expertise of the IP provider is crucial. A deep understanding of the implications of low power design is required to make critical decisions, such as selecting which analog sections are to run at high supply voltages, and the best type of transistor (thin oxide, thick oxide or compound) to be used in order to circumvent one of the main roadblocks in 65nm CMOS technologies—the low nominal supply voltage. [6]

In today's hyper-power-sensitive market, it is not sufficient to supply just the lowest power IP. The IP provider must also provide their customers with all of the associated data necessary to integrate this IP into their designs. For soft IP, the advent of new power intent specification standards such as the Unified Power Formation (UPF) require delivery of an additional data file, which contains the definition of the power connectivity and power behavior of the IP.

SW Development

There are several ways to consider power in software (SW) design. Two key aspects of power awareness are: 1) writing SW to consume less power in hardware (HW), and 2) using SW to control and leverage the low-power-enabled capabilities in HW. One approach for writing software to consume less power is to employ thread-level parallelism, which is directly related to the advent of multi-core microprocessors. If this trend continues, new applications will have to be designed to utilize multiple threads in order to benefit from the increase in potential computing power provided by multiple lower-frequency, lower-power processor cores, instead of just relying on existing code to be automatically sped up when executed on a newer, faster – and higher powered -- processor core.

At the SW level, knowledge of upcoming instructions to be processed affords the ideal opportunity to adjust the operating voltage and clock frequency of a design, based on predicted device needs. Using SW-based power management control of the HW capabilities such as power gating and dynamic voltage and frequency scaling (DVFS), the SW designer can reduce the frequency or voltage, or shut off the power altogether, for HW subsystems that are not in use. This higher level of low power management can provide the most significant impact on a device's battery life, since it is reducing power over a period of time, thus reducing overall energy consumption.

Validating all of these techniques is a complex challenge. For an efficient and productive low-power design environment, tools must be consistent and correlated, and power information must be shared across the design flow. Specifying power intent in the early stages of RTL design, and carrying it through to downstream implementation and verification tasks, is fundamental to achieving a comprehensive and well correlated low-power design flow. A comprehensive approach requires a design flow that is power-aware at every stage of the design cycle, including verification, RTL synthesis, test, physical implementation and sign-off. When advanced low power implementation techniques such as Multi-Vdd and power shutdown are employed, corresponding advanced techniques for verification are required. For example, in functional simulation, not only is it necessary to validate the design in its various power states, but also it is essential to validate the transitions between those power states and the proper sequencing of those transitions. For static analysis in signoff, it is essential that timing, signal integrity, and power consumption all be considered concurrently, such that the interaction between these design aspects is accounted for in the analysis. Furthermore, given the impact of process variation on leakage power at the smaller technology points, it is essential that signoff analysis take this variability into account.

Conclusion

For almost a decade, designers and semiconductor vendors have been worriedly forecasting an end to the applicability of Moore's Law, and have laid blame for that demise on the issue of power. While recent advances in semiconductor technology and design techniques for power management have enabled Moore's Law to remain in effect, it has become increasingly clear that only a holistic approach to power reduction and management, from system level approaches all the way down to process technology modifications, will carry Moore's prediction into the next decade and beyond.

References

[1] "Moore's Law 40th Anniversary", Intel Press Kit, May 2005 (http://www.intel.com/pressroom/kits/events/moores_law_40th)

[2] "Power-Efficient System-on-Chip power Trends, System Drivers", International Technology Roadmap for Semiconductors (ITRS) 2005 (www.itrs.net/reports.html)

[3] A.B. Kahng, "Design Challenges at 65nm and Beyond", DATE07 proceedings, p. 1466.

[4] D. Linsalata, A. Slawsby, "Addressing Growing Handset Complexity with Software Solutions", IDC whitepaper, August 2005. (www.adobe.com/mobile/news_reviews/articles/2005/idc_whitepaper.pdf)

[5] M. LaPedus, "IC firms announce gate material breakthroughs", EETimesAsia, April 2, 2007

[6] N. Nanda, "Dealing with IP at 65nm and below", EETimes, November 1, 2007.