INTEGRATION, the VLSI journal 47 (2014) 1–11

Contents lists available at ScienceDirect

INTEGRATION, the VLSI journal

journal homepage: www.elsevier.com/locate/vlsi

Energy efficient computation: A silicon perspective

Kevin Brelsford [a], Serafín A. Pérez López [b], Santiago Fernandez-Gomez [c,*]

[a] Synopsys, Inc, USA
[b] EE, University of Vigo, Spain
[c] nanoASICs, Inc, USA

ARTICLE INFO

Article history:
Received 19 December 2012
Received in revised form
7 July 2013
Accepted 9 July 2013
Available online 16 July 2013

Keywords:
Energy efficiency
Low power
Computation
Microprocessors
Integrated circuits

ABSTRACT

Building and deploying an energy-efficient infrastructure for computation requires coordination across a large number of disciplines, from the end-user application software to the device that performs the actual computation through layers of virtualization software, operating systems, communication networks, data center architectures, arrays of servers, and others. While the most optimal savings can be achieved by careful coordination among all of the elements, there are opportunities for improvement on each individual component of the system. In particular, at the very core of computation, the processing elements are silicon devices embedded in integrated circuits. In this paper, an overview of the state of the art in building efficient silicon for computational applications is presented, including the techniques actively used by the industry, the upcoming new technologies, and the research initiatives geared toward the future.

Contents

1. At the core of computation

The vast majority of computation is currently performed by general-purpose microprocessors, which are defined as *multi-tasking programmable processor(s) contained on an integrated-circuit chip*. These elements are at the core of any type of computer infrastructure and, in particular, support cloud computing everywhere. Fig. 1 provides an overview of the entire system, from the end user to the smallest building element, the integrated circuit, which is generally manufactured using silicon technologies.

Any task performed by the end user, from browsing the web to generating a report querying a database, requires virtually all of the elements in Fig. 1, in addition to an extensive software base that typically includes multiple operating systems and applications. Invariably, one or more processors will be involved for the management, control, or execution of the task. When addressing the improvement of the amount of energy consumed by this system, the largest opportunities clearly arise from full-system coordinated strategies. For example, an interactive client terminal operated by a user could retain a certain configuration and state while at idle and allow for powering off the processor, server, and even the network access, moving the entire system into a very power efficient state, which would require an extremely tight coordination between all the elements, starting with the application software and operating system that must be able to convey the state of the terminal and manage powering up or down a large infrastructure. The greater power savings in this example are accompanied by a much larger complexity in the architecture, control, and management of the system.

At a much smaller scale, it is easier to find opportunities to improve the power efficiency of the different components in Fig. 1. Whether in a single component or within a small aggregation, there are general strategies such as locality or power states that

* Corresponding author. Tel.: +1 4086049310.
E-mail address: sfdez@mindsps.com (S. Fernandez-Gomez).

0167-9260/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.vlsi.2013.07.001

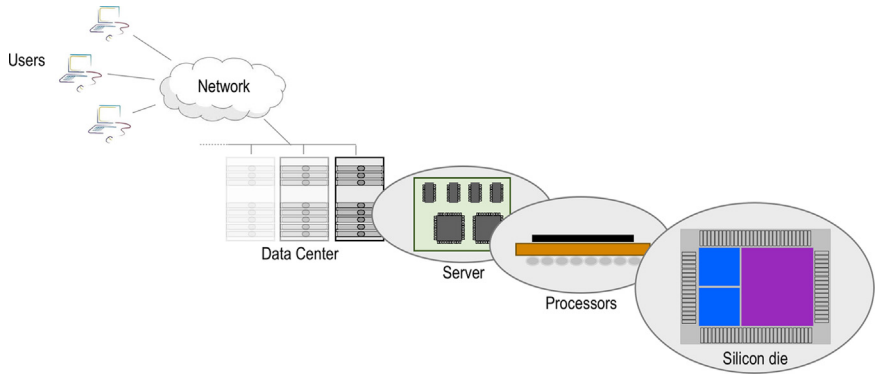

Fig. 1. Place of the silicon die in the hierarchy of computation.

can greatly reduce their power consumption without the need for complex coordination within the system. A number of companies have been successfully applying these concepts, providing hardware and software that allow, for instance, a decrease in the distance between a computational cluster and its main memory, drastically reducing the power required for common tasks [1]. All data centers manage multiple power states, where servers can be placed in reduced power modes, or even powered off, depending on the overall load of the system.

At an even more granular level, this article focuses on the processing elements built into integrated circuits (ICs), which are at the core of the computation. ICs are also significant consumers of power in the entire system, either directly with the energy they dissipate or indirectly through the support infrastructure necessary to supply power and remove their thermal dissipation. Depending on the target application, a large number of options exist to improve the efficiency of this type of IC, which can be applied throughout their design cycle and even later during their useful life. In recent years, with the advent of the Systems-on-Package, other power savings opportunities have become available; however, they are out of the scope of this article.

This paper is organized into five sections that follow the typical development phases of computational elements constructed using silicon technologies. First, an overview of how power is consumed in ICs is presented, including a discussion of their main components. Then, an overview of the main techniques for building efficient silicon is outlined for each of the main phases of silicon development: architecture and design, physical realization, and process technology. The article concludes with a look ahead to upcoming and future technologies under research or development that may help to further reduce power consumption in processors in the coming years.

2. Building energy efficient silicon

Power consumption in silicon-based integrated circuits can be divided into two main types: static power and dynamic power. Static power is defined as the power consumed by the different devices on-chip when they are not switching. They consume energy simply by being powered up, which essentially reflects one of the aspects of losses in current manufacturing technologies. The second source of power consumption is dynamic power. It is usually defined in CMOS technology as the power consumed by transistors when switching from one state to another. Eq. (1) summarizes both sources of consumption and the main factors that affect them [2,3].

$$\text{Power consumption} = P_S(V, process) + P_D(C, V^2, f) \quad (1)$$

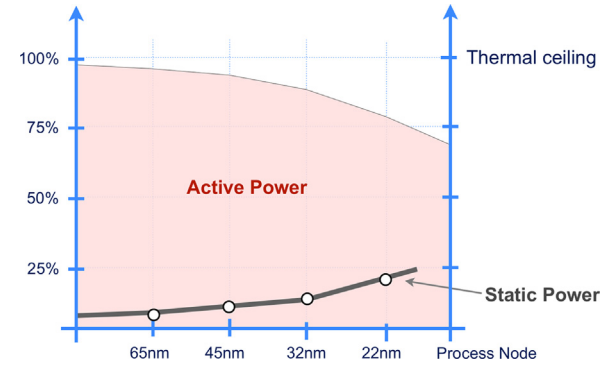

Fig. 2. Static vs. maximum available dynamic power in CMOS devices.

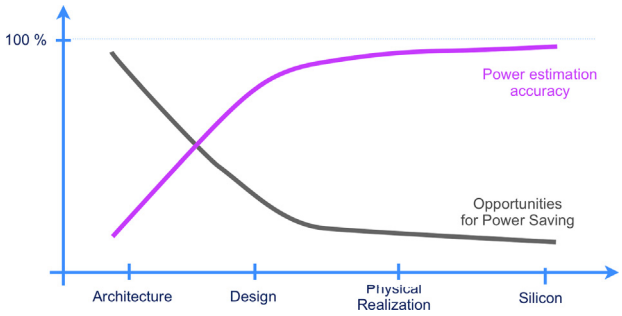

Fig. 3. Silicon design phases vs. opportunities for power savings.

In Eq. (1), the voltage, which has a quadratic effect on dynamic power, is one of the key levers to reduce power. Other factors, such as capacitance and frequency, have linear effects, but in the case of frequency, there is a direct impact on the performance that can be expected from the design. The capacitance factor is process and design dependent and one of the main targets of physically aware techniques to reduce power. In addition, the cycle time, which is not explicitly part of the equation, is a fundamental factor that needs to be considered. For applications where the workload does not require the IC to operate 100% of the time, other techniques are available for reducing the power, such as decreasing the IC voltage/frequency during idle times or even partially or completely powering off. Lastly, as the process technologies shrink, new effects are bringing static power to a first plane. In current and upcoming technologies, the typical amount of static power consumed by an IC can be greater than 25% (see the expected trend in Fig. 2, from the industry work of the authors and others [4,5]) of the total useful power that the component can handle, usually limited by its thermal envelope, which presents an additional challenge in which the amount of power available to perform computation (active power) is lower.