Project 2: Processor Power Consumption

Computer architects are constantly trying to increase performance while also trying to maintain or reduce power consumption figures. This is done using a variety of techniques in architecture, manufacturing, and software. The current trend is to add more cores to increase performance by utilizing parallelism, but this cannot last forever. To continue increasing performance and efficiency, new technologies will have to be developed, different from anything we have today.

Before we talk about where things are headed, let’s define what we mean when we talk about power. Power is the rate of energy consumed per unit of time. When we talk about power consumption, we are referring to the amount of energy being consumed at a given point in time or on average. Power consumption is important for many reasons. For devices that run on batteries, such as laptops and mobile devices, power consumption dictates how long the energy in those batteries will last. For datacenters and supercomputers, electricity and cooling are major costs, and should not be secondary concerns [1]. Beyond the monetary impact of energy bills and batteries, however, lies another important consideration. Power consumption is a limiting factor of computer performance as well. Up until a few years ago, the computer industry was concerned primarily with achieving higher clock rates to improve performance, and because frequency of operation and power are proportional to each other, other trade-offs had to be made to keep power consumption down. Now that components are smaller and have higher power density, significantly increasing clock speed is no longer feasible, and the industry has turned toward multicore computing.

The advantage of multicore computing is that highly parallel workloads can be done with a speedup close to equal to the number of cores devoted to the task using close to the same amount of power as doing the entire operation on a single core. If a parallelizable task can be done on one core in a certain amount of time at a fraction of the effort to the core, then
certainly it can be done on multiple cores with a certain level of performance at a fraction of the time. The question then becomes one of how many cores to use to get the most out of a system. This is highly dependent on the workload, since there are some tasks that simply have to be executed sequentially. If the architects have a general sense as to how much of a system’s workload is going to be parallelizable, Amdahl’s law can be used to determine the ideal number and configuration of cores to get the most performance while using the least amount of power [2].

As multicore technology improves, natural progression dictates that the number of cores on a chip will continue to grow, but this trend cannot continue forever. To fit more cores onto a chip, die sizes need to continue to shrink. The smaller transistors get, the more they collectively leak power. To mitigate this, not all of the transistors on a chip can be active at once, and the inactive portions of the chip are referred to as dark silicon. As transistors get smaller, and more of them are packed onto a chip, the ratio of dark silicon to total chip size grows exponentially. While the industry may be able to continue to shrink die sizes and fit more transistors on a chip, the increase in performance is not proportional to the amount of transistors being added. To avoid this problem, new architectures will have to be designed that are significantly different from the way computers are designed today [3].

The notion that the current path computer technology is on is unsustainable is a shocking one, to say the least. For decades, the industry has strived to fulfill Moore’s Law, the idea that the number of transistors that could be packed on a chip would double every 18 months or so. Now that we seem to be approaching a hard limit imposed by physics, it looks like Moore’s Law might not hold up much longer. For the time being, we are in the multicore era, and the battle is between latency- and throughput-oriented architectures and how to achieve the best balance, but soon the multicore era will end and computer architects will be forced to come up with the next great idea. Power consumption will always be a limiting factor in computer design, and coming up with a way to wring more computational performance out of a computer system without it approaching the temperature of the sun will be a great achievement.
References:

