1-1-2012

Generalizing Amdahl's Law for Power and Energy

Rong Ge
Marquette University, rong.ge@marquette.edu

Kirk W. Cameron
Virginia Polytechnic Institute and State University

Accepted version. *IEEE Computer*, Vol. 45, No. 3 (2012): 75-77. DOI.

Generalizing Amdahl's Law for Power and Energy

Kirk W. Cameron
Department of Computer Science,
Virginia Polytechnic Institute and State University,
Blacksburg, VA
Rong Ge
Department of Mathematics, Statistics and Computer Science,
Marquette University
Milwaukee, WI

Abstract: Extending Amdahl's law to identify optimal power-performance configurations requires considering the interactive effects of power, performance, and parallel overhead.

IEEE Computer, Vol 45, No. 3 (2012): pg. 75-77. DOI.

In the late 1960s, Gene Amdahl had the foresight to observe and predict one of computer science's most notable laws. At the time, Amdahl's law seemed somewhat discouraging to the community that was developing parallel systems and applications. He noted that increasing the number of processors working on a task resulted in two primary effects: the introduced parallelism sped up a portion of the task, but this had no effect on the remaining portion of the task.

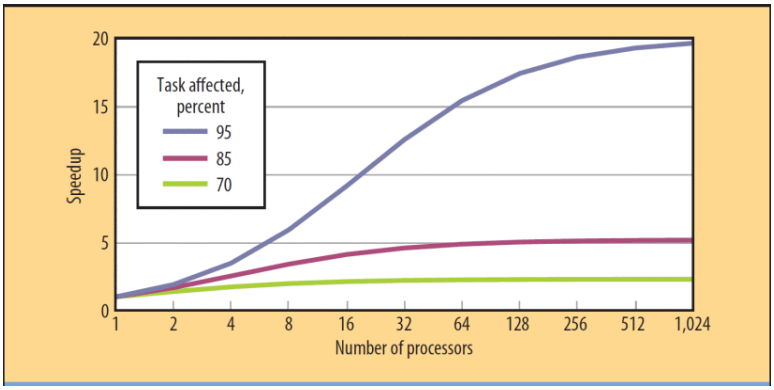

Figure 1. Amdahl's law for the effects of parallelism.

Amdahl highlighted the limitations of parallelism for making tasks run faster, which painted a bleak picture for the usefulness of parallel systems. The implication was that a task could be made faster to a point, but was limited by the portion of the task that was unaffected by parallelism.

Figure 1 shows the effective speedup for a task on a parallel system with different portions of the task affected by parallelism. Even with a large amount of the task affected (95 percent), the maximum speedup for the task is about 20 times faster, despite the use of more than 500 processors (systems or cores).

Despite these dire predictions, parallel systems proliferated, culminating in what we today deem the "multicore era." In the late 1980s, John Gustafson articulated the reason for the continued development of such systems. He noted that Amdahl's law applied to fixed workloads on scalable systems while parallel systems were being developed to solve problems that were previously intractable. In other words, the workload itself grew as the systems scaled. Thus, performance wasn't just defined as making a task run faster, but as completing more tasks in a fixed amount of time while a system scales.

IEEE Computer, Vol 45, No. 3 (2012): pg. 75-77. DOI.

Energy and Amdahl

A little over a decade ago, we started a journey that would ultimately bring together the performance analysis capabilities of Amdahl's law with the emerging notion of energy efficiency in parallel systems. The year 2000 brought major shifts in microarchitecture design that we postulated would have significant impact on the highperformance servers used as building blocks for large parallel systems. The emergence of controllable power modes meant system performance would vary with energy efficiency. Systems emerging with additional power states-for example, turning cores on and off-amplified the implications for parallel performance and for Amdahl's law.

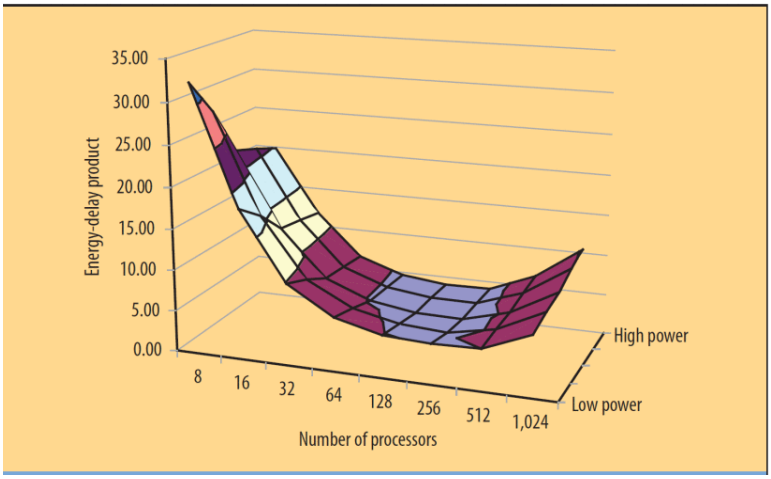

Figure 2. Power-aware speedup.

Over the years, we've been able to demonstrate how to maintain high performance on parallel systems while reducing total energy consumption. Initially, we tried (and failed) to use Amdahl's law to explain the effects of power modes on performance. But, after hundreds of experiments and with the aid of advanced analytical models of performance, we came to understand the interactive effects of performance and power.

IEEE Computer, Vol 45, No. 3 (2012): pg. 75-77. DOI.

Power-Aware Speedup

Amdahl's law is also called *fixed load speedup* because it describes a situation in which the workload is fixed and the amount of parallelism changes. Amdahl's law is simply $S_n = T(1)/T(n)$, where $T(1)$ is the sequential time and $T(n)$ is the parallel time on n processing elements for the same workload. Power-aware speedup, in the simplest case, describes the situation in which the workload is fixed and both the power modes and the amount of parallelism change. Power modes apply to different power-performance pairings where lower power is assumed to mean slower performance of a component. Strictly speaking, power-aware speedup is a generalized version of Amdahl's law because we can reduce it to Amdahl's law when power modes are fixed.

We can describe power-aware speedup as $S_n = T_1(w, f)/T_n(w, f')$, where $T_1(w, f)$ is the sequential time for workload w and fixed power mode f and $T_n(wf')$ is the parallel time for workload w and variable power mode f' on n processing elements.

Overhead?

Though a single article doesn't provide enough room to discuss all the details, suffice it to say that the trick in applying power-aware speedup for effective analysis of power scalable systems is capturing the parameters (R. Ge and K.W. Cameron, "Power-Aware Speedup," *Proc. Parallel and Distributed Processing Symp.*, IEEE CS,2007:doi/0.1109/IPDPS.2007.370246).

Of note, and critical to capturing the essence of why generalizing Amdahl's law for energy is important, is that all descriptions we've found in textbooks and the literature gloss over what turns out to be an essential link between Amdahl's law and power and energy. Namely, all of the aforementioned laws ignore the overhead of parallelism. Algorithm developers or programmers introduce parallel overhead when they redesign a

program for a parallel system. For example, data might need to be exchanged between parallel processes where no such communications were required in the sequential version.

Our early work in energy optimization showed that during communication phases (particularly during parallel overhead), there were great opportunities for saving energy without affecting performance. For example, during a busy waiting cycle, the processor might have spun for a prolonged period waiting on communication. These phases were excellent times to slow down the processor to save power.

Our initial attempts to use Amdahl's law to capture the effects of power changes during communication phases didn't capture the nuances of the overhead, so they effectively ignored active power management.

Without considering overhead, power efficiency in parallel systems is fairly straightforward: using more systems consumes more power. Calculating a bound for energy requires multiplying nodal power by duration for sequential and parallel systems. Conserving energy requires minimizing for energy across all combinations of systems.

Applying Amdahl's law can capture such cases. For a power scalable system, lowering the total power used (that is, using fixed minimal power states for power modes) minimizes energy use, which has an effect on power, performance, and then energy.

Power efficiency becomes infinitely more interesting in a power scalable system with dynamic power rnanagement—that is, a system that dynamically matches power use to performance demand. If the goal is to maintain performance while minimizing energy use in a power-scalable system, understanding the implications of dynamic power management requires capturing the overhead.

IEEE Computer, Vol 45, No. 3 (2012): pg. 75-77. DOI.

We thus modify the power-aware speedup formula to include parallel overhead:

$$S_n = \frac{T_1(w, f)}{T_n(w, f') + T_n(w_{PO}, f')},$$

where $T_n(w_{PO}f')$ is the parallel overhead and $T_n(wf')$ is parallel execution time without overhead.

Using a slight variation of this formula, it's possible to analyze the optimal configurations for energy and performance in parallel systems and applications. This technique can capture the effects of minimizing power using static configurations in the low-power state and their impact on energy and performance. More importantly, the model is accurate enough to capture active, dynamic power optimization curves where power is minimized during portions of the code, as in communication phases.

Figure 2 shows the dynamic power optimization plane for multiple power states and highly parallel configurations. Finding the minimization points on the resulting plane will identify the optimal configurations for a given system and application.

Amdahl's law still provides significant, useful insights to parallel applications and systems. The power-aware speedup model generalizes Amdahl's law by broadening the parameter space to consider the effects of active power-management strategies in power-scalable systems. In particular, introducing more accurate estimates of computational overhead allows us to truly capture the tradeoffs in performance and energy.

The resulting strategies for identifying optimal energy-efficient operating modes are applicable to tools including dynamic voltage and frequency scaling of processors, power throttling of memory, nodal power management, and core power management-basically. any system with dynamic power configurations that correspond to performance changes.

IEEE Computer, Vol 45, No. 3 (2012): pg. 75-77. DOI.