JouleSort: A Balanced Energy-Efficiency Benchmark

Suzanne Rivoire
Stanford University

Mehul A. Shah
HP Labs

Parthasarathy Ranganathan
HP Labs

Christos Kozyrakis
Stanford University

ABSTRACT

The energy efficiency of computer systems is an important concern in a variety of contexts. In data centers, reducing energy use improves operating cost, scalability, reliability, and other factors. For mobile devices, energy consumption directly affects functionality and usability. We propose and motivate *JouleSort*, an external sort benchmark, for evaluating the energy efficiency of a wide range of computer systems from clusters to handhelds. We list the criteria, challenges, and pitfalls from our experience in creating a fair energy-efficiency benchmark. Using a commercial sort, we demonstrate a JouleSort system that is over 3.5x as energy-efficient as last year's estimated winner. This system is quite different from those currently used in data centers. It consists of a commodity mobile CPU and 13 laptop drives, connected by server-style I/O interfaces.

Categories and Subject Descriptors

H.2.4 [**Information Systems**]: Database Management—*Systems*

General Terms

Design, Experimentation, Measurement, Performance

Keywords

Benchmark, Energy-Efficiency, Power, Servers, Sort

1. INTRODUCTION

In contexts ranging from large-scale data centers to mobile devices, energy use in computer systems is an important concern.

In data center environments, energy efficiency affects a number of factors. First, power and cooling costs are significant components of operational and up-front costs. Today, a typical data center with 1000 racks, consuming 10MW total power, costs $7M to power and $4-$8M to cool per year, with $2-$4M of up-front costs for cooling equipment [28]. These costs vary depending upon the installation, but they are growing rapidly and have the potential eventually to outstrip the cost of hardware [2]. Second, energy use has implications for density, reliability, and scalability. As data centers house more servers and consume more energy, removing heat from the data center becomes increasingly difficult [27]. Since the reliability of servers and disks decreases with increased temperature, the power consumption of servers and other components limits the achievable density, which in turn limits scalability. Third, energy use in data centers is starting to prompt environmental concerns of pollution and excessive load placed on local utilities [28]. Energy-related concerns are severe enough that companies like Google are starting to build data centers close to electric plants in cold-weather climates [24]. All these concerns have led to improvements in cooling infrastructure and in server power consumption [28].

For mobile devices, battery capacity and energy use directly affect usability. Battery capacity determines how long devices last, constrains form factors, and limits functionality. Since battery capacity is limited and improving slowly, device architects have concentrated on extracting greater energy efficiency from the underlying components, such as the processor, the display, and the wireless subsystems in isolation [20, 29, 31].

To drive energy-efficiency improvements, we need benchmarks to assess their effectiveness. Unfortunately, there has been no focus on a complete benchmark, including a workload, metric, and guidelines, to gauge the efficacy of energy optimizations from a whole-system perspective. Some efforts are under way to establish benchmarks for energy efficiency in data centers [33, 35] but are incomplete. Other work has emphasized metrics such as the energy-delay product or performance per Watt to capture energy efficiency for processors [13, 21, 27] and servers [34] without fixing a workload. Moreover, while past emphasis on processor energy efficiency has led to improvements in overall power consumption, there has been little focus on the I/O subsystem, which plays a significant role in total system power for many important workloads and systems.

In this paper, we propose *JouleSort* as a holistic benchmark to drive the design of energy-efficient systems. JouleSort uses the same workload as the other external sort benchmarks [1, 17, 25], but its metric incorporates total energy, which is a combination of power consumption and performance. The benchmark can be summarized as follows:

- Sort a fixed number of randomly permuted 100-byte records with 10-byte keys.

SIGMOD'07, June 12–14, 2007, Beijing, China.

- The sort must start with input in a file on non-volatile store and finish with output in a file on non-volatile store.
- There are three scale categories for JouleSort: 10^8 ($\sim$ 10GB), 10^9 ($\sim$ 100GB), and 10^{10} ($\sim$ 1TB) records
- The winner in each category is the system with the minimum total energy use.

We choose sort as the workload for the same basic reason that the Terabyte Sort, MinuteSort, PennySort, and Performance-price Sort benchmarks do [16, 17, 25]: it is simple to state and balances system component use. Sort stresses all core components of a system: memory, CPU, and I/O. Sort also exercises the OS and filesystem. Sort is a portable workload; it is applicable to a variety of systems from mobile devices to large server configurations. Another natural reason for choosing sort is that it represents sequential I/O tasks in data management workloads.

JouleSort is an I/O-centric benchmark that measures the energy efficiency of systems at peak use. Like previous sort benchmarks, one of its goals is to gauge the end-to-end effectiveness of improvements in system components. To do so, JouleSort allows us to compare the energy efficiencies of a variety of disparate system configurations. Because of the simplicity and portability of sort, previous sort benchmarks have been technology trend bellwethers, for example, foreshadowing the transition from supercomputers to clusters. Similarly, an important purpose of JouleSort is to chart past trends and gain insight into future trends in energy efficiency.

Beyond the benchmark definition, our main contributions are twofold. First, we motivate and describe pitfalls surrounding the creation of a fair energy-efficiency benchmark. We justify our fairest formulation, which includes three scale factors that correspond naturally to the dominant classes of systems found today: mobile, desktop, and server. Although we support both Daytona (commercially supported) and Indy ("no-holds-barred") categories for each scale, we concentrate on Daytona systems in this paper. Second, we present the winning 100GB JouleSort system that is over 3.5x more efficient ($\sim$ 11300 SortedRecs/Joule for 100GB) than last year's estimated winner ($\sim$ 3200 SortedRecs/Joule for 55GB). This system shows that a focus on energy efficiency leads to a unique configuration that is hard to find pre-assembled. Our winner balances a low-power, mobile processor with numerous laptop disks connected via server-class PCI-e I/O cards and uses a commercial sort, NSort [26].

The rest of the paper is organized as follows. In Section 2, we estimate the energy efficiency of past sort benchmark winners, which suggests that existing sort benchmarks cannot serve as surrogates for an energy-efficiency benchmark. Section 3 details the criteria and challenges in designing JouleSort and lists issues and guidelines for proper energy measurement. In Section 4, we *measure* the energy consumption of unbalanced and balanced systems to motivate our choices in designing our winning system. The balanced system shows that the I/O subsystem is a significant part of total power.

Section 5 provides an in-depth study of our 100GB JouleSort system using NSort [26]. In particular, we show that the most energy-efficient, cost-effective, and best-performing configuration for this system is when the sort is CPU-bound.

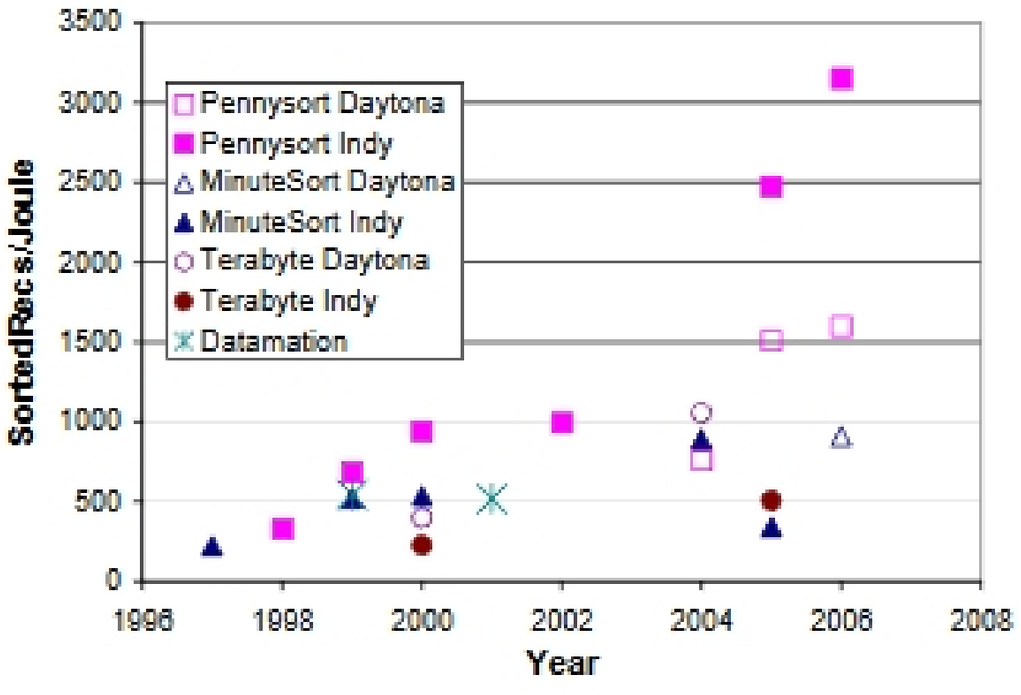

Figure 1: Estimated energy-efficiency of previous winners of sort benchmarks.

We also find that both the choice of filesystem and in-memory sorting algorithm affect energy efficiency. Section 6 discusses the related work, and Section 7 presents limitations and future directions.

2. HISTORICAL TRENDS

In this section, we seek to understand if any of the existing sort benchmarks can serve as a surrogate for an energy-efficiency benchmark. To do so, we first estimate the SortedRecs/Joule ratio, a measure of energy efficiency, of the past decade's sort benchmark winners. This analysis reveals that the energy efficiency of systems designed for pure performance (i.e. MinuteSort, Terabyte Sort, and Datamation winners) has improved slowly. Moreover, systems designed for price-performance (i.e. PennySort winners) are comparatively more energy-efficient, and their energy efficiency is growing rapidly. However, since our 100GB JouleSort system's energy efficiency is well beyond what growth rates would predict for this year's PennySort winner, we conclude that existing sort benchmarks do not inherently provide an incentive to optimize for energy efficiency, supporting the need for JouleSort.

2.1 Methodology

Figure 1 shows the estimated SortedRecs/Joule metric for the past sort benchmark winners since 1997. We compute these metrics from the published performance records and our own estimates of power consumption since energy use was not reported. We obtain the performance records and hardware configuration information from the Sort Benchmark website and the winners' posted reports [16].

We estimate total energy during system use with a straightforward approach from the power-management community. Since CPU, memory, and disk are usually the main power-consuming system components, we use individual estimates of these to compute total power. For memory and disks, we use the HP Enterprise Configurator [19] power calculator to yield a fixed power of 13W per disk and 4W per DIMM. Some of the sort benchmark reports only mention total memory capacity and not the number of DIMMs; in those cases, we assume a DIMM size appropriate to the era of the report. The maximum power specs for CPUs, usually

quoted as thermal design power (TDP), are much higher than the peak numbers seen in common use; thus, we derate these power ratings by a 0.7 factor. Although a bit conservative, this approach allows reasonable approximations for a variety of systems. When uncertain, we assume the newest possible generation of the reported processor as of the sort benchmark record because a given CPU's power consumption improves with shrinking feature sizes. Finally, to account for power supplies inefficiencies, which can vary widely [3, 5], and other components, we scale total system power derived from component-level estimates by 1.2 for single-node systems. We use a higher factor, 1.6, for clusters to account for additional components, such as networking, management hardware, and redundant power supplies.

Our power estimates are intended to illuminate coarse historical trends and are accurate enough to support the high-level conclusions in this section. We experimentally validated this approach against some server and desktop-class systems, and its accuracy was between 2% and 25%.

2.2 Analysis

Although previous sort benchmark winners were not configured with power consumption in mind, they roughly reflect the power characteristics of desktop and higher-end systems in their day. Thus, from the data in Figure 1, we can infer qualitative information about the relative improvements in performance, price-performance, and energy efficiency in the last decade. Figure 1 compares the energy efficiency of previous sort winners using the SortedRecs/Joule ratio and supports the following observations.

Systems optimized for price-performance, i.e. PennySort winners, clearly are more energy-efficient than the other sort benchmark winners, which were optimized for pure performance. There are two reasons for this effect. First, the price-performance metric motivates system designers to use fewer components, and thus less power. Second, it provides incentive to use cheaper, commodity components which, for a given performance point, traditionally have used less energy than expensive, high-performance components.

The energy efficiency of cost-conscious systems has improved faster than that of performance-optimized systems, which have hardly improved. Others have also observed a flat energy-efficiency trend for cluster hardware [2]. Much of the growth in the PennySort curve is from the last two Indy winners, which have made large leaps in energy efficiency. In 2005, algorithmic improvements and a minimal hardware configuration played a role in this improvement, but most importantly, CPU design trends had finally swung toward energy efficiency. The processor used in the 2005 PennySort winner has 6x the clock frequency of its immediate predecessor, while only consuming 2x the power. Overall, the 2005 sort had 3x better performance than the previous data point, while using 2x the power. The 2006 PennySort winner, GPUTeraSort, increased energy efficiency by introducing a new system component, the graphics processing unit (GPU), and utilizing it very effectively. The chosen GPU is inexpensive and comparable in power consumption (57W) to the CPU (80W), but it provides better streaming memory bandwidth than the CPU.

This latest winner, in particular, shows the danger of relying on energy benchmarks that focus only on specific hardware like CPU or disks, rather than end-to-end efficiency. Such specific benchmarks would only drive and track improvements of existing technologies and may fail to anticipate the use of potentially disruptive technologies.

Benchmark	SRecs/sec	SRecs/$	SRecs/J
PennySort	50%/yr.	57%/yr.	24%/yr.
Minute, Terabyte, and Datamation	37%/yr.	n/a	12%/yr.

Table 1: This table shows the estimated yearly growth in pure performance, price-performance, and energy efficiency of past winners.

Since price-performance winners are more energy-efficient, we next examine whether the most cost-effective sort implies the best achievable energy-efficient sort. To do so, we first estimate the growth rate of sort winners along multiple dimensions. Table 1 shows the growth rate of past sort benchmark winners along three dimensions: performance (SortedRecs/sec), price-performance (SortedRecs/$), and energy efficiency (SortedRecs/Joule). We separate the growth rates into two categories based on the benchmark's optimization goal: price- or pure performance, since the goal drives the system design. For each category, we calculate the growth rate as follows. We choose the best system (according to the metric) in each year and fit the result with an exponential. Table 1 shows that PennySort systems are improving almost at the pace of Moore's Law along the performance and price-performance dimensions. The pure performance systems, however, are improving much more slowly, as noted elsewhere [16].

More importantly, our analysis shows much slower estimated growth in energy efficiency than in the other two metrics for both benchmark categories. Given last year's estimated PennySort winner provides $\sim$ 3200 SRecs/J, our current JouleSort winner at $\sim$ 11300 SRecs/J is nearly 3x the expected value of $\sim$ 4000 SRecs/J for this year. This result suggests that we need a benchmark focused on energy efficiency to promote development of the most energy-efficient sorting systems and allow for disruptive technologies in energy efficiency irrespective of cost.

3. BENCHMARK DESIGN

In this section, we detail the criteria and challenges in designing an energy-efficiency benchmark. We describe some of the pitfalls of our initial specifications and how the benchmark has evolved. We also specify rules of the benchmark with respect to both workload and energy measurement.

3.1 Criteria

Although past studies have proposed energy-efficiency metrics [13, 21, 34, 27] or power measurement techniques [9], none provide a complete benchmark: a workload, a metric of comparison, and rules for running the workload and measuring energy consumption. Moreover, these studies traditionally have focused on comparing existing systems rather than providing insight into future technology trends. We set out to design an energy-oriented benchmark that addresses these drawbacks with the criteria below in mind. While achieving all these criteria simultaneously is hard, we strive to encompass them as much as possible.

Energy-efficiency: The benchmark should measure a system's "bang for the buck," where bang is work done and the cost reflects some measure of power use, e.g. average