Mobile Platform Benchmarks

A Methodology for Evaluating Mobile Computing Devices

Daniel McKenna
Transmeta Corporation

February 3, 2000

Property of:
Transmeta Corporation
3940 Freedom Circle
Santa Clara, CA 95054 USA
(408) 919-3000
http://www.transmeta.com

Overview

Today's truly mobile computing devices operate in an untethered environment characterized by wireless Internet connectivity and battery-powered operation. This mobile computing paradigm is significantly different from the conventional desktop computing model, and introduces a new set of evaluation criteria for platform characterization and comparison. Truly mobile platform user's requirements and expectations include mobile application compatibility, performance sufficiency, and long battery life, in addition to the traditional peak performance requirement of desktop computing.

Other distinguishing features of truly mobile computing are the application software popular in these environments, the variety of ways mobile users choose to configure and operate the devices, and the trade-offs that users face in attempting to get useful work done while striving to maximize the operating time on batteries. Truly mobile computing also covers a broad range of platform types, including hand-held Windows® PCs, thin-clients, Internet appliances, and PDAs. This large variety of platform types, operating system and application environments, and usage profiles complicates the task of platform benchmarking and comparison.

Computer industry benchmarks for system characterization evolved out of the desktop-workstation-server-mainframe environment that demands continuous peak performance. These traditional benchmarks are inadequate to address the requirements of the new mobile computing paradigm. Truly mobile platforms should be gauged with the attributes important to mobile users. The benchmarks appropriate for truly mobile computer products are different than benchmarks appropriate for desktop and desktop-replacement portable systems.

Mobile computer benchmarks should measure performance in combination with the energy consumption and battery life penalty for that performance, and should address the issue of performance sufficiency as well as peak performance. Mobile platform benchmarks should measure performance and energy consumption using real mobile application workloads in mobile systems configured and operating the way users configure and operate them, under battery power.

This paper describes a methodology for developing benchmarks for mobile platforms that address the shortcomings of traditional computer benchmarks and incorporate the criteria above.

Introduction

Characterizing mobile platform performance requires a different viewpoint than the traditional desktop-workstation-server-mainframe model that demands the delivery of continuous peak performance as the single goal of the system. Performance sufficiency and energy efficiency have become the predominant platform requirements for today's battery-powered mobile computing devices. Improvements in mobile

system usability must take into consideration the battery life penalty that is the natural consequence of the peak performance model of processing. Mobile systems must deliver performance sufficient to satisfy the application workload and no more. Each increment of performance beyond the instantaneous requirement of the application workload results in an unacceptable degradation in battery life. If application performance sufficiency is achieved, as is common with today's high-speed system designs, energy efficiency becomes the overwhelming mobile platform value attribute.

Mobile Platforms

The rapidly evolving market for highly mobile Internet computing devices is driven by the need to work and communicate in a free-roaming, battery-powered operating mode. The communications and computing technology infrastructure, combined with the rapid growth of an Internet e-commerce business model, is enabling a revolution in mobile computing. Mobile Internet computing platforms are appearing at a rapid pace in a proliferation of new categories, including thin-and-light PCs, mini- and hand-held PCs, mobile thin clients, Internet appliances, and PDAs. Each of these categories addresses a different constellation of user needs, with a variety of features and operating characteristics. Evaluating and comparing these new mobile devices has become a challenge because of the multitude of platform-specific operating systems, applications, features, and value attributes. In addition to these complexities, some of the new technologies developed for truly mobile platforms cannot be evaluated accurately by existing industry benchmark methodologies.

Computer Benchmarks

Computer industry benchmarks for system characterization and comparison evolved from the same desktop-workstation-server-mainframe environment that demands continuous peak performance. The central benchmarking metric, for almost the entire history of computer benchmarking, has been a peak performance metric (often referred to as throughput), that is simply the computational workload completed per unit of time. These benchmarks, the direct descendents of yesterday's mainframe-era batch-processing benchmarks, are completely inadequate to address the requirements of the new mobile computing paradigm.

One of the major limitations of classic computer benchmarks has been the synthetic nature of the benchmark software and usage profiles. Synthetic benchmarks are artificial programs, usually carefully written using run-time instruction frequency statistics, with the expectation they are representative of the run-time operating characteristics of real application software. These synthetic benchmarks have been helpful in some limited circumstances, e.g. examining processor micro-architectural features and trade-offs, and examining processor/cache/memory interactions and trade-offs. For characterizing and comparing PC platforms, however, they have been of very limited usefulness, primarily because synthetic benchmark results do not correlate well with the vast majority of PC software usage, and are very poor at driving representative system-level and IO activities.

TRANSMETA
Corporation

More recently, attempts to create application benchmark suites for evaluating computer systems have become common. These application benchmark suites are now widely used for evaluating and comparing PC systems. Application benchmarks were developed in an attempt to overcome the limitations of synthetic benchmarks. By using real application code, the difficulty in creating representative synthetic applications is overcome. Some of the significant limitations of these new application-level benchmarks are the use of synthetic application usage profiles, score weighting, and result aggregation.

A fundamental requirement for an effective computer benchmark is the benchmark methodology must not introduce uncharacteristic system behavior. Whatever it is the benchmark is doing in the system, it must be doing something very similar, if not identical, to what the system normally does when running user applications. If this is not the case, then how is this benchmark representative of what users really experience? This gets to the heart of the issue with application benchmark usage profiles, score weighting, and aggregation.

It is inherently extremely difficult, if not impossible, to derive a representative set of applications and usage profiles, with representative user interactions, and construct a script that puts this all together, times and weights the component results, and outputs a single number that tells us anything very meaningful. Compounding this difficulty are the new requirements imposed by mobile computing platforms with a diversity of system characteristics, applications, usage patterns, and concern for battery life.

There is an additional hazard with any benchmark technique, that is observable in the correlation of the benchmark results with actual user experience. Benchmarks are fundamentally tools for evaluating and comparing alternatives. The resulting scores, rankings, and output data from the benchmarks must somehow be representative of some real-world computer system behavior, and must show a strong correlation with the user-perceived operating attributes of the software and systems being benchmarked. Without this correlation, the benchmarks become essentially meaningless, and the benchmark methodology used can be judged by the strength of this correlation with user experience.

The requirements for an adequate mobile platform benchmark methodology are that all the important platform operating attributes are evaluated appropriately, including application compatibility, performance, and energy efficiency; that the methodology enables comparison of mobile platform alternatives, at the component, subsystem, system, and software levels; that the results correlate with real-user real-world experience in all the significant platform attributes; that the methodology comprehends platform specific features and attributes; and that the benchmark metrics are easily measured, calculated, and the results reproducible.

Evaluating mobile systems using yesterday's mainframe benchmark technology has left the industry and the user community with very poor visibility into the operating characteristics and attributes so important to the users of mobile computer products. As the old saying goes, "When your only tool is a hammer, all your problems look like nails," so it is when your only yardstick is peak performance (mainframe benchmarks), everything you measure looks like a mainframe. Mobile platforms should be measured using the attributes important to mobile users. In the absence of adequate industry-standard mobile

benchmarks, and in conjunction with the development of a new type of microprocessor addressing the needs of the mobile computer user, Transmeta has also developed a new benchmark methodology for evaluating mobile computer products.

Workload Characteristics

Application workloads for truly mobile computing devices, under battery powered operating conditions, are distinctly different from the workloads encountered in the desktop environment. The users of mobile platforms have different requirements and expectations for these mobile computing devices compared to line-powered desktop systems, and the type of work that mobile users perform on their systems is diverging rapidly away from typical desktop usage scenarios. Failure to recognize these differences has resulted in the mistaken application of desktop benchmark measurement tools to mobile platform evaluation.

An analysis of the variety of workloads common in mobile applications is helpful in understanding the operating characteristics of the different application types, and for highlighting the significant mobile platform attributes that must be addressed by an adequate mobile platform benchmark methodology. Application workloads fall into distinctly identifiable categories, as described below.

Fixed Demand Workloads

Fixed demand workloads, where the required system performance is relatively constant over a predetermined interval of time, are becoming a very common workload category on mobile platforms. Multimedia is the best example of this type of application. For instance, a DVD movie requires sufficient performance to display the movie without dropping frames, i.e. the system must deliver 30 frames/second. The processor performance required may vary slightly depending on the changing content of the frames, but any extra performance beyond what is required only degrades battery life.

Variable Demand Workloads

The second workload category is variable demand workloads, where the required system performance is determined by a variety of factors, including user response time, internal and external device response times, and processor/memory response time. Traditional productivity applications are good examples in this category. An application such as Excel may be paced by human interaction rate and response time when a spreadsheet is being created and data entered. The processor performance required during this time is usually quite low. On the other hand, during a large spreadsheet 'recalc' or the creation of a graph from spreadsheet data, the processor performance required is much higher in order to complete the task in a reasonable time from the users perspective. This variability in application task demand is very common for mobile platform applications.

Mobile Platform Benchmarks

Benchmarks, in general, are supposed to allow comparison of alternatives and yield results that correlate with the real-world experience of users. Mobile platform benchmarks should allow comparison of various component and system-level mobile platform solutions, and the results should correlate with real-world and real-user mobile platform operating experience. The benchmark methodology described in this document attempts to address these requirements, and is referred to as the mobile platform benchmark methodology.

The core of the mobile platform benchmark methodology is to run real-world mobile x86-compatible application software (workloads) in usage profiles that correspond with the way users typically interact with the software and the mobile platform. A set of benchmark metrics was selected that incorporate energy usage and efficiency, peak performance (throughput), and performance sufficiency in a way that is plausible, measurable, reproducible, and correlates with real-world experience.

Workload Completion Rate (Peak Performance) Metric

The first mobile platform benchmark metric is called Workload Completion Rate (WCR). WCR is a classic application-level peak performance metric. Another common name for this type of metric is throughput. Workload completion rate is defined as:

$$WCR = (Mobile\ Workload\ Completed)\ /\ (Time\ to\ Complete\ Workload)$$

The measurement units for WCR is workload units per hour. This metric is significant for variable demand workloads, and processor/memory-paced workloads in particular.

Workload Completion Efficiency (Energy Efficiency) Metric

The second mobile platform benchmark metric is called Workload Completion Efficiency (WCE). WCE describes an energy consumed-work completed relationship, and may be thought of as the energy efficiency for a given workload. Workload completion efficiency is defined as:

$$WCE = (Mobile\ Workload\ Completed)\ /\ (Energy\ Consumed\ to\ Complete\ Workload)$$

The measurement units for WCE is workload units per Watt-hour. This metric is significant for both fixed demand workloads and variable demand workloads. The metric is very important for mobile platforms because it measures energy efficiency and directly relates to battery life.

WCE provide a much better measure of mobile system operation then WCR (peak performance) alone, because mobile users experience the energy penalty of accelerated system activity as shortened operating life under battery operation. WCR is important, however, to establish thresholds for application

performance sufficiency. Some mobile applications will be unusable if WCR (peak performance) is not sufficiently large.

About the Metrics

The mobile platform benchmark metrics were selected to address the relationships between three independent variables: workload (demand characteristics), energy consumption, and time. Workload demand characteristics vary widely across the range of mobile applications. Inclusion of this variable in the methodology addresses a major limitation of synthetic benchmarks and current industry application (aggregated result) benchmarks.

The three variables were combined in the metrics to have plausible intuitive meaning, as described above. The metrics have the property that larger values represent relative improvement in the attributes addressed by the metric. For example, a larger WCR (peak performance) value indicates a "faster" device or system. Likewise, a larger WCE value indicates a more energy efficient solution.

Selection of Mobile Platform Workloads

The mobile platform benchmark methodology is focused on mobile platform components, systems, and applications, and benchmarks constructed in accordance with this methodology must reflect real-world mobile platform usage. The fundamental characteristic of mobile platform usage is untethered operation, which implies battery operation and either wireless or no connectivity. The applications selected for mobile platform benchmarks must be typical for the mobile platforms tested, and should be run in battery-powered operating mode. For Windows-based mobile PC platforms, these applications can include Microsoft® Word®, Excel®, PowerPoint®, Outlook®, and Internet Explorer®. Other applications useful for Windows-based mobile platforms can include games, MP3, soft DVD and other media players, as well as operating system functions.

Mobile thin clients and Internet appliances based on Linux require a different set of applications. Note that this methodology allows selection of any operating system and applications for testing as long as the software selected is appropriate for the particular platform being characterized and the benchmark tests are run on the system in a manner consistent with typical use.

Measurements

Energy efficiency benchmark results are captured by running mobile application program workloads on instrumented test platforms and measuring the energy consumption of components (e.g. CPU), subsystems, and/or the total system. Performance related benchmarks also capture time for completion of

TRANSMETA
Corporation

application workloads. Application workloads that are user rate-limited are scripted at typical user interaction rates. Systems are setup to operate in untethered mobile configurations (i.e. power management enabled and operational) whenever possible and appropriate.

Reporting

To cover the space of mobile applications and usage profiles, reporting of metrics is not aggregated into composite benchmark results. Rather, a large family of results is reported, which in total provides a much more detailed picture of what the particular operating characteristics are for the mobile platform components and systems evaluated. See "Crusoe^{TM} Processor Benchmark Report" [2] for examples of actual result reporting formats.

Ongoing Development

It is expected that a significant and ongoing effort will be required for selection and tuning of application workloads, scripts, and usage profiles, as well as instrumentation development, test platform modifications, and the actual data measurements. Fully instrumented mobile platform benchmark measurement systems are being used to collect the wide array of data necessary for characterizing current and future mobile platform processor components and systems. The current methodology will be expanded to incorporate device and system-level thermal characterization, user-observable system response-time measurements (e.g. time to wake-up from low-power inactive states), and other platform usability characteristics. See [2] for current benchmark results.

Conclusion

The methodology and metrics for the mobile platform benchmarks described here are a significant improvement over existing industry standard PC benchmark efforts, particularly with respect to objective mobile platform characterization. The metrics chosen are extremely plausible, easily measured, and show strong correlation to real-world mobile platform operating characteristics. The metrics can be applied at the component, subsystem, and system level, and are thus extremely useful tools for system comparison, design trade-off analysis, competitive analysis, and system hardware and software performance and energy efficiency tuning.

References

[1] Alexander Klaiber, "The Technology Behind Crusoe™ Processors", *Transmeta Corporation White Paper, January 2000.*

[2] "Crusoe™ Processor Benchmark Report", *Transmeta Corporation.*

[3] Thomas D. Burd and Robert W. Brodersen, "Processor Design for Portable Systems", *Dept. EECS, University of California Berkeley.*

[4] Thomas Burd and Robert Brodersen, "Energy Efficient CMOS Microprocessor Design", *Proceedings of the 28th Hawaii International Conference on System Sciences, January 1995.*

[5] Trevor Pering, Tom Burd, and Robert Brodersen, "The Simulation and Evaluation of Dynamic Voltage Scaling Algorithms", *Proceedings of the 1998 International Symposium on Low Power Electronics Design.*

[6] Trevor Pering, Tom Burd, and Robert Brodersen, "Dynamic Voltage Scaling and the Design of a Low-Power Microprocessor System", *Electronic Research Laboratory, University of California Berkeley.*

To learn more about Transmeta Corporation, Crusoe processors, and mobile platform benchmarks, visit http://www.transmeta.com.

Microsoft, Windows, Word, Excel, PowerPoint, Outlook, and Internet Explorer are registered trademarks of Microsoft Corporation.

Acknowledgments