Cloud Infrastructure Nirvana: Brawny vs. Wimpy Cores

Jason Waxman
General Manager,
Cloud Infrastructure Group, Intel
June 20, 2012

Brawny vs. Wimpy Cores

Google™

Brawny cores still beat wimpy cores, most of the time

Urs Hölzle
Google

Slower but energy efficient "wimpy" cores only win for general workloads if their single-core speed is reasonably close to that of mid-range "brawny" cores.

At Google, we've been long-term proponents of multicore architectures and throughput-oriented computing. In warehouse-scale systems[1] throughput is more important than single-threaded peak performance, because no single processor can handle the full workload. In addition, maximizing single-threaded performance costs power through larger die areas (for example, for larger reorder buffers or branch predictors) and higher clock frequencies. Multicore architectures are great for warehouse-scale systems because they provide ample parallelism in the request stream as well as data parallelism for search or analysis over petabyte data sets.

We classify multicore systems as *brawny-core systems*, whose single-core performance is fairly high, or *wimpy-core systems*, whose single-core performance is low. The latter are more power efficient. Typically,

DOI:10.1145/1965724.1965747

FAWN: A Fast Array of Wimpy Nodes

By David G. Andersen, Jason Franklin, Michael Kaminsky, Amar Phanishayee, Lawrence Tan, and Vijay Vasudevan

Abstract

This paper presents a fast array of wimpy nodes—FAWN—an approach for achieving low-power data-intensive data-center computing. FAWN couples low-power processors to small amounts of local flash storage, balancing computation and I/O capabilities. FAWN optimizes for per node energy efficiency to enable efficient, massively parallel access to data.

The key contributions of this paper are the principles of the FAWN approach and the design and implementation of FAWN-KV—a consistent, replicated, highly available, and high-performance key-value storage system built on a FAWN prototype. Our design centers around purely log-structured datastores that provide the basis for high performance on flash storage, as well as for replication and consistency obtained using chain replication on a consistent hashing ring. Our evaluation demonstrates that FAWN clusters can handle roughly 350 key-value *queries per Joule* of energy—two orders of magnitude more than a disk-based system.

1. INTRODUCTION

Large-scale data-intensive applications, such as high-performance key-value storage systems, are growing in both size and importance; they now are critical parts of major Internet services such as Amazon (Dynamo[7]), LinkedIn (Voldemort), and Facebook (memcached).

The workloads these systems support share several characteristics: They are I/O, not computation, intensive, requiring random access over large datasets; they are massively parallel, with thousands of concurrent, mostly independent operations; their high load requires large clusters to support them; and the size of objects stored is typically small, for example, 1KB values for thumbnail images, hundreds of bytes for wall posts, and twitter messages.

The clusters that serve these workloads must provide both high performance and low cost operation. Unfortunately, small-object random-access workloads are particularly ill served by conventional disk-based or memory-based clusters. The poor seek performance of disks makes disk-based systems inefficient in terms of both system performance and performance per Watt. High-performance DRAM-based clusters, storing terabytes or petabytes of data, are expensive and power-hungry: Two high-speed DRAM DIMMs can consume as much energy as a 1TB disk.

The power draw of these clusters is becoming an increasing fraction of their cost—up to 50% of the 3 year total cost of owning a computer. The density of the datacenters that house them is in turn limited by their ability to supply and cool 10–20 kW of power per rack and up to 10–20 MW per datacenter.[13] Future datacenters may require as much as 200 MW,[12] and datacenters are being constructed today with dedicated electrical substations to feed them.

These challenges necessitate the question: Can we build a cost-effective cluster for data-intensive workloads that uses less than a tenth of the power required by a conventional architecture, but that still meets the same capacity, availability, throughput, and latency requirements?

The FAWN approach is designed to address this question. FAWN couples low-power, efficient CPUs with flash storage to provide efficient, fast, and cost-effective access to large, random-access data. Flash is faster than disk, cheaper than DRAM, and consumes less power than either. Thus, it is a particularly suitable choice for FAWN and its workloads. FAWN represents a class of systems that targets both system balance and per node energy efficiency: The 2008-era FAWN prototypes used in this work used embedded CPUs and CompactFlash, while today a FAWN node might be composed of laptop processors and higher-speed SSDs. Relative to today's highest-end computers, a contemporary FAWN system might use dual or quad-core 1.6 GHz CPUs with 1–4GB of DRAM.

To show that it is *practical* to use these constrained nodes as the core of a large system, we designed and built the FAWN-KV cluster-based key-value store, which provides storage functionality similar to that used in several large enterprises.[7] FAWN-KV is designed to exploit the advantages and avoid the limitations of wimpy nodes with flash memory for storage.

The key design choice in FAWN-KV is the use of a *log-structured per-node datastore* called FAWN-DS that provides high-performance reads and writes using flash memory. This append-only data log provides the basis for replication and strong consistency using *chain replication*[23] between nodes. Data is distributed across nodes using consistent hashing, with data split into contiguous ranges on disk such that all replication and node insertion operations involve only a fully in-order traversal of the subset of data that must be copied to a new node. Together with the log structure, these properties combine to provide fast failover and fast node insertion, and they minimize the time the affected datastore's key range is locked during such operations.

The original version of this paper was published in *Proceedings of the 22nd ACM Symposium of Operating Systems Principles*, October 2009.

intel®

Brawny vs. Wimpy Cores

Hosting

Virtualize across cores — Dedicated / Lightweight

Content Delivery

Scale (eg perf, memory) — Low performance, High I/O

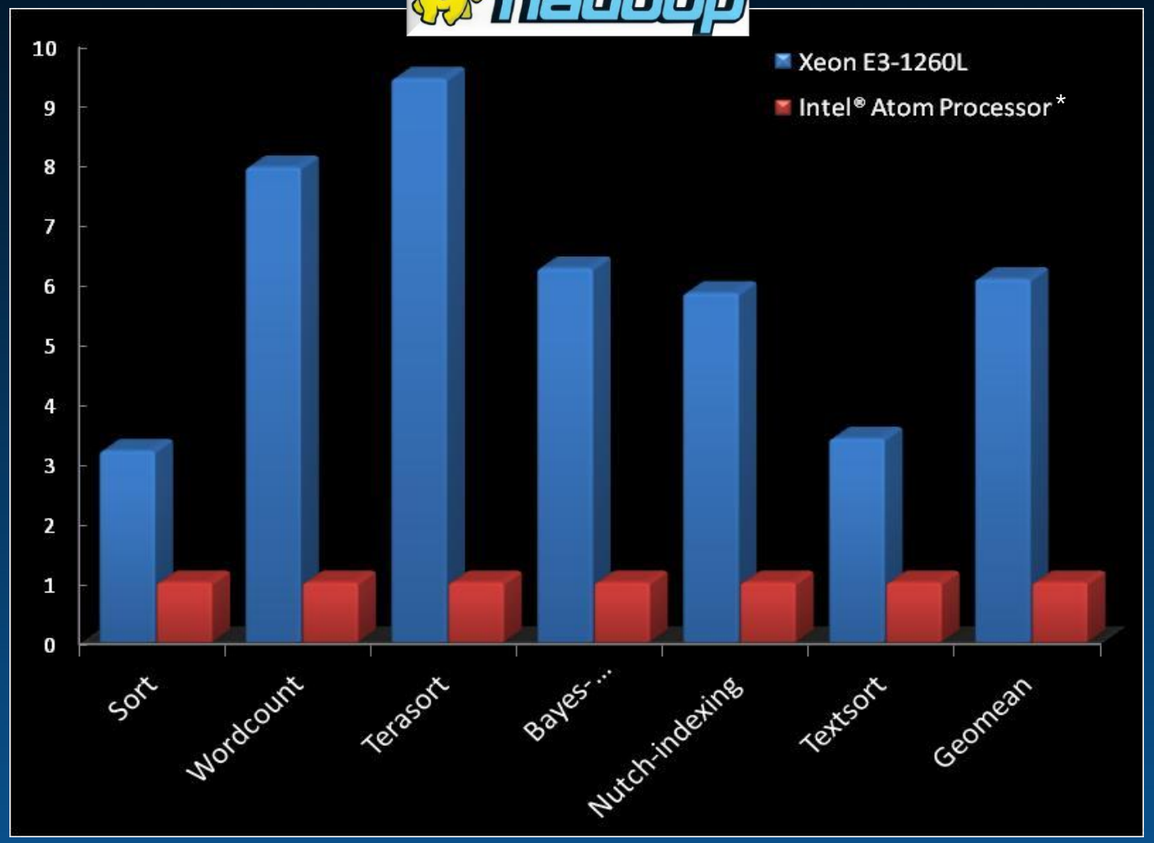
Analytics

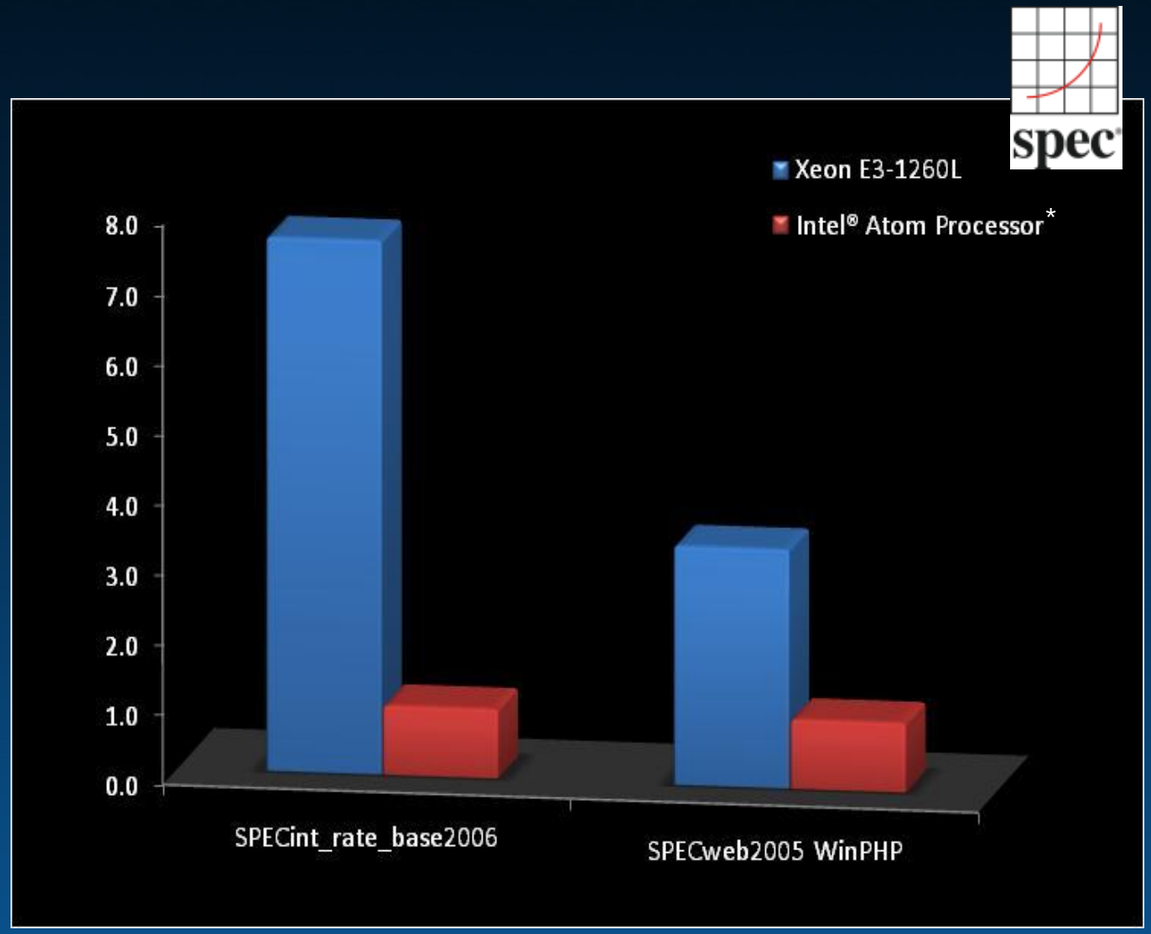
CPU Intensive — I/O Intensive

Distributed Memory Caching

Lower Latency — Small "Blast" Radius

Brawny vs. Wimpy Cores





Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

1 SPECint & SPECweb are registered trademarks of the Standard Performance Evaluation Corporation. For more information see

* Measured on Intel® Atom D525 processor

It Doesn't Matter

Leadership Building Blocks for Micro Servers

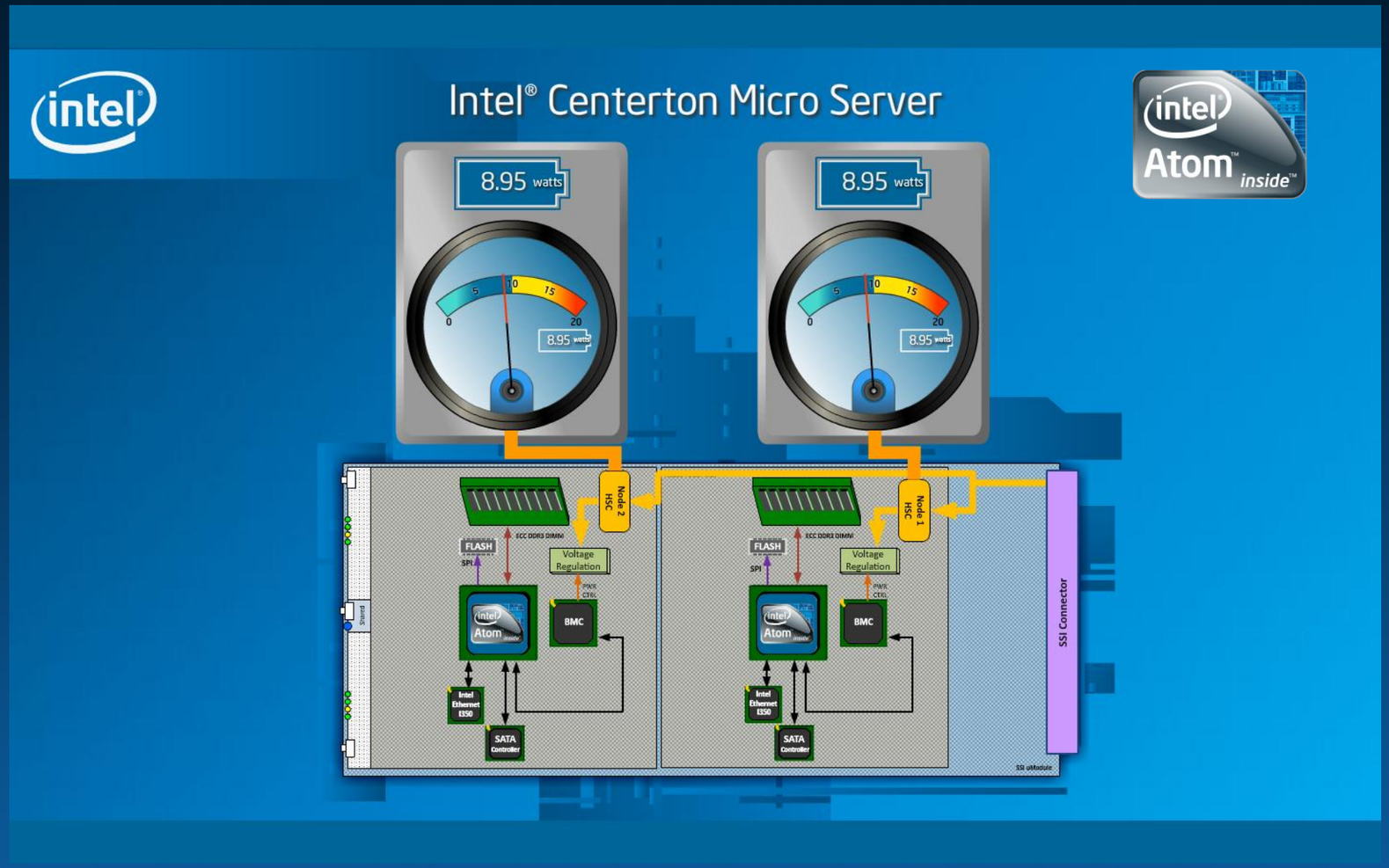
2010	2011	2012	2013
45 Watt	45 Watt	45 Watt	Haswell 22nm
30 Watt	20 Watt	17 Watt	
	15 Watt		
		"Centerton" 6W (On track for 2H'12)	Avoton 22nm

Available Now

64 bit
Intel® VT
ECC Memory
SW Compatibility

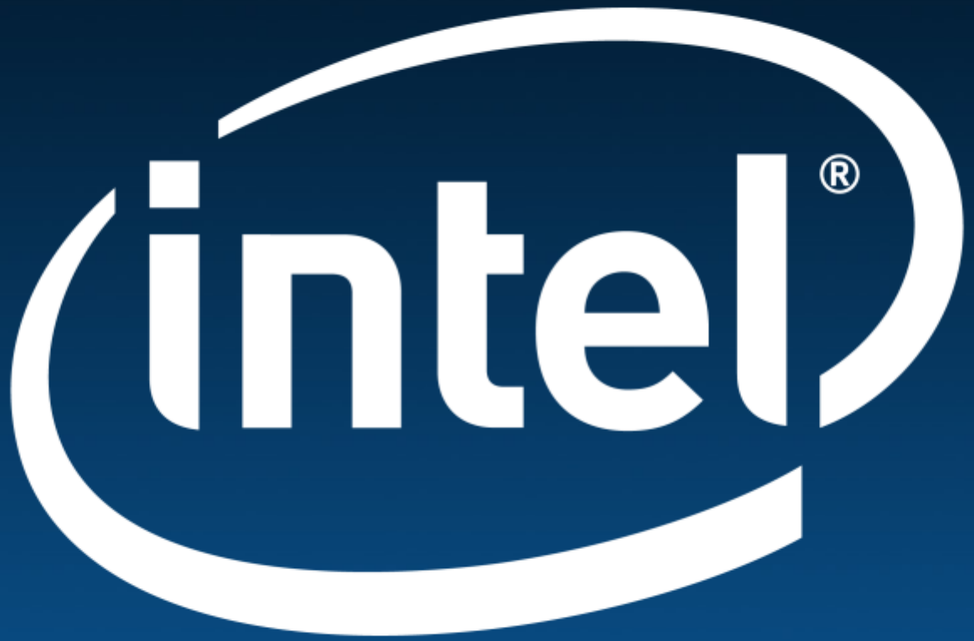
All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.
*2011 Intel® server processors are based on Intel® microarchitecture codename Sandy Bridge

First "Centerton" live demo with sub-9W power / node[1]



1 Node consists of Centerton SoC, 4GB memory, SATA controller, GbE controller, VRs & BMC

Industry Developments


intel

Legal Disclaimer

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL® PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. INTEL PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.

Intel may make changes to specifications and product descriptions at any time, without notice.

All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Any code names featured are used internally within Intel to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Intel to use code names in advertising, promotion or marketing of any product or services and any such use of Intel's internal code names is at the sole risk of the user.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to http://www.intel.com/performance

Intel, Intel Inside, the Intel logo, Centrino, Centrino Inside, Intel Core, Intel Atom and Pentium are trademarks of Intel Corporation in the United States and other countries.

Material in this presentation is intended as product positioning and not approved end user messaging.

This document contains information on products in the design phase of development.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. Click here for details

Hyper-Threading Technology requires a computer system with a processor supporting HT Technology and an HT Technology-enabled chipset, BIOS and operating system. Performance will vary depending on the specific hardware and software you use. For more information including details on which processors support HT Technology, see here

Intel® Turbo Boost Technology requires a PC with a processor with Intel Turbo Boost Technology capability. Intel Turbo Boost Technology performance varies depending on hardware, software and overall system configuration. Check with your PC manufacturer on whether your system delivers Intel Turbo Boost Technology. For more information, see http://www.intel.com/technology/turboboost.

No computer system can provide absolute security under all conditions. Intel® Trusted Execution Technology (Intel® TXT) requires a computer system with Intel® Virtualization Technology, an Intel TXT-enabled processor, chipset, BIOS, authenticated Code Modules and an Intel TXT-compatible measured launched environment (MLE). The MLE could consist of a virtual machine monitor, an OS or an application. In addition, Intel TXT requires the system to contain a TPM v1.2,as defined by the Trusted Computing Group and specific software for some uses. For more information, see here

The original equipment manufacturer must provide TPM functionality, which requires a TPM-supported BIOS. TPM functionality must be initialized and may not be available in all countries.

Roadmap not reflective of exact launch granularity and timing – please refer to ILU guidance

Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM). Functionality, performance or other benefits will vary depending on hardware and software configurations. Software applications may not be compatible with all operating systems. Consult your PC manufacturer. For more information, visit http://www.intel.com/go/virtualization

Intel® AES-NI requires a computer system with an AES-NI enabled processor, as well as non-Intel software to execute the instructions in the correct sequence. AES-NI is available on select Intel® processors. For availability, consult Your reseller or system manufacturer. For more information, see http://software.intel.com/en-us/articles/intel-advanced-encryption-standard-instructions-aes-ni/

Intel product plans in this presentation do not constitute Intel plan of record product roadmaps. Please contact your Intel representative to obtain Intel's current plan of record product roadmaps.

*Other names and brands may be claimed as the property of others.

Legal Information: Performance

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, Go to: http://www.intel.com/performance/resources/benchmark_limitations.htm.

Intel does not control or audit the design or implementation of third party benchmarks or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmarks are reported and confirm whether the referenced benchmarks are accurate and reflect performance of systems available for purchase.

Relative performance is calculated by assigning a baseline value of 1.0 to one benchmark result, and then dividing the actual benchmark result for the baseline platform into each of the specific benchmark results of each of the other platforms, and assigning them a relative performance number that correlates with the performance improvements reported.

SPEC, SPECint, SPECfp, SPECrate. SPECpower, SPECjAppServer, SPECjEnterprise, SPECjbb, SPECompM, SPECompL, and SPEC MPI are trademarks of the Standard Performance Evaluation Corporation. See http://www.spec.org for more information.

TPC Benchmark is a trademark of the Transaction Processing Council. See http://www.tpc.org for more information.

SAP and SAP NetWeaver are the registered trademarks of SAP AG in Germany and in several other countries. See http://www.sap.com/benchmark for more information.

INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS". NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, reference www.intel.com/software/products.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.