Image source: San Diego Supercomputer Center, Alan Decker

The San Diego Supercomputer Center Cloud is a world class storage repository that enables data sharing among the global community of researchers. Open-source solutions help facilitate flexible access to the cloud, whether it's by humans using URLs or machines using APIs.

RELEASING CLOUDS OF DATA TO FACILITATE THE SPREAD OF HUMAN KNOWLEDGE

The San Diego Supercomputer Center has created a massive storage cloud that makes research data of virtually any type accessible by others, easily and at a low cost.

The need to share vast bodies of data has long been a staple of research on subjects from the celestial to the terrestrial. Whether it's hyper-spectral imagery from deep space or temperature readings from an ordinary salt marsh, that information is often the key to advancing scientific understanding.

Researchers have traditionally archived data on tape or other removable media, typically on isolated storage that is chosen with safekeeping in mind, instead of optimizing its accessibility by others. The limits of that approach are all the more apparent in the context of increasingly mainstream Big Data capabilities for data mining and analysis on massive data sets such as those created by large-scale simulations in astrophysics or genetic engineering.

To make scientific data sets readily available to researchers all over the world, engineers at the San Diego Supercomputer Center (SDSC) have created a storage cloud with an initial raw capacity of 5.5 petabytes (PB), which is able to scale to tens of petabytes. Open-source-based solutions, with CentOS* and OpenStack* at their core, are the software technologies of choice for this project, and nodes based on Intel® Xeon® processors deliver cost-effective performance that helps drive success.

CHALLENGE

Make research data—much of which is unstructured—available to researchers all over the world, moving beyond the notion of "write once, read never" data archiving to facilitate greater data accessibility and sharing, in support of academic and commercial research.

SOLUTION

The San Diego Supercomputer Center (SDSC) at the University of California, San Diego, created a very large cloud based on OpenStack* and other open-source software running on Intel® Xeon® processor-based servers that provides low-cost, fast access using either HTTP queries or API-based access.

RESULT

Researchers have a simple means to share data, allowing them to focus on their research interests instead of storage infrastructure. The SDSC Cloud provides very high interoperability with other clouds, software cost savings without vendor lock-in, and agility from a robust community and a fast open-source development cycle.

intel®

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

FACILITATING RESEARCH WITH DATA:

SHARING AND PROTECTING THE OUTCOMES OF SCIENTIFIC INQUIRY

An aging tape archive at SDSC that housed several petabytes of unstructured research data needed to be refreshed or replaced. Engineers at the Center considered their options for a suitable object store and concluded that hard-disk-based storage might be a better fit for their mission than the older tape-based approach, particularly in light of the market forces driving down the price of hard-disk capacity.

A key goal of SDSC is to promote the preservation and sharing of scientific data. While the tape archive provided a reliable means to preserve data almost indefinitely, it was not well suited to broad-based sharing. A cloud approach, by contrast, held promise as a large-scale data archive that would also facilitate access to its contents by the researchers who generated the data, as well as by others with the appropriate permissions.

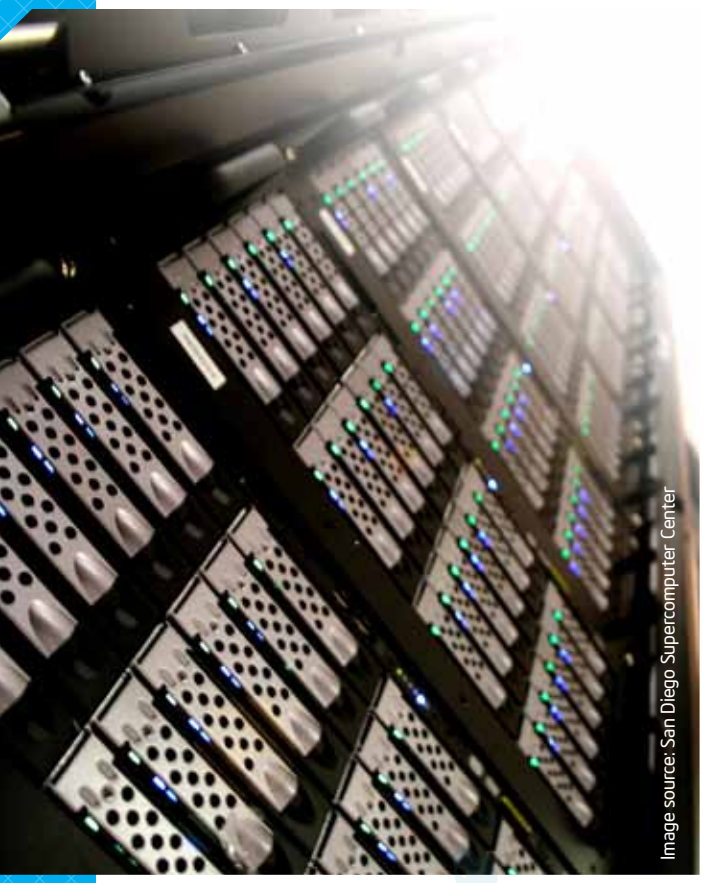


Image source: San Diego Supercomputer Center

The trend in research institutions is to not only provide the outcome and findings of studies through published papers, but to also make the underlying data sets available. As a result of that requirement and broader trends, SDSC initiatives are supporting both local (on-campus) and national research computing needs, including an infrastructure comprising high-bandwidth networking, high-performance computational resources, and large-capacity data storage, to further scientific research.

For example, this infrastructure supports data collection and analysis from the rapidly proliferating high-throughput scientific instruments in a variety of disciplines, such as mass spectrometers, astronomical instruments, and DNA sequencers. SDSC helps collect, transfer, store, and analyze the large bodies of data associated with these instruments and also makes the data available to help facilitate other research efforts. Some key components of this effort include the following:

- **High-performance computing clusters**, including Gordon, a 16,384-core supercomputer
- **Data Oasis, a large-scale parallel distributed file system based on Lustre***, which facilitates staging data for low-latency access during computation
- **SDSC Cloud, an OpenStack-based environment** for the preservation and sharing of research data

Using these systems, research facilities can transfer their large data sets to SDSC for analysis, storage, and sharing. The result is low-cost, long-term storage that is more suitable than tape and that facilitates requirements such as those imposed by the National Science Foundation to maintain robust data-management plans.

MASSIVE DATA, GLOBAL ACCESS:

CHOOSING TECHNOLOGIES FOR A WORLD-CLASS OPEN STORAGE CLOUD

To unlock the value of the data stored in a difficult-to-access state in robotic tape libraries, the SDSC team began to investigate various cloud-based object storage systems. It became apparent early on that the majority of these systems were based on proprietary APIs and hardware, which would limit flexibility and carry significant cost over the life of the solution.

As the team turned its attention toward open-source, standards-based software to overcome the inflexibility and cost hurdles of the proprietary solutions, OpenStack emerged as the consensus preference, particularly in terms of interoperability. Freely available under the Apache license, OpenStack uses software technologies to provide data replication across nodes, which allows for the use of low-cost, off-the-shelf hard drives and other system components.

Another advantage proved to be Intel's involvement in the OpenStack community, which has included significant contributions to the project. This relationship has helped ensure high optimization for the features and capabilities of Intel Xeon processors.

Intel's broad industry and ecosystem involvement extends to collaboration with many equipment providers, such as Aberdeen, Arista Networks, and Dell. Those relationships benefit the SDSC Cloud, the main hardware components of which include the following:

- **Dell PowerEdge* R610 and R620 Servers**, rack-optimized, 1U systems based on the Intel Xeon processor 5600 series and Intel Xeon processor E5-2600 product family, respectively, used as both proxy and storage nodes.
- **Aberdeen x539 Storage Servers**, rack-optimized 5U systems based on the Intel Xeon processor 5500 series and configured with 24 SATA hard drives, each with 2 terabyte (TB) capacity (48 TB total per server).
- **Intel® Ethernet Network Daughter Card X520-DA2 / I350-T2 (Dell 430-4935)**, with two 10 Gigabit Ethernet SFP+ Direct Attach ports and two Gigabit Ethernet 1000BASE-T ports.
- **Redundant Arista Networks 7508 switches**, each providing 384 10 Gigabit Ethernet ports for more than 10 terabits per second of non-blocking, IP-based connectivity.

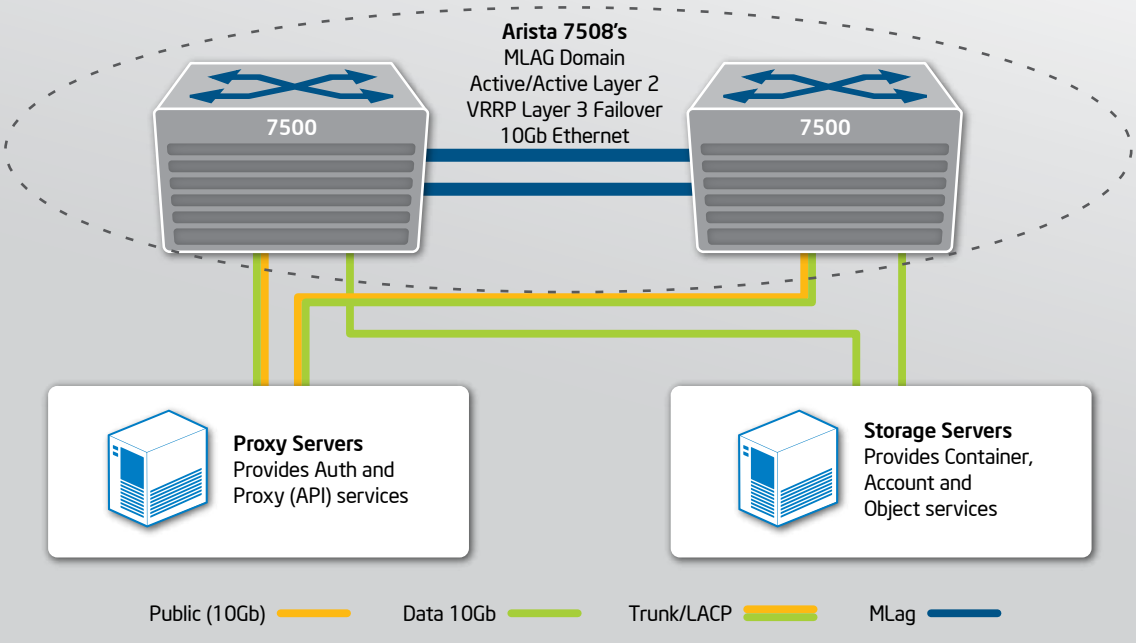


Image source: San Diego Supercomputer Center

The San Diego Supercomputer Center Cloud provides an initial total of 5.5 petabytes (PB) of raw storage, designed to scale up to 100 PB. It is built on OpenStack* Swift object storage, and it uses nodes based on Intel® Xeon® processors, networked using Intel® Ethernet Network Daughter Cards and Arista 7508 switches.

ENVISIONING AND ENACTING CLOUD CAPABILITIES:

THE BROAD-BASED DESIGN AND FEATURES OF THE SDSC CLOUD

Beyond the Limitations of Existing Systems

Developing the SDSC Cloud was, in part, a response to some common limitations of the storage systems at traditional supercomputer centers. Those systems were not designed to support a worldwide research-data repository, because of considerations that include the following:

- **Access**. It is often difficult to access archival data on traditional storage systems because of high-latency, low-bandwidth user interfaces.
- **Sharing**. Multiple users find it difficult to share archival data, and mechanisms such as open APIs to overcome that limitation are typically lacking.
- **Intent**. Large bodies of data, particularly high-performance computing (HPC) simulations, are often archived with no intention of ever being retrieved, in a "write once, read-never" modality.

As the sizes of data stores increase exponentially and technologies evolve to analyze them quickly and flexibly, collaborative research endeavors make access to shared data increasingly valuable.

The Arrival of a New Approach

Cloud technologies are an invitation to reimagine data-repository designs. Using OpenStack Swift, the SDSC Cloud design team addressed the latency and multi-user issues of traditional tape-based systems, creating an approach that assumes archived data will be accessed again and again.

The team established simple interfaces to access data and manage permissions to it. An open range of mechanisms can be used for access, including interfaces to various commercial and open-source-based products. Encryption and transaction logging help support regulatory requirements for sensitive data, setting the stage for applications such as the following:

- Shared, published, and curated data collections
- HPC simulation data storage and sharing
- Web/portal applications and site hosting
- Application integration using supported APIs
- Serving images, videos, and other media
- Providing backup services

Customers across academia and industry can access resources from the SDSC Cloud, on either an on-demand basis or on dedicated, customer-purchased hardware.

Image source: San Diego Supercomputer Center

For more information on the SDSC Cloud, including system information, pricing, and usage information, see **http://cloud.sdsc.edu/**

JOINING UP WITH POWERFUL FORCES:
OPENSTACK CAPABILITIES AND INTEL'S CLOUD COMMITMENT

Working with the Community

Intel shares its passion with the rest of the open-source community to inspire fresh ideas, accelerate innovation, and lower costs. For over two decades, Intel's contributions to open-source projects have enabled a breadth of solutions to run exceptionally well on Intel® architecture.

As a member of the OpenStack community, Intel helps advance the OpenStack project to create an enterprise-ready, open-source cloud infrastructure. Our contributions have been instrumental in enhancing its security and efficiency, while also helping streamline deployment and optimize the platform's ability to deliver optimal value from data center hardware.

Intel's upstream contributions to collaborative open-source projects such as OpenStack, coupled with its downstream collaboration with ecosystem partners, helps open cloud solutions such as the SDSC Cloud break new ground.

Broad Effort to Enable the Open Cloud

Intel recognizes that a strong, broadly adopted set of industry standards is necessary for cloud infrastructures such as the SDSC Cloud to thrive. With the benefit of such standards, open, interoperable solutions will develop in cloud computing and related fields.

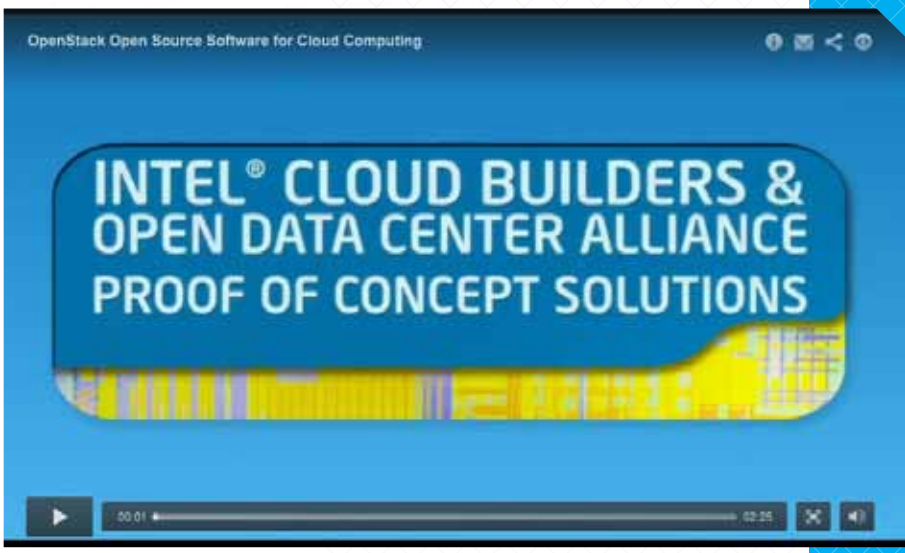
To that end, Intel maintains a broad range of industry involvement, dedicated to enabling the present and future of the open cloud.

- **Cross-industry initiatives**. Intel plays a major role in efforts to advance the ecosystem as a whole, such as acting as a technical advisor to the Open Data Center Alliance (ODCA). The mission of this alliance is to define a roadmap of requirements and usage models in areas such as security, automation, common management, and service transparency.
- **Helping support a fast path to the cloud**. By delivering targeted tools, programs, and resources, Intel is helping organizations rapidly adopt cloud technologies. Based on Intel's own early adoption of OpenStack, the Intel® Cloud Builders program is an alliance of industry thought leaders creating best-of-breed cloud reference architectures.

Through these and many other efforts, Intel is helping IT organizations reduce the risk, complexity, and cost associated with adopting cloud infrastructures.

To learn more about how Intel is helping to advance the promise of cloud computing based on OpenStack, check out the video at **http://software.intel.com/en-us/videos/openstack-open-source-software-for-cloud-computing**.

openstack™



COLLABORATE

DRIVING ADVANCEMENTS FOR THE CLOUD:

INNOVATIONS FROM THE INTEL® XEON® PROCESSOR E5-2600 PRODUCT FAMILY

The SDSC Cloud relies on the Intel Xeon processor to store, share, and protect the data entrusted to it. Many of the first nodes deployed as part of the project were based on the previous-generation Intel Xeon processor 5600 series. The hardware presently being deployed is based on the Intel Xeon processor E5-2600 product family, which may deliver an increase in performance of up to 80 percent in some implementations.[1]

Embracing the new processor technology complements the ongoing enablement and contributions for the open-source community, including the OpenStack project, the Linux kernel, and many others. That combination of hardware and software commitment positions environments such as the SDSC Cloud for increasing benefits as time goes on.

Performance innovations are built into the Intel Xeon processor E5-2600 product family that benefit the entire environment, creating a balanced set of advancements that permeates all of the key aspects of the computing platform.

- **Processing**. The Intel Xeon processor E5-2600 product family supports up to eight cores (16 threads) per socket, plus Intel® Turbo Boost Technology 2.0[2] to help handle peak workloads, powering the ongoing growth of the cloud.
- **Memory**. Support for up to 768 gigabytes of system memory and 20 megabyte last-level cache help deliver state-of-the art headroom in the memory subsystem for even the most data-intensive workloads.
- **I/O**. Intel® Integrated I/O with PCI Express* 3.0 can up to triple[3] the movement of data into and out of the processor, keeping the execution cores supplied with data, even from I/O-intensive cloud workloads.

Hardware-based Security Features Help Protect Sensitive Data

The SDSC Cloud houses a wide variety of research data. While most scientific data is openly shared, some data is increasingly subject to regulatory requirements such as the Health Insurance Portability and Accountability Act (HIPAA) and the Federal Information Security Management Act (FISMA).

Contributions by Intel to the OpenStack project provide support for security features built into the Intel Xeon processor E5-2600 product family, providing the ability to control data access and meet regulatory requirements.

- **Intel® Advanced Encryption Standard New Instructions** accelerates data encryption and decryption, dramatically reducing the overhead associated with pervasive encryption of sensitive data.
- **Intel® Trusted Execution Technology** helps provide hardened protection for data—including cloud object stores—by creating a trusted foundation to reduce the infrastructure exposure to malicious attacks.

COST-EFFECTIVE, LONG-TERM INTEROPERABILITY:
ESTABLISHING ONGOING CAPABILITIES FOR INNOVATION

Using OpenStack as the basis of the SDSC Cloud has helped offer storage to customers at a lower cost than would be possible with proprietary solutions. That low cost, in turn, helps drive wider adoption, enabling economies of scale that add to the financial soundness and overall success of the project and leading to a sustainable business model.

OpenStack and the other open-source tools used by the SDSC Cloud enable customers to access durable, high-speed, on-demand storage services using a wide variety of access methods. OpenStack supports APIs, command-line operation, and multiple client options.

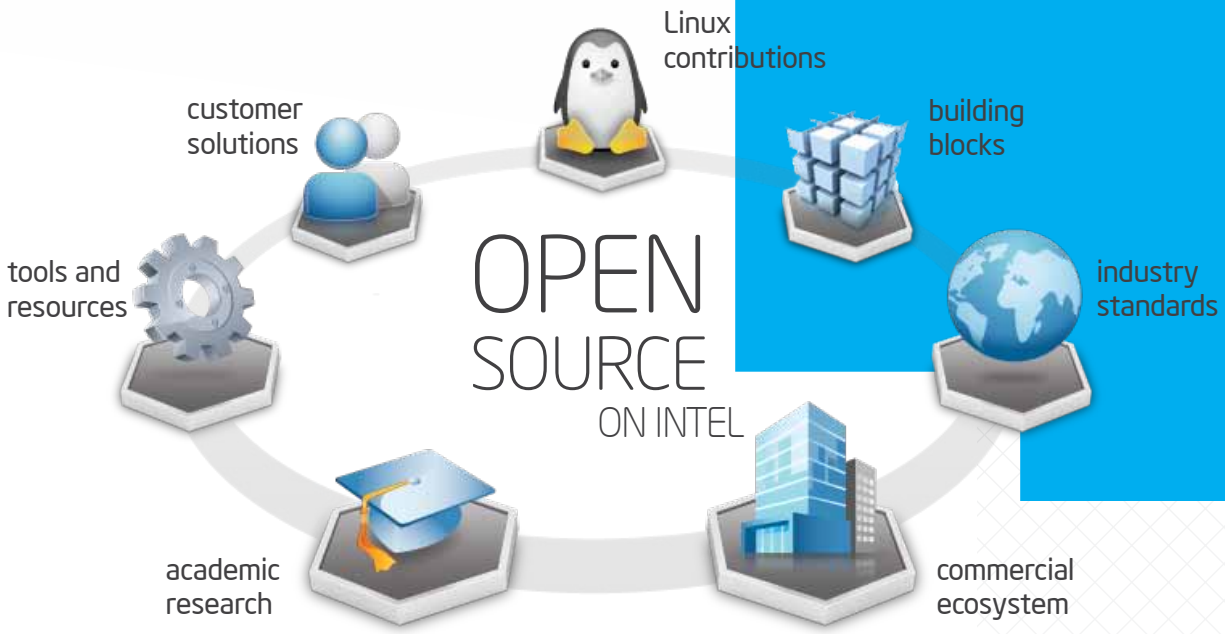
Users can collect data directly from scientific devices and store it in the cloud by leveraging the API, groups can store and share data with community accounts, and data can be accessed from anywhere with the custom web interface. SDSC Cloud storage relieves end users from the overhead of purchasing and managing their own large-scale, on-site storage solutions.

The open-source basis for the SDSC Cloud has also delivered advantages in terms of adding new features and capabilities when needed on a shorter development cycle than would be possible with commercial software. Developing or sourcing capabilities from the community has proved to be a viable means of extending functionality to meet specific emerging needs.

> "The OpenStack object store offers extremely scalable capacity for any type of data, accessible by a wide range of APIs. It's also automatically optimized for the performance and security features of Intel Xeon processors."
>
> — Ron Hawkins, Industry Relations Director, San Diego Supercomputer Center

Image source: San Diego Supercomputer Center

Intel takes pride
in being a long-standing member of the open-source community. We believe in open source development as a means to create rich business opportunities, advance promising technologies, and bring together top talent from diverse fields to solve computing challenges. Our contributions to the community include reliable hardware architectures, professional development tools, work on essential open-source components, collaboration and co-engineering with leading companies, investment in academic research and commercial businesses, and helping to build a thriving ecosystem around open source.



www.intel.com/opensource

[1] Performance comparison using best submitted/published 2-socket server results on the SPECfp*_rate_base2006 benchmark as of 6 March 2012. Baseline score of 271 published by Itautec on the Servidor Itautec MX203* and Servidor Itautec MX223* platforms based on the prior generation Intel® Xeon® processor X5690. New score of 492 submitted for publication by Dell on the PowerEdge* T620 platform and Fujitsu on the PRIMERGY RX300 S7* platform based on the Intel® Xeon® processor E5-2690 product family. For additional details, please visit www.spec.org. Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase.

[2] Requires a system with Intel® Turbo Boost Technology. Intel Turbo Boost Technology and Intel Turbo Boost Technology 2.0 are only available on select Intel® processors. Consult your PC manufacturer. Performance varies depending on hardware, software, and system configuration. For more information, visit **www.intel.com/go/turbo**.

[3] Intel internal measurements of maximum achievable I/O R/W bandwidth (512B transactions, 50-percent reads, 50-percent writes) comparing Intel® Xeon® processor E5-2680 product family-based platform with 64 lanes of PCIe* 3.0 (66 GB/s) versus Intel Xeon processor X5670-based platform with 32 lanes of PCIe* 2.0 (18 GB/s). Baseline configuration: Green City system with two Intel® Xeon® processor X5670 (2.93 GHz, 6 cores), 24 GB memory at 1333 MHz, 4 x8 Intel internal PCIe 2.0 test cards. New configuration: Rose City system with two Intel Xeon processor E5-2680 product family (2.7 GHz, 8 cores), 64 GB memory at 1600 MHz, 2 x16 Intel internal PCIe 3.0 test cards on each node (all traffic sent to local nodes).

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information. The products described in this document 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. Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order. Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or by visiting Intel's Web Site **www.intel.com**.

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 **www.intel.com/performance**.

0513/NKR/MESH/PDF 328629-001US