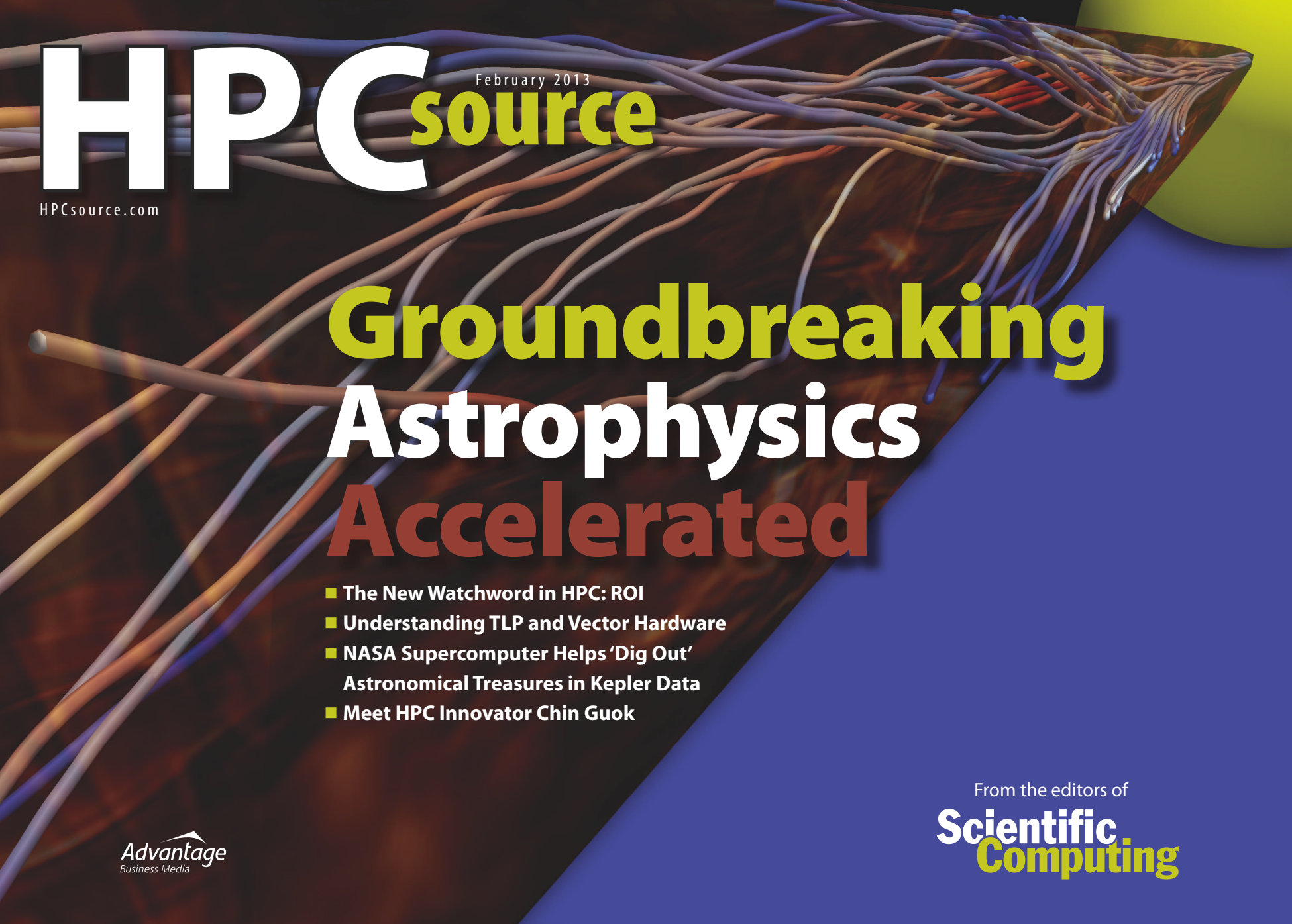
HPCsource

February 2013

HPCsource.com



Groundbreaking Astrophysics Accelerated

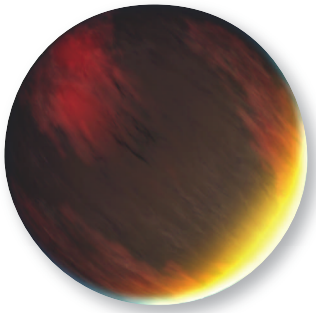
- The New Watchword in HPC: ROI
- Understanding TLP and Vector Hardware
- NASA Supercomputer Helps 'Dig Out' Astronomical Treasures in Kepler Data
- Meet HPC Innovator Chin Guok

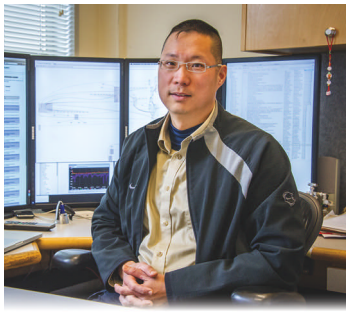
Advantage Business Media

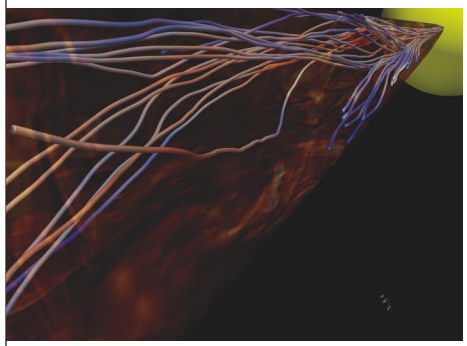
From the editors of Scientific Computing


"Dell's high-performance computing solution lets our researchers spend 50% more time focusing on the big picture."
Dr. Nathan Crawford
Director, Molecular Modeling
University of California, Irvine
Do more with Dell cloud solutions
See how we helped the University of California, Irvine's School of Physical Sciences do 50% more research in 50% of the computing footprint with a customized high-performance computing solution by Dell, Intel® and QLogic.
Learn more at Dell.com/casestudies.
DELL
The power to do more

Contents







NASA Supercomputer Helps 'Dig Out' Astronomical Treasures in Kepler Data

Pleiades handles pipeline's most data-intensive segments

Jill Dunbar

It seems like every time you turn around, NASA reveals a new astronomical discovery made possible by its Kepler space observatory. Most recently, in January of this year the Kepler mission team announced the discovery of 461 new planet candidates over the past 12 months — a 20 percent increase over the previous year. A year ago, 33 candidates in the Kepler data had been confirmed as planets. Today, there are 105 confirmed planets.

To find new planets, Kepler looks for tiny dips in the brightness of a star (as viewed from Earth) that occur when a planet crosses in front of, or transits, its host star. The observatory's large photometer continuously measures the brightness of more than 150,000 stars in the Milky Way to search for these transiting planets, while back on Earth NASA's most powerful supercomputer, an SGI system named Pleiades, crunches the data that makes these discoveries possible.

Now in its third year of operation, the Kepler mission relays about 100 gigabytes of data back to Earth every month in its quest to discover Earth-size and smaller planets — especially those in the "habitable zone" of their stars, where liquid water might exist. Some of that data is processed in the Kepler Science Operations Center (SOC) at NASA Ames Research Center, Moffett Field, CA. Pixels are calibrated, combined to form light curves, and corrected for systematic errors.

The data is then copied to Pleiades to handle the most data-intensive segments of the SOC pipeline — the pre-search data conditioning, transiting planet search, and data validation modules. The processing of light curves on Pleiades typically utilizes about 3,000 nodes for 20 hours, versus more than a month on SOC computers.

Background image: NASA announced that the Kepler space telescope, designed to find Earth-sized planets in the habitable zone of sun-like stars, discovered its first five new exoplanets, Kepler 4b, 5b, 6b, 7b and 8b, on January 4, 2010. These planets are known as "hot Jupiters" because of their high masses and extreme temperatures. The exoplanets range in size from similar to Neptune to larger than Jupiter.

The compute capability of the Pleiades system, which was recently expanded to 129,024 cores running at 1.24 petaflops sustained performance, "allows us to analyze Kepler data in hours or days rather than weeks or months," said Shawn Seader, a scientific programmer who develops software code for the SOC pipeline. With this speed-up, Seader explained, the SOC team can take advantage of algorithm improvements much more quickly and "dig out more of the transit signals buried in the data."

One of the most exciting aspects of the Kepler mission is that it's not just NASA scientists who are digging out these astronomical treasures. Thanks to a decision to make Kepler SOC pipeline results and data available to the public for analysis, amateur astronomers and scientists from around the globe are contributing to the search for other worlds beyond our solar system by discovering their own planet candidates, which now number 2,740.

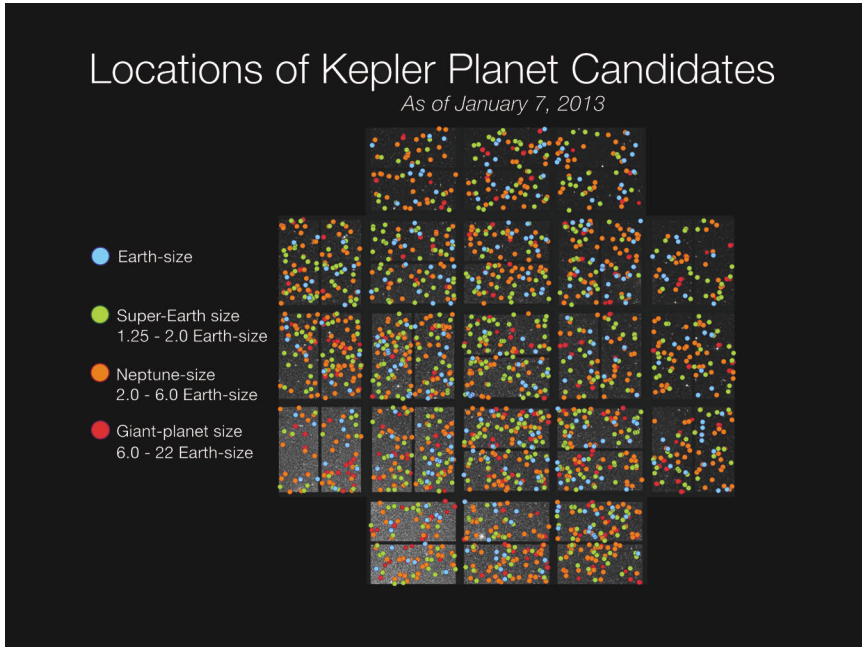
Among the most recent discoveries made possible by Kepler and Pleiades are:

- a planet candidate with a radius just 1.5 times that of Earth — the first habitable-zone super-Earth around a sun-type star (KOI-172.02)
- a planet orbiting a double star that is orbited by a second distant pair of stars (PH1, also known as Kepler-64b)
- the first planet confirmed to orbit within a star's habitable zone, the region around a star where liquid water might exist (Kepler-22b)
- the first transiting circumbinary system — two planets (Kepler-47b and c) orbiting two suns, which proved that more than one planet can form and persist in the stressful realm of a binary star system

The most dramatic increase in discoveries since February 2012 has been in the number of Earth-sized and "super Earth" (just slightly larger than the Earth) candidates, which grew by 43 and 21 percent respectively. Coinciding with the January release of Kepler mission data, Francois Fressin, of the Harvard-Smithsonian Center for Astrophysics, presented findings from a Kepler data analysis at the 221st meeting of the American Astronomical Society in Long Beach, CA. Results show that 50 percent of stars from the 2012 candidate catalog have a planet of Earth-size or larger in a close orbit. By including even larger planets, which have been detected in wider orbits up to that of the Earth, this number reaches 70 percent.

Courtesy of NASA Ames Research Center/Kepler Mission

With astronomers and enthusiasts from around the globe analyzing the data gathered from Kepler and processed by Pleiades, stay tuned for plenty of exciting NASA announcements in the next few years. Perhaps we will soon confirm the discovery of the first habitable Earth-sized planet and transform our understanding of the universe. HPC



Kepler's planet candidates. This relatively small patch of sky being observed by Kepler's spacecraft is potentially teeming with planetary systems, with planets spanning a vast range of sizes compared to Earth. *Courtesy of Natalie Batalha, San Jose State University*

FOR FURTHER INFORMATION

- Kepler Mission: www.kepler.nasa.gov
- Pleiades Supercomputer: www.nas.nasa.gov/hecc/resources/pleiades.html
- Publicly available Kepler data and planet searching tools: www.planethunters.org

Jill Dunbar is a senior writer at NASA Ames Research Center, NASA Advanced Supercomputing (NAS) Division. She may be reached at editor@ScientificComputing.com.

The New Watchword in HPC: ROI

To secure big money, it is becoming more important to talk about returns on investments

Steve Conway

The history of computing machines dates at least as far back as the 19th century, but the development of modern electronic computers, including HPC systems, was spurred mainly by the national security/military requirements of World War II (1939-1945) and the subsequent Cold War (1945-1991).

For half a century, therefore, the primary rationale for funding scientific computers — which would come to be called supercomputers in the 1960s — was national security. The primary funders were the governments of the U.S. and its closest allies. By the end of this period, total spending in the worldwide HPC ecosystem for servers, storage, management software, applications and service reached about $2 billion per year.

The long-desired end of the Cold War altered this upward trajectory. It brought about a sudden, substantial drop in funding, especially from the world's largest HPC customer: the U.S. government. As those of us who witnessed it remember, the early 1990s were a wrenching period in HPC history. Established vendors struggled to survive, while many newer vendors disappeared. Annual SC conferences during this period tended to be somber affairs where concerns were often expressed about the long-term viability of the supercomputing market.

These concerns turned out to be unnecessary. Thanks mainly to standards-based clusters replacing vector supercomputers as the dominant species of HPC systems, the global HPC market skyrocketed during the decade of

the 2000s. In 2011, the HPC ecosystem was worth a record $20.2 billion, 10 times its 1990 value. And the high-end supercomputer market segment looks stronger than ever, driven by the worldwide petascale/exascale race.

The major change that the Cold War ending helped set in motion was a multi-year, ongoing shift in the primary driver of HPC funding and market growth, away from national security and toward expected returns-on-investment (ROI). HPC spending related to defense and national security remains important — it is one of the half dozen largest market segments. But, in today's greatly expanded HPC market, it no longer occupies the kingpin position it formerly held.

In many industrial sectors, such as the aerospace and automotive markets, expected ROI has long been an important rationale for funding HPC. What's newer is the escalating trend toward ROI-based rationales in government and academic HPC centers, especially large centers that purchase leadership-class HPC systems.

One important sign of the ROI trend is the growing number of HPC programs and centers that have added industrial outreach programs. These include the U.S. Department of Energy's INCITE and SciDAC programs, industrial seminars run by PRACE and individual nations in Europe, NCSA's Private Sector Program, and many others. An equally important sign is the growing conviction among the many scientists who participate in IDC's global HPC studies that providing industry with access to leadership-class HPC systems is necessary to advance the competitiveness of their countries.

IDC predicts that politicians and other government officials who control funding for these large systems will increasingly need ROI arguments to gain approval for supercomputers that can cost as much as $400 to 500 million (as opposed to top prices of about $30 million in 1990), particularly in today's still-difficult economies.

ROI arguments are starting to be presented with the added bonus that HPC is a proven innovation accelerator that can help propel economies out of recessions.

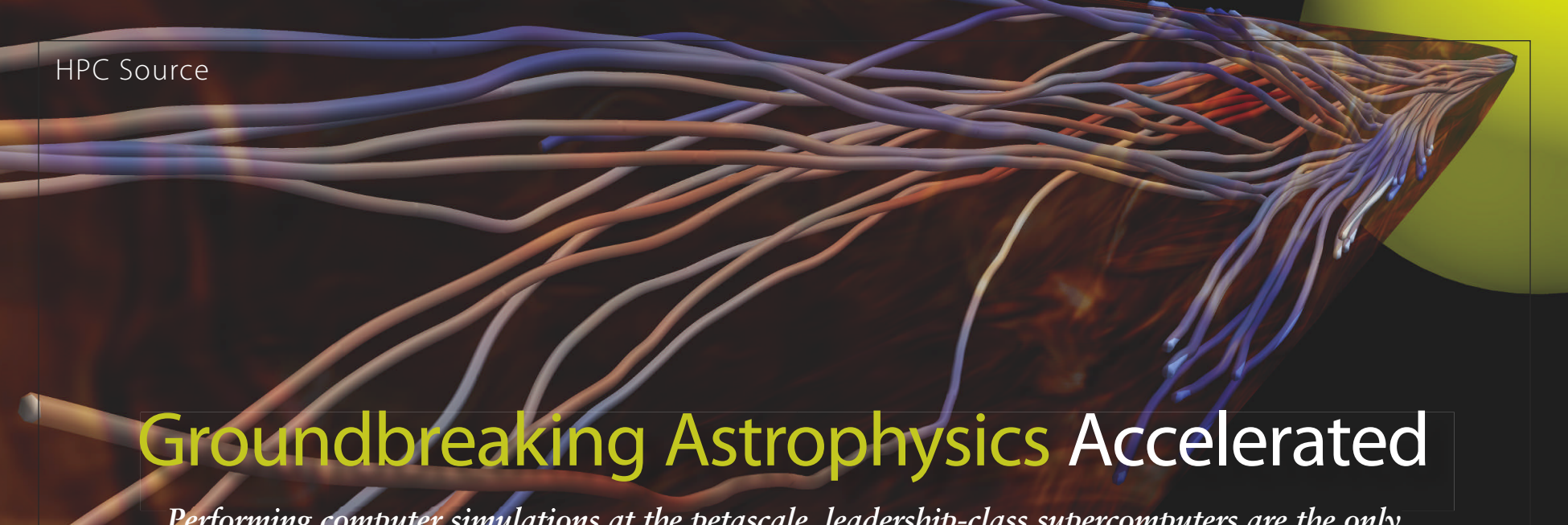
A major glitch is that solid, quantified examples of HPC-based ROI are few and far between today. The same handful of shining examples keeps being rolled out to make the ROI case for HPC.

To help remedy this situation, two years ago, IDC launched the HPC Innovation Excellence Award Program, which to date has added dozens of quantified ROI examples to those that were previously available.

In November 2012, the U.S. Department of Energy's Office of Science and National Nuclear Security Agency awarded IDC an eight-month grant to create two economic models for tracking and predicting HPC innovation and ROI. As part of this grant, IDC also will begin to populate the models with real-world examples.

As the HPC community traverses the Petascale Era en route to the Exascale Era, we also will be moving deeper into the ROI Era. It will be more important than ever to learn how to communicate the inherent value of fundamental science. Without adequate investments in fundamental science to expand our understanding, advances in applied science and industrial engineering would eventually slow to a crawl. The good news is that increased investments in HPC-supported applied science and engineering are likely to benefit funding for fundamental science as well. HPC

Steve Conway is Research VP, HPC at IDC. He may be reached at editor@ScientificComputing.com.



Groundbreaking Astrophysics Accelerated

Performing computer simulations at the petascale, leadership-class supercomputers are the only practical tools for some types of computational science

Imelda G. Francis and Cheryl Drugan

Scientists who provide insights about the physical world, attempt to answer the "big" questions and address the complex problems of our time have a new high-performance computing (HPC) tool literally at their fingertips. Mira — the new petascale IBM Blue Gene/Q supercomputer that was installed at the Argonne Leadership Computing Facility (ALCF) at the end of 2012 — ushers in a new era in scientific supercomputing. It ranks among the world's fastest computers. The 10-petaflops computer is capable of an astounding 10 quadrillion calculations per second. The scale of computation possible on Mira enables current researchers using the ALCF to attempt to answer key questions.

CONDUCTING SCIENCE AT THE ALCF

One team of scientists is investigating the nature of solar wind and the reasons for solar coronal heating. The heliophysics project is led by principal investigator Jean Carlos Perez, of the Space Science Center at the University of New Hampshire. Calculations started at the ALCF in 2011 using Intrepid, an earlier generation Blue Gene/P. The project has been awarded an ALCF computer time grant through the Department of Energy (DOE)-

Figure 1: A snapshot of turbulent magnetic field lines inside a coronal hole that expands from a small patch on the solar surface to five solar radii. Alfven Waves (AWs) (analogous to traveling waves on a string, where the magnetic field lines play the role of the string) are launched by convective motions on the photosphere and propagate in the inhomogeneous Solar atmosphere. As these AWs propagate, they generate reflected AWs that drive turbulence, causing the wave energy to cascade from large- to small-scale motions and heating the ambient plasma. The surface colors around the flux tube depict typical magnetic field fluctuations associated to outward-propagating AWs.

sponsored INCITE (Innovative and Novel Computational Impact on Theory Experiment) program each year since.

The team is conducting unprecedented direct numerical simulations of Alfvén wave (AW) turbulence in the extended solar atmosphere. These simulations are the first to simultaneously account for the inhomogeneities in the density, flow speed and background magnetic field within a narrow magnetic flux tube extending from approximately one solar radius to 11 solar radii, which is precisely the region within which most of the heating and acceleration occur. Because of the extremely large radial extent of the region to be simulated, leadership-class supercomputers like Mira and Intrepid are the only practical tools for some types of computational science.

Researchers will use the ALCF computer simulations to test existing theories of MHD (magnetohydrodynamic) turbulence, develop new theoretical models, and investigate the viability of AW turbulence as a mechanism for generating the solar wind. The results from the simulations will answer a number of key unanswered questions that are being intensely debated in the heliophysics community.

"Our work will provide new insights into the basic properties of inhomogeneous AW turbulence and make a major contribution to scientists' understanding of coronal heating and the origin of the solar wind," says Dr. Perez.

The simulation results are being compared against experimental observations, such as measurements of coronal Faraday rotation fluctuations taken by NASA's old Helios spacecraft.

The results are of added interest to the space physics community, given the preparations now underway for NASA's Solar Probe Plus mission. Researchers will use the simulations to make detailed predictions of the future measurements from the Solar Probe Plus mission. With a planned launch date of 2018, this mission will send a spacecraft to a distance of the closest approach to the Sun of 9.5R$_\odot$, (where R$_\odot$ = 695,990 kilometers is one solar radius inside the region of the Sun that the researchers are simulating numerically).

Supercomputing Sheds Light on the Dark Universe

At Argonne National Laboratory, scientists are using supercomputers to shed light on one of the great mysteries in science today, the Dark Universe. With Mira, a petascale supercomputer at the Argonne Leadership Computing Facility, a team led by physicists Salman Habib and Katrin Heitmann will run the largest, most complex simulation of the universe ever attempted. By contrasting the results from Mira with state-of-the-art telescope surveys, the scientists hope to gain new insights into the distribution of matter in the universe, advancing future investigations of dark energy and dark matter into a new realm. The team's research was named a finalist for the 2012 Gordon Bell Prize, an award recognizing outstanding achievement in high-performance computing.

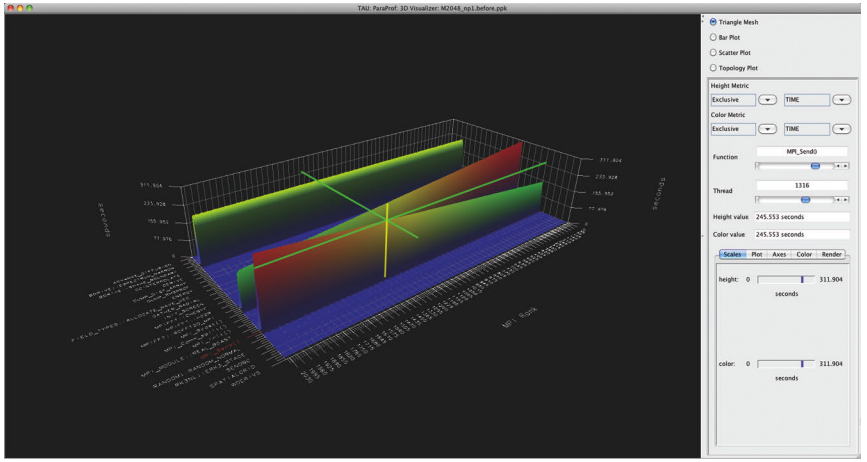


Figure 2: ParProf 3-D window shows the shape of communication routines for a 2,048-core execution.

Perez conducts simulations for the project with the Inhomogeneous Reduced Magnetohydrodynamics (IRMHD) code that he developed to carry out this research. Figure 1 shows a snapshot of the turbulence from recent simulations.

Perez's project does their simulation with the Inhomogeneous Reduced Magnetohydrodynamics (IRMHD) code developed by Perez to carry out this research. The code has excellent weak and strong scaling properties and was extensively tested and benchmarked on Intrepid.

PROBLEM SOLVING

The ALCF encourages collaboration between the researchers using its facility and a team of on-site computational scientists that are there to assist with code refinements and optimization. Collaboration between these groups being a foremost goal, the ALCF holds annual workshops to provide an opportunity to bring the research scientists together with system experts, including also non-ALCF experts on tools and mathematical libraries of importance in HPC. In preparing for its Winter Workshop in January 2012, the ALCF contacted the attendees (including Dr. Perez) and asked them what they hoped to accomplish.

This workshop was almost entirely focused on hands-on work, where workshop participants obtained help from ALCF Catalysts and Performance Engineers to port and optimize the codes for Blue Gene. At least one ALCF computational scientist is assigned to each attendee — Tim Williams was paired with Perez. Sameer Shende, Director of the Performance Research Laboratory, University of Oregon, offered his expertise to those who needed help using the TAU (Tuning and Analysis Utilities) Performance System. Using TAU, a portable profiling and tracing toolkit for performance analysis of parallel programs, the team evaluated Perez's code.

It was after instrumenting the code using TAU and studying the profile information it generated that the group identified a communication volley taking up a suspiciously large amount of run time. Shende reviewed the results with the view shown in Figure 2 and saw telltale slopes in the MPI function calls that indicated an inefficient communication pattern. After some refinements in the code, it was clear significant improvements had been achieved in the asynchronous model of communications being used.

Shende explains: "Before PIs run their applications on the full scale of the machine, it is a good software engineering practice to compare the performance of various communication strategies. Performance evaluation tools such as TAU help with the process."

THE SOLUTION

TAU's profile visualization tool, ParaProf, allowed the research team to quickly identify sources of bottlenecks in the application using the graphical interface. Figure 2 shows the shape of four key routines in ParaProf's 3-D display. Using the information gathered, a non-blocking communication

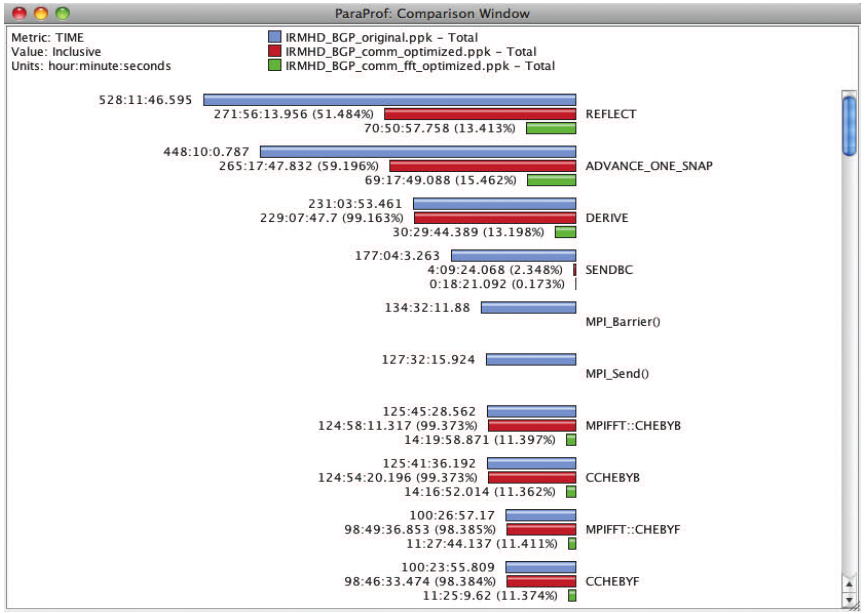


Figure 3: TAU's ParaProf shows the impact of performance optimization after restructuring the communication and FFT routines.

substrate was designed and a more efficient implementation of the underlying FFT (fast Fourier transform) library was deployed. The code ran so much faster after optimizing this communication volley — and using a faster FFT routine optimized specifically for Blue Gene architecture — that Perez was able to conduct significantly more runs in his simulation campaign during 2012. Figure 3 shows how the overall execution time was reduced from 528.18 CPU hours, to just 70.85 CPU hours (13.41 percent) using TAU, for a small 2,048-processor execution on Intrepid.

"By choosing the best communication strategy for Intrepid, the time in some key routines was substantially reduced. This performance improvement translated into big savings — millions of CPU hours. This allowed a more complete and well-resolved physics study with the project's INCITE hours," noted the ALCF's Williams.

CONCLUSION

The refinements made to the code will prove useful in continuing with new INCITE project research. The Perez team was given an INCITE allocation of 53 million hours for 2013 (up from 10 million in 2012) to continue its solar wind and coronal heating research. The new research will promote large-scale numerical simulations of AW turbulence in the extended solar atmosphere and solar wind. The simulations will substantially extend previous, low-resolution simulations to an unprecedented 35 billion-point mesh, to capture more realistic radial variations of the solar atmosphere, as well as the small scale turbulent dynamics that are present in the solar-wind acceleration region.

HPC centers, like the ALCF, will continue to attract computational scientists from around the world and provide them with state-of-the-art tools at the cutting edge of scientific innovation. HPC is the tool of choice for modern scientists because their research often involves extreme environments. There are few, if any, practical ways of gathering insights about these environments other than through computer simulations at the petascale. The era of machines like Mira will provide insights about the physical world at the atomic and subatomic, biological and cosmic levels and, perhaps, answer some of the biggest questions and solve some of the most complex problems along the way. The ALCF is committed to delivering 786 million core hours on Mira in 2013, enabling a wide range of interdisciplinary scientific research to be conducted at Argonne National Laboratory. HPC

Imelda G. Francis is a science writer and editor at Argonne National Laboratory, and Cheryl Drugan is Manager, Business & Technical Communications at Argonne National Laboratory. They may be reached at editor@ScientificComputing.com.

Meet HPC Innovator
Chin Guok

Helping to run the world's fastest science network wasn't originally on Guok's career horizon

Photo: Roy Kaltschmidt, LBNL

Jon Bashor

Although the subject of software defined networking, or SDN, is now a hot topic in the networking community, Chin Guok of the Department of Energy's ESnet (the Energy Sciences Network) has been helping lay the foundation for SDN since 2004. Guok is the technical lead for OSCARS, ESnet's On-Demand Secure Circuits and Advance Reservation System.

OSCARS allows users to quickly configure multi-domain, high-bandwidth virtual circuits that guarantee end-to-end network data transfer performance. Although the networking industry has fielded competing SDN technologies as companies seek leadership in the field, ESnet has been developing and deploying OSCARS since soon after it was first proposed as a research project in August 2004.

OSCARS has grown into a robust production service used by more than 40 other networks, exchange points and testbeds, making it the most widely adopted inter-domain dynamic circuit services application within the global research and networking community.

Broadly defined, SDN makes it easier for software applications to automatically configure and control the various layers of the network. ESnet has been conducting localized tests to innovate, experiment and demonstrate the value of SDN when applied toward end-to-end support of data-intensive science collaborations and applications.

SDN gives users a measure of predictability by giving them more control over their dataflows. Without this ability, the larger science data flows could compete unpredictably with other data transfers for bandwidth making it difficult for scientists to get critical data when it is needed, especially if the data is stored across multiple sites, which is increasingly common.

"When we proposed OSCARS, it was intended to be just within our network and not communicate with any outside networks. We were only operating at Layer 3 of the network providing IP circuits," Guok said. "But, as we worked with more collaborators, we also gained new requirements. We are now running it on Layer 2 as Ether-

net VLAN circuits and are extending the capabilities across multiple domains. The focus is now how to grow this service out beyond the 40 networks as well as up to the software applications."

That extensibility is important, as 80 percent of the traffic currently carried by ESnet has only one end in the ESnet domain and the other end somewhere else. Not only is OSCARS getting bigger, it's also getting faster and easier.

"For SC10, we ran the numbers on what it took to build a virtual circuit to Europe — it required 10 hours of phone calls — about 100 e-mails over three months," Guok said. "Now, with OSCARS, it takes one person about five minutes and one e-mail."

In order to work, OSCARS has to be "technology agnostic" and work independent of the underlying transport resources available. But to succeed, it has to overcome other hurdles, such as the fact that some nations' networks have a pay-per-use model, while DOE researchers are not charged for ESnet. The result has been a number of international agreements working out trades-in-kind.

DOWN TO THE NEXT LEVEL

In November 2012, ESnet partnered with network equipment vendor Infinera on a test to see if OSCARS could be used at Layer 1, the optical transport upon which the higher network layers are built. Using ESnet's testbed in New York, they successfully moved data between Manhattan and Brookhaven National Laboratory on Long Island. What the demo showed is that Infinera's optical transport switch could be dynamically configured by software at that level. This would allow data transfers to reroute themselves based on network demand, finding unused bandwidth and using the resource more efficiently.

The demonstration marked the first time an open architecture with SDN was used to provide traffic-engineered paths at the optical layer and was accomplished through extensions to the OpenFlow protocol. The open source OpenFlow application was developed to work on Layer 2. However, in this demo, it was configured to run on Layer 1 as well.

"Conceptually, this was a very big deal, as it makes the management of network devices much easier," Guok said. "When you allow a Layer 1 device to speak the same 'language' as the rest of the network, it's a very powerful tool for controlling and managing the network."

Computational science administrators and users reflect on 25 years of ESnet, the data network that ties 25,000 scientists to Department of Energy laboratories, computers and instruments. During its anniversary year, ESnet celebrated an upgrade to carry even more data even faster.

There is also an economic factor at play. At each higher layer, the equipment is more expensive as features and capabilities are added. For example, a 100 Gb interface for the optical layer may cost about $50,000, while a 100 Gbps-capable router for Layer 3 would cost about $250,000. This means the cost of sending a bit across the network goes up by 1.5 to 5 times as the traffic moves up to the next layer.

"You really want to get down to the lowest layer you need," Guok said. "One of the prime features of OSCARS is its 'intelligence' — it doesn't matter

OSCARS Timeline

Aug 2004: OSCARS proposal submitted.

Sep 2004: Initial network protocols testing completed (for Juniper M-series platform), first Layer-3 VC manually configured between BNL and FNAL, policed at 50Mb/s.

Feb 2005: Started collaboration with I2's BRUW project.

Apr 2005: First production use of OSCARS VC. Transatlantic LHC 10GE connection between CERN and Chicago severed by fishing boat, causing LHC Service Challenge data to reroute through NY and flow over FNAL production OC12. OSCARS VC set up to carry LHC Service Challenge traffic from NY to Chicago onto FNAL non-production connection at Starlight.

Jun 2005: End-user beta testing with Les Cottrell (SLAC), Sean Flanagan (GA), and Dantong Yu (BNL); jitter measurement testing with SLAC's Datagrid Wide Area Monitoring Infrastructure (DWMI).

Jan 2006: More jitter measurement testing with SLAC.

Mar 2006: Adoption of subset of GÉANT AMPS WSDL service descriptions; Formation of DICE (Dante, Internet2, Canarie, ESnet) Control Plane WG; Venue UCLPv2 meeting in Ontario.

Apr 2006: First Layer-3 interdomain VC dynamically negotiated between I2 (BRUW) and ESnet (OSCARS). Unidirectional 25Mb/s VC from I2 test host in Indianapolis, IN, to ESnet host in Sunnyvale, CA.

Jul 2006: Start of rewrite of OSCARS/BRUW software into Java.

Aug 2006: Successful Layer-3 interdomain reservation between BNL TeraPaths and ESnet OSCARS.

May 2007: Adoption of OGF NMWG topology schema in consensus with DICE Control Plane WG; Collaborative measurements of Hybrid Multilayer Network Data Plane testing with Internet2, USN, and ESnet (using OSCARS VCs)

Aug 2007: First ESnet Layer-2 VC (Ethernet VLAN) configured by OSCARS

Sep 2007: First interdomain topology exchange between GÉANT2, Internet2 and ESnet

Oct 2007: First Layer-2 interdomain VC dynamically negotiated between I2 (HOPI) and ESnet (OSCARS)

Nov 2007: Successful Layer-2 interdomain reservation between BNL TeraPaths and ESnet OSCARS ; FNAL LambdaStation and ESnet OSCARS; GÉANT2 AutoBAHN and ESnet OSCARS; Nortel DRAC and ESnet OSCARS; demonstrated token based authorization concept with OSCARS VC setup.

Feb 2008: Formation of the GLIF GNI-API Task Force of which OSCARS is one of the core representatives.

Apr 2008: Active collaboration with OGF NML-WG to combine work of NMWG and NDL.

May 2008: OSCARS VC operational change to remark over-subscribed packets to scavenger service instead of discarding; OSCARS operational change to support site coordinator role; DICE IDCP v1.0 specification completed.

Dec 2008: Successful control plane interdomain interoperability between ESnet OSCARS and g-Lambda using GLIF GNI-API GUSI.

Jan 2009: Formation of the GLIF NSI-WG. OSCARS is a core contributor in the writing of the OGF NSI-Architecture document; Draft architecture designs for OSCARS v0.6.

Oct 2009: Successful control plane interdomain interoperability between IDC (ESnet OSCARS, g-Lambda and Harmony using GLIF GNI-API Fenius (GUSI (GLIF Unified Service Interface))).

Nov 2009: Successful control and data plane interdomain interoperability between IDC (ESnet OSCARS, g-Lambda and Harmony using GLIF GNI-API Fenius); OSCARS used by SC09 SCinet to manage bandwidth challenges.

Feb 2010: DICE IDCP v1.1 released to support brokered notification.

Aug 2010: OSCARS is selected for NSF-funded Dynamic Network System (DYNES) project.

Nov 2010: OSCARS used by SC10 SCinet to manage demo and bandwidth challenges.

Jan 2011: OSCARS v0.6 PCE SDK released allowing developers to test OSCARS v0.6 flexible PCE framework.

Sep 2011: OSCARS interoperates with OGF NSI protocol v1 using an adapter at NSI Plugfest at GLIF in Rio.

Nov 2011: PSS for ALUs is developed and tested to support ESnet ANI 100G prototype network and SC11 SCinet.

Dec 2011: OSCARS v0.6 RC1 is released.

Jun 2012: OSCARS v0.6 Final is released.

Nov 2012: Successful demonstration of interdomain control-plane signalling using OGF NSI CS v2.0 (between OSCARS, g-Lambda, OpenDRAC, BoD, and AutoBHAN) at SC12.

if it's on Layer 1, 2 or 3."

Helping to run the world's fastest science network — ESnet just upgraded its national network to 100 Gbps — wasn't originally on Guok's career horizon. Growing up in Singapore, he was unsure if he could master Mandarin well enough to pass the compulsory exam to enter the local university. So, he came to the United States and earned his undergraduate degree in computer science from the University of the Pacific in California.

While there, he met Joe Burrescia, who was working at ESnet. After graduating, Guok returned to Singapore to fulfill his military service obligation. He subsequently returned to the United States, and was working as a teaching assistant at the University of Arizona while earning his master's in computer science when Burrescia found him via a Web search, asked "Are you the Chin who attended UOP?" and then encouraged him to apply for a network engineer job at ESnet.

"I was really thinking of going into operating systems, but Joe convinced me to give networking a try," Guok said. "My whole career at ESnet is based on a Web search, which I guess is fitting now that I'm a network engineer." HPC

Jon Bashor is a communications manager at Lawrence Berkeley National Laboratory. He may be reached at editor@ScientificComputing.com.

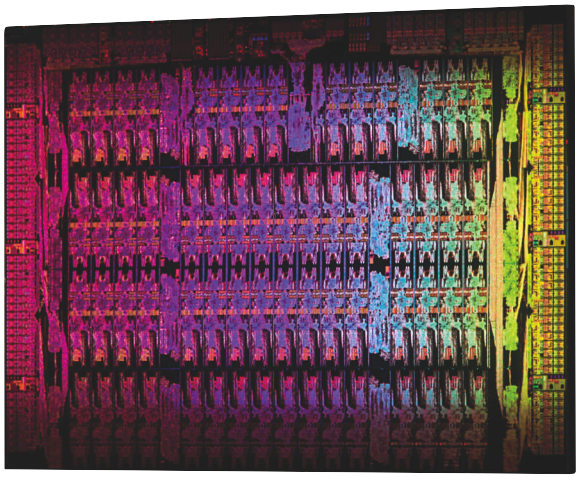
Understanding
TLP and Vector Hardware

GPUs rely on thread level parallelism, while the Intel Xeon Phi product family relies on traditional thread and vector programming models

Rob Farber

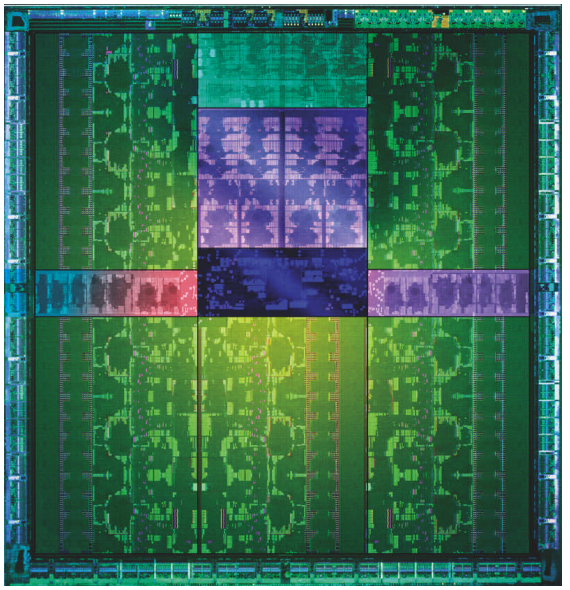
The good news for scientists, managers and engineers is that GPUs and the new Intel Xeon Phi coprocessors provide teraflop/s hardware capability in small, power-efficient and cost-effective PCIe packages. When considering these devices, it is important to understand the distinctions between GPU and Intel Xeon Phi coprocessor threads and how they are used to achieve high performance on these massively parallel devices. Applications written for both device architectures must utilize a large number of concurrent threads of execution to exploit high degrees of hardware parallelism and achieve high performance.

A key challenge for those who wish to run on Intel Xeon Phi coprocessors and/or GPUs lies in reconciling the difference between the 60 to 240



Intel Xeon Phi Coprocessor

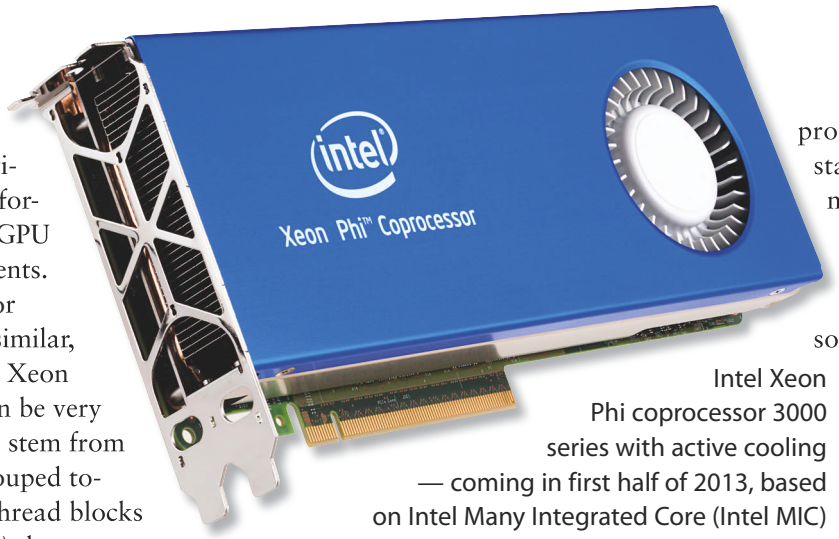
concurrent threads supported by the Intel Xeon Phi coprocessors and the thousands to millions of threads they are strongly encouraged to use when writing a GPU application. OpenCL and



NVIDIA Kepler GK110 Die

OpenACC development platforms that can transparently compile application source code to run on both GPUs and the Intel Xeon Phi product family will be available this year.

From a programming standpoint, a thread is a thread regardless of the device. Both GPU and Intel Xeon Phi coprocessor threads are regions of serial code that have been identified by the programmer or compiler as candidates for parallel execution. As discussed in my article "The Intel Xeon Phi Coprocessor for CUDA Programmers" on Dr. Dobbs, individual threads on both devices are computational universal[1] — meaning that any computable function can theoretically be implemented within a single



Intel Xeon Phi coprocessor 3000 series with active cooling — coming in first half of 2013, based on Intel Many Integrated Core (Intel MIC)

GPU or Intel Xeon Phi coprocessor thread. This does not mean that identical threads will deliver equivalent performance when running in the different GPU and Intel Xeon Phi parallel environments.

While programming a single GPU or Intel Xeon Phi coprocessor thread is similar, programming multiple GPU and Intel Xeon Phi coprocessor threads in parallel can be very different. Ultimately, these differences stem from the fact that GPU threads must be grouped together into blocks of threads (called thread blocks in CUDA or work-groups in OpenCL) that execute concurrently on the GPU streaming multiprocessors (SMs) according to a single instruction multiple data (SIMD)[2] execution model, as opposed to the Intel Xeon Phi coprocessor approach that utilizes individual threads that run concurrently on individual processor cores but must utilize a vector unit to achieve high performance. (Note that both devices also support high degrees of multiple instruction multiple data (MIMD[3]) parallelism.)

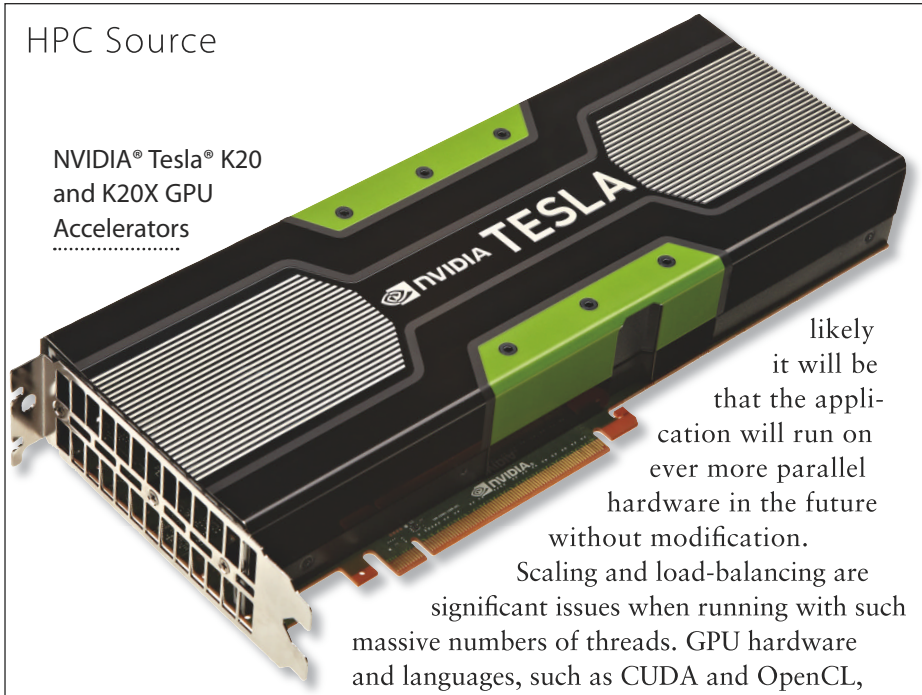
The Intel Xeon Phi coprocessor hardware designers followed a traditional approach by ramping up their design to "bet the product line" on many-core parallelism combined with a per-core wide vector unit. It is now possible to manufacture multi-core processor chips that contain a large number of simple Pentium class processing cores. Vector processors are used to augment these Pentium cores in the Intel Xeon Phi product family to provide additional operations per clock. Vector processor design is well-understood, so these new devices can pack a large amount of floating-point and integer capability into a small space on a chip in a power-efficient manner. As a result, the Intel Xeon Phi coprocessors exhibit a very high degree of floating-point and integer parallelism (essentially the number of cores multiplied by the number concurrent per core vector operations). Further, vector processing is a well-established programming model that was the de facto standard for parallel programming for many years. As a result, the Intel Xeon Phi coprocessor design ramps up performance and lowers the cost of access for users running decades worth of legacy software written for vector hardware.

It is important to stress that high performance can only be achieved on the Intel Xeon Phi coprocessors when there is sufficient vector and thread parallelism in the application code. In other words, the application code must be able to scale to 60+ cores and to effectively utilize the 512-bit vector units. For non-vector codes, the fallback is to use the Intel Xeon Phi coprocessors as support processors that run conventional x86 applications. In this way, the Intel design team is betting that customers will be able to leverage both their processor and co-processor investments with minimal software effort.

Most Intel Xeon Phi coprocessor programmers rely on OpenMP directives or Intel's Cilk Plus[4] to parallelize and vectorize their code. Optimized libraries, such as MKL, also are popular. It is also possible to directly program the vector units using compiler intrinsic operations or assembly language.

In contrast, GPU hardware manufacturers have "bet the company" on hardware designed to run many thousands of threads at one time. For this reason, GPU programmers are taught to use thousands to millions of parallel threads of execution when writing codes for GPU devices. The GPU reasoning is twofold:

1. the more threads an application provides to the device, the better the chances are that the device will be fully utilized to deliver high application performance, and
2. the more parallelism a developer uses today, means the more



NVIDIA® Tesla® K20 and K20X GPU Accelerators

likely it will be that the application will run on ever more parallel hardware in the future without modification.

Scaling and load-balancing are significant issues when running with such massive numbers of threads. GPU hardware and languages, such as CUDA and OpenCL, get around these issues by bundling groups of threads into a thread block and enforcing the restriction that only threads within a thread block can communicate with each other. The benefit of this restricted form of bundling is that all the thread blocks are free to run independently of each other, in any order, either sequentially or in parallel. This is a *really* big deal because it means the GPU hardware scheduler can arbitrarily "deal" thread blocks to all the SM on a device, much like a deck of cards. The independence of the thread blocks is fundamental to massively parallel GPU programming because it means

1. scalability is not an issue, since adding threads simply creates more thread blocks to run in parallel on the device, and
2. load balancing is not an issue, as any SM that is waiting for work can be dealt one or more of the independent thread block "cards" to keep it busy

GPU hardware designers have leveraged the characteristics of the software thread block abstraction to create power-efficient, high-performance, massively parallel devices. The common computational building block of a GPU is the SM. Each SM is a SIMD processor that performs the thread block computations in terms of a warp number of threads per SIMD instruction. Current GPUs operate on a warp size of 32, meaning a single SIMD instruction processes 32 threads at the same time. From a hardware perspective, the simplicity of the SIMD design means that each SM can pack a lot of power-efficient computational capability into a small space. Replication of the SM defines the hardware parallelism and performance of the device. In other words, the greater the number of SM on a GPU, the more instructions a GPU can run in parallel at one time.

Internally, each SM maintains a queue of the "active" thread blocks it needs to process. Those thread blocks that have all their data dependencies satisfied for their current instruction are marked as ready-to-run. Each SM can then process any ready-to-run instructions as computational resources become available, which means each SM independently has the ability to keep all its internal integer, floating-point and special function units busy. In this way, the hardware scalability of large numbers of SM replicated on a GPU is preserved.

In a nutshell, programming with a large numbers of threads means the programmer is attempting to maximize the thread level parallelism (TLP) performance bet by ensuring that the queues inside each SM will contain as many active thread blocks as possible. (The actual number, or occupancy, depends on resource utilization in the SM.) The greater the number of thread blocks queued on each SM, the better the chances are that the programmer will win their bet that at least one instruction per SM will be ready to run at any moment on each SM. The more threads used, the better the chances are that the device will be fully utilized so it can deliver high performance. Resource allocation logic inside the SM perform a best fit algorithm to preserve high performance GPU performance, regardless of the mix of integer and floating-point operations required by an application.

Succinctly:

- Intel Xeon Phi coprocessor runs individual threads that place no

restriction on how threads run or communicate. Posix pthreads[5] are a popular example of an API that supports this generic capability. It is the programmer's responsibility to ensure that threads do not deadlock, enter race conditions or limit scalability. OpenMP is a pragma-based approach that simplifies the use of these generic threads.

• CUDA programmers are required to group threads into blocks, where only threads within a thread block can communicate. This restriction frees the programmer from scaling and load-balancing concerns so they can program with thousands to millions of threads. All threads within a block execute simultaneously according to a SIMD execution model. MIMD execution is possible at the granularity of a thread block. Blocks can communicate through global memory on the GPU, but this is discouraged, as it can cause deadlock, race conditions and scalability issues. CUDA and OpenCL are the gateway development platforms for programming GPU architectures. OpenACC is a new pragma-based approach that simplifies GPU programming, just as OpenMP simplifies parallel programming on multi-core processors.

Looking to the future, we are seeing a clear convergence of programming platforms for both GPU and multi-core architectures like x86, ARM and the many-core Intel Xeon Phi product family. It is likely that developers will be writing software that can transparently be recompiled to run on GPUs, Intel Xeon Phi coprocessors, x86 processors and ARM processors.

OpenACC, a new programming standard, uses pragma-based source code annotations similar to OpenMP to direct parallel code generation. Both PGI and CAPS demonstrated OpenACC code compiling and running on Intel Xeon Phi, NVIDIA and AMD in their booths at Supercomputing 2012. Since their software translates from OpenACC to OpenCL, CAPS demonstrated OpenACC running on ARM processors as well.

Currently, OpenACC and OpenMP pragmas can be mixed in the same source code. Depending on what happens in the standards committees, language platforms such as OpenACC and OpenMP will likely converge at some point, so the same pragmas can be used to program any device. Thus, only a change of compiler switches will be required to build code that can run on a GPU, Intel Xeon Phi coprocessor and other devices.

The Intel OpenCL compiler for Intel Xeon Phi processors is currently in pre-release testing. Thus far, comments have been positive. Once released, OpenCL programmers will be able to generate binaries that can run on NVIDIA and AMD GPUs, as well as Intel Xeon Phi devices plus x86 and ARM processors. While certainly limited to some extent, there is even an OpenCL compiler for FPGA devices.

CUDA also is evolving into a multi-platforms language. For example, the PGI CUDA-x86 compiler already transparently compiles CUDA to run on x86 processors. Wu Feng's CU2CL project at the University of Virginia is an interesting technology to watch, because it provides an alternative pathway to automatically translate CUDA to OpenCL, which can be compiled to run on all the devices.

It will be interesting to see how the OpenCL standards committee responds to the CUDA[5] "dynamic parallelism" capability to more efficiently program divide-and-conquer algorithms on massively parallel devices. Version 2 of the OpenACC standard already provides for dynamic parallelism. While not mainstream, the CUDA C++ Thrust API also can be compiled with both OpenMP and Intel TBB (Thread Building Blocks) backends to run on x86 and presumably Intel Xeon Phi devices. HPC

REFERENCES

1. http://en.wikipedia.org/wiki/Turing_completeness
2. http://en.wikipedia.org/wiki/SIMD
3. http://en.wikipedia.org/wiki/MIMD
4. http://en.wikipedia.org/wiki/Cilk_Plus
5. http://en.wikipedia.org/wiki/POSIX_Threads

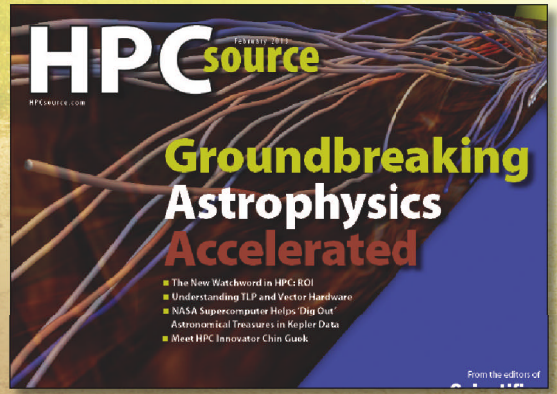
Rob Farber is an independent HPC expert to startups and fortune 100 companies, as well as government and academic organizations. He may be reached at editor@ScientificComputing.com.

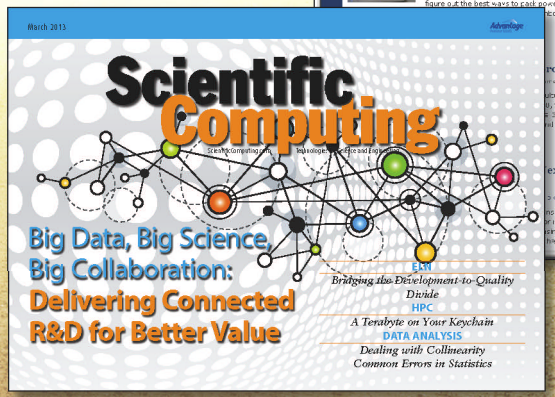
Scientific Computing HPC source

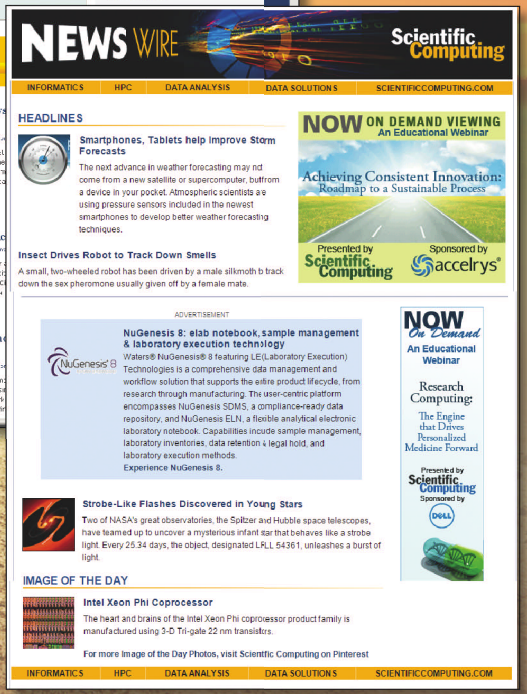
The *Scientific Computing* brand – digital magazine, *HPC Source* supplements, enewsletters, website, webcasts, and more – is designed to keep you current with today's ever-evolving technologies, delivering content in the formats you prefer to ensure your success.

Keep pace with solutions that can help optimize your workflows and expedite operations.

Visit **www.ScientificComputing.com** to subscribe.







SUBSCRIBE TODAY TO OUR NEWSLETTERS AND SUPPLEMENTS