Use of IU parallel computing resources and high performance file systems July 2013 to Dec 2014

Matthew R. Link
Robert Henschel
David Y. Hancock
Craig A. Stewart

Indiana University

PTI Technical Report PTI-TR15-006

1 February 2015

Citation:

Link, M.R., R. Henschel, D.Y. Hancock, C.A. Stewart. “Use of IU parallel computing resources and high performance file systems July 2013 to Dec 2014,” Indiana University, Bloomington, IN. PTI Technical Report PTI-TR15-006, Feb 2015. http://hdl.handle.net/2022/20334

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

RESEARCH TECHNOLOGIES
INDIANA UNIVERSITY
University Information Technology Services
Pervasive Technology Institute

1. Introduction

The Office of the Vice President for Information Technologies – and particularly the Research Technologies Division of UITS – made two very strong commitments when acquiring Big Red II, a 1 PetaFLOPS Cray Supercomputer, and the Data Capacitor II, a 5 petabyte high performance file system. Specifically, Research Technologies leadership committed to:

- Revolutionizing the way IU thinks about and uses supercomputing and advanced cyberinfrastructure
- Promoting Big Red II so effectively that its use by the IU community would span 150 or more disciplines and sub-disciplines at the university

The Big Red II dedication took place in late April 2013. It was available for use in "friendly user" mode in July, and in production mode on September 1, 2013. It is appropriate now to look back and determine UITS' success in fulfilling commitments made when requesting funds – and, more importantly, the impact of Big Red II and Data Capacitor II on IU's research and creative output.

2. Parallel versus serial workloads on Big Red II and Quarry

UITS operates four large-scale computer systems for general use by the IU community: Mason, Quarry, Karst, and Big Red II. Table 1 shows the characteristics of these systems.

Name	Architecture	Clock rate	Total nodes	TFLOPS	Total RAM (TB)
Big Red II	344 nodes Cray XE6 dual socket, 16 core (AMD x86-64) 676 nodes XK6 single socket, 16 core (AMD x86-64) and one NVIDIA K20 GPU	2.8 GHz 2.7 GHz	1020	1000.3	47
Karst	Dual socket Intel Xeon E5-2650 v2 8-core; in production as of 1/1/2014	2.6 GHz	272	93.2	11
Quarry	140 nodes IBM e1350 Intel Xeon (HS21 blades) 230 dx360 IBM iDataPlex; to be phased out 5/1/2015	2.0 GHz	370	26.11	4.9
Mason	HP DL580 G7 Intel Xeon	1.86 GHz	20	4.29	8
Totals				**1,123.9**	**70.9**

Table 1. Summary of supercomputers and supercomputer clusters operated by Research Technologies.

Mason is a large memory system, and its use is effectively (and straightforwardly) focused on computational work that requires large memory – primarily in the biological sciences. Big Red II is a supercomputer with a high speed interconnect and GPUs. When deploying Big Red II, UITS intended to focus its supercomputing power on large-scale computation and on bringing parallel approaches to sub-disciplines that have not used such approaches previously.

There is a natural tension between the goal of supporting large-scale parallel workloads and increasing the overall use of advanced cyberinfrastructure across the university community. People new to supercomputing tend to initially start at relatively lower levels of parallelism. However, UITS commonly starts people off using Big Red II as their first parallel computing resource because the attention that Big Red II has generated leads to interest in using the system. Statements such as "This supercomputer is capable of more than a quadrillion mathematical operations per cycle and Meryl Streep and Sylvia McNair both attended its dedication" and "You can use a system so interesting that an Oscar winner and a

Grammy winner attended its dedication" go a long way. Big Red II catches the attention of IU faculty and generates interest in use of advanced cyberinfrastructure in ways that other smaller and less glamorous systems do not.

Research Technologies has worked to focus the use of Big Red II on large parallel jobs through a combination of:

- **Code optimization.** The Scientific Applications and Performance Tuning (SciAPT) group has worked on 5 different applications to improve scalability of codes and 20 to improve job level optimization on Big Red II.
- **Classes on parallel programming and scalability.** Research Technologies has taught a total of 11 one-day or multi-day classes on parallel programming since the launch of Big Red II, with over 350 people attending those classes.
- **Queue policy management.** Big Red II policies are set to give high priority to parallel jobs and lowest priority to uniprocessor or single node jobs.
- **Uniprocessor workload redistribution.** The new Karst system will replace Quarry and better meet demand for uniprocessor and single node jobs. Karst will be about as fast, or faster, than Big Red II for these jobs. This will encourage people to run uniprocessor jobs on Karst rather than Big Red II, and we will further limit the extent to which Big Red II can be used for uniprocessor jobs.

Big Red II has predominantly been used for large parallel jobs, even with many researchers, scholars, and artists using the supercomputer for their first forays into parallel computing. As of the end of June, more than half of computing time on Big Red II takes advantage of 16 nodes (512 processor cores) or more. A third uses 64 or more nodes – that's a maximum of 2,048 processor cores and a minimum of 64 cores used in parallel on a single computational task. (The scaling of nodes reflects needs for processor cores and needs for memory). Figure 1 shows that the general trend in use of Big Red II is increased levels of parallelism of jobs over time. In sum, the workloads on Big Red II are using a significant fraction of the total nodes available on the system in large-scale parallel workloads.

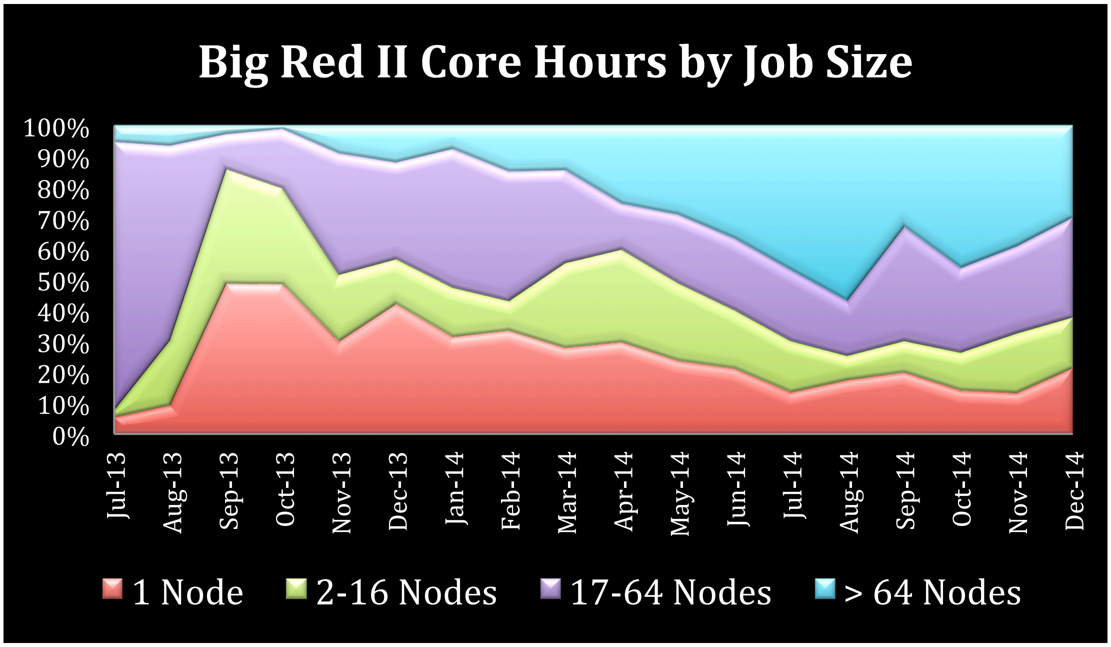


Figure 1. How much of BRII's time was used to support real parallel supercomputing applications? This graph shows BRII usage in FY14, subdivided by the number of nodes from start of production operations in September 2013 to the end of December 2014.

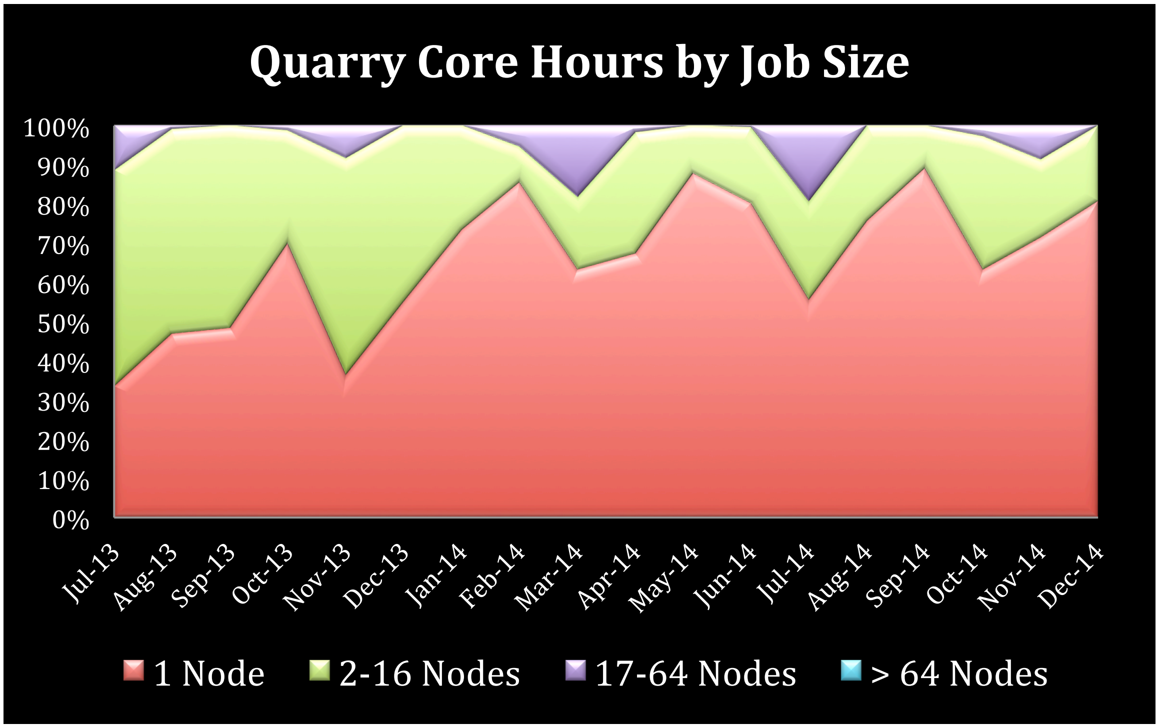


Figure 2. Workload mix on Quarry from Sept 2013 to Dec 2014.

There was a short-term increase in the number of single processor or single node jobs during the fall of 2014. Understanding Big Red II use over time is only possible through also considering systems intended to support uniprocessor workloads. In the eyes of researchers running jobs best done as uniprocessor jobs, their research is as valuable as anyone else's research. Left to their own devices, such researchers will simply look for the combination of queue wait times and processor speeds that let them finish their work as quickly as possible. The demand for uniprocessor analyses was so great in fall 2014, and so far in excess of the Quarry system capacity intended to meet those workloads, that some of this demand moved to workloads run on Big Red II in spite of queuing policies that give priority to large parallel jobs. Figures 2 and 3 show job mixes on Quarry and queue wait times for Big Red II and Quarry. (Since Big Red II went into production, overall utilization has essentially been at full capacity on the CPU-only nodes with significant GPU node utilization over time.)

Big Red II operates in two different modes: Extreme Scalability Mode (ESM), and Cluster Compatibility Mode (CCM). ESM is the "yes, we are really using this system as a supercomputer" mode. This means that the Cray Gemini Interconnect is being used to its fullest capabilities, and the code is running in the lightweight Unix-like OS (not a full kernel) that gives the best possible performance. Cluster Computing Mode uses a full Linux kernel and takes some advantage of the advanced computing capability of the Cray architecture, but this work could reasonably be done on a cluster. As Figure 3 shows, in the last year over 80% of Big Red II's monthly usage has been in the "Extreme Scalability Mode" – taking full advantage of the Cray supercomputing capabilities.

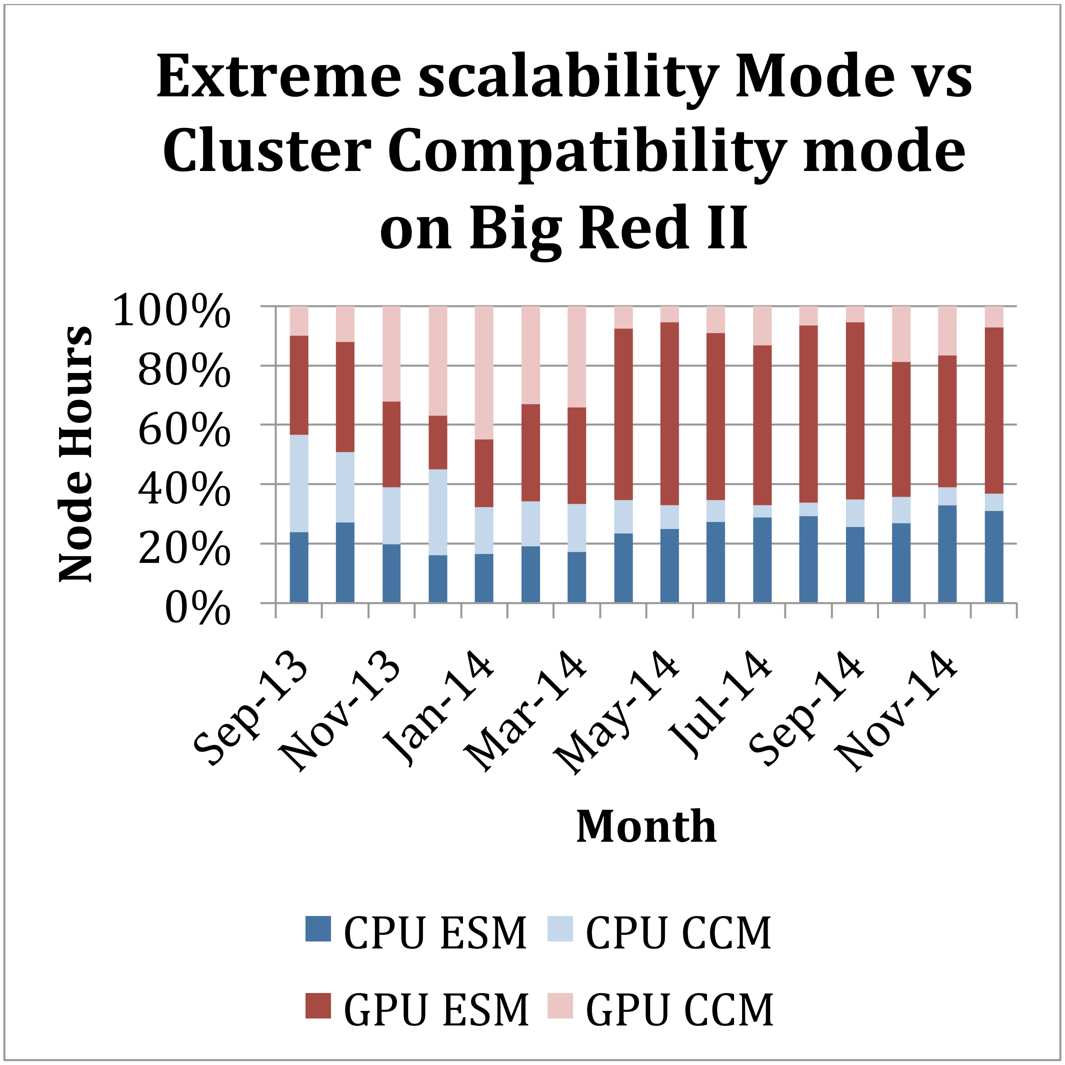


Figure 3. Percentages of node hours used on Big Red II in the CPU and GPU Extreme Scalability Mode or Cluster Compatibility mode.

3. Disciplinary diversity

We are very, very close to the goal of having 150 different IU disciplines and sub-disciplines derive benefit from Big Red II. As of the end of 2014, a total of 144 different disciplines and sub-disciplines, ranging from Ancient Studies to Vision Science, had used Big Red II. See Table 2 for details.

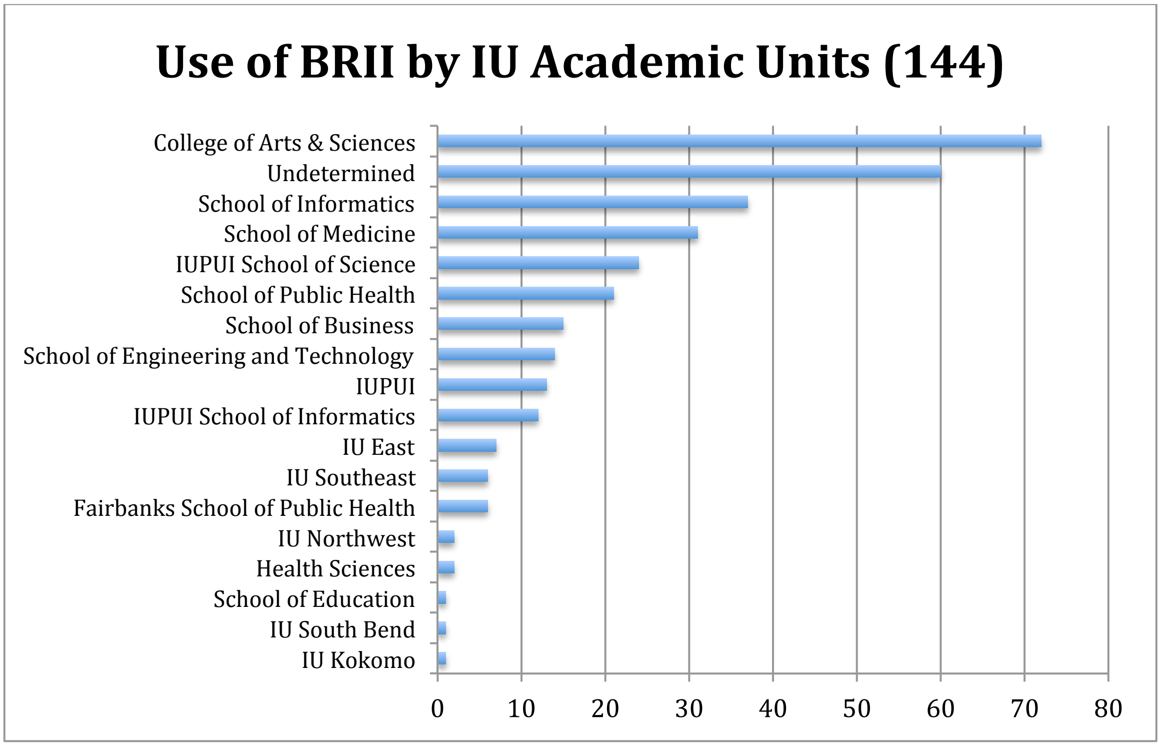


Table 2. Use of Big Red II by IU academic units.

Big Red II has also served as a springboard to the use of major national systems. For example, Steve Gottlieb used tremendous amounts of time on Big Red II's GPU nodes to optimize and tune Intel Many Integrated Core (MIC) Architecture code to run on the Nvidia GPUs. Gottlieb also used this work to successfully request time on the NSF Blue Waters system and the DOE Cray supercomputer (called Titan) at Oak Ridge National Labs. Gottlieb and the MILC consortium have an allocation of more than 30 million node hours on Blue Waters – the single largest award on this system. (The second largest allocation is 18 million node hours). Based on IU internal core hour rates, the value of this award is roughly $3M – and IU's core hour charge is less expensive than Blue Waters core hours. There is a large award in review for use of Titan as well through the DOE Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program

The most difficult aspect of managing Big Red II usage over time has proved to be use of the GPUs. They provide tremendous computational advantages, but for a very small number of codes. We had great utilization of Big Red II's GPUs when Professor Gottlieb was using Big Red II to tune his MILC applications. Of course, as a consequence of his success in getting allocations on national systems, he is no longer using Big Red II or its GPUs. We had been working with Professor Sara Pryor to port her atmospheric codes to run effectively on Big Red II's GPUs. Her recruitment (along with Rebecca Barthelmie) by Cornell University is a grievous loss to IU's intellectual capital and has at the same time taken away the person we expected to be the single biggest user of Big Red II's GPUs this year. We continue to work with a number of researchers to identify codes that will run at high speeds on CPUs, and port those codes to GPUs.

4. High-speed file systems

Research Technologies operates two high-speed storage systems – the Data Capacitor II and DC-WAN. Data Capacitor is a 5 petabyte (unformatted), high-throughput, high-bandwidth Lustre-based file system serving all IU campuses. It is directly accessible from Big Red II, Karst, Quarry, and Mason research computing systems. DCII's massive storage capacity and high-speed I/O enable Big Red II in particular to function as IU's main "big data" system. The DC-WAN (Data Capacitor - Wide Area Network) file system lets researchers access remote data as if that data were stored locally, making it easy to share large amounts of data with researchers at multiple remote sites. It's smaller than DC II at about 1 petabyte total capacity, but provides particularly valuable services to IU researchers who have lab instruments that produce large amounts of data. See Figure 5 for utilization trends.

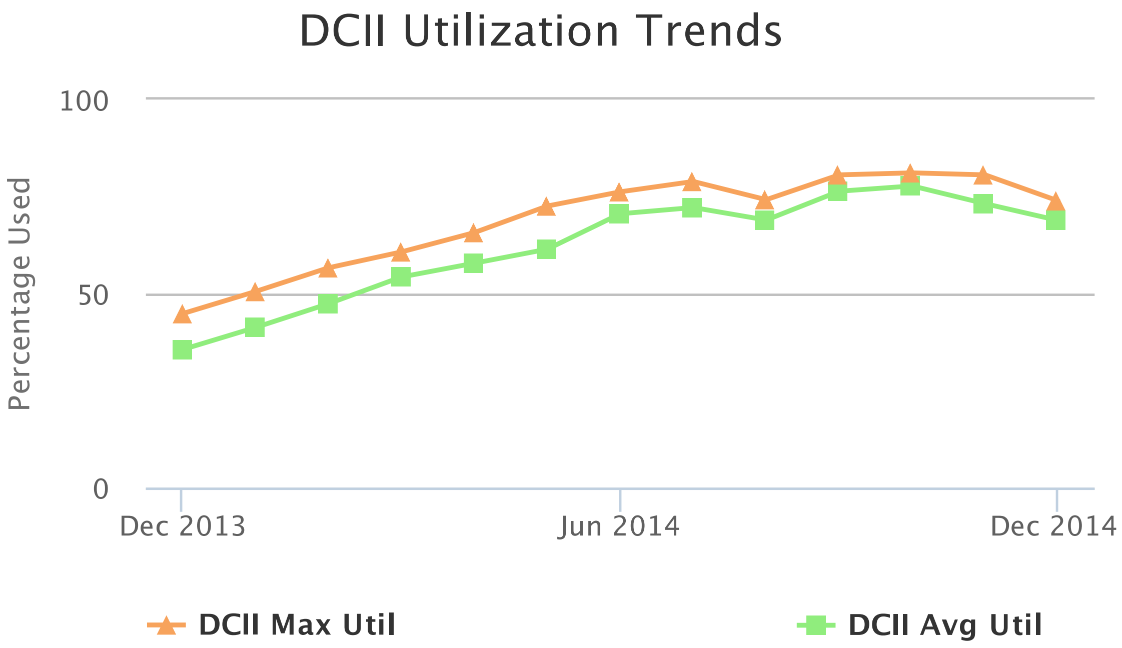


Figure 5. Utilization trends for Data Capacitor II over time.

5. Select research highlights featuring use of Big Red II or Quarry

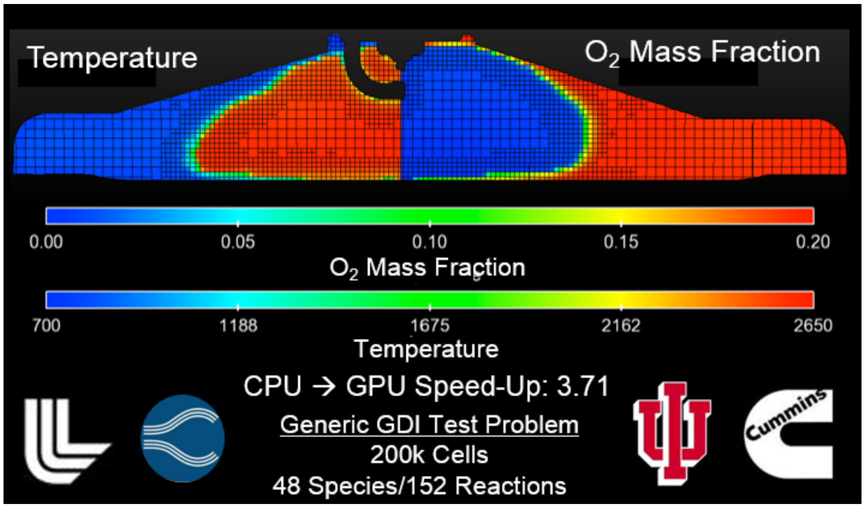
Big Red II's GPUs help Cummins improve diesel engines

Diesel engines are a major source of greenhouse gases and harmful nitrogen-bearing compounds in the atmosphere. Located in Columbus, Indiana, Cummins Inc. is a Fortune 500 corporation that designs and manufactures diesel engines widely used the world over. Cummins is deeply committed to increasing the fuel efficiency of its diesel engines, and to decreasing the extent to which its engines release pollutants into the atmosphere.

Cummins sold one million diesel engines in 2014. John Deuer, director of combustion research at Cummins, puts this in context related to their fuel efficiency and pollution reduction goals: "A 5% increase in fuel efficiency produces 14 fewer tons CO2 per year and saves $4,000 in fuel costs per year for one 18-wheeler. With a fleet of around a quarter-million trucks over five years, this means a savings of 3.5 billion gallons of fuel and 40 million tons of CO2."

Cummins design engineers use sophisticated computer models to help improve the design of the pistons and combustion chambers in its diesel engines. This is very complicated because diesel fuel combustion involves hundreds of different chemicals and thousands of reactions among those chemicals. The goal in optimizing diesel engine design is to ensure that fuels are burned as cleanly as possible. The more thorough and realistic a simulation can be, the more engine performance can be optimized.

Using a traditional computer CPU, only a simulation with about 50 chemicals and 150 reactions is practical. The speed and economy of GPU-based computing offers the potential to employ much more detailed reaction mechanisms. A partnership of Lawrence Livermore National Laboratory, Indiana University, and Cummins is adapting engine simulation software from a company called Convergent Science (the top maker of engine simulation software) to the GPUs. The figure below shows an engine simulation performed with this new software running on Big Red II's GPU processors.



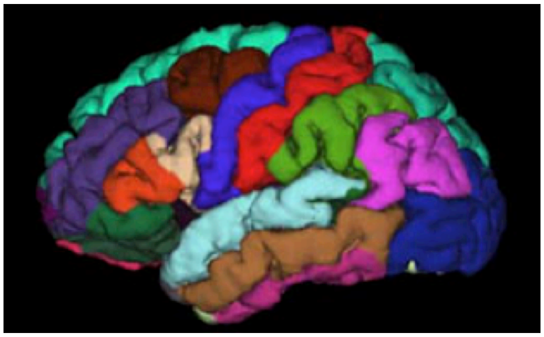
With this new software, Cummins hopes to increase the precision of its simulations through inclusion of significantly more chemicals and reactions. This will help build better, more efficient engines that will aid the US in its efforts to decrease carbon emissions, limit release of harmful nitrogen compounds into the air, and achieve energy independence.

Finding the sources of Alzheimer's disease (Big Red II)

Nearly 44 million people worldwide may be living with Alzheimer's disease or other dementias, according to the BrightFocus Foundation. Alzheimer’s robs people of their memories and causes tens of thousands of deaths a year. Millions of people watch their loved ones fade away as the disease first robs people of their short-term memory, then long-term memories, and even changes personalities.

Since 2005, the Alzheimer’s Disease Neuroimaging Initiative (ADNI) has been studying genetic causes of Alzheimer’s and environmental factors that contribute to the disease. ADNI's immediate goal is to find ways to make early diagnoses of Alzheimer’s. Its ultimate goals are to be able to treat and finally prevent this disease.

Dr. Andrew Saykin, Raymond C. Beeler Professor of Radiology and director of the IU Center for Neuroimaging, is using the advanced cyberinfrastructure of PTI and the UITS Research Technologies Division to untangle the causes of Alzheimer’s disease. Saykin’s current ADNI studies involve the entire genomes of 818 study volunteers.



Brain scan created using IU supercomputers. Courtesy Andrew Saykin

To understand Alzheimer’s, Saykin first needed to assemble the raw genetic data on each volunteer into a full and properly aligned genome. This would take approximately two weeks on a standard scientific workstation – and a total of 400 months for the entire study. That is simply impractical without a supercomputer.

"Data sets of unprecedented scope can facilitate new discoveries regarding the brain, genome, disease and therapies but computational power has become a major bottleneck to scientific progress," said Saykin. "To analyze the entire human genome in relation to longitudinal changes on brain MRI and PET scans in over 800 individuals, we need significant computing power."

Thanks to IU’s Big Red II, Saykin was able to sequence these genomes in roughly eight months – using up to one petabyte of data storage. (To put that in context, a nearly 80-foot-tall stack of single-sided DVDs would be required to store this data.) The 818 assembled genomes are allowing Saykin and colleagues to relate the genetic sequences of healthy individuals and Alzheimer’s victims to their genes. They can then use brain scans and behavioral data to track the disease's progress.

We are still a long way away from preventing or curing Alzheimer’s. But thanks to Saykin and IU’s advanced cyberinfrastructure, we are close to understanding its causes.

What do pasta and waffles have to do with neutron stars? (Big Red II)

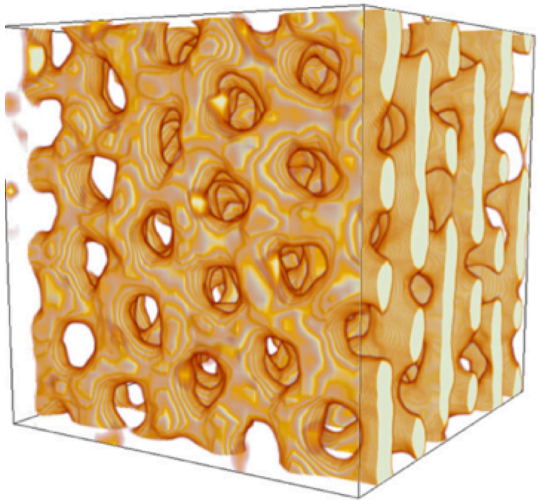
Understanding the structure of the interior of neutron stars provides clues about extraordinary physics that could help develop stronger materials like burr-inspired Velcro and sharkskin-inspired swimsuits. Nuclear pasta's design may also have an impact on the energy industry, according to IU Physics Professor Chuck Horowitz, who performed simulations on the university's Big Red II supercomputer.

A neutron star is the remnant of a massive star after it undergoes a core-collapse (or Type II) supernova at the end of its life. What remains is a compact, super-dense ball of nuclear matter with more mass than the sun in a sphere with just a 10-kilometer radius.

Below about 1,000 meters, a neutron star consists of uniform neutron matter. Above that depth, neutrons and protons bind into densely packed variants of elements found on Earth. However, physicists have long known that there is a 100-meter thick transition layer between these two domains, in the inner crust of the neutron star.

Nearly spherical nuclei begin to merge in the transition layer, their neutrons and protons combining to form long cylindrical shapes. Because of this, the layer has been dubbed "nuclear pasta." Nuclear pasta is impossible to study on earth, and cannot be observed directly. Physicists are turning to theoretical calculations and computer simulations such as molecular dynamics to understand it.

By running large-scale molecular dynamics simulations on Big Red II, Horowitz and colleagues can follow individual neutrons and protons as they interact. Such simulations shed much light on the nature of nuclear pasta – which, compressed to 10^{14} grams per cubic centimeter (100 trillion times more dense than water), forms waffle-like layers.



Simulation of the waffle-like structure of a neutron star

The recent discovery of this waffle phase by Horowitz and colleagues will help physicists better understand nuclear pasta and neutron stars. The detailed form of matter at this depth is important because it determines many of the overall properties of a neutron star.

With further study of the thermal conductivity of the star and the structure of its interior, physicists envision pursuing new material and energy applications. After all, the complex arrangement of these waffle-like layers is 10 billion times stronger than steel.

Shedding new light on snake anatomy (Quarry)

As high-performance computing resources reshape the future, scientists have greater abilities to look into the past and unlock its secrets. Analyses that would have been impossible even a decade ago now provide fascinating insights into questions about nature like: How did snakes end up without legs?

IU researcher David Polly and University of Nebraska-Lincoln's Jason Head used IU's Quarry supercomputer cluster and some very innovative approaches to paleontology to shed new light on this question. Conventional wisdom holds that snakes developed long, legless bodies by losing regions in their spinal column over time. Scientists assumed that snake ancestors had limbs even though today's snakes no longer have shoulders or shoulders blades, or hips for that matter. They explained these simplified bodies as the result of disruptions in Hox genes, which determine regions of the body – from head to tail, including structures like legs or wings – in birds, reptiles, and mammals.

Polly and Head used IU's Quarry system to do complex analyses of 56 vertebral bones from snakes, lizards, alligators, and mice. They began by photographing and digitizing the bones, then chose specific landmarks on each spinal segment. Using these coordinates, they were able to determine the size and shape of vertebra – and, after statistical analysis, where one segment ended and the next began.

What they discovered is fascinating, simple, and profound: The body regions of snakes are as complex as the body regions of other vertebrates. "Our findings turn the [assumed] sequence of evolutionary events on its head," Polly said. "It isn't that snakes have lost regions and Hox expression; it is that mammals and birds have independently gained distinct regions by augmenting the ordinary Hox expression."

A boa constrictor body shown over the shadow of a lizard body: The regions of the spine and body of the snake are as complex as the regions of the lizard – which is not what people previously believed.

Image courtesy Craig Chandler, Angie Fox, and Jason Head, University of Nebraska-Lincoln

By studying fossils from collections worldwide, Head and Polly were able to show how complexity evolved independently in each group. The computational power required was well beyond any desktop system – with 7.2 million different models making up the data for their study, nothing less than a supercomputer would do.

“Our supercomputing environments serve a broad base of users and purposes,” noted David Hancock, manager of IU’s high performance systems. “We often support research done in the hard sciences and math such as Polly’s, but we also see analytics done for business faculty, marketing and modeling for interior design projects, and lighting simulations for theater productions.”

Analyses of the scale Polly and Head needed would have been unapproachable even a decade ago. “A lot of the big jobs ran on Quarry,” says Polly. “To run one of these exhaustive models on a single snake took about three and a half days. Ten years ago we could barely have scratched the surface.”

While there is much more to understand about the evolution of the body structure of birds, reptiles, and mammals, at least we now know the right order of some of the most critical events in the process!