Report about the collaboration between UITS/Research Technologies at Indiana University and the Center for Information Services and High Performance Computing at Technische Universität Dresden, Germany

Reporting Period: July 2011 to June 2012

Robert Henschel
Craig A. Stewart
Thomas William
Wolfgang Nagel

Indiana University

PTI Technical Report PTI-TR13-005

8 July 2013

Citation:

Henschel, R., C.A. Stewart, T. William, M. Müller, and W. Nagel. "Report about the collaboration between UITS/Research Technologies at Indiana University and the Center for Information Services and High Performance Computing at Technische Universität Dresden, Germany," Indiana University, Bloomington, IN. PTI Technical Report PTI-TR13-005, Jul 2013. Available from: http://hdl.handle.net/2022/16670

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

RESEARCH TECHNOLOGIES
INDIANA UNIVERSITY
University Information Technology Services
Pervasive Technology Institute

This material is based upon work supported in part by the National Science Foundation under Grant No. 0910812 to Indiana University for "FutureGrid: An Experimental, High-Performance Grid Test-bed." Partners in the FutureGrid project include San Diego Supercomputer Center at UC San Diego, University of Chicago, University of Florida, University of Southern California, University of Tennessee at Knoxville, University of Texas at Austin, Purdue University, University of Virginia, and T-U Dresden. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

More information is available at: http://futuregrid.org/

Table of Contents

1. Summary

This report lists the activities and outcomes of the collaboration between Research Technologies (RT), a division of the University Information Technology Services (UITS) at Indiana University (IU) and the Center for Information Services and High Performance Computing (ZIH) at Technische Universität Dresden (TUD).

This collaboration was initiated by Craig Stewart in 2006 and was formalized in 2008 through the signing of a memorandum of understanding (MOU) that provided the framework for a tighter collaboration. The collaboration has already produced a number of results, including awards won at international conferences, peer-reviewed papers, awarded grants, and the exchange of researchers in order to foster information sharing between the two institutions. After the National Science Foundation awarded IU the FutureGrid grant in 2009, the collaboration entered a new phase with direct funding of a full-time equivalent at ZIH through the grant.

Three major projects were the focus during this reporting period.

First, the 100Gbit projects at TUD and IU's 100GBit SCinet Research Sandbox at the 2012 International Conference for High Performance Computing Networking, Storage, and Analysis (SC12) are a perfect opportunity for this collaboration.

Second, the application performance analysis of the IU Molecular Dynamics (IUMD) code was ultimately successful and led to several publications at the Cray User Group (CUG 2012) and before that as an electronic poster at SC11.

Third, in collaboration with the Broad Institute and the National Center for Genome Analysis Support (NCGAS), we analyzed and optimized the Trinity software; the results were published at the Extreme Science and Engineering Discovery Environment 2012 summit (XSEDE12).

The report is structured as follows. The history of the collaboration between IU and ZIH is outlined first, covering notable activities before this reporting period. In the following section, detailed information about the three major projects of this reporting period is presented. The next section briefly outlines smaller projects. The report concludes with a section about ongoing activities and an outlook into the next reporting period.

2. Success stories

This section describes projects that have been completed successfully during this reporting period.

2.1. Application performance analysis

This section outlines all projects that are related to performance analysis of scientific applications.

2.1.1. Hands-on support with application tracing

Thomas William provides hands-on support with tracing parallel scientific applications, both to members of RT and to researchers at IU. This includes the maintenance of the various Vampir tool chain installations at IU as well as user-support in tracing applications. Projects that have benefited from his involvement include the initial analysis of the GPU performance of the NBODY6 code, an n-body code used in the astrophysics community; I/O optimizations of the IU-developed hydro code; the RNA-Seq de novo Assembly tool Trinity; and mlRho, a program for estimating the population mutation and recombination rates.

2.1.2. Trinity

Trinity is a "best in class" de novo sequence assembly tool for RNA-sequencing data. Although it consistently is shown to provide superior de novo assemblies, for most computational uses it requires high performance computing infrastructure beyond that of a high-end workstation or a departmental cluster. During the stay of Matthias Lieber (ZIH) at IU in Bloomington, the Trinity code was analyzed using the Vampir tool suite. This led to substantial performance improvements to be seen in Figure 1 below. We optimized and expanded the OpenMP parallelization, tuned I/O, and applied serial optimizations to the algorithms in several of Trinity's components. We also added the capability to choose whether to use the default GNU compiler or the commercial compiler developed by Intel.

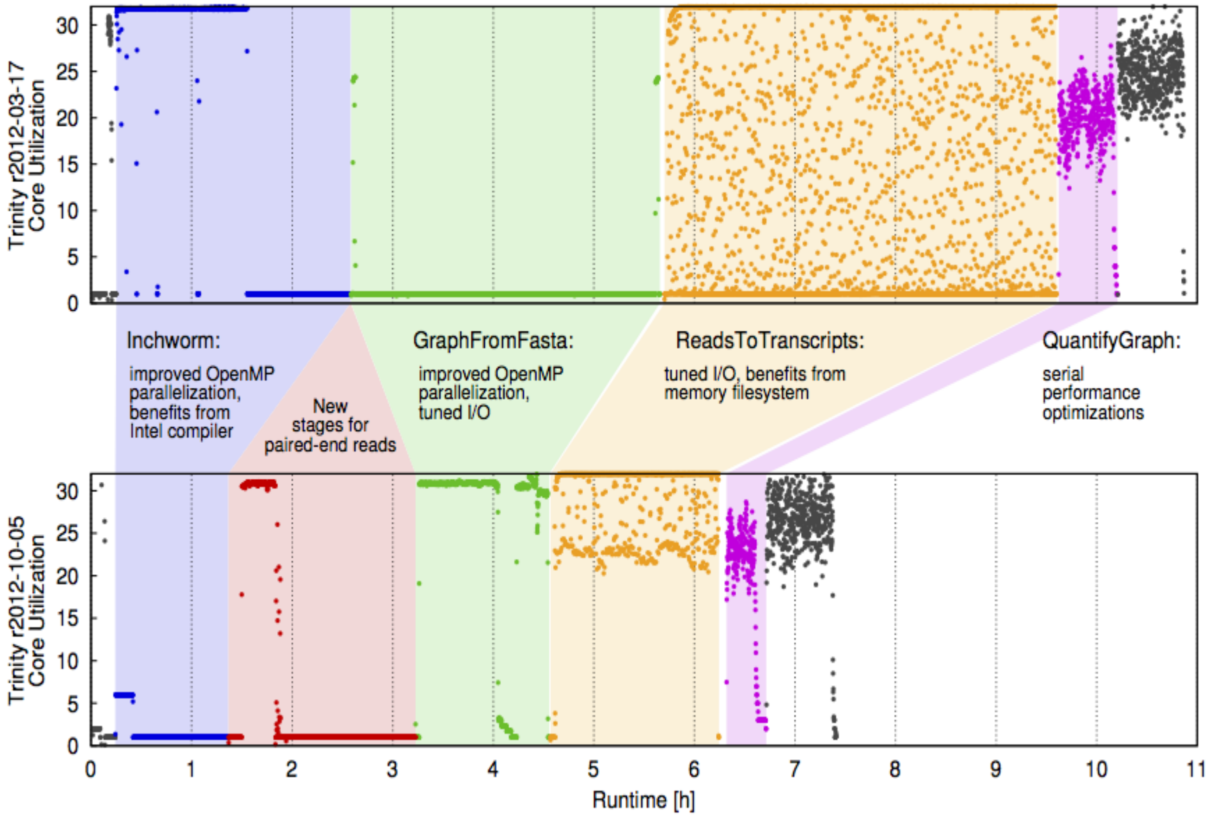


Figure 1: Trinity Performance improvements (trinityrnaseq_r2012-10-05)

As Trinity is in ongoing development, new stages are added constantly (e.g. red part in Figure 1), which then in turn need new analysis to optimize them for execution on HPC systems. A paper presented at the XSEDE12 conference by Robert Henschel (http://dx.doi.org/10.1145/2335755.2335842) shows a decrease in runtime by a factor of 3.9 due to the before mentioned optimizations. This is an ongoing project involving the RT Scientific Applications and Performance Tuning (SciAPT) group, NCGAS, and the Broad Institute.

2.1.3. IUMD

The IU Molecular Dynamics software was written to simulate certain physical properties of neutron stars and white dwarfs. C.J. Horowitz's group develops IUMD in-house at IU. It consists of several versions of the same semantics implemented in various ways; for example, making use of new language constructs available in newer Fortran versions. IUMD can be compiled as a serial, OpenMP, MPI or MPI+OpenMP program. The variants have different names, (md, md_omp, md_mpi, md_mpi_omp) and are selected during compile time. Together with a wide range of compile time flags that influence the arithmetic of the simulation, this creates several hundred possible application-runs to look at. All the measurements are done on XRay, a Cray XT5m provided by the FutureGrid project (NSF grant 0910812).

The code is altered to allow for a Performance Application Programming Interface (PAPI) counter to be set during compile time. Inside the newton subroutine, the PAPI counter is measured for each call to the acceleration subroutine. The acceleration subroutine consists of a case statement that switches between the nucleon, pure-ion, and ion-mixture simulation type. By wrapping the subroutine we get a PAPI-value at each simulation time step t + δt. The values are written to a file and summed up at the end of the simulation run. Creating different binaries for the different PAPI counters allows us to run all the binaries with the same input dataset in one batch-job. The Vampir suite is then used to analyze all these versions of the code.

Running the MPI-only version had only a 3% overhead. So the focus is on the OpenMP pragmas. The trace shows that most of the time spent inside OpenMP is actually spent in thread management instead of computation. Although the reality is distorted by the fact that VampirTrace also uses these functions to create the necessary trace data to log the threads, which further adds to the overhead, this behavior indicates that the workload per thread is too small and performance is impeded. These findings were reported back to the developer leading to a change in the algorithm. The new version exhibits much better behavior regarding the parallel efficiency.

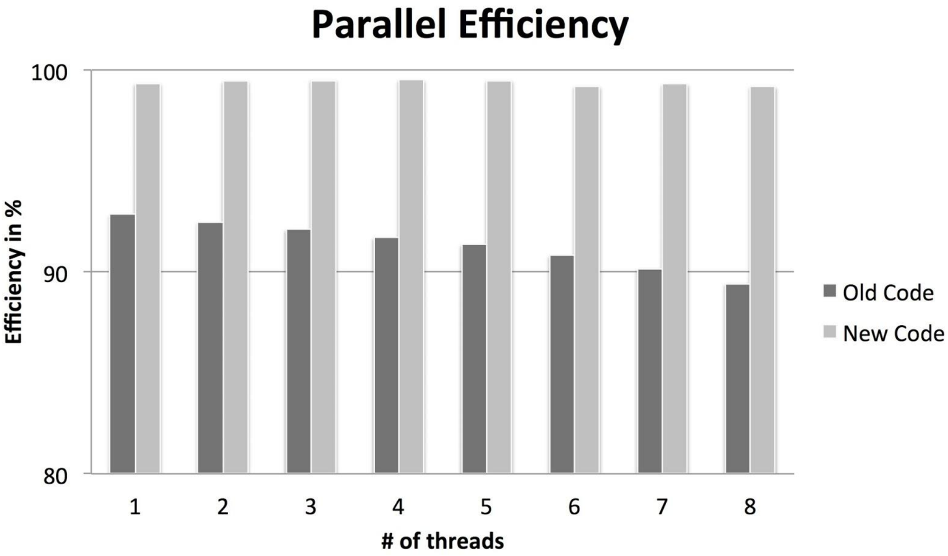


Figure 2: IUMD Performance Comparison (Version 6.1)

Figure 2 above shows that even the single thread performance benefits from the change to three loops. The serial version of the old code needed 14504 seconds for a single core, 55k particles run. Due to the OpenMP overhead this increased to 15622 seconds running one OpenMP thread. With the new version the OpenMP overhead is lower and this now takes 14602 seconds to complete.

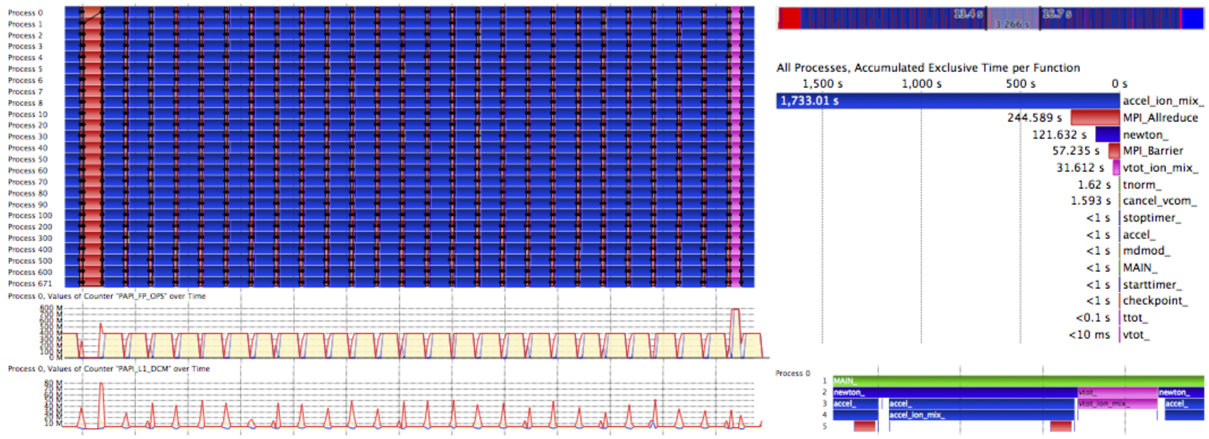


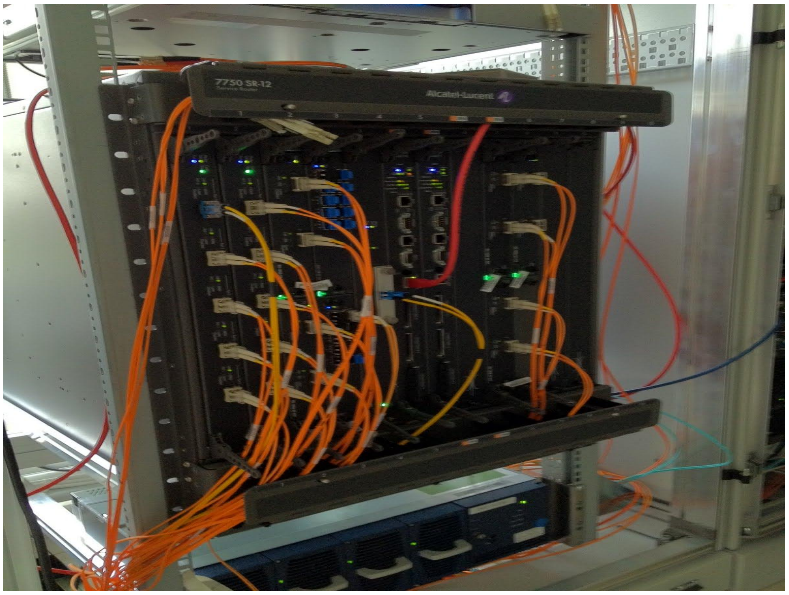
Figure 18. Vampir showing the MPI-only version running on all 672 cores.

Figure 3: Vampir Screenshot showing a full machine run on XRay (IU owned Cray XT5m)

While the old version drops below 90% efficiency with eight threads, the new version starts at 99.3% with a single thread, peaks at three threads with 99.5% and achieves 99.2% with eight threads. This is the maximum number of threads on the Cray XT5m. Newer hardware architectures with more cores, such as the AMD bulldozer, will benefit more from this. To maximize the performance on the given hardware (XT5m), OpenMP and MPI are combined. We start with eight cores on one node using four OpenMP threads and two MPI processes to resemble the dual socket quad core hardware. This yields 98.2%, which is 1% less than the pure OpenMP version. All subsequent measurements were therefore taken using MPI inter-node and OpenMP intra-node. Using the full machine XRay and the optimized OpenMP code, an efficiency of 97.2% can be achieved, compared to 79.2% for the MPI-only version.

First results of the analysis and optimization were published as an ePoster at the SC11 conference. A more in-depth explanation is available as a paper for CUG 2012 and can be accessed through their digital library at https://cug.org/publications.

2.2. 100Gbps DRS/FG



The 100Gbps test project finished with a workshop presenting the results, held in Mannheim on the 28th of September 2011 in Germany. (For more details, see last year's report at http://hdl.handle.net/2022/14295.) IU's Stephen Simms participated in this workshop. A journal paper about the 100Gbps work in Dresden, titled "Performance and quality of service of data and video movement over a 100 Gbps testbed," written jointly by IU and TUD authors, was submitted and published in Future Generation Computer Systems 29 (2013), pages 230–240. We are currently in the planning phase for switching from a testbed mode to pilot operation.

2.3. 100Gbps IND/SEA

Building on the experiences in Dresden, IU decided to have a 100Gbps showcase at the SC11 conference in Seattle. To prepare for this event, Thomas William joined IU's team seven weeks prior to the conference.

Work started on setting up both clusters for the two endpoints of the 100Gbps lane (one on the SC11 show floor and the other in Bloomington). A set of demonstration applications was agreed upon and software installation begun. Again, the IUMD code was chosen as one of the applications due to the deep understanding of the code, gained in the previously mentioned projects. Besides background traffic, which was simulated with "iperf" TCP/UDP streams, the following applications were used to simulate traffic on the 100Gbps lane:

Application/Workflow	Domain	Number of Nodes
Heat3D	Heat Diffusion	8
VampirTrace	Application Performance Analysis	7
Enzo	Astronomy	6
NCGAS	Genomics	3
OLAM	Weather	2
CMES	Computational Neuroscience	2
Gromacs	Molecular Dynamics	1
ODI-PPA	Astronomy	1

Table 1: Applications used in 100Gbps test

Parallel to this, the High Performance File System group (HPFS) installed, configured and tuned a Lustre file system to be used on the clusters using the knowledge gained in the Dresden testbed.

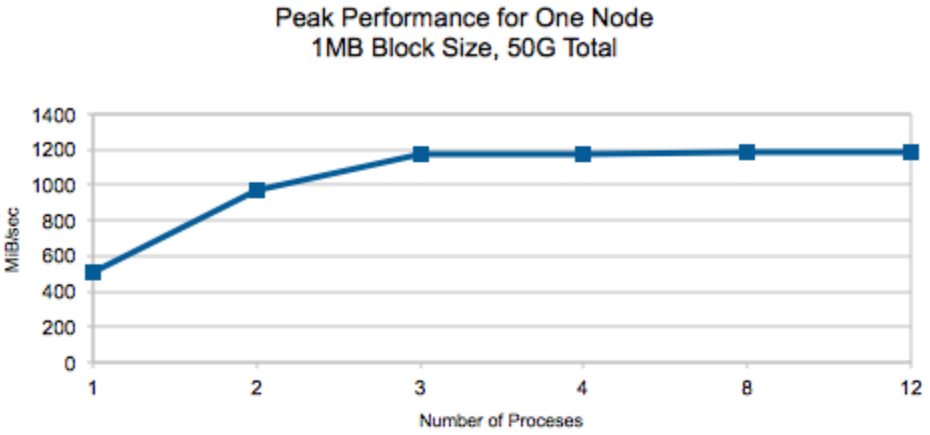


Figure 4: Single Node I/O Performance (IOR run)

After a successful set of demonstrations on the show floor of SC11, several papers were published describing various aspects of the work done:

Michael, Scott; Zhen, Liang; Henschel, Robert; Simms, Stephen; Barton, Eric; Link, Matthew. A study of Lustre networking over a 100 gigabit wide area network with 50 milliseconds of latency.

Knepper, Richard; Michael, Scott; Johnson, William; Henschel, Robert; Link, Matthew. The Lustre File System and 100 Gigabit Wide Area Networking: An Example Case from SC11.

Henschel, Robert; Simms, Stephen; Hancock, David; Michael, Scott; Johnson, William; Heald, Nathan; William, Thomas; Berry, Donald; Allen, Matt; Knepper, Richard; Davy, Matthew; Link, Matthew. Demonstrating Lustre over a 100Gbps wide area network of 3,500km.

3. Short-term projects

This section lists short-term activities that have been performed in this reporting period.

3.1. Vampir on the display wall in the CIB

The entrance hall of IU Bloomington's Cyberinfrastructure Building (CIB) features a wall of 6x4 full HD monitors interconnected to form one giant screen. Vampir was installed on this system as the multi display interface of the Vampir GUI can make efficient use of these multiple screens while analyzing huge datasets.

3.2. Module logging (IU and FutureGrid)

A large fraction of today's high performance computing (HPC) systems uses software modules. The Environment Modules package provides for the dynamic modification of a user's environment via module files. To track software usage in FutureGrid, we augmented the modules' sources to provide logs about used modules per HPC system. A log entry looks like this:

```
Apr 23 06:17:46 gw41 module_cmd: william load mpfr/2.4.2
```

This denotes that user `william` loaded the module `mpfr/2.4.2` on the system named `gw41`. The same functionality was ported to the IU monitoring website (stats tracking).

3.3. Intel Sandybridge and AMD Interlagos test systems

In preparation for the next procurement at IU, two loaner systems using AMD and Intel CPUs were benchmarked and evaluated.

Betty is a single node Intel system containing two Intel Xeon 8-core E5-2680 2.70 (base frequency) GHz processors and 64 GB RAM, for a total of 16 processing cores totaling 345.6 gigaflops per node at the base frequency. Wilma is a HP ProLiant DL585 G7 6282 SE system containing four AMD Opteron™ Model 6282 SE 2.6 (base frequency) GHz processors and 128 GB RAM, for a total of 64 processing cores totaling 665.6 gigaflops per node at the base frequency. Both run RedHat Enterprise Linux Server release 6.2 (Santiago).

The following benchmarks were selected to run on both systems.

- HPC Challenge Benchmark from University of Tennessee Knoxville
- High-Performance Linpack Benchmark from University of Tennessee Knoxville
- SPEC OpenMP2001 Benchmark Suite Version 3.2 (SPEC OMP2001-3.2)
- SPEC MPI2007 Benchmark Suite Version 2.0 (SPEC MPI2007-2.0)
- Indiana University Molecular Dynamics (IUMD) code

As IUMD was already examined (see success stories section), it made sense to use the gained experience in the benchmarks. Running IUMD with a reasonable set of parameters, we generated numbers for different compiler flags on both systems:

Time		Time	
Betty	838.64	Wilma	859.74
-fast	780.62	-fast	460.02
-mavx	821.04	-tp bulldozer-64	828.49
-xAVX	822.24	–Mvect=simd:128	821.21

Table 2: Comparison between Intel and AMD CPUs

For Wilma, not using the -fast flag, the runtime of the 64-core AMD is similar to running on the 16-core Intel chip of Betty (~830 seconds). Using 32 cores, Wilma finishes in 720 seconds, comparable to using the "-fast" flag on Betty. Using all 64 cores and the "-fast" flag on the Interlagos CPU, the code finishes in 460 seconds, a parallel efficiency of 75%.

For Betty, the table shows that explicitly using the AVX Flags yield a small performance gain of 2% compared to using the compiler without any optimizer flags. Fast (-xHOST -O3 -ipo -no-prec-div -static) achieves 7%.

The Betty system performs as expected, especially in terms of the HPL performance. The compiler switch "-xAVX" helps to improve performance slightly for the cases like SPECMPI, but this really depends on the application. A typical improvement can be in the range of 5%. Wilma's performance was relatively below our expectation. For the HPL performance, however, it looks reasonable. From our experience, the AMD CPUs usually produce HPL performance in range of 65 to 70 percent of peak.

3.4. Dedicated circuit test between IU and Dresden (speed.hps.iu.edu)

As part of high performance computing collaboration, performance analysis of parallel scientific applications involving the capturing of full details of the runtime behavior and storage of this information in trace files is done. Trace files can be anywhere from a few megabytes to several gigabytes in size. There is the need to transfer the trace files from one institution to the other. The need for the file transfer depends on the nature of the performance problem and cannot be predicted in advance. This is also true for the amount of files and the frequency with which they need to be transferred. In 2006, IU deployed the Data Capacitor, a site-wide Lustre file system spanning Bloomington and Indianapolis (50 miles). Experiments with mounting the file system across much longer distances nationally (San Diego Supercomputer Center) and abroad (Technische Universität Dresden, Germany) were successful in mounting the Data Capacitor at many TeraGrid resource providers (dedicated network). However, there were inconsistencies in the network path from Dresden to IU, asymmetric routes, seemingly arbitrary packet breakups, and bandwidth limitations. Those issues could be solved eventually, but required constant attention, as the tuning needed to be redone every few weeks. So the current state is again that secure copy (ssh) is used to transfer files between both institutions, where the latency of 130 msec yields poor results and underpins the need for a dedicated circuit connection. We therefore evaluated the impact of a dedicated circuit connection on file transfers between the two institutions. In the long run we hope to use on-demand dedicated circuits to speed up file transfers.

Two endpoints were set up in Dresden and in Bloomington. Later Tucson was added. Several tests were implemented (RFS/AFS, Lustre, ssh, ping …) and run hourly over a period of more than two months. This showed a lot of unpredicted effects like changing routes that cause the latency to change:

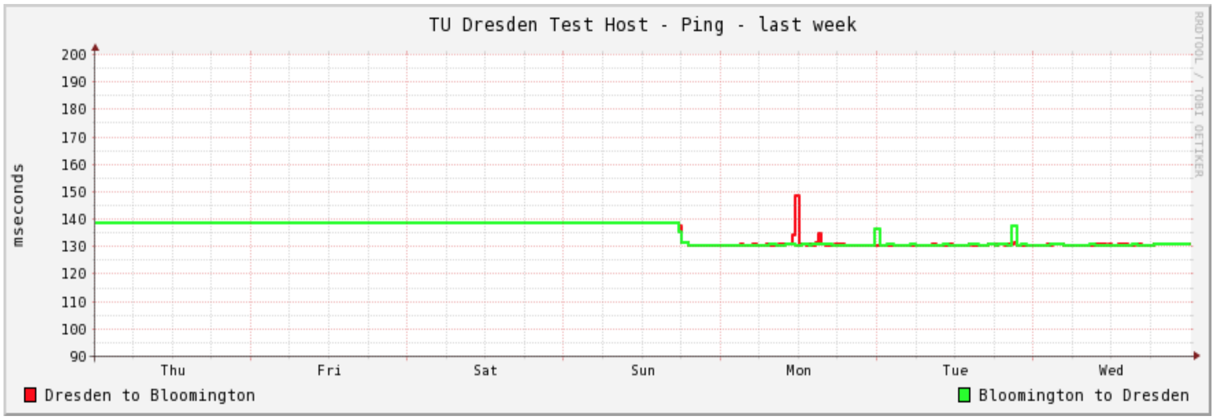


Figure 5: Latency change due to changes in the route

It also showed that the available bandwidth changes depending on the date and time and that there is more traffic coming from the US than going to the US, which impacts the available bandwidth during our tests:

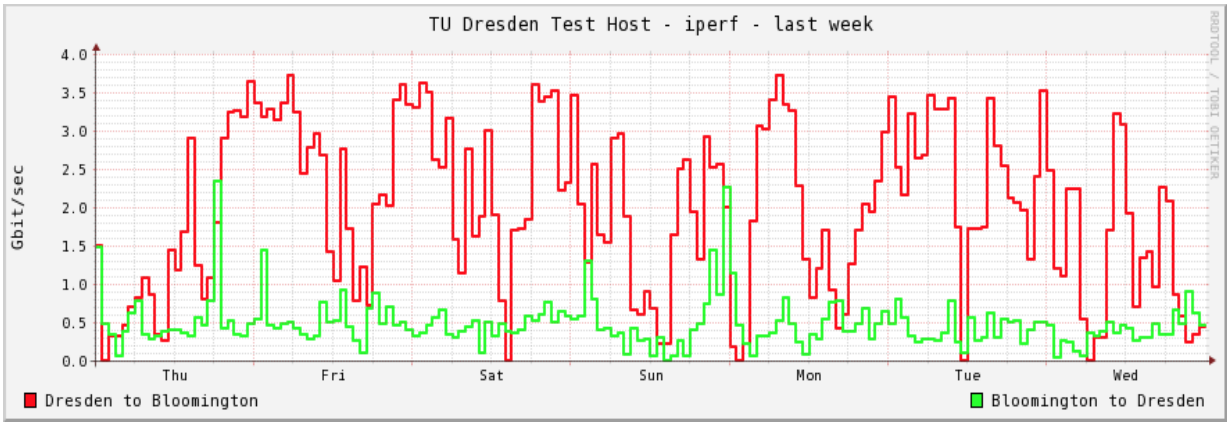


Figure 6: Bandwidth achieved using "iperf" between IU and ZIH

Comparing "iperf" above with "scp" below shows why larger file transfers are not feasible with secure copy:

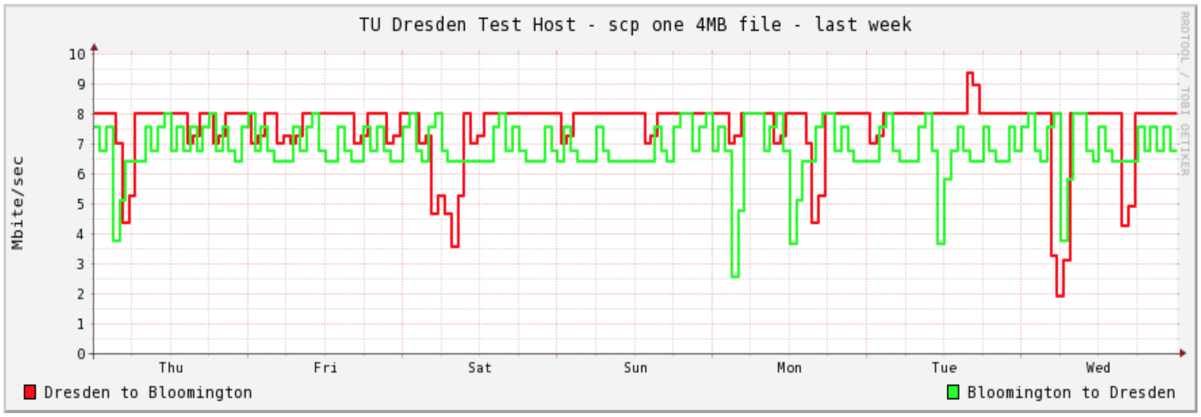


Figure 7: Bandwidth achieved using "scp" between IU and ZIH

It is the goal of the envisioned Dedicated Circuit Project with Jim Williams to find a solution to these problems.

3.5. *Xen and KVM comparison*

As part of our tasks in FutureGrid we benchmarked different virtualization software packages. We compared the host ("bare metal") with Xen and KVM concerning CPU performance, bandwidth and file I/O.

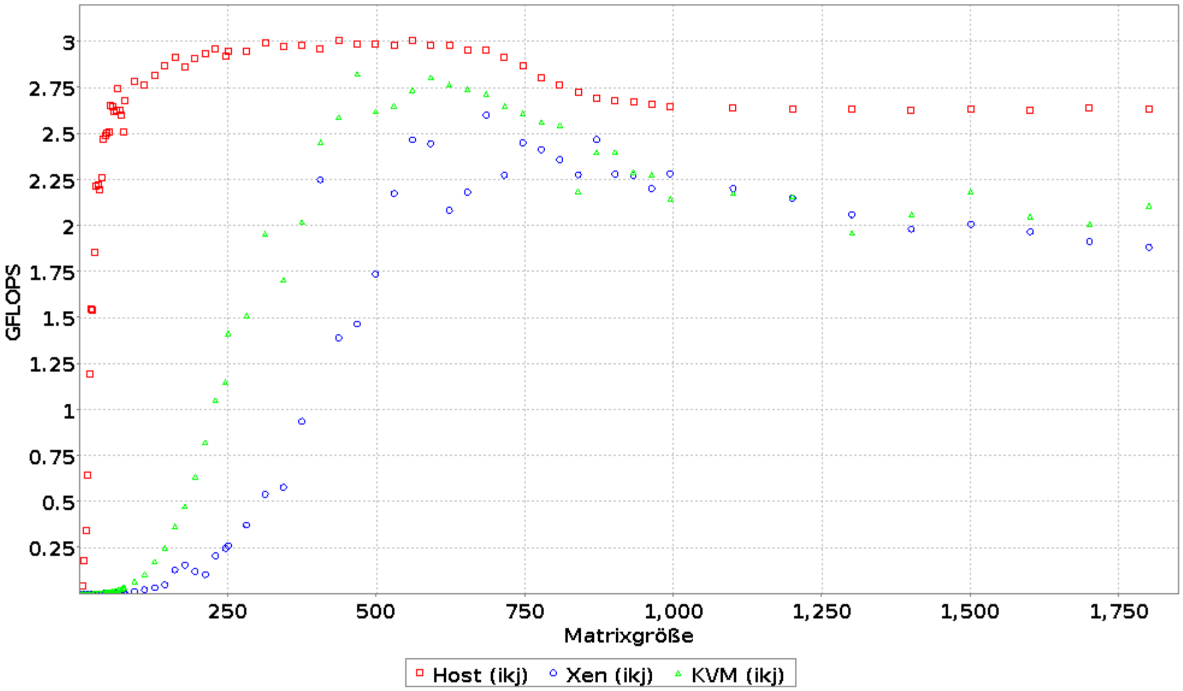


Figure 8: Matrix Multiplication benchmark

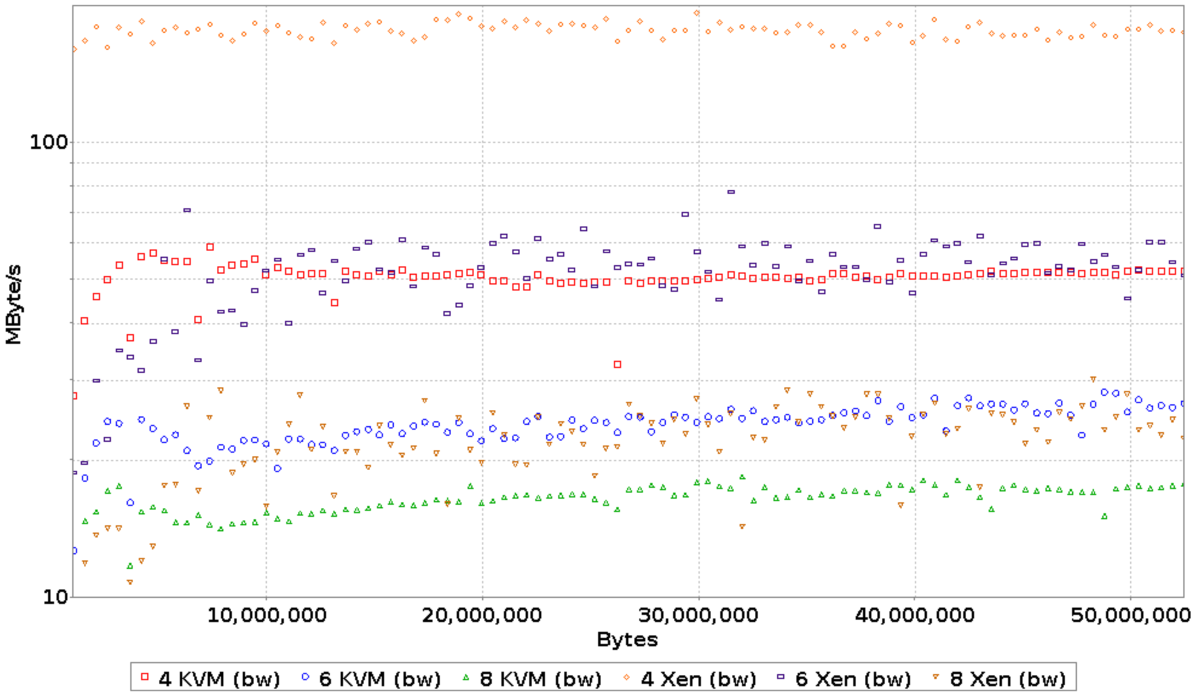


Figure 9: Communication bandwidth

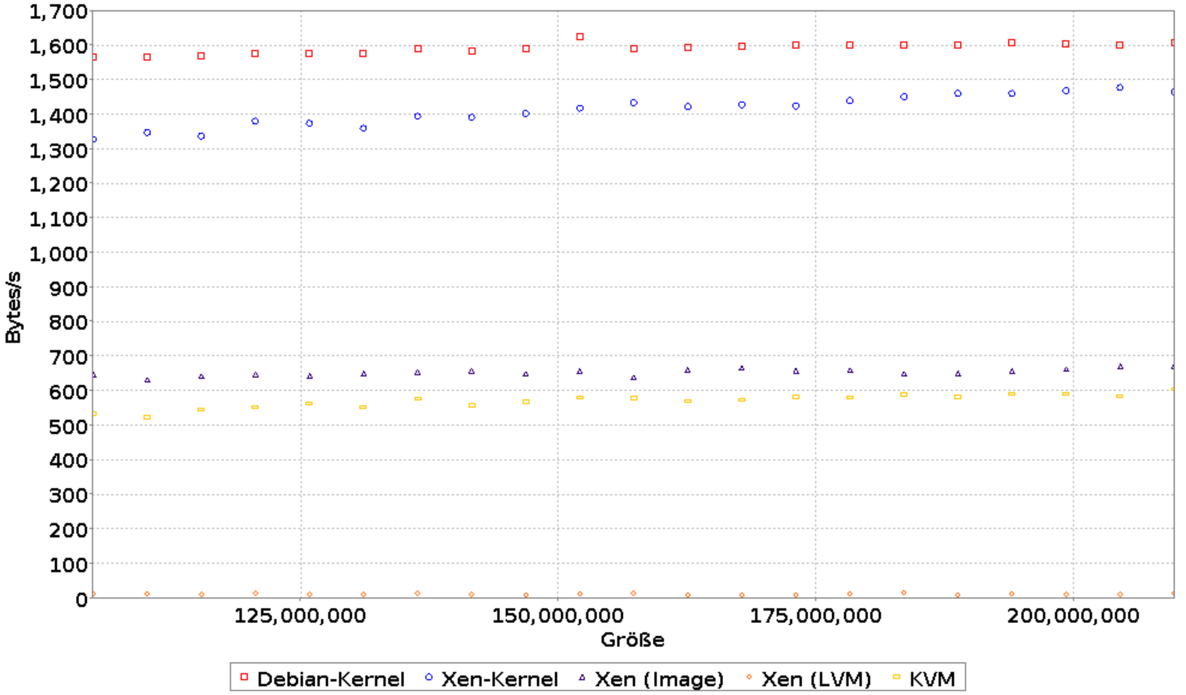


Figure 10: File I/O doing a simple file copy

The CPU benchmark is a naive implementation of the matrix multiplication. Serial jobs showed no significant differences, but using an OpenMP-parallelized version of the benchmark, KVM outperformed Xen. Xen achieved much higher bandwidth ratios compared to KVM but both suffered severely from oversubscription. Also raw I/O performance of KVM was lower compared to Xen or the native Debian Kernel running on the host.

4. Ongoing activities

To foster active participation in this collaboration we have created various recurring activities. The purpose of those meetings is sharing information on new opportunities for collaboration as well as reporting on the status of existing projects.

Thomas William from ZIH participates in the weekly meetings of RT's SciAPT group. This ensures that he is kept in the loop about activities at IU, as well as allowing everyone in SciAPT to engage Thomas in application performance analysis related activities. This has sparked a number of smaller projects and has led to a wider adoption of the Vampir toolchain, which IU has licensed from ZIH for eight years. This has also led to IU becoming more active in providing feedback about the Vampir tool to ZIH, allowing us to positively influence the development roadmap.

In addition to the weekly group meetings, Thomas and members of SciAPT get together as needed to talk about specific projects. The widespread availability of video conferencing technology at IU and ZIH dramatically improves the quality and outcomes of such meetings, allowing us to not only see and hear each other but at the same time also share computer screens for analyzing data.

We have organized a monthly meeting that is dedicated to this collaboration. Here we discuss longer-term goals and activities. Invitations to participate are distributed widely on both sides.

Personal interactions are an important part of this collaboration. Representatives from both institutions get a chance to see each other in person at least twice a year, at the International Supercomputing Conference in Germany in summer and at the Supercomputing Conference in the United States in the fall. However, due to the large number of activities at these conferences, those interactions are usually very short. To further build personal ties, it is our goal to have at least one visit per year to the other institution. During this reporting period, members from IU have visited ZIH to work on the Trinity project (July/August 2012) and on general projects related to this collaboration. Also, a member of ZIH was in Bloomington in April 2012 to kick-start the collaborative work on Trinity.

4.1. Vampir updates on Quarry, Mason, Big Red, and XRay

All relevant packages are kept up to date. The most important are:

- gcc-4.2.2
- openmpi-1.4.1-gcc-4.2.2-64
- vampirserver-2-openmpi-64
- vampirtrace-5.11.1-gcc-openmpi-64
- vtlibwrap generator (+how-to for quarry)
- PAPI, Vampir, and VampirTrace on mason

4.2. SPEC work of IU and ZIH

SPEC benchmarking is an ongoing effort. We plan to publish SPEC OpenMP2012 in Oct. 2012. It will be discussed in next year's report; however, work towards this was done in this reporting period.

4.3. Research visits

As part of the MOU, there are research visits each year to boost the projects listed in this report.

Thomas visited in October 2011 to help with the SC11 preparations, especially with the 100Gbps demonstrations. This work included cluster installation (PAPI, VampirTrace, lm-sensors), SFA10K optimizations, Lustre at large (tools, bandwidth and so on), Lustre reformat optimizations for 100Gbps and IOR + VampirTrace measurements for Stephen Simms' Group. The work was presented in the SCinet Research Sandbox Test.

Matthias Müller visited Bloomington in April 2012 to work together with SciAPT and NCGAS on kick-starting the Trinity collaboration.

Robert Henschel visited Dresden in July/August 2012 (outside the report period).