From job submission support to advanced performance tuning of parallel applications.

A case study from a university with an open access policy to high performance computing.

Robert Henschel
Director, Science Community Tools
Research Technologies, UITS
Indiana University
June 22nd, 2017
Contents

• Indiana University and HPC@IU
• What is Performance Tuning
• Examples of Performance Engineering
• SPEC High Performance Group
Indiana University
IU – Campuses and Medical School Centers

IU Campuses

IU School of Medicine campuses and clinics
IU Overview

Overall
- Operating budget - $3.5B
- Grant Awards of $614M in 2016

Centralized IT Org - UITS
- 700+ professional staff
  - 130 Research Technologies
- 500+ part time staff

<table>
<thead>
<tr>
<th>Fall 2016</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduate</td>
<td>93,740</td>
</tr>
<tr>
<td>Graduate</td>
<td>12,397</td>
</tr>
<tr>
<td>Doctoral - Research</td>
<td>4,323</td>
</tr>
<tr>
<td>Doctoral - Practice</td>
<td>3,700</td>
</tr>
<tr>
<td>Total Students</td>
<td>114,160</td>
</tr>
<tr>
<td>Staff</td>
<td>11,498</td>
</tr>
<tr>
<td>Faculty</td>
<td>9,005</td>
</tr>
<tr>
<td>Grand Total</td>
<td>134,633</td>
</tr>
</tbody>
</table>
# Research Technologies

**Associate Dean, RT, and Executive Director, PTI Craig A. Stewart**

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systems</td>
<td>Matt Link</td>
</tr>
<tr>
<td>Visualization and Analytics</td>
<td>Eric Wernert</td>
</tr>
<tr>
<td>Science Community Tools</td>
<td>Robert Henschel</td>
</tr>
<tr>
<td>Community Engagement and Interoperability</td>
<td>Therese Miller</td>
</tr>
<tr>
<td>Advanced Cyberinfrastructure</td>
<td>Dave Hancock</td>
</tr>
<tr>
<td>High Performance File Systems</td>
<td>Stephen Simms</td>
</tr>
<tr>
<td>Research Analytics</td>
<td>Scott Michael</td>
</tr>
<tr>
<td>Scientific Applications and Performance Tuning</td>
<td>Abhinav Thota</td>
</tr>
<tr>
<td>Campus Bridging and Research Infrastructure</td>
<td>Joe Buttler</td>
</tr>
<tr>
<td>High Performance Systems</td>
<td>Peg Lindenlaub</td>
</tr>
<tr>
<td>Research Storage</td>
<td>Charles McClary</td>
</tr>
<tr>
<td>Advanced Visualization Lab</td>
<td>Michael Boyles</td>
</tr>
<tr>
<td>National Center for Genome Analysis Support</td>
<td>Tom Doak</td>
</tr>
<tr>
<td>Grant Support and Outreach</td>
<td>Winona Snapp-Childs</td>
</tr>
<tr>
<td>Advanced Parallel Applications</td>
<td>Ray Sheppard</td>
</tr>
<tr>
<td>Education, Outreach, Training</td>
<td>Robert Ping</td>
</tr>
<tr>
<td>Jetstream Cyberinfrastructure</td>
<td>Georg Turner</td>
</tr>
<tr>
<td>Application Desktop Virtualization</td>
<td>Stephanie Cox</td>
</tr>
<tr>
<td>Research Data Services</td>
<td>Esen Tuna</td>
</tr>
<tr>
<td>Advanced Parallel Applications</td>
<td></td>
</tr>
<tr>
<td>Scalable Compute Archive</td>
<td>Arvind Gopu</td>
</tr>
<tr>
<td>Jetstream Project Management and Outreach</td>
<td></td>
</tr>
<tr>
<td>Advanced Biomedical IT Core</td>
<td>Richard Meraz</td>
</tr>
<tr>
<td>Digital Humanities Cyberinfrastructure</td>
<td>Tassie Gniady</td>
</tr>
<tr>
<td>High Throughput Computing</td>
<td>Robert Quick</td>
</tr>
<tr>
<td>National Center for Genomic Analysis Support</td>
<td>Tom Doak</td>
</tr>
<tr>
<td>Digital Humanities Cyberinfrastructure</td>
<td>Tassie Gniady</td>
</tr>
<tr>
<td>Scalable Compute Archive</td>
<td>Arvind Gopu</td>
</tr>
<tr>
<td>Jetstream Project Management and Outreach</td>
<td></td>
</tr>
<tr>
<td>Advanced Biomedical IT Core</td>
<td>Richard Meraz</td>
</tr>
</tbody>
</table>
**HPC @ IU - Compute**

- **Big Red II** – Cray XE6/XK7
  - 1020 nodes, 1 PFLOPS
  - CPUs/GPUs
  - CLE 5 up 02
  - Torque/Moab
  - 22 LNET Routers (QDR)
  - 4 DVS nodes (10Gb)

- Available to all Faculty, Staff, and Graduate Students
- Support/consulting available
**HPC @ IU - Compute**

- **Big Red II+** – Cray XC30
  - 560 nodes, 286 TFLOPS
  - Only CPUs
  - CLE 5, soon CLE 6
  - SLURM
  - 6 LNET Routers (2x FDR)
  - 2 DVS Nodes (40Gb)
- Available to Grand Challenge Projects
- Jobs >= 256 node desired
HPC @ IU - Compute

- **Karst** – standard cluster available for expansion
  - General purpose Intel Linux cluster
  - Condo nodes may be purchased for special needs or greater response

- Started at ~275 nodes -> ~400
- Upgrade in progress
- First nodes installed in Fall 2014
- NextScale nx360 M4 & M5
- 10/40Gb networking
- Memory profiles from 32GB -> 1024GB
- Using xCAT
- RHEL6 soon with some RHEL7
HPC @ IU – Interactive Compute

• 13 “fat” nodes with ThinLinc remote desktop
• Serving users with interactive needs and users new to HPC
• Test bed for HPC convenience features
HPC @ IU – Cloud Compute

- **Jetstream** – NSF production cloud
- NSF’s first cloud dedicated to science and engineering research across all areas of activity supported by the NSF
- Interactive/On-Demand System
- User-selectable library of VMs
- Supporting 9 science gateways currently
  - Galaxy, CyVerse, SEAGrid, others
- >1,500 users in 1\textsuperscript{st} year
- 20% new to XSEDE
HPC @ IU – Storage

Data Capacitor, DC-WAN, DC-RAM

• Data storage on disk, not backed up (scratch & projects)
• Temporary storage of research data – purged regularly
• 5.3 PB DCII / 1.1 PB DC-WAN / 35 TB DC-RAM
• Wrangler (dual-site 20 PB environment with TACC)

• Lustre-based file systems
• In the midst of storage procurement
• Will add 1-2 file systems and ~2x capacity
HPC @ IU – Storage

Scholarly Data Archive (SDA)

• Distributed tape storage for large-scale archival/near-line storage
• Mirrored – 2 copies (IUB and IUPUI)
• Open to IU community – undergrads/non-IU must have sponsor
• Supports collaborative activities

• 43 PB of tape storage capacity
• Supports SFTP, HSI, HPSS API
• HIPAA-aligned
Contents

• Indiana University and HPC@IU
• What is Performance Tuning
• Examples of Performance Engineering
• SPEC High Performance Group
What is Performance Tuning?

• Decrease the resource need or increase the output of the application/workflow.
• ... of scientific applications.
  – Make them run faster.
  – Make them run at all.
  – Run problem sizes impossible without tuning.
What is Performance Tuning?

• ... of scientific workflows.
  – Make the whole computational workflow run faster.
  – Work with a research group to enable research otherwise impossible.
What is Performance Tuning?

Time

Development

Testing

Production / Research

Performance Analysis
What is Performance Tuning?

Time

- Development
- Testing
- Production / Research
- Performance Analysis
What is Performance Tuning?

Time

Development
Testing
Production / Research
Performance Analysis
Contents

• Indiana University and HPC@IU
• What is Performance Tuning
• Examples of Performance Engineering
• SPEC High Performance Group
Examples of Performance Engineering

• Karst Desktop – Entry Level HPC
• Workflow Tuning for WGS
• Trinity Performance Tuning
• Agro-IBIS Performance Tuning
Entry Level HPC

• A way to make supercomputing more user friendly
• A new way to login and interact with the Karst cluster
• A GUI/desktop instead of a terminal
• Based on ThinLinc, a Linux remote desktop solution using VNC and SSH
Entry Level HPC

User Desktop/Laptop

SSH

Karst Login Nodes (3)

SSH

VNC through SSH

Karst Desktop Nodes (13)

SSH

Karst Compute Nodes (256)
Karst Desktop - Features

• More user friendly interface than a terminal
  – A new front end to Karst, with new capabilities
  – Filesystem browser and file editors/viewers
• Graphical access to compute nodes (indirectly)
• Works more seamlessly compared to X forwarding
  – Addresses latency issues
  – Really great for GUI based applications
• Convenient data transfer/share options
• Supports long running tasks (disconnect / reconnect)
• Supports ssh keys
Karst Desktop – Use Cases

- Running mathematical and statistical applications
- GUIs of HPC applications such as Vampir, Allinea MAP, TotalView
- Visualization
- COMSOL Multiphysics Client/Server
- Data Enclave
- Desktop environment for crystallography tool suite
- Easy access to compilers for classes
- Long running data movement jobs
- Facilitates collaboration
Karst Desktop – Use Cases
Workflow Tuning for WGS

- Broad Reference Pipeline (with very minor modifications)
  - 19 stages, 10 days of runtime
  - Going from 200 GB to 1 TByte per subject
- 818 Alzheimers patients
  - 150 TByte of total data
  - 100x coverage
- Final result:
  - Reduced pipeline runtime by 30% and output volume by 20%
Runtime per Pipeline Step

- 1-2 $HTSUTILS bamshuf/bam2fq: 0
- 3-4 sed R1/R2: 8
- 5-6 $BWA aln R1/R2: 20
- 7-8-9 $BWA sampe $SAMTOOLS view/$SAMTOOLS sort: 42
- 10 $BAMUTILS filter: 11
- 11 java $PICARD: 6
- 12 $SAMTOOLS index: 14
- 13 java GATK RealignerTargetCreator: 1
- 14 java GATK IndelRealigner: 1
- 15 java GATK BaseRecalibrator: 18
- 16 java GATK PrintReads: 1
- 17-18 java GATK ReduceReads/BaseRecalibrator: 1
- 19 java GATK AnalyzeCovariates: 69
Runtime Comparision by Application

- $HTSUTILS
- sed
- $BWA
- $SAMTOOLS
- $BAMUTILS
- java - PICARD
- java - GATK

Runtime [hours]

Original
Tuned
Trinity Performance Tuning

• Work done in 2012, together with ZIH and the BROAD Institute.
  – Matthias Lieber, Richard LeDuc, Brian Haas

• Resulted in a successful NIH grant proposal with BROAD and ZIH.
Trinity

• A bioinformatics code
  – Actually... a Perl script that calls a whole bunch of binaries – a workflow.

• Runtime can be hours, days, or even weeks, depending on input data and compute resource

• Open source with 3rd party dependencies
Our Plan

• Reproduce results from previous performance paper
• Perform general optimizations
• Optimize components
• Publish results
• Push modified source code into official repository
Performance Visualization - CollectL

Jellyfish

Inchworm  GraphFromFastA  ReadsToTranscripts  QuantifyGraph  Butterfly
General Optimizations

• Only global optimizations that can be applied by end users
  – Compiler and runtime options

• Building with the Intel Compiler where possible
  – Using "-fast" compiler flags:
    -ipo -O3 -no-prec-div -static -xHost

• Thread placement and pinning using KMP_AFFINITY and "numactl"

• Input/Output and temporary files on "/dev/shm"
General Optimizations

Jellyfish
Inchworm
GraphFromFastA
ReadsToTranscripts
QuantifyGraph
Butterfly
Optimizing Components

• Inchworm
• GraphFromFastA
• QuantifyGraph
• Other components
Optimizing Inchworm

- Intel’s OpenMP runtime seems superior to GCC’s, for this workload
Optimizing GraphFromFastA

- Parallelizing read counting phase
- Optimized file input to reduce OpenMP critical section
- 10x faster on 32 cores
Optimizing GraphFromFastA

- Improved scalability.
Optimizing QuantifyGraph

• Thousands of embarrassingly parallel tasks, with runtimes of 160 ms to 25 min
• Optimized relational operator “<”
• Reducing “system()” calls
• Reducing the read buffer from 200MB to 1kB
• 5x faster on 32 cores
Optimizing QuantifyGraph

- Improved scalability.
Optimizing Other Components

- Increasing the maximum for the “--CPU” parameter from 22 to 64
- Converting input files in parallel
- Setting “--max_memory” for Jellyfish to 20G, which reduces the number of times it flushes data
- Reduce Java GC threads for Butterfly to 4 per JVM
Final Results

- **Original (Mason)**
- **General opt. (Mason)**
- **Optimized (Mason)**
- **Optimized (Blacklight)**

**Runtime (hours)** vs. **Data set size (Gbp)**
Agro-IBIS

• Simulates agricultural ecosystems
  – Inputs include climate and weather data, farming decisions, and landscape properties
  – Outputs include physical state variables, fluxes, and agricultural parameters
  – Widely validated results for Midwestern US

• Serial, Fortran code

• Data is available to simulate at much larger scale

• Need to develop an HPC implementation of Agro-IBIS to solve large-scale models
Agro-IBIS

• Development: Gains and Constraints
  – Strong desire to maintain consistency with community code
  – Optimizations desired to be drop-in or easily integrated with downloaded code
  – Implementation of netCDF library for standardized, optimized data storage
  – Parallel MPI wrapper written in C++ to manage domain decomposition and job launching
  – Previously unrecognized I/O bottleneck waiting to show up in parallel runs
Agro-IBIS

- Method for running IBIS puts a lot of strain on the filesystems used
  - Inputs and outputs for each IBIS run are separate file trees
  - IBIS instances scale perfectly – if you could ignore I/O cost
  - Very easy to tax the MDS without realizing it
  - This is just an example of what any conventional, serial app would do when domain decomposition doesn’t take I/O into account.
Contents

• Indiana University and HPC@IU
• What is Performance Tuning
• Examples of Performance Engineering
• SPEC HPG
SPEC is a non-profit corporation formed to "establish, maintain and endorse a standardized set of relevant benchmarks that can be applied to the newest generation of high-performance computers"

- Composed of four groups
  - Graphics and Workstation Performance Group (GWPG)
  - High Performance Group (HPG)
  - Open Systems Group (OSG)
  - Research Group (RG)

- [https://www.spec.org](https://www.spec.org)
SPEC High Performance Group

• Develops benchmarks to represent high-performance computing applications for standardized, cross-platform performance evaluation.

• Benchmarks
  • SPEC OMP2012
  • SPEC MPI2007
  • SPEC ACCEL
SPEC ACCEL 1.2 – OpenMP Target

• Version 1.2 of the SPEC ACCEL benchmark was released this week.

• Addition of OpenMP suite with target directives
SPEC HPG Search Program

• We are building a new benchmark.
• MPI+X; Where X can be:
  – Nothing
  – Accelerator paradigms: CUDA, OpenACC, OpenMP4, ...
  – Parallel paradigms: OpenMP, Threads, ...
  – Libraries like Kokkos, TBB, MKL, ...
• https://www.spec.org/hpg/search/
Acknowledgement

• Tom Doak, Arvind Gopu, Abhinav Thota, Dave Hancock, Richard Meraz for providing material for this presentation.
• Matthias Lieber for his work on Trinity.
• Huian Li for his work on the Alzheimer project.
• Holger Brunst, Shawn Slaving, Steven Simms, Cicada Dennis for their work on Agro-IBIS.
• RT leadership for all the support over the last couple of years.

This material is based in part upon work supported by the National Science Foundation under Grant Numbers ACI-1445604, DBI-1458641 and ABI-1062432. 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 National Science Foundation.

This research was supported by the National Cancer Institute Information Technology in Cancer Research program of the National Institutes of Health under award number 5U24CA180922-03.
Thank you!

Questions?