

Big Data, Big Red II, Data
Capacitor II, Wrangler,
Jetstream, and Globus Online



funded by the National Science Foundation
Award #ACI-1445604

Big Data, Big Red II, Data Capacitor II, Wrangler, Jetstream, and Globus Online: A national science & engineering cloud

Craig A. Stewart, Ph.D. (Wittenberg class of '81)

Orcid ID: 0000-0003-2423-9019

Executive Director, Pervasive Technology Institute

Associate Dean, Research Technologies

Indiana University

Please cite as: Stewart, C.A. 2014. Big Data, Big Red II, Data Capacitor II, Wrangler, Jetstream, and Globus Online: A national science & engineering cloud. Presentation before the IU School of Public and Environmental Affairs (SPEA), Bloomington, IN. February 2015. Bloomington, IN.

<http://hdl.handle.net/2022/19680>



License specifics in last slide



Award #1445604



pti.iu.edu/jetstream

An Intro to the UITS Research Technologies Division of UITS and Pervasive Technology Institute

- The mission of Research Technologies ... is to develop, deliver and support advanced technology solutions that improve the productivity of and enable new possibilities in research, scholarly endeavors, and creative activity at Indiana University and beyond; and to complement this with education and technology translation activities to improve the quality of life of people in Indiana, the nation, and the world.
- **RT is a mission- and value-driven organization. We are not a technology-driven organization.**
- PTI is a collaborative organization within IU involving OVPIT, UITS, SOIC, Maurer School of Law, and the College of Arts and Sciences. It puts the 'basic CS and Informatics' research at the front end of 'Discover, Develop, Deploy, Deliver, Support' for research
- PTI and RT identify needs, identify possibilities, and discover new ways to meet those needs, realize those possibilities, and create new ones. In so doing, we create, deploy, and support technology. **We are technology driving organizations.**



Big Data (5Vs) Characteristics

- Characterized by (from http://en.wikipedia.org/wiki/Big_data):
 - Volume
 - Variety
 - Velocity
 - Variability
 - Veracity



Big Data analytics (6Cs)

- Consist of (from http://en.wikipedia.org/wiki/Big_data):
 - Connection (sensor and networks)
 - Cloud (computing and data on demand)
 - Cyber (model and memory)
 - Content/Context (meaning and correlation)
 - Community (sharing and collaboration)
 - Customization (personalization and value)



Is there an accepted definition of Big Data?

- Not really, not yet
- Common purposes for Big Data Analyses
 - Marketing, spying, and security (social media, email, video feeds, web page analytics, video feeds)
 - Credit Card fraud detection
 - Manufacturing
 - Medical service delivery
 - Scientific research: LHC, telescope, Earthscope
 - Humanities research and scholarship: text analysis (e.g. when did “the United States” become a singular noun as opposed to plural in common English usage in the US?)
- Lack of definition doesn't matter that much if Big Data analyses techniques help you get new insights from data

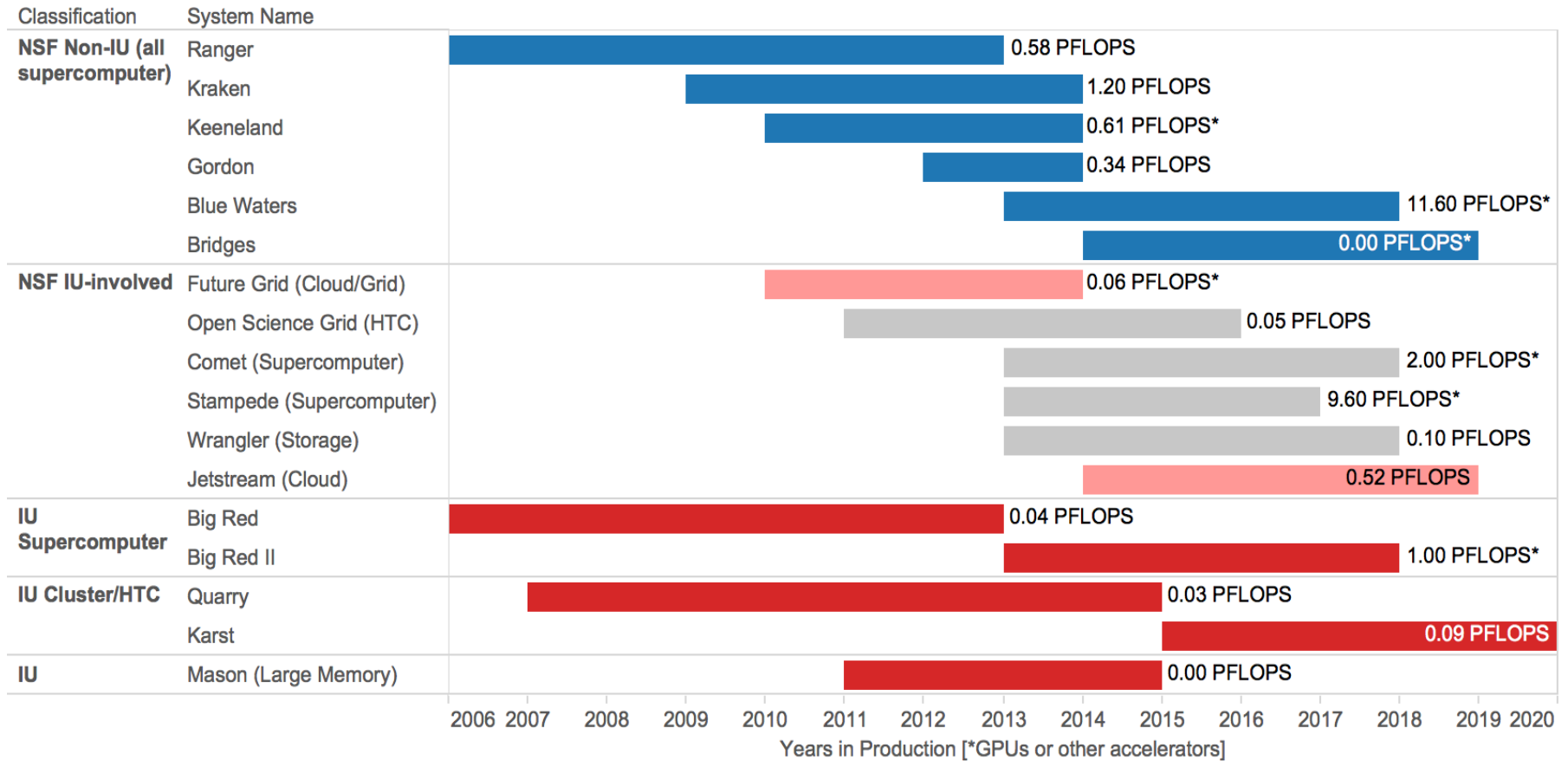


Software and systems for Big Data

- Many sorts of big data are just lots and lots of instances of little data (LHC is a good example)
- In many cases one does want to analyze large amounts of data at once
- MapReduce – programming model that filters (maps) and analyzes (reduce) large amounts of distributed data in parallel
- Hadoop is a software suite (<http://hadoop.apache.org/>) that includes MapReduce, a file system (HDFS), and other utilities
- Microsoft Azure started of as an analytics/Hadoop environment and has morphed
- Amazon Web Services
- NoSQL Databases (MongoDB is most widely known_
- GIS systems – Public Health
- Lustre – open source object store file system



A bit about the world of big systems



Legend

- IU System
- NSF IU Lead
- NSF IU Partner
- NSF Non-IU

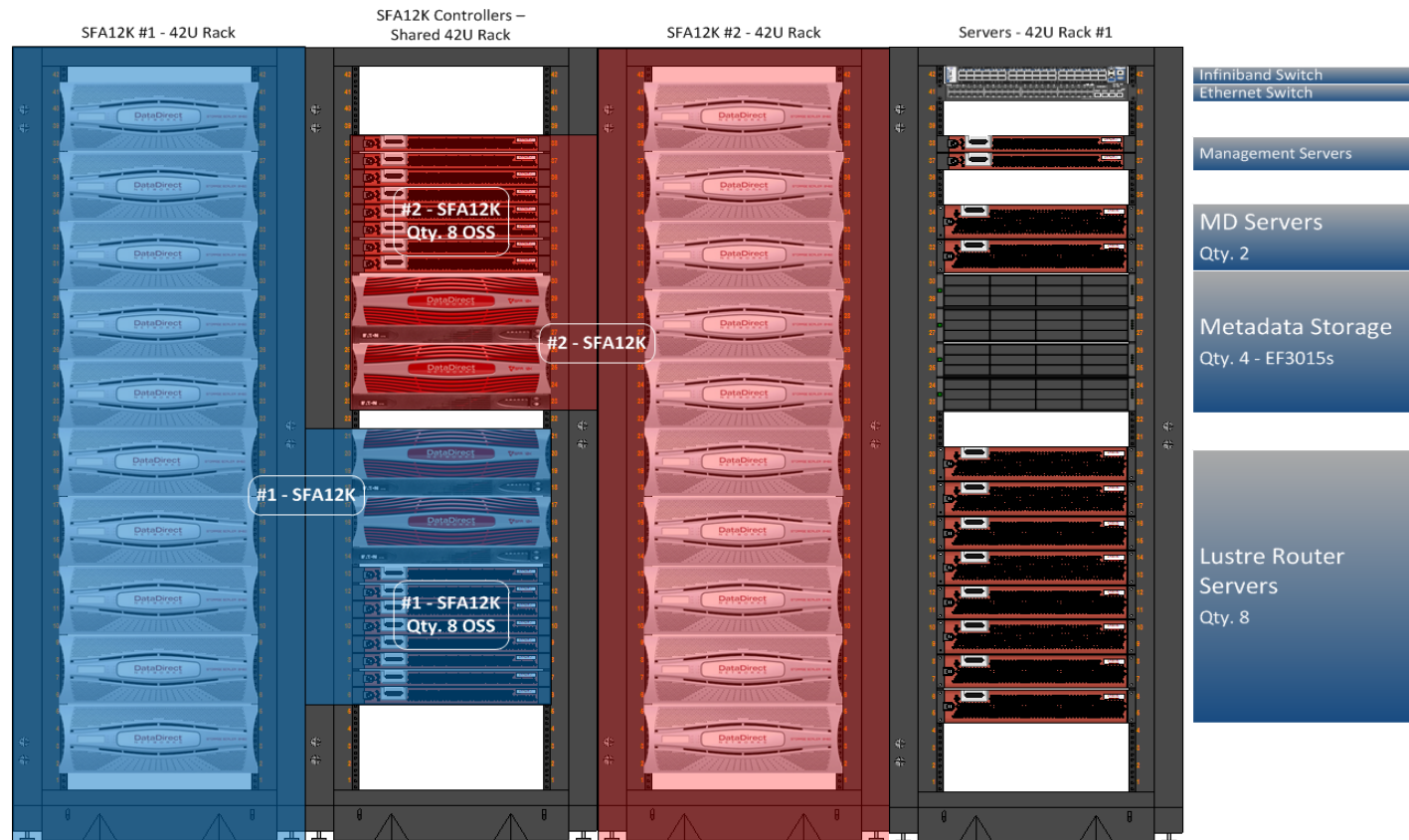


Award #1445604



pti.iu.edu/jetstream

Data Capacitor II



- 5 PB total disk space
- 40 GBps I/O

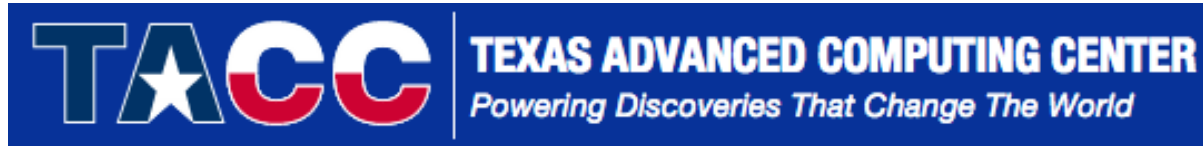


Scholarly Data Archive (now that's REALLY big)

- Archival Storage – 15 PB
- Replicated by default in Indianapolis and Bloomington
- Capability for out-of-region storage in Texas
- Aggregate I/O 10 GB/sec
- Runs under High Performance Storage System software – very secure



Wrangler



- Geographically replicated, high performance data storage (10PB each site)
- Large scale flash storage tier for analytics with bandwidth of 1TB/s and 250M IOPS
- More than 3,000 embedded processor cores for data analysis
- Wide range of software stacks, including Hadoop®, R, relational and noSQL databases
- Globus services for rapid, reliable data transfer, sharing, and publication
- Partnership of Texas Advanced Computing Center, IU Pervasive Technology Institute, University of Chicago, Dell



Jetstream

- Geographically Distributed Cloud, 0.5 PetaFLOPS
- High-speed connections to Internet2 and local connections to Wrangler disk storage at IU and TACC
- Globus-based large scale file movement



Jetstream will . . .

- be NSF's first cloud for science and engineering research across all areas of activity supported by the NSF
- be a user-friendly cloud environment designed to give researchers and research students access to interactive computing and data analysis resources "on demand"
- leverage Globus tools for data movement and authentication
- provide a user-selectable library of virtual machines that users can select from to do their research.
- enable software creators and researchers to create their own customized virtual machines or their own "private computing system" within Jetstream.
- ... and store them in IUScholarWorks with a DOI for publication and support of replicability of research
- enable discoveries across disciplines such as biology, atmospheric science, economics, network science, observational astronomy, and social sciences.

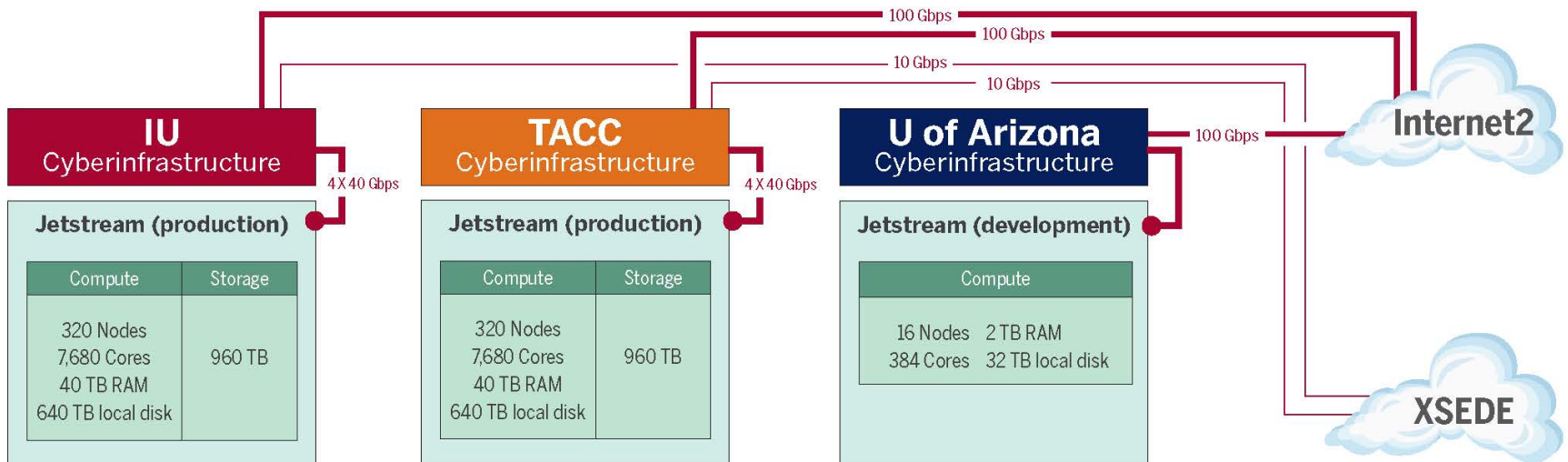


What does the name mean? And is it really a cloud?

- Name
 - In the atmosphere the Jetstream lies at the border of two different air masses
 - The Jetstream system stands at the border of the existing NSF-funded XD program and advanced cyberinfrastructure resources and users who have not previously used such NSF funded infrastructure before.
- Yep, it's really a cloud, or at least a cloud environment (one could quibble over the definition of cloud vis-à-vis expansibility). Software layers:
 - Atmosphere interface
 - KVM
 - OpenStack
 - CentOS Linux (probably)



Jetstream System Diagram



Award #1445604



pti.iu.edu/jetstream

Dashboard

Images

Favorites

My Images

Projects

Cloud Providers

Quotas

Settings

Search Images

Search by App Images, Tag, OS, and more

Popular Searches: [R](#) [Bisque](#) [NGS](#) [Community: Astrophysics](#)


Quick Sort: Popularity Recency Rating

[Advanced Search Options](#)

Quick Filter:

View as:




Popular Images from All Communities


 ☆

Math Kernel Library

[blas](#) [fft](#) [fortran](#) [lapack](#)

Community: Mathematics

 52  0  7




 ★


RNASeq Analysis Tools

[bowtie2](#) [blast](#) [blat](#) [edgeR](#)

[R](#) [rnaseq](#) [tophat2](#)

Community: Biology




 30  2  4


 ☆

Atmospheric Dispersion Modeling

[aermod](#) [aermet](#) [aermap](#)

Community: Atmospheric Sciences

 20  0  0




 ☆


MrBayes with TreeMix


[bayesian inference](#) [mrbayes](#)


[treemix](#)

Community: Phylogenetics

 25  1  10

 ★

 ☆

 ☆

 ☆



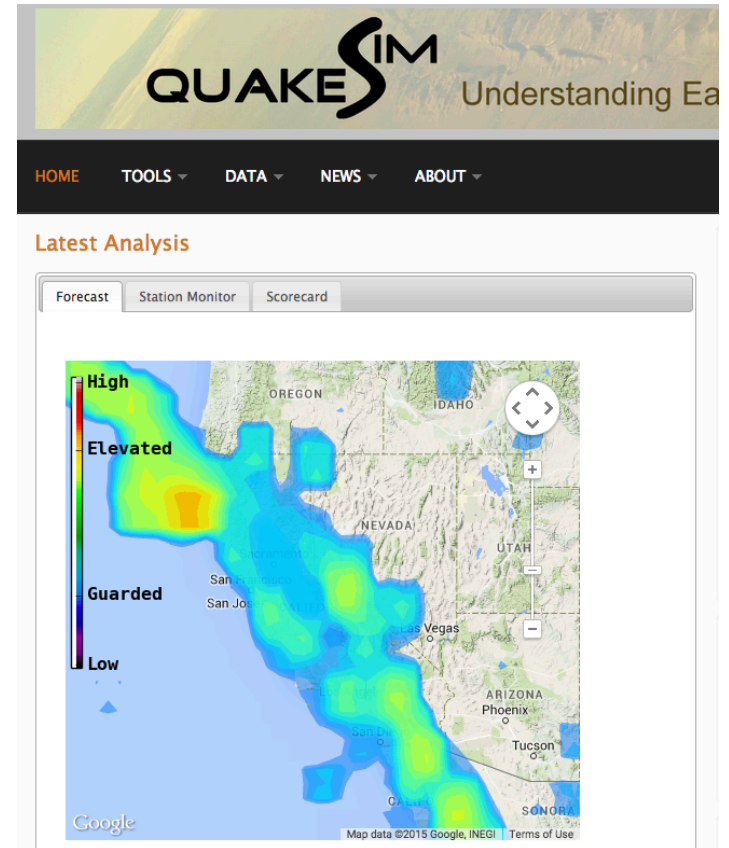
Science Domains and Users

- Biology
- Earth Science/Polar Science
- Field Station Research
- Geographical Information Systems
- Network Science
- Observational Astronomy
- Social Sciences
- Jetstream will be particularly focused on researchers working in the “long tail” of science with born digital data
- Enabling analysis of field-collected empirical data on the impact and effects of global climate change will be one of the specific foci of Jetstream
- Whatever *you* do



Types of applications supported

- Interactive, VM-based work
- Persistent science gateways
- Hadoop at modest scale



21st-century workforce development

- Jetstream will include virtual Linux desktops and applications specifically aimed to enable research and research education at small colleges and universities including HBCUs (Historically Black Colleges and Universities), MSIs (Minority Serving Institutions), Tribal colleges, and higher-ed institutions in EPSCoR States
- Jetstream will also support deployment of user-friendly Science Gateways



Jetstream Deployment Partner Organizations

A seasoned team of organizations and experts:

- University of Texas Austin (TACC)
- University of Chicago (Argonne National Lab)
- (The above trio should look familiar)
- University of Arizona
- University of Texas at San Antonio (Open Cloud Lab)
- Johns Hopkins University
- Penn State University



Globus Online – the biggest thing in big data movement



Products ▾

News ▾

About ▾

Support ▾

Lo



Research data
management
simplified.

79,658,453,221 MB
TRANSFERRED



Award #1445604



pti.iu.edu/jetstream

Portal to IU computing resources

- Get updates on Big Red II, Karst, and Mason status.
- Move files from your desktop to IU's Scholarly Data Archive and Data Capacitor.
- Monitor and manage running jobs on Big Red II, Karst, and Mason.
- Find information on available software.

System Status

Resource	Status	Utilization	Job
Big Red II	Healthy	Total Nodes: 1020 Running Nodes: 0 Busy Nodes: 928 Idle Nodes: 68 Drained Nodes: 0 Other Nodes: 24	Total: 456 Running: 377 Idle: 71 Not Queued: 0 Completed: 0 Other: 8
Karst	Healthy	Total Nodes: 249 Running Nodes: 22 Busy Nodes: 82 Idle Nodes: 142 Drained Nodes: 2 Other Nodes: 1	Total: 190 Running: 121 Idle: 66 Not Queued: 0 Completed: 3 Other: 0
Mason	Healthy	Total Nodes: 18 Running Nodes: 4 Busy Nodes: 14 Idle Nodes: 0 Drained Nodes: 0 Other Nodes: 0	Total: 202 Running: 60 Idle: 142 Not Queued: 0 Completed: 0 Other: 0

Protected Health Information

- IU cyberinfrastructure is HIPAA aligned
- Globus Transfer is not . . . Yet
- HIPAA alignment of IU systems potentially very useful in public health / public policy research



Seeing big data – IQ-Wall



Award #1445604



pti.iu.edu/jetstream

Note: we've talked a lot about gear and support professionals. IU has many “Big Data” faculty experts

- SOIC
 - Katy Boerner - Mapping
 - Johan Bollen – Tweet feed analyses
 - Geoffrey Fox - FutureGrid
 - Fil Menczer - Truthy
 - Beth Plale –HathiTrust Research Center
 - Judy Qiu - Twister
- Not to mention the College and particularly the Digital Humanities program
- And of course SPEA has experts in relevant areas of research, relevant history, and many areas of potential application of Big Data approaches



Questions and discussion?



Award #1445604



pti.iu.edu/jetstream

License Terms

- Stewart, C.A. 2014. Big Data, Big Red II, Data Capacitor II, Wrangler, Jetstream, and Globus Online: Jetstream - a national science & engineering cloud. Presentation before the IU School of Public and Environmental Affairs (SPEA), Bloomington, IN. February 2015. Bloomington, IN.
<http://hdl.handle.net/2022/19680>
- Items indicated with a © are under copyright and used here with permission. Such items may not be reused without permission from the holder of copyright except where license terms noted on a slide permit reuse.
- Except where otherwise noted, contents of this presentation are copyright 2014 by the Trustees of Indiana University.
- This document is released under the Creative Commons Attribution 3.0 Unported license (<http://creativecommons.org/licenses/by/3.0/>). This license includes the following terms: You are free to share – to copy, distribute and transmit the work and to remix – to adapt the work under the following conditions: attribution – you must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). For any reuse or distribution, you must make clear to others the license terms of this work.
- This research was supported in part by the National Science Foundation through Award ACI-1445604. This research was supported in part by the Indiana University Pervasive Technology Institute, which was established with the assistance of a major award from the Lilly Endowment, Inc. Opinions presented here are those of the author(s) and do not necessarily represent the views of the NSF, IUPTI, IU, or the Lilly Endowment, Inc.

