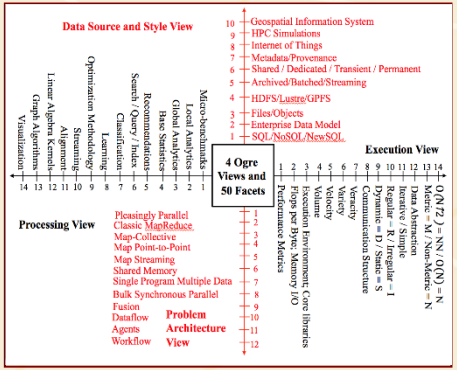
Digital Science Center

Geoffrey C. Fox, David Crandall, Judy Qiu, Gregor von Laszewski, Fugang Wang, Badi' Abdul-Wahid
Saliya Ekanayake, Supun Kamburugamuva, Jerome Mitchell, Bingjing Zhang
School of Informatics and Computing, Indiana University

Digital Science Center Research Areas

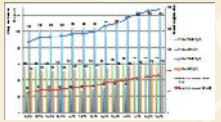
- Digital Science Center **Facilities**
- **RaPyDLI** Deep Learning Environment
- **SPIDAL** Scalable Data Analytics Library and applications including Ice Sheet
- **MIDAS** Big Data Software
- **Big Data Ogres** Classification and Benchmarks
- **CloudIOT** Internet of Things Environment
- **Cloudmesh** Cloud and Bare metal Automation
- **XSEDE TAS** Monitoring citations and system metrics
- **Data Science Education** with MOOC's

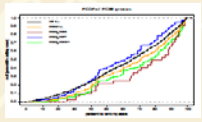


Kaleidoscope of (Apache) Big Data Stack (ABDS) and HPC Technologies		
Cross-Cutting Functions	17) **Workflow-Orchestration:** ODE, ActiveBPEL, Airavata, Pegasus, Kepler, Swift, Taverna, Triana, Trident, BioKepler, Galaxy, IPython, Dryad, Naiad, Oozie, Tez, Google FlumeJava, Crunch, Cascading, Scalding, e-Science Central, Azure Data Factory, Google Cloud Dataflow, Nifi (NSA), Jitterbit, Talend, Pentaho, Apatar, Docker Compose	
7) Message and Data Protocols: Avro, Thrift, Protobuf	18) **Application and Analytics:** Mahout , MLlib , MLbase, DataFu, R, pbdR, Bioconductor, ImageJ, OpenCV, Scalapack, PetSc, Azure Machine Learning, Google Prediction API & Translation API, mlpy, scikit-learn, PyBrain, CompLearn, DAL(Intel), Caffe, Torch, Theano, DL4J, H2O, IBM Watson, Oracle PGX, GraphLab, GraphX, IBM System G, GraphBuilder(Intel), TinkerPop, Google Fusion Tables, CINET, NWB, Elasticsearch, Kibana, Logstash, Graylog, Splunk, Tableau, D3.js, three.js, Potree, DC.js	
8) Distributed Coordination: Google Chubby, Zookeeper, Giraffe, Jgroups	19) **Application Hosting Frameworks:** Google App Engine, AppScale, Red Hat OpenShift, Heroku, Aerobatic, AWS Elastic Beanstalk, Azure, Cloud Foundry, Pivotal, IBM BlueMix, Ninefold, Jelastic, Stackato, appfog, CloudBees, Engine Yard, CloudControl, dotCloud, Dokku, OSGi, HUBzero, OODT, Agave, Atmosphere	
9) Security & Privacy: InCommon, Eduroam OpenStack Keystone, LDAP, Sentry, Sqrd, OpenID, SAML OAuth	12A) **High level Programming:** Kite, Hive, HCatalog, Tajo, Shark, Phoenix, Impala, MRQL, SAP HANA, HadoopDB, PolyBase, Pivotal HD/Hawq, Presto, Google Dremel, Google BigQuery, Amazon Redshift, Drill, Kyoto Cabinet, Pig, Sawzall, Google Cloud DataFlow, Summingbird	
	14B) **Streams:** Storm, S4, Samza, Granules, Google MillWheel, Amazon Kinesis, LinkedIn Databus, Facebook Puma/Ptail/Scribe/ODS, Azure Stream Analytics, Floe	
	14A) **Basic Programming model and runtime, SPMD, MapReduce:** Hadoop, Spark, Twister, MR-MPI, Stratosphere (Apache Flink), Reef, Hama, Giraph, Pregel, Pegasus, Ligra, GraphChi, Galois, Medusa-GPU, MapGraph, Totem	
	15) **Inter process communication Collectives, point-to-point, publish-subscribe:** MPI, Harp, Netty, ZeroMQ, ActiveMQ, RabbitMQ, NaradaBrokering, QPid, Kafka, Kestrel, JMS, AMQP, Stomp, MQTT, Marionette Collective, **Public Cloud:** Amazon SNS, Lambda, Google Pub Sub, Azure Queues, Event Hubs	
	12) **In-memory databases/caches:** Gora (general object from NoSQL), Memcached, Redis, LMDB (key value), Hazelcast, Ehcache, Infinispan	
	12) **Object-relational mapping:** Hibernate, OpenJPA, EclipseLink, DataNucleus, ODSC/JDBC	
	12) **Extraction Tools:** UIMA, Tika	
	11C) **SQL(NewSQL):** Oracle, DB2, SQL Server, SQLite, MySQL, PostgreSQL, CUBRID, Galera Cluster, SciDB, Rasdaman, Apache Derby, Pivotal Greenplum, Google Cloud SQL, Azure SQL, Amazon RDS, Google F1, IBM dashDB, NIQL, BlinkDB	
10) Monitoring: Ambari, Ganglia, Nagios, Inca	11B) **NoSQL:** Lucene, Solr, Solandra, Voldemort, Riak, Berkeley DB, Kyoto/Tokyo Cabinet, Tycoon, Tyrant, MongoDB, Espresso, CouchDB, CouchBase, IBM Cloudant, Pivotal GemFire, HBase, Google Bigtable, LevelDB, Megastore and Spanner, Accumulo, Cassandra, RYA, Sqrd, Neo4J, Yarcdata, AllegroGraph, Blazegraph, Facebook Tao, Titan:db, Jena, Sesame	
	Public Cloud: Azure Table, Amazon Dynamo, Google DataStore	
	11A) **File management:** iRODS, NetCDF, CDF, HDF, OPeNDAP, FITS, RCFile, ORC, Parquet	
21 layers Over 350 Software Packages May 15 2015	10) **Data Transport:** BitTorrent, HTTP, FTP, SSH, Globus Online (GridFTP), Flume, Sqoop, Pivotal GPLOAD/GPFDIST	
	9) **Cluster Resource Management:** Mesos, Yarn, Helix, Llama, Google Omega, Facebook Corona, Celery, HTCondor, SGE, Platform, Moab, Slurm, Torque, Globus Tools, Pilot Jobs	
	8) **File systems:** HDFS, Swift, Haystack, 9i, Cinder, Ceph, PUSE, Gluster, Lustre, GPFS, GFFS	Green implies HPC Integration
	Public Cloud: Amazon S3, Azure Blob, Google Cloud Storage	
	7) **Interoperability:** Libvirt, Libcloud, JClouds, TOSCA, OCCI, CDMI, Whirr, Saga, Genesis	
	6) **DevOps:** Docker (Machine, Swarm), Puppet, Chef, Ansible, SaltStack, Boto, Cobbler, Xcat, Razor, CloudMesh, Juju, Foreman, OpenStack Heat, Sahara, Rocks, Cisco Intelligent Automation for Cloud, Ubuntu MaaS, Facebook Tupperware, AWS OpsWorks, OpenStack Ironic, Google Kubernetes, Buildstep, Gitreceive, OpenTOSCA, Winery, CloudML, Blueprints, Terraform, DevOpSlang, Any2Api	
	5) **IaaS Management from HPC to hypervisors:** Xen, KVM, Hyper-V, VirtualBox, OpenVZ, LXC, Linux-Vserver, OpenStack, OpenNebula, Eucalyptus, Nimbus, CloudStack, CoreOS, rkt, VMware ESXi, vSphere and vCloud, Amazon, Azure, Google and other public Clouds	
	Networking: Google Cloud DNS, Amazon Route 53	

Scientific Impact Metrics

We developed a software framework and process to evaluate scientific impact for XSEDE. We calculate and track various Scientific Impact Metrics of XSEDE., BlueWaters, and NCAR. In recently conducted peer comparison study we showed how well XSEDE and its support services (ECSS) perform judging by citations received. During this process we retrieved and processed millions of data entries from multiple sources in various formats to obtain the result.



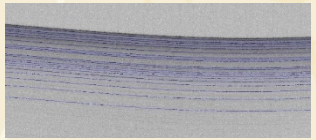


	# Publications	Rank - Average	Rank - Median	# Citation - Average	# Citation - Median
XSEDE	2349	61	65	26	11
Peers	168422	49	48	13	5

Ice Layer Detection Algorithm

The polar science community has built radars capable of surveying the polar ice sheets, and as a result, have collected terabytes of data and is increasing its repository each year as signal processing techniques improve and the cost of hard drives decrease enabling a new-generation of high resolution ice thickness and accumulation maps.

Manually extracting layers from an enormous corpus of ice thickness and accumulation data is time-consuming and requires sparse hand-selection, so developing image processing techniques to automatically aid in the discovery of knowledge is of high importance.



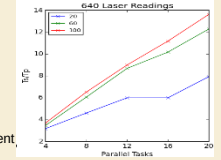
Data analytics for IoT devices in Cloud

We developed a framework to bring data from IoT devices to a cloud environment for real time data analysis. The framework consists of; Data collection nodes near the devices, Publish-subscribe brokers to bring data to cloud and Apache Storm coupled with other batch processing engines for data processing in cloud. Our data pipe line is Robot → Gateway → Message Brokers → Apache Storm.

Simultaneous Localization and Mapping(SLAM) is an example application built on top of our framework, where we exploit parallel data processing to speedup the expensive SLAM computation.

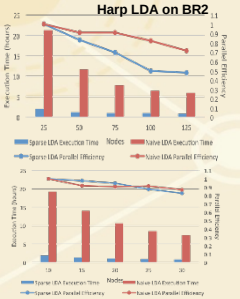
Robot

Map of the Environment



Parallel Sparse LDA

- Original LDA (orange) compared to LDA exploiting sparseness (blue)
- Note data analytics making use of Infiniband (i.e. limited by communication!)
- Java code running under Harp – Hadoop plus HPC plugin
- Corpus: 3,775,554 Wikipedia documents, Vocabulary: 1 million words; Topics: 10k topics;
- BR II is Big Red II supercomputer with Cray Gemini interconnect
- Juliet is Haswell Cluster with Intel (switch) and Mellanox (node) infiniband (not optimized)



Harp LDA on Juliet (36 core nodes)

High Performance Data Analytics with Java + MPI on Multicore HPC Clusters

We find it is challenging to achieve high performance in HPC clusters for big data problems. We approach this with **Java** and **MPI**, but improves further using Java **memory maps** to and **off heap data structures.** We achieve **zero intra-node messaging**, **zero GC**, and **minimal memory** footprint. We present performance results of running it on a latest **Intel Haswell HPC cluster** consisting **3456 cores** total.