NSF14-43054 start October 1, 2014

Datanet: CIF21 DIBBs: Middleware and High Performance Analytics Libraries for Scalable Data Science

- Indiana University (Fox, Qiu, Crandall, von Laszewski),
- Rutgers (Jha)
- Virginia Tech (Marathe)
- Kansas (Paden)
- Stony Brook (Wang)
- Arizona State (Beckstein)
- Utah (Cheatheam)

Overview by Gregor von Laszewski April 6 2016

http://spidal.org/
http://news.indiana.edu/releases/ia/2014/10/big-data-dibbs-grant.shtml
Some Important Components of SPIDAL Dibbs

- **NIST Big Data Application Analysis**: features of data intensive apps
- **HPC-ABDS**: Cloud-HPC interoperable software with performance of HPC (High Performance Computing) and the rich functionality of the commodity Apache Big Data Stack.
  - Reservoir of software subsystems – nearly all from outside project and mix of HPC and Big Data communities
  - Leads to Big Data – Simulation - HPC Convergence
- **MIDAS**: Integrating Middleware – from project
- **Applications**: Biomolecular Simulations, Network and Computational Social Science, Epidemiology, Computer Vision, Spatial Geographical Information Systems, Remote Sensing for Polar Science and Pathology Informatics.
- **SPIDAL (Scalable Parallel Interoperable Data Analytics Library)**: Scalable Analytics for
  - Domain specific data analytics libraries – mainly from project
  - Add Core Machine learning Libraries – mainly from community
  - Performance of Java and MIDAS Inter- and Intra-node
- **Benchmarks** – project adds to community; See WBDB 2015 Seventh Workshop on Big Data Benchmarking
Java MPI performs better than Threads
128 24 core Haswell nodes on SPIDAL DA-MDS Code

Best Threads intra node
Best MPI; inter and intra node
Best Threads intra node
Best MPI, inter node
Kaleidoscope of (Apache) Big Data Stack (ABDS) and HPC Technologies

Cross-Cutting Functions

1) Message and Data Protocols: Avro, Thrift, Protobuf
2) Distributed Coordination: Google Chubby, Zookeeper, Giraffe, JGroups
3) Security & Privacy: InCommon, Eduroam, OpenStack, Keystone, LDAP, Sentry, Sqrrl, OpenID, SAML, OAuth
4) Monitoring: Ambari, Ganglia, Nagios, Inca

Monitoring:


2) Application and Analytics:
   - Mahout, MLlib, MLbase, DataFu, R, pbDr, Bioconductor, ImageJ, OpenCV, Scalapack, PetSc, PLASMA MAGMA, Azure Machine Learning, Google Prediction API & Translation API, mllib, scikit-learn, PyBrain, CompLearn, DAAL (Intel), Caffe, Torch, Theano, DL4j, H2O, IBM Watson, Oracle PGX, GraphLab, GraphX, IBM System G, GraphBuilder (Intel), TinkerPop, Parasol, Dream:Lab, Google Fusion Tables, CINET, NWB, Elasticsearch, Kibana, Logstash, Graylog, Splunk, Tableau, D3.js, three.js, Potree, DC.js, TensorFlow, CNTK

3) Inter process communication Collectives, point-to-point, publish-subscribe: MPI, HPX-5, Argo BEAST HPX-5 BEAST PULSAR, 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

4) In-memory databases/caches: Gora (general object from NoSQL), Memcached, Redis, LMDB (key value), Hazelcast, Ehcache, Infinispan, VoltDB, H-Store

5) Object-relational mapping: Hibernate, OpenJPA, EclipseLink, DataNucleus, ODBC/JDBC

6) Extraction Tools: UIMA, Tika

7) SQL (NoSQL):
   - Oracle, DB2, SQL Server, SQLite, MySQL, PostgreSQL, CUBRID, Galera Cluster, SciDB, Rasdaman, Apache Derby, Pivotal Greenplum, Google Cloud SQL, Azure SQL, Amazon RDS, MySQL, IBM DB2, N1QL, BlinkDB, Spark SQL

8) NoSQL:
   - Lucene, Solr, Solandra, Voldemort, Riak, ZHT, 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, Sqrrl, Neo4J, graphdb, Yarcdata, AllegroGraph, Blazegraph, Facebook Tajo, Titan:db, Jena, Sesame

9) Public Cloud: Amazon S3, Google Cloud Storage

10) File management: iRODS, NetCDF, CDF, HDF, OPeNDAP, FITS, RCFile, ORC, Parquet

11) Data Transport: BitTorrent, HTTP, FTP, SSH, Globus Online (GridFTP), Flume, Sqoop, Pivotal GLOAD/GPFDIST

12) Cluster Resource Management: Mesos, Yarn, Helix, Llama, Google Omega, Facebook Corona, Celery, HTCondor, SGE, OpenPBS, Moab, Slurm, Torque, Globus Tools, Pilot Jobs

13) File systems: HDFS, Swift, Haystack, f4, Cinder, Ceph, FUSE, Gluster, Lustre, GPFS, GFFS

14) Public Cloud: Amazon S3, Azure Blob, Google Cloud Storage

15) Interoperability: Libvirt, Libelcloud, JClouds, TOSCA, OCCI, CDMI, Whirr, Saga, Genesis


17) IaaS Management from HPC to hypervisors: Xen, KVM, QEMU, 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

January 29, 2016

21 layers

Over 350 Software Packages
HPC-ABDS Mapping of Activities

- **Level 17: Orchestration**: Apache Beam (Google Cloud Dataflow) integrated with Cloudmesh
- **Level 16: Applications**: Datamining for molecular dynamics, Image processing for remote sensing and pathology, graphs, streaming
- **Level 16: Algorithms**: MLlib, Mahout, SPIDAL
  - Release SPIDAL Clustering, Dimension Reduction (multi-dimensional scaling), Web point set visualization
- **Level 14: Programming**: Storm, Heron, Hadoop, Spark, Flink. Inter and Intra-node performance
- **Level 13: Communication**: Enhanced Storm and Hadoop using HPC runtime technologies
- **Level 11: Data management**: Hbase and MongoDB integrated via use of Beam and other Apache tools
- **Level 9: Cluster Management**: Integrate Pilot Jobs with Yarn, Spark, Hadoop
- **Level 6: DevOps**: Python Cloudmesh virtual Cluster Interoperability