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(Cheatham)*

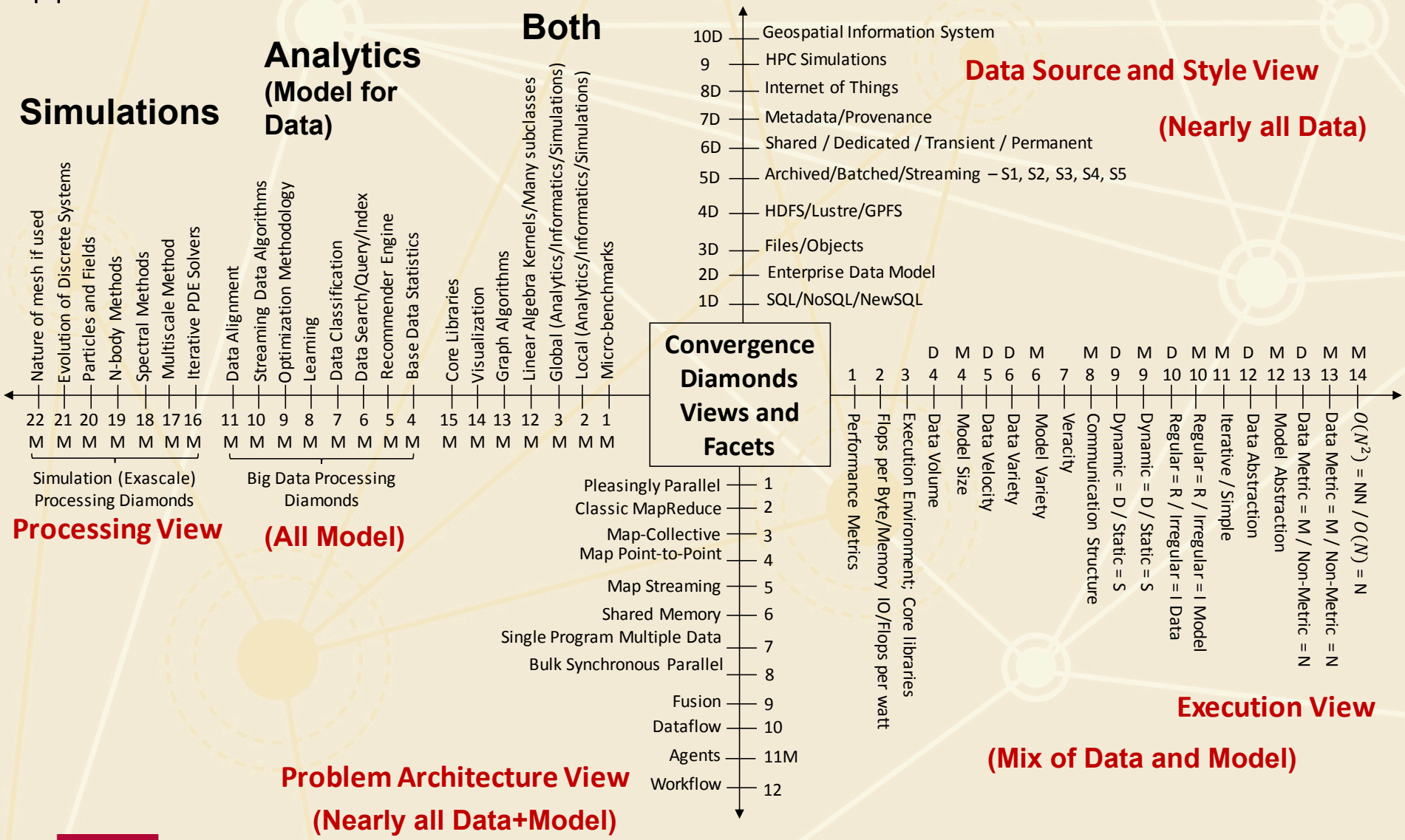
Overview by *Gregor von Laszewski* April 6 2016

http://spidal.org/
http://news.indiana.edu/releases/iu/2014/10/big-data-dibbs-grant.shtml
http://www.nsf.gov/awardsearch/showAward?AWD_ID=1443054

Some Important Components of SPIDAL Dibbs

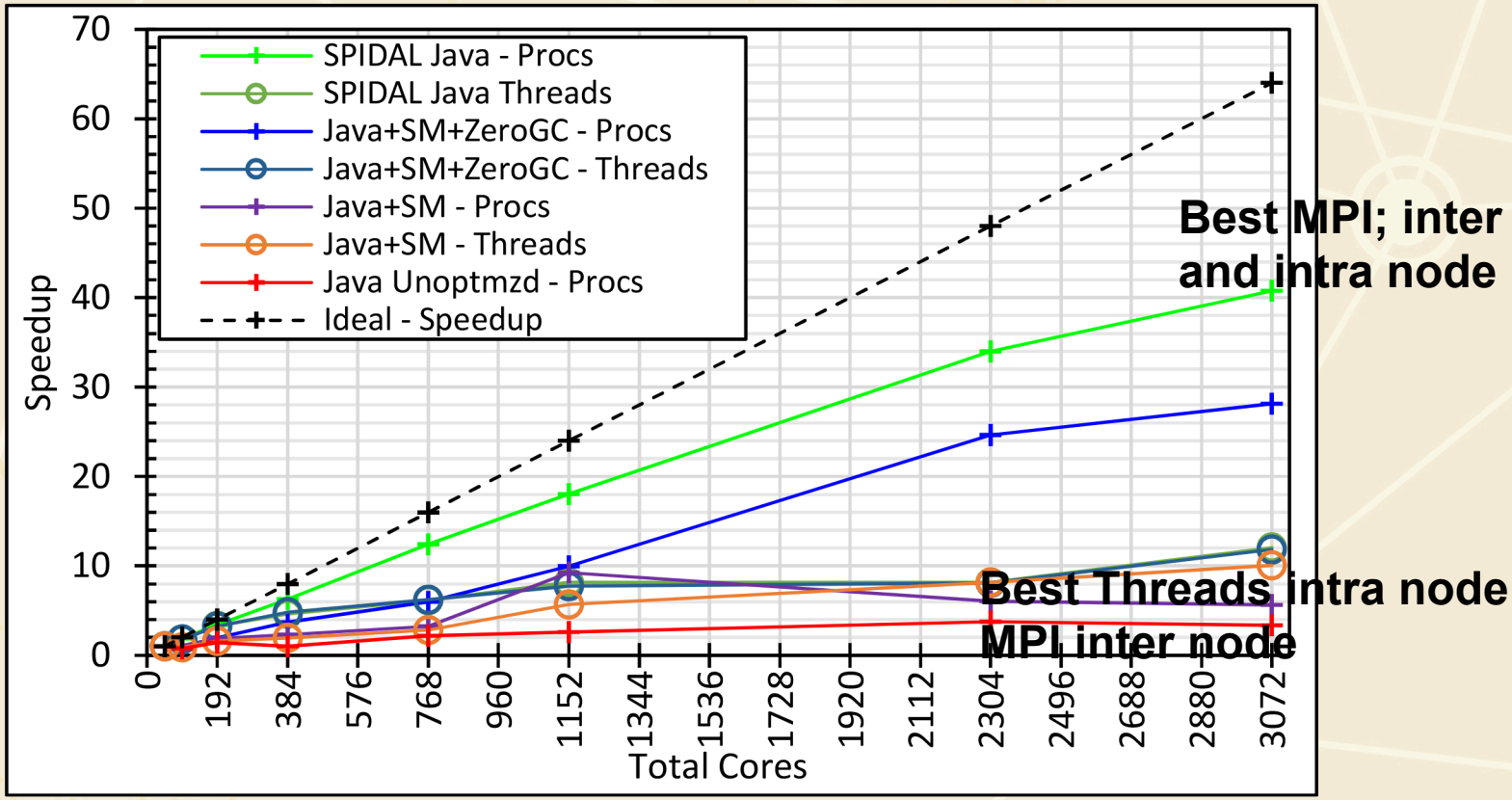
- **NIST Big Data Application Analysis:** features of data intensive apps
- **HPC-ABDS:** Cloud-HPC interoperable software with performance of HPC (High Performance Computing) and the rich functionality of the commodity Apache Big Data Stack.
  - Reservoir of software subsystems – nearly all from outside 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

64 Features in 4 views for Unified Classification of Big Data and Simulation Applications



Java MPI performs better than Threads

128 24 core Haswell nodes on SPIDAL DA-MDS Code



Kaleidoscope of (Apache) Big Data Stack (ABDS) and HPC Technologies

Cross-Cutting Functions	
1) **Message and Data Protocols:** Avro, Thrift, Protobuf	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, KeystoneML
2) **Distributed Coordination :** Google Chubby, Zookeeper, Giraffe, JGroups	16) **Application and Analytics:** Mahout , MLlib , MLbase, DataFu, R, pbdR, Bioconductor, ImageJ, OpenCV, Scalapack, PetSc, PLASMA MAGMA, Azure Machine Learning, Google Prediction API & Translation API, mlpy, 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) **Security & Privacy:** InCommon, Eduroam OpenStack Keystone, LDAP, Sentry, Sqrrl, OpenID, SAML OAuth	15B) **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 15A) **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
4) **Monitoring:** Ambari, Ganglia, Nagios, Inca	14B) **Streams:** Storm, S4, Samza, Granules, Neptune, Google MillWheel, Amazon Kinesis, LinkedIn, Twitter Heron, Databus, Facebook Puma/Ptail/Scribe/ODS, Azure Stream Analytics, Floe, Spark Streaming, Flink Streaming, DataTurbine 14A) **Basic Programming model and runtime**, **SPMD, MapReduce:** Hadoop, Spark, Twister, MR-MPI, Stratosphere (Apache Flink), Reef, Disco, Hama, Giraph, Pregel, Pegasus, Ligra, GraphChi, Galois, Medusa-GPU, MapGraph, Totem
	13) **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
	12) **In-memory databases/caches:** Gora (general object from NoSQL), Memcached, Redis, LMDB (key value), Hazelcast, Ehcache, Infinispan, VoltDB, H-Store
	12) **Object-relational mapping:** Hibernate, OpenJPA, EclipseLink, DataNucleus, ODBC/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, N1QL, BlinkDB, Spark SQL
	11B) **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 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
	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, OpenPBS, Moab, Slurm, Torque, Globus Tools, Pilot Jobs
	8) **File systems:** HDFS, Swift, Haystack, f4, Cinder, Ceph, FUSE, Gluster, Lustre, GPFS, GFFS **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, 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

21 layers
Over 350 Software Packages

January 29 2016

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