The importance of advanced analytics to derive insight and knowledge from increasing volumes of complex data will continue to grow. The enterprise community have made impressive gains and seem to have converged around the Apache stack, a distinctive feature is the existence of many implementations of the specific components of the Apache stack, providing sufficient richness in the trade-off between performance and capability. In contrast, within the scientific computing community, progress has been reliant either on long-term foundational advances or short-term hardware fixes. as opposed to integrated approaches that marry the relative technical strengths of the two communities yet deliver these as implementations usable on high performance and distributed computing HPDC infrastructure such as XSEDE, OSG and other domain-specific infrastructure. In both domains, scalable yet general-purpose and broadly applicable solutions in the form of analytic libraries and abstractions are noticeable by their absence.

To remedy this major gap and proffer an integrated solution that brings the best of recent advances to the service of extreme-scale science requirements on current and future science production platforms, we are developing HPC-ABDS – a first implementation of a high-performance Big Data stack (HPBDS) that integrates the best of the Apache developments and HPC capabilities. HPC-ABDS will utilize and expose the integrated relative technical strengths of the two hitherto disjoint approaches and communities, yet it will focus on delivering these as production grade implementations that will bring the best-of-both to shared-infrastructure – such as NSF's XSEDE, DOE's leadership machine, OSG and other domain-specific infrastructure, as well as the software developments underway as part of the SI2 software program. HPC-ABDS will translate these applications characteristics, infrastructural requirements and existing capabilities into well-defined and implemented building blocks.

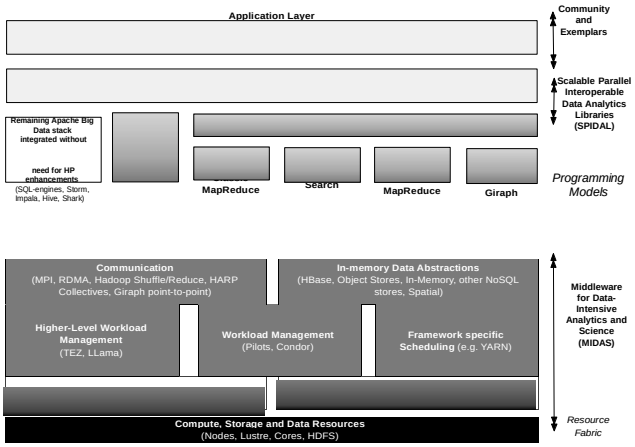

Fig. 1. Key components of integrated HPBDS stack. Many capabilities unaffected by integration are not shown explicitly

The key components of such an integrated platform are shown in Fig. 1. The aim of HPC-ABDS is to aim for the performance of HPC and the breadth and productivity of ABDS. The resultant integrated architecture is targeted at both production high-end computing platforms (such as leadership machines and XSEDE), as well as (commercial) cloud computing. As part of HPC-ABDS, we propose two fundamental building blocks, Middleware for Data-Intensive Analytics and Science (MIDAS) and the Scalable Parallel Interoperable Data Analytics Library (SPIDAL).

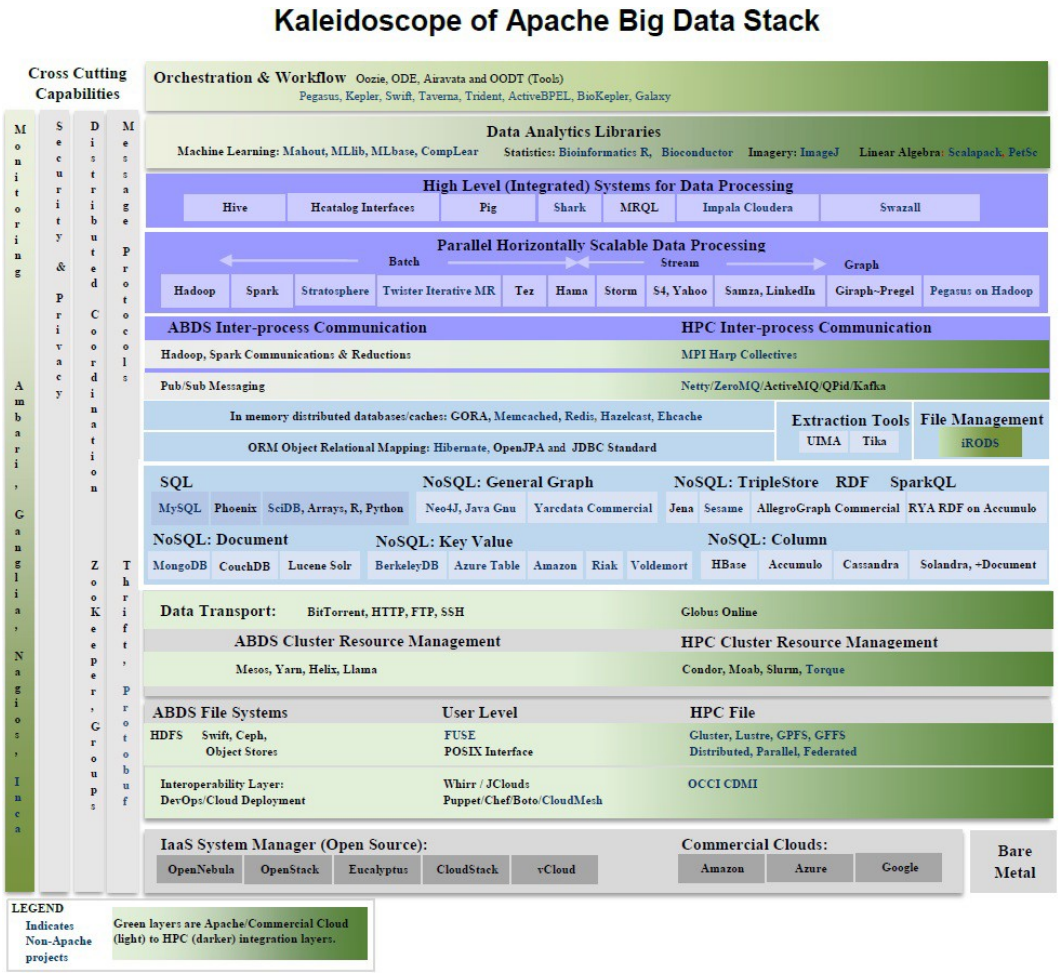

Fig. 5. For updated figure see: http://hpc-abds.org/kaleidoscope/

7. APPENDIX

REFERENCES

2012. CRAN Task View: Cluster Analysis and Finite Mixture Models. (July 21 2012). http://cran.cnr.berkeley.edu/web/views/Cluster.html

2013a. ArcGIS. (2013). http://www.esri.com/software/arcgis

2013b. Database Partitioning,Table Partitioning, and MDC for DB2 9. (2013). http://www.redbooks.ibm.com/redbooks/pdfs/sg247467.pdf

2013c. Delivering Location Intelligence with Spatial Data. (2013). http://www.microsoft.com/sql/techinfo/whitepapers/spatialdata.mspx

2013d. Greenplum. (2013). http://www.greenplum.com/products/greenplum-database

2013e. Hadoop-GIS Wiki. (2013). https://web.cci.emory.edu/confluence/display/HadoopGIS

2013f. IBM DB2 Spatial. (2013). http://www-01.ibm.com/software/data/spatial/

2013g. IBM Netezza. (2013). http://www-01.ibm.com/software/data/netezza/

2013h. Oracle Spatial and Oracle Locator. (2013). http://www.oracle.com/us/products/database/options/spatial/overview/index.html
2013i. PostGIS. (2013). http://postgis.refractions.net/
2013j. Teradata. (2013). http://www.teradata.com/
Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. 2009. Building Rome in a day. (Sept. 29-Oct. 2 2009). DOI:http://dx.doi.org/10.1109/ICCV.2009.5459148
Ablimit Aji, Fusheng Wang, Hoang Vo, Rubao Lee, Qiaoling Liu, Xiaodong Zhang, and Joel Saltz. 2013. Hadoop-GIS: A High Performance Spatial Data Warehousing System Over MapReduce. *Proc. VLDB Endow.* 6, 11 (2013), 1009–1020.
Afsin Akdogan, Ugur Demiryurek, Farnoush Banaei-Kashani, and Cyrus Shahabi. 2010. Voronoi-Based Geospatial Query Processing with MapReduce. (2010).
Maksudul Alam, Maleq Khan, and Madhav V. Marathe. 2013. Distributed-memory parallel algorithms for generating massive scale-free networks using preferential attachment model. (2013). DOI:http://dx.doi.org/10.1145/2503210.2503291
N. Alon, P. Dao, I. Hajirasouliha, F. Hormozdiari, and S. Sahinalp. 2008. Biomolecular network motif counting and discovery by color coding. *Bioinformatics* 24, 13 (2008), i241.
Apache Mahout. 2012. Apache Mahout Scalable machine learning and data mining. http://mahout.apache.org/. (2012).
V. Aravind and V. Raman. 2002. Approximate counting of small subgraphs of bounded treewidth and related problems. (2002).
Shaikh Arifuzzaman, Maleq Khan, and Madhav Marathe. 2013. PATRIC: a parallel algorithm for counting triangles in massive networks. (2013). DOI:http://dx.doi.org/10.1145/2505515.2505545
D.A. Bader. 2010. Analyzing Massive Social Networks using Multicore and Multithreaded Architectures. *Facing the Multicore-Challenge: Aspects of New Paradigms and Technologies in Parallel Computing, Lecture Notes in Computer Science* 6310, 1 (2010).
D.A. Bader and G. Cong. 2004. Fast Shared-Memory Algorithms for Computing the Minimum Spanning Forest of Sparse Graphs. (April 26-30 2004).
D.A. Bader and K. Madduri. 2008. A Graph-Theoretic Analysis of the Human Protein-Interaction Network Using Multi-core Parallel Algorithms. *Parallel Comput.* 34, 11 (2008), 627–639.
David A. Bader and Guojing Cong. 2005. A fast, parallel spanning tree algorithm for symmetric multiprocessors (SMPs). *J. Parallel Distrib. Comput.* 65, 9 (2005), 994–1006. DOI:http://dx.doi.org/10.1016/j.jpdc.2005.03.011
D. A. Bader and . J. JJ. 1996. Parallel Algorithms for Image Histogramming and Connected Components with an Experimental Study. *J. Parallel and Distrib. Comput.* 35, 2 (1996), 173–190.
Seung-Hee Bae. 2012. *SCALABLE HIGH PERFORMANCE MULTIDIMENSIONAL SCALING*. Thesis. http://grids.ucs.indiana.edu/ptliupages/publications/SeungheeBae_Dissertation.pdf
Jaime Ballesteros, Ariel Cary, and Naphtali Rishe. 2011. SpSJoin: Parallel Spatial Similarity Joins. (2011). DOI:http://dx.doi.org/10.1145/2093973.2094054
R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. Van der Vorst. 1994. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, 2nd Edition*. SIAM, Philadelphia, PA.
V. Batagelj and A. Mrvar. 1998. Pajek - Program for Large Network Analysis. *Connections* 21, 2 (1998), 47–57. http://vlado.fmf.uni-lj.si/pub/networks/doc/pajek.pdf

Ariel Cary, Zhengguo Sun, Vagelis Hristidis, and Naphtali Rishe. 2009. Experiences on Processing Spatial Data with MapReduce. (2009). DOI:http://dx.doi.org/10.1007/978-3-642-02279-1_24

Kang-Tsung Chang. 2003. *Introduction to Geographic Information Systems*. Prentice Hall PTR. 384 pages.

Wo Chang. 2014. ISO/IEC JTC 1 Study Group on Big Data. http://jtc1bigdatasg.nist.gov/flyer_BigDataEcosystem_Workshop_US.pdf. In *1st Big Data Interoperability Framework Workshop: Building Robust Big Data Ecosystem*, Vol. 2014. NIST.

Jong Youl Choi. 2012. *Unsupervised Learning Of Finite Mixture Models With Deterministic Annealing For Large-scale Data Analysis*. Thesis. http://grids.ucs.indiana.edu/ptliupages/publications/damix.final.v1.pdf

Jong Youl Choi, Mohammad H. Abbasi, David Pugmire, Scott Klasky, Judy Qiu, and Geoffrey Fox. 2012. Mining Hidden Mixture Context With ADIOS-P To Improve Predictive Pre-fetcher Accuracy. (October 8-12 2012). http://grids.ucs.indiana.edu/ptliupages/publications/hcming(1).pdf

Jong Youl Choi, Seung-Hee Bae, Judy Qiu, Bin Chen, and David Wild. 2011. Browsing Large Scale Cheminformatics Data with Dimension Reduction. *Concurr. Comput. : Pract. Exper.* Special Issue on ECMLS2010 (2011). http://grids.ucs.indiana.edu/ptliupages/publications/plotviz.v6.pdf

David Crandall, Andrew Owens, Noah Snavely, and Daniel P. Huttenlocher. 2011. Discrete-continuous optimization for large-scale structure from motion. (2011).

David Crandall and Noah Snavely. 2012. Modeling people and places with internet photo collections. *Commun. ACM* 55, 6 (2012), 52–60. DOI:http://dx.doi.org/10.1145/2184319.2184336

Navneet Dalal and Bill Triggs. 2005. Histograms of Oriented Gradients for Human Detection. (2005). DOI:http://dx.doi.org/10.1109/cvpr.2005.177

P. Daszak. 2000. Emerging Infectious Diseases of Wildlife– Threats to Biodiversity and Human Health. *Science* 287, 5452 (2000), 443–449. DOI:http://dx.doi.org/10.1126/science.287.5452.443

Rob F. Van der Wijngaart, Srinivas Sridharan, and Victor W. Lee. 2012. Extending the BT NAS parallel benchmark to exascale computing. (2012).

D. Ediger, K. Jiang, E.J. Riedy, and D.A. Bader. 2012. GraphCT: Multithreaded Algorithms for Massive Graph Analysis. *'IEEE Transactions on Parallel & Distributed Systems* (2012).

Jaliya Ekanayake, Thilina Gunarathne, Judy Qiu, Geoffrey Fox, Scott Beason, Jong Youl Choi, Yang Ruan, Seung-Hee Bae, and Hui Li. 2010a. *Applicability of DryadLINQ to Scientific Applications*. Report. Community Grids Laboratory, Indiana University.

J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S. Bae, J. Qiu, and G. Fox. 2010b. Twister: A Runtime for iterative MapReduce. (2010). http://grids.ucs.indiana.edu/ptliupages/publications/hpdc-camera-ready-submission.pdf

Ahmed Eldawy and Mohamed Mokbel. 2013. A Demonstration of SpatialHadoop: An Efficient MapReduce Framework for Spatial Data. (2013).

Christos Faloutsos. 2012. Project Pegasus Peta-Scale graph mining. (2012). http://www.cs.cmu.edu/~pegasus/

S. Fortunato. 2010. Community detection in graphs. *Physics Reports* 486, 3-5 (2010), 75–174.

Geoffrey Fox. 2013. Robust Scalable Visualized Clustering in Vector and non Vector Semimetric Spaces. *Parallel Processing Letters* 23, 2 (2013). DOI:http://dx.doi.org/doi/abs/10.1142/S0129626413400069

Geoffrey Fox, Seung-Hee Bae, Jaliya Ekanayake, Xiaohong Qiu, and Huapeng Yuan. 2009a. *Parallel Data Mining from Multicore to Cloudy Grids*.

IOS Press, Amsterdam. http://grids.ucs.indiana.edu/ptliupages/publications/CetraroWriteupJune11-09.pdf

Geoffrey Fox, Xiaohong Qiu, Scott Beason, Jong Youl Choi, Mina Rho, Haixu Tang, Neil Devadasan, and Gilbert Liu. 2009b. Biomedical Case Studies in Data Intensive Computing. (December 1-4 2009). http://grids.ucs.indiana.edu/ptliupages/publications/SALSACloudCompaperOct10-09.pdf

Judy Fox Fox, Shantenu Jha and Andre Luckow. 2014. Towards an Understanding of Facets and Exemplars of Big Data Applications. (October 13-14 2014).

Rudolf Fruhwirth, D. R. Mani, and Saumyadipta Pyne. 2011. Clustering with position-specific constraints on variance: Applying redescending M-estimators to label-free LC-MS data analysis. *BMC Bioinformatics* 12, 1 (2011), 358. http://www.biomedcentral.com/1471-2105/12/358

I. G. Goldberg, C. Allan, J. M. Burel, D. Creager, A. Falconi, H. Hochheiser, J. Johnston, J. Mellen, P. K. Sorger, and J. R. Swedlow. 2005. The Open Mi- croscopy Environment (OME) Data Model and XML file: open tools for informat- ics and quantitative analysis in biological imaging. *Genome Biol* 6, 5 (2005), R47. DOI:http://dx.doi.org/gb-2005-6-5-r47[pii];10.1186/gb-2005-6-5-r47

M. Gonen and Y. Shavitt. 2009. Approximating the number of network motifs. (2009).

O. Green, R. McColl, and D.A. Bader. 2012. A Fast Algorithm for Incremental Betweenness Centrality. (September 3-5 2012).

Danhuai Guo, Kaichao Wu, Jianhui Li, and Yuwei Wang. 2010. Spatial scene similarity assessment on Hadoop. (2010). DOI:http://dx.doi.org/10.1145/1869692.1869700

A. A. Hagberg, D. A. Schult, and P. J. Swart. 2008. Exploring network structure, dynamics, and function using NetworkX. (2008). http://conference.scipy.org/proceedings/scipy2008/paper_2/

M. Handcock, D. Hunter, C. Butts, S. Goodreau, and M. Morris. 2003. statnet: Software tools for the Statistical Modeling of Network Data, Version 2.0. (2003). http://statnetproject.org

Linda Hayden, Geoffrey Fox, and Prasad Gogineni. 2007. CYBERINFRASTRUCTURE FOR REMOTE SENSING OF ICE SHEETS. (June 4-8 2007). http://grids.ucs.indiana.edu/ptliupages/publications/TeraGrid07_paper.pdf

James Hays and Alexei A. Efros. 2007. Scene completion using millions of photographs. *ACM Trans. Graph.* 26, 3 (2007), 4. DOI:http://dx.doi.org/10.1145/1276377.1276382

James Hays and Alexei A. Efros. 2008. IM2GPS: Estimating Geographic Information from a Single Image. (2008).

Hortonworks. 2014. Apache Hadoop YARN is a sub-project of Hadoop introduced in Hadoop 2.0. (2014). http://hortonworks.com/hadoop/yarn/

Adam Hughes, Yang Ruan, Saliya Ekanayake, Seung-Hee Bae, Qunfeng Dong, Mina Rho, Judy Qiu, and Geoffrey Fox. 2012. Interpolative Multidimensional Scaling Techniques for the Identification of Clusters in Very Large Sequence Sets. *BMC Bioinformatics* 13(Suppl 2):S9, Special Issue of for Proceedings of GLBIO Great Lakes Bioinformatics Conference Ohio University Athens Ohio May 2-4 2011 (2012). DOI:http://dx.doi.org/10.1186/1471-2105-13-S2-S9

S. Jha, M. Cole, D. Katz, O. Rana, M. Parashar, and J. Weissman. 2013. Distributed Computing Practice for Large-Scale Science & Engineering Applications. *Concurrency and Computation: Practice and Experience* 25, 11 (2013), 1559–1585. DOI:http://dx.doi.org/10.1002/cpe.2897

Shantenu Jha, Neil Chue Hong, Simon Dobson, Daniel S. Katz, Andre Luckow, Omer Rana, and Yogesh Simmhan. 2014a. Introducing Distributed Dynamic Data-intensive (D3) Science: Understanding Applications and Infrastructure. (2014).

Shantenu Jha, Judy Qiu, Andre Luckow, Pradeep Mantha, and Geoffrey C. Fox. 2014b. A Tale of Two Data-Intensive Approaches: Applications, Architectures and Infrastructure. http://arxiv.org/abs/1403.1528. (June 27- July 2 2014).

K. Jiang, D. Ediger, and D.A. Bader. 2009. Generalizing k-Betweenness Centrality Using Short Paths and a Parallel Multithreaded Implementation. (September 22-25 2009).

U Kang, Charalampos Tsourakakis, and Christos Faloutsos. 2009. PEGASUS: A Peta-Scale Graph Mining System - Implementation and Observations. (December 2009).

U Kang, Charalampos E. Tsourakakis, Ana Paula Appel, Christos Faloutsos, and Jure Leskovec. 2011. HADI: Mining Radii of Large Graphs. *ACM Trans. Knowl. Discov. Data* 5, 2 (2011), 1–24. DOI:http://dx.doi.org/10.1145/1921632.1921634

Howard Karloff, Siddharth Suri, and Sergei Vassilvitskii. 2010. A model of computation for MapReduce. (2010).

Anthony J. Kearsley, Richard A. Tapia, and Michael W. Trosset. 1995. *The Solution of the Metric STRESS and SSTRESS Problems in Multidimensional Scaling Using Newtons Method*. Report. Rice University.

Hartmut Klauck, Danupon Nanongkai, Gopal Pandurangan, and Peter Robinson. 2013. *The Distributed Complexity of Large-scale Graph Processing*. Report. http://arxiv.org/abs/1311.6209

J. Leskovec. 2012. Stanford Network Analysis Project. (2012). http://snap.stanford.edu/

Wengen Li, Weili Wang, and Ting Jin. 2012. *Evaluating Spatial Keyword Queries under the MapReduce Framework Database Systems for Advanced Applications*. Lecture Notes in Computer Science, Vol. 7240. Springer Berlin / Heidelberg, 251–261. DOI:http://dx.doi.org/10.1007/978-3-642-29023-7_26

Yunpeng Li, David Crandall, and Daniel P. Huttenlocher. 2009. Landmark Classification in Large-scale Image Collections. (2009).

Yan Liu, Kaichao Wu, Shaowen Wang, Yanli Zhao, and Qian Huang. 2010. A MapReduce Approach to Gi*(d) Spatial Statistic. (2010). DOI:http://dx.doi.org/10.1145/1869692.1869695

David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision* 60, 2 (2004), 91–110. DOI:http://dx.doi.org/10.1023/b:visi.0000029664.99615.94

Andre Luckow, Mark Santcroos, and Shantenu Jha. 2014. Pilot-Data: An Abstraction for Distributed Data. *J. Parallel and Distrib. Comput.* In press (2014). http://arXiv.org/abs/1301.6228

Andre Luckow, Mark Santcroos, Ole Weidner, Andre Merzky, Pradeep Man- tha, and Shantenu Jha. 2012. P*: A Model of Pilot-Abstractions. (2012). DOI:http://dx.doi.org/10.1109/eScience.2012.6404423

Qiang Ma, Bin Yang, Weining Qian, and Aoying Zhou. 2009. Query Processing of Massive Trajectory Data Based on Mapreduce. (2009). DOI:http://dx.doi.org/10.1145/1651263.1651266

K. Madduri and D.A. Bader. 2009. Compact Graph Representations and Parallel Connectivity Algorithms for Massive Dynamic Network Analysis. (May 25-29 2009).

K. Madduri, D.A. Bader, J.W. Berry, and J.R. Crobak. 2007. An Experimental Study of A Parallel Shortest Path Algorithm for Solving Large-Scale Graph Instances. (January 6 2007).

Kamesh Madduri, David Ediger, Karl Jiang, David A. Bader, and Daniel Chavarria-Miranda. 2009. A faster parallel algorithm and efficient multithreaded implementations for evaluating betweenness centrality on massive datasets. (2009). DOI:http://dx.doi.org/10.1109/ipdps.2009.5161100

M. E. Martone, J. Tran, W. W. Wong, J. Sargis, L. Fong, S. Larson, S. P. Lamont, A. Gupta, and M. H. Ellisman. 2008. The cell centered database project: an update on building community resources for managing and sharing 3D imaging data. *J Struct Biol* 161, 3 (2008), 220–31. DOI:http://dx.doi.org/S1047-8477(07)00254-7[pii];10.1016/j.jsb.2007.10.003

M. E. Martone, S. Zhang, A. Gupta, X. Qian, H. He, D. L. Price, M. Wong, S. Santini, and M. H. Ellisman. 2003. The cell-centered database: a database for multiscale structural and protein localization data from light and electron microscopy. 1, 4 (2003), 379–95. DOI:http://dx.doi.org/NI:1:4:379[pii];10.1385/NI:1:4:379

Manish Mehta and David J. DeWitt. 1995. Managing Intra-operator Parallelism in Parallel Database Systems. (1995).

Henning Meyerhenke, David Ediger, and David A. Bader. 2011. Parallel Community Detection for Massive Graphs. (September 2011).

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. Network motifs: simple building blocks of complex networks. *Science* 298, 5594 (2002), 824.

MLLib. 2014. Machine Learning Library (MLlib). http://spark.apache.org/docs/0.9.0/mllib-guide.html. (2014).

A. Lumsdaine N. Edmonds, T. Hoefler. 2010. A space-efficient parallel algorithm for computing betweenness centrality in distributed memory. (2010).

Donald Nguyen, Andrew Lenharth, and Keshav Pingali. 2013. A lightweight infrastructure for graph analytics. (2013). DOI:http://dx.doi.org/10.1145/2517349.2522739

NIST. 2013a. Big Data Initiative Reports from V1. http://bigdatawg.nist.gov/V1_output_docs.php. (2013).

NIST. 2013b. NIST Big Data Public Working Group (NBD-PWG) Use Cases and Requirements. http://bigdatawg.nist.gov/usecases.php, (2013).

Jignesh Patel, JieBing Yu, Navin Kabra, Kristin Tufte, Biswadeep Nag, Josef Burger, Nancy Hall, Karthikeyan Ramasamy, Roger Lueder, Curt Ellmann, Jim Kupsch, Shelly Guo, Johan Larson, David De Witt, and Jeffrey Naughton. 1997. Building A Scaleable Geo-Spatial DBMS: Technology, Implementation and Evaluation. *SIGMOD Rec.* 26, 2 (1997), 336–347. DOI:http://dx.doi.org/10.1145/253262.253342

A. Pavlo, E. Paulson, A. Rasin, D. J. Abadi, D. J. DeWitt, S. Madden, and M. Stonebraker. 2009. A Comparison of Approaches to Large-Scale Data Analysis. (2009).

Alexei Pozdnoukhov and Christian Kaiser. 2011. Scalable Local Regression for Spatial Analytics. (2011). DOI:http://dx.doi.org/10.1145/2093973.2094023

Dimitrios Prountzos and Keshav Pingali. 2013. Betweenness centrality: algorithms and implementations. *SIGPLAN Not.* 48, 8 (2013), 35–46. DOI:http://dx.doi.org/10.1145/2517327.2442521

G. Qin and L. Gao. 2012. An algorithm for network motif discovery in biological networks. *International Journal of Data Mining and Bioinformatics* 6, 1 (2012), 1–16.

Judy Qiu, Jaliya Ekanayake, Thilina Gunarathne, Jong Youl Choi, Seung-Hee Bae, Hui Li, Bingjing Zhang, Tak-Lon Wu, Yang Ryan, Saliya Ekanayake, Adam Hughes, and Geoffrey Fox. 2010. Hybrid cloud and cluster computing paradigms for life science applications. *BMC Bioinformatics* Proceedings of BOSC 2010 (2010). http://grids.ucs.indiana.edu/ptliupages/publications/HybridCloudandClusterComputingParadigmsforLifeScienceApplications_Pub.pdf

Judy Qiu, Jaliya Ekanayake, Thilina Gunarathne, Jong Youl Choi, Seung-Hee Bae, Yang Ruan, Saliya Ekanayake, Stephen Wu, Scott Beason, Geoffrey Fox, Mina Rho, and Haixu Tang. 2011. *Data Intensive Computing for Bioinformatics*. IGI Publishers. DOI:http://dx.doi.org/10.4018/978-1-6152971-2

Judy Qiu, Thilina Gunarathne, Jaliya Ekanayake, Jong Youl Choi, Seung-Hee Bae, Hui Li, Bingjing Zhang, Yang Ryan, Saliya Ekanayake, Tak-

Lon Wu, Scott Beason, Adam Hughes, and Geoffrey Fox. 2010. Hybrid Cloud and Cluster Computing Paradigms for Life Science Applications. (July 9-10 2010). http://grids.ucs.indiana.edu/ptliupages/publications/HybridCloudandClusterComputingParadigmsforLifeScienceApplications.pdf

Judy Qiu and Bingjing Zhang. 2013. *Clustering Social Images with MapReduce and High Performance Collective Communication*. IOS Press. http://grids.ucs.indiana.edu/ptliupages/publications/MammothDataintheCloudClusteringSocialImage.pdf

Judy Qiu and Bingjing Zhang. 2014. Harp: a runtime for efficient in-memory communication. (2014). http://salsaproj.indiana.edu/harp/

R Project. 2012. R open source statistical library. http://www.r-project.org/. (2012).

P. Ribeiro, F. Silva, and L. Lopes. 2012. Parallel discovery of network motifs. *J. Parallel and Distrib. Comput.* 72, 2 (2012), 144–154.

Ken Rose. 1998. Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems. *Proc. IEEE* 86 (1998), 2210–2239.

Ken Rose, Eitan Gurewitz, and Geoffrey Fox. 1990. A deterministic annealing approach to clustering. *Pattern Recogn. Lett.* 11 (1990), 589–594.

Yang Ruan, Saliya Ekanayake, Mina Rho, Haixu Tang, Seung-Hee Bae, Judy Qiu, and Geoffrey Fox. 2012a. DACIDR: Deterministic Annealed Clustering with Interpolative Dimension Reduction using Large Collection of 16S rRNA Sequences. (October 7-10 2012). http://grids.ucs.indiana.edu/ptliupages/publications/DACIDR_camera_ready_v0.3.pdf

Yang Ruan and Geoffrey Fox. 2013. A Robust and Scalable Solution for Interpolative Multidimensional Scaling with Weighting. (October 22-25 2013). DOI:http://dx.doi.org/10.1109/eScience.2013.30

Yang Ruan, Zhenhua Guo, Yuduo Zhou, Judy Qiu, and Geoffrey Fox. 2012b. HyMR: a Hybrid MapReduce Workflow System. (June 18 2012). http://grids.ucs.indiana.edu/ptliupages/publications/HyMR_submission_HPDC_workshop_final.pdf

Yang Ruan, Geoffrey L. House, Saliya Ekanayake, Ursel Schtte, James D. Bever, Haixu Tang, and Geoffrey Fox. 2014. Integration of Clustering and Multidimensional Scaling to Determine Phylogenetic Trees as Spherical Phylograms Visualized in 3 Dimensions. (May 26-29 2014). http://grids.ucs.indiana.edu/ptliupages/publications/PhylogeneticTreeDisplayWithClustering.pdf

S. Sarkar and A. Dong. 2011. Community detection in graphs using singular value decomposition. *Physical Review E* 83, 4 (2011), 04611.

V. Satuluri and S. Parthasarathy. 2009. Scalable graph clustering using stochastic ows: applications to community discovery. (2009).

Venu Satuluri, Srinivasan Parthasarathy, and Yiye Ruan. 2011. Local graph sparsification for scalable clustering. (2011). DOI:http://dx.doi.org/10.1145/1989323.1989399

D. A. Schiffmann, D. Dikovskaya, P. L. Appleton, I. P. Newton, D. A. Creager, C. Allan, I. S. Nathke, and I. G. Goldberg. 2006. Open microscopy environment and findspots: integrating image informatics with quantitative multidimensional image analysis. *Biotechniques* 41, 2 (2006), 199–208. DOI:http://dx.doi.org/000112224[pii]

C. Seshadhri, T. Kolda, and A. Pinar. 2012a. Community structure and scale-free collections of Erdos–Renyi graphs. *Physical Review E* (2012). http://arxiv.org/abs/1112.3644

C. Seshadhri, A. Pinar, and T. Kolda. 2012b. Fast Triangle Counting through Wedge Sampling. (2012). http://arxiv.org/abs/1202.5230

Larissa Stanberry, Roger Higdon, Winston Haynes, Natali Kolker, William Broomall, Saliya Ekanayake, Yang Ruan, Judy Qiu, Eugene Kolker, Geoffrey Fox, and Adam Hughes. 2012. Visualizing the Protein Sequence Universe. (June 18 2012). http://grids.ucs.indiana.edu/ptliupages/publications/paperDelft_final.pdf

Antonio Torralba, Kevin P Murphy, William T Freeman, and Mark A Rubin. Context-based vision system for place and object recognition. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 273–280.

Charalampos E. Tsourakakis, U Kang, Gary L. Miller, and Christos Faloutsos. 2009. DOULION: Counting triangles in massive graphs with coin. (2009).

Stanford University. 2014. ImageNet image database organized according to the WordNet hierarchy. (2014). http://www.image-net.org/

Fusheng Wang, Jun Kong, Jingjing Gao, Lee A.D. Cooper, Tahsin Kurc, Zhengwen Zhou, David Adler, Cristobal Vergara-Niedermayr, Bryan Katigbak, Daniel J Brat, and Joel H Saltz. 2013. A high-performance spatial database based approach for pathology imaging algorithm evaluation. *J Pathol Inform.* 4, 5 (2013).

FWang, TKurc, PWidener, TPan, JKong, L Cooper, D Gutman, A Sharma, S Cholleti, V Kumar, and J Saltz. 2010. High-performance Systems for In Silico Microscopy Imaging Studies. *The 7th International Conference on Data Integration in the Life Sciences, Gothenburg, Sweden* (2010).

Kaibo Wang, Yin Huai, Rubao Lee, Fusheng Wang, Xiaodong Zhang, and Joel Saltz. 2012. Accelerating Pathology Image Data Cross Comparison on CPU-GPU Hybrid Systems. *Proc. VLDB Endow.* 5, 11 (2012), 1543–1554.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 3485–3492.

Shubin Zhang, Jizhong Han, Zhiyong Liu, Kai Wang, and Shengzhong Feng. Spatial Queries Evaluation with MapReduce. In *Eighth International Conference on Grid and Cooperative Computing*. 287–292. DOI:http://dx.doi.org/10.1109/gcc.2009.16

Shubin Zhang, Jizhong Han, Zhiyong Liu, Kai Wang, and Zhiyong Xu. SJMR: Parallelizing Spatial Join with MapReduce on Clusters. In *IEEE International Conference on Cluster Computing*. 1–8. DOI:http://dx.doi.org/10.1109/clustr.2009.5289178

Y. Zhang, Z. Wang, Y. Wang, and L. Zhou. 2009. Parallel community detection on large networks with propinquity dynamics. (2009).

Zhao Zhao, Maleq Khan, V. S. Anil Kumar, and Madhav V. Marathe. 2010. Subgraph Enumeration in Large Social Contact Networks Using Parallel Color Coding and Streaming. (2010). DOI:http://dx.doi.org/10.1109/icpp.2010.67

Z. Zhao, G. Wang, A. Butt, M. Khan, V. S. Anil Kumar, and M. Marathe. 2012. SAHAD: Subgraph Analysis in Massive Networks Using Hadoop. (May 2012).