Dominique A. Heger Ph.D.
Email: dheger@dhtusa.com, dheger@mlanalytica.com
www.dhtusa.com, www.mlanalytica.com, www.hotshotanalytics.com

Professional Summary

- *Dominique Heger has over 30 years of professional experience in systems performance, capacity planning, UNIX internals, algorithms and data structures, systems design and architecture, reliability and availability, operations research, compilers, software engineering, machine learning, robotics, artificial neural networks (ANN), fuzzy logic, data analytics/mining, petri nets, and communication technologies.*
- *He has designed several systems stability related methodologies to analyze and quantify hardware and software performance/capacity and availability/reliability related issues. Over the years, he has conducted several large-scale systems performance and capacity planning projects in a wide range of professional environments such as science & research, aerospace, manufacturing, banking, insurance, oil & gas, telecom, and the chemical industry.*
- *Specific areas of expertise include IBM, HP, Oracle/Sun, Dell, and SGI SMP and MPP/HPC systems architecture, cluster (Big Data & Hadoop), GRID, and Cloud Computing, UNIX/Linux operating systems, I/O subsystem performance, algorithms and data structures, performance modeling and simulation, operations research, network and cluster technologies, telecommunication (VoIP), and all phases of the software development life cycle.*
- *Throughout his career, he has taught UNIX/Linux internals and performance management classes for DHT, IBM, Hewlett-Packard, and Unisys, respectively. In addition, he is now also teaching Machine Learning, AI, and Big Data & Predictive Analytics classes for Data Analytica.*

Technical Summary

Hardware:	IBM, HP, Oracle/Sun, Hitachi, and Dell SMP, cluster, and SAN Subsystems
Operating Systems:	AIX, HP-UX, Solaris, Linux, Android, IRIX, and Windows
File Systems:	GPFS, Lustre, GFS, HDFS, XFS, JFS/J2, UFS, VxFS, Btrfs, ZFS, NFS
Languages/Compilers:	C (MPI), Pascal, Fortran, ADA, C++, Java, Python, Scala, Clojure, Lisp, and Prolog
Software:	Oracle (RAC), DB2, MathCad, Matlab, Octave, R, Petri-Nets, ANN, MPC, NoSQL
Big Data & Hadoop:	MapReduce, Storm, HBase, Hive, Pig, Cassandra, MongoDB, CouchDB, Sqoop, Neo4j, Mahout, Spark, Hama, Impala, Flume, Avro, Kafka, S4, Tez, SAMOA, H2O

Professional Experience

DHTechnologies (DHT) (Since 2005), Owner/CEO
Data Analytica (Since 2011), Owner/CEO
HotShot Sports Analytics (Since 2017), Part-Owner/CTO

Conduct comprehensive Big Data (including Deep Learning - TensorFlow, Keras, Theano), data analytics (descriptive, predictive, and prescriptive) and data mining projects, focusing on the actual (machine learning) models being used in these studies. Models based on artificial neural networks, inductive learning, or ensemble Kalman filter techniques have been successfully designed,

implemented, tested, and utilized for estimation and forecasting related projects (some of these projects specifically target the Hadoop/HDFS/Tez IT Infrastructures while others focus on real-time predictive analysis via Kafka/Storm/ Spark/S4/H2O/Cassandra components). Data filtering via singular value decomposition, principal component analysis, SOM, or k-means algorithms have been applied to streamline the data sets. Some of the robotics projects focus on self driving objects, image recognition, and robot movement.

Conduct comprehensive robotics studies focusing on self driven objects (land and water) and (real-time) SLAM algorithms. Work on improving the efficiency and effectiveness of moving robots in extreme working conditions. Develop deep learning software used for warehouse security (land robots) as well as inspections and damage assessments (drones).

Conduct and successfully complete comprehensive Big Data & Cloud Computing studies. The studies focus on design, proof of concept, performance, security, availability, reliability, data storage (distributed file systems - HDFS), scalability, as well as the business aspects of Big Data & Cloud projects. Designed and implemented a Hadoop MapReduce model (YARN) to quantify task execution time and cluster setup optimization under various workload conditions. Designed and implemented a real-time predictive analytics model based on Storm and Mahout. Designed and implemented comprehensive CEP environments based on Esper/Drools/Storm. Other modeling projects are using ANN (traditional and Deep Learning), MPC, and/or Fuzzy Logic based methods and some focus on In-Memory Computing to optimize the performance behavior.

Conduct comprehensive performance modeling, capacity planning, and scalability studies in the area of large SMP, cluster (HPC as well as Hadoop & Big Data), Cloud, N-Tier, and Grid environments, respectively. Transform complex IT processes into state-transition diagrams, analytical, neural network, model predictive control, and petri net based models, and hence mathematically abstract the environment to conduct performance, capacity, reliability, and availability related sensitivity studies. One strategic area of expertise in DHT revolves around scalable IO performance, focusing on local, distributed, as well as cluster file systems in the commercial, as well as the scientific arena. Some studies incorporate large SAN subsystems from Hitachi, IBM, EMC, Dell, or Oracle.

Conduct large-scale database migration projects (Oracle, DB2) from UNIX based cluster systems (Solaris, AIX, HP-UX) to Linux powered cluster environments. The projects utilize sophisticated mathematical abstractions to design, size, and optimize the new environments. Next to a baseline analysis, the models are used to quantify the performance behavior (headroom) under increased workload conditions, as well as the scalability potential of the design.

Research new methodologies to enhance the scalability of cluster (HPC & Big Data, including Hadoop), Cloud, and GRID systems. Conduct studies in Operating Systems performance, Scalable IO performance, and Operations Research. Teach Big Data & Predictive Analytics, UNIX Internals & Architecture, Cluster and Grid Technology, and Performance Tuning and Capacity Planning classes.

Conduct and successfully complete vulnerability, reliability, maintainability, and availability projects in heterogeneous IT environments. Outline and quantify the potential issues and recommend ways to improve the security and stability behavior of the IT environment.

Design and develop the 1st OS centric Android performance and stress-testing benchmark suite. The benchmark suite (for smart phones and tablets) focuses on the CPU, memory, IO, as well as the network subcomponents of Android.

IBM Corporation, (1996 - 2005), IBM Certified Systems Performance IT Specialist

Conducted a comprehensive, large-scale GPFS I/O performance analysis, focusing on metadata and large sequential I/O operations. The study incorporated evaluating I/O design choices, tuning aspects, identifying and resolving performance issues, as well as conducting empirical studies on the actual ASC Purple and Blue Gene hardware. The ASC Purple system, a 1,536-node AIX HPC supercomputer was successfully delivered to the Lawrence Livermore National Laboratory (LLNL) in Q4 2005.

Designed an entirely new read-ahead methodology for Linux 2.6. Designed new hash functions for the directory (dcache) and the inode cache (icache) subsystems for Linux 2.6. Designed and implemented a treap data structure component for the new Linux dcache design. Designed and implemented a threaded red-black data structure for the Linux 2.6 VMA subsystems. Designed a new Logistic Map based random number generator and an optimized red-black tree structure for the IBM flexible file system (FFSB) application I/O benchmark tool that I designed in 2000.

Designed and implemented a methodology to quantify the capacity driven asymptotic lower bound of disk subsystems. Designed and implemented a methodology to evaluate application driven I/O performance in a RAID environment. Both performance models are being used by the IBM LTC to substantiate the results of empirical studies.

Designed a methodology to quantify infrastructure-oriented systems stability that allows determining the Quality of Service provided to the user community, and enables identifying the dimension that reveals the greatest potential for improvements. The methodology focuses on the interrelationships among systems dependability, performability (scalability), and maintainability. Designed a methodology to quantify the performance characteristics of Web-based server systems. Designed and implemented analytical performance models to simulate a Web based server environment. The methodology and the models were utilized at the AT&T account to evaluate different design alternatives in a large-scale 3-Tier Web Application projects (Six Sigma – DMADV based).

Designed a *Factorized Scaleup Model* that allows evaluating the throughput scalability of scientific parallel applications. The model has been used as a performance evaluation and capacity planning tool to analyze and quantify the scaleup on SMP as well as MPP systems. As an artifact of the study, I designed a methodology to evaluate the CPU communication overhead and its impact on SMP capacity.

Researched and solved complex I/O performance limitations in GPFS, NSD, VSD, IP, KLAPI, and KHAL on IBM AIX supercomputers as well as on Linux based cluster systems (optimized Linux device drivers for GPFS/FastT environment). The conducted research resulted in higher scalability of GPFS, improved throughput, faster parallel inserts, and a much more stable production environment. Conducted performance studies for ASCI White, evaluating the efficiency of parallel inserts in GPFS.

Designed and implemented a methodology to evaluate workload dependent systems performance for kernel intensive applications (the project was focused on Linux on Intel and PowerPC architecture). The methodology allows studying systems performance in the context of performance paths that include the application, the operating system, as well as the hardware.

Researched performance related issues on UNIX SMP and MPP server systems. Defined methodologies to analyze performance as well as availability and reliability issues on parallel server systems. Researched and evaluated new operating system functionality's and features in AIX, HP-UX, Linux, and Solaris. Analyzed and evaluated their performance and made application specific and workload dependent tuning recommendations.

Evaluated HP, IBM, and Sun server systems, architected high-performance UNIX server solutions for IBM customers. Managed and conducted Systems Performance and Capacity Planning projects (modeling, simulation, and measurement based) in heterogeneous UNIX server environments. Analyzed systems and application benchmarks on UNIX server systems. Implemented kernel programs to evaluate and monitor systems performance. Taught UNIX Internals and Performance Management classes for IBM. Established and implemented operational processes for Performance Management and Capacity Planning.

Hewlett – Packard, (1989 – 1994), *UNIX Systems Engineer*

Conducted UNIX and network related projects on HP, Sun, IBM, SGI and TI server systems. Major responsibilities included tuning and performance analysis on the kernel level, as well as analyzing systems core dumps. Troubleshooting on file system and memory management level. C source code and compiler support (C, C++, Fortran, Pascal, ADA). Conducted performance evaluations and capacity planning projects on SMP and MPP UNIX server systems at CERN in Geneva (on the CERN campus

from 1990 – 1993, where I worked on a 64-CPU HP MPP 'snake farm'). Conducted strategic information systems planning and network design projects for CERN.

Project tasks for the RCO Computer Centers in Europe and the US included developing several client-server applications (such as a Remote Systems Performance and Network monitor). Worked on C-compiler and HP-UX source code in Cupertino, CA, USA.

Taught UNIX Internals and Performance Management classes for HP.

UNISYS, (1986 – 1989), *UNIX Systems Analyst*

Developed CAD/CAM software on SUN-OS UNIX workstations. Provided UNIX and programming support for contract customers. Taught UNIX Internal classes.

US Patent

POU920040076US1 - Communication Resource Reservation System for Improved Message Passing

Miscellaneous

- Keynote Drexel University, Deep Learning, SLAM Robotics, KIE2017, Philadelphia, 2017
- Keynote at the Big Data Symposium (Deep Learning - KIE2016), Berlin, 2016
- Keynote at the International Conference on Big Data and Predictive Analytics, Minsk, 2015
- Keynote at the Big Data Symposium (KIE2014), Riga, 2014
- Keynote at the Int. Conference on Knowledge, Innovation, & Enterprise, London, UK, 2013
- Keynote at the Hadoop & Big Data Symposium, University of Greenwich, UK, 2013

- Performance Tuning for Linux Server (3 chapters), Prentice Hall, ISBN 013144753X, 2005
- Recipient of the best paper award at the CITSA04 Conference, Orlando, FL, 2004
- Mentored 2 IBM Extreme Blue research projects in 2000 and 2001, respectively
- Designed FFSB as part of the IBM Extreme Blue research project in 2000
- Selected by IBM to be sponsored for a Ph.D. program
- On the advisory board of the American Association of Big Data Professionals (since 2013)

Educational Summary

BSCS, Bern University of Applied Science, School of Engineering & Information Technology, Switzerland
MBA (MIS), Maryville University, St. Louis, MO, USA
Ph.D. (IS), Nova Southeastern University, Fort Lauderdale, FL, USA

Major Publications/Books (Since 2001)

1. *Boris Zibitsker, Alex Lupersolsky, Yuri Balasanov, Mouttayen Manivassakam and Dominique Heger, Dynamic Performance Management of Big Data Clusters, Proceedings of the CMG imPACT 2017 conference, New Orleans, November 2017*

2. *Boris Zibitsker, Alex Lupersolsky, Dominique Heger, Yuri Balasanov, Jianghui (Cherish) Wen, Minghao Bian and Zhiyin Shi, Benchmarking ML Algorithms and Libraries for Big Data Applications, Proceedings of the CMG imPACT 2017 conference, New Orleans, November 2017*

3. *Heger D., "Big Data Analytics - Missing or Messy Data, What Now?, Proceedings of the 3d International Conference on Big Data Advanced Analytics, Minsk, Belarus, 2017*

4. *Heger, D. "Machine Learning in the Realm of Big Data Analytics", Fundcraft Publication, ISBN 978-0-578-19095-2, March 2017*

5. *Heger, D. " Visualizing data captured by nmon in Good Time", ADMIN Network and Security Journal, Volume 34, August 2016*

6. *Heger, D., Ogunleye, J., "Big Data, the Cloud and Challenges of Operationalising Big Data Analytics", Journal of Current Studies in Comparative Education, Science, and Technologies, Volume 22, Number 2, pp. 427-435, December 2015*

7. *Heger, D., "Big Data & Predictive Analytics - Algorithms, Applications, and Cluster Systems", Fundcraft Publication, ISBN 978-1-61422-951-3, January 2015*

8. *Heger, D., "Big Data Analytics—Where to go from Here", International Journal of Developments in Big Data and Analytics, Volume 1 No. 1, 2014, pp. 42—58*

9. *Heger, D., "Workload Dependent Hadoop MapReduce Application Performance Modeling", Performance & Capacity Measure IT Journal #13, July 2013*

10. *Heger, D., "Hadoop Performance Tuning - A Pragmatic & Iterative Approach", Performance & Capacity CMG Journal of Computer Resource Management, March 2013*

11. *Heger, D., "Hadoop Design, Architecture & MapReduce Performance", CMG Journal of Computer Resource Management, December 2012*

12. *Heger, D., "Optimized Resource Allocation & Task Scheduling Challenges in Cloud Computing Environments", CMG Journal of Computer Resource Management, December 2012*

13. *Heger, D., "Data Mining - The Gaining Knowledge Progression", CMG MeasureIT Journal, August 2012*

14. *Heger, D., "Quo Vadis Cloud Computing - Issues & Opportunities in the Cloud", eBook, Fundcraft Publishing, April 2012*

15. *Heger, D., "SSD Write Performance – IOPS Confusion Due to Poor Benchmarking Techniques", CMG MeasureIT Journal, Issue #7, August 2011*

16. *Heger, D., "Mobile Devices – An Introduction to the Android Operating Environment – Design, Architecture, and Performance Implications", CMG Journal of Computer Resource Management, 2011*

17. *Heger D., Quinn, R. "Linux 2.6 IO Performance Analysis, Quantification, and Optimization", Proceeding of the International Conference for Performance and Capacity Management - CMG2010, Orlando, FL*

18. *Heger, D., "Quantifying IT Stability 2nd Edition - Grid, Cloud, Cluster, and SMP Systems", Fundcraft Publication, ISBN 978-0-578-05264-9, April 2010*

19. *Heger, D. "Workload Dependent Performance Evaluation of the Btrfs and ZFS Filesystems", Proceeding of the International Conference for Performance and Capacity Management - CMG2009, Dallas, TX*

20. *Heger, D. "Characterization of the Underlying Behavioral Model in a Polynomial Mapping Environment", CMG Journal, July 2009*

21. *Heger, D., "Quantifying IT Stability", iUniverse Publication, ISBN 978-1-4401-0697-2, December 2008*

22. *Heger, D. Carinhas, P., "Parallel File System Technologies in a Cluster and GRID environment", Proceeding of the International Conference for Performance and Capacity Management – CMG2007, San Diego, CA, December 2007*

23. *Heger, D. "Deterministic Stochastic Petri Net Based IO Subsystem Performance Quantification", CMG Journal, 2007*

24. *Heger, D. Carinhas, P., Simco G., "GRID Technology – Vision, Architecture, and Node Capacity Considerations", Proceeding of the International Conference for Performance and Capacity Management – CMG2006, Reno, NV, December 2006*

25. *Heger, D. Carinhas, P., "A Cohesive Framework to Quantify Computer Systems Assurance", Proceeding of the International Conference for Performance and Capacity Management – CMG2006, Reno, NV, December 2006*

26. *Heger, D. "Quantitative Disk IO Performance – A Mathematical Abstraction & Analysis", CMG Journal, May 2006*

27. *Heger, D., Rao S., Pratt, S., "Examining the Linux 2.6 Page-Cache Performance", Proceeding of the 2005 Linux Symposium (OLS), July 2005*

28. *Heger, D., Tankeh, A., "The Design of a Dynamic Zero-Copy Communication Model for Cluster Based Systems", The European Journal for the Informatics Professional – June 2005*

29. *Heger D., Simco G., "Quantifying the Cluster Speedup Behavior in the Realm of Internode Communication", Proceeding of the IEEE Southeast Conference, Fort Lauderdale. April 2005*

30. *Heger D., "A Discourse on the Design and Analysis of Data Algorithms", CMG Journal of Computer Resource Management,, Fall Issue 2004*

31. *Heger, D., "A Disquisition on the Performance Behavior of Binary Search Tree Data Structures", Mosaic – Journal on Software Process Technology, Upgrade Volume V, Issue Nr. 5, October 2004*

32. *Heger D., "Methodology to Quantify I/O Performance Based on Analytical Models", Proceeding of 10th International Conference on Information Systems Analysis and Synthesis, CITSA/IEEE 04, Orlando, 2004*

33. *Heger D., Pratt, S., "Workload Dependent Performance Evaluation of the Linux 2.6 I/O Schedulers", Proceeding of the 2004 Linux Symposium (OLS), Ottawa, 2004*

34. *Heger, D., Johnson, S., Anand, M., Peloquin, M., Sullivan, M., Theurer, M., Wong. P., "An Application Centric Performance Evaluation of the Linux 2.6 Operating System", IBM Red Book White Paper, Austin, TX, 2004*

35. *Heger, D., "Methodology to Quantify the Performance Characteristics of Web Based Server Systems", Proceeding of the 29th International Conference for the Resource Management and Performance Evaluation of Enterprise Computing Systems (CMG), Dallas, December 2003*

36. *Heger, D., "Modeling and Predicting Load-Dependent I/O Performance in a ZBR Environment", CMG Journal, Issue 111, Summer Edition, 2003*

37. *Heger, D., "A Workload Dependent Scalability Model for Scientific Parallel Applications", CMG Journal, Issue 109, Winter Edition, 2002*

38. *Heger, D., Simco. G., "The Interrelationship Among Speedup Models and Performance Measurements", Proceeding of the 17th International Conference on Computers and Their Applications (CATA-02), San Francisco, 2002*

39. *Heger, D., 'The Design of a Logarithmic File Data Allocation Algorithm for Extent Based File Systems', Ph.D. Thesis, UMI Publication Number 3039329, 2001*

40. *Heger, D., Shah, G., 'GPFS 1.4 – Architecture and Performance', IBM Performance – IBM Red Book White Paper, Poughkeepsie, NY, 2001*