Références

[1] Amazon cloud to break the 1 billion dollars barrier ? http://www.crn.com/news/cloud/231002515/amazon-cloud-to-break-the-1-billion-barrier.htm read the 29/03/2012.

[2] Appdomain trick. http://code.google.com/p/lokad-cloud/wiki/ExceptionHandling read the 26/04/2012.

[3] Azure pricing. http://www.microsoft.com/windowsazure/pricing/.

[4] Azure pricing calculator. http://www.windowsazure.com/en-us/pricing/calculator/advanced/ read the 25/04/2012.

[5] Azure pricing details. http://www.windowsazure.com/en-us/pricing/details/ read the 25/04/2012.

[6] Azure scope. http://azurescope.cloudapp.net/.

[7] Azure service level agreement. http://www.microsoft.com/windowsazure/sla/.

[8] Azure storage resources. http://blogs.msdn.com/b/windowsazurestorage/archive/2010/03/28/windows-azure-storage-resources.aspx.

[9] Cloud survey. http://assets1.csc.com/newsroom/downloads/CSC_Cloud_Usage_Index_Report.pdf read the 28/03/2012.

[10] http ://nosql-database.org/. http://nosql-database.org/censoredNoSQLdatabases.

[11] Lokad-cloud. http://code.google.com/p/lokad-cloud/.

[12] Lokad-cqrs. http://lokad.github.com/lokad-cqrs/.

[13] Microsoft cloud investment. http://www.bloomberg.com/news/2011-04-06/microsoft-s-courtois-says-to-spend-90-of-r-d-on-cloud-strategy.html read the 28/03/2012.

[14] Mpi cluster on ec2. http://datawrangling.s3.amazonaws.com/elasticwulf_pycon_talk.pdf read the 03/05/2012.

[15] Problems with acid and how to fix them. http://dbmsmusings.blogspot.com/2010/08/problems-with-acid-and-how-to-fix-them.html read the 28/03/2012.

[16] Should i expose synchronous wrappers for asynchronous methods ? http://blogs.msdn.com/b/pfxteam/archive/2012/04/13/10293638.aspx read the 26/04/2012.

[17] Sort benchmark home page. http://sortbenchmark.org/.

[18] Websitemonitoring. http://www.website-monitoring.com/.

[19] *VL2 : A Scalable and Flexible Data Center Network*, 2011.

[20] D. Abadi. Problems with cap, and yahoo's little known nosql system. http://dbmsmusings.blogspot.com/2010/04/problems-with-cap-and-yahoos-little.html read the 28/03/2012.

[21] C. Abraham, P. A. Cornillon, E. Matzner-Løber, and N. Molinari. Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics*, 30 :581–595, 2003.

[22] R. Agrawal and J. C. Shafer. Parallel mining of association rules : Design, implementation, and experience. *IEEE Trans. Knowledge and Data Eng., 8(6) :962-969*, 1996.

[23] J. Albrecht, C. Tuttle, A. C. Snoeren, and A. Vahdat. Loose synchronization for large-scale networked systems. In *Proceedings of the annual conference on USENIX '06 Annual Technical Conference*, ATEC '06, pages 28–28, Berkeley, CA, USA, 2006. USENIX Association.

[24] D. P. Anderson. Boinc : A system for public-resource computing and storage. In *5th IEEE/ACM International Workshop on Grid Computing*, pages 4–10, 2004.

[25] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. Above the clouds : A berkeley view of cloud computing. Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley, Feb 2009.

[26] D. Arthur and S. Vassilvitskii. How slow is the k-means method ? *Construction*, 2006.

[27] S. Ayende. Slaying relational dragons.

[28] J. Bahi, S. Contassot-Vivier, and R. Couturier. Evaluation of the asynchronous iterative algorithms in the context of distant heterogeneous clusters. *Parallel Computing*, 31(5) :439–461, 2005.

[29] R. Bekkerman, M. Bilenko, and J. Langford. *Scaling up Machine Learning*. Cambridge University Press, 2012.

[30] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*. Springer-Verlag, 1990.

[31] S. Bermejo and J. Cabestany. The effect of finite sample size on on-line k-means. *Neurocomputing*, 48 :511–539, 2002.

[32] Gérard Biau, Luc Devroye, and Gábor Lugosi. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54(2) :781–790, 2008.

[33] L. Bottou and Y. Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems 7*, pages 585–592. MIT Press, 1995.

[34] L. Bottou and Y. LeCun. Large scale online learning. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

[35] L. Bottou and Y. LeCun. On-line learning for very large datasets. *Applied Stochastic Models in Business and Industry*, 21 :137–151, 2005.

[36] P. S. Bradley and U. M. Fayyad. *Refining Initial Points for K-Means Clustering*, volume 727, pages 91–99. Morgan Kaufmann, San Francisco, CA, 1998.

[37] B. Calder, J. Wang, A. Ogus, N. Nilakantan, A. Skjolsvold, S. McKelvie, Y. Xu, S. Srivastav, J. Wu, H. Simitci, J. Haridas, C. Uddaraju, H. Khatri, A. Edwards, V. Bedekar, S. Mainali, R. Abbasi, A. Agarwal, M. F. Haq, M. I. Haq, D. Bhardwaj, S. Dayanand, A. Adusumilli, M. McNett, S. Sankaran, K. Manivannan, and L. Rigas. Windows azure storage : a highly available cloud storage service with

strong consistency. In *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*, SOSP '11, pages 143–157, New York, NY, USA, 2011. ACM.

[38] F. Cappello, E. Caron, M. Dayde, F. Desprez, Y. Jegou, P. Primet, E. Jeannot, S. Lanteri, J. Leduc, N. Melab, G. Mornet, R. Namyst, B. Quetier, and O. Richard. Grid'5000 : A large scale and highly reconfigurable grid experimental testbed. In *Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing*, GRID '05, pages 99–106, Washington, DC, USA, 2005. IEEE Computer Society.

[39] R. Chaiken, B. Jenkins, P. Larson, B. Ramsey, D. Shakib, S. Weaver, and J. Zhou. Scope : easy and efficient parallel processing of massive data sets. *Proc. VLDB Endow.*, 1(2) :1265–1276, August 2008.

[40] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable : A distributed storage system for structured data. In *in proceedings of the 7th conference on usenix symposium on operating systems design and implementation - volume 7*, pages 205–218, 2006.

[41] C. T. Chu, S. K. Kim, Y. A. Lin, Y. Yu, G. R. Bradski, A. Y. Ng, and K. Olukotun. Map-Reduce for Machine Learning on Multicore. In Bernhard Schölkopf, John C. Platt, and Thomas Hoffman, editors, *NIPS*, pages 281–288. MIT Press, 2006.

[42] I. Clarke, O. Sandberg, B. Wiley, and T. W. Hong. Freenet : A distributed anonymous information storage and retrieval system. In *International workshop on designing privacy enhancing technologies : design issues in anonymity and unobservability*, pages 46–66. Springer-Verlag New York, Inc., 2001.

[43] S. Contassot-Vivier, T. Jost, and S. Vialle. Impact of asynchronism on gpu accelerated parallel iterative computations. In Kristján Jónasson, editor, *PARA 2010 : State of the Art in Scientific and Parallel Computing*, LNCS. Springer, Heidelberg, 2011. To be published.

[44] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears. Benchmarking cloud serving systems with ycsb. In *Proceedings of the 1st ACM symposium on Cloud computing*, SoCC '10, pages 143–154, New York, NY, USA, 2010. ACM.

[45] J. Dai and B. Huang. *Design Patterns for Cloud Services*, volume 74, pages 31–56. Springer Berlin Heidelberg, 2011.

[46] C. de Boor. On calculating with b-splines. *Journal of Approximation Theory*, 6 :50–62, 1972.

[47] C. de Boor. *A practical guide to splines*. Springer-Verlag, 1978.

[48] J. Dean and S. Ghemawat. Mapreduce : simplified data processing on large clusters. In *OSDI'04 : Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation*, pages 10–10, Berkeley, CA, USA, 2004. USENIX Association.

[49] J. Dean and S. Ghemawat. Mapreduce : simplified data processing on large clusters. In *Communications of the ACM, vol. 51, no. 1*, pages 107–113, 2008.

[50] O. Dekel, R. Gilad-Bachrach, O Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13 :165–202, March 2012.

[51] O. Delalleau and Y. Bengio. Parallel stochastic gradient descent. *Proceedings of SPIE*, 6711 :67110F–67110F–14, 2007.

[52] I. S. Dhillon and D. S. Modha. A data-clustering algorithm on distributed memory multiprocessors. In *Revised Papers from Large-Scale Parallel Data Mining, Workshop on Large-Scale Parallel KDD Systems, SIGKDD*, pages 245–260, London, UK, 2000. Springer-Verlag.

[53] On Ibm Sp (draft, Gang Cheng, and Marek Podgorny. The high performance switch and programming interfaces on ibm sp2, 1995.

[54] U. Drepper. What every programmer should know about memory. *Changes*, 3(4) :114, 2007.

[55] J.C. Fort, M. Cottrell, and P. Letremy. Stochastic on-line algorithm versus batch algorithm for quantization and self organizing maps. *Neural Networks for Signal Processing XI Proceedings of the 2001 IEEE Signal Processing Society Workshop IEEE Piscataway NJ USA*, 00(C) :43–52, 2001.

[56] A. Freeman. *Pro .NET 4 Parallel Programming in C#*. Apress, 2010.

[57] S. Garfinkel. Commodity grid computing with amazon s3 and ec2. *Usenix*, 32 :7–13, 2007.

[58] V. Georgitsis and J. Sobolewski. Performance of mpl and mpich on the sp2 system.

[59] A. Gersho and R. M. Gray. *Vector quantization and signal compression*. Kluwer, 1992.

[60] S. Ghemawat, H. Gobioff, and S. Leung. The google file system. *SIGOPS Oper. Syst. Rev.*, 37(5) :29–43, October 2003.

[61] S. Gilbert and N. Lynch. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. *SIGACT News*, 33(2) :51–59, 2002.

[62] James R. Goodman, Mary K. Vernon, and Philip J. Woest. Efficient synchronization primitives for large-scale cache-coherent multiprocessors. *SIGARCH Comput. Archit. News*, 17(2) :64–75, April 1989.

[63] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel. The cost of a cloud : research problems in data center networks. *SIGCOMM Comput. Commun. Rev.*, 39(1) :68–73, December 2008.

[64] J. Gregorio. Sharding counters. `https://developers.google.com/appengine/articles/sharding_counters` read the 17/05/2012.

[65] W. H. Greub. *Linear algebra*. Springer-Verlag, 4th edition, 1975.

[66] W. Gropp and E. Lusk. Reproducible measurements of mpi performance characteristics. In *Proceedings of the 6th European PVM/MPI Users' Group Meeting on Recent Advances in Parallel Virtual Machine and Message Passing Interface*, pages 11–18, London, UK, UK, 1999. Springer-Verlag.

[67] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24 :8–12, 2009.

[68] P. Helland. *Life beyond distributed transactions : an apostate's opinion*, volume Asilomar,, pages 132–141. 2007.

[69] C. Hennig. Models and methods for clusterwise linear regression. In *Proceedings in Computational Statistics*, pages 3–0. Springer, 1999.

[70] M. Herlihy, B. H. Lim, and N. Shavit. Scalable concurrent counting. *ACM Transactions on Computer Systems*, 13 :343–364, 1995.

[71] M. Herlihy and N. Shavit. *The Art of Multiprocessor Programming*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.

[72] T. Hey, S. Tansley, and K. Tolle. *The Fourth Paradigm : Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington, 2009.

[73] Z. Hill, J. Li, M. Mao, A. Ruiz-Alvarez, and M. Humphrey. Early observations on the performance of windows azure. *Sci. Program.*, 19(2-3) :121–132, April 2011.

[74] T. Hoefler, W. Gropp, R. Thakur, and J. L. Träff. *Toward Performance Models of MPI Implementations for Understanding Application Scaling Issues*, pages 21–30. Springer-Verlag Berlin Heidelberg, 2010.

[75] B. Hong and Z. He. An asynchronous multithreaded algorithm for the maximum network flow problem with nonblocking global relabeling heuristic. *IEEE Transactions on Parallel and Distributed Systems*, 22 :1025–1033, 2011.

[76] M. Isard. Autopilot : automatic data center management. *SIGOPS Oper. Syst. Rev.*, 41(2) :60–67, April 2007.

[77] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. Dryad : distributed data-parallel programs from sequential building blocks. In *EuroSys '07 : Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007*, pages 59–72, New York, NY, USA, 2007. ACM.

[78] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering : A review. In *ACM computing surveys, Vol.31, no.3, September*, 1999.

[79] M. N. Joshi. Parallel k-means algorithm on distributed memory multiprocessors. *Cities*, spring 2003.

[80] S. Kantabutra and A. L. Couch. Parallel k-means clustering algorithm on nows. *NecTec Technical Journal*, January 2000.

[81] T. Kohonen. Analysis of a simple self-organizing process. *Biological Cybernetics*, 44 :135–140, 1982.

[82] J. G. Koomey. Worldwide electricity used in data centers. *Environmental Research Letters*, 3(3) :034008, 2008.

[83] P. Kraj, A. Sharma, N. Garge, R. Podolsky, and R. A. McIndoe. Parakmeans : implementation of a parallelised k-means algorithm suitable for laboratory use. *BMC BioInformatics*, april 2008.

[84] H. J. Kushner and D. S. Clark. *Stochastic approximation for constrained and unconstrained systems*. Springer-Verlag, 1978.

[85] S. M. Larson, C. D. Snow, M. Shirts, and V. S. Pande. Folding@home and genome@home : Using distributed computing to tackle previously intractable problems in computational biology. *Security*, 2009.

[86] A. Lenk, M. Klems, J. Nimis, S. Tai, and T. Sandholm. What's inside the cloud ? an architectural map of the cloud landscape. In *ICSE Workshop on Software Engineering Challenges of Cloud Computing, 2009. CLOUD 09.* IEEE Press, Mai 2009.

[87] W. Liao. Parallel k-means data clustering. 2005.

[88] J. Lin. Brute force and indexed approaches to pairwise document similarity comparisons with mapreduce. In *In Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009*, pages 155–162, 2009.

[89] J. Lin. The Curse of Zipf and Limits to Parallelization : A Look at the Stragglers Problem in MapReduce. In *LSDS-IR workshop*, 2009.

[90] J. Lin and C. Dyer. *Data-intensive text processing with MapReduce*. Morgan & Claypool Publishers, 2010.

[91] H. Liu and D. Orban. Cloud mapreduce : a mapreduce implementation on top of a cloud operating system. Technical report, Accenture Technology Labs, 2009. `http://code.google.com/p/cloudmapreduce/`.

[92] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28 :129–137, 2003.

[93] G. Louppe and P. Geurts. A zealous parallel gradient descent algorithm. In *NIPS 2010 Workshop on Learning on Cores, Clusters and Clouds*, 2010.

[94] G. Louppe and P. Geurts. A zealous parallel gradient descent algorithm. In *NIPS 2010 Workshop on Learning on Cores, Clusters and Clouds*, 2010.

[95] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein. Graphlab : A new parallel framework for machine learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, California, July 2010.

[96] J. B. MacQueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[97] M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is np-hard. In *Proceedings of the 3rd International Workshop on Algorithms and Computation*, WALCOM '09, pages 274–285, Berlin, Heidelberg, 2009. Springer-Verlag.

[98] D. C. Marinescu. *Cloud Computing : Theory and Practice*. 2012.

[99] R. Masud. High performance computing with clouds.

[100] G. W. Milligan and P. D. Isaac. The validation of four ultrametric clustering algorithms. *Pattern Recognition*, 12 :41–50, 1980.

[101] B. Mirkin. *Clustering for data mining : a data recovery approach*. Chapman & Hall/CRC, 2005.

[102] J. Napper and P. Bientinesi. Can cloud computing reach the top500 ? In *UCHPC-MAW 09 : Proceedings of the combined workshops on UnConventional high performance computing workshop plus memory access workshop*, pages 17–20, New York, NY, USA, 2009. ACM.

[103] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov. The eucalyptus open-source cloud-computing system. In *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, CCGRID '09, pages 124–131, Washington, DC, USA, 2009. IEEE Computer Society.

[104] O. O Malley. Terabyte sort on apache hadoop. `http://www.hpl.hp.com/hosted/sortbenchmark/YahooHadoop.pdf`.

[105] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. Pig latin : a not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pages 1099–1110, New York, NY, USA, 2008. ACM.

[106] G. Pagès. A space vector quantization for numerical integration. *Journal of Applied and Computational Mathematics*, 89 :1–38, 1997.

[107] B. Patra. Convergence of distributed asynchronous learning vector quantization algorithms. *ArXiv e-prints*, December 2010.

[108] A. D. Peterson, A. P. Ghosh, and R. Maitra. A systematic evaluation of different methods for initializing the k-means clustering algorithm. Technical report, 2010.

[109] R. Pike, S. Dorward, R. Griesemer, and S. Quinlan. Interpreting the data : Parallel analysis with sawzall. *Sci. Program.*, 13(4) :277–298, October 2005.

[110] D. Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, January 1981.

[111] D. Pritchett. Base : An acid alternative. *Queue*, 6(3) :48–55, May 2008.

[112] Y. Raz. The dynamic two phase commitment (d2pc) protocol. In *Database Theory - ICDT '95, Lecture Notes in Computer Science, Volume 893 Springer, ISBN 978-3-540-58907-5*, pages 162–176, 1995.

[113] F. Rossi, B. Conan-Guez, and A. El Golli. Clustering functional data with the som algorithm. In *Proceedings of ESANN 2004*, 2004.

[114] S. Saini, D. H. Bailey, and S. Origin. Nas parallel benchmark (version 1.0) results 11-96. Technical report, 1996.

[115] J. C. Shafer, R. Agrawal, and M. Mehta. A scalable parallel classifier for data mining. *Proc. 22nd International Conference on VLDB, Mumbai, India*, 1996.

[116] Y. Simmhan, C. van Ingen, G. Subramanian, and J. Li. Bridging the gap between the cloud and an escience application platform. *Science*, 64(1) :56–68, 2003.

[117] M. Snir, S. Otto, S. Huss-Lederman, W. David, and J. Dongarra. *MPI : The Complete Reference*. MIT Press, Boston, 1996.

[118] V. S. Sunderam. Pvm : A framework for parallel distributed computing. *Concurrency : Practice and Experience*, 2 :315–339, 1990.

[119] D. Thain, T. Tannenbaum, and M. Livny. Distributed computing in practice : The condor experience. *Concurrency and Computation : Practice and Experience*, 17 :2–4, 2005.

[120] J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31 :803–812, 1986.

[121] L. G. Valiant. A bridging model for parallel computation. *Communications of the ACM*, 33(8) :103–111, 1990.

[122] W. Vogels. Eventually consistent. *Queue*, 6(6) :14–19, October 2008.

[123] H. Wada, A. Fekete, L. Zhao, K. Lee, and A. Liu. Data consistency properties and the trade-offs in commercial cloud storages : the consumers'perspective. *Reading*, pages 134–143, 2011.

[124] E. Walker. Benchmarking Amazon EC2 for high-performance scientific computing. *LOGIN*, 33(5) :18–23, October 2008.

[125] Guohui Wang and T. S. Eugene Ng. The impact of virtualization on network performance of amazon ec2 data center. In *Proceedings of the 29th conference on Information communications*, INFOCOM'10, pages 1163–1171, Piscataway, NJ, USA, 2010. IEEE Press.

[126] Tom White. *Hadoop : The Definitive Guide*. O'Reilly Media, Inc., 1st edition, 2009.

[127] E. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. In *Communications in Pure and Applied Mathematics vol. 13, No. I February*, 1960.

[128] D.Randall Wilson and Tony R. Martinez. The general inefficiency of batch training for gradient descent learning. *Neural Networks*, 16(10) :1429 – 1451, 2003.

[129] P.C. Yew, N.F. Tzeng, and D.H. Lawrie. Distributing hot-spot addressing in large-scale multiprocessors. *IEEE Transactions on Computers*, pages 388–395, April 1987.

[130] Y. Yu, M. Isard, D. Fetterly, M. Budiu, U. Erlingsson, P. Kumar, and G. J. Currey. Dryadlinq : a system for general-purpose distributed data-parallel computing using a high-level language. In *Proceedings of the 8th USENIX conference on Operating systems design and implementation*, OSDI'08, pages 1–14, Berkeley, CA, USA, 2008. USENIX Association.

[131] M. Zinkevich, A. Smola, and J. Langford. Slow learners are fast. In *Advances in Neural Information Processing Systems 22*, pages 2331–2339, 2009.

[132] M. Zinkevich, M. Weimer, A. Smola, and L. Li. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems 23*, 2010.