The Google Stack

Source: Malte Schwarzkopf. “Operating system support for warehouse-scale computing”. PhD thesis. University of Cambridge Computer Laboratory (to appear), 2015, Chapter 2.

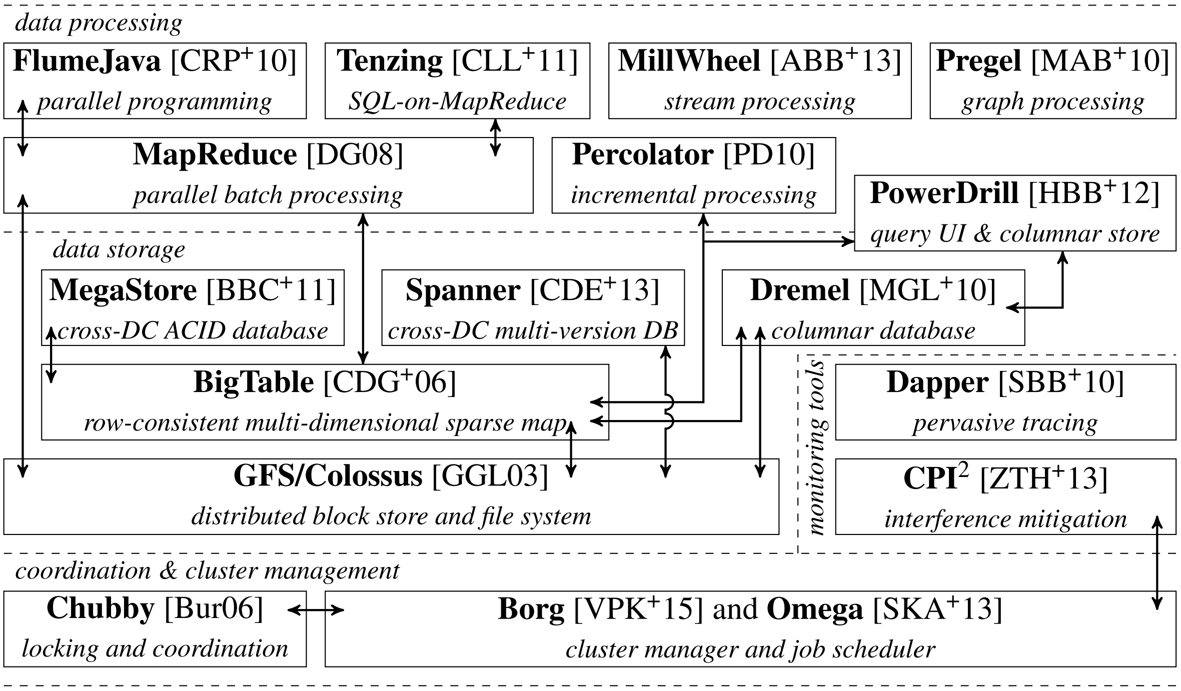

Figure 1: The Google infrastructure stack. I omit the F1 database [SOE+12] (the back-end of which was superseeded by Spanner), and unknown front-end serving systems. Arrows indicate data exchange and dependencies between systems; simple layering does *not* imply a dependency or relation.

In addition, there are also papers that do not directly cover systems in the Google stack:

- An early-days (2003) high-level overview of the Google architecture [BDH03].
- An extensive description of Google’s General Configuration Language (GCL), sadly with some parts blackened [Bok08].
- A study focusing on tail latency effects in Google WSCs [DB13].
- Several papers characterising Google workloads from public traces [MHC+10; SCH+11; ZHB11; DKC12; LC12; RTG+12; DKC13; AA14].
- Papers analysing the impact of workload co-location [MTH+11; MT13], hyperthreading [ZZE+14], and job packing strategies on workloads [VKW14].

Bibliography

[AA14] Omar Arif Abdul-Rahman and Kento Aida. “Towards understanding the usage behavior of Google cloud users: the mice and elephants phenomenon”. In: *Proceedings of the IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*. Singapore, Dec. 2014, pp. 272–277 (cited on page 1).

[ABB+13] Tyler Akidau, Alex Balikov, Kaya Bekirolu, Slava Chernyak, Josh Haberman, Reuven Lax, et al. “MillWheel: Fault-tolerant Stream Processing at Internet Scale”. In: *Proceedings of the VLDB Endowment* 6.11 (Aug. 2013), pp. 1033–1044 (cited on page 1).

[BBC+11] Jason Baker, Chris Bond, James C Corbett, JJ Furman, Andrey Khorlin, James Larson, et al. “Megastore: Providing Scalable, Highly Available Storage for Interactive Services”. In: *Proceedings of the 5th Biennial Conference on Innovative Data Systems Research (CIDR)*. Asilomar, California, USA, Jan. 2011, pp. 223–234 (cited on page 1).

[BDH03] Luiz André Barroso, Jeff Dean, and Urs Hölzle. “Web search for a planet: The Google cluster architecture”. In: *IEEE Micro* 23.2 (Mar. 2003), pp. 22–28 (cited on page 1).

[Bok08] Ibrahim Bokharouss. “GCL Viewer: A study in improving the understanding of GCL programs”. Master’s thesis. TU Eindhoven, 2008 (cited on page 1).

[Bur06] Mike Burrows. “The Chubby Lock Service for Loosely-coupled Distributed Systems”. In: *Proceedings of the 7th Symposium on Operating Systems Design and Implementation (OSDI)*. Seattle, Washington, USA, 2006, pp. 335–350 (cited on page 1).

[CRP+10] Craig Chambers, Ashish Raniwala, Frances Perry, Stephen Adams, Robert R. Henry, Robert Bradshaw, et al. “FlumeJava: Easy, Efficient Data-parallel Pipelines”. In: *Proceedings of the 2010 ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*. Toronto, Ontario, Canada, 2010, pp. 363–375 (cited on page 1).

[CDG+06] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, et al. "Bigtable: A Distributed Storage System for Structured Data". In: *Proceedings of the 7th USENIX Symposium on Operating System Design and Implementation (OSDI)*. 2006 (cited on page 1).

[CLL+11] Biswapesh Chattopadhyay, Liang Lin, Weiran Liu, Sagar Mittal, Prathyusha Aragonda, Vera Lychagina, et al. "Tenzing: A SQL Implementation On The MapReduce Framework". In: *Proceedings of the 37th International Conference on Very Large Data Bases (VLDB)*. Seattle, WA, USA, Aug. 2011, pp. 1318–1327 (cited on page 1).

[CDE+13] James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, J. J. Furman, et al. "Spanner: Google's Globally Distributed Database". In: *ACM Transactions on Computer Systems* 31.3 (Aug. 2013), 8:1–8:22 (cited on page 1).

[DB13] Jeffrey Dean and Luiz André Barroso. "The Tail at Scale". In: *Communications of the ACM* 56.2 (Feb. 2013), pp. 74–80 (cited on page 1).

[DG08] Jeffrey Dean and Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters". In: *Communications of the ACM* 51.1 (Jan. 2008), pp. 107–113 (cited on page 1).

[DKC13] Sheng Di, Derrick Kondo, and Franck Cappello. "Characterizing Cloud Applications on a Google Data Center". In: *Proceedings of the 42nd International Conference on Parallel Processing (ICPP)*. Lyon, France, Oct. 2013, pp. 468–473 (cited on page 1).

[DKC12] Sheng Di, Derrick Kondo, and Walfredo Cirne. "Characterization and Comparison of Cloud versus Grid Workloads". In: *Proceedings of the 2012 IEEE International Conference on Cluster Computing (CLUSTER)*. Sept. 2012, pp. 230–238 (cited on page 1).

[GGL03] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. "The Google File System". In: *Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP)*. Bolton Landing, NY, USA, 2003, pp. 29–43 (cited on page 1).

[HBB+12] Alexander Hall, Olaf Bachmann, Robert Büssow, Silviu Gnceanu, and Marc Nunkesser. "Processing a Trillion Cells Per Mouse Click". In: *Proceedings of the VLDB Endowment* 5.11 (July 2012), pp. 1436–1446 (cited on page 1).

[LC12] Zitao Liu and Sangyeun Cho. "Characterizing machines and workloads on a Google cluster". In: *Proceedings of the 8th International Workshop on Scheduling and Resource Management for Parallel and Distributed Systems (SRMPDS)*. Pittsburgh, Pennsylvania, USA, Sept. 2012, pp. 397–403 (cited on page 1).

[MAB^{+}10] Grzegorz Malewicz, Matthew H. Austern, Aart J.C Bik, James C. Dehnert, Ilan Horn, Naty Leiser, et al. "Pregel: A System for Large-scale Graph Processing". In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data (SIGMOD)*. Indianapolis, Indiana, USA, 2010, pp. 135–146 (cited on page 1).

[MT13] Jason Mars and Lingjia Tang. "Whare-map: Heterogeneity in "Homogeneous" Warehouse-scale Computers". In: *Proceedings of the 40th Annual International Symposium on Computer Architecture (ISCA)*. Tel-Aviv, Israel, 2013, pp. 619–630 (cited on page 1).

[MTH^{+}11] Jason Mars, Lingjia Tang, Robert Hundt, Kevin Skadron, and Mary Lou Soffa. "Bubble-up: Increasing utilization in modern warehouse scale computers via sensible co-locations". In: *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. 2011, pp. 248–259 (cited on page 1).

[MGL^{+}10] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, et al. "Dremel: Interactive Analysis of Web-scale Datasets". In: *Proceedings of the VLDB Endowment* 3.1-2 (Sept. 2010), pp. 330–339 (cited on page 1).

[MHC^{+}10] Asit K. Mishra, Joseph L. Hellerstein, Walfredo Cirne, and Chita R. Das. "Towards Characterizing Cloud Backend Workloads: Insights from Google Compute Clusters". In: *SIGMETRICS Performance Evalution Review* 37.4 (Mar. 2010), pp. 34–41 (cited on page 1).

[PD10] Daniel Peng and Frank Dabek. "Large-scale Incremental Processing Using Distributed Transactions and Notifications". In: *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation (OSDI)*. Vancouver, BC, Canada, 2010, pp. 1–15 (cited on page 1).

[RTG^{+}12] Charles Reiss, Alexey Tumanov, Gregory R. Ganger, Randy H. Katz, and Michael A. Kozuch. "Heterogeneity and dynamicity of clouds at scale: Google trace analysis". In: *Proceedings of the 3rd ACM Symposium on Cloud Computing (SoCC)*. San Jose, California, 2012, 7:1–7:13 (cited on page 1).

[Sch15] Malte Schwarzkopf. "Operating system support for warehouse-scale computing". PhD thesis. University of Cambridge Computer Laboratory (to appear), 2015 (cited on page 1).

[SKA^{+}13] Malte Schwarzkopf, Andy Konwinski, Michael Abd-El-Malek, and John Wilkes. "Omega: flexible, scalable schedulers for large compute clusters". In: *Proceedings of the 8th ACM European Conference on Computer Systems (EuroSys)*. Prague, Czech Republic, Apr. 2013, pp. 351–364 (cited on page 1).

[SCH^{+}11] Bikash Sharma, Victor Chudnovsky, Joseph L. Hellerstein, Rasekh Rifaat, and Chita R. Das. “Modeling and synthesizing task placement constraints in Google compute clusters”. In: *Proceedings of the 2nd ACM Symposium on Cloud Computing (SoCC).* Cascais, Portugal, 2011, 3:1–3:14 (cited on page 1).

[SOE^{+}12] Jeff Shute, Mircea Oancea, Stephan Ellner, Ben Handy, Eric Rollins, Bart Samwel, et al. “F1: The Fault-tolerant Distributed RDBMS Supporting Google’s Ad Business”. In: *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (SIGMOD).* Scottsdale, Arizona, USA, 2012, pp. 777–778 (cited on page 1).

[SBB^{+}10] Benjamin H. Sigelman, Luiz André Barroso, Mike Burrows, Pat Stephenson, Manoj Plakal, Donald Beaver, et al. *Dapper, a Large-Scale Distributed Systems Tracing Infrastructure*. Technical report. Google, Inc., 2010 (cited on page 1).

[VKW14] A. Verma, M. Korupolu, and J. Wilkes. “Evaluating job packing in warehouse-scale computing”. In: *Proceedings of the 2014 IEEE International Conference on Cluster Computing (CLUSTER).* Madrid, Spain, Sept. 2014, pp. 48–56 (cited on page 1).

[VPK^{+}15] Abhishek Verma, Luis David Pedrosa, Madhukar Korupolu, David Oppenheimer, and John Wilkes. “Large scale cluster management at Google”. In: *Proceedings of the 10th ACM European Conference on Computer Systems (EuroSys).* To appear. Bordeaux, France, Apr. 2015 (cited on page 1).

[ZZE^{+}14] Yan Zhai, Xiao Zhang, Stephane Eranian, Lingjia Tang, and Jason Mars. “HaPPy: Hyperthread-aware Power Profiling Dynamically”. In: *Proceedings of the 2014 USENIX Annual Technical Conference (ATC).* Philadelphia, PA, USA, 2014, pp. 211–218 (cited on page 1).

[ZHB11] Qi Zhang, Joseph L. Hellerstein, and Raouf Boutaba. “Characterizing task usage shapes in Google’s compute clusters”. In: *Proceedings of the 5th Workshop on Large Scale Distributed Systems and Middleware (LADIS).* 2011 (cited on page 1).

[ZTH^{+}13] Xiao Zhang, Eric Tune, Robert Hagmann, Rohit Jnagal, Vrigo Gokhale, and John Wilkes. “CPI^{2}: CPU Performance Isolation for Shared Compute Clusters”. In: *Proceedings of the 8th ACM European Conference on Computer Systems (EuroSys).* Prague, Czech Republic, 2013, pp. 379–391 (cited on page 1).