

Cache Characterization and Performance Studies Using Locality Surfaces Annotated Bibliography

Elizabeth S. Sorenson

February 2003

This annotated bibliography represents the important papers and books dealing with the topics of cache memories, cache performance, and locality.

1 Caches, Cache Simulations, Analytical Cache Models

These papers give a background for caches and why they are important [1] [2]. They give examples of cache simulation studies and how such studies are used [2] [3]. A couple of papers also present analytical cache models that predict cache simulation results [4] [5] [6]. These prediction results can be compared with our proposed method of predicting cache simulation results using the locality surface.

2 Locality

These papers give a brief overview of the history of locality studies [7]. They introduce the original locality surface [8] and cover other issues, such as stack distance [9], related to our new locality surface. They give a number of examples of the drawbacks of previous locality measures [10] [11] [12] and how locality is used in relation to cache studies [13] [14].

3 Synthetic Traces, Motivation for Synthetic Traces

These papers present a number of methods for creating synthetic memory reference traces, from the simplest, earliest models [15] to more complex, more recent models [16] [17] [18] [19].

4 Trace Methods

These papers describe a number of methods for collecting traces [20] [21] and detail the BACH method that has been used for all the traces involved with this dissertation [22]. They show the complexity of collecting traces of sufficient length for modern cache studies [20], giving motivation for synthetic trace generation models. They also give examples of trace compression methods [23].

5 Workload Characterization

These papers examine issues related to workload characterization and why it is important [24]. They describe why a good locality metric can be useful for characterizing workloads [25].

References

- [1] Alan Jay Smith. Cache memories. *Computing Surveys*, 14(3):473–530, September 1982. The basic, standard, cache studies paper. Presents a large number of trace driven simulation results and examines a variety of cache issues. Has good explanations of the cost versus performance tradeoffs.
- [2] John L. Hennessy and David A. Patterson. *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann Publishers, San Fransisco, CA, third edition, 2003. The definitive book on systems analysis. Gives the most accepted definition of locality and the basic terminology for most cache studies. Describes column associative and victim caches and why compiler optimizations are useful to improve locality.

- [3] Sangyeun Cho, Pen-Chung Yew, and Gyungho Lee. Decoupling local variable accesses in a wide-issue superscalar processor. In *Proceedings of the 26th annual international symposium on Computer architecture*, pages 100–110. IEEE Computer Society Press, 1999. Recent example of cache studies using simulations. Compares several architecture choices, such as cache size and number of cache ports, using simulation.
- [4] A. Agarwal, M. Horowitz, and J. Hennessy. An analytical cache model. *ACM Transactions on Computer Systems*, 7(2):184–215, May 1989. Presents a well respected analytical model that gives the miss rate for a trace given a cache size, associativity, block size, and other cache parameters. The authors separate start-up effects from other causes of miss rates.
- [5] Abraham Mendelson, Dominique Thiebaut, and Dhiraj K. Pradhan. Modeling live and dead lines in cache memory systems. *IEEE Transactions on Computers*, 42(1):1–14, January 1993. Presents an analytic cache model for live and dead cache lines in direct mapped, set associative, and fully associative caches. Assumes hyperbolic model of program behavior.
- [6] Che-Chi Weng and Eric E. Johnson. The time-space model for instruction reference behavior. In *Proceedings of the 1994 IEEE International Phoenix Conference on Computers and Communications*, pages 220–226, April 1994. Model of instruction references used to predict miss ratio curves without simulation. Uses sojourn times and working spaces of program executions. More accurate predictions for larger cache sizes.
- [7] P. J. Denning. The working set model for program behavior. *Communications of the ACM*, 11(5):323–333, May 1968. Largely considered the original paper on locality. Denning introduces the concept of 'working set,' or the most recently used pages from a paged memory.
- [8] K. Grimsrud. *Visualizing Locality*. PhD thesis, Brigham Young University, 1993. In this dissertation, Grimsrud quantifies locality and creates the original locality surface. He presents a few uses of the surface, including workload characterization, synthetic trace evaluation, and memory hierarchy design choices.

- [9] Thomas M. Conte and Wen mei W. Hwu. Benchmark characterization for experimental system evaluation. In *Proceedings of the 1990 Hawaii International Conference on System Sciences (HICSS)*, volume I of *Architecture Track*, pages 6–18, 1990. Introduces some locality measures that are stack based, meaning the measures use the unique number of references for the delay.
- [10] A. Wayne Madison and Alan P. Batson. Characteristics of program localities. *Communications of the ACM*, 19(5):285–294, May 1976. An early quantitative definition of locality and methods for detecting it. Examines how locality changes throughout a trace and defines a phase of localized reference behavior.
- [11] F. J. Sanchez and A. Gonzalez. Data locality analysis of the specfp95. *Digest of Performance Analysis and its Impact on Design (PAID) Workshop*, pages 78–84, 1998. Detailed locality analysis of specfp95 demonstrating the complexity of multiple 2 dimensional graphs. Introduces a tool that provides information about a variety of cache configurations and many locality statistics with just one traversal of a given benchmark.
- [12] Kathryn S. McKinley and Olivier Temam. Quantifying loop nest locality using spec’95 and the perfect benchmarks. *ACM Transactions on Computer Systems*, 17(4):288–336, November 1999. Example of the use of locality in examining loop nests, using spec95 and perfect benchmarks. Demonstrates the complexity of multiple definitions of locality (spatial, temporal, self, and group locality) and multiple 2 dimensional graphs.
- [13] Michael E. Wolf and Monica S. Lam. A data locality optimizing algorithm. In *Proceedings of the ACM SIGPLAN ’91 Conference on Programming Language Design and Implementation*, Toronto, Ontario, Canada, June 1991. Attempts to modify the compiler to improve loop nest locality. Differentiates between ’locality’ and ’reuse’ and attempts to quantify both terms.
- [14] Bruce L. Jacob, Peter M. Chen, Seth R. Silverman, and Trevor N. Mudge. An analytical model for designing memory hierarchies. *IEEE Transactions on Computers*, 45(10), October 1996. Uses a model of locality, including stack distance curves, to determine best sizes for different levels of the memory heirarchy. Takes into consideration cost as well as performance.

- [15] J. Spirn. *Program Behavior: Models and Measurements*. Elsevier North-Holland, Inc., New York, NY, 1977. The definitive book on modeling. Presents several classic synthetic trace models including the Independent Reference Mode, the Distance String Model, and the LRU Stack Model.
- [16] C. Fricker and P. Robert. A memory reference model for the analysis of cache memories. *Performance '90*, pages 255–269, 1990. Presents a mathematical description of a trace that can be used for predicting miss rates using trace statistics. Focuses on direct-mapped caches, but the work can be expanded to associative caches as long as they use LRU replacement.
- [17] Dominique Thiebaut, Joel L. Wolf, and Harold S. Stone. Synthetic traces for trace-driven simulation of cache memories. *IEEE Transactions on Computers*, 41(4):388–410, April 1992. A synthetic trace generation model based on random walks using a quantified scalar of locality and working set size. Reasons why synthetic traces are useful. Uses miss rate for determining the accuracy of the synthetic traces.
- [18] Vidyadhar Phalke and Bhaskarpillai Gopinath. An inter-reference gap model for temporal locality in program behavior. In *Proceedings of the 1995 ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems*, pages 291–300. ACM Press, 1995. Introduces the Inter-Reference Gap (IRG) model, based on a k order Markov chain. This model can be used for trace compaction.
- [19] Anup Mathur. *A Stochastic Process Model for Transient Trace Data*. PhD thesis, Virginia Polytechnic Institute and State University, 1996. Uses stochastic processes to model trace data. Uses the Kruskal-Wallis test and Correlation Analysis to validate the model.
- [20] A. Borg, R. E. Kessler, and D. W. Wall. Generation and analysis of very long address traces. In *Proceedings of the 17th International Symposium on Computer Architecture*, pages 270–279, May 1990. Presents the Epoxie method for collecting traces using link-time code modification. Demonstrates how long traces are necessary for today’s cache sizes.
- [21] Richard A. Uhlig and Trevor N. Mudge. Trace-driven memory simulation: A survey. *ACM Computing Surveys*, 29(2), June 1997. Thorough

overview of the strengths and weaknesses of various methods of trace collection, reduction, and processing.

- [22] J. K. Flanagan, B. Nelson, J. Archibald, and K. Grimsrud. BACH: BYU address collection hardware; the collection of complete traces. In *Proceedings of the 6th International Conference On Modeling Techniques and Tools for Computer Performance Evaluation*, September 1992. Good, basic description of BACH. Compares BACH traces with other traces and shows how BACH creates longer, contiguous, and more complete traces.
- [23] Eric E. Johnson, Jiheng Ha, and M. Baqar Zaidi. Lossless trace compression. *IEEE Transactions on Computers*, 50(2), February 2001. Surveys a range of loss-less trace compression schemes and their effectiveness. Looks at reductions in both space and time.
- [24] A. K. Agrawala, J. M. Mohr, and R. M. Bryant. An approach to the workload characterization problem. *IEEE Computer*, 9(6):18–32, June 1976. Isolates the problems and issues associated with workload characterization and why it is useful. Points out the workload models can be used for selecting an appropriate test workload.
- [25] Lizy Kurian John, Purnima Vasudevan, and Jyotsna Sabarinathan. Workload characterization: Motivation, goals and methodology. In *Workload Characterization: Methodology and Case Studies*, Dallas, Texas, November 1998. Talks about the goals of a locality metric and how locality can be used for workload characterization. Examines some previous definitions of locality.