Pacific Northwest
NATIONAL LABORATORY

PNNL-33259

Fixing Amdahl's Law within the Limits of Accelerated Systems

FALLACY

August 2022

Andrés Márquez
Nathan Tallent
Kilic Ozgur
Chenhao Xie
Yasodha Suriyakumar

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-33259

Fixing Amdahl's Law within the Limits of Accelerated Systems

FALLACY

August 2022

Andrés Márquez
Nathan Tallent
Kilic Ozgur
Chenhao Xie
Yasodha Suriyakumar

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

The performance of **Data Model Convergence Initiative** (DMC) applications on parallel machines is far below the limit set by Amdahl's law. Whether the machine is based on many-core, GPUs, FPGAs, or a heterogeneous combination, usually the most significant bottleneck is accessing data from the memory system. Aligning with DMC's HW/architecture thrust, this project developed a set of memory-centric tools called 'MemGaze' that inform the HW/SW stack about an application's memory behavior, including data access latency and diagnosing poor data layout and data composition. Our approach uses architectural modeling and analysis of workload data accesses.

Acknowledgments

This research was supported by the **Data Model Convergence Initiative**, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

The computational work was performed using PNNL Computing at Pacific Northwest National Laboratory.

Contents

Figures

Tables

1.0 Introduction

The performance of the state-of-the-art applications on parallel machines is far below the limit set by Amdahl's law. Whether the machine is based on many-core, GPUs, FPGAs, or a heterogeneous combination, usually the most significant bottleneck is accessing data from the memory system (Amdahl 1967), (Wulf and McKee 1995), (Esmaeilzadeh et al. 2011). Thus, memory analysis and optimization are critical. Figure 1 showcases the pitfalls that might ensue by not considering memory movement in the context of parallelization efforts.

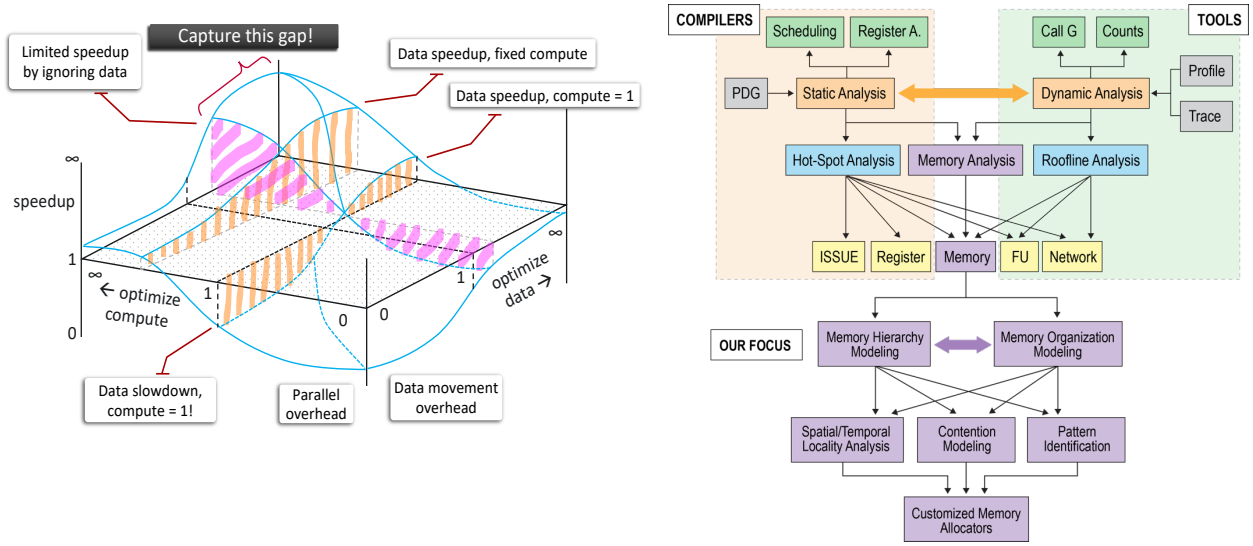

Figure 1 Amdahl's Law including Overheads due to Parallelization and Data Movement (left)
Figure 2 Toolchain Breakdown (FALLACY in magenta) (right)

The goal of this project is to provide a new low overhead tool 'MemGaze' https://github.com/pnnl/memgaze to conduct performance analysis. Furthermore, the tool must be capable of dual use, allowing the SW as well as the HW practitioner apply this tool for their analysis. The tool's breakdown is depicted in Figure 2, where the FALLACY's tool is colored in magenta. FALLACY's approach is to take a memory-centric vantage point and pursue a cascading tool flow where it switches judiciously from static to dynamic analysis. FALLACY trades off analysis overhead for approximate analysis characterization.

The project has collaborated with other DMC projects, including 'SO(DA)^2' to evaluate Data Flow engines and 'PACER' to build & analyze a suite of kernels suitable for Co-Design.

As of August'22, a large part of this work has transitioned via business development efforts to a collaborative effort with industry called 'Advanced Memory to Support AI for Science'.

2.0 MemGaze Introduction

The performance of the state-of-the-art applications on parallel machines is far below the limit set by Amdahl's law. Whether the machine is based on many-core, GPUs, FPGAs, or a heterogeneous combination, usually the most significant bottleneck is accessing data from the memory system (Amdahl 1967), (Wulf and McKee 1995), (Esmaeilzadeh et al. 2011). Thus, memory analysis and optimization are critical. The major challenge of memory analysis tools is delivering detailed insight without orders-of-magnitude of additional time, space, and execution resources. Detailed application insight requires analysis of both movement and data reuse (Beyls and D'Hollander 2001), (Weinberg et al. 2005), (Anghel et al. 2013), (Badr et al. 2020), (Carlson, Heirman, and Eeckhout 2011), (Binkert et al. 2011), (Li et al. 2020), (Kim, Yang, and Mutlu 2015). Data movement includes access frequencies and costs. Data reuse includes temporal and spatial locality, footprint, and access patterns. Such insight is typically gathered with memory reuse and modeling tools (Carlson, Heirman, and Eeckhout 2011), (Binkert et al. 2011), (Xiang et al. 2013) or memory simulators (Li et al. 2020), (Kim, Yang, and Mutlu 2015) but requires orders-of-magnitude of additional resources (time and execution) for tracing and space for data (intermediate and final). Recent measurement techniques permit low-overhead application analysis, but address only one form of locality, usually reuse distance (Wang, Liu, and Chabbi 2019).

Fallacy developed MemGaze, a new tool that provides low-overhead, load-level analysis of memory accesses and data reuse for applications. MemGaze differs from prior low-overhead tools in two ways. First, it analyzes sampled access traces, where each sample is a long address sequence (≈1K), collected with emerging hardware support for processor tracing. Such sequences are useful because they enable temporal, spatial, location, and access pattern analyses. Second, via sampled traces, it provides a broad set of memory analyses that include locations vs. operations, accesses vs. spatio-temporal reuse, and reuse (as distance, rate, and volume) vs. access patterns.

Prior low-overhead memory analyses do not analyze memory address sequences because prior collection methods incur very high overheads. Software-based sampling can capture sequences of memory addresses by switching between instrumented and non-instrumented execution, but easily incurs time overhead of 100× (Xiang et al. 2013), (Kilic, Tallent, and Friese 2020). Hardware for performance monitoring was not designed for detailed data reuse analysis. Many tools analyze cache misses or statistics from load-store queues. Costly data accesses (Liu and Mellor-Crummey 2014), (Choi, Blagodurov, and Tseng 2021) can be identified by sampling loads and capturing data addresses with AMD's IBS (or LWP) (AMD 2021), Intel's PEBS-DLA (Intel Corporation 2020), IBM's Marked Events (IBM 2018), and ARM's SPE (ARM 2021). Temporal reuse can be captured by sampling reuse intervals (Wang, Liu, and Chabbi 2019), (Eklov and Hagersten 2010), (Sasongko et al. 2021). However, none of these methods simultaneously permit temporal, spatial, location, and access pattern analyses. To enable low-overhead and high-resolution memory analysis, MemGaze uses Processor Tracing (PT) to collect sampled and compressed memory address traces. PT was originally designed to collect control flow to assist debugging and is supported by different vendors, e.g., Intel x64 (Intel Corporation 2013) and ARM (ARM 2011). Both Intel's and ARM's PT also support gathering data addresses, though neither is widely available. MemGaze leverages Intel's ptwrite instruction, which can write an arbitrary word data packet, such as an address, to a pinned OS buffer without OS

intervention. That is, PT can be entirely enabled or disabled by hardware. ptwrite is currently supported by Intel 5th and 6th gen Core CPUs and Goldmont based Atom CPUs such as the Gemini Lake that we use in our experiments; it very recently appeared with Alder Lake ('desktop') and is scheduled for Sapphire Rapids ('server') (Intel Corporation 2021), which will be used in Argonne National Laboratory's Aurora supercomputer.

Although ptwrite presents an interesting opportunity for collecting memory traces, there are several challenges. First, analysis of such traces must account for gaps due to sampling. Second, straightforward use of ptwrite for memory generates huge traces — larger than PT for control flow (Linux perf 2021b) — that require frequent copying from OS memory to application memory to storage, which consumes memory bandwidth and introduces blocking delays unless throttled. For example, Linux perf (Linux perf 2021a) drops an unpredictable 30–50% of data, even when CPU frequency is throttled. Thus, satisfactory solutions should reduce time overhead and data rates.

Salient features of MemGaze are as follows: First, we demonstrate feasibility of low-overhead sequence-aware memory analysis. Specifically, we show how to use ptwrite to collect sampled memory access traces, i.e., samples of memory accesses and their data addresses. The sampled traces enable instruction-level analysis of access sequences and their data address. Because both the sampling rate and sequence size are controllable, trace size is also controllable. Although our prototype is Intel-specific, the method is easily generalizable. Second, we describe static analysis and binary instrumentation process that enables a compressed (non-lossy) sample representation via selective instrumentation of memory instruction. Although most data reduction comes from sampling, the compression adds another factor of 1.2–2×. This analysis enables rapid decomposition of footprint by access patterns without expensive sequence analysis. Third, we describe multi-resolution analysis for locations vs. operations, accesses vs. spatio-temporal reuse, and reuse (distance, rate, volume) vs. access patterns. Because trace size is controllable, analysis times are reasonable, even with prototype implementation. Fourth, we evaluate MemGaze on benchmarks and parallel applications with different access patterns. We elucidate the memory effects of different data structure implementations and algorithms. For a sampled trace that is ≈1% of a full one, analysis metrics have 1-25% MAPE (mean absolute percentage error) for histograms of varying dynamic sequence lengths. With current suboptimal kernel support (PT runs continuously), *MemGaze's* time overhead is typically 10– 95%; 7× worst case. However, when PT runs only during samples, overhead is 10–35% on memory intensive regions and correlates with executed *ptwrites*, which are expensive.

3.0 MemGaze Overview

MemGaze is based on collecting and analyzing sampled memory traces. Detailed memory analysis of applications is often not feasible using other methods either with respect to time, data volume, or resource usage. Figure 3 shows an overview of collecting and analyzing memory traces. The first two steps implement sampled tracing. Binary instrumentation (Step 1) inserts Processor Tracing instructions (*ptwrite*) that record sampled sequences of memory addresses (Step 2). The instrumentation can be entirely enabled or disabled by hardware.

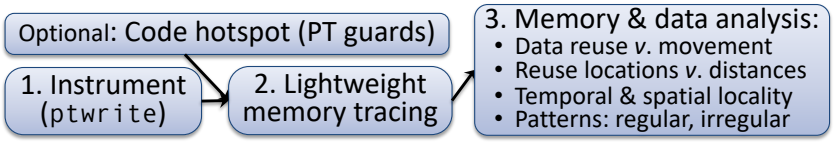

Figure 3 Collecting and Analyzing Memory Traces

To help focus results, one may optionally perform standard hotspot analysis based on time or memory loads. This result defines a region of interest (set of functions) that are used to limit tracing. The region of interest limits tracing using one of two methods: selective instrumentation (Step 1) or Processor Tracing's hardware guards (Step 2). With PT's hardware guards, the region of interest can change without reinstrumentation. Figure 4 shows a sampled trace fragment. A sample is a sequence of w recorded accesses followed by z non-recorded accesses. Since $(w + z) \gg w$, e.g., $(w + z)/w \approx 10^{3 \cdots 5}$, our traces are a fraction of a full memory trace. The result is address- and sequence-aware traces for very low time and space overhead. To diagnose data reuse and memory movement problems, we provide analyses to characterize data movement, temporal and spatial reuse, locations and footprints, and access patterns. The analyses fall into four categories. Analysis of data access frequency vs. reuse compares memory hotspots (movement)with poor locality and accesses, potential causes of unnecessary movement. Second, reuse analysis can be performed on memory locations or memory operations. The former highlights a specific memory region or data object; the latter highlights a specific sequence of memory operations. Third, we characterize both temporal and spatial locality to highlight good use of cache blocks and caches. Finally, we characterize access patterns, both regular and irregular, to distinguish between accesses that are expected to have good and poor performance. For example, the former (regular) can hide data movement with prefetching; the latter (irregular) cannot.

4.0 MemGaze High-Resolution Memory Traces

Very large traces affect (a) application execution (frequent memory copies, saturated memory bandwidth, blocking events), (b) trace integrity (unpredictable throttling and data drops), (c) downstream analysis time and memory, and (d) trace storage. This section describes how to collect sampled, high-resolution memory traces using ptwrite. It also describes static analysis for selective instrumentation and compressed traces. A sampled trace is shown in Figure 5. To collect such traces, *ptwrite* must be inserted into the instruction execution stream, which we do using static binary instrumentation.

4.1 Instrumenting Loads

Our instrumentor, which leverages DynInst (Meng and Miller 2016), takes as input an executable's important load modules, i.e., an executable and relevant library. It outputs a new executable and an auxiliary annotation file, whose contents are described below. The off-line approach ensures that the static code analysis we perform for trace compression and access pattern decomposition has no time or space overhead for executions. A benefit of binary instrumentation is that it can instrument libraries such as the C++ standard library, which can be a significant source of complex memory behavior. For each load module, the instrumentor analyzes memory accesses within each procedure separately. On x64 there are several addressing modes. The two main categories are:

```
ptwrite rs ; load rd ← [rs] + o

ptwrite rs1 ; ptwrite rs2 ; load rd ← [rs1] + k[rs2] + o
```

where `rs` indicates source registers, brackets indicate dereference, and `k` and `o` are literals for scale and offset, respectively. *ptwrites* are only inserted for source registers, i.e., dynamic info. The literals are extracted, keyed by instruction address, and placed in the auxiliary annotation file. *ptwrites* should precede loads because the source address can be overwritten when `rd = rs`. The effect of this strategy is that all instrumentation can be masked by hardware. The instrumentation is a single instruction without side effects on the CPU's architectural state. Consequently, it does not require a function call or register saves/restores. Further, the instrumentation can be enabled or disabled by hardware. In some situations, it would be possible to pack multiple *ptwrite* payloads into a single 64bit payload. Examples are two 32-bit registers or two registers whose contents are known to have a common 32-bit prefix. However, packing significantly complicates instrumentation by requiring an additional register, triggering either a trampoline call or register reallocation.

4.2 Compression via load classes

For load-based analysis we can ignore stores. Rather than instrumenting every load, we compress traces based on a load's expected access pattern. Programs often reuse constant pools of data that are uninteresting from the perspective of dynamic footprint analysis. Our analysis views as uninteresting both scalars on the execution stack and scalar global data. Thus, if a load is Constant, there is no need to instrument its register as long as we know it executed. Load access patterns are also helpful during analysis because they highlight differences in expected access latencies. To analyze access patterns, the instrumentor analyzes data dependencies for each procedure's object code. From data dependencies, the instrumentor classifies each load to distinguish three load classes.

Constant loads access scalar data either within a stack frame or to global data. Specifically, this means scalar loads (offset of 0) that are relative to a frame pointer or to a global section. All constant loads are viewed as accessing the same address, using total space of 1 unit.

Strided loads are relative to a loop induction variable (loop-carried dependency) with constant stride.

Irregular. All other loads are classified as irregular. Typically, they are indirect loads through pointers.

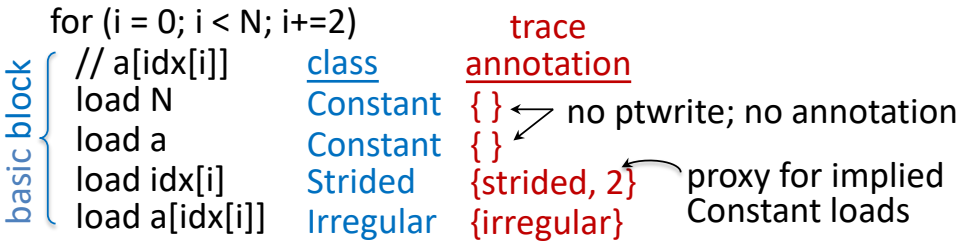

Figure 4 Trace Compression using Load Access Classes

Figure 4 shows an example of load classification and trace compression. For Strided and Irregular loads, memory addresses are always instrumented, and the load class (blue) added to the auxiliary annotations (red). For constant loads, it is sufficient to know its basic block executed. Basic blocks divide code into straight-line sequences such that an instruction is executed if and only if any other is executed. Thus, the instrumentor selects a proxy instruction within the basic block. If the block contains a Strided or Irregular load, one is selected and annotated with the number of implied Constant loads. Otherwise, the instrumentor selects the first Constant load as the proxy and instruments it. As a result, in Figure 4, only half the loads are instrumented. Our results (§ VI-C) show that with this scheme, compression of non-optimized and optimized code is about 2× and 1.2×, respectively. The difference makes sense because of the higher rate of frame loads without optimization. One can also imagine a compressed representation of strided loads using a tuple of ⟨begin-address, stride, end-address⟩. We choose to forgo this because it results in complex instrumentation (e.g., conditionals, register spilling).

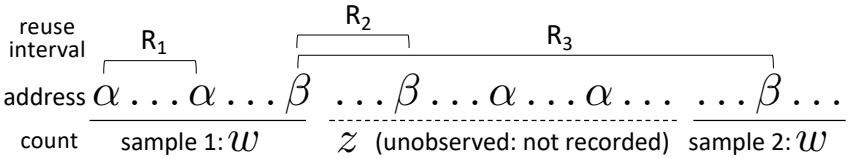

Figure 5 Sampled Memory Address Trace with Reuse Interval Categories R1-R3

4.3 Sampled memory traces

We collect traces using an extended Linux perf (Linux perf 2021a). Figure 5 shows two samples from a resulting trace. A sample is a sequence of w recorded accesses followed by z non-recorded accesses. Each recorded access is associated with an instruction pointer, memory address, and timestamp. Each sample shows memory accesses (loads) in temporal context. The samples are uniform in memory accesses because any set of accesses is equally likely to appear. The samples' accesses represent common reuse and access patterns. With Processor Tracing, the sample window w corresponds to the contents of a fixed-size circular buffer. The average sampling period $w + z$ specifies that a sampling trigger is generated to read the buffer. Usually, $(w + z) \gg w$, e.g., ratios of $10^{3\ldots5}$ to 1. Table 1 tracks important symbols. Let σ be a set of samples with a total of $|\sigma|$ samples. Averaging across σ, each sample has w loads. Let $A(\sigma)$ be the number of observed memory accesses in σ and $\mathcal{A}(\sigma)$ the (estimated) value for all (uncompressed) accesses. Then, if σ is a single sample, $w = A(\sigma)$ and $w + z = \mathcal{A}(\sigma)$.

Decompressing samples: With compression, observed accesses A can differ from accesses $\mathcal{A}$ directly implied by the observation. To reason about compression, we introduce the sample ratio ρ of all executed to all sampled memory accesses, or $\rho = \hat{A}(\sigma)/A(\sigma)$. We also introduce the compression ratio $\kappa(\sigma)$ of all to selected (compressed) accesses in σ. (see Table 1) When monitoring all memory accesses (non-selective instrumentation), ρ is simply the ratio of all to observed accesses or $w+z/w$. With selective instrumentation, we account for the constant loads directly implied by the sample:

Equation 1

$$\rho = \frac{\hat{\mathcal{A}}(\sigma)}{\mathcal{A}(\sigma)} = \overset{\textit{non-selective}}{\frac{|\sigma|(w+z)}{A(\sigma)}} = \overset{\textit{selective}}{\frac{|\sigma|(\kappa(w)+z)}{\kappa(\sigma)A(\sigma)}}$$

We calculate $\kappa(\sigma)$ using the relation $\kappa(\sigma)A(\sigma) = A(\sigma) + A_{const}(\sigma)$, which yields

Equation 2

$$\kappa(\sigma) = 1 + \frac{A_{\text{const}}(\sigma)}{A(\sigma)}$$

It is easy to calculate `Aconst(σ)` from the combination of the trace and auxiliary annotations generated during binary instrumentation. `κ(w)` is analogous.

Table 1 Important Symbols

σ, $\lvert\sigma\rvert$	Sample of access sequences; number of samples
$\hat{}$	Population estimate from sample (vs. observation)
F_{str}, F_{irr}	Footprint with {strided, irregular} access pattern
F_{str}%, F_{irr}%	Fraction of {strided, irregular} footprint
A_{const}%	Fraction of accesses to constant-sized data
ΔF	Footprint growth rate; footprint per access
ΔF_{str}%, ΔF_{irr}%	Fraction of strided (irregular) footprint growth
$w + z$	Sample period (memory accesses)
$\mathcal{A}$	All (uncompressed) accesses
A	Observed (possibly compressed) accesses
ρ	Sample ratio. All executed : all sampled accesses
κ	Compression ratio. All : selected accesses
W	Effective trace window or interval size
C	Captures, addresses with reuse
S	Survivals, addresses without reuse
D	Spatio-temporal block reuse distance
F	Footprint

4.4 Enabling source code attribution

A challenge when using binary instrumentation is attributing memory analysis results to source code. The reason is that the new (instrumented) instruction stream is no longer aligned with the load module's source-line mapping. To recover the source code mapping, we extended the functionality in DynInst with an interface that reads the newly recorded mapping between the new object code and source code.

5.0 MemGaze Analyzing Sampled Traces

This section describes analysis for traces of sampled access sequences. Compared to low-overhead methods that sample address regions or reuse instances, these traces enable temporal, spatial, location, and access pattern analyses. However, there are limitations that must be understood. We describe these limitations and then summarize our analyses.

5.1 Sampling Limitations

An obvious limitation of sampled traces is that they may miss very short or infrequent behaviors. More subtly, a uniform sample of memory accesses may not have a uniform sample of reuse intervals. A reuse interval is the number of loads (or instructions) between a pair of references to the same address (see (Yuan et al. 2019)). (In contrast, reuse distance, or stack distance (Mattson et al. 1970), is the number of unique addresses in the interval.) It turns out that some reuse intervals may not be captured, so that the samples do not represent a uniform sample of reuse intervals. To see this, we classify the ability to observe reuse intervals into three categories (Figure 5). We say a reuse interval is captured if both accesses appear within sampled data.

(R1) Within a sample (`w`), some reuse intervals `2 . . . w-1` can be captured. Some cannot because one element of the interval could occur at the end of a sample buffer.

(R2) Within a sample period `w+z`, it is not possible to capture reuse intervals `w, . . ., z`. Similarly, it is impossible to capture intervals `z+1, . . . , w+z-1`.

(R3) Between samples we have a generalization of the prior categories. It is impossible to capture intervals such that `d mod w=0, . . ., z`. Further, although it may be possible to capture some intervals greater than z+1, it is impossible to distinguish a single complete interval from multiple incomplete instances to the same address.

5.2 Reducing error with sample aggregation

Although some reuse intervals may not be observed, with sufficient samples, effective reuse analysis is very likely. Most of our analyses aggregate all samples across a certain dimension, such as a function instance, reuse distance range, or address region. Sample aggregation is important not only because it highlights diagnostic trends, but also because the accumulation of more samples reduces blind spots and statistical error. Recall that trace windows have blind spots for (R2) and potentially sparse coverage for (R3).

When trace windows are aggregated to program functions over many samples, forming code windows, we expect some observability of blind spots (region z in Figure 5) and therefore useful estimates for (R2) and (R3). The reason is that estimates for `captures (C)` and `survivals (S)` — addresses with and without reuse, respectively — significantly improve.

With sufficient samples, we can use standard estimators to scale statistics from samples and apply them to the population. The typical case is scaling metrics by the ratio of executed to

sampled memory accesses, or ρ; see Eq. (3). (It may be necessary to account for trace compression, κ.) For hotspots, we expect these diagnostics to be highly informative. If error is a concern, recall that there is often no practical non-sampling methodology for detailed memory analysis. Even with PT hardware assistance, full traces are not feasible without random drops. Rather, the choice is between coping with microbenchmarks and small data sets, devoting enormous time and machine resources to simulation, or ad hoc methods.

5.3 Multi-resolution time & location analysis

To quickly find interesting time intervals and memory regions, we use a multi-resolution analysis that recursively adjusts the granularity of execution time and access location using tree structures.

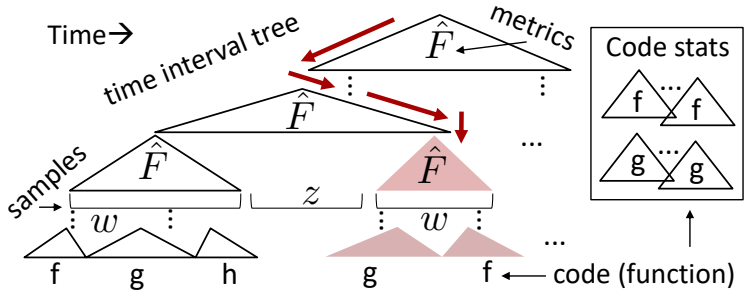

Figure 6 Multi-Resolution Time Analysis finds Time Regions with Poor Locality

1) Execution time: To find time intervals of operations with poor reuse, we analyze *accesses* and *static code* (e.g., functions) over execution time. Figure 6 shows an *execution interval tree* representing sets of samples at varying window sizes (time intervals). The execution interval tree is built bottom-up, beginning with samples. Tree nodes above samples correspond to increasing inter-sample intervals. Nodes below samples correspond to intra-sample intervals. The leaf function nodes group access sequences from the same function. Metrics are associated with each node. Inter-sample metrics are estimates, whereas intra-sample metrics are exact. The red sequence of arrows descending from the root "zooms" to a hot interval (many accesses) with poor reuse (e.g., large footprint growth).

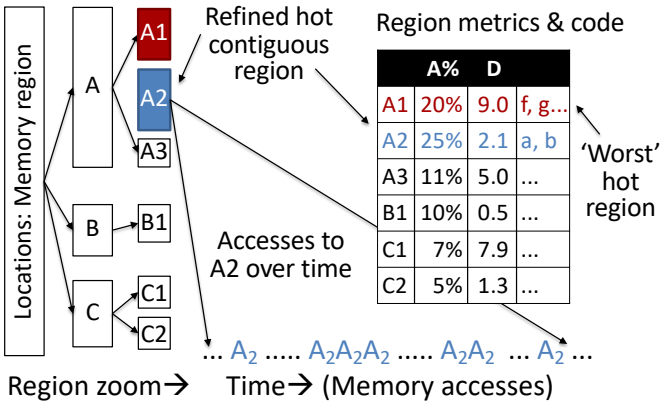

Figure 7 Location Zooming finds Hot Memory Regions with Poor Locality

2) *Location regions*: To find *memory regions* with poor spatio-temporal locality, we use location-based zooming. Figure 7 shows this analysis. The zoom tree proceeds top-down from a single memory region to its hot sub regions (left-to-right in the figure). The tree's leaves show the final regions, e.g., A2. For each final region, we show hotness (% total accesses), spatial-temporal reuse distance `D` for accesses to that region, and the code (function, line) for those accesses. The figure's table shows this data. Accesses for region A2 are shown over time (blue), which correspond to functions a and b, account for 25% of the total. However, the 'worst' hot region is A1, which accounts for 20% of total accesses but has a far worse reuse distance (`D`). The recursive zoom procedure proceeds as follows. Given a region, it is divided into fixed-sized pages and access blocks. An access block b_a, represents the unit access size for spatio-temporal reuse distance `D`; we default to the cache line size. The page size b_p is used to identify subregions and recursively reduces with tree level. A hot subregion is a maximal set of contiguous pages, each with at least 1 access, where the set's total access is at least `t%` the region's accesses. The zoom stops when a subregion reaches a minimum threshold. The contiguous property of a hot region is important. Although some parts of a hot region may be cold, including them tends to capture a single object or collections of related objects. In this way, reuse distance (`D`) represents the spatio-temporal locality of the *entire* object. In contrast, only focusing on a region's hot blocks filters all other accesses to the region, frequently making spatio-temporal locality appear very good. The stopping threshold is also important: when too large it can capture semantically different objects and average many behaviors. When too small, the analysis is resource intensive and potentially noisy.

6.0 MemGaze Memory and Data Reuse Analyses

We develop several methods for characterizing reuse within sampled traces. Together they provide a broad set of diagnostics that capture locations vs. operations, accesses vs. spatio-temporal reuse, and reuse (as distance, rate, and volume) vs. access patterns.

6.1 Data movement and access frequency

Characterizing data movement between different system components is a complex process. In this paper, we view memory as a single system. To find hot memory regions, we calculate access frequencies for each region, focusing on accesses (`A`) classified as non-Constant. These accesses also represent data that must be moved by the memory system.

6.2 Spatio-temporal reuse distance and interval

Memory performance is related to the temporal and spatial reuse of cache lines. We therefore capture spatio-temporal locality using reuse distance and reuse interval with respect to a configurable access block size. Reuse distance `D` (or stack distance) (Beyls and D'Hollander 2001), (Weinberg et al. 2005) is defined as follows. With execution time analysis, `D(w)` it is the unique memory blocks between two operations or a window of accesses `w`. With location region analysis, `D(b)` for a block `b` is the unique memory blocks between two subsequent accesses to `b`. The reuse interval for that same block measures number of accesses; it is easier to calculate but only an estimate of unique blocks. To adapt `D` to sampled traces, we either focus solely on intra-sample windows or calculate the average unique blocks accessed between samples based on footprint growth. For cache-friendly data structures, we focus on intra-sample reuse where blocks are cache lines. For working-set analysis, we use inter-sample reuse and blocks of OS page size.

6.3 Data volume: Footprint

In time analysis, it is important to understand an access sequence's data footprint and new data per access. Footprint is the amount of unique data accessed by a series of operations. The observed footprint `F(w)` for a single sample of size `w` is the unique addresses in `w`. The estimated footprint `^F(w+z)` for the sampled and unsampled addresses assumes `F(w)` is representative of `w+z`, i.e., that the ratios of unique addresses between the sample and total population are similar. With sufficient samples within a uniform sample of loads, this assumption of proportionality holds. Therefore, ˆF should scale the sample footprint F by the

ratio ρ of expected and observed footprint. Thus, over a set of many samples σ, the estimated footprint `^F(σ)` for window size `W(σ)`

Equation 3

$$\hat{F}(\sigma) = \begin{cases} F(\sigma) = C(\sigma) + S(\sigma) & \text{intra-window} \\ \rho F(\sigma) = \rho\,(C(\sigma) + S(\sigma)) & \text{inter-window} \end{cases}$$

Statistically, intra-sample intervals (R1) can be interpreted in two ways. Besides the intra-window portion of Eq. (3), they can also be viewed as a sample end point (e.g., smaller average `w`) and scaled using the inter-window portion.

6.4 Data reuse rates: Footprint growth

Footprint growth is footprint's rate of change. Let `A(σ)` be number of addresses in a sample. Then, average footprint growth `Δ^F(σ)` over the sample of window size `W(σ)` is

Equation 4

$$\Delta\hat{F}(\sigma) = \frac{\hat{F}(\sigma)}{W(\sigma)} = \frac{F(\sigma)}{\kappa(\sigma)A(\sigma)}$$

An alternative way to view footprint growth is as normalized footprint, i.e., average footprint per load. Note that the final equation form does not depend on window classes.

6.5 Access Patterns

Many applications tend to frequently alternate between regular execution phases with structured memory access patterns and irregular phases with unpredictable memory behaviors.

a) Footprint access diagnostics: To highlight large differences in expected access latencies, we decompose footprint into strided (prefetchable) and irregular (non-prefetchable) access components. The footprint categories are called footprint access diagnostics and represent the most common patterns that affect memory performance. Specifically, we define the following metrics: strided and irregular portion of footprint (`Fstr, Firr`), their growth rate (`ΔFstr, ΔFirr`), fraction of footprint growth due to them (`ΔFstr%, ΔFirr%`), and fraction of constant loads (`Aconst%`). This analysis is constant time per operation, without any pattern analysis, using the load classes of described above.

b) General irregularity: For a more general measure of irregularity, we use spatio-temporal reuse distance (above).

7.0 MemGaze Evaluation

This section evaluates metric accuracy, time overhead, and space reduction. The next section provides case studies.

Platform: To evaluate MemGaze we use an Intel Pentium Silver J5005 (Gemini Lake) CPU with four cores and 16GB memory. (We are procuring an Alder Lake machine, but it is not yet available.)

Benchmarks: We use a set of microbenchmarks and application benchmarks. The microbenchmarks simulate accesses to both dense and sparse data structures and vary access patterns, data reuse, access sparsity, and access likelihood. Microbenchmarks test analysis of (very) short-lived access sequences that become hotspots (repeated 100 times). The microbenchmark names give access patterns using "str" (strided with stride step) and "irr" (irregular). The different access patterns can be composed conditionally ('/') or in series ('|'). For applications we use the graph benchmarks miniVite (Ghosh et al. 2018) (Louvain Community Detection) and GAP (Beamer, Asanovic, and Patterson 2015) (Connected Components and Page Rank); and the machine learning Darknet (Redmon 2015). The graph benchmarks test analysis of reuse and access patterns within load-intensive applications that include highly irregular accesses. We vary compiler optimization levels (O0 vs. O3). For graph benchmarks, we also vary data structure implementations and algorithms. For Darknet, we vary network types for inferencing on images.

All application benchmarks support OpenMP and are executed with and without parallelism. However, note that our analysis focuses on memory behavior and is *orthogonal* to CPU parallelism. (Future work includes exploring the relationship between CPU concurrency and memory systems.)

Sampling configuration: For microbenchmarks, we use a small period (10K loads) and a large buffer (16 KiB yielding ≈1150 addresses) to capture short-lived phenomena. The miniVite and GAP benchmarks use a larger (typical) period (10M and 5M loads, respectively) and smaller buffer (8 KiB yielding ≈500 addresses). The reason buffers do not yield the expected addresses (size / 8 bytes) is due to the suboptimal kernel support (buffer fill and flushes occur asynchronously with the sampling trigger that record the next buffer).

7.1 Validation of data reuse analysis

This section validates data reuse analysis of sampled traces. Figure 8 shows results for each microbenchmark and graph benchmark. The data series represent footprint access diagnostics: footprint (`F`), irregular footprint (`Firr`) and strided footprint (`Fstr`). We exclude reuse distance (`D`) as we prefer intra-sample calculations. The three series for MAPE show mean absolute percentage error over different *trace windows*, i.e., metric histograms with power-of-2 windows. The three series for percentage error represent *code windows*, i.e., aggregated samples for functions, which reduce error.

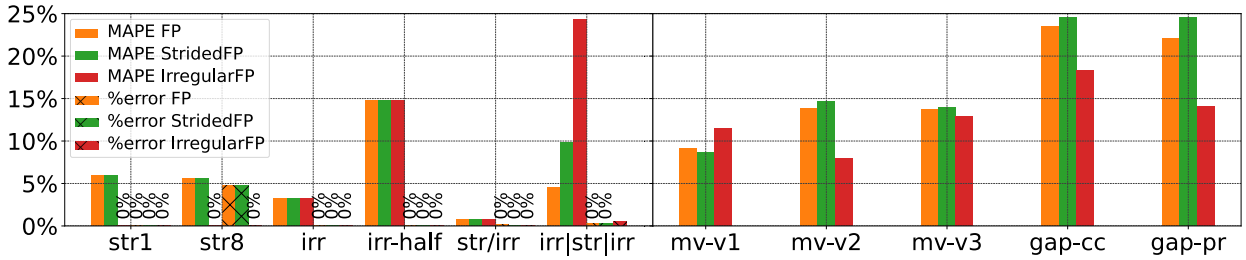

Figure 8 Validating Sampled Footprint Access Diagnostics

For microbenchmarks, we validate against *full* traces. For graph benchmarks, we validate against sampling data that is 10× more frequent. Collecting full traces for the graph benchmarks is problematic for three reasons. First, using PT is not feasible: the data copy rate between PT's pinned kernel buffer and user memory is too high for real-time, resulting in random drops of 30–50%. (CPU frequency throttling makes little difference.) Second, when we attempted to collect full trace data with our validation tool, we found it would take days per benchmark. Finally, full traces are huge (order 1 TB), which requires substantial resources to process per benchmark.

For *trace windows* (first three series), MAPE is <25%. Errors tend to be higher for benchmarks that are more irregular or that include more short-lived behavior, which is expected. For *code windows* (second three series), error <5%. In general, we expect code windows to be more accurate than trace windows because they accumulate more samples.

The errors within trace windows can be understood as follows. First, as expected, errors are *quantitative* overestimates rather than *qualitative*. Second, from our analysis of reuse intervals, we know traces have some blind spots. Third, with benchmarks that include irregular and short-lived behavior, we expect that characterizing trace windows from samples will include error within the reuse analysis (captures C and survivals S).

It should be possible to automatically detect most under- sampling by analyzing sample density and forming confidence intervals. One could flag regions with insufficient samples.

7.2 Time overhead

Recall that MemGaze's toolchain consists of (1) binary instrumentation, (2) memory tracing, and (3) analysis (Figure 3). We evaluate each step. Figure 9 focuses on memory tracing while Table 2 shows sample times for the first and third.

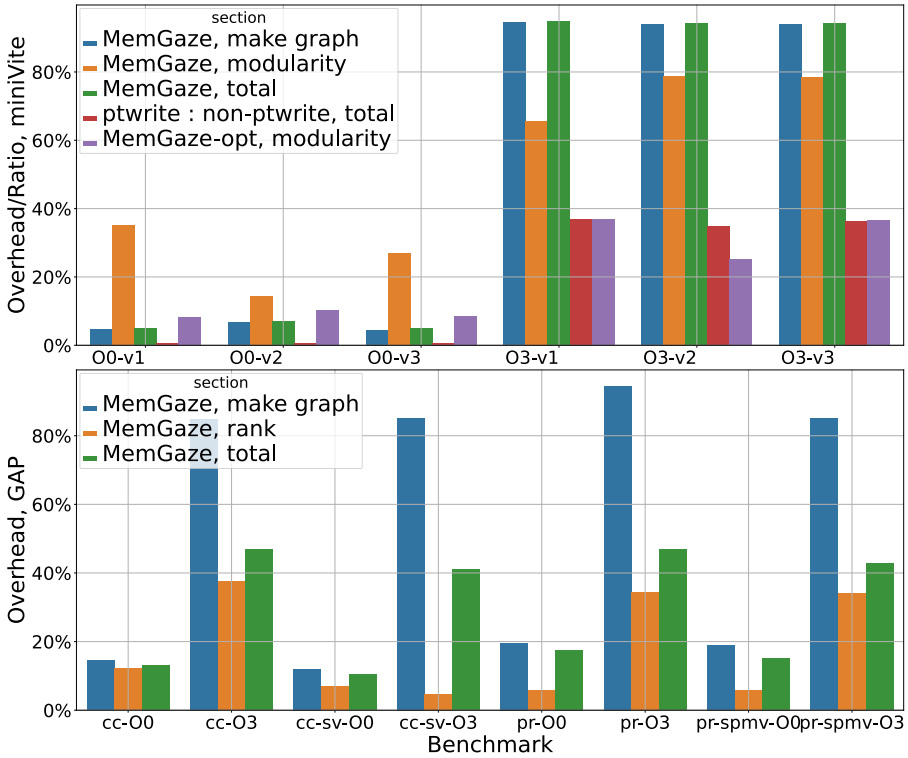

Figure 9 Time Overhead for Memory Tracing. Top show miniVite; bottom, GAP

1) *Memory Tracing*: Our primary interest is memory tracing because low overhead implies (a) fewer timing disturbances (which enables analysis of task interleaving and memory parallelism) and (b) smaller trace sizes, which improves analysis time. Figure 9 shows MemGaze's run time overhead. We show results for two versions of MemGaze. MemGaze is the full implementation that relies on suboptimal kernel support where PT runs continuously. MemGaze-opt is proof-of-concept (in user-space) that enables Processor Tracing only during samples. The figure also breaks down overhead by application phase. The miniVite and GAP benchmarks include a graph generation phase that has distinctly different memory accesses than the second phases ('modularity' and 'rank').

The first three series of both figures show per-phase and total overhead for MemGaze. Even with sub-optimal driver support, MemGaze's overhead is typically 10–95%. Overhead is higher with more compiler optimization (O3) because (a) the rate of instrumented loads is higher and (b) the total run time is smaller (making it harder to amortize overhead). We hypothesize DarkNet's overhead of 5×–7× is due to `ptwrite` interfering with its much higher store rate. The fourth (red) series illustrate (a) using the ratio of `ptwrites` to non-`ptwrite` instructions. The ratio highly correlates with total overhead. The MemGaze-opt series (fifth, purple) reduces overhead to the near-minimum for our scheme by enabling PT only when sampling miniVite/modularity, a load intensive hotspot. MemGaze-opt reduces overhead from 80% to less than 40%, which is very close to the execution rate of `ptwrite` instructions. This makes sense because `ptwrites` are expensive to decode and trigger data copies (Linux perf 2021b). It may be possible to further reduce overhead with 32-bit packets and additional compression that reduces `ptwrites` for Strided loads. In contrast, current methods for tracing sequences of memory

addresses easily incur time overhead of 100× (Xiang et al. 2013), (Kilic, Tallent, and Friese 2020).

2) *Binary Instrumentation and Analysis*: Table 2 shows sample run times for MemGaze's instrumentation ('Instrument') and analysis ('Analysis/1' + 'Analysis/2') steps. These times are 'extreme worst cases' in that neither step has been optimized or parallelized. Even so, we can easily analyze memory intensive applications that are unfeasible to analyze using full traces.

As MemGaze's binary instrumentor takes a binary to analyze and instrument, its run time depends on the size and complexity of the binary. The different run times between microbenchmarks and miniVite show the effect of miniVite's much larger binary. The different run times between miniVite and PR show the effect of PR's increased routine complexity.

Analysis time is divided into two sub-steps: trace building (Analysis/1) and trace analysis (Analysis/2). The building step is from converting Linux `perf's` trace into one for our trace analysis. The run times of both depend on trace size (which is shown in Table III).

Table 2 Time Overhead: Binary Instrumentation & Analysis

Benchmark	Binary Size	Instrument	Analysis/1	Analysis/2
μbenchmarks	19 kB	1.1s	76.1s	79.2s
miniVite-O3-v1	1900 kB	135.5s	357.0s	284.5s
GAP pr-O3	95 kB	144.2s	37.4s	22.9s
GAP cc-O3	100 kB	122.1s	32.3s	21.5s
Darknet-AlexNet	2700 kB	83.2s	87.4s	21.1s
Darknet-ResNet	2700 kB	83.2s	898.2s	289.5s

7.3 Space reduction

This section evaluates the space savings of sampled memory traces. Table 3 compares MemGaze's traces against the sizes of full memory access traces ('Full'). The table shows three versions of a 'Full' trace.

- 'Rec' shows a compressed full trace that suffers data loss because of unpredictable throttling and data drop, due to inability to copy between PT's pinned kernel buffer and user memory.
- 'All' shows a *compressed* full trace without data drops by correcting based on `perf's` 'DROP' records.
- 'All+' show the full size *without trace compression*, i.e., by including the suppressed Constant loads.

The 'MemGaze' column shows size of a sampled, compressed trace. The 'Ratio' column shows the ratio as a *percentage* between MemGaze and the corresponding 'Full' size. A glance at the

'All' or 'All+' data shows that a straightforward use of `ptwrite` would generates huge traces — if it were possible to do without significant drops. Note that, except for the microbenchmarks, we did not use the 'Rec' traces for validation because drops are most likely to occur in the most load-intensive regions, which is exactly where we want to validate. (We captured a full trace by inserting OS sleeps after each load to substantially reduce the load rate.) Comparing the 'All' and 'All+' shows that our compression method gives an average of 1.2× and 2× space savings for compiler optimization levels O3 and O0, respectively.

Clearly, the major benefit of MemGaze's trace size comes from sampling. MemGaze's trace size is generally around 1% of the full trace for applications with O3 optimization. The size is controllable by changing the sample buffer size and the sampling period. In general, 'All' traces are proportional to the number of non-Constant loads in the execution, where loads with two source registers take twice as much space. In contrast, MemGaze's traces are proportional to the product of the number of samples, $|\sigma|$, and the sample buffer size, where $|\sigma|$ is the total executed loads divided by sample period $w + z$. Finally, trace sizes are independent of the analysis methodology adopted.

Table 3 Space Savings of MemGaze's Memory Traces

Benchmark	Full (GB)			MemGaze	Ratio (%)		
	Rec	All	All+	(MB)	Rec	All	All+
all μbench-O0 (1×)	1.9	1.9	3.5	63	3.3	3.3	1.8
all μbench-O3 (1×)	1.9	1.9	1.91	20	1.1	1.1	1
all μbench-O3	112	112	113	865	0.8	0.8	0.7
miniVite-O0-v1	77	163	316.5	1620	2.1	0.9	0.5
miniVite-O0-v2	71	198	387.9	1697	2.4	0.9	0.4
miniVite-O0-v3	79	150	292.7	1660	2.1	1.1	0.6
miniVite-O3-v1	19	41	41.1	310	1.6	0.8	0.7
miniVite-O3-v2	22	43	54.9	310	1.4	0.7	0.6
miniVite-O3-v3	13	23	29.4	341	2.6	1.5	1.1
GAP-cc-O0	2.3	3.4	6.6	355	15.4	10.4	5.3
GAP-cc-O3	4.9	7.9	9.5	31	0.6	0.4	0.3
GAP-cc-sv-O0	4.4	6.4	12.5	377	8.6	5.9	3
GAP-cc-sv-O3	6.7	10.8	13	35	0.5	0.3	0.3
GAP-pr-O0	5.1	7.5	14.6	377	7.4	5.0	2.5
GAP-pr-O3	5.4	7.9	9.5	35	0.7	0.4	0.4
GAP-pr-spmv-O0	6.3	8.9	17.4	385	6.1	4.3	2.2
GAP-pr-spmv-O3	6.5	10.1	12.1	36	0.6	0.4	0.3
Darknet-AlexNet	4.6	11.2	16.9	71	1.6	0.6	0.4
Darknet-ResNet	29	59	66	748	2.6	1.3	1.2

8.0 MemGaze Case Studies

This section provides case studies for miniVite (Ghosh et al. 2018) (Louvain Community Detection), Darknet (Redmon 2015), and GAP (Beamer, Asanovic, and Patterson 2015) (Connected Components and Page Rank). Effective analysis requires an understanding of access frequency (e.g., accesses are costly), spatio-temporal reuse (e.g., for cache performance) and access patterns (e.g., strided accesses leverage prefetching). Our analyses therefore combine time-centric and location-centric results

8.1 miniVite

miniVite is a benchmark (Ghosh et al. 2018) for Louvain Community Detection. We use three variants that show effects of different hash table implementations (map object) on a hotspot that inspects (`buildMap`) neighboring communities for each vertex. v1 uses a C++ `unordered_map`. As an open hash table — an array of keys, each containing a linked list for items — it creates irregular accesses. v2 and v3 use TSL `hopscotch` (Tessil 2019), (Herlihy, Shavit, and Tzafrir 2008) a closed hash table that replaces many irregular accesses with strided ones. v2 uses the default table size. v3 right-sizes each table instance — there are many — to avoid dynamic resizing.

Time and location analyses for O3 variants are shown in Table 4 and Table 5, respectively. The hotspot analysis clearly highlights `buildMap` and the map's logical insert function. It also highlights `getMax`, which uses map. Runtime for each variant is shown to the right and indicate increasing improvement between v1, v2, and v3, showing that the different hash tables configurations have an effect. The location analysis highlights three hot regions used within `buildMap`: the map object; a vector of remote edges for local graph vertices; and other objects in `buildMap` coming from the caller. The region's size corresponds to allocation sizes; and accesses/block corresponds to block hotness. We discuss results by variant. Although v1 has the fewest accesses ($\mathcal{A}$), it has poor cache behavior due to irregular accesses. This is shown in Table 4 with the highest footprint growth (ΔF) and highest percentage of irregular accesses (lowest $\Delta F_{str}\%$). Table 5 shows that spatial-temporal reuse distance (D) is highest or high. We created v2 to address the poor cache behavior. v2's closed hash table results in repeated (strided) traversals, indicated by an increase in $\Delta F_{str}\%$, that leverage hardware prefetchers to hide more memory latency. The access pattern also improves (lowers) footprint growth (ΔF) and tends to improve (lower) average reuse distance (D). The trade-off for the better performance and contiguous accesses is that v2 and v3's closed design uses more memory. In contrast v1's open hash table grows only as large as needed as suggested by footprint (F). A defect with v2 is that it significantly increases accesses: see $\mathcal{A}$ for `map.insert` and `A` for map. v3's right-sized closed hash tables avoid extra accesses from resizing (copies) and searches (over-allocation). The result reduces accesses compared to v2 but retains the benefits of strided accesses. Interestingly, because accesses decrease so much, ΔF for `map.insert` increases. v3's smaller sizes and

contiguous accesses generally improve reuse distance (`D`). Our analysis uncovers a common tendency: *Sparse structures have smaller footprint but more irregular access patterns, whereas dense structures have larger footprints but more regular access patterns. Further, several locality metrics are important for capturing a complete picture.*

Table 4 Data Locality of Hot Function Accesses

Function	Variant	*F*	Δ*F*	$F_{str\%}$	A
buildMap (make *map*)	v1	2.3G	0.156	66.4	291K
	v2	2.1G	0.151	66.9	273K
	v3	2.1G	0.160	66.8	270K
map.insert	v1	>0.7G	0.011	73.3	106K
	v2	2.4G	0.003	93.7	318K
	v3	0.5G	0.009	92.8	67.8K
getMax (use *map*)	v1	0.4G	0.150	0.5	44.7K
	v2	1.3G	0.040	98.4	182K
	v3	1.5G	0.040	97.8	194K

v1	8.60 s
v2	5.15 s
v3	3.88 s

Table 5 Spatio-Temporal Reuse of Hot Memory (64B Block)

Object	Variant	Reuse (*D*)	# blocks	*A*	*A* / block
map (hash table)	v1	2.65	768	55K	71.9
	v2	2.79	768	119K	155.2
	v3	1.97	768	85K	111.3
remote edges of local vertices	v1	8.71	4864	24K	4.9
	v2	4.90	4864	19K	3.9
	v3	3.32	4864	19K	3.9
other objs in buildMap (from caller)	v1	0.37	104K	19235	0.2
	v2	0.15	101K	21362	0.2
	v3	0.24	110K	22306	0.2

8.2 Darknet AlexNet & ResNet

Darknet (Redmon 2015) is an open neural network framework written in C/OpenMP. We analyze image classification (inferencing) with two pre-trained models for AlexNet (Krizhevsky, Sutskever, and Hinton 2012) and ResNet152 (He et al. 2016) for a single image.

Time and location analysis for the hottest kernels (`gemm` and `im2col`) are shown in Table 6 and Table 7 respectively. To capture differences between the neural network layers, Table 8 compares `gemm` across memory access intervals. Table 6 characterizes the primary hotspot, `gemm`, matrix multiplication specialized for neural networks. It dominates total footprint (> 90%) and nearly all accesses are strided, as expected. The location analysis (Table 7) independently highlights `gemm` matrices as the primary hotspot. `gemm` takes matrices **A** (M ×K) and **B** (K ×N)

and produces **C** (M ×N). Due to memory allocator decisions, all AlexNet's gemm matrices are in a single region whereas ResNet152's hot region contains only **B**.

Interestingly, we can compare `gemm`'s behavior between AlexNet and ResNet152. The differences in footprint (`F`) and footprint growth (`ΔF`) between AlexNet and ResNet152 correspond to differences in number and type of layers, which result in different combinations of matrix multiplication input sizes. The differences in reuse distance (Table 7) correspond to the different matrices in the hot regions. Due to the loop order in `gemm`, there is long-term reuse of **B** that is unlikely to occur within a sample that intra-sample reuse distance (our current interest) will not capture.

Table 8 shows how `gemm`'s locality metrics differ across access intervals over time. There are three observations. First, AlexNet's `ΔF` varies more than ResNet152's. This corresponds to AlexNet's varying convolutional, fully connected, and pooling layers compared to ResNet's more consistent convolutional structure. Second, for both networks, spatio-temporal reuse distance (`D`) *over all objects* increases over time. This corresponds to a decrease in dimension N (gemm's innermost loop) as the networks synthesize data from higher-level feature filters; and for AlexNet N decreases *very* rapidly. Finally, observe that, in contrast to AlexNet, ResNet152's footprint growth (`ΔF`) tends to decrease over time. This correlates with the changes in dimensions N (decrease) and K (smaller increase), corresponding to `gemm`'s inner two loops.

Our analysis captures memory behavior of the well-known matrix multiplication kernel by time, location, and across time-location. These differing perspectives are critical for capturing a complete picture.

We evaluated optimizations for `gemm` kernel. Without per-layer specialization, the current algorithm uses correct loop ordering and appropriate inner-loop reuse via unrolling. We do not expect tiling to be effective because the matrices are relatively small, and it would significantly complicate outer-loop parallelism.

Table 6 Darknet: Data Locality of Hot Function Accesses

Function	Model	*F*	Δ*F*	$F_{str\%}$	A
gemm	AlexNet	69M	0.113	100	1.2M
	ResNet152	3855M	0.478	100	2.6M
im2col	AlexNet	3.4M	0.138	100	0.05M
	ResNet152	244M	0.813	100	0.09M

Table 7 Darknet: Spatio-Temporal Reuse of Hot Memory (64B Block)

Object	Model	Reuse (*D*)	# blocks	*A*	*A* / block
gemm's **A**,**B**,**C**	AlexNet	0.76	66048	977K	14.8
gemm's **B**	ResNet152	0.01	38400	598K	15.6
hot region in	AlexNet	1.87	8192	167K	20.4
im2_col	ResNet152	2.54	3328	7K	1.9

8.3 GAP's Connected Components & Page Rank

We study the memory effects of different algorithms for each of GAP's PageRank (PR) and ConnectedComponents (CC) specifications. For PR we use pr (Gauss-Seidel) and pr-spmv (Jacobi-style). For CC we use cc (Afforest) and cc-sv (Shiloach-Vishkin). In both cases, the former represents an optimization of the latter. We run each with the same graph size of 2^{22}, i.e., 4 M vertices and 64 M edges. We ignore graph building and focus on the respective algorithm. Location and time analyses are shown in Table 9 and Figure 10 & Figure 11. The time analysis is a histogram plot showing data locality (average) with respect to hot access interval size.

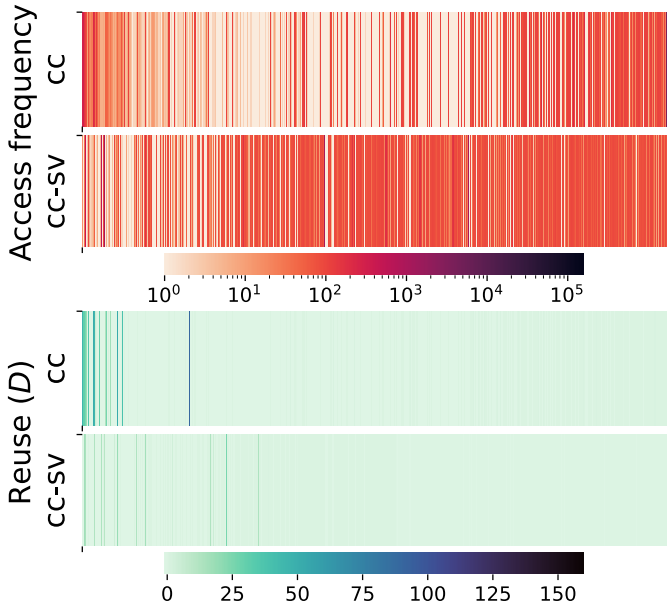

Figure 10 GAP: Distribution of Spatio-Temporal Metrics for Hot Memory

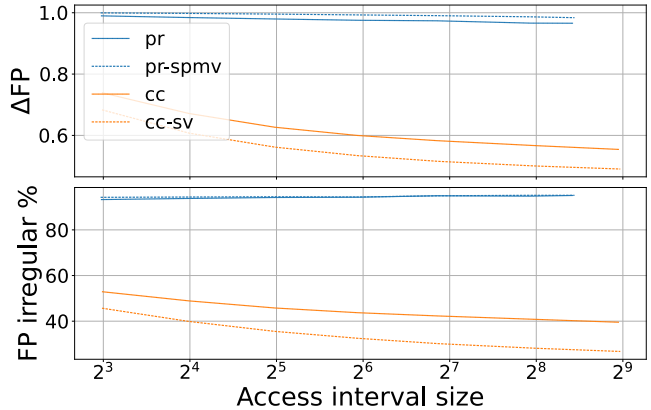

Figure 11 GAP: Data Locality of Hot Access Intervals (intra-sample)

Table 8 Darknet/GEMM: Data Locality over Time of Hot Access Intervals

Access	AlexNet				ResNet152			
Interval	*F*	Δ*F*	*D*	A	*F*	Δ*F*	*D*	A
0	28M	0.475	0.01	30K	639M	0.747	0.47	286K
1	55M	0.675	0.02	30K	772M	0.799	0.57	293K
2	89M	0.983	0.02	25K	640M	0.617	2.71	302K
3	64M	0.794	0.14	26K	620M	0.599	2.62	304K
4	39M	0.489	1.64	29K	591M	0.574	2.69	302K
5	55M	0.627	1.66	26K	638M	0.618	2.65	302K
6	41M	0.493	1.66	29K	648M	0.625	2.63	304K
7	38M	0.644	1.49	17K	549M	0.514	2.66	312K

PR's results show the effects of pr's optimized algorithm. With both PR variants, Table 9 shows that the hot memory object is `o-score`, representing intermediate contributions to each vertex's final score (rank). With pr-spmv, updates to `o-score` are saved until the next iteration whereas with pr, updates occur immediately. As a result, Table IX shows that pr's spatio-temporal reuse distance (`D`) is noticeably smaller (better); and its footprint growth (`ΔF`) slightly smaller. The pr variant requires fewer total iterations and, hence, accesses (`A`).

Table 9 Spatio-Temporal Reuse of Hot Memory (64B Block)

Object	Algorithm	Reuse (*D*)	Max *D*	*A*	*A*/block	Time
o-score	pr	1.13	152	64K	0.76	57.2 s
o-score	pr-spmv	2.41	132	82K	1.14	80.1 s
cc	cc	5.21	154	581K	8.87	2.7 s
cc	cc-sv	0.83	36	476K	8.65	45.5 s

CC's results show the value of detailed location analysis for understanding cc's optimized algorithm. The cc variant uses subgraph sampling (Sutton, Ben-Nun, and Barak 2018), which requires more accesses (`A`) but can improve performance for NUMA parallelism vs. cc-sv. Interestingly, a subset of locality metrics could suggest that cc-sv has better locality: cc has higher average reuse distance (`D`), footprint growth (`ΔF`), and irregular footprint ratio (`F`$_{irr}$`%`). However, Figure 10's detailed heatmaps help explain the difference. The heatmaps show the distributions of access frequencies and reuse distances (`D`), where darker is higher. For accesses, cc has fewer and smaller dark bands, indicating more access locality than cc-sv. For reuse distance, the heatmaps show that the table's summary metrics are affected by outliers: the average behavior is relatively similar. *Thus, to understand CC's differences, it is important to understand memory behavior from many angles and at different resolutions.*

9.0 MemGaze Related Work

MemGaze provides low-overhead, high-resolution memory analysis with load-level, sequence-aware analysis of data reuse. High-resolution memory reuse. Current methods for collecting load-level sequences of memory accesses use software-based instruction instrumentation, either with compilers (Beyls and D'Hollander 2006) or binary instrumentation (Xiang et al. 2013), (Marin, Dongarra, and Terpstra 2014). (Xiang et al. 2013) sample address sequences by enabling and disabling dynamic instruction tracing (Luk et al. 2005) which results in time overheads of 100×. For more detailed access analysis memory modeling tools (Carlson, Heirman, and Eeckhout 2011) and simulators (Binkert et al. 2011), (Li et al. 2020), (Kim, Yang, and Mutlu 2015) can be used. Low-overhead reuse interval analysis. One class of well- known low-overhead methods for analyzing memory accesses target reuse intervals, or the time between two accesses to the same address. Reuse intervals can be converted into reuse distance (Shen et al. 2007). (Eklov and Hagersten 2010) sample reuse intervals, measured in loads, using virtual memory traps. For lower overhead, debug registers can be used to sample reuse intervals (Wang, Liu, and Chabbi 2019), (Sasongko et al. 2021). Since there are only 4 such registers, this technique obtains a uniform sample using reservoir sampling. (W. Zhao et al. 2011) disable expensive memory tracking of reuse distances when a program's working set size is stable. In contrast, we sample address sequences rather than reuse intervals, which enables analysis of patterns and sequences in addition to reuse. Further, the incorporation of static analysis to select instrumentation points immediately suggests a variety of extensions to specialize for different use. Various methods for reuse distance analysis have been designed for storage systems (Waldspurger et al. 2015), (Wires et al. 2014). The SHARDS method uses spatial sampling (Waldspurger et al. 2015), or monitoring of sampled portions of the address space, which is like StatStack (Eklov and Hagersten 2010). Other low-overhead methods. To capture interaction effects between instructions, ProfileMe monitoring hardware (Dean et al. 1997), (Fields et al. 2004) could monitor pairs of instructions. Our work extends this to sequences and data reuse. Another low-overhead method for analyzing data reuse in memory is to use CPU performance monitoring units to collect sparse address sets or monitor cache behavior. For example, AMD's IBS (or LWP) (AMD 2021, 64) and Intel's PEBS (Intel Corporation 2020, 64) can collect data addresses. However, the results are so sparse that detecting reuse, much less access patterns, is difficult (Roy et al. 2018), (Liu and Mellor-Crummey 2014). Another method estimates data reuse by inferring bounds on footprint through profiles of accesses to each memory hierarchy level (Kilic, Tallent, and Friese 2020). This method can also make some inferences about access patterns. However, the footprint estimates are coarse and qualitative. In contrast, MemGaze enables far more resolution for both data reuse and access patterns. Thread-level data reuse can be captured by monitoring OS-level virtual memory events and thread interactions. NumaPerf (X. Zhao et al. 2021), couples this OS-level data with source-code compiler analysis in order to diagnose NUMA-related performance problems. Performance counters for accesses (loads) and cache locality (cache misses) are frequently used for system adaptation (Chen et al. 2020) or compiler feedback (Oh et al. 2021). Analyzing data reuse. Snir et al.(Snir and Yu 2005) establish limits on data reuse analysis. Yuan et al.(Yuan et al. 2019) provide an excellent overview of temporal data locality of scalar accesses within a trace. Our analysis extends the latter in two ways. First, footprint access patterns provide information on spatial locality and

expected behavior of an access stream or dynamic access sequence. Second, we consider static access sequences. We describe multi-resolution analysis for footprint (volume), footprint growth (rate), and reuse distance; and decompose each by access pattern (strided vs. irregular). Our analysis differs from Xiang et al.'s footprint analysis of sampled traces (Xiang et al. 2013), (Yuan et al. 2019). The latter's analysis is based on reuse intervals and therefore samples access sequences of varying lengths. In contrast, we sample constant-size sequences of accesses and therefore avoid long reuse intervals (sequences). As mentioned, we also identify common access patterns for insight into spatial locality and expected behavior. Prior work has combined analysis of spatial locality and temporal locality (Weinberg et al. 2005), (Anghel et al. 2013). However, this work relies on static analysis or coarse (virtual memory) page-level statistics.

10.0 MemGaze Conclusion

The time and space costs (≈100×) of prior methods for sequence-aware analysis of data reuse make them unattractive or unfeasible for many working sets or execution scenarios. We show that Processor Tracing (PT) can be an effective technique for low-overhead, high-resolution memory analysis that includes locations vs. operations, accesses vs. spatio-temporal reuse, and reuse (as distance, rate, and volume) vs. access patterns. We demonstrate load-level analysis of applications, not kernels, because both trace size and resolution are controllable. Using sampled traces that are ≈1% of full ones, we elucidate the memory effects of different data structures and algorithms, including explaining performance differences between prefetching and irregular accesses with good spatio-temporal locality. We show that with a straightforward driver optimization, time overhead for collecting access traces is 10– 35% on memory intensive regions and highly correlates with executed `ptwrites`. We plan to leverage our work for hardware/software co-design. Using models of different memory systems, we can obtain insight into memory system performance and concurrency with respect to data location, data movement, and workload accesses. We also believe our work motivates further attention to hardware/software co-design of Performance Monitoring Units (PMUs). `ptwrite` functionality generalizes the notion of 'state gathering' when using Performance Monitoring Units and could be used to analyze arbitrary execution state. Further, `ptwrite` generalizes the notion of sequence, permitting both, points (length 1) or true sequence analysis, something that has not thus far been generally possible without binary instrumentation. Finally, `ptwrite` can substantially reduce the number of OS interrupts needed to gather state from PMU registers if coupled with judicious (e.g., sampled, masked, or predicated) execution.

11.0 References

AMD. 2021. “AMD64 Architecture Programmer’s Manual Volume 2: System Programming, 3.37.” https://www.amd.com/system/files/TechDocs/24593.pdf.

Amdahl, Gene M. 1967. “Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities.” In *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*, 483–85.

Anghel, Andreea, Gero Dittmann, Rik Jongerius, and R Luijten. 2013. “Spatio-Temporal Locality Characterization.” In *1st Workshop on Near-Data Processing*, 1–5.

ARM. 2011. “Embedded Trace Macrocell Architecture Specification.” https://documentation-service.arm.com/static/5f90158b4966cd7c95fd5b5e.

———. 2021. “ARM Architecture Reference Manual.” https://documentation-service.arm.com/static/.

Badr, Mario, Carlo Delconte, Isak Edo, Radhika Jagtap, Matteo Andreozzi, and Natalie Enright Jerger. 2020. “Mocktails: Capturing the Memory Behaviour of Proprietary Mobile Architectures.” In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, 460–72. IEEE.

Beamer, Scott, Krste Asanovic, and David A. Patterson. 2015. “The GAP Benchmark Suite.” *CoRR* abs/1508.03619. http://arxiv.org/abs/1508.03619.

Beyls, Kristof, and Erik D’Hollander. 2001. “Reuse Distance as a Metric for Cache Behavior.” In *Proceedings of the IASTED Conference on Parallel and Distributed Computing and Systems*, 14:350–60. Citeseer.

Beyls, Kristof, and Erik H. D’Hollander. 2006. “Intermediately Executed Code Is the Key to Find Refactorings That Improve Temporal Data Locality.” In *Proc.\ of the 3rd Conf. on Computing Frontiers*, 373–82. CF ’06. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1128022.1128071.

Binkert, Nathan, Bradford Beckmann, Gabriel Black, Steven K. Reinhardt, Ali Saidi, Arkaprava Basu, Joel Hestness, et al. 2011. “The Gem5 Simulator.” *SIGARCH Comput. Archit. News* 39 (2): 1–7. https://doi.org/10.1145/2024716.2024718.

Carlson, Trevor E, Wim Heirman, and Lieven Eeckhout. 2011. “Sniper: Exploring the Level of Abstraction for Scalable and Accurate Parallel Multi-Core Simulation.” In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–12.

Chen, Yu, Ivy B. Peng, Zhen Peng, Xu Liu, and Bin Ren. 2020. “ATMem: Adaptive Data Placement in Graph Applications on Heterogeneous Memories.” In *Proc. of the 18th ACM/IEEE Intl. Symp on Code Generation and Optimization*, 293–304. CGO 2020. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3368826.3377922.

Choi, Jinyoung, Sergey Blagodurov, and Hung-Wei Tseng. 2021. “Dancing in the Dark: Profiling for Tiered Memory.” In *2021 IEEE International Parallel and Distributed Processing Symposium*, 13–22. https://doi.org/10.1109/IPDPS49936.2021.00011.

Dean, Jeffrey, James E. Hicks, Carl A. Waldspurger, William E. Weihl, and George Chrysos. 1997. “ProfileMe: Hardware Support for Instruction-Level Profiling on Out-of-Order Processors.” In *Proc.\ of the 30th Annual ACM/IEEE Intl. Symp. on Microarchitecture*, 292–302. Washington, DC, USA: IEEE Computer Society.

Eklov, D., and E. Hagersten. 2010. “StatStack: Efficient Modeling of LRU Caches.” In *IEEE Intl. Symp. on Performance Analysis of Systems Software*, 55–65. https://doi.org/10.1109/ISPASS.2010.5452069.

Esmaeilzadeh, H., E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger. 2011. “Dark Silicon and the End of Multicore Scaling.” In *2011 38th Annual International Symposium on Computer Architecture (ISCA)*, 365–76.

Fields, Brian A., Rastislav Bodik, Mark D. Hill, and Chris J. Newburn. 2004. “Interaction Cost and Shotgun Profiling.” *ACM Trans. Archit. Code Optim.* 1 (3): 272–304. https://doi.org/10.1145/1022969.1022971.

Ghosh, Sayan, Mahantesh Halappanavar, Antonino Tumeo, Ananth Kalyanaraman, and Assefaw H Gebremedhin. 2018. “MiniVite: A Graph Analytics Benchmarking Tool for Massively Parallel Systems.” In *Proc. of the 9th Intl. Workshop on Performance Modeling, Benchmarking, and Simulation of High-Performance Computer Systems*.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. “Deep Residual Learning for Image Recognition.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–78.

Herlihy, Maurice, Nir Shavit, and Moran Tzafrir. 2008. “Hopscotch Hashing.” In *Distributed Computing*, edited by Gadi Taubenfeld, 350–64. Berlin, Heidelberg: Springer Berlin Heidelberg.

IBM. 2018. “POWER9 Performance Monitor Unit User’s Guide.” https://wiki.raptorcs.com/w/images/6/6b/POWER9_PMU_UG_v12_28NOV2018_pub.pdf.

Intel Corporation. 2013. “Intel Processor Tracing.” https://software.intel.com/content/www/us/en/develop/blogs/processor-tracing.html.

———. 2020. “Intel 64 and IA-32 Architectures Software Developer’s Manual, Combined Volumes: 1, 2A, 2B, 2C, 2D, 3A, 3B, 3C, 3D and 4.”

———. 2021. “Intel Architecture Instruction Set Extensions and Future Features (319433-044).” https://software.intel.com/content/www/us/en/develop/download/intel-architecture-instruction-set-extensions-programming-reference.html.

Kilic, Ozgur O., Nathan R. Tallent, and Ryan D. Friese. 2020. “Rapid Memory Footprint Access Diagnostics.” In *Proc. of the 2020 IEEE Intl. Symp. on Performance Analysis of Systems and Software*. IEEE Computer Society.

Kim, Yoongu, Weikun Yang, and Onur Mutlu. 2015. “Ramulator: A Fast and Extensible DRAM Simulator.” *IEEE Computer Architecture Letters* 15 (1): 45–49.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” *Advances in Neural Information Processing Systems* 25: 1097–1105.

Li, Shang, Zhiyuan Yang, Dhiraj Reddy, Ankur Srivastava, and Bruce Jacob. 2020. “DRAMsim3: A Cycle-Accurate, Thermal-Capable DRAM Simulator.” *IEEE Computer Architecture Letters* 19 (2): 106–9.

Linux perf. 2021a. “Perf-Mem.” http://man7.org/linux/man-pages/man1/perf-mem.1.html.

———. 2021b. “Perf-Intel-Pt.” https://man7.org/linux/man-pages/man1/perf-intel-pt.1.html.

Liu, Xu, and John Mellor-Crummey. 2014. “A Tool to Analyze the Performance of Multithreaded Programs on NUMA Architectures.” *SIGPLAN Not.* 49 (8): 259–72. https://doi.org/10.1145/2692916.2555271.

Luk, Chi-Keung, Robert Cohn, Robert Muth, Harish Patil, Artur Klauser, Geoff Lowney, Steven Wallace, Vijay Janapa Reddi, and Kim Hazelwood. 2005. “Pin: Building Customized Program Analysis Tools with Dynamic Instrumentation.” In *Proc. of the 2005 ACM SIGPLAN Conf. on Programming Language Design and Implementation*, 190–200. New York, NY, USA: ACM. http://doi.acm.org/10.1145/1065010.1065034.

Marin, G., J. Dongarra, and D. Terpstra. 2014. “MIAMI: A Framework for Application Performance Diagnosis.” In *Proc. of the 2014 IEEE Intl. Symp. on Performance Analysis of Systems and Software*, 158–68. https://doi.org/10.1109/ISPASS.2014.6844480.

Mattson, R. L., J. Gecsei, D. R. Slutz, and I. L. Traiger. 1970. “Evaluation Techniques for Storage Hierarchies.” *IBM Systems Journal* 9 (2): 78–117. https://doi.org/10.1147/sj.92.0078.

Meng, Xiaozhu, and Barton P Miller. 2016. “Binary Code Is Not Easy.” In *Proc.\ of the 25th Intl. Symp. on Software Testing and Analysis*, 24–35.

Oh, Deok-Jae, Yaebin Moon, Eojin Lee, Tae Jun Ham, Yongjun Park, Jae W. Lee, and Jung Ho Ahn. 2021. “MaPHeA: A Lightweight Memory Hierarchy-Aware Profile-Guided Heap Allocation Framework.” In *Proceedings of the 22nd ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems*, 24–36. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3461648.3463844.

Redmon, Joseph. 2015. “Darknet: Open Source Neural Networks in C.” https://pjreddie.com/darknet.

Roy, Probir, Shuaiwen Leon Song, Sriram Krishnamoorthy, and Xu Liu. 2018. “Lightweight Detection of Cache Conflicts.” In *Proc. of the 2018 Intl. Symp. on Code Generation and Optimization*, 200–213. CGO 2018. New York, NY, USA: ACM. https://doi.org/10.1145/3168819.

Sasongko, Muhammad Aditya, Milind Chabbi, Mandana Bagheri Marzijarani, and Didem Unat. 2021. “ReuseTracker: Fast Yet Accurate Multicore Reuse Distance Analyzer.” *ACM Trans. Archit. Code Optim.* 19 (1). https://doi.org/10.1145/3484199.

Shen, Xipeng, Jonathan Shaw, Brian Meeker, and Chen Ding. 2007. “Locality Approximation Using Time.” In *Proc.\ of the 34th Annual ACM SIGPLAN-SIGACT Symp. on Principles of Programming Languages*, 55–61. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1190216.1190227.

Snir, Marc, and Jing Yu. 2005. “On the Theory of Spatial and Temporal Locality.” UIUCDCS-R-2005-2611, http://hdl.handle.net/2142/11077. University of Illinois at Urbana-Champaign.

Sutton, Michael, Tal Ben-Nun, and Amnon Barak. 2018. “Optimizing Parallel Graph Connectivity Computation via Subgraph Sampling.” In *2018 IEEE International Parallel and Distributed Processing Symposium*, 12–21. https://doi.org/10.1109/IPDPS.2018.00012.

Tessil, Thibaut G. 2019. “Tessil Hopscotch-Map.”

Waldspurger, Carl A., Nohhyun Park, Alexander Garthwaite, and Irfan Ahmad. 2015. “Efficient MRC Construction with SHARDS.” In *13th USENIX Conf. on File and Storage Technologies (FAST 15)*, 95–110. Santa Clara, CA: USENIX Association. https://www.usenix.org/conference/fast15/technical-sessions/presentation/waldspurger.

Wang, Q., X. Liu, and M. Chabbi. 2019. “Featherlight Reuse-Distance Measurement.” In *2019 IEEE Intl. Symp. on High Performance Computer Architecture*, 440–53. https://doi.org/10.1109/HPCA.2019.00056.

Weinberg, J., M.O. McCracken, E. Strohmaier, and A. Snavely. 2005. “Quantifying Locality In The Memory Access Patterns of HPC Applications.” In *Proc.\ of the 2005 ACM/IEEE Conf. on Supercomputing*, 50–50. https://doi.org/10.1109/SC.2005.59.

Wires, Jake, Stephen Ingram, Zachary Drudi, Nicholas J. A. Harvey, and Andrew Warfield. 2014. “Characterizing Storage Workloads with Counter Stacks.” In *11th USENIX Symp. on Operating Systems Design and Implementation*, 335–49. Broomfield, CO: USENIX Association. https://www.usenix.org/conference/osdi14/technical-sessions/presentation/wires.

Wulf, Wm A, and Sally A McKee. 1995. “Hitting the Memory Wall: Implications of the Obvious.” *ACM SIGARCH Computer Architecture News* 23 (1): 20–24.

Xiang, Xiaoya, Chen Ding, Hao Luo, and Bin Bao. 2013. “HOTL: A Higher Order Theory of Locality.” *SIGARCH Comput. Archit. News* 41 (1): 343–56. https://doi.org/10.1145/2490301.2451153.

Yuan, Liang, Chen Ding, Wesley Smith, Peter Denning, and Yunquan Zhang. 2019. “A Relational Theory of Locality.” *ACM Trans. Archit. Code Optim.* 16 (3). https://doi.org/10.1145/3341109.

Zhao, Weiming, Xinxin Jin, Zhenlin Wang, Xiaolin Wang, Yingwei Luo, and Xiaoming Li. 2011. “Low Cost Working Set Size Tracking.” In *Proc.\ of the 2011 USENIX Conf. on USENIX Annual Technical Conference*. USENIXATC’11. USA: USENIX Association.

Zhao, Xin, Jin Zhou, Hui Guan, Wei Wang, Xu Liu, and Tongping Liu. 2021. “NumaPerf: Predictive NUMA Profiling.” In *Proc.\ of the ACM Intl. Conf. on Supercomputing*, 52–62. ICS ’21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3447818.3460361.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov