Mnemo: Boosting Memory Cost Efficiency in Hybrid Memory Systems

Thaleia Dimitra Doudali
Georgia Institute of Technology
thdoudali@gatech.edu

Ada Gavrilovska
Georgia Institute of Technology
ada@cc.gatech.edu

ABSTRACT

Mnemo is an application profiling tool specialized for data serving and caching workloads, which retrieve data from cloud in-memory key-value stores. The increasing demand to boost application performance via in-memory data retrieval and the resulting spike in the overall system hosting cost, lead to the promise that cheaper but slower memory technologies, such as NVDIMMs (Non Volatile Memory), are going to co-exist with the currently predominant ones, i.e. DRAM. In such future cloud systems where the memory substrate is going to include heterogeneous hardware, Mnemo comes as the necessary memory sizing and data tiering consultant. Mnemo permits quick exploration of the trade-offs between the system cost and application performance, due to the various possible sizings of the hybrid memory system components.

CCS CONCEPTS

• **Hardware** → *Analysis and design of emerging devices and systems*; *Emerging architectures*; *Emerging tools and methodologies*;

KEYWORDS

Hybrid Memory Systems, Capacity Sizing, Cost-Benefit Analysis, In-Memory Key-Value Stores, Cloud Hosting Cost

ACM Reference Format:
Thaleia Dimitra Doudali and Ada Gavrilovska. 2018. Mnemo: Boosting Memory Cost Efficiency in Hybrid Memory Systems. In *Proceedings of ACM Symposium on Cloud Computing , Carlsbad, CA, USA, October 11–13, 2018 (SoCC '18),* 1 pages.
https://doi.org/10.1145/3267809.3275465

Motivation. Applications that execute on hybrid memory systems will have to face performance degradation from the ideal case of infinite DRAM-only capacity. Although there's been a lot of research effort into optimizing performance via runtime-, operating system- and hardware-level solutions for systems with *fixed* memory capacities, Mnemo is providing a solution for right-sizing the various memory components, so as to have good application performance together with reasonable system cost. Initial experiments on an emulated hybrid memory system consisting of DRAM and NVM, show that the performance of key-value store workloads highly depends on the key access pattern, read:write operation ratio and data record size. Most importantly, we observe that for certain widely used workloads, there is scope to significantly reduce the overall cloud hosting cost for trivial application performance slowdown.

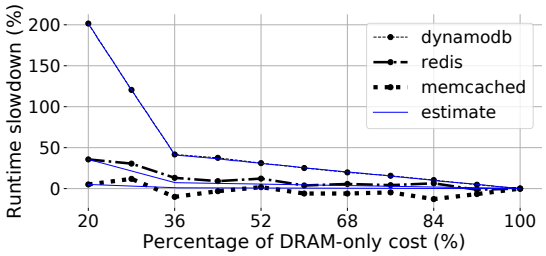

Figure 1: Application slowdown compared to memory cost reduction from DRAM-only availability, for the three top used in-memory key-value stores across incremental DRAM to NVM capacity sizing.

Design. Mnemo is an open-source[1], easy to setup tool, designed for capacity sizing analysis of key value stores on hybrid memory systems. Mnemo does not require any application level modification, but needs a workload descriptor so that it can execute it "as-is" over DRAM-only and NVM-only servers, which will act as baselines of best performance and highest cost vs worst performance and lowest cost. Next, Mnemo uses a very simple but extremely accurate model to estimate the performance slowdown for incremental sizing of DRAM compared to NVM. In this way, Mnemo quickly produces the application performance slowdown trendline of a workload that runs on a hybrid memory system, so that the user can decide which memory capacity ratio provides the required performance guarantees in exchange for minimum memory cost.

Evaluation. We evaluate the tool's accuracy for a variety of representative workloads across the currently most used key-value stores. Figure 1 summarizes the slowdown to cost trendline across key-value stores for workloads with a small set of hot keys, based on the assumption that NVM $/byte is only 20% of the DRAM $/byte. Mnemo's estimate (blue line) is extremely accurate, with only 0.75% median error across Redis, Memcached and DynamoDB. Also, Mnemo identifies that Memcached could operate solely on NVM for less that 10% application performance degradation across workloads, reducing cost down to the minimum possible. Finally, Mnemo suggests that Redis can operate within 10% performance slowdown and minimal use of DRAM for workloads that contain a small set of hot keys, such as retrieval of *trending* data.

[1]https://github.com/Thaleia-DimitraDoudali/mnemo

SoCC '18, October 11–13, 2018, Carlsbad, CA, USA

ACM ISBN 978-1-4503-6011-1/18/10…$15.00
https://doi.org/10.1145/3267809.3275465