Enabling machine intelligent system-level support for data management over extremely heterogeneous hardware

The profile of my research. As a researcher I design and build systems that bring intelligence in the management of the heterogeneous hardware resources of emerging platforms, thereby achieving greater performance and efficiency. Technology has embraced a trend of extreme heterogeneity, and conventional methods to resource management are not adequate for the increased complexity that heterogeneity brings forward and will fail to deliver the full potential of the new hardware. What sets my research approach apart is the fact that I strive to identify when machine intelligence is necessary, versus when our human intelligence and optimized traditional approaches are sufficient. Using observation-driven insights my research lays the grounds for the practical integration of machine learning into system-level resource management, enabling seamless cooperation with existing mechanisms and limiting learning overheads. My dissertation research has developed new techniques focused on managing heterogeneous memory resources. Moving forward I am excited to broaden the impact of my research across the software and hardware stack, to further establish the practical coexistence of machine and human intelligence at the system level.

The relevance of my research. Heterogeneous hardware emerged to address the slowdown of Moore's Law and the exponentially growing demand for compute and data by popular Big Data analytics, applications of artificial intelligence and scientific simulations. More specifically, new types of hardware such as specialized accelerators, persistent memory and smart, programmable interconnects, are now used alongside traditional components in the configurations of exascale supercomputers, datacenters, all the way to personal and edge devices. Although these new hardware technologies deliver acceleration and massive data storage capacities, they introduce new challenges into their management. Not only there is a bigger number of different hardware units to configure and manage, but also there is a substantial disparity in the performance and efficiency tradeoffs they expose, making the decision of which technology to use at what times even more intricate. Naive assumptions that traditional heuristics and approaches are robust to this extreme heterogeneity, lead to significant application performance degradation and major resource inefficiencies. Therefore, it is critical that we revisit well-established resource management approaches and maximize the utility of the hybrid hardware to unlock its full potential.

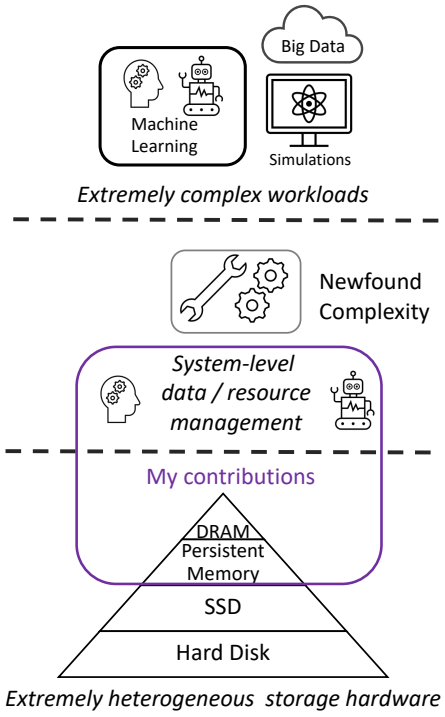

Figure 1: My contributions leverage machine intelligence to tackle the newfound complexity in system-level resource management of heterogeneous memory hardware.

The importance of my research. Recent research on improving the management of heterogeneous systems, has proposed approaches with limited applicability to specific application classes, or which rely on complex performance models, heavy profiling and fine-tuned heuristics, limiting their practical adoption. It is not surprising, therefore, that commercial systems still use simple approaches and empirically set configurations, since these are the most lightweight to use across application domains. My research exposes significant performance gaps left by such systems, proving the need for more sophisticated, yet practical solutions. In response, my research identifies *which* resource management decisions require machine intelligence and determines ways for their *practical* deployment at the system level, under permissible learning overheads. The novelty of my research lies in enabling existing lightweight solutions and sophisticated machine learning methods to operate synergistically, tied together under observation-driven insights. In other words, my research uses machine intelligence whenever it is the only way to reveal performance improvements otherwise feasible under oracular resource management. Next, I describe in more detail the specific contributions of my research so far, that improve upon the resource management of heterogeneous *memory* hardware [1, 2, 3, 4, 5, 6, 7].

1 Research Contributions

Summary. Persistent memory realizes massive memory capacities at a fraction of the cost of traditional DRAM-only systems. The recent commercial release of persistent memory (PMEM), following upon years of emulation and simulation prototypes, unveiled some unexpected performance behaviors [5], with respect to data locality and data management granularity. Regardless, the most well established system-level approach [8, 9, 10] to improve application performance when both DRAM and PMEM are part of a flat memory address space, is to optimize the initial and dynamic data tiering. In more detail, data access behavior is dynamically monitored and frequently accessed (hot) data is moved from DRAM to PMEM and cold data from DRAM to PMEM. However, the decision of how to tier data, which data to move and at what times is non trivial. The decision becomes even harder for popular workloads with irregular memory access patterns, such as graph analytics and machine learning algorithms, and existing lightweight approaches leave a significant performance gap. In response, my early work optimizes upon static data tiering for shared hybrid memory systems [7] and maximizes the system's cost efficiency with appropriate memory capacity sizing [6]. Regarding dynamic data movements, my first dissertation contribution integrates machine intelligence into the selection of which data to move, enabling accurate predictions of future data access behaviors. Second, my research unveils the impact of empirically misconfigured data movement frequencies and proposes a fast and effective tuning solution, based on observation-driven insights regarding application data reuse. Lastly, my research leverages these insights, to further reduce the machine learning overheads and realize their practical system-level integration. In this way, my research brings together human and machine intelligence into a holistic hybrid memory management solution, that eliminates the performance gap left by current state-of-the-art approaches.

1.1 Machine intelligent data movement selection

A big challenge regarding the dynamic data movement selection across heterogeneous hardware, is how to accurately predict which data will be frequently accessed in the future, so as to timely move it in DRAM. Existing hybrid memory management approaches rely on past data access history to predict future behaviors. Although this is a lightweight solution, since it uses readily available system-level information, it is not robust to sudden changes in access patterns or complete randomness. Consequently, this creates up to 50% performance degradation from the case of perfectly accurate pattern prediction.

To address this challenge, I developed *Kleio*; a hybrid memory page scheduler with machine intelligence [4], that was a **best paper award finalist at HPDC'19**. Kleio deploys Recurrent Neural Networks (RNNs) to learn memory access patterns, that existing history-based solutions fail to accurately predict. The deployment of RNNs for the prediction of page access patterns, such as for cache prefetching [11], cannot be *practically* adapted 'as-is' to provide accurate predictions for page scheduling; instead Kleio deploys RNNs at the granularity of individual pages, to learn the pattern of their access frequency across periods of time. The novelty in the design of Kleio comes from the selection of a small page subset, whose machine intelligent management reveals most of the application performance improvement, while using existing history-based management for the remaining pages. The resulting hybrid management approach lays the ground for Kleio's practical integration at the system-level with permissible learning overheads. Kleio's impact is extremely promising, since it bridges on average 80% and up to 95% of the existing performance gap.

1.2 Data movement frequency tuning

Another challenging task in hybrid memory management, is to properly set the data movement frequency, to strike the right balance between maximizing DRAM's utility under acceptable data migration overheads. The extreme heterogeneity of current systems and the ever exploding number of configuration parameters, makes it hard to fine-tune all possible ones. Thus, existing hybrid memory management solutions empirically set the data movement frequency at the system-level at fixed values that apply across application domains and inputs. Even more interestingly, the selected values range within orders of magnitude across solutions.This results in 10%-100% application performance degradation compared to an optimally chosen value.

To address this challenge, I built *Cori*; a system-level solution for tuning the operational frequency of periodic page schedulers for hybrid memories [2], that will appear in **IPDPS '21**. The novelty of Cori derives from observations on data reuse times and their alignment with the data movement frequency. Cori synthesizes information on data reuse to properly identify the data movement frequencies to be tested, reducing by 5×

the number of tuning trials compared to existing empirical or insight-less tuning approaches, and realizing application performance levels within only 3%, on average, from the case of optimally selected frequency.

1.3 Reducing machine intelligent data management overheads

An important challenge in the integration of machine intelligence inside the system-level hybrid memory management is to allow for low training overheads and learning times. As a first step in that direction, my research sets the example of how such systems should be built, as per Kleio's [4] design. Yet, the ever growing data sizes and scales of current analytics, together with the increasing randomness in application data access behaviors, may require a non-trivial number of pages to be managed intelligently, thus RNNs to be deployed. In fact, this number varies between 3× - 4× across applications and depending upon the available resources and time constraints, can be prohibitive.

To address this challenge, I built a grouping mechanism [1] to identify pages that share the same access patterns throughout application execution. This enables the deployment of a single RNN model per page group, reducing the total number of RNNs to be trained, thus aggregate learning overheads. The novelty of this work derives from utilizing insights on data reuse times, as highlighted in Cori [2], to identify page clusters with similar access behavior across application runtime intervals. This allows the machine intelligent management of 2× more pages than Kleio, under the same training overheads, making it possible to extend the benefits of Kleio to more complex workloads in a practical manner.

2 Future Research Directions

Summary. I am passionate about continuing to work at the intersection of machine learning and systems, with a primary focus on machine intelligent system-level resource management of complex systems. I have a plan for my immediate next steps and I have built proper foundations to extend the impact of my research more broadly across heterogeneous types of hardware and data management layers. I look forward to fostering collaborations with experts in artificial intelligence to better understand the usability of these algorithms. In addition, teamwork with researchers that specialize in solutions in other layers of the software stack, such as application- or compiler-level, will be critical to facilitate the support for the appropriate data collection that will enable practical machine learning augmented systems. Venturing out of my area of expertise, I believe that proper data visualization and user-centric tool designs can be of tremendous help toward better guiding the design of the new class of system-level solutions needed for extremely heterogeneous hardware and emerging workloads.

2.1 Online adaptive hybrid memory management

The immediate next step in my research agenda is to extend my dissertation research toward building practical online adaptive data management techniques for systems with heterogeneous memories. I will build upon the design principles, whose effectiveness I established in my dissertation, and deliver practical robust solutions that realize benefits in performance and resource efficiency. This is feasible due to my approach of selecting the key design points of my systems. More specifically, the online training capabilities of RNNs and the online tracking of data reuse facilitate the straight-forward extension of my recent work into an online solution. I am particularly interested in extending my current system prototypes to actual deployments in the context of real exascale and datacenter server systems, with a complete set of state-of-the-art accelerator, compute and memory infrastructure. I hope to maintain and create new collaborations with supercomputing and industry labs and facilities, so as to obtain access to such hardware resources. In conclusion, my goal is to see Kleio, Cori, and the future tools and systems my research will develop, in practical deployments that realize the tremendous benefits of machine intelligent hybrid memory management.

2.2 Practical machine learning for system-level heterogeneous hardware management

Moving forward I aspire to be one of the pioneers in establishing the practical system-level integration of machine learning methods to better manage the extreme complexity of the heterogeneous hardware. I am able to do so because the design principles of my machine intelligent hybrid memory management, can be adapted to any level of heterogeneous resources. More specifically, the same questions of data tiering, data movement

selection and frequency become significantly more complex to solve when considering the combination of heterogeneous, multi-tiered memory and storage technologies. In addition, the growing number of specialized accelerators, such as GPUs, TPUs, DPUS, IPUs, further perplexes their data affinity and the trade-offs associated with the different data movement paths that interconnect them. Although my research so far lays the grounds, there is a lot more to be done in establishing robust and effective machine intelligent data management at the system-level. The major challenge is to enable fast machine learning inference, so as to preserve the performance guarantees of a system-level solution. However, model training and inference is an intricate workload that requires heavy resource use and possibly prohibitive aggregate runtimes for the purpose of online management of enormous data scales. Therefore, even the use of custom hardware accelerators may not be sufficient to enable permissible inference times.

In response, I plan to develop a research program that tackles these management challenges in a multi-pronged manner. First, I strongly believe in the judicious use of machine learning for when its critical and its benefits outweight any costs and complexities. This naturally leads to heterogeneous solutions over heterogeneous hardware. For example, my hybrid memory management solution - Kleio - is hybrid itself. Such hybrid solutions allow to take advantage of well established system-level solutions that are state-of-the-art for homogeneous hardware and introduce machine intelligent management for the cases where the prior ones fail to deliver the attainable performance levels. The challenge, and what particularly excites me, is how to identify such cases where machine intelligence is necessary. This is where our human intelligence and observation-driven insights can be detrimental to synthesize such a hybrid solution. Understanding data access behaviors, which current heuristics work best at what times, is critical into creating a sophisticated combination of machine intelligence with current state-of-the-art approaches.

Next, I plan to explore opportunities to leverage advances in machine learning techniques itself, that facilitate their practical deployment. In particular, the method of transfer learning allows for the application of the acquired knowledge in one task to another similar task, helping to amortize model training costs. This can be particularly beneficial for the purpose of hardware independent machine learning training. For example, if the available storage capacity changes or the hardware technology, given the ever increasing number of options or shared use of resources, this shouldn't require the retrain of models, whose inference is on the critical path of a resource or data management decision. In addition, transfer learning can also be beneficial for applications with similar access patterns or executions under larger input sizes. Given the ever increasing complexity in the data access behaviors of emerging workloads, it will be daunting to train different models for each specific one. Thus, focusing learning of different classes of applications, such as graph analytics or neural networks, and transfer learning for different sizes and configurations, can further reduce training overheads.

Finally, considering the critical nature of input data and feature selection for the efficacy of the resulting machine intelligence, I will explore collaborations with researchers across the department, particularly those working at the architecture, compiler/runtime, and programming language levels, to identify new principles and methodologies that will streamline the process of generating lightweight and relevant resource management data and resulting models.

In conclusion, I am excited to work on enabling a truly practical and robust system-level integration of machine learning for the purpose of heterogeneous hardware management. The continued trends toward extreme heterogeneity across the computing landscape introduce tremendous management complexity. Without judicious use of machine intelligent methods, this complexity can render management ineffective and obviate the benefits that the new technologies are expected to deliver. The successful design principles of my work so far, show great promise for impactful future contributions across hardware technologies and emerging workloads.

2.3 Data visualization for effective system-level solutions for complex hardware and workloads

I am a big proponent of the phrase that 'one picture is worth a thousand words', meaning that humans comprehend more effectively visual rather than textual information. Visualization of system-level information is primarily used in commercial software products, such as virtual machines and cloud solutions, usually in the form of a visual board of performance statistics and summary of behaviors, so that the users can fine-tune the configurations of their systems and deployments. Yet in computer systems research, data visualization is not part of the system design process, since the decision space so far has been limited and we were able to generate simple heuristics and methods to support traditional hardware and application classes.

However, the extreme heterogeneity of current hardware introduces new complexity in the decision space of their system-level management, given the increase in choices and performance statistics to analyze. Similarly,

the access behavior of modern analytics, such as applications of artificial intelligence, are significantly different from traditional ones, making it hard to predict. The proper visualization of application-level behaviors or system-level performance metrics, can unlock new observations that create a new class of more sophisticated system designs. More specifically, I believe that data visualization and in general observation-based insights, can be the key in understanding when human versus machine intelligence is necessary to use. More importantly, it justifies the need for a more intelligent solution. From a personal experience, the visualization of memory access patterns in relation to the timing of data management decisions, has been a major determinant in the design of my system-level hybrid memory management solutions [4, 2]. For instance, the visualization of the data movement frequency and data reuse, led to the insight that these should be aligned to deliver best performance, which has been completely neglected by well established yet insight-less approaches. Based on these observations, in the longer term, I'm intrigued by the opportunities for systems researchers to benefit from data visualization research, and more general information about the workloads or platforms that exists outside of traditional system-level layers. I plan to seek out collaboration opportunities to further explore this space, with the goal of designing next generation management infrastructure for complex systems with increased effectiveness, usability and interpretability.

References

[1] **Thaleia Dimitra Doudali** and Ada Gavrilovska. Learning at the right beat of periodic data movements over hybrid memory systems. Under preparation.

[2] **Thaleia Dimitra Doudali**, Daniel Zahka, and Ada Gavrilovska. Cori: Dancing to the right beat of periodic data movements over hybrid memory systems. In *2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2021.

[3] **Thaleia Dimitra Doudali**, Daniel Zahka, and Ada Gavrilovska. The case for optimizing the frequency of periodic data movements over hybrid memory systems. In *Proceedings of the International Symposium on Memory Systems*, MEMSYS '20, 2020.

[4] **Thaleia Dimitra Doudali**, Sergey Blagodurov, Abhinav Vishnu, Sudhanva Gurumurthi, and Ada Gavrilovska. Kleio: A hybrid memory page scheduler with machine intelligence. In *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, HPDC '19, pages 37–48, New York, NY, USA, 2019. ACM.

[5] Tony Mason, **Thaleia Dimitra Doudali**, Margo Seltzer, and Ada Gavrilovska. Unexpected performance of intel optane dc persistent memory. *IEEE Computer Architecture Letters*, 19(1):55–58, 2020.

[6] **Thaleia Dimitra Doudali** and Ada Gavrilovska. Mnemo: Boosting memory cost efficiency in hybrid memory systems. In *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 412–421, 2019.

[7] **Thaleia Dimitra Doudali** and Ada Gavrilovska. Comerge: Toward efficient data placement in shared heterogeneous memory systems. In *Proceedings of the International Symposium on Memory Systems*, MEMSYS '17, pages 251–261, New York, NY, USA, 2017. Association for Computing Machinery.

[8] Neha Agarwal and Thomas F. Wenisch. Thermostat: Application-transparent page management for two-tiered main memory. In *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '17, pages 631–644, New York, NY, USA, 2017. Association for Computing Machinery.

[9] Subramanya R. Dulloor, Amitabha Roy, Zheguang Zhao, Narayanan Sundaram, Nadathur Satish, Rajesh Sankaran, Jeff Jackson, and Karsten Schwan. Data tiering in heterogeneous memory systems. In *Proceedings of the Eleventh European Conference on Computer Systems*, EuroSys '16, New York, NY, USA, 2016. Association for Computing Machinery.

[10] Sudarsun Kannan, Ada Gavrilovska, Vishal Gupta, and Karsten Schwan. Heteroos: Os design for heterogeneous memory management in datacenter. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, ISCA '17, pages 521–534, New York, NY, USA, 2017. Association for Computing Machinery.

[11] Milad Hashemi, Kevin Swersky, Jamie Smith, Grant Ayers, Heiner Litz, Jichuan Chang, Christos Kozyrakis, and Parthasarathy Ranganathan. Learning memory access patterns. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1919–1928, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.