IP4 Interactive Presentations

Date: Thursday, March 28, 2019
Time: 10:00 - 10:30
Location / **Room:** Poster Area

Interactive Presentations run simultaneously during a 30-minute slot. Additionally, each IP paper is briefly introduced in a one-minute presentation in a corresponding regular session

Label	Presentation Title Authors

IP4-1 **AN EFFICIENT MAPPING APPROACH TO LARGE-SCALE DNNS ON MULTI-FPGA ARCHITECTURES**

Speaker:
Jiaxi Zhang, Peking University, CN
Authors:
Wentai Zhang[1], Jiaxi Zhang[1], Minghua Shen[2], Guojie Luo[1] and Nong Xiao[3]
[1]Peking University, CN; [2]Sun Yat-sen University, CN; [3]Sun Yat-Sen University, CN

Abstract

FPGAs are very attractive to accelerate the deep neural networks (DNNs). While single FPGA can provide good performance for small-scale DNNs, support for large-scale DNNs is limited due to higher resource demand. In this paper, we propose an efficient mapping approach for accelerating large-scale DNNs on asymmetric multi-FPGA architectures. In this approach, the neural network mapping can be formulated as a resource allocation problem. We design a dynamic programming-based partitioning to solve this problem optimally. Experimental results using the large-scale ResNet-152 demonstrate that our approach deploys sixteen FPGAs to provide an advantage of 16.4x GOPS over the state-of-the-art work.

Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-2 **A WRITE-EFFICIENT CACHE ALGORITHM BASED ON MACROSCOPIC TREND FOR NVM-BASED READ CACHE**

Speaker:
Ning Bao, Renmin University of China, CN
Authors:
Ning Bao[1], Yunpeng Chai[1] and Xiao Qin[2]
[1]Renmin University of China, CN; [2]Auburn University, US

Abstract

Compared with traditional storage technologies, non-volatile memory (NVM) techniques have excellent I/O performances, but high costs and limited write endurance (e.g., NAND and PCM) or high energy consumption of writing (e.g., STT-MRAM). As a result, the storage systems prefer to utilize NVM devices as read caches for performance boost. Unlike write caches, read caches have greater potential of write reduction because their writes are only triggered by cache updates. However, traditional cache algorithms like LRU and LFU have to update cached blocks frequently because it is difficult for them to predict data popularity in the long future. Although some new algorithms like SieveStore reduce cache write pressure, they still rely on those traditional cache schemes for data popularity prediction. Due to the bad long-term data popularity prediction effect, these new cache algorithms lead to a significant and unnecessary decrease of cache hit ratios. In this paper, we propose a new Macroscopic Trend (MT) cache replacement algorithm to reduce cache updates effectively and maintain high cache hit ratios. This algorithm discovers long-term hot data effectively by observing the macroscopic trend of data blocks. We have conducted extensive experiments driven by a series of real-world traces, and the results indicate that compared with LRU, the MT cache algorithm can achieve 15.28 times longer lifetime or less energy consumption of NVM caches with a similar hit ratio.

Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-3 **SRAM DESIGN EXPLORATION WITH INTEGRATED APPLICATION-AWARE AGING ANALYSIS**

Speaker:
Alexandra Listl, TUM, DE
Authors:
Alexandra Listl[1], Daniel Mueller-Gritschneder[2], Sani Nassif[3] and Ulf Schlichtmann[2]
[1]Chair of Electronic Design Automation, DE; [2]TUM, DE; [3]Radyalis, US

Abstract

On-Chip SRAMs are an integral part of safetycritical System-on-Chips. At the same time however, they are also most susceptible to reliability threats such as Bias Temperature Instability (BTI), originating from the continuous trend of technology shrinking. BTI leads to a significant performance degradation, especially in the Sense Amplifiers (SAs) of SRAMs, where failures are fatal, since the data of a whole column is destroyed. As BTI strongly depends on the workload of an application, the aging rates of SAs in a memory array differ significantly and the incorporation of workload information into aging simulations is vital. Especially in safety-critical systems precise estimation of application specific reliability requirements to predict the memory lifetime is a key concern. In this paper we present a workload-aware aging analysis for On-Chip SRAMs that incorporates the workload of real applications executed on a processor. According to this workload, we predict the performance degradation of the SAs in the memory. We integrate this aging analysis into an aging-aware SRAM design exploration framework that generates and characterizes memories of different array granularity to select the most reliable memory architecture for the intended application. We show that this technique can mitigate SA degradation significantly depending on the environmental conditions and the application workload.

Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-4 **FROM MULTI-LEVEL TO ABSTRACT-BASED SIMULATION OF A PRODUCTION LINE**

Speaker:
Stefano Centomo, University of Verona, IT
Authors:
Stefano Centomo, Enrico Fraccaroli and Marco Panato, University of Verona, IT

Abstract

This paper proposes two approaches for the integration of cyber-physical systems in a production line in order to obtain predictions concerning the actual production, core operation in the context of Industry 4.0. The first approach relies on the Multi-Level paradigm where multiple descriptions of the same CPS are modeled with different levels of details. Then, the models are switched at runtime. The second approach relies on abstraction techniques of CPS maintaining a certain levels of details. The two approaches are validated and compared with a real use case scenario to identify the most effective simulation strategy.

Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-5 **ACCURATE DYNAMIC MODELLING OF HYDRAULIC SERVOMECHANISMS**

Speaker:
Manuel Pencelli, Yanmar R&D Europe S.r.l., IT

Authors:
Manuel Pencelli[1], Renzo Villa[2], Alfredo Argiolas[1], Gianni Ferretti[2], Marta Niccolini[1], Matteo Ragaglia[1], Paolo Rocco[2] and Andrea Maria Zanchettin[2]
[1]YANMAR R&D EUROPE S.R.L, IT; [2]Politecnico di Milano, IT

Abstract
In this paper, the process of modelling and identification of a hydraulic actuator is discussed. In this framework a simple model based on the classical theory have been derived and a first experimental campaign has been performed on a test bench. These tests highlighted the presence of unmodelled phenomena (dead-zone, hysteresis, etc.), therefore a second and more extensive set of experiments has been done. With the acquired knowledge a new improved model is presented and its parameter identified. Finally several test has been performed in order to experimentally validate the model.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-6 **PLANNING WITH REAL-TIME COLLISION AVOIDANCE FOR COOPERATING AGENTS UNDER RIGID BODY CONSTRAINT**

Speaker:
Federico Vesentini, University of Verona, IT

Authors:
Nicola Piccinelli, Federico Vesentini and Riccardo Muradore, University of Verona, IT

Abstract
In automated warehouses, path planning is a crucial topic to improve automation and efficiency. This kind of planning is usually computed off-line knowing the planimetry of the warehouse and the starting and target points of each agent. However, this global approach is not able to manage unexpected static/dynamic obstacles and other agents moving in the same area. For this reason in multi-robot systems global planners are usually integrated with local collision avoidance algorithms. In this paper we use the Voronoi diagram as global planner and the Velocity Obstacle (VO) method as collision avoidance algorithm. The goal of this paper is to extend such hybrid motion planner by enforcing mechanical constraints between agents in order to execute a task that cannot be performed by a single agent. We will focus on the cooperative task of carrying a payload, such as a bar. Two agents are constrained to move at the end points of the bar. We will improve the original algorithms by taking into account dynamically the constrained motion both at the global and at the collision avoidance level.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-7 **THE CASE FOR EXPLOITING UNDERUTILIZED RESOURCES IN HETEROGENEOUS MOBILE ARCHITECTURES**

Speaker:
Nikil Dutt, University of California, Irvine, US

Authors:
Chenying Hsieh, Nikil Dutt and Ardalan Amiri Sani, UC Irvine, US

Abstract
Heterogeneous architectures are ubiquitous in mobile plat-forms, with mobile SoCs typically integrating multiple processors along with accelerators such as GPUs (for data-parallel kernels) and DSPs (for signal processing kernels). This strict partitioning of application execution on heterogeneous compute resources often results in underutilization of resources such as DSPs. We present a case study executing popular data-parallel workloads such as convolutional neural networks (CNNs), computer vision application and graphics kernels on mobile devices, and show that both performance and energy consumption of mobile platforms can be improved by synergistically deploying these underutilized DSPs. Our experiments on a mobile Snapdragon 835 platform under both single and multiple application scenarios executing CNNs and graphics workloads, demonstrates average performance and energy improvements of 15-46% and 18-80% respectively by synergistically deploying all available compute resources, especially the underutilized DSP.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-8 **ONLINE RARE CATEGORY DETECTION FOR EDGE COMPUTING**

Speaker:
Yufei Cui, City University of Hong Kong, HK

Authors:
Yufei Cui[1], Qiao Li[1], Sarana Nutanong[2] and Chun Jason Xue[1]
[1]City University of Hong Kong, HK; [2]Vidyasirimedhi Institute of Science and Technology, TH

Abstract
Abstract — Identifying rare categories is an important data management problem in many application fields including video surveillance, ecological environment monitoring and precision medicine. Previous solutions in literature require all data instances to be first delivered to the server. Then, the rare categories identification algorithms are executed on the pool of data to find informative instances for human annotators to label. This incurs large bandwidth consumption and high latency. To deal with the problems, we propose a light-weight rare categories identification framework. At the sensor side, the designed online algorithm filters less informative data instances from the data stream and only sends the informative ones to human annotators. After labeling, the server only sends labels of the corresponding data instances in response. The sensor-side algorithm is extended to enable cooperation between embedded devices for the cases that data is collected in a distributed manner. Experiments are conducted to show our framework dramatically outperforms the baseline. The network traffic is reduced by 75% on average.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-9 **RAGRA: LEVERAGING MONOLITHIC 3D RERAM FOR MASSIVELY-PARALLEL GRAPH PROCESSING**

Speaker:
Yu Huang, Huazhong University of Science and Technology, CN

Authors:
Yu Huang, Long Zheng, Xiaofei Liao, Hai Jin, Pengcheng Yao and Chuangyi Gui, Huazhong University of Science and Technology, CN

Abstract
With the maturity of monolithic 3D integration, 3D ReRAM provides impressive storage-density and computational-parallelism with great opportunities for parallel-graph processing acceleration. In this paper, we present RAGra, a 3D ReRAM based graph processing accelerator, which has two significant technical highlights. First, monolithic 3D ReRAM usually has the complexly-intertwined feature with shared input wordlines and output bitlines for different layers. We propose a novel mapping scheme, which can guide to apply graph algorithms into 3D ReRAM seamlessly and correctly for exposing the massive parallelism of 3D ReRAM. Second, consider the sparsity of real-world graphs, we further propose a row- and column-mixed execution model, which can filter invalid subgraphs for exploiting the massive parallelism of 3D ReRAM. Our evaluation on 8-layer stacked ReRAM shows that RAGra outperforms state-of-the-art planar (2D) ReRAM-based graph accelerator GraphR by 6.18x performance improvement and 2.21x energy saving, on average. In particular, RAGra significantly outperforms GridGraph (a typical CPU-based graph system) by up to 293.12x speedup.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-10 **ACCURATE COST ESTIMATION OF MEMORY SYSTEMS INSPIRED BY MACHINE LEARNING FOR COMPUTER VISION**

Speaker:
Lorenzo Servadei, Infineon Technologies, DE

Authors:
Lorenzo Servadei[1], Elena Zennaro[1], Keerthikumara Devarajegowda[1], Martin Manzinger[1], Wolfgang Ecker[1] and Robert Wille[2]
[1]Infineon AG, DE; [2]Johannes Kepler University Linz, AT

Abstract
Hardware/software co-designs are usually defined at high levels of abstractions at the beginning of the design process in order to allow plenty of options how to eventually realize a system. This allows for design exploration which in turn heavily relies on knowing the costs of different design configurations (with respect to hardware usage as well as firmware metrics). To this end, methods for cost estimation are frequently applied in industrial practice. However, currently used methods for cost estimation oversimplify the problem and ignore important features - leading to estimates which are far off from the real values. In this work, we address this problem for memory systems. To this end, we borrow and re-adapt solutions based on Machine Learning (ML) which have been found suitable for problems from the domain of Computer Vision (CV) - in particular age determination of persons depicted in images. We show that, for an ML approach, age determination from the CV domain is actually very similar to cost estimation of a memory system.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-11 **PRACTICAL CAUSALITY HANDLING FOR SYNCHRONOUS LANGUAGES**

Speaker:
Steven Smyth, Kiel University, DE
Authors:
Steven Smyth, Alexander Schulz-Rosengarten and Reinhard von Hanxleden, Dept. of Computer Science, Kiel University, DE

Abstract
A key to the synchronous principle of reconciling concurrency with determinism is to establish at compile time that a program is causal, which means that there exists a schedule that obeys the rules put down by the language. In practice it can be rather cumbersome for the developer to cure causality problems. To facilitate causality handling, we propose, first, to enrich the scheduling regime of the language to also consider explicit scheduling directives that can be used by either the modeler or model-to-model transformations. Secondly, we propose to enhance programming environments with dedicated causality views to guide the developer in finding causality issues. Our proposals should be applicable for synchronous languages; we here illustrate them for the SCCharts language and its open source development platform KIELER.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-12 **APPLICATION PERFORMANCE PREDICTION AND OPTIMIZATION UNDER CACHE ALLOCATION TECHNOLOGY**

Speaker:
Yeseong Kim, UCSD, US
Authors:
Yeseong Kim[1], Ankit More[2], Emily Shriver[2] and Tajana Rosing[1]
[1]University of California San Diego, US; [2]Intel, US

Abstract
Many applications running on high-performance computing systems share limited resources such as the last-level cache, often resulting in lower performance. Intel recently introduced a new control mechanism, called cache allocation technology (CAT), which controls the cache size used by each application. To intelligently utilize this technology for automated management, it is essential to accurately identify application performance behavior for different cache allocation scenarios. In this work, we show a novel approach which automatically builds a prediction model for application performance changes with CAT. We profile the workload characteristics based on Intel Top-down Microarchitecture Analysis Method (TMAM), and train the model using machine learning. The model predicts instructions per cycle (IPC) across available cache sizes allocated for the applications. We also design a dynamic cache management technique which utilizes the prediction model and intelligently partitions the cache resource to improve application throughput. We implemented and evaluated the proposed framework in Intel PMU profiling tool running on Xeon Platinum 8186 Skylake processor. In our evaluation, we show that the proposed model accurately predicts the IPC changes of applications with 4.7% error on average for different cache allocation scenarios. Our predictive online cache managements achieves improvements on application performance of up to 25% as compared to a prediction-agnostic policy.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-13 **GENERALIZED MATRIX FACTORIZATION TECHNIQUES FOR APPROXIMATE LOGIC SYNTHESIS**

Speaker:
Sherief Reda, Brown University, US
Authors:
Soheil Hashemi and Sherief Reda, Brown University, US

Abstract
Approximate computing is an emerging computing paradigm, where computing accuracy is relaxed for improvements in hardware metrics, such as design area and power profile. In circuit design, a major challenge is to synthesize approximate circuits automatically from input exact circuits. In this work, we extend our previous work, BLASYS, for approximate logic synthesis based on matrix factorization, where an arbitrary input circuit can be approximated in a controlled fashion. Whereas our previous approach uses a semi-ring algebra for factorization, this work generalizes matrix-based circuit factorization to include both semi-ring and field algebra implementations. We also propose a new method for truth table folding to improve the factorization quality. These new approaches significantly widen the design space of possible approximate circuits, effectively offering improved trade-offs in terms of quality, area and power consumption. We evaluate our methodology on a number of representative circuits showcasing the benefits of our proposed methodology for approximate logic synthesis.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-14 **CARS: A MULTI-LAYER CONFLICT-AWARE REQUEST SCHEDULER FOR NVME SSDS**

Speaker:
Tianming Yang, Huanghuai University, CN
Authors:
Tianming Yang[1], Ping Huang[2], Weiying Zhang[3], Haitao Wu[1] and Longxin Lin[4]
[1]Huanghuai University, CN; [2]Temple University, US; [3]Northeastern University, CN; [4]Jinan University, CN

Abstract
NVMe SSDs are nowadays widely deployed in various computing platforms due to its high performance and low power consumption, especially in data centers to support modern latency-sensitive applications. NVMe SSDs improve on SATA and SAS interfaced SSDs by providing a large number of device I/O queues at the host side and applications can directly manage the queues to concurrently issue requests to the device. However, the currently deployed request scheduling approach is oblivious to the states of the various device internal components and thus may lead to suboptimal decisions due to various resource contentions at different layers inside the SSD device. In this work, we propose a Conflict Aware Request Scheduling policy named CARS for NVMe SSDs to maximally leverage the rich parallelism available in modern NVMe SSDs. The central idea is to check possible conflicts that a fetched request might be associated with before dispatching that request. If there exists a conflict, it refrains from issuing the request and move to check a request in the next submission queue. In doing so, our scheduler can evenly distribute the requests among the parallel idle components in the flash chips, improving performance. Our evaluations have shown that our scheduler can reduce the slowdown metric by up to 46% relative to the de facto round-robin scheduling policy for a variety of patterned workloads.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP4-15 **QUEUE BASED MEMORY MANAGEMENT UNIT FOR HETEROGENEOUS MPSOCS**

Speaker:
Robert Wittig, Technische Universität Dresden, DE
Authors:
Robert Wittig, Mattis Hasler, Emil Matus and Gerhard Fettweis, Technische Universität Dresden, DE

Abstract
Sharing tightly coupled memory in a multi-processor system-on-chip is a promising approach to improve the programming flexibility as well as to ease the constraints imposed by area and power. However, it poses a challenge in terms of access latency. In this paper, we present a queue based memory management unit which combines the low latency access of shared tightly coupled memory with the flexibility of a traditional memory management unit. Our passive conflict detection approach significantly reduces the critical path compared to previously proposed methods while preserving the flexibility associated with dynamic memory allocation and heterogeneous data widths.
Download Paper (PDF; Only available from the DATE venue WiFi)
