



Published on DATE 2019 (<https://past.date-conference.com>)

[Home](#) > [Printer-friendly PDF](#) > [Printer-friendly PDF](#)

## IP2 Interactive Presentations

**Date:** Wednesday, March 27, 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
IP2-1	<p><b>TRANSREC: IMPROVING ADAPTABILITY IN SINGLE-ISA HETEROGENEOUS SYSTEMS WITH TRANSPARENT AND RECONFIGURABLE ACCELERATION</b></p> <p><b>Speaker:</b> Marcelo Brandalero, Universidade Federal do Rio Grande do Sul (UFRGS), BR</p> <p><b>Authors:</b> Marcelo Brandalero<sup>1</sup>, Muhammad Shafique<sup>2</sup>, Luigi Carro<sup>1</sup> and Antonio Carlos Schneider Beck<sup>1</sup> <sup>1</sup>UFRGS - Universidade Federal do Rio Grande do Sul, BR; <sup>2</sup>Vienna University of Technology (TU Wien), AT</p> <p><b>Abstract</b> <i>Single-ISA heterogeneous systems, such as ARM's big.LITTLE, use microarchitecturally-different General-Purpose Processor cores to efficiently match the capabilities of the processing resources with applications' performance and energy requirements that change at run time. However, since only a fixed and non-configurable set of cores is available, reaching the best possible match between the available resources and applications' requirements remains a challenge, especially considering the varying and unpredictable workloads. In this work, we propose TransRec, a hardware architecture which improves over these traditional heterogeneous designs. TransRec integrates a shared, transparent (i.e., no need to change application binary) and adaptive accelerator in the form of a Coarse-Grained Reconfigurable Array that can be used by any of the General-Purpose Processor cores for on-demand acceleration. Through evaluations with cycle-accurate gem5 simulations, synthesis of real RISC-V processor designs for a 15nm technology, and considering the effects of Dynamic Voltage and Frequency Scaling, we demonstrate that TransRec provides better performance-energy tradeoffs that are otherwise unachievable with traditional big.LITTLE-like designs. In particular, for less than 40% area overhead, TransRec can improve performance in the low-energy mode (LITTLE) by 2.28x, and can improve both performance and energy efficiency by 1.32x and 1.59x, respectively, in high-performance mode (big).</i></p> <p><a href="#">Download Paper (PDF; Only available from the DATE venue WiFi)</a></p>
IP2-2	<p><b>CADE: CONFIGURABLE APPROXIMATE DIVIDER FOR ENERGY EFFICIENCY</b></p> <p><b>Speaker:</b> Mohsen Imani, University of California, San Diego, US</p> <p><b>Authors:</b> Mohsen Imani, Ricardo Garcia, Andrew Huang and Tajana Rosing, University of California San Diego, US</p> <p><b>Abstract</b> <i>Approximate computing is a promising solution to design faster and more energy efficient systems, which provides an adequate quality for a variety of functions. Division, in particular, floating point division, is one of the most important operations in multimedia applications, which has been implemented less in hardware due to its significant cost and complexity. In this paper, we proposed CADE, a Configurable Approximate Divider which performs floating point division operation with a runtime controllable accuracy. The approximation of the CADE is accomplished by removing the costly division operation and replacing it with a subtraction of the input operands mantissa. To increase the level of accuracy, CADE analyses the first N bits (called tuning bits) of both input operands mantissa to estimate the division error. If CADE determines that the first approximation is unacceptable, a pre-computed value is retrieved from memory and subtracted from the first approximation mantissa. At runtime, CADE can provide a higher accuracy by increasing the number of tuning bits. The proposed CADE was integrated on the AMD GPU architecture. Our evaluation shows that CADE is at least 4.1x more energy efficient, 1.5x faster, and 1.7x higher area efficient as compared to state-of-the-art approximate dividers while providing 25% lower error rate. In addition, CADE gives a new knob to GPU in order to configure the level of approximation at runtime depending on the application/user accuracy requirement.</i></p> <p><a href="#">Download Paper (PDF; Only available from the DATE venue WiFi)</a></p>
IP2-3	<p><b>HCFTL: A LOCALITY-AWARE PAGE-LEVEL FLASH TRANSLATION LAYER</b></p> <p><b>Speaker:</b> Hao Chen, University of Science and Technology of China, CN</p> <p><b>Authors:</b> Hao Chen<sup>1</sup>, Cheng Li<sup>1</sup>, Yubiao Pan<sup>2</sup>, Min Lyu<sup>1</sup>, Yongkun Li<sup>1</sup> and Yinlong Xu<sup>1</sup> <sup>1</sup>University of Science and Technology of China, CN; <sup>2</sup>Huaqiao University, CN</p> <p><b>Abstract</b> <i>The increasing capacity of SSDs requires a large amount of built-in DRAM to hold the mapping information of logical-to-physical address translation. Due to the limited size of DRAM, existing FTL schemes selectively keep some active mapping entries in a Cached Mapping Table (CMT) in DRAM, while storing the entire mapping table on flash. To improve the CMT hit ratio with limited cache space on SSDs, in this paper, we propose a novel FTL, a hot-clusterity FTL (HCFTL) that clusters mapping entries recently evicted from the cache into dynamic translation pages (DTPs). Given the temporal localities that those hot entries are likely to be visited in near future, loading DTPs will increase the CMT hit ratio and thus improve the FTL performance. Furthermore, we introduce an index structure to speedup the lookup of mapping entries in DTPs. Our experiments show that HCFTL can improve the CMT hit ratio by up to 41.1% and decrease the system response time by up to 33.3%, compared to state-of-the-art FTL schemes.</i></p> <p><a href="#">Download Paper (PDF; Only available from the DATE venue WiFi)</a></p>
IP2-4	<p><b>MODEL CHECKING IS POSSIBLE TO VERIFY LARGE-SCALE VEHICLE DISTRIBUTED APPLICATION SYSTEMS</b></p> <p><b>Speaker:</b> Haitao Zhang, School of Information Science and Engineering, Lanzhou University, CN</p> <p><b>Authors:</b> Haitao Zhang<sup>1</sup>, Ayang Tuo<sup>1</sup> and Guoqiang Li<sup>2</sup> <sup>1</sup>Lanzhou University, CN; <sup>2</sup>Shanghai Jiao Tong University, CN</p> <p><b>Abstract</b> <i>OSEK/VDX is a specification for vehicle-mounted systems. Currently, the specification has been widely adopted by many automotive companies to develop a distributed vehicle application system. However, the ever increasing complexity of the developed distributed application system has created a challenge for exhaustively ensuring its reliability. Model checking as an exhaustive technique has been applied to verify OSEK/VDX distributed application systems to discover subtle errors. Unfortunately, it faces a poor scalability for practical systems because the verification models derived from such systems are highly complex. This paper presents an efficient approach that addresses this problem by reducing the complexity of the verification model such that model checking can easily complete the verification.</i></p> <p><a href="#">Download Paper (PDF; Only available from the DATE venue WiFi)</a></p>

- IP2-5 **AUTOMATIC ASSERTION GENERATION FROM NATURAL LANGUAGE SPECIFICATIONS USING SUBTREE ANALYSIS**  
**Speaker:**  
Ian Hamis, University of California, Irvine, US  
**Authors:**  
Junchen Zhao and Ian Harris, University of California Irvine, US  
**Abstract**  
*We present an approach to generate assertions from natural language specifications by performing semantic analysis of sentences in the specification document. Other techniques for automatic assertion generation use information found in the design implementation, either by performing static or dynamic analysis. Our approach generates assertions directly from the specification document, so bugs in the implementation will not be reflected in the assertions. Our approach parses each sentence and examines the resulting syntactic parse trees to locate subtrees which are associated with important phrases, such as the antecedent and consequent of an implication. Formal assertions are generated using the information inside these subtrees to fill a set of assertion templates which we present. We evaluate the effectiveness of our approach using a set of statements taken from a real specification document.*  
[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)
- IP2-6 **DETECTION OF HARDWARE TROJANS IN SYSTEMC HLS DESIGNS VIA COVERAGE-GUIDED FUZZING**  
**Speaker:**  
Niklas Bruns, Cyber-Physical Systems, DFKI GmbH, DE  
**Authors:**  
Hoang M. Le, Daniel Grosse, Niklas Bruns and Rolf Drechsler, University of Bremen, DE  
**Abstract**  
*High-level Synthesis (HLS) is being increasingly adopted as a mean to raise design productivity. HLS designs, which can be automatically translated into RTL, are typically written in SystemC at a more abstract level. Hardware Trojan attacks and countermeasures, while well-known and well-researched for RTL and below, have been only recently considered for HLS. The paper makes a contribution to this emerging research area by proposing a novel detection approach for Hardware Trojans in SystemC HLS designs. The proposed approach is based on coverage-guided fuzzing, a new promising idea from software (security) testing research. The efficiency of the approach in identifying stealthy behavior is demonstrated on a set of open-source benchmarks.*  
[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)
- IP2-7 **DESIGN OPTIMIZATION FOR HARDWARE-BASED MESSAGE FILTERS IN BROADCAST BUSES**  
**Speaker:**  
Lea Schönberger, TU Dortmund University, DE  
**Authors:**  
Lea Schönberger, Georg von der Brüggen, Horst Schirmeier and Jian-Jia Chen, Technical University of Dortmund, DE  
**Abstract**  
*In the field of automotive engineering, broadcast buses, e.g., Controller Area Network (CAN), are frequently used to connect multiple electronic control units (ECUs). Each message transmitted on such buses can be received by each single participant, but not all messages are relevant for every ECU. For this purpose, all incoming messages must be filtered in terms of relevance by either hardware or software techniques. We address the issue of designing hardware filter configurations for clients connected to a broadcast bus in order to reduce the cost, i.e., the computation overhead, provoked by undesired but accepted messages. More precisely, we propose an SMT formulation that can be applied to i) retrieve a (minimal) perfect filter configuration, i.e., no undesired messages are received, ii) optimize the filter quality under given hardware restrictions, or iii) minimize the hardware cost for a given type of filter component and a maximum cost threshold.*  
[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)
- IP2-8 **VEHICLE SEQUENCE REORDERING WITH COOPERATIVE ADAPTIVE CRUISE CONTROL**  
**Speaker:**  
Yun-Yun Tsai, National Tsing Hua University, TW  
**Authors:**  
Ta-Wei Huang<sup>1</sup>, Yun-Yun Tsai<sup>1</sup>, Chung-Wei Lin<sup>2</sup> and Tsung-Yi Ho<sup>1</sup>  
<sup>1</sup>National Tsing Hua University, TW; <sup>2</sup>National Taiwan University, TW  
**Abstract**  
*With Cooperative Adaptive Cruise Control (CACC) systems, vehicles are allowed to communicate and cooperate with each other to form platoons and improve the traffic throughput, traffic performance, and energy efficiency. In this paper, we take into account the braking factors of different vehicles so that there is a desired platoon sequence which minimizes the platoon length. We formulate the vehicle sequence reordering problem and propose an algorithm to reorder vehicles to their desired platoon sequence.*  
[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)
- IP2-9 **USING STATISTICAL MODEL CHECKING TO ASSESS RELIABILITY FOR BATHTUB-SHAPED FAILURE RATES**  
**Speaker and Author:**  
Josef Strnad, Brno University of Technology, CZ  
**Abstract**  
*Ideally, the reliability can be assessed analytically, provided that an analytical solution exists and its presumptions are met. Otherwise, alternative approaches to the assessment must apply. This paper proposes a novel, simulation based approach that relies on stochastic timed automata. Based on the automata, our paper explains principles of creating reliability models for various scenarios. Our approach expects that a reliability model is then processed by a statistical model checking method, used to assess the reliability by statistical processing of simulation results over the model. Main goal of this paper is to show that instruments of stochastic timed automata and statistical model checking are capable of facilitating the assessment process even for adverse conditions such as bathtub shaped failure rates.*  
[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)
- IP2-10 **EMPIRICAL EVALUATION OF IC3-BASED MODEL CHECKING TECHNIQUES ON VERILOG RTL DESIGNS**  
**Speaker:**  
Aman Goel, University of Michigan, US  
**Authors:**  
Aman Goel and Karem Sakallah, University of Michigan, US  
**Abstract**  
*IC3-based algorithms have emerged as effective scalable approaches for hardware model checking. In this paper we evaluate six implementations of IC3-based model checkers on a diverse set of publicly-available and proprietary industrial Verilog RTL designs. Four of the six verifiers we examined operate at the bit level and two employ abstraction to take advantage of word-level RTL semantics. Overall, the word-level verifier employing data abstraction outperformed the others, especially on the large industrial designs. The analysis helped us identify several key insights on the techniques underlying these tools, their strengths and weaknesses, differences and commonalities, and opportunities for improvement.*  
[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)

IP2-11 **CO-DESIGN IMPLICATIONS OF ON-DEMAND-ACCELERATION FOR CLOUD HEALTHCARE ANALYTICS: THE AEGLE APPROACH**

**Speaker:**

Konstantina Koliogeorgi, National Technical University of Athens, GR

**Authors:**

Dimosthenis Masouros<sup>1</sup>, Konstantina Koliogeorgi<sup>1</sup>, Georgios Zervakis<sup>1</sup>, Alexandra Kosvyra<sup>2</sup>, Achilleas Chytas<sup>2</sup>, Sotirios Xydis<sup>1</sup>, Ioanna Chouvarda<sup>2</sup> and Dimitrios Soudris<sup>3</sup>

<sup>1</sup>National Technical University of Athens, GR; <sup>2</sup>Aristotle University of Thessaloniki, GR; <sup>3</sup>Democritus University of Thrace, GR

**Abstract**

Nowadays, big data and machine learning are transforming the way we realize and manage our data. Even though the healthcare domain has recognized big data analytics as a prominent candidate, it has not yet fully grasped their promising benefits that allow medical information to be converted to useful knowledge. In this paper, we introduce AEGLE's big data infrastructure provided as a Platform as a Service. Utilizing the suite of genomic analytics from the Chronic Lymphocytic Leukaemia (CLL) use case, we show that on-demand acceleration is profitable w.r.t a pure software cloud-based solution. However, we further show that on-demand acceleration is not offered as a "free-lunch" and we provide an in-depth analysis and lessons learnt on the co-design implications to be carefully considered for enabling cost-effective acceleration at the cloud-level.

[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)

IP2-12 **MODULAR FPGA ACCELERATION OF DATA ANALYTICS IN HETEROGENOUS COMPUTING**

**Speaker:**

Christoforos Kachris, ICCS-NTUA, GR

**Authors:**

Christoforos Kachris, Dimitrios Soudris and Elias Koromilas, Democritus University of Thrace, GR

**Abstract**

Emerging cloud applications like machine learning, AI and big data analytics require high performance computing systems that can sustain the increased amount of data processing without consuming excessive power. Towards this end, many cloud operators have started deploying hardware accelerators, like FPGAs, to increase the performance of computational intensive tasks but increasing the programming complexity to utilize these accelerators. VINEYARD has developed an efficient framework that allows the seamless deployment and utilization of hardware accelerators in the cloud without increasing the programming complexity and offering the flexibility of software packages. This paper presents a modular approach for the acceleration of data analytics using FPGAs. The modular approach allows the automatic development of integrated hardware designs for the acceleration of data analytics. The proposed framework shows the data analytics modules can be used to achieve up to 10x speedup compared to high performance general-purpose processors.

[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)

IP2-13 **ACDC: AN ACCURACY- AND CONGESTION-AWARE DYNAMIC TRAFFIC CONTROL METHOD FOR NETWORKS-ON-CHIP**

**Speaker:**

Siyuan Xiao, South China University of Technology, CN

**Authors:**

Siyuan Xiao<sup>1</sup>, Xiaohang Wang<sup>1</sup>, Maurizio Palesi<sup>2</sup>, Amit Kumar Singh<sup>3</sup> and Terrence Mak<sup>4</sup>

<sup>1</sup>South China University of Technology, CN; <sup>2</sup>University of Catania, IT; <sup>3</sup>University of Essex, GB; <sup>4</sup>University of Southampton, GB

**Abstract**

Many applications exhibit error forgiving features. For these applications, approximate computing provides the opportunity of accelerating the execution time or reducing power consumption, by mitigating computation effort to get an approximate result. Among the components on a chip, network-on-chip (NoC) contributes a large portion to system power and performance. In this paper, we exploit the opportunity of aggressively reducing network congestion and latency by selectively dropping data. Essentially, the importance of the dropped data is measured based on a quality model. An optimization problem is formulated to minimize the network congestion with constraint of the result quality. A lightweight online algorithm is proposed to solve this problem. Experiments show that on average, our proposed method can reduce the execution time by as much as 12.87% and energy consumption by 12.42% under strict quality requirement, speedup execution by 19.59% and reduce energy consumption by 21.20% under relaxed requirement, compared to a recent work on approximate computing approach for NoCs.

[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)

IP2-14 **POWER AND PERFORMANCE OPTIMAL NOC DESIGN FOR CPU-GPU ARCHITECTURE USING FORMAL MODELS**

**Speaker:**

Nader Bagherzadeh, University of California Irvine, US

**Authors:**

Lulwah Alhubail and Nader Bagherzadeh, University of California - Irvine, US

**Abstract**

Heterogeneous computing architectures that fuse both CPU and GPU on the same chip are common now-a-days. Using homogeneous interconnect for such heterogeneous processors each with different network demands can result in performance degradation. In this paper, we focused on designing a heterogeneous mesh-style network-on-chip (NoC) to connect heterogeneous CPU-GPU processors. We tackled three problems at once; mapping Processing Elements (PEs) to the routers of the mesh, assigning the number of virtual channels (VC), and assigning the buffer size (BS) for each port of each router in the NoC. By relying on formal models, we developed a method based on Strength Pareto Evolutionary Algorithm2 (SPEA2) to obtain the Pareto optimal set that optimizes communication performance and power consumption of the NoC. By validating our method on a full-system simulator, results show that the NoC performance can be improved by 17% while minimizing the power consumption by at least 2.3x and maintaining the overall system performance.

[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)

IP2-15 **DEEP LEARNING-BASED CIRCUIT RECOGNITION USING SPARSE MAPPING AND LEVEL-DEPENDENT DECAYING SUM CIRCUIT REPRESENTATION**

**Speaker:**

Massoud Pedram, University of southern california, US

**Authors:**

Arash Fayyazi<sup>1</sup>, Soheil Shababi<sup>2</sup>, Pierluigi Nuzzo<sup>2</sup>, Shahin Nazarian<sup>2</sup> and Massoud Pedram<sup>1</sup>

<sup>1</sup>University of southern california, US; <sup>2</sup>University of Southern California, US

**Abstract**

Efficiently recognizing the functionality of a circuit is key to many applications, such as formal verification, reverse engineering, and security. We present a scalable framework for gate-level circuit recognition that leverages deep learning and a convolutional neural network (CNN)-based circuit representation. Given a standard cell library, we present a sparse mapping algorithm to improve the time and memory efficiency of the CNN-based circuit representation. Sparse mapping allows encoding only the logic cell functionality, independently of implementation parameters such as timing or area. We further propose a data structure, termed level-dependent decaying sum (LDDS) existence vector, which can compactly represent information about the circuit topology. Given a reference gate in the circuit, an LDDS vector can capture the function of the gates in the input and output cones as well as their distance (number of stages) from the reference. Compared to the baseline approach, our framework obtains more than an-order-of-magnitude reduction in the average training time and 2x improvement in the average runtime for generating CNN-based representations from gate-level circuits, while achieving 10% higher accuracy on a set of benchmarks including EPFL and ISCAS'85 circuits.

[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)

IP2-16 PARTIAL ENCRYPTION OF BEHAVIORAL IPS TO SELECTIVELY CONTROL THE DESIGN SPACE IN HIGH-LEVEL SYNTHESIS

**Speaker:**

Farah Taher, The University of Texas at Dallas, US

**Authors:**

Zi Wang and Benjamin Carrion Schaefer, The University of Texas at Dallas, US

**Abstract**

*Abstract—Commercial High-Level Synthesis(HLS) tool vendors have started to enable ways to protect Behavioral IP (BIPs) from being unlawfully used. The main approach is to provide tools to encrypt these BIPs which can be decrypted by the HLS tool only. The main problem with this approach is that encrypting the IP does not allow BIP users to insert synthesis directives into the source code in the form of pragmas (comments), and hence cancels out one of the most important advantages of C-based VLSI design: The ability to automatically generate micro-architectures with unique design metrics, e.g. area, power and performance. This work studies the impact to the search space when synthesis directives are not able to be inserted in to the encrypted IP source code while other options are still available to the BIP users (e.g. setting global synthesis options and limiting the number and type of functional units) and proposes a method that selectively controls the search space by encrypting different portions of the BIP. To achieve this goal we propose a fast heuristic based on divide and conquer method. Experimental results show that our proposed method works well compared to an exhaustive search that leads to the optimal solution.*

[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)

IP2-17 SOFTWARE-HARDWARE CO-DESIGN OF MULTI-STANDARD DIGITAL BASEBAND PROCESSOR FOR IOT

**Speaker:**

Carolynn Bernier, CEA-Leti, FR

**Authors:**

Hela Belhadj Amor and Carolynn Bernier, CEA, LETI, FR

**Abstract**

*This work demonstrates an ultra-low power, software-defined wireless transceiver designed for IoT applications using an open-source 32-bit RISC-V core. The key driver behind this success is an optimized hardware/software partitioning of the receiver's digital signal processing operators. We benchmarked our architecture on an algorithm for the detection of FSK-modulated frames using a RISC-V compatible core and ARM Cortex-M series processors. We use only standard compilation tools and no assembly-level optimizations. Our results show that Bluetooth LE frames can be detected with an estimated peak core power consumption of 1.6 mW on a 28 nm FDSOI technology, and falling to less than 0.6 mW (on average) during symbol demodulation. This is achieved at nominal voltage. Compared to state of the art, our work offers a power efficient alternative to the design of dedicated baseband processors for ultra-low power software-defined radios with a low software complexity.*

[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)

IP2-18 TAMING DATA CACHES FOR PREDICTABLE EXECUTION ON GPU-BASED SOCS

**Speaker:**

Björn Forsberg, ETH Zürich, CH

**Authors:**

Björn Forsberg<sup>1</sup>, Luca Benini<sup>2</sup> and Andrea Marongiu<sup>2</sup>

<sup>1</sup>ETH Zürich, CH; <sup>2</sup>Università di Bologna, IT

**Abstract**

*Heterogeneous SoCs (HeSoCs) typically share a single DRAM between the CPU and GPU, making workloads susceptible to memory interference, and predictable execution troublesome. State-of-the-art predictable execution models (PREM) for HeSoCs prefetch data to the GPU scratchpad memory (SPM), for computations to be insensitive to CPU-generated DRAM traffic. However, the amount of work that the small SPM sizes allow is typically insufficient to absorb CPU/GPU synchronization costs. On-chip caches are larger, and would solve this issue, but have been argued too unpredictable due to self-evictions. We show how self-eviction can be minimized in GPU caches via clever managing of prefetches, thus lowering the performance cost, while retaining timing predictability.*

[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)

IP2-19 DESIGN AND EVALUATION OF SMALLFLOAT SIMD EXTENSIONS TO THE RISC-V ISA

**Speaker:**

Giuseppe Tagliavini, Università di Bologna, IT

**Authors:**

Giuseppe Tagliavini<sup>1</sup>, Stefan Mach<sup>2</sup>, Davide Rossi<sup>1</sup>, Andrea Marongiu<sup>1</sup> and Luca Benini<sup>1</sup>

<sup>1</sup>Università di Bologna, IT; <sup>2</sup>ETH Zurich, CH

**Abstract**

*RISC-V is an open-source instruction set architecture (ISA) with a modular design consisting of a mandatory base part plus optional extensions. The RISC-V 32IMFC ISA configuration has been widely adopted for the design of new-generation, low-power processors. Motivated by the important energy savings that smaller-than-32-bit FP types have enabled in several application domains and related compute platforms, some recent studies have published encouraging early results for their adoption in RISC-V processors. In this paper we introduce a set of ISA extensions for RISC-V 32IMFC, supporting scalar and SIMD operations (fitting the 32-bit register size) for 8-bit and two 16-bit FP types. The proposed extensions are enabled by exposing the new FP types to the standard C/C++ type system and an implementation for the RISC-V GCC compiler is presented. As a further, novel contribution, we extensively characterize the performance and energy savings achievable with the proposed extensions. On average, experimental results show that their adoption provide benefits in terms of performance (1.64x speedup for 16-bit and 2.18x for 8-bit types) and energy consumption (30% saving for 16-bit and 50% for 8-bit types). We also illustrate an approach based on automatic precision tuning to make effective use of the new FP types.*

[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)

IP2-20 vDARM: DYNAMIC ADAPTIVE RESOURCE MANAGEMENT FOR VIRTUALIZED MULTIPROCESSOR SYSTEMS

**Speaker:**

Jianmin Qian, Shanghai Jiao Tong University, CN

**Authors:**

Jianmin Qian, Jian Li, Ruhui Ma and Haibing Guan, Shanghai Jiao Tong University, CN

**Abstract**

*Modern data center servers have been enhancing their computing capacity by increasing processor counts. Meanwhile, these servers are highly virtualized to achieve efficient resource utilization and energy savings. However, due to the shifting of server architecture to non-uniform memory access (NUMA), current hypervisor-level or OS-level resource management methods continue to be challenged in their ability to meet the performance requirement of various user applications. In this work, we first build a performance slowdown model to accurately identify the current system overheads. Based on the model, we finally design a dynamic adaptive virtual resource management method (vDARM) to eliminate the runtime NUMA overheads by re-configuring virtual-to-physical resource mappings. Experiment results show that, compared with state-of-art approaches, vDARM can bring up an average performance improvement of 36.2% on a 8-node NUMA machines. Meanwhile, vDARM only incurs extra CPU utilization no more than 4%.*

[Download Paper \(PDF; Only available from the DATE venue WiFi\)](#)