PhD Thesis Proposal Form
China Scholarship Council (CSC)/ENS Rennes
Call for projects 2021

FIELD: Computer science

THESIS SUBJECT TITLE: Imprecise Computation for Mixed-Critical systems on Multicore platforms

1. Single French PhD proposal:

- Laboratory name: IRISA (http://www.irisa.fr/en)
  CAIRN team (https://team.inria.fr/cairn/)
- PhD director (contact person):
  - Name: Emmanuel Casseau
  - Position: Professor
  - E-mail: emmanuel.casseau@irisa.fr
  - Phone number: + 33 2 96 46 91 76

2. Co-directed or a joint PhD, please specify:

- Joint PhD (cotutelle): ~~YES~~ or **NO**
- Co-directed PhD: **YES** or ~~NO~~

- Partner university name: Southeast University, China
- Laboratory name and web site: School of Automation (https://automation.seu.edu.cn/)
-
- PhD director-s in partner university (contact person):
  - Name: Lei Mo
  - Position: Associate Professor
  - E-mail: lmo@seu.edu.cn
  - Phone number: +86-18668109963

- If previous collaborations with the Chinese co-director/university, please detail:

  We have active research collaboration with Assoc. Prof. Lei Mo, who is the co-director of this proposal and who was previously a post-doctoral researcher in our IRISA lab and now is with the Southeast University of China. We have 11 common publications related to the subject of this thesis proposal, with 6 articles published in top-ranked ranked internationals journals (4 x Q1 (IEEE TII, IEEE JIOT, IEEE JETCAS) and 2 x Q2 (IEEE TCAD, Springer P2PNA)), 4 articles published in highly ranked conferences (DATE, ICCD, SAMOS, AdHoc-Now) and 1 book chapter.

  We have also joint publications with following universities: the Zhejiang University, China (Prof. Shibo He) in collaborative scheduling and control on Networked Control Systems, and the Southeast University, China (Prof. Xianghui Cao) in resource optimization and task allocation on Cyber-Physical Systems. The PhD director also has active collaboration in combinatorial optimization problems for custom processors with his former PhD student, Chenglong Xiao, who is currently a Professor in Shantou University.

  Furthermore, the University of Rennes 1, where the PhD director belongs to, had a common Master program (M2 Embedded System) with the Southeast University in Nanjing, China. Several members of our team, CAIRN, have travelled to China and given courses under this collaboration.

- Interest of the Joint PhD for the French co-director, for his/her laboratory, for ENS Rennes:

  Our motivation for the PhD is to establish a long-term research collaboration with Chinese Universities, especially with the Southeast University of China, in the domain of electronics and embedded systems and especially multicore-manycore architectures. We strongly believe that this PhD will promote international exchanges and cooperation, as it allows our team to recruit students with high technical skills required in our domain and at the same time opens opportunities for potential collaboration in international research projects and student exchanges with Chinese universities, research centers and innovation companies.

- Thesis proposal (max 1500 words):

  The embedded systems from **safety-critical** domain industries, such as the avionics, automotive, space, healthcare or robotics industries, usually consist of mixed-critical systems, where high-criticality applications are executed with low-criticality applications. For example, avionics systems consist of applications with different design assurance levels, space systems consist of navigation control applications and scientific applications, automotive systems have applications in the same wheel sensor for stability control and for the acceleration regulation. Embedded systems have to provide **timing guarantees**, at least for the high-criticality applications, in order to be safe, i.e., guarantee that tasks are completed before their respective deadlines and/or the total execution does not exceed a given latency requirement.

  Within last thirty years, the code size of automotive, space and avionics applications has significantly increased. The mixed-critical systems face exponential growth in performance requirements, whereas future automotive and aerospace applications will require higher performance computing resources. This constantly growing processing demand has led the processor manufacturing industry towards **multicore platforms**. These platforms have multiple processor elements, called cores, providing massive computing power by concurrently executing a high volume of tasks. While such architectures can successfully meet the demands for the majority of computing systems, the same cannot be argued for hard real-time systems, due to the unpredictable timing behavior of multicore platforms.

  Unpredictable timing behaviour originates from resource sharing among cores. Integrating high-criticality and low-criticality applications on the same platform leads to concurrent accesses to shared resources, introducing timing delays (interferences), highly affecting applications' timing behaviour in a non-deterministic manner. To provide a predictable timing behaviour, either spatio-temporal isolation is enforced to avoid interferences, or Worst-Case Execution Time (WCET) bounds have to be overestimated. In both cases, the result is a, highly undesired, sub-optimal use of resources, and **system over-provisioning**. Actually, the former solution assigns resources to critical applications and prohibits their use by other applications, even in idle state. The latter solution has to consider the worst case during WCET estimations: for every access to a shared resource by a core, it is assumed that the remaining cores access the same resource at the same time. Consequently, the WCET estimations are overly pessimistic. This WCET over-approximation practice has led to the "one-out-of-m processors" problem, i.e., the processing capacity of multicore architectures is negated by the WCET pessimism. As a result, sequential execution on a single core may provide better timing guarantees than any parallel execution, seriously undermining the advantages of using multicores.

  However, it is shown that applications in mixed critical systems can accept approximate results as long as the baseline Quality of Service (QoS) is satisfied. For instance, frames of fuzzy images in image processing applications and rough estimates of location in tracking applications produced in time are better results than perfect images and accurate location produced too late. In these domains, the applications can be **imprecisely computed**: ideally, a task is logically decomposed into a set of mandatory subtasks and a set of optional subtasks. The mandatory subtasks must be completed before the deadline to have an acceptable result, while the optional subtasks can be left incomplete, if there are no available resources or time left, at the cost of reduced quality. When either resources or time is available, the QoS of the intermediate results can be increased, as the optional subtasks are executed longer.

On the one hand, existing mixed-criticality scheduling and assignment approaches focus on providing guaranties for high-criticality tasks, while considering precise computation for both high-criticality and low-criticality tasks. Typical examples of such approaches come from scheduling theory, such as approaches considering pessimistic WCET and approaches that incorporate interference analysis. As the tasks have to be precisely computed, no exploitation is possible that reduces the executed code, and thus the WCET, while the baseline QoS is still satisfied. On the other hand, scheduling approaches based on imprecise computation have as goal to maximize QoS under energy supply and real-time constraints. However, most of these works consider only tasks of the same criticality. As no difference exist in the criticality of tasks, no exploitation can occur by prioritizing the high-criticality tasks.

The aim of this thesis is to bridge this gap in the state-of-the-art by combining mixed-critical systems and imprecise computation on multicore architectures. The goal of the thesis is to design novel methodologies to efficiently solve the problem of task execution on multicore platforms, under real-time constraints, with the objective of reducing the WCET pessimism and maximize the system QoS.

In order to summarize the main methodology steps to achieve this goal, first a simulation and profiling framework will be developed to estimate the required QoS for an application based on a given context, inspired by system-scenario techniques. This QoS estimation is required, since the same application running on different input data (a.k.a. context) does not require the same QoS. For instance, aircraft collision detection systems compute trajectories using radar data. Trajectories for stationary obstacles are computationally simpler that those of moving obstacles. Designing the system to compute trajectories always considering moving obstacles, even when it is not required, increases the WCET and thus over-designs the system. Then, an innovative methodology will be developed to parametrically link different application regions with the obtained QoS. Applications have critical regions, for which correct execution is required, and forgiving regions, which can tolerate changes. This link will be established based on information obtained by exploring i) loop iteration skipping via QoS-aware profilers or statistical approaches, and ii) code skipping via similar methods to task skipping. Using this correlation, the application will be modelled as a set of Mandatory (M) subtasks, that provide the required QoS, and optional subtasks, able to further increase QoS. As a next step, a parametric WCET model will be proposed based on the application context. Last, but not least, innovative methodologies will be developed to perform near-optimal subtasks scheduling on the multicore architecture taking into account interference impact. We are interested in both approaches that provide the optimal solution and approaches able to degrade the obtained solution in a controlled way.

The benefits of this thesis are the integration of mixed-critical applications on a multicore platform meeting real-time and QoS requirements, while supporting the replacement of the over-provisioning existing infrastructures, i.e. leading to a significant reduction of the physical system components. As a result, the system cost, energy and maintenance are reduced.

- Publications of the laboratory in the field (max 5):

[1] M. Cui, L. Mo, A. Kritikakou, and E. Casseau, “Energy-aware Partial-Duplication Task Mapping under Real-Time and Reliability Constraints”, Int’l Conference on Embedded Computer Systems: Architectures, Modeling and Simulation, SAMOS XX, 2020.

[2] L. Mo, A. Kritikakou and O. Sentieys, “Approximation-aware task deployment on asymmetric multicore processors”, in Proc. Design, Automation & Test in Europe (DATE), pp. 25-29, 2019.

[3] P. Dobias, E. Casseau, and O. Sinnen, “Comparison of Different Methods Making Use of Backup Copies for Fault-Tolerant Scheduling on Embedded Multiprocessor Systems”, Int’l Conference on Design and Architectures for Signal and Image Processing (DASIP), pp. 1-6, 2018.

[4] L. Mo, A. Kritikakou, and O. Sentieys, “Controllable QoS for imprecise computation tasks on DVFS multicores with time and energy constraints”, IEEE Journal on Emerging and Selected Topics in Circuits and Systems (IEEE JETCAS), vol. 8, no. 4, pp. 708-721, 2018.

[5] A. Kritikakou, T. Marty, and M. Roy. “DYNASCORE: DYNAmic Software COntroller to increase REsource utilization in mixed-critical systems”, ACM Transactions on Design Automation of Electronic Systems (TODAES), vol. 23, no. 2, pp. 1-25, 2017.