Offre n°2024-07893

PhD Position F/M FPGA-based Near Memory Computing Architectures

Le descriptif de l'offre ci-dessous est en Anglais

Type de contrat : CDD

Niveau de diplôme exigé : Bac + 5 ou équivalent

Fonction : Doctorant

Contexte et atouts du poste

Context and background:

Moore's law has been driving computer performance for decades through CMOS down-scaling and architecture enhancements, resulting in doubled performance every 18 months. However, current technology encounters three significant challenges, including the leakage wall, reliability wall, and cost wall. Similarly, computer architectures are confronted with three walls: the memory wall, power wall, and instruction-level parallelism (ILP) wall. Numerous novel technologies and architectures are being researched to overcome these walls and enhance performance [1], [2].

Architects and designers are compelled to seek breakthroughs in computer architecture as the **total computation costs are dominated by the energy and performance costs of moving data** between the memory subsystem and the CPU. Indeed, modern computing systems experience a significant disparity between the performance and energy efficiency of computation and memory units. Such systems adopt a processor-centric method where data must travel to and from memory units through a relatively slow and power-intensive off-chip bus to computation units for processing. Consequently, workloads that heavily rely on data necessitate constant data movement between memory and CPU, leading to a **substantial overhead in execution time and energy efficiency** [3].

Modern field-programmable gate array (FPGA) with their on-chip URAMs and BRAMs, as well as access to high-bandwidth off-chip memory, can efficiently handle irregular memory access patterns and provide significantly higher memory bandwidth than CPUs. Their architectural flexibility makes them a popular choice for various data processing tasks, including machine learning and deep learning [4]–[6], database processing [7], [8], and networking [9].

Near-memory computing (NMC) is a ***memory-centric computing*** paradigm that has emerged as a promising solution to overcome the challenges mentioned above [1]. The goal of NMC is to **perform processing in proximity to the data** location - by placing compute units near the data - and it aims to minimize costly data transfers [2], [10].

Interestingly, **FPGA accelerators exhibit near-memory computation capabilities** , enabled by high-bandwidth memory (HBM), which are much closer to the logic than traditional DDR4-based FPGA boards. HBM technologies that use three-dimensional stacking enable processing close to memory by integrating logic and memory on a single chip. This innovative approach involves packing several layers vertically using through-silicon vias, which reduces memory access latency and energy consumption while increasing bandwidth [2].

In a typical architecture of an FPGA-based NMC system, two main entities are tightly integrated: a host CPU - operating at GHz frequencies and communicating with host memory in the order of 10-100 Gbps - and an FPGA fabric - operating at hundreds of MHz frequencies and communicating with its dedicated memory in the order of 400 Gbps. This architecture enables **efficient offloading of acceleration kernels from the host CPU to the FPGA**, allowing them to be **processed close to the memory** [2] [3]. FPGAs have been used to accelerate major kernels of modern data-intensive applications – such as genome analysis and weather prediction – by using the Near-memory computing paradigm [3].

Utilizing the full potential of FPGAs to accelerate an application is not a straightforward task. To achieve significant performance gains, FPGAs must exhibit at least ten times more parallelism than a typical workload. This necessitates extensive knowledge of FPGA programming to effectively map the application kernel and optimize the design for the FPGA microarchitecture.

[1] A. Gebregiorgis *et al.*, “A Survey on Memory-Centric Computer Architectures,” *ACM J. Emerg. Technol. Comput. Syst.*, p. 3544974, Jun. 2022, doi: 10.1145/3544974.

[2] V. Iskandar, M. A. A. E. Ghany, and D. Göhringer, “Near-memory Computing on FPGAs with 3D-stacked Memories: Applications, Architectures, and Optimizations,” *ACM Trans. Reconfigurable Technol. Syst.*, vol. 16, no. 1, p. 16:1-16:32, Dec. 2022, doi: 10.1145/3547658.

[3] G. Singh *et al.*, “FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications,” *IEEE Micro*, vol. 41, no. 4, pp. 39–48, Jul. 2021, doi: 10.1109/MM.2021.3088396.

[4] J. Fowers *et al.*, “A Configurable Cloud-Scale DNN Processor for Real-Time AI,” in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, Jun. 2018, pp. 1–14. doi: 10.1109/ISCA.2018.00012.

[5] K. Kara, D. Alistarh, G. Alonso, O. Mutlu, and C. Zhang, “FPGA-Accelerated Dense Linear Machine Learning: A Precision-Convergence Trade-Off,” in *2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, Apr. 2017, pp. 160–167. doi: 10.1109/FCCM.2017.39.

[6] Y. Umuroglu *et al.*, “FINN: A Framework for Fast, Scalable Binarized Neural Network Inference,” in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-*

Programmable Gate Arrays, Monterey California USA, Feb. 2017, pp. 65–74. doi: 10.1145/3020078.3021744.

[7] Y. Choi, Y. Chi, W. Qiao, N. Samardzic, and J. Cong, “HBM Connect: High-Performance HLS Interconnect for FPGA HBM,” in *The 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, New York, NY, USA, Feb. 2021, pp. 116–126. doi: 10.1145/3431920.3439301.

[8] K. Kara, J. Giceva, and G. Alonso, “FPGA-based Data Partitioning,” in *Proceedings of the 2017 ACM International Conference on Management of Data*, New York, NY, USA, May 2017, pp. 433–445. doi: 10.1145/3035918.3035946.

[9] M. Ruiz, D. Sidler, G. Sutter, G. Alonso, and S. López-Buedo, “Limago: An FPGA-Based Open-Source 100 GbE TCP/IP Stack,” in *2019 29th International Conference on Field Programmable Logic and Applications (FPL)*, Sep. 2019, pp. 286–292. doi: 10.1109/FPL.2019.00053.

[10] G. Singh *et al.*, “Near-memory computing: Past, present, and future,” *Microprocess. Microsyst.*, vol. 71, p. 102868, Nov. 2019, doi: 10.1016/j.micpro.2019.102868.

Mission confiée

Ph.D. thesis goal:

Through this Ph.D. thesis, we want to **investigate different FPGA-powered NMC architectural choices to accelerate data-intensive applications** to identify the most suitable ones. Moreover, we aim to **provide high-level mechanisms** to enable users to make informed choices on **the most suitable architecture for their kernel/application to accelerate**.

Principales activités

As an example, among others, we want to explore the NMC architecture structured as follows.

Multiple highly parallel NMC Processing Elements (PEs) inside the FPGA fabric, whose number, dimension, communication type, and computation capabilities depend on the kernel/application to accelerate. Each PE will be composed of a RISC-V-based core and of a memory array and the PEs will be connected through a communication infrastructure. Memory array dimension, communication infrastructure type (e.g. shared/hierarchical/matrix bus, Network-On-Chip), and core computational capability will depend on the final kernel/application to accelerate: applications in different domains will need (i) different numbers of PEs, (ii) different ISA instructions to be implemented by the core, (iii) different dimensions of memory array depending on their degree of parallelism, and (iv) a different communication infrastructure to manage the correct amount of communications between the PEs.

High Level Synthesis (HLS) techniques will contribute to the high-level abstraction of these concepts and allow engineers and developers to **accelerate their applications as transparently as possible.**

In fact, low-level hardware description languages demand proficiency in hardware design and hands-on experience. Conversely, higher-level approaches like HLS can decrease the programming efforts needed by FPGA developers. Nevertheless, while HLS enhances programmability, **attaining high performance requires careful consideration of the hardware** pipeline and memory subsystem design, which will be one of the focuses of the thesis.

Compétences

Hardware design: VHDL/Verilog, HW synthesis flow (design, simulation, synthesis, and deployment through commercial tools for FPGA)

Experience in **Computer architecture**: Instruction Set Architecture (ISA), Microarchitecture, and Systems design.

SW Programming/Scripting: C/C++, Python, Linux scripting

Experience with **High-Level Synthesis (HLS)** and related tools (e.g., Vivado/Vitis HLS or Siemens Catapult) is a plus.

Experience with **Design Space Exploration** and **Optimization** approaches is a plus.

Languages : proficiency in written **English** is required. Fluency in spoken English or French is required.

Relational skills : the candidate will work in a research team, where regular meetings will be set up. The candidate has to be able to present the progress of their work in a clear and detailed manner.

Other valued appreciated : Open-mindedness, strong integration skills and team spirit.

The Ph.D. thesis will be carried out in the context of the PEPR Cloud, Project Archi-CESAM. The Ph.D. candidate will be supervised by the TARAN team(https://team.inria.fr/taran/) at the Inria Centre at Rennes University, IRISA laboratory, in France. The Ph.D. salary will follow standard French rates.

For more information, contact the following people at Inria Rennes / IRISA laboratory:

• Marcello Traiola: marcello.traiola@inria.fr

• Olivier Sentieys: olivier.sentieys@inria.fr

Avantages

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Rémunération

Monthly gross salary amounting to 2200€

Informations générales

- **Thème/Domaine :** Architecture, langages et compilation
  Système & réseaux (BAP E)
- **Ville :** Rennes
- **Centre Inria :** Centre Inria de l'Université de Rennes
- **Date de prise de fonction souhaitée :** 2025-06-01
- **Durée de contrat :** 3 ans
- **Date limite pour postuler :** 2025-03-31

Contacts

- **Équipe Inria :** TARAN
- **Directeur de thèse :**
  Sentieys Olivier / Olivier.Sentieys@irisa.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'e?orce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

L'essentiel pour réussir

Candidates with knowledge and experience in Embedded Systems and Hardware Design and Synthesis (especially FPGA-oriented synthesis flow) are highly appreciated.

We seek highly motivated and passionate candidates. Autonomy is a highly appreciated quality.

Essential qualities to fulfill a Ph.D. thesis are feeling at ease in an environment of scientific dynamics and wanting to learn, listen, and share.

Candidates must have a Master's degree (or equivalent) in Computer Engineering or Electronic Engineering.

Talented last year Master's students may start as 6-month interns and continue as Ph.D. researchers after graduation

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Please submit online : your resume, cover letter and letters of recommendation eventually

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.