Hybrid SORS: A Continuum of Matrix Multiplications: From Scientific Computing to Deep Learning

Objectives

Download here the slides of the presentation.

Abstract: Matrix multiplication (GEMM) is a key, pervasive computational kernel that spans across multiple domains. On the one hand, many applications arising in scientific computing require the solution of linear systems of equations, least-square problems, and eigenvalue problems. For portability, these applications often rely on linear algebra routines from LAPACK (linear algebra package). In turn, in order to deliver high performance, LAPACK heavily relies on GEMM and other Basic Linear algebra subroutines (BLAS). On the other hand, to a large extent, the computational cost for the convolutional neural networks (CNNs) that dominate machine learning algorithms for signal processing and computer vision tasks, as well as the transformers behind recent deep learning (DL) applications, such as ChatGPT, is largely determined by the performance of GEMM.

In this talk we will first expose caveats of current instances of GEMM in linear algebra libraries for conventional multicore architectures: suboptimal performance and missing support for DL-oriented data types. Starting from that point, we will then demonstrate how these problems can be overcome via tools for the (semi-)automatic generation of the only architecture-specific piece of GEMM, known as micro-kernel, together with an analytical-based model to capture the cache hierarchy configuration. In addition, we will show that this approach carries over to more "exotic" architectures, from high-end vector accelerators and the Xilinx artificial intelligence engine (AIE) to low-power designs such as RISC-V processors and ARM-based (Arduino) micro-controllers.

Short bio: Enrique S. Quintana-Orti received his bachelor and Ph.D. degrees in computer sciences from the Universitat Politecnica de Valencia (UPV), Spain, in 1992 and 1996,

respectively. After 20+ years at the Universitat Jaume I of Castellon, Spain, he came back to UPV in 2019, where he is now Professor in Computer Architecture. For his research, he received the NVIDIA 2008 Professor Partnership Award and two awards from the USA National Space Agency (NASA). He has published 400+ articles in journals and international conferences. Currently he participates in the EU projects APROPOS (approximate computing), RED-SEA (exascale computer networks), eFLOWS4HPC (workflows for HPC and AI) and Nimble AI (neuromorphic chip for sensing & processing). His research interests include parallel programming, linear algebra, energy consumption, transprecision computing and deep learning as well as advanced architectures and hardware accelerators.

Speakers

Speaker: Enrique S. Quintana-Ortí, Universitat Politècnica de València
Host: Miquel Moretó, High Performance Domain-Specific Architectures Associated Researcher, Computer Sciences

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---

Source URL (retrieved on *14 jul 2024 - 05:56*): https://www.bsc.es/ca/research-and-development/research-seminars/hybrid-sors-continuum-matrix-multiplications-scientific-computing-deep-learning