July 3rd, 2018

Vector Architecture for HPC and ML

Alejandro Rico
Arm Research

Abstract

Vector computing is one of the historical ways to improve performance per watt, area and cost. The features of the instruction set architecture are fundamental to drive the capabilities of the compiler and programmer to enable vectorization of codes. Without vector code generation, vector execution units remain underutilized and irrelevant for performance. A vector architecture must therefore include features that enable vectorization of a larger set of codes and provide support for the market segments it is targeted to. Around this philosophy, Arm and its partners designed the Scalable Vector Extension (SVE) for the Arm architecture targeting high-performance computing and machine learning. The talk will cover SVE features that enhance vectorization and enable implementation for multiple markets, such as vector-length agnostic programming, speculative vectorization, gather-scatter support and non-temporal loads and stores. The talk will also cover microarchitectural aspects that must be considered when designing a vector architecture and the set of SVE-enabled tools available to explore that microarchitectural design space.

Short bio

Alejandro Rico is a Staff Research Engineer in Arm Research working on processor architecture and microarchitecture for high-performance computing. His work at Arm focuses on scalability of parallel applications in large SoCs, and enablement, exploitation and performance analysis of vector processing using the Scalable Vector Extension. Before joining Arm, he worked in multi-core simulation methodologies and task-based parallel programs at the Barcelona Supercomputing Center. He holds a PhD from Universitat Politecnica de Catalunya and has co-authored over 20 scientific publications.