High performance tensor–vector multiples on shared memory systems

Filip Pawlowski[1,2] , Bora Uçar[2,3] , Albert-Jan Yzelman[1]
[1]Huawei Technologies France
20 Quai du Point du Jour, 92100 Boulogne-Billancourt, France
[2]ENS Lyon, France
[3]CNRS and LIP, Lyon, France
`{filip.pawlowski1,albertjan.yzelman}@huawei.com`
`bora.ucar@ens-lyon.fr`

Tensor–vector multiplication is one of the core components in tensor computations. We have recently investigated high performance, single core implementation of this bandwidth-bound operation. In this work, we investigate efficient, shared memory algorithms to carry out this operation. Upon carefully analyzing the design space, we implement a number of alternatives using OpenMP and compare them experimentally. Experimental results on up to 8 socket system show near peak performance for the proposed algorithm