TTLG: Tensor Transpose Library for GPUs*

Arjun Suresh*, Jyothi Vedurada†, Jinsung Kim*, Aravind Sukumaran-Rajam*, Ajay Panyalal, Krishna Nandivada†, Sriram Krishnamoorthyl, and P. Sadayappan*

* Ohio State University † Indian Inst. Tech., Madras l PNNL

INTRODUCTION

- General tensor transpose library
- $B_{i_1 i_2 \ldots i_n} \leftarrow A_{j_1 j_2 \ldots j_m}$
- Key kernel needed in many domains, e.g. Computational Chemistry, Machine Learning
- Multiple data movement schemes, based on distinct/non-distinct fastest-varying-indices (FVIs) in input/output tensor
- Library includes performance prediction model that can be queried
  - Useful for higher level optimizers, e.g., Tensor Contraction library/code-generator

Tensor Transposition

DISTINCT FVI: $i_0 > 32$, $O_0 > 32$

Experimental Evaluation

Current status & ongoing work

- Initial release of TTLG planned for mid Q4, 2017
- GPU tensor contraction kernels for CCSD(T) using TTLG included in the next (Nov. 2017) release of NWChem
- General tensor contraction library using TTLG now under development

Tensor Transposition

DISTINCT FVI: $i_0 < 32$, $O_0 < 32$

2D Matrix Transposition

*Funded in part by DOE-ECP project: “Exascale Code Generation Toolkit”, Dan Quinlan, PI