Abstract: A Performant Matrix of Pearson's Correlation Coefficient (MPCC) Calculations with Support for Missing Data on Emerging HPC Architectures, Y1

The work presented is motivated by the increasing computational requirements of GeneNetwork.org (GN). GN is a free and open source (FOSS) framework for web-based mouse genetics and allows biologists to upload and analyze high-throughput experimental data, such as expression data from microarrays and RNA Sequencing, as well as 'classic' phenotypes. Currently GN contains over 30 years of accumulated experimental data. One of the main analysis methods used by GN is the calculation of Pearson's Correlation Coefficient (PCC) within and between different data sets. Due to the nature of biological data PCC calculations are performed in the presence of missing data to find relationships between and among genotypes and phenotypes in mouse strains. The calculations are a bottleneck for moderate to large problem sizes. Calculating PCC is pervasive across bioinformatics, data analysis, phylogenetics, statistics, stochastics, and anthropology. Our approach can be used anywhere a matrix of PCC calculations is computed.
Results: Our solution is a reformulation of the problem such that it can be solved using matrix-matrix products, the canonical arithmetically intensive algorithm, preliminary benchmarks achieved 4.3 TFlop/s in single precision (77% of the theoretical peak) on a single Intel Xeon Gold 6148 CPU @ 2.4 GHz (Skylake) compute node. A rough estimate shows that this translates into as much as a 90x speedup over the previous approach and is independent of the percentage of missing data.