Autotuning OpenCL Workgroup Sizes

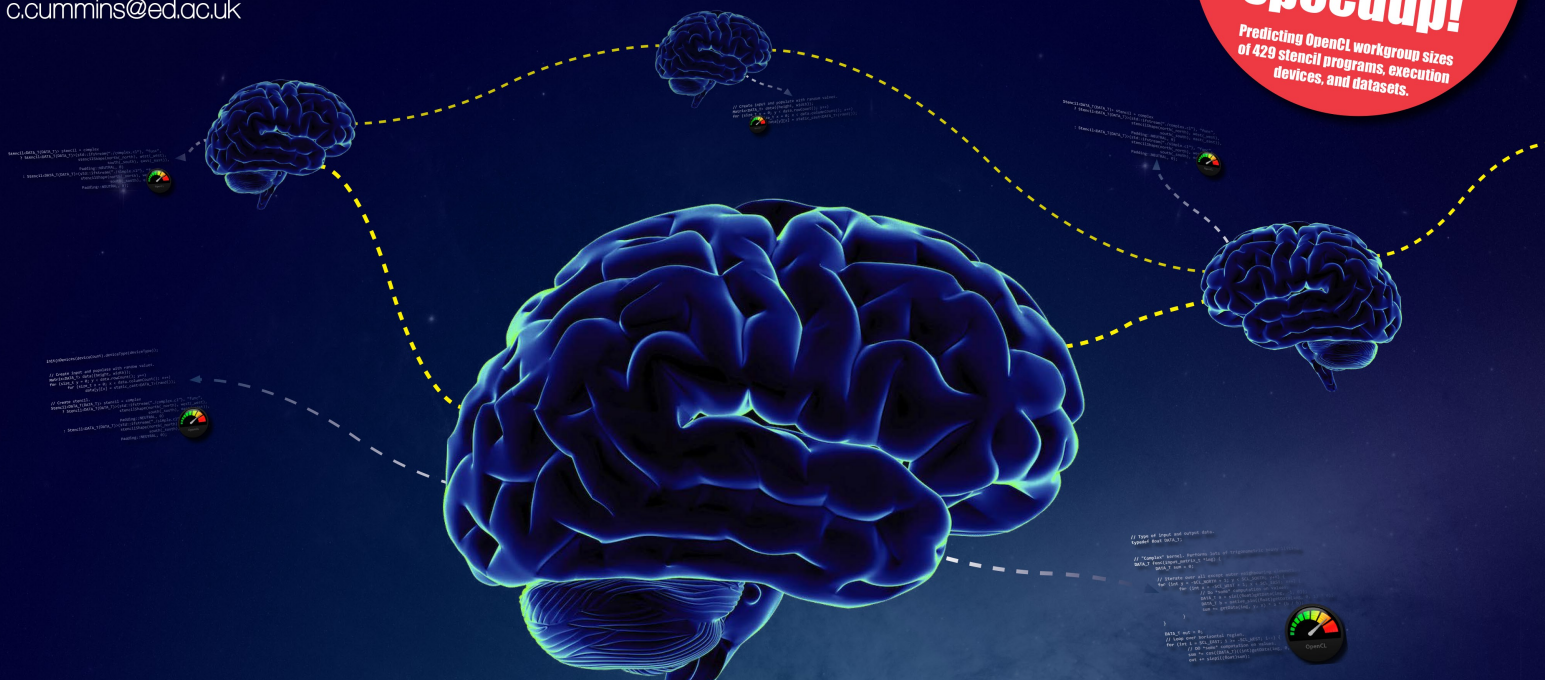
Tuning GPU Stencils with machine learning outperforms human experts

3.79x speedup!

Predicting OpenCL workgroup sizes of 429 stencil programs, execution devices, and datasets.

Chris Cummins Pavlos Petoumenos Michel Steuwer Hugh Leather

c.cummins@ed.ac.uk

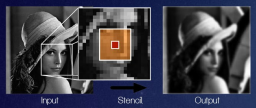


Hand tuning programs is expensive and time consuming

We automate this tuning using collaborative machine learning

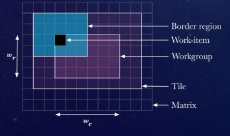
Stencil Skeletons

Stencil Skeletons are a common data parallel pattern with a range of applications from image processing to partial differential equations and cellular automata.



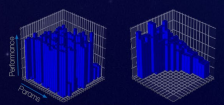
OpenCL Workgroup Size

OpenCL stencil skeletons are parameterised with a workgroup size, which controls grouping of hardware threads and local memory utilisation.

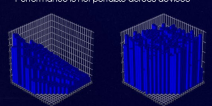


Optimization Space

Choosing the right OpenCL workgroup size for stencil kernels depends on the program, device and dataset:



Performance is not portable across devices



Performance is not portable across programs

implemented using SkelCL skelcon library

Introducing OmniTune ...

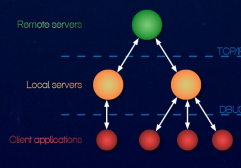
OmniTune generates synthetic benchmark programs to use for empirical testing

OmniTune collaboratively gathers performance data by testing different parameter values

OmniTune uses machine learning to predict parameters for unseen programs at runtime

Machine Learning features

OmniTune architecture



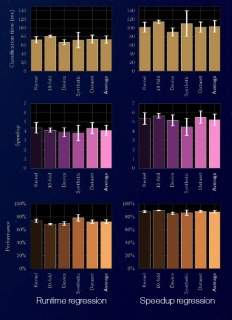
Remotes collate machine learning data across devices, accessible by a remote API. Implemented using MySQL, AWS.

Local servers train models on machine learning data, and serve requests from client applications for autotuning. Implemented using Python.

Client applications request autotuning parameters, and submit observed performance results. Implemented using SkelCL (C++ & OpenCL).

Results

Evaluated using 429 combinations of programs, devices, and datasets. Machine learning classification using linear regressors to predict program runtime of program speedup. Prediction quality evaluated across devices, programs, and datasets.



Runtime regression Speedup regression

Read more ...

C. Cummins, P. Petoumenos, M. Steuwer, H. Leather "Autotuning OpenCL Workgroup Size for Stencil Patterns" ADAPT 2016.

C. Cummins, P. Petoumenos, M. Steuwer, H. Leather "Towards Collaborative Performance Tuning of Algorithmic Skeletons" HLPGPU 2016.

http://chriscummins.cc

THE UNIVERSITY of EDINBURGH informatics

EPSRC Centre for Doctoral Training in Pervasive Parallelism

Image credits: Grant Department of Molecular & Cell Biology, Berkeley, NASA.

EPSRC Engineering and Physical Sciences Research Council