BONSAI: an OpeN Software Autotuning Infrastructure

Jakub Kurzak (PI), Piotr Luszczek (co-PI), Mark Gates, Yaohung Tsai, Matthew Bachstein

THE PREMISE

The goal of the BONSAI project is to develop a software infrastructure for using parallel hybrid systems at any scale to carry out large, concurrent autotuning sweeps in order to dramatically accelerate the optimization process of computational kernels for GPU accelerators and many-core coprocessors.

Test Many Data Sets

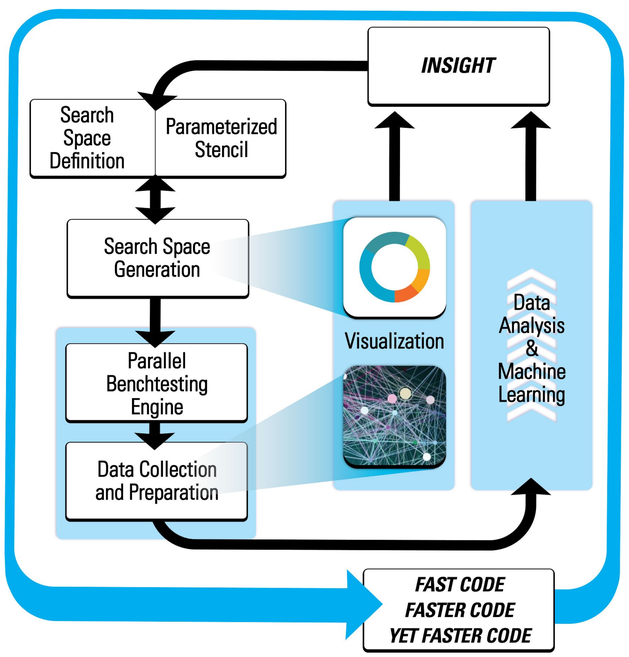
In the course of our work on accelerating the ALS algorithm for collaborative filtering, we discovered that the optimal parameter configuration depends heavily on the properties of the input data set, which motivates tuning sweeps over many datasets. In particular, consider tuning a sparse matrix kernel by making a sweep over all matrices in the University of Florida matrix collection (currently 2757 problems and growing).

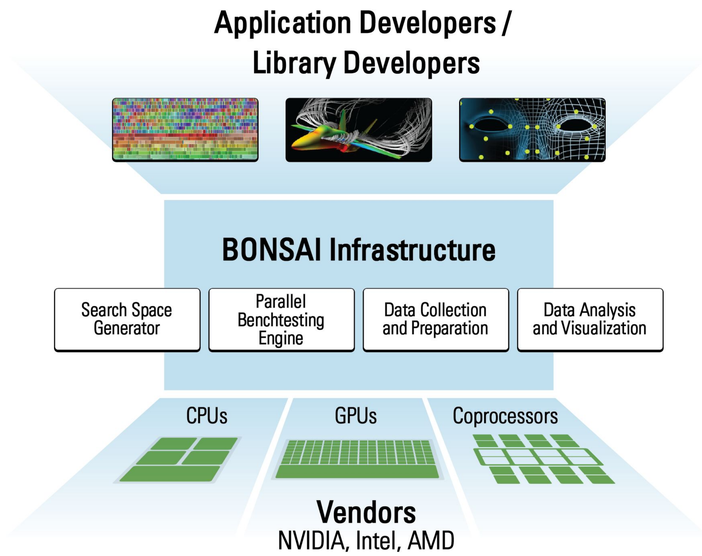
Test Many Data Layouts

Modern hardware is increasingly sensitive to the layout of data in memory. A number of different layouts have been proposed for dense linear algebra (row and column major, tile, space-filling curves), sparse linear algebra (CSC/CSR, ELLPACK, SELL-C/SELL-P, Sell-C-Sigma, BCSR, DIA, COO), deep learning (NCHW, NHWC), PDE discretizations (structure of arrays or array of structures), etc.

Collect Lots of Performance Metrics

Nvidia Maxwell can collect 111 different hardware counter metrics, based on 75 different events, and only a few can be collected in a single run. This forces many reruns of the kernel to collect all relevant metrics. Similarly, Intel Xeon Phi Knights Landing can collect 119 native events, using 5 counters, also making it necessary to rerun the kernel multiple times.





DELIVERABLE

We will develop a parallel, distributed benchmarking engine capable of scaling to tens of thousands of nodes, to benchmark millions of combinations of kernel configurations, problem sizes, and representative input datasets, while collecting hundreds of performance metrics such as time, energy consumption, cache misses, and memory bandwidth.

Compilation

BONSAI will perform parallel compilation, both across distributed memory nodes and within each node. Compilation of a large number of kernels takes a significant time in the autotuning process. Numerical kernels are frequently heavily unrolled, which contributes to long compilation times. Also, compilation time can be extremely nonuniform. Therefore, BONSAI will dynamically balance the workload.

Benchmarking

We will use a standard MPI parallel job designed to work in existing batch queuing systems such as TORQUE PBS. This will make BONSAI deployable on a wide variety of systems, from small university clusters, to cloud computing, to national supercomputer centers. When available, we will take advantage of multiple accelerators within each node.

Data Collection

BONSAI will provide a framework to simplify the process of collecting hardware counters and performance data. We will leverage the various open-source and vendor-specific libraries such as Nvidia's CUPTI API, AMD's CodeXL, Intel's VTune, and the open source PAPI library. BONSAI will simplify the task of instrumenting the kernel and provide a simple interface for selecting the counters to be collected.

Data Analysis

We will provide a number of analytical tools and examples to guide the developer in analyzing their code. The analytical tools provided with BONSAI will include statistical and machine-learning tools in addition to a number of visualization utilities. These tools will leverage open-source data analysis libraries such as the PyData stack, R, and Spark tools such as MLlib.

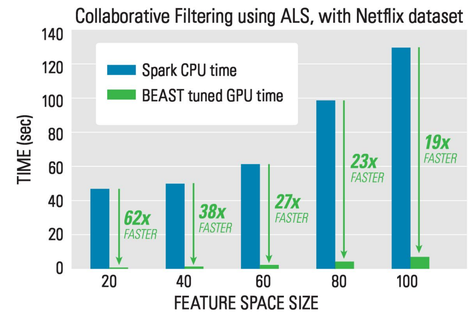
PUBLICATION TRACK RECORD

M. Tsai, P. Luszczek, J. Kurzak, J. Dongarra
Performance-Portable Autotuning of OpenCL Kernels for Convolutional Layers of Deep Neural Networks
MLHPC'16: Second Workshop on Machine Learning in HPC Environments
SC16: International Conference for High Performance Computing, Networking, Storage and Analysis, Salt Lake City, UT, 2016
DOI: 10.1109/MLHPC.2016.5

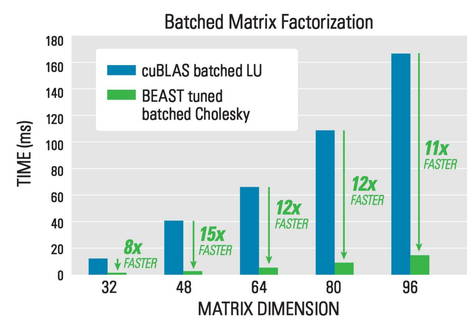
P. Luszczek, M. Gates, J. Kurzak, A. Danalis, J. Dongarra
Search Space Generation and Pruning System for Autotuners
iWAPT'16: The International Workshop on Automatic Performance Tuning
IPDPS'16: 30th IEEE International Parallel & Distributed Processing Symposium, Chicago, IL, 2016
DOI: 10.1109/IPDPSW.2016.197

H. Anzt, B. Haugen, J. Kurzak, P. Luszczek, J. Dongarra
Experiences in Autotuning Matrix Multiplication for Energy Minimization on GPUs
Concurrency and Computation: Practice and Experience, 27(17):5096-5113, 2015
DOI: 10.1002/cpe.3516

B. Haugen, J. Kurzak, J. Dongarra
Search Space Pruning Constraints Visualization
VISSOFT'14: Second IEEE Working Conference on Software Visualization, Victoria, BC, Canada, 2014
DOI: 10.1109/VISSOFT.2014.15



M. Gates, H. Anzt, J. Kurzak, J. Dongarra
Accelerating Collaborative Filtering Using Concepts from High Performance Computing
BigData'15: IEEE International Conference on Big Data, Santa Clara, CA, 2015
DOI: 10.1109/BigData.2015.7363811



J. Kurzak, H. Anzt, M. Gates, J. Dongarra
Implementation and Tuning of Batched Cholesky Factorization and Solve for NVIDIA GPUs
Transactions on Parallel and Distributed Systems, 27(7):2036-2048, 2015
DOI 10.1109/TPDS.2015.2481890