ARGONNE **LEADERSHIP COMPUTING** FACILITY

Exploring the Performance of Machine Learning on Kove XPDs

Melanie Cornelius[1], William Allcock[1], Brian Toonen[1], Zhiling Lan[2], Zack Cornelius[3]

[1]Leadership Computing Facility, Argonne National Laboratory (USA); [2]Illinois Institute of Technology (USA), [3]Kove (USA)

Abstract

Machine learning is a common large-scale application. As HPC moves toward the exascale, the requirements of machine learning are only increasing. Memory consumption and computational load increase superlinearly as dataset and algorithm complexity grow, but the predictive learning ability also tends to increase with these factors. Thus high-performance machine learning requires large, performant systems and extensive time.

Disaggregated memory is a technique similar to SAN in the 90s; here, memory is de-coupled from the CPU. The benefits are myriad and hotly under debate, but with disparity in resource demand caused by the memory wall problem, and given the memory requirements of large-scale machine learning, machine learning may be an ideal application for disaggregated memory systems.

To that end, this project works toward an understanding of the relationship between machine learning conditions and disaggregated memory setup to determine how machine learning might be made more performant at large-scales in a disaggregated memory datacenter.

Machine Learning

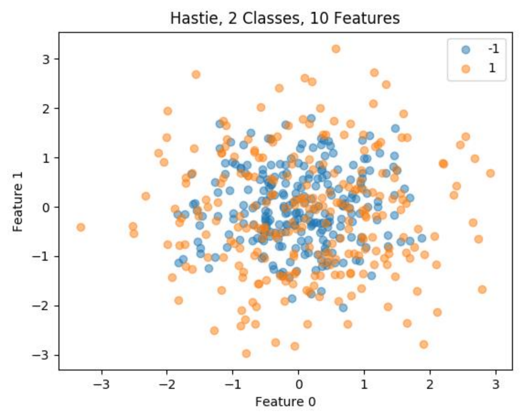
Simple machine learning algorithms were trained and tested on engineered datasets. Algorithms were selected to be mathematically predictable but computationally diverse. Datasets were engineered for different feature/class relationships, and all were widely configurable.

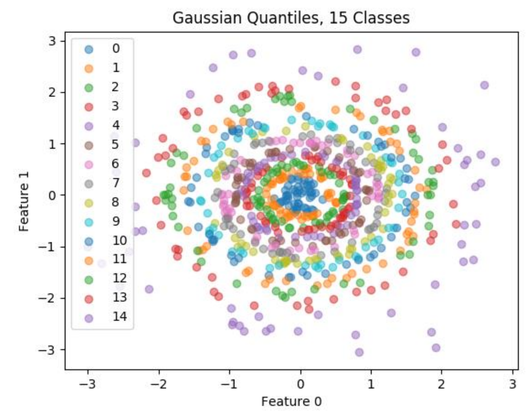
Algorithms:

- AdaBoost
- Decision Tree
- K-Nearest Neighbors (k = 3, 5)
- Naïve Bayes
- Multilayer Perceptron Neural Network
- Random Forest

Datasets:

- Hastie 10-2
  - 10 features
  - Binary classification
  - Classes arranged in two noisy, 10-dimensional hyperspheres
- Gaussian Quantiles
  - Binary features
  - 15 classes
  - Classes arranged in 15 noisy, two-dimensional hyperspheres (circles of increasing radius)
- Moons
  - Binary features
  - Binary classification
  - Classes shaped like two adjacent crescents arranged to intersect given noise





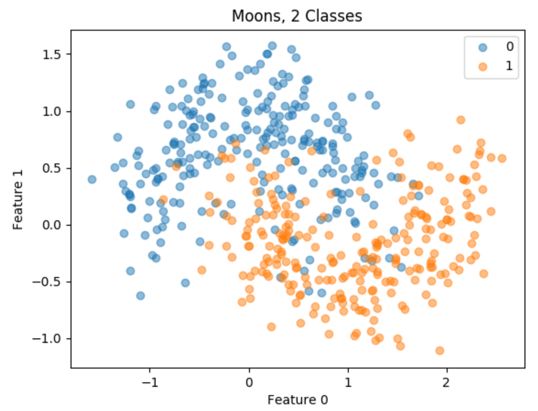


Figure 1 (top): Relationship between two (of ten) features in the Hastie dataset.
Figure 2 (middle): Relationship between 15 classes over two features in the Gaussian Quantiles dataset.
Figure 3 (bottom): Relationship between two classes over two features in the Moons dataset.

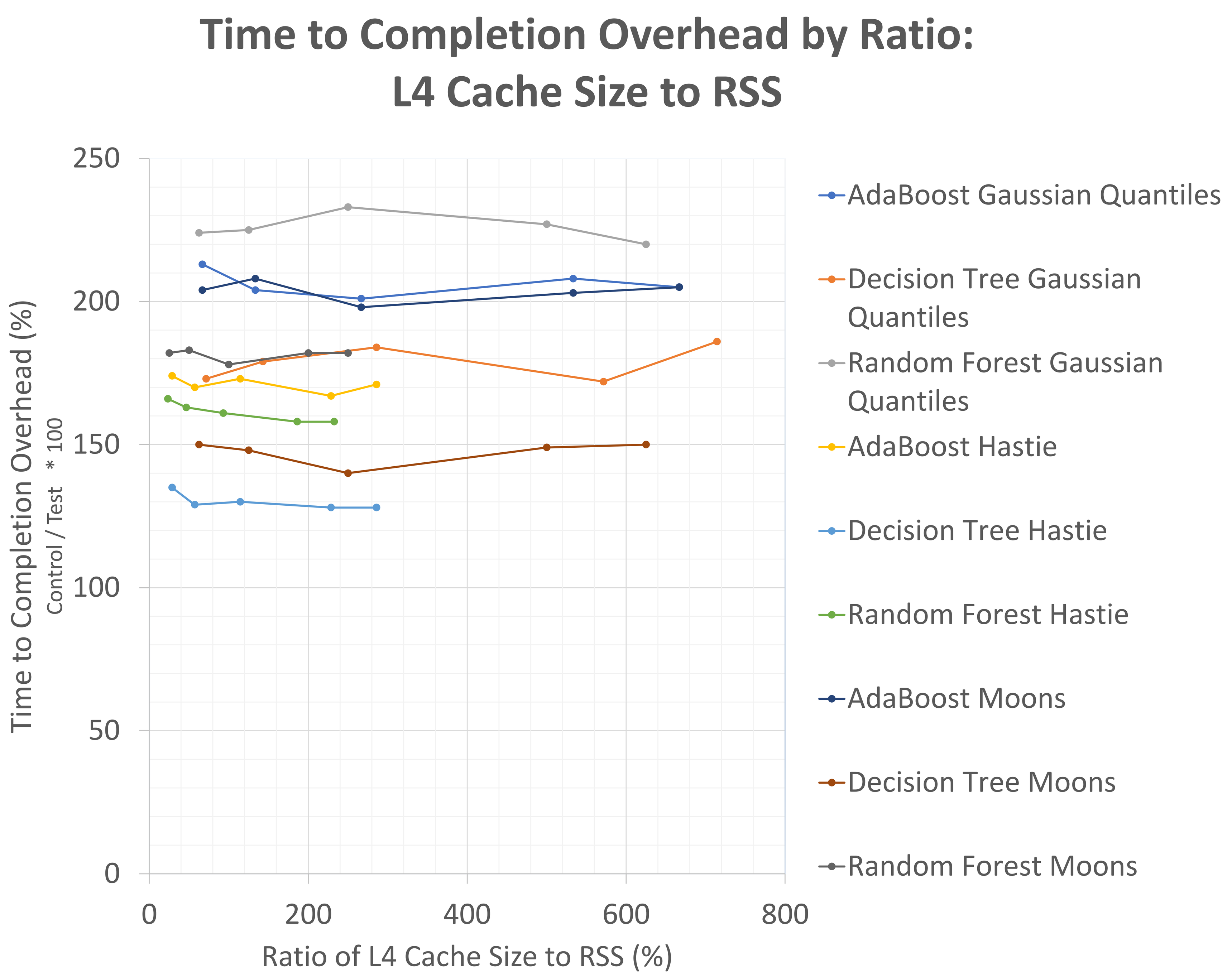


Figure 4: Overhead (time to completion) by percentage of RSS able to fit into the L4 cache. The nearly-horizontal lines demonstrate the overhead plateau – the lack of variation in the overhead per dataset-algorithm combination. Here, the computational intensity of each dataset-algorithm combination *increases* as the overhead plateau lowers; more computationally intensive algorithms have less overhead. This is encouraging; real-world machine learning is increasingly computationally intensive.

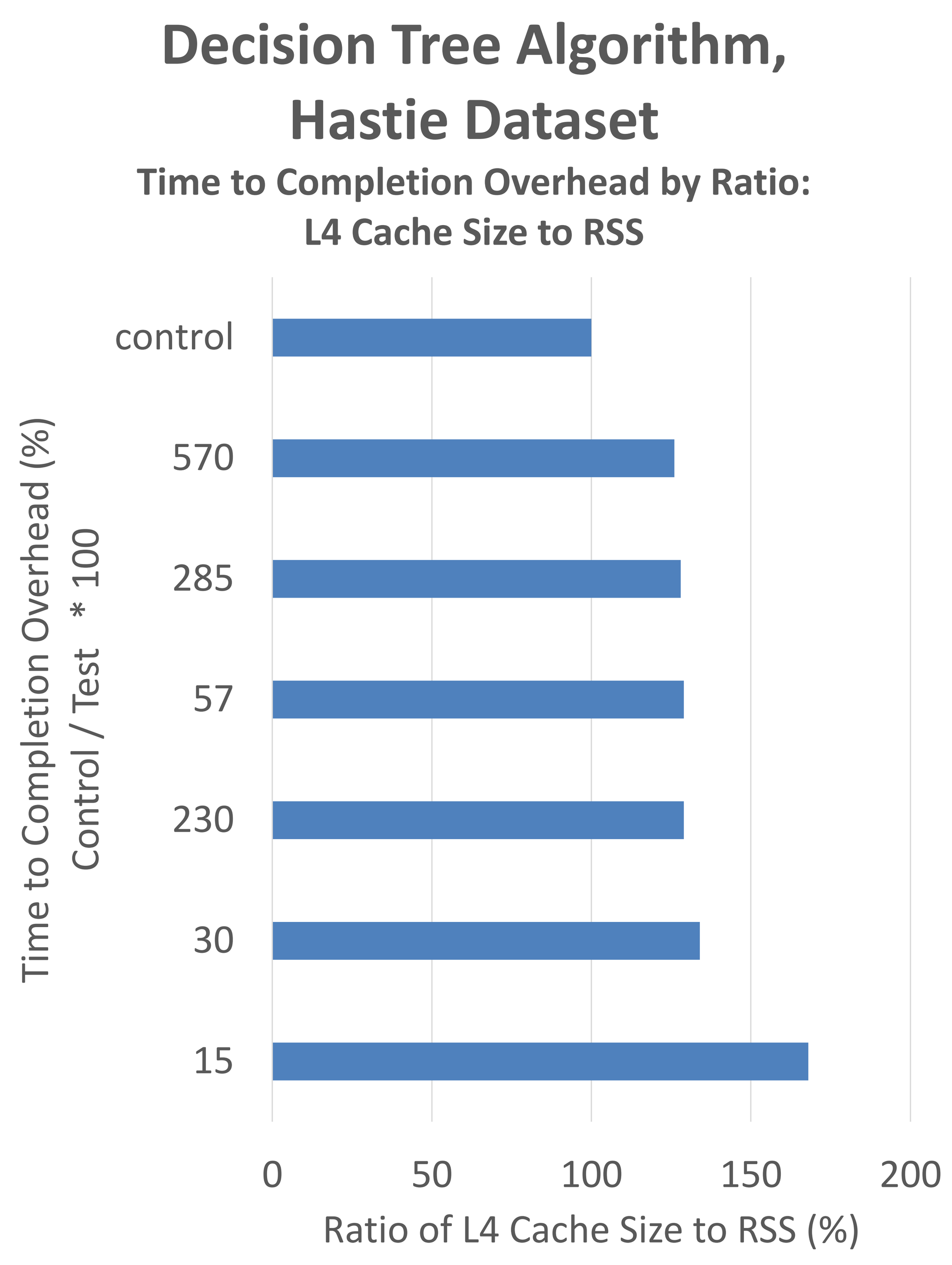
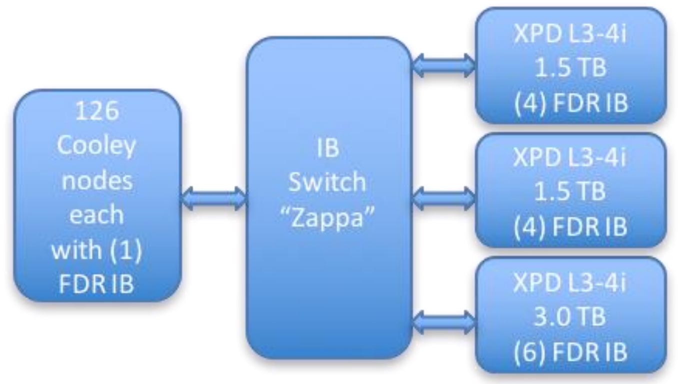


Figure 5: Overhead (time to completion) by percentage of RSS able to fit into the L4 cache. Here there is extended data for smaller L4 / RSS ratios showing the breakdown of the plateau seen when up to 30% of the RSS fits into the local L4 cache.

The RAM Area Network and Kove XPDs

Argonne National Laboratory (ANL) has a production setup for disaggregated memory. The project directing this system is called RAN – the Ram Area Network – and is attached to ANL's cluster Cooley. Cooley has 126 nodes, each equipped with dual Intel Haswell 6-core processors, 384 GiB local RAM, and one InfiniBand (IB) HCA connecting the cluster over an FDR IB network.

Disaggregated memory is served from three Kove XPDs connected to Cooley's IB network. Two XPDs have 4 IB links and 1.5 TiB DDR3 RAM, and the third has 6 links and 3 TiB of memory.



A software package provided by Kove, termed xmem, allows users to configure the behavior of volumes on the XPDs as well as to set aside a subset of local RAM on the node. This subset of local memory is used both by the process using xmem and by the volume as a cache for paging out to the remote memory. Effectively, such a setup extends the cache hierarchy to include an L4 cache composed of some amount of local DDR3 memory and replacing the lowest-level cache with remote memory on the XPD volume.

Experimental Configuration and Results

Tests were run setting the local RAM cache (the L4 cache) to different sizes. Results are presented as ratios of the resident set size (RSS) to the L4 cache size. This represents how much of the RSS was local. Results indicate the performance cost when using the disaggregated system as caused by the RSS / L4 cache size ratio.

For each dataset-algorithm combination, a constant overhead in time to completion results from using disaggregated memory in this setup. The overhead cost holds constant with L4 cache sizes sufficient to hold only 30% of the RSS.

Major Conclusions

1. Using disaggregated memory in this experiment's setup has an overhead cost in time to completion. Accuracy is unaffected.
2. The overhead cost plateaus for each dataset-algorithm combination regardless of if the RSS can fit many times in the local RAM cache or if only 30% of the RSS fits in the local cache.
3. The more computationally intensive the dataset-algorithm combination, the less paging the calculation experiences, and the lower the overhead plateau.

(place images describing conclusion #3 here)

Future Work

L4 Cache Size

To continue exploring the overhead plateau, additional tests should be run with L4 cache sizes below 30% the RSS of each dataset-algorithm combination. These tests are currently underway.

Batch Size

Modern machine learning frequently depends on batch training. Here, subsets of the training dataset are processed per step, and tuning this parameter often affects performance (time to completion and accuracy) wildly. Use of disaggregated memory and the availability of a large bank of remote memory might dramatically change how batch size selection is configured. Further testing here is needed.

Resident Set Size

Modern machine learning frequently uses enormous datasets and complex algorithms designed to fit within the limits of the system. Tests should be performed where the resident set size is made one, two, and (if possible) three orders of magnitude larger.

Neural Networks and Deep Learning

Neural networks and deep learning are major topics in modern HPC and data science. Simple neural networks were tested in these experiments with the goal of understanding the effect of altering batch size. More complex neural networks should be trained and tested to understand if the above conclusions hold in this related but differing application of machine learning.

XPD Configuration

Kove's xmem driver brings the user a wide variety of configuration options to control the behavior of paging to remote memory. In these tests, only the L4 cache size parameter was altered. It is possible altering other xmem settings could affect performance, likely reducing the overhead cost plateau. These tests are currently underway with significant input from Kove developers.

Argonne National is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

SC18 | November, 12, 2018, Dallas, TX