Intelligent HPC Cloud

Illustrated by Harp and Harp-DAAL at Indiana University
IEEE Cloud Computing Conference
June 26, 2017

Judy Qiu

Intelligent Systems Engineering Department, Indiana University
Email: xqiu@indiana.edu
Intelligent HPC Cloud

1. General Purpose Machine Intelligence requires both Cloud and HPC Technologies
2. HPC-Apache Big Data Stack supports AI and IoT
3. Illustration of Harp (Hadoop plug in) and Intel’s High Performance Data Analytics
50 Billion Devices by 2020

World Popular will be 7.6 billion by 2020

HPC Cloud
(e.g. CPU, GPU, FPGA)
**High Performance – Apache Big Data Stack**

**HPC-ABDS** as Cloud-HPC interoperable software with performance of HPC (High Performance Computing) and the rich functionality of the commodity Apache Big Data Stack was a bold idea developed.

Many of them have fixed communication patterns!
Harp is an open-source project developed at Indiana University [6], it has:

- MPI-like **collective communication** operations that are highly optimized for big data problems.
- Harp has efficient and innovative **computation models** for different machine learning problems.

DAAL is an open-source project that provides:

• Algorithms Kernels to Users
  • Batch Mode (Single Node)
  • **Distributed Mode (multi nodes)**
  • Streaming Mode (single node)

• Data Management & APIs to Developers
  • Data structure, e.g., Table, Map, etc.
  • HPC Kernels and Tools: MKL, TBB, etc.
  • Hardware Support: Compiler
Harp-DAAL enable faster Machine Learning Algorithms with Hadoop Clusters on Multi-core and Many-core architectures

- Bridge the gap between HPC hardware and Big data/Machine learning Software
- Support Iterative Computation, Collective Communication, Intel DAAL and native kernels
- Portable to new many-core architectures like Xeon Phi and run on Haswell and KNL clusters
Performance of Harp-DAAL on KNL Single Node

Harp-DAAL vs. Spark vs. NOMAD

Harp-DAAL-Kmeans vs. Spark-Kmeans:

~20x speedup
1) Harp-DAAL-Kmeans invokes MKL matrix operation kernels at low level
2) Matrix data stored in contiguous memory space, leading to regular access pattern and data locality

Harp-DAAL-SGD vs. NOMAD-SGD

1) Small dataset (MovieLens, Netflix): comparable perf
2) Large dataset (Yahooomusic, Enwiki): 1.1x to 2.5x, depending on data distribution of matrices

Harp-DAAL-ALS vs. Spark-ALS

20x to 50x speedup
1) Harp-DAAL-ALS invokes MKL at low level
2) Regular memory access, data locality in matrix operations
Hadoop/Harp-DAAL: Prototype and Production Code

Source codes became available on Github at Harp-DAAL project in February, 2017.

- Harp-DAAL follows the same standard of DAAL’s original codes
- Six Applications
  - Harp-DAAL Kmeans
  - Harp-DAAL MF-SGD
  - Harp-DAAL MF-ALS
  - Harp-DAAL SVD
  - Harp-DAAL PCA
  - Harp-DAAL Neural Networks
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Category</th>
<th>Applications</th>
<th>Features</th>
<th>Computation Model</th>
<th>Collective Communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>Clustering</td>
<td>Most scientific domain</td>
<td>Vectors</td>
<td>AllReduce</td>
<td>allreduce, regroup+allgather, broadcast+reduce, push+pull</td>
</tr>
<tr>
<td>Multi-class Logistic Regression</td>
<td>Classification</td>
<td>Most scientific domain</td>
<td>Vectors, words</td>
<td>Rotation</td>
<td>regroup, rotate, allgather</td>
</tr>
<tr>
<td>Random Forests</td>
<td>Classification</td>
<td>Most scientific domain</td>
<td>Vectors</td>
<td>AllReduce</td>
<td>allreduce</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>Classification, Regression</td>
<td>Most scientific domain</td>
<td>Vectors</td>
<td>AllReduce</td>
<td>allgather</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>Classification</td>
<td>Image processing, voice recognition</td>
<td>Vectors</td>
<td>AllReduce</td>
<td>allreduce</td>
</tr>
<tr>
<td>Latent Dirichlet Allocation</td>
<td>Structure learning (Latent topic model)</td>
<td>Text mining, Bioinformatics, Image Processing</td>
<td>Sparse vectors; Bag of words</td>
<td>Rotation</td>
<td>rotate, allreduce</td>
</tr>
<tr>
<td>Matrix Factorization</td>
<td>Structure learning (Matrix completion)</td>
<td>Recommender system</td>
<td>Irregular sparse Matrix; Dense model vectors</td>
<td>Rotation</td>
<td>rotate</td>
</tr>
<tr>
<td>Multi-Dimensional Scaling</td>
<td>Dimension reduction</td>
<td>Visualization and nonlinear identification of principal components</td>
<td>Vectors</td>
<td>AllReduce</td>
<td>allgather, allreduce</td>
</tr>
<tr>
<td>Subgraph Mining</td>
<td>Graph</td>
<td>Social network analysis, data mining, fraud detection, chemical informatics, bioinformatics</td>
<td>Graph, subgraph</td>
<td>Rotation</td>
<td>rotate</td>
</tr>
<tr>
<td>Force-Directed Graph Drawing</td>
<td>Graph</td>
<td>Social media community detection and visualization</td>
<td>Graph</td>
<td>AllReduce</td>
<td>allgather, allreduce</td>
</tr>
</tbody>
</table>