

TonY: An Orchestrator for Distributed Machine Learning Jobs

Anthony Hsu, Keqiu Hu, Jonathan Hung, Arun Suresh, Zhe Zhang

LinkedIn

{ahsu,khu,jhung,asuresh,zezhang}@linkedin.com

Abstract

Training machine learning (ML) models on large datasets requires considerable computing power. To speed up training, it is typical to distribute training across several machines, often with specialized hardware like GPUs or TPUs. Managing a distributed training job is complex and requires dealing with resource contention, distributed configurations, monitoring, and fault tolerance. In this paper, we describe TonY, an open-source orchestrator for distributed ML jobs built at LinkedIn to address these challenges.

1 Introduction

The past couple of decades have seen an explosion in "Big Data" systems for storing and processing data. Some widely used systems include MapReduce [10], Hadoop Distributed File System [19], and Spark [21]. The scale of these systems has made it possible to store petabytes of data and do large-scale ML.

Many features on the LinkedIn website are powered by ML, including People You May Know, Job Recommendations, the News Feed, and Learning Recommendations. Many of these features are powered by ML techniques such as boosted decision trees [7] and generalized linear models [22].

To boost the accuracy of predictions, ML engineers have started experimenting with non-linear models such as neural networks [11] to capture more complex relationships in the data. Programming these neural networks in a generic language is tedious and error-prone. To address this, many frameworks have been created to simplify the construction and training of neural networks. These frameworks include DistBelief [9] and its successor TensorFlow [4], Theano [6], Caffe [13], PyTorch [17], and Keras [8].

An ML engineer will often begin model development by developing on a single machine. One popular tool is a "notebook" program such as Jupyter [15] or Zeppelin [1] that allows an ML engineer to interactively explore the data and test out fragments of their models. This works when experimenting on a sample of the data. However, to validate a new model, they generally need to train and test their model on the full dataset, which may be petabytes in size and would take too long to train on a single machine. To scale up their training, they need to divide the data across multiple machines and train in parallel [5].

Most ML frameworks provide APIs for doing distributed training. However, to make use of multiple machines, an ML engineer still has to copy their program to each host, set the appropriate environment variables and configurations for distributed training on each host, and then launch their training program on each host. This ad-hoc process faces several challenges:

- **Resource contention.** ML engineers sharing the same pool of unmanaged machines fight for the same memory, CPU, and GPU resources. Consequently, jobs may fail with out-of-memory exceptions or errors allocating GPUs.
- **Tedious and error-prone configuration.** Setting up a distributed training job requires copying configurations to all hosts and it is hard to verify and update these configurations.
- **Lack of monitoring.** While the job is running, it is difficult to monitor its global progress.
- **Lack of fault tolerance.** Transient errors are hard to debug and require manual restarts.

To address these challenges, we built and open-sourced TonY [12], an orchestrator that interacts with a cluster scheduler to launch and manage distributed training jobs.

2 Architecture

TonY consists of a client for submitting jobs to a scheduler and an application that runs in the scheduler. Users use the client to submit their ML jobs, and the application handles allocating resources, setting up configurations, and launching the ML job in a distributed fashion. **The client interface is generic and its implementation can support submitting to multiple schedulers.** The scheduler implementation can be changed without requiring users to update their ML or client submission code.

For our initial implementation of TonY, we added support for running distributed TensorFlow jobs on Hadoop YARN (Yet Another Resource Negotiator) [20] (hence the name TonY), as these were the most commonly used ML framework and scheduler, respectively, at LinkedIn.

The overall architecture of TonY is presented in Figure 1. We present the client and cluster components of TonY in more detail in the following subsections.

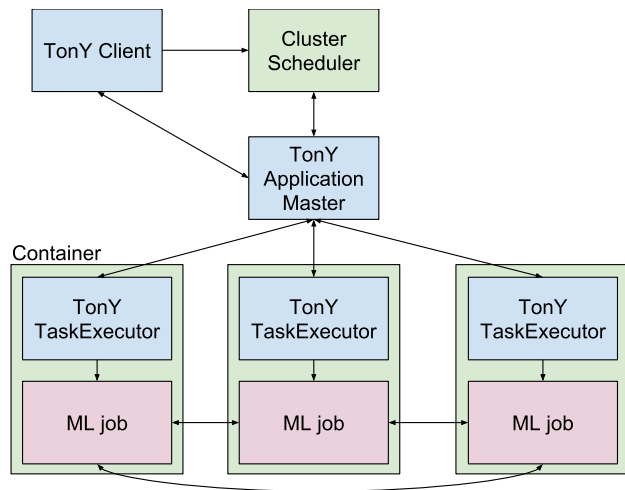


Figure 1: TonY’s architecture.

2.1 TonY Client

The TonY client is the library users use to launch their distributed ML jobs. Users describe in an XML file the resources required by their job. For TensorFlow, this might include the number of worker and parameter server instances as well as how much memory and how many GPUs per instance. If needed, users can also specify additional configurations for the underlying scheduler. In the case of YARN, this might include specifying the queue [3] or node label [14] (e.g.: high-memory) to run on.

Users will also provide the path to their ML program as well as the virtual environment or Docker image [16] in which their program should run on the cluster. Additionally, users can specify properties such as model-specific hyperparameters, input data, and output location via command-line arguments passed to the TonY client.

Often, distributed ML jobs will be run as part of a larger workflow that includes data preprocessing and model deployment. To simplify integration into existing workflows, we built a TonY plugin for one such workflow manager, Azkaban [2], that lets users add distributed ML jobs in the same workflow alongside Spark, MapReduce, and other jobs.

2.2 TonY Cluster Application

When the user runs the TonY Client to submit their job, the client will package the user configurations, ML program, and virtual environment into an archive file that it submits to the cluster scheduler.

The TonY Client will launch a master program in the cluster scheduler. In our initial implementation supporting Hadoop’s YARN scheduler, we launch a TonY ApplicationMaster (AM) in a YARN container. The AM then negotiates with YARN’s ResourceManager (RM) to request all the other containers (e.g.: worker and parameter server tasks) needed by the ML

job. The AM handles heterogeneous resource requests for different task types, such as requesting containers with GPUs for the worker tasks but requesting CPU-only containers for the parameter server tasks.

Once the task containers are allocated by the RM to the TonY AM, it then launches a TaskExecutor in each task container. This TaskExecutor will allocate a port for its task to run on and register this port with the AM. Upon receiving registration from all TaskExecutors, the AM will construct a global cluster spec that it will then send back to every TaskExecutor. Each TaskExecutor will then set the global cluster spec along with task-specific configuration in environment variables before spawning the ML job as a child process. Once all the ML jobs start up, they will communicate and coordinate with one another via the ML framework’s distributed protocol (whether that be RPC, MPI, etc.), and the TaskExecutors will monitor the task processes and heartbeat back to the AM. When the task processes finish, the TaskExecutor will register the final exit status with the AM before terminating.

The TaskExecutor for the first worker task will also allocate a port for launching a visualization user interface such as TensorBoard for monitoring the running job. This also gets registered with TonY AM. This user interface URL, along with links to all the other task logs, is sent back to the TonY Client so that users can directly access the visualization UI and task logs from one place.

Finally, if any task fails, the TonY AM will automatically tear down the remaining tasks, request new task containers, setup a new global cluster spec, and relaunch the tasks. The ML tasks can then restore from the last checkpoint and continue training.

3 Discussion

Previously, ML engineers had to write ad-hoc scripts to launch distributed ML jobs on a pool of machines, with no resource guarantees or isolation from other users. Now, using TonY, users can configure their job once and rely on TonY to negotiate with a cluster scheduler for guaranteed resources.

The TonY master handles all the distributed setup and provides a central place to monitor and visualize the training job. It also ensures fault tolerance by restarting distributed jobs in case of transient task failures.

The master and TaskExecutor orchestration framework is also an ideal place to instrument the ML tasks and collect metrics about the tasks’ performance and resource utilization. These statistics could be aggregated and analyzed in a UI such as Dr. Elephant [18] to suggest new settings for the ML jobs that would improve performance and resource utilization. We are currently implementing these new features in TonY and plan to discuss them more in future work.

References

- [1] Apache zeppelin. <https://zeppelin.apache.org/>. Accessed: 2019-02-04.
- [2] Azkaban: Open-source workflow manager. <https://azkaban.github.io/>. Accessed: 2019-02-04.
- [3] Orgqueue for easy capacityscheduler queue configuration management. <https://issues.apache.org/jira/browse/YARN-5734>, 2016.
- [4] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *OSDI*, 2016.
- [5] Tal Ben-Nun and Torsten Hoefler. Demystifying parallel and distributed deep learning: An in-depth concurrency analysis. *arXiv preprint arXiv:1802.09941*, 2018.
- [6] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: A cpu and gpu math compiler in python. In *SciPy*, 2010.
- [7] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *KDD*, 2016.
- [8] François Chollet et al. Keras. <https://github.com/fchollet/keras>, 2015.
- [9] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, et al. Large scale distributed deep networks. In *NIPS*, 2012.
- [10] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. In *OSDI*, 2004.
- [11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [12] Jonathan Hung. Open sourcing tony: Native support of tensorflow on hadoop. <https://engineering.linkedin.com/blog/2018/09/open-sourcing-tony--native-support-of-tensorflow-on-hadoop>, 2018.
- [13] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, 2014.
- [14] Konstantinos Karanasos, Arun Suresh, and Chris Douglas. Advancements in yarn resource manager. *Encyclopedia of Big Data Technologies*, 2018.
- [15] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. Jupyter notebooks – a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press, 2016.
- [16] Dirk Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014(239):2, 2014.
- [17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.
- [18] Akshay Rai. Open sourcing dr. elephant: Self-serve performance tuning for hadoop and spark. <https://engineering.linkedin.com/blog/2016/04/dr-elephant-open-source-self-serve-performance-tuning-hadoop-spark>, 2016.
- [19] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *MSST*, 2010.
- [20] Vinod Kumar Vavilapalli, Arun C Murthy, Chris Douglas, Sharad Agarwal, Mahadev Konar, Robert Evans, Thomas Graves, Jason Lowe, Hitesh Shah, Siddharth Seth, et al. Apache hadoop yarn: Yet another resource negotiator. In *SoCC*, 2013.
- [21] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *NSDI*, 2012.
- [22] XianXing Zhang, Yitong Zhou, Yiming Ma, Bee-Chung Chen, Liang Zhang, and Deepak Agarwal. Glmix: Generalized linear mixed models for large-scale response prediction. In *KDD*, 2016.