Building Scalable Deep Learning Systems with RPC

Jeff Hajewski
Department of Computer Science
University of Iowa
Iowa City, IA, USA
jeffrey-hajewski@uiowa.edu

Suely Oliveira
Department of Computer Science
University of Iowa
Iowa City, IA, USA
suely-oliveira@uiowa.edu

***Abstract*—word word**

***Index Terms*—distributed deep learning, neural architecture search, artificial intelligence**

I. Introduction

Building robust and scalable distributed applications is challenging — getting communications patterns correct, handling node failures, and allowing for elastic compute resources all contribute to a high level of complexity. These challenges are exacerbated for long-runnining applications such as training large deep learning models or neural architecture search. Among the many popular frameworks for distributed programming, MPI [**?**] combined with additional accelerator paradigms such as OpenMP [**?**], OpenACC [**?**], and CUDA [**?**] is the common choice for performance-critical numerical workloads. In the deep learning setting, MPI is less popular with many favoring mutli-threading in combination with multiple GPUs, and more recently experimenting with Kubernetes [**?**], [**?**]. Although part of this is simplicity of communication patterns, the fault tolerance required for long-running model training (which can last on the order of months in the industrial setting) makes MPI a poor choice for the underlying communication infrastructure.

In this work we propose an RPC-based system for distributed deep learning. We experiment with the proposed architecture in the domain of neural architecture search (NAS), a computationally intensive problem that trains thousands of deep neural networks in search of an optimal network architecture. We use this problem domain to illustrate four primary advantages to building a computationally intensive distributed system using RPC:

- RPC's higher level of abstraction over MPI simplifies the process of building more complex systems and communication patterns.
- The user can avoid lower-level message serialization and deserialization through RPC coupled with framework such as Protocol Buffers or Thrift.
- RPC systems don't require underlying software or resource managers — a user can create an AWS instance and immediately run their RPC-based code.
- RPC-based systems can offer elastic compute abilities, adding or removing nodes as needed without needing to stop or restart the system.

Or system consists of four separate pieces: a *model* that directs the search for a network architecture, a number of *workers* that perform the computational work of training the models, a number of *brokers* that form the backbone of the data pipeline from model to workers, and a *nameserver* that simplifies the process of adding new brokers, workers, or models to the system. It also offers elastic compute resources, allowing an arbitrary number of workers to join during high computational loads as well as allowing workers to leave the system, decreasing the overall available compute, without needing to restart or manual intervention. The system is fault-tolerant to the loss of workers or brokers, and is highly scalable due to the ability of the brokers to share work and compute resources. Perhaps most importantly from a usability perspective, our system is language agnostic. In our experiments, we use Python for our model and workers, which we use to build and train or deep neural networks via PyTorch [**?**], and use Go to build the data pipeline of brokers. We use gRPC [**?**] to handle the generation of RPC stubs, but could have just as easily used Apache Thrift [**?**], which generates stubs in a larger range of languages such as Ocaml, Haskell, and Rust.

Although we are proposing RPC as an alternative to MPI, it is important to note that we are not claiming RPC is superior to MPI in every problem domain. In the domain of distributed deep learning, RPC offers many useful features such as simple fault tolerance and allowing compute resources to join and leave the system at will. In other domains, such as distributed linear algebra computations, MPI is probably a better choice due to its built in ability to efficiently broadcast messages to nodes within a system in a manner that takes advantage of the network architecture.

II. Neural Architecture Search

The goal of neural architecture search (NAS) is to find an optimal neural network architecture for a given problem. NAS is computationally intensive due to the requirement of having to train a candidate network in order to evaluate its

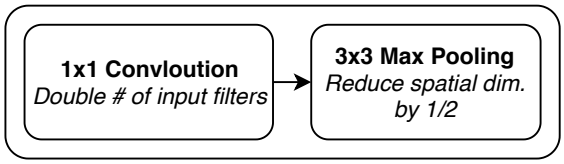

Fig. 1. Reduction layer.

effectiveness. Although recent novel approaches have dramatically reduced this cost [**?**], [**?**], these techniques fix certain elements of the design process, somewhat limiting the available architectures. Despite the computationally intense nature of the NAS problem, the task itself is trivially parallelizable across the network evaluations — two separate networks can be trained simultaneously before being evaluated against eachother.

The two common approaches to NAS are reinforcement learning based approaches such as [**?**], [**?**], [**?**], and evolutionary approaches such as [**?**], [**?**], [**?**]. In our experiments we focus on the evolutionary approach due to its simplicity both in understanding and implementation.

A. Evolutionary Algorithms for NAS

We use a relatively simple approach to evolving neural network architectures in that we only construct linear networks, rather than allow an arbitrary number of incoming connections for any given layer. The motivation behind this is two-fold: it reduces the size of search space for the neural network architecture and simplifies the implementation. Because this the focus of this work is the system architecture, rather than neural architecture search, we feel the trade-off is reasonable.

The problem domain we focus on is computer vision, so we restrict ourselves to convolutional layers for the hidden layers. The networks are constructed by duplicating modules one after the other a pre-determined (by the user) number of times, where layers composing the module unit are determined via evolutionary search. This approach allows us to control the decrease in spatial dimensions, which is useful in the implementation of the network (all layers within a module have the same spatial dimension) and also allows us to force the heuristic of doubling the number of filters via a 1x1 convultional layer prior to reducing the spatial dimension by a factor of two via a 3x3 max pooling layer. The intuition behind this technique is to try and preserve as much information contained within a tensor as possible.

Module evolution proceeds by appending layers to the module or performing crossover, where two network are split at random positions and recombined to form two new networks. The maximum number of layers in a module is fixed (by the user), and evolution can add any of the following 2D convolutions:

- 1x1 Convolution
- 3x3 Convolution
- 3x1-1x3 Convolution
- 5x5 Convolution
- 5x1-1x5 Convolution

A ReLU activation is used between layers

Algorithm 1 High-level outline of evolutionary algorithm.

```
procedure ProduceOffsprint(N_1, N_2)
    if |N_1| ≠ |N_2| then
        o_1 ← N_1.mutate()
        o_2 ← N_2.mutate()
        return ReturnFittest(o_1, o_2)
    end if
    o ← N_1.crossover(N_2)
    return o.mutate()
end procedure
procedure Mutate
    // Possibly append layer
    if sampleUniform(0, 1) < append_p then
        self.layers.append(randomLayer())
    end if
end procedure
```

III. RPC-based Communication

Remote Procedure Call (RPC) offers a method of invoking a function on a remote computer with a given set of arguments. Most RPC frameworks involve a DSL used to define the RPC service, that is, the API available to the caller, and some type of data serialization format. For example, gRPC uses Protocol Buffers (ProtoBufs) [**?**] as the serialization format for data sent across the network. RPC offers a number of advantages for network communication. It is robust to node failures or network partitions (the RPC invokation simply fails). The data sent across the network is compactly represented, giving way to high bandwidth and low latency communication. The point-to-point communication allows for diverse communication patterns and paradigms. RPC forms the network communication infrastructure at Google [**?**], Facebook [**?**], as well as Hadoop [**?**], [**?**].

RPC frameworks typically use an IDL to represent the RPC service. Figure 3 gives an example of a gRPC service definition for a heartbeat service. The ProtoBuf compiler uses this definition to generate server stubs and service clients in a number of languages (C++, Java, Python, Go, etc.). The receiver of the RPC call must complete the server stubs by implementing the defined interface. For example, in the heartbeat service defined in Figure 3, the receiver of the heartbeat message would implement a `SendHeartbeat` function whose body would handle the logic of *receiving* a heartbeat from another process, such as updating a timestamp for the given process ID. The caller of `SendHeartbeat` uses the generated Heartbeat service client and is only responsible for constructing the `HeartbeatMsg`.

While the robustness to node failures and finer-grained point-to-point communication capabilities of RPC are core to building resilient distributed systems, the more powerful feature we capitalize on is the ability to build a system that is agnostic to the type of data flowing through its pipes. Figure 4

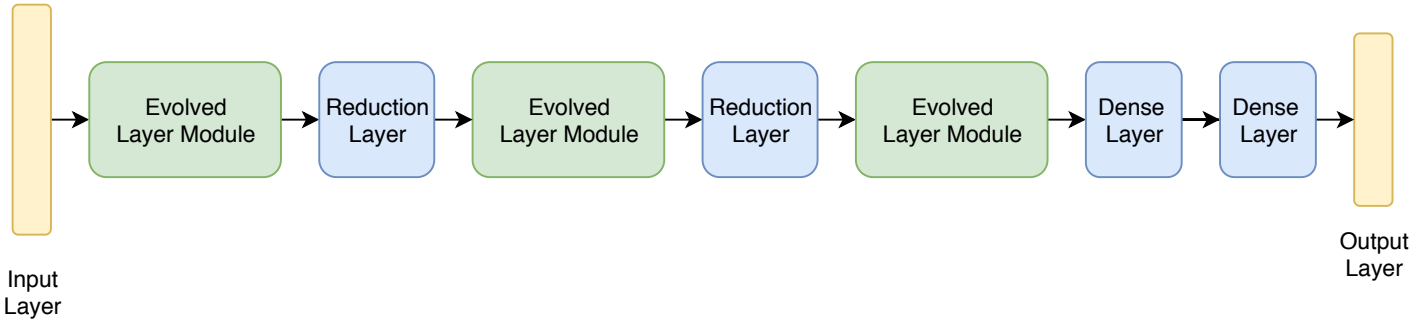

Fig. 2. Example of a constructed network.

```
service Heartbeat {
  rpc SendHeartbeat(Heartbeat) returns (HBResp) {}
}
message Heartbeat {
  string id = 1;
}
message HBResp {}
```

Fig. 3. Example gRPC service definition of a heartbeat service.

```
message Task {
  string id = 1;
  enum type = 2; // or string type
  bytes task_obj = 3;
}
```

Fig. 4. Example of a data agnostic message type.

gives an example of constructing a Protocol Buffer[CITE] message type that can be used to transport arbitrary data types through the system. Protocol Buffers (and similarly, Thrift messages) support arbitrary length byte arrays [1] This means the user can send a serialized object stored in the `Task`'s `task_obj` field without having to modify the system. A user can switch between running models and tasks using Java to running models and tasks using Python without modifying the system. In fact, the system can transport and run these tasks (in both Java and Python) simultaneously. By encapsulating the tasks (and results) as serialized objects stored as byte arrays the entire system can be data type agnostic.

There is a slight catch. If information within the serialized object is needed to properly schedule or transport the task/result to its destination, this information will need to be added to the message definition. This requires recompiling the message types and regenerating the gRPC (or Thrift) stubs. This is a minor inconveninece, as incorporating this new information into the system requires modifying the system infrastructure.

Another advantage of using gRPC or Thrift to define the RPC API of the system is the service definition itself acts as documentation on the flow of information within the system.

[1]In practice these arrays are limited in size. The failure rate of the RPC system typically increases with the size of the messages.

```
service Task {
  // Called by a worker to request a task.
  rpc RequestTask(TaskRequest)
    returns (TaskResponse) {}
}
service Result {
  // Called by a worker to send a result
  // back to the broker.
  rpc SendResult(ResultMsg)
    returns (ResultResponse) {}
}
```

Fig. 5. Example service definitions for tasks and results.

A user can look at these definitions and see how information is meant to flow.

A. *RPC vs MPI*

From a performance perspective, some prior work [**?**] has found RPC to provide lower latency and higher bandwidth for some tasks. That does not tell the entire story. MPI comes with built-in complex communication primitives that are capable of taking advantage the physical network architecture of the cluster. Additionally, MPI is capable of bulk broadcasts to groups of nodes using custom communicator definitions. These primitives save the programmer a lot of overhead when building HPC applications. Settings such as large, distributed matrix multiplications or iterative optimization algorithms are particuarly well suited for MPI's communication primitives.

However, applications that benefit from specialized communication patterns or that are asynchronous in nature are typically better suited for RPC-based communication. Consider, for example, an application that sends some amount of work to a number of worker nodes and waits for the work to be completed. In the synchronous setting, the entire system will only be as fast as its slowest node and will cause idle resources as the faster nodes wait for the slowest node to finish. In the asynchronous setting, the faster nodes can continue to process work. This scenario was a motivation for [**?**].

Another short-comig of RPC is message size limitations. This can be particularly troublesome in distributed machine learning settings where highly parameterized models such as deep neural networks are sent across the network. The solution

is to send a representation of the model, rahter than the model itself, across the network. The advantage of this approach is reduced network bandwidth utilization. The disadvantage of course is the loss of trained models at the worker nodes. In practice, this may not be a major issue as final model architectures are typically trained in a specialized manner (e.g., for a large number of epochs). Additionally, models can be saved to a distributed filesystem such as HDFS.

Of course, MPI does offer asynchronous communication, but it is not as flexible as that of RPC-based systems. In an asynchronous setting there are usually classes of nodes whose behavior are class-dependent. In the system we propose, there are four classes of processes: the model, the broker, the nameserver, and the worker. It is conceptually simpler to think of these as four different processes/programs and build each of them separately, rather than a single program that determines its behavoir based on its assigned communicator rank (as typically done in MPI).

IV. System Architecture

As previously mentioned, our system consists of four different components. Figure 6 shows an overview of the system architecture. A model is a problem specific implementation that controls what is sent to the system for evaluation and handles the result it receives. The brokers form the data pipeline of the system, moving work to available workers. Workers form the other customizable part of the system because they need to know how to perform their assigned work. Lastly, the nameserver maps known brokers to their network address – this is useful for connecting to brokers, such as a model connecting to a broker, a broker connecting to a broker (for broker-broker peering), or a worker connecting to a broker. The following sections will detail each components functionality individual and within the system as a whole.

A. Model

The model is the problem-specific, user-defined logic that determines what work should be performed next and how the results of previously assigned work should be processed. The only requirement of the model is that is uses a broker client stub (generated by gRPC) to push work to the system and implements the model service interface to allow the broker to push results back to the model.

The model needs to track outstanding tasks that have been sent to the broker. While the system is fault-tolerant for most brokers and all workers, if the broker the model is sending work to fails, the work the model is waiting to receive will be lost and the model will need to resend to a new broker.

B. Worker

Workers are the other user-defined and implemented portion of the system. While a single worker implementation can work for multiple model implementations, there is no general worker implementation that will work across all languages and models.

Using an API similar to that shown in Figure 8, one can use the same worker implementation for any task that inherits from the `BaseTask` class.

Figure 7 shows the communication pattern between the broker and a worker. There are two interesting aspects of this diagram: all communication starts at the worker and there is no heartbeat exchange between the worker and the broker. Both of these details are what allow the system to treat workers as ephemeral compute resources. This means an arbitrary number of workers can join and leave the system without the system knowing or caring. The downside to this approach is it requires a little extra bookkeeping at the broker. The broker needs to track what tasks are outstanding and decide if they should be moved back in to the work queue.

C. Broker

Brokers form the data pipeline of our system. Work is sent from a model to a broker, which in turn sends the work to a free worker and returns the result to the original model. At its core, a broker is essentially just a process with a owned task queue, a helper task queue (tasks received from other brokers), a processing queue, and a results queue. Work in the owned task queue is work that was sent from a model directly to the broker – this is the work that will be lost if the broker crashes. Work in the helper task queue is work that has been sent from other brokers that the respective broker has agreed to help with. If the broker crashes, this work will *not* be lost as the other brokers will see the failure and can recover the task from their processing queue. The processing queue stores tasks that have been sent to workers or other brokers. When a result is received from a worker or another broker, the corresponding ID will be removed from the processing queue and the result will be added to the result queue. The broker pulls tasks from the result queue and sends the result to its owner, which may be a model or another broker.

Brokers can establish links with other brokers.

V. Experiments

VI. Related Work

The idea of a brokered message queue is not new. RabbitMQ [CITE] is a general purpose message broker that supports the same functionality demonstrated in this paper, but messages are delivered based on a routing key, rather than the first available worker. Similar to the RPC message definitions used in this work, RabbitMQ uses a collection of bytes as the message body. For streaming data, Apache Kafka [**?**] is a good choice. Kafka is requires a ZooKeeper [**?**] instance and is generally more complex to setup and run. Both RabbitMQ and Kafka are robust to failures. At the other end of the spectrum, ∅MQ is a low-level messaging library that can be used to build a performant, brokered messaging system similar to the one described in this paper.

A natural question at this point in the work is why we would bother building our own brokered system when there are industrial strength alternatives available. These message

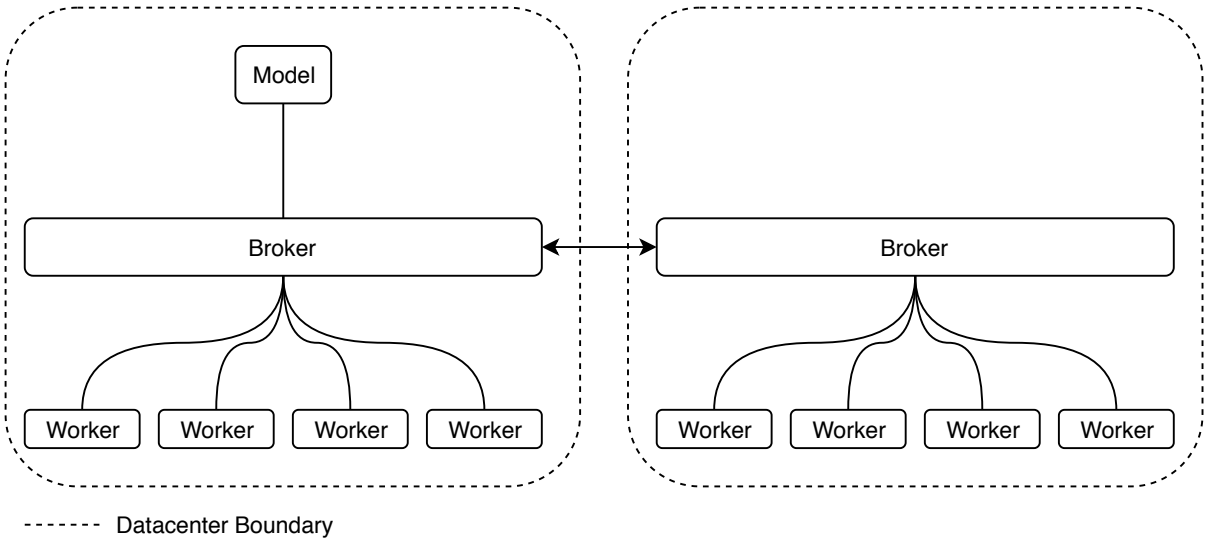

Fig. 6. Diagram of system architecture.

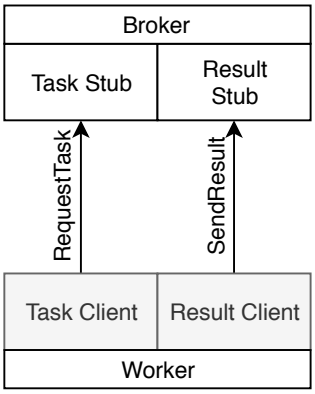

Fig. 7. Communication pattern between a worker and broker.

```
class BaseTask:
  def run(self):
    raise NotImplementedError()

class Worker:
  def process_task(self, task):
    result = task.run()
    self.broker_client.send_result(result)
```

Fig. 8. Example Task API in Python.

passing systems[2] stricly deal with sending and delivering messages in a scalable and robust manner. Our broker does more than simply relay messages—the broker is determining retry logic, hadling lost tasks, [ADD OTHER STUFF].

Tensorflow [**?**] offers some distributed training facilities. Other work such as Horovod [**?**] has improved upon Tensorflow's built-in distributed training facilities. Work has also been done on training deep learning models using MPI [**?**],

[2]We use this as a generic term, since Kafka is not really a message queue.

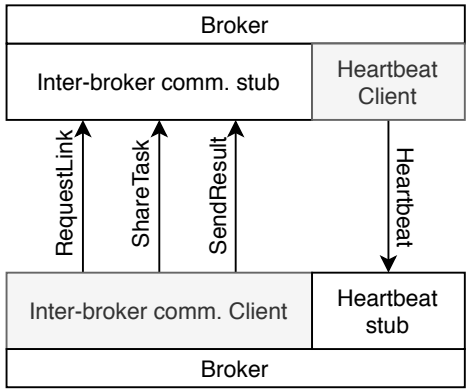

Fig. 9. Communication pattern for broker-broker communication.

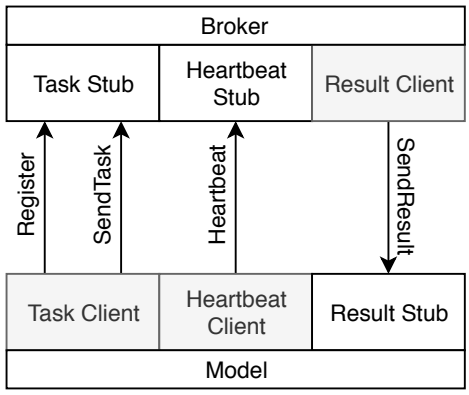

Fig. 10. Communication pattern between the broker and model.

[**?**]. However, all of these works suffer from fault tolerance of workers and elastic compute environments. Additionally, they require that MPI is installed on the cluster, which is not an issue if the software is running on a University research cluster or at a national lab, but requires the user to setup the underlying MPI installation if run on a provisioned AWS, GCP, or Azure cluster.

MXNet [**?**] introduces a declarative language used to build a computation graph, which is then optimized and evaluated by the MXNet engine, similar to Tensorflow pre-2.0. [NEEDS ADDITIONAL WORK]

Much of the previous work on neural architecture search uses some form of a distributed architecture consisting of a model, a coordinator, and workers. The coordinator handles assignment of work to the workers. While many of these works don't detail the specifics of their system, we reviewed some of the available code on Github. A popular paradigm is using Python's `multiprocessing` module to run multiple models on a multi-GPU machine. These GPUs are fed from a thread-safe queue. We expect some of the larger works from Google and Facebook likely use RPC for communication, as this is the standard approach for the rest of their infrastructure and is what their engineers are familiar with.

VII. Conclusion

It can be tempting to think of system design as an all-or-nothing decision—either build a system with MPI or build a system using RPCs. An interesting avenue for future work is combining the two. Consider a workload whose core unit of work is amenable to an MPI-based system but the individual units of work are independent of eachother with the exception of possible boundary interactions. A combination of RPC and MPI communication might be ideal—using MPI within a single cluster to complete individual units of work and RPC to communicate between clusters. In this way, individual clusters become the workers of the system and can join and leave at will, while the broker backbone manages sending tasks to the individual clusters and returning their results to the model.