Ray: A Distributed Framework for Emerging AI Applications

Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I. Jordan, Ion Stoica
University of California, Berkeley

Abstract

The next generation of AI applications will continuously interact with the environment and learn from these interactions. These applications impose new and demanding systems requirements, both in terms of performance and flexibility. In this paper, we consider these requirements and present Ray—a distributed system to address them. Ray implements a unified interface that can express both task-parallel and actor-based computations, supported by a single dynamic execution engine. To meet the performance requirements, Ray employs a distributed scheduler and a distributed and fault-tolerant store to manage the system’s control state. In our experiments, we demonstrate scaling beyond 1.8 million tasks per second and better performance than existing specialized systems for several challenging reinforcement learning applications.

1 Introduction

Over the past two decades, many organizations have been collecting—and aiming to exploit—ever-growing quantities of data. This has led to the development of a plethora of frameworks for distributed data analysis, including batch [20, 54, 28], streaming [15, 39, 31], and graph [44, 45, 24] processing systems. The success of these frameworks has made it possible for organizations to analyze large data sets as a core part of their business or scientific strategy, and has ushered in the age of “Big Data.”

More recently, the scope of data-focused applications has expanded to encompass more complex artificial intelligence (AI) or machine learning (ML) techniques [50]. The paradigm case is that of supervised learning, where data points are accompanied by labels, and where the workhorse technology for mapping data points to labels is provided by deep neural networks. The complexity of these deep networks has led to another flurry of frameworks that focus on the training of deep neural networks and their use in prediction. These frameworks often leverage specialized hardware (e.g., GPUs and TPUs), with the goal of reducing training time in a batch setting. Examples include TensorFlow [7], MXNet [18], and PyTorch [46].

The promise of AI is, however, far broader than classical supervised learning. Emerging AI applications must increasingly operate in dynamic environments, react to changes in the environment, and take sequences of actions to accomplish long-term goals [8, 33]. They must aim not only to exploit the data gathered, but also to explore the space of possible actions. These broader requirements are naturally framed within the paradigm of reinforcement learning (RL). RL deals with learning to operate continuously within an uncertain environment based on delayed and limited feedback [56]. RL-based systems have already yielded remarkable results, such as Google’s AlphaGo beating a human world champion [54], and are beginning to find their way into dialogue systems, UAVs [42], and robotic manipulation [25, 60].

The central goal of an RL application is to learn a policy—a mapping from the state of the environment to a choice of action—that yields effective performance over time, e.g., winning a game or piloting a drone. Finding effective policies in large-scale applications requires three main capabilities. First, RL methods often rely on simulation to evaluate policies. Simulations make it possible to explore many different choices of action sequences and to learn about the long-term consequences of those choices. Second, like their supervised learning counterparts, RL algorithms need to perform distributed training to improve the policy based on data generated through simulations or interactions with the physical environment. Third, policies are intended to provide solutions to control problems, and thus it is necessary to serve the policy in interactive closed-loop and open-loop control scenarios.

These characteristics drive new systems requirements: a system for RL must support fine-grained computations (e.g., rendering actions in milliseconds when interacting with the real world, and performing vast numbers of sim-
To satisfy these requirements, Ray implements a unified framework that can express both stateless and stateful computations. This framework consists of two components that are typically centralized in existing systems: (1) the task scheduler and (2) a directory for data objects. This allows Ray to schedule millions of tasks per second with millisecond-level latencies. Furthermore, Ray provides lineage-based fault tolerance for tasks and actors, and replication-based fault tolerance for the metadata store.

While Ray supports training, serving, and simulation in the context of RL applications, this does not mean that it should be viewed as a replacement for systems that provide solutions for these workloads in other contexts. In particular, Ray does not aim to substitute for generic data-parallel frameworks, such as Spark, as it currently lacks the rich functionality and APIs (e.g., straggler mitigation, query optimization) that these frameworks provide.

We make the following contributions:

- We design and build the first distributed framework that unifies training, simulation, and serving—necessary components of emerging RL applications.
- To support these workloads, we unify the actor and task-parallel abstractions on top of a dynamic task execution engine.
- To achieve scalability and fault tolerance, we propose a system design principle in which control state is stored in a sharded metadata store and all other system components are stateless.
- To achieve scalability, we propose a bottom-up distributed scheduling strategy.

2 Motivation and Requirements

We begin by considering the basic components of an RL system and fleshing out the key requirements for Ray. As shown in Figure 1, in an RL setting, an agent interacts repeatedly with the environment. The goal of the agent is to learn a policy that maximizes a reward. A policy is
To evaluate the policy, the agent interacts with the environment. A serving system aims to minimize latency, and maximize the number of decisions per second. To scale, load is typically balanced across multiple nodes serving the policy.

Figure 2: Typical RL pseudocode for learning a policy.

---

Flexible computation model. RL applications require both stateless and stateful computations. Stateless computations can be executed on any node in the system, which makes it easy to achieve load balancing and movement of computation to data, if needed. Thus stateless computations are a good fit for fine-grained simulation and data processing, such as extracting features from images or videos. In contrast, stateful computations are a good fit for implementing parameter servers, performing repeated computation on GPU-backed data, or running third-party simulators that do not expose their state.

Dynamic execution. Several components of RL applications require dynamic execution, as the order in which computations finish is not always known in advance (e.g., the order in which simulations finish), and the results of a computation can determine future computations (e.g., the results of a simulation will determine whether we need to perform more simulations).

We make two final comments. First, to achieve high utilization in large clusters, such a framework must handle millions of tasks per second. Second, such a framework is not intended for implementing deep neural networks or complex simulators from scratch. Instead, it should enable seamless integration with existing simulators and deep learning frameworks.

---

*Assume 5ms single-core tasks and a cluster of 200 32-core nodes. This cluster can run (1s/5ms) \times 32 \times 200 = 1.28M tasks/sec.
3 Programming and Computation Model

Ray implements a dynamic task graph computation model, i.e., it models an application as a graph of dependent tasks that evolves during execution. On top of this model, Ray provides both an actor and a task-parallel programming abstraction. This unification differentiates Ray from related systems like CIEL, which only provides a task-parallel abstraction, and from Orleans [14] or Akka [1], which primarily provide an actor abstraction.

3.1 Programming Model

Tasks. A task represents the execution of a remote function on a stateless worker. When a remote function is invoked, a future representing the result of the task is returned immediately. Futures can be retrieved using ray.get() and passed as arguments into other remote functions without waiting for their result. This allows the user to express parallelism while capturing data dependencies. Table 1 shows Ray’s API.

Remote functions operate on immutable objects and are expected to be stateless and side-effect free: their outputs are determined solely by their inputs. This implies idempotence, which simplifies fault tolerance through function re-execution on failure.

Actors. An actor represents a stateful computation. Each actor exposes methods that can be invoked remotely and are executed serially. A method execution is similar to a task, in that it executes remotely and returns a future, but differs in that it executes on a stateful worker. A handle to an actor can be passed to other actors or tasks, making it possible for them to invoke methods on that actor.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>futures = f.remote(args)</td>
<td>Execute function f remotely. f.remote() can take objects or futures as inputs and returns one or more futures. This is non-blocking.</td>
</tr>
<tr>
<td>objects = ray.get(futures)</td>
<td>Return the values associated with one or more futures. This is blocking.</td>
</tr>
<tr>
<td>ready_futures = ray.wait(futures, k, timeout)</td>
<td>Return the futures whose corresponding tasks have completed as soon as either k have completed or the timeout expires.</td>
</tr>
<tr>
<td>actor = Class.remote(args)</td>
<td>Instantiate class Class as a remote actor, and return a handle to it. Call a method on the remote actor and return one or more futures. Both are non-blocking.</td>
</tr>
<tr>
<td>futures = actor.method.remote(args)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Ray API

Table 2 summarizes the properties of tasks and actors. Tasks enable fine-grained load balancing through leveraging load-aware scheduling at task granularity, input data locality, as each task can be scheduled on the node storing its inputs, and low recovery overhead, as there is no need to checkpoint and recover intermediate state. In contrast, actors provide much more efficient fine-grained updates, as these updates are performed on internal rather than external state, which typically requires serialization and deserialization. For example, actors can be used to implement parameter servers [32] and GPU-based iterative computations (e.g., training). In addition, actors can be used to wrap third-party simulators and other opaque handles that are hard to serialize.

To satisfy the requirements for heterogeneity and flexibility (Section 2), we augment the API in three ways. First, to handle concurrent tasks with heterogeneous durations, we introduce ray.wait(), which waits for the first k available results, instead of waiting for all results like ray.get(). Second, to handle resource-heterogeneous tasks, we enable developers to specify resource requirements so that the Ray scheduler can efficiently manage resources. Third, to improve flexibility, we enable nested remote functions, meaning that remote functions can invoke other remote functions. This is also critical for achieving high scalability (Section 4), as it enables multiple processes to invoke remote functions in a distributed fashion.

3.2 Computation Model

Ray employs a dynamic task graph computation model [31], in which the execution of both remote functions and actor methods is automatically triggered by the system when their inputs become available. In this section, we describe how the computation graph (Figure 4) is constructed from a user program (Figure 3). This program uses the API in Table 1 to implement the pseudocode from Figure 2.

Ignoring actors first, there are two types of nodes in a computation graph: data objects and remote function invocations, or tasks. There are also two types of edges: data edges and control edges. Data edges capture the de-
we add a third type of edge: a stateful edge. If method
with one key difference. To capture the state dependency
we add a data edge from
M
across subsequent method invocations on the same actor,
in the computation graph. They are identical to tasks
mote functions (Section 3.1): if task
the computation dependencies that result from nested re-
pendencies between data objects and tasks. More pre-
ately, if data object
D
is an output of task
T
, we add a
data edge from
T
to
D
. Similarly, if
D
is an input to
T
, we add a data edge from
D
to
T
. Control edges capture the computation dependencies that result from nested re-
remote functions (Section 4.1): if task
T
1
invokes task
T
2
, then we add a control edge from
T
1
to
T
2
.
Actor method invocations are also represented as nodes
in the computation graph. They are identical to tasks
with one key difference. To capture the state dependency
across subsequent method invocations on the same actor,
we add a third type of edge: a stateful edge. If method
M
i
is called right after method
M
j
on the same actor,
then we add a stateful edge from
M
i
to
M
j
. Thus, all

Figure 3: Python code implementing the example in Figure 2
in Ray. Note that @ray.remote indicates remote functions and
actors. Invocations of remote functions and actor methods return
futures, which can be passed to subsequent remote functions or
actor methods to encode task dependencies. Each actor has an
environment object self.env shared between all of its methods.

methods invoked on the same actor object form a chain
that is connected by stateful edges (Figure 4). This chain
captures the order in which these methods were invoked.

Stateful edges help us embed actors in an otherwise
stateless task graph, as they capture the implicit data de-
pendency between successive method invocations sharing
the internal state of an actor. Stateful edges also enable
us to maintain lineage. As in other dataflow systems [44],
we track data lineage to enable reconstruction. By explicitly
including stateful edges in the lineage graph, we can
easily reconstruct lost data, whether produced by remote
functions or actor methods (Section 4.2.3).

4 Architecture
Ray’s architecture comprises (1) an application layer
implementing the API, and (2) a system layer providing high
scalability and fault tolerance.

4.1 Application Layer
The application layer consists of three types of processes:

• Driver: A process executing the user program.

• Worker: A stateless process that executes tasks
(remote functions) invoked by a driver or another
The primary reason for the GCS and its design is to maintain the entire control state of the system, and it is a unique feature of our design. At its core, GCS is a key-value store with sub-functionality. We use sharding to achieve scale, and per-shard chain replication to provide fault tolerance. The primary reason for the GCS and its design is to maintain fault tolerance and low latency for a system that can dynamically spawn millions of tasks per second.

Fault tolerance in case of node failure requires a solution to maintain lineage information. Existing lineage-based solutions focus on coarse-grained parallelism and can therefore use a single node (e.g., master, driver) to store the lineage without impacting performance. However, this design is not scalable for a fine-grained and dynamic workload like simulation. Therefore, we decouple the durable lineage storage from the other system components, allowing each to scale independently.

Maintaining low latency requires minimizing overheads in task scheduling, which involves choosing where to execute, and subsequently task dispatch, which involves retrieving remote inputs from other nodes. Many existing dataflow systems couple these by storing object locations and sizes in a centralized scheduler, a natural design when the scheduler is not a bottleneck. However, the scale and granularity that Ray targets requires keeping the centralized scheduler off the critical path. Involving the scheduler in each object transfer is prohibitively expensive for primitives important to distributed training like allreduce, which is both communication-intensive and latency-sensitive. Therefore, we store the object metadata in the GCS rather than in the scheduler, fully decoupling task dispatch from task scheduling.

In summary, the GCS significantly simplifies Ray’s overall design, as it enables every component in the system to be stateless. This not only simplifies support for fault tolerance (i.e., on failure, components simply restart and read the lineage from the GCS), but also makes it easy to scale the distributed object store and scheduler independently, as all components share the needed state via the GCS. An added benefit is the easy development of debugging, profiling, and visualization tools.

4.2.1 Global Control Store (GCS)

As discussed in Section 2, Ray needs to dynamically schedule millions of tasks per second, tasks which may take as little as a few milliseconds. None of the cluster schedulers we are aware of meet these requirements. Most cluster computing frameworks, such as Spark, CIEL, and Dryad implement a centralized scheduler, which can provide locality but at latencies in the tens of ms. Distributed schedulers such as work stealing, Sparrow, and Canary can achieve high scale, but they either don’t consider data locality, or assume tasks belong to independent jobs, or assume the computation graph is known.

To satisfy the above requirements, we design a two-level hierarchical scheduler consisting of a global scheduler and per-node local schedulers. To avoid overloading the global scheduler, the tasks created at a node are submitted first to the node’s local scheduler. A local scheduler schedules tasks locally unless the node is overloaded (i.e., its local task queue exceeds a predefined threshold), or it cannot satisfy a task’s requirements (e.g., lacks a GPU). If a local scheduler decides not to schedule a task locally, it forwards it to the global scheduler. Since this scheduler attempts to schedule tasks locally first (i.e., at the leaves of the scheduling hierarchy), we call it a bottom-up scheduler.

4.2.2 Bottom-Up Distributed Scheduler
writes its outputs to the local object store. Replication
To minimize task latency, we implement an in-memory
This makes our scheduler architecture highly scalable.

The global scheduler considers each node’s load and
task’s constraints to make scheduling decisions. More pre-
cisely, the global scheduler identifies the set of nodes that
have enough resources of the type requested by the task,
and of these nodes selects the node which provides the
lowest estimated waiting time. At a given node, this time
is the sum of (i) the estimated time the task will be queued
at that node (i.e., task queue size times average task ex-
ecution), and (ii) the estimated transfer time of task’s
remote inputs (i.e., total size of remote inputs divided by
average bandwidth). The global scheduler gets the queue
size at each node and the node resource availability via
heartbeats, and the location of the task’s inputs and their
sizes from GCS. Furthermore, the global scheduler com-
putes the average task execution and the average transfer
bandwidth using simple exponential averaging. If the
global scheduler becomes a bottleneck, we can instantiate
more replicas all sharing the same information via GCS.
This makes our scheduler architecture highly scalable.

4.2.3 In-Memory Distributed Object Store
To minimize task latency, we implement an in-memory
distributed storage system to store the inputs and outputs
of every task, or stateless computation. On each node, we
implement the object store via shared memory. This al-

ows zero-copy data sharing between tasks running on
the same node. As a data format, we use Apache Arrow.

If a task’s inputs are not local, the inputs are replicated
to the local object store before execution. Also, a task
writes its outputs to the local object store. Replication
eliminates the potential bottleneck due to hot data objects
and minimizes task execution time as a task only read-
s/writes data from/to the local memory. This increases
throughput for computation-bound workloads, a profile
shared by many AI applications. For low latency, we keep
objects entirely in memory and evict them as needed to
disk using an LRU policy.

As with existing cluster computing frameworks, such as
Spark and Dryad, the object store is limited to immutable
data. This obviates the need for complex
consistency protocols (as objects are not updated), and
simplifies support for fault tolerance. In the case of node
failure, Ray recovers any needed objects through lineage
re-execution. The lineage stored in the GCS tracks both
stateless tasks and stateful actors during initial execution;
we use the former to reconstruct objects in the store.

For simplicity, our object store does not support dis-
tributed objects, i.e., each object fits on a single node. Dis-
tributed objects like large matrices or trees can be imple-
mented at the application level as collections of futures.

4.2.4 Implementation
Ray is an active open source project developed at the Un-
iversity of California, Berkeley. Ray fully integrates with
the Python environment and is easy to install by simply
running pip install ray. The implementation comprises
≈ 40K lines of code (LoC), 72% in C++ for the
system layer, 28% in Python for the application layer. The
GCS uses one Redis key-value store per shard, with
entirely single-key operations. GCS tables are sharded
by object and task IDs to scale, and every shard is chain-
replicated for fault tolerance. We implement both
the local and global schedulers as event-driven, single-
threaded processes. Internally, local schedulers maintain
cached state for local object metadata, tasks waiting for
inputs, and tasks ready for dispatch to a worker. To trans-
fer large objects between different object stores, we stripe
the object across multiple TCP connections.

4.3 Putting Everything Together
Figure 7 illustrates how Ray works end-to-end with a
simple example that adds two objects a and b, which
could be scalars or matrices, and returns result c. The
remote function add() is automatically registered with the
GCS upon initialization and distributed to every worker
in the system (step 0 in Figure 7).

Figure 7 shows the step-by-step operations triggered
by a driver invoking add.remote(a, b), where a and b are
stored on nodes N1 and N2, respectively. The driver sub-
mits add(a, b) to the local scheduler (step 1), which for-
wards it to a global scheduler (step 2). Next, the global
scheduler looks up the locations of add(a, b)’s arguments
in the GCS (step 3) and decides to schedule the task on
node N2, which stores argument b (step 4). The local
scheduler at node N2 checks whether the local object
store contains add(a, b)’s arguments (step 5). Since the

\*\*https://github.com/ray-project/ray
\*\*Note that N1 could also decide to schedule the task locally.
local store doesn’t have object $a$, it looks up $a$’s location in the GCS (step 6). Learning that $a$ is stored at $N1$, $N2$’s object store replicates it locally (step 7). As all arguments of $\text{add()}$ are now stored locally, the local scheduler invokes $\text{add()}$ at a local worker (step 8), which accesses the arguments via shared memory (step 9).

Figure 7 shows the step-by-step operations triggered by the execution of $\text{ray.get()}$ at $N1$, and of $\text{add()}$ at $N2$, respectively. Upon $\text{ray.get(id_c)}$’s invocation, the driver checks the local object store for the value $c$, using the future $id_c$ returned by $\text{add()}$ (step 1). Since the local object store doesn’t store $c$, it looks up its location in the GCS. At this time, there is no entry for $c$, as $c$ has not been created yet. As a result, $N1$’s object store registers a callback with the Object Table to be triggered when $c$’s entry has been created (step 2). Meanwhile, at $N2$, $\text{add()}$ completes its execution, stores the result $c$ in the local object store (step 3), which in turn adds $c$’s entry to the GCS (step 4). As a result, the GCS triggers a callback to $N1$’s object store with $c$’s entry (step 5). Next, $N1$ replicates $c$ from $N2$ (step 6), and returns $c$ to $\text{ray.get()}$ (step 7), which finally completes the task.

While this example involves a large number of RPCs, in many cases this number is much smaller, as most tasks are scheduled locally, and the GCS replies are cached by the global and local schedulers.

Figure 8: (a) Tasks leverage locality-aware placement. 1000 tasks with a random object dependency are scheduled onto one of two nodes. With locality-aware policy, task latency remains independent of the size of task inputs instead of growing by 1-2 orders of magnitude. (b) Near-linear scalability leveraging the GCS and bottom-up distributed scheduler. Ray reaches 1 million tasks per second throughput with 60 nodes. $x \in \{70, 80, 90\}$ omitted due to cost.

5 Evaluation

In our evaluation, we study the following questions:

1. How well does Ray meet the latency, scalability, and fault tolerance requirements listed in Section 2? (Section 5.1)

2. What overheads are imposed on distributed primitives (e.g., allreduce) written using Ray’s API? (Section 5.1)

3. In the context of RL workloads, how does Ray compare against specialized systems for training, serving, and simulation? (Section 5.2)

4. What advantages does Ray provide for RL applications, compared to custom systems? (Section 5.3)

All experiments were run on Amazon Web Services. Unless otherwise stated, we use m4.16xlarge CPU instances and p3.16xlarge GPU instances.

5.1 Microbenchmarks

Locality-aware task placement. Fine-grain load balancing and locality-aware placement are primary benefits of tasks in Ray. Actors, once placed, are unable to move their computation to large remote objects, while tasks can. In Figure 8a, tasks placed without data locality awareness (as is the case for actor methods), suffer 1-2 orders of magnitude latency increase at 10-100MB input data sizes. Ray unifies tasks and actors through the shared object store, allowing developers to use tasks for e.g., expensive postprocessing on output produced by simulation actors.

End-to-end scalability. One of the key benefits of
the Global Control Store (GCS) and the bottom-up distributed scheduler is the ability to horizontally scale the system to support a high throughput of fine-grained tasks, while maintaining fault tolerance and low-latency task scheduling. In Figure 5b, we evaluate this ability on an embarrassingly parallel workload of empty tasks, increasing the cluster size on the x-axis. We observe near-perfect linearity in progressively increasing task throughput. Ray exceeds 1 million tasks per second throughput at 60 nodes and continues to scale linearly beyond 1.8 million tasks per second at 100 nodes. The rightmost datapoint shows that Ray can process 100 million tasks in less than a minute (54s), with minimum variability. As expected, increasing task duration reduces throughput proportionally to mean task duration, but the overall scalability remains linear. While many realistic workloads may exhibit more limited scalability due to object dependencies and inherent limits to application parallelism, this demonstrates the scalability of our overall architecture under high load.

Object store performance. To evaluate the performance of the object store (Section 4.2.3), we track two metrics: IOPS (for small objects) and write throughput (for large objects). In Figure 9 the write throughput from a single client exceeds 15GB/s as object size increases. For larger objects, memcopy dominates object creation time. For smaller objects, the main overheads are in serialization and IPC between the client and object store.

GCS fault tolerance. To maintain low latency while providing strong consistency and fault tolerance, we build a lightweight chain replication layer on top of Redis. Figure 10a simulates recording Ray tasks to and reading tasks from the GCS, where keys are 25 bytes and values are 512 bytes. The client sends requests as fast as it can, having at most one in-flight request at a time. Failures are reported to the chain master either from the client (having received explicit errors, or timeouts despite retries) or from any server in the chain (having received explicit errors). Overall, reconfigurations caused a maximum client-observed delay of under 30ms (this includes both failure detection and recovery delays).

GCS flushing. Ray is equipped to periodically flush the contents of GCS to disk. In Figure 10b, we submit 50 million empty tasks sequentially and monitor GCS memory consumption. As expected, it grows linearly with the number of tasks tracked and eventually reaches the memory capacity of the system. At that point, the system becomes stalled and the workload fails to finish within a reasonable amount of time. With periodic GCS flushing, we achieve two goals. First, the memory footprint is capped at a user-configurable level (in the microbenchmark we employ an aggressive strategy where consumed memory is kept as low as possible). Second, the flushing mechanism provides a natural way to snapshot lineage to disk for long-running Ray applications.

Recovering from task failures. In Figure 11a we
primitive important to many machine learning workloads. Here, we evaluate whether Ray can natively support a ring allreduce \cite{allreduce} implementation with low enough overhead to match existing implementations \cite{wick2010mpi}. We find that Ray completes allreduce across 16 nodes on 100MB in \sim 200ms and 1GB in \sim 1200ms, surprisingly outperforming OpenMPI (v1.10), a popular MPI implementation, by 1.5× and 2× respectively (Figure 12a). We attribute Ray’s performance to its use of multiple threads for network transfers, taking full advantage of the 25Gbps connection between nodes on AWS, whereas OpenMPI sequentially sends and receives data on a single thread \cite{wick2010mpi}.

For smaller objects, OpenMPI outperforms Ray by switching to a lower overhead algorithm, an optimization we plan to implement in the future. Ray’s scheduler performance is critical to implementing primitives such as allreduce. In Figure 12b we inject artificial task execution delays and show that performance drops nearly 2× with just a few ms of extra latency. Systems with centralized schedulers like Spark and CIEL typically have scheduler overheads in the tens of milliseconds \cite{spark, ciel}, making such workloads impractical. Scheduler throughput also becomes a bottleneck since the number of tasks required by ring reduce scales quadratically with the number of participants.

5.2 Building blocks

End-to-end applications (e.g., AlphaGo \cite{alphago}) require a tight coupling of training, serving, and simulation. In this section, we isolate each of these workloads to a setting that illustrates a typical RL application’s requirements. Due to a flexible programming model targeted to RL, and a system designed to support this programming model, Ray matches and sometimes exceeds the performance of dedicated systems for these individual workloads.
We used OpenMPI 3.0, TF 1.8, and NCCL2 for all runs.

5.2.1 Distributed Training

We implement data-parallel synchronous SGD leveraging the Ray actor abstraction to represent model replicas. Model weights are synchronized via allreduce \(^5\) or parameter server, both implemented on top of the Ray API.

In Figure 13, we evaluate the performance of the Ray (synchronous) parameter-server SGD implementation against state-of-the-art implementations \(^4\), using the same TensorFlow model and synthetic data generator for each experiment. We compare only against TensorFlow-based systems to accurately measure the overhead imposed by Ray, rather than differences between the deep learning frameworks themselves. In each iteration, model replica actors compute gradients in parallel, send the gradients to a sharded parameter server, then read the summed gradients from the parameter server for the next iteration.

Figure 13 shows that Ray matches the performance of Horovod and is within 10% of distributed TensorFlow (in distributed_replicated mode). This is due to the ability to express the same application-level optimizations found in these specialized systems in Ray’s general-purpose API. A key optimization is the pipelining of gradient computation, transfer, and summation within a single iteration. To overlap GPU computation with network transfer, we use a custom TensorFlow operator to write tensors directly to Ray’s object store.

5.2.2 Serving

Model serving is an important component of end-to-end applications. Ray focuses primarily on the embedded serving of models to simulators running within the same dynamic task graph (e.g., within an RL application on Ray). In contrast, systems like Clipper \(^7\) focus on serving predictions to external clients.

In this setting, low latency is critical for achieving high utilization. To show this, in Table 4 we compare the server throughput achieved using a Ray actor to serve a policy versus using the open source Clipper system over REST. Here, both client and server processes are colocated on the same machine (a p3.8xlarge instance). This is often the case for RL applications but not for the general web serving workloads addressed by systems like Clipper. Due to its low-overhead serialization and shared memory abstractions, Ray achieves an order of magnitude higher throughput for a small fully connected policy model that takes in a large input and is also faster on a more expensive residual network policy model, similar to one used in AlphaGo Zero, that takes smaller input.

5.2.3 Simulation

Simulators used in RL produce results with variable lengths (“timesteps”) that, due to the tight loop with training, must be used as soon as they are available. The task heterogeneity and timeliness requirements make simulations hard to support efficiently in BSP-style systems. To demonstrate, we compare (1) an MPI implementation that submits \(3n\) parallel simulation runs on \(n\) cores in 3 rounds, with a global barrier between round 1 to (2) a Ray program that issues the same \(3n\) tasks while concurrently gathering simulation results back to the driver. Table 4 shows that both systems scale well, yet Ray achieves up to \(1.8\times\) throughput. This motivates a programming model that can dynamically spawn and collect the results of fine-grained simulation tasks.

Table 3: Throughput comparisons for Clipper \(^9\), a dedicated serving system, and Ray for two embedded serving workloads. We use a residual network and a small fully connected network, taking 10ms and 5ms to evaluate, respectively. The server is queried by clients that each send states of size 4KB and 100KB respectively in batches of 64.

Table 4: Timesteps per second for the Pendulum-v0 simulator in OpenAI Gym \(^1\). Ray allows for better utilization when running heterogeneous simulations at scale.

\(^5\)Note that experts can use MPI’s asynchronous primitives to get around barriers—at the expense of increased program complexity—we nonetheless chose such an implementation to simulate BSP.
5.3 RL Applications

Without a system that can tightly couple the training, simulation, and serving steps, reinforcement learning algorithms today are implemented as one-off solutions that make it difficult to incorporate optimizations that, for example, require a different computation structure or that utilize different architectures. Consequently, with implementations of two representative reinforcement learning applications in Ray, we are able to match and even outperform custom systems built specifically for these algorithms. The primary reason is the flexibility of Ray’s programming model, which can express application-level optimizations that would require substantial engineering effort to port to custom-built systems, but are transparently supported by Ray’s dynamic task graph execution engine.

5.3.1 Evolution Strategies

To evaluate Ray on large-scale RL workloads, we implement the evolution strategies (ES) algorithm and compare to the reference implementation—a system specially built for this algorithm that relies on Redis for messaging and low-level multiprocessing libraries for data-sharing. The algorithm periodically broadcasts a new policy to a pool of workers and aggregates the results of roughly 10000 tasks (each performing 10 to 1000 simulation steps).

As shown in Figure 14(a), an implementation on Ray scales to 8192 cores. Doubling the cores available yields an average completion time speedup of 1.6×. Conversely, the special-purpose system fails to complete at 2048 cores, where the work in the system exceeds the processing capacity of the application driver. To avoid this issue, the Ray implementation uses an aggregation tree of actors, reaching a median time of 3.7 minutes, more than twice as fast as the best published result (10 minutes).

Initial parallelization of a serial implementation using Ray required modifying only 7 lines of code. Performance improvement through hierarchical aggregation was easy to realize with Ray’s support for nested tasks and actors. In contrast, the reference implementation had several hundred lines of code dedicated to a protocol for communicating tasks and data between workers, and would require further engineering to support optimizations like hierarchical aggregation.

5.3.2 Proximal Policy Optimization

We implement Proximal Policy Optimization (PPO) in Ray and compare to a highly-optimized reference implementation that uses OpenMPI communication primitives. The algorithm is an asynchronous scatter-gather, where new tasks are assigned to simulation actors as they return rollouts to the driver. Tasks are submitted until 320000 simulation steps are collected (each task produces between 10 and 1000 steps). The policy update performs 20 steps of SGD with a batch size of 32768. The model parameters in this example are roughly 350KB. These experiments were run using p2.16xlarge (GPU) and m4.16xlarge (high CPU) instances.

As shown in Figure 14(b), the Ray implementation outperforms the optimized MPI implementation in all experiments, while using a fraction of the GPUs. The reason is that Ray is heterogeneity-aware and allows the user to utilize asymmetric architectures by expressing resource requirements at the granularity of a task or actor. The Ray implementation can then leverage TensorFlow’s single-process multi-GPU support and can pin objects in GPU memory when possible. This optimization cannot be easily ported to MPI due to the need to asynchronously gather rollouts to a single GPU process. Indeed, includes two custom implementations of PPO, one using MPI for large clusters and one that is optimized for GPUs but that is restricted to a single node. Ray allows for an implementation suitable for both scenarios.

Ray’s ability to handle resource heterogeneity also decreased PPO’s cost by a factor of 4.5, since CPU-only tasks can be scheduled on cheaper high-CPU instances. In contrast, MPI applications often exhibit symmetric architectures, in which all processes run the same code and require identical resources, in this case preventing the use of CPU-only machines for scale-out. Furthermore, the MPI implementation requires on-demand instances since it does not transparently handle failure. Assuming 4× cheaper spot instances, Ray’s fault tolerance and resource-aware scheduling together cut costs by 18×.
6 Related Work

Dynamic task graphs. Ray is closely related to CIEL \cite{40} and Dask \cite{49}. All three support dynamic task graphs with nested tasks and implement the futures abstraction. CIEL also provides lineage-based fault tolerance, while Dask, like Ray, fully integrates with Python. However, Ray differs in two aspects that have important performance consequences. First, Ray extends the task model with an actor abstraction. This is necessary for efficient stateful computation in distributed training and serving, to keep the model data collocated with the computation. Second, Ray employs a fully distributed and decoupled control plane and scheduler, instead of relying on a single master storing all metadata. This is critical for efficiently supporting primitives like allreduce without system modification. At peak performance for 100MB on 16 nodes, allreduce on Ray (Section 5.1) submits 32 rounds of 16 tasks in 200ms. Meanwhile, Dask reports a maximum scheduler throughput of 3k tasks/s on 512 cores \cite{3}. With a centralized scheduler, each round of allreduce would then incur a minimum of \( \sim 5\text{ms} \) of scheduling delay, translating to up to \( 2 \times \) worse completion time (Figure 12a). Even with a decentralized scheduler, coupling the control plane information with the scheduler leaves the latter on the critical path for data transfer, adding an extra roundtrip to every round of allreduce.

Dataflow systems. Popular dataflow systems, such as MapReduce \cite{20}, Spark \cite{45}, and Dryad \cite{28} have widespread adoption for analytics and ML workloads, but their computation model is too restrictive for a fine-grained and dynamic simulation workload. Spark and MapReduce implement the BSP execution model, which assumes that tasks within the same stage perform the same computation and take roughly the same amount of time. Dryad relaxes this restriction but lacks support for dynamic task graphs. Furthermore, none of these systems provide an actor abstraction, nor implement a distributed scalable control plane and scheduler. Finally, Naiad \cite{49} is a dataflow system that provides improved scalability for some workloads, but only supports static task graphs.

Machine learning frameworks. TensorFlow \cite{17} and MXNet \cite{18} target deep learning workloads and efficiently leverage both CPUs and GPUs. While they achieve great performance for training workloads consisting of static DAGs of linear algebra operations, they have limited support for the more general computation required to tightly couple training with simulation and embedded serving. TensorFlow Fold \cite{33} provides some support for dynamic task graphs, as well as MXNet through its internal C++ APIs, but neither fully supports the ability to modify the DAG during execution in response to task progress, task completion times, or faults. TensorFlow and MXNet in principle achieve generality by allowing the programmer to simulate low-level message-passing and synchronization primitives, but the pitfalls and user experience in this case are similar to those of MPI. OpenMPI \cite{22} can achieve high performance, but it is relatively hard to program as it requires explicit coordination to handle heterogeneous and dynamic task graphs. Furthermore, it forces the programmer to explicitly handle fault tolerance.

Actor systems. Orleans \cite{14} and Akka \cite{41} are two actor frameworks well suited to developing highly available and concurrent distributed systems. However, compared to Ray, they provide less support for recovery from data loss. To recover stateful actors, the Orleans developer must explicitly checkpoint actor state and intermediate responses. Stateless actors in Orleans can be replicated for scale-out, and could therefore act as tasks, but unlike in Ray, they have no lineage. Similarly, while Akka explicitly supports persisting actor state across failures, it does not provide efficient fault tolerance for stateless computation (i.e., tasks). For message delivery, Orleans provides at-least-once and Akka provides at-most-once semantics. In contrast, Ray provides transparent fault tolerance and exactly-once semantics, as each method call is logged in the GCS and both arguments and results are immutable. We find that in practice these limitations do not affect the performance of our applications. Erlang \cite{10} and C++ Actor Framework \cite{17}, two other actor-based systems, have similarly limited support for fault tolerance.

Global control store and scheduling. The concept of logically centralizing the control plane has been previously proposed in software defined networks (SDNs) \cite{16}, distributed file systems (e.g., GFS \cite{23}), resource management (e.g., Omega \cite{52}), and distributed frameworks (e.g., MapReduce \cite{20}, BOOM \cite{9}), to name a few. Ray draws inspiration from these pioneering efforts, but provides significant improvements. In contrast with SDNs, BOOM, and GFS, Ray decouples the storage of the control plane information (e.g., GCS) from the logic implementation (e.g., schedulers). This allows both storage and computation layers to scale independently, which is key to achieving our scalability targets. Omega uses a distributed architecture in which schedulers coordinate via globally shared state. To this architecture, Ray adds global schedulers to balance load across local schedulers, and targets ms-level, not second-level, task scheduling.

Ray implements a unique distributed bottom-up scheduler that is horizontally scalable, and can handle dynamically constructed task graphs. Unlike Ray, most existing cluster computing systems \cite{20, 64, 40} use a centralized scheduler architecture. While Sparrow \cite{45} is decentralized, its schedulers make independent decisions, limiting the possible scheduling policies, and all tasks of a job are handled by the same global scheduler. Mesos \cite{46} implements a two-level hierarchical scheduler, but its top-level scheduler manages frameworks, not individual tasks.
work-stealing scheduler achieves provably efficient load-balancing for dynamic task graphs. However, with no central coordinator like Ray’s global scheduler, this fully parallel design is also difficult to extend to support data locality and resource heterogeneity in a distributed setting.

7 Discussion and Experiences

Building Ray has been a long journey. It started two years ago with a Spark library to perform distributed training and simulations. However, the relative inflexibility of the BSP model, the high per-task overhead, and the lack of an actor abstraction led us to develop a new system. Since we released Ray roughly one year ago, several hundreds of people have used it and several companies are running it in production. Here we discuss our experience developing and using Ray, and some early user feedback.

API. In designing the API, we have emphasized minimalism. Initially we started with a basic task abstraction. Later, we added the wait() primitive to accommodate rollouts with heterogeneous durations and the actor abstraction to accommodate third-party simulators and amortize the overhead of expensive initializations. While the resulting API is relatively low-level, it has proven both powerful and simple to use. We have already used this API to implement many state-of-the-art RL algorithms on top of Ray, including A3C [56], PPO [51], DQN [57], ES [49], DDPG [55], and Ape-X [27]. In most cases it took us just a few tens of lines of code to port these algorithms to Ray. Based on early user feedback, we are considering enhancing the API to include higher level primitives and libraries, which could also inform scheduling decisions.

Limitations. Given the workload generality, specialized optimizations are hard. For example, we must make scheduling decisions without full knowledge of the computation graph. Scheduling optimizations in Ray might require more complex runtime profiling. In addition, storing lineage for each task requires the implementation of garbage collection policies to bound storage costs in the GCS, a feature we are actively developing.

Fault tolerance. We are often asked if fault tolerance is really needed for AI applications. After all, due to the statistical nature of many AI algorithms, one could simply ignore failed rollouts. Based on our experience, our answer is “yes”. First, the ability to ignore failures makes applications much easier to write and reason about. Second, our particular implementation of fault tolerance via deterministic replay dramatically simplifies debugging as it allows us to easily reproduce most errors. This is particularly important since, due to their stochasticity, AI algorithms are notoriously hard to debug. Third, fault tolerance helps save money since it allows us to run on cheap resources like spot instances on AWS. Of course, this comes at the price of some overhead. However, we found this overhead to be minimal for our target workloads.

GCS and Horizontal Scalability. The GCS dramatically simplified Ray development and debugging. It enabled us to query the entire system state while debugging Ray itself, instead of having to manually expose internal component state. In addition, the GCS is also the backend for our timeline visualization tool, used for application-level debugging.

The GCS was also instrumental to Ray’s horizontal scalability. In Section 5, we were able to scale by adding more shards whenever the GCS became a bottleneck. The GCS also enabled the global scheduler to scale by simply adding more replicas. Due to these advantages, we believe that centralizing control state will be a key design component of future distributed systems.

8 Conclusion

No general-purpose system today can efficiently support the tight loop of training, serving, and simulation. To express these core building blocks and meet the demands of emerging AI applications, Ray unifies task-parallel and actor programming models in a single dynamic task graph and employs a scalable architecture enabled by the global control store and a bottom-up distributed scheduler. The programming flexibility, high throughput, and low latencies simultaneously achieved by this architecture is particularly important for emerging artificial intelligence workloads, which produce tasks diverse in their resource requirements, duration, and functionality. Our evaluation demonstrates linear scalability up to 1.8 million tasks per second, transparent fault tolerance, and substantial performance improvements on several contemporary RL workloads. Thus, Ray provides a powerful combination of flexibility, performance, and ease of use for the development of future AI applications.

9 Acknowledgments

This research is supported in part by NSF CISE Expeditions Award CCF-1730628 and gifts from Alibaba, Amazon Web Services, Ant Financial, Arm, CapitalOne, Ericsson, Facebook, Google, Huawei, Intel, Microsoft, Scotiabank, Splunk and VMware as well as by NSF grant DGE-1106400. We are grateful to our anonymous reviewers and our shepherd, Miguel Castro, for thoughtful feedback, which helped improve the quality of this paper.
References


[10] Armstrong, J., Virding, R., Wikström, C., and Williams, M. Concurrent programming in ERLANG.


