# distributed Documentation

***Release 1.9.5***

**Matthew Rocklin**

October 01, 2016

Distributed is a lightweight library for distributed computing in Python. It extends both the `concurrent.futures` and `dask` APIs to moderate sized clusters. Distributed provides data-local computation by keeping data on worker nodes, running computations where data lives, and by managing complex data dependencies between tasks.

See the quickstart to get started.

CHAPTER 1

# Motivation

Why build yet-another-distributed-system?

`Distributed` serves to complement the existing PyData analysis stack. In particular it meets the following needs:

- **Low latency:** Each task suffers about 1ms of overhead. A small computation and network roundtrip can complete in less than 10ms.
- **Peer-to-peer data sharing:** Workers communicate with each other to share data. This removes central bottlenecks for data transfer.
- **Complex Scheduling:** Supports complex workflows (not just map/filter/reduce) which are necessary for sophisticated algorithms used in nd-arrays, machine learning, image processing, and statistics.
- **Pure Python:** Built in Python using well-known technologies. This eases installation, improves efficiency (for Python users), and simplifies debugging.
- **Data Locality:** Scheduling algorithms cleverly execute computations where data lives. This minimizes network traffic and improves efficiency.
- **Familiar APIs:** Compatible with the concurrent.futures API in the Python standard library. Compatible with dask API for parallel algorithms
- **Easy Setup:** As a Pure Python package distributed is `pip` installable and easy to set up on your own cluster.

CHAPTER 2

# Contents

## 2.1 Install Distributed

You can install distributed with `conda`, with `pip`, or by installing from source.

### 2.1.1 Conda

Install distributed from the dask channel using conda:

```
conda install distributed -c dask
```

### 2.1.2 Pip

Or install distributed with `pip`:

```
pip install distributed --upgrade
```

### 2.1.3 Source

To install distributed from source, clone the repository from github:

```
git clone https://github.com/dask/distributed.git
cd distributed
python setup.py install
```

### 2.1.4 Notes

**Note for Macports users:** There is a known issue. with python from macports that makes executables be placed in a location that is not available by default. A simple solution is to extend the *PATH* environment variable to the location where python from macports install the binaries:

```
$ export PATH=/opt/local/Library/Frameworks/Python.framework/Versions/3.5/bin:$PATH

or

$ export PATH=/opt/local/Library/Frameworks/Python.framework/Versions/2.7/bin:$PATH
```

## 2.2 Quickstart

### 2.2.1 Install

```
$ pip install distributed --upgrade
```

See installation document for more information.

### 2.2.2 Setup Cluster

Set up a fake cluster on your local computer:

```
$ dscheduler
$ dworker 127.0.0.1:8786
$ dworker 127.0.0.1:8786
$ dworker 127.0.0.1:8786
```

Or if you can ssh into your own computer (or others) then use the `dcluster` command, providing hostnames or IP addresses:

```
$ dcluster 127.0.0.1 127.0.0.1 127.0.0.1 127.0.0.1
```

See setup for advanced use or EC2 quickstart for instructions on how to deploy on Amazon's Elastic Compute Cloud.

### 2.2.3 Launch Executor

Launch an Executor to interact with the network. Point to the center IP/port.:

```
$ ipython
```

```
>>> from distributed import Executor
>>> executor = Executor('127.0.0.1:8786')
```

#### Map and Submit Functions

Use the `map` and `submit` methods to launch computations on the cluster. The `map/submit` functions send the function and arguments to the remote workers for processing. They return `Future` objects that refer to remote data on the cluster. The `Future` returns immediately while the computations run remotely in the background.

```
>>> def square(x):
        return x ** 2

>>> def neg(x):
        return -x

>>> A = executor.map(square, range(10))
>>> B = executor.map(neg, A)
>>> total = executor.submit(sum, B)
>>> total.result()
-285
```

#### Gather

The `map/submit` functions return `Future` objects, lightweight tokens that refer to results on the cluster. By default the results of computations *stay on the cluster*.

```
>>> total  # Function hasn't yet completed
<Future: status: waiting, key: sum-58999c52e0fa35c7d7346c098f5085c7>

>>> total  # Function completed, result ready on remote worker
<Future: status: finished, key: sum-58999c52e0fa35c7d7346c098f5085c7>
```

Gather results to your local machine either with the `Future.result` method for a single future, or with the `Executor.gather` method for many futures at once.

```
>>> total.result()      # result for single future
-285
>>> executor.gather(A) # gather for many futures
[0, 1, 4, 9, 16, 25, 36, 49, 64, 81]
```

#### Restart

When things go wrong, or when you want to reset the cluster state, call the `restart` method.

```
>>> executor.restart()
```

See executor for advanced use.

## 2.3 Setup Network

A `distributed` network consists of one `Scheduler` node and several `Worker` nodes. One can set these up in a variety of ways

### 2.3.1 Using the Command Line

We launch the `dscheduler` executable in one process and the `dworker` executable in several processes, possibly on different machines.

Launch `dscheduler` on one node:

```
$ dscheduler
Start scheduler at 192.168.0.1:8786
```

Then launch `dworker` on the rest of the nodes, providing the address to the node that hosts `dscheduler`:

```
$ dworker 192.168.0.1:8786
Start worker at:            192.168.0.2:12345
Registered with center at:  192.168.0.1:8786

$ dworker 192.168.0.1:8786
Start worker at:            192.168.0.3:12346
Registered with center at:  192.168.0.1:8786

$ dworker 192.168.0.1:8786
Start worker at:            192.168.0.4:12347
Registered with center at:  192.168.0.1:8786
```

There are various mechanisms to deploy these executables on a cluster, ranging from manualy SSH-ing into all of the nodes to more automated systems like SGE/SLURM/Torque or Yarn/Mesos.

### 2.3.2 Using SSH

The convenience script `dcluster` opens several SSH connections to your target computers and initializes the network accordingly. You can give it a list of hostnames or IP addresses:

```
$ dcluster 192.168.0.1 192.168.0.2 192.168.0.3 192.168.0.4
```

Or you can use normal UNIX grouping:

```
$ dcluster 192.168.0.{1,2,3,4}
```

Or you can specify a hostfile that includes a list of hosts:

```
$ cat hostfile.txt
192.168.0.1
192.168.0.2
192.168.0.3
192.168.0.4

$ dcluster --hostfile hostfile.txt
```

### 2.3.3 Using the Python API

Alternatively you can start up the `distributed.scheduler.Scheduler` and `distributed.worker.Worker` objects within a Python session manually. Both are `torando.tcpserver.TCPServer` objects.

Start the Scheduler, provide the listening port (defaults to 8786) and Tornado IOLoop (defaults to `IOLoop.current()`)

```
from distributed import Scheduler
s = Scheduler(loop=loop)
s.start(port)
```

On other nodes start worker processes that point to the IP address and port of the scheduler.

```
from distributed import Worker
w = Worker('192.168.0.1', 8786, loop=loop)
w.start(0)  # choose randomly assigned port
```

Alternatively, replace `Worker` with `Nanny` if you want your workers to be managed in a separate process by a local nanny process.

If you do not already have a Tornado event loop running you will need to create and start one, possibly in a separate thread.

```
from tornado.ioloop import IOLoop
loop = IOLoop()

from threading import Thread
t = Thread(target=loop.start)
t.start()
```

### 2.3.4 Using Amazon EC2

See the EC2 quickstart for information on the `dec2` easy setup script to launch a canned cluster on EC2.

### 2.3.5 Cleanup

It is common and safe to terminate the cluster by just killing the processes. The workers and scheduler have no persistent state.

Programmatically you can use the client interface (`rpc`) to call the `terminate` methods on the workers and schedulers.

## 2.4 EC2 Startup Script

First, add your AWS credentials to `~/.aws/credentials` like this:

```
[default]
aws_access_key_id = YOUR_ACCESS_KEY
aws_secret_access_key = YOUR_SECRET_KEY
```

For other ways to manage or troubleshoot credentials, see the boto3 docs.

Now, you can quickly deploy a scheduler and workers on EC2 using the `dec2` quickstart application:

```
pip install dec2
dec2 up --keyname YOUR-AWS-KEY --keypair ~/.ssh/YOUR-AWS-SSH-KEY.pem
```

This provisions a cluster on Amazon's EC2 cloud service, installs Anaconda, and sets up a scheduler and workers. In then prints out instructions on how to connect to the cluster.

### 2.4.1 Options

The `dec2` startup script comes with the following options for creating a cluster:

```
$ dec2 up --help
Usage: dec2 up [OPTIONS]

Options:
  --keyname TEXT                Keyname on EC2 console  [required]
  --keypair PATH                Path to the keypair that matches the keyname [required]
  --name TEXT                   Tag name on EC2
  --region-name TEXT            AWS region  [default: us-east-1]
  --ami TEXT                    EC2 AMI  [default: ami-d05e75b8]
  --username TEXT               User to SSH to the AMI  [default: ubuntu]
  --type TEXT                   EC2 Instance Type  [default: m3.2xlarge]
  --count INTEGER               Number of nodes  [default: 4]
  --security-group TEXT         Security Group Name  [default: dec2-default]
  --volume-type TEXT            Root volume type  [default: gp2]
  --volume-size INTEGER         Root volume size (GB)  [default: 500]
  --file PATH                   File to save the metadata  [default: cluster.yaml]
  --provision / --no-provision  Provision salt on the nodes  [default: True]
  --dask / --no-dask            Install Dask.Distributed in the cluster [default: True]
  --nprocs INTEGER              Number of processes per worker  [default: 1]
  -h, --help                    Show this message and exit.
```

### 2.4.2 Connect

Connection instructions follow successful completion of the `dec2 up` command. The involve the following:

```
dec2 ssh            # SSH into head node
ipython             # Start IPython console on head node
```

```
>>> from distributed import Executor, s3, progress
>>> e = Executor('127.0.0.1:8786')
```

This executor now has access to all the cores of your cluster.

### 2.4.3 Destroy

You can destroy your cluster from your local machine with the destroy command:

```
dec2 destroy
```

## 2.5 Web Interface

Information about the current state of the network helps to track progress, identify performance issues, and debug failures.

Dask.distributed includes a web interface to help deliver this information over a normal web page in real time. This web interface is launched by default wherever the scheduler is launched if the scheduler machine has Bokeh installed (`conda install bokeh`). The web interface is normally available at `http://scheduler-address:8787/status/` and can be viewed any normal web browser.

The web UI shows basic statistics on all worker machines, grouped by physical address. This includes information like CPU/memory load, active tasks, latency, network and disk usage, etc.. The tabular statistics are updated about once a second.

It also shows the progress of all groups of tasks currently running on the cluster. Dark blue is used for tasks that are completed and in memory, light blue for tasks that are completed and have been released, gray for not yet completed, and black for erred. The progress bar is updated every 100ms.

There is a resource plot showing total CPU and memory use of the cluster over the last few minutes.

Finally there is a plot of tasks as they complete, showing their start and end times, start and end transfer times (in red), as well as which worker they were run on. Hovering over any of the tasks gives the task name as well as more precise information. This plot can be invaluable to determine performance issues. It is updated every 200ms. It only includes the most recent thousand tasks. For the most recent 20000 tasks visit http://my-scheduler-address:8787/tasks , although beware that this page is not updated in real time.

## 2.6 Examples

### 2.6.1 Word count in HDFS

#### Setup

In this example, we'll use `distributed` with the `hdfs3` library to count the number of words in text files (Enron email dataset, 6.4 GB) stored in HDFS.

Copy the text data from Amazon S3 into HDFS on the cluster:

```
$ hadoop distcp s3n://AWS_SECRET_ID:AWS_SECRET_KEY@blaze-data/enron-email hdfs:///tmp/enron
```

where `AWS_SECRET_ID` and `AWS_SECRET_KEY` are valid AWS credentials.

Start the `distributed` scheduler and workers on the cluster.

### Code example

Import `distributed`, `hdfs3`, and other standard libraries used in this example:

```
>>> import hdfs3
>>> from collections import defaultdict, Counter
>>> from distributed import Executor, progress
```

Initalize a connection to HDFS, replacing `NAMENODE_HOSTNAME` and `NAMENODE_PORT` with the hostname and port (default: 8020) of the HDFS namenode.

```
>>> hdfs = hdfs3.HDFileSystem('NAMENODE_HOSTNAME', port=NAMENODE_PORT)
```

Initalize a connection to the `distributed` executor, replacing `EXECUTOR_IP` and `EXECUTOR_PORT` with the IP address and port of the `distributed` scheduler.

```
>>> e = Executor('EXECUTOR_IP:EXECUTOR_PORT')
```

Generate a list of filenames from the text data in HDFS:

```
>>> filenames = hdfs.glob('/tmp/enron/*/*')
>>> print(filenames[:5])

['/tmp/enron/edrm-enron-v2_nemec-g_xml.zip/merged.txt',
'/tmp/enron/edrm-enron-v2_ring-r_xml.zip/merged.txt',
'/tmp/enron/edrm-enron-v2_bailey-s_xml.zip/merged.txt',
'/tmp/enron/edrm-enron-v2_fischer-m_xml.zip/merged.txt',
'/tmp/enron/edrm-enron-v2_geaccone-t_xml.zip/merged.txt']
```

Print the first 1024 bytes of the first text file:

```
>>> print(hdfs.head(filenames[0]))

b'Date: Wed, 29 Nov 2000 09:33:00 -0800 (PST)\r\nFrom: Xochitl-Alexis Velasc
o\r\nTo: Mark Knippa, Mike D Smith, Gerald Nemec, Dave S Laipple, Bo Barnwel
l\r\nCc: Melissa Jones, Iris Waser, Pat Radford, Bonnie Shumaker\r\nSubject:
 Finalize ECS/EES Master Agreement\r\nX-SDOC: 161476\r\nX-ZLID: zl-edrm-enro
n-v2-nemec-g-2802.eml\r\n\r\nPlease plan to attend a meeting to finalize the
 ECS/EES  Master Agreement \r\ntomorrow 11/30/00 at 1:30 pm CST.\r\n\r\nI wi
ll email everyone tomorrow with location.\r\n\r\nDave-I will also email you
the call in number tomorrow.\r\n\r\nThanks\r\nXochitl\r\n\r\n**********\r\n
EDRM Enron Email Data Set has been produced in EML, PST and NSF format by ZL
 Technologies, Inc. This Data Set is licensed under a Creative Commons Attri
bution 3.0 United States License <http://creativecommons.org/licenses/by/3.0
/us/> . To provide attribution, please cite to "ZL Technologies, Inc. (http:
//www.zlti.com)."\r\n**********\r\nDate: Wed, 29 Nov 2000 09:40:00 -0800 (P
ST)\r\nFrom: Jill T Zivley\r\nTo: Robert Cook, Robert Crockett, John Handley
, Shawna'
```

Create a function to count words in each file:

```
>>> def count_words(fn):
...     word_counts = defaultdict(int)
...     with hdfs.open(fn) as f:
...         for line in f:
...             for word in line.split():
...                 word_counts[word] += 1
...     return word_counts
```

Before we process all of the text files using the distributed workers, let's test our function locally by counting the number of words in the first text file:

```
>>> counts = count_words(filenames[0])
>>> print(sorted(counts.items(), key=lambda k_v: k_v[1], reverse=True)[:10])

[(b'the', 144873),
 (b'of', 98122),
 (b'to', 97202),
 (b'and', 90575),
 (b'or', 60305),
 (b'in', 53869),
 (b'a', 43300),
 (b'any', 31632),
 (b'by', 31515),
 (b'is', 30055)]
```

We can perform the same operation of counting the words in the first text file, except we will use `e.submit` to execute the computation on a `distributed` worker:

```
>>> future = e.submit(count_words, filenames[0])
>>> counts = future.result()
>>> print(sorted(counts.items(), key=lambda k_v: k_v[1], reverse=True)[:10])

[(b'the', 144873),
 (b'of', 98122),
 (b'to', 97202),
 (b'and', 90575),
 (b'or', 60305),
 (b'in', 53869),
 (b'a', 43300),
 (b'any', 31632),
 (b'by', 31515),
 (b'is', 30055)]
```

We are ready to count the number of words in all of the text files using `distributed` workers. Note that the `map` operation is non-blocking, and you can continue to work in the Python shell/notebook while the computations are running.

```
>>> futures = e.map(count_words, filenames)
```

We can check the status of some `futures` while all of the text files are being processed:

```
>>> len(futures)

161

>>> futures[:5]

[<Future: status: finished, key: count_words-5114ab5911de1b071295999c9049e941>,
 <Future: status: pending, key: count_words-d9e0d9daf6a1eab4ca1f26033d2714e7>,
```

```
 <Future: status: pending, key: count_words-d2f365a2360a075519713e9380af45c5>,
 <Future: status: pending, key: count_words-bae65a245042325b4c77fc8dde1acf1e>,
 <Future: status: pending, key: count_words-03e82a9b707c7e36eab95f4feec1b173>]

>>> progress(futures)

[########################################] | 100% Completed |  3min  0.2s
```

When the `futures` finish reading in all of the text files and counting words, the results will exist on each worker. This operation required about 3 minutes to run on a cluster with three worker machines, each with 4 cores and 16 GB RAM.

Note that because the previous computation is bound by the GIL in Python, we can speed it up by starting the `distributed` workers with the `--nprocs 4` option.

To sum the word counts for all of the text files, we need to gather some information from the `distributed` workers. To reduce the amount of data that we gather from the workers, we can define a function that only returns the top 10,000 words from each text file.

```
>>> def top_items(d):
...     items = sorted(d.items(), key=lambda kv: kv[1], reverse=True)[:10000]
...     return dict(items)
```

We can then `map` the futures from the previous step to this culling function. This is a convenient way to construct a pipeline of computations using futures:

```
>>> futures2 = e.map(top_items, futures)
```

We can `gather` the resulting culled word count data for each text file to the local process:

```
>>> results = e.gather(iter(futures2))
```

To sum the word counts for all of the text files, we can iterate over the results in `futures2` and update a local dictionary that contains all of the word counts.

```
>>> all_counts = Counter()
>>> for result in results:
...     all_counts.update(result)
```

Finally, we print the total number of words in the results and the words with the highest frequency from all of the text files:

```
>>> print(len(all_counts))

8797842

>>> print(sorted(all_counts.items(), key=lambda k_v: k_v[1], reverse=True)[:10])

[(b'0', 67218380),
 (b'the', 19586868),
 (b'-', 14123768),
 (b'to', 11893464),
 (b'N/A', 11814665),
 (b'of', 11724827),
 (b'and', 10253753),
 (b'in', 6684937),
 (b'a', 5470371),
 (b'or', 5227805)]
```

The complete Python script for this example is shown below:

```
# word-count.py

import hdfs3
from collections import defaultdict, Counter
from distributed import Executor, progress

hdfs = hdfs3.HDFileSystem('NAMENODE_HOSTNAME', port=NAMENODE_PORT)
e = Executor('EXECUTOR_IP:EXECUTOR_PORT')

filenames = hdfs.glob('/tmp/enron/*/*')
print(filenames[:5])
print(hdfs.head(filenames[0]))


def count_words(fn):
    word_counts = defaultdict(int)
    with hdfs.open(fn) as f:
        for line in f:
            for word in line.split():
                word_counts[word] += 1
    return word_counts

counts = count_words(filenames[0])
print(sorted(counts.items(), key=lambda k_v: k_v[1], reverse=True)[:10])

future = e.submit(count_words, filenames[0])
counts = future.result()
print(sorted(counts.items(), key=lambda k_v: k_v[1], reverse=True)[:10])

futures = e.map(count_words, filenames)
len(futures)
futures[:5]
progress(futures)


def top_items(d):
    items = sorted(d.items(), key=lambda kv: kv[1], reverse=True)[:10000]
    return dict(items)

futures2 = e.map(top_items, futures)
results = e.gather(iter(futures2))

all_counts = Counter()
for result in results:
    all_counts.update(result)

print(len(all_counts))

print(sorted(all_counts.items(), key=lambda k_v: k_v[1], reverse=True)[:10])
```

## 2.7 Executor

The Executor is the primary entry point for users of `distributed`.

After you setup a cluster, initialize an `Executor` by pointing it to the address of a `Scheduler`:

```
>>> from distributed import Executor
>>> executor = Executor('127.0.0.1:8786')
```

### 2.7.1 Usage

#### submit

You can submit individual function calls with the `executor.submit` method

```
>>> def inc(x):
        return x + 1

>>> x = executor.submit(inc, 10)
>>> x
<Future - key: inc-e4853cffcc2f51909cdb69d16dacd1a5>
```

The result is on one of the distributed workers. We can continue using `x` in further calls to `submit`:

```
>>> type(x)
Future
>>> y = executor.submit(inc, x)
```

#### Gather results

We can collect results in a variety of ways. First, we can use the `.result()` method on futures

```
>>> x.result()
2
```

Second, we can use the gather method on the executor

```
>>> executor.gather([x, y])
(2, 3)
```

Third, we can use the `as_completed` function to iterate over results as soon as they become available.

```
>>> from distributed import as_completed
>>> seq = as_completed([x, y])
>>> next(seq).result()
2
>>> next(seq).result()
3
```

But, as always, we want to minimize communicating results back to the local process. It's often best to leave data on the cluster and operate on it remotely with functions like `submit`, `map`, `get` and `compute`. See efficiency for more information on efficient use of distributed.

#### map

We can map a function over many inputs at once

```
>>> L = executor.map(inc, range(10))
```

The `map` method returns a list of futures. This is a break with the `concurrent.futures` API, which returns the results directly. We keep the results as futures so that they can stay on the distributed cluster.

Additionally, we don't do any kind of batching so every function application will be a new task which will have a couple milliseconds of overhead. It is unwise to use `executor.map` for small, fast functions where scheduling overhead is likely to be more expensive than the cost of the function itself. For example, our function `inc` is actually a *terrible* function to parallelize in practice.

#### dask

Distributed provides a dask compliant task scheduling interface. It provides this through two methods, `get` (synchronous) and `compute` (asynchronous).

**get**

We provide dask graph dictionaries to the scheduler:

```
>>> dsk = {'x': 1, 'y': (inc, 'x')}
>>> executor.get(dsk, 'y')
2
```

This function pulls results back by default. This is so that it can integrate with existing dask code.

```
>>> import dask.array as da
>>> x = da.random.random(1000000000, chunks=(1000000,))
>>> x.sum().compute()  # use local threads
499999359.23511785
>>> x.sum().compute(get=executor.get)  # use distributed cluster
499999359.23511785
```

**compute**

We can also provide dask collections (arrays, bags, dataframes, imperative values) to the executor with the `compute` method.

```
>>> type(x)
dask.array.Array
>>> type(df)
dask.dataframe.DataFrame

>>> x_future, df_future = executor.compute(x, df)
```

This immediately returns standard `Future` objects as would be returned by `submit` or `map`.

#### restart

When things go wrong, restart the cluster with the `.restart()` method.

```
>>> executor.restart()
```

This both resets the scheduler state and all of the worker processes. All current data and computations will be lost. All existing futures set their status to `'cancelled'`.

See resilience for more information.

### 2.7.2 Internals

#### Data Locality

By default the executor does not bring results back to your local computer but leaves them on the distributed network. As a result, computations on returned results like the following don't require any data transfer.

```
>>> y = executor.submit(inc, x)  # no data transfer required
```

In addition, the internal scheduler endeavors to run functions on worker nodes that already have the necessary input data. It avoids worker-to-worker communication when convenient.

### Pure Functions by Default

By default we assume that all functions are pure. If this is not the case you should use the `pure=False` keyword argument.

The executor associates a key to all computations. This key is accessible on the Future object.

```
>>> from operator import add
>>> x = executor.submit(add, 1, 2)
>>> x.key
'add-ebf39f96ad7174656f97097d658f3fa2'
```

This key should be the same accross all computations with the same inputs and across all machines. If you run the computation above on any computer with the same environment then you should get the exact same key.

The scheduler avoids redundant computations. If the result is already in memory from a previous call then that old result will be used rather than recomputing it. Calls to submit or map are idempotent in the common case.

While convenient, this feature may be undesired for impure functions, like `random`. In these cases two calls to the same function with the same inputs should produce different results. We accomplish this with the `pure=False` keyword argument. In this case keys are randomly generated (by `uuid4`.)

```
>>> import numpy as np
>>> executor.submit(np.random.random, 1000, pure=False).key
'random_sample-fc814a39-ee00-42f3-8b6f-cac65bcb5556'
>>> executor.submit(np.random.random, 1000, pure=False).key
'random_sample-a24e7220-a113-47f2-a030-72209439f093'
```

### Garbage Collection

Prolonged use of `distributed` may allocate a lot of remote data. The executor can clean up unused results by reference counting.

The executor reference counts `Future` objects. When a particular key no longer has any Future objects pointing to it it will be released from distributed memory if no active computations still require it.

In this way garbage collection in the distributed memory space of your cluster mirrors garbage collection within your local Python session.

Known future keys and reference counts can be found in the following dictionaries:

```
>>> executor.futures
>>> executor.refcount
```

The scheduler also cleans up intermediate results when provided full dask graphs. You can always use the lower level `delete` or `clear` functions in `distributed.client` to manage data manually.

### Dask Graph

The executor and scheduler maintain a dask graph of all known computations. This graph is accessible via the `.dask` attribute. At times it may be worth visualizing this object.

```
>>> executor.dask

>>> from dask.base import visualize
>>> visualize(executor, filename='executor.pdf')
```

All functions like `.submit`, `.map`, and `.get` just add small subgraphs to this graph. Functions like `.result`, `as_completed`, or `.gather`, wait until their respective parts of the graph have completed. All of these actions are asynchronous to the actual execution of the graph, which is managed in a background thread.

The dask graph is also used to recover results in case of failure.

#### Coroutines

If you are operating in an asynchronous environment then all blocking functions listed above have asynchronous equivalents. Currently these have the exact same name but are prepended with an underscore (_) so, `.result` is synchronous while `._result` is asynchronous. If a function has no asynchronous counterpart then that means it does not significantly block. The `.submit` and `.map` functions are examples of this; they return immediately in either case.

## 2.8 Efficiency

Parallel computing done well is responsive and rewarding. However, several speed-bumps can get in the way. This section describes common ways to ensure performance.

### 2.8.1 Leave data on the cluster

Wait as long as possible to gather data locally. If you want to ask a question of a large piece of data on the cluster it is often faster to submit a function onto that data then to bring the data down to your local computer.

For example if we have a numpy array on the cluster and we want to know its shape we might choose one of the following options:

1. **Slow:** Gather the numpy array to the local process, access the `.shape` attribute
2. **Fast:** Send a lambda function up to the cluster to compute the shape

```
>>> x = executor.submit(np.random.random, (1000, 1000))
>>> type(x)
Future
```

**Slow**

```
>>> x.result().shape()  # Slow from lots of data transfer
(1000, 1000)
```

**Fast**

```
>>> executor.submit(lambda a: a.shape, x).result()  # fast
(1000, 1000)
```

### 2.8.2 Use larger tasks

The scheduler adds about *one millisecond* of overhead per task or Future object. While this may sound fast it's quite slow if you run a billion tasks. If your functions run faster than 100ms or so then you might not see any speedup from using distributed computing.

A common solution is to batch your input into larger chunks.

**Slow**

```
>>> futures = executor.map(f, seq)
>>> len(futures)  # avoid large numbers of futures
1000000000
```

**Fast**

```
>>> def f_many(chunk):
...     return [f(x) for x in chunk]

>>> from toolz import partition_all
>>> chunks = partition_all(1000000, seq)  # Collect into groups of size 1000

>>> futures = executor.map(f_many, chunks)
>>> len(futures)  # Compute on larger pieces of your data at once
1000
```

### 2.8.3 Adjust between Threads and Processes

By default a single `Worker` runs many computations in parallel using as many threads as your compute node has cores. When using pure Python functions this may not be optimal and you may instead want to run several separate worker processes on each node, each using one thread. When configuring your cluster you may want to use the options to the `dworker` executable as follows:

```
$ dworker ip:port --nprocs 8 --nthreads 1
```

Note that if you're primarily using NumPy, Pandas, SciPy, Scikit Learn, Numba, or other C/Fortran/LLVM/Cython-accelerated libraries then this is not an issue for you. Your code is likely optimal for use with multi-threading.

### 2.8.4 Don't go distributed

Consider the dask and concurrent.futures modules, which have similar APIs to distributed but operate on a single machine. It may be that your problem performs well enough on a laptop or large workstation.

Consider accelerating your code through other means than parallelism. Better algorithms, data structures, storage formats, or just a little bit of C/Fortran/Numba code might be enough to give you the 10x speed boost that you're looking for. Parallelism and distributed computing are expensive ways to accelerate your application.

## 2.9 Data Locality

*Data movement often needlessly limits performance.*

This is especially true for analytic computations. `Distributed` minimizes data movement when possible and enables the user to take control when necessary. This document describes current scheduling policies and user API around data locality.

### 2.9.1 Current Policies

#### Task Submission

In the common case distributed runs tasks on workers that already hold dependent data. If you have a task `f(x)` that requires some data `x` then that task will very likely be run on the worker that already holds `x`.

If a task requires data split among multiple workers, then the scheduler chooses to run the task on the worker that requires the least data transfer to it. The size of each data element is measured by the workers using the `sys.getsizeof` function, which depends on the `__sizeof__` protocol generally available on most relevant Python objects.

#### Data Scatter

When a user scatters data from their local process to the distributed network this data is distributed in a round-robin fashion grouping by number of cores. So for example If we have two workers `Alice` and `Bob`, each with two cores and we scatter out the list `range(10)` as follows:

```
futures = e.scatter(range(10))
```

Then Alice and Bob receive the following data

- Alice: `[0, 1, 4, 5, 8, 9]`
- Bob: `[2, 3, 6, 7]`

### 2.9.2 User Control

Complex algorithms may require more user control.

For example the existence of specialized hardware such as GPUs or database connections may restrict the set of valid workers for a particular task.

In these cases use the `workers=` keyword argument to the `submit`, `map`, or `scatter` functions, providing a hostname, IP address, or alias as follows:

```
future = e.submit(func, *args, workers=['Alice'])
```

- Alice: `[0, 1, 4, 5, 8, 9, new_result]`
- Bob: `[2, 3, 6, 7]`

Required data will always be moved to these workers, even if the volume of that data is significant. If this restriction is only a preference and not a strict requirement, then add the `allow_other_workers` keyword argument to signal that in extreme cases such as when no valid worker is present, another may be used.

```
future = e.submit(func, *args, workers=['Alice'],
                  allow_other_workers=True)
```

Additionally the `scatter` function supports a `broadcast=` keyword argument to enforce that the all data is sent to all workers rather than round-robined. If new workers arrive they will not automatically receive this data.

```
futures = e.scatter([1, 2, 3], broadcast=True)  # send data to all workers
```

- Alice: `[1, 2, 3]`
- Bob: `[1, 2, 3]`

Valid arguments for `workers=` include the following:

- A single IP addresses, IP/Port pair, or hostname like the following:

```
192.168.1.100, 192.168.1.100:8989, alice, alice:8989
```

- A list or set of the above:

```
['alice'], ['192.168.1.100', '192.168.1.101:9999']
```

If only a hostname or IP is given then any worker on that machine will be considered valid. Additionally, you can provide aliases to workers upon creation.:

```
$ dworker scheduler_address:8786 --name worker_1
```

And then use this name when specifying workers instead.

```
e.map(func, sequence, workers='worker_1')
```

See the efficiency page to learn about best practices.

## 2.10 Joblib Frontend

Joblib is a library for simple parallel programming primarily developed and used by the Scikit Learn community. As of version 0.10 it contains a plugin mechanism to allow Joblib code to use other parallel frameworks to execute computations. The `distributed` scheduler implements such a plugin in the `distributed.joblib` module and registers it appropriately with Joblib. As a result, any joblib code (including many scikit-learn algorithms) will run on the distributed scheduler if you enclose it in a context manager as follows:

```
import distributed.joblib
from joblib import Parallel, parallel_backend

with parallel_backend('distributed', scheduler_host='HOST:PORT'):
    # normal Joblib code
```

For example you might distributed a randomized cross validated parameter search as follows:

```
import distributed.joblib
from joblib import Parallel, parallel_backend
from sklearn.datasets import load_digits
from sklearn.grid_search import RandomizedSearchCV
from sklearn.svm import SVC
import numpy as np

digits = load_digits()

param_space = {
    'C': np.logspace(-6, 6, 13),
    'gamma': np.logspace(-8, 8, 17),
    'tol': np.logspace(-4, -1, 4),
    'class_weight': [None, 'balanced'],
}

model = SVC(kernel='rbf')
search = RandomizedSearchCV(model, param_space, cv=3, n_iter=50, verbose=10)

with parallel_backend('distributed', scheduler_host='localhost:8786'):
    search.fit(digits.data, digits.target)
```

## 2.11 API

**Executor**

| | |
|---|---|
| *Executor*(address[, start, loop, timeout]) | Drive computations on a distributed cluster |
| *Executor.cancel*(futures) | Cancel running futures |
| *Executor.compute*(args[, sync]) | Compute dask collections on cluster |
| *Executor.gather*(futures[, errors]) | Gather futures from distributed memory |
| *Executor.get*(dsk, keys, **kwargs) | Compute dask graph |
| *Executor.has_what*([workers]) | Which keys are held by which workers |
| *Executor.map*(func, *iterables, **kwargs) | Map a function on a sequence of arguments |
| *Executor.ncores*([workers]) | The number of threads/cores available on each worker node |
| *Executor.persist*(collections) | Persist dask collections on cluster |
| *Executor.rebalance*([futures, workers]) | Rebalance data within network |
| *Executor.replicate*(futures[, n, workers, ...]) | Set replication of futures within network |
| *Executor.restart*() | Restart the distributed network |
| *Executor.run*(function, *args, **kwargs) | Run a function on all workers outside of task scheduling system |
| *Executor.scatter*(data[, workers, broadcast]) | Scatter data into distributed memory |
| *Executor.shutdown*([timeout]) | Send shutdown signal and wait until scheduler terminates |
| *Executor.submit*(func, *args, **kwargs) | Submit a function application to the scheduler |
| *Executor.upload_file*(filename) | Upload local package to workers |
| *Executor.who_has*([futures]) | The workers storing each future's data |

**Future**

| | |
|---|---|
| *Future*(key, executor) | A remotely running computation |
| *Future.cancel*() | Returns True if the future has been cancelled |
| *Future.cancelled*() | Returns True if the future has been cancelled |
| *Future.done*() | Is the computation complete? |
| *Future.exception*() | Return the exception of a failed task |
| *Future.result*() | Wait until computation completes. |
| *Future.traceback*() | Return the traceback of a failed task |

**Other**

| | |
|---|---|
| *as_completed*(fs) | Return futures in the order in which they complete |
| *distributed.diagnostics.progress*(*futures, ...) | Track progress of futures |
| *wait*(fs[, timeout, return_when]) | Wait until all futures are complete |

### 2.11.1 Executor

**class** distributed.executor.**Executor**(*address*, *start=True*, *loop=None*, *timeout=3*)

Drive computations on a distributed cluster

The Executor connects users to a distributed compute cluster. It provides an asynchronous user interface around functions and futures. This class resembles executors in `concurrent.futures` but also allows `Future` objects within `submit/map` calls.

**Parameters address: string, tuple, or ''Scheduler''**

This can be the address of a `Scheduler` server, either as a string `'127.0.0.1:8787'` or tuple `('127.0.0.1', 8787)` or it can be a local

`Scheduler` object.

**See also:**

***distributed.scheduler.Scheduler*** Internal scheduler

### Examples

Provide cluster's head node address on initialization:

```
>>> executor = Executor('127.0.0.1:8787')
```

Use `submit` method to send individual computations to the cluster

```
>>> a = executor.submit(add, 1, 2)
>>> b = executor.submit(add, 10, 20)
```

Continue using submit or map on results to build up larger computations

```
>>> c = executor.submit(add, a, b)
```

Gather results with the `gather` method.

```
>>> executor.gather([c])
33
```

**cancel** (*futures*)
Cancel running futures

This stops future tasks from being scheduled if they have not yet run and deletes them if they have already run. After calling, this result and all dependent results will no longer be accessible

**Parameters** **futures: list of Futures**

**compute** (*args*, *sync=False*)
Compute dask collections on cluster

**Parameters** **args: iterable of dask objects or single dask object**

Collections like dask.array or dataframe or dask.value objects

**sync: bool (optional)**

Returns Futures if False (default) or concrete values if True

**Returns** List of Futures if input is a sequence, or a single future otherwise

**See also:**

***Executor.get*** Normal synchronous dask.get function

### Examples

```
>>> from dask import do, value
>>> from operator import add
>>> x = dask.do(add)(1, 2)
>>> y = dask.do(add)(x, x)
>>> xx, yy = executor.compute([x, y])
>>> xx
<Future: status: finished, key: add-8f6e709446674bad78ea8aeecfee188e>
```

```
>>> xx.result()
3
>>> yy.result()
6
```

Also support single arguments

```
>>> xx = executor.compute(x)
```

**gather** (*futures*, *errors='raise'*)
Gather futures from distributed memory

Accepts a future, nested container of futures, iterator, or queue. The return type will match the input type.

**Returns** Future results

**See also:**

***Executor.scatter*** Send data out to cluster

#### Examples

```
>>> from operator import add
>>> e = Executor('127.0.0.1:8787')
>>> x = e.submit(add, 1, 2)
>>> e.gather(x)
3
>>> e.gather([x, [x], x])  # support lists and dicts
[3, [3], 3]
```

```
>>> seq = e.gather(iter([x, x]))  # support iterators
>>> next(seq)
3
```

**get** (*dsk*, *keys*, ***kwargs*)
Compute dask graph

**Parameters** **dsk: dict**

**keys: object, or nested lists of objects**

**restrictions: dict (optional)**

A mapping of {key: {set of worker hostnames}} that restricts where jobs can take place

**See also:**

***Executor.compute*** Compute asynchronous collections

#### Examples

```
>>> from operator import add
>>> e = Executor('127.0.0.1:8787')
>>> e.get({'x': (add, 1, 2)}, 'x')
3
```

**has_what** (*workers=None*)
Which keys are held by which workers

**Parameters** **workers: list (optional)**

A list of worker addresses, defaults to all

**See also:**

*Executor.who_has*, *Executor.ncores*

### Examples

```
>>> x, y, z = e.map(inc, [1, 2, 3])
>>> e.has_what()
{'192.168.1.141:46784': ['inc-1c8dd6be1c21646c71f76c16d09304ea',
                         'inc-fd65c238a7ea60f6a01bf4c8a5fcf44b',
                         'inc-1e297fc27658d7b67b3a758f16bcf47a']}
```

**map** (*func*, **iterables*, ***kwargs*)
Map a function on a sequence of arguments

Arguments can be normal objects or Futures

**Parameters** **func: callable**

**iterables: Iterables, Iterators, or Queues**

**pure: bool (defaults to True)**

Whether or not the function is pure. Set `pure=False` for impure functions like `np.random.random`.

**workers: set, iterable of sets**

A set of worker hostnames on which computations may be performed. Leave empty to default to all workers (common case)

**Returns** List, iterator, or Queue of futures, depending on the type of the

inputs.

**See also:**

***Executor.submit*** Submit a single function

### Examples

```
>>> L = executor.map(func, sequence)
```

**ncores** (*workers=None*)
The number of threads/cores available on each worker node

**Parameters** **workers: list (optional)**

A list of workers that we care about specifically. Leave empty to receive information about all workers.

**See also:**

*Executor.who_has*, *Executor.has_what*

#### Examples

```
>>> e.ncores()
{'192.168.1.141:46784': 8,
 '192.167.1.142:47548': 8,
 '192.167.1.143:47329': 8,
 '192.167.1.144:37297': 8}
```

**persist**(*collections*)

Persist dask collections on cluster

Starts computation of the collection on the cluster in the background. Provides a new dask collection that is semantically identical to the previous one, but now based off of futures currently in execution.

**Parameters** **collections: sequence or single dask object**

Collections like dask.array or dataframe or dask.value objects

**Returns** List of collections, or single collection, depending on type of input.

**See also:**

*Executor.compute*

#### Examples

```
>>> xx = executor.persist(x)
>>> xx, yy = executor.persist([x, y])
```

**rebalance**(*futures=None*, *workers=None*)

Rebalance data within network

Move data between workers to roughly balance memory burden. This either affects a subset of the keys/workers or the entire network, depending on keyword arguments.

This operation is generally not well tested against normal operation of the scheduler. It it not recommended to use it while waiting on computations.

**Parameters** **futures: list, optional**

A list of futures to balance, defaults all data

**workers: list, optional**

A list of workers on which to balance, defaults to all workers

**replicate**(*futures*, *n=None*, *workers=None*, *branching_factor=2*)

Set replication of futures within network

This performs a tree copy of the data throughout the network individually on each piece of data.

This operation blocks until complete. It does not guarantee replication of data to future workers.

**Parameters** **futures: list of futures**

Futures we wish to replicate

**n: int, optional**

Number of processes on the cluster on which to replicate the data. Defaults to all.

**workers: list of worker addresses**

Workers on which we want to restrict the replication. Defaults to all.

**branching_factor: int, optional**

The number of workers that can copy data in each generation

**See also:**

*Executor.rebalance*

#### Examples

```
>>> x = e.submit(func, *args)
>>> e.replicate([x])  # send to all workers
>>> e.replicate([x], n=3)  # send to three workers
>>> e.replicate([x], workers=['alice', 'bob'])  # send to specific
>>> e.replicate([x], n=1, workers=['alice', 'bob'])  # send to one of specific workers
>>> e.replicate([x], n=1)  # reduce replications
```

**restart**()

Restart the distributed network

This kills all active work, deletes all data on the network, and restarts the worker processes.

**run**(*function*, **args*, ***kwargs*)

Run a function on all workers outside of task scheduling system

This calls a function on all currently known workers immediately, blocks until those results come back, and returns the results asynchronously as a dictionary keyed by worker address. This method if generally used for side effects, such and collecting diagnostic information or installing libraries.

**Parameters function: callable**

***args: arguments for remote function**

****kwargs: keyword arguments for remote function**

**workers: list**

Workers on which to run the function. Defaults to all known workers.

#### Examples

```
>>> e.run(os.getpid)
{'192.168.0.100:9000': 1234,
 '192.168.0.101:9000': 4321,
 '192.168.0.102:9000': 5555}
```

Restrict computation to particular workers with the `workers=` keyword argument.

```
>>> e.run(os.getpid, workers=['192.168.0.100:9000',
...                           '192.168.0.101:9000'])
{'192.168.0.100:9000': 1234,
 '192.168.0.101:9000': 4321}
```

**scatter**(*data*, *workers=None*, *broadcast=False*)

Scatter data into distributed memory

**Parameters data: list, iterator, dict, or Queue**

Data to scatter out to workers. Output type matches input type.

**workers: list of tuples (optional)**

Optionally constrain locations of data. Specify workers as hostname/port pairs, e.g. `('127.0.0.1', 8787)`.

**broadcast: bool (defaults to False)**

Whether to send each data element to all workers. By default we round-robin based on number of cores.

**Returns** List, dict, iterator, or queue of futures matching the type of input.

**See also:**

***Executor.gather*** Gather data back to local process

### Examples

```
>>> e = Executor('127.0.0.1:8787')
>>> e.scatter([1, 2, 3])
[<Future: status: finished, key: c0a8a20f903a4915b94db8de3ea63195>,
 <Future: status: finished, key: 58e78e1b34eb49a68c65b54815d1b158>,
 <Future: status: finished, key: d3395e15f605bc35ab1bac6341a285e2>]
```

```
>>> e.scatter({'x': 1, 'y': 2, 'z': 3})
{'x': <Future: status: finished, key: x>,
 'y': <Future: status: finished, key: y>,
 'z': <Future: status: finished, key: z>}
```

Constrain location of data to subset of workers >>> e.scatter([1, 2, 3], workers=[('hostname', 8788)]) # doctest: +SKIP

Handle streaming sequences of data with iterators or queues >>> seq = e.scatter(iter([1, 2, 3])) # doctest: +SKIP >>> next(seq) # doctest: +SKIP <Future: status: finished, key: c0a8a20f903a4915b94db8de3ea63195>,

Broadcast data to all workers >>> [future] = e.scatter([element], broadcast=True) # doctest: +SKIP

**shutdown** (*timeout=10*)
Send shutdown signal and wait until scheduler terminates

**start** (***kwargs*)
Start scheduler running in separate thread

**submit** (*func*, **args*, ***kwargs*)
Submit a function application to the scheduler

**Parameters func: callable**

***args:**

****kwargs:**

**pure: bool (defaults to True)**

Whether or not the function is pure. Set `pure=False` for impure functions like `np.random.random`.

**workers: set, iterable of sets**

A set of worker hostnames on which computations may be performed. Leave empty to default to all workers (common case)

**Returns** Future

**See also:**

*Executor.map* Submit on many arguments at once

**Examples**

```
>>> c = executor.submit(add, a, b)
```

**upload_file** (*filename*)
Upload local package to workers

This sends a local file up to all worker nodes. This file is placed into a temporary directory on Python's system path so any .py or .egg files will be importable.

**Parameters filename: string**

Filename of .py or .egg file to send to workers

**Examples**

```
>>> executor.upload_file('mylibrary.egg')
>>> from mylibrary import myfunc
>>> L = e.map(myfunc, seq)
```

**who_has** (*futures=None*)
The workers storing each future's data

**Parameters futures: list (optional)**

A list of futures, defaults to all data

**See also:**

*Executor.has_what*, *Executor.ncores*

**Examples**

```
>>> x, y, z = e.map(inc, [1, 2, 3])
>>> e.who_has()
{'inc-1c8dd6be1c21646c71f76c16d09304ea': ['192.168.1.141:46784'],
 'inc-1e297fc27658d7b67b3a758f16bcf47a': ['192.168.1.141:46784'],
 'inc-fd65c238a7ea60f6a01bf4c8a5fcf44b': ['192.168.1.141:46784']}
```

```
>>> e.who_has([x, y])
{'inc-1c8dd6be1c21646c71f76c16d09304ea': ['192.168.1.141:46784'],
 'inc-1e297fc27658d7b67b3a758f16bcf47a': ['192.168.1.141:46784']}
```

### 2.11.2 CompatibleExecutor

**class** distributed.executor.**CompatibleExecutor** (*address*, *start=True*, *loop=None*, *timeout=3*)
A concurrent.futures-compatible Executor

A subclass of Executor that conforms to concurrent.futures API, allowing swapping in for other Executors.

**map** (*func*, **iterables*, ***kwargs*)
Map a function on a sequence of arguments

**Returns** iter_results: iterable

Iterable yielding results of the map.

**See also:**

*Executor.map* for more info

### 2.11.3 Future

**class** distributed.executor.**Future**(*key*, *executor*)

A remotely running computation

A Future is a local proxy to a result running on a remote worker. A user manages future objects in the local Python process to determine what happens in the larger cluster.

**See also:**

*Executor* Creates futures

#### Examples

Futures typically emerge from Executor computations

```
>>> my_future = executor.submit(add, 1, 2)
```

We can track the progress and results of a future

```
>>> my_future
<Future: status: finished, key: add-8f6e709446674bad78ea8aeecfee188e>
```

We can get the result or the exception and traceback from the future

```
>>> my_future.result()
```

**cancel**()

Returns True if the future has been cancelled

**cancelled**()

Returns True if the future has been cancelled

**done**()

Is the computation complete?

**exception**()

Return the exception of a failed task

**See also:**

*Future.traceback*

**result**()

Wait until computation completes. Gather result to local process

**traceback**()

Return the traceback of a failed task

This returns a traceback object. You can inspect this object using the `traceback` module. Alternatively if you call `future.result()` this traceback will accompany the raised exception.

**See also:**

*Future.exception*

**Examples**

```
>>> import traceback
>>> tb = future.traceback()
>>> traceback.export_tb(tb)
[...]
```

### 2.11.4 Other

distributed.executor.**as_completed**(*fs*)
Return futures in the order in which they complete

This returns an iterator that yields the input future objects in the order in which they complete. Calling `next` on the iterator will block until the next future completes, irrespective of order.

This function does not return futures in the order in which they are input.

distributed.diagnostics.**progress**(**futures*, ***kwargs*)
Track progress of futures

This operates differently in the notebook and the console

- Notebook: This returns immediately, leaving an IPython widget on screen
- Console: This blocks until the computation completes

**Parameters** **futures: Futures**

A list of futures or keys to track

**notebook: bool (optional)**

Running in the notebook or not (defaults to guess)

**multi: bool (optional)**

Track different functions independently (defaults to True)

**complete: bool (optional)**

Track all keys (True) or only keys that have not yet run (False) (defaults to True)

**Examples**

```
>>> progress(futures)
[########################################] | 100% Completed |  1.7s
```

distributed.executor.**wait**(*fs*, *timeout=None*, *return_when='ALL_COMPLETED'*)
Wait until all futures are complete

**Parameters** **fs: list of futures**

**Returns** Named tuple of completed, not completed

## 2.12 Foundations

You should read through the quickstart before reading this document.

Distributed computing is hard for two reasons:

1. Consistent coordination of distributed systems requires sophistication
2. Concurrent network programming is tricky and error prone

The foundations of `distributed` provide abstractions to hide some complexity of concurrent network programming (#2). These abstractions ease the construction of sophisticated parallel systems (#1) in a safer envirotnment. However, as with all layered abstractions, ours has flaws. Critical feedback is welcome.

### 2.12.1 Concurrency with Tornado Coroutines

Worker and Scheduler nodes operate concurrently. They serve several overlapping requests and perform several overlapping computations at the same time without blocking. There are several approaches for concurrent programming, we've chosen to use Tornado for the following reasons:

1. Developing and debugging is more comfortable without threads
2. Tornado's documentation is excellent
3. Stackoverflow coverage is excellent
4. Performance is satisfactory

### 2.12.2 Communication with Tornado Streams (raw sockets)

Workers, the Scheduler, and clients communicate with each other over the network. They use *raw sockets* as mediated by tornado streams. We separate messages by a sentinel value.

`distributed.core.`**`read`**(*stream*)
Read a message from a stream

`distributed.core.`**`write`**(*stream*, *msg*)
Write a message to a stream

### 2.12.3 Servers

Worker and Scheduler nodes serve requests over TCP. Both Worker and Scheduler objects inherit from a `Server` class. This Server class thinly wraps `tornado.tcpserver.TCPServer`. These servers expect requests of a particular form.

**class** `distributed.core.`**`Server`**(*handlers*, *max_buffer_size=2069891072.0*, ***kwargs*)
Distributed TCP Server

Superclass for both Worker and Center objects. Inherits from `tornado.tcpserver.TCPServer`, adding a protocol for RPC.

**Handlers**

Servers define operations with a `handlers` dict mapping operation names to functions. The first argument of a handler function must be a stream for the connection to the client. Other arguments will receive inputs from the keys of the incoming message which will always be a dictionary.

```
>>> def pingpong(stream):
...     return b'pong'
```

```
>>> def add(stream, x, y):
...     return x + y
```

```
>>> handlers = {'ping': pingpong, 'add': add}
>>> server = Server(handlers)
>>> server.listen(8000)
```

**Message Format**

The server expects messages to be dictionaries with a special key, *'op'* that corresponds to the name of the operation, and other key-value pairs as required by the function.

So in the example above the following would be good messages.

- `{'op': 'ping'}`
- `{'op': 'add': 'x': 10, 'y': 20}`

### 2.12.4 RPC

To interact with remote servers we typically use `rpc` objects.

**class** `distributed.core.`**rpc**(*arg=None*, *stream=None*, *ip=None*, *port=None*, *addr=None*, *timeout=3*)
Conveniently interact with a remote server

Normally we construct messages as dictionaries and send them with read/write

```
>>> stream = yield connect(ip, port)
>>> msg = {'op': 'add', 'x': 10, 'y': 20}
>>> yield write(stream, msg)
>>> response = yield read(stream)
```

To reduce verbosity we use an `rpc` object.

```
>>> remote = rpc(ip=ip, port=port)
>>> response = yield remote.add(x=10, y=20)
```

One rpc object can be reused for several interactions. Additionally, this object creates and destroys many streams as necessary and so is safe to use in multiple overlapping communications.

When done, close streams explicitly.

```
>>> remote.close_streams()
```

### 2.12.5 Example

Here is a small example using distributed.core to create and interact with a custom server.

#### Server Side

```
from tornado import gen
from tornado.ioloop import IOLoop
from distributed.core import write, Server
```

```
def add(stream, x=None, y=None):  # simple handler, just a function
    return x + y

@gen.coroutine
def stream_data(stream, interval=1):  # complex handler, multiple responses
    data = 0
    while True:
        yield gen.sleep(interval)
        data += 1
        yield write(stream, data)

s = Server({'add': add, 'stream': stream_data})
s.listen(8888)

IOLoop.current().start()
```

### Client Side

```
from tornado import gen
from tornado.ioloop import IOLoop
from distributed.core import connect, read, write

@gen.coroutine
def f():
    stream = yield connect('127.0.0.1', 8888)
    yield write(stream, {'op': 'add', 'x': 1, 'y': 2})
    result = yield read(stream)
    print(result)

>>> IOLoop().run_sync(f)
3

@gen.coroutine
def g():
    stream = yield connect('127.0.0.1', 8888)
    yield write(stream, {'op': 'stream', 'interval': 1})
    while True:
        result = yield read(stream)
        print(result)

>>> IOLoop().run_sync(g)
1
2
3
...
```

### Client Side with rpc

RPC provides a more pythonic interface. It also provides other benefits, such as using multiple streams in concurrent cases. Most distributed code uses rpc. The exception is when we need to perform multiple reads or writes, as with the stream data case above.

```
from tornado import gen
from tornado.ioloop import IOLoop
from distributed.core import rpc
```

```
@gen.coroutine
def f():
    # stream = yield connect('127.0.0.1', 8888)
    # yield write(stream, {'op': 'add', 'x': 1, 'y': 2})
    # result = yield read(stream)
    r = rpc(ip='127.0.0.1', 8888)
    result = yield r.add(x=1, y=2)

    print(result)

>>> IOLoop().run_sync(f)
3
```

### 2.12.6 Everything is a Server

Workers, Scheduler, and Nanny objects all inherit from Server. Each maintains separate state and serves separate functions but all communicate in the way shown above. They talk to each other by opening connections, writing messages that trigger remote functions, and then collect the results with read.

## 2.13 Client Interaction

As discussed in the quickstart users can interact with the worker-center network with the Executor abstraction.

This is built with lower level functions described below.

### 2.13.1 Scatter/Gather

Users rarely create RemoteData objects by hand. They are created by other client libraries or functions like `gather` and `scatter`.

`distributed.client.`**`scatter`**(*center*, *data*, *serialize=True*)

Scatter data to workers

**Parameters center:**

(ip, port) tuple or Stream, or rpc object designating the Center

**data: dict or iterable**

either a dictionary of key: value pairs or an iterable of values

**key:**

if data is an iterable of values then we use the key to generate keys as key-0, key-1, key-2, ...

**See also:**

*distributed.client.gather*, distributed.client._scatter, distributed.client.scatter_to_workers

#### Examples

```
>>> remote_data = scatter('127.0.0.1:8787', [1, 2, 3])
>>> local_data = gather('127.0.0.1:8787', remote_data)
```

distributed.client.**gather**(*center*, *needed*)
Gather data from distributed workers

This operates by first asking the center who has all of the state keys and then trying those workers directly.

Keys not found on the network will not appear in the output. You should check the length of the output against the input if concerned about missing data.

**Parameters** **center:**

(ip, port) tuple or Stream, or rpc object designating the Center

**needed: iterable**

A list of required keys

**Returns** result: dict

A mapping of the given keys to data values

**See also:**

*distributed.client.scatter*, distributed.client._gather, distributed.client.gather_from_workers

#### Examples

```
>>> remote_data = scatter('127.0.0.1:8787', [1, 2, 3])
>>> local_data = gather('127.0.0.1:8787', remote_data)
```

distributed.client.**delete**(*center*, *keys*)
Delete keys from all workers

distributed.client.**clear**(*center*)
Clear all data from all workers' memory

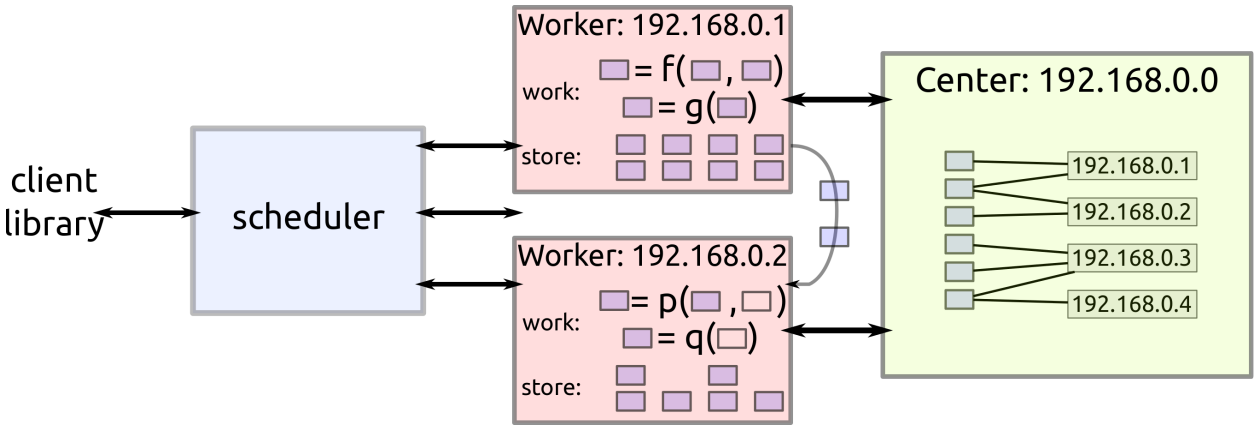
**See also:**

*distributed.client.delete*

## 2.14 Worker

We build a distributed network from two kinds of nodes.

- A single scheduler node
- Several Worker nodes




This page describes the worker nodes.

### 2.14.1 Serve Data

Workers serve data from a local dictionary of data:

```
{'x': np.array(...),
 'y': pd.DataFrame(...)}
```

Operations include normal dictionary operations, like get, set, and delete key-value pairs. In the following example we connect to two workers, collect data from one worker and send it to another.

```
alice = rpc(ip='192.168.0.101', port=8788)
d = yield alice.get_data(keys=['x', 'y'])

bob = rpc(ip='192.168.0.102', port=8788)
yield alice.update_data(data=d)
```

However, this is only an example, typically one does not manually manage data transfer between workers. They handle that as necessary on their own.

### 2.14.2 Compute

Workers evaluate functions provided by the user on their data. They evaluate functions either on their data or can automatically collect data from peers (as shown above) if they don't have the necessary data but their peers do:

```
z <- add(x, y)  # can be done with only local data
z <- add(x, a)  # need to find out where we can get 'a'
```

The result of such a computation on our end is just a response `b'OK'`. The actual result stays on the remote worker.

```
>>> response, metadata = yield alice.compute(function=add, keys=['x', 'a'])
>>> response
b'OK'
>>> metadata
{'nbytes': 1024}
```

The worker also reports back to the center/scheduler whenever it completes a computation. Metadata storage is centralized but all data transfer is peer-to-peer. Here is a quick example of what happens during a call to `compute`:

```
client:  Hey Alice!   Compute ``z <- add(x, a)``

Alice:   Hey Center!  Who has a?
Center:  Hey Alice!   Bob has a.
Alice:   Hey Bob!     Send me a!
Bob:     Hey Alice!   Here's a!

Alice:   Hey Client!  I've computed z and am holding on to it!
Alice:   Hey Center!  I have z!
```

**class** distributed.worker.**Worker**(*center_ip*, *center_port*, *ip=None*, *ncores=None*, *loop=None*, *local_dir=None*, *services=None*, *service_ports=None*, *name=None*, ***kwargs*)

Worker Node

Workers perform two functions:

1. **Serve data** from a local dictionary
2. **Perform computation** on that data and on data from peers

Additionally workers keep a Center informed of their data and use that Center to gather data from other workers when necessary to perform a computation.

You can start a worker with the `dworker` command line application:

```
$ dworker scheduler-ip:port
```

**State**

- **data: {key: object}:** Dictionary mapping keys to actual values
- **active: {key}:** Set of keys currently under computation
- **ncores: int:** Number of cores used by this worker process
- **executor: concurrent.futures.ThreadPoolExecutor:** Executor used to perform computation
- **local_dir: path:** Path on local machine to store temporary files
- **center: rpc:** Location of center or scheduler. See `.ip/.port` attributes.
- **name: string:** Alias
- **services: {str: Server}:** Auxiliary web servers running on this worker
- **service_ports:** {str: port}:

**See also:**

`distributed.center.Center`

#### Examples

Create centers and workers in Python:

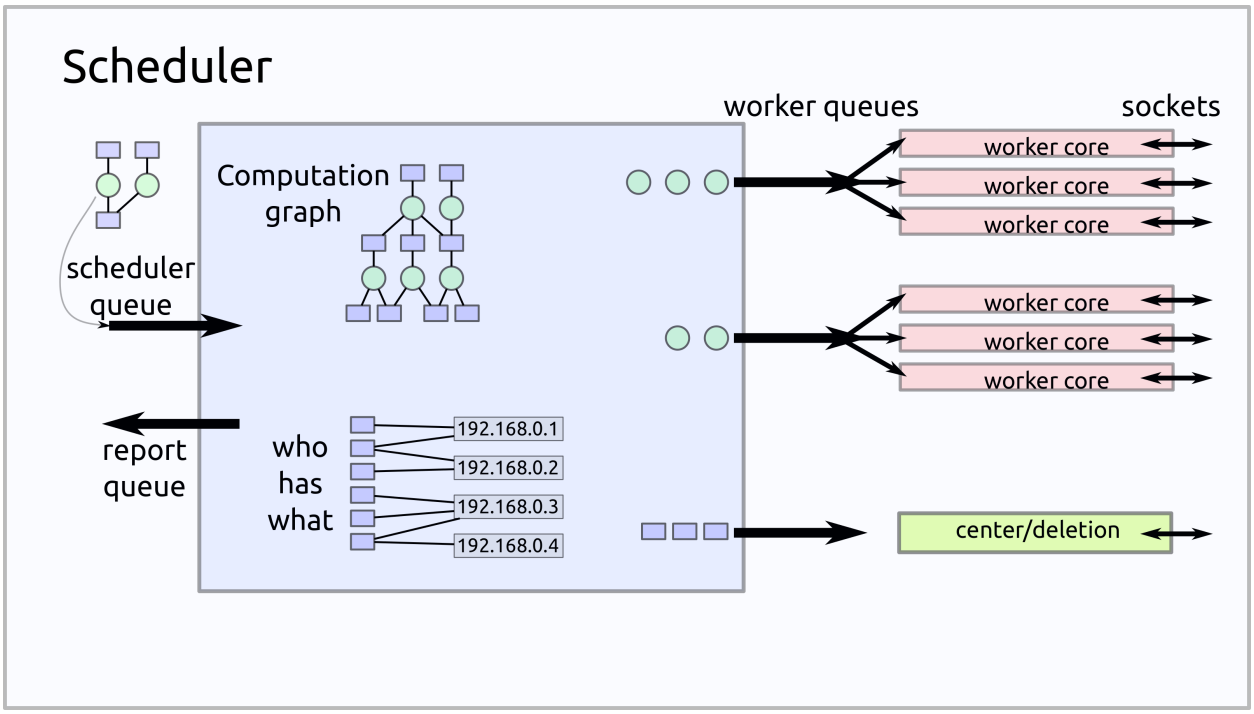
```
>>> from distributed import Center, Worker
>>> c = Center('192.168.0.100', 8787)
>>> w = Worker(c.ip, c.port)
>>> yield w._start(port=8788)
```

Or use the command line:

```
$ dcenter
Start center at 127.0.0.1:8787

$ dworker 127.0.0.1:8787
Start worker at:            127.0.0.1:8788
Registered with center at:  127.0.0.1:8787
```

## 2.15 Scheduler




The scheduler orchestrates which workers work on which tasks in what order. It tracks the current state of the entire cluster. It consists of several coroutines rounning in a single event loop.

**class** `distributed.scheduler.`**`Scheduler`**(*center=None*, *loop=None*, *resource_interval=1*, *resource_log_size=1000*, *max_buffer_size=2069891072.0*, *delete_interval=500*, *ip=None*, *services=None*, *heartbeat_interval=500*, ***kwargs*)

Dynamic distributed task scheduler

The scheduler tracks the current state of workers, data, and computations. The scheduler listens for events and responds by controlling workers appropriately. It continuously tries to use the workers to execute an ever growing dask graph.

All events are handled quickly, in linear time with respect to their input (which is often of constant size) and generally within a millisecond. To accomplish this the scheduler tracks a lot of state. Every operation maintains the consistency of this state.

The scheduler communicates with the outside world either by adding pairs of in/out queues or by responding to a new IOStream (the Scheduler can operate as a typical distributed `Server`). It maintains a consistent and valid view of the world even when listening to several clients at once.

A Scheduler is typically started either with the `dscheduler` executable:

```
$ dscheduler 127.0.0.1:8787  # address of center
```

Or as part of when an Executor starts up and connects to a Center:

```
>>> e = Executor('127.0.0.1:8787')
>>> e.scheduler
Scheduler(...)
```

Users typically do not interact with the scheduler except through Plugins. See http://distributed.readthedocs.org/en/latest/plugins.html

**State**

- **tasks: {key: task}:** Dictionary mapping key to task, either dask task, or serialized dict like: `{'function': b'xxx', 'args': b'xxx'}` or `{'task': b'xxx'}`
- **dependencies: {key: {key}}:** Dictionary showing which keys depend on which others
- **dependents: {key: {key}}:** Dictionary showing which keys are dependent on which others
- **waiting: {key: {key}}:** Dictionary like dependencies but excludes keys already computed
- **waiting_data: {key: {key}}:** Dictionary like dependents but excludes keys already computed
- **ready: deque(key)** Keys that are ready to run, but not yet assigned to a worker
- **ncores: {worker: int}:** Number of cores owned by each worker
- **idle: {worker}:** Set of workers that are not fully utilized
- **services: {str: port}:** Other services running on this scheduler, like HTTP
- **worker_info: {worker: {str: data}}:** Information about each worker
- **host_info: {hostname: dict}:** Information about each worker host
- **who_has: {key: {worker}}:** Where each key lives. The current state of distributed memory.
- **has_what: {worker: {key}}:** What worker has what keys. The transpose of who_has.
- **who_wants: {key: {client}}:** Which clients want each key. The active targets of computation.
- **wants_what: {client: {key}}:** What keys are wanted by each client.. The transpose of who_wants.
- **nbytes: {key: int}:** Number of bytes for a key as reported by workers holding that key.
- **processing: {worker: {keys}}:** Set of keys currently in execution on each worker
- **stacks: {worker: [keys]}:** List of keys waiting to be sent to each worker
- **released: {keys}** Set of keys that are known, but released from memory
- **unrunnable: {key}** Keys that we are unable to run
- **retrictions: {key: {hostnames}}:** A set of hostnames per key of where that key can be run. Usually this is empty unless a key has been specifically restricted to only run on certain hosts. These restrictions don't include a worker port. Any worker on that hostname is deemed valid.
- **loose_retrictions: {key}:** Set of keys for which we are allow to violate restrictions (see above) if not valid workers are present.
- **keyorder: {key: tuple}:** A score per key that determines its priority
- **scheduler_queues: [Queues]:** A list of Tornado Queues from which we accept stimuli
- **report_queues: [Queues]:** A list of Tornado Queues on which we report results

- **streams: [IOStreams]:** A list of Tornado IOStreams from which we both accept stimuli and report results
- **coroutines: [Futures]:** A list of active futures that control operation
- **exceptions: {key: Exception}:** A dict mapping keys to remote exceptions
- **tracebacks: {key: list}:** A dict mapping keys to remote tracebacks stored as a list of strings
- **exceptions_blame: {key: key}:** A dict mapping a key to another key on which it depends that has failed
- **deleted_keys: {key: {workers}}** Locations of workers that have keys that should be deleted
- **loop: IOLoop:** The running Torando IOLoop

**add_client**(*stream*, *client=None*)
Listen to messages from an IOStream

**add_plugin**(*plugin*)
Add external plugin to scheduler

See http://http://distributed.readthedocs.org/en/latest/plugins.html

**broadcast**(*stream=None*, *msg=None*, *workers=None*)
Broadcast message to workers, return all results

**cancel**(*stream*, *keys=None*, *client=None*)
Stop execution on a list of keys

**cleanup**()
Clean up queues and coroutines, prepare to stop

**clear_data_from_workers**()
This is intended to be run periodically,

The `self._delete_periodic_callback` attribute holds a PeriodicCallback that runs this every `self.delete_interval` milliseconds‘‘.

**close**(*stream=None*)
Send cleanup signal to all coroutines then wait until finished

**See also:**

*Scheduler.cleanup*

**coerce_address**(*addr*)
Coerce possible input addresses to canonical form

Handles lists, strings, bytes, tuples, or aliases

**correct_time_delay**(*worker*, *msg*)
Apply offset time delay in message times

Operates in place

**ensure_idle_ready**()
Run ready tasks on idle workers

**Work stealing policy**

If some workers are idle but not others, if there are no globally ready tasks, and if there are tasks in worker stacks then we start to pull preferred tasks from overburdened workers and deploy them back into the global pool in the following manner.

We determine the number of tasks to reclaim as the number of all tasks in all stacks times the fraction of idle workers to all workers. We sort the stacks by size and walk through them, reclaiming half of each stack until we have enough task to fill the global pool. We are careful not to reclaim tasks that are restricted to run on certain workers.

**ensure_in_play**(*key*)
Ensure that a key is on track to enter memory in the future

This will only act on keys currently in self.released.

**ensure_occupied**(*worker*)
Send tasks to worker while it has tasks and free cores

These tasks may come from the worker's own stacks or from the global ready deque.

We update the idle workers set appropriately.

**finished**()
Wait until all coroutines have ceased

**forget**(*key*)
Forget a key if no one cares about it

This removes all knowledge of how to produce a key from the scheduler. This is almost exclusively called by release_held_data

**gather**(*stream=None*, *keys=None*)
Collect data in from workers

**handle_messages**(*in_queue*, *report*, *client=None*)
Master coroutine. Handles inbound messages.

This runs once per Queue or Stream.

**handle_queues**(*scheduler_queue*, *report_queue*)
Register new control and report queues to the Scheduler

**identity**(*stream*)
Basic information about ourselves and our cluster

**log_state**(*msg=''*)
Log current full state of the scheduler

**mark_failed**(*key*, *failing_key=None*)
When a task fails mark it and all dependent task as failed

**mark_key_in_memory**(*key*, *workers=None*, *type=None*)
Mark that a key now lives in distributed memory

**mark_missing_data**(*keys=None*, *key=None*, *worker=None*, ***kwargs*)
Mark that certain keys have gone missing. Recover.

**See also:**

*recover_missing*

**mark_ready_to_run**(*key*)
Mark a task as ready to run.

If the task should be assigned to a worker then make that determination and assign appropriately. Otherwise place task in the ready queue.

Trigger appropriate workers if idle.

**See also:**

*decide_worker*, *Scheduler.ensure_occupied*

**mark_task_erred**(*key=None*, *worker=None*, *exception=None*, *traceback=None*, ***kwargs*)
Mark that a task has erred on a particular worker

**See also:**

*Scheduler.mark_failed*

**mark_task_finished**(*key=None*, *worker=None*, *nbytes=None*, *type=None*, ***kwargs*)
Mark that a task has finished execution on a particular worker

**put**(*msg*)
Place a message into the scheduler's queue

**recover_missing**(*key*)
Recover a recently lost piece of data

This assumes that we've already removed this key from who_has/has_what.

**release_held_data**(*keys=None*)
Mark that a key is no longer externally required to be in memory

**release_live_dependencies**(*key*)
We no longer need to keep data in memory to compute this

This occurs after we've computed it or after we've forgotten it

**remove_worker**(*stream=None*, *address=None*)
Mark that a worker no longer seems responsive

**See also:**

*Scheduler.recover_missing*

**replicate**(*stream=None*, *keys=None*, *n=None*, *workers=None*, *branching_factor=2*)
Replicate data throughout cluster

This performs a tree copy of the data throughout the network individually on each piece of data.

**Parameters keys: Iterable**

list of keys to replicate

**n: int**

Number of replications we expect to see within the cluster

**branching_factor: int, optional**

The number of workers that can copy data in each generation

**See also:**

Scheduler.rebalance

**report**(*msg*)
Publish updates to all listening Queues and Streams

**restart**()
Restart all workers. Reset local state

**rpc**(*arg=None*, *ip=None*, *port=None*, *addr=None*)
Cached rpc objects

**scatter**(*stream=None*, *data=None*, *workers=None*, *client=None*, *broadcast=False*)
Send data out to workers

**start** (*port=8786*, *start_queues=True*)
Clear out old state and restart all running coroutines

**update_data** (*who_has=None*, *nbytes=None*, *client=None*)
Learn that new data has entered the network from an external source

**See also:**

*Scheduler.mark_key_in_memory*

**update_graph** (*client=None*, *tasks=None*, *keys=None*, *dependencies=None*, *restrictions=None*, *loose_restrictions=None*)
Add new computations to the internal dask graph

This happens whenever the Executor calls submit, map, get, or compute.

**workers_list** (*workers*)
List of qualifying workers

Takes a list of worker addresses or hostnames. Returns a list of all worker addresses that match

distributed.scheduler.**decide_worker** (*dependencies*, *stacks*, *who_has*, *restrictions*, *loose_restrictions*, *nbytes*, *key*)
Decide which worker should take task

```
>>> dependencies = {'c': {'b'}, 'b': {'a'}}
>>> stacks = {'alice:8000': ['z'], 'bob:8000': []}
>>> who_has = {'a': {'alice:8000'}}
>>> nbytes = {'a': 100}
>>> restrictions = {}
>>> loose_restrictions = set()
```

We choose the worker that has the data on which 'b' depends (alice has 'a')

```
>>> decide_worker(dependencies, stacks, who_has, restrictions,
...               loose_restrictions, nbytes, 'b')
'alice:8000'
```

If both Alice and Bob have dependencies then we choose the less-busy worker

```
>>> who_has = {'a': {'alice:8000', 'bob:8000'}}
>>> decide_worker(dependencies, stacks, who_has, restrictions,
...               loose_restrictions, nbytes, 'b')
'bob:8000'
```

Optionally provide restrictions of where jobs are allowed to occur

```
>>> restrictions = {'b': {'alice', 'charile'}}
>>> decide_worker(dependencies, stacks, who_has, restrictions,
...               loose_restrictions, nbytes, 'b')
'alice:8000'
```

If the task requires data communication, then we choose to minimize the number of bytes sent between workers. This takes precedence over worker occupancy.

```
>>> dependencies = {'c': {'a', 'b'}}
>>> who_has = {'a': {'alice:8000'}, 'b': {'bob:8000'}}
>>> nbytes = {'a': 1, 'b': 1000}
>>> stacks = {'alice:8000': [], 'bob:8000': []}
```

```
>>> decide_worker(dependencies, stacks, who_has, {}, set(), nbytes, 'c')
'bob:8000'
```

## 2.16 Resilience

Software fails, Hardware fails, network connections fail, user code fails. This document describes how `distributed` responds in the face of these failures and other known bugs.

### 2.16.1 User code failures

When a function raises an error that error is kept and transmitted to the executor on request. Any attempt to gather that result *or any dependent result* will raise that exception.

```
>>> def div(a, b):
...     return a / b

>>> x = executor.submit(div, 1, 0)
>>> x.result()
ZeroDivisionError: division by zero

>>> y = executor.submit(add, x, 10)
>>> y.result()  # same error as above
ZeroDivisionError: division by zero
```

This does not affect the smooth operation of the scheduler or worker in any way.

### 2.16.2 Closed Network Connections

If the connection to a remote worker unexpectedly closes and the local process appropriately raises an `IOError` then the scheduler will reroute all pending computations to other workers.

If the lost worker was the only worker to hold vital results necessary for future computations then those results will be recomputed by surviving workers. The scheduler maintains a full history of how each result was produced and so is able to reproduce those same computations on other workers.

This has some fail cases.

1. If results depend on impure functions then you may get a different (although still entirely accurate) result
2. If the worker failed due to a bad function, for example a function that causes a segmentation fault, then that bad function will repeatedly be called on other workers, and proceed to kill the distributed system, one worker at a time.
3. Data ‘‘scatter‘‘ed out to the workers is not kept in the scheduler (it is often quite large) and so the loss of this data is irreparable.

### 2.16.3 Hardware Failures

It is not clear under which circumstances the local process will know that the remote worker has closed the connection. If the socket does not close cleanly then the system will wait for a timeout, roughly three seconds, before marking the worker as failed and resuming smooth operation.

### 2.16.4 Scheduler Failure

The process containing the scheduler might die. There is currently no persistence mechanism to record and recover the scheduler state. The data will remain on the cluster until cleared.

### 2.16.5 Restart and Nanny Processes

The executor provides a mechanism to restart all of the workers in the cluster. This is convenient if, during the course of experimentation, you find your workers in an inconvenient state that makes them unresponsive. The `Executor.restart` method does the following process:

1. Sends a soft shutdown signal to all of the coroutines watching workers
2. Sends a hard kill signal to each worker's Nanny process, which oversees that worker. This Nanny process terminates the worker process ungracefully and unregisters that worker from the Scheduler.
3. Clears out all scheduler state and sets all Future's status to `'cancelled'`
4. Sends a restart signal to all Nanny processes, which in turn restart clean Worker processes and register these workers with the Scheduler. New workers may not have the same port as their previous iterations. The `.nannies` dictionary on the Executor serves as an accurate set of aliases if necessary.
5. Restarts the scheduler, with clean and empty state

This effectively removes all data and clears out all computations from the scheduler. Any data or computations not saved to persistent storage are lost. This process is very robust to a number of failure modes, including non-responsive or swamped workers but not including full hardware failures.

Currently the user may experience a few error logging messages from Tornado upon closing their session. These can safely be ignored.

## 2.17 Journey of a Task

We follow a single task through the user interface, scheduler, worker nodes, and back. Hopefully this helps to illustrate the inner workings of the system.

### 2.17.1 User code

A user computes the addition of two variables already on the cluster, then pulls the result back to the local process.

```
e = Executor('host:port')
x = e.submit(...)
y = e.submit(...)

z = e.submit(add, x, y)  # we follow z

print(z.result())
```

### 2.17.2 Step 1: Executor

`z` begins its life when the `Executor.submit` function sends the following message to the `Scheduler`:

```
{'op': 'update-graph',
 'tasks': {'z': (add, x, y)},
 'keys': ['z']}
```

The executor then creates a `Future` object with the key `'z'` and returns that object back to the user. This happens even before the message has been received by the scheduler. The status of the future says `'pending'`.

### 2.17.3 Step 2: Arrive in the Scheduler

A few milliseconds later, the scheduler receives this message on an open socket.

The scheduler updates its state with this little graph that shows how to compute `z`.:

```
scheduler.tasks.update[msg['tasks']]
```

The scheduler also updates *a lot* of other state. Notably, it has to identify that `x` and `y` are themselves variables, and connect all of those dependencies. This is a long and detail oriented process that involves updating roughly 10 sets and dictionaries. Interested readers should investigate `distributed/scheduler.py::update_state()`. While this is fairly complex and tedious to describe rest assured that it all happens in constant time and in about a millisecond.

### 2.17.4 Step 3: Select a Worker

Once the latter of `x` and `y` finishes, the scheduler notices that all of `z`'s dependencies are in memory and that `z` itself may now run. Which worker should `z` select? We consider a sequence of criteria:

1. First, we quickly downselect to only those workers that have either `x` or `y` in local memory.
2. Then, we select the worker that would have to gather the least number of bytes in order to get both `x` and `y` locally. E.g. if two different workers have `x` and `y` and if `y` takes up more bytes than `x` then we select the machine that holds `y` so that we don't have to communicate as much.
3. If there are multiple workers that require the minimum number of communication bytes then we select the worker that is the least busy

`z` considers the workers and chooses one based on the above criteria. In the common case the choice is pretty obvious after step 1. `z` waits on a stack associated with the chosen worker. The worker may still be busy though, so `z` may wait a while.

*Note: This policy is under flux and this part of this document is quite possibly out of date.*

### 2.17.5 Step 4: Transmit to the Worker

Eventually the worker finishes a task, has a spare core, and `z` finds itself at the top of the stack (note, that this may be some time after the last section if other tasks placed themselves on top of the worker's stack in the meantime.)

We place `z` into a `worker_queue` associated with that worker and a `worker_core` coroutine pulls it out. `z`'s function, the keys associated to its arguments, and the locations of workers that hold those keys are packed up into a message that looks like this:

```
{'op': 'compute',
 'function': execute_task,
 'args': ((add, 'x', 'y'),),
 'who_has': {'x': {(worker_host, port)},
             'y': {(worker_host, port), (worker_host, port)}},
 'key': 'z'}
```

This message is serialized and sent across a TCP socket to the worker.

### 2.17.6 Step 5: Execute on the Worker

The worker unpacks the message, and notices that it needs to have both `x` and `y`. If the worker does not already have both of these then it gathers them from the workers listed in the `who_has` dictionary also in the message. For each key that it doesn't have it selects a valid worker from `who_has` at random and gathers data from it.

After this exchange, the worker has both the value for `x` and the value for `z`. So it launches the computation `add(x, y)` in a local `ThreadPoolExecutor` and waits on the result.

*In the mean time the worker repeats this process concurrently for other tasks. Nothing blocks.*

Eventually the computation completes. The Worker stores this result in its local memory:

```
data['x'] = ...
```

And transmits back a success, and the number of bytes of the result:

```
Worker: Hey Scheduler, 'z' worked great.
        I'm holding onto it.
        It takes up 64 bytes.
```

The worker does not transmit back the actual value for `z`.

### 2.17.7 Step 6: Scheduler Aftermath

The scheduler receives this message and does a few things:

1. It notes that the worker has a free core, and sends up another task if available
2. If `x` or `y` are no longer needed then it sends a message out to relevant workers to delete them from local memory.
3. It sends a message to all of the clients that `z` is ready and so all client `Future` objects that are currently waiting should, wake up. In particular, this wakes up the `z.result()` command executed by the user originally.

### 2.17.8 Step 7: Gather

When the user calls `z.result()` they wait both on the completion of the computation and for the computation to be sent back over the wire to the local process. Usually this isn't necessary, usually you don't want to move data back to the local process but instead want to keep in on the cluster.

But perhaps the user really wanted to actually know this value, so they called `z.result()`.

The scheduler checks who has `z` and sends them a message asking for the result. This message doesn't wait in a queue or for other jobs to complete, it starts instantly. The value gets serialized, sent over TCP, and then deserialized and returned to the user (passing through a queue or two on the way.)

### 2.17.9 Step 8: Garbage Collection

The user leaves this part of their code and the local variable `z` goes out of scope. The Python garbage collector cleans it up. This triggers a decremented reference on the executor (we didn't mention this, but when we created the `Future` we also started a reference count.) If this is the only instance of a Future pointing to `z` then we send a message up to the scheduler that it is OK to release `z`. The user no longer requires it to persist.

The scheduler receives this message and, if there are no computations that might depend on `z` in the immediate future, it removes elements of this key from local scheduler state and adds the key to a list of keys to be deleted periodically. Every 500 ms a message goes out to relevant workers telling them which keys they can delete from their local memory. The graph/recipe to create the result of `z` persists in the scheduler for all time.

### 2.17.10 Overhead

The user experiences this in about 10 milliseconds, depending on network latency.

## 2.18 Protocol

The scheduler, workers, and clients pass messages between each other. Semantically these messages encode commands, status updates, and data, like the following:

- Please compute the function `sum` on the data `x` and store in `y`
- The computation `y` has been completed
- Be advised that a new worker named `alice` is available for use
- Here is the data for the keys `'x'`, and `'y'`

In practice we represent these messages with dictionaries/mappings:

```
{'op': 'compute',
 'function': ...
 'args': ['x']}

{'op': 'task-complete',
 'key': 'y',
 'nbytes': 26}

{'op': 'register-worker',
 'address': '192.168.1.42',
 'name': 'alice',
 'ncores': 4}

{'x': b'...',
 'y': b'...'}
```

When we communicate these messages between nodes we need to serialize these messages down to a string of bytes that can then be deserialized on the other end to their in-memory dictionary form. For simple cases several options exist like JSON, MsgPack, Protobuffers, and Thrift. The situation is made more complex by concerns like serializing Python functions and Python objects, optional compression, cross-language support, large messages, and efficiency.

This document describes the protocol used by `dask.distributed` today. Be advised that this protocol changes rapidly as we continue to optimize for performance.

### 2.18.1 Overview

We may split a single message into multiple message-part to suit different protocols. Generally small bits of data are encoded with MsgPack while large bytestrings are handled specially by a custom format. Each message-part gets its own header, which is always encoded as msgpack. After serializing all message parts we have a sequence of bytestrings or *frames* which we send along the wire, prepended with length information.

The application doesn't know any of this, it just sends us Python dictionaries with various datatypes and we produce a list of bytestrings that get written to a socket. This format is fast both for many frequent messages and for large messages.

### 2.18.2 MsgPack for Messages

Most messages are encoded with MsgPack, a self describing semi-structured serialization format that is very similar to JSON, but smaller, faster, not human-readable, and supporting of bytestrings and (soon) timestamps. We chose MsgPack as a base serialization format for the following reasons:

- It does not require separate headers, and so is easy and flexible to use which is particularly important in an early stage project like `dask.distributed`
- It is very fast, much faster than JSON, and there are nicely optimized implementations, particularly within the `pandas.msgpack` module. With few exceptions (described later) MsgPack does not come anywhere near being a bottleneck, even under heavy use.
- Unlike JSON it supports bytestrings
- It covers the standard set of types necessary to encode most information
- It is widely implemented in a number of languages (see cross language section below)

However, MsgPack fails (correctly) in the following ways:

- It does not provide any way for us to encode Python functions or user defined data types
- It does not support bytestrings greater than 4GB and is generally inefficient for very large messages.

Because of these failings we supplement it with a language-specific protocol and a special case for large bytestrings.

### 2.18.3 CloudPickle for Functions and Data

Pickle and CloudPickle are Python libraries to serialize almost any Python object, including functions. We use these libraries to transform the users' functions and data into bytes before we include them in the dictionary/map that we pass off to msgpack. In the introductory example you may have noticed that we skipped providing an example for the function argument:

```
{'op': 'compute',
 'function': ...
 'args': ['x']}
```

That is because this value `...` will actually be the result of calling `cloudpickle.dumps(myfunction)`. Those bytes will then be included in the dictionary that we send off to msgpack, which will only have to deal with bytes rather than obscure Python functions.

### 2.18.4 Cross Language Specialization

The Client and Workers must agree on a language-specific serialization format. In the standard `dask.distributed` client and worker objects this ends up being the following:

```
bytes = cloudpickle.dumps(obj, protocol=pickle.HIGHEST_PROTOCOL)
obj = cloudpickle.loads(bytes)
```

This varies between Python 2 and 3 and so your client and workers must match their Python versions and software environments.

However, the Scheduler never uses the language-specific serialization and instead only deals with MsgPack. If the client sends a pickled function up to the scheduler the scheduler will not unpack function but will instead keep it as bytes. Eventually those bytes will be sent to a worker, which will then unpack the bytes into a proper Python function. Because the Scheduler never unpacks language-specific serialized bytes it may be in a different language.

**The client and workers must share the same language and software environment, the scheduler may differ.**

This has a few advantages:

1. The Scheduler is protected from unpickling unsafe code
2. The Scheduler can be run under `pypy` for improved performance. This is only useful for larger clusters.

3. We could conceivably implement workers and clients for other languages (like R or Julia) and reuse the Python scheduler. The worker and client code is fairly simple and much easier to reimplement than the scheduler, which is complex.
4. The scheduler might some day be rewritten in more heavily optimized C or Go

### 2.18.5 Compression

Fast compression libraries like LZ4 or Snappy may increase effective bandwidth by compressing data before sending and decompressing it after reception. This is especially valuable on lower-bandwidth networks.

If either of these libraries is available (we prefer LZ4 to Snappy) then for every message greater than 1kB we try to compress the message and, if the compression is at least a 10% improvement, we send the compressed bytes rather than the original payload. We record the compression used within the header as a string like `'lz4'` or `'snappy'`.

To avoid compressing large amounts of uncompressable data we first try to compress a sample. We take 10kB chunks from five locations in the dataset, arrange them together, and try compressing the result. If this doesn't result in significant compression then we don't try to compress the full result.

### 2.18.6 Header

The header is a small dictionary encoded in msgpack that includes some metadata about the message, such as compression.

### 2.18.7 Large Bytestrings

Whenever a message comes in with very large byte values like the following:

```
{'key': 'x',
 'address': 'alice',
 'data-1': b'...'  # very long bytestring
 'data-2': b'...'  # very long bytestring
 }
```

We separate the message into two messages, one encoding all of the large bytestrings, and one encoding everything else:

```
{'key': 'x', 'addresss': 'alice'}
{'data-1': b'...', 'data-2': b'...'}
```

The first message we pass normally with msgpack, the second we pass in multiple parts, including a header that contains the keys and compression used for each value:

```
{'keys': ['data-1', 'data-2'],
 'compression': ['lz4', None]}
b'...'
b'...'
```

### 2.18.8 Frames

At the end of the pipeline we have a sequence of bytestrings or frames. We need to tell the receiving end how many frames there are and how long each these frames are. We order the frames and lengths of frames as follows:

1. The number of frames, stored as an 8 byte unsigned integer

2. The length of each frame, each stored as an 8 byte unsigned integer
3. Each of the frames

In the following sections we describe how we create these frames.

### 2.18.9 Performance

For large numpy arrays we currently suffer three memory copies. On a nice machine this ends up being a 1-1.5 GB/s bottleneck, which is almost always faster than the network bandwidth. These copies come from NumPy (two memcopies) and Tornado (one memcopy).

For small messages we generally serialize in around 5 microseconds.

## 2.19 Scheduler Plugins

**class** distributed.diagnostics.plugin.**SchedulerPlugin**

Interface to extend the Scheduler

The scheduler operates by triggering and responding to events like `task_finished`, `update_graph`, `task_erred`, etc..

A plugin enables custom code to run at each of those same events. The scheduler will run the analagous methods on this class when each event is triggered. This runs user code within the scheduler thread that can perform arbitrary operations in synchrony with the scheduler itself.

Plugins are often used for diagnostics and measurement, but have full access to the scheduler and could in principle affect core scheduling.

To implement a plugin implement some of the methods of this class and add the plugin to the scheduler with `Scheduler.add_plugin(myplugin)`.

#### Examples

```
>>> class Counter(SchedulerPlugin):
...     def __init__(self):
...         self.counter = 0
...
...     def task_finished(self, scheduler, key, worker, nbytes):
...         self.counter += 1
...
...     def restart(self, scheduler):
...         self.counter = 0
```

```
>>> c = Counter()
>>> scheduler.add_plugin(c)
```

**restart**(*scheduler*, ***kwargs*)

Run when the scheduler restarts itself

**task_erred**(*scheduler*, *key=None*, *worker=None*, *exception=None*, ***kwargs*)

Run when a task is reported failed

**task_finished**(*scheduler*, *key=None*, *worker=None*, *nbytes=None*, ***kwargs*)

Run when a task is reported complete

**update_graph**(*scheduler*, *dsk=None*, *keys=None*, *restrictions=None*, ***kwargs*)
Run when a new graph / tasks enter the scheduler

## 2.20 Related Work

Writing the "related work" for a project called "distributed", is a Sisyphean task. We'll list a few notable projects that you've probably already heard of down below.

You may also find the dask comparison with spark of interest.

### 2.20.1 Big Data World

- The venerable Hadoop provides batch processing with the MapReduce programming paradigm. Python users typically use Hadoop Streaming or MRJob.
- Spark builds on top of HDFS systems with a nicer API and in-memory processing. Python users typically use PySpark.
- Storm provides streaming computation. Python users typically use streamparse.

This is a woefully inadequate representation of the excellent work blossoming in this space. A variety of projects have come into this space and rival or complement the projects above. Still, most "Big Data" processing hype probably centers around the three projects above, or their derivatives.

### 2.20.2 Python Projects

There are dozens of Python projects for distributed computing. Here we list a few of the more prominent projects that we see in active use today.

#### Task scheduling

- Celery: An asynchronous task scheduler, focusing on real-time processing.
- Luigi: A bulk big-data/batch task scheduler, with hooks to a variety of interesting data sources.

#### Ad hoc computation

- IPython Parallel: Allows for stateful remote control of several running ipython sessions.
- Scoop: Implements the concurrent.futures API on distributed workers. Notably allows tasks to spawn more tasks.

#### Direct Communication

- MPI4Py: Wraps the Message Passing Interface popular in high performance computing.
- PyZMQ: Wraps ZeroMQ, the gentleman's socket.

#### Venerable

There are a couple of older projects that often get mentioned

- Dispy: Embarrassingly parallel function evaluation
- Pyro: Remote objects / RPC

### 2.20.3 Relationship

In relation to these projects `distributed`...

- Supports data-local computation like Hadoop and Spark
- Uses a task graph with data dependencies abstraction like Luigi
- In support of ad-hoc applications, like IPython Parallel and Scoop

### 2.20.4 In depth comparison to particular projects

#### IPython Parallel

**Short Description**

IPython Parallel is a distributed computing framework from the IPython project. It uses a centralized hub to farm out jobs to several `ipengine` processes running on remote workers. It communicates over ZeroMQ sockets and centralizes communication through the central hub.

IPython parallel has been around for a while and, while not particularly fancy, is quite stable and robust.

IPython Parallel offers parallel `map` and remote `apply` functions that route computations to remote workers

```
>>> view = Client(...)[:]
>>> results = view.map(func, sequence)
>>> result = view.apply(func, *args, **kwargs)
>>> future = view.apply_async(func, *args, **kwargs)
```

It also provides direct execution of code in the remote process and collection of data from the remote namespace.

```
>>> view.execute('x = 1 + 2')
>>> view['x']
[3, 3, 3, 3, 3, 3]
```

**Brief Comparison**

Distributed and IPython Parallel are similar in that they provide `map` and `apply/submit` abstractions over distributed worker processes running Python. Both manage the remote namespaces of those worker processes.

They are dissimilar in terms of their maturity, how worker nodes communicate to each other, and in the complexity of algorithms that they enable.

**Distributed Advantages**

The primary advantages of `distributed` over IPython Parallel include

1. Peer-to-peer communication between workers
2. Dynamic task scheduling

Distributed workers share data in a peer-to-peer fashion, without having to send intermediate results through a central bottleneck. This allows distributed to be more effective for more complex algorithms and to manage larger datasets in a more natural manner. IPython parallel does not provide a mechanism for workers to communicate with each other, except by using the central node as an intermediary for data transfer or by relying on some other medium, like a shared file system. Data transfer through the central node can easily become a bottleneck and so IPython parallel has been mostly helpful in embarrassingly parallel work (the bulk of applications) but has not been used extensively for more sophisticated algorithms that require non-trivial communication patterns.

The distributed executor includes a dynamic task scheduler capable of managing deep data dependencies between tasks. The IPython parallel docs include a recipe for executing task graphs with data dependencies. This same idea is core to all of distributed, which uses a dynamic task scheduler for all operations. Notably, distributed.Future objects can be used within submit/map/get calls before they have completed.

```
>>> x = executor.submit(f, 1)  # returns a future
>>> y = executor.submit(f, 2)  # returns a future
>>> z = executor.submit(add, x, y)  # consumes futures
```

The ability to use futures cheaply within submit and map methods enables the construction of very sophisticated data pipelines with simple code. Additionally, distributed can serve as a full dask task scheduler, enabling support for distributed arrays, dataframes, machine learning pipelines, and any other application build on dask graphs. The dynamic task schedulers within distributed are adapted from the dask task schedulers and so are fairly sophisticated/efficient.

**IPython Parallel Advantages**

IPython Parallel has the following advantages over distributed

1. Maturity: IPython Parallel has been around for a while.
2. Explicit control over the worker processes: IPython parallel allows you to execute arbitrary statements on the workers, allowing it to serve in system administration tasks.
3. Deployment help: IPython Parallel has mechanisms built-in to aid deployment on SGE, MPI, etc.. Distributed does not have any such sugar, though is fairly simple to set up by hand.
4. Various other advantages: Over the years IPython parallel has accrued a variety of helpful features like IPython interaction magics, @parallel decorators, etc..

### concurrent.futures

The distributed.Executor API is modeled after concurrent.futures and PEP-3184. It has a few notable differences:

- distributed accepts Future objects within calls to submit/map. It is preferable to submit Future objects directly rather than wait on them before submission.
- The map function returns Future objects, not concrete results. The map function returns immediately.
- It is not yet possible to cancel a Future (though this is theoretically possible please raise an issue if this is of concrete importance to you.)
- Distributed generally does not support timeouts or callbacks

distributed.CompatibleExecutor is a subclass of distributed.Executor that does conform to the concurrent.futures API, allowing it to be used as a drop-in replacement for other Executors using the common API.

## N



## P



## R



## S



## T



## U



## W