Chapter 16. Testing, Debugging, and Optimizing

You're not finished with a programming task when you're done writing the code; you're finished when the code runs correctly, with acceptable performance. *Testing* (covered in "Testing") means verifying that code runs correctly, by automatically exercising the code under known conditions and checking that results are as expected. *Debugging* (covered in "Debugging") means discovering causes of incorrect behavior and repairing them (repair is often easy, once you figure out the causes).

Optimizing (covered in "Optimization") is often used as an umbrella term for activities meant to ensure acceptable performance. Optimizing breaks down into *benchmarking* (measuring performance for given tasks to check that it's within acceptable bounds), *profiling* (*instrumenting* the program with extra code to identify performance bottlenecks), and optimizing proper (removing bottlenecks to make overall program performance acceptable). Clearly, you can't remove performance bottlenecks until you've found out where they are (using profiling), which in turn requires knowing that there *are* performance problems (using benchmarking).

This chapter covers the subjects in the natural order in which they occur in development: testing first and foremost, debugging next, and optimizing last. Most programmers' enthusiasm focuses on optimization: testing and debugging are often (wrongly, in our opinion) perceived as being chores, while optimization is seen as being fun. Thus, were you to read only one section of the chapter, we might suggest that section be "Developing a Fast-Enough Python Application", which summarizes the Pythonic approach to optimization—close to Jackson's classic "Rules of Optimization: Rule 1: Don't do it. Rule 2 (for experts only): Don't do it *yet*."

All of these tasks are large and important, and each could fill at least a book by itself. This chapter does not even come close to exploring every related technique and implication; it focuses on Python-specific techniques, approaches, and tools.

Testing

In this chapter, we distinguish between two different kinds of testing: unit testing and system testing. Testing is a rich, important field: many more distinctions could be drawn, but we focus on the issues of most importance to software developers. Many developers are reluctant to spend time on testing, seeing it as time stolen from "real" development, but this is short-sighted: defects are easier to fix the earlier you find out about them—an hour spent developing tests can amply pay for itself when you find defects ASAP, saving you many hours of debugging that would otherwise have been needed in later phases of the software development cycle.

Unit Testing and System Testing

Unit testing means writing and running tests to exercise a single module, or an even smaller unit, such as a class or function. *System testing* (also known as *functional* or *integration* or *end-to-end* testing) involves running an entire program with known inputs. Some classic books on testing also draw the distinction between *white-box testing*, done with knowledge of a program's internals, and *black-box testing*, done without such knowledge. This classic viewpoint parallels, but does not exactly duplicate, the modern one of unit versus system testing.

Unit and system testing serve different goals. Unit testing proceeds apace with development; you can and should test each unit as you're developing it. One relatively modern approach (first proposed in 1971 in Weinberg's immortal classic *The Psychology of Computer Programming*) is known as *test-driven development* (TDD): for each feature that your program must have, you first write unit tests, and only then do you proceed to write code that implements the feature and makes the tests pass. TDD may seem upside-down, but it has advantages; for example, it ensures that you won't omit unit tests for some feature. Developing test-first is helpful because it urges you to focus first on what tasks a certain function, class, or method should accomplish, dealing only afterward with *how* to implement that function, class, or method. An innovation along the lines of TDD is behavior-driven development.

In order to test a unit—which may depend on other units not yet fully developed—you often have to write *stubs*, also known as *mocks*—fake implementations of various units' interfaces giving known, correct responses in cases needed to test other units. The `mock` module (part of v3's standard library, in the package `unittest`; backport available for v2, from PyPI) helps you implement such stubs.

System testing comes later, as it requires the system to exist, with at least some subset of system functionality believed (based on unit testing) to be working. System testing offers a sanity check: each module in the program works properly (passes unit tests), but does the *whole* program work? If each unit is okay but the system is not, there's a problem in the integration between units—the way the units cooperate. For this reason, system testing is also known as *integration* testing.

System testing is similar to running the system in production use, except that you fix inputs in advance so that any problems you may find are easy to reproduce. The cost of failures in system testing is lower than in production use, since outputs from system testing are not used to make decisions, serve customers, control external systems, and so on. Rather, outputs from system testing are systematically compared with the outputs that the system *should* produce given the known inputs. The purpose is to find, in cheap and reproducible ways, discrepancies between what the program *should* do and what the program actually *does*.

Failures discovered by system testing (just like system failures in production use) may reveal some defects in unit tests, as well as defects in the code. Unit testing may have been insufficient: a module's unit tests may have failed to exercise all needed functionality of the module. In that case, the unit tests need to be beefed up. Do that *before* you change your code to fix the problem, then run the newly enhanced unit tests to confirm that they now show the problem. Then, fix the problem, and run unit tests again to confirm they show no problem anymore. Finally, rerun the system tests to confirm that the problem has indeed gone away.

Bug-fixing best practice

This best practice is a specific application of test-driven design that we recommend without reservation: never fix a bug before having added unit tests that would have revealed the bug (this practice is excellent, cheap insurance against software regression bugs).

Often, failures in system testing reveal communication problems within the development team: a module correctly implements a certain functionality, but another module expects different functionality. This kind of problem (an integration problem in the strict sense) is hard to pinpoint in unit testing. In good development practice, unit tests must run often, so it is crucial that they run fast. It's therefore essential, in the unit-testing phase, that each unit can assume other units are working correctly and as expected.

Unit tests run in reasonably late stages of development can reveal integration problems if the system architecture is hierarchical, a common and reasonable organization. In such an architecture, low-level modules depend on no others (except library modules, which you can assume to be correct), so the unit tests of such low-level modules, if complete, suffice to assure correctness. High-level modules depend on low-level ones, and thus also depend on correct understanding about what functionality each module expects and supplies. Running complete unit tests on high-level modules (using true low-level modules, not stubs) exercises interfaces between modules, as well as the high-level modules' own code.

Unit tests for high-level modules are thus run in two ways. You run the tests with stubs for the low levels during the early stages of development, when the low-level modules are not yet ready or, later, when you only need to check the correctness of the high levels. During later stages of development, you also regularly run the high-level modules' unit tests using the true low-level modules. In this way, you check the correctness of the whole subsystem, from the high levels downward. Even in this favorable case, you *still* need to run system tests to ensure the system's functionality is exercised and checked, and no interface between modules is neglected.

System testing is similar to running the program in normal ways. You need special support only to ensure that known inputs are supplied and that outputs are captured for comparison with expected outputs. This is easy for programs that perform I/O (input/output) on files, and hard for programs whose I/O relies on a GUI, network, or other communication with external entities. To simulate such external entities and make them predictable and entirely observable, you generally need platform-dependent infrastructure. Another useful piece of supporting infrastructure for system testing is a *testing framework* to automate the running of system tests, including logging of successes and failures. Such a framework can also help testers prepare sets of known inputs and corresponding expected outputs.

Both free and commercial programs for these purposes exist, but they are not dependent on which programming languages are used in the system under test. System testing is a close kin to what was classically known as black-box testing, or testing that is independent from the implementation of the system under test (and therefore, in particular, independent from the programming languages used for implementation). Instead, testing frameworks usually depend on the operating system platform on which they run, since the tasks they perform are platform-dependent: running programs with given inputs, capturing their outputs, and particularly simulating and capturing GUI, network, and other interprocess communication I/O. Since frameworks for system testing depend on the platform and not on programming languages, we do not cover them further in this book. For a thorough list of Python testing tools, see the Python wiki.

The doctest Module

The `doctest` module exists to let you create good examples in your code's docstrings, by checking that the examples do in fact produce the results that your docstrings show for them. `doctest` recognizes such examples by looking within the docstring for the interactive Python prompt `'>>> '`, followed on the same line by a Python statement, and the statement's expected output on the next line(s).

As you develop a module, keep the docstrings up to date and enrich them with examples. Each time a part of the module (e.g., a function) is ready, or partially ready, make it a habit to add examples to its docstring. Import the module into an interactive session, and use the parts you just developed in order to provide examples with a mix of typical cases, limit cases, and failing cases. For this specific purpose only, use `from module import *` so that your examples don't prefix `module.` to each name the module supplies. Copy and paste the interactive session into the docstring in an editor, adjust any glitches, and you're almost done.

Your documentation is now enriched with examples, and readers have an easier time following it, assuming you choose a good mix of examples, wisely seasoned with nonexample text. Make sure you have docstrings, with examples, for the module as a whole, and for each function, class, and method the module exports. You may choose to skip functions, classes, and methods whose names start with `_`, since (as their names indicate) they're meant to be private implementation details; `doctest` by default ignores them, and so should readers of your module.

Match reality

Examples that don't match the way your code works are worse than useless. Documentation and comments are useful only if they match reality; docs and comments that lie can be seriously damaging.

Docstrings and comments often get out of date as code changes, and thus become misinformation, hampering, rather than helping, any reader of the source. Better to have no comments and docstrings at all, poor as such a choice would be, than to have ones that lie. `doctest` can help you through the examples in your docstrings. A failing `doctest` run should prompt you to review the docstring that contains the failing examples, thus reminding you to keep the whole docstring updated.

At the end of your module's source, insert the following snippet:

```
if __name__ == '__main__':
    import doctest
    doctest.testmod()
```

This code calls the function `testmod` of the module `doctest` when you run your module as the main program. `testmod` examines docstrings (the module docstring, and docstrings of all public functions, classes, and methods thereof). In each docstring, `testmod` finds all examples (by looking for occurrences of the interpreter prompt `'>>> '`, possibly preceded by whitespace) and runs each example. `testmod` checks that each example's results match the output given in the docstring right after the example. In case of exceptions, `testmod` ignores the traceback, and just checks that the expected and observed error messages are equal.

When everything goes right, `testmod` terminates silently. Otherwise, it outputs detailed messages about examples that failed, showing expected and actual output. Example 16-1 shows a typical example of `doctest` at work on a module *mod.py*.

Example 16-1. Using doctest

```
"""
This module supplies a single function reverse_words that
reverses
a string by
words.

>>> reverse_words('four score and seven
years')
'years seven and score
four'
>>>
reverse_words('justoneword')
'justoneword'
>>>
reverse_words('')
''

You must call reverse_words with one argument, a
string:

>>>
reverse_words()
Traceback (most recent call
last):

...
```

```
...
TypeError: reverse_words() takes exactly 1 argument (0
given)
>>> reverse_words('one',
'another')
Traceback (most recent call
last):

...
TypeError: reverse_words( ) takes exactly 1 argument (2
given)
>>>
reverse_words(1)
Traceback (most recent call
last):

...
AttributeError: 'int' object has no attribute
'split'
>>> reverse_words(u'however, unicode is all right too')  # v2
check
u'too right all is unicode
however,'

As a side effect, reverse_words eliminates any redundant
spacing:

>>> reverse_words('with   redundant
spacing')
'spacing redundant
with'

"""
def reverseWords(astring):
    words = astring.split()
    words.reverse()
          '
    return '  .join(words)

if __name__=='__main__':
    import doctest
    doctest.testmod()
```

In this module's docstring, we snipped the tracebacks from the docstring and replaced them with ellipses (`...`): this is good practice, since `doctest` ignores tracebacks, which add nothing to the explanatory value of a failing case. Apart from this snipping, the docstring is the copy and paste of an interactive session, plus some explanatory text and empty lines for readability. Save this source as *mod.py*, and then run it with `python mod.py`. It produces no output, meaning that all examples work right. Try `python mod.py -v` to get an account of all tests tried and a verbose summary at the end. Finally, alter the example results in the module docstring, making them incorrect, to see the messages `doctest` provides for errant examples.

While `doctest` is not meant for general-purpose unit testing, it can be tempting to use it for that purpose. The recommended way to do unit testing in Python is with the module `unittest`, covered in "The unittest Module".

However, unit testing with `doctest` can be easier and faster to set up, since that requires little more than copying and pasting from an interactive session. If you need to maintain a module that lacks unit tests, retrofitting such tests into the module with `doctest` is a reasonable compromise (although you should plan to eventually upgrade to full-fledged tests with `unittest`). It's better to have `doctest`-based unit tests than not to have any unit tests at all, as might otherwise happen should you decide that setting up tests properly with `unittest` from the start would take you too long.

If you do decide to use `doctest` for unit testing, don't cram extra tests into your module's docstrings. This would damage the docstrings by making them too long and hard to read. Keep in the docstrings the right amount and kind of examples, strictly for explanatory purposes, just as if unit testing were not in the picture. Instead, put the extra tests into a global variable of your module, a dictionary named `__test__`. The keys in `__test__` are strings used as arbitrary test names, and the corresponding values are strings that `doctest` picks up and uses in just the same way as it uses docstrings. The values in `__test__` may also be function and class objects, in which case `doctest` examines their docstrings for tests to run. This latter feature is a convenient way to run `doctest` on objects with private names, which `doctest` skips by default.

The `doctest` module also supplies two functions that return instances of the `unittest.TestSuite` class based on doctests so that you can integrate such tests into testing frameworks based on `unittest`. The documentation for this advanced functionality is online.

The unittest Module

The `unittest` module is the Python version of a unit-testing framework originally developed by Kent Beck for Smalltalk. Similar, widespread versions of the framework also exist for many other programming languages (e.g., the `JUnit` package for Java) and are often collectively referred to as `xUnit`.

To use `unittest`, don't put your testing code in the same source file as the tested module: write a separate test module for each module to test. A popular convention is to name the test module like the module being tested, with a prefix such as `'test_'`, and put it in a subdirectory of the source's directory named *test*. For example, the test module for *mod.py* can be *test/test_mod.py*. A simple, consistent naming convention makes it easy for you to write and maintain auxiliary scripts that find and run all unit tests for a package.

Separation between a module's source code and its unit-testing code lets you refactor the module more easily, including possibly recoding its functionality in C, without perturbing unit-testing code. Knowing that *test_mod.py* stays intact, whatever changes you make to *mod.py*, enhances your confidence that passing the tests in *test_mod.py* indicates that *mod.py* still works correctly after the changes.

A unit-testing module defines one or more subclasses of `unittest`'s `TestCase` class. Each subclass specifies one or more test cases by defining *test-case methods*, methods that are callable without arguments and whose names start with `test`.

The subclass may override `setUp`, which the framework calls to prepare a new instance just before each test case, and `tearDown`, which the framework calls to clean things up right after each test case; this setup-teardown arrangement is known as a *test fixture*.

Have setUp use addCleanup when needed

When `setUp` propagates an exception, `tearDown` does not execute. So, if `setUp` prepares several things needing cleanup, and some preparation steps might cause uncaught exceptions, `setUp` must not rely on `tearDown` for the clean-up work; rather, right after each preparation step succeeds, call `self.addCleanup(f, *a, **k)`, passing a clean-up callable `f` (and optionally positional and named arguments for `f`). In this case, `f(*a, **k)` does get called after the test case (after `tearDown` when `setUp` propagates no exception, but unconditionally even when `setUp` does propagate), so the needed clean-up code always executes.

Each test case calls, on `self`, methods of the class `TestCase` whose names start with `assert` to express the conditions that the test must meet. `unittest` runs the test-case methods within a `TestCase` subclass in arbitrary order, each on a new instance of the subclass, running `setUp` just before each test case and `tearDown` just after each test case.

`unittest` provides other facilities, such as grouping test cases into test suites, per-class and per-module fixtures, test discovery, and other, even more advanced functionality. You do not need such extras unless you're defining a custom unit-testing framework or, at the very least, structuring complex testing procedures for equally complex packages. In most cases, the concepts and details covered in this section are enough to perform effective and systematic unit testing. Example 16-2 shows how to use `unittest` to provide unit tests for the module *mod.py* of Example 16-1. This example uses `unittest` to perform exactly the same tests that Example 16-1 uses as examples in docstrings using `doctest`.

Example 16-2. Using unittest

```
""" This module tests function reverseWords
provided by module mod.py.
"""
import unittest
import mod

class ModTest(unittest.TestCase):

    def testNormalCaseWorks(self):
        self.assertEqual(
                              'four score and seven
            mod.reverse_words(years'
),
            'years seven and score
            four'                       )

    def testSingleWordIsNoop(self):
        self.assertEqual(
            mod.reverse_words('justoneword'),
            'justoneword')

    def testEmptyWorks(self):
        self.assertEqual(mod.reverse_words(''), '')

    def testRedundantSpacingGetsRemoved(self):
        self.assertEqual(
                              'with   redundant
            mod.reverse_words(spacing'
),
            'spacing redundant
            with'                   )

    def testUnicodeWorks(self):
        self.assertEqual(
                               'unicode is all right
            mod.reverse_words(utoo'                      ),
             'too right all is
            uunicode'                   )

    def testExactlyOneArgumentIsEnforced(self):
        self.assertRaises(TypeError, mod.reverse_words)
        with self.assertRaises(TypeError):
            mod.reverse_words('one', 'another')

    def testArgumentMustBeString(self):
        with self.assertRaises((AttributeError,TypeError)):
            mod.reverse_words(1)

if __name__=='__main__':
    unittest.main()
```

Running this script with `python test/test_mod.py` (or, equivalently, `python -m test.test_mod`) is a bit more verbose than using `python mod.py` to run `doctest`, as in Example 16-1. *test_mod.py* outputs a `.` (dot) for each test case it runs, then a separator line of dashes, and finally a summary line, such as "Ran 7 tests in 0.110s," and a final line of "OK" if every test passed.

Each test-case method makes one or more calls to methods whose names start with `assert`. Here, the method `testExactlyOneArgumentIsEnforced` is the only one with two such calls. In more complicated cases, multiple calls to assert methods from a single test-case method are quite common.

Even in a case as simple as this, one minor aspect shows that, for unit testing, `unittest` is more powerful and flexible than `doctest`. In the method `testArgumentMustBeString`, we pass as the argument to `assertRaises` a pair of exception classes, meaning we accept either kind of exception. *test_mod.py* therefore accepts as valid multiple implementations of *mod.py*. It accepts the implementation in Example 16-1, which tries calling the method `split` on its argument, and therefore raises `AttributeError` when called with an argument that is not a string. However, it also accepts a different hypothetical implementation, one that raises `TypeError` instead when called with an argument of the wrong type. It's possible to code such functionality with `doctest`, but it would be awkward and nonobvious, while `unittest` makes it simple and natural.

This kind of flexibility is crucial for real-life unit tests, which to some extent are executable specifications for their modules. You could, pessimistically, view the need for test flexibility as meaning the interface of the code you're testing is not well defined. However, it's best to view the interface as being defined with a useful amount of flexibility for the implementer: under circumstance `X` (argument of invalid type passed to function `reverseWords`, in this example), either of two things (raising `AttributeError` or `TypeError`) is allowed to happen.

Thus, implementations with either of the two behaviors are correct, and the implementer can choose between them on the basis of such considerations as performance and clarity. By viewing unit tests as executable specifications for their modules (the modern view, and the basis of test-driven development), rather than as white-box tests strictly constrained to a specific implementation (as in some traditional taxonomies of testing), you'll find that the tests become an even more vital component of the software development process.

The TestCase class

With `unittest`, you write test cases by extending `TestCase`, adding methods, callable without arguments, whose names start with `test`. Such test-case methods, in turn, call methods that your class inherits from `TestCase`, whose names start with `assert`, to indicate conditions that must hold for the test to succeed.

The `TestCase` class also defines two methods that your class can optionally override to group actions to perform right before and after each test-case method runs. This doesn't exhaust `TestCase`'s functionality, but you won't need the rest unless you're developing testing frameworks or performing other advanced tasks. The frequently called methods in a `TestCase` instance `t` are the following:

assertAlmostEqual	`t.assertAlmostEqual(first,second,places=7,msg=None)` Fails and outputs `msg` when `first!=second` to within `places` decimal digits; otherwise, does nothing. Almost always, this method is preferable to `assertEqual` when what you are comparing are `float`s, because, due to floating-point computation vagaries, equality in that realm is usually approximate, not 100% exact.
assertEqual	`t.assertEqual(first,second,msg=None)` Fails and outputs `msg` when `first!=second`; otherwise, does nothing.

assertFalse	`t.assertFalse(condition, msg=None)` Fails and outputs `msg` when `condition` is true; otherwise, does nothing.
assertNotAlmostEqual	`t.assertNotAlmostEqual(first,second,places=7,msg=None)` Fails and outputs `msg` when `first==second` to within `places` decimal digits; otherwise, does nothing.
assertNotEqual	`t.assertNotEqual(first,second,msg=None)` Fails and outputs `msg` when `first==second`; otherwise, does nothing.

assertRaises	`t.assertRaises(exceptionSpec,callable,*args,**kwargs)`
	
	Calls `callable(*args,**kwargs)`. Fails when the call doesn't raise any exception. When the call raises an exception that does not meet `exceptionSpec`, `assertRaises` propagates the exception. When the call raises an exception that meets `exceptionSpec`, `assertRaises` does nothing. `exceptionSpec` can be an exception class or a tuple of classes, just like the first argument of the `except` clause in a `try`/`except` statement.
	
	An alternative, and usually preferable, way to use `assertRaises` is as a *context manager*—that is, in a `with` statement:
	
	`with self.assertRaises(exceptionSpec):    ...a block of code...`
	
	Here, the "block of code" indented within the `with` statement executes, rather than just the `callable` being called with certain arguments. The expectation (which avoids the construct failing) is that the block of code raises an exception meeting the given exception specification (an exception class or a tuple of classes). This alternative approach is more general, natural, and readable.

assertTrue	`t.assertTruecondition,msg=None)` Fails and outputs `msg` when `condition` is false; otherwise, does nothing. Do not use this method when you can use a more specific one, such as `assertEqual`: specific methods provide clearer messages.

fail	`t.fail(msg=None)`
	
	Fails unconditionally and outputs `msg`. An example snippet might be:
	
	`if not complex_check_if_its_ok(some, thing):`
	`    self.fail('Complex checks failed on {}, {}`
	`'`
	`    .format(some, thing))`

setUp	`t.setUp()` The framework calls `t.setUp()` just before calling a test-case method. `setUp` in `TestCase` does nothing. `setUp` exists only to let your class override the method when your class needs to perform some preparation for each test.

tearDown	`t.tearDown()` The framework calls `t.tearDown()` just after a test-case method. `tearDown` in `TestCase` does nothing. `tearDown` exists only to let your class override the method when your class needs to perform some cleanup after each test.

In addition, a `TestCase` instance maintains a last-in, first-out list (*LIFO* stack) of *clean-up functions*. When code in one of your tests (or in `setUp`) does something that requires cleanup, call `addCleanup`, passing a clean-up callable `f` and optionally positional and named arguments for `f`. To perform the stacked-up cleanups, you may call `doCleanups`; however, the framework itself calls `doCleanups` after `tearDown`. Here are the signatures of the two cleanup methods:

addCleanup	`.addCleanup(`*`func`*`, *`*`a`*`, t**`*`k`*`)` Appends `(`*`func`*`, `*`a`*`, `*`k`*`)` at the end of the cleanups' list.

doCleanups	`t.doCleanups()`
	
	Perform all cleanups, if any is stacked. Substantially equivalent to:
	
	`while self.list_of_cleanups:`
	`    func, a, k = self.list_of_cleanups.pop()`
	`    func(*a, **k)`
	
	for a hypothetical stack `self.list_of_cleanups`, plus, of course, error-checking and reporting.

Unit tests dealing with large amounts of data

Unit tests must be fast: run them often as you develop. So, unit-test each aspect of your modules on small amounts of data, when feasible. This makes unit tests faster, and lets you embed the data in the test's source code. When you test a function that reads from or writes to a file object, use an instance of the class `io.TextIO` for a text—that is, Unicode file (`io.BytesIO` for a byte, i.e., binary file, as covered in "In-Memory "Files": io.StringIO and io.BytesIO")—to get a "file" with the data in memory: faster than writing to disk, and no clean-up chore (removing disk files after the tests).

In rare cases, it may be impossible to exercise a module's functionality without supplying and/or comparing data in quantities larger than can be reasonably embedded in a test's source code. In such cases, your unit test must rely on auxiliary, external data files to hold the data to supply to the module it tests and/or the data it needs to compare to the output. Even then, you're generally better off using instances of the above-mentioned `io` classes, rather than directing the tested module to perform actual disk I/O. Even more important, we strongly suggest that you generally use stubs to unit-test modules that interact with external entities, such as databases, GUIs, or other programs over a network. It's easier to control all aspects of the test when using stubs rather than real external entities. Also, to reiterate, the speed at which you can run unit tests is important, and it's faster to perform simulated operations in stubs, rather than real operations.

To test, make randomness reproducible by supplying a seed

If your code uses pseudorandom numbers (e.g., as covered in "The random Module"), you can make it easier to test by ensuring its "random" behavior is *reproducible*: specifically, ensure that it's easy for your tests to have random.seed called with a known argument, so that the ensuing pseudorandom numbers become fully predictable. This also applies when you're using pseudorandom numbers to set up your tests by generating random inputs: such generation should default to a known seed, to be used in most testing, keeping the extra flexibility of changing seeds for specific techniques such as fuzzing.

Debugging

Since Python's development cycle is fast, the most effective way to debug is often to edit your code to output relevant information at key points. Python has many ways to let your code explore its own state in order to extract information that may be relevant for debugging. The inspect and traceback modules specifically support such exploration, which is also known as reflection or introspection.

Once you have debugging-relevant information, print is often the way to display it (pprint, covered in "The pprint Module", is also often a good choice). Better, log debugging information to files. Logging is useful for programs that run unattended (e.g., server programs). Displaying debugging information is just like displaying other information, as covered in Chapter 10. Logging such information is like writing to files (covered in Chapter 10); however, to help with the frequent task of logging, Python's standard library supplies a logging module, covered in "The logging package". As covered in Table 7-3, rebinding excepthook in the module sys lets your program log error info just before terminating with a propagating exception.

Python also offers hooks to enable interactive debugging. The pdb module supplies a simple text-mode interactive debugger. Other, powerful interactive debuggers for Python are part of integrated development environments (IDEs), such as IDLE and various commercial offerings, as mentioned in "Python Development Environments"; we do not cover these debuggers further.

Before You Debug

Before you embark on lengthy debugging explorations, make sure you have thoroughly checked your Python sources with the tools mentioned in Chapter 2. Such tools catch only a subset of the bugs in your code, but they're much faster than interactive debugging: their use amply repays itself.

Moreover, again before starting a debugging session, make sure that all the code involved is well covered by unit tests, covered in "The unittest Module". As mentioned earlier in the chapter, once you have found a bug, *before* you fix it, add to your suite of unit tests (or, if need be, to the suite of system tests) a test or two that would have found the bug had they been present from the start, and run the tests again to confirm that they now reveal and isolate the bug; only once that is done should you proceed to fix the bug. By regularly following this procedure, you get a much better suite of tests; learn to write better, more thorough tests; and gain much sounder assurance about the overall, *enduring* correctness of your code.

Remember, even with all the facilities offered by Python, its standard library, and whatever IDEs you fancy, debugging is still *hard*. Take this into account even before you start designing and coding: write and run plenty of unit tests, and keep your design and code *simple*, to reduce to the minimum the amount of debugging you will need! Classic advice about this, by Brian Kernighan: "Debugging is twice as hard as writing the code in the first place. Therefore, if you write the code as cleverly as you can, you are, by definition, not smart enough to debug it." This is part of why "clever" is not a positive word when used to describe Python code, or a coder…

The inspect Module

The inspect module supplies functions to get information about all kinds of objects, including the Python call stack (which records all function calls currently executing) and source files. The most frequently used functions of inspect are as follows:

getargspec, formatargspec

`getargspec(f)`

Deprecated in v3: still works in Python 3.5 and 3.6, but will be removed in some future version. To introspect callables in v3, use `inspect.signature(f)` and the resulting instance of class `inspect.Signature`, covered in "Introspecting callables in v3".

`f` is a function object. `getargspec` returns a named tuple with four items: `(args, varargs, keywords, defaults)`. `args` is the sequence of names of `f`'s parameters. `varargs` is the name of the special parameter of the form `*a`, or `None` when `f` has no such parameter. `keywords` is the name of the special parameter of the form `**k`, or `None` when `f` has no such parameter. `defaults` is the tuple of default values for `f`'s arguments. You can deduce other details of `f`'s signature from `getargspec`'s results: `f` has `len(args)-len(defaults)` mandatory positional arguments, and the names of `f`'s optional arguments are the strings that are the items of the list slice `args[-len(defaults):]`.

`formatargspec` accepts one to four arguments, same as the items of the named tuple that `getargspec` returns, and returns a string with this information. Thus, `formatargspec(*getargspec(f))` returns a string with `f`'s parameters (also known as `f`'s *signature*) in parentheses, as in the `def` statement that created `f`. For example:

```
import inspectdef f(a,b=23,**c): passprint(inspect.formatargspec(*
                                         (a, b=23,
inspect.getargspec(f)))# prints: **c)
```

---

getargvalues, formatargvalues

`getargvalues(f)`

Deprecated in v3: still works in Python 3.5 and 3.6, but will be removed in some future version. To introspect callables in v3, use `inspect.signature(f)` and the resulting instance of class `inspect.Signature`, covered in "Introspecting callables in v3".

`f` is a frame object—for example, the result of a call to the function `_getframe` in module `sys` (covered in "The frame Type") or to function `currentframe` in module `inspect`. `getargvalues` returns a named tuple with four items: `(args, varargs, keywords, locals)`. `args` is the sequence of names of `f`'s function's parameters. `varargs` is the name of the special parameter of form `*a`, or `None` when `f`'s function has no such parameter. `keywords` is the name of the special parameter of form `**k`, or `None` when `f`'s function has no such parameter. `locals` is the dictionary of local variables for `f`. Since arguments, in particular, are local variables, the value of each argument can be obtained from `locals` by indexing the `locals` dictionary with the argument's corresponding parameter name.

`formatargvalues` accepts one to four arguments that are the same as the items of the named tuple that `getargvalues` returns, and returns a string with this information. `formatargvalues(*getargvalues(f))` returns a string with `f`'s arguments in parentheses, in named form, as used in the call statement that created `f`. For example:

```
def f(x=23): return inspect.currentframe()print(inspect.
formatargvalues(  *inspect.getargvalues(f())))# prints: (x=23)
```

---

currentframe	`currentframe()` Returns the frame object for the current function (the caller of `currentframe`). `formatargvalues(*getargvalues(currentframe()))`, for example, returns a string with the arguments of the calling function.
getdoc	`getdoc(obj)` Returns the docstring for `obj`, a multiline string with tabs expanded to spaces and redundant whitespace stripped from each line.
getfile, getsourcefile	`getfile(obj)` Returns the name of the file that defined `obj` ; raises `TypeError` when unable to determine the file. For example, `getfile` raises `TypeError` when `obj` is built-in. `getfile` returns the name of a binary or source file. `getsourcefile` returns the name of a source file and raises `TypeError` when all it can find is a binary file, not the corresponding source file.

getmembers	`getmembers(obj, filter=None)`
	
	Returns all attributes (members), both data and methods (including special methods) of `obj`, a sorted list of `(name,value)` pairs. When `filter` is not `None`, returns only attributes for which callable `filter` is true when called on the attribute's `value`, like:
	
	`sorted((n, v) for n, v in getmembers(obj) if filter(v))`

getmodule	`getmodule(obj)` Returns the module object that defined `obj`, or `None` when it is unable to determine it.

getmro	`getmro(c)`
	
	Returns a tuple of bases and ancestors of class `c` in method resolution order. `c` is the first item in the tuple. Each class appears only once in the tuple. For example:
	
	`class newA(object): passclass newB(newA): passclass newC(newA): pass class newD(newB, newC): passfor c in inspect.getmro(newD): print(c. __name__, end=' ')newD newB newC newA object`

getsource, getsourcelines	`getsource(obj)` Returns a multiline string that is the source code for `obj`; raises `IOError` if it is unable to determine or fetch it. `getsourcelines` returns a pair: the first item is the source code for `obj` (a list of lines), and the second item is the line number of first line within its file.

isbuiltin, isclass, iscode, isframe, isfunction, ismethod, ismodule, isroutine	`isbuiltin(obj)` Each of these functions accepts a single argument `obj` and returns `True` when `obj` is of the kind indicated in the function name. Accepted objects are, respectively: built-in (C-coded) functions, class objects, code objects, frame objects, Python-coded functions (including `lambda` expressions), methods, modules, and—for `isroutine`—all methods or functions, either C-coded or Python-coded. These functions are often used as the `filter` argument to `getmembers`.
stack	`stack(context=1)` Returns a list of six-item tuples. The first tuple is about `stack`'s caller, the second about the caller's caller, and so on. Each tuple's items, are: frame object, filename, line number, function name, list of `context` source lines around the current line, index of current line within the list.

Introspecting callables in v3

In v3, to introspect a callable's signature, it's best not to use deprecated functions like `inspect.getargspect(f)`, which are scheduled for removal in some future Python version. Rather, call `inspect.signature(f)`, which returns an instance `s` of class `inspect.Signature`.

`s.parameters` is an `OrderedDict` mapping parameter names to `inspect.Parameter` instances. Call `s.bind(*a, **k)` to bind all parameters to the given positional and named arguments, or `s.bind_partial(*a, **k)` to bind a subset of them: each returns an instance `b` of `inspect.BoundArguments`.

For detailed information and examples on how to introspect callables' signatures through these classes and their methods, see PEP 362.

An example of using inspect

Suppose that somewhere in your program you execute a statement such as:

```
x.f()
```

and unexpectedly receive an `AttributeError` informing you that object `x` has no attribute named `f`. This means that object `x` is not as you expected, so you want to determine more about `x` as a preliminary to ascertaining why `x` is that way and what you should do about it. Change the statement to:

```
try: x.f()
except AttributeError:
    import sys, inspect
                        ,
    print('x is type {}(  {!r})'.format(type(x), x), file=sys.stderr
)
          "x's methods
    print(are:"             , file=sys.stderr, end='')
    for n, v in inspect.getmembers(x, callable):
        print(n, file=sys.stderr, end='')
    print(file=sys.stderr)
    raise
```

This example uses `sys.stderr` (covered in Table 7-3), since it displays information related to an error, not program results. The function `getmembers` of the module `inspect` obtains the name of all the methods available on `x` in order to display them. If you need this kind of diagnostic functionality often, package it up into a separate function, such as:

```
import sys, inspect
def show_obj_methods(obj, name, show=sys.stderr.write):
    show('{} is type {}({!r})\n'.format(name,obj,type(obj
)))
            's methods are:
    show("{}"                .format(name))
    for n, v in inspect.getmembers(obj, callable):

       show('{}' .format(n))
    show('\n')
```

And then the example becomes just:

```
try: x.f()
except AttributeError:
    show_obj_methods(x, 'x'
)
    raise
```

Good program structure and organization are just as necessary in code intended for diagnostic and debugging purposes as they are in code that implements your program's functionality. See also "The __debug__ Built-in Variable" for a good technique to use when defining diagnostic and debugging functions.

The traceback Module

The `traceback` module lets you extract, format, and output information about tracebacks as normally produced by uncaught exceptions. By default, the `traceback` module reproduces the formatting Python uses for tracebacks. However, the `traceback` module also lets you exert fine-grained control. The module supplies many functions, but in typical use you need only one of them:

print_exc `print_exc(limit=None, file=sys.stderr)`

Call `print_exc` from an exception handler or a function called, directly or indirectly, by an exception handler. `print_exc` outputs to file-like object `file` the traceback that Python outputs to `stderr` for uncaught exceptions. When `limit` is an integer, `print_exc` outputs only `limit` traceback nesting levels. For example, when, in an exception handler, you want to cause a diagnostic message just as if the exception propagated, but stop the exception from propagating further (so that your program keeps running and no further handlers are involved), call `traceback.print_exc()`.

The pdb Module

The `pdb` module uses the Python interpreter's debugging and tracing hooks to implement a simple command-line interactive debugger. `pdb` lets you set breakpoints, single-step on source code, examine stack frames, and so on.

To run code under `pdb`'s control, import `pdb`, then call `pdb.run`, passing as the single argument a string of code to execute. To use `pdb` for post-mortem debugging (debugging of code that just terminated by propagating an exception at an interactive prompt), call `pdb.pm()` without arguments. To trigger `pdb` directly from your application code, import `pdb`, then call `pdb.set_trace()`.

When `pdb` starts, it first reads text files named *.pdbrc* in your home directory and in the current directory. Such files can contain any `pdb` commands, but most often you put in them `alias` commands in order to define useful synonyms and abbreviations for other commands that you use often.

When `pdb` is in control, it prompts with the string `'(Pdb) '`, and you can enter `pdb` commands. The command `help` (which you can enter in the abbreviated form `h`) lists available commands. Call `help` with an argument (separated by a space) to get help about any specific command. You can abbreviate most commands to the first one or two letters, but you must always enter commands in lowercase: `pdb`, like Python itself, is case-sensitive. Entering an empty line repeats the previous command. The most frequently used `pdb` commands are listed in Table 16-1.

Table 16-1.

! `! statement`

Executes Python statement `statement` in the currently debugged context.

alias, unalias `alias [ name [ command ] ]`

`alias` with no arguments lists currently defined aliases. `alias name` outputs the current definition of alias `name`. In the full form, `command` is any `pdb` command, with arguments, and may contain `%1`, `%2`, and so on to refer to specific arguments passed to the new alias `name` being defined, or `%*` to refer to all such arguments. Command `unalias name` removes an alias.

args, a `args`

Lists all arguments passed to the function you are currently debugging.

break, b	`break [ [ location , condition ] ]` `break` with no arguments lists currently defined breakpoints and the number of times each breakpoint has triggered. With an argument, `break` sets a breakpoint at the given `location`. `location` can be a line number or a function name, optionally preceded by `filename:` to set a breakpoint in a file that is not the current one or at the start of a function whose name is ambiguous (i.e., a function that exists in more than one file). When `condition` is present, it is an expression to evaluate (in the debugged context) each time the given line or function is about to execute; execution breaks only when the expression returns a true value. When setting a new breakpoint, `break` returns a breakpoint number, which you can then use to refer to the new breakpoint in any other breakpoint-related `pdb` command.
clear, cl	`clear [ breakpoint-numbers ]` Clears (removes) one or more breakpoints. `clear` with no arguments removes all breakpoints after asking for confirmation. To deactivate a breakpoint without removing it, see `disable`, covered below.
condition	`condition breakpoint-number [ expression ]` `condition n expression` sets or changes the condition on breakpoint `n`. `condition n`, without `expression`, makes breakpoint `n` unconditional.
continue, c, cont	`continue` Continues execution of the code being debugged, up to a breakpoint, if any.
disable	`disable [ breakpoint-numbers ]` Disables one or more breakpoints. `disable` without arguments disables all breakpoints (after asking for confirmation). This differs from `clear` in that the debugger remembers the breakpoint, and you can reactivate it via `enable`.
down, d	`down` Moves down one frame in the stack (i.e., toward the most recent function call). Normally, the current position in the stack is at the bottom (i.e., at the function that was called most recently and is now being debugged), so command `down` can't go further down. However, command `down` is useful if you have previously executed command `up`, which moves the current position upward.
enable	`enable [ breakpoint-numbers ]` Enables one or more breakpoints. `enable` without arguments enables all breakpoints after asking for confirmation.

ignore `ignore breakpoint-number [ count ]`

Sets the breakpoint's ignore count (to `0` if `count` is omitted). Triggering a breakpoint whose ignore count is greater than `0` just decrements the count. Execution stops, presenting you with an interactive `pdb` prompt, when you trigger a breakpoint whose ignore count is `0`. For example, say that module *fob.py* contains the following code:

```
def f():
    for i in range(1000
):
        g(i)
def g(i):
    pass
```

Now consider the following interactive `pdb` session (minor formatting details may change depending on the Python version you're running):

```
>>> import pdb>>> import fob>>> pdb.run('fob.f()')> <string>(1)?()
                   Breakpoint 1 at                    (Pdb) ignore 1
(Pdb) break fob.gC:\mydir\fob.py:5                    500
Will ignore next 500 crossings of breakpoint   (Pdb)
1.                                            continue      >
                              (Pdb)
C:\mydir\fob.py(5)g()-> passprint(i)      500
```

The `ignore` command, as `pdb` says, tells `pdb` to ignore the next `500` hits on breakpoint `1`, which we set at `fob.g` in the previous `break` statement. Therefore, when execution finally stops, function `g` has already been called `500` times, as we show by printing its argument `i`, which indeed is now `500`. The ignore count of breakpoint `1` is now `0`; if we give another `continue` and `print i`, `i` shows as `501`. In other words, once the ignore count decrements to `0`, execution stops every time the breakpoint is hit. If we want to skip some more hits, we must give `pdb` another `ignore` command, setting the ignore count of breakpoint `1` at some value greater than `0` yet again.

list, l `list [ [ first , last ] ]`

`list` without arguments lists `11` (eleven) lines centered on the current one, or the next `11` lines if the previous command was also a `list`. Arguments to the `list` command can optionally specify the first and last lines to list within the current file. The `list` command lists physical lines, including comments and empty lines, not logical lines.

next, n `next`

Executes the current line, without "stepping into" any function called from the current line. However, hitting breakpoints in functions called directly or indirectly from the current line does stop execution.

print, p `p` *`expression`* or `print(`*`expression`*`)`

Evaluates `expression` in the current context and displays the result.

quit, q	`quit` Immediately terminates both `pdb` and the program being debugged.
return, r	`return` Executes the rest of the current function, stopping only at breakpoints, if any.
step, s	`step` Executes the current line, stepping into any function called from the current line.

tbreak	`tbreak [ ]`
	`[ location , condition ]`
	
	Like `break`, but the breakpoint is temporary (i.e., `pdb` automatically removes the breakpoint as soon as the breakpoint is triggered).

up, u	`up` Moves up one frame in the stack (i.e., away from the most recent function call and toward the calling function).
where, w	`where` Shows the stack of frames and indicates the current one (i.e., in which frame's context command `!` executes statements, command `args` shows arguments, command `print` evaluates expressions, etc.).

The warnings Module

Warnings are messages about errors or anomalies that aren't serious enough to disrupt the program's control flow (as would happen by raising an exception). The `warnings` module affords fine-grained control over which warnings are output and what happens to them. You can conditionally output a warning by calling the function `warn` in the `warnings` module. Other functions in the module let you control how warnings are formatted, set their destinations, and conditionally suppress some warnings or transform some warnings into exceptions.

Classes

Exception classes that represent warnings are not supplied by `warnings`: rather, they are built-ins. The class `Warning` subclasses `Exception` and is the base class for all warnings. You may define your own warning classes; they must subclass `Warning`, either directly or via one of its other existing subclasses, which include:

`DeprecationWarning`

Use of deprecated features supplied only for backward compatibility

`RuntimeWarning`

Use of features whose semantics are error-prone

`SyntaxWarning`

Use of features whose syntax is error-prone

`UserWarning`

Other user-defined warnings that don’t fit any of the above cases

Objects

Python supplies no concrete warning objects. A warning is made up of a `message` (a string), a `category` (a subclass of `Warning`), and two pieces of information to identify where the warning was raised from: `module` (name of the module that raised the warning) and `lineno` (line number of the source code line raising the warning). Conceptually, you may think of these as attributes of a warning object `w`: we use attribute notation later for clarity, but no specific object `w` actually exists.

Filters

At any time, the `warnings` module keeps a list of active filters for warnings. When you import `warnings` for the first time in a run, the module examines `sys.warnoptions` to determine the initial set of filters. You can run Python with the option `-W` to set `sys.warnoptions` for a given run. Do not rely on the initial set of filters being held specifically in `sys.warnoptions`, as this is an implementation aspect that may change in future versions of Python.

As each warning `w` occurs, `warnings` tests `w` against each filter until a filter matches. The first matching filter determines what happens to `w`. Each filter is a tuple of five items. The first item, `action`, is a string that defines what happens on a match. The other four items, `message`, `category`, `module`, and `lineno`, control what it means for `w` to match the filter, and all conditions must be satisfied for a match. Here are the meanings of these items (using attribute notation to indicate conceptual attributes of `w`):

`message`

> A regular expression pattern string; the match condition is `re.match(`*`message`*`,w.message, re.I)` (the match is case-insensitive).

`category`

> `Warning` or a subclass of `Warning`; the match condition is `issubclass(w.category,category)`.

`module`

> A regular expression pattern string; the match condition is `re.match(`*`module`*`,w.module)` (the match is case-sensitive).

`lineno`

> An `int`; the match condition is `lineno in (0, w.lineno)`: that is, either `lineno` is `0`, meaning `w.lineno` does not matter, or `w.lineno` must exactly equal `lineno`.

Upon a match, the first field of the filter, the `action`, determines what happens:

`'always'`

> `w.message` is output whether or not `w` has already occurred.

`'default'`

> `w.message` is output if, and only if, this is the first time `w` occurs from this specific location (i.e., this specific `w.module`, `w.location` pair).

`'error'`

`w.category(w.message)` is raised as an exception.

`'ignore'`

`w` is ignored.

`'module'`

`w.message` is output if, and only if, this is the first time `w` occurs from `w.module`.

`'once'`

`w.message` is output if, and only if, this is the first time `w` occurs from any location.

__warningsgregistry__

When a module issues a warning, `warnings` adds to that module's global variables a `dict` named `__warningsgregistry__`, if that `dict` is not already present. Each key in the `dict` is a pair `(`*`message, category`*`)`, or a tuple with three items `(`*`message, category, lineno`*`)`; the corresponding value is `True` when further occurrences of that message are to be suppressed. Thus, for example, you can reset the suppression state of all warnings from a module `m` by executing *`m`*`.__warningsregistry__.clear()`: when you do that, all messages get output again (once) even if, for example, they've previously triggered a filter with an `action` of `'module'`.

Functions

The `warnings` module supplies the following functions:

filterwarnings	`filterwarnings(action,message='.*',category=Warning, module='.*',lineno=0, append=False)` Adds a filter to the list of active filters. When `append` is true, `filterwarnings` adds the filter after all other existing filters (i.e., appends the filter to the list of existing filters); otherwise, `filterwarnings` inserts the filter before any other existing filter. All components, save `action` , have default values that mean "match everything." As detailed above, `message` and `module` are pattern strings for regular expressions, `category` is some subclass of `Warning`, `lineno` is an integer, and `action` is a string that determines what happens when a message matches this filter.
formatwarning	`formatwarning(message,category,filename,lineno)` Returns a string that represents the given warning with standard formatting.
resetwarnings	`resetwarnings()` Removes all filters from the list of filters. `resetwarnings` also discards any filters originally added with the `-W` command-line option.

showwarning	`showwarning(message,category,filename,lineno, file=sys.stderr)` Outputs the given warning to the given file object. Filter actions that output warnings call `showwarning`, letting the argument `file` default to `sys.stderr`. To change what happens when filter actions output warnings, code your own function with this signature and bind it to `warnings.showwarning`, thus overriding the default implementation.
warn	`warn(message,category=UserWarning,stacklevel=1)` Sends a warning so that the filters examine and possibly output it. The location of the warning is the current function (caller of `warn`) if `stacklevel` is `1`, or the caller of the current function if `stacklevel` is `2`. Thus, passing `2` as the value of `stacklevel` lets you write functions that send warnings on their caller's behalf, such as:

```
def toUnicode(bytestr):
    try:
        return bytestr.decode()
    except UnicodeError:
        warnings.warn(
              'Invalid characters in {}'.format(bytestr
),
              stacklevel=2)
        return bytestr.decode(errors='ignore')
```

Thanks to the parameter `stacklevel=2`, the warning appears to come from the caller of `toUnicode`, rather than from `toUnicode` itself. This is very important when the `action` of the filter that matches this warning is `default` or `module`, since these actions output a warning only the first time the warning occurs from a given location or module.

Optimization

"First make it work. Then make it right. Then make it fast." This quotation, often with slight variations, is widely known as "the golden rule of programming." As far as we've been able to ascertain, the quotation is by Kent Beck, who credits his father with it. This principle is often quoted, but too rarely followed. A negative form, slightly exaggerated for emphasis, is in a quotation by Don Knuth (who credits Hoare with it): "Premature optimization is the root of all evil in programming."

Optimization is premature if your code is not working yet, or if you're not sure what, precisely, your code should be doing (since then you cannot be sure if it's working). First make it work: ensure that your code is correctly performing exactly the tasks it is *meant* to perform.

Optimization is also premature if your code is working but you are not satisfied with the overall architecture and design. Remedy structural flaws before worrying about optimization: first make it work, then make it right. These steps are not optional; working, well-architected code is *always* a must.

In contrast, you don't always need to make it fast. Benchmarks may show that your code's performance is already acceptable after the first two steps. When performance is not acceptable, profiling often shows that all performance issues are in a small part of the code, perhaps 10 to 20 percent of the code where your program spends 80 or 90 percent of the time. Such performance-crucial regions of your code are known as *bottlenecks*, or *hot spots*. It's a waste of effort to optimize large portions of code that account for, say, 10 percent of your program's running time. Even if you made that part run 10 times as fast (a rare feat), your program's overall runtime would only decrease by

9 percent, a speedup no user would even notice. If optimization is needed, focus your efforts where they matter: on bottlenecks. You can optimize bottlenecks while keeping your code 100 percent pure Python, thus not preventing future porting to other Python implementations. In some cases, you can resort to recoding some computational bottlenecks as Python extensions (as covered in Chapter 24), potentially gaining even better performance (possibly at the expense of some potential future portability).

Developing a Fast-Enough Python Application

Start by designing, coding, and testing your application in Python, using available extension modules if they save you work. This takes much less time than it would with a classic compiled language. Then benchmark the application to find out if the resulting code is fast enough. Often it is, and you're done—congratulations! Ship it!

Since much of Python itself is coded in highly optimized C (as are many of its standard library and extension modules), your application may even turn out to already be faster than typical C code. However, if the application is too slow, you need, first and foremost, to rethink your algorithms and data structures. Check for bottlenecks due to application architecture, network traffic, database access, and operating system interactions. For typical applications, each of these factors is more likely than language choice to cause slowdowns. Tinkering with large-scale architectural aspects can often dramatically speed up an application, and Python is an excellent medium for such experimentation.

If your program is still too slow, profile it to find out where the time is going. Applications often exhibit computational bottlenecks: small areas of the source code—often 20% or less of it—account for 80% or more of the running time. Optimize the bottlenecks, applying the techniques suggested in the rest of this chapter.

If normal Python-level optimizations still leave some outstanding computational bottlenecks, you can recode those as Python extension modules, as covered in Chapter 24. In the end, your application runs at roughly the same speed as if you had coded it all in C, C++, or Fortran—or faster, when large-scale experimentation has let you find a better architecture. Your overall programming productivity with this process is not much less than if you coded everything in Python. Future changes and maintenance are easy, since you use Python to express the overall structure of the program, and lower-level, harder-to-maintain languages for only a few specific computational bottlenecks.

As you build applications in a given area following this process, you accumulate a library of reusable Python extension modules. You therefore become more and more productive at developing other fast-running Python applications in the same field.

Even if external constraints eventually force you to recode the whole application in a lower-level language, you're still better off for having started in Python. Rapid prototyping has long been acknowledged as the best way to get software architecture just right. A working prototype lets you check that you have identified the right problems and taken a good path to their solution. A prototype also affords the kind of large-scale architectural experiments that can make a real difference in performance. Starting your prototype with Python allows a gradual migration to other languages by way of extension modules, if need be. The application remains fully functional and testable at each stage. This ensures against the risk of compromising a design's architectural integrity in the coding stage. The resulting software is faster and more robust than if all of the coding had been lower-level from the start, and your productivity—while not quite as good as with a pure Python application—is still better than if you had been coding at a lower level throughout.

Benchmarking

Benchmarking (also known as *load testing*) is similar to system testing: both activities are much like running the program for production purposes. In both cases, you need to have at least some subset of the program's intended functionality working, and you need to use known, reproducible inputs. For benchmarking, you don't need to capture and check your program's output: since you make it work and make it right before you make it fast, you're already fully confident about your program's correctness by the time you load-test it. You do need inputs that are representative of typical system operations, ideally ones that may be most challenging for your program's performance. If your program performs several kinds of operations, make sure you run some benchmarks for each different kind of operation.

Elapsed time as measured by your wristwatch is probably precise enough to benchmark most programs. Programs with hard real-time constraints are another matter, but they have needs very different from those of normal programs in most respects. A 5 or 10 percent difference in performance, except in programs with very peculiar constraints, makes no practical difference to a program's real-life usability.

When you benchmark "toy" programs or snippets in order to help you choose an algorithm or data structure, you may need more precision: the `timeit` module of Python's standard library (covered in "The timeit module") is quite suitable for such tasks. The benchmarking discussed in this section is of a different kind: it is an approximation of real-life program operation for the sole purpose of checking whether the program's performance at each task is acceptable, before embarking on profiling and other optimization activities. For such "system" benchmarking, a situation that approximates the program's normal operating conditions is best, and high accuracy in timing is not all that important.

Large-Scale Optimization

The aspects of your program that are most important for performance are large-scale ones: your choice of overall architecture, algorithms, and data structures.

The performance issues that you must often take into account are those connected with the traditional big-O notation of computer science. Informally, if you call `N` the input size of an algorithm, big-O notation expresses algorithm performance, for large values of `N`, as proportional to some function of `N`. (In precise computer science lingo, this should be called big-Theta, but in real life, programmers call this big-O, perhaps because an uppercase Theta looks like an O with a dot in the center!)

An `O(1)` algorithm (also known as "constant time") is one that takes a time not growing with `N`. An `O(N)` algorithm (also known as "linear time") is one where, for large enough `N`, handling twice as much data takes about twice as much time, three times as much data three times as much time, and so on, proportionally to `N`. An $O(N^2)$ algorithm (also known as a "quadratic time" algorithm) is one where, for large enough `N`, handling twice as much data takes about four times as much time, three times as much data nine times as much time, and so on, growing proportionally to `N` squared. Identical concepts and notation are used to describe a program's consumption of memory ("space") rather than of time.

To find more information on big-O notation, and about algorithms and their complexity, any good book about algorithms and data structures can help; we recommend Magnus Lie Hetland's excellent book *Python Algorithms: Mastering Basic Algorithms in the Python Language*, 2nd edition (Apress, 2014).

To understand the practical importance of big-O considerations in your programs, consider two different ways to accept all items from an input iterable and accumulate them into a list in reverse order:

```
def slow(it):
    result = []
    for item in it: result.insert(0, item
)
    return result

def fast(it):
    result = []
    for item in it: result.append(item)
    result.reverse()
    return result
```

We could express each of these functions more concisely, but the key difference is best appreciated by presenting the functions in these elementary terms. The function `slow` builds the result list by inserting each input item before all previously received ones. The function `fast` appends each input item after all previously received ones, then reverses the result list at the end. Intuitively, one might think that the final reversing represents extra work, and therefore `slow` should be faster than `fast`. But that's not the way things work out.

Each call to `result.append` takes roughly the same amount of time, independent of how many items are already in the list `result`, since there is (nearly) always a free slot for an extra item at the end of the list (in pedantic terms, `append` is *amortized* `O(1)`, but we don't cover amortization in this book). The `for` loop in the function `fast` executes `N` times to receive `N` items. Since each iteration of the loop takes a constant time, overall loop time is `O(N)`. `result.reverse` also takes time `O(N)`, as it is directly proportional to the total number of items. Thus, the total running time of `fast` is `O(N)`. (If you don't understand why a sum of two quantities, each `O(N)`, is also `O(N)`, consider that the sum of any two linear functions of `N` is also a linear function of `N`—and "being `O(N)`" has exactly the same meaning as "consuming an amount of time that is a linear function of `N`.")

On the other hand, each call to `result.insert` makes space at slot `0` for the new item to insert, moving all items that are already in list `result` forward one slot. This takes time proportional to the number of items already in the list. The overall amount of time to receive `N` items is therefore proportional to `1+2+3+...N-1`, a sum whose value is $O(N^2)$. Therefore, total running time of `slow` is $O(N^2)$.

It's almost always worth replacing an $O(N^2)$ solution with an `O(N)` one, unless you can somehow assign rigorous small limits to input size `N`. If `N` can grow without very strict bounds, the $O(N^2)$ solution turns out to be disastrously slower than the `O(N)` one for large values of `N`, no matter what the proportionality constants in each case may be (and, no matter what profiling tells you). Unless you have other $O(N^2)$ or even worse bottlenecks elsewhere that you can't eliminate, a part of the program that is $O(N^2)$ turns into the program's bottleneck, dominating runtime for large values of `N`. Do yourself a favor and watch out for the big O: all other performance issues, in comparison, are usually almost insignificant.

Incidentally, you can make the function `fast` even faster by expressing it in more idiomatic Python. Just replace the first two lines with the following single statement:

```
result = list(it)
```

This change does not affect `fast`'s big-O character (`fast` is still `O(N)` after the change), but does speed things up by a large constant factor.

Simple is better than complex, and usually faster!

More often than not, in Python, the simplest, clearest, most direct and idiomatic way to express something is also the fastest.

Choosing algorithms with good big-O is roughly the same task in Python as in any other language. You just need a few hints about the big-O performance of Python's elementary building blocks, and we provide them in the following sections.

List operations

Python lists are internally implemented as *vectors* (also known as *dynamic arrays*), not as "linked lists." This implementation choice determines just about all performance characteristics of Python lists, in big-O terms.

Chaining two lists `L1` and `L2`, of length `N1` and `N2` (i.e., `L1+L2`) is `O(N1+N2)`. Multiplying a list `L` of length `N` by integer `M` (i.e., `L*M`) is `O(N*M)`. Accessing or rebinding any list item is `O(1)`. `len()` on a list is also `O(1)`. Accessing any slice of length `M` is `O(M)`. Rebinding a slice of length `M` with one of identical length is also `O(M)`. Rebinding a slice of length `M1` with one of different length `M2` is `O(M1+M2+N1)`, where `N1` is the number of items *after* the slice in the target list (in other words, such length-changing slice rebindings are relatively cheap when they occur at the *end* of a list, more costly when they occur at the *beginning* or around the middle of a long list). If you need first-in, first-out (FIFO) operations, a list is probably not the fastest data structure for the purpose: instead, try the type `collections.deque`, covered in "deque".

Most list methods, as shown in Table 3-3, are equivalent to slice rebindings and have equivalent big-O performance. The methods `count`, `index`, `remove`, and `reverse`, and the operator `in`, are `O(N)`. The method `sort` is generally `O(N*log(N))`, but is highly optimized to be `O(N)` in some important special cases, such as when the list is already sorted or reverse-sorted except for a few items. `range(a,b,c)` in v2 is `O((b-a)/c)`. `xrange(a,b,c)` in v2, and `range` in v3, is `O(1)`, but looping on all items of the result is `O((b-a)/c)`.

String operations

Most methods on a string of length `N` (be it bytes or Unicode) are `O(N)`. `len(astring)` is `O(1)`. The fastest way to produce a copy of a string with transliterations and/or removal of specified characters is the string's method `translate`. The single most practically important big-O consideration involving strings is covered in "Building up a string from pieces".

Dictionary operations

Python `dict`s are implemented with hash tables. This implementation choice determines all performance characteristics of Python dictionaries, in big-O terms.

Accessing, rebinding, adding, or removing a dictionary item is `O(1)`, as are the methods `has_key`, `get`, `setdefault`, and `popitem`, and operator `in`. `d1.update(d2)` is `O(len(d2))`. `len(adict)` is `O(1)`. The methods `keys`, `items`, and `values` are `O(N)` in v2. The same methods in v3, and the methods `iterkeys`, `iteritems`, and `itervalues` in v2, are `O(1)`, but looping on all items of the iterators those methods return is `O(N)`, and looping directly on a `dict` has the same big-O performance as v2's `iterkeys`.

Never code if x in d.keys():

Never test `if x in d.keys():`! That's `O(N)`, while the equivalent test `if x in d:` is `O(1)` (`if d.has_key(x):` is also `O(1)`, but is slower than `if x in d:`, has no compensating advantage, and is deprecated—so, never use it either).

When the keys in a dictionary are instances of classes that define `__hash__` and equality comparison methods, dictionary performance is of course affected by those methods. The performance indications presented in this section hold when hashing and equality comparison are `O(1)`.

Set operations

Python sets, like `dict`s, are implemented with hash tables. All performance characteristics of sets are, in big-O terms, the same as for dictionaries.

Adding or removing a set item is `O(1)`, as is operator `in`. `len(aset)` is `O(1)`. Looping on a set is `O(N)`. When the items in a set are instances of classes that define `__hash__` and equality comparison methods, set performance is of course affected by those methods. The performance hints presented in this section hold when hashing and equality comparison are `O(1)`.

Summary of big-O times for operations on Python built-in types

Let `L` be any list, `T` any string (plain/bytes or Unicode), `D` any dict, `S` any set, with (say) numbers as items (just for the purpose of ensuring `O(1)` hashing and comparison), and `x` any number (ditto):

`O(1)`

: `len(L)`, `len(T)`, `len(D)`, `len(S)`, `L[i]`, `T[i]`, `D[i]`, `del D[i]`, `if x in D`, `if x in S`, `S.add(x)`, `S.remove(x)`, appends or removals to/from the very right end of `L`

`O(N)`

: Loops on `L`, `T`, `D`, `S`, general appends or removals to/from `L` (except at the very right end), all methods on `T`, `if x in L`, `if x in T`, most methods on `L`, all shallow copies

`O(N log N)`

: `L.sort`, mostly (but `O(N)` if `L` is already nearly sorted or reverse-sorted)

Profiling

Most programs have hot spots (i.e., relatively small regions of source code that account for most of the time elapsed during a program run). Don't try to guess where your program's hot spots are: a programmer's intuition is notoriously unreliable in this field. Instead, use the module `profile` to collect profile data over one or more runs of your program, with known inputs. Then use the module `pstats` to collate, interpret, and display that profile data.

To gain accuracy, you can calibrate the Python profiler for your machine (i.e., determine what overhead profiling incurs on your machine). The `profile` module can then subtract this overhead from the times it measures so that the profile data you collect is closer to reality. The standard library module `cProfile` has similar functionality to `profile`; `cProfile` is preferable, since it's faster, which imposes less overhead. Yet another profiling module in Python's standard library (v2 only) is `hotshot`; unfortunately, `hotshot` does not support threads, nor v3.

The profile module

The `profile` module supplies one often-used function:

run `run(code,filename=None)`

`code` is a string that is usable with `exec`, normally a call to the main function of the program you're profiling. `filename` is the path of a file that `run` creates or rewrites with profile data. Usually, you call `run` a few times, specifying different filenames, and different arguments to your program's main function, in order to exercise various program parts in proportion to what you expect to be their use "in real life." Then, you use module `pstats` to display collated results across the various runs.

You may call `run` without a `filename` to get a summary report, similar to the one the `pstats` module could give you, on standard output. However, this approach gives no control over the output format, nor any way to consolidate several runs into one report. In practice, you should rarely use this feature: it's best to collect profile data into files.

The `profile` module also supplies the class `Profile` (mentioned in the next section). By instantiating `Profile` directly, you can access advanced functionality, such as the ability to run a command in specified local and global dictionaries. We do not cover such advanced functionality of the class `profile.Profile` further in this book.

Calibration

To calibrate `profile` for your machine, you need to use the class `Profile`, which `profile` supplies and internally uses in the function `run`. An instance `p` of `Profile` supplies one method you use for calibration:

calibrate `p.calibrate(N)`

Loops `N` times, then returns a number that is the profiling overhead per call on your machine. `N` must be large if your machine is fast. Call `p.calibrate(10000)` a few times and check that the various numbers it returns are close to each other, then pick the smallest one of them. If the numbers vary a lot, try again with larger values of `N`.

The calibration procedure can be time-consuming. However, you need to perform it only once, repeating it only when you make changes that could alter your machine's characteristics, such as applying patches to your operating system, adding memory, or changing Python version. Once you know your machine's overhead, you can tell `profile` about it each time you import it, right before using `profile.run`. The simplest way to do this is as follows:

```
import profileprofile.Profile.bias = ...the overhead you measured...profile.
run('main()', 'somefile')
```

The pstats module

The `pstats` module supplies a single class, `Stats`, to analyze, consolidate, and report on the profile data contained in one or more files written by the function `profile.run`:

Stats `class Stats(      filename,*filenames)`

Instantiates `Stats` with one or more filenames of files of profile data written by function `profile.run`.

An instance `s` of the class `Stats` provides methods to add profile data and sort and output results. Each method returns `s`, so you can chain several calls in the same expression. `s`'s main methods are described in Table 16-2.

Table 16-2.

add	`s.add(filename)` Adds another file of profile data to the set that `s` is holding for analysis.
print_callees, print_callers	`s.print_callees(*restrictions)` Outputs the list of functions in `s`'s profile data, sorted according to the latest call to `s.sort_stats` and subject to given restrictions, if any. You can call each printing method with zero or more `restrictions`, to be applied one after the other, in order, to reduce the number of output lines. A restriction that is an `int n` limits the output to the first `n` lines. A restriction that is a `float f` between `0.0` and `1.0` limits the output to a fraction `f` of the lines. A restriction that is a string is compiled as a regular expression pattern (covered in "Regular Expressions and the re Module"); only lines that satisfy a `search` method call on the regular expression are output. Restrictions are cumulative. For example, `s.print_callees(10,0.5)` outputs the first 5 lines (half of 10). Restrictions apply only after the summary and header lines: the summary and header are output unconditionally. Each function `f` that is output is accompanied by the list of `f`'s callers (the functions that called `f`) or `f`'s callees (the functions that `f` called) according to the name of the method.
print_stats	`s.print_stats(*restrictions)` Outputs statistics about `s`'s profile data, sorted according to the latest call to `s.sort_stats` and subject to given restrictions, if any, as covered in **print_callees, print_callers**, above. After a few summary lines (date and time on which profile data was collected, number of function calls, and sort criteria used), the output—absent restrictions—is one line per function, with six fields per line, labeled in a header line. For each function `f`, `print_stats` outputs six fields: • Total number of calls to `f` • Total time spent in `f`, exclusive of other functions that `f` called • Total time per call to `f` (i.e., field `2` divided by field `1`) • Cumulative time spent in `f`, and all functions directly or indirectly called from `f` • Cumulative time per call to `f` (i.e., field `4` divided by field `1`) • The name of function `f`

sort_stats	`s.sort_stats(key*, keys)`	

Gives one or more keys on which to sort future output, in priority order. Each key is a string. The sort is descending for keys that indicate times or numbers, and alphabetical for key `'nfl'`. The most frequently used keys when calling `sort_stats` are:

`'calls'`

Number of calls to the function (like field `1` covered in **print_stats**, above)

`'cumulative'`

Cumulative time spent in the function and all functions it called (like field `4` covered in **print_stats**, above)

`'nfl'`

Name of the function, its module, and the line number of the function in its file (like field `6` covered in **print_stats**, above)

`'time'`

Total time spent in the function itself, exclusive of functions it called (like field `2` covered in **print_stats**, above)

strip_dirs `s.strip_dirs()`

Alters `s` by stripping directory names from all module names to make future output more compact. `s` is unsorted after `s.strip_dirs()`, and therefore you normally call `s.sort_stats` right after calling `s.strip_dirs`.

Small-Scale Optimization

Fine-tuning of program operations is rarely important. Tuning may make a small but meaningful difference in some particularly hot spot, but it is hardly ever a decisive factor. And yet, fine-tuning—in the pursuit of mostly irrelevant micro-efficiencies—is where a programmer's instincts are likely to lead. It is in good part because of this that most optimization is premature and best avoided. The most that can be said in favor of fine-tuning is that, if one idiom is *always* speedier than another when the difference is measurable, then it's worth your while to get into the habit of always using the speedier way.

Most often, in Python, if you do what comes naturally, choosing simplicity and elegance, you end up with code that has good performance as well as clarity and maintainability. In other words, "let Python do the work": when Python provides a simple, direct way to do a task, chances are that it's also the fastest way to perform that task. In a few cases, an approach that may not be intuitively preferable still offers performance advantages, as discussed in the rest of this section.

The simplest optimization is to run your Python programs using `python -O` or `-OO`. `-OO` makes little difference to performance compared to `-O`, but may save memory, as it removes docstrings from the bytecode, and memory is sometimes (indirectly) a performance bottleneck. The optimizer is not powerful in current releases of Python, but it may gain you performance advantages on the order of 5 percent, sometimes as large as 10 percent (potentially larger if you make use of `assert` statements and `if __debug__:` guards, as suggested in "The assert Statement"). The best aspect of `-O` is that it costs nothing—as long as your optimization isn't premature, of course (don't bother using `-O` on a program you're still developing).

The timeit module

The standard library module `timeit` is handy for measuring the precise performance of specific snippets of code. You can import `timeit` to use `timeit`'s functionality in your programs, but the simplest and most normal use is from the command line:

```
python -m timeit -s'setup statement(s)' 'statement(s) to be
timed'
```

The "setup statement" is executed only once, to set things up; the "statements to be timed" are executed repeatedly, to carefully measure the average time they take.

For example, say you're wondering about the performance of `x=x+1` versus `x+=1`, where `x` is an `int`. At a command prompt, you can easily try:

```
$ python -m timeit -s 'x=0' 'x=x+1'
1000000 loops, best of 3: 0.0416 usec per loop
$ python -m timeit -s 'x=0' 'x+=1'
1000000 loops, best of 3: 0.0406 usec per loop
```

and find out that performance is, to all intents and purposes, the same in both cases.

Building up a string from pieces

The single Python "anti-idiom" that's likeliest to kill your program's performance, to the point that you should *never* use it, is to build up a large string from pieces by looping on string concatenation statements such as `big_string +=piece`. Python strings are immutable, so each such concatenation means that Python must free the `M` bytes previously allocated for `big_string`, and allocate and fill `M+K` bytes for the new version. Doing this repeatedly in a loop, you end up with roughly $O(N^2)$ performance, where `N` is the total number of characters. More often than not, $O(N^2)$ performance where `O(N)` is available is a disaster—even though Python bends over backward to help with this specific, terrible but common, anti-pattern. On some platforms, things may be even bleaker due to memory fragmentation effects caused by freeing many areas of progressively larger sizes.

To achieve `O(N)` performance, accumulate intermediate pieces in a list, rather than build up the string piece by piece. Lists, unlike strings, are mutable, so appending to a list is `O(1)` (amortized). Change each occurrence of `big_string+=piece` into `temp_list.append(piece)`. Then, when you're done accumulating, use the following code to build your desired string result in `O(N)` time:

```
big_string = ''.join(temp_list)
```

Using a list comprehension, generator expression, or other direct means (such as a call to `map`, or use of the standard library module `itertools`) to build `temp_list` may often offer further (substantial, but not big-O) optimization over repeated calls to `temp_list.append`. Other `O(N)` ways to build up big strings, which some Python programmers find more readable, are to concatenate the pieces to an instance of `array.array('u')` with the array's `extend` method, use a `bytearray`, or write the pieces to an instance of `io.TextIO` or `io.BytesIO`.

In the special case where you want to output the resulting string, you may gain a further small slice of performance by using `writelines` on `temp_list` (never building `big_string` in memory). When feasible (i.e., when you have the output file object open and available in the loop, and the file is buffered), it's just as effective to perform a `write` call for each `piece`, without any accumulation.

Although not nearly as crucial as `+=` on a big string in a loop, another case where removing string concatenation may give a slight performance improvement is when you're concatenating several values in an expression:

```
                ' eggs and           ' slices of     '
oneway = str(x)+'            +str(y)+'            +k+ham'
another = '{} eggs and {} slices of {} ham'.format(x, y, k)
```

Using the `format` method to format strings is often a good performance choice, as well as being more idiomatic and thereby clearer than concatenation approaches.

Searching and sorting

The operator `in`, the most natural tool for searching, is `O(1)` when the righthand side operand is a `set` or `dict`, but `O(N)` when the righthand side operand is a string, list, or tuple. If you must perform many searches on a container, you're much better off using a `set` or `dict`, rather than a list or tuple, as the container. Python `set`s and `dict`s are highly optimized for searching and fetching items by key. Building the `set` or `dict` from other containers, however, is `O(N)`, so, for this crucial optimization to be worthwhile, you must be able to hold on to the `set` or `dict` over several searches, possibly altering it apace as the underlying sequence changes.

The method `sort` of Python lists is also a highly optimized and sophisticated tool. You can rely on `sort`'s performance. Performance dramatically degrades, however, if, in v2, you pass `sort` a custom callable to perform comparisons (in order to sort a list based on custom comparisons). Most functions and methods that perform comparisons accept a `key=` argument to determine how, exactly, to compare items. If you only have a function suitable as a `cmp` argument, you can use `functools.cmp_to_key`, covered in Table 7-4, to build from it a function suitable as the `key` argument, and pass the new function thus built as the `key=` argument, instead of passing the original function as the `cmp=` argument.

However, most functions in the module `heapq`, covered in "The heapq Module", do not accept a `key=` argument. In such cases, you can use the *decorate-sort-undecorate (DSU)* idiom, covered in "The Decorate-Sort-Undecorate Idiom". (Heaps are well worth keeping in mind, since in some cases they can save you from having to perform sorting on all of your data.)

The `operator` module supplies the functions `attrgetter` and `itemgetter` that are particularly suitable to support the `key` approach, avoiding slow `lambda`s.

Avoid exec and from … import *

Code in a function runs faster than code at the top level in a module, because access to a function's local variables is very fast. If a function contains an `exec` without explicit dictionaries, however, the function slows down. The presence of such an `exec` forces the Python compiler to avoid the modest but important optimization it normally performs regarding access to local variables, since the `exec` might alter the function's namespace. A `from` statement of the form:

```
from MyModule import *
```

wastes performance, too, since it also can alter a function's namespace unpredictably, and therefore inhibits

Python’s local-variable optimizations.

`exec` itself is also quite slow, and even more so if you apply it to a string of source code rather than to a code object. By far the best approach—for performance, for correctness, and for clarity—is to avoid `exec` altogether. It’s most often possible to find better (faster, more robust, and clearer) solutions. If you *must* use `exec`, *always* use it with explicit `dict`s. If you need to `exec` a dynamically obtained string more than once, `compile` the string just once and then repeatedly `exec` the resulting code object. But avoiding `exec` altogether is *far* better, if at all feasible.

`eval` works on expressions, not on statements; therefore, while still slow, it avoids some of the worst performance impacts of `exec`. With `eval`, too, you’re best advised to use explicit `dict`s. If you need several evaluations of the same dynamically obtained string, `compile` the string once and then repeatedly `eval` the resulting code object. Avoiding `eval` altogether is even better.

See “Dynamic Execution and exec” for more details and advice about `exec`, `eval`, and `compile`.

Optimizing loops

Most of your program’s bottlenecks will be in loops, particularly nested loops, because loop bodies execute repeatedly. Python does not implicitly perform any *code hoisting*: if you have any code inside a loop that you could execute just once by hoisting it out of the loop, and the loop is a bottleneck, hoist the code out yourself. Sometimes the presence of code to hoist may not be immediately obvious:

```
def slower(anobject, ahugenumber):
    for i in range(ahugenumber): anobject.amethod(i
)
def faster(anobject, ahugenumber):
    themethod = anobject.amethod
    for i in range(ahugenumber): themethod(i)
```

In this case, the code that `faster` hoists out of the loop is the attribute lookup `anobject.amethod`. `slower` repeats the lookup every time, while `faster` performs it just once. The two functions are not 100 percent equivalent: it is (barely) conceivable that executing `amethod` might cause such changes on `anobject` that the next lookup for the same named attribute fetches a different method object. This is part of why Python doesn’t perform such optimizations itself. In practice, such subtle, obscure, and tricky cases happen very rarely; you’re safe in performing such optimizations to squeeze the last drop of performance out of some bottleneck.

Python is faster with local variables than with global ones. If a loop repeatedly accesses a global whose value does not change between iterations, cache the value in a local variable, and access the local instead. This also applies to built-ins:

```
def slightly_slower(asequence, adict):
    for x in asequence: adict[x] = hex(x)
def slightly_faster(asequence, adict):
    myhex = hex
    for x in asequence: adict[x] = myhex(x
)
```

Here, the speedup is very modest, on the order of 5 percent or so.

Do not cache `None`. `None` is a keyword, so no further optimization is needed.

List comprehensions and generator expressions can be faster than loops, and, sometimes, so can `map` and `filter`. For optimization purposes, try changing loops into list comprehensions, generator expressions, or perhaps `map` and `filter` calls, where feasible. The performance advantage of `map` and `filter` is nullified, and worse, if you have to use a `lambda` or an extra level of function call. Only when you pass to `map` or `filter` a built-in function, or a function you'd have to call anyway even from an explicit loop, list comprehension, or generator expression, do you stand to gain some tiny speed-up.

The loops that you can replace most naturally with list comprehensions, or `map` and `filter` calls, are ones that build up a list by repeatedly calling `append` on the list. The following example shows this optimization in a micro-performance benchmark script (of course, we could use the module `timeit` instead of coding our own time measurement, but the example is meant to show how to do the latter):

```
import time, operator

def slow(asequence):
    result = []
    for x in asequence: result.append(-x)
    return result

def middling(asequence):
    return list(map(operator.neg, asequence))

def fast(asequence):
    return [-x for x in asequence]

biggie = range(500*1000)
tentimes = [None]*10
def timit(afunc):
    lobi = biggie
    start = time.clock()
    for x in tentimes: afunc(lobi)
    stend = time.clock()
    return '{:<10}: {:.2f}'.format(afunc.__name__, stend-start
)

for afunc in slow, middling, fast, fast, middling, slow:
    print(timit(afunc))
```

Running this example in v2 on an old laptop shows that `fast` takes about 0.36 seconds, `middling` 0.43 seconds, and `slow` 0.77 seconds. In other words, on that machine, `slow` (the loop of `append` method calls) is about 80 percent slower than `middling` (the single `map` call), and `middling`, in turn, is about 20 percent slower than `fast` (the list comprehension).

The list comprehension is the most direct way to express the task being micro-benchmarked in this example, so, not surprisingly, it's also fastest—about two times faster than the loop of `append` method calls.

Optimizing I/O

If your program does substantial amounts of I/O, it's likely that performance bottlenecks are due to I/O, not to computation. Such programs are said to be *I/O-bound*, rather than *CPU-bound*. Your operating system tries to optimize I/O performance, but you can help it in a couple of ways. One such way is to perform your I/O in chunks of a size that is optimal for performance, rather than simply convenient for your program's operations. Another way is to use threading. Often the very best way is to "go asynchronous," as covered in Chapter 18.

From the point of view of a program's convenience and simplicity, the ideal amount of data to read or write at a time is often small (one character or one line) or very large (an entire file at a time). That's often okay: Python and your operating system work behind the scenes to let your program use convenient logical chunks for I/O, while arranging for physical I/O operations to use chunk sizes more attuned to performance. Reading and writing a whole file at a time is quite likely to be okay for performance as long as the file is not *very* large. Specifically, file-at-a-time I/O is fine as long as the file's data fits very comfortably in physical RAM, leaving ample memory available for your program and operating system to perform whatever other tasks they're doing at the same time. The hard problems of I/O-bound performance tend to come with huge files.

If performance is an issue, *never* use a file's `readline` method, which is limited in the amount of chunking and buffering it can perform. (Using `writelines`, on the other hand, gives no performance problem when that method is convenient for your program.) When reading a text file, loop directly on the file object to get one line at a time with best performance. If the file isn't too huge, and so can conveniently fit in memory, time two versions of your program —one looping directly on the file object, the other reading the whole file into memory. Either may prove faster by a little.

For binary files, particularly large binary files whose contents you need just a part of on each given run of your program, the module `mmap` (covered in "The mmap Module") can sometimes give you both good performance and program simplicity.

Making an I/O-bound program multithreaded sometimes affords substantial performance gains, if you can arrange your architecture accordingly. Start a few worker threads devoted to I/O, have the computational threads request I/O operations from the I/O threads via `Queue` instances, and post the request for each input operation as soon as you know you'll eventually need that data. Performance increases only if there are other tasks your computational threads can perform while I/O threads are blocked waiting for data. You get better performance this way only if you can manage to overlap computation and waiting for data by having different threads do the computing and the waiting. (See "Threads in Python" for detailed coverage of Python threading and a suggested architecture.)

On the other hand, a possibly even faster and more scalable approach is to eschew threads in favor of asynchronous (event-driven) architectures, as covered in Chapter 18.

Terminology in this area is confused and confusing: terms like dummies, fakes, spies, mocks, stubs, and "test doubles" are used by different authors with different distinctions. For an authoritative approach (though not the exact one we use), see Martin Fowler's essay.

That's partly because the structure of the system tends to mirror the structure of the organization, per Conway's Law.

However, be sure you know exactly what you're using doctest for in any given case: to quote Peter Norvig, writing precisely on this subject: "know what you're aiming for; if you aim at two targets at once you usually miss them both."