Parallel Python

Embracing the Future with Sub-Interpreters and Free Threading

Shekhar Koirala

PyCon Ireland 2024

About me

- Working as ML Engineer at Identv LLC
- Studying Masters at Georgia Tech
- Volunteer at EuroPython and other Python/ PyData Conferences

Before …….

1. Don’t Blindly go with the Benchmarks
2. Play around with releases but wait for safety
3. Embrace the Incremental Process*

* https://peps.python.org/pep-0703/ [703]
* https://peps.python.org/pep-0684/ [684]
* https://peps.python.org/pep-0734/ [734]

Why care about Parallelism ?

Approaches for Concurrency and Parallelism

threading_example.py

```python
import threading
import time

def fibonacci(num):
    a, b = 0, 1
    for i in range(num):
        a, b = b, a + b
    return a

thread1 = threading.Thread(target=fibonacci, args=("Thread 1", 10))
thread2 = threading.Thread(target=fibonacci, args=("Thread 2", 10))

thread1.star
thread2.star

thread1.join
thread2.join
```

concurrent_example.py

```python
import concurrent.futures
import time

def fibonacci(num):
  a,b = 0,1
  for i in range(num):
    a,b = b, a+b
  return a

with concurrent.futures.ThreadPoolExecutor() as executor:
    futures = [
        executor.submit(fibonacci, "Thread 1", 10),
        executor.submit(fibonacci, "Thread 2", 10)
    ]

    for future in concurrent.futures.as_completed(futures):
        print(future.result())
```

multiprocessing_example.py

```python
from multiprocessing import Process, Array

def fibonacci(num, index, shared_array):
    a, b = 0, 1
    for _ in range(num):
        a, b = b, a + b
    shared_array[index] = a

if __name__ == "__main__":
    numbers = [10, 15, 20, 25]
    shared_array = Array('i', len(numbers))
    processes = []

    for i, num in enumerate(numbers):
        p = Pro
        process
        p.start

    for p in pr
        p.join(

    for i, num
        print(f
```

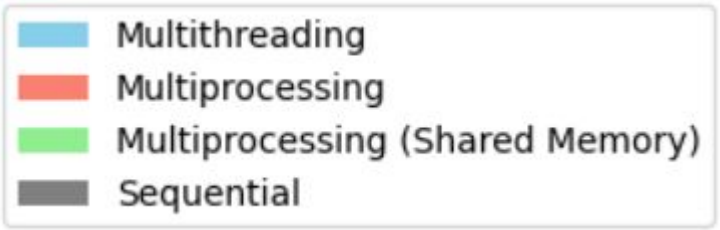
asyncio_example.py

```python
import asyncio

async def fibonacci(num):
    a, b = 0, 1
    for _ in range(num):
        a, b = b, a + b
    return a

async def main():
    tasks = [
        asyncio.create_task(fibonacci(20)),
        asyncio.create_task(fibonacci(20)),
    ]
    results = await asyncio.gather(*tasks)
    for i, result in enumerate(results, start=1):
        print(f"Result of Task {i}: Fibonacci = {result}")

asyncio.run(main())
```

Quiz Time

element_wise_operation.py

```python
arr = np.random.rand(int(1e8))

def process_large_numpy_arr(arr):
    return np.sin(arr) ** 2 + np.cos(arr) ** 2
```



shared_memory.py

```python
from multiprocessing import shared_memory

shm = shared_memory.SharedMemory(
                    create=True,
                    size=num_processes * np.finfo(np.float64).bits)

results = np.ndarray((num_processes,), dtype=np.float64,
                     buffer=shm.buf)
```

* issue: https://github.com/python/cpython/issues/82300

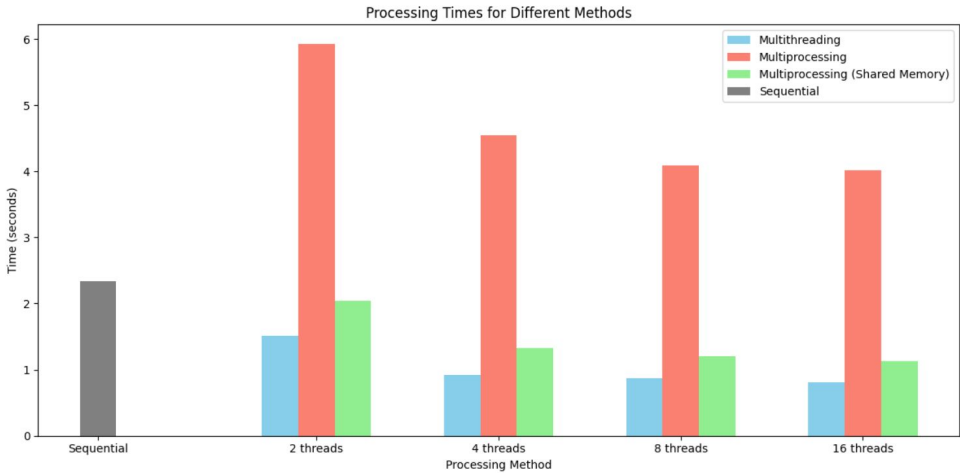
*https://medium.com/@turbopython/multithreading-vs-multiprocessing-with-numpy-which-one-is-faster-afebf027a0fa

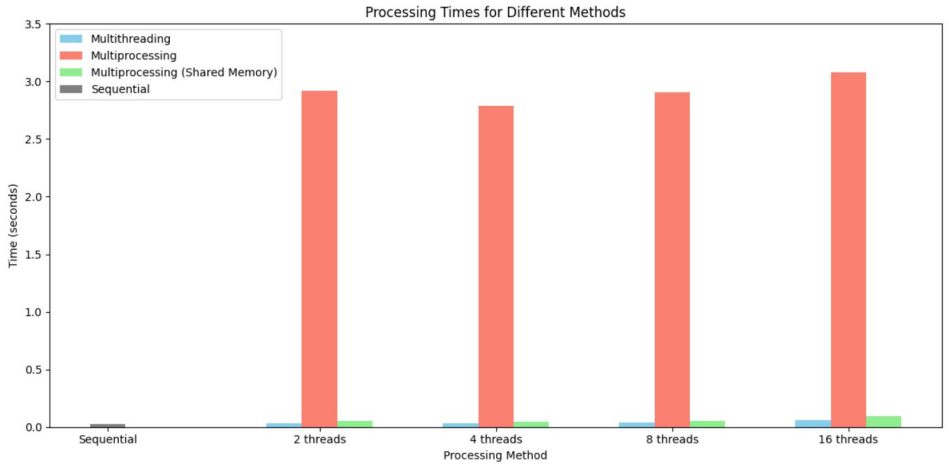
element_wise_operation.py

```python
arr = np.random.rand(int(1e8))

def process_large_numpy_arr(arr):
    return np.sin(arr) ** 2 + np.cos(arr) ** 2
```

vector_operation.py

```python
arr = np.random.rand(int(1e8))

def process_large_numpy_arr(arr):
    return np.linalg.norm(arr)
```





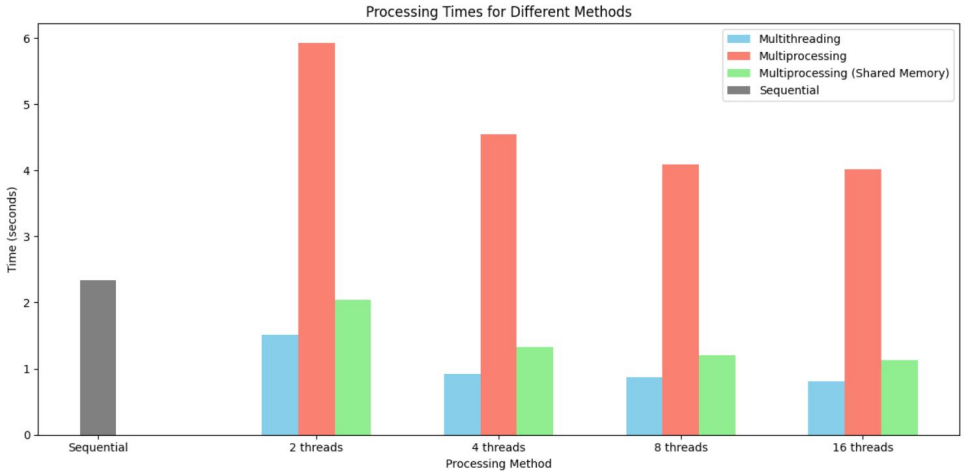
*https://medium.com/@turbopython/multithreading-vs-multiprocessing-with-numpy-which-one-is-faster-afebf027a0fa

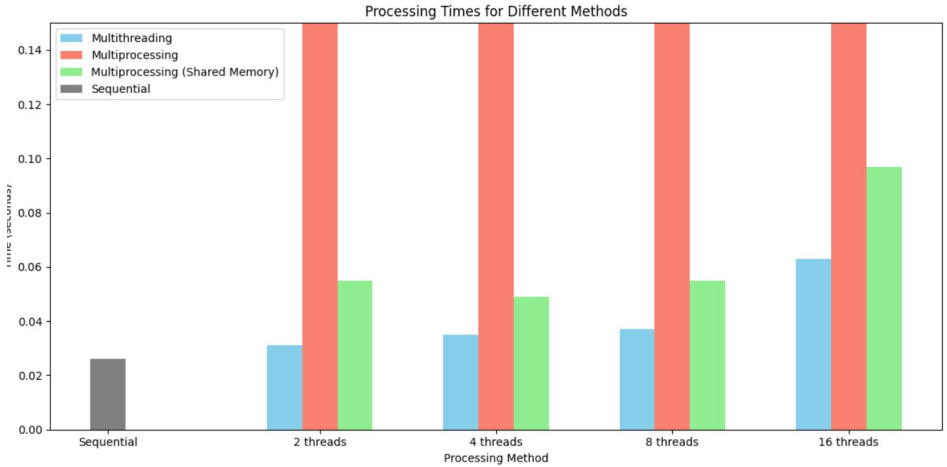
element_wise_operation.py

```python
arr = np.random.rand(int(1e8))

def process_large_numpy_arr(arr):
    return np.sin(arr) ** 2 + np.cos(arr) ** 2
```

vector_operation.py

```python
arr = np.random.rand(int(1e8))

def process_large_numpy_arr(arr):
    return np.linalg.norm(arr)
```





Why it happened ?

Note that operations that *do not* release the GIL will see no performance gains from use of the **threading** module,
and instead might be better served with **multiprocessing**.
In particular, operations on arrays with dtype=object do not release the GIL.

- * https://numpy.org/devdocs/reference/thread_safety.html

ndarraytypes.h

```
#define NPY_BEGIN_ALLOW_THREADS Py_BEGIN_ALLOW_THREADS
#define NPY_END_ALLOW_THREADS Py_END_ALLOW_THREADS
#define NPY_BEGIN_THREADS do {_save = PyEval_SaveThread();} while (0);
#define NPY_END_THREADS   do { if (_save) \
            { PyEval_RestoreThread(_save); _save = NULL;} } while (0);
#define NPY_BEGIN_THREADS_THRESHOLDED(loop_size) do { if ((loop_size) > 500) \
            { _save = PyEval_SaveThread();} } while (0);
```

cpython/socketModule.c

```
Py_BEGIN_ALLOW_THREADS
res = bind(s->sock_fd, SAS2SA(&addrbuf), addrlen);
Py_END_ALLOW_THREADS
```

Why it happened ?

src/arrow/python/common.h

```
// Same as OwnedRef, but ensures the GIL is taken when it goes out of scope.
// This is for situations where the GIL is not always known to be held
// (e.g. if it is released in the middle of a function for performance reasons)
class ARROW_PYTHON_EXPORT OwnedRefNoGIL : public OwnedRef {
 public:
  OwnedRefNoGIL() : OwnedRef() {}
  OwnedRefNoGIL(OwnedRefNoGIL&& other) : OwnedRef(other.detach()) {}
  explicit OwnedRefNoGIL(PyObject* obj) : OwnedRef(obj) {}

  ~OwnedRefNoGIL() {
    PyAcquireGIL lock;
    reset();
  }
};
```

ndarraytypes.h

```
#define NPY_BEGIN_ALLOW_THREADS Py_BEGIN_ALLOW_THREADS
#define NPY_END_ALLOW_THREADS Py_END_ALLOW_THREADS
#define NPY_BEGIN_THREADS do {_save = PyEval_SaveThread();} while (0);
#define NPY_END_THREADS   do { if (_save) \
            { PyEval_RestoreThread(_save); _save = NULL;} } while (0);
#define NPY_BEGIN_THREADS_THRESHOLDED(loop_size) do { if ((loop_size) > 500) \
            { _save = PyEval_SaveThread();} } while (0);
```

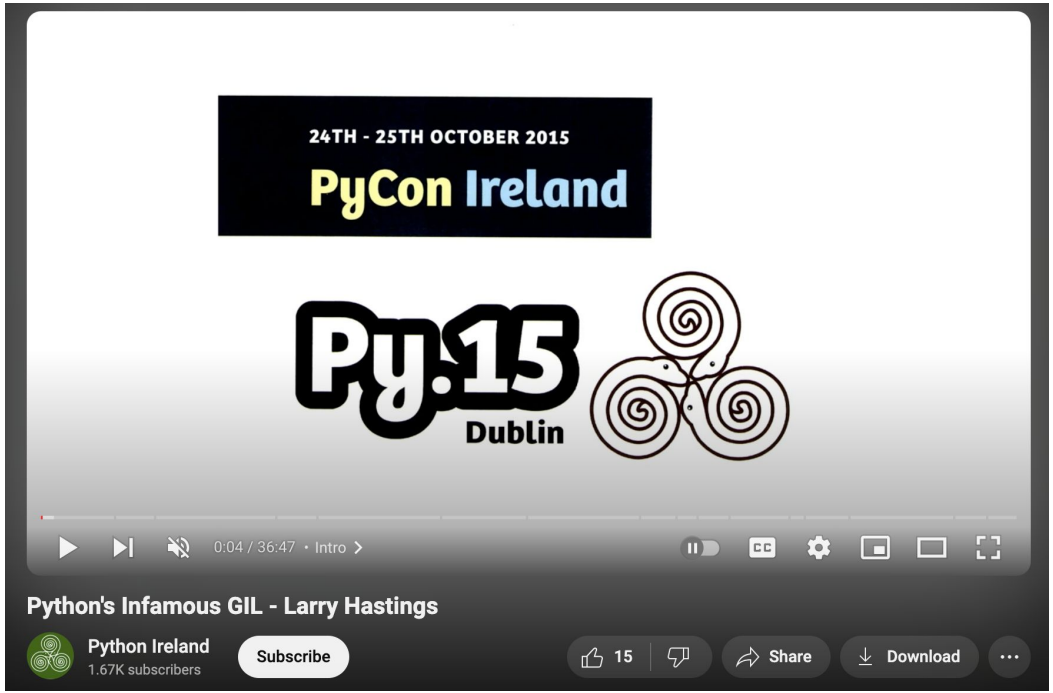
cpython/socketModule.c

```
Py_BEGIN_ALLOW_THREADS
res = bind(s->sock_fd, SAS2SA(&addrbuf), addrlen);
Py_END_ALLOW_THREADS
```

GIL

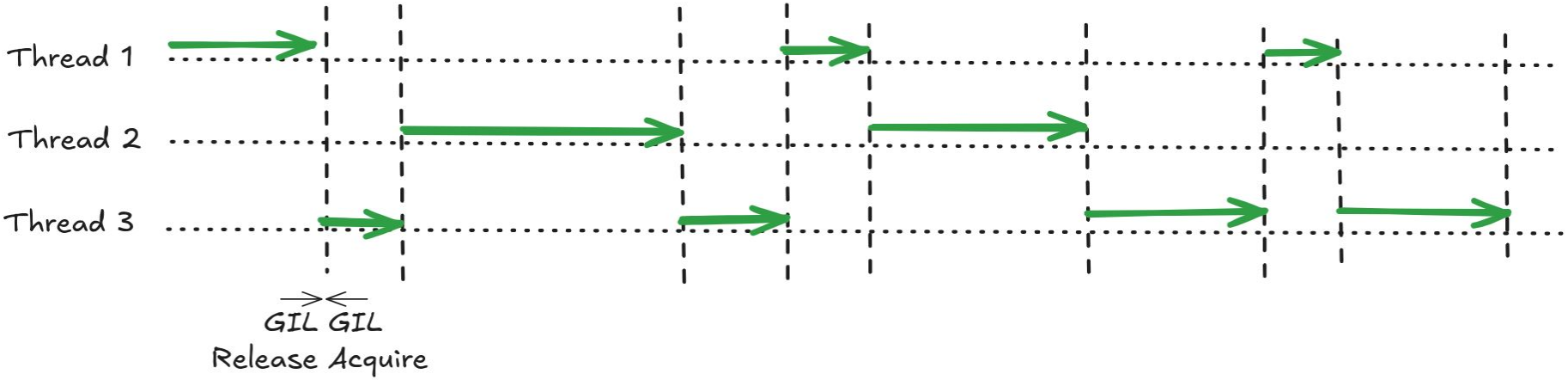
*https://www.youtube.com/watch?v=Obt-vMVdM8s

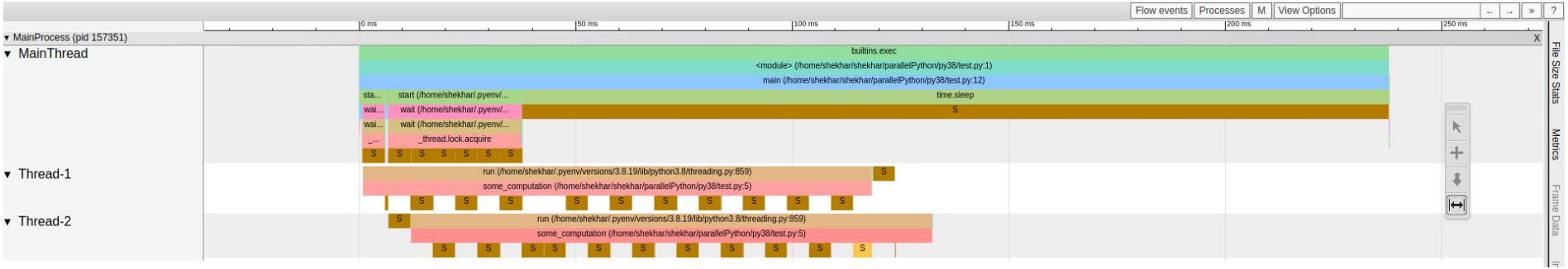
GIL



*https://www.youtube.com/watch?v=sxMl4DsYgpw

Visualizing GIL





*https://www.maartenbreddels.com/perf/jupyter/python/tracing/gil/2021/01/14/Tracing-the-Python-GIL.html

More on GIL

- Lock used by Python runtimes to protect Global states and variables
- Prevents race condition when running its ByteCode.
- Protect C extension module
- This was in the 90’s where many devices where single core.
- Gil was attempted to removed but it was slower than single thread.*

*https://www.youtube.com/watch?v=P3AyI_u66Bw

GIL was/is hard to remove

- Lots of C extensions assumes GIL exists
- ABI changes

* No GIL but use Locks or immortal objects* !!!

* https://peps.python.org/pep-0683/

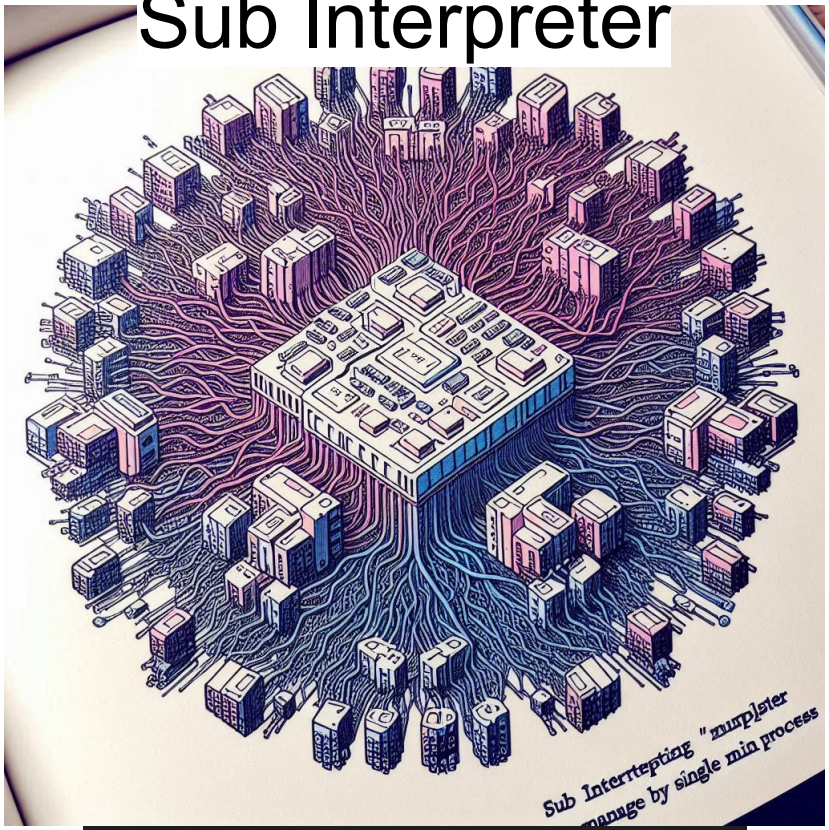
Other Parallelism Project

- CPython alternative : Jython / IronPython
- Gilectomy [“no-gil” project]
- PyParallel Project
- Other parallelism tool : Dask/ Taichi
- MultiProcessing
- Give Up Multi Core , ASYNCIO …. !!!

Free Threading

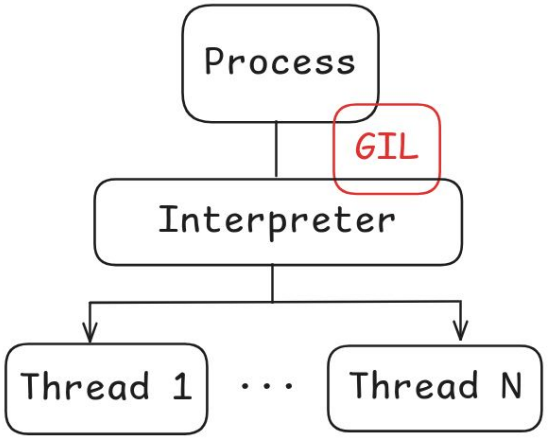
```
$ pyenv install 3.13.0t
```

Sub Interpreter

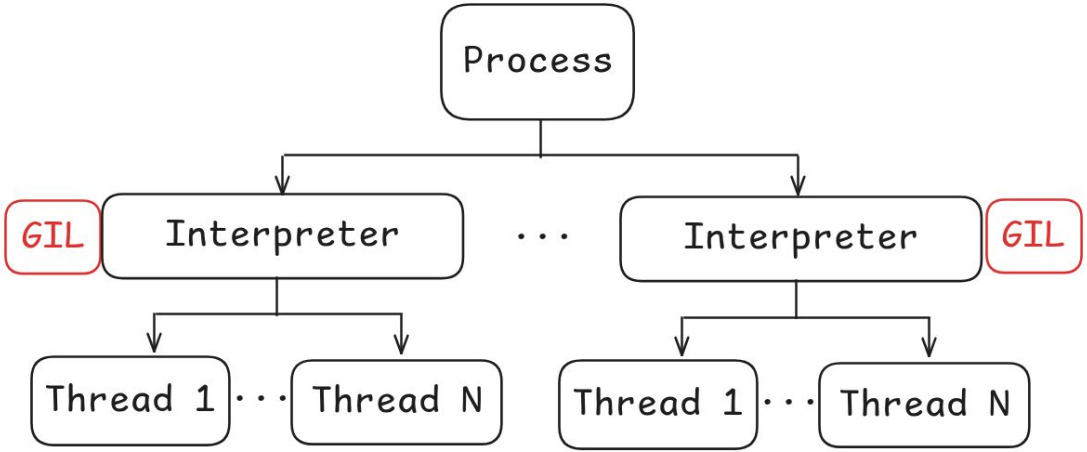


```
$ pip install interpreters-pep-734
```

Sub interpreters



Python <=3.11



*Python >=3.12

* https://youtu.be/z4MaqTbM_gY?si=VIwHv72WkwQCApxG Anthony Shaw

Differences

Feature	Multiprocessing	Sub Interpreter
Execution	Separate OS Processes	Multiple Interpreters within one process
Memory Isolation	Separate Memory per process	Same Memory space but separate Python states
Communication	Inter Process Communication [Pipe, Queue, Manager]	Shared Memory Access [Queues Channel] *
Overhead / creation	Higher to create	Lower to create [less memory footprint]

*https://www.youtube.com/watch?v=fwRMdncVOnA

Start-up Time in Sub Interpreter

Around 5x improvement compared to multiprocessing.*

```
multiprocessing.set_start_method("spawn", force=True)
```

Speed of starting up sub interpreters lies between the “fork” and “spawn” method.

*https://discuss.python.org/t/expected-performance-characteristics-of-subinterpreters/53251

Sharable Objects in Sub Interpreter

- str
- bytes
- int
- float
- bool (True/False)
- None
- tuple (only with shareable items)
- interpreters.Queue
- Memoryview (underlying buffer actually shared)

IMMUTABLE

→Synchronization using locks can be used for Thread safety.

Interpreter.get_current(“situation”)

Only in Python 3.12*

```
import _xxsubinterpreters as interpreters

interpreters.run('''
print("PyCon Ireland 2024 .... !!!")
''')
```

PEP 734 – Multiple Interpreters in the Stdlib

Status: Deferred

```
    def exec(self, code, /):
        """Run the given source code in the interpreter.

        This is essentially the same as calling the builtin "exec"
        with this interpreter, using the __dict__ of its __main__
        module as both globals and locals."""

    def call(self, callable, /):
        """Call the object in the interpreter with given args/kwargs.

        Only functions that take no arguments and have no closure are supported."""

    def call_in_thread(self, callable, /):
        """Return a new thread that calls the object in the interpreter.

        The return value and any raised exception are discarded.
        """
```

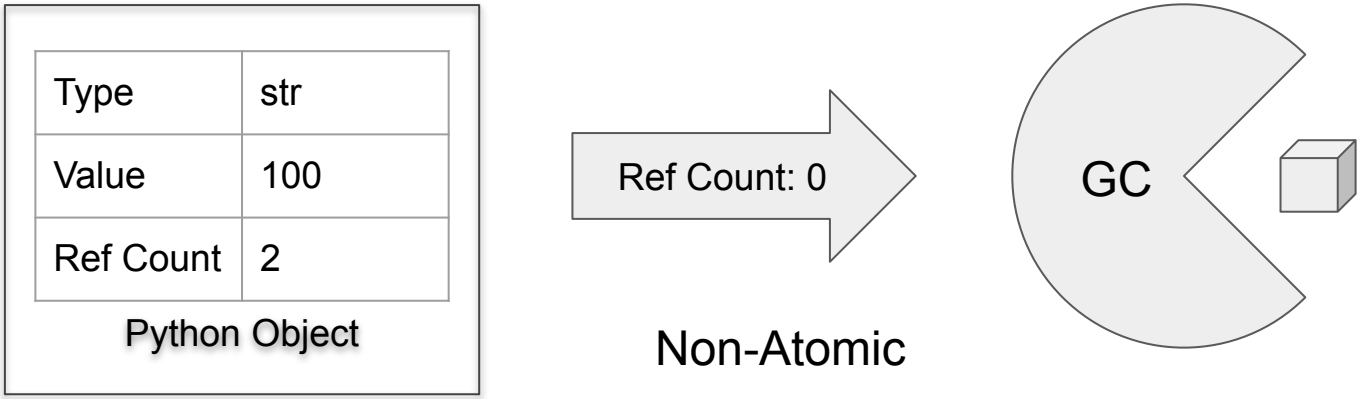
For Python 3.13*

```
$ pip install interpreters-pep-734
```

*will be added in
concurrent.futures.InterpreterPoolExecutor

Free Threading

Reference Counting



Biased reference count = sum(local ref count + shared ref count)

Free Threading

Reference Counting

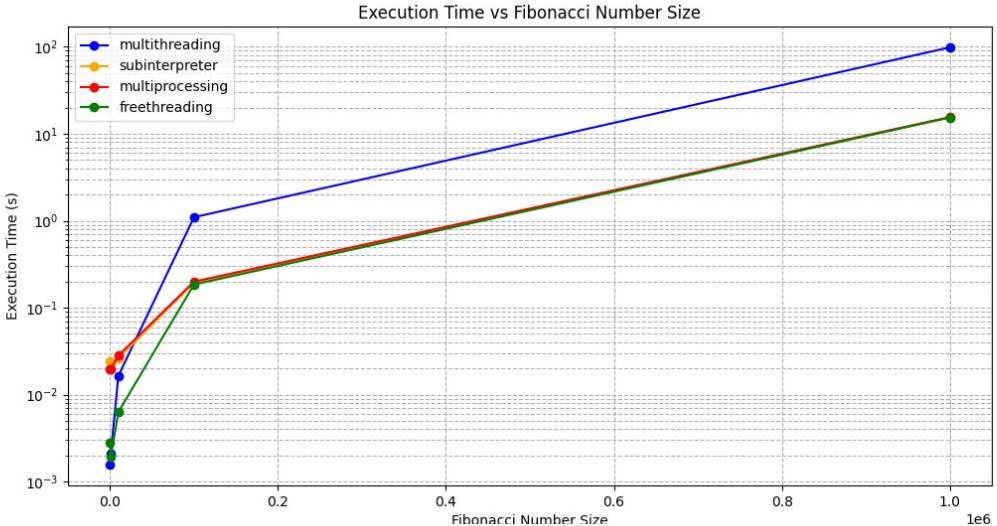
Memory Allocation

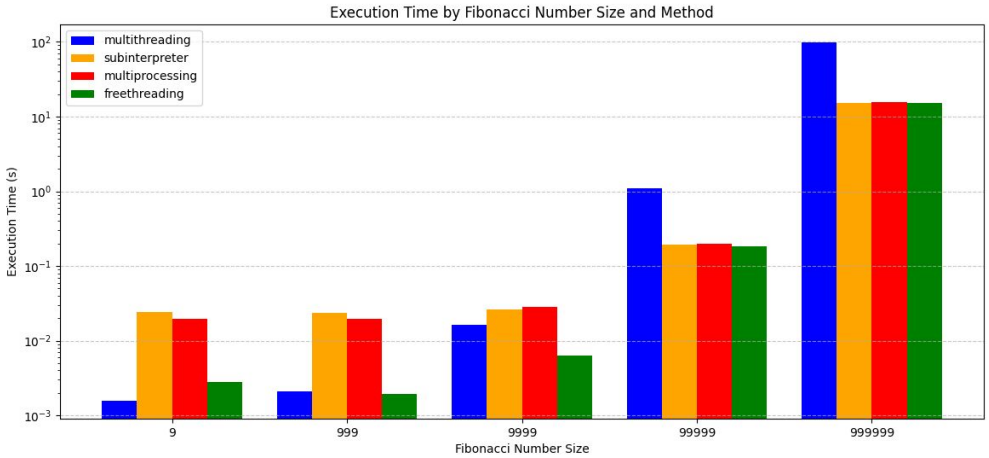
Thread Safety

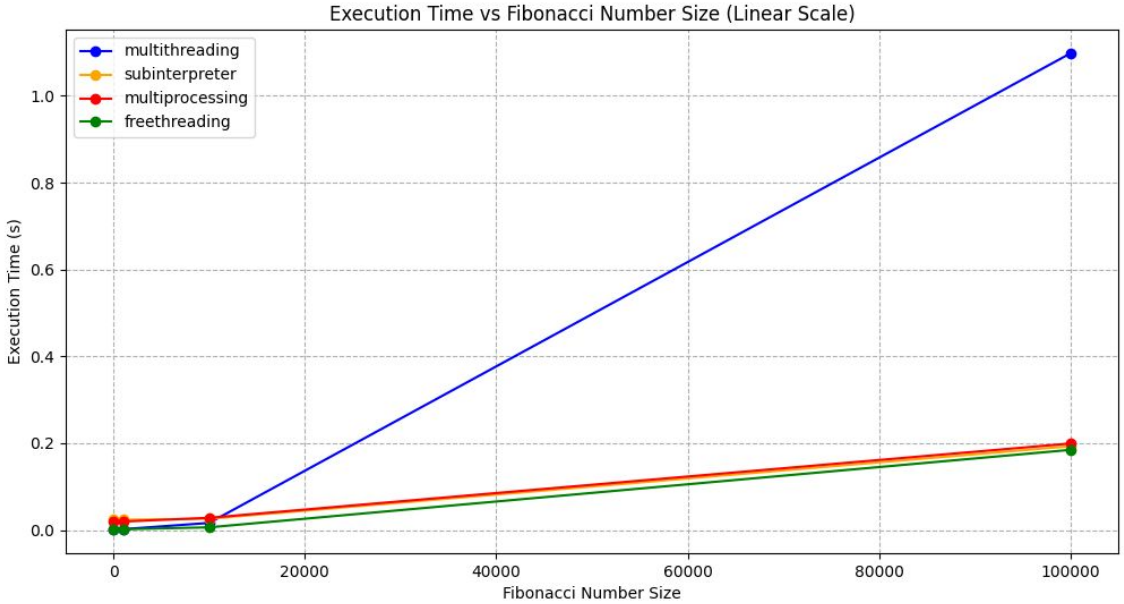
Special wheel and special build of Cpython

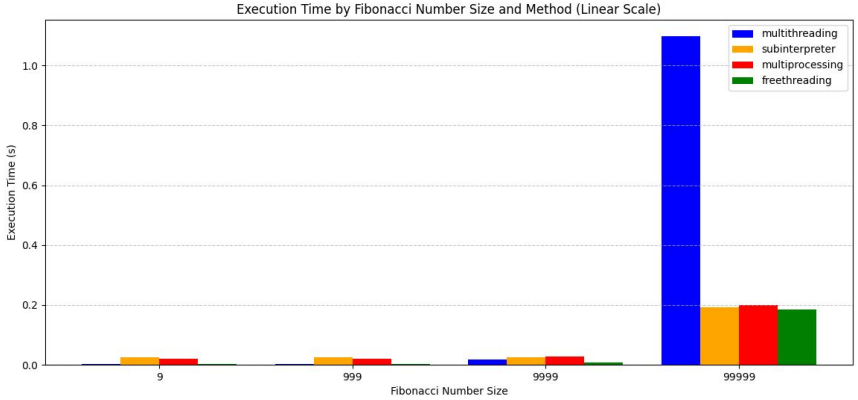
Introduces new ABI

Benchmarks

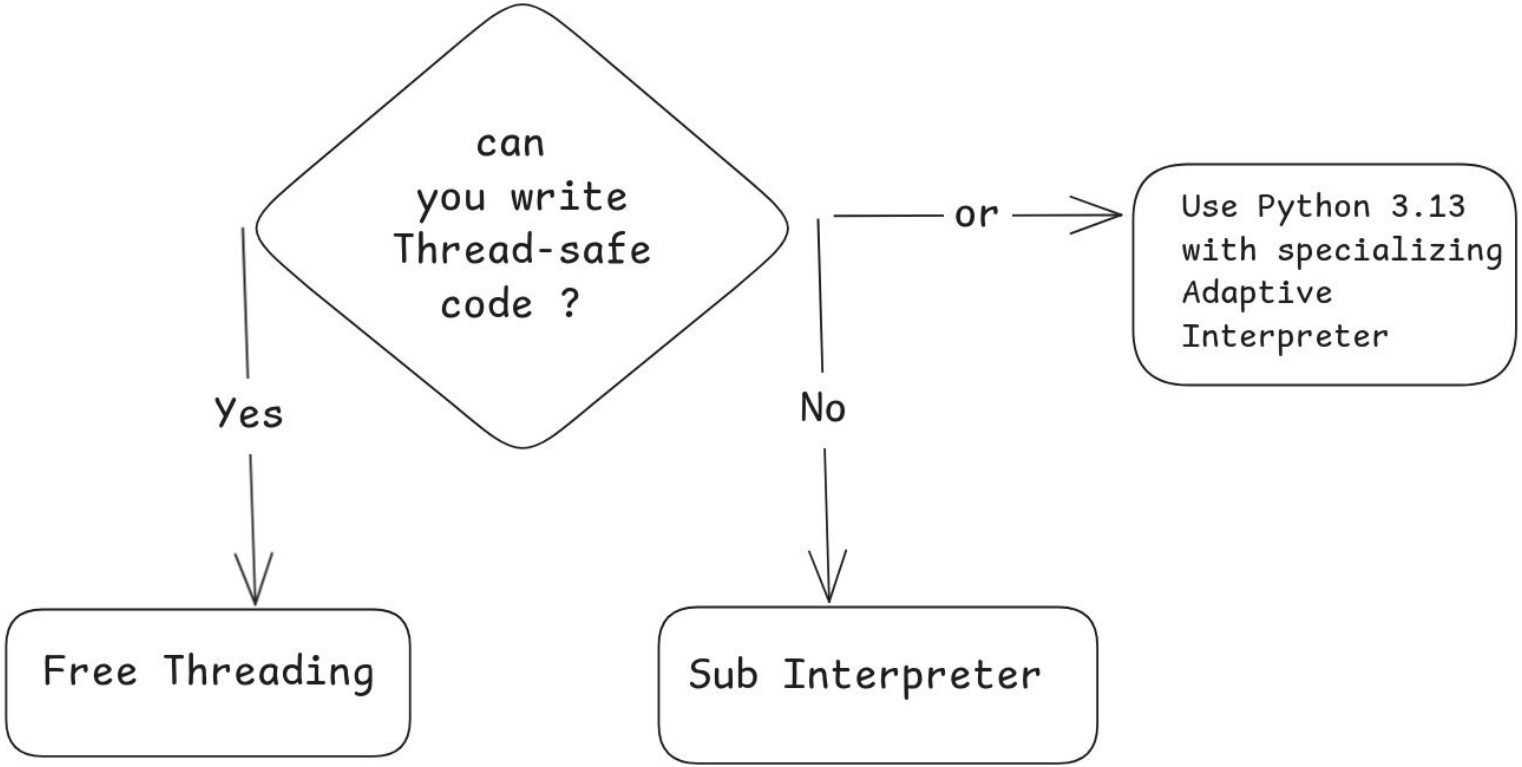

Execution Time vs Fibonacci Number Size
multithreading
subinterpreter
multiprocessing
freethreading
Execution Time (s)
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Fibonacci Number Size

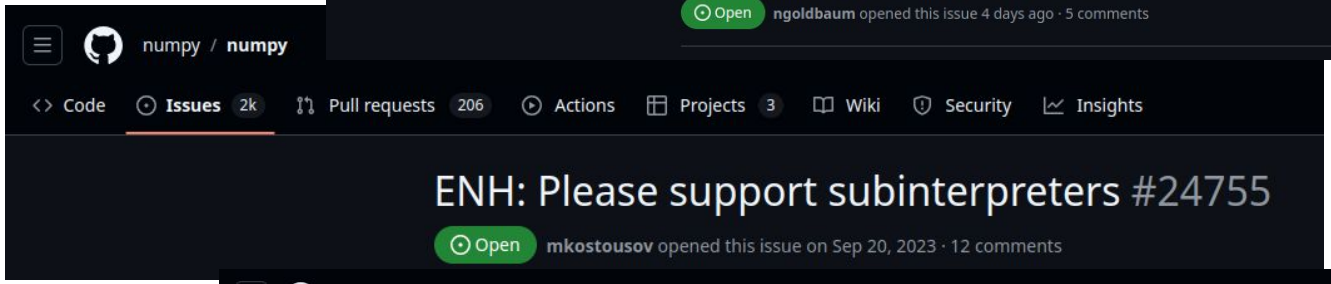

Execution Time by Fibonacci Number Size and Method
multithreading
subinterpreter
multiprocessing
freethreading
Execution Time (s)
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
9
999
9999
99999
999999
Fibonacci Number Size

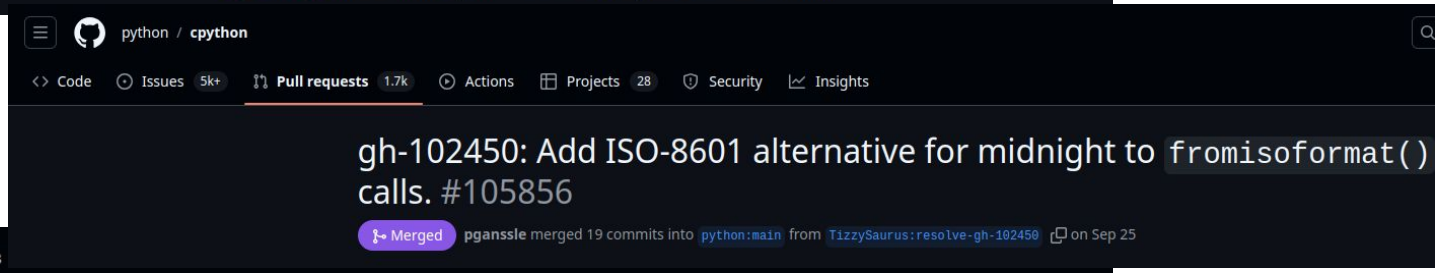

Execution Time vs Fibonacci Number Size (Linear Scale)
multithreading
subinterpreter
multiprocessing
freethreading
Execution Time (s)
0.0
0.2
0.4
0.6
0.8
1.0
0
20000
40000
60000
80000
100000
Fibonacci Number Size

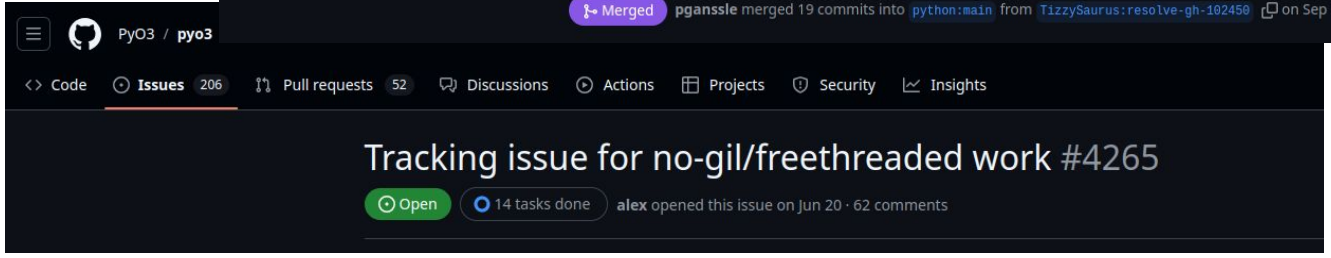

Execution Time by Fibonacci Number Size and Method (Linear Scale)
multithreading
subinterpreter
multiprocessing
freethreading
Execution Time (s)
0.0
0.2
0.4
0.6
0.8
1.0
9
999
9999
99999
Fibonacci Number Size

Free Threading or Sub Interpreter ?







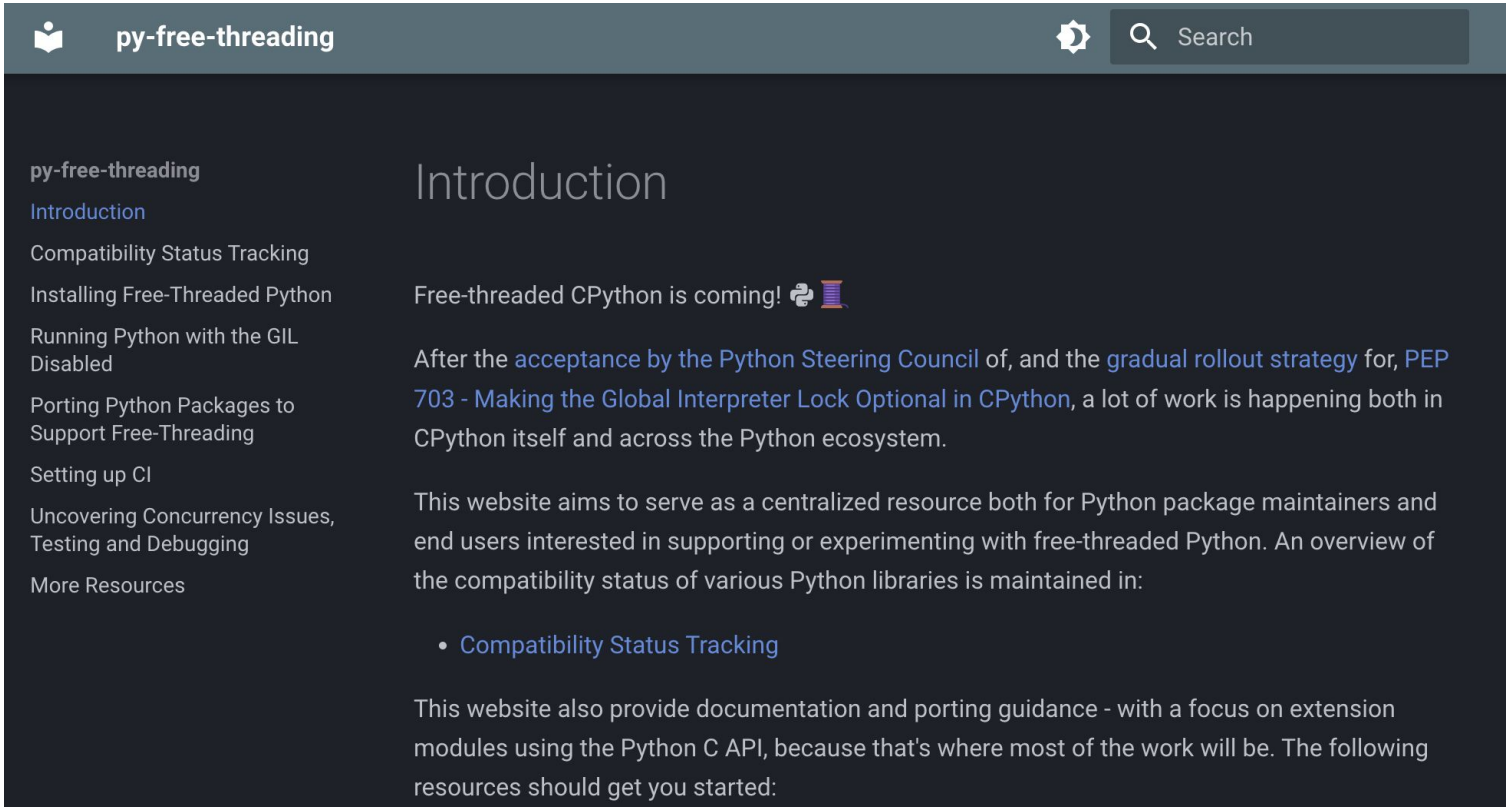


* https://github.com/numpy/numpy/issues/24755
* https://github.com/jupyterlab/jupyterlab/issues/16915
* https://github.com/PyO3/pyo3/issues/4265
* https://github.com/python/cpython/pull/105856

Check few things before using Free Threading

Free-Threading Info : https://py-free-threading.github.io

Py-free-threading



Check few things before using Free Threading

Free-Threading Info : https://py-free-threading.github.io

All of the underlying libraries supports Free Threading ?

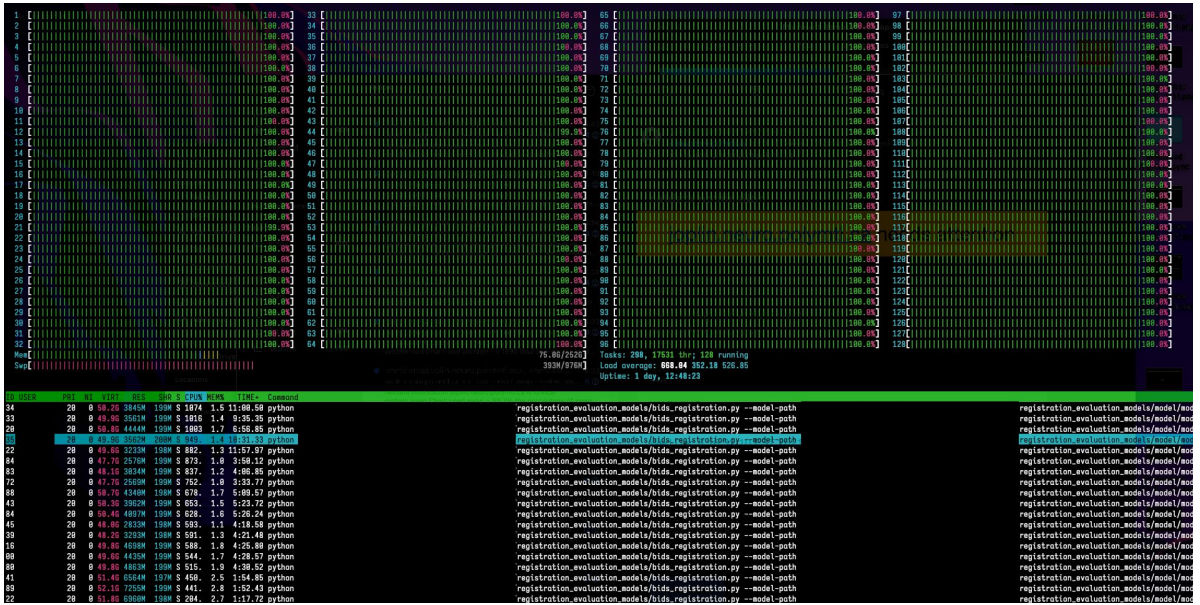
Test Yourself using : pytest-run-parallel / pytest-freethreaded

gil_check

```
>>> import sys
>>> sys._is_gil_enabled()
False
```

* https://github.com/Quansight-Labs/pytest-run-parallel
* https://github.com/tonybaloney/pytest-freethreaded

Thank You



https://www.linkedin.com/in/shekharkoirala/