

Nebenläufigkeit in Python



🚩 Concurrency in Python

Dozent: Prof. Dr. Michael Eichberg
Kontakt: michael.eichberg@dhbw.de
Version: 1.0.1.1

Folien: <https://delors.github.io/ds-nebenlaeufigkeit-in-python/folien.de.md.html>
<https://delors.github.io/ds-nebenlaeufigkeit-in-python/folien.de.md.html.pdf>

Fehler melden: <https://github.com/Delors/delors.github.io/issues>

KI Verwendung: Bei der Erstellung der Unterlagen wurden KI Assistenten (insbesondere Claude aber ggf. auch ChatGPT, Ollama mit Gemma/LLama/Qwen oder OpenCode mit Kimi/Qwen/Deepseek...) unterstützend eingesetzt. Dies erfolgte insbesondere zur Unterstützung bei der Generierung von Grafiken (d. h. SVG Dateien), oder um sich Übersichtstabellen generieren zu lassen. Weiterhin wurde KI zur allgemeinen Qualitätssicherung eingesetzt. Inhalte, die ggf. von der KI vorgeschlagen wurden, wurden im Falle der Übernahme explizit validiert und angepasst.

Hintergrund

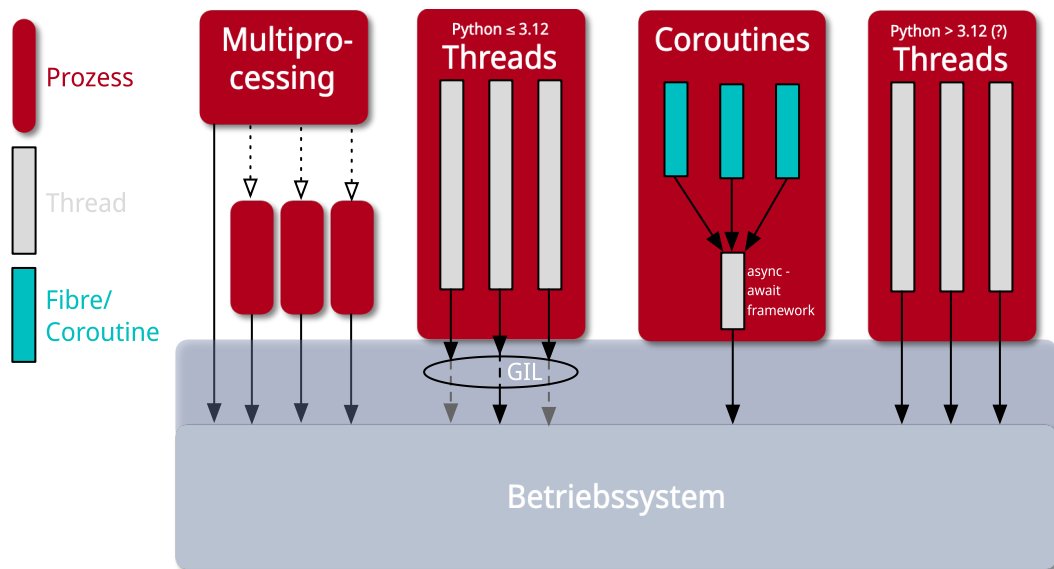
Nebenläufigkeit ist notwendig, um:

- die Reaktionsfähigkeit von Anwendungen zu verbessern
- eine bessere Auslastung von (Mehrkern-)prozessoren zu erreichen

Ein gutes Verständnis von nebenläufiger Programmierung ist für die Entwicklung von verteilten Anwendungen unerlässlich, da Server immer mehrere Anfragen gleichzeitig bearbeiten.

1. Nebenläufigkeitsmodelle in Python

Prozesse vs. Threads vs. Coroutines in Python



- Prozesse sind voneinander isoliert und können nur über explizite Mechanismen miteinander kommunizieren (z. B. Pipes und Queues); Prozesse teilen sich *nicht* denselben Adressraum.
- Alle Threads eines Prozesses teilen sich denselben Adressraum. Python Threads sind vom Betriebssystem unterstützte Threads, die direkt vom Betriebssystem verwaltet werden. Python (d. h. der Standardinterpreter CPython bis (mind.) einschließlich Version 3.12) führt aber immer nur einen Thread aus aufgrund des *Global Interpreter Locks* (GIL).

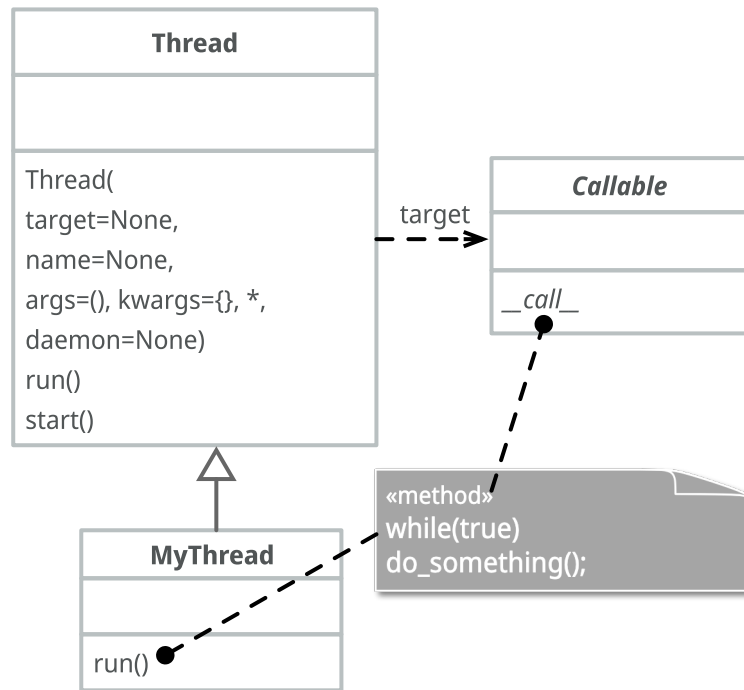
Der GIL existiert(e) insbesondere, da dadurch die Implementierung von Python einfacher wurde (z. B. kann problemlos *Reference Counting* verwendet werden und Probleme mit externen Bibliotheken sind auch minimiert.)

Andere Python-Implementierungen (wie Jython und IronPython) haben keinen GIL und können daher mehrere Threads (echt) parallel ausführen.

- *Coroutines* (auch *Fibres*) nutzen immer kooperatives Multitasking. D. h. ein Fibre gibt die Kontrolle an eine andere Fibre explizit ab. (Früher wurden *Fibres* auch als *Green Threads* bezeichnet.) Diese sind für das Betriebssystem unsichtbar.

Coroutines erfordern explizite Unterstützung in den Bibliotheken. Alle auf Koroutinen basierenden Tasks werden in von der Event-Loop verwaltet und von einem einzigen Thread ausgeführt.

Nebenläufigkeit in Python



♦ Bemerkung

Die `Process` API bietet eine vergleichbare Schnittstelle.

- Threads werden in Python über die vordefinierte Klasse `threading.Thread` bereitgestellt. `target` ist hierbei die Referenz auf das `Callable` Objekt, dass ausgeführt wird als Reaktion auf ein `start` Aufruf.
- Alternativ kann ein `Callable` an ein `Thread`-Objekt übergeben werden.
- Threads beginnen ihre Ausführung erst, wenn die `start`-Methode in der `Thread`-Klasse aufgerufen wird. Die `Thread.start`-Methode ruft die `run`-Methode auf. Ein direkter Aufruf der `run`-Methode führt nicht zu einer nebenläufigen Ausführung.
- Der aktuelle Thread kann mittels der statischen Methode `Thread.currentThread()` ermittelt werden.
- Ein Thread wird beendet, wenn die Ausführung seiner `run`-Methode entweder normal oder als Ergebnis einer unbehandelten Ausnahme endet.
- Python unterscheidet *User-Threads* und *Daemon-Threads*.

Daemon-Threads sind Threads, die allgemeine Dienste bereitstellen und normalerweise nie beendet werden. Jeder Thread, der eine Endlosschleife ausführt sollte als *Daemon-Thread* gekennzeichnet werden bei Erzeugung.

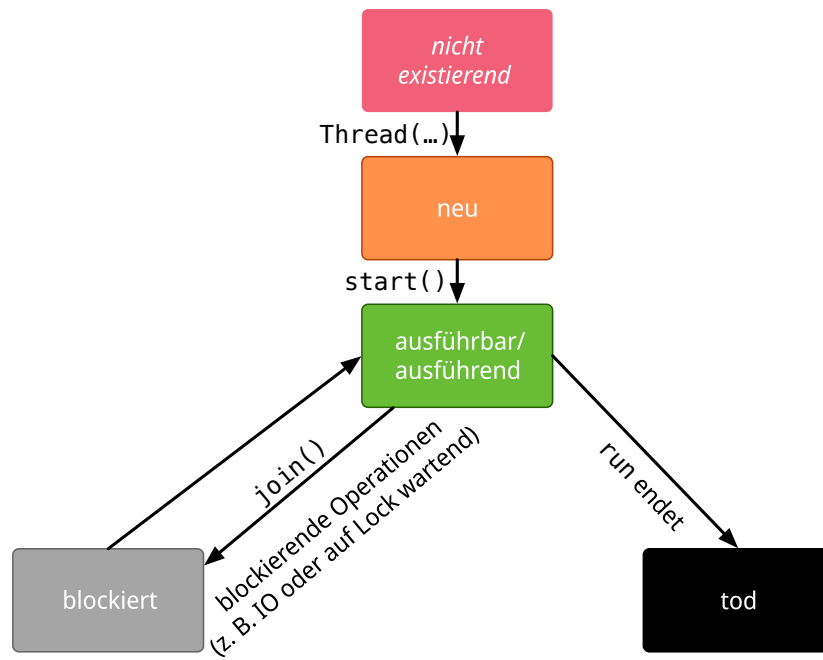
Wenn alle Benutzer-Threads beendet sind, werden die *Daemon-Threads* automatisch beendet, und das Hauptprogramm endet.

Der Thread kann beim Erzeugen als *Daemon-Thread* gekennzeichnet werden, indem der Parameter `daemon` auf `True` gesetzt wird.

Inter-Thread/Process- Koordination

- Ein `Thread/Process` kann (mit oder ohne Zeitüberschreitung) auf die Beendigung eines anderen `Threads/Processes` (des Ziels) warten, indem er die `join`-Methode für das `Thread/Process`-Objekt des Ziels aufruft.
- Mit der Methode `is_alive` kann ein Thread feststellen, ob der Ziel-Thread beendet wurde.

Python Thread States



Beispiel: Multiprocessing - „IO-Bound“

```
1 import time
2 from multiprocessing \
3     import Process, current_process
4
5 def busy_sleep():
6     time.sleep(10)
7
8 print(current_process().name)
9
10 if __name__ == '__main__':
11     p1 = Process(target=busy_sleep)
12     p2 = Process(target=busy_sleep)
13     p1.start() ; p2.start()
14     p1.join()
15     p2.join()
```

```
$ time ./processes_sleep.py
MainProcess
Process-2
Process-1
./processes_sleep.py
 0.07s user
 0.02s system
 0% cpu
10.070 total
```

Beispiel: Multiprocessing - CPU-Bound

```
1 from multiprocessing \
2     import Process, current_process
3
4 def computation():
5     j = 1
6     for i in range(100*1000*1000):
7         j += (i/j)
8     print("Done:"+str(j))
9
10 print(current_process().name)
11
12 if __name__ == '__main__':
13     p1 = Process(target=computation)
14     p2 = Process(target=computation)
15     p1.start()
16     p2.start()
17     p1.join()
18     p2.join()
```

```
$ time ./processes_computation.py
MainProcess
Process-1
Process-2
Done:100000000.0
Done:100000000.0
./processes_computation.py
 5.60s user
 0.02s system
194% cpu
 2.899 total
```

Hinweise

Je nach Betriebssystem werden die Kindprozesse ggf. anders ausgeführt (`fork` oder `spawn`). Linux/Posix bietet die beste Unterstützung gefolgt von MacOS und Windows.

Beispiel: Threading - „IO-Bound“

```
1 import time
2 from threading import Thread, current_thread
3
4 def busy_sleep():
5     # ts_print(current_thread().name)
6     time.sleep(10)
7
8 if __name__ == '__main__':
9     t1 = Thread(target=busy_sleep)
10    t2 = Thread(target=busy_sleep)
11    t1.start()
12    t2.start()
13    t1.join()
14    t2.join()
```

```
$ time ./threads_sleep.py
0.02s user
0.01s system
0% cpu
10.188 total
```

Beispiel: Threading - CPU-Bound

```
1 import time
2 from threading \
3     import Thread, current_thread
4
5 def computation():
6     ts_print(current_thread().name)
7     j = 1
8     for i in range(100*1000*1000):
9         j += (i/j)
10    ts_print("Done:"+str(j))
11
12 if __name__ == '__main__':
13     t1 = Thread(target=computation)
14     t2 = Thread(target=computation)
15     t1.start()
16     t2.start()
17     t1.join()
18     t2.join()
```

```
$ time ./threads_computation.py 16:10:15
Thread-1 (computation)
Thread-2 (computation)
Done:100000000.0
Done:100000000.0
Done.
./threads_computation.py
5.27s user
0.02s system
96% cpu
5.450 total
```

Beispiel: Coroutines

```
1 import asyncio
2
3 async def busy_sleep(id):
4     print(f"Task {id} started")
5     await asyncio.sleep(10)
6     print(f"Task {id} completed")
7
8 async def main():
9     t1 = asyncio.create_task(busy_sleep(1))
10    t2 = asyncio.create_task(busy_sleep(2))
11
12    print("Both initialized.")
13    await t1
14    await t2
15    print("Done.")
16
17 if __name__ == '__main__':
18     asyncio.run(main())
```

```
$ time ./async.py
Both initialized.
Task 1 started
Task 2 started
Task 1 completed
Task 2 completed
Done.
./async.py
 0.05s user
 0.01s system
 0% cpu
10.063 total
```

-
- Beide Tasks werden von dem gleichen Thread ausgeführt. Der Thread gibt „die Kontrolle an die Event-Loop ab“, wenn er auf eine entsprechende blockierende Methode trifft. Die Event-Loop kann dann die Kontrolle an einen anderen Task übergeben.
 - Warten (`await`) ist nur möglich in asynchronen Methoden (`async def`).
 - `asyncio.run(<fn>)` startet die Event-Loop und führt die übergebene asynchrone Methode aus.
 - Die Verwendung von Koroutinen erfordert explizite Unterstützung in den Bibliotheken.

2. Sperren und Bedingungsvariablen

Synchronisation mit Hilfe von *Sperren*

- Zugriff auf gemeinsam genutzte Ressourcen muss synchronisiert werden, um  *Race Conditions* ( *Wettlaufsituationen*) zu vermeiden.

(Unabhängig davon ob Threads echt parallel oder nur scheinbar parallel ausgeführt werden.)

- Eine *Sperre* (`Lock`) ist ein Objekt, das es erlaubt Code im wechselseitigen Ausschluss (engl. *mutual exclusion*) auszuführen.

D. h. ein Thread blockiert, wenn er versucht eine Sperre zu erwerben, die bereits von einem anderen Thread gehalten wird.

- Der Code, der von einer Sperre geschützt wird, wird als kritischer Abschnitt bezeichnet.

Eine *Race Condition* liegt vor, wenn der Zustand eines (Software-)Systems von der Abfolge oder dem Zeitpunkt anderer unkontrollierbarer Ereignisse abhängt. Eine Race Condition führt ggf. zu unerwarteten oder inkonsistenten Ergebnissen.

Verwendung von *Sperren*[\[1\]](#)

- Am Anfang des kritischen Abschnitts wird die Sperre angefordert mit `<Lock>.acquire()`.
- Am Ende des kritischen Abschnitts wird die Sperre freigegeben mit `<Lock>.release()`.
- Um sicherzustellen, dass eine gehaltene Sperre immer aufgehoben wird, sollte `try-finally` oder ein passendes `with`-Statement verwendet werden. (Lock implementiert z. B. das Protokoll von *Context-Managern*)

```
lock = Lock()
lock.acquire()
try:
    # critical section
finally:
    lock.release()
```

```
lock = Lock()

with lock:
    # critical section
```

[\[1\]](#) Die APIs von `threading` und `multiprocessing` sind in weiten Teilen vergleichbar.

Beispiel: Thread-safe Shared Counter

```
1 from threading import Thread, Lock
2
3 class SharedCounter:
4
5     def __init__(self):
6         self._value = 0
7         self.lock = Lock()
8
9     def value(self):
10         return self._value
```

```
1 # Thread-sichere Implementierungen
2 # von increment und decrement
3
4
5 def increment(self):
6     self.lock.acquire()
7     try:
8         self._value += 1
9     finally:
10         self.lock.release()
11
12 def decrement(self):
13     with self.lock:
14         self._value -= 1
```

⚠ Warnung

Code, der eine konkrete Sperre erzeugt, anfordert und freigibt, sollte immer lokal sein; d. h. nicht über die Codebasis verteilt sein. Auch wenn es möglich ist eine Instanz eines Locks weiterzureichen und Sperren in einer Methode anzufordern und in einer anderen Methode freizugeben, so ist dies eine schlechte Praxis, da es zu ((sehr,) sehr) schwer zu findenden Fehlern führen kann.

Sperrern und komplexe Rückgabewerte

```
1 from threading import Thread, Lock
2
3 class SharedCoordinate:
4
5     def __init__(self, x, y):
6         self.x = x
7         self.y = y
8         self.lock = Lock()
```

```
1 def update(self, x, y):
2     self.lock.acquire()
3     try:
4         self.x = x
5         self.y = y
6     finally:
7         self.lock.release()
8
9 def value(self):
10     with self.lock:
11         return (self.x, self.y)
```

Beide Methoden müssen synchronisiert werden, damit es nicht dazu kommen kann, dass man einen ungültigen Zustand beobachten kann. Ein ungültiger Zustand wäre ein paar Koordinaten, die nicht zusammengehören. Z. B. wenn der Wert x von einem Aufruf kommt (update(100,100)) und der Wert y von einem anderen (update(200,200)); d.h. der Wert, den `value` zurückliefert: 100, 200 wäre.

Bedingungsvariablen

Bedingte Synchronisation

- drückt eine Bedingung für die Reihenfolge der Ausführung von Operationen aus.
- z. B. können Daten erst dann aus einem Puffer entfernt werden, wenn Daten in den Puffer eingegeben wurden.
- Python unterstützt optionale Bedingungs-Variablen (Instanzen von `Condition`), mit den klassischen Methoden `wait` und `notify` bzw. `notify_all`.

Diese Methoden erlauben es auf bestimmte Bedingungen zu warten und andere Threads zu benachrichtigen, wenn sich die Bedingung geändert hat.

Programmierung mit `Conditions`

- Die Methoden `wait` und `notify(_all)` können nur verwendet werden, wenn die Sperre gehalten wird; andernfalls wird eine `RuntimeError` ausgelöst.
- Die `wait`-Methode blockiert immer den aufrufenden Thread und gibt die mit dem Objekt verbundene Sperre frei.
- Die `notify(n=1)`-Methode weckt (mind.) n wartende Threads auf. Welcher Thread aufgeweckt wird, ist nicht spezifiziert.
`notify` gibt die Sperre nicht frei; daher muss der aufgeweckte Thread warten, bis er die Sperre erhalten kann, bevor er fortfahren kann.
- Um alle wartenden Threads aufzuwecken, muss die Methode `notify_all` verwendet werden.
Warten die Threads aufgrund unterschiedlicher Bedingungen, so ist immer `notify_all` zu verwenden.
- Wenn kein Thread wartet, dann haben `notify` und `notify_all` keine Wirkung.

⚠ Warnung

Wenn ein Thread aufgeweckt wird, kann er nicht davon ausgehen, dass seine Bedingung erfüllt ist!

Die Bedingung ist immer in einer Schleife zu prüfen und der Thread muss sich ggf. wieder in den Wartezustand versetzen.

Beispiel: Implementation eines *BoundedBuffer*

■ Ein *BoundedBuffer* hat (z. B.) traditionell zwei Bedingungsvariablen:

- *not_full* und
- *not_empty*.

In diesem Fall würde gelten, dass, wenn ein Thread auf eine Bedingung wartet, kein anderer Thread auf die andere Bedingung warten kann, da sich die Bedingungen gegenseitig ausschließen.

Beispiel: Synchronisation mit Bedingungsvariablen

```
1 from threading \
2     import Condition, Lock
3
4 class BoundedBuffer:
5
6     def __init__(self, capacity):
7         self.capacity = capacity
8         self.buffer = []
9         self.lock = Lock()
10        self.not_empty = Condition(self.lock)
11        self.not_full = Condition(self.lock)
```

```
1     def put(self, item):
2         with self.not_full:
3             while len(self.buffer) == \
4                 self.capacity:
5                 self.not_full.wait()
6             self.buffer.append(item)
7             self.not_empty.notify()
8
9     def get(self):
10        with self.not_empty:
11            while len(self.buffer) == 0:
12                self.not_empty.wait()
13            item = self.buffer.pop(0)
14            self.not_full.notify()
15        return item
```

Beispiel: Synchronisation mit nur einer Bedingung

Im Folgenden sehen wir eine Implementierung mit nur einer Bedingungsvariablen, um bestimmte Synchronisationsfehler demonstrieren zu können.

```
1 from threading import Thread, Lock, Condition
2
3 class BoundedBuffer:
4
5     def __init__(self, capacity):
6         self.capacity = capacity
7         self.buffer = []
8         self.lock = Lock()
9         self.not_used = Condition(self.lock)
10
11 ...
11
12 def put(self, item):
13     with self.not_used:
14         while len(self.buffer) == self.capacity:
15             self.not_used.wait()
16         self.buffer.append(item)
17         self.not_used.notify_all() # notify_all() !
18
19 def get(self):
20     with self.not_used:
21         while len(self.buffer) == 0:
22             self.not_used.wait()
23         item = self.buffer.pop(0)
24         self.not_used.notify_all() # notify_all() !
25     return item
```

Fehler, der bei der Verwendung von `notify` (statt `notify_all`) auftreten könnte.

```
1 bb = BoundedBuffer(1);
2 p1 = Thread(target=lambda: bb.put(1)); p2 = Thread(target=lambda: bb.put(2))
3 c1 = Thread(target=lambda: bb.get()); c2 = Thread(target=lambda: bb.get())
4 c1.start(); c2.start(); p1.start(); p2.start();
```

Aktionen	(Änderung des) Zustand(s) des Buffers	Auf die Sperre (Lock) wartend	An der Bedingung wartend
c1:bb.get(), 1 c2:bb.get(), p1:bb.put(), p2:bb.put()	empty	{c2,p1,p2}	{c1}
2 c2:bb.get()	empty	{p1,p2}	{c1,c2}
3 p1:bb.put(1)	empty → not empty	{p2,c1}	{c2}
4 p2:bb.put(2)	not empty	{c1}	{c2,p2}
5 c1:bb.get()	not empty → empty	{c2}	{p2}
6 c2:bb.get()	empty	∅	{c2,p2}

In Schritt 5 wurde (z. B.)- aufgrund des Aufrufs von `notify` durch `c1` - der Thread `c2` aufgeweckt - anstatt des Threads `p2`. Der aufgeweckte Thread `c2` prüft die Bedingung (Schritt 6) und stellt fest, dass der Puffer leer ist. Er geht wieder in den Wartezustand. Jetzt warten sowohl ein Thread, der ein Wert schreiben möchte, als auch ein Thread, der einen Wert lesen möchte.

Best Practices in Hinblick auf Synchronisation

- ◉ Code, der eine Sperre hält ( Lock) sollte so kurz (zeitlich) wie möglich gehalten werden.

(D. h. der Code zwischen `Lock.acquire()` und `Lock.release()`)

- ◉ Verschachtelte Anforderungen von Sperren sollten vermieden werden, da die äußere Sperre nicht freigegeben wird, wenn man an der Inneren wartet. Dies kann leicht zum Auftreten eines Deadlocks führen.

Wenn zwei (oder mehr) Threads bzw. Prozesse auf die gleichen Ressourcen in unterschiedlicher Reihenfolge zugreifen und entsprechende Sperren halten bzw. anfordern, kann es zu einem Deadlock kommen.

⚠ Warnung

Ressourcen immer in der gleichen Reihenfolge sperren, um Deadlocks zu vermeiden.

※ Hinweis

Sperren (d. h. `Locks`) in Verbindung mit Bedingungsvariablen sind nur eine Möglichkeit, um die Synchronisation von Threads zu ermöglichen. Es ist jedoch ein sehr häufiges Modell. (Alternativen sind zum Beispiel: *Semaphoren*, *Nachrichtenübermittlung*)

3. Ausgewählte Aspekte der Nebenläufigkeit

Thread-lokaler Speicher

Thread-lokaler Speicher (`threading.local()`) ermöglicht es, dass jeder Thread eine lokale Kopie einer bestimmten Variable hat

```
import time
import threading

stop = False # shared global variable
local_data = threading.local()

def f(v):
    setattr(local_data, "value", 0)
    while(not stop):
        print(local_data.value)
        local_data.value += v
        time.sleep(1)

# "main" thread
t1 = threading.Thread(target=f, args=(1,))
t2 = threading.Thread(target=f, args=(-1,))
t1.start()
t2.start()
time.sleep(3);
print("Attributes of local_data: " + \
      str(local_data.__dict__.keys()))
stop = True
print("Stop set to True.")
t1.join()
t2.join()
```

```
$ ./ThreadLocal.py
```

```
0
0
-1
1
-2
2
```

```
Attributes of local_data: []
```

```
Stop set to True. Waiting for threads to finish.
```

Reentrant Locks

- *Reentrant Locks* (`RLock`) sind Sperren, die von demselben Thread mehrmals erworben werden können.
- Implementierungen: `threading.RLock` oder `multiprocessing.RLock`.

Thread-/ProcessPools

- *ThreadPool*s und *ProcessPool*s bieten eine höherwertige Abstraktion, um eine große Anzahl von Aufgaben nebenläufig zu verarbeiten.

- Beide erben von `concurrent.futures.Executor`; zentrale Methoden:

- `submit(fn, *args, **kwargs)`: Fügt eine Aufgabe hinzu und gibt ein `Future`-Objekt zurück.

Auf `Futures` sind die Hauptfunktionen:

- `done()`: Gibt zurück, ob die Aufgabe abgeschlossen ist.
- `result(timeout=None)`: Gibt das Ergebnis zurück, wenn die Aufgabe abgeschlossen ist; blockiert ggf..
- `map(func, *iterables, timeout=None, chunksize=1)`: Führt die Funktion für jedes Element in `iterables` aus und gibt die Ergebnisse in der Reihenfolge zurück, in der sie abgeschlossen wurden.

Nachrichtenaustausch

Motivation: Nachrichtenaustausch

- Locks haben das große Potential eigentlich nebenläufige Programme effektiv zu serialisieren (und zu verlangsamen).
- Prozesse nutzen keinen gemeinsamen Adressraum.
- Eine Möglichkeit auf Locks weitgehend zu verzichten ist der Nachrichtenaustausch.

Generell ist der Austausch zwischen Prozessen über `Queues`, `Pipes` und (explizitem) `SharedMemory` möglich; d. h. in diesen Fällen ist Inter-Prozess-Kommunikation (*Interprocess Communication (IPC)*) notwendig.

Queues

`queue.Queue` **oder** `multiprocessing.JoinableQueue`

Die grundlegenden Methoden von `Queues` sind:

- `Queue(maxsize=0)`

Erzeugt eine neue Queue-Instanz welche `maxsize` Elemente speichern kann. 0 bedeutet, dass die Queue unendlich groß ist.

(Pythons `Queue` realisiert einen *Bounded Buffer*.)

- `put(item)` : Fügt ein Element in die Queue ein.

- `get()` : Entfernt und gibt das erste Element aus der Queue zurück.

- `task_done()` : Signalisiert, dass ein Element aus der Queue *abgearbeitet* wurde.

- `join()` : Blockiert bis alle Elemente aus der Queue *abgearbeitet* wurde.

Beispiel - Verwendung von Queues für Thread-Sichere Konsolenausgabe

Setup

```
1 import threading
2 from queue import Queue
3
4 print_queue = Queue()
5
6 def ts_print(msg):
7     print_queue.put(msg)
8
9 def print_handler():
10     while True:
11         msg = print_queue.get()
12         # there will ever be only one thread
13         print(msg)
14         print_queue.task_done()
```

Verwendung

```
Thread(target=print_handler, daemon=True).\
    start()
:
# <thread 1:> ts_print("Hello")
:
# <thread 2:> ts_print("World")
:
print_queue.join()
```

※ Hinweis

- nur ein Thread darf die print_queue abarbeiten
- wir müssen überall ts_print verwenden

Verwendung von Queues für die Kommunikation zwischen Prozessen

```
1 from random import randint
2 from multiprocessing import current_process, Process, JoinableQueue as MPQueue
3 from threading import Thread
4 from queue import Queue as TQueue
5 import time
6
7 def print_queue_handler(print_queue):
8     while True:
9         msg = print_queue.get()
10        print(msg)
11        print_queue.task_done()
12
13 def read_from_ip_queue(ip_queue, print_queue): # ip =(here) interprocess
14     while True:
15         msg = ip_queue.get()
16         print_queue.put(msg)
17         ip_queue.task_done()
18
19 def f(c_to_p_ip_queue):
20     time.sleep(randint(1, 3)) # just some fuzzing
21
22     c_to_p_ip_queue.put("I'm alive: " + current_process().name)
23
24     time.sleep(randint(1, 3)) # just some fuzzing
25
26     c_to_p_ip_queue.put("Hell World from " + current_process().name)
27
28 if __name__ == "__main__":
29     print_queue = TQueue()
30     c_to_p_ip_queue = MPQueue()
31     p1 = Process(target=f, args=(c_to_p_ip_queue,))
32     p1.start()
33     p2 = Process(target=f, args=(c_to_p_ip_queue,))
34     p2.start()
35     Thread(
36         target=read_from_ip_queue,
37         args=(c_to_p_ip_queue, print_queue, ),
38         daemon=True,
39     ).start()
40     Thread(target=print_queue_handler, args=(print_queue,), daemon=True).start()
41     c_to_p_ip_queue.join()
42     print_queue.join()
43     p2.join()
44     p1.join()
```

Thread Safety

🇩🇪 *Threadsicherheit*

Thread Safety - Voraussetzung

Damit eine Klasse thread-sicher ist, muss sie sich in einer single-threaded Umgebung korrekt verhalten.

D. h. wenn eine Klasse korrekt implementiert ist, dann sollte keine Abfolge von Operationen (Lesen oder Schreiben von öffentlichen Feldern und Aufrufen von öffentlichen Methoden) auf Objekten dieser Klasse in der Lage sein:

- das Objekt in einen ungültigen Zustand versetzen,
- das Objekt in einem ungültigen Zustand zu beobachten oder
- eine der Invarianten, Vorbedingungen oder Nachbedingungen der Klasse verletzen.

Die Klasse muss das korrekte Verhalten auch dann aufweisen, wenn auf sie von mehreren Threads aus zugegriffen wird.

- Unabhängig vom *Scheduling* oder der Verschachtelung der Ausführung dieser Threads durch die Laufzeitumgebung,
- Ohne zusätzliche Synchronisierung auf Seiten des aufrufenden Codes.

Dies hat zur Folge, dass Operationen auf einem thread-sicheren Objekt für alle Threads so erscheinen als ob die Operationen in einer festen, global konsistenten Reihenfolge erfolgen würden.

Da sich Prozesse den Adressraum mit Threads nicht teilen, ist es nicht möglich, dass ein Prozess den Speicher eines anderen Prozesses direkt manipuliert. Dies bedeutet jedoch nicht, dass keine Inter-Prozess-Koordination notwendig ist. Insbesondere wenn auf gemeinsame Ressourcen - wie zum Beispiel die Konsole - zugegriffen wird, ist eine Koordination notwendig.

Thread Safety Level

Immutable *Unveränderlich*:

Die Objekte sind konstant und können nicht geändert werden.

Thread-sicher: Die Objekte sind veränderbar, unterstützen aber nebenläufigen Zugriff, da die Methoden entsprechende Sperren und Bedingungen verwenden.

Bedingt Thread-sicher:

All solche Objekte bei denen jede einzelne Operation thread-sicher ist, aber bestimmte Sequenzen von Operationen eine externe Synchronisierung erfordern können.

Thread-kompatibel:

Alle Objekte die keinerlei Synchronisierung aufweisen. Der Aufrufer kann die Synchronisierung jedoch ggf. extern übernehmen.

Thread-hostile „Thread-schädlich“:

Objekte, die nicht thread-sicher sind und auch nicht thread-sicher gemacht werden können, da sie zum Beispiel globalen Zustand manipulieren.

Ein Beispiel bzgl. *bedingt Thread-sicher* wäre die Verwendung eines Iterators, bei dem die Methoden für sich genommen thread-sicher sind, aber die Iteration über die Elemente als ganzes zusätzliche Synchronisation erfordert, damit die Ergebnisse konsistent sind.

Ein Beispiel für eine *thread-schädliche* Klasse (Code) wäre eine Klasse, die auf eine globale Variable zugreift bzw. globalen Zustand ändert, der von mehreren Threads verwendet wird, ohne dass eine Synchronisierung stattfindet.

⚠ Warnung

Wenn Nebenläufigkeit nicht richtig umgesetzt wird, dann kann dies nicht nur zu schwer zu findenden Fehlern führen sondern auch **zu langsam(er)en Programmen**.

Im Allgemeinen sollte Parallelisierung auf *höchstmöglicher Ebene* erfolgen.

Threads und Prozesse nicht terminieren

⚠ Warnung

Auch wenn es technisch möglich ist Threads und Prozesse explizit zu terminieren (z. B. durch `Process.terminate()`) so sollte man darauf verzichten.

Das Hauptproblem sind nicht freigegebene Locks und Ressourcen, die sich in einem inkonsistenten Zustand befinden können.

Auch in anderen Programmiersprachen sollte man niemals Threads oder Prozesse explizit terminieren.

⚠ Warnung

Nebenläufigkeit macht *nichts* einfacher! Entwickle und teste immer erst eine single-threaded Version des Programms.

Übung

Implementieren Sie einen einfachen *DelayedBuffer*, der es ermöglicht Aufgaben (d. h. Objekte vom Typ `Callable`) erst nach einer bestimmten Zeit auszuführen. Die Klasse muss zwei Funktionen zur Verfügung stellen:

```
submit(self, delay, fn, *args, **kwargs):
```

Die Funktion `fn` wird nach `delay` Sekunden ausgeführt wobei `delay` vom Typ `Float` ist. `args` und `kwargs` sind die Argumente, die an `fn` übergeben werden.

```
join(self):
```

Wartet bis alle Aufgaben abgearbeitet wurden.

Im Folgenden sehen Sie eine mögliche Verwendung des Puffers:

```
buffer = DelayedBuffer()
buffer.submit(100 / 1000, ts_print, "Hello ", **{"end": "", "flush": True})
buffer.submit(1000 / 1000, ts_print, "World!")
buffer.submit(500 / 1000, ts_print, "of the ", **{"end": "", "flush": True})
buffer.submit(200 / 1000, ts_print, "from ", **{"end": "", "flush": True})
buffer.submit(300 / 1000, ts_print, "the other side ", **{"end": "", "flush": True})
# ggf. await buffer.join() im Falle von Koroutinen
buffer.join()
print("Done.")
```

3.1. Implementation mit Threads

Implementieren Sie die Klasse `DelayedBuffer` mit Hilfe von `Threads` (und ggf. `Queues` bzw. `Locks`).

Implementieren Sie `ts_print` als Thread-sichere Variante von `print`.

3.2. Implementation mit Threadpool

Implementieren Sie die Klasse `DelayedBuffer` mit Hilfe eines `concurrent.futures.ThreadPool`s (und ggf. `Queues` bzw. `Locks`).

Implementieren Sie `ts_print` als Thread-sichere Variante von `print`. Wählen Sie ggf. eine andere Implementierung als in der vorherigen Aufgabe.

3.3. Implementation mit Koroutinen

Implementieren Sie die Klasse `DelayedBuffer` mit Hilfe von Koroutinen (und ggf. `asyncio.Queue`).