Parallel Computing in Python: multiprocessing

Konrad HINSEN

Centre de Biophysique Moléculaire (Orléans)

and

Synchrotron Soleil (St Aubin)

Parallel computing: Theory

Parallel computers

- **Multiprocessor/multicore:**
  several processors work on data stored in shared memory
- **Cluster:**
  several processor/memory units work together by exchanging data over a network
- **Co-processor:**
  a general-purpose processor delegates specific tasks to a special-purpose processor (GPU, FPGA, ...)
- **Other:**
  - Cluster of multicore nodes with GPUs
  - NUMA (non-uniform memory access) architectures
  - ...

Almost all computers made today are parallel!

Parallelism vs. concurrency

Parallelism:

use multiple processors to make a computation faster.

Concurrency:

permit multiple tasks to proceed without waiting for each other.

Different goals that share *implementation aspects*.

Scientific computing cares more about ***parallelism***.
Concurrency is rarely needed.

Parallel Programming

Parallel Programming

- Decomposition of the complete task into independent subtasks and the data flow between them.

Parallel Programming

- Decomposition of the complete task into independent subtasks and the data flow between them.
- Distribution of the subtasks over the processors minimizing the total execution time.

Parallel Programming

- Decomposition of the complete task into independent subtasks and the data flow between them.
- Distribution of the subtasks over the processors minimizing the total execution time.
- For clusters: distribution of the data over the nodes minimizing the communication time.

Parallel Programming

- Decomposition of the complete task into independent subtasks and the data flow between them.
- Distribution of the subtasks over the processors minimizing the total execution time.
- For clusters: distribution of the data over the nodes minimizing the communication time.
- For multiprocessors: optimization of the memory access patterns minimizing waiting times.

Parallel Programming

- Decomposition of the complete task into independent subtasks and the data flow between them.
- Distribution of the subtasks over the processors minimizing the total execution time.
- For clusters: distribution of the data over the nodes minimizing the communication time.
- For multiprocessors: optimization of the memory access patterns minimizing waiting times.
- Synchronization of the individual processes.

Difficulties

Correctness:

- Verifying that subtasks are indeed independent.
- Making synchronization patterns deadlock-free.
- Clusters: verifying communication patterns.

Difficulties

Correctness:

- Verifying that subtasks are indeed independent.
- Making synchronization patterns deadlock-free.
- Clusters: verifying communication patterns.

Efficiency:

- Attributing equal workloads to all processors.
- Taking into account computation *and* communication.
- Optimizing for a specific parallel computer.

Synchronization issues

Deadlock:

- Two processes are waiting for each other to finish.
- Usually caused by **locks** or by **blocking communication**.

Race condition:

- Two or more processes modify a shared resource (variable, file, ...)
- Result depends on which process comes first.
- Can be avoided using **locks**, but...
- ... handling locks is ***very*** difficult and mistakes often cause **deadlocks**.

Parallel Programming Models

Many parallel programming models have been proposed, more are currently being developed. This is a very active field of research. The ideal programming model should be

- **High-level:** express parallel algorithms rather than parallel computing hardware.
- **Deterministic:** a program should have a unique result that depends only on its inputs and not on the number of processors or on random factors (machine load etc.).
- **Universal:** efficiently applicable to clusters, multiprocessors, etc.
- **General:** applicable to all parallel algorithms.
- **Modular:** permit the construction of composable libraries.
- **Simple:** both the meaning and the cost (time and memory) of each statement or construct should be clear.

Parallel Programming Models

Popular models in scientific computing:

- **Message-passing:** a low-level non-deterministic general model for programming clusters. Used by the very popular MPI library. Suitable but less efficient for multiprocessors.

- **Multi-threading:** a low-level non-deterministic general model for multiprocessors. Not suitable for clusters. Popular implementations: Posix threads, OpenMP.

- **Data-parallel:** a high-level deterministic specialized model. Suitable for all parallel computers.

- **Task farming:** a high-level almost-deterministic specialized model. Suitable for all parallel computers.

Parallel Programming Models

Popular models in scientific computing:

- **Message-passing:** a low-level non-deterministic general model for programming clusters. Used by the very popular MPI library. Suitable but less efficient for multiprocessors.

- **Multi-threading:** a low-level non-deterministic general model for multiprocessors. Not suitable for clusters. Popular implementations: Posix threads, OpenMP.

- As low-level as C programming. Sometimes necessary for top performance, but try to avoid them as long as possible!

- **Task farming:** a high-level almost-deterministic specialized model. Suitable for all parallel computers.

Parallel Programming Models

Popular models in scientific computing:

- **Message-passing:** a low-level non-deterministic general model for programming clusters. Used by the very popular MPI library. Suitable but less efficient for multiprocessors.

- **Multi-threading:** a low-level non-deterministic general model for multiprocessors. Not suitable for clusters. Popular implementations: Posix threads, OpenMP.

- **Data-parallel:** a high-level deterministic specialized model. Suitable for all parallel computers.

- **Task farming:** a high-level almost-deterministic specialized model. Suitable for all parallel computers.

Task farming

One **master process** supervises the execution of the program. It defines independent **tasks** and puts them on a **to-do list**. It also **collects** the **results** of these tasks.

Any number of **slave processes** each take a task from the to-do list, execute it, and put the result into the master's mailbox.

Advantages:

- Very simple model

- No deadlocks, since only the master process ever waits for another process to finish

Limitations:

- Tasks cannot delegate work to sub-tasks. Adding this possibility would introduce the possibility of deadlocks.

- Rigid communication pattern, no optimization possible.

- Distributed data storage impossible.

Parallel computing:
Python practice

Work in progress

Parallel computing in Python (as in most other languages) is recent. There are many complications and limitations in the parallelization libraries that needn't be there and will probably disappear in the future. Please be patient.

Parallel computing libraries

Lots of Python libraries:

http://wiki.python.org/moin/ParallelProcessing

- **In the Python standard library:**
  - threading: thread handling and locks
  - multiprocessing (Python 2.6): process-based multithreading
- **In ScientificPython:**
  - Scientific.BSP: "Bulk Synchronous Parallel" model
  - Scientific.DistributedComputing: task farming
- **In IPython:**
  - interactive shell for working with clusters
- **Other:**
  - pypar, pyMPI, mpi4py implement MPI-like message passing

Communication

Python programming libraries use two mechanisms for exchanging data between processes/threads/nodes:

1) **Shared memory** (threading, multiprocessing)

   Requires locks for safe modification !

2) **Communication streams** (multiprocessing, MPI, ...)

   Data is passed as byte streams through sockets or TCP connections. Non-string data requires *serialization* before being sent and deserialization for reconstruction.

   This adds overhead that can be important.

   Moreover, ***not all Python objects can be serialized.***
   Sometimes for a good technical reason, sometimes because before parallel computing nobody needed it.

A first parallel program

```
from multiprocessing import Pool
import numpy

def sqrt(x):
    return numpy.sqrt(x)

if __name__ == '__main__':
    pool = Pool()
    roots = pool.map(sqrt, range(100))
    print roots
```

for Windows compatibility

Why not directly:

```
squares = pool.map(numpy.sqrt, range(100))
```

Because `numpy.sqrt` is not serializable (yet).

At the moment, only Python functions defined at the top level of a module are serializable.

What happens...

1. `pool = Pool()` launches one slave process per physical processor on the computer. On Unix systems, the slaves are forked from the master process. Under Windows, a new process is started that imports the script.
2. `pool.map(sqrt, range(100))` divides the input list into chunks of roughly equal size and puts the tasks (function + chunk) on a todo list.
3. Each slave process takes a task (function + a chunk of data) from the todo list, runs `map(function, chunk)`, and puts the result on a result list.
3. `pool.map` on the master process waits until all tasks are handled and returns the concatenation of the result lists.

The todo list

The todo list is actually a *queue*, i.e. a data structure to which items are added at one end and taken off at the other end.

The todo list must be accessible by all processes (master and slaves).

Access to the todo list must be synchronized to prevent data corruption.

The todo list is stored on the master process. A special thread of the master process waits for task requests from slave processes and returns the task function and arguments. This requires *serialization*.

Processes vs. threads

A *process* consists of

- a block of memory
- some executable code
- one or more threads that execute code independently but work on the same memory

Multithreading: using multiple threads in the same process for concurrency or parallelism

Multiprocessing: using multiple processes with separate memory spaces for concurrency or parallelism

But... why use processes rather than threads?

The Global Interpreter Lock (GIL)

The Python interpreter is not thread safe.

A few critical internal data structures may only be accessed by one thread at a time. Access to them is protected by the GIL.

This is not a requirement of the Python language, but an implementation detail of the CPython interpreter. Jython, IronPython, and PyPy don't have a GIL and are fully thread-safe.

Attempts at removing the GIL from CPython have failed until now. The main difficulty is maintaining the C API for extension modules.

Multiprocessing avoids the GIL by having separate processes which each have an independent copy of the interpreter data structures.

The price to pay: serialization of tasks, arguments, and results.

Explicit task definition

```
from multiprocessing import Pool
import numpy

def sqrt(x):
    return numpy.sqrt(x)

if __name__ == '__main__':
    pool = Pool()
    results = [pool.apply_async(sqrt, (x,))
               for x in range(100)]
    roots = [r.get() for r in results]
    print roots
```

1. pool.apply_async returns a proxy object immediately
2. proxy.get() waits for task completion and returns the result

Use for:

- launching *different* tasks in parallel
- launching tasks with more than one argument
- better control of task distribution

Shared memory

Under Unix, it is possible to share blocks of memory between processes. This eliminates the serialization overhead.

Multiprocessing can create shared memory blocks containing C variables and C arrays. A NumPy extension adds shared NumPy arrays. It it *not* possible to share arbitrary Python objects.

Caveats:

1) Portability: there is no shared memory under Windows.
2) If you care about your mental sanity, **don't modify shared memory contents in the slave processes**. You will end up debugging race conditions.

Use shared memory only to transfer data from the master to the slaves!

Shared memory

```
from multiprocessing import Pool
from parutils import distribute
import numpy
import sharedmem

def apply_sqrt(a, imin, imax):
    return numpy.sqrt(a[imin:imax])

if __name__ == '__main__':

    pool = Pool()
    data = sharedmem.empty((100,), numpy.float)
    data[:] = numpy.arange(len(data))
    slices = distribute(len(data))
    results = [pool.apply_async(apply_sqrt, (data, imin, imax))
               for (imin, imax) in slices]
    for r, (imin, imax) in zip(results, slices):
        data[imin:imax] = r.get()
    print data
```

PARUTILS.DISTRIBUTE

Distributes a sequence equally (as much as possible) over the available processors. Returns a list of index pairs (imin, imax) that delimit the slice to give to one task.

```
from multiprocessing import cpu_count

default_nprocs = cpu_count()
```

```
def distribute(nitems, nprocs=None):
    if nprocs is None:
        nprocs = default_nprocs
    nitems_per_proc = (nitems+nprocs-1)/nprocs
    return [(i, min(nitems, i+nitems_per_proc))
            for i in range(0, nitems, nitems_per_proc)]
```

Shared memory with in-place modification

```
from multiprocessing import Pool
from parutils import distribute
import numpy
import sharedmem

def apply_sqrt(a, imin, imax):
    a[imin:imax] = numpy.sqrt(a[imin:imax])

if __name__ == '__main__':

    pool = Pool()
    data = sharedmem.empty((100,), numpy.float)
    data[:] = numpy.arange(len(data))
    tasks = [pool.apply_async(apply_sqrt, (data, imin, imax))
             for (imin, imax) in distribute(len(data))]
    for t in tasks:
        t.wait()
    print data
```

Debugging: monoprocessing

Parallel debugging is a mess. You don't want to do it.

The module `monoprocessing` contains a class `Pool` with the same methods as `multiprocessing.Pool`, but all tasks are executed immediately and in the same process. This permits debugging with standard tools.

If your programs works with `monoprocessing` but not with `multiprocessing`, explore the following possibilities:

- Serialization: some object cannot be serialized
- The code of a task refers to a global variable in the master process
- The code of a tasks modifies data in shared memory

Exercises

Application 1: système solaire

Dans le simulateur du système solaire, la fonction responsable pour la majeure partie du temps CPU est calc_forces. Parallélisez-la.

0) Prenez la version NumPy du simulateur comme point de départ.
1) Donnez deux arguments supplémentaires (jmin, jmax) à calc_forces. Au lieu de calculer les forces entres tous les pairs (i, j), elle doit calculer seulement les interactions des pairs (i, jmin:jmax).
2) Renommez calc_forces et faites une nouvelle fonction calc_forces qui appelle l'ancienne version pour plusieurs tranches (jmin:jmax) par pool.apply_async, attend les résultats, et le combine dans un seul tableau.

Après chaque modification, vérifiez que le programme fonctionne encore correctement !

Application 2: plaque chauffée

Dans le simulateur de la plaque chauffée, la méthode responsable pour la majeure partie du temps CPU est laplacien. Parallélisez-la.

0) Prenez la version NumPy du simulateur comme point de départ.
1) Créez une fonction laplacian hors de la classe Grid (pour la sérialisation) qui prend comme seul argument le tableau des températures et qui retourne l'erreur residuelle ainsi que le tableau des températures à la fin de l'itération. Modifiez la classe Grid tel qu'elle utilise cette fonction.
2) Créez une méthode laplacian_parallel qui découpe le tableau en bandes, les passes à des tâches indépendantes, récupère les résultats, et les combine pour obtenier le tableau de températures total ainsi que l'erreur résiduelle.

Après chaque modification, vérifiez que le programme fonctionne encore correctement !