MSML 605 - Lecture 10

Parallel Processing

Process

- *A unit of work, for example, Jupyter notebook*
- *An OS can run multiple processes at the same time.*
- *By default Python interpreter executes instructions serially.*
- *The size of the datasets has in increased.*
- *The algorithms are more complex and need to process more, hence the need for multi-processing*

Parallel processing

- *To speed up a process we want to split it to distribute across many CPUs*
- *Faster or/and efficiently*
- *Many tasks are suited for parallel processing, for example matrix multiplication*
- *A process can have multiple threads.*

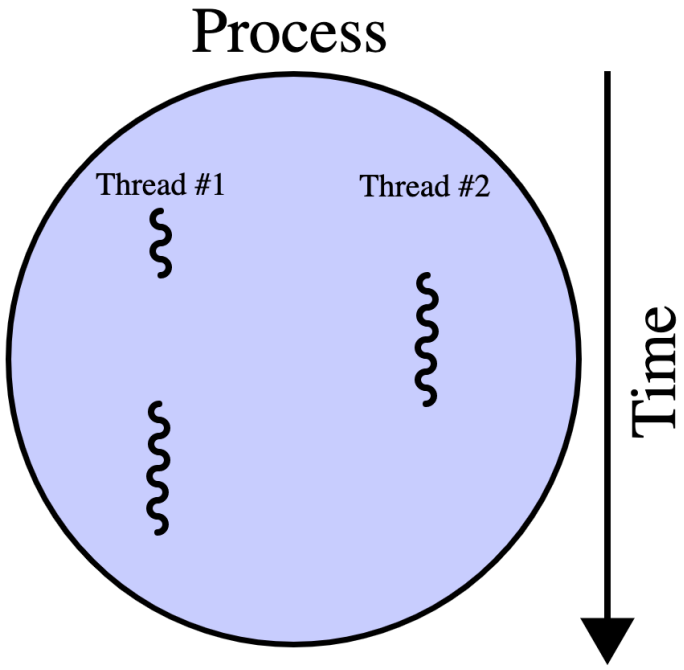
Parallel processing

- *It can be achieved in two ways: Multiprocessing and Threading*
- *Process: An instance of a program*
  Uses its own memory space

- *Threads: components of a process, which can run in parallel*
  - *Multiple threads*
  - *Share parent process memory space*

Processes and Threads

- *Threads live in the same memory space*
- *Processes have their separate memory space*
- *Spawning processes is slower than spawning threads.*
- *Sharing objects between threads is easier.*
- *Inter-process communication between processes.*



Source

Cons of parallel processing

- *Race Condition:*
  - *For threads same memory and access to variables.*
  - *To avoid, use mutex (mutual exclusion) lock around code.*
- *Starvation: A thread is denied access to a resource for a long duration.*
- *Deadlock: Mutex overuse can cause deadlocks. A thread has to wait for another thread to release a lock.*
- *Livelock: threads keep running in a loop but don't make any progress.*

Threading

- *Use threading if network bound and multiprocessing if it's CPU bound.*
- *threading is perfect*
- *for I/O operations such as web scraping*
- *GUI programs, for example one for text editing, another for recording and a third one to do spell-checking.*
- *Tensorflow uses thread pool to transform data in parallel.*

Multiprocessing

- *Useful when the program is CPU intensive and not dependent on IO or user interaction.*
  - *For example, processing numbers*
  - *Pytorch Dataloaded loads data into GPU*