Introduction to Parallel Computing

Luis Villaseñor
Universidad de Morelia

2020 Annual Meeting of the
Cosmic Ray Division
of the Mexican Physical Society

November 23rd to 26th, 2020

Link to the Jupyter notebook that contains the code for this workshop (open in Google Colab):

https://drive.google.com/file/d/1iizpk5gle48esIfEXRv44SQFdU9SOuVM/view?usp=sharing

Part 1: CPU

Google Colab is a free cloud service that, in it's own words, "allows you to share Jupyter notebooks with others without having to download, install, or run anything on your own computer other than a browser!" Now you may be asking,

What Are Jupyter Notebooks?

Jupyter Notebook is an open-source web application that gives you the power to create and share documents that have live code, equations, visualizations, and narrative text. Jupyter Notebook is a interactive notebook that has many applications. It supports over 40 programming languages including Python, R, Julia, Scala, and even more (meaning that it can run code written in all those languages) It's even been used for machine learning! Your Jupyter Notebooks are stored in somewhere aptly named as "my binder" which holds all those notebooks.

https://medium.com/@dinaelhanan/an-absolute-beginners-guide-to-google-colaboratory-d55c0eb375de

Fastest Computers

Fugaku, Tokio, Japan, 442 petaFLOPS

More than 7.6 million cores

See part 2 for an introduction to parallel computing with GPUs

Summit, ORNL, USA 148.6 petaFLOPS

9216 POWER9 22-core CPUs
27.648 Nvidia Tesla Volta GV100 GPUs

Summit is now playing a critical role in the global race to discover treatments and vaccines against COVID-19.

Fastest Computers

LNS, BUAP, Puebla, Mexico, > 200 TeraFLOPS

1. Cuetlaxcoapan: Se conforma por un clúster con procesadores Intel Xeon Haswell

El clúster Intel Xeon familia Haswell cuenta con:

228 nodos de cálculo Thin (5472 núcleos). 2 CPU de 12 cores y 128 GB RAM.

20 nodos de cálculo Fat (480 núcleos). 2 CPU de 12 cores y 512 GB de RAM.

20 nodos de cálculo semi Fat (480 núcleos). 2 CPU de 12 cores y 256 GB de RAM.

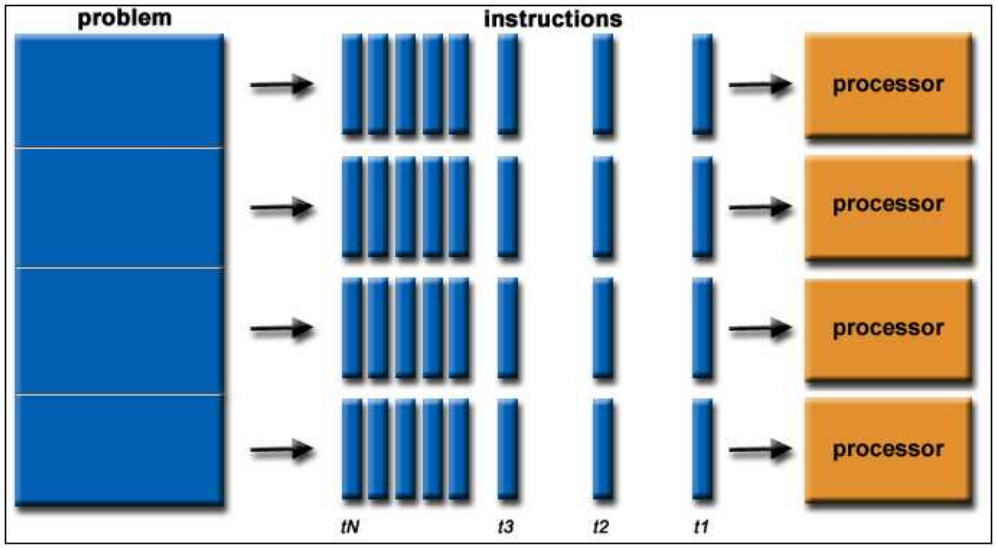
2. Centepetl. Clúster con procesadores Intel Knights Landing.

45 nodos de cálculo (3060 núcleos). UN CPU Knights Landing de 68 cores y 384 GB de RAM

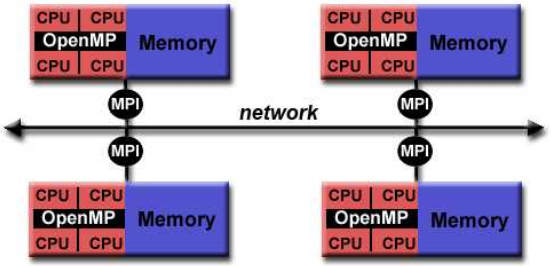
Parallel Computing Definition

Parallel computing is a type of computing architecture in which several processors simultaneously execute multiple, smaller calculations broken down from an overall larger, complex problem.

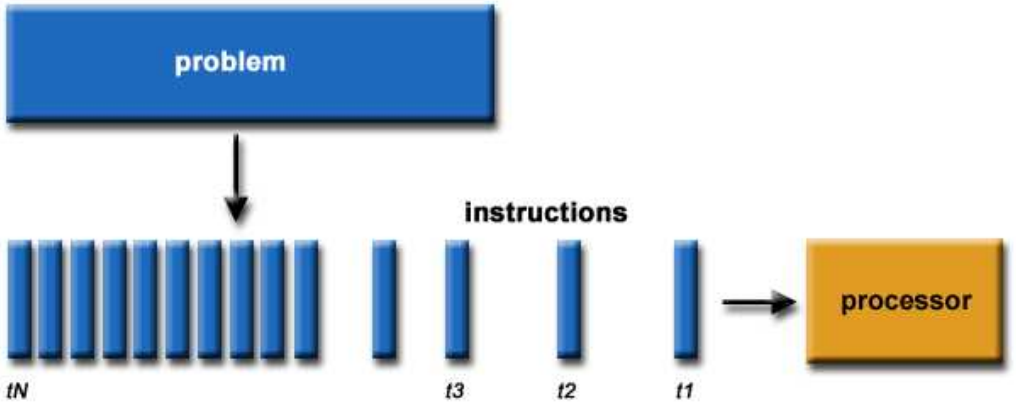
Parallel Computing



Shared (OpenMP) and Distributed Memory (MPI) parallel computing



Serial Computing



Flynn's Classical Taxonomy

SISD Single Instruction stream Single Data stream	**SIMD** Single Instruction stream Multiple Data stream
MISD Multiple Instruction stream Single Data stream	**MIMD** Multiple Instruction stream Multiple Data stream

In this workshop we will study an introduction to MPI

https://computing.llnl.gov/tutorials/parallel_comp/

Degree of parallelizability

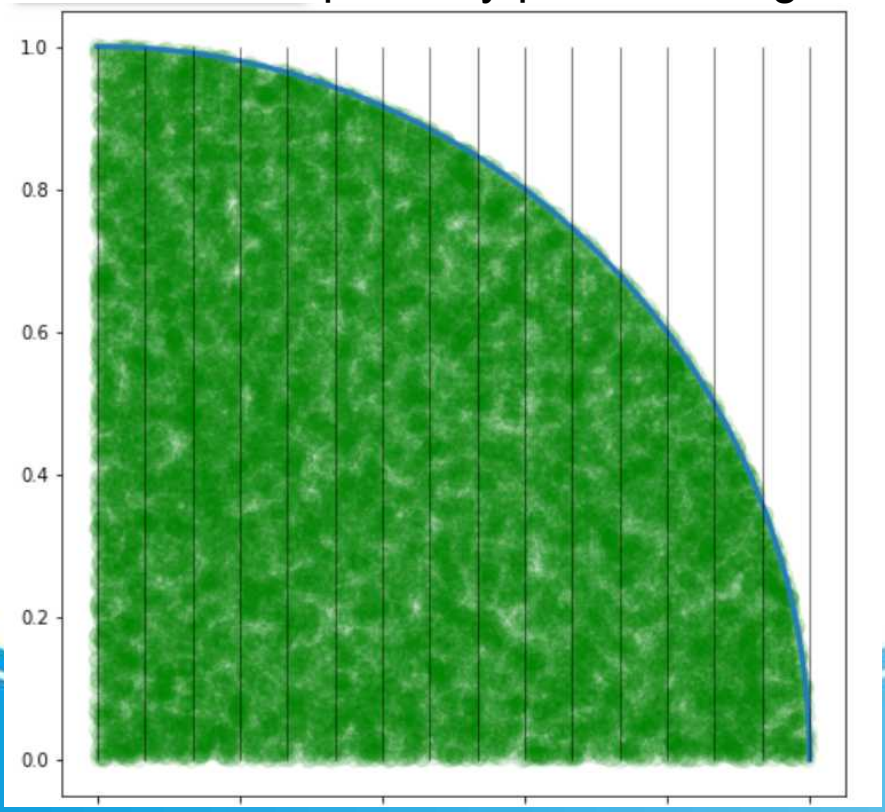
M P I = Message Passing Interface

We will use an integral to illustrate these topics

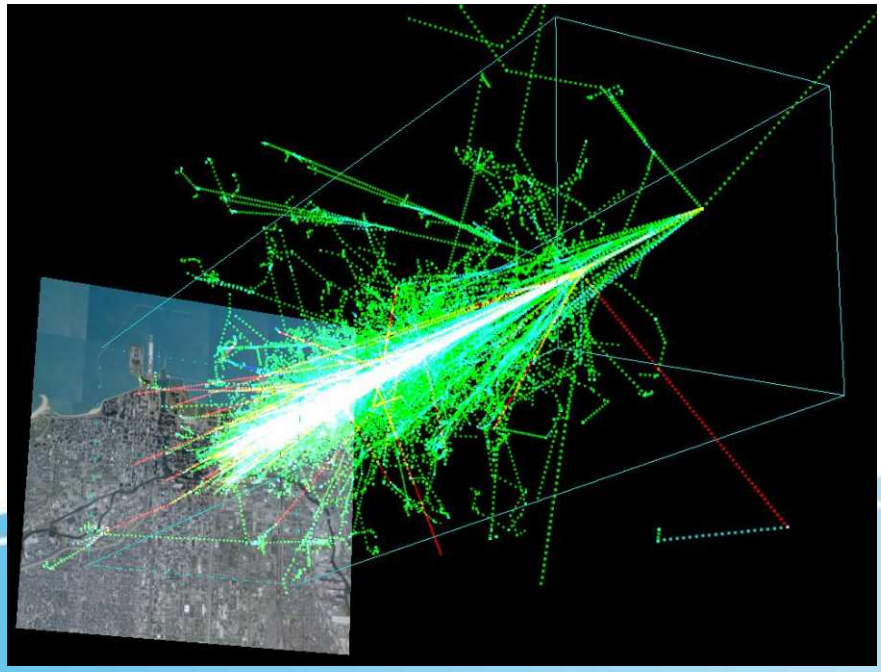
$$\pi = 4 \int_0^1 (1 - x^2)^{1/2} dx$$

Embarrassingly parallel

also called perfectly parallel, delightfully parallel or pleasingly parallel



Embarrassingly parallel



In the other extreme are the **inherently sequential or serial problems**

Part 2: GPU

Graphics processing unit, a specialized processor originally designed to accelerate graphics rendering. GPUs can process many pieces of data simultaneously, making them useful for machine learning, video editing, and gaming applications.

Nvidia TESLA K80 24GB GDDR5 PCIe x16 (900-22080-0000-000)

Special offer. New/unused item.

Key Features

- 560 MHz Core - Boostable to 876 MHz
- 4992 CUDA Cores
- 24GB GDDR5 vRAM
- 10 GHz Effective Memory Clock

CPU vs GPU

Frontier

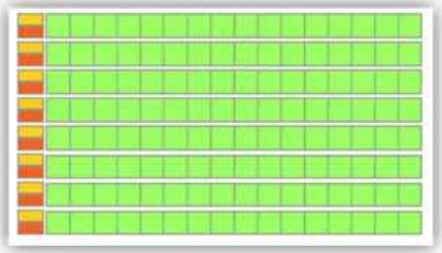
Operators	Oak Ridge National Laboratory and U.S. Department of Energy
Location	Oak Ridge Leadership Computing Facility (planned)
Power	30 MW
Space	2225 m^2 (7,300 sq ft)
Speed	1.5 exaFLOPS (estimated speed)
Cost	US$600M (estimated cost)
Purpose	Scientific research

SYSTEM SPECS	TITAN	SUMMIT	FRONTIER
Peak Performance	27 PF	200 PF	> 1.5 EF
Cabinets	200	256	> 100
Node	1 AMD Opteron CPU 1 NVIDIA K20X Kepler	2 IBM POWER9™ CPUs 6 NVIDIA Volta GPUs	1 HPC and AI Optimized AMD EPYC CPU 4 Purpose Built AMD Radeon Instinct GPU

CPU

* Low compute density
* Complex control logic
* Large caches (L1$/L2$, etc.)
* Optimized for serial operations
  - Fewer execution units (ALUs)
  - Higher clock speeds
* Shallow pipelines (<30 stages)
* Low Latency Tolerance
* Newer CPUs have more parallelism

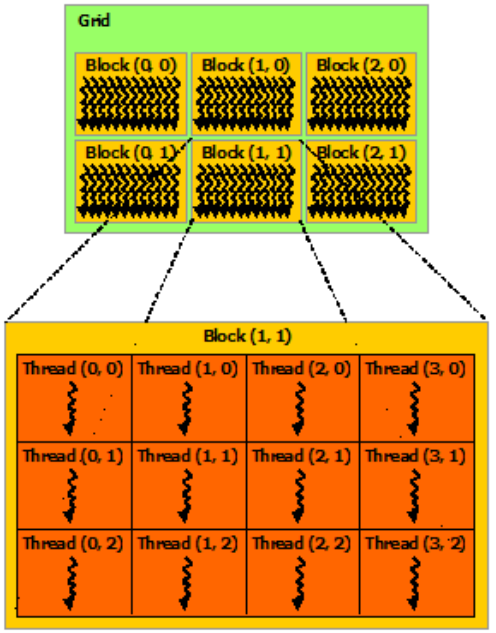
GPU



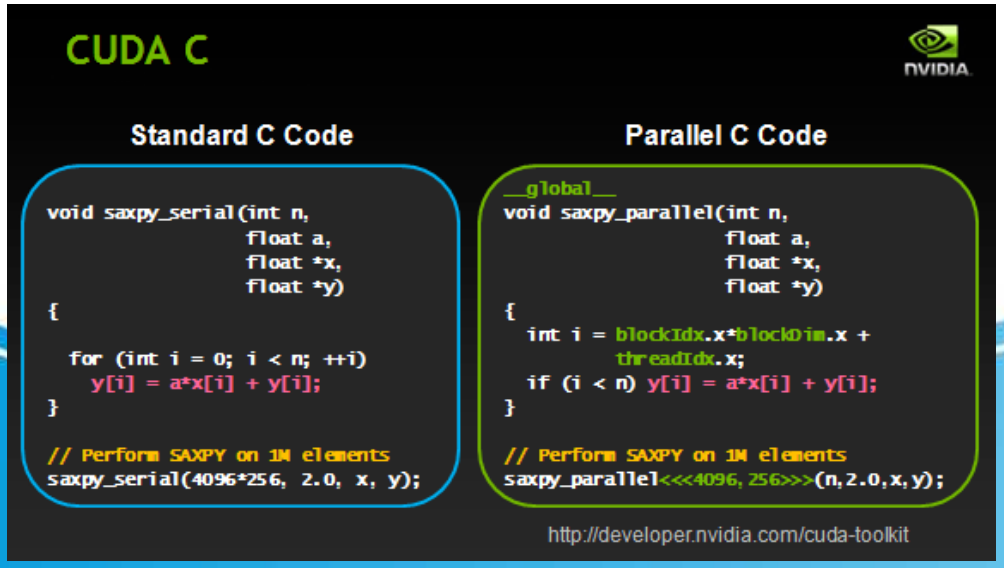
* High compute density
* High Computations per Memory Access
* Built for parallel operations
  - Many parallel execution units (ALUs)
  - Graphics is the best known case of parallelism
* Deep pipelines (hundreds of stages)
* High Throughput
* High Latency Tolerance
* Newer GPUs:
  - Better flow control logic (becoming more CPU-like)
  - Scatter/Gather Memory Access
  - Don't have one-way pipelines anymore

Compute Unified Device Architecture (CUDA)

CUDA C (2006) is essentially C/C++ with a few extensions that allow one to execute functions on the GPU using many threads in parallel. Third party wrappers are also available for Python, Perl, Fortran, Java, Ruby, Lua, Common Lisp, Haskell, R, MATLAB, IDL, Julia, and native support in Mathematica.



Example
3x2x1 grid
4x3x1 blocks



Example
2x2x2 grid
3x3x3 blocks