Rapids - cuDF

SPEEDING UP PANDAS WITH YOUR GPU

WHAT IS RAPIDS ?

- **RAPIDS is an open source project supported by NVIDIA, geared towards Data Science applications**
- **RAPIDS is a suite of libraries which aims at accelerating analytical and data science tasks by putting to work your computer's GPU**
- **The RAPIDS collection of libraries includes libraries such as:**
  - **cuGraph : a graph analytics library**
  - **cuML : a library for machine learning**
  - **cuDF : a dataframe manipulation library**

cuDF : The RAPIDS library for dataframes

- **cuDF is a Python GPU library focused on accelerating dataframe manipulation**
- **Very similar to the PANDAS library, both in terms of functionality and actual code writing**
  → ***switching an existing code from pandas to cuDF will be intuitive and require minimal code changes***

Ex: opening a csv file:

With Python:

```
df1 = pd.read_csv('effect_covid.csv')
```

With cuDF:

```
df2 = cudf.read_csv('effect_covid.csv')
```

DOCUMENTATION ON cuDF

- **Link to the cuDF documentation : https://docs.rapids.ai/api/cudf/stable/api.html#dataframe**

- **In this link, you will be able to find all the functions there is in the cuDF library, which you can then apply on your data**

`isna` ()	Identify missing values.
`isnull` ()	Identify missing values.
`iteritems` ()	Iterate over column names and series pairs
`join` (other[, on, how, lsuffix, rsuffix, ...])	Join columns with other DataFrame on index or on a key column.
`keys` ()	Get the columns.
`kurt` ([axis, skipna, level, numeric_only])	Return Fisher's unbiased kurtosis of a sample.
`kurtosis` ([axis, skipna, level, numeric_only])	Return Fisher's unbiased kurtosis of a sample.
`label_encoding` (column, prefix, cats[, ...])	Encode labels in a column with label encoding.
`log` ()	Get the natural logarithm of all elements, element-wise.
`mask` (cond[, other, inplace])	Replace values where the condition is True.
`max` ([axis, skipna, level, numeric_only])	Return the maximum of the values in the DataFrame.
`mean` ([axis, skipna, level, numeric_only])	Return the mean of the values for the requested axis.

WHAT DIFFERENTIATES cuDF FROM PANDAS ?

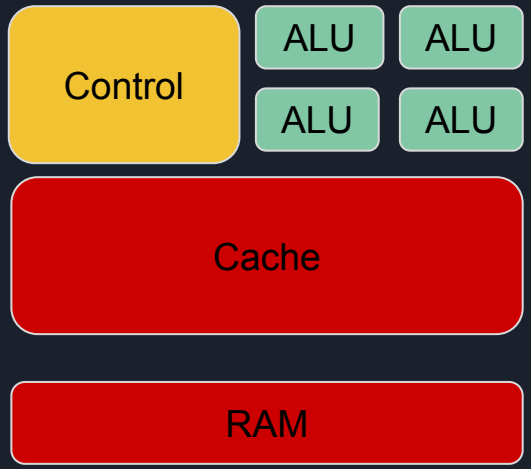
- **The key difference between PANDAS and cuDF when manipulating data is the speed at which you will be able to do so**

- **PANDAS functions are already optimized in terms of speed : using PANDAS functions to manipulate data instead of basic Python code is already a significant step up. However PANDAS is running on the CPU only. Given the size of modern datasets, it can quickly show its limitation.**

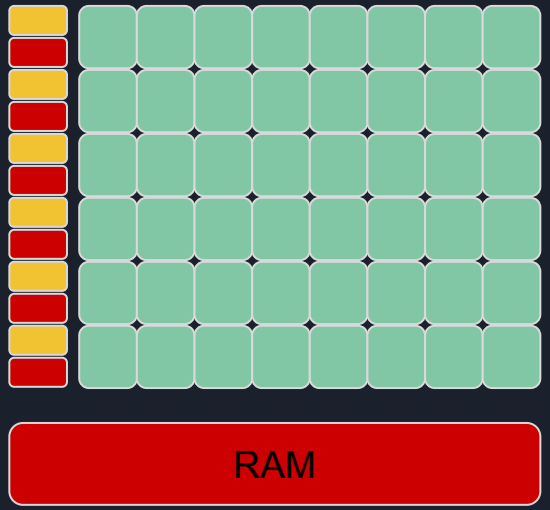
- **On the other hand, cuDF relies on the GPU which allows it to benefit from significant acceleration since GPUs have more cores than CPUs typically do**

HOW DOES IT WORK ?

CPU

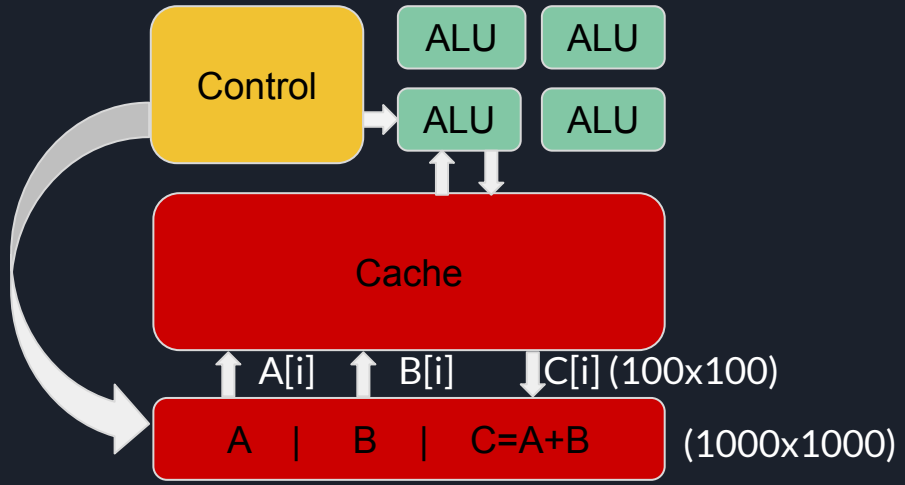


GPU

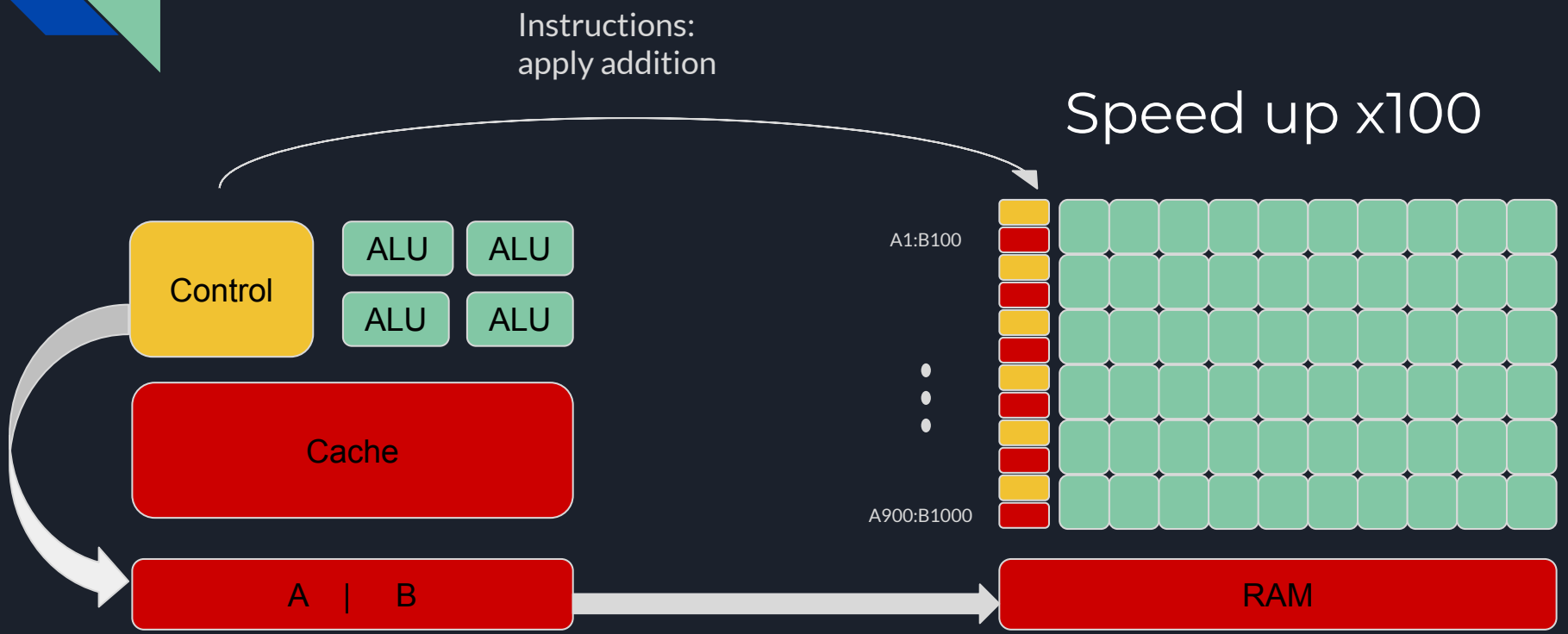


Pandas : HOW DOES IT WORK ?

Pandas always uses only one ALU unless the multiprocessing python library is used.



cuDF : HOW DOES IT WORK ?



➢ *→ going to the GPU allows a higher amount of parallel processing*

HOW TO INSTALL RAPIDS / cuDF?

Next: Google Colab