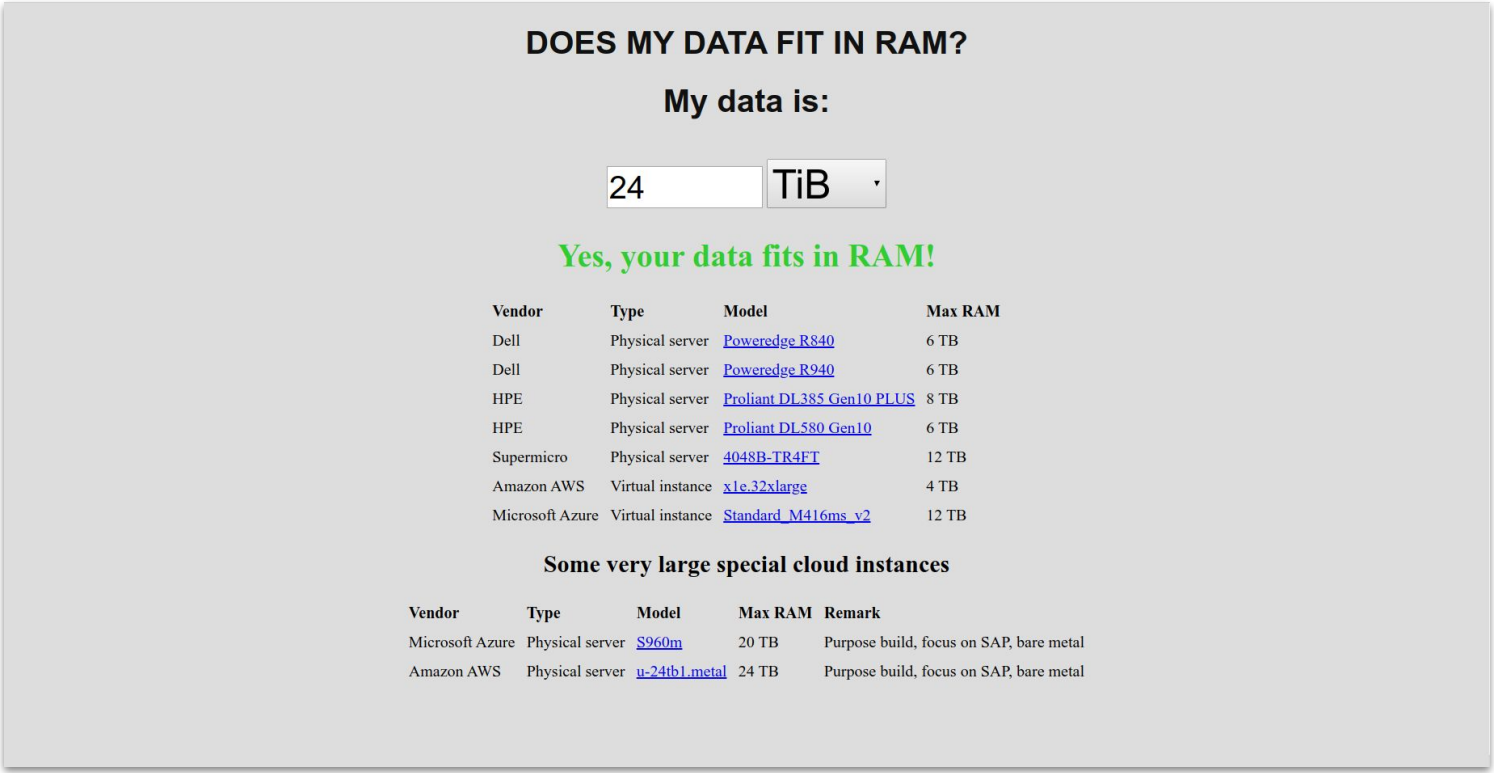
Introduction to using GPUs for Analytics

PyData PHL - February 18, 2020

Randy Zwitch
Senior Director of Community

@randyzwitch
randy.zwitch@omnisci.com
/in/randyzwitch/
/randyzwitch

Your Data Fits in RAM...but that isn't the point



DOES MY DATA FIT IN RAM?

My data is:

24 TiB

Yes, your data fits in RAM!

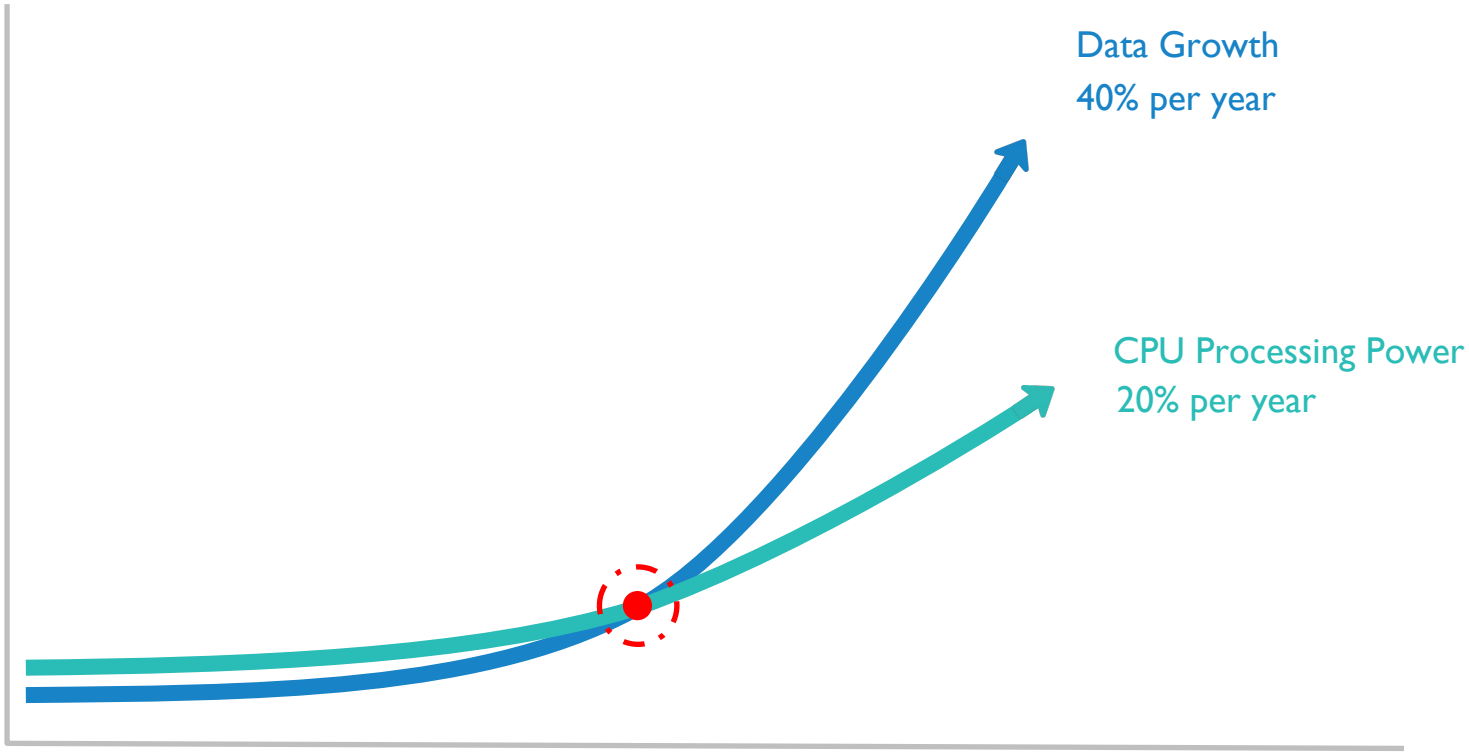
Vendor	Type	Model	Max RAM
Dell	Physical server	Poweredge R840	6 TB
Dell	Physical server	Poweredge R940	6 TB
HPE	Physical server	Proliant DL385 Gen10 PLUS	8 TB
HPE	Physical server	Proliant DL580 Gen10	6 TB
Supermicro	Physical server	4048B-TR4FT	12 TB
Amazon AWS	Virtual instance	x1e.32xlarge	4 TB
Microsoft Azure	Virtual instance	Standard_M416ms_v2	12 TB

Some very large special cloud instances

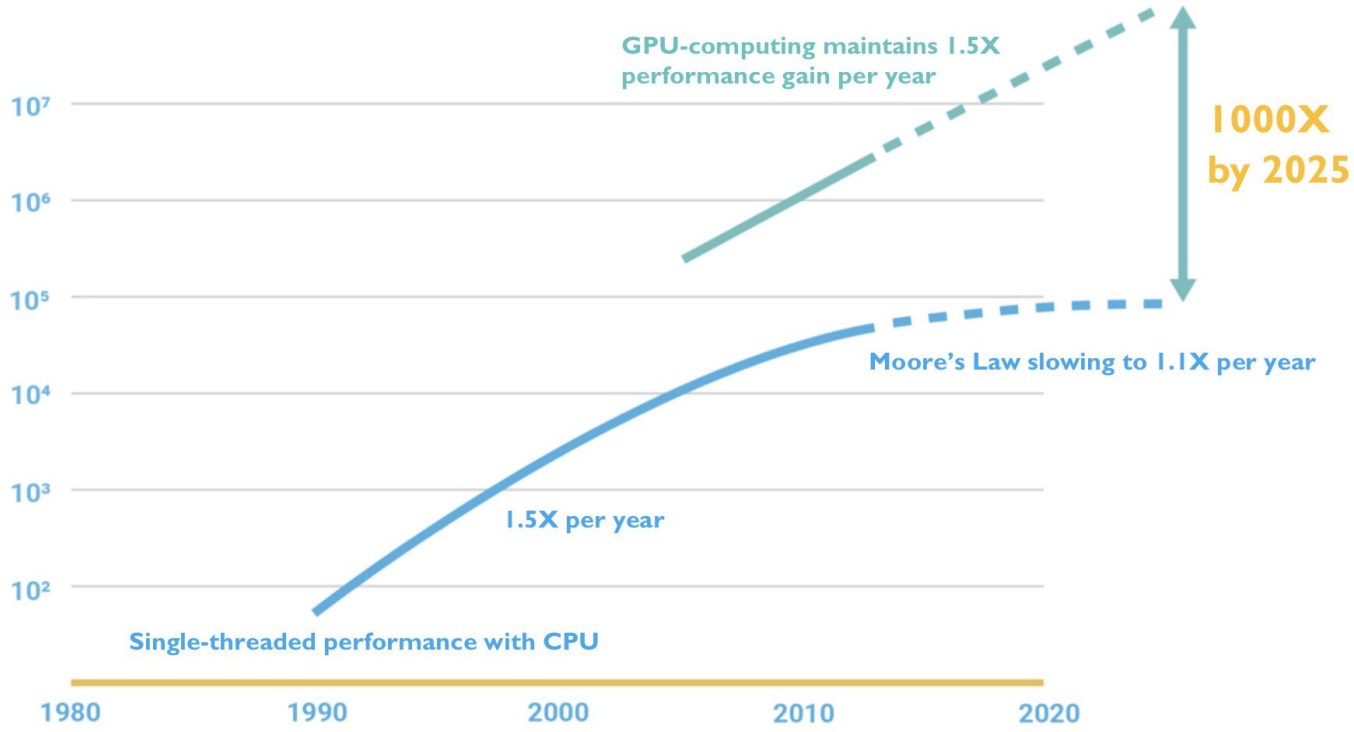
Vendor	Type	Model	Max RAM	Remark
Microsoft Azure	Physical server	S960m	20 TB	Purpose build, focus on SAP, bare metal
Amazon AWS	Physical server	u-24tb1.metal	24 TB	Purpose build, focus on SAP, bare metal

Source: https://yourdatafitsinram.net/

Data Grows Faster Than CPU Processing

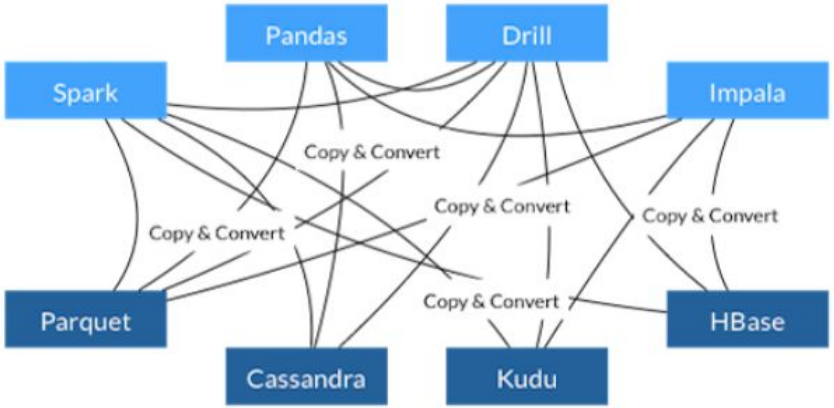


GPU Processing Keeps Moore's Law Alive

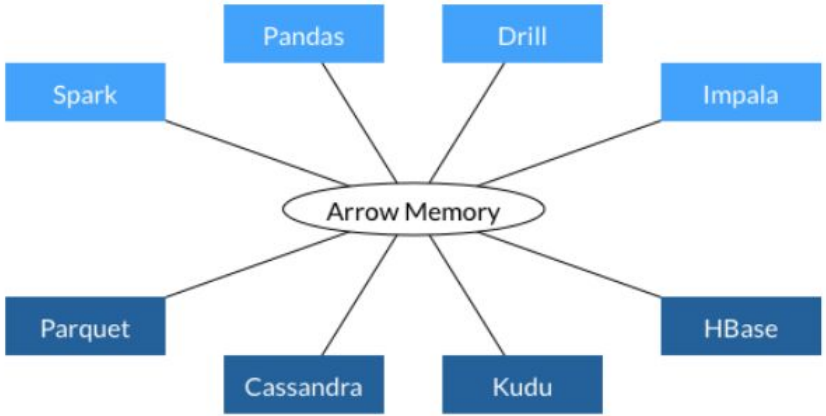


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

Apache Arrow: Shared Memory Layout



- Each system has its own internal memory format
- 70-80% computation wasted on serialization and deserialization
- Similar functionality implemented in multiple projects



- All systems utilize the same memory format
- No overhead for cross-system communication
- Projects can share functionality (eg, Parquet-to-Arrow reader)

Source: https://arrow.apache.org/

RAPIDS and the GPU DataFrame

Born from the GPU Open Analytics Initiative – fusing Machine Learning and GPU analytics

CONTRIBUTORS

ANACONDA. BLAZINGDB GUNROCK NVIDIA. QUANSIGHT Walmart

ADOPTERS

Chainer databricks graphistry $H_2O.ai$ IBM iguazio Inria MAPR omnisci PyTorch Uber URSA LABS

OPEN SOURCE

Apache Arrow blazingSQL DASK GOAi Numba scikit learn dmlc XGBoost

By adopting a common memory layout in Apache Arrow, these tools can work seamlessly with one another with zero-copy memory transfer!

Demo: Parallelized examples using Python

GitHub: https://github.com/randyzwitch/pydataphl

Exploring Parallel Computation Using Python

Introduction to using GPUs for Analytics

Speaker: Randy Zwitch, Senior Director of Community at OmniSci
PyData PHL: https://www.meetup.com/PyData-PHL/events/268253667/
Feb 18, 2020

This notebook demostrates some of the basic principles for using GPUs to accelerate computations. It is not intended to be a primer on machine learning; rather, the intent is to help users gain an intuition about code that can be parallelized in general, then show the speed up from moving computation from CPU to GPU.

0. Example Data

[1]:
```
import pandas as pd

#1 month of bikshare data from Baywheels (SF)
#~295k records not that large, but useful for example
#full dataset: https://s3.amazonaws.com/baywheels-data/index.html
baywheels_df = pd.read_csv("https://s3.amazonaws.com/baywheels-data/202001-baywheels-tripdata.csv.zip", low_memory=False)
baywheels_df.shape
```

[1]:
```
(295854, 15)
```

[2]:
```
baywheels_df.head()
```

[2]:

	duration_sec	start_time	end_time	start_station_id	start_station_name	start_station_latitude	start_station_longitude	end_station_id	end_station_name
0	83118	2020-01-31 15:23:47.7330	2020-02-01 14:29:06.2630	400.0	Buchanan St at North Point St	37.804272	-122.433537	400.0	Buchanan St at North Point St
1	68239	2020-01-31 15:40:31.6160	2020-02-01 10:37:51.0000	99.0	Folsom St at 15th St	37.767037	-122.415443	99.0	Folsom St at 15th St
2	55407	2020-01-31	2020-02-01	197.0	El Embarcadero at	37.808848	-122.249680	197.0	El Embarcadero at

Questions

Randy Zwitch
Senior Director of Community

@randyzwitch
randy.zwitch@omnisci.com
/in/randyzwitch/
/randyzwitch