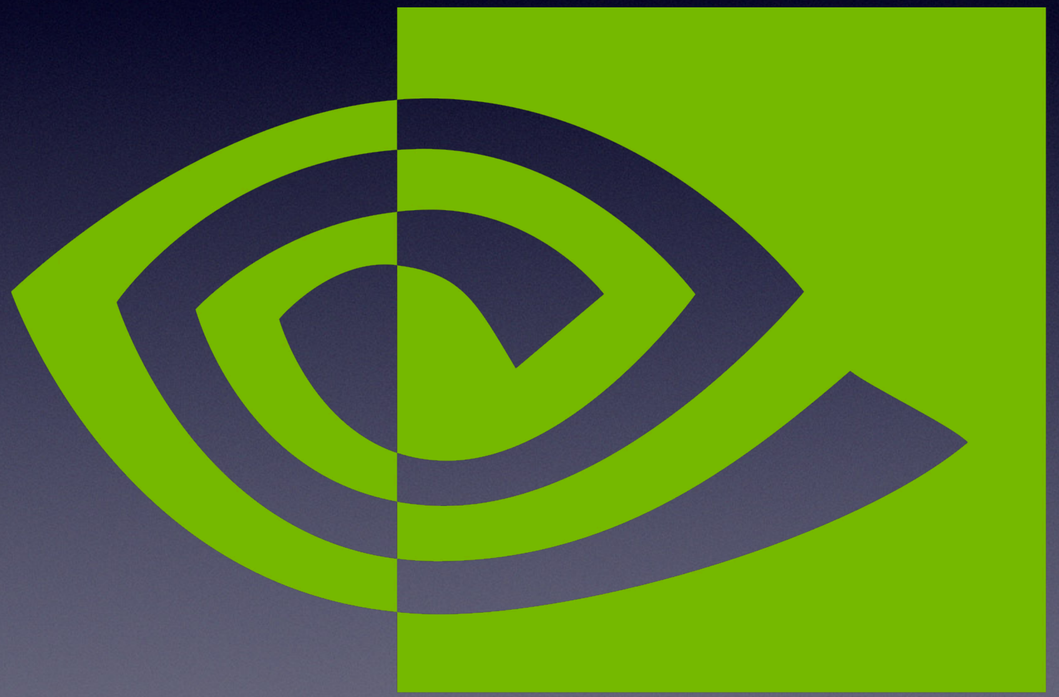
sli.do: #GPUSFTW

GPU Computing



Greg Chase

About Me

- Data Scientist, MeetMindful
- Focus Areas
  - Deep Learning
  - GPU Computing
- GOAI community member

Galvanize Capstones

- g49 scrum groups
- Trend in project completion time

Why aren’t all data scientists using GPU’s?

THE *DATA* STRUGGLE IS REAL...

DATA PROCESSING EVOLUTION

Faster Data Access Less Data Movement

Hadoop Processing, Reading from disk

DATA PROCESSING EVOLUTION

Faster Data Access Less Data Movement

SPARK IS NOT ENOUGH

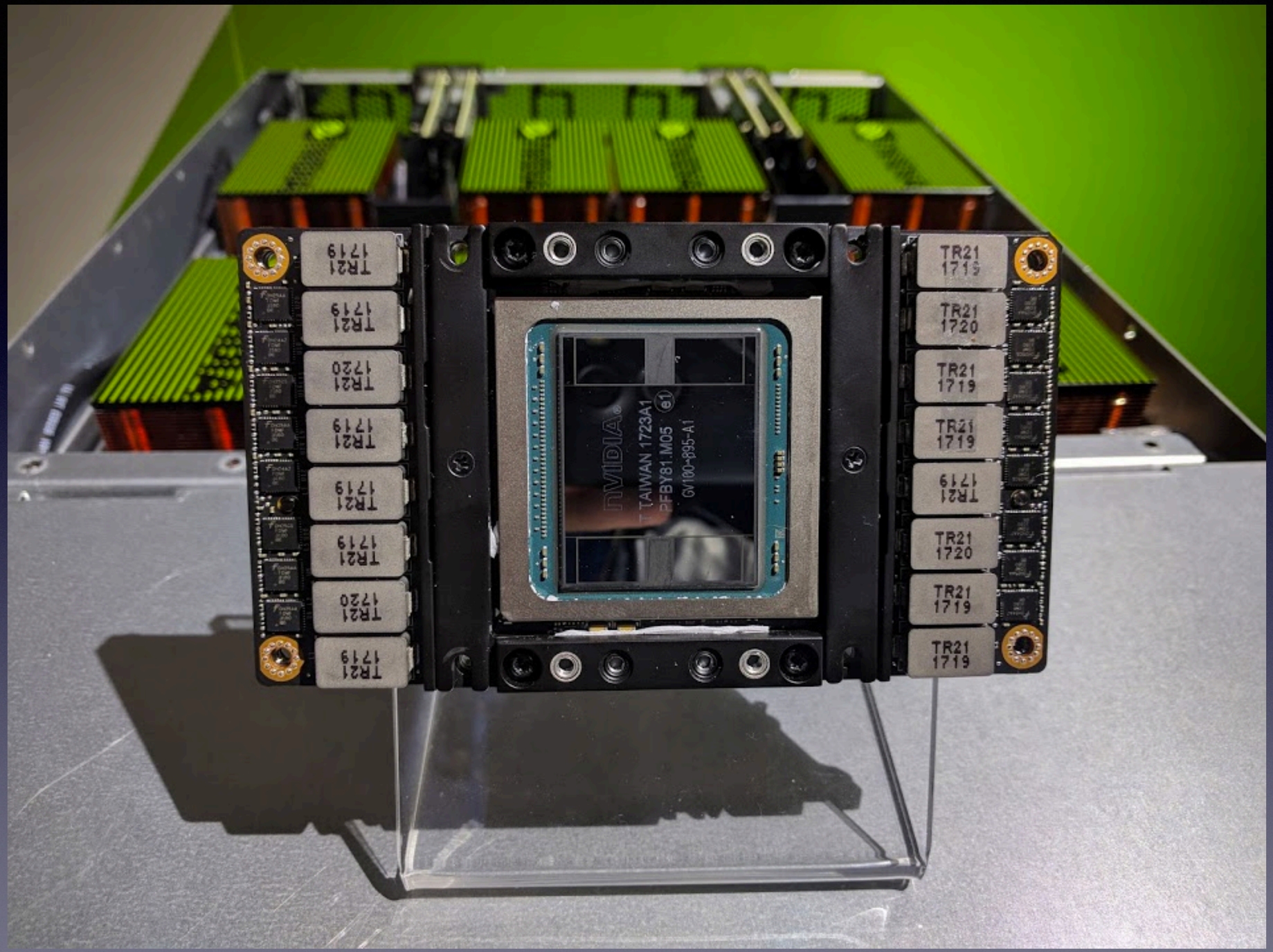
Basic workloads are bottlenecked by the CPU

- In a simple benchmark consisting of aggregating data, the **CPU is the bottleneck**
- This is after the data is parsed and cached into memory which is another common bottleneck
- The CPU bottleneck is even worse in more complex workloads!

SELECT cab_type, count(*)
FROM trips_orc
GROUP BY cab_type;

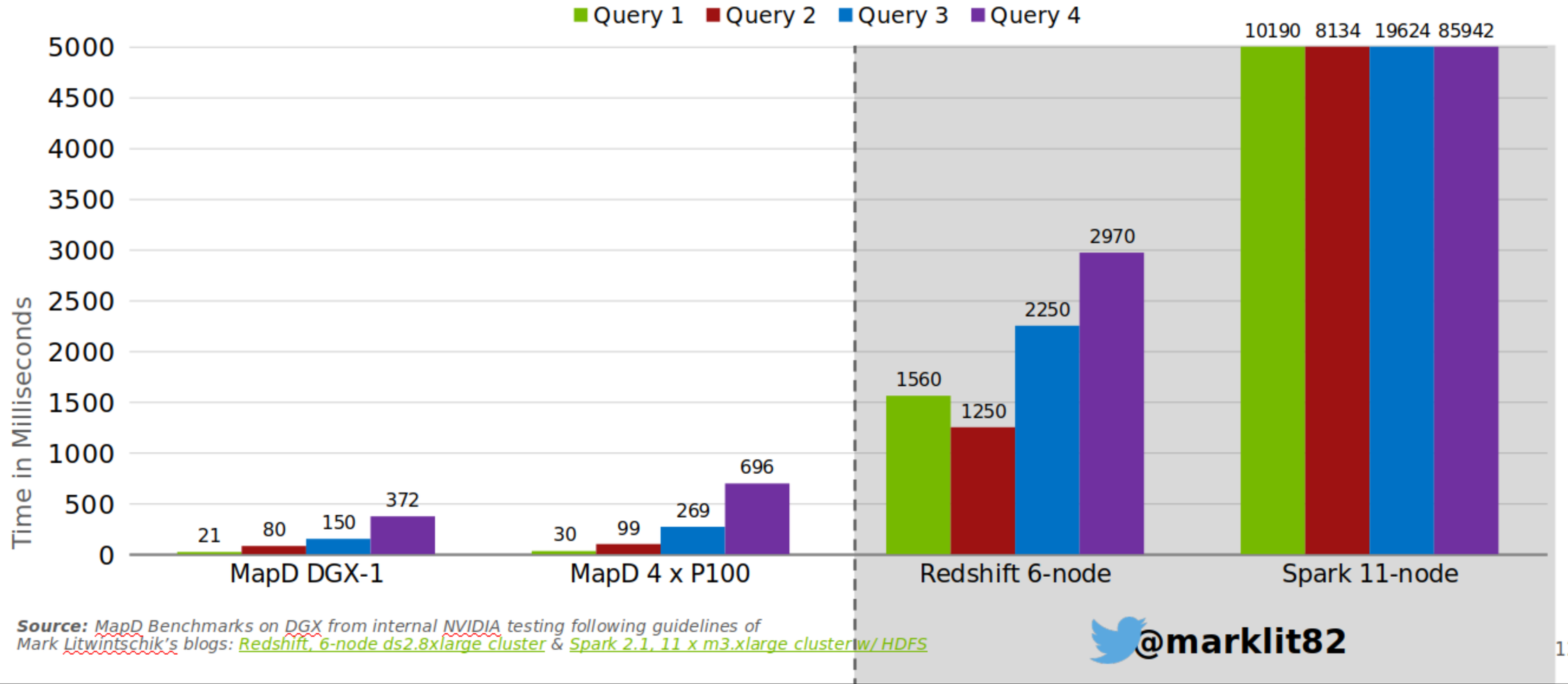
```
top - 08:54:14 up  1:50,  4 users,  load average: 0.20, 1.64, 6.43
Tasks: 360 total,   2 running, 358 sleeping,   0 stopped,   0 zombie
%Cpu0  : 94.7 us,  1.7 sy,  0.0 ni,  3.3 id,  0.0 wa,  0.0 hi,  0.3 si,  0.0
%Cpu1  : 95.0 us,  1.7 sy,  0.0 ni,  3.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu2  : 98.3 us,  0.3 sy,  0.0 ni,  1.3 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu3  : 87.3 us,  4.3 sy,  0.0 ni,  8.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu4  : 95.0 us,  1.3 sy,  0.0 ni,  3.7 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu5  : 98.3 us,  0.0 sy,  0.0 ni,  1.7 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu6  : 96.7 us,  1.3 sy,  0.0 ni,  2.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu7  : 92.7 us,  1.0 sy,  0.0 ni,  5.6 id,  0.3 wa,  0.0 hi,  0.3 si,  0.0
%Cpu8  : 93.7 us,  1.3 sy,  0.0 ni,  5.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu9  : 92.3 us,  0.7 sy,  0.0 ni,  7.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu10 : 97.3 us,  0.7 sy,  0.0 ni,  2.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu11 : 97.3 us,  0.7 sy,  0.0 ni,  2.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu12 : 92.0 us,  3.0 sy,  0.0 ni,  5.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu13 : 94.9 us,  1.0 sy,  0.0 ni,  4.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu14 : 88.3 us,  3.0 sy,  0.0 ni,  8.7 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu15 : 92.6 us,  2.3 sy,  0.0 ni,  4.7 id,  0.0 wa,  0.0 hi,  0.3 si,  0.0
%Cpu16 : 94.7 us,  2.3 sy,  0.0 ni,  2.6 id,  0.0 wa,  0.0 hi,  0.3 si,  0.0
%Cpu17 : 93.0 us,  0.7 sy,  0.0 ni,  6.0 id,  0.0 wa,  0.0 hi,  0.3 si,  0.0
%Cpu18 : 93.0 us,  3.7 sy,  0.0 ni,  3.3 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu19 : 91.2 us,  0.7 sy,  0.0 ni,  8.1 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
```

Source: *Mark Litwintschik's blog: 1.1 Billion Taxi Rides: EC2 versus EMR*


NVIDIA
TAIWAN 1723A1
PFBY81.M05
GV100-895-A1
TR21 1719
TR21 1720

GPUS ARE FAST

1.1 Billion Taxi Ride Benchmark



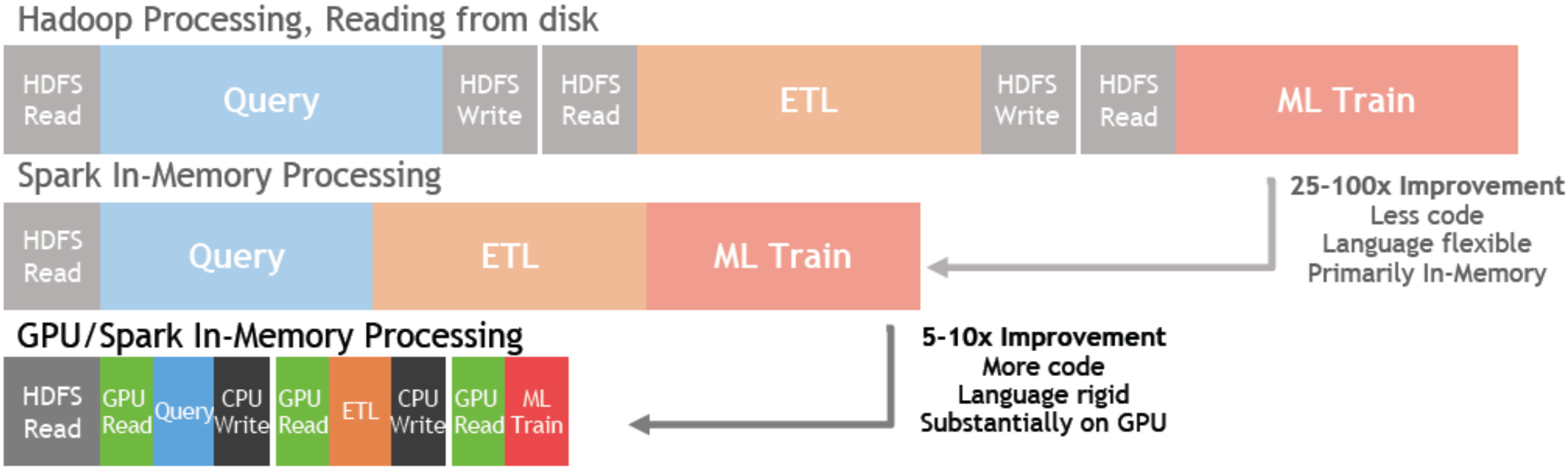
Source: *MapD Benchmarks on DGX from internal NVIDIA testing following guidelines of Mark Litwintschik's blogs: Redshift, 6-node ds2.8xlarge cluster & Spark 2.1, 11 x m3.xlarge cluster w/ HDFS*

@marklit82

NVIDIA

DATA PROCESSING EVOLUTION

Faster Data Access Less Data Movement

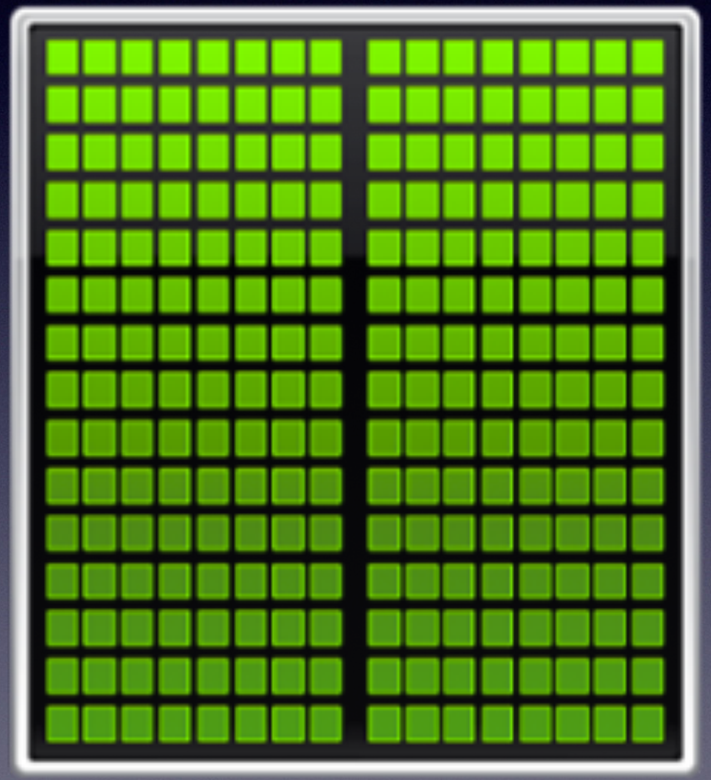


CPU

Optimized for Serial Tasks

GPU

Optimized for Many Parallel Tasks



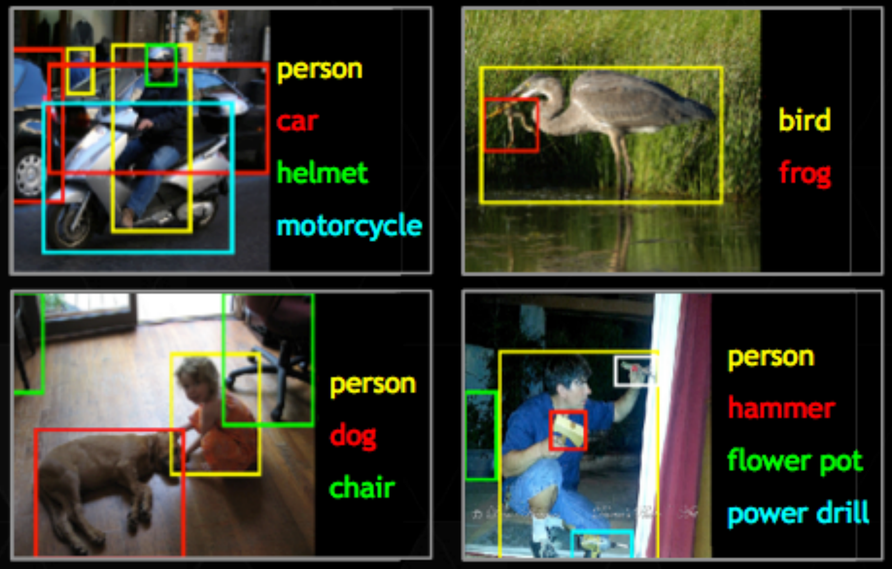
GPUs — THE PLATFORM FOR DEEP LEARNING

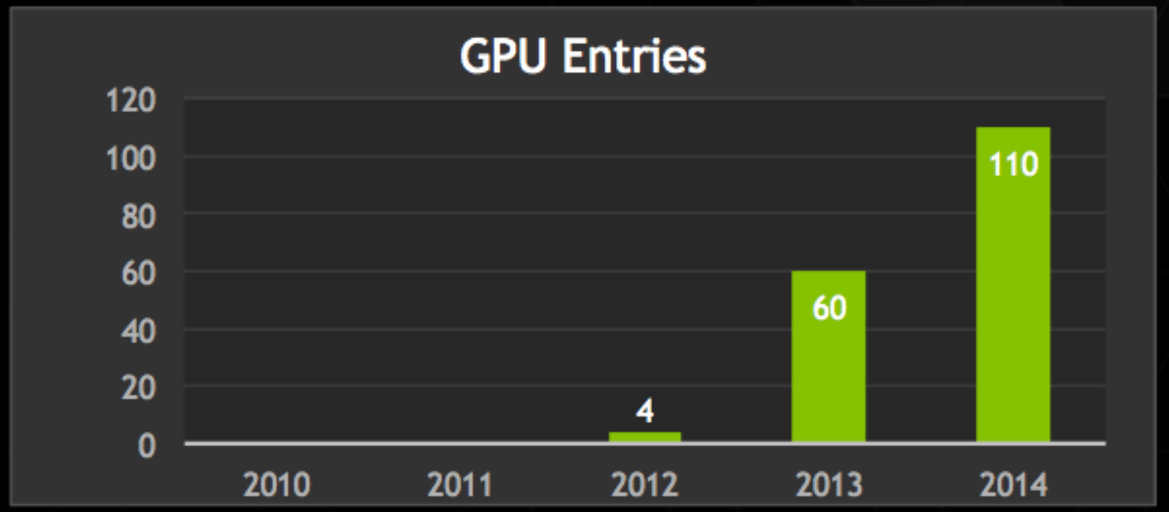
Image Recognition Challenge

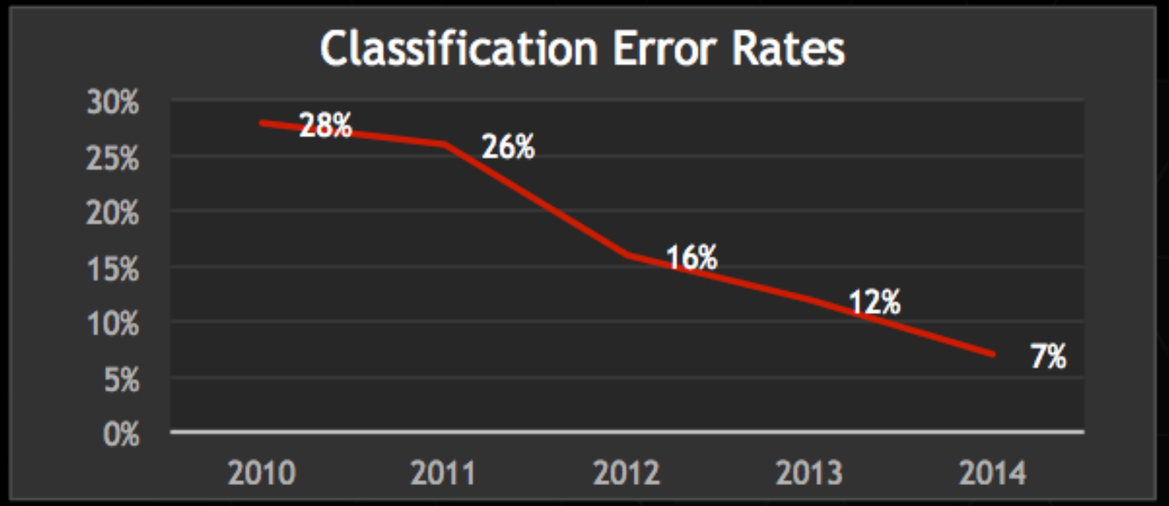
1.2M training images • 1000 object categories

Hosted by

IMAGENET







GPUS MAKE DEEP LEARNING ACCESSIBLE

Deep learning with COTS HPC systems

A. Coates, B. Huval, T. Wang, D. Wu, A. Ng, B. Catanzaro

ICML 2013

"Now You Can Build Google's $1M Artificial Brain on the Cheap"

WIRED

GOOGLE DATACENTER

1,000 CPU Servers
2,000 CPUs • 16,000 cores

600 kWatts
$5,000,000

STANFORD AI LAB

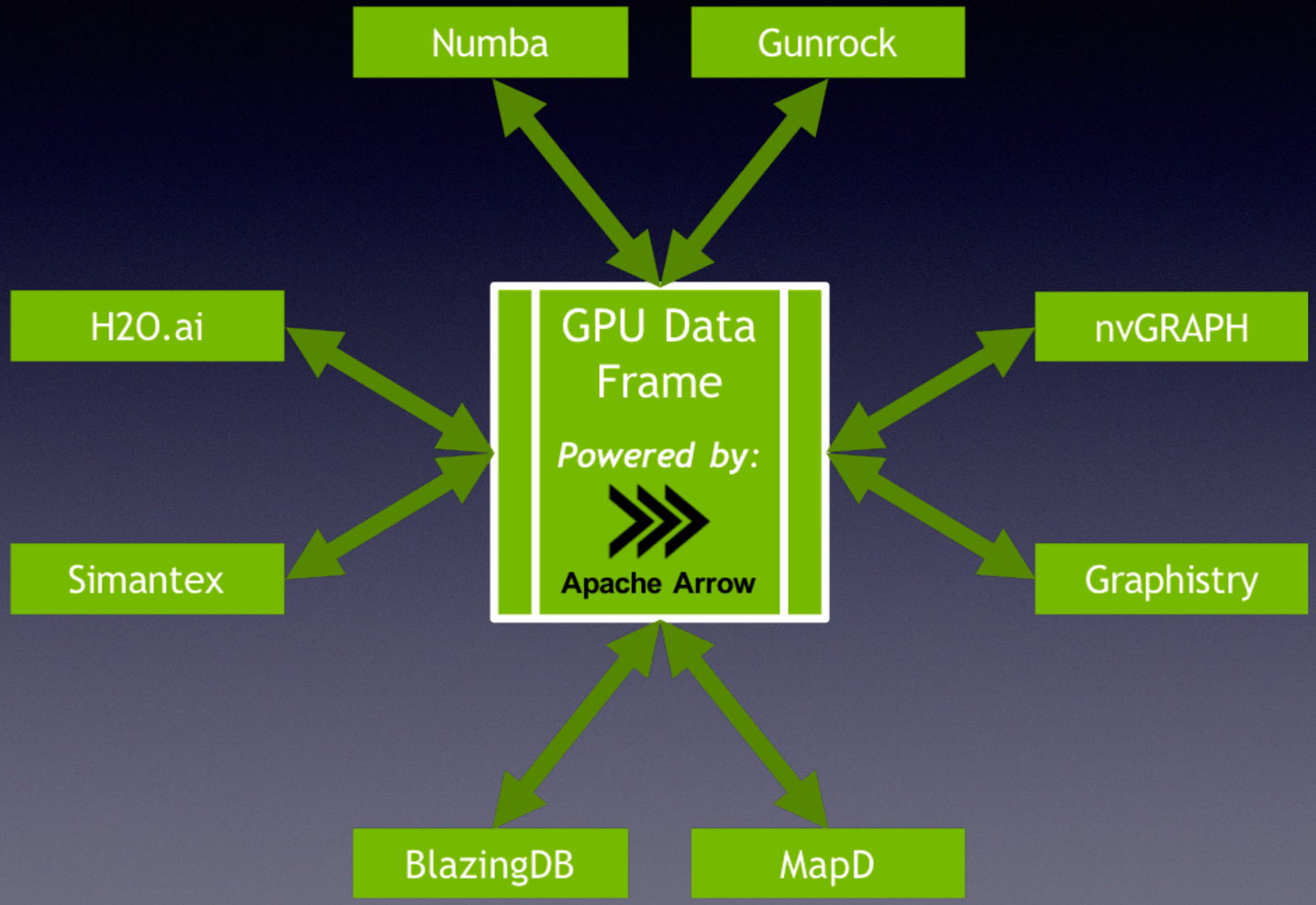
3 GPU-Accelerated Servers
12 GPUs • 18,432 cores

4 kWatts
$33,000

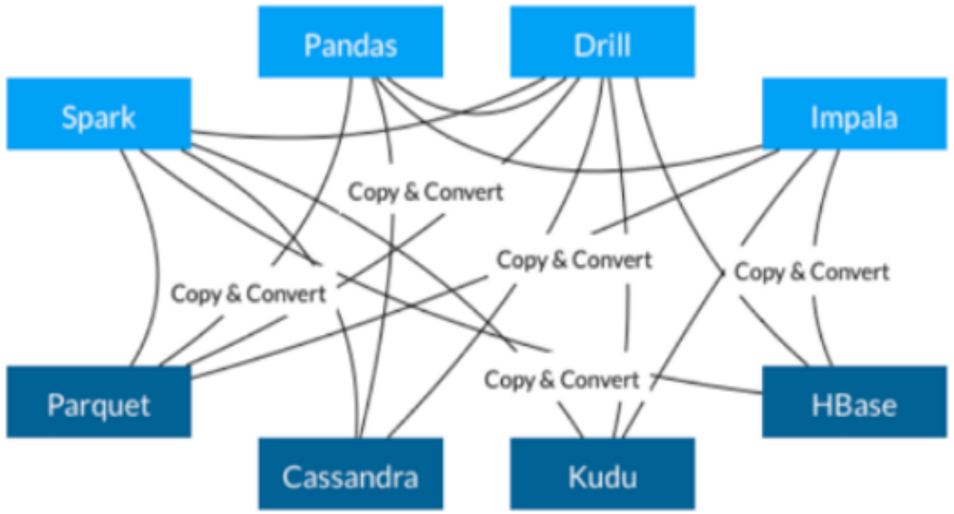
Workstation	Year Released	GPU's	Cost	Savings
Google "Brain"	2011	0	$5,000,000	$0
Stanford AI Lab	2012	12	$33,000	$4,967,000
NVIDIA DGX Station	2017	4	$50,000	$4,950,000
NVIDIA DGX-1	2017	8	$150,000	$4,850,000

GPU Libraries

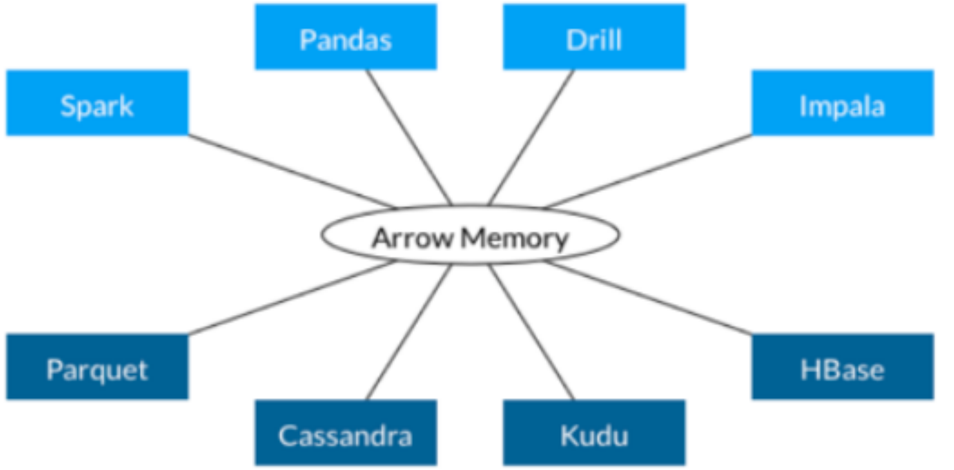
PyGDF - Pandas on GPU



APACHE ARROW COMMON DATA LAYER



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

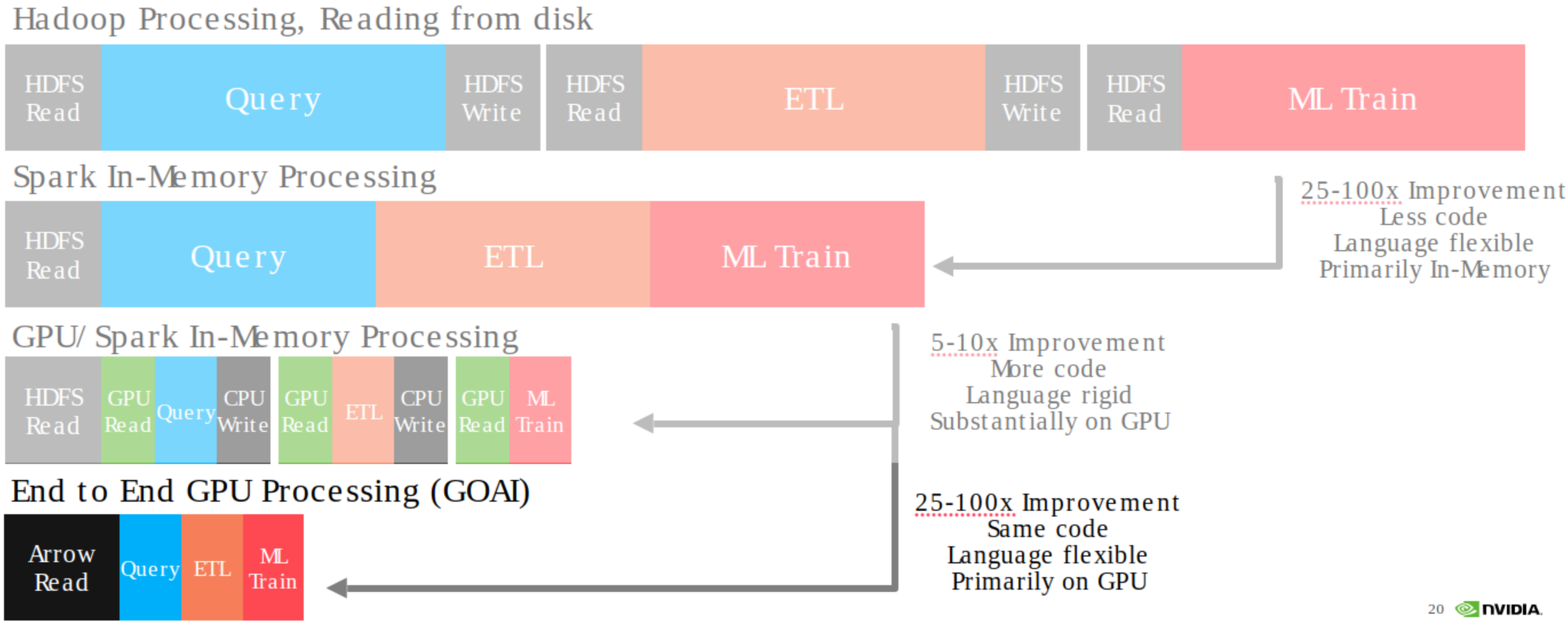


- All systems utilize the same memory format
- No overhead for cross-system communication
- Projects can share functionality (eg, Parquet-to-Arrow reader)

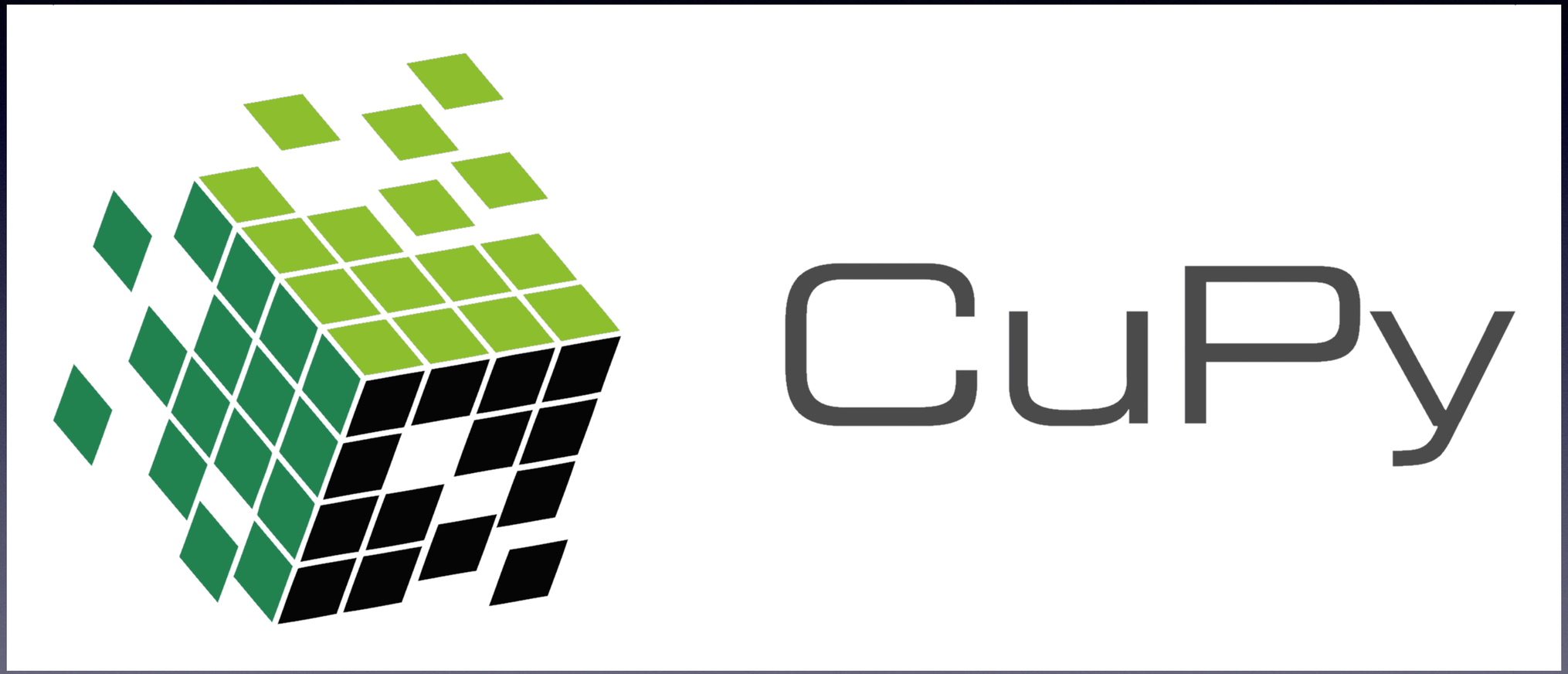
From Apache Arrow Home Page - https://arrow.apache.org/

GPU DATA FRAME

Faster Data Access Less Data Movement



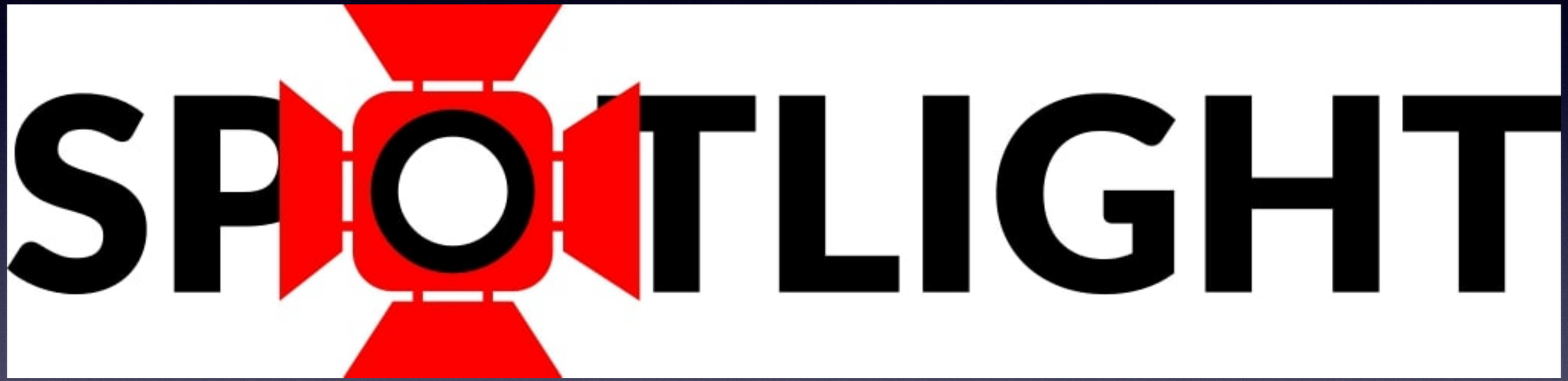
CuPy - NumPy on GPU

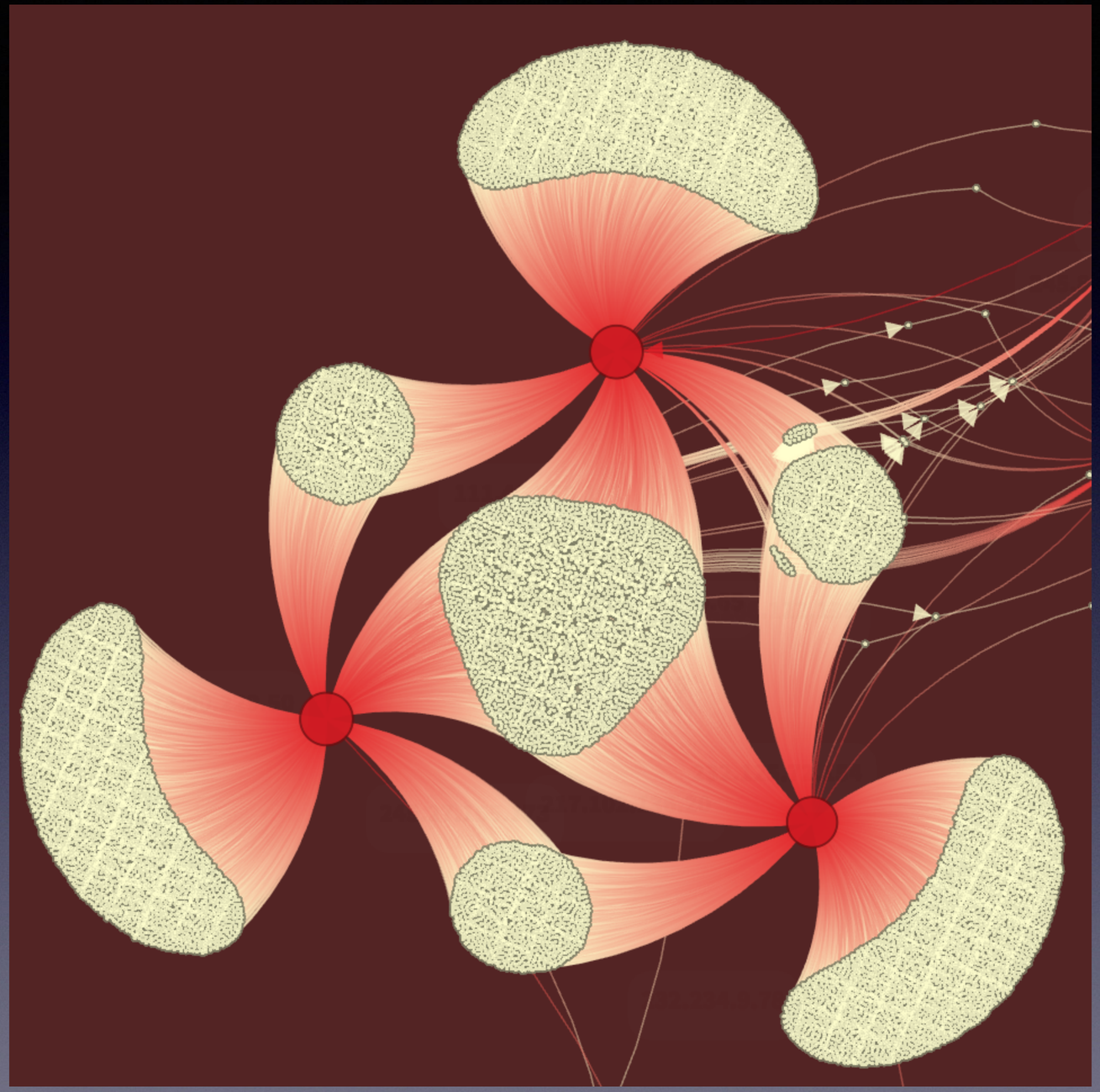


CuPy - CPU/GPU Agnostic

```
>>> # Stable implementation of log(1 + exp(x))
>>> def softplus(x):
...     xp = cp.get_array_module(x)
...     return xp.maximum(0, x) + xp.log1p(xp.exp(-abs(x)))
```

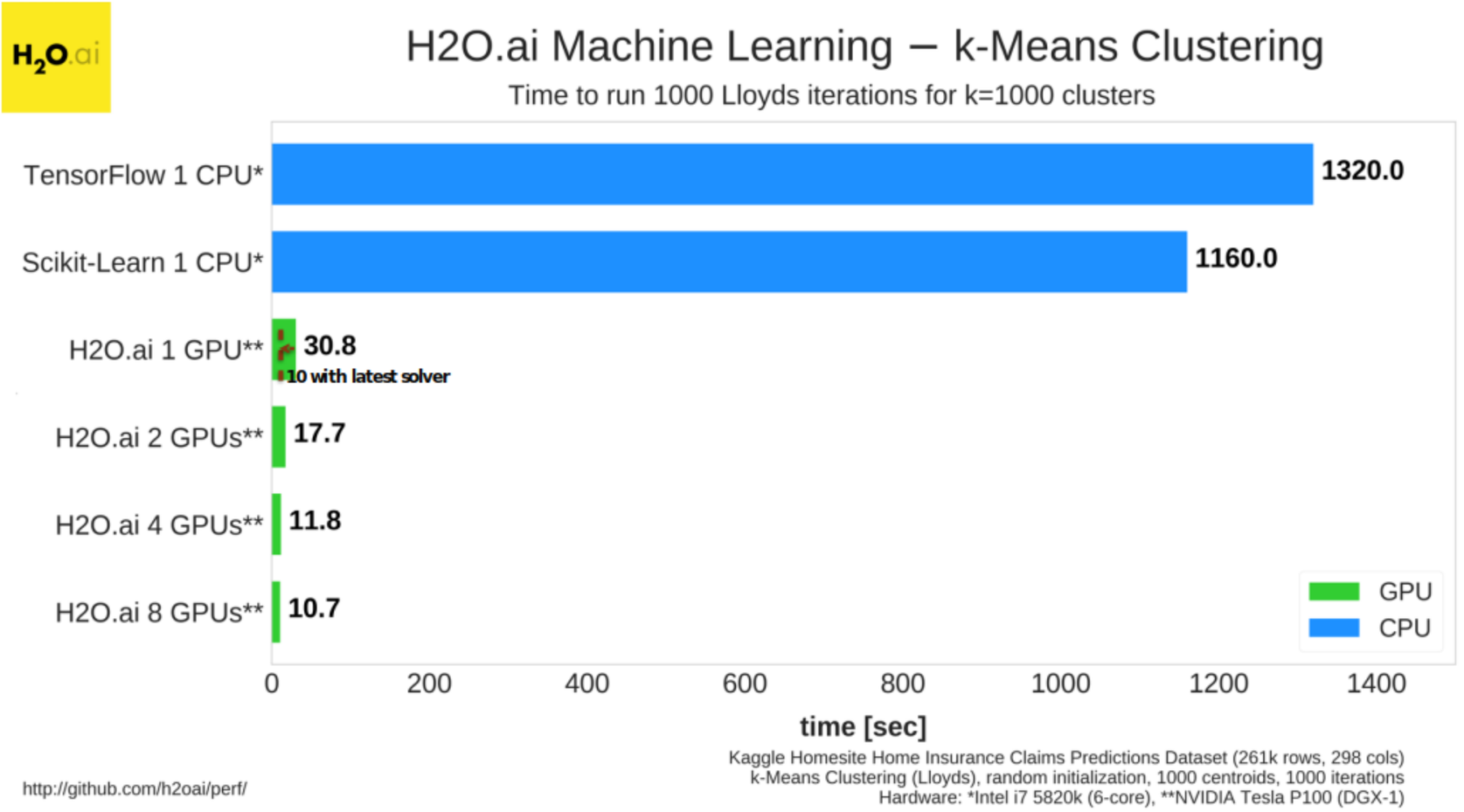
GPU Recommenders





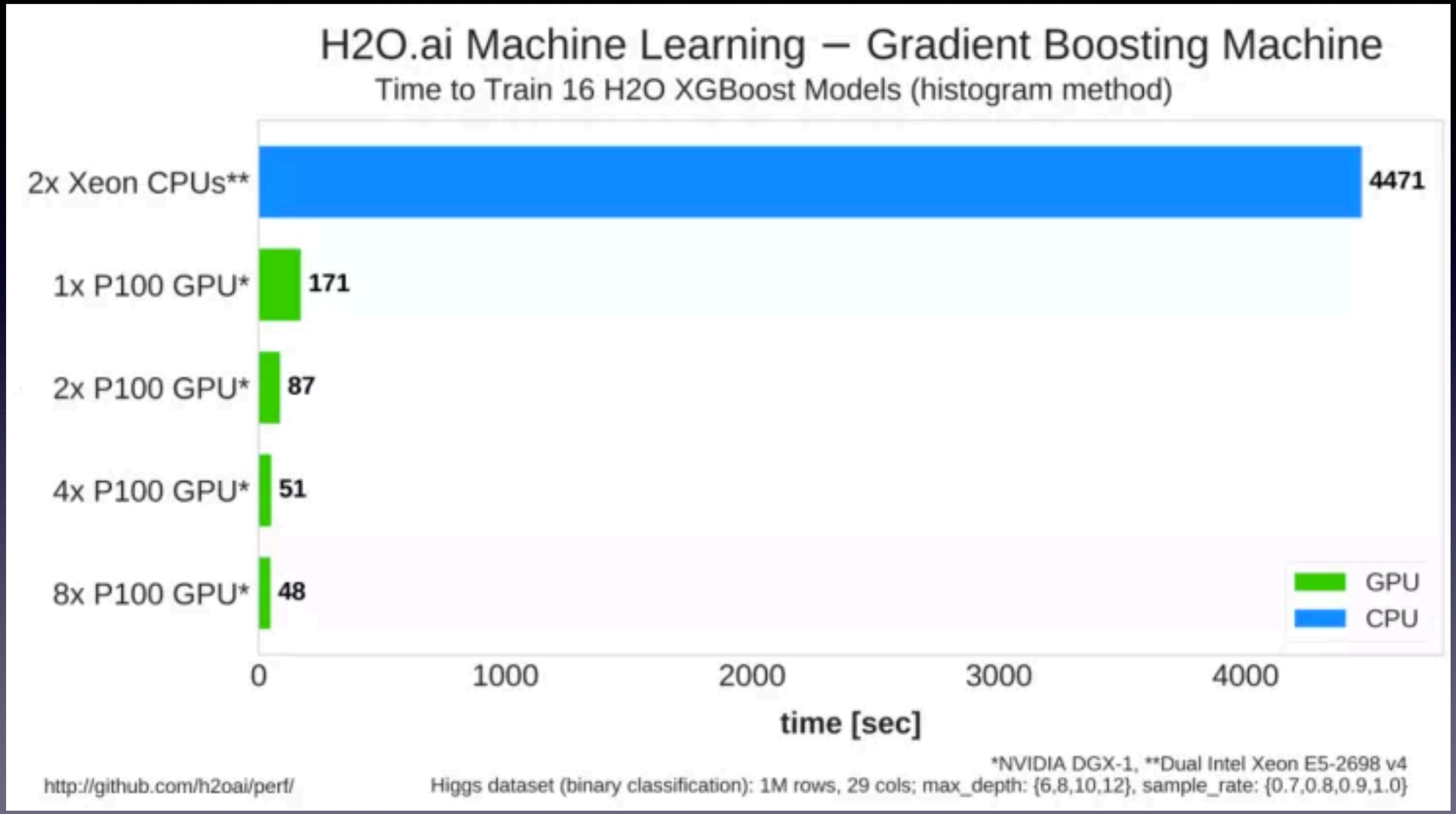
H_2O4GPU

K-Means Benchmark



Kaggle Homesite Home Insurance Claims Predictions Dataset (261k rows, 298 cols)
k-Means Clustering (Lloyds), random initialization, 1000 centroids, 1000 iterations
Hardware: *Intel i7 5820k (6-core), **NVIDIA Tesla P100 (DGX-1)

http://github.com/h2oai/perf/


H2O.ai Machine Learning – Gradient Boosting Machine
Time to Train 16 H2O XGBoost Models (histogram method)
2x Xeon CPUs**
4471
1x P100 GPU*
171
2x P100 GPU*
87
4x P100 GPU*
51
8x P100 GPU*
48
GPU
CPU
0
1000
2000
3000
4000
time [sec]
*NVIDIA DGX-1, **Dual Intel Xeon E5-2698 v4
http://github.com/h2oai/perf/
Higgs dataset (binary classification): 1M rows, 29 cols; max_depth: {6,8,10,12}, sample_rate: {0.7,0.8,0.9,1.0}

H2O4GPU Roadmap

Currently Available - Q3 (09-30-2017)
GLM (POGS)
Python API for training & scoring
GBM
Inference on GPU (GLM)
Random Forest
Inference on GPU (GBM)
k-Means Clustering

API Support
Python API for training & scoring
Sckit learn API compatibility

Q4 2017 - (12-31-2017)
k-Nearest Neighbors
PCA
SVD
Quantiles
Kalman Filters
Sort
Aggregator

API Support
R API for training & scoring
GOAI API support
Data.table

Performance & Scalability
Fastest single GPU performance
Multi GPU
Multi machine

2018-19
Kernel Methods
Recommendation Engines - Non-Negative Matrix Factorization
Recommendation Engines - Bayesian Neural Nets
MCMC Solver
Time Series
SVM
Text Analysis - TF-IDF
Text Analysis - Word2Vec
Text Analysis - Doc2Vec
Automatic K for K-means
H2O GLM - Lasso
Simulation Techniques
Sampling Techniques

Domain Specific Algorithms
Life Sciences
Financial Services
Underwriting
Sampling Techniques

H2O.ai

Why aren’t all data scientists using GPU’s?

You’re no longer limited by time.

You’re now limited by ingenuity.

GPU's help you accomplish your life's work.

Special Thanks

- NVIDIA
  - Keith Kraus (slides)
  - Eric Harper (mentor)

Contact

in/gregwchase

gregwchase

greg@meetmindful.com