CVML Software Development Tools summary

N. Tsapanos, P. Bassia, P. Giannakeris,
Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr

Version 2.7

Artificial Intelligence &
Information Analysis Lab

Distributed computing (under development)

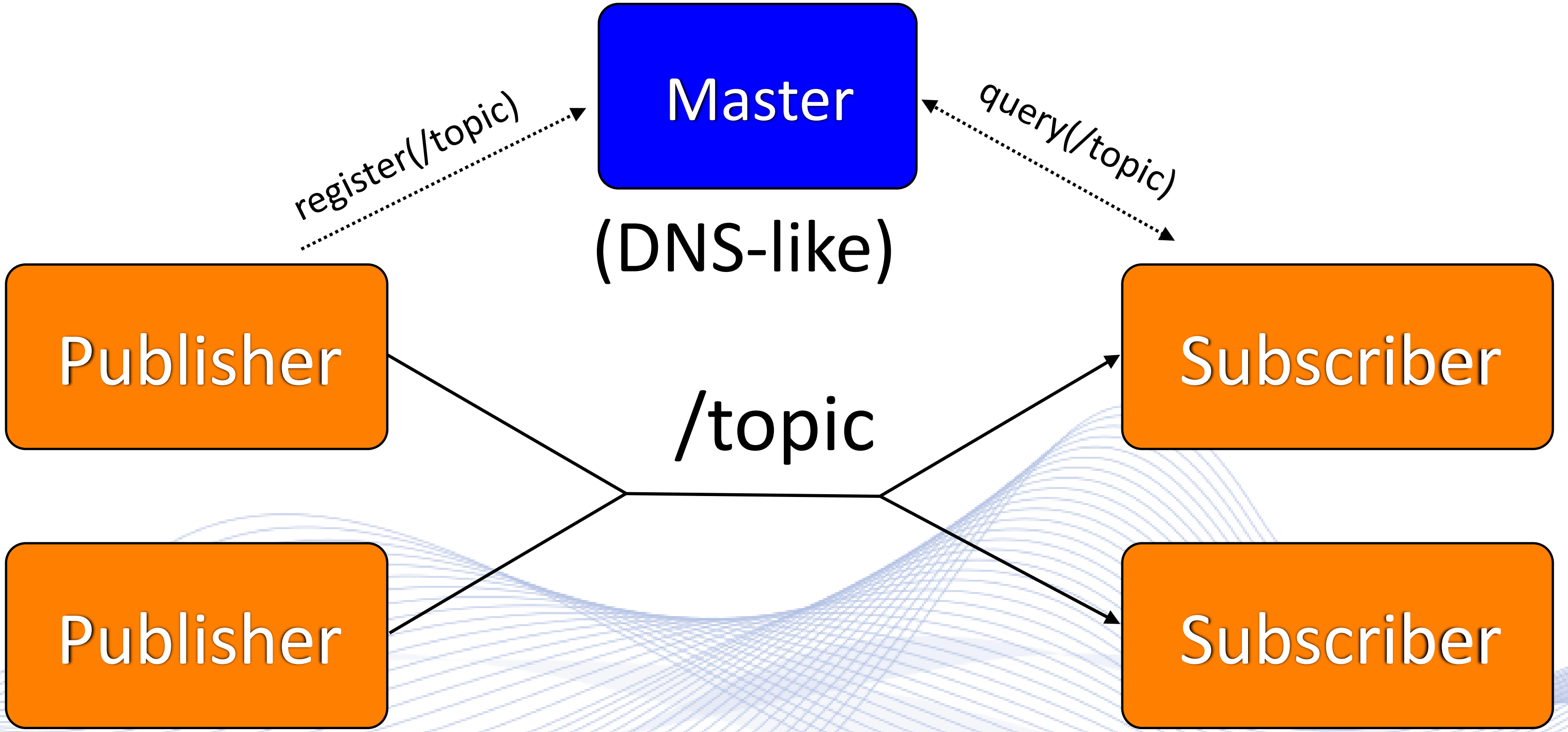
- ROS
- Libraries
  - OpenCV
  - BLAS, cuBLAS
  - MKL DNN, cuDNN
- DNN Frameworks
  - Neon, Tensorflow, Pytortch, Keras, MXNet
- Distributed/cloud computing
  - MapReduce programming model
  - Apache Spark
- Collaborative SW Development
  - GitHub/Bitbucket.

Robotic Operating System (ROS)

ROS is used in developing robotic applications.

- ROS architecture:
  - ROS master, ROS nodes.
- Node communications:
  - Topics
  - Publisher and Subscriber nodes.
- Functionalities:
  - Node graph visualization
  - Logging/plotting
  - Checks and diagnostics.

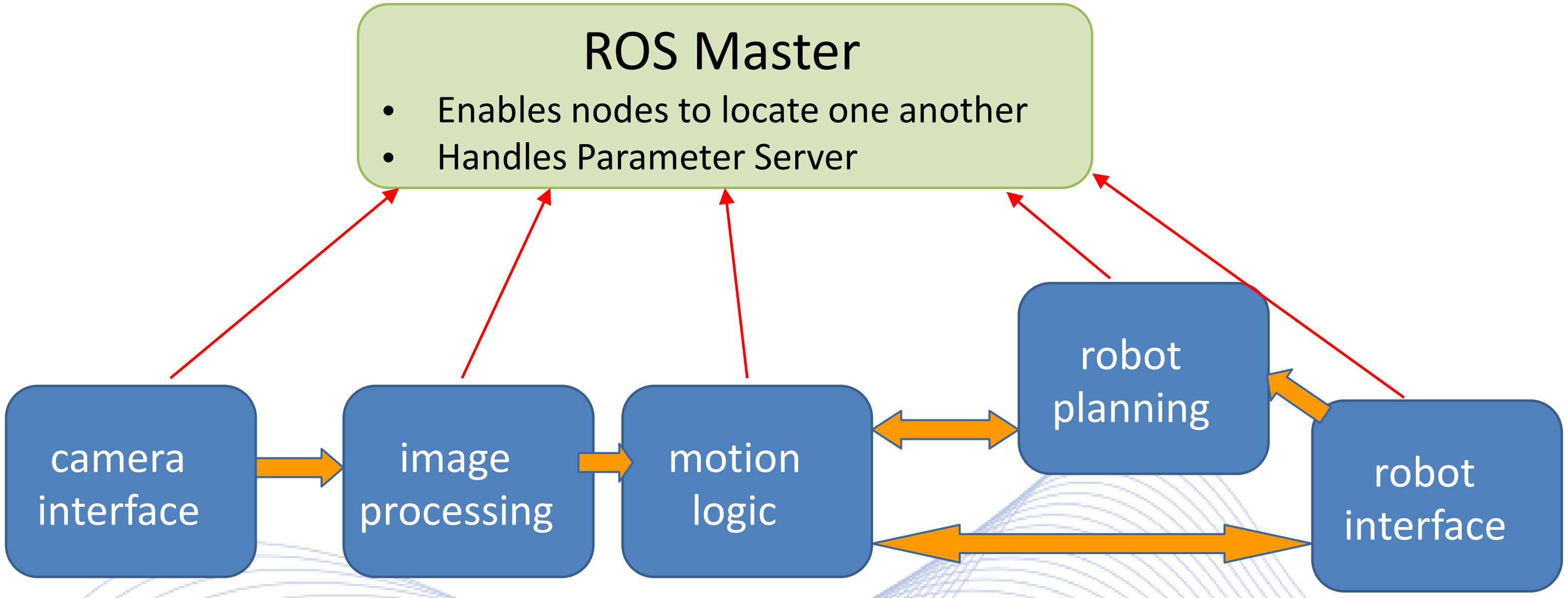
ROS architecture



(Adapted from Willow Garage's "What is ROS?" Presentation)

Artificial Intelligence &
Information Analysis Lab

ROS architecture



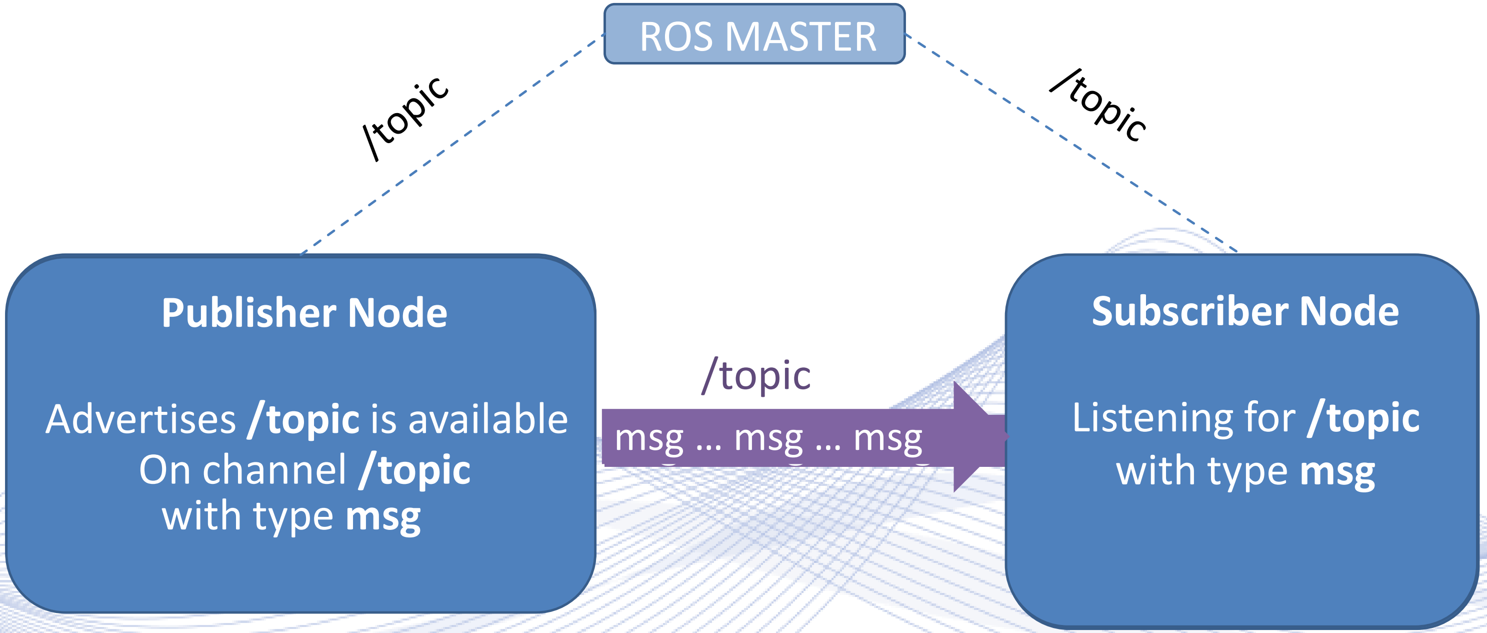
ROS Master

- Enables nodes to locate one another
- Handles Parameter Server

- A ***Node*** is a single ROS-enabled program.
- Most communication happens ***between*** nodes.
- Nodes can run on many different ***devices*.**
- One ***Master*** per system.

ROS Communications

Topics are for **Streaming Data.**



Publisher Node

Advertises **/topic** is available
On channel **/topic**
with type **msg**

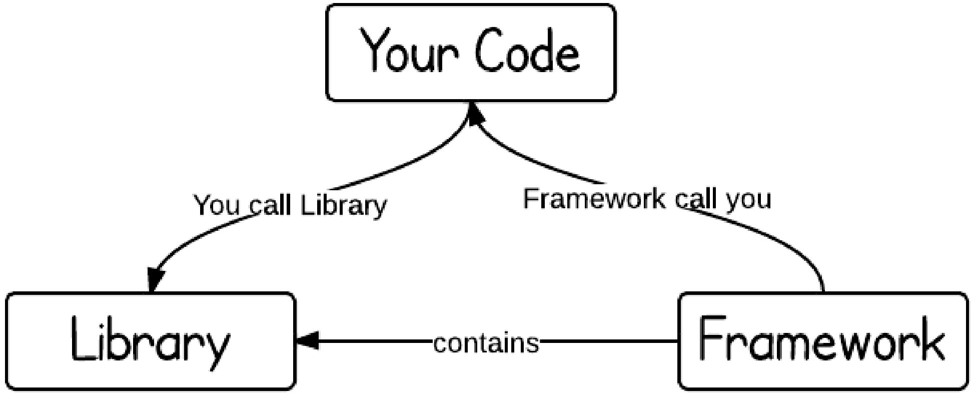
Subscriber Node

Listening for **/topic**
with type **msg**

Artificial Intelligence &
Information Analysis Lab

Libraries vs Frameworks

- When you use a library, ***you are in charge of the flow of the application***. You are choosing when and where to call the library.
- When you use a framework, ***the framework is in charge of the flow***. It provides some places for you to plug in your code, but it calls the code you plugged in as needed.



https://www.programcreek.com/

Artificial Intelligence & Information Analysis Lab

OpenCV

Open Computer vision library:

- Image processing.
- Video analysis.

- Deep NN, Machine Learning
  - Object detection.

- Computer vision
  - Camera calibration, 3D world modeling.

Basic Linear Algebra Subprograms (BLAS) Library

Basic Linear Algebra Subprograms (***BLAS***) is a software library of high performance Linear Algebra routines:

- BLAS has three routine sets (***BLAS levels***).
- They correspond to both the chronological order of definition and publication, as well as the degree of algorithm complexity.

cuBLAS

- NVIDIA cuBLAS library is a fast GPU-accelerated implementation of the standard basic linear algebra subroutines (BLAS).
- Speed up is achieved by deploying compute-intensive operations to a single GPU or scale up and ***distribute work across multi-GPU configurations efficiently***.
- Data needs to be copied to the ***GPU memory*** prior to calling cuBLAS functions. Results need to be copied backwards to the CPU memory.

KL DNN

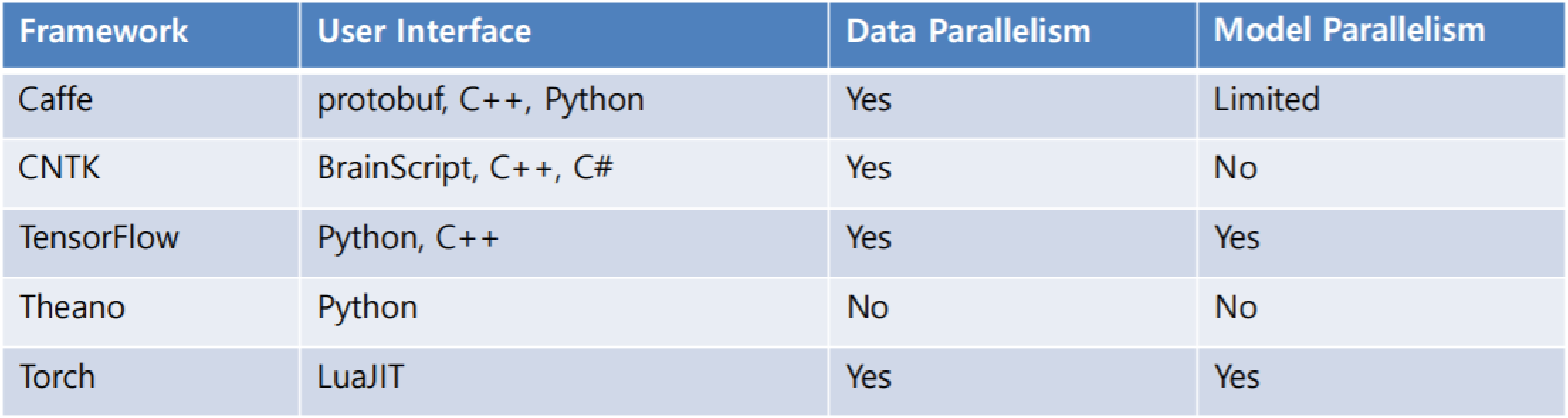
- Intel ***Math Kernel Library*** – Deep Neural Network is an ***open source, performance-enhancing library*** for accelerating deep learning frameworks on Intel Architecture.
- Optimizations are abstracted and integrated directly into PyTorch, MXNet frameworks.
- Makes use of SIMD instructions available on Intel processors
- Engine choice available (CPU/GPU).

cuDNN

- The NVIDIA CUDA® Deep Neural Network library (cuDNN) is a ***GPU-accelerated library*** of primitives for deep neural networks.
- Standard routines:
  - Forward and backward convolution.
  - Pooling.
  - Normalization layers.
  - Activation layers.
- cuDNN accelerates widely used deep learning frameworks:
  - Caffe, Keras, MATLAB, TensorFlow, Torch.

https://developer.nvidia.com/cudnn

DNN Frameworks

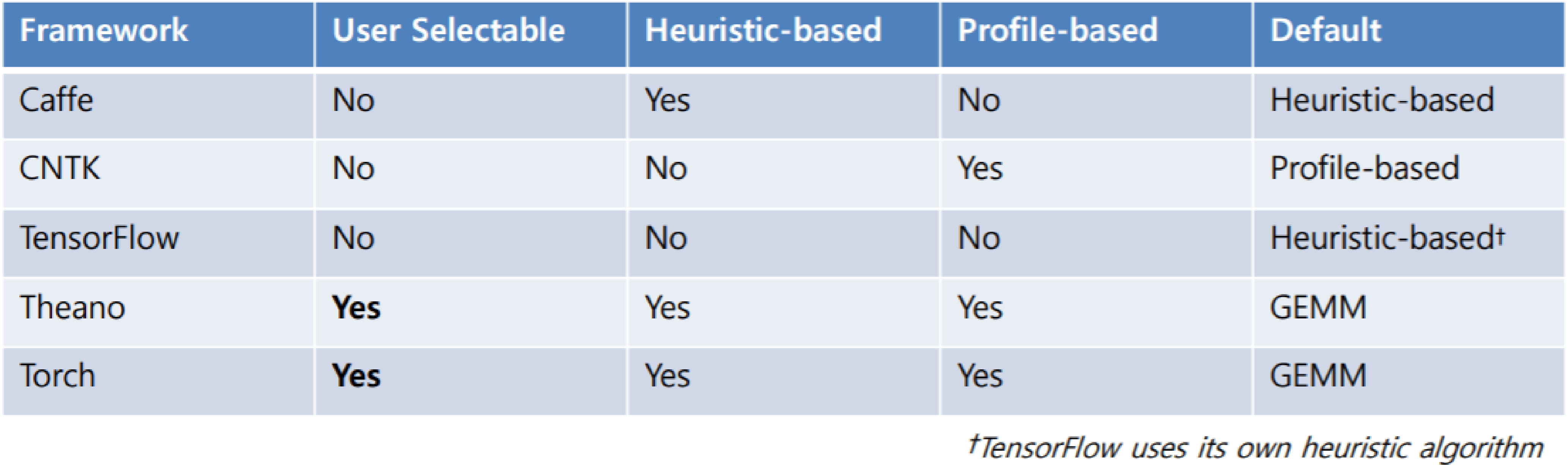


Framework	User Interface	Data Parallelism	Model Parallelism
Caffe	protobuf, C++, Python	Yes	Limited
CNTK	BrainScript, C++, C#	Yes	No
TensorFlow	Python, C++	Yes	Yes
Theano	Python	No	No
Torch	LuaJIT	Yes	Yes

Image Source: Heehoon Kim, Hyoungwook Nam, Wookeun Jung, and Jaejin Le - Performance Analysis of CNN Frameworks for GPUs

DNN Frameworks

- ***All 5 frameworks work with cuDNN as backend.***
- cuDNN unfortunately is not open source.
- cuDNN supports FFT and Winograd convolutions.



Framework	User Selectable	Heuristic-based	Profile-based	Default
Caffe	No	Yes	No	Heuristic-based
CNTK	No	No	Yes	Profile-based
TensorFlow	No	No	No	Heuristic-based†
Theano	**Yes**	Yes	Yes	GEMM
Torch	**Yes**	Yes	Yes	GEMM

†TensorFlow uses its own heuristic algorithm

Image Source: Heehoon Kim, Hyoungwook Nam, Wookeun Jung, and Jaejin Le - Performance Analysis of CNN Frameworks for GPUs

Artificial Intelligence & Information Analysis Lab

DNN Frameworks

Neon

- Intel Neon is a modern deep learning framework created by Nervana Systems.
- Implemented in Python, while Nervana Caffe framework is written in C and C++.
- ***Impressively fast compared to other frameworks.***
- Image processing oriented (not general purpose enough).

DNN Frameworks

TensorFlow **(*TF*)**

- For Python (but also APIs in C++, C#, JavaScript, Java).
- It operates with static computation graphs.
- Recommended for production (cloud, mobile).
- Supported by Google.
- Scalable.

DNN Frameworks

Pytorch

- Main competitor of TF.
- Dynamically updated graph (advantage over TF).
- ***Suited for research, small projects and prototyping.***
- Supports distributed learning.

DNN Frameworks

Keras

- High-level API for TF.
- Suitable for quick testing, prototyping.
- ***Very easy to use, good for beginners.***
- Many plug-and-play architectures available.

DNN Frameworks

MXNet

- Highly scalable.
- Very effective parallelization on multiple GPUs and machines.
- Supports many languages (C ++, Python, R, Julia, JavaScript, Scala, Go, Perl).
- Supported by Apache.

Distributed/Cloud computing

- An alternative to expensive Graphics cards and supercomputers.
- ***It can consist of any combination of PCs of various processing power.***
- Extremely flexible and scalable.
- Designed to handle Big Data.

Edge/Fog Computing

- Autonomous systems contain massive data sources:
  - Video, 3D point clouds.
- Data cannot be processing locally, due to energy and computing power constraints.
- They should not be streamed deep for cloud computing.
- ***Edge computing*** is very important, as it resides at the communication edge.

MapReduce programming model

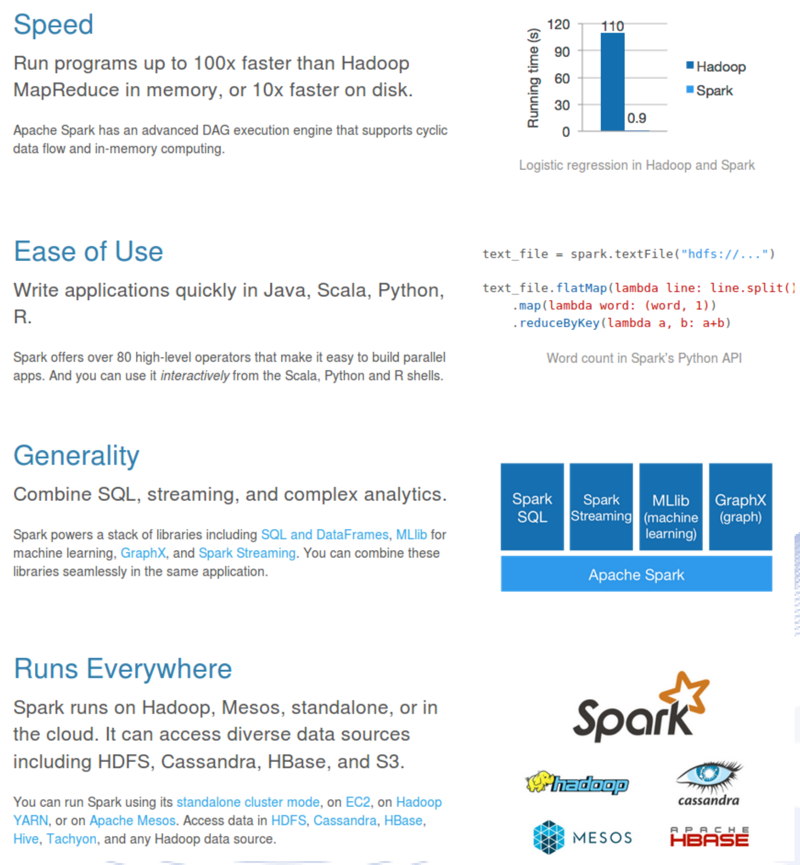
- This programming model was developed to implement the Map and Reduce commands, which can, in turn, be used to write distributed programs with ease.

- The ***Map command applies a function on every element*** of the distributed dataset. This can be used to filter, sort, or transform the data.

- The ***Reduce command collects the distributed dataset***, or the results of a previous Map or Reduce command, in a summary operation, such as addition.

Apache Spark

- Open software for cluster computing.
- Developed by Apache.
- ***Distributed big data computing.***
- It is based on MapReduce.
- Up to 100 times faster than Hadoop MapReduce.
- Machine Learning, Graph Processing, SQL, Streaming modules.



Speed

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Apache Spark has an advanced DAG execution engine that supports cyclic data flow and in-memory computing.

Logistic regression in Hadoop and Spark

Ease of Use

Write applications quickly in Java, Scala, Python, R.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala, Python and R shells.

```
text_file = spark.textFile("hdfs://...")

text_file.flatMap(lambda line: line.split()
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

Word count in Spark's Python API

Generality

Combine SQL, streaming, and complex analytics.

Spark powers a stack of libraries including SQL and DataFrames, MLlib for machine learning, GraphX, and Spark Streaming. You can combine these libraries seamlessly in the same application.

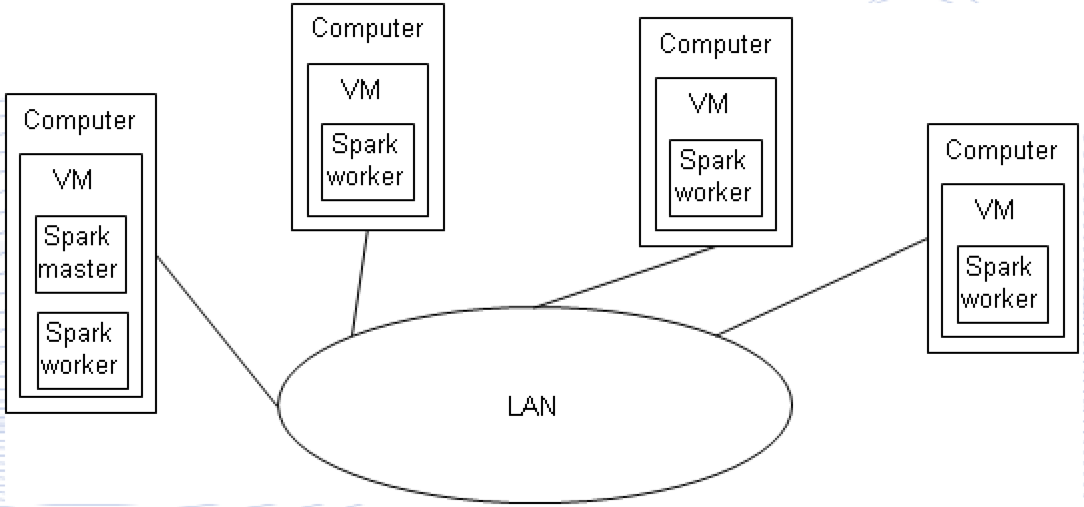
Runs Everywhere

Spark runs on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3.

You can run Spark using its standalone cluster mode, on EC2, on Hadoop YARN, or on Apache Mesos. Access data in HDFS, Cassandra, HBase, Hive, Tachyon, and any Hadoop data source.

Apache Spark

- In Spark, a ***master node*** coordinates several worker nodes.
- We used VirtualBox VMs running Ubuntu to set up our computing cluster.



Collaborative SW Development

- Several alternatives exist.

- Most popular front-end services:
  - GitHub.
  - Bitbucket.

- Most popular underlying open-source version control systems:
  - Git.
  - Mercurial.

Q & A

Thank you very much for your attention!

Contact: Prof. I. Pitas
pitas@csd.auth.gr