# Machine Learning Models Deployment

## Hardware or Software ?

Antoine CARME (2022)
Antoine.CARME@outlook.com

# Software vs Hardware Deployment ?

- Software is easy to maintain, update, correct…
  - Only when released as Open Source, by experience.
- Specialized Hardware is not reusable
- The following slides will (try to) give an overview of the past and present systems for deploying analytic/data processes
- Some hardware systems are listed here only to keep track of the “Hall of Fame of bad Ideas in Hardware Analytics” !!!
- Some ~~Lobbying~~ arguments for “Software Supremacy”:
  - https://github.com/antoinecarme/sklearn2sql-demo/blob/master/notebooks/sql_rationale.md

# Software Deployment Systems 1/4

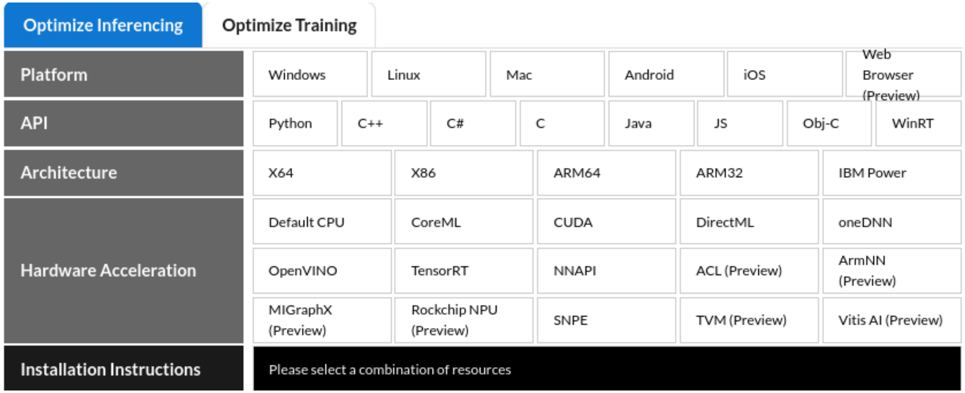
- Predictive Model Markup Language (PMML)
  - https://en.wikipedia.org/wiki/Predictive_Model_Markup_Language
  - XML Schema developed by the Data Mining Group, A consortium of proprietary data mining software companies (SAS, SPSS, ...)
  - Deployment is made through specialized vendor-independent software (PMML runtime).

# Software Deployment Systems 2/4

- Java PMML API
  - Open Source PMML software.
    - https://github.com/jpmml
    - Developed by https://openscoring.io
    - Very Actively Developed.
    - Works with R and Scikit-Learn Models

# Software Deployment Systems 3/4

- Open Neural Network Exchange (ONNX)
  - https://onnx.ai/index.html
  - The open standard for machine learning interoperability
  - Allows building and Deploying models.
  - Supports many ML/DL frameworks (Scikit-Learn TensorFlow, PyTorch, Caffe2, …)
  - Needs a runtime for deploying models
    - One runtime for each target environmet/(programming language).
    - Not all environments are equal ...
  - Actively Developed.
    - https://github.com/onnx/sklearn-onnx



Optimize Inferencing | Optimize Training

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Platform | Windows | Linux | Mac | Android | iOS | Web Browser (Preview) | | |
| API | Python | C++ | C# | C | Java | JS | Obj-C | WinRT |
| Architecture | X64 | X86 | ARM64 | ARM32 | IBM Power | | | |
| Hardware Acceleration | Default CPU | CoreML | CUDA | DirectML | oneDNN | | | |
| | OpenVINO | TensorRT | NNAPI | ACL (Preview) | ArmNN (Preview) | | | |
| | MIGraphX (Preview) | Rockchip NPU (Preview) | SNPE | TVM (Preview) | Vitis AI (Preview) | | | |
| Installation Instructions | Please select a combination of resources | | | | | | | |

# Software Deployment Systems 4/4

- Vendor Specific Systems
  - The software used to train ML models can be used to deploy these models.
  - SAS and SPSS have some kind of in-Database Scoring (SQL-based)
    - https://github.com/antoinecarme/sklearn2sql-demo/blob/master/notebooks/limitations.md
    - Often limited in the supported models and databases.
  - TFLite : Deploy TF models on mobile and edge devices
    - Google : https://www.tensorflow.org/lite?hl=fr
    - TensorFlow Lite for Micro-controllers currently supports a limited subset of TensorFlow operations
  - PyTorch Mobile
    - Allow building apps to deploy PyTorch models on iOS and Android devices.
    - https://pytorch.org/mobile/home/

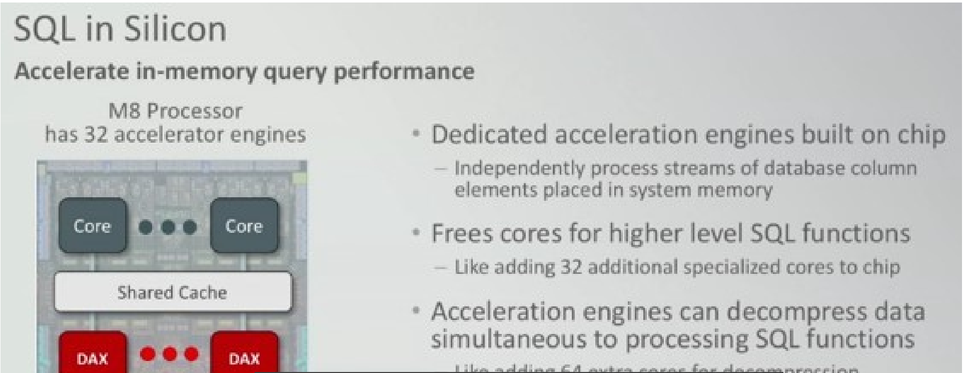
# Hardware Deployment Systems 1/5

- History
  - The use of hardware systems to deploy Machine Learning systems is a very old idea.
  - Experimental Fax + OCR + Speech.
    - Apple. Mimetics (199x)
    - https://techmonitor.ai/technology/mimetics_shows_fax_into_voice_product_at_comdex_98
    - https://www.manualsdir.com/manuals/548581/apple-fax.html?page=101
  - Defense systems.
  - Many categories:
    - Database Accelerators / Data Caching
    - Deep Learning GPUs / TPUs / ASICs / FPGAs
    - New trends. NPUs

# Hardware Deployment Systems 2/5

- Database Appliances
  - Netezza, etc …
- Database Analytics Accelerators
  - Oracle DAX, T7 and M8 Sparc CPUs
  - ZD-XL SQL Server Accelerator
- Database + GPU/FPGA/…
  - Kinetica, swarm64, …

kinetica swarm64



## SQL in Silicon

**Accelerate in-memory query performance**

M8 Processor has 32 accelerator engines

- Dedicated acceleration engines built on chip
  - Independently process streams of database column elements placed in system memory
- Frees cores for higher level SQL functions
  - Like adding 32 additional specialized cores to chip
- Acceleration engines can decompress data simultaneous to processing SQL functions
  - Like adding 64 extra cores for decompression

tion

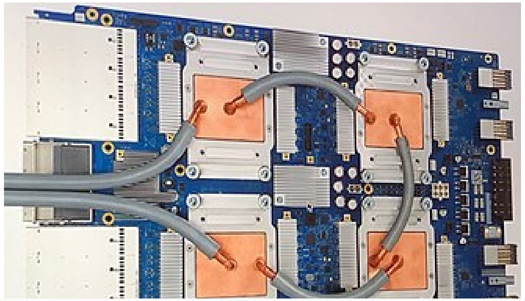
e-SP

# Hardware Deployment Systems 3/5

- The Use of GPUs
  - The use of GPUs in Deep Learning training and deployment follows years of use of GPUs as a tool for accelerating video graphics and scientific computing.
    - https://en.wikipedia.org/wiki/General-purpose_computing_on_graphics_processing_units
  - The main DL frameworks (Theano, TF, PyTorch) allow using GPUs to speed up computations otherwise using system CPUs.
  - Nvidia is the main hardware manufacturer in this area.
    - Nvidia provides embedded systems (Jetson) dedicated to ML/DL.
      - https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/
  - The main Cloud providers (AWS, IBM, …) have specialized instances with GPUs

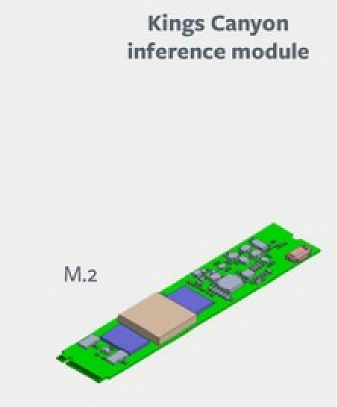
# Hardware Deployment Systems 4/5

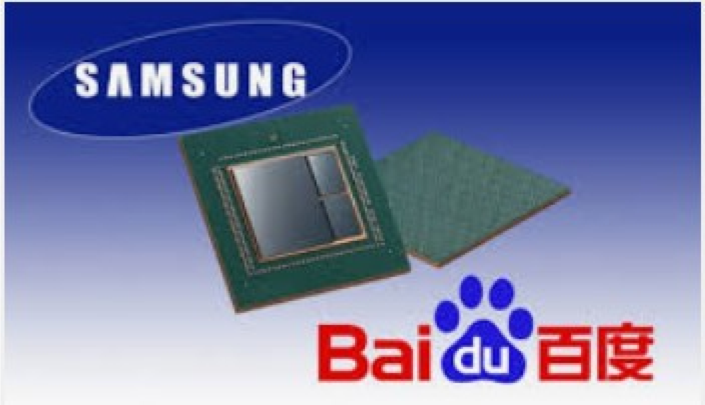
- The Use of Specialized Hardware (ASICs)
  - A current trend is to build  application-specific integrated circuit (ASIC) for DL data (tensor-processing units)
    - https://en.wikipedia.org/wiki/Tensor_Processing_Unit
  - Google Uses its own TPUs instead of GPUs for training and deploying DL/TF Models
    - https://cloud.google.com/tpu
  - Google Edge TPUs are available for microcontrollers.
    - https://coral.ai/products/
    - Uses TFLite. Limited.



# Hardware Deployment Systems 5/5

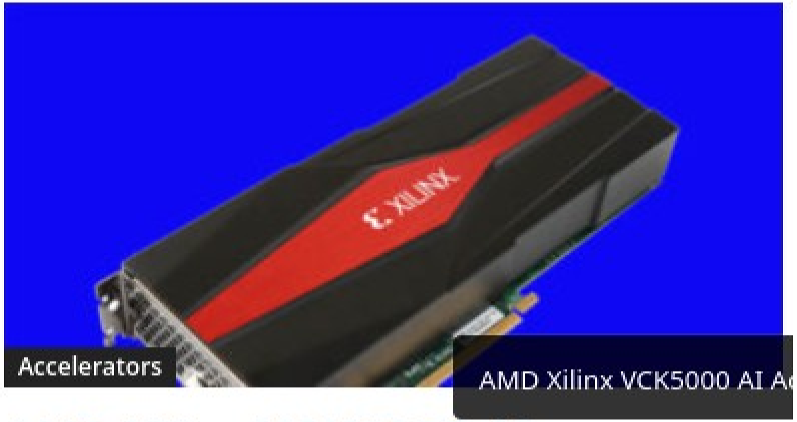
- Deep learning hardware accelerators (+NPUs)
  - GAFAM and BATX race to AI chips.
  - https://syncedreview.com/2019/03/14/facebook-releases-a-trio-of-new-ai-hardware-designs/





- https://en.wikipedia.org/wiki/AI_accelerator

Intel Habana Greco AI Inference PCIe Card at Vision 2022



AMD Xilinx VCK5000 AI Accelerator Launched

# 谢谢 !!!!