Deep Learning in Industry

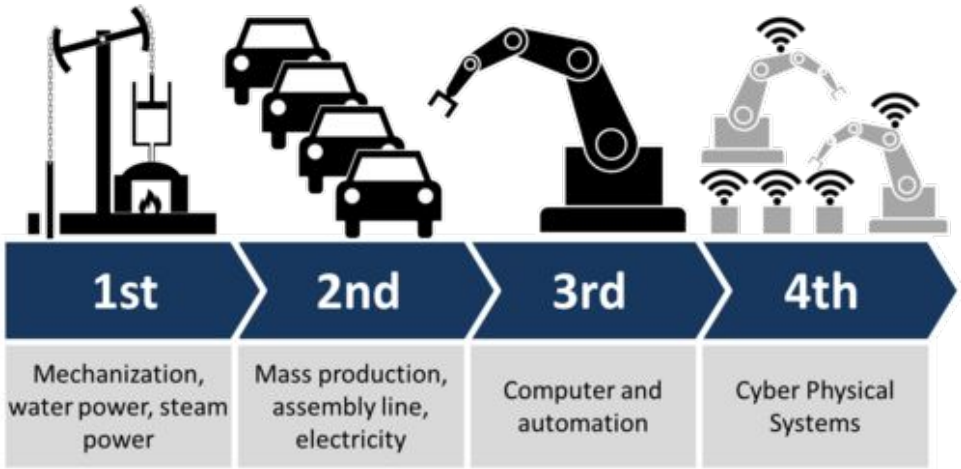
Shenjun Zhong, PhD

Data Scientist, Telstra Big Data
Research Officer, Monash University

19 March, 2018

NO MATH :)

Industry is Embracing Deep Learning



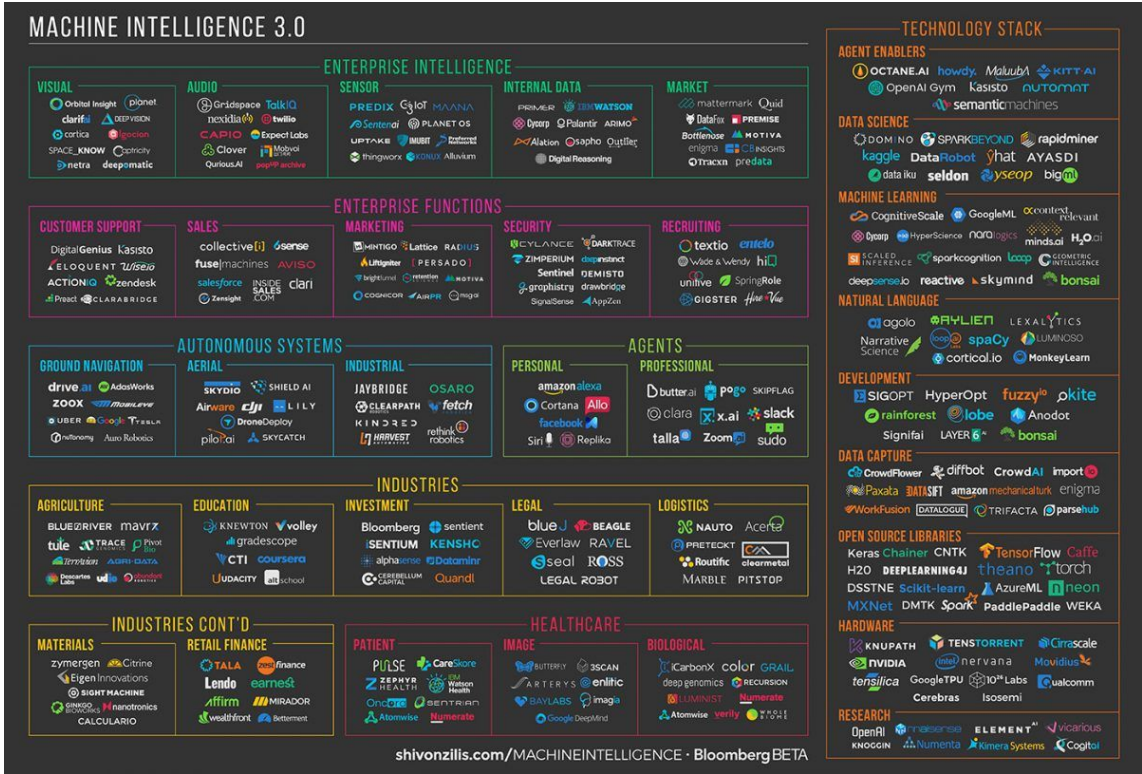
"the G20 countries should support coordinated research, development and deployment activities on AI for the ***fourth industrial revolution***."

- G20 Conference 2016

The 4th Industrial Revolution ≈ AI ?

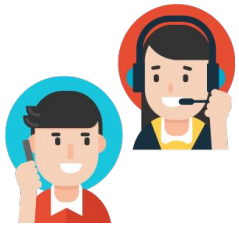
Industry is Embracing Deep Learning

- Healthcare
- Transportation
- Personalized AI Assistants
- Robotics
- Speech/Video Recognition and Generation
- Security
- etc.

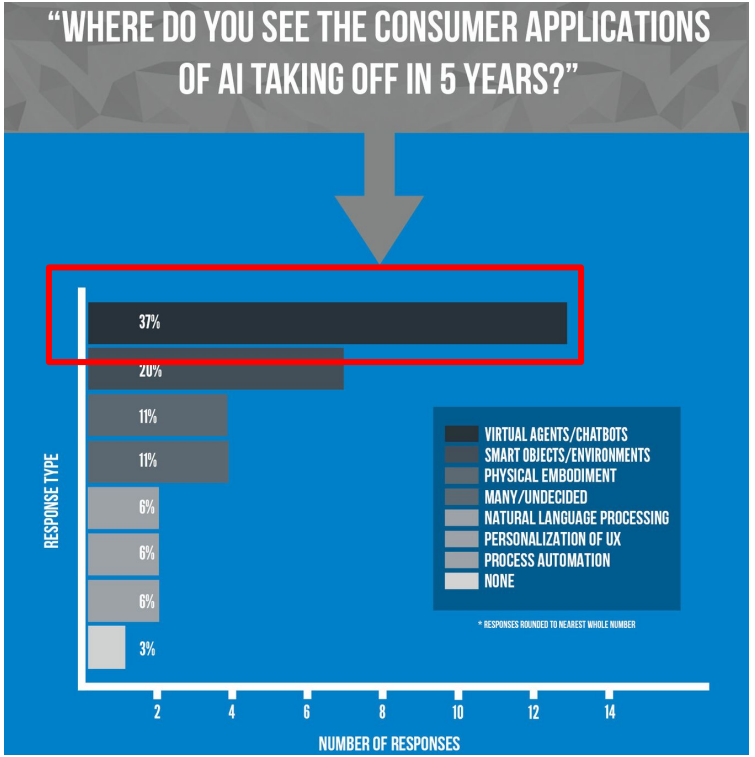


So Are We - Build a “CHATBOT”

Customer Experience



Customer Service



Source: www.techemergence.com

Deep Learning Day & Night

Model Training

- Data Preparation
- Model building
- Optimization
- Fine Tuning
- etc.

Model Serving

- Model Archiving
- Model Version Control
- Performance
- Scalability
- etc.

Deep Learning in ~~Industry~~ *Production*

Optimization in Serving DL Model

Outline

- Goal
- How
  - Options
- A Small Real-life Case Study
- Future works / Potential Collaboration

Goal

To make a serving system

- High throughput
- Scalable
- Fault-tolerant
- Easy managing
- CICD

HOW?

- Option 1
  - Third party SaaS
- Option 2
  - Cloud Serving Platform
- Option 3
  - DIY

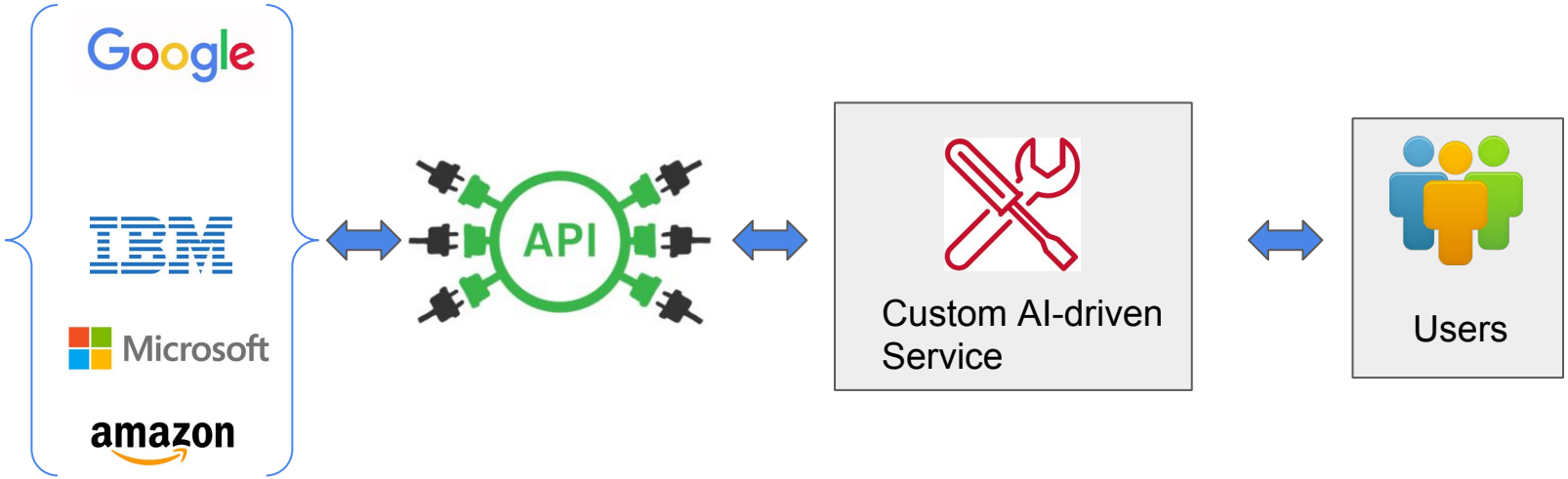
Option 1: Let the “Expert” do the hard work

Machine Learning as a Service (MLaaS)

- Google
  - Google APIs
    - Vision, Video, Speech, Translation, NLP
  - Cloud AutoML (Alpha)
    - Hyperparameters optimization and modern optimizers
- IBM Watson
  - Conversation, NLP, Visual Recognition, Translator, Speech2Text, Text2Speech, Tone Analyzer, etc.
- Microsoft
  - Cognitive-services: Vision, Speech, Language
- Amazon
  - Amazon Rekognition: Video, Image
  - Amazon Lex: Conversation - Alexa
  - Amazon Language

- Out-of-the-box
- Robust
- Auto Scaling

Option 1: MLaaS



Need Customization?

- Train your model
- Serve it somewhere

Option 2: Cloud Model Serving Service

- Google
  - Cloud ML Engine
    - TPU Service(beta)

- AWS
  - Model Server for Apache MXnet
  - ONNX - Open Neural Network Exchange
    - Caffe2, Torch/PyTorch, CNTK, MXnet and Chainer

Need More Control and Flexibility?

- Do everything yourself !

Option 3: Do it yourself - A Case Study

Trained Model

Serving Stack

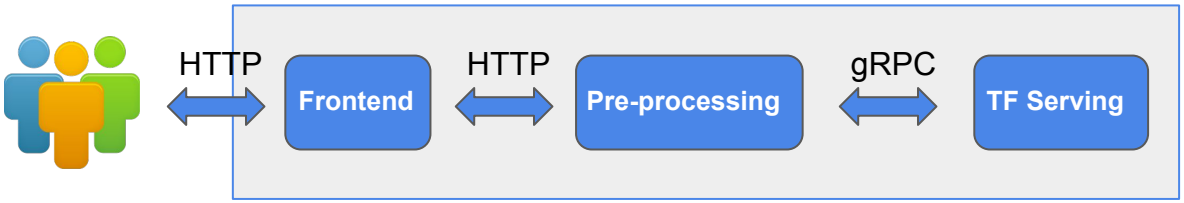
TensorFlow + docker

kubernetes

amazon web services™

- Caffe2
- Spark
- TensorRT
- Apache's PredictionIO
- Redis Labs's Redis-ML
- Deeplearning4j's model server

Evolution of the Serving Architecture



- REST API
- Authentication

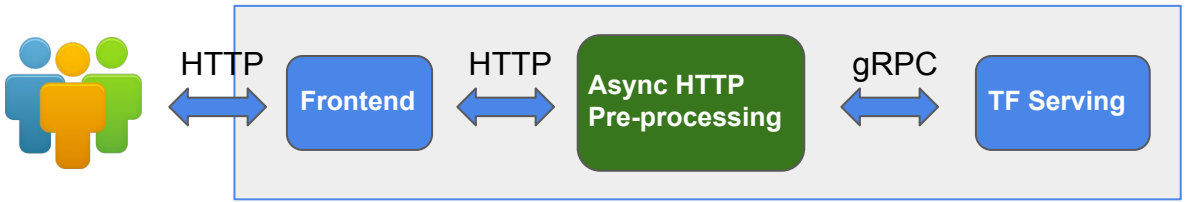
- Tokenize
- Embedding
- etc.

- Inference

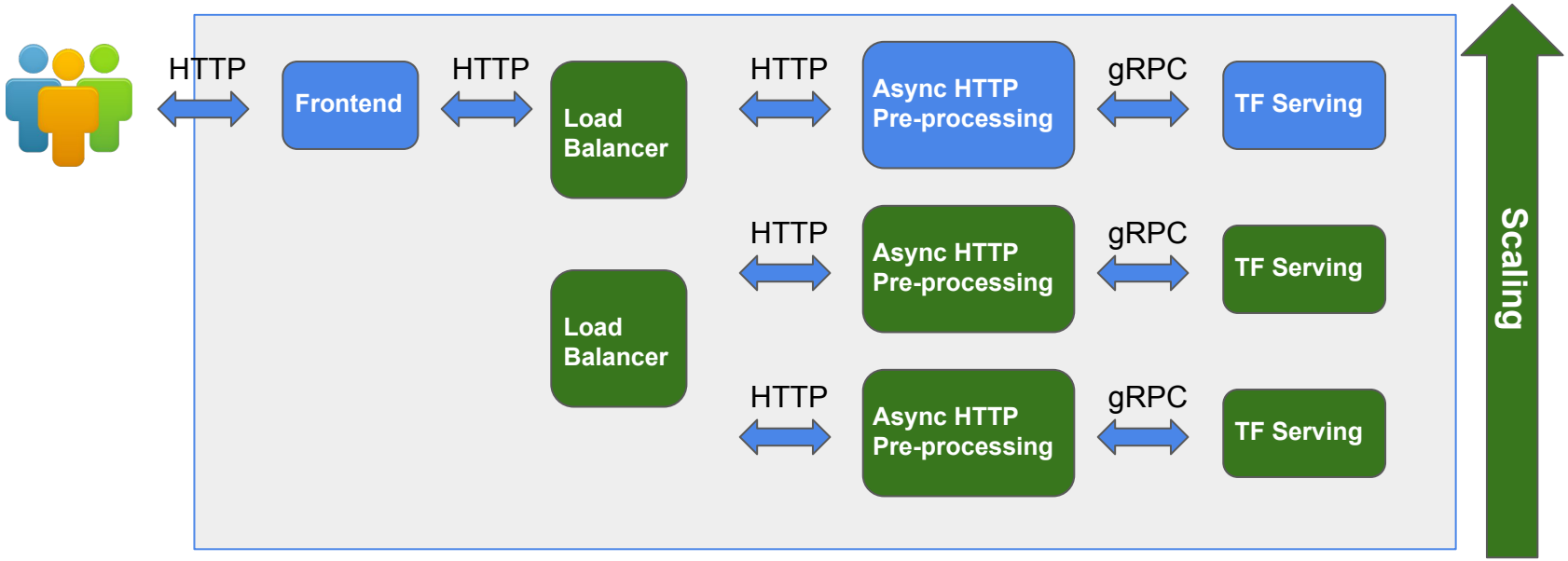
Baseline

- Blocking IO
- Single Request may take up to several seconds

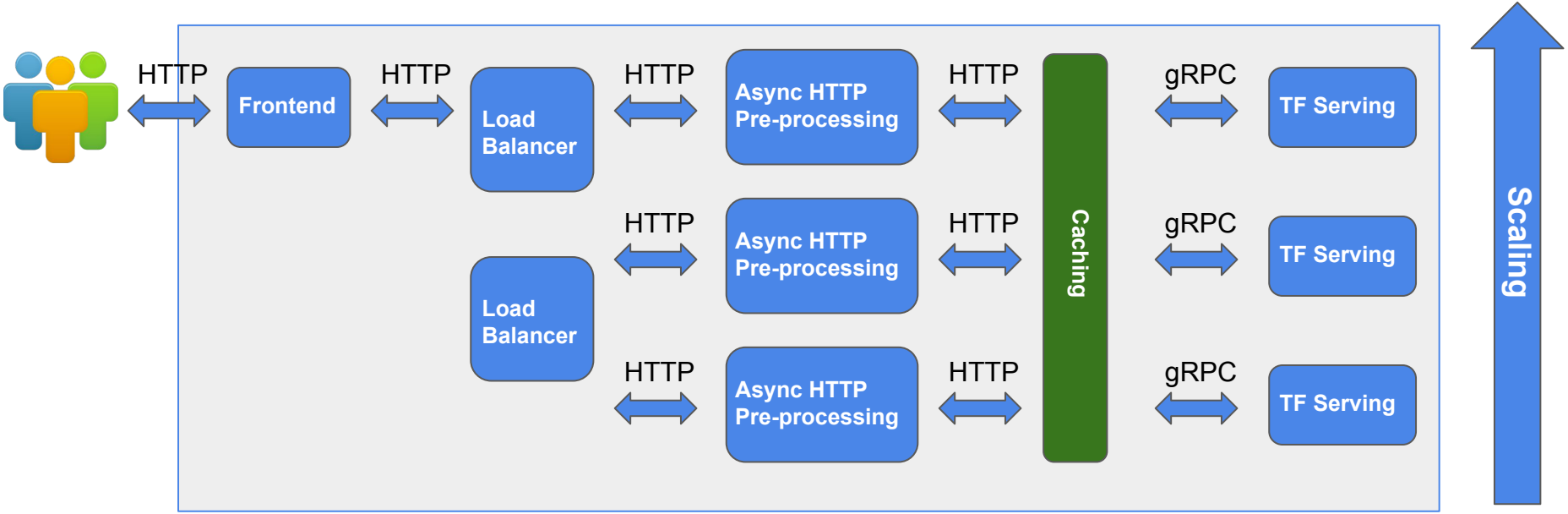
Evolution of the Serving Architecture



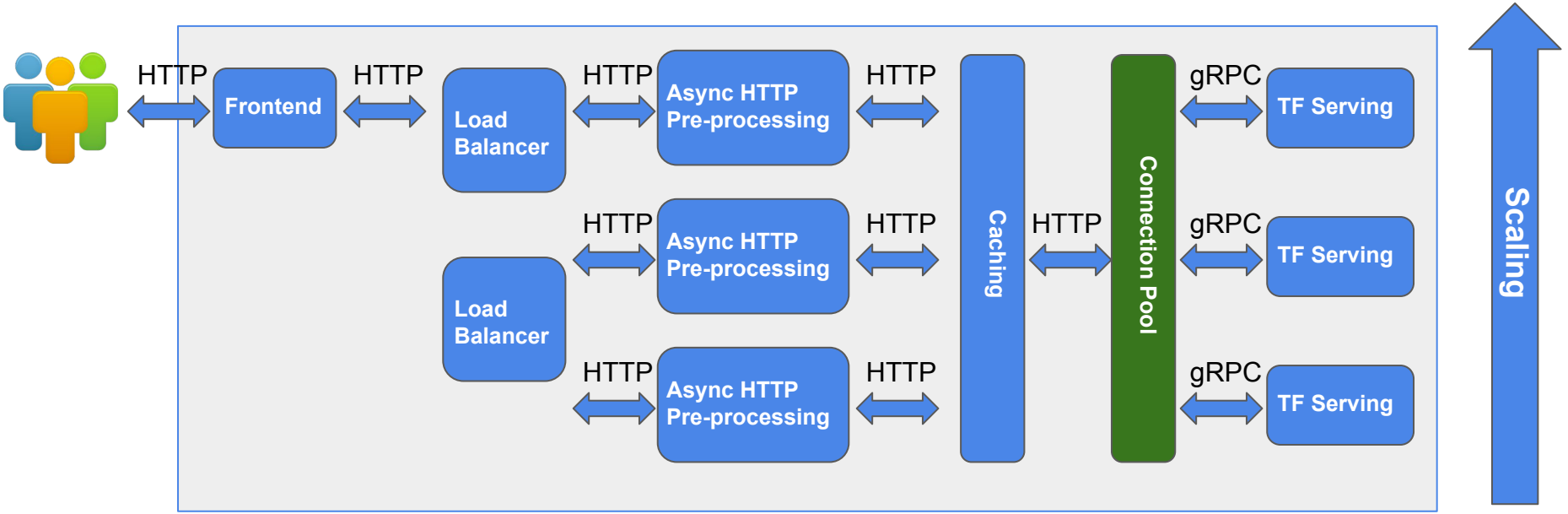
Evolution of the Serving Architecture



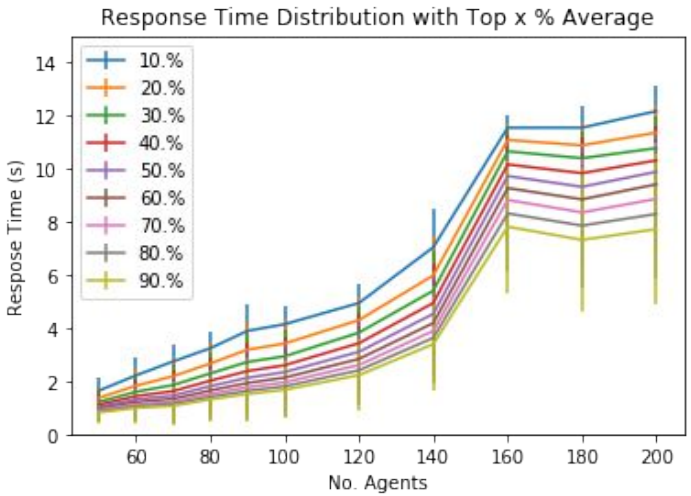
Evolution of the Serving Architecture

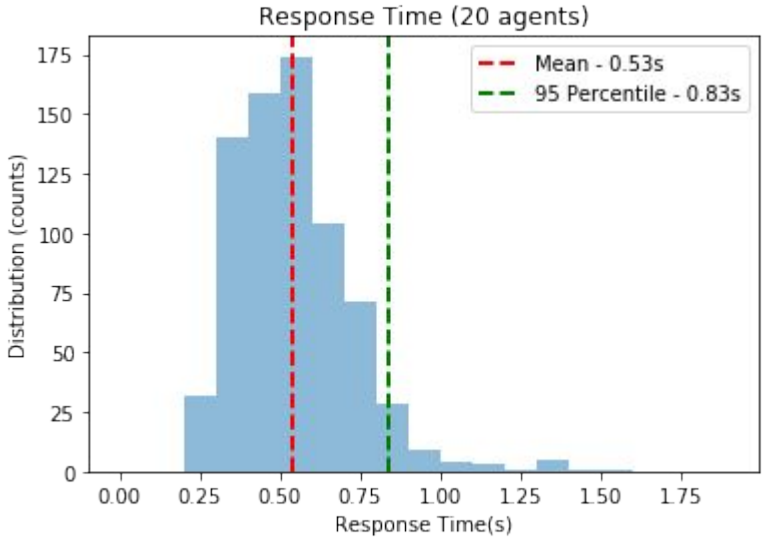


Evolution of the Serving Architecture



Performance Metrics

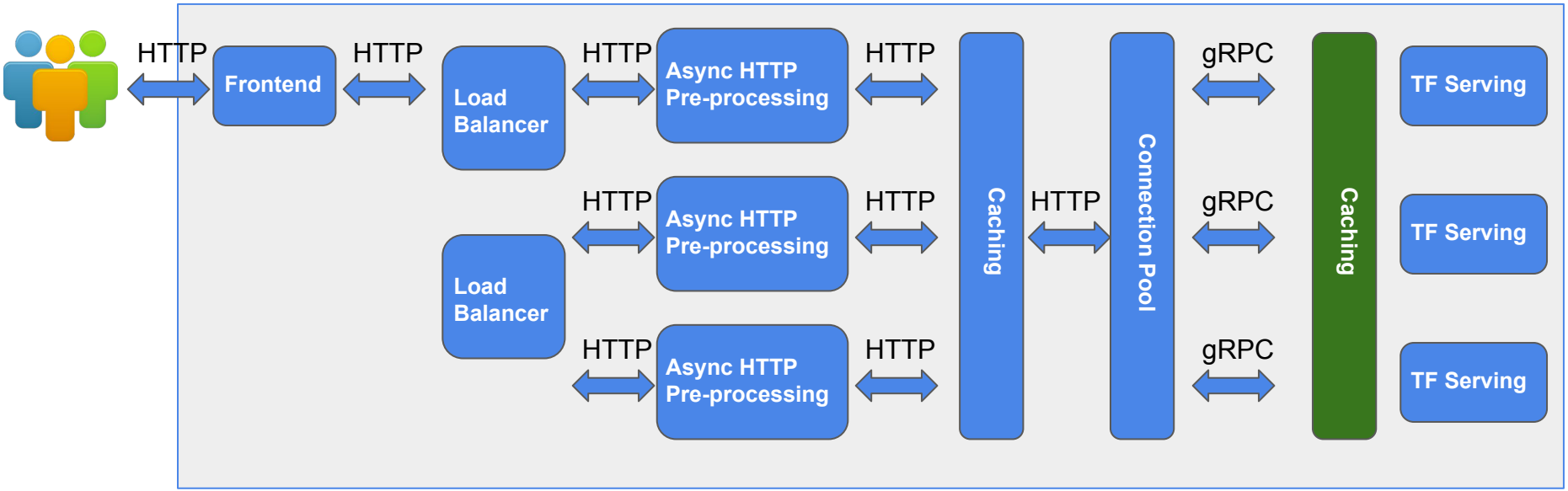




- **12** AWS X.large Instances
- 2 x 2.35 CPU cores
- 8G Ram

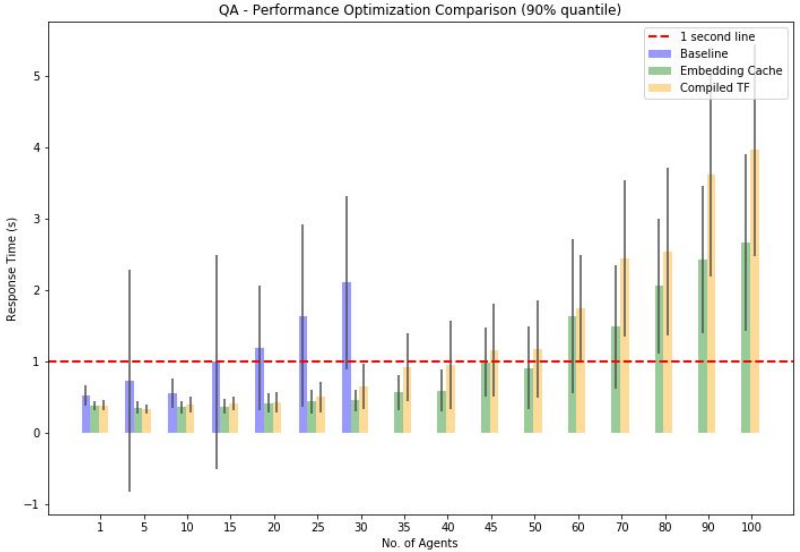
- Sub-second latency
- Each AWS instance can handle **3** users concurrently

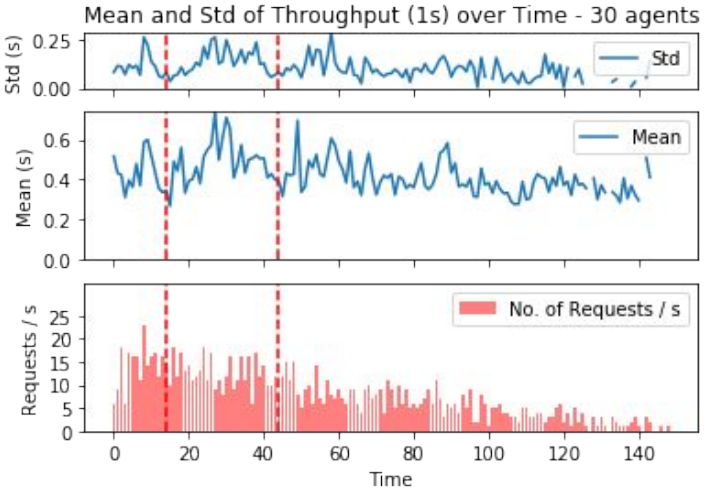
Evolution of the Serving Architecture

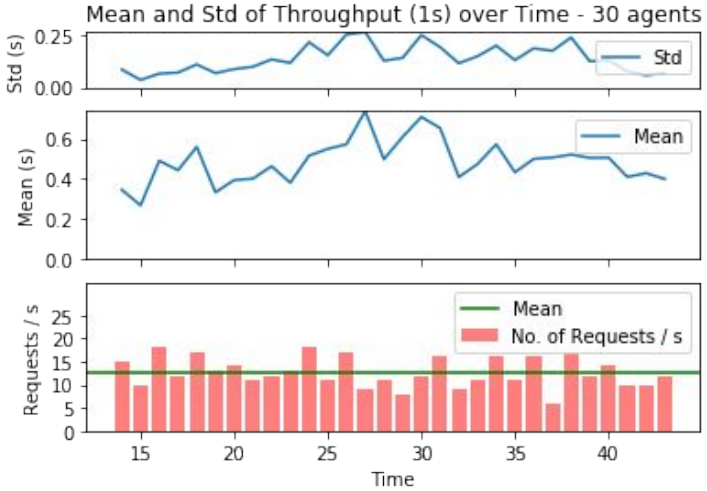


Performance Metrics

- **1** AWS Instance
- Can serve up to 40 users with <1 second latency
- > **13x** speedup
- Huge budget saving

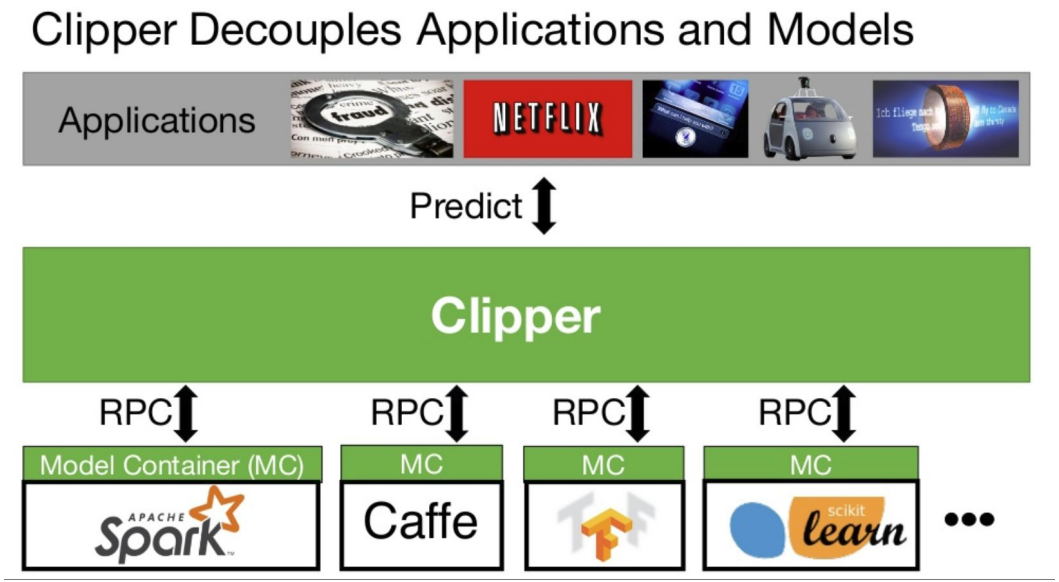






Service Orchestration - Clipper

- Clipper [1]
  - A Low-Latency Online Prediction Serving System
  - Made by RISE Lab in UC Berkeley
  - A middleware managing serving containers
  - Compatible to most of the major frameworks
  - Very early stage (version 0.2)

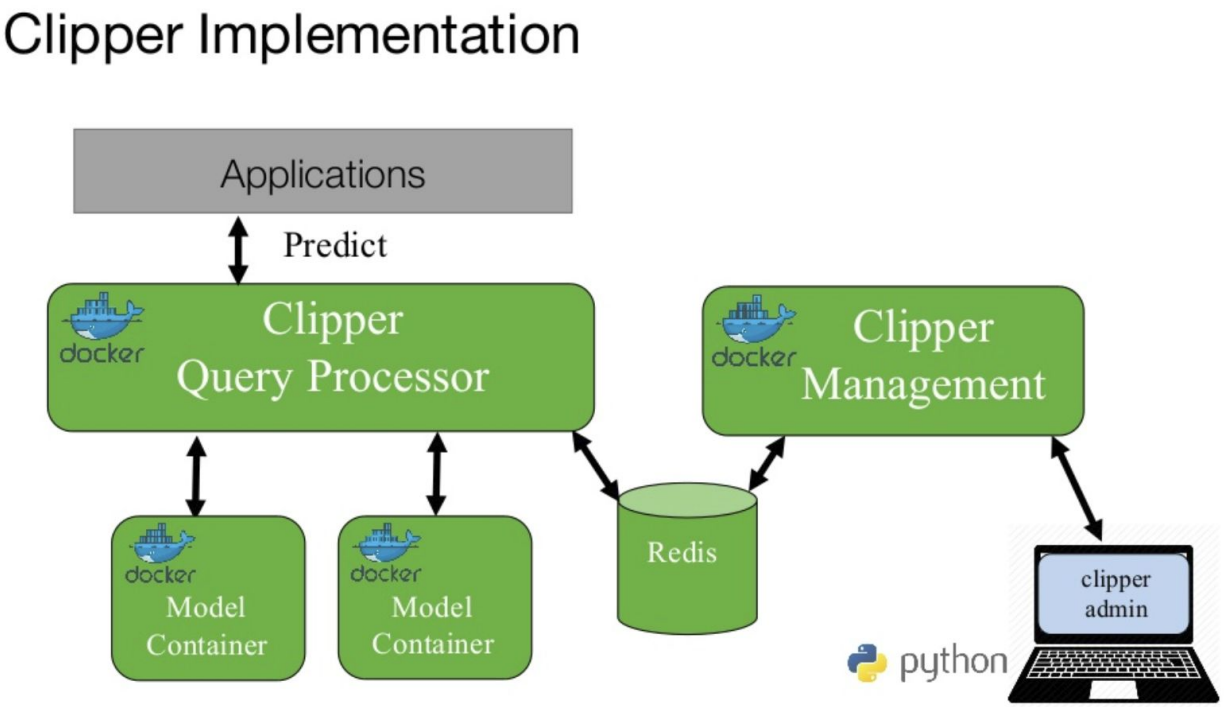


Source: http://clipper.ai/

[1] Crankshaw, D., Wang, X., Zhou, G., Franklin, M. J., Gonzalez, J. E., & Stoica, I. (2017, March). Clipper: A Low-Latency Online Prediction Serving System. In NSDI (pp. 613-627).

Service Orchestration - Clipper

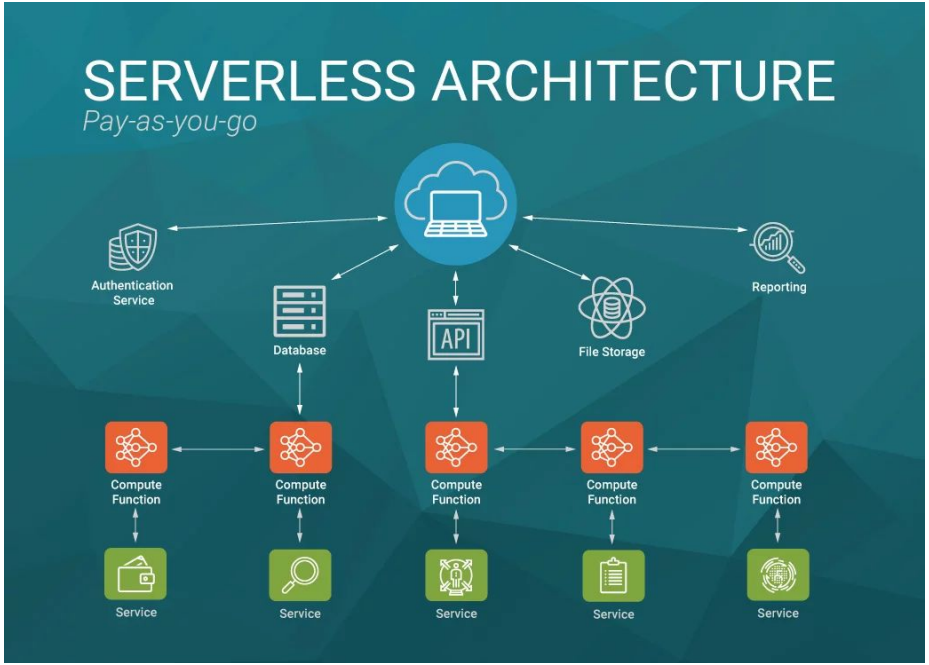
- An unified solution
- Auto Scaling
- Caching
- Adaptive batching
- Cross-framework batching



Source: http://clipper.ai/

Serverless Architecture for Model Serving

- Serverless Architecture
  - Stateless container
  - Outsource the management
- Providers
  - AWS Lambda
  - Google Functions (beta)
- How
  - Package model + weights into a docker image
  - Deploy onto Serverless Service
  - Call the service with the API
- Preliminary results not shown



Other Optimizations

- Batching on TF serving
- Source Compilation
- Model Compression (2015)
  - Quantization
  - Model Pruning [1]
    - LSTM Pruning [2]

[1] Han, S., Mao, H., & Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149.
[2] Tang, S., & Han, J. A pruning based method to learn both weights and connections for LSTM.

Progress and Future works

- Model Pruning / Compression
- Serverless Architecture
- DL Model for auto scaling
- TVM : An End-to-End Optimization Stack - Published in Feb,2018 [1]
- Model Serving on BlockChain

[1] Chen, T., Moreau, T., Jiang, Z., Shen, H., Yan, E., Wang, L., ... & Krishnamurthy, A. (2018). TVM: End-to-End Optimization Stack for Deep Learning. arXiv preprint arXiv:1802.04799.

Summary

Model serving is a Challenge / Nightmare

- Lack of good model serving platform / service
- Industrial Standard
  - CICD
  - Security check
  - *System robustness*
  - *Scalability*
- *One of the next big things in AI industries*