

Deep Learning on AWS

Julien Simon
Principal Technical Evangelist
Amazon Web Services

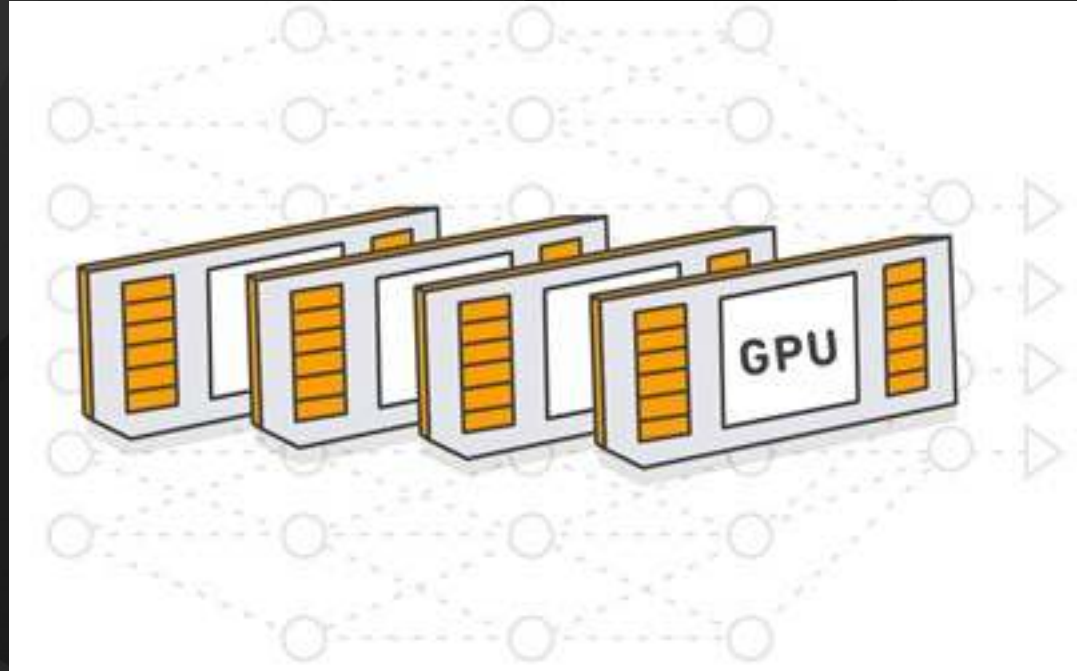
julsimon@amazon.fr
@julsimon

09/11/2016



Agenda

- GPU instances
- Deep Learning AMI
- Personalization @ Amazon.com
- Amazon DSSTNE
- Demo !



AWS GPU Instances

- **G2** (2xlarge, 8xlarge)
 - 32 vCPUs, 60 GB RAM
 - 4 NVIDIA K520 GPUs
 - 16 GB of GPU memory, 6144 cores
- **P2** (xlarge, 8xlarge, 16xlarge)
 - Launched in 09/16
 - 64 vCPUs, 732 GB RAM
 - 16 NVIDIA GK210 GPUs
 - 192 GB of GPU memory, 39936 cores
 - 20 Gbit/s networking

EC2 Instance Type ⓘ	Total
g2.2xlarge	\$0.65/hr
g2.8xlarge	\$2.60/hr
p2.8xlarge	\$7.20/hr
p2.xlarge	\$0.90/hr
p2.16xlarge	\$14.40/hr

<https://aws.amazon.com/blogs/aws/new-g2-instance-type-with-4x-more-gpu-power/>
<https://aws.amazon.com/blogs/aws/new-p2-instance-type-for-amazon-ec2-up-to-16-gpus/>



AWS Deep Learning AMI

- **Deep Learning Frameworks** – 5 popular Deep Learning Frameworks (MXNet, Caffe, Tensorflow, Theano, and Torch) all prebuilt and pre-installed
- **Pre-installed components** – Nvidia drivers, cuDNN, Anaconda, Python2 and Python3
- **AWS Integration** – Packages and configurations that provide tight integration with Amazon Web Services like Amazon EFS (Elastic File System)

Personalization @ Amazon.com

Generating Recommendations at Amazon Scale with Apache Spark and Amazon DSSTNE

by Kiuk Chung | on 09 JUL 2016 | [Permalink](#) | [Comments](#)

Kiuk Chung is a Software Development Engineer with the Amazon Personalization team

In [Personalization](#) at Amazon, we use neural networks to generate personalized product recommendations for our customers. Amazon's product catalog is huge compared to the number of products that a customer has purchased, making our datasets extremely sparse. And with hundreds of millions of customers and products, our neural network models often have to be distributed across multiple GPUs to meet space and time constraints.

For this reason, we have created and open-sourced [DSSTNE](#), the Deep Scalable Sparse Tensor Neural Engine, which runs entirely on the GPU. We use DSSTNE to train neural networks and generate recommendations that power various personalized experiences on the retail website and Amazon devices.

<https://aws.amazon.com/blogs/big-data/generating-recommendations-at-amazon-scale-with-apache-spark-and-amazon-dsstne/>

Amazon DSSTNE (aka 'Destiny')

- Deep Scalable Sparse Tensor Network Engine
- Open source library for deep neural networks using GPUs
<https://github.com/amznlabs/amazon-dsstne>
- Multi-GPU scale for training and prediction
- Larger networks than are possible with a single GPU
- Optimized for fast performance on sparse datasets
- Can run locally, in a Docker container or on AWS

Amazon Destiny vs Google TensorFlow

First DSSTNE Benchmarks TLDR: Up to Almost 15x Faster than TensorFlow

“DSSTNE on a single virtualized K520 GPU (released in 2012) is faster than TensorFlow on a bare metal Tesla M40 (released in 2015)”

“TensorFlow does not provide the automagic model parallelism provided by DSSTNE”

<https://medium.com/@scottlegrand/first-dsstne-benchmarks-tldr-almost-15x-faster-than-tensorflow-393db80c0f>

Demo: Amazon Destiny

<http://grouplens.org/datasets/movielens/>

27,000 movies

138,000 users

20 million movie recommendations

(Matrix is 99.5% sparse)

→ Start a **g2.8xlarge** instance

→ Train a neural network

Input & output layers: 27,000 neurons

3 hidden layers: 128 neurons each

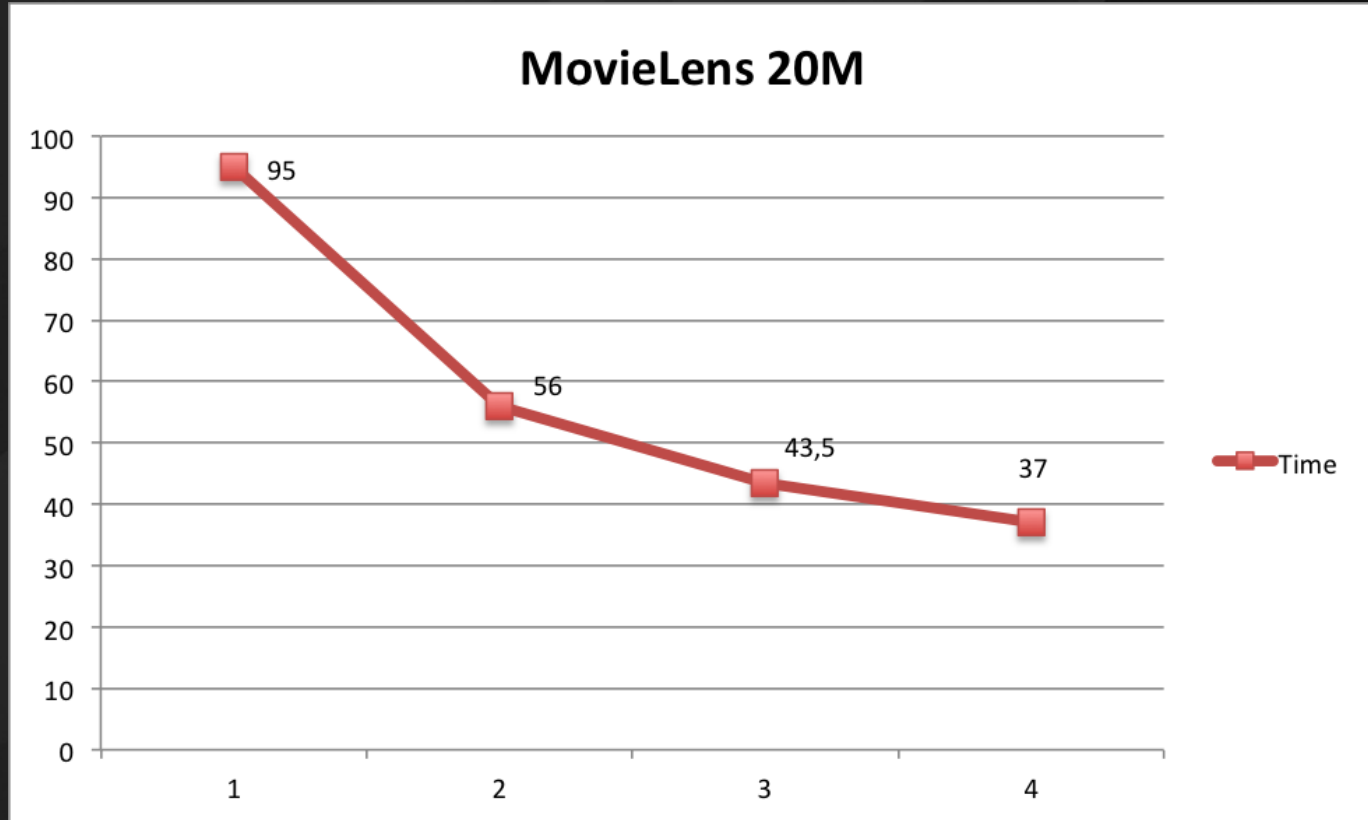
→ Recommend 10 movies per user

	Alice	Bob	Charlie	David	Ernest
Star Wars		1		1	1
Lord of the Rings	1		1		1
Inception		1			
Bambi	1			1	
Pretty Woman	1				1



	Alice	Bob	Charlie	David	Ernest
Star Wars	0.23	1	0.12	1	1
Lord of the Rings	1	0.34	1	0.89	1
Inception	0.8	1	0.43	0.76	0.45
Bambi	1	0.42	0.5	1	0.34
Pretty Woman	1	0.09	0.67	0.04	1

Training Amazon Destiny on multiple GPUs



Resources

Big Data Architectural Patterns and Best Practices on AWS

<https://www.youtube.com/watch?v=K7o5OIRLtvU>

Real-World Smart Applications With Amazon Machine Learning

<https://www.youtube.com/watch?v=sHJx1KJf8p0>

Deep Learning: Going Beyond Machine Learning

<https://www.youtube.com/watch?v=Ra6m70d3t0o>

DSSTNE: A new Deep Learning Framework For Large Sparse Datasets

<https://www.youtube.com/watch?v=LbYR6Mzq6FE>

AWS Big Data blog: <https://blogs.aws.amazon.com/bigdata/>

AWS User Groups



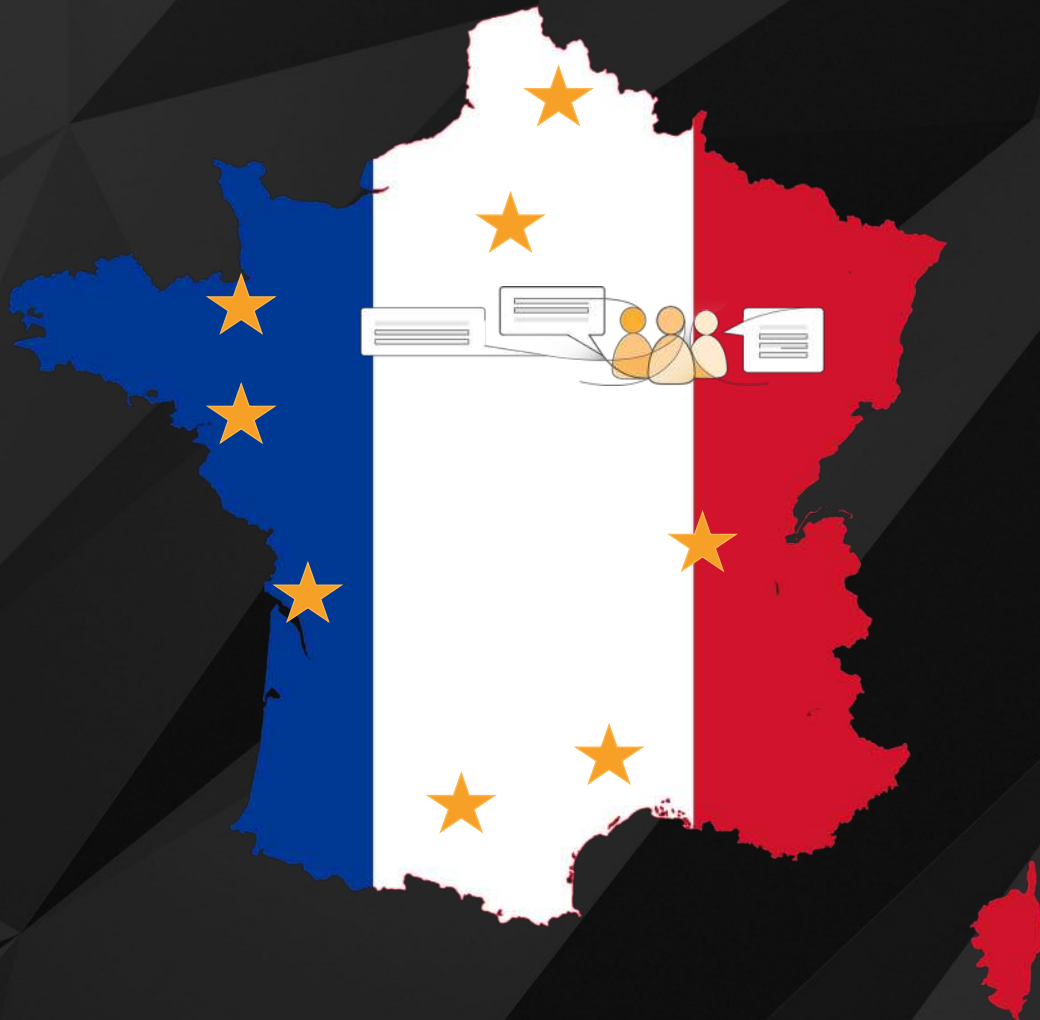
Lille
Paris
Rennes
Nantes
Bordeaux
Lyon
Montpellier
Toulouse



facebook.com/groups/AWSFrance/



[@aws_actus](https://twitter.com/aws_actus)



Thank you!

Julien Simon
Principal Technical Evangelist
Amazon Web Services

julsimon@amazon.fr
@julsimon

