GRACE HOPPER CELEBRATION

ANITA B.ORG

Proof-of-concept to Production

How to scale your Deep Learning Models

Presenters: Swetha Mandava, Alex Qi

Contributor: Michael Mendelson

#GHC19

AGENDA

Deep Learning at Scale

- Why scaling
- Tips and Tricks

Desired outcomes

- Understand when to and how to scale
- Know the typical techniques to apply
- Understand the theory/concepts of typical scaling techniques

Deep Learning At Scale

Proof-of-concept

- Model architecture search
- Hyperparameter Optimization
- Toy dataset/Toy models
- Low frequency of retraining
- Non-optimized resource utilization

Production

- Models need to be trained and retrained with shorter times and higher frequency.
- Massive datasets
- Big & complex models

Pain Points

DL models training:

- **Time consuming**: can take days, weeks...
- **Capability:** is limited by memory capacity on batch size and model size

Scaling is necessary but hard:

- Convergence and Stability
- Computation and Scaling efficiency
- Hardware Limits

Src: https://mc.ai/ibm-offering-gpu-powered-notebooks/

Techniques for Scaling Deep Learning Training

- **Convergence and Stability**
  - Warmup
  - Linear Scaling Rule
  - LARS

- **Computation and Scaling efficiency**
  - Automatic Mixed Precision

- **Hardware limits on Dataset and Model Size**
  - User profiles/preference of millions of users
  - Data Parallelism
  - Model Parallelism

Scaling Success

BERT PreTraining on DGX SuperPOD

# V100 GPUs	Time to train (hours)
16	58.4 (2.4 days)
256	3.9
1024	1.2
1472	0.9 (53 min)

GOAL :

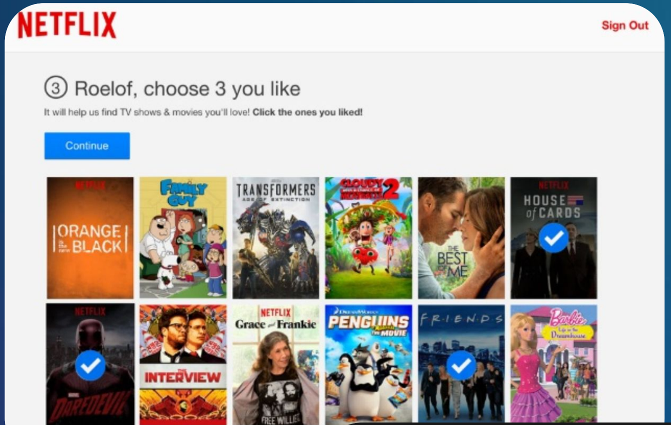
1. Maintain Accuracy
2. Decrease Time to Train

Let’s do it on a smaller model: NCF
BERT has 110 million parameters!

Recommender System: Neural Collaborative Filtering

Recommendation engines are everywhere…

- Personalized
- Convenient
- More efficient

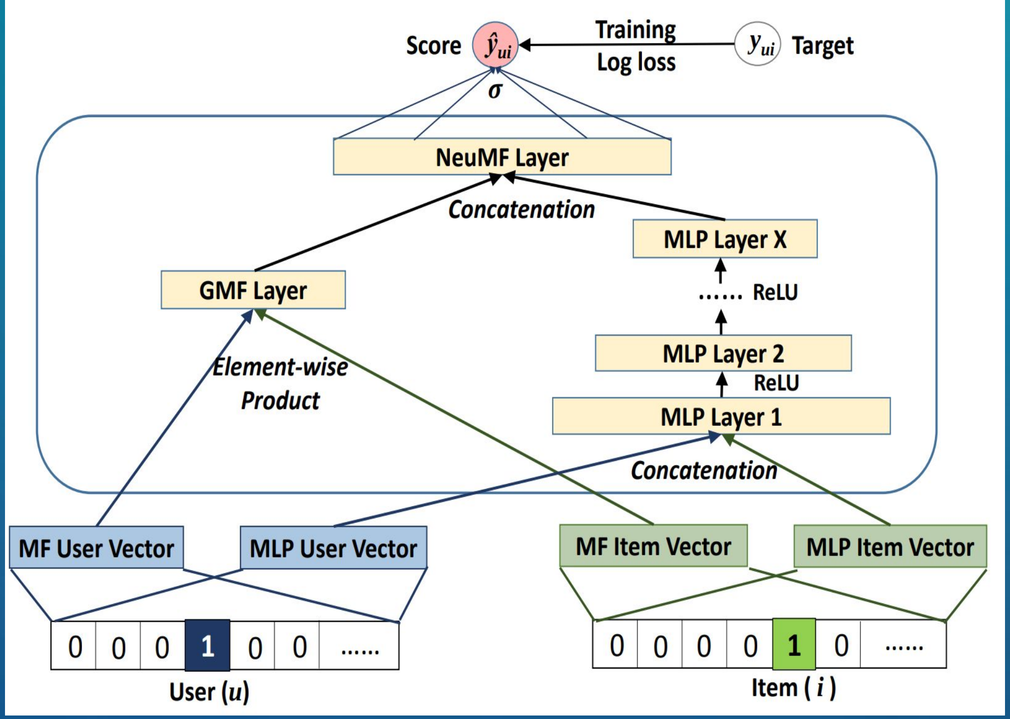


③ Roelof, choose 3 you like

It will help us find TV shows & movies you'll love! **Click the ones you liked!**

Watch All Episodes Now

First they got the coke. Then they got the money. Now the Colombian cartels want the power. Let the drug wars begin.



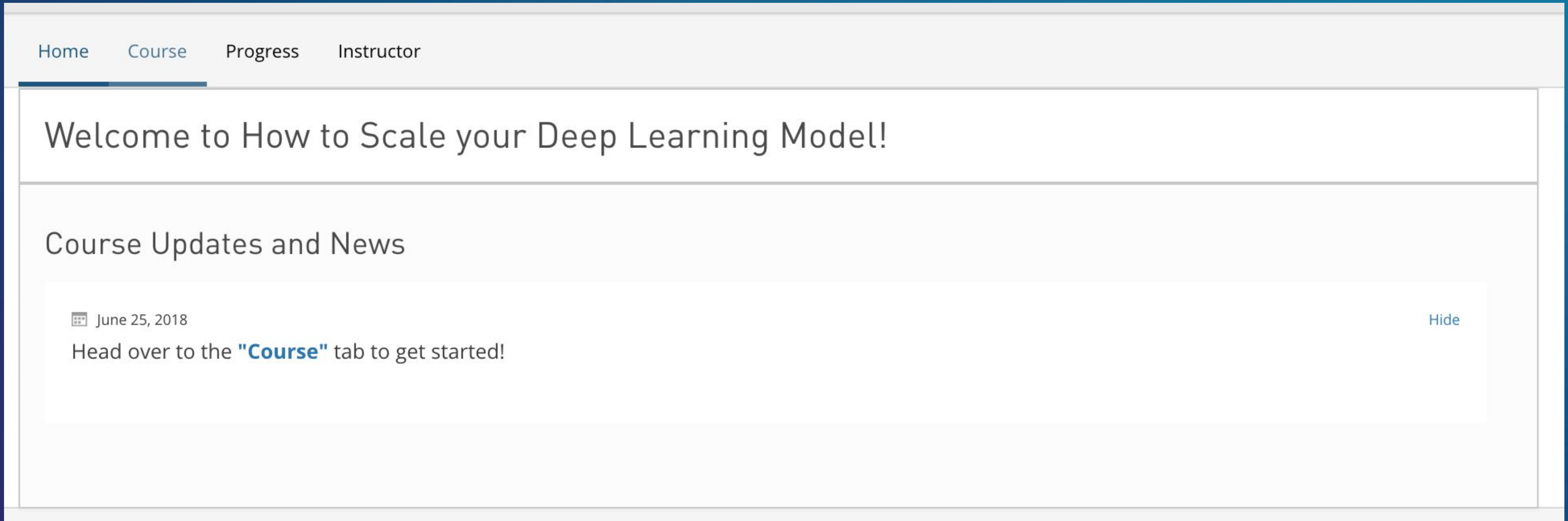
Deep Learning based Recommendation system model architecture

#GHC19

Launch Hands-on Task

- Create Account at https://courses.nvidia.com/join
- Go to courses.nvidia.com/dli-event
- Browser Recommendation: Chrome
- Use event code and create an account

Select the “Course” Tab



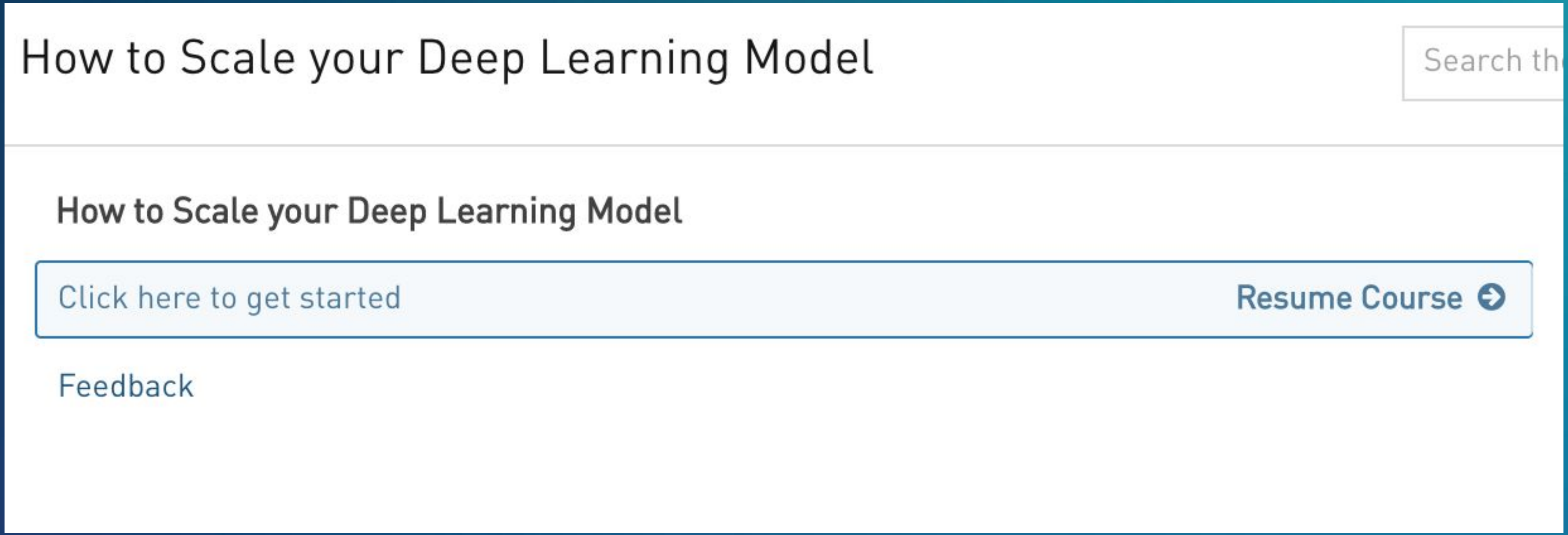
Welcome to How to Scale your Deep Learning Model!

Course Updates and News

June 25, 2018

Head over to the **"Course"** tab to get started!

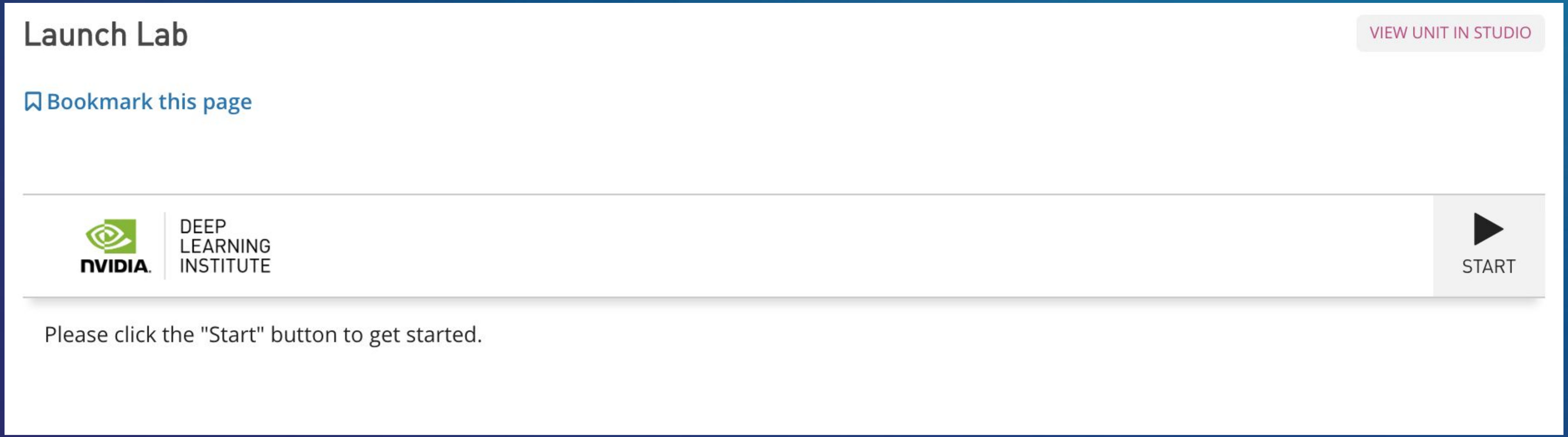
Open the first hands-on section



How to Scale your Deep Learning Model

How to Scale your Deep Learning Model

Select the Start button and wait



Launch Lab

Please click the "Start" button to get started.

What else can we do?

- Is 192 still the maximum batch size we can use?
- Can we scale further than 1 GPU?
  - Data Parallelism

https://github.com/NVIDIA/DeepLearningExamples

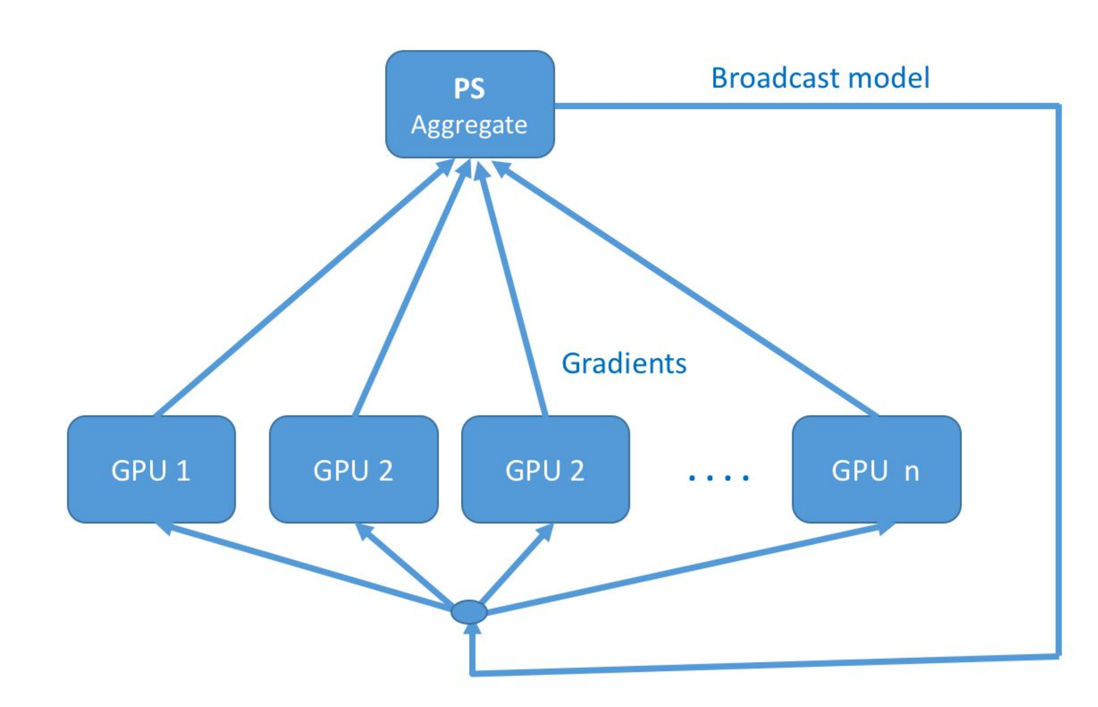
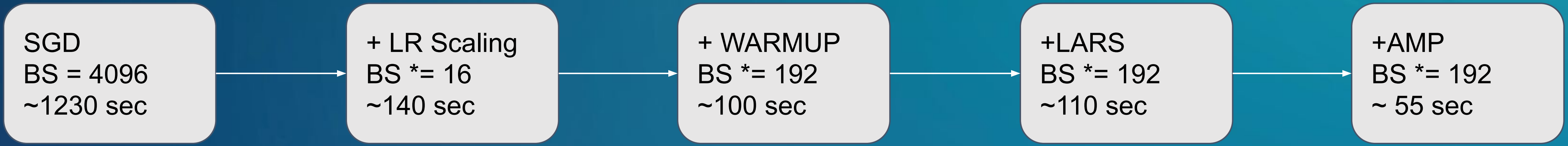


*Image from Towards Data Science

Summary

Productivity matters : teams with better tools/scaling can try out more ideas



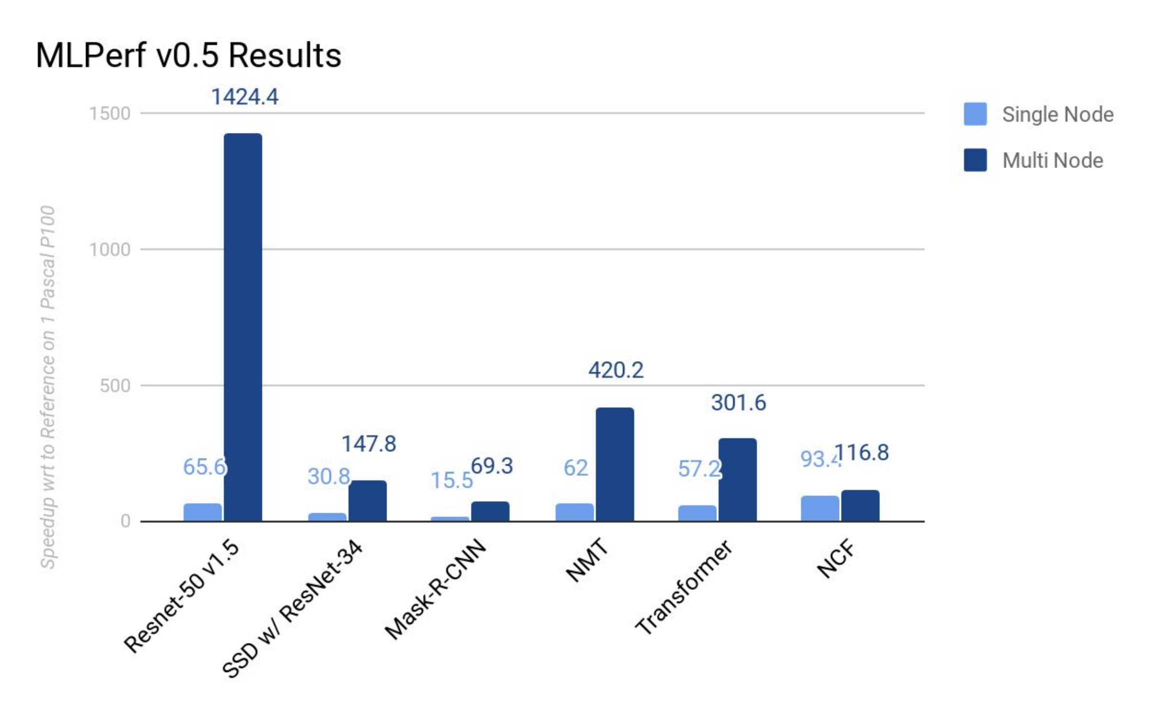
SGD
BS = 4096
~1230 sec

+ LR Scaling
BS *= 16
~140 sec

+ WARMUP
BS *= 192
~100 sec

+LARS
BS *= 192
~110 sec

+AMP
BS *= 192
~ 55 sec



Thank You!

Please remember to complete the session survey in the mobile app.

Follow us @

https://www.linkedin.com/in/alexlingqi/

https://www.linkedin.com/in/swethmandava/

@swethmandava
@mrmendelson

GRACE HOPPER CELEBRATION
ANITA B.ORG

#GHC19