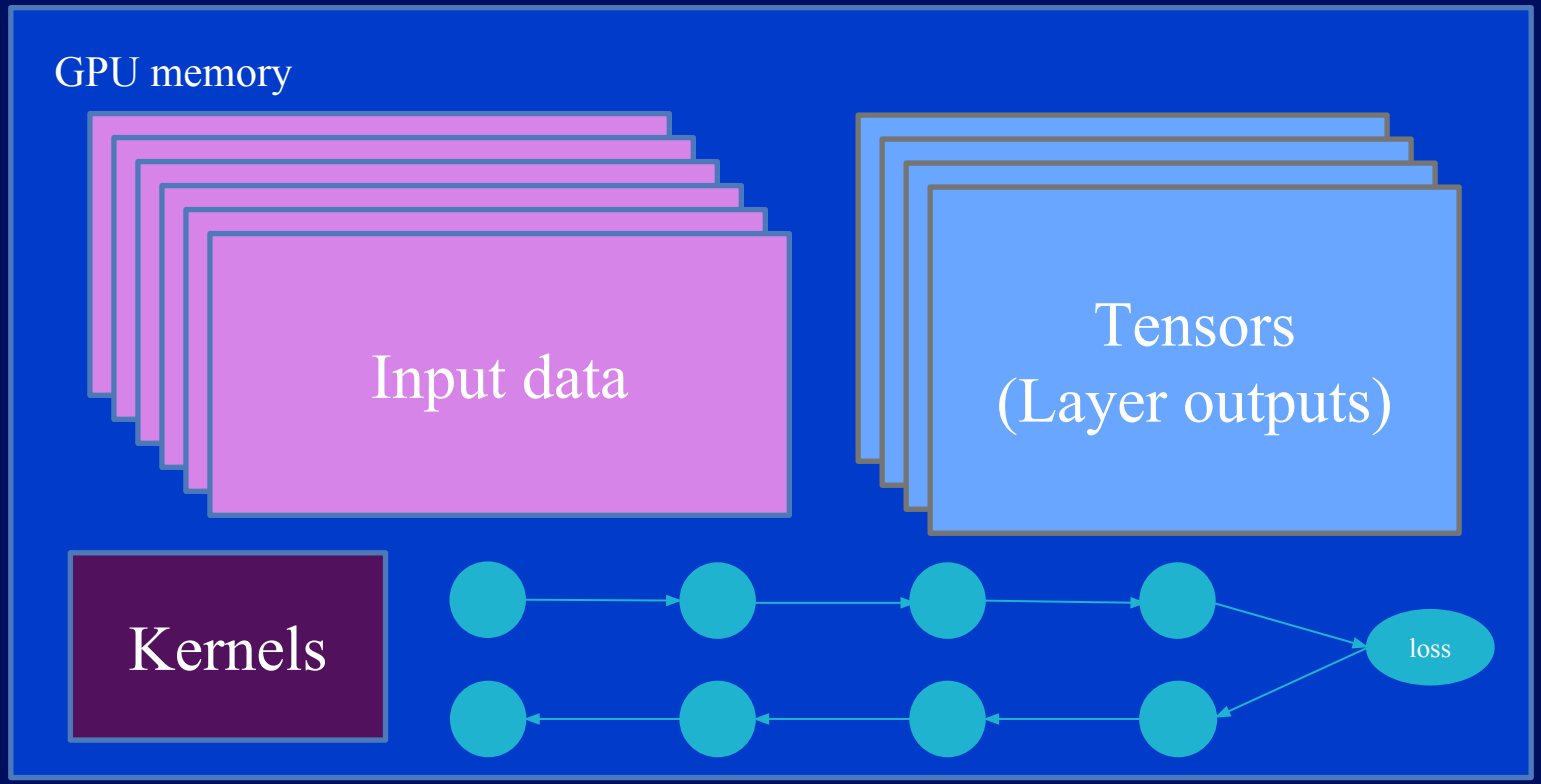
Using Tensor Swapping and NVLink to Overcome GPU Memory Limits with TensorFlow

Sam Matzek

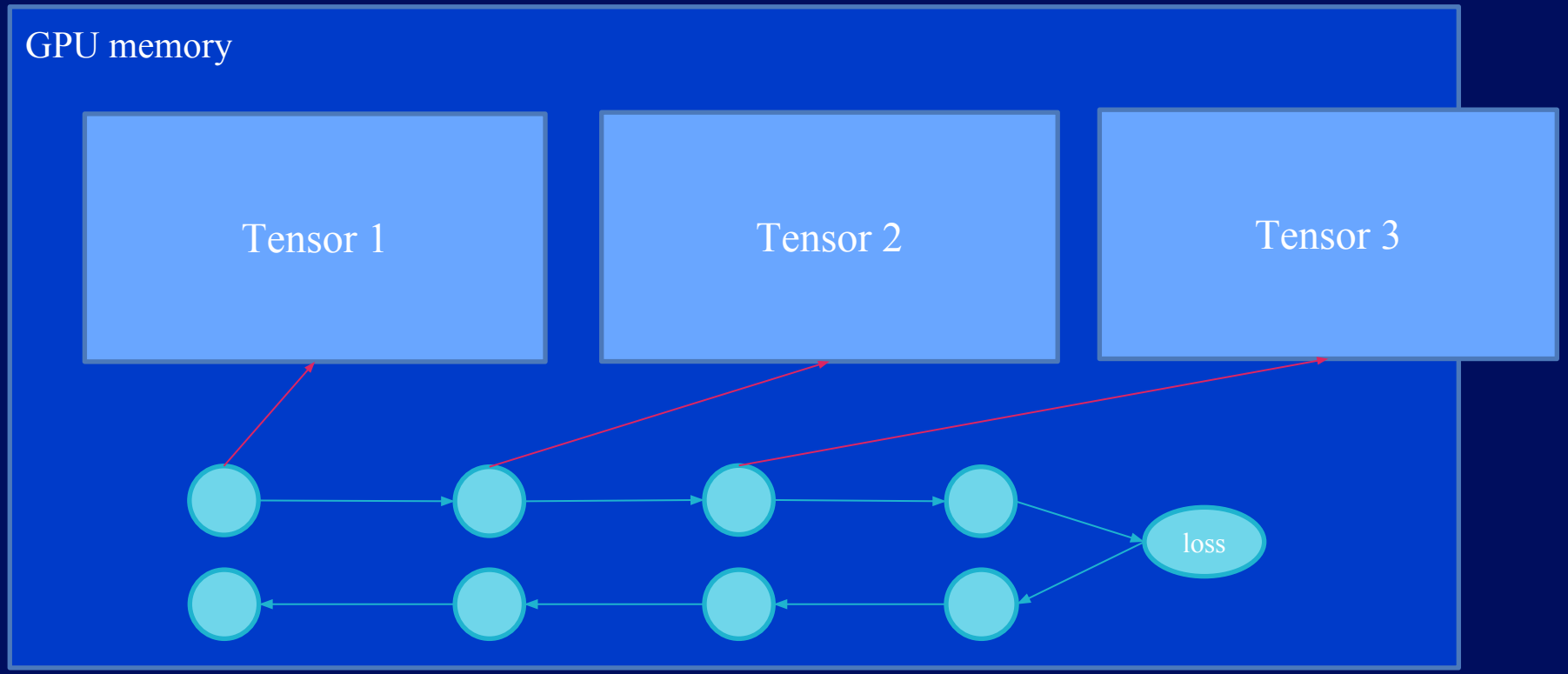
Deep learning is memory constrained

- GPUs have limited memory
- Neural networks are growing deeper and wider
- Amount and size of data to process is always growing

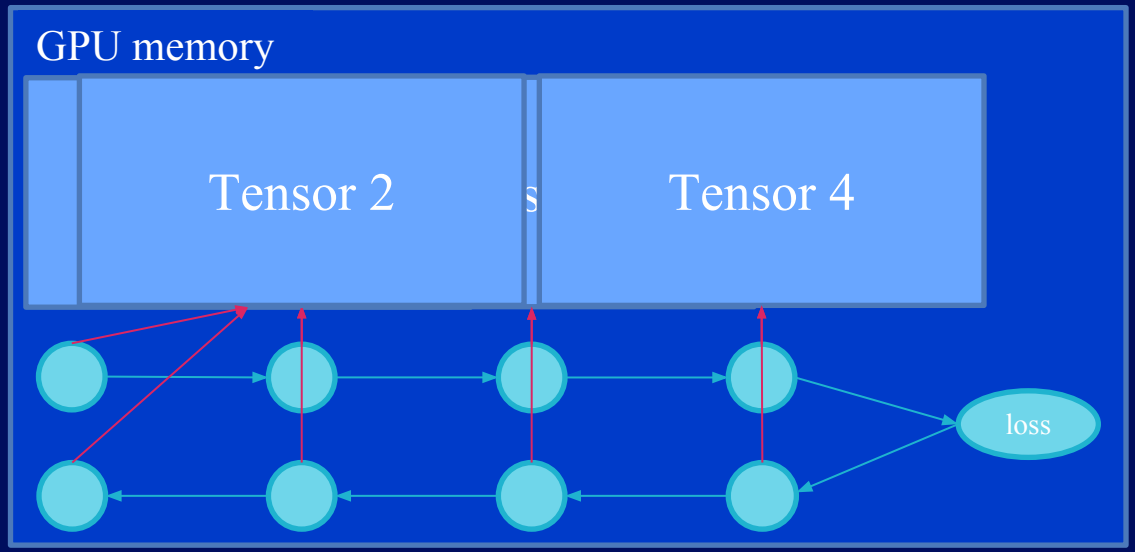
GPU Memory Usage



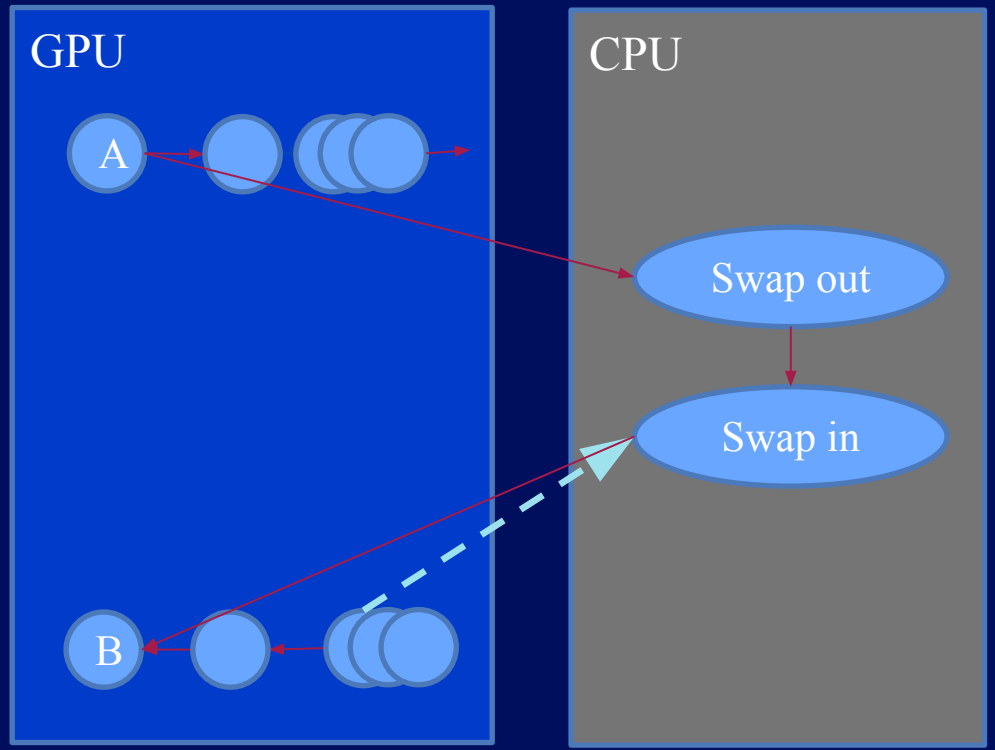
Model Training in GPU Memory



Model Training with Tensor Swapping



TensorFlow Large Model Support Graph Modifications



https://arxiv.org/pdf/1807.02037.pdf

Enabling TensorFlow Large Model Support

Keras API

```
from tensorflow_large_model_support import LMS
lms = LMS()
lms.batch_size = 1
# ...
model.fit_generator(generator=training_gen,
                    callbacks=[lms])
```

Estimator API

```
from tensorflow_large_model_support import LMS
lms = LMS()
# ...
mnist_classifier.train(input_fn=train_input_fn, steps=20000,
                       hooks=[logging_hook, lms])
```

* Examples for TFLMS v2.0.0

What's possible with Large Model Support?

- 10x image resolution - Keras ResNet50
- 10x image resolution - DeepLabV3 2D image segmentation
- 5x MRI resolution - 3D U-Net 3D image segmentation

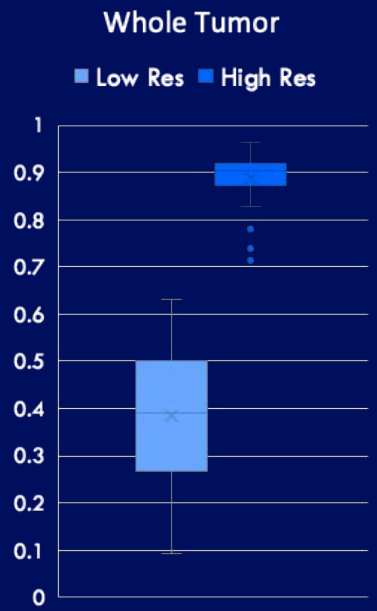
Measured with TFLMS v2.0.0 on TensorFlow 1.13, CUDA 10.1, cuDNN 7.5

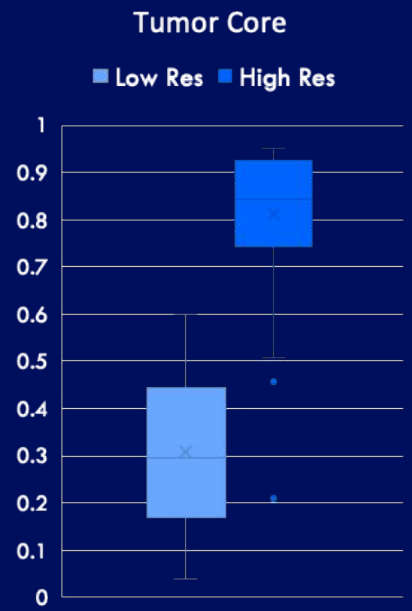
3D U-Net image segmentation

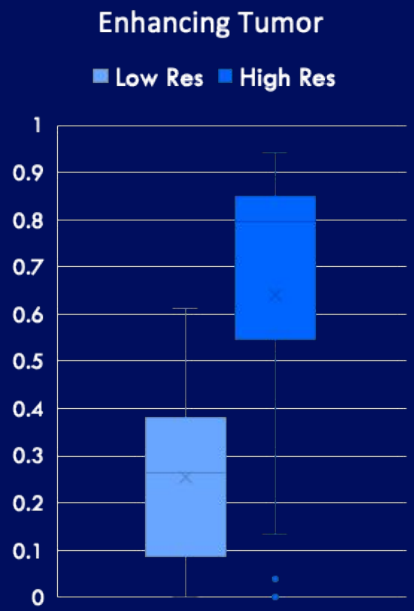
- 3D U-Net generally has high memory usage requirements
- International Multimodal Brain Tumor Segmentation Challenge (BraTS)
- Existing Keras model with TensorFlow backend

Effect of 2x resolution on Dice Coefficients

(higher is better)

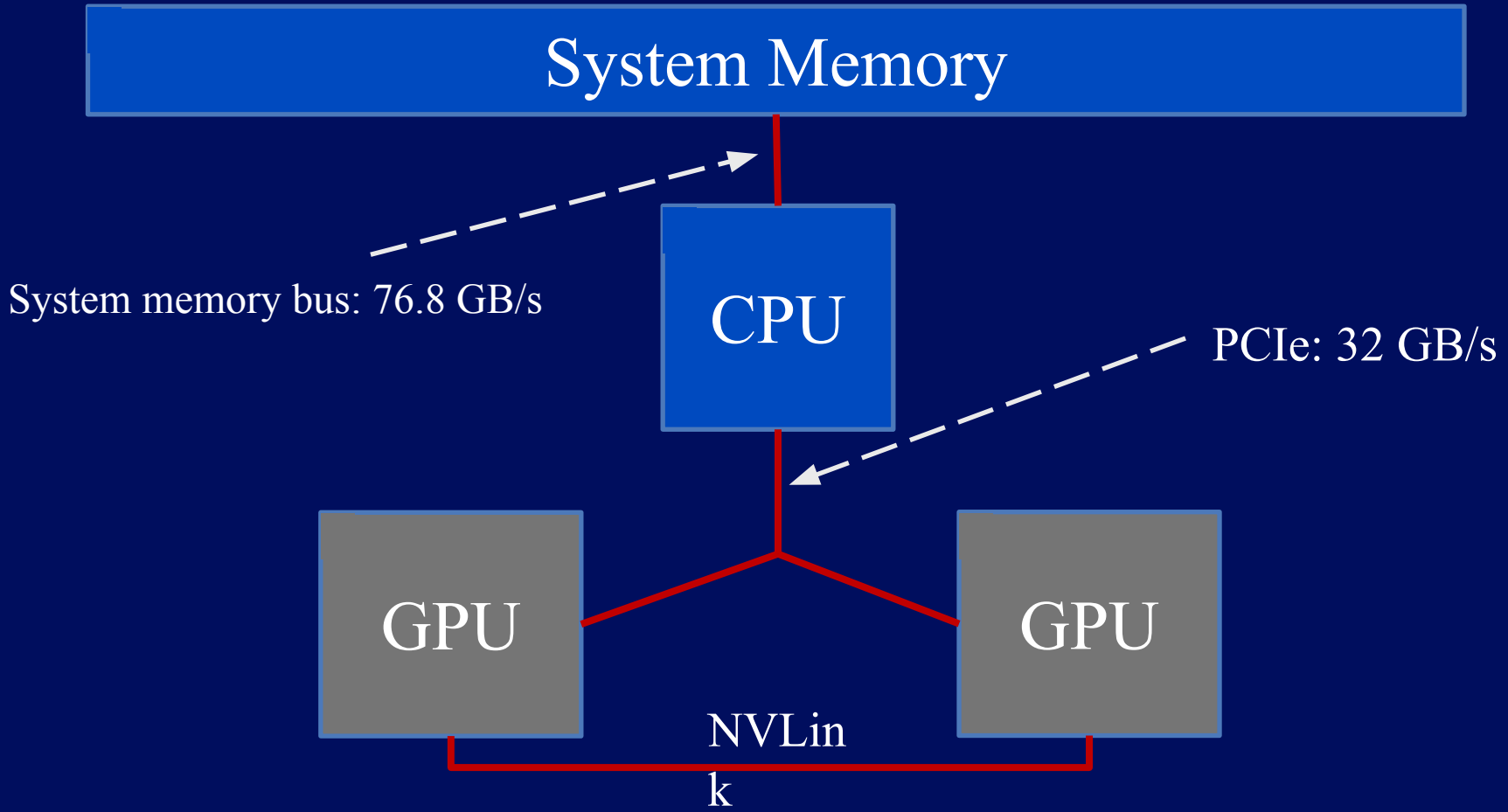




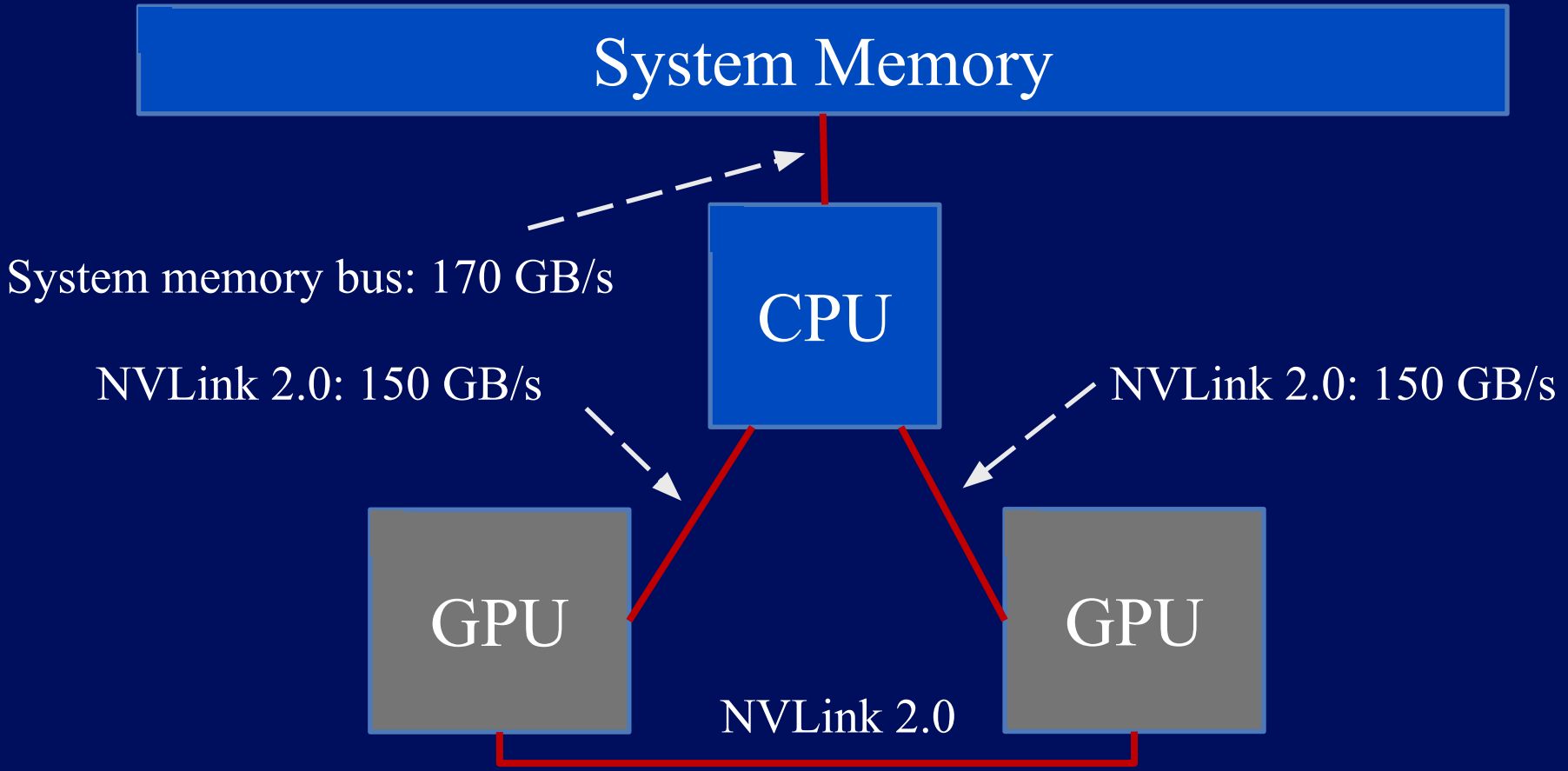


“Swapping makes everything slow”

Typical GPU connectivity



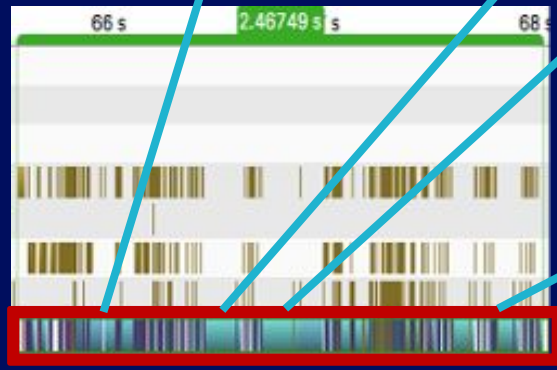
POWER9 CPU to GPU connectivity



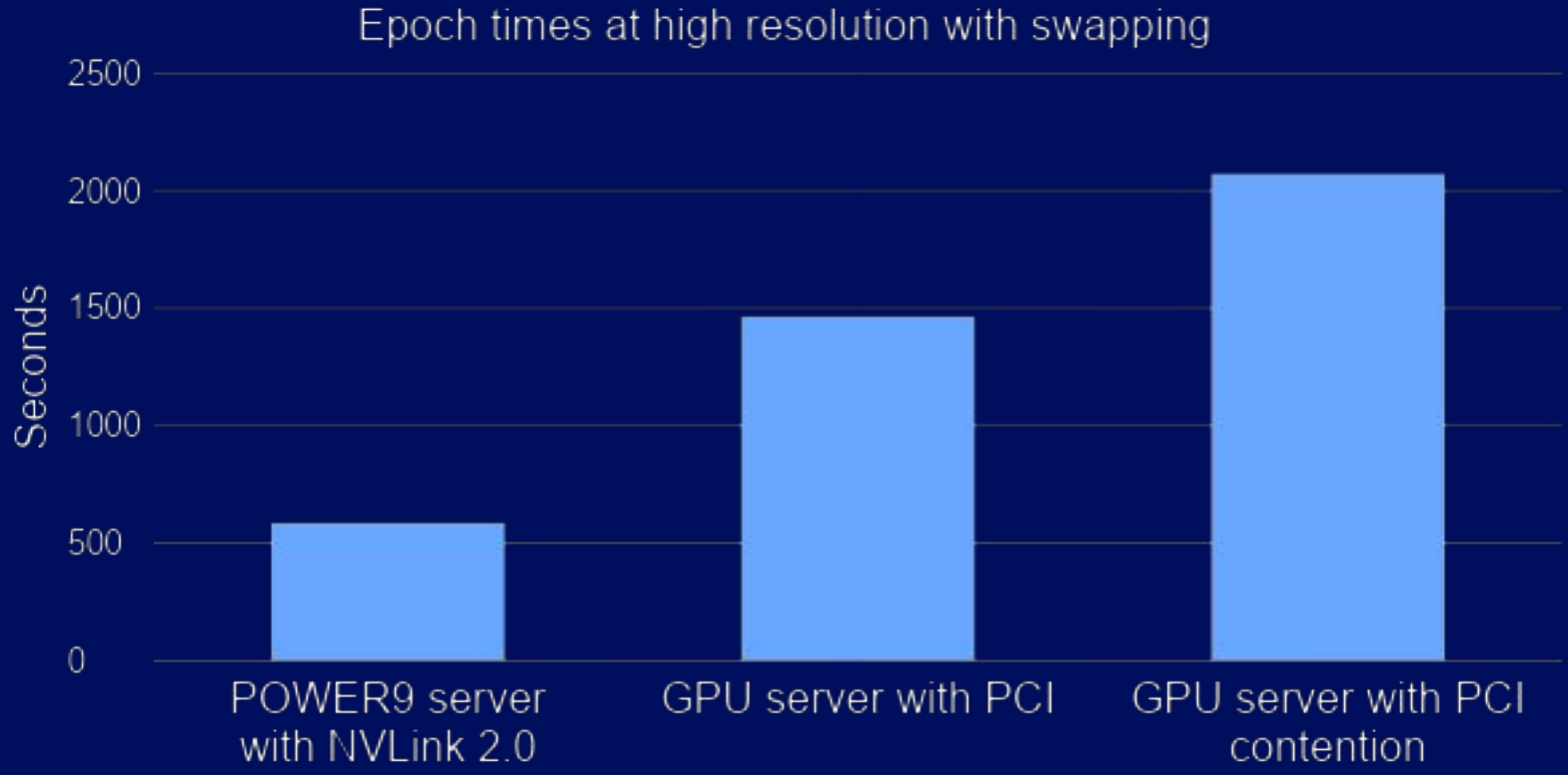
Effects of NVLink 2.0 on Large Model Support

PCIe connected GPU training one high res 3D MRI with large model support

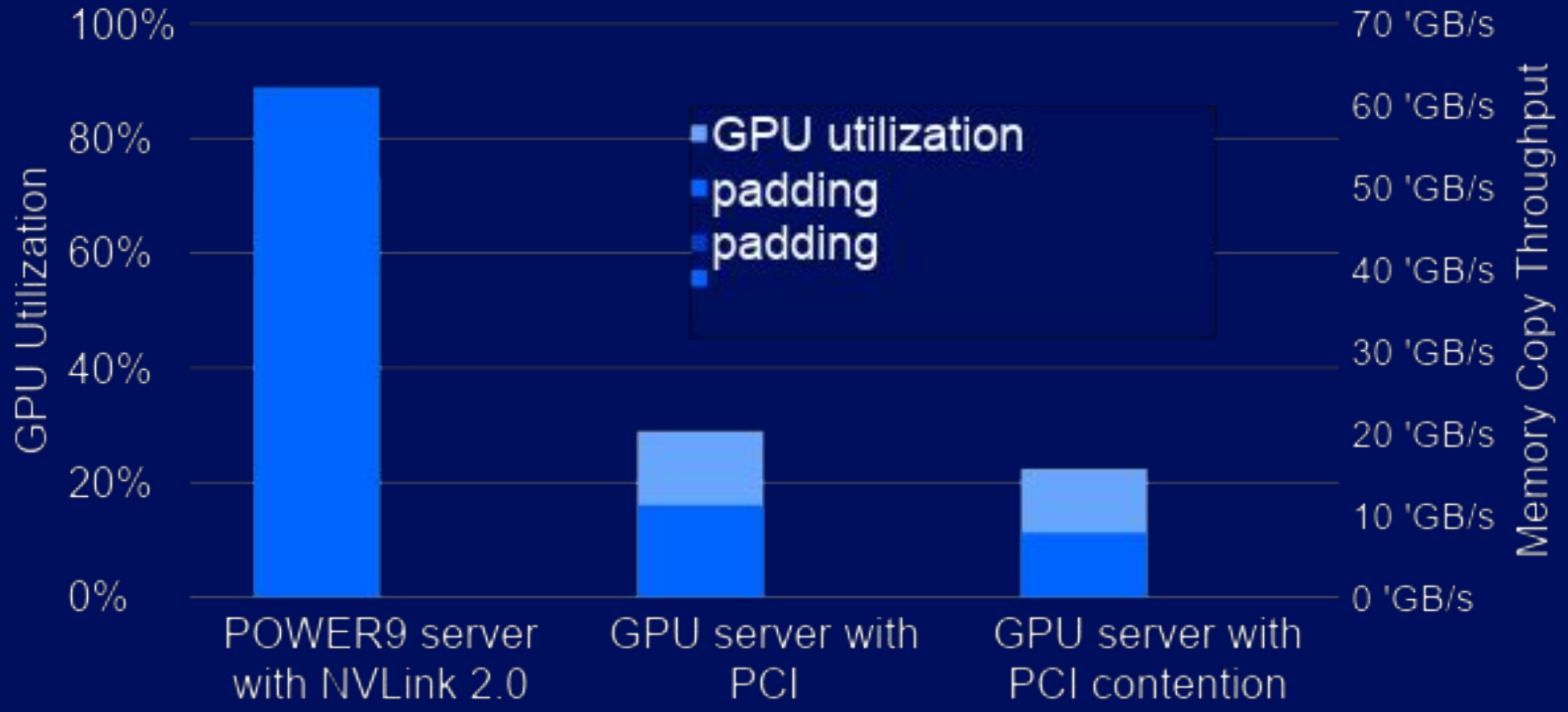
NVLink 2.0 connected GPU training one high res 3D MRI with large model support



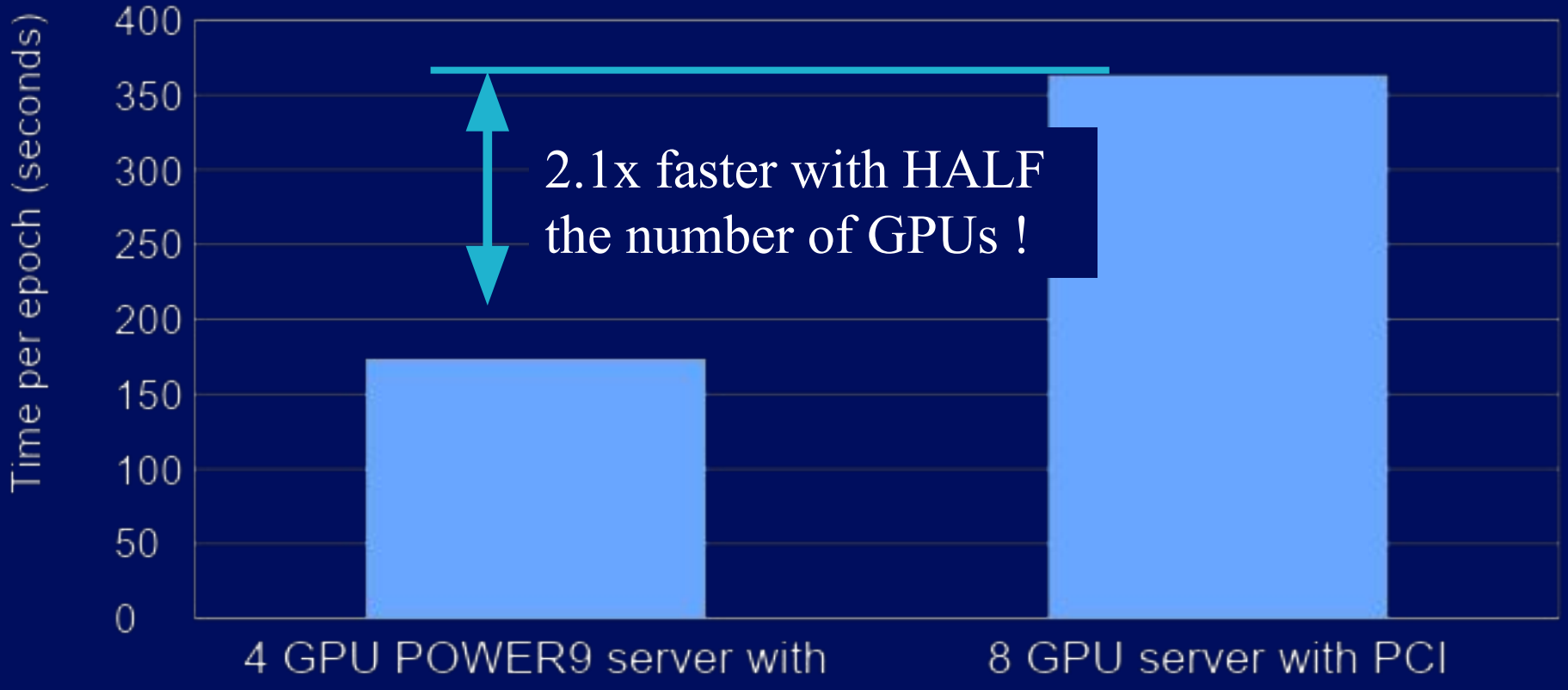
Effects of NVLink 2.0 on epoch times



Effects of NVLink 2.0 on GPU Utilization

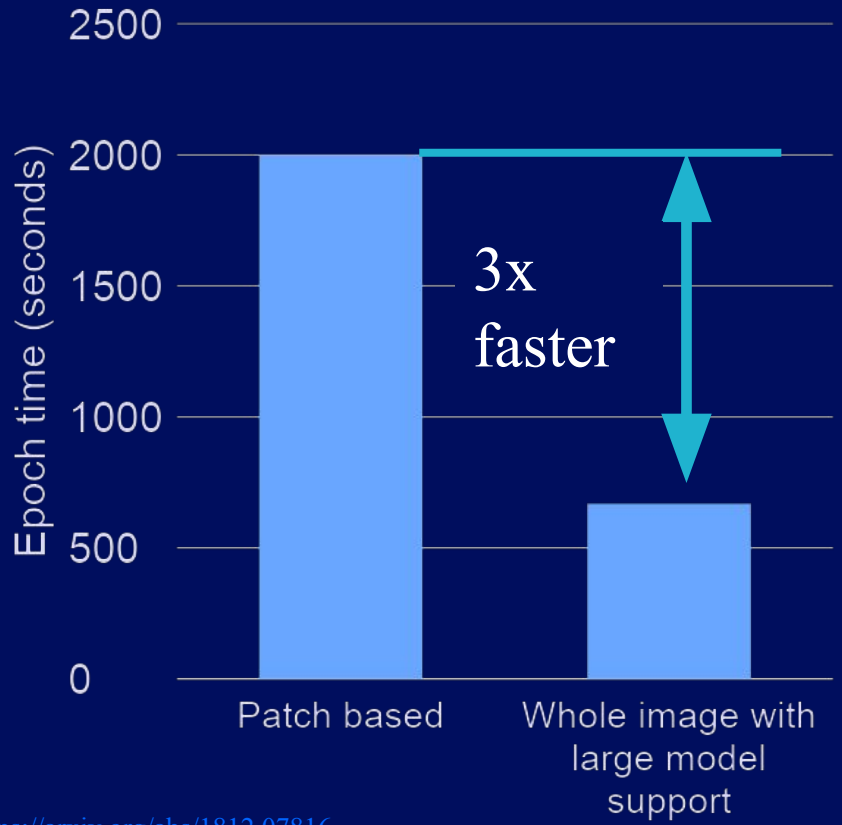


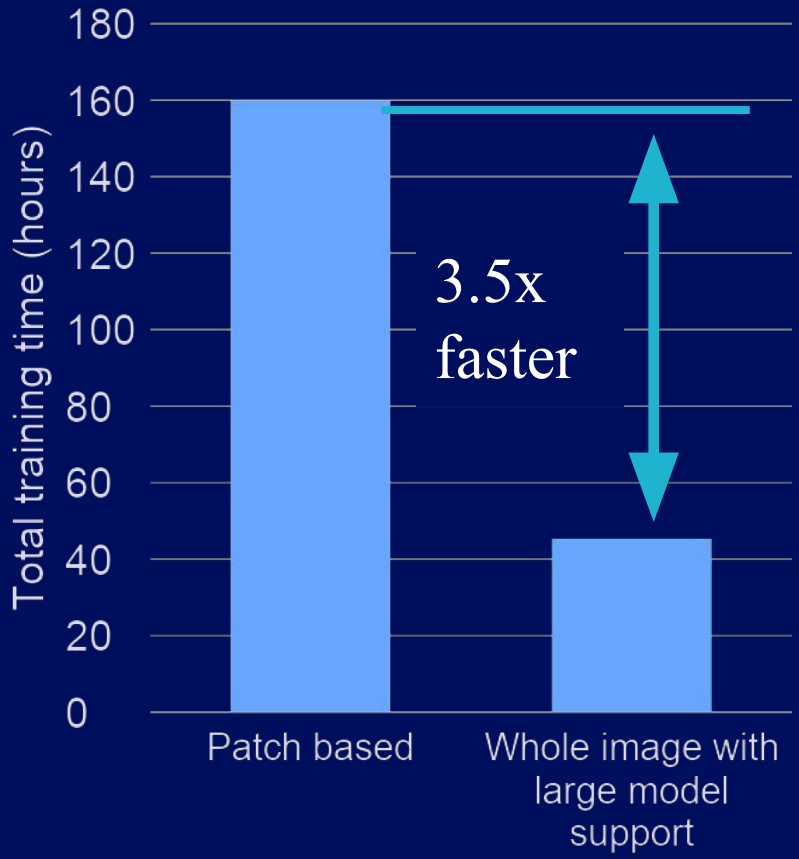
Multi-GPU model training with NVLink 2.0



http://ibm.biz/3dunet-tflms-multigpu

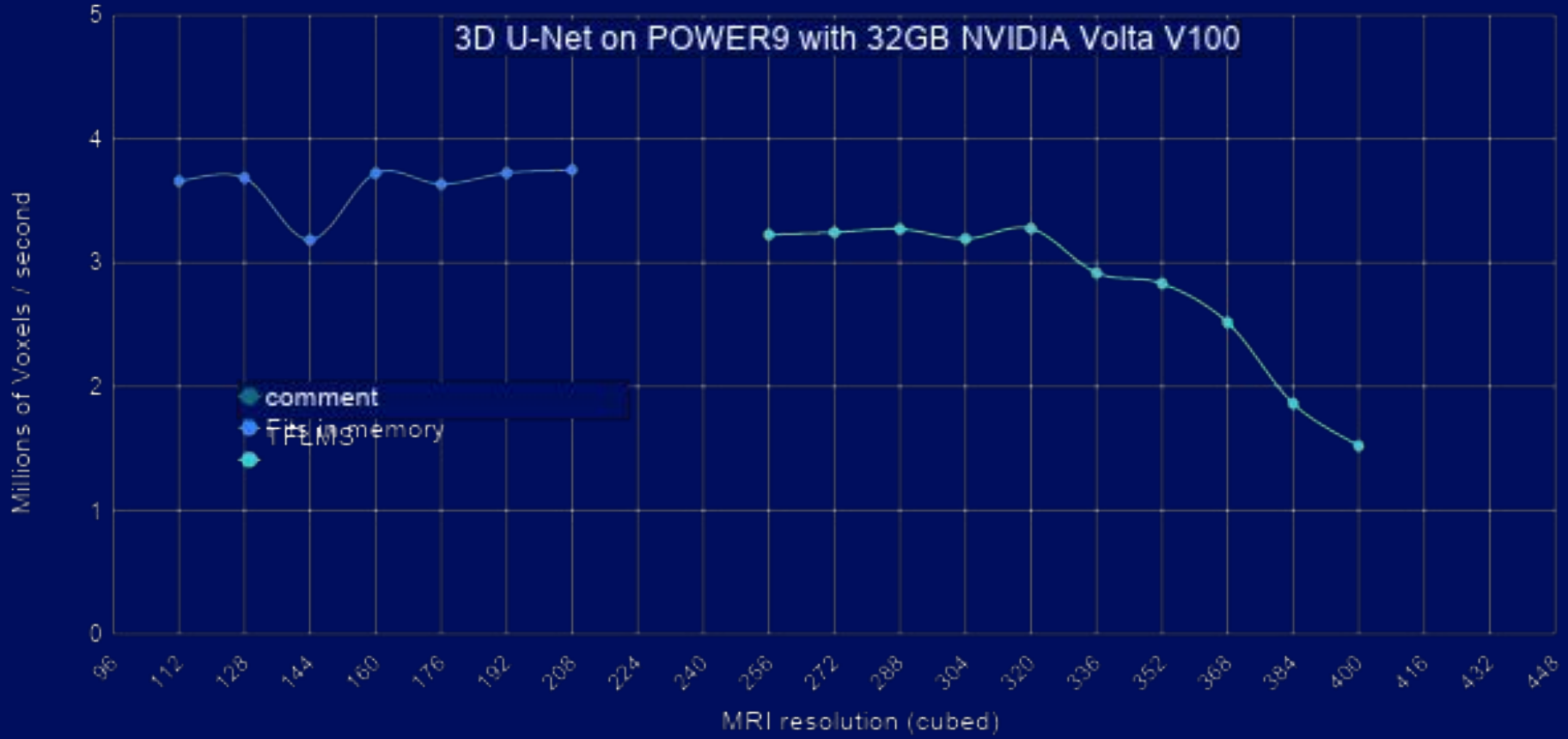
Patches versus whole image





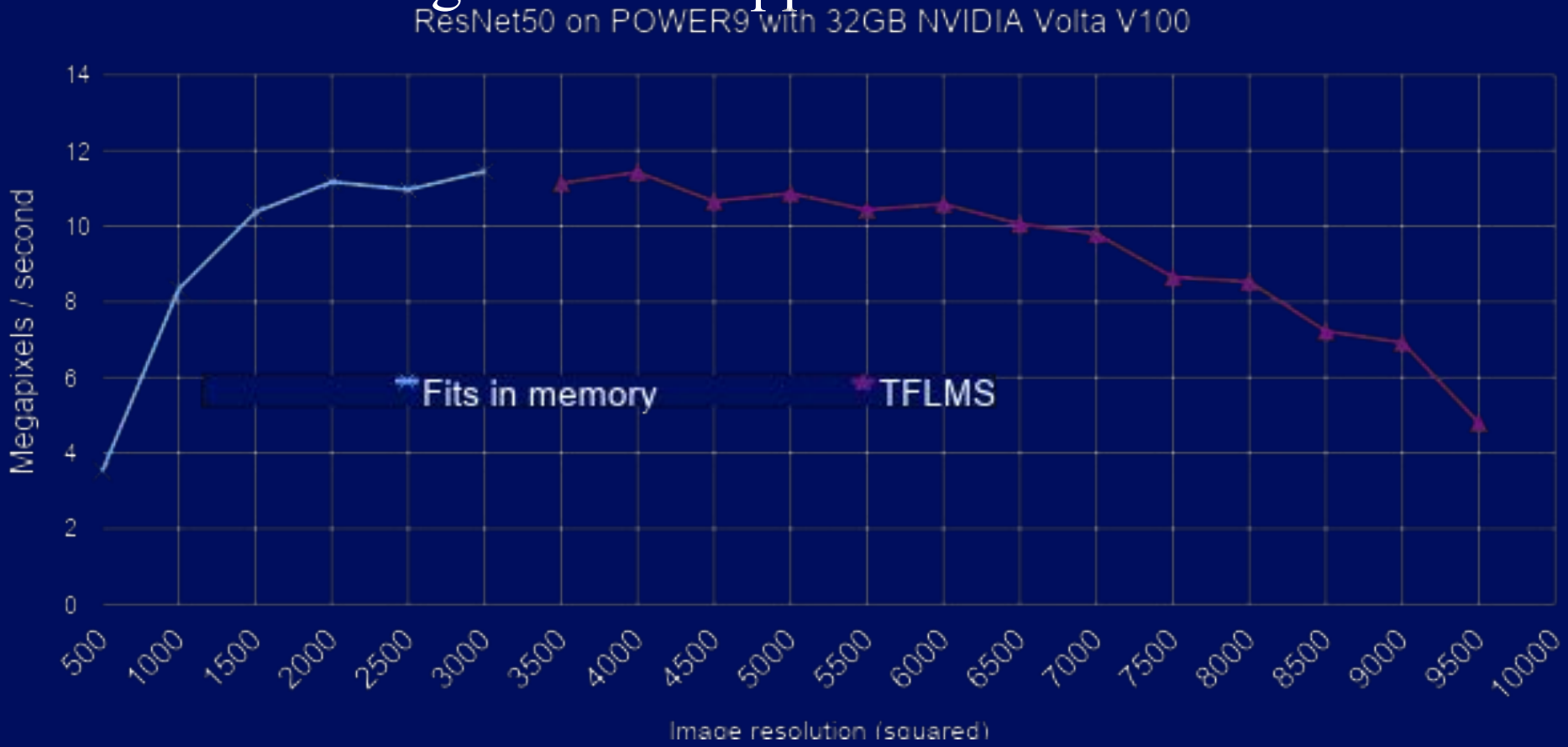
https://arxiv.org/abs/1812.07816

Overhead of Large Model Support with NVLink 2.0



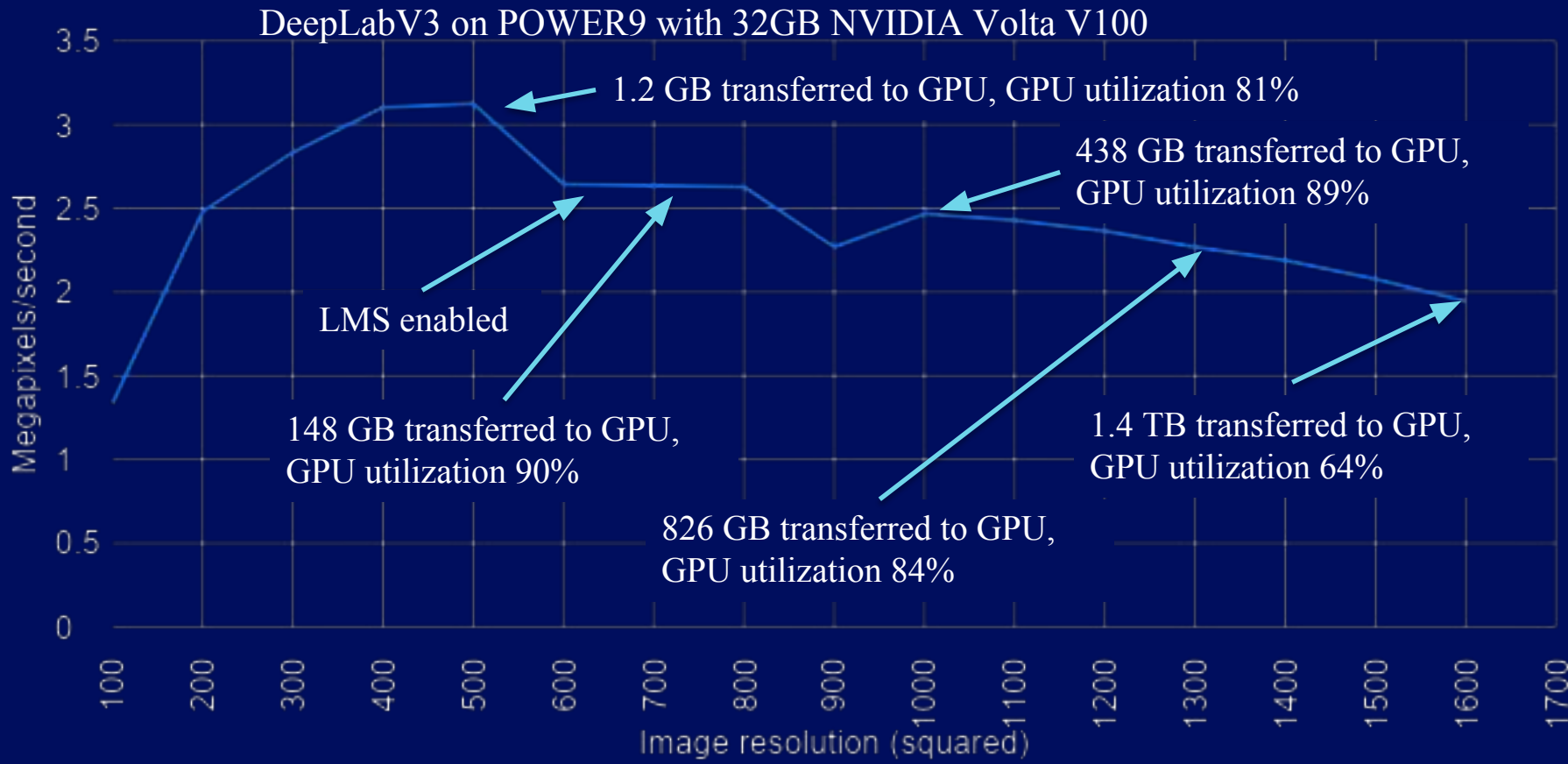
Measured with TFLMS v2.0.0 on TensorFlow 1.13, CUDA 10.1, cuDNN 7.5

Overhead of Large Model Support with NVLink 2.0



Measured with TFLMS v2.0.0 on TensorFlow 1.13, CUDA 10.1, cuDNN 7.5

Overhead of Large Model Support with NVLink 2.0



Using bs=16, fine_tune_batch_norm=true, measured on 32GB GPU with TensorFlow 1.13, CUDA 10.1, cuDNN 7.5

Large Model Support with NVLink 2.0

- Tensor swapping can be used to overcome GPU memory limits
- Allows training of:
  - deeper models
  - higher resolution data
  - larger batch sizes
- NVLink 2.0 between CPU and GPU allow tensor swapping with minimal overhead

More information

TensorFlow Large Model Support
https://github.com/IBM/tensorflow-large-model-support

TFLMS: Large Model Support in TensorFlow by Graph Rewriting
https://arxiv.org/pdf/1807.02037.pdf

TensorFlow Large Model Support Case Study
https://developer.ibm.com/linuxonpower/2018/07/27/tensorflow-large-model-support-case-study-3d-image-segmentation/

Performance of 3DUnet Multi GPU Model for Medical Image Segmentation using TensorFlow Large Model Support
http://ibm.biz/3dunet-tflms-multigpu

Fast and Accurate 3D Medical Image Segmentation with Data-swapping Method
https://arxiv.org/abs/1812.07816

Data-parallel distributed training of very large models beyond GPU capacity
https://arxiv.org/abs/1811.12174

POWER9 server with NVLink 2.0 connections between CPU and GPU (IBM AC922):
https://www.ibm.com/us-en/marketplace/power-systems-ac922