DASYA DAPHNE IT UNIVERSITY OF COPENHAGEN

www.dasya.dk https://daphne-eu.github.io/ www.itu.dk

toward hardware-conscious data science

(or how is academia going for me so far?)

Pınar Tözün

IT University of Copenhagen

pito@itu.dk, www.pinartozun.com

FGDB 2022

25/03/2022

IT UNIVERSITY OF COPENHAGEN

www.dasya.dk https://daphne-eu.github.io/ www.itu.dk

toward hardware-conscious machine learning

(or how is academia going for me so far?)

Pınar Tözün

IT University of Copenhagen

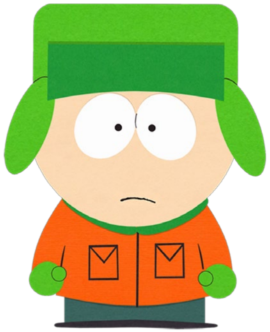
pito@itu.dk, www.pinartozun.com

FGDB 2022 25/03/2022

how did i get into machine learning?

sebastian baunsgaard

me



Could you supervise our MSc thesis?

What would you like to work on?

Automatic speech recognition

Why are you talking to me?

We want to make it scalable

sebastian benjamin wrede

ok then

how did i get into machine learning?

sebastian baunsgaard

me

Could you supervise our MSc thesis?

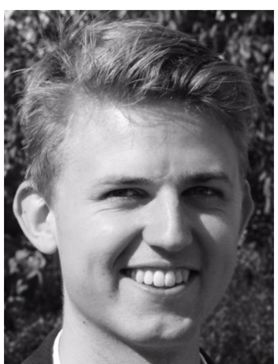
What would you like to work on?

Automatic speech recognition

Why are you talking to me?

We want to make it scalable

ok then

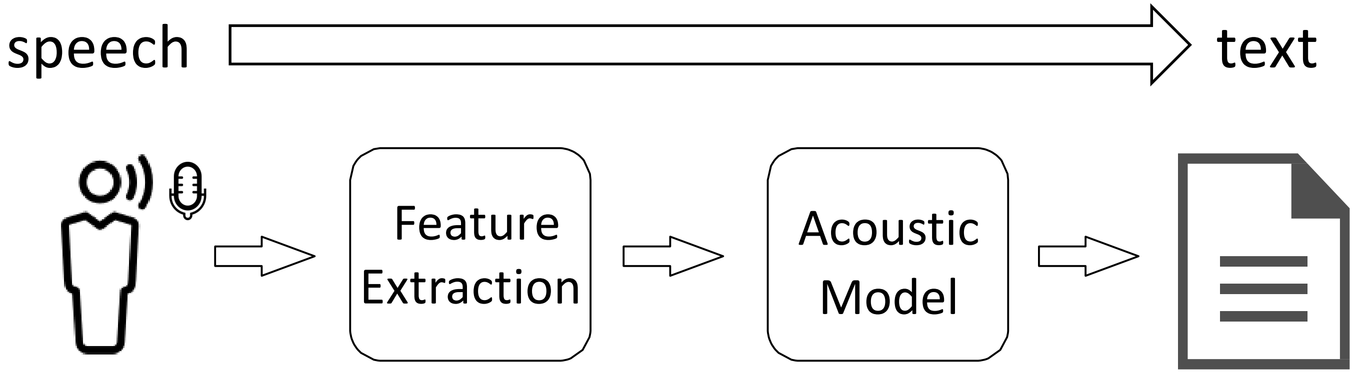


sebastian benjamin wrede

agenda

- training speech recognition on co-processors[ADMS2020]
- studying workload co-location
- challenges & opportunities

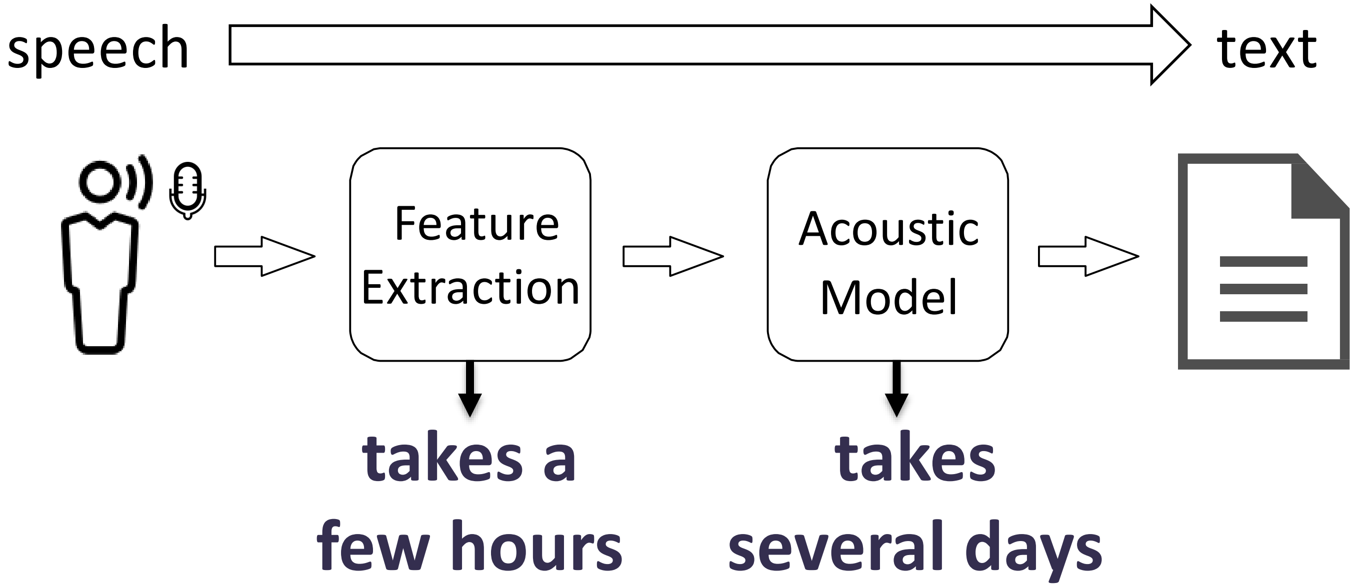
speech recognition



- human-computer & human-human interactions
- hospitals, call-centers, etc.

state-of-the-art *acoustic models* are based on neural networks in recent years → natural fit for GPUs

speech recognition

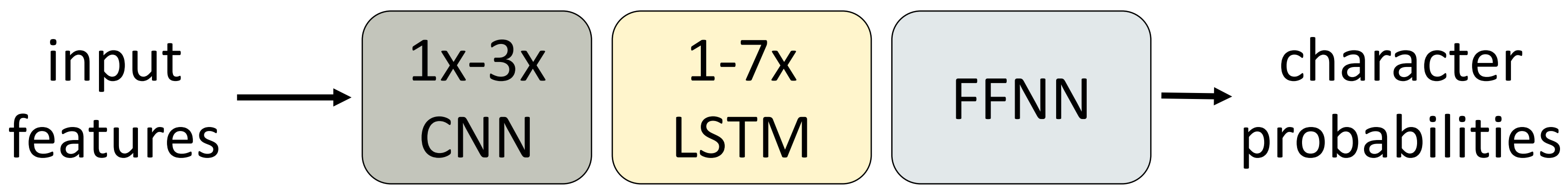


state-of-the-art *acoustic models* are based on neural networks in recent years → natural fit for GPUs

acoustic model

our final model

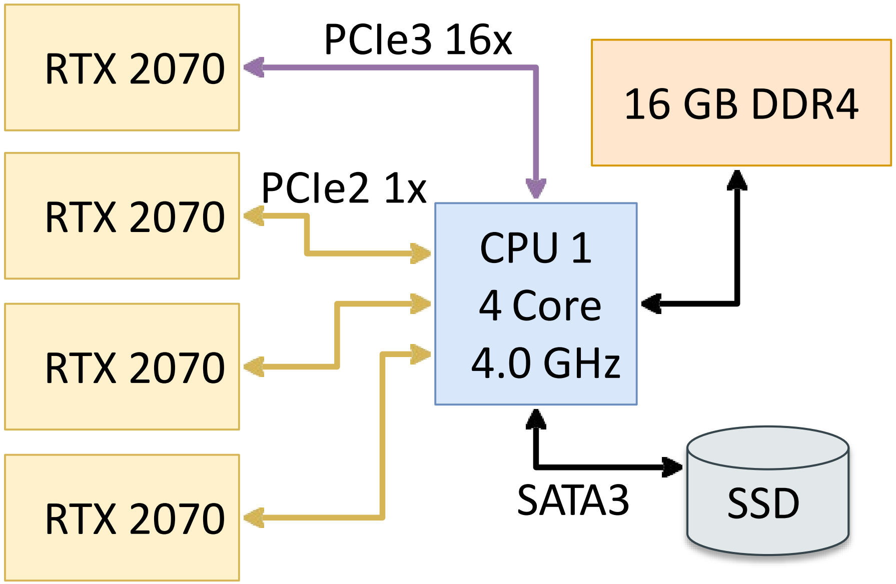
- inspired by Baidu Research, Deep Speech 2, ICML 2016
- basis for MLPerf’s speech recognition benchmark as well



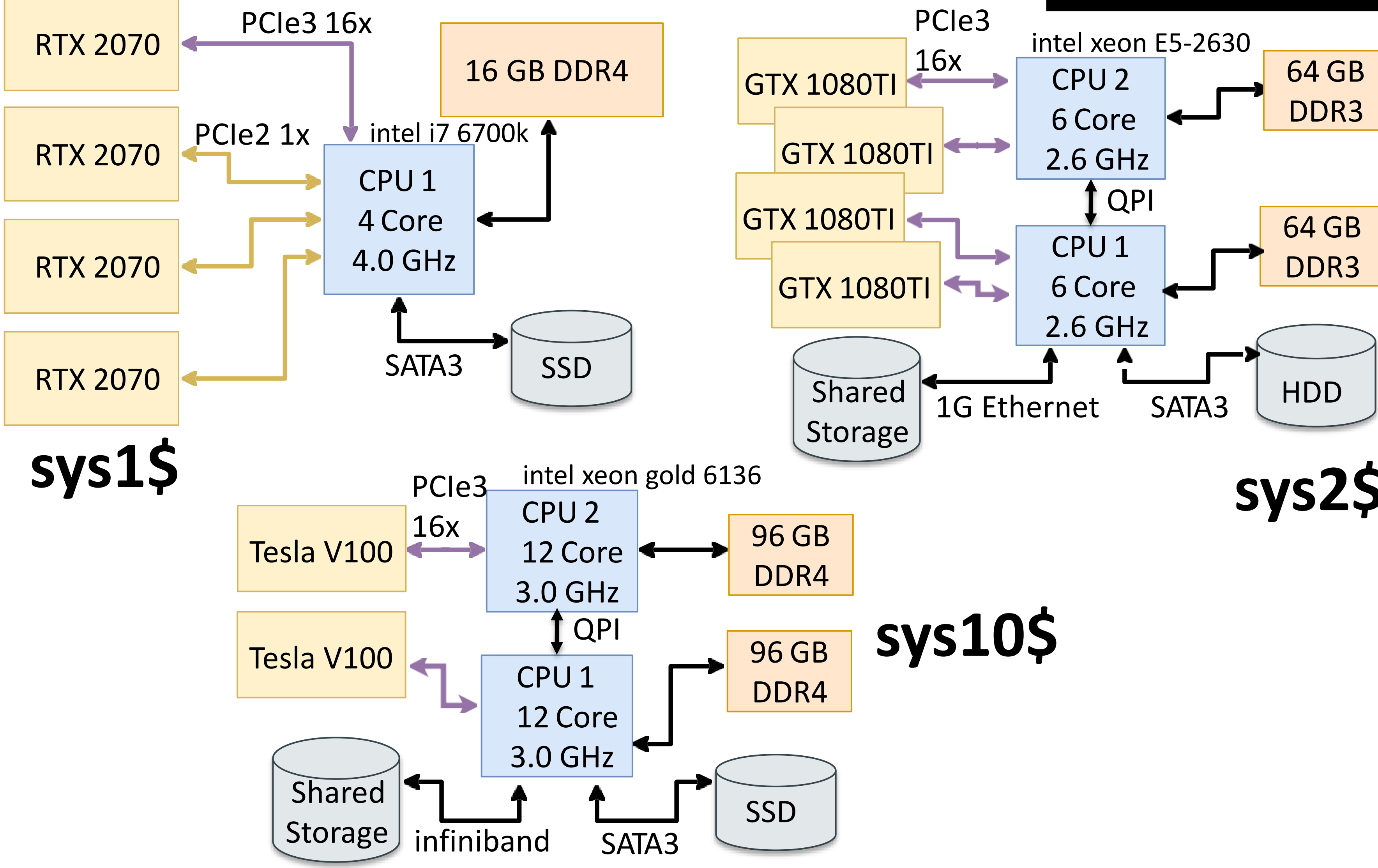
deep speech 2

process of determining the right set of layers is heavily based on trial-&-error

sebastians built *rebelrig*



motherboard = repurposed
cheap crypto-mining rig

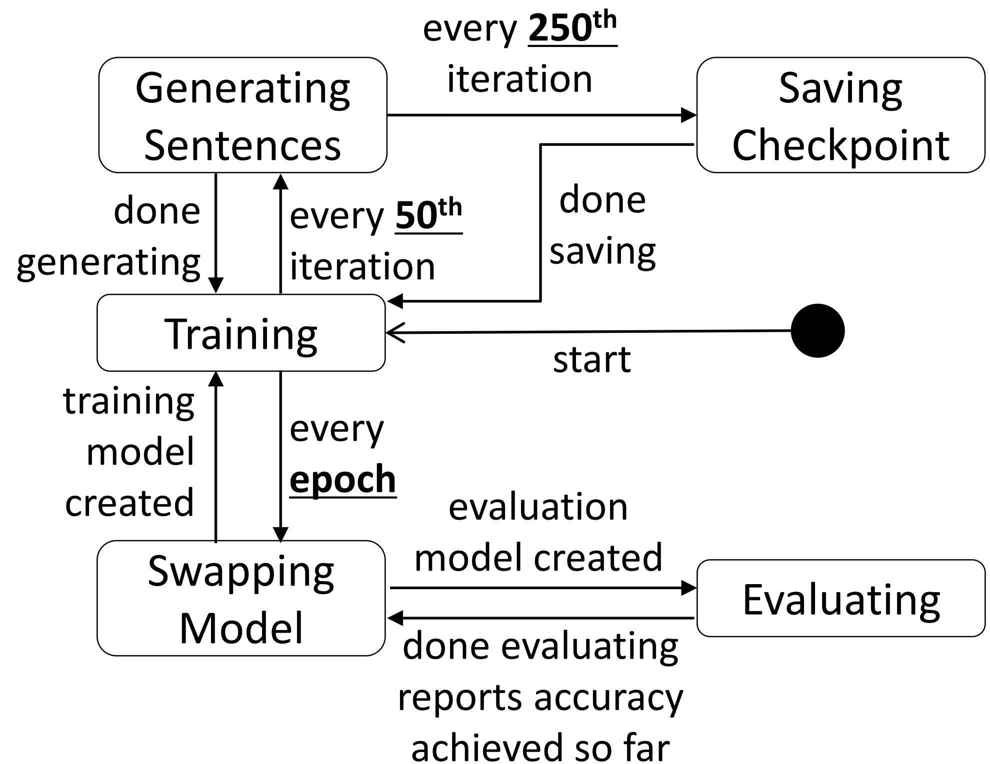

RTX 2070
RTX 2070
RTX 2070
RTX 2070
PCIe3 16x
PCIe2 1x
intel i7 6700k
CPU 1
4 Core
4.0 GHz
16 GB DDR4
SATA3
SSD
sys1$
PCIe3
16x
GTX 1080TI
GTX 1080TI
GTX 1080TI
GTX 1080TI
intel xeon E5-2630
CPU 2
6 Core
2.6 GHz
QPI
CPU 1
6 Core
2.6 GHz
64 GB
DDR3
64 GB
DDR3
Shared
Storage
1G Ethernet
SATA3
HDD
sys2$
PCIe3
16x
intel xeon gold 6136
Tesla V100
Tesla V100
CPU 2
12 Core
3.0 GHz
QPI
CPU 1
12 Core
3.0 GHz
96 GB
DDR4
96 GB
DDR4
Shared
Storage
infiniband
SATA3
SSD
sys10$

experimental setup

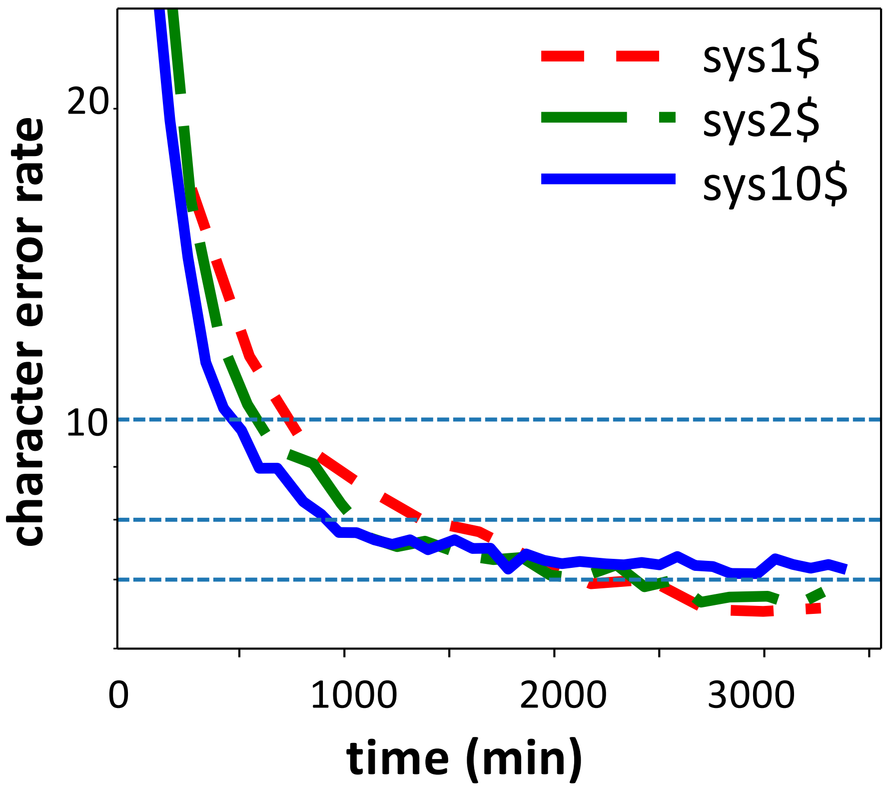
acoustic model implemented using TensorFlow 1.14

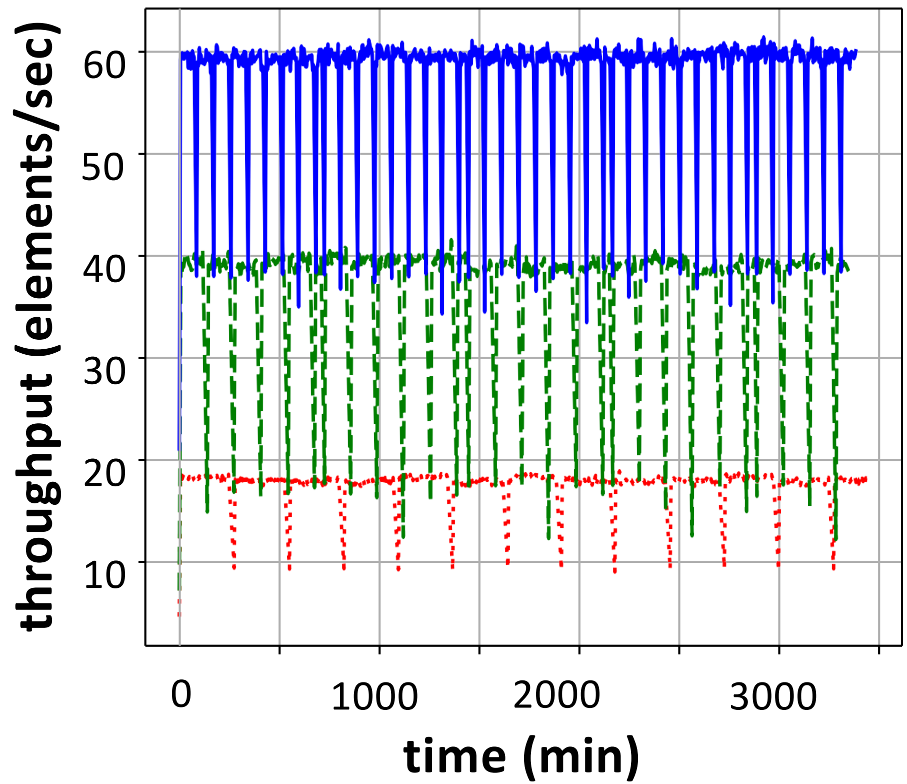
training over three platforms

dataset : LibriSpeech audiobooks
~1000 hours of speech
(both clean & noisy)



results

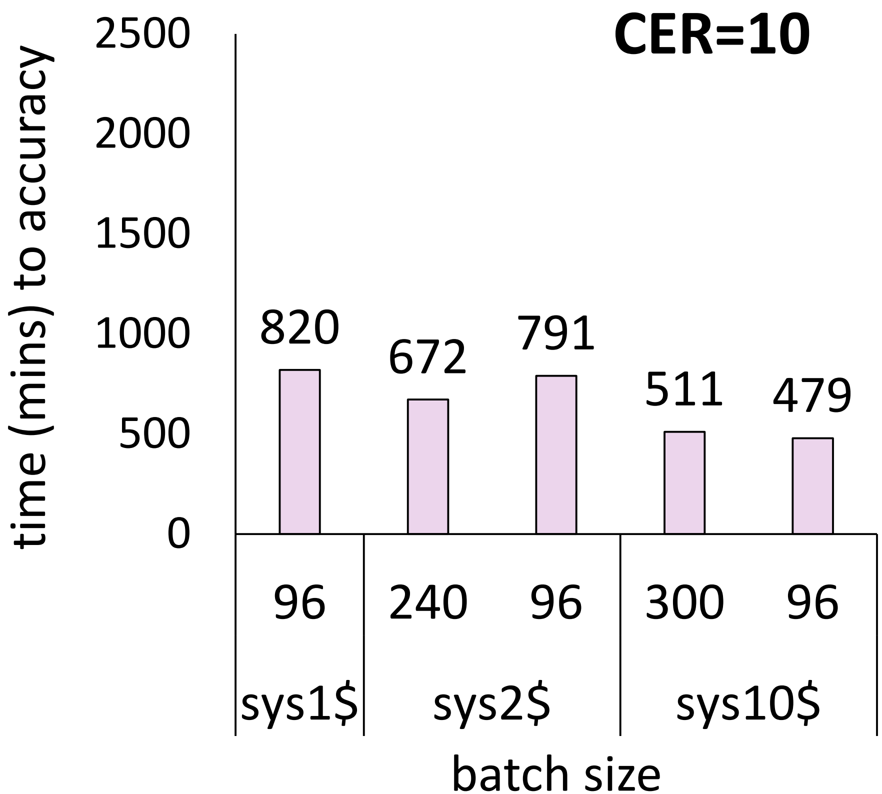


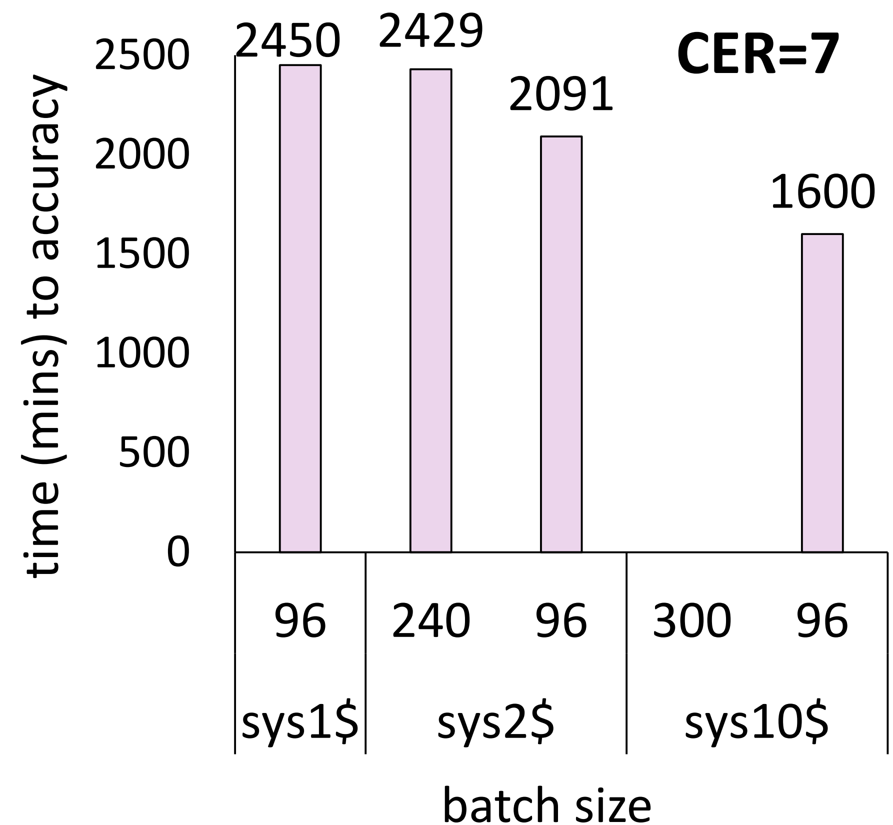


no huge difference in accuracy across platforms

high throughput != faster time-to-accuracy

impact of batch size





larger batch size increases hardware utilization, but may not help with time-to-accuracy

word error rate comparison

platforms	test data	
	clean	**noisy**
sys1\$	17.32	45.04
sys2\$	18.68	48.15
sys10\$	19.45	49.43
	after 2 days 8 hours	
deep speech 2	5.15	12.73
their paper says, this requires 3-6 weeks to execute on a single GPU		

published results in this domain can be very vague when it comes to time-to-accuracy

lessons learned

- very powerful co-processors more and more widely available for machine learning
- but takes a lot to exploit, no free lunch as usual
- need to invest further in improving ML libraries or resource managers for ML on heterogeneous hardware
- on the other hand, low-budget platforms may be good enough for your needs

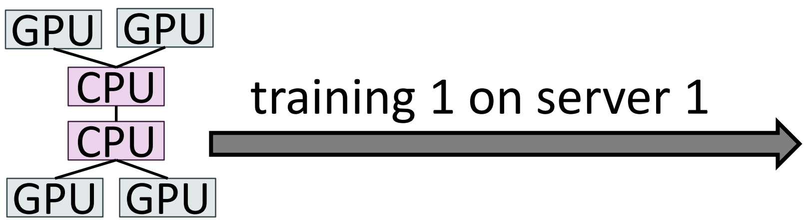
same old challenge, different workload & hardware

agenda

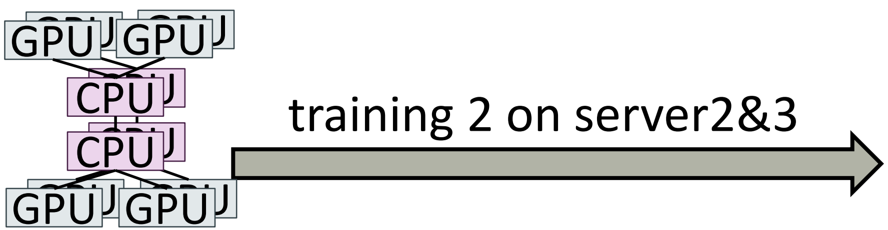
- training speech recognition on co-processors[ADMS2020]
- studying workload co-location
- challenges & opportunities

how to better utilize things?

conventional wisdom

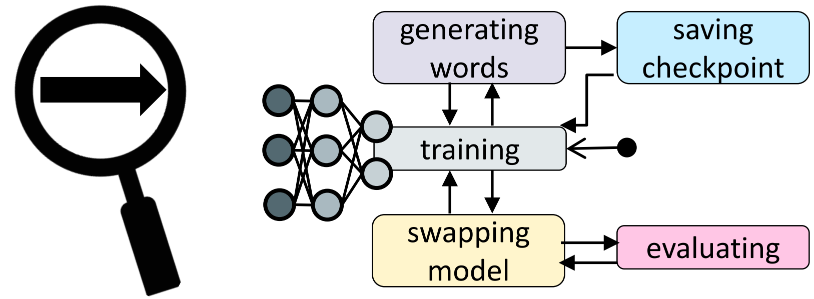


static resource allocation
for the whole for simplicity



strict separation of training
tasks due to fear of interference

resource-aware learning

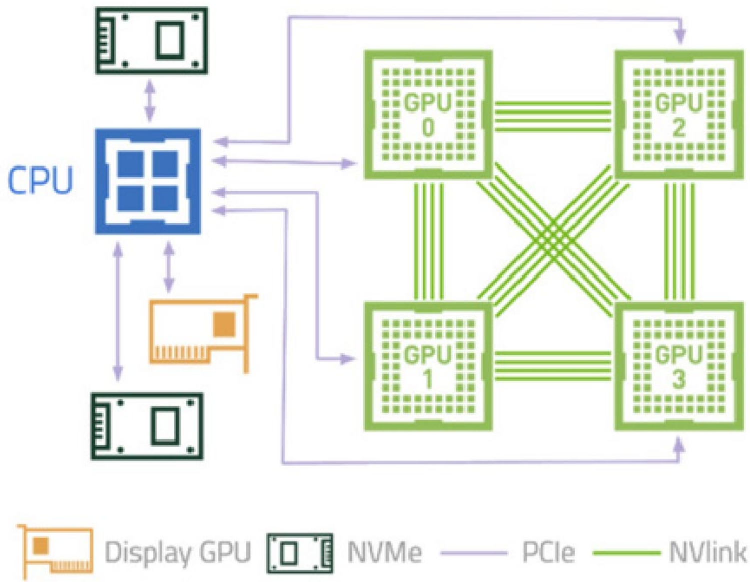


dynamic & fine-grained view

collaborative
work & resource sharing

opportunity for fine-grained co-location

figure from ***NVIDIA DGX Station A100***
System Architecture Technical White Paper



CPU = AMD 7742 – 512 GB RAM
64 physical cores
GPU = NVIDIA A100 – 40 GB RAM
allows ***multi-instance GPU (MIG)***

MSc thesis work of
Stilyan Petrov Paleykov
& Anders Friis Kaas

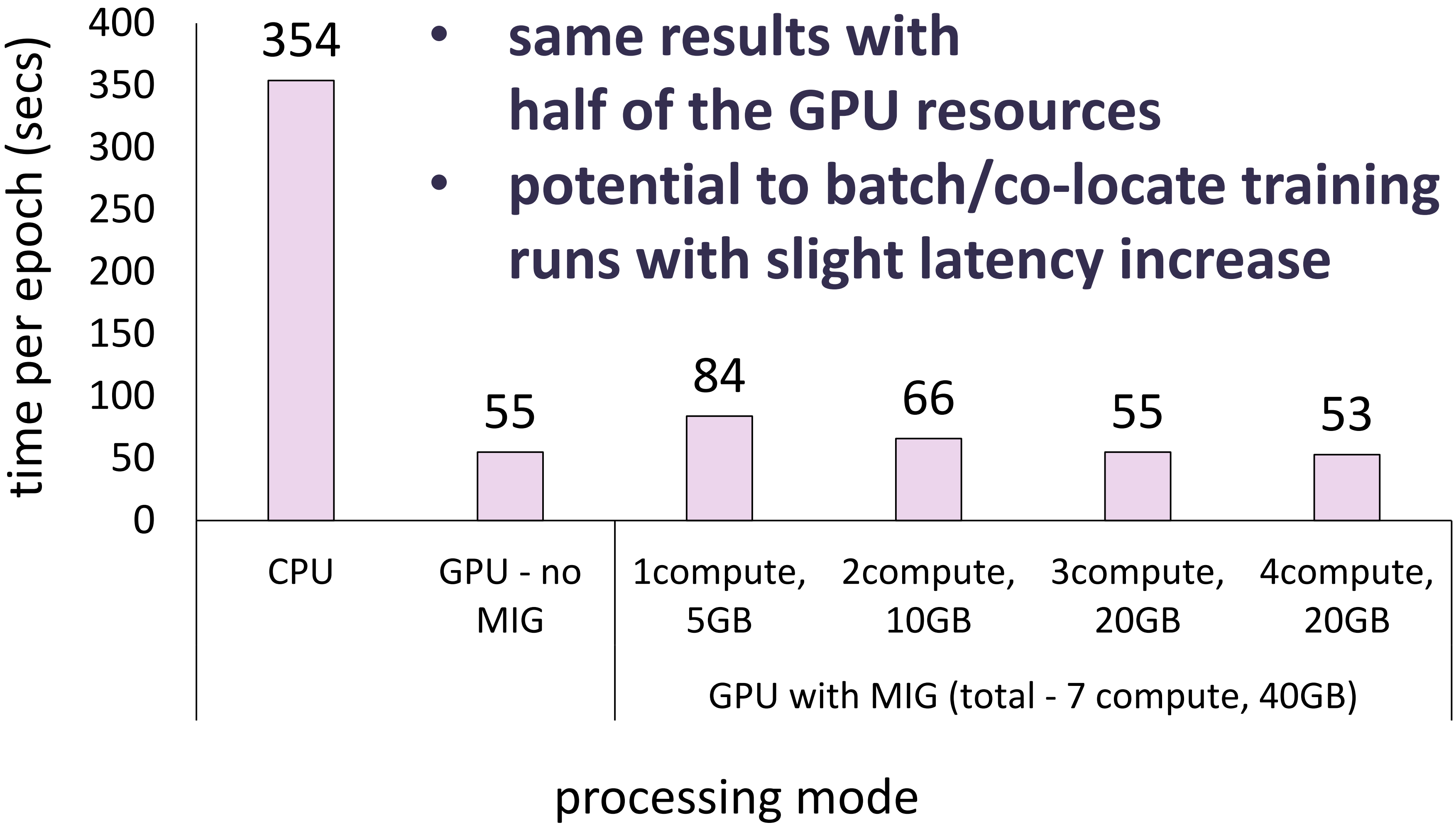
small use case; TensorFlow
training ResNet50 on CIFAR-10 dataset

tools: dcgmi, nvidia-smi, top [, nsight]

initial step: get familiar with MIG & resource monitoring on DGX

impact of multi-instance GPU

- **same results with half of the GPU resources**
- **potential to batch/co-locate training runs with slight latency increase**



challenges & opportunities

thank you!

challenges

- workloads
- experimental duration
- state-of-the-art models changing fast
- measuring computational footprint with many parameters to set
- profiling & co-location granularity

opportunities

- devices that allow finer-grained scheduling & space management
- diversity of applications, hardware, & end-users

***ongoing:* workload characterization on different platforms**

team *RAD*

https://itu-dasyalab.github.io/RAD/

Ties Robroek

Ehsan Yousefzadeh-Asl-Miandoab

Jon Voigt Tøttrup

Robert Bayer

Lottie Greenwood

servers with CPU-GPU co-processors

edge devices

IT system admin

challenges & opportunities

thank you!

challenges

- workloads
- experimental duration
- state-of-the-art models changing fast
- measuring computational footprint with many parameters to set
- profiling & co-location granularity

opportunities

- devices that allow finer-grained scheduling & space management
- diversity of applications, hardware, & end-users

***ongoing:* workload characterization on different platforms**