AI UKRAINE 2019

Як оптимізувати моделі та «залізо» під швидкий інференс на highload проекті

Олег Черній
CTO RIA.com

oleg.cherniy@ria.com

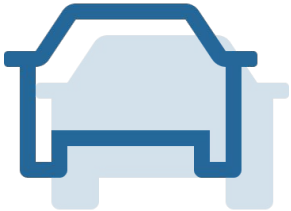
RIA.com auto RIA.com dom RIA.com

Про компанію AUTO.RIA.com

№1 автосайт в Україні, №9 серед світових автосайтів
(за данними similarWeb)

11 000 000

користувачів
в місяць



10 000

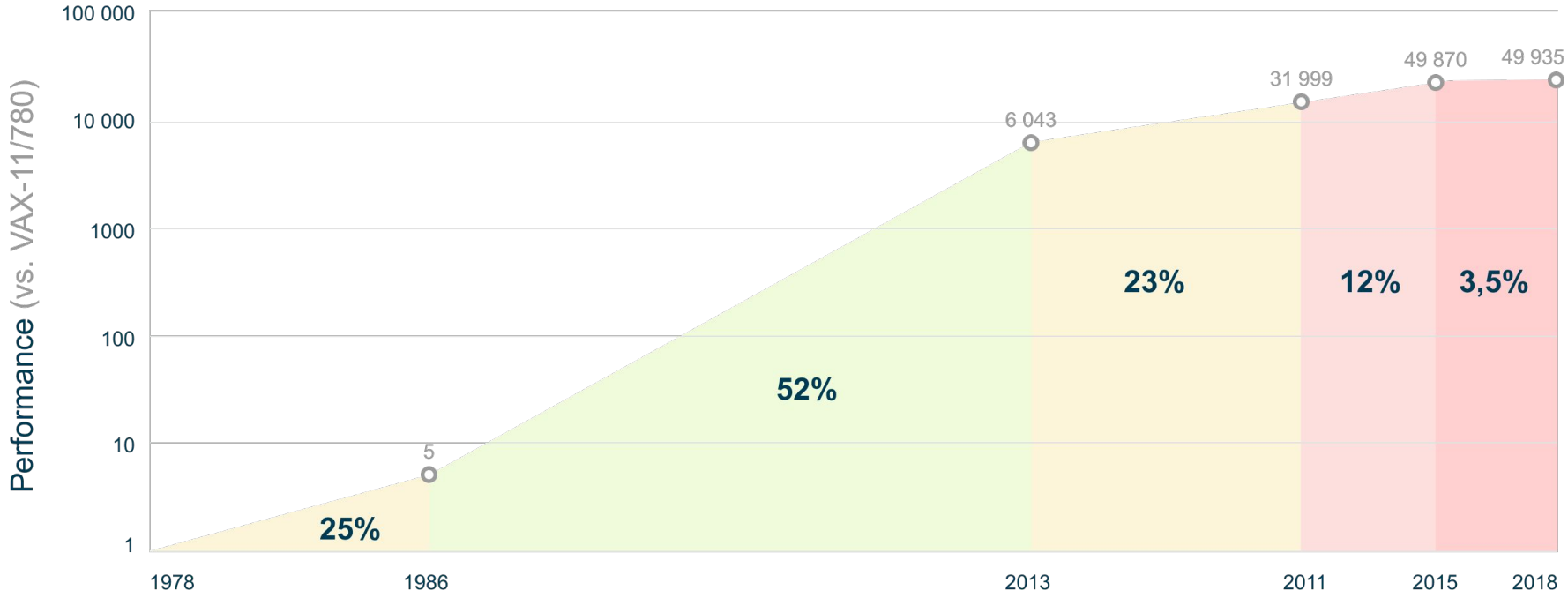
оголошень
щоденно

200 000

фотографій до оголошень
щоденно

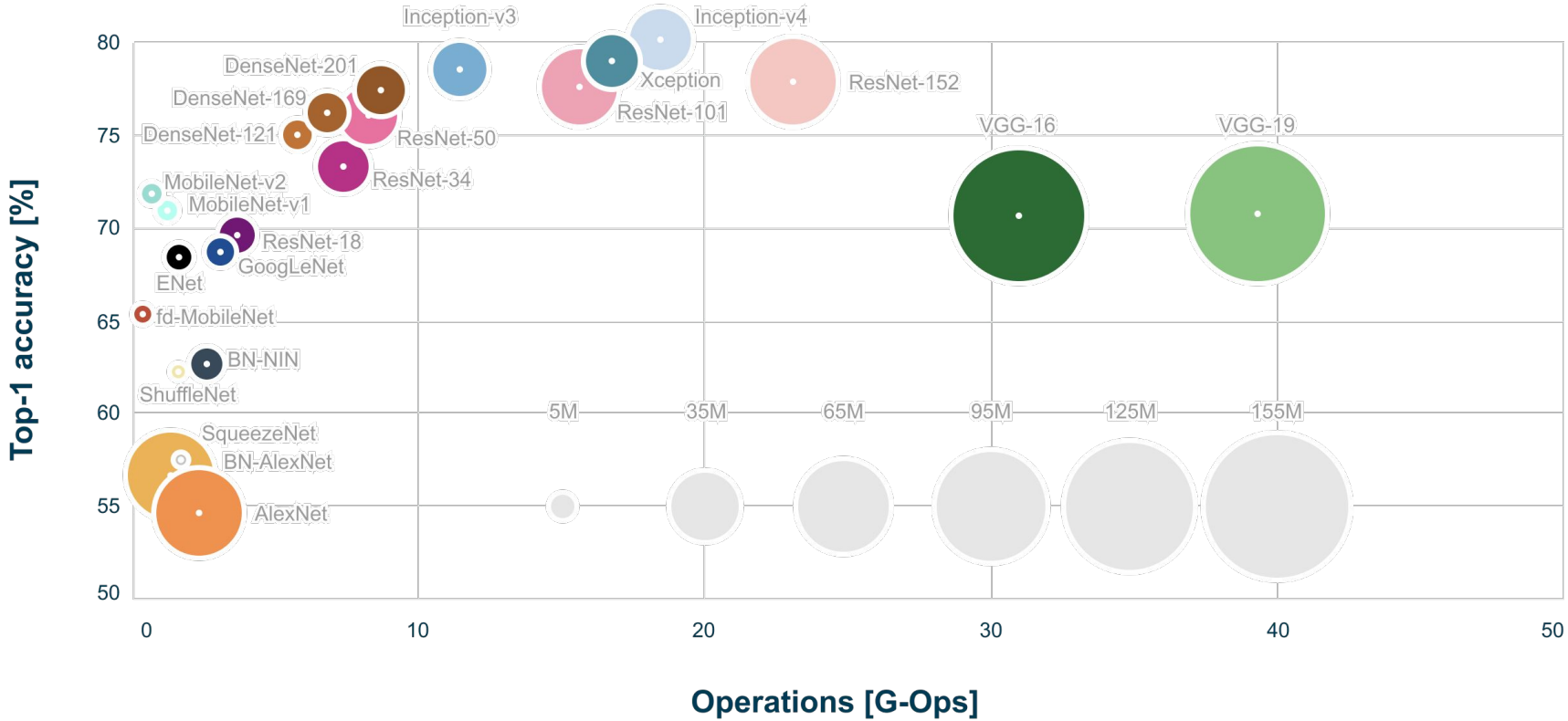
Закон Мура більше не працює

Дженсен Хуанг • CEO NVIDIA



nextbigfuture.com/2019/02/the...

Складність мереж росте



towardsdatascience.com/neura...

Вимоги до обчислень та пам’яті ростуть

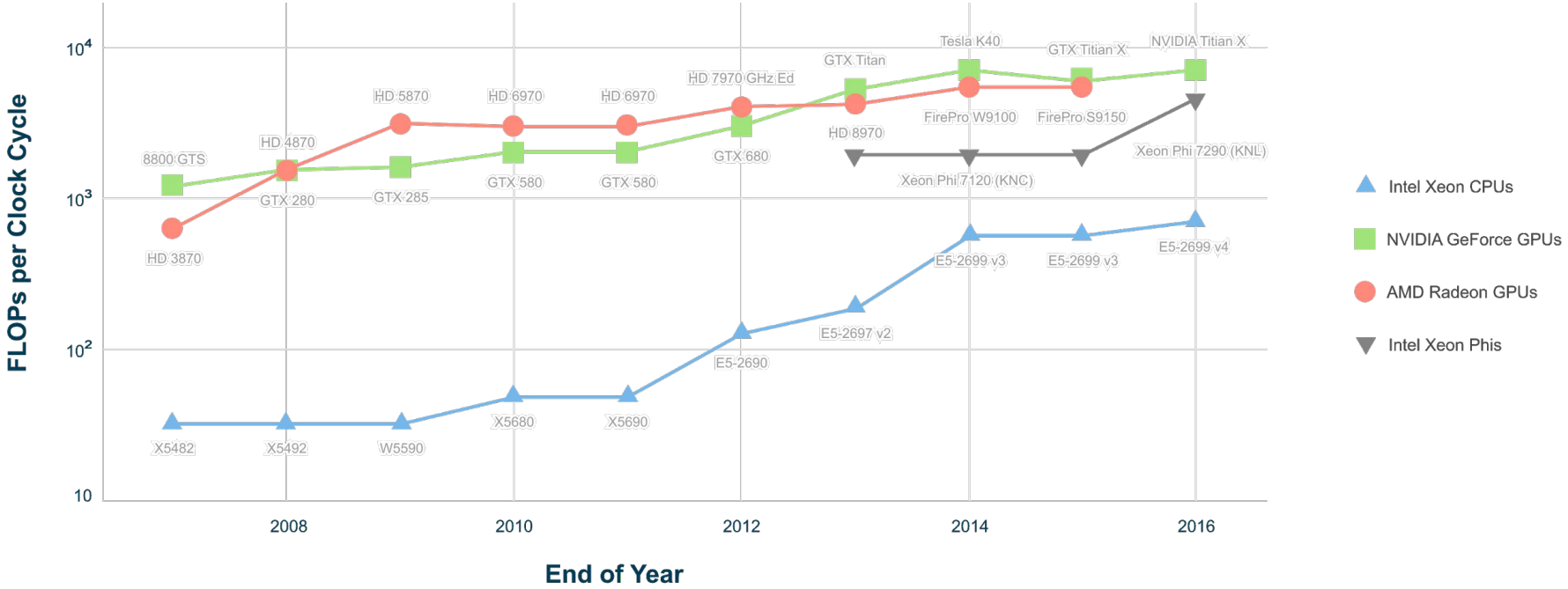
Object Detection Architectures

Model	Input size	Param memory	Feature memory	Flops
rfcn-res50-pascal	600 x 850	122 MB	1 GB	79 GFLOPS
rfcn-res101-pascal	600 x 850	194 MB	2 GB	117 GFLOPS
ssd-pascal-vggvd-300	300 x 300	100 MB	116 MB	31 GFLOPS
ssd-pascal-vggvd-512	512 x 512	104 MB	337 MB	91 GFLOPS
ssd-pascal-mobilenet-ft	300 x 300	22 MB	37 MB	1 GFLOPs
faster-rcnn-vggvd-pascal	600 x 850	523 MB	600 MB	172 GFLOPS

github.com/albanie/convnet-bur...

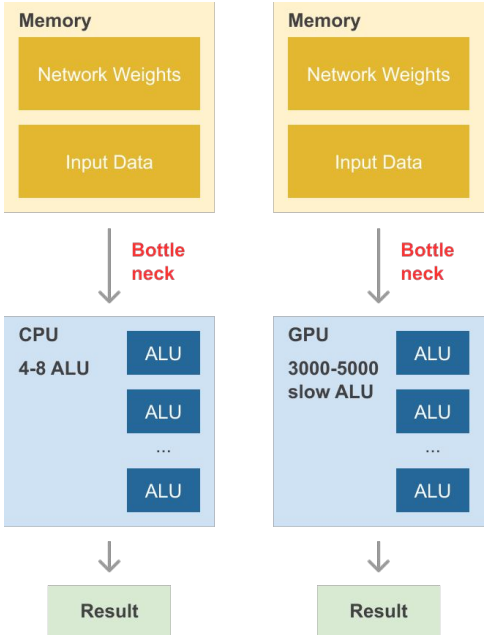
CPU, NVIDIA i AMD, Intel Xeon Phi

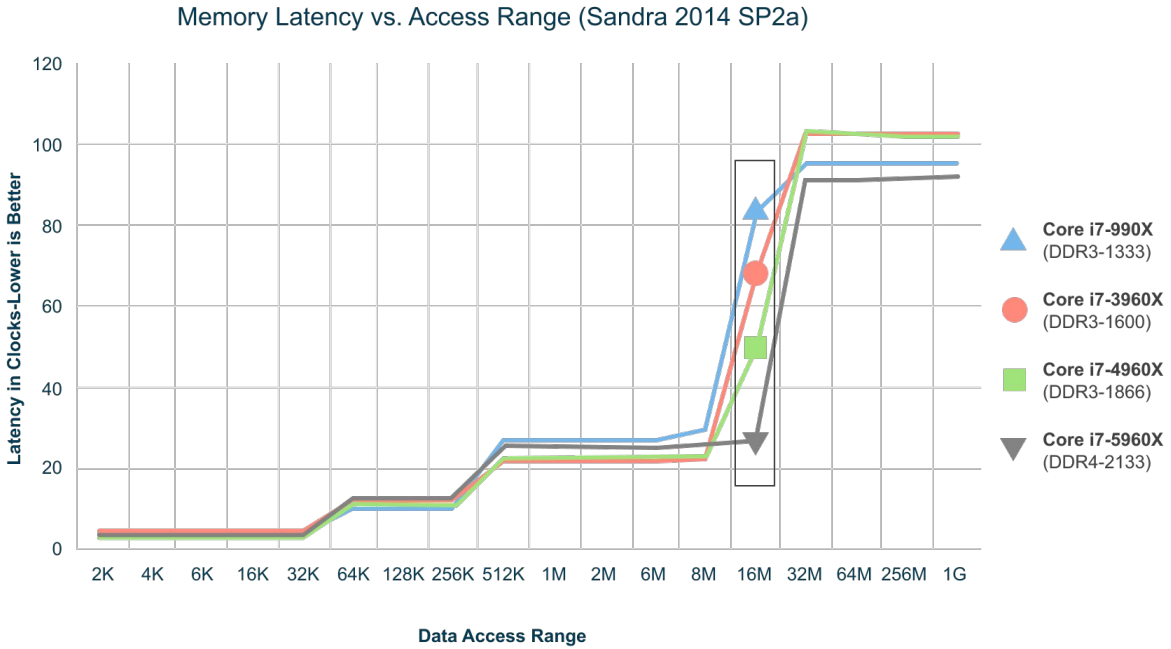
Theoretical Peak Floating Point operations per Clock Cycle, Single Precision



www.karlrupp.net/2013/06/cpu-...

CPU vs GPU

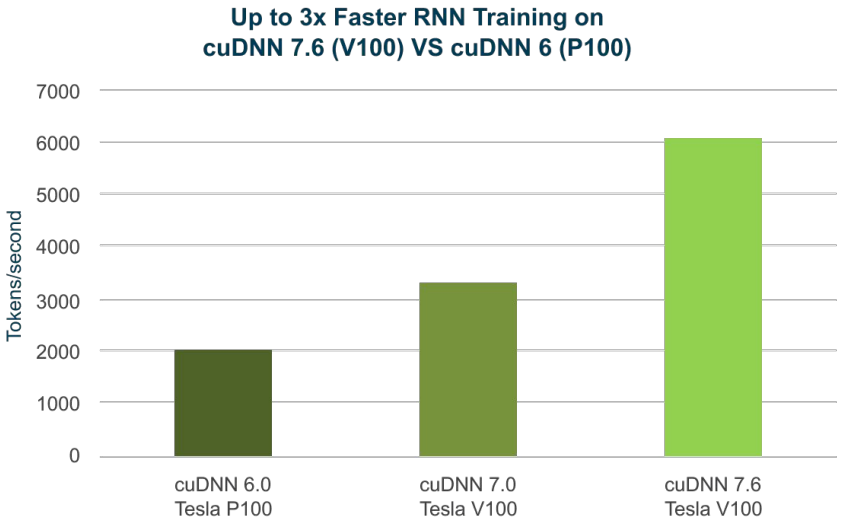




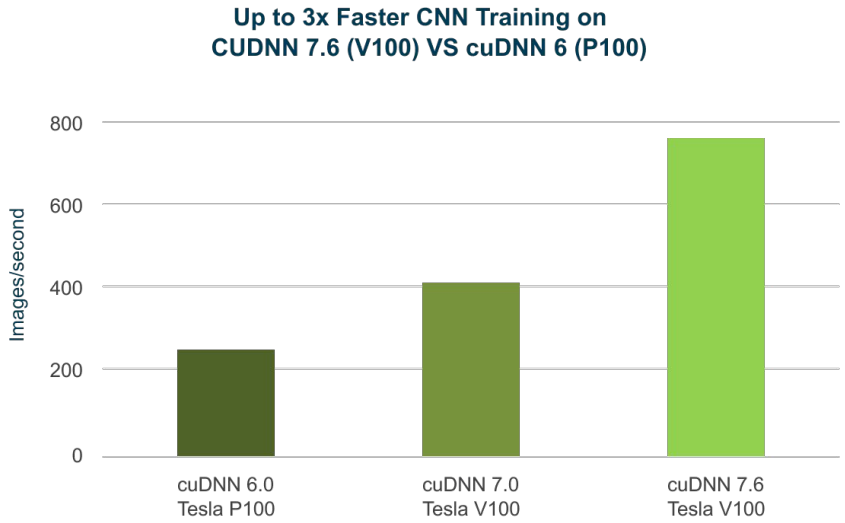
habr.com/ru/post/455353/

anandtech.com/show/842...

GPU NVIDIA - лідер



TensorFlow performance tokens/sec), Tesla P100+ cuDNN 6 (FP32) on 17 12 NGC container, Tesla V100+ CUDNN 70 (Mixed) on 18.02 NGC container, Tesla V100 + cuDNN 7.6 (Mixed) on 19.05 NGC container, OpenSeq2Seq (GNMT), Batch Size: 64

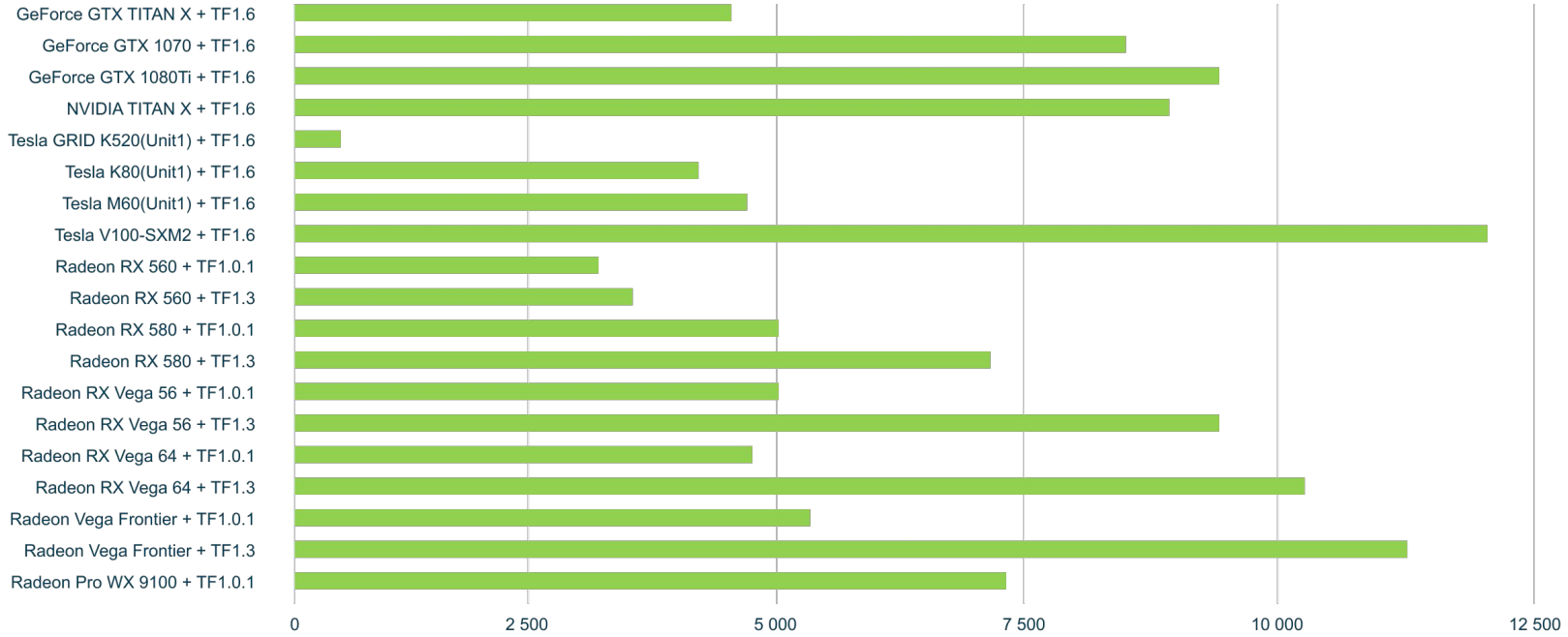


TensorFlow performance (images/sec), Tesla P100+ cuDNN 6 (FP32) on 17.12 NGC container, Tesla V100+ cuDNN 7.0 (Mixed) on 18.02 NGC container, Telsa V100 + cuDNN 7.6 (Mixed) on 19.05 NGC container, ResNet-50, Batch Size: 128

developer.nvidia.com/cudnn

GPU AMD - потужна, проте недостатня підтримка

Benchmark CIFAR10 on TensorFlow



http://blog.gpueater.com/en/201...

AMD INSTINCT - майбутнє поряд

AMD Radeon Instinct™ MI60

Lithography	TSMC 7nm FinFET
Кол-во потоковых процессоров	4096
Peak Single Precision (FP32) Performance	14.7 TFLOPs
Memory Size	HBM2 32 GB

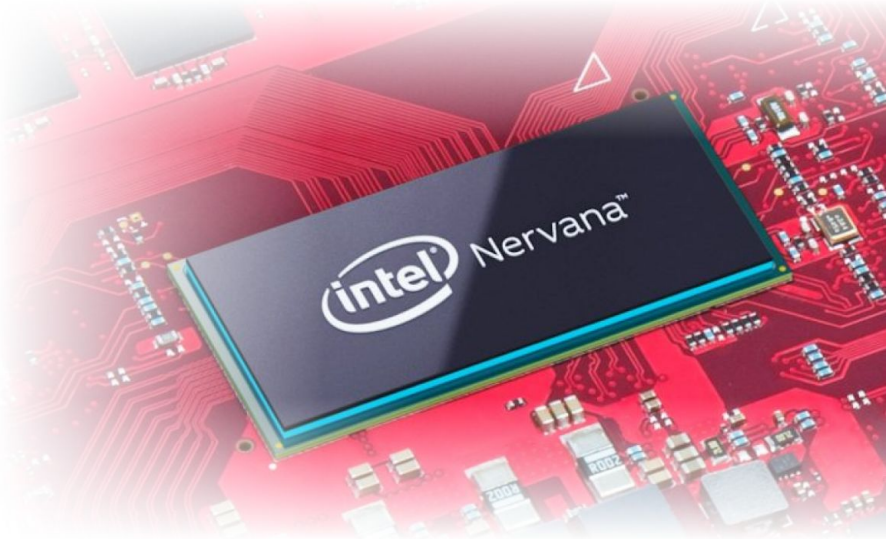
amd.com/ru/graphics/servers-...

CPU INTEL - доганяє потяг

Intel® Xeon Phi™ Processor 7290F

Intel® Xeon® E5-2699A v4

Intel® Nervana® ?



venturebeat.com/2019/01/07/int...

TPU Huawei - китай не відстає!

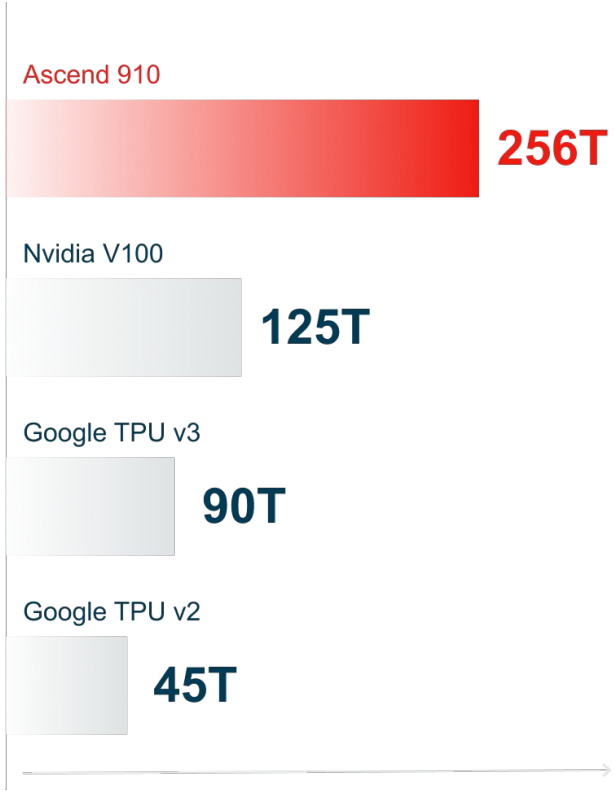
Ascend 910

Greatest computing density in a single chip

Ascend-Max
Architecture: Da Vinci

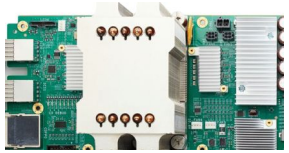
Half-precision (FP16):256 TeraFLOOPS
Ingeter-Precision (INT8): 512 TeraOPS
128 Channel FHD Video Decoder - H.264/265

Max Power: 350W
7nm
2019 Q2



habr.com/ru/post/455353/

Google TPU - зрілий і потужний



TPUv2 Chip

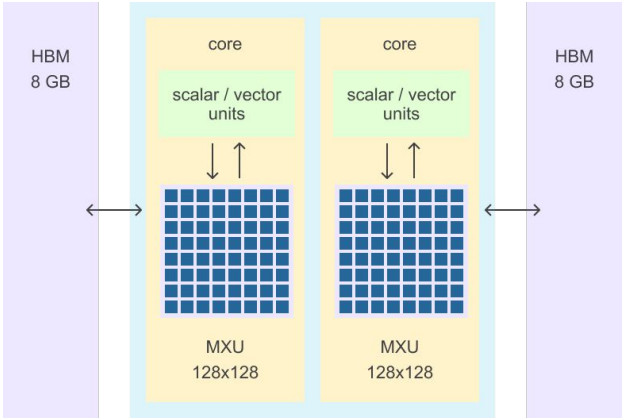
16 GB of HBM

600 GB/s mem BW

Scalar/vector units: 32b float

MXU: 32b float accumulation but reduced precision for multipliers

45 TFLOPS



TPU Version	On-demand price	Preemptible price
Cloud TPU v2	**$4.95** / TPU hour	**$1.485** / TPU hour
Cloud TPU v3	**$8.80** / TPU hour	**$2.64** / TPU hour

cloud.google.com/tpu/pricing?h...

Amazon AWS і Microsoft Azure: Надійні, але дорогі

Amazon SageMaker

Инстансы с графическим процессором – текущее поколение

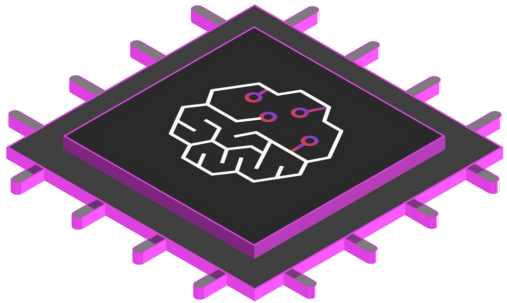
ml.p2.xlarge	**1 361** USD
ml.p2.8xlarge	**10 886** USD
ml.p2.16xlarge	**21 773** USD
ml.p3.2xlarge	**4 627** USD
ml.p3.8xlarge	**18 508** USD
ml.p3.16xlarge	**37 016** USD

Самий "дешевий" (річний бюджет): ml.p2.xlarge (1 x K80 12G) 1.361*24*30*12 = **11 759**
Для V100 (річний бюджет): ml.p3.2xlarge (1 x V100 16G) 4.627*24*30*12 = **39 977**

aws.amazon.com/ru/sagemake...

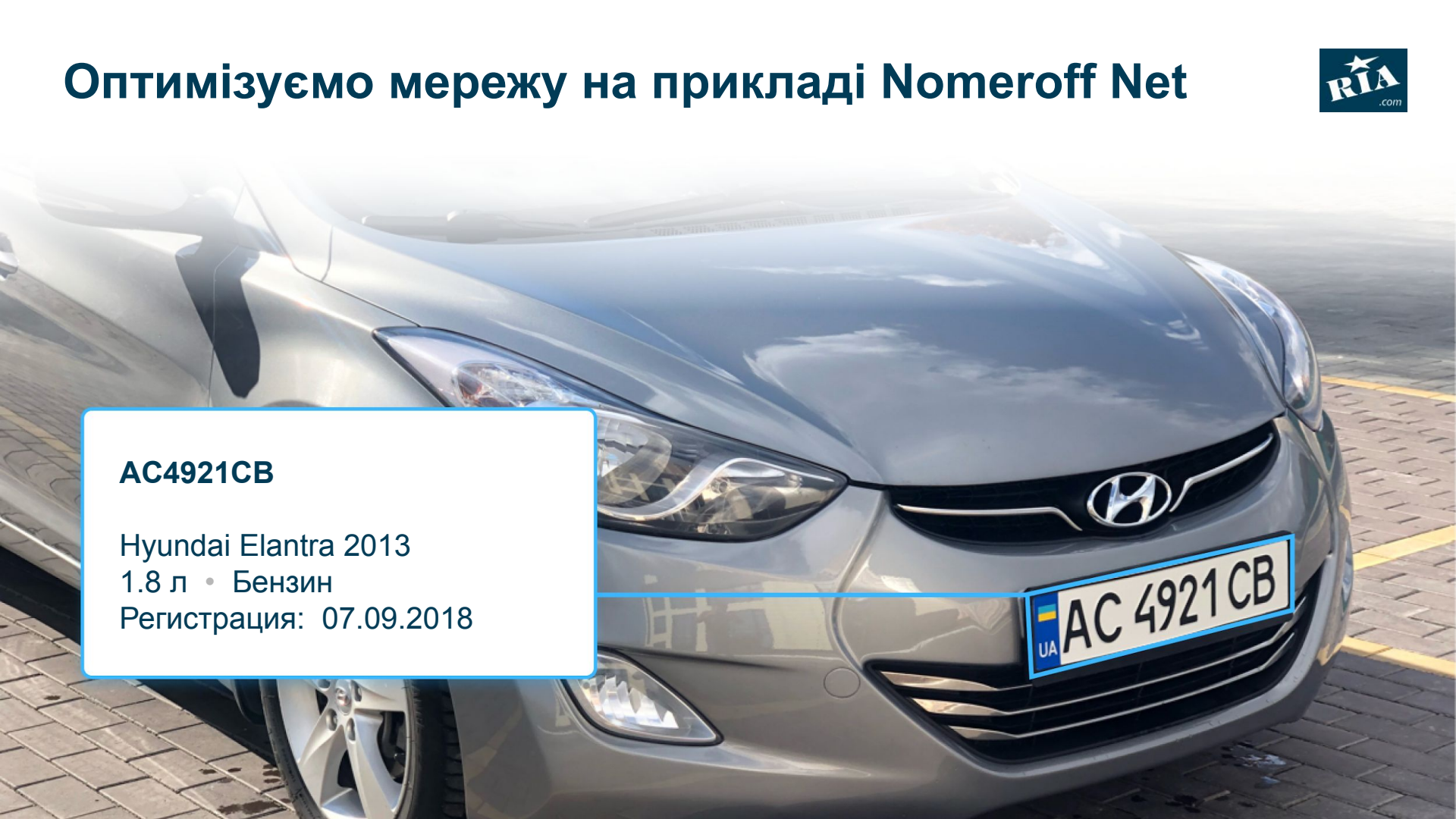
Детальний опис

Amazon Elastic Inference



aws.amazon.com/ru/machine-le...

Оптимізуємо мережу на прикладі Nomeroff Net



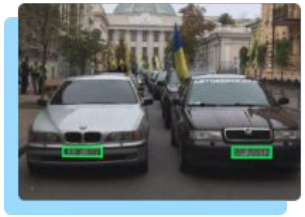
Nomeroff Net
nomeroff.net.ua

RIA .com

Instance segmentation
Визначення областей з номером

Classification
Якого типу номер, чи замальований,
скільки стрічок у тексті

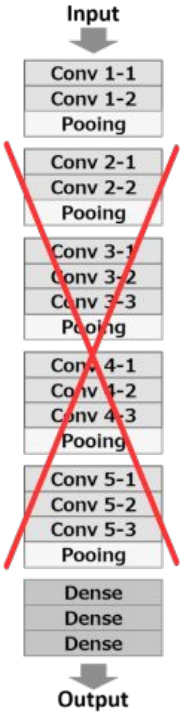
OCR, Постобробка



Архітектура класифікатора

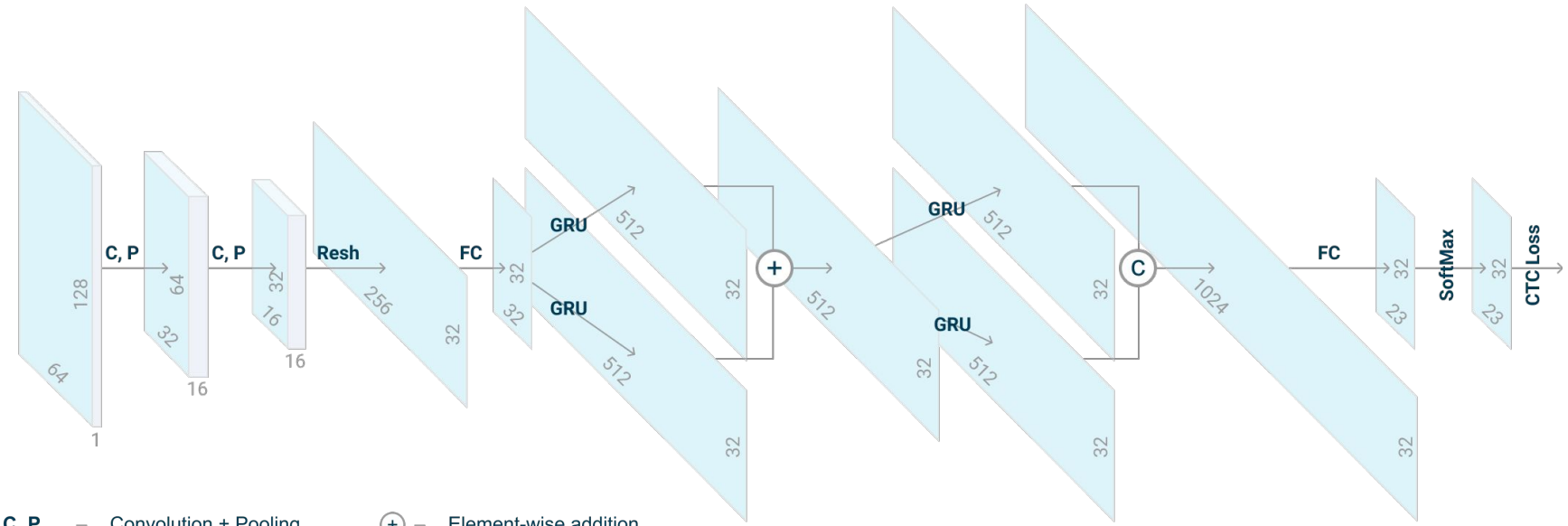
Замість **VGG16** створили кастомну архітектуру:

```
def create_simple_conv(self, inp):
  conv_base = layers.Conv2D(32, (3, 3), activation='relu')(inp)
  conv_base = layers.MaxPooling2D((2, 2))(conv_base)

  conv_base = layers.Conv2D(64, (3, 3), activation='relu')(conv_base)
  conv_base = layers.MaxPooling2D((2, 2))(conv_base)

  conv_base = layers.Conv2D(128, (3, 3), activation='relu')(conv_base)
  conv_base = layers.MaxPooling2D((2, 2))(conv_base)

  conv_base = layers.Conv2D(128, (3, 3), activation='relu')(conv_base)
  conv_base = layers.MaxPooling2D((2, 2))(conv_base)

  return conv_base
```



github.com/ria-com/nomeroff-ne...

Архітектура OCR



C, P – Convolution + Pooling
Resh – Reshape
FC – Fully Connected
GRU – Gated recurrent unit

(+) – Element-wise addition
(C) – Concentration

hackernoon.com/latest-deep-le...

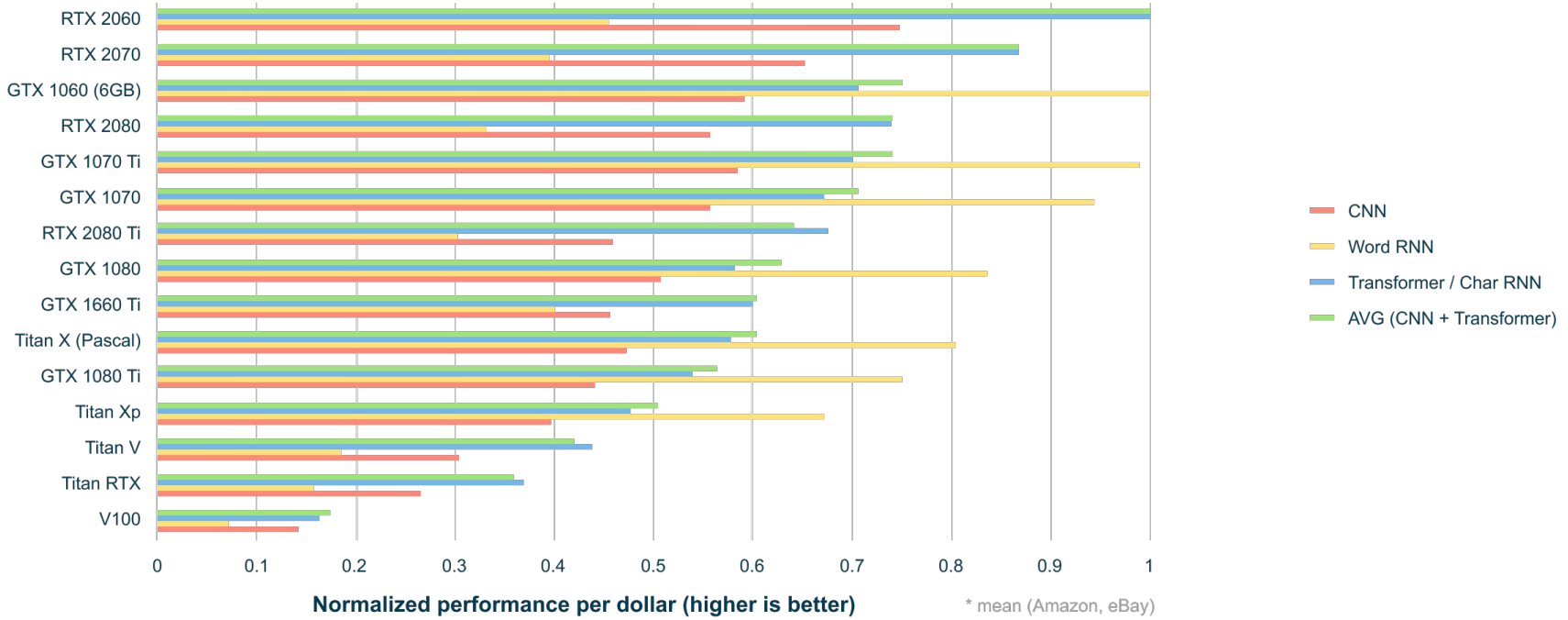
Оптимізуємо залізо: Server vs Desktop

У більшості задач немає сенсу суттєво переплачувати за серверне обладнання

neurohive.io/ru/osnovy-data-sci...

Оптимізуємо залізо: вибір GPU

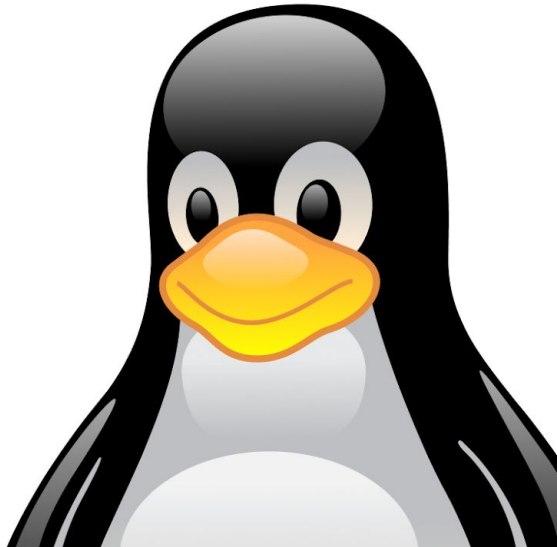
Performance per Dollar*



timdettmers.com/2019/04/03/w...

Оптимізуємо OS

- Linux зі свіжим ядром (kernel >= 5.1), напр. Ubuntu, Fedora
- Свіжі драйвера (nvidia >= 430.26, cuda >= 10.1, cuda-cudnn >= 7.6)
- Політики для ядер:
  cat /sys/devices/system/cpu/cpufreq/policy*/scaling_available_governors
  Перемикаємо в perfomance:
  echo "performance" | tee
  /sys/devices/system/cpu/cpufreq/policy*/scaling_governor
- Якщо платформа двосокетна ставимо nvidia-persistenced
- Моніторим стан GPU:
  pip3 install gpustat



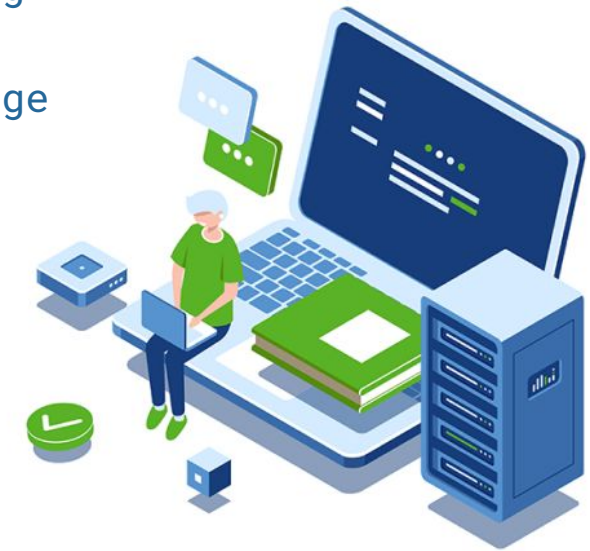
Оптимізуємо софт

- Свіжий python >= 3.7
- Збирайте самостійно останню версію tensorflow >= 1.15.0rc0 з оптимізаціями для CPU і GPU

```
bazel build --config=opt --config=cuda --config=mkl
--copt=-march=native
//tensorflow/tools/pip_package:build_pip_package
```

- Для доступу до GPU у docker-контейнері ставим nvidia-docker
- XLA: Optimizing Compiler for TensorFlow

```
$ TF_XLA_FLAGS=--tf_xla_auto_jit=2
path/to/your/tf/program
```



Оптимізуємо софт

Апаратно-залежна оптимізація інференсу:

- **NVIDIA TensorRT**

  youtube.com/watch?v=NYUoew...

- **Intel OpenVINO**

  (6th-8th Generation Intel® Core™, Intel® Xeon® v5, v6 family, Intel® Movidius™ Neural Compute Stick, Intel® Neural Compute Stick 2, Intel® Vision Accelerator Design with Intel® Movidius™ VPUs)

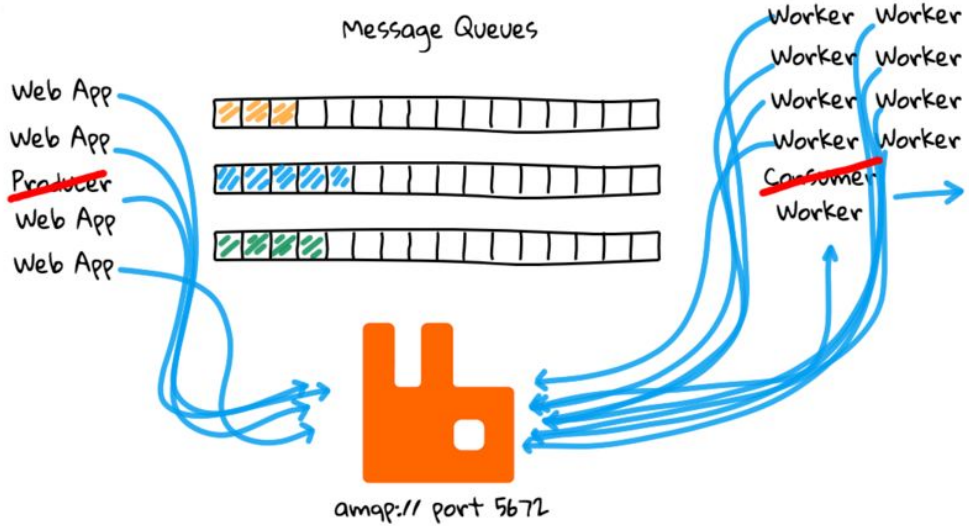
  dlology.com/blog/how-to-run-ke...

Оптимізуємо модель

- Ресайз фото перед передачою на інференс працює швидко через cv2.resize
- Спрощуємо архітектуру, це завжди приводить до втрати якості
- Використовуємо оптимізовані слої на TF (для NVIDIA GRU -> CuDNNGRU, ...)
- Керуємо розпралелюванням
  CUDA_VISIBLE_DEVICES=0 taskset -c 0-7 python3 ./test
- Заморожуємо індекси і конвертуємо в pb (model size)

Оптимізуємо навантаження на GPU

- Розпаралелювання в рамках 1 GPU. Запускайте декілька інстансів для інфенсу.
- По можливості не робіть інференс у режимі online, набагато дешевше ставити задачі у чергу та завантажувати систему рівномірно (організовуйте чергу за допомогою Rabbit MQ або Apache Kafka).



Nomeroff Net до і після оптимізації

- Оптимізація тренування OCR

  github.com/ApelSYN/aiukraine...

- Бенчмарк інференсу Nomeroff Net для CPU та GPU

  github.com/ria-com/nomeroff-n...

Корисні посилання

- Nomeroff Net Online форма розпізнавання номерів

  nomeroff.net.ua/onlinedemo.html

- Аппарате прискорення глибоких нейромереж: GPU, FPGA, ASIC, TPU, VPU, IPU, DPU, NPU, RPU, NNP та інші літери

  habr.com/ru/post/455353/

Час для запитань