

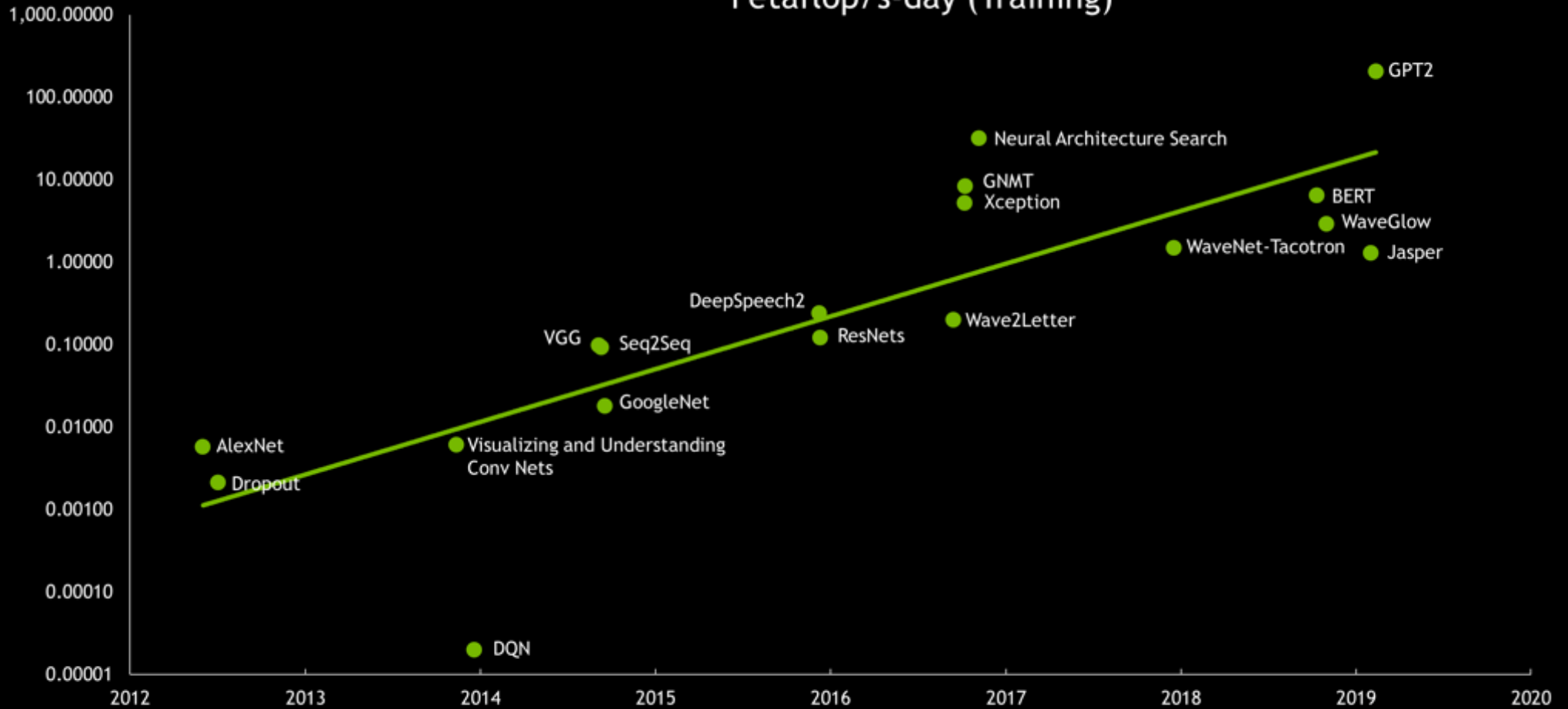


ПРИКЛАДНОЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Антон Джораев

ВЗРЫВНОЙ РОСТ СЛОЖНОСТИ НЕЙРОСЕТЕЙ

Petaflop/s-day (Training)



ЧТО ТАКОЕ РАЗГОВОРНЫЙ ИИ?

Диалоговый, не транзакционный

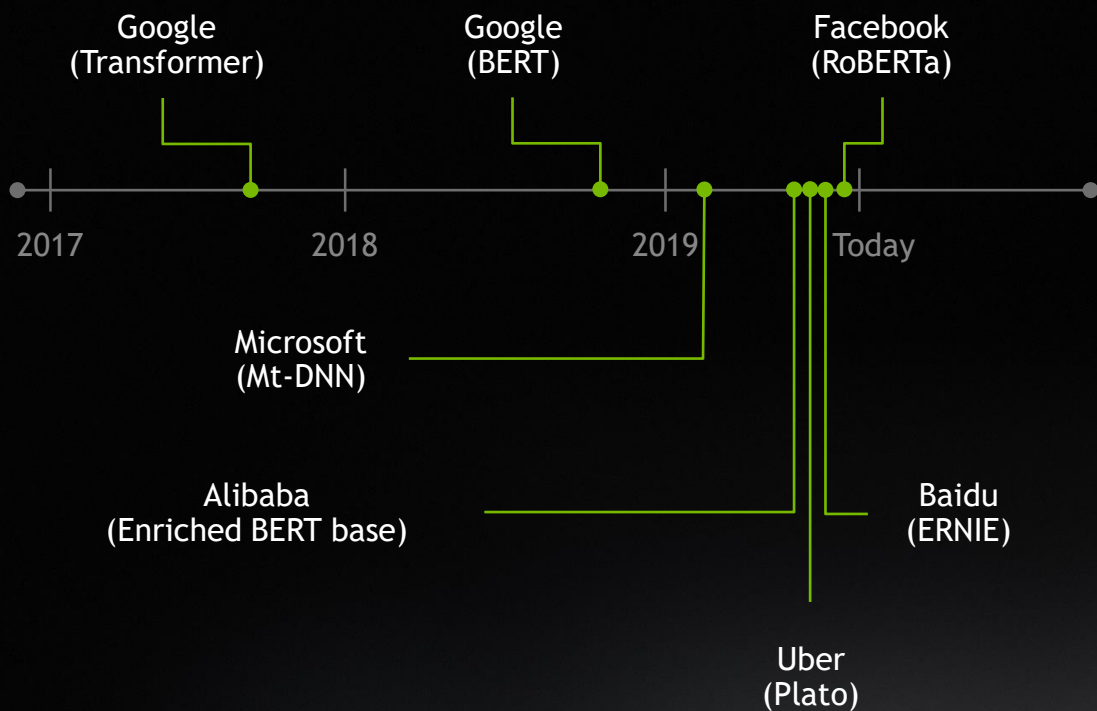
Используется контекст

Ответы должны быть моментальными

Модели становятся очень большими



ГОНКА ЗА РАЗГОВОРНЫМ ИИ



Лидеры рейтинга GLUE

Rank	Name	Model	Score
1	Google	ALBERT (ensemble)	89.4
2	Microsoft	ADV-RoBERTa (ensemble)	88.8
3	Facebook AI	RoBERTa	88.5
6	GLUE Human Baselines	GLUE Human Baselines	87.1

<https://gluebenchmark.com/leaderboard>

NVIDIA DGX-2

Самая мощная в мире DL-система для самых сложных задач
глубокого обучения

- Первая в мире 2 PFLOPS система
- 16 Tesla V100 32GB GPUs с общим интерконнектом
- NVSwitch: 2.4 TB/s пропускная способность
- 0.5 TB унифицированной памяти GPU



www.nvidia.ru/dgx

NVIDIA DGX SUPERPOD

Лидерство в ИИ невозможно без лидерства в инфраструктуре

Автопилот | Речевой ИИ | Здравоохранение | Графика | HPC

Испытательный стенда для масштабируемых HPC систем

- 9.4 PF Linpack | ~200 PF AI | #22 в списке Top500
- <2 минут на обучение RN-50

Модульная архитектура DGX SuperPOD

- Собран и запущен за 3 недели
- Оптимизации в Compute, Networking, Storage & Software

Интегрированный софтверный стек

- Свободно доступен через NGC

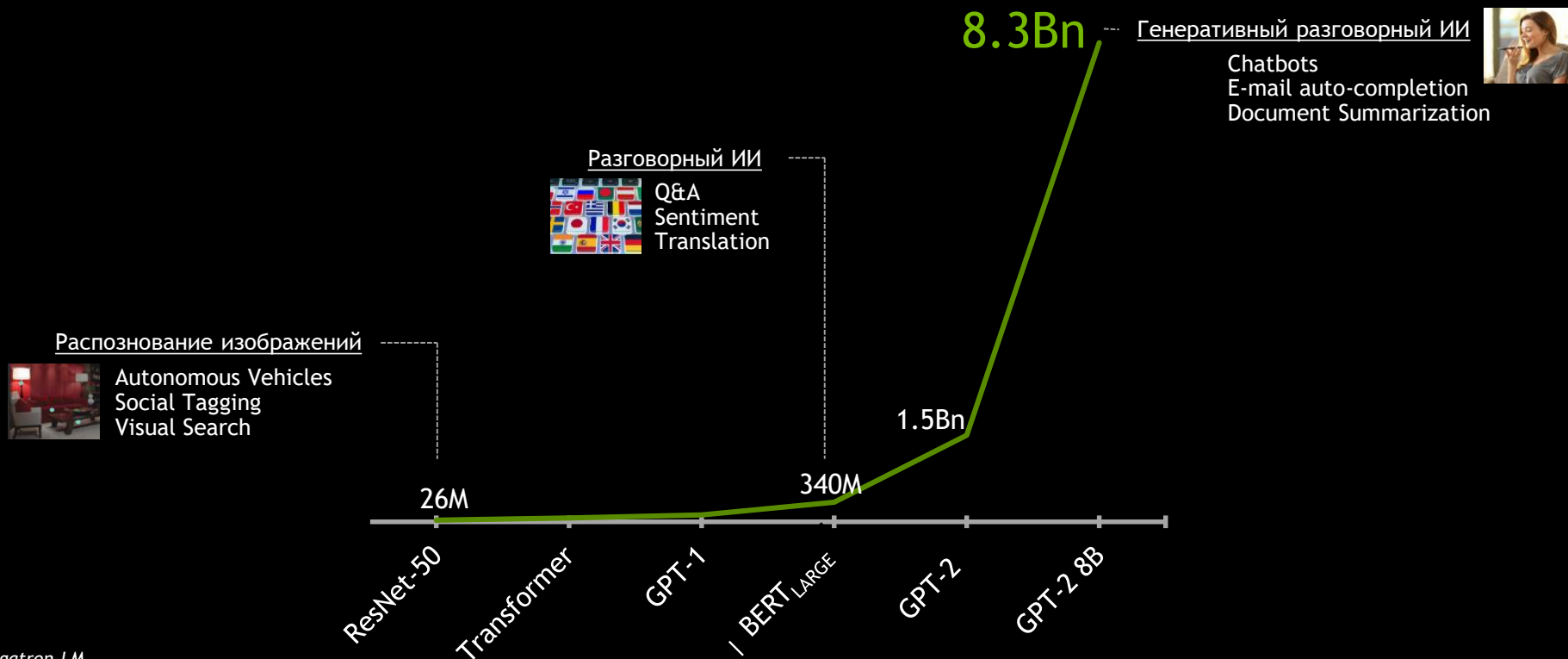


- 96 DGX-2H
- 10 Mellanox EDR IB per node
- 1,536 V100 Tensor Core GPUs
- 1 megawatt of power

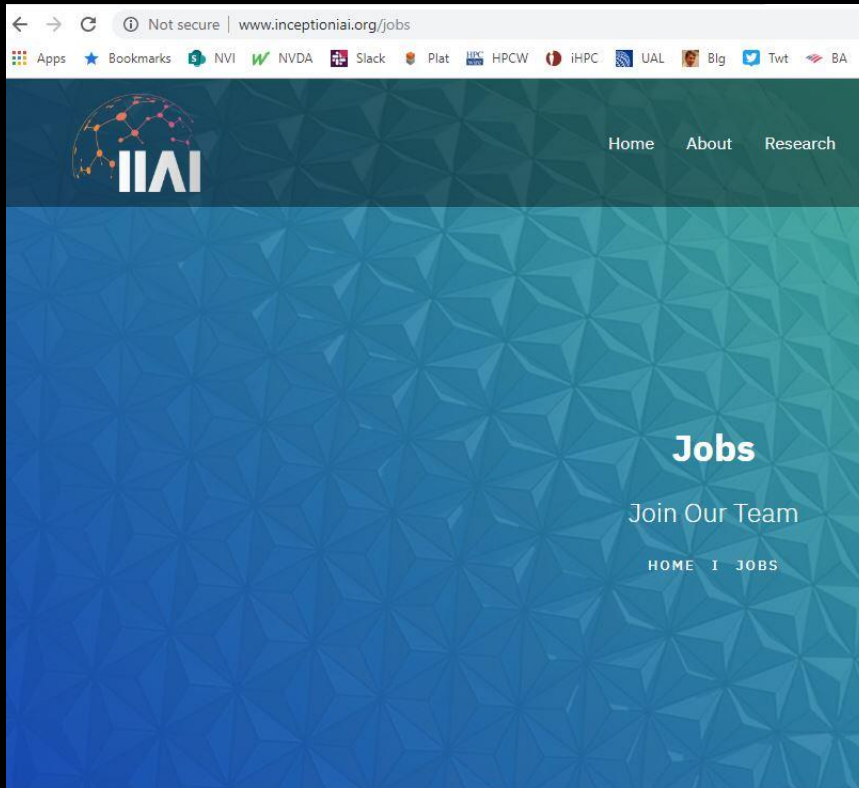
САМЫЕ СОВРЕМЕННЫЕ РЕЧЕВЫЕ МОДЕЛИ

Разработка разговорного ИИ требует гигантских объемов вычислений

Количество параметров нейросети



НАЙМ DATA SCIENTIST'ОВ - ЧТО ВАЖНО?



in the country.

Location

Our office is situated in the central business area of Abu Dhabi, within walking distance of shopping malls, various restaurants, the Cleveland Clinic that offers the world's best healthcare and multiple sea view luxury residential blocks. Being the home to 175 nationalities, the UAE provides an open-minded and hospitable living and working environment to all expats. From clothes to cuisines, in the UAE, you can always find the humblest and the most extravagant.

Minimum requirements

- PhD degree in Computer Science, Data Science, Mathematics, Statistics or related streams.
- Proven experience in Computer Vision, Machine/Deep Learning, Natural Language Processing, Data Mining, Multimedia, Medical Imaging, Bioinformatics or related fields.
- Strong publication record at conferences and in journals, such as TPAMI, IJCV, NIPS, ICML, ICCV, CVPR, ECCV, ICLR, KDD, ACL and MICCAI.

Benefits

- A highly motivating position that will drive you to broaden your horizon and learn from senior peers.
- Freedom of academic research without the pressure of grants and teaching in academia.
- Huge amounts of data across the country that can help you fulfill and fast-track your research into real-world scenarios.
- A GPU farm of **100 NVIDIA DGX-1** servers and other GPU machines (**1000 Tesla V100 GPUs**).
- An enjoyable modern working environment with a spectacular sea view.
- World-leading standard of healthcare, education and hospitality industry.
- All year round sunshine and amazing beaches.
- Extremely competitive package (tax-free salary).

APPLY NOW

NVIDIA DGX-1

Самый эффективный инструмент для ИИ

960 Tensor TFLOPS | 8x Tesla V100 | NVLink Hybrid Cube

Заменяет сотню традиционных серверов на задачах ИИ

60 ТФлопс FP64

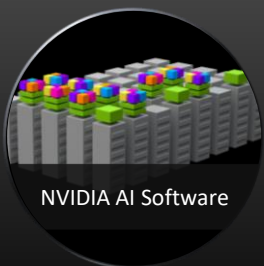
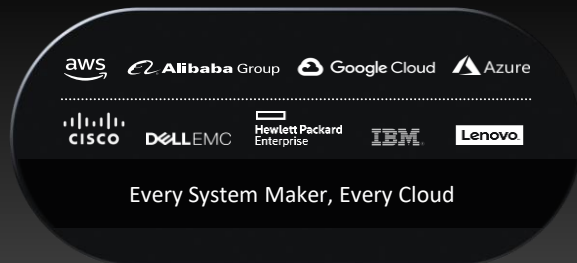
3U в стойке, 3.2 кВт

Интерконнект 4 EDR IB



www.nvidia.ru/dgx

Партнерство с NVIDIA для развития компетенции в ИИ



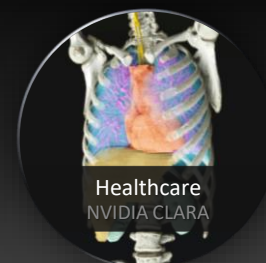
Технологии и экосистема

Оборудование и ПО компании NVIDIA является стандартом де-факто для построения ИИ-систем во всем мире. NVIDIA GPU доступны от любого производителя серверов, компьютеров и из любого облака.



Экспертиза и инвестиции

Ведущие специалисты мира в области ИИ работают на NVIDIA, ведущие ВУЗы сотрудничают с NVIDIA, программа NVIDIA Inception поддерживает ИИ-стартапы по всему миру, а институт NVIDIA DLI обучает специалистов по всему миру.



Отраслевые платформы

Четыре наши промышленные платформы объединяют технологии и экосистемы с тем чтобы способствовать развитию фундаментальных отраслей.

ВЫВОДЫ

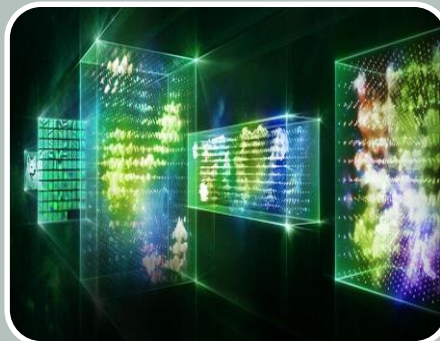
1. Ни один ВУЗ или институт больше не будет восприниматься всерьез без экспертизы в области ИИ
2. Организации без вычислительного ресурса для ИИ - неконкурентоспособны
3. NVIDIA DGX и партнерство с NVIDIA - самый быстрый способ получить действующую систему и начать исследования, разработку, курсы

NVIDIA Deep Learning Accelerator

NVDLA — процессор с открытым исходным кодом



Разработан как часть Xavier — платформы NVIDIA для автономных автомобилей



Оптимизирован для инференса сверточных нейросетей, машинного зрения



Открытая архитектура и RTL. Открытый исходный код компилятора



Законченное решение с Verilog и C-моделью, Linux драйвером, тестами, kernel и user-mode ПО и виртуальной платформой

Детали на WWW.NVDLA.ORG

secure | nvidia.org/index.html



[NVDLA Open Source Project](#) » [Documentation](#) »

NVDLA

The NVIDIA Deep Learning Accelerator (NVDLA) is a free and open source hardware accelerator. With its modular architecture, NVDLA is scalable and supports a wide range of IoT devices. Delivered as open source hardware, and documentation will be available on GitHub.

- **Open Source:** Developed [on GitHub](#) in an open community where contributions are encouraged.
- **Complete Solution:** Comes complete with a C-model, Linux drivers, test benches and test suites, kernel- and user-mode software, and software development tools. Easily portable to other hardware systems.
- **Scalable:** Well-suited to scale across a wide range of devices.



Learn
Gain a full
Learn More

Accelerating Deep Learning Inference

NVDLA introduces a modular architecture designed to simplify configuration to accelerate core Deep Learning inference operations. NVDLA hardware consists of several key components:

- Convolution Core – optimized high-performance convolution engine
- Single Data Processor – single-point lookup engine for activation functions
- Planar Data Processor – planar averaging engine for pooling.
- Channel Data Processor – multi-channel averaging engine for feature maps
- Dedicated Memory and Data Reshape Engines – memory-to-memory operations.

Each of these blocks are separate and independently configurable. A system can be configured to use a single planar averaging engine entirely; or, a system that needs additional convolution units can be configured to use a convolution unit without modifying other units in the accelerator. Scheduling is managed by a CPU; they operate on extremely fine-grained scheduling boundaries with a shared memory. Memory management can be made part of the NVDLA sub-system with a dedicated memory management implementation, or this functionality can be fused with the higher-level memory management implementation. This enables the same NVDLA hardware architecture to be used in a variety of system configurations.

NVDLA hardware utilizes standard practices to interface with the rest of the system. A standard interrupt interface, and a pair of standard AXI bus interfaces are used to connect to the system's wider memory system, including system DRAM and I/O peripherals. The second memory interface is optional, and allocated either to dedicated NVDLA or to a computer vision subsystem in general. This provides flexibility for scaling between different types of host systems.

[NVDLA Open Source Project](#) » [Documentation](#) »

NVDLA Index of Documentation

Welcome to the NVDLA open source project! Following, you will find an list of documents that have been written about the project.

- [NVDLA Primer](#) – an introduction to the concepts behind NVDLA, the solution that NVDLA provides, the basics of the architecture, and what's included in the NVDLA release. Intended for audiences of all levels.
- [Open Source Roadmap](#) – a look into what's next in NVDLA releases.
- [Hardware Manual](#) – hardware documents
- [Software Manual](#) – an exploration of the software ecosystem that supports NVDLA.
- [Virtual Platform](#) – an introduction of virtual platform for NVDLA.
 - [Virtual Platform On AWS FPGA](#) – an introduction of virtual platform for software to run on NVDLA (small scale) on AWS FPGA.
- [Open NVDLA Repository Updates](#) – a list of updates to the NVDLA repository
- [NVIDIA Open NVDLA License and Agreement v1.0](#) – the license under which NVDLA hardware is released.
- [Contributing to NVDLA](#) – some guidelines for contributing changes.
- This project adheres to the [Open NVDLA Code of Conduct](#); by participating, you are expected to uphold this code.
- [Glossary And Acronyms](#)



Антон Джораев, adzhoraev@nvidia.com