deci

RTiC

Run-time Inference Container

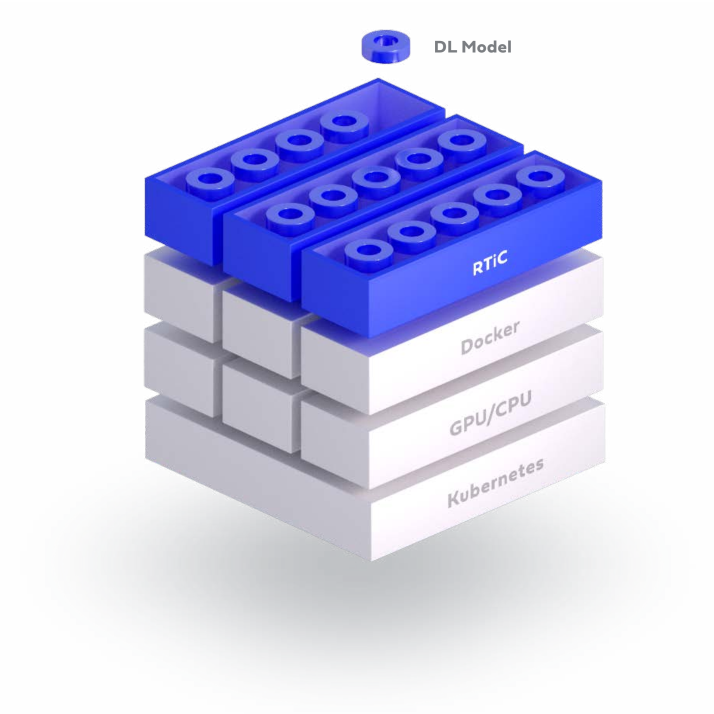
SOLUTION OVERVIEW

Take your DL models to production seamlessly while maximizing performance and resource utilization

Introducing RTiC

Based on various reports and surveys, only 20% to 30% of trained models find their way to production. Efficiently meeting the business and technical requirements to execute these models in their runtime environments and at scale, has become a complex task.

Considering how critical the inference of DL models is in production, it is vital that the performance of these models meets the specifications of the target hardware. It is also essential that the model integrates seamlessly with customer applications, using well-defined communication (APIs) alongside efficient data import/export methods between the model's and application's source code. Finally, it is imperative that both data-scientists and DevOps/engineering teams have options for the rapid deployment and scaling of model revisions. All these preconditions have turned the task of gearing-up models for production into a challenging, time-consuming, and unpredictable task.



At Deci, we are dedicated to helping you unlock new deep learning use cases, achieve faster time-to-market, and maximize both your resource utilization and the success rate of models reaching production. That's why we are excited to introduce the Deci Runtime Inference Container **(RTiC)**, a containerized deep learning runtime engine that turns any model into a siloed run-time server.

Deci RTiC enables the efficient inference and seamless deployment of models, at scale, to production on any hardware. It will give your data scientists the flexibility to work on any framework and boost their models to achieve state-of-the-art performance while enabling agile DevOps integration to any target environment.

Deploy & integrate any model framework into any environment with groundbreaking performance

Seeking the best way to prepare your models for production? Containerize your models with Deci's RTiC to ensure performance optimization and portability across multiple HW, platforms, and frameworks.

01.

Deploy and scale seamlessly

02.

Accelerate model performance

03.

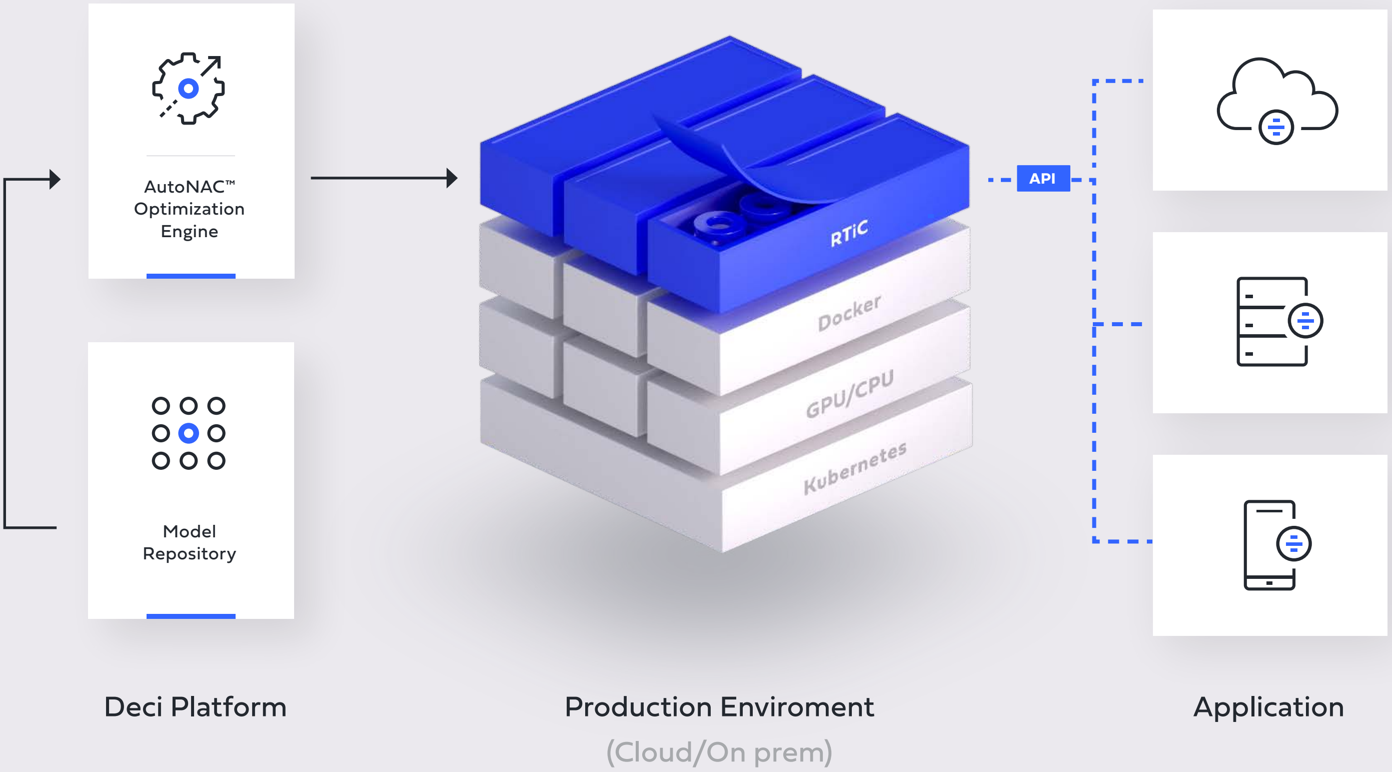
Any framework, for any environment

Package your DL models into production-ready docker containers, so they are easily deployed and ready for scaling.

Optimize for any given target hardware (CPU or GPU) to make the most of your cloud / hardware resources.

Enjoy full portability of your model development across frameworks, hardware types, and deployment environments, for easy integration to your applications.

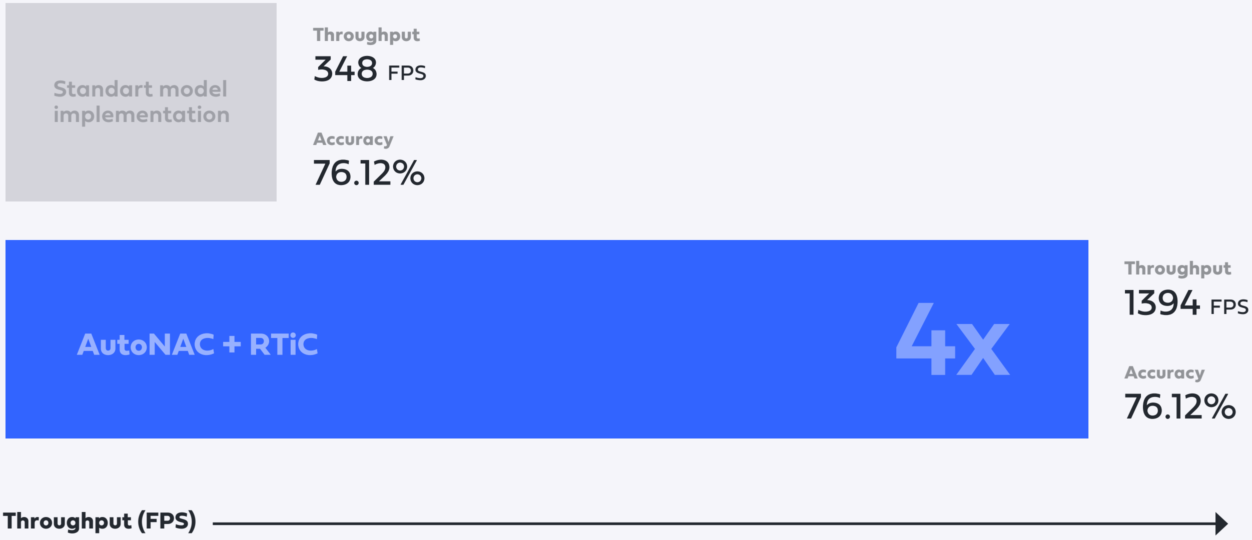
01. Deploy and scale seamlessly



Key Features

- Enable developers and DevOps to effortlessly deploy a container with the packaged models as part of their cloud microservices or native software architecture.
- Automatically and efficiently coordinate clusters of containers at scale in production, by using RTiC with Docker and orchestrating RTiC with standard infrastructures such as Kubernetes.
- Streamline your DL CI/CD by seamlessly integrating with Deci's platform model repository, allowing agile management, versioning, and deployment of your models across different production environments.
- Embrace portability and avoid vendor lock-in by easily importing models from your public/private cloud storage (such as AWS S3) or local storage.

02. Accelerate model performance

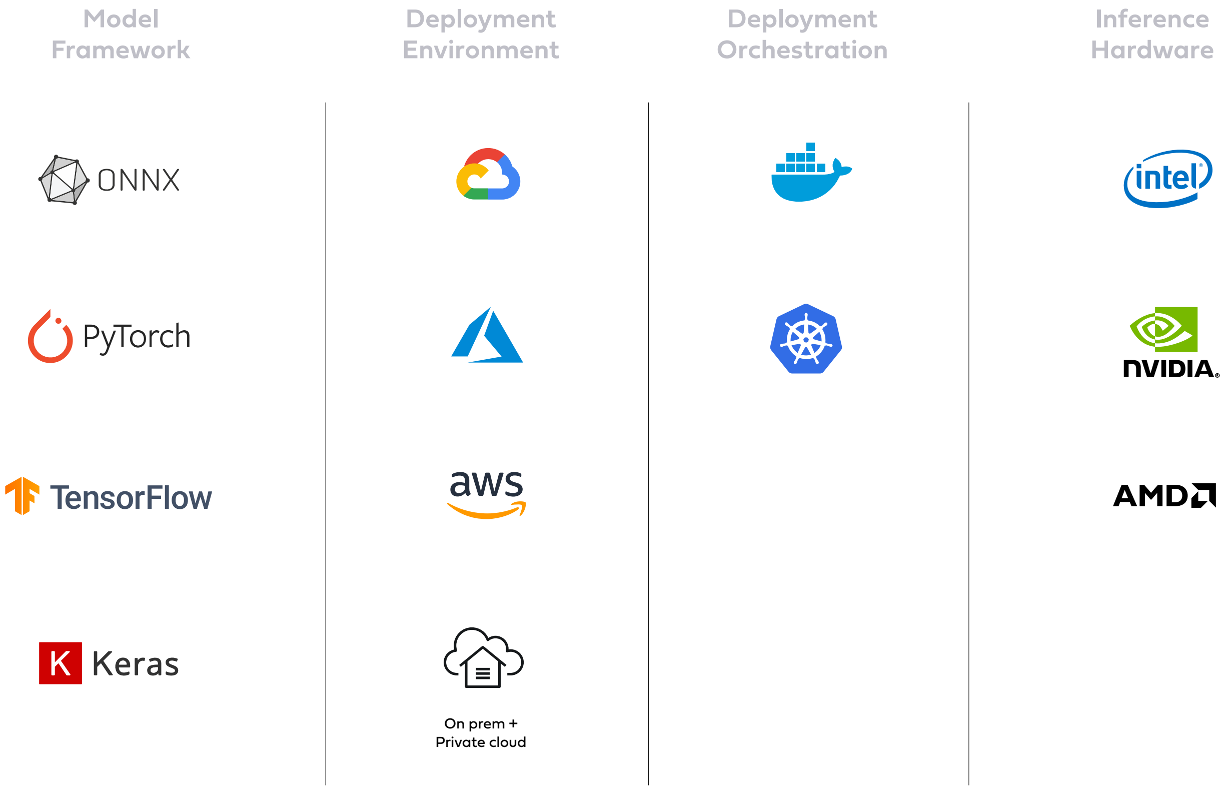


ResNet-50, Imagenet, Nvidia T4 GPU

Key Features

- Leverage Deci's patent-pending AutoNAC technology engine to boost any model's inference performance by up to 10x on any hardware, without compromising accuracy.
- Benefit from our runtime engine, which includes proprietary optimization technology alongside best-of-breed open-source graph compilers and hardware drivers.
- Maximize resource utilization with efficient model serving, using advanced pre-fetching and batching methods.
- Scale performance by running multiple models on the same hardware, pipelining several models on a single GPU/CPU to achieve maximum hardware utilization.
- Leverage concurrent execution of multiple models (or the same one) on the hardware.
- Realize close to zero server latency overhead using Deci's advanced communication protocols.

03. Any framework, for any environment



Key Features

- Use RTiC with all major DL frameworks, including TensorFlow, PyTorch, ONNX, and Keras, enabling data scientists to work on their framework of choice while maintaining full model portability at the organizational level.

- Serve and maintain different models and different versions of the same model, simultaneously within the same container.

- Smoothly integrate RTiC into your code with its client SDK available for any software applications in various programming languages (e.g., Python, Javascript, GO, C++, etc.).

- We use standard API communication for inference requests to/from the DL models within the RTiC. Communcication method can be either HTTP, gRPC or IPC.

- Achieve consistent and efficient data pre/post-processing to meet model formats.

Getting started

RTiC is built with simplicity in mind, to ensure easy, intuitive, and seamless deployment. Get started with a few lines of code, full detailed documentation, Jupyter notebook examples, and tutorials:

Contact us for more information >

Technical Specifications

System	Architecture	On-premise Cloud - GCP, AWS, AZURE
		Container-based software-only solution
User interface		CLI - API-based communication
		Python client
Communication		HTTP, gRPC, IPC
Integrations	Frameworks & Compilers	PyTorch (TorchScript), ONNX runtime, Keras, TensorFlow (Frozen Graph) TRT, OpenVino
	HW	Nvidia GPUs, Intel CPUs, AMD CPUs and GPUs
	Drivers	Nvidia Driver with support for CUDA 10.2+
	Orchestration	Kubernetes, GKS, AKS, EKS

About Deci

Deci is ushering in a new AI paradigm by using AI to build and operate AI models.

Deci's deep learning platform enables data scientists to transform their AI models into production-grade solutions on any hardware, crafting the next generation of AI for enterprises across the board.

Deci's proprietary AutoNAC (Automated Neural Architecture Construction) technology autonomously redesigns an enterprise's deep learning models to squeeze the maximum utilization out of its hardware.

Founded in 2019 and based in Tel Aviv, Deci's team of deep learning experts are dedicated to eliminating production-related bottlenecks across the AI lifecycle to allow developers and engineers the time to do what they do best - create innovative AI solutions for our world's complex problems.