



Kako Nvidia podstiče revoluciju u veštačkoj inteligenciji (AI)

Nikola Spasojević
Nvidia

Matematička gimnazija

17. 05. 2022.

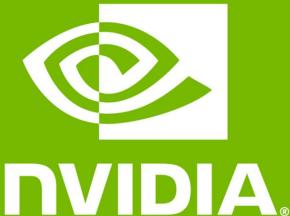
Uvod: Gde sam radio kao softverski inženjer



SAMSUNG



THOMSON REUTERS



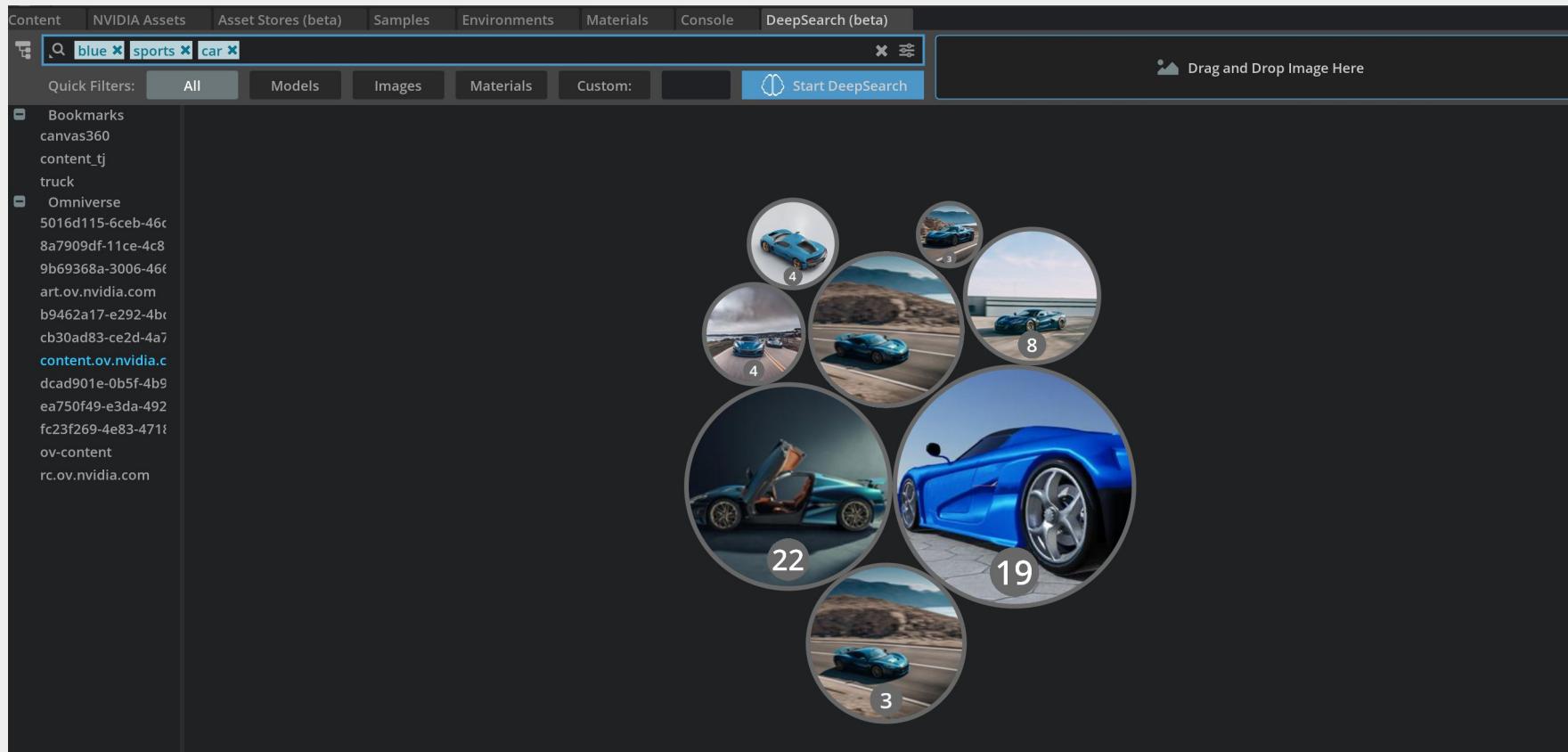
amazon alexa

CREDIT SUISSE

Trenutni projekti u Nvidia

Omniverse:

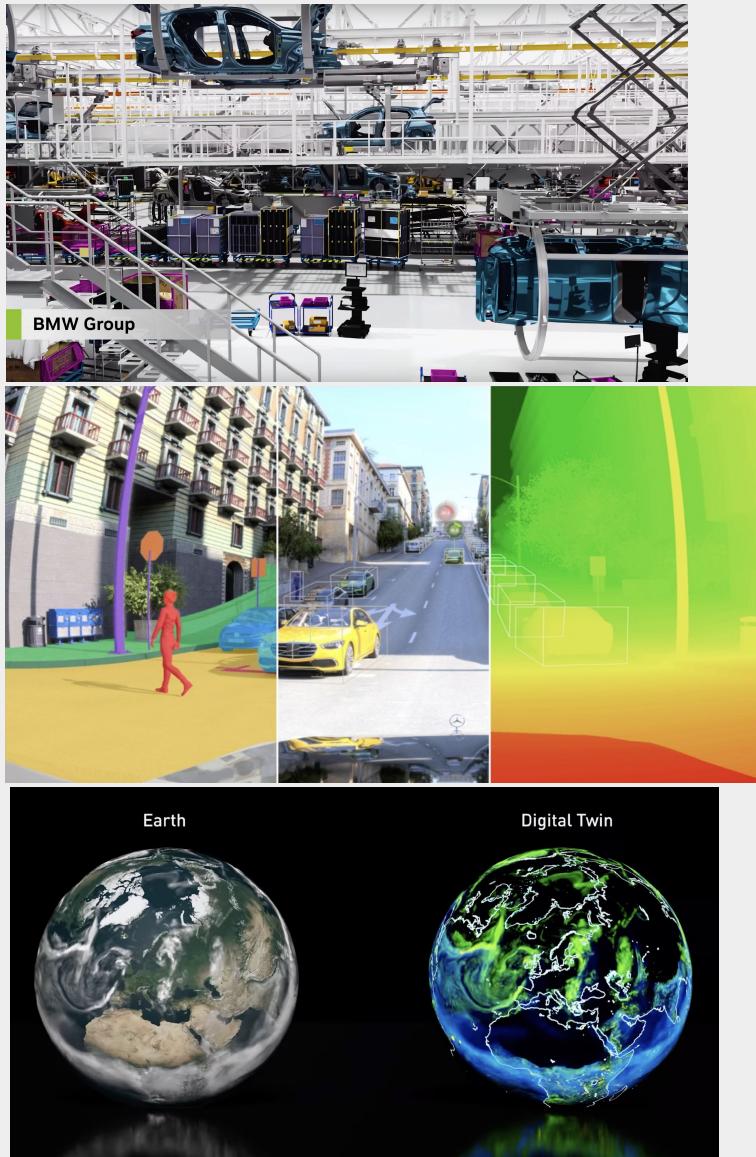
- Pretraga 3D objekata sa tekstom



Trenutni projekti u Nvidia

Digital Twins:

- Virtualne fabrike
- “Veštačke” datoteke za samovozeće automobile
- Simulacija klime





Large Language Model (LLM)

- [ChatGPT](#)
- [Github Copilot](#)

Open Source Machine Learning

- [HuggingFace Model Hub](#)

Stable Diffusion

- [Text-to-Image models](#)



Kako softverski inženjer koristi AI u svakodnevnici

ChatGPT

- Chatbot koji govori ljudskim jezikom (natural language)
- Razume “kontekst” teksta i slika

which is the fastest json serializer/deserializer?

The performance of JSON serializers/deserializers can vary depending on the programming language, implementation, and use case. However, here are some commonly used JSON serializers and deserializers that are known for their high performance:

1. Jackson: Jackson is a popular JSON library for Java, which is known for its high performance. It uses a streaming API to read and write JSON data, which makes it memory-efficient and fast.
2. GSON: GSON is another popular JSON library for Java, which is also known for its high performance. It uses a simple API to

for python?

For Python, there are several JSON libraries available, and the performance can vary depending on the library and implementation. However, here are some commonly used JSON libraries for Python that are known for their high performance:

1. ujson: ujson is a fast JSON library for Python that uses C extensions to accelerate its performance. It is known to be one of the fastest JSON libraries available for Python.
2. simplejson: simplejson is another popular JSON library for Python, which is known for its



Kako softverski inženjer koristi AI u svakodnevnici

Github Copilot

- Ljudskim jezikom preporučuje kod/opis celih funkcija
- Uči “kontekst” tvog koda

```
def binary_search(arr: list, target: int) -> int:
    """
    Binary search implementation.
    """

    low = 0
    high = len(arr) - 1

    while low <= high:
        mid = (low + high) // 2

        if arr[mid] == target:
            return mid
        elif arr[mid] < target:
            low = mid + 1
        elif arr[mid] > target:
            high = mid - 1

    return -1
```



Kako softverski inženjer koristi AI u svakodnevnici

HuggingFace Model Hub - “Demokratizacija” AI modela

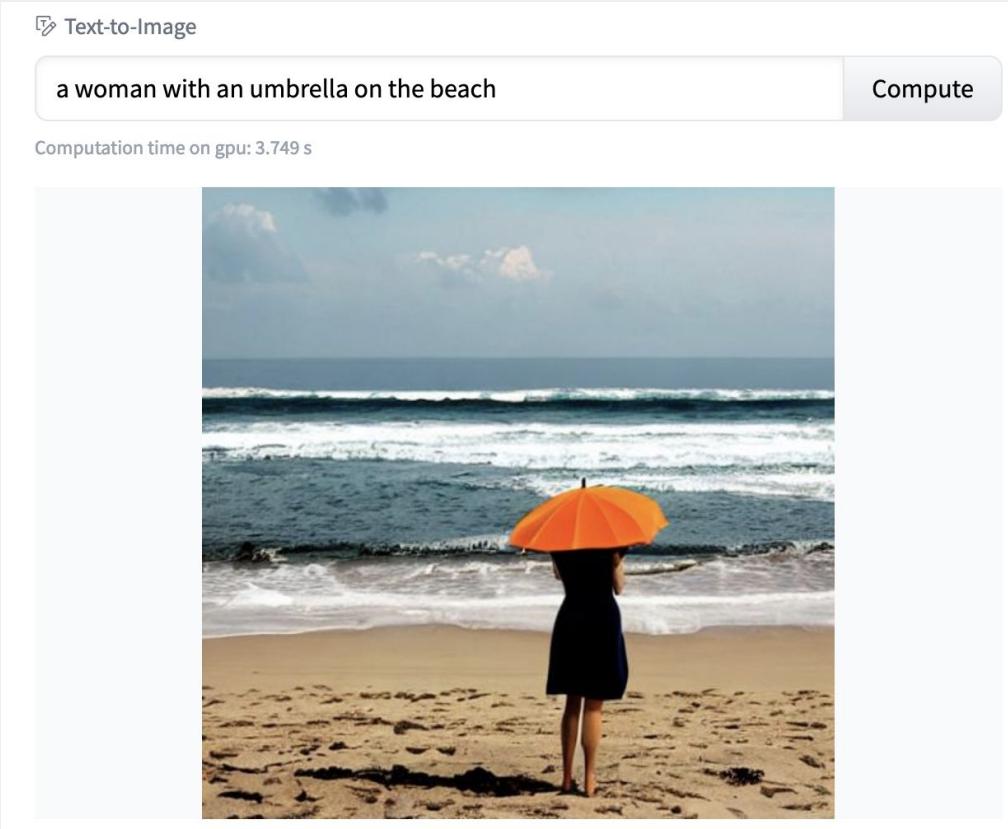
The screenshot shows the HuggingFace Model Hub interface. On the left, there's a sidebar with a search bar labeled "Filter Tasks by name". Below it, categories include Multimodal (Feature Extraction, Text-to-Image, Image-to-Text, Text-to-Video, Visual Question Answering, Document Question Answering, Graph Machine Learning), Computer Vision (Depth Estimation, Image Classification, Object Detection, Image Segmentation, Image-to-Image, Unconditional Image Generation, Video Classification, Zero-Shot Image Classification), Natural Language Processing (Text Classification, Token Classification, Table Question Answering, Question Answering, Zero-Shot Classification, Translation, Summarization, Conversational, Text Generation, Text2Text Generation, Fill-Mask, Sentence Similarity), Audio (Text-to-Speech, Automatic Speech Recognition, Audio-to-Audio, Audio Classification, Voice Activity Detection), Tabular (Tabular Classification, Tabular Regression), and Reinforcement Learning (Reinforcement Learning, Robotics). On the right, the main area is titled "Models 199,134" with a "Filter by name" input field. It lists various AI models with their details: bert-base-uncased (updated Nov 16, 2022, 67.7M, 827 stars), facebook/dino-vitb16 (updated about 24 hours ago, 35.4M, 12 stars), xlm-roberta-base (updated Apr 7, 22.3M, 281 stars), openai/clip-vit-large-patch14 (updated Oct 4, 2022, 13.5M, 384 stars), facebook/convnext-large-224 (updated Mar 2, 2022, 11.2M, 12 stars), facebook/convnext-base-224 (updated Feb 26, 2022, 9.91M, 7 stars), microsoft/layoutlmv3-base (updated 29 days ago, 9.8M, 143 stars), t5-base (updated Apr 6, 8.65M, 228 stars), xlm-roberta-large (updated Apr 6, 7.74M, 138 stars), microsoft/deberta-base (updated Sep 26, 2022, 6.01M, 41 stars), roberta-large (updated Mar 22, 5.4M, 105 stars), and albert-base-v2 (updated 23 days ago, 4.58M, 50 stars).

Model	Last Updated	Size	Stars
bert-base-uncased	Nov 16, 2022	67.7M	827
facebook/dino-vitb16	About 24 hours ago	35.4M	12
xlm-roberta-base	Apr 7	22.3M	281
openai/clip-vit-large-patch14	Oct 4, 2022	13.5M	384
facebook/convnext-large-224	Mar 2, 2022	11.2M	12
facebook/convnext-base-224	Feb 26, 2022	9.91M	7
microsoft/layoutlmv3-base	29 days ago	9.8M	143
t5-base	Apr 6	8.65M	228
xlm-roberta-large	Apr 6	7.74M	138
microsoft/deberta-base	Sep 26, 2022	6.01M	41
roberta-large	Mar 22	5.4M	105
albert-base-v2	23 days ago	4.58M	50

Kako softverski inženjer koristi AI u svakodnevnici

Stable Diffusion (Text-to-Image models)

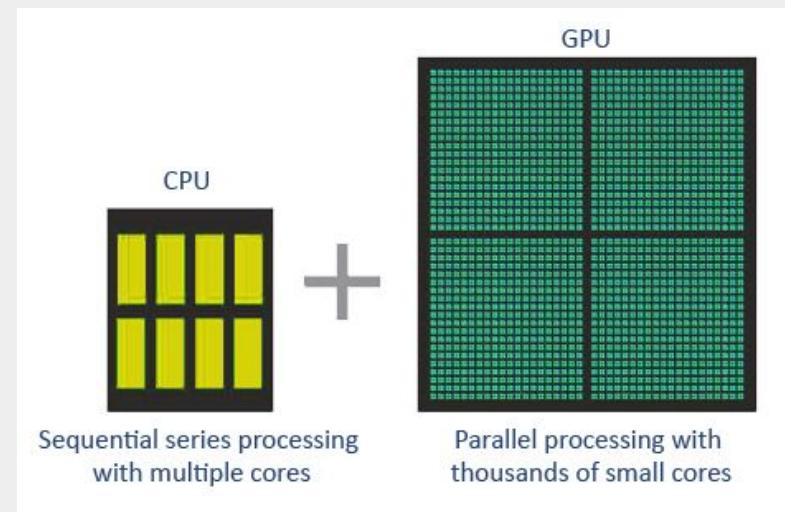
- Model koji razume “kontekst” u slici



Kako je ovo moguće?

- Šta je GPU?

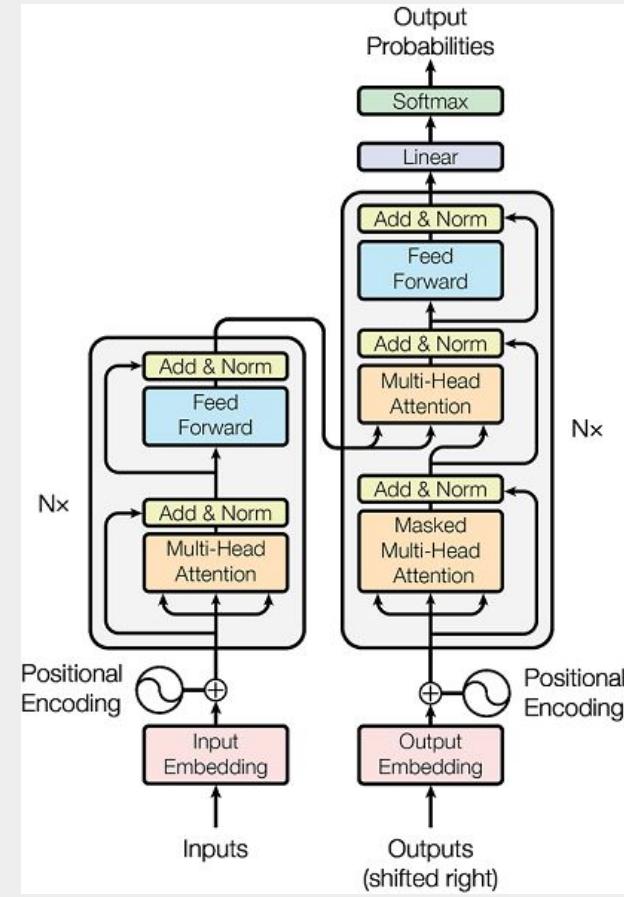
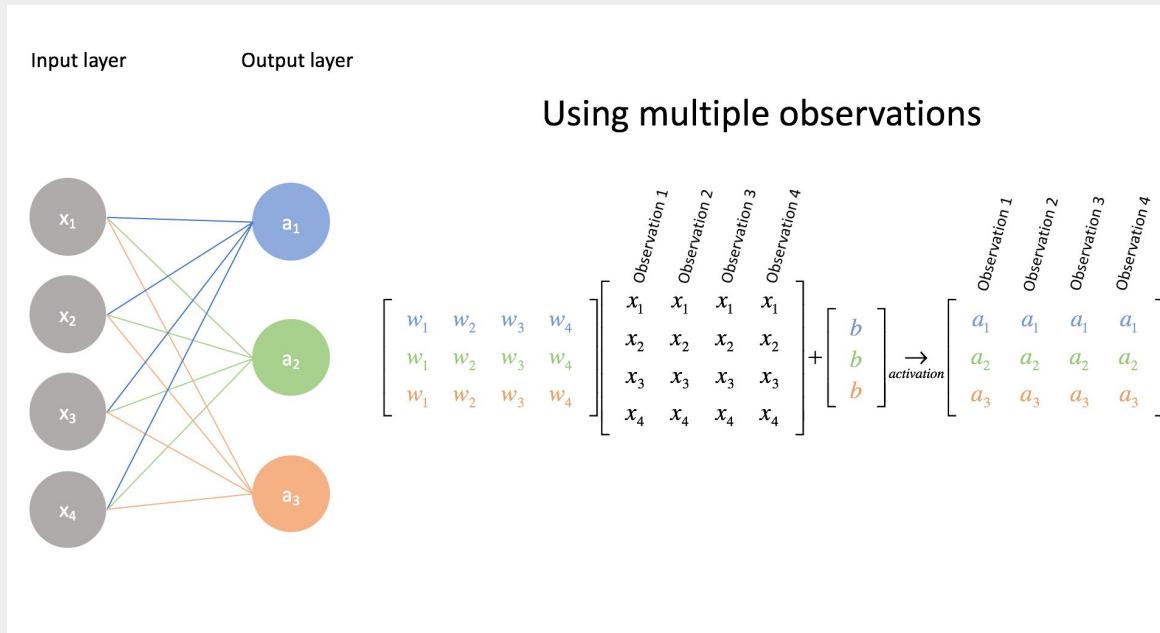
- Dizajnirano primarno za paralelizaciju matematičkih operacija
- Za razliku od CPU-a, koji radi operacije sekvencijalno
- Originalno korišćeno za grafiku i video



Kako je ovo moguće?

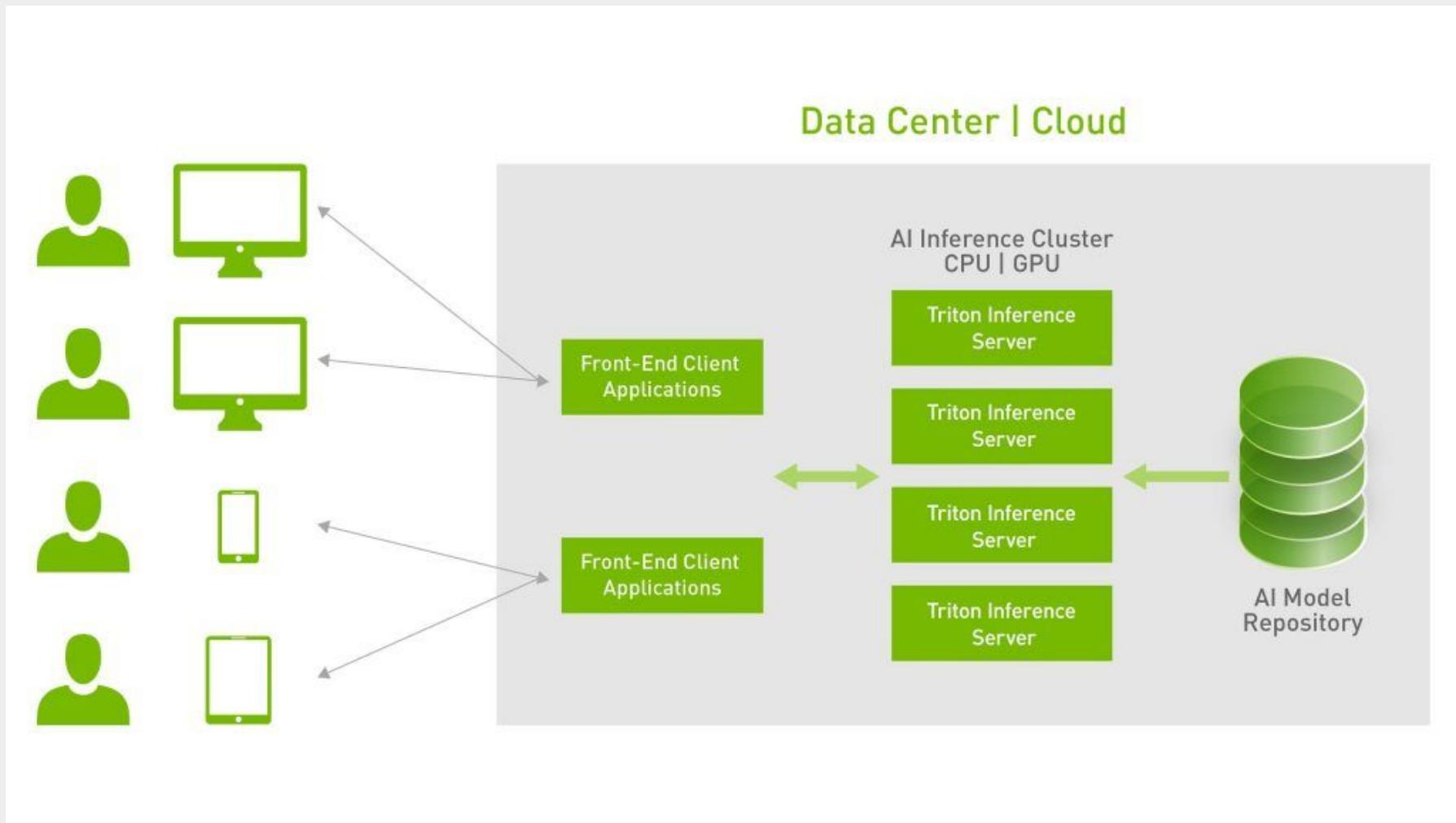
- Zašto se koristi GPU za “treniranje” AI modela?

- Jer modeli su samo matrice za koje tražimo tačne “weights”
- Možemo i da paralelizujemo više GPU-a



Deploying a model into production

Nije samo dovoljno trenirati model, treba ih “deploy”-ovati i “serve”-irati:





Deploying a model into production

Zašto je ovo najteži deo?

- Veliki su modeli
 - “Mala” verzija GPT4-a ima 100 milione parametra
 - Ne može svaki kompjuter/GPU da ih “drži” u (RAM) memoriju
 - Treba vremena da se izračuna rezultat
 - Treba imati dobru arhitekturu sistema da podnosi mnogo zahteva (npr. google prima stotine hiljade svake sekunde)
- Skupi su GPU-i

Mala napomena

- Programiranje je nova pismenost:
 - Trebamo svi biti spremni za digitalni svet
 - Sa kompjuterima možemo mnogo više da postignem
 - Copilot/[leetcode](#) kombinacija za učenje programiranja
- AI je već svuda oko nas:
 - Neće AI da nam ukrade poslove, nago ljudi koje koriste AI
 - Bitno je da znamo kako AI modeli funkcionišu
 - Za šta su sposobni
 - Za šta nisu sposobni



Hvala na pažnji!

Pitanja?