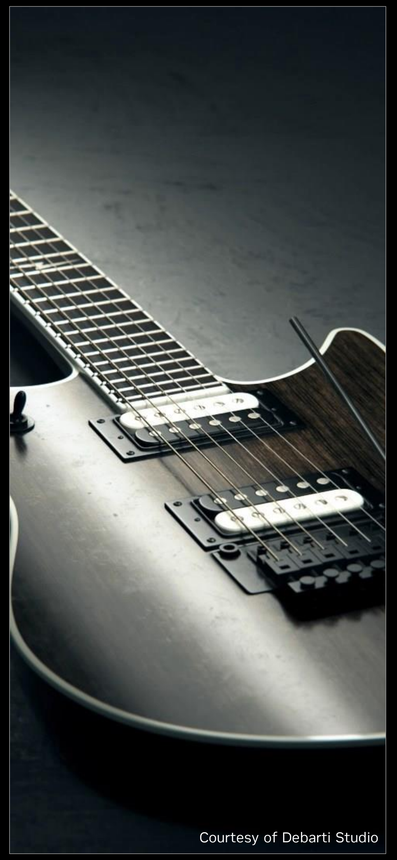
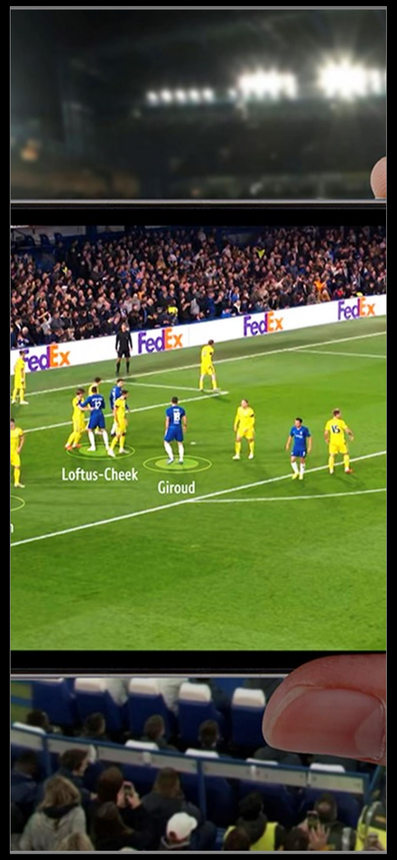
NVIDIA RTX-Powered AI Workstations for AI Training

Generative AI Transforming Workflows Across Industries

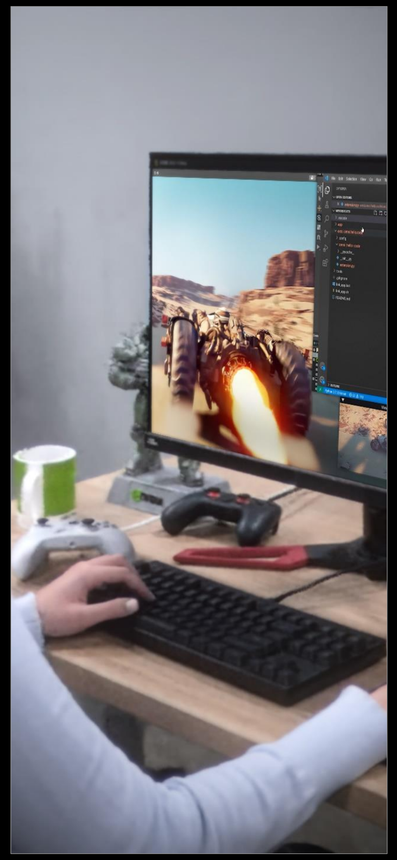
Architecture



Product Design



Film / Video



3D FX / Game Dev

Marketing

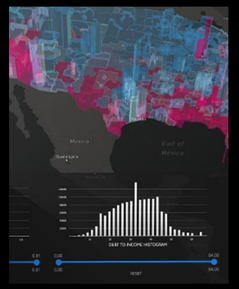
Photography

NVIDIA RTX

Built for AI and Graphics Intensive Workflows

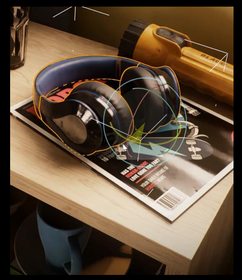
AI TRAINING & DEVELOPMENT

INFERENCE



GENERATIVE AI

CONTENT CREATION



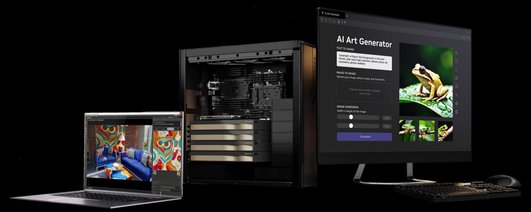
COLLABORATION

SIMULATION

INDUSTRIAL DIGITALIZATION

NVIDIA AI ENTERPRISE

NVIDIA Omniverse Enterprise



WORKSTATION

DATA CENTER

CLOUD

NVIDIA RTX-Powered AI Workstations: Key Workloads

The new generation of workstations powered by NVIDIA RTX Ada Generation professional GPUs are ideal for today's demanding AI workflows

AI Training & Development

AI Inference, Generative AI-Augmented Applications & Workflows

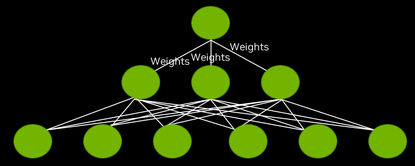
Data Science

NVIDAI RTX-Powered AI Workstations for AI Training

AI Vocabulary

- **Models**:
  - Program that creates a trained AI neural network. There are many types of AI models, common Generative AI models include:
    - Large Language Models (LLMs)
      - Models: GPT3, GPT4, Llama, Llama2,
      - Implementations: ChatGPT, CoPilot
    - Diffusion models
      - Models: DALL-E, Stable Diffusion XL
      - Implementations: Automatic 1111, Midjourney

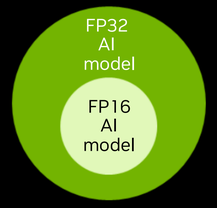
- **Parameters**:
  - the values the AI model can change as it learns. Examples of parameters include: the weights in a neural network or the coefficients in a linear regression or logistic regression. The final values of these parameters will be included in the final version of the model.



Some models include the number of parameters in the model's name:
- Llama 2 70B – 70 billion parameters
- Llama 2 13B – 13 billion parameters
- GPT3 40B – 40 billion parameters
- Falcon 7B – 7 billion parameters

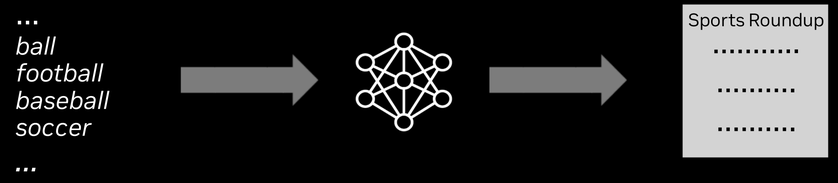
- **Data Formats**
  - Numerical format of the data used by the model. Formats can be 32-bits, 16-bits, 8-bits, or 4-bits in size. Common data formats are FP32, TF32, FP16, BF16, FP8, INT8, INT4 (FP stands for “floating point”, number with a decimal point). The more bits, the larger number you can represent, but it takes up more compute memory space.

8-bits per byte, so:
32-bit = 4 bytes
16-bit = 2 bytes
8-bit = 1 byte



An AI model using FP16 data will be half the size of the same model using FP32, using half the amount of GPU memory

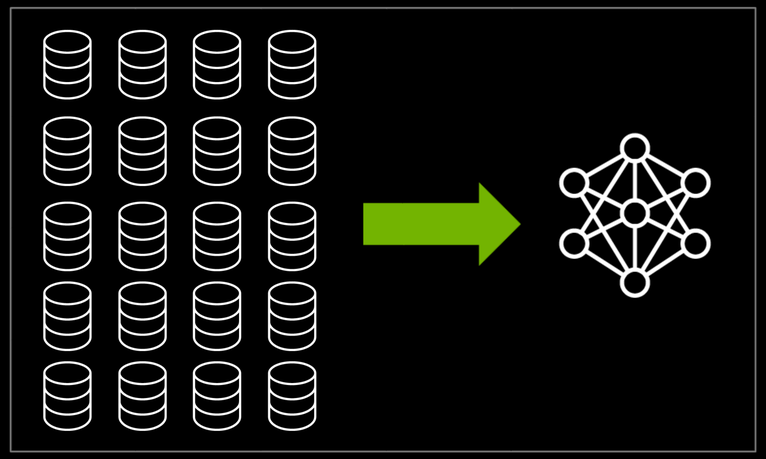
- **Tokens**:
  - The basic units of text or code that an AI model uses to learn, process, and generate output. Tokens can be characters, words, sub words, or even segments of text or code, numbers, etc., depending on the method or scheme used by the AI model implementation



What is AI Training?

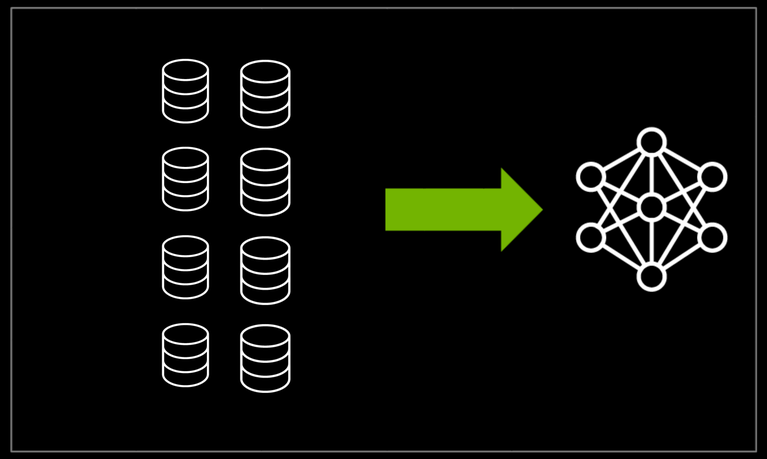
- Artificial intelligence (AI) training is the process of teaching an AI system with large amounts of data so that it can interpret and learn from the data so that it will be capable of making decisions base in the new information it is provided (inferencing). Training is done in 3 steps:

1 Training



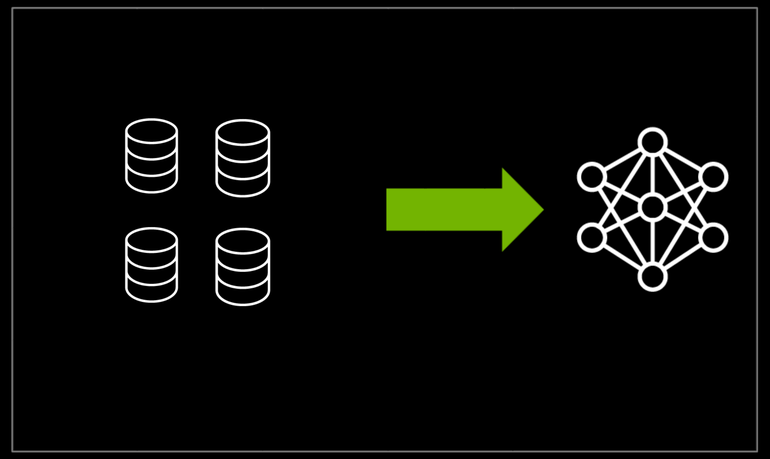
Large amounts of data are fed into the AI model to teach it

2 Validation



A smaller data set is used to validate how well the train model performs on data it has not seen before

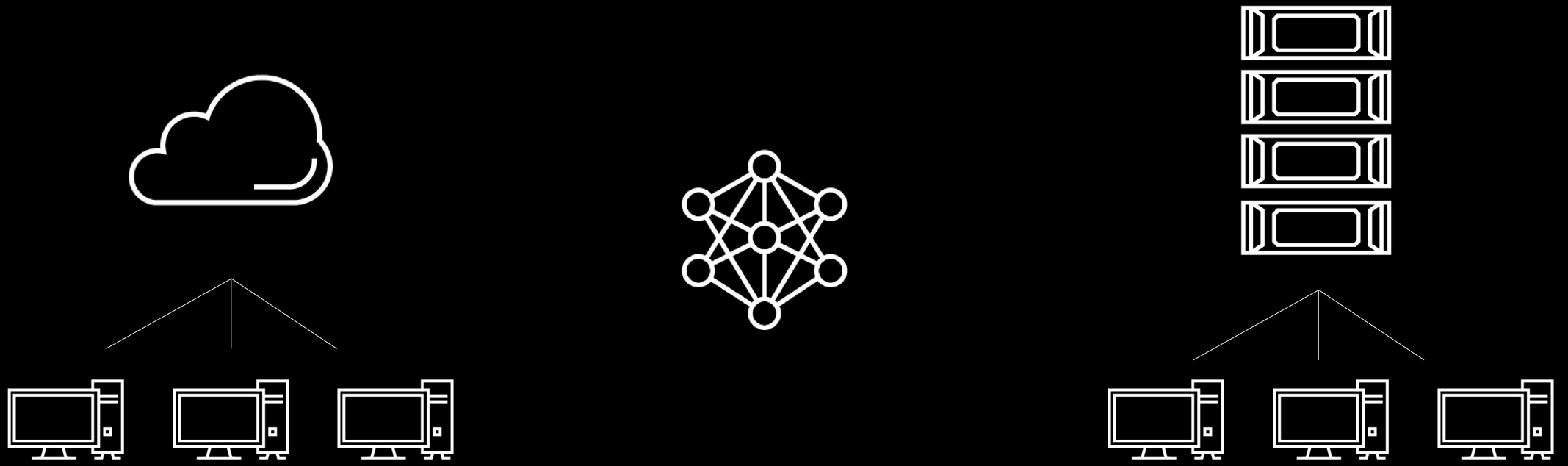
3 Testing



Trained model is given new data as it would receive in the real world to evaluate performance

NVIDIA RTX-Powered AI Workstations for AI Training & Development

- The explosion of large language models (LLMs) and generative AI is causing incredible demand for AI computing resources. As data centers and cloud service providers (CSPs) add and augment AI computing capacity, server-based resources will continue to be in high-demand. AI Workstations can augment data center and cloud-based resources for AI model training, research and development tasks, providing critical AI compute resources to help meet the demand.

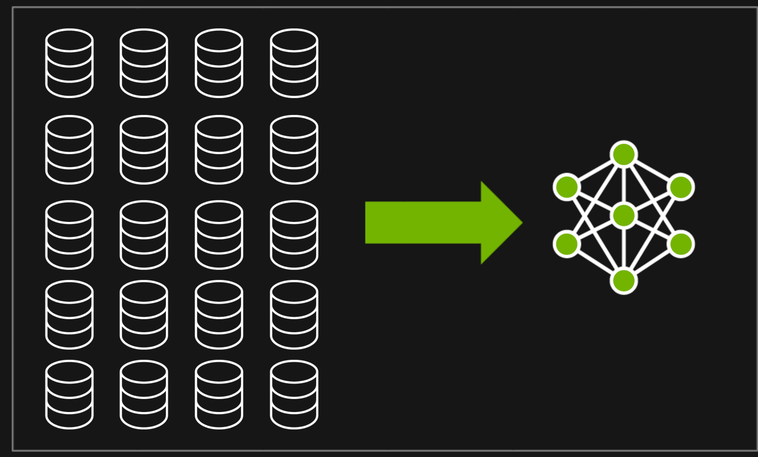


Types of AI Training

We will focus on two types of training for this session

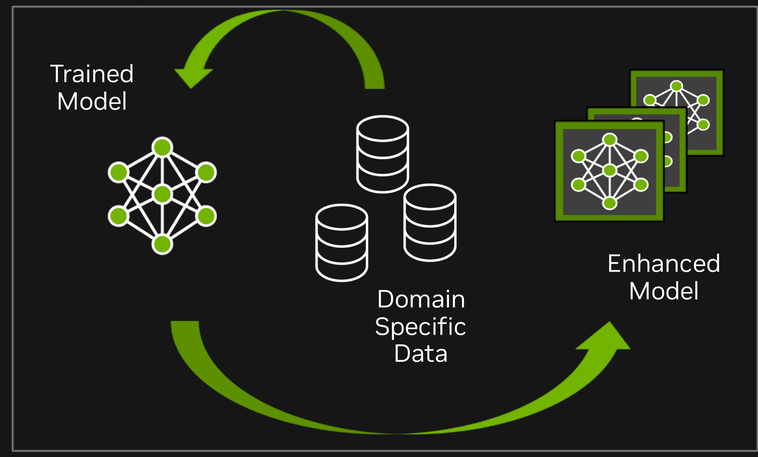
Foundation model training
(Train a model from scratch)

- AI models trained on massive amounts of unlabeled datasets, usually through self-supervised learning.
- Examples of these kinds of models are large language models (LLMs) such as ChatGPT, and DALL-E, diffusion models such as Stable Diffusion



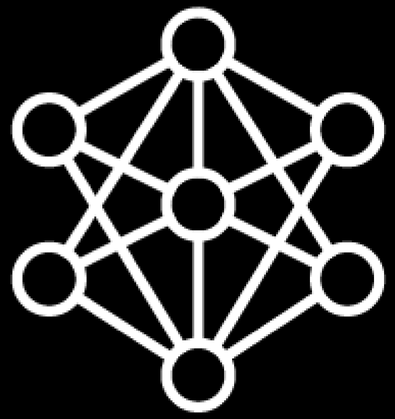
Fine Tuning Training
(Train a pre-existing model)

- A pre-trained AI model is given additional training with new data to fine tune the results
- Models can be fine tuned by additional model training or using prompts to provide the new information



AI Foundation Model Training

- The latest AI models are very large and continue to grow in size and complexity

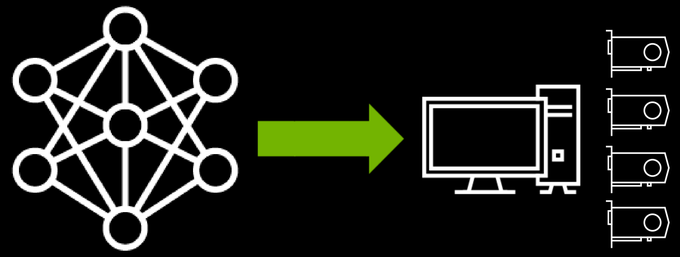


- Over 1 trillion parameters
- Training data sets of 100s of gigabytes of data
- Training can take several months on very large clusters of servers

- How can we use workstations for training with such large models?

AI Workstations for Training Workloads

- Workstations are ideal for training smaller models, doing local research and experimentation



- Workstations are ideal for training smaller models, local experimentation
  - Example: NVIDIA Tiny CUDA NN – used for Instant Neural Graphics Primitives
    - https://github.com/NVlabs/tiny-cuda-nn

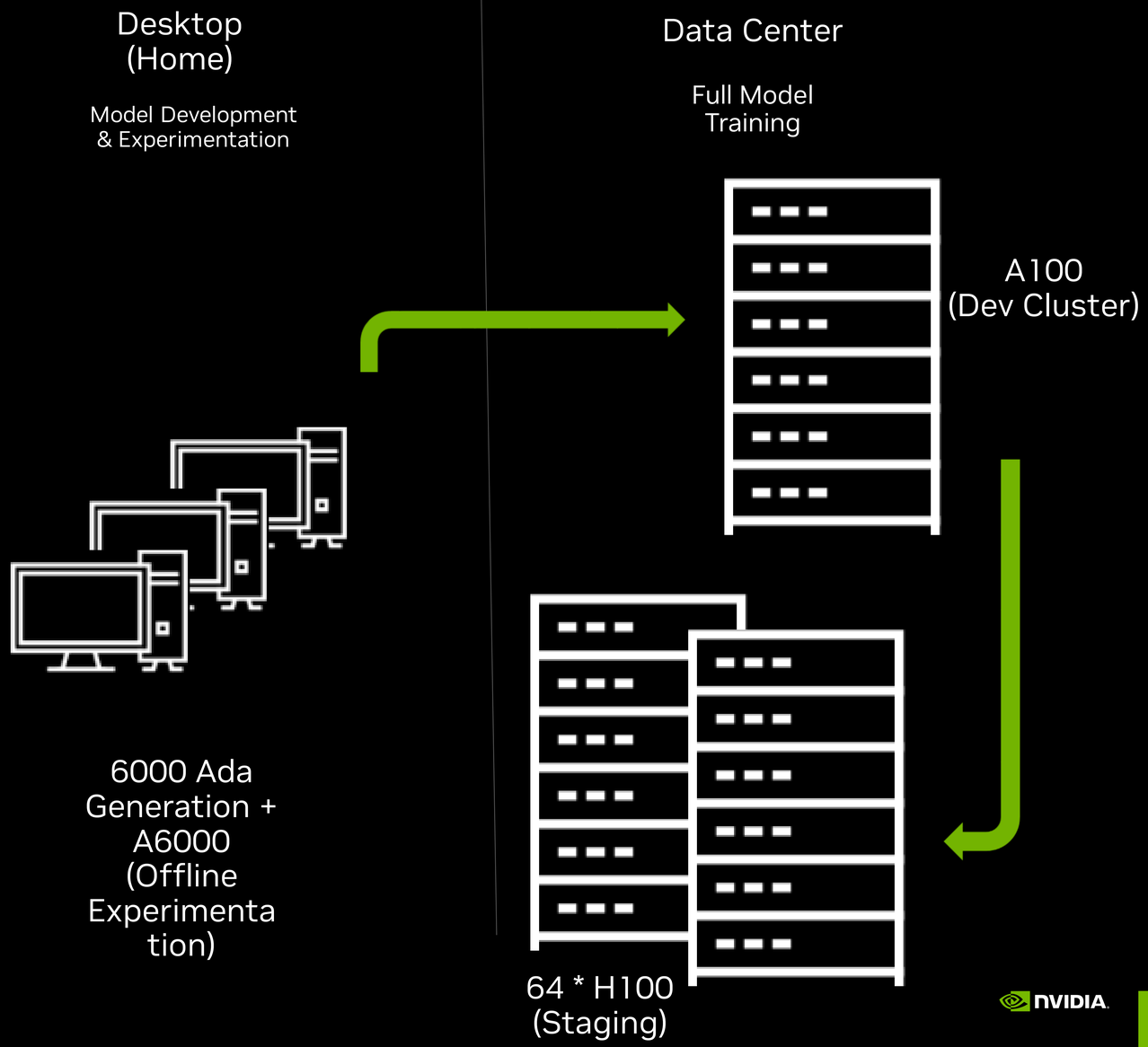
- Research on creating smaller models and small language models (SLMs) is building momentum
  - Examples
    - *LLaMA (Large Language Model Meta AI) trained four model sizes: 7, 13, 33, and 65 billion parameters*
    - *Falcon: Falcon-40B, 40B parameters, Falcon-7B, 7B parameters*
    - *Orca (Microsoft) : 13B parameters*
    - The Power of Small Language Models: A Quiet Revolution

  - TF32, FP16, BF16, FP8, and mixed precision data formats can reduce memory footprint

- For larger models, researchers and developers can experiment locally with reduced parameters and data sets to experiment and tests
  - Work can be moved to data center or cloud for full training if required

AI Workstation Training Example: Irreverent Labs

“NVIDIA GPUs uniquely equip Irreverent Labs to manage our advanced deep learning workloads. The RTX 6000 Ada Generation GPU is a game-changer for local experimentation, significantly speeding our pace of AI innovation, tasks that would’ve previously taken two to three days now take less than 12 hours. The NVIDIA H100 and A100 Tensor Core GPUs are ideal for large-scale training and online inference, giving us up to 4x performance improvements for 4D video prediction.”



Customers to Target for AI Workstation Training Opportunities

- Customers who are doing AI development:
  - AI researchers, application developers, software developers, traditional ISVs, HER students & researchers
- Industries:
  - Large enterprises across industries looking to integrate generative AI into their business processes and investigating or currently working on creating their own AI models
  - Established ISVs across industries, AI startups, HER
  - Most ISVs are working to integrate generative AI into their applications

Conversation Starters

- *What is your company's AI strategy?*
- *Do you create AI models or do any AI model training?*
  - If "Yes"
    - *What is your current model training infrastructure? (server / data center / cloud)*
    - *What systems do your AI researchers use for experimentation?*
    - *Do you have requirements to secure data locally?*
    - *Are you looking to increase your model training compute capacity?*
      - *Do you have servers on order?*

"Workstations can augment your existing data center and cloud resources and provide a cost-effective solution to augment these resources, providing additional computing power and enable AI researchers to work locally for development and experimentation and seamlessly move to the data center or cloud for more computing power if necssary. We have workstations ready to ship to help accelerate your AI development efforts."

If customers want to discuss AI frameworks, AI model specifics, tool chains or other technical details, time to bring in your technical resources or contact your NVIDIA business partner to involve NVIDIA technical experts

AI Model Fine Tuning

- AI model fine tuning is additional training applied to a pre-trained model with a small subset of data to give the model specific knowledge
- Example:
  - Prompt: “Create a picture of Toy Jensen in space”

Toy Jensen

ORIGINAL RESULTS

TUNED WITH 8 IMAGES

TUNED WITH 50 IMAGES

Why NVIDIA RTX-Powered AI Workstations for AI Model Fine Tuning?

Large GPU memory and scalable computing platform

- AI model fine tuning is AI training using an existing model and a dataset that is smaller than the original training data set, but it still requires significant computing resources to perform the training

- AI Workstations can be configured with up to 4 NVIDIA RTX 6000 Ada Generation GPUs
  - 48GB of GPU memory per GPU, 192GB total workstation system GPU memory
  - Up to 1457 TFLOPs of AI compute per GPU, 5,828 TFLOPS total compute per workstation

- Example fine tuning training:
  - 4x NVIDIA RTX 6000 Ada Generation workstation
  - Fine-tuning of GPT3-40B with 860M tokens: 15 hours to complete

- Significant research on tuning models on desktop systems is building momentum
  - “QLoRA: Fine-Tune a Large Language Model on Your GPU”, Towards Data Science, June 1, 2023 https://towardsdatascience.com/qlora-fine-tune-a-large-language-model-on-your-gpu-27bed5a03e2b

What Customers are Doing AI Model Fine Tuning?

- Customers across vertical segments: Designers, artists, creators, researchers, scientists, students
  - Most customers are using publicly or commercially available models – they will need to fine tune for their specific use cases
  - Some customers may need to fine tune models on a regular basis
    - Adoption of new models
    - New products, information, etc., will require additional fine tuning for existing models

- Industries:
  - Generative AI has the potential to impact every industry, some will adopt more rapidly
    - M&E industry has historically been an early adopter
    - Manufacturing/Product design – creative teams will likely adopt sooner than engineering teams
    - AECO – architects, creative teams will likely adopt quickly to help with concept & asset creation
    - HER – students, researchers will continue to explore and fine tune models as part of their coursework and research
    - ISVs / software developers will need to fine tune models for specific products and product features
      - Models may need to be updated/fine tuned for product releases/updates

Conversation Starters

- *“How do you use AI in your business?”*
- *“Are you using generative AI as part of your workflows?”*
- *If “Yes”:*
  - *“Do you need to fine tune models to get the desired results?”*
    - If no, *“Would fine tuning AI models help give you better results?” (you may need to explain fine tuning if they are unfamiliar with it)*
    - If yes, *“How often do you need to fine tune AI models?”*
      - *“How are you fine tuning your models?” (laptop/desktop/data center/cloud)*
        - *“How long does it take to fine tune your models?”*
        - *“What tools are you using to fine tune your models?”*

“Workstations can provide a cost-effective solution to augment data center and cloud resources to accelerate your AI model fine tuning tasks. They provide a scalable solution that can be used by multiple teams for model fine tuning”

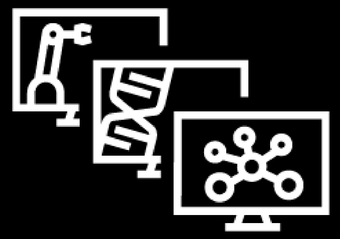
If customers want to discuss specific AI models, fine tuning tools, time to bring in your technical resources or contact your NVIDIA business partner to involve NVIDIA technical experts

AI Training Tools - Resources

- NVIDIA has built the most extensive AI development ecosystem available. To learn more, you can guide your customers to:
  - NVIDIA AI web pages: overview of latest AI technologies and solutions
  - NVIDIA AI Enterprise: enterprise-grade software that powers the NVIDIA AI platform, accelerated the development and deployment of production-ready generative AI, computer vision, speech AI, data science, and more.
  - NVIDIA NGC™: online portal of enterprise services, software, management tools, and support for end-to-end AI workflows. NGC provides access to GPU-accelerated software for AI model development, pretrained AI models, and industry-specific SDKs.
  - NVIDIA Developer Program: provides access to developer tools, SDKs, training, workshops, and documentation for developers working with AI, HPC, graphics, rendering, simulation, video/broadcast/display, and many other technologies.
  - NVIDIA RTX Professional GPUs webpages: solutions for enterprise level desktop and mobile workstations
  - NVIDIA AI Workbench: unified, easy-to-use toolkit that allows developers to quickly create, test and customize pretrained generative AI models and LLMs on a PC or workstation -- then scale them to virtually any data center, public cloud or NVIDIA DGX Cloud.

NVIDIA AI Workbench

Accelerating Generative AI Workflows



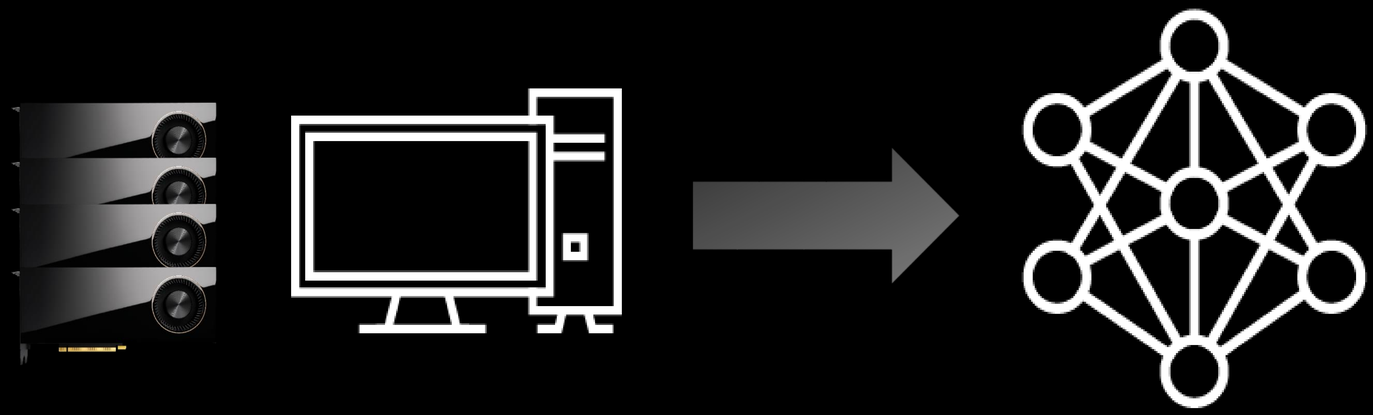
- Set up containers and developer environments on Windows and Linux machines with one click
- No setup access to the best GPU-optimized frameworks through JupyterLab and VS Code

- Streamlined containerization to quickly and easily build quality models with NVIDIA accelerating frameworks and libraries
- Open-source software from GitHub, Hugging Face, and NVIDIA NGC™ runs on Windows and Linux without extra effort

- Collaborate and share work projects locally and remotely with ease
- Effortlessly move workloads across laptops, GPU servers, cloud instances, and NVIDIA DGX™ Cloud

Easily grab and deploy pre-trained models – AI workflow automation for beginners and experts

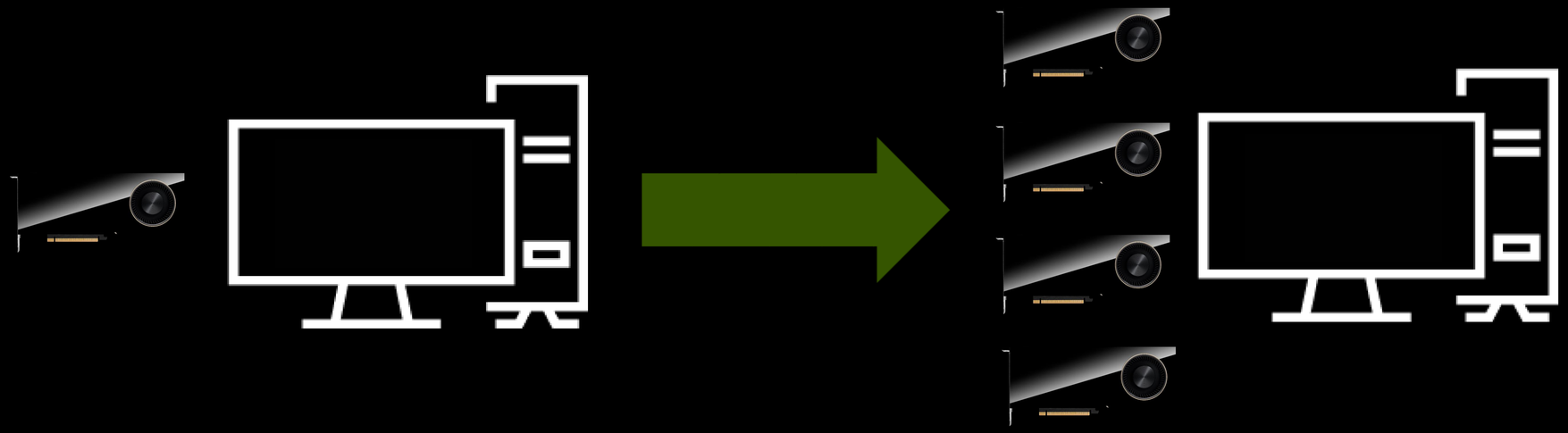
NVIDIA RTX-Powered AI Workstations for Training



- Workstations can help your customers with their AI training efforts, offloading over subscribed data center resource, expensive cloud instances, and provide compute resources for local training
  - Provide a local compute resource for AI training research, development, and model & data exploration
  - Local training for smaller models
  - Enable local compute resources for model fine tuning training

NVIDIA RTX-Powered AI Workstations

Recommended Configurations for AI Training, R&D, Data Science Workloads



Latest generation workstations plus:

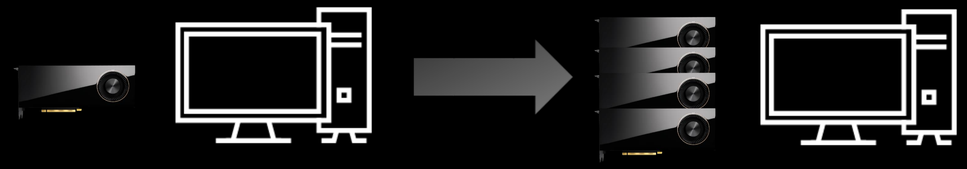
- 1 – 4 RTX 6000 Ada Generation GPUs

Latest Generation Mobile workstations plus:

- RTX 5000 Ada Generation Laptop GPU

NVIDIA RTX-Powered AI Solutions

NVIDIA RTX-Powered AI Workstations
AI Training/Development, Local Inference, and Data Science Workloads



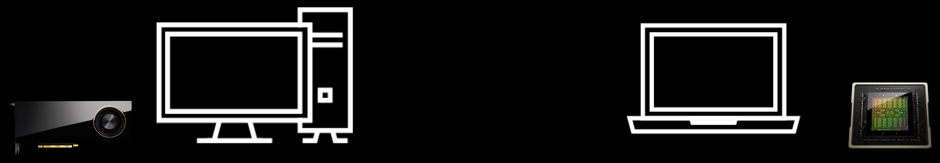
Latest generation workstations plus:

- 1 – 4 RTX 6000 Ada Generation GPUs

Latest Generation Mobile workstations plus:

- RTX 5000 Ada Generation GPU

NVIDIA RTX-Powered Workstations
AI Enabled Applications (Inference)

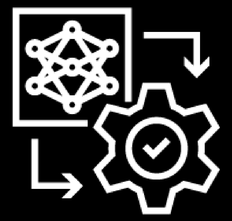


	Latest generation workstations plus:	Latest generation laptops plus:
Large Workloads	RTX 6000, 5000 series, Multi-GPU	RTX 5000, 4000 series
Medium Workloads	RTX 4000 series, Multi-GPU	RTX 3000, 2000 series
Basic Workloads	RTX 2000 series	RTX 1000, A500 series

NVIDIA RTX Powered Workstations for AI

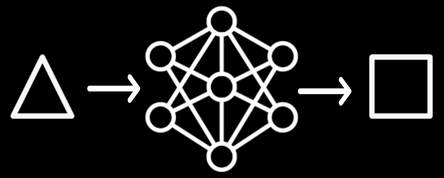
Ideal platform for professional AI workflows

AI Training



- Provides additional development resources to augment data center and cloud resources
- Additional compute resource for smaller model training, model fine tuning, experimentation, and application development

AI Inference



- Provides performance and large GPU memory for today's rapidly evolving Gen AI augmented workflows
- NVIDIA RTX acceleration of AI-enabled applications
- Small form factor solutions for powerful inferencing at the edge

Data Science

- Cost effective compliment to data center and cloud resources
- Provides local compute resources for data science projects at the desktop for data, feature, and model experimentation and exploration

Workstations Value Proposition

- Compute power, large GPU memory capacity (with ECC) and scalability required for today’s demanding professional multi-application workflows
- Tuned, optimize, and certified for professional ISV applications
- Enterprise level hardware/software support, deployment tools – maximize uptime and minimize IT support requirements
- NVIDIA Enterprise Drivers are tuned, tested and validated by OEM partners to provide enterprise level stability, reliability, and support for mission critical workstation deployments
- Professional features designed to accelerate, optimize, and improve the efficiency of today’s modern professional workflows.
- Form factors, power requirements, and cooling designed for workstation deployments and multi-GPU configurations.

Resources

NVIDIA RTX-Powered AI Workstation deck (customer facing version)

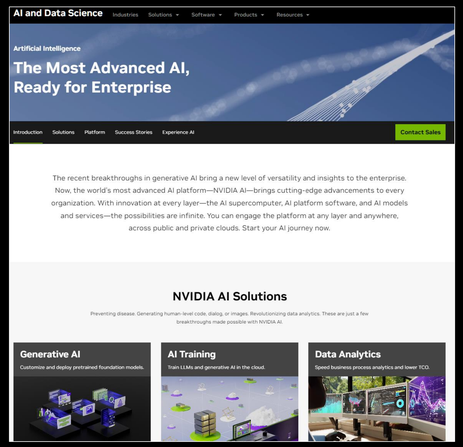
Product Decks & Go-to-Market Kits

https://nvid.nvidia.com/dashboard/

NVIDIA RTX Powered AI Workstations

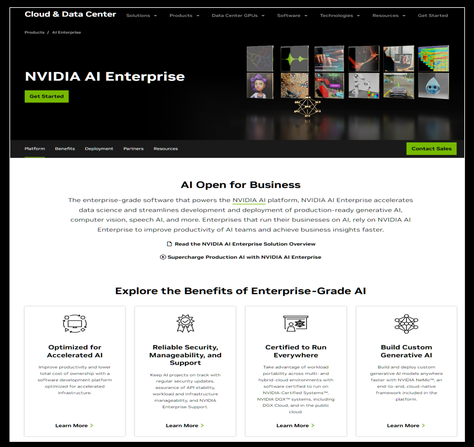
https://www.nvidia.com/en-us/ai-data-science/workstations/

NVIDIA AI



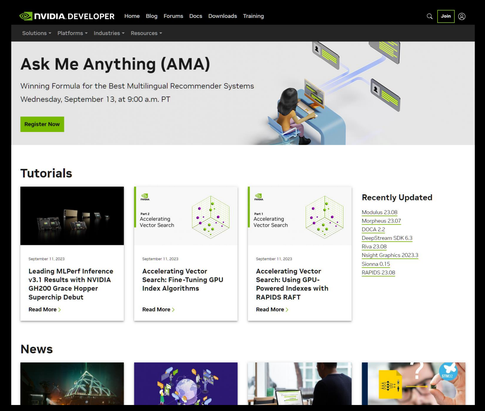
https://www.nvidia.com/en-us/ai-data-science/

NVIDIA AI Enterprise



https://www.nvidia.com/en-us/data-center/products/ai-enterprise/

NVIDIA Developer Program



https://developer.nvidia.com/


NVIDIA