BOSTON
Servers | Storage | Solutions
A DIVISION OF SOURCECODE

NVIDIA Elite Partner

WEBCAST: NVIDIA GTC RECAP

Produkt-Highlights, Roadmap Updates,Inference Scaling

21.05.2026 | 10:00 AM CET

BOSTON
Servers | Storage | Solutions
A DIVISION OF SOURCECODE

NVIDIA Elite Partner

REFERENTEN

TILMAN STRÜBIG
Senior Solutions Architect AI
Boston IT

ANGELIKA HARRER
Head of Marketing
Boston IT

BOSTON
Servers | Storage | Solutions
A DIVISION OF SOURCECODE

NVIDIA Elite Partner

AGENDA

- GTC 2026 Overview
- Hyperscaler Sidenote
- Relevant Shifts for non-Hyperscalers
- Choosing the Right Hardware for the Model
- Unified Software Stack
- Q&A

AI - THE NEW INDUSTRIAL REVOLUTION

APPLICATIONS

MODELS

INFRASTRUCTURE

GPUS

ENERGY

CHATBOTS

DIGITAL BIOLOGY

ROBOTAXI

ENTERPRISE AI AGENTS

SCIENCE

ROBOTICS

MANUFACTURING

AI CODER

LLM | VLM | VLA | MMLLM | GPT

DM | GNN | MOE | SSM | LBM

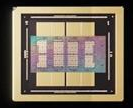
AI FACTORIES

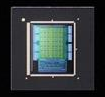
NVIDIA Vera Rubin

7 Chips – 5 Rack Systems

AI Factory for the Agentic AI Frontier

1 GW AI Factory	X86 + Hopper	Vera Rubin
# of GPUs	600K	300K
AI FLOPS	1.2 ZFLOPS	16 ZFLOPS
All-to-All Scale-up	7.2 TB/s	260 TB/s
Memory BW-per-Domain (GROQ SRAM)	2 EB/s	100 EB/s
Tokens per Second	2M	700M

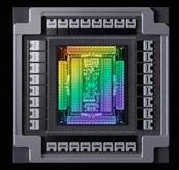

Rubin GPU


Vera CPU


CX9


BF4


NVLink Switch


Spectrum CPO


Groq 3 LPU


Vera Rubin Compute Tray


NVLink Switch Tray


Vera Compute Tray


BF4 STX Server


Spectrum Switch


Groq 3 Compute Tray

NVIDIA Extreme Co-Design Delivering X-Factors Every Year
From Chips to Racks to AI Factories

Develop Once, Deploy Anywhere

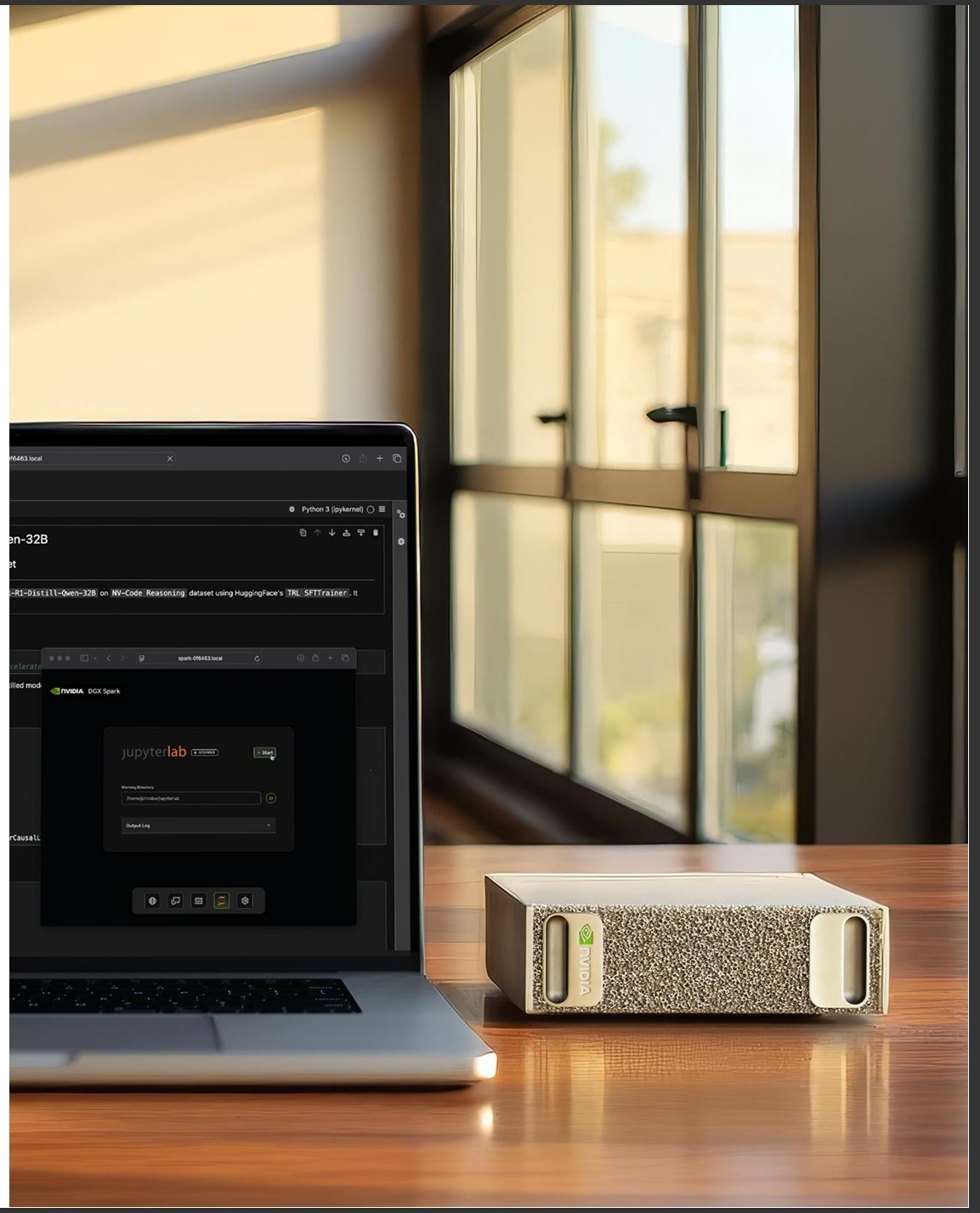
One Architecture – From Desk to Edge, Enterprise to Cloud

DGX Spark

Designed and build for AI

The ultimate harmony of cutting-edge, AI-optimized hardware & NVIDIA's industry-leading AI software stack

- Work with large workloads locally
  - Latest Generation Blackwell Architecture
  - 128GB Coherent Unified Memory
- Provides the required AI software stack
  - NVIDIA AI software stack
  - Extensive 3rd party ecosystem support

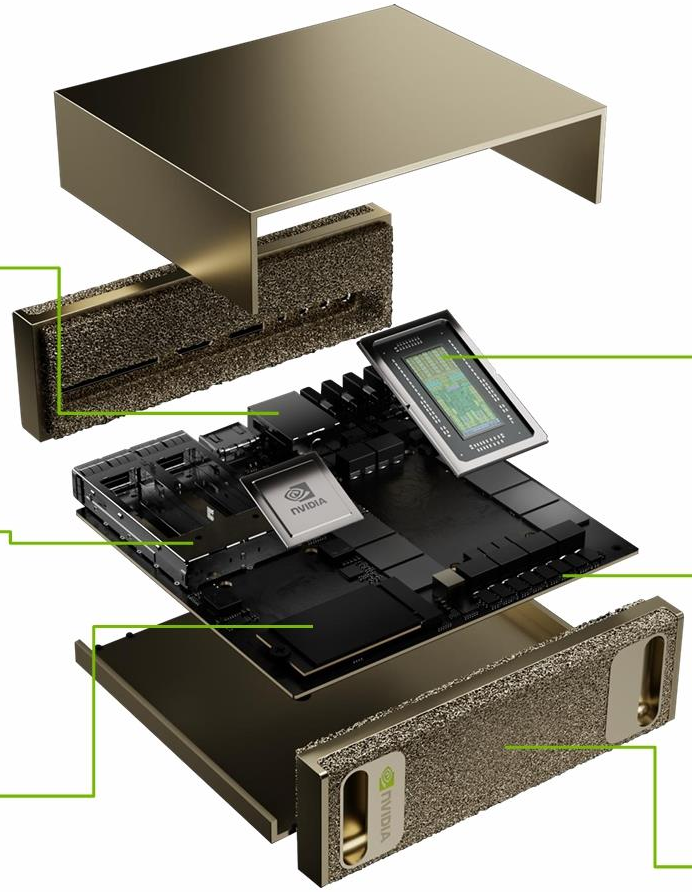


NVIDIA DGX Spark: Key Features

Connectivity
10Gb Ethernet
Wi-Fi 7, Bluetooth, USB

ConnectX 7
Cluster GB10 systems
100Gb Ethernet x 2

1 - 4TB SSD Storage
OEM option



NVIDIA GB10 Superchip
Latest Gen Blackwell GPU
20 Arm Cores
1 Petaflop FP4 AI Compute

High-Bandwidth Unified Memory
128 GB Low Power LPDDR5x

NVIDIA AI Software
Blueprints, NIMs, NEMO, & Toolkits etc.

NVIDIA DGX Base OS
Built on Ubuntu

DGX Station

The Ultimate Workstation
for AI and Data Science

- GB300 Superchip
- 784 GB Unified System Memory
- 20,000 AI TFLOPS
- ConnectX-8 SuperNIC

ASUS BOXX DELL Technologies hp Lambda

Develop Once, Deploy Anywhere

One Architecture – From Desk to Edge, Enterprise to Cloud

NVIDIA RTX PRO 6000 Blackwell Server Edition

The most powerful universal GPU for
AI and visual computing in the data center

Breakthrough Multimodal AI Inference

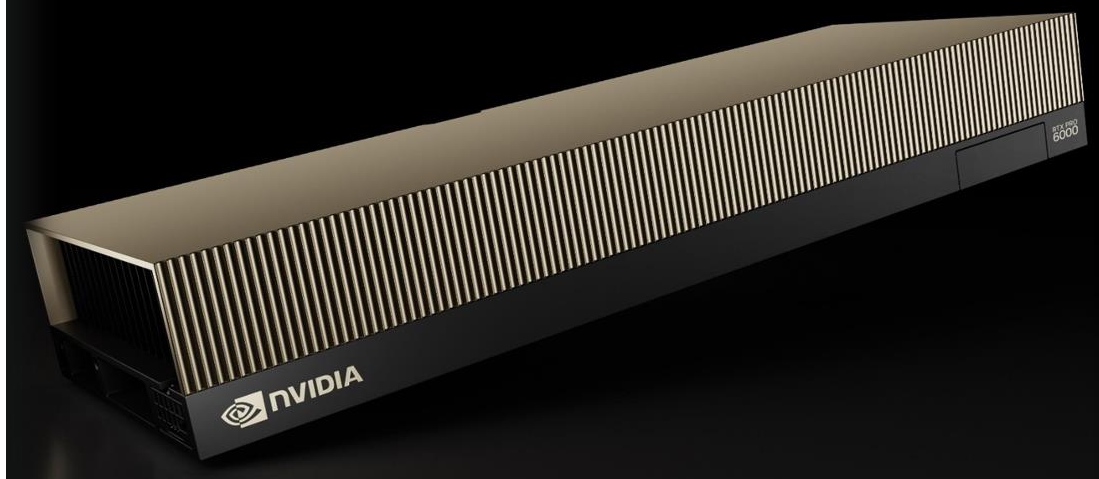
- 5th-Gen Tensor, 2nd-Gen Transformer Engine, FP4
- Full Media Pipeline: 4 NVENC/ NVDEC/ NVJPEG

Powerful Graphics and Visual Computing

- 4th-Gen RTX, Neural Shaders, DLSS 4
- vGPU Support, AI Virtual Workstations (vWS)

Data Center Ready

- 96GB GDDR7,1.6 TB/s Memory BW, 128MB L2 Cache
- Multi-Instance GPU (MIG), TEE Confidential Compute



RTX PRO 6000 Blackwell Server Edition GPU

Dual Slot, FHFL I Up to 600W

NVIDIA RTX PRO 4500 Blackwell Server Edition

Power-efficient NVIDIA Blackwell performance for any enterprise workload.

Universal Accelerator for AI Inference, Video, Data Processing

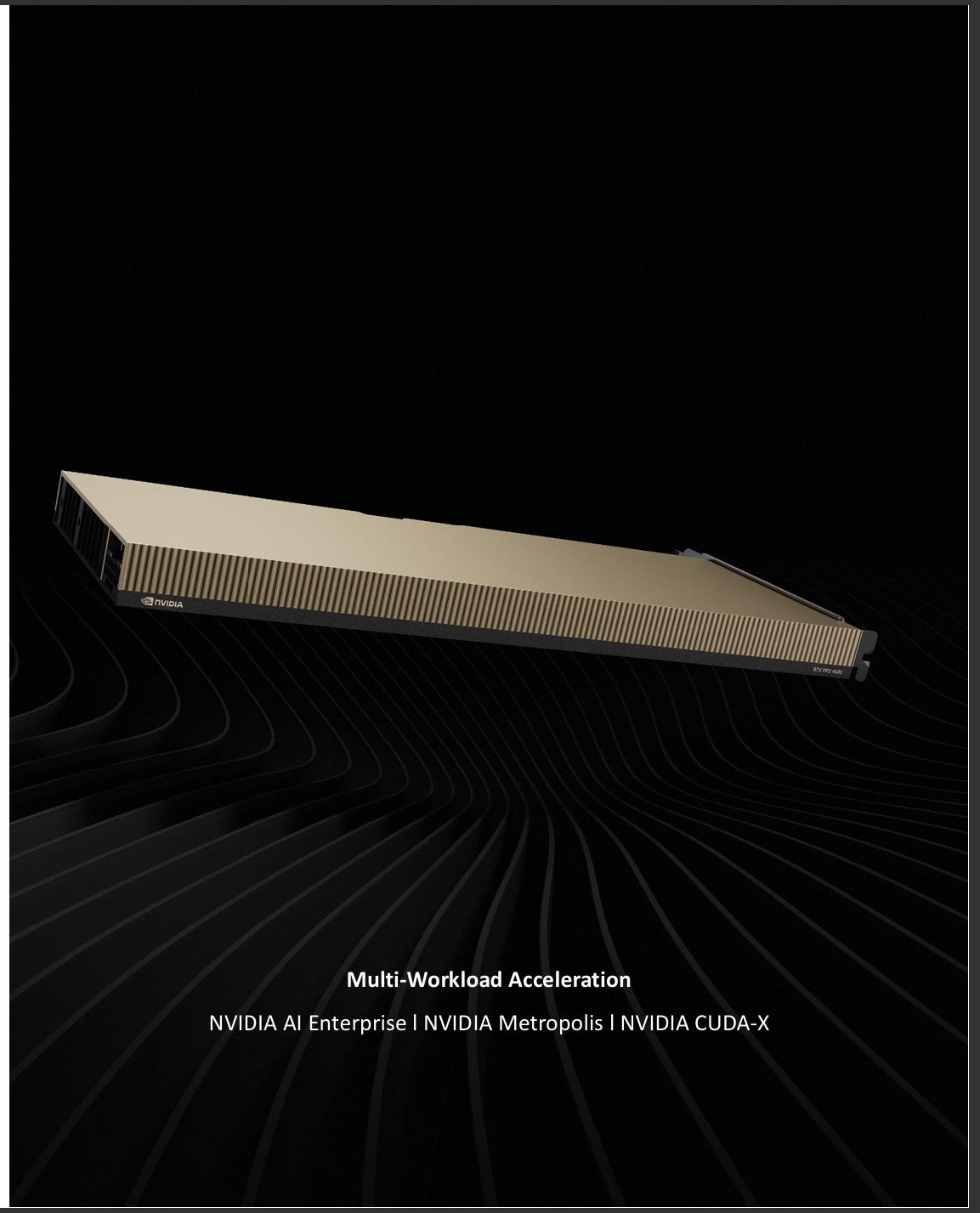
- 5th-Gen Tensor Cores, 2nd-Gen Transformer Engine, FP4
- 3 NVENC/NVDEC, Multi-Instance GPU (MIG)

Professional Graphics and Visualization

- 4th-Gen RT Cores, Neural Shaders, DLSS 4
- vGPU, AI Virtual Workstations, Visualization

Optimized for Mainstream Enterprise and Edge Servers

- 32GB GDDR7, 800 GB/s Bandwidth
- Single-Slot, 165W, FHFL. PCIe Gen 5



Multi-Workload Acceleration

NVIDIA AI Enterprise l NVIDIA Metropolis l NVIDIA CUDA-X

NVIDIA RTX PRO Blackwell Lineup

Designed for Multi-workload | Flexible CPU:GPU ratio

Highest Performance Universal Compute + Graphics

Omniverse Physical AI

vWS + PC Cloud Gaming

GenAI Image/Video

LLM Fine Tuning

Inference

FP32 HPC

7kW 8-GPU in < 10kW rack
4 nodes in 30kW rack
128 MIG users per rack

RTX PRO 6000
Blackwell Server Edition

600 | 96GB | 2S FHFL

Compact, Power-Efficient Universal Compute + Graphics

Data Processing + Analytics

Small Model Inference

GenAI

vWS + Mobile Cloud Gaming

Edge AI, Video Analytics

Transcoding, Video AI FX

3kW 8-GPU in < 5kW rack
10 nodes in 30kW rack
160 MIG users per rack

RTX PRO 4500
Blackwell Server Edition

165W | 32GB | 1S FHFL

NEW!

DGX B300

Accelerated Infrastructure for the Era of AI Reasoning

- Newest air-cooled DGX system with NVIDIA Blackwell Ultra GPUs
- All new system design seamlessly integrates into NVIDIA MGX or traditional enterprise racks
- **2.3TB of GPU memory**, enabling training and inference of complex models
- Equipped with NVIDIA ConnectX-8 high speed networking at **800Gb/s**
- Delivers **72 PFLOPS AI training** and **144 PFLOPS AI inference** performance
- Purpose-built platform for the era of AI reasoning, setting a new bar for LLM inference

10U Chassis | ~14 kW system
Designed for the modern data center

Open-Source Model Landscape

Frontier-class open weights, May 2026

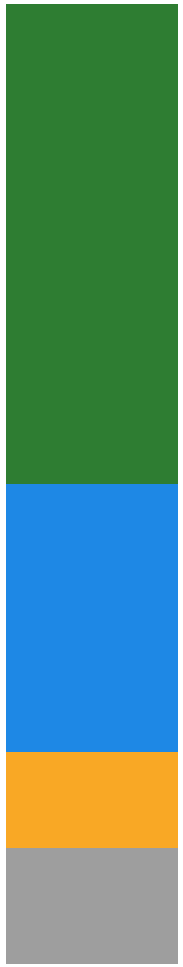
Model	Total	Active	License	Strengths
Nemotron 3 Nano	30B	3.5B	NVIDIA Open	Local agents, edge, 1M context
Gemma 4 26B	26B	dense	Gemma	Strong local default, 256K context
Llama 3.3 70B	70B	dense	Llama 3	Mature, broad ecosystem
gpt-oss 120B	120B	MoE	Apache 2.0	Math/reasoning, easy to deploy
Nemotron 3 Super	120B	12B	NVIDIA Open	Agentic AI, 1M context, FP4-native
Qwen 3 235B	235B	22B	Apache 2.0	Reasoning, multilingual, permissive
DeepSeek V4 Flash	284B	13B	DeepSeek	Cost-efficient reasoning, 1M context
Mistral Large 3	~123B	dense	Mistral Research	EU-friendly, strong coding
Llama 4 Scout	109B	17B	Llama 4	10M context window (extreme long-context)
Llama 4 Maverick	400B	17B	Llama 4	Frontier-class, multimodal
Nemotron 3 Ultra	~500B	~50B	NVIDIA Open	Frontier reasoning, enterprise agents
GLM-5.1	744B	40B	MIT	Long-horizon agentic coding, 200K context
Kimi K2.6	1T	32B	Modified MIT	Best open coding, 256K context
MiMo-V2.5 Pro	1.02T	42B	Xiaomi	Coding agents, 32K context
DeepSeek V4 Pro	1.6T	49B	DeepSeek	Maximum capability open-weight, 1M context

Workstation / RTX PRO · DGX Spark / Station · HGX B300 territory

Memory at FP8 ≈ 1 GB / 1B params · FP4 ≈ 0.5 GB / 1B

How VRAM Gets Used

Four buckets decide whether your model actually fits



Model weights

`Parameters × bytes/param`

FP16 = 2 B · FP8 = 1 B · FP4 = 0.5 B

KV cache

`≈ 2 × layers × KV-heads × head-dim × context × bytes × batch`

Grows with context length AND concurrent users

Activations + overhead

`~0.55 GB CUDA workspace + ≈8% of weights`

Real for large models, ignore at your peril

Headroom

`Reserve 10–15%`

Worked example

Llama 3.3 70B · FP8
32K context · 8 users

Weights	**70 GB**
KV cache (8 × 32K)	**~24 GB**
Activations + overhead	**~6 GB**
Headroom (15%)	**~15 GB**
Total needed	**~115 GB**

Fits on:

RTX PRO 6000 (96 GB)	tight
DGX Spark (128 GB)	**fits**
DGX Station (784 GB)	**easy**
HGX B300 (2.3 TB)	**+100 users**

KV cache is the hidden monster: serving 50 users at long context can quadruple your VRAM need vs single-user inference.

NVIDIA AI Developer Systems

Largest AI Model Size – FP4*

Extremely large

Data Center/ Cloud

~1T

DGX Station
(784GB)

600B

4x GPU Workstation
(384GB)

405B

2x DGX Spark**
(200GB)

200B

DGX Spark**
(100GB)

153B

1x GPU Workstation
(96GB)

51B

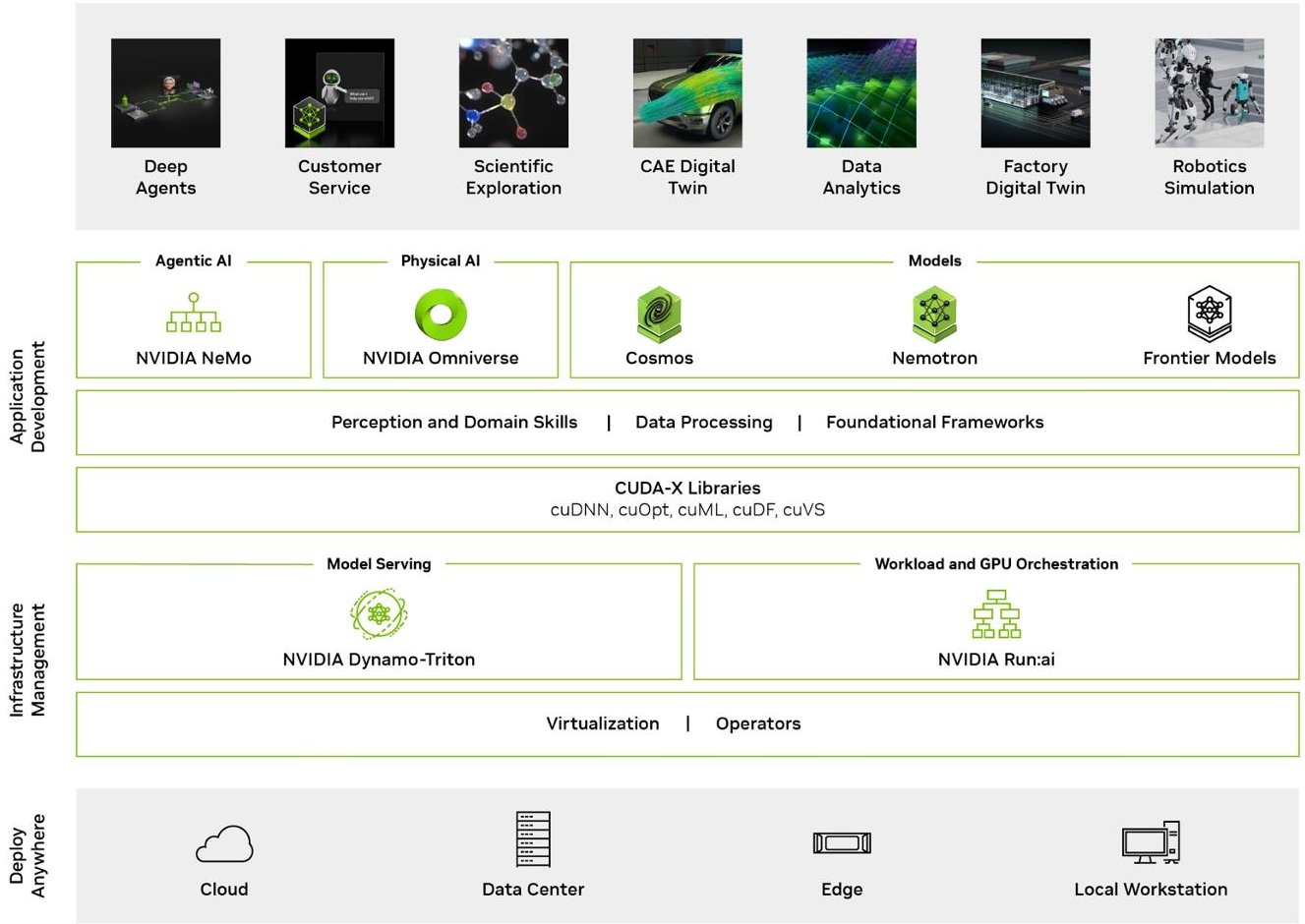
1x GPU PC
(32GB)

* Estimates based on available GPU memory in bytes * 2 for FP4 data minus 20% for overhead running model
**DGX Spark 128GB total system memory, 100GB available for user data

NVIDIA

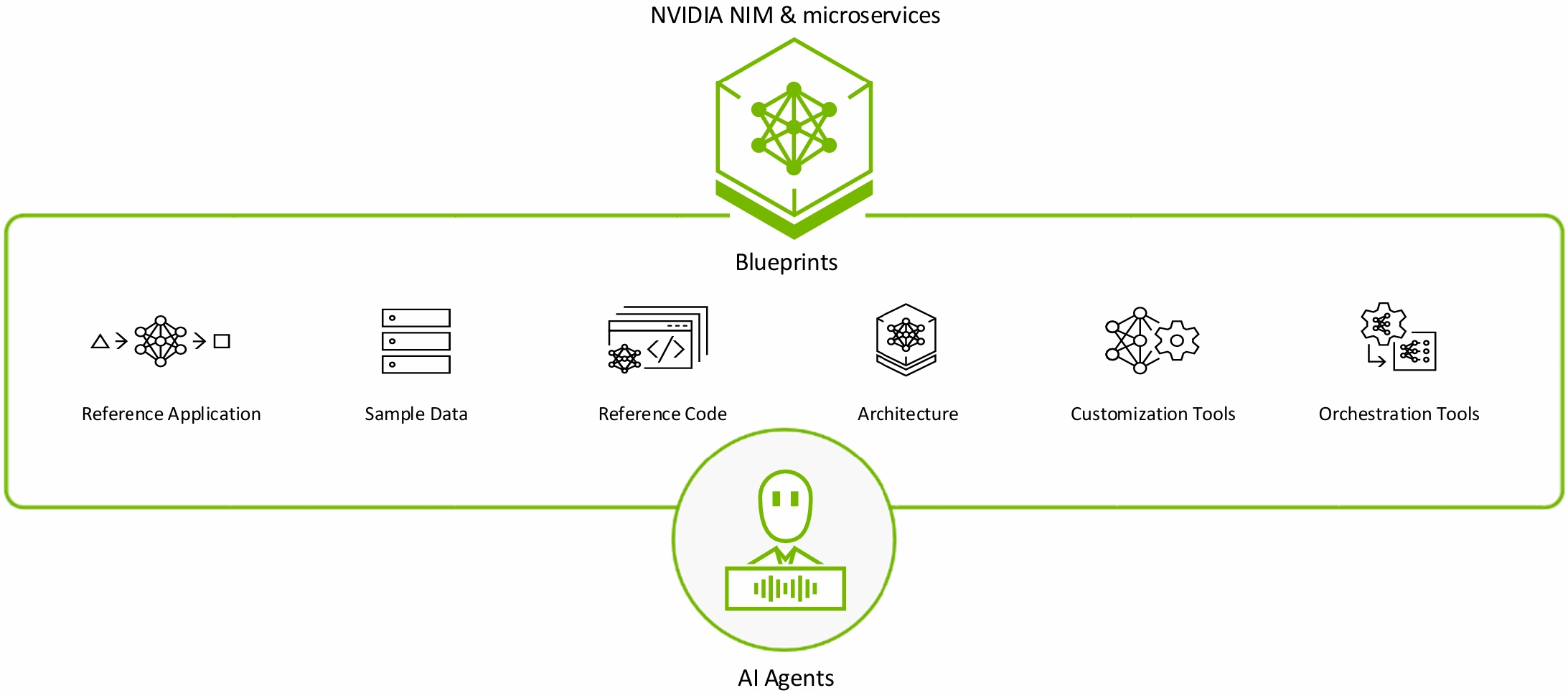
NVIDIA AI Enterprise

Cloud Native Software Platform for Production AI



NVIDIA Blueprints

Available on build.nvidia.com



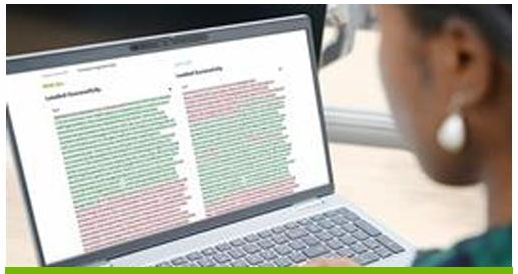
NVIDIA Blueprints

Available on build.nvidia.com

Digital Humans
for Customer Service

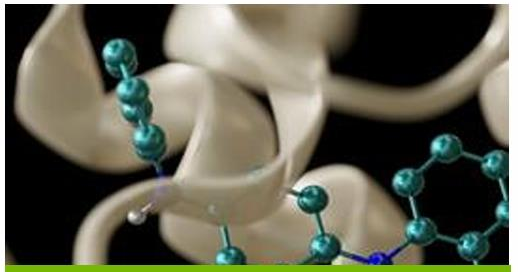
NVIDIA AI Blueprint

Multimodal PDF Data Extraction
for Enterprise RAG



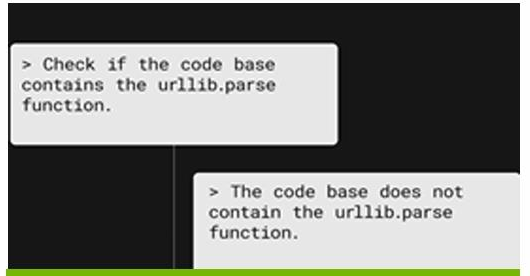
NVIDIA AI Blueprint

Generative Virtual Screening
for Drug Discovery



NVIDIA BioNeMo Blueprint

Vulnerability Analysis
for Container Security



NVIDIA AI Blueprint

AI Virtual Assistants
for Customer Service

NVIDIA AI Blueprint

Visual AI Agent
for Video Search and Summarization

NVIDIA AI Blueprint

3D Conditioning for
Precise Visual Generative AI

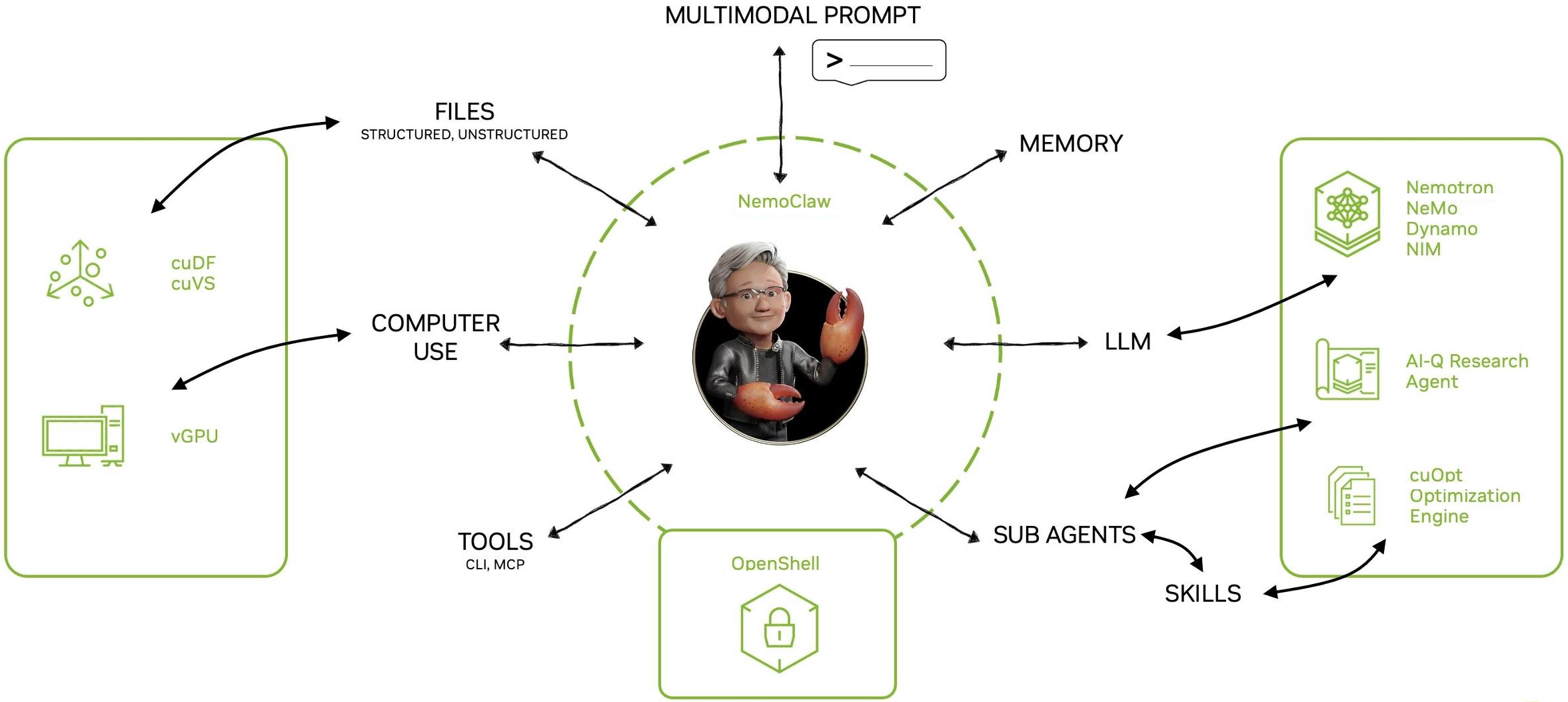
NVIDIA Omniverse Blueprint

Build a Digital Twin for
Interactive Fluid Simulation

NVIDIA Omniverse Blueprint

Announcing NVIDIA Agent Toolkit

NVIDIA Accelerates Autonomous Agents



THANK YOU!

HIER TREFFEN SIE UNS IN 2026

Boston TID

18. Juni 2026
Vaterstetten

DAS jährliche Boston Event
in der DACH-Region!
Mehr erfahren

Innovation erleben.
Zukunft gestalten.

ISC High Performance

22.-26. Juni 2026
Messe Hamburg

Boston mit Partnern
im Foyer 3, Z05.
Mehr erfahren

Treffen Sie unser Team und
unsere Technologiepartner.

GITEX Europe

30.-01. Juli 2026
Messe Berlin
Halle 1.2 |Stand 5

Die Boston wieder mit
eigenem Stand auf der
Gitex Europe

Wir sind mit vielen neuen
Technologien mit dabei.

ProspectX

08. Oktober 2026
Landshut

Die Boston ist vor Ort bei der
ProspectX in Landshut.

Jetzt Termin vormerken
zur prospectX 2026.

BOSTON
Servers | Storage | Solutions
A DIVISION OF SOURCECODE

NVIDIA Elite Partner

TRETEN SIE IN KONTAKT MIT UNS

Bei Fragen ist unser Team in der DACH-Region für Sie da:

- Deutschland: +49 (0) 89 9090 1993 | sales@boston-it.de (DE)
- Österreich: +43 660 2090400 | sales@boston-it.at (AT)
- Schweiz: +41 71 5542275 | sales@boston-it.ch (CH)

TILMAN STRÜBIG
Senior Solutions Architect AI
Boston IT
tilman.struebig@boston-it.de

ANGELIKA HARRER
Head of Marketing
Boston IT
angelika.harrer@boston-it.de