TEAMRGE EVENT 2024
WHERE FUTURE OF END USER COMPUTING MEETS REALITY

10+ community sessions around GPUs, VDI, DaaS, DEX, Remoting Protocols and AI

15th February 2024
16:00 CEST / 10:00AM EDT / 07:00AM PDT

Register Now

www.teamrge.com/events

Dr. Benny Tritsch
Managing Director at
Dr. Tritsch IT Consulting

Bram Wolfs
Consultant at
Wolfs IT Solutions

Eltjo van Gulik
Principal Product Manager
for HDX Graphics & Seamless
at Citrix

Esther Barthel
Solutions Architect
at Cognition IT

Joe DaSilva
PMTS, Solutions Architect, Cloud
Graphics at AMD

Johan van Amersfoort
Technologist EUC & AI
at ITQ

Magnar Johnson
Manager | Solution Architect
Sopra Steria

Rody Kossen
Senior Principal Quality
Engineer at Citrix

Ruben Spruijt
Field CTO
at Dizzion

Ryan Ververs-Bijkerk
Technical Evangelist
at GO-INIT

Shawn Bass
Start-up advisor and
former EUC CTO of Desktop
Technologies at VMware

Thomas Poppelgaard
Independent Consultant and
Technology Evangelist at
Poppelgaard.com

This FREE community event is made possible with support of:

DIZZION itq EUC Score

TeamRGE
Remoting Graphics Experts

TWO VOICES, ONE FUTURE
END USER COMPUTING, DAAS, AND GPU TRENDS FOR 2024

Dr. Benny Tritsch
Managing Director
at Dr. Tritsch IT Consulting

Ruben Spruijt
Field CTO at Dizzion

This FREE community event is made possible with support of:

TeamRGE
Remoting Graphics Experts

DIZZION

itq

EUC Score

AGENDA

1. What the Hex is happening in EUC?
2. GPU Evolution and the rise of AI
3. EUC from on-premises to cloud service
4. Remoting Protocol evolution
5. Shift from infrastructure to DEX

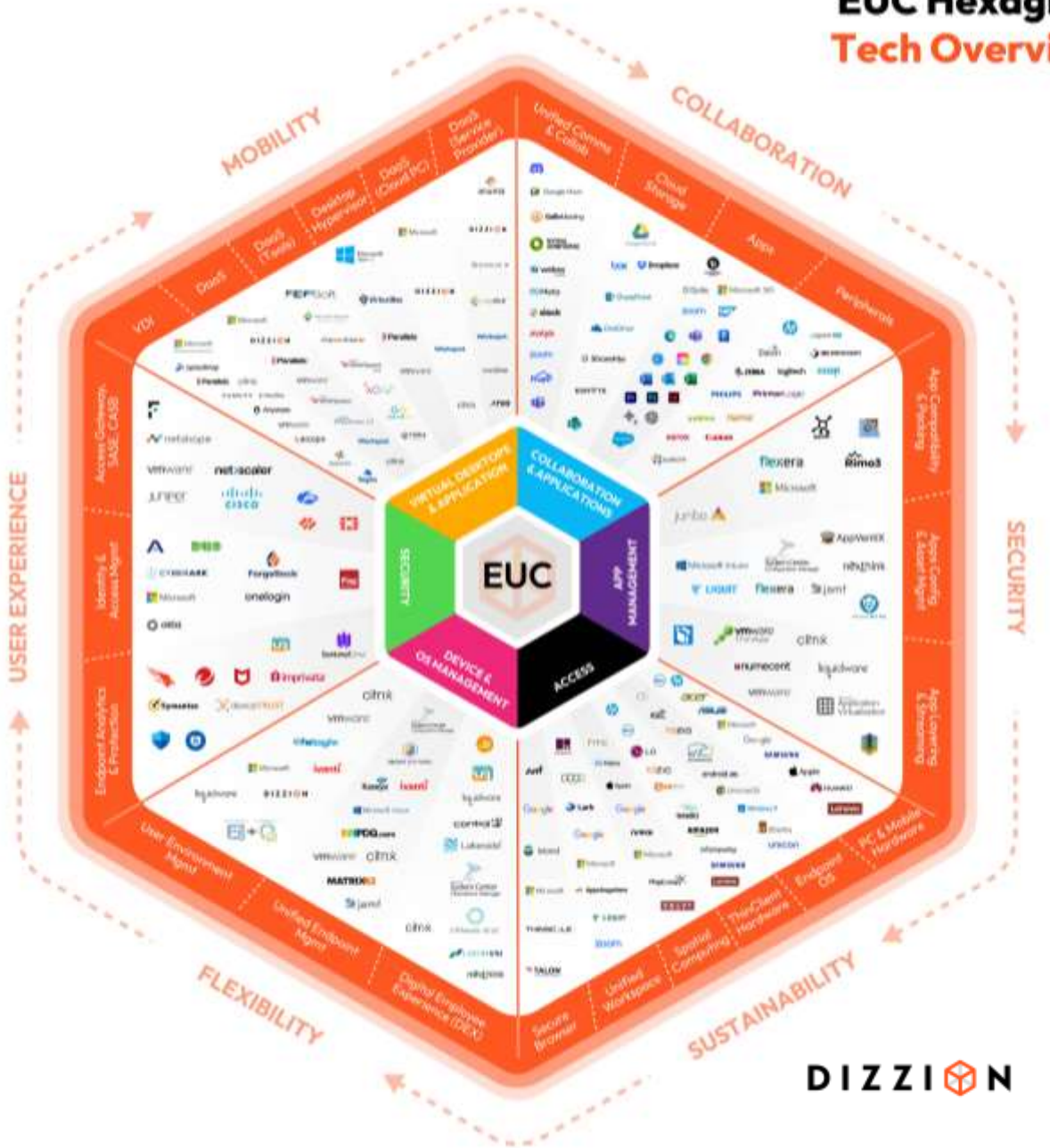

AI interpretation of 5 topics captured in 1 picture

ANNOUNCING

EUC HEXAGRID

6 MAJOR EUC CATEGORIES
24 SUB-CATEGORIES
220+ EUC VENDORS

EUC Hexagrid:
Tech Overview



TeamRGE
Remoting Graphics Experts

DIZZION

DOWNLOAD EUC HEXAGRID

https://www.dizzion.com/resources/euc-hexagrid

TeamRGE
Remoting Graphics Experts

WHAT THE HEX IS HAPPENING IN EUC?

- Generative AI in the Workspace
  - Buzz, excitement, and hype for sure
    - While GenAI isn't new…. 2022 for many was the "ChatGPT – Midjourney - aha" moment
    - We are in the middle of the hype; reality hasn't kicked in yet.
    - Vendor focus from GPU for 'EUC' moving to 'GPU for AI enterprise'. NVIDIA = GPU Hardware = 'AI Software' with 450 developer software building blocks

Text-to-Image

A photo of a cute cat with lots of Holi colors

Text-to-Video

Purple bioluminescent jellyfish swimming in space

https://blogs.nvidia.com/blog/chat-with-rtx-available-now/

TeamRGE
Remoting Graphics Experts

WHAT THE HEX IS HAPPENING IN EUC?

- Generative AI in the Workspace
  - OpenAI, Microsoft, Google, Meta, Slack, Zoom and many others blending GenAI into Workspace
    - "Co-Pilot for (almost) everything" – what is next? CoPilot for Notepad? ;-)
    - Many of us getting our hands dirty / using AI – e.g. https://www.youtube.com/watch?v=hXTZBGoqnJA (I am AI here).
    - Also me:
      - a Natural Language Interface for: Idea generation, prototyping, "co-pilot for software development, data analytics
      - content summarization, generating visualization, text-to-voice,
      - AI-generated avatars, unified comms 'buddy'.
    - AI – tools to get work done faster / most productive - efficient / more fun.
    - "still" need to know your stuff – otherwise, you end up with fluffy/shitty results.
    - GenAI in the business world is still narrow and tactical.
  - What are the real-world examples that provide real value? What are the real-use cases?!
  - What can the technology do?!

TeamRGE
Remoting Graphics Experts

WHAT THE HEX IS HAPPENING IN EUC?

- Azure Virtual Desktop – Cloud PC
  - From "niche" with multi-user to "mainstream' with Cloud PC; 100M versus 500+M TAM
  - AVD on Azure Stack HCI - adoption?!
  - Microsoft and artificial licensing boundaries - Win10/11 EMS on-premises and public clouds
  - Microsoft 365 Apps "Office 365 support" - Windows Server 2025 Multi-User
  - Monoculture in the making? – Identity – Management – IaaS – OS – Productivity - Collaboration
- Digital Employee Experience (DEX)
  - It is about employee satisfaction, measuring how good a user experience is in a digital world.
  - From IT-centric monitoring to user-centric monitoring and analytics. What is the perceived performance?
  - "I want to predict the problem before IT (and the end-user) sees it."

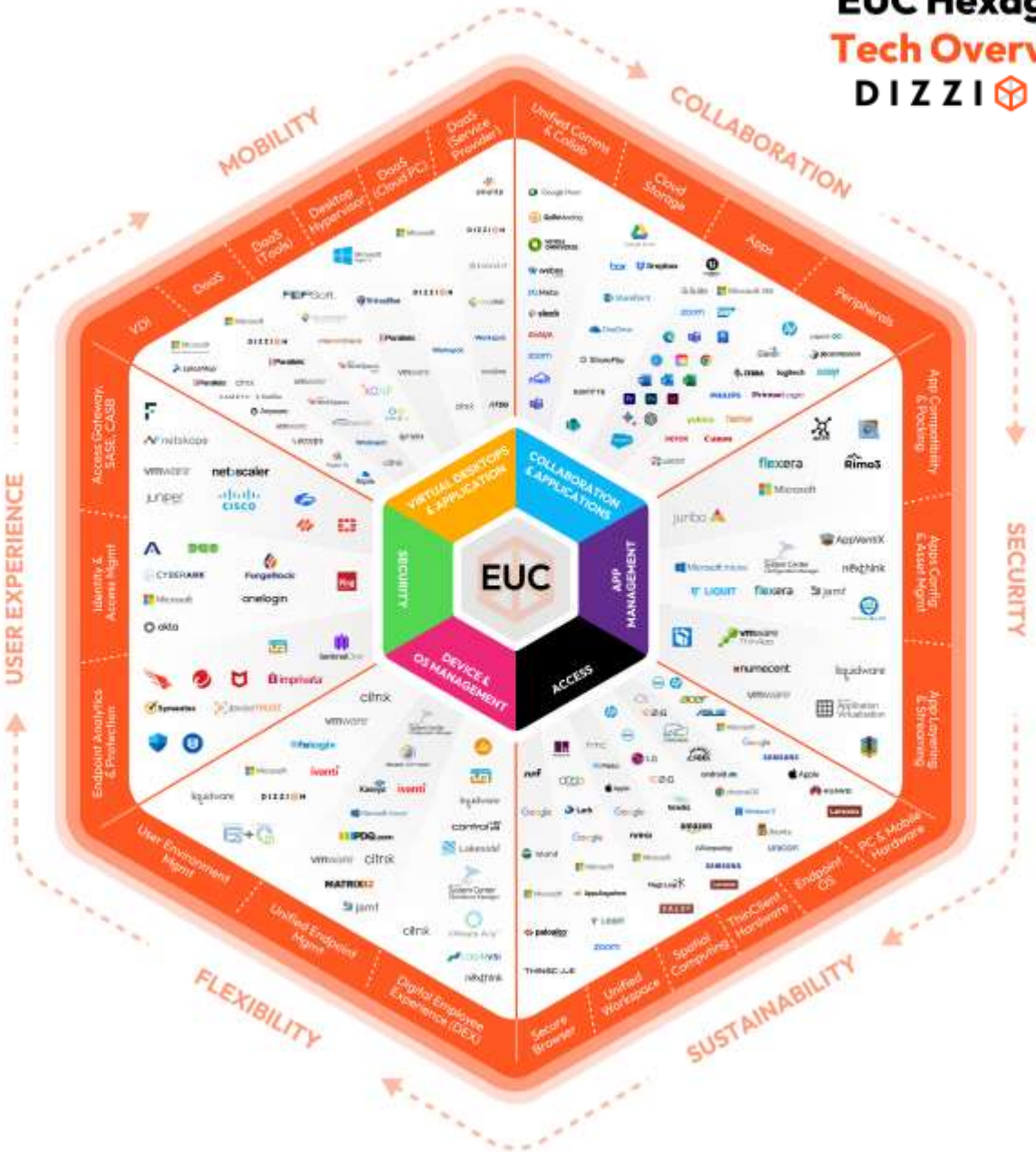
WHAT THE HEX IS HAPPENING IN EUC?

- Thin Clients - old wine in new bottles?
  - Revival of Thin Clients - Amazon Thin Client
  - From 'VDI/DaaS' jump board to Web/SaaS
  - Google ChromeOS & Windows compete
  - "RepurpOS" – hardware & sustainability
  - IGEL, UniCon Stratodesk, 10Zig, Zeetim –
- Virtual Apps and Desktops – stormy weather or perfect storm?
  - Citrix refocus – impact on customers, partners, and community
  - VMware > Broadcom - impact on customers, partners, and community
  - VMware EUC Business Unit will be 'Contoso' or 'ACME'
  - "VMware EUC" - "Horizon on Nutanix AHV"? ;-)
  - The rise of Microsoft and other DaaS solutions
  - Merger between Dizzion & Frame

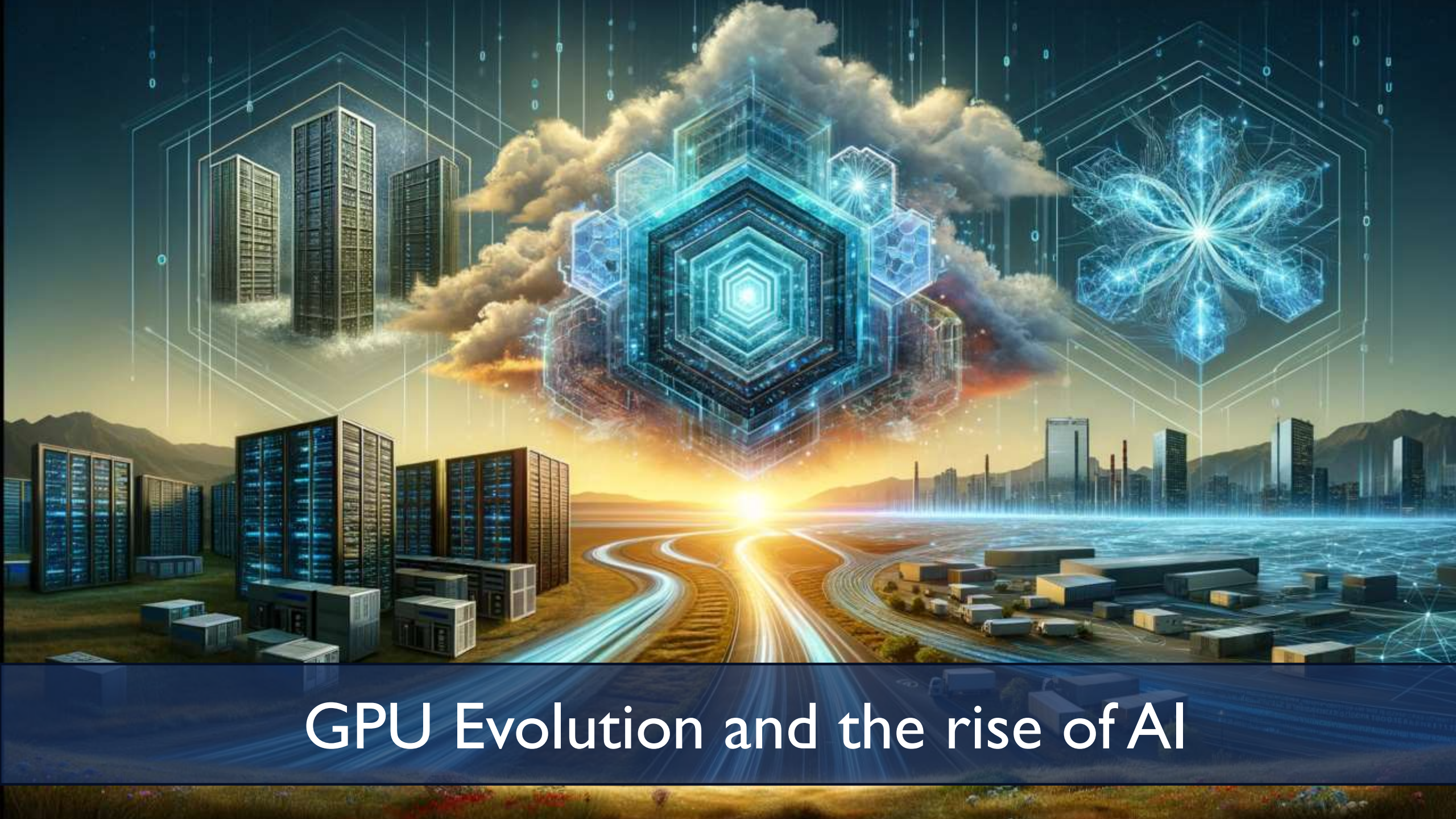
TeamRGE
Remoting Graphics Experts

EUC Hexagrid:
Tech Overview

DIZZION



TeamRGE
Remoting Graphics Experts

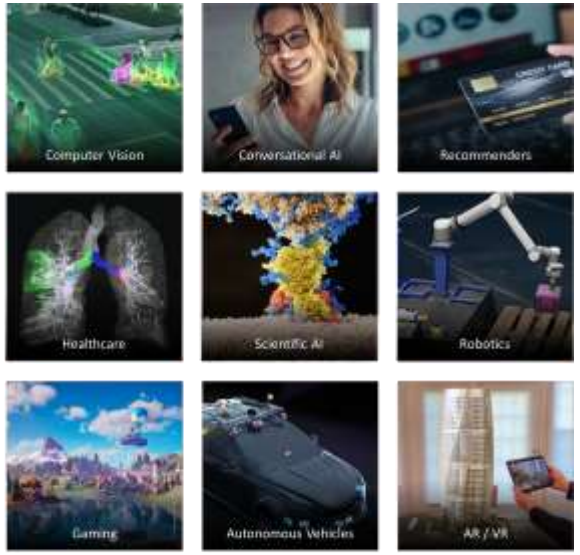

GPU Evolution and the rise of AI

GPU EVOLUTION AND THE RISE OF AI

- GPU evolution
  - More cost-effective GPU partitioning options in the public cloud – Azure leading here with NVIDIA (NVadsA10) and AMD (NGadsV620).
  - AWS and GCP lagging behind in 'GPU partitioning'.
  - GCP leading with newest 'Ada Lovelace' L4 GPUs (CPU's performance still is 'meh').
  - AWS G4ad (AMD V520) good cost/performance, AWS G5 (NVIDIA A10G) very good performance.
  - What about Intel GPU Flex140/170 – AMD Radeon V620 – NVIDIA in the data center!?
  - "Cloud workstations for CAD, BIM and visualization" – AEC Magazine 'Cloud Workstation Special Report'.

GPU EVOLUTION AND THE RISE OF AI

- Rise of AI
  - VDI by day AI at Night
  - Changing role of 'EUC IT Professional' > architect/technical/business w/ containers and container management?!
  - GPU vendors like NVIDIA – From Hardware Company to more 'sticky' Software powerhouse
    - 450 different SDK/Frameworks – building blocks for developers
    - Developer, start-up, research center, universities, point solutions, ISVs – overall, not the typical EUC partners
    - AI enterprise and Omniverse is focus

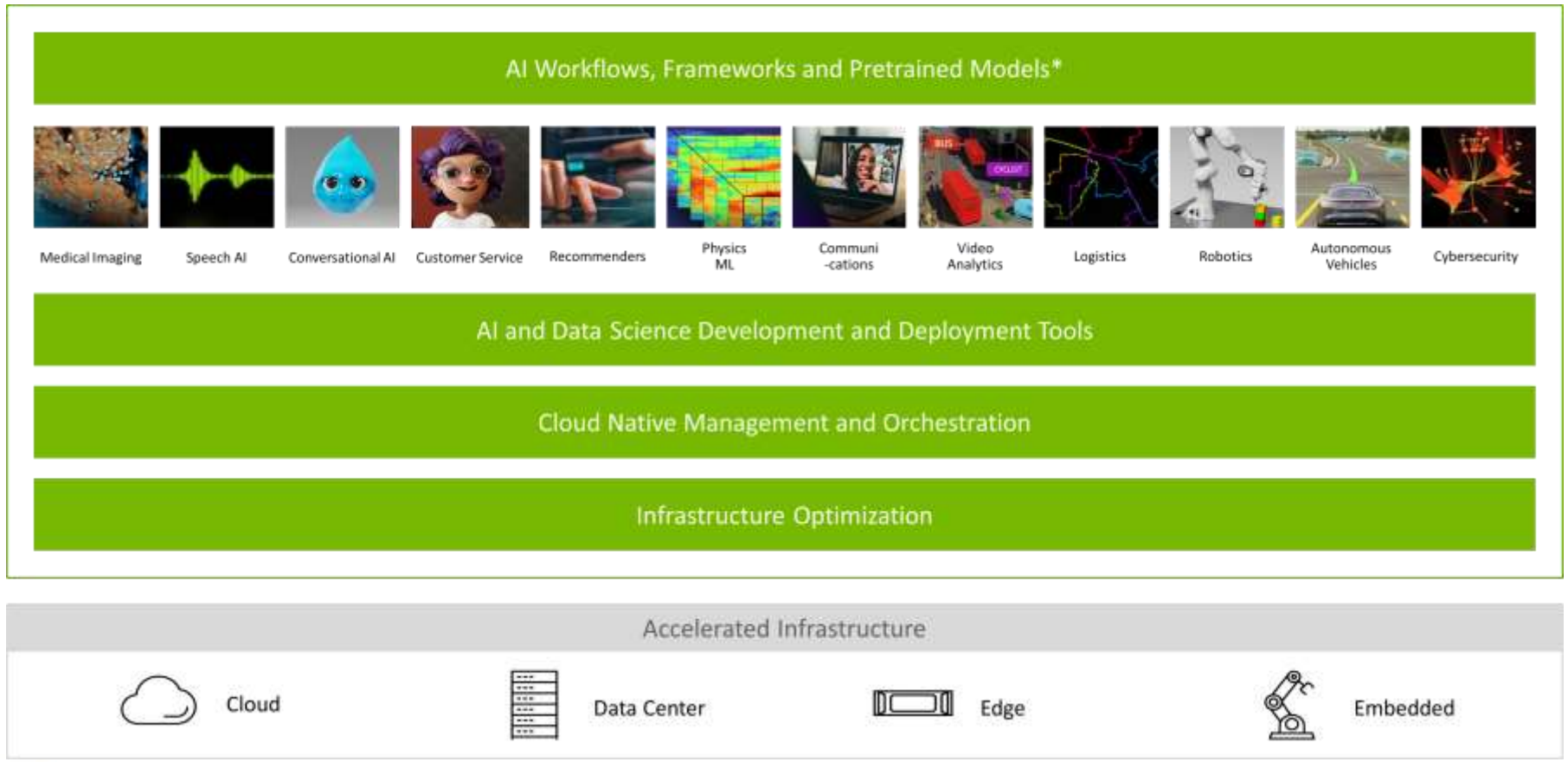


Transforming Industries

TeamRGE
Remoting Graphics Experts

NVIDIA AI ENTERPRISE

- End to End AI Software Includes Over 50 Frameworks and Pretrained Models



NVIDIA AI INFERENCE PLATFORM IN THE CLOUD

	aws	Google Cloud	Microsoft Azure	ORACLE CLOUD Infrastructure
NVIDIA Optimized Cloud Machine Images	• NVIDIA GPU optimized AMI • Free with option to purchase enterprise support through NVIDIA AI Enterprise	• NVIDIA Cloud Native Stack VMI • Free with option to purchase enterprise support through NVIDIA AI Enterprise	• NVIDIA GPU-Optimized VMI • Free with option to purchase enterprise support through NVIDIA AI Enterprise	• NVIDIA GPU Cloud Machine Image • Free with option to purchase enterprise support through NVIDIA AI Enterprise
NVIDIA Triton Inference Server	• Available on AWS Deep Learning Containers • Deploy with SageMaker Python SDK, AWS CLI ,Boto3 • Supports AWS SageMaker Multi-Model Endpoint (MME) API Contract	• Vertex AI Prediction supports deploying models on Triton running on custom NGC container • Deploy Triton as a containerized microservice on GKE managed cluster using One-Click Triton Inference Server App for GKE	• Deploy using Azure CLI, Python SDK v2, and Azure ML studio (specify Type as "triton_model" in YAML deployment file) • Supports No-code deployment in managed online endpoints and Kubernetes online endpoints	• Seamless integration with OCI Data Science's model deployment (pass Triton as env variable) • Push Triton image to OCI Container Registry and save the model to the model catalog • Use with OCI software developer kits (SDKs), APIs, or the Oracle Cloud Console
NVIDIA GPU Powered Cloud Instances	• NVIDIA H100 | P5 • NVIDIA A100 | P4D • NVIDIA V100 | P3 • NVIDIA A10G | G5 • NVIDIA T4(G) | G4, G5g • NVIDIA H200 | Coming Soon • Just announced GH200, L40S, L4	• NVIDIA H100 | A3 • NVIDIA L4 | G2 • NVIDIA A100 | A2 • NVIDIA V100 | N1 • NVIDIA T4 | N1 • NVIDIA H200 | Coming Soon	• NVIDIA H100 | ND H100 v5, NCads H100 v5 • NVIDIA A100 | ND, NC v4 • NVIDIA V100 | NC v3, ND v2 • NVIDIA A10 | NVadsA10 v5 • NVIDIA T4 | NCasT4_v3 • NVIDIA H200 | Coming Soon	• NVIDIA H100 | BM.GPU.H100.8 • NVIDIA A100 | GPU.A100 • NVIDIA A10 | VM.GPU.A10 • NVIDIA V100 | VM.GPU3 • NVIDIA L40S | Coming soon • NVIDIA H200 | Coming Soon • NVIDIA GH200 | Coming Soon

TeamRGE
Remoting Graphics Experts

NVIDIA VGPU DATACENTER GPU SOLUTIONS

NVIDIA A16

Knowledge Worker VDI
w/ NVIDIA Virtual PC
Entry RTX vWS

Office Productivity,
streaming video

NVIDIA L4

Graphics-rich VDI with vPC,
Entry to Mid RTX vWS

Medium size/complexity
CAD models, Basic DCC,
Medical Imaging, PLM

Enterprise Acceleration for
Graphics, Analytics, Inference

NVIDIA L40/L40S

High-End RTX vWS

Large/complex CAD models,
CAE, Seismic exploration,
complex DCC effects, rendering,
3D Medical Imaging Recon

NVIDIA A30

Mainstream Virtual Compute
Inferencing

AI inferencing at scale,
high-performance computing

NVIDIA A100

High-End Virtual Compute

Deep Learning Training,
HPC, AI, Data Science

TeamRGE
Remoting Graphics Experts

NVIDIA DATA CENTER GPUS

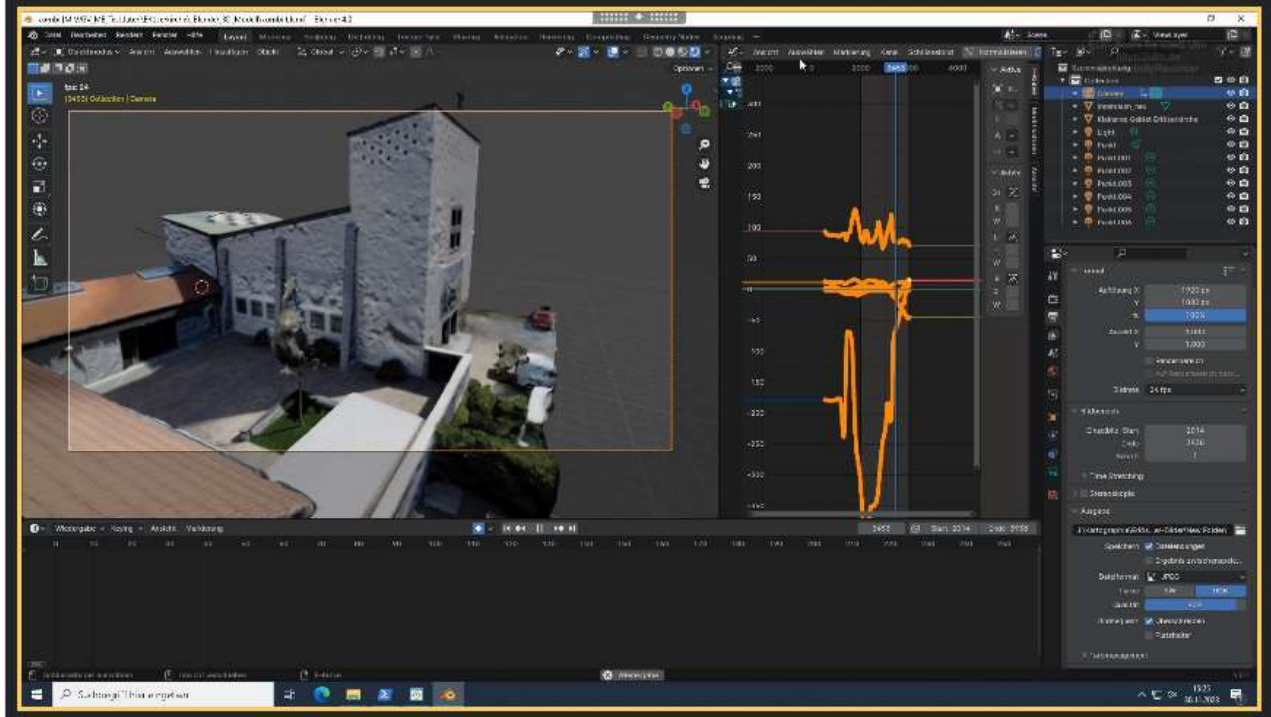
	H200	H100		L40S	L40	L4	A100		A30	A40	A10	A16	A2
Design	Highest Perf AI, HPC, DA	Mainstream Training and HPC		Highest Perf Universal	Powerful Graphics + AI	Universal AI, Video, and Graphics	High Perf Compute		Mainstream Compute	High Perf Graphics	Mainstream Graphics & Video with AI	High Density Virtual Desktop	Entry-Level Small Footprint
Form Factor	SXM5	SXM5	PCIe Gen5 x16 2 Slot FHFL 3 NVLink Bridge	NVL 2x PCIe Gen5 x16 2x 2 Slot FHFL NVLink Bridged	PCIe Gen4 x16 2 Slot FHFL	PCIe Gen4 x16 1 slot LP	SXM4	PCIe Gen4 x16 2 Slot FHFL 3 NVLink Bridge	PCIe Gen4 x16 2 Slot FHFL 1 NVLink Bridge	PCIe Gen4 x16 2 Slot FHFL 1 NVLink Bridge	PCIe Gen4 x16 1 slot FHFL	PCIe Gen4 x16 2 Slot FHFL	PCIe Gen4 x8 1 Slot LP
Max Power	700W	700W	350W	400W	350W	300W	72W	500W	300W	165W	300W	150W	250W	60W
FP64 TC | FP32 TFLOPS[1]	67 | 67	67 | 67	51 | 51	134 | 134	NA | 91.6	NA | 90	NA | 30	19.5 | 19.5	19.5 | 19.5	10 |10	NA | 37	NA | 31	NA | 4x4.5	NA | 4.5
TF32 TC | FP16 TC TFLOPS[2]	989 | 1979	989 | 1979	756 | 1513	1979 | 3958	366 | 733	181 | 362	120 | 242	312 | 624	312 | 624	165 | 330	150 | 300	125 | 250	4x18 | 4x36	18 | 36
FP8 TC | INT8 TC TFLOPS/TOPS[2]	3958 | 3958	3958 | 3958	3026 | 3026	7916 | 7916	1466 | 1466	724 | 724	485 | 485	NA | 1248	NA | 1248	NA | 661	NA | 600	NA | 500	NA | 4x72	NA | 72
GPU Memory	141GB HBM3 e	80GB HBM3	80GB HBM2e	188GB HBM2e	48GB GDDR6	48GB GDDR6	24GB GDDR6	80GB HBM2e	80GB HBM2e	24GB HBM2	48GB GDDR6	24GB GDDR6	4x 16GB GDDR6	16GB GDDR6
Multi-Instance GPU (MIG)	Up to 7	Up to 7	Up to 7	Up to 14	-	-	-	Up to 7	Up to 7	Up to 4	-	-	-	-
NVLink Connectivity	Up to 256	Up to 256	2 cards	2 cards	-	-	-	Up to 8	2 cards	2 cards	2 cards	-	-	-
Media Acceleration	7 JPEG Decoder 7 Video Decoder	7 JPEG Decoder 7 Video Decoder	7 JPEG Decoder 7 Video Decoder	14 JPEG Decoder 14 Video Decoder	3 Video Encoder 3 Video Decoder 4 JPEG Decoder	3 Video Encoder 3 Video Decoder 4 JPEG Decoder	2 Video Encoder3 4 Video Decoder3 4 JPEG Decode	1 JPEG Decoder 5 Video Decoder	1 JPEG Decoder 5 Video Decoder	1 JPEG Decoder 4 Video Decoder	1 Video Encoder 2 Video Decoder (+AV1 decode)	1 Video Encoder 2 Video Decoder (+AV1 decode)	4 Video Encoder 8 Video Decoder (+AV1 decode)	1 Video Encoder 2 Video Decoder (+AV1 decode)
Ray Tracing	-	-	-	-	-	Yes	Yes	-	-	-	Yes	Yes	Yes	Yes
Transformer Engine	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-	-	-	-	-	-	-
DPX Instructions	Yes	Yes	Yes	Yes	-	-	-	-	-	-	-	-	-	-
Graphics	For in-situ viz (no vPC or RTX vWS)	For in-situ viz (no vPC or RTX vWS)	For in-situ viz (no vPC or RTX vWS)	For in-situ viz (no vPC or RTX vWS)	Top-of-Line	Top-of-Line	Better	For in-situ viz (no vPC or RTX vWS)	For in-situ viz (no vPC or RTX vWS)	For in-situ viz (no vPC or RTX vWS)	Best	Better	Good	Good
vGPU	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hardware Root of Trust	Internal and External	Internal and External	Internal and External	Internal and External	Internal and External	Internal	Internal with Option for External	Internal with Option for External	Internal with Option for External	Internal with Option for External	Internal with Option for External	Internal with Option for External	Internal with Option for External	Internal with Option for External
Confidential Computing	Yes	Yes	Yes	Yes	-	-	-	(1)	-	-	-	-	-	-
NVIDIA AI Enterprise	Add-on	Add-on	Included	Included	Add-on	Add-on	Add-on	Add-on	Add-on	Add-on	Add-on	Add-on	Add-on	Add-on

1. *Supported on Azure NVIDIA A100 with reduced performance compared to A100 without Confidential Computing or H100 with Confidential Computing.*
2. *All Tensor Core numbers with sparsity. Without sparsity is ½ the value.*
3. *Includes AV1 in addition to H.265, H.264, VP9, VP8, MPEG4*

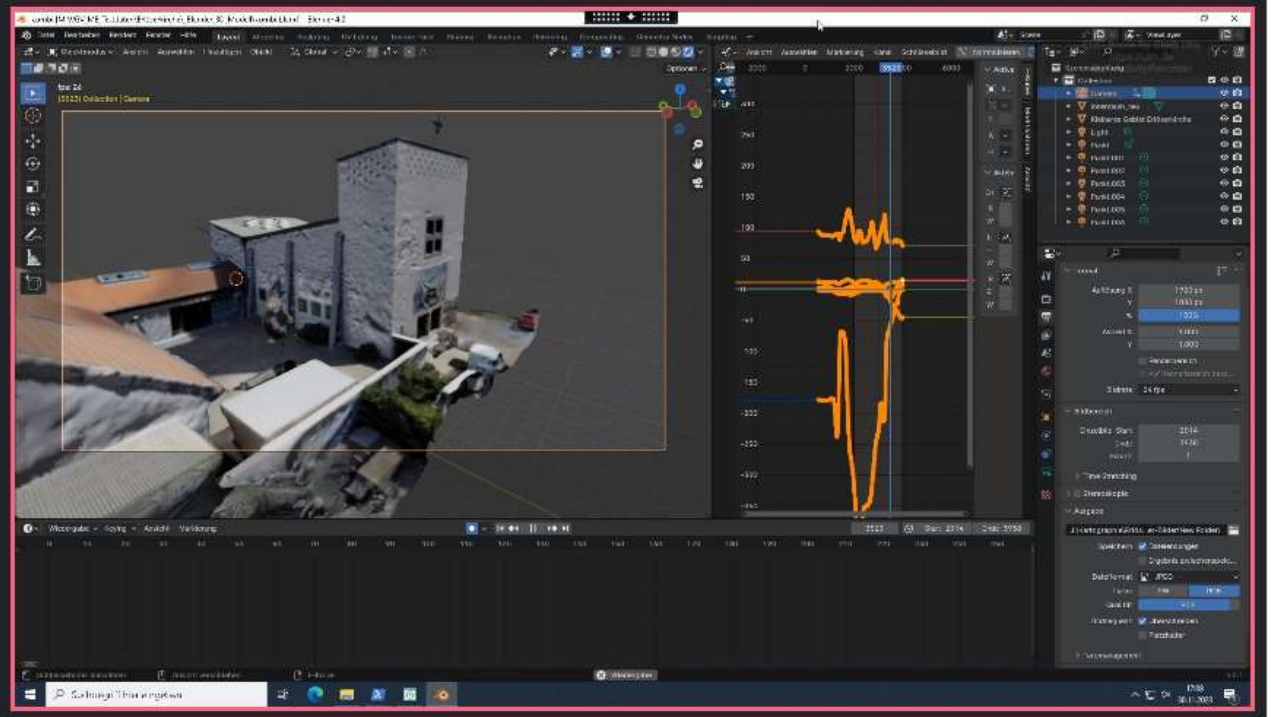
WHAT THE HEX IS HAPPENING IN EUC?

TeamRGE
Remoting Graphics Experts

SLX-Blender on Citrix Session | Windows 10 Enterprise | RTX 4000 | HDX 100MB/s



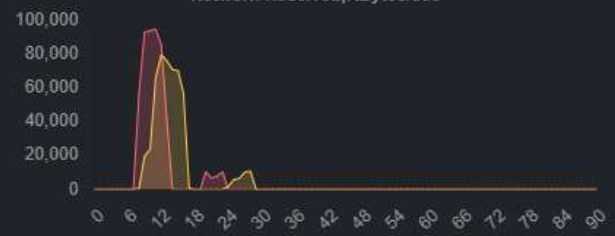
SLX-Blender on Citrix Session | Windows 10 Enterprise | RTX 4000 | HDX 12MB/s 100ms 0.1%



CPU|%

Memory Available|MBytes

Network Received|KBytes/sec



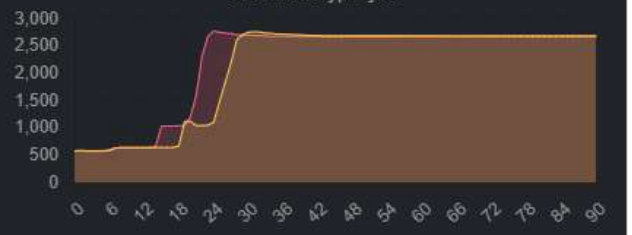
Network Sent|KBytes/sec

GPU 3D|%

GPU Video Decode|%

GPU Video Processing|%

GPU Memory|Mbytes

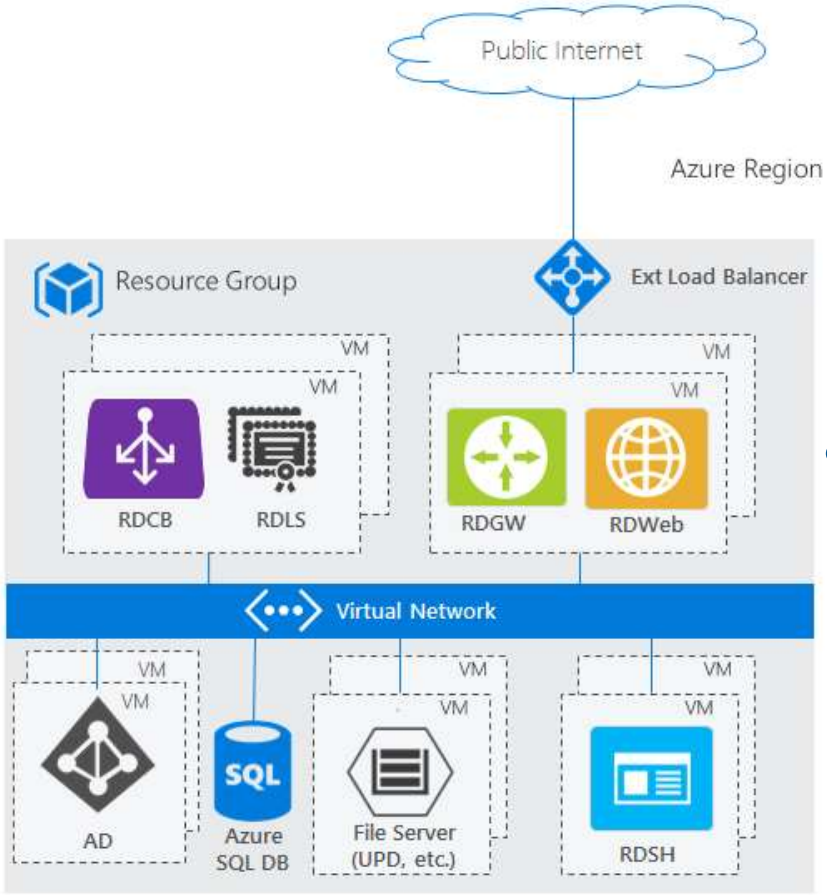


00:01:29 00:01:29 Help Report

FROM PHYSICAL TO ON-PREM EUC TO CLOUD DAAS – AND BACK…

TeamRGE
Remoting Graphics Experts

SHIFT FROM INFRASTRUCTURE TO PERCEIVED USER EXPERIENCE



TeamRGE
Remoting Graphics Experts

REMOTING PROTOCOL IMPROVEMENTS

- Self-adaptive when network conditions change
- Reduced impact of latency and packet loss
- Advanced caching
- Modern video codecs
- UDP-enabled
- GPU-enabled on sender and on receiver side
- Allowing reverse connect

Microsoft Remote Desktop Protocol

Amazon Workspace Streaming Protocol

Teradici/HP PCoIP

Citrix ICA/HDX

VMware Blast

Parsec

Frame Remoting Protocol

VNC Remote Framebuffer Protocol

TeamRGE
Remoting Graphics Experts

COMBINATION OF RENDERING/ENCODING AND AI

- GPU-enabled virtual/remote desktops with specific capabilities designed to run generative AI and ML tasks locally
- Adds another use case to high-end rendering, video encoding, mining and gaming
- Copilot example: "Paint an image that combines GPU-enabled 3D rendering and machine learning"

TeamRGE
Remoting Graphics Experts

THANK YOU

Ruben Spruijt
Field CTO at Dizzion
ruben@dizzion.com

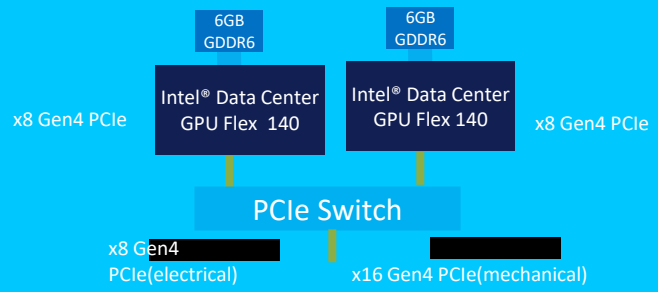
Dr. Benny Tritsch
Managing Director
at Dr. Tritsch IT Consulting
benny@drtritsch.com

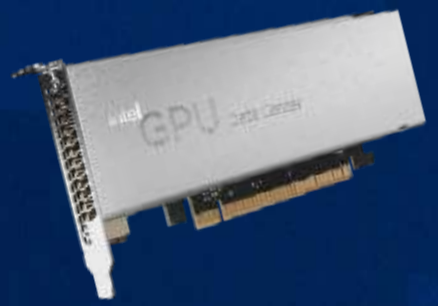
TeamRGE
Remoting Graphics Experts
WWW.TEAMRGE.COM

This FREE community event is made possible with support of:

DIZZION itq EUC Score

Intel® Data Center GPU Flex 140 Card Overview

Card Design	Intel® Data Center GPU Flex 140 
Card TDP	Board Power: 75W
Card Specifications	Half height, half length, single-wide, Passive cooling
GPU	Intel® Data Center GPU Flex 140
GPU's Per Card	2
Memory w/ECC	Capacity: 12GB (6GB/GPU) Mem xfer Rate: 1750GT/s Mem Bus Width: 96 bits/GPU
Fixed Function Media Units (Per Card)	4 (2 per GPU): 28 transcode streams H.265 1080p60 1:1
Supported Usecases	Media transcode, Visual Inference/Media Analytics, Mobile Cloud Gaming, PC Cloud Gaming, VDI
GPU Throughput (Peak)	• FP32: 8.0 TOPs • FP16: 52 TOPs • INT8: 105 TOPs • INT4: 210 TOPs
Product Availability	3 years*
Operating System	Linux: Ubuntu, CentOS, Debian, RHEL Windows: WinServer 2019 & 2022, WinClient
Host CPU Support	3rd Gen and 4th Gen Intel® Xeon® Scalable Processors
Branding/Channel Partners	Intel Branded Card

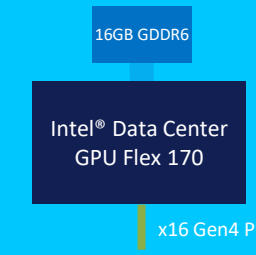


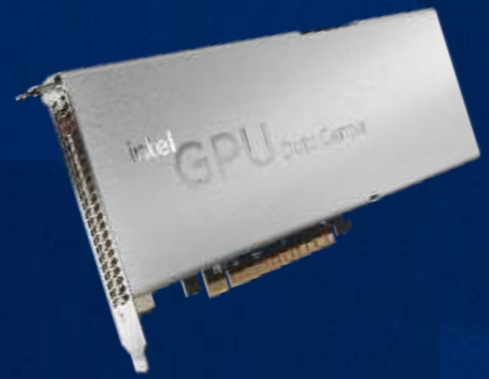
Built for **high density,** multipurpose use cases

- Optimized for lower TCO
- 2nd Gen Xe Media Engine with AV1 delivers 30%-60% bit rate savings
- Up to 62 virtual functions using HW SR-IOV with no SW licensing fee

*/ From launch date Sep'22

Intel® Data Center GPU Flex 170 Card Overview

	Intel® Data Center GPU Flex 170
Card Design	
Card TDP	Board Power: 150W
Card Specifications	Full Height, ¾ length, single-wide, Passive cooling
GPU	Intel® Data Center GPU Flex 170
GPU's Per Card	1
Memory w/ECC	Capacity: 16GB Mem xfer Rate: 2250GT/s Mem Bus Width: 256 bits
Fixed Function Media Units (Per Card)	2 (2 per GPU): 14 transcode streams H.265 1080p60 1:1
Supported Usecases	Media transcode, Visual Inference/Media Analytics, Mobile Cloud Gaming, PC Cloud Gaming, VDI
GPU Throughput (Peak)	• FP32: 16.8 TOPs • FP16: 128 TOPs • INT8: 256 TOPs • INT4: 512 TOPs
Product Availability	3 years*
Operating System	Linux: Ubuntu, CentOS, Debian, RHEL Windows: WinServer 2019 & 2022, WinClient
Host CPU Support	3rd Gen and 4th Gen Intel® Xeon® Scalable Processors
Branding/Channel Partners	Intel Branded Card



Built for multipurpose use cases requiring **maximum peak performance**

- AAA Gaming support with Ray Tracing
- Compute up to 500 TOPs for visual inference and media analytics workloads
- Up to 31 virtual functions using HW SR-IOV with no SW licensing fee

*/ From launch date Sep'22