Virtual Desktops, GPUs, and Things You Didn't Know You Could Do

Tony Foster (WonderNerd)

- Sr. Principal Engineering Technologist – Dell Technologies
- Adjunct Instructor of Computer Systems at K-State
- vExpert, NVIDIA vGPU Community Advisor, Omnissa Tech Insider
- Virtualized since 2005 (ESX 2.0)
- Find out more at: wondernerd.net
- **X:** @wonder_nerd **LinkedIn:** linkedin.com/in/wondernerd

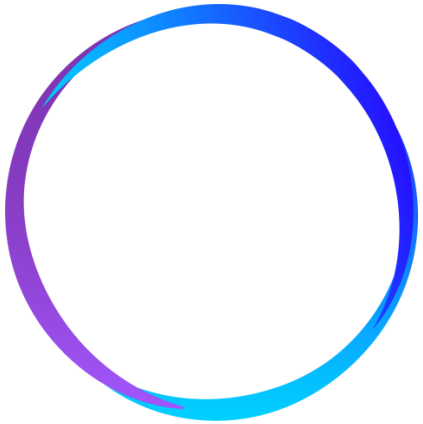
vmware by Broadcom vEXPERT

omnissa TECH INSIDER 2024

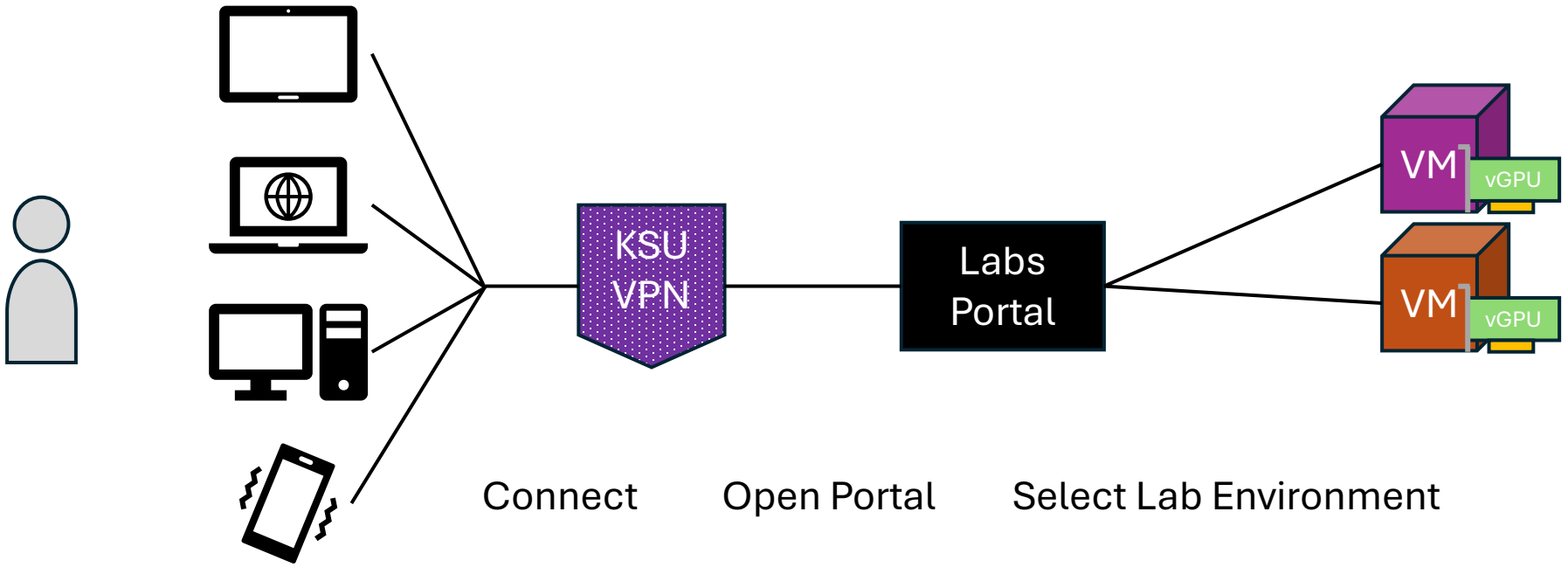
NVIDIA COMMUNITY ADVISOR

Agenda

- AI in VDI
- Picking the Right GPU for the Job
- New Fangled Workloads (GenAI, RAG, LLM, SLM, etc.)
- Consumption Models
- Why VDI?
- VDI Superpowers for Mild Mannered Data Scientists
- Trying Things Out
- Resources



What Does It Look Like?



The Right GPU for the Job

- Data Center Grade GPUs
- Display Heads: The A100 VS A10
  - AI != graphics
  - VDI = display graphics

NVIDIA A100

	A100 80GB PCIe	A100 80GB SXM
FP64	9.7 TFLOPS	
FP64 Tensor Core	19.5 TFLOPS	
FP32	19.5 TFLOPS	
Tensor Float 32 (TF32)	156 TFLOPS | 312 TFLOPS*	
BFLOAT16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
FP16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
INT8 Tensor Core	624 TOPS | 1248 TOPS*	
GPU Memory	80GB HBM2e	80GB HBM2e
GPU Memory Bandwidth	1,935GB/s	2,039GB/s
Max Thermal Design Power (TDP)	300W	400W***
Multi-Instance GPU	Up to 7 MIGs @ 10GB	Up to 7 MIGs @ 10GB
Form Factor	PCIe dual-slot air cooled or single-slot liquid cooled	SXM
Interconnect	NVIDIA® NVLink® Bridge for 2 GPUs: 600GB/s ** PCIe Gen4: 64GB/s	NVLink: 600GB/s PCIe Gen4: 64GB/s
Server Options	Partner and NVIDIA-Certified Systems™ with 1-8 GPUs	NVIDIA HGX™ A100-Partner and NVIDIA-Certified Systems with 4,8, or 16 GPUs NVIDIA DGX™ A100 with 8 GPUs

* With sparsity

** SXM4 GPUs via HGX A100 server boards; PCIe GPUs via NVLink Bridge for up to two GPUs

*** 400W TDP for standard configuration. HGX A100-80GB CTS (Custom Thermal Solution) SKU can support TDPs up to 500W

NVIDIA A10

FP32	31.2 TF
TF32 Tensor Core	62.5 TF | 125 TF*
BFLOAT16 Tensor Core	125 TF | 250 TF*
FP16 Tensor Core	125 TF | 250 TF*
INT8 Tensor Core	250 TOPS | 500 TOPS*
INT4 Tensor Core	500 TOPS | 1000 TOPS*
RT Cores	72
Encode / Decode	1 encoder 2 decoders (+AV1 decode)
GPU Memory	24 GB GDDR6
GPU Memory Bandwidth	600 GB/s
Interconnect	PCIe Gen4: 64 GB/s
Form Factor	1-slot FHFL
Max TDP Power	150W
vGPU Software Support	NVIDIA vPC/vApps, NVIDIA RTX™ vWS, NVIDIA AI Enterprise
Secure and Measured Boot with Hardware Root of Trust	Yes (optional)
NEBS Ready	Level 3
Power Connector	PEX 8-pin

*with sparsity

https://resources.nvidia.com/en-us-gpu-resources

Where Does VDI Fit In?

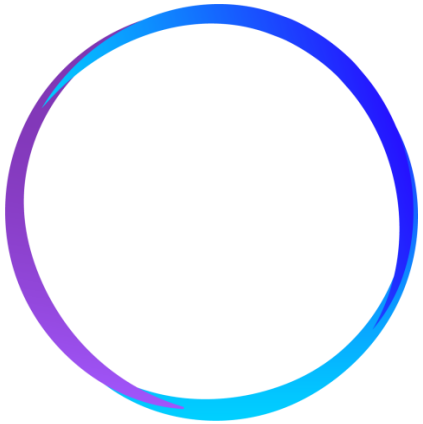
- To access a VM with a vGPU

CAUTION:

Output from the VM console is not available for VMs that are running vGPU. Make sure that you have installed an alternate means of accessing the VM (such as a VNC server) before you configure vGPU.

https://docs.nvidia.com/ai-enterprise/latest/user-guide/index.html

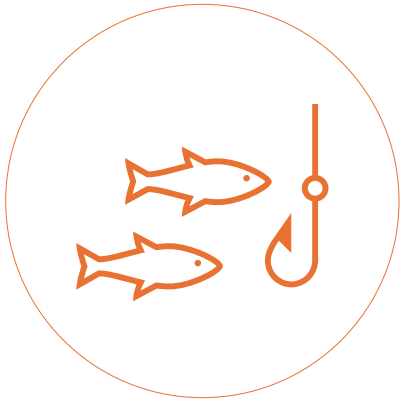
- Consolidation of resources
- Standardization of IT resources
- Ease of access / Access Controls



New Fangled Workloads

Customer Support & Chatbots

Coding Assistants



Spear Phishing Detection

Marketing Support

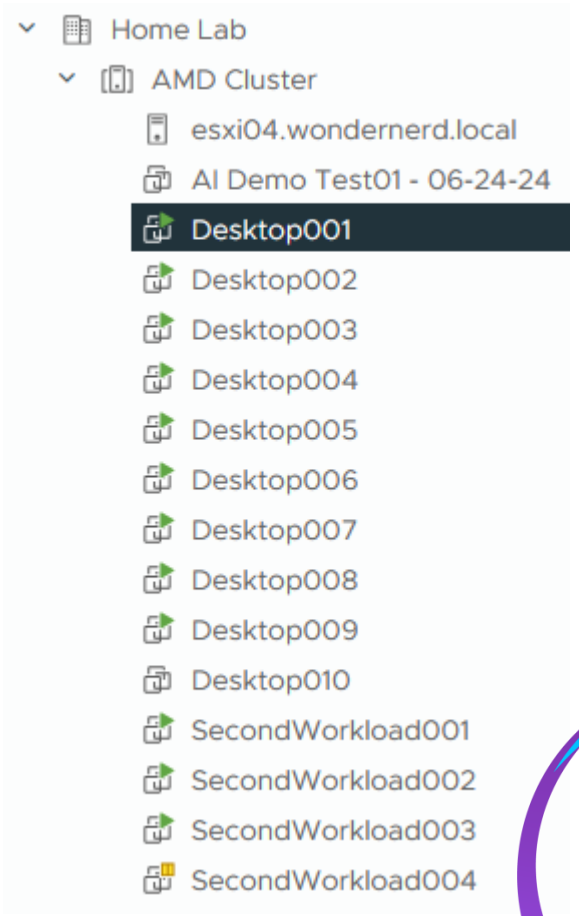
Consumption Models – Containers

Licensed through: Getty Images

- Trained Models
- Purchased/OS Models
  - NVIDIA NIM: build.nvidia.com
  - NVIDIA NGC: catalog.ngc.nvidia.com

Consumption Models – VMs

- Model Training & Development
  - NVIDIA Triton
  - NVIDIA Riva
  - NVIDIA Morpheus
- Education



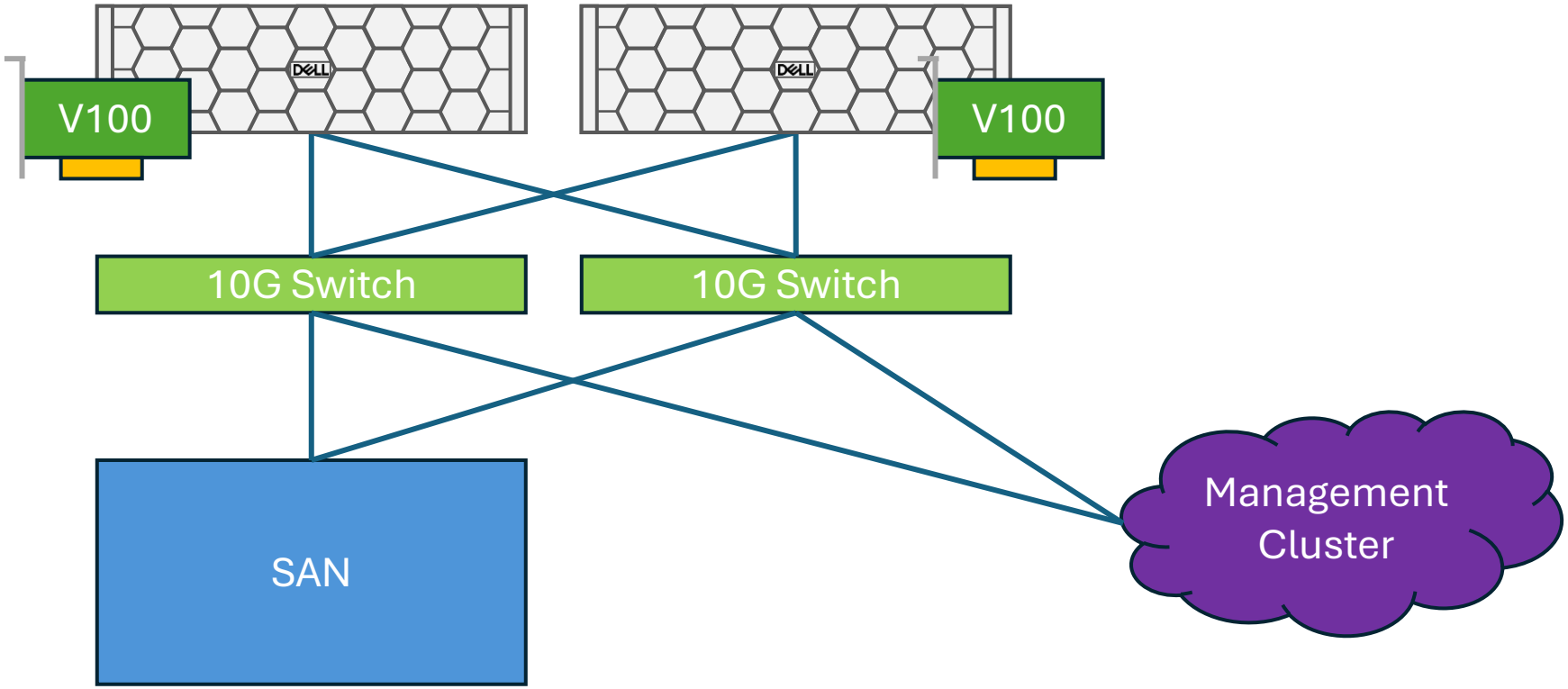
VDI Superpowers for Data Scientists

- Provisioning
- Repeatability
- Disaster Recovery

Generated in Adobe Photoshop, Prompt:
Computer super hero cartoon data scientist with a purple cape in the data center

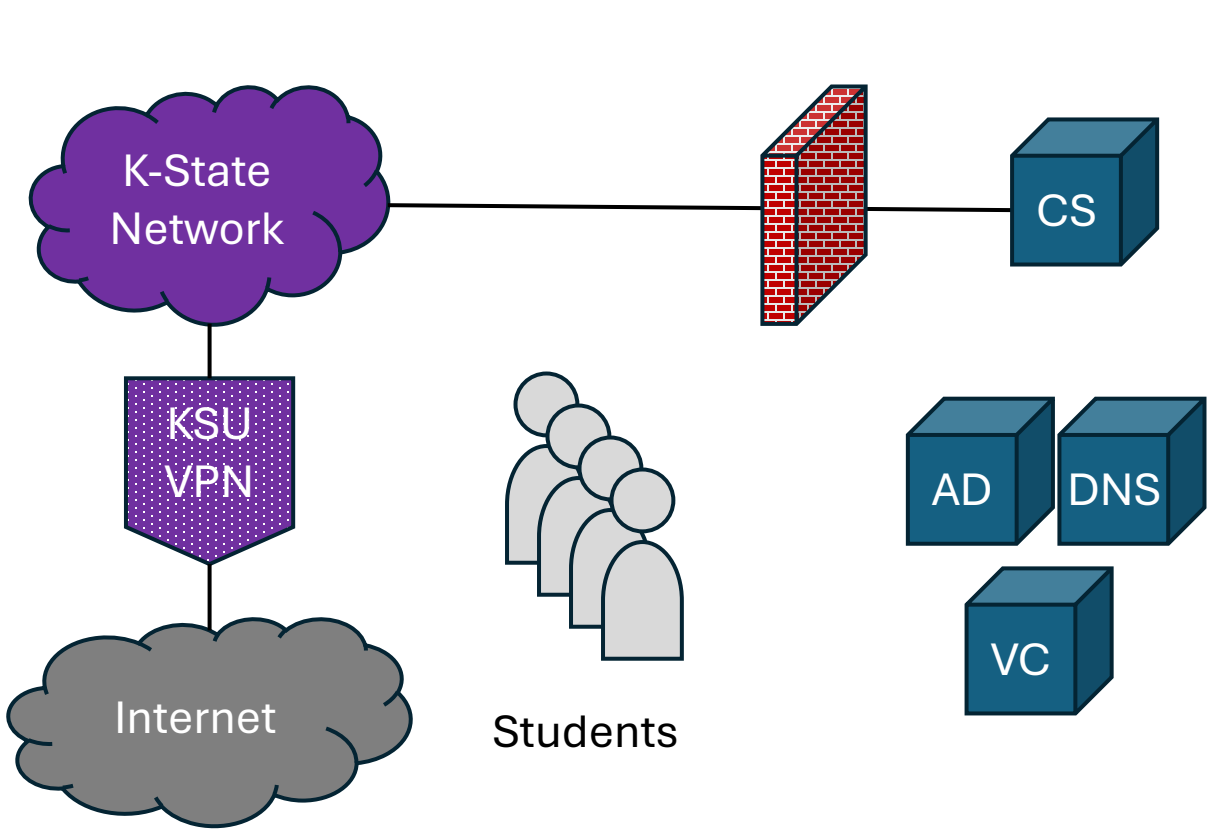
What Does It Look Like? Part 1

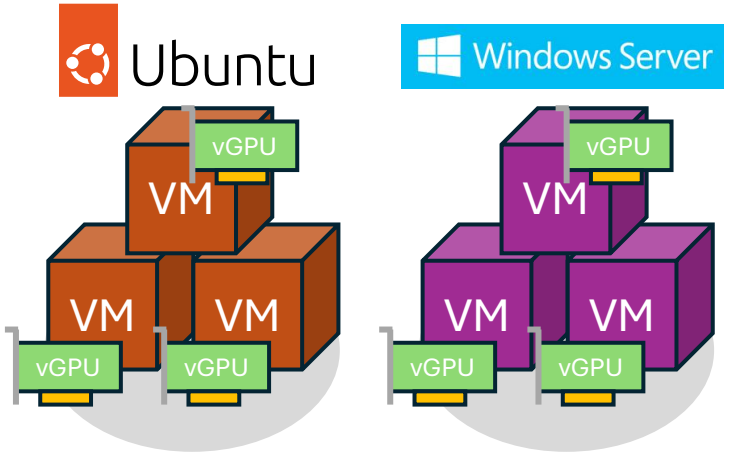
- Physical Architecture – VMTN3081LV

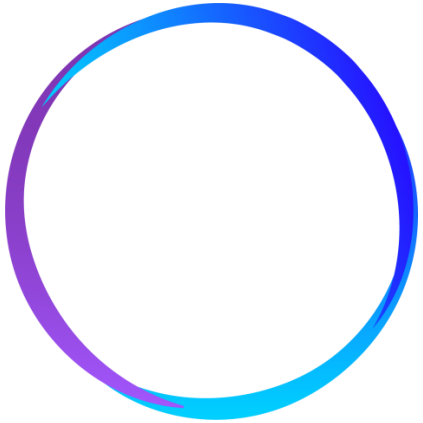


What Does It Look Like? Part 2

- Logical Architecture – VMTN3081LV

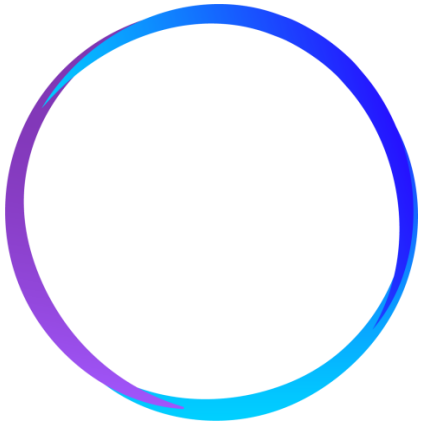






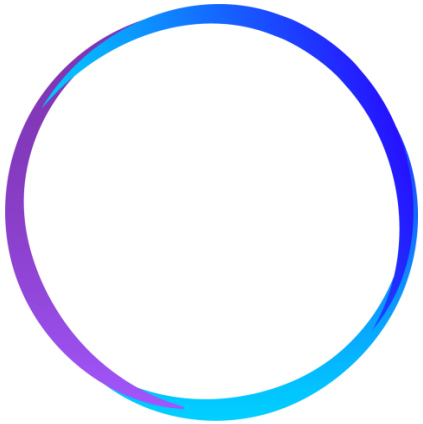
Trying Things Out

- NVIDIA NIM: build.nvidia.com
- NVIDIA GPU Cloud (NGC): ngc.nvidia.com
- NVIDIA Test Drive: https://www.nvidia.com/en-us/data-center/data-center-gpus/gpu-test-drive/



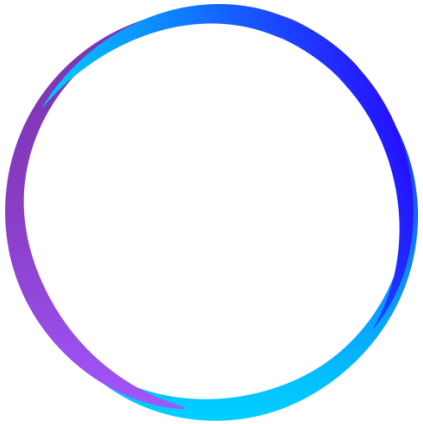
Resources

- Wondernerd.net
  - VMware Explore 2023 VMTN3081LV – AI on the Horizon: Delivering Virtualized AI Environments to Students
  - VMware Explore 2024 – Unlocking the Magic of Gen-AI in VMware VCF: Where Dreams Meet Reality
- NVIDIA AI Enterprise Quick Start Guide: https://docs.nvidia.com/ai-enterprise/latest/quick-start-guide/index.html
- VMware Private AI Foundation with NVIDIA Guide: https://docs.vmware.com/en/VMware-Cloud-Foundation/5.2/vmware-private-ai-foundation-nvidia/



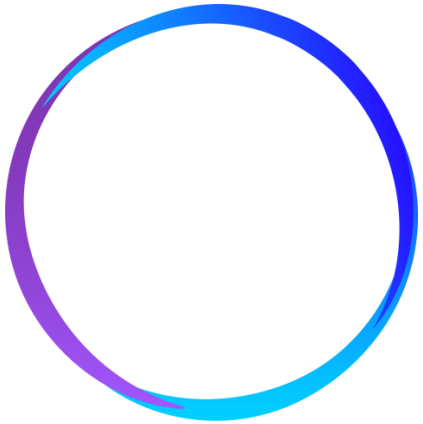
Resources

- VMware Explore Content:
  - HOL: Accelerate Machine Learning in vSphere Using GPUs [SPL2521LV]
  - CMTY1709LV - Building cost-effective homelabs for AI/ML workloads https://www.youtube.com/live/k_6Z5dYUnpE?si=xyYNLcs_LkBZyNDb



Bringing It All Together

- GPUs, More Than VDI
- Use the Right GPU
- Containers, VMs, Both
- Access the VM
- VDI Superpowers



VMUG VIRTUAL

Thank you for attending!

Slides available at:
wondernerd.net