TeamRGE

Remoting Graphics Experts

GPU Smackdown

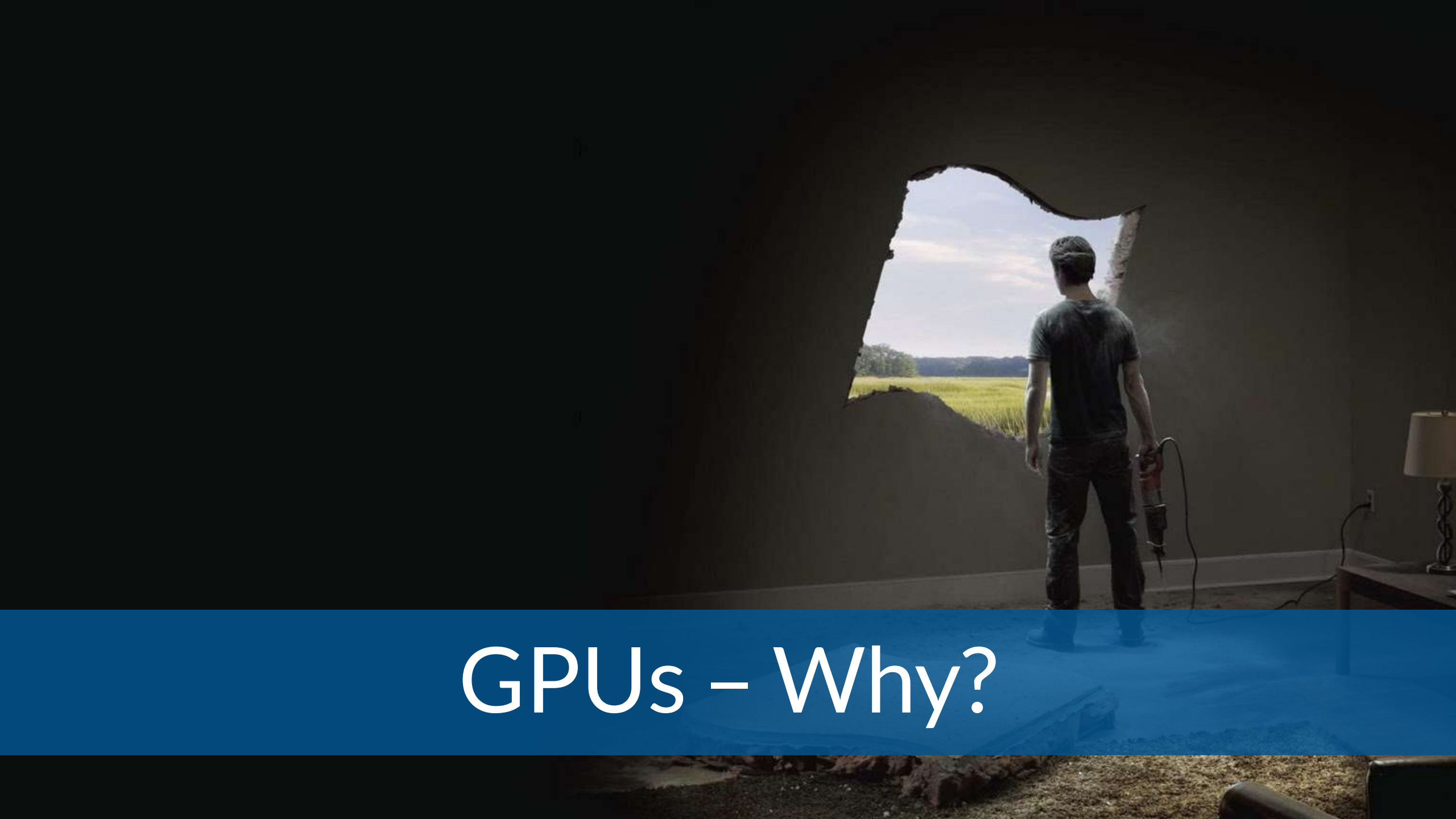
(on-premises and public clouds)

Benny Tritsch – benny@rdsgurus.com

Ruben Spruijt – ruben@rspruijt.com

Supported by

xenappblog.com



GPUs – Why?


Microsoft Office
Windows 7
Windows 10
Adobe Creative Cloud
Bohemia Interactive SIMULATIONS
Bentley MicroStation
INTERGRAPH GeoMedia
Wolfram Mathematica
ANSYS
creo A PTC Product
MATLAB SIMULINK
Schlumberger
CATIA
SOLIDWORKS
Paradigm
flame
HEXAGON GEOSPATIAL
Lightworks
PLM Software SIEMENS NX
AUTODESK PRODUCT DESIGN SUITE
AUTODESK INVENTOR
SYNAPSE PACS
AUTODESK MUDBOX
AUTODESK NAVISWORKS
AUTODESK REVIT
AUTODESK 3DS MAX
AUTODESK AUTOCAD 2014
AUTODESK INFRAWORKS
AUTODESK BUILDING DESIGN SUITE
ESRI ArcGIS
AUTODESK FABRICATION
AUTODESK SHOWCASE
AUTODESK ALIAS
acceleware PROCESSING SUPERPOWER
AUTODESK AUTOCAD CIVIL 3D
AUTODESK SIMULATION
Autodesk Maya


CITRIX®


Microsoft


NUTANIX™
Xi Frame


amazon

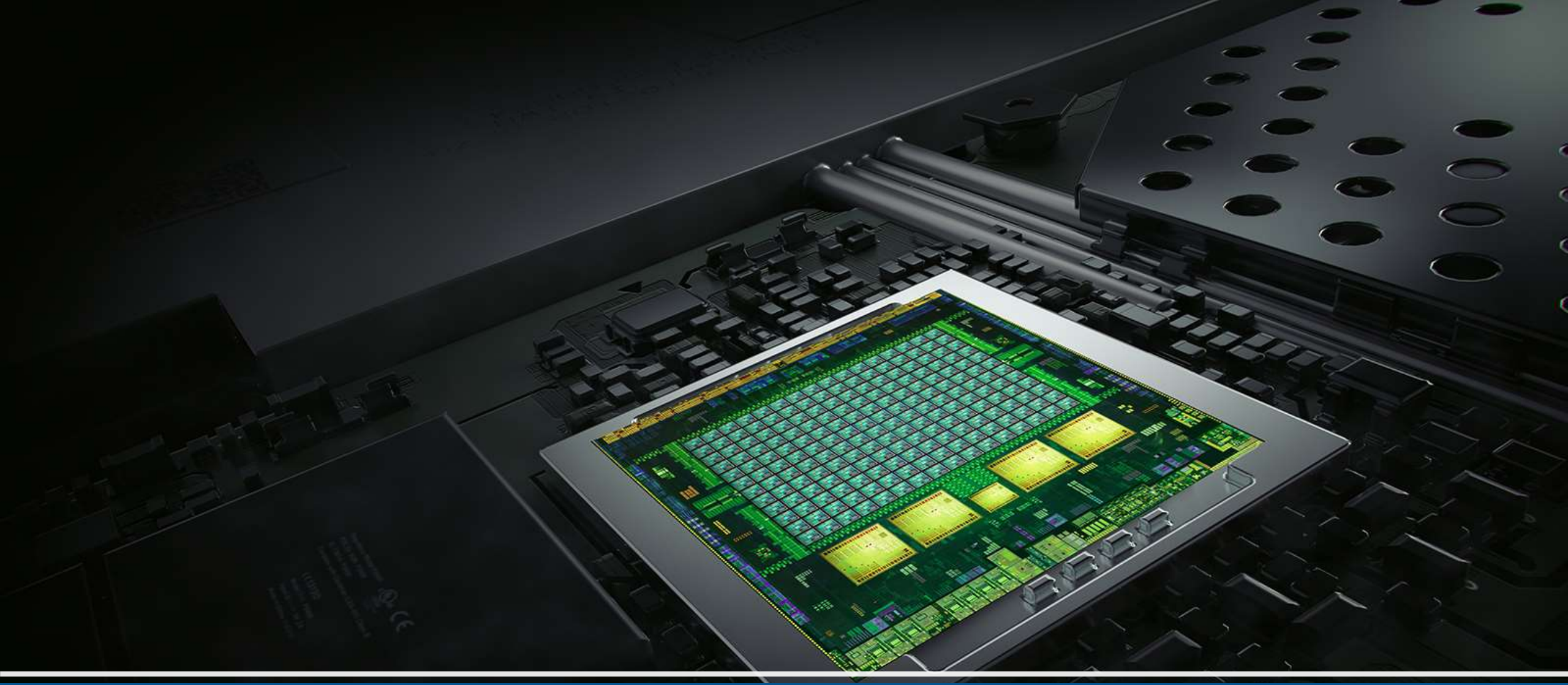

ERICOM
BE CONNECTED, BE SECURE


vmware®


Workspot


Parallels™


What are the options?



GPU options on-premises


AMD
intel
NVIDIA

GPU options public clouds


Microsoft
Azure


Alibaba Cloud


aws


Google Cloud Platform

GPU optimized virtual machine sizes

06/11/2019 • 10 minutes to read • Contributors all

GPU optimized VM sizes are specialized virtual machines available with single or multiple NVIDIA GPUs. These sizes are designed compute-intensive, graphics-intensive, and visualization workloads. This article provides information about the number and type GPUs, vCPUs, data disks, and NICs. Storage throughput and network bandwidth are also included for each size in this grouping.

- NC, NCv2, NCv3 sizes are optimized for compute-intensive and network-intensive applications and algorithms. Some exam are CUDA- and OpenCL-based applications and simulations, AI, and Deep Learning. The NCv3-series is focused on high- performance computing workloads featuring NVIDIA's Tesla V100 GPU. The NC-series uses the Intel Xeon E5-2690 v3 2.60G v3 (Haswell) processor, and the NCv2-series and NCv3-series VMs use the Intel Xeon E5-2690 v4 (Broadwell) processor.
- ND, and NDv2 The ND-series is focused on training and inference scenarios for deep learning. It uses the NVIDIA Tesla P40 and the Intel Xeon E5-2690 v4 (Broadwell) processor. The NDv2-series uses the Intel Xeon Platinum 8168 (Skylake) processo
- NV and NVv3 sizes are optimized and designed for remote visualization, streaming, gaming, encoding, and VDI scenarios u frameworks such as OpenGL and DirectX. These VMs are backed by the NVIDIA Tesla M60 GPU.

Google Cloud Platform

NV6, NC12, NV24 – Cloud Workstation Workhorse

- CPU: Xeon v3 – 2.60GHz
- CPU: 6-
- RAM: 56GB-224GB
- GPU: 1-4 NVIDIA M60 GPU
- Storage: Standard SSD – Azure

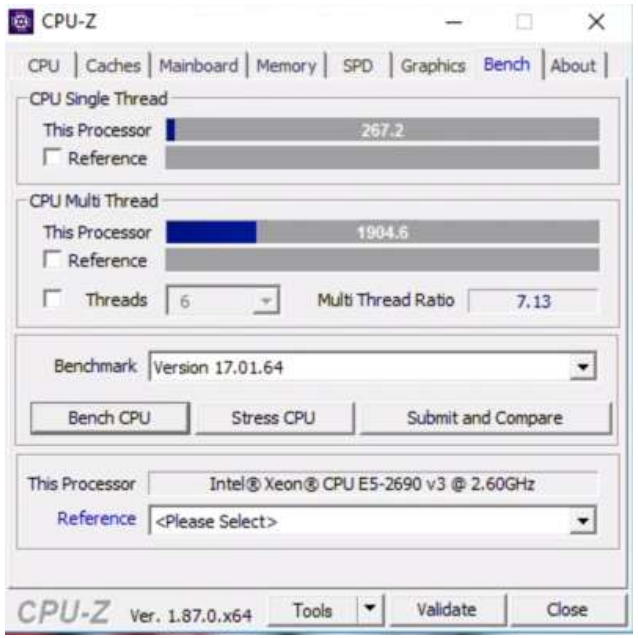
Size	vCPU	Memory: GiB	Temp storage (SSD) GiB	GPU	GPU memory: GiB	Max data disks	Max NICs	Virtual Workstations	Virtual Applications
Standard_NV6	6	56	340	1	8	24	1	1	25
Standard_NV12	12	112	680	2	16	48	2	2	50
Standard_NV24	24	224	1440	4	32	64	4	4	100

1 GPU = one-half M60 card.

Microsoft Azure

Personal notes:

- Intel v3 CPUs – only 2.6GHz
- NVIDIA Maxwell based GPUs
- No ‘slicing’ of GPUs | No vGPU
- No Smaller-size (cheaper) instance
- Storage = options and be-aware!
- Benchmark below = NV6



G2 – Kepler based
G3 – Maxwell based
Elastic Graphics
Graphics (AMD) within AppStream

G3s – G3.16XL – Cloud Workstation Workhorse

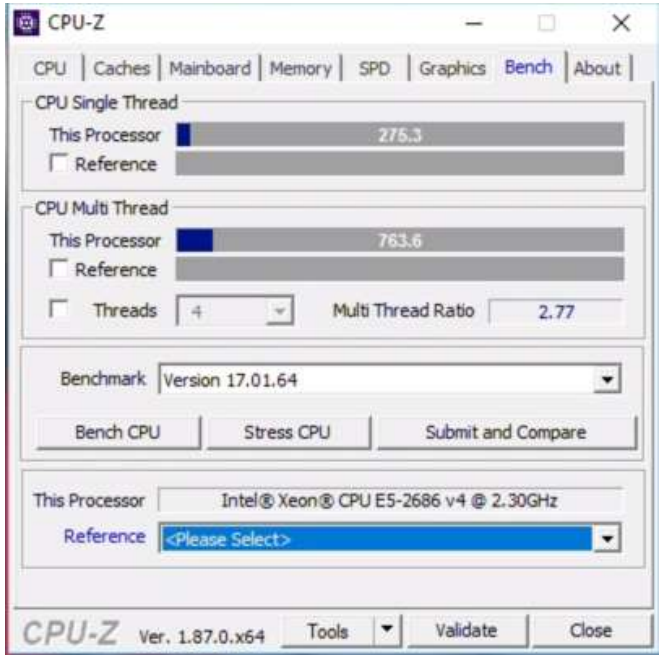
- CPU: Xeon v4 – 2.30GHz, 4-64 vCPU
- RAM: 30-488GB
- GPU: 1-4 NVIDIA M60 GPU
- Storage: EBS

Name	GPUs	vCPU	Memory (GiB)	GPU Memory (GiB)	Price/hr* (Linux)	Price/hr* (Windows)	1-yr Reserved Instance Effective Hourly* (Linux)	3-yr Reserved Instance Effective Hourly* (Linux)
g3s.xlarge	1	4	30.5	8	$0.75	$0.93	$0.525	$0.405
g3.4xlarge	1	16	122	8	$1.14	$1.876	$0.741	$0.538
g3.8xlarge	2	32	244	16	$2.28	$3.752	$1.482	$1.076
g3.16xlarge	4	64	488	32	$4.56	$7.504	$2.964	$2.152

*Prices shown are for US East (Northern Virginia) AWS Region. Prices for 1-year and 3-year reserved instances are for "Partial Upfront" payment options or "No upfront" for instances without the Partial Upfront option.

Personal notes G3:

- Intel v4 CPUs – only 2.3GHz
- NVIDIA Maxwell based GPUs
- No ‘slicing’ of GPUs | No vGPU
- Great to have small size (G3s.XL) workstation
- EBS is good
- Benchmark below = G3S.XL



Google Cloud Platform

For graphics workloads, GPU models are available in the following stages:

- NVIDIA® Tesla® T4 Virtual Workstations: `nvidia-tesla-t4-vws` : **Generally Available**
- NVIDIA® Tesla® P100 Virtual Workstations: `nvidia-tesla-p100-vws` : **Generally Available**
- NVIDIA® Tesla® P4 Virtual Workstations: `nvidia-tesla-p4-vws` : **Generally Available**

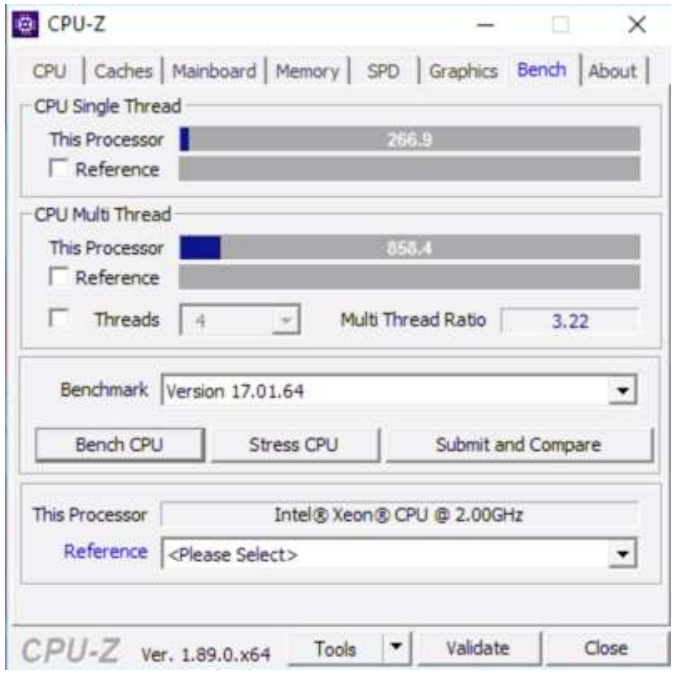
Google Cloud Platform

NVIDIA® Tesla® P4	1 GPU	8 GB GDDR5	1 - 24 vCPUs	1 - 156 GB	• us-west2-c • us-west2-b • us-central1-a • us-central1-c • us-east4-a • us-east4-b • us-east4-c • northamerica-northeast1-a • northamerica-northeast1-b • northamerica-northeast1-c • europe-west4-b • europe-west4-c • australia-southeast1-a • australia-southeast1-b • asia-southeast1-b • asia-southeast1-c
	2 GPUs	16 GB GDDR5	1 - 48 vCPUs	1 - 312 GB	
	4 GPUs	32 GB GDDR5	1 - 96 vCPUs	1 - 624 GB	

Google Cloud Platform

Personal notes :

- Intel CPUs – 2.0GHz – Meh!!
- Choice! (P4,T4,P100) NVIDIA based GPUs
- No ‘slicing’ of GPUs | No vGPU
- Great to have flexibility in CPU/RAM config
- Benchmark below = 1_x_nvidia-tesla-p4-vws



vgn5i, light-weight compute optimized type fa...

gn6i, compute optimized type family with GPUs

gn6v, compute optimized type family with GPUs

gn5, compute optimized type family with GPU

gn5i, compute optimized type family with GPU

gn4, compute optimized type family with GPU

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	GPU	GPU memory (GiB)	Bandwidth (Gbit/s)**	Packet forwarding rate (thousand pps)***	NIC queues****	ENIs*****
ecs.vgn5i-m1.large	2	6	N/A	P4*1/8	1	1	300	2	2
ecs.vgn5i-m2.xlarge	4	12	N/A	P4*1/4	2	2	500	2	3
ecs.vgn5i-m4.2xlarge	8	24	N/A	P4*1/2	4	3	800	2	4
ecs.vgn5i-m8.4xlarge	16	48	N/A	P4*1	8	5	1,000	4	5

Features

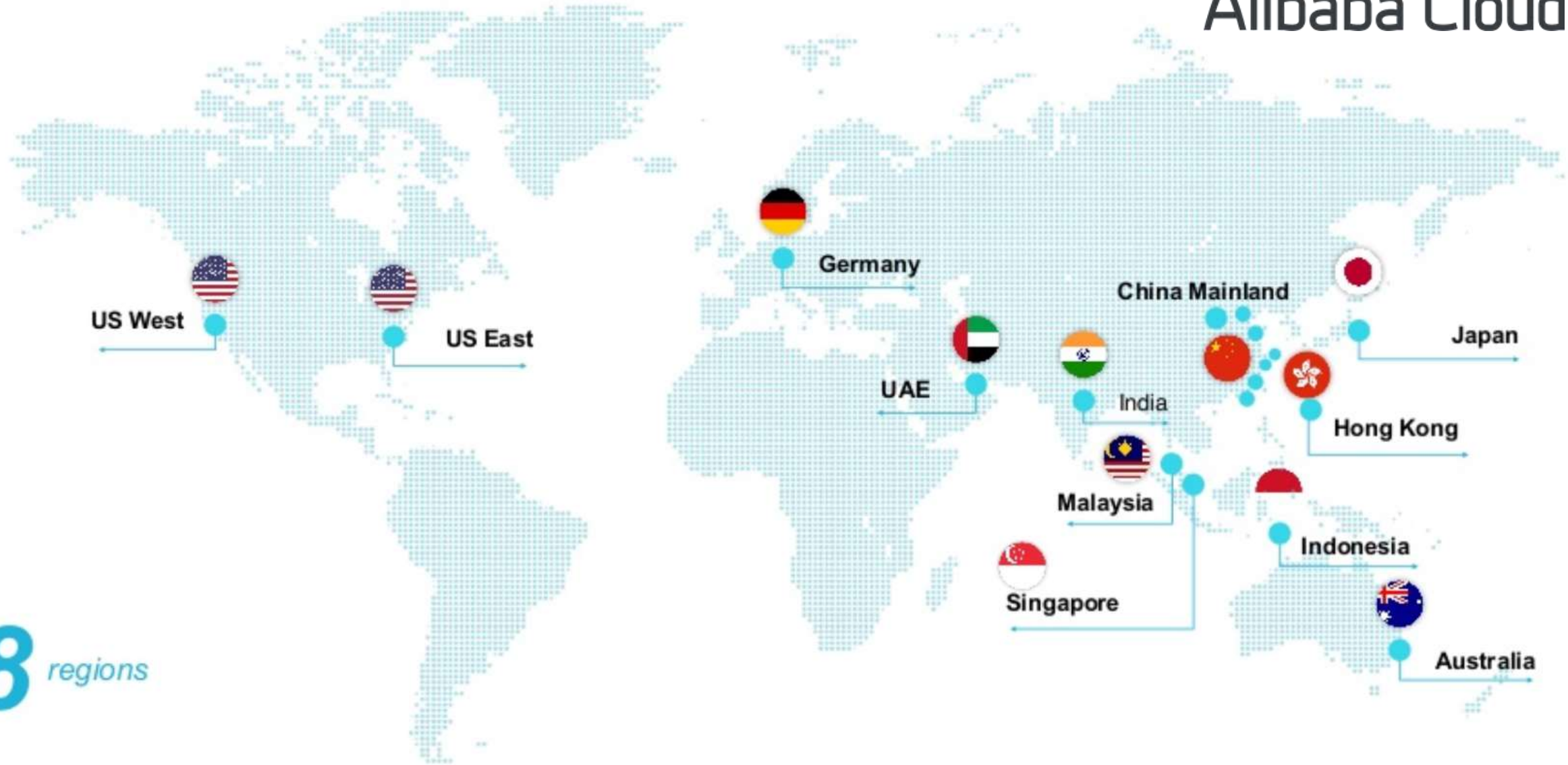
- I/O-optimized
- Supports SSD Cloud Disks and Ultra Disks
- Use an NVIDIA P4 GPU computation accelerator
- Contains a virtual GPU (which is the result of partitioned virtualization)
  - Supports the 1/8, 1/4, 1/2, and 1:1 computing capacity of NVIDIA Tesla P4 GPUs
  - Supports 1, 2, 4, and 8 GiB of video memory
- Equipped with a vCPU to memory ratio of 1:3
- Equipped with 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- Supports strong network performance through sufficient computing capacity
- Suitable for the following scenarios:
  - Real-time online rendering required for cloud gaming and AR/VR applications
  - AI reasoning (including deep and machine learning), used in the elastic deployment of Internet services that use AI reasoning and computing
  - Educational and modeling experiment environments that use deep learning

Instance types

Instance types	vCPU	Memory (GiB)	Local disks (GiB)*	GPU	GPU memory (GiB)	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	IPv6-ready?	NIC queues****	ENIs*****
ecs.gn6i-c4g1.xlarge	4	15	N/A	T4*1	16	4	500	Yes	2	2
ecs.gn6i-c8g1.2xlarge	8	31	N/A	T4*1	16	5	800	Yes	2	2
ecs.gn6i-c16g1.4xlarge	16	62	N/A	T4*1	16	6	1,000	Yes	4	3
ecs.gn6i-c24g1.6xlarge	24	93	N/A	T4*1	16	7.5	1,200	Yes	6	4
ecs.gn6i-c24g1.12xlarge	48	186	N/A	T4*2	32	15	2,400	Yes	12	6
ecs.gn6i-c24g1.24xlarge	96	372	N/A	T4*4	64	30	4,800	Yes	24	8

Alibaba Cloud Regions All Around the World

Alibaba Cloud



18 *regions*

TeamRGE

Remoting Graphics Experts

Thanks!

Benny Tritsch – benny@rdsgurus.com

Ruben Spruijt – ruben@rspruijt.com

Supported by

xenappblog.com