Vienna Deep Learning Meetup

Deep Learning Hardware Overview: What and where to buy or rent

Jan Schlüter
JKU Linz

René Donner
contextflow

Thomas Lidy
Musimap

Outline

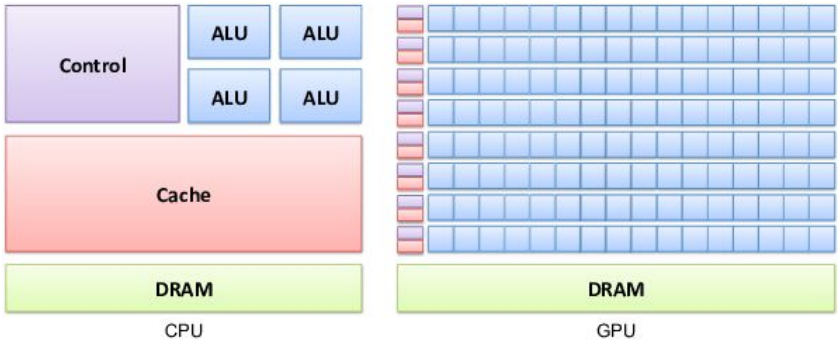
- NVIDIA GPUs
- Intel/AMD GPU-ready Servers
- AMD GPUs
- GPU Cloud Providers

NVIDIA GPUs

NVIDIA GPUs

Why GPUs?

- Deep learning: lots of linear algebra (matrix-matrix products, convolutions)
  - Embarrassingly parallel
- GPUs designed to run *same computation* on different pieces of data
  - multi-core CPUs designed to run *different computations* in parallel
  - CPUs have SIMD (single instruction, multiple data) only on smaller scale
- Larson/McAllister 2001: “Fast Matrix Multiplies using Graphics Hardware”



Lounis et al., 2015

https://doi.org/10.1145/582034.582089

NVIDIA GPUs

Why NVIDIA?

- saw the potential of GPGPUs (general-purpose graphical processing units)
- started selling GPUs *without graphics output* in 2007
- invested a lot in software to make it accessible: CUDA, cuBLAS, cuDNN, …
- academic hardware grant program to donate GPUs to universities

➢ led to wide adoption in academia, and academic software libraries

https://doi.org/10.1145/582034.582089

NVIDIA GPUs

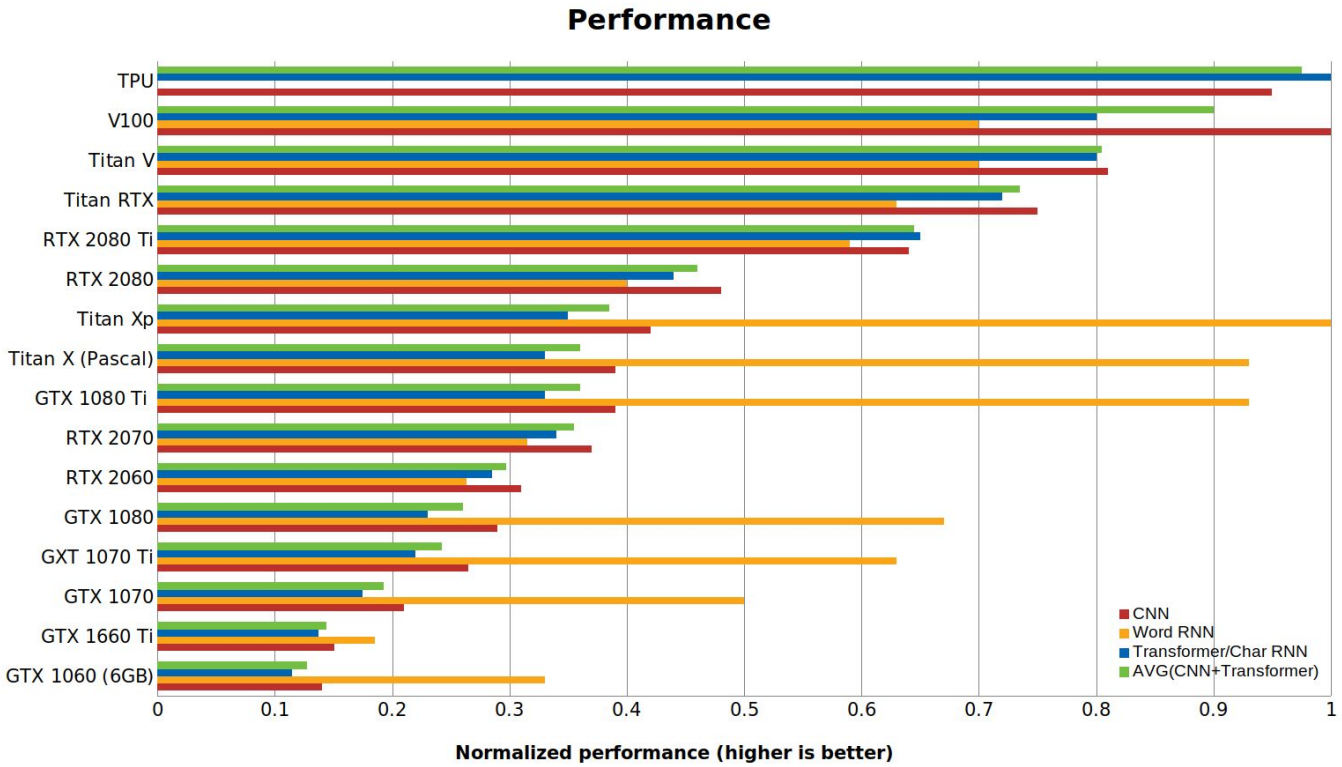
Main criteria

- Number of CUDA cores (32-bit and/or 64-bit precision)
- Number of Tensor cores (32-bit and 16-bit precision, possibly int16/8/4)
- Memory size
- Memory bandwidth

Other criteria

- 64-bit performance
- power draw / cooling requirements

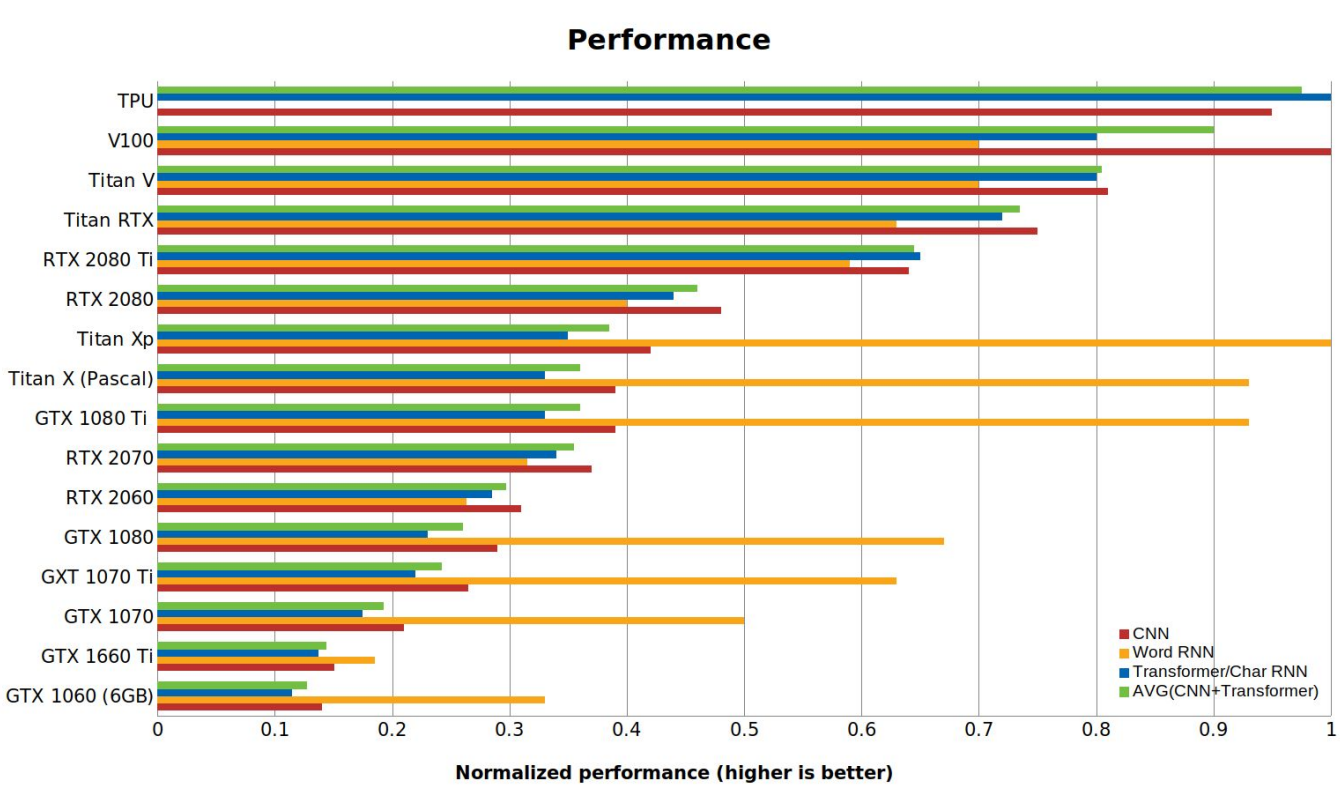
NVIDIA GPUs



https://timdettmers.com/2019/04/03/which-gpu-for-deep-learning/

Vienna Deep Learning Meetup

NVIDIA GPUs



by Google, not for sale

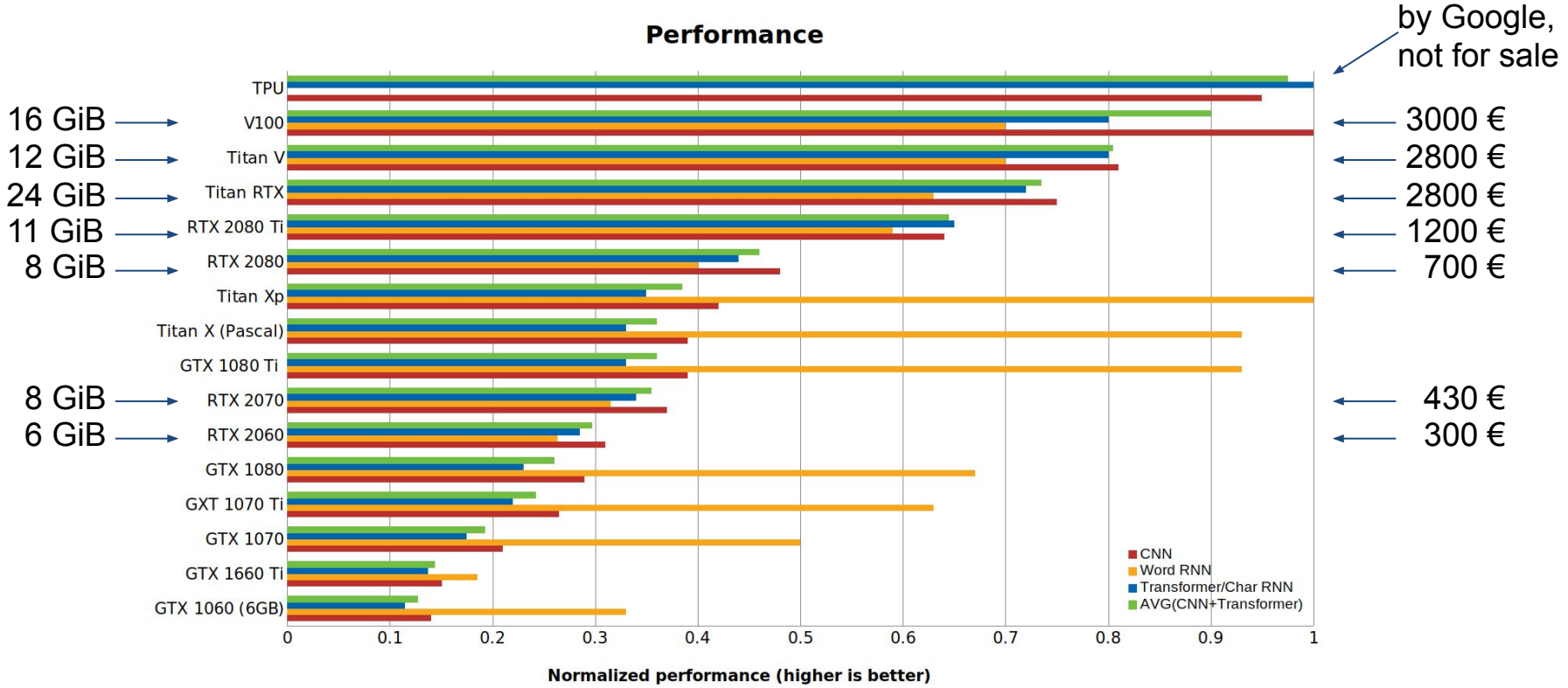
3000 €
2800 €
2800 €
1200 €
700 €

430 €
300 €

https://timdettmers.com/2019/04/03/which-gpu-for-deep-learning/

Vienna Deep Learning Meetup

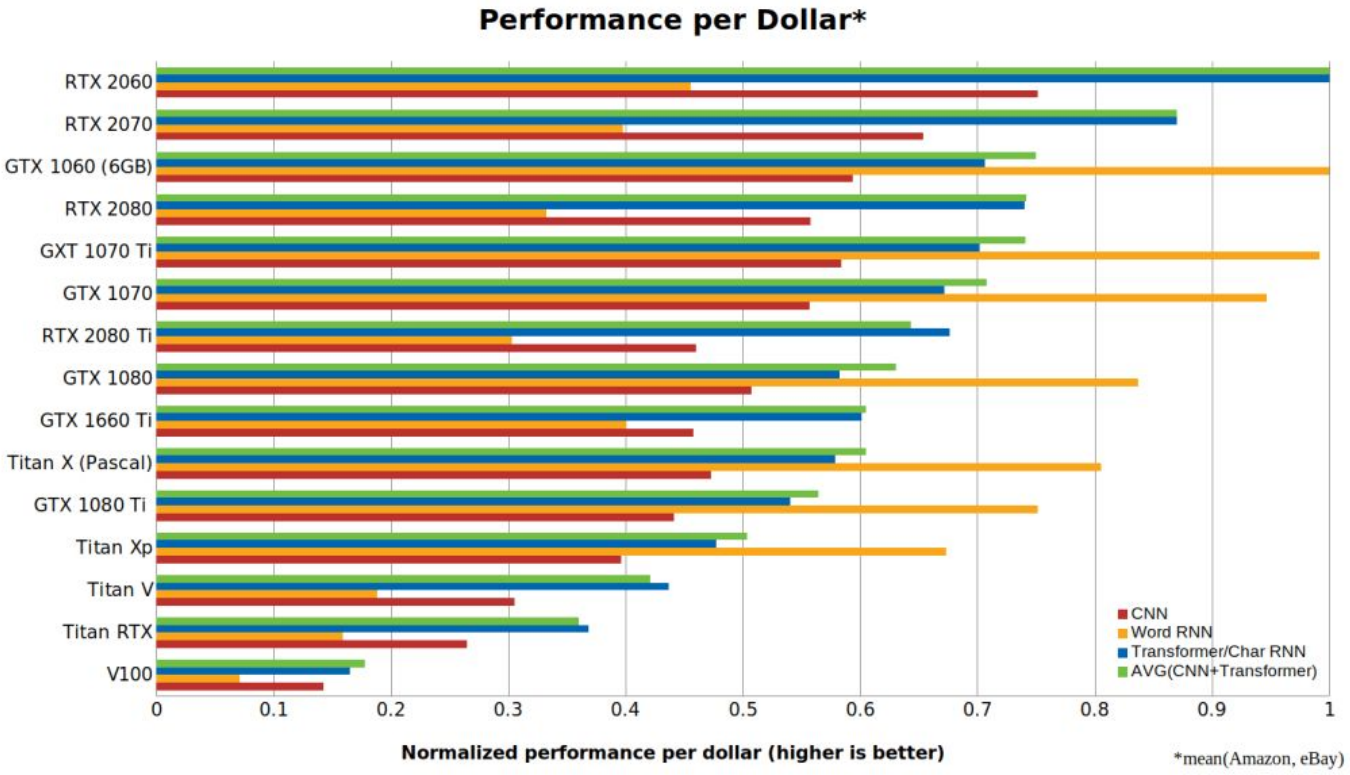
NVIDIA GPUs



https://timdettmers.com/2019/04/03/which-gpu-for-deep-learning/

Vienna Deep Learning Meetup

NVIDIA GPUs

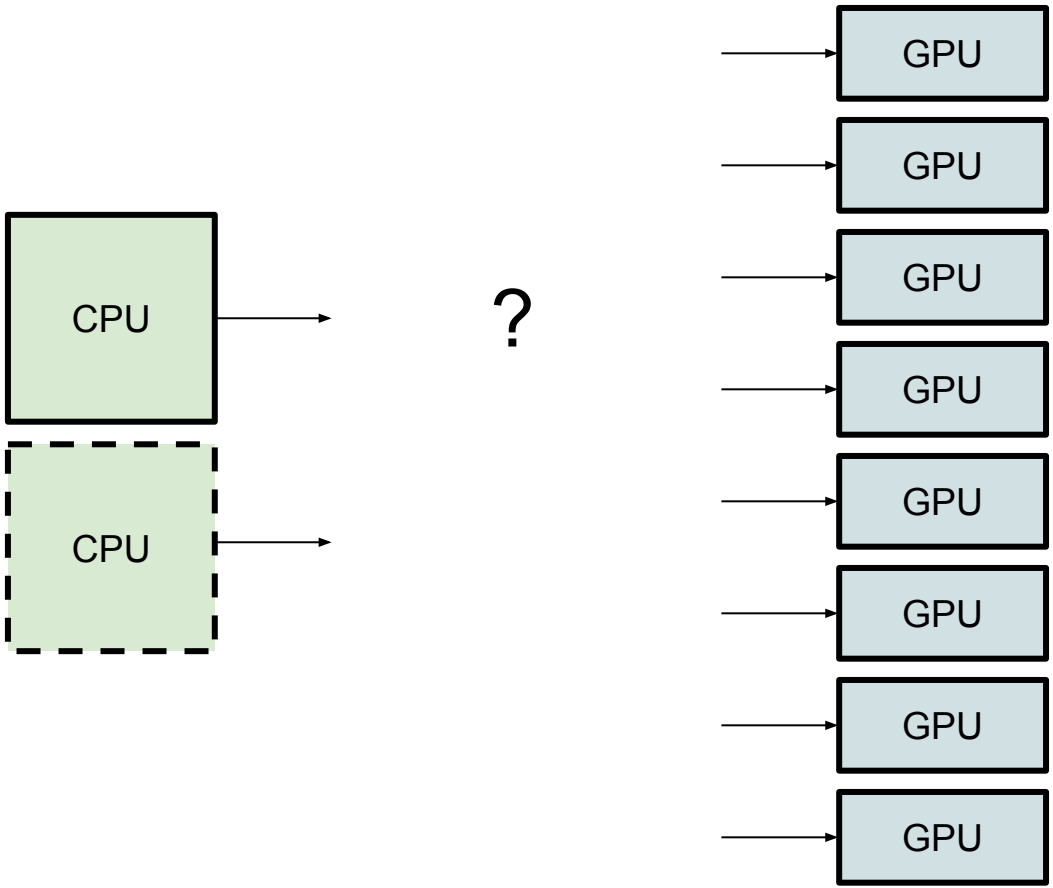


https://timdettmers.com/2019/04/03/which-gpu-for-deep-learning/

Vienna Deep Learning Meetup

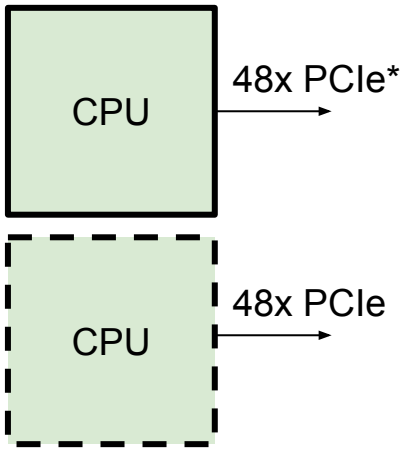
Intel/AMD GPU-ready servers

Intel/AMD GPU-ready servers

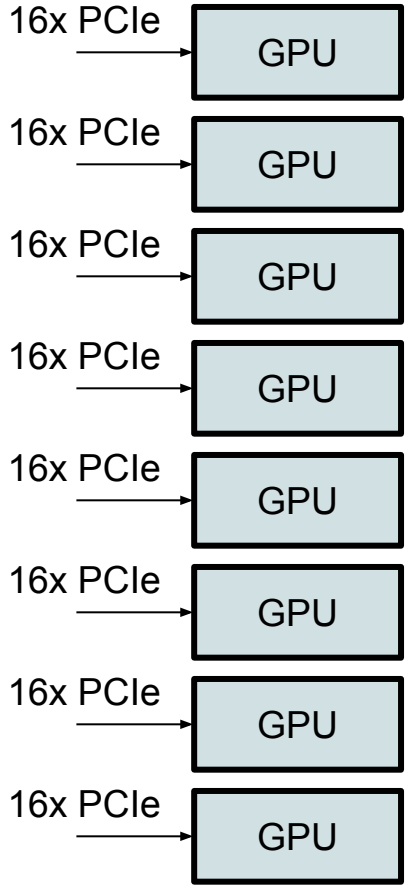


Vienna Deep Learning Meetup

Intel/AMD GPU-ready servers



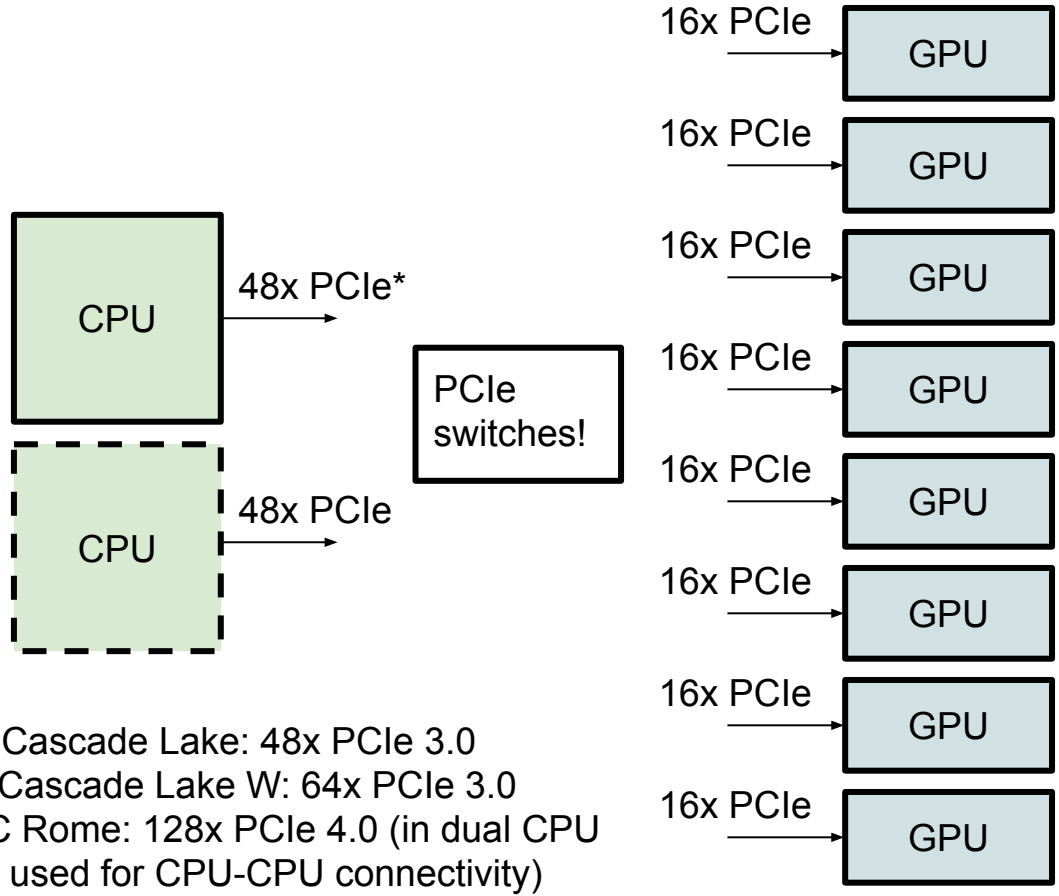
?



* Intel Xeon Cascade Lake: 48x PCIe 3.0
Intel Xeon Cascade Lake W: 64x PCIe 3.0
AMD EPYC Rome: 128x PCIe 4.0 (in dual CPU system: 64 used for CPU-CPU connectivity)

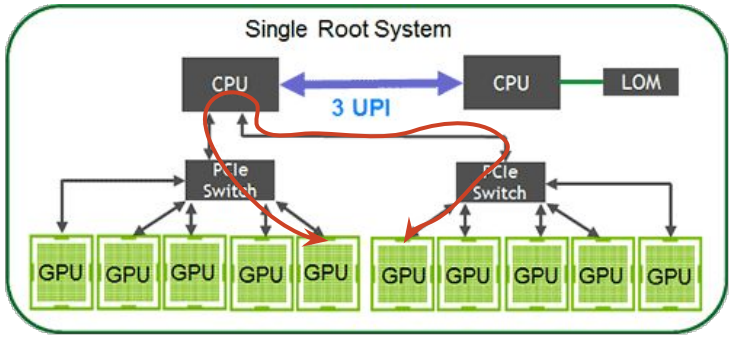
Vienna Deep Learning Meetup

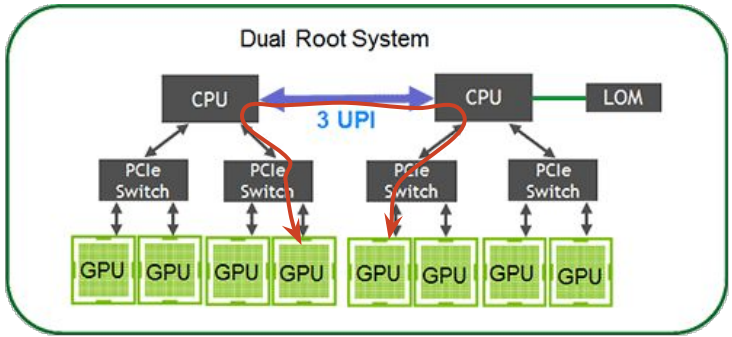
Intel/AMD GPU-ready servers



* Intel Xeon Cascade Lake: 48x PCIe 3.0
  Intel Xeon Cascade Lake W: 64x PCIe 3.0
  AMD EPYC Rome: 128x PCIe 4.0 (in dual CPU system: 64 used for CPU-CPU connectivity)

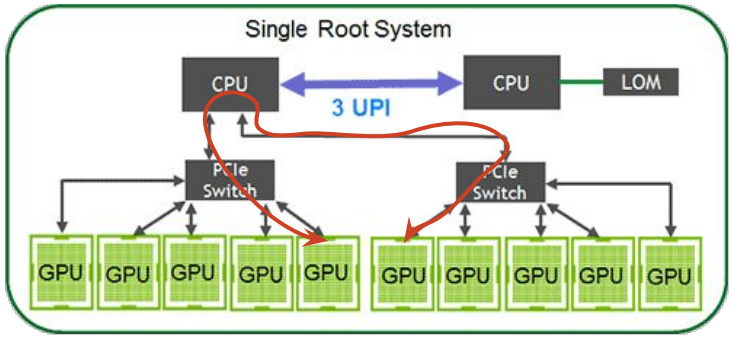
Single root complex vs. dual root complex



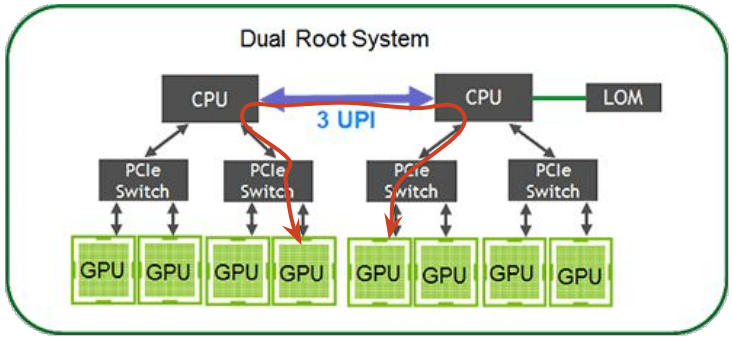


https://www.supermicro.com/products/system/4U/4029/PCIe-Root-Architecture.cfm

Single root complex vs. dual root complex



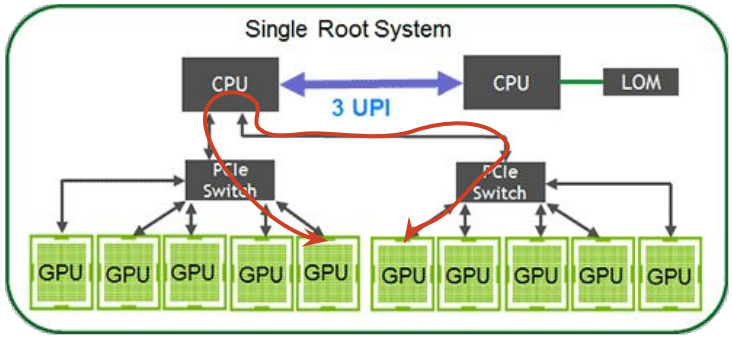
high GPU-GPU bandwidth



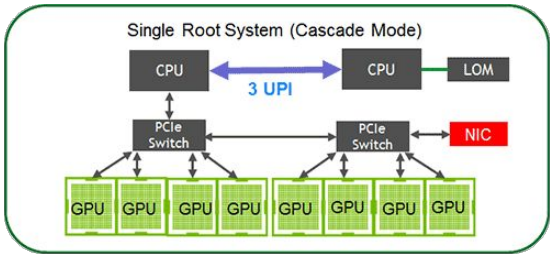
high GPU-CPU bandwidth

https://www.supermicro.com/products/system/4U/4029/PCIe-Root-Architecture.cfm

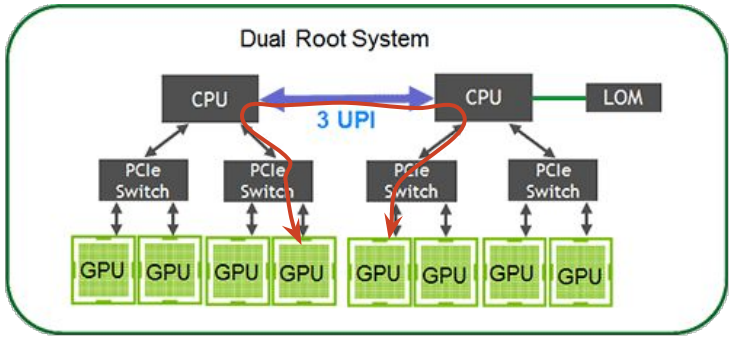
Single root complex vs. dual root complex



high GPU-GPU bandwidth



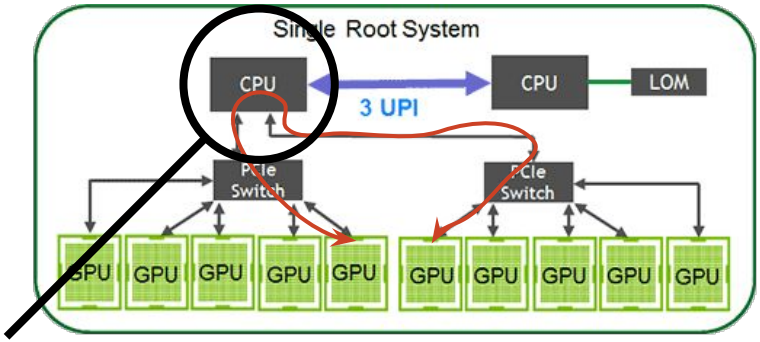
high GPU-GPU bandwidth, but only 16 lanes between CPU and all GPUs



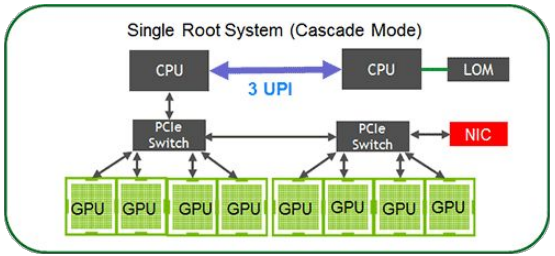
high GPU-CPU bandwidth

https://www.supermicro.com/products/system/4U/4029/PCIe-Root-Architecture.cfm

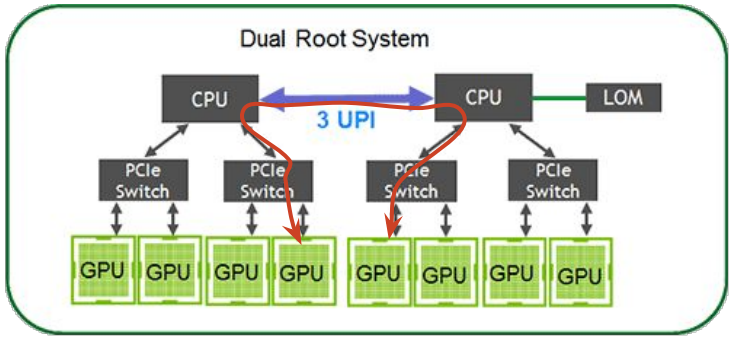
Single root complex vs. dual root complex



high GPU-GPU bandwidth



high GPU-GPU bandwidth, but only 16 lanes between CPU and all GPUs



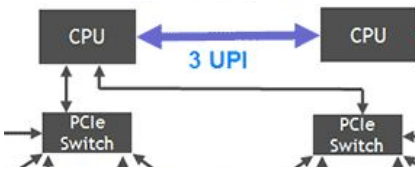
high GPU-CPU bandwidth

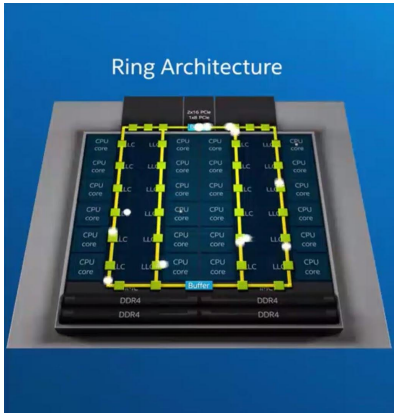
https://www.supermicro.com/products/system/4U/4029/PCIe-Root-Architecture.cfm

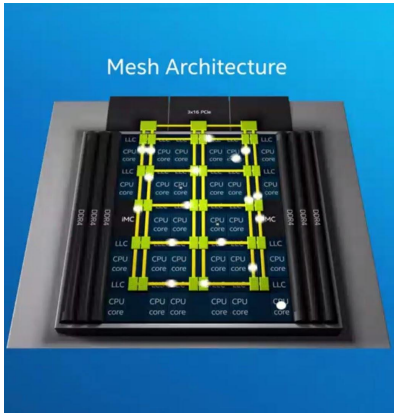
Single root complex vs. dual root complex

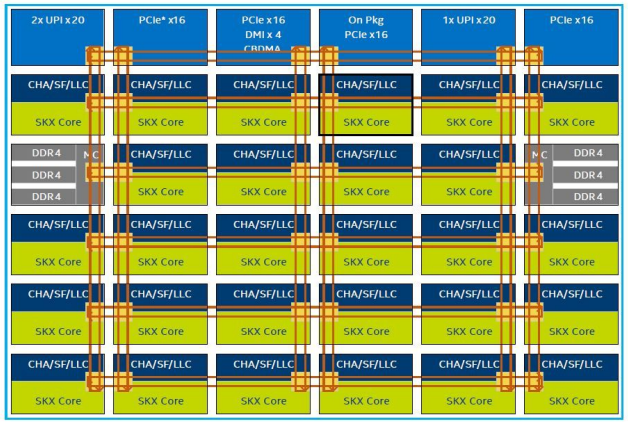
Intel Broadwell:
Ring architecture
Two PCIe controllers in single root

Intel Cascade Lake:
Mesh architecture
Two PCIe controllers on different hops





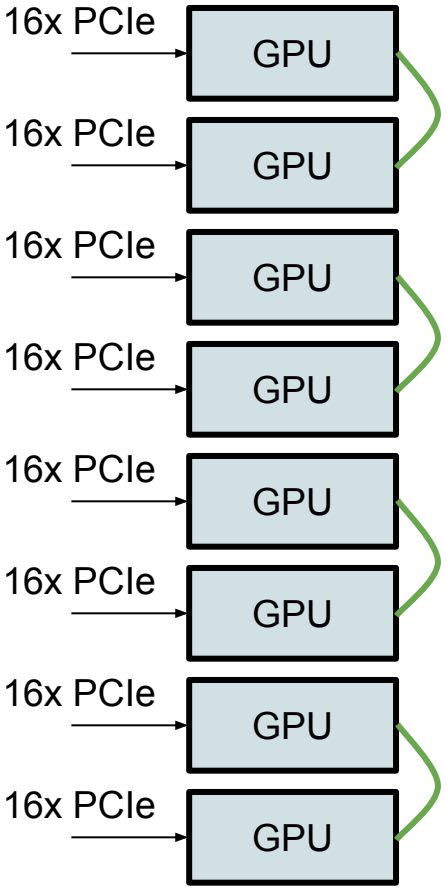




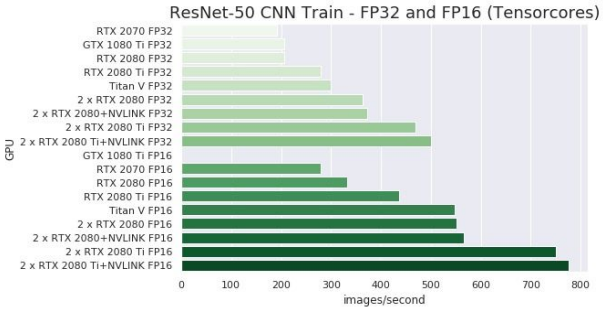
CHA – Caching and Home Agent ; SF – Snoop Filter; LLC – Last Level Cache;
SKX Core – Skylake Server Core; UPI – Intel® UltraPath Interconnect

https://www.servethehome.com/how-intel-skylake-sp-changes-impacted-single-root-pcie-deep-learning-servers/

NVlink



- High-bandwidth interconnect between pairs of GPUs
- 3-4% performance improvement in dual-GPU ResNet training



https://www.pugetsystems.com/labs/hpc/RTX-2080Ti-with-NVLINK---TensorFlow-Performance-Includes-Comparison-with-GTX-1080Ti-RTX-2070-2080-2080Ti-and-Titan-V-1267/

Vienna Deep Learning Meetup

Server sizes

1U (rack height unit)

- up to 4 GPUs
- hard to cool
- only for inference, e.g. NVIDIA T4
- some vendors will sell it with RTX 2080 Ti, not recommended!

2U

- up to 8 GPUs
- still hard to cool, need to choose carefully

4U

- up to 10 GPUs
- has space for NVlink

Server recommendations

Gigabyte G291-Z20

- 2U server for 8 GPUs
- AMD EPYC Rome
- cool enough if only putting 4 GPUs
- ~10k€ all in all

Supermicro 4029-TRT2

- 4U server for 10 GPUs
- Intel Cascade Lake
- need special top cover for RTX cards and NVlink (makes it 4.5U)
- ~20k€ all in all

Other options

- smaller Supermicro servers: not cooled well enough
- Tyan, Asus, Asrock: more expensive
- Zstor GS-P206G: custom 2U 6GPU design with good cooling, but expensive

Vienna Deep Learning Meetup

Possible vendors

- server-bau.de
- deltacomputer.de
- sysgen.de
- bressner.de
- thomas-krenn.com
- zstor.de

AMD GPUs

AMD GPUs

- Good value / money
- Focussed on gaming so far
- CUDA -> ROCm ecosystem

AMD GPUs

- Vega 20 chip
- Navi 10/14 chips

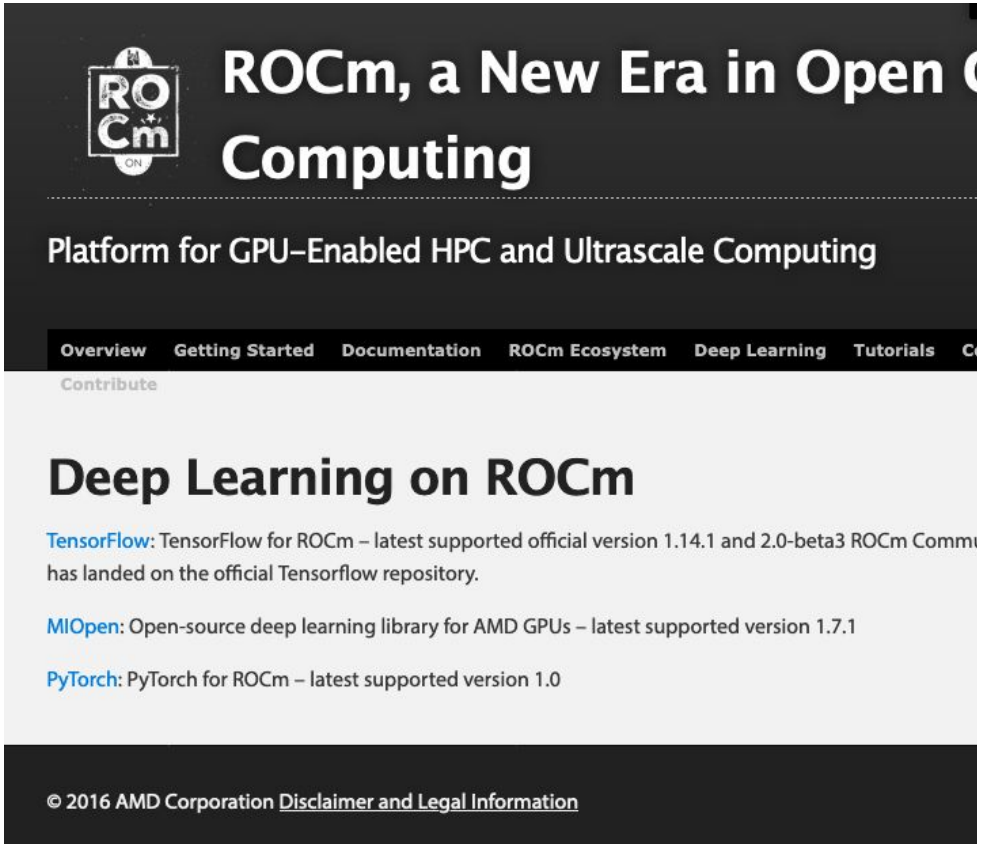
	Score	GPU	Base/Boost	Memory	Power	Buy
Nvidia Titan RTX	100	TU102	1350/1770 MHz	24GB GDDR6	280W	US$2,449 VIEW DEAL AMAZON
Nvidia GeForce RTX 2080 Ti	98.4	TU102	1350/1635 MHz	11GB GDDR6	260W	US$1,186.94 VIEW DEAL AMAZO
Nvidia GeForce RTX 2080 Super	98.2	TU104	1650/1815 MHz	8GB GDDR6	250W	US$696.16 VIEW DEAL AMAZON
Nvidia GeForce RTX 2080	96.1	TU104	1515/1800 MHz	8GB GDDR6	225W	US$1,102.64 VIEW DEAL AMAZO
Nvidia Titan X	96.0	GP102	1405/1480 MHz	12GB GDDR5X	250W	CHECK AMAZON
Nvidia GeForce GTX 1080 Ti	96.0	GP102	1480/1582 MHz	11GB GDDR5X	250W	US$769.99 VIEW DEAL AMAZON
AMD Radeon RX 5700 XT	95.8	Navi 10	1605/1905 MHz	8GB GDDR6	225W	US$389.99 VIEW DEAL AMAZON
Nvidia GeForce RTX 2070 Super	94.1	TU104	1605/1770 MHz	8GB GDDR6	215W	US$599.42 VIEW DEAL AMAZON
AMD Radeon VII	92.4	Vega 20	1400/1750 MHz	16GB HBM2	300W	US$599.99 VIEW DEAL AMAZON

AMD ROCm software stack

- Replaces CUDA
- cuDNN -> MIOpen

https://rocm.github.io/

AMD ROCm software stack



https://rocm.github.io/

Debugger

- ROCm debugger binary build
- ROCm GDB source code
- ROCm GPU debugger SDK;ROCm debug run time that services GBD

Profiling Tools

- rocprof
- rocProfiler
- rocTracer

Math Libraries

- rocBLAS
- rocFFT
- Tensile
- Eigen C++ Math Library with HIP Based GPU Acceleration
- clBLAS
- clFFT
- clSparse
- clRNG

Vienna Deep Learning Meetup

Benchmarks

▲ bonoboTP 88 days ago | parent | favorite | on: Tensorflow 2.0 AMD Supp

I just ran LambdaLabs' ResNet50 training benchmark (I could do more m

TensorFlow 1.14.4

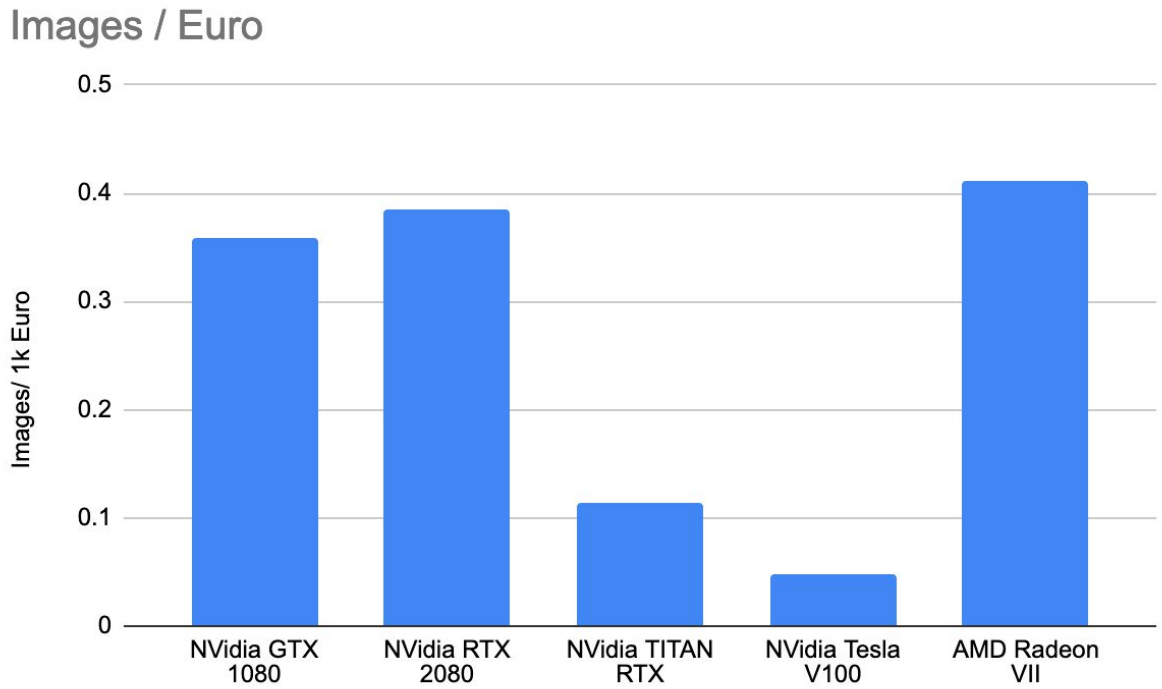
```
git clone https://github.com/lambdal/lambda-tensorflow-be
python lambda-tensorflow-benchmark/benchmarks/scripts/tf_
```

FP32
- NVIDIA GTX 1080 Ti: ~215 images/sec
- NVIDIA RTX 2080 Ti: ~300 images/sec
- NVIDIA TITAN RTX: ~320 images/sec
- NVIDIA Tesla V100: ~383 images/sec
- AMD Radeon VII: ~275 images/sec

FP16
- NVIDIA GTX 1080 Ti: ~277 images/sec
- NVIDIA RTX 2080 Ti: ~495 images/sec
- NVIDIA TITAN RTX: ~518 images/sec
- NVIDIA Tesla V100: ~725 images/sec
- AMD Radeon VII: ~373 images/sec

https://news.ycombinator.com/item?id=21666411

Value for money?



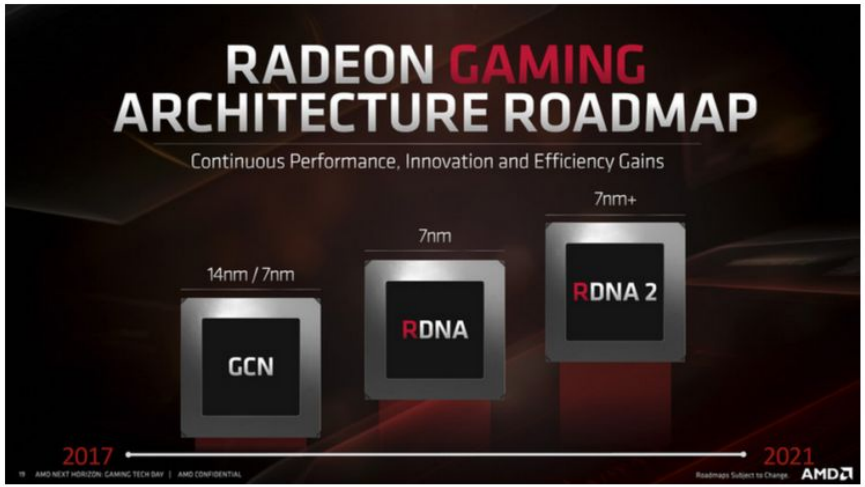
AMD GPU cloud provider

GPU EATER

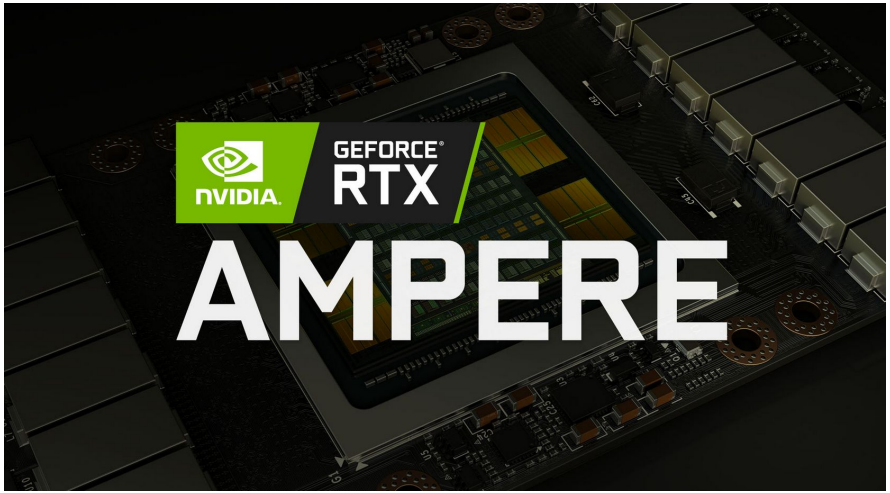
Provider	Plan	TFlops	Cores	GPU Mem	Architecture	Hourly	Monthly
amazon web services	p2.xlarge	4.37	2496	12GB	NVIDIA Kepler	$0.9000/h	$648/m
amazon web services	g3.4xlarge	4.82	2048	8GB	NVIDIA Maxwell	$1.1400/h	$820/m
GPU EATER	**a1.rx580**	**6.1**	**2304**	**8GB**	**AMD RADEON RX**	**$0.3458/h**	**$249/m**
GPU EATER	**a1.vega56**	**10.5**	**3584**	**8GB**	**AMD RADEON VEGA**	**$0.4794/h**	**$345/m**
GPU EATER	**a1.vegafe**	**13.1**	**4096**	**16GB**	**AMD RADEON VEGA**	**$0.6164/h**	**$443/m**
amazon web services	p3.2xlarge	14.0	5120	16GB	NVIDIA Volta	$3.0600/h	$2203/m

https://www.gpueater.com/index.html

AMD vs NVidia 2020



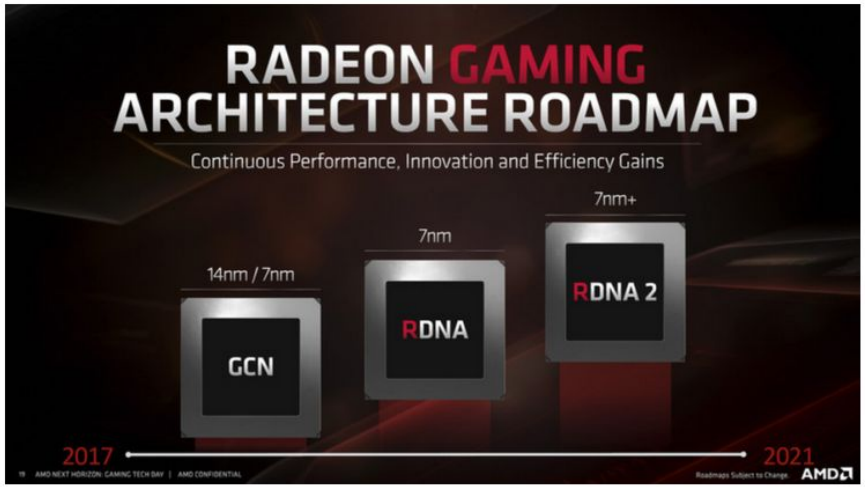
New AMD “Big Navi” chip



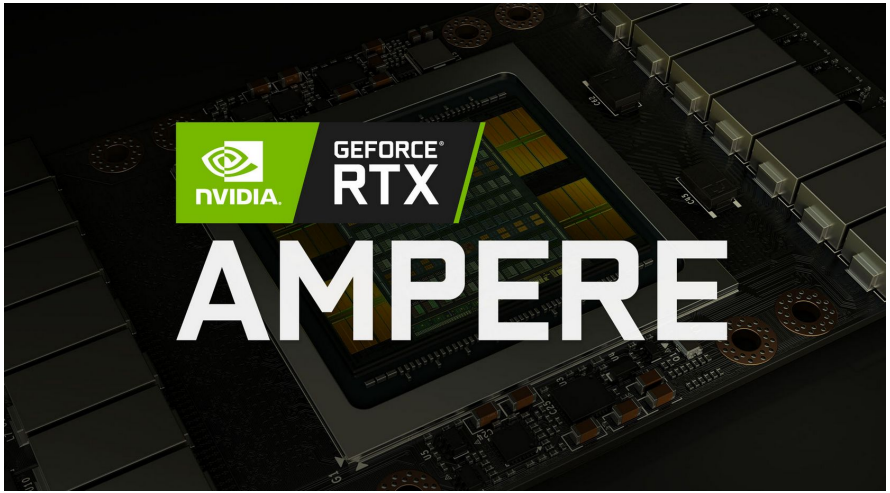
New NVidia “Ampere” architecture

Vienna Deep Learning Meetup

AMD vs NVidia 2020



New AMD “Big Navi” chip



New NVidia “Ampere” architecture

AMD Radeon Navi 20 Could Pack 5,120 Cores, 24GB of HBM2: Report

By Joel Hruska on February 26, 2020 at 7:47 am | 1 Comment

Vienna Deep Learning Meetup

GPU Cloud Providers

GPU Cloud Providers: Overview

- Amazon Web Services
- Floydhub
- Linode
- Exoscale
- Paperspace
- LeaderGPU
- Packet
- OVHcloud
- LambdaLabs
- Hetzner
- Vast.ai

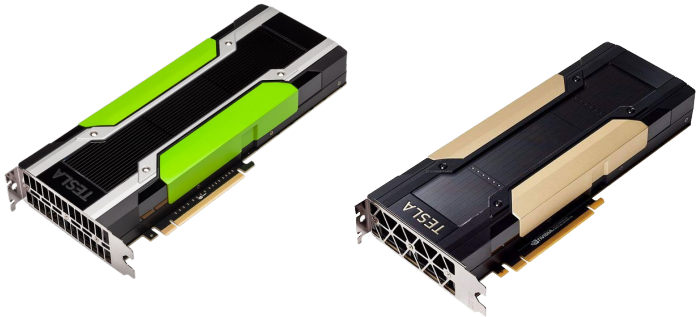
Amazon Web Services

Accelerated Computing:

- P2 Instances: 1, 8 or 16 **NVIDIA K80 GPUs** (12 GB VRAM each)
- P3 Instances: 1, 4 or 8 **NVIDIA Tesla V100 GPUs** (16 GB VRAM each)

Intel Xeon E5-2686 v4 (Broadwell) processors with 2.3 GHz (2.7 GHz turbo)
(or Xeon P-8175M processors for p3dn.24xlarge)

https://aws.amazon.com/ec2/instance-types/



Vienna Deep Learning Meetup

Amazon Web Services

Accelerated Computing - P2 Instances (NVidia K80):

Instance	GPUs	vCPU	Mem (GiB)	GPU Memory (GiB)	Network Performance	Pricing per hour	Pricing per month
p2.xlarge	1	4	61	12	High	$0,97	$700
p2.8xlarge	8	32	488	96	10 Gigabit	$7,78	$5.599
p2.16xlarge	16	64	732	192	25 Gigabit	$15,55	$11.197

Pricing for on-demand in EU-Ireland Region: https://aws.amazon.com/ec2/pricing/on-demand/
(up to 90% reduction using “Spot Instances”)

Amazon Web Services

Accelerated Computing - P3 Instances (NVidia Tesla V100):

Instance	GPUs	vCPU	Mem (GiB)	GPU Mem (GiB)	GPU P2P	Storage (GB)	Dedicated EBS Bandwidth	Networking Performance	Pricing per hour	Pricing per month
p3.2xlarge	1	8	61	16	-	EBS-Only	1.5 Gbps	Up to 10 Gigabit	$3,31	$2.380
p3.8xlarge	4	32	244	64	NVLink	EBS-Only	7 Gbps	10 Gigabit	$13,22	$9.518
p3.16xlarge	8	64	488	128	NVLink	EBS-Only	14 Gbps	25 Gigabit	$26,44	$19.037
p3dn.24xlarge	8	96	768	256	NVLink	2 x 900 NVMe SSD	19 Gbps	100 Gigabi	$33,71	$24.272

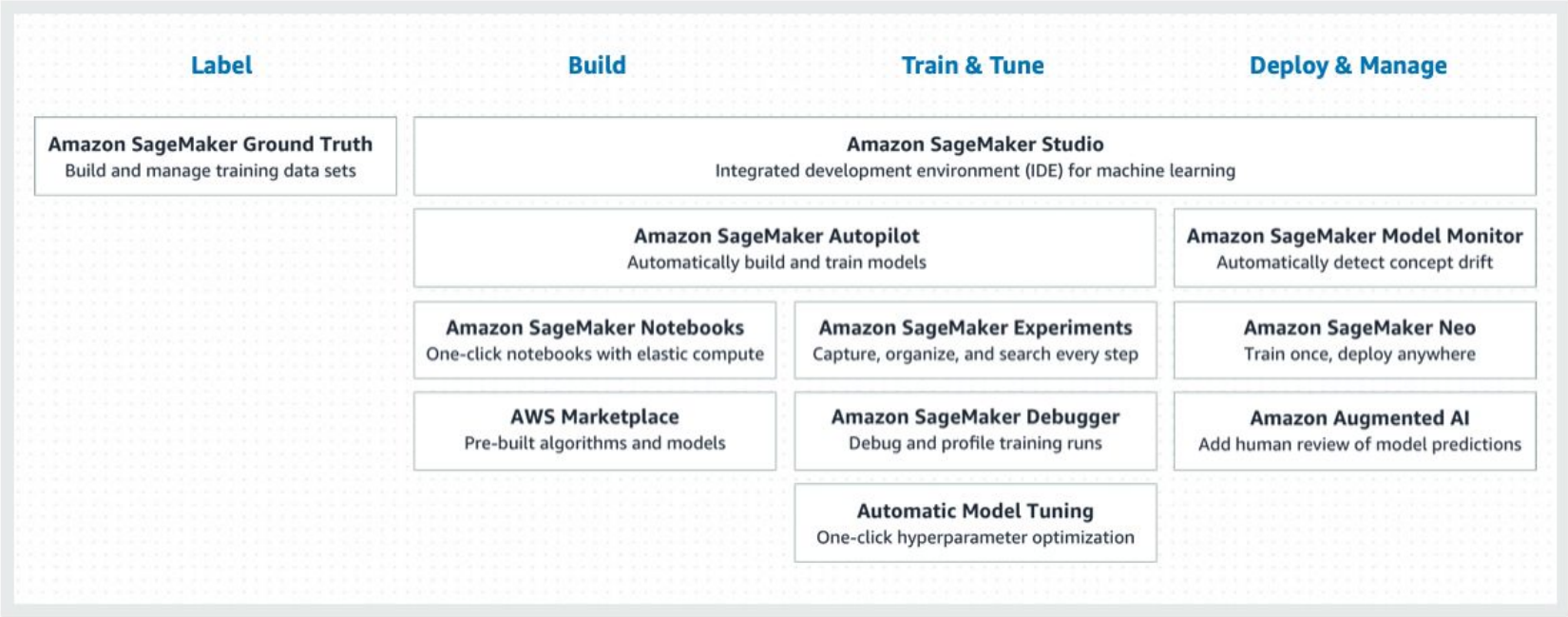
Pricing for on-demand in EU-Ireland Region: https://aws.amazon.com/ec2/pricing/on-demand/
(up to 90% reduction using “Spot Instances”)

Vienna Deep Learning Meetup

Amazon Web Services

Amazon SageMaker

fully managed service to build, train, and deploy machine learning models quickly



https://aws.amazon.com/sagemaker/

FLOYDHUB

Dedicated deep learning machines, ready for your next project.

MACHINE	GPU	CPU	RAM	DISK	PRICE
gpu	Tesla K80 12 GB Memory	4 vCPUs	61 GB	200 GB SSD	$1.20/hr $0.00033/sec
gpu2	Tesla V100 16 GB Memory	8 vCPUs	61 GB	200 GB SSD	$4.20/hr $0.00117/sec
cpu		2 vCPUs	8 GB	200 GB SSD	$0.19/hr $0.00005/sec
cpu2		8 vCPUs	32 GB	200 GB SSD	$0.48/hr $0.00013/sec

https://www.floydhub.com/pricing

FLOYDHUB

	Beginner	Data Scientist	Teams
INFRASTRUCTURE			
GPU Access	Tesla K80, Tesla V100	Tesla K80, Tesla V100	Tesla K80, Tesla V100
Machine Priority	Low priority	High priority	High priority
Deep Learning Frameworks	TensorFlow, PyTorch, Keras, others	TensorFlow, PyTorch, Keras, others	TensorFlow, PyTorch, Keras, others
Install Python Packages	✔	✔	✔
Customer Support	Forum / Email	Chat	Chat
Private Projects and Datasets		Unlimited	Unlimited
Public Projects and Datasets	Unlimited	Unlimited	All projects / datasets are private
TRAINING			
Training Duration	1 day	7 days	Custom
Storage	10 GB	100 GB	100 GB
Pricing (per month): https://www.floydhub.com/pricing	$0	$9	$99

Vienna Deep Learning Meetup

FLOYDHUB

Everything you need for deep learning

- **Jupyter Notebooks**
  One-click GPU-powered interactive notebooks.
- **Framework Support**
  TensorFlow, PyTorch, Keras, and more - fully-configured.
- **Per Second Billing**
  Pay only for what you use, per second.
- **Version Control**
  Full environment reproducibility.
- **Parameter Sweeping**
  Run concurrent jobs for optimization.
- **Public Datasets**
  Train using data from the open source community.
- **Deploy Trained Models**
  Create model-serving REST API in a single command.
- **Tensorboard**
  Visualize and explore your live experiments.

GPU Plans

Dedicated Virtual Machines to Speed Up Complex Compute Jobs with RTX 6000 GPUs

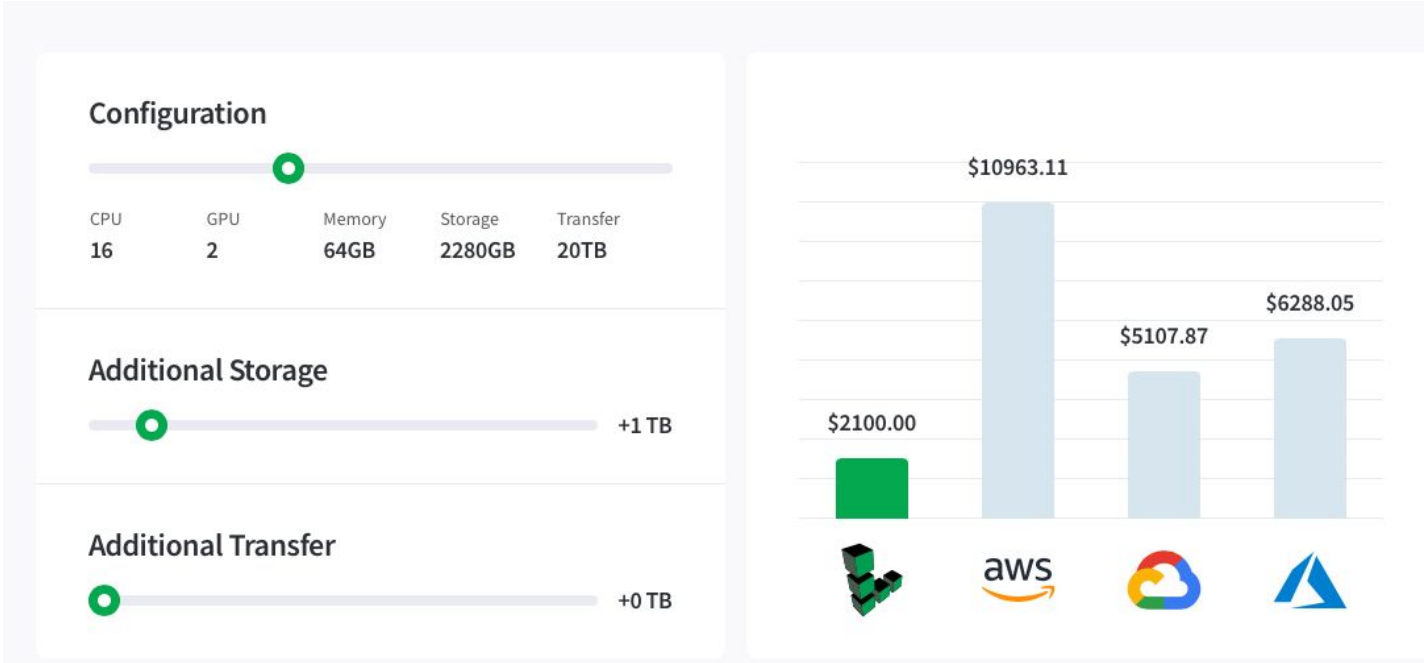
Learn More >

RAM	CPU	Storage	GPU Cards	Transfer	Network In	Network Out	Price
32 GB	8 Cores	640 GB SSD	1	16 TB	40 Gbps	10000 Mbps	$1000 / mo ($1.5 / hr)
64 GB	16 Cores	1280 GB SSD	2	20 TB	40 Gbps	10000 Mbps	$2000 / mo ($3.0 / hr)
96 GB	20 Cores	1920 GB SSD	3	20 TB	40 Gbps	10000 Mbps	$3000 / mo ($4.5 / hr)
128 GB	24 Cores	2560 GB SSD	4	20 TB	40 Gbps	10000 Mbps	$4000 / mo ($6.0 / hr)

https://www.linode.com/products/gpu/

linode

Pricing Comparison (Example):



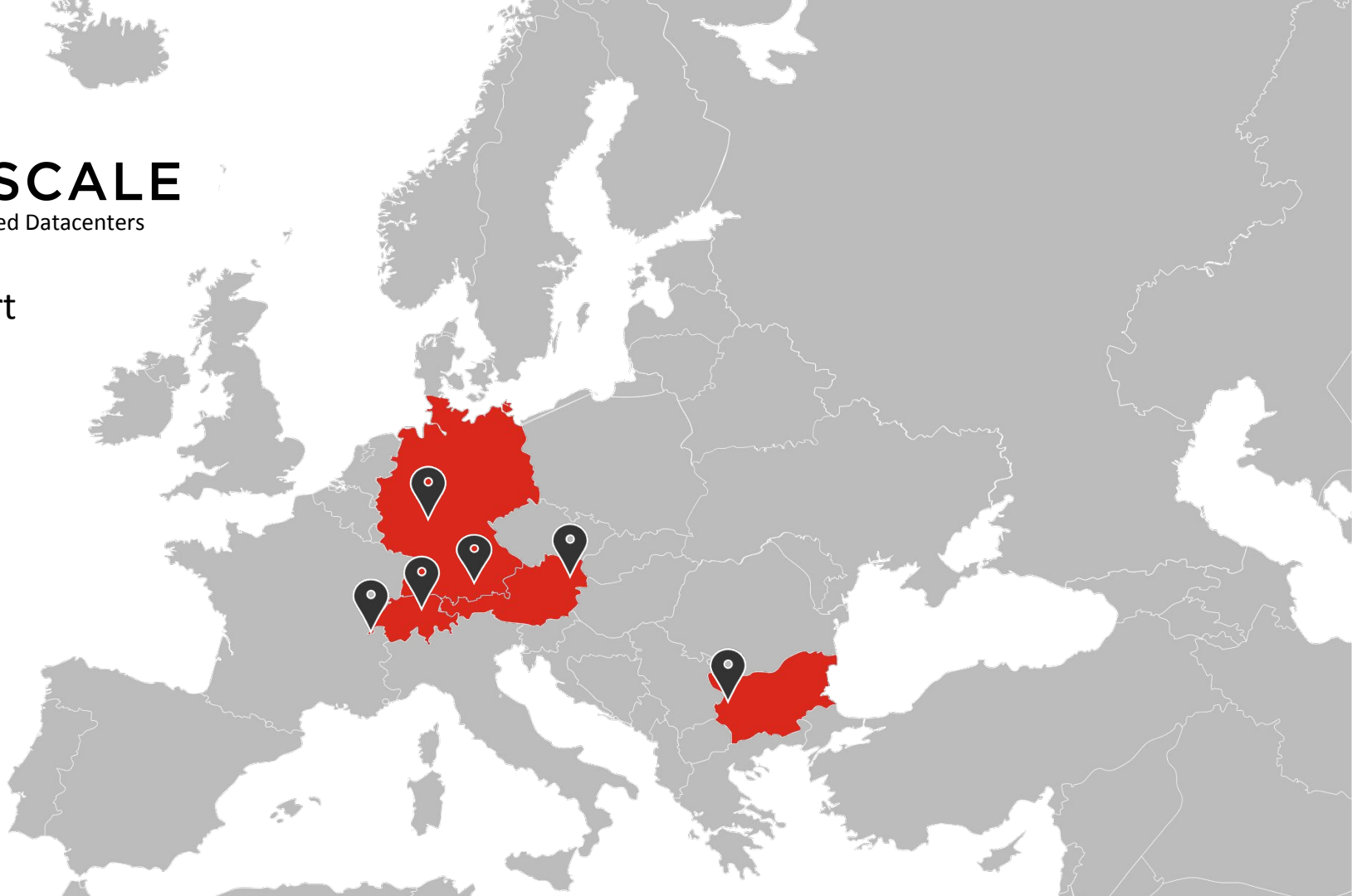
https://www.linode.com

Vienna Deep Learning Meetup

EXOSCALE

ISO 27001 Certified Datacenters

Frankfurt
Munich
Vienna
Zurich
Geneva
Sofia



NVIDIA Tesla V100

EXOSCALE

- From 1 up to 4 **NVIDIA Tesla V100** per GPU2 instance
- 5,120 CUDA cores per card
- 640 dedicated Tensor Cores per card
- 16 GB of dedicated RAM per card
- Up to 48 CPUs per GPU2 instance
- Up to 1.6 TB of SSD local storage

	RAM	CPU Cores	GPU Cards	Price (EUR)
Small	56 GB	12 Cores	1 GPU	1.25015/hr
Medium	90 GB	16 Cores	2 GPU	1.83103/hr
Large	120 GB	24 Cores	3 GPU	2.41191/hr
Huge	225 GB	48 Cores	4 GPU	3.01803/hr

NVIDIA Tesla P100

EXOSCALE

- From 1 up to 4 **NVIDIA Tesla P100** per GPU instance
- 3584 CUDA cores per card
- 16 GB of dedicated RAM per card
- Up to 48 CPUs per GPU instance
- Up to 1.6 TB of SSD local storage

	RAM	CPU Cores	GPU Cards	Price (EUR)
Small	56 GB	12 Cores	1 GPU	1.06200/hr
Medium	90 GB	16 Cores	2 GPU	1.55574/hr
Large	120 GB	24 Cores	3 GPU	2.04949/hr
Huge	225 GB	48 Cores	4 GPU	3.01803/hr

Real-World Benchmarks

EXOSCALE

CIFAR10 Tensorflow Example

Instance type	Boot time	Processing Time	Costs / hour	Total Costs
Exoscale V100 GPU Instance	30 sec	10 sec	1.25015 EUR	~0.013 EUR
Exoscale Extra-Large	30 sec	5 mins	0.16164 EUR	~0.014 EUR
Our Office Laptop	Don’t ask…	20 mins		Our Sanity

https://github.com/tensorflow/models/tree/master/tutorials/image/cifar10

Why Exoscale?

EXOSCALE

European
Built & Managed in Europe

Transparent Billing
Simple, understandable cost structure

Performance
Beats most cloud providers

Compliant
GDPR, ISO27001, & more

Affordable
30-40% lower costs than big hyperscalers

Fast Startup Times
New Virtual Machines in under 30 seconds

Automated
Terraform, Ansible, & more

Free Internal Bandwidth
Even between zones!

GPUs
Tesla V100 and P100 GPUs for Machine Learning

SLA
99.95%

No Vendor Lock-In
Standardized, vendor-independent systems.

Dedicated Resources
Get real CPU cores and dedicated resources.

Interested?

EXOSCALE

Talk to this guy

Janos Pasztor

Cloud Solution Architect @ Exoscale

janos.pasztor@a1.digital

Paperspace

Dedicated GPU

GPU+	P5000	P6000	V100
$0.51 / hour	$0.78 / hour	$1.10 / hour	$2.30 / hour
+ $5 monthly storage fee	+ $5 monthly storage fee	+ $5 monthly storage fee	+ $5 monthly storage fee
NVIDIA Quadro P4000 with 1792 CUDA cores.	NVIDIA Quadro P5000 with 2560 CUDA cores.	3840 CUDA cores and up to 432 GB/s memory bandwidth.	112 TeraFLOPs VOLTA chipset. The most powerful GPU in the world.
• 30GB RAM • 8 x CPU • 8 GB GPU DEDICATED	• 30GB RAM • 8 x CPU • 16 GB GPU DEDICATED	• 30GB RAM • 8 x CPU • 24 GB GPU DEDICATED	• 30GB RAM • 8 x CPU • 16 GB GPU DEDICATED
SIGN UP	SIGN UP	SIGN UP	SIGN UP

+ Collaboration Tools, Workflow, Templates etc.

Paperspace

GPU Instances

A range of GPU types for high-end workloads

K80	P100	V100	V100 x8
$0.25 / hour	$0.59 / hour	$1.15 / hour	$8.43 / hour
12GB GDDR5 DEDICATED	16GB GDDR5 DEDICATED	16GB GDDR5 DEDICATED	128GB GDDR5 DEDICATED
12GB RAM	24GB RAM	30GB RAM	130GB RAM
2 vCPU	4 vCPU	8 vCPU	20 vCPU
480 GB/s memory bandwidth	732 GB/s memory bandwidth	900 GB/s memory bandwidth	7,200 GB/s memory bandwidth
2,496 CUDA cores	3,584 CUDA cores	5,120 CUDA cores	40,960 CUDA cores

New! TPU

advanced AI accelerator chip

TPUv2
$8.42 / hour
16GB DEDICATED
2,400 GB/s memory bandwidth
180 TeraFLOPs

https://www.paperspace.com/pricing

gradient
by Paperspace

Product Learn Enterprise Pricing Sign in Sign up free

NEW Forbes covers launch of Gradient Multi-Cloud →

Production tools for Machine Learning.

From exploration to deployment, Gradient enables individuals and teams to quickly develop, track, and collaborate on Machine Learning models of any size and complexity.

GET STARTED CONTACT SALES

LeaderGPU PRICES NEWS CATALOG HELP API ABOUT US

SALE 20%

on 8 x 2080 Ti 384 GB

~~€ 2 291,95 month~~	€ 1 833,56 month
~~€ 572,99 week~~	€ 458,39 week
~~€ 0,11 minute~~	€ 0,08 minute

ORDER NOW

Vienna Deep Learning Meetup

LeaderGPU

4 x 1080 Ti 128Gb RAM

RECOMMENDED

GPU speed:

●●●○○

Server Configuration:

GPU: 4 pcs 1080Ti (Each GPU card has: 3584 CUDA cores GeForce GTX1080 Ti)

CPU: 2 x Intel Xeon E5-2609v4 1.7 GHz

RAM: 128 GB RAM

SSD: 480 GB SSD

Internal network: 10 Gbps Port

Best for:

» Blockchain processing
» Universal GPU card

OS:

- CentOS 7
- CentOS 8
- Ubuntu 16.04
- Ubuntu 18.04
- Windows 2016

Prices:

- *893.7 € / month*
- *223.43 € / week*
- *0.04 € / minute*
- *0 € - setup fee*

(all prices without VAT)

For **1 MONTHS** you will pay: **893.7 EURO**

Available to run

ORDER NOW

2 x P100 PCI

GPU speed:

●●●○○

Server Configuration:

GPU: 2 pcs P100 (Each GPU card has: 3584 CUDA cores)

CPU: 2 x Intel Xeon E5-2609v4 1.7 GHz

RAM: 256 GB RAM

SSD: 960 GB SSD

Internal network: 10 Gbps Port

Best for:

» Deep Learning
» AI Training
» AI Inference
» HPC

OS:

- CentOS 7
- CentOS 8
- Ubuntu 16.04
- Ubuntu 18.04
- Windows 2016

Prices:

- *1119.4 € / month*
- *279.85 € / week*
- *0.05 € / minute*
- *0 € - setup fee*

(all prices without VAT)

For **1 MONTHS** you will pay: **1119.4 EURO**

Available to run

ORDER NOW

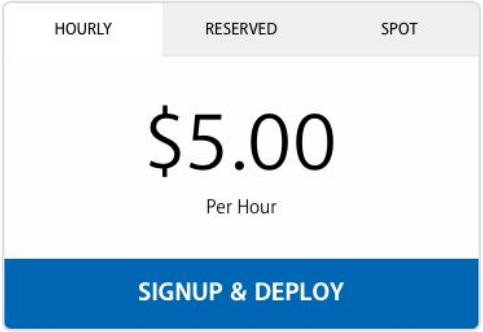
+ many more options: https://www.leadergpu.com

Vienna Deep Learning Meetup

packet

We all know that this is the GPU server we would build in our basement...if we had a small powerplant and a lot of cooling down there! Fast cores (thanks to some Intel Xeon 6126 processors), fast RAM, and of course dual Nvidia V100 32GB cards with NVLINK.

Hardware Specs	
CPU	24 Physical Cores @ 2.6 GHz (2 X XEON GOLD 6126)
Memory	192 GB of DDR4 ECC RAM
Storage	2 × 120 GB SSD boot
Storage	2 x 480GB SSD
GPU	2 x Nvidia V100 32GB GPU w/NVLINK



2x Nvidia V100

Currently only this GPU option. https://www.packet.com/cloud/servers/g2-large/

Vienna Deep Learning Meetup

OVHcloud

GPU

Experience our most powerful public cloud instances, up to 1,000 times faster than a CPU for parallel processing

Find out more

Name	Memory	vCore	GPU	Storage	Public network	Private network	Price
t1-45	45 GB	8	Tesla V100 16 GB	400 GB SSD	2 Gbps guaranteed	4 Gbps max.	**$2.661 /hour**
t1-90	90 GB	18	2×Tesla V100 16 GB	800 GB SSD	4 Gbps guaranteed	4 Gbps max.	**$5.44 /hour**
t1-180	180 GB	32	4×Tesla V100 16 GB	50 GB SSD + 2 TB	10 Gbps	4 Gbps max.	**$10.88 /hour**

Up to 1,000 times faster than a CPU on parallel computing, NGC integration makes it easy to use Tensorflow, Caffe, Mxnet, and much more.

https://www.ovhcloud.com

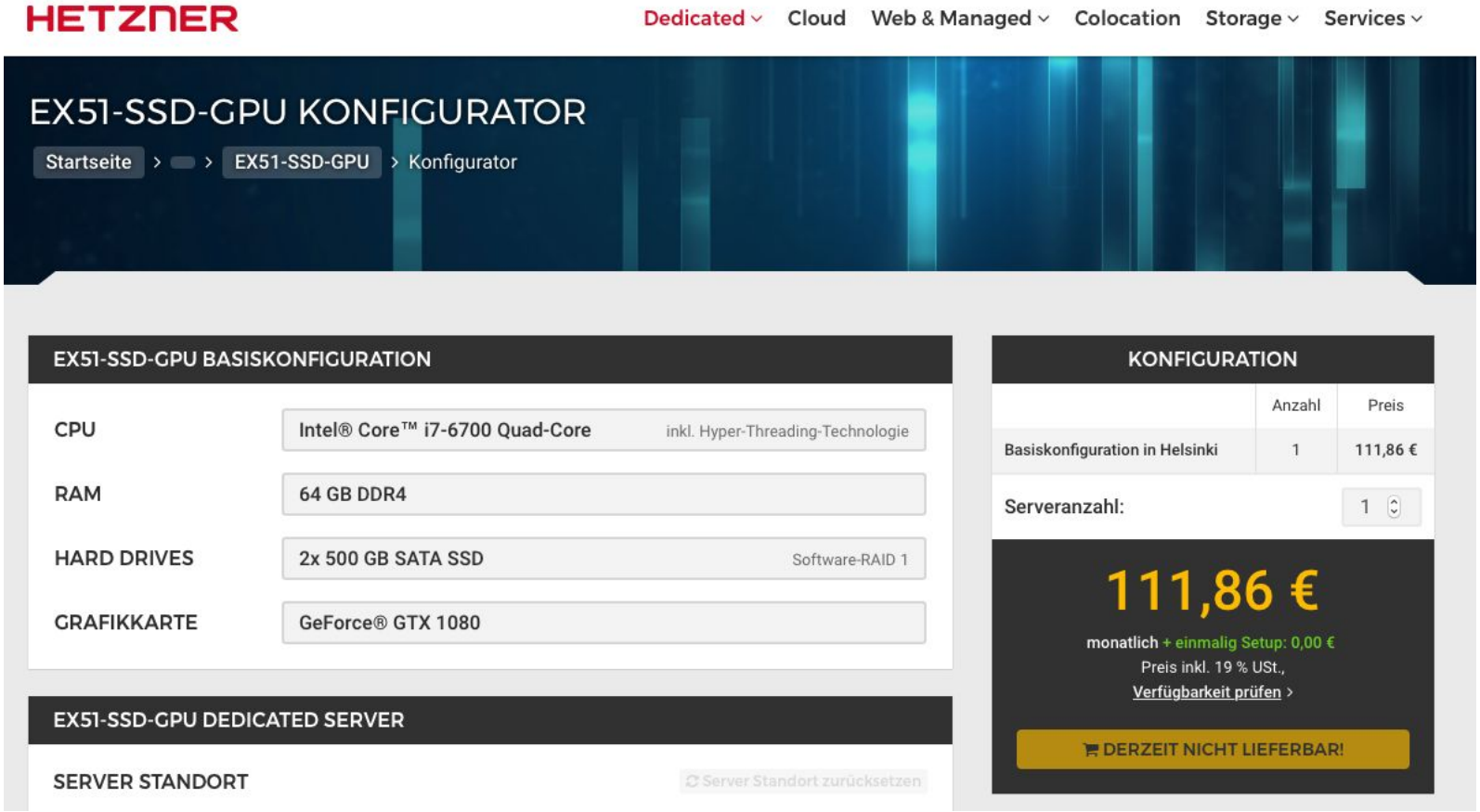
Lambda

Each Lambda gpu.4x instances has four (4) vGPUs with 11 GB of VRAM, 3584 CUDA Cores, and performance equivalent to 4x 1080 Ti GPUs. For exact details see below:

Instance Type	GPUs	GPU Memory (GB)	vCPUs	System Memory (GB)	Instance Speed	On-Demand Hourly Rate	1-yr Reserved Effective Hourly Rate
Lambda gpu.4x	4 GPUs	11 GB / GPU	8 vCPU Cores @ 3.5 GHz	32 GB DDR4	Up to 10 Gbps	$1.50	$1.25

- GPU servers optimized for deep learning
- On-demand Pricing: $0.375 / hour / GPU, Reserved Pricing: $0.3125 / hour / GPU
- TensorFlow, PyTorch, Keras, and Caffe preinstalled on every server
- Less than half the cost of AWS and Azure

https://lambdalabs.com/service/lambda-gpu-cloud-faq

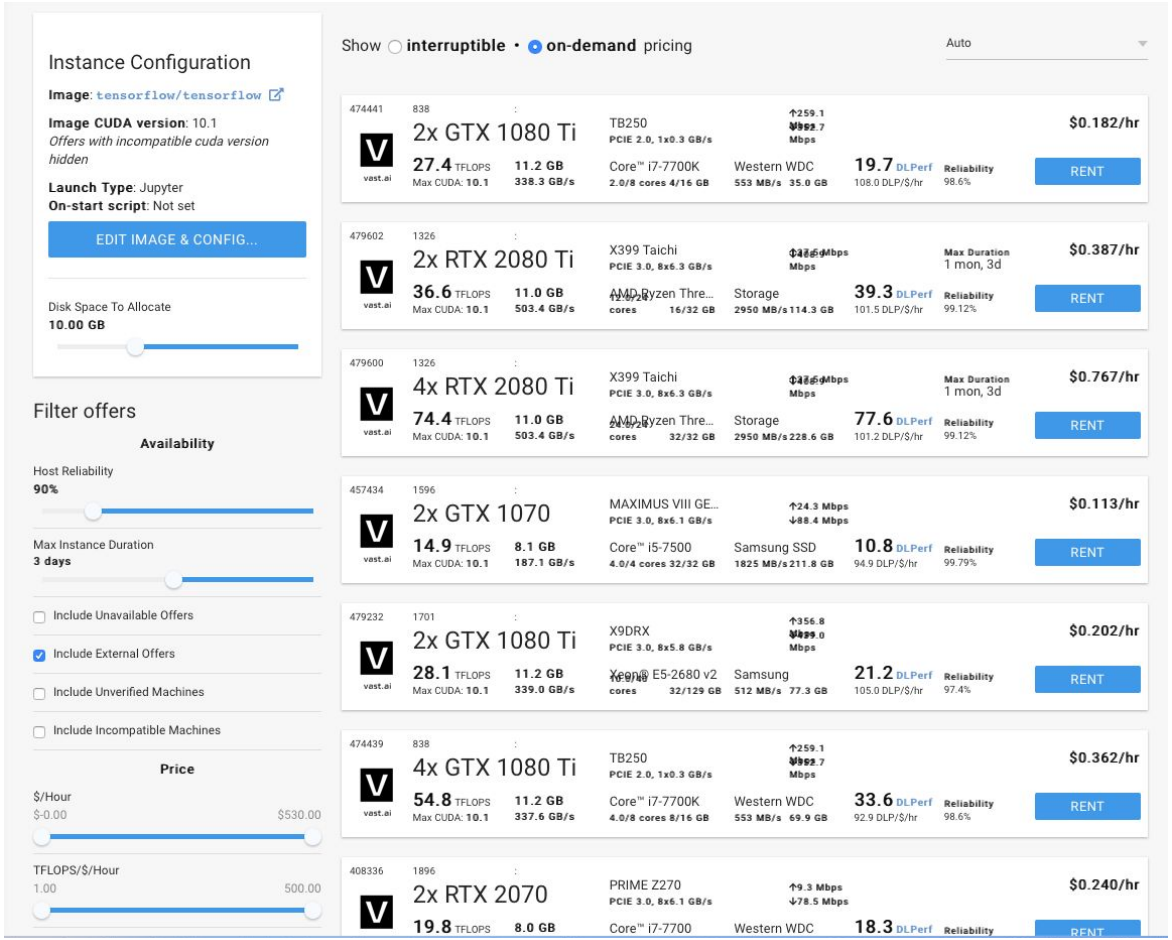


Note: only 1 NVidia GTX 1080 GPU per instance possible
https://www.hetzner.de/dedicated-rootserver/ex51-ssd-gpu/konfigurator

Vienna Deep Learning Meetup

vast.ai

GPU Compute “Search Engine” for GPU power from different vendors



https://vast.ai

Vienna Deep Learning Meetup

Who provides which GPU types?

Provider	K80	V100	P100	Quadro	GTX 1080Ti	RTX 2080Ti	Other
AWS	✔️	✔️					
FloydHub	✔️	✔️					
Linode				**RTX 6000**			
Exoscale		✔️	✔️				
Paperspace	✔️	✔️	✔️	**P4/5/6000**			**TPU**
LeaderGPU		✔️	✔️		✔️	✔️	**Tesla T4**
Packet		✔️					
OVHcloud		✔️					
LambdaLabs					✔️ equivalent		
Hetzner					**GTX 1080**		
GPU Eater							**AMD Radeon**

Vienna Deep Learning Meetup

Deep Learning Hardware Overview: What and where to buy or rent

Jan Schlüter
JKU Linz

René Donner
contextflow

Thomas Lidy
Musimap