USING GPUS IN THE ELKH CLOUD

THE SECOND PART

MEZO.GYORGY@WIGNER.HU

KADLECSIK.JOZSEF@WIGNER.HU

SUMMARY OF THE PRESENTATION

- The existing GPU infrastructure
- The selected projects
- Analysis of the selected projects from the Data Center's perspective
- The planned expansion of GPU capacities and services

Wigner Data Center

One Supermicro Server with 4 NVIDIA V100 SXM2 16 GB Cards

- 2 x Intel Xeon Processor E5-2620 v4
- 16 x 32 GB DDR4-2400 RAM
- 8 x Intel S3520 480 GB SSD

The GPU server was used separately from OpenStack Cloud. It is available via VPN connection.

Technologies used:

- oVirt with PCI passtrough
- OpenShift
- Docker
- Centos 7 linux Bare Metal

Project based support and software installation

Institute for Computer Science and Control

Three supermicro servers

- 4(1+1+2) x NVIDIA Tesla K80 Cards
- 512 GB RAM

The GPU servers are in the Openstack Cloud

Technologies used:

- PCI passtrough
- Special Flavor for GPU instances

Project	Institute	Software used	Technology
Planet Formation	Konkoly Observatory, Zsolt Regály	Fargo, Fargo3D, C and C++ programs with multi GPU	Docker, Headless VNC, No Machine, VirtualGL TurboVNC for remote accelerated graphics
3D Rendering for 3D Animated Movies of the Surface of Mars	Konkoly Observatory, Vilmos Steinmann	Blender with multi GPU	Centos 7 Linux on Bare Metal
Image Based Classification of Variable Stars	Konkoly Observatoy, Tamás Szklenár	Tensorflow, Keras, Jupyter	Centos 7 Linux on Bare Metal
Digital Pathology with AI	Eötvös Lóránd University, Péter Pollner	TensorFlow, Keras, Jupyter	Centos 7 Linux on Bare Metal
Stress related sleep disorders	Institute of Experimental Medicine, László Acsády	DeepLabCut for animal (mice) motion detection, Tensorflow, Python 3	Centos 7 Linux on Bare Metal
The development of digital twin model in Ansys for the new bridge of Komárom	Budapest University of Technology and Economics, Attila Joó	Ansys Mechanical (FEA)	Centos 7 Linux on Bare Metal
Ansys Benchmarking and Optimalization with multi GPU	eCon Engineering Ltd. (Official dis, János Szegletes	Ansys Mechanical, Ansys Fluent, Ansys CFX, Ansys Electronics	Centos 7 Linux on Bare Metal

EXPERIENCES AND USER NEEDS 5/1

Planet Formation project	**- oVirt without fast file server and with a single GPU server is too slow** **- GPU server used via Docker as fast as a bare metal machine** **- Remote rendering - VirtualGl, TurboVNC - combining with docker is a very complex environment but works** **- It is difficult to select which card works on rendering and which one on calculations, but since the Docker is developed**
	- These type of projects with custom C/C++ programs can use any number of GPUs and can use any amount of allocated time interval very effectively also. There is almost no idle time. - The GPU's role in these calculations is much more pronounced than the calculation on the CPU. Double-Precision performance is important. - Based on the data center's experience, the project uses docker technology in all of its own GPU's servers in the Konkoly Observatory
3D Rendering for 3D Animated Movies of the Surface of Mars	- The Blender rendering program used the multiple GPU cards very effectively. High single precision performance is enough. NVIDIA cards are the standard, but recently AMD cards also good performers. - The CPU's were also used heavily - The high Data IO was assisted by local SSD array

EXPERIENCES AND USER NEEDS 5/2

Image Based Classification of variable Stars	**- The Machine Learning Group of the Konkoly Observatory started the calculations on WDC's GPU server with the Darknet's neural networks** **- It was interesting to see the needs of the projects as the function of the learning curve** **- Significant idle times between the searching and choosing the right environment for development** **- High Single Precision performance is important**
	- Many research groups are starting to use ML for data discoveries. To start their own development , a "small" GPU's "part" is enough: virtual GPU technology. NVIDIA is more matured in it than AMD. - A preinstalled environment is very helpful for these users: Tensorflow, Keras, PyTorch - JupyterHub with GPU's on Kubernetes for many users - OpenShift to cover all the development phases but too complex - It is important to have courses on how to use the GPU and ML'technics
	- ELKH's course for the novice Cloud' users is scheduled on 2020.10.22. from 13:00-16:45 PM.

EXPERIENCES AND USER NEEDS 5/3

Digital Pathology with AI	**- ELTE Department of Physics of Complex Systems with cooperation of Semmelweis University** **- Labelled scanned tissue images of patients. So privacy is a basic requirement.** **- Expert users** **- TensorFlow, Keras … Single Precision** **- Serious challenge with huge images and plenty of data so every available resource is required**
	- The projects of expert users are also in development phases so no need to immediately allocate huge resources - With the help of orchestration “burst” infrastructure can be used. It is possible to schedule the calculation depending on the project phase and on the current resources - Emphasis should be placed on planning the project and architecting the planned infrastructure

EXPERIENCES AND USER NEEDS 5/4

Stress Related Sleep Disorders	**- The researchers had the concrete program for the GPU : DeepLabCut. It is for animal pose estimation.** **- Application of pretrained deep neural networks** **- Docker version is also available** **- It is for NVIDIA**
	- The support was the installation of the program and solving the data transfer. Previously the researcher installed this program on windows desktop without GPU. - Process videos about mice on a regular basis
The development of digital twin model in Ansys for the new bridge of Komárom	- Operation based on the digital model. Determine the time of maintenance from the collected data of the field sensors. - Needs huge capacity or dedicated infrastructure
	- Ansys Mechanical installation on Centos 7 linux - Running and developing the basic building block models - Study how to scale up with the help of GPU and clustering the calculations

EXPERIENCES AND USER NEEDS 5/5

Ansys Benchmarking and Optimalization with multi GPU	**- Ansys has many products: Mechanical, Electromagnetics, CFD, Fluent …** **- How to scale the model calculations up with GPU** **- How to scale up with clustering** **- Which solution is cheaper and more effective considering the license fee: more GPU or more CPU?** **- Find a convenient way tor engineers to remotely work with 3D CAD programs** **- How to measure the performance on GPU? How to measure the CLOUD performance?** **- Compare HPC CLOUD providers: Rescale, Penguin Computing, UberCloud, CADFEM Engineering Simulation** **- Privacy is very important in automotive industry**
	- Finding and running benchmark programs with a field specialist engineer - Graphical Engineering Desktop on the CLOUD - GPU and CPU balance - NVIDIA Virtual Compute Server software

PLANNED EXPANSION OF GPU CAPACITIES AND SERVICES

WDC

4 GPU Servers

- Minimum 4 NVIDA A100 40 GB (Maximum 8 cards)
- 100 G network card
- 2 high performance CPUs
- NVIDIA Virtual Compute Server (vCS) licenses
- 512 GB RAM

IAAS

-oVirt/OpenStack PCI passtrough

- OpenStack Bare Metal
- Special Flavors

PAAS

- Kubernetes, Jupyterhub
- Orchestrated Solutions

Institute for Computer Science and Control

8 GPU Servers

- 4 NVIDIA V100 32 GB NVLINK
- 2 high performance CPUs
- NVIDIA Virtual Compute Server (vCS) licenses
- 768 GB RAM

IAAS

- OpenStack PCI passtrough
- Special Flavors

PAAS

- Many orchestrated Solutions