Running Applications on a Hybrid Cluster

A.V. Bogdanov[1], I.G. Gankevich[1],
V.Yu. Gaiduchok[2], N.V. Yuzhanin[1]

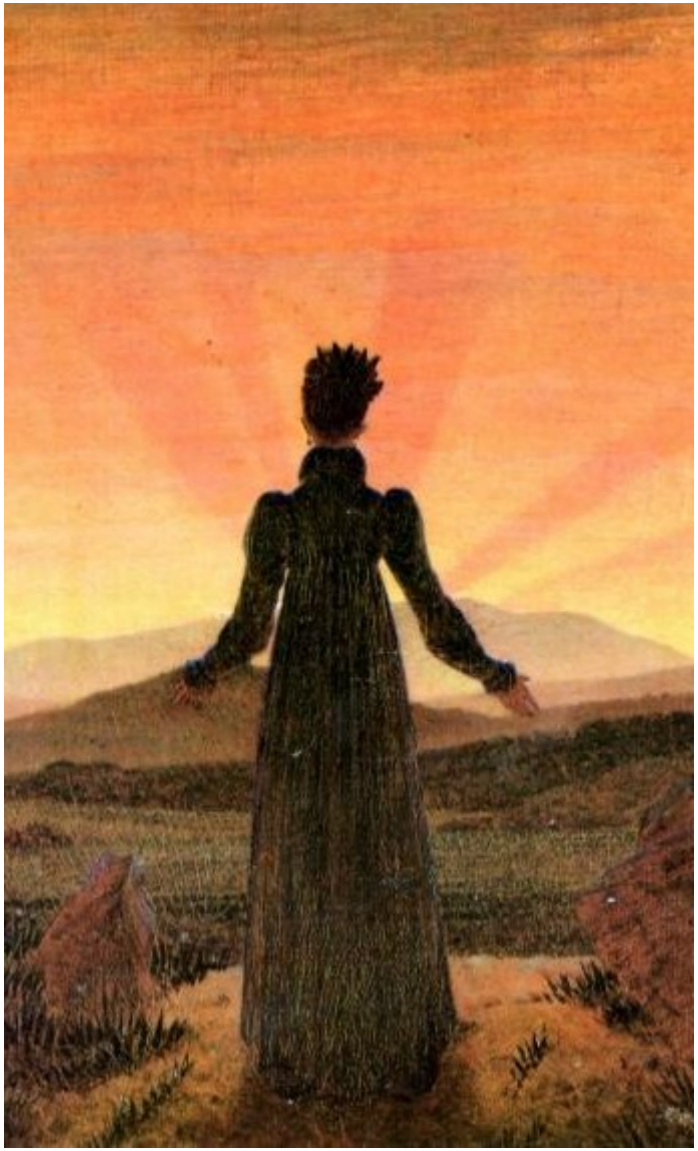
[1] *Saint Petersburg State University, Russia*
[2] *Saint Petersburg Electrotechnical University "LETI", Russia*

GRID

2014

Agenda

- Introduction
- Use Cases
- Platform Specifications
- Basic Tests
- Applications
- Conclusions



Introduction

Cluster is a set of connected computers that are used as a single system and dedicated to solve particular task.

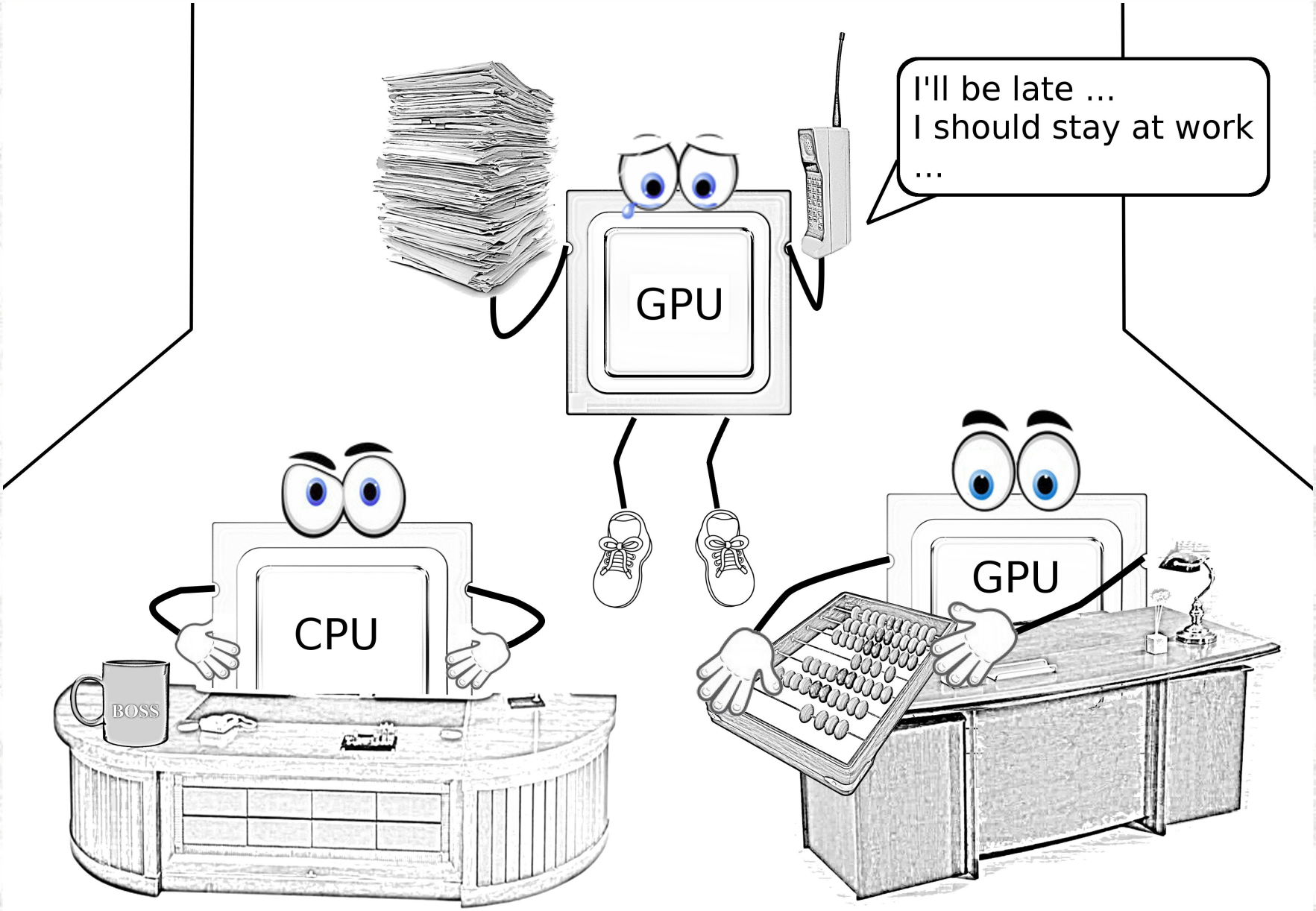
Cluster example

Hybrid Cluster is a computational cluster with nodes that have special purpose accelerators:

- GPUs (NVIDIA, AMD);
- Cell (IBM);
- MIC (Intel);
- ...

We will use GPUs NVIDIA **CUDA**.

Introduction

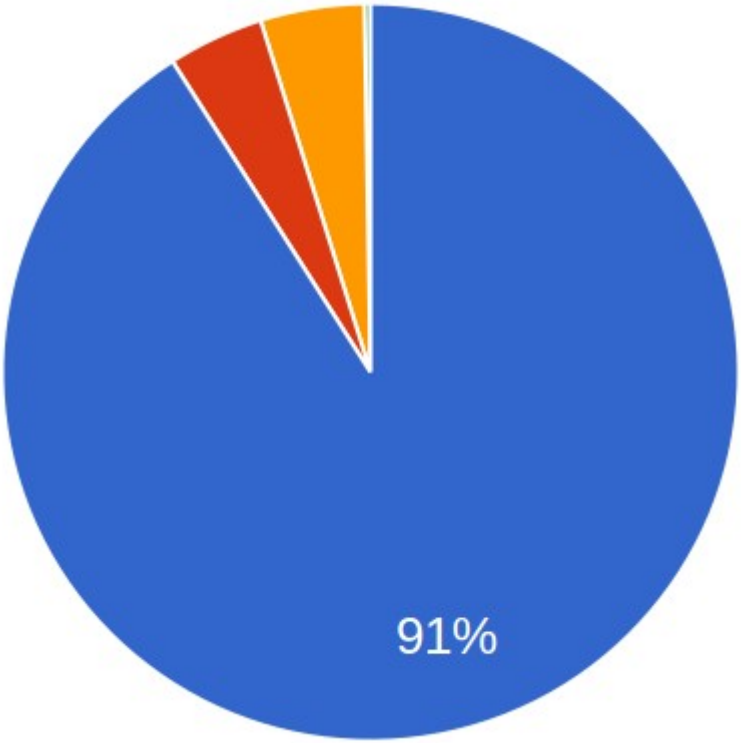


Hybrid Node Inc.

Introduction

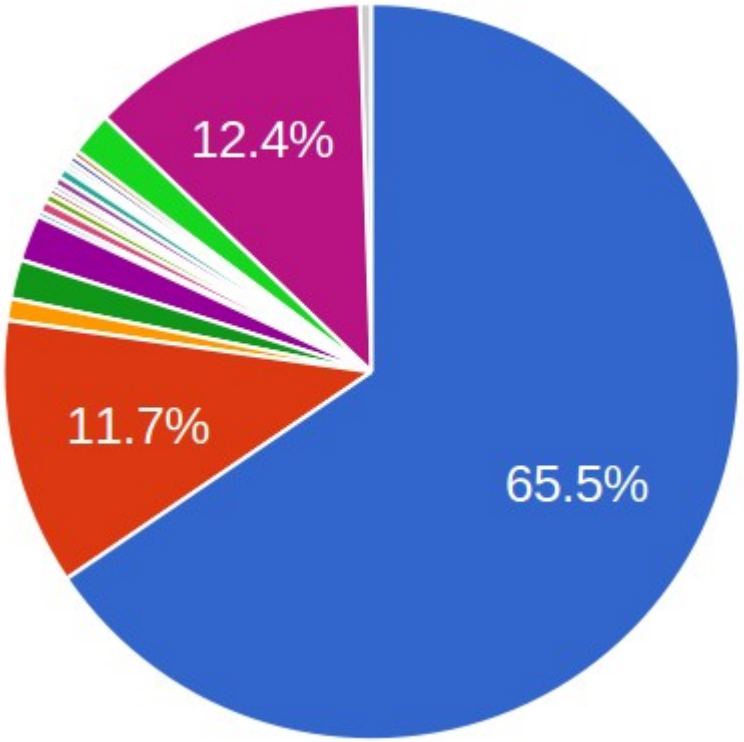
TOP 500: Accelerator/Co-Processor Performance Share

June 2010



None
IBM PowerXCell 8i
NVIDIA 2050
Clearspeed CSX600

June 2014



None
NVIDIA K20x
NVIDIA 2090
Intel Xeon Phi 5110P
NVIDIA 2050
Nvidia K40m
Nvidia K20m
...

Use Cases

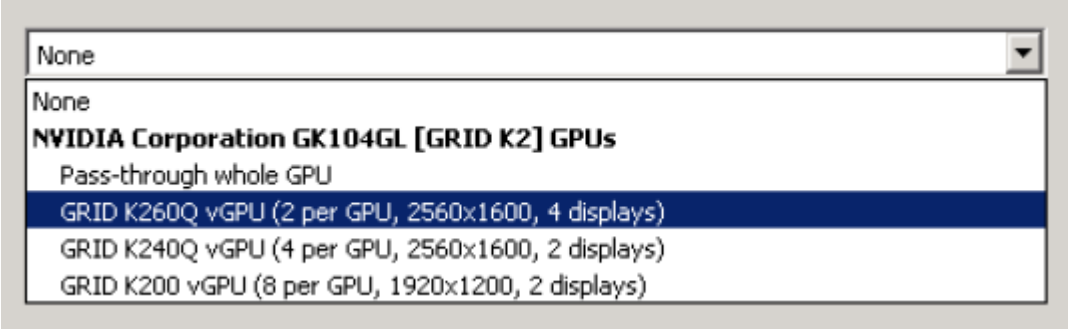
GPU use cases:

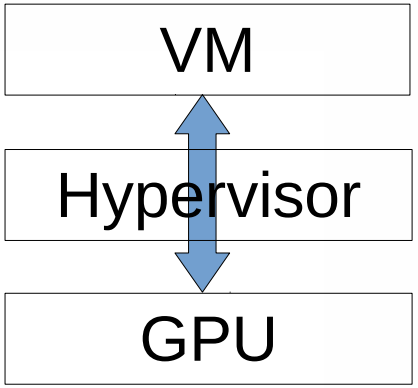
- Conventional GPU usage on PCs;
- Virtual machines with GPUs (visualization):
  - Dedicated GPUs (passthrough);
  - Virtual GPUs (virtual device for VM, e.g virtual GPUs with XenServer);
- GPGPU (computations):
  - Dedicated servers;
  - VMs;
  - **Clusters**.

Use Cases

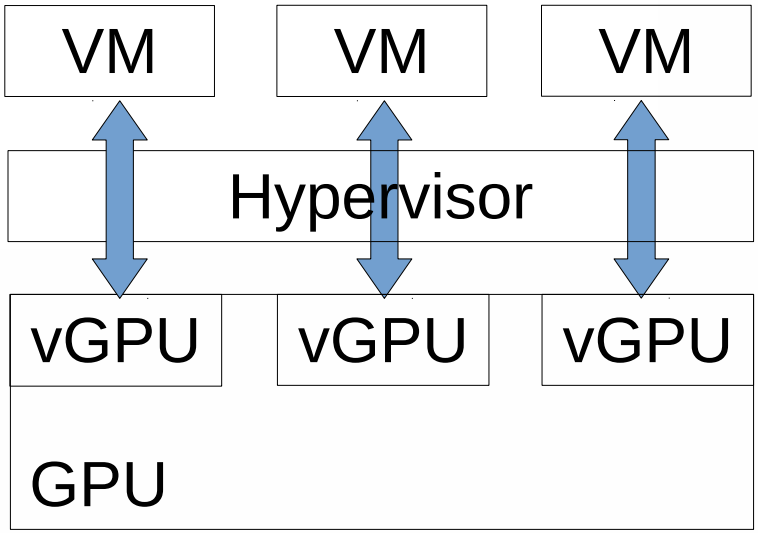
Virtual machines with GPUs

Dedicated GPUs (passthrough)





Virtual GPUs



Use Cases

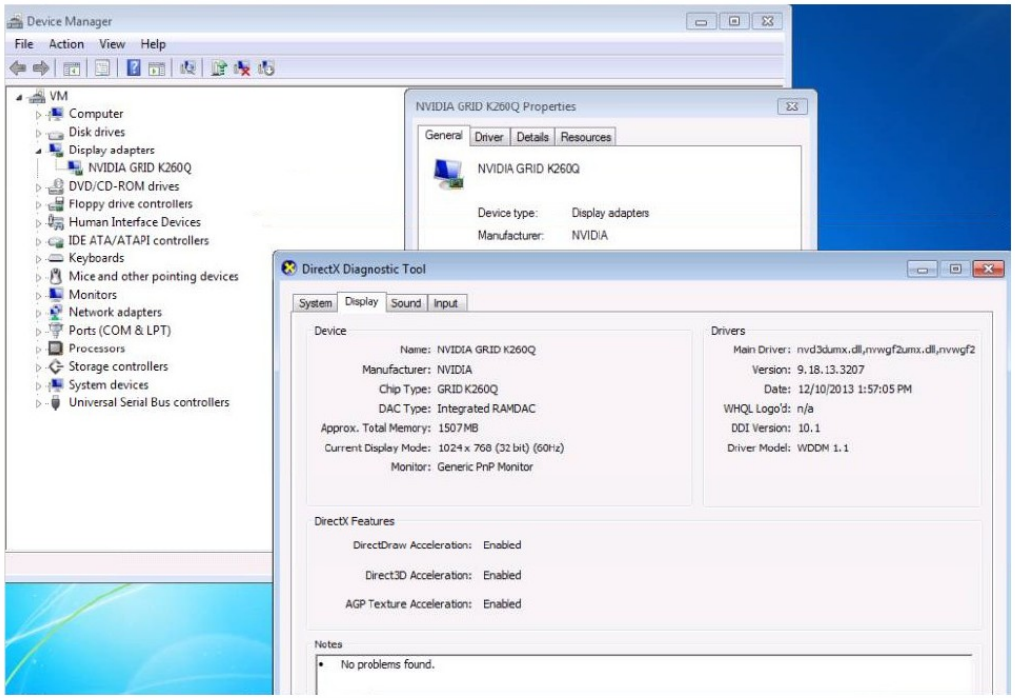
Virtual machines with GPUs

Virtual GPUs are supported only for windows VMs (there is only NVIDIA driver for windows).

Actions to do:

Server side:
- Install XenServer
- Install NVIDIA Virtual GPU Manager.

VM side:
- Install XenTools;
- Install NVIDIA GPU driver for VM.

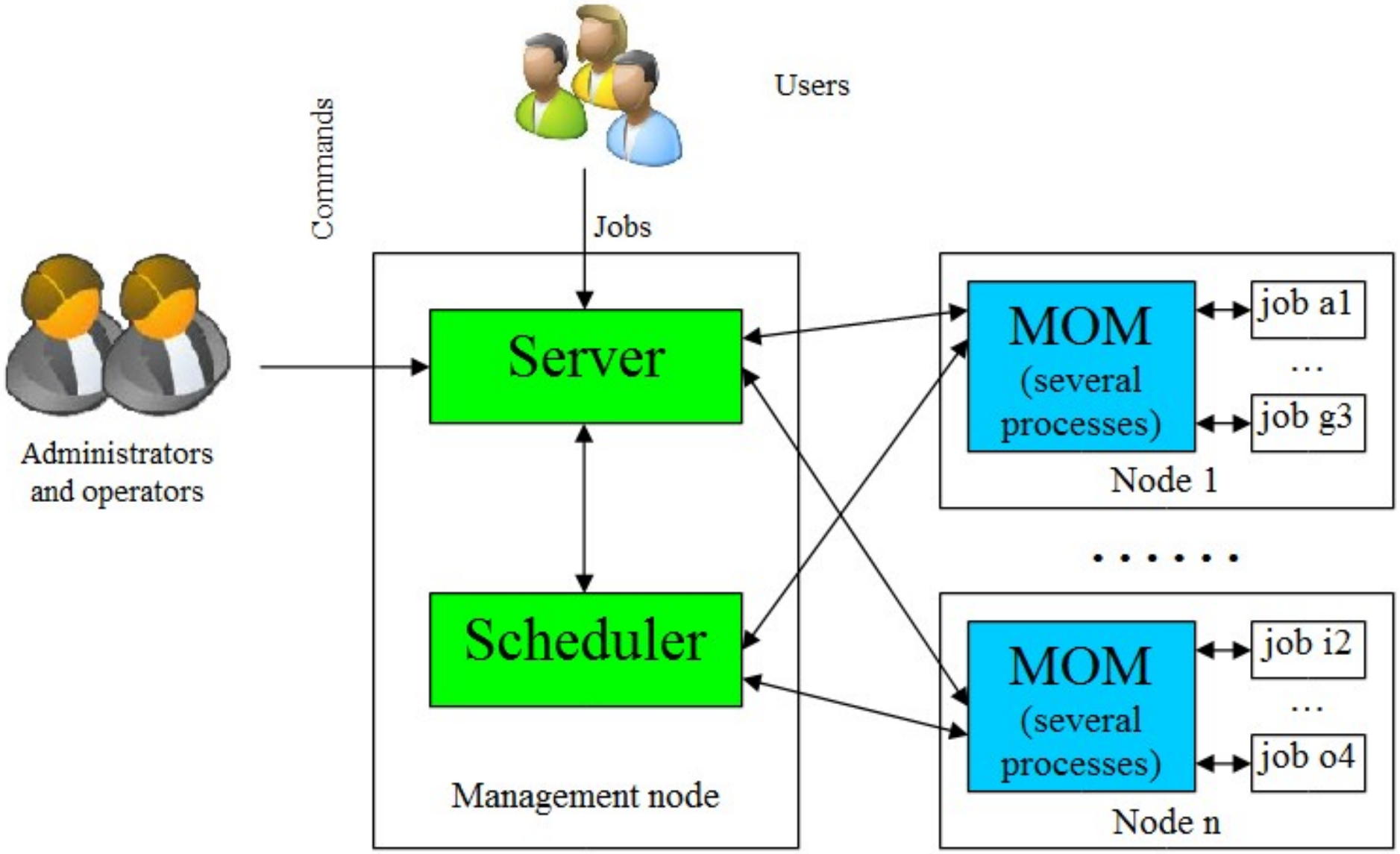


NVIDIA

Xen Project

Use Cases

Cluster with PBS



Use Cases

- Administrator specifies available resources for each node.
- User writes a script and submits the script as a job to a queue requesting some resources.
- If user is allowed to do it, the job is enqueued.
- PBS scheduler decides whether this job can be started taking into account resources requested by user, nodes load, queue priority, limitations for current user (or his group).
- When it’s possible job is started on cluster node(s).
- After the job finishes user receives stdout and stderr of the job (as 2 files).

Use Cases

PBS queue summary.

- Queues have a priority.
- Queues are associated with nodes.
- The same node can be assigned to several queues.
- Queues have different limitations.
- One can specify authorization rules for queues.
- Queues are configured with PBS server commands.

Platform Specifications

Node characteristics:

- CPU: 2 x Intel Xeon X5650 2.67 GHz
  (total 12 cores);
- RAM: 96 GB;
- GPU: 3 or 8 NVIDIA Tesla M2050 per node;
- NET: InfiniBand QDR (40 Gb/sec);
- Cluster management: PBS
- OS: CentOS 6.4;
- CUDA Toolkit 5.5.

PBS Works™

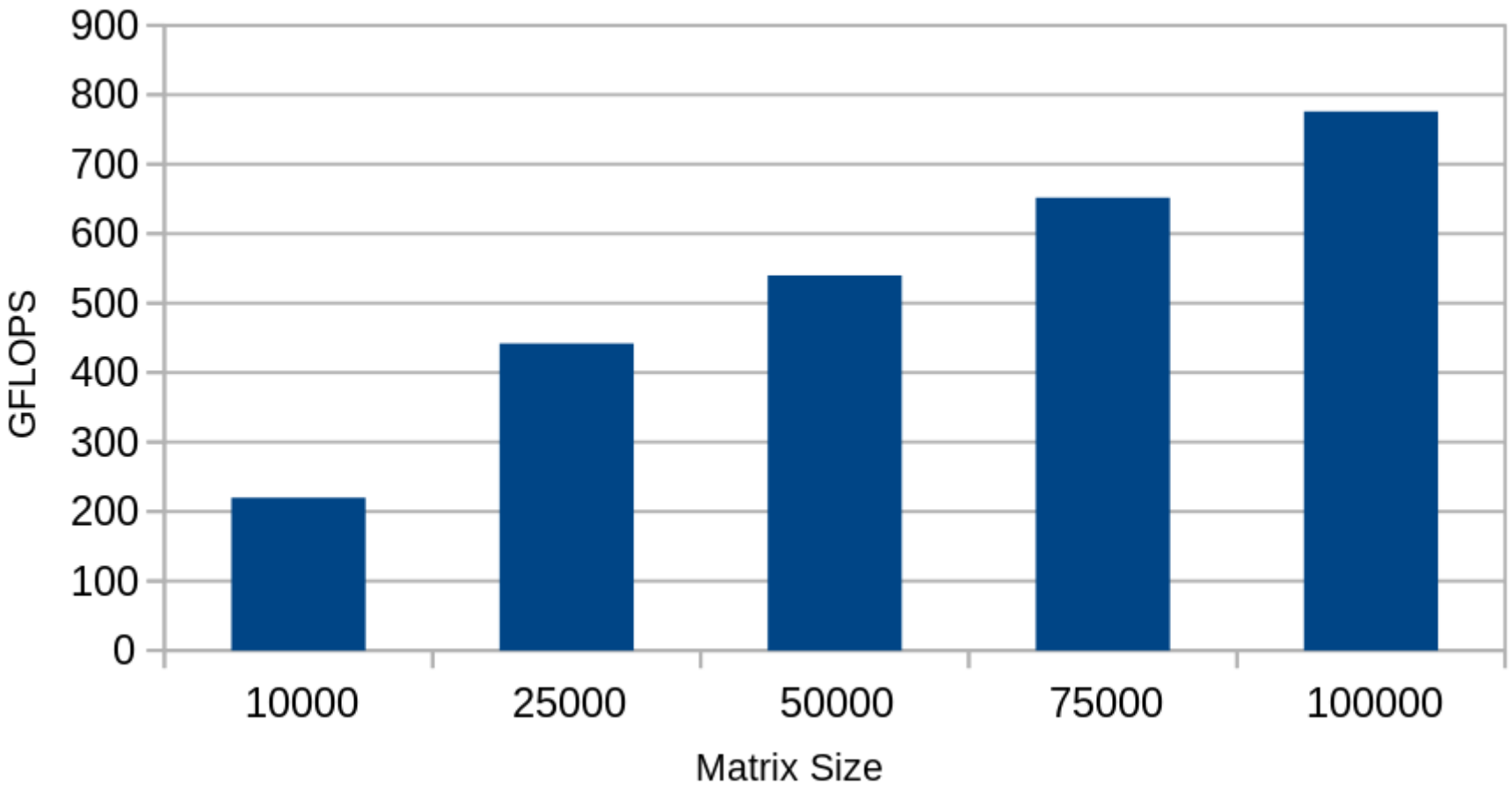
Platform Specifications

Access to clusters

- Virtual machines are used to access clusters.
- It's easy for users.
- Users can store experiment results safely and have uniform access to them.
- Users can freely customize their environments.
- Organizations have less security issues.
- Virtual machine can be used as a computational resource, even hybrid (VMs can be migrated to the powerful HPC hosts).

Basic Tests: LINPACK

- De facto standard for HPC systems benchmarking.
- LINPACK benchmark: solve a system of linear equations.
- It can demonstrate problems of a hybrid cluster.



Basic Tests: LINPACK

- One should choose good test parameters in order to get good performance.
- The main parameter is a matrix size.
- This matrix should be allocated in RAM only (swap memory should not be used).
- Data transfers can slow the computations down.
- In order to get good performance with LINPACK, one should increase the amount of memory, upgrade network, think about communication within a node (CPU-GPU).

```
================================================================================
T/V                N      NB     P     Q               Time                 Gflops
--------------------------------------------------------------------------------
WR10L2L2      100000   512     1     3             859.08              7.760e+02
--------------------------------------------------------------------------------
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=0.0031244 .... PASSED
================================================================================
```

Test results for 1 node (3 GPUs).

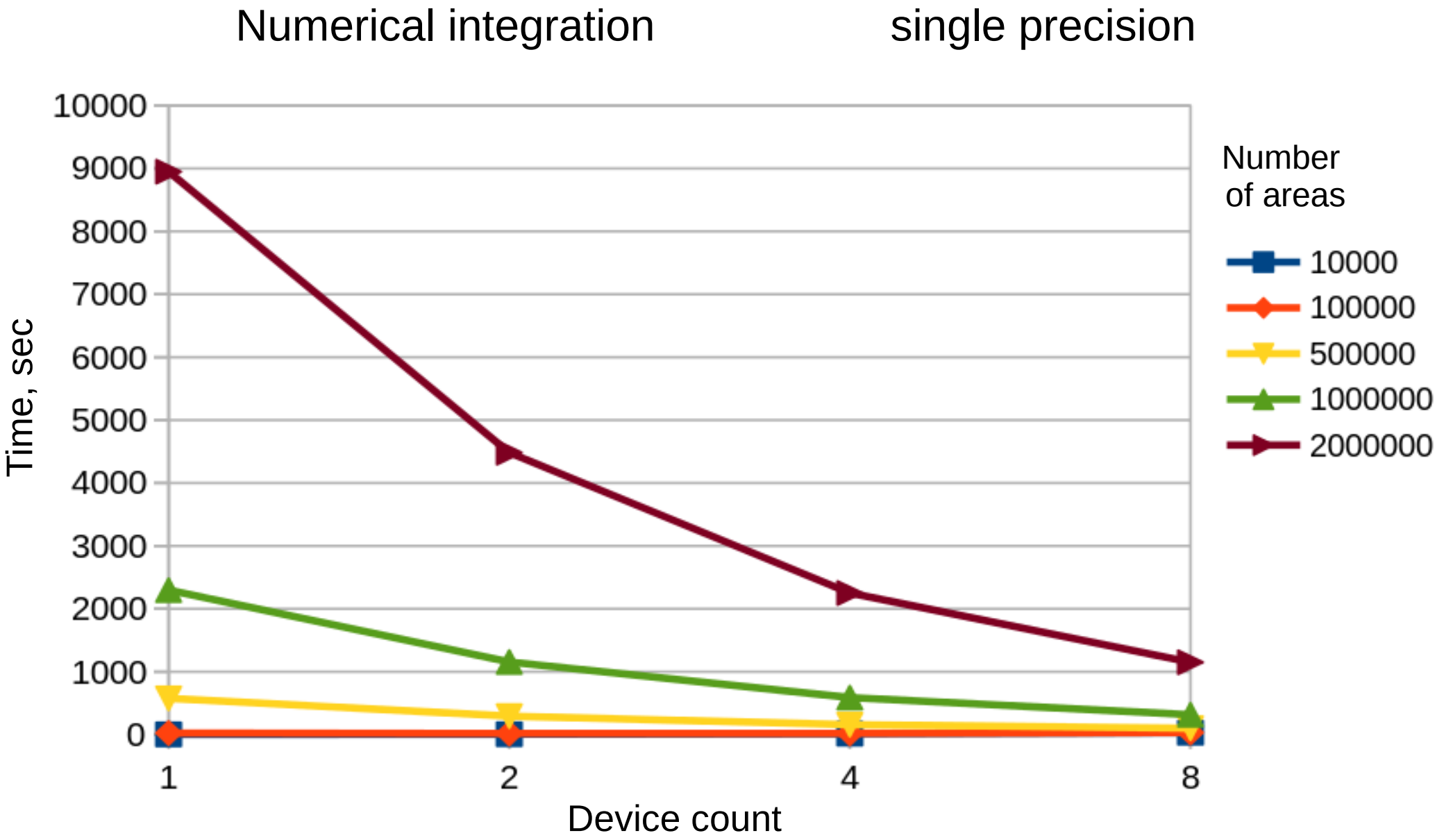
Basic Tests

Are hybrid clusters suitable for any task?

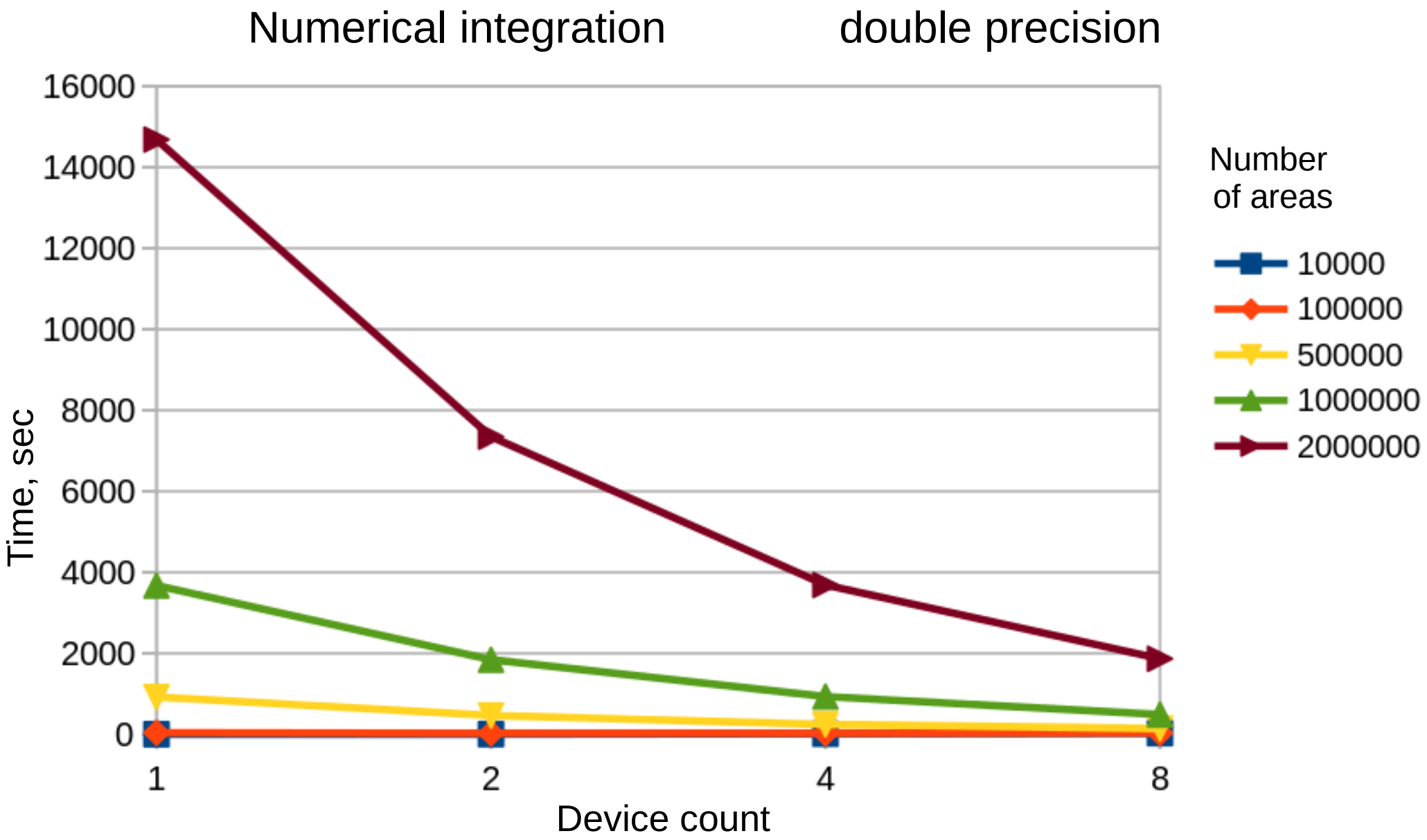
You're just too good to be true ...

- Heterogeneous calculations are an efficient way for solving computationally intensive problems.
- But not every task can be smoothly mapped onto a hybrid system.
- One can achieve substantial speedup on big tasks.
- Data transfer should not be very intensive.
- Coarse grained algorimths is a good choice.

Basic Tests


Numerical integration
single precision
Time, sec
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1
2
4
8
Device count
Number of areas
10000
100000
500000
1000000
2000000

Basic Tests


Numerical integration
double precision
16000
14000
12000
10000
8000
6000
4000
2000
0
Time, sec
1
2
4
8
Device count
Number of areas
10000
100000
500000
1000000
2000000

Applications

There are many applications that support **GPGPU**.
The list of such **applications** is constantly growing:

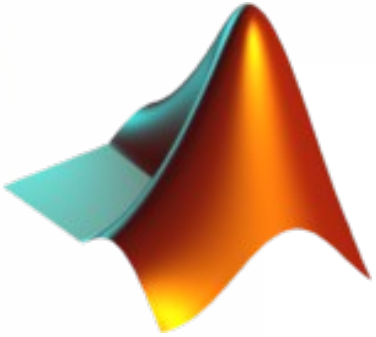
- Abinit
- ANSYS
- GROMACS
- MATLAB
- OpenFOAM (ofgpu)
- QuantumEspresso
- …

- Compilers: NVCC, PGI ...

There are many **commercial** GPGPU products as well as free and **open source**.

Applications: Mathematics

MALTAB is a large commercial software package for computations.

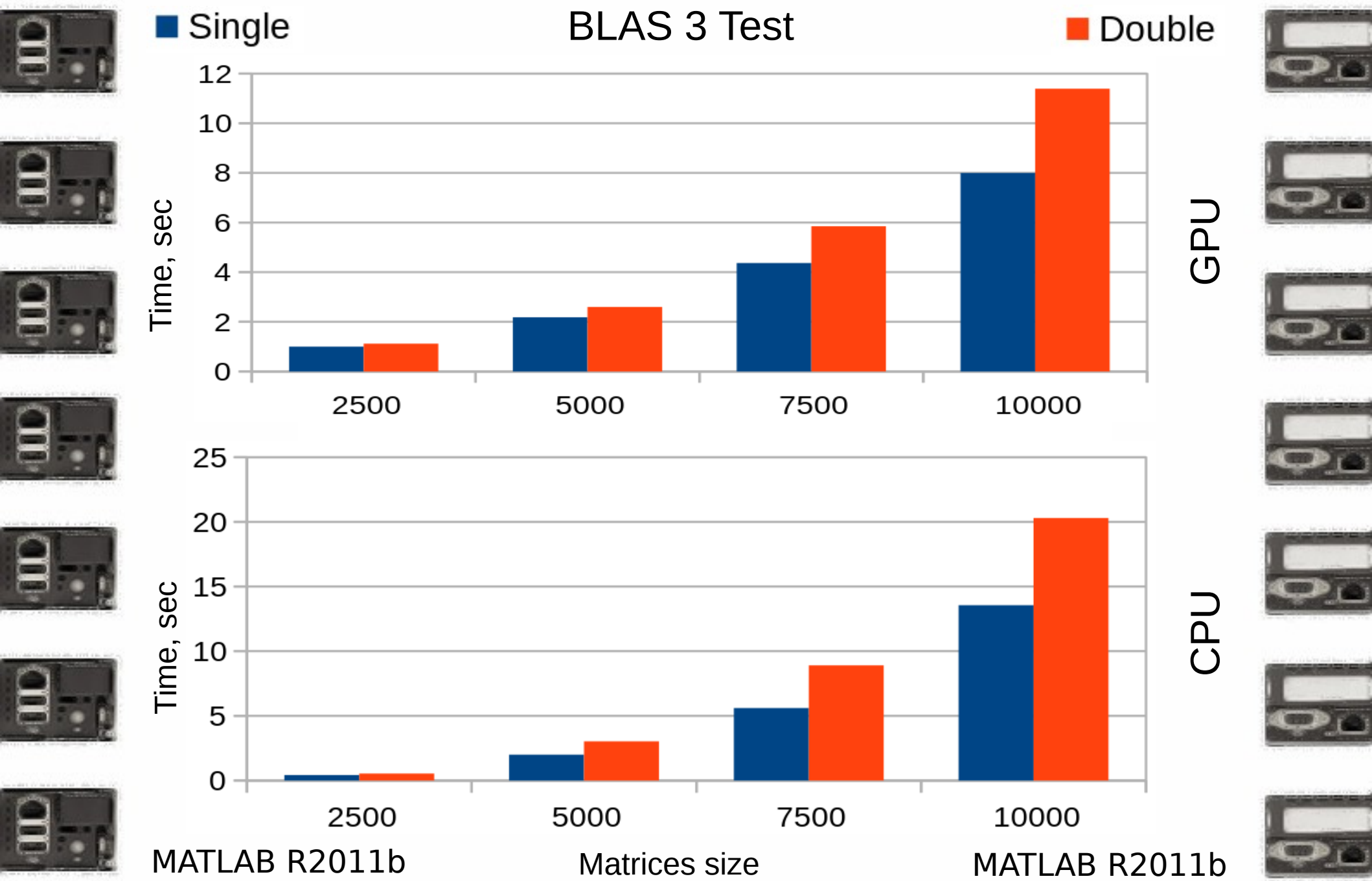


- Only NVIDIA CUDA GPUs are supported;
- It provides users with GPU-enabled functions;
- It allows CUDA kernel integration in applications.

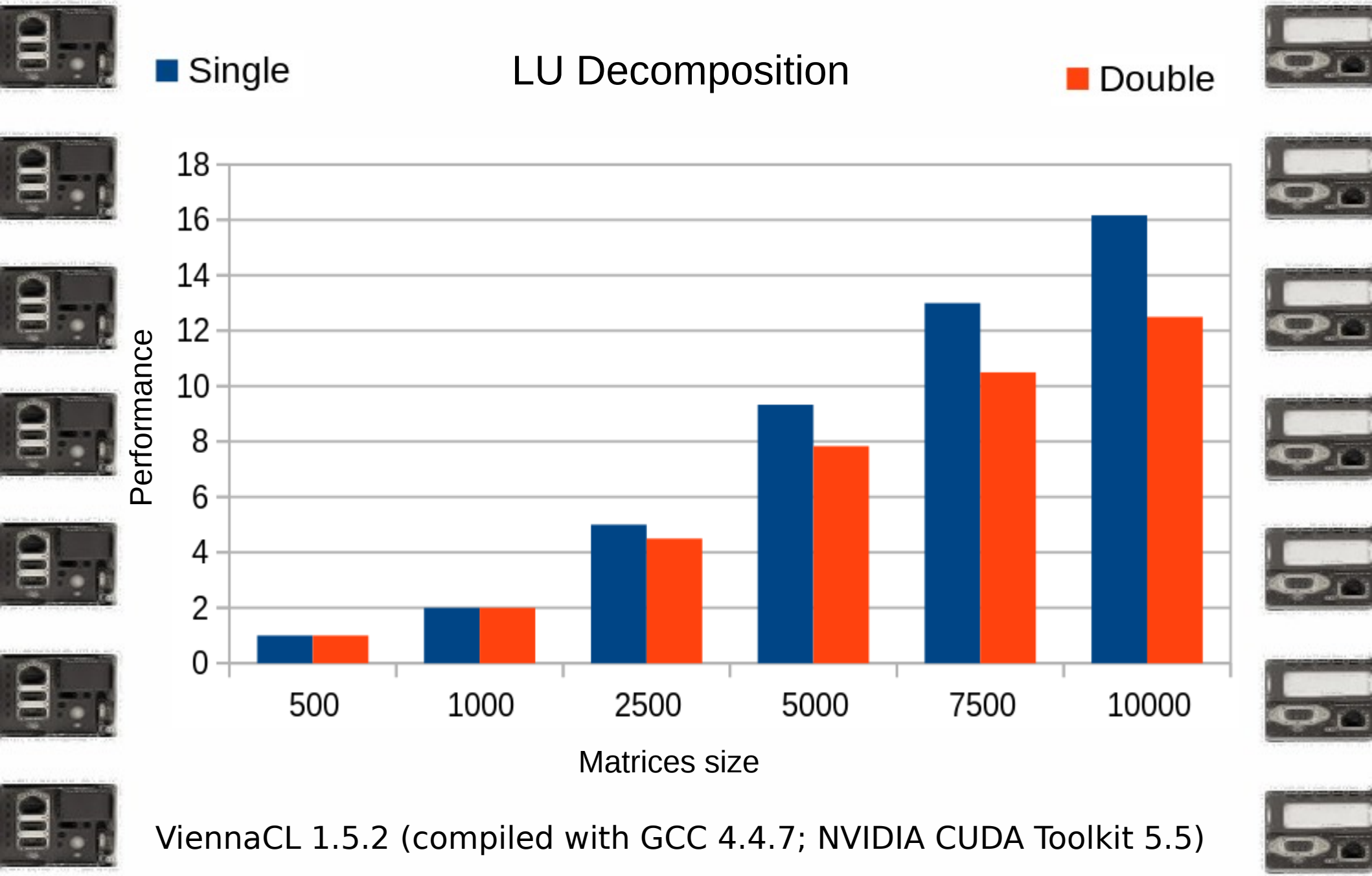
ViennaCL is an open source linear algebra library.

- Several accelerator architectures are supported (GPUs, MIC);
- Written in C++ (CUDA, OpenCL, OpenMP variants).

Applications: MATLAB



Applications: ViennaCL



ViennaCL 1.5.2 (compiled with GCC 4.4.7; NVIDIA CUDA Toolkit 5.5)

Applications: CFD

OpenFOAM is an open source platform for solving CFD problems.

There are several libraries for running OpenFOAM on GPU.

Ofgpu is an open source library for GPU computations.

- It provides users with GPU linear system solvers.
- It uses CUSP to work with matrices.

SpeedIT is a commercial library for GPGPU.

- It contains accelerated linear system solvers.
- It has a free version, but that version has restrictions.

Applications: OpenFOAM

OFGPU

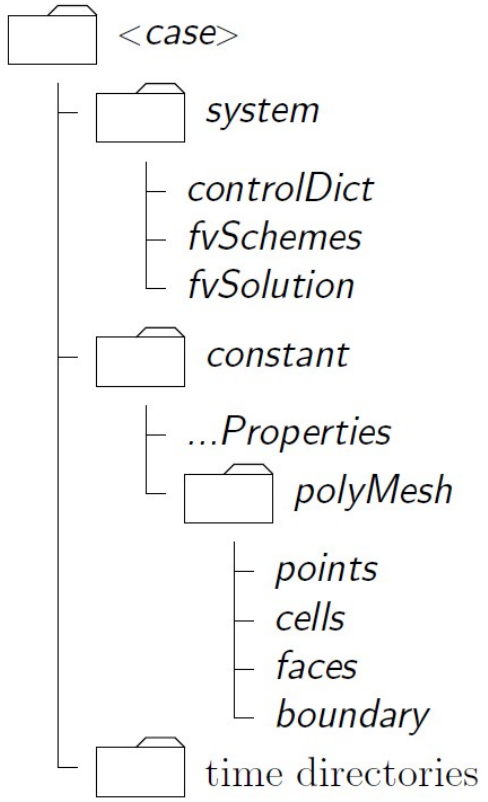
OFGPU provides users with 2 linear system solvers:

- **PCGgpu** - preconditioned conjugate gradient solver for symmetric matrices for GPGPU;
- **PBiCGgpu** - preconditioned biconjugate gradient solver for asymmetric matrices for GPGPU.

These solvers should be specified in the file:

<case>/sysem/fvSolution

```
...
U
{
    solver          PBiCGgpu;
    preconditioner  diagonal;
    ...
}
...
```

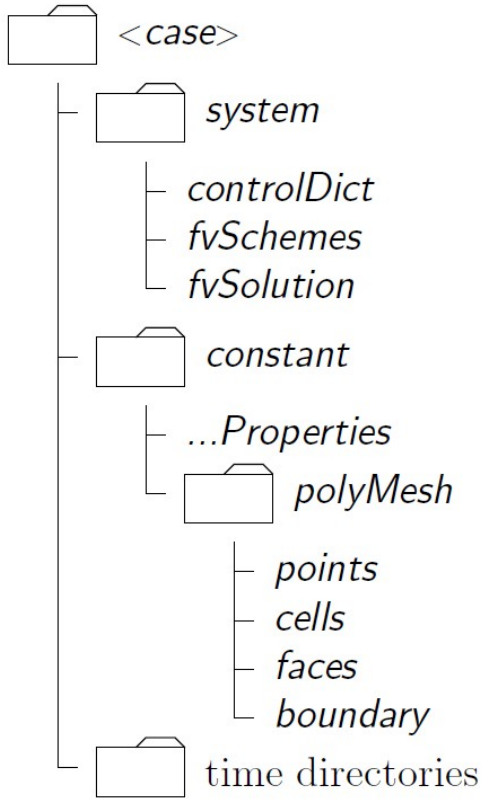


Applications: OpenFOAM

OFGPU is built as a separate library and it should be loaded when OF solver use **gpu* linear system solvers.

Additional OFGPU settings can be specified in the file:
<case>/sysem/controlDict

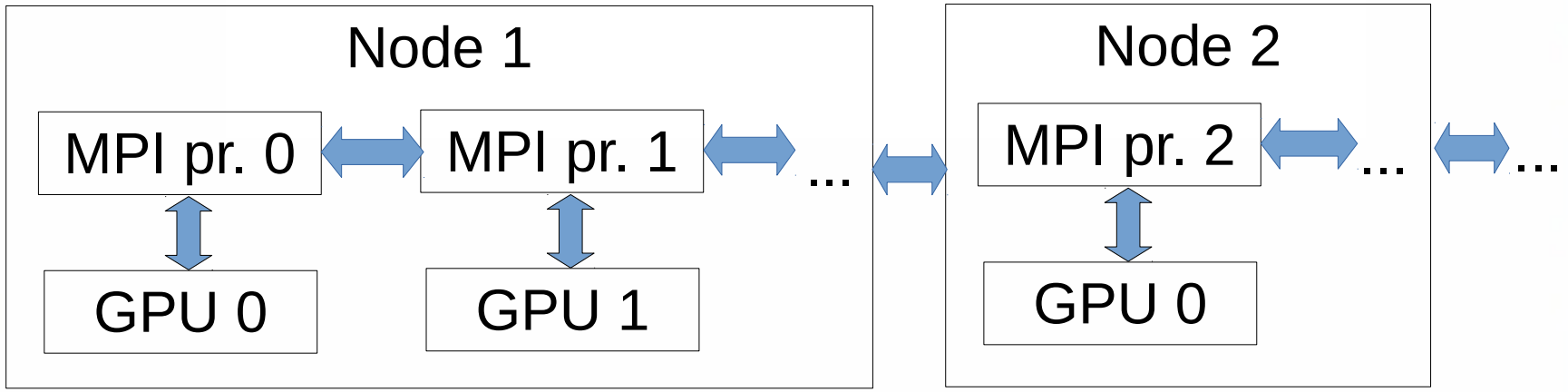
```
functions
{
  cudaGpu
  {
    type cudaGpu;
    functionObjectLibs ("gpu");
    cudaDevice 0;
  }
}
```



User can specify GPU device to use ("cudaDevice" parameter).

Applications: OpenFOAM

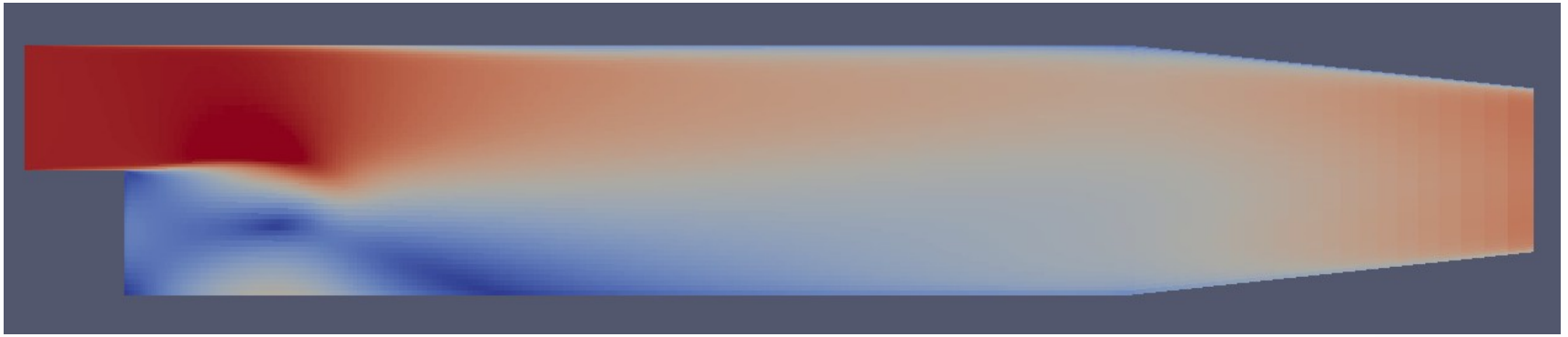
- OFGPU allows users to run OF computations only on one GPU. Multi-GPU systems are not supported.
- We rewrite OFGPU code to harness multiple GPUs on different nodes on a hybrid cluster (work in progress).
- We use OF approach for parallelizing tasks — domain decomposition: initial mesh is decomposed between N MPI processes (solvers).
- Available GPUs are distributed between MPI processes (each process works with 1 GPU). Number of GPUs that will be used is equal to the number of MPI processes.
- Each process initiates GPGPU calculation ons its own GPU for solving system of linear equations.



Applications: OpenFOAM

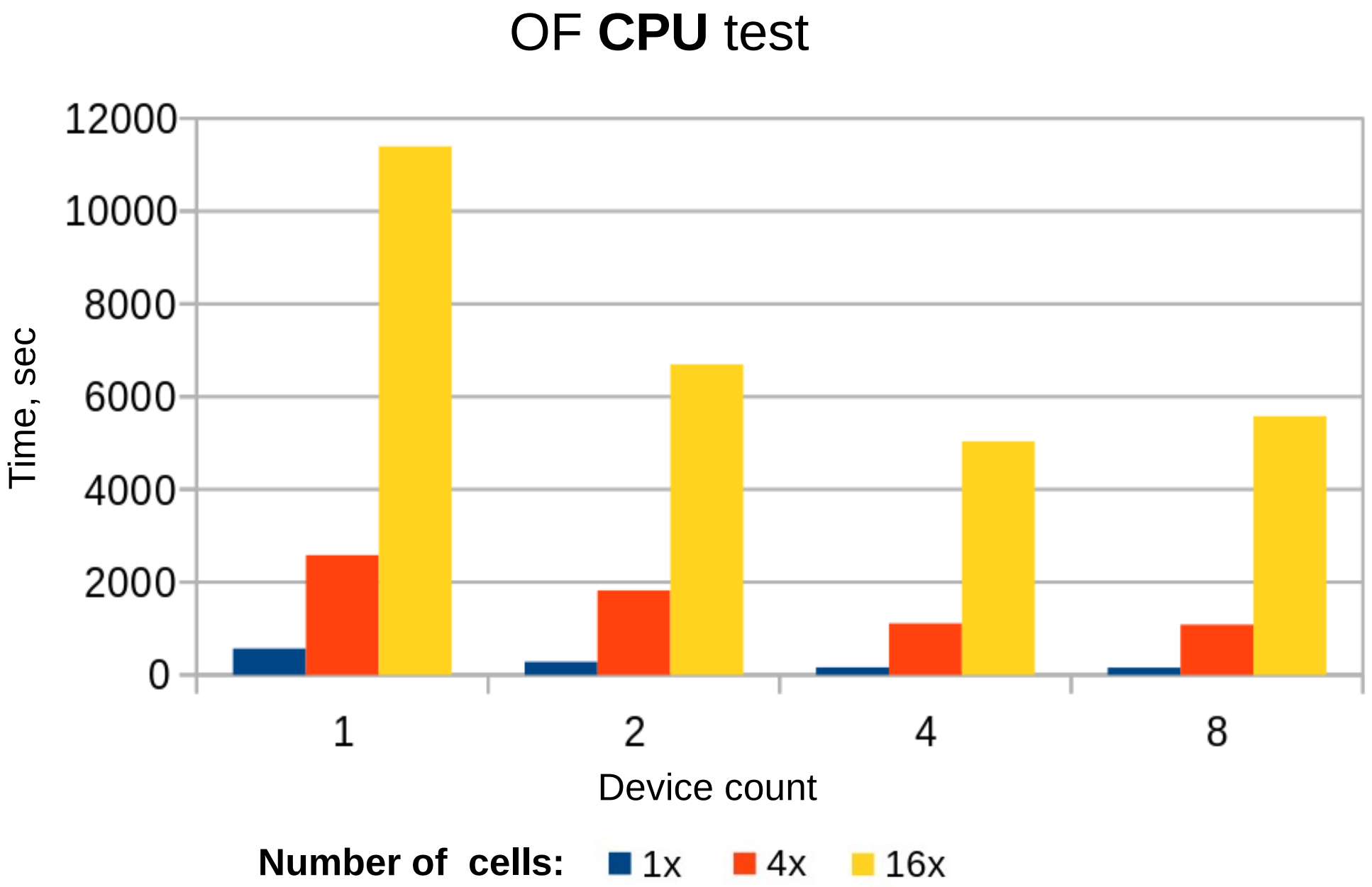
Tests

- Modified test from OpenFOAM package.
- Steady turbulent flow over a backward-facing step.
- It was run with different number of cells in the mesh.
- Solver: simpleFoam.
- Linear system solvers: PCG and PBiCG for CPU, PCGgpu and PBiCGgpu for GPU.

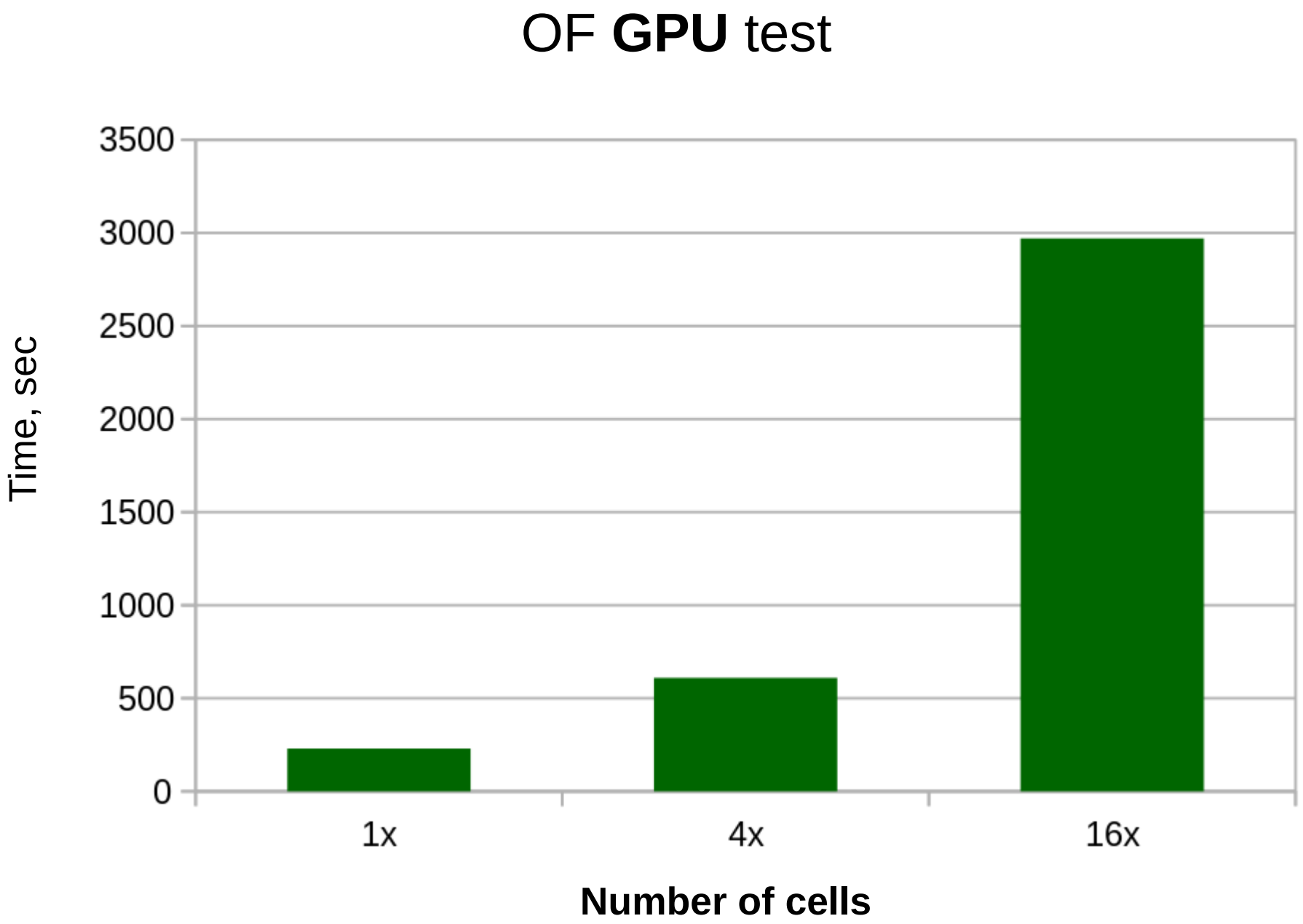


- OpenFOAM 2.2.2 with OFGPU 1.1 .
- Compiled using Intel compilers (Intel Cluster Studio 2013; ICC 13.0.1).
- MPI: IMPI 4.1.0.024 from Intel Cluster Studio 2013.
- CUDA Toolkit 5.5 .

Applications: OpenFOAM



Applications: OpenFOAM

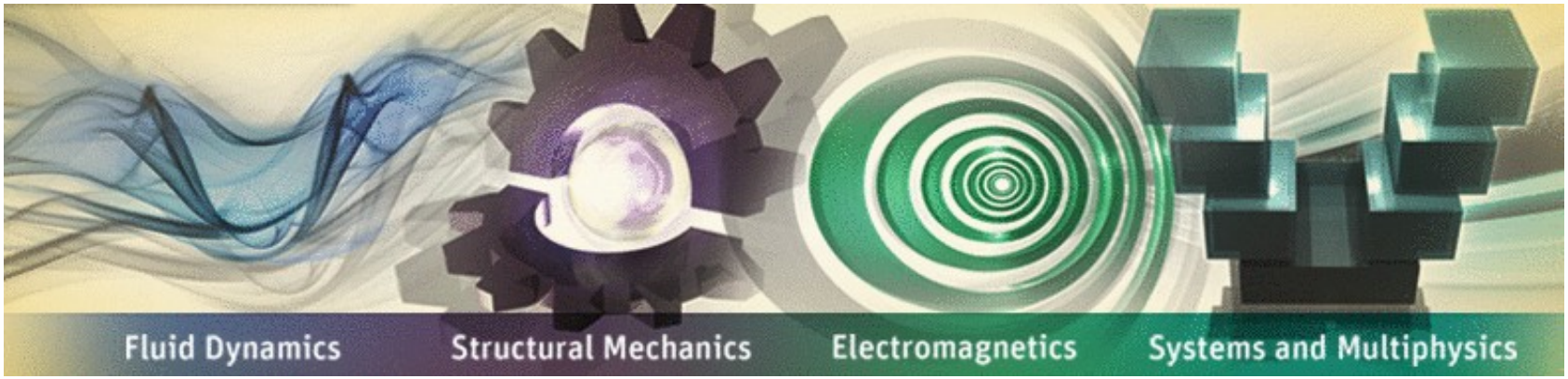


Applications: ANSYS

ANSYS is a simulaion software solution for computer-aided engineering.

ANSYS Fluent is a software package that is used to model flow, turbulence, heat transfer, and reactions.

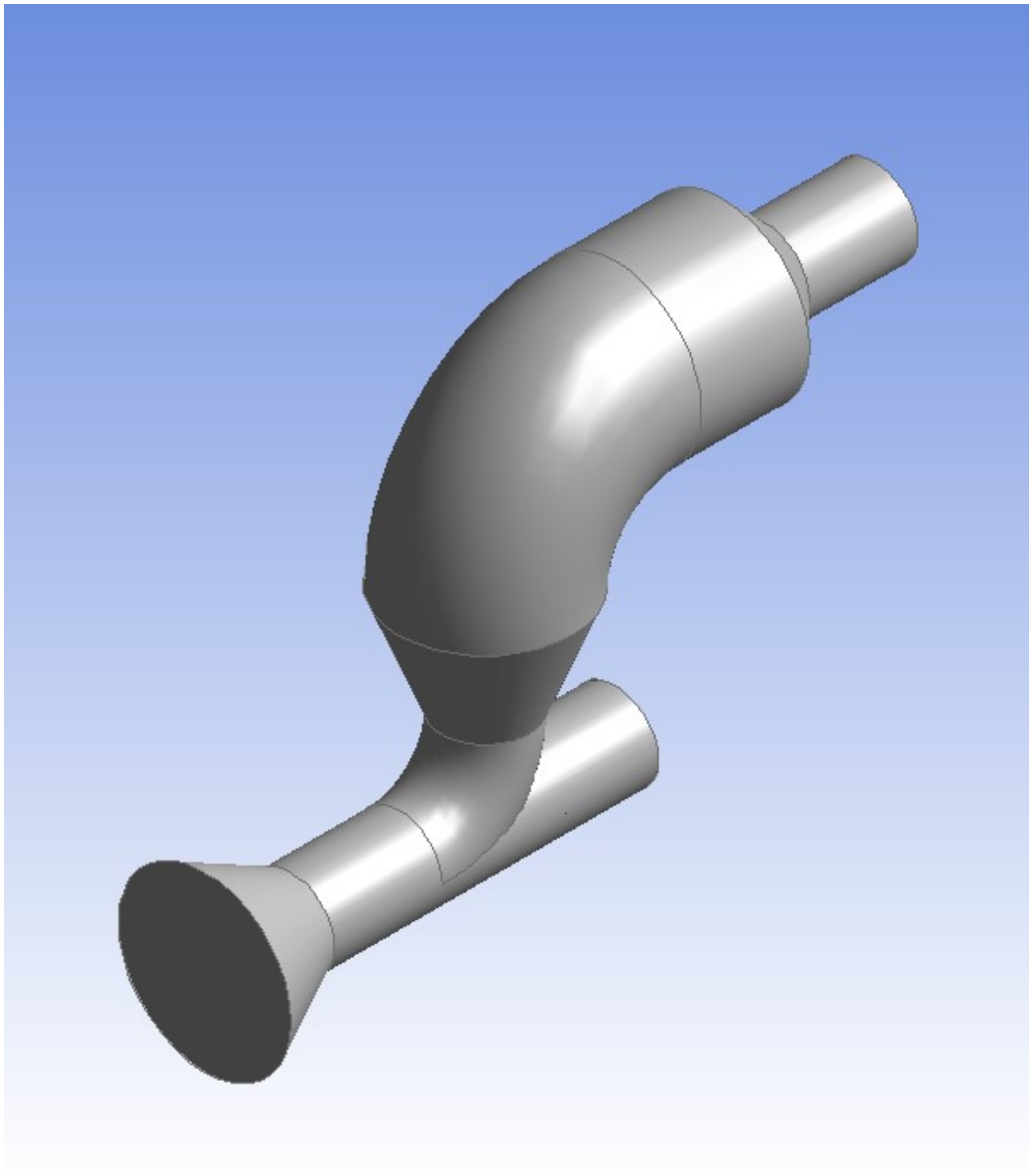
- Fluent offloads copmutationally intensive tasks to GPUs.
- Multi-GPU systems are supported in the latest version (15).



Applications: ANSYS

Flow in a tube case (small case)

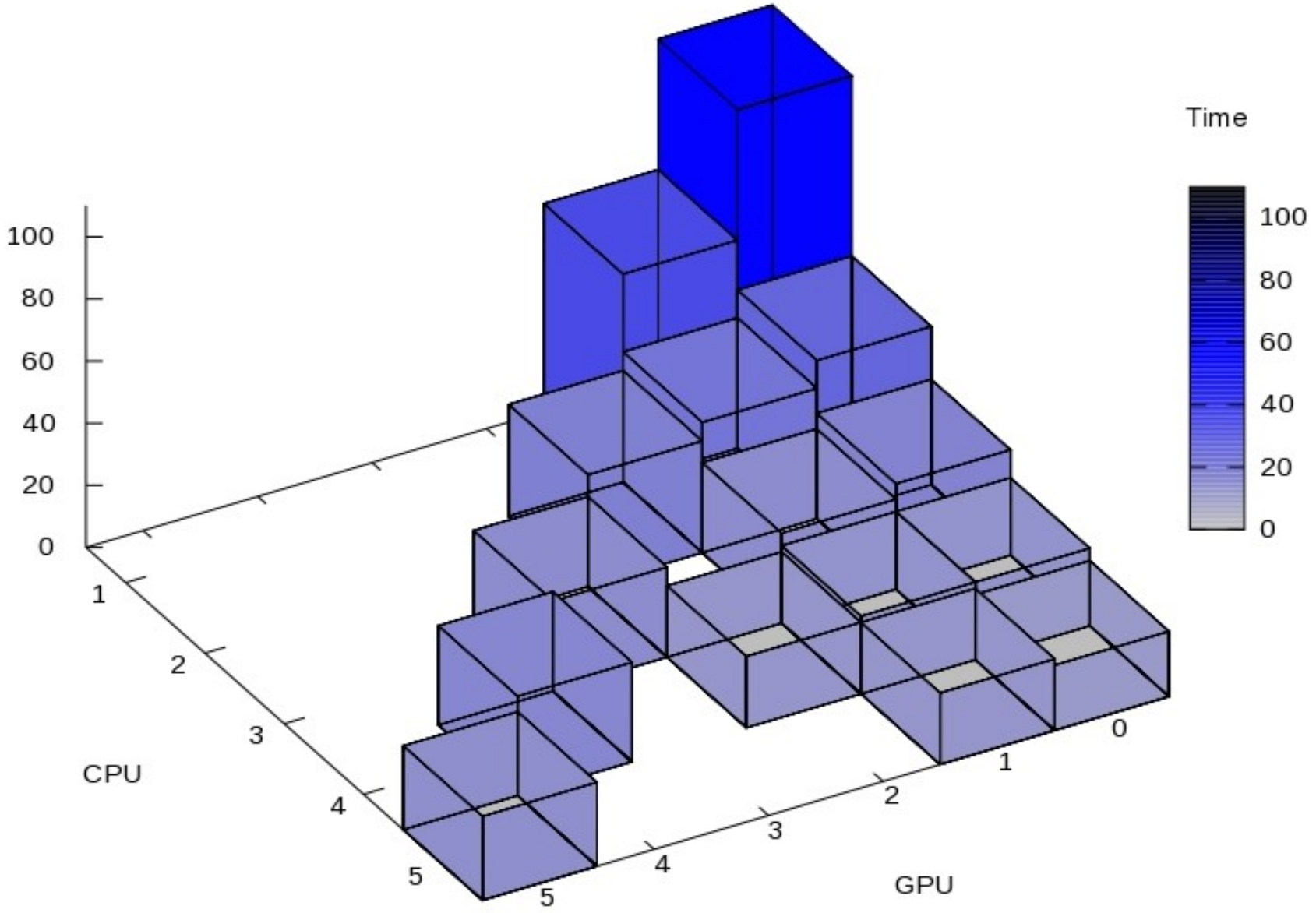
ANSYS 15



ANSYS 15

Applications: ANSYS

Flow in a tube case (small case)



Conclusions

- Hybrid clusters offer great peak performance.
- Real performance depends on task.
- It is advisable to run big tasks on such systems.
- Good choice is coarse grained algorithms.
- Data transfer can be a bottleneck (data transfer between nodes via network; data transfer between CPU and GPU).
- In case of MPI apps choose MPI implementation that works fine with the fastest interconnect available on the system (e.g. implementation that works with InfiniBand better than others).
- Cluster management system can ease maintenance of such complexes (e.g. PBS with good scheduler).
- When writing an application memory size should be taken into account.

The Last Slide

Thank you!

Questions?

Acknowledgements:

the research was carried out using computational resources of Resource Center "Computational Center of Saint-Petersburg State University"