Acceleration of Heat Diffusion Simulation on Single-Node Systems

David Bayer

Faculty of Information Technology, Brno University of Technology, Centre of Excellence IT4Innovations, CZ

BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Overview

k-Wave is a MATLAB toolbox for simulating the propagation of acoustic waves in the time domain field. It consists of several computationally intensive simulations, one of which is a heat diffusion simulation. To reduce the overall computation time and cost, an accelerated implementation targeting single-node systems was created supporting CPU and single or multiple GPUs computation.

What is computed?

The simulation computes a first order differential equation based on *Pennes' bioheat equation* which describes heat transfer in living tissues. The equation is solved in the time domain using the *k-space* pseudospectral method for spatial gradient calculation. The equation is defined as

$$\frac{\partial T}{\partial t} = \frac{\nabla^{-1}(k \cdot \nabla T) - \rho_b c_b \varkappa_b \cdot (T - T_{ba}) + Q}{\rho c}$$

where

T is temperature [°C],
t is time [s],
∇ is spatial gradient operator,
ρ is medium density $\left[\frac{\mathrm{kg}}{\mathrm{m}^3}\right]$,
c is medium specific heat capacity $\left[\frac{\mathrm{J}}{\mathrm{m \cdot K}}\right]$,
k is medium thermal conductivity $\left[\frac{\mathrm{W}}{\mathrm{m \cdot K}}\right]$,
ρ_b is blood density $\left[\frac{\mathrm{kg}}{\mathrm{m}^3}\right]$,
c_b is blood specific heat capacity $\left[\frac{\mathrm{J}}{\mathrm{m \cdot K}}\right]$,
$\varkappa_b$ is blood perfusion rate $\left[\frac{\mathrm{W}}{\mathrm{m \cdot K}}\right]$,
T_{ba} is blood ambient temperature [°C],
Q is volume rate of heat deposition $\left[\frac{\mathrm{W}}{\mathrm{m}^3}\right]$.

Together with the heat diffusion simulation, the *Cumulative Equivalent Minutes at* 43 °C (CEM43) integral is computed. It is a measure of the thermal dose that is used to predict the thermal damage of the tissue. It is defined as

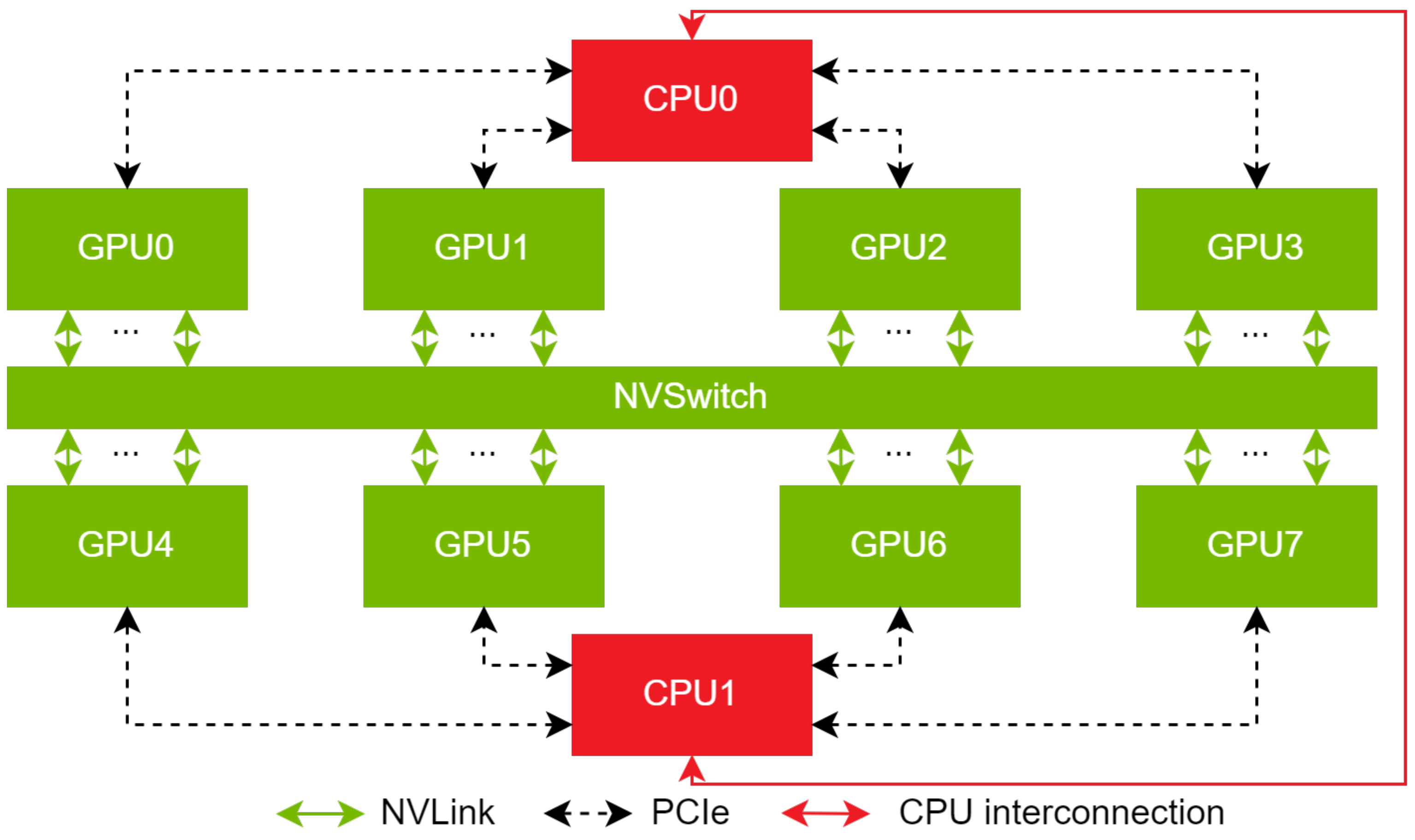
$$\mathrm{CEM43} = \int_0^t \frac{1}{60} \cdot R^{T_{43} - T} dt$$

where

$\mathrm{CEM43}$ is cumulative equivalent 43 °C,
T is temperature [°C],
T_{43} is temperature threshold 43 °C,
R is the experimentally determined factor to compensate per 1 °C temperature change.

The calculation of the spatial gradient via *k-space* pseudospectral method depends on the boundary conditions. For periodic boundary condition, the spatial gradient is computed via fast Fourier transform (FFT). In case of insulating bouldary condition, the discrete cosine (DCT) transform can be used, and for the conducting boundary condition, the discrete sine transform (DST) can be used.

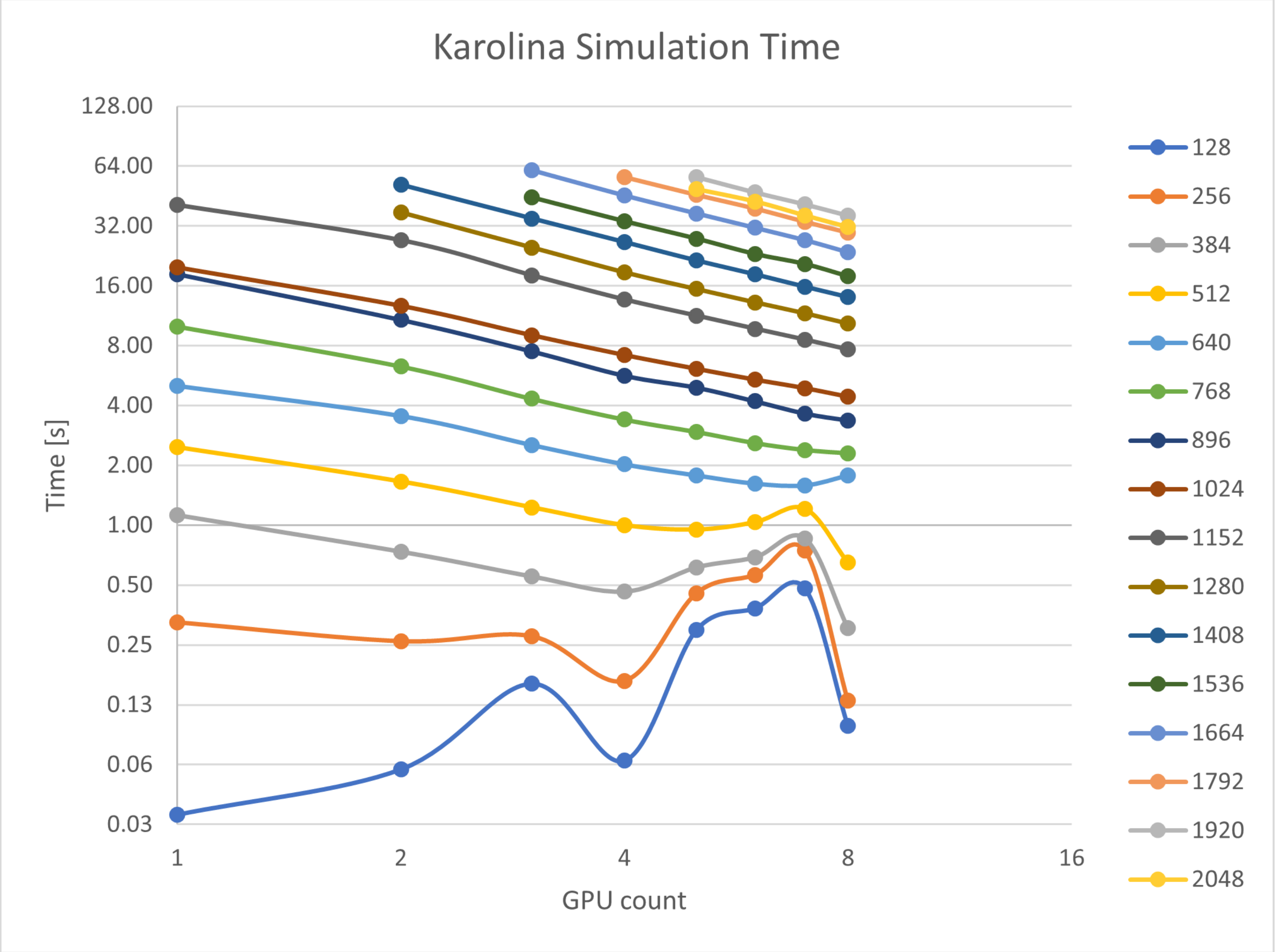
Technologies



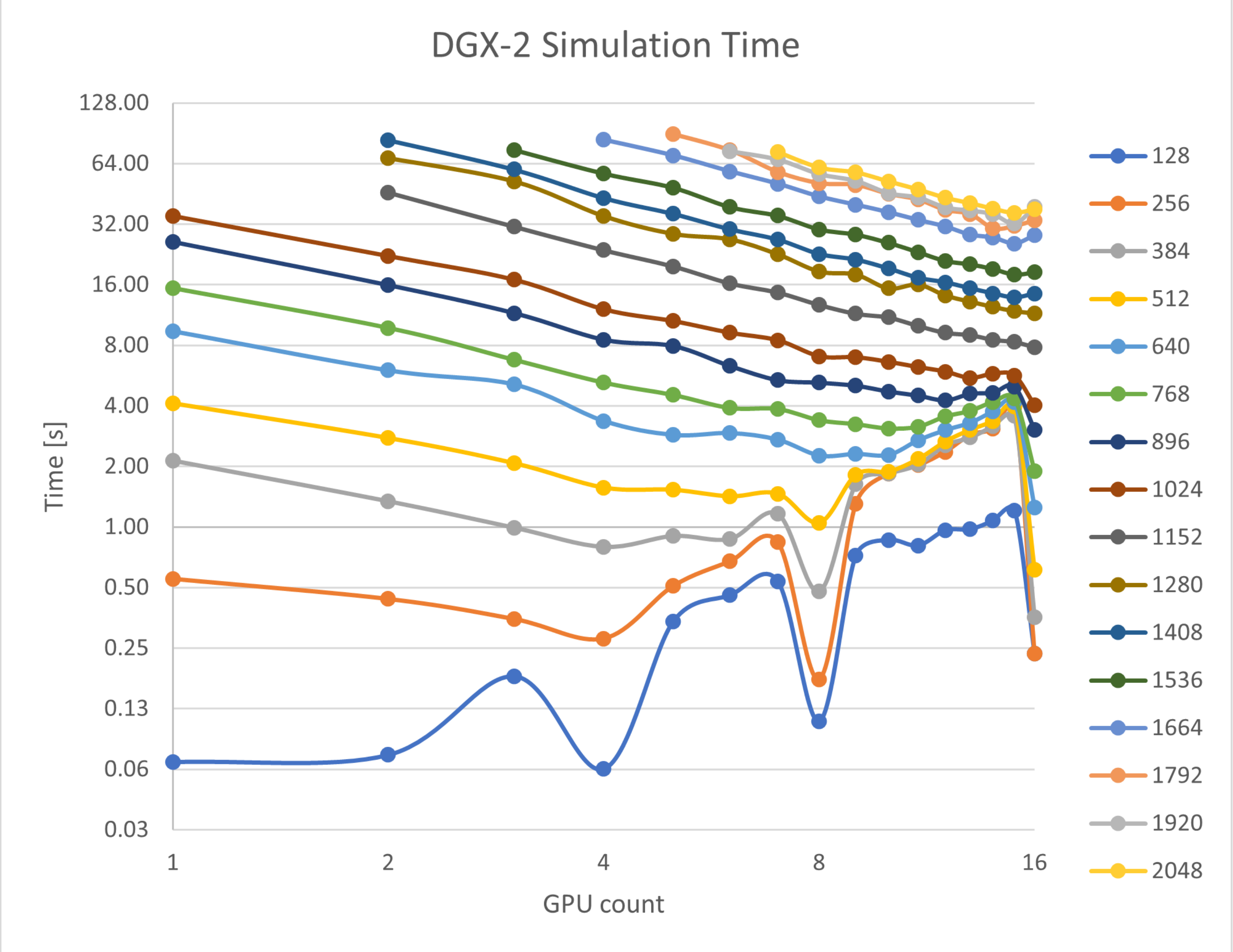
In our program, we use high level techniques for SIMD code generation and multi-threading through *OpenMP*. For FFT (and other transformations) computation we use libaries with *FFTW*-like API that support multi-threading and SIMD units utilization. Computation on GPUs is supported for NVIDIA and AMD platforms. For NVIDIA GPUs, we use *CUDA* and *cuFFT* library. For AMD GPUs, we use *HIP* and *rocFFT* library. Additionally on NVIDIA GPUs, multi-GPU job splitting is supported as well. To obtain the best GPU performance, technologies like *Graphs* and *real-time compilation* (RTC) are used.

Results

In this section we present the results of the simulation time measurements on a 3D homogenous simulation without perfusion on different multi-GPU machines for cubes of side lenghts 128 – 2048. The Karolina supercomputer's GPU accelerated nodes contain 8x NVIDIA A100 GPU, each equiped with high bandwidth 40 GiB of memory. From the results can be seen that for simulations larger than 384^3 the efficiency is around 80-85 % for 2 to 4 GPUs. In case of more than 4 GPUs, the efficiency starts slowly dropping towards 60 %. The highest acceleration 5.43 times is achieved using 8 GPUs for size 896^3.



DGX-2 is a high performance machine containing 16x NVIDIA V100 GPU with 32 GiB of memory each. These GPUs are a generation older than those on Karolina. This has a significant impact on the performance mostly because of a smaller bandwidth between devices. The best efficiency around 75 % was achieved for 2-4 GPUs on simulations larger than 384^3. The 4-8 GPUs still offer acceptable efficiency around 60-70 % and for more than 8 GPUs the efficiency drops rapidly. The best acceleration of 8.76x was achieved with all 16 GPUs and grid size of 1024^3.



Conclusion

The results show that if a user wants to spent as little money as possible, it is convinient to use a single GPU of the latest architecture available. If the data does not fit into a single GPU's memory or one wants to reduce the compute time, we would recommend to use up to 8 GPUs, in rare cases more.

VSB TUO IT4I

k-Wave
A MATLAB toolbox for the time-domain simulation of acoustic wave fields

European Innovation Council

citrus

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254). This project has received funding from the European Unions Horizon Europe research and innovation programme under grant agreement No 101071008.