SYCL vs CUDA vs HIP: Evaluasi Performa dan Portabilitas Berbagai Model Pemrograman GPU pada GPU NVIDIA dan AMD = SYCL vs CUDA vs HIP: Evaluating the Performance and Portability of Different GPU Programming Models on NVIDIA and AMD GPUs

Valerian Salim, author

Abstrak

Dalam HPC, pemanfaatan GPU untuk kapabilitas pemrosesan paralelnya dapat mempercepat komputasi secara masif, terutama untuk masalah yang embarrassingly parallel. Namun, saat memilih model pemrograman GPU, portabilitas model pemrograman dan sistem vendor GPU harus dipertimbangkan. Untuk memahaminya dengan lebih baik, makalah ini menganalisis waktu eksekusi baseline CUDA, HIP, dan SYCL pada GPU NVIDIA dan AMD. Program UVaFTLE, sebuah program yang digunakan untuk menentukan Lagrangian Coherent Structures melalui ekstraksi Finite-Time Lagrangian Exponents (FTLE), digunakan untuk mengukur waktu eksekusi. Eksperimen ini menunjukkan kinerja CUDA dan SYCL pada kedua platform GPU, yang secara konsisten mengalahkan HIP dalam waktu eksekusi. Upaya untuk mengoptimalkan waktu eksekusi fungsi kernel GPU di seluruh platform juga dilakukan, secara drastis memangkas waktu eksekusi kernel preproc hingga lebih dari 90%. Setelah optimasi, SYCL tetap menjadi yang terbaik, sementara CUDA berada di posisi kedua, dan HIP yang terlihat jelas paling lambat. Makalah ini juga membahas tantangan development yang dihadapi. CUDA dan SYCL membanggakan dokumentasi dan dukungan komunitas yang sangat baik, sementara dokumentasi HIP tertinggal dan tidak memberikan pengalaman development yang positif seperti kedua model lainnya.

.....In HPC, leveraging GPUs for their parallel processing capabilities can massively accelerate computation, especially for embarrassingly parallel problems. However, when choosing a GPU programming model, one must take into consideration the portability of the programming model and the GPU vendor of their system. To understand them better, this paper analyzes the baseline execution time of CUDA, HIP, and SYCL on both NVIDIA and AMD GPUs. The UVaFTLE program, a program used to determine Lagrangian Coherent Structures through the extraction of Finite-Time Lagrangian Exponents (FTLE), is used to benchmark execution time. The experiment showcases the performance of CUDA and SYCL on both GPU platforms, which consistently beat HIP in execution time. An effort to optimize the execution time of the GPU kernel functions across is made, drastically cutting the execution time of the preproc kernel by over 90%. After the optimizations, SYCL remains as the champion, while CUDA comes second, and HIP is clearly the slowest. This paper also discusses the development challenges encountered. CUDA and SYCL boast excellent documentation and community support, while HIP’s documentation falls behind and does not provide a developer experience as positive as the other two.