Реализация схемы переноса пассивной примеси на графических ускорителях при использовании половинной точности

Гащук Е.М.[1, 2], Дебольский А.В.[3], Мортиков Е.В.[3,2]

[1] Мехмат МГУ им. М.В. Ломоносова, Москва.

[2]Институт вычислительной математики им. Г.И. Марчука РАН, Москва.

[3]НИВЦ МГУ им. М.В. Ломоносова, Москва.

Зачем?

Задачи прогноза погоды и климата, вычислительной гидродинамики требуют использования суперкомпьютеров

Современные суперкомпьютеры потребляют огромное количество электроэнергии

Необходимо разрабатывать вычислительно-эффективные алгоритмы, учитывающие архитектуру суперкомпьютеров

Как?

1.перенос вычислений на графические ускорители в разы сокращает время выполнения вычислений

2.Понижение точности вычислений вплоть до половинной (поддерживается аппаратно на современных GPU)

Моделирование турбулентных течений

Прямое численное моделирование (DNS) – эффективный подход детального изучения турбулентности

Система уравнений Навье-Стокса для вязкой несжимаемой жидкости:

$$\frac{\partial u_i}{\partial t} + \frac{\partial u_i u_j}{\partial x_j} = -\frac{1}{\rho_0}\frac{\partial p}{\partial x_i} + \frac{1}{Re}\frac{\partial^2 u_i}{\partial x_j \partial x_j},$$

$$\frac{\partial u_i}{\partial x_i} = 0,$$

$$\frac{\partial C_k}{\partial t} + \frac{\partial u_i C_k}{\partial x_i} = \frac{1}{ReSc_k}\frac{\partial^2 C_k}{\partial x_i \partial x_i} - T_k^{-1} C_k.$$

Явная схема интегрирования уравнения переноса:

Схемы 2-ого и 4-ого порядка точности

$$\mathrm{ADV}^n = -\Delta t \left[\frac{\partial u_i C_k}{\partial x_i}\right]_h^n,$$

$$\mathrm{DIFF}^n = \Delta t \left[\frac{1}{ReSc_k}\frac{\partial^2 C_k}{\partial x_i \partial x_i},\right]_h^n,$$

$$\mathrm{RHS}^n = \frac{3}{2}(\mathrm{ADV}+\mathrm{DIFF})^n - \frac{1}{2}(\mathrm{ADV}+\mathrm{DIFF})^{n-1} - \left[\Delta t T_k^{-1} C_k\right]_h^n,$$

$$C_k^{n+1} = C_k^n + \mathrm{RHS}^n.$$

Все данные и вся арифметика в fp16

Алгоритм компенсационного суммирования Кэхэна

Результаты

Ускорение вычислений на GPU в 1.42 раза на сетке из 3 млн. ячеек

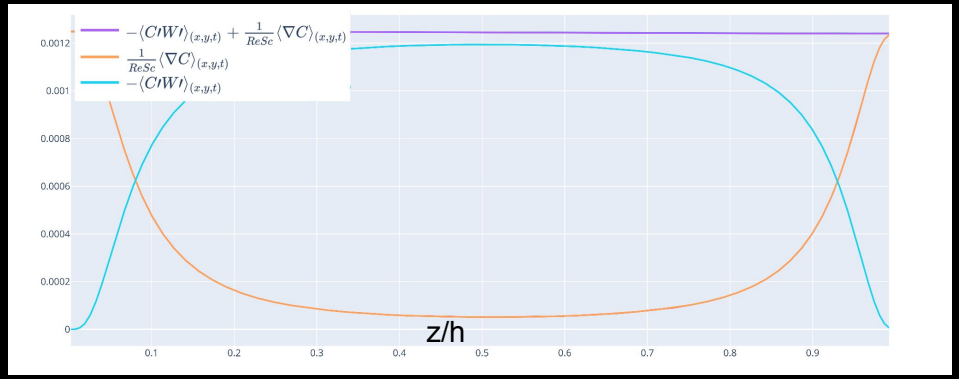

Рис 1. Полный поток скаляра и его компоненты: турбулентная и вязкая часть в fp16 и fp32 (визуально совпадают).

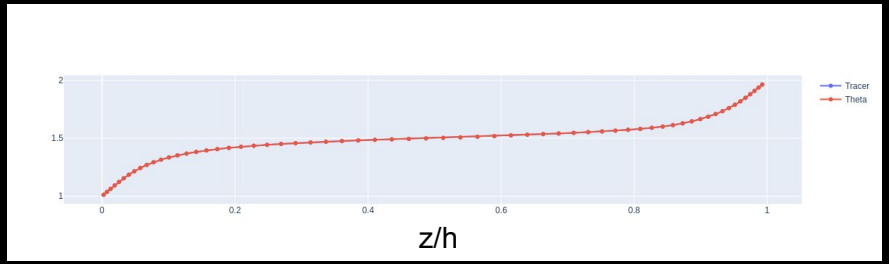

Рис. 3 Средний профиль концентрации примеси в fp16 и в fp32(визуально совпадают: норма ошибки 1.45×10^{-4}).

$$\frac{\|res_{fp32} - res_{fp16}\|_{L_2}}{\|res_{fp32}\|_{L_2}} \cdot 100\%, \quad \|\mathbf{x} = (x_1, \ldots, x_n)^T\|_{L_2} = \sqrt{x_1^2 + \ldots + x_n^2}$$

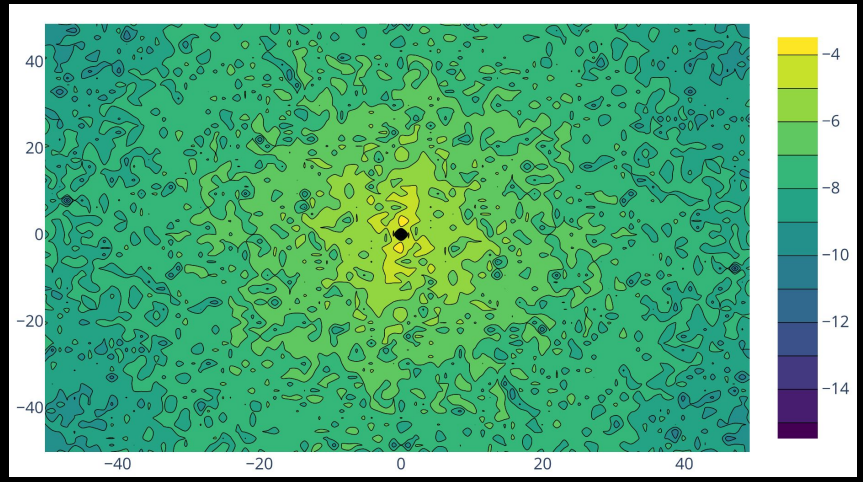

Рис. 2 Спектральная плотность дисперсии примеси в плоскости Oxy на высоте канала 0.3H в fp16.

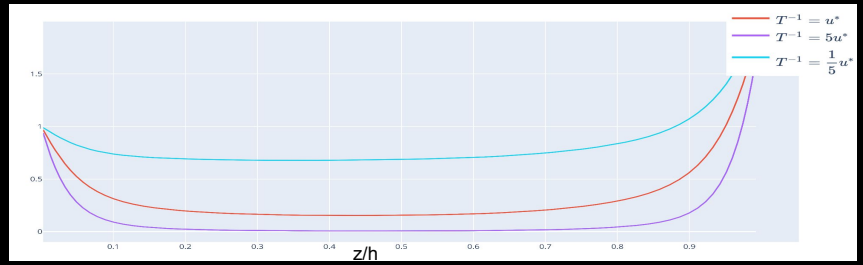

Рис. 4 Средний профиль примеси в fp16 и в fp32 (визуально совпадают: норма ошибки для каждого не превышает 10^{-4}) при различном “времени” жизни.

Заключение

Основные результаты:

- полная реализация блока переноса примеси в fp16 с использованием алгоритма Кэхэна дает достаточно точные численные результаты
- ускорение вычислений на GPU до 1.42 раз
- уменьшение используемой памяти в 1,5 раза в сравнении с fp32 с учетом дополнительного массива для хранения ошибки округления в алгоритме Кэхэна
- Ускорение исполнения MPI-обменов засчет уменьшения объема передаваемых данных

Спасибо за внимание!