Sciara-fv2 on CUDA

Massively Parallel Programming on GPU

Christian Bruni Francesco Tieri

December 13, 2025

Indice

Implementations Overview

Implementations

We introduced 5 different implementations of the Sciara-fv2 model on CUDA:

- **Naive Implementation**, by using global memory only.
- **Tiled implementation**, without using halo cells.
- **Tiled implementation**, with halo cells.
- **Cfame Implementation**.
- **Cfamo Implementation**, by memorizing outflows cells in shared memory.

Let's analyze them and how they perform.

Naive Implementation 1

Naive Implementation 1/2

First we take a look at the **16x16** block implementation without using shared memory.

Operational Intensity

Questo è un blocco "normale". Utile per definizioni.

- FLOP: 34
- Bytes: 208
- Operational Intensity: 0.163 FLOP/Byte

Now the 32x32 block implementation without using shared memory.

- FLOP: 34
- Bytes: 208
- Operational Intensity: 0.163 FLOP/Byte

Operational Intensity

Operation Intensity calculated for the 32x8 block dimension.

Naive Implementation 2/2

Qui la roofline

Tiled without Halo

Layout a due colonne

16x16 Block dimension without Halo
OPERATIONAL INTENSITY:

Roofline:

32x8 block dimension without Halo

- Punto A
- Punto B
- Punto C

Tiled with Halo

Tiled with Halo

16x16 Block dimension with Halo
OPERATIONAL INTENSITY:

Roofline:

32x8 block dimension with Halo

- Punto A
- Punto B
- Punto C

Cfame Implementation

Cfame Implementation

16x16 Block dimension with Halo
OPERATIONAL INTENSITY:

Roofline:

32x8 block dimension with Halo

- Punto A
- Punto B
- Punto C

Cfamo Implementation

Cfamo

16x16 Block dimension with Halo
OPERATIONAL INTENSITY:

Roofline:

32x8 block dimension with Halo

- Punto A
- Punto B
- Punto C

Differences between sequential

Riepilogo

Grafico a barre

Grazie!