Parallel Reduction

Roberto Marino

May 13, 2025

1 Introduzione

La **parallel reduction** è una tecnica fondamentale della programmazione parallela e distribuita, usata per calcolare l'aggregazione (come la somma, il massimo, il minimo o il prodotto) di un array di elementi in modo efficiente. È alla base di algoritmi avanzati come il `prefix sum`, il prodotto scalare, il training distribuito etc ...

2 Definizione Formale

Sia $A = [a_0, a_1, \ldots, a_{n-1}]$ un vettore di n elementi. Vogliamo calcolare:

$$R = \bigoplus_{i=0}^{n-1} a_i$$

dove $\oplus$ è un'operazione binaria associativa, come la somma ($+$) o il massimo (max).

3 Versione Sequenziale

In modo sequenziale, calcoliamo:

$$R = (((a_0 \oplus a_1) \oplus a_2) \cdots \oplus a_{n-1})$$

con complessità temporale $O(n)$ e spaziale $O(1)$ (spazio in memoria costante).

4 Distribuzione: Approccio Ad Albero

La parallel reduction sfrutta un approccio **a riduzione binaria**, dove in ogni livello dell'albero:

- I nodi elaborano in parallelo coppie di elementi.

- Il numero di elementi da elaborare si dimezza a ogni passaggio.
- La profondità dell'albero è $\log_2 n$.

5 Caso Dispari

Quando n non è una potenza di 2, l'elemento spaiato può essere:

- propagato al livello successivo senza modificarlo,
- o unito a un risultato parziale alla fine.

6 Complessità

- **Complessità temporale**: $O(\log n)$
- **Complessità spaziale**: $O(n)$ (se implementato bene)

7 Applicazioni

- Somma/massimo/minimo
- Prodotto scalare (dot product)
- Aggregazione per training distribuito (es. FedAvg)