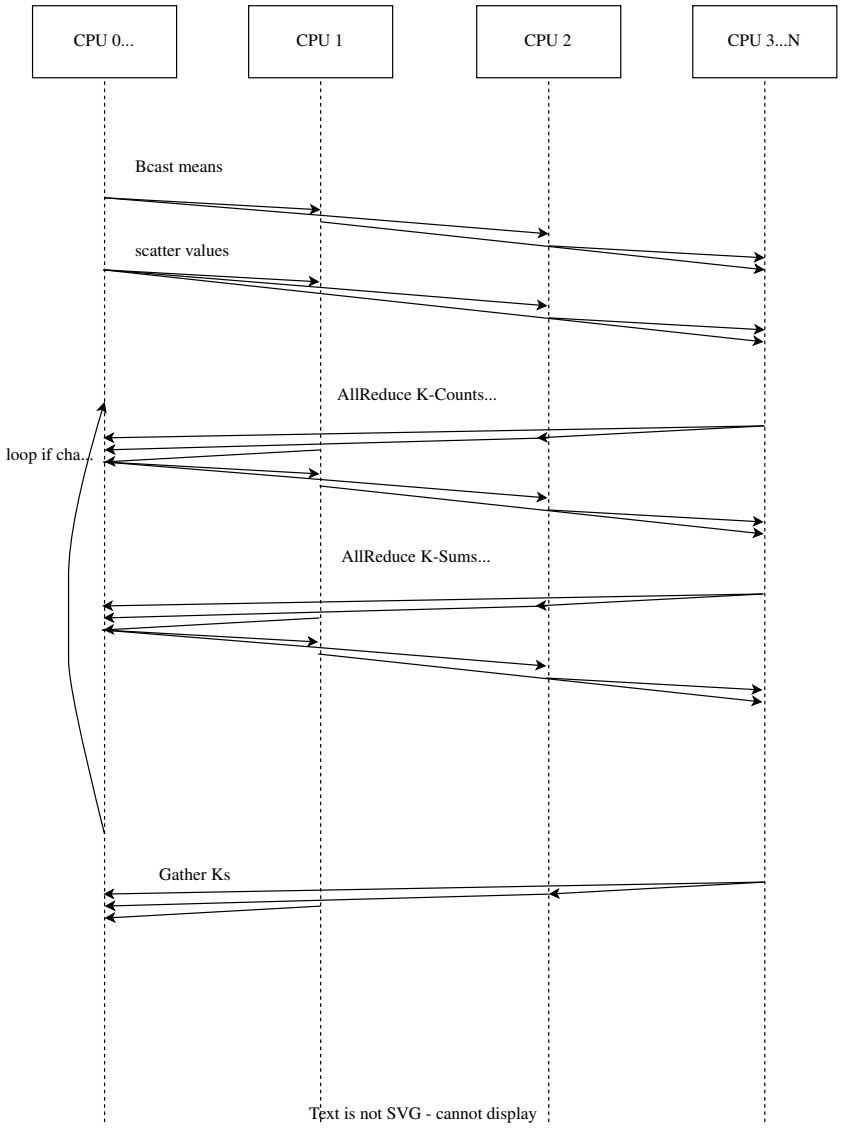
# 1 Implementace paralelního algoritmu K-means PRL

Autor: Ladislav Dokoupil




Figur 1: diagram komunikace

Program lze pro potřebu analýzy rozdělit na 3 části.:

1. načítaní vstupu
2. paralelní iterativní výpočet clusterů
3. shromážděni dat a jejich výpis

Po počátečním načtení dat a vytvoření iniciálních centroidů ROOT procesem ze souboru numbers, je každému procesoru zasláno jeho příslušné číslo a odhady centroidu. Ve smyčce poté zjišťuji nejbližší centroid do proměnné argMin. Následně si procesory vytvoří a zašlou

počet a sumu hodnot patřící ke každému centroidu. Z těchto hodnot si každý procesor vytvoří nové centroidy a pokud došlo ke změně cyklus se opakuje. Nakonec jsou ROOT procesoru zaslány hodnoty argMin, které vypíše na obrazovku.

## 1.1 Časová analýza

V první části potřebuje ROOT proces O(n) pro načtení vstupu. Následně zasílá data ostatním procesům pomocí Broadcast O(log n) a Scatter O(log n). hlavní smyčka má předem neznámý počet iterací (T), který není přímo závislý na velikosti vstupu. V těle smyčky se hledá index nejbližšího clusteru a následně se počítá průměr pomoci SUM Allreduce O(log n). V poslední části ROOT proces sbírá clustery příslušných dat pomoci Gather O(log n) , které následně vypíše O(n). Pokud zanedbáme častí načítaní a výpisu dat dostaneme celkovou složitost O(T * log n), kde T je počet iterací.

## 1.2 Prostorová analýza

ROOT proces potřebuje O(n) pro udrženi, distribuci a sběr vstupu. Avšak ostatním procesům stačí O(1) pro udržení vlastní hodnoty a K průměrů.

## 1.3 Cena

Procesorů je O(n) přičemž každý pracuje po dobu O(T * log n), tudíž celková cena algoritmu je O(T * n log n).