Výkonnostní měřítka paralelních algoritmů, PRAM model, APRAM model, škálovatelnost

velikost vstupních dat

$T_A^K(n)$ - časová složitost seq. algoritmu A, řešícího problém K

$SL(n)$ - spodní mez časové složitosti (triviální dána velikostí n)

$SU(n)$ - horní mez nejlepšího existujícího algoritmu (*)

počet procesorů (jader, vláken)

$T(n,p)$ - paralelní čas

*) $T_A(n) = \Theta\, SU^K(n) = \Theta\, SL^K(n)$ = asymptoticky optimální algoritmus

Paralelní zrychlení:

$$\boxed{\underset{\text{speedup}}{S}(n,p) = \frac{SU(n)}{T(n,p)} \leq p}$$

→ lineární pokud $S(n,p) = \Theta p$

– superlineární zrychlení ....

Spodní mez:

$$\boxed{\underset{\text{lower bound}}{L}(n,p) = \frac{SL(n)}{p}}$$

Paralelní cena:

$$\boxed{\underset{\text{cost}}{C}(n,p) = p \cdot T(n,p)} = \Omega(SU(n))$$

→ cenově optimální algoritmus pokud $C(n,p) = \Theta(SU(n))$

Paralelní efektivnost:

$$\boxed{E(n,p) = \frac{SU(n)}{C(n,p)}} = \frac{S(n,p)\cdot T(n,p)}{p\cdot T(n,p)} = \frac{S(n,p)}{p} \leq 1$$

– algoritmus je konstantně efektivní

$$E(n,p) \geq E_0 \quad \text{pro} \quad 0 < E_0 < 1$$

cenová optimalita <=> lineární zrychlení <=> konstantní efektivnost

PRAM Model

RAM model - Random access machine
- instrukce s jednotkovým časem
- časová složitost = počet provedených instrukcí
- paměťová složitost - počet použitých buněk nekonečné pásky

PRAM: množina p procesorů
- ∀ procesor má index a lokální paměť
- m sdílených paměťových buněk - každý má do jakékoliv access v $O(1)$
  → konflikty se musí explicitně řešit
- typy operací: READ, WRITE, LOCAL
  - jednotkový model: všechny operace trvají 1
  - globální model: L trvá 1, R/W konst. čas $d > 1$

EREW - PRAM - žádné dva procesory nemohou R/W do stejné buňky najednou
CREW - PRAM - čtení více najednou je povoleno
CRCW - PRAM - i zápis je možný najednou
- common - všechny zapisované hodnoty musí být stejné
- priority - pevné přidělení priority každému procesoru
- arbitrary - náhodné pořadí

APRAM Model

- asynchronní zpracování
- je nutná explicitní synchronizace bariérou
- není jednotková doba přístupu do sdílené paměti
- výpočet: posloupnost globálních fází oddělených bariérou

Paralelní škálovatelnost

- schopnost par. algoritmu držet paralelní optimalitu při změnách p a n

Silná: jak klesá E při pevném n pokud roste p, slabá: pevné p a roste n

Amdahlův zákon saturace paralelizace

$$S(n,p) = \frac{T(n)}{f_s \cdot T(n) + \frac{1-f_s}{p} \cdot T(n)} = \frac{1}{f_s + \frac{1-f_s}{p}} \le \frac{1}{f_s}$$

zrychlení nemůže přesáhnout

inherentně sekvenční podíl úlohy

paralelní podíl

př. $f_s = 10\%$ => $S(n,p) \le \frac{1}{0,1} = 10$ pro jakékoliv p

Gustavsonův zákon

$$S(n,p) = \frac{t_{seq} + t_{par}(n,1)}{t_{seq} + t_{par}(n,p)}$$

„s rostoucím p musíme úměrně zvyšovat i velikost problému n"

- inherentně sekvenční část trvá t_{seq} (konstantní)

→ inherentně paralelní část bude lineárně škálovat