Tečajevi naprednog računarstva u Križevcima

HACK2020 | Hub for Advanced Computing Križevci

Osnove računarstva visokih performansi

Vol. 9: MPI – Message Passing Interface

Andrej Dundović

Križevci, 17. 7. 2021.

Organizator:

udruga point križevci

Pokrovitelj:

Embassy of the United States
Zagreb - Croatia

Paralelizacija na razini više umreženih računala

- Ako smo iskoristili sve mogućnosti da skratimo vrijeme izvršavanja našeg programa na jednom računalu, ostaje nam samo da se program izvršava na više računala paralelno...
- U nekim slučajevima nije samo cilj smanjiti vrijeme izvršavanja, već povećati dostupnu memoriju (32-bitna računala: $2^{32} - bit = 4\text{GB}$, 64-bitna: $2^{64} = 16\text{EB}$, exabyte - u teoriji, u praksi - 4TB)
- Uvođenjem više računala kompleksnost cijelog sustava još se više povećava, a najznačanije je da imamo novo usko grlo koje moramo uzeti u razmatranje: računalna mreža koja povezuje ta računala
- U *distribuiranom računarstvu* nema (hardverske, globalne) dijeljenje memorije, već svako računalo – *čvor* ili *node* – ima svoju lokalnu memoriju, a komunikacija se vrši razmjenom poruka (*message passing*)
- Ekstremni primjer: The Berkeley Open Infrastructure for Network Computing (BOINC)[1] koji je isprva razvijen za projekt SETI@home
- Po nekim klasifikacijama, nasuprot *DR* imamo *paralelno računarstvo* gdje računala mogu dijeliti memoriju

[1] guinnessworldrecords.com

Sumirano

- Danas imamo procesor koji može obraditi više podataka primjenjujući istu instrukciju odjednom (vektorizacija, odnosno SIMD, zadnje je AVX-512: 16x float)
- Više "procesora" integrirano je na jednu pločicu i imaju ulogu jezgre u novom višejezgrenom procesoru (max. 64 jezgre kod AMD EPYC 7763)
- U jednom računalu možemo ugraditi više takvih procesora (2 ili 4)...
- Paralelizacija na razini više računala – nema definirane gornje granice
- Trenutni predvodnik na TOP500: Supercomputer Fugaku (Japan): Cores: 7,630,848; Memory: 5,087,232 GB; Processor: A64FX 48C 2.2GHz [2]
- Ovdje se nismo ni dotakli specijaliziranih procesora, kao što su GPU ili FPGA (*stream processing* $\Leftrightarrow$ SIMD) – za neku drugu prigodu :-)

[2] TOP500.org: Supercomputer Fugaku

Saturacija: Amdahlov *zakon*

- Ako program podijelimo na dio koji se može paralelizirati (tj. može iskoristiti dodatne resurse) s udjelom p i onaj koji se ne može s udjelom $1 - p$ možemo izračunati teoretsko maksimalno ubrzanje *Amdahlov zakon*
- Vrijeme izvršavanja programa dano je s:

$$T = (1 - p)T + pT \tag{1}$$

- pT dio može se uzbrzati za faktor s pa je radi toga ukupno vrijeme skraćeno na:

$$T(s) = (1 - p)T + \frac{p}{s}T \tag{2}$$

- Iz čega izračunamo faktor uzbranja

$$S_{\text{speed}-\text{up}} = \frac{T}{T(s)} = \frac{1}{1 - p + \frac{p}{s}} \tag{3}$$

- Primjer $s \to \infty$ (ako dodajemo beskonačno mnogo paralelnih računala) saturiramo na $1/(1 - p)$.

MPI – Message Passing Interface

- Iako možemo i sami razviti protokol i definirati komunikaciju između procesa na različitim računalima...
- ...MPI je standardiziran – to mu je najveća prednost – dostupan je na gotovo svakom grozdu (*clusteru*)

```
mpirun -n 4 ./program
```

- Terminologija
  - Rank – identifikator procesa (ima ga svaki MPI proces), počinje od 0 do $n-1$ gdje je $-n$ broj paralelizacija (specificirano s `-n`); identifikator je spremljen u varijablu **MPI_Comm_rank**, a ukupan n u **MPI_Comm_size**
  - test