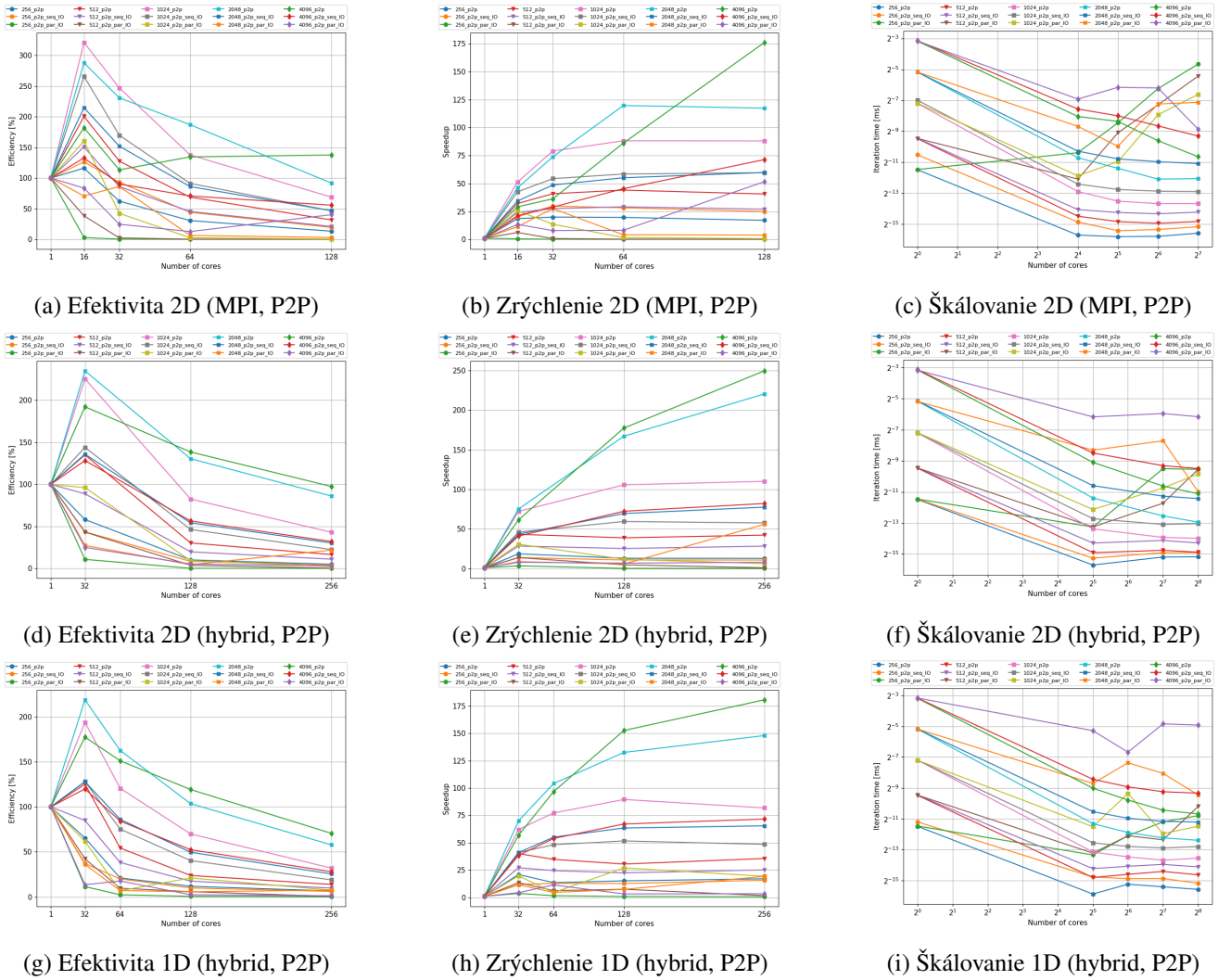
Praktické paralelné programovanie
Správa k projektu MPI a paralelné I/O

Vypracoval: Pavel Kratochvíl (xkrato61)

Aký je rozdiel medzi škálovaním/efektivitou 1D a 2D dekompozície a čím je tento rozdiel spôsobený?

Dvojdimenzionálna dekompozícia sa podľa očakávania ukazuje ako efektívnejšia v niekoľkých aspektoch. Objem a frekvencia medzi susedmi sa zdá optimálnejšia ako pri 1D dekompozícii a navyše je pri niektorých volaniach (napr. výpočet priemernej teploty v strednom stĺpci) potrebný výpočet rozdelený medzi viacero procesov (resp. vlákien). Taktiež dochádza k silnému škálovaniu vďaka efektívnejšiemu prekrytiu výpočtu a komunikácie.



(a) Efektivita 2D (MPI, P2P) (b) Zrýchlenie 2D (MPI, P2P) (c) Škálovanie 2D (MPI, P2P)

(d) Efektivita 2D (hybrid, P2P) (e) Zrýchlenie 2D (hybrid, P2P) (f) Škálovanie 2D (hybrid, P2P)

(g) Efektivita 1D (hybrid, P2P) (h) Zrýchlenie 1D (hybrid, P2P) (i) Škálovanie 1D (hybrid, P2P)

Obr. 1: Grafy silného škálovania pre 2D v konfigurácii MPI (a, b, c) a 2D (d, e, f) a 1D dekompozíciu (g, h, i) v hybridnej konfigurácii (OpenMP + OpenMPI).

Pri porovnaní MPI a hybridnej konfigurácie by mal byť teoreticky výraznejší rozdiel v čase potrebnom na komunikáciu medzi procesmi v prospech hybridnej konfigurácie, keďže dochádza k menšiemu množstvu komunikácie medzi fyzickými uzlami.

Najväčší pokles efektivity nastáva pri kombinácii 1D dekompozície, paralelného I/O, väčšej domény, RMA a väčšieho počtu jadier (výrazný pokles nastáva pri 64 jadrách), kedy pravdepodobne dochádza k saturácii množstva zápisov na OSTs (Object Storage Targets) kvôli mnohým malým žiadostiam o zápis.

RMA grafy sú v odovzdanom archíve. Všeobecne pri RMA konfigurácii klesala efektivita, čo je zrejme spôsobené privysokou réžiou na vytváranie okien a následnou výmenou relatívne malého množstva dát medzi malým počtom uzlov.

Aký je vplyv paralelného I/O v porovnání so sekvenčným?

Pri paralelnom I/O pozorujeme zhoršenie škálovania a výrazný pokles v efektivite a nárast doby behu (v porovnaní s behmi programu so sekvenčným zápisom). Najväčší pokles pozorujeme pri ukladaní mnohých veľmi malých častí, takže doba behu programu je priamo úmerná počtu jadier a nepriamo úmerná veľkosti domény. Zaujímavé je, že v žiadnej konfigurácii nedošlo ku zlepšeniu doby behu pri využití paralelného I/O pri porovnaní so sekvenčným I/O.

Akým spôsobom možno zefektívniť paralelné I/O?

Vhodným nastavením LFS (Lustre File System) možno výrazne zefektívniť zápis výstupných dát. Zvolil som hodnotu `stripe count=5`, čo je počet zhodný s počtom OSTs na clustri Barbora a `stripe size=512k`, ktorý sa rovná približne priemeru veľkosti časti ukladanej jedným rank-om. Týmto nastaveniam som prispôsobil aj zarovnanie pamäti pri ukladaní (`H5Pset_alignment`) a chunking pri ukladaní na veľkosť lokálnej dlaždice.

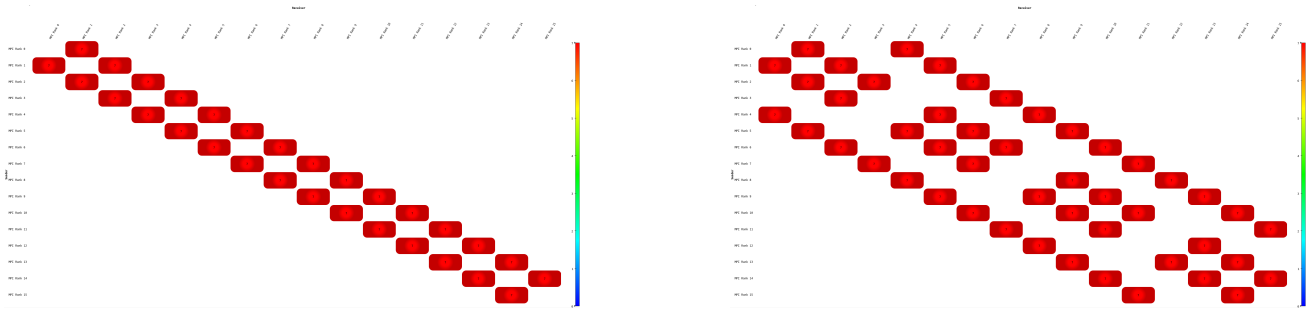
Ako sa líši množstvo komunikácie medzi jednotlivými procesmi v 1D a 2D dekompozícii. Je záťaž vyrovnaná?

Pri 1D dekompozícii komunikácie vymieňajú jednotlivé ranky dáta s jedným až dvomi susedmi, kým pri 2D dekompozícii komunikujú s dvomi až štyrmi susedmi. Pri zvolení P2P komunikácie a 1D dekompozície je počet samotných volaní `MPI_Isend` a `MPI_Irecv` menší, no objem zasielaných dát je väčší, čo sa prejaví na skoršom dosiahnutí eager limitu a tým pádom aj v navýšení času strávenom na dokončenie komunikácie rendezvous protokolom.

(a) 1D dekompozícia (b) 2D dekompozícia

Obr. 2: Rozloženie objemu komunikácie medzi rank-ami v rôznych dekompozíciach (snímky sú z nástroja Cube).

Záťaž jednotlivých procesov ako aj individuálnych vlákien je vyvážená až na sekvenčnú časť, ktorú vykonáva root rank (načítanie vstupu, distribúcia dát, zozbieranie výsledkov ako aj progress reporting). Avšak procesy, ktoré majú na starosť krajné časti domény musia logicky menej komunikovať a výpočet halo zón im zaberá výrazne kratšie.

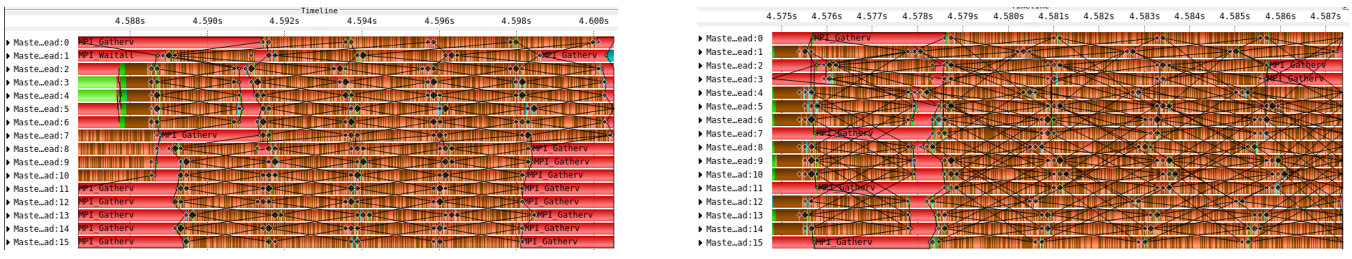


(a) 1D dekompozícia (b) 2D dekompozícia

Obr. 3: Matice komunikácie (histogram počtu zaslaných správ medzi rank-ami) pre rôzne dekompozície (snímky sú z nástroja Vampir).

Aký prínos má prekrytie komunikácie a výpočtu?

Prínos spočíva vo využití asynchrónnej komunikácie a efektívnom zúžitkovaní času, ktorý by sme inak stratili pri použití synchrónnej komunikácie a následnom čakaní na jej dokončenie. Aj vďaka prekrytiu výmeny halo zón s výpočtom vnútornej časti dlaždice dosahuje program niekoľkonásobného zrýchlenia v porovnaní so sekvenčnou verziou.



(a) 1D dekompozícia (b) 2D dekompozícia

Obr. 4: Vizualizácia komunikácie pri rôznych dekompozíciach. Čierne čiary znázorňujú zaslané MPI správy medzi rank-ami. V oboch dekompozíciach je tiež vidno vykonávanie výpočtu v čase, kedy zároveň prebiehala P2P komunikácia.