OPTIMIZATION OF THE DISTRIBUTED I/O SUBSYSTEM OF THE K-WAVE PROJECT

Ondřej Vysocký
Bachelor Degree Programme (3BIT), FIT BUT
E-mail: xvysoc01@stud.fit.vutbr.cz

Supervised by: Jiří Jaroš
E-mail: jarosjir@fit.vutbr.cz

Abstract: This paper deals with an effective solution of parallel writting of variable amount of data on supercomputer facilities. The work will be used by the k-Wave project, designed for time domain acoustic and ultrasound simulations. Because the simulation is computationally and data intensive, the project requires to be implemented with libraries for parallel computig (OpenMPI) and large data processing (HDF5) and it must run on a supercomputer. The result application is implemented in C++ and uses previously mentioned libraries, as well as k-Wave project. The paper presents first experimental results on the IO optimisation.

Keywords: parallel programming, C++, OpenMPI, Message Passing Interface, HDF5, Hierarchical Data Format, k-Wave, supercomputer, I/O, Lustre, read, write, large amounts of data, optimalization

1 ÚVOD

Bakalářská práce se zabývá efektivním paralelním zápisem velkého množství dat ve formátu HDF5 za pomocí rozhraní pro komunikaci paralelních procesů OpenMPI. Jedná se o problematiku přeuspořádání velkého objemu dat mezi jednotlivé procesy tak, aby bylo dosaženo co možná nejrychlejšího zápisu do HDF5 souboru na souborovém systému Lustre. Výsledný program je součástí projektu k-Wave [1] (kde nahrazuje původní třídu zajišťující zápis do souboru), který pracuje na simulaci akustických a ultrazvukových vln, avšak jeho využití je možné ve všech aplikacích s vysokými nároky na zápis a čtení dat z pevných disků.

2 NÁSTROJE

- Message Passing Interface
  MPI (Message Passing Interface) [2] je specifikace rozhraní pro zasílání zpráv v paralelních programech. V tomto projektu pracuji s implementací OpenMPI.

  MPI specifikuje také způsob zápisu do souboru. Tato část se nazývá MPI I/O. Ve vyvíjené aplikaci se však nevyužívá těchto funkcí přímo, nýbrž pomocí funkcí knihovny HDF5 [3].

- Hierarchical Data Format
  Hierarchický datový formát (HDF) [4] je samo-popisný souborovým formátem pro usnadnění práce a uložení výzkumných dat nezávisle na operačním systému a architektuře počítače. Tento formát umožňuje paralelní zápis do souboru pomocí takzvaných hyperslabů, vymezujících prostor pro data jednotlivého zapisujícího procesu.

- Lustre
  Lustre [5] je distribuovaný souborový systém pojící disková pole RAID v jeden celek. Tento systém využívaný na počítačových clustrech poskytuje možnost prokládání, čímž je umožněn vysoký výkon při paralelním přístupu.

3 PROBLEMATIKA

V průběhu simulace dochází k progresivnímu ukládání velkých objemů dat. Z hlediska efektivity není výhodné rozdělit zátěž na všechny procesy, ani ji směřovat pouze na jeden proces.

V případě, že bychom tato data rozdělili mezi všechny procesy, kde každý zapisuje přibližně KB dat, dochází k zahlcení pevných disků velkým množstvím malých požadavků a v důsledku toho nastává výrazné zpomalení.

Naopak můžeme tento problém řešit přesunem všech dat na jediný proces, který následně provede zápis na disk. V tomto případě dojde nejprve k zahlcení síťové karty, jež zajišťuje přenos dat mezi procesy a následně tento jeden proces bude na pevný disk zapisovat data v řádu GB, zatímco zbylé procesy nebudou využity vůbec.

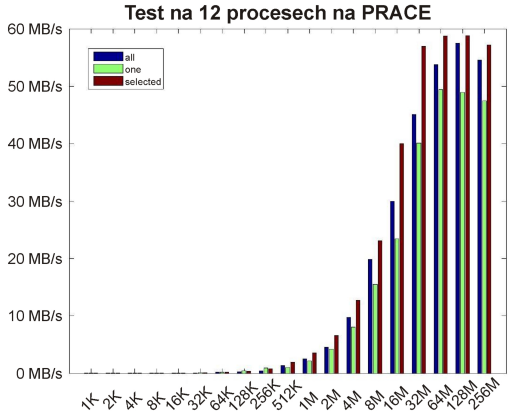
Proto jsem vytvořil program, v němž se provede rozdělení procesů do skupin (pomocí MPI_Comm_Split), následně v každém kroku simulace provede přesun dat do jednoho procesu v každé skupině (pomocí MPI_Gatherv). Tyto procesy následně provedou paralelní zápis do souboru.

Při simulaci šíření vln často dochází k neuniformnímu rozdělení dat mezi jednotlivé procesy (v případě že nás zajímá stav pouze v určité části z celé simulační domény), tudíž některé procesy pracují s megabajty, další mohou zpracovávat kilobajt dat, ale také nepracovat vůbec. Pokud necháme procesy zapisovat data tak, jak byla rozdělena bez vyvážení zátěže, nastává problém s velkým množstvím menších požadavků. Je nutno zdůraznit, že toto nastavení zápisu výrazně zpomaluje režie zajišťující spolehlivý zápis do souboru, která je s paralelním zápisem dat spojena. Ta narůstá především v případě, že všechny procesy nezapisují stejné množství dat. Přesná míra režie MPI při tomto výpočtu bude určena pomocí profileru, který bude nainstalován správci počítače Anselm v Ostravě.

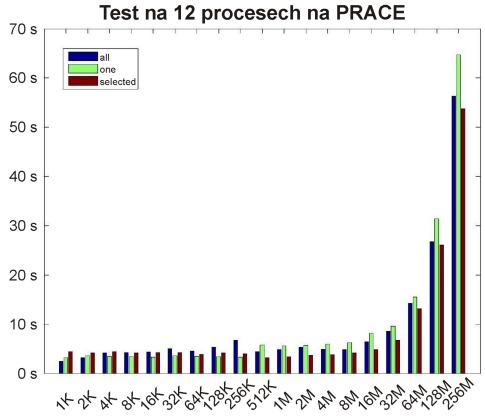
Je důležité podotknout, že ukládání dat je blokující operace, což znamená, že všechna jádra nemohou pokračovat ve výpočtu, dokud nedokončí zápis nejpomalejší jádro. Proto chceme docílit co nejrychlejšího zápisu a při práci na větším počtu jader, chceme aby pracovaly po stejně dlouhou dobu. Cílem práce tedy je vytvořit metriku pro vyvážení zátěže mezi jednotlivé procesy a zajistit tak co nejrychlejší zápis dat z pevných disků.

4 TESTOVÁNÍ

Na základě dříve zmíněných tří možností zápisu dat jsem vytvořil program testující rychlosti zápisu (MB/s) jednotlivých způsobů při různém množství dat na každý jeden proces. Výsledkem tohoto testu je graf 1.



Obrázek 1: Porovnání rychlosti zápisu různého množství dat na proces v MB/s.



Obrázek 2: Porovnání doby zápisu různého množství dat na proces v sekundách.

V grafu můžeme pozorovat velmi nízkou rychlost zápisu při malých datových objemech. Ztráta rychlosti je způsobena latencí distribuovaného diskového pole Lustre velkým množstvím režie, která se musí vždy provést. Z toho odvozujeme, že v případě zápisu malého množství dat je výhodné zápis odložit a provést jej až po nastřádání dalších dat - například při dalším kroku simulace.

Dále bychom měli v grafu porovnávat především zápis pomocí všech procesů (vyobrazen modrou barvou) a zápis pouze vybranými procesy (červenou barvou), které vykazovaly nejlepší výsledky. Při posledních testech (viz 1) dosahoval zápis pomocí vybraných procesů lepších výsledků, avšak veškeré testy nebyly takto jednoznačné. Bude nutné provést měření s mnohem větším počtem procesorových jader (128-1024).

Tento test nám dává uvedené informace, ale musíme však vzít v potaz, že všechny procesy zapisovaly stejné množství dat, což jsou ideální podmínky pro paralelní zápis do souboru, a to zlepšuje především rychlost zápisu všech procesů. Při zápisu pomocí vybraných procesů, jsem zjistil, že čas potřebný na přesun dat mezi procesy znamenal v určitých případech až 50% z celkové doby zápisu. V neprospěch tohoto řešení dále mluví, fakt že test probíhal na pouhých dvanácti procesech (2 uzly o 6 jádrech) a režie MPI s počtem procesů roste.

Z těchto důvodů budu nyní provádět testy zjišťující vliv počtu procesů na rychlost zápisu, stejně tak budu testovat vliv proměnného množství dat na jednotlivých procesech. Také budu testovat, jak ovlivní rychlost zápisu nastavení dělení dat do chunků (nejmenší zapisovatelná jednotka) a možnosti nastavení Lustre.

5 ZÁVĚR

Z již provedených a nyní naplánovaných testů je zapotřebí vyvodit výsledky, které určí za jakých podmínek bude nejvýhodnější určitý způsob zápisu. Pokud tyto podmínky nebudou přesně dané, pravděpodobně bude zapotřebí využít fuzzy logiky.

V současné době je tento program testován pouze na superpočítači Supernova v Polsku a na strojích MetaCentra, bude však vhodné porovnat výsledky testů na dalších superpočítačích, aby byla zajištěna univerzalita řešení.

PODĚKOVÁNÍ

Tato práce využívá výpočetní zdroje v rámci projektu "Simulating the effect of fiducial markers on high-intensity focussed ultrasound treatments of the prostate", PRACE, 11DECI0130, 2013-2014. Dále pak přístup k výpočetním a datovým zdrojům the National Grid Infrastructure MetaCentrum, pod záštitou projektu "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005). Tento článek vznikl za podpory projektu VUT v Brně FIT-S-14-2297.

REFERENCE

[1] WWW stránky: Oficiální stránky projektu k-Wave [online]. http://www.k-wave.org.

[2] WWW stránky: Oficiální stránky The Message Passing Interface Standard [online]. http://www.mcs.anl.gov/research/projects/mpi/.

[3] Cheng, A.: Parallel HDF5 Tutorial. http://www.speedup.ch/workshops/w37_2008/HDF5-Tutorial-PDF/PSI-HDF5-PARALLEL.pdf,2008-09-08.

[4] WWW stránky: HDF5 [online]. http://www.hdfgroup.org/HDF5.

[5] WWW stránky: Oficiální stránky The Lustre [online]. http://wiki.lustre.org/index.php/Main_Page.