



UNIUNEA EUROPEANĂ



GUVERNUL ROMÂNIEI



Fondul Social European
POSDRU 2007-2013



Instrumente Structurale
2007-2013



OIPOSDRU



UNIVERSITATEA
TEHNICĂ
CLUJ-NAPOCA

Investește în oameni!

Proiect cofinanțat din Fondul Social European prin Programul Operațional Sectorial pentru Dezvoltarea Resurselor Umane 2007 – 2013

Axa prioritară: 1 „Educația și formarea profesională în sprijinul creșterii economice și dezvoltării societății bazate pe cunoaștere”

Domeniul major de intervenție: 1.5 „Programe doctorale și postdoctorale în sprijinul cercetării”

Titlul proiectului: „Parteneriat inter-universitar pentru excelența în inginerie - PARTING”

Cod Contract: POSDRU/159/1.5/S/137516

Beneficiar: Universitatea Tehnică din Cluj-Napoca

ing. Adrian SABOU

TEZĂ DE DOCTORAT

-Rezumat-

SIMULAREA PARALELĂ A MODELELOR TRIDIMENSIONALE BAZATE PE PARTICULE

Conducător științific

Prof. dr. ing. Dorian GORGAN

UNIVERSITATEA TEHNICĂ
DIN CLUJ-NAPOCA
FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE

2014

Rezumat

1. Introducere, context și motivație

Modelarea și simularea suprafețelor tridimensionale dinamice reprezintă una din cele mai importante direcții de cercetare din domeniul graficii pe calculator și probabil una din cele mai interesante. Tehnicile cele mai proeminente pentru astfel de simulări sunt metodele bazate pe modele fizice, care au devenit o componentă importantă a industriei de divertisment și sunt utilizate regulat în jocuri pe calculator, filme și diferite animații bazate pe calculator, pentru a obține efecte vizuale ce implică interacțiunea cu fluide, materiale textile, corpuri deformabile și rigide. Unul din cele mai importante subdomenii ale modelării bazate pe modele fizice este modelarea bazată pe particule, în cadrul căreia suprafețele sunt approximate prin mulțimi de puncte discrete având diferite proprietăți fizice cum ar fi masă, volum, viteză, accelerație, iar comportamentul acestor suprafețe împreună cu interacțiunea lor cu mediul înconjurător este guvernată de legile fizicii, mai specific de forțele care acționează asupra particulelor. Modelarea bazată pe particule este o alegere naturală din moment ce, în lumea reală, interacțiunile care implică suprafețe deformabile au loc la un nivel molecular, iar, teoretic, având un număr suficient de particule, orice astfel de suprafață poate fi simulată cu precizie. Astfel de tehnici de modelare depășesc mult tehnicile de animație folosind cadre cheie în ceea ce privește nivelulul de realism și complexitatea.

Obiectele maleabile, deformabile există peste tot în jurul nostru și nu ne mai putem imagina o simulare computerizată a lumii reale fără simularea comportamentului acestor suprafețe cât mai realist cu putință. Modelele fizice au fost utilizate cu succes încă de când au fost introduse pentru dezvoltarea de conținut vizual pe calculator pentru un public din ce în ce mai greu de impresionat. Cu toate acestea, goana după realism a atras după sine o nouă provocare, și anume goana după putere de calcul și eficiență. Odată cu evoluția componentelor hardware secvențiale precum unitatea centrală de procesare (CPU), algoritmi și tehnicile utilizate au suferit puține modificări, creșterile de performanță fiind obținute în principal prin frecvențe crescute de operare, dar, odată cu plafonarea vitezelor de procesare ale CPU la 4 GHz, temuta „barieră termală” a împiedicat progresele ulterioare în această direcție. Totuși, pentru o parte a utilizatorilor acest lucru a fost insuficient, astfel începând „era paralelă”. CPU a evoluat spre arhitecturi cu nuclee multiple și au apărut componente hardware specializate pentru calcul paralel în forma unităților de procesare grafică (GPU). Cu această nouă direcție de calcul a apărut nevoia dezvoltării unor tehnici paralele eficiente pentru a putea utiliza resursele existent într-un mod optimal.

Printre aplicațiile care pot beneficia de pe urma accelerării GPU sunt și simulările bazate pe particule, natura inherent paralelă a modelelor bazate pe particule făcându-le potrivite pentru implementări bazate pe GPU. Astfel, în ultimul deceniu, cercetările în această direcție au rezultat în dezvoltarea de tehnici și metode paralele foarte eficiente pentru accelerarea etapelor cele mai costisitoare ale simulării, anume integrarea numerică și detecția coliziunilor. Cu toate acestea, odată cu mărirea rezoluției modelelor simulate în goana după realism, unitățile individuale GPU nu mai

pot ține pasul cu scenariile din ce în ce mai complicate, în special din cauza limitelor memoriei. Astfel, pentru a continua să forțăm limitele modelării bazate pe modele fizice, este necesar să ne îndreptăm atenția spre arhitecturi de calcul mai scalabile precum clusterelor grafice.

Este convingerea noastră că simulările bazate pe particule ar putea să beneficieze într-o mare măsură de pe urma utilizării unor astfel de arhitecturi de calcul performant, astfel, motivația acestei teze centrându-se în jurul acestei idei, deși există provocări, cu precădere la nivelul dezvoltării de aplicații. În primul rând, datorită construcției și modului de operare, paradigme specifice și memoriei partajate și memoriei distribuite se aplică în dezvoltarea aplicațiilor ce vor rula pe clustere GPU. Adăugând la aceasta atributul de timp real pe care, de obicei, astfel de simulări îl necesită, împreună cu nevoia de a transforma datele brute procesate într-o formă vizuală relevantă pentru utilizatori, obținem probabil una din cele mai complexe clase de aplicații pentru a fi implementate pe astfel de arhitecturi de calcul. Principalele provocări includ paralelizarea modelelor fizice, implementarea integrării numerice paralele, detecția paralelă a coliziunilor, procesarea distribuită a modelelor pe mai multe noduri ale clusterului, vizualizarea și interacțiunea cu scena de la distanță și, cel mai important, asigurarea comunicării și sincronizării între toate aceste procese. Astfel, orice abordare a simulării bazate pe particule pe clustere GPU trebuie să nu se limiteze la a aborda individual fiecare din provocările de mai sus și la optimizarea lor individuală, ci trebuie să descopere interdependențele dintre acestea și să le trateze ca pe module care contribuie la o simulare eficientă, ținând cont de particularitățile hardware menționate. Mai mult, acest nou tip de calcul de înaltă performanță ar trebui să fie disponibil unui public mai larg prin interfețe de programare de nivel înalt. Spre exemplu, algoritmi puternici derivați din analiza numerică ar putea fi dezvoltați pentru simularea bazată pe particule de către cercetători din afara sferei științei calculatoarelor, care nu ar vrea să fie împovărați cu toate detaliile de nivel scăzut al funcționării clusterelor. În acest context, ne propunem să cercetăm particularitățile unei arhitecturi de dezvoltare software optimizată, modulară și puternic scalabilă pentru dezvoltarea scenariilor de simulare bazată pe particule pe clustere GPU.

2. Obiectivele cercetării

Obiectivul principal al acestei cercetări este optimizarea aplicațiilor de calcul performant vizuale și interactive precum simulările bazate pe particule pentru rularea acestora pe clustere GPU. Obiectivul este axat în jurul cercetării particularităților unei arhitecturi de dezvoltare software optimizată, paralelă, distribuită, modulară și puternic scalabilă pentru dezvoltarea scenariilor de simulare bazată pe particule.

Alte obiective importante, care derivă din obiectivul principal, sunt:

- identificarea unei posibile structuri modulare pentru arhitectura de dezvoltare software;
- identificarea tiparelor de comunicație dintre modulele individuale;
- propunerea unei arhitecturi conceptuale bazată pe proiectarea modulară optimală și tiparele de comunicație identificate;
- implementarea și optimizarea modulelor individuale prin adaptarea tehnicilor paralele utilizate în prezent pentru simularea bazată pe modele fizice în vederea folosirii acestora pe clustere GPU;

- extinderea modelelor bazate pe particule și a arhitecturii de dezvoltare software în vederea utilizării acestora în afara sferei animației pe calculator;
- testarea și evaluarea arhitecturii propuse;

3. Structura tezei

Teza este structurată în opt capitole, având 153 de referințe bibliografice. Contribuțiile subliniate pe parcursul lucrării au fost publicate în 1 articol în jurnal indexat B+, 5 articole la conferințe internaționale indexate ACM sau IEEE și 3 prezentări la diferite conferințe, workshop-uri și training-uri.

Capitolul 1 reprezintă o introducere în domeniul cercetării, care descrie contextul și motivația prezentei cercetării, alături de principalele obiective.

Capitolul 2 oferă o analiză asupra arhitecturilor de calcul paralele și distribuite. Se axează pe sisteme cu memorie partajată și memorie distribuită, evidențiind avantajele și dezavantajele pe care fiecare categorie le oferă, din punctul de vedere al modelării prin particule. Sunt analizate arhitecturi moderne cu memorie partajată precum unitatea de procesare grafică și, la final, concluzionăm că, pentru a putea oferi o soluție rapidă și scalabilă pentru simulatoarele bazate pe particule, arhitecturile hibride precum clusterelor grafice ar putea oferi cele mai multe avantaje, prin combinarea punctelor forte ale sistemelor cu memorie partajată și memorie distribuită.

Capitolul 3 reprezintă un studiu critic al modelelor și tehnicilor utilizate în simulările bazate pe particule. Acesta oferă o imagine de ansamblu asupra cercetărilor similare existente în domeniul simulării bazate pe particule, precum și o încercare de descriere formală, la nivel înalt, a modelelor de particule. Sunt analizate mai multe modele pentru simularea țesăturilor și fluidelor din punctul de vedere al potențialului de paralelizare, precum modelele bazate pe particule și modelele bazate pe dinamica mediilor continue, precum și cele mai utilizate tehnici de integrare numerică explicită și implicită și de detecție a coliziunilor. Descrierea formală a modelelor de particule abstractizează conceptul de particulă în vederea facilitării unei posibile extinderi ale simulatoarelor bazate pe particule în afara domeniului animației pe calculator. O structură modulară inițială pentru arhitectura de dezvoltare software, bazată pe principalele etape ale simulărilor bazate pe particule, este descrisă spre finalul capitolului.

Capitolul 4 prezintă în detaliu tehnicile de paralelizare dezvoltate în vederea accelerării simulărilor bazate pe particule pe cluster grafice. Capitolul cuprinde cele mai importante probleme ale simulării paralele bazată pe particule precum descompunerea și distribuirea modelului la nivel GPU și la nivel de cluster, paralelism hibrid CPU/GPU și integrarea numerică paralelă. Pentru partiționarea modelului am propus o metodă de descompunere statică a domeniului care asigură frontiere de lungime minimă și astfel minimizează traficul prin rețea necesar sincronizării. Am propus de asemenea și o tehnică ce asigură modificări minime asupra nucleelor folosite pe un singur nod, prin păstrarea unui model extins pe fiecare nod de procesare. Paralelismul CPU/GPU a fost utilizat în vederea minimizării timpilor de inactivitate ale CPU în timp ce se așteaptă finalizarea execuției GPU. Prin utilizarea unei tehnici care permite decuplarea anumitor pași ai simulării, am putut rula simularea într-o manieră defazată și să efectuăm sarcinile de sincronizare dependente de CPU în timp

ce GPU execută sarcinile complexe de calcul. Integrarea explicită paralelă este obținută folosind o abordarea paralelă la nivel de date în două treceri, în timp ce, pentru integrarea implicită, ne-am bazat pe o variantă paralelă a algoritmului gradientului conjugat. Pentru a accelera și mai mult calculele și pentru reducerea cerințelor de memorie, am dezvoltat o tehnică eficientă pentru actualizarea matricelor rare de dimensiuni mari implicate în integrarea implicită direct în formatul Compressed Sparse Row, prin exploatarea structurilor regulate ale acestora.

Capitolul 5 introduce *ParTSim*, o arhitectură de dezvoltare software paralelă, distribuită și modulară pentru rularea simulărilor bazate pe particule pe clustere grafice. Arhitectura se bazează pe o structură client-server și, la nivel de server, este proiectată pe baza modulelor principale identificate ca indispensabile pentru acest tip de simulări, împreună cu modulele introduse de către particularitățile procesului de calcul pe aceste arhitecturi distribuite și la distanță. Bazată pe standarde deschise precum OpenCL și OpenGL, această arhitectură utilizează interoperabilitatea între calculul general și trasarea grafică pe același GPU, reducând astfel transferurile de memorie între dispozitiv și gazdă și îmbunătățind considerabil performanțele. Structura modulară permite o flexibilitate crescută în combinarea diferitelor modele cu diferite tehnici de integrare numerică și metode de detecție și rezolvare a coliziunilor. Acest capitol tratează și problemele care apar în contextul clusterelor grafice din cauza naturii distribuite și la distanță a procesării. Pentru a obține vizualizare centralizată performantă, am prezentat o tehnică optimizată de trasare paralelă. Bazată pe o abordare de trasare paralelă „sortare-după” cuplată cu un algoritm „zonă-de-interes”, această tehnică îmbunătățește considerabil performanțele simulării când vizualizarea întregii scene și a modelului trebuie făcută pe un singur dispozitiv de afișare. Vizualizarea și interacțiunea la distanță este asigurată printr-o soluție bazată pe protocolul VNC care oferă performanțe bune pentru rularea aplicațiilor grafice interactive la distanță chiar și în lipsa dispozitivelor fizice de afișare, cum se întâmplă la majoritatea clusterelor GPU.

Capitolul 6 își propune extinderea modelelor bazate pe particule din punctul de vedere al scopului inițial al acestora. Bazându-ne pe descrierea formală din capitolul 3, identificăm corespondențe posibile între modelele bazate pe particule utilizate în grafica pe calculator și modele care aparțin altor domenii științifice, precum sociofizica și econofizica. Acest lucru ne permite să utilizăm arhitectura *ParTSim* pentru accelerarea simulărilor din afara sferei dinamicii corpurilor deformabile. Ca studiu de caz, simulăm un model existent pentru adopția de tehnologie folosind o abordare vizuală bazată pe particule. Identificăm corespondențele dintre elementele principale ale modelelor de particule tradiționale și elementele modelului sociofizic și rulăm simularea interactiv utilizând arhitectura *ParTSim*.

Capitolul 7 descrie validarea și evaluarea experimentală a arhitecturii *ParTSim* și a tuturor tehnicilor de simulare paralelă prezentate în capitolele precedente. Pentru evaluarea performanțelor utilizăm metrici specifice atât calculului paralel cât și graficii pe calculator, precum timpul de execuție, speedup-ul și rata de cadre.

Capitolul 8 subliniază concluziile alături de principalele contribuții și direcțiile de dezvoltare viitoare.

4. Contribuții

Principala contribuție a acestei teze este dezvoltarea unor tehnici paralele eficiente care permit simularea interactivă a modelelor bazate pe particule pe clustere grafice. Alte contribuții originale cu impact direct asupra atingerii acestui obiectiv principal sunt:

- descrierea formală a modelelor bazate pe particule, bazată pe elementele constitutive ale acestora, care facilitează posibila extindere a acestora din punctul de vedere al scopului lor inițial, înafara sferei graficii pe calculator [2];
- descrierea unei tehnici de descompunere a domeniului pentru execuția paralelă a simulărilor bazate pe particule pe GPU și pe clustere grafice, care minimizează lungimea frontierelor și păstrează ordonarea globală a particulelor [5];
- descrierea și implementarea unei tehnici de paralelizare hibridă CPU/GPU care îmbunătățește performanțele simulării prin minimizarea timpilor de inactivitate ai CPU [4];
- descrierea și implementarea unei metode de simulare defazată care permite decuplarea execuției nucleelor GPU și a sincronizării bazate pe CPU [4];
- descrierea și implementarea unei tehnici paralele de integrare explicită Velocity Verlet în două treceri pentru clustere grafice [4] [5];
- descrierea și implementarea unei tehnici de actualizare paralelă a matricelor rare de dimensiuni mare implicate în integrarea implicită care permite actualizarea rapidă și eficientă direct în formatul Compressed Sparse Row [3];
- descrierea și implementarea arhitecturii de dezvoltare software *ParTSim*, optimizată, paralelă, distribuită, modulară și puternic scalabilă, pentru simularea scenariilor bazate pe particule pe clustere grafice [1];
- identificarea tiparelor de comunicație între modulele individuale ale arhitecturii *ParTSim* [1];
- descrierea și implementarea modului de interoperabilitate GPGPU/Pipeline grafic bazat pe OpenCL/OpenGL care evită transferul suplimentar de memorie între gazdă și dispozitiv în vederea trasării grafice a modelelor de particule [5];
- descrierea și implementarea unei tehnici de trasare paralelă „sortează-după” optimizată care utilizează un algoritm „regiune-de-interes” pentru a reduce traficul prin rețea necesar sincronizării [4];
- adaptarea tehnicilor de vizualizare și interacțiune la distanță existente în literatura de specialitate pentru simularea interactivă bazată pe particule pe clustere grafice [5];
- extinderea modelelor bazate pe particule din punctul de vedere al scopului inițial al acestora pentru a permite simularea modelelor din afara sferei graficii pe calculator, bazată pe descrierea formală a acestora [2];
- validarea și evaluarea creșterilor de performanță obținute prin utilizarea tehnicilor de paralelizare propuse pentru simularea bazată pe particule pe clustere GPU [3] [4] [5];
- validarea și evaluarea arhitecturii *ParTSim* prin efectuarea unei comparații vizuale de performanță între integrarea explicită și implicită [1];
- validarea și evaluarea simulării bazate pe particule a modelelor sociofizice [2];

5. Lista de lucrări

Articole în jurnale

- D. Copândeian, A. Sabou and D. Gorgan, „Tehnici de interacțiune utilizator cu obiecte 3D modelate prin particule”, in Revista Română de Interacțiune Om-Calculator, vol. 7, no. 1, 2014, pp. 71-88;

Articole la conferințe

- [1] Sabou and D. Gorgan, „A parallel, distributed, high-performance architecture for simulating particle-based models”, in 2014 16th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Sept 2014, in press;
- [2] Sabou and D. Gorgan, „Interactive particle-based simulation of sociophysics models”, in 2014 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), Sept 2014, pp. 411--416.
- [3] Sabou, D. Gorgan, and I. R. Peter, „Parallel implicit time integration for particle-based models on graphics clusters”, in 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 2014, pp. 336--341;
- [4] Sabou and D. Gorgan, „Physical simulation of 3d dynamical surfaces on graphics clusters”, in 2013 36th International Convention on Information Communication Technology Electronics Microelectronics (MIPRO), May 2013, pp. 292--297;
- [5] Sabou, C. Mocan, and D. Gorgan, „Particle based modeling and processing of high resolution and large textile surfaces”, in 2012 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), Sept 2012, pp. 355—360;

Prezentări la conferințe, workshop-uri și training-uri

- Sabou, „Parallel simulation of three-dimensional physically-based models”, in The Second graduate school within the EU COST Action IC0805 ``Heterogeneous computing - impact on algorithms", 3-7 June 2013, Uppsala, Sweden;
- Sabou and D. Gorgan, „Parallel Real Time Simulation of Mass-Spring Models on Graphic Cluster Architectures”, in 3rd Workshop of COST 0805 - Open Network for High-Performance Computing on Complex Environments, 17-18 April 2012, Genova, Italy;
- Sabou and D. Gorgan, „Exploring a graphic cluster based solution for real-time virtual surgery”, in 2012 International Conference on Medical Education Informatics, 6-7 April 2012, Thessaloniki, Greece;