High performance N-body simulation on computational grids

Groen, D.J.

Publication date
2010

Nederlandse Samenvatting

De zwaartekracht is van fundamenteel belang voor de vorming van sterrenhopen, sterrenstelsels en het universum als geheel. Zij zorgt ervoor dat deeltjes elkaar aantrekken en samenklonteren tot complexe structuren. Door te begrijpen hoe de zwaartekracht werkt zijn we in staat om het verleden van ons Universum te ontrafelen en voorspellingen te doen over hoe deze er in de toekomst uit zal zien. Het voorspellen van de beweging van materie in het universum is een ingewikkelde taak, en wordt ook wel het *N-body probleem* genoemd. Het N-body probleem heeft een exacte wiskundige oplossing voor systemen met 2 deeltjes, maar niet voor systemen met 3 of meer deeltjes. De bewegingen van deze N-body systemen worden daarom op numerieke wijze benaderd door N-body simulaties.

Vrijwel alle grootschalige N-body simulaties worden uitgevoerd met behulp van een enkele parallelle computer, terwijl er duizenden parallelle computers in de wereld aanwezig zijn. Dit proefschrift bevat een uitgebreid onderzoek naar de haalbaarheid van het uitvoeren van N-body simulaties over een wereldwijd netwerk van parallelle computers. We hebben hiertoe computers in verschillende werelddelen aan elkaar gekoppeld om N-body simulaties van sterrenhopen, sterrenstelsels en donkere materie mogelijk te maken. Daarnaast hebben we een tijdscomplexiteitsmodel ontwikkeld van enkele methoden die gebruikt worden om N-body systemen te simuleren. Deze modellen passen we toe om te analyseren onder welke omstandigheden N-body systemen efficiënt gesimuleerd kunnen worden op een wereldwijd netwerk van computers.

Hoofdstuk 2 bevat een analyse van N-body experimenten over een wereldwijd netwerk van GRAPEs. De GRAPE (GRAvity PipE) is een hardware component die speciaal ontwikkeld is om N-body simulaties te versnellen. We hebben een gedistribueerd systeem opgezet van drie GRAPE sites in drie verschillende continenten. Hierop hebben we een aantal simulaties van sterrenhopen uitgevoerd, en de prestaties van deze simulaties gemeten. Daarnaast hebben we een tijdscomplexiteitsmodel gemaakt van onze simulaties. De simulaties gebruikten de directe integratiemethode, waarbij de

krachtsuitwisselingen tussen alle deeltjes expliciet wordt uitgerekend. Uit onze resultaten blijkt dat de hoge responstijd van de intercontinentale netwerken een overheersende factor is in de uitvoertijd van simulaties met $\lesssim 10^4$ deeltjes. Ook de uitvoertijd van simulaties met meer dan 10^4 deeltjes wordt gedomineerd door communicatie, maar bij deze experimenten is het vooral de beschikbare bandbreedte die van groot belang is. Doordat we de communicaties tussen de GRAPE sites over regulier internet uitgevoerd hebben, en niet over een optisch netwerk, was de bandbreedte tussen sites beperkt. Op basis van onze metingen en voorspellingen concluderen we dat een N-body simulatie met de directe methode niet geschikt is om efficiënt over een wereldwijd gedistribueerd systeem uit te voeren. Wel is het mogelijk om een directe N-body berekening efficiënt op een nationaal netwerk van GRAPEs (of grafische kaarten) uit te voeren, mits de simulatie tenminste een paar miljoen deeltjes bevat.

In Hoofdstuk 3 beschrijven we een alternatieve manier om N-body simulaties over meerdere sites in een wereldwijd grid uit te voeren. We presenteren de *living simulation*, een dynamisch simulatieprogramma dat in staat is om zelfstandig te kiezen tussen verschillende N-body simulatiecodes en zich zelfstandig kan verplaatsen tussen verschillende grid sites. Om dit te bewerkstelligen hebben we de simulatie mogelijkheden gegeven om tijdelijk gebruikersrechten te krijgen voor het grid. Deze rechten worden dan door het programma gebruikt om bestanden te verplaatsen, en zichzelf te "klonen" naar een andere site. De mogelijkheid om zelfstandig te wisselen tussen N-body codes en te wisselen tussen locaties maakt het mogelijk om elke N-body code uit te voeren op de site die daarvoor het meest geschikt is. We hebben het living simulation concept toegepast om een hybride simulatie van de samensmelting van twee sterrenstelsels uit te voeren. Deze simulatie maakt gebruik van een Barnes-Hut tree N-body code met beperkte precisie op het moment dat de sterrenstelsels ver van elkaar verwijderd zijn, en stapt over naar het gebruik van een directe N-body code met hogere precisie op het moment dat de sterrenstelsels dicht bij elkaar komen. Wanneer de tree code gebruikt wordt draait de simulatie op een GPU in Nederland, maar zodra de directe code toegepast wordt verplaatst de simulatie zich naar een GRAPE in de Verenigde Staten. We hebben enkele experimenten met deze code uitgevoerd, waarbij we maximaal 65538 deeltjes gebruikt hebben. Tijdens de meeste simulaties wisselde de simulatie enkele keren tussen de verschillende sites, en was de totale overhead die door de wisselingen veroorzaakt werd minder dan 5% van de totale uitvoertijd.

Op basis van onze eerste experimenten concluderen we dat N-body simulaties efficiënter uitgevoerd kunnen worden over een wereldwijd netwerk van computers naarmate deze simulaties meer deeltjes bevatten. De tijdscomplexiteit van directe N-body simulaties schaalt echter met het kwadraat van de hoeveelheid deeltjes, waardoor simulaties met meer dan een paar miljoen deeltjes onpraktisch veel rekentijd vergen. Een alternatieve manier om N-body systemen te modelleren is met behulp van een *Tree/Particle-Mesh* (TreePM) code. Simulaties met een TreePM code hebben een lagere nauwkeurigheid, maar schalen qua tijdscomplexiteit met $O(N \log N)$, waarbij N gelijk is aan het aantal deeltjes in de simulatie. Door de lagere tijdscomplexiteit van TreePM is het mogelijk om enkele miljarden deeltjes te simuleren op een supercomputer. De TreePM methode wordt voornamelijk toegepast voor het simuleren van

grootschalige systemen met veel deeltjes, zoals kosmologische volumes bestaande uit meerdere sterrenstelsels.

In Hoofdstuk 4 doen we verslag van onze ervaringen bij het implementeren en uitvoeren van TreePM N-body simulaties over twee supercomputers. Om dit mogelijk te maken hebben we de *GreeM* simulatie code gekoppeld aan een communicatiebibliotheek. Deze bibliotheek hebben we specifiek ontwikkeld om te communiceren over lange afstand via optische netwerken. Onze experimenten zijn uitgevoerd met een IBM Power6 supercomputer in Amsterdam en een Cray-XT4 supercomputer in Tokyo, welke verbonden zijn met een 10 Gigabit/s optisch netwerk. We hebben een simulatie met 2048^3 deeltjes uit kunnen voeren over beide supercomputers. Hierbij is $\sim 90\%$ van de totale simulatietijd besteed aan berekeningen en $\sim 10\%$ aan communicaties. Volgens onze voorspellingen is het mogelijk om de simulatie op efficiënte wijze over 10 of meer supercomputers uit te voeren.

Hoofdstuk 5 bevat een uitgebreide analyse van N-body simulaties die uitgevoerd worden over een wereldwijd netwerk van supercomputers. We hebben de kosmologische code, die beschreven is in Hoofdstuk 4, uitgebreid om simulaties over 3 of meer supercomputers mogelijk te maken. De code, genaamd *SUSHI*, simuleert een plak van het kosmologisch deeltjesvolume op elke supercomputer en communiceert tussen de verschillende machines door de machines te schakelen in een ring structuur. Daarnaast is SUSHI in staat om tijdens de simulatie de werklast te herverdelen zodat de rekentijd op alle sites gelijk blijft. We hebben SUSHI getest op een nationaal netwerk van 5 computer clusters en op een wereldwijd netwerk van 4 supercomputers. Onze simulaties behalen een efficiëntie van 87% voor een simulatie met 1024^3 deeltjes over 3 supercomputers, en een efficiëntie van 73% voor een simulatie met 512^3 deeltjes over 4 supercomputers. We hebben een tijdscomplexiteitsmodel opgesteld voor SUSHI en dit toegepast om de uitvoertijd te voorspellen van onze simulaties over een groter aantal supercomputers. Aan de hand van onze voorspellingen concluderen we dat een TreePM simulatie met 2048^3 deeltjes met een raster van 256^3 mesh cellen efficiënt uitgevoerd kan worden over maximaal ~ 16 supercomputers. Ter vergelijking hebben we ook de tijdscomplexiteitsmodellen van simulaties met de tree en directe methode toegepast voor een netwerk van meerdere supercomputers. Uit deze voorspellingen blijkt dat de simulaties die een schema van tijdstappen in blokken toepassen niet efficiënt uitgevoerd kunnen worden over meerdere sites. We concluderen dat het technisch mogelijk is om meer rekenkracht te verkrijgen door een wereldwijd netwerk van supercomputers te gebruiken. Het op elkaar afstemmen van het beleid van individuele supercomputercentra, voor het reserveren en plannen van simulaties, vormt de grootste hindernis om langdurige productiesimulaties over meerdere supercomputers mogelijk te maken.

In hoofdstuk 6 presenteren we de *MPWide* communicatiebibliotheek. We hebben MPWide ontwikkeld om een efficiënte uitwisseling van berichten tussen supercomputers mogelijk te maken. MPWide is geoptimaliseerd voor gebruik met optische netwerken over lange afstanden. Het is een compact programma met weinig vereisten dat gemakkelijk geïnstalleerd kan worden op verschillende soorten supercomputers. Met MPWide is het mogelijk om meerdere parallelle MPI programma's te combineren tot een groter programma dat over meerdere sites uitgevoerd kan worden. Daarnaast is het mogelijk

om vanuit de applicatie de instellingen van individuele verbindingen aan te passen. We hebben de bibliotheek getest over een lokaal netwerk, tussen twee sites in een nationaal netwerk en tussen twee supercomputers in een internationaal netwerk. De twee supercomputers staan respectievelijk in Amsterdam (Nederland) en Helsinki (Finland). We hebben een gemiddelde snelheid van 4.8 Gigabit/s behaald op het 10 Gigabit/s gedeelde netwerk tussen Amsterdam en Helsinki. Daarnaast hebben we de bibliotheek toegepast om kosmologische simulaties over meerdere supercomputers uit te voeren.

We concluderen dat het in veel gevallen mogelijk is om N-body simulaties uit te voeren op een wereldwijd netwerk van computers. De grootste beperkende factor voor de efficiëntie van gedistribueerde experimenten is de hoge responstijd van intercontinentale netwerken. Dit maakt een wereldwijd netwerk vooral geschikt voor simulaties met relatief weinig communicatiestappen, waarin relatief grote data buffers uitgewisseld worden. De technologie is reeds aanwezig om langdurige productie simulaties over meerdere supercomputers uit te voeren. Er zullen echter aanvullende organisatorische en politieke inspanningen nodig zijn om een wereldwijd gedistribueerde supercomputer klaar te stomen voor productie.