Research Article

Haifeng Ma*

Development of a CPU-GPU heterogeneous platform based on a nonlinear parallel algorithm

https://doi.org/10.1515/nleng-2022-0027
received December 10, 2021; accepted April 24, 2022

Abstract: In order to seek a refined model analysis software platform that can balance both the computational accuracy and computational efficiency, a CPU-GPU heterogeneous platform based on a nonlinear parallel algorithm is developed. The modular design method is adopted to complete the architecture construction of structural nonlinear analysis software, clarify the basic analysis steps of nonlinear finite element problems, so as to determine the structure of the software system, conduct module division, and clarify the function, interface, and call relationship of each module. The results show that when the number of model layers is 10, the GPU is 210.5/s and the CPU is 1073.2/s, and the computational time of the GPU is significantly better, with an acceleration ratio of 5.1. For all the models, the GPU calculation time is much less than that of the CPU, and when the number of model degrees of freedom increases, the acceleration effect of the GPU becomes more obvious. Therefore, the CPU-GPU heterogeneous platform can more accurately describe the nonlinear behavior in the complex stress states of the shear walls, and is computationally efficient.

Keywords: heterogeneous platform, GPU parallel computing, nonlinear finite element analysis program

1 Introduction

With the rapid growth of engineering and scientific computing, even the current multi-core processors can not meet the required computational complexity [1]. Although supercomputers with high-performance servers or even clustered systems can meet certain computing needs, these machines are expensive for maintenance and energy consumption, and only a few large companies and research units have them. In order to meet the needs of the computing capabilities of most small and medium enterprises, research institutes, and individuals, various dedicated computing chips were created, such as Field Programmable Gate Array (FPGA) [2]. Although the computational power of these dedicated computing chips is ten times more than a general CPU, its function is fixed, it is often only possible to accelerate a particular type of algorithm. In addition, its price is more expensive, complicated, and is not suitable for personal and small laboratories. The emergence of parallel computing is the result of social development needs, more accurate weather forecast, analysis of human gene sequences, large-scale engineering system design simulation, massive data, information mining, etc. Relevant science and engineering areas require high strength calculations, earlier the serial calculation method was very hard in the face of such problems. There is a large number of intrinsic parallelities in the calculation of these applications, so these scientific puzzles can be solved by powerful computing performance provided by parallel computing [3]. The essence of parallel computing is to decompose a complex large problem into several simple small problems, which can be solved by multi-threaded computing on multiple processors, so as to solve complex problems more quickly.

Based on this, this article proposes a CPU-GPU heterogeneous platform development based on nonlinear parallel algorithm. According to the nonlinear finite element theory, it is fully understood that the basic numerical solution of the nonlinear finite element of the material and the basic numerical solution of nonlinear equations are determined to determine the structure of the software, and the function interface and the calling relationship between modules are defined.

According to the characteristics of strong CPU logical computing power, sufficient cache space, and superior GPU floating-point computing performance, task is assigned to the heterogeneous platform as shown in Figure 1. The part of calling GPU for calculation is implemented by calling

* Corresponding author: Haifeng Ma, Department of Information Engineering, Qinhuangdao Vocational and Technical College, Qinhuangdao, 066100, China, e-mail: HaifengMa823@126.com

Open Access. © 2022 Haifeng Ma, published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
cuBLAS library and cuSPARSE library, which function to complete the calculation between the vectors and the operation between matrix and vectors, respectively.

In order to shorten the development cycle and make full use of the existing open source finite element program code, the reverse engineering analysis of the large-scale general open source finite element program is carried out to master its overall architecture and functional module design, and compared with the nonlinear finite element software designed in this paper. Based on object-oriented design ideas, the details of each module are designed in C++ language, and the structural nonlinear finite element analysis problem and its numerical solution are transformed into computer programs, and based on modular methods, the structural elastoplastic analysis software architecture is completed. This method demonstrates that the CPU-GPU heterogeneous platform can more accurately describe nonlinear behavior in the complex strategy of shear walls, and with higher computational efficiency.

2 Literature review

In order to successfully complete the parallel calculation tasks, first three essential conditions are to be satisfied: first, there must be equipment with parallel computing power. Second, the task that needs to be calculated can be solved by breaking it into multiple relatively independent sub-tasks, then these independent sub-tasks can be performed in parallel to achieve parallel solutions for the target [4]. Third, in the parallel equipment, the corresponding programming environment is configured, and the corresponding algorithm is designed according to the specific problem design, and the program executed in parallel is implemented by a valid programming language, and ultimately solved in parallel. Based on GPU-based general calculations, the problem is solved, it is an important branch of parallel computing. Under the double stimulation of the user’s more visual experience and the demand for military field simulation, GPU’s computational power increased, far exceeding the current CPU’s development speed. The main responsibility of the GPU is the calculation and rendering of complex graphics. The graphical rendering process has a high degree of parallelism. The generation of the vertex, the generation of the element generation, and the generation process of the unit are almost independent, which makes it perform their calculations in parallel [5]. Given the GPU in parallelism and powerful computing power, it is desirable to handle data calculations other than graphical calculations by means of its powerful computing power, which is GPGPU (based on GPU-based general calculation). Modern computers basically consist of CPU and GPU, which bring gospel to most parallel program developers, but the relatively low price GPU can not provide powerful computing power. Hou et al. introduced the parallel tabu search algorithm (GPT) based on GPU-based PATPUL for HW/SW partitions. A single GPU core compacted in the neighborhood is proposed, which theoretically reduces the amount of GPU global memory. A kernel fusion strategy is further proposed to reduce GPT GPU global memory visits. In order to further minimize the transfer overhead of the GPT between the CPU and GPUs, an optimized transfer strategy based on the GPU-based TABU assessment is proposed, which considers that all candidates do not satisfy a given constraint. Experiments have shown that GPTS is better than tabu search, and it is competitive with other HW/SW partitioning methods. When considering a normal GPU platform, the proposed parallelization is significant [6]. A framework, supporting perceptively delayed data initialization on an integrated heterogeneous platform, is proposed by Wang et al. The framework not only includes three data initialization modes, central processor initialization, GPU initialization, and initialization and mixing, but also wisely utilizes an affinity estimation model to determine the best application initialization mode, which can optimize the initialization delay performance of the application. Design was evaluated on the NVIDIA TX2 and AGX platforms. It is shown that the proposed framework is able to accurately select data initialization patterns for a given application, thus significantly reducing the initialization delay. We envision that this delay-aware data initialization framework will be adopted in a future full version of an autonomous solution (such as Autoware) [7]. Currently, GPU achieves
high-throughput computation by running large numbers of threads. Li et al. introduced xflow, which enables streamlined execution by leveraging the hardware mechanism in the new generation of GPU. The xflow significantly reduces the cost of explicit copying and kernel startup in existing ways. As an alternative, xflow introduces a persistent operator that continuously processes data by sharing topics and establishes efficient interprocessor data channels through hardware page failures. To demonstrate its potential, two applications are also evaluated for GPU acceleration, including data encoding and OLAP queries [8]. Hybrid multi-core processors will dominate the next-generation of computing. By integrating several types of cores into a single chip, designers expect continued performance growth, while reducing reliance on raw circuit speed and reducing the power demand per unit of performance. The performance/energy tradeoff problem of a heterogeneous platform consisting of Intel multicore processors and multiple NVIDIA GPU coprocessors was investigated by Gadou et al. Using CMT-bone, an agent application designed by the University of Florida, we explored the load-balancing strategies for various combinations of CPU and GPU architectures to optimize performance, power, and energy measurements. (C) 2017 Elsevier Inc. All rights reserved [9].

3 Research methods

3.1 Modular design of the nonlinear software for a heterogeneous platform

Nonlinear finite element problems can usually be divided into three kinds: (1) material nonlinearity, structural nonlinearity of the material constitutive relationship still satisfies the assumption of infinite small displacement, (2) geometric nonlinearity, mainly shows the relationship between displacement and strain is nonlinear including large strain and geometric small deformation, and (3) boundary nonlinearity, the structural nonlinear effects of the boundary conditions, in which the contact problem is most typical. In the field of civil engineering, structural nonlinearity is generally caused by material nonlinearity. The nonlinearity mentioned in this article considers only the nonlinearity of the material. The material nonlinearity problem is a small deformation elastoplastic problem, considering only the nonlinearity of the material constitutive relation, satisfying the assumption of infinite small node displacement, and thus not considering the change in node coordinates [10]. The solution of the structural nonlinear problem by finite element method is essentially the solution of a system of equilibrium equations nonlinear with the node shift as basically unknown quantity, as shown in formula (1):

\[ K(\delta)\delta = P. \]  (1)

where \( \delta \) is the displacement vector; \( K(\delta) \) is the structural total rigid matrix; and \( P \) is the load vector.

The iterative method and incremental method are commonly used in nonlinear finite element analysis. The basic process of the iterative method is first determined to determine the initial displacement \( u^0 \), the next step \( u^{n+1} \) can be determined using the formula \( u^{n+1} = (K)^{-1}F \), where \( K^n \) needs to be determined and updated according to the unit state of each step. The basic idea of the incremental method is to specifically divide the load into several increments when performing structural nonlinear finite element analysis. Each load incremental step is formed by linear processing, solving the displacement increment of each node, the cumulative load of the current step is transferred, and the state of the unit is active, and the structural rigidity matrix is reprocessed according to the state, and the new level load increment is applied to the structure or member. Nonlinear finite element analysis basic flow is shown in Figure 2.

3.2 Nonlinear software FEM model is modular based on CPU platform architecture

The finite element domain class describes the finite element model, manages and stores the instance objects of node class, unit class, material class, section class, and load class, providing a unified interface for the information exchange between node class, unit class, material class, and analysis class. In the finite element analysis process, the information exchange includes unit class and load class to transfer the corresponding stiffness matrix and node load information to the analysis class, and return the updated node displacement information to the node class in the analysis class [11]. Also, the finite element domain class creates and initializes its own objects based on the information provided by the modeling class. When using elastoplastic analysis, the relationship between stress and strain is nonlinear, with the difficulty of the tangent modulus. In order to solve this problem, the procedure used to solve the stress increment to determine the current
stress, according to the ratio of strain increment, which is an effective algorithm to solve the material constitutive matrix \cite{12,13}. The main calculation steps are as follows:

1. Material elastic modulus calculation stress $\sigma_{\text{trial}}^{n+1}$, as shown in formula (2):

$$\sigma_{\text{trial}}^{n+1} = \sigma^n + D_{\varepsilon} \Delta \varepsilon.$$  \hspace{1cm} (2)

2. Calculate the value of the production function $f_{\text{trial}}^{n+1}$ under the test force.

3. Test yield conditions

4. If in the yield surface, $\sigma^{n+1} = \sigma_{\text{trial}}^{n+1}$, return the elastic modulus, the $\sigma^{n+1}$ and tangent modulus should be recalculated according to the constitutive relationship according to the plastic stage.

5. Update the material stress and tangential modulus, and the status variable of the material.

6. In order to achieve the application of software in the field of civil engineering, the process of concrete one-dimensional primary model under monotonous load is prepared, derived from material, inheriting the basic properties and methods of materials, and used for materials the method of material stress and tangent rigidity is overloaded \cite{14,15}. As shown in the formulas (3)–(6):

$$y = \frac{\sigma}{f_{\text{ck}}} = \frac{nx}{\alpha_c(x - 1)^2 + x^n}, \quad x > 1,$$  \hspace{1cm} (3)

$$x = \frac{\varepsilon}{\varepsilon_{\text{fk}}},$$  \hspace{1cm} (4)

$$\varepsilon_{\text{fk}} = f_{\text{ck}}^{0.54} \times 65 \times 10^{-6},$$  \hspace{1cm} (5)

$$a_t = 0.312f_{t}^2.$$  \hspace{1cm} (6)

For the compression stage: as shown in formulas (7)–(12):

$$y = \frac{nx}{n - 1 + x^n}, \quad x \leq 1,$$  \hspace{1cm} (7)

$$y = \frac{x}{\alpha_c(x - 1)^2 + x^n}, \quad x > 1,$$  \hspace{1cm} (8)

$$x = \frac{\varepsilon}{\varepsilon_{\text{ck}}},$$  \hspace{1cm} (9)

$$y = \frac{\sigma}{f_{\text{ck}}},$$  \hspace{1cm} (10)

$$\alpha_c = 0.715f_{c}^{0.785} - 0.905,$$  \hspace{1cm} (11)

$$n = \frac{E_{\text{ck}}}{E_{\text{ck}} - f_c},$$  \hspace{1cm} (12)

In the design of the nonlinear analysis program, when the strain reaches the ultimate concrete pressure strain, consider that the structure concrete is damaged, the structure reaches the ultimate bearing capacity, the unit related to the material quits, without considering the concrete softening effect \cite{16,17}.

### 3.3 Design and implementation of analytical modules based on heterogeneous platform

The key to the development of finite element computing software based on heterogeneous platform is to coordinate the CPU platform and the GPU platform work together and reasonably call the implementation of the finite element calculation process. This article designed software which is a high-efficiency structure nonlinear analysis software based on heterogeneous platforms by completing the use of CUDA’s use of the GPU. The use of CUDA first requires an anti-Weida driver for the current operating system \cite{18,19}. Second, configure the environment you need to run in the program development tool visual studio. Finally, the CUDA library is added to the GPU-based mathematical computational library cuSPARSE (sparse matrix) and cuBLAS (linear algebraic libraries), so that the CPU and GPU collaborative completion analysis calculations are achieved when the program is running \cite{20,21}. According to the CPU logic capacity, the cached space is sufficient and the GPU floating
point computational performance is superior, and the heterogeneous platform is assigned as shown in Figure 3. The part in which the cuBLAS library and the cuSPARSE library implementation are called by calling the cuBLAS library and the cuSPARSE library, and the functions are completed between the vectors and the operation between the matrices and the vector, respectively [22,23].

4 Result analysis

4.1 Based on static elastic analysis of beam and column structure of heterogeneous platform

Example 1 is a high-rise reinforced concrete frame, whose plane layout and elevation layout are shown in Figure 4. The column section is square with concrete with C40 strength grade, and the beam section is rectangular with C30 concrete. The floor constant load is 3.0 kN/m² and the live load is calculated as 2.0 kN/m². The load condition adopts 1.0 times constant load + 0.5 times live load combination. The structure is analyzed under gravity load [24,25].

The framework model is analyzed by ABAQUS and the development software, the results are shown in Table 1. The calculation results of the software are highly consistent with the ABAQUS analysis results and the error control is within 1%, verifying the correctness of the program developed in this article [26,27]. It also shows that the software is verified in the software. By analyzing the framework model under the CPU solver of the software and the heterogeneous platform-based solver, the CPU computing time is 22.7 s and the heterogeneous platform computing time is only 4.496 s. The computing efficiency of the software is significantly improved, and the efficiency of the nonlinear software based on the heterogeneous platform structure is initially verified. To further verify the computing efficiency of the software, more example analysis is needed in the later stage [28,29].

Example 2 is a multi-storey reinforced concrete frame with a height of 3.6 m with 6 layers, and the structure layout is shown in Figure 5. Column section is square, area 0.25 m², beam section width is 250 mm, and height is 600 mm. The floor constant load is 5.0 kN/m² and the live load is calculated as 2.0 kN/m². The characteristic period is 0.35 s, structure is 0.912 s. C30 concrete and grade HRB33 reinforcement were used. Statistic elastic-plastic analysis of the frame structure under gravity load and horizontal earthquake [30,31] were performed.

<table>
<thead>
<tr>
<th>Floor number</th>
<th>SOR-PCG (mm)</th>
<th>ABAQUS (mm)</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.80774</td>
<td>-0.80142</td>
<td>0.788</td>
</tr>
<tr>
<td>2</td>
<td>-1.53207</td>
<td>-1.51839</td>
<td>0.901</td>
</tr>
<tr>
<td>3</td>
<td>-2.17417</td>
<td>-2.15399</td>
<td>0.937</td>
</tr>
<tr>
<td>4</td>
<td>-2.73468</td>
<td>-2.71007</td>
<td>0.908</td>
</tr>
<tr>
<td>5</td>
<td>-3.21418</td>
<td>-3.18456</td>
<td>0.930</td>
</tr>
<tr>
<td>6</td>
<td>-3.61317</td>
<td>-3.57718</td>
<td>1.006</td>
</tr>
<tr>
<td>7</td>
<td>-3.93204</td>
<td>-3.89241</td>
<td>0.972</td>
</tr>
<tr>
<td>8</td>
<td>-4.17113</td>
<td>-4.12959</td>
<td>1.006</td>
</tr>
<tr>
<td>9</td>
<td>-4.33062</td>
<td>-4.28749</td>
<td>1.006</td>
</tr>
<tr>
<td>10</td>
<td>-4.44112</td>
<td>-4.39915</td>
<td>0.954</td>
</tr>
</tbody>
</table>
Established model and the generated data file, and the ABAQUS and this software are shown in Table 2 and Figure 6. From the error between the figure and ABAQUS analysis, we prove the accuracy of the software in structural static nonlinear calculation [32,33].

In order to test the effectiveness of high-performance computing software, this article analyzes the high-level framework model with the same structure layout using the CPU platform and the heterogeneous platform based software, and calculates the calculation time required for the same model based on the two platforms [18,33]. The results are shown in Table 3. It is seen from Table 3 that when the number of model layers is 10, GPU is 210.5/s, CPU is 1073.2/s, and GPU’s computational time is significantly superior, with an acceleration ratio of 5.1. For all models, GPU computation time is much less than CPU, especially when the acceleration effect of GPU becomes more obvious, the number of model degrees of freedom increases [34,35].

### 5 Conclusion

Based on the combination of CPU serial computing and GPU high-performance parallel computing, we established a heterogeneous platform of CPU-GPU hybrid programming, proposed the CPU-GPU heterogeneous platform development based on nonlinear parallel algorithm, conducted program implementation, and successfully added to the solution library, accelerating the solution speed of the software. Using multiple parallel optimization strategies suitable for GPU computing, we improved the computational efficiency of large sparse linear equations through the elastic analysis of the solver based on CPU, and verified the reliability and effectiveness of the heterogeneous platform. The results show that the spatial shell element established by CPU-GPU heterogeneous platform can accurately describe the nonlinear behavior in the complex stress state of shear wall and have high computational efficiency. Although this article is based on CPU-GPU heterogeneous platform for building structure elastically, plastic dynamic time range reaction refined numerical model and numerical algorithm made more detailed research and made some research results, but based on CPU-GPU heterogeneous platform parallel software development, need to combine the latest computer technology, finite element theory, algorithm research, software tools, is a long-term and persistent work.
**Funding information:** The author states no funding involved.

**Author contributions:** The author has accepted responsibility for the entire content of this manuscript and approved its submission.

**Conflict of interest:** The author states no conflict of interest.

**References**


Liu Y, Chen M, Xu X. Forest rainfall characteristics based on heterogeneous computing and influencing factors of athletes’ physical supplement. Arab J Geosci. 2021;14(15):1516.

