Multi-GPU Computing with Abaqus: Benchmarking and scaling for multiphysics applications in mechatronics

Aaron D. Price

ABB Corporate Research Center Germany
Wallstadter Str. 59
68526 Ladenburg, Germany
aaron.price@de.abb.com

Abstract: Mechatronic systems encountered in the power and automation industries exhibit very complex behavior for a variety of applications. These problems require solutions derived from diverse physical phenomena, and hence are considered to be multiphysics problems. One such problem includes computing the coupled electromechanical response of electroactive polymer actuators. Due to the complex nature of the problem, obtaining the transient response of electroactive polymer actuator-enabled mechanical systems is a computationally expensive task. In this study, we investigate how multi-GPU acceleration using Abaqus/Standard 6.12 may improve the overall simulation speed. Benchmark calculations are performed on a 16-core workstation with 3 GPUs. This approach is applied to two classes of problem size in order to elucidate scaling effects. As a result, some best practice guidelines will be given for efficient use of GPU acceleration in the given context.

Keywords: GPU acceleration, dielectric elastomer actuator, UMAT, user subroutine, multiphysics, high performance computing.

1. Introduction

The benefits of workstations equipped with a single general-purpose graphics processing unit (GPGPU, herein referred to simply as GPU) as applied for the acceleration of finite element analysis using Abaqus 6.11 have been reported (Shyu, 2012). The release of Abaqus 6.12 introduced the ability to utilize multiple GPUs on a single workstation, and thus a new and largely unexplored opportunity exists to accelerate simulations to an even higher degree. In this paper, we quantify the performance gains of the multi-GPU approach against an equivalent computation without GPU acceleration for the multiphysics simulation of dielectric elastomer actuators.

The remainder of this section is devoted to the background and introductory theory governing dielectric elastomer actuators. Section 2 provides an overview of the modeling approach implemented to capture the non-linear electromechanical response of the actuator. Finally, Sections 3 and 4 discuss simulation results and the impact of multiple GPU acceleration respectively.

1.1 Dielectric elastomer actuators

A variety of dielectric elastomer actuator configurations have been proposed. Dielectric elastomer stacks are a particularly promising class of electroactive polymer actuator due to their uniquely favorable specific work densities in terms of actuator output force and stroke capabilities. These performance characteristics are highly dependent on many interrelated factors including layer geometry, mechanical and electrical properties of the electrode and dielectric layers, driving voltage, and the ancillary electrical system and interconnection behavior. Furthermore, these considerations are often made more complex by varying environmental factors encountered in service. Thus, the specification of an optimal actuator design remains a challenging task. Simulation of the transient electromechanical actuation response of stacked dielectric elastomers in three dimensions is an intensive and time consuming process, and hence these problems are ideal candidates for GPU acceleration.

1.2 Stacked actuator configurations

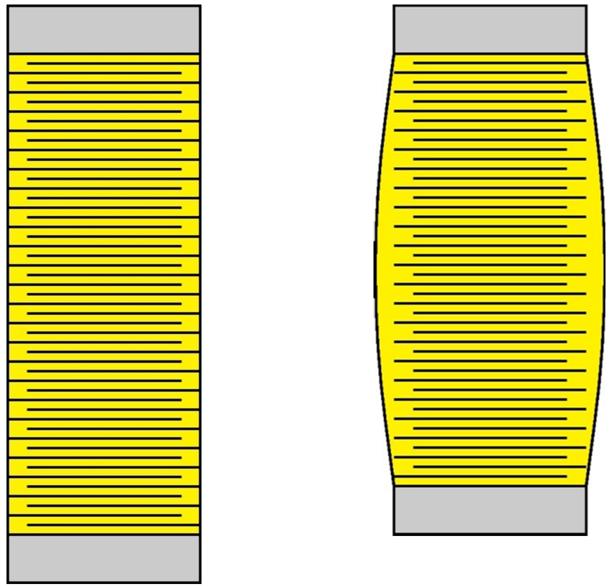
In general, the electroactive response of dielectric elastomer actuators is primarily attributed to electrostatic stresses defined by the Maxwell pressure, p, across the membrane:

$$p = \epsilon_r \epsilon_0 E^2 \tag{1}$$

where ϵ_r is the material relative permittivity (dielectric constant), ϵ_0 is the vacuum permittivity, and E is the electric field strength (Suo, 2010, Pelrine, 1998). This stress can be related to mechanical deformations by the application of a suitable elastic or hyper-elastic material law. Several such material models have been applied including the Hookean, neo-Hookean, Ogden, Yeoh, Mooney-Rivlin, and others (Brochu, 2010). Additional methods have been reported to modify the response equation in the case of large deformations and to obtain derivations from an energy balance standpoint, which is particularly beneficial for the incorporation of boundary conditions (Suo, 2010, Goulbourne, 2007, Wissler, 2007).

As illustrated in Figure 1, stacked dielectric elastomer actuators consist of multiple layers of dielectric elastomer membranes stacked or folded on top of each other and terminated by relatively rigid end-cap fixtures (Carpi, 2007, Kovacs, 2009a, Kovacs, 2009b). This configuration has been shown to exhibit a favorable balance of force and stroke (work densities approaching 13 J/kg) relative to alternative dielectric elastomer actuator configurations, and is therefore the motivation for ongoing research activities. Unfortunately, in their current state these actuators require kilovolt-level input potentials that inhibit their wide-spread commercial adoption (primarily due to the cost of high-voltage power supplies and the associated safety concerns). Hence, optimized stacked actuator design specifications using improved elastomer materials and geometries are currently sought.

An important consequence of Equation 1 is that the Maxwell pressure is proportional to the square of the electric field, thus the induced stress for a given input voltage is maximized by reducing the film thickness (a geometric design consideration). Additionally, the stress can be increased linearly by increasing the dielectric constant of the elastomer (a material science consideration). Special attention must be paid that increases in the dielectric constant imparted through the introduction of highly polar particle additives to the elastomer does not overly stiffen the material, as the resulting loss in compliance will be detrimental to the actuator stroke. Alternatively, these



(a) Relaxed (b) Energized

Figure 1. A simplified representation of a dielectric elastomer stack.

adjustments fundamentally suggest that equivalent actuation performance can be obtained at much-reduced voltages if the film thickness is reduced, the dielectric constant is increased, and the stiffness is decreased; thereby mitigating the necessity for a high-voltage power source and its associated complications.

2. Overview of modeling approach

In order to assess the impact of various stack geometries and material properties on the dynamic performance of a stacked electroactive polymer stack, a multiphysics finite element model is proposed. This section describes how the dielectric elastomer stack model was implemented using Abaqus/Standard 6.12.

2.1 Neo-Hookean material model

The elastomer material is modelled as a neo-Hookean hyperelastic material (Boyce, 2000, Charlton, 1994). Material properties have been prescribed in order to capture typical performance of the VHB 4910 elastomer manufactured by 3M commonly used by the electroactive polymer research community. Estimates of the relevant parameters are defined as follows: density, ρ, of 1000 kg/m^3, neo-Hookean strain energy potential parameters, C_{10} of 25 kPa and D_1 of 8×10^{-7} Pa^{-1} and respectively, and finally a dielectric permittivity, ϵ, of 4.16×10^{-11} corresponding to a dielectric constant, ϵ_r, of 4.7 (Abaqus Theory Manual, 2012). Viscoelastic effects have not yet been incorporated in the model.

2.2 User subroutine capturing electroactive response

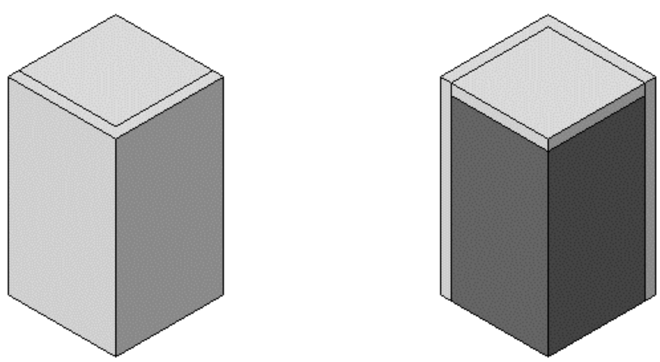
The multiphysics aspect of the simulation that couples the Maxwell pressure to the hyperelastic deformation of the elastomer is implemented using modified free-energy functions for the neo-Hookean finite-strain model (Qu, 2012, Zhao, 2008). With this approach, a custom written User Material (UMAT) subroutine written in FORTRAN is called from the Abaqus model code. For the model under consideration herein, the nominal electric field is prescribed solely in the vertical direction and the resulting Maxwell pressures and deformation gradients are used to determine the true electric field and corresponding material deformations.

2.3 Part geometry and meshing

The actuator geometry consists of a rectangular prism representing a complete stack with a square cross-sectional side length of 2.0 cm and overall height of 3.5 cm. Using symmetry planes, a section of the geometry is modeled by a prism of 1.0 cm side half-length, and half-height of 1.75 cm to reduce redundant computational effort as illustrated in Figure 2.

The prism is partitioned into an active core region within which the material responds to the applied electric field, and a passive exterior region consisting of a material of equal compliance that does not respond to the electric field. In practice, this arrangement arises due to the requirement to encapsulate the stack for safety, and to realize the alternating arrangement of inter-layer electrodes.

The part is initially meshed using a global seed size of 0.5 mm resulting in 14 000 elements of type C3D20R (continuum stress-displacement, 20-node quadratic brick, reduced integration) for a total of 61 551 nodes and 184 653 variables. The effects of mesh density on simulation acceleration are discussed in Section 4.



(a) Isometric view (b) Reverse isometric

Figure 2. Part geometry indicating the passive outer layers and active polymer core. Symmetries are exploited to model only 1/8th of the total stack.

2.4 Boundary conditions

Symmetry boundary conditions are used to capture the complete stack behavior using a $1/8^{th}$ symmetric section by prescribing XSYMM, YSYMM, and ZSYMM boundary conditions on the corresponding cut planes.

In practice, the stack is coupled to the actuated load via relatively rigid end-caps adhered to the stack end-faces which are constrained to move only in the vertical direction. These boundary conditions are idealized in the model via fixed horizontal displacement constraints applied to the nodes of the end-cap interface in combination with a constraint equation that enforces uniform vertical displacements.

2.5 Step definition and applied loads

The model consists of two dynamic-implicit simulation steps:

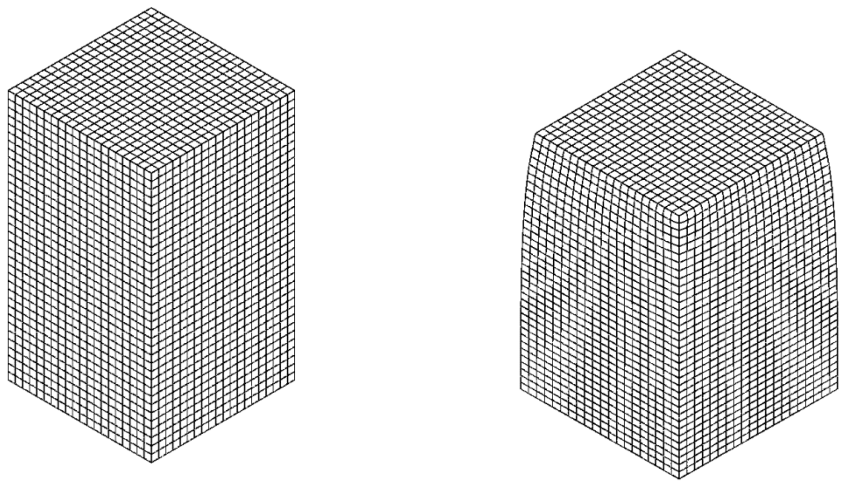
- Step 1 consists of an optional pre-stretching mechanism typically employed for thin elastomer membranes. In the current stack configuration, pre-strains are not explicitly prescribed and thus the step serves only for equilibration. The step duration is 0.5 s.
- Step 2 defines the electroactive response of the material, wherein the electric field is gradually applied using a smoothed ramp amplitude. The step duration is 1.1 s. An applied electric field of approximately 22.9 MV/m is prescribed corresponding to an applied voltage of 800 V for a polymer layer thickness of 35 µm.

3. Simulation results

Figure 3 illustrates the deformed mesh as a result of the induced Maxwell stresses arising between the electrode layers in the stacked actuator. Figure 4 indicates the typical compressive stresses arising in the vertical direction within the active stack at full deflection are approximately 10 MPa. In this case the electrostatic stresses are balanced by the internal deformation stresses within the elastomer.

The history output of the vertical nodal displacement of the end-cap reference node indicates a vertical contraction of 2.6 mm, corresponding to an overall actuator stroke of 5.2 mm (Figure 4). This displacement corresponds to an effective actuator contractile strain of 14.8% under zero external load. Note the difference in applied electric field and displacement curves due to the field-squared dependence introduced in Section 1.

An important limitation of the current version of the model is the absence of viscoelastic effects, which are known to be significant at high actuation frequencies.



(a) Relaxed (b) Energized

Figure 3. Meshed dielectric elastomer stack in the (a) relaxed state and the (b) deformed state in response to an applied voltage of 800 V.

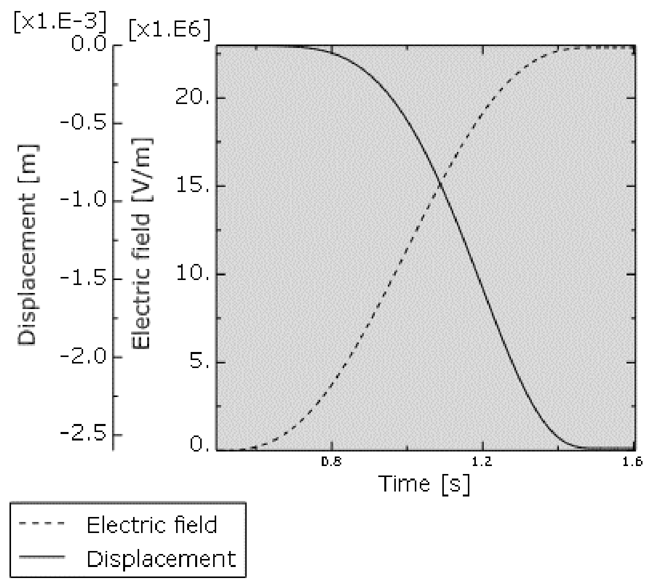


Figure 4. Time history outputs indicating the evolution of total actuator stroke in response to the smoothly applied electric field.

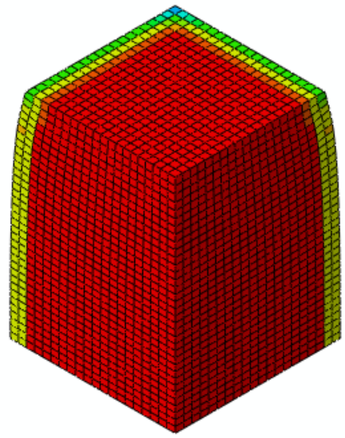


Figure 5. Deformed mesh indicating the principle stress in the vertical direction. Note the stress differences between the electroactive core and passive exterior.

4. Impact of multiple GPU acceleration

4.1 Workstation configuration

To investigate the impact of multiple GPUs on simulation acceleration, the solver was executed on a custom-built high-performance workstation operating on openSuse 12.1. Processing capability is provided via two Intel Xeon Sandy Bridge EP 8 core CPUs (16 cores total) at 3.1 GHz which are supplemented with three Nvidia Tesla C2075 GPUs (in addition to an Nvidia GTX 550 Ti GPU dedicated solely to graphics output which is not employed for GPGPU acceleration). A total of 256 GB of RAM is installed to reduce swapping operations during the computation of large jobs and to facilitate multiple simultaneous users. Finally, several hard drives are installed in a high-speed RAID 0 configuration with 3.5 TB of total capacity.

4.2 Comparison of hardware acceleration configurations

This section presents a comparison of simulation times using the model of the dielectric elastomer actuator with various quantities of CPUs and GPUs enabled. Two mesh densities are considered as separate cases to observe the scaling effects. In each case, the wallclock solution time, required number of increments (since a dynamic analysis is being performed some variation in the number of increments required is observed, and should be considered when comparing hardware acceleration performance), the tokens required for a given hardware configuration, a metric deemed herein as *token efficiency* which is determined as the quotient of solution time and token requirement (a lower figure is better), and finally a *speedup factor* which relates the solution time to the base case of a single CPU with GPU acceleration fully disabled.

The first case examines a relatively coarse mesh consisting of 1800 elements with 8657 nodes for a total number of variables of 25 971. Figures 6–8 indicate a summary of the simulation

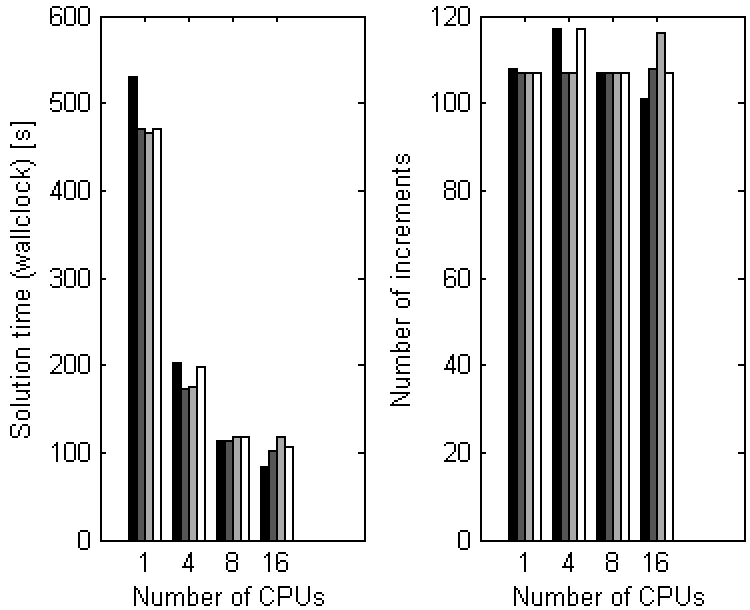


Figure 6. Solution time and required number of increments as a function of number of CPUs and GPUs employed. Impact of GPUs is shown within each group ranging from 0 (black) to 3 (white) GPUs enabled.

acceleration results for the coarse mesh. The results indicate that for this relatively small problem, performance gains are most appreciable when the number of CPU cores is small and when only one additional GPU is enabled. In some cases, enabling the GPU hardware reduces the speedup factor due to the additional overhead associated with the parallelization and recombination of the problem. In these cases, the GPU is largely underutilized and thus there is a net impediment on the overall solution time. Hence, for this small problem the speedup factor of GPU-enabled tests on a per-token basis is consistently outperformed by the CPU-only baseline.

The second case examines a relatively fine mesh consisting of 112 630 elements with 472 448 nodes for a total number of variables of 1 417 344. In these tests, the memory requirement is approximately two orders of magnitude higher relative to the coarse mesh. Figures 9–11 indicate a summary of the simulation acceleration results for the fine mesh. In this case, the impact of GPU acceleration is significantly more prevalent, though under certain conditions with many processors utilized the addition of a single GPU impeded the solution time. In the case of the single CPU calculation, the addition of a single GPU improved the performance by more than a factor of 5 while enabling a second or third GPU did not improve the acceleration performance.

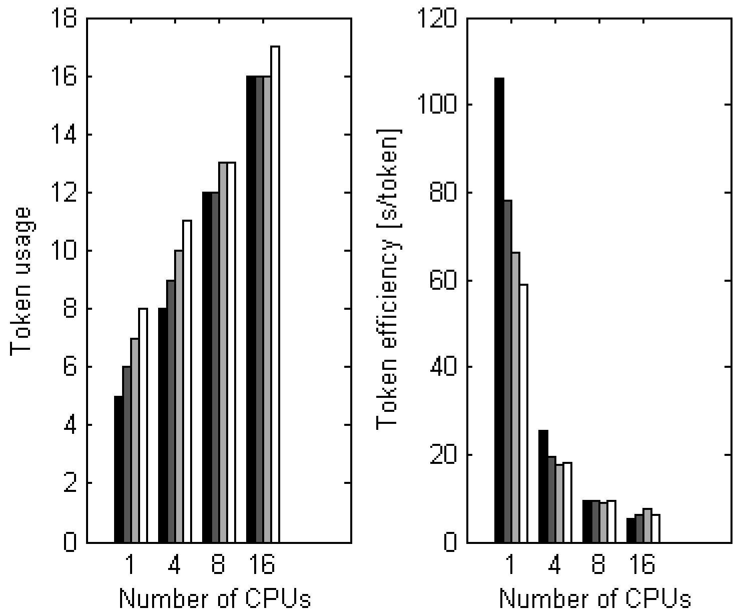


Figure 7. Token usage and token efficiency as a function of hardware configuration.

The Nvidia System Management Interface program was invoked to monitor the memory and GPU utilization, and in this case (1 CPU, one or more GPUs) it was found that in general only one of enabled GPUs could be fully utilized at a given time (alternating between all available GPU devices). It is suggested that the lone CPU cannot sufficiently cope with the operations required to support multiple GPUs simultaneously, and hence the performance of two or three GPUs is essentially reduced to the response of a single GPU-enabled system. Conversely, in the case of 4 or more CPU cores the GPU utilization was at times reported to simultaneously exceed 95% for all three GPUs.

Unlike the trend observed for the small problem case, for the large problem appreciable gains in speedup factor are generally observed when multiple GPUs are employed. Figure 11 also indicates that in certain cases the best speedup factor does not always correlate with expending the most license tokens, however this behavior is highly problem dependent. Hence, when considering hardware acceleration opportunities for new simulation problems, it is recommended to perform some initial tests to determine what combination of hardware provides the best balance of performance and license utilization. Periodic online monitoring using the Nvidia System Management Interface is also recommended to capture the utilization of each GPU in order to actively manage the acceleration hardware configuration with respect to memory and GPU utilization.

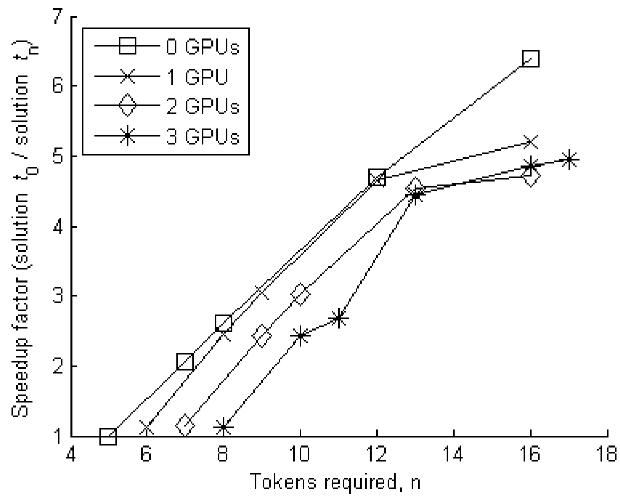


Figure 8. Speedup factor as a function of token usage for the small problem. Reference solution is 1 CPU, 0 GPUs.

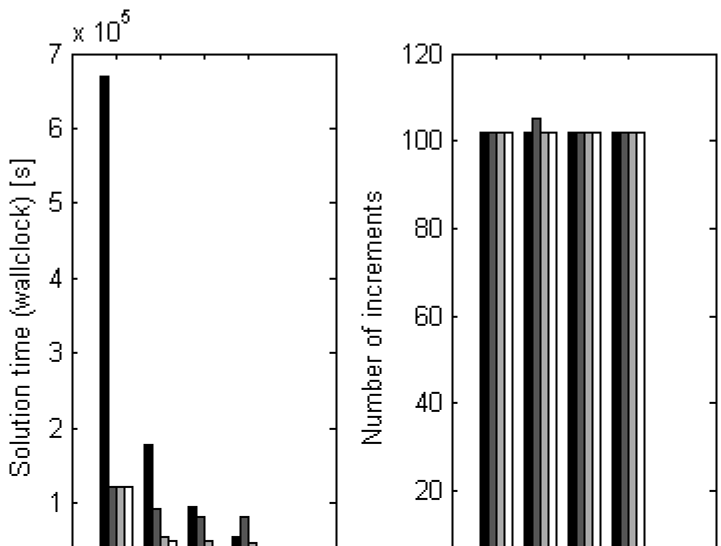


Figure 9. Solution time and required number of increments as a function of number of CPUs and GPUs employed.

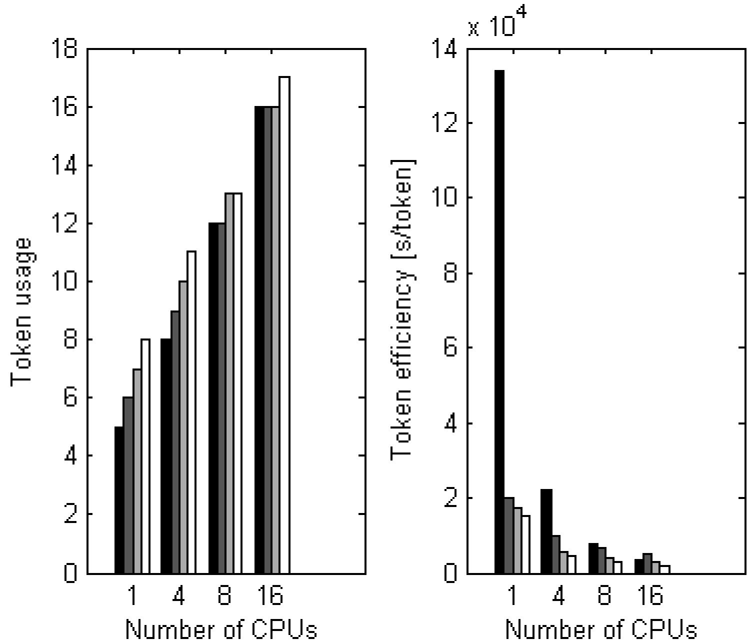


Figure 10. Token usage and token efficiency as a function of number of CPUs and GPUs employed.

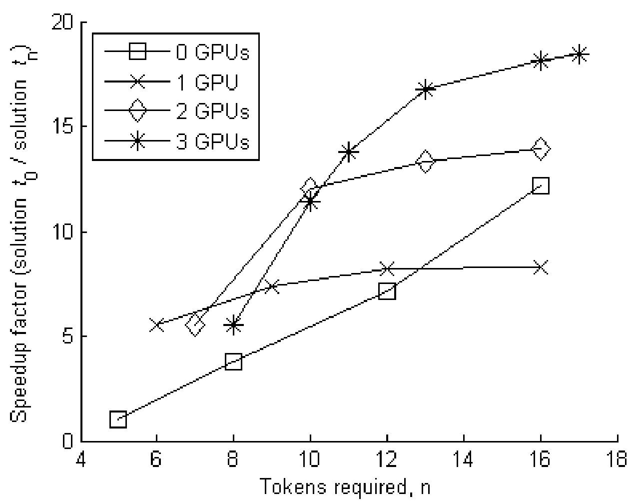


Figure 11. Speedup factor as a function of token usage for various numbers of GPUs enabled. The reference solution time is 1 CPU and 0 GPUs.

5. Conclusion

This study investigated performance acceleration and scaling considerations associated with multi-GPU computing using Abaqus. Benchmarking tests were performed in the context of a non-linear multiphysics model of a dielectric elastomer actuator system and simulation acceleration results were characterized according to solution time, token efficiency, and speedup factor. These outcomes confirm that multiple GPU acceleration may be employed to greatly reduce solution times for relatively large problems, in particular when only a few CPU cores are available or when license token availability is limited.

6. Acknowledgments

Support for this activity is provided in part within the project "DIELASTAR - Dielektrische Elastomere für Stellaktoren" (Dielectric Elastomer Actuators), funded by the Federal Ministry of Education and Research (BMBF) of Germany under grant number 03X4011. Technical discussions providing valuable inputs for this activity were facilitated by the European Cooperation in Science and Technology (COST) in the framework of ESNAM (European Scientific Network for Artificial Muscles, COST Action MP1003). Special gratitude is extended to Dr. Jörg Gebhardt of ABB Corporate Research for providing insightful direction on the execution of these experiments.

7. References

Abaqus Theory Manual, §4.6.1 Hyperelastic material behavior, Version 6.12-3, Dassault Systémes Simulia Corp., Providence, RI (2012).

Boyce, M. and Arruda, E., "Constitutive models of rubber elasticity: A review," *Rubber Chemistry and Technology* 73(3), pp. 504–523 (2000).

Brochu, P. and Pei, Q., "Advances in dielectric elastomers for actuators and artificial muscles," *Macromolecular rapid communications* 31(1), 10-36 (2010).

Carpi, F. and De Rossi, D., "Contractile folded dielectric elastomer actuators," in Electroactive Polymer Actuators and Devices (EAPAD) 2007, Y. Bar-Cohen, ed., *Proceedings of SPIE - The International Society for Optical Engineering* 6524, p. 65240D, SPIE Press, Bellingham, Washington (2007).

Charlton, D., Yang, J. and Teh, K., "Review of methods to characterize rubber elastic behavior for use in finite element analysis," *Rubber Chemistry and Technology* 67(3), pp. 481–503 (1994).

Goulbourne, N., Mockensturm, E. and Frecker, M., "Electro-elastomers: Large deformation analysis of silicone membranes," *International Journal of Solids and Structures* 44(9), pp. 2609–2626 (2007).

Kovacs, G. and Düring, L., "Contractive tension force stack actuator based on soft dielectric EAP," in *Proceedings of SPIE - The International Society for Optical Engineering* 7287, 72870A (2009a).

Kovacs, G., Düring, L., Michel, S. and Terrasi, G., "Stacked dielectric elastomer actuator for tensile force transmission," *Sensors and Actuators A: Physical* 155(2), 299-307 (2009b).

Pelrine, R., Kornbluh, R. and Joseph, J., "Electrostriction of polymer dielectrics with compliant electrodes as a means of actuation," *Sensors and Actuators A: Physical* 64(1), 77-85 (1998).

Qu, S. and Suo, Z., "A finite element method for dielectric elastomer transducers," *Acta Mechanica Solida Sinica* 25(5), 459-466 (2012).

Shyu, G. and Deng, G., "GPU Computing with Abaqus 6.11 and its Application in Oil & Gas Industry," *SIMULIA Community Conference*, Providence, Rhode Island, (2012).

Suo, Z., "Theory of dielectric elastomers," *Acta Mechanica Solida Sinica* 23(6), 549-578 (2010).

Wissler, M. and Mazza, E., "Electromechanical coupling in dielectric elastomer actuators," *Sensors and Actuators, A: Physical* 138(2), pp. 384–393 (2007).

Zhao, X. and Suo, Z., "Method to analyze programmable deformation of dielectric elastomer layers," *Applied Physics Letters* 93(25), 251902 (2008).