The implementation of an edge-based three-dimensional Reynolds Average Navier–Stokes solver for unstructured grids able to run on multiple graphics processing units (GPUs) is presented. Loops over edges, which are the most time-consuming part of the solver, have been written to exploit the massively parallel capabilities of GPUs. Non-blocking communications between parallel processes and between the GPU and the central processor unit (CPU) have been used to enhance code scalability. The code is written using a mixture of C++ and OpenCL, to allow the execution of the source code on GPUs. The Message Passage Interface (MPI) library is used to allow the parallel execution of the solver on multiple GPUs. A comparative study of the solver parallel performance is carried out using a cluster of CPUs and another of GPUs. It is shown that a single GPU is up to 64 times faster than a single CPU core. The parallel scalability of the solver is mainly degraded due to the loss of computing efficiency of the GPU when the size of the case decreases. However, for large enough grid sizes, the scalability is strongly improved. A cluster featuring commodity GPUs and a high bandwidth network is ten times less costly and consumes 33% less energy than a CPU-based cluster with an equivalent computational power.