A Parallel Approach for 3D-Variational Data Assimilation on GPUs in Ocean Circulation Models

Rossella Arcucci, Luisa D’Amore, Simone Celestino, Giuseppe Scotti, Giuliano Laccetti

Abstract—This work is the first dowel in a rather wide research activity in collaboration with Euro Mediterranean Center for Climate Changes, aimed at introducing scalable approaches in Ocean Circulation Models. We discuss designing and implementation of a parallel algorithm for solving the Variational Data Assimilation (DA) problem on Graphics Processing Units (GPUs). The algorithm is based on the fully scalable 3DVar DA model, previously proposed by the authors, which uses a Domain Decomposition approach (we refer to this model as the DD-DA model). We proceed with an incremental porting process consisting of 3 distinct stages: requirements and source code analysis, incremental development of CUDA kernels, testing and optimization. Experiments confirm the theoretic performance analysis based on the so-called scale up factor demonstrating that the DD-DA model can be suitably mapped on GPU architectures.

Keywords—Data Assimilation, Parallel Algorithm, GPU architectures, Ocean Models.

I. INTRODUCTION

DATA Assimilation (DA) is an Uncertainty Quantification technique widely used in simulation science to incorporate observational data into a prediction model [15]. Due to the scale of the forecasting area and the number of state variables used to describe ocean or atmosphere for climate or weather predictions, DA are large scale problems that should be solved in near real-time. During the last 20 years, parallel algorithms for data assimilations reached a widespread interests at many federal research institutes as well as at many universities [NCAR (National Center for Atmospheric Research), NCEP (National Centers for Environmental Prediction), DWD (Deutscher Wetterdienst), UK Met Office, JMA (Japan Meteorological Agency), CMC (Canadian Association of Management Consultants) and the CMCC (Euro-Mediterranean Center for Climate Changes)]. The CMCC makes use of a 3D Variational (3DVar) DA software, called OceanVar, for assimilating data in Mediterranean Forecasting System (MFS) context [10], [3]. MFS is a daily 10-day forecast system in operational use since 1998, and its ocean general circulation model (OGCM) is based on the Ocean Parallelise (OPA) code, which has subsequently been set up for the Mediterranean Sea (NEMO framework) [17].

Together with University of Naples Federico II, CMCC has developed a fully scalable 3DVar DA model which is based on a Domain Decomposition approach (called DD-DA model) [5], [4]. The resulted parallel algorithm consists of several copies of a slightly modified 3D-Var algorithm, each one requiring approximately the same amount of computations on each sub domain and an exchange of boundary conditions between adjacent sub domains. Data flow across the surfaces and a so-called surface-to-volume effect is produced [1].

Over the last few years, the rapid evolution of Graphics Processing Units (GPUs) into powerful, cost-efficient, programmable computing architectures for general purpose computations has provided application potential beyond the primary purpose of graphics processing. As the number of supercomputers equip GPUs is massively increasing [19], large scale problems are embracing GPUs for massive thread level parallelism. GPUs have enjoyed rapid adoption within the high-performance computing (HPC) community because GPUs enable high levels of fine-grain data parallelism. The latest GPU programming interfaces such as NVIDIA’s Compute Unified Device Architecture (CUDA) [14], and more recently Open Computing Language (OpenCL) [16] provide the programmer a flexible model while exposing enough of the hardware for optimization. GPU clusters, where fast network connected compute-nodes are augmented with latest GPUs, [18] are now being used to solve challenging problems from various domains. These new systems are designed for high performance as well as high power efficiency, which is a crucial factor in future exascale computing.

However, GPU architecture is unlike that of any other, and designing algorithms to fully harness the capabilities of a GPU is not a straightforward task, especially when one considers the advantages and disadvantages of the various resources that a GPU has available to it. To best utilize the computing capabilities provided by the graphic processors, it is highly desired to study how to map algorithms and programs on them. Briefly, the goal is to reduce the total data transfer time as much as possible, meaning reducing the amount of data that is transferred back and forth between host (the CPU) and device (the GPU).

In this article we describe how DD-DA model is well-suited for efficiently using GPU architecture. The paper is organized as follows. In Section II our parallel approach is presented. The mathematical model we implement is reviewed in Section II-A. In Section II-B a brief overview of the GPU architecture and some programming basics required to designing algorithms to fully harness the capabilities of a GPU is presented. In Section II-C shows the implementative strategy we used for efficiently using GPU architecture to develope DD-DA model. In Section III...
we present selected numerical results to demonstrate the effectiveness of the GPU-based parallel DD-DA algorithm. Section IV concludes the paper and outlines possible future work.

II. DOMAIN DECOMPOSITION DA MODEL FOR GPUs

The DD-DA model is based on a Domain Decomposition approach for solving Variational Data Assimilation problem [5]. The model uses a partition of the global domain into sub domains. On these sub domains we define local 3D-Var functionals and we prove that the minimum of the global 3D-Var functional can be obtained by collecting the minimum of each local functional. The (global) problem is decomposed into (local) sub problems in such a way. The resulted algorithm consists of several copies of a slightly modified 3D-Var algorithm, each one requiring approximately the same amount of computations on each sub domain and an exchange of boundary conditions between adjacent sub domains.

Fig. 1 shows a simple example of how the DD-DA model works on a decomposition of the global domain in six subdomains. Red points represent the observed data which are distributed geographically as the physical subdomains. Green lines represent the overlapping regions between different physical subdomains. With this decomposition, local DA problem are solved concurrently, each subdomain is processed on a processor node of the supercomputer, which make the resulted algorithm, each one requiring approximately the same amount of computations on each sub domain and an exchange of boundary conditions between adjacent sub domains.

A. DD-DA computational model

Let \( t_k \), \( k = 0,1,\ldots,n \) be a sequence of observation times and, for each \( k \), let \( x_{M_k} \equiv x_M(t_k) \in \mathbb{R}^N \) be the vector denoting the state of the Mediterranean sea system at time \( t_k \) as defined in (1) where \( T \) is the three-dimensional temperature field, \( S \) the three-dimensional salinity field, \( \eta \) the two-dimensional free surface elevation, and \( u,v \) are the total horizontal velocity components and where with \( * \) we denote the vector transposed.

\[
x_{M_k} = [T,S,\eta,u,v]^* \quad (1)
\]

At each time step \( t_k \), let \( y_k \) be the observations vector as defined in (2) where \( H_k : \mathbb{R}^N \mapsto \mathbb{R}^m \) is a non-linear operator collecting the observations at time \( t_k \) [3].

\[
y_k = H_k(x_k) \quad (2)
\]

Let (3) be an overlapping decomposition of the physical domain \( \Omega \) such that \( \Omega_i \cap \Omega_j \neq \emptyset \) if \( \Omega_i \) and \( \Omega_j \) are adjacent and \( \Omega_{ij} \) is called overlapping region.

\[
\Omega = \bigcup_{i=1}^{N_{sub}} \Omega_i \quad (3)
\]

According to this decomposition the DD-DA computational model is a system of \( N_{sub} \) non-linear least square problems [10], [5] described in (4)-(6) where \( J_i \) in (6) is a cost-function.

\[
x_{DA}(t_k) = \sum_{i=1}^{N_{sub}} \tilde{x}_{DA_i}(t_k) \quad (4)
\]

\[
\tilde{x}_{DA_i} = \left\{ \begin{array}{ll}
\operatorname{argmin}_{x_k} J_i(x_k) & \text{on } \Omega_i \\
0 & \text{on } \Omega - \Omega_i 
\end{array} \right. \quad (5)
\]

\[
J_i(x_k) = \left\| H_{i,k}(x_{k_i}) - y_{k_i} \right\|_{R_i}^2 + \left\| x_{k_i} - x_{M_k} \right\|_{B_i}^2 + \left\| x_{k_i/\Omega_{ij} - x_{k_j/\Omega_{ij}} \right\|_{B_{ij}}^2 \quad (6)
\]

\( x_{DA} \) in (4) is the analysis (i.e. the estimation of the vector \( x_{k_i} \) at time \( t_k \)). The variables \( x_{M_k} \) and \( y_{k_i} \) in (6) are the same vectors in (1) and (2) defined on the subdomain \( \Omega_i \), \( R_i \) and \( B_i \) are the covariance matrices defined in (7) and (8), whose elements provide the estimate of the errors on \( y_{k_i} \) and on \( x_{M_k} \), respectively. Also the variables \( x_{k_i/\Omega_{ij}} \), \( x_{k_j/\Omega_{ij}} \), \( B_{ij} \) are the restriction on \( \Omega_{ij} \) of these quantities.

\[
R_i = \sigma_o^2 I_p \quad (7)
\]

where \( \sigma_o^2 \) is the observational error variance.

\[
B_i = \sigma_b^2 C \quad (8)
\]

where

\[
c_{ij} = \rho^{|i-j|}, \quad \rho = \exp \left( \frac{\Delta x^2}{2L^2} \right), \quad |i-j| < N/2
\]

\( N \) is the size of domain, \( C \) denotes the Gaussian correlation structure of the background errors while \( \sigma_b^2 \) is the background error variance. As a consequence:

\[
\mu(B) = \mu(C)
\]

where \( \mu() \) denote the condition number.

The ill conditioning of the DA inverse problem [11] (i.e. the sensitivity of the analysis to small perturbations in the data),
depends on the conditioning of the Hessian of each $J_i$ in (6). Small errors in the Hessian lead to large errors in its inverse, so the computed solution to the DA problem may be very inaccurate. In designing of the DA schemes, it is important to ensure that the conditioning of the Hessian is as small as possible, or it is essential to use preconditioning techniques to improve the conditioning.

In our model, matrix $B_i$ is decomposed as in (9) to have a preconditioner.

$$B_i = U_i D_i^2 U_i^T \in \Omega_i$$

The matrix $V_i = U_i D_i^2$ such that (10) is a preconditioner.

$$B_i = V_i V_i^T.$$ \hspace{1cm} (10)

Let $d = [y_k - H(x_k)]$ be the misfit, by using the following linearization of $\mathcal{H}$:

$$\mathcal{H}(x) = \mathcal{H}(x + \delta x) + H \delta x$$

where $H$ is the matrix obtained by the first order approximation of the Jacobian of $\mathcal{H}$ and, by setting $v_i = V_i^T \delta x_i$, the preconditioned (see [2]) cost function is:

$$J_i(v_i) = \frac{1}{2} v_i^T v_i + \frac{1}{2} (H V_i v_i - d_i)^T R_i^{-1} (H V_i v_i - d_i) +$$

$$+ \frac{1}{2} (V_i v_i^+ - V_i v_i^-)^T (V_i v_i^+ - V_i v_i^-)$$ \hspace{1cm} (11)

where $v_i^+$ and $v_i^-$ are shown in Fig. 2.

On each subdomain of $\Omega$, the function $J_i$ ($\forall i = 1, \ldots, N_{sub}$), is minimized using the L-BFGS method [20], [8].

### C. Mapping of DD-DA model on GPU architecture

Let $n_x \times n_y \times n_z$ be the size of a computational grid which discretize $\Omega$ and let $N_{sub} = p \times q$ be the number of sub domains partition of $\Omega$. We group gridpoints into blocks, we partition the computational grid into $p \times q$ three-dimensional (3-D) blocks of size $n_{xloc} \times n_{yloc} \times n_z$, each of which can be viewed as consisting of $q$ two-dimensional (2-D) blocks (see Fig. 4) with $n_{xloc}$ and $n_{yloc}$ defined in (12). These dimensions include overlapping $(2o_x \times 2o_y)$.

$$n_{xloc} = \frac{n_x}{p} + 2o_x, \quad n_{yloc} = \frac{n_y}{q} + 2o_y.$$ \hspace{1cm} (12)

In Fig. 4, orange points represent overlapping regions. From the viewpoint of the blocks, the overlapping region are not completely part of blocks, but come from adjacent blocks: the North, South, East, West overlapping are from neighboring blocks in those respectively directions.

Let us now describe the mapping of the DD-DA model on the GPU architecture. In any GPU implementation, the CPU (the host) runs the program and unloads some kernel functions (generally the more computationally demanding code parts) to the GPU (the device). In our algorithm, the CPU acquires the input data (data from forecasting model and observations) of
one computes the minimum of the function $f$, and the computations/work associated with all grid-points. Hence, physical grid-points to a thread-block which carries out all the computations associated with all grid-points. Therefore, we decided to assign each 3-D block of size $n_{xloc} \times n_{yloc} \times n_{zloc} = 6 \times 5 \times 3$. In our algorithm, we decided to assign each 3-D block of size $n_{xloc} \times n_{yloc} \times n_{zloc} = 12 \times 10 \times 3$, with $n_{xloc} = 6$ and $n_{yloc} = 5$. In this example, the computational domain is partitioned into 3-D blocks of size $n_{xloc} \times n_{yloc} \times n_{zloc} = 6 \times 5 \times 3$.

In any GPU implementation when a C program, using CUDA extensions and running on the CPU, invokes a kernel, many copies of this kernel (which are referred to as threads) are distributed to the available multiprocessors, where they are executed. Threads are grouped into thread-blocks, which are in turn arranged on a grid. Threads in a thread-block are able to communicate with each other very efficiently via the shared memory and are able to synchronize their executions. Thread-blocks can execute in any order relative to each other, which allows transparent scalability in the parallelism of CUDA kernels.

In our algorithm, we decided to assign each 3-D block of physical grid-points to a thread-block which carries out all the computations/work associated with all grid-points. Hence, thread-blocks are solved on the GPU concurrently. Each thread-block is able to communicate with each other very efficiently via the shared memory and are able to synchronize their executions. Thread-blocks can execute in any order relative to each other, which allows transparent scalability in the parallelism of CUDA kernels.

In our algorithm, we decided to assign each 3-D block of physical grid-points to a thread-block which carries out all the computations/work associated with all grid-points. Hence, thread-blocks are solved on the GPU concurrently. Each one computes the minimum of the function $J_i$ defined in (11) on its part of the computational grid, by using a CUDA version of the L-BFGS routine [8]. The boundary conditions between adjacent subdomains are efficiently communicated via the shared memory by introducing a synchronization of thread-blocks, as described in Algorithm 1. The global solution $x_DA$ defined in (4) is computed on the host finally.

III. NUMERICAL RESULTS

In this section, we present selected numerical results to demonstrate the effectiveness of the GPU-based parallel DD-DA algorithm. We used the CUDA 3.2 driver and toolkit, and all the experiments with the GPU code were conducted on a NVIDIA Tesla K20, which allows double-precision computations, and is connected to a quad-core Intel i7 CPU running at 3.07GHz, 12 GB of RAM. Out test case is based on shallow water equations which are a simplified version of NEMO forecasting model. According with NEMO, we have a variable time $t$, and space coordinates $(x$ and $y$) as independent variables. The dependent variables are the fluid height or depth $h$ and the two-dimensional fluid velocity field $u$ and $v$. The state variable is given in (13).

$$x_{_{DA}} = [h, u, v]^T.$$  

We assume $n_x = n_y = n$ and $n_z = 3$ which implies a problem size $N = n^3 \times 3$. The time step used for the temporal discretization of model is $d_t = 0.01$.

For executing Algorithm 1 on shallow water test case, we need to acquire $x_{DA}$ in (13) by running shallow water forecasting code. Observations vector $y_k$ is obtained by randomly choosing and randomly perturbing values of $x_{DA}$ [5]. We assume $H_k$, in (2) is a piecewise linear interpolation function and operators $B_k$ and $R_k$ defined in (14) and (15).

$$B_k = \sigma_k^2 C, \quad \sigma_k^2 = 0.5, \quad \sigma_o^2 = 0.5.$$  

For fixed values of $N_{_{sub}}$, the size of each subdomain used is $N_{_{sub}}^{-1}$ as explained in Section II-C.

In [5] the authors provided a formal mathematical proof of the reliability of DD-DA model and accuracy of its solution. Also, our implementation on GPU does not affect accuracy of numerical results as the arithmetic system we are using is double precision.

Let $T_{N_{_{sub}}}(N)$ be the execution time of Algorithm 1 for a problem size $N$ defined in (16) where $T_{H}(N)$ is the execution time of algorithm running on the CPU, $T_{\text{com}}(H+D)(N)$ is the communication time between host and device and with $T_{D}(N)$ execution time of algorithm running on the GPU. Then

$$T_{N_{_{sub}}}(N) = T_{H}(N) + T_{\text{com}}(H+D)(N) + T_{D}(N).$$  

(16)
As explained at the end of Section II-C, in Algorithm 1, the host acquires input data, it computes operators $H_k$, $d_k$, $R_k$, and $V_k$, and sends all these data to device which solves concurrently $N_{sub}$ thread-blocks for computing the minimum of function $J$ in (11). In practice, the host computer operators which are input for computations on device and the device solves the DD-DA model (4)-(6). Hence, $T_{D}^{N_{sub}}(N)$ is execution time that CPU needs for building data. These data are transferred just once as well as output data $x^D_k$. So we have that $T_{com}(H+D)(N)$ is reduced to the time of I/O transfer. For this reasons we evaluate the performance of DD-DA implementation on GPU by analysing $T_{D}^{N_{sub}}(N)$. Table I shows execution time of algorithm running on GPU for $N_{sub}=128$ which give a problem size $O(10^7)$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$p$</th>
<th>$T_{D}^{N_{sub}}(N)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O(10^7)$</td>
<td>1</td>
<td>0.144</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.024</td>
</tr>
</tbody>
</table>

Theoretical bandwidth $B_W$ can be calculated using hardware specifications available in the product literature as in (19) where $M_{CR}$ is the memory clock rate and $M_{SW}$ is the wide memory interface. In (19) we convert the memory clock rate to Hz, multiply it by the interface width (divided by 8, to convert bits to bytes) and multiply by 2 due to the double data rate. Finally, we divide by $10^9$ to convert the result to GB/secs.

$$B_W = M_{CR} \cdot 10^6 \cdot \frac{M_{SW}}{8} \cdot \frac{2}{10^9} \text{ GB/secs}.$$  

(19)

In our case, the NVIDIA Tesla K20 GPU uses DDR (double data rate) RAM with a memory clock rate of 2.6 MHz and a 320-bit wide memory interface. Using these data items, the peak theoretical memory bandwidth of the NVIDIA Tesla K20 is 208 GB/secs, as computed in the following

$$B_W = 2600 \cdot 10^6 \cdot \frac{320}{8} \cdot \frac{2}{10^9} \text{ GB/s = 208 GB/secs}.$$  

For a domain of dimension $O(10^7)$ we have $D_N = 3.7 \text{ GB}$ which gives $T_{com}(D) \approx 3.7/208 \text{ secs} \approx 0.017 \text{ secs}$.

The Scale-Up factor (see [5]) of algorithm running on the GPU is function of $T_{flop}(D)(N)$ as given in (20)

$$S_{DD-DA}^{N_{sub}} = \frac{T_{N_{sub}}^{flop}(D)(N)}{T_{mem}(D)(N)}.$$  

(20)

$$S_{DD-DA}^{N_{sub}} = \frac{T_{N_{sub}}^{flop}(D)(N)}{T_{mem}(D)(N)}.$$  

Where $T_{N_{sub}}^{flop}(D)(N)$ is the time for global and local memories transfers into the device, $T_{N_{sub}}^{flop}(D)(N)$ is the computing time required for execution of floating point operations.

$$T_{D}^{N_{sub}}(N) = T(N) \times T_{flop}. $$  

(21)

With this assumption, the scale-up factor of algorithm is given in (22) as we have in our case $T(N) = O(N^2)$.

$$S_{DD-DA}^{N_{sub}} = \frac{T(N)}{2p T(N^2)} = O \left( \frac{N^2}{2p} \right) = O (2p).$$  

(22)

Table 2 shows values of $T_{N_{sub}}^{flop}(D)(N)$, and values of measured Scale-up factor compared with theoretical one. Finally, we observe that measured values of Scale-up factor are defined as in (23) with $\alpha_1 < 1$ and $\alpha_{N_{sub}} < 1$ as the parallel implementation we have by using CUDA.

$$\text{measured } S_{DD-DA}^{N_{sub}} \approx \frac{T_{N_{sub}}^{flop}(D)(N)}{T_{mem}(D)(N)} \approx \frac{T_{N_{sub}}^{flop}(D)(N)}{T_{mem}(D)(N)} \approx \frac{T_{N_{sub}}^{flop}(D)(N)}{T_{mem}(D)(N)} \approx \frac{T_{N_{sub}}^{flop}(D)(N)}{T_{mem}(D)(N)}.$$  

(23)
IV. CONCLUSIONS AND FUTURE WORK

Moving forward to exascale will put heavier demands on algorithms in at least two areas: the need for increasing amounts of data locality in order to perform computations efficiently, and the need to obtain much higher factors of fine-grained parallelism as high-end systems support increasing numbers of compute threads. As a consequence, parallel algorithms must adapt to this environment, and new algorithms and implementations must be developed to extract the computational capabilities of the new hardware.

We presented a parallel algorithm on GPU which is based on a domain decomposition approach. The standard approach for reducing the execution time of an algorithm on GPU is to place concurrency inside the most time-consuming computational kernels, i.e., to introduce a parallelism at the level of fine-grained computations. Furthermore, in order to reduce the data movement between host and device, thus increasing the computation/communication ratio, the parallel algorithm relies on a domain decomposition approach that introduces a coarse-grained data decomposition which have more favorable performance characteristics. Finally, coarse and fine grained computations are suitably mapped on the processing elements of our target architecture, that is the multiprocessors and the ALUs respectively [9].

We are currently working on the deployment of this algorithm in a concrete scenario. Mainly, we are working on the variational DA systems used with the NEMO ocean model, on emerging exascale computing architectures [7].

REFERENCES


Rossella Arcucci Her area of expertise is in Numerical Analysis, Scientific Computing and development of methods, algorithms and software for scientific applications on high performance architectures including parallel and distributed computing. During her master degree in Mathematics she started to work on mathematical models to study real phenomena and in her master degree thesis in 2008, she used discretization methods to solve partial differential equations (PDE). She gave particular attention to discretization methods for elliptic, parabolic and hyperbolic problems which are the basis of all the models that describe real phenomena. During and after her PhD (obtained in 2012), she worked on Data Assimilation problem. Her main contribution was been the development of a fully scalable mathematical model for Variational Data Assimilation based on Domain Decomposition method. Another interest of her is on high performance computing (HPC). In this field she worked on a project aimed at the development of fully scalable software for Data Assimilation able to effectively take advantage of the available HPC resources and handle big data. By working on numerical models which describe real phenomena, she realised the importance of producing solutions as accurate as possible. She became aware of the importance for a scientist of knowing how errors generated in hardware propagate in numerical solutions. She has worked on models for error detection in circuits in the context of reliable systems with electrical and electronic engineers. At the moment, she is combining her competences in numerical analysis and hardware architectures by working on numerical and parallel techniques for both improving accuracy of solution and reducing execution time for solving Data Assimilation models on supercomputers. She is a researcher of University of Naples Federico II and she is collaborating with a group of physics at Imperial College London on Data Assimilation problems for oceanographic data.

Luisa D’Amore Degree in Mathematics, in 1988. PhD in Applied Mathematics and Computer Science, in 1995. Researcher in Numerical Analysis, in 1996. Associate professor of Numerical Analysis, at University of Naples Federico II (IT), since 2001. Associate staff researcher of the ASC (Advanced Scientific Computing) Division of CMCC (Euro Mediterranean Center on Climate Changes), since 2011. Member of the Academic Board of the PhD in Mathematics and Informatics, of University of Naples Federico II. Teacher of courses in Numerical Analysis, Scientific Computing and Parallel Computing. Research activity focuses on Scientific Computing and it is addressed to the numerical solution of ill-posed inverse problems with applications in image analysis, medical imaging, astronomy, digital restoration of films and data assimilation. The need of computing the numerical solution on a suitable time, induced by the applications, often requires the use of advanced computing architectures. This involves designing and development of algorithms and software capable of exploiting the high performance of emerging computing infrastructures. Research produces a total of about 100 publications in refereed journals and conference proceedings.
Simone Celestino He is currently a PhD student in Mathematics and Computer Science at University of Naples “Federico II”. His area of interest is the High Performance Computing with particular attention to parallel implementations on General Purpose Graphic Processing Unit (GPGPU). He had a Master Degree in Computer Science obtained at the same University where he is a PhD student, the subject of him thesis was on the development of a parallel software implementing a rendering algorithm on GPUs.

Giuseppe Scotti Degree in Computer Science at University of Naples “Parthenope”. His thesis was on an implementation of a biological model on Graphics Processing Units (GPU). He is currently a master degree student in Computer Science at University of Naples Federico II. During his degree and master degree he also worked as software developer in High Performance Computing.

Giuliano Laccetti is presently Full professor of computer science at the University of Naples Federico II, in Naples, Italy. He received his Laurea degree (cum laude) in Physics from the University of Naples; his main research interests are Mathematical Software, Scientific Computing, High Performance Architecture for Scientific Computing, Distributed Computing, Grid Computing, Cloud Computing, Algorithms on emerging hybrid architectures (CPU+GPU, etc.). He is author (or co-author) of about 90 papers published on refereed international Journals, chapters of books, Conference Proceedings and Technical Reports. He has been involved in several EU funded Projects (EGEE, EGEE II, EGEE III; in this last case he served as University of Naples scientific coordinator). He has been involved also in National (EU funded) Projects as SCOPE, and, presently, RECAS; in this case, he is member of the Scientific and Management Board and he is also coordinator of the curriculum Master Degree of the University of Naples about Technologies for The High Performance Scientific Computing, funded by the RECAS Project itself. In the recent past, he organised and chaired international workshops (joined to the PPAM Conference, from 2007 to 2013) in Gdansk, Torun, Warsaw, as well as he is planning to organise another one next year, in Krakow, about methodologies, algorithms and software for hybrid computer architectures; most recently, he (co)organised and (co)chaired, in Naples, the International Conference Advances in Pure and Applied Mathematics.

Giuliano Laccetti is head of the High Performance Scientific Computing Lab of the Department of Mathematics of the University of Naples; he is also member of the editorial board of the journal Advances in Computer Science and Engineering. Presently, he teaches Computer Programming I and Computer Programming II to Computer Science undergraduate students, and Parallel and Distributed Computing to Computer Science graduate students, as well as to Mathematics graduate students. He is also member of the Teaching and Steering Committee of the Ph.D. school on Mathematics and Computer Science of the University of Naples. Giuliano Laccetti is (or has been) member of ACM, IEEE-Computer Society, SIAM, SIMAI, AICA.