INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS
Int. J. Numer. Meth. Fluids 2003; **42**:265–295 (DOI: 10.1002/fld.493)

Multidimensional FEM-FCT schemes for arbitrary time stepping

D. Kuzmin[*,†], M. Möller and S. Turek

Institute of Applied Mathematics (LS III), University of Dortmund, Vogelpothsweg 87, D-44227, Dortmund, Germany

SUMMARY

The flux-corrected-transport paradigm is generalized to finite-element schemes based on arbitrary time stepping. A conservative flux decomposition procedure is proposed for both convective and diffusive terms. Mathematical properties of positivity-preserving schemes are reviewed. A nonoscillatory low-order method is constructed by elimination of negative off-diagonal entries of the discrete transport operator. The linearization of source terms and extension to hyperbolic systems are discussed. Zalesak's multidimensional limiter is employed to switch between linear discretizations of high and low order. A rigorous proof of positivity is provided. The treatment of non-linearities and iterative solution of linear systems are addressed. The performance of the new algorithm is illustrated by numerical examples for the shock tube problem in one dimension and scalar transport equations in two dimensions. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: high resolution; finite elements; unstructured grids; flux-corrected-transport; hyperbolic conservation laws

1. INTRODUCTION

An adequate treatment of convection-dominated transport problems remains a major challenge in numerical simulation of both compressible and incompressible flows. As a rule, approximate solutions produced by linear discretization schemes are corrupted by non-physical oscillations and/or excessive numerical diffusion. Thus, the use of a nonlinear shock-capturing viscosity is indispensable if a good resolution of singularities is to be achieved without sacrificing important properties of the exact solution such as positivity and monotonicity. The pioneering work of Boris and Book [1] has established the basic principles for the construction of high-resolution schemes. In particular, the authors introduced the concept of *flux-corrected-transport* (FCT) which essentially amounts to using a low-order method in regions with steep gradients and a high-order method elsewhere.

[*] Correspondence to: D. Kuzmin, Institute of Applied Mathematics, LS III, University of Dortmund, Vogelpothsweg 87, D-44227, Dortmund, Germany.

[†] E-mail: kuzmin@math.uni-dortmund.de

Received 20 March 2002
Revised 7 January 2003

The original FCT algorithm named SHASTA was a rather specialized one-dimensional finite-difference scheme. It was dramatically improved by Zalesak [2] who proposed a genuinely multidimensional generalization applicable to arbitrary combinations of high- and low-order discretizations. In the finite-element framework, flux correction was first exploited by Parrott and Christie [3] and promoted to maturity by Löhner and his coworkers [4, 5]. The classical FEM-FCT methodology builds on Zalesak's formulation with antidiffusive element contributions in place of fluxes. An alternative approach is based on applying the flux limiter edge by edge [6, 7].

Some modern compressible flow solvers abandon the conventional finite-element data structure altogether in favour of an edge-based data structure. Peraire *et al.* [8] developed a procedure for the conservative decomposition of Galerkin integrals into fluxes assigned to the edges of a triangular or tetrahedral mesh. The transition to an edge-based data structure reduces the overhead incurred by indirect addressing and offers considerable savings in terms of both CPU time and memory requirements [9]. Moreover, it facilitates the extension of the one-dimensional theory to unstructured meshes. In particular, popular upwind-biased schemes based on flux difference splitting [10, 11] or flux vector splitting (e.g. References [12–14]) can be readily implemented in the finite-element context [15]. The spatial accuracy can be enhanced by using non-linear discretizations of MUSCL, TVD or LED type utilizing flux/slope limiters [13, 16–19]. To this end, a local one-dimensional stencil is reconstructed for each edge by the insertion of two dummy nodes. The solution values at these nodes are obtained by appropriate gradient recovery and/or linear interpolation techniques.

In this paper, we present a coherent methodology for the design of multidimensional FCT schemes employing either the standard or the edge-based data structure for the finite-element mesh. Its foundations were laid in References [20, 21] where we applied the theory of positivity-preserving schemes [13, 22] to derive a conservative FEM-FCT formulation valid for arbitrary time stepping. No other implicit high-resolution finite-element schemes seem to be available to date. Explicit methods are typically more accurate than implicit ones, but the severe stability limitation makes them extremely inefficient for problems with strongly varying velocities and/or mesh sizes. Likewise, steady-state computations based on time marching call for a fully implicit time discretization. The details of the temporal evolution are immaterial in this case, so that the (artificial) time step should be chosen as large as possible to minimize the computational cost.

While the new FEM-FCT procedure was implemented using the traditional data structure, the discrete diffusion/antidiffusion terms were decomposed into numerical fluxes which were treated in an essentially one-dimensional fashion. In the case of simplex elements, the fluxes can be associated with edges of the finite-element mesh. At the same time, interacting nodes of multilinear elements do not have to be connected by an edge. The low-order transport operator was constructed from the high-order operator by the elimination of negative off-diagonal matrix entries. An advantage of this approach is that it is completely independent of the underlying discretization procedure and of the spatial mesh. Moreover, it automatically yields an upper bound for admissible time steps.

Below we extend the FEM-FCT schemes introduced in Reference [21] to problems with source terms and systems of non-linear hyperbolic conservation laws. Another contribution of this work is a universal strategy for the decomposition of Galerkin-type discretizations into skew-symmetric internodal fluxes. Unlike the widespread algorithm of Peraire *et al.* [8], the new decomposition method carries over to general finite-element approximations on quadrilateral and hexahedral meshes. Therefore, it appears to be a promising tool for the development of 'edge-based' flow solvers. Last but not least, we elaborate on various implementation aspects and discuss the efficiency of iterative solution techniques. Encouraging results obtained for a number of one- and two-dimensional test problems demonstrate the potential of the implicit formulation.

2. FEM FOR SCALAR CONSERVATION LAWS

Consider a generic time-dependent conservation law for a scalar quantity u

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f} = q \quad \text{in } \Omega \tag{1}$$

where q is a source term, and $\mathbf{f}$ is a flux function, which may depend on the solution in a nonlinear way. Typically we can distinguish between convective and diffusive fluxes:

$$\mathbf{f} = \mathbf{v}u - \varepsilon \nabla u$$

Here the first term represents the convective transport with a characteristic velocity $\mathbf{v}$. The second one describes the diffusive transfer of the conserved quantity (e.g. mass or heat) from regions of high concentration into regions of low concentration. If the diffusion coefficient ε vanishes or is small as compared to $\mathbf{v}$, the flow is dominated by convection. In this case, the hyperbolic nature of the equation at hand makes it notoriously difficult to treat numerically.

The problem setting is completed by specifying initial and boundary conditions. The variational formulation of Equation (1) reads

$$\int_\Omega w \left[\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f} - q \right] \mathrm{d}\mathbf{x} = 0, \quad \forall w \tag{2}$$

The standard Galerkin space discretization is performed by using an approximation of u in a suitable finite-dimensional space and substituting the basis functions φ_i for w. For customary finite elements, the sum of basis functions equals unity: $\sum_i \varphi_i \equiv 1$. Summing all equations and invoking the divergence theorem, we recover the underlying integral form of the conservation law:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega u \, \mathrm{d}\mathbf{x} = \int_\Omega q \, \mathrm{d}\mathbf{x} - \int_{\partial\Omega} \mathbf{f} \cdot \mathbf{n} \, \mathrm{d}s \tag{3}$$

where $\mathbf{n}$ denotes the unit outward normal. This relation implies that the total amount of u in Ω may only change due to fluxes through the boundary and internal sources or sinks. The integral formulation is more general than the differential one, because it remains valid for discontinuous solutions.

In light of the above, the Galerkin finite-element method is conservative in an integral sense. Mass conservation is a very important property which ensures that if a consistent numerical method does converge, then it converges to a weak solution of the conservation law. At the same time, the uniqueness of weak solutions is not guaranteed for non-linear hyperbolic equations. Hence, an additional condition might be needed to pick out the physically relevant entropy solution obtained in the limit of vanishing viscosity [23].

Finite volume and discontinuous Galerkin methods apply formulation (3) directly to each element of the triangulation, so that mass conservation is enforced not only globally but also locally. Flux correction for such discontinuous approximations is fairly straightforward. The objective of this paper is to extend the available FCT tools to continuous (linear and multilinear) finite elements and systems of hyperbolic conservation laws.

3. GALERKIN FLUX DECOMPOSITION

It is well known from the theory of finite-difference methods that a numerical scheme is conservative if it admits decomposition into a sum of fluxes from one node into another. Indeed, as long as the internodal fluxes are equal in magnitude and opposite in direction, the total mass of the system may only change due to boundary fluxes. Hence, it is highly desirable to represent the numerical method in conservation form whenever possible. At the same time, it has been largely unclear how to accomplish this for finite-element discretizations on unstructured meshes.

Peraire *et al.* [8] demonstrated that the flux decomposition is feasible for the Galerkin method employing triangular or tetrahedral elements with linear basis functions which have a constant gradient. The authors advocated the transition to an edge-based data structure which offers certain computational advantages as compared to the conventional element-based formulation. The derivation of the decomposition procedure is quite tedious, though. The interested reader is referred to the monographs by Lyra [15] and Löhner [9] for details. An excellent presentation of edge-based finite element methods catering for high resolution on unstructured grids can also be found in the review article by Morgan and Peraire [19]. Unfortunately, the approach proposed by Peraire *et al.* [8] is limited to simplex elements with linear interpolation of the unknown solution and of the associated flux function. In what follows, we will work out an alternative flux decomposition technique applicable to general finite-element approximations on arbitrary meshes including quadrilateral and hexahedral ones.

Integration by parts of the weak formulation (2) yields

$$\int_\Omega w\frac{\partial u}{\partial t}\,\mathrm{d}\mathbf{x} - \int_\Omega \nabla w\cdot\mathbf{f}\,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} w\mathbf{f}\cdot\mathbf{n}\,\mathrm{d}s - \int_\Omega wq\,\mathrm{d}\mathbf{x} = 0, \quad \forall w \tag{4}$$

A common practice in finite-element computations of compressible flow is to approximate the fluxes in the same way as the desired solution. This approach termed the *group finite-element formulation* by Fletcher [24] provides an efficient matrix assembly leading to a considerable reduction in the computational cost. Surprisingly enough, it was also found to produce a small gain in accuracy e.g. when applied to the Burgers equation on a uniform grid [24]. The resulting savings in CPU time become increasingly pronounced for multidimensional problems and/or strong non-linearities.

Let the solution, the fluxes and the source terms be represented in the form

$$u = \sum_j u_j\varphi_j, \quad \mathbf{f} = \sum_j \mathbf{f}_j\,\varphi_j, \quad q = \sum_j q_j\varphi_j \tag{5}$$

In fact, it is not compelling to use the same approximations for u and $\mathbf{f}$. For instance, one can think of a hybrid method, whereby the unknown solution is sought in the space of continuous piecewise-linear or multilinear functions, while the fluxes are interpolated using non-conforming Crouzeix–Raviart [25] or Rannacher–Turek [26] elements. In this case, the solution values would be defined at the nodes, whereas the degrees of freedom for the fluxes would reside on the edges of the finite-element mesh.

After the substitution of expressions (5) and the weighting functions $w = \varphi_i$ into the variational formulation (4), we obtain

$$\sum_j \left[\int_\Omega \varphi_i \varphi_j \, \mathrm{d}\mathbf{x} \right] (\dot{u}_j - q_j) - \sum_j \left[\int_\Omega \nabla \varphi_i \varphi_j \, \mathrm{d}\mathbf{x} - \int_{\partial\Omega} \varphi_i \varphi_j \, \mathbf{n} \, \mathrm{d}s \right] \cdot \mathbf{f}_j = 0 \tag{6}$$

which can be written compactly in matrix form as

$$M_\mathrm{C}(\dot{u} - q) = K_x f_x + K_y f_y + K_z f_z \tag{7}$$

in the three-dimensional case. For some applications (e.g. steady-state flows), it may be worthwhile to replace the consistent mass matrix M_C by its diagonal counterpart M_L obtained by the conservative row-sum mass lumping, which can be interpreted as using a low-order quadrature rule for the numerical integration [27]. This modification essentially decouples the solution increments and results in a finite-difference-like discretization. In particular, no linear systems have to be solved for explicit schemes. The utility of the group formulation is illustrated by the fact that the matrices K_x, K_y and K_z engendered by the corresponding first-order derivatives can be assembled once and for all at the beginning of the simulation, as long as the mesh does not change. This is in contrast to the standard finite-element approach, whereby the discrete operators for the linearized convective terms have to be updated in each time step.

By construction, the discretized flux term consists of an interior part and a boundary part. The former is given by the integral

$$\sum_j \left[\int_\Omega \nabla \varphi_i \, \varphi_j \, \mathrm{d}\mathbf{x} \right] \cdot \mathbf{f}_j = \sum_j \mathbf{c}_{ij} \cdot \mathbf{f}_j, \quad \mathbf{c}_{ij} = \int_\Omega \nabla \varphi_i \, \varphi_j \, \mathrm{d}\mathbf{x} \tag{8}$$

where the coefficient matrices $c_{ij}^x, c_{ij}^y, c_{ij}^z$ possess the zero column sum property, since it is assumed that the sum of basis functions equals unity. Therefore, it is possible to express the diagonal coefficients in terms of off-diagonal ones:

$$\sum_i \mathbf{c}_{ij} = 0 \;\Rightarrow\; \mathbf{c}_{ii} = -\sum_{j \neq i} \mathbf{c}_{ji} \tag{9}$$

It follows that the interior flux term (8) can be rewritten as

$$\sum_j \mathbf{c}_{ij} \cdot \mathbf{f}_j = \sum_{j \neq i} g_{ij} \quad \text{where } g_{ij} := \mathbf{c}_{ij} \cdot \mathbf{f}_j - \mathbf{c}_{ji} \cdot \mathbf{f}_i \tag{10}$$

The newly introduced quantity g_{ij} represents the *Galerkin flux* from node j into node i. It is important that $g_{ji} = -g_{ij}$, so that node j receives the same contribution with the opposite sign. Roughly speaking, g_{ij} is nothing else but the 'projection' of an averaged flux onto the segment joining the two nodes. It is worth noting that for one-dimensional linear finite elements the weighting coefficients are simply $c_{ij}=-c_{ji}=\frac{1}{2}$, so that $g_{ij}=(f_i+f_j)/2$. This stems from the well-known fact that the Galerkin method is equivalent to the central difference approximation of differential operators.

By virtue of relation (10), the terms resulting from the Galerkin discretization can be decomposed into a sum of numerical fluxes similar to those encountered in conservative finite-difference methods. Galerkin fluxes can be associated with edges of the graph representing the connectivity of the global finite-element matrix. For linear triangles or tetrahedra, the graph edges match the physical edges of the element, while multilinear or high-order approximations will also give rise to 'internal' edges, which merely link the interacting degrees of freedom. As a rule, each node exchanges mass with other nodes sharing an element with it. The net flux between any pair of nodes is zero, so that mass conservation is guaranteed. The contribution of boundary fluxes is given by the surface integral in Equation (6), which can be evaluated using an appropriate quadrature rule.

Importantly, the flux decomposition procedure is applicable to *generalized diffusion operators* [21] which are defined as symmetric matrices having zero row and column sums. The purely diffusive Galerkin flux assumes a remarkably simple form

$$\sum_i d_{ij}=\sum_j d_{ij}=0, \quad d_{ij}=d_{ji}, \;\Rightarrow\; g_{ij}=d_{ij}(u_j-u_i) \tag{11}$$

Note that generalized diffusion operators are not required to have continuous counterparts. Some typical examples are the discrete Laplacian, the streamline diffusion operator and the matrix $M_{\mathrm{C}}-M_{\mathrm{L}}$ sometimes referred to as 'mass diffusion'. As we will see shortly, the properties of discrete diffusion operators render them a valuable tool for the design of non-oscillatory low-order methods to be combined with high-order ones within the flux-corrected-transport algorithm.

Another promising approach to the derivation of high-resolution finite-element schemes involves the replacement of the original Galerkin flux by another *consistent numerical flux*. Its potential is demonstrated by numerous publications [15, 18, 19] in which one-dimensional limiters are successfully applied on unstructured meshes in conjunction with the edge-based data structure of Peraire *et al.* [8]. In particular, approximate Riemann solvers with upwind-biased interpolations, scalar-limited dissipation schemes and other essentially one-dimensional discretization tools developed for systems of hyperbolic conservation laws can be integrated into the finite-element framework. For high-resolution schemes based on the reconstruction of a local one-dimensional stencil, the limiter depends not only on the unknowns and fluxes but also on the algorithm employed to obtain the solution values at dummy nodes. Simulation results are strongly affected by the choice of the recovery procedure, especially in the case of highly irregular meshes [15].

4. POSITIVITY CRITERIA

In this section, we introduce some mathematical tools which are of importance for the development of high-resolution schemes. Since many physical quantities are inherently non-negative, it is natural to impose this constraint on the numerical solution as well. Moreover, it is known that positivity-preserving schemes do not give rise to non-physical phenomena. In particular, they encompass the important class of *monotone* methods which guarantee that a converged solution of the conservation law does satisfy the entropy inequality. A very handy positivity criterion is provided by the concept of an *M-matrix* as explained below.

Definition
A non-singular discrete operator $A \in \mathbb{R}^{n\times n}$ is called an M-matrix if $a_{ij} \leqslant 0$ for $i \neq j$ and all the entries of A^{-1} are non-negative.

If A is strictly diagonally dominant and $a_{ii} > 0$, while $a_{ij} \leqslant 0$ for $i \neq j$, then A is an M-matrix. Note that for M-matrices $Ax \geqslant 0$ implies that $x \geqslant 0$. This property leads to the following fundamental lemma:

Lemma
Let the numerical scheme be represented in abstract matrix operator form as

$$Lu^{n+1} = Ru^n \tag{12}$$

A sufficient condition for such a scheme to preserve positivity is that L be an M-matrix and all entries of R be non-negative ($R \geqslant 0$).

The conditions of the lemma are sufficient (but not necessary) to ensure that the numerical solution satisfies the discrete maximum principle. Furthermore, it seems expedient to require that the steady-state counterpart of L be an M-matrix as well. Otherwise non-physical ripples might emerge even though the solution remains positive.

To introduce another useful concept, consider a semi-discrete problem of the form

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_j c_{ij} u_j, \quad \sum_j c_{ij} = 0 \tag{13}$$

where u_i are the nodal values, and c_{ij} are some coefficients depending on the procedure employed for spatial discretization. In particular, the lumped-mass Galerkin discretization of the transport equation admits such a representation if the flow is incompressible.

Since the coefficient matrix has zero row sum, the scheme can be rewritten as

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j\neq i} c_{ij}(u_j - u_i) \tag{14}$$

Furthermore, suppose that all coefficients are non-negative: $c_{ij} \geqslant 0$, $j \neq i$. Then this scheme is stable in the L_∞-norm. Indeed, if u_i is a maximum, then $u_j - u_i \leqslant 0$, $\forall j$, so that $\mathrm{d}u_i/\mathrm{d}t \leqslant 0$. Hence, a maximum cannot increase, and similarly a minimum cannot decrease. As a rule, coefficient matrices are sparse, so that $c_{ij} = 0$ unless i and j are adjacent nodes. Arguing as above, one can show that in this case a *local* maximum cannot increase, and a *local* minimum cannot decrease. Schemes which possess this property are called local extremum diminishing (LED).

The LED criterion was introduced by Jameson [13, 22] as a convenient tool for the design of high-resolution schemes on unstructured meshes. It implies positivity, since if the solution is positive everywhere, then so is the global minimum which cannot decrease by definition. Hence, the LED property provides an effective mechanism for preventing the birth and growth of non-physical oscillations. In the one-dimensional case, it guarantees that the total variation of the solution defined as

$$TV(u) = \int_{-\infty}^{+\infty} \left| \frac{\partial u}{\partial x} \right| \mathrm{d}x \tag{15}$$

does not increase. For the sake of simplicity, consider homogeneous Dirichlet boundary conditions at both endpoints. Then the total variation is given by

$$\mathrm{TV}(u) = 2(\sum \max u - \sum \min u) \tag{16}$$

Thus, a one-dimensional LED scheme is necessarily total variation diminishing (TVD). This is a highly advantageous property, which has formed the basis for the development of a whole class of non-oscillatory schemes. The advantage of the LED principle as compared to TVD concepts is its applicability to multidimensional problems on both structured and unstructured meshes.

Recall that Equations (13) and (14) correspond to the problem discretized in space only. Let us now investigate the conditions under which an LED scheme will remain positive after the time discretization. If the standard one-step θ-scheme is employed, the fully discretized equation reads

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \theta \sum_{j \neq i} c_{ij}(u_j^{n+1} - u_i^{n+1}) + (1-\theta) \sum_{j \neq i} c_{ij}(u_j^n - u_i^n), \quad 0 \leqslant \theta \leqslant 1 \tag{17}$$

The choice of the parameter θ specifies the type of time stepping. The extreme cases $\theta = 0$ and $\theta = 1$ define the well-known forward and backward Euler methods. Both of them are first-order accurate with respect to the time step Δt. The method corresponding to $\theta = 0.5$ is known as the Crank–Nicolson scheme, which is second-order accurate.

The application of our lemma to Equation (17) yields the following theorem [21].

Positivity Theorem

A local extremum diminishing scheme discretized in time by the backward Euler method is unconditionally positive. Other time-stepping schemes ($0 \leqslant \theta < 1$) preserve positivity under the CFL-like condition

$$1 + \Delta t(1-\theta) \min_i c_{ii} \geqslant 0 \tag{18}$$

An important message delivered by this theorem is that the positivity criterion at our disposal makes it possible to obtain rigorous estimates of the largest admissible time step for explicit schemes. Remarkably, the derivation of the upper bound does not require any knowledge of the underlying partial differential equation and of the employed spatial mesh. It is sufficient to examine the diagonal coefficients c_{ii} of the semi-discrete scheme. Upper bounds for non-LED schemes can be readily derived in the same way.

5. CLASSIFICATION OF NON-LINEAR SCHEMES

The positivity concepts introduced above lay the groundwork for the construction of high-resolution numerical schemes. The desired properties of discrete operators can be realized by the introduction of artificial diffusion or by the use of upwind biasing. However, it was shown by Godunov [28] that no linear discretization method of order higher than first can guarantee monotonicity of the numerical solution. In practice, this means that the results produced by such schemes are overly diffusive. Higher accuracy can be attained by sophisticated non-linear methods with coefficients depending on the solution. The control of artificial diffusion is typically executed by means of flux or slope limiters which adaptively switch between high- and low-order methods. The high-order approximation is used in regions where the solution is smooth, whereas the order is reduced in the vicinity of discontinuities so as to dampen non-physical undershoots and overshoots.

Non-linear high-resolution schemes can be classified into *smoothness monitoring* (SM) and *diffusion-antidiffusion* (DAD) methods [29]. The former approach relies on some kind of smoothness sensors to assess the minimum amount of artificial diffusion that must be applied to preserve monotonicity or at least positivity. Both the amplitudes and the phases of the Fourier modes are predetermined by the SM procedure. By contrast, numerical schemes falling into the DAD class employ ‘operator splitting’ to separate the effects of convective transport and (limited) antidiffusion. At the first stage, sufficient *constant* artificial diffusion is built into the discretization of transport terms, so as to maintain monotonicity. This modification reduces the phase errors but leads to a pronounced damping of the harmonics. The second step corrects the amplitudes by introducing a properly tuned non-linear antidiffusion. Remarkably, the phases are not affected by this correction, so that the improvement of the phase accuracy gained in the first step persists. Dietachmayer [29] argued that the better phase properties of DAD methods make them superior to SM techniques. Indeed, since the tendency of the solution to oscillate is alleviated in the first step, more compensating antidiffusion can be added in the second step without loss of positivity. Hence, DAD schemes can be generally expected to exhibit better accuracy and contain less (net) artificial dissipation than SM methods.

In the design of flux-corrected-transport algorithms, the SM and DAD techniques have been used interchangeably. The SHASTA scheme of Boris and Book [1] is a classical DAD method. At the same time, its generalization proposed by Zalesak [2] is of the SM type. Löhner *et al.* [4] employed mass lumping and added constant mass diffusion to transform an explicit high-order method into a low-order one. This approach is very attractive from the view point of computational efficiency, but the involved free parameter should be chosen with care. The effective diffusion coefficient is inversely proportional to the time step, so that the scheme becomes increasingly overdiffusive as the time step dwindles. This is acceptable, because excessive smearing is likely to be cured by the antidiffusion step. On the contrary, the case of insufficient artificial diffusion is very alarming, since spurious extrema may arise and be transmitted to the final solution.

Explicit time stepping stipulates the use of small time steps due to the CFL condition. In this case, the variable diffusion coefficients corresponding to SM schemes are smaller than the constant value for a typical DAD method. Some extra diffusion may even prove to be beneficial as explained above. On the other hand, implicit schemes are usually operated at large Courant numbers, which rules out the constant-diffusion approach to the construction of the low-order method. Due to the failure of DAD schemes to cope with large time steps, we will adhere to the SM methodology. Finite-element methods utilizing smoothness sensors and/or flux limiters to modulate artificial diffusion were proposed in the late 1980s by several authors [7, 16, 30]. As we are about to see, the flexibility of locally modulated dissipation makes it a valuable tool for the derivation of finite-element discretizations satisfying the positivity constraint [20, 21].

6. LOW-ORDER DISCRETIZATION

To a large extent, the performance of the flux-corrected-transport procedure depends on the quality of the underlying low-order method which is supposed to preserve positivity and withstand the formation of numerical wiggles. In the realm of finite-difference and finite-volume discretizations, a perfect candidate for this job is certainly the upwind scheme. At the same time, it has been largely unclear how to perform upwinding in the finite-element framework. Most upwind-like finite-element methods encountered in the literature resort to a finite-volume discretization for the convective terms [16, 31]. An alternative derivation of the least diffusive positivity-preserving scheme can be carried out by adding discrete diffusion depending solely on the magnitude and position of negative entries in the finite-element matrix [21].

The scalar conservation law (1)–(2) discretized in space by the Galerkin method can be written in the form

$$M_C \frac{\mathrm{d}u}{\mathrm{d}t} = K^H u + q \tag{19}$$

In order to render this semi-discrete scheme positive, we must perform mass lumping, so as to remove the implicit antidiffusion built into the consistent mass matrix. Furthermore, the discrete transport operator K^H should be modified by adding a proper amount of artificial diffusion. We define its low-order counterpart as $K^L = K^H + D$, where the dissipation tensor D is designed so as to eliminate all negative off-diagonal entries of the high-order operator [21]:

$$d_{ii} = -\sum_{k\neq i} d_{ik}, \quad d_{ij} = d_{ji} = \max\{0, -k_{ij}^H, -k_{ji}^H\}, \quad \forall i<j \tag{20}$$

In essence, this corresponds to applying one-dimensional diffusion operators associated with the (fictitious) edges connecting the adjacent nodes. It is easy to verify that D is characterized by zero row and column sums, and thus enjoys all properties of generalized diffusion operators including mass conservation. Note that if physical diffusion is strong enough, so that the coefficients are non-negative from the outset, then no artificial diffusion is added. Hence, in diffusion-dominated cases the matrices K^H and K^L are identical. For systems of hyperbolic conservation laws, the modulation parameter d_{ij} is set proportional to the spectral radius of the corresponding Roe matrix (see Example 2 below).

The time discretization of the modified scheme yields

$$(M_L - \theta \Delta t K^L) u^{n+1} = (M_L + (1-\theta)\Delta t K^L) u^n + \Delta t\, q^{n+\theta} \tag{21}$$

This fully discretized equation differs from (12) by the presence of the source term, which may take negative values. In order to prevent the violation of the positivity constraint, source terms can be linearized as proposed by Patankar [32]:

$$q = q_C + q_P u \quad \text{where } q_C \geqslant 0,\ q_P \leqslant 0 \tag{22}$$

A simple way to perform such a splitting is based on the identity

$$q = q^+ - q^- = q^+ - \left(\frac{q^-}{u}\right) u \tag{23}$$

in which q^+ is the positive part of the source term, and $-q^-$ is the negative one. Thus, we can adopt $q_C = q^+$, $q_P = -q^-/u^*$, where u^* denotes the best approximation to u currently available.

Note that the actual state u about which the source term is linearized, is yet to be specified. For $q = q^{n+\theta}$ it is natural to take $u = \theta u^{n+1} + (1-\theta)u^n$. Alternatively, we can simply linearize about $u = u^{n+1}$ regardless of the choice of θ. An advantage of this approach is that Equation (21) assumes the convenient form (12) with

$$L = M_L - \theta\Delta t K^L + \Delta t S^-, \quad S^- = \mathrm{diag}\{q^-/u^*\}$$

$$R = M_L + (1-\theta)\Delta t K^L + \Delta t S^+, \quad S^+ = \mathrm{diag}\{q^+/u^n\}$$

For this kind of splitting, the diagonal matrices S^- and S^+ engendered by the source term are seen to reinforce the properties of L and R required by the lemma.

By construction, all off-diagonal entries of the matrix L are non-positive, while those of the matrix R are non-negative. Therefore, it remains to secure the positivity of diagonal coefficients. Since the elements of M_L and the contributions of source terms (if any) are positive, this condition can always be realized by choosing the time step to be small enough. In particular, the time step for the low-order discretization of the incompressible convection–diffusion equation without source terms is bounded by

$$\Delta t \leqslant \frac{1}{1-\theta}\min_i\{-m_i/k_{ii}^L \mid k_{ii}^L < 0\} \tag{24}$$

where m_i denote the diagonal entries of the lumped mass matrix. This CFL-like condition, which follows from the positivity theorem, gives a sharp estimate of the maximum admissible time step. It can be used to steer adaptive time stepping for (semi-) explicit schemes. The upper bound depends on the degree of implicitness θ and on the ratio m_i/k_{ii}^L. Hence, excessive artificial diffusion not only degrades the accuracy of the method but also requires taking impractically small time steps.

Example 1 (*One-dimensional scalar convection*)
Let us illustrate the construction of low-order operators by a one-dimensional example. Consider the pure convection equation

$$\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = 0 \tag{25}$$

discretized on a uniform mesh of linear elements. For the sake of simplicity assume that the velocity v is constant and positive. The involved element matrices have the form

$$\hat{M}_\mathrm{L} = \frac{\Delta x}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \hat{K}^H = \frac{v}{2}\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} \tag{26}$$

After the global matrix assembly, the central difference approximation of the convective term is recovered at interior nodes:

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = -v\,\frac{u_{i+1} - u_{i-1}}{2\,\Delta x} \tag{27}$$

The minimum amount of artificial dissipation sufficient to enforce positivity is proportional to $\hat{d}_{12} = v/2$. The corresponding discrete diffusion operator restricted to one element is given by

$$\hat{D} = \frac{v}{2}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \Rightarrow \hat{K}^L = v\begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix} \tag{28}$$

The resulting low-order scheme is seen to be equivalent to the upwind finite-difference method in the interior:

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = -v\,\frac{u_i - u_{i-1}}{\Delta x} \tag{29}$$

Obviously, this is the least diffusive linear scheme which preserves positivity. The associated CFL condition reads

$$v\frac{\Delta t}{\Delta x} \leqslant \frac{1}{1-\theta} \tag{30}$$

This example demonstrates that our low-order discretization reduces to standard upwinding for pure convection in one dimension. At the same time, its derivation based on the post-processing of the discrete transport operator remains valid for arbitrary meshes and multidimensional problems. Moreover, physical diffusion (if any) is automatically detected, and the amount of artificial diffusion is reduced accordingly.

Example 2 (*One-dimensional Euler equations*)
To elucidate the derivation of positive low-order schemes for systems of hyperbolic conservation laws, we consider the one-dimensional Euler equations of gas dynamics

$$\frac{\partial U}{\partial t} + \frac{\partial F}{\partial x} = 0 \tag{31}$$

written in terms of the conservative variables and fluxes

$$U = \begin{bmatrix} \rho \\ \rho v \\ E \end{bmatrix}, \quad F = \begin{bmatrix} \rho v \\ \rho v^2 + p \\ v(E+p) \end{bmatrix} \tag{32}$$

where ρ, v, p and E represent the density, velocity, pressure and total energy of the fluid, respectively. This system is completed by specifying the equation of state relating the energy to pressure and density:

$$E = \frac{p}{\gamma - 1} + \frac{1}{2}\rho v^2 \tag{33}$$

in which γ denotes the ratio of specific heats for a polytropic gas.

Alternatively, the Euler equations may be represented in the quasi-linear form

$$\frac{\partial U}{\partial t} + A\frac{\partial U}{\partial x} = 0 \tag{34}$$

The Jacobian matrix $A = \partial F/\partial U$ is given by

$$A = \begin{bmatrix} 0 & 1 & 0 \\ (\gamma-3)v^2/2 & (3-\gamma)v & \gamma-1 \\ v[(\gamma-1)v^2/2 - h] & h-(\gamma-1)v^2 & \gamma v \end{bmatrix} \tag{35}$$

where $h = (E+p)/\rho$ is the total enthalpy.

Interestingly enough, the flux vector F is a homogeneous function of the conservative variables U, so that the following useful identity holds:

$$F = \frac{\partial F}{\partial U}U = AU \tag{36}$$

It is worth mentioning that the flux vectors F_x, F_y, F_z for compressible inviscid flow in three dimensions also possess this property [15].

The strict hyperbolicity of the Euler equations follows from the fact that A is diagonalizable with distinct real eigenvalues. Indeed, it admits the decomposition

$$A = R\Lambda R^{-1} \quad \text{where } \Lambda = \text{diag}\{v-c, v, v+c\} \tag{37}$$

is the diagonal matrix of eigenvalues, and

$$R = \begin{bmatrix} 1 & 1 & 1 \\ v-c & v & v+c \\ h-vc & v^2/2 & h+vc \end{bmatrix} \tag{38}$$

is the matrix of right eigenvectors. Here $c = \sqrt{\gamma p/\rho}$ stands for the local speed of sound.

If we apply the Galerkin discretization to the weak formulation of Equation (31) without integrating by parts, the semi-discrete problem can be cast into the form

$$M_C \frac{dU}{dt} = \sum_{j\neq i} k_{ij}(F_j - F_i) = \sum_{j\neq i} k_{ij}\hat{A}(U_j - U_i) \tag{39}$$

where $k_{ij} = \pm 1/2$ in one dimension (see above), and $\hat{A}$ is the so-called Roe matrix obtained by evaluating the Jacobian at the intermediate state [10]

$$\hat{\rho} = \sqrt{\rho_i\rho_j}, \qquad \hat{v} = \frac{\sqrt{\rho_i}v_i + \sqrt{\rho_j}v_j}{\sqrt{\rho_i}+\sqrt{\rho_j}}, \qquad \hat{h} = \frac{\sqrt{\rho_i}h_i + \sqrt{\rho_j}h_j}{\sqrt{\rho_i}+\sqrt{\rho_j}} \tag{40}$$

The density-averaged quantities $\hat{\rho}$, $\hat{v}$ and $\hat{h}$ are called the Roe mean values.

Note that the transition from the fluxes to the nodal solution values in (39) makes it possible to calculate the contributions of edges to the global finite-element matrices explicitly. The coefficients of the nine blocks C^{kl} are augmented as follows:

$$\begin{aligned} c_{ii}^{kl} &= c_{ii}^{kl} - k_{ij}\hat{a}_{kl}, \quad & c_{ij}^{kl} &= c_{ij}^{kl} + k_{ij}\hat{a}_{kl} \\ c_{ji}^{kl} &= c_{ji}^{kl} + k_{ji}\hat{a}_{kl}, \quad & c_{jj}^{kl} &= c_{jj}^{kl} - k_{ji}\hat{a}_{kl} \end{aligned} \tag{41}$$

where $\hat{a}_{kl}$ denote the entries of the 3×3 matrix $\hat{A}$ corresponding to the edge e_{ij}. The assembly process remains the same in multidimensions. A distinct advantage of this approach is that the coefficients k_{ij} are typically fixed, so that the matrices can be efficiently assembled edge by edge without resorting to numerical integration.

A usable low-order method can be constructed by adding artificial diffusion proportional to the spectral radius of the Roe matrix:

$$d_{ij} = \lambda_{\max}|k_{ij} - k_{ji}|/2 \quad \text{where } \lambda_{\max} = |\hat{v}| + \hat{c} \tag{42}$$

This scalar dissipation is to be inserted into the three diagonal blocks:

$$\begin{aligned} c_{ii}^{kk} &= c_{ii}^{kk} - d_{ij}, \quad c_{ij}^{kk} = c_{ij}^{kk} + d_{ij} \\ c_{ji}^{kk} &= c_{ji}^{kk} + d_{ij}, \quad c_{jj}^{kk} = c_{jj}^{kk} - d_{ij} \end{aligned} \tag{43}$$

In the one-dimensional case, $d_{ij} = \lambda_{\max}/2$. Thus, for scalar convection problems it reduces to the parameter value derived in Example 1.

One of the most popular upwind-biased methods for the numerical solution of the Euler equations is Roe's approximate Riemann solver, which can be implemented by using the modified flux

$$G_{ij}^{*} = \frac{F_i + F_j}{2} + \frac{1}{2}|\hat{A}|(U_j - U_i), \tag{44}$$

where

$$|\hat{A}| = \hat{R}|\hat{\Lambda}|\hat{R}^{-1}, \quad |\hat{\Lambda}| = \operatorname{diag}\{|\hat{v} - \hat{c}|, |\hat{v}|, |\hat{v} + \hat{c}|\} \tag{45}$$

instead of the centered Galerkin flux $G_{ij} = (F_i + F_j)/2$ in decomposition (10). In essence, this kind of flux difference splitting corresponds to adding tensorial artificial dissipation $|\hat{A}|(U_j - U_i)/2$, which affects both the diagonal and the off-diagonal blocks of the discrete transport operator. Roe's Riemann solver constitutes a good low-order method *per se* but it results in considerable overhead costs and is not to be recommended for the use in the FEM-FCT environment.

7. GENERALIZED FEM-FCT FORMULATION

Another cornerstone of the FEM-FCT algorithm is the linear high-order method. A variety of finite-element schemes employing streamline diffusion to stabilize the troublesome convective terms were proposed in References [33–35]. For instance, Taylor–Galerkin methods attribute this stabilization to high-order time derivatives in the Taylor series expansion. This leads to improved time-stepping schemes which are combined with the standard Galerkin spatial discretization. The most popular representative of such stabilized methods is the well-known Lax–Wendroff scheme. An investigation of the modified equation for its finite-element counterpart reveals that the introduced dissipation just counterbalances the intrinsic negative diffusion which renders the explicit Euler/Galerkin scheme unstable for pure convection problems. For an in-depth study of the Lax–Wendroff and higher-order Taylor–Galerkin methods the reader is referred to References [35, 36].

While the stabilization of convective terms is mandatory for the fully explicit time discretization, implicit finite-element schemes based on the Crank–Nicolson and backward Euler time stepping are unconditionally stable. Therefore, they can be used as a high-order method without being stabilized by streamline diffusion. Linear discretization schemes of this type are of little use, because they are prone to non-physical oscillations. The incorporation of a flux limiter makes it possible to get rid of oscillations in the framework of non-linear Crank–Nicolson/FCT and backward Euler/FCT methods.

The high-order transport operator can be transformed into a low-order one as explained in the previous section. For simplicity, let us omit the (linearized) source terms. The resulting methods of high and low order discretized in time by the standard θ-scheme are related by the formula

$$(M_{\mathrm{L}} - \theta \Delta t K^L) u^H = (M_{\mathrm{L}} + (1-\theta)\Delta t K^L) u^n + F(u^H, u^n) \tag{46}$$

where the antidiffusion responsible for high spatial accuracy is given by

$$F(u^H, u^n) = -(M_{\mathrm{C}} - M_{\mathrm{L}}) \Delta u^H - \Delta t (K^L - K^H)[\theta u^H + (1-\theta) u^n] + \Delta t\, S u^n \tag{47}$$

Here the superscript H refers to the high-order solution, and S stands for the streamline diffusion operator which is required only for the fully explicit scheme. If the antidiffusive term $F(u^H, u^n)$ is omitted, then the positive low-order scheme is recovered, whereas retaining it yields the original high-order method.

It can readily be seen that all the matrices in (47) represent discrete (anti-) diffusion operators and thereby lend themselves to decomposition into fluxes

$$\begin{aligned} f_{ij} &= -m_{ij}(\Delta u_j^H - \Delta u_i^H) - \Delta t\, d_{ij}[\theta(u_j^H - u_i^H) + (1-\theta)(u_j^n - u_i^n)] \\ &\quad + \Delta t\, s_{ij}(u_j^n - u_i^n), \quad f_{ji} = -f_{ij}, \quad i<j \end{aligned} \tag{48}$$

These raw antidiffusive fluxes offset the errors induced by mass lumping, ‘upwinding’ and first-order time discretization (for the explicit scheme). The coefficients m_{ij}, d_{ij} and s_{ij} denote the entries of the consistent mass matrix, artificial diffusion and streamline diffusion operators, respectively.

The crux of the FCT procedure consists in adding as much antidiffusion as possible without generating non-physical undershoots and overshoots. The flux-corrected version of scheme (46) can be written in the form

$$m_i u_i^{n+1} - \theta \Delta t \sum_j k_{ij}^L u_j^{n+1} = m_i \tilde{u}_i + \sum_{j \neq i} \alpha_{ij} f_{ij}, \quad \alpha_{ji} = \alpha_{ij} \tag{49}$$

where α_{ij} denote the correction factors (see the next section), while $\tilde{u}$ represents the positivity-preserving solution to the explicit subproblem

$$m_i \tilde{u}_i = m_i u_i^n + (1-\theta)\Delta t \sum_j k_{ij}^L u_j^n \tag{50}$$

In essence, $\tilde{u}$ corresponds to an intermediate solution computed at the time instant $t^{n+1-\theta}$ by the explicit low-order scheme. It reduces to the old solution u^n for the backward Euler method and to the low-order solution u^L for the forward Euler method.

It is obvious that the success of the FCT algorithm depends on the positivity of the provisional solution $\tilde{u}$ and on the choice of the correction factors α_{ij}. For $\tilde{u}$ to be positive, the time step must satisfy the CFL-like condition (24) unless the scheme is fully implicit. As long as the left-hand side operator is an M-matrix, our positivity criteria ensure that scheme (49) can be rendered positive by tuning the correction factors.

The new family of FEM-FCT schemes distinguishes itself in that it is applicable to explicit and implicit time discretizations alike. The fully explicit scheme is consistent with the standard FCT methodology. Note that implicit schemes require solving *two* non-symmetric linear systems per time step: one for the high-order solution (which is needed to compute the antidiffusive fluxes) and one for the final solution. Nevertheless, implicit methods are typically more efficient than explicit ones because larger time steps can be taken. If iterative solvers are employed, the high-order solution provides a reasonable initial guess for the final solution.

The majority of practical applications are described by *non-linear* conservation laws. In this case, the matrices K^H and K^L depend on the unknown solution, so that additional outer iterations are necessary for implicit schemes. It will be noted that the linearization of the problem using extrapolation in time can entail a loss of mass and alter the shock speed. The simplest iterative treatment of non-linearities is afforded by a fixed point defect correction method. If we consider an abstract non-linear system of the form

$$A(u)u = b \tag{51}$$

then the basic non-linear iteration can be formulated as

$$u^{(l+1)} = u^{(l)} - [C(u^{(l)})]^{-1}(A(u^{(l)})u^{(l)} - b) \tag{52}$$

where l is the outer iteration counter, and C is a suitably chosen 'preconditioner' (an approximate Fréchet derivative) which should be easy to invert. The iteration process is terminated when the relative solution changes are small enough or l exceeds a given limit. As a rule, the 'inversion' of C is also performed by some iterative procedure. Hence, a certain number of inner iterations per cycle is required. It is worth mentioning that the problem does not have to be solved very accurately at each outer iteration. A moderate improvement of the residual (1–2 digits) is sufficient to obtain a good overall accuracy.

For a non-linear problem of form (46), it is reasonable to use the low-order operator as preconditioner:

$$C(u^{(l)}) = M_{\mathrm{L}} - \theta \Delta t K^L(u^{(l)})$$

This yields an iterative FEM-FCT algorithm, whereby the approximate solution and the transport operator are successively updated as follows:

$$(M_{\mathrm{L}} - \theta \Delta t K^L(u^{(l)}))u^{(l+1)} = (M_{\mathrm{L}} + (1-\theta)\Delta t K^L(u^n))u^n + F(u^{(l)}, u^n) \tag{53}$$

The last term is composed from the (limited) antidiffusive fluxes. Flux correction can be performed after each outer iteration or just once after the high-order solution has converged. In either case, positivity of the numerical solution is guaranteed.

8. LIMITING STRATEGY

The flux limiter is a key element of the FEM-FCT paradigm. By varying the correction factors α_{ij} between zero and unity, it is possible to obtain the diffusive low-order solution, the oscillatory high-order solution or anything in-between. Clearly, it is desirable to utilize the antidiffusive terms to the greatest extent possible without generating wiggles and violating the positivity constraint. Following Löhner *et al.* [4], we employ Zalesak's limiter to select the 'optimal' correction factors. Kuzmin and Turek [21] demonstrated that this multidimensional limiter can be readily extended to implicit time discretizations. This enables us to adopt a unified limiting strategy for explicit and implicit schemes. The ins and outs of the flux correction process are elucidated below.

8.1. Prelimiting step

Let us start with an optional but important component of the flux limiter. It turns out that explicit FCT schemes may benefit from canceling all antidiffusive fluxes directed down the gradient of $\tilde{u}$:

$$f_{ij} := 0 \quad \text{if} \quad f_{ij}(\tilde{u}_i - \tilde{u}_j) < 0 \tag{54}$$

This test should be applied *before* the flux correction step. Its purpose is to ensure that the flux does not smooth the low-order solution. To put it another way, an antidiffusive flux is not allowed to be diffusive. When this happens, small-scale numerical ripples can be produced even though the solution remains positive. Hence, the limiter is positivity preserving but not monotonicity preserving [37].

The prelimiting of antidiffusive fluxes can be traced back to the original SHASTA scheme [1]. Zalesak also mentioned this approach in passing [2] but did not promote its regular use. He argued that the majority of antidiffusive fluxes act to steepen the gradient, while the effect of amendment (54) is minimal and cosmetic in nature. This remark has discouraged the use of prelimiting in FCT algorithms based on Zalesak's multidimensional limiter. Apparently, this is not the sole reason why this step is missing in the FEM-FCT procedure of Löhner *et al.* [4]. The replacement of antidiffusive fluxes by element contributions makes the prelimiting impossible to carry out for multidimensional problems. Only the comeback of a flux-based formulation enables us to apply this technique in the finite-element context.

DeVore [37] has rediscovered the preprocessing of antidiffusive fluxes as a way to achieve monotonicity and demonstrated that it can lead to a dramatic qualitative improvement of dynamic simulation results. Even for simple test problems with discontinuous solutions, remarkable 'esthetic' improvements are observed [21]. Therefore, the prelimiting step is to be included in explicit FCT algorithms. In our experience, it remains relevant also for the implicit schemes presented in this paper.

8.2. Zalesak's limiter

Zalesak's limiter remains the only genuinely multidimensional high-resolution scheme available to date. In order to elucidate its operating principles and internal structure, we restrict ourselves to the fully explicit case in which $\tilde{u} = u^L$. In the sequel, we will show that the same limiting strategy can be applied to implicit FEM-FCT schemes. The basic ingredients of Zalesak's limiter are sketched in Figure 1 using the close-up of a uniform one-dimensional mesh.

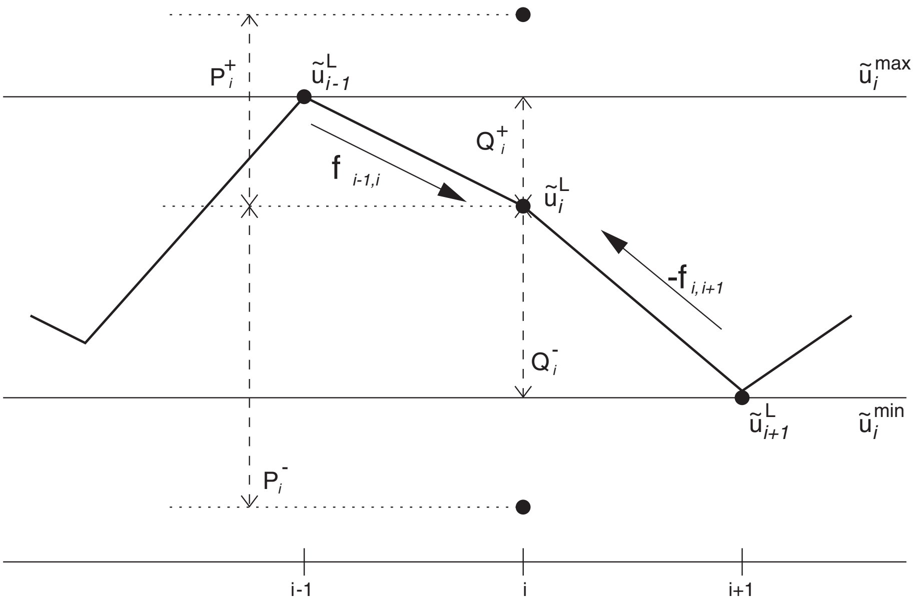


Figure 1. Zalesak's limiter in one dimension.

Let $\tilde{u}_i^{\max}$ and $\tilde{u}_i^{\min}$ denote the maximum and minimum solution values at the stencil S_i which consists of node i and its nearest neighbours:

$$\tilde{u}_i^{\substack{\max\\\min}} = \substack{\max\\\min}\, \tilde{u}_j, \quad j \in S_i \tag{55}$$

In the classical theory of explicit FCT schemes [1], these quantities represent the upper and lower bounds for the nodal values of the final solution. The task of the limiter is to guarantee that the antidiffusive fluxes cannot conspire to create new extrema or accentuate already existing ones.

In fact, Zalesak proposed that both u^n and u^L be involved in the computation of solution bounds. The screening of the old solution was intended to alleviate 'peak clipping' inherent to the FCT limiter. This modification was shown to produce the desired effect for a number of test configurations. At the same time, the use of outdated information on the magnitude of local maxima and minima may lead to the formation of numerical ripples in other situations. For instance, physical extrema may decay with time due to negative source terms or a variable velocity field. In this case, the resurrection of old peaks would result in an overshoot. Hence, it is prudent to search for extrema only in the low-order solution as in the SHASTA scheme of Boris and Book [1].

Zalesak's limiter can be elucidated as follows. The solution value at node i is affected by incoming antidiffusive fluxes from the neighbouring nodes. In the worst case, these fluxes have the same sign and threaten to generate or enhance a local extremum. Let us denote the sum of all positive/negative contributions to node i by

$$P_i^{\pm} = \frac{1}{m_i} \sum_{j \neq i} \substack{\max\\\min}\{0, f_{ij}\} \tag{56}$$

The maximum/minimum increment which node i is allowed to accept is given by the distance to the local extremum:

$$Q_i^{\pm} = \tilde{u}_i^{\substack{\max\\ \min}} - \tilde{u}_i \tag{57}$$

A geometric interpretation of the auxiliary quantities $P_i^{\pm}$ and $Q_i^{\pm}$ is presented in Figure 1. It can be seen that unlimited antidiffusive fluxes acting in concert may force the solution beyond the physically admissible range, thus leading to spurious overshoots and undershoots. The maximum percentage of a flux into node i which can be retained reads

$$R_i^{\pm} = \begin{cases} \min\{1, Q_i^{\pm}/P_i^{\pm}\} & \text{if } P_i^{\pm} \neq 0 \\ 0 & \text{if } P_i^{\pm} = 0 \end{cases} \tag{58}$$

Since the nodes exchange mass on a bilateral basis, a positive flux f_{ij} into node i is always balanced by a negative flux $f_{ji} = -f_{ij}$ into node j and *vice versa*. For the final solution to remain within the bounds at both nodes, we must check the sign of the flux and take the minimum of the nodal correction factors:

$$\alpha_{ij} = \begin{cases} \min\{R_i^+, R_j^-\} & \text{if } f_{ij} \geqslant 0 \\ \min\{R_j^+, R_i^-\} & \text{if } f_{ij} < 0 \end{cases} \tag{59}$$

This choice of the coefficients α_{ij} is safe enough to guarantee that the constraint $\tilde{u}_i^{\min} \leqslant u_i^{n+1} \leqslant \tilde{u}_i^{\max}$ is satisfied at all nodes. Hence, the final solution will preserve positivity if the low-order one does. The resulting flux limiter is independent of the number of spatial dimensions and can easily be implemented as a 'black-box' routine which computes the correction factors given an array of antidiffusive fluxes for each pair of neighbouring nodes.

8.3. Positivity proof

In order to prove the positivity of Zalesak's limiter for arbitrary time-stepping, we have to apply the mathematical theory of positivity-preserving schemes rather than the heuristic considerations presented above. The crucial point of the proof is the following representation of the right-hand side of our FEM-FCT schemes:

$$\text{RHS} = m_i\tilde{u}_i + \sum_{j\neq i} \alpha_{ij} f_{ij} = m_i\tilde{u}_i + c_i Q_i, \quad c_i = \frac{\sum_{j\neq i} \alpha_{ij} f_{ij}}{Q_i} \tag{60}$$

where the intermediate solution $\tilde{u} = u^L(t^{n+1-\theta})$ depends on the concrete time-stepping scheme, and the multiplier Q_i is chosen to be

$$Q_i = \begin{cases} Q_i^+ = \tilde{u}_i^{\max} - \tilde{u}_i & \text{if } \sum_{j\neq i} \alpha_{ij} f_{ij} > 0 \\ Q_i^- = \tilde{u}_i^{\min} - \tilde{u}_i & \text{if } \sum_{j\neq i} \alpha_{ij} f_{ij} < 0 \\ 1 & \text{if } \sum_{j\neq i} \alpha_{ij} f_{ij} = 0 \end{cases} \tag{61}$$

In accordance with the FCT theory, all antidiffusive fluxes which try to accentuate a local maximum or minimum must be completely cancelled:

$$\alpha_{ij} = 0 \quad \text{if} \ \tilde{u}_i = \tilde{u}_i^{\max}, \quad f_{ij} > 0 \quad \text{or} \quad \tilde{u}_i = \tilde{u}_i^{\min}, \quad f_{ij} < 0 \tag{62}$$

This relation implies that $Q_i \neq 0$, so that no division by zero takes place in (60).

Note that the auxiliary coefficient c_i is always non-negative. Let the local extremum $\tilde{u}_i^{\substack{\max \\ \min}}$ be attained at some node k adjacent to node i. Then the antidiffusive term exhibits a local extremum diminishing structure, and we obtain

$$\text{RHS} = m_i \tilde{u}_i + c_i(\tilde{u}_k - \tilde{u}_i) = (m_i - c_i)\tilde{u}_i + c_i \tilde{u}_k, \quad c_i \geqslant 0 \tag{63}$$

According to our lemma, both explicit and implicit FEM-FCT schemes of form (49) will preserve positivity as long as $m_i \geqslant c_i$. This important observation frames a general rule for the selection of the correction factors α_{ij}.

It remains to show that Zalesak's limiter does possess the desired properties. The side condition (62) is automatically satisfied, since $Q_i^{\pm} = 0$ implies $R_i^{\pm} = 0$ and $\alpha_{ij} = 0$. Hence, any enhancement of local extrema is neutralized by the limiter. Furthermore, the following estimate holds:

$$\sum_{j \neq i} \alpha_{ij} f_{ij} \leqslant \sum_{j \neq i} \alpha_{ij} \max\{0, f_{ij}\} \leqslant m_i R_i^+ P_i^+ \leqslant m_i Q_i^+ \tag{64}$$

In much the same way, it can be verified that

$$\sum_{j \neq i} \alpha_{ij} f_{ij} \geqslant \sum_{j \neq i} \alpha_{ij} \min\{0, f_{ij}\} \geqslant m_i R_i^- P_i^- \geqslant m_i Q_i^- \tag{65}$$

This proves that the corrected antidiffusive fluxes satisfy the constraint $m_i \geqslant c_i$. Therefore, the right-hand side poses no hazard to positivity. Recall that the matrix of the left-hand side was assumed to be an M-matrix. Therefore, the positivity of u^n is inherited by u^{n+1} provided that the time step is small enough. According to (24), the backward Euler time stepping is unconditionally positive. The Crank–Nicolson scheme is subject to the CFL-like condition for the auxiliary problem (50), but the admissible Courant numbers are twice as large as those for the fully explicit scheme.

8.4. Limiting for systems of equations

Despite remarkable progress made in the development of FCT schemes for scalar equations, the issue of flux correction for systems of hyperbolic conservation laws remains largely unresolved. Of course, it is possible to use an operator-splitting approach, whereby the coupled equations are solved in a segregated manner within a block-iterative loop. However, independent limiting of intricately related variables was found to produce unsatisfactory results in some cases. This has led the FCT community to devise a common flux limiter for the whole system by merging individual limiters for different variables. The resulting improvements in the numerical solutions can be attributed to the fact that the phase errors for the involved equations become synchronized [4]. Nevertheless, there is still a large degree of empiricism in the construction of such limiters, and their performance is highly problem dependent.

Flux correction for the system of Euler equations was addressed by Löhner [4, 9]. He singled out the following approaches to the design of a synchronized limiter:

- Use of a limiter for a single 'indicator variable'.
- Use of the minimum of limiters obtained for some group of variables.

The combination of limiters for the density and energy is to be recommended for the treatment of highly dynamic flows characterized by propagating and/or interacting shock waves. The minimum of correction factors for the density and pressure is also claimed to perform fairly well, especially for steady-state problems [4]. As a matter of fact, the synchronized limiter may be formulated in terms of variables other than those being solved for. A general algorithm for the construction of a flux limiter based on arbitrary derived quantities is presented in Reference [9].

9. SUMMARY OF THE ALGORITHM

As we have seen, Zalesak's limiter can be integrated into the generalized FEM-FCT formulation and applied to a wide range of CFD problems described by (systems of) conservation laws of form (1). The proposed high-resolution finite-element schemes can be implemented on arbitrary unstructured grids using the conventional or edge-based data structure. The main algorithmic steps can be summarized as follows:

1. Discretize the governing equation by a high-order finite-element method.
2. Perform mass lumping and eliminate negative off-diagonal entries of the transport operator to construct the associated low-order scheme.
3. For $\theta < 1$, examine the diagonal entries of the low-order operator and adapt the time step so as to comply with the positivity condition.
4. Advance the solution in time by the high-order scheme to obtain u^H.
5. Assemble the raw antidiffusive fluxes f_{ij} for each pair of nodes.
6. Compute the positivity-preserving auxiliary solution $\tilde{u} = u^L(t^{n+1-\theta})$.
7. Cancel all antidiffusive fluxes directed down the gradient of $\tilde{u}$.
8. Apply Zalesak's limiter to calculate the correction factors α_{ij}.
9. Add the contribution of the limited antidiffusive fluxes $\alpha_{ij} f_{ij}$ to the right-hand side of the low-order scheme.
10. For $\theta = 0$, scale the right-hand side by the diagonal matrix M_{L}^{-1}. Otherwise, solve the linear system for the end-of-step solution u^{n+1}.

In the non-linear version of the FEM-FCT algorithm the high-order solution u^H is replaced by the last iterate $u^{(l)}$ so that just one linear system per outer iteration has to be solved. Furthermore, only the low-order matrix $C(u^{(l)})$ needs to be assembled and stored.

A remark is in order regarding the iterative solution of linear systems. Explicit schemes do not require any advanced linear algebra tools, since the consistent mass matrix can be efficiently 'inverted' e.g. by just a few Jacobi-like iterations using the lumped mass matrix as a preconditioner [5]. Similarly, for relatively small time steps the non-symmetric linear systems engendered by implicit schemes can be solved by BiCGSTAB or multigrid methods with basic components like Jacobi, Gauß-Seidel or SOR smoothers. However, the large time steps afforded by the unconditionally positive backward Euler/FCT method may cause a severe

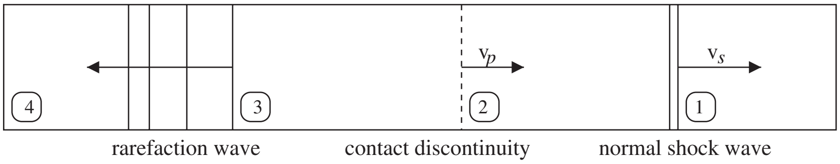


Figure 2. Flow structure in the shock tube.

deterioration of the matrix for the high-order system, so that an iterative method may fail to converge. This can be rectified by resorting to defect correction, which approximates the original matrix by a well-behaved preconditioner. The M-matrix properties of the low-order operator make it particularly amenable to iterative solution. The robustness of the solver can be further enhanced by using appropriate renumbering techniques in conjunction with the ILU decomposition as a smoother/preconditioner.

10. NUMERICAL EXAMPLES

In the examples which follow, we study the behaviour of the implicit Crank–Nicolson (CN/FCT) and backward Euler (BE/FCT) schemes. Strictly speaking, the numerical solutions corresponding to the former method were obtained using the value $\theta = 0.55$ of the implicitness parameter. This is a standard trick which enhances the stability of the Crank–Nicolson discretization without incurring any appreciable loss of accuracy. The fully explicit Lax–Wendroff (LW/FCT) method produces similar results. Many other test problems featuring both smooth and discontinuous solutions were considered in the preceding paper [21] which should also be consulted for a discussion of certain numerical difficulties resulting from an improper treatment of outflow boundaries.

10.1. Shock tube problem

A simple one-dimensional example that is frequently used to evaluate compressible flow solvers is the shock tube problem of gas dynamics [23, 38]. Its physical prototype is a closed tube initially filled with a quiescent gas separated by a membrane into two regions. A higher gas pressure is maintained on the left of the tube than on the right. The removal of the membrane brings about a net motion of gas in the direction of lower pressure. Provided that the gas is distributed uniformly across each cross-section of the tube, the evolution of the flow is described by the one-dimensional Euler equations (31).

The flow structure sketched in Figure 2 is characterized by three distinct waves travelling with different speeds and delimiting regions in which the state variables are constant. After the membrane is abruptly removed at time $t = 0$, a normal shock wave sets off for the region of lower pressure with velocity v_s satisfying the Rankine–Hugoniot conditions. All of the primitive variables are discontinuous across the shock. The pressure jump propels the mass in the same direction with velocity v_p. The moving interface between the regions of different densities but constant velocity and pressure represents a contact discontinuity. Finally, a rarefaction wave propagates in the opposite direction providing a smooth transition to the original values of the state variables in the region of high pressure. In fact, this flow structure

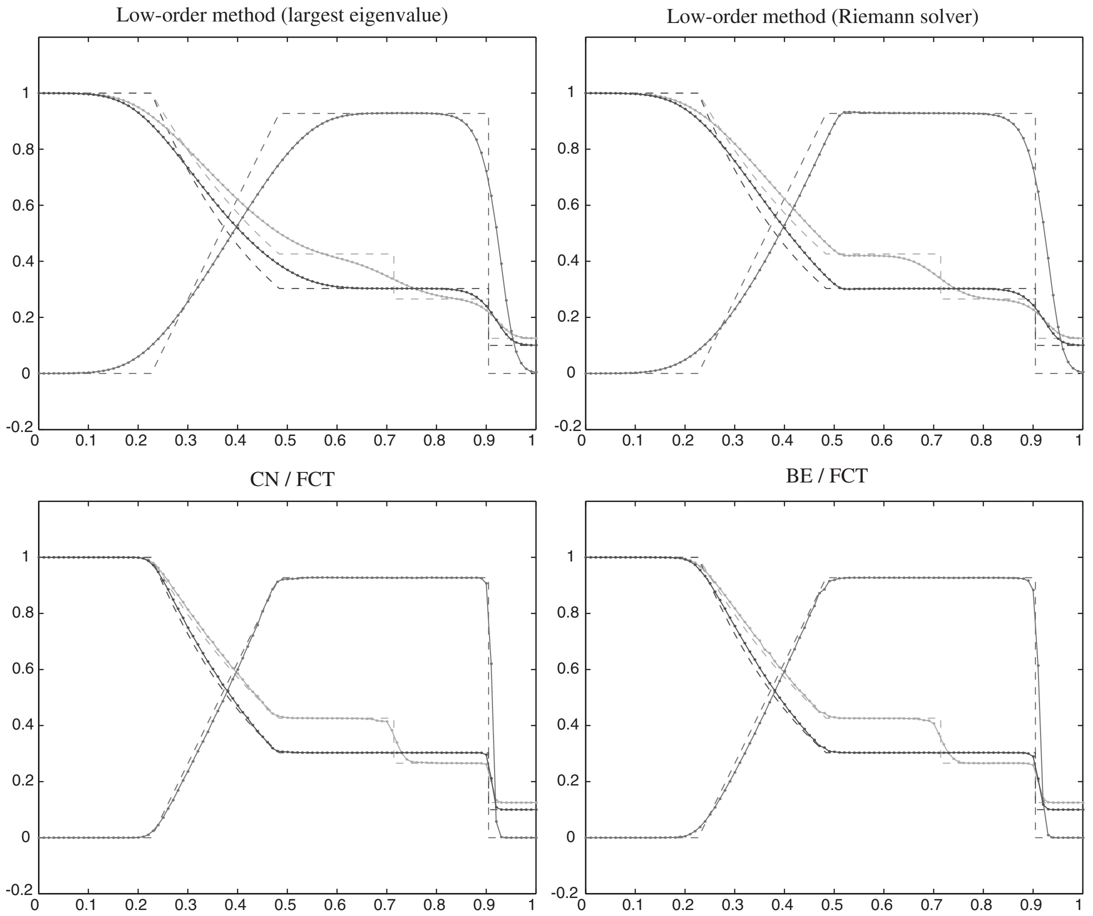


Figure 3. Shock tube problem. Numerical solutions at $t = 0.231$.

prevails only until the shock wave impinges on the right lid of the tube or the rarefaction wave reaches the left lid. If these phenomena are to be captured, it is necessary to deal with reflections of shocks and/or rarefaction waves.

Let us consider the following initial data for the Riemann problem:

$$\begin{bmatrix} \rho_L \\ v_L \\ p_L \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.0 \\ 1.0 \end{bmatrix} \quad \text{for } x \in [0, 0.5], \qquad \begin{bmatrix} \rho_R \\ v_R \\ p_R \end{bmatrix} = \begin{bmatrix} 0.125 \\ 0.0 \\ 0.1 \end{bmatrix} \quad \text{for } x \in (0.5, 1]$$

The numerical solutions displayed in Figure 3 were computed on a uniform mesh of 100 linear finite elements with a fixed time step $\Delta t = 10^{-3}$. In all diagrams, the dotted line designates the exact solution, which was obtained using the technique presented in Reference [39]. The snapshots correspond to the time instant $t = 0.231$. The upper plots show the results produced by the low-order methods based on the spectral radius of the Roe matrix (left) and on Roe's approximate Riemann solver (right). Even though the former strategy leads to a slightly stronger smearing, both solutions are of comparable quality and reproduce the flow behaviour fairly well without generating oscillations.

In the FEM-FCT framework, scalar dissipation proportional to the largest eigenvalue appears to perform better than the flux difference splitting or the brute-force addition of constant artificial diffusion. This can be explained by the already mentioned fact that some (but not too much) extra diffusion is beneficial as long as it can be removed in the antidiffusive step. For this reason, it is not advisable to employ scalar upwinding for individual variables constituting the hyperbolic system. In our experience, this segregated approach to the construction of the low-order scheme yields a sensible low-order solution, but the flux-corrected version may be polluted by some minor ripples.

Having adopted the artificial diffusion proportional to the spectral radius of the Roe matrix, we ought to investigate the ability of the flux limiter to curtail it. The numerical solutions produced by the CN/FCT and BE/FCT schemes are presented at the bottom of Figure 3. Both methods are seen to provide a sharp resolution of discontinuities while keeping the solution free of oscillations. The first-order accurate BE/FCT scheme (right) tends to be diffusive at large time steps. This deficiency is partially compensated by the unconditional positivity of the fully implicit scheme, which makes it a natural choice for steady-state computations and/or CFD solvers incorporating adaptive mesh refinement (see the two-dimensional example below).

Following Löhner *et al.* [4], we applied the synchronized flux limiter defined by the minimum of the correction factors for the density and energy. The results produced by the minimum of the limiters for all three conservative variables are slightly more diffusive than those shown in Figure 3. Nevertheless, they still look quite reasonable, which is not the case if constant diffusion is used to construct the low-order scheme.

10.2. Rotation of a slotted cylinder

Our first two-dimensional example deals with the solid body rotation of a slotted cylinder. This classical benchmark problem, which was proposed by Zalesak [2] for testing transport algorithms, turns out to be rather challenging because of its discontinuities and small-scale features.

In our simulations, the initial data was taken to be

$$u(x,y,0)=\begin{cases}1 & R<1/3 \text{ and } (|x|>0.05 \text{ or } y>0.5)\\ 0 & \text{otherwise}\end{cases}$$

where $R=\sqrt{x^2+(y-1/3)^2}$. The cylinder defined thereby is exposed to the non-uniform velocity field $\mathbf{v}=(-y,x)$ and undergoes a counterclockwise rotation about the centre of the square domain $(-1,1)\times(-1,1)$. The local Courant number increases with distance from the origin. Homogeneous Dirichlet boundary conditions are prescribed on the inflow parts of the boundary where the normal velocity is directed into the domain.

The transport equation is discretized in space on a uniform mesh of 129×129 grid points. Figure 4 shows the numerical solutions produced by the CN/FCT and BE/FCT schemes using quadrilateral Q1 elements (left) or triangular P1 elements (right) at the time instant $t=2\pi$ which corresponds to one full revolution of the cylinder. Note that the exact solution coincides with the initial data in this case. Both FEM-FCT methods under consideration

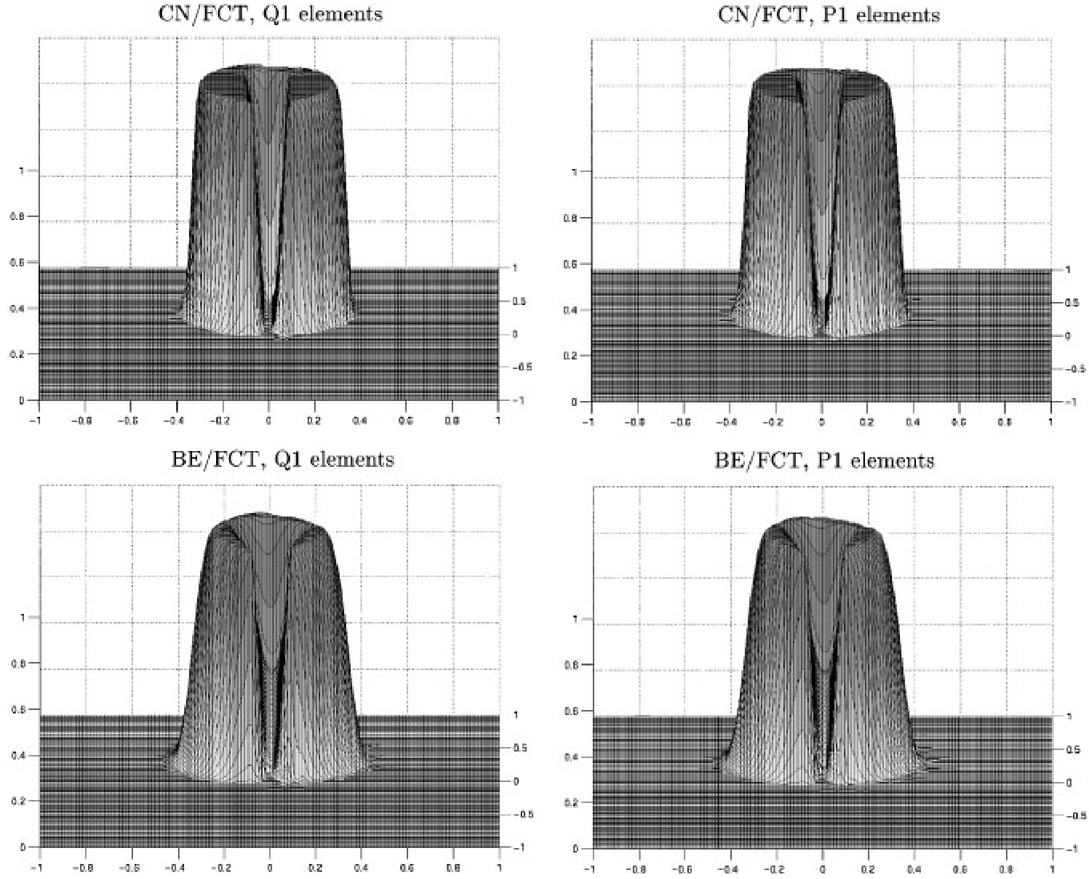


Figure 4. Rotation of a slotted cylinder. FEM-FCT solution at $t = 2\pi$.

succeed in getting rid of non-physical oscillations. As a matter of fact, the prelimiting of antidiffusive fluxes has proved its worth for this test. If it is omitted, the computational results are contaminated by innocuous but ugly ripples.

No appreciable differences were observed in the performance of quadrilateral and triangular elements (see Figure 4). The employed time step $\Delta t = 2.5 \times 10^{-3}$ was deliberately chosen relatively large to provide a fair comparison between the first-order accurate and second-order accurate time stepping. While the CN/FCT scheme resolves the discontinuity very well, the BE/FCT method brings about a considerable erosion of the cylinder. At the same time, the bridge is preserved, and the fill-in of the slot is insignificant, since it is located in the low Courant number zone, where the accuracy of the backward Euler method is sufficient.

As the time step is refined beyond $\Delta t = 10^{-3}$, the impact of the temporal error diminishes, and the numerical solutions delivered by the CN/FCT and BE/FCT schemes become virtually indistinguishable. In practice, it would be wasteful to use an implicit scheme with such small time steps because the computational cost per time step is rather high. Indeed, it is the ability to handle large Courant numbers which makes implicit methods attractive in the first place.

The above example serves as an evidence that genuinely time-dependent problems should be treated explicitly for accuracy reasons. However, this does not undermine the utility of unconditionally positive implicit schemes, which lend themselves to an efficient treatment of less dynamic applications.

10.3. Convection–diffusion of a Gaussian hill

In order to assess the numerical dissipation introduced by our FEM-FCT schemes, let us consider the rotation of a two-dimensional Gaussian hill being gradually smeared by diffusion as it orbits the origin. The computational domain and the velocity field are the same as in the previous example.

In the rotating Lagrangian reference frame, the governing equation

$$\frac{\partial u}{\partial t} + \mathbf{v} \cdot \nabla u = \varepsilon \Delta u$$

reduces to a pure diffusion problem which can be solved analytically. The exact solution we are after is a normal distribution function

$$u(x, y, t) = \frac{1}{4\pi\varepsilon t}\, \mathrm{e}^{-r^2/4\varepsilon t}, \quad r^2 = (x - \hat{x})^2 + (y - \hat{y})^2$$

where $\hat{x}$ and $\hat{y}$ denote the time-dependent peak co-ordinates

$$\hat{x}(t) = \hat{x}(0)\cos t - \hat{y}(0)\sin t, \quad \hat{y}(t) = -\hat{x}(0)\sin t + \hat{y}(0)\cos t$$

The initial condition is given by the Dirac delta function

$$u(x, y, 0) = \delta(r_0)$$

Naturally, it is not possible to specify a delta function as initial condition in a finite-element code. Instead, it is reasonable to concentrate the whole mass at a single node. The integral of a discrete function over the domain Ω can be computed as the sum of nodal values multiplied by the entries of the lumped mass matrix:

$$\int_\Omega u_h \,\mathrm{d}\mathbf{x} = \int_\Omega \sum_i u_i \varphi_i \,\mathrm{d}\mathbf{x} = \sum_i u_i m_i$$

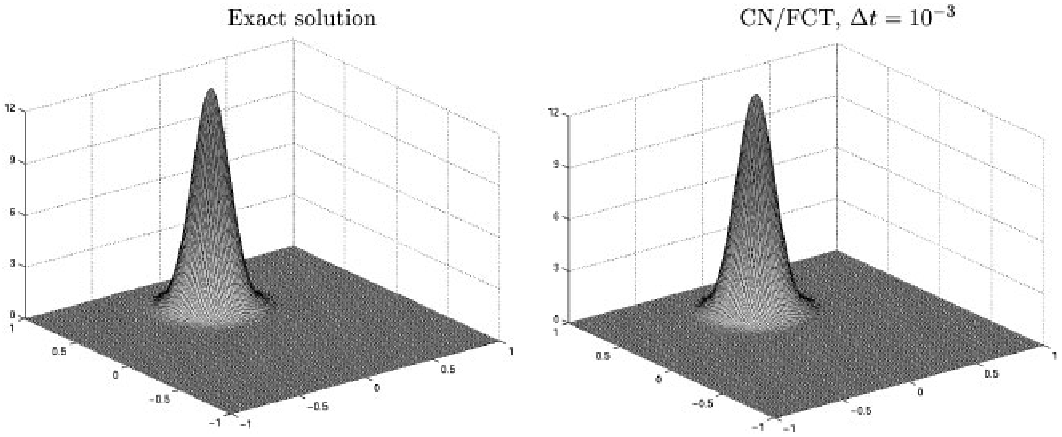
The total mass of a delta function equals unity. Hence, we should find the node i_0 closest to the peak location $(\hat{x}_0, \hat{y}_0)$ and set $u_{i_0}^0 = 1/m_{i_0}$, $u_i^0 = 0$, $i \neq i_0$.

Furthermore, the actual co-ordinates of the peak may differ from those presented above. They can be calculated as the mathematical expectation of the centre of mass under the probability distribution defined by the numerical solution:

$$\hat{x}_h(t) = \int_\Omega x u_h(x, y, t) \,\mathrm{d}\mathbf{x}, \quad \hat{y}_h(t) = \int_\Omega y u_h(x, y, t) \,\mathrm{d}\mathbf{x}$$

The quality of approximation may be assessed by considering the standard deviation

$$\sigma_h^2(t) = \int_\Omega r_h^2 u_h(x, y, t) \,\mathrm{d}\mathbf{x}, \quad r_h^2 = (x - \hat{x}_h)^2 + (y - \hat{y}_h)^2$$


Exact solution

CN/FCT, $\Delta t = 10^{-3}$

Figure 5. Convection–diffusion of a Gaussian hill, 128 × 128 Q_1-elements.

which quantifies the rate of smearing caused by artificial diffusion and equals $\sigma^2 = 4\varepsilon t$ for the exact solution. The dissipative properties of a discretization scheme are characterized by the difference between the exact and numerical values of the variance [40].

Let the initial peak be located at the point $(0, 1/2)$ and take the diffusion coefficient to be $\varepsilon = 10^{-3}$. The exact and numerical solutions after one complete revolution $(t = 2\pi)$ are presented in Figure 5. It is instructive to examine the dependence of the numerical variance on the employed time step. Figure 6 illustrates the behaviour of the relative error $\Delta\sigma_{\rm rel} = \sigma_h^2/(4\varepsilon t) - 1$ produced by the BE/FCT and CN/FCT schemes for time steps in the range 10^{-3}–10^{-2}. The backward Euler method is only first-order accurate in time. Therefore, it proves extremely diffusive at large time steps. The amount of numerical diffusion decreases linearly as the time step is refined, and the accuracy approaches that of the second-order-accurate Crank–Nicolson scheme.

10.4. Steady-state convection–diffusion

As we have seen, the fully implicit BE/FCT scheme is quite diffusive for transient transport problems. At the same time, it appears to be very attractive as an iterative solver for (quasi-) steady-state convection–diffusion equations. Indeed, the steady-state solution can be obtained by applying an FEM-FCT method to the associated time-dependent problem. Possible nonlinearities can be treated in the same iterative loop. The temporal accuracy of the method does not matter in this case, since the time step is merely an artificial parameter which determines the convergence rates. In fact, local time stepping can be employed. As long as the accuracy of the converged solution depends entirely on the spatial discretization, it is expedient to choose the time steps as large as possible, so as to reduce the computational cost. This makes explicit schemes non-competitive, since they must satisfy a restrictive CFL condition. Moreover, numerical solutions produced e.g. by the Lax–Wendroff method are affected by the streamline diffusion depending on the artificial time step. Hence, steady-state problems call for an implicit treatment.

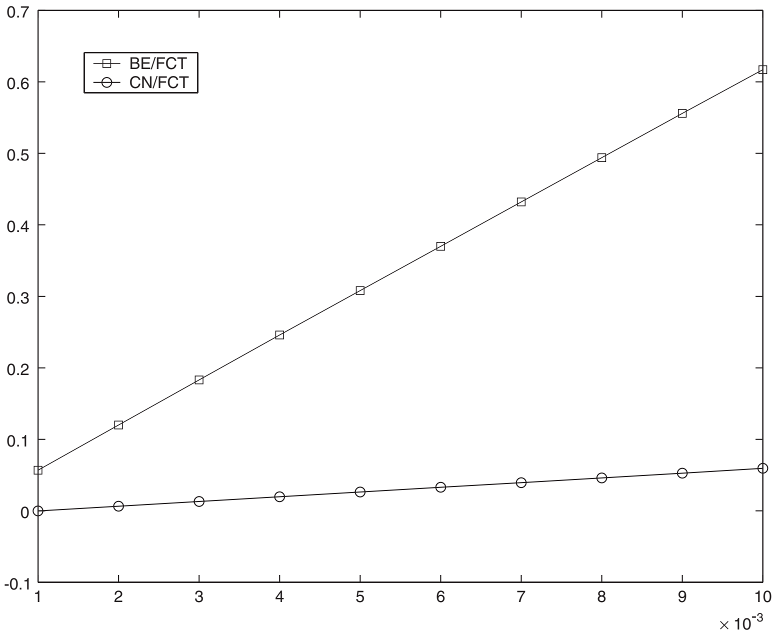


Figure 6. Relative variance error vs the time step.

Let us illustrate the advantages of the BE/FCT method by a two-dimensional steady-state example. Consider the singularly perturbed convection–diffusion equation

$$\mathbf{v}\cdot\nabla u-\varepsilon\Delta u=0 \quad \text{in } \Omega=(0,1)\times(0,1)$$

where $\mathbf{v}=(\cos 10^{\mathrm{o}},\sin 10^{\mathrm{o}})$ and $\varepsilon=10^{-3}$. The concomitant boundary conditions read

$$\frac{\partial u}{\partial y}(x,1)=0, \quad u(x,0)=u(1,y)=0, \quad u(0,y)=\begin{cases}1 & y\geqslant 0.5\\ 0 & y<0.5\end{cases}$$

The solution to this elliptic problem is characterized by the presence of a sharp front next to the line $x=1$. The boundary layer develops because the solution of the reduced problem $(\varepsilon=0)$ does not satisfy the homogeneous Dirichlet boundary condition imposed for the full problem.

A reasonable initial approximation for the pseudo-time-stepping loop is given by

$$u^0(x,y)=\begin{cases}1-x & y\geqslant 0.5\\ 0 & y<0.5\end{cases}$$

For practical applications, it is worthwhile to compute the stationary low-order solution using any direct or iterative solver, and then activate the time-dependent FEM-FCT algorithm. In this case, the cost of flux correction is minimized, since the initial guess should be close enough to the steady-state limit. Furthermore, the use of the consistent mass matrix is not

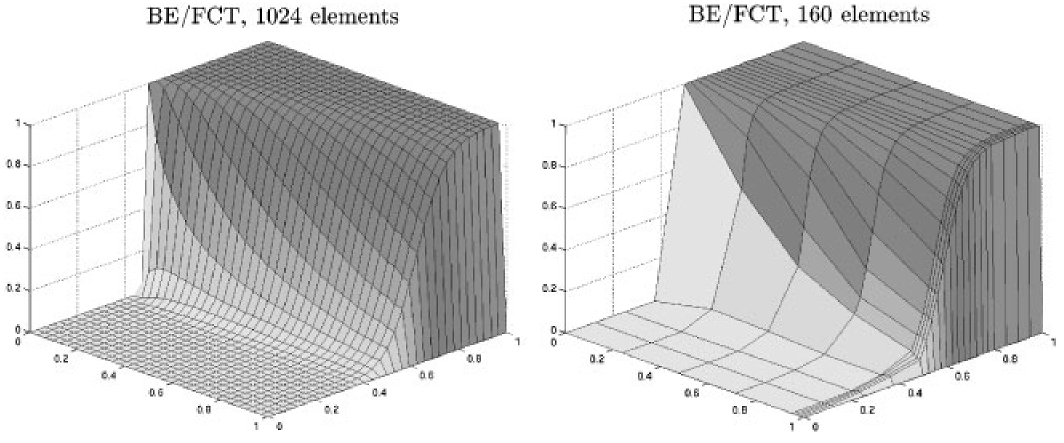


Figure 7. Steady-state convection–diffusion in 2D, $\varepsilon = 10^{-3}$.

justified for stationary problems, so that mass lumping is appropriate also for the high-order scheme.

The numerical solutions depicted in Figure 7 indicate that the backward Euler method equipped with flux correction is capable of producing accurate solutions to steady-state problems at a very low computational cost. It seems to be particularly efficient when used in combination with adaptive mesh refinement. The solution displayed on the left of Figure 7 was computed on a uniform mesh of 32×32 bilinear elements. It is completely non-oscillatory and exhibits a sharp resolution of the boundary layer. Remarkably, comparable accuracy can be achieved with an adaptive mesh consisting of just 160 elements (see Figure 7, right). In addition, the BE/FCT method can be operated with very large time steps, unlike its explicit counterparts which are subject to a CFL condition based on the smallest mesh size. This confirms that the fully implicit approach constitutes a natural framework for the incorporation of an adaptive grid strategy aimed at providing high resolution in the most efficient manner.

11. CONCLUSIONS

A new family of finite-element methods based on the flux-corrected-transport procedure was presented. It was shown that the transition to an edge-based data structure is feasible not only for linear triangles and tetrahedra but also for high-order approximations on arbitrary meshes. The proposed approach to decomposition of convective and diffusive terms resulting from the Galerkin discretization into fluxes is very straightforward as opposed to the technique of Peraire *et al.* [8] underlying some modern compressible flow solvers. The skew symmetry of internodal fluxes guarantees strict mass conservation and makes it possible to apply essentially one-dimensional flux correction tools. In fact, an edge-based representation of antidiffusive fluxes can be used in conjunction with the conventional data structure for other terms. Hence, flux limiters can be built into existing finite-element software without major modifications.

Another highlight of this paper was the algorithm for the construction of low-order schemes to be combined with high-order ones in the FEM-FCT framework. Row-sum mass lumping was employed to remove implicit antidiffusion from the consistent mass matrix. An oscillation-prone transport operator was modified by adding discrete diffusion so as to eliminate negative matrix entries. The source terms were linearized in order to satisfy a rigorous positivity criterion building on the concept of an M-matrix. The derivation of positivity-preserving low-order schemes for systems of hyperbolic conservation laws was discussed. Numerical examples for the one-dimensional shock tube problem indicate that the low-order method based on the spectral radius of the Roe matrix produces reasonable results even as a stand-alone solver for the Euler equations of gas dynamics.

Zalesak's limiter was embedded into a generalized FEM-FCT formulation and backed by a solid mathematical background. A readily computable upper bound for admissible time steps was provided. The presented numerical examples testify that the CN/FCT scheme outperforms the BE/FCT method when it comes to time-accurate simulation of transient flows. At the same time, the backward Euler scheme is unconditionally positive and constitutes an excellent solver for steady-state problems. The fully implicit treatment is also appropriate if a non-uniform distribution of Courant numbers (due to adaptive mesh refinement or strongly varying velocities) makes the CFL condition too restrictive. In other situations, a second-order time discretization of Lax–Wendroff or Crank–Nicolson type should be employed for accuracy reasons. Hence, the unified FEM-FCT formulation encompassing both explicit and implicit schemes represents a very flexible approach to be recommended for a wide range of CFD applications. Its extension to the multidimensional Euler equations will be presented in a forthcoming paper [41].

REFERENCES

1. Boris JP, Book DL. Flux-corrected transport. I. SHASTA, A fluid transport algorithm that works. *Journal of Computational Physics* 1973; **11**:38–69.
2. Zalesak ST. Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of Computational Physics* 1979; **31**:335–362.
3. Parrott AK, Christie MA. FCT applied to the 2-D finite element solution of tracer transport by single phase flow in a porous medium. *Proceedings of ICFD Conference on Numerical Methods in Fluid Dynamics*. Oxford University Press: Oxford, 1986; 609–619.
4. Löhner R, Morgan K, Peraire J, Vahdati M. Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier–Stokes equations. *International Journal for Numerical Methods in Fluids* 1987; **7**:1093–1109.
5. Löhner R, Morgan K, Vahdati M, Boris P, Book DL. FEM-FCT: combining unstructured grids with high resolution. *Communications in Applied Numerical Methods* 1988; **4**:717–729.
6. Selmin V. Finite element solution of hyperbolic equations. I. One-dimensional case. *INRIA Research Report*, vol. 655, 1987.
7. Selmin V. Finite element solution of hyperbolic equations. II. Two-dimensional case. *INRIA Research Report*, vol. 708, 1987.
8. Peraire J, Vahdati M, Peiro J, Morgan K. The construction and behaviour of some unstructured grid algorithms for compressible flows. *Numerical Methods for Fluid Dynamics*, vol. IV. Oxford University Press: Oxford, 1993; 221–239.
9. Löhner R. *Adaptive CFD Techniques*. Wiley: New York, 2001.
10. Roe PL. Approximate Riemann solvers, parameter vectors and difference schemes. *Journal of Computational Physics* 1981; **43**:357–372.
11. Osher S, Solomon F. Upwind difference schemes for hyperbolic systems of conservation laws. *Mathematics of Computation* 1982; **38**:339–374.
12. Steger JL, Warming RF. Flux vector splitting of the inviscid gasdynamic equations with application to finite-difference methods. *Journal of Computational Physics* 1981; **40**:263–293.
13. Jameson A. Computational algorithms for aerodynamic analysis and design. *Applied Numerical Mathematics* 1993; **13**:383–422.
14. Liou MS, Steffen CJ. A new flux splitting scheme. *Journal of Computational Physics* 1993; **107**:23–39.
15. Lyra PRM. Unstructured grid adaptive algorithms for fluid dynamics and heat conduction. *Ph.D. Thesis*, University of Wales, Swansea, 1994.

16. Arminjon P, Dervieux A. Construction of TVD-like artificial viscosities on 2-dimensional arbitrary FEM grids. *INRIA Research Report*, vol. 1111, 1989.
17. Lyra PRM, Morgan K, Peraire J. A high resolution flux splitting scheme for the solution of the compressible Navier–Stokes equations on triangular grids. In *Numerical Methods for the Navier–Stokes Equations*. Hebeker FK, Rannachr R, Wittum G (eds). Vieweg: Heidelberg, 1994; 167–180.
18. Lyra PRM, Morgan K, Peraire J, Peiro J. TVD algorithms for the solution of the compressible Euler equations on unstructured meshes. *International Journal for Numerical Methods in Fluids* 1994; **19**:827–847.
19. Morgan K, Peraire J. Unstructured grid finite element methods for fluid mechanics. *Reports on Progress in Physics* 1998; **61**:569–638.
20. Kuzmin D. Positive finite element schemes based on the flux-corrected transport procedure. In *Computational Fluid and Solid Mechanics*. Bathe KJ (ed.). Elsevier: Amsterdam, 2001; 887–888.
21. Kuzmin D, Turek S. Flux correction tools for finite elements. *Journal of Computational Physics* 2002; **175**:525–558.
22. Jameson A. Positive schemes and shock modelling for compressible flows. *International Journal for Numerical Methods in Fluids* 1995; **20**:743–776.
23. LeVeque RJ. *Numerical Methods for Conservation Laws*. Birkhäuser: Basel, 1992.
24. Fletcher CAJ. The group finite element formulation. *Computer Methods in Applied Mechanics and Engineering* 1983; **37**:225–243.
25. Crouzeix M, Raviart PA. Conforming and nonconforming finite elements for solving the stationary Stokes equations. *RAIRO*, Série Rouge Anal. Num., vol. 7, 1973; R-3, 33–76.
26. Rannacher R, Turek S. A simple nonconforming quadrilateral Stokes element. *Numerical Methods in PDEs* 1992; **8**(2):97–111.
27. Hansbo P. Aspects of conservation in finite element flow computations. *Computer Methods in Applied Mechanics and Engineering* 1994; **117**:423–437.
28. Godunov SK. Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics. *Matematicheskii Sbornik* 1959; **47**:271–306.
29. Dietachmayer GS. A comparison and evaluation of some positive definite advection schemes. *Computational Techniques and Applications*: *CTAC-85*. Elsevier: Amsterdam, 1986; 217–232.
30. Donea J, Selmin V, Quartapelle L. Recent developments of the Taylor–Galerkin method for the numerical solution of hyperbolic problems. *Numerical Methods for Fluid Dynamics III*. Inst. Math. Appl. Conf. Ser: Oxford, 1988; 171–185.
31. Turek S. *Efficient Solvers for Incompressible Flow Problems*: *An Algorithmic and Computational Approach*. Lecture Notes in Computer Science, vol. 6. Springer: Berlin, 1999.
32. Patankar SV. *Numerical Heat Transfer and Fluid Flow*. McGraw-Hill: New York, 1980.
33. Brooks AN, Hughes TJR. Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations. *Computer Methods in Applied Mechanics and Engineering* 1982; **32**:199–259.
34. Carey GF, Jiang BN. Least-squares finite elements for first-order hyperbolic systems. *International Journal for Numerical Methods in Fluids* 1988; **26**:81–93.
35. Donea J, Quartapelle L, Selmin V. An analysis of time discretization in the finite element solution of hyperbolic problems. *Journal of Computational Physics* 1987; **70**:463–499.
36. Donea J, Roig B, Huerta A. *High-order Accurate Time-stepping Schemes for Convection–Diffusion Problems.* International Center for Numerical Methods in Engineering: Barcelona, 1998.
37. DeVore CR. An improved limiter for multidimensional flux-corrected transport. *NASA Technical Report AD*-A360122, 1998.
38. Sod G. A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *Journal of Computational Physics* 1978; **27**:1–31.
39. Anderson JD Jr. *Modern Compressible Flow*. McGraw-Hill: New York, 1990.
40. Lapin A. University of Stuttgart. Private communication.
41. Kuzmin D, Möller M, Turek S. Implicit flux-corrected transport algorithm for finite element simulation of the compressible Euler equations. *Technical Report No*. 221, University of Dortmund, 2002. In *Proceedings of the Conference* "*Finite Element Methods*: *50 Years of Conjugate Gradients*", University of Jyväskylä, Finland, 11–12 June, 2002, to appear.