

Computational Physics

Cache blocking for flux reconstruction: Extension to Navier-Stokes equations and anti-aliasing [☆]Semih Akkurt ^{a,*}, Freddie Witherden ^b, Peter Vincent ^a^a Department of Aeronautics, Imperial College London, SW7 2AZ, United Kingdom^b Department of Ocean Engineering, Texas A&M University, College Station, TX 77843, USA

ARTICLE INFO

Keywords:

Cache blocking
Kernel fusion
High performance computing
Tensor product factorisation
Flux reconstruction
Computational fluid dynamics

ABSTRACT

In this article, cache blocking is implemented for the Navier Stokes equations with anti-aliasing support on mixed grids in PyFR for CPUs. In particular, cache blocking is used as an alternative to kernel fusion to eliminate unnecessary data movements between kernels at the main memory level. Specifically, kernels that exchange data are grouped together, and these groups are then executed on small sub-regions of the domain that fit in per-core private data cache. Additionally, cache blocking is also used to efficiently implement a tensor product factorisation of the interpolation operators associated with anti-aliasing. By using cache blocking, the intermediate results between application of the sparse factors are stored in per-core private data cache, and a significant amount of data movement from main memory is avoided. In order to assess the performance gains a theoretical model is developed, and the implementation is benchmarked using a compressible 3D Taylor-Green vortex test case on both hexahedral and prismatic grids, with third-, fourth-, and fifth-order solution polynomials. The expected performance gains based on the theoretical model range from 1.99 to 2.83, and the speedups obtained in practice range from 1.51 to 3.91 compared to PyFR v1.11.0.

1. Introduction

Achieving a target accuracy level with as little computational cost as possible has always been an important goal in Computational Fluid Dynamics (CFD). To this end, various numerical discretizations have been developed and improved over the years. For example, compact high-order discretizations for unstructured grids, such as Discontinuous Galerkin (DG), Spectral Difference (SD), and Flux Reconstruction (FR) schemes have been developed to enable accurate scale-resolving simulations of turbulent flow in a tractable cost envelope. However, scale resolving simulations of various industry-relevant high Reynolds number flows are still beyond the reach of high-order methods even when using national laboratory scale supercomputers. This has led researchers to investigate ways to utilize available computational power more efficiently. Changing the arithmetic intensity of kernels can be a way to achieve this. For example, an alternative mathematical form of a given operation may require fewer FLOPs and this may lead to a better arithmetic intensity for a particular hardware. Another example is kernel fusion, which reduces the data movements between processor and mem-

ory, and therefore increases the arithmetic intensity [1–3]. Depending on the arithmetic intensity profile of the kernels and the target hardware, these methods can be used to achieve better performance.

In recent work, the authors implemented a cache blocking strategy as an alternative to kernel fusion to reduce bandwidth requirement for an FR code for Euler equations without anti-aliasing on hexahedral meshes [4]. Cache blocking reduces main memory accesses by using L2 cache as a buffer storage for intermediate and temporary data in a set of kernels that exchange data, which we will refer as a kernel group in the remainder of this article. As long as the L2 cache is big enough for storing the temporary data, cache blocking has all the benefits of kernel fusion in terms of bandwidth reduction without any downsides [4]. Furthermore, the cache blocking approach is easy to implement, and changing the configuration of kernel groups or adding new kernels into the existing kernel groups is simple and straightforward. Here, we extend this previous work to support Navier-Stokes equations with anti-aliasing on mixed element meshes. Extension to Navier-Stokes solver enables real-world viscous simulations and it is necessary to use anti-aliasing in order to remain numerically stable for certain physical simulations

[☆] The review of this paper was arranged by Prof. David W. Walker.

* Corresponding author.

E-mail address: s.akkurt18@imperial.ac.uk (S. Akkurt).

Nomenclature

Indices

e	Element type
n	Element number
α	Field variable number
i, j, k	Summation indices
ζ, σ, ν	Summation indices

Domains

Ω	Solution domain
Ω_e	All elements in Ω of type e
$\hat{\Omega}_e$	A <i>standard</i> element of type e
Ω_{en}	Element n of type e
$ \Omega_e $	Number of elements of type e

Expansions

p	Polynomial order
-----	------------------

p_q	Quadrature order
N_D	Number of spatial dimensions
N_V	Number of field variables

Adornments and suffixes

$\tilde{\square}$	A quantity in transformed space
$\hat{\square}$	A vector quantity of unit magnitude
$\square^{(u)}$	A quantity at a solution point
$\square^{(q)}$	A quantity at a quadrature point
$\square^{(f)}$	A quantity at a flux point
$\square^{(sq)}$	A quantity at a surface-quadrature point
$\square^{(f\perp)}$	A normal quantity at a flux point
$\square^{(sq\perp)}$	A normal quantity at a surface-quadrature point

Operators

$\mathfrak{C}_\alpha$	Common solution at an interface
$\mathfrak{F}_\alpha$	Common normal flux at an interface

such as turbulent flows that are marginally resolved [5]. Additionally, supporting mixed element meshes provides flexibility for complex geometries.

There are three important aspects to consider for the aforementioned extensions. First, compared to an Euler solver, a Navier-Stokes solver has more kernels and this introduces more constraints on how they can be grouped together. Next, a direct implementation of anti-aliasing results in additional dense operators as well as changing many of the existing sparse operators in the Navier-Stokes solver into dense operators. This results in FLOP limited kernels due to the sufficiently high arithmetic intensity associated with the dense operators. However, an alternative method for implementing anti-aliasing is using sum factorisation, which was initially proposed by Orszag [6] for spectral methods and later adapted and used for spectral/ hp element methods by Sherwin and Karniadakis [7]. A detailed analysis of sum factorisation carried out by Cantwell [8] shows that sum factorisation can be effective only at $p = 7$ and beyond. Further studies also confirmed a similar trend [9,10]. Świrydowicz et al. [11] employed a slightly different strategy which factors the tensor product operators into sparse components by doing 1D interpolations in every dimension one-by-one. Employing such a tensor product factorisation strategy naively for these dense operators where each factor is applied sequentially to the entire data set without cache blocking results in significant increase in the bandwidth requirement, and makes the operations bandwidth bound rather than FLOP bound. Kernel fusion is a solution to this, but it requires not only a complex implementation, but also makes the code architecture specific. Therefore, we propose using the cache blocking strategy in conjunction with the tensor product factorisation in order to eliminate the excess bandwidth requirement. In short, the methodology we propose for anti-aliasing effectively trades off FLOP requirements with bandwidth requirements by using a tensor product factorisation strategy, and then eliminates the excess bandwidth requirement by employing a cache blocking strategy. As a result, anti-aliasing kernels require significantly less FLOPs while requiring similar or even reduced access to main memory. Finally, we enable mixed element support including hexahedra, prisms, tetrahedra and pyramids where tensor product factorisation is only used for hexahedral and prism elements.

The paper is structured as follows. The FR formulation for the Navier-Stokes equations including anti-aliasing is given in Section 2 alongside with the kernels that implement these mathematical operations. Next, current kernel execution order and the data movement requirements are presented in Section 3. Anti-aliasing and its effects in terms of computational cost are discussed in Section 4. Then, an efficient implementation of anti-aliasing in FR within the context of cache blocking is described

in Section 5. A kernel grouping strategy for the Navier-Stokes equations including anti-aliasing support is given alongside the speedups obtained in practice in Section 6. Finally, conclusions are discussed in Section 7.

2. Navier-Stokes formulation and kernels

Flux Reconstruction (FR) was first developed by Huynh in 2007 [12]. An overview of the FR approach for solving the Navier-Stokes equations including anti-aliasing is presented in this section.

2.1. Formulation

The Navier-Stokes equations can be written in conservative form as,

$$\frac{\partial u_\alpha}{\partial t} + \nabla \cdot \mathbf{f}_\alpha = 0, \quad (1)$$

where α is the field variable index, $u_\alpha = u_\alpha(\mathbf{x}, t)$ are the conservative field variables thus,

$$\mathbf{u} = \begin{bmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho v_z \\ E \end{bmatrix}, \quad (2)$$

and $\mathbf{f} = \mathbf{f}(u, \nabla u)$ is the flux term that incorporates inviscid and viscous fluxes thus $\mathbf{f} = \mathbf{f}^e - \mathbf{f}^v$. First, the inviscid flux, $\mathbf{f}^e$, is defined as

$$\mathbf{f}^e(u) = \begin{bmatrix} \rho v_x \\ \rho v_x^2 + p \\ \rho v_x v_y \\ \rho v_x v_z \\ v_x(E + p) \end{bmatrix} \mathbf{i} + \begin{bmatrix} \rho v_y \\ \rho v_y^2 + p \\ \rho v_y v_x \\ \rho v_y v_z \\ v_y(E + p) \end{bmatrix} \mathbf{j} + \begin{bmatrix} \rho v_z \\ \rho v_z^2 + p \\ \rho v_z v_x \\ \rho v_z v_y \\ v_z(E + p) \end{bmatrix} \mathbf{k}, \quad (3)$$

where ρ is density, v_x, v_y, v_z are velocity, p is pressure, E is total energy per unit volume, and $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are orthogonal unit vectors. The relation between the pressure and the total energy for a perfect gas is,

$$E = \frac{p}{\gamma - 1} + \frac{1}{2} \rho (v_x^2 + v_y^2 + v_z^2), \quad (4)$$

where γ is the specific heat ratio, c_p/c_v , c_p is the specific heat at constant pressure, and c_v is the specific heat at constant volume.

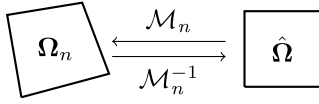


Fig. 1. Mapping function from physical space to reference space for a quadrilateral element.

Next, the viscous flux, $\mathbf{f}^v$ is given as

$$\mathbf{f}^v(u, \Delta u) = \begin{bmatrix} 0 \\ \mathcal{T}_{xx} \\ \mathcal{T}_{xy} \\ \mathcal{T}_{xz} \\ v_i \mathcal{T}_{ix} + \phi_x \end{bmatrix} \mathbf{i} + \begin{bmatrix} 0 \\ \mathcal{T}_{yx} \\ \mathcal{T}_{yy} \\ \mathcal{T}_{yz} \\ v_i \mathcal{T}_{iy} + \phi_y \end{bmatrix} \mathbf{j} + \begin{bmatrix} 0 \\ \mathcal{T}_{zx} \\ \mathcal{T}_{zy} \\ \mathcal{T}_{zz} \\ v_i \mathcal{T}_{iz} + \phi_z \end{bmatrix} \mathbf{k}, \quad (5)$$

where $\mathcal{T}$ is the stress tensor and defined as,

$$\mathcal{T} = \mu(\partial_i v_j + \partial_j v_i) - \frac{2}{3} \mu \delta_{ij} \nabla \cdot \mathbf{v}, \quad (6)$$

and ϕ is the heat flux defined as,

$$\phi = \mu \frac{c_p}{P_r} \nabla T. \quad (7)$$

The ideal gas law relates the temperature T with pressure and density as

$$T = \frac{1}{c_v} \frac{1}{1 - \gamma} \frac{p}{\rho}. \quad (8)$$

With the flux term defined, ∇u can be substituted by $\mathbf{q}$ so that Equation (1) can be rewritten as a first order system thus,

$$\begin{aligned} \frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}_a(u, \mathbf{q}) &= 0, \\ \mathbf{q}_a - \nabla u &= 0. \end{aligned} \quad (9)$$

When applying FR to solve the Navier-Stokes equations in a domain Ω , the first step is to tessellate the domain with non-overlapping, conforming elements thus

$$\Omega = \bigcup_{e \in \mathcal{E}} \Omega_e, \quad \Omega_e = \bigcup_{n=1}^{|\Omega_e|} \Omega_{en}, \quad \bigcap_{e \in \mathcal{E}} \bigcap_{n=1}^{|\Omega_e|} \Omega_{en} = \emptyset. \quad (10)$$

Where e is a particular element type in the list of available elements $\mathcal{E}$, therefore Ω_e refers to all elements of type e , and $|\Omega_e|$ is the number of elements of type e . Each element Ω_{en} is then mapped to a reference element $\hat{\Omega}_e$ via a mapping function $\mathcal{M}_{en}$ defined as

$$\mathbf{x} = \mathcal{M}_{en}(\tilde{\mathbf{x}}), \quad \tilde{\mathbf{x}} = \mathcal{M}_{en}^{-1}(\mathbf{x}),$$

and this is illustrated for a quadrilateral element in Fig. 1.

Geometric Jacobian matrices can be defined from the mapping functions thus

$$\begin{aligned} \mathbf{J}_{en} &= J_{en\tilde{j}} = \frac{\partial \mathcal{M}_{en\tilde{i}}}{\partial \tilde{x}_j}, & J_{en} &= \det \mathbf{J}_{en}, \\ \mathbf{J}_{en}^{-1} &= J_{en\tilde{j}}^{-1} = \frac{\partial \mathcal{M}_{en\tilde{i}}^{-1}}{\partial \tilde{x}_j}, & J_{en}^{-1} &= \det \mathbf{J}_{en}^{-1} = \frac{1}{J_{en}}. \end{aligned}$$

The definitions above will be used to transform quantities to and from reference element space. In order to transform Equation (9) into reference space, transformed solution $\tilde{u}_{en}$, transformed flux $\tilde{\mathbf{f}}_{en}$, and transformed gradient $\tilde{\mathbf{q}}_{en}$ are defined thus

$$\tilde{u}_{en} = \tilde{u}_{en}(\tilde{\mathbf{x}}, t) = J_{en}(\tilde{\mathbf{x}}) u_{en}(\mathcal{M}_{en}(\tilde{\mathbf{x}}), t), \quad (11a)$$

$$\tilde{\mathbf{f}}_{en} = \tilde{\mathbf{f}}_{en}(\tilde{\mathbf{x}}, t) = J_{en}(\tilde{\mathbf{x}}) \mathbf{J}_{en}^{-1}(\mathcal{M}_{en}(\tilde{\mathbf{x}})) \mathbf{f}_{en}(\mathcal{M}_{en}(\tilde{\mathbf{x}}), t), \quad (11b)$$

$$\tilde{\mathbf{q}}_{en} = \tilde{\mathbf{q}}_{en}(\tilde{\mathbf{x}}, t) = \mathbf{J}_{en}^T(\tilde{\mathbf{x}}) \mathbf{q}_{en}(\mathcal{M}_{en}(\tilde{\mathbf{x}}), t), \quad (11c)$$

and for a more compact definition let $\tilde{\nabla} = \partial / \partial \tilde{x}_i$. Then Equation (9) can be written as,

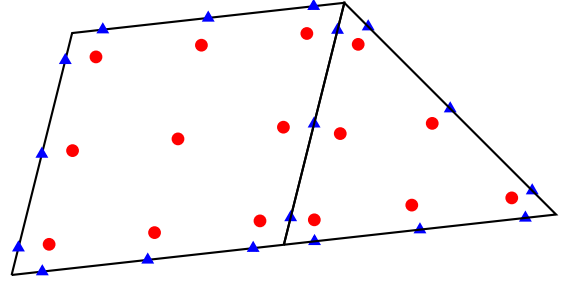


Fig. 2. Solution points and flux points for a quadrilateral and a triangular elements. (For interpretation of the colours in the figure(s) and table(s), the reader is referred to the web version of this article.)

$$\begin{aligned} \frac{\partial u_{en\alpha}}{\partial t} + J_{en}^{-1} \tilde{\nabla} \cdot \tilde{\mathbf{f}}_{en\alpha} &= 0, \\ \tilde{\mathbf{q}}_{en\alpha} - \tilde{\nabla} u_{en\alpha} &= 0. \end{aligned} \quad (12)$$

Now that the equation in the discretized domain is defined, the next step is to define set of solution points inside the elements and flux points at element interfaces. FR method represents the solution inside the elements by using a polynomial of order p . The solution across the element interfaces is allowed to be discontinuous. Selection of the solution and flux point sets affects the properties of the numerical scheme as indicated in various publications such as [13–15]. The formulation given in this section is valid for any solution and flux point set.

First, let $\tilde{\mathbf{x}}_{e\zeta}^{(u)}$ be the set of solution points for each reference element $e \in \mathcal{E}$. ζ is the solution point index in an element and satisfies $0 \leq \zeta < N_e^{(u)}$, where $N_e^{(u)}$ is the number of solution points in element type e . Now a nodal basis set $\ell_{e\zeta}^{(u)}(\tilde{\mathbf{x}})$ can be defined where the nodal basis polynomial $\ell_{e\zeta}^{(u)}$ satisfies $\ell_{e\zeta}^{(u)}(\tilde{\mathbf{x}}_{e\sigma}^{(u)}) = \delta_{\zeta\sigma}$. Next, a set of flux points $\tilde{\mathbf{x}}_{e\zeta}^{(f)}$ on $\partial \hat{\Omega}_e$ is also defined, where $0 \leq \zeta < N_e^{(f)}$ and $N_e^{(f)}$ is the number of flux points at the interfaces of the element type e . These flux points are constrained such that all paired flux points across interfaces match in terms of global coordinates, $\mathcal{M}_{en}(\tilde{\mathbf{x}}_{e\zeta}^{(f)}) = \mathcal{M}_{\tilde{en}}(\tilde{\mathbf{x}}_{\tilde{e}\zeta}^{(f)})$, where $\sim$ denotes the corresponding interface from the neighbouring element. Furthermore, there is an associated outward-pointing normal vector for each flux point given by $\hat{\mathbf{n}}_{e\zeta}^{(f)}$. The solution points, flux points, and an interface between a quadrilateral element and triangular element are all demonstrated in Fig. 2.

The first step in the FR method is obtaining the discontinuous solution at the flux points $u_{e\sigma n\alpha}^{(f)}$ from the solution state at the solution points $u_{e\zeta n\alpha}^{(u)}$ as

$$u_{e\sigma n\alpha}^{(f)} = u_{e\zeta n\alpha}^{(u)} \ell_{e\zeta}^{(u)}(\tilde{\mathbf{x}}_{e\sigma}^{(f)}) \quad (13)$$

Then, the discontinuous solution at the interfaces can be used to obtain a common solution via the LDG method [16,17] which is defined as

$$\mathfrak{C}(u_L, u_R) = (\frac{1}{2} - \beta) u_L + (\frac{1}{2} + \beta) u_R. \quad (14)$$

It is important to note that $\mathfrak{C}(u_L, u_R)$ is not in general equal to $\mathfrak{C}(u_R, u_L)$. Therefore it is important to visit all the flux point pairs in the domain only once and assign the common value to both points. Consequently, the following equalities hold

$$\mathfrak{C}_\alpha(u_{e\sigma n\alpha}^{(f)}, u_{\tilde{e}\sigma n\alpha}^{(f)}) = \mathfrak{C}_\alpha u_{e\sigma n\alpha}^{(f)} = \mathfrak{C}_\alpha u_{\tilde{e}\sigma n\alpha}^{(f)}. \quad (15)$$

A correction procedure is then used to obtain a C^0 continuous solution across elements. For this purpose, there is a correction function associated with each flux point in the reference element. Specifically, $\mathbf{g}_{e\zeta}^{(f)}$ is the vector correction function which satisfies $\hat{\mathbf{n}}_{e\sigma}^{(f)} \cdot \mathbf{g}_{e\zeta}^{(f)}(\tilde{\mathbf{x}}_{e\sigma}^{(f)}) = \delta_{\zeta\sigma}$. The form of $\mathbf{g}_{e\zeta}^{(f)}$ determines a particular FR scheme. Further details on the correction function can be found in [18].

Combining the above, Equation (11c) can be expressed as

$$\tilde{\mathbf{q}}_{e\zeta n\alpha}^{(u)} = \left[\hat{\mathbf{n}}_{e\sigma}^{(f)} \cdot \tilde{\mathbf{g}}_{e\sigma}^{(f)}(\tilde{\mathbf{x}}) \{ \mathbf{c}_\alpha u_{e\sigma n\alpha}^{(f)} - u_{e\sigma n\alpha}^{(f)} \} + \tilde{\mathbf{u}}_{evn\alpha}^{(u)} \tilde{\nabla} \ell_{ev}^{(u)}(\tilde{\mathbf{x}}) \right]_{\tilde{\mathbf{x}}=\tilde{\mathbf{x}}_{e\zeta}^{(u)}}. \quad (16)$$

This equation simply computes the transformed gradient by using the analytic derivative of the basis polynomial ℓ_e together with the solution at solution points, and adds the contribution of the jump between the discontinuous solution at the flux points and the common solution at flux points.

The physical gradients at solution points and at flux points can be computed using approaches of Kopriva [19] and Sun et al. [20] as

$$\mathbf{q}_{e\zeta n\alpha}^{(u)} = \mathbf{J}_{en}^{-T}(\tilde{\mathbf{x}}_{e\sigma}^{(u)}) \tilde{\mathbf{q}}_{e\zeta n\alpha}^{(u)}, \quad (17)$$

$$\mathbf{q}_{e\zeta n\alpha}^{(f)} = \ell_{e\zeta}^{(u)}(\tilde{\mathbf{x}}_{e\sigma}^{(f)}) \tilde{\mathbf{q}}_{e\zeta n\alpha}^{(u)}. \quad (18)$$

The auxiliary equation is now solved and the transformed flux can be obtained as

$$\tilde{\mathbf{f}}_{e\sigma n\alpha}^{(u)} = \mathbf{J}_{e\sigma n}^{(u)} \mathbf{J}_{e\sigma n}^{-1(u)} \mathbf{f}_\alpha(u_{e\sigma n}^{(u)}, \mathbf{q}_{e\sigma n}^{(u)}), \quad (19)$$

where $\mathbf{J}_{e\sigma n}^{(u)} = \det(\mathbf{J}_{en}^{-T}(\tilde{\mathbf{x}}_{e\sigma}^{(u)}))$. Now a procedure similar to that used in Equation (13) can be applied here to obtain the normal transformed flux at flux points as

$$f_{e\sigma n\alpha}^{(f_\perp)} = \ell_{e\zeta}^{(u)}(\tilde{\mathbf{x}}_{e\sigma}^{(f)}) \hat{\mathbf{n}}_{e\sigma}^{(f)} \cdot \tilde{\mathbf{f}}_{e\zeta n\alpha}^{(u)}. \quad (20)$$

With the discontinuous solution and the physical gradient at the flux points, the next step is to obtain the common normal flux for each flux point pair in the domain. The inviscid component of the flux is computed using a Riemann solver [21,22]. The viscous component of the flux is obtained by using the Local Discontinuous Galerkin (LDG) method as described in [16] and it is defined as

$$\tilde{\mathbf{f}}^v(\mathbf{f}_L^v, \mathbf{f}_R^v, \hat{\mathbf{n}}_L) = \hat{\mathbf{n}}_L \cdot ((1/2 + \beta) \mathbf{f}_L^v + (1/2 - \beta) \mathbf{f}_R^v) + \tau(u_L - u_R), \quad (21)$$

where τ is a penalty parameter, $\mathbf{f}_L^v = \mathbf{f}^v(u_L, \mathbf{q}_L)$, $\mathbf{f}_R^v = \mathbf{f}^v(u_R, \mathbf{q}_R)$, and β controls the upwinding or downwinding.

$\tilde{\mathcal{F}}_\alpha f_{e\sigma n\alpha}^{(f_\perp)}$ is the operator for the common normal flux at the interface, and $f_{e\sigma n\alpha}^{(f_\perp)}$ denotes the normal flux at a flux point. Note that interfaces between elements conform and we have $\hat{\mathbf{n}}_{\sigma n}^{(f)} = -\hat{\mathbf{n}}_{\sigma n}^{(f)}$. Due to the conservation property of the numerical scheme, we always have $\tilde{\mathcal{F}}_\alpha f_{e\sigma n\alpha}^{(f_\perp)} = -\tilde{\mathcal{F}}_\alpha f_{e\sigma n\alpha}^{(f_\perp)}$, and the full form of the common normal flux operator can be given as,

$$\tilde{\mathcal{F}}_\alpha f_{e\sigma n\alpha}^{(f_\perp)} = -\tilde{\mathcal{F}}_\alpha f_{e\sigma n\alpha}^{(f_\perp)} = \tilde{\mathcal{F}}_\alpha f_{e\sigma n\alpha}^{(f_\perp)}(u_{e\sigma n}^{(f)}, u_{e\sigma n}^{(f)}, \mathbf{q}_{e\sigma n}^{(f)}, \mathbf{q}_{e\sigma n}^{(f)}, \hat{\mathbf{n}}_{e\sigma n}^{(f)}). \quad (22)$$

Next, the difference between the flux and the common normal flux at flux points is used in the correction procedure analogous to the Equation (16). Finally, assembling all of the steps defined above, we can define an expression for the divergence of the continuous flux as,

$$(\tilde{\nabla} \cdot \tilde{\mathbf{f}})_{e\zeta n\alpha}^{(u)} = \left[\tilde{\nabla} \cdot \tilde{\mathbf{g}}_{e\sigma}^{(f)}(\tilde{\mathbf{x}}) \{ \tilde{\mathcal{F}}_\alpha f_{e\sigma n\alpha}^{(f_\perp)} - f_{e\sigma n\alpha}^{(f_\perp)} \} + \tilde{\mathbf{f}}_{evn\alpha}^{(u)} \cdot \tilde{\nabla} \ell_{ev}^{(u)}(\tilde{\mathbf{x}}) \right]_{\tilde{\mathbf{x}}=\tilde{\mathbf{x}}_{e\zeta}^{(u)}}, \quad (23)$$

which results in a semi-discretised form of the governing system.

2.1.1. Anti-aliasing formulation

FR uses collocation based projections to obtain the transformed flux at the solution points. This approach is efficient but may result in aliasing errors where the energy from unresolved modes gets erroneously transferred into resolved modes and consequently this results in stability problems as well as accuracy issues. One way to overcome aliasing errors is using L^2 based projections with suitable quadrature rules [5].

Two types of anti-aliasing are considered here, flux anti-aliasing and surface-flux anti-aliasing. Flux anti-aliasing uses fluxes at quadrature points to perform an L^2 projection of flux and then samples this projection at the original solution points. Surface-flux anti-aliasing applies the flux correction operation in a similar way. First, solution gradients at solution points are used to obtain the solution gradients at the

surface-quadrature points. Then, common flux and common solution are evaluated at the surface-quadrature points, and finally common flux and solution are sampled at the original flux points. These procedures require extra operators and changes in some of the existing operators. First, flux anti-aliasing requires two new matrix multiplication operations. One of them operates on the solution and the other operates on the gradients at the solution points. Next, surface-flux anti-aliasing requires changes in many of the existing operators in the FR algorithm.

First, for flux anti-aliasing, interpolation from solution points to quadrature points can be expressed as

$$u_{e\sigma n\alpha}^{(q)} = u_{e\zeta n\alpha}^{(u)} \ell_{e\zeta}^{(u)}(\tilde{\mathbf{x}}_{e\sigma}^{(q)}), \quad (24)$$

$$\mathbf{q}_{e\sigma n\alpha}^{(q)} = \mathbf{q}_{e\zeta n\alpha}^{(u)} \ell_{e\zeta}^{(u)}(\tilde{\mathbf{x}}_{e\sigma}^{(q)}). \quad (25)$$

Then the solution and gradients at the quadrature points are used to evaluate the flux at quadrature points, replacing Equation (19) thus

$$\tilde{\mathbf{f}}_{e\sigma n\alpha}^{(q)} = \mathbf{J}_{e\sigma n}^{(q)} \mathbf{J}_{e\sigma n}^{-1(q)} \mathbf{f}_\alpha(u_{e\sigma n}^{(q)}, \mathbf{q}_{e\sigma n}^{(q)}). \quad (26)$$

Then, in order to sample the fluxes at the original solution points the following operation can be used

$$\tilde{\mathbf{f}}_{e\zeta n\alpha}^{(u)} = \psi_{ev}(\tilde{\mathbf{x}}_{e\zeta}^{(u)}) \omega_{e\sigma}^{(q)} \psi_{ev}(\tilde{\mathbf{x}}_{e\sigma}^{(q)}) \tilde{\mathbf{f}}_{e\sigma n\alpha}^{(q)}, \quad (27)$$

where ψ_{ev} is the polynomial basis function and $\omega_{e\sigma}^{(q)}$ is the weights associated with quadrature points.

Next, a similar strategy can be used for the surface-flux anti-aliasing. Surface-flux anti-aliasing first evaluates the solution $u_{e\sigma n\alpha}^{(sq)}$ at the surface-quadrature points $\tilde{\mathbf{x}}_{e\sigma}^{(sq)}$ from the solution at the solution points $u_{e\zeta n\alpha}^{(u)}$ as

$$u_{e\sigma n\alpha}^{(sq)} = u_{e\zeta n\alpha}^{(u)} \ell_{e\zeta}^{(u)}(\tilde{\mathbf{x}}_{e\sigma}^{(sq)}), \quad (28)$$

and evaluates the gradients of the solution at the surface-quadrature points $\tilde{\mathbf{q}}_{e\sigma n\alpha}^{(sq)}$ from the gradients of the solution at the solution points $\tilde{\mathbf{q}}_{e\zeta n\alpha}^{(u)}$ as

$$\tilde{\mathbf{q}}_{e\sigma n\alpha}^{(sq)} = \tilde{\mathbf{q}}_{e\zeta n\alpha}^{(u)} \ell_{e\zeta}^{(u)}(\tilde{\mathbf{x}}_{e\sigma}^{(sq)}). \quad (29)$$

Then, we obtain the common normal flux $\tilde{\mathcal{F}}_\alpha f_{e\sigma n\alpha}^{(sq_\perp)}$ as described in Equation (22), but at the surface-quadrature points. Finally, the common normal flux at the surface-quadrature points is projected back to the flux points via

$$\tilde{\mathcal{F}}_\alpha f_{e\zeta n\alpha}^{(f_\perp)} = \psi_{ev}(\tilde{\mathbf{x}}_{e\zeta}^{(f)}) \omega_{e\sigma}^{(sq)} \psi_{ev}(\tilde{\mathbf{x}}_{e\sigma}^{(sq)}) \tilde{\mathcal{F}}_\alpha f_{e\sigma n\alpha}^{(sq_\perp)}. \quad (30)$$

2.2. Formulation in terms of kernels

The operations required to implement the FR algorithm described in Section 2 can be cast in terms of individual kernels. These can be grouped into two main types, matrix-matrix multiplication and point-wise kernels. In PyFR, various third party libraries such as LIBXSMM [23] and GiMMiK [24] are made available for the matrix-matrix multiplication kernels. Point-wise kernels are implemented through PyFR-Mako, a bespoke templating language derived from Mako, which is then rendered into low-level platform-specific code.

In order to formulate the FR algorithm in terms of kernels, we define the various constant operator matrices and point-wise kernels, alongside various storage matrices for state variables and some intermediate data.

First, we define the *disu* kernel, which corresponds to Equation (13). This kernel evaluates the solution at flux points using the solution at solution points, and can be formulated as a matrix multiplication. The constant operator matrix of element type e for the *disu* kernel, $\mathbf{M}_e^0$, is defined as

$$(\mathbf{M}_e^0)_{\sigma\zeta} = \ell_{e\zeta}^{(u)}(\tilde{\mathbf{x}}_{e\sigma}^{(f)}), \quad \dim \mathbf{M}_e^0 = N_e^{(f)} \times N_e^{(u)}, \quad (31)$$

the array $\mathbf{U}_e^{(u)}$ for storing the solution state at solution points is defined as

$$(\mathbf{U}_e^{(u)})_{\zeta(n\alpha)} = u_{e\zeta n\alpha}^{(u)}, \quad \dim \mathbf{U}_e^{(u)} = N_e^{(u)} \times N_V |\Omega_e|, \quad (32)$$

where N_V is the number of state variables. Therefore, the *disu* kernel can be cast as a matrix multiplication as,

$$\mathbf{U}_e^{(f)} = \mathbf{M}_e^0 \mathbf{U}_e^{(u)}, \quad (33)$$

where $\mathbf{U}_e^{(f)}$ is the storage for the solution state at flux points which has dimension $\dim \mathbf{U}_e^{(f)} = N_e^{(f)} \times N_V |\Omega_e|$.

The next five kernels are dedicated to obtaining the gradients of the solution which is a requirement for calculating the flux and common flux. The first operation related to the gradients is obtaining the common solution at the flux points which is defined in Equation (15). The *con_u* kernel is defined to obtain the common solution as

$$\mathbf{C}_e^{(f)} = \text{con}_u(\mathbf{U}_e^{(f)}, \mathbf{n}_e^{(f)}), \quad (34)$$

which is a point-wise kernel with considerable indirect memory access to main memory. The *con_u* kernel uses up-winding or down-winding to determine the common solution at the element interfaces, and writes the solution into a sub-region of the $\mathbf{F}_e^{(f)}$ array in order to save memory usage. For clarity we will refer to this sub-region of the $\mathbf{F}_e^{(f)}$ matrix as $\mathbf{C}_e^{(f)}$ after its data is replaced by the *con_u* kernel. $\mathbf{n}_e^{(f)}$ refers to the outward pointing normal vector at the flux points.

Afterwards, the corrections have to be applied in order to obtain the gradients. The procedure to apply corrections is described in Equation (16). Therefore the following two constant matrices are defined thus

$$(\mathbf{M}_e^4)_{\zeta\sigma} = [\tilde{\mathbf{V}}_{e\zeta}^{(u)}(\tilde{\mathbf{x}})]_{\tilde{\mathbf{x}}=\tilde{\mathbf{x}}_{e\sigma}}^T, \quad \dim \mathbf{M}_e^4 = N_e^{(u)} \times N_D N_e^{(u)}, \quad (35)$$

$$(\mathbf{M}_e^6)_{\zeta\sigma} = [\tilde{\mathbf{g}}_{e\zeta}^{(f)} \cdot \tilde{\mathbf{V}} \cdot \mathbf{g}_{e\zeta}^{(f)}(\tilde{\mathbf{x}})]_{\tilde{\mathbf{x}}=\tilde{\mathbf{x}}_{e\sigma}}, \quad \dim \mathbf{M}_e^6 = N_e^{(u)} \times N_e^{(f)}, \quad (36)$$

where N_D is the number of dimensions. Using the matrices defined above, Equation (16) can be written in matrix multiplication form as

$$\begin{aligned} \tilde{\mathbf{Q}}_e^{(u)} &= \mathbf{M}_e^6 \{ \tilde{\mathbf{C}}_e^{(f)} - \tilde{\mathbf{U}}_e^{(f)} \} + \mathbf{M}_e^4 \tilde{\mathbf{U}}_e^{(u)} \\ &= \mathbf{M}_e^6 \{ \tilde{\mathbf{C}}_e^{(f)} - \mathbf{M}_e^0 \tilde{\mathbf{U}}_e^{(u)} \} + \mathbf{M}_e^4 \tilde{\mathbf{U}}_e^{(u)} \\ &= \mathbf{M}_e^6 \tilde{\mathbf{C}}_e^{(f)} - \{ \mathbf{M}_e^4 - \mathbf{M}_e^6 \mathbf{M}_e^0 \} \tilde{\mathbf{U}}_e^{(u)}, \end{aligned} \quad (37)$$

where the $\tilde{\mathbf{Q}}_e^{(u)}$ array is the storage for the gradients at solution points and has a size $\dim \tilde{\mathbf{Q}}_e^{(u)} = N_D N_e^{(u)} \times N_V |\Omega_e|$. The size of $\tilde{\mathbf{Q}}_e^{(u)}$ is N_D times bigger compared to that of $\tilde{\mathbf{U}}_e^{(u)}$ as there are N_D many components of gradient for each state variable. As there are two arrays and two constant matrix multiplications involved in this operation, it can be cast as two consecutive matrix multiplication kernels. Corresponding to the first and second terms in the equation, *tgradcoru* and *tgradpcoru* kernels are defined respectively. The *tgradpcoru* is called first and the output is stored at $\tilde{\mathbf{Q}}_e^{(u)}$ thus

$$\tilde{\mathbf{Q}}_e^{(u)} = \{ \mathbf{M}_e^4 - \mathbf{M}_e^6 \mathbf{M}_e^0 \} \tilde{\mathbf{U}}_e^{(u)}. \quad (38)$$

Next, the *tgradcoru* kernel is called and the result is added to $\tilde{\mathbf{Q}}_e^{(u)}$ as

$$\tilde{\mathbf{Q}}_e^{(u)} = \tilde{\mathbf{Q}}_e^{(u)} + \mathbf{M}_e^6 \tilde{\mathbf{C}}_e^{(f)}. \quad (39)$$

Then, a point-wise kernel, *gradcoru* is defined in order to obtain the gradients of the solution at solution points in physical space as formulated in Equation (17) thus

$$\mathbf{Q}_e^{(u)} = \text{gradcoru}(\mathbf{J}_e^{(u)}, \tilde{\mathbf{Q}}_e^{(u)}). \quad (40)$$

Afterwards, the *gradcoru_fpts* kernel is defined in order to evaluate the physical gradients at the flux points corresponding to Equation (18). This is also a matrix multiplication kernel and the associated constant operator matrix is $\mathbf{M}_e^5$, which is defined as

$$(\mathbf{M}_e^5)_{\sigma\zeta} = \text{diag}(\mathbf{M}_e^0, \dots, \mathbf{M}_e^0), \quad \dim \mathbf{M}_e^5 = N_D N_e^{(f)} \times N_D N_e^{(u)}. \quad (41)$$

Therefore, *gradcoru_fpts* kernel can be defined as

$$\mathbf{Q}_e^{(f)} = \mathbf{M}_e^5 \mathbf{Q}_e^{(u)}, \quad (42)$$

where $\mathbf{Q}_e^{(f)}$ is the storage array for the gradients at flux points with the dimension $\dim \mathbf{Q}_e^{(f)} = N_D N_e^{(f)} \times N_V |\Omega_e|$.

Next, with the solution and its gradient available at the element interfaces, it is now possible to obtain the common interface flux at the element interfaces as in Equation (22). For this purpose, the *comm_flux* kernel is defined thus

$$\mathbf{D}_e^{(f)} = \text{comm_flux}(\mathbf{U}_e^{(f)}, \mathbf{Q}_e^{(f)}, \mathbf{n}_e^{(f)}), \quad (43)$$

which is a point-wise kernel with a significant indirect memory access requirement. The *comm_flux* kernel calls a Riemann flux function on pairs of flux points reading $\mathbf{U}_e^{(f)}$ and $\mathbf{F}_e^{(f)}$, then returning $\mathbf{D}_e^{(f)}$. In practice it overwrites the data over $\mathbf{U}_e^{(f)}$. However for clarity, we will refer to the $\mathbf{U}_e^{(f)}$ matrix as $\mathbf{D}_e^{(f)}$ after its data is replaced by the *comm_flux* kernel.

Then, the next step is obtaining the transformed flux at the solution points as in Equation (19). A point-wise kernel, *tdisf*, is defined to evaluate the flux at solution points thus

$$\tilde{\mathbf{F}}_e^{(u)} = \text{tdisf}(\mathbf{U}_e^{(u)}, \mathbf{Q}_e^{(u)}, \mathbf{x}_e^{(u)}). \quad (44)$$

This kernel takes the solution $\mathbf{U}_e^{(u)}$, gradient at the solution points $\mathbf{Q}_e^{(u)}$, and the element coordinates $\mathbf{x}_e^{(u)}$, and returns the transformed flux $\tilde{\mathbf{F}}_e^{(u)}$, where $\tilde{\mathbf{F}}_e^{(u)}$ has a size $\dim \tilde{\mathbf{F}}_e^{(u)} = N_D N_e^{(u)} \times N_V |\Omega_e|$. Because the gradients at the solution points are not needed any more, the $\mathbf{Q}_e^{(u)}$ array is overwritten by $\mathbf{F}_e^{(u)}$ and the array is referred as $\mathbf{F}_e^{(u)}$ from this point onwards.

In order to rewrite the final semi-discrete form in Equation (23) in matrix multiplication format, the following constant operator matrices are defined,

$$(\mathbf{M}_e^1)_{\zeta\sigma} = [\tilde{\mathbf{V}}_{e\zeta}^{(u)}(\tilde{\mathbf{x}})]_{\tilde{\mathbf{x}}=\tilde{\mathbf{x}}_{e\sigma}}^T, \quad \dim \mathbf{M}_e^1 = N_e^{(u)} \times N_D N_e^{(u)} \quad (45)$$

$$(\mathbf{M}_e^2)_{\zeta\sigma} = [\mathcal{C}_{e\zeta}^{(u)}(\tilde{\mathbf{x}}_{e\sigma}) \hat{\mathbf{n}}_{e\sigma}^{(f)}]^T, \quad \dim \mathbf{M}_e^2 = N_e^{(f)} \times N_D N_e^{(u)} \quad (46)$$

$$(\mathbf{M}_e^3)_{\zeta\sigma} = [\tilde{\mathbf{V}} \cdot \mathbf{g}_{e\zeta}^{(f)}(\tilde{\mathbf{x}})]_{\tilde{\mathbf{x}}=\tilde{\mathbf{x}}_{e\sigma}}, \quad \dim \mathbf{M}_e^3 = N_e^{(u)} \times N_e^{(f)}. \quad (47)$$

Then Equation (23) can be rewritten using the matrices defined above as

$$\begin{aligned} \tilde{\mathbf{R}}_e^{(u)} &= \mathbf{M}_e^3 \{ \tilde{\mathbf{D}}_e^{(f)} - \mathbf{M}_e^2 \tilde{\mathbf{F}}_e^{(u)} \} + \mathbf{M}_e^1 \tilde{\mathbf{F}}_e^{(u)} \\ &= \mathbf{M}_e^3 \tilde{\mathbf{D}}_e^{(f)} + \{ \mathbf{M}_e^1 - \mathbf{M}_e^3 \mathbf{M}_e^2 \} \tilde{\mathbf{F}}_e^{(u)}, \end{aligned} \quad (48)$$

where the $\tilde{\mathbf{R}}_e^{(u)}$ array has a size $\dim \tilde{\mathbf{R}}_e^{(u)} = N_e^{(u)} \times N_V |\Omega_e|$, equivalent to the size of the solution array $\mathbf{U}_e^{(u)}$. The kernel implementation of Equation (23) consists of two matrix multiplication kernels that correspond to the two terms in the right hand side of the equation, *tdivtconf* and *tdivpcorf* respectively. The *tdivpcorf* kernel is called first and the data in the $\tilde{\mathbf{R}}_e^{(u)}$ array is overwritten by the result as

$$\tilde{\mathbf{R}}_e^{(u)} = \{ \mathbf{M}_e^1 - \mathbf{M}_e^3 \mathbf{M}_e^2 \} \tilde{\mathbf{F}}_e^{(u)}. \quad (49)$$

Next, the *tdivtconf* kernel is executed and the result is added to $\tilde{\mathbf{R}}_e^{(u)}$ thus

$$\tilde{\mathbf{R}}_e^{(u)} = \tilde{\mathbf{R}}_e^{(u)} + \mathbf{M}_e^3 \tilde{\mathbf{D}}_e^{(f)}. \quad (50)$$

Finally, the *negdivconf* kernel is defined in order to obtain the divergence of flux in the physical space thus

$$\mathbf{R}_e^{(u)} = \text{negdivconf}(\det \mathbf{J}_e^{(u)}, \tilde{\mathbf{R}}_e^{(u)}). \quad (51)$$

2.2.1. Anti-aliasing kernels

Flux anti-aliasing Flux anti-aliasing requires two additional kernels which evaluate the solution state and gradients at quadrature points within each element, such that the *tdisf* kernel can then evaluate these fluxes at these same quadrature points. For this purpose *uqpts* and *gradcoru_qpts* kernels are defined which interpolate the solution and gradients at solution points to quadrature points. They are both matrix

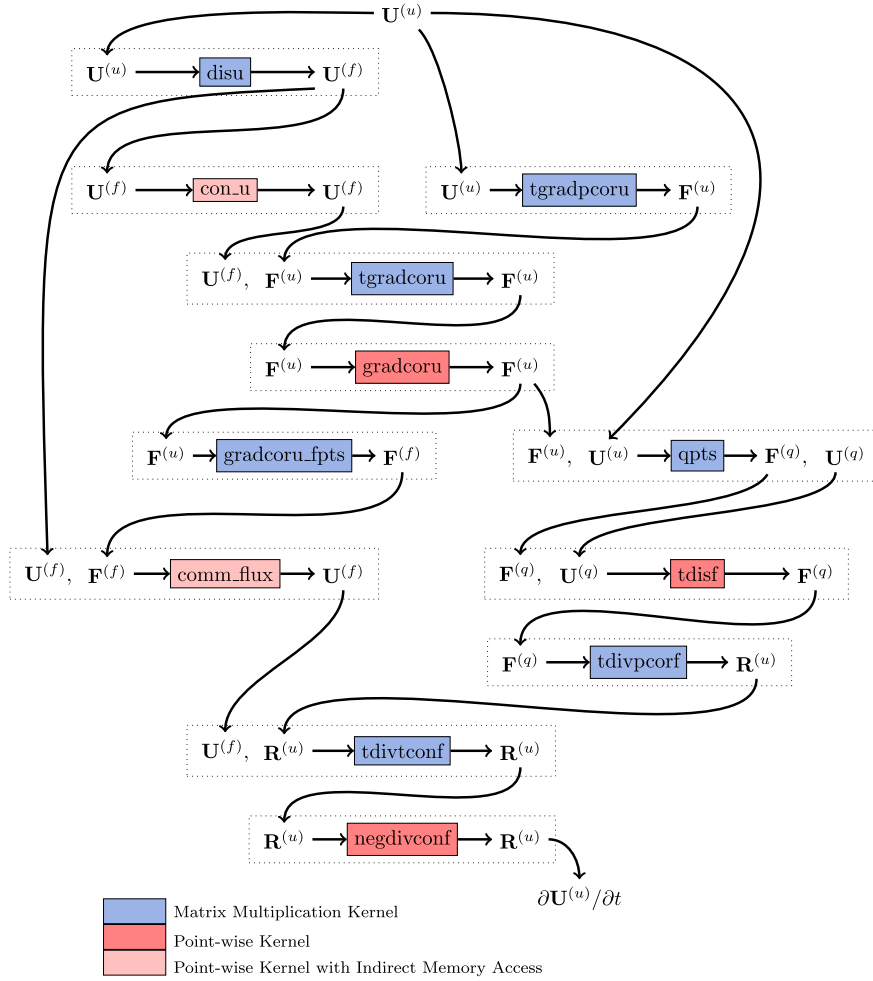


Fig. 3. Kernel prerequisites map for Navier-Stokes Solver including optional flux anti-aliasing kernels.

multiplication kernels corresponding to Equations (24) and (25) respectively. The associated operator matrix $\mathbf{M}_e^7$ for the $uqpts$ kernel is defined as

$$(\mathbf{M}_e^7)_{\sigma\zeta} = \ell_{e\zeta}^{(u)}(\tilde{\mathbf{x}}_{e\sigma}^{(q)}), \quad \dim \mathbf{M}_e^7 = N_e^{(u)} \times N_e^{(q)}, \quad (52)$$

and the $uqpts$ kernel is defined as

$$\mathbf{U}_e^{(q)} = \mathbf{M}_e^7 \mathbf{U}_e^{(u)}. \quad (53)$$

The operator matrix $\mathbf{M}_e^{10}$ for $gradcoru_qpts$ is given as

$$(\mathbf{M}_e^{10})_{\sigma\zeta} = \text{diag}(\mathbf{M}_e^7, \dots, \mathbf{M}_e^7), \quad \dim \mathbf{M}_e^{10} = N_D N_e^{(u)} \times N_D N_e^{(q)}, \quad (54)$$

and the $gradcoru_qpts$ kernel can be expressed as

$$\mathbf{Q}_e^{(q)} = \mathbf{M}_e^{10} \mathbf{Q}_e^{(u)}. \quad (55)$$

When flux anti-aliasing is turned on, the $tdisf$ kernel evaluates the flux at quadrature points using the solution and gradients at quadrature points thus it can be redefined as

$$\tilde{\mathbf{F}}_e^{(q)} = tdisf(\mathbf{U}_e^{(q)}, \mathbf{Q}_e^{(q)}, \mathbf{x}_e). \quad (56)$$

Then, the operator matrix $\mathbf{M}_e^8$ is defined to convert these modal coefficients into nodal values at the solution points.

$$(\mathbf{M}_e^8)_{\sigma\zeta} = \psi_{ev}(\tilde{\mathbf{x}}_{e\zeta}^{(u)}) \omega_{e\sigma}^{(q)} \psi_{ev}(\tilde{\mathbf{x}}_{e\sigma}^{(q)}), \quad \dim \mathbf{M}_e^8 = N_e^{(q)} \times N_e^{(u)}. \quad (57)$$

$\mathbf{M}_e^9$ is the block diagonal form of the $\mathbf{M}_e^8$ matrix and it is defined as

$$(\mathbf{M}_e^9)_{\sigma\zeta} = \text{diag}(\mathbf{M}_e^8, \dots, \mathbf{M}_e^8), \quad \dim \mathbf{M}_e^9 = N_D N_e^{(q)} \times N_D N_e^{(u)}. \quad (58)$$

This operation is carried out by the $tdivtconf$ kernel with a small update therefore when flux anti-aliasing is in use $tdivtconf$ kernel becomes

$$\tilde{\mathbf{R}}_e^{(u)} = \{\mathbf{M}_e^1 - \mathbf{M}_e^3 \mathbf{M}_e^2\} \mathbf{M}_e^9 \tilde{\mathbf{F}}_e^{(q)}. \quad (59)$$

Surface-flux anti-aliasing Surface-flux anti-aliasing on the other hand changes many of the constant operator matrices that act on flux points. Thus, surface-flux anti-aliasing can be implemented by changing the existing operator matrices such that they act on surface-quadrature points rather than flux points. For example, the operator matrix associated with the $disu$ kernel, $\mathbf{M}_e^0$, can be updated to obtain the discontinuous solution at the surface-quadrature points instead of at the flux points.

A schematic that includes all the Navier-Stokes kernels and the additional flux anti-aliasing kernels is given in Fig. 3 in order to provide a visual guide for all the operations and dependencies,

3. Kernel execution order and data movement requirements

As a baseline for comparison, data movement and bandwidth requirements are first analysed for PyFR version v1.11.0, which has no cache blocking support. Similar to analysis carried out for the Euler solver in [4], time integration related operations are not included in data movement requirements and speedups therefore will only reflect the performance of evaluating the RHS. In reality, a time integrator such as RK4 will need to be used to move forward in time which requires additional data movement that consumes bandwidth and it will slightly reduce the performance per RHS call.

Table 1
Total data movement requirement for PyFR version v1.11.0.

Data Movement Req. Per RHS Per Element [KiB/RHS/element]						
Total Bandwidth	Anti-aliasing off			Full anti-aliasing		
	$p = 3$	$p = 4$	$p = 5$	$p = 3$	$p = 4$	$p = 5$
Hexahedral	143.625	258.773	422.250	227.781	385.242	603.688
Prism	94.375	163.875	260.063	147.117	240.867	367.945

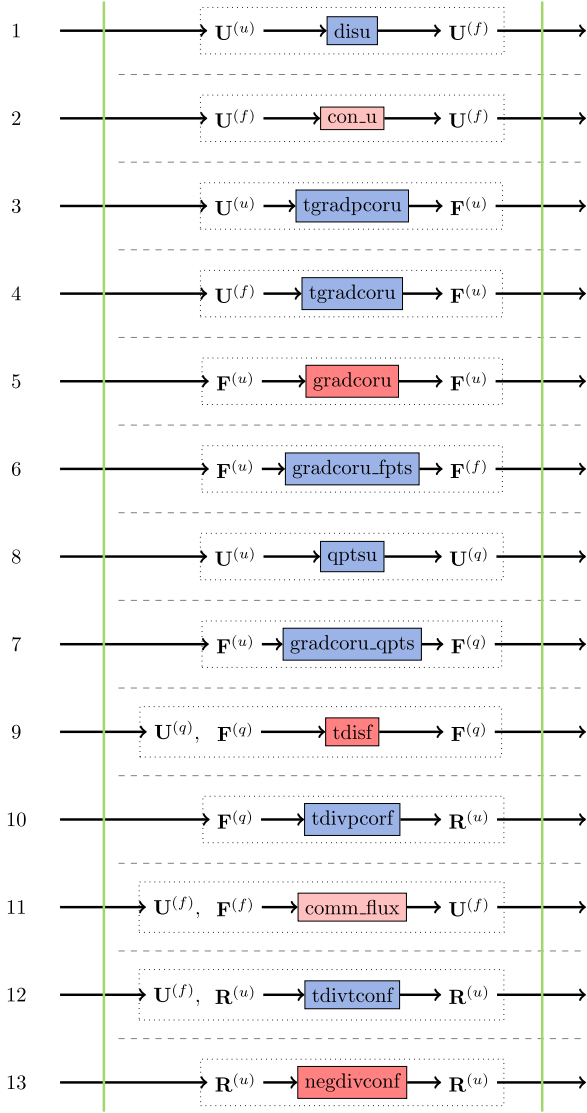


Fig. 4. Current kernel execution order. Arrows going over green line indicate main memory access.

Kernel execution order for the Navier-Stokes solver in PyFR version v1.11.0 is demonstrated in Fig. 4 including the flux anti-aliasing kernels. Without cache blocking support, each kernel is executed for the entire input and output arrays and therefore the data movement requirement is directly related to the size of input and output arrays. Data movement requirements of each kernel are tabulated in Table A.8 for hexahedral and prism meshes, normalized per element, at polynomial orders $p = 3$, $p = 4$, and $p = 5$, double precision, and including flux and surface-flux anti-aliasing on/off with quadrature point polynomial order $p_q = p + 1$ when anti-aliasing is on. We decided the span of the performance investigation regarding polynomial orders based on Verniere et al. [25] where it was demonstrated that PyFR is most efficient in terms of cost and accuracy at polynomial orders from $p = 3$ to 5, and regarding the

choice of quadrature orders based on Mengaldo et al. [26] where it was demonstrated that $p_q = p + 1$ is effective for anti-aliasing. Non-temporal stores are used whenever possible for the matrix multiplication kernels through the LIBXSMM library [23] to reduce the data movement requirements for the qualifying kernels. Total data movement requirements for polynomial orders $p = 3$, $p = 4$, and $p = 5$ for anti-aliasing on and off is summarised in Table 1.

4. Tensor product factorisation of anti-aliasing operators

Turning on anti-aliasing in PyFR brings in additional operators and changes in the existing kernels. A straightforward implementation of these new operators and changes in the existing kernels results in dense matrix operators. However, it is well known in the literature that for tensor product elements these operators can be cast as multiple sparse factors as opposed to a single dense matrix multiplication. Converting a single dense operator into a chain of sparse operations reduces the FLOP requirement significantly, but a direct implementation can make the code overly bandwidth bound and the performance gains would be limited. However for tensor product elements such as hexahedra and prisms, it is possible to factor the dense operator into sparse components. In this section, we examine ways to factor the new dense operators associated with flux anti-aliasing and the existing operators that are converted into a dense form due to surface-flux anti-aliasing. A breakdown of all the affected kernels due to flux and surface-flux anti-aliasing is shown in Table 2.

First, flux anti-aliasing brings in two new operators, M^7 and M^9 . M^7 operator interpolates data in solution points to quadrature points. If implemented directly, it is a single dense operator. However, for a hexahedral element this process can be factored into three sparse components. The process is explained in detail in [11], and can be formulated for the M^7 operator as

$$M^7 = M_z^7 M_y^7 M_x^7, \quad (60)$$

where M_z^7 , M_y^7 , and M_x^7 components correspond to the unique sparse factors that carry out interpolations in z , y , and x directions respectively. Sampling the nodal coefficients at solution points using the modal representation at quadrature points can also be factored into three sparse components. M^9 operator applies the reverse operation for the gradients at solution points, so it's a block diagonal matrix that applies the reverse interpolation for three components of the gradient at each solution point.

The interpolations between solution points and quadrature points are the main building blocks for decomposing the rest of the operators that are associated with the surface-flux anti-aliasing. The only extra operation involved with surface-flux anti-aliasing is the interpolation to flux points at the interfaces. The M^0 operator without anti-aliasing that interpolates data from solution points to flux points will be used together with the M^7 or M^9 operator to form the sparse factors of the M^0 operator with surface-flux anti-aliasing.

4.1. Extension to prism elements

The decomposition into sparse factors is apparent for hexahedral elements, however, there is no generic way of obtaining sparse factors for tetrahedral or pyramid elements. Prisms on the other hand are

Table 2

Changes in operator matrices due to anti-aliasing. Blue colour indicates a change due to flux anti-aliasing, and red colour indicates a change due to surface-flux anti-aliasing.

Kernel	Operator	Changes in the kernels due to anti-aliasing (AA)		
		flux AA	surf-flux AA	full AA
<i>disu</i>	M^0		M^0	M^0
<i>con_u</i>	Pointwise			
<i>tgradpcoru</i>	$M^4 - M^6 M^0$		$M^4 - M^6 M^0$	$M^4 - M^6 M^0$
<i>tgradcoru</i>	M^6		M^6	M^6
<i>gradcoru</i>	Pointwise			
<i>gradcoru_fpts</i>	$I_3 \otimes M^0$		$I_3 \otimes M^0$	$I_3 \otimes M^0$
<i>comm_flux</i>	Pointwise			
<i>uqpts</i>	M^7	M^7		M^7
<i>gradcoru_qpts</i>	$I_3 \otimes M^7$	$I_3 \otimes M^7$		$I_3 \otimes M^7$
<i>tdisf</i>	Pointwise			
<i>tdivtpconf</i>	$M^1 - M^3 M^2$	$(M^1 - M^3 M^2) M^9$	$M^1 - M^3 M^2$	$(M^1 - M^3 M^2) M^9$
<i>tdivtconf</i>	M^3		M^3	M^3
<i>negdivtconf</i>	Pointwise			

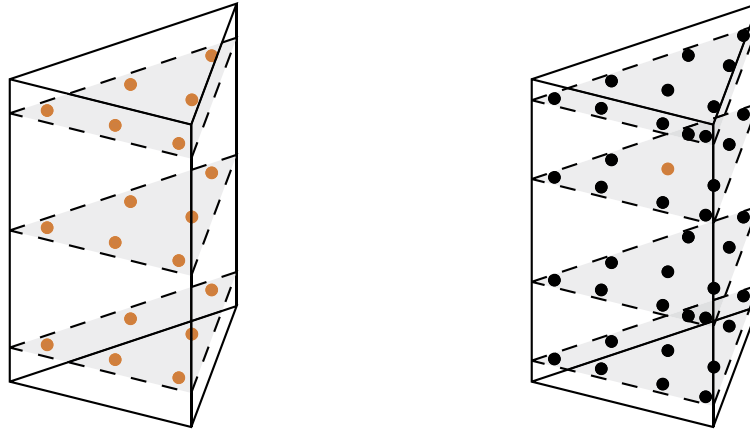


Fig. 5. Solution points for prism elements with polynomial orders $p = 2$ and $p = 3$. Point distribution is based on William-Shun at the triangular cross sections and based on Gauss-Legendre along the z -direction. A direct interpolation between $p = 2$ and $p = 3$ prisms solution points results in a dense operator matrix as any point in $p = 3$ depends on all the points in $p = 2$ as indicated with orange points.

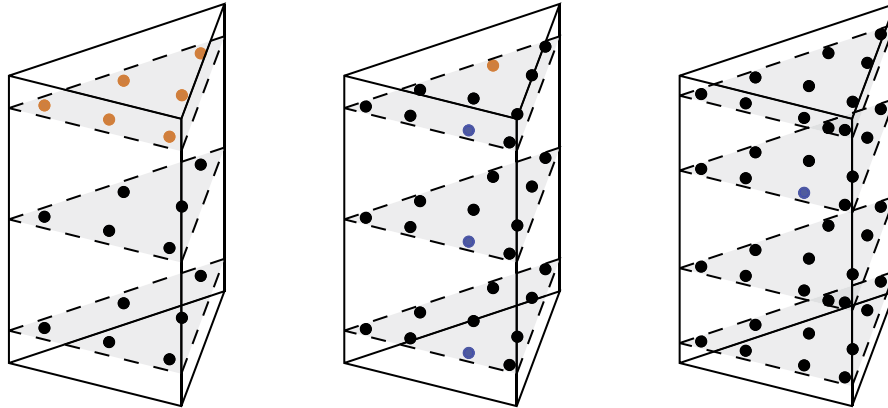


Fig. 6. An intermediate state shown in the middle is introduced to decompose interpolation from $p = 2$ to $p = 3$ into one block diagonal and one sparse factor. First, an interpolation is carried out on triangular planes only resulting in a block diagonal operator as indicated with the orange points. The next step is interpolating values along the z -direction that results in a sparse operator as indicated with the blue points.

slightly different. A direct approach for obtaining the modal coefficients at quadrature points by using the nodal coefficients at solution points in a prism would require a dense operator matrix as shown in Fig. 5.

However, a prism can be generated by a tensor production of a triangle and a line that is perpendicular to the triangle. Depending on the polynomial order, there are a number of triangular planar regions inside a prism and the solution points are located on these regions. Therefore,

obtaining modal coefficients at quadrature points using the nodal coefficients at solution points, which is an operation carried out by the *qptsu* kernel can be simplified and reformulated in a two part operation. The first operation obtains the solution at the quadrature points within triangular planes in a prism, and then a second operation obtains the solution at a set of triangular planes inside the prism. The procedure is illustrated in Fig. 6. The former requires a block structured operator

```

for (int ib = 0; ib < nblocks; ib++){
    exec(M7x, u + ib*u_blocksz, buff_x);
    exec(M7y, buff_x, buff_y);
    exec(M7z, buff_y, q + ib*q_blocksz);
}
exec(M7x, u, q);
exec(M7y, q, q_temp);
exec(M7z, q_temp, q);

```

(a) Direct implementation.

(b) Cache blocked implementation.

Fig. 7. Before (7a) and after (7b) cache blocking. M7x, M7y, and M7z correspond to the sparse factors defined in Equation (60). q_temp is a domain sized array, while buff_x and buff_y are small cache-block sized arrays.

Table 3

FLOP and Data movement requirements for the *tdivtconf* kernel on hexahedral elements for three different implementations.

	Factored		
	Dense	Direct	Cache Blocked
FLOP [GFLOPS]	30.00	4.50	4.50
Data Movement [GiB]	1.95	7.81	1.95
FLOP/Byte	14.48	0.54	2.16

while the latter is a sparse operation. As there is a sparse component in the factored formulation, cache blocking can be used effectively to implement this.

5. Application of cache blocking in tensor product factorisation

There are a few different ways to implement the tensor product operations after obtaining the sparse factorisation. The naive implementation would simply call a matrix multiplication library for each sparse factor. This would reduce the FLOP requirement significantly due to reduced number of non-zero entries in the sparse factors provided that a sparse matrix multiplication library is used. However, it would result in increased bandwidth use because the intermediate solutions between each of the separate kernel invocations have to be stored in memory and this results in unnecessary reads and writes. The alternative is using cache blocking strategies. These make it possible to store the intermediate results in CPU cache, and reduce the main memory access and data movement requirements significantly. Using sparse factors rather than a single dense matrix reduces the FLOP requirement, and cache blocking strategy eliminates the excess data movement by utilizing CPU cache as an intermediate storage. Therefore, it is possible to get the best of both worlds by enabling cache blocking for the tensor product operations. A code snippet is provided in Fig. 7 to demonstrate the differences in implementation. Note that cache blocked implementation requires a change in the underlying data structure as explained in [4]. The blocked AoSoA data structure enables data locality in a group of elements thus makes it possible to store the intermediate results in CPUs cache between sparse factor multiplications.

Table 3 shows how the FLOP and data movement requirements differ between the regular approach where a single dense matrix is used, the naive approach where a matrix multiplication library is called for each sparse factor, and the ideal approach where cache blocking is enabled together with the sparse factors. Furthermore, a roofline model [27] demonstrating the FLOP/Byte ratio for a selection of Navier-Stokes kernels with full anti-aliasing is given in Fig. 8 before and after applying tensor product factorisation using the cache blocking strategy. It is clear from Table 3 and Fig. 8 that the tensor product factorisation using the cache blocking strategy for the standalone tensor product kernels reduces the FLOP requirement while keeping the data movement requirement from main memory the same. As a result, the profile of these individual kernels shifted towards a smaller FLOP/Byte region where they would normally be bandwidth bound from main memory. However, now that most of the kernels are bandwidth bound, a kernel grouping strategy can be implemented to eliminate the unnecessary

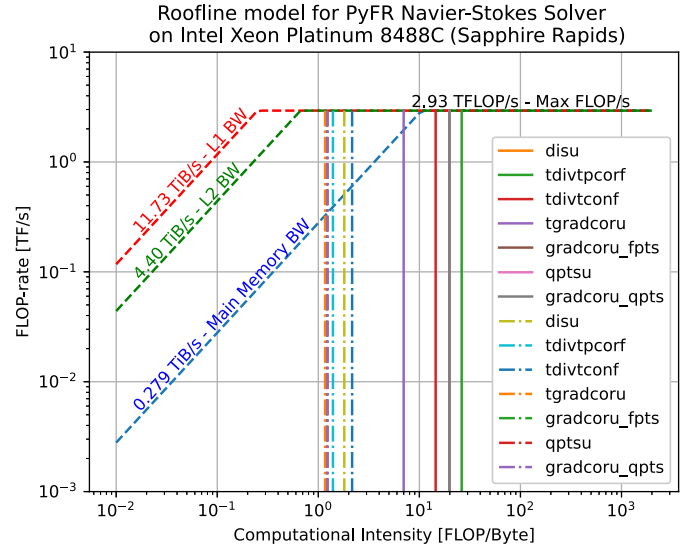


Fig. 8. A selection of the Navier-Stokes kernels are shown in the figure for anti-aliasing off and full anti-aliasing on a full hexahedral mesh. The remaining kernels have similar properties if not exactly the same. Solid and dashed vertical lines indicate the theoretical arithmetic intensity of the given kernels before and after tensor product factorisation.

data movements between kernels that exchange data. When these approaches are implemented together, there is a reduction in both FLOP and data movement requirements.

In summary, tensor product factorisation with cache blocking reduces the FLOP requirement significantly while keeping the amount of data movement from main memory the same. Then, a kernel grouping strategy can be implemented to reduce the total amount of data movement requirements to improve the performance even further.

6. Theoretical predictions and benchmarking in PyFR

A kernel grouping configuration for the Navier-Stokes solver is examined in this section both for anti-aliasing on. The test case used for the performance comparisons is the compressible Taylor-Green Vortex. It is a typical test case for the validation of high-order software and it is used in [25] for comparing PyFR performance against industry standard codes. Also, it is one of the test cases in the high-order workshops [28]. The configuration and mesh files used here are taken from Vermeire [25], and the numerical results are compared against spectral DNS results from van Rees et al. [29]. Taylor-Green Vortex case is initialized with the following initial conditions

$$\begin{aligned}
 v_x &= +U_0 \sin(x/L) \cos(y/L) \cos(z/L), \\
 v_y &= +U_0 \cos(x/L) \sin(y/L) \cos(z/L), \\
 v_z &= 0, \\
 p &= p_0 + \frac{\rho_0 U_0^2}{16} (\cos(2x/L) + \cos(2y/L)) (\cos(2z/L) + 2), \\
 \rho &= \frac{p}{RT_0}
 \end{aligned} \tag{61}$$

in a cube domain with periodic boundary conditions in all the faces with dimensions $[-\pi L, +\pi L]^3$. Performance comparisons of the cache blocking approach are executed at three different polynomial orders, $p = 3$, $p = 4$, and $p = 5$, with anti-aliasing on and off, and on hexahedral and prism meshes. The Reynolds number $Re = 1600$ based on the length L and velocity U_0 . The Mach number based on U_0 is $M = 0.1$ so that the simulation is effectively incompressible. Firstly, a simulation is carried out using the cache blocking approach in order to validate the imple-

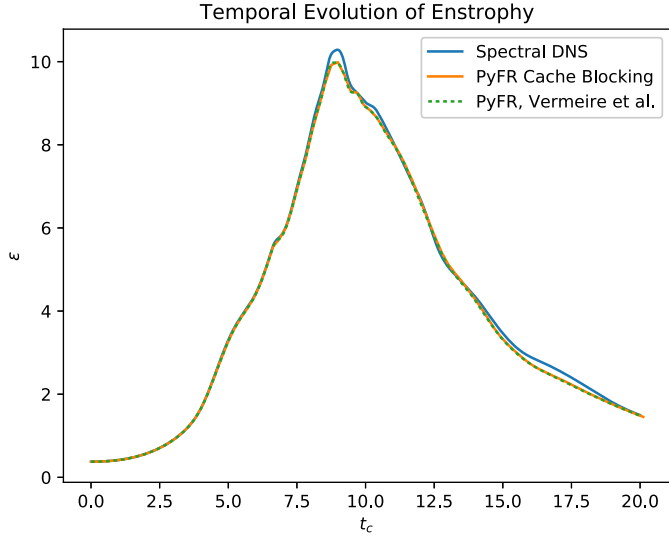


Fig. 9. Temporal evolution of enstrophy obtained with cache blocking approach implemented in PyFR. Simulation is carried out on a 52^3 all hexahedral mesh at $p = 4$ with anti-aliasing off and compared against the reference results of Vermeire et al. [25].

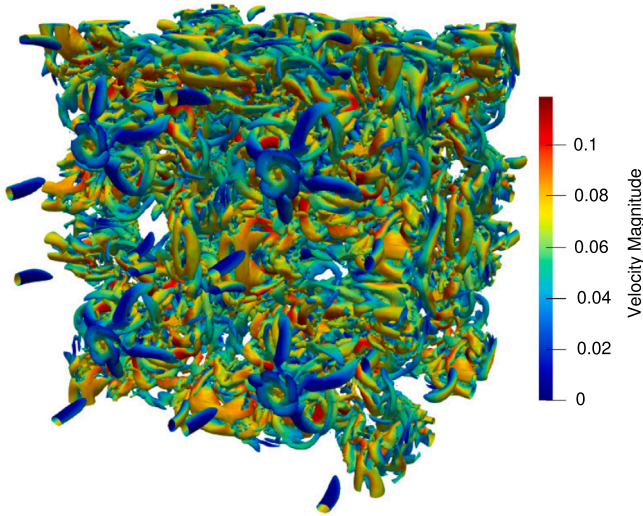


Fig. 10. Q-criterion isosurfaces coloured based on velocity magnitude for a snapshot at $t = 10t_c$.

mentation. The evolution of enstrophy is compared with the spectral DNS result [29] and shown in Fig. 9. The results from cache blocking show a good agreement with the spectral DNS results. A snapshot at $t = 10t_c$ is shown in Fig. 10 with isosurfaces based on Q-criterion [30] and coloured by velocity magnitude.

6.1. Kernel grouping configuration

The main consideration when forming kernel groupings is the amount of bandwidth saving. Bandwidth saving is maximised when a chain of kernels that write and read from the same location in the memory grouped together. The more data kernels in a group exchange by writing and reading the same piece of memory the better the savings and therefore speedups. Therefore the first step is analysing the kernels in the Navier-Stokes solver and determines chain of kernels that exchange data. When compared with the Euler solver, there are more constraints in Navier-Stokes solver and this limits the number of possible kernel grouping configurations. As with the Euler solver, eliminating

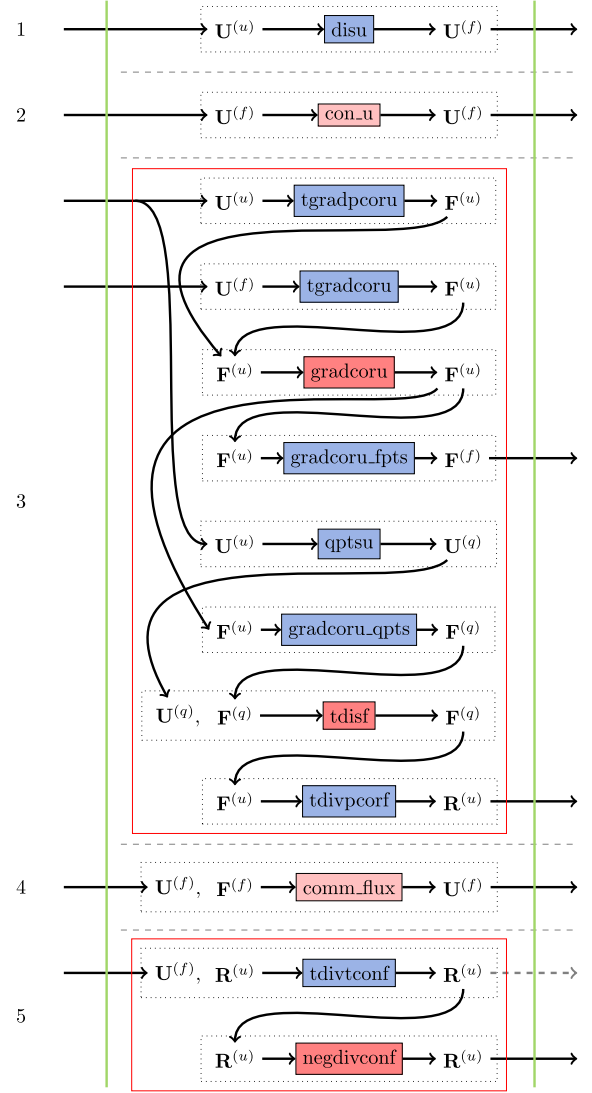


Fig. 11. Kernel grouping configuration of the Navier-Stokes solver with flux anti-aliasing. Arrows going over green line indicate main memory access. Dashed gray arrow indicates write orders avoided due to cache blocking, while solid arrow indicates a definite read/write operation. Constant arrays such as x , which are not changed by the kernels, are omitted for simplicity.

the intermediate storage array $\mathbf{F}^{(u)}$ is possible in Navier-Stokes solver as well and it saves a considerable amount of bandwidth. There are 6 kernels in total that operate on $\mathbf{F}^{(u)}$, and a kernel group with all of them can save a significant amount of data movement. The kernels are *tgradpcoru*, *tgradcoru*, *gradcoru*, *gradcoru_fpts*, *tdisf*, and *tdivpcorf*. In case of flux anti-aliasing the additional kernels, *qptsu* and *gradcoru_qpts*, are also added to this kernel grouping as they also make use of $\mathbf{F}^{(u)}$ as well as introducing a new temporary storages $\mathbf{U}^{(q)}$ and $\mathbf{F}^{(q)}$ that can be stored entirely in the CPU cache for a single block. Afterwards, there is only one other kernel grouping possibility is left which consist of *tdivtconf* and *negdivconf*. A schematic of the kernel grouping configuration is given in Fig. 11 for full anti-aliasing. Data movement requirements of the present kernel grouping configuration with anti-aliasing on and off are tabulated in Table A.9. Further, the speedups obtained in three different platforms with an Arm based AWS Graviton3, Intel Xeon Platinum 8488C and AMD EPYC 9R14 are given in Tables 4 and 5. The highest speedup is 3.91x compared to PyFR v1.11.0 and it is obtained when running at $p = 5$ with full anti-aliasing on Intel Xeon Platinum 8488C CPU. GDOF/s parameters are also tabulated in Tables 6 and 7. ‘GDOF/s’ is the num-

Table 4

Theoretical speedup predictions compared with observed speedups on hexahedral meshes on AMD EPYC 9R14, Arm based AWS Graviton3, and Intel Xeon Platinum 8488C CPUs.

Hexa	Anti-aliasing off			Full anti-aliasing		
	$p = 3$	$p = 4$	$p = 5$	$p = 3$	$p = 4$	$p = 5$
Prediction	2.09	2.30	2.50	2.29	2.57	2.83
AMD	2.45	2.77	2.97	3.07	3.46	3.60
Arm	2.17	2.53	2.56	2.82	2.03	1.91
Intel	2.61	2.76	2.92	2.83	3.67	3.91

Table 5

Theoretical speedup predictions compared with observed speedups on prism meshes on AMD EPYC 9R14, Arm based AWS Graviton3, and Intel Xeon Platinum 8488C CPUs.

Prism	Anti-aliasing off			Full anti-aliasing		
	$p = 3$	$p = 4$	$p = 5$	$p = 3$	$p = 4$	$p = 5$
Prediction	1.99	2.16	2.31	2.14	2.37	2.58
AMD	1.79	1.86	2.04	1.80	2.02	2.13
Arm	1.76	1.78	2.03	1.68	1.67	1.51
Intel	1.83	1.90	2.15	1.95	2.08	2.25

Table 6

Comparison of GDOF/s: PyFR with Cache Blocking on hexahedral meshes on AMD EPYC 9R14 CPU, Arm based AWS Graviton3 CPU, Intel Xeon Platinum 8488C CPU.

Hexa	Anti-aliasing off			Full anti-aliasing		
	$p = 3$	$p = 4$	$p = 5$	$p = 3$	$p = 4$	$p = 5$
AMD	1.393	1.680	1.893	0.846	0.965	0.800
Arm	0.921	1.162	1.215	0.538	0.339	0.237
Intel	0.830	0.983	1.093	0.499	0.596	0.561

Table 7

Comparison of GDOF/s: PyFR with Cache Blocking on prism meshes on AMD EPYC 9R14 CPU, Arm based AWS Graviton3 CPU, Intel Xeon Platinum 8488C CPU.

Prism	Anti-aliasing off			Full anti-aliasing		
	$p = 3$	$p = 4$	$p = 5$	$p = 3$	$p = 4$	$p = 5$
AMD	0.928	1.072	1.148	0.519	0.589	0.596
Arm	0.617	0.702	0.768	0.347	0.359	0.253
Intel	0.550	0.632	0.679	0.310	0.348	0.355

ber of degree of freedom processed per second, specifically calculated here as

$$\text{GDOF/s} = \sum_e \frac{N_V \cdot |\Omega_e| \cdot N_e^{(u)}}{10^9 \cdot t_{\text{RHS}}},$$

where t_{RHS} is wall time per right hand side evaluation. We use this normalised parameter to compare performance of PyFR on different platforms.

7. Conclusion

In this article, a methodology for applying cache blocking in FR has been extended and implemented for the Navier-Stokes solver with full anti-aliasing support on mixed grids in PyFR for CPUs. Cache blocking was used to eliminate unnecessary data movements between kernels that exchange data at the main memory level, as an alternative to kernel fusion. In particular, per-core private data cache on CPUs was used as a temporary storage to enable data exchanges between consecutive

kernels. As the amount of cache is limited, these kernels are executed on small sub-regions of the domain that fit in the available cache. In addition to eliminating data exchanges between kernels, cache blocking was used to implement dense interpolation kernels associated with anti-aliasing as well. These dense interpolation kernels first decomposed into sparse matrix factors using tensor product factorisation, and cache blocking is used as a temporary storage for the intermediate results when applying these chain of sparse matrix factors consecutively. As a result, cache blocking saved a significant amount of data movement from main memory. The amount of savings in data movements is then used to construct a theoretical model to predict the performance gains. Finally, the implementation was benchmarked using a compressible 3D Taylor-Green vortex test case on both hexahedral and prismatic grids, with third-, fourth-, and fifth-order solution polynomials.

CRedit authorship contribution statement

Semih Akkurt: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Freddie Witherden:** Conceptualization, Funding acquisition, Methodology, Writing – review & editing. **Peter Vincent:** Conceptualization, Funding acquisition, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the code patches, configuration files, and mesh files in the attach file step as supplementary material.

Acknowledgements

The first author would like to thank Imperial College London for their support via the President's PhD Scholarship (2900572) and Engineering and Physical Sciences Research Council for their support via a Doctoral Training Partnership (EP/R513052/1). The authors would like to thank the Engineering and Physical Sciences Research Council for their support via an Early Career Fellowship (EP/R030340/1), the European Union for their support via the SciFiTurbo Project (101138080), the Air Force Office of Scientific Research for their support via grants FA9550-21-1-0190 ("Enabling next-generation heterogeneous computing for massively parallel high-order compressible CFD") and FA9550-23-1-0232 ("Next Generation High-Order Methods for Multi-Physics Multi-Scale Problems") under the direction of Dr. Fariba Fahroo.

Appendix A. Data movement requirements

A breakdown of data movement requirements for PyFR v1.11.0 and PyFR with kernel grouping are provided in Tables A.8 and A.9.

Appendix B. Supplementary material

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.cpc.2024.109332>.

Table A.8
Current Navier-Stokes Data Movement Requirements.

Data Movement Requirements Per RHS Per Element [KiB/RHS/element]															
Kernel	Input	Output	Formula	Hexahedral						Prism					
				Anti-aliasing off			Full anti-aliasing			Anti-aliasing off			Full anti-aliasing		
				$p = 3$	$p = 4$	$p = 5$	$p = 3$	$p = 4$	$p = 5$	$p = 3$	$p = 4$	$p = 5$	$p = 3$	$p = 4$	$p = 5$
<i>disu</i>	$\mathbf{U}_e^{(u)}$	$\mathbf{U}_e^{(f)}$	$S(\mathbf{U}_e^{(u)}) + S(\mathbf{U}_e^{(f)})$	6.25	10.74	16.88	8.36	13.32	19.92	4.22	7.03	10.78	5.66	8.79	12.85
<i>con_u</i>	$\mathbf{D}_e^{(f)}, \mathbf{n}_e^{(f)}$	$\mathbf{D}_e^{(f)}$	$S(\mathbf{n}_e^{(f)}) + 2S(\mathbf{D}_e^{(f)})$	10.50	16.41	23.63	16.41	23.63	32.16	7.44	11.48	16.41	11.48	16.41	22.20
<i>tgradpcoru</i>	$\mathbf{U}_e^{(u)}$	$\mathbf{F}_e^{(u)}$	$S(\mathbf{U}_e^{(u)}) + S(\mathbf{F}_e^{(u)})$	10.00	19.53	33.75	10.00	19.53	33.75	6.25	11.72	19.69	6.25	11.72	19.69
<i>tgradcoru</i>	$\mathbf{U}_e^{(f)}$	$\mathbf{F}_e^{(u)}$	$S(\mathbf{U}_e^{(f)}) + 2S(\mathbf{F}_e^{(u)})$	18.75	35.16	59.06	20.86	37.73	62.11	12.03	21.68	35.39	13.48	23.44	37.46
<i>gradcoru</i>	$\mathbf{F}_e^{(u)}, \mathbf{x}_e, \mathbf{J}_e^{(u)}$	$\mathbf{F}_e^{(u)}$	$S(\mathbf{J}_e^{(u)}) + S(\mathbf{x}_e) + 2S(\mathbf{F}_e^{(u)})$	15.69	30.46	52.50	15.69	30.46	52.50	9.83	18.30	30.66	9.83	18.30	30.66
<i>gradcoru_fpts</i>	$\mathbf{F}_e^{(u)}$	$\mathbf{F}_e^{(f)}$	$S(\mathbf{F}_e^{(u)}) + S(\mathbf{F}_e^{(f)})$	18.75	32.23	50.63	25.08	39.96	59.77	12.66	21.09	32.34	16.99	26.37	38.55
<i>uqpts</i>	$\mathbf{U}_e^{(u)}$	$\mathbf{U}_e^{(q)}$	$S(\mathbf{U}_e^{(u)}) + S(\mathbf{U}_e^{(q)})$	-	-	-	7.38	13.32	21.84	-	-	-	4.49	7.85	12.58
<i>gradcoru_qpts</i>	$\mathbf{F}_e^{(u)}$	$\mathbf{F}_e^{(q)}$	$S(\mathbf{F}_e^{(u)}) + S(\mathbf{F}_e^{(q)})$	-	-	-	22.15	39.96	65.51	-	-	-	13.48	23.55	37.73
<i>tdisf</i>	$\mathbf{U}_e^{(u/q)}, \mathbf{x}_e, \mathbf{F}_e^{(u/q)}$	$\mathbf{F}_e^{(u/q)}$	$S(\mathbf{U}_e^{(u/q)}) + S(\mathbf{x}_e) + 2S(\mathbf{F}_e^{(u/q)})$	17.69	34.37	59.25	34.37	59.25	93.98	11.08	20.65	34.59	20.65	34.59	53.73
<i>tdivpcorf</i>	$\mathbf{F}_e^{(u/q)}$	$\mathbf{R}_e^{(u)}$	$S(\mathbf{F}_e^{(u/q)}) + S(\mathbf{R}_e^{(u)})$	10.00	19.53	33.75	17.15	30.20	48.63	6.25	11.72	19.69	10.35	17.70	27.89
<i>comm_flux</i>	$\mathbf{D}_e^{(f)}, \mathbf{F}_e^{(f)}, \mathbf{n}_e^{(f)}$	$\mathbf{D}_e^{(f)}$	$S(\mathbf{n}_e^{(f)}) + S(\mathbf{F}_e^{(f)}) + 2S(\mathbf{D}_e^{(f)})$	21.75	33.98	48.94	33.98	48.94	66.61	15.41	23.79	33.98	23.79	33.98	45.99
<i>tdivtconf</i>	$\mathbf{U}_e^{(f)}$	$\mathbf{R}_e^{(u)}$	$S(\mathbf{U}_e^{(f)}) + 2S(\mathbf{R}_e^{(u)})$	8.75	15.63	25.31	10.86	18.20	28.36	5.78	9.96	15.70	7.23	11.72	17.77
<i>negdivconf</i>	$\mathbf{R}_e^{(u)}, \mathbf{J}_e^{(u)}$	$\mathbf{R}_e^{(u)}$	$S(\mathbf{J}_e^{(u)}) + 2S(\mathbf{R}_e^{(u)})$	5.50	10.74	18.56	5.50	10.74	18.56	3.44	6.45	10.83	3.44	6.45	10.83
Total Bandwidth				143.63	258.77	422.25	227.78	385.24	603.69	94.38	163.88	260.06	147.12	240.87	367.95

Table A.9
Navier-Stokes Data Movement Requirements with kernel grouping.

Data Movement Requirements Per RHS Per Element [KiB/RHS/element]															
Kernel	Input	Output	Formula	Hexahedral						Prism					
				Anti-aliasing off			Full anti-aliasing			Anti-aliasing off			Full anti-aliasing		
				$p = 3$	$p = 4$	$p = 5$	$p = 3$	$p = 4$	$p = 5$	$p = 3$	$p = 4$	$p = 5$	$p = 3$	$p = 4$	$p = 5$
<i>disu</i>	$\mathbf{U}_e^{(u)}$	$\mathbf{U}_e^{(f)}$	$S(\mathbf{U}_e^{(u)}) + S(\mathbf{U}_e^{(f)})$	6.25	10.74	16.88	8.36	13.32	19.92	4.22	7.03	10.78	5.66	8.79	12.85
<i>con_u</i>	$\mathbf{D}_e^{(f)}, \mathbf{n}_e^{(f)}$	$\mathbf{D}_e^{(f)}$	$S(\mathbf{n}_e^{(f)}) + 2S(\mathbf{D}_e^{(f)})$	10.50	16.41	23.63	16.41	23.63	32.16	7.44	11.48	16.41	11.48	16.41	22.20
<i>tgradpcoru</i>	$\mathbf{U}_e^{(u)}$	$\mathbf{F}_e^{(u)}$	$S(\mathbf{U}_e^{(u)})$	2.50	4.88	8.44	2.50	4.88	8.44	1.56	2.93	4.92	1.56	2.93	4.92
<i>tgradcoru</i>	$\mathbf{U}_e^{(f)}$	$\mathbf{F}_e^{(u)}$	$S(\mathbf{U}_e^{(f)})$	3.75	5.86	8.44	5.86	8.44	11.48	2.66	4.10	5.86	4.10	5.86	7.93
<i>gradcoru</i>	$\mathbf{F}_e^{(u)}, \mathbf{x}_e, \mathbf{J}_e^{(u)}$	$\mathbf{F}_e^{(u)}$	$S(\mathbf{J}_e^{(u)}) + S(\mathbf{x}_e)$	0.69	1.16	1.88	0.69	1.16	1.88	0.45	0.73	1.13	0.45	0.73	1.13
<i>gradcoru_fpts</i>	$\mathbf{F}_e^{(u)}$	$\mathbf{F}_e^{(f)}$	$S(\mathbf{F}_e^{(f)})$	11.25	17.58	25.31	17.58	25.31	34.45	7.97	12.30	17.58	12.30	17.58	23.79
<i>uqpts</i>	$\mathbf{U}_e^{(u)}$	$\mathbf{U}_e^{(q)}$	0	-	-	-	0.00	0.00	0.00	-	-	-	0.00	0.00	0.00
<i>gradcoru_qpts</i>	$\mathbf{F}_e^{(u)}$	$\mathbf{F}_e^{(q)}$	0	-	-	-	0.00	0.00	0.00	-	-	-	0.00	0.00	0.00
<i>tdisf</i>	$\mathbf{U}_e^{(u/q)}, \mathbf{x}_e, \mathbf{F}_e^{(u/q)}$	$\mathbf{F}_e^{(u/q)}$	$S(\mathbf{x}_e)$	0.19	0.19	0.19	0.19	0.19	0.19	0.14	0.14	0.14	0.14	0.14	0.14
<i>tdivpcorf</i>	$\mathbf{F}_e^{(u)}$	$\mathbf{R}_e^{(u)}$	$S(\mathbf{R}_e^{(u)})$	2.50	4.88	8.44	2.50	4.88	8.44	1.56	2.93	4.92	1.56	2.93	4.92
<i>comm_flux</i>	$\mathbf{D}_e^{(f)}, \mathbf{F}_e^{(f)}, \mathbf{n}_e^{(f)}$	$\mathbf{D}_e^{(f)}$	$S(\mathbf{n}_e^{(f)}) + S(\mathbf{F}_e^{(f)}) + 2S(\mathbf{D}_e^{(f)})$	21.75	33.98	48.94	33.98	48.94	66.61	15.41	23.79	33.98	23.79	33.98	45.99
<i>tdivtconf</i>	$\mathbf{U}_e^{(f)}$	$\mathbf{R}_e^{(u)}$	$S(\mathbf{U}_e^{(f)}) + S(\mathbf{R}_e^{(u)})$	6.25	10.74	16.88	8.36	13.32	19.92	4.22	7.03	10.78	5.66	8.79	12.85
<i>negdivconf</i>	$\mathbf{R}_e^{(u)}, \mathbf{J}_e^{(u)}$	$\mathbf{R}_e^{(u)}$	$S(\mathbf{J}_e^{(u)}) + S(\mathbf{R}_e^{(u)})$	3.00	5.86	10.13	3.00	5.86	10.13	1.88	3.52	5.91	1.88	3.52	5.91
Total				68.63	112.29	169.13	99.42	149.93	213.61	47.50	75.98		66.60	101.65	142.63
Predicted Speedup				2.09	2.30	2.50	2.29	2.57	2.83	1.99	2.16	2.31	2.14	2.37	2.58

References

- [1] M. Wahib, N. Maruyama, Scalable kernel fusion for memory-bound gpu applications, in: SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2014, pp. 191–202.
- [2] A. Li, B. Zheng, G. Pekhimenko, F. Long, Automatic horizontal fusion for gpu kernels, in: 2022 IEEE/ACM International Symposium on Code Generation and Optimization (CGO), 2022, pp. 14–27.
- [3] W. Sun, A. Li, S. Stuijk, H. Corporaal, How much can we gain from tensor kernel fusion on gpus?, IEEE Access (2024) 1, <https://doi.org/10.1109/ACCESS.2024.3411473>.
- [4] S. Akkurt, F. Witherden, P. Vincent, Cache blocking strategies applied to flux reconstruction, Comput. Phys. Commun. 271 (2022) 108193, <https://doi.org/10.1016/j.cpc.2021.108193>.
- [5] J.S. Park, F.D. Witherden, P.E. Vincent, High-order implicit large-eddy simulations of flow over a naca0021 aerofoil, AIAA J. 55 (7) (2017) 2186–2197, <https://doi.org/10.2514/1.J055304>, [arXiv:https://doi.org/10.2514/1.J055304](https://arxiv.org/abs/10.2514/1.J055304).
- [6] S.A. Orszag, Spectral methods for problems in complex geometries, J. Comput. Phys. 37 (1) (1980) 70–92, [https://doi.org/10.1016/0021-9991\(80\)90005-4](https://doi.org/10.1016/0021-9991(80)90005-4), <https://www.sciencedirect.com/science/article/pii/S0021999180900054>.
- [7] S.J. Sherwin, G.E. Karniadakis, A new triangular and tetrahedral basis for high-order (hp) finite element methods, Int. J. Numer. Methods Eng. 38 (22) (1995) 3775–3802, <https://doi.org/10.1002/nme.1620382204>, [arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/nme.1620382204](https://onlinelibrary.wiley.com/doi/pdf/10.1002/nme.1620382204), <https://onlinelibrary.wiley.com/doi/abs/10.1002/nme.1620382204>.
- [8] C.D. Cantwell, S.J. Sherwin, R.M. Kirby, P.H.J. Kelly, From h to p efficiently: selecting the optimal spectral/hp discretisation in three dimensions, Math. Model. Nat. Phenom. 6 (3) (2011) 84–96, <https://doi.org/10.1051/mmnp/20116304>.
- [9] D. Moxey, C. Cantwell, R. Kirby, S. Sherwin, Optimising the performance of the spectral/hp element method with collective linear algebra operations, Comput. Methods Appl. Mech. Eng. 310 (2016) 628–645, <https://doi.org/10.1016/j.cma.2016.07.001>, <https://www.sciencedirect.com/science/article/pii/S0045782516306739>.
- [10] A. Bolis, C.D. Cantwell, R.M. Kirby, S.J. Sherwin, From h to p efficiently: optimal implementation strategies for explicit time-dependent problems using the spectral/hp element method, Int. J. Numer. Methods Fluids 75 (8) (2014) 591–607, <https://doi.org/10.1002/fld.3909>, [arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/fld.3909](https://onlinelibrary.wiley.com/doi/pdf/10.1002/fld.3909), <https://onlinelibrary.wiley.com/doi/abs/10.1002/fld.3909>.
- [11] K. Świrydowicz, N. Chalmers, A. Karakus, T. Warburton, Acceleration of tensor-product operations for high-order finite element methods, Int. J. High Perform. Comput. Appl. 33 (4) (2019) 735–757, <https://doi.org/10.1177/1094342018816368>, [arXiv:https://doi.org/10.1177/1094342018816368](https://arxiv.org/abs/10.1177/1094342018816368).
- [12] H.T. Huynh, A flux reconstruction approach to high-order schemes including discontinuous Galerkin methods, in: 18th AIAA Computational Fluid Dynamics Conference, 2007.
- [13] F. Witherden, P. Vincent, On nodal point sets for flux reconstruction, J. Comput. Appl. Math. 381 (2021) 113014, <https://doi.org/10.1016/j.cam.2020.113014>, <https://www.sciencedirect.com/science/article/pii/S0377042720303058>.
- [14] F. Witherden, P. Vincent, On the identification of symmetric quadrature rules for finite element methods, Comput. Math. Appl. 69 (10) (2015) 1232–1241, <https://doi.org/10.1016/j.camwa.2015.03.017>, <https://www.sciencedirect.com/science/article/pii/S0898122115001224>.
- [15] F. Witherden, J. Park, P. Vincent, An analysis of solution point coordinates for flux reconstruction schemes on tetrahedral elements, J. Sci. Comput. 69 (2) (2016) 905–920, <https://doi.org/10.1007/s10915-016-0204-y>, <https://link.springer.com/article/10.1007/s10915-016-0204-y>.
- [16] J.S. Hesthaven, T. Warburton, Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications, Springer Science & Business Media, 2007.
- [17] P. Castonguay, D. Williams, P. Vincent, A. Jameson, Energy stable flux reconstruction schemes for advection–diffusion problems, Comput. Methods Appl. Mech. Eng. 267 (2013) 400–417, <https://doi.org/10.1016/j.cma.2013.08.012>, <https://www.sciencedirect.com/science/article/pii/S0045782513002156>.
- [18] P. Vincent, P. Castonguay, A. Jameson, Insights from von Neumann analysis of high-order flux reconstruction schemes, J. Comput. Phys. 230 (22) (2011) 8134–8154, <https://doi.org/10.1016/j.jcp.2011.07.013>, <https://www.sciencedirect.com/science/article/pii/S0021999111004323>.
- [19] D.A. Kopriva, A staggered-grid multidomain spectral method for the compressible Navier–Stokes equations, J. Comput. Phys. 143 (1) (1998) 125–158.
- [20] Y. Sun, Z. Wang, Y. Liu, High-order multidomain spectral difference method for the Navier–Stokes equations on unstructured hexahedral grids, Commun. Comput. Phys. 2 (2) (2007) 310–333.
- [21] E.F. Toro, Riemann Solvers and Numerical Methods for Fluid Dynamics a Practical Introduction, 3rd edition, Springer, Berlin, 2006.
- [22] V. Rusanov, The calculation of the interaction of non-stationary shock waves and obstacles, USSR Comput. Math. Math. Phys. 1 (2) (1962) 304–320.
- [23] A. Heinecke, G. Henry, M. Hutchinson, H. Pabst, LIBXSMM: accelerating small matrix multiplications by runtime code generation, in: SC '16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2016, pp. 981–991.
- [24] B.D. Wozniak, F.D. Witherden, F.P. Russell, P.E. Vincent, P.H. Kelly, GiMMiK—generating bespoke matrix multiplication kernels for accelerators: application to high-order computational fluid dynamics, Comput. Phys. Commun. 202 (2016) 12–22, <https://doi.org/10.1016/j.cpc.2015.12.012>.
- [25] B. Vermeire, F. Witherden, P. Vincent, On the utility of gpu accelerated high-order methods for unsteady flow simulations: a comparison with industry-standard tools, J. Comput. Phys. 334 (2017) 497–521, <https://doi.org/10.1016/j.jcp.2016.12.049>, <https://www.sciencedirect.com/science/article/pii/S0021999116307136>.
- [26] G. Mengaldo, D. De Grazia, D. Moxey, P. Vincent, S. Sherwin, Dealiasing techniques for high-order spectral element methods on regular and irregular grids, J. Comput. Phys. 299 (2015) 56–81, <https://doi.org/10.1016/j.jcp.2015.06.032>, <https://www.sciencedirect.com/science/article/pii/S0021999115004301>.
- [27] S. Williams, A. Waterman, D. Patterson, Roofline: an insightful visual performance model for multicore architectures, Commun. ACM 52 (4) (2009) 65–76, <https://doi.org/10.1145/1498765.1498785>.
- [28] Z. Wang, K. Fidkowski, R. Abgrall, F. Bassi, D. Caraeni, A. Cary, H. Deconinck, R. Hartmann, K. Hillewaert, H. Huynh, N. Kroll, G. May, P.-O. Persson, B. van Leer, M. Visbal, High-order CFD methods: current status and perspective, Int. J. Numer. Methods Fluids 72 (8) (2013) 811–845, <https://doi.org/10.1002/fld.3767>, [arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/fld.3767](https://onlinelibrary.wiley.com/doi/pdf/10.1002/fld.3767), <https://onlinelibrary.wiley.com/doi/abs/10.1002/fld.3767>.
- [29] W.M. van Rees, A. Leonard, D. Pullin, P. Koumoutsakos, A comparison of vortex and pseudo-spectral methods for the simulation of periodic vortical flows at high Reynolds numbers, J. Comput. Phys. 230 (8) (2011) 2794–2805, <https://doi.org/10.1016/j.jcp.2010.11.031>, <https://www.sciencedirect.com/science/article/pii/S0021999110006467>.
- [30] J. Hunt, A. Wray, P. Moin, Eddies, streams, and convergence zones in turbulent flows, in: Proceedings of the Summer Program, Center for Turbulence Research, Stanford Univ., 1998, pp. 193–208.