Towards Cross-Domain Domain-Specific Compiler Architecture

Paul Kelly
Group Leader, Software Performance Optimisation
Department of Computing
Imperial College London

Joint work with David Ham (Imperial Maths), Gerard Gorman (Imperial ESE), Lawrence Mitchell (now with NVIDIA), Sophia Vorderwuelbecke, George Bisbas, Edward Stow (Imperial), Fabio Luporini (Devito Codes Ltd), Florian Rathgeber (now with Google), Doru Bercea (now with IBM Research), Michael Lange (now with ECMWF), Andrew McRae (now at University of Oxford), Graham Markall (now at NVIDIA), Tianjiao Sun (now at Cerebras), Thomas Gibson (NCSA Illinois), Kaushik Kulkarni (UIUC), Andreas Klockner (UIUC), Tobias Grosser, Michel Steuwer (University of Edinburgh), Larisa Stolzfus, Amrey Krause, Nick Brown (EPCC), Navjot Kukreja (University of Liverpool) And many others....

Who am I and what do I do?

- I've worked on a lot of things....
 - GPUs, FPGAs, vector/matrix ISAs, cache coherency, large-scale SIMD, precision optimisation...
- I have worked on general-purpose compilers
 - Notably pointer analysis
 - adopted into GCC, go compiler
 - (actually the work of my PhD student David Pearce)
- But the benefits were incremental
- Meanwhile I engaged with applications specialists
 - Who know they have major performance optimisation opportunities
 - So I got interested in automating domain-specific optimisations

Imperial College Power tools for performance programming

By capturing domain-specific representation....

- We can deliver domain-specific optimisations
- We collect and automate all the performance techniques that are known for a family of problems
- If we get it right.... we get
 - **Productivity** by generating low-level code from a high-level specification
 - Performance by automating optimisations
 - **Performance portability** with multiple back-ends





The work of my research group

Vectorisation, parametric polyhedral tiling

Tiling for unstructuredmesh stencils

Lazy, data-driven compute-communicate

Runtime code generation

Multicore graph worklists

Tensor contractions

Generalised loop-invariant code motion

Sparsity in Fourier transforms

Functional Variational Inference

Search-based optimisation

Processor/accelerator microarchitecture, codesign

MLIR

Technologies

Finite-volume CFD

Finite-element

Finite-difference

Real-time 3D scene understanding

Adaptive-mesh CFD

Contour trees, Reeb graphs

Unsteady CFD - higher-order flux-reconstruction

Ab-initio computational chemistry (ONETEP)

Gaussian belief propagation

Uncertainty in DNNs

Near-camera processing

Quantum computing

Contexts

Automating domain-specific performance optimisations

Exploiting higher-level language to get better performance than low level code

PyOP2/OP2

Unstructured-mesh stencils

Firedrake

Finite-element

Devito: finite difference

SLAMBench:

3D vision, dense SLAM

SuperEight: octree SLAM

PRAgMaTic: Unstructured mesh adaptation

GiMMiK: small matrix multiply

TINTL: Fourier interpolation

Hypermapper: design optimisation

RobotWeb: distributed localisation

Projects

Aeroengine turbomachinery

Weather and climate

Glaciers

Domestic robotics, augmented reality

Tidal turbine placement

Formula-1, UAVs, buildings

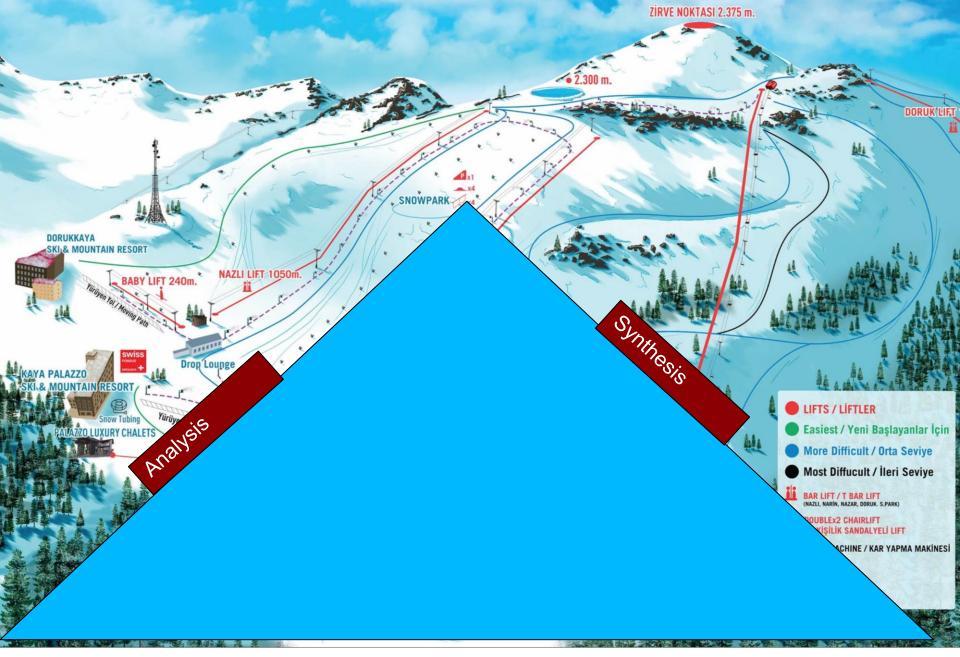
Solar energy, drug design

Medical imaging

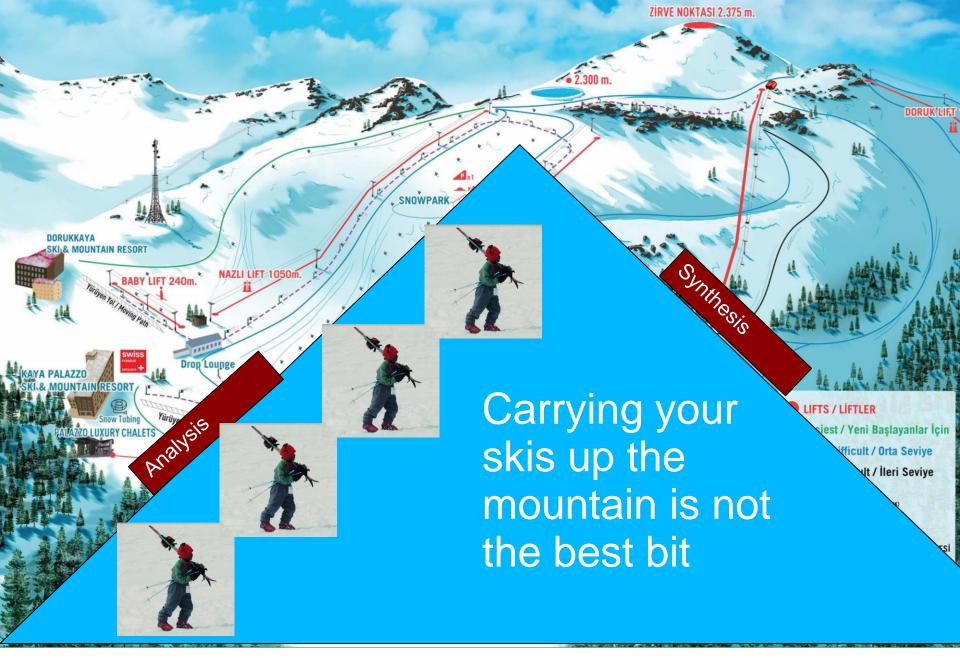
Applications



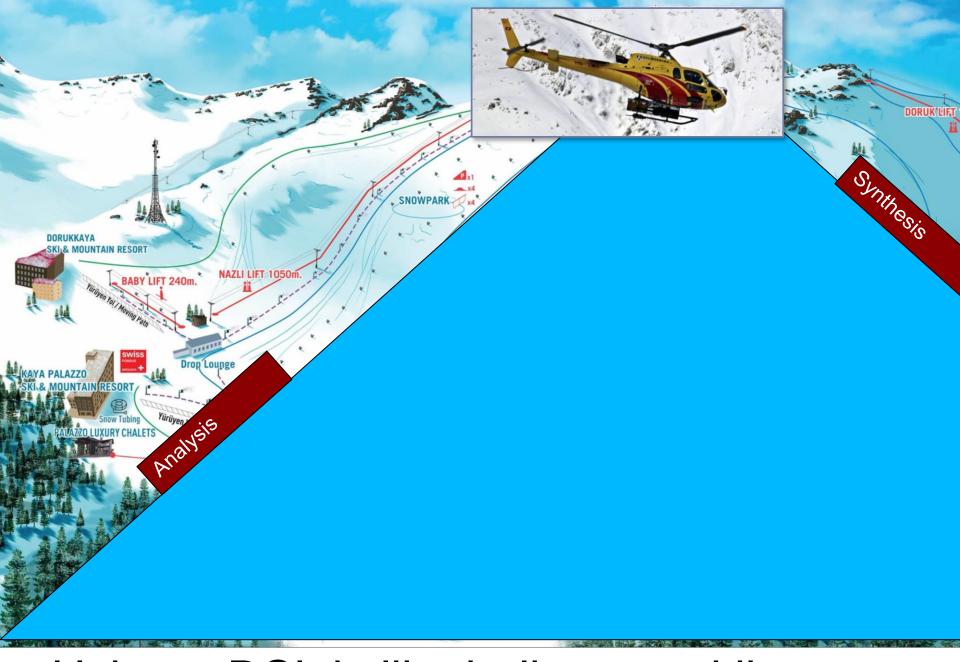
Compilation is like skiing



Compilation is like skiing



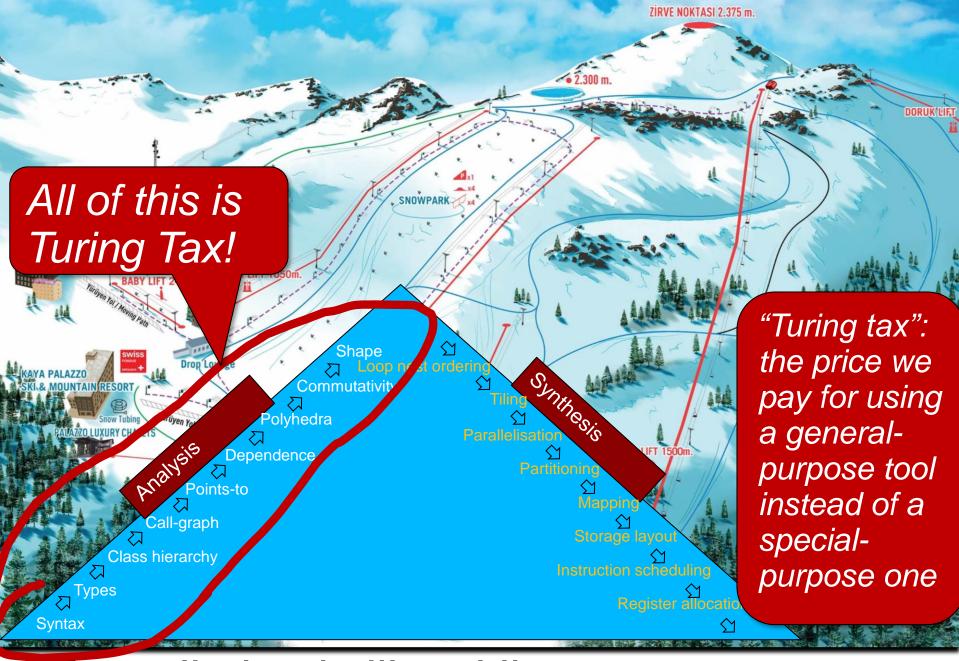
Compilation is like skiing



Using a DSL is like helicopter skiing



- Compiling is like skiing
 - Analysis is an uphill struggle
- "Turing Tax"
 - The price you pay for running on a general-purpose computer rather than a specialised one
- What do we call...
 - The price you pay for using a general-purpose programming language rather than a DSL?
- This talk:
 - DSLs really can deliver
 - DSL compiler architecture: how do DSLs win?
 - Why we have to make the DSL ecosystem work!



Compilation is like skiing



Documentation

Download

Team

Citing Publications

Events

Funding

Contact

GitHub Jenkins

Firedrake is an automated system for the solution of partial differential equations using the finite element method (FEM). Firedrake uses sophisticated code generation to provide mathematicians, scientists, and engineers with a very high productivity way to create sophisticated high performance simulations.

Features:

- Expressive specification of any PDE using the Unified Form Language from the FEniCS Project.
- Sophisticated, programmable solvers through seamless coupling with PETSc.
- Triangular, quadrilateral, and tetrahedral unstructured meshes.
- Layered meshes of triangular wedges or hexahedra.
- Vast range of finite element spaces.
- Sophisticated automatic optimisation, including sum factorisation for high order elements, and vectorisation.
- · Geometric multigrid.
- · Customisable operator preconditioners.
- Support for static condensation, hybridisation, and HDG methods.

Latest commits to the Firedrake master branch on Github

Merge pull request #1520 from firedrakeproject/wence/feature/assemble-diagonal

Lawrence Mitchell authored at 22/10/2019, 09:14:34

tests: Check that getting diagonal of matrix works

Lawrence Mitchell authored at 21/10/2019, 13:04:04

matfree: Add getDiagonal method to implicit matrices

Lawrence Mitchell authored at 18/10/2019, 10:19:48

assemble: Add option to assemble diagonal of 2-form into Dat

Lawrence Mitchell authored at 18/10/2019, 10:08:37

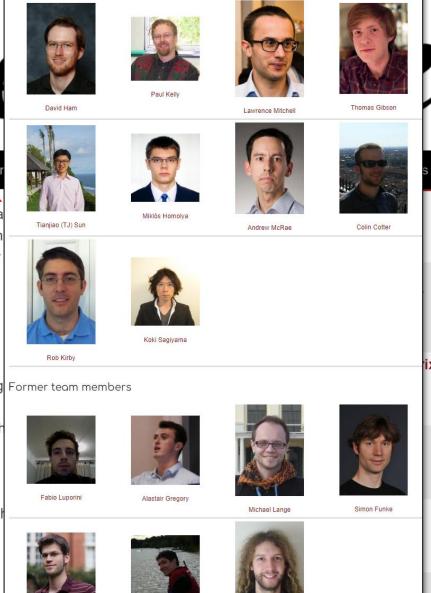
Merge pull request #1509 from firedrakeproject/wence/patch-c-wrapper



Firedrake is an automated system for the solution of partial differential equa finite element method (FEM). Firedrake uses sophisticated code generation mathematicians, scientists, and engineers with a very high productivity way sophisticated high performance simulations.

Features:

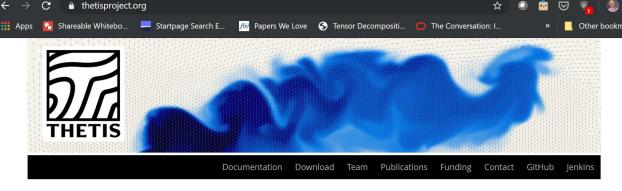
- Expressive specification of any PDE using the Unified Form Languag Former team members
 Project.
- · Sophisticated, programmable solvers through seamless coupling with
- Triangular, quadrilateral, and tetrahedral unstructured meshes.
- Layered meshes of triangular wedges or hexahedra.
- · Vast range of finite element spaces.
- Sophisticated automatic optimisation, including sum factorisation for lelements, and vectorisation.
- Geometric multigrid.
- · Customisable operator preconditioners.
- Support for static condensation, hybridisation, and HDG methods.



Doru Bercea

Graham Markal

- Firedrake is used in:
 - Thetis:
 unstructured
 grid coastal
 modelling
 framework

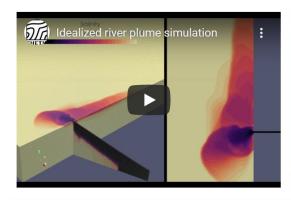


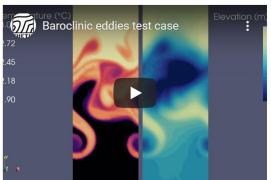
The Thetis project

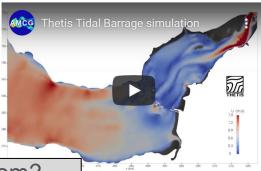
Thetis is an unstructured grid coastal ocean model built using the Firedrake finite element framework. Currently Thetis consists of 2D depth averaged and full 3D baroclinic models.

Some example animations are shown below. More animations can be found in the Youtube channel.

Current development status Latest status: build passing Thetis source code is hosted on Github and is being continually tested using Jenkins.



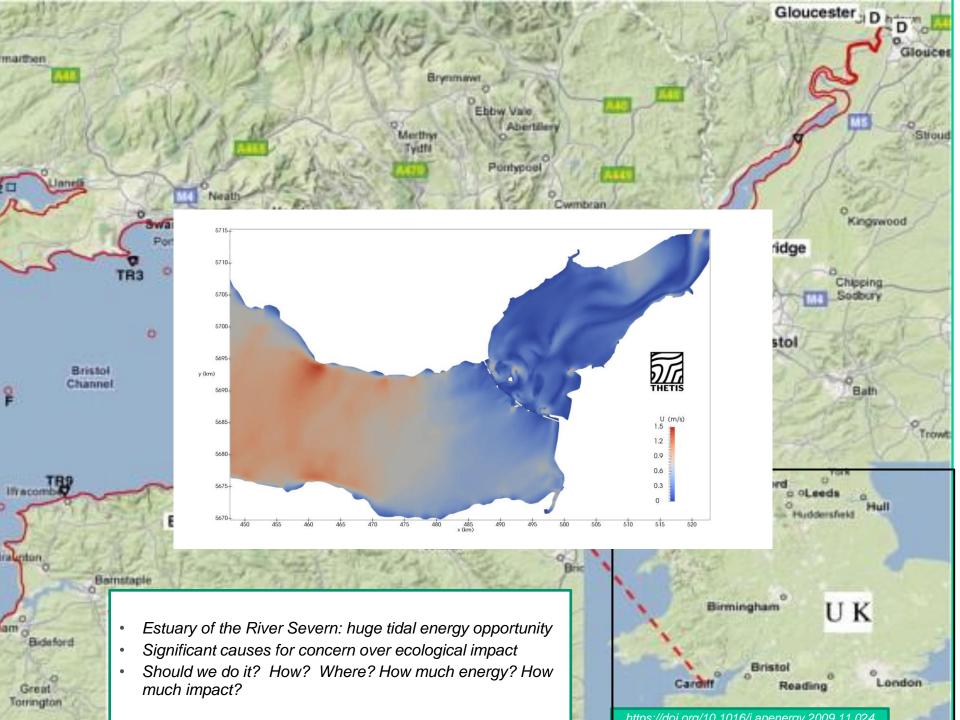




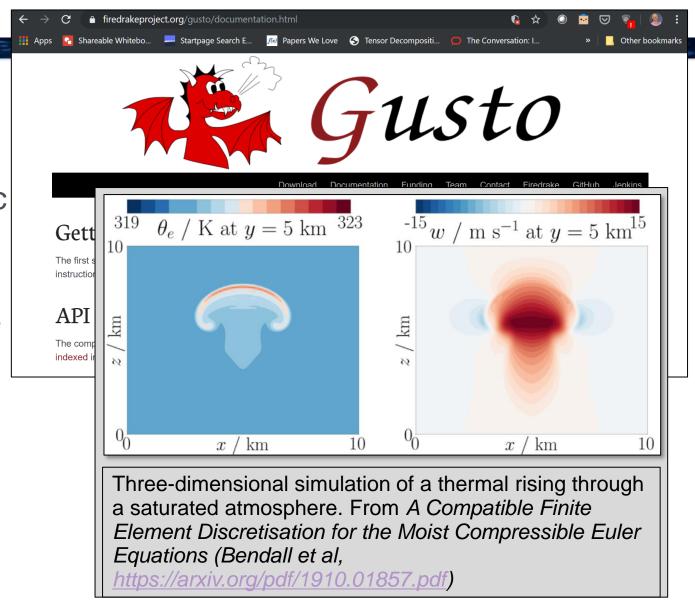


What is it used for? By whom?



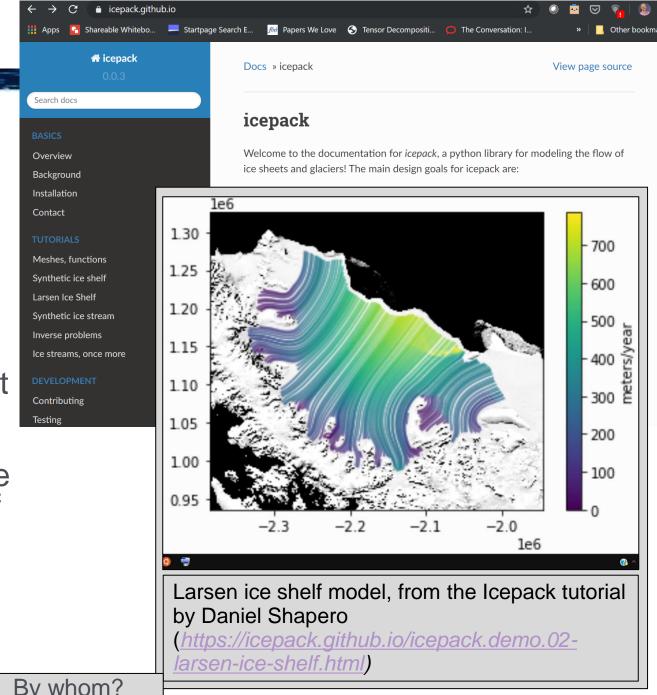


- Firedrake is used in:
 - Gusto: atmospheric modelling framework being used to prototype the next generation of weather and climate simulations for the UK Met Office



What is it used for? By whom?

- Firedrake is used in:
 - Icepack: a framework for modeling the flow of glaciers and ice sheets, developed at the Polar Science Center at the University of Washington



What is it used for? By whom?

Example: Burgers equation

$$\int_{\Omega} rac{u^{n+1}-u^n}{dt} \cdot v + ((u^{n+1}\cdot
abla)u^{n+1}) \cdot v +
u
abla u^{n+1} \cdot
abla v \, \mathrm{d}x = 0.$$

- From the weak form of the PDE, we derive an equation to solve, that determines the state at each timestep in terms of the previous timestep
- Transcribe into Python u is u^{n+1} , u_ is u^n :

Set up the equation and solve for the next timestep u: solve(F == 0, u)

At this point, Firedrake generates code to assemble a linear system, runs it and calls a linear solver (we use PetSC)

```
from firedrake import *
                                        mesh = UnitSquareMesh(n, n)
n = 50
mesh = UnitSquareMesh(n, n)
# We choose degree 2 continuous Lagrange polynomials
# piecewise linear space for output purposes::
                                          = VectorFunctionSpace(mesh, "CG", 2)
V = VectorFunctionSpace(mesh, "CG", 2)
                                        V out = VectorFunctionSpace(mesh, "CG", 1)
V out = VectorFunctionSpace(mesh, "CG", 1)
# We also need solution functions for the current and
                                            = Function(V, name="Velocity")
u = Function(V, name="Velocity")
u = Function(V, name="VelocityNext")
                                          = Function(V, name="VelocityNext")
V = TestFunction(V)
# We supply an initial condition::
                                         # set up initial conditions for u and u_
x = SpatialCoordinate(mesh)
ic = project(as_vector([sin(pi*x[0]), 0]), V)
# Start with current value of u set to
# initial condition as our starting gue # Define the residual of the equation::
u_.assign(ic)
                             F = (inner((u - u)/timestep, v)
u.assign(ic)
                                   + inner(dot(u,nabla_grad(u)), v) + nu*inner(grad(u), grad(v)))*dx
:math:`\nu` is set to a (fairly arbit
nu = 0.000
           = 0.0
timestep =
# Define t
F = (inner while (t <= end):
   + inn
                solve(F == 0, u)
outfile =
                u_.assign(u)
outfile.wr
# Finally,
                t += timestep
t = 0.0
                outfile.write(project(u, V_out, name="Velocity")
end = 0.5
   solve(\bar{F} == 0, u)
   u .assign(u)
                                                           What does its DSL actually look like?
   t += timestep
   outfile.write(project(u, V out, name="Velocity"))
```

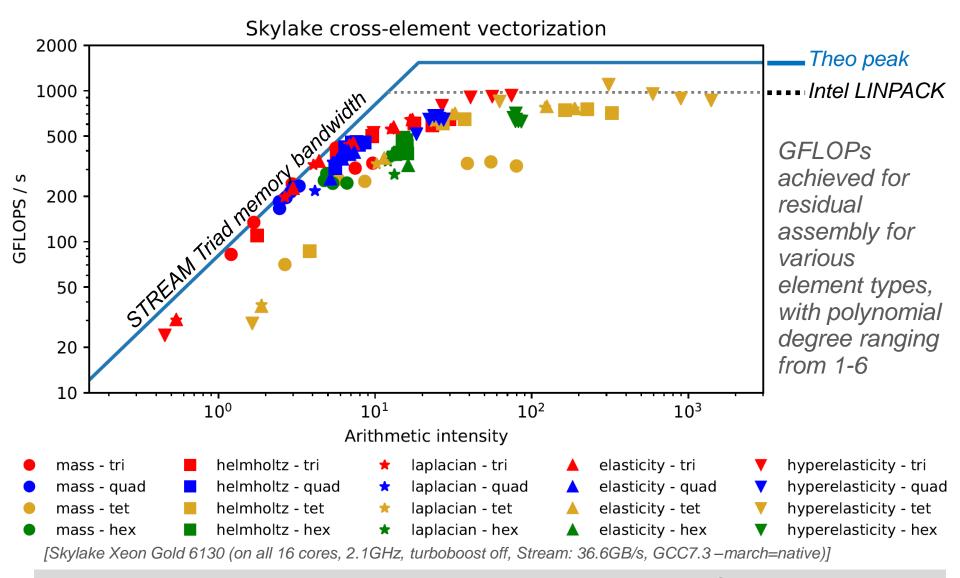
```
#include <math.h>
void wrap form00 cell integral otherwise(int const start, int const end, Mat const mat0, double const * restrict dat1, double const * restrict dat0, int const * restrict map0, int const * restrict map0, int const * restrict map1)
  double const form_t17[7] = { ... };
  double const form_t18[7 * 6] = { ... };
double const form_t19[7 * 6] = { ... };
  double form t2;
  double const form_t20[7 * 6] = { ... };
                                                                                                                                                                                  Generated code
  double form t21...t37;
  double form_t38[6];
  double form_t39[6];
  double form_t4;
  double form t40...t45;
                                                                                                                                                                                                to assemble the
  double form t5...t9;
  double t0[6 * 2];
  double t1[3 * 2];
  double t2[6 * 2 * 6 * 2];
                                                                                                                                                                                                resulting linear
  for (int n = start; n <= -1 + end; ++n)
     for (int i4 = 0; i4 \le 5; ++i4)
        for (int i5 = 0; i5 <= 1; ++i5)
          for (int i6 = 0; i6 <= 5; ++i6)

for (int i7 = 0; i7 <= 1; ++i7)

t2[24 * i4 + 12 * i5 + 2 * i6 + i7] = 0.0;
                                                                                                                                                                                                system matrix
     for (int i2 = 0; i2 \le 2; ++i2)
       for (int i3 = 0; i3 <= 1; ++i3)
t1[2 * i2 + i3] = dat1[2 * map1[3 * n + i2] + i3];
     for (int i0 = 0; i0 <= 5; ++i0)
for (int i1 = 0; i1 <= 1; ++i1)
                                                                                                                                                                                  Executed at each
     t0[2 * i0 + i1] = dat0[2 * map0[6 * n + i0] + i1];
form t0 = -1.0 * t1[1];
     form_t1 = form_t0 + t1[3];
form_t2 = -1.0 * t1[0];
                                                                                                                                                                                               triangle in the
     form t3 = form \ t2 + t1[2];
     form t4 = form_t\theta + t1[5]
     form t5 = form t2 + t1[4];
     form t6 = form t3 * form_t4 + -1.0 * form_t5 * form_t1;
     form_t7 = 1.0 7 form_t6;
form_t8 = form_t7 * -1.0 * form_t1;
form_t9 = form_t4 * form_t7;
                                                                                                                                                                                                mesh
     form t10 = form t3 * form t7;
     form t11 = form t7 * -1.0 * form t5:
     form t12 = 0.0001 * (form t8 * form t9 + form t10 * form t11);
                                                                                                                                                                                            Accesses
     form t13 = 0.0001 * (form t8 * form t8 + form t10 * form t10);
     form t14 = 0.0001 * (form t9 * form t9 + form t11 * form t11);
     form_t15 = 0.0001 * (form_t9 * form_t8 + form_t11 * form_t10);
     form t16 = fabs(form t6);
     for (int form ip = 0; form ip <= 6; ++form ip)
                                                                                                                                                                                                degrees of
        form t26 = 0.0; form t25 = 0.0; form t24 = 0.0; form t23 = 0.0; form t22 = 0.0; form t21 = 0.0;
        for (int form_i = 0; form_i <= 5; ++form_i)</pre>
           form_t21 = form_t21 + form_t20[6 * form_ip + form_i] * t0[1 + 2 * form_i];
                                                                                                                                                                                                freedom shared
          form t22 = form t22 + form t19[6 * form ip + form i] * t0[1 + 2 * form i];
           form t23 = form t23 + form t20[6 * form ip + form i] * t0[2 * form i];
           form t24 = form t24 + form t19[6 * form ip + form i] * t0[2 * form i];
           form t25 = form t25 + form t18[6 * form ip + form i] * t0[1 + 2 * form i];
           form t26 = form t26 + form t18[6 * form ip + form i] * t0[2 * form i];
                                                                                                                                                                                               with neighbour
        form t27 = form t17[form ip] * form t16;
        form t28 = form t27 * form t15;
        form t29 = form t27 * form t14;
        form_t30 = form_t27 * (form_t26 * form_t9 + form_t25 * form_t11);
                                                                                                                                                                                               triangles through
        form_t31 = form_t27 * form_t13;
        form_t32 = form_t27 * form_t12;
        form t33 = form t27 * (form t26 * form t8 + form t25 * form t10);
form t34 = form t27 * (form t11 * form t24 + form t10 * form t23);
        form_t35 = form_t27 * (form_t9 * form_t22 + form_t8 * form_t21);
                                                                                                                                                                                                indirection map
        form_t36 = form_t27 * (50.0 + form_t9 * form_t24 + form_t8 * form_t23)
        form_t37 = form_t27 * (50.0 + form_t11 * form_t22 + form_t10 * form_t21);
        for (int form_k0 = 0; form_k0 \le 5; ++form_k0)
           form t38[form k0] = form t18[6 * form ip + form k0] * form t37;
           form_t39[form_k0] = form_t18[6 * form_ip + form_k0] * form_t36;
        for (int form_j0 = 0; form_j0 <= 5; ++form_j0)
           form_t40 = form_t18[6 * form_ip + form_j0] * form_t35;
          form_t41 = form_t1816 * form_ip + form_j0] * form_t34; form_t41 = form_t1816 * form_ip + form_j0] * form_t34; form_t42 = form_t2016 * form_ip + form_j0] * form_t31 + form_t0216 * form_ip + form_j0] * form_t32; form_t42 = form_t2016 * form_ip + form_j0] * form_t28; form_t028 = form_t2016 * form_t029; form_t028; form_t0
           for (int form_k0_0 = 0; form_k0_0 <= 5; ++form_k0_0)
             MatSetValuesBlockedLocal(matθ, 6, &(mapθ[6 * n]), 6, &(mapθ[6 * n]), &(t2[θ]), ADD_VALUES);
```

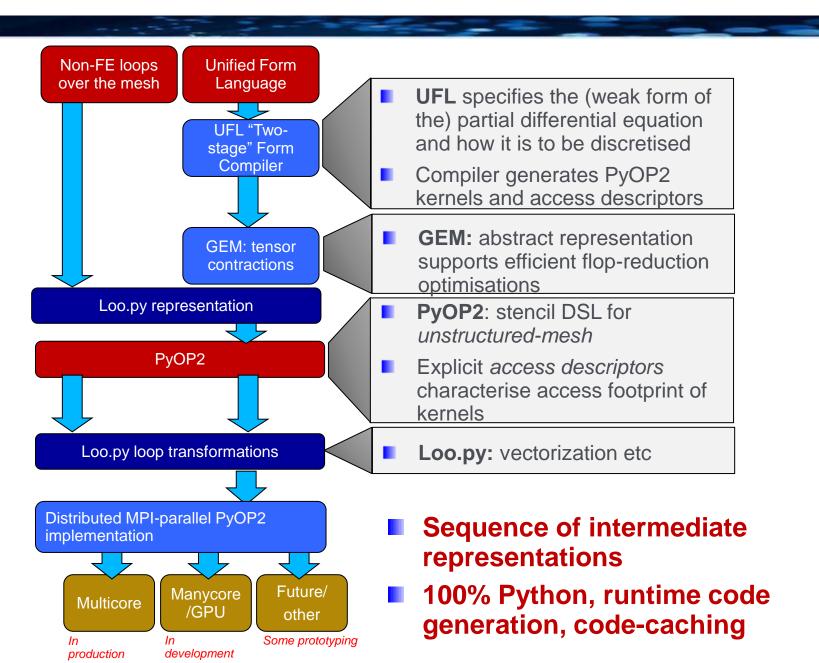
Firedrake: single-node AVX512 performance

Does it generate good code?



A study of vectorization for matrix-free finite element methods, Tianjiao Sun et al IJHPCA 2020 https://arxiv.org/abs/1903.08243

Firedrake: compiler architecture





Devito: Symbolic Finite Difference Computation

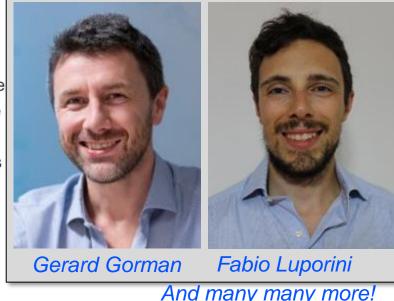
Devito is a domain-specific Language (DSL) and code generation framework for the design of highly optimised finite difference kernels for use in inversion methods. Devito utilises SymPy to allow the definition of operators from high-level symbolic equations and generates optimised and automatically tuned code specific to a given target

architecture.

Symbolic computation is a powerful tool that allows users to:

- Build complex solvers from only a few lines of high-level code
- Use automated performance optimisation for generated code
- Adjust stencil discretisation at runtime as required
- (Re-)development of solver code in hours rather than months

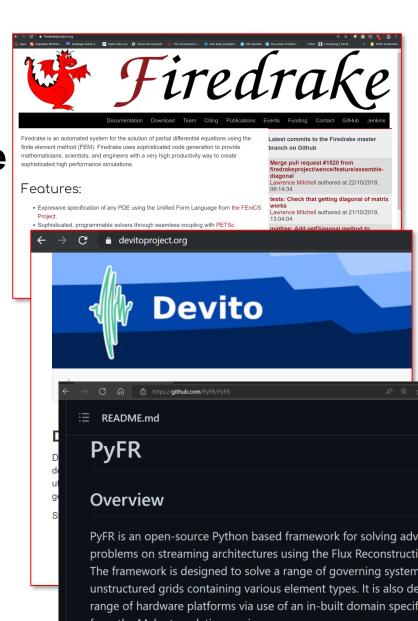




- Engaging with applications to exploit domain-specific optimisations can be incredibly fruitful
 - Compiling general purpose languages is worthy but usually incremental
- Compiler architecture is all about designing intermediate representations that make hard things look easy
 - Tools to deliver domain-specific optimisations often have domain-specific representations
 - Premature lowering is the constant enemy (appropriate lowering is great)
- Along the way, we learn something about building better general-purpose compilers and programming abstractions
 - Drill vertically, expand horizontally

How can we change the world?

- The real value of Firedrake and Devito is in supporting the applications users in exploring *their* design space
- We enable them to navigate rapidly through alternative solutions to their problem
- In the future, we will have automated pathways from maths to code for many classes of problem, and many alternative solution techniques



The actual message of this talk....

Domain-specific code generation tools really work

But we can't afford to continue like this!



DSLs are a dysfunctional ecosystem

- Users have really good reasons not to adopt DSLs:
 - Longevity: will the tool support outlive my project?
 - Walled garden: if later I discover I need something the DSL can't do, I'm in trouble
 - Interoperability: how do I plug DSL stuff together with other parts of my code?
- Devito shares no code with Firedrake above the C compiler
- For example: dozens of stencil DSLs
 - Almost all are dead academic projects
 - None share any code

Stencil DSLs:

- Halide
- Gridtools/Stella
- Open Earth Compiler
- Psyclone
- YASK
- OPS
- Artemis
- StencilGen
- Lift
- SDSLc
- ExaStencils
- Patus
- Physis
- Mint
- Pochoir
- ShiftCalculus
- HPF
- MSF
- ...

We have to work together...

- What do you need to do to get leading-edge performance with stencils?
 - Spatial tiling (hierarchical)
 - Temporal tiling (wavefront, diamond....)
 - Data layout transformations ("dimension lifting")
 - Arrow scheduling (use associativity to control register pressure)
 - Cross-iteration redundancy elimination (CIRE)
 - Shuffling ("Software Systolic Array")
 - Matrix ISA optimisations, tensor cores
 - Loop fusion/fission/distribution
 - Compute-communicate overlap
 - Wavefield compression
 - Precision management
 - And many many more!!!

We have to work together...

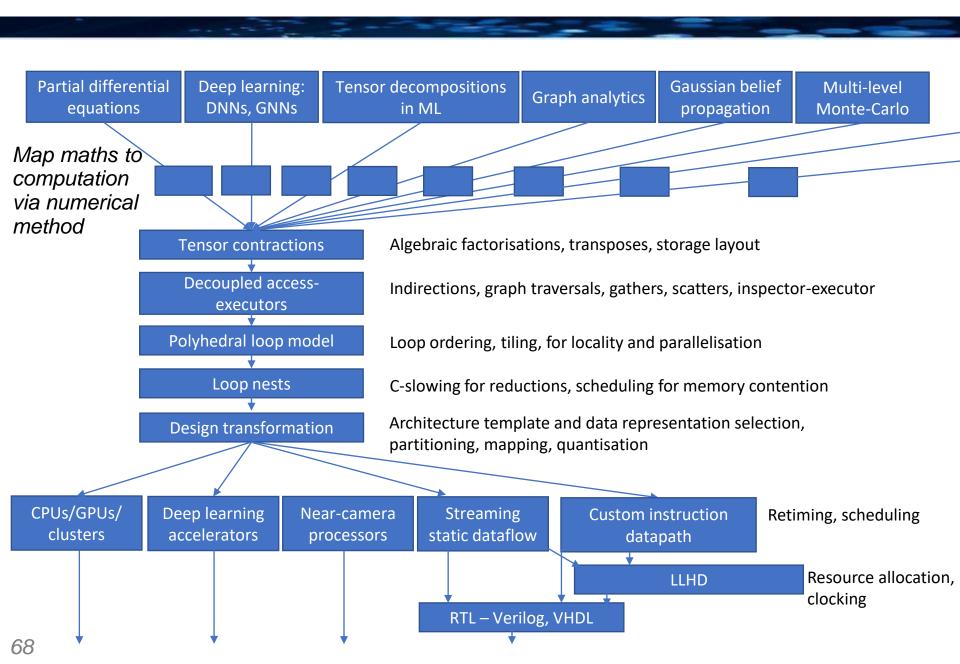
- What do you need to do to get leading-edge performance with stencils?
 - Spatial tiling (hierarchical)
 - Temporal tiling (wavefront, diamond....)
 - Data layout transformations ("dimension lifting")
 - Arrow scheduling (use associativity to control register pressure)
 - Cross-iteration redundancy elimination (CIRF)
 - Shuffling ("Software Systolic Array")
 - Matrix ISA optimisations, tensor core
 - Loop fusion/fission/distribution
 - Compute-communicate overlap
 - Wavefield compression
 - Precision management
 - And many many more!!!

No one team can expect to integrate all the known tehniques into a single silo of code

Imperial College have to build code that works together...

- What does it mean to compose DSLs?
- Target-specific DSLs
 - **Eg**: CUDA, HLS, Graphcore's Poplar, Maxeler's MaxJ
 - Composition: eg MPI+CUDA, MPI+X, multi-target OpenMP
- Data-structure-specific DSLs
 - **Eg** OP2/PyOP2, OPS, Psyclone, SAMR DSLs like Chombo, BVH DSLs like Optix, Octtree DSLs like SuperEight, Tensorflow
 - Composition: Particle-in-cell, Particle mesh Ewald, Multiphysics frameworks like Uintah, visualisation tools like VTK
- DSLs that automate numerical methods
 - **Eg:** FeNICS, Firedrake, Devito, PyFR, OpenSBLI, SPIRAL, GPFlow
 - Composition:
 - Coupled problems: fluid-structure, PME, Model Coupling Toolkit
 - Outer-loop: PDE-constrained optimisation (FWI), assimilation (4DVAR), uncertainty quantification (MLMC), parameterisation with deep learning
 - At this level we have access to the maths!

Vision for the future



Acknowledgements

Thank you to our many many collaborators!

Partly funded/supported by

- SysGenX: Composable software generation for system-level simulation at Exascale (EP/W026066/1)
- XDSL: Efficient Cross-Domain DSL Development for Exascale (EP/W007789/1)
- NERC Doctoral Training Grant (NE/G523512/1)
- EPSRC "MAPDES" project (EP/I00677X/1)
- EPSRC "PSL" project (EP/I006761/1)
- Rolls Royce and the TSB through the SILOET programme
- EPSRC "PAMELA" Programme Grant (EP/K008730/1)
- EPSRC "PRISM" Platform Grants (EP/I006761/1 and EP/R029423/1)
- EPSRC "Custom Computing" Platform Grant (EP/I012036/1)

https://github.com/xdslproject/xdsl

- EPSRC "Application Customisation" Platform Grant (EP/P010040/1)
- EPSRC "A new simulation and optimisation platform for marine technology" (EP/M011054/1)
- Basque Centre for Applied Mathematics (BCAM)
- Schloss Dagstuhl
- Code:
 - http://www.firedrakeproject.org/, https://www.devitoproject.org/
 - http://op2.github.io/PyOP2/, https://github.com/OP-DSL/OP2-Common
- 69