Tales from the trenches: preparing scientific computing software for emerging compute and data intensive challenges.

Dr. Cathal Ó Broin MInstP

LERO: The Irish Software Research Centre

The Irish Centre for High-End Computing (ICHEC)

Introduction

1. **PhD:** Single atoms and molecules in intense laser fields
2. **Current:** Extreme Scale Seismic Data
3. Software Engineering Practices

- Sweeping generalisations from my own experience.
- Bonus, Fortran is mentioned on one slide!

Part 1: Single atoms and molecules in intense laser fields

Physics Picture

Laser Interacts with an atom or a molecule

Image: xfel.eu (femtosecond laser)

R-Matrix Incorporating Time for Atoms and Molecules

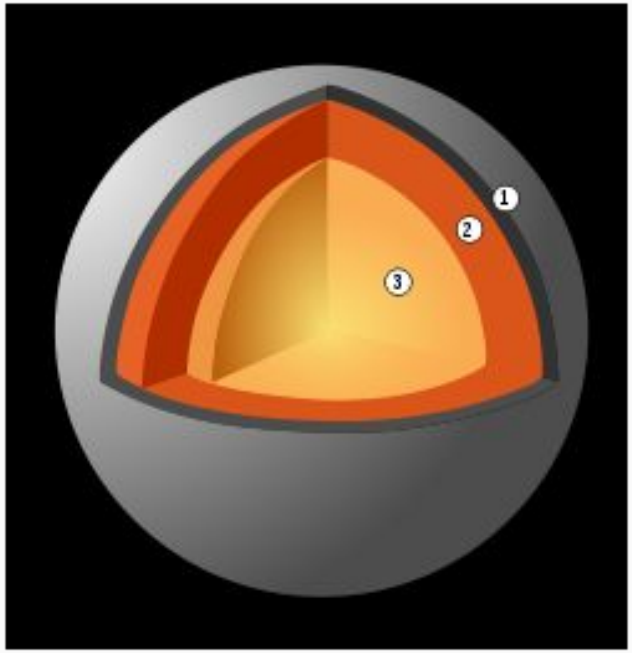


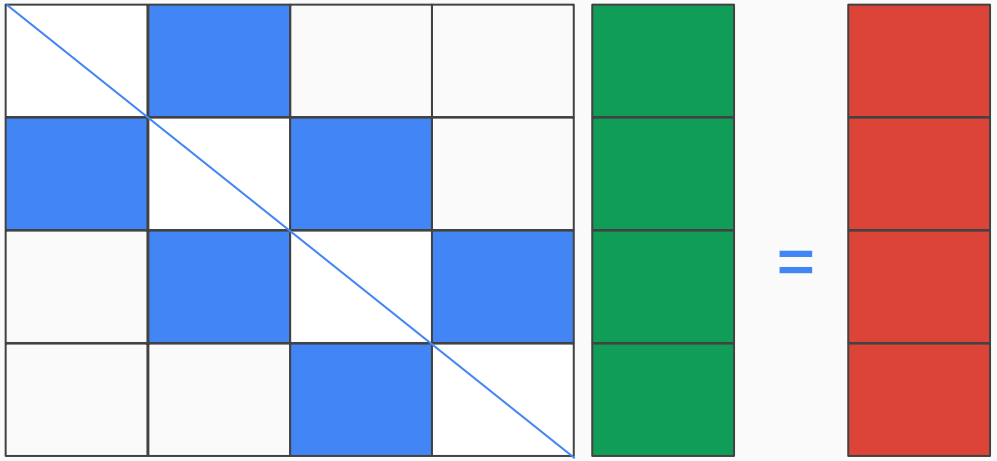
Image: https://commons.wikimedia.org/wiki/File:Mercury_Internal_Structure.svg

- Small inner region and large outer region.
- Outer switches to asymptotic form
- Each region has different representations of the space.
- Communication over a boundary between two regions

C. Ó Broin. L.A.A Nikolopoulos *Physical Review A*, *92*(6), p.063428. 2015
C. Ó Broin. L.A.A Nikolopoulos *2016 (submitted)*

Numerical bottleneck

- Solving TDSE: initial value problem.
  - Taylor series
  - Arnoldi-Lanczos
  - Runge-Kutta methods.
- Bottleneck: matrix-vector calculation sparse and (blocked) dense matrices.



Accelerating the Bottleneck!

- GPUs → compute bound, internal memory bandwidth.
- Perfect match!
- GPUs can be much cheaper per flop than CPUs.

Image: amd.com

Choice of GPU programming models

- OpenCL looked promising as an open standard.
- Portability across vendors and low cost of consumer AMD GPUs.
- AMD consumer card dwarfed the performance of my cheap Intel CPU.
- AMD consumer card’s DP performance was competitive with expensive NVIDIA Tesla cards.

Image: developer.apple.com

Software Design during a PhD

- Often very fluid requirements, hard to design at the start.
- Lots of new feature requirements appear.
- Feature additions aren’t tracked.

Image: cliffpro.com

Why not test?

PhD practices:

- No unit testing, only system tests.
- Testing against results from published papers, comparisons against other codes.
- Testing not automated, no testing for regressions.
- Overhead of rigorous testing and coding too high?

Scientific Software Practices in Atomic Physics

- Slow adoption of new technologies.
- Many codes are not well designed. Continual use exposes bugs.
- Software architecture issues!

… but things worked out for me in the end.

At the end of my PhD, I implemented the necessary features for my research work and publications, finished my thesis and secured a postdoc (not in that order).

- C. Ó Broin, L.A.A. Nikolopoulous, Comput. Phys. Commun. 183 (10), 2071-2080, 2012
- Benis *et al*, Phys. Rev. A 86 (4), 043428, 2014
- C. Ó Broin, L.A.A. Nikolopoulous, Comput. Phys. Commun. 185 (6), 1791-1807, 2014
- C. Ó Broin, L.A.A. Nikolopoulous, Phys. Rev. A 92 (6), 063428, 2015
- C. Ó Broin, L.A.A. Nikolopoulous, submitted, 2016

What now for the long term maintenance and development of the code?

Is OpenCL still a good choice?

- AMD DP instructions reduced from 1 in every 4 cycles to 1 in every 16.
- Expensive GPUs → budgets of small academic groups.
- Will OpenCL be supported on 2nd gen Intel Xeon Phi?
- NVIDIA OpenCL support limited.
- Not clear where things will go from here.

Part 2: Extreme Scale Seismic Data (ExSeisDat)

Seismic image processing

- Geophysicists analyze seismic data to look for oil and gas deposits.
- Seismic migration from lightweight to the very demanding.
- We focus on the huge data requirements!

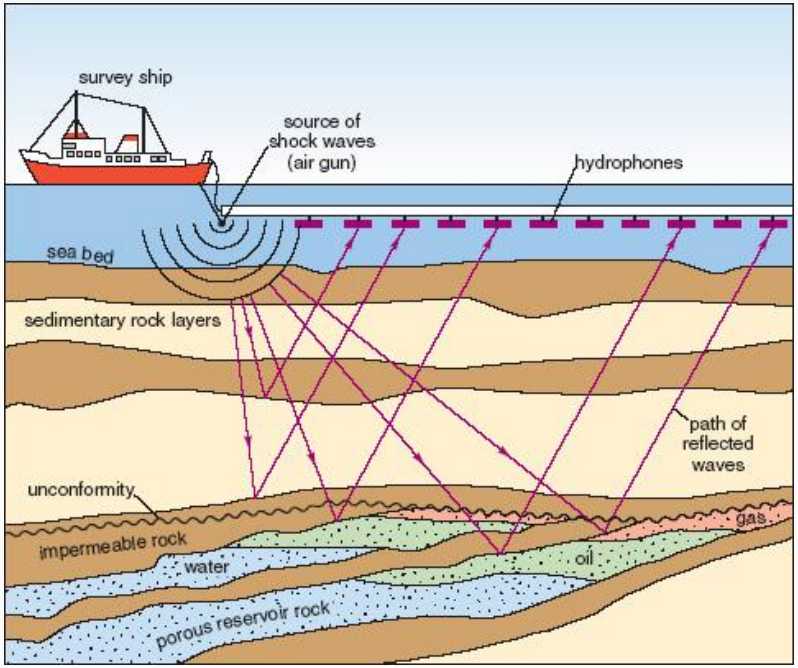


Image:epa.org

Tullow Oil plc

- Upstream Oil & Gas company originally founded in Tullow, Ireland.
- Tullow Oil internally develop applications for seismic processing.
- Much of their geophysics and geoscience takes place in Dublin and London.

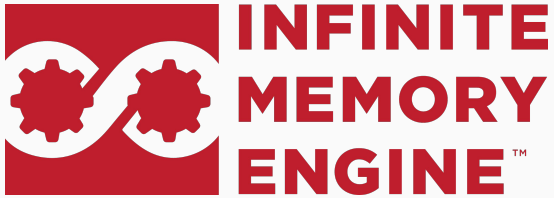
Image: FPSO, Tullow Oil Brochure for TEN Project
“Floating Production, Storage and Offloading”

Project Context

- Tullow Oil seismic datasets are large, parallel data-access needed.
- Geophysicist productivity spent on parallel data-access instead of geophysics.
  - Productivity lost
  - Redundant work
- Opportunities with next generation tiered storage

Project Goals

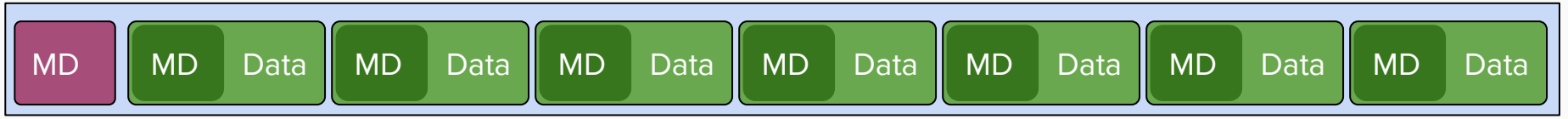
- A dedicated team develops a parallel I/O library.
- Good software engineering practices.
- Research on parallel I/O for O&G workflows.
- Next-gen tiered storage: DDN IME burst buffer.
- Workflow Integration with existing proprietary software.



The SEG-Y format

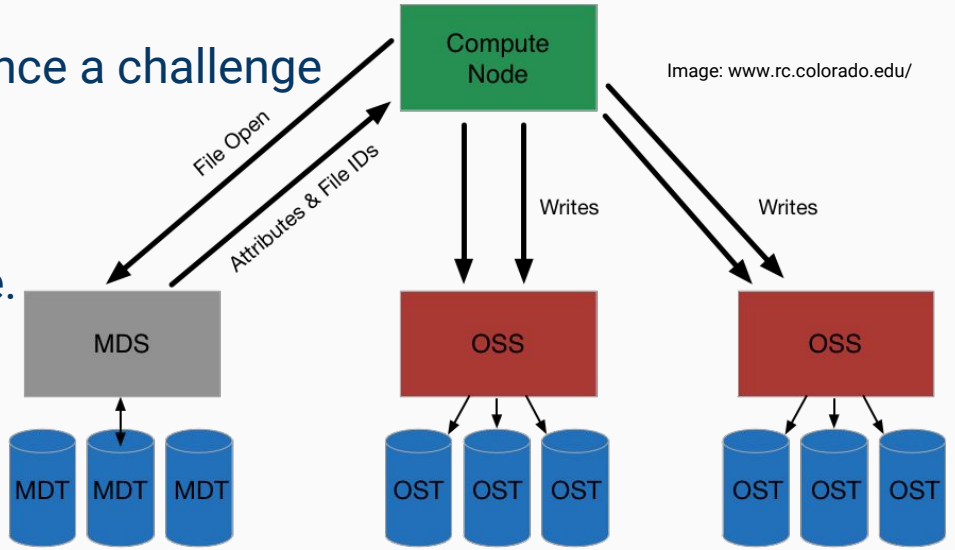
- SEG-Y file format was standardised in 1975
- Layout for slow tape access: metadata and data together.
- Very large industry inertia. Used across exploration companies.
- Loosely followed

Image: ibm.com

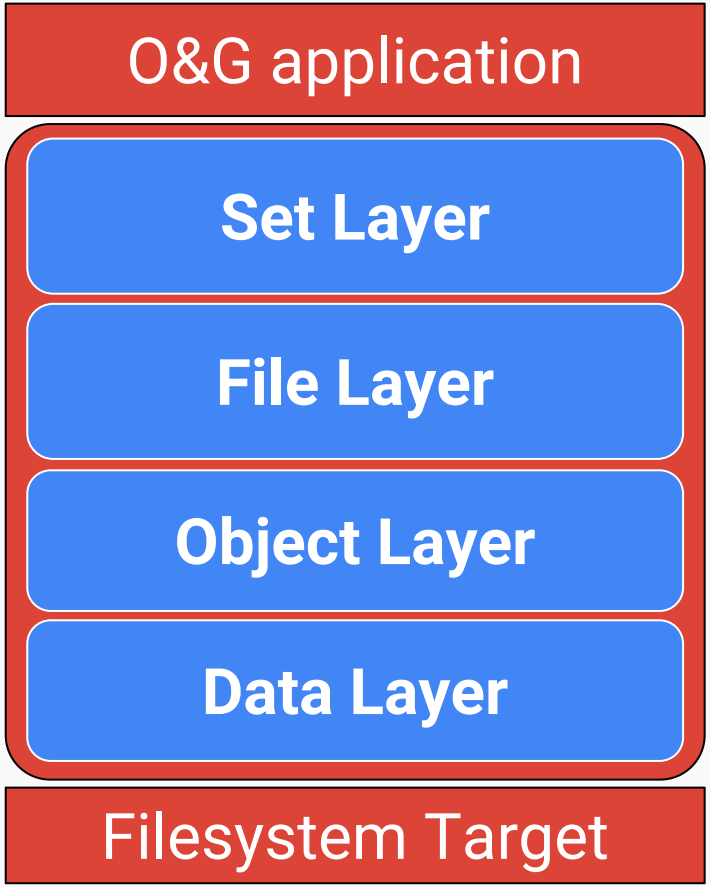


ExSeisPIOL: Parallel I/O for SEG-Y

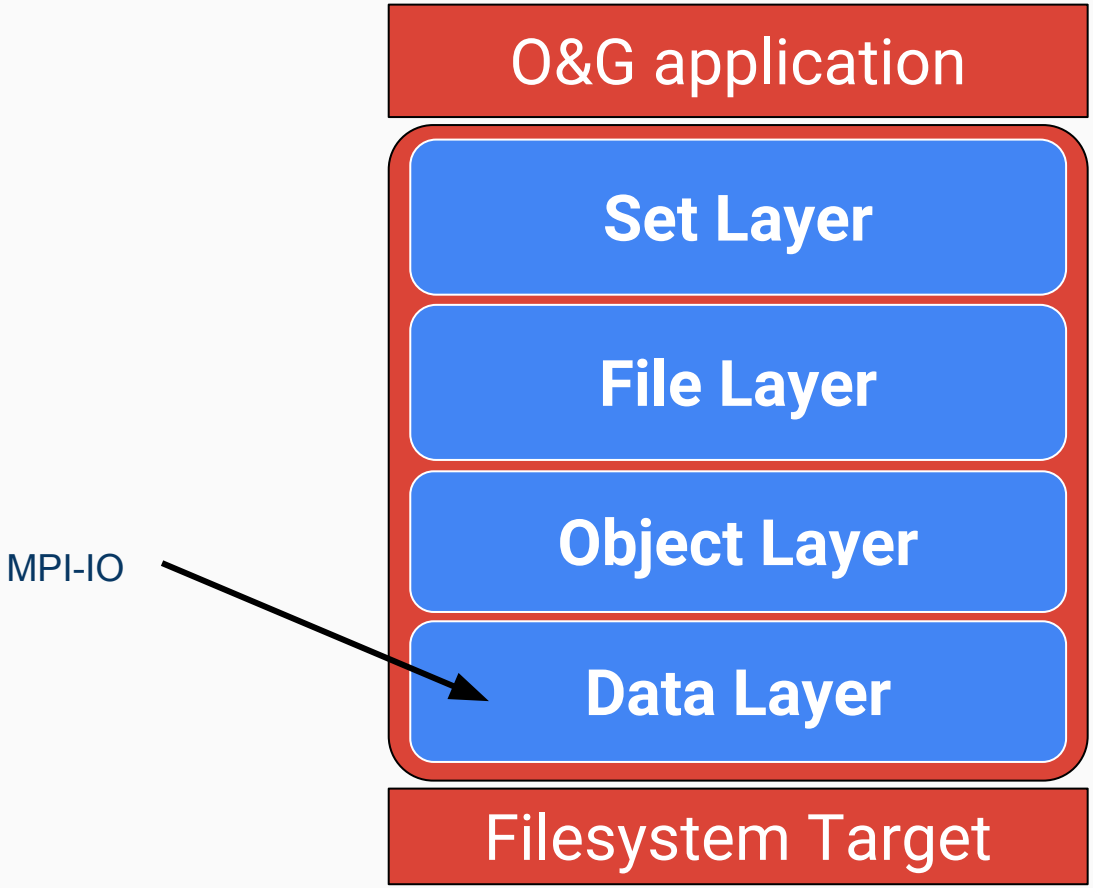
- 1st stage → parallel I/O library (PIOL).
- Maximising parallel I/O performance a challenge
- Work with Lustre and others.
- Non-POSIX for next-gen hardware.



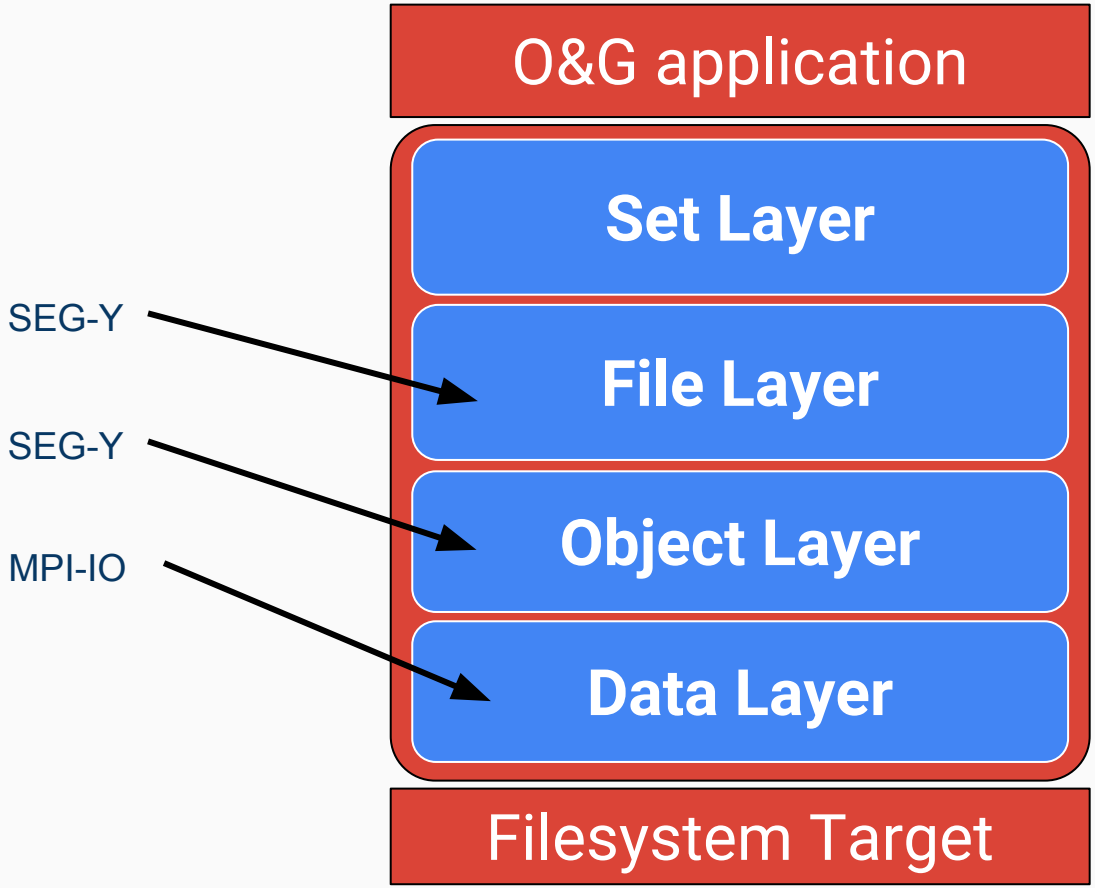
ExSeisPIOL: Parallel I/O Stack



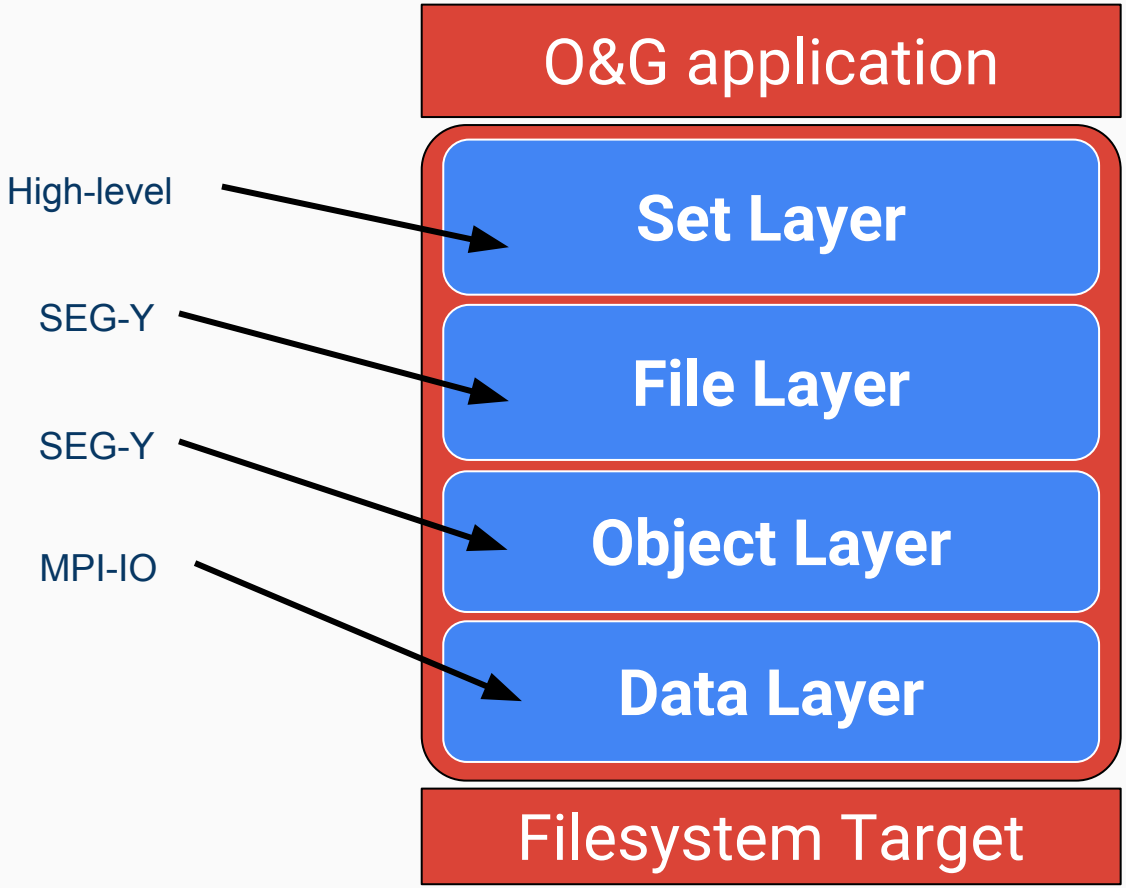
ExSeisPIOL: Parallel I/O Stack



ExSeisPIOL: Parallel I/O Stack



ExSeisPIOL: Parallel I/O Stack



Part 3. Software Engineering Practices

1. Scrum (Agile)

We use Scrum:

- Agile development approach

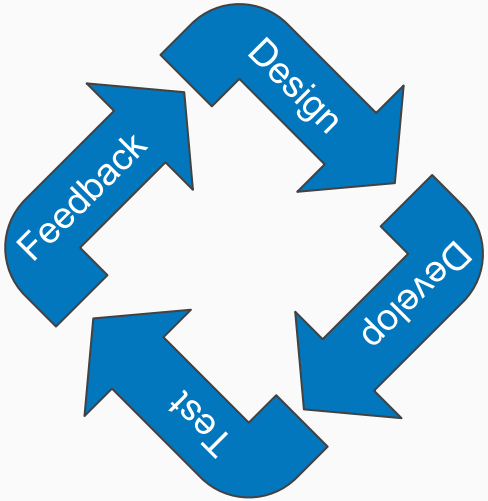
Image: http://www.therugbyblog.com/

Emphasises:

- Continuous client engagement, early delivery
- Evidence-driven development priorities; deliver value
- U*ser stories*: basic unit of development.
- People over process.

Software Design

- Project had two months of familiarisation and requirements analysis
- 1 month of planning and prototyping.
- Plans are still important!



Requirements Analysis

- Development should be evidence based.
- Geophysics workflows provides context.
- Identify shortcomings and where ***value*** is generated.
- Period of analysis to extract broad long-term requirements and detailed short term requirements.

Testing

Tests of the code come in a variety of forms.

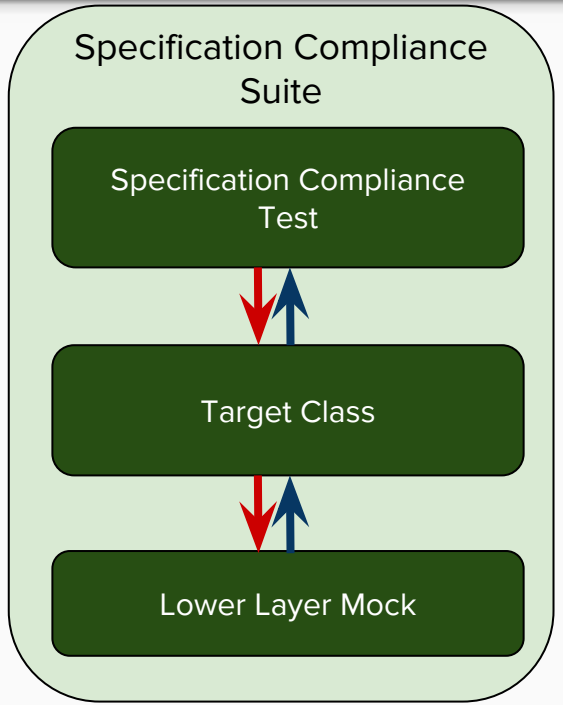
- Unit tests
- Specification compliance tests using *mocks*
- Integration tests.
- System & Performance tests.
  - All pairs testing used to generate system test parameters.

Unit Tests

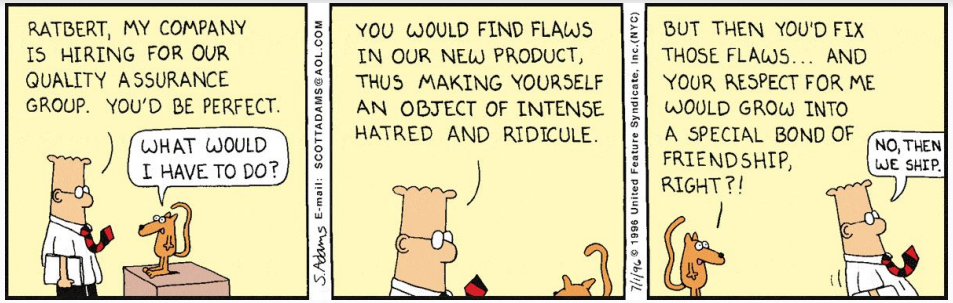
- Unit tests work on a small unit of code in isolation.
- Test extremes
- Unit tests help with early detection of code regressions.

Mock objects

- Help to isolate bugs to a layer.
- Each layer can be tested in isolation with mocks.
- Mocks → actions, expectations, no implementation.
- Ensure a layer interacts with another as expected.



All-Pairs Testing



- System tests → many parameters.
- Many parameters → many tests
- 2 node counts, 3 PPN, 3 MPI libraries, 3 input files, 2 compilers, 6 system test codes = 2*3*3*3*2*6 = 648 test case possibilities
- All-pairs testing only tests pairs of parameters → 19 test cases

Developer Teamwork

- Pair programming
- Code Reviews
- Software walkthrough
- Detail planning
- PhD students: Single-developer projects exchange code-review with related projects.

Image: https://www.flickr.com/photos/78453620@N00/552208117/

Unknown mixture: Research & Agile

- Who is the customer?
  - The journal editor?
  - Other researchers/supervisor?
  - The student?
- What user stories?
- Can PhD students keep up with a dev team, while doing their own research?

User documentation

User guide provides the end-user API with examples.

Doxygen → inline comments useful

PhD code: doxygen, peer-reviewed user guide equivalent

Language Choice

- C++, C and Fortran → similar performance.
- **C99 for PhD**: Libraries, low-level simplicity, documentation.
- **Fortran for Postdoc** - Existing code was Fortran
  - Modern Fortran (2003+) nice for compute heavy work (retro feel).
  - Ecosystem?
- **C++ for ExSeisDat**: Good language for designing libraries with complex structures, C interoperability. Vibrant ecosystem.

Conclusions

- I’ve adopted new practices as I move in my career.
- Students may be aware of good practices but requires PI lead!
- Academic groups with strong scientific software dev could work well as agile groups: much development work fits.

Acknowledgements

Many thanks to:

- Thanks to the IOP CPG committee for the *Annual PhD Thesis Prize* and the organisers for inviting me!
- Dr. Lampros Nikolopoulos (PhD supervisor in DCU, Ireland)
- Prof. Piero Decleva, (postdoctoral PI in the University of Trieste, Italy)
- Dr. Michael Lysaght, (PI for the ExSeisDat project, Scrum Master).

ExSeisDat Notices

- We want to expand our team to a second developer (C++, HPC I/O). See the advertisement at ichec.ie/about_us/employment for more details. **Apply soon**!
- We are also recruiting two PhD students, see lero.ie/phdposition/ml01
- Any information about how others use SEG-Y or other seismic formats welcome.