Building ONETEP
Scientific software design and development practices

James C. Womack

University of Bristol, UK

Manby group meeting
Nov 05, 2019
Preamble

Slides will soon be available at jcwomack.com

References — ask me, or see slides online
Outline for this talk

How we build ONETEP and how you can use it

1. Thoughts on scientific software development
   ▶ What defines scientific software?
   ▶ What is “good” scientific software?

2. ONETEP as a research tool
   ▶ The motivating problem
   ▶ Linear-scaling density functional theory
   ▶ A tour of capabilities

3. ONETEP as a software package
   ▶ Distribution, usage and governance
   ▶ Design and supporting infrastructure
   ▶ Software development practices
Scientific software
What defines scientific software?

Some common attributes of scientific software packages:

1. Motivated by needs of developers  
   - Developers and users often overlap

2. Not developers’ main “product”  
   - A means to perform research

3. Continually developed, extended  
   - Always a prototype, never finished
   - Nature of problem changes

These are some general characteristics I have observed: they are not unique to scientific software (e.g. Linux kernel), nor are they universal to all scientific software.
What defines scientific software?

Contrast this with software to control an ATM . . .

1. Designed for users, not developers

2. Software is developers’ product

3. Clear, fixed specification

Image: “Nice ATM” by Flickr user Jason Cupp, used under CC BY 2.0
Scientific software: some key principles

Key principles

Version control
- Scientific record
- Multiple developers
- Avoid regression
- Confidence in results
- Simplified debugging

Modularity
- Easier to extend
- Code reuse
- Scientific record
- Help future developers
- Support users
- Validate new functionality

Testing
- Testing
- Confidence in results
- Validate new functionality

Documentation
- Documentation
- Help future developers
- Support users
Scientific software is scientific infrastructure

A scientific software package is similar to a large shared scientific facility:

1. Designers (of experiments) are users
2. The product is research, not the instrument itself
3. New experiments add new capabilities

Image: Google Maps, Imagery (c)2018 Google, Map data (c)2018 Google
Scientific software is scientific infrastructure

Not just a metaphor...

▶ Research tools for scientific communities
▶ Significant investment of human-time
▶ Requires maintenance and long-term planning
▶ Enables applications beyond those possible for any individual researcher or group

e.g. EPSRC’s “Software as Infrastructure” strategy
Scientific software: the good, the bad and the ugly

A. Developer
*J. Gen. Sci.* 2018
Scientific software: the good, the bad and the ugly

D, E, F et al. Nature, 2019

D, E, F et al. Nature, 2019
Scientific software: the good, the bad and the ugly

▶ Be experimental, but then refine and clean up

▶ Plan development to avoid unmaintainable messes

▶ Well-written code is easier to extend and maintain
ONETEP as a research tool
Quantum chemistry

Time-independent electronic Schrödinger equation

\[ \hat{H} \Psi = E \Psi \]

- Solutions provide information about properties of materials and molecules
- Intractable many-body problem for all but the smallest atomic system

Self-consistent field approach (SCF)

\[ \hat{H} \psi_i = \varepsilon_i \psi_i \]

- Convert \( N \)-body problem into \( N \) more-tractable 1-body problems
- Electrons interact through a mean field which is self-consistently solved for
Density functional theory (DFT)

\[ E[n] = T_s[n] + E_{es}[n] + E_{xc}[n] \]

▶ A popular and widely adopted SCF method
▶ Balances accuracy and computational expense
▶ Conventional DFT: \( O(N^3) \) scaling

The \( O(N^3) \) scaling of conventional DFT rapidly becomes problematic for > 1000 atoms.

Image: Fig. 1 from D.J. Cole et al., J. Phys. Chem. Lett. 4, 4206 (2013) [CC-BY-4.0 license] [1]
Linear-scaling DFT in ONETEP

Motivation: Practical quantum chemical methods for modelling large systems

Premise: Reformulate DFT to achieve $O(N)$ scaling

- Possible due to “nearsightedness” [2] of electronic interactions
- Locality of interactions manifested in the density matrix, $\rho(r, r')$
- ONETEP expresses DFT as a direct energy minimization wrt $\rho(r, r')$

$$E = \min_{\rho} E[\rho(r, r')]$$

Many interesting theoretical and computational techniques are used to implement this in practice [3, 4].

We will focus on the result...
Linear-scaling DFT in action (2005)

64 processors
2606 atoms

Linear-scaling DFT in action (2016)

![Graph showing total calculation time vs. number of atoms for different methods: B97M-V, PBE, and PKZB. The x-axis represents the number of atoms, N, ranging from 0 to 12,000, and the y-axis represents the total calculation time in hours, ranging from 0 to 8. The graph includes data points for each method, with distinct markers for B97M-V (stars), PBE (green pluses), and PKZB (red crosses). The graph indicates a linear relationship between the number of atoms and the calculation time for each method.]

Iridis 4
512 cores
13696 atoms
A tool for researchers

ONETEP computes and analyses quantities relevant to physical scientists

- Many distinct capabilities
- All built upon LS-DFT framework
- See website for documentation

www.onetep.org
An incomplete list of capabilities...

- *Ab initio* MD [5, 6]
- Bandstructure calculations [7, 8, 9]
- Constrained DFT [10, 11]
- Core loss spectroscopy (e.g. EELS) [12]
- Distributed multipole analysis (DMA) [5, 13]
- DFT+U [14, 15, 16]
- DDEC charge partitioning [17, 18]
- Dynamical mean field theory (DMFT) [19, 20, 21]
- Electron localisation descriptors (e.g. ELF)
- Electronic transport [22]
- Embedded mean-field theory (EMFT)
- Energy decomposition analysis (EDA) [23, 24]
- Ensemble DFT [25, 26, 27]; (incl. $O(N)$ variant [28])
- Empirical dispersion corrections (e.g. D2) [29]
- Linear-response TDDFT [30, 31]
- Local and projected DOS [32]
- Natural bond orbitals (NBOs) [33]
- Nudged elastic band (NEB) TS searching
  - Implicit solvent modelling [34, 35, 36, 37]
  - Open BC electrostatics [38]
  - Orbital dependent XC (hybrids & meta-GGAs) [39, 40]
- Phonons [41]
- Projector augmented waves [42]
- QM/MM and polarisable embedding [13, 43]
- Quantified natural transition orbitals (QNTOs) [44]
- vdW density functionals [45]
ONETEP as a software package
Distribution and licensing

<table>
<thead>
<tr>
<th>License</th>
<th>Cost</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaborator - Individual, specific project</td>
<td>-</td>
<td>Individual, specific project</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Includes source</td>
</tr>
<tr>
<td>Academic</td>
<td>£500</td>
<td>Entire research group</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Includes source</td>
</tr>
<tr>
<td>Commercial</td>
<td>Purchase of Materials Studio</td>
<td>Non-academic users</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GUI interface</td>
</tr>
</tbody>
</table>

- All members of the [CCP9](#) community are entitled to collaborator agreements.

- Daussault Systèmes [BIOVIA Materials Studio](#) packages ONETEP into a suite of materials science tools with a single GUI.
Daussault Systèmes BIOVIA’s products have high market penetration in top chemical companies (pharma, oil, automotive, materials, etc.)

Data from ~2016–17
Developers

**ONETEP Developer’s Group (ODG)**

Group of core developers who manage licensing, releases, relationship with BIOVIA:

Jacek Dziedzic, Peter D. Haynes, Nicholas D. M. Hine, Arash A. Mostofi, Mike C. Payne and Chris-Kriton Skylaris

1 University of Southampton, 2 Gdańsk University of Technology, 3 Imperial College London, 4 University of Warwick, 5 University of Cambridge

41 code contributors (past and present)

37 active users of repository (developers and expert users)
Source code

Source code

154 source files

450,000 lines of code

- Written in Fortran (2003 standard), built using GNU Make
- Bundled support utilities written in other languages (e.g. Bash scripts)
- Fortran modules used to encapsulate related procedures and data types
Parallelism

- Designed to run on parallel (super)computers
- Necessary for large-scale calculations
- Combined MPI [46] and OpenMP [47] parallelism.

**MPI** Work divided between processes which communicate by *message passing*

**OpenMP** Processes subdivided into threads, with *shared memory* access
Interaction with other software

Designed for Unix-like systems: Linux, Mac OS X (Windows via WSL)

Supporting software is required to build and run ONETEP:

<table>
<thead>
<tr>
<th>Software/library</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fortran compiler (F2003, OpenMP)</td>
<td>Intel Fortran, GNU Fortran</td>
</tr>
<tr>
<td>MPI library</td>
<td>Intel MPI, OpenMPI</td>
</tr>
<tr>
<td>Linear algebra libraries</td>
<td>BLAS, LAPACK</td>
</tr>
<tr>
<td>Fast Fourier transform (FFT) library</td>
<td>ScaLAPACK, Intel MKL</td>
</tr>
<tr>
<td>Multigrid solver library</td>
<td>FFTW, FFTW3</td>
</tr>
<tr>
<td></td>
<td>Vendor-specific FFT libraries</td>
</tr>
<tr>
<td></td>
<td>DL_MG</td>
</tr>
</tbody>
</table>
Hardware

- ONETEP runs well on conventional CPU-based machines
- Scales to 1000s of cores [47]
- More cores and more memory per core are generally advantageous
- GPU-port is a work-in-progress [49]
- Can also run on desktops/laptops
Aside: DL_MG

DL_MG [37] is an open-source Poisson solver library distributed with ONETEP

- Solves several variants of the Poisson equation in real-space:
  \[
  \nabla \cdot (\varepsilon(r) \nabla \phi(r)) = -4\pi (n(r) + n_{ions}[\phi](r))
  \]

- Designed for use in large scale electronic structure calculations
- Multigrid approach [48] scales well with problem size and converges rapidly
- Iterative high-order defect-correction reduces discretization error

Enables **open BC electrostatics** and **implicit solvent** functionality in ONETEP
Aside: DL_MG

Generalized Poisson equation
\[ \nabla \cdot (\varepsilon(\mathbf{r}) \nabla \phi(\mathbf{r})) = -4\pi n(\mathbf{r}) \]

Model system
(known solution)

64 MPI processes
2 OpenMP per MPI

Total execution time demonstrates \( O(N_{\text{grid}}) \) scaling up to \( \sim 1000^3 \) grid points
Aside: DL_MG

Source code features
▶ Written in modern Fortran and built using GNU Make
▶ Code organized into Fortran modules
▶ Combines distributed (MPI) and shared memory parallelism (OpenMP)

Modularity in practice
▶ Could have been part of ONETEP, but instead built an open-source library
▶ DL_MG can be utilized by other codes
▶ Has recently been interfaced with CASTEP [36, 50] & PSI4 [51, 52]

Source code available at
www.dlmg.org
How do we manage day-to-day development?

Follow the key principles:

1. Version control
2. Testing
3. Documentation
4. Modularity
We have a Git repository hosted on Bitbucket and use a forking workflow:

- Single “official” server-side repository
- Private fork for each contributor
- Changes contributed via pull request
- Pull requests undergo review

New developers don’t need to worry about breaking the official repository!
Testing

We have a growing suite of regression tests:

- When new functionality is added, require a new regression test (“QC” test)
- Each test runs a calculation to produce known results (benchmarks)
- QC test suite is run by a Python program called `testcode`
- A test fails if outputs are outside of allowed tolerance (wrt benchmarks)

The test suite allows new builds of ONETEP to be validated.
We use the Buildbot package to automate testing of the software:

- A Buildbot instance regularly pulls the official master branch and starts a build
- Each build includes compilation, code quality checks and QC test suite
- If any step fails, core developers are notified by e-mail
Documentation

Developer documentation:

- Template module file to encourage consistent documentation
- Each new module/procedure should be documented
- Encourage developers to comment their code as they write it

User documentation:

- User documentation is required for substantial new functionality
- Distributed with the code and at www.onetep.org
- Daussault Systèmes BIOVIA provide documentation for Materials Studio
Modularity

ONETEP’s source code is organized as a loose hierarchy of modules:

- Modules encapsulate related functionality
- High-level modules depend on low-level modules
- Low-level modules provide basic functionality
- High-level modules perform sophisticated tasks

A module can often be developed without needing to consider the behaviour of the entire program.
Closing remarks
Summary

**Building ONETEP**

1. ONETEP is a tool for large-scale quantum simulation of matter
2. An extensive and expanding list of capabilities
3. Development of ONETEP follows a few key principles
4. These principles keep ONETEP usable and maintainable
A thought: Research vs. software development

There is a tension between research outputs and good software development practice:

- Often we don’t know what the solution will look like until we create it.
- This “exploratory” development can make specification and testing difficult.
- Being overly strict may inhibit the creativity needed to solve scientific problems.
- **BUT** being too relaxed leads to unmaintainable nightmare code!

Scientific software development needs to balance developer freedom and creativity against software development rigour.
Further details

Introducing ONETEP: Linear-scaling density functional simulations on parallel computers

DL_MG: A parallel multigrid Poisson and Poisson-Boltzmann solver for electronic structure calculations in vacuum and solution
Acknowledgements

Funding and compute resources

- EPSRC-funded CCP9 flagship ([EP/P02209X/1](#))
- UK MMM Hub “Thomas” ([EP/P020194/1](#))
- IRIDIS HPC Facility (Southampton)
- ARCHER UK National Supercomputing Service (UKCP consortium [EP/P022561/1](#))

Thank you for your attention!

... any questions?
Bibliography I


Bibliography V


Bibliography VIII


