Outline

• Background
• Why is open science important?
• Opening up chemistry over the last four years
• The Visualization Toolkit (VTK)
• ParaView – a client-server Qt based VTK GUI
• New frontiers – web, mobile and tablets
• Future directions
My Background

- Ph.D. (Physics) – University of Sheffield
- Google Summer of Code – Avogadro
- Postdoc (Chemistry) – University of Pittsburgh
- R&D engineer – Kitware, Inc
- Passionate about physics, chemistry, and the growing need to improve computational tools
- See the need for powerful open source, cross platform frameworks and applications
- Develop(ed): Gentoo, KDE, Kalzium, Avogadro, Open Babel, VTK, ParaView, Titan, CMake
Kitware

- Founded in 1998: 5 former GE Research employees
- 95 employees: 42% PhD
- Privately held, profitable from creation, no debt
- Rapidly Growing: >30% in 2010, 7M web-visitors/quarter
- Offices
  - Albany, NY
  - Carrboro, NC
  - Lyon, France
  - Bangalore, India

- 2011 Small Business Administration’s Tibbetts Award
- HPCWire Readers and Editor’s Choice
- Inc’s 5000 List: 2008 to 2010
Kitware: Core Technologies

- CMake
- CDash
- MEDICAL IMAGING
- 3D Slicer
- SOFTWARE PROCESS
- CMake
- CDash
- COMPUTER VISION
- OpenCV
- VXL
- MIDAS
- VTK
- ITK
- Supercalculus
What Is “Open Science”? 

“Open science is the idea that scientific knowledge of all kinds should be openly shared as early as is practical in the discovery process.”

openscience.org
What Is The Problem?

“…when the journal system was developed in the 17th and 18th centuries it was an excellent example of open science. The journals are perhaps the most open system for the dissemination of knowledge that can be constructed — if you’re working with 17th century technology. But, of course, today we can do a lot better.”

openscience.org
Opening Up Chemistry

• Computational chemistry is currently one of the more closed sciences
• Lots of black box proprietary codes
  – Only a few have access to the code
  – Publishing results from black box codes
  – Many file formats in use, little agreement
• More papers should be including data
• Growing need for open standards
Movements for Open Chemistry

• Formed an “unorganization” – Blue Obelisk
  – Published first article in 2005
  – Open data, open standards and open source
  – Meet at ACS and other conferences when possible
  – Follow-up article currently in press

• Quixote collaboration more recently
  – Provide meaningful data storage and exchange
  – Principally targeting computational chemistry
Typical Chemistry Workflow

- **Edit/Analyze**
  - Data
  - Results
  - Log File

- **Job Submission**
  - Input File
  - Local
  - Remote

- **Calculation**
  - ezhpc

- **Results**
  - Data

- **Data**
  - Input File

- **Log File**
  - Edit/Analyze

- **ezhpc**
  - Job Submission
  - Remote

- **Kitware**
  - Desktop Summit
Problem: Pretty Complex/Manual

- Most steps require user intervention
- Obtain starting structure (previous work, databases)
- Edit structure
- Write input file
- Move input file to cluster
- Submit to queue
- Wait for completion
- Retrieve input file
- Analyze output file
- Extract the relevant data, change formats
- Store results
- Repeat
Improved Chemistry Workflow
Avogadro

• Project began 2006
• Split into library and application (plugin based)
• One of very few open source editors
• Designed to be extensible from the start
• Generate input & read output from many codes
• An active and growing community
• Chemistry needs a free, open framework
Avogadro’s Roots

• Avogadro projected started in 2006
• First funded work in 2007 by Marcus Hanwell
  – Google Summer of Code student
  – Final year of Ph.D. spent the summer coding
  – Funded as part of KDE project – Kalzium editor
• Built on several other open source projects
  – Qt, Eigen, Open Babel, Blue Obelisk Data Repository
• Also uses open standards, e.g. OpenGL
• Cross platform, open source stack
Avogadro Vital Statistics

- Supports Linux, Windows and Mac OS X
- Contributions from over 20 developers
- Over 180,000 downloads over 4 years
- Translated into 19 languages
- Used by Kalzium for molecular editor
- Featured by Trolltech/Nokia,
  - Qt in use
  - Qt ambassador program
Desktop Database

- Use of “document store” NoSQL
  - Doesn’t force too much structure
    - Some entries have experimental data available
    - Some have computational jobs
  - Employ a “pile of stuff” approach
    - Can store both source and derived data
    - Calculate identifiers, QSAR properties, etc
- MongoDB is a scalable, open solution
  - Proven scaling with large web applications
Chemistry Data Explorer

- Qt application
- Connects to local or remote database
- Uses VTK for visual data exploration
- Can ingest new data
  - Uses Open Babel to generate descriptors
  - Standard InChI, SMILES, molecular weight
  - More could be added
    - All derived from files stored in the database
Chemistry Data Explorer
Database Interaction on the Web

- Avogadro directly accesses some (read-only) public databases:
  - PDB, NIH “fetch by name”
  - More could be added
- ChemData follows this approach
- Quixote aims to support both public and private sharing models – open framework
Z-Matrix/Cartesian Molecule Editor
Calling Stand Alone Programs

• Many already supported:
  • GAMESS, GAMESS-UK, Molpro, Q-Chem, MOPAC, NWChem, Gaussian, Dalton
  • Easy to add more
• Some codes writing Avogadro based custom applications,
  • Q-Chem, Molpro…
• DLPOLY author approached me:
  • Open sourced DLPOLY2, want a GUI
GAMESS Input Generation

File created by the GAMESS Input Deck Generator Plugin for Avogadro

$BASIS
GBASIS=N311 NGAUSS=6 NDFUNC=1 NPRUNC=1 DIFFSP=.TRUE. DIFFS=.TRUE. $END

$CONSTRL
SCFTYP=RHF RUNTYP=OPTIMIZE DFTYP=B3LYP $END

$STATPT
OPTTOL=0.0001 NSTEP=20 $END

$DATA
Optimization of methylbenzenethiol

Cl
C   6.0  3.39655  0.00076  -0.00682
C   6.0  1.88686 -0.00066  -0.02246
C   6.0  1.16513  1.19785  -0.01793
C   6.0  -0.22818  1.20208  -0.00139
C   6.0  -0.93378  0.00305  -0.00343
C   6.0  -0.22614 -1.20687  0.00075
C   6.0  1.16727 -1.20031  -0.01499
S   16.0  2.73763  0.00595  0.08134
H   1.0  3.77792  0.00559  1.02052
OpenQube – Quantum Data

• Reads in key quantum data
  – Basis set used in calculation
  – Eigenvectors for molecular orbitals
  – Density matrix for electron density
  – Standard geometry

• Multithreaded calculation
  – Produce regular grids of scalar data
  – Molecular orbitals, electron density…
Molecular Orbitals and Electron Density

- Quantum files store basis sets and matrices

\[
GTO = ce^{-\alpha r^2}
\]

\[
\phi_i = \sum_{\mu} c_{\mu i} \phi_\mu
\]

\[
\rho(r) = \sum_{\mu} \sum_{\nu} P_{\mu \nu} \phi_\mu \phi_\nu
\]

- Using these equations, and the supplied matrices – calculate cubes
Advanced Visualization: VTK

• New Avogadro plugin:
  • Takes volumetric data from Avogadro
  • Uses GPU accelerated rendering in VTK
• Widespread excitement from many in the chemistry community
• Several groups interested in collaborating
• Google Summer of Code project
• Leverage significant capabilities in VTK
Volume Rendered With Contours
Electron Density Volume Render
Electron Density Ray Tracing
VTK: The Toolkit

• Collection of C++ libraries
  – Leveraged by many applications
  – Divided into logical areas, e.g.
    • Filtering – data processing in visualization pipeline
    • InfoVis – informatics visualization
    • Widgets – 3D interaction widgets
    • VolumeRendering – 3D volume rendering

• Cross platform, using OpenGL
• Wrapped in Python, Tcl and Java
From Ohloh: **Very large, active development team:** Over the past twelve months, **100 developers** contributed new code to VTK. This is one of the largest open-source teams in the world, and is in the **top 2%** of all project teams on Ohloh.

and many others...
ParaView

• Parallel visualization application
• Open source, BSD licensed
• Turn-key application wrapper around VTK
• Parallel data processing and rendering
Large Data Visualization

- **BlueGene/L at LLNL**
  - 65,536 compute nodes (32 bit PPC)
  - 1,024 I/O nodes (32 bit PPC)
  - 512 MB of RAM per node

- **Sandia Red Storm**
  - 12,960 compute nodes (AMD Opteron dual)
  - 640 service and I/O nodes
  - 40 TB of DDR RAM per node
1 Billion Cell Asteroid Simulation
Tiled Displays
Parallel Processing/Rendering
3D Chemistry Visualization

• Some existing features specific to chemistry
  – Gaussian cube, PDB, and a few others
• Excellent handling of volumetric data:
  – Marching cubes
  – Volume rendering
  – Contouring
• Advanced rendering:
  – Point sprites
  – Manta – real time ray tracing
Titan: VTK and Informatics

- Led by Sandia National Laboratories
- Substantial expansion of VTK:
  - Informatics & analysis
- Actively developed, growing feature set
- Improved 2D rendering and API
- Database connectivity, client-server, pipeline based approach
- Uses web technologies such as ProtoViz
- Scalable, interactive infoviz
Manta: Real Time Ray Tracing
New Frontiers

• New work porting VTK
  – Use C++ as the common core
    • iOS port in the early stages
    • Android port
  – Use OpenGL ES 2.0 – new rendering code

• Also ParaViewWeb – delivering over web
  – Use image delivery and rendering on server
  – Also using WebGL for rendering (optionally)
Future Directions

• VTK modularization (in progress)
  – Developing more agile build systems
  – Automating more with CMake

• Using Git more fully to improve stability
  – Use of master and next
  – Topic branches - merge when ready

• Code review using Gerrit
  – Integration with continuous integration
  – Test before merge
Standard Representations
Standard Representations
Volumetric Data: Molecular Orbitals
Biomolecules
Nanomaterials
Periodic Systems
Simplified Views
Hybrid Views: CPK + MO + Ball & Stick
Linked Views of Live Data
2D: Graphs and Charts
Informatics