The Open Chemistry Project

February 2, 2013
FOSDEM

Dr. Marcus D. Hanwell
Technical Leader, Kitware, Inc.
marcus.hanwell@kitware.com
http://openchemistry.org/

Outline

- Kitware
- History
- Open Chemistry
  - Avogadro
  - MoleQueue
  - MongoChem
- Final thoughts

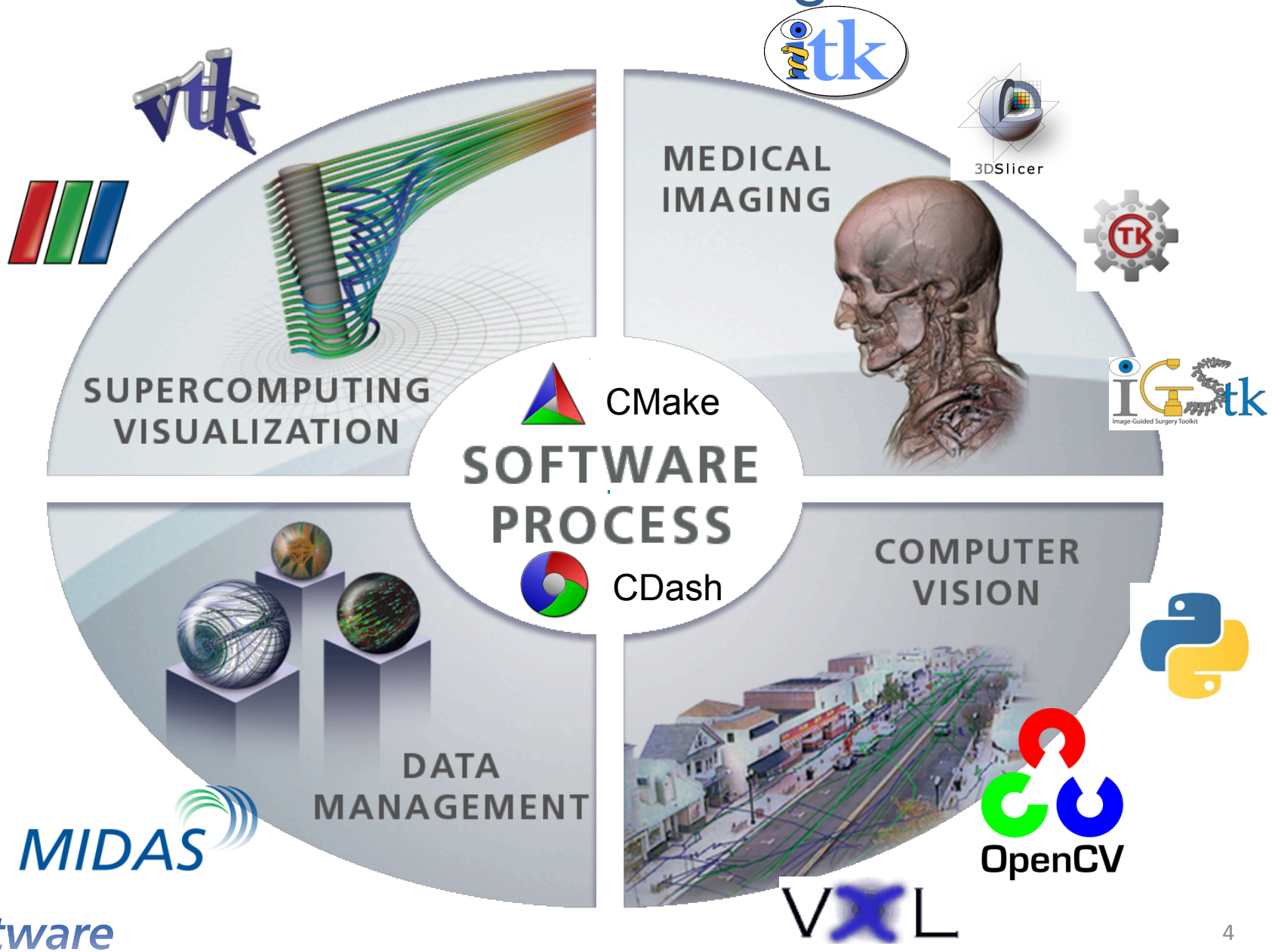
Kitware

- Founded in 1998 by 5 former GE Research employees
- 111 employees: 38 with PhDs
- Privately held, profitable from creation, no debt
- Rapidly Growing: >30% in 2011, 7M web-visitors/quarter
- Offices
  - Albany, NY
  - Carrboro, NC
  - Santa Fe, NM
  - Lyon, France

- 2011 Small Business Administration's Tibbetts Award
- HPCWire Readers and Editor's Choice
- Inc's 5000 List: 2008 to 2011

Kitware

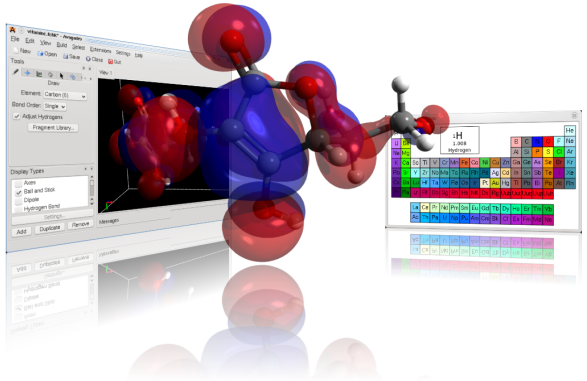
Kitware: Core Technologies



Kitware

Beginnings of Open Chemistry

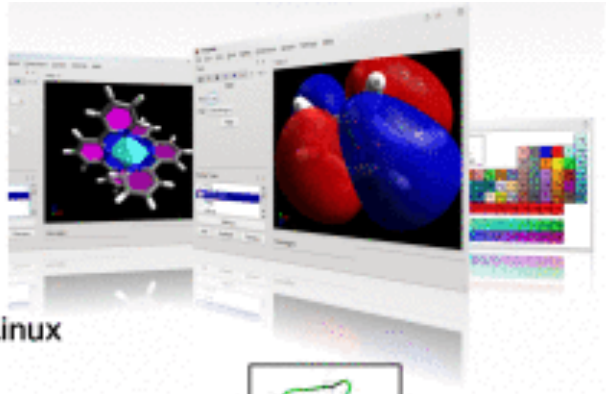
- The Avogadro project began in 2006
- One of very few open source 3D chemical editors
  - Draw/edit structure
  - Generate input for codes
  - Analyze output of codes
- Open source, GPLv2 GUI
- Google Summer of Code in 2007 (KDE)
- Used by Kalzium in KDE educational tool
- Over 300,000 downloads, 20+ translations



Avogadro Paper Published 8/13/12

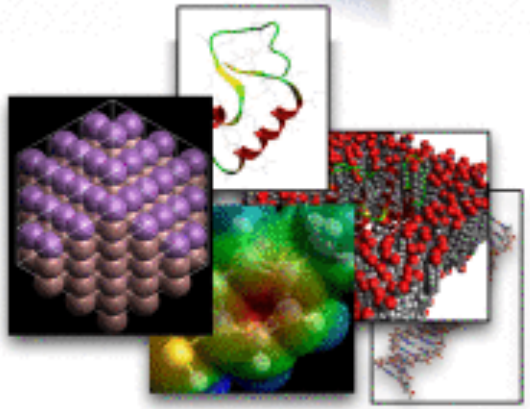
Features

Intuitive "Drawing"
Fast Optimization
Results + Analysis
20+ Languages
Windows + Mac + Linux



Extendability

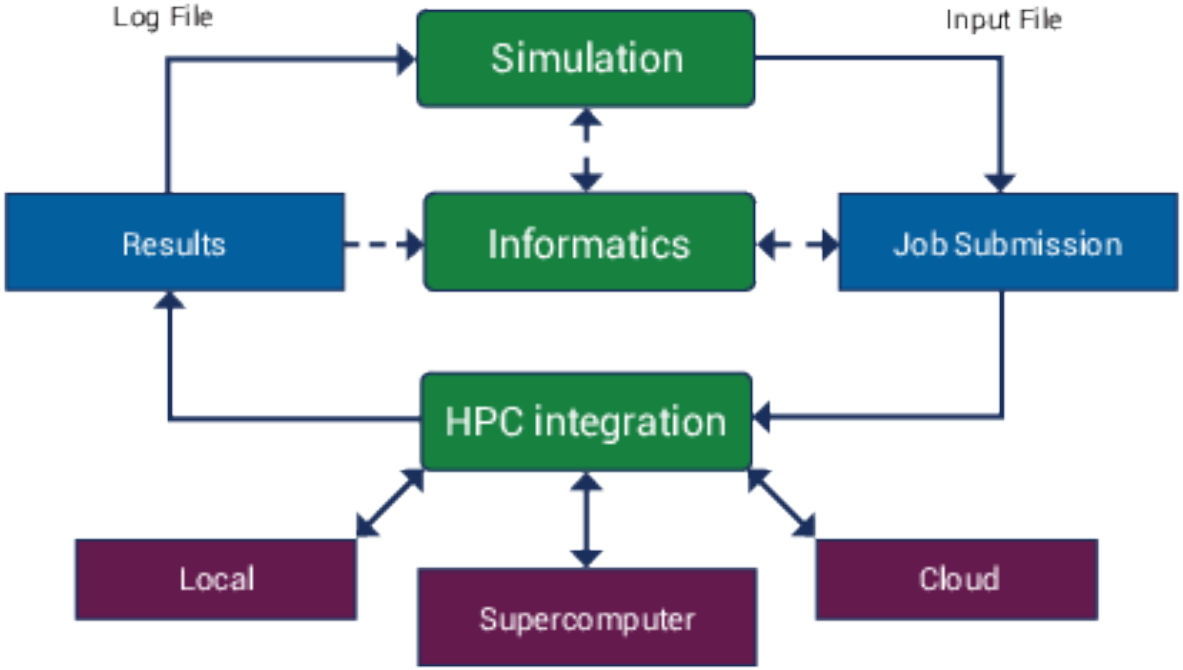
C++ Plugins
Python Scripting
Open Babel library
Input Generation for simulation packages



http://www.jcheminf.com/content/4/1/17

Introduction to Open Chemistry

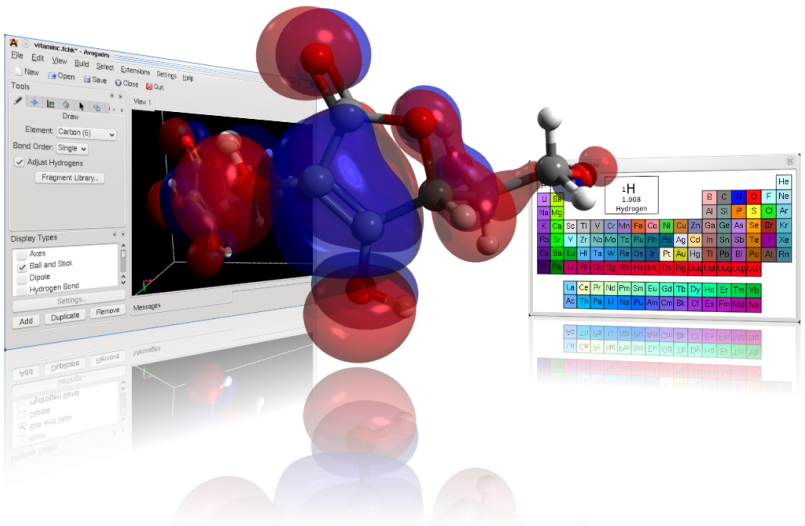
- User-friendly integration with
  - Computational codes
  - HPC/cloud resources
  - Database/informatics resources



Kitware

Vision

- Advancing the state-of-the-art
- Tight integration is needed
  - Computational codes
  - Clusters/supercomputers
  - Data repositories
  - Reduce, reuse, recycle!
- Facilitate sharing and searching of data
- Embracing data-centric workflows



Overview

- Desktop chemistry application suite
  - 3D structure editor, pre- and post-processing
  - HPC integration – easily run codes
  - Cheminformatics to store, index, and analyze
- Each can work independently
  - Enhanced functionality when used together
  - One-click HPC job submission
  - Easily open structure found in database
  - Coordination of job submission

Open Chemistry Project Approach

- Open approach to chemistry software
  - Open source frameworks
  - Developed openly
  - Cross platform
  - Tested, verified
  - Contribution model
  - Supported by Kitware experts
- BSD licensed to facilitate research/reuse

Open Chemistry Development Team

- Assembled an inter-disciplinary team
- Domain specialists: quantum chemistry, biology, solid-state materials
- Computer scientists: build systems, queuing, graphics, process
- Marcus, Kyle, David, and Chris.




Kitware

OpenChemistry.org

- Central site to promote Open Chemistry
- Hosting of project-specific pages
- Providing an identity for related projects
- Promoting shared ownership of projects
  - Website
  - Code submission/review
  - Testing infrastructure
  - Wiki, mailing lists, news, galleries

Open Chemistry

Kitware

Search

PROJECTS RESOURCES HELP OPEN SOURCE

The **Open Chemistry** project is a collection of open source, cross platform libraries and applications for the exploration, analysis and generation of chemical data. The project builds upon various efforts by collaborators and innovators in open chemistry such as the Blue Obelisk, Quixote and the associated projects. We aim to improve the state of the art, and facilitate the open exchange of ideas and exchange of chemical data leveraging the best technologies ranging from quantum chemistry codes, molecular dynamics, informatics and visualization.

News **More News >**

08.17.2012 Avogadro Featured in Journal of Cheminformatics

01.24.2012 Kitware Receives Phase II Funding for the Development of a Comput...

Events

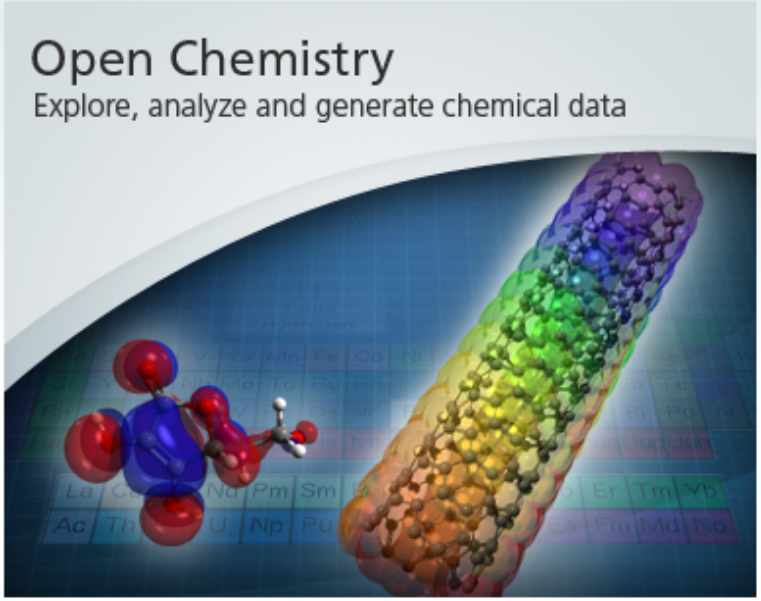
11.10.2012 Supercomputing 2012

Blog Posts

10.24.2012 Open Access Week and the Scholarly Poor

09.07.2012 Open Chemistry at the American Chemical Society Meeting

08.20.2012 PhDelta: Guest Post on Nature's Soapbox Science



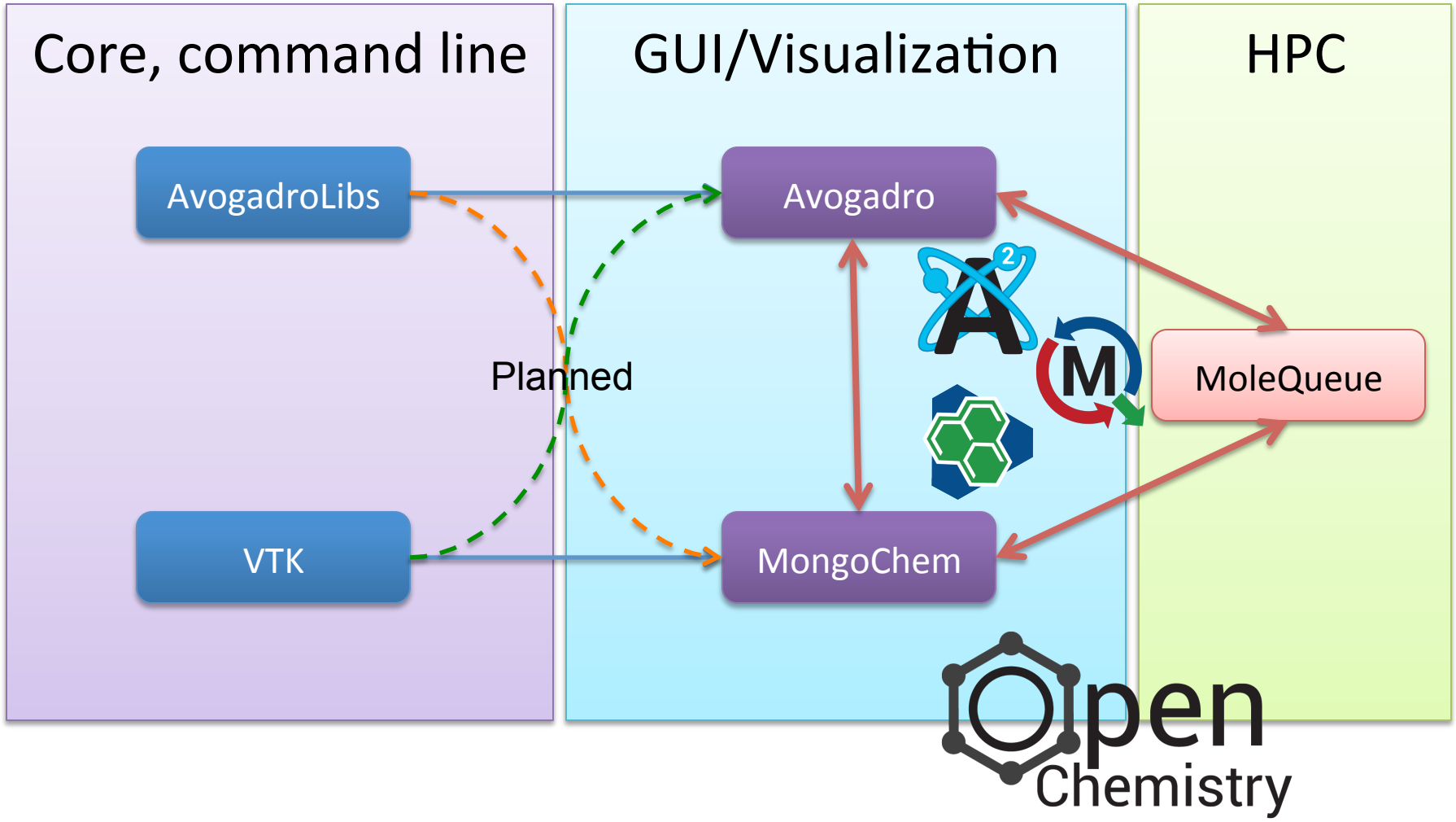
Applications Being Developed

- Three independent applications
- Communication handled with local sockets
- Avogadro 2 – structure editing, input generation, output viewing and analysis
- MoleQueue – running local and remote jobs in standalone programs, management
- MongoChem – Storage of data, searching, entry, annotation

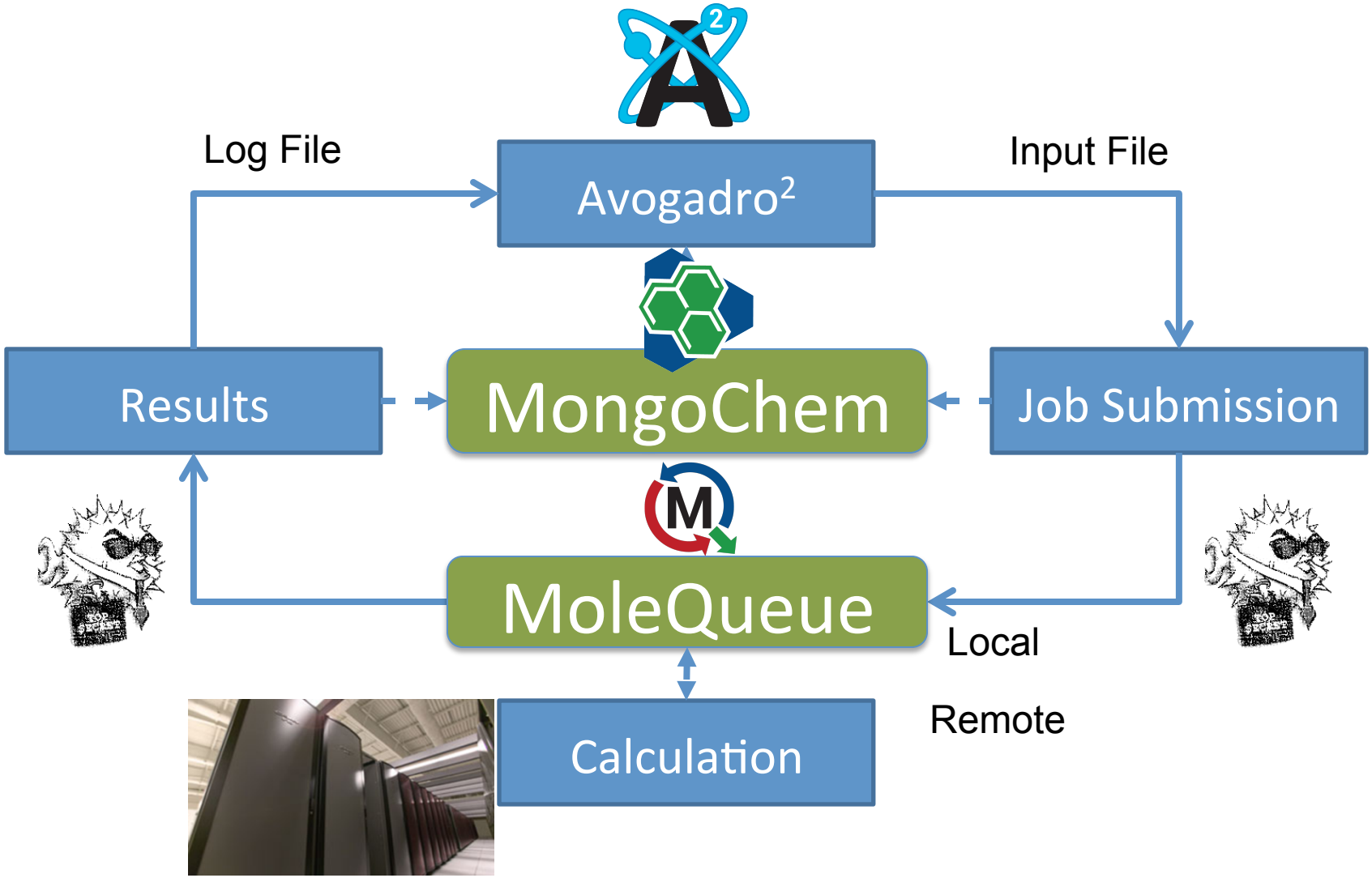
Open Frameworks

- AvogadroLibs – core data structures and algorithms shared across codes
  - Split into dedicated libraries, e.g. core, io, rendering, qtgui, qtopengl, qtplugins, quantum
  - Core maintains a minimal dependency set
  - Intended for use on server, command line, and in a full-blown desktop application
- VTK – specialized chemistry visualization/ data structures, use of above

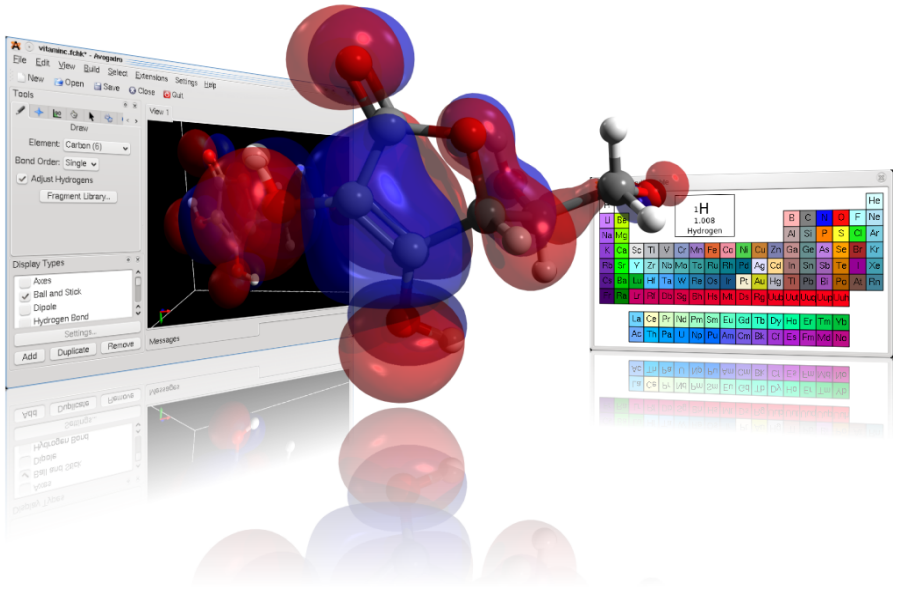
Project Diagram: Libraries/Apps



Workflow in Open Chemistry



Avogadro^2



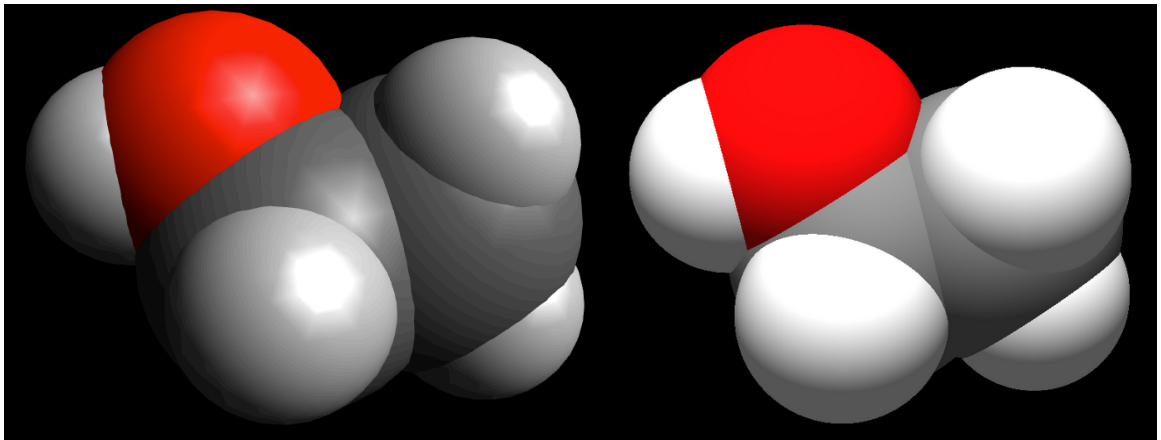
- Rewrite of Avogadro
- Split into libraries and application (plugin based)
- Still one of very few open source **editors**
- Still using Qt, C++, Eigen, OpenGL, CMake
- Use AvogadroLibs for core data
- Introduce client-server dataflow/patterns
- New, efficient rendering code
- More liberally licensed – from GPL to BSD

Avogadro: Visualization

- GPU accelerated rendering
- VTK for advanced visualization
- Support for 2D and 3D plots of data
- Optimized data structures
  - Large data
  - Streaming
- Reworked interface
  - Tighter database/workflow integration

Advanced Impostor Rendering

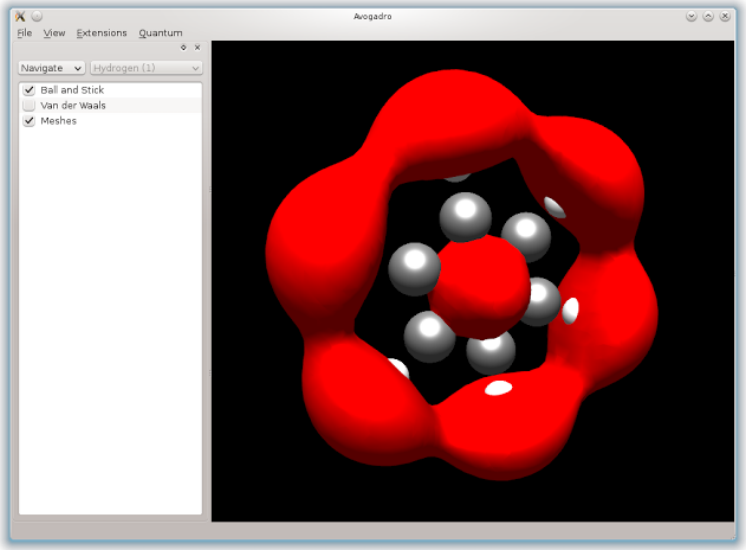
- Using a scene, vertex buffer objects, and OpenGL shading language
- Impostor techniques
  - Sphere goes from 100s of triangles to 2!
  - No artifacts from triangulation
  - Scales to millions of spheres on modest GPU



Kitware

Electronic Structure Visualization

- Read quantum output files
  - Calculate cubes for molecular orbitals
  - Show isosurface or volume rendering
  - Multithreaded C++ code to perform calculations – scales very well



Kitware

Scriptable Simulation Input Generator

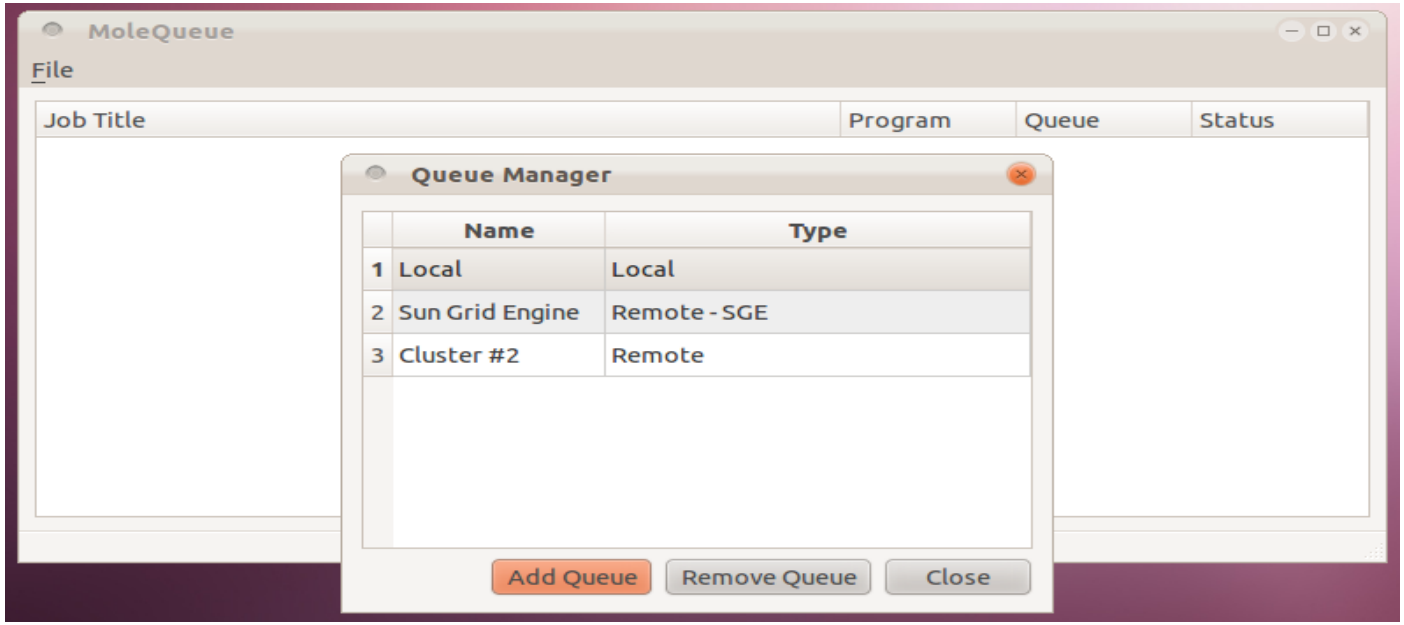
- Previous input generators were C++
- Execute a simple Python script
  - Script can output JSON with parameters
  - Input is parameters specified by user
  - Chemical JSON with full structure
- New input generator is as simple as adding a new Python script
  - Implement 2-3 entry points and done

MoleQueue: Job Management

- Tighter integration with remote queues
- Integration with databases
  - Retain full log of computational jobs
  - Trigger actions on completion
- Plugin based system
  - Easy addition of new codes
  - Easy addition of new queue systems
- Provide client API for applications

MoleQueue

- Support configuration for a variety of remote clusters and queuing software



Kitware

MoleQueue: Queue Types

- Several transports implemented
  - Command line SSH/plink (Windows)
  - libssh2 (experimental)
  - HTTPS (SOAP)
- Several queue types
  - Sun Grid Engine
  - PBS
  - UIT (ezHPC with largely PBS dialect)

Using JSON

- MongoDB stores data as BSON
  - JSON: JavaScript Object Notation
  - BSON: Binary form, type safe
- JSON is very compact, standardized

```
{
  “name”: “water”,
  “atoms”: {
    “elementType”: [“H”, “H”, “O”],
  }
  “properties”: { “molecular weight”: 18.0153 }
}
```

JSON-RPC interface

Applications can submit jobs via a local socket or ZeroMQ connection:

Client request:

```
{ "jsonrpc": "2.0",
  "method": "submitJob",
  "params": {
    "queue": "Remote cluster PBS",
    "program": "MOPAC",
    "description": "PM6 H2 optimization",
    "inputAsString": "PM6\n\nH 0.0 0.0 0.0\nH 1.0 0.0 0.0\n"
  },
  "id": "XXX” }
```

Server reply:

```
{ "jsonrpc": "2.0",
  "result": {
    "moleQueueId": 17,
    "queueId": 123456,
    "workingDirectory": "/tmp/MoleQueue/17/"
  },
  "id": "XXX” }
```

Kitware

Chemical JSON

- Stores molecular structure, geometry, identifiers, and descriptors as a JSON object
- Benefits:
  - More compact than XML/CML
  - Native language of MongoDB and JSON-RPC
  - Easily converted to a binary representation (BSON)


```
{
  "chemical json": 0,
  "name": "ethane",
  "inchi": "1/C2H6/c1-2/h1-2H3",
  "formula": "C 2 H 6",
  "atoms": {
    "elements": {
      "number": [  1,   6,   1,   1,   6,   1,   1,   1 ]
    },
    "coords": {
      "3d": [  1.185080, -0.003838,  0.987524,
               0.751621, -0.022441, -0.020839,
               1.166929,  0.833015, -0.569312,
               1.115519, -0.932892, -0.514525,
              -0.751587,  0.022496,  0.020891,
              -1.166882, -0.833372,  0.568699,
              -1.115691,  0.932608,  0.515082,
              -1.184988,  0.004424, -0.987522 ]
    }
  },
  "bonds": {
    "connections": {
      "index": [ 0, 1,
                 1, 2,
                 1, 3,
                 1, 4,
                 4, 5,
                 4, 6,
                 4, 7 ]
    },
    "order": [ 1, 1, 1, 1, 1, 1, 1 ]
  },
  "properties": {
    "molecular weight": 30.0690,
    "melting point": -172,
    "boiling point": -88
  }
}
```

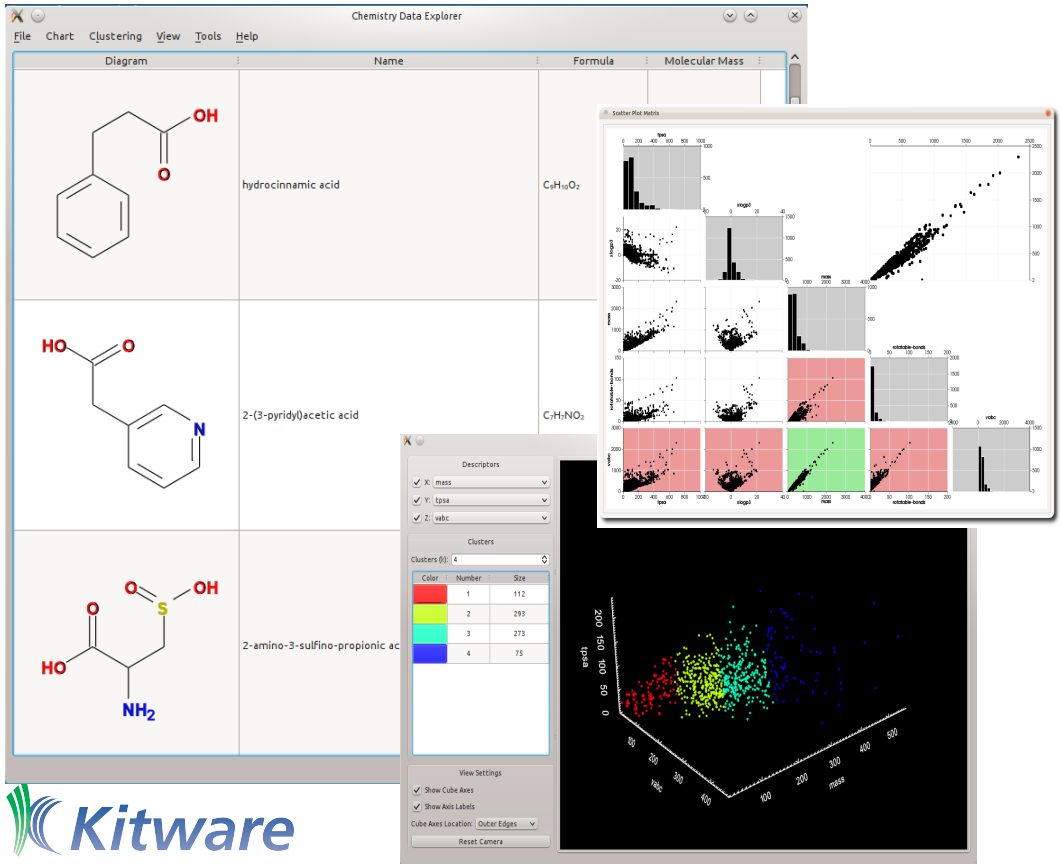

Kitware

MongoChem Overview

- A desktop cheminformatics tool
  - Chemical data exploration and analysis
  - Interactive, editable, and searchable database
- Leverages several open-source projects
  - Qt, VTK, MongoDB, Chemkit, Open Babel
- Designed to look at many molecules
- Spot patterns, outliers, run many jobs
- Scaling studies with ~3 million structures

Architecture Overview

- Native, cross-platform C++ application built with Qt
- Stores chemical data in a NoSQL MongoDB database
- Uses VTK for 2D and 3D dataset visualization



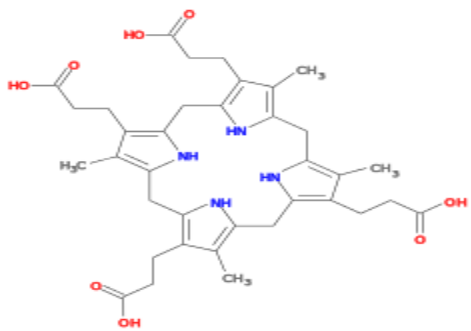
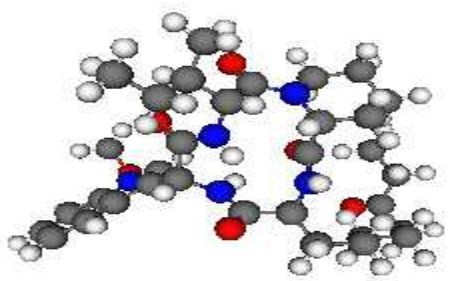
ParaViewWeb and Open Chemistry

paraviewweb.kitware.com/OpenChemistry/

Open Chemistry

Query our Open Database

Name ▾ contains ▾ Search

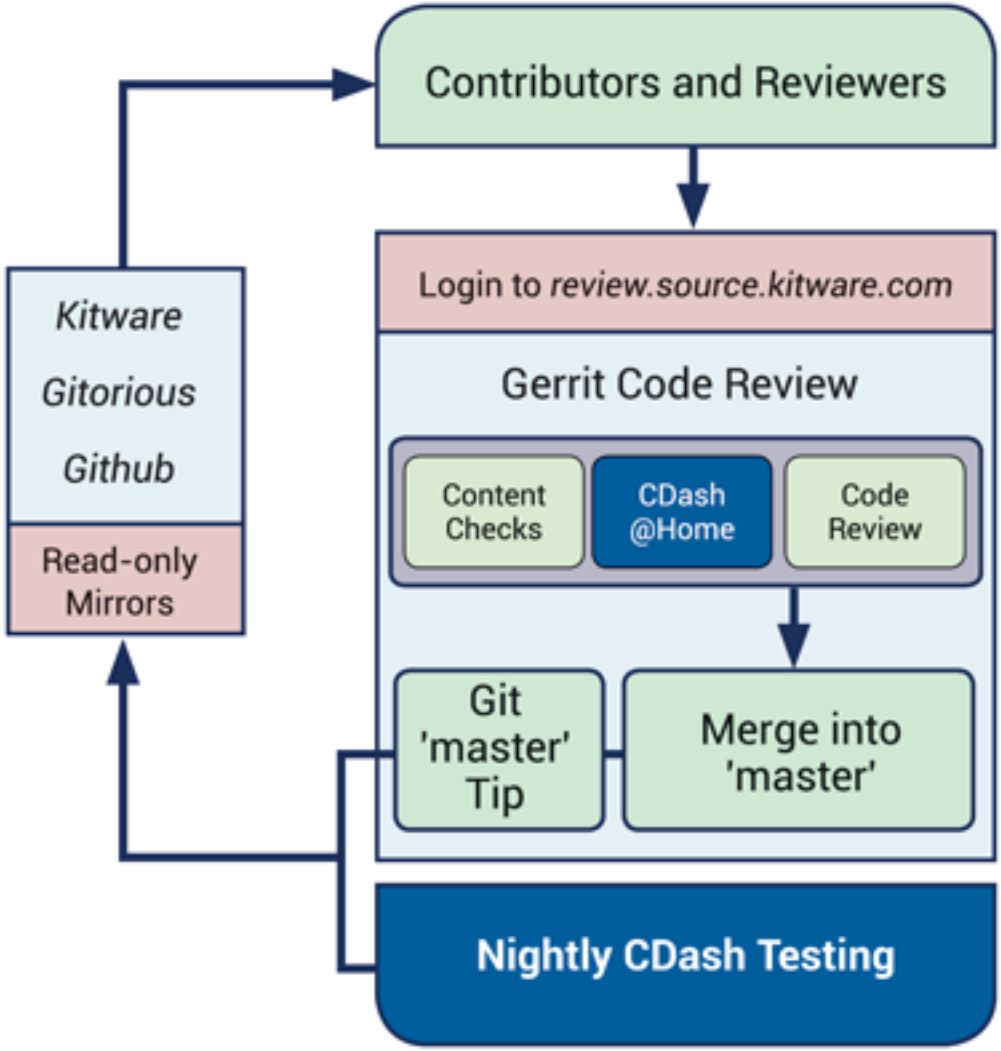
Molecule	Information
2D 3D 	**3-[8,12,17-tris(2-carboxyethyl)-3,7,13,18-tetramethyl-5,10,15,20,21,22,23,24-octahydroporphin-2-yl]propionic acid** **Formula:** $C_{36}H_{44}N_4O_8$ **Mass:** 660.75656 **InChi:** InChI=1S/C36H44N4O8/c1-17-21(5-9-33(41)42)29-14-27-19 **InChiKey:** NIUVHXTXUXOFEB-UHFFFAOYSA-N Fullscreen 3D
2D 3D 	**9-(6-ketooctyl)-6-(1-methoxyindol-3-yl)-3-sec-butyl-1,4,7,10-tetrazabicyclo[10.4.0]hexadecane-2,5,8,11-diquinone** **Formula:** $C_{33}H_{47}N_5O_6$ **Mass:** 609.75618 **InChi:** InChI=1S/C33H47N5O6/c1-5-21(3)28-33(43)37-19-13-12-18 **InChiKey:** XWKJTSOFFKCRMH-UHFFFAOYSA-N Fullscreen 3D

Kitware

Software Process

- Source code publicly hosted using Git
- Gerrit for code review
- CTest/CDash for testing/summary
  - Gerrit can use CDash@Home
    - Test proposed changes before merge
- CDash can now provide binaries
  - Built nightly, available for direct download
- Wiki, mailing list, bug tracker

Software Process



Final Thoughts

- Real opportunity to make an impact
- Bringing best practices to chemistry
- Improve research, industry and teaching
- Semantic data at the center of our work
  - Storage
  - Search
  - Interaction with computational codes
  - Comparison with experimental data
- Liberal, BSD licensed, cross platform codes